ABSTRACT

Title of Document:	**MULTIPLE TESTING PROCEDURES FOR THE ANALYSIS OF MICROARRAY DATA.**
	Ayala Kimel Nuriely, Master of Arts, 2013
Directed By:	**Professor Paul J. Smith, Statistics Program, Department of Mathematics**

We reviewed literature about various multiple testing techniques, especially addressing microarray analyses and small sample sizes, and reanalyzed data from Yuen et al. (*Physiological Genomics*, 2006) which compared the effect of $HgCl_2$ and Ischemia/Reperfusion injuries on rat kidney tissues. Our analysis uses only 22 rats with small numbers of rats in each treatment group, and 9,501 genes under study. We used empirical Bayes (EB) and permutation testing (implemented in Bioconductor) in an effort to identify differentially expressed genes. EB identified a large number of genes as differentially expressed, including both previously identified and newly identified genes. The newly identified genes appear to have biological functions similar to those previously identified. We also recognized power differences between EB and permutation tests, possibly due to nonnormality of the data but also because permutation tests do not make use of all available information in the data.

MULTIPLE TESTING PROCEDURES FOR THE ANALYSIS OF MICROARRAY DATA

By

Ayala Kimel Nuriely.

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2013

Advisory Committee:
Professor Paul J. Smith, Chair
Professor Abram Kagan
Professor Benjamin Kedem

Dedication

This thesis is dedicated to my amazing kids, Ron and Liya, to whom I gave birth in the course of completing my degree, and to my loving husband, Avi, for his continued support.

Acknowledgements

I would like to thank Dr. Robert Star, for allowing me access to the data and for the kind advice provided in the course of analyzing the data.

I would also like to thank Prof. Paul Smith, for the truly enjoyable process of analysis and writing of the thesis, for the guidance, teaching, advice and mainly a lot of patience and support.

Table of Contents

List of Tables

List of Figures

Chapter 1: Overview

Statistical methodology has advanced over the last 10 years in handling problems of multiplicity which inevitably arise in genomic research. In this context, a large number of hypothesis tests (typically thousands) is carried out, potentially leading to a large number of falsely significant results due to an increased chance of committing at least one false positive, that is, at least one Type I error. Small unadjusted *p*-values, which would lead to the rejection of a single hypothesis, may no longer correspond to significant findings. To control this multiplicity effect, classical multiple comparison procedures aim to control the probability of committing one or more type I error in families of comparisons under simultaneous consideration. In this thesis, we are going to review that literature, emphasizing recent developments, and we will apply some of the new multiple testing methodology to an available dataset.

This dataset consists of microarray analyses from kidney tissues obtained from 31 Sprague-Dawley lab rats which were submitted to mercuric chloride ($HgCl_2$) and ischemia reperfusion (IR) treatments, with approximately 10,000 genes per rat.

The dataset had been previously analyzed by Yuen et al. (2006) using a combination of ad hoc data screening techniques and off-the-shelf statistical analyses. We will take up the analysis of these data using several different approaches which have been incorporated in the R Bioconductor ensemble. The **limma** approach assumes normally distributed data, and relies on a combination of normal theory analysis of linear models together with an empirical Bayes approach and other shrinkage methods which borrow information across genes to stabilize the analyses even for experiments with small number of arrays (Smyth, 2004). The **multtest** approach is permutation-based, and its algorithm makes no assumptions about the data distribution (Pollard, Dudoit and van der Laan, 2004).

Our **limma** analysis identified a large number of differentially expressed genes, many of which were not identified by previous workers. On the other hand, the **multtest** approach identified very few differentially expressed genes. It appears as if the **limma** approach has much more power than has the **multtest** approach.

Future chapters will review the literature of statistical methodology for handling multiple testing (Chapter 2), present our analysis and results (Chapter 3) and then summarize our findings and discuss our conclusions (Chapter 4). The Appendix includes detailed documentation and numerical results of our differentially expressed gene findings, as produced by the **limma** and **multtest** procedures.

Chapter 2: Literature Review

Statistical methodology has advanced over the last 10 years in handling problems of multiplicity which inevitably arise in genomic research. In this context, a large number of hypothesis tests (typically thousands) is carried out, potentially leading to a large number of falsely significant results due to an increased chance of committing at least one false positive, that is, at least one Type I error. Small unadjusted *p*-values, which would lead to the rejection of a single hypothesis, may no longer correspond to significant findings. To control this multiplicity effect, classical multiple comparison procedures aim to control the probability of committing one or more type I error in families of comparisons under simultaneous consideration.

The dataset analyzed in this thesis is a microarray analysis of kidney tissues obtained from 31 Sprague-Dawley lab rats which were subjected to mercuric chloride ($HgCl_2$) or ischemia reperfusion (IR) treatments, with approximately 10,000 genes per rat.

In this Chapter, we review literature about the analysis of genomic data with a small amount of background from Biology. In subsequent sections, we will review literature in the following topics: bioinformatics background; mathematical framework; permutation-based versus bootstrap-based tests (as advocated by Dudoit et al. (2004)), and packaged by the Bioconductor project into **`multtest`**); Smyth's (2004) article – the basis for **`limma`**, where by fitting a linear model to the expression data for each gene, this package allows analyses of contrasts of interest analysis. Empirical Bayes and other shrinkage methods are used to borrow information across genes making the analyses stable even for experiments with small number of array; Overview of Multiple Testing Procedures; The Bioconductor project; **`Multtest`** package which mechanizes Dudoit's approach and **`Limma`** package which mechanizes Smyth's approach.

In this chapter, we make extensive use of the following two books: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* by Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry and Sandrine Dudoit; and *Multiple Testing Procedures with Applications to Genomics* by Sandrine Dudoit and Mark J. van der Laan. Therefore these two books will not be cited individually. Other references will be cited as needed.

2.1 Bioinformatics Background

Microarray technology takes advantage of hybridization properties of nucleic acid and uses complementary molecules attached to a solid surface, referred to as probes, to measure the quantity of specific nucleic acid transcripts of interest that are present in a sample, referred to as the target. The molecules in the target are labeled (via a cascade of biochemical reactions), and a specialized scanner is used to measure the amount of hybridized target at each probe, which is reported as an intensity. The biochemical reactions and optical detection are performed in parallel, allowing up to a million measurements on one array. The raw or probe-level data are the intensities read for each of these components.

Subtle variations between arrays, in the reagents used, and in the environmental conditions lead to slightly different measurements even for the same sample. The effects of these variations may be divided into two classes. Systematic effects affect a large number of measurements (for examples, the measurements for all probes on one array, or the measurements from one probe across several arrays) simultaneously. Such effects can be identified and approximately removed. Other kinds of effects are completely random, with no well understood pattern. These effects are commonly called stochastic components or noise.

Complementary DNA (cDNA) microarrays are used to compare gene expression in different samples of cells, and they permit us to study the expression of thousands of genes simultaneously. They are now used in many different contexts to compare mRNA levels between two or more samples of cells. The technique has a wide range of applications including learning how genes interact, which genes are used in different cell types, and which genes change their expression in cells due to disease or drug stimuli. Microarray experiments typically yield expression measurements on a large number of genes, on a scale of 10,000-20,000, but with few, if any, replicates for each gene. [Lönnstedt and Speed, (2002)].

Fundamental to the task of analyzing gene expression (microarray) data is the need to identify genes whose patterns of expression differ according to phenotype (e.g., normal cell vs. cancer cell) or experimental condition. Gene expression is a wellcoordinated system, and hence measurements on different genes are in general not statistically independent. Given more complete knowledge of the specific interactions and transcriptional controls, it is conceivable that meaningful comparisons between samples can be made by considering the joint distribution of specific sets of genes. However, the high dimension of gene expression space prohibits a comprehensive exploration, while the fact that that our understanding of biological systems is only at its infancy means that in many cases we do not know which relationships are important and should be studied. In current practice, differential gene expression analysis will therefore at least start with a gene-by-gene approach, ignoring the dependencies among genes.

A simple approach is to select differentially expressed genes using a fold-change (ratio of intensities) criterion. This may be the only possibility in cases where no, or very few replicates, are available. An analysis solely based on fold change, however, does not allow the assessment of significance of expression differences in the presence of biological and experimental variation, which may differ from gene to gene. This is the main reason for using statistical tests to assess differential expression.

Generally, one may look at various parameters of the distributions of a gene's expression levels under different conditions, though most often location parameters of these distributions, such as the mean or the median, are considered. One may distinguish between parametric tests, such as the *t*-test, and nonparametric tests, such as the Mann-Whitney test or permutation tests.

Parametric tests usually have a higher power if the underlying model assumptions, such as normality in the case of the *t*-test, are at least approximately satisfied. Nonparametric tests do have the advantage of requiring less stringent assumptions on the data generating distribution. In many microarray studies, however, a small sample size leads to insufficient power for nonparametric tests. A pragmatic approach in these situations is to employ parametric tests, but to use the resulting p-values cautiously to rank genes by their evidence for differential expression.

When performing statistical analysis of microarray data, an important question is determining on which scale to analyze the data. Often the logarithmic scale is used in order to make the distribution of replicated measurements per gene roughly symmetric and close to normal. A variance stabilizing transformation derived from an error model for microarray measurements may be employed to make the variance of the measured intensities independent of their expected value. This can be advantageous for gene-wise statistical tests that rely on variance homogeneity, because it will diminish differences in variance between experimental conditions that are due to differences in the intensity level. Of course, differences in variance between conditions may also have gene-specific biological reasons, and these will remain untouched. One or two group *t*-test comparisons, multiple group ANOVA, and more general trend tests are all instances of linear models that are frequently used for assessing differential gene expression.

The approach of conducting a statistical test for each gene is popular, largely because it's relatively straightforward and a standard repertoire of methods can be applied.

However, this unguarded use of single-inference procedures results in a greatly increased false positive (Type I error) rate. A large number of hypothesis tests (typically thousands) is carried out, potentially leading to a large number of falsely significant results due to an increased chance of committing at least one false positive, that is, at least one Type I error. Small unadjusted *p*-values, which would lead to the rejection of a single hypothesis (e.g., $p = 0.001$), may no longer correspond to significant findings.

To control this multiplicity (selection) effect, classical multiple comparison procedures (MCP's) aim to control the probability of committing one or more type I error in families of comparisons under simultaneous consideration. The control of this familywise error rate (FWER) is usually required in a strong sense, that is, under all configurations of true and false hypotheses.

A common criticism of multiple testing procedures designed to control parameters of the distribution of the number of Type I errors (e.g., FWER) is their lack of power, especially for large-scale testing problems such as those encountered in biomedical and genomics research. In many situations, control of the FWER can lead to unduly conservative procedures. In microarray experiments, thousands of tests are performed simultaneously and a fairly large proportion of null hypotheses are expected to be false. In this context, one may be prepared to tolerate Type I errors, provided their number is small in comparison to the number of rejected hypotheses. Error rates based on the proportion of false positives among the rejected hypotheses are especially appealing for large scale testing problems, compared to error rates based on the number of false positives (FWER), as they remain stable with an increasing number of tested hypotheses. These considerations have led Benjamini and Hochberg (1995), Genovese and Wasserman (2004a, b), Korn et al. (2004), van der Laan et al. (2004b, 2005), Lehmann and Romano (2005), Romano and Wolf (2005), among others, to consider controlling parameters of the proportion of false positives (PFP) among the rejected hypotheses. For current high-dimensional applications, this less stringent Type I error control may be more appropriate than error rates based on the absolute number of Type I errors and therefore presents promising alternatives to FWER-controlling approaches.

2.2 Mathematical Framework

Let $\underset{\sim}{\mathbf{X}} = \{\underset{\sim}{\mathbf{X}}_{\mathbf{1}}, ..., \underset{\sim}{\mathbf{X}}_{\mathbf{k}}\}$ be a random sample of *k* independent random vectors, where the data generating distribution *P* is an element of a particular statistical model $\mathcal{M}$. In a microarray experiment each $\underset{\sim}{\mathbf{X}}_{\mathbf{i}}$, *i=1, …, k* is a vector of gene expression measurements, which we observe for each of *k* arrays.

Define *N* null hypotheses $H_0(n) \equiv I[P \in \mathcal{M}(n)]$ in terms of a collection of submodels, $\mathcal{M}(n)$, *n=1, …., N,* for the data generating distribution *P.*

In a microarray experiment, $H_0(n)$ would state that gene *n* is equally expressed on each of the *k* arrays. A testing procedure is a data-driven rule for deciding whether or not to reject each of the *N* null hypotheses $\mathrm{H}_0(n)$ based on an *N*-vector of test statistics, $T_k = (T_k(n)$: n=1, …., N), which are functions of the observed data. Denote the typically unknown (finite sample) joint distribution of the test statistic T_k by $Q_k = Q_k(P)$.

A multiple testing procedure (MTP) provides rejection regions, $C_k(n)$, that is, sets of values for each test statistic $T_\mathrm{k}(n)$, that lead to the decision to reject the null hypothesis $H_0(n)$. In other words, an MTP produces a random (i.e., data-dependent) subset R_k of rejected hypotheses that estimates the set of true positives,

$R_k = R(T_{k,} Q_{ok,} \alpha) \equiv \{n : H_0(n)$ is rejected $\} = \{ n : T_k(n) \in C_k(n)\}$,

where the long notation $R(T_{k,} Q_{ok,} \alpha)$ emphasizes that the MTP depends on (i) the data through the test statistics T_k; (ii) the (estimated) null distribution, Q_o , of the test statistic T_k, which is used to derive rejection regions; and (iii) the nominal level α, that is, the desired upper bound for a suitably defined Type I error rate.

Given an MTP $R_k(\alpha) = R(T_{\mathrm{k},} Q_{ok,} \alpha)$, the unadjusted p-value $P_{0k}(n) = P(T_k(n), Q_{0,n})$, for the single test of null hypothesis $H_0(n)$, is defined as

$$P_{0k}(n) \equiv \inf \{ \alpha \in [0,1] : \text{Reject } H_0(n) \text{ at single test nominal level } \alpha\}$$

$$= \inf \{ \alpha \in [0,1] : T_\mathrm{k}(n) \in C_k(n) \}, \quad n=1, …., N.$$

That is, the unadjusted p-value $P_{0k}(n)$, for null hypothesis $H_0(n)$, is the smallest nominal type I error level of the single hypothesis testing procedure at which one could reject $H_0(n)$, given $T_k(n)$. Unadjusted p-values may also be referred to as marginal or raw p-values.

Let $O_k(n)$ denote the indices for the ordered unadjusted p-values, $P^{\circ}_{0k}(n) \equiv P_{0k}(O_k(n))$, so that $P_{0k}(O_k(1)) \leq \cdots \leq P_{0k}(O_k(N))$.

Given an MTP $R_k(\alpha) = R(T_k, Q_{ok}, \alpha)$, the adjusted p-value $\tilde{P}_{ok}(n) = \tilde{P}(T_k, Q_{ok})(n)$ for null hypothesis $H_0(n)$, is defined as

$$\tilde{P}_{ok}(n) \equiv \inf \{ \alpha \in [0,1] : H_0(n) \text{ is rejected at nominal MTP level } \alpha\}$$

$$= \inf \{ \alpha \in [0,1] : n \in R_k(\alpha)\}$$

$$= \inf \{ \alpha \in [0,1] : T_k(n) \in C_k(n) \}, \quad n=1, \ldots, N.$$

That is, the adjusted p-value $\tilde{P}_{ok}(n)$, for null hypothesis $H_0(n)$, is the smallest nominal type I error level of the multiple hypothesis testing procedure at which one could reject $H_0(n)$, given $T_k(n)$.

Let V_k be the number of Type I errors (false positives, or rejected null hypotheses that are true), while R_k is the number of rejected hypotheses.

Then the generalized family-wise error rate (gFWER), or probability of at least (r+1) Type I errors is:

$$\text{gFWER}\ (r) \equiv \Pr\ (V_k > r).$$

When $r=0$, the gFWER is the usual family-wise error rate (FWER), or probability of at least one Type I error: FWER $\equiv$ Pr ($V_k > 0$).

The tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses are defined by

$$\text{TPPFP}(q) \equiv \Pr\ (V_k / R_k > \text{q}), \quad 0<q<1\ .$$

The false discovery rate (FDR), or the expected value of the proportion of false positives among the rejected hypotheses (Benjamini and Hochberg, 1995) is

$$\text{FDR} \equiv E[V_k\,/\,R_k]\ . \qquad\qquad (V_k\,/\,R_k \text{ is defined as 0 if } R_k = 0).$$

The false discovery rate (FDR) is the expected value $E[V_n/R_n]$ of the proportion of false positives among the rejected hypotheses, while TPPFP controls tail probabilities for the proportion of false positives, $\Pr(V_n/R_n > q)$ among the rejected hypotheses.

Error controls based on the proportion of false positives (e.g., TPPFP and FDR) are especially appealing for large-scale testing problems such as those encountered in genomics, compared to error control based on the number of false positives (e.g., gFWER), as they do not increase rapidly with the number of tested hypotheses.

One usually distinguishes between two classes of MTPs, single-step and stepwise procedures, depending on whether the rejection regions for the test statistics are constant or random (given a test statistic null distribution or an estimator thereof), that is, whether or not they are independent of the data.

In **single-step procedures**, each null hypothesis is tested using a rejection region that is independent of the results of the tests of other hypotheses and thus not a function of the data X_n. An example is the Bonferroni procedure. Controlling the FWER at level *α,* the single-step Bonferroni procedure rejects any null hypothesis, $H_0(n)$ with unadjusted p-value $P_{ok}(n)$ less than or equal to the common single-step cut-off $a_n(\alpha) \equiv \alpha / N$.

Improvement in power, while preserving Type I error control, may be achieved by **stepwise procedures**, in which the decision to reject a particular null hypothesis depends on the outcome of the tests of other hypotheses. That is, the test procedure is applied to a sequence of successively smaller nested random (i.e., data-independent) subsets of ordered null hypotheses, defined by the ordering of the test statistics (common cut-off MTPs) or unadjusted p-values (common-quantile MTPs). The rejection regions are therefore allowed to depend on the data X_n via the test statistics T_n. An example is the Holm (1979) procedure. For controlling the FWER at level *α,* the unadjusted p-value cut-offs for the step-down Holm (1979) procedure are as follows,

$$a_n(\alpha) \equiv \frac{1}{N-n+1}\alpha, \qquad n= 1, \ldots, N,$$

and the set of rejected null hypotheses is

$$R_k(\alpha) \equiv \left\{ O_k(n) : \tilde{P}_{ok}(O_k(h)) \le \frac{1}{N-h+1}\alpha, \forall h \le n \right\}.$$

The corresponding adjusted p-values are thus given by

$$\tilde{P}_{ok}(O_k(n)) = \max_{h=1, \ldots, n} \left\{ \min\left\{ (N-h+1) P_{ok}(O_k(h)), 1 \right\} \right\} \qquad n=1, \ldots, N.$$

In *step-down* procedures, the *most significant* null hypotheses (i.e., the null hypotheses with the largest test statistics for common-cut-off MTPs or smallest unadjusted p-values for common-quantile MTPs) are considered successively, with further tests depending on the outcome of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, for step-up procedures, the least significant null hypotheses are considered successively, again with further tests depending in the outcome of earlier ones. As soon as one null hypothesis is rejected, all remaining more significant hypotheses are rejected.

The main difference between step-down and step-up procedures is the order in which null hypotheses are tested: from most significant to least significant for the step-down approach vs. from least significant to most significant for the step-up approach.

The above single-step procedures are based solely on the marginal distributions of the test statistics. In many situations, the test statistics, and hence the corresponding unadjusted p-values, have complex and unknown dependence structures. This is the case, for example, in microarray data analysis, where groups of genes tend to have highly correlated expression measures due to co-regulation. Gains in power may be achieved by taking into account the joint distribution of the test statistics.

The next two procedures are less conservative multiple testing procedures that account for the dependence structure of the test statistics and that control the FWER for arbitrary test statistics joint null distributions. These procedures are based, respectively, on minima of unadjusted p-values (common quantile minP MTP) and maxima of test statistics (common-cut-off maxT MTP).

Let $O_k(n)$ denote the indices for the ordered unadjusted p-values, $P_{0k}^{\circ}(n) \equiv P_{0k}(O_k(n))$, so that $P_{0k}(O_k(1)) \leq \cdots \leq P_{0k}(O_k(N))$.

<u>*Definition: FWER-controlling common-cut-off maxT procedure*</u>

The common-cut-off maxT procedure is based on the maximum test statistic, $Z^{\circ}(1) \equiv \max_n Z(n)$, for the N-vector $Z = (Z(n): n = 1,..., N) \sim Q_0$. Adjusted p-values are given by

$\tilde{p}_{ok}(n) = \Pr_{Q_0}(\max_{n=1,...,N} Z(n) \geq t_k(n)), \qquad n = 1,..., N.$

Definition: FWER-controlling common-quantile minP procedure

The common-quantile minP procedure is based on the minimum unadjusted p-value, $P^{\circ}(1) \equiv \min_n P_0(n)$,where $P_0(n) \equiv \bar{Q}_{0,n}(Z(n))$ denote unadjusted p-values under the test statistics null distribution Q_0, i.e. for $Z = (Z(n): n = 1, \ldots, N) \sim Q_0$. Adjusted p-values are given by

$$\tilde{p}_{ok}(n) = \Pr_{Q_0}(\min_{n=1,\ldots,N} P_0(n) \geq p_{0k}(n)), \qquad n = 1, \ldots, N.$$

For common-quantile MTPs, the *n*th most significant null hypothesis refers to the hypothesis $H_0(O_k(n))$ with the *n*th smallest unadjusted p-value $P_{0k}^{\circ}(n)$, that is, to the hypothesis with *p*-value rank *n*; in contrast, for common-cut-off MTPs, the *n*th most significant null hypothesis is that with the *n*th largest test statistic.

Single step common-cut-off maxT procedure and common-quantile minP procedures are based, respectively, on the distributions of the maximum test statistic and minimum unadjusted p-value over all N null hypotheses. In contrast, the step-down common-cut-off maxT procedure and common-quantile minP procedures are based, respectively, on the distributions of maxima of test statistics and minima of unadjusted p-values over successively smaller nested random subsets of ordered null hypotheses.

2.3 Permutation-based versus Bootstrap-based Tests

Parametric modeling for genomics data is not necessarily accurate. This results in a need for procedures that produce p-values which are reliable in case where null hypothesis is true. One way to accomplish this goal is to use permutation-based *p*-values rather than parametric-based *p*-values.

When performing a permutation test, the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. The theory has evolved from the work of Fisher in the 1930s.

The null hypothesis, under the permutations scheme, is that the data follow the same distribution, which is unknown. In the classic case of the two-sample problem, where $X_1,....,X_m \sim K$ and $Y_1,....,Y_m \sim G$, the null hypothesis is $H_0 : K = G$.

In order to perform the Fisher permutation test based on a test statistic $T(X_1,...,X_m;Y_1,...,Y_n)$, we combine *n*+*m* measurements, draw a sample $Z_1,...,Z_m$ without replacement from the pool, each with probability $\frac{1}{\binom{n+m}{n}}$ and treat it as a control or *X* sample. The remaining observations $Z_{m+1},...,Z_{m+n}$ are treated as a treatment or *Y* sample. Under H_0, each set of *n* observations has the same probability of appearing in a treatment set. Next, a test statistic $T(\underset{\sim}{Z}) = Z_1,...,Z_m;Z_{m+1},...,Z_{m+n})$ is computed. The significance level is the proportion of $T(\underset{\sim}{Z})$ values such that $T(\underset{\sim}{Z}) \geq T(X_1,...,X_m;Y_1,...,Y_n)$. The permutation method works for any statistic.

For the bootstrap testing, one would draw samples of size *n*+*m* from a combined pool, with replacement. Assign the first *m* to control, the next *n* to treatment. Then, compute the test statistic. Repeat this process *B* times. Next, compute the significance level, as above.

When testing on the basis of permutations, the test statistic is computed for each permutation. Since the number of all possible permutations is so huge, even for small sample sizes, the approach is to randomly sample possible permutations for each sample. For the two sample problem, the only difference between permutation tests and bootstrap tests is that samples are drawn with replacement in the bootstrap case.

When performing bootstrap testing, samples are drawn at random with replacement, whereas in the permutations case they are randomly sampled with no replacement. For this reason, when the permutation approach is appropriate, it tends to provide less variable estimators of the test statistic's null distribution than the nonparametric bootstrap.

The relative performances of the various MTPs differ for bootstrap- and permutation-based test statistics null distributions. Differences between the two can be attributed to the fact that the set $\{T_n^B(\cdot,b): b=1,...,B\}$ of *B* bootstrap test statistics does not necessarily include the observed test statistics T_n. This allows bootstrap unadjusted p-values to be zero for some null hypotheses. In contrast, for a null distribution based on all possible $B = \binom{n+m}{n}$ permutations of the treatment and control labels, the observed test statistics T_n are included in the set of B permutations test statistics.

2.4 Smyth's 2004 Article – The Basis for **limma**

The purpose of this paper is to develop the hierarchical model of Lönnstedt and Speed (2002) into a practical approach for microarray experiments with arbitrary numbers of treatments and RNA samples. The first step is to reset it in the context of general linear models with arbitrary coefficients and contrasts of interest. The second step is to derive consistent, closed form estimators for the hyperparameters using the marginal distributions of the observed of the observed statistics. The estimators proposed by Smyth (2004) have robust behavior even for small numbers of arrays. The third step is to reformulate the posterior odds statistic in terms of a moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations.

This approach makes explicit what was implicit in Lönnstedt and Speed (2002), that the hierarchical model results in a shrinkage of the gene-wise residual sample variances towards a common value, resulting in far more stable inference when the number of arrays is small. The use of the moderated t-statistic has the advantage over the posterior odds of reducing the number of hyperparameters which need to be estimated under the hierarchical model; in particular, knowledge of the non-null prior for the fold changes are not required. The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests involving two or more contrasts through the use of moderated F-statistics.

In general we assume that we have a set of k microarrays yielding a response vector $y_g^T = \left[y_{g1}, ..., y_{gk} \right]$ for the gth gene.

Smyth's hierarchical model is:

$$\underset{\sim}{\mathbf{y}}_{\mathbf{g}} = \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\alpha}}_{\mathbf{g}} + \underset{\sim}{\mathbf{e}}_{\mathbf{g}} \quad ; \quad Var - Cov\ \underset{\sim}{\mathbf{e}} \ = \underset{\sim}{\mathbf{W}}_{\mathbf{g}}\sigma_g^2 \quad , \quad ,$$

where $\underset{\sim}{\mathbf{X}}$ is the design matrix of full column rank, $\underset{\sim}{\boldsymbol{\alpha}}_{\mathbf{g}}$ is a parameter vector, and $\underset{\sim}{\mathbf{W}}_{\mathbf{g}}$ is a known non-negative definite weight matrix. The vector $\underset{\sim}{\mathbf{y}}_{\mathbf{g}}$ may contain missing values and the matrix $\underset{\sim}{\mathbf{W}}_{\mathbf{g}}$ may contain diagonal weights which are zero.

Certain contrasts of the coefficients are assumed to be of biological interest and these are defined by $\underset{\sim}{\boldsymbol{\beta}}_{\mathbf{g}} \mathbf{=} \underset{\sim}{\mathbf{C}}^{\mathbf{T}}\underset{\sim}{\boldsymbol{\alpha}}_{\mathbf{g}}$. We assume that it is of interest to test whether individual contrast values β_{gj} are equal to zero.

We assume that the linear model is fitted to the responses for each gene to obtain coefficient estimators $\hat{\alpha}_g$, estimators s_g^2 of σ_g^2 and estimated covariance matrices $Var \cdot Cov(\hat{\alpha}_g) = V_g s_g^2$ where $\underset{\sim}{\mathbf{V}}_{\mathbf{g}}$ is a positive definite matrix not depending on s_g^2.

The contrast estimators are $\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{g}} = \underset{\sim}{\mathbf{C}}^{\mathbf{T}}\underset{\sim}{\hat{\boldsymbol{\alpha}}}_{\mathbf{g}}$. The $\hat{\boldsymbol{\beta}}_{\mathbf{g}}$ are normally distributed,

$\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{g}} \sim N(\underset{\sim}{\boldsymbol{\beta}}_{\mathbf{g}}, \boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{2}}\underset{\sim}{\mathbf{C}}^{\mathbf{T}}\underset{\sim}{\mathbf{V}}_{\mathbf{g}}\underset{\sim}{\mathbf{C}})$ with $Var \cdot Cov\ \underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{g}} = \boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{2}}\underset{\sim}{\mathbf{C}}^{\mathbf{T}}\underset{\sim}{\mathbf{V}}_{\mathbf{g}}\underset{\sim}{\mathbf{C}}$.

The responses y_g are not necessarily assumed to be normal and the fitting of the linear model is not assumed to be by least squares. Nevertheless, the contrast estimators are assumed to be approximately normal and the sample residual variances s_g^2 are assumed to follow approximately a scaled chi-square distribution.

Let v_{gj} be the *j*th diagonal element of $C^T V_g C$. The distributional assumptions made in this paper about the data imply

$$\hat{\beta}_{gj} \Big| \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj,} v_{gj} \sigma_g^2)$$

and

$$s_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{dg}^2$$

where d_g is the residual degrees of freedom in the linear model for gene *g*.

Prior information is assumed on σ_g^2, *g*=1, …., *G,* equivalent to having a prior estimator s_0^2 with d_0 degrees of freedom. That is, the σ_g^2 are i.i.d. with $\dfrac{1}{\sigma_g^2} \sim \dfrac{1}{d_0 s_0^2}\chi_{d_0}^2$,

and $\tilde{s}_g^2 = \dfrac{d_g s_g^2 + d_0 s_0^2}{d_g + d_0} = E\left[\sigma_g^2 \middle| s_g^2\right]$.

Smyth defines the moderated t-statistic as

$$\tilde{t} = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}} \sim t_{d_0+d_g} \text{ under } H_0 : \beta_g = 0 .$$

The added degrees of freedom for $\tilde{t}$ over t reflect the extra information which is borrowed, on the basis of the hierarchical model, from the ensemble of genes for inference about each individual gene. Note that this distributional result assumes d_0 and s_0^2 to be given values. In practice these values need to be estimated from the data.

The article shows that $\tilde{t}$ and s^2 are independent with $s^2 \sim s_0^2 F_{d,d_0}$ and $\tilde{t}\ \beta| = 0 \sim t_{d_0+d}$.

2.5 Overview of Multiple Testing Procedures

2.5.1 Bonferroni (1936)

Bonferroni's (1936) classical procedure for FWER is perhaps the best-known procedure in the multiple testing literature. It controls the FWER for arbitrary test statistics joint null distribution.

The Bonferroni inequality is often used when conducting multiple tests of significance to set an upper bound on the overall significance level α (Miller, 1981, pp. 67-70).

The Bonferroni inequality,

$$P\left\{\bigcup_{true} (P_i \leq \alpha / N)\right\} \leq \sum_{true} P[P_i \leq \alpha / N] \leq \sum_{i}^{n} P[P_i \leq \alpha / N] = N(\alpha / N) = \alpha$$

$(0 \leq \alpha \leq 1)$,

offers strong FWER control and ensures that the probability of rejecting at least one true hypothesis is no greater than *α*.

Definition: FWER-controlling single-step Bonferroni (1936) Procedure

If $T_1, \ldots., T_n$ is a set of *n* statistics with corresponding p-values $P_1, \ldots., P_n$ for testing hypotheses $H_1, \ldots.., H_n$, the classical Bonferroni multiple test procedure is usually performed by rejecting the combined null hypothesis $H_0 = \{H_1 \cap H_2 \cap \cdots \cap H_n\}$ if any p-value is less than *α/N*. Furthermore, the specific hypothesis H_n is rejected for each *n* such that $P_n \leq \alpha/N$ (n=1, ….,*N*).

To control the FWER at level *α,* the single-step Bonferroni procedure rejects any null hypothesis $H_0(n)$ with unadjusted p-value $P_{ok}(n)$ less than or equal to the common single-step cut-off $a_n(\alpha) \equiv \alpha / N$. That is, the set of rejected null hypotheses is:

$$R_k(\alpha) \equiv \left\{n : P_{0k}(n) \leq \frac{1}{N}\alpha\right\}.$$

The corresponding adjusted p-values are thus given by

$$\tilde{P}_{ok}(n) = min\ \{N\ P_{0k}\ ,\ 1\}, \qquad n=1, \ldots., N.$$

Although several multivariate methods have been developed for multiple statistical inference, the Bonferroni procedure is still valuable, being simple to use, requiring no distributional assumptions and enabling individual alternative hypotheses to be tested. Nevertheless, the procedure is conservative and lacks power if several highly correlated tests are undertaken.

2.5.2 Šidák (1967)

Closely related to the Bonferroni procedure is Šidák's (1967) single-step procedure, which controls the FWER for test statistic null distributions Q_0 that satisfy, for the true null hypotheses H_0, an inequality known as Šidák's Inequality.

Consider a random N-vector $Z = (Z(n): n=1,\ldots, N)$, with joint distribution Q_0, and an N-vector of constants $c = (c(n) : n = 1, .., N) \in \mathbb{R}^N$. Then, under conditions described below, Šidák's inequality states that

$$\Pr_{Q0}\left(\bigcap_{n=1}^{N} \{Z(n) \le c(n)\}\right) \ge \prod_{n=1}^{N} \Pr_{Q0}\,(Z(n) \le c(n)).$$

To control the FWER at level α, the single-step Šidák (1967) procedure rejects any null hypothesis $H_0(n)$ with unadjusted p-value $P_{ok}(n)$ less than or equal to the common single-step cut-off $a_n(\alpha) \equiv 1-(1-\alpha)^{1/N}$. That is, the set of rejected null hypotheses is:

$$R_k\,(\alpha) \equiv \{n\!: P_{0k}\,(n) \le 1-(1-\alpha)^{1/N}\}.$$

The corresponding adjusted p-values are thus given by

$$\tilde{P}_{ok}(n) = 1-(1-P_{ok}\ (n))^N, \qquad n=1, \ldots., N.$$

Šidák's inequality holds for independent test statistics and for test statistics with certain parametric distributions. Specifically, the inequality was initially derived by Dunn (1958) for multivariate Gaussian distributions with mean vector zero and certain types of covariance matrices. Šidák (1967) extended the result to multivariate Gaussian distributions with arbitrary covariance matrices and Jogdeo (1977) showed that the inequality holds for a larger class of distributions, including some multivariate *t*- and *F*- distributions.

While the above single-step procedure is simple to implement, it tends to be conservative for control of the FWER. Improvement in power can be achieved by stepwise procedures. Stepwise multiple testing procedures apply the testing procedure to a sequence of successively smaller nested random subsets of null hypotheses, defined by the ordering of the test statistics (common cut-off MTPs) or unadjusted p-values (common quantile MTPs). Step-down MTPs start with the most significant null hypothesis; as soon as one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, step-up MTPs start with the least significant hull hypothesis; as soon as one rejects a null hypothesis, all remaining more significant null hypotheses are rejected.

Note that single-step common cut-off maxT and common-quantile minP procedures are based, respectively, on the distributions of the maximum test statistic and minimum unadjusted p-value over all *N* null hypotheses. In contrast, step-down common cut off maxT and common-quantile minP procedures are based, respectively, on the distributions of maxima of test statistics and minima of unadjusted p-values over successively smaller nested random subsets of ordered null hypotheses.

2.5.3 Holm (1979)

For controlling the FWER at level *α*, the unadjusted p-value cut-offs for the step-down Holm (1979) procedure are as follows:

Let $P_{(1)}, \ldots., P_{(N)}$ be the ordered p-values for testing hypotheses $H_{(1)}, \ldots, H_{(N)}$. Then H_n is rejected if $P_{(n)} \leq \alpha / (N\text{-}n+1)$ for any n=1, …, N.

Definition: FWER-controlling step-down Holm (1979) Procedure

For controlling the FWER at level *α,* the unadjusted p-value cut-offs for the step-down Holm (1979) procedure are as follows:

$$a_n(\alpha) \equiv \frac{1}{N-n+1}\alpha, \qquad n= 1, \ldots, N.$$

The set of rejected null hypotheses is

$$R_k(\alpha) \equiv \left\{ O_k(n) : \tilde{P}_{ok}(O_k(h)) \leq \frac{1}{N-h+1}\alpha, \forall h \leq n \right\}.$$

The corresponding adjusted p-values are thus given by

$$\tilde{P}_{ok}(O_k(n)) = \max_{h=1, \ldots, n}\left\{\min\left\{(N-h+1)P_{ok}(O_k(h)), 1\right\}\right\} \qquad n=1, \ldots, N.$$

Holm's procedure is the step-down analogue of classical single-step Bonferroni procedure and also controls the FWER for arbitrary joint null distributions of the test statistics. The step-down Holm p-value cut-offs, $a_n(\alpha) = \dfrac{\alpha}{N-n+1}$, are greater (i.e., less conservative) than the single-step Bonferroni cut-offs, $a_n(\alpha) = \alpha / N$.

2.5.4. Simes Inequality

Type I error control for commonly used step-up procedures is typically established under the assumption that the test statistics satisfy the following inequality, known as Simes' Inequality (Simes, 1986):

Consider a random *N*-vector $Z = (Z(n){:}n=1,\ldots, N)$, with joint distribution Q_0, unadjusted p-values $P_{(0)} = (P_{(0)}(n){:}\ n = 1, \ldots, N)$, and ordered p-values $P_0^{\circ}(n)$ such that $P_0^{\circ}(1) \le \ldots. \le P_0^{\circ}(N)$. Then, Simes' Inequality states that

$$\Pr_{Q_0}\left(\bigcup_{n=1}^{N} \left\{P_0^{\circ}(n) \le \frac{n}{N}\alpha\right\}\right) \le \alpha .$$

Simes (1986) introduced a modified Bonferroni procedure:

Let $P_{(1)}, \ldots., P_{(n)}$ be the ordered p-values for testing hypotheses $H_{(1)}, \ldots, H_{(N)}$. Then H_n is rejected if $P_{(n)} \le n\ \alpha/N$ for any n=1, …, *N*.

This test procedure, based on the ordered p-values of the individual tests, has FWER equal to α for independent tests.

The modified test procedure is conservative provided

$$Pr\left\{\bigcup_{n=1}^{N} P_{(n)} \leq \frac{n\alpha}{N}\right\} \leq \alpha .$$

This inequality is not true in general as counterexamples can be found. Nevertheless, it may well be true for a large family of multivariate distributions as suggested by Simes' simulation studies (Simes 1986). In this paper, Simes proved the above inequality for independent test statistics, with equality in the continuous case. Although Simes' inequality does not hold for all joint distributions Q_0, the simulations studies in Simes (1986) suggest that the inequality is conservative for a variety of multivariate Gaussian and Gamma test statistics distributions.

The modified Bonferroni procedure should be advantageous by having an actual significance level much closer to the nominal level and consequently a lower type II error probability. Since the Bonferroni procedure leads to a conservative test procedure, there have been several attempts to improve on the method. Šidák (1968, 1971) has shown that the significance level for each test, α/N, can be improved by using 1-(1- $\alpha)^{1/N}$ under certain conditions, although the degree of improvement for N<10 and $\alpha = 0.05$ is slight. (Simes, 1986).

Sarkar (2005) notes that Simes' Inequality also holds for test statistics that satisfy a positive regression dependence on subset (PRDS), as considered by Benjamini and Yekutieli (2001) in the context of FDR-controlling step-up procedures.

The following property, which Benjamini and Yekutieli (2001) call positive regression dependency on each one from a subset I_0, or PRDS on I_0, captures the positive dependency structure for which our main result holds. Recall that a set *D* is called increasing if $x \in D$ and $y \geq x$, implies that $y \in D$ as well.

Property PRDS: For any increasing set *D*, and for each $i \in I_0$, $P\left(X \in D \middle| X_i = x\right)$ is nondecreasing in *x*.

The PRDS property is a relaxed form of the positive regression dependency property. The latter means that for any increasing set *D*, $P\left(X \in D \middle| X_1 = x_1, ..., X_i = x_i\right)$ is nondecreasing in $\left(x_1, ..., x_i\right)$ (Sarkar (1969)). In PRDS the conditioning is on one variable only, each time, and required to hold only for a subset of the variables.

2.6 False Discovery Rate (FDR)

In a now classical article, Benjamini and Hochberg (1995) suggested a new point of view on the problem of multiplicity. In many multiplicity problems the number of erroneous rejections should be taken into account and not only the question of whether any error was made. Yet, at the same time, the seriousness of the loss incurred by erroneous rejections is inversely related to the number of hypotheses rejected. From this point of view, a desirable error rate to control may be the expected proportion of Type I errors among the rejected hypotheses, which Benjamini and Hochberg termed the False Discovery Rate (FDR). This criterion integrates Spjøtvoll's (1972) concern about the number of errors committed in multiple comparison problems, with Soriç's (1989) concern about the probability of a false rejection given a rejection.

Consider the problem of testing simultaneously N (null) hypotheses, of which n_0 are true. Let $\boldsymbol{R}$ be the number of hypotheses rejected. Table 2.1 summarizes the situation in a traditional form.

Table 2.1

Number of errors committed when testing *N* null hypotheses

	Declared non-significant	Declared significant	Total
True Null Hypotheses	U	V	n_0
False Null hypotheses	T	S	N-n_0
	$N - R$	R	N

The specific N hypotheses are assumed to be known in advance. The quantity R is an observable random variable; U, V, S, T are unobservable random variables. If each individual null hypothesis is tested separately at level α, then $R = R(\alpha)$ is increasing in α.

In terms of these random variables, the per comparison error rate is $E(V/N)$ and the FWER is $P(V\geq 1)$. Testing individually each hypothesis at level α guarantees that $E(V/N) \leq \alpha$. Testing individually each hypothesis at level α/N guarantees that $P(V\geq 1) \leq \alpha$.

Definition: False Discovery rate

The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $Q = V/(V+S)$ – the proportion of the rejected null hypotheses which are erroneously rejected. Naturally, Q is defined to be zero when $V + S = 0$, as no error of false rejection can be committed. The ratio Q is an unobserved (unknown) random variable, as we do not know V or S, and thus $Q = V/(V+S)$, even after experimentation and data analysis. The FDR Q_e is defined to be the expectation of Q:

$$Q_e = E(Q) = E\{V/(V+S)\} = E(V/R).$$

Two properties of this error rate are easily shown, yet are very important.

(a) If all null hypotheses are true, the FDR is equivalent to the FWER: in this case $S = 0$ and $V = R$, so if $V = 0$ then $Q = 0$, and if $V>0$ then $Q = 1$, leading to $P(V\geq 1) = E(Q) = Q_e$. Therefore control of the FDR implies control of the FWER in the weak sense.

(b) When $n_o<N$, the FDR is smaller than or equal to the FWER: in this case, if $V>0$ then $V/R\leq 1$, leading to $\chi_{(V\geq 1)}\geq Q$. Taking expectations on both sides, the following is obtained: $P(V\geq 1) \geq Q_e$, and the two can be quite different. As a result, any procedure that controls the FWER also controls the FDR. However, if a procedure controls the FDR only, it can be less stringent, and a gain in power may be expected. In particular, the larger the number of the false null hypotheses is, the larger S tends to be, and so the larger the difference between the error rates tends to be. As a result, the potential for increase in power is larger when more of the hypotheses are false.

Benjamini and Hochberg (1995) propose the following FDR-controlling marginal step-up procedure, based on Simes' unadjusted p-value cut-offs.

Definition: FDR-controlling step-up Benjamini and Hochberg (1995) procedure

For controlling the FDR at level *α,* the unadjusted p-value cut-offs for the step-up Benjamini and Hochberg (1995) procedure are as follows,

$$a_n(\alpha) \equiv \frac{n}{N}\alpha\,, \qquad n = 1, \ldots, N,$$

and the set of rejected null hypotheses is

$$R_k(\alpha) \equiv \{O_k(n) : \exists h \geq n \text{ such that } P_{0k}(O_k(h)) \leq \frac{h}{N}\alpha\}.$$

The corresponding adjusted p-values are thus given by

$$\tilde{P}_{ok}(O_k(n)) = \min_{h=n, \ldots., N}\{\min\{(\frac{N}{h}P_{ok}(O_k(h)),1\}\} \qquad n=1, \ldots.,N.$$

Benjamini and Hochberg (1995) proved that their procedure controls the FDR for independent test statistics. The subsequent article of Benjamini and Yekutieli (2001) established FDR control for test statistics with more general independence structures, such as positive regression dependence.

Most FDR-controlling procedures proposed thus far do not exploit the dependence structure of the test statistics; That is, they are based solely on the (marginal) unadjusted p-values. In addition, FDR control results are generally derived under the assumption that the test statistics are either independently distributed or have certain forms of dependence such as positive regression dependence (Benjamini and Yekutieli, 2001).

2.6.1 Tail Probability for the Proportion of False Positives (TPPFP)

In contrast to FDR-controlling approaches which focus on the expected value of the proportion of false positives (PFP) among the rejected hypotheses, Genovese and Wasserman (2004 a,b), Korn et al. (2004), van der Laan et al. (2004b, 2005), Lehmann and Romano (2005), and Romano and Wolf (2005) proposed procedures that control tail probabilities for this proportion. These authors argue that although FDR-controlling approaches control the PFP (proportion of false positives) on average, they do not preclude large variations in the PFP. When one wishes to have high confidence (i.e. chance at least (1-α)) that the set of rejected null hypotheses contains at most a specified proportion *q* of false positives, control of the tail probability for the proportion of false positives (TPPFP) among the rejected hypotheses,

$TPPFP(q) = \Theta\left(F_{V_n/R_n}\right) = 1 - F_{V_n/R_n}(q) = \Pr\left(V_n / R_n > q\right)$, is the appropriate form of Type I error control. The parameter *q* confers flexibility to TPPFP-controlling MTPs and can be tuned to achieve an acceptable level of false positives.

Multiple testing procedures proposed thus far for controlling a parameter of the distribution of the proportion of false positives among the rejected hypotheses suffer from one or both of the following limitations: (i) they are based solely on the marginal distributions of the test statistics; (ii) they rely on a number of assumptions concerning the joint distribution of the test statistics, such as independence, positive regression dependence or normality. Van der Laan et al. (2004b) showed that *any* FWER-controlling procedure can be straightforwardly augmented to control the

TPPFP, for general data generating distributions and, hence, arbitrary dependence structures for the test statistics.

Van der Laan et al. (2004a), and subsequently Dudoit et al. (2004a) and Dudoit and van der Laan (2004), proposed the augmentation multiple testing procedure (AMTP), obtained by adding suitably chosen null hypotheses to the set of null hypotheses already rejected by an initial gFWER-controlling MTP. Adjusted p-values for the AMTP are shown to be simply shifted versions of the adjusted p-values for the original MTP. Denote the adjusted p-values for the initial FWER-controlling procedure R_k (α) by $\tilde{P}_{ok}\,(n)$. Order the N null hypotheses according to these p-values, from smallest to largest. That is, define indices $O_k(n)$, so that

$\tilde{P}_{ok}\,[O_k(1)] \leq \ldots. \leq \tilde{P}_{ok}\,[O_k(N)]$.

<u>Definition: TPPFP-controlling augmentation multiple testing procedure (Van der Laan et al. (2004b)</u>

For control of TPPFP(q) at level α, given an initial FWER-controlling procedure R_k (α), reject the R_k (α) = $\left|R_k\left(\alpha\right)\right|$ null hypotheses specified by this MTP, as well as the next A_k (α) most significant hypotheses,

$$A_k\,(\alpha) \;=\; \max\left\{ n \in \{0,\ldots.,\; N - \;R_k\left(\alpha\right)\} \;: \frac{n}{n + R_k(\alpha)} \leq q \right\}$$

$$=\; \min\left\{ \left\lfloor \frac{qR_k\left(\alpha\right)}{1-q} \right\rfloor, N - \;R_k\left(\alpha\right) \right\},$$

where the floor function $\lfloor x \rfloor$ denotes the greatest integer less than or equal to *x,* that is, $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. That is, keep rejecting null hypotheses until the ratio of additional rejections to the total number of rejections reaches the allowed proportion *q* of false positives. The adjusted p-values $\tilde{P}^{+}_{ok}[\, O_k(n)]$ for the new TPPFP-controlling AMTP are simply *nq*-shifted versions of the adjusted p-values of the initial FWER-controlling MTP. That is,

$$\tilde{P}^{+}_{ok}[O_k(n)] = \tilde{P}_{ok}(O_k(\lceil (1-q)n \rceil)), \qquad n = 1, \ldots.., N,$$

where the ceiling function $\lceil x \rceil$ denotes that least integer greater than or equal to *x.*

It is interesting to note the parallels between TPPFP-controlling step-down procedures and the FDR-controlling step-up procedures of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). The penalty to guarantee Type I error control for general dependence structures tends to be more severe for FDR-controlling procedures than for the TPPFP-controlling procedures.

2.7 The Bioconductor Project

Current statistical inference problems in biomedical and genomic data analysis routinely involve the simultaneous test of thousands, or even millions, of null hypotheses. These testing problems share the following general characteristics: inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables; a broad range of parameters of interest, for example, regression coefficients and correlations; many null hypotheses, in the thousands or even millions; and complex dependence structures among test statistics.

The Bioconductor project started in 2001, and is an open source, open development software project to provide tools for the analysis and comprehension of high-throughput genomic data. It is based primarily on the R programming language, and most of the Bioconductor components are distributed as R packages. It provides widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data. Bioconductor's software can be used for: microarray analysis (data import, quality assessment, normalization, differential expression analysis, clustering, classification, and many more applications); annotation (using microarray probe, gene, pathway, gene ontology, homology and other annotations); high throughput assays (importing, transforming, editting, analyzing and visualizing various types of assays); and Transcription factors analysis (finding candidate binding sites for known transcription factors via sequence matching). [www.Bioconductor.org].

2.7.1 `Multtest`

Pollard, Dudoit and van der Laan (2004) developed the Bioconductor R package **`multtest`,** which implements widely applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates. Nonparametric bootstrap and permutation resampling-based multiple testing procedures (including empirical Bayes methods) for controlling the family-wise error rate (FWER), generalized family-wise error rate (gFWER), tail probability of the proportion of false positives (TPPFP), and false discovery rate (FDR) are implemented. Permutation tests based on a variety of *t*- and *F*-statistics (including *t*-statistics based on regression parameters from linear and survival models as well as those based on correlation parameters) are included. Results are reported in terms of adjusted p-values, confidence regions and test statistic cutoffs. The procedures are directly applicable to identifying differentially expressed genes in DNA microarray experiments.

2.7.2 `Limma`

An alternative software package under Bioconductor is **`limma`**, which was developed by Smyth (2004). It is designed to analyze complex microarray experiments involving comparisons between many RNA targets simultaneously. By fitting a linear model to the expression data for each gene, this package allows analyses of contrasts of interest. Empirical Bayes and other shrinkage methods are used to borrow information across genes, making the analyses stable even for experiments with small number of arrays (Smyth, 2004; Smyth et al., 2005).

`Limma` uses linear models to analyze designed microarray experiments (Yang and Speed, 2003; Smyth, 2004). This approach allows very general experiments to be analyzed nearly as easily as a simple replicated experiment.

Mathematically, we assume a linear model $E[y_j] = X\alpha_j$, where y_j contains the expression data for gene *j*, X is the design matrix, and α_j is a vector of coefficients.

Here y_j^T is the *j*th row of the expression matrix and contains either log-ratios or log-intensities. The contrasts of interest are given by $\beta_j = C^T \alpha_j$ where C is the contrast matrix. The coefficients component of the fitted model contains estimated values for the α_j. After applying the contrast step, the coefficients component now contains estimated values for the β_j.

With common reference microarray data, linear modeling is much the same as ordinary ANOVA or multiple regression except that a model is fitted for every gene.

The basic statistic used for significant analysis is the moderated t-statistic, which is computed for each probe and for each contrast. This has the same interpretation as an ordinary t-statistic except that the standard error is estimated by pooling variance estimates across genes, that is, shrunk toward a common value, using a simple Bayesian model. This has the effect of borrowing information from the ensemble of genes to aid with inference about each individual gene (Smyth, 2004).

Moderated t-statistics (for specified t-tests among 2-groups) lead to p-values in the same way that ordinary t-statistics do except that the degrees of freedom are increased, reflecting the greater reliability associated with the smoothed standard errors. These *p*-values are adjusted for multiple testing. The most popular form of adjustment is “FDR”, which is Benjamini and Hochberg’s method to control the false discovery rate (Benjamini and Hochberg, 1995).

The *B* statistic (lods or *B*) is the log odds that the gene is differentially expressed (Smyth, 2004, Section 5). The *B* statistic is automatically adjusted for multiple testing by assuming that 1% of all genes, or some other percentage specified, are expected to be differentially expressed. The *p*-values and *B*-statistics will normally rank genes in the same order.

The empirical Bayes step computes one more useful statistic. The moderated *F*-statistic (*F*) combines the *t*-statistics for all the contrasts into an overall test of significance for that gene. The *F*-statistic tests whether all contrasts are non-zero for that gene against a general alternative. The denominator degrees of freedom is the same as that of the moderated *t*. It is similar to the ordinary *F*-statistic from analysis of variance except that the denominator mean squares are shrunken, as described above.

Chapter 3: Analysis

In this chapter, we analyze gene expression data on a study comparing the responses of kidney cells in rats exposed to ischemia and toxic assaults.

3.1 Data Collection

The data used in this paper were originally collected in Dr. Robert A. Star's laboratory, at the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland. The results of this study were published in Yuen et al. 2006 paper in the Physiological Genomics journal.
The data is publically available at:
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3219.

In Yuen et al. (2006), the authors used microarrays to identify early biomarkers that distinguish ischemic from nephrotoxic acute renal failure or biomarkers that detect both injury types. A total of 31 male rats were assigned to 9 different experimental groups, in which rat kidney transcriptomes were compared at 2 and 8 hours after ischemia/reperfusion and after mercuric chloride injection. A control group was also included in the study. The nine different experimental groups were: normal; volume depletion; sham (40 minutes sham surgery and then harvested after 2h or 8h); ischemia/reperfusion (40 minutes bilateral ischemia, 2h or 8h reperfusion, then harvest); ischemia/reperfusion + α -melanocyte hormone (α -MSH), 2h; mercuric chloride (inject 4 mg/kg mercuric chloride, s.c., harvest after 2h or 8h); and cisplatine, 2h. To minimize the effect of circadian rhythms in gene expression, all treatments began at the same time of day. Processing of the total RNA and subsequent hybridization were performed in two lots, according to the CodeLink Gene Expression Bioarray user guide. Five of the total RNA samples were processed in duplicate.

3.2 Data Processing and Analysis Methods Employed by Yuen et.al

3.2.1 Data Processing

The scanned data were processed by CodeLink Expression Analysis software and the signal intensity of all genes on a microarray was normalized by median centering within each microarray (where the hybridization intensity for each gene in a microarray was multiplied by a scaling factor for that array, so that the median intensity became 1) and the signal intensities were $\log_2$ transformed.

Genes with missing values were then removed (9,251 remained from the original 9,988). Quality control was initially assessed by a histogram analysis of all possible pairs of arrays and was confirmed by Bland-Altman analysis and principal component analysis (PCA). PCA is a method used to transform gene expression information into variance-based information. Even though 9,251-dimensional gene space is converted into 9,251 principal components, the first 3 principal components appeared to contain most of the variance-based information and could be visualized in 3-dimensional space. Because the results from the first and second microarray lots were dramatically different and the second lot had larger animal-to-animal variation, Yuen et al. (2006) normalized the second lot on a gene-by-gene basis to equalize the normal genes in both lots as follows:

Adjusted signal intensity$_{\text{gene x}}$(lot 2) = signal intensity$_{\text{gene x}}$(lot 2)
Χ mean normal signal intensity$_{\text{gene x}}$(lot 1)/mean normal signal intensity$_{\text{gene x}}$(lot 2)

Yuen et al. (2006) used a dissimilarity matrix (Euclidian distance heat map) to rapidly assess the overall quality of each microarray, and this analysis also provided an initial estimate of microarray clustering. The dissimilarity matrix, histogram, and Bland-Altman analyses gave the same result, that there were three outlier microarrays (one normal, one volume depletion and one ischemia/reperfusion, 8h), which they removed from further analysis. The three analyses also revealed that two biological groups, cisplatin 2h, and ischemia/reperfusion + α-MSH, 2h. were not different from their corresponding control microarrays. They removed these groups from further microarray analysis, and thus a total of 22 microarrays remained.

3.2.2 Analysis Methods

After lot normalization and removal of outlier microarrays and groups, the PCA and hierarchical clustering analyses were repeated, with improved outcomes (*n*=4, normal; *n*=3, sham; *n*=4, volume depletion; *n*=4, ischemia/reperfusion, 2h; *n*=2, ischemia/reperfusion, 8h; *n*=3, mercuric chloride, 2h; *n*=3 mercuric chloride, 8h). The normal groups from lot 1 and lot 2 were indistinguishable by either method. However, the PCA and hierarchical clustering analyses classified the injury groups differently. Hierarchical clustering showed more segregation by time (2 vs. 8 h) than type of injury (ischemia/reperfusion vs. mercuric chloride). In contrast, PCA indicated that the mercuric chloride groups are closer to each other and to the normal/sham/volume depletion groups, and the ischemia/reperfusion groups are farthest from each other and from the normal/sham/volume depletion groups. These results may be attributed to the different distance metrics used by the two methods.

Two methods of filtering were applied to the lot-normalized data set (after removing outliers). After removal of duplicate microarrays, an unsupervised one-way ANOVA was applied, gene by gene, to the remaining 22 microarrays, using a Dunn-Sidak corrected *p*-value of <0.001 as a cutoff. The 615 genes showing significant effects were subjected to PCA, and the first three principal components accounted for 58% + 14% + 10% = 82% of the total variation. The Ischemia genes yielded a 3-dimensional plot which appeared to be close to a straight line in the PCA space, while the Mercury result appeared to be very close to a distinct straight line in the PCA space.

Because the PCA analysis suggested that the treatment groups were distinct, Yuen et al. (2006) performed a two-stage filtering protocol, where the first stage was an unsupervised gene by gene ANOVA with a less stringent cutoff of P<0.05 (Dunn-Sidak), resulting in 1,596 genes, followed by a series of pre-specified *t*-tests between the normal group and each injury group, with a cutoff of P < 0.01, combined with a twofold change in the mean level of gene expression. The two stage filtering protocol culminated in a total of 728 genes, which were categorized by individual or combined conditions and summarized in a table. Each condition was expressed as exclusive of other groups or nonexclusive.

3.3 Data Processing and Analysis Methods employed in the present Research

3.3.1 Data Processing

The initial screening performed in this analysis of the data is different from the methods of Yuen et al. (2006). Both screening protocols started with signal intensity of all genes on a microarray being normalized by median centering within each microarray and $\log_2$ transformed.

The data used for the analysis presented in this paper is based on the 22 microarrays that remained after the first stage of Yuen et al.'s preprocessing of the data, with only one significant difference: genes with missing values were removed from the 22 microarrays (9,501 remained from the original 9,988) and not from the original 36 microarrays (9,251 remained in Yuen et al.'s study). This resulted in 250 more genes (roughly 2.5% of the original data collected) to analyze and compare in this analysis, information that was lost in Yuen et al.'s study.

3.3.2 Methods of Analysis

The analyses presented in this paper employed modern multiple-comparison procedures aimed to control the proportion of type I errors among the rejected hypotheses in families of comparisons under simultaneous consideration.

These analyses were performed using two Bioconductor R packages (refer to Bioconductor website), **`limma`** and **`multtest`**, which represent two different approaches of analysis. The first relies on classical linear models theory while the second uses permutation-based tests.

The Bioconductor R package, **`limma`**, is designed to analyze complex microarray experiments involving comparisons between many RNA targets simultaneously. By fitting a linear model to the expression data for each gene, this package allows analyzing contrasts of interest. Empirical Bayes and other shrinkage methods are used to borrow information across genes making the analyses stable even for experiments with small number of arrays. **`Limma`** then adjusts for multiple testing by applying the False Discovery Rate (FDR) procedure [Benjamini and Hochberg (1995)]. (Smyth, 2004; Smyth et al., 2005),

On the other hand, the **`multtest`** R package implements widely applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates. False Discovery Rate (FDR) [Benjamini and Hochberg (1995)], and the Tail Probability for the Proportion of False Positives (TPPFP) [van der Laan et al. (2004b, 2005)] procedures were applied and compared by their results yield.

3.4 Results

3.4.1 **`Limma`** Analysis

The differentially expressed genes found by Yuen et al. (2006) exclusively for each treatment had the following quantities: 417 IR genes, 90 mercuric chloride ($HgCl_2$) genes, 0 genes for sham treatment and 0 genes for volume depletion. Guided by the choice of Yuen et al. (2006) to focus attention on the ischemic (Ischemia Reperfusion) and nephrotoxic (mercuric chloride) treatments due to their biologically significant medical effects and also due to their genes comprising the vast majority of differentially expressed genes found in their study, the next step in this analysis was to study the different effects each of these two types of treatments had on the genes which reacted to them to the most significant extent. **`Limma`** analysis was performed comparing normal groups to the 2 hour- and 8 hour- treatments of mercuric chloride ($HgCl_2$) and Ischemia/Reperfusion (IR). The **`limma`** procedure compared the average difference between treatment groups and normal arrays, yielding a list of top differentially expressed genes based on the adjusted *p*-values. The default adjustment method used to adjust the *p*-values for multiple testing is the Benjamini-Hochberg method, which controls the expected false discovery rate). The top 50 genes were clustered and plotted into the heatmaps depicted in Figures 3.1-3.3.

The R function **`hclust`** was used for the hierarchical clustering to be performed. This function performs a hierarchical cluster analysis using a set of dissimilarities for the *n* objects being clustered. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance–Williams dissimilarity update formula according to the particular clustering method being used.

A number of different clustering methods are available through the **`hclust`** function. *Ward's* minimum variance method aims at finding compact, spherical clusters. The *complete linkage* method finds similar clusters, and is the clustering method that was chosen. The *single linkage* method (which is closely related to the minimal spanning tree) adopts a ‘friends of friends’ clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods.

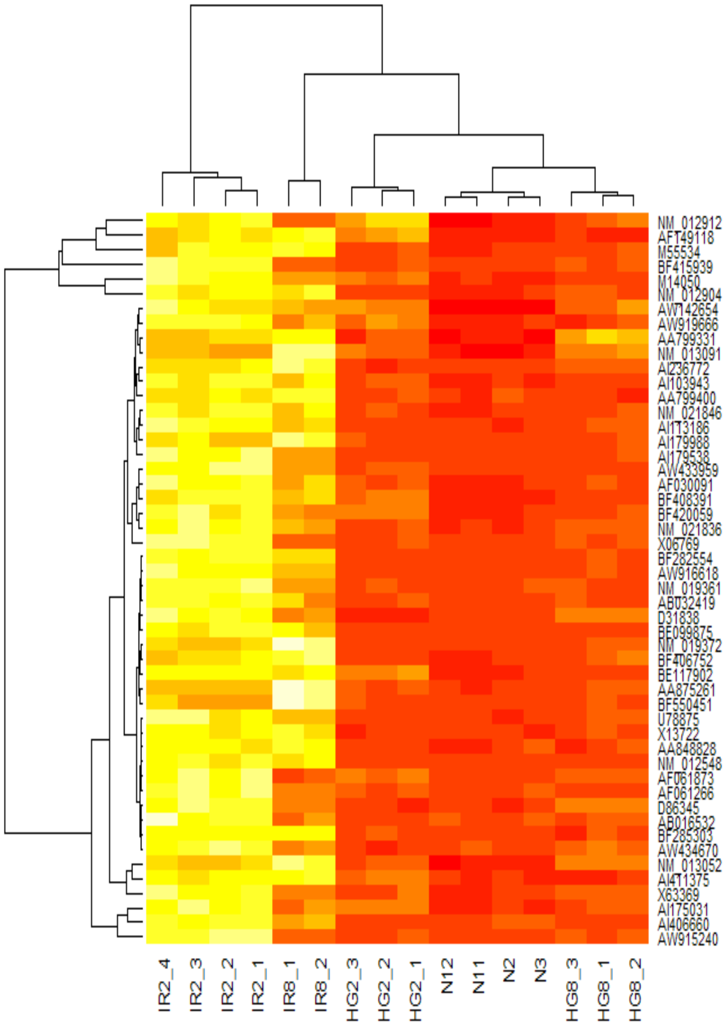

Figure 3.1: Heatmap for hierarchical clustering of top 50 differentially expressed genes yielded by **`limma`** comparison of $HgCl_2$ at 2 and 8 hours (HG2 and HG8), IR at 2 and 8 hours and Normal treatments. Color scale ranges from white, transitioning through yellow and orange, to red, where white represents highest expression values and red represents the lowest expression values.

The heatmap in Figure 3.1 shows the expression levels and clustering of the 50 most differentially-expressed genes based on contrasts that tested the average difference between the treatment groups ($HgCl_2$ at 2 and 8 hours and IR at 2 and 8 hours) and the normal groups. The clustering seems to separate well the IR treatment groups from the others, while not doing the same in separating the $HgCl_2$ treatment groups from the controls. In general, these top 50 differentially expressed genes seem to show a consistent pattern of up-regulation to the highest degree by the IR treatments and to a much lesser degree by the $HgCl_2$ treatment. The clustering of the genes reveals a very big separation of the first cluster from all the rest, which are not so diverse as measured by height. The height of that cluster differs greatly from the height of the rest of the clusters of genes, indicating a significantly different pattern of expression for that clustered collection of 6 genes (NM_012912, AF149118, M55534, BF415939, M14050, NM_012904). Three of these genes, AF149118, M55534, and BF415939 are differentially expressed genes newly discovered by this analysis. Their p-values are very small (ranging between 1.62e-14 and 4.89e-09), and their corresponding fold changes are highly significant (ranging between +4.26 and +100.68). Genes NM_012912 and AF149118 appear in the top 50 differentially expressed genes for $HgCl_2$. One of them, NM_012912 , also appears in an unusual cluster in the corresponding dendogram. Genes M55534, BF415939, and M14050 and NM_012904 appear in the top 50 differentially expressed genes for IR and also appear in the same unusual cluster in the corresponding dendogram.

Among these top 50 differentially expressed genes for both IR and $HgCl_2$ treatments, two genes (AW142654 and BF420059) also appear in Yuen et al.'s list of 23 genes found to be differentially expressed for the two treatments.

A more detailed analysis was performed by testing the IR and $HgCl_2$ arrays separately by **limma**, in order to study the effects of each treatment on the expression behavior as detected by the contrasts we tested (average difference between treatment groups and normal arrays).

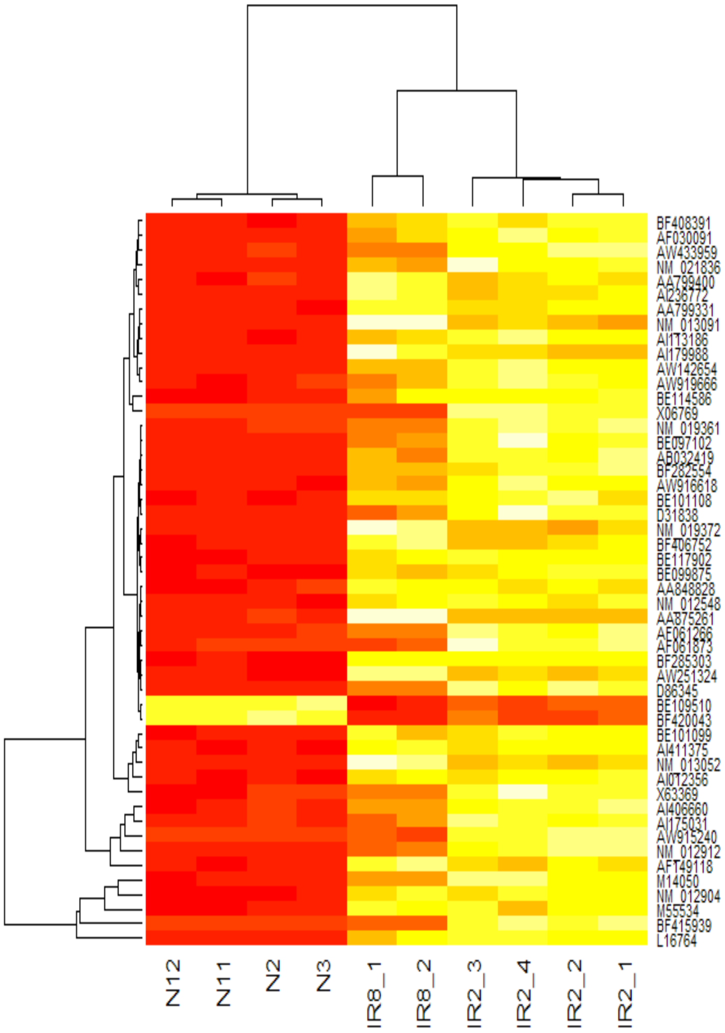

Figure 3.2: Heatmap for hierarchical clustering of top 50 differentially expressed genes yielded by **limma** comparison of IR 2 and 8 hours to the Normal treatments. Color scale ranges from white, transitioning through yellow and orange, to red, where white represents highest expression values and red represents the lowest expression values.

Figure 3.2 shows the expression levels and clustering of the 50 most differentially-expressed genes for the IR treatments. It shows that most genes are up-regulated by the IR treatment, with the exception of two genes (BE109510 and BF42004) which show an opposite pattern (down-regulated by the IR treatment). A high percentage of the genes seem to be over-expressed to the highest degree at 2 hours, and to transition toward normal levels again at 8 hours. Located in the Appendix section is Table A1, which is a complete, detailed list of the Top 50 differentially expressed genes for the IR treatment, and includes their *p*-values, biological identity/function, and their corresponding fold-changes in reference to normal. The biological identity/function was searched in Entrez, which is NIH's NCBI Life Sciences search-engine (http://www.ncbi.nlm.nih.gov/sites/gquery). The fold-change is calculated as the ration of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

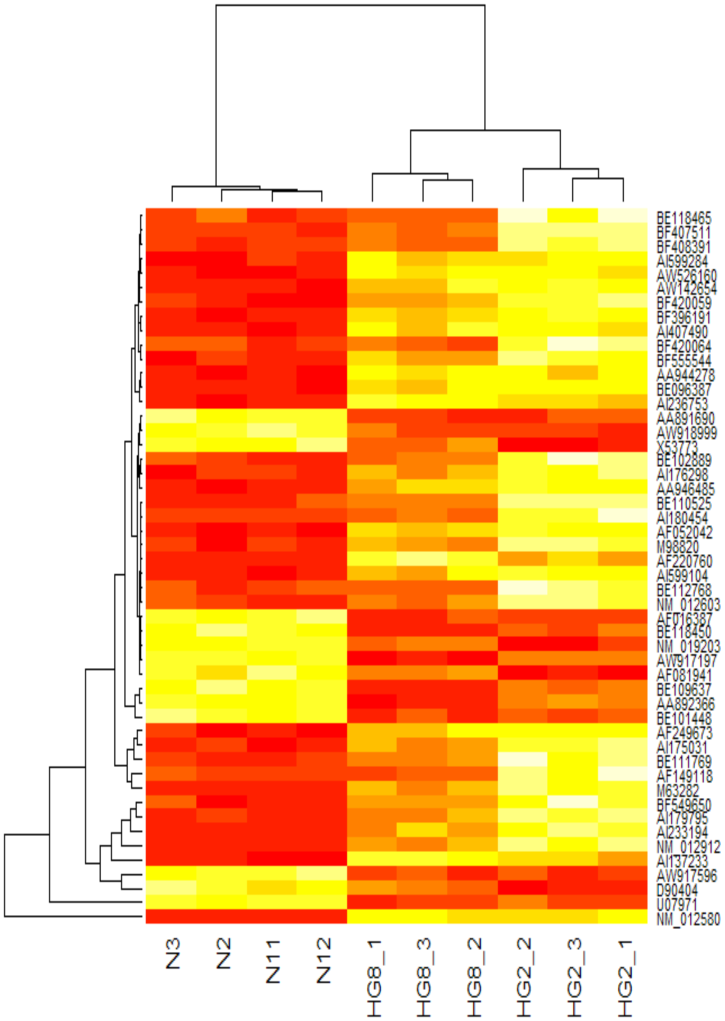

Figure 3.3: Heatmap for hierarchical clustering of top 50 differentially expressed genes yielded by **`limma`** comparison of $HgCl_2$ at 2 and 8 hours (HG2 and HG8) to the Normal treatments. Color scale ranges from white, transitioning through yellow and orange, to red, where white represents highest expression values and red represents the lowest expression values.

Figure 3.3 shows the expression levels and clustering of the 50 most differentially-expressed genes for the $HgCl_2$ treatments. It clearly shows different expression patterns for separate chunks of genes. Some of them show over-expression to the highest degree at 2 hours while transitioning back to close-to normal levels at 8 hours. Some show under-expression at 2 and 8 hours in response to the $HgCl_2$ treatment. This clear and significant separation of the 50 most differentially-expressed genes into groups can be biologically interpreted as an indication for their common function. Located in the Appendix section is Table A2, which is a complete, detailed list of the Top 50 differentially expressed genes for the $HgCl_2$ treatment, and includes their p-values, biological identity/function, and their corresponding fold-changes in reference to normal. . The biological identity/function was searched in Entrez, which is NIH's NCBI Life Sciences search-engine (http://www.ncbi.nlm.nih.gov/sites/gquery). The fold-change is calculated as the ration of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

The heatmaps in Figures 3.2 and 3.3 depict a very different image of the pattern of expression of the top 50 differentially-expressed genes for the two different treatments (IR and $HgCl_2$).

The clustering of these top 50 differentially-expressed genes (yielded by **`limma`**'s analysis and ranked by the **`toptable`** function) for each treatment is presented in Figures 3.4 and 3.5 (the **`hclust`** function is the clustering method, the same as in the heatmaps, and the information portrayed is identical in both the heatmaps and the dendograms).

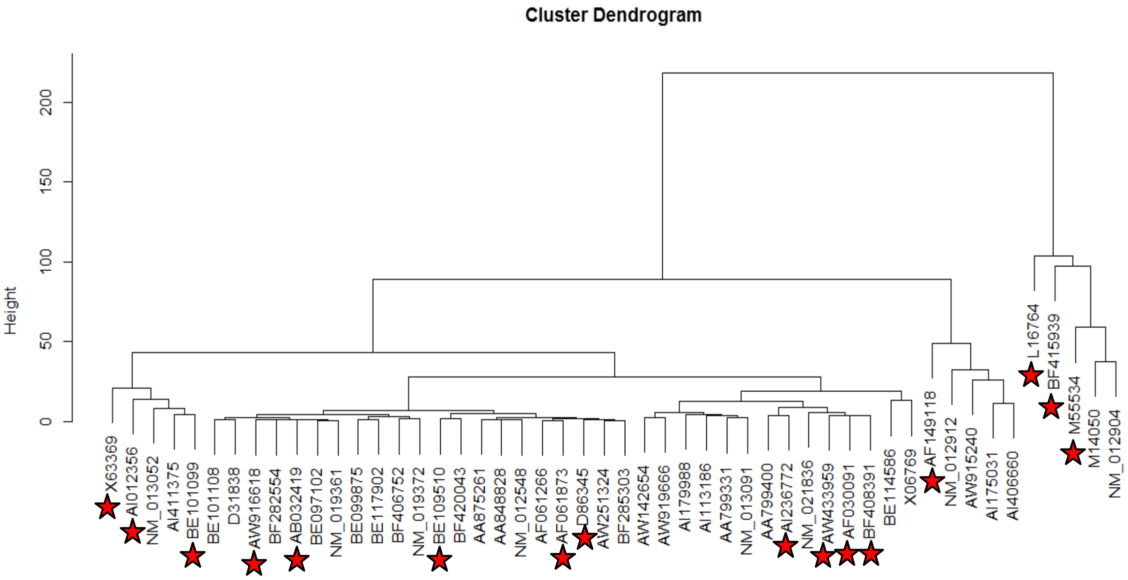

Figure 3.4: Cluster dendogram of top 50 differentially expressed genes yielded by `limma` comparison of IR 2 and 8 hours to the Normal treatments. Newly discovered genes are marked on the dendogram by a red star ★.

Figure 3.4 shows the Cluster dendogram of top 50 differentially expressed genes yielded by `limma` comparison of IR 2 and 8 hours to the Normal treatments. Newly discovered genes are marked on the dendogram by a red star . The cluster dendogram in this figure shows a clear separation of two right-most clusters from the rest of the clusters. These two clusters contain newly discovered genes and hence the dendogram may serve as a potential tool for finding new genes and figuring out their function. In the first cluster, five genes appear. Three of these genes, L16764, M55534, and BF415939 are newly discovered genes by this analysis. A search in Entrez, NIH's NCBI Life Sciences search-engine, revealed that new gene L16764 is a

Heat shock 70kD protein 1A (Hspa1a), located on chromosome number 20 of the Rattus Norvegicus (Norway rat) genome. The search also identified new gene M55534 as Crystallin, alphaB (Cryab), which seems to be highly similar to a heat-shock protein that's alpha-crystallin-related. Within that same cluster appears gene M14050, which had been identified by Yuen et al. as Heat shock 70kD protein (Hspa5), located on chromosome number 3 of the Rattus Norvegicus genome. The common functionality of the clustered genes demonstrates how the clustering served as a helpful tool in identifying the deciphering the function of newly discovered, unknown genes. It also supports the validity of these results.

The newly discovered gene BF415939 is identified as FBJ osteosarcoma oncogene (Fos), located on chromosome number 6 of the Rattus Norvegicus genome. The **limma** procedure for top differentially expressed IR genes assigned a significantly small p-value to it (4.88e-06), and it shows impressive fold changes of +100.68 and +12.35 at IR 2 hours and 8 hours respectively.

Another newly discovered gene, appearing in the second cluster, is AF149118. Entrez identifies it as ADAM metallopeptidase with thrombospondin type 1 motif. It shows fold-changes of +7.63 and +9.69 at 2 and 8 hours of IR treatment respectively. It was also found as an $HgCl_2$ differentially expressed gene, showing fold-changes of +5.68 and +1.44 at 2 and 8 hours of $HgCl_2$ treatment respectively. This gene is positioned next to gene NM_012912 (known as Activating transcription factor 3, Atf3), which is also a differentially expressed gene under both IR and $HgCl_2$ treatments (showing fold-changes of +14.6 at 2 hours of IR treatment and +11.9 and

+9.80 at 2 and 8 hours of $HgCl_2$ treatment respectively). The tight clustering of these two genes may hint at a common or related pathway shared by the two genes.

Two genes, BE109510 (newly discovered gene) and BF420043, are the only two genes which showed an expression behavior opposite to the rest of the genes in the IR heatmap. They are still grouped together with other genes (which show the opposite pattern) in the same cluster. Gene BE109510 is identified by Entrez as Transmembrane and coiled-coil domains 6 (Tmco6). Gene BF420043 had been found by Yuen et al. as an IR differentially expressed gene, but its function hasn't been identified. They both show a negative fold change at IR 2 and 8 hours (greater negative change at 8 hours).

Two other interesting newly discovered genes are AF061873 and D86345, which are clustered together. Entrez identifies them as having a similar function as receptors: Transient receptor potential cation channel, subfamily C, member 1, Trpc1, for gene AF061873, and Leukemia inhibitory factor receptor alpha, Lifr, for gene D86345.

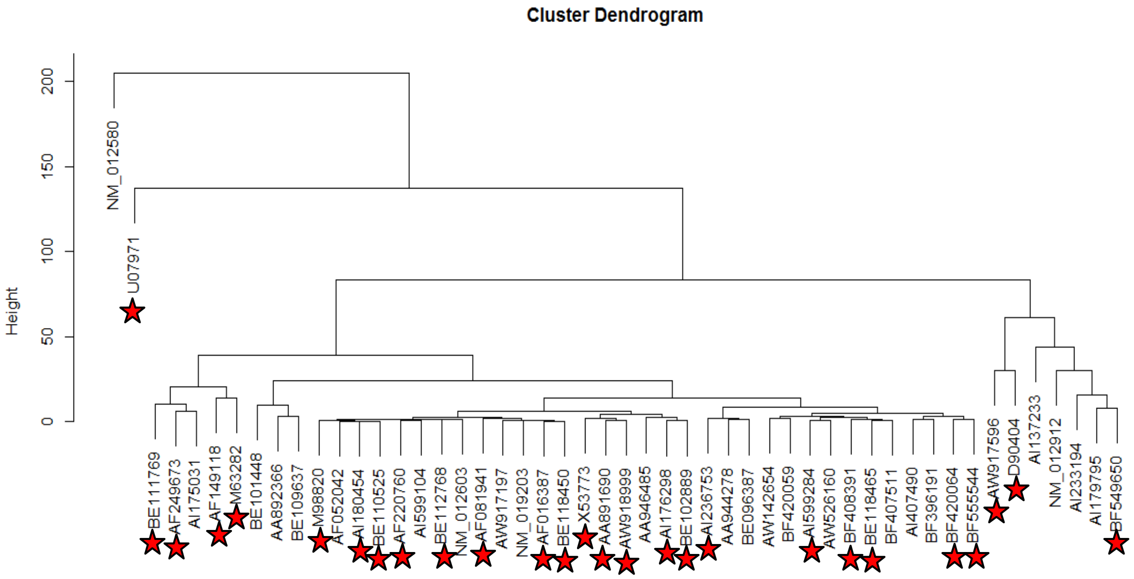

Figure 3.5: Cluster dendogram of top 50 differentially expressed genes yielded by `limma` comparison of $HgCl_2$ 2 and 8 hours to the Normal treatments. Newly discovered genes are marked on the dendogram by a red star ★

Figure 3.5 shows the Cluster dendogram of top 50 differentially expressed genes yielded by **limma** comparison of $HgCl_2$ at 2 and 8 hours to the Normal treatments. Newly discovered genes are marked on the dendogram by a red star. The cluster dendogram in this figure shows a clear separation of one cluster (located at the far right) from the rest of the clusters. The first newly discovered gene belonging to this cluster is BF549650. This gene is still unknown in terms of its identity and/or functionality. It is positioned closest to genes AI179795 and AI233194. Both genes are known as zinc transporters and are affected to a similar extent by the $HgCl_2$ treatment at 2 hours (approximately a +2.50 fold change). (Note that zinc and mercury occupy the same column in the Periodic Table and hence share similar properties). Could this hint at a common or closely related functionality or pathway of these genes?

As in the previous Cluster dendogram (Figure 3.4), for the top 50 differentially expressed genes under IR treatment at 2 and 8 hours, here in this Cluster dendogram (Figure 3.5) for the corresponding top 50 differentially expressed genes under the $HgCl_2$ treatment, the newly discovered gene, AF149118, appears again (in the cluster on the far left). Entrez identifies it as ADAM metallopeptidase with thrombospondin type 1 motif. It shows fold-changes of +7.63 and +9.69 at 2 and 8 hours of IR treatment respectively. It was also found as an $HgCl_2$ differentially expressed gene, showing fold-changes of +5.68 and +1.44 at 2 and 8 hours of $HgCl_2$ treatment respectively. In this Cluster dendogram, similar to the Cluster dendogram for the IR treatment, this AF149118 gene is positioned, again, closest to another gene which has the same function as in the previous dendogram, although not carrying the same accession number. The M63282 gene, (which was also newly discovered by our analysis), is now identified by Entrez as Activating transcription factor 3, Atf3 (same as gene NM_012912 from Figure 3.4), which is also a differentially expressed gene under both IR and $HgCl_2$ treatments (showing fold-changes of +72.52 and +38.78 at 2 and 8 hours of $HgCl_2$ treatment and +128.49 and +44.37 at 2 and 8 hours of IR treatment respectively). The consistently tight clustering of the newly discovered gene, AF149118, with the two Atf3 (NM_012912 and M63282 genes) may hint at a common or related pathway shared by these genes.

The dendogram also positions two single genes (appearing on the left) as separate than all the rest of the clusters. These are considered outliers. One of them, NM_012580 is known as Heme oxygenase (decycling)1 (Hmox1). It is highly affected by the $HgCl_2$ treatment, showing a +79.6 fold-change at 2 hours, and +91.2 fold-change at 8 hours. The gene next to it, U07971, is a newly discovered gene, showing a -2.23 fold-change at 2 hours and -2.61 fold-change at 8 hours of $HgCl_2$ treatment. These two show opposite expression behavior both in terms of up versus down regulation and in terms of order of magnitude. Their p-values are small, 3.57e-05 for NM_012580, and 8.65e-04 for U07971. While the NM_012580 gene is clearly an outlier, gene U07971 is not, and it is unclear why it was positioned as such.

In order to test for the effect of time on expression patterns of the genes, we compared the average difference between normal arrays (serving as time 0 hours measurements) and treatment groups at 2 hours and at 8 hours separately. The results were first displayed as Q-Q plots. A Q-Q plot is a probability plot, which is a graphical method for comparing two sample probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles are chosen. A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). A Q-Q plot is an order statistic plot, so information regarding the identity of genes is lost.

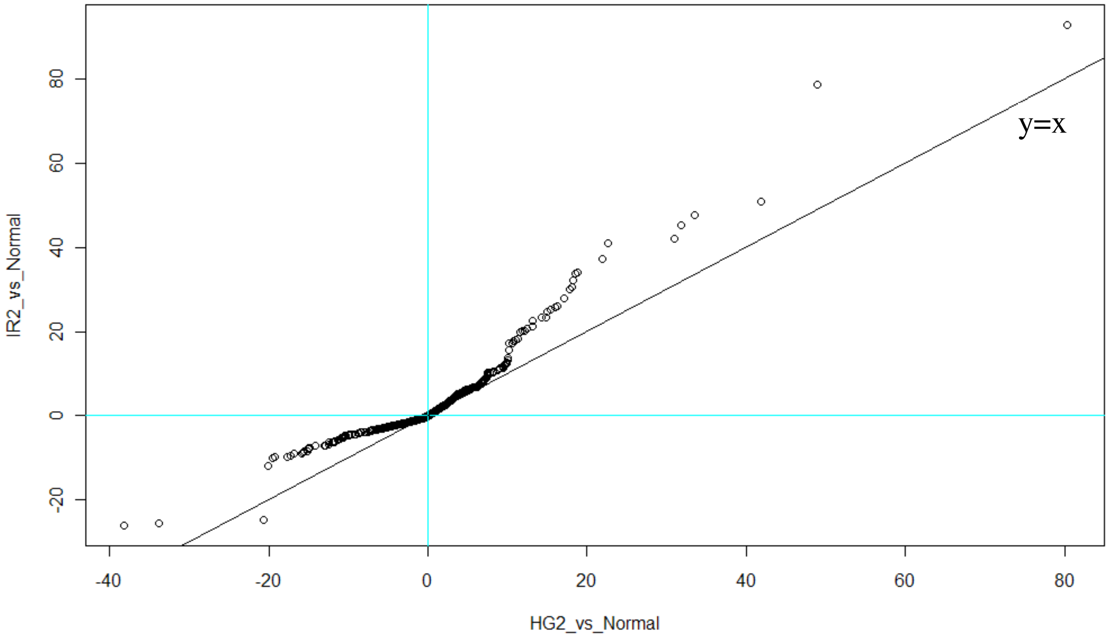

Figure 3.6: Q-Q plot comparing average difference between $HgCl_2$ and IR treatment groups at 2 hours (HG2 and IR2) and Normal arrays (o hours) yielded by **`limma`**

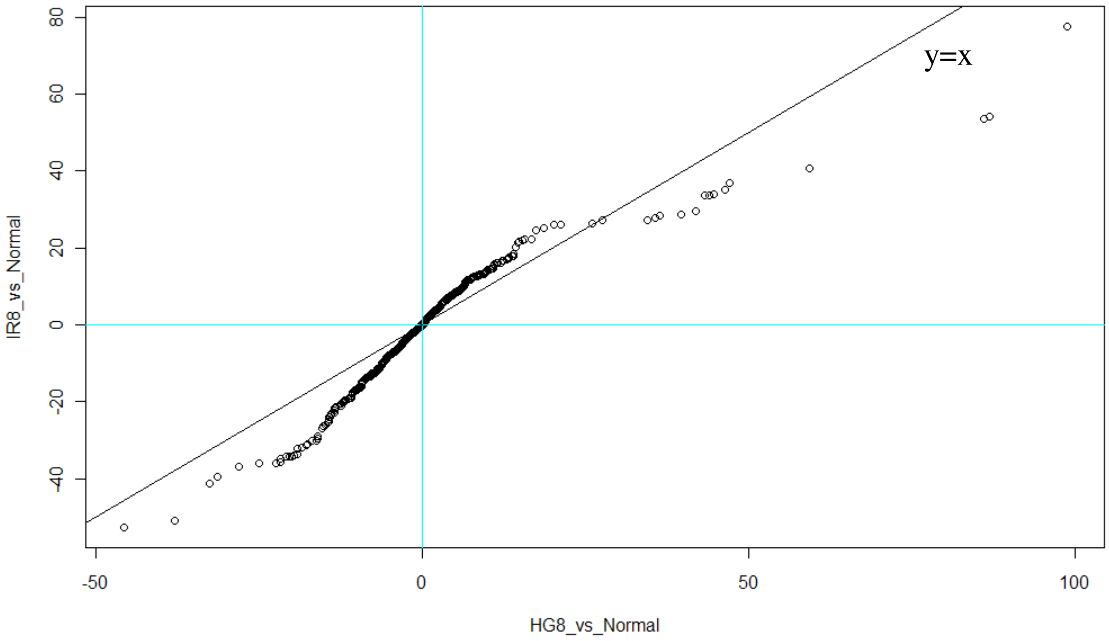

Figure 3.7: Q-Q plot comparing average difference between $HgCl_2$ and IR treatment groups at 8 hours (HG8 and IR8) and Normal arrays (o hours) yielded by **`limma`**

The line $y = x$ is used as a reference indicating the case where the two distributions being compared are similar.

The Q-Q plots for the 2 hour and 8 hours show very different patterns and the shape of distribution appears to be very different.

The 2 hour Q-Q plot shows different effect for $HgCl_2$ than for IR. The 8 hour Q-Q plot shows no unusual behavior in its middle region, but out in the edges expression levels are up in first quadrant, and also the lines bend. The graph is above the $y = x$ line. The slope of the $y = x$ line represents the ratio of standard errors and the Q-Q plot is indicating that there is a difference in variation. IR8 has more variation than HG8 ($HgCl_2$ at 8 hours).

The difference in scale between the $HgCl_2$ at 2 hours (-40 to +80) and the $HgCl_2$ at 8 hours (-50 to +100) may indicate that the effect of $HgCl_2$ may be greater at 8 hours. The two densities also show obvious differences in terms of more outliers/genes that respond more strongly at 8 hours.

The nature of response seems to be different between IR and $HgCl_2$. The graphs show that genes that are affected by IR are different than those affected by $HgCl_2$.

A normal Q-Q plot comparing randomly generated, independent standard normal data on the vertical axis to quantiles of a standard normal population on the horizontal axis will fluctuate randomly from the $y = x$ (45º) line. The linearity of the points along the line would suggest in that case that the data are normally distributed. The Q-Q plots in Figures 3.6 and 3.7 show a difference from what a normally-distributed graph will look like. This may serve as indication to a departure from normality of the data.

Whereas a Q-Q plot is an ordered statistic plot in which information regarding the identity of genes is lost, a scatter-plot portrays that information as well. A scatter-plot will tell us to what extent there might be an association between response and $HgCl_2$ vs. IR.

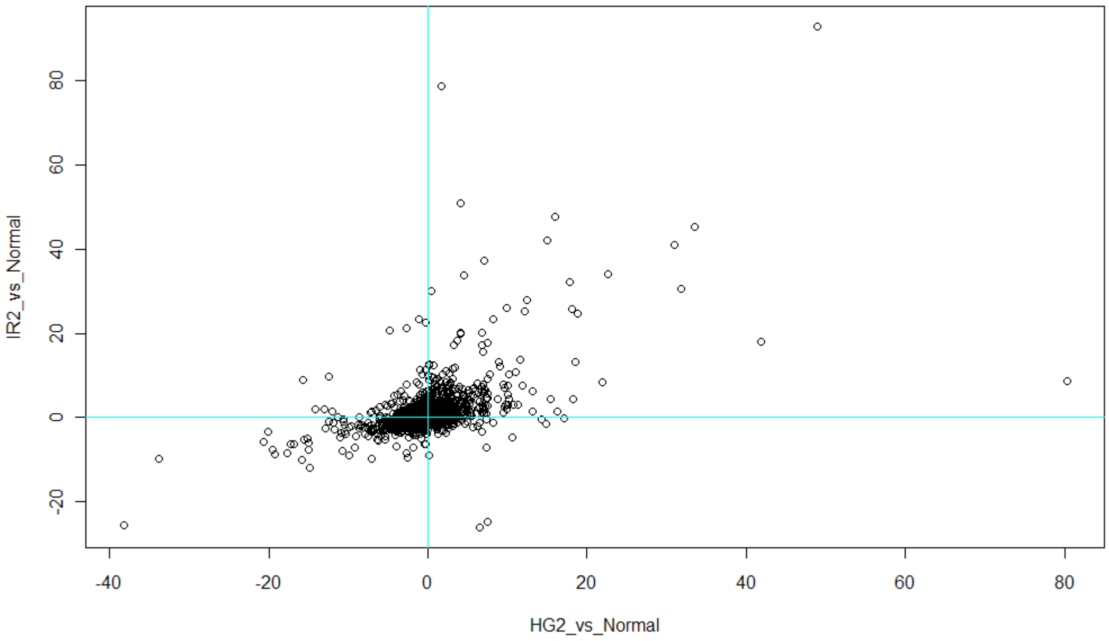

Figure 3.8: Scatter plot comparing average difference between $HgCl_2$ and IR treatment groups at 2 hours (HG2 and IR2) and Normal arrays (o hours) yielded by **`limma`**

Figure 3.8 is a scatter plot comparing the average difference between treatment groups at 2 hours ($HgCl_2$ and IR) and Normal arrays (o hours) yielded by **`limma`**. The scatter plot shows that there are outliers in both directions (far out in the diagonal). The outliers located along the y axis are hardly affected by $HgCl_2$. Genes affected by either treatment are concentrated around the y=x line. Appearing mostly in 1st and 3rd quadrants, most outliers seem to be either up- or down- regulated under both treatments at 2 hours. The graph indicates that the tendency is for over-expression (because of the asymmetry in the scatogram). Scale is -40 to +80.

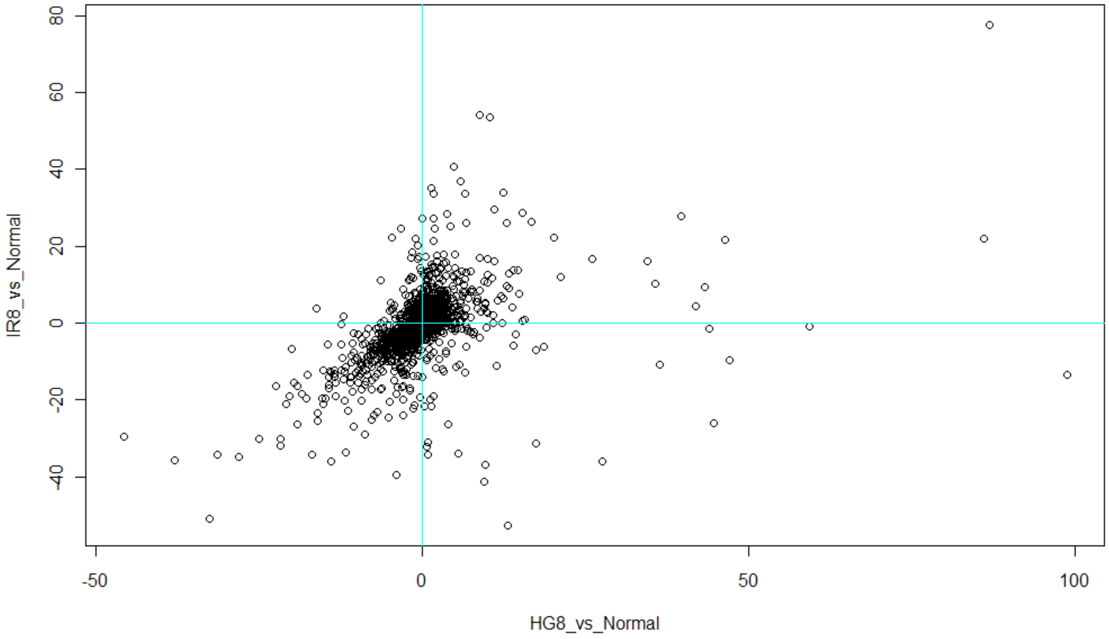

Figure 3.9: Scatter plot comparing average difference between $HgCl_2$ and IR treatment groups at 8 hours (HG8 and IR8) and Normal arrays (o hours) yielded by **`limma`**

Figure 3.9 is a scatter plot comparing the average difference between treatment groups at 8 hours ($HgCl_2$ and IR) and Normal arrays (o hours) yielded by **`limma`**. The scatter plot shows that there are outliers in both directions (far out in the diagonal). The outliers located along the y axis are hardly affected by $HgCl_2$. Genes affected by either treatment are concentrated around the y=x line. Appearing in 1st , 2nd, and 3rd quadrant, most outliers seem to be either up/down- regulated under both treatments at 8 hours, or up-regulated by $HgCl_2$ and down-regulated by IR at 8 hours. The graph indicates that the tendency is for over-expression (because of the asymmetry in the scatogram). Scale is -50 to +100.

The scatter plots show much more points in the second and third quadrants at 8 hours versus at 2 hours. Also more extreme measurements on second quadrant appear on the 8 hours graph, though it has different scale (Scale of HG8 is -50 versus -40 in 2 hours for x axis).

These differences between the scatter-plots at 2 and 8 hours may be further explained as to their significance by biologists.

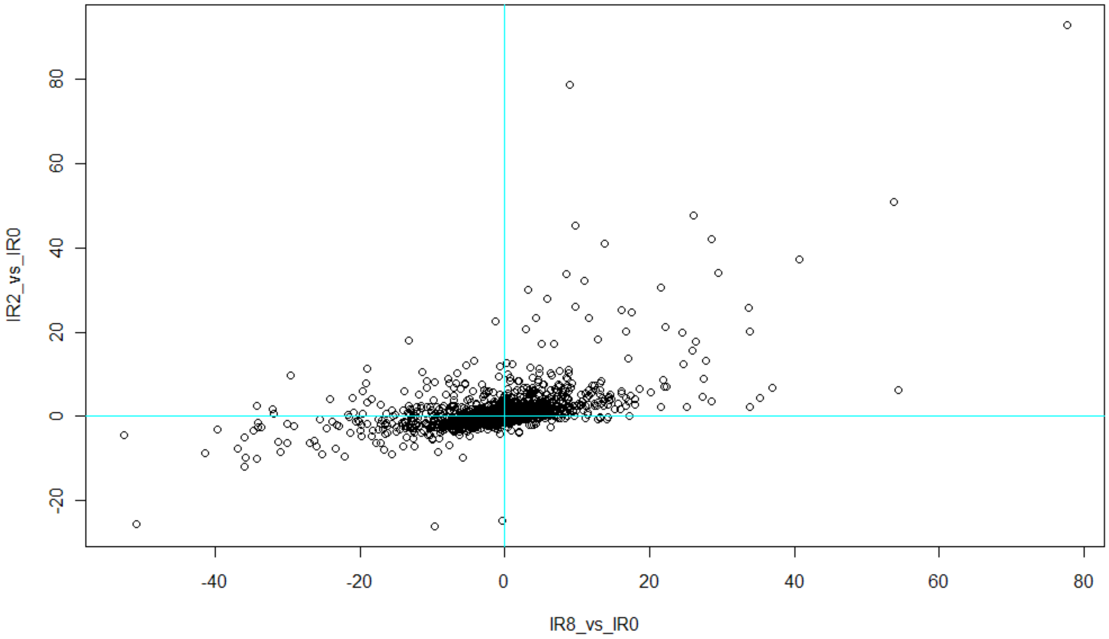

Figure 3.10: Scatter plot comparing average difference between IR at 2 and 8 hours and Normal arrays (o hours) yielded by **`limma`**

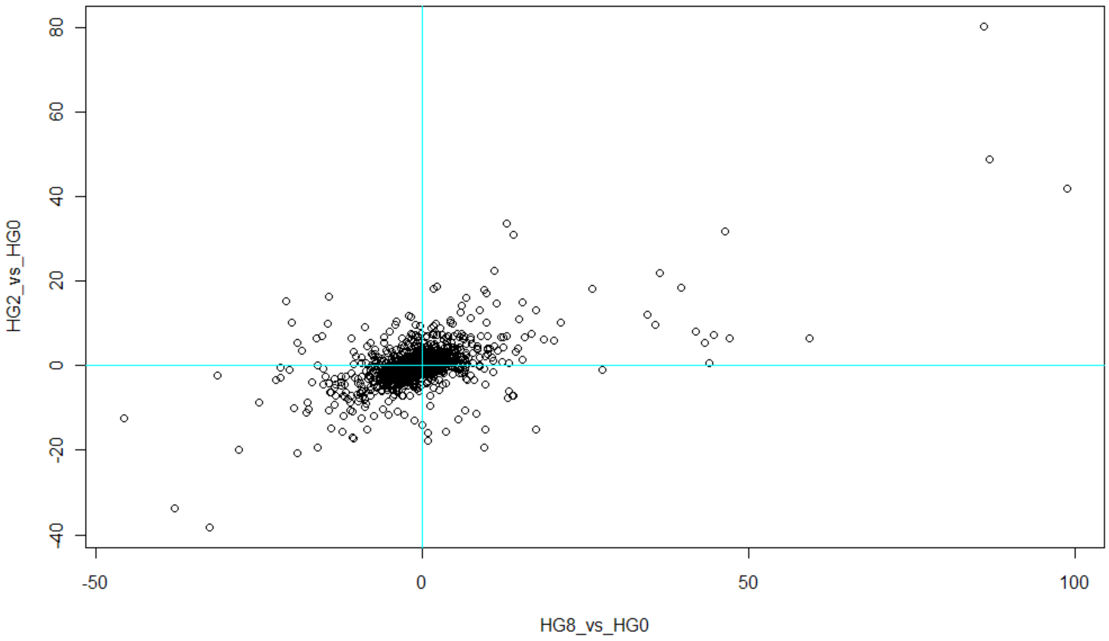

Figure 3.11: Scatter plot comparing average difference between $HgCl_2$ (HG) at 2 and 8 hours and Normal arrays (o hours) yielded by **`limma`**

Figures 3.10 and 3.11 are time-effect graphs. These scatter plots compare the average difference between a common treatment group (either IR or $HgCl_2$) at 2 and 8 hours and Normal arrays (o hours) yielded by **`limma`**.

The $HgCl_2$ scatter-plot looks as expected. It shows a response which is concentrated approximately along a 45 degree line. This shape of the graph typically portrays a response mechanism in which a longer exposure to a toxin results in a more extreme response. Given that the $HgCl_2$ treatment protocol includes injecting 4 mg/kg mercuric chloride, and then harvesting after 2h or 8h, the graph seems to clearly be showing the expected response to that type of treatment.

The IR graph is surprisingly looking. It looks very flat. If it was a flat line completely it would have indicated not much correlation between x and y. But the influence of outliers is substantial. The bulk of the points are showing less difference. There seems to be much variation in IR8 scale, not much in IR2 scale, which may indicate a tendency for more variation at 8 hours versus 2 hours.

There seem to be many outliers showing a peak at 2 hours and a decline at 8 hours, but at the same time, a significant amount of outlier genes show the opposite pattern (peaking at 8 hours and to a lesser degree at 2 hours).

The IR treatment protocol includes 40 minutes bilateral ischemia, 2h or 8h reperfusion, then harvest. Ischemia is a restriction in blood supply to tissues, causing a shortage of oxygen and glucose needed for cellular metabolism. The ischemia is then followed by reperfusion, which is the tissue damage caused when blood supply returns to the tissue after a period of ischemia or lack of oxygen. There could be a cell repair mechanism operating, in which some genes are needed for cellular pathways being completed within 2 hours, while other (repair?) genes get activated only after 8 hours. Also, since the cells are then harvested at either 2 or 8 hours, at which point their gene expression is measured, we are only seeing the measurements of those two time points. It is possible that if measurements had been taken at time point = 6 hours, that information could supply additional details and insights as to how the cellular repair mechanism operates.

While the two scatter-plots in Figures 3.10 and 3.11 both show that the extreme genes are mostly over-expressed, they, at the same time clearly depict two very different repair mechanisms for the IR and $HgCl_2$ treatments.

3.4.2 **`Multtest`** Analysis

In the analysis based on **`multtest`**, step-down maxT was chosen for FWER control, and the test statistics' null distribution was estimated by the permutation resampling method (the bootstrap option was not an available application yet).

The MTP procedure performed compared the 22 treatments (Normal, n=4; $HgCl_2$ at 2 hours, n=3; $HgCl_2$ at 8 hours, n=3; Volume Depletion, n=3; Ischemia/Reperfusion at 2 hours, n=4; Ischemia/Reperfusion at 8 hours, n=2; and Sham, n=3) and found, at level alpha = 0.05, 928 rejections (equivalent to differentially expressed genes) under the TPPFP procedure (with q = 0.1, which is the default), and 622 rejections under the FDR procedure.

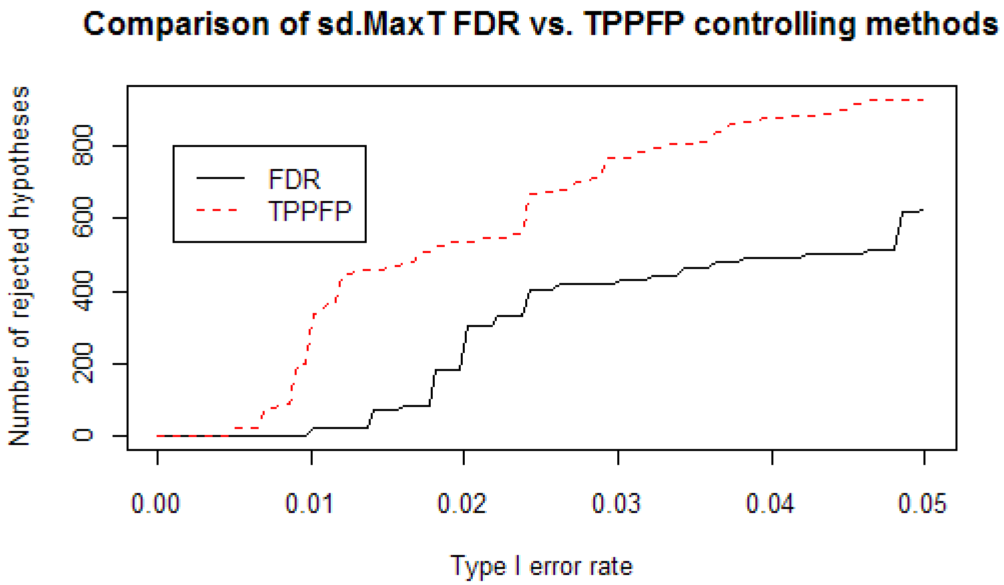

Figure 3.12: Comparison of TPPFP and FDR procedures, based on the FWER-controlling permutation-based step-down maxT procedure.

Figure 3.12 depicts the comparison in performance between TPPFP and FDR. It shows that for significant levels of alpha, TPPFP finds more rejected hypotheses than FDR, which seems to be more conservative. This power advantage of TPPFP seems to persist up to Type I error rate of 0.65.

In contrast to the FDR-controlling approach that focuses on the expected value of the proportion of false positives among the rejected hypotheses, the TPPFP procedure controls the tail probabilities for this proportion. According to Genovese and Wasserman (2004a,b), Korn et al. (2004), van der Laan et al. (2004b, 2005), Lehmann and Romano (2005), and Romano and Wolf (2005), although FDR-controlling approaches control the proportion of false positives (PFP) on average, they do not preclude large variations in PFP. When one wishes to have high confidence (i.e., chance at least (1-α)) that the set of rejected null hypotheses contains at most a specified proportion q of false positive, control of the tail probability for the proportion of false positives (TPPFP) is the appropriate form of Type I error control. The parameter q confers flexibility to TPPFP-controlling MTPs and can be tuned to achieve a desired level of false positives.

The graphs in Figure 3.12 also show an interesting finding: for very small α's, TPPFP seems to find a significantly greater amount of rejected hypotheses. For example, running the MTP function with both TPPFP and FDR procedures at α level equal 0.01, yielded 337 rejections under the TPPFP procedure and 23 rejections under the FDR procedure.

The hypotheses rejected by either TPPFP or FDR procedures correspond to genes whose expression levels were found to be significantly different from the rest of the genes. These are considered the differentially expressed genes and the ones with the respectively smallest p-values are the most significant ones.

Table A3, found in the Appendix section, specifies the 50 most significant differentially expressed genes among the 22 arrays as found by the TPPFP procedure, at 0.05 alpha level (with q = 0.1, which is the default). The table includes their p-values, biological identity/function, and the corresponding fold-changes in reference to normal. The biological identity/function was searched in Entrez, which is NIH's NCBI Life Sciences search-engine (http://www.ncbi.nlm.nih.gov/sites/gquery). The fold-change is calculated as the ration of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

All 50 genes appearing on this TPPFP-based **multtest** analysis seem to be different than their normal level measurements, showing significant differential expression both in terms of their fold-changes (ranging from 2.01 to 91.2) and in terms of their p-values (ranging from 0.002 to 0.007). Since this is a permutation-based test (with a finite number of permutations), a discrete significance level is observed. Most of these genes are differentially expressed genes under the IR treatment, a fact which coincides with Yuen et al.'s findings as well.

The majority of the genes appearing in Table A3 (36 out of 50 or 72%) were found by Yuen et al. as being differentially expressed genes. The other 14 genes (28%) are new differentially expressed genes discovered by this analysis. All of them showing significant differential expression both in terms of their fold-changes (ranging from 2.38 to 100.68) and in terms of their p-values (ranging from 0.002 to 0.007).

The second most significant gene in Table A3 is BF415939, which is a new differentially expressed gene discovered by this analysis. Under the IR treatment, it shows a fold-change of +100.68 at 2 hours, and a fold-change of +12.35 at 8 hours. The **multtest** procedure assigns it a p-value of 0.002, and Entrez identifies it as FBJ osteosarcoma oncogene. This gene was actually identified also by the **limma** analysis for top differentially expressed genes for the IR treatment, as the most differentially expressed gene, with a p-value of 4.88e-06.

Some other interesting new differentially expressed genes discovered by this analysis are: gene L12025, which is ranked number 18 on Table A3. It has a p-value of 0.005, and the following fold-changes: +5.0 under IR2, +30.0 under IR8, and +4.60 under HG8 ($HgCl_2$ at 8 hours) Gene AF149118, is ranked number 31 on this list. It has a p-value of 0.007, and the following fold-changes: +7.63 under IR2, +9.69 under IR8, and +5.68 under HG2 ($HgCl_2$ at 2 hours). It is identified by Entrez as ADAM metallopeptidase with thrombospondin type 1 motif, 1, Adamts1. This gene was identified also by **limma** as a differentially expressed gene for both the IR and $HgCl_2$ treatments. Gene AF269251, ranked number 33 on the list, has a p-value of 0.007 and the shows fold-changes of +3.33 and +30.95 under the IR treatment at 2 and 8 hours respectively.

These genes clearly are highly significant differentially expressed genes, and may not have been identified by Yuen et al. partly due to the loss of information of 250 genes incurred at the step where genes with missing values were filtered out from the original 36 microarrays, and not from the 22 microarrays (as was done by this paper's analysis).

3.5 Difference in Performance between **limma** and **multtest**

3.5.1 Difference in Performance

Next, the relative performances of **limma** versus **multtest** were compared. Both procedures were run using FDR as their Type I error control, and two sets of data were analyzed: $HgCl_2$ treatments (2 hours and 8 hours) against the Normals, and comparison of all 22 treatments ($HgCl_2$, IR, Volume depletion, Sham and Normal treatments). The output of these analyses were compared, in terms of numbers of differentially expressed genes found, range of adjusted p-values, range of fold-change, and genes identified as differentially expressed by the two procedures.

Table A4, found in the Appendix section, specifies the 50 most significant differentially expressed genes among the 22 arrays as found by the **limma** procedure. The table includes their p-values, biological identity/function, and their corresponding fold-changes in reference to normal. . The biological identity/function was searched in Entrez, which is NIH's NCBI Life Sciences search-engine (http://www.ncbi.nlm.nih.gov/sites/gquery). The fold-change is calculated as the ration of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported).

When analyzing the $HgCl_2$ treatments against Normals, **multtest** found only 4 rejections, equivalent to identifying only 4 differentially expressed genes, at the alpha = 0.05 level. These genes are: AW251878, BF407511, NM_012580, and AW917197, and their adjusted p-values, as assigned by **multtest**, are: 0.022, 0.022, 0.022, and 0.030 respectively. **Limma**'s analysis of the same data set yielded a list of genes, whose top 100 differentially expressed genes had an adjusted p-value range of 3.57e-05 to 2.69e-03. Three of the four genes found by **multtest**, BF407511 (fold-change of +2.54 under $HgCl_2$ at 2 hours), NM_012580 (fold-change of +79.6 under $HgCl_2$ at 2 hours and +91.2 under $HgCl_2$ at 8 hours), and AW917197 (fold-change of -3.21 under $HgCl_2$ at 8 hours) appear on the **limma** list as its first, second and seventh most differentially expressed genes, having adjusted p-values of 3.57e-05, 3.57e-05, and 4.95e-04 respectively. The gene found as the most differentially expressed gene by **multtest**, AW251878 (fold-change of +3.24 under $HgCl_2$ at 8 hours) does not appear in the top 100 differentially expressed genes found by **limma**.

This comparison between **multtest** and **limma** revealed a significant difference in the number of genes identified as differentially expressed by the two procedures. Only 3% of **limma**'s list of top 100 differentially expressed genes were identified by **multtest**. For the same genes found by both procedures as differentially expressed, their respective adjusted *p*-values are significantly smaller as assigned by **limma** relative to the **multtest** procedure.

The analysis of all 22 treatments ($HgCl_2$, IR, Volume depletion, Sham and Normal treatments) by **limma** and **multtest** revealed the following differences:

At level alpha = 0.05, **multtest** found 622 rejections. Its list of top 25 differentially expressed genes had an adjusted p-value range of 0.004 to 0.014.

Limma's analysis of the same data set yielded a list of genes, whose top 25 had an adjusted *p*-value range of 1.52e-14 to 3.28e-09. The two lists of top 25 differentially expressed genes produced by the two different procedures had 9 genes (36%) in common. Comparison of the top 50 differentially expressed genes' lists had 21 genes (42%) in common, also following the same pattern of much smaller adjusted p-values range assigned by the **limma** procedure. Both lists produced by the two procedures included significant differentially expressed genes to the same extent, in terms of their fold-changes.

The output of **limma**, relative to that of **multtest**, seems to consistently have a very different order of magnitude in terms of its adjusted *p*-values. Under the **limma** procedure, there is also a significant difference between the adjusted and unadjusted values for p (unadjusted p-values tend to be smaller than their respective adjusted values by an average order of 10^{-3}). The gene lists as produced by the two procedures were different. At the very extreme, the two procedures don't necessarily identify same genes, but some are significant enough to appear in both. The ranking of significance levels between **multtest** and **limma** is different. There appears to be a loss of power with **multtest**, which seems to be a more conservative procedure.

3.5.2 The Difference in Mechanism

The **multtest** package implements widely applicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates, in testing problems involving general data generating distributions (with arbitrary dependence structures

among variables), null hypotheses, and test statistics. In this study, the permutation-based estimator of the null distributions of the test statistics (*t*- or *F*-statistics) null distribution was chosen. Procedures are provided to control Type I error rates defined as tail probabilities and expected values of arbitrary functions of the numbers of Type I errors and rejected hypotheses. These error rates include: the generalized family wise error rate, tail probabilities for the proportion of false positives among the rejected hypotheses, and the false discovery rate. Single-step and step-down common-cut-off (maxT) and common-quantile (minP) procedures, that take into account the joint distribution of the test statistics, are implemented to control the FWER. In addition, augmentation procedures are provided to control the gFWER, TPPFP, and FDR, based on any initial FWER-controlling procedure.
[http://www.bioconductor.org].

Limma uses linear models to analyze designed microarray experiments (Yang and Speed, 2003; Smyth, 2004). The approach requires two matrices to be specified, the design matrix and the contrast matrix. The first step is to fit a linear model. Each row of the design matrix corresponds to an array in the experiment and each column corresponds to a coefficient. One purpose of this step is to estimate the variability in the data. The contrast step allows the fitted coefficients to be compared in as many ways as there are questions to be answered, regardless of how many or how few these might be. The comparison of interest is the average difference between treatment groups and normal arrays.

The most popular form of adjustment is "FDR", which is Benjamini and Hochberg's method to control the false discovery rate (Benjamini and Hochberg, 1995).

3.5.3 Discussion

The **`limma`** and **`multtest`** procedures are clearly different. The *F*-test performed by the two procedures is not the same. The **`multtest`** procedure permutes the 9,900 dimensions of gene expression, yielding a calculated *F* statistic for the data. Then, it adjusts for multiplicity.

`Limma`, on the other hand, has an empirical Bayes step (that **`multtest`** does not) which shrinks the dnominators of the *F* statistics making them smaller testing significance. Its approach improves on the *t*-statistic in terms of not giving high rank to genes only because they have small sample variances. The empirical Bayes (eBayes) step also has an effect of down-weighing an outlier, so that the wilder p-values tends to get shrunk. **`Limma`** first modify *p*-values by eBayes, then identify significance among them by FDR.

`Multtest` uses permutation-based tests performed on ANOVA-like statistics. **`Limma`** performs ANOVA-like analyses, based on classical linear models theory. Permutation tests are valid under any distribution of the *Y*'s when the null hypothesis is true. Therefore *P*(Type I error) = under non normality. With nonnormal data, $F = MST/MSE$ may not follow the theoretical *F* distribution, so that we're not sure if we control *P*(Type I error). This is the appeal of permutation tests. Empirical Bayes is applied to the denominator of $F = MST/MSE$, shrinking MSE toward a prior value and increasing the denominator degrees of freedom. Hence values of F_{mod} are more significant than F.

The two procedures also represent two different approaches: **`limma`** assumes normally distributed data (classical linear model), and **`multtest`** is permutation-based. The Q-Q plots analyzing the data showed departure from normality in extremes. It can be speculated that one thing affecting the performance of **`limma`** and **`multtest`** is the non normality of the data. The loss of power occurring with **`multtest`** is a consequence of a protection against type I errors in non-normal data. The price one pays is reduced power (small sample size in our case). Are we finding such a severe nonnormality in data that **`limma`** is finding false findings? On the other hand, if we knew normality assumptions were met, then it could be concluded that **`limma`** is superior to **`multtest`**.

Is the difference in performance of the two procedures evidence of conservatism of permutations or liberalism of the linear models –based test? Since it is not known how much the apparent distribution affected the results, it is not possible to determine whether the reason for the difference in performance between **`limma`** and **`multtest`** is due to the superiority of the **`limma`** mechanism or to an inherent bias of **`limma`**.

Table 3.1 - Selected Newly Discovered Genes

	Accession #	**Adj_p**	**Found by Yuen et al**	**Description**	**Fold Change**
1	BF415939	4.88e-06	NOT	FBJ osteosarcoma oncogene (Fos)	+100.68 IR2 +12.35 IR8
2	AF269251	0.007	NOT	Interleukin 24 Il24	+3.33 IR2 +30.95 IR8
3	AB032419	7.61e-09	NOT	Early growth response 2 (Egr2)	+21.02 IR2 +11.81 IR8
4	AI179538	7.36e-09	NOT	Kruppel-like factor 4 (gut) (Klf4)	+9.54 IR2 +4.16 IR8
5	AF065147	0.005	NOT	Cd44 molecule (Cd44)	+8.71 IR8
6	U95368	0.005	NOT	Gamma-aminobutyric acid (GABA-A) receptor, pi	+7.71 IR8
7	AF001417	3.25e-08	NOT	Kruppel-like factor 6 (Klf6)	+16.67 IR2 +12.05 IR8 +4.21 HG2 +5.62 HG8
8	L12025	2.54e-08	NOT	Poliovirus receptor (PVR)	+8.50 IR2 +25.37 IR8 +4.93HG8
9	AF149118	4.07e-09	NOT	ADAM metallopeptidase with thrombospondin type 1 motif, 1 (Adamts1)	+7.63 IR2 +9.69 IR8 +5.68 HG2
10	BF408391	1.60e-10	NOT	BMP and activin membrane-bound inhibitor, homolog (Xenopus laevis) (Bambi)	+5.39 IR2 +4.01 IR8 +2.42 HG2

The table above portrays a selection of 10 most differentially expressed genes, which were discovered in the course of this paper's analysis.

In the top of the list, gene BF415939 (FBJ osteosarcoma oncogene, Fos), showing an enormous fold change under IR treatment, of +100.68 at 2 hours and +12.35 at 8 hours. It was identified by all types of multiple testing procedures employed in our study as either first or second most differentially expressed gene either among all the IR treatments and/or among all treatment groups.

Next on the list, genes AF269251 and AB032419, affected by the IR treatment at both 2 and 8 hours, and showing impressing fold changes at both time points. Their fold change pattern is opposite, one peaking at 8 hours (+30.95 fold change) while the other peaking at 2 hours (+21.02 fold change).

Another couple of interesting genes is gene AI795538 and gene AF001417 (the forth and seventh on the list). The first of which is affected by the IR treatment only at 2 and 8 hours, which the second is affected under both IR and $HgCl_2$ treatments, both to a significant degree. While their description characterizes them as closely related in functionality, Kruppel-like factors 4 and 6, both are significantly affected by the IR treatment but differently affected by the $HgCl_2$ treatment.

Gene AF065147, is identified as Cd44 molecule. Another gene, having the same function, had been identified by Yuen et al (and by another study – Huang Q et. al), accession number NM_012924.2, showing a similar fold change +8.22 under the same treatment (IR) at the same time point (8 hours). Could the two genes be related?

Chapter 4: Summary and Conclusions

4.1 Summary

In this dissertation, we have reviewed methods for dealing with multiple testing that arise in microarray analysis, involving thousands of genes and few subjects. We applied newly developed multiple testing methods to a real world microarray data set that was originally collected in Dr. Robert Star's NIH laboratory and was then analyzed by Dr. Star and his colleagues (Yuen et al. 2006), in an effort to identify differentially expressed genes serving as biomarkers that distinguish ischemic from nephrotoxic injury types. In the study, a total of 31 male rats were assigned to different experimental groups, in which rat kidney transcriptomes were compared at 2 and 8 hours after ischemia/reperfusion and after mercuric chloride injection.

After collecting the data, Yuen et al. (2006) preprocessed the data using classical statistical methods combined with a heuristic data-screening approach. Understanding the need for multiple testing, they eliminated genes that they thought were obviously non-differentially expressed and then subjected each of the remaining genes to several ANOVA-based analyses. Their protocols reduced the data to a total of 728 genes, which were categorized by individual or combined conditions and summarized in a table.

Our goal was to analyze the same data using more sophisticated tools. The analyses presented in this paper employed modern multiple-comparison procedures designed to control the proportion of type I errors among the rejected hypotheses in families of comparisons under simultaneous consideration. Our approach is entirely statistical, so it reduces any type of subjectivity in preprocessing the data. We eliminated genes where the data were incomplete, as did Yuen et al. (2006), but we used a more refined method so that we were able to include additional genes. We were then able to identify differentially expressed genes using well developed statistical methodologies, applying both the **`limma`** and **`multtest`** procedures in the R Bioconductor software ensemble, while addressing error control in the form of false discovery rate.

4.2 Conclusions

`Limma` analysis was performed comparing normal groups to the 2 hour- and 8 hour-treatments of Mercuric Chloride ($HgCl_2$) and Ischemia/Reperfusion (IR). The **`limma`** procedure compared average differences between treatment and normal groups, yielding a list of most differentially expressed genes, based on the adjusted *p*-values.

Our statistical analysis yielded a collection of differentially expressed genes, under both treatments and for each treatment separately as well. Many of these have been previously identified by Yuen et al. and other researchers, but others are new. The top 50 differentially expressed genes for the Mercury Chloride and Ischemia

Reperfusion treatments were found to show distinct patterns of expression, while presenting a substantial difference in magnitude. Their heatmap plots showed a clear distinction between the two groups. These genes were clustered using a hierarchical clustering algorithm, which yielded a notable separation between the clusters, while presenting a different expression pattern as well. Normal quantile plots and scatterplots both indicate different mechanisms operating for the two different treatments, as well as hints of non-normality of the data.

`Limma` analysis comparing $HgCl_2$ at 2 and 8 hours (HG2 and HG8), IR at 2 and 8 hours and Normal treatments yielded several newly discovered differentially expressed genes, among which are **AF149118**, **M55534**, and **BF415939**, which were clearly classified as belonging to a separate cluster by the clustering algorithm. The height of that cluster differs greatly from the height of the rest of the clusters of genes, indicating a significantly different pattern of expression for that clustered collection of genes. Their p-values are very small (ranging between 1.62e-14 and 4.89e-09), and their corresponding fold changes are highly significant (ranging between +4.26 and +100.68).

The Cluster dendogram of top 50 differentially expressed genes yielded by **`limma`** comparison of IR 2 and 8 hours to the Normal treatments describes two of these genes, **M55534** and **BF415939,** as clustered together. A search in Entrez, NIH's NCBI Life Sciences search-engine identified new gene **M55534** as Crystallin, alphaB (Cryab), which seems to be highly similar to a heat-shock protein that's alpha-crystallin-related. Within that same cluster appears gene M14050, which had been identified by Yuen et al. as Heat shock 70kD protein (Hspa5), located on chromosome number 3 of the Rattus Norvegicus genome. And a another newly discovered gene within the same cluster, L16764, identified by Entrez as a Heat shock 70kD protein 1A (Hspa1a), located on chromosome number 20 of the Rattus Norvegicus (Norway rat) genome.

The common functionality of the clustered genes demonstrates how the clustering served as a helpful tool in identifying the deciphering the function of newly discovered, unknown genes.

The newly discovered gene **BF415939** is identified as FBJ osteosarcoma oncogene (Fos), located on chromosome number 6 of the Rattus Norvegicus genome. The **`limma`** procedure for top differentially expressed IR genes assigned a significantly small p-value to it (4.88e-06), and it shows impressive fold changes of +100.68 and +12.35 at IR 2 hours and 8 hours respectively. This gene is also the second most significant gene in **`multtest`**'s top 50 differentially expressed genes list, where it was assigned a p-value of 0.002.

Gene **AF149118** was identified by Entrez as ADAM metallopeptidase with thrombospondin type 1 motif. It shows fold-changes of +7.63 and +9.69 at 2 and 8 hours of IR treatment respectively. It was also found as an $HgCl_2$ differentially expressed gene, showing fold-changes of +5.68 and +1.44 at 2 and 8 hours of $HgCl_2$ treatment respectively. This gene is positioned next to gene NM_012912 (known as Activating transcription factor 3, Atf3), which is also a differentially expressed gene under both IR and $HgCl_2$ treatments (showing fold-changes of +14.6 at 2 hours of IR treatment and +11.9 and +9.80 at 2 and 8 hours of $HgCl_2$ treatment respectively). The tight clustering of these two genes may hint at a common or related pathway shared by the two genes.

In the $HgCl_2$ top 50 differentially expressed cluster dendogram, similar to equivalent IR treatment, this **AF149118** gene is positioned, again, closest to another gene which has the same function as in the previous dendogram, although not carrying the same accession number. The M63282 gene, (which was also newly discovered by our analysis), is now identified by Entrez as Activating transcription factor 3, Atf3 (same as gene NM_012912 from Figure 3.4), which is also a differentially expressed gene under both IR and $HgCl_2$ treatments (showing fold-changes of +72.52 and +38.78 at 2 and 8 hours of $HgCl_2$ treatment and +128.49 and +44.37 at 2 and 8 hours of IR treatment respectively). The consistently tight clustering of the newly discovered gene, AF149118, with the two Atf3 (NM_012912 and M63282 genes) may hint at a common or related pathway shared by these genes. Gene **AF149118** was also ranked number 31 on the **`multtest`** list showing a p-value of 0.007.

By using purely statistical analyses of the microarray results we are able to point out differentially expressed genes that went unrecognized before, that is, previous false negatives. Different types of assaults yield cell damage which can be detected in DNA. We conclude that once such a treatment is applied, technology can detect its effects. Using microarray technology and sophisticated multiple testing machinery we can pinpoint and quantify these effects.

4.3 Comparison of **`limma`** *and* **`multtest`**

In the course of our analysis, we came across an interesting finding regarding the relative performance of the **`limma`** versus the **`multtest`** procedures.

Application of **`multtest`** and **`limma`** to the same data revealed a significant difference in the numbers of genes identified by the two procedures as differentially expressed. While **`limma`** identified a large number of genes as differentially expressed, at a given overall FDR level, **`multtest`** identified only a few. For the genes identified by both procedures as differentially expressed, their respective adjusted *p*-values are significantly smaller as assigned by **`limma`** relative to the **`multtest`** procedure. When analyzing the $HgCl_2$ treatments against Normals, for example, only 3% of **`limma`**'s list of top 100 differentially expressed genes were identified by **`multtest`**. **`Multtest`** identified only 4 differentially expressed genes, having adjusted p-values, as assigned by **`multtest`**, ranging between 0.022 and 0.030. On the other hand, **`limma`**'s analysis of the same data set yielded a list of genes, whose top 100 differentially expressed genes had an adjusted p-value range of $3.57\text{x}10^{-5}$ to $2.69\text{x}10^{-3}$. Three of the four genes found by **`multtest`** appeared on the **`limma`** list as its first, second and seventh most differentially expressed genes, having adjusted p-values of $3.57\text{x}10^{-5}$, $3.57\text{x}10^{-5}$, and $4.95\text{x}10^{-4}$ respectively. The gene found as the most differentially expressed gene by **`multtest`**, AW251878 (fold-change of +3.24 at $HgCl_2$ 8 hours) does not appear among the top 100 differentially expressed genes found by **limma**.

The output of **limma**, compared to that of **multtest**, consistently has very different orders of magnitude of its adjusted *p*-values. The gene lists produced by the two procedures were different. The ranking of significance levels between **multtest** and **limma** is different. There appears to be a loss of power with **multtest**, which seems to be a more conservative procedure.

The two procedures are based on very different principles and represent two different approaches: **limma** assumes normally distributed data. It relies on a combination of normal theory analysis of linear models together with an empirical Bayes approach and other shrinkage methods used to borrow information across genes making the analyses stable even for experiments with small number of arrays (Smyth, 2004). The **multtest** algorithm makes no assumptions. It is permutation-based and as such, it has reduced power (Pollard, Dudoit and van der Laan , 2004). The Q-Q plots analyzing the data showed departure from normality in extremes. It can be speculated that one thing affecting the performance of **limma** and **multtest** is the nonnormality of the data. The loss of power occurring with **multtest** is a consequence of a protection against type I errors in non-normal data. The price one pays is reduced power, particularly in our small sample study. Are we finding such a severe nonnormality in data that **limma** is finding false findings? On the other hand, if we knew that normality assumptions were met, then it could be concluded that **`limma`** is superior to **`multtest`**.

Is the difference in performance of the two procedures evidence of conservatism of permutations or liberalism of the linear models –based test? Since it is not known how much the apparent distribution affected the results, it is not possible to determine whether the reason for the difference in performance between **`limma`** and **`multtest`** is due to the superiority of the **`limma`** mechanism or to an inherent bias of **`limma`**.

4.4 Future Research

Multiple testing procedures allow one to assess the overall significance of the results of a family of hypothesis tests. They focus on specificity by controlling type I error rates such as the family-wise error rate or the false discovery rate (Dudoit et al., 2003). Still, multiple testing remains a problem, because an increase in specificity, as provided by the p-value adjustment methods, is coupled with a loss of sensitivity, that is, a reduced chance of detecting true positives. Furthermore, the genes with the most drastic changes in expression are not necessarily the "key players" in the relevant biological processes. This problem can only be addressed by incorporating prior biological knowledge into the analysis of microarray data, by Bayesian techniques, which may lead to focusing the analysis on a specific set of genes.

In summary, microarrays are used in a wide variety of experimental settings for the detection of differential gene expression. Although the goals and design concern of these experiments vary, concepts including gene filtering, multiple comparisons adjustment, and gene selection according to the appropriate test statistic apply in general to these experiments. The Bioconductor packages help to address these concerns, thereby providing insight into biological pathways and providing a platform for future hypothesis development.

Appendix

Table A1 – Top 50 DE genes under IR treatment at_2 and 8 hours as found by limma procedure

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
1	BF415939	4.88e-06	NOT	FBJ osteosarcoma oncogene (Fos)	+100.68 IR2 +12.35 IR8
2	AW915240	5.73e-06	✓	V-fos FBJ murine osteosarcoma viral oncogene homolog (Fos)	+ 63.70 IR2 + 8.13 IR8
3	L16764	5.02e-05	NOT	Heat shock 70kD protein 1A (Hspa1a)	+69.24 IR2 +58.04 IR8
4	BF285303	5.93e-05	✓	Enigma homolog (Enh)	+ 2.59 IR2 + 2.68 IR8
5	AI236772	5.93e-05	✓		+3.78 IR2
6	D86345	5.93e-05	NOT	Leukemia inhibitory factor receptor alpha (Lifr)	+3.99 IR2
7	BE117902	5.93e-05	✓		+ 3.70 IR2
8	X06769	6.03e-05	✓	c-Fos	+ 27.6 IR2 + 3.48 IR8
9	BF282554	6.30e-05	✓		+ 5.51 IR2 + 4.26 IR8
10	AI406660	8.74e-05	✓		+ 3.34 IR2
11	AI411375	8.74e-05	✓	V-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (Ets2)	+ 2.66 IR2 + 2.79 IR8
12	NM_012548	9.98e-05	✓	Endothelin 1 (Edn1)	+ 9.31 IR2
13	AA799331	1.11e-04	✓	Pelota homolog (Pelo)	+ 2.34 IR2 + 2.44 IR8
14	AW919666	1.11e-04	✓	Similar to LIM and cysteine-rich domains 1	+ 2.58 IR2
15	NM_019361	1.11e-04	✓	Activity regulated cytoskeletal-associated protein (Arc)	+ 4.79 IR2 + 2.35 IR8

1. **Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.**

2. **Blank entries in the Description column represent an unidentified gene.**

3. **Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.**

4. **Blank entries in the Fold Change column indicate a non-significant fold change.**

Table A1 – Top 50 DE genes under IR treatment at_2 and 8 hours as found by limma procedure – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
16	M14050	1.11e-04	✓	Heat shock 70kD protein 5 (Hspa5)/Immunoglobulin heavy chain binding protein (BiP)	+ 2.71 IR2
17	AW142654	1.11e-04	✓		+ 5.16 IR2 + 4.17 IR8
18	BF408391	1.26e-04	NOT	BMP and activin membrane-bound inhibitor, homolog (Xenopus laevis) (Bambi)	+5.39 IR2 +4.01 IR8
19	BF406752	1.26e-04	✓	Similar to uridine phosphorylase	+ 4.33 IR2 + 5.79 IR8
20	AA875261	1.26e-04	✓	CSX-associated LIM (Cal)	+ 2.42 IR2 + 3.93 IR8
21	AF061266	1.38e-04	✓	Transient receptor protein 1 (Trrp1)	+ 2.75 IR2
22	AI012356	1.38e-04	NOT	Signal transducer and activator of transcription 3 (acute-phase response factor) (Stat3)	+1.95 IR2 +1.91 IR8
23	AI175031	1.38e-04	✓	Similar to DnaJ homolog subfamily B member 4	+ 4.88 IR2
24	NM_019372	1.38e-04	✓	Protein phosphatase 2C, magnesium-dependent, catalytic subunit (Ppm2c)	+ 2.23 IR2 + 3.14 IR8
25	BF420043	1.38e-04	✓		- 2.40 IR2 - 4.35 IR8
26	BE099875	1.38e-04	✓	Inositol 1,4,5-trisphosphate 3-kinase C (Itpkc)	+ 10.0 IR2
27	NM_013091	1.38e-04	✓	Tumor necrosis factor receptor superfamily, member 1a (Tnfrsf1a)	+ 2.27 IR2 + 3.27 IR8
28	AW251324	1.38e-04	✓	Similar to methylenetetrahydrofolate dehydrogenase (NAD) (EC 1.5.1.15) methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9) precursor – mouse	+ 8.02 IR2 + 12.5 IR8
29	BE109510	1.38e-04	NOT	Transmembrane and coiled-coil domains 6 (Tmco6)	-1.43 IR2 -1.74 IR8

1. **Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.**

2. **Blank entries in the Description column represent an unidentified gene.**

3. **Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.**

4. **Blank entries in the Fold Change column indicate a non-significant fold change.**

Table A1 – Top 50 DE genes under IR treatment at_2 and 8 hours as found by limma procedure – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
30	AF061873	1.38e-04	NOT	Transient receptor potential cation channel, subfamily C, member 1 (Trpc1)	+2.75 IR2
31	NM_012912	1.38e-04	✓	Activating transcription factor 3 (Atf3)	+ 14.6 IR2
32	NM_013052	1.38e-04	✓	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide (Ywhah)	+ 2.06 IR8
33	D31838	1.40e-04	✓	Wee1 tyrosine kinase	+ 3.20 IR2
34	BE101108	1.40e-04	✓		+ 2.43 IR2 + 2.12 IR8
35	NM_012904	1.52e-04	✓	Annexin A1 (Anxa1)	+ 3.33 IR2
36	AF149118	1.52e-04	NOT	ADAM metallopeptidase with thrombospondin type 1 motif, 1 (Adamts1)	+7.63 IR2 +9.69 IR8
37	AW916618	1.56e-04	NOT	Sphingosine kinase 1 (Sphk1)	+4.60 IR2 +3.15 IR8
38	BE097102	1.62e-04	✓	Similar to neuronal tyrosine threonine phosphatase 1	+ 5.69 IR2 + 3.15 IR8
39	AB032419	2.00e-04	NOT	Early growth response 2 (Egr2)	+21.02 IR2 +11.81 IR8
40	BE114586	2.00e-04	✓	Cyclin-dependent kinase inhibitor 1A (Cdkn1a)	+ 6.89 IR2
41	NM_021836	2.05e-04	✓	Jun-B oncogene (Junb)	+ 10.1 IR2
42	AI179988	2.05e-04	✓	Ectodermal-neural cortex 1 (Enc1)	+ 5.32 IR2 + 7.80 IR8
43	X63369	2.05e-04	NOT		+6.08 IR2 +2.99 IR8
44	AA848828	2.05e-04	✓		+ 2.04 IR2
45	AA799400	2.44e-04	✓	Similar to UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase III	+ 2.10 IR2 + 2.51 IR8
46	AF030091	2.46e-04	NOT	Cyclin L1 (Ccnl1)	+4.42 IR2 +3.24 IR8

1. **Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.**

2. **Blank entries in the Description column represent an unidentified gene.**

3. **Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.**

4. **Blank entries in the Fold Change column indicate a non-significant fold change.**

Table A1 – Top 50 DE genes under IR treatment at_2 and 8 hours as found by limma procedure – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
47	AI113186	2.46e-04	✓	Ras association domain family 1 (Rassf1)	+ 4.45 IR2 + 3.62 IR8
48	M55534	2.46e-04	NOT	Crystallin, alpha B (Cryab)	+4.00 IR2 +4.26 IR8
49	BE101099	2.48e-04	NOT	Zinc finger protein 36, C3H type-like 2 (Zfp36l2)	+1.99 IR2 +1.89 IR8
50	AW433959	2.52e-04	NOT	Myeloid-associated differentiation marker (Myadm)	+3.41 IR2

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A2 – Top 50 DE genes under HG treatment at_2 and 8 hours as found by limma procedure

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
1	BF407511	3.57e-05	✓	Ubiquitin specific protease 36 (predicted)	+ 2.54 HG2
2	NM_012580	3.57e-05	✓	NM_012580.1 Heme oxygenase (decycling) 1 (Hmox1)	+ 79.6 HG2 + 91.2 HG8
3	BF408391	1.29e-04	NOT	Bambi: BMP and activin membrane-bound inhibitor, homolog (Xenopus laevis)	+ 2.42 HG2 + 1.29 HG8
4	BE096387	4.95e-04	✓		+ 2.12 HG8
5	NM_019203	4.95e-04	✓	NM_019203.1 Testis specific X-linked gene (Tsx)	-2.69 HG2
6	AW526160	4.95e-04	✓	Myocyte enhancer factor 2D (Mef2d)	+ 2.07 HG8
7	AW917197	4.95e-04	✓		- 3.21 HG8
8	AF052042	4.95e-04	✓	AF052042 Zinc finger protein Y1 (RLZF-Y) (Rlzfy)	+ 2.33 HG2
9	AF016387	4.95e-04	NOT	Rattus norvegicus retinoid X receptor gamma (RXRgamma) mRNA, partial cds	-1.69 HG2 -1.78 HG8
10	AI179795	4.95e-04	✓	AI179795 Solute carrier family 30 (zinc transporter), member 1 (Slc30a1)	+ 2.62 HG2
11	AA946485	4.95e-04	✓	AA946485 Similar to TG interacting factor (Tgif)	+ 2.86 HG2 + 2.21 HG8
12	AI236753	7.20e-04	NOT	EST233315 Normalized rat ovary, Bento Soares Rattus sp. cDNA clone ROVDK16 3- end, mRNA sequence	+1.52 HG2 +1.65 HG8
13	AI175031	7.20e-04	✓	Similar to DnaJ homolog subfamily B member 4	+ 2.43 HG2 + 2.35 HG8
14	BF420059	7.20e-04	✓		+ 3.07 HG2 + 3.22 HG8
15	AA944278	7.20e-04	✓	AA944278 Similar to Isoleucyl-tRNA synthetase, cytoplasmic (Isoleucine--tRNA ligase) (IleRS) (IRS)	+ 2.37 HG8
16	BF420064	7.20e-04	NOT	Lkap: Limkain b1	+1.48 HG2 +1.04 HG8
17	BF396191	7.20e-04	✓		+ 2.34 HG2

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP $R_k(\alpha) = R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A2 – Top 50 DE genes under HG treatment at_2 and 8 hours as found by limma procedure – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
18	AW918999	7.20e-04	NOT	Dguok: Deoxyguanosine kinase	-1.79 HG2 -1.52 HG8
19	NM_012603	7.20e-04	✓	V-myc avian myelocytomatosis viral oncogene homolog (Myc)	+ 4.46 HG2 + 13.7 HG8
20	AF149118	8.14e-04	NOT	ADAM metallopeptidase with thrombospondin type 1 motif	+5.68 HG2
21	BE111769	8.27e-04	NOT	Trafficking protein, kinesin binding 1 (Trak1)	+1.84 HG2
22	AI599104	8.27e-04	✓		+ 5.17 HG2 + 6.12 HG8
23	BF555544	8.27e-04	NOT	Membrane protein, palmitoylated 5 (MAGUK p55 subfamily member 5) (Mpp5)	+1.29 HG2 +1.17 HG8
24	AA891690	8.27e-04	NOT	Tumor necrosis factor (ligand) superfamily, member 13 (Tnfsf13)	-1.67 HG2 -1.73 HG8
25	NM_012912	8.27e-04	✓	Activating transcription factor 3 (Atf3)	+ 11.9 HG2 + 9.80 HG8
26	AF220760	8.61e-04	NOT	Thioredoxin reductase 1 (Txnrd1)	+2.85 HG2 +4.00 HG8
27	AI137233	8.65e-04	✓	AI137233 Similar to sudD, suppressor of bimD6 homolog	+ 2.17 HG2 + 3.99 HG8
28	BE110525	8.65e-04	NOT		+1.54 HG2
29	M63282	8.65e-04	NOT	Activating transcription factor 3 (Atf3)	+72.52 HG2 +38.78 HG8
30	AI176298	8.65e-04	NOT		+1.73 HG2
31	BE102889	8.65e-04	NOT	Zinc finger protein 451 (Zfp451)	+1.70 HG2
32	AI599284	8.65e-04	NOT	Similar to hypothetical protein MGC30618 (RGD1305572)	+1.72 HG2 +1.64 HG8
33	U07971	8.65e-04	NOT	Glycine amidinotransferase (L-arginine:glycine amidinotransferase) (Gatm)	-2.23 HG2 -2.61 HG8
34	X53773	8.77e-04	NOT	Adaptor-related protein complex 2, alpha 2 subunit (Ap2a2)	-1.60 HG2 -1.28 HG8
35	AI407490	8.77e-04	✓	Similar to tyrosyl-tRNA synthetase	+ 2.32 HG2

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP $R_k\ (\alpha) = R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A2 – Top 50 DE genes under HG treatment at_2 and 8 hours as found by limma procedure – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
36	AF081941	9.34e-04	NOT	Adenylate cyclase 10 (soluble) (Adcy10)	-1.90 HG2 -1.33 HG8
37	BE101448	9.34e-04	✓	Similar to Cartilage-associated protein precursor	- 2.57 HG2 - 3.24 HG8 - 2.26 IR8
38	AW917596	9.34e-04	NOT	Transcribed locus, strongly similar to XP_003754401.1 PREDICTED: keratin, type II cytoskeletal 7 [Rattus norvegicus]	-1.71 HG2 -1.72 HG8
39	AI233194	9.56e-04	✓	Solute carrier family 30 (zinc transporter), member 1 (Slc30a1)	+ 2.47 HG2 + 2.09 HG8
40	BE109637	9.82e-04	✓	BE109637	- 3.19 HG8
41	BE112768	9.82e-04	NOT		+1.48 HG2
42	BE118465	1.09e-03	NOT		+1.91
43	AW142654	1.09e-03	✓		+ 3.42 HG2 + 2.52 HG8
44	AF249673	1.12e-03	NOT	Solute carrier family 38, member 2 (Slc38a2)	+2.18 HG2 +1.94 HG8
45	AA892366	1.12e-03	✓		- 2.98 HG8
46	AI180454	1.12e-03	NOT	Insulin-like growth factor 2 mRNA binding protein 2 (Igf2bp2)	+1.85 HG2
47	BF549650	1.12e-03	NOT		+2.06 HG2 +1.46 HG8
48	D90404	1.12e-03	NOT	Cathepsin C (Ctsc)	-1.82 HG2 -1.37 HG8
49	M98820	1.12e-03	NOT	Interleukin 1 beta (Il1b)	+3.00 HG2 +1.91 HG8
50	BE118450	1.35e-03	NOT	Retinoid X receptor gamma (Rxrg)	-1.73 HG2 -2.27 HG8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A3: Top 50 DE genes as found by multtest comparing all 22 treatments TPPFP-based procedure at α = 0.05

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
1	AA851327	0.002	✓	Similar to membrane protein expressed in epithelial-like lung adenocarcinoma	+ 10.1 IR8 +2.22 HG8
2	BF415939	0.002	NOT	FBJ osteosarcoma oncogene (Fos)	+100.68 IR2 +12.35 IR8
3	NM_012992	0.003	✓	Nucleophosmin 1 (Npm1)	+ 2.72 IR8
4	AF031483	0.005	NOT	Basic leucine zipper and W2 domains 2 (Bzw2)	+4.24 IR8
5	AF035963	0.005	✓	Kidney Injury Molecule 1 (Kim1)	+ 15.6 IR8
6	AF065147	0.005	NOT	Cd44 molecule (Cd44)	+8.71 IR8
7	AI169903	0.005	✓		+2.39 HG8 +2.14 IR8
8	AI236772	0.005	✓	Testis derived transcript (Tes)	+3.78 IR2
9	AI406499	0.005	✓	S100 calcium binding protein A16 (S100a16)	+3.94 IR8
10	AI575026	0.005	NOT	SH3 domain binding glutamic acid-rich protein-like 3	+3.52 IR8
11	AA875261	0.005	✓	CSX-associated LIM (Cal)	+ 2.42 IR2 + 3.93 IR8
12	AW915240	0.005	✓	V-fos FBJ murine osteosarcoma viral oncogene homolog (Fos)	+5.70 HG8 + 63.7 IR2 + 8.13 IR8
13	BE113365	0.005	✓	Ribosomal RNA processing 15 homolog (S. cerevisiae) (Rrp15)	+3.10 IR8
14	BE117902	0.005	✓		+2.07 HG2 +3.70 IR2
15	BF282554	0.005	✓		+5.51 IR2 +4.26 IR8
16	BF417071	0.005	✓	Similar to RING finger protein	+2.94 IR8
17	BF550451	0.005	✓	Similar to retinoic acid inducible protein 3	+11.0 IR2 +25.9 IR8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A3: Top 50 DE genes as found by multtest comparing all 22 treatments TPPFP-based procedure at α = 0.05 – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
18	L12025	0.005	NOT	Poliovirus receptor (PVR)	+8.50 IR2 +25.37 IR8 +4.93HG8
19	NM_012548	0.005	✓	Endothelin 1 (Edn1)	+9.31 IR2
20	NM_012580	0.005	✓	Heme oxygenase (decycling) 1 (Hmox1)	+ 9.93 IR2 + 23.2 IR8 + 79.6 HG2 + 91.2 HG8
21	NM_017022	0.005	✓	Integrin beta 1 (Itgb1)	+2.30 IR8
22	NM_019372	0.005	✓	Protein phosphatase 2C, magnesium-dependent, catalytic subunit (Ppm2c)	+ 2.23 IR2 + 3.14 IR8
23	U22893	0.005	✓	Cold shock domain protein A (Csda)	+3.81 IR8
24	U95368	0.005	NOT	Gamma-aminobutyric acid (GABA-A) receptor, pi	+7.71 IR8
25	X06769	0.005	✓	CFos	+ 27.6 IR2 + 3.48 IR8
26	AI406660	0.006	✓		+3.34 IR2
27	AF061266	0.007	✓	Transient receptor protein 1 (Trrp1)	+2.75 IR2
28	AF061873	0.007	NOT	Transient receptor potential cation channel, subfamily C, member 1 Trpc1	+2.73 IR2
29	AF063447	0.007	NOT	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A, Ddx39a, RNA helicase	+1.84 IR2 +3.44 IR8
30	AF063939	0.007	NOT	Trophoblast glycoprotein Tpbg	+1.89 IR2 +4.05 IR8
31	AF149118	0.007	NOT	ADAM metallopeptidase with thrombospondin type 1 motif, 1 Adamts1	+7.63 IR2 +9.69 IR8 +5.68 HG2
32	AF248543	0.007	NOT	Alpha 1,3-galactosyltransferase 2 A3galt2	+4.74 IR8
33	AF269251	0.007	NOT	Interleukin 24 Il24	+3.33 IR2 +30.95 IR8
34	AI009780	0.007	NOT	Hypothetical protein LOC682999	+2.38 IR8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP $R_k(\alpha) = R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A3: Top 50 DE genes as found by multtest comparing all 22 treatments TPPFP-based procedure at α = 0.05 – cont.

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
35	AI012474	0.007	✓	Estrogen-regulated protein CBL20, 20.4kD	+4.69 IR8
36	AI013474	0.007	✓	Similar to alpha/beta hydrolase-2 fold protein	+4.49 IR8
37	AI113186	0.007	✓	Ras association domain family 1 (Rassf1)	+ 4.45 IR2 + 3.62 IR8
38	AI137233	0.007	✓	Similar to sudD, suppressor of bimD6 homolog	+ 2.17 HG2 + 3.99 HG8
39	AI175031	0.007	✓	Similar to DnaJ homolog subfamily B member 4	+ 2.43 HG2 + 2.35 HG8 +4.88 IR2
40	AI177706	0.007	✓		+2.19 IR2 +3.77 IR8
41	AI179795	0.007	✓	Solute carrier family 30 (zinc transporter), member 1 (Slc30a1)	+2.62 HG2
42	AI179988	0.007	✓	Ectodermal-neural cortex 1 (Enc1)	+5.32 IR2 +7.80 IR8
43	AI180454	0.007	✓	Similar to IGF-II mRNA-binding protein 2	+ 2.84 IR8
44	AI406520	0.007	NOT		+3.44 IR8
45	AI407064	0.007	✓	Nucleolar protein 12 (Nol12)	+3.03 IR8
46	AI409108	0.007	✓	Transcribed locus, strongly similar to NP_001102090.1 nuclear pore complex protein Nup205 [Rattus norvegicus]	+2.14 IR8
47	AI411375	0.007	✓	V-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (Ets2)	+2.66 IR2 +2.79 IR8
48	AI579555	0.007	✓	Seryl-aminoacyl-tRNA synthetase (Sars1)	+2.01 HG2 +3.21 HG8
49	AJ011811	0.007	✓	Claudin 7 (Cldn7)	+3.00 IR2 +8.63 IR8
50	AW523504	0.007	✓	Similar to putative protein, with at least 9 transmembrane domains, of eukaryotic origin (43.9 kD) (2G415) (RGD1309228)	+2.30 IR8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP $R_k(\alpha) = R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A4: Top 50 DE genes as found by limma FDR-based procedure comparing all 22 treatments

	Accession #	Adj_p	Found by Huen et al	Description	Fold Change
1	BF415939	1.52e-14	NOT	FBJ osteosarcoma oncogene (Fos)	+100.68 IR2 +12.35 IR8
2	AW915240	3.00e-14	✓	V-fos FBJ murine osteosarcoma viral oncogene homolog (Fos)	+ 63.7 IR2 + 8.13 IR8 + 5.70 HG8
3	X06769	1.74e-11	✓	CFos	+ 27.6 IR2 + 3.48 IR8
4	BF282554	2.57e-11	✓		+ 5.51 IR2 + 4.26 IR8
5	AI236772	4.24e-11	✓	Testis derived transcript (Tes)	+ 3.78 IR2
6	BE117902	5.15e-11	✓		+ 3.70 IR2 + 2.07 HG2
7	NM_012548	1.16e-10	✓	Endothelin 1 (Edn1)	+ 9.31 IR2
8	AI406660	1.44e-10	✓		+ 3.34 IR2
9	BF408391	1.60e-10	NOT	BMP and activin membrane-bound inhibitor, homolog (Xenopus laevis) (Bambi)	+5.39 IR2 +4.01 IR8 +2.42 HG2
10	D86345	2.27e-10	NOT	Leukemia inhibitory factor receptor alpha (Lifr)	+3.98 IR2
11	BE099875	2.27e-10	✓	Inositol 1,4,5-trisphosphate 3-kinase C (Itpkc)	+ 10.0 IR2
12	NM_019361	2.27e-10	✓	Activity regulated cytoskeletal-associated protein (Arc)	+ 2.06 HG8 + 4.79 IR2 + 2.35 IR8
13	AI175031	2.49e-10	✓	Similar to DnaJ homolog subfamily B member 4	+ 2.43 HG2 + 2.35 HG8 + 4.88 IR2
14	M14050	6.34e-10	✓	Heat shock 70kD protein 5 (Hspa5)/Immunoglobulin heavy chain binding protein (BiP)	+ 2.71 IR2
15	AF061266	8.92e-10	✓	Transient receptor protein 1 (Trrp1)	+ 2.75 IR2
16	AW916618	1.09-09	NOT	Sphingosine kinase 1 (Sphk1)	
17	NM_012904	1.10e-09	✓	Annexin A1 (Anxa1)	+ 3.33 IR2

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}$, α).

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A4: Top 50 DE genes as found by limma FDR-based procedure comparing all 22 treatments – cont.

18	AF061873	1.25e-09	NOT	Transient receptor potential cation channel, subfamily C, member 1 (Trpc1)	+2.73 IR2
19	AW142654	1.27e-09	✓		+ 5.16 IR2 + 4.17 IR8 + 3.42 HG2 + 2.52 HG8
20	AI411375	1.80e-09	✓	V-ets erythroblastosis virus E26 oncogene homolog 2 (avian) (Ets2)	+ 2.66 IR2 + 2.79 IR8
21	AI113186	1.80e-09	✓	Ras association domain family 1 (Rassf1)	+ 4.45 IR2 + 3.62 IR8
22	AW433959	2.56e-09	NOT	Myeloid-associated differentiation marker (Myadm)	+3.40 IR2
23	BF285303	2.56e-09	✓	Enigma homolog (Enh)	+ 2.59 IR2 + 2.68 IR8
24	NM_019372	2.68e-09	✓	Protein phosphatase 2C, magnesium-dependent, catalytic subunit (Ppm2c)	+ 2.23 IR2 + 3.14 IR8
25	BF420059	3.29e-09	✓	Immediate early response 2 (Ier2)	+ 3.07 HG2 + 3.22 HG8 + 7.14 IR2 +3.83 IR8
26	AF149118	4.07e-09	NOT	ADAM metallopeptidase with thrombospondin type 1 motif, 1 (Adamts1)	+7.63 IR2 +9.69 IR8 +5.68 HG2
27	AF030091	4.85e-09	NOT	Cyclin L1 (Ccnl1)	+3.52 IR2
28	D31838	5.07e-09	✓	Wee1 tyrosine kinase	+ 3.20 IR2
29	NM_021836	5.57e-09	✓	Jun-B oncogene (Junb)	+ 10.1 IR2 + 2.96 HG8
30	NM_012912	7.19e-09	✓	Activating transcription factor 3 (Atf3)	+ 11.9 HG2 + 9.80 HG8 + 14.6 IR2
31	AI179538	7.36e-09	NOT	Kruppel-like factor 4 (gut) (Klf4)	+9.54 IR2 +4.16 IR8
32	AB032419	7.61e-09	NOT	Early growth response 2 (Egr2)	+21.02 IR2 +11.81 IR8
33	AI179988	9.63e-09	✓	Ectodermal-neural cortex 1 (Enc1)	+ 5.32 IR2 + 7.80 IR8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP R_k (α) = $R(T_k, Q_{ok}, α)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Table A4: Top 50 DE genes as found by limma FDR-based procedure comparing all 22 treatments – cont.

34	AA799400	1.00e-08	✓	Similar to UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase III	+ 2.10 IR2 + 2.51 IR8
35	NM_021846	1.17e-08	✓	Myeloid cell leukemia sequence 1 (Mcl1)	+ 2.97 IR2
36	U78875	1.52e-08	NOT	Kruppel-like factor 10 (Klf10)	
37	BF550451	1.60e-08	✓	Similar to retinoic acid inducible protein 3	+ 11.0 IR2 + 25.9 IR8
38	AI103943	1.60e-08	NOT	Ras association (RalGDS/AF-6) domain family member 1 (Rassf1)	+5.23 IR2 +4.64 IR8
39	AA875261	1.60e-08	✓	CSX-associated LIM (Cal)	+ 2.42 IR2 + 3.93 IR8
40	M55534	1.75e-08	NOT	Crystallin, alpha B (Cryab)	+4.00 IR2 +4.26 IR8
41	BF406752	1.94e-08	✓	Similar to uridine phosphorylase	+ 4.33 IR2 + 5.79 IR8 + 2.47 HG8
42	NM_013091	2.06e-08	✓	Tumor necrosis factor receptor superfamily, member 1a (Tnfrsf1a)	+ 2.27 IR2 + 3.27 IR8
43	AA848828	2.16e-08	✓		+ 2.04 IR2
44	L12025	2.54e-08	NOT	Poliovirus receptor (PVR)	+8.50 IR2 +25.37 IR8 +4.93 HG8
45	X59601	2.62e-08	NOT	Plectin (Plec)	+3.05 IR2
46	NM_012633	2.91e-08	✓	Peripherin 1 (Prph1)	+ 4.25 IR2 + 3.71 IR8 + 2.82 HG8 + 2.66 VD
47	AW434670	3.02e-08	✓	Similar to Xin	+ 13.4 IR2 + 5.84 IR8
48	AI176298	3.06e-08	✓		+2.74 IR2
49	X13722	3.22e-08	✓	Low density lipoprotein receptor (Ldlr)	+ 3.49 IR2
50	AF001417	3.25e-08	NOT	Kruppel-like factor 6 (Klf6)	+16.67 IR2 +12.05 IR8 +4.21 HG2 +5.62 HG8

1. Adj_p value is defined as the smallest Type I error level α at which one could reject $H_0(n)$, given an MTP $R_k(\alpha) = R(T_k, Q_{ok}, \alpha)$.

2. Blank entries in the Description column represent an unidentified gene.

3. Fold Change is calculated as the ratio of expression under treatment to expression under control circumstances. If ratio<1, then its reciprocal with a negative sign is reported.

4. Blank entries in the Fold Change column indicate a non-significant fold change.

Bibliography

1. Peter S. T. Yuen, Sang-Kyung Jo, Mikaela K. Holly, Xuzhen Hu, and Robert A. Star. "Ischemic and nephrotoxic acute renal failure are distinguished by their broad transcriptomic responses". *Physiological Genomics*. 25(3): 375-386, 2006.

2. Yuen PS, Jo S, Star RA. "Ischemic vs. nephrotoxic acute renal failure, early time points (2h and 8h)".
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3219.
17 Mar. 2006. Web. 12 July. 2010.

3. Gordon K. Smyth. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*. 3(1): Article 3, 2004.

4. K. S. Pollard, S. Dudoit, and M. J. van der Laan. *Multiple testing procedures and applications to genomics*. Technical Report 164, Division of Biostatistics, University of California, Berkeley, 2004.
URL www.bepress.com/ucbbiostat/paper164.

5. R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 2005.

6. S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Springer, New York, 2004.

7. I. Lönnstedt and T. Speed. "Replicated microarray data". *Statistica Sinica*, 12:31-46, 2002.

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

9. C. R. Genovese and L. Wasserman. "A stochastic process approach to false discovery control". *Annals of Statistics*, 32(3):1035-1061, 2004a.

10. C. R. Genovese and L. Wasserman. "Exceedance control of the false discovery proportion". Technical Report 807, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, July 2004b. Available at www.stat.cmu.edu/tr/tr807/tr807.html.

11. E. L. Korn, J. F. Troendle, L. M. McShane, and R. Simon. “Controlling the number of false discoveries; Application to high-dimensional genomic data”. *Journal of Statistical Planning and Inference*, 124(2):379-398, 2004.

12. M. J. van der Laan, S. Dudoit, and K. S. Pollard. “Multiple testing. Part II. Step-down procedures for control of the family-wise error rate”. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, 2004a. Available at www.bepress.com/sagmb/vol3/iss1/art14.

13. M. J. van der Laan, S. Dudoit, and K. S. Pollard. “Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives”. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004b.
Available at www.bepress.com/sagmb/vol3/iss1/art15.

14. M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. “Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives”. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 29, 2005.
Available at www.bepress.com/sagmb/vol4/iss1/art29.

15. E. L. Lehman and J. P. Romano. “Generalizations of the familywise error rate”. *Annals of Statistics*, 33(3):1138-1154, 2005.

16. J. P. Romano and M. Wolf. “Control of generalized error rates in multiple testing”. Technical Report 2005-12, Department of Statistics, Stanford University, Stanford, CA 94305, 2005.

17. C. E. Bonferroni. “Teoria Statistica Delle Classi e Calcolo Delle Probabilità”. *Pubblicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, pages 3-62, 1936.

18. R. G. Miller. *Simultaneus Statistical Inference*, 2nd ed. New York: Springer-Verlag, pages 67-70, 1981.

19. Z. Šidák. “Rectangular confidence regions for the means of multivariate normal distributions”. *Journal of the American Statistical Association*, 62:626-633, 1967.

20. O. J. Dunn. “Estimation of the means of dependent variables. *Annals of Mathematical Statistics*, 29:1095-111, 1958.

21. K. Jogdeo. “Association and probability inequalities”. *Annals of Statistics*, 5:495-504, 1977.

22. S. Holm. “A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65-70, 1979.

23. R. J. Simes. “An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751-754, 1986.

24. Z. Šidák. “On multivariate normal probabilities of rectangles: their dependence on correlations”. *Annals of Mathematical Statistics*, 39, 1425-34, 1968.

25. Z. Šidák. “On probabilities of rectangles in multivariate Student distributions: their dependence on correlations”. *Annals of Mathematical Statistics*, 42, 169-75, 1971.

26. S. K. Sarkar. “Generalizing Simes’ test and Hochberg’s stepup procedure”. Technical Report, Fox School of Business and Management, Temple University, Philadelphia, PA 19122, August 2005.

27. Y. Benjamini and D. Yekutieli. “The control of false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165-1188, 2001.

28. T. K. Sarkar. “Some lower bounds of reliability”. Technical Report, 124, Dept. Operation Research and Statistics, Stanford University, 1969.

29. E. Spjøtvoll. “On the Optimality of Some Multiple Comparison Procedures”. *The Annals of Mathematical Statistics*, 43(2), 398-411, 1972.

30. B. Soriç. “Statistical “discoveries” and effect-size estimation”. *Journal of the American Statistical Association*, 84(406):608-610, 1989.

31. S. Dudoit. M. J. van der Laan, and M. D. Birkner. “Multiple testing procedures for controlling tail probability error rates”. Technical Report 166, Division of Biostatistics, University of California, Berkeley, 2004a. URL www.bepress.com/ucbbiostat/paper166.

32. G. K. Smyth, J. Michaud, and H. Scott. “The use of within-array replicate spots for assessing differential expression in microarray experiments”. *Bioinformatics*, 21(9), 2067-2075, 2005.

33. Y. H. Yang and T. P. Speed. “Design and analysis of comparative microarray experiments”. In T. P. Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*, pages 35-91. Chapman & Hall/CRC Press, Boca Raton, 2003.

34. Katherine S. Pollard, Sandrine Dudoit, Mark J. van der Laan. "Multiple Testing Procedures". www.bioconductor.org. June 22, 2007. April 1st, 2013.

35. Q. Huang, R. T. Dunn, S. Jayadev, O. DiSorbo, F. D. Pack, S. B. Farr, R. E. Stoll, and K. T. Blanchard. "Assessment of cisplatin-induced nephrotoxicity by microarray technology". *Toxicological Sciences*, 63:196-207, 2001.