

STATISTICAL SIGNIFICANCE ASSESSMENT IN
COMPUTATIONAL SYSTEMS BIOLOGY

LI JUNTAO

NATIONAL UNIVERSITY OF SINGAPORE

2012

STATISTICAL SIGNIFICANCE ASSESSMENT IN
COMPUTATIONAL SYSTEMS BIOLOGY

LI JUNTAO

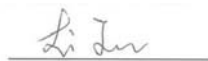
(Master of Science, Beijing Normal University, China)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
2012

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, appearing to read "Li Juntao", is written above a horizontal line.

LI JUNTAO

15 April 2012

Acknowledgements

I would like to thank my supervisor Prof. Choi Kwok Pui for his guidance on my study and his valuable advice on my research work. As a part-time PhD student, I have encountered many difficulties in balancing my job and study. At these moments, Prof. Choi always encouraged me to keep pursuing my goal and showed great patience in tolerating my delay in making progress.

I would also like to thank my supervisor in Genome Institute of Singapore, Dr. R Krishna Murthy Karuturi who is my mentor and my friend. During the past seven years in GIS, he consistently supported me and encouraged me. I would not have finished my PhD thesis without his advice and help.

Thanks go to my colleagues in the Genome Institute of Singapore, Paramita, Huaien, Ian, Max and Sigrid who work together with me and share many helpful ideas and discussions. I thank Dr. Liu Jianhua from GIS and Dr. Jeena Gupta from NIPER, India who provided their beautiful datasets for my analysis. Spe-

cially, I thank GIS and A*STAR for giving me the opportunity to pursue my PhD study.

Last but not least, I would like to give my most heartfelt thanks to my family: my parents, my wife and my baby. Their encouragement and support have been my source of strength and power.

Li Juntao

January 2012

Contents

Acknowledgements	ii
Summary	viii
List of Tables	ix
List of Figures	xi
1 Introduction	1
1.1 Overview of microarray data analysis and multiple testing	2
1.2 Error rates for multiple testing in microarray studies	4
1.3 p -value distribution and π_0 estimation	9
1.4 Significance analysis of microarrays	11

1.5	Problems and approaches	13
1.5.1	Constrained regression recalibration	17
1.5.2	Iterative piecewise linear regression	18
1.6	Organization of the thesis	19
2	ConReg-R: Constrained regression recalibration	20
2.1	Background	20
2.2	Methods	24
2.2.1	Uniformly distributed p -value generation	24
2.2.2	Constrained regression recalibration	26
2.3	Results	31
2.3.1	Dependence simulation	31
2.3.2	Combined p -values simulation	37
3	iPLR: Iterative piecewise linear regression	44
3.1	Background	44
3.2	Methods	48

<i>CONTENTS</i>	vi
3.2.1 Re-estimating the expected statistics	49
3.2.2 Iterative piecewise linear regression	53
3.2.3 iPLR for one-sided test	57
3.3 Results	58
3.3.1 Two-class simulations	58
3.3.2 Multi-class simulations	63
4 Applications of ConReg-R and iPLR in Systems Biology	67
4.1 Yeast environmental response data	67
4.2 Human RNA-seq data	72
4.3 Fission yeast data	74
4.4 Human Ewing tumor data	75
4.5 Integrating analysis in type2 diabetes	79
5 Conclusions and future works	86
5.1 Conclusions	86
5.2 Limitations and future works	89

<i>CONTENTS</i>	vii
5.2.1 Some special p -value distributions	89
5.2.2 Parametric recalibration method	91
5.2.3 Discrete p -values	91
5.2.4 π_0 estimation for ConReg-R and iPLR	93
5.2.5 Other regression functions for iPLR	93
Bibliography	93

Summary

In systems biology, high-throughput omics data, such as microarray and sequencing data, are generated to be analyzed. Multiple testing methods always are employed to interpret the omics data. In multiple testing problems, false discovery rates (FDR) are commonly used to assess statistical significance. Appropriate tests are usually chosen for the underlying data sets. However the statistical significance (p -values and error rates) may not be appropriately estimated due to the complex data structure of the microarray.

In this thesis, we proposed two methods to improve the false discovery rate estimation in computational systems biology. The first method, called *constrained regression recalibration* (ConReg-R), recalibrates the empirical p -values by modeling their distribution in order to improve the FDR estimates. Our ConReg-R method is based on the observation that accurately estimated p -values from true null hypotheses follow uniform distribution and the observed distribution of p -values is indeed a mixture of distributions of p -values from true null hypotheses

and true alternative hypotheses. Hence, ConReg-R recalibrates the observed p -values so that they exhibit the properties of an ideal empirical p -value distribution. The proportion of true null hypotheses (π_0) and FDR are estimated after the recalibration. ConReg-R provides an efficient way to improve the FDR estimates. It only requires the p -values from the tests and avoids permutation of the original test data. We demonstrate that the proposed method significantly improves FDR estimation on several gene expression datasets obtained from microarray and RNA-seq experiments.

The second method, called *iterative piecewise linear regression* (iPLR), in the context of SAM to re-estimate the expected statistics and FDR for both one-sided as well as two-sided statistics based tests. We demonstrate that iPLR can accurately assess the statistical significance in batch confounded microarray analysis. It can successfully reduce the effects of batch confounding in the FDR estimation and elicit the true significance of differential expression. We demonstrate the efficacy of iPLR on both simulated as well as several real microarray datasets. Moreover, iPLR provides a better interpretation of the linear model parameters.

List of Tables

1.1	Four possible hypothesis testing outcomes.	5
2.1	Combined p -values methods.	39
3.1	Illustration of batch confounding.	46
3.2	Parameters used to simulate the 4 different datasets A, B, C and D.	59
3.3	Significant gene tables for dataset ABCD.	61
3.4	Parameters used to simulate the 4 different datasets MA, MB, MC and MD.	63
3.5	Significant gene tables for Multi-class simulated dataset MA-MD. .	65
4.1	List of datasets used for ConReg-R and iPLR application.	68
4.2	Significant gene tables for yeast datasets.	75
4.3	Significant gene tables for human Ewing tumor datasets.	79

List of Figures

1.1	Four different p -value density plot examples.	15
1.2	Three different Q-Q plot examples.	17
2.1	Illustration of choosing k_{best} using k vs. $\hat{\pi}_0(k)$ plot.	30
2.2	Density histograms of dependent datasets and independent datasets.	33
2.3	Procedural steps for the independent and dependent datasets. . . .	34
2.4	Procedural steps for the independent and dependent datasets with random dependent effect.	36
2.5	Boxplots of FDR estimation errors.	37
2.6	Density histograms for “Min”, “Max”, “Sqroot”, “Square” and “Prod” datasets.	40
2.7	Procedure details for “Min”, “Max”, “Sqroot”, “Square” and “Prod” datasets at $\pi_0 = 0.7$	41
2.8	Procedure details for “Min”, “Max”, “Sqroot”, “Square” and “Prod” datasets at $\pi_0 = 0.9$	42
2.9	Boxplots of FDR estimation errors.	43
3.1	Examples for Q-Q plot slope approximation.	53
3.2	Work flow for iPLR.	54
3.3	Illustration of first two iterations in iPLR.	56
3.4	FDR comparison for simulation data sets A, B, C and D.	62

3.5	FDR comparison for simulation data sets MA, MB, MC and MD.	66
4.1	p -value density histograms for 10 stress response data sets.	69
4.2	Improvements in FDR estimation for yeast environmental response datasets.	71
4.3	p -value density histograms for meta-analysis (“Max”) before and after applying ConReg-R using yeast environmental response datasets.	71
4.4	p -value density histograms for RNA-seq and Affymetrix datasets.	73
4.5	Overlap between significantly differentially expressed genes identified by sequencing and microarray technologies.	73
4.6	SAM plot and FDR comparison (before and after iPLR re-estimation) for <i>S. pombe</i> data set.	76
4.7	Clustering of all arrays from Ewing <i>et al.</i> data using all the genes.	77
4.8	SAM plots and FDR comparison (before and after iPLR re-estimation) for human Ewing tumor data set.	78
4.9	SAM plots before and after iPLR re-estimation for 30min gene expression data for type2 diabetes and integrating cluster heat map for gene expression and histone marks.	82
4.10	RT-PCR validation on Histone H3 acetylation, lysine 4 mono methylation and lysine 9 mono methylation levels on coding regions of the chromatin modification regulating genes.	84
5.1	Hump shape and U-shape p -value density histograms.	90
5.2	p -value Density histograms from Wilcoxon test for various sample sizes.	92

Chapter 1

Introduction

In recent years, a number of novel biotechnologies have enabled biologists to readily monitor genome-wide expression levels. For instance, microarray technology is one of the most popular technologies. To analyze microarray data, many statistical methods are employed and multiple hypothesis testing procedure is one of the major approaches. In multiple hypothesis testing problem, p -values and false discovery rates (FDR) are commonly used to assess statistical significance. In this thesis, we develop two methods to assess the statistical significance in microarray studies. One method is extrapolative recalibration of the empirical distribution of p -value to improve FDR estimation. The second method is iterative piecewise linear regression to accurately assess the statistical significance in batch confounded microarray analysis.

1.1 Overview of microarray data analysis and multiple testing

A common question in microarray data analysis is the identification of differentially expressed genes, i.e., genes whose expression levels are associated with possibly censored biological and clinical covariates and outcomes. Most microarray studies include identifying disease genes (Diao *et al.*, 2004) or differentially expressed genes between wild type cell and mutant cell (Chu *et al.*, 2007a); finding differential patterns by time course microarray experiments (Chu *et al.*, 2007b; Li *et al.*, 2007). Moreover, microarray technology can be applied in comparative genomic hybridization (Pollack *et al.*, 1999), SNP (single nucleotide polymorphism) detection (Hacia *et al.*, 1999), Chromatin immunoprecipitation on Chip (Li *et al.*, 2009) and even DNA replication studies (Eshaghi *et al.*, 2007; Li *et al.*, 2008a).

The biological question in microarray data analysis can be restated as a multiple hypothesis testing problem: simultaneous testing for each gene or each probe in microarray, with the null hypothesis of no association between the expression measures and the covariates.

In microarray data analysis, parametric or non-parametric tests are employed. The two sample t-test and ANOVA (Baggerly *et al.*, 2001; Kerr *et al.*, 2004; Park *et al.*, 2003) are among the most widely used techniques in microarray studies. Although the usage of their basic form, possibly without justification of their main

assumptions, is not advisable (Jafari and Azuaje, 2006). Modifications to the standard t-test to deal with small sample size and inherent noise in gene expression datasets include a number of t-test like statistics and a number of Bayesian framework based statistics (Baldi and Long, 2001; Fox and Dimmic, 2006). In limma (linear model for microarray data), Smyth (2004) cleverly borrowed information from the ensemble of genes to make inference for individual gene based on the moderate t-statistic. Some other researchers also took advantages of shared information by examining data jointly. Efron et al. (2001) proposed a mixture model methodology implemented via an empirical Bayes approach. Similarly, Broet et al. (2002), Edwards et al. (2005), Do et al. (2005) used Bayesian mixture model to identify differentially expressed genes. Although Gaussian assumptions have dominated the field, other types of parametrical approaches can also be found in the literature, such as Gamma distribution models (Newton *et al.*, 2001).

Due to the uncertainty about the true underlying distribution of many gene expression scenarios, and the difficulties to validate distributional assumptions because of small sample sizes, non-parametric methods have been widely used as an attractive alternative to make less stringent distributional assumptions, such as the Wilcoxon rank-sum test (Troyanskaya *et al.*, 2002).

1.2 Error rates for multiple testing in microarray studies

Each time a statistical test is performed, one of four outcomes occurs, depending on whether the null hypothesis is true and whether the statistical procedure rejects the null hypothesis (Table 1.1): the procedure rejects a true null hypothesis (i.e. a false positive or type I error); the procedure fails to reject a true null hypothesis (i.e. a true negative); the procedure rejects a false null hypothesis (i.e. a true positive); or the procedure fails to reject a false null hypothesis (i.e. a false negative or type II error).

Therefore, there is some probability that the procedure will suggest an incorrect inference. When only one hypothesis is to be tested, the probability of each type of erroneous inference can be limited to tolerable levels by carefully planning the experiment and the statistical analysis. In this simple setting, the probability of a false positive can be limited by preselecting the p -value threshold for rejecting the null hypothesis. The probability of a false negative can be limited by performing an experiment with adequate replications. Statistical power calculations are performed to determine the number of replications required to achieve a desired level of control of the probability of a false negative result (pawitan *et al.*, 2005). When multiple tests are performed, as in the analysis of microarray data, it is even more critical to carefully plan the experiment and statistical analysis to reduce

Table 1.1: Four possible hypothesis testing outcomes.

Statistical inference	Fail to reject the null hypothesis	Reject the null hypothesis	Total
True null hypotheses	U (True negative)	V (False positive)	m_0
False null hypotheses	O (False negative)	S (True positive)	m_1
Total	W	R	m

the occurrence of erroneous inferences.

Every multiple testing procedure uses some error rate to measure the occurrence of incorrect inferences. Most error rates focus on the occurrence of false positives. Some error rates that have been used in the multiple testing are described next.

Classical multiple testing procedures use the family-wise error rate (FWER) control. The FWER is the probability of at least one Type I error,

$$\text{FWER} = \Pr(V > 0) = 1 - \Pr(V = 0), \quad (1.1)$$

where V is defined in Table 1.1.

The FWER was quickly recognized as being too conservative for the analysis of genome scale data, because in many applications, the probability that any of thousands of statistical tests yield a false positive inference is close to 1 and no result is deemed significant. A similar, but less stringent, error rate is the generalized family-wise error rate ($g\text{FWER}$). The $g\text{FWER}$ is the probability that more than k of the significant findings are actually false positives.

$$g\text{FWER}(k) = \Pr(V > k). \quad (1.2)$$

When $k = 0$, the g FWER reduces to the usual family-wise error rate, FWER. Recently, some procedures have been proposed to use the g FWER to measure the occurrence of false positives (Dudoit *et al.*, 2004).

The false discovery rate (Benjamini and Hochberg, 1995) (FDR) control is now recognized as a very useful measure of the relative occurrence of false positives in omics studies (Storey and Tibshirani, 2003). The FDR is the expected value of the proportion of Type I errors among the rejected hypotheses,

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R} 1_{\{R>0\}}\right], \quad (1.3)$$

where V and R are defined in Table 1.1. If all null hypotheses are true, all R rejected hypotheses are false positives, hence $V/R = 1$ and $\text{FDR} = \text{FWER} = \Pr(V > 0)$. FDR-controlling procedures therefore also control the FWER in the weak sense. In general, because $V/R \leq 1$, the FDR is less than or equal to the FWER for any given multiple testing procedure.

If we are only interested in estimating an error rate when positive findings have occurred, then the positive false discovery rate (p FDR) (Storey, 2002) is appropriate. It is defined as the conditional expectation of the proportion of type I errors among the rejected hypotheses, given that at least one hypothesis is rejected

$$p\text{FDR} = \mathbb{E}\left[\frac{V}{R} | R > 0\right]. \quad (1.4)$$

This definition is intuitively pleasing and has a nice Bayesian interpretation. Suppose that identical hypothesis tests are performed with independent statistic

T and rejection region Γ . Also suppose that a null hypothesis is true with a priori probability π_0 . Then

$$p\text{FDR}(\Gamma) = \frac{\pi_0 \Pr(T \in \Gamma | H = 0)}{\Pr(T \in \Gamma)} = \Pr(H = 0 | T \in \Gamma) \quad (1.5)$$

where $\Pr(T \in \Gamma) = \pi_0 \Pr(T \in \Gamma | H = 0) + (1 - \pi_0) \Pr(T \in \Gamma | H = 1)$. Here H is an indicator variable where $H = 1$ if the alternative hypothesis is true and $H = 0$ if the null is true. We denote $\Pr(H = 0)$ by π_0 .

The conditional false discovery rate (Tsai *et al.*, 2003) ($c\text{FDR}$) is the FDR conditional on the observed number of rejections $R = r$, is defined as

$$c\text{FDR} = \mathbb{E}(V/R | R = r) = \mathbb{E}(V | R = r) / r \quad (1.6)$$

provided that $r > 0$, and $c\text{FDR} = 0$, for $r = 0$.

The $c\text{FDR}$ is a natural measure of proportion of false positives among the r most significant tests. Further, under Storey's mixture model (Storey, 2002), Tsai *et al.* (2003) have shown that

$$c\text{FDR}(\alpha) = p\text{FDR}(\alpha) = \pi_0 \alpha m / r. \quad (1.7)$$

A major criticism of FDR is that it is a cumulative measure for a set of r most significant tests. An r^{th} significance test may have an acceptable FDR only due to it being part of the r most significant tests. To address this anomaly, Efron *et al.* (2001) introduced the local false discovery rate ($l\text{FDR}$), a variant of Benjamini-Hochberg's FDR. It gives each tested null hypothesis its own false

discovery rate. While the FDR is defined for one rejection region, the l FDR is defined for a particular value of the test statistic. The definition of l FDR is:

$$l\text{FDR}(t) = Pr(H = 0 | T = t). \quad (1.8)$$

The local nature of the l FDR is an advantage for interpreting results from individual test statistic. Moreover, l FDR is the average of global FDR given $T \in \Gamma$ i.e.

$$\text{FDR}(\Gamma) = E(l\text{FDR}(T) | T \in \Gamma). \quad (1.9)$$

In recent years, many methods are developed to estimate l FDR. For example, constrained polynomial regression procedure (Dalmasso *et al.*, 2007), unified approach (Strimmer, 2008) or semi-parametric kernel-based approach (Guedj *et al.*, 2009).

Ploner *et al.* (2006) generalized the local FDR as a function of multiple statistics, which combining a common test statistics with its standard error information and proposed 2D- l FDR. If two different statistics Z_1 and Z_2 capture different aspects of the information contained in the data, the 2D- l FDR can be defined as

$$2\text{D-}l\text{FDR}(z_1, z_2) = \pi_0 \frac{f_0(z_1, z_2)}{f(z_1, z_2)}, \quad (1.10)$$

where $f(z)$ is the density function of the statistics z , and $f_0(z) = f(z | z \in H_0)$.

2D- l FDR is very useful to deal with small standard error problems.

The FDR, c FDR, p FDR, l FDR and 2D- l FDR are reasonable error rates because they can naturally be translated into the costs of attempting to validate

false positive results. In practice the first three concepts lead to similar values, and most statistical software will usually report only one of the three (Li *et al.*, 2012b).

1.3 p -value distribution and π_0 estimation

P -value is the smallest level of significance where the hypothesis is rejected with probability one (Lehmann and Romano, 2005) and the definition is following,

Definition 1. *Suppose X has distribution P_θ for some $\theta \in \Omega$, and the null hypothesis H_0 specifies $\theta \in \Omega_{H_0}$. Assume the rejection regions S_α are nested in the sense that*

$$S_\alpha \subset S_{\alpha'} \text{ if } \alpha < \alpha', \quad (1.11)$$

p -value is defined as follows:

$$p = p(X) = \inf\{\alpha : X \in S_\alpha\}. \quad (1.12)$$

A general property of p -values is given in the following lemma.

Lemma 1.1. *Suppose the p -value p follows the definition 1, and assume the rejection regions S_α satisfy (1.11).*

(i) *If*

$$\sup_{\theta \in \Omega_{H_0}} P_\theta\{X \in S_\alpha\} \leq \alpha \text{ for all } 0 < \alpha < 1, \quad (1.13)$$

then the distribution of p under $\theta \in \Omega_{H_0}$ satisfies

$$P_\theta\{p \leq u\} \leq u \text{ for all } 0 \leq u \leq 1. \quad (1.14)$$

(ii) If, for $\theta \in \Omega_{H_0}$,

$$P_\theta\{X \in S_\alpha\} = \alpha \text{ for all } 0 < \alpha < 1, \quad (1.15)$$

then

$$P_\theta\{p \leq u\} = u \text{ for all } 0 \leq u \leq 1; \quad (1.16)$$

i.e. p is uniformly distributed over $(0, 1)$.

Proof. (i) If $\theta \in \Omega_{H_0}$, then the event $\{p \leq u\}$ implies $\{X \in S_v\}$ for all $u < v$.

The result follows by letting $v \rightarrow u$.

(ii) Since the event $\{X \in S_u\}$ implies $\{p \leq u\}$, it follows that

$$P_\theta\{p \leq u\} \geq P_\theta\{X \in S_u\}.$$

Therefore, if (1.15) holds, then $P_\theta\{p \leq u\} \geq u$, and the result follows from (i). \square

From Lemma 1.1, p -values from multiple testing is assumed to follow a mixture model with two components, one component follows a uniform distribution on $[0, 1]$ under the null hypotheses (Casella and Berger, 2001), and other component under the true alternative hypotheses (Pounds and Morris, 2003). A density plot (or histogram) of p -values is a useful tool for determining when problems are

present in the analysis. This simple graphical assessment can indicate when crucial assumptions of the methods operating on p -values have been radically violated (Pounds, 2006).

Additionally, it can be helpful to add a horizontal reference line to the p -value density plot at the value of the estimated π_0 , null proportion. A line falling far below the height of the shortest bar suggests that the estimate of the null proportion may be downward biased. Conversely, a line high above the top of the shortest bar may suggest that the method is overly conservative. It is appropriate to add this line to the density plot to assess the reliability of the π_0 estimates (Storey, 2002).

Furthermore, adding the estimated density curves to the p -value histogram can aid in assessing model fit (Pounds and Cheng, 2004). Large discrepancies between the density of the fitted model and the histogram indicate a lack of fit. This diagnostic can identify when some methods produce unreliable results. This is a good graphic diagnostic for any of the smoothing based and model-based methods that operate on p -values.

1.4 Significance analysis of microarrays

SAM (Significance Analysis of Microarrays) is a statistical technique for finding significant genes in a set of microarray experiments. It was proposed by (Tusher

et al., 2001). SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. The p -value for each gene is computed by repeated permutations of the data and the estimation of π_0 (Storey, 2002) is given below:

$$\hat{\pi}_0 = \min\left(\frac{\#\{d_i \in (q_{25}, q_{75})\}}{0.5n}, 1\right) \quad (1.17)$$

where the d_i are the original score for gene i ($i = 0, 1, \dots, n$), and q_{25}, q_{75} are 25% and 75% points of the permuted scores.

q -value (Storey, 2002) and local FDR ($lFDR$) (Efron *et al.*, 2001) are used in SAM. q -value is the lowest FDR at which the gene is called significant. The q -value measures how significant the gene is, as score increases, the corresponding q -value decreases. $lFDR$ is the false discovery rate for genes with scores that fall in a window around the score for the given gene. This is in contrast to the usual (global) FDR, which is the false discovery rate for a list of genes, whose scores exceed a given threshold.

GSA (Gene Set Analysis) (Efron and Tibshirani, 2007), a variation on the Gene Set Enrichment Analysis technique of (Subramanian *et al.*, 2005), is a function in SAM. The idea is to make inferences not about individual genes, but pre-defined sets of genes. GSA mentions most gene set enrichment scores S appear significantly large compared to the permutation values S^* . To address this kind of permutation bias, GSA use “Restandardization” method to adjust the permutation values as

follow,

$$S^{**} = \mu + \frac{\sigma}{\sigma^*}(S^* - \mu^*) \quad (1.18)$$

where S^{**} is restandardized permutation value, (μ, σ) and (μ^*, σ^*) are the overall means and standard deviations for S and S^* .

This approach is very simple and effective when π_0 is extremely close to 1 such that the test statistic S will almost come from null hypothesis and follow the unique asymptotically normal distribution. In GSA, only few gene sets will significantly enrich out of thousands gene sets for most cases, therefore, the permutation bias can be easily removed in GSA.

1.5 Problems and approaches

In microarray data analysis, multiple hypothesis testing is employed to address certain biological problems (e.g., gene selection, binding site selection and selection of gene sets). Appropriate tests are usually chosen for the particular microarray data sets, however the statistical significance (p -values and error rates) may not be appropriately estimated due to the complicated data structure of the microarray.

There are many factors influencing statistical significance in microarray studies. Dependence in the data is one of the major factors. Usually microarray data have large number of genes (variables) but few samples, and there are many groups of genes having similar expression patterns. Each array also has global

effect which will influence the dependence of the data. FDR controlling procedure for independent test statistics may still control the false discovery rate, however it requires that the test statistics have positive regression dependency on each of the test statistics corresponding to the true null hypotheses (Benjamini and Yekutieli, 2001). For example, batch and cluster effects often occur in the experiments and sometimes it may mainly affect the significance i.e. underestimate or overestimate the statistical significance. Besides these major factors, approximate p -value estimation, violation of test assumptions, over or under estimation of some parameters and other unaccounted variations may also influence the FDR estimation.

Batch effects (Lander *et al.*, 1999) are commonly observed across multiple batches of microarray experiments. There are many different kinds of effects, RNA batch effect (experimenter, time of day, temperature), array effect (scanning level, pre/postwashing), location effect (chip, coverslip, washing), dye effect (dye, unequal mixing of mixtures, labeling, intensity), print pin effect, spot effect (amount of DNA in the spot printed on slide) (Wit and McClure, 2003) and even the atmospheric ozone level (Fare *et al.*, 2003). Local batch effects (such as location, print pin, dye effect and spot effect) may be removed by using one of the many local normalization methods available in the literature (Smyth and Speed, 2003). However global batch effects are too complicated. It is difficult to detect and not easy to eliminate across all circumstances.

If the test statistics from multiple testing can be well modeled using certain

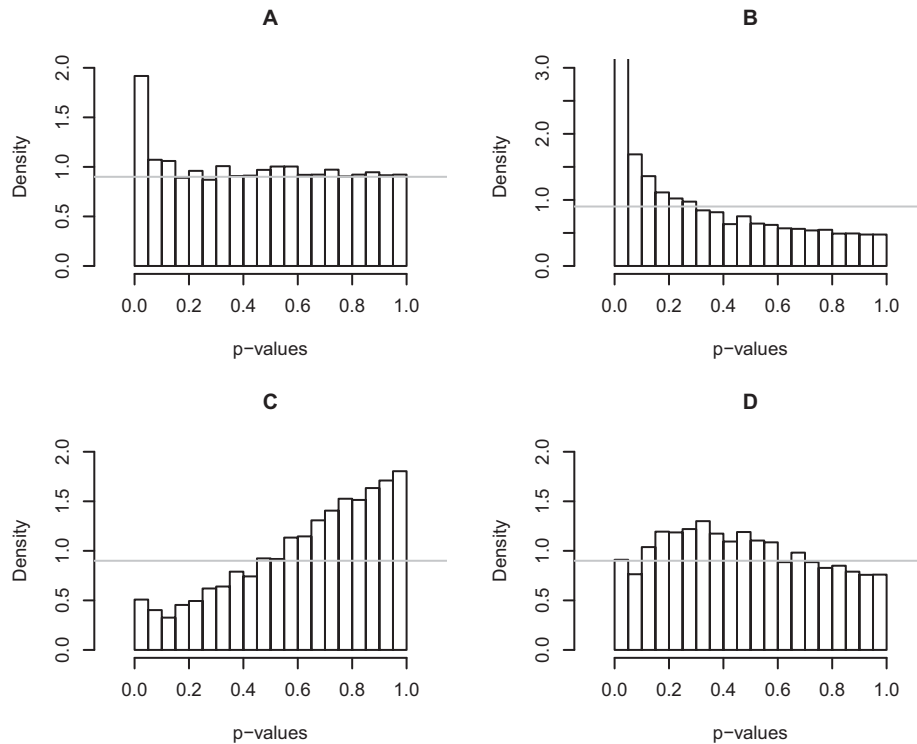


Figure 1.1: Four different p -value density plot examples.

distribution and p -values are appropriately computed, the p -value distribution can be used to validate whether the statistical significance is appropriately estimated or not. In Figure 1.1, there are four different p -value density plot examples. The most desirable shape of the p -value density plot is the one in which the p -values are most dense near zero, become less dense as the p -values increase, and have near-uniform tail towards 1 (Figure 1.1A). This shape does not indicate violation of the assumptions of methods operating on p -values and suggests that several features are differentially expressed, though they may not be statistically significant after adjusting for multiple testing. A very sharp p -value density plot without near-uniform tail close to 1 (Figure 1.1B) and $g(1) < 0.5$ may indicate over-assessment

of significance i.e. under-measured p -values where $g(.)$ is the density function of p -value. It suggests that fewer features are significant than observed. A right triangle p -value density plot with $g(0) < g(1)$ and $g(1) > 1$ (Figure 1.1C) may also indicate over-measure p -values, suggesting that more features are differentially expressed than observed. A p -value density plot with one or more humps in the middle (Figure 1.1D) can indicate that an inappropriate statistical test was used to compute the p -values, some heterogeneity data were included in the analysis, or a strong and extensive correlation structure is present in the data set (Pounds, 2006).

Sometimes the tests can be modified to increase the stability of the testing power (for example, modified t-test) and the test statistics may not follow any well-defined distribution. Re-sampling method is usually used to measure the statistical significance. Re-sampling p -values mostly are not highly precise and its distribution is difficult to model. We can use Q-Q plot between observed test statistics and expected test statistics to validate whether the statistical significance is appropriately estimated. In Figure 1.2A, the expected score (expected test statistics) and observed score (test statistics) are aligned with the diagonal. This indicates the statistical significance is appropriately estimated. If the expected test statistics deviate much from observed test statistics (Figure 1.2B and 1.2C), the statistical significance will be over/under-estimated.

Therefore, we develop two methods which focus on p -values and re-sampling

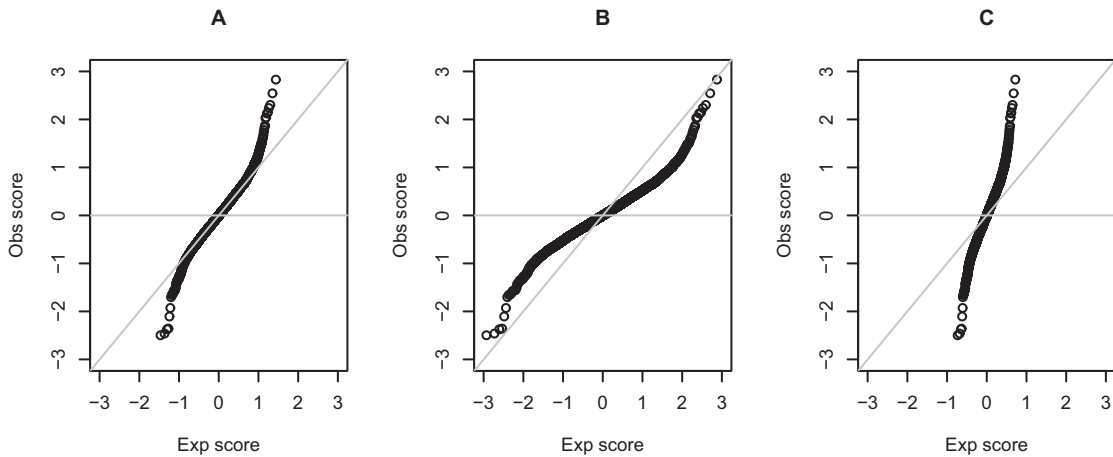


Figure 1.2: Three different Q-Q plot examples.

statistics respectively to assess the statistical significance in microarray studies. One method is extrapolative recalibration of the empirical distribution of p -value to improve FDR estimation (Li *et al.*, 2011). The second method is iterative piecewise linear regression to accurately assess the statistical significance in batch confounded microarray analysis (Li *et al.*, 2012a).

1.5.1 Constrained regression recalibration

In multiple hypothesis testing problems, the most appropriate error control may be false discovery rate (FDR) control. The precise FDR depends on the accurate p -values from each test and validity of independent assumption. However, in many practical testing problems such as in genomics, the p -values could be under-measured or over-measured for many known or unknown reasons. Consequently, FDR estimation would then be influenced and lose its veracity.

We propose a regression method to model the empirical distribution of p -values and transform the conservative or optimistic p -values to well-defined p -values to improve the FDR estimation. Our approach first generates the theoretical p -values following uniform distribution, and then performs the constrained polynomial regression between the p -values supposedly to have come from the null hypotheses and the theoretical p -values. The constrained polynomial regression can be posed as a quadratic programming problem. Finally, the overall p -values will be transformed using the normalized regression function and output the adjusted p -values. FDR is estimated using the adjusted p -values and the π_0 can be determined during this procedure. We have demonstrated that our procedure can well estimate the FDR by adjusted p -values from both dependency data and meta-analyzed data.

1.5.2 Iterative piecewise linear regression

Batch dependent variation in microarray experiments may be manifested through systematic shift in expression measurements from batch to batch. Such a systematic shift could be taken care of by using an appropriate model for differential expression analysis. However, it poses greater challenge in the estimation of statistical significance and false discovery rate (FDR), if the batches are confounded (collinear) with the biological groups of interest. Batch confounding problem occurs commonly in the analysis of time-course data or data from different laboratories.

We demonstrate that batch confounding may lead to incorrect estimation of the expected statistics. We propose an *iterative piecewise linear regression* (iPLR) method, a major extension of our previously published *Stepped Linear Regression* (SLR) method, in the context of SAM to re-estimate the expected statistics and FDR. iPLR can be applied to one-sided or two-sided statistics based tests. We demonstrate the efficacy of iPLR on both simulated and real microarray datasets. iPLR also provides a better interpretation of the linear model parameters.

1.6 Organization of the thesis

This thesis consists of 5 chapters. The next chapter, Chapter 2, is focused on the details of ConReg-R method to model and recalibrate the p -value distribution. In Chapter 3, we propose iterative piecewise linear regression (iPLR) method to address batch confounding problem. In Chapter 4, we study the application of our methods in few real microarray data studies such as yeast datasets, human tumor datasets, human RNA-seq datasets and ChIP-chip studies. Finally, in Chapter 5, we summarize the achievements in the thesis work, discuss the limitations of the methods, and propose a few potential directions for future work.

Chapter 2

ConReg-R: Constrained regression recalibration

This chapter describes the ConReg-R procedure to recalibrate p -values for accurate assessment of FDR and simulation results.

2.1 Background

In high-throughput biological data analysis, multiple hypothesis testing is employed to address certain biological problems. Appropriate tests are chosen for the data, and the p -values are then computed under some distributional assumptions. Due to the large number of tests performed, error rate controls (which focus on the occurrence of false positives) are commonly used to measure the statistical

significance. False discovery rate (FDR) control is accepted as the most appropriate error control. Other useful error rate controls include conditional FDR (cFDR) (Tsai *et al.*, 2003), positive FDR (pFDR) (Storey, 2002) and local FDR (lFDR) (Efron *et al.*, 2001) which have similar interpretations as that of FDR. However, appropriate FDR estimation depends on the precise p -values from each test and the validity of the underlying assumptions of the distribution.

The p -values from multiple hypothesis testing, for n hypotheses, can be described by a mixture model $g(p)$ (2.1) with two components: one component $g_0(p)$ originates from true null hypotheses and follows uniform distribution $U(0, 1)$, and the other component $g_1(p)$ results from true alternative hypotheses and follows a distribution confined to the p -values close to 0 (Lehmann and Romano, 2005; Pounds and Morris, 2003). The mixing parameter, π_0 , is the proportion of true null hypotheses in the data.

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_1(p) \quad (2.1)$$

where $g_0(p) = 1$ is a uniform distribution over $(0, 1)$ and $g_1(p)$ will be approximately 0 for p close to 1 which is expected to be true in most practical situations. Therefore, $g(p)$ will be close to a constant (i.e. π_0) for p close to 1.

FDR in multiple hypothesis testing for a given p -value threshold α is estimated as

$$\widehat{\text{FDR}}_\alpha = \frac{\hat{\pi}_0 \alpha n}{\#\{p < \alpha\}}.$$

π_0 can be estimated from the mixture model in equation (2.1) as (Storey, 2002)

$$\hat{\pi}_0 = \frac{\#\{p > \beta\}}{[(1 - \beta)n]}$$

where β is typically chosen to be 0.25, 0.5 or 0.75. These estimates are reasonable under the uniform distribution assumption of $g_0(p)$ component in this mixture model (Pawitan *et al.*, 2005).

However, in many applied testing problems, the p -values could be under-measured or over-measured for many known or unknown reasons. The violation of p -value distribution assumptions may lead to inaccurate FDR estimation. There are many factors influencing FDR estimation in the analysis of high-throughput biological data such as microarray and sequencing studies. Dependence among the test statistics is one of the major factors (Efron, 2007; Qiu *et al.*, 2005). Usually in microarray data, there are many groups of genes having similar expression patterns and the test statistics (for example, t-statistic) are not independent within one group. The global effects in the array may also influence the dependence in the data. For example, batch and cluster effects (Johnson *et al.*, 2007; Li *et al.*, 2008b) always occur in the experiments and sometimes they may be the major cause of incorrectly estimated FDR.

Further, due to the “large p , small n ” problem (Ochs *et al.*, 2001) for the gene expression data, some parameters such as mean and variance for each gene cannot be well estimated, or the test assumptions are not satisfied or the distribution of the statistic under null hypotheses may not be accurate. Therefore, many applied

testing methods modified the standard testing methods (for example, modifying t-statistic to moderated t-statistic (Smyth, 2004) to increase their usability. As the modified test statistics only approximately follow some known distribution, the approximate p -value estimation may influence the FDR estimation. Resampling strategies may better estimate the underlying distributions of the test statistics. However, due to small sample size and data correlation, the limited number of permutations and resampling bias (Efron and Tibshirani, 2007) also influence the FDR estimation.

To address the above problems, we propose a novel extrapolative recalibration procedure called *Constrained Regression Recalibration* (ConReg-R) which models the empirical distribution of p -values in multiple hypothesis testing and recalibrates the imprecise p -value calculation to better recalibrated p -values to improve the FDR estimation. Our approach focuses on p -values as the p -values from true null hypotheses are expected to follow the uniform distribution and the interference from the distribution of p -values from alternative hypotheses is expected to be minimal towards $p=1$. In contrast, the estimation of the empirical null distributions of test statistics may not be accurate as their parametric form may not be known beforehand and their accuracy may depend on the data and the resampling strategy used. ConReg-R first maps the observed p -values to predefined uniformly distributed p -values preserving their rank order and estimates the recalibration mapping function by performing constrained polynomial regression to the k highest p -values. The constrained polynomial regression is implemented

by quadratic programming solvers. Finally, the p -values will be recalibrated using the normalized recalibration function. FDR is estimated using the recalibrated p -values and the $\hat{\pi}_0$ can be determined during ConReg-R procedure. We demonstrate that our ConReg-R procedure can significantly improve the estimation of FDR on simulated data, and also the environmental stress response time course microarray datasets in yeast and a human RNA-seq dataset.

2.2 Methods

Under the null hypotheses, the p -values are uniformly distributed. Hence, ConReg-R first generates the uniformly distributed p -values within $[0, 1]$ range.

2.2.1 Uniformly distributed p -value generation

Let p_i denotes the p -value of the i^{th} test ($i = 1, \dots, n$), without loss of generality, we assume $p_1 \geq p_2 \geq \dots \geq p_n$. If we choose a suitable $k < n$ such that the i^{th} null hypothesis $H_0^{(i)}(i \leq k)$ is most likely true, then p_1, \dots, p_k correspond to the order statistics of k independent uniformly distributed random variables provided p_i 's $i(i = 1, \dots, k)$ are correctly estimated.

Let p'_i are conditional expectations of the corresponding order statistics of p -values under $H_0^{(i)}(i \leq k)$, and suppose p'_k is known. $p'_i(i \leq k)$ can be defined

as

$$p'_i = 1 - \frac{i-1}{k-1}(1-p'_k), i = 1, \dots, k. \quad (2.2)$$

Then

$$\hat{\pi}_0 = \frac{k}{n(1-p'_k)}. \quad (2.3)$$

Using (2.3), (2.2) becomes

$$p'_i = 1 - \frac{i-1}{k-1} \cdot \frac{k}{n\hat{\pi}_0}, i = 1, \dots, k. \quad (2.4)$$

Since k is usually large, $k/(k-1)$ is almost 1, therefore p'_i in (2.4) can be approximated as

$$p'_i \doteq 1 - \frac{i-1}{n\hat{\pi}_0}, i = 1, \dots, k. \quad (2.5)$$

We can estimate the recalibration function $f(\cdot)$, to be described below, between $\{p'_i\}_{i=1}^k$ and $\{p_i\}_{i=1}^k$ and apply it to all input p -values to output the recalibrated p -values, p_i^{cal} ($i = 1, \dots, n$) i.e.

$$p_i^{cal} = f(p_i), i = 1, \dots, n. \quad (2.6)$$

By Stone-Weierstrass theorem (Bishop, 1961), polynomial functions can well approximate any continuous function in the interval $[0, 1]$. Therefore we use polynomial regression to estimate the recalibration function $f(\cdot)$ satisfying appropriate boundary and monotone constraints.

2.2.2 Constrained regression recalibration

Let $y_i = p'_i$ and $x_i = p_i$ ($i = 1 \dots k$), and the recalibration polynomial function $f(\cdot)$ is defined as follows,

$$y_i = f(x_i) = \sum_{j=0}^t \beta_j x_i^j + \varepsilon_i. \quad (2.7)$$

The constraints $f(0) = 0$, $f(1) = 1$ and $f'(x) > 0$ should be imposed to ensure the orders of the p -values remain the same after the transformation. Furthermore, the constraint for either $f''(x) > 0$ or $f''(x) < 0$ indicates the function f should also be a monotonic convex or monotonic concave function to deal with the situations with under-measured or over-measured p -values separately and helps in good extrapolation.

The constraints $f(0) = 0$ and $f(1) = 1$ can be easily met by scaling and shifting the regression function. Therefore, the regression function only depends on the other two constraints which can be combined into one constraint during the regression procedure.

Quadratic programming (QP) (Nocedal and Wright, 2000) is employed to estimate the regression function as follows: Let $\mathbf{y} = (y_1, \dots, y_k)^T$, $\beta = (\beta_0, \dots, \beta_t)^T$ and

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^t \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_k & x_k^2 & \dots & x_k^t \end{pmatrix}.$$

Equation (2.7) can be rewritten more succinctly as

$$\mathbf{y} = f(X) = X\beta + \epsilon \quad (2.8)$$

and the constraints for the first and second order derivatives of $f(X)$ will be $A\beta \geq \mathbf{b}$

where $\mathbf{b} = (0, \dots, 0)^T$ is a $2l \times 1$ vector and

$$A = \begin{pmatrix} 0 & 1 & 2a_1 & \dots & ta_1^{t-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 2a_l & \dots & ta_l^{t-1} \\ 0 & 0 & (-1)^c 2 & \dots & (-1)^c t(t-1)a_1^{t-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & (-1)^c 2 & \dots & (-1)^c t(t-1)a_l^{t-2} \end{pmatrix}$$

is a $2l \times (t+1)$ matrix, where a_1, \dots, a_l are l randomly generated numbers following $U(0, 1)$ to guarantee this constraint is valid in $(0, 1)$, and c is chosen to be 0 (or 1) if f is desired to be convex (or concave respectively).

The least squares procedure for (2.8) will minimize

$$\begin{aligned} \|\mathbf{y} - X\beta\|_2^2 &= (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta) \\ &= (\mathbf{y}^T - \beta^T X^T) (\mathbf{y} - X\beta) = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T X\beta - \beta^T X^T \mathbf{y} + \beta^T X^T X\beta \\ &= \beta^T X^T X\beta - 2\mathbf{y}^T X\beta + \mathbf{y}^T \mathbf{y}. \end{aligned} \quad (2.9)$$

Minimizing (2.9) under $A\beta \geq \mathbf{b}$ is equivalent to minimizing

$$h(\beta) = \frac{1}{2} \beta^T Q \beta + \mathbf{q}^T \beta \quad (2.10)$$

under $A\beta \geq \mathbf{b}$, where $Q = X^T X$ and $\mathbf{q} = -X^T \mathbf{y}$. Therefore, the constrained polynomial regression problem can be reformulated as a quadratic programming problem. Dalmasso *et al.* (2007) have used similar ideas to estimate IFDR. However, they used entire data for fitting and meant to estimate the densities of g_0 and g_1 which can be used to estimate IFDR. In contrast, ConReg-R is an extrapolative procedure to generate well calibrated p-values which can be used for multitude of purposes e.g. meta-analysis, FDR computation, IFDR computation, effect size estimation, etc.

Two further modifications

We use QuadProg package in *R* to solve the quadratic programming problem (Goldfarb and Idnani, 1983). Due to floating point errors (Press *et al.*, 2007), $Q = X^T X$ tends to be positive semidefinite instead of being positive definite. To get around this, we add a sufficiently small positive value ($\lambda = 10^{-10}$) to the diagonal of Q to guarantee $Q' = Q + \lambda \mathbf{I}_{t+1}$ is positive definite and Q' replaces Q in (2.10).

Furthermore, the polynomial function may not accurately fit the data due to the limitation of the polynomial maximal power (usually set the maximal power $t = 10$). We can add the fraction of the power (i.e. a non-integer power) to increase the accuracy of the fit. For example, let $f(x_i) = \sum_{j=0}^{mt} \beta_j x_i^{j/m} + \varepsilon_i$, where $m = 1, 2$ or more.

Computational procedure

For any given k , after applying ConReg-R, the estimation of $\hat{\pi}_0$ and its variation (error) are given by

$$\hat{\pi}_0(k) = \text{median}\left(\left\{\frac{i}{n(1-p_i^{cal})}\right\}_{i=1}^k\right) \quad (2.11)$$

and

$$e_{\hat{\pi}_0(k)} = \text{MAD}\left(\left\{\frac{i}{n(1-p_i^{cal})}\right\}_{i=1}^k\right) \quad (2.12)$$

where MAD denotes the median absolute deviation. The final regression function and optimal $k(k_{best})$ are determined by examining $\hat{\pi}_0(k)$ and $e_{\hat{\pi}_0(k)}$ over k . Figure 2.1 illustrates how to choose k_{best} from the function $\hat{\pi}_0(k)$. Ideally, $\hat{\pi}_0(k)$ is not expected to change over a range of k (as shown by the blue dashed line in Figure 2.1) such that p_1, \dots, p_k are most likely to be from null hypotheses. If k is too large, p_1, \dots, p_k may contain too many p -values from alternate hypotheses and $\hat{\pi}_0(k)$ may be wrongly estimated to be close to 1, in an extreme case if k is chosen to be n then $\hat{\pi}_0(k) = 1$. However, the extrapolation in recalibration procedure may be unreliable if only a small number of p -values (i.e. small k) are used for the regression and $\hat{\pi}_0(k)$ may fluctuate near the real π_0 (the red curve in Figure 2.1). Therefore, we aim to choose optimal $k(k_{best})$ as a trade-off to include just enough p -values from null hypotheses for the regression to achieve good extrapolation. The k that gives stable estimate ($e_{\hat{\pi}_0(k)} < \delta$) and the last minimum of $\hat{\pi}_0(k)$ is chosen to be the k_{best} . The regression function, extrapolation and $\hat{\pi}_0(k)$ corresponding to $k = k_{best}$ are chosen for recalibrating p -values and re-estimating FDR.

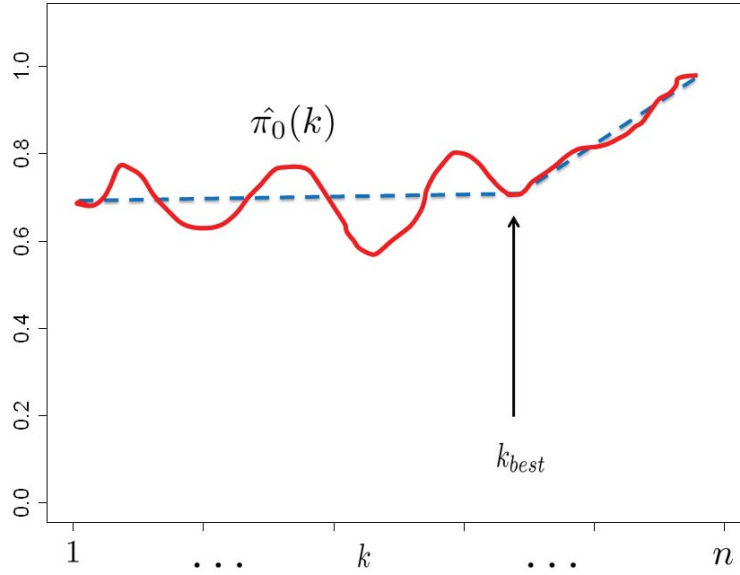


Figure 2.1: Illustration of choosing k_{best} using k vs. $\hat{\pi}_0(k)$ plot. The blue dashed line indicates the ideal π_0 estimated for different choice of k . The red curve indicates the actual $\hat{\pi}_0(k)$.

The following is the computational procedure for a given $\{p_i\}_{i=1}^n$ in descending order:

1. For each $k = v, 2v, \dots, \lceil \frac{n}{v} \rceil v$ (v is the interval over k and default setting is $v = \lceil n/100 \rceil$), let $\hat{\pi}_0 = 1$.
2. Use equation (2.5) to compute $\{p'_i\}_{i=1}^k$.
3. Use quadratic programming to obtain regression function h_k , where c can be predefined or estimated by checking whether more than half of points for (p_i, p'_i) are above the diagonal (line from origin to $(1, 1)$) ($c = 1$) or below the diagonal ($c = 0$).

4. Transform h_k to $f_k(\cdot) = \frac{h_k(\cdot) - h_k(0)}{h_k(1) - h_k(0)}$ to satisfy constraints $f_k(0) = 0$, and $f_k(1) = 1$.
5. Repeat steps 2-4 for all k , and compute the $\hat{\pi}_0(k)$ and $e_{\hat{\pi}_0(k)}$ for each k . Let k_{best} be the maximal of k which locally minimizes $\hat{\pi}_0$ under the constraint of small $e_{\hat{\pi}_0}$, where the cutoff of $e_{\hat{\pi}_0}$ and local minimization criteria should be predefined.
6. Choose the final regression function $f(\cdot)$ under k_{best} and output recalibrated p -values.
7. Re-estimate the FDR using recalibrated p -values and $\hat{\pi}_0 = \hat{\pi}_0(k_{best})$.

R-code for ConReg-R is attached as Appendix A.

2.3 Results

2.3.1 Dependence simulation

Data dependence is one of the major causes for under-measured or over-measured p -values. We simulated an expression data, with dependence, $Z = (z_{ij})(i = 1, \dots, n, j = 1, \dots, r)$ with $n(n = 10000)$ genes and $r(r = 10)$ replicates using the formula as follows,

$$z_{ij} = b_i + d_{ij} + \varepsilon_{ij}$$

where b_i denotes the biological effect, d_{ij} denotes the dependence effect. Set $b_i = 1$, if $i \leq n(1 - \pi_0)$ and $b_i = 0$ if $i > n(1 - \pi_0)$. $d_{i.} = (1, 1, 1, 0, 0, 0, 0, -1, -1, -1)$ if $i \leq \lfloor \frac{n}{2} \rfloor$ and $d_{i.} = (-1, -1, -1, 0, 0, 0, 0, 1, 1, 1)$ if $i > \lfloor \frac{n}{2} \rfloor$. $\varepsilon_{ij} \sim N(0, 1)$ is the background noise.

To compare the result, we also simulated a data set with no dependence using the same procedure but with the dependence effect $d_{ij} = 0$. One sample t-test was performed to generate p -values. Figure 2.2 shows the p -value density histograms for $\pi_0 = 0.7$ and $\pi_0 = 0.9$. As can be seen in the plots B and D in Figure 2.2, the p -value histograms from independent data have constant frequency for $p \geq 0.5$ and the density near 1 indicates the $\hat{\pi}_0$. However, the p -value histograms from dependent data (the plots A and C in Figure 2.2) do not have such constant frequency and p -value density increases as p -value increases in the neighborhood of 1. The density near 1 exceeds the respective π_0 .

ConReg-R used the above p -values as input and output the recalibrated p -values. The results are shown in Figure 2.3. For the independent data sets, the algorithm chose $k = 0.71n$ for $\pi_0 = 0.7$ and $k = 0.64n$ for $\pi_0 = 0.9$ since it locally minimized $\hat{\pi}_0$ under $error(\hat{\pi}_0) < 0.05$. The p -values do not significantly change after regression. As such, the regression curves almost overlap with the diagonals, and the input p -value histogram and the output p -value histogram are very similar to each other. The FDR estimation errors (the absolute difference between FDR estimated by p -values and real FDR) also do not significantly change

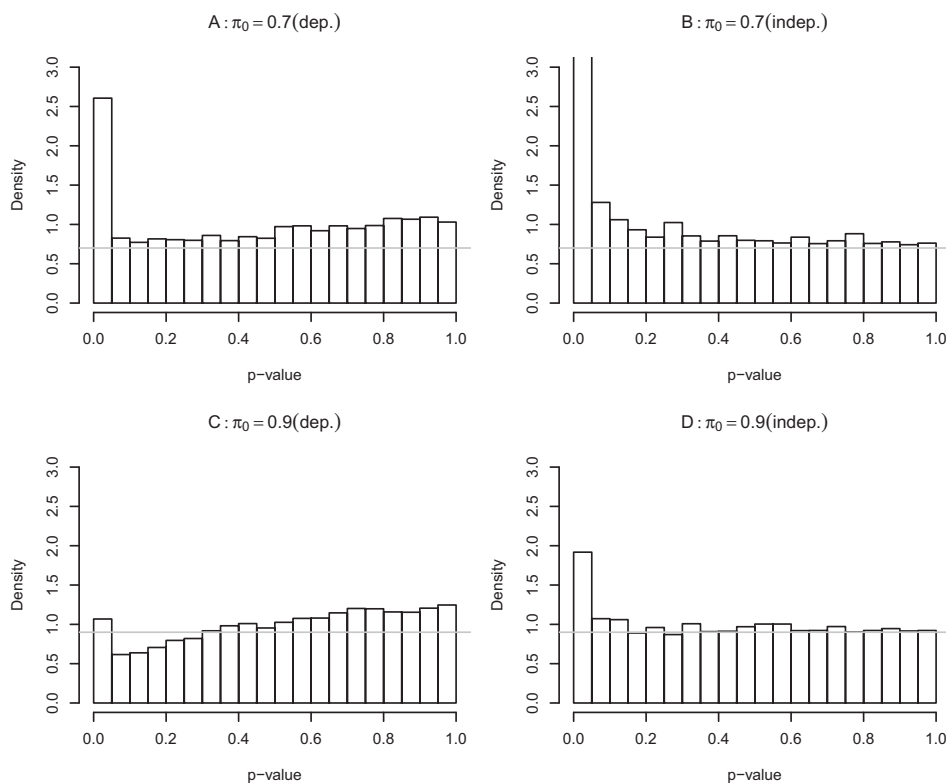


Figure 2.2: Density histograms of dependent datasets and independent datasets at $\pi_0 = 0.7$ and $\pi_0 = 0.9$, and the gray horizontal line indicates the π_0 for each dataset.

after applying ConReg-R and the estimation of FDR is very close to the real FDR. However, for the dependent data sets, the algorithm chose $k = 0.62n$ for $\pi_0 = 0.7$ and $k = 0.88n$ for $\pi_0 = 0.9$. The regression curves are all below the diagonals and the output p -value histograms after applying ConReg-R appears more like the ones obtained for the independent data. The accuracy of estimated FDR after applying ConReg-R is substantially improved.

To study more complicated dependency situations, we generated dependent

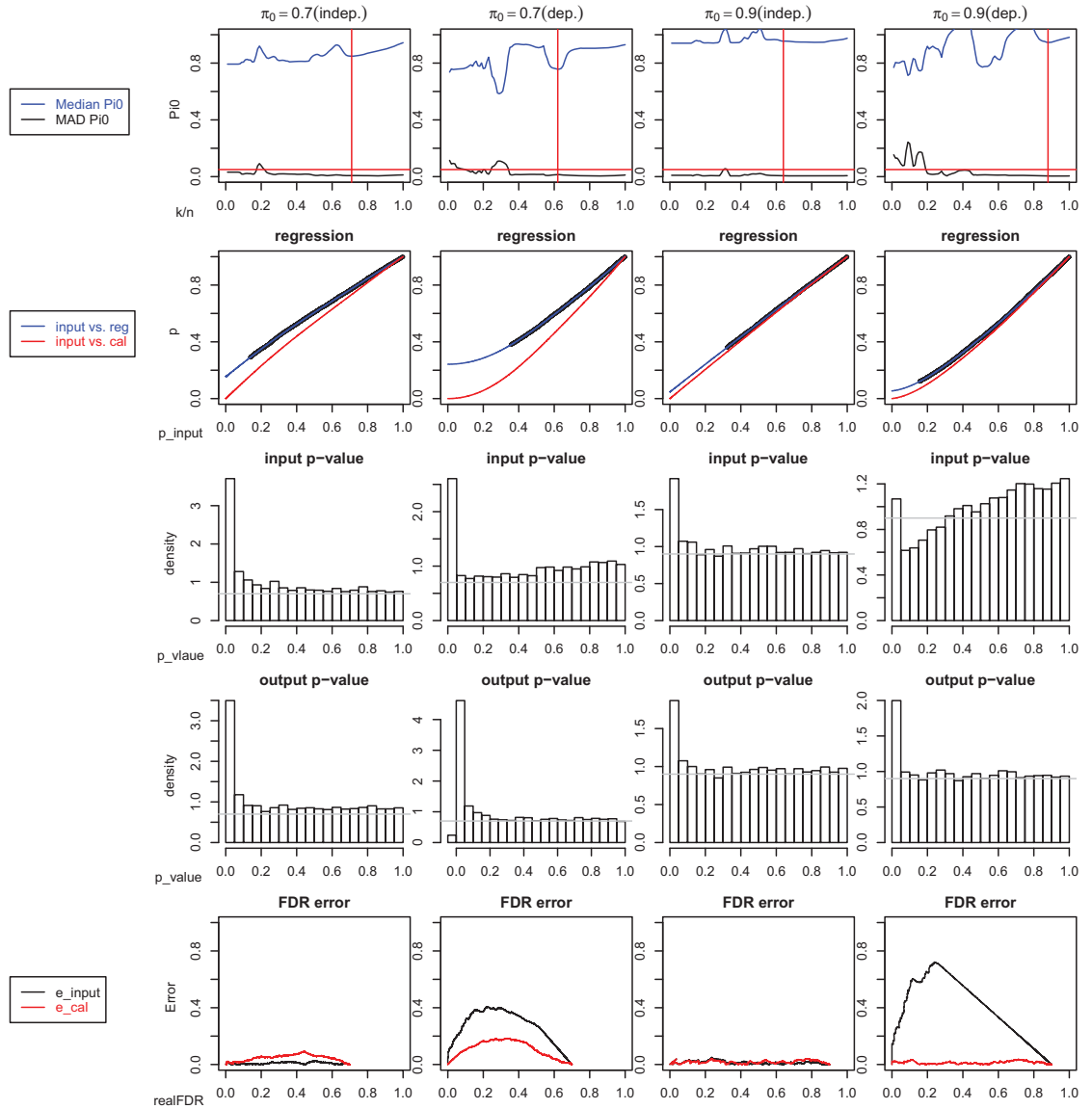


Figure 2.3: Procedural steps for the independent and dependent datasets at $\pi_0 = 0.7$ and $\pi_0 = 0.9$. The plots in first row show the $\hat{\pi}_0$ and $e_{\hat{\pi}_0}$ at different k/n . The blue curve indicates $\hat{\pi}_0$ and the black curve indicates $e_{\hat{\pi}_0}$, the red horizontal line indicates the cutoff of $e_{\hat{\pi}_0}$ (here we used 0.05), the red vertical line indicates the choice of k/n at which locally minimized $\hat{\pi}_0$ under $e_{\hat{\pi}_0} < 0.05$ is obtained. The plots in second row show the regression procedure. The black thick curve indicates the $(p_i, p'_i), i = 1, \dots, k$ and the blue curve is the regression line $h_k(\cdot)$, and the red curve is the regression line $f(\cdot)$ after transformation. The plots in third and fourth row show the p -value histograms before and after applying ConReg-R and the gray horizontal line indicates the π_0 . The plots in last row show the FDR estimation errors between real FDR and the FDR estimated by p -values before (black) and after applying ConReg-R (red).

datasets with random dependence effect (Qiu *et al.*, 2005) as follows,

$$z_{ij} = \rho(b_i + d_j) + (1 - \rho)\varepsilon_{ij}$$

where ρ is the correlation constant (here we set $\rho = 0.5$) which determines the correlation coefficient between genes. Here b_i denotes the biological effect, and d_j denotes the random dependence effect. Set $b_i = 1$, if $i \leq n(1 - \pi_0)$ and $b_i = 0$ if $i > n(1 - \pi_0)$, and $d_j \sim N(0, 1)$. Let $\varepsilon_{ij} \sim N(0, 1)$ be the background noise. The result for $\pi_0 = 0.7$ and $\pi_0 = 0.9$ are shown in Figure 2.4. Similar to the simulations of fixed dependence effect, the estimated FDR after applying ConReg-R is closer to real FDR.

The results of our procedure for 100 repeated simulations are summarized in the box-and-whisker plots in Figure 2.5. As shown in this figure, for the independent data sets, the FDR estimation errors (the mean absolute difference between real FDR and the FDR estimated by p -values using Benjamini-Hochberg method) after applying ConReg-R is slightly higher. However, it is still acceptable since most simulations resulted in errors below 0.05. For the dependent data sets with fixed and random dependence effects, the FDR estimation errors after applying ConReg-R are significantly less than those without applying ConReg-R. The FDR estimation for $\pi_0 = 0.9$ is even closer to real FDR after applying ConReg-R compared with the result for $\pi_0 = 0.7$ because of more p -values used for regression and less number of p -values for extrapolating in datasets of $\pi_0 = 0.9$.

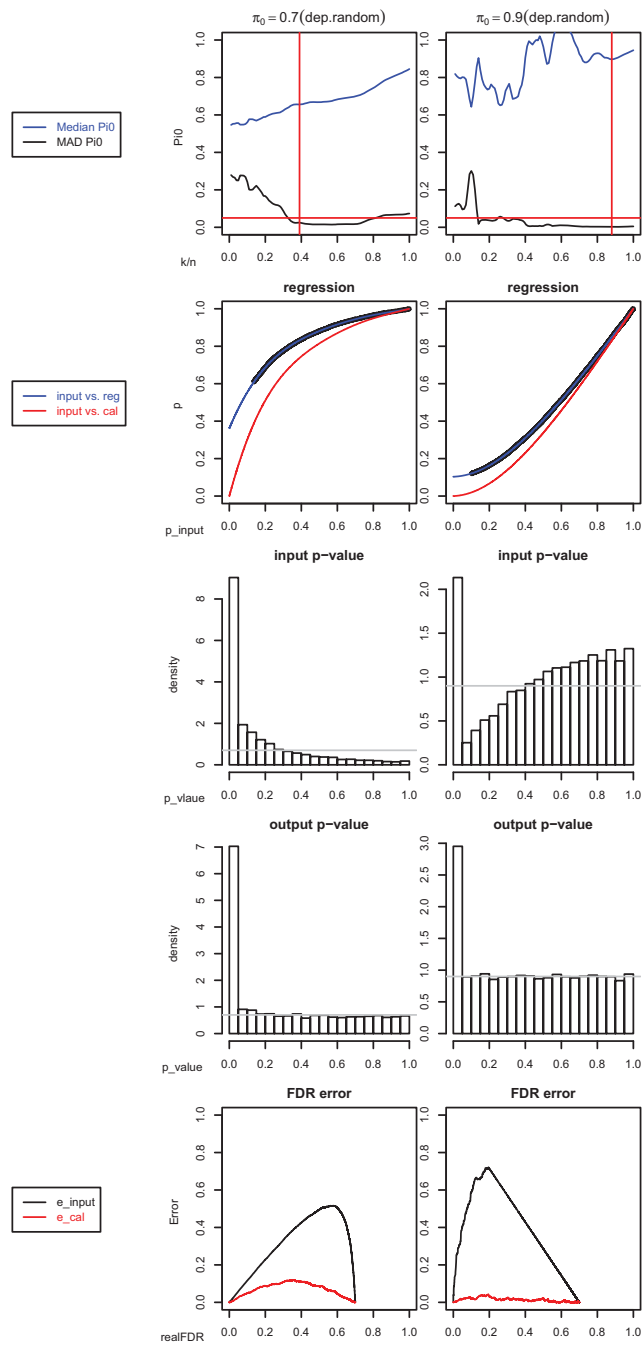


Figure 2.4: Procedural steps for the independent and dependent datasets with random dependent effect at $\pi_0 = 0.7$ and $\pi_0 = 0.9$. The detail description for plots in each row is same as Figure 2.2.

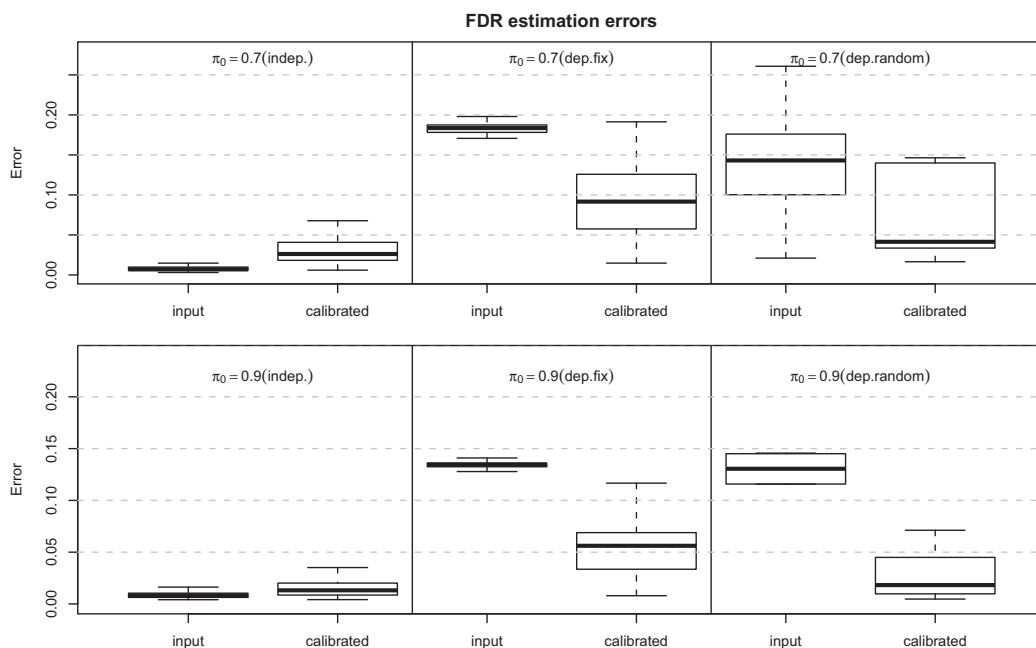


Figure 2.5: Boxplots of FDR estimation errors (the mean difference between real FDR and the FDR estimated by p -values) for 100 simulations of independent and dependent datasets at $\pi_0 = 0.7$ and $\pi_0 = 0.9$ before (input) and after applying ConReg-R (calibrated).

2.3.2 Combined p -values simulation

In many analyses, more than one dataset are involved and a meta-analysis by combining p -values from different studies or datasets is needed to estimate the overall significance for each gene. For example, (i) to find genes which are significant in at least one experiment, minimal p -values will be of interest; (ii) to identify genes which are significant across all the experiments, the maximal p -values will be of interest; and (iii) in order to detect genes which are significant on average, the product of p -values will be appropriate.

The distribution of combined p -values will not be uniform even under true null

hypotheses (Hedges and Olkin, 1985). For currently used meta-analysis methods, such as “minimal”, “maximal” or “product”, we can obtain the transformation functions to recalibrate the combined p -values to satisfy the condition of p -values are uniform distributed under true null hypotheses. However, for other more complicated meta-analysis methods, the transformation function cannot be determined accurately leading to under- or over-estimation of significance, and ConReg-R can provide the polynomial function approximation for the unknown transformation.

Suppose for gene i , the p -values $p_{ij}(j = 1, 2, \dots, L)$ follow the uniform distribution over $(0, 1)$, then $1 - (1 - p_{\min})^L \sim U(0, 1)$ and $p_{\max}^L \sim U(0, 1)$, where $p_{\min} = \min(p_{i1}, p_{i2}, \dots, p_{iL})$ and $p_{\max} = \max(p_{i1}, p_{i2}, \dots, p_{iL})$. For the p -values from “product” method, $-2 \sum_{j=1}^L \log(p_{ij}) \sim \chi_{2L}^2$ according to Fisher’s method (Fisher, 1948).

For each meta-analysis method, we simulated two data sets $Z^0 = (\delta_{ij})$, $Z = (z_{ij})(i = 1, \dots, n, j = 1, \dots, r)$ with $n(n = 10000)$ genes and $r(r = 10)$ repeats based on the formula as follows,

$$z_{ij} = b_i + \varepsilon_{ij}$$

where b_i ($b_i = 1$, if $i \leq n(1 - \pi_0)$ and $b_i = 0$ if $i > n(1 - \pi_0)$) denotes the biological effect, both $\delta_{ij} \sim N(0, 1)$ and $\varepsilon_{ij} \sim N(0, 1)$ are the background noise. The individual p -values are computed from two-sample t-test and the combined p -values are calculated by $L(L = 3)$ simulations.

Table 2.1: Combined p -values methods (Hedges and Olkin, 1985).

Method	Formula	Transformation
Min	$p_{\min} = \min(p_{i1}, p_{i2}, \dots, p_{iL})$	$1 - (1 - p_{\min})^L$
Max	$p_{\max} = \max(p_{i1}, p_{i2}, \dots, p_{iL})$	p_{\max}^L
Square	$p_{\text{sq}} = p_{i1}^2$	$\sqrt{p_{\text{sq}}}$
Sqroot	$p_{\text{sqrt}} = \sqrt{p_{i1}}$	p_{sqrt}^2
Prod	$p_{\text{prod}} = \prod_{j=1}^L p_{ij}$	$-2 \sum_{j=1}^L \log(p_{ij}) \sim \chi_{2L}^2$

To compare the results, we also included two other transformation methods, “square” and “square root”. All methods are listed in Table 2.1.

The two p -value histograms for each $\pi_0 = 0.7$ and $\pi_0 = 0.9$, and for each of five different methods are plotted in Figure 2.6. It can be seen from Figure 2.6 that the p -value histograms after theoretical transformation have constant frequency after 0.5 and the p -value density near 1 indicates the $\hat{\pi}_0$. However, the p -value histograms from “Min”, “Square”, “Prod” shifted towards 0 and the p -value histograms from “Max”, “Sqroot” shifted towards 1.

ConReg-R used the above combined p -values as input and the results are shown in Figure 2.7 ($\pi_0 = 0.7$) and Figure 2.8 ($\pi_0 = 0.9$). From Figure 2.7 and Figure 2.8, the regression curves are monotonic concave functions for “Min”, “Square”, “Prod” and monotonic convex functions for “Max”, “Sqroot”. The histograms after applying ConReg-R are also very similar to the theoretical transformed p -value histograms. The FDR estimation improved significantly after applying ConReg-R. It shows that the estimated FDR after applying ConReg-R is more likely to be the real FDR.

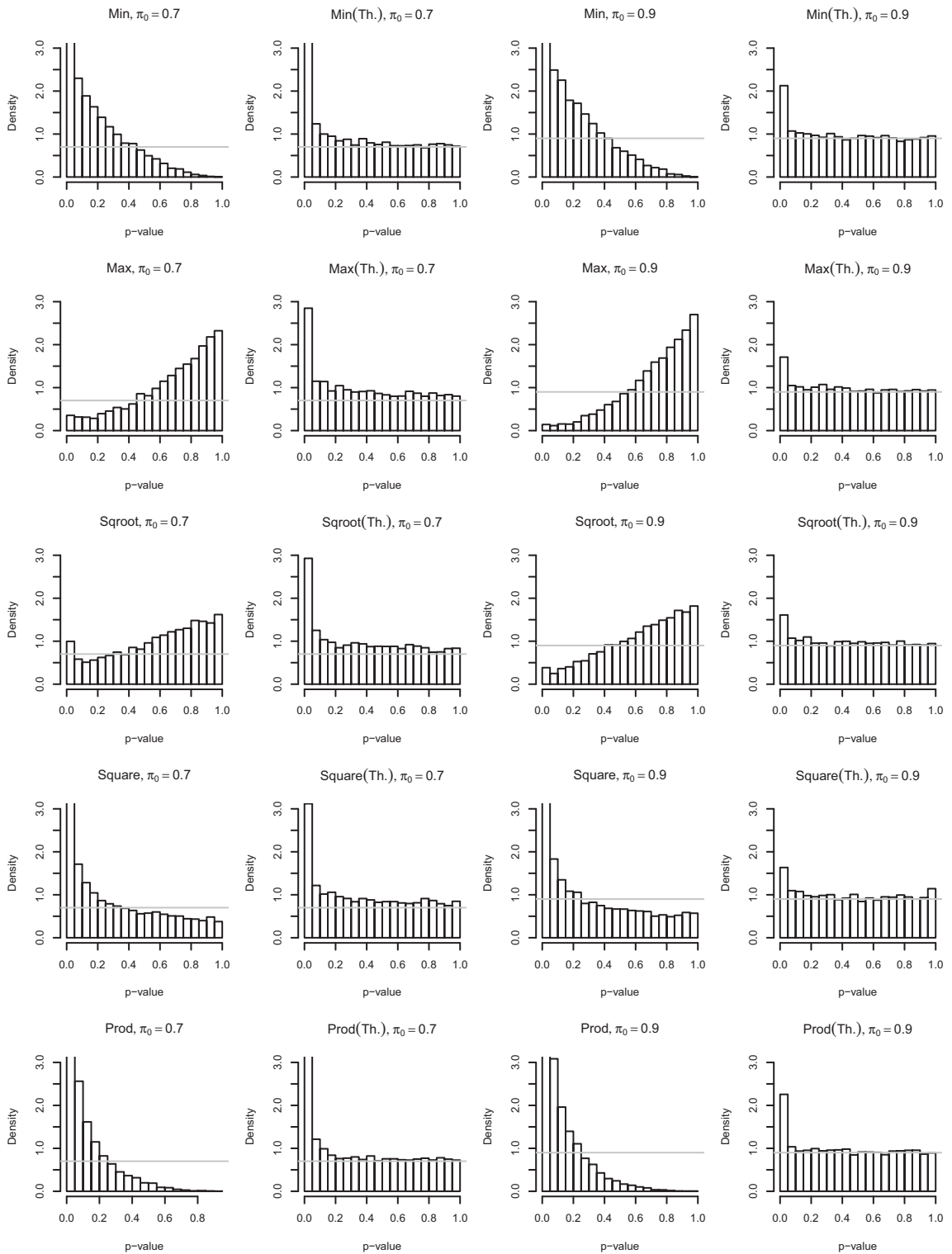


Figure 2.6: Density histograms for “Min”, “Max”, “Sqrt”, “Square” and “Prod” datasets at $\pi_0 = 0.7$ and $\pi_0 = 0.9$. (Th.) indicates the density histograms for each method after theoretical transformation. The gray horizontal line indicates the π_0 for each plot.

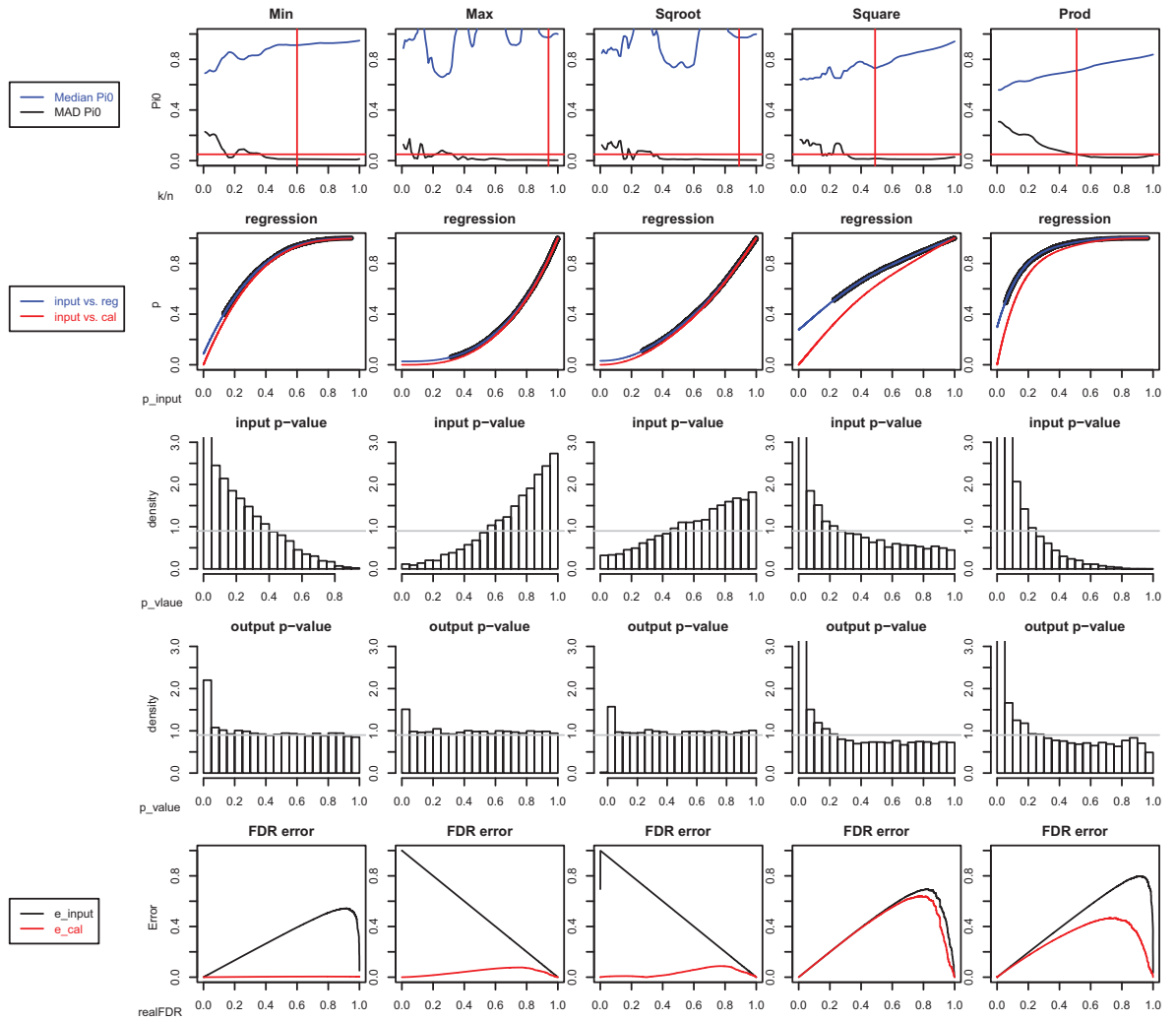


Figure 2.7: Procedure details for “Min”, “Max”, “Sqrt”, “Square” and “Prod” datasets at $\pi_0 = 0.7$. The detail description for plots in each row is same as Figure 2.2.

The results of using our procedure for 100 repeated simulations are summarized in Figure 2.9. The FDR estimation errors after applying ConReg-R are significantly less than those obtained without applying ConReg-R.

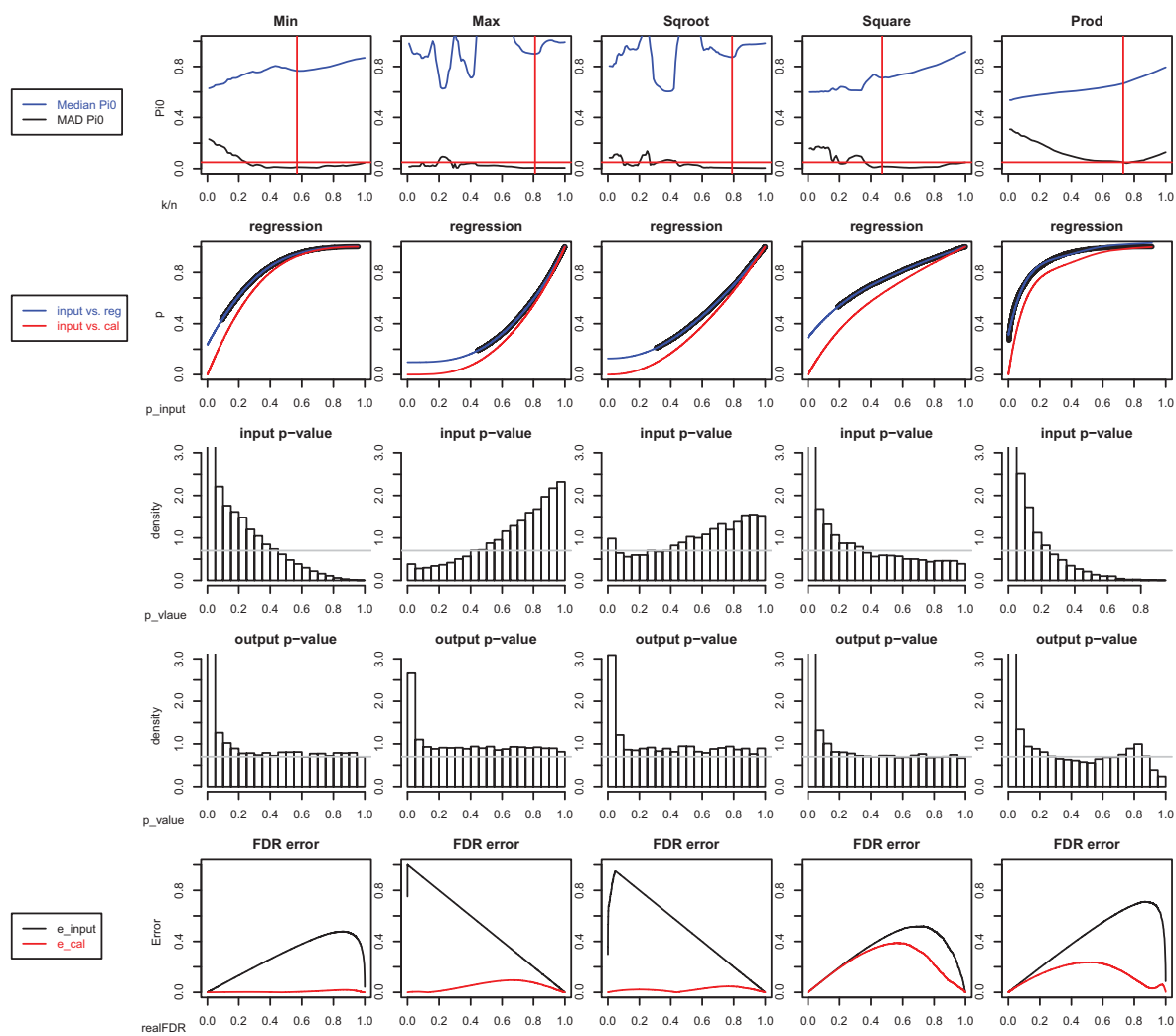


Figure 2.8: Procedure details for “Min”, “Max”, “Sqrt”, “Square” and “Prod”

datasets at $\pi_0 = 0.9$. The detail description for plots in each row is same as Figure

2.2.

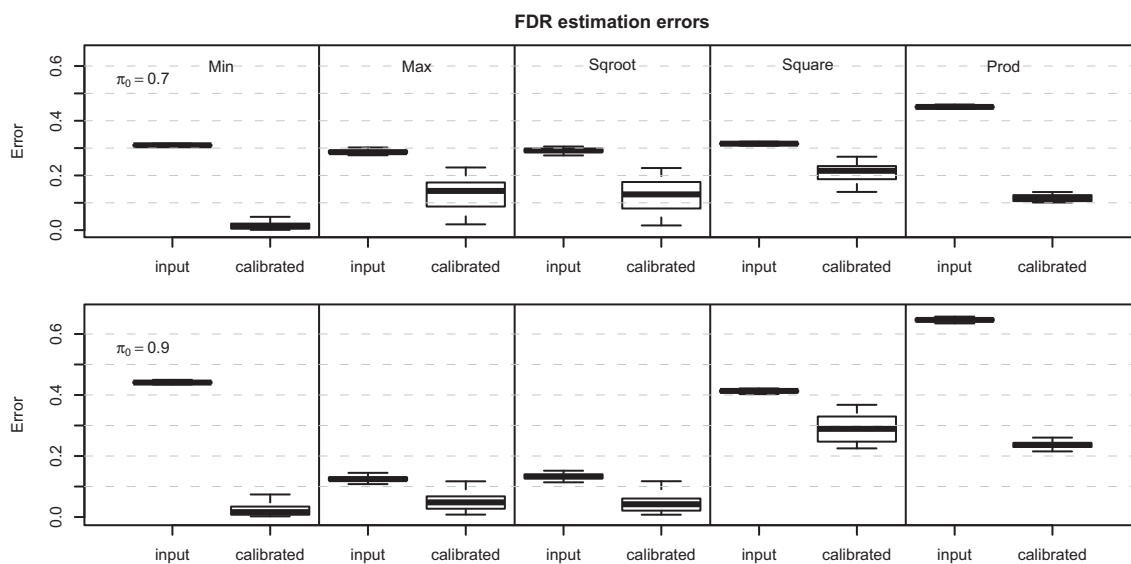


Figure 2.9: Boxplots of FDR estimation errors for 100 simulations of “Min”, “Max”, “Sqrt”, “Square” and “Prod” datasets at $\pi_0 = 0.7$ and $\pi_0 = 0.9$ before (input) and after applying ConReg-R (calibrated).

Chapter 3

iPLR: Iterative piecewise linear regression

This chapter describes the iPLR procedure to re-estimate null distribution from resampling procedures and the simulation results.

3.1 Background

Batch dependent systematic variations or batch effects (Lamb *et al.*, 2006) are commonly observed across multiple batches of microarray experiments. Batch effect influences the expression measurements of all genes in the arrays; the effect on a single gene is random but similar across all arrays in the batch and different from other genes and batches (Li and Wong, 2003). It has been observed by

many researchers that the established normalization and preprocessing methods cannot fully eliminate batch effects (Johnson *et al.*, 2007) and hence developed procedures to account for batch effects at probe level in the differential expression models (Alter *et al.*, 2000; Nielsen *et al.*, 2002; Benito *et al.*, 2004).

A few popular methods are SVD (Singular Value Decomposition) / PCA (Principal Component Analysis) (Alter *et al.*, 2000; Nielsen *et al.*, 2002), DWD (Distance Weighted Discrimination) (Benito *et al.*, 2004) and empirical Bayes methods (Johnson *et al.*, 2007) which treat batch as a factor assuming that the experimental batches are not confounded with the biological groups of interest i.e. batch and treatment variables are not collinear. In other words, each batch contains arrays of samples from different biological groups (see the row titled “ideal batch” in Table 3.1). However the problem is not amenable to such analysis if the biological groups are confounded with that of the batches i.e., the arrays in a batch receive samples from one biological group of interest and the arrays in the other batch contain samples from the other biological group of interest (see the row titled “batch confounding” in Table 3.1 for illustration). It results in collinearity of batch and treatment variable which means the above methods are not applicable. It is unavoidable in many practical situations as one wants to compare the data from one experiment or laboratory to the data from another experiment or laboratory which essentially means batch confounded biological groups. Time course experiments spread over long time horizons may also result in batch confounding when samples from different time-points are compared for change of expression.

Table 3.1: Illustration of batch confounding. s1-8 indicate sample 1-8, c1(2) indicates class 1(2), and b1(2) indicates batch 1(2).

samples	s1	s2	s3	s4	s5	s6	s7	s8
class	c1	c1	c1	c1	c2	c2	c2	c2
batch confounding	b1	b1	b1	b1	b2	b2	b2	b2
ideal batch	b1	b1	b2	b2	b1	b1	b2	b2

Similarly, batch confounding is unavoidable in huge experiments even though all groups were generated in the same laboratory.

Batch confounding has severe influence on differential expression analysis as the biologically differentially expressed genes are mixed up with large number of mere batch affected expression measurements. Even after microarray data preprocessing and normalization, batch confounding still exists in the data. It may lead to gross incorrect estimation of statistical significance, i.e. false discovery rate (FDR), to an intolerable limit as several batch affected biologically irrelevant genes will also have significantly lower p -values. This is true irrespective of whether the statistical significance is assessed using resampling as in SAM (Significance Analysis of Microarrays) (Tusher *et al.*, 2001) or parametric distribution as in LIMMA (Linear Models for Microarray Data) (Smyth, 2004). In the absence of gold standard positive and negative gene sets in genome-wide expression studies, FDR, being an important parameter, needs to be accurately estimated. For example, FDR has been used to estimate the effects of certain treatment or condition on a cell culture via the number of genes passed the FDR cut-off (Storey and Tib-

shirani, 2003). Hence it is important to estimate FDR as accurately as possible even in the batch confounded data analysis to facilitate correct conclusions on the significantly affected genes.

To address these issues, we developed a method called *stepped linear regression* (SLR) (Li *et al.*, 2008b) to improve FDR estimation in batch confounded data. After increasing the accuracy and usability of SLR, we upgraded SLR to *iterative piecewise linear regression* (iPLR) which is major modification of SLR. iPLR re-estimates the expected differential expression statistics under the assumption that the expression difference due to batch variation is smaller than that of the biological variation. FDR is estimated based on the re-estimated expected statistics i.e. the null distribution is re-estimated. After applying iPLR, we can get accurate significance assessment and biologically significant genes. Moreover, our method provides a better interpretation for the linear model in this paper and incorporated procedure to handle one-sided tests.

We present our iPLR in the context of SAM (Tusher *et al.*, 2001). SAM is a statistical technique for finding significantly differentially expressed genes in microarray experiments. SAM assigns a score called d-score to each gene on the basis of change in gene expression relative to the standard deviation of replicated measurements. The genes with d-scores greater than certain threshold are declared to be differentially expressed. This threshold corresponds to certain false discovery rate (FDR), the percentage of genes identified by chance for the given

d-score by permuting the class labels. Those genes will be regarded as significantly biologically relevant genes according to the data. However, in the case of batch confounding, many of them may not be actually relevant to the underlying biology of interest. Our iPLR helps correct this artifact.

Though iPLR is presented in the context of SAM analysis for simplicity, the method is equally applicable to any reasonable statistical procedure based on resampling strategy. Our results show that iPLR is effective in estimating FDR accurately both in simulated as well as real data with batch confounding. We demonstrate how iPLR corrects for the incorrectly magnified assessment of effects of certain conditions or treatments on gene expression.

3.2 Methods

The iterative piecewise linear regression (iPLR) is based on the following assumptions: (a) for those biologically differentially expressed genes, the biological influence is much greater than those of the batch effects' influence; (b) the batch effect is independent of biological effect; and (c) the proportion of biologically non-differentially expressed genes (π_0) is larger than 0.5.

3.2.1 Re-estimating the expected statistics

For a SAM analysis based on a two-sided test statistics (Chu *et al.*), we first obtain the SAM computed statistic $d_i = \frac{r_i}{s_i + s_0}$ for each gene $g_i (i = 1, 2, \dots, n)$, where r_i is a score, s_i is a standard deviation, and s_0 is an exchangeability factor. Without loss of generality we assume $d_1 \leq d_2 \leq \dots \leq d_n$. To compute FDR, SAM performs permutations by random labeling each sample for as many times as defined by the user to estimate the expected values of these order statistics $\bar{d}_1 \leq \bar{d}_2 \leq \dots \leq \bar{d}_n$.

When batch effect exists in the data and is confounded with biological effect, we propose the linear model between observed statistics d_i and expected statistics \bar{d}_i as follows:

$$d_i = a\bar{d}_i + b + c_i + e_i, \quad (3.1)$$

where a and b are batch effect factors, c_i is the biological effect factor and e_i is the model error ($i = 1, 2, \dots, n$). $c_i = 0$ if gene g_i has no differential expression between different classes of the experiment.

It is difficult to estimate batch effect factors without knowing biological effect factor c_i . Therefore, we simply approximate c_i by a linear function in \bar{d}_i when $c_i \neq 0$. Based on (3.1), we are led to consider

$$d_i = \begin{cases} a\bar{d}_i + b + e_i & \text{if } c_i = 0 \\ a\bar{d}_i + b + c_{a+}\bar{d}_i + c_{b+} + e_i & \text{if } c_i > 0 \\ a\bar{d}_i + b + c_{a-}\bar{d}_i + c_{b-} + e_i & \text{if } c_i < 0 \end{cases} \quad (3.2)$$

where c_{a+}, c_{b+} and c_{a-}, c_{b-} are the coefficients of the linearity between c_i and \bar{d}_i . From (3.2), we perform the iterative piecewise linear regression (iPLR) to estimate the batch effect factors a and b . In iPLR, we use iterative approach to identify the regression section with $c_i = 0$ to estimate the batch effect factors a and b . The proportion of non-differentially expressed genes, π_0 , will be estimated by $\frac{\#\{c_i=0\}}{n}$.

After estimating the batch effect factors a and b , we can re-estimate the expected statistics to eliminate the batch effect in FDR estimation. The model in (3.1) is about comparing quantiles or ordered statistics (origin is about the 50th percentile or median which is usually close to 0) for the observed test statistics (a combination of null hypotheses and alternative hypotheses) and the test statistics of null distribution obtained by resampling. If π_0 is very close to 1, then quantiles of both distributions will be very close to each other. So the test statistics and the expected statistics will lie close to the diagonal of the observed statistic quantiles versus the expected statistic quantiles plot. Then we can set $a = 1$ to eliminate the batch effect. However, when π_0 is not very close to 1 (for example, $\pi_0 = 0.7$), we have to consider the fact that the distribution of the test statistics is a mixture of the distribution of statistics under null hypothesis (null distribution multiplied by π_0) and the alternate distribution. If the null distribution of test statistics is uniform, the slope for the observed statistic quantiles versus the expected statistic quantiles plot for $c_i = 0$ is π_0^{-1} . In typical hypothesis testing in microarray experiments, the null distributions are unimodal. In these typical cases, we can approximate this slope as π_0^{-1} to achieve a better estimate of the FDR based on

Proposition 3.1. Therefore, we set $\hat{a} = \hat{\pi}_0^{-1}$ to eliminate the batch effect. Then, we define the re-estimated expected order statistics as

$$\bar{d}_i^* = \hat{\pi}_0(\hat{a}\bar{d}_i + \hat{b}), \quad (3.3)$$

and then (3.1) will be rewritten as

$$d_i = \hat{\pi}_0^{-1}\bar{d}_i^* + c_i + e_i. \quad (3.4)$$

This is the linear model after eliminating the batch effect.

Below is a proof that the slope of the Q-Q plot at the 50th quantile is π_0^{-1} under some mild conditions. Let g be the function which maps the q th quantile of the null cumulative distribution function to the q th quantile of the mixture cumulative distribution function. We shall identify g below after introducing some notations.

Let f_0 and f_1 be symmetric probability density functions. For $\pi_0 \in (0, 1)$, we define a probability density function f as

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x).$$

Let F, F_0 and F_1 be the cumulative distribution functions of f, f_0 and f_1 respectively.

To see why g is given as in (3.5). We let x_q and y_q be the q th quantile relative to F_0 and F respectively. That is, $x_q = F_0^{-1}(q)$ and $y_q = F^{-1}(q)$. Equivalently, $q = F_0(x_q)$ and $y_q = F^{-1}(F_0(x_q))$. As $y_q = g(x_q)$, this leads to $g(x) = F^{-1}(F_0(x))$.

Proposition 3.1. *With f_0, f_1, f, F_0, F_1 and F as defined above, we assume further that*

1. $f_0(x)$ is continuous at $x = 0$ and $f_0(0) > 0$, and
2. $f_1(x) = 0$ for x in the neighborhood of 0.

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be defined as

$$g(x) = F^{-1}(F_0(x)). \quad (3.5)$$

Then $g'(0) = \pi_0^{-1}$.

Proof. Since $0 = x_{0.5} \mapsto y_{0.5} = 0$, we have $g(0) = 0$. From (3.5), we obtain

$$F_0(x) = F(g(x)). \quad (3.6)$$

Differentiate (3.6) with respect to x in the neighborhood of 0 where $f_1 = 0$, we obtain

$$f_0(x) = f(g(x))g'(x) = [\pi_0 f_0(g(x)) + (1 - \pi_0)f_1(g(x))]g'(x).$$

Since $g(0) = 0$ and $f_1(0) = 0$, evaluating the above at $x = 0$ gives

$$f_0(0) = \pi_0 f_0(0)g'(0)$$

which leads to $g'(0) = 1/\pi_0$ after canceling out $f_0(0)$ which is positive. \square

Examples: We simulated $x \sim f_0$ and $y \sim \pi_0 f_0 + (1 - \pi_0)f_1$, where we set f_0 is $N(0, 1)$ and f_1 is $-\chi^2(1) - 1$ and $\chi^2(1) + 1$. We can see the example Q-Q plots for $\pi_0 = 0.9$ and $\pi_0 = 0.7$. The red lines are approximately $y_q = \pi_0^{-1}x_q$.

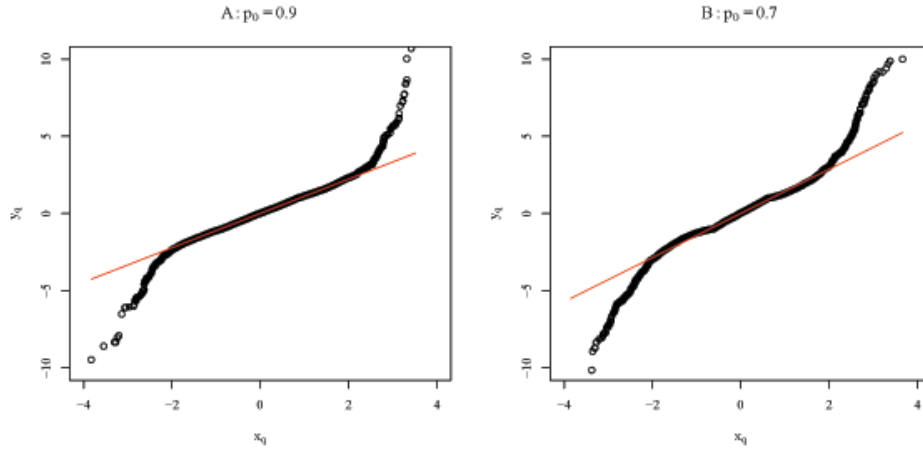


Figure 3.1: Examples for Q-Q plot slope approximation.

3.2.2 Iterative piecewise linear regression

The iPLR takes observed statistics d_i and expected statistics \bar{d}_i as input data, and uses iterative approach to search for the best piecewise linear regression model fit in (3.2). The batch effect factors and π_0 are estimated by this model, then iPLR re-estimates the expected order statistics \bar{d}_i^* by (3.3). Finally, iPLR outputs the re-estimated FDR. The work flow for iPLR is illustrated in Figure 3.2.

By assumption (c), there are more than 50% non-differentially expressed genes in the dataset. Therefore, the baseline, regression line for $c_i = 0$ part in iPLR, will include more than half of the data and batch effect factors a and b will be estimated from this portion of the data.

At the initialization step, we set $\hat{\pi}_0^{(0)} = 1$, and define the data split

$$S_0 = \{D^{(0)}, D_-^{(0)}, D_+^{(0)}\}, \quad (3.7)$$

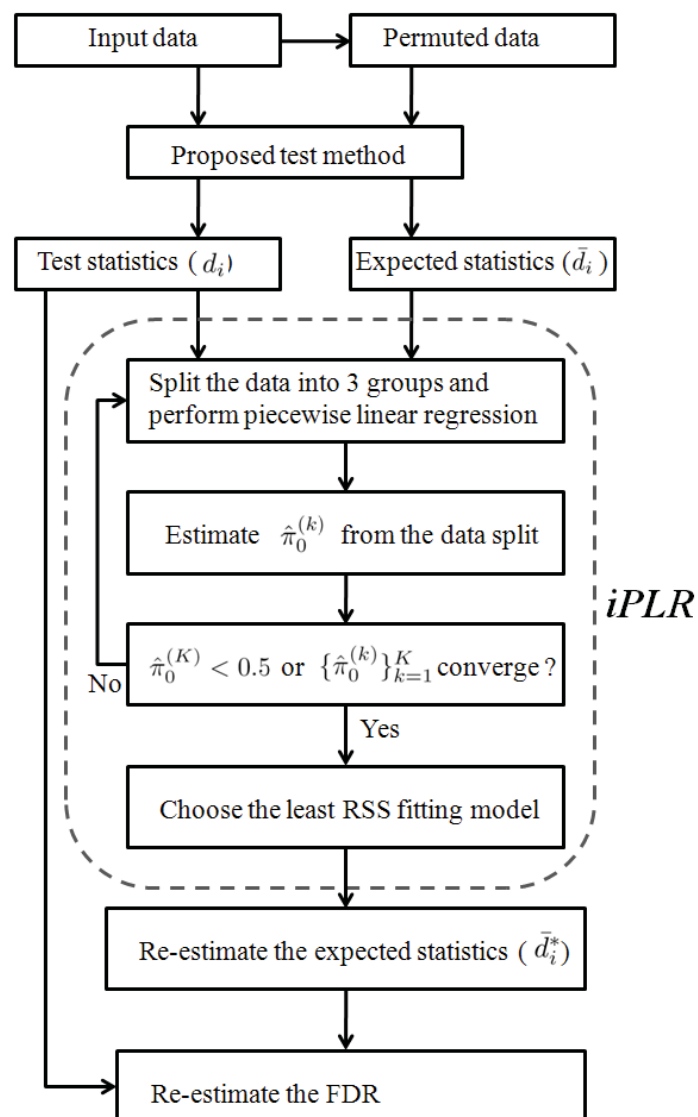


Figure 3.2: Work flow for iPLR.

where $D^{(0)} := \{(d_i, \bar{d}_i), i = 1, 2, \dots, n\}$ is the initial baseline dataset and $D_-^{(0)} = D_+^{(0)} = \phi$ (empty set).

We perform the linear regression in $D^{(0)}$ and this is the baseline in the initial step. Let $\delta^{(0)}$ be the standard deviation of baseline regression errors. The next baseline dataset $D^{(1)}$ is generated by excluding the data points which are far away from the baseline. We define the split $S_1 = \{D^{(1)}, D_-^{(1)}, D_+^{(1)}\}$ as

$$\begin{aligned} D^{(1)} &= \{(d_i, \bar{d}_i) | -z\delta^{(0)} \leq l_i^{(0)} \leq z\delta^{(0)}\}, \quad \hat{\pi}_0^{(1)} = \frac{\#\{D^{(1)}\}}{n} \\ D_-^{(1)} &= \{(d_i, \bar{d}_i) | l_i^{(0)} < -z\delta^{(0)}\}, \quad D_+^{(1)} = \{(d_i, \bar{d}_i) | l_i^{(0)} > z\delta^{(0)}\}, \end{aligned} \quad (3.8)$$

where $l_i^{(0)}$ is the distance between (d_i, \bar{d}_i) to the regression baseline, z is a pre-defined boundary cutoff. $D_-^{(1)}$ and $D_+^{(1)}$ indicate $c_i < 0$ and $c_i > 0$ and generally distributed at the left and right tails of the data. Then we perform 3-piece linear regression for $D^{(1)}$, $D_-^{(1)}$ and $D_+^{(1)}$ separately. We repeat the above procedure to generate $S_k = \{D^{(k)}, D_-^{(k)}, D_+^{(k)}\}$ and $\hat{\pi}_0^{(k)}$ from $S_{k-1} = \{D^{(k-1)}, D_-^{(k-1)}, D_+^{(k-1)}\}$ and $\hat{\pi}_0^{(k-1)}$ until convergence is reached. The procedure is said to converge at $k = K$ if $\hat{\pi}_0^{(K)} < 0.5$ or the sequence $\{\hat{\pi}_0^{(k)}\}_{k=1}^K$ converges to a constant i.e. $|\hat{\pi}_0^{(K)} - \hat{\pi}_0^{(K-1)}| < 10^{-3}$.

Among the sequence of data-splits S_k for $k = 1, \dots, K$, we choose the split that gives lowest fitting RSS (residual sum of squares) for 3-piece linear regression, and $\hat{\pi}_0$ is estimated by this split. In iPLR, there are only two free parameters to be pre-selected, boundary cutoff z and stopping cutoff 10^{-3} . We can set $z = 3$ implying the 3 standard deviation boundary. Different choices of z and stopping cutoff

influence the search bandwidth and the number of iterations without affecting the outcome significantly. The first two steps (from $D^{(0)}$ to S_1 , and S_1 to S_2) for iPLR are illustrated in Figure 3.3.

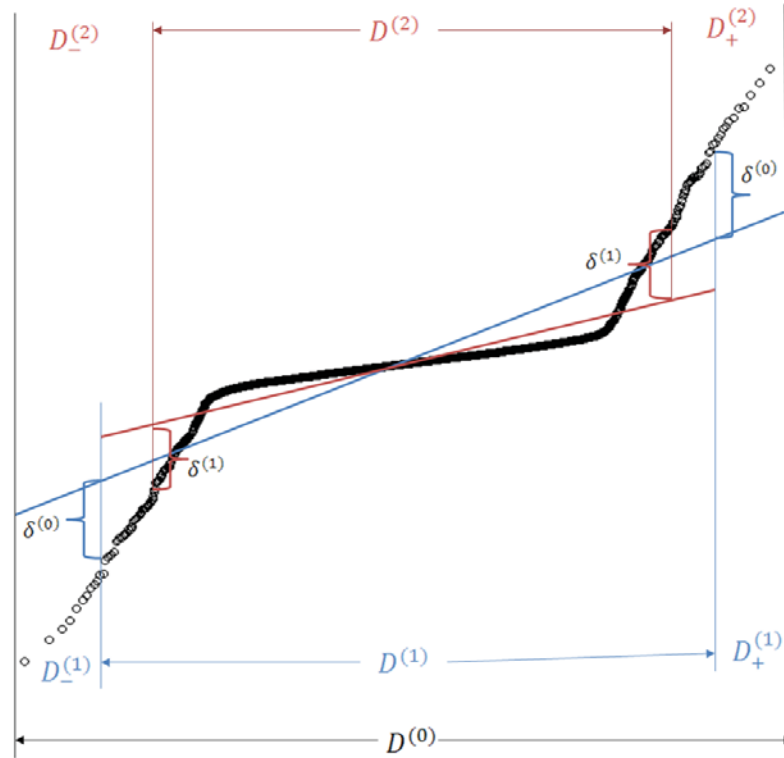


Figure 3.3: Illustration of first two iterations in iPLR. They can be generalized to the following iterations.

The following steps detail the computational procedure for iPLR.

1. Set $\hat{\pi}_0^{(0)} = 1$. Perform the linear regression for $D^{(0)}$ to determine the baseline. Compute the distance of each point to the baseline and the standard deviation of baseline regression errors $\delta^{(0)}$.
2. Calculate $\hat{\pi}_0^{(1)}$ using (3.8) and perform a 3-piece linear regression for $D^{(1)}$,

- $D_-^{(1)}$ and $D_+^{(1)}$ separately. Compute the standard deviation of baseline regression errors $\delta^{(1)}$ and RSS of the 3-piece regression.
3. Repeat step 2 to obtain S_k and $\hat{\pi}_0^{(k)}$ from $S_{(k-1)}$ until $k = K$ for which $\hat{\pi}_0^{(K)} < 0.5$ or $|\hat{\pi}_0^{(K)} - \hat{\pi}_0^{(K-1)}| < 10^{-3}$.
 4. Choose the estimation of $\hat{\pi}_0$ as $\hat{\pi}_0^{(k)}$ with the least RSS fitting for the iPLR, and the batch effect factors a and b are estimated in the baseline regression using this $\hat{\pi}_0$.
 5. Re-estimate the expected statistics using (3.3). Re-estimate the FDR for each gene.

3.2.3 iPLR for one-sided test

The above procedure is designed for 3-piece linear regression for a two-sided test. If the test statistics are from one-sided test, the biological effect $c_i \geq 0, i = 1, 2, \dots, n$. Therefore, (3.2) becomes

$$d_i = \begin{cases} a\bar{d}_i + b + e_i & \text{if } c_i = 0 \\ a\bar{d}_i + b + c_{a+}\bar{d}_i + c_{b+} + e_i & \text{if } c_i > 0 \end{cases}. \quad (3.9)$$

Then iPLR procedure can be modified to take care of this one-sided test. Indeed, we only need to set $D_-^{(k)} = \phi, k = 1, 2, \dots, K$ in the definition of the split. Subsequently, iPLR performs a 2-piece linear regression by removing one piece for $D_-^{(k)}, k = 1, 2, \dots, K$ and the rest of iPLR procedure remains the same.

3.3 Results

We demonstrate that the effects of batch confounding on FDR estimation and the efficacy of iPLR in alleviating it using both simulated data and real data. Using simulated data, we show that iPLR does not introduce any artifacts in FDR estimation for data without batch confounding, and that iPLR corrects the influence of batch confounding if it is present in the data.

3.3.1 Two-class simulations

A two-group data was simulated using the following rule

$$x_{ijk} = \mu_{ik} + \eta_{ik} + \epsilon_{ijk}, \quad (3.10)$$

where x_{ijk} is an expression measurement of gene g_i ($i = 1, 2, \dots, n = 10000$) in sample S_j ($j = 1, 2, \dots, 10$) in group G_k ($k = 1, 2$), and ϵ_{ijk} are standard normal noise. The biological effect μ_{ik} and global batch effect η_{ik} (η_{ik} is the effect of batch on the gene expression x_{ijk} which is different from the effect of batch confounding on the relationship between d_i and \bar{d}_i) are defined as follows:

$$\begin{aligned} \mu_{i1} &= 0 \text{ for } 1 \leq i \leq n, \\ \eta_{i1} &= 0 \text{ for } 1 \leq i \leq n, \\ \mu_{i2} &= \begin{cases} \theta_{i1} \sim N(0, \sigma_\mu^2) & \text{for } 1 \leq i \leq m \\ 0 & \text{for } m < i \leq n \end{cases}, \\ \eta_{i2} &= \theta_{i2} \sim N(0, \sigma_\eta^2) \text{ for } 1 \leq i \leq n, \end{aligned} \quad (3.11)$$

Table 3.2: Parameters used to simulate the 4 different datasets A, B, C and D.

	Dataset Simulation Parameters			
Datasets	Batch Effect	σ_η	σ_μ	π_0
A	No	0	4	0.95
B	Yes	2	4	0.95
C	No	0	4	0.7
D	Yes	2	4	0.7

where m is the number of differentially expressed genes and n is the total number of genes. The model parameters signify that the batch effect and biological effect are independent and the level of differential expression and batch effect varies from gene to gene. The fraction $1 - (m/n)$ is π_0 : the fraction of non-differentially expressed genes or genes not affected by biological treatments.

We simulated four different datasets of $n = 10000$ genes each using two different settings for two choices of σ_η and π_0 as shown in Table 3.2 while keeping $\sigma_\mu = 4$. Datasets A and C are simulated without batch effects and analyzed with our procedure in order to find out whether our procedure would introduce any artifacts in FDR estimation or not (i.e., FDR estimates before and after re-estimation of \bar{d}_i should be close to each other for non-batch affected data). Datasets B and D are batch effect confounded with reasonably different values of π_0 whose FDR estimates before re-estimation are expected to be far from reality while the FDR estimates after re-estimation are expected to be close to the reality. We used SAM on each of the four datasets to obtain d-statistics for original as well as permuted data. We applied our procedure on each pair of d-statistic sets.

Table 3.3 gives the estimates of π_0 , the number of genes identified significant, and the FDR estimates from applying SAM only, and from SAM followed by applying iPLR. The estimates of π_0 after applying iPLR are markedly more accurate than not applying iPLR for datasets B and D which are batch confounded. This shows that iPLR succeeds in reducing the batch effects in FDR estimation. Both estimates (before and after applying iPLR) of π_0 for datasets A and C, which are not batch confounded, agree quite well with the true values. Indeed, the estimates after applying iPLR are slightly better than that of not using iPLR. This shows that iPLR does not introduce any bias in the absence batch effect. The other π_0 estimation methods which are only based on the p -value distribution will also be affected by batch confounded data. For example, we used a cross-validated approach (Celisse and Robin, 2010) to estimate π_0 for datasets A-D and the results ($\hat{\pi}_0\{A, B, C, D\} = \{0.9593, 0.3365, 0.8037, 0.3192\}$) are similar to SAM estimation. It shows that p -value distribution based procedures cannot solve the bias introduced by batch confounding.

The estimated FDR and true FDR for both before and after iPLR re-estimation procedure are plotted in Figures 3.4(A) and 3.4(B). Ideally, the estimated FDR should be close to the real FDR (smooth black curve), the closer the better. Both plots for dataset A are similar and are close to the real FDR curve. The estimated FDR after iPLR is slightly higher than real FDR at $FDR > 0.6$ which is not important for practical purposes. However, FDR plots for original SAM and after iPLR re-estimation are quite different for dataset B which is batch confounded.

Table 3.3: Significant gene tables for dataset ABCD.

delta	#sig.genes	$\widehat{\text{FDR}}$	#sig.genes	$\widehat{\text{FDR}}$
dataset A ($\pi_0 = 0.95$)	SAM ($\hat{\pi}_0 = 0.975$)		SAM+iPLR ($\hat{\pi}_0 = 0.9673$)	
0.1	672	0.4832	575	0.4880
0.5	354	0.0137	345	0.0168
1	289	0	287	0
5	34	0	34	0
dataset B ($\pi_0 = 0.95$)	SAM ($\hat{\pi}_0 = 0.2822$)		SAM+iPLR ($\hat{\pi}_0 = 0.9726$)	
0.1	9327	0.2293	1480	0.7055
0.5	6785	0.0394	308	0.2242
1	4300	0.0002	112	0.0434
5	50	0	5	0
dataset C ($\pi_0 = 0.7$)	SAM ($\hat{\pi}_0 = 0.7902$)		SAM+iPLR ($\hat{\pi}_0 = 0.7833$)	
0.1	4918	0.6025	4101	0.6095
0.5	2376	0.0349	2299	0.0537
1	1967	0.0004	1921	0.0008
5	248	0	227	0
dataset D ($\pi_0 = 0.7$)	SAM ($\hat{\pi}_0 = 0.236$)		SAM+iPLR ($\hat{\pi}_0 = 0.8599$)	
0.1	9470	0.1963	5431	0.7831
0.5	7419	0.0396	2110	0.3637
1	5216	0.0002	1081	0.0823
5	313	0	27	0

Plot for SAM is almost 0 showing how severely the FDR was underestimated. But FDR after iPLR re-estimation is closer to the real FDR, even though FDR still underestimated the real FDR due to the influence of batch confounding on the permutation procedure (Xie *et al.*, 2005).

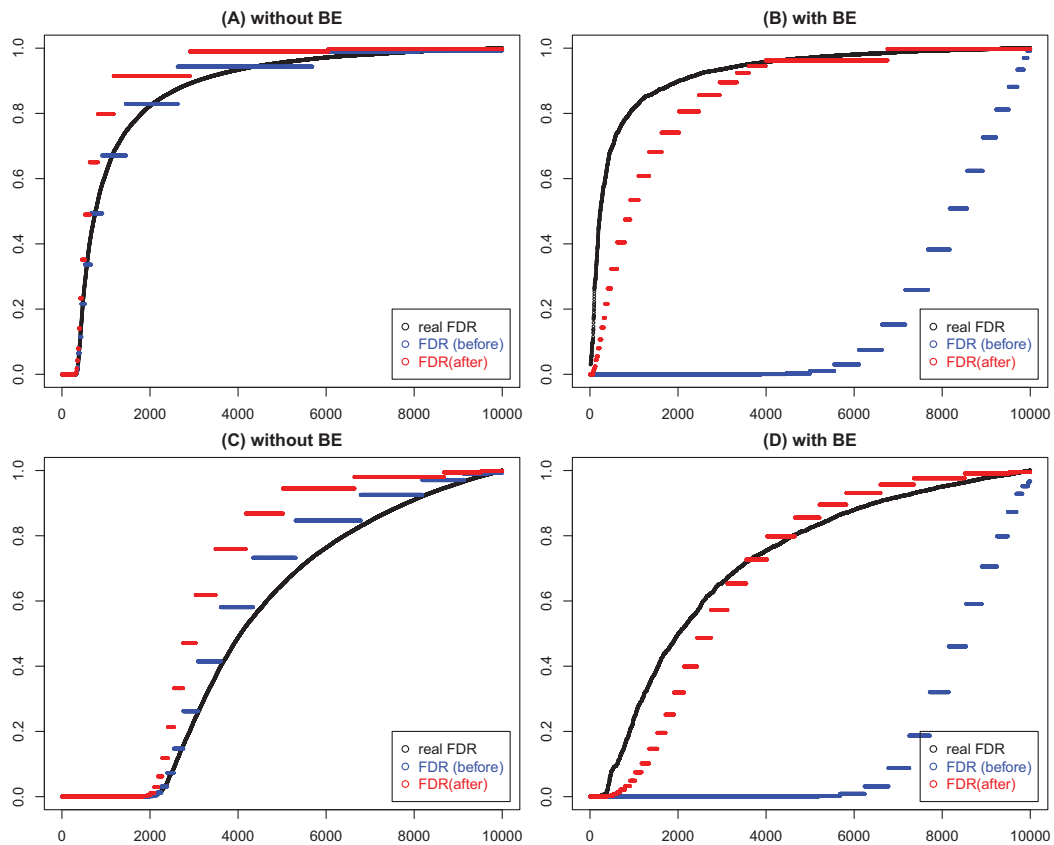


Figure 3.4: FDR comparison for simulation data sets A, B, C and D. Black points indicate the real FDR. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation.

Similar results are shown for datasets C and D in Table 3.3. The estimates of π_0 are consistently closer to the true value of 0.7 irrespective of the presence or the absence of batch effect after applying our iPLR procedure. In the presence of batch effect, the original SAM FDR estimates are severely biased by batch effect.

Table 3.4: Parameters used to simulate the 4 different datasets MA, MB, MC and MD.

Datasets	Dataset Simulation Parameters				
	Batch Effect	$\sigma_{\eta1}$	$\sigma_{\eta2}$	σ_{μ}	π_0
MA	No	0	0	4	0.95
MB	Yes	1	1	4	0.95
MC	No	0	0	4	0.7
MD	Yes	1	1	4	0.7

Similar differences were observed even for FDR estimates for various values of δ . The estimated FDR and true FDR for both before and after iPLR re-estimation are plotted in Figures 3.4(C) and 3.4(D). Both plots for dataset C are similar and close to real FDR for FDR in $[0, 0.5]$ except towards FDR=1 which is not critical in differential expression analysis. However, FDR plots for original SAM and that of iPLR adjusted are quite different for the batch confounded dataset D. Plot for SAM is almost 0 showing how erroneous the FDR estimation could be. FDR after iPLR adjustment, on the other hand, is much closer to the real FDR.

3.3.2 Multi-class simulations

iPLR is not only applicable to two-class analysis with two batches. We can easily adapt it to analyze multi-class dataset with more than two batches. We do this by considering a 2-piece linear regression instead of a 3-piece linear regression since up-regulated and down-regulated gene groups in two-class analysis will be merged into one single differentially expressed gene group in multi-class analysis.

We simulated three-class datasets using similar procedure as in two-class simulations by adding one more class for each dataset. As in two-class simulations, we simulated 2 different datasets without batch effect ($\sigma_\mu = 4, \sigma_{\eta_1} = \sigma_{\eta_2} = 0$) and 2 datasets with batch effect ($\sigma_\mu = 4, \sigma_{\eta_1} = \sigma_{\eta_2} = 1$) for $\pi_0 = 0.95$ and $\pi_0 = 0.7$. The parameter settings are listed in Table 3.4. We used SAM multi-class method to generate d-score and permuted score, and then performed iPLR (2-piecewise linear regression) to re-estimate the FDR. Results for these 4 datasets are listed in Table 3.5. FDR comparisons are shown in Figure 3.5. Similar to the results of two-class simulations, iPLR can accurately estimate the batch effect factors in datasets and the FDR.

Table 3.5: Significant gene tables for Multi-class simulated dataset MA-MD.

delta	#sig.genes	$\widehat{\text{FDR}}$	#sig.genes	$\widehat{\text{FDR}}$
dataset MA ($\pi_0 = 0.95$)	SAM ($\hat{\pi}_0 = 0.9762$)		SAM+iPLR ($\hat{\pi}_0 = 0.9567$)	
0.05	572	0.3413	529	0.3418
0.1	415	0.1011	400	0.1112
0.25	336	0.0000	332	0.0029
0.5	254	0.0000	250	0.0000
dataset MB ($\pi_0 = 0.95$)	SAM ($\hat{\pi}_0 = 0.2194$)		SAM+iPLR ($\hat{\pi}_0 = 0.9744$)	
0.05	9663	0.1845	639	0.5542
0.1	8734	0.1103	369	0.3366
0.25	4604	0.0027	182	0.0669
0.5	928	0.0000	107	0.0000
dataset MC ($\pi_0 = 0.7$)	SAM ($\hat{\pi}_0 = 0.8168$)		SAM+iPLR ($\hat{\pi}_0 = 0.7492$)	
0.1	2666	0.2045	2837	0.2084
0.25	2080	0.0023	2122	0.0031
0.5	1629	0.0000	1653	0.0000
1	908	0.0000	925	0.0000
dataset MD ($\pi_0 = 0.7$)	SAM ($\hat{\pi}_0 = 0.2008$)		SAM+iPLR ($\hat{\pi}_0 = 0.8135$)	
0.1	8957	0.1099	4129	0.5334
0.25	5659	0.0040	1791	0.1423
0.5	2474	0.0000	1098	0.0051
1	1006	0.0000	478	0.0000

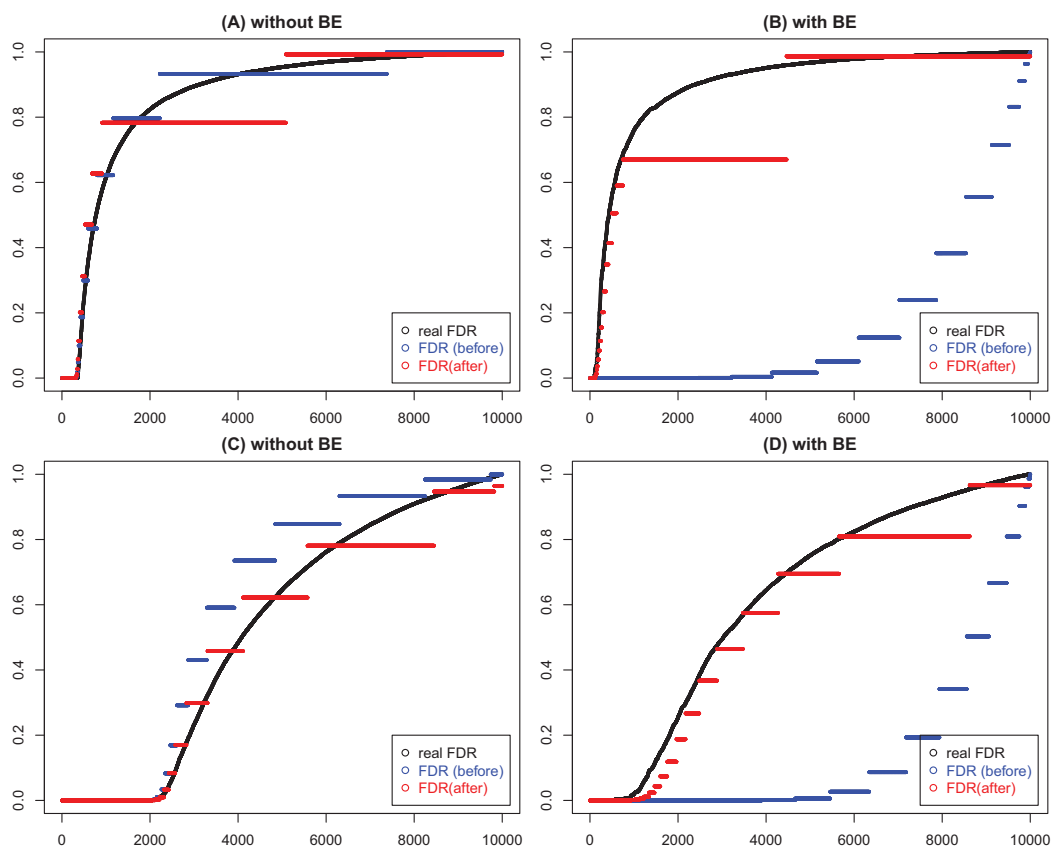


Figure 3.5: FDR comparison for simulation data sets MA, MB, MC and MD. Black points indicate the real FDR. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation.

Chapter 4

Applications of ConReg-R and iPLR in Systems Biology

In this chapter, we present the analysis of five high-throughput biological datasets (Table 4.1) using ConReg-R and iPLR methods. The datasets were obtained from different technologies and from different species. The analysis demonstrates the efficacy and usefulness of ConReg-R and iPLR in systems biology.

4.1 Yeast environmental response data

Yeast environmental stress response gene expression data generated by (DeRisi *et al.*, 1997; Gasch *et al.*, 2000) for nearly 6000 genes of yeast (*S. cerevisiae*) was aimed at understanding how yeast adopts or reacts to various stresses present in its

Table 4.1: List of datasets used for ConReg-R and iPLR application.

Dataset	platform	Method	Reference
Yeast environmental response data	DNA microarray	ConReg-R	(DeRisi <i>et al.</i> , 1997; Gasch <i>et al.</i> , 2000)
Human RNA-seq data	RNA-seq and DNA microarray	ConReg-R	(Marioni <i>et al.</i> , 2008)
Fission yeast data	DNA microarray	iPLR	(Chu <i>et al.</i> , 2007a)
Human Ewing tumor data	DNA microarray	iPLR	(Stegmaier <i>et al.</i> , 2007)
Data for type2 diabetes	cDNA microarray	iPLR	NIPER, India (Unpublished)

environment. We selected 10 datasets: (1) Heat shock from $25^{\circ}C$ to $37^{\circ}C$ response; (2) Hydrogen peroxide treatment; (3) Menadione exposure; (4) DTT exposure response; (5) Diamide treatment response; (6) Hyper-osmotic shock response; (7) Nitrogen source depletion; (8) Diauxic shift study; and, (9-10) two nearly identical experiments on stationary phase. We used Limma (Linear Models for Microarray Data) (Smyth, 2004) package in R to compute p -values for responsiveness of genes in each dataset.

The p -value distribution for each dataset is shown in Figure 4.1. As can be seen in Figure 4.1, the majority of the p -value histograms do not have similar frequency after $p = 0.5$, and the density near $p = 1$ is less than $\pi_0 = 0.5$. This implies that the p -values were under-measured and the number of significantly responsive genes under these environmental stresses should be less than observed. We applied ConReg-R on the p -values of each dataset. Our result shows that the histograms of recalibrated p -values obtained by applying ConReg-R are better than without recalibration, and π_0 estimations are all above 0.5 (Figure 4.1).

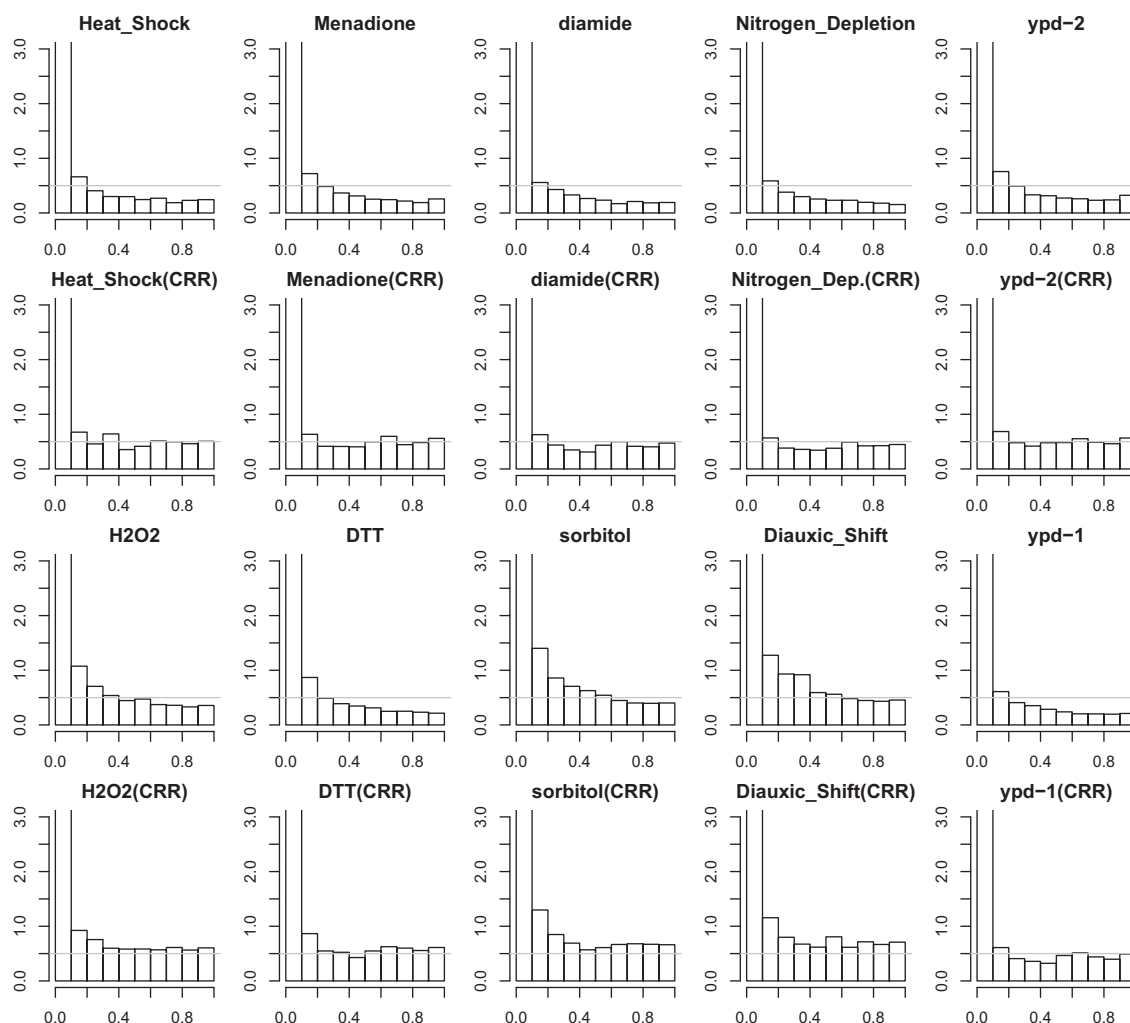


Figure 4.1: p -value density histograms for 10 stress response data sets. (CRR) indicates the re-estimated p -values after ConReg-R. The gray horizontal line indicates $\pi_0 = 0.5$ for each plot.

We use a true positive set of 270 genes from (Chen *et al.*, 2003) to compute true FDR (FDR^r). This is the intersection of core environmental stress response genes obtained by co-regulation study in (Gasch *et al.*, 2000) and the yeast orthologs of *S. pombe* stress response genes. These 270 genes have been used as the true

positive sets in other studies (Han *et al.*, 2004; Li *et al.*, 2007). The true FDR is calculated based on this 270 gene list and we calculated the improvement of FDR estimation (FDR_{im}) for each dataset after applying ConReg-R. The FDR_{im} is defined as followed:

$$FDR_{im} = \frac{\sum_{i=1}^n |FDR_i^0 - FDR_i^r| - \sum_{i=1}^n |FDR_i^1 - FDR_i^r|}{\sum_{i=1}^n |FDR_i^0 - FDR_i^r|}$$

where FDR_i^1 (respectively, FDR_i^0) is the estimated FDR by recalibration (respectively, input) p -values for gene i ($i = 1 \dots n$); and FDR_i^r is the true FDR for gene i .

The improvements in FDR estimation for all 10 datasets are shown in Figure 4.2. After applying ConReg-R, FDR estimation improved by 15% to 25% which means that the FDR estimation will be closer to the real FDR.

We performed the meta-analysis of 10 datasets to detect the core environmental stress response genes using “maximal” method. The combined p -values are computed by the maximal p -values across 10 datasets, and then transferred to meta analysis p -values by transformation function in Table 2.1. The p -value density histograms for meta-analysis before and after applying ConReg-R are shown in Figure 4.3. The meta-analysis p -values show better distribution after first applying ConReg-R to each dataset and then perform the meta analysis. Moreover, FDR estimation improved by 38.5% after applying ConReg-R.

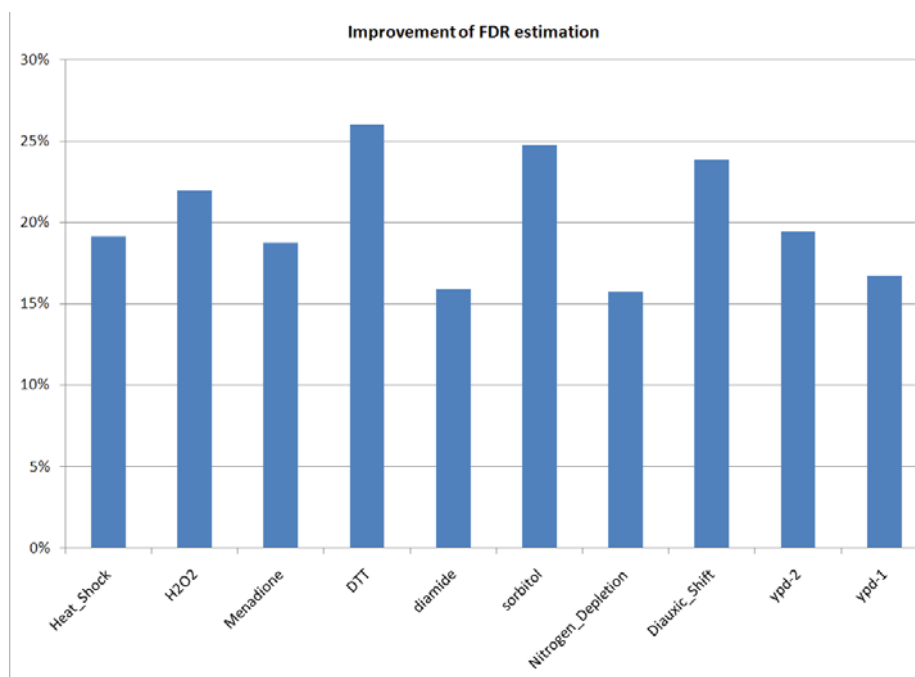


Figure 4.2: Improvements in FDR estimation for yeast environmental response datasets.

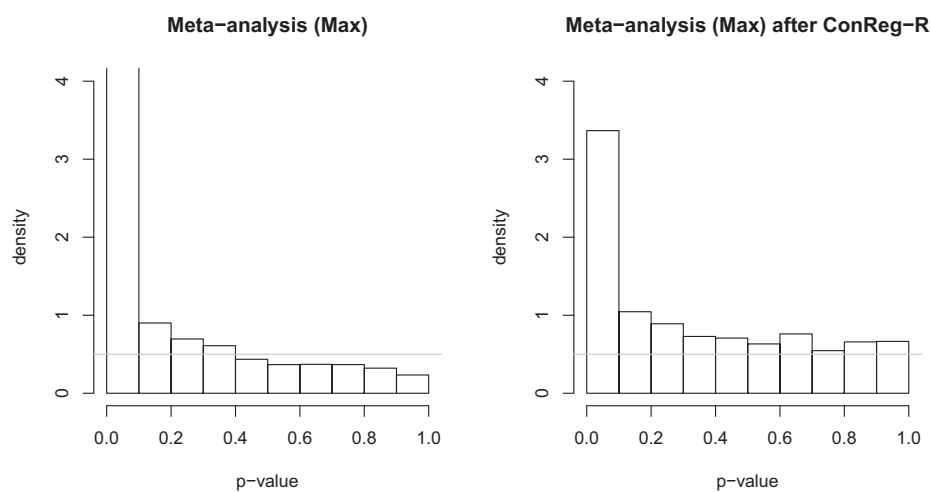


Figure 4.3: The p -value density histograms for meta-analysis (“Max”) before and after applying ConReg-R using yeast environmental response datasets. The gray horizontal line indicates $\pi_0 = 0.5$ for each plot.

4.2 Human RNA-seq data

The next-generation sequencing technologies have been used for gene expression measurement. In (Marioni *et al.*, 2008), the authors compared RNA-seq and Affymetrix microarray experiments and claimed that the sequencing data identified many more differentially expressed genes between human kidney and liver tissue samples than microarray data using the same FDR cutoff. In total, 11,493 significant genes were identified by RNA-seq (3380 more genes than Affymetrix), only 6534 (56.9%) genes were also identified by Affymetrix experiments. Upon checking the p -value histograms for RNA-seq dataset, we found that majority of p -values are very significant and its frequencies are very non-uniform for $p > 0.5$. However, the p -value histogram for Affymetrix datasets is close to uniform for $p > 0.5$ (Figure 4.4).

We applied ConReg-R to recalibrate the p -values obtained from RNA-seq datasets and re-estimated the FDR. We found 9481 significantly differentially expressed genes (only 1368 more genes than affymetrix) at $FDR \leq 0.1\%$. Among them, 6266 genes (66.1%) were also identified by Affymetrix experiments. There is an increase of 9.2% overlap after application of ConReg-R (Figure 4.5). The FDR estimation is improved by 20% after applying ConReg-R if we used significant genes identified by affymetrix experiments as the true positive set.

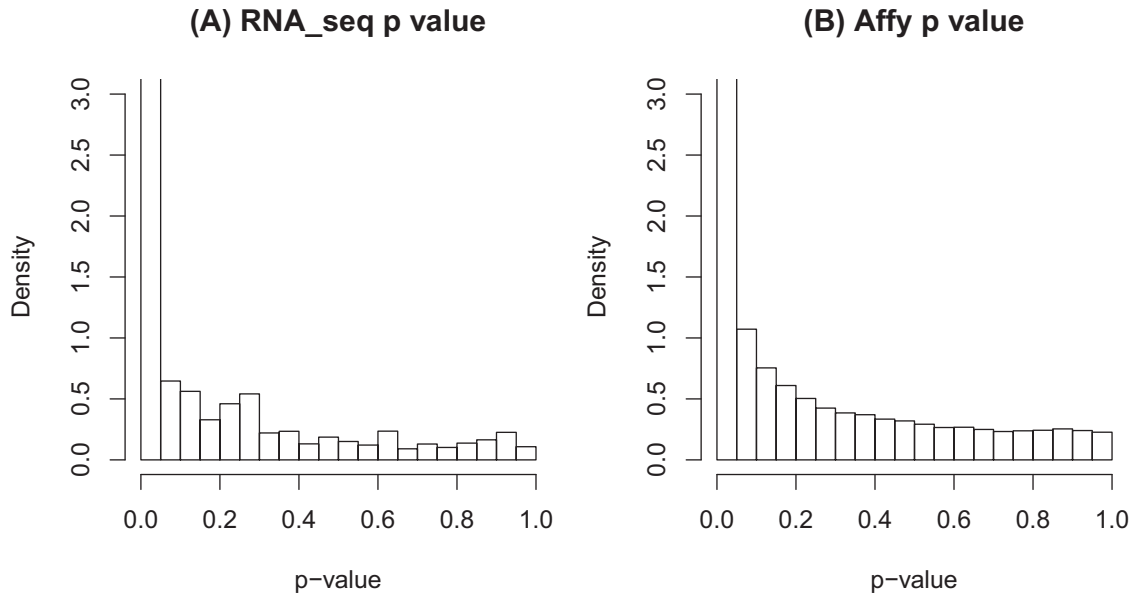


Figure 4.4: p -value density histograms for RNA-seq and Affymetrix datasets.

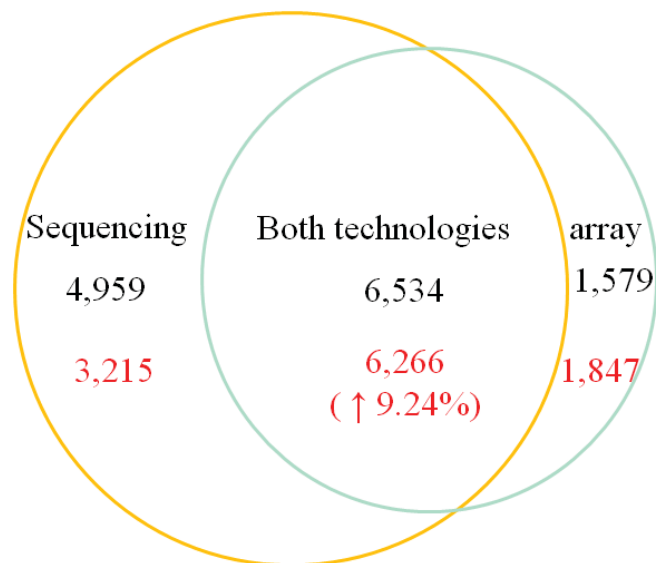


Figure 4.5: Overlap between significantly differentially expressed genes identified by sequencing (left circle) and microarray (right circle) technologies. The numbers in black are the numbers reported in (Marioni *et al.*, 2008). The numbers in red are the numbers after applying ConReg-R.

4.3 Fission yeast data

Having shown the efficacy of iPLR re-estimation on simulated data, we next demonstrate the utility of our iPLR re-estimation on real gene expression data, *mip1* mutant ($\Delta mip1$) differential expression in *S. pombe* (or fission yeast) compared to its wild-type. The data was obtained from (Chu *et al.*, 2007a) containing 28 wt/wt spotted two-color array data and 6 $\Delta mip1$ /wt data for ~ 5000 open reading frames (ORFs). The purpose is to find the genes influenced by *mip1* mutation ($\Delta mip1$). The data have been global and local normalized. The wt/wt data contains two batches of equal number of arrays which we call wt1/wt1 (or wt-rep1) and wt2/wt2 (or wt-rep2). The application of SAM on wt1/wt1 vs. wt2/wt2 data are shown in Table 4.2. It estimates π_0 to be 0.255, while it should be 1 as both wt1 and wt2 samples are the same except that they were hybridized into two different batches of arrays at two different times. The corresponding SAM plot is shown in Figure 4.6(A1). After application of our iPLR procedure, the estimate of π_0 is improved to 0.997. The respective SAM plot was shown in Figure 4.6(A2).

We analyzed wt/wt (combining wt1/wt1 and wt2/wt2) versus $\Delta mip1$ /wt using SAM to identify differentially expressed genes, the results are shown in Table 4.2 and the corresponding SAM plot is shown in Figure 4.6(B1-2). $\Delta mip1$ /wt was hybridized altogether on a different batch of arrays at completely different time. This resulted in batch effects again and the underestimation of π_0 (0.22). FDR estimates and results are shown in Table 4.2 and Figure 4.6(B1). We applied our

Table 4.2: Significant gene tables for yeast datasets.

delta	#sig.genes	$\widehat{\text{FDR}}$	#sig.genes	$\widehat{\text{FDR}}$
wt vs. wt	SAM ($\hat{\pi}_0 = 0.2548$)		SAM+iPLR ($\hat{\pi}_0 = 0.9972$)	
0.5	4092	0.099	550	0.5657
1	3554	0.0369	146	0.3449
2	2241	0.0007	13	0.0767
3	1099	0	10	0
wt vs. Δmip1	SAM ($\hat{\pi}_0 = 0.2194$)		SAM+iPLR ($\hat{\pi}_0 = 0.9744$)	
0.5	3984	0.0426	813	0.8258
1	3116	0.0001	428	0.307
2	1918	0	185	0.0225
3	934	0	113	0

iPLR procedure on this dataset. The corresponding estimates of π_0 and FDR are shown in Table 4.2 and Figure 4.6(B2). π_0 estimate is closer to 1 (0.974) than the otherwise unrealistic estimate (0.22) by SAM alone. Results of comparing estimated FDR before and after iPLR are shown in Figures 4.6(A3) and 4.6(B3). As shown in these figures, FDR estimation before iPLR is extremely low for most genes, but after iPLR procedure, they are closer to what are expected. This results showed that iPLR is a practically useful technique.

4.4 Human Ewing tumor data

Another dataset we analyzed is human Ewing tumor data from (Stegmaier *et al.*, 2007). It is affymetrix microarray data of A673 cells treated with DMSO vehicle control expression profiled at 24 hours (6 replicates), 3 days (5 replicates), and 5 days (6 replicates). The data have been quantile normalized. Since the

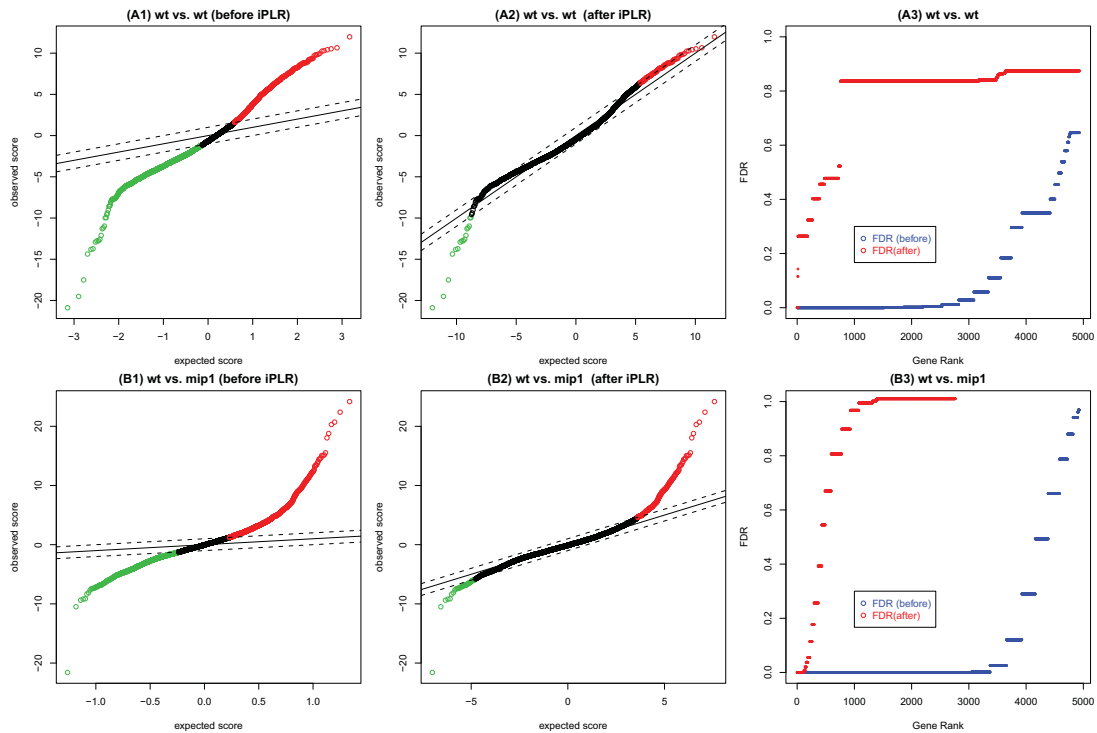


Figure 4.6: SAM plot and FDR comparison (before and after iPLR re-estimation) for *S. pombe* data set. (A1) The SAM plot before iPLR re-estimation for wt1/wt1 vs. wt2/wt2 dataset. (A2) The SAM plot after re-estimation for wt1/wt1 vs. wt2/wt2 dataset. (A3) FDR comparison for wt1/wt1 vs. wt2/wt2 dataset. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation. (B1) The SAM plot before re-estimation for wt/wt vs. Δ mip1/wt dataset. (B2) The SAM plot after re-estimation for wt/wt vs. Δ mip1/wt dataset. (B3) FDR comparison for wt/wt vs. Δ mip1/wt dataset. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation. The results are encouraging and iPLR is a practically useful technique.

experiments were performed on three different days, it is unrealistic to assume no batch confounding effect. In fact, from the array clustering result in Figure 4.7, the data from 3 different days are well separated by the day of the sample. It suggests that biological effect and batch effect are confounded.

First we analyzed the datasets from 24 hours and 3 days using SAM alone, and

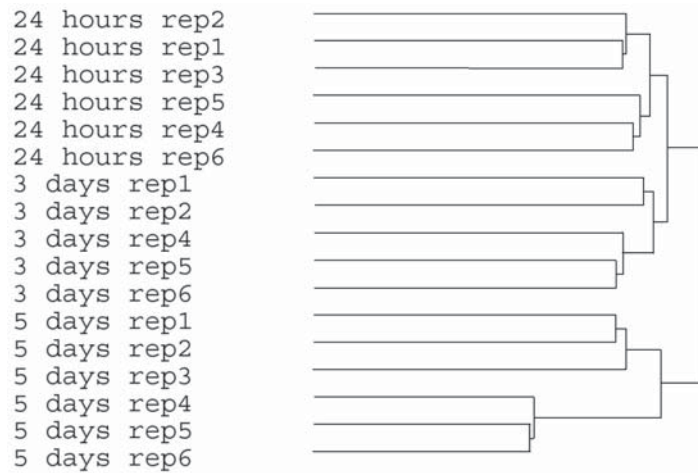


Figure 4.7: Clustering of all arrays from Ewing *et al.* data using all the genes.

using iPLR combined with SAM. The result is shown in Table 4.3. SAM estimated π_0 to be 0.68, while we expect a higher π_0 as the vehicle control data between two days should not be very different. After application of iPLR, the estimation of π_0 is 0.84 closer to what is expected. We repeated the same procedure to compare 24 hours vs. 5 days, and obtained similar result (Table 4.3). The estimation of π_0 is 0.81 and it is less than the π_0 estimated in comparison of 24 hours vs. 3 days. This is to be expected since there should be more differently expressed genes in 5 days versus 24 hours than 3 days versus 24 hours. The SAM plots for these two comparisons and the comparisons of estimated FDR before and after applying iPLR are shown in Figures 4.8 (A1-3) and (B1-3).

We also compared these three groups: 24 hours, 3 days and 5 days. Results are shown in Table 4.3. It is seen that FDR is improved after applying iPLR and estimation of π_0 is closer to real π_0 . The SAM plots and the comparisons of estimated FDR before and after applying iPLR are shown in Figure 4.8 (C1-3).

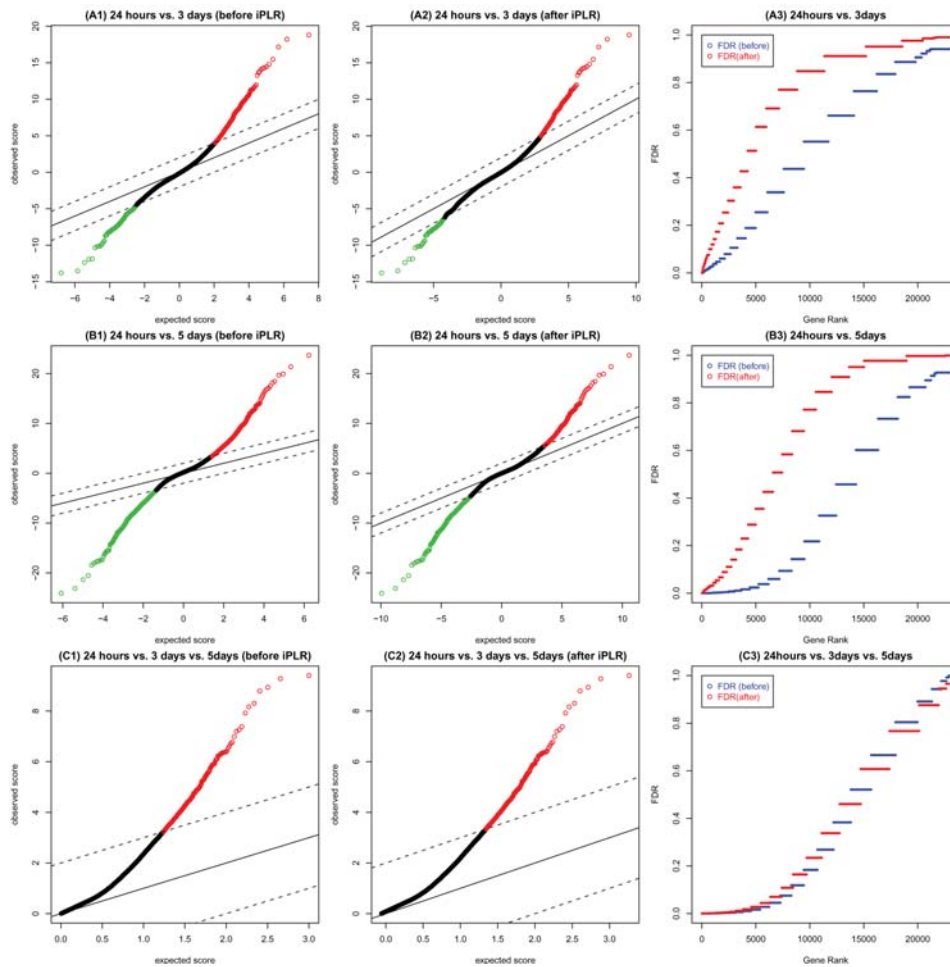


Figure 4.8: SAM plots and FDR comparison (before and after iPLR re-estimation) for human Ewing tumor data set. (A1) The SAM plot before iPLR re-estimation for 24 hours vs. 3 days dataset. (A2) The SAM plot after re-estimation for 24 hours vs. 3 days dataset. (A3) FDR comparison for 24 hours vs. 3 days dataset and 24 hours vs. 5 days dataset. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation. (B1) The SAM plot before re-estimation for 24 hours vs. 5 days dataset. (B2) The SAM plot after re-estimation for 24 hours vs. 5 days dataset. (B3) FDR comparison for 24 hours vs. 3 days dataset and 24 hours vs. 5 days dataset. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation. (C1) The SAM plot before iPLR re-estimation for 24 hours vs. 3 days vs. 5 days dataset. (C2) The SAM plot after re-estimation for 24 hours vs. 3 days vs. 5 days dataset. (C3) FDR comparison for simulation 24 hours vs. 3 days vs. 5 days dataset. Blue points indicate estimated FDR before iPLR re-estimation and the red points indicate estimated FDR after iPLR re-estimation.

Table 4.3: Significant gene tables for human Ewing tumor datasets.

delta	#sig.genes	$\widehat{\text{FDR}}$	#sig.genes	$\widehat{\text{FDR}}$
24H vs. 3D	SAM ($\hat{\pi}_0 = 0.6844$)		SAM+iPLR ($\hat{\pi}_0 = 0.8416$)	
0.5	8905	0.2789	4091	0.3801
1	3822	0.0865	1675	0.1706
2	1024	0.0167	435	0.0474
3	368	0.0065	168	0.0251
24H vs. 5D	SAM ($\hat{\pi}_0 = 0.6052$)		SAM+iPLR ($\hat{\pi}_0 = 0.8063$)	
0.5	8905	0.2789	4091	0.3801
1	3822	0.0865	1675	0.1706
2	1024	0.0167	435	0.0474
3	368	0.0065	168	0.0251
24H vs. 3D vs. 5D	SAM ($\hat{\pi}_0 = 0.5102$)		SAM+iPLR ($\hat{\pi}_0 = 0.6372$)	
0.25	13873	0.2319	14101	0.3348
0.5	8421	0.0537	8131	0.0921
1	3498	0.0037	3144	0.0083
2	605	0.0000	494	0.0000
24H: 24 hours; 3D: 3 days; 5D: 5 days.				

4.5 Integrating analysis in type2 diabetes

To understand Histone-DNA interaction mechanism in type2 diabetes, our collaborators from National Institute of Pharmaceutical Education and Research (NIPER, India) performed the H3K4/ H3K9 mono methylation experiments with the alteration in gene expression in 3T3 adipocytes under hyperglycaemic/hyperinsulinemic conditions. The mouse 15K microarray (Microarray centre, University Health Care, Toronto) used in this study consisted of 15,264 genes spotted in duplicate. The experiments generate H3Ac (Histone H3 acetylation), H3K4me (H3 lysine 4 mono methylation), H3K9me (H3 lysine 9 mono methylation) ChIP-chip data and 30min gene expression data (each experiment have 3 biological replicates and each

replicate have 2 technical replicates due to the duplicate probes in one array). This array confounding effect which is similar to batch confounding effect occurred in this study mainly because of the array design.

Since this array is the cDNA microarray, the signal (intensity) is weaker than that from DNA microarray. Therefore, we add a small positive value to each channel (CH5 & CH3) to achieve more stable data. The procedure is described as followed,

$$X_{ij} = \text{Log}_2 \frac{I_{ij}^{C5} + c_j}{I_{ij}^{C3} + c_j},$$

where X_{ij} is the gene expression for gene i in array j and I_{ij}^{C5} and I_{ij}^{C3} are the intensity in CH5 and CH3 for gene i in array j . $c_j \leq 100$ is the predefined positive value for array j . We can chose c_j by maximizing the Pearson correlation coefficient between the duplicate in array j . Since the median intensity of microarray is around few thousands, the gene expression ratio do not change so much for the majority of genes. It will reduce the variation for low intensity genes (the intensity below 1000). We performed the LOWESS normalization for each array.

The SAM plot for 30min gene expression data is shown in Figure 4.9 (A). The curve of expected score vs. observed score is all below the diagonal which may be naively interpreted as that only down-regulated genes are identified and no up-regulated genes. That result grossly deviated from our biological knowledge and array confounding effect play a major role to generate this unexpected result. Therefore, we performed iPLR to re-estimate the expected statistics, and the SAM

plot after iPLR re-estimation is shown in Figure 4.9 (B). As shown in this figure, we can obtain the up-regulated genes and down-regulated genes. There are total 1536 genes which are differentially expressed with at least 1.5 fold difference and $FDR < 10\%$.

To get further insight into the level of H3K9me, H3K4me and H3Ac across the coding regions of the mouse genome, we performed ChIP-cDNA analysis using 15K cDNA array after 30 minutes of the insulin stimulation under the high glucose condition. Using same procedure, SAM analysis with iPLR re-estimation, we identified 844 targets for H3Ac, 215 targets for H3K4me and 999 targets for H3K9me with differential status in high glucose as compared to no glucose condition in coding regions of the genes.

To understand the role of these histone H3 modifications in regulation of the genes under hyperglycaemic/hyperinsulinemic conditions, we identified the genes that underwent changes in any of these three histone modifications along with change in their gene expression levels. To do so, we set up a criterion and to select only the genes that were common in cDNA expression analysis with differential change in status in any one of either H3Ac or H3K4me or H3K9me. This stringent criterion might result in false negatives but it also reduces the number of genes to a manageable size for further validation analysis and reduces the chance of having false positives. With this criterion we identified 831 genes with significant differential H3Ac or H3K4me or H3K9me status and also change in their mRNA

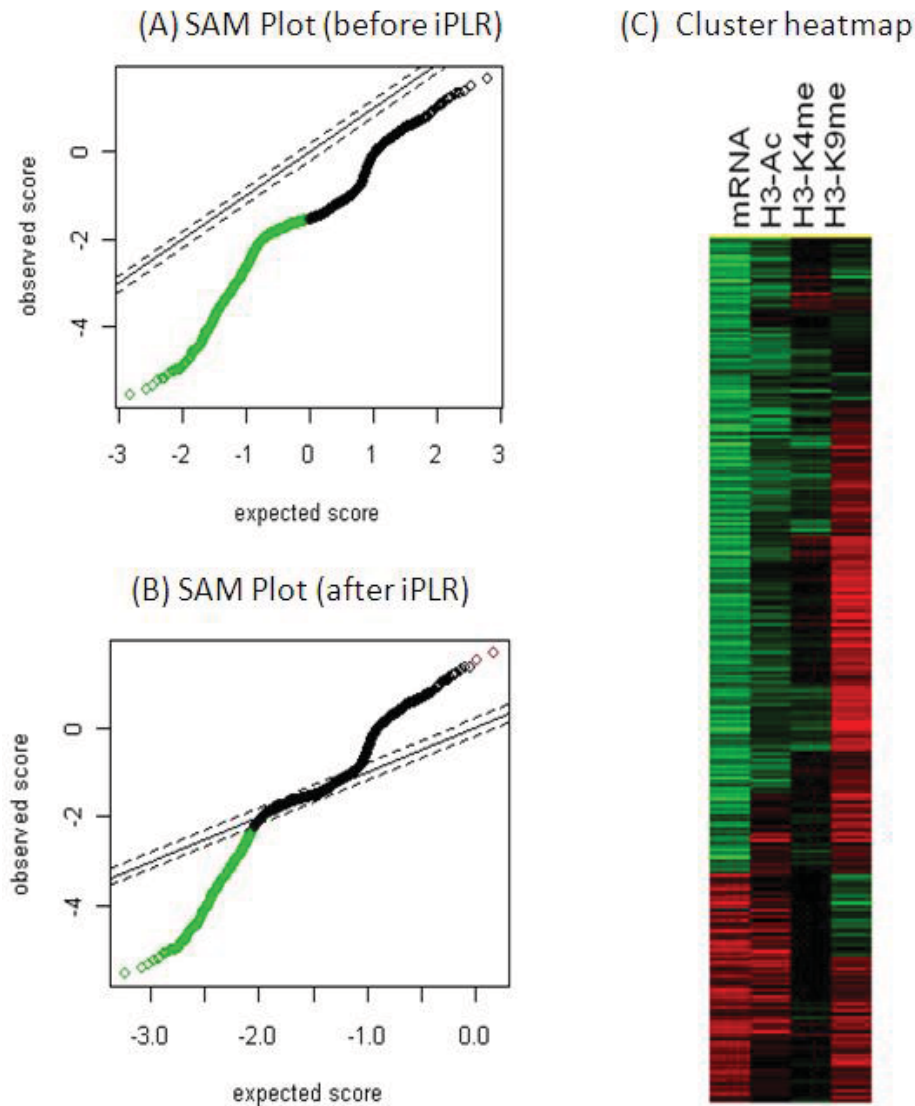


Figure 4.9: SAM plots before and after iPLR re-estimation for 30min gene expression data for type2 diabetes and integrating cluster heat map for gene expression and histone marks. (A) The SAM plot before iPLR re-estimation for 30min gene expression data. (B) The SAM plot after iPLR re-estimation for 30min gene expression data. (C) Hierarchical cluster analysis of mRNA, H3Ac, H3K4me and H3K9me profiles on coding regions of genes altered by the insulin (100 nM) stimulation under high glucose as compared to low glucose conditions.

expression levels. Of these, 608 genes were down regulated and 223 genes were up regulated. The integrating cluster heat map for gene expression and histone

marks for these 831 genes is shown in Figure 4.9 (C).

With this analysis we demonstrated that histone H3Ac levels in the coding regions of the genes very well correlates with the mRNA expression level of the respective genes signifying H3Ac as a mark of gene activation even in the coding regions of the genes. Furthermore, mRNA expression of most of the genes were inversely proportional to H3K9me levels, suggesting that increased H3K9me occupancy in the coding regions of the genes is associated with gene inactivation. However, very few genes are enriched for H3K4me in the coding regions and we also failed to observe much overlap between H3K4me and mRNA expression levels (4.9 (C)). This indicates that the genes with increased occupancy of H3Ac and H3K9me in the coding region are not enriched for H3K4me.

Out of differentially expressed genes identified by cDNA microarray and ChIP-chip analysis, we observed significant change in the expression of 9 genes that are responsible for mediating chromatin remodeling by insulin under high glucose condition. These include down regulation of *Myst4* and *Ep400* (histone acetyl transferases, HAT), *Jmjd2b* and *Jarid2* (histone methyl transferases, HMT) and *Dyrk2* (histone kinase). In addition to the above mentioned genes, *Brd4* gene which is involved in reorganization of acetylated chromatin was also found to be down regulated. Increase in the expression of *Set* gene (HAT inhibitor) and also genes responsible for histone H3K4 demethylation (*Jarid1a* and *Aof1*) further supports our earlier observation. The change in expression of these genes observed in the

present study was in accordance with our previous findings that shows decrease in levels of H3Ac, H3K4me and H3K9me after 30 minutes of insulin stimulation under high glucose condition (Kabra *et al.*, 2009).

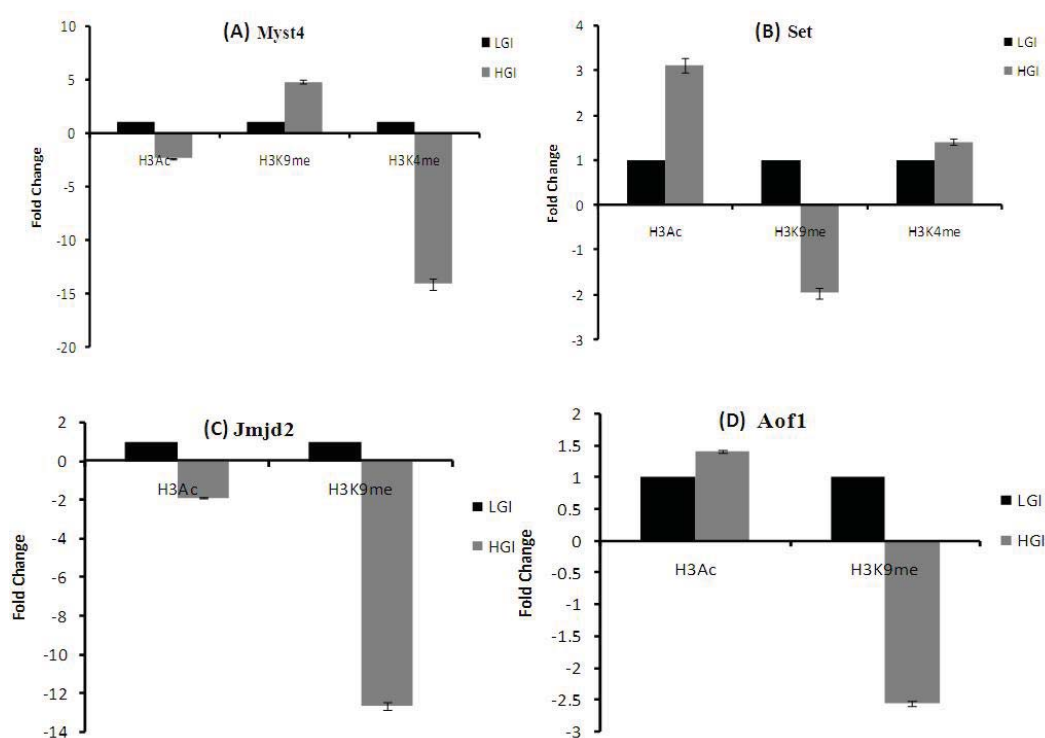


Figure 4.10: RT-PCR validation on Histone H3 acetylation, lysine 4 mono methylation and lysine 9 mono methylation levels on coding regions of the chromatin modification regulating genes. (A) H3Ac, H3K4me and H3K9me levels on *Myst4*; (B) H3Ac, H3K4me and H3K9me levels on *Set*; (C) H3Ac and H3K4me levels on *Jmjd2b* and (D) H3Ac and H3K4me levels on *Aof1*. Relative fold change was calculated after normalization with input. Similar results were obtained in the three independent sets of experiments. All the values were represented as Mean \pm S.E.M. ($n=3$), $***p < 0.001$, $**p < 0.01$ and $*p < 0.05$, Vs LGI.

Further, we selected 4 chromatin remodeling genes, *Myst4*, *Jmjd2b*, *Set* and *Aof1* and confirmed the change in H3Ac, H3K4me and H3K9me levels on their coding regions by performing ChIP-RT-PCR analysis (Figure 4.10). We observed a decrease in the level of H3Ac on *Myst4* and *Jmjd2b* and an increase on *Set* and *Aof1*

genes confirming our ChIP-chip data. However, we failed to observe any change in H3K9me levels on the coding regions of histone H3K9 demethylase (Jmjd2b) and H3K4 demethylase (Aof1). Decreased H3K4me levels on *Myst4* and *Jmjd2b* and increased H3K4me levels on *Set* and *Aof1* further confirmed our ChIP-chip analysis. These results suggest a novel mechanism of regulating the level of H3Ac and H3K4me by each other under hyperinsulinemic/hyperglycemic conditions. However, levels of H3K9me were only changed on histone acetylase (*Myst4*) and deacetylase (*Set*), highlighting the role of this modification in regulating histone acetylation only.

Chapter 5

Conclusions and future works

In this chapter, we first summarize the two methods presented in the thesis and then discuss their limitations and potential directions of future work.

5.1 Conclusions

In the first method, to eliminate the dependency effect in microarray studies, we developed *Constrained Regression Recalibration* (ConReg-R) which focuses on the uniformity of p -values under null hypotheses and uses constrained polynomial regression to recalibrate the empirical p -value distribution to more well-defined p -value distribution. Therefore, the FDR estimation can be improved after the recalibration since the assumption of FDR estimation is that the input p -values should follow such an ideal empirical p -value distribution under null hypothesis.

If the input p -values follow the properties of ideal empirical p -values distribution, the regression function tends to be diagonal line (i.e., $y = x$) and the p -values do not change considerably after recalibration.

Though our method is discussed in the context of global FDR control, it is equally applicable to the other FDR like controls such as local FDR. Our method does not provide any new FDR control, but inputs better calibrated p -values to the existing FDR estimators to improve their efficacy.

In the second method, to remove the batch confounding effect in microarray studies, we proposed *iterative piecewise linear regression* (iPLR) to correct the bias introduced in the estimation of null distribution when experimental batches are confounded with treatment groups of interest. In FDR estimation, this correction is critical in gene expression studies where one wants to compare data obtained from different laboratories or from the same laboratory but collected at different times. Our results on the real data, which was preprocessed and normalized appropriately, demonstrated that the effect of batch confounding continues to exist in the normalized data also and leads to erroneous FDR estimation. iPLR plays an important role in such a case, it works at the downstream of a resampling based method such as SAM. In iPLR, we assume that batch effects are small and influences all spots on the array in unexpected but definite manner which varies from batch to batch. Under this assumption which was used in the popularly used location/scale model for batch effects (Johnson *et al.*, 2007), the influence is mainly

on the estimation of FDR via badly estimated null distribution, underestimated proportion of non-differentially expressed genes and by the inevitable influence of change of mean value on permutation procedure. The SAM manual cites this behavior as one that could be biologically more meaningful to be left to the biologists to decide. When it is reasonable to assume in gene expression studies that π_0 is more than 0.5, and under realistic assumptions of low batch effects, we proposed iPLR method to resolve this problem. iPLR procedure is equally applicable to any differential expression analysis procedure for any number of classes. It is only for the sake of simplicity in describing our methodology and evaluating the results in the context of SAM (a widely used method for differential expression analysis).

Similar problem has been addressed in the evaluation of enrichment of gene sets in a list of genes (Efron and Tibshirani, 2007), the GSA (Gene Set Analysis) algorithm. GSA handles the problem by making the mean and standard deviations of the distributions of both observed statistics and permutation statistics to be the same. The idea is simple and effective for GSA because π_0 in GSA is generally close to 1. However, it may not work well in several gene expression studies if π_0 is well below 1. This may lead to severe overestimation of standard deviation and make the idea ineffective for this purpose. Hence, iPLR may play an important contribution.

We have shown the efficacy of our iPLR method on both simulated and real data. These results demonstrate that iPLR combined with SAM is robust to batch

confounding effects of treatments. Results in Table 3.3 suggest that iPLR improves the estimate of π_0 to some extent than using SAM alone even in the absence of batch confounding effects. More extensive experiments will be conducted in the future to verify this hypothesis. Furthermore, there is still room to improve iPLR. As shown in Figure 3.4, re-estimated FDR deviates considerably from real FDR for dataset C. However, iPLR in its current form is still useful in making the right choice of differential expression significance threshold in the wake of better and meaningful FDR estimation.

5.2 Limitations and future works

There are several limitations and potential future works of the methods proposed in this thesis.

5.2.1 Some special p -value distributions

In most common cases, the p -values are under-estimated or over-estimated and p -value distribution is biased towards 1 or 0 respectively (e.g., Figure 1.1B & 1.1C). ConReg-R can be useful to deal with these two cases by setting the regression function is convex or concave function.

There are two special p -value distributions with mixture under-estimated or over-estimated p -values in one experiment. One is mixture of over-estimating

p -values from H1 and under-estimating p -values from H0 (Hump shape p -value distribution in Figure 5.1(A)). Another is mixture of under-estimating p -values from H1 and over-estimating p -values from H0 (U-shape p -value distribution in Figure 5.1(B)).

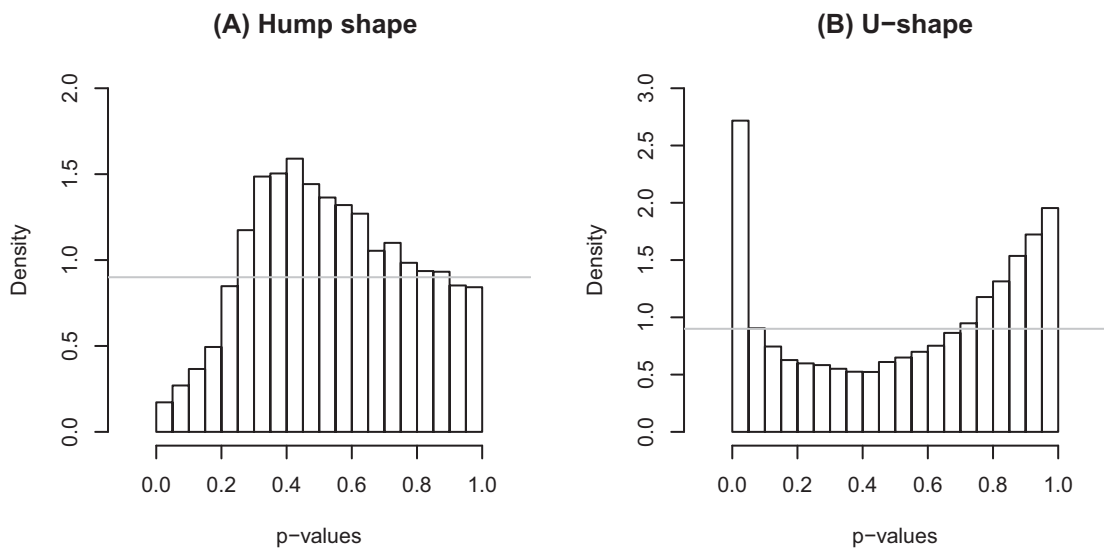


Figure 5.1: Hump shape and U-shape p -value density histograms. (A) Hump shape p -value density histogram. The gray horizontal line indicates the $\pi_0 = 0.9$. (B) U-shape p -value density histogram. The gray horizontal line indicates the $\pi_0 = 0.9$.

The regression function for hump shape p -value distribution should be convex for p -values from H1 and concave for p -values from H0. Similarly, The regression function for U shape p -value distribution should be concave for p -values from H1 and convex for p -values from H0. However, how to distinguish the p -values from H1 and H0 or define the regression function is a difficult problem. This may be

one potential future work.

5.2.2 Parametric recalibration method

The distribution of p -values from microarray experiment can be modeled by beta-uniform mixture (BUM) distribution (Pounds and Morris, 2003). The probability density function for BUM distribution is

$$f(x|a, \pi_0) = \pi_0 + (1 - \pi_0)ax^{a-1},$$

where $0 < x \leq 1$, $0 < \pi_0 < 1$ and $0 < a < 1$. Therefore, the parametric recalibration method similar to ConReg-R can be developed. Though this procedure, the estimation of π_0 and a can be obtain by inputting any kind of p -value distribution. The false discover rate can be estimate by BUM distribution.

To more accurately estimate p -value distribution, we can use mixture of more than 2 beta distributions to model the p -value distribution (uniform distribution is the special case of beta distribution) (Allison *et al.*, 2002). It is sufficient to estimate all the parameters for multiple mixture beta distribution if we have large number of p -values.

5.2.3 Discrete p -values

ConReg-R is only applicable for continues p -values from parametric test. If the p -values from permutation or non-parametric test and the sample size is relatively

small, the uniformity property of p -values may not fit well. For example, the p -values from Wilcoxon tests for sample size = 3, 5, 10, 50 in Figure 5.2.

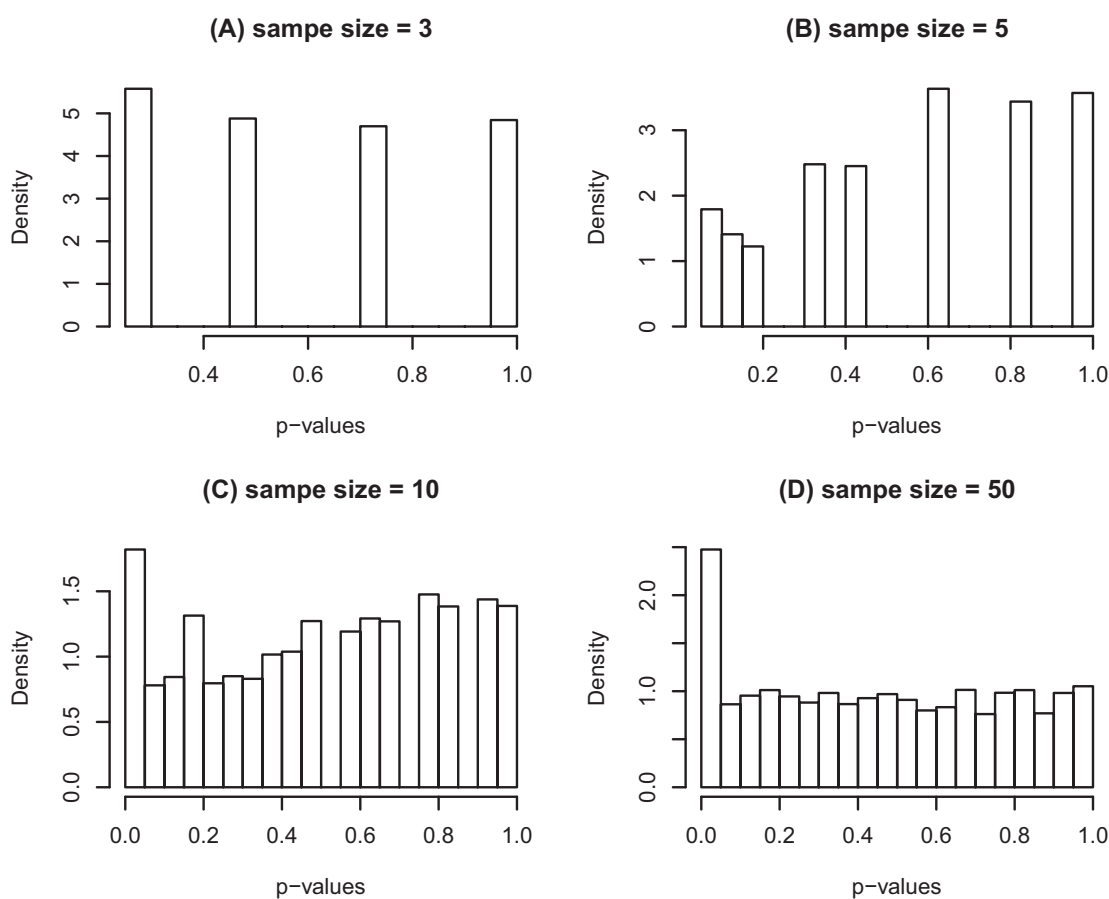


Figure 5.2: p -value Density histograms from Wilcoxon test for various sample sizes (3, 5, 10, 50).

The regression function in ConReg-R cannot be estimated by the discrete p -values with small sample size because those p -values are distributed within few blocks (Figure 5.2(A-C)). If the sample size is very large (Figure 5.2(D)), ConReg-R may still work well. The new procedure to handle discrete p -values is one of

potential future work.

5.2.4 π_0 estimation for ConReg-R and iPLR

π_0 estimation is very important in multiple testing problem. In ConReg-R, we used (2.9) to estimate π_0 . And in iPLR, we used iterative approach to estimate π_0 . However, our goal in this thesis is to improve FDR estimation. Next natural goal will be better π_0 estimation. Validation of π_0 estimation in those two procedures and more comprehensive comparisons to other exiting π_0 estimation methods will be explored. Another potential future work is that whether the π_0 estimation can be improved if we consider the input raw data.

5.2.5 Other regression functions for iPLR

In iPLR, we simply approximate the biological effect factor c_i to be linearly related with the permutation statistics \bar{d}_i when $c_i \neq 0$. However, to more accurately fit the iPLR model, other regression functions can be apply, such as, 2 or 3 degree polynomial function.

Bibliography

- Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39(1):1–20.
- Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA*, 97(18):10101–10106.
- Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and Zhang, W. (2001). Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology*, 8(6):639–659.
- Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519.
- Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society Series B (Methodological)*, 57:289–300.

- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188.
- Bishop, E. (1961). A generalization of the stone-weierstrass theorem. *Pacific Journal of Mathematics*, 11(3):777–783.
- Broet, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, 9(4):671.
- Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury Press, 2 edition.
- Celisse, A. and Robin, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132–3147.
- Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N., and Bähler, J. (2003). Global transcriptional responses of fission yeast to environmental stress. *Molecular Biology of the Cell*, 14(1):214–229.
- Chu, G., Narasimhan, B., Tibshirani, R., and Tusher, V. SAM, “significance analysis of microarrays”. *Users guide and technical document*.
- Chu, Z., Li, J., Eshaghi, M., Karuturi, R. K. M., Lin, K., and Liu, J. (2007a). Adaptive expression responses in the pol-gamma null strain of *S. pombe* depleted of mitochondrial genome. *BMC Genomics*, 8:323.

- Chu, Z., Li, J., Eshaghi, M., Peng, X., Karuturi, R. K. M., and Liu, J. (2007b). Modulation of cell cycle-specific gene expressions at the onset of s phase arrest contributes to the robust DNA replication checkpoint response in fission yeast. *Molecular Biology of the Cell*, 18(5):1756–1767.
- Dalmasso, C., Bar-Hen, A. and Broet, P. (2007). A constrained polynomial regression procedure for estimating the local False Discovery Rate. *BMC Bioinformatics*, 8:229.
- DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686.
- Diao, Q., Hu, W., Zhong, H., Li, J., Xue, F., Wang, T., and Zhang, Y. (2004). Disease gene explorer: Display disease gene dependency by combining bayesian networks with clustering. In *Computational Systems Bioinformatics Conference*, 574–575.
- Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. part I. single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):article13.
- Do, K., Muller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society. Series C*, 54(3):627.
- Edwards, J.W., Page, G.P., Gadbury, G., Heo, M., Kayo, T., Weindruch, R., and

- Allison, D.B. (2006). Empirical Bayes estimation of gene-specific effects in micro-array research. In *Functional and Integrative Genomics*, 6(3):261.
- Efron, B. (2007). Correlation and Large-Scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103.
- Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129.
- Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.
- Eshaghi, M., Karuturi, R. K. M., Li, J., Chu, Z., Liu, E. T., and Liu, J. (2007). Global profiling of DNA replication timing and efficiency reveals that efficient replication/firing occurs late during s-phase in *s. pombe*. *PloS One*, 2(1):e722.
- Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y., and Wang, Y. (2003). Effects of atmospheric ozone on microarray data quality. *Analytical Chemistry*, 75:4672–4675.
- Fisher, R. A. (1948). Combining independent tests of significance. *American Statistician*, 2(5):30.
- Fox, R. J. and Dimmic, M. W. (2006). A two-sample bayesian t-test for microarray data. *BMC Bioinformatics*, 7:126.

- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257.
- Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33.
- Guedj, M., Robin, S., Celisse, A. and Nuel, G. (2009). Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10:84.
- Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P., and Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, 22(2):164–167.
- Han, X., Sung, W., and Feng, L. (2007). Identifying differentially expressed genes in time-course microarray experiment without replicate. *Journal of Bioinformatics and Computational Biology*, 05(02a):281.
- Hedges, L. V. and Olkin, I. (1985). Test of statistical significance of combined results. In *Statistical methods for meta-analysis*, pages 28–46. Academic Press, 6th edition.

- Jafari, P. and Azuaje, F. (2006). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, 6:27.
- Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8:118–127.
- Kabra, D. G., Gupta, J., and Tikoo, K. (2009). Insulin induced alteration in post-translational modifications of histone h3 under a hyperglycemic condition in l6 skeletal muscle myoblasts. *Biochimica Et Biophysica Acta*, 1792(6):574–583. PMID: 19327396.
- Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837.
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.
- Lander, E. S. (1999). Array of hope. *Nature Genetics*, 21(1 Suppl):3–4.

- Lehmann, E. and Romano, J. P. (2005). p -values. In *Testing Statistical Hypotheses*, pages 63–65. Springer, New York, 3rd edition.
- Li, C. and Wong, W. H. (2003). DNA-chip analyzer (dchip). In *The Analysis of Gene Expression Data: Methods and Software*, pages 28–46. Springer.
- Li, J., Choi, K. P., and Karuturi, R. K. M. (2012a). Iterative piecewise linear regression to accurately assess statistical significance in batch confounded differential expression analysis. In *Proceedings of the 8th international conference on Bioinformatics research and applications*.
- Li, J., Choi, K. P., Pawitan, Y., and Karuturi, R. K. M. (2012b). Statistical significance assessment for biological feature selection: methods and issues. In *Handbook of Biological Knowledge Discovery*. Wiley (In press).
- Li, J., Eshaghi, M., Liu, J., and Karuturi, R. K. M. (2008a). Near-sigmoid modeling to simultaneously profile genome-wide DNA replication timing and efficiency in single DNA replication microarray studies. In *Proceedings of 6th Asia-Pacific Bioinformatics Conference*, pages 383–392.
- Li, J., Liu, J., and Karuturi, R. K. M. (2007). Data-driven smoothness enhanced variance ratio test to unearth responsive genes in 0-time normalized time-course microarray data. In *Proceedings of the 3th international conference on Bioinformatics research and applications*, pages 25–36.
- Li, J., Liu, J., and Karuturi, R. K. M. (2008b). Stepped linear regression to accu-

- rately assess statistical significance in batch confounded differential expression analysis. In *Proceedings of the 4th international conference on Bioinformatics research and applications*, pages 481–491.
- Li, J., Paramita, P., Choi, K. P., and Karuturi, R. K. M. (2011). Congreg-r: Extrapolative recalibration of the empirical distribution of p-values to improve false discovery rate estimates. *Biology Direct*, 6:27.
- Li, J., Yunus, F., Lei, Z., Eshaghi, M., Liu, J., and Karuturi, R. K. M. (2009). Modeling and visualizing heterogeneity of spatial patterns of protein-DNA interaction from high-density chromatin precipitation mapping data. In *Proceedings of the 5th international conference on Bioinformatics research and applications*, pages 236–247.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.
- Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52.
- Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O’Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., Brown, P. O., Botstein,

- D., and van de Rijn, M. (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, 359(9314):1301–1307.
- Nocedal, J. and Wright, S. (2000). *Numerical Optimization*. Springer.
- Ochs, M. F. (2010). Knowledge-based data analysis comes of age. *Briefings in Bioinformatics*, 11(1):30–39.
- Park, T., Yi, S., Lee, S., Lee, S. Y., Yoo, D., Ahn, J., and Lee, Y. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6):694–703.
- Pawitan, Y., Karuturi, R. K. M., Michiels, S., and Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, 21(20):3865–3872.
- Ploner, A., Calza, S., Gusnanto, A., and Pawitan, Y. (2006). Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, 22(5):556–565.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–46.
- Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics*, 20:1737–1745.

- Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19:1236–1242.
- Pounds, S. B. (2006). Estimation and control of multiple testing error rates for microarray studies. *Briefings in Bioinformatics*, 7:25–36.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 3 edition.
- Qiu, X., Klebanov, L., and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4(1).
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).
- Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4):265–273.
- Stegmaier, K., Wong, J. S., Ross, K. N., Chow, K. T., Peck, D., Wright, R. D., Lessnick, S. L., Kung, A. L., and Golub, T. R. (2007). Signature-based small

- molecule screening identifies cytosine arabinoside as an EWS/FLI modulator in ewing sarcoma. *PLoS Medicine*, 4(4):e122.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of The Royal Statistical Society Series B*, 64:479–498.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 100:9440–9445.
- Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545–15550.
- Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461.
- Tsai, C., Hsueh, H., and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, 59:1071–1081.

- Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98:5116–5121.
- Wit, E. and McClure, J. (2003). Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics*, 19(9):1055–1060.
- Xie, Y., Pan, W., and Khodursky, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280–4288.