Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations

Graduate School

2012

Statistical Methods for Normalization and Analysis of High-Throughput Genomic Data

Tobias Guennel
Virginia Commonwealth University

Follow this and additional works at: http://scholarscompass.vcu.edu/etd

Part of the Biostatistics Commons

© The Author

Downloaded from

http://scholarscompass.vcu.edu/etd/2647

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Statistical Methods for Normalization and Analysis of High-Throughput Genomic Data

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at Virginia Commonwealth University.

by

Tobias Guennel

Dipl.-Math. techn., Chemnitz University of Technology, Germany, 2008

B.S. Mathematics, Longwood University, USA, 2006

Director: Mark Reimers, Ph.D., Assistant Professor, Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia, USA

December, 2011

Acknowledgements

I would like to thank my adviser, Dr. Mark Reimers, for his patience, guidance, and friendship over the last two years. He challenged me to not only grow as a statistician but also as an instructor and independent thinker and it has been a privilege to work with him. I would like to thank Dr. Kellie Archer, who got me interested in molecular biology and in pursuing a degree in Biostatistics in the summer of 2006 . I would to thank my doctoral committee members Dr. Nitai Mukhopadhyay and Dr. Shirley Taylor for their patience and time. I would like to thank Dr. Michael Neale not only for being on my doctoral committee, but more importantly for accepting me to the National Institute for Drug Abuse training grant (grant number R25DA026119) and providing the financial support that allowed me to work on this project.

I also would like to thank my peers in the Department of Biostatistics, the Graduate Student Association, and the VCU Alumni Association for inspiring me outside the classroom. Joining the student government gave me the opportunity to meet many wonderful and dedicated people and was a life lesson I would not like to miss.

I would like to thank my friends in Richmond for keeping me sane and reminding me that there is a life outside academia. You will make it difficult to leave Richmond, but so many good memories will remain. I would also like to thank my friends back home, who despite my continued absence welcome me back with open arms whenever I have the chance to visit.

Lastly, but most importantly, I want to thank my parents, Dagmar and Thomas, and my brother Florian for their unconditional love and support of what turned out to become a ten year long journey through three universities on two continents. My parent's sacrifices over the last 30 years gave me the privilege to pursue my dreams and my family's encouragement and belief in my abilities ultimately gave me the strength to complete this project.

Table of Contents

List of Tables **v**

List of Figures **viii**

List of Abbreviations **xi**

Abstract **xv**

1 Introduction **1**

2 Normalization of aCGH Microarrays **7**

2.1 Introduction 7

2.2 Materials and Methods 11

2.2.1 Datasets 11

2.2.2 Algorithm 12

2.2.3 Performance Measures 21

2.2.4 Implementation 26

2.3 Results 26

2.3.1 Spatial Artifacts 27

2.3.2 Derivative Log2Ratio Spread 28

2.3.3 Improvement in Signal to Noise Ratios 29
2.3.4 Improvement in Agreement between Technologies 31
2.3.5 Improved Resolution of Called Segments 33
2.3.6 Interpretation and Utility of LOESS Regression and Principal Components 36
2.4 Discussion 45

3 Assessment of Current Methods for Analyzing RNA-Seq Studies 48
3.1 Introduction 48
3.2 Methods 50
3.2.1 edgeR 50
3.2.2 DESeq 53
3.2.3 Simulation Studies 54
3.2.4 Application to Internal and Publicly Available Datasets 58
3.3 Results 59
3.4 Conclusions 60

4 Analyzing RNA-Seq Studies Using a Negative Binomial Model with Zero-inflated Component 64
4.1 Introduction 64
4.2 Methods 69
4.2.1 Negative binomial model with zero-inflated component 69
4.2.2 Model Assessment 71
4.3 Results 72
4.3.1 Statistical Properties 72
4.3.2 Application to internal and publicly available data 80

4.3.3 Impact of zero-inflation on power and type I error 81
4.3.4 Implementation 82
4.4 Conclusions and Discussion 87

5 Alternative Usage of mRNA polyadenylation sites 91

5.1 Introduction 91
5.2 Materials and Methods 92
5.2.1 3' UTR Database Used And Corresponding Read Counts 92
5.2.2 Testing for Differential 3' UTR Usage Between Individual Samples . 94
5.2.3 Testing for Differential 3' UTR Usage Between Two Conditions With Biological Replicates 95
5.3 Results 96
5.4 Testing for Differential 3' UTR Usage Between Individual Samples 96
5.5 Testing for Differential 3' UTR Usage Between Two Conditions With Biological Replicates 97
5.6 Conclusions and Discussion 98

References 101

A Supplementary Tables 114

B Supplementary Figures 123

C Supplementary Code 126

List of Tables

2.1 Summary of Agilent and NimbleGen Datasets 12

2.2 Derivative Log2Ratio Spread: Comparison of DLRS estimates and mean algorithm efficiency 28

2.3 Mean fold change between signal to noise ratios for dye swaps 29

2.4 Signal to noise ratios for technical replicates 30

2.5 Agreement between technologies: Mean correlations between NimbleGen and Agilent probe intensities for normalization methods used 31

2.6 Variance explained in log_2 ratio intensity differences between technologies . 32

2.7 Proportion of Probes in Agreement (PPA) 32

2.8 Improvement in Resolution: Median aberrant region length (MARL), proportion of probes called aberrant (PPCA), and median MAD across all cell lines after segmentation (MMAS) 34

4.1 Proportion of genes with empirical p-values smaller than 0.001 67

4.2 Number of genes with AIC differences showing moderate and strong evidence for improved model fit among genes for which AIC_{ZINB} was smaller than AIC_{NB} 68

4.3 Median bias and IQR obtained from simulated datasets of sample size 15 based on Cheung data structure 76
4.4 Genes called significant at a 10% FDR cutoff when testing for gender differences. 81
4.5 Convergence rates for the Chen, Marioni, and Cheung datasets 82

5.1 Genes with evidence for correlation between differential 3' UTR usage and brain pH 102

A.1 Size estimates at significance level 0.001 across 64 simulated data sets under various dispersion distribution scenarios and sample sizes 115
A.2 Median (MAD) number of false positives at FDR cutoff of 10% across 64 simulated data sets under various dispersion distribution scenarios and sample sizes 116
A.3 Power estimates at FDR cutoff of 10% across 64 simulated data sets under various dispersion distribution scenarios and sample sizes 117
A.4 Size estimates at significance level 0.001 across 64 simulated data sets in scenario 2 with zero inflation 118
A.5 Size estimates at significance level 0.001 across 64 simulated data sets in scenario 4 with zero inflation 119
A.6 Power estimates at 10% FDR cutoff across 64 simulated data sets in scenario 2 with zero inflation 120
A.7 Power estimates at 10% FDR cutoff across 64 simulated data sets in scenario 4 with zero inflation 121

A.8 Percentage of genes showing evidence for differential 3' UTR usage and percentage of those genes with using longer 3' UTRs when compared to brain tissue for UTRs of type III 122

A.9 Percentage of genes showing evidence for differential 3' UTR usage and percentage of those genes with using longer 3' UTRs when compared to brain tissue for UTRs of type II 122

List of Figures

1.1 Worflow of a typical aCGH experiment [101]. 2

1.2 Example of intensity ratios for aCGH experiment. 3

1.3 Workflow of a typical RNA-Seq experiment [106]. 6

2.1 Spatial Artifacts in Reference Channel for NimbleGen Arrays: Plots show deviations from average reference channel intensities for two NimbleGen arrays 15

2.2 Impact of Normalization on Spatial Artifacts: Plots show deviations from average log_2 intensity ratios a) before and b) after normalization 16

2.3 Principal Components for Agilent and NimbleGen Data: Scree plots obtained by robust principal components analysis of residual differences between neighboring probes 22

2.4 Visualization of Pre- and Post-normalized Data: Raw and Normalized Data for a Segment of Chromosome 17 of Cell Line SK-Mel-5 35

2.5 Density Plots of Signal Intensities Before and After Normalization: Density plots for cell line OVCAR-8 are shown pre- and post-normalization. . . . 35

2.6 Scatter plot of each probe's first principle component loading against its GC content where the PCA step was applied directly to the raw NimbleGen data. The correlation between the PC loadings and GC content is 0.247. . 38

2.7 Scatter plot of each probe's first principle component loading against its GC content where the PCA step was applied to the LOESS normalized NimbleGen data. The correlation between the PC loadings and GC content is 0.071. 39

2.8 Scatter plot of each probe's first principle component loading against its genomic position for the Agilent data set. 40

2.9 Scatter plot of each probe's first principle component loading against its genomic position for the NimbleGen data set. 41

2.10 Scatter plot of each probe's LOESS residual against its genomic position for the NimbleGen data set. 42

2.11 Scatter plot of each probe's first principle component loading for Agilent and Agilent Dye Swap data sets. The correlation between the loadings is 0.647. 43

2.12 Scatter plot of each probe's first principle component loadings for randomly divided NimbleGen data sets (36 arrays each). The correlation between the loadings is 0.384. 44

3.1 Dispersion estimates against mean gene count on a log scale for the Marioni dataset 56

3.2 Dispersion estimates against mean gene count on a log scale for the Cheung dataset 57

3.3 Dispersion estimates against true dispersion value on log scale 62

3.4 Dispersion values against mean gene count on a log scale 63

4.1 Heatmap of counts for Marioni dataset 66
4.2 Size estimates for simulation studies at 0.001 significance level. The red line indicates nominal level. 73
4.3 Median number of false positives when testing for species differences. . . . 74
4.4 Power against sample size for Cheung simulation studies. 75
4.5 Test statistics against dispersion estimates on log scale for the Marioni dataset. 78
4.6 Test statistics against dispersion estimates on log scale for the Cheung dataset. 79
4.7 Size estimates for scenario 2 at 0.001 significance level. The red line indicates nominal level. 83
4.8 Size estimates for scenario 4 at 0.001 significance level. The red line indicates nominal level. 84
4.9 Power estimates for scenario 2 at 10% FDR cutoff. 85
4.10 Power estimates for scenario 4 at 10% FDR cutoff. 86

5.1 Scenarios considered for differential 3' UTR usage analysis [49]. 92
5.2 P-value plots for age, PMI, brain pH and diagnostic group omitting a random effect term. 99
5.3 P-value plots for age PMI, brain pH and diagnostic group including a random effect term. 100

B.1 Schematic of Illumina Genome Analyzer 2 workflow[50]. 124
B.2 Schematic of ABI SOLiD workflow [50]. 125

List of Abbreviations

aCGH	array comparative genomic hybridization
AIC	Akaike information criterion
bp	base pair
cCGH	chromosomal comparative genomic hybridization
cDNA	complementary DNA
CNV	Copy number variation
DLRS	Derivative Log2Ratio Spread
DNA	Deoxyribonucleic acid
FDR	false discovery rate
GAM	generalized additive models
GLM	generalized linear model
IQR	interquartile range
LOESS	locally weighted scatterplot smoothing

LRT	likelihood ratios test
MAD	median absolute deviation
MARL	median aberrant region length
ML	maximum likelihood
MMAS	median MAD across all cell lines after segmentation
mRNA	mature RNA
NB	negative binomial
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	next-generation sequencing
PCA	principle component analysis
PMI	post mortem interval
PPCA	proportion of probes called aberrant
qCML	quantile adjusted conditional maximum likelihood
RA	Running Average
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
SNPs	single nucleotide polymorphisms

UTR	untranslated region
WL	weighted conditional maximum likelihood
ZINB	zero-inflated negative binomial

Abstract

STATISTICAL METHODS FOR NORMALIZATION AND ANALYSIS OF HIGH-THROUGHPUT GENOMIC DATA

By Tobias Guennel, Dipl.-Math. techn.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2012

Major Director: Mark Reimers, Ph.D., Assistant Professor, Department of Biostatistics

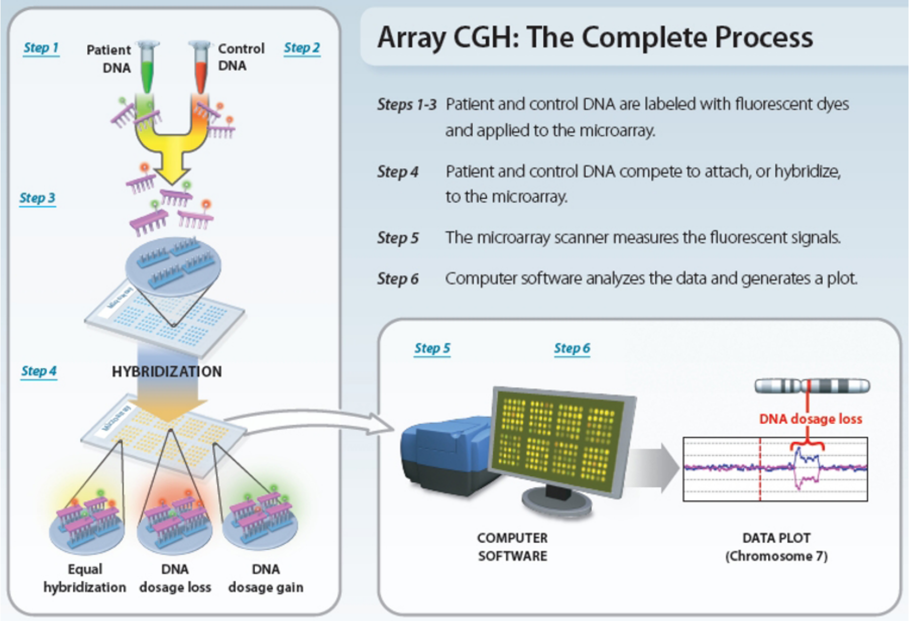
High-throughput genomic datasets obtained from microarray or sequencing studies have revolutionized the field of molecular biology over the last decade. The complexity of these new technologies also poses new challenges to statisticians to separate biological relevant information from technical noise. Two methods are introduced that address important issues with normalization of array comparative genomic hybridization (aCGH) microarrays and the analysis of RNA sequencing (RNA-Seq) studies. Many studies investigating copy number aberrations at the DNA level for cancer and genetic studies use comparative genomic hybridization (CGH) on oligo arrays. However, aCGH data often suffer from low signal to noise ratios resulting in poor resolution of fine features. Bilke *et al.* [11] showed that the commonly used running average noise reduction strategy performs poorly when errors are dominated by systematic components. A method called pcaCGH is proposed that significantly reduces noise using a non-parametric regression on technical covariates of probes to estimate systematic bias. Then a robust principal components analysis (PCA) estimates any remaining systematic bias not explained by technical covariates used in the preceding regression. The proposed algorithm is demonstrated on two CGH datasets measuring the NCI-60 cell lines utilizing NimbleGen and Agilent microarrays. The method achieves a nominal error variance reduction of 60%-65% as well as an 2-fold increase in signal to noise ratio on average, resulting in more detailed copy number estimates. Furthermore, correlations of signal intensity ratios of NimbleGen and Agilent arrays are increased by 40% on average, indicating a significant improvement in agreement between the technologies.

A second algorithm called gamSeq is introduced to test for differential gene expression in RNA sequencing studies. Limitations of existing methods are outlined and the proposed algorithm is compared to these existing algorithms. Simulation studies and real data are used to show that gamSeq improves upon existing methods with regards to type I error control while maintaining similar or better power for a range of sample sizes for RNA-Seq studies. Furthermore, the proposed method is applied to detect differential 3' UTR usage.

Chapter 1

Introduction

A new era in science is usually heralded by the adoption of a new technology that increases throughput by an order of magnitude and exponentially reduces cost compared with existing approaches. In the field of molecular biology, scientists have been witnesses and drivers of two intertwined eras over the last two decades: the eras of microarrays and next-generation sequencing (NGS). Microarrays evolved from Southern Blotting, where fragmented DNA is attach to a substrate and then probed with a known gene or fragment and were first reported in the late 1980's and early 1990's [4, 5, 53]. It took microarray technology until the mid to late 1990's, when miniaturized microarrays were first used for gene expression profiling and a complete eukaryotic genome could be fit onto a microarray [43, 90], to develop the throughput and cost efficiency to revolutionize how scientists were investigating molecular processes. Since then the uses for microarrays have increased exponentially to provide new insights into DNA methylation, copy number variation via DNA genotyping and array comparative genomic hybridization (aCGH), and protein binding site usage through chromatin immunoprecipitation assays to name a few [28, 30, 91]. In a typical microarray experiment, a mRNA or DNA sample from a given cell type or



Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures the fluorescent signals.

Step 6 Computer software analyzes the data and generates a plot.

Figure 1.1. Worflow of a typical aCGH experiment [101].

tissue is used to generate a labeled sample, sometimes termed the 'target', which is hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered array. The density of DNA sequences that can be attached to the microarray surface allows the detection and quantification of up to several millions of targets simultaneously.

One application for microarray technology is the detection of copy number variations in DNA samples through aCGH. Figure 1.1 illustrates the workflow of an aCGH microarray experiment. First, DNA is extracted from a test and a normal control (reference) sample and both are labeled with a fluorescent dye of different colors after DNA fragmentation. The two genomic DNA fragment samples are then washed over a microarray and since the DNA has been denatured, its fragments are single stranded and attempt to hybridize with the arrayed single-strand probes. Next, digital imaging systems are used to capture and quantify the relative fluorescence intensities of the labeled DNA probes that have

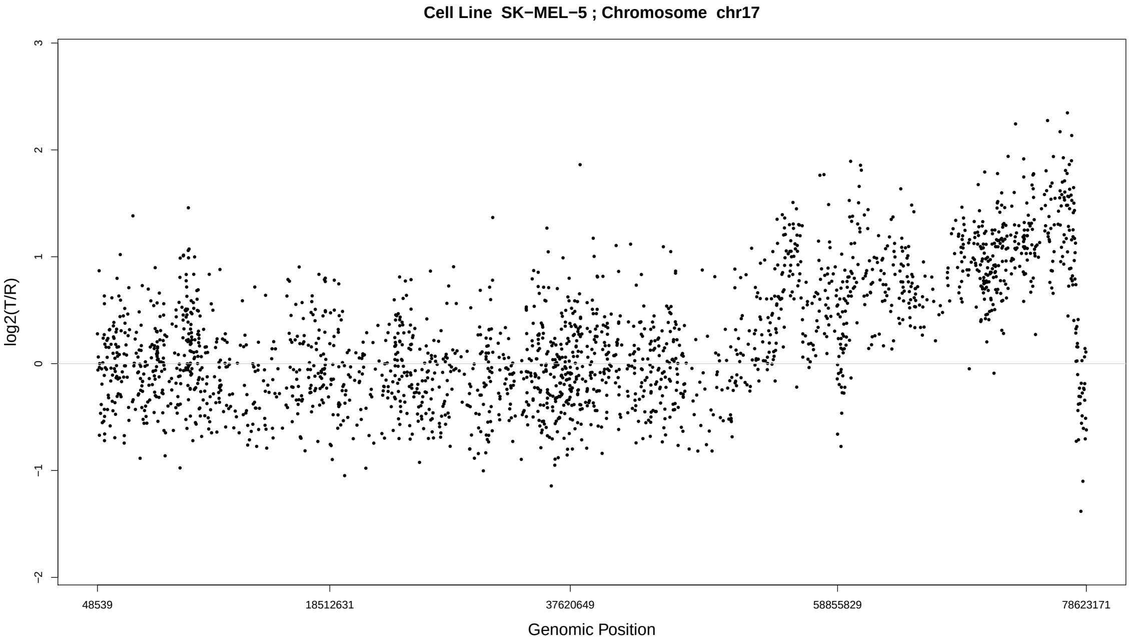


Figure 1.2. Example of intensity ratios for aCGH experiment.

hybridized to each target. The fluorescence ratio of the test and reference hybridization signals is determined at different positions along the genome and provides information on the relative copy number of sequences in the test genome as compared to the normal genome. Intensity ratios are log_2 transformed and an example how these ratios are represented is shown in Figure 1.2. Here, log ratios of zero indicate no copy number variation while positive and negative ratios indicate copy number gains or losses, respectively. Detecting copy number variations ultimately transforms into a change point problem where scientists want to pin-point genomic regions of elevated or decreased copy number ratios as exactly as possible. It is apparent in Figure 1.2 that these ratios are quite noisy and normalization methods that improve signal to noise ratios allow more accurate detection of change points are needed. Chapter 2 introduces a novel algorithm to normalize aCGH data and compares it to existing methods.

After microarrays had become the most common technology for investigating molecular processes on a genome-wide scale, scientists were looking for a technology that would overcome some of the pitfalls of microarray technology including for example a limited dynamic range and reliance on manufactures to provide accurate assays for their biological question of interest. This new technology took center stage in 2008 when the first complete human diploid genome was sequenced using Roche's 454 sequencing platform [108] to perform massively parallel DNA sequencing. Since then, high-throughput or next-generation sequencing has revolutionized the field of molecular biology in recent years replacing microarrays as the method of choice for many genome-wide studies of transcription levels (RNA-Seq), DNA-protein interactions (ChIP-Seq), chromatin structure and DNA methylation (Methyl-Seq) [22, 29, 51, 58, 60]. The basic principle behind DNA sequencing is that DNA fragments are not hybridized to probes attached to a glass surface but rather sequenced directly. Sanger sequencing [89] had long been the choice for sequencing small numbers of DNA fragments, but the need for low-cost sequencing spurred the development of the above mentioned technologies that can sequence millions of DNA fragments in parallel. Numerous sequencing platforms lead by Illumina's HighSeq and GenomeAnalyzer systems [57] and Applied Biosystem's SOLiD system [54] have emerged and provide scientists with a myriad of possibilities to quantify molecular processes on a genome-wide scale with high resolution. For schematics how these two platforms sequence DNA fragments, please refer to Figures B.1 and B.2 in Appendix B. A detailed review of these and other NGS technologies can be found in [54]. The increase in throughput and cost reduction that could be achieved through massive parallel sequencing was enormous. While scientists could sequence six Megabases per day at \$500 per Megabase sequenced with Sanger sequencing, Roche's 454 platform has a throughput of 750 Megabases a day at \$10 per Megabase while Illumina's HighSeq platform and ABI's SOLiD platform achieve up to 35 Gigabases per day at \$0.10 per Megabase [25]. These cost savings have led to an exponential increase in experiments utilizing NGS and new methods to normalize and analyze these datasets are needed. Although early studies claimed that NGS is less prone to technical artifacts than microarrays [51, 106], it has become clear that this technology has its own pitfalls [15, 22, 27, 57, 85, 87]. While these studies show that quality assessment and normalization are important for NGS data, the second part of this thesis focuses on the subsequent step, i.e. the analysis of quality assured data. More specifically, methods for analyzing gene expression data obtained through RNA sequencing (RNA-Seq) are considered. Figure 1.3 shows a schematic of a typical RNA-Seq experiment. First, mRNA from a tissue of interest is extracted, fragmented and reverse transcribed. These cDNA fragments are then pre-processed and sequenced according to protocols supplied by the manufacturer of the sequencing platform. The short sequence reads obtained from the sequencing run are then aligned to a reference genome. Using this reference genome, the researcher can then define a gene model of choice by defining genomic regions that represent a gene. Gene expression levels are determined by counting how many reads fall into these pre-defined regions. Once count data for each gene and sample are obtained, the data is ready for analysis.

In Chapter 3, existing methods to analyze gene expression data obtained through RNA-Seq are assessed. A new algorithm to analyze RNA-Seq data is proposed and compared to existing methods in Chapter 4. Furthermore, the flexibility of the proposed is illustrated in Chapter 5.

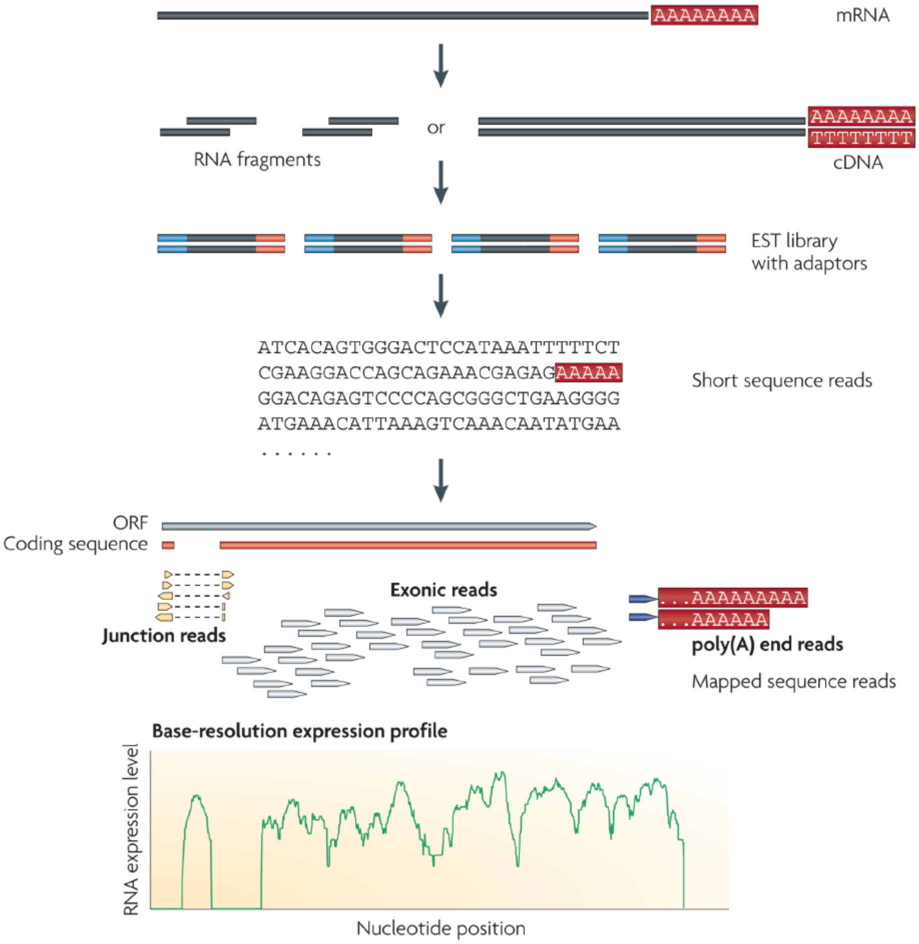


Figure 1.3. Workflow of a typical RNA-Seq experiment [106].

Chapter 2

Normalization of aCGH Microarrays

2.1 Introduction

Copy number variation (CNV) of DNA sequences has long been suspected to be a form of normal genetic variation and to play an important role in many genetic disorders. However, only recently has its importance in human diversity been demonstrated by Sebat *et al.* [92] and Iafrate *et al.* [33]. In a subsequent study, Redon *et al.* [79] found quantitative evidence showing that CNV regions cover at least 12% of the human genome and more nucleotide content per genome than single nucleotide polymorphisms (SNPs). DNA copy number changes are also consistently observed in cancer cells where different types of cancers show different copy number structures [35, 55]. More interestingly, Weiss *et al.* [107] and Van Wieringen *et al.* [103] observed correlations between CNV and clinical outcomes such as patient survival and responsiveness to certain treatments. In this light, obtaining accurate and reliable estimates of chromosomal copy numbers has become increasingly important.

Chromosomal comparative genomic hybridization (cCGH) was introduced by Kallioniemi *et al.* [36] as method to investigate DNA copy number alterations on a genome-wide scale. cCGH is capable of detecting loss, gain, and amplification of genomic regions with different sensitivities for different copy number alterations. Copy number amplifications can be detected in regions of less than 1 Mb, while a single copy loss can be difficult to detect in regions of less than 5 Mb in length. With the emergence of microarrays, array comparative genomic hybridization (aCGH) offered improved resolution, a higher dynamic range, and improved throughput, along with a convenient way to access the location of copy number alterations on the genomic map [70]. In a typical aCGH experiment, genomic DNA from test and reference cells is isolated, labeled with two different fluorescent dyes and hybridized on an oligo microarray. After removing excess dye particles, a high resolution camera takes pictures of each channel, which are subsequently scanned and transformed into test and reference channel intensities. Typically, the ratio of test to reference channel intensity (T/R) on a log_2 scale is used to investigate copy number variations. Since the reference genome is assumed to have very few copy number variations, any significant departures from zero in the $log_2(T/R)$ ratios indicate copy number aberrations in the test sample.

Many early studies using aCGH have focused on amplifications where 10 or more extra DNA copies are present. Lately detecting heterozygous deletions and subtle gains of one or more extra DNA copies or detecting short to moderate length copy number aberrations has become of increasing interest [10, 111]. However, low level copy number changes are difficult to detect due to the low signal to noise ratios of current aCGH technologies. The most commonly used approaches to increase signal to noise ratios during the analysis of CNV are variants of neighbor dependent methods that average signal intensities of neighboring probes [11, 47]. For example, the running average (RA) method [32, 71] calculates an average signal intensity from all neighboring probes contained in a windows of size W.

Then, a sliding window is used to define regions of size W across the genome that are investigated for copy number changes. The success of those methods in detecting small segments of CNV depends on the length of the window selected, which in turn depends on the signal to noise ratio in the data after pre-processing. Smaller signal to noise ratios allow for smaller segments to be averaged, which consequently results in higher resolution.

Conducting a microarray experiment is a multi-step process during which technical variation can be introduced by several sources. Technical variation is considered to be any variation induced by differences in sample handling or the hybridization process. A technical covariate is a variable that indexes differential effects of these differences in handling on log-ratio intensities. If the magnitude of technical variation is comparable in size to the magnitude of biological variation, the experiment is unlikely to yield statistically significant biological differences. The first stage of the proposed algorithm aims to reduce variation which can be indexed or predicted by known technical covariates. Normalization methods adjusting for technical covariates have been developed for ChIP-chip arrays [48] as well as Affymetrix gene expression [109]. Chen *et al.* [17] pointed out that normalization methods for aCGH data are rare and that researchers often apply methods developed for gene expression data in a slightly modified form to aCGH data. These authors demonstrate using simulated aCGH data that this approach is problematic due to the different nature of aCGH data, where the most important differences lie in smaller dynamic ranges and dependencies among probes based on genomic position. Neuvial *et al.* [63] uncovers continuous spatial biases as well as local spatial biases in BAC aCGH data, and stresses the importance of correcting for spatial artifacts for meaningful biological inferences.

Several methods have been proposed to increase signal to noise ratios during pre-processing of aCGH data. Median and quantile normalization [14] usually used for normalizing gene expression microarray data have also been used with aCGH data. Median normalization centers the distribution of log intensity ratios at its median and thus does not affect signal to noise ratios as it only shifts the distribution. Quantile normalization calculates a reference distribution of log_2 intensity ratios using all arrays in the data set and then replaces intensity ratios for individual arrays with the corresponding quantile of the reference distribution. The reference distribution is usually obtained by averaging corresponding quantiles from individual arrays across all arrays. Lepretre et al. [46] have proposed their waves aCGH correction algorithm (WACA) that uses a locally weighted scatterplot smoothing (LOESS) fit [19] based on GC content and fragment size correction to improve accuracy. Staaf et al. [99] propose an algorithm that performs population-based intensity-based LOESS (popLowess) smoothing. The algorithm first stratifies data into populations of copy numbers and then performs LOESS normalization based on M-A plots [96].

The proposed algorithm consists of two parts. First, a non-parametric locally weighted linear regression (LOESS) is used to estimate systematic bias due to variables that induce technical variation. Second, a principal component analysis (PCA) is employed to account for any unknown technical covariates that could further cloud the biological information contained in the data. A detailed description of the algorithm is provided in the following section.

2.2 Materials and Methods

2.2.1 Datasets

In the late 1980's, the National Cancer Institute (NCI) prepared a drug testing pipeline whose first screening stage measured the effect of putative anti-cancer agents on various cancer cell lines. To this end they gathered 60 cell lines, several of which (such as MCF-7) had been previously widely used in cancer research. These 60 cell lines came from nine distinct tissues of origin (although some were misclassified initially): breast, brain, colon, lung, kidney, ovary, prostate, lymphocytes and melanocytes.

Samples of those 60 cell lines were hybridized to Agilent Human Genome CGH Microarray 44K [1] and NimbleGen HG17 CGH 385K WG Tiling [64] arrays. A summary of each dataset is shown in Table 2.1. Thus, three datasets were available for analysis: two consisting of signal intensities obtained from NimbleGen [81] and Agilent arrays with standard dye assignments to test and control samples and one consisting of signal intensities obtained from Agilent arrays where the dye assignments to test and control samples were reversed. Note that the Agilent datasets contained four replicates for cell line A549-ATCC and also that cell line NCI-H226 was excluded from analysis due to low data quality. Furthermore, the experiment using NimbleGen arrays contained ten replicates for cell line A549-ATCC, including four dye swaps, and four replicates for cell lines SF-268 and OVCAR-8. The intensities used for analysis were the perfect match (PM) intensities provided by Nimblescan v2.3.4 for the NimbleGen arrays and the processed intensities provided by Agilent's Feature Extraction Software v8.1.18. In a later stage each probe's neighbors on the genome are identified by locating each probe by BLAT search.

The probes on the Agilent and NimbleGen aCGH platforms used here were based on the UCSC version HG19 [38, 42]. All probes on either platform that did not align uniquely to the most current human assembly built were dropped from the analysis reducing the effective number of probes per array (see Table 2.1). The data sets are publicly available on the CellMiner webpage [93] at http://discover.nci.nih.gov/cellminer.

Table 2.1. Summary of Agilent and NimbleGen Datasets

	Agilent	NimbleGen
Array Description	Human Genome CGH Microarray 44K	HG17 CGH 385K 2005-03-16_HG17_WG_CGH
Number of Probes per Array	44000	385000
Effective Number of Probes (I)	42853	378779
Length of Probe Sequence	60 mer	45-85 mer
Median Probe Spacing	43000 bp	5000 bp
Number of Arrays (K)	124	72
Number of Dye Swaps	62	4

2.2.2 Algorithm

The principle idea behind the proposed algorithm is the following partition of a data matrix $\boldsymbol{M}$ of intensity ratios:

$$\boldsymbol{M} = \boldsymbol{M}^{biological} + \boldsymbol{M}^{systematic} + \boldsymbol{M}^{random}. \tag{2.1}$$

Here, $\boldsymbol{M}^{biological}$ represents signal due to biological differences of interest, $\boldsymbol{M}^{systematic}$ represents systematic noise due to technical covariates, and $\boldsymbol{M}^{random}$ represents random noise. The goal is to normalize the raw data matrix $\boldsymbol{M}$ by estimating $\boldsymbol{M}^{systematic}$ and removing the corresponding residuals:

$$\boldsymbol{M}^{*} = \boldsymbol{M} - \widehat{\boldsymbol{M}}^{systematic}. \tag{2.2}$$

The following sections introduce a two step algorithm that aims to obtain good estimates of $\boldsymbol{M}^{systematic}$ to remove as much systematic noise as possible.

Technical Covariate Normalization

A non-parametric regression approach using LOESS was employed to account for suspected technical variables inducing non-biological variation. Locally weighted least squares regression (LOESS regression) is a technique to fit a smoothing surface to the data using second order polynomials. After carefully investigating the dependence of log-ratios on a variety of potentially informative technical covariates, the following probe specific technical covariates for aCGH experiments were identified:

1. horizontal and vertical coordinates indexed by probe position X and Y on the array,
2. average reference channel intensity $\overline{G}$ across all arrays,
3. difference of reference channel intensity G_k to average reference channel intensity $\overline{G}$ across all arrays, and
4. probe GC content GC.

For the purposes of this algorithm, the dependent variable is $M_{ki} = log_2(T_{ki}/R_{ki})$, the ratio of the raw probe intensities of probe i, $i = 1, \ldots, I$, for array $k = 1, \ldots, K$, on a log_2 scale. Then the LOESS model is specified by

$$M_{ki} = f(X_i, Y_i, GC_i, G_{ki} - \overline{G_i}, \overline{G_i}) + \epsilon \tag{2.3}$$

where f represents a nonlinear function, in this case a local regression surface, and ϵ represents biological signal plus error not predictable from the technical covariates used.

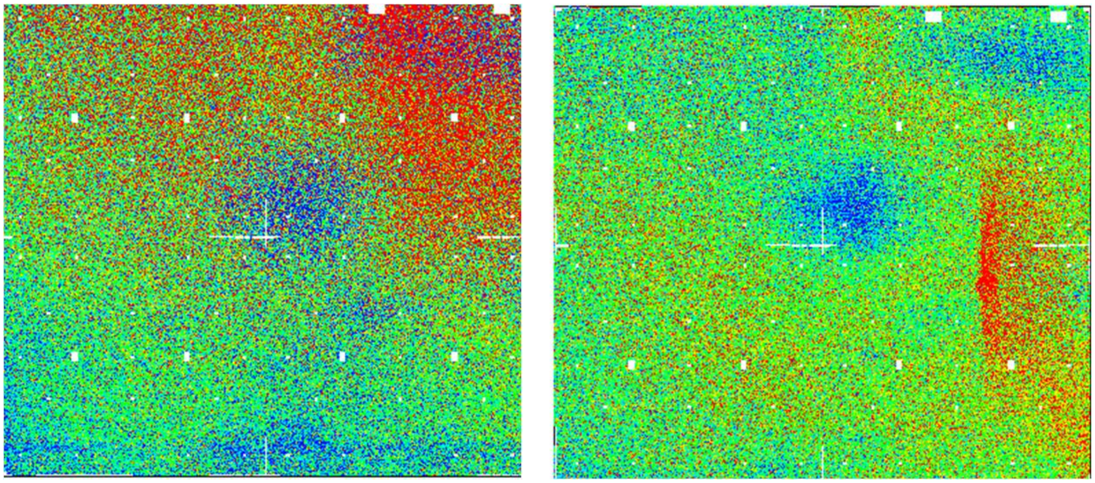
The normalized log_2 intensity ratios are obtained by

$$M_{ki}^{LOESS} = M_{ki} - \widehat{M}_{ki}, \tag{2.4}$$

where $\widehat{M}_{ki}$ are the log_2 intensity ratios as predicted from the technical covariates alone by the LOESS model in Equation 2.3.

The following paragraphs describe the empirical evidence that led to the choice of technical covariates. By design, microarray probes are randomly distributed across an array to avoid spatial biases, i.e. probes that represent adjacent regions on the genome are not located physically adjacent on an array. Therefore, in the absence of spatial artifacts, one would expect a random pattern of high ratios, low ratios, and ratios of one, across an array. Furthermore the distribution of reference channel intensities should be fairly homogeneous across the whole array as reference DNA was extracted from healthy cells, i.e. cells for which the vast majority of DNA regions should have a copy number of two. It is known that even healthy cells contain some DNA regions with deletions or amplification, but probes measuring copy numbers for those sparse regions should again be randomly distributed across the array. Reimers and Weinstein [80] introduced quality assessment plots for microarray data to visualize spatial artifacts of regional biases. Figures 2.1 and 2.2.a) show plots of log_2 reference channel intensities and log_2 intensity ratios, respectively, by physical probe position on two specific NimbleGen arrays. Two types of non-random patterns can clearly be identified, continuous spatial gradients and distinct local spatial artifacts. Considering log_2 intensity ratios, continuous spatial gradients were detected on numerous NimbleGen as well as Agilent arrays and a significant number of NimbleGen arrays and a few Agilent arrays showed local spatial artifacts. Those patterns

Figure 2.1. Spatial Artifacts in Reference Channel for NimbleGen Arrays: Plots show deviations from average reference channel intensities for two NimbleGen arrays



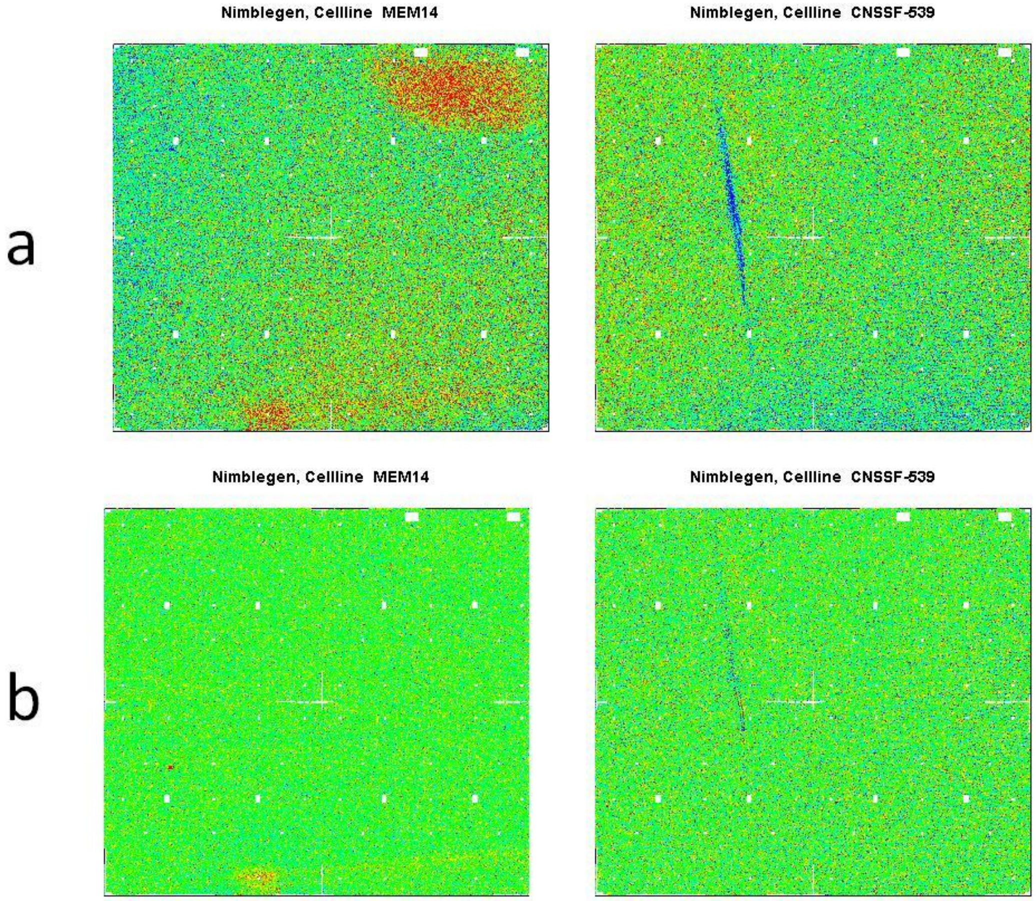
seem to be even more pronounced for reference channel intensities on NimbleGen arrays. To account for spatial artifacts in the reference channel of aCGH data, the reference channel intensity for each array, as well as the difference of reference channel intensity to average reference channel intensity across all arrays were used in the LOESS regression. Let i, $i = 1, \ldots, I$, denote the ith probe on a aCGH array. The average reference channel intensity $\overline{G_i}$ for probe i is calculated as a 10% trimmed mean of G_{ki} for all $k = 1, \ldots, K$ where K is the number of arrays in an aCGH dataset and

$$\overline{G_i} = \frac{1}{0.8\,K}[(L - 0.2 * K)(G_{(L)i} + G_{(K-L+1)i}) + \sum_{j=L+1}^{K-L} G_{(j)i}], \tag{2.5}$$

where $L = \lfloor 0.2K \rfloor + 1$ and $G_{(j)i}$ denotes the j^{th} ordered reference channel intensity. Equation 2.5 can similarly be used to calculate trimmed means of continuous data vectors $\boldsymbol{x}$ and this calculation will be abbreviated by $Mean_{10\%}(\boldsymbol{x})$ in the following.

Agilent arrays also show non-random patterns (plots not shown), but the frequency and magnitude of those patterns were much smaller compared to NimbleGen arrays. Fur-

Figure 2.2. Impact of Normalization on Spatial Artifacts: Plots show deviations from average log_2 intensity ratios a) before and b) after normalization



thermore, Agilent's Feature Extraction Software v8.1.18 pre-processing step of the raw intensities produces similar effects as stage 1 of the proposed algorithm and therefore only stage 2 was applied to the Agilent intensities. It should be noted that newer versions of NimbleScan (v2.5 and higher) also include LOESS based spatial correction to adjust signal intensities based on physical feature position. The algorithm should be adjusted to the corresponding software version to avoid overcorrection of real biological signal.

Wu *et al.* [109] have shown that melting temperature can influence probe intensities. The melting temperature T_m of a DNA strand is defined as the temperature where one-half of its nucleotides are paired with their complement while one-half are unpaired. Ideally, each probe on the array should have the same melting temperature. However, since each probe must uniquely identify one specific region of the genome, heterogeneous melting temperatures across an array are common and need to be accounted for. A probe's GC content is a good proxy of its melting temperature. Furthermore, several studies [52, 61, 98] have identified probe GC content as an important source of technical variation.

The LOESS curve is fit using the `loess` function implemented in the **R** programming environment [77]. Since the `loess` function can only fit four predictors at a time, two LOESS surfaces were fit and the corresponding residuals were subtracted sequentially. The first LOESS surface fits predictors related to spatial properties, i.e. horizontal and vertical coordinates indexed by probe position X and Y on the array, average reference channel intensity $\overline{G}$ across all arrays and difference of reference channel intensity G_k to average reference channel intensity $\overline{G}$ across all arrays. The second LOESS surface fits GC content by itself as this is a sequence specific predictor. Due to the evident fine-scale structure in Figures 1 and 2, a span of 1% is used, i.e. 1% of the data is used to fit each local regression, for the first LOESS fit. On the other hand, the dependence on probe GC content is smoother, and therefore a span of 30% in the second LOESS fit was used. In principle the size of the LOESS span was chosen to use up approximately 1% of the total available degrees of freedom. It was found however that using a span smaller than 1% increases the computational burden substantially while only marginal improvement in normalization performance could be achieved.

Robust Principal Component Analysis of Residuals

The novel part proposed here is the combination of technical covariate normalization with principal component analysis (PCA) to uncover any remaining systematic patterns due to unaccounted technical variation after accounting for specific variables that are expected to index unwanted technical variation in stage 1. The goal of PCA is to find a few principal components that explain the majority of variance inherent in the data, i.e. to perform a dimensionality reduction of the data. Specifically, a data matrix $\boldsymbol{X} \in \mathbb{R}^{I \times K}$ is modeled as

$$\boldsymbol{X} = \boldsymbol{D}^{\tau} \boldsymbol{Y}^{\tau} + \boldsymbol{E}. \tag{2.6}$$

Here, $\boldsymbol{D}^{\tau} \in \mathbb{R}^{I \times p}$ denotes the transpose of the reduced rotation matrix containing only the first p principal directions, $\boldsymbol{Y}^{\tau} \in \mathbb{R}^{p \times K}$ contains coordinates of each data point in the new coordinate system, i.e. $\boldsymbol{Y}^{\tau}$ constitutes the transpose of the rotated data matrix, and $\boldsymbol{E} \in \mathbb{R}^{I \times K}$ are the corresponding residuals [45].

PCA is a special case of projection pursuit which maximizes the magnitude of a projection index PI after projection multidimensional data into a one-dimensional subspace.

Formally, the jth principal direction $\boldsymbol{a_j}$ is defined by

$$\boldsymbol{a_j} = \underset{\|\boldsymbol{a}\|=1, \boldsymbol{a}\perp\boldsymbol{a_1},\ldots,\boldsymbol{a}\perp\boldsymbol{a_{j-1}}}{\arg\max} PI(\boldsymbol{a}^t\boldsymbol{x}_1, \ldots, \boldsymbol{a}^t\boldsymbol{x}_n) \tag{2.7}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are I-dimensional data points. Using the sample variance as projection index corresponds to PCA, i.e. finding the eigenvalues of the sample covariance matrix and the corresponding eigenvectors are the principal directions. PCA is not a robust method as outliers can significantly influence the obtained principal components and directions, Croux *et al.* [21] suggested using the median absolute deviation (MAD) as projection index to deal with the high incidence of outliers in microarray experiments [44, 59]. An efficient version of this algorithm is implemented in the **R** function `PCAGrid` within the package `pcaPP` [23].

The proposed approach is based on the inherent characteristic of aCGH data, that signal intensities of neighboring probes are highly correlated. Specifically, the approach takes advantage of the common situation that probes that are in close proximity measure the same copy number. The difference $M_{ki}^{LOESS} - \overline{M}_{ki}'$ between ratios of a specific probe and its neighboring probes is partitioned into a systematic part and a random part, i.e. $\boldsymbol{X} = \boldsymbol{M}^{LOESS} - \overline{\boldsymbol{M}}^{LOESS}$ in Equation 2.6 where $\boldsymbol{M}^{LOESS} = \{M_{ki}^{LOESS}\}_{i=1,\ldots,I;k=1,\ldots,K}$ and $\overline{\boldsymbol{M}}^{LOESS} = \{\overline{M}_{ki}^{LOESS}\}_{i=1,\ldots,I;k=1,\ldots,K}$. Here, $\overline{M}_{ki}^{LOESS}$ denotes the 10% trimmed mean of probes within a window of window size W around probe i. The difference $\boldsymbol{M}^{LOESS} - \overline{\boldsymbol{M}}^{LOESS} \in \mathbb{R}^{I\times K}$ can then be viewed as a matrix of bias estimates, one for each probe of each array. The accuracy of the bias estimates dependents on the window size W as well as the validity of the assumption that neighboring probes measure the same copy number. Therefore an iterative algorithm is used to find a window large enough to provide an accurate bias estimate while testing the validity of the above mentioned assumption. The window size W is calculated by iteratively including up- and downstream probes x_{i-h} and x_{i+j}, $h = 1, \ldots, H; j = 1, \ldots, J$, closest to probe i up to a maximum window size of $W_{max} = 30$ for Agilent arrays and $W_{max} = 60$ for NimbleGen arrays, i.e. $H_{max} = J_{max} = 15$ and $H_{max} = J_{max} = 30$, respectively. The maximum window sizes are chosen based on the probe density of the respective array to ensure that probes in relative proximity to the probe for which the bias ought to be estimated are chosen. At each step of the iterative inclusion process, an ad-hoc test is used to compare whether the inclusion of additional probes could possibly lead to the inclusion of a true copy number change due to a biological process. If an actual change in copy number is suspected, the current probe and all probes farther away from probe i are not included in the corresponding window. This comparison is performed for the up- and downstream regions independently and therefore the resulting window may not be symmetric. The ad-hoc test used to test for a true change in copy number after each iteration is based on a Smith-Waterman algorithm described in [74] and implemented in the **R** package `cgh` by Price *et al.* [75].

The normalized probe log_2 intensity ratios M_{ki}^{PCA} are then given by subtracting the systematic bias predicted by the principal components from the original log_2 ratio intensities:

$$\boldsymbol{M}^{PCA} = \boldsymbol{M}^{LOESS} - \boldsymbol{D}^{\tau}\boldsymbol{Y}^{\tau} = \overline{\boldsymbol{M}}^{LOESS} + \boldsymbol{E}, \tag{2.8}$$

where $\boldsymbol{M}^{PCA} = \{M_{ki}^{PCA}\}_{i=1,\ldots,I;k=1,\ldots,K}$. $\boldsymbol{E}$ is obtained by performing a robust PCA as described above on $\boldsymbol{M}^{LOESS} - \overline{\boldsymbol{M}}^{LOESS}$ and retaining the first p principal components, i.e.

$$\boldsymbol{E} = \boldsymbol{M}^{LOESS} - \overline{\boldsymbol{M}}^{LOESS} - \boldsymbol{D}^{\tau}\boldsymbol{Y}^{\tau}. \tag{2.9}$$

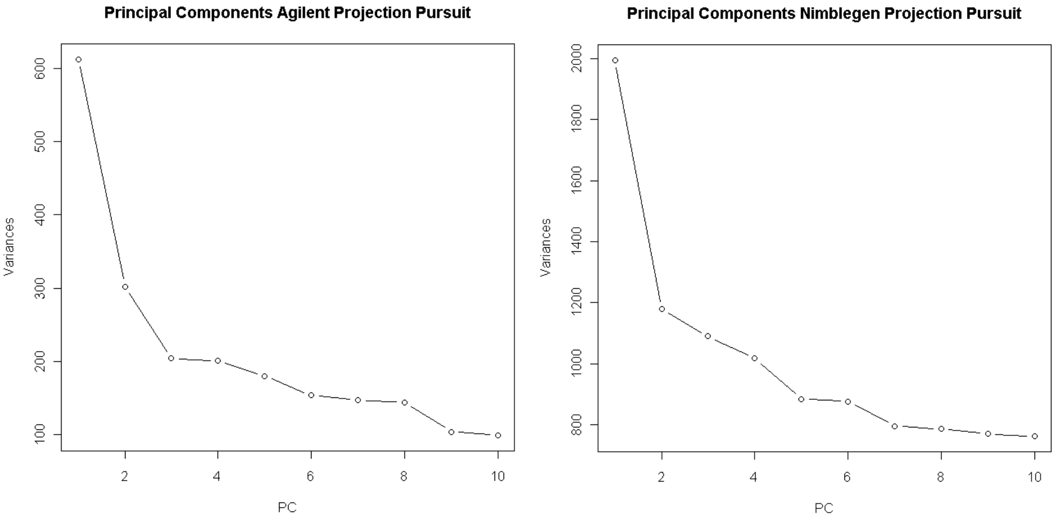
The number of principal components kept is determined by scree plots. Scree plots show the size of the principal components in decreasing order. If there is a strong systematic component, the scree plot will level off at some point and all principal components up to this point are retained. While it is not exact, the scree plot method is a reasonable approach. A great difference in performance was not observed when one or two additional PCs judged to be optimal from the scree plot were used. The scree plots for both Agilent and NimbleGen data are shown in Figure 2.3. The plots level off after the third and fourth principal component and thus three and four principal components are retained for Agilent and NimbleGen data, respectively. For further details on PCA, please refer to Johnson & Wichern [34] and Croux *et al.* [21]. The procedure outlined above assumes that probes that are in close proximity measure the same copy number in the majority of cases, and therefore that $\boldsymbol{M}^{LOESS} - \overline{\boldsymbol{M}}^{LOESS}$ measures variability due to technical covariates and not biological signal. The iterative algorithm used to estimate the error structure was designed to automatically assess this assumption during the estimation process and ensure the validity of the PCA performed.

2.2.3 Performance Measures

The following qualitative and quantitative measures to assess the performance of the proposed normalization method were used:

- quality assessment plots visualizing spatial artifacts before and after normalization,
- the Derivative Log2Ratio Spread (DLRS) investigating the variance reduction in differences between adjacent probes,
- signal to noise ratios of technical replicates of cell lines A549-ATCC and SF-268,

Figure 2.3. Principal Components for Agilent and NimbleGen Data: Scree plots obtained by robust principal components analysis of residual differences between neighboring probes



- signal to noise ratios of arrays and their corresponding dye swaps of the same cell line in the Agilent dataset,
- concordance between NimbleGen and Agilent arrays, and
- median aberrant region lengths as a measure of resolution depth.

Several of these performance measures have also been used in a recent study comparing several aCGH platforms [68].

Derivative Log2Ratio Spread

The Derivative Log2Ratio Spread (DLRS) was introduced by Kincaid *et al.* [39] and is implemented in Agilent's own DNAanalytics software as the metric of choice for noise quantification. It calculates a robust variance estimate of the difference in log_2 intensities of neighboring (with respect to chromosomal location) probes. The principal assumption is that the majority of adjacent probes measure the same copy number. In fact, the DLRS assumes that less the 50% of probes delimit breakpoints [46]. Let

$$D^{method}_{i,i+1;k} = M^{method}_{k(i+1)} - M^{method}_{ki}; i = 1, \ldots, I-1; k = 1, \ldots, K, \tag{2.10}$$

denote the difference in log_2 intensity ratio of adjacent probes i and $i+1$ on array k for a specific normalization method (or the raw data). The DLRS for array k and a specific method is given by

$$DLRS^{method}_{k} = \frac{Q_3(\boldsymbol{D}^{method}_{k}) - Q_1(\boldsymbol{D}^{method}_{k})}{1.349 * \sqrt{2}} \tag{2.11}$$

where $\boldsymbol{D}^{method}_{k} = D^{method}_{i,i+1;k}\,{}_{i=1,\ldots,I-1;k=1,\ldots,K}$ and $Q_n(\boldsymbol{D}^{method}_{k})$ denotes the nth quartile of the corresponding distribution of differences. The DLRS, presented as a robust method of estimating noise from the sample array alone, can range from under 0.2 for an excellent array to higher than 0.3 for poor experiments. One way to evaluate the mean efficiency of an algorithm can then be written as

$$DLRS^{method}_{eff} = \frac{1}{K} \sum_{k=1}^{K} \frac{DLRS^{raw}_{k} - DLRS^{method}_{k}}{DLRS^{raw}_{k}}. \tag{2.12}$$

Signal to Noise Ratios

Signal to noise ratios for technical replicates and dye swaps were estimated. For the four replicates of the A549-ATCC sample, the mean variance within replicates estimated using a robust estimator based on the median absolute deviation (MAD) was used as an estimate of the dynamic range (signal) while noise was estimated by the median of the variance across replicates calculated for all probes. Specifically, the signal to noise ratio was calculated using

$$SN = \frac{\frac{1}{K^*}\sum_{k^*=1}^{K^*}(1.4826 * MAD(\boldsymbol{M}_{k^*}^{method}))^2}{Median(\{Var(\boldsymbol{M}_i^{method})\}_{i=1,\ldots,I})} \tag{2.13}$$

where K^* is the number of replicates, $\boldsymbol{M}_{k^*}^{method} = \{M_{k^*}^{methodi}\}_{i=1,\ldots,I}$, $\boldsymbol{M}_i^{method} = \{M_{k^*}^{methodi}\}_{k^*=1,\ldots,K^*}$, and $Var(\boldsymbol{x})$ is the usual variance estimator.

Agreement between Technologies

To assess if there was an improvement in agreement between the two different aCGH platforms, detection call intensities obtained from StepGram [47] with default settings were used to compare technologies. Since the HG17 385K NimbleGen arrays have significantly more probes than the CGH 44K WG Agilent array, the comparison was made between each Agilent probe and the closest NimbleGen probe within a 15kb window. If no probe was located within a 15kb window, the corresponding Agilent probe was omitted in this analysis. The following metrics were used to assess improvement between technologies:

1. the proportion of variance (scaled by geometric mean of total variance in the two data sets) explained in differences between matched Agilent and NimbleGen probes with respect to the raw data, i.e.

$$Var_{explained} = 1 - \frac{\frac{Var(A^*_{Norm} - N^*_{Norm})}{\sqrt{Var(A_{Norm}) * Var(N_{Norm})}}}{\frac{Var(A^*_{raw} - N^*_{raw})}{\sqrt{Var(A_{raw}) * Var(N_{raw})}}},$$

2. the correlation between signal intensities of the two platforms before and after normalization, and

3. the proportion of corresponding probes with the same detection call, i.e. pairs of

probes from different platforms that were both called a deletion, amplification, or neither.

While reporting correlation is more common, the first measure of agreement, the proportion of scaled variance explained, seems preferable since the correlation measure is often influenced by the large dynamic range of the array measures while the this measure is scaled by total variance.

Resolution after Segmentation

While improvements in the quantitative performance measures mentioned above are useful to assess how well normalization methods reduce noise levels, researchers are ultimately interested in called regions with copy number aberrations. Improvements in this outcome are difficult to measure, absent an independent assessment of true copy number aberrations. Nevertheless the following measures calculated using copy number ratio calls obtained from StepGram are useful empirical measures that can be used to assess whether reduced noise levels have an impact on the ability to detect regions with copy number aberrations:

1. median aberrant region length (MARL),
2. proportion of probes called aberrant (PPCA), and
3. median MAD across all cell lines after segmentation (MMAS).

The first two measures assess the resolution after segmentation. Reduced noise levels can have a significant impact on copy number ratio calling. Although the true aberration lengths are not known, it would be expected that if the MARL is reduced in the presence of a similar PPCA, the segmentation algorithm is able to better distinguish between regions of differing copy numbers. Furthermore, the MMAS is a useful measure of the dynamic range of the data since segmentation algorithms smooth the data and remove noise based on statistical algorithms and segmented data therefore represent biological information for the most part. While normalization methods aim to reduce noise levels, the biological information that is of interest should not be normalized out by any normalization procedure. Thus, after segmentation, the MAD, which is expected to be driven mostly by biological differences between disease samples and genomic references, should be of similar magnitude before and after normalization. Any significant reduction in MMAS indicates that the normalization procedure does not only remove noise due to technical variation but also biological interesting information.

2.2.4 Implementation

The approach outlined above is implemented in the **R** package `pcaCGH` available at www.people.vcu.edu/∼mreimers. Basic parallel computing capabilities were implemented to ensure efficiency. Currently, the NimbleGen Human Genome HG17 CGH 385K (remapped to HG19) and the Agilent CGH Microarray 44K 2005-03-16_HG17_WG_CGH (remapped to HG19) are supported. Output from Nimblescan v2.3.4 for the NimbleGen arrays and the processed intensities provided by Agilent's Feature Extraction Software v8.1.18 were used in this study. Additional chip types as well as output from the Nimblegen's Nimblescan and Agilent's Feature Extraction Software can be added by request.

2.3 Results

The proposed method was applied to the three datasets introduced in the previous section and evaluated each performance measure for the proposed method as well as the following existing methods:

1. median normalization, Bolstad *et al.* [14], **R** package `limma` version 3.4.5,
2. quantile normalization, Bolstad *et al.* [14], **R** package `limma` version 3.4.5,
3. popLowess, Staaf *et al.* [99], **R** package `popLowess` version 1.0.2, and
4. WACA, Lepretre *et al.* [46], **R** code provided by author.

Each method was applied as recommended by the authors. WACA was designed specifically for Agilent aCGH microarrays and therefore was not used with the NimbleGen arrays.

2.3.1 Spatial Artifacts

Stage 1 of the proposed algorithm is targeted at specific technical covariates such as melting temperature, reference channel artifacts, and spatial artifacts. To demonstrate the effectiveness of technical covariate LOESS normalization in removing spatial artifacts, Figure 2.2 shows log_2 intensity ratios of two arrays before and after LOESS normalization. It can clearly be seen that the majority of spatial artifacts are removed and the expected random patterns of amplifications and deletion are observable. Note that the ability of LOESS normalization to remove very sharp local biases such as the scratch seen on the right of Figure 2.2.a) is limited due to the limited flexibility of the algorithm at reasonable numbers of degrees of freedom. Although the effect of the scratch could be dampened, it could not be removed completely.

2.3.2 Derivative Log2Ratio Spread

Table 2.2 summarizes DLRS as well as mean efficiency estimates for each dataset and normalization method. The proposed pcaCGH approach reduces DLRS significantly and outperforms existing methods by up to an order of magnitude in terms of efficiency. Especially striking is the improvement over the commonly used quantile normalization approach for the Agilent data sets. While quantile normalization seems to increase the DLRS, the pcaCGH method achieves a similar reduction in DLRS as in the NimbleGen data set.

Table 2.2. Derivative Log2Ratio Spread: Comparison of DLRS estimates and mean algorithm efficiency

Dataset	Method	DLRS	95% CI	DLRSe	95% CI
Agilent	Processed Signal	0.176	(0.170,0.182)	N/A	N/A
	pcaCGH	0.110	(0.105,0.116)	0.366	(0.334,0.398)
	quantile	0.183	(0.176,0.190)	-0.041	(-0.059,-0.023)
	popLowess	0.171	(0.165,0.176)	0.030	(0.021,0.039)
	WACA	0.170	(0.164,0.175)	0.035	(0.026,0.043)
Agilent Dye Swap	Processed Signal	0.174	(0.167,0.182)	N/A	N/A
	pcaCGH	0.114	(0.109,0.118)	0.334	(0.301,0.368)
	quantile	0.180	(0.173,0.187)	-0.039	(-0.065,-0.001)
	popLowess	0.168	(0.161,0.175)	0.036	(0.030,0.043)
	WACA	0.166	(0.159,0.173)	0.045	(0.035,0.055)
NimbleGen	Raw Signal	0.155	(0.149,0.161)	N/A	N/A
	pcaCGH	0.104	(0.010,0.104)	0.316	(0.290,0.343)
	quantile	0.145	(0.138,0.151)	0.065	(0.042,0.087)
	popLowess	0.132	(0.124,0.140)	0.088	(0.067,0.110)

2.3.3 Improvement in Signal to Noise Ratios

Table 2.4 summarizes signal to noise ratios for each normalization method and data set as well as fold changes in signal to noise ratios with respect to the raw data. Signal to noise ratios for each array and its corresponding dye swap were calculated in a similar way. Table 2.3 lists fold changes in signal to noise ratios between Agilent's original and dye swap raw data and normalized data. The proposed pcaCGH method is the only approach that consistently improves signal to noise ratios significantly, well above improvements (if any) of existing methods.

Table 2.3. Mean fold change between signal to noise ratios for dye swaps

Method	Mean FC	95% CI
pcaCGH	1.87	(1.66,2.09)
median	1.01	(1.01,1.02)
quantile	0.96	(0.94,0.98)
popLowess	0.90	(0.83,0.96)
WACA	0.99	(0.96,1.01)

Table 2.4. Signal to noise ratios for technical replicates

Dataset	Cell Line	Method	S/N Ratio	Fold Change
Agilent	A549	Processed Signal	12.60	N/A
		pcaCGH	26.99	2.14
		median	12.67	1.01
		quantile	12.06	0.96
		popLowess	12.88	1.02
		WACA	14.34	1.14
Agilent Dye Swap	A549	Processed Signal	18.42	N/A
		pcaCGH	29.32	1.60
		median	20.10	1.09
		quantile	16.23	0.88
		popLowess	18.39	1.00
		WACA	20.11	1.09
NimbleGen	A549	Raw Signal	2.03	N/A
		pcaCGH	7.04	3.47
		median	5.42	2.67
		quantile	5.67	2.79
		popLowess	5.41	2.67
	SF-268	Raw Signal	3.11	N/A
		pcaCGH	8.15	2.62
		median	6.61	2.12
		quantile	6.78	2.18
		popLowess	6.57	2.11
	OVCAR-8	Raw Signal	5.07	N/A
		pcaCGH	11.25	2.21
		median	9.34	1.84
		quantile	8.40	1.65
		popLowess	8.14	1.61

2.3.4 Improvement in Agreement between Technologies

Table 2.5 shows the mean correlation across cell lines between log_2 ratio intensities from different platforms before and after normalization. While all normalization approaches improved correlations significantly, the pcaCGH approach outperformed existing methods easily. Furthermore, the variance in differences between corresponding probes from both platforms could be reduced by up to 45% while the next best method popLowess could only reduce the variance by up to 14% (see Table 2.6). Lastly, a similar increase in proportion of matching probes with the same detection call, i.e. matched pairs of probes from different platforms that were both called a deletion, amplification, or neither, was observed across normalization procedures (see Table 2.7).

Table 2.5. Agreement between technologies: Mean correlations between NimbleGen and Agilent probe intensities for normalization methods used

	Agilent/NimbleGen		Agilent DS/NimbleGen	
Method	Mean Correlation	95% CI	Mean Correlation	95% CI
Raw Data	0.08	(0.06,0.09)	0.07	(0.05,0.09)
pcaCGH	0.50	(0.46,0.53)	0.49	(0.45,0.52)
median	0.35	(0.32,0.38)	0.35	(0.31,0.38)
quantile	0.37	(0.34,0.40)	0.37	(0.34,0.39)
popLowess	0.39	(0.36,0.42)	0.39	(0.36,0.42)

Table 2.6. Variance explained in log_2 ratio intensity differences between technologies

	Agilent/NimbleGen		Agilent DS/NimbleGen	
Method	Var. Explained	95% CI	Var. Explained	95% CI
pcaCGH	0.12	(0.08,0.16)	0.13	(0.10,0.16)
median	0	N/A	0	N/A
quantile	0.06	(0.03,0.8)	0.07	(0.04,0.10)
popLowess	0.02	(0.01,0.4)	0.03	(0.01,0.05)

Table 2.7. Proportion of Probes in Agreement (PPA)

	Agilent/NimbleGen	Agilent DS/NimbleGen
Method	PPA	PPA
raw data	0.53	0.55
pcaCGH	0.72	0.72
median	0.72	0.73
quantile	0.76	0.76
popLowess	0.67	0.67

2.3.5 Improved Resolution of Called Segments

Figure 2.4 shows detection calls for parts of chromosome 17 of cell line SK-MEL-5 for the Agilent dataset where blue lines represent calls for the raw and green calls for normalized log_2 intensity ratios. It can be seen that regions of called copy number aberrations are shorter and more frequent indicating an increased resolution. Table 2.8 quantifies this first impression by listing the performance measures introduced in Section 2.2.3. The shorter median aberrant region lengths after normalization indicate that along with the expected long regions of copy number aberrations, there are also many more short regions present in cancer cell lines that have not been detected due to low signal to noise ratios in the raw data and that cannot be picked up by normalizing with existing methods. Further evidence of improved resolution can be seen in Figure 2.5, which shows density plots of intensity ratios before and after normalization for cell line OVCAR-8 in the Agilent dataset. While only two clear peaks are distinguishable in pre-normalized data, three clear peaks that are significantly sharper are present in post-normalized data.

Furthermore, the MMAS estimates for segmented data normalized by the proposed pcaCGH method are not significantly smaller than those from either segmented raw data or those from segmented data normalized with existing method. The lack of reduction in dynamic range together with similar proportions of probes called aberrant across normalization methods is strong evidence that the pcaCGH approach does not remove biological relevant information.

Table 2.8. Improvement in Resolution: Median aberrant region length (MARL), proportion of probes called aberrant (PPCA), and median MAD across all cell lines after segmentation (MMAS)

Dataset	Method	MARL	PPCA	MMAS
Agilent	Processed Signal	123	0.74	0.2815
	pcaCGH	63	0.76	0.3041
	median	122	0.73	0.2819
	quantile	163	0.76	0.2648
	popLowess	96	0.59	0.3011
	WACA	115	0.70	0.2497
Agilent Dye Swap	Processed Signal	108	0.74	0.2959
	pcaCGH	61	0.76	0.2617
	median	109	0.72	0.2846
	quantile	139	0.71	0.2532
	popLowess	83	0.59	0.2865
	WACA	95	0.69	0.2356
NimbleGen	Raw Signal	277	0.76	0.1621
	pcaCGH	92	0.57	0.1779
	median	245	0.57	0.1695
	quantile	197	0.59	0.1888
	popLowess	185	0.55	0.1789

Figure 2.4. Visualization of Pre- and Post-normalized Data: Raw and Normalized Data for a Segment of Chromosome 17 of Cell Line SK-Mel-5

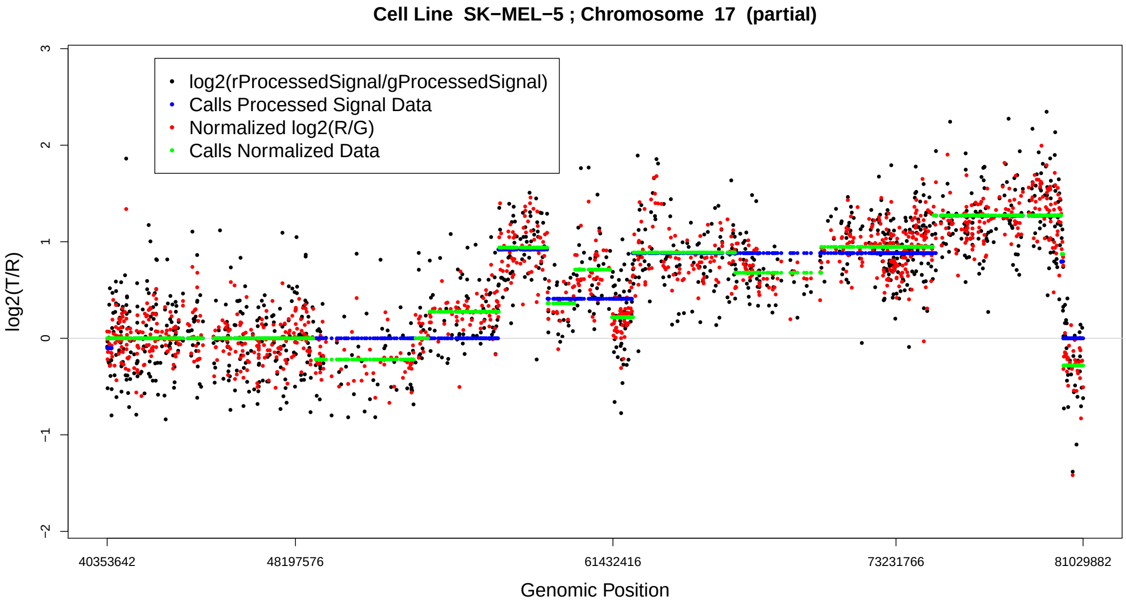
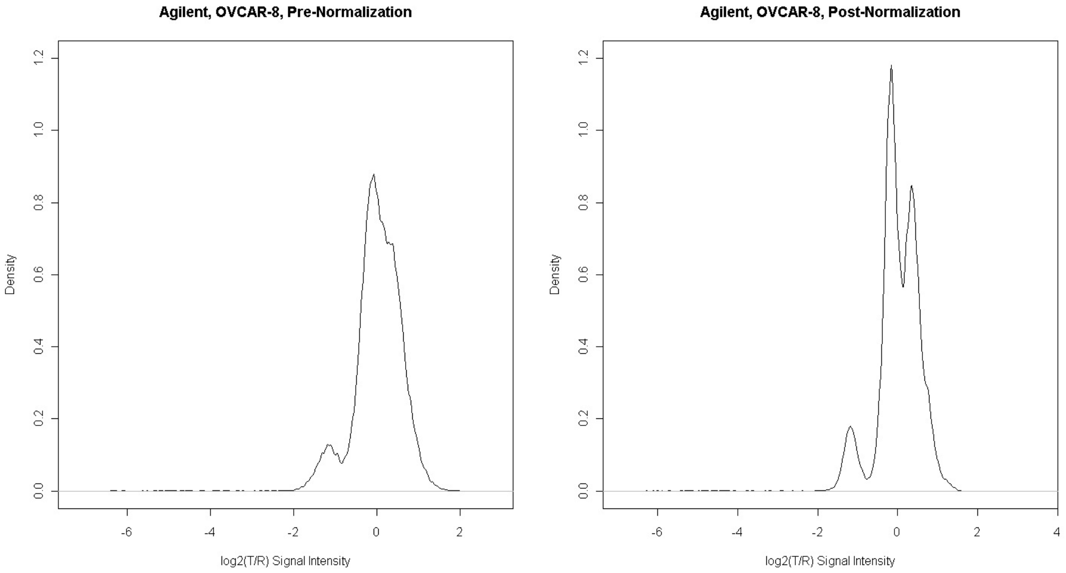


Figure 2.5. Density Plots of Signal Intensities Before and After Normalization: Density plots for cell line OVCAR-8 are shown pre- and post-normalization.



2.3.6 Interpretation and Utility of LOESS Regression and Principal Components

To illustrate the utility of the LOESS regression and the PCA step in the proposed algorithm, the loadings of the strongest principle component recovered from each data set and their correlation to GC content as a known technical covariate were investigated. To that end, the PCA step was applied directly to the raw NimbleGen data without performing LOESS normalization first. Figures 2.6 and 2.7 show each probe's loading against its GC content recovered from the raw and LOESS normalized NimbleGen data, respectively. The corresponding correlations between loadings and GC content were 0.247 and 0.071, respectively, indicating that the LOESS normalization does reduce systematic variability due to technical covariate and also that the PCA step does pick up that same variability if not already removed in step 1 of the algorithm. These observations and the previous work on PCA like methods beg the question of whether the LOESS step is needed at all. This was addressed by computing DLRS and DLRSe measures without the initial LOESS step. The DLRS and DLRSe for the normalized data obtained from applying the PCA step directly to the raw NimbleGen data is 0.136 and 0.111 while the DLRS and DLRSe of the PCA step applied to the LOESS normalized data is 0.104 and 0.316, respectively. This shows that both LOESS regression and PCA contribute substantively to reducing systematic variability. It was also of interest to verify that neither the LOESS regression nor the PCA step are picking up significant amounts of biological signal. If that were the case, one would expect that high principal component loadings would cluster on a few genomic positions. Figures 2.8 and 2.9 show the loadings of the first principle components by genomic position for Agilent and NimbleGen data while Figure 2.10 shows the LOESS residuals against genomic position for the NimbleGen data.

None of these three figures do show chromosome location patterns in the distribution of the loadings suggesting that biological signal is not a major component of the variance removed.

Note that the loadings were highly reproducible for the replicated Agilent data sets with a correlation of 0.647 (see Figure 2.11). Furthermore, the NimbleGen data set was randomly divided into two sets of 36 arrays and performed the PCA step on both data sets separately. The resulting loadings of the first principal components were moderately correlated with a correlation of 0.384 (see Figure 2.12).

Figure 2.6. Scatter plot of each probe's first principle component loading against its GC content where the PCA step was applied directly to the raw NimbleGen data. The correlation between the PC loadings and GC content is 0.247.

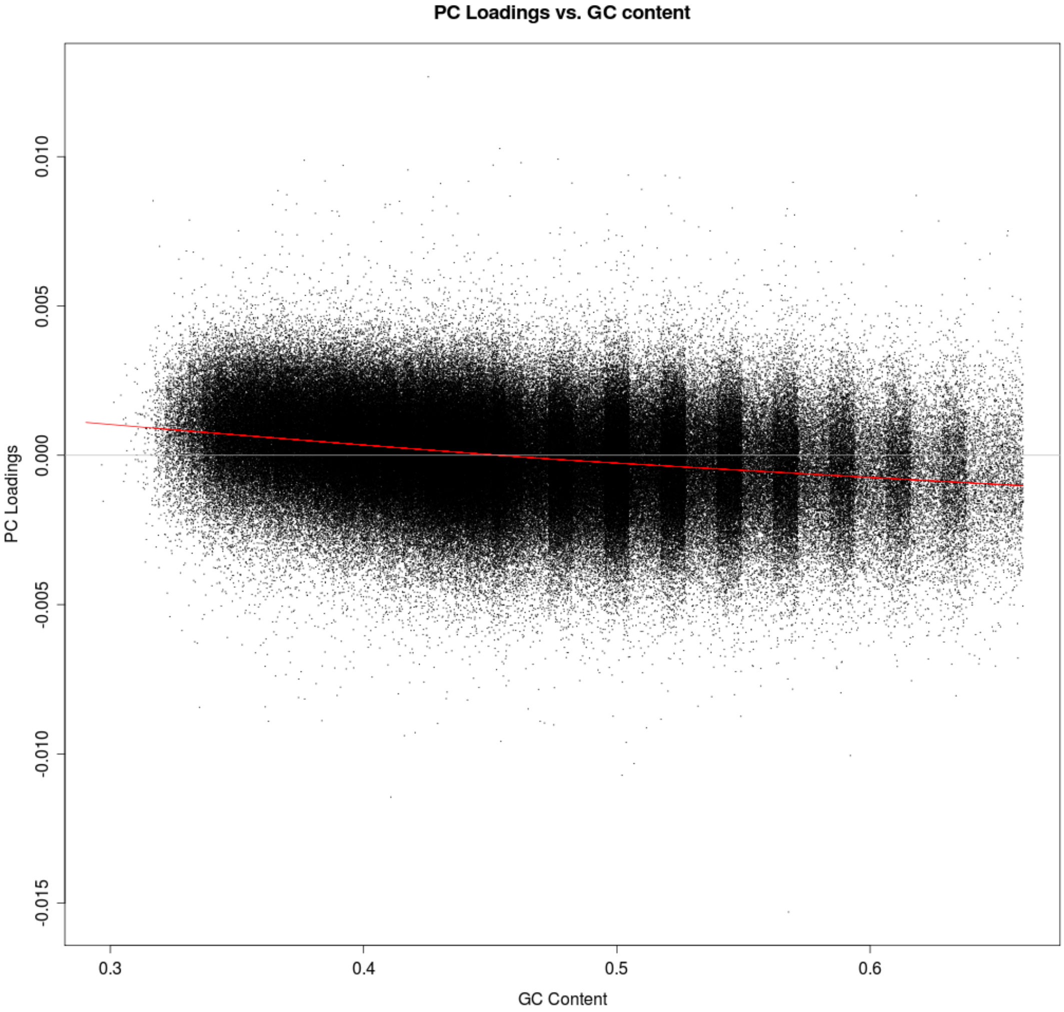


Figure 2.7. Scatter plot of each probe's first principle component loading against its GC content where the PCA step was applied to the LOESS normalized NimbleGen data. The correlation between the PC loadings and GC content is 0.071.

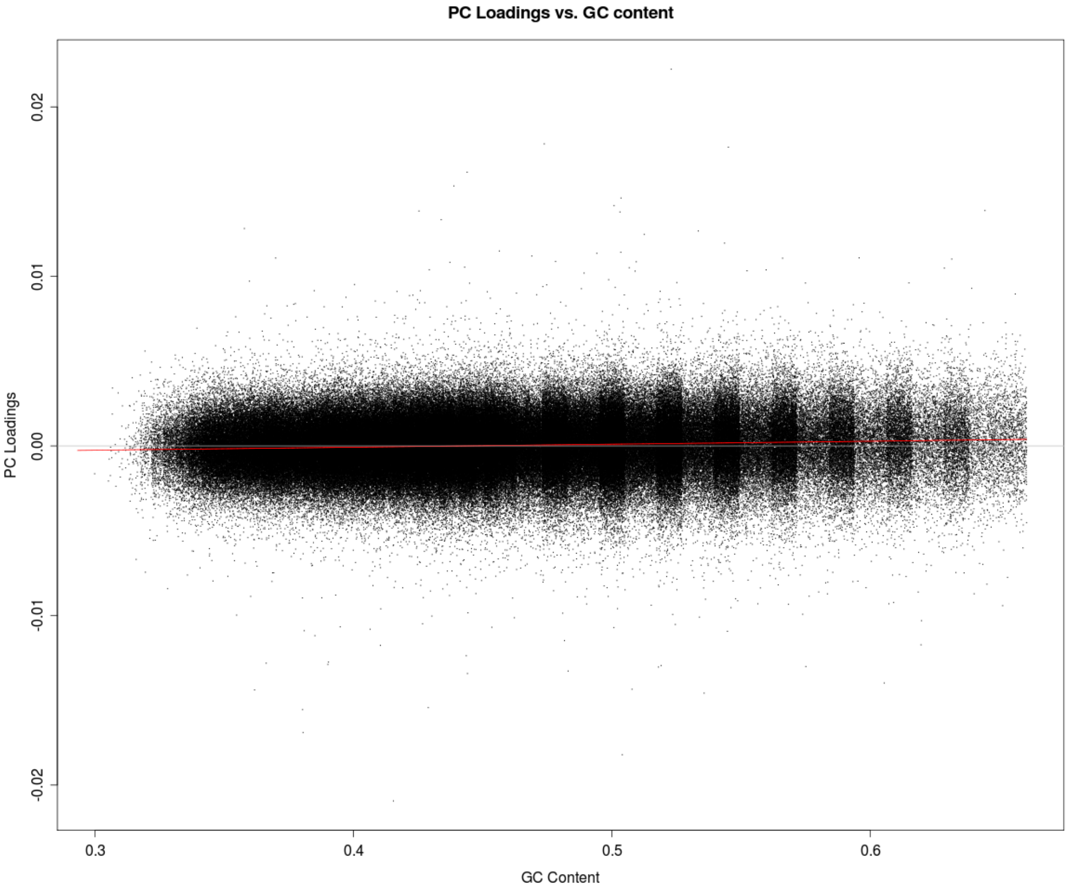


Figure 2.8. Scatter plot of each probe's first principle component loading against its genomic position for the Agilent data set.

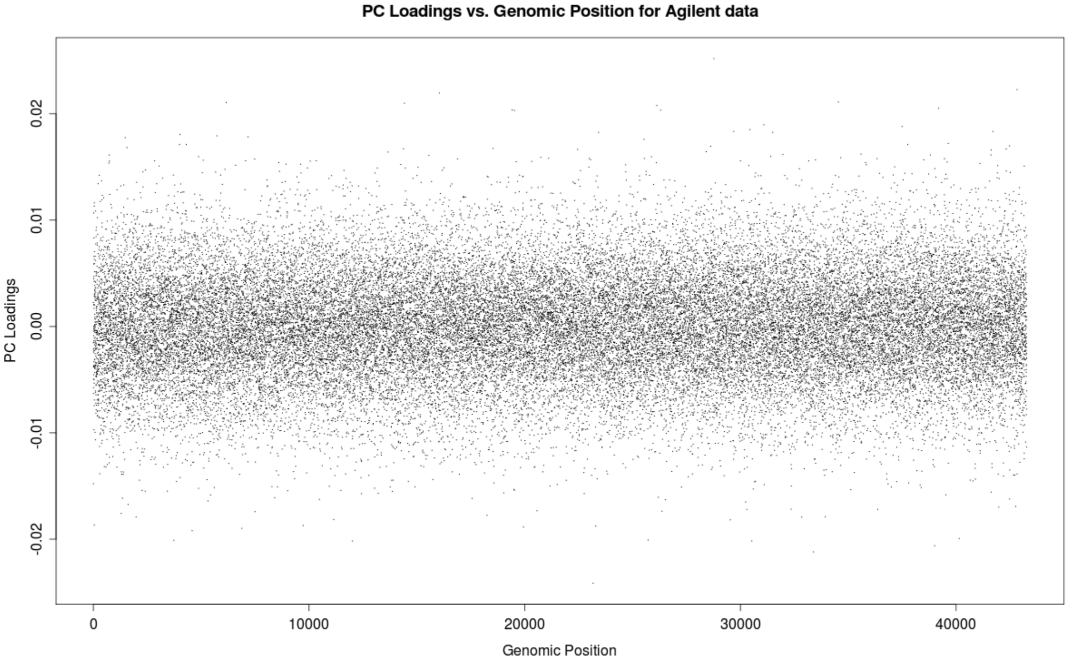


Figure 2.9. Scatter plot of each probe's first principle component loading against its genomic position for the NimbleGen data set.

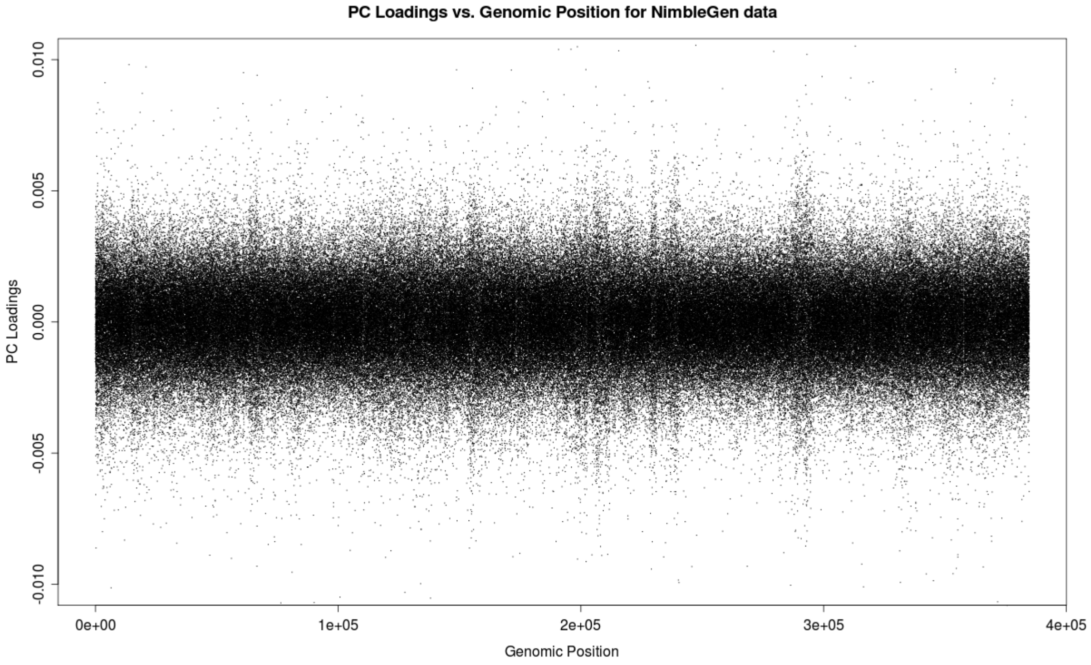


Figure 2.10. Scatter plot of each probe's LOESS residual against its genomic position for the NimbleGen data set.

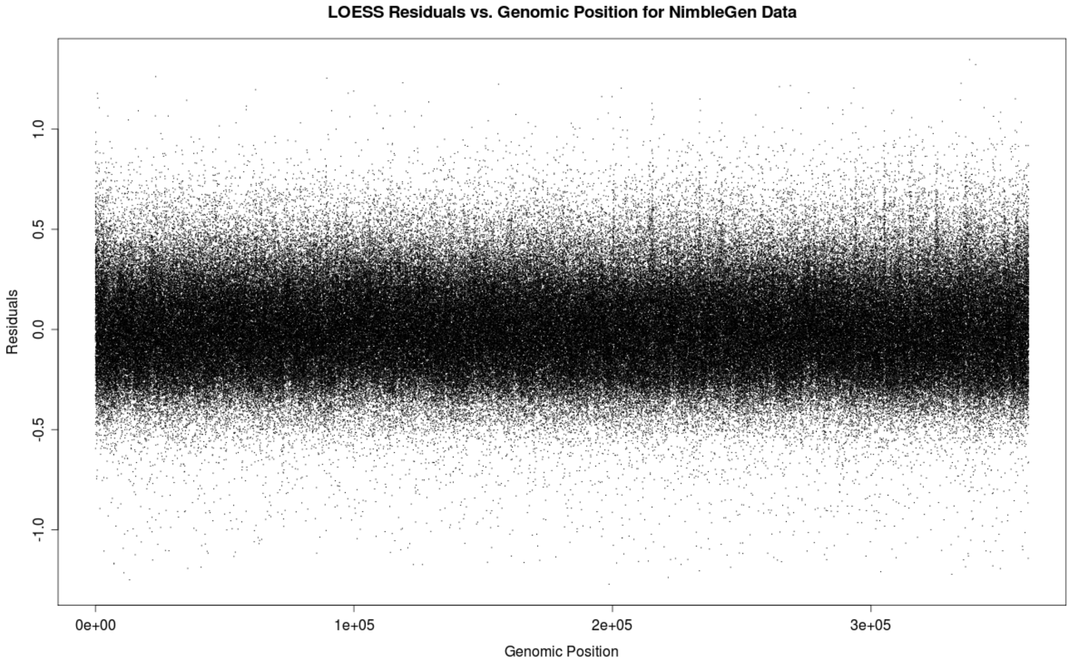


Figure 2.11. Scatter plot of each probe's first principle component loading for Agilent and Agilent Dye Swap data sets. The correlation between the loadings is 0.647.

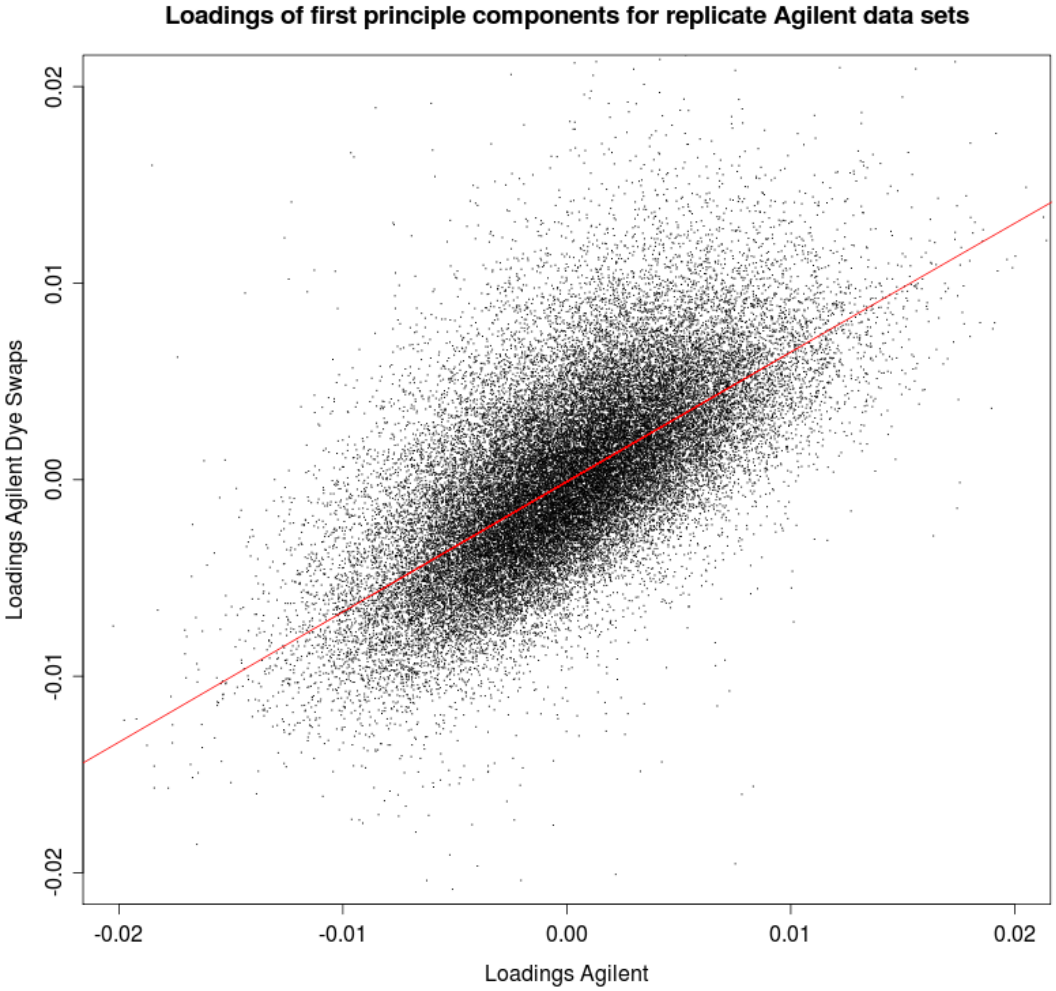
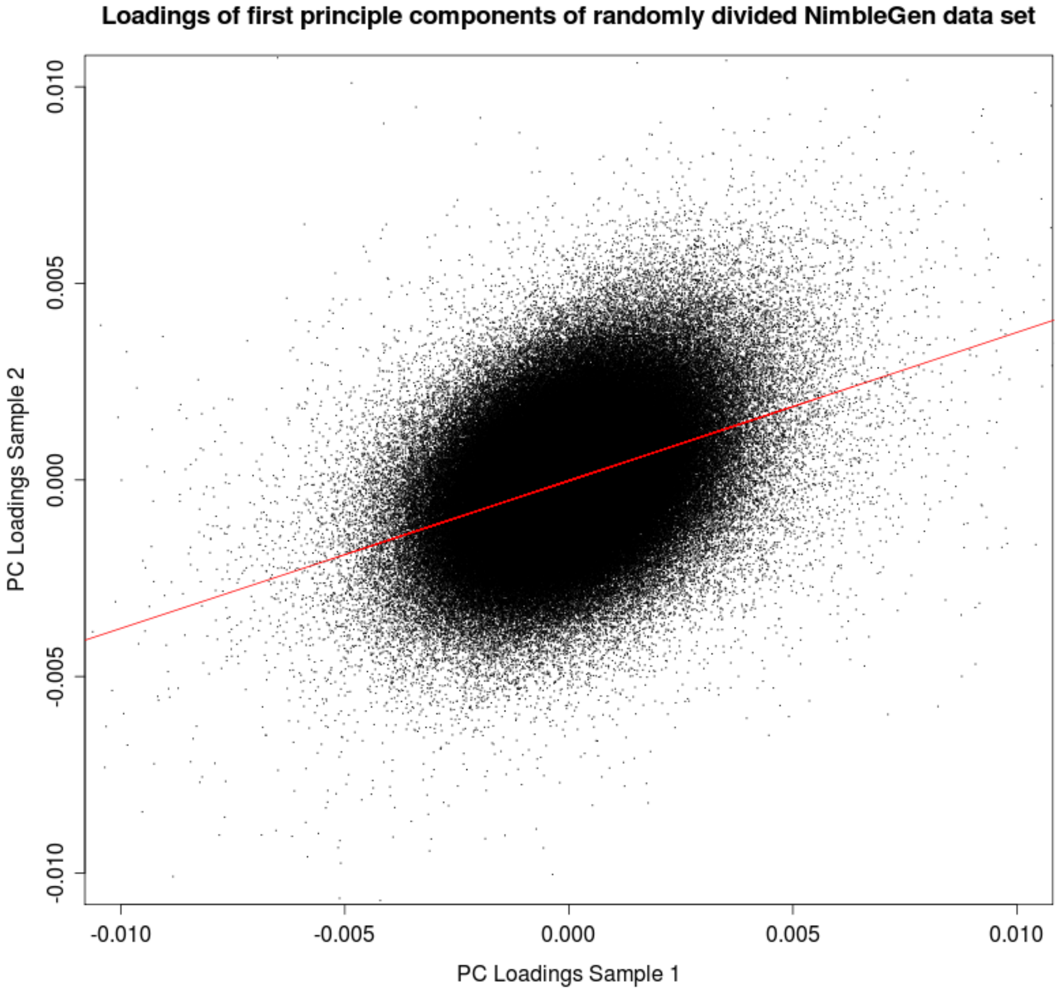


Figure 2.12. Scatter plot of each probe's first principle component loadings for randomly divided NimbleGen data sets (36 arrays each). The correlation between the loadings is 0.384.



2.4 Discussion

Detecting copy number aberration has become an integral part of uncovering the underlying processes of many genetic diseases and is becoming more prevalent in personalized patient care [65]. Array CGH technology has enabled researchers to investigate copy number aberrations across the whole genome in a high throughput fashion with an improved resolution, a higher dynamic range, and a convenient way to access the data. While large fold changes can be easily detected using standard methods to analyze the raw data, many scientists suspect that shorter and more subtle copy number changes play an important role in many genetic disorders. Microarray experiments involve numerous complex procedures including DNA extraction, DNA hybridization, and image scanning, that contribute to non-biological variation. Standard methods do not adjust for the technical error sources and therefore lack the resolution to reliably detect those subtle changes and thus more sophisticated approaches are needed.

A method called pcaCGH was proposed to normalize aCGH data using technical covariates and a robust PCA. Qualitative and quantitative evidence showing the efficiency of the proposed algorithm was presented. Furthermore its performance was compared to existing methods commonly used for normalization of aCGH data and it was shown that the pcaCGH approach significantly improves on those methods.

A good measure of how well a normalization method performs is how well the biological differences stand out above the 'noise' of differences due to factors other than biology. This 'technical noise' is believed to be mostly systematic bias due to variation in technical aspects of processing the samples. The technical noise due to variation in technical aspects of sample processing is well measured by the differences between replicate arrays which measure the same sample relative to the differences between arrays measuring different samples, which are compounded of biological differences and technical noise. Therefore the ratio of the variance across biologically distinct samples, relative to the variance between replicate samples, is a good measure of how well a normalization method has succeeded in its goals. Note that by construction the procedure reduces overall variance in a dataset; however it is not believed that the proposed procedure simply reduces variation. It was shown that the ratios of variance attributable to combinations of biological and technical variation, to the differences attributable solely to technical variation, are significantly better after applying the proposed algorithm than without normalization as well as better than after applying existing normalization procedures.

Using this metric on technical replicates as well as dye swap replicates, it was demonstrated that the proposed method significantly increases signal to noise ratio in comparison to existing methods. The increase in signal to noise ratio seem to lead to a higher resolution, i.e. the ability to detect smaller copy number changes, which is apparent when comparing median lengths of aberrant regions. Thus the proposed algorithm allows for a more detailed picture of copy number structure across the whole genome. Furthermore it was shown that while technical noise is reduced significantly the dynamic range is preserved.

Another strong argument for the proposed algorithm is an improved agreement across platforms. To date, few previous analyses have investigated concordance between different aCGH technologies. Compelling evidence was presented that the algorithm not only increases signal to noise ratios significantly, but also notably improves agreement between Agilent and NimbleGen data well above above improvements existing methods can achieve.

Chapter 3

Assessment of Current Methods for Analyzing RNA-Seq Studies

3.1 Introduction

As mentioned in the introduction, the decrease in cost per Megabase sequenced and the increase in throughput let to a significant increase in NGS datasets. The following two chapters focus on the analysis of quality assured data obtained from RNA-Seq studies investigating transcription levels through some form of experimental design. The main advantage of RNA-Seq over microarrays for gene-expression studies is a higher dynamic range as the only limiting factor for genes expressed at low levels is the number of total reads obtained from the mRNA sample, which correlates to the total cost of a study, while the lack of background noise due to cross hybridization as in microarray experiments allows detection of weaker signals [106]. Furthermore, scientists do not have to rely on probe annotations supplied by manufacturers anymore since the sequence of cDNA molecules derived from the mRNA sample is obtained directly and can then be aligned to the scientist's reference sequence or gene model of choice. Theses advantages have lead to numerous RNA-Seq studies on a variety of organisms and cell types [24]. Scientists most commonly look for differences in transcription levels for genes, where a gene is defined as the union of its exons and this definition will be adopted for the following two chapters. However, the methods described to analyze gene-level data could also be applied to exon-by-exon analyses or any other analysis that summarizes reads aligning to a pre-specified genomic region into count data as demonstrated in Chapter 5.

A natural choice to analyze count data is the Poisson distribution. Marioni et al. [51] first investigated the properties of RNA-Seq data and concluded that the Poisson distribution is suitable to describe count data obtained from sequencing technical replicates, i.e. sequencing the same RNA sample repeatedly. However, it became clear very quickly that the Poisson distribution could not explain extra variation seen in the data when sequencing biological replicates, i.e. sequencing RNA samples extracted from the same tissue but from separate individuals [16, 60, 86]. A natural extension to the Poisson distribution is the negative binomial (NB) distribution that models extra variation above that expected from a Poisson distribution through an additional parameter called the overdispersion parameter. Two methods, *DESeq* by Anders et al. [3] and *edgeR* by Robinson et al. [87], based on the NB distribution have been adopted by the scientific community as preferred approaches to analyze RNA-Seq data. Both approaches use some form of information sharing across genes (see Sections 3.2.1 and 3.2.2) and were initially developed to simply detect differences between two groups of samples, e.g. RNA samples from individuals with a specific disease and normal controls. As experimental designs have recently become more complex, these methods were recently extended to handle potential confounders in addition to the covariate of interest. To date, there has not been an evaluation of these extensions' performance with regards to type I error control and power and the purpose of this chapter is to extensively test DESeq and edgeR across a wide range of scenarios and data structures through simulations and application to real data. The existing methods are also compared to a newly proposed method to introduced in Chapter 4.

3.2 Methods

The following sections summarize briefly the approaches by Anders and Huber [3], implemented in the R package DESeq, and Robinson et al. [87], implemented in the R package edgeR, and also outline the strategy to evaluate those methods. R version 2.14 (released on 10/31/2011), DESeq version 1.6.0 and edgeR version 2.4.0 were used to obtain the results presented in this study. Standard settings as described in the package's vignettes were used and the an example of the R code used to obtain presented results can be found in Appendix C. Both methods are currently considered the "gold-standard" in analyzing RNA-Seq data and were evaluated by performing extensive simulation studies under a variety of scenarios as well as applying the methods to publicly available datasets and data obtained from a RNA-Seq experiment conducted at Virginia Commonwealth University.

3.2.1 edgeR

Robinson et al. have developed their software edgeR [87] to analyze count data from high throughput sequencing studies based on previous papers by Smyth and Verbyla [97] as well as Robinson and Smyth [86]. Their model is based on the negative binomial (NB) distribution with probability mass function (pmf)

$$f_{NB}(y|r,p)\binom{y+r-1}{k}(1-p)^r p^k, \tag{3.1}$$

for $k = 0, 1, 2, \ldots$, $p \in (0,1)$ and $r > 0$. Here $E(Y) = \mu = \frac{pr}{1-p}$ and $Var(Y) = \sigma^2 = \frac{pr}{(1-p)^2}$. The distribution can be re-parameterized in terms of μ and σ^2:

$$f_{NB}(y|\mu, \sigma^2) = \binom{y + \mu^2/(\sigma^2 - \mu) - 1}{\mu^2/(\sigma^2 - \mu) - 1} (\mu/\sigma^2)^{\mu^2/(\sigma^2-\mu)} (1 - \mu/\sigma^2)^y, y \geq 0, \mu \geq 0, \sigma^2 > 0. \tag{3.2}$$

Note that the NB distribution is equivalent to the Poisson distribution when $\sigma^2 = \mu$. Furthermore, let $l_{NB}(\mu, \sigma^2|y)$ denote the likelihood function of the NB distribution. The mean value μ_{ij} of the observed counts for gene i, $i = 1, \ldots, I$, and sample j, $j = 1, \ldots, J$, is parameterized as

$$\mu_{ij} = q_{i,\rho(j)} s_j, \tag{3.3}$$

where $q_{i,\rho(j)}$ is proportional to the expected value of the true (but unknown) concentration of fragments from gene i under condition $\rho(j)$ and s_j represents a normalizing factor based on the total number of reads from sample j compared to total number of reads from the other samples. The authors define the variance the commonly used parameterization $\sigma^2_{ij} = \mu_{ij} + \theta_i \mu^2_{ij}$ where θ_i is called the dispersion parameter. Assuming equal library sizes for all samples, the authors estimate the dispersion θ_i for gene i by weighted conditional maximum likelihood:

$$WL(\theta_i) = l_i(\theta_i) + \alpha l_C(\theta_i), \tag{3.4}$$

where θ_i is the genewise conditional log-likelihood derived in [86] and $l_C(\theta_i) = \sum_{i=1}^{I} l_i(\theta_i)$ is the common likelihood over all genes. Since the assumption of equal library sizes is unattainable for real HTS studies, Robinson et al. use a method called quantile adjusted conditional maximum likelihood (qCML). Quantile-adjusted CML uses an iterative algorithm to estimate θ_i that adjusts observed counts as if all observations come from a $NB(q_{i,\rho(j)} s, \sigma^2_{ij})$ distribution where s is the geometric mean of the library sizes s_j, i.e.

$s = \left(\prod_{j=1}^{J} s_j\right)^{\frac{1}{J}}$.

The parameter α, estimated using an empirical Bayes procedure, determines how much a gene's dispersion estimate based on its counts alone is shrunken towards a common dispersion estimate $\hat{\theta}_C$ obtained by maximizing l_C. If $\alpha = 0$ then the likelihood of that gene's data is maximized while if α is sufficiently large, the estimate for θ_i is close to the estimate of a common dispersion θ_C. The Bayes procedure assumes that $\hat{\theta}_i|\theta_i \sim N(\theta_i, \tau_i^2)$ and $\theta_i \sim N(\theta_0, \tau_0^2)$. The author's strategy to estimate α relies on choosing α such that $WL(\theta_i)$ coincides with an empirical Bayes rule using the posterior mean estimator of θ_i:

$$\hat{\theta}_i^B = E(\theta_i|\hat{\theta}_i) = \frac{\hat{\theta}_i/\tau_i^2 + \theta_C/\tau_0^2}{1/\tau_i^2 + /\tau_0^2}, \tag{3.5}$$

where the hyperparameters θ_0 and τ_0^2 can be estimated from the marginal distribution of $\hat{\theta}_i$. For a detailed derivation of the approach, please refer to the author's original papers. Robinson's method has been extended to generalized linear models (GLMs) [62] using Cox-Reid approximate conditional inference [20] to estimate dispersion and use those values to fit NB models to each gene using their own fitting procedure. GLMs are used to model the relationship between mean μ_{ij} and explanatory variables through a link function g:

$$g(\mu_{ij}) = g(q_{ij}s_j) = \beta_{i0} + \beta_{i1}x_{j1} + \ldots + \beta_{ip}x_{jp} = \boldsymbol{x_j}\boldsymbol{\beta_i}, \tag{3.6}$$

where $X_1, \ldots, X_p$ represent either covariates of interest or confounders that need to be adjusted for. Note that the rate q_{ij} for gene i and sample j does now not only depend on conditions $\rho(j)$ as proposed originally by Robinson et al. in Equation 3.3, but rather on the values of all p covariates of that sample. The most commonly used link function is the log link, which is used as link function of choice for all models presented hereafter in this work. The software supplies p-values for testing differential expression based on standard likelihood ratio tests (LRTs).

3.2.2 DESeq

Anders et al. [3] also based their work on the NB distribution described in Equation 3.2 where $\mu_{ij} = q_{i,\rho(j)}s_j$ is similarly defined as in edgeR. The variance σ^2_{ij} however is defined by the authors as a sum of a "shot noise" term and a raw variance term:

$$\sigma^2_{ij}(q_{i,\rho(j)}) = s_j q_{i,\rho(j)} + s_j^2 \nu_\rho(q_{i,\rho(j)}) = \mu_{ij} + s_j^2 \nu_\rho(q_{i,\rho(j)}). \tag{3.7}$$

Note that the term shot noise is not used properly here as it not depend on magnitude of the actual signal, which in this case is $q_{i,\rho(j)}$. Furthermore note that Anders et al. assume that the per-gene raw variance parameter $\nu_{\rho(j)}$ is a smooth function of q_i and $\rho(j)$ and consequently the variance σ^2_{ij} is a function of these parameters. This assumption is designed to obtain more precise estimates of the variance for gene i using data from genes with similar gene expression. In general, Anders et al. fit the smooth function $\nu_{i,\rho(j)}$ empirically by first estimating $q_{i,\rho(j)}$ and σ^2_{ij} and then fitting a smooth curve through those estimates. The smooth curve was first fit using a local regression but now uses a parametric form as default fitting procedure. Specifically, the parametric form is given by

$$\sigma^2_{ij}(q_{i,\rho(j)}) = a_0 + \frac{a_1}{s_j q_{i,\rho(j)}}, \tag{3.8}$$

where a_0 and a_1 are estimated using a robust gamma-family GLM.

Their algorithm to fit the proposed model and estimate all parameters is described in great detail in [3] using the functional relationships described in Equations 3.7 and 3.8. Two options are available in `DESeq` to fit the mean-variance relationship proposed in Equation 3.7. The first approach estimates the variance for each gene without taking into account each sample's group membership, i.e. $\nu_{i,\rho(j)} = \nu_i$, enabling this approach to be applied to data without biological replicates. The second approach first calculates variance estimates within each condition and then pools the estimates across conditions taking into account biological variation. The two approaches are labeled ”blind” and “pooled” dispersion estimation.

Using Equation 3.6, Anders et al. have since extended both approaches for use with GLMs using a custom negative binomial family implementing Equation 3.7 in conjunction with R's `glm` function. DESeq also offers an ad hoc adjustment to dispersion estimates called ”sharing” modes. The first mode labeled *fit-only* uses the variance estimate obtained by using the proposed algorithm to estimate the variance parameter σ_{ij}^2. The second mode labeled *maximum* estimates σ_{ij}^2 using the proposed method as well as the variance estimator used to fit $\nu_{i,\rho(j)}$ and proceeds using the maximum of the two variance estimates in subsequent analyses. Note that there does not seem to be a theoretical justification why this adjustment is needed, but rather relies on empirical observations made by the authors that using sharing mode fit-only can lead to false positives as mentioned in the help file for the package. Two methods were evaluated and will be referred to as *DESeq Liberal* and *DESeq Conservative*, respectively. DESeq Liberal uses “blind“ dispersion estimation and sharing mode fit-only while DESeq Conservative uses pooled dispersion estimation and sharing mode maximum. The software supplies p-values for testing differential expression based on standard LRTs.

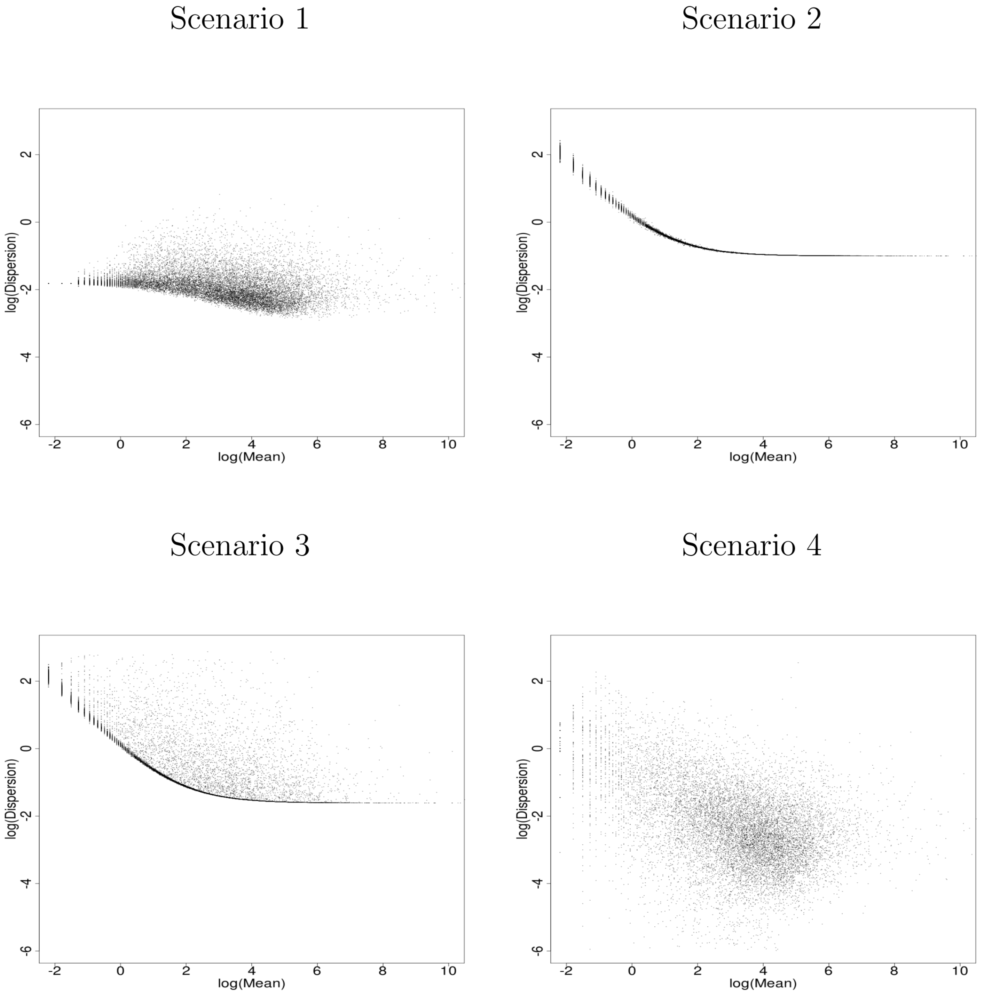
3.2.3 Simulation Studies

The simulation studies were designed to test the methods under realistic scenarios. To that end, two publicly available datasets were used. The first contained RNA-Seq count data on humans, chimpanzees, and rhesus macaques using liver RNA samples from three males and three females from each species [12] and was downloaded from the NCBI Gene Expression Omnibus [6] under accession number GSE17274. The dataset contained counts for 17,254 genes. Count data for 12,410 genes expressed in human B-cell RNA samples of 17 females and 24 males sequenced by Cheung et al.[18] was downloaded from the ReCount database [24]. Since we do not know the true underlying distribution of the dispersion parameters, the datasets were used to obtain four sets of dispersion estimates using the following algorithms:

1. edgeR's tagwise Cox-Reid dispersion estimation algorithm,
2. DESeq's blind dispersion estimation algorithm with sharing mode fit-only,
3. DESeq's pooled dispersion estimation algorithm with sharing mode maximum, and
4. maximum likelihood estimation using Ripley's and Venables' *theta.ml* R function [105].

The four sets of dispersion estimates were then taken to be the true underlying distribution of dispersion parameters and together with the estimated mean count for each gene were used to create four simulation scenarios. Figures 3.1 and 3.2 illustrate the relationship between mean-dispersion relationship for the four scenarios. By design, DESeq prescribes a functional relationship between mean and dispersion while the correlation between mean and dispersion for edgeR and ML dispersion estimates is significantly weaker.

Sixty-four datasets each with sample size 18 for the Marioni data and sample sizes of 41, 18, 15, 12 and 9 for the Cheung data were simulated using the four simulation scenarios described above. To assess power, 10% of genes were randomly chosen to have log fold changes drawn from a $N(0, 2)$ distribution and those fold changes were used to generate gender differences. For each dataset, the three methods were used to obtain p-values testing for gender differences. The proportion of genes with true gender differences



Scenario 1

Scenario 2

Scenario 3

Scenario 4

Figure 3.1. Dispersion estimates against mean gene count on a log scale for the Marioni dataset

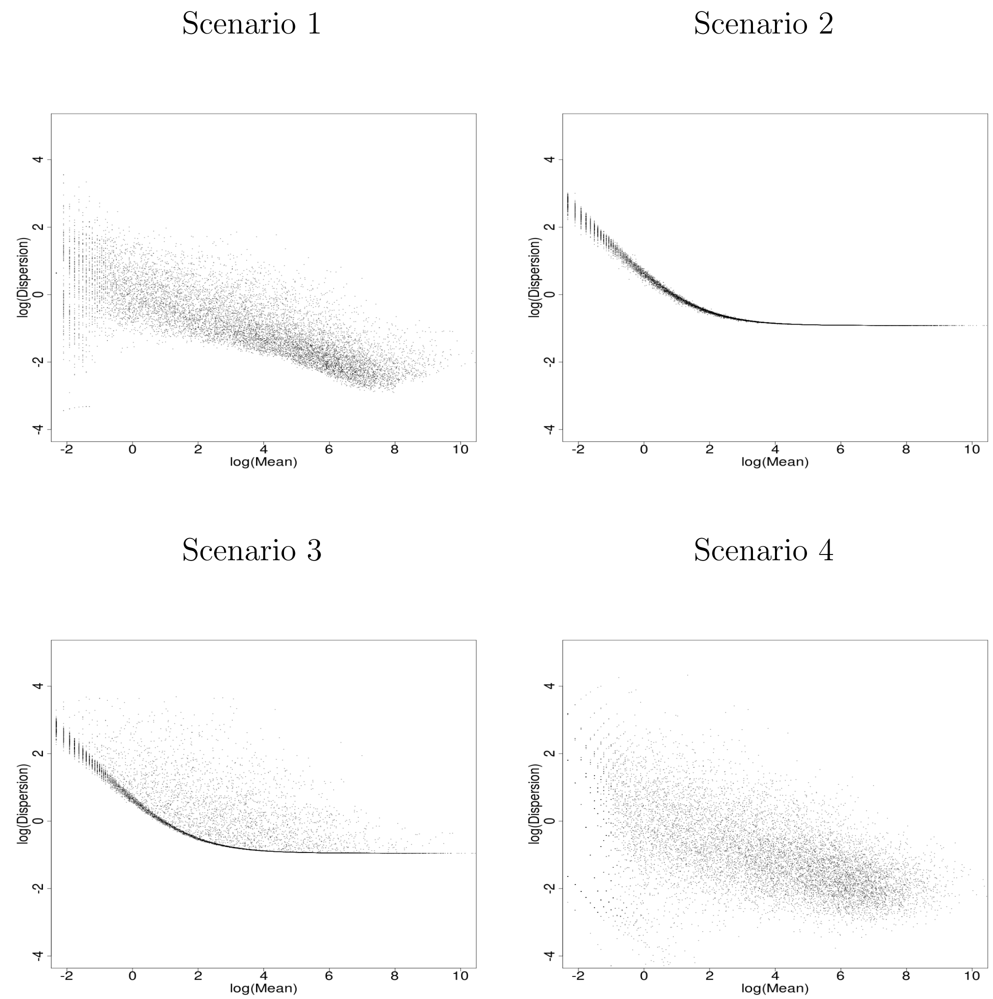


Figure 3.2. Dispersion estimates against mean gene count on a log scale for the Cheung dataset

called significant at a 10% false discovery rate (FDR) after Benjamini-Hochberg multiple comparison correction [7] was used as an estimate of power. Additionally, a categorical covariate with three levels distributed evenly across gender was included in each model and was used to assess type I error. The size of each test at a 0.001 level was calculated as the proportion of genes with p-values smaller than 0.001 where genes for which the models did not converged were disregarded. Furthermore, the number of genes called significant at a 10% false discovery rate (FDR) after Benjamini-Hochberg multiple comparison correction was recorded.

3.2.4 Application to Internal and Publicly Available Datasets

Gender-specific gene expression has been studied extensively over the last years [66, 95, 104, 110, 112]. These studies have shown that the vast majority of genes showing evidence for gender-specific gene expression are located on chromosome Y, few are located on chromosome X, and rarely any are located on autosomal chromosomes. Therefore testing for gender-specific gene expression is well suited to compare methods in terms of false positive rate and power on real data.

In addition to the Marioni primate dataset and the Cheung HapMap dataset, one additional publicly available dataset and one additional internal dataset with phenotypic information on gender were used. The public dataset is a subset of the Marioni primate dataset that was re-aligned and summarized by Frazee at al. [24] and is available from the Recount database under the name “Gilad“. This dataset contains count data on 10,525 genes for three male and three female samples. The internal dataset from Xiangning Chen’s lab at Virginia Commonwealth University contained count data on 21,134 genes for 82 sequenced brain samples from normal controls (n=26) and patients with bipolar disorder (n=25) or schizophrenia (n=31). The Chen data included information on age, brain pH, and post mortem interval (PMI), which were used as covariates during model fitting in addition to gender and diagnostic group. Furthermore, to generate a dataset with smaller sample size, a subset was taken from the Cheung dataset by randomly choosing ten male and ten female samples.

3.3 Results

Supplementary Table A.1 and Figure 4.2 show average size estimates at the 0.001 significance level as well as average and maximum number of genes called significant at a 10% FDR cutoff over 64 simulated datasets. It can be seen that edgeR and DESeq Liberal only control the size of the test under one scenario. In the worst case, size estimates for edgeR and DESeq Liberal are up to 3-fold and 8-fold over nominal size. This leads to a significant number of false positives when applying a FDR procedure as shown in Supplementary Table A.2 and Figure 4.3. To further evaluate the reason for the increased number of false positives, the dispersion estimates were compared to the true dispersion values used to simulate the data. Figures 3.3.a and 3.3.b show dispersion estimates from edgeR and DESeq Liberal, respectively, against the true dispersion values in blue on a log scale over 64 simulations for scenario 4. Red points indicate the average dispersion estimate for a specific gene while green points indicate average dispersion estimates for genes falsely identified as differentially expressed. The plots show that for the overwhelming majority of genes falsely called differentially expressed the dispersion estimate was smaller than the true dispersion value. It was also of interest whether false positives occur only for genes expressed at high levels, low levels or a mixture of both. Figure 3.4 shows true dispersion values against mean gene count in black on a log scale for scenario 4. Red points indicate mean dispersion estimates obtained from DESeq Liberal over 64 simulations while magenta points indicate false positives, which seem to occur for the entire range of mean gene counts. The size estimates for DESeq Conservative were found to be below nominal level for the majority of scenarios and sample sizes. However, for scenario 4 and small sample sizes, size estimates were increased 2-fold above nominal level. Similar observations to those made in the simulation studies could be made when both methods were used to test for gender-specific gene expression in real data. Table 4.4 shows the number of genes called differentially expressed at a 10% FDR cutoff after Benjamini-Hochberg multiple comparison correction. DESeq Liberal identifies a significant number of genes not located on chromosome Y and also identify genes on chromosome X that are not known to show gender-specific expression for all datasets while edgeR performs well for the Chen data, but not for the remaining datasets. DESeq Conservative is more conservative, but also identifies a number of false positives for the Gilad dataset, which features a small sample size.

3.4 Conclusions

The evaluation of methods using moderation when estimating the dispersion parameter for a NB distribution has shown that DESeq Liberal and to a lesser extend edgeR have significant problems to control type I error under certain scenarios. An explanation for increased false positive rates was given by showing that these methods underestimate the dispersion parameter for a subset of genes. DESeq Conservative is conservative for the majority of scenarios and sample sizes tested, but also tends to increased false positive rates when sample sizes are small. In Chapter 4 a new algorithm will be proposed that does not use moderation and can handle categorical and continuous covariates. It will be shown that this algorithm maintains size under a variety of scenarios and has comparable or better power for sample sizes of twelve or larger when compared to existing methods. The results obtained in this chapter will be put into further perspective in Section 4.4.

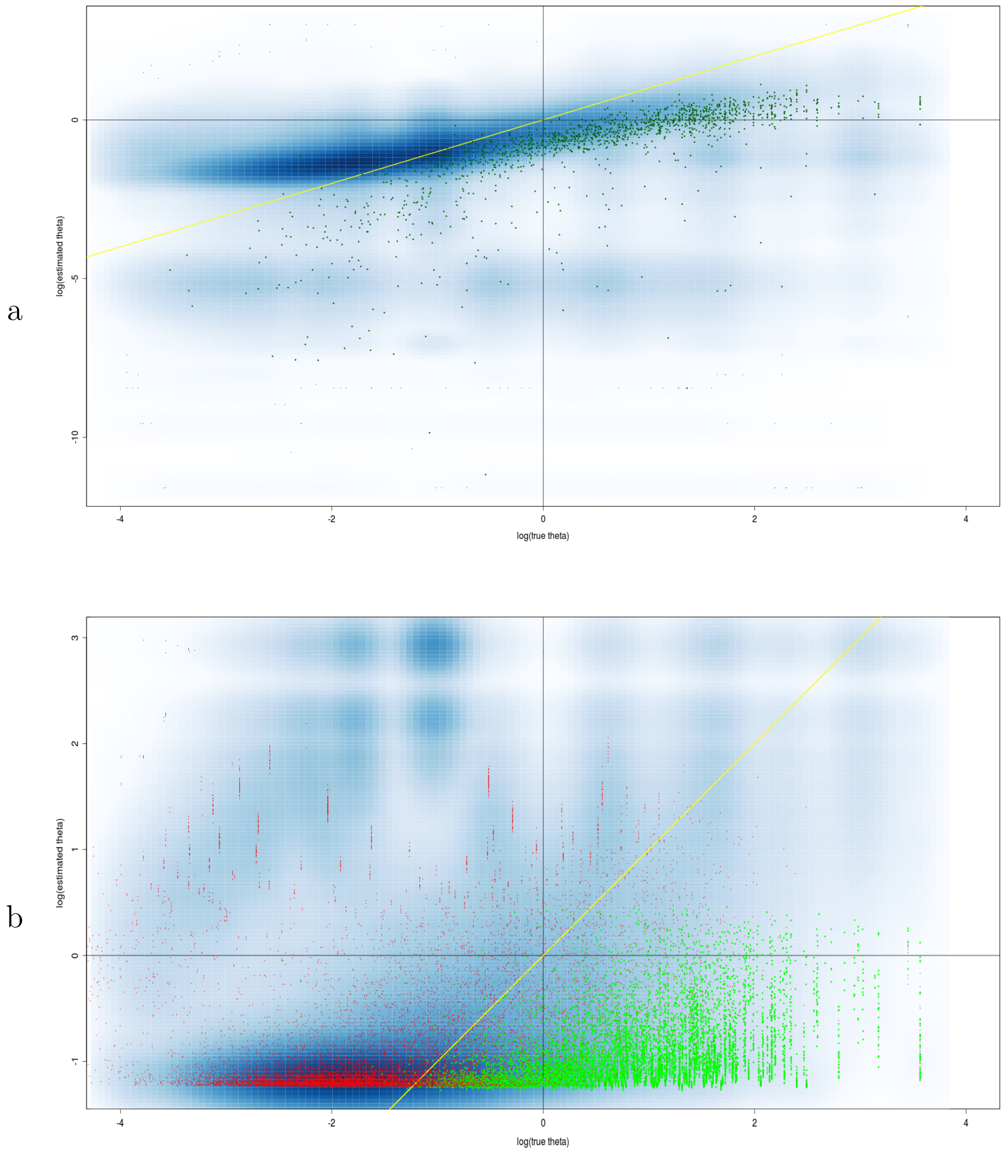


Figure 3.3. Dispersion estimates against true dispersion value on log scale

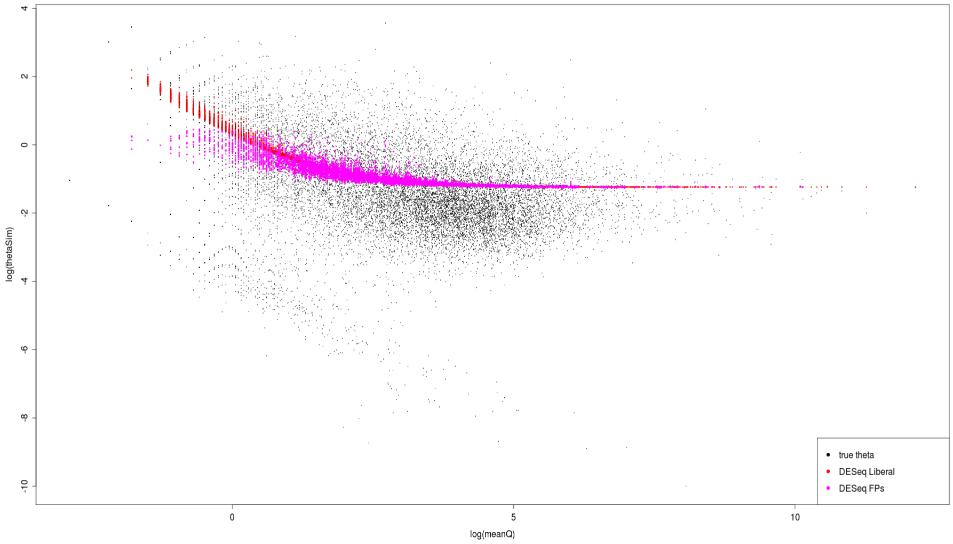


Figure 3.4. Dispersion values against mean gene count on a log scale

Chapter 4

Analyzing RNA-Seq Studies Using a Negative Binomial Model with Zero-inflated Component

4.1 Introduction

Gene expression analyses utilizing data obtained through sequencing of RNA molecules (RNA-Seq [58, 60]) have gained widespread popularity over the last years. Due to the high costs of sequencing when first introduced, early RNA-Seq studies employed very simple experimental designs, i.e. compared a small number of biological replicates across two conditions without regards for potential confounders. The two most popular software packages DESeq [3] and edgeR [87] described in Sections 3.2.1 and 3.2.2 to analyze RNA-Seq data were designed to analyze studies with few biological replicates by leveraging information across genes. As sequencing costs decrease rapidly [40], more complex, and ultimately more interesting, experimental designs can be utilized to investigate biological questions of interest. For example, due to its complexity, the Chen dataset introduced in Section 3.2.4 requires the inclusion of several covariates such as age, gender, brain pH, and PMI in addition to diagnostic group, which is the covariate of interest. The Marioni dataset introduced in Section 3.2.3 has multiple biological replicates for males and females across three species. A subset of a recently published gene expression study by Kang et al. [37] using exon arrays to investigate the development of the human brain across 32 brain regions is currently being sequenced using RNA-Seq to investigate differences in gene expression profiles across several brain regions. Covariates such as brain pH, PMI, age and gender played an important role in the original study and will need to be adjusted for in the RNA-Seq study as well. The emergence of these complex RNA-Seq datasets stresses the need for models that can handle additional covariates.

The GLM framework introduced in Section 3.2.1 is well suited for complex datasets. Consequently, Robinson et al. and Anders et al. extended their methods to handle additional covariates using the GLM framework. As outlined in Chapter 3, these methods seem to have trouble with controlling type I error under certain realistic scenarios.

More complex datasets have also unearthed an additional phenomenon that affects a subset of genes. Consider Figure 4.1 that shows a heatmap of counts for a subset of genes of the Marioni dataset. It can be seen that there are a number of genes for which there are many samples with zero counts (blue color) while other samples have counts between 70-1,000 (red color). These large differences do not seem to correlate with gender or species differences for the majority of genes and therefore a GLM with gender and species as explanatory variables will not be able to explain the variance present for these genes.

In statistics this property is called zero-inflation, i.e. having more zero counts than

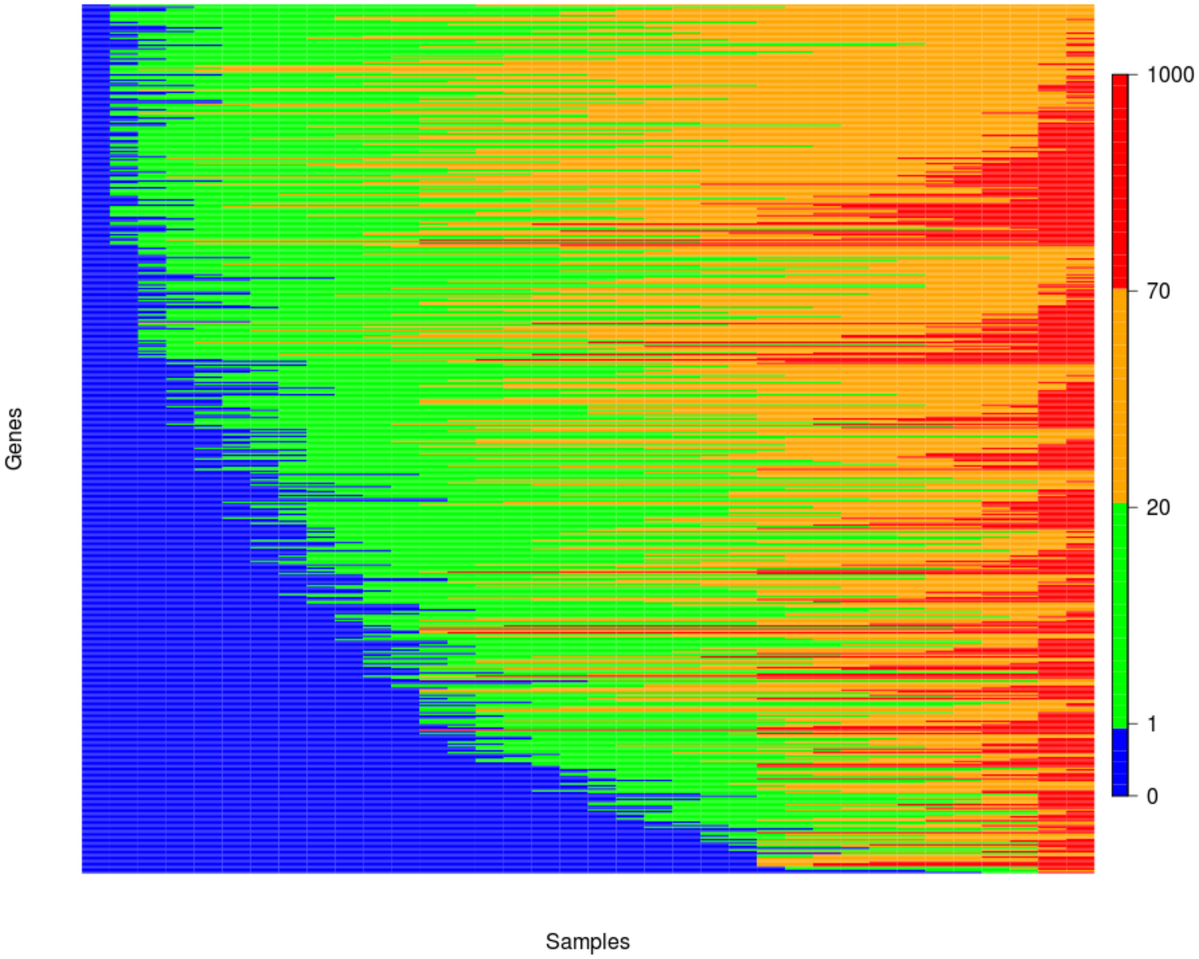


Figure 4.1. Heatmap of counts for Marioni dataset

one would expect from a NB distribution with given mean and dispersion. To further quantify to what degree and how many genes are affected, simulation studies using NB models with mean gene counts obtained from the Chen, Marioni, and Cheung datasets and the four sets of dispersion estimates described in Section 3.2.3 were conducted to determine the proportion of genes with zero counts above the expected number of zero counts. One thousand datasets were simulated for each scenario and original dataset and the number of zeros occurring for each gene were recorded. For each gene an empirical p-value for the number of zeros occurring in the original dataset was computed as the proportion of simulated datasets that showed more zeros than the observed zero count. Table 4.1 lists the proportion of genes with empirical p-values smaller than 0.001 for three datasets across all four scenarios. One would expect of course that 0.1% of the genes have empirical p-values of 0.001 or smaller. The results of the simulation studies however indicate that between 0.7% and 5% of genes could be affected depending on the true underlying distribution of dispersion parameters and the structure of the dataset.

Table 4.1. Proportion of genes with empirical p-values smaller than 0.001

Dataset	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Chen	0.05	0.03	0.02	0.01
Cheung	0.017	0.016	0.009	0.007
Marioni	0.023	0.019	0.011	0.012

Furthermore, a NB GLM and a NB GLM with additional zero-inflation parameter (see Section 4.2.1 for more detail) were fit to the Chen, Cheung, and Marioni datasets and the Akaike information criterion (AIC [2]) for both models was recorded. Since $AIC_{NB} - AIC_{ZINB} - 2$ is distributed according to a χ^2-distribution with one degree of freedom, a p-value that quantifies how much the extra parameter in the zero-inflated NB model improves the fit of the model can be obtained. Table 4.2 lists the number of genes with moderate (p-value<0.2) and strong (p-value<0.05) evidence for an improved model fit among genes for which AIC_{ZINB} was smaller than AIC_{NB}.

Table 4.2. Number of genes with AIC differences showing moderate and strong evidence for improved model fit among genes for which AIC_{ZINB} was smaller than AIC_{NB}

Dataset	Total	P-value<0.2	P-value<0.05
Chen	1074	477	206
Cheung	749	244	123
Marioni	424	250	107

A zero-inflated NB model was introduced by Rashid et al. [78] to identify genomic regions enriched in ChIP-seq and DNA-Seq data. In this chapter a comprehensive method fitting a NB GLM with zero-inflation component is proposed that improves upon existing methods with regards to type I error control while maintaining similar or better power for a range of sample sizes for RNA-Seq studies. The zero-inflation component also improves power for the subset of genes identified above. The method is called *gamSeq* as it is based on a fitting algorithm used to fit generalized additive models (GAM) for location, scale, and shape introduced by Rigby and Stasinopoulos [84]. The statistical properties of the proposed model are evaluated and compared to existing methods for analyzing RNA-Seq data.

4.2 Methods

4.2.1 Negative binomial model with zero-inflated component

Using the GLM framework introduced in Section 3.2.2 and the parameterization used by edgeR, i.e. $\mu_{ij} = q_{ij}s_j$ and $\sigma^2_{ij} = \mu_{ij} + \theta_i\mu^2_{ij}$, the following negative binomial model with zero-inflation component (ZINB) is proposed to analyze RNA-Seq studies with more sophisticated experimental designs:

$$f_{ZINB}(y_{ij}|\mu_{ij}, \theta_i, \boldsymbol{x_j}, \boldsymbol{\beta_i}, \pi_i, s_j) = \pi_i * I_0(y_{ij}) + (1 - \pi_i) * f_{NB}(y_{ij}|\mu_{ij}, \theta_i, \boldsymbol{x_j}, \boldsymbol{\beta_i}, s_j), \quad (4.1)$$

where $y = 0, \ldots, 1;\ j = 1, \ldots, N;\ i = 1, \ldots, I$ and

- $\pi_i \in [0, 1]$ is unobserved probability of belonging to the point mass component of zero counts,
- $I_0(y_{ij}) = 1$ if $y_{ij} = 0$ and $I_0(y_{ij}) = 0$ if $y_{ij} > 0$),
- $f_{NB}(y|\mu_{ij}, \theta_i, \boldsymbol{x_j}, \boldsymbol{\beta}, s_j)$ is the negative binomial pmf with log link function

$$\log(\mu_{ij}) = \log(q_{ij}s_j) = \log(s_j) + \beta_{i0} + \beta_{i1}x_{j1} + \ldots + \beta_{ip}x_{jp} = \log(s_j) + \boldsymbol{x_j}\boldsymbol{\beta_i}. \quad (4.2)$$

Recall that $X_1, \ldots, X_p$ represent either covariates of interest or confounders that need to be adjusted for and $\log(s_j)$ is used as a constant offset to adjust for total number of reads for sample j in the model statement. Note that for $\pi \equiv 0$, $f_{ZINB} \equiv f_{NB}$.

Since the proposed model only ought to be used when evidence for zero-inflation exists, the following algorithm is proposed to obtain parameter estimates for $\pi_i, \boldsymbol{\beta_i}$, and θ_i for gene i:

1. Fit the NB model, i.e. set $\pi_i \equiv 0$.
2. If at least one $y_{ij} = 0$, fit ZINB model.
3. If $AIC_{ZINB} < AIC_{NB}$, use the parameter estimates and their estimated variance-covariance matrix obtained from fitting the ZINB model. Otherwise, use the parameter estimates and their variance-covariance matrix obtained from fitting the NB model.
4. Use the variance-covariance matrix of parameter estimates of the corresponding model supplied by the fitting algorithm to test for significance of covariates of interest using a Wald-type test.

Both models are fit using an algorithm by Rigby and Stasinopoulos [84] implementing generalized additive models for location, scale and shape (gamlss) in their R package gamlss [100]. The gamlss algorithm implements a wide variety of distributions and allows all parameters of the distribution to be modeled as a combination of parametric and/or additive nonparametric functions of explanatory variables as well as random-effect terms. As shown in Equation 4.1, the ZINB distribution is defined by three parameters, the location parameter μ_{ij}, the scale (or dispersion) parameter θ_i and the zero-inflation parameter π_i. To achieve acceptable convergence rates in a timely and computational efficient manner, the standard log link function shown in Equation 4.2 was used to model the location parameter μ_{ij} while no additional covariates were used to model the dispersion parameter θ_i and zero-inflation parameter π_i. The proposed algorithm will be referred to by *gamSeq*, named after the accompanying R package, in the remainder of this text.

For the proposed model, Rigby and Stasinopoulos recommend the use of their own fitting procedure introduced in [26] and [83]. For gene i, $i = 1, \ldots, I$, contingent on the model used, the algorithm maximizes the likelihood $l_{ZINB}(\boldsymbol{\beta_i}, \theta_i, \pi_i | \boldsymbol{y_i}, \boldsymbol{X})$ or $l_{NB}(\boldsymbol{\beta_i}, \theta_i | \boldsymbol{y_i}, \boldsymbol{X})$, corresponding to Equations 3.2 and 4.1 where $\boldsymbol{y_i} = \{y_{ij}\}_{j=1,\dots,N}$ and $\boldsymbol{X} = \{x_{jk}\}_{j=1,\dots,N;k=1,\dots,p}$. An iterative Newton-Raphson algorithm is used to find the maximum of the corresponding likelihoods using the observed information matrix. The observed information matrix is also used to provide standard errors on parameter estimates and standard Wald-type t-tests are used to test hypotheses for individual parameters. A custom grid search is employed to find valid starting values. For a detailed description of the fitting algorithm, please refer to Appendix B.2 in [84].

4.2.2 Model Assessment

To assess the performance of the proposed model, the same strategy outlined in Sections 3.2.3 and 3.2.4 was used. Four hypothetical true dispersion distributions based on dispersion estimates obtained by using DESeq Conservative, DESeq Liberal, edgeR, and ML estimation were used to simulate data according to a NB distribution. These simulation studies aim to evaluate the performance model with regards to type I error and power of the proposed algorithm when the underlying distribution does not incorporate any zero-inflation, i.e. under favorable conditions for the current approaches. Furthermore, scenarios 2 and 4 were used to simulate data according to Equation 4.1, i.e. a negative binomial distribution with zero-inflation parameter π. Values of 0.05, 0.1, 0.15, and 0.2 based on simulation studies outlined in Section 4.1 were chosen for π. Sixty-four datasets were simulated for both scenarios using mean and dispersion estimates based on the Cheung data and the size of the test at a 0.001 significance level, power at a 10% FDR cutoff, and the number of false positives when testing for species differences were recorded. The proposed model was also used to test for gender differences in the five datasets described in Sections 3.2.3 and 3.2.4.

4.3 Results

In this section the results from the simulation studies and application to real data described in Section 4.2.2 are presented.

4.3.1 Statistical Properties

First the statistical properties of the proposed method are evaluated and compared to existing methods. Supplementary Table A.1 lists size estimates at a 0.001 significance level across the four different scenarios and datasets used and Figure 4.2 illustrates those results. The horizontal red line in Figure 4.2 indicates the nominal level and bars above the red line indicate increased type I error. In scenario 2, size estimates for all methods are fairly close to nominal level and do not give reason for concern. In scenarios 1, 3 and 4 however, size estimates for edgeR are up to 2-fold, 3-fold and 4-fold above nominal level, size estimates for DESeq Liberal are up to 5-fold and 8-fold above nominal level while size estimates for DESeq Conservative are again at nominal level or below and size estimates for gamSeq hover around nominal level. Supplementary Table A.2, listing the average (maximum) number of false positives when testing for the categorical nuisance variable included in the model using a 10% FDR cutoff, and Figure 4.3 provide further evidence for increased type I error rates for edegR and DESeq Liberal for scenarios 1, 3, and 4.

Supplementary Table A.3 lists power estimates at a 10% FDR cutoff and Figure 4.4 shows power estimates against sample size for simulation studies based on the Cheung data. Comparing power estimates for gamSeq to those for edegR and DESeq Liberal,

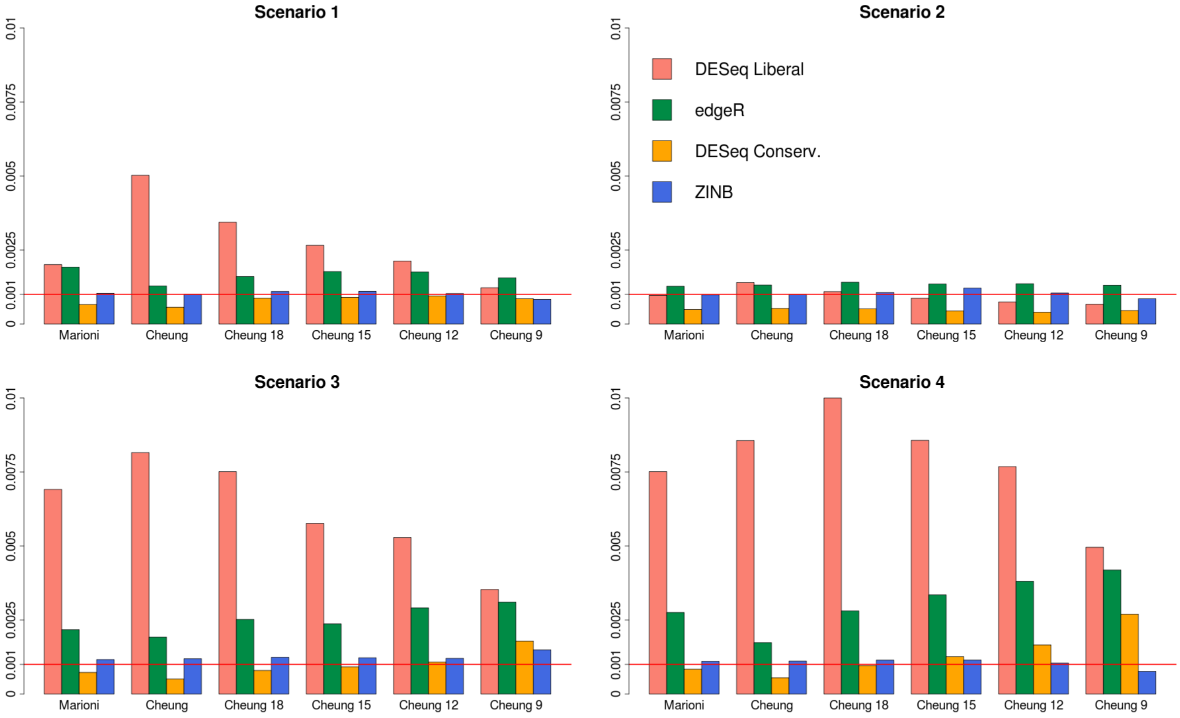


Figure 4.2. Size estimates for simulation studies at 0.001 significance level. The red line indicates nominal level.

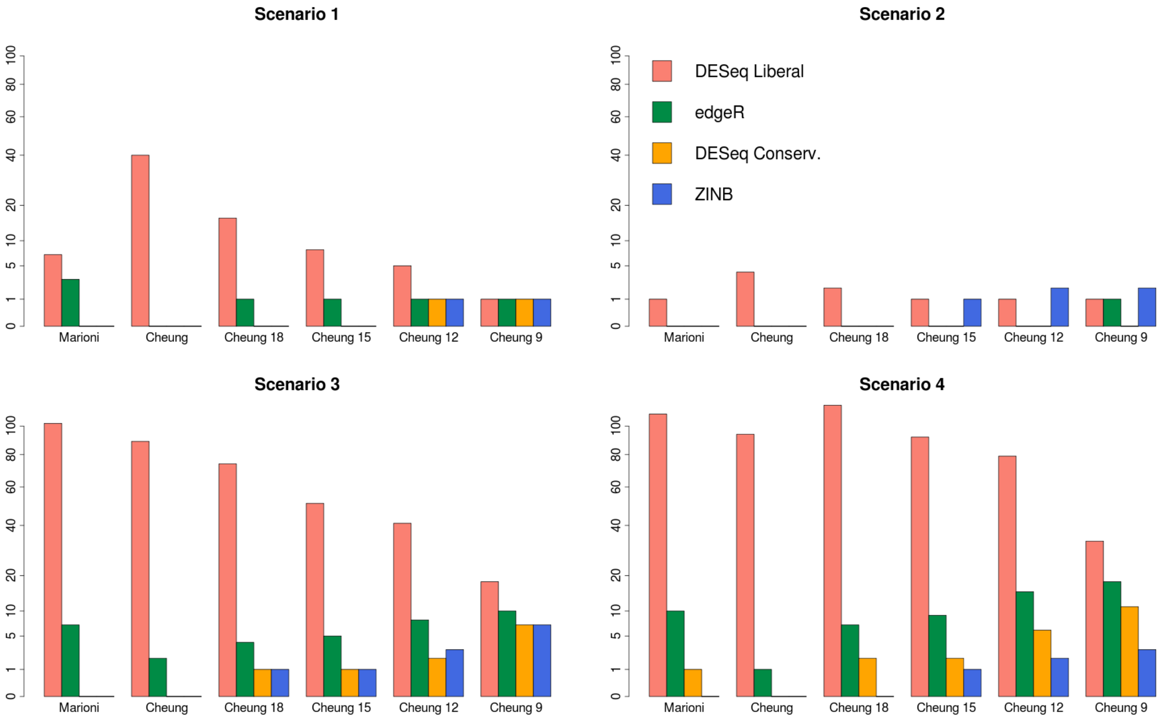


Figure 4.3. Median number of false positives when testing for species differences.

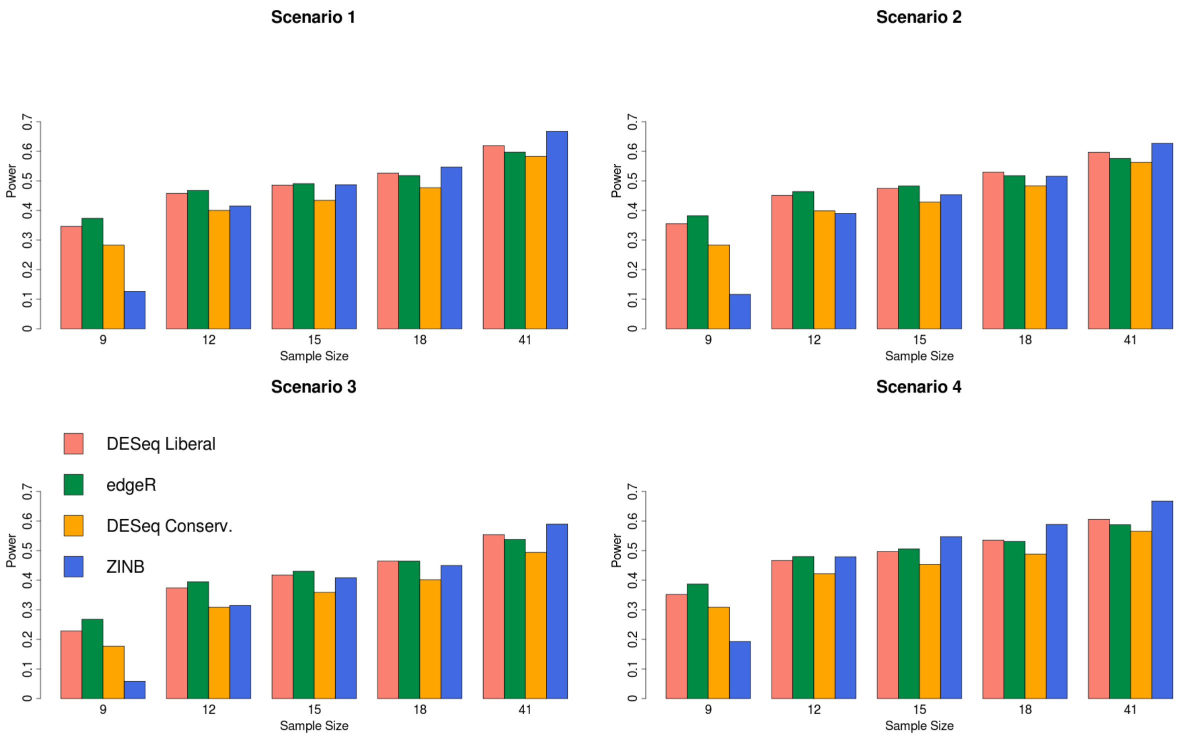


Figure 4.4. Power against sample size for Cheung simulation studies.

gamSeq provides similar or better power down to sample size 15, a slight power loss at sample size 12 and a significant loss of power at sample size 9 for scenarios 1, 2, and 3 while for scenario 4 gamSeq provides higher power down to sample size 15, similar power for sample size 12 and a loss of power at sample size 9. Compared to DESeq Conservative, gamSeq provides higher power for sample sizes of 15 or larger, comparable power at sample size 12, and shows a moderate power loss at sample size 9 across all scenarios.

To further investigate the reason for the increased numbers of false positives for DESeq Liberal and edgeR observed in scenarios 3 and 4, biases of dispersion estimates obtained from the simulations studies of sample size 15 based on the Cheung dataset were calculated. As shown in Table 4.3, which lists the interquartile range (IQR) and median bias, all four methods perform fairly well in recovering the true underlying dispersion where gamSeq slightly underestimates and the remaining methods slightly overestimate the true dispersion. For genes that were wrongly identified by DESeq Liberal as differentially expressed in scenarios 3 and 4 however, the median bias of DESeq Liberal's dispersion estimates was determined to be -3.09 and -1.83, respectively, while the median bias of gamSeq's dispersion estimates was calculated as -1.18 and -0.71, respectively. Similar observations could be made when looking at false positives identified by edgeR in scenarios 3 and 4 where the median bias for edgeR's dispersion estimates was -1.79 and -1.018, respectively, compared to a median bias of -1.25 and -0.72 for gamSeq's dispersion estimates. It has been shown by Hubbard and Allen [31] that the LRT has inflated type I error when the overdispersion parameter θ is significantly underestimated. Thus, these findings indicate that underestimated dispersion for a subset of genes under certain scenarios leads to increased type I error rates for DESeq Liberal and edgeR while gamSeq's estimates for these genes are less biased allowing for better control of type I error.

Table 4.3. Median bias and IQR obtained from simulated datasets of sample size 15 based on Cheung data structure

Scenario		DESeq Liberal	edgeR	DESeq Conservative	gamSeq
1	Median Bias	0.1470	0.0279	0.1588	-0.0777
	IQR	(-0.01,0.2654)	(-0.0323,0.1214)	(0.0344,0.4605)	(-0.1982,-0.0191)
2	Median Bias	0.0695	-0.0057	0.0487	-0.0817
	IQR	(0.0629,0.1013)	(-0.0542,0.0537)	(0.0181,0.2758)	(-0.1780,0.0100)
3	Median Bias	0.2212	0.0452	0.1697	-0.1146
	IQR	(0.1695,0.2944)	(-0.1006,0.1550)	(0.1496,0.4856)	(-0.2469,-0.0573)
4	Median Bias	0.2391	0.0770	0.1989	-0.0597
	IQR	(0.0749,0.4603)	(-0.0260,0.2565)	(0.0997,0.7446)	(-0.2218,-0.0132)

Finally it was of interest to determine whether there is a relationship between test statistics and dispersion estimates. When testing for differential gene expression the interest lies in detecting differences in population means, i.e. the null hypothesis $\mu_A = \mu_B$ is tested against the alternative hypothesis $\mu_A \neq \mu_B$ for two populations A and B. This implies that the dispersion parameter θ in the negative binomial distribution is a true nuisance parameter in the sense that it is not of interest for testing the null hypothesis. Thus, a desirable statistical property is the ability to detect true differences in populations means independently from the value of the overdispersion parameter θ. Therefore test statistics should in principle be independent from dispersion estimates obtained from the fitting procedure. Figures 4.5 and 4.6 show test statistics against dispersion estimates for the Marioni and Cheung data. Systematic patterns can be observed for the DESeq methods and to a lesser extend for edgeR while gamSeq does not show any relationship between test statistics and dispersion estimates.

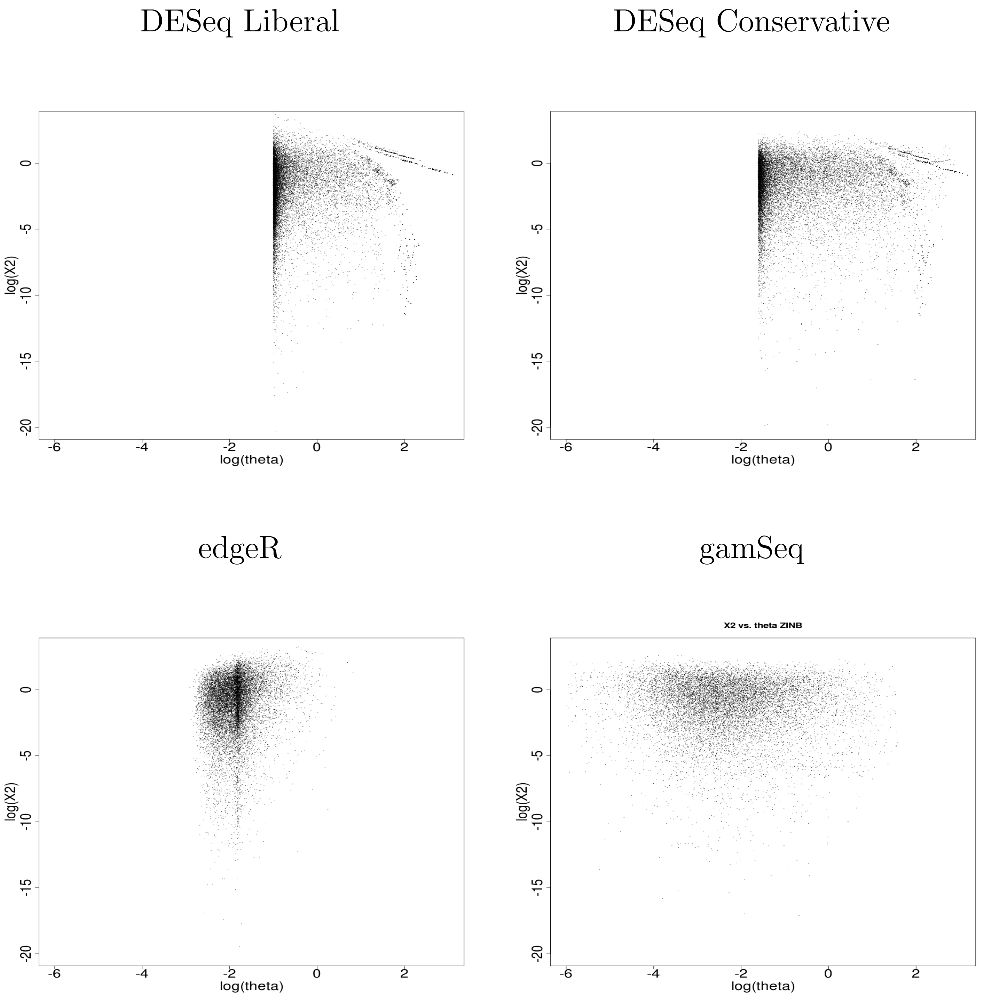
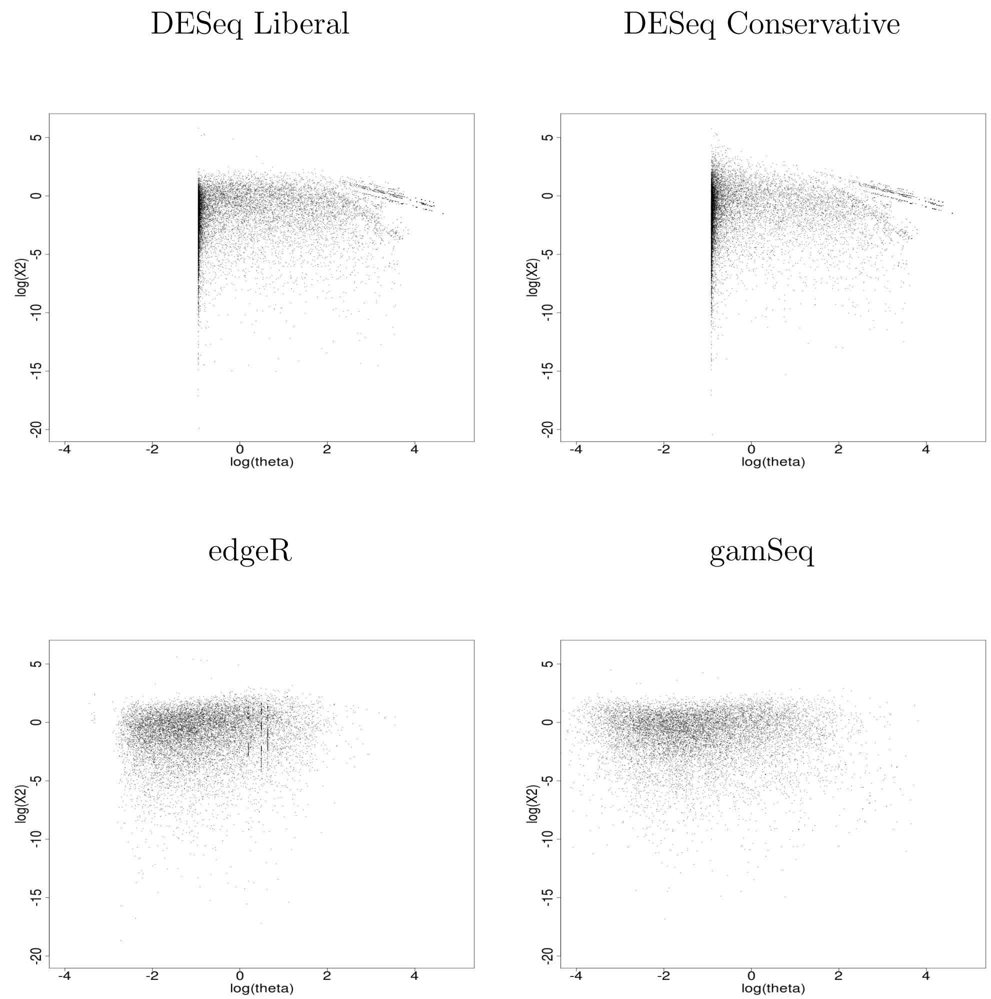


Figure 4.5. Test statistics against dispersion estimates on log scale for the Marioni dataset.


DESeq Liberal

DESeq Conservative

edgeR

gamSeq

Figure 4.6. Test statistics against dispersion estimates on log scale for the Cheung dataset.

4.3.2 Application to internal and publicly available data

While simulation studies provide valuable insight into a method's performance and robustness under various conditions, the ultimate goal is to analyze experiments designed to answer biological questions of interest. To that end the proposed method was used to test for gender differences using three publicly available datasets and one internal dataset described in Sections 3.2.4 and 3.2.3. Table 4.4 shows the distribution of genes called significant across chromosome Y, chromosome X, and autosomal chromosomes. For the Chen dataset, the majority of genes identified to show evidence for gender specific gene expression by DESeq Conservative, edgeR, and gamSeq are located on chromosome Y. All genes located on chromosome X identified by gamSeq have been indicated to show gender specific gene expression profiles by Xu at al. [110], Zhang et al. [112], and Preumont et al.[73]. Genes identified by edgeR follow a similar pattern where four out of five genes located on chromosome X overlap with those identified by gamSeq and therefore have been indicated to show gender specific gene expression while the fifth gene and the two genes located on other chromosomes have not been indicated before. DESeq Conservative and gamSeq perform slightly better than edgeR as they identify more genes on chromosome Y. DESeq Liberal performs worse with regards to false positives as the method identified two genes on chromosome X and 18 genes on autosomal chromosomes that are not known to show gender specific gene expression.

For the full Cheung dataset, DESeq Conservative and gamSeq perform similarly well, each identifying six genes that showed evidence for gender specific gene expression in previous studies, whereas the majority of genes identified by edgeR and DESeq Liberal have not been indicated before and are most likely false positives. A random subset, i.e. ten male and ten female samples, of the Cheung data was used to investigate the robustness of the findings from the full dataset. Results obtained with gamSeq proved to be the most robust although five genes without prior evidence for gender specific gene expression were identified. However, gamSeq's increased false positive rate is modest when compared to the breakdown in performance of edgeR and to a lesser extend DESeq Conservative.

Similar observations with regard to false positive findings could be made using the Marioni and Gilad data where gamSeq performs well in both cases while DESeq Conservative performs well for the larger dataset but does produce a significant number of false positive findings when applied to the smaller dataset. DESeq Liberal and edgeR do both show the same tendency towards false positive findings that was observed in the Cheung data.

Table 4.4. Genes called significant at a 10% FDR cutoff when testing for gender differences.

Dataset	**Chromosome**	**edgeR**	**DESeq Lib.**	**DESeq Cons.**	**gamSeq**
Chen	chr X	5	3	2	4
	chr Y	10	17	16	15
	autosomal	2	18	0	1
Cheung Full	chr X	2	10	0	1
	chr Y	5	7	6	5
	autosomal	22	231	0	0
Cheung SS20	chr X	12	8	1	0
	chr Y	2	5	6	5
	autosomal	240	182	31	5
Marioni	chr X	0	0	0	0
	chr Y	0	0	0	0
	autosomal	10	19	0	0
Gilad	chr X	1	1	0	0
	chr Y	1	1	0	0
	autosomal	57	39	37	0

4.3.3 Impact of zero-inflation on power and type I error

As indicated in the introduction, zero-inflation can play a significant role for a subset of genes and it was of interest to determine how the statistical properties of the four methods change when zero-inflation is present. To that end, the simulation studies described in Section 4.2.2 were used to assess power at a 10% FDR cutoff and size at a 0.001 significance level. Figures 4.7 and 4.8 as well as Supplementary Tables A.4 and A.5 show size estimates for four levels of zero-inflation across 64 simulated datasets for scenario 2 and 4 based on dispersion estimates obtained from the Cheung dataset. Zero-inflation seems to impact DESeq and edgeR differently. Both DESeq methods show inflated type I error rates for increasing values of the zero-inflation parameter π and decreasing sample size while edgeR is highly conservative for even small levels of zero-inflation. gamSeq on the other hand hovers around nominal size across all scenarios. The proposed method has also the edge with regards to power when zero-inflation is present. Figures 4.9 and 4.10 as well as Supplementary Tables A.6 and A.7 show power estimates for aforementioned simulations. At the same nominal level, gamSeq outperforms the other methods for sample sizes of twelve or larger across all scenarios. The difference in performance becomes more pronounced with larger values of the zero-inflation parameter.

4.3.4 Implementation

The method is implemented in the R package gamSeq available at www.people.vcu.edu/∼mreimers. The package depends on the gamlss package version 4.1. Convergence rates for the different approaches are shown in Table 4.5.

Table 4.5. Convergence rates for the Chen, Marioni, and Cheung datasets

Dataset	**Method**			
	edgeR	DESeq Lib.	DESeq Cons.	gamSeq
Chen	95.5%	96.4%	96.3%	96.7%
Marioni	81.0%	100.0%	99.9%	93.5%
Cheung	89.9%	100.0%	99.9%	99.0%

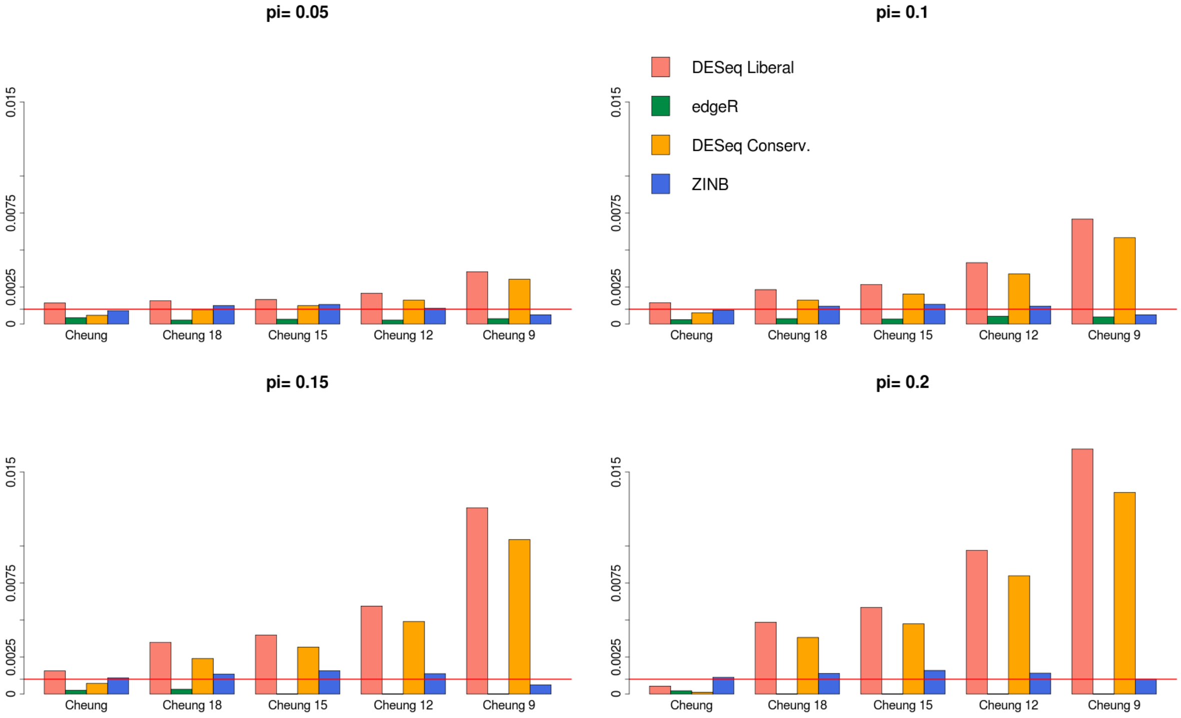


Figure 4.7. Size estimates for scenario 2 at 0.001 significance level. The red line indicates nominal level.

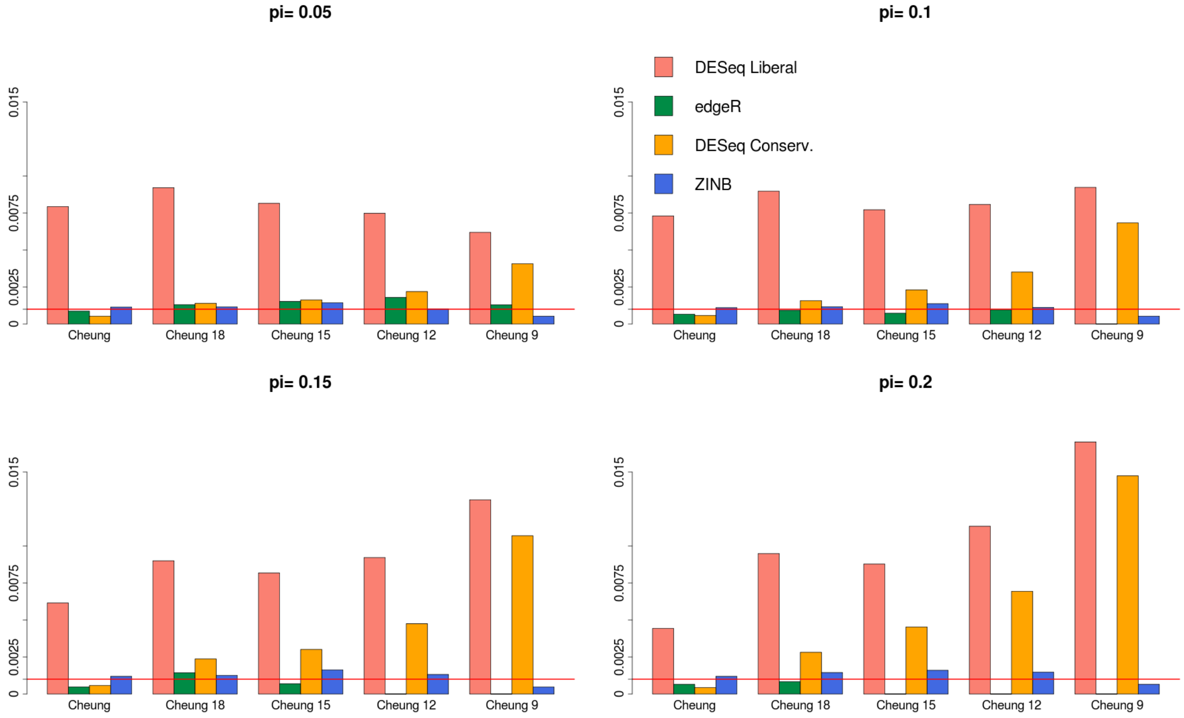


Figure 4.8. Size estimates for scenario 4 at 0.001 significance level. The red line indicates nominal level.

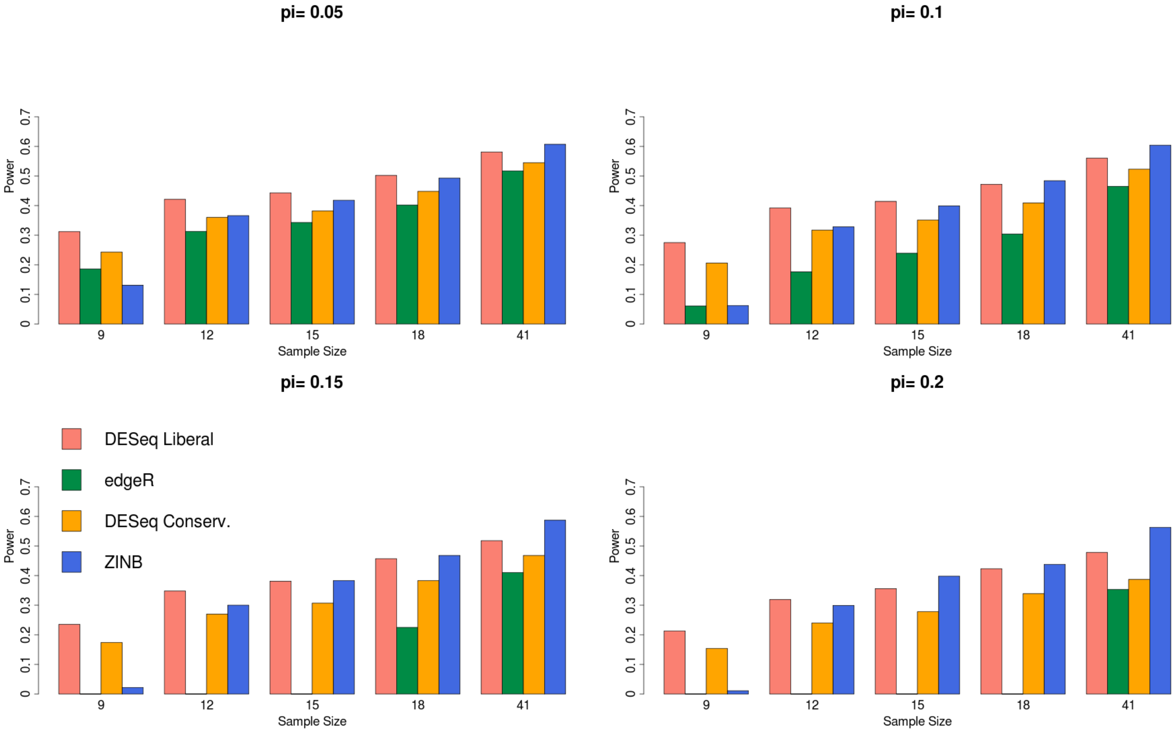


Figure 4.9. Power estimates for scenario 2 at 10% FDR cutoff.

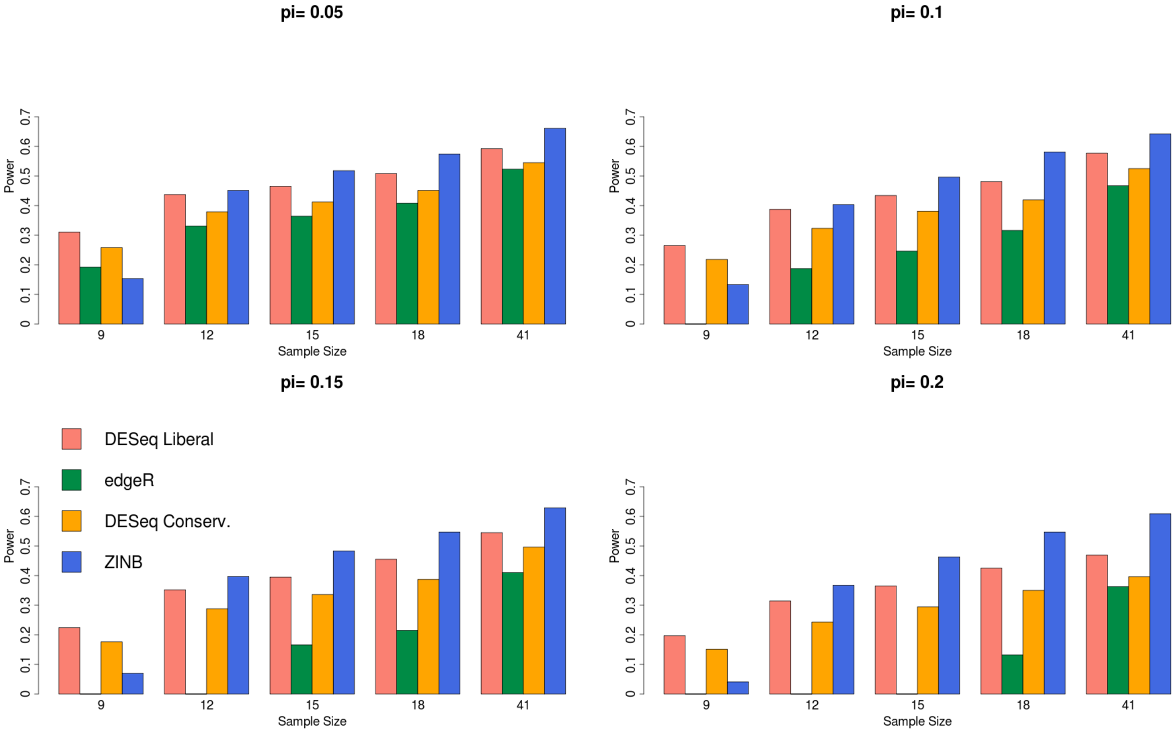


Figure 4.10. Power estimates for scenario 4 at 10% FDR cutoff.

4.4 Conclusions and Discussion

RNA-Seq studies are a powerful tool to test hypotheses regarding differences in transcription levels on a genome-wide basis. In contrast to microarray data that contains continuous measures of light intensity emitted due to hybridization of cDNA fragments to probes representing a gene as proxy for its transcription level, RNA-Seq experiments measure a gene's transcription abundance through counts of cDNA fragments mapping to that gene. The discrete nature of RNA-Seq data led to new applications of existing statistical models for analyzing count data in a high-throughput fashion. After the Poisson distribution was deemed as insufficient to explain biological variation in gene counts, the negative binomial distribution with its additional overdispersion parameter θ has been the distribution of choice to analyze RNA-Seq data. Early methods such as DESeq and edgeR used the properties of the NB distribution and empirical or Bayes approaches to increase power through sharing information across genes in studies with few biological replicates. Through the framework of GLMs these methods have been extended to analyze more complex experimental designs that include multiple potential confounders in addition to the covariate of interest.

In this chapter, *gamSeq*, a method based on the NB distribution with a zero-inflation component, was proposed to analyze complex RNA-Seq type studies and compared to existing methods across a wide range of sample sizes and hypothetical distributions of the dispersion parameter θ derived from real data since its true distribution is unknown. Several statistical properties including type I error, power, and bias of parameter estimates were investigated. Through simulation studies it was shown that the proposed method is more robust than existing methods with regards to controlling type I error across a number of different dispersion distributions. The existing methods edgeR and

DESeq Liberal were shown have increased size at a 0.001 significance level for a number of scenarios while gamSeq reliably controlled across all scenarios. A third method, DESeq Conservative, was considered and shown to be a conservative alternative that maintained size below the nominal level. The importance of controlling type I error was demonstrated by testing for gender differences in one internal and three publicly available datasets. DESeq Liberal and to a lesser extend edgeR identified a significant number of false positives for a number of datasets while gamSeq reliably identified genes known to show gender specific gene expression. DESeq Conservative performed similar to gamSeq for a number of datasets, but was shown to have the same tendency for false positives as DESeq Liberal when multiple cells had only one biological replicate, e.g. in datasets with few biological replicates per parameter fitted. The simulation studies also provided a reasonable explanation why existing methods fail control type I error in certain scenarios by showing that in general dispersion estimates obtained through information sharing across genes slightly overestimate the true underlying dispersion, but for a subset of genes, these methods significantly underestimate dispersion. Since the estimate of variance for a NB distribution is linearly related to the estimate of the dispersion parameter θ, underestimating the dispersion translates into an underestimate of variance that in turn can lead to false positives. In contrast, the proposed method gamSeq generally underestimates the true underlying dispersion slightly, but was shown to be less prone to significantly underestimate dispersion resulting in maintaining type I error at nominal level.

While controlling type I error is a necessary and desirable property for any statistical method used for inference, it should not be achieved through sacrificing statistical power, i.e. the ability to detect true biological differences. The same simulation studies used to investigate type I error were used to demonstrate that the proposed method achieves similar or better power compared to edgeR and DESeq Liberal and is superior to DESeq

Conservative for sample sizes of twelve or larger. A significant loss of power could be observed for a sample size of nine when fitting three categorical covariates, which translates into three observations per parameter fitted. In any traditional statistical analysis fitting a GLM this sample size would be considered insufficient to attain reasonable power when fitting a single model, not to mention fitting 10,000-20,000 models. For these small sample sizes methods that share information across genes seem to provide merit although positive results should be examined critically as it was shown that false positives are still a concern.

The gamSeq algorithm also addresses the phenomenon of zero-inflation, i.e. a higher number of zeros than expected from a NB distribution with given mean and dispersion, observed in more complex datasets. The gamSeq algorithm uses a screening procedure to identify genes with potential zero-inflation and introduces an additional zero-inflation parameter π when needed. Empirical and statistical evidence was presented to illustrate the problem and simulation studies were used to investigate its impact on power and type I error. It was shown that gamSeq performs significantly better than existing methods when zero-inflation is present and sample sizes are larger than or equal to twelve indicating that the same sample size limitations as observed for data generated from a NB distribution apply.

The gamSeq algorithm relies on the gamlss fitting algorithm, which was chosen due to its robustness and potential to fit more complex models that can include random and nonparametric effects. With the decreasing costs of RNA-Seq studies, longitudinal studies, studies exploring gene expression from different tissues of the same individual, or studies with technical replicates will become, or already are, feasible and will require explicit modeling of the dependencies among observations. The flexibility of the gamlss algorithm in principle allows implementation of such modeling approaches and extending the current

R package to include such terms in RNA-Seq analyses will be considered in the future. Recent studies have also uncovered that GC-content biases across samples may pose a serious problem [27, 67]. It may well turn out that some of the variance observed in the data is due to these biases. The gamlss algorithm allows explicit modeling of distribution parameters such as the overdispersion parameter through additional terms. For example, an adjustment for GC-content biases by modeling overdispersion as a function of gene specific GC-content could be implemented.

First exploratory analyses using random effects and non-parametric terms indicate that maintaining satisfactory convergence rates, computational efficiency, and desired statistical properties will be challenging. Issues that come with mixed models such as negative estimates of variance components, slow convergence of fitting algorithms, and difficulties in accurately estimating the variance-covariance matrix of parameter estimates multiply when fitting 10,000-20,000 generalized mixed models at a time and innovative solutions are needed.

In summary, insight into the behavior of existing methods to analyze RNA-Seq data quantifying transcription levels is provided and their limitations are demonstrated. A method is proposed that addresses the majority of limitations as well as the issue of zero-inflation that affects a subset of genes. Furthermore, a brief outlook to future developments in the field is provided along with an outline of major challenges existing methods are not able to address.

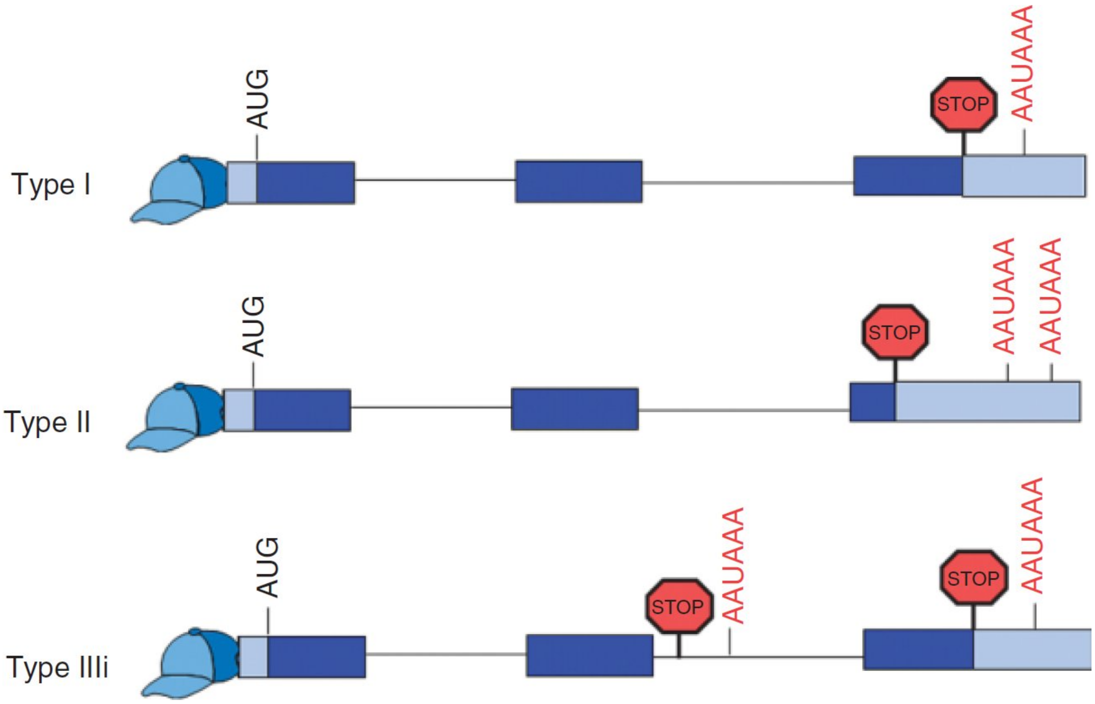
Chapter 5

Alternative Usage of mRNA polyadenylation sites

5.1 Introduction

Termination of transcription at the 3' end of genes has long been suspected to play an important role in regulation of gene expression [56, 69]. Lutz and Moreira [49] state that alternative usage of polyadenylation sites can affect important molecular processes such as RNA stability, translation, gene expression silencing, cell development and differentiation or genomic maintenance. The authors categorize polyadenylation events into three types shown in Figure 5.1. Biologically it is of interest whether the ratio of events of type II and III varies significantly between different groups, e.g. different tissues or disease groups. The high resolution of RNA-Seq enables researches to investigate differential termination of gene transcription and detect novel 3' untranslated regions (UTRs) or polyadenylation sites. This chapter introduces statistical methods to test for differential usage of polyadenylation sites using count data obtained through RNA-Seq experiments.

Figure 5.1. Scenarios considered for differential 3' UTR usage analysis [49].



Note that events of type II represent a differential use of polyadenylation sites while events of type III represent differential use of 3' UTRs. In the following differential usage of 3' UTRs and differential usage of polyadenylation sites will be used interchangeably keeping the difference in mind.

5.2 Materials and Methods

5.2.1 3' UTR Database Used And Corresponding Read Counts

The AceView database curated by Danielle and Jean Thierry-Mieg [102] provides a comprehensive and non-redundant sequence representation of all public mRNA sequences (mRNAs from GenBank [8] or RefSeq [76], and single pass cDNA sequences from dbEST [13] and Trace [9]). These experimental cDNA sequences are first co-aligned on the genome then clustered into a minimal number of alternative transcript variants and grouped into genes. AceView is arguably the richest transcript database currently available, containing more than 250,000 transcripts along with corresponding UTR regions. The UCSC Genome Browser [82] incorporates AceView annotations and therefore was utilized to create a database of 3' UTRs matching scenarios II and III in Figure 5.1. Since different datasets were aligned to different builds of the NCBI reference sequence, the following steps were repeated for the NCBI36/hg18 assembly as well as the NCBI37/hg19 assembly:

1. extract genomic coordinates for all AceView transcripts and the corresponding 3'UTR regions as well as coding regions with respect to NCBI36/hg18 (AceView build April 2007) and NCBI37/hg19 (AceView build February 2011)

2. For each gene, extract all combinations of 3' UTR regions that correspond to type II and III in Figure 5.1.

3. For each combination, remove regions that overlap coding regions, i.e. regions that are used as protein coding regions in a different transcript. If the entire range of either one of the 3' UTR regions in a combination is used as a protein coding region, then remove the combination from the putative 3' UTR regions.

4. From the remaining combinations for each gene, choose the combination of 3' UTR regions that is supported by the most cDNA clones in the AceView database as the 3' UTR region of interest.

Step 3 is necessary to ensure that sequenced reads that map to 3' UTR regions are representative of a UTR rather than a coding region of a transcript overlapping a specific UTR. After step 4, the database contains two 3' UTR regions for each gene and scenario. 10,184 genes had 3' UTR combinations structured according to type II while 3,827 genes had 3'UTR combinations structured according to type III. Read counts for each 3' UTR region and sample were obtained by counting how many uniquely mappable reads with at most one mismatch mapped to the corresponding genomic coordinates with a minimal overlap of three base pairs to account for possible base call errors.

5.2.2 Testing for Differential 3' UTR Usage Between Individual Samples

Due to cost constraints of HTS, many early RNA-Seq studies lack biological replicates. Often it is still of interest to make inferences about differential 3' UTR usage between individual mRNA samples extracted from two separate conditions or tissues. To illustrate how to statistically analyze differential 3' UTR usage using data from individual samples, RNA-Seq data obtained by Illumia from their HiSeq 2000 platform [24] was used to extract read counts for 15 different human tissue types (see Table A.8) sequenced with 75bp single end reads for a total of 15 lanes with 50-70 million reads per lane. For each tissue, read counts were obtained by using the algorithm outlined in Section 5.2.1. A multiplicative model for ratios of Poisson rates between UTRs based on the GLM framework introduced in Section 3.2.1 was then used to test for differential 3' UTR usage:

$$\log(n * \lambda) = \alpha + \beta_1 * tissue + \beta_2 * region + \beta_{12} * region * tissue \tag{5.1}$$

Since this is a multiplicative model, testing for the interaction between region and tissue, i.e. testing

$$H_0 : \beta_{12} = 0 \text{ vs. } H_a : \beta_{12} \neq 0, \tag{5.2}$$

is equivalent to testing whether the ratios in read counts for the two UTR regions differ significantly between tissues. A LRT test was used to test the interaction. Note that this model assumes that observations from overlapping or independent 3' UTR regions from the same gene are statistically independent. Realistically this assumption is violated for both types of polyadenylation events since two read counts are obtained from the same sample, one for each region. This equates to a model with two repeated measures per sample. Since there are only four degrees of freedom available when fitting the full model described in Equation 5.1, additional terms such as a random effect that could account for repeated measures designs could only be added with additional replicates. This concern is given further consideration in the next section.

5.2.3 Testing for Differential 3' UTR Usage Between Two Conditions With Biological Replicates

In more complex study designs it is often of more interest to compare two groups, e.g. disease patients against normal controls, with each group having multiple biological replicates. The framework presented in Section 5.2.2 can easily be extended to this scenario. The Chen dataset introduced in Section 3.2.4 was used to illustrate the proposed model. Recall that the Chen dataset contained reads for 82 sequenced brain samples from normal controls (n=26) and patients with bipolar disorder (n=25) or schizophrenia (n=31) and included information on age, brain pH, and post mortem interval (PMI), which were used during model fitting in addition to gender and diagnostic group. Since the some of the samples were sequenced at very low coverage due to multiplexing, all samples with less than one million uniquely mappable reads were excluded leaving 62 samples, 19 samples from normal controls, 13 samples from patients with bipolar disorder, and 30 samples from patients with schizophrenia. It was of interest to test whether read count ratios are significantly different between diagnostic groups.

A multiplicative model similar to that shown in Equation 5.1 was fit using the zero-inflated model introduced in Chapter 4.

$$\begin{aligned}\log(n * \lambda) = \alpha + \beta_1 * group + \beta_2 * region + \beta_{12} * region * group \\ + \beta_3 * age + \beta_4 * sex + \beta_5 * pH + \beta_6 * PMI\end{aligned} \tag{5.3}$$

To further account for the repeated measures design, the model described in Equation 5.3 was also fit with an additional random term $\gamma_i \sim N(0, \sigma_r^2 \boldsymbol{I})$ where i=1,...,N and $\boldsymbol{I}$ is a 2×2 identity matrix, using the glmmADMB package. Again, since this is a multiplicative model, testing for the interaction between region and group is equivalent to testing whether ratios in read counts between two 3' UTR regions differ significantly between diagnostic groups. P-values based on Wald-type tests were used to test the interaction.

5.3 Results

5.4 Testing for Differential 3' UTR Usage Between Individual Samples

The method outlined in Section 5.2.2 was used to test for differential 3' UTR usage between tissues in the Bodymap dataset. The focus was the comparison of 3' UTR usage between brain and the remaining tissues since Sandberg et al. [88] suggested that mRNA transcripts in brain tissue have longer 3' UTRs. For polyadenylation events of type III, 35%-50% of genes with detectable expression levels showed evidence for differential usage at a 10% FDR cutoff when comparing brain tissue to the other 14 tissues. Among these genes, 53%-65% used longer 3' UTRs in brain. Similar observations could be made for polyadenylation events of type II where 35%-62% of genes showed evidence for differential usage at a 10% FDR cutoff with 53%-69% of those genes using longer 3' UTRs in brain tissue. Please refer to Supplementary Tables A.8 and A.9 for more details.

5.5 Testing for Differential 3' UTR Usage Between Two Conditions With Biological Replicates

Using models similar to the model described in Equation 5.3, the influence of diagnostic group, brain pH, age, and PMI on differential usage of 3' UTRs was investigated. Brain pH was chosen as it has been shown to affect RNA integrity [94]. RNA degrades first from the 3' end [72] and therefore low RNA integrity could affect the number of fragments in 3' UTRs. PMI and age have not been found to be correlated with RNA integrity [41] and are not expected to have an effect on apparent 3' UTR usage. Diagnostic group was the variable of interest as differences in 3' UTR usage in patients with bipolar disorder or schizophrenia when compared to normals would be an interesting biological phenomenon. Figure 5.2 shows p-value plots for age, PMI, brain pH and diagnostic group after testing for the interaction in model 5.3 without random effect term. No genes showed evidence for an effect of age, PMI, or diagnostic group on 3' UTR usage at a 10% FDR cutoff. However, significant evidence for an effect of brain pH was found for 13 genes after multiple testing correction. Similar results were found when a random effect was included in the model (see p-value plots in Figure 5.3. Again, no genes showed evidence for differential 3' UTR usage depending on age, PMI, or diagnostic group while 26 genes, including all 13 genes found when omitting a random effect term, were found to show significant effect of brain pH on 3' UTR usage.

5.6 Conclusions and Discussion

In this chapter, a method based on the GLM framework was introduced to test for differential usage of 3' UTRs in two scenarios. The method was applied to two datasets for which count data for UTR usage were extracted from reads aligned to the human reference genome. The results obtained for the Bodymap dataset, i.e. a majority of genes show evidence for differential 3' UTR usage and transcripts in brain tissue have a slight preference towards longer UTRs, were consistent with observations made in previous studies indicating that the proposed method gives sensible results.

Sensible results were also obtained when applied to the more complex Chen dataset. While age and PMI do not seem to have an effect on differential 3' UTR usage, a number of genes showed evidence for an effect of brain pH. Interestingly, brain pH has been previously indicated to affect RNA integrity, which is often low due to RNA degradation at the 3' end of transcripts. Popova et al. [72] indicate that longer transcripts are more affected by RNA degradation at the 3' end. The mean length of transcripts associated with genes listed in Table 5.1 is 3920bp while the mean length of all other transcripts in RefSeq is 2970bp (t=1.79, df=47,p=0.0397 for one-sided test) supporting this conclusion. While positive results could be obtained, the statistical power of the proposed method is seriously limited in the Chen dataset by its lack of sequencing coverage. Seventy-five percent of the shorter region of 3' UTRs have mean counts of 30 or lower with 50% having mean counts of 5 or lower. Future datasets with significant higher sequencing coverage will allow a more detailed investigation of 3' UTR usage using the proposed method.

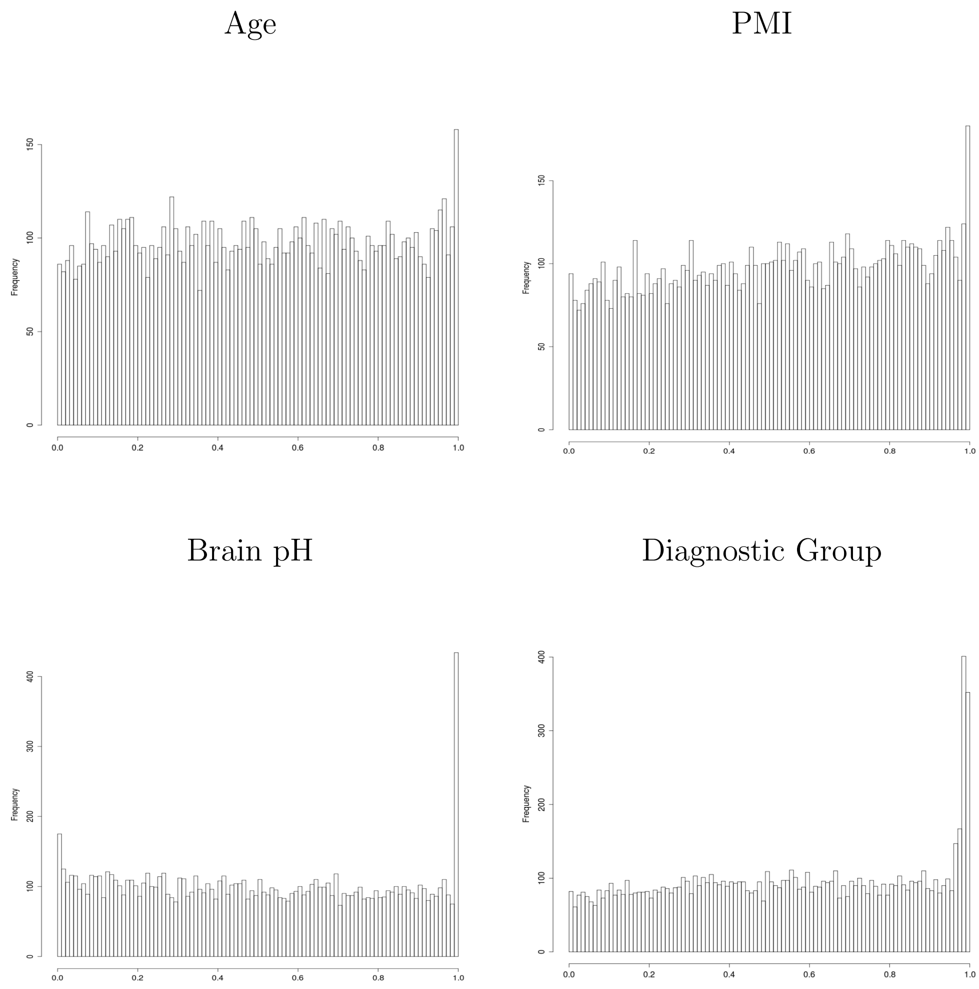


Figure 5.2. P-value plots for age, PMI, brain pH and diagnostic group omitting a random effect term.

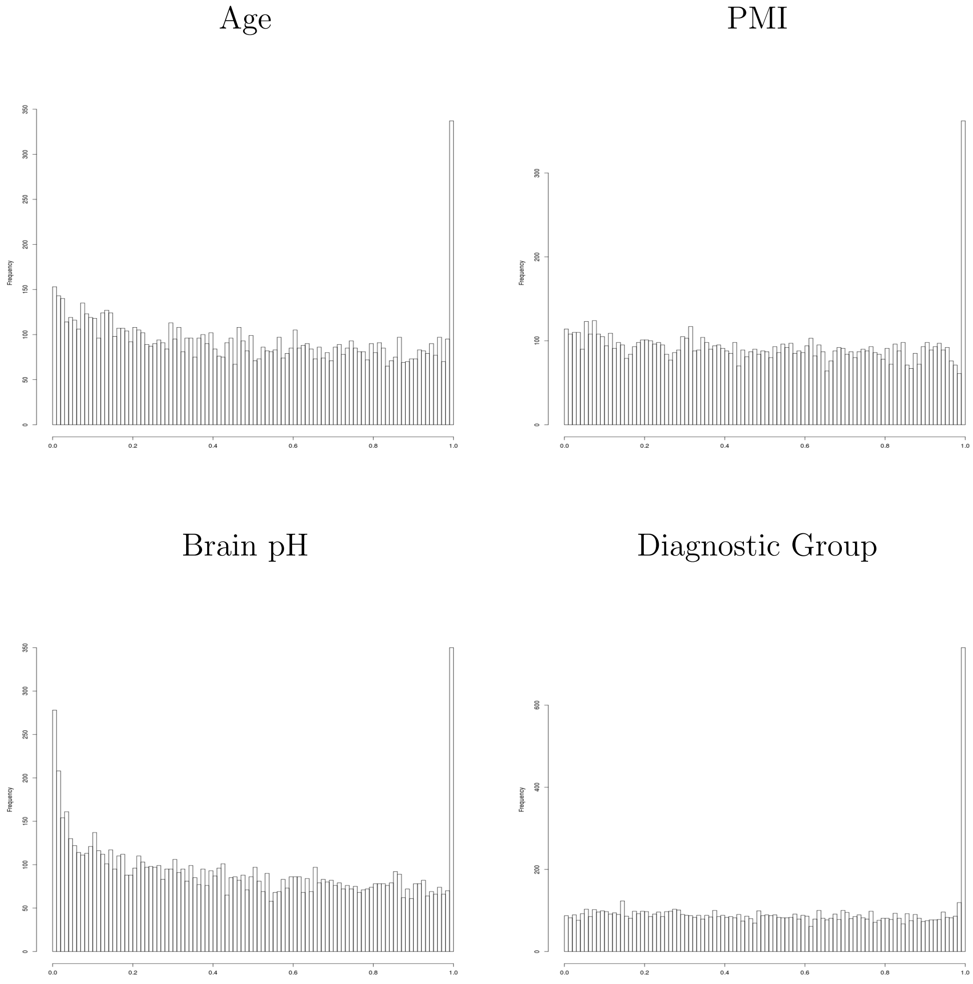


Figure 5.3. P-value plots for age PMI, brain pH and diagnostic group including a random effect term.

Since by design of the proposed method observations are correlated, the merit of including a random effect term in model 5.3 was evaluated. To that end, the glmmADMB was used and the number of significant genes showing evidence for 3' UTR usage influenced by brain pH could be doubled. These results indicate that including a random effect should be considered for study designs with repeated measures in space or time. However, first exploratory analyses regarding power, type I error, and convergence rates show that the glmmADMB package might not perform well for small sample sizes. Further research in this direction, i.e. how to fit 10,000-20,000 generalized linear mixed models at a time with satisfactory statistical properties and convergence rates along with computational efficiency, is needed.

Table 5.1. Genes with evidence for correlation between differential 3' UTR usage and brain pH

Gene	Length of exonic regions (bp)
ANXA5	1599
ARF3	3537
BAD	1493
BPNT1	2461
DDX55	2622
FABP3	1097
GDAP1	4113
GSTO1	1310
GTPBP2	2979
HSPA5	3970
KLC1	3158
LMO4	5406
MGAT4A	10915
CLVS1	3492
NRN1	2056
PCLO	22498
PTPRS	7347
PUM1	5514
RAB5C	2031
RTN3	6691
SREBF1	5001
STX17	6908
TOLLIP	3660
TSGA13	1653
VAMP1	6044
ZCCHC9	1994

Bibliography

1. Agilent. Humane genome cgh microarray 44k, March 2009.

2. Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

3. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

4. L.H. Augenlicht, L. Anderson, J. Taylor, and M. Lipkin. Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer. *PNAS*, 88(8):3286–3289, 1991.

5. L.H. Augenlicht, M.Z. Wahrman, H. Halsey, L. Anderson, J. Taylor, and Lipkin M. Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer Research*, 47(22):6017–6021, 1987.

6. T Barrett, DB Troup, SE Wilhite, P Ledoux, C Evangelista, IF Kim, M Tomashevsky, KA Marshall, KH Phillippy, PM Sherman, RN Muertter, O Holko M, Ayanbule, A Yefanov, and Soboleva A. Ncbi geo: archive for functional genomics data setsŮ10 years on. *Nucl. Acids Res.*, 39:D10005–10, 2011.

7. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

8. Dennis A. Benson, Ilene Karsch-mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. Genbank: update. *Nucleic Acids Res*, 32:23–26, 2004.

9. Eugene Berezikov, Ronald H.A. Plasterk, and Edwin Cuppen. Genotrace: cdna-based local genome assembly from trace archives. *Bioinformatics*, 18(10):1396–1397, 2002.

10. Graham R. Bignell, Jing Huang, Joel Greshock, Stephen Watt, Adam Butler, Sofie West, Mira Grigorova, Keith W. Jones, Wen Wei, Michael R. Stratton, P. Andrew Futreal, Barbara Weber, Michael H. Shapero, and Richard Wooster. High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. *Genome Research*, 14(2):287–295, 2004.

11. Sven Bilke, Qing-Rong Chen, Craig C. Whiteford, and Javed Khan. Detection of low level genomic alterations by comparative genomic hybridization based on cdna micro-arrays. *Bioinformatics*, 21:1138–1145, 2005.

12. Ran Blekhman, John C. Marioni, Paul Zumbo, Matthew Stephens, and Yoav Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research*, 20:180–189, 2010.

13. M. S. Boguski, T. M. J. Lowe, and C. M. Tolstoshev. Dbest-database for expressed sequence tags. 1993.

14. B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19:185–193, 2003.

15. Héctor Corrada Bravo and Rafael A. Irizarry. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, 10:1541–0420, 2009.

16. J. H. Bullard, E. A. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11:94, 2010.

17. Hung-I Harry Chen, Fang-Han Hsu, Yuan Jiang, Mong-Hsun Tsai, Pan-Chyr Yang, Paul S. Meltzer, Eric Y. Chuang, and Yidong Chen. A probe-density based analysis method for array cgh data: Simulation, normalization and centralization. *Bioinformatics*, 24(16):1749–56, 2008.

18. VG Cheung, RR Nayak, IX Wang, S Elwyn, SM Cousins, M Morley, and RS Spielman. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biology*, 8(9), 2010.

19. W.S. Cleveland and S.J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

20. D.R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, B(49):1–39, 1987.

21. C. Croux, P. Filzmoser, and M. R. Oliveira. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87:218–225, 2007.

22. Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucl. Acids Res.*, 36(16):e105–, 2008.

23. Peter Filzmoser, Heinrich Fritz, and Klaudius Kalcher. *pcaPP: Robust PCA by Projection Pursuit*, 2009. R package version 1.7.

24. Alyssa Frazee, Ben Langmead, and Jeff Leek. Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 2011. http://bowtie-bio.sourceforge.net/recount/.

25. Travis C. Glenn. Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, 11(5):759–769, 2011.

26. Wolfgang Haerdle and Michael Schimek, editors. *Statistical Theory and Computational Aspects of Smoothing*. Physica, 1996.

27. Kasper D. Hansen, Rafael A. Irizarry, and Zhijin Wu. Removing technical variability in rna-seq data using conditional quantile normalization. Johns Hopkins University, Dept. of Biostatistics Working Papers, Working Paper 227, May 2011.

28. Michael J. Heller. Dna microarray technology: Devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4:129–153, 2002.

29. P. A. C.Št Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucl. Acids Res.*, 36(21), 2008.

30. Joerg D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, 7:200–210, 2006.

31. Dean J. Hubbard and Brian Allen. Robustness of the sprt for a negative binomial to misspecification of the dispersion parameter. *Biometrics*, 47(2):419–427, 1991.

32. E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, O. Sauter, G.and Monni, A. Elkahloun, O.P. Kallioniemi, and A. Kallioniemi. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Research*, 62:6240–6245, 2002.

33. AJ Iafrate, L Feuk, MN Rivera, ML Listewnik, PK Donahoe, Y Qi, SW Scherer, and C Lee. Detection of large-scale variation in the human genome. *Nat. Genet.*, 36(9):949–951, 2004.

34. Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, 6th edition, 2007.

35. K. Jong, E. Marchiori, A. van der Vaart, S. F. Chin, B. Carvalho, M. Tijssen, P. P. Eijk, P. van den Ijssel, H. Grabsch, P. Quirke, J. J. Oudejans, G. A. Meijer, C. Caldas, and B. Ylstra. Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. *Oncogene*, 26(current):1499–1506, 2007.

36. A Kallioniemi, OP Kallioniemi, D Sudar, D Rutovitz, JW Gray, F Waldman, and D Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, 1992.

37. Hyo Jung Kang, Yuka Imamura Kawasawa, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li, Andre M. Sousa, Mihovil Pletikos, Kyle A. Meyer, Goran Sedmak, Tobias Guennel, Yurae Shin, Matthew B. Johnson, Zeljka Krsnik, Simone Mayer, Sofia Fertuzinhos, Sheila Umlauf, Steven N. Lisgo, Alexander Vortmeyer, Daniel R. Weinberger, Shrikant Mane, Thomas M Hyde, Anita Huttner, Mark Reimers, Joel E Kleinman, and Nenad Sestan. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, 2011.

38. D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database. *Nucl. Acids Res.*, 31(1):51–54, 2003.

39. Robert H. Kincaid, Jayati Ghosh, and Bo U. Curry. Analyzing cgh data to identify aberrations, February 2007.

40. Martin Kircher and Janet Kelso. High-throughput dna sequencing - concepts and limitations. *Bioessays*, 32:524–536, 2010.

41. Antje Koppelkamm, Benedikt Vennemann, Sabine Lutz-Bonengel, Tony Fracasso, and Marielle Vennemann. Rna integrity in post-mortem samples: influencing parameters and implications on rt-qpcr assays. *International Journal of Legal Medicine*, 125(4):573–580, 2011.

42. R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P.

Barber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2009. *Nucl. Acids Res.*, 37:D755–761, 2009.

43. D.A. Lashkari, J.L. DeRisi, J.H. McCusker, A.F. Namath, C. Gentile, S.Y. Hwang, P.O. Brown, and R.W. Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *PNAS*, 94(24):13057–13062, 1997.

44. Mei-Ling Ting Lee, Frank C. Kuo, G. A. Whitmore, and Jeffrey Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):9834–9839, 2000.

45. Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 09 2007.

46. F. Lepretre, C. Villenet, S. Quief, O. Nibourel, C. Jacquemin, X. Troussard, F. Jardin, F. Gibson, J. P. Kerckaert, C. Roumier, and M. Figeac. Waved acgh: to smooth or not to smooth. *Nucleic Acids Research*, 38(7):e94, 2010.

47. Doron Lipson, Yonatan Aumann, Amir Ben-Dor, Nathan Linial, and Zohar Yakhini. Efficient calculation of interval scores for dna copy number data analysis. *Journal of Computational Biology*, 13(2):215–228, 2006.

48. X. Shirley Liu, W. Evan Johnson, Clifford A. Meyer Wei Li and, Raphael Gottardo, Jason S. Carroll, and Myles Brown. Model-based analysis of tiling-arrays for chip-chip. *Proc. Natl. Acad. Sci. U. S. A.*, 103:12457–12462, 2006.

49. Carol S. Lutz and Alexandra Moreira. Alternative mrna polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley Interdisciplinary Reviews: RNA*, 2(1):22–31, 2011.

50. Elaine R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133 – 141, 2008.

51. J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

52. John C Marioni, Natalie P Thorne, Armand Valsesia, Tomas Fitzgerald, Richard Redon, Heike Fiegler, T Daniel Andrews, Barbara E Stranger, Andrew G Lynch, Emmanouil T Dermitzakis, Nigel P Carter, Simon Tavare, and Matthew E Hurles. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology*, 8:R:228, 2007.

53. U. Maskos and E.M. Southern. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucl. Acids Res.*, 20(7):1679–1684, 1992.

54. Michael L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11, 2010.

55. Leonardo A. Meza-Zepeda, Stine H. Kresse, Ana H. Barragan-Polania, Bodil Bjerkehagen, Hege O. Ohnstad, Heidi M. Namlos, Junbai Wang, Bjorn E. Kristiansen, and Ola Myklebost. Array Comparative Genomic Hybridization Reveals Distinct DNA Copy Number Differences between Gastrointestinal Stromal Tumors and Leiomyosarcomas. *Cancer Research*, 66(18):8984–8993, 2006.

56. Flavio Mignone, Carmela Gissi, Sabino Liuni, and Graziano Pesole. Untranslated regions of mrnas. *Genome Biology*, 3(3):reviews0004.1–reviews0004.10, 2002.

57. Andre E. Minoche, Juliane C. Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology*, 12:R112, 2011.

58. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5:621–628, 2008.

59. Robert Nadon and Jennifer Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends in Genetics*, 18(5):265–271, May 2002.

60. U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Sciene*, 320:1344–1349, 2008.

61. Yasuhito Nannya, Masashi Sanada, Kumi Nakazaki, Noriko Hosoya, Lili Wang, Akira Hangaishi, Mineo Kurokawa, Shigeru Chiba, Dione K. Bailey, Giulia C. Kennedy, and Seishi Ogawa. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research*, 65(14):6071–6079, 2005.

62. J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, A(135):370–384, 1972.

63. Pierre Neuvial, Philippe Hupé, Isabel Brito, Stéphane Liva, Élodie Manié, Caroline Brennetot, François Radvanyi, Alain Aurias, and Emmanuel Barillot. Spatial normalization of array-cgh data. *Bioinformatics*, 7:264, 2006.

64. NimbleGen. Hg18 cgh 385k whole genome tiling v2.0, March 2009.

65. Justin Petrone. In some european clinical cytogenetic labs, arrays have become primary sample-analysis tool. BioArray News, October 27 2009.

66. X Piao, P Cai, S Liu, N Hou, L Hao, F Yang, H Wang, J Wang, Q Jin, and Q Chen. Global expression analysis revealed novel gender-specific gene expression features in the blood fluke parasite schistosoma japonicum. *PLoS One*, 6(4), 2011.

67. Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464:768–722, 2010.

68. Dalila Pinto, Katayoon Darvishi, Xinghua Shi, Diana Rajan, Diane Rigler, Tom Fitzgerald, Anath C Lionel, Bhooma Thiruvahindrapuram, Jeffrey R MacDonald, Ryan Mills, Aparna Prasad, Kristin Noonan, Susan Gribble, Elena Prigmore, Patricia K Donahoe, Richard S Smith, Ji Hyeon Park, Matthew E Hurles, Nigel P Carter, Charles Lee, Stephen W Scherer, and Lars Feuk. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, 29:512–520, 2011.

69. Terry Platt. Transcription termination and the regulation of gene expression. *Annual Review of Biochemistry*, 55:339–372, 1986.

70. Jonathan R Pollack, Charles M Perou, Ash A Alizadeh, Michael B Eisen, Alexander Pergamenschikov, Cheryl F Williams, Stefanie S Jeffrey, David Botstein, and Patrick O Brown. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nat. Genet.*, 23:41–46, 1999.

71. J.R Pollack, T. Sø rlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Borresen-Dale, and P.O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. U. S. A.*, 99:12963–12968, 2002.

72. Tatiana Popova, Detlev Mennerich, Andreas Weith, and Karsten Quast. Effect of rna quality on transcript intensity levels in microarray analysis of human postmortem brain tissues. *BMC Genomics*, 9(91), 2008.

73. A. Preumont, R. Rzem, D. Vertommen, and E. Van Schaftingen. Hdhd1, which is often deleted in x-linked ichthyosis, encodes a pseudouridine-5'-phosphatase. *Biochemistry Journal*, 431(2):237–244, 2010.

74. Thomas S. Price, Regina Regan, Richard Mott, Asa Hedman, Ben Honey, Rachael J. Daniels, Lee Smith, Andy Greenfield, Ana Tiganescu, Veronica Buckle,

Nicki Ventress, Helena Ayyub, Anita Salhan, Susana Pedraza-Diaz, John Broxholme, Jiannis Ragoussis, Douglas R. Higgs, Jonathan Flint, and Samantha J. L. Knight. Sw-array: a dynamic programming solution for the identification of copy-number changes in genomic dna using array comparative genome hybridization data. *Nucleic Acids Research*, 33(11):3455–3464, 2005.

75. Tom Price. *cgh: Microarray CGH analysis using the Smith-Waterman algorithm*, 2010. R package version 1.0-7.1.
76. Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue):D61–D65, January 2007.
77. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.
78. Naim U. Rashid, Paul G. Giresi, Joseph G. Ibrahim, Wei Sun, and Jason D. Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, 12:r67, 2011.
79. Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, Michael H. Shapero, Andrew R. Carson, Wenwei Chen, Eun K. Cho, Stephanie Dallaire, Jennifer L. Freeman, Juan R. Gonzalez, Monica Gratacos, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R. MacDonald, Christian R. Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J. Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F. Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P. Carter, Hiroyuki Aburatani, Charles Lee, Keith W. Jones, Stephen W. Scherer, and Matthew E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006.
80. Mark Reimers and John N Weinstein. Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *Bioinformatics*, 6:166–174, 2005.
81. William C. Reinhold, Jean-Louis Mergny, Hongfang Liu, Michael Ryan, Thomas D. Pfister, Robert Kinders, Ralph Parchment, James Doroshow, John N. Weinstein, and Yves Pommier. Exon array analyses across the nci-60 reveal potential regulation of top1 by transcription pausing at guanosine quartets in the first intron. *Cancer Research*, 70(6), 2010.

82. Brooke Rhead, Donna Karolchik, Robert M. Kuhn, Angie S. Hinrichs, Ann S. Zweig, Pauline A. Fujita, Mark Diekhans, Kayla E. Smith, Kate R. Rosenbloom, Brian J. Raney, Andy Pohl, Michael Pheasant, Laurence R. Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A. Harte, Belinda Giardine, Timothy R. Dreszer, Hiram Clawson, Galt P. Barber, David Haussler, and W. James Kent. The UCSC Genome Browser database: update 2010. *Nucleic Acids Research*, 38(suppl 1):D613–D619, 2010.

83. R. A. Rigby and D. M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Staistics and Computing*, 6(1):57–65, 1996.

84. R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Applied Statistics*, 54(3):507–554, 2005.

85. Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011.

86. M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:21:2881–2887, 2007.

87. Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11:R25, 2010.

88. Rickard Sandberg, Joel R. Neilson, Arup Sarma, Phillip A. Sharp, and Christopher B. Burge. Proliferating cells express mrnas with shortened 3' utrs and fewer microrna target sites. *Science*, 320(5883):1643–1647, 2008.

89. F. Sanger and A.R. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

90. M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

91. Almut Schulze and Julian Downward. Navigating gene expression using microarrays - a technology review. *Nature Cell Biology*, 3, 2001.

92. Jonathan Sebat, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, T. Conrad Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305(5683):525–528, July 2004.

93. Uma T Shankavaram, Sudhir Varma, David Kane, Margot Sunshine, Krishna K Chary, William C Reinhold, Yves Pommier, and John N Weinstein. Cellminer: a relational database and query tool for the nci-60 cancer cell lines. *BMC Genomics. 2009; 10: 277.*, 10:277, 2009.

94. K. R. Sherwood, M. W. Head, R. Walker, C. Smith, J. W. Ironside, and J. K. Fazakerley. Rna integrity in post mortem human variant creutzfeldt-jakob disease (vcjd) and control brain tissue. *Neuropathology and Applied Neurobiology*, 37:633–642, 2011.

95. Han Si, Ramandeep S. Banga, Pinelopi Kapitsinou, Manjunath Ramaiah, Janis Lawrence, Ganesh Kambhampati, Antje Gruenwald, Erwin Bottinger, Daniel Glicklich, Vivian Tellis, Stuart Greenstein, David B. Thomas, James Pullman, Melissa Fazzari, and Katalin Susztak. Human and murine kidneys show gender- and species-specific gene expression differences in response to injury. *PLoS One*, 4(3), 2009.

96. G.K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, 2003.

97. Gordon K. Smyth and Arunas Verbyla. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Statis. Soc.*, 58:3:565–572, 1996.

98. Jun S Song, W Evan Johnson, Xiaopeng Zhu, Xinmin Zhang, Wei Li, Arjun K Manrai, Jun S Liu, Runsheng Chen, and X Shirley Liu. Model-based analysis of two-color arrays (ma2c). *Genome Biology*, 8:R178, 2007.

99. Johan Staaf, Goran Jonsson, Markus Ringner, and Johan Vallon-Christersson. Normalization of array-cgh data: influence of copy number imbalances. *BMC Genomics*, 8:382, 2007.

100. D. Mikis Stasinopoulos and Robert A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7), 2007.

101. A. Theisen. Microarray-based comparative genomic hybridization (acgh). *Nature Education*, 1(1), 2008.

102. Danielle Thierry-Mieg and Jean Thierry-Mieg. Aceview: a comprehensive cdna-supported gene and transcripts annotation. *Genome Biology*, 7(Suppl 1):S12, 2006.

103. Wessel N N. Van Wieringen, Mark A A. Van De Wiel, and Bauke Ylstra. Weighted clustering of called array cgh data. *Biostatistics*, 9(3):484–500, December 2007.

104. Marquis P Vawter, Simon Evans, Prabhakara Choudary, Hiroaki Tomita, Jim Meador-Woodruff, Margherita Molnar, Jun Li, Juan F Lopez, Rick Myers, David Cox, Stanley J Watson, Huda Akil, Edward G Jones, and William E Bunney. Gender-specific gene expression in post-mortem human brain: Localization to sex chromosomes. *Neuropsychopharmacology*, 29(2):373–384, 2004.

105. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

106. Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, 10(1):57–63, 2009.

107. M. Weiss, E. Kuipers, C. Postma, A. Snijders, I. Siccama, D. Pinkel, J. Westerga, S. Meuwissen, D. Albertson, and G Meijer. Genomic profiling of gastric cancer predicts lymph node status and survival. *Oncogene*, 22:1872–1879, 2003.

108. David A. Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G. Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L. Turcotte, Gerard P. Irzyk, James R. Lupski, Craig Chinault, Xing-zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M. Muzny, Marcel Margulies, George M. Weinstock, Richard A. Gibbs, and Jonathan M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008.

109. Z Wu, R Irizarry, R Gentleman, and FM Murilloand F Spencer. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, 99:909–917, 2004.

110. Jun Xu, Paul Burgoyne, and Arthur Arnold. Sex differences in sex chromosome gene expression in mouse brain. *Human Molecular Genetics*, 11:1409–1419, 2002.

111. Saliha Yilmaz, Hervé Fontaine, Karène Brochet, Marie-José Grégoire, Marie-Dominique Devignes, Jean-Luc Schaff, Christophe Philippe, Christophe Nemos, John Louis McGregor, and Philippe Jonveaux. Screening of subtle copy number changes in aicardi syndrome patients with a high resolution x chromosome array-cgh. *European Journal of Medical Genetics*, 50(5):386 – 391, 2007.

112. Wei Zhang, R. Stephanie Huang, Shiwei Duan, and M. Eileen Dolan. Gene set enrichment analyses revealed differences in gene expression patterns between males and females. *In Sililico Biology*, 9, 2009.

Appendix A

Supplementary Tables

Table A.1. Size estimates at significance level 0.001 across 64 simulated data sets under various dispersion distribution scenarios and sample sizes

Method	Scenario			
	1	2	3	4
	Marioni (N=18)			
edgeR	0.00192	0.00127	0.00217	0.00275
DESeq Liberal	0.00201	0.00096	0.00691	0.00751
DESeq Conservative	0.00066	0.00049	0.00073	0.00084
gamSeq	0.00103	0.00098	0.00116	0.00110
	Cheung (N=41)			
edgeR	0.00129	0.00132	0.00192	0.00173
DESeq Liberal	0.00502	0.00140	0.00815	0.00856
DESeq Conservative	0.00057	0.00053	0.00050	0.00055
gamSeq	0.00100	0.00099	0.00119	0.00111
	Cheung (N=18)			
edgeR	0.00160	0.00141	0.00252	0.00281
DESeq Liberal	0.00344	0.00109	0.00751	0.01000
DESeq Conservative	0.00087	0.00051	0.00080	0.00096
gamSeq	0.00110	0.00106	0.00124	0.00114
	Cheung(N=15)			
edgeR	0.00177	0.00135	0.00237	0.00335
DESeq Liberal	0.00266	0.00087	0.00576	0.00857
DESeq Conservative	0.00090	0.00044	0.00091	0.00126
gamSeq	0.00110	0.00121	0.00122	0.00114
	Cheung (N=12)			
edgeR	0.00175	0.00136	0.00291	0.00381
DESeq Liberal	0.00213	0.00075	0.00528	0.00768
DESeq Conservative	0.00095	0.00040	0.00107	0.00166
gamSeq	0.00103	0.00105	0.00120	0.00105
	Cheung (N=9)			
edgeR	0.00156	0.00131	0.00310	0.00419
DESeq Liberal	0.00123	0.00067	0.00353	0.00496
DESeq Conservative	0.00085	0.00045	0.00179	0.00269
gamSeq	0.00083	0.00085	0.00149	0.00076

Table A.2. Median (MAD) number of false positives at FDR cutoff of 10% across 64 simulated data sets under various dispersion distribution scenarios and sample sizes

Method	Scenario			
	1	2	3	4
	Marioni (N=18)			
edgeR	3 (3)	0 (0)	7 (3)	10 (5.9)
DESeq Liberal	7 (3.7)	1 (1.5)	102 (11.9)	109 (15.6)
DESeq Conservative	0 (0)	0 (0)	0 (0)	1 (1.5)
gamSeq	0 (0)	0 (0)	0 (0)	0 (0)
	Cheung (N=41)			
edgeR	0 (0)	0 (0)	2 (1.5)	1 (1.5)
DESeq Liberal	40 (8.2)	4 (3)	89 (13.3)	94 (14.1)
DESeq Conservative	0 (0)	0 (0)	0 (0)	0 (0)
gamSeq	0 (0)	0 (0)	0 (0)	0 (0)
	Cheung (N=18)			
edgeR	1 (1.5)	0 (0)	4 (3)	7 (5.9)
DESeq Liberal	16 (8.2)	2 (2.2)	74 (13.3)	116 (16.3)
DESeq Conservative	0 (0)	0 (0)	1 (1.5)	2 (1.5)
gamSeq	0 (0)	0 (0)	1 (0.7)	0 (0)
	Cheung(N=15)			
edgeR	1 (1.5)	0 (0)	5 (4.4)	9 (5.2)
DESeq Liberal	8 (4.4)	1 (1.5)	51 (11.1)	92 (13.3)
DESeq Conservative	0 (0)	0 (0)	1 (1.5)	2 (3)
gamSeq	0 (0)	1 (1.5)	1 (1.5)	1 (0.7)
	Cheung (N=12)			
edgeR	1 (1.5)	0 (0)	8 (4.4)	15 (5.9)
DESeq Liberal	5 (4.4)	1 (1.5)	41 (7.4)	79 (11.1)
DESeq Conservative	1 (0.7)	0 (0)	2 (1.5)	6 (3)
gamSeq	1 (1.5)	2 (1.5)	3 (3)	2 (1.5)
	Cheung (N=9)			
edgeR	1 (1.5)	1 (0)	10 (4.4)	18 (6.7)
DESeq Liberal	1 (1.5)	1 (1.5)	18 (7.4)	33 (8.2)
DESeq Conservative	1 (0)	0 (0)	7 (3)	11 (6.7)
gamSeq	1 (1.5)	2 (1.5)	7 (3)	3 (1.5)

Table A.3. Power estimates at FDR cutoff of 10% across 64 simulated data sets under various dispersion distribution scenarios and sample sizes

Method	Scenario			
	1	2	3	4
	Marioni (N=18)			
edgeR	0.544	0.485	0.493	0.508
DESeq Liberal	0.569	0.503	0.511	0.521
DESeq Conservative	0.545	0.453	0.460	0.472
gamSeq	0.543	0.462	0.499	0.538
	Cheung (N=41)			
edgeR	0.597	0.576	0.538	0.587
DESeq Liberal	0.619	0.597	0.554	0.606
DESeq Conservative	0.583	0.563	0.494	0.566
gamSeq	0.667	0.627	0.589	0.668
	Cheung (N=18)			
edgeR	0.518	0.518	0.464	0.531
DESeq Liberal	0.526	0.529	0.464	0.536
DESeq Conservative	0.477	0.483	0.402	0.488
gamSeq	0.547	0.516	0.450	0.588
	Cheung(N=15)			
edgeR	0.491	0.483	0.430	0.506
DESeq Liberal	0.486	0.475	0.418	0.497
DESeq Conservative	0.434	0.428	0.359	0.454
gamSeq	0.487	0.453	0.408	0.547
	Cheung (N=12)			
edgeR	0.468	0.464	0.395	0.480
DESeq Liberal	0.458	0.451	0.374	0.467
DESeq Conservative	0.400	0.398	0.309	0.422
gamSeq	0.415	0.390	0.315	0.479
	Cheung (N=9)			
edgeR	0.373	0.382	0.268	0.387
DESeq Liberal	0.346	0.355	0.228	0.352
DESeq Conservative	0.283	0.283	0.177	0.309
gamSeq	0.126	0.116	0.058	0.193

Table A.4. Size estimates at significance level 0.001 across 64 simulated data sets in scenario 2 with zero inflation

Method	π			
	0.05	0.10	0.15	0.2
	Cheung (N=41)			
edgeR	0.00042	0.00029	0.00025	0.00022
DESeq Liberal	0.00142	0.00144	0.00156	0.00053
DESeq Conservative	0.00059	0.00076	0.00072	0.00012
gamSeq	0.00090	0.00093	0.00109	0.00113
	Cheung (N=18)			
edgeR	0.00027	0.00036	0.00032	NA
DESeq Liberal	0.00157	0.00232	0.00349	0.00485
DESeq Conservative	0.00096	0.00161	0.00240	0.00382
gamSeq	0.00124	0.00120	0.00134	0.00139
	Cheung(N=15)			
edgeR	0.00032	0.00033	NA	NA
DESeq Liberal	0.00165	0.00267	0.00398	0.00585
DESeq Conservative	0.00124	0.00203	0.00317	0.00474
gamSeq	0.00132	0.00133	0.00157	0.00159
	Cheung(N=12)			
edgeR	0.00027	0.00052	NA	NA
DESeq Liberal	0.00208	0.00414	0.00594	0.00970
DESeq Conservative	0.00161	0.00338	0.00490	0.00799
gamSeq	0.00106	0.00121	0.00137	0.00141
	Cheung (N=9)			
edgeR	0.00036	0.00047	NA	NA
DESeq Liberal	0.00353	0.00709	0.01258	0.01655
DESeq Conservative	0.00302	0.00583	0.01043	0.01362
gamSeq	0.00062	0.00062	0.00062	0.00100

Table A.5. Size estimates at significance level 0.001 across 64 simulated data sets in scenario 4 with zero inflation

Method	π			
	0.05	0.10	0.15	0.2
	Cheung (N=41)			
edgeR	0.00087	0.00066	0.00048	0.00065
DESeq Liberal	0.00793	0.0073	0.00615	0.00443
DESeq Conservative	0.00053	0.00058	0.00058	0.00043
gamSeq	0.00114	0.0011	0.00119	0.00119
	Cheung (N=18)			
edgeR	0.00131	0.00092	0.00144	0.00083
DESeq Liberal	0.00921	0.00897	0.009	0.00949
DESeq Conservative	0.0014	0.00158	0.00237	0.00282
gamSeq	0.00115	0.00116	0.00126	0.00145
	Cheung(N=15)			
edgeR	0.00152	0.00073	0.00069	NA
DESeq Liberal	0.00815	0.00772	0.00818	0.00878
DESeq Conservative	0.00163	0.00231	0.00301	0.00453
gamSeq	0.00143	0.00137	0.00163	0.0016
	Cheung(N=12)			
edgeR	0.00179	0.00096	NA	NA
DESeq Liberal	0.00748	0.00808	0.00922	0.01133
DESeq Conservative	0.00219	0.00351	0.00475	0.00694
gamSeq	0.00101	0.00111	0.00132	0.00147
	Cheung (N=9)			
edgeR	0.00129	NA	NA	NA
DESeq Liberal	0.00619	0.00923	0.01312	0.01703
DESeq Conservative	0.00407	0.00683	0.01069	0.01474
gamSeq	0.00052	0.00053	0.00047	0.00065

Table A.6. Power estimates at 10% FDR cutoff across 64 simulated data sets in scenario 2 with zero inflation

Method	π			
	0.05	0.10	0.15	0.2
	Cheung (N=41)			
edgeR	0.517	0.465	0.41	0.353
DESeq Liberal	0.581	0.56	0.518	0.478
DESeq Conservative	0.545	0.523	0.468	0.387
gamSeq	0.607	0.604	0.587	0.563
	Cheung (N=18)			
edgeR	0.402	0.304	0.225	0
DESeq Liberal	0.502	0.472	0.457	0.423
DESeq Conservative	0.448	0.409	0.383	0.339
gamSeq	0.493	0.484	0.468	0.438
	Cheung(N=15)			
edgeR	0.343	0.239	NA	NA
DESeq Liberal	0.443	0.414	0.381	0.356
DESeq Conservative	0.382	0.351	0.307	0.278
gamSeq	0.418	0.399	0.383	0.398
	Cheung(N=12)			
edgeR	0.313	0.176	NA	NA
DESeq Liberal	0.421	0.392	0.348	0.319
DESeq Conservative	0.36	0.317	0.27	0.24
gamSeq	0.366	0.328	0.3	0.299
	Cheung (N=9)			
edgeR	0.186	0.061	NA	NA
DESeq Liberal	0.312	0.275	0.235	0.213
DESeq Conservative	0.243	0.206	0.174	0.154
gamSeq	0.131	0.062	0.022	0.011

Table A.7. Power estimates at 10% FDR cutoff across 64 simulated data sets in scenario 4 with zero inflation

Method	π			
	0.05	0.10	0.15	0.2
	Cheung (N=41)			
edgeR	0.523	0.467	0.41	0.363
DESeq Liberal	0.592	0.577	0.545	0.469
DESeq Conservative	0.545	0.525	0.496	0.396
gamSeq	0.661	0.642	0.629	0.609
	Cheung (N=18)			
edgeR	0.408	0.316	0.215	0.132
DESeq Liberal	0.508	0.481	0.455	0.425
DESeq Conservative	0.451	0.419	0.387	0.35
gamSeq	0.574	0.581	0.547	0.547
	Cheung(N=15)			
edgeR	0.364	0.246	0.166	NA
DESeq Liberal	0.465	0.434	0.395	0.365
DESeq Conservative	0.412	0.381	0.336	0.294
gamSeq	0.518	0.496	0.483	0.463
	Cheung(N=12)			
edgeR	0.331	0.187	NA	NA
DESeq Liberal	0.437	0.387	0.352	0.314
DESeq Conservative	0.379	0.323	0.288	0.243
gamSeq	0.451	0.403	0.397	0.367
	Cheung (N=9)			
edgeR	0.192	NA	NA	NA
DESeq Liberal	0.31	0.265	0.224	0.197
DESeq Conservative	0.258	0.218	0.176	0.151
gamSeq	0.153	0.133	0.07	0.041

Table A.8. Percentage of genes showing evidence for differential 3' UTR usage and percentage of those genes with using longer 3' UTRs when compared to brain tissue for UTRs of type III

Tissue	Differential Usage	Length
Blood	57.68%	66.91%
Muscle	62.11%	64.44%
Breast	41.18%	58.46%
Kidney	47.90%	58.61%
Lymphnode	58.33%	57.42%
Colon	46.38%	61.76%
Prostate	53.51%	60.38%
Testes	46.16%	69.02%
Hear	39.86%	69.12%
Lung	54.71%	64.77%
Ovary	34.85%	53.86%
Thyroid	39.97%	62.11%
Adrenal	40.96%	52.66%
Adipose	41.90%	68.42%

Table A.9. Percentage of genes showing evidence for differential 3' UTR usage and percentage of those genes with using longer 3' UTRs when compared to brain tissue for UTRs of type II

Tissue	Differential Usage	Length
Blood	53.51%	55.89%
Muscle	50.59%	56.71%
Breast	48.70%	63.61%
Kidney	36.20%	59.83%
Lymphnode	47.69%	57.16%
Colon	36.89%	53.31%
Prostate	37.12%	64.13%
Testes	43.23%	60.85%
Hear	44.21%	54.67%
Lung	36.58%	58.96%
Ovary	41.86%	64.92%
Thyroid	43.13%	53.36%
Adrenal	48.99%	57.73%
Adipose	42.65%	61.45%

Appendix B

Supplementary Figures

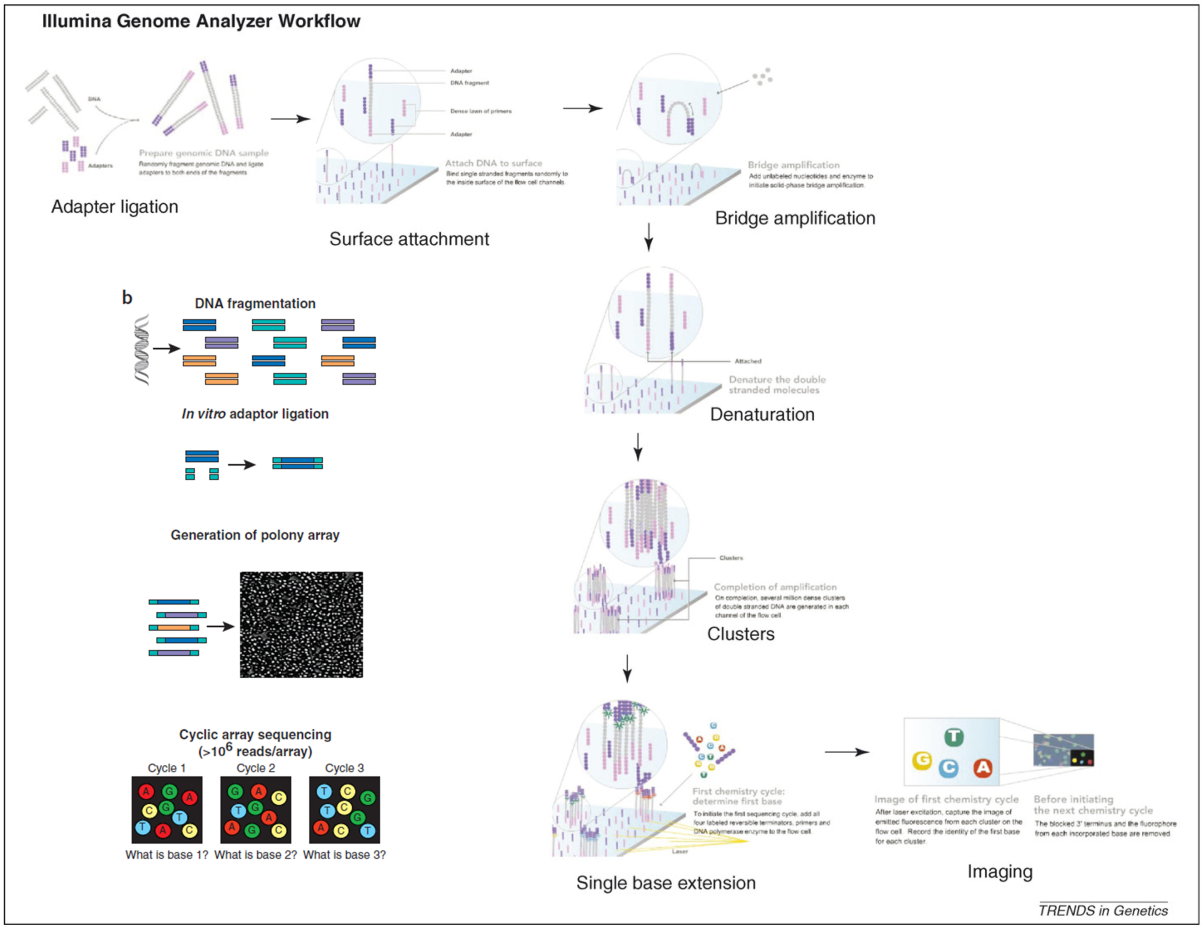

Illumina Genome Analyzer Workflow
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
Adapter ligation
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels
Surface attachment
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification
Bridge amplification
Attached
Denature the double stranded molecules
Denaturation
Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell
Clusters
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell
Laser
Single base extension
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed
Imaging
b
DNA fragmentation
In vitro adaptor ligation
Generation of polony array
Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3
What is base 1?
What is base 2?
What is base 3?
TRENDS in Genetics

Figure B.1. Schematic of Illumina Genome Analyzer 2 workflow[50].

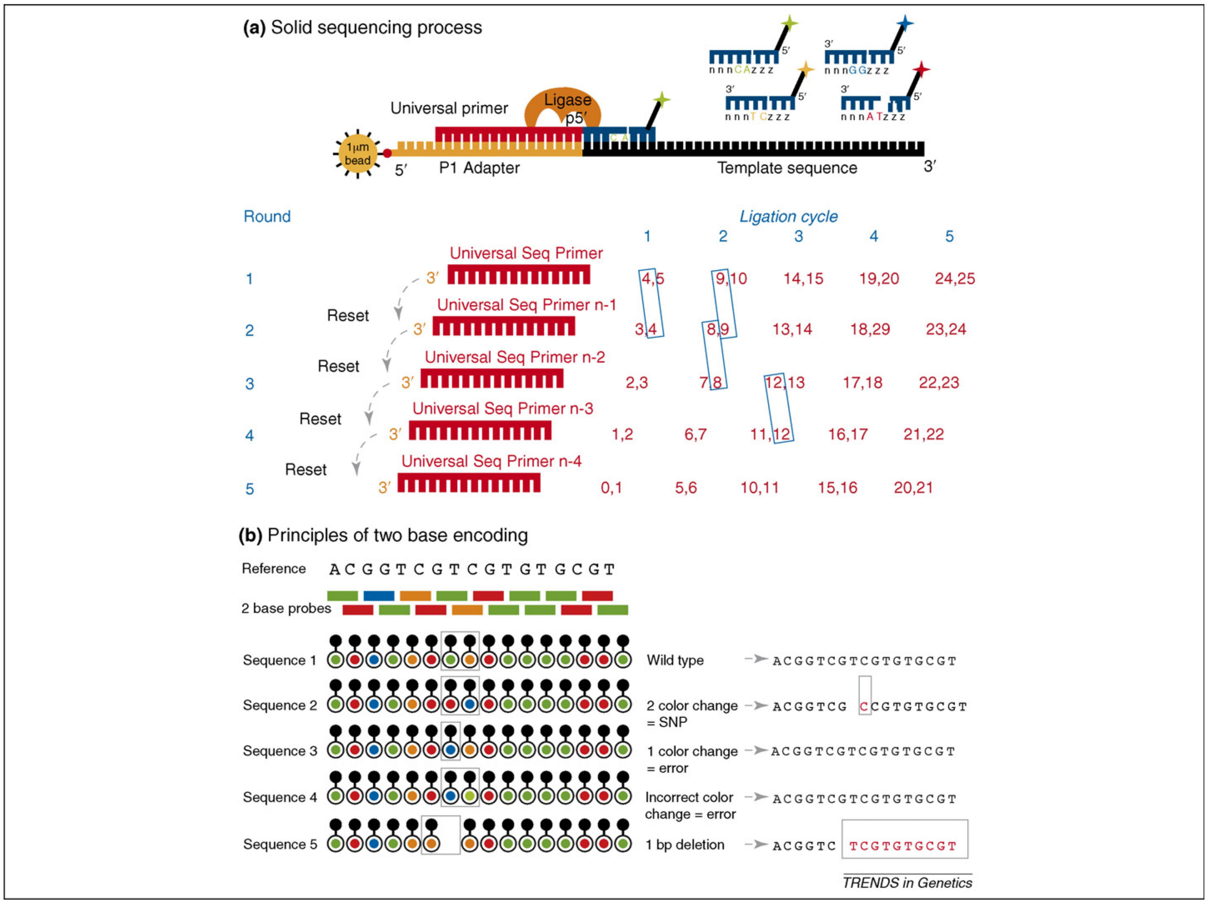


Figure B.2. Schematic of ABI SOLiD workflow [50].

Appendix C

Supplementary Code

This appendix demonstrates how the different methods were used to analyze the Marioni dataset described in Section 3.2.3. The count file is available from `http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE17274`. The other datasets and simulation studies were analyzed in similar fashion.

First, the necessary libraries are loaded and a function `geneChr` is defined that extracts the chromosome a gene is located on given an gene identifier.

```
> library(multicore)
> library(DESeq)
> library(edgeR)
> library(gamSeq)
> geneChr <- function(ids,counts,prin=F){
+   require(org.Hs.eg.db)
+   cat("Number of genes with evidence for DGE:",length(ids),"\n")
+   if(length(ids)>0){
+     cat("Distribution of genes across chromosomes:")
+     print(table(unlist(lapply(rownames(counts)[ids],function(x)
+     tryCatch(get(get(x,org.Hs.egENSEMBL2EG),org.Hs.egCHR),
+     error=function(x) NA)))))
+   }
+ }
```

The Marioni dataset is read in and pre-processed. The Marioni dataset contains one technical replicate for each sample. Only one replicate was used in the analyses presented.

```
> marioniCounts <- read.table("/home/tguennel/ZIM/
+ GSE17274_ReadCountPerLane.txt",as.is=T,sep="\t",header=T)
> rownames(marioniCounts) <- marioniCounts[,1]
> marioniCounts <- marioniCounts[,-1]
> # order by species and gender
> marioniCounts <- marioniCounts[,order(sub("^.*\\.", "",
+ colnames(marioniCounts)))]
> # remove zero rows and replicates
> marioniCounts <- marioniCounts[-which(rowSums(
+ marioniCounts[,seq(2,36,by=2)])==0),-seq(1,36,by=2)]
> species <- rep(c("HS","PT","RM"),each=6)
> gender <- rep(rep(c("F","M"),each=3),3)
> # create data frame
> dataF2 <- data.frame(gender=gender,species=species)
```

Now gamSeq is run in parallel using eight CPUs and the chromosomes for genes called significant are shown.

```
> nCPUs <- 8
> fitGamSeq <- gamSeq(counts=as.matrix(marioniCounts),
+ covariates="gender+species", data=dataF2, offSet=colSums(marioniCounts),
+ numCPUs=nCPUs)

Starting analysis
 1 ... 2 ... 3 ... 4 ... 5 ... 6 ... 7 ... 8 ... 9 ... 10 ... 11 ...
 12 ... 13 ... 14 ... 15 ... 16 ... 17 ...
Analysis completed

> geneChr(which(p.adjust(fitGamSeq$pValue$genderM,method="BH")<0.1),
+ counts=marioniCounts)

Number of genes with evidence for DGE: 0
```

The same is repeated for edgeR and DESeq using code as suggested by the package's authors.

```
> design <- model.matrix(~gender + species ,data=dataF2)
> edgeRdata <- DGEList(counts=marioniCounts,
+ lib.size=colSums(marioniCounts), group=dataF2$gender)
> edgeRdata <- calcNormFactors(edgeRdata)
> edgeRdata <- estimateGLMCommonDisp(edgeRdata,design)
> edgeRdata <- estimateGLMTrendedDisp(edgeRdata,design)
> edgeRdata <- estimateGLMTagwiseDisp(edgeRdata,design)
> fitER <- glmFit(edgeRdata,design=design)
> lrtX1 <- glmLRT(edgeRdata,fitER,coef=c("genderM"))
> geneChr((1:nrow(marioniCounts))[-fitER$not.converged][which(p.adjust
+ (lrtX1$table$p.value[-fitER$not.converged],method="BH")<0.1)],
+ counts=marioniCounts)
```

```
Number of genes with evidence for DGE: 10
Distribution of genes across chromosomes:
 1 10 11 12 14 15 19  2 20  6
 1  1  1  1  1  1  1  1  1  1
```

First DESeq Conservative is run.

```
> dataF3 <- data.frame(condition=gender,species=species)
> cds <- newCountDataSet( marioniCounts, conditions=dataF3)
> cds <- estimateSizeFactors(cds)
> cds <- estimateDispersions(cds,method="pooled",sharingMode="max")
> fit1M <- fitNbinomGLMs(cds,count~condition+species,
+ glmControl=list(maxit=1000))
```

```
.................
```

```
> fit0M <- fitNbinomGLMs(cds,count~species,glmControl=list(maxit=1000))
```

```
.................
```

```
> pGenderDESPooled <- nbinomGLMTest(fit1M,fit0M)
> pGenderDESPooled[union(which(!fit0M$converged),
+ which(!fit1M$converged))] <- NA
> geneChr(which(p.adjust(pGenderDESPooled,method="BH")<0.1),
+ counts=marioniCounts)
```

```
Number of genes with evidence for DGE: 0
```

Now the same for DESeq Liberal.

```
> # DESeq-Blind
>
> cds2 <- newCountDataSet( marioniCounts, conditions=dataF3)
> cds2 <- estimateSizeFactors( cds2 )
> cds2 <- estimateDispersions(cds2, method = "blind",
+ sharingMode="fit-only")
> fit1M2 <- fitNbinomGLMs(cds2,count~condition+species,
+ glmControl=list(maxit=1000))

................

> fit0M2 <- fitNbinomGLMs( cds2, count ~   species,
+ glmControl=list(maxit=1000))

................

> pGenderDESBlind<- nbinomGLMTest( fit1M2, fit0M2 )
> pGenderDESBlind[union(which(!fit1M2$converged),
+ which(!fit0M2$converged))] <- NA
> geneChr(which(p.adjust(pGenderDESBlind,method="BH")<0.1),
+ counts=marioniCounts)

Number of genes with evidence for DGE: 19
Distribution of genes across chromosomes:
 1 10 11 12 13 14 15 17 19  2 20  4  6
 2  1  1  2  1  1  1  1  2  2  1  1  2
```

The follwing R version and package versions were used for all analyses.

- R version 2.14.0 (2011-10-31), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`

- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: akima 0.5-4, AnnotationDbi 1.16.0, Biobase 2.14.0, DBI 0.2-5, DESeq 1.6.0, edgeR 2.4.0, gamlss 4.1-1, gamlss.data 4.0-5, gamlss.dist 4.1-0, gamSeq 0.1.0, lattice 0.20-0, locfit 1.5-6, MASS 7.3-16, multicore 0.1-7, nlme 3.1-102, org.Hs.eg.db 2.6.4, RSQLite 0.10.0
- Loaded via a namespace (and not attached): annotate 1.32.0, genefilter 1.36.0, geneplotter 1.32.1, grid 2.14.0, IRanges 1.12.1, limma 3.10.0, RColorBrewer 1.0-5, survival 2.36-10, tools 2.14.0, xtable 1.6-0

Vita

Virginia Commonwealth University
Department of Biostatistics
Box 980032
Richmond, VA 23298-0032
Date of Birth: April 28, 1982
Citizenship: Germany

Phone: (804) 822-5082
Fax: (804) 828-8900
Email: `tobiasguennel@gmail.com`

Education

Ph.D. Biostatistics, Virginia Commonwealth University, *expected* December 2011.

- *Concentration*: Genomics and Statistical Genetics.

Dipl.-Math. techn., Chemnitz University of Technology, Germany, 2001–2008.

- equivalent to M.S. Applied Mathematics.
- *Minors*: Mechanical Engineering and Computer Science.
- *Thesis*: Ordinal Classification Approach using Bagged Classification Trees and the Proportional
- Odds Model as Splitting Criteria.

B.S. Mathematics, Longwood University, 2004–2006.

- *Minor*: Computer Science.
- *Honors*: Summa Cum Laude, Phi Kappa Phi.

Research Interests

Microarray Data Analysis, High Throughput Sequencing Data Analysis, Statistical Challenges in Epigenetics, Statistical Genetics, Pharmacogenomics, Classification Systems.

Academic and Professional Experience

BioStat Solutions Inc., Mount Airy, MD

- Statistical Intern, January 2011-present.

Virginia Commonwealth University, Department of Biostatistics

- Predoctoral Fellow supported by National Institute on Drug Abuse Training Grant, Mark Reimers, Ph.D., January 2010-Present.
- Graduate Assistant, Kellie J. Archer, Ph.D., Summer 2008-December 2009.
- Teaching Assistant, Al M. Best, Ph.D., Fall 2007-Spring 2008.
- Trainee, Kellie J. Archer, Ph.D., Summer 2006.

Longwood University, Department of Mathematics and Computer Sciences

- Mathematics Tutor, Learning Center, Fall 2004-Spring 2006.

Research

Work in Progress

- Technical variable normalization of two color CGH arrays, with Mark A. Reimers, Ph.D..
- Quality assessment, normalization, and analysis of High-throughput Sequencing data.

Publications in Refereed Journals

- Integrating RNA-Seq and genome-wide association to identify risk genes for schizophrenia, Chen X, Zhong, Y, XU J, Peng Z, Guennel T, Reimers M, Bacanu S, Zhongming Z, and Kendler KS, *Abstract for International Society of Psychiatric Genetics Symposium*, 2011.
- Comparing performance of multi-class classification systems with ROC manifolds: When volume and correct classification fails, Schubert CA and Guennel T, *Communications in Statistics - Simulation and Computation*, under review, 2011.
- Spatiotemporal transcriptome of the human brain, Kang HM, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa1 AM, Pletikos M, Meyer KM, Guennel T, Sedmak G, Shin Y, Johnson MB, Krsnik Z, Fertuzinhos S, Umlauf S, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE, and Šesta N, *Nature*, Vol. 478(7370): 483-489, 2011.
- Noise reduction for aCGH data using technical covariates and probe level information, Guennel T, Reinhold WC, Pommier Y, Selzer R, Weinstein JN, and Reimers

M, *Bioinformatics*, under review, 2011.

- Identifying genes progressively silenced in preneoplastic and neoplastic liver tissues, Archer KJ, Zhao Z, Guennel T, Maluf DG, Fisher RA, and Mas VR, *International Journal of Computational Biology and Drug Design*, Vol. 3, No. 1, 52-67, 2010.
- An application for assessing quality of RNA hybridized to Affymetrix GeneChips, Archer KJ and Guennel T, *Bioinformatics* 22: 2699-2701, 2006.

Scientific Software

- logitT R package, A BioConductor R package implementing the Logit-t algorithm introduced in ”A high performance test of differential gene expression for oligonucleotide arrays” by William J Lemon, Sandya Liyanarachchi and Ming You for use with Affymetrix data stored in an AffyBatch object in R. (2008).
- PixelAnalyzer, An application for assessing quality of RNA hybridized to Affymetrix GeneChips with Kellie J. Archer (2006).

Conference Presentations

- ”Noise Reduction for array CGH Data Using Technical Covariates and Probe-Level Information”, National Institute for Health, Bethesda, MD, September 2, 2009.
- ”Noise Reduction for array CGH Data Using Technical Covariates and Probe-Level Information”, Joint Statistical Meetings, Washington, D.C., August 5, 2009.
- ”Statistical Issues in High-throughput Sequencing”, 2^{nd} SEQC Face-to-face Meeting, Little Rock, Arkansas, March 9, 2009.
- ”Ordinal Classification Approach using Bagged Classification Trees and the Proportional Odds Model as Splitting Criteria”, Daniel T. Watts Research Symposium, Virginia Commonwealth University, October 28, 2008.
- ”Ordinal Classification Approach using Bagged Classification Trees and the Proportional Odds Model as Splitting Criteria”, Joint Statistical Meetings, Denver, Colorado, August 4, 2008.

Teaching

Professional Workshops

- Lab manager for Integrative Statistical Analysis of Genome Scale Data, Cold Spring Harbor Laboratory, June 2010.

Graduate Courses

- Teaching assistant for Statistical Methods I and II, Virginia Commonwealth University,
  Fall 2007–Spring 2008.

Undergraduate Courses

- Tutor for Pre-Calculus and Finite Mathematics, Longwood University, Fall 2004–Spring 2006.
- Pre-Calculus, Longwood University, Fall 2005.

Professional Activities

Service

- President of the Graduate Student Association at Virginia Commonwealth University , March 2010–May 2011.
- Student representative serving on the Virginia Commonwealth University Alumni Association Board of Directors, August 2010–present.
- Member of the Virginia Commonwealth University Strategic Plan Recalibration Task Force, May 2010–May 2011.
- Student representative serving on the Virginia Commonwealth University Student Leadership Committee, March 2010–May 2011.

- Graduate School representative serving on the Virginia Commonwealth University Council, August 2010–May 2011.
- Communications Director serving on the Virginia Commonwealth University Graduate Student Association Executive Committee, January 2010–May 2011.
- Student representative serving on the Virginia Commonwealth University Student Health Advisory Committee, May 2009–May 2010.
- Member of the Virginia Commonwealth University Graduate Student Association Executive Council, May 2009–present.

Professional Memberships

- American Statistical Association, 2007–Present.

Conferences and Workshops Attended

- ENAR Spring Meetings, Miami, Fl, March 20–23, 2011.
- 3^{rd} SEQC Face-to-face Meeting, Bethesda, Maryland, December 6, 2010.
- International Workshop on Statistical Genetics and Methodology of Twin and Family Studies, Boulder, CO, March 1–5, 2010.
- Critical Assessment of Massive Data Analysis, Chicago, IL, October 5–6, 2009.
- Joint Statistical Meetings, Washington, D.C., August 1–6, 2009.
- 2^{nd} SEQC Face-to-face Meeting, Little Rock, Arkansas, March 9–10, 2009.
- Joint Statistical Meetings, Denver, Colorado, August 3–7, 2008.

Honors & Awards

Awards

- University Leadership Award, Virginia Commonwealth University, 2011.
- MCV Alumni Association of VCU Scholarship, Virginia Commonwealth University, 2011.
- Who's Who Among Students in American Universities, Virginia Commonwealth University, 2011.
- Phi Kappa Phi Scholarship, School of Medicine, Virginia Commonwealth University, 2010.
- Student Summer Research Project Award, Department of Biostatistics, Virginia Commonwealth University, 2009.
- John C. Forbes Research Colloquium Award, School of Medicine, Virginia Commonwealth University, 2009.
- Charles C. Clayton Fellowship, School of Medicine, Virginia Commonwealth University, 2009.
- GlaxoSmithKline Scholar Award, Department of Biostatistics, Virginia Commonwealth University, 2007.
- President's List, Longwood University, Spring 2006.
- Dean's List, Longwood University, Fall 2004–Fall 2005.

Honorary Societies

- Golden Key International Honor Society, 2008–present.
- Phi Kappa Phi, 2006–present.

Computing Experience

- *Mastery*: ASP, C, C++, JMP, JSP, Perl, LaTeX, Matlab, R, SAS, SQL Server, Unix, Visual Basic.
- *Familiar*: Fortran, Java, Mathematica, Maple, Python.

Relevant Coursework

- **Statistics**: Advanced Inference I/II, Analysis of Biomedical Data, Analysis of Categorical Data, Applied Bayesian Biostatistics, Applied Statistics I/II, Biostatistical Computing, Biostatistical Consulting, Clinical Trials, High Throughput Data Analysis, Linear Models, Mathematical Statistics I/II, Multivariate Analysis I/II, Nonlinear Models, Statistical Methods for Microarray Data I/II.
- **Genetics**: Advanced Human Genetics, Molecular Genetics, Sequence Analysis in Biological Systems, Statistical Genetics.
- **Computer Science**: C++, Computer Organization, Databases, Fortran, Java, Java Server Pages, Matlab, Maple, Network Theory.
- **Mathematics**: Algebra, Analysis I/II/III, Complex Analysis, Finite Element Analysis, Linear Algebra, Numeric Analysis, Ordinary Differential Equations, Partial

Differential Equations, Probability Theory, Queuing Theory.

- **Engineering**: Control Engineering, Higher Technical Mechanics, Measurement Engineering, Stochastic Modeling of Complex Production Systems, Technical Mechanics I/II/III.