Università degli Studi di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

SIMULTANEOUS INFERENCE FOR RNA-SEQ DATA

Direttore della Scuola: Prof.ssa Alessandra Salvan

Supervisore: Prof.ssa Monica Chiogna

Co-supervisori: Prof.ssa Sandrine Dudoit, Dott.ssa Chiara Romualdi

Dottorando: DAVIDE RISSO

31 Gennaio 2012

Riassunto

Negli ultimi anni il sequenziamento massivo di RNA (RNA-Seq) è diventato una scelta frequente per gli studi di espressione genica. Questa tecnica ha il potenziale di superare i *microarray* come tecnica standard per lo studio dei profili trascrizionali. A livello genico, i dati di RNA-Seq si presentano sotto forma di conteggi, al contrario dei *microarray* che stimano l'espressione su una scala continua. Questo porta alla necessità di sviluppare nuovi metodi e modelli per l'analisi di dati di conteggio in problemi con dimensionalità elevata.

In questa tesi verranno affrontati alcuni problemi relativi all'esplorazione e alla modellazione dei dati di RNA-Seq. In particolare, verranno introdotti metodi per la visualizzazione e il riassunto numerico dei dati. Inoltre si definirà un nuovo algoritmo per il raggruppamento dei dati e alcune strategie per la normalizzazione, volte a eliminare le distorsioni specifiche di questa tecnologia. Infine, verrà definito un modello gerarchico Bayesiano per modellare l'espressione di dati RNA-Seq e verificarne le eventuali differenze in diverse condizioni sperimentali. Il modello tiene in considerazione la profondità di sequenziamento e la sovra-dispersione e automaticamente sviluppa diversi tipi di normalizzazione.

Abstract

In the last few years, RNA-Seq has become a popular choice for high-throughput studies of gene expression, revealing its potential to overcome microarrays and become the new standard for transcriptional profiling. At a gene-level, RNA-Seq yields counts rather than continuous measures of expression, leading to the need for novel methods to deal with count data in high-dimensional problems.

In this Thesis, we aim at shedding light on the problems related to the exploration and modeling of RNA-Seq data. In particular, we introduce simple and effective ways to summarize and visualize the data; we define a novel algorithm for the clustering of RNA-Seq data and we implement simple normalization strategies to deal with technology-related biases. Finally, we present a hierarchical Bayesian approach to the modeling of RNA-Seq data. The model accounts for the difference in sequencing depth, as well as for overdispersion, automatically accounting for different types of normalization.

Acknowledgments

First of all, I would like to thank my supervisor, Monica Chiogna, for her support and advice, and for giving me the opportunity to work on such an interesting topic.

I am very grateful to Chiara Romualdi for conveying to me her passion for research: without her I wouldn't have started my PhD program.

When I arrived in Berkeley for the first time I didn't imagine to find such a great person and mentor as Sandrine Dudoit: I was very lucky to meet her and work with her and I'm really grateful for everything she taught me.

I am grateful to Gavin Sherlock for all the help with the cell biology, which is still magic for me, and to Martin Morgan and the Bioconductor people for the help with the implementation of the R package.

A special mention goes to my lab mates Enrica, Gabriele and Paolo for all the amazing time we spent together inside and outside the lab and for everything they taught me even if they don't know! Well... Gabriele knows...

To the XXIV cycle of the PhD school of Statistics: Antonio, Francesca, Marlies, Nicola and Riccardo for the discussions and the laughs we had, especially during the first year.

To Abha, Yuval and all the Speed/Dudoit group members for all the discussions about work and for those over a beer, and for the great singing performance at the retreat!

To my family, to my parents Graziella and Mauro and to my sister Giulia for their support.

To all my friends in Genova, Padova and Berkeley,

Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Gene expression studies aim to quantify the messenger RNA (mRNA) in a cell as a key purpose for better understanding gene functions, regulations, and interactions. Abnormally high (or low) gene expression can lead to pathologies and different gene profiles can cause different responses to drugs and treatments. The simplest way to identify genes of potential interest through several related experiments is to search for those that are consistently either up- or down-regulated, i.e., differentially expressed between two (or more) conditions.

Microarrays have been the technology of choice for large-scale studies of gene expression since the mid-1990s. Recently, a new technology, the so-called next-generation sequencing (also known as ultra high-throughput sequencing, deep sequencing), has been successfully applied to study genome-wide transcription levels, in what is called RNA-Seq (Wang *et al.*, 2009). The main difference between the two technologies is that RNA-Seq yields counts rather than continuous measures of expression.

Although there is a vast statistical literature on microarray data, little has been done for the understanding and the analysis of RNA-Seq. The first article on RNA-Seq has been published in 2008; since then, several authors have assessed technical aspects of the assay, showing good reproducibility and significant improvements over microarrays in terms of greater dynamic range and more accurate expression estimation (Bullard *et al.*, 2010a; Marioni *et al.*, 2008; Mortazavi *et al.*, 2008). Nonetheless, major technology-related artifacts and biases affect the expression measures (Bullard *et al.*, 2010a; Hansen *et al.*, 2011; Pickrell *et al.*, 2010) and, as in microarrays, normalization remains an important issue, despite initial optimistic claims such as: "One particularly powerful advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets" (Wang *et al.*, 2009).

From a modeling perspective, the Poisson distribution has been proposed to find differentially expressed genes, as it is the simplest model to deal with count data (Bullard *et al.*, 2010a; Marioni *et al.*, 2008). However, several authors have observed overdispersion when the experiment involves biological replicates (i.e., different individuals for each condition), proposing the negative binomial as a more flexible alternative (Anders and Huber, 2010; Robinson *et al.*, 2010). When the sample size is small, the dispersion parameter is estimated by pooling the expression of all genes and considering it either common (Robinson *et al.*, 2010) or as a smooth function that varies across the mean expression range (Anders and Huber, 2010). Despite such results, there is still a lack of understanding on the peculiarities and biases of the technology, as well as on the best way to model the data.

This Thesis aims at shedding light on the problems related to the exploration and modeling of RNA-Seq data. In particular, tools for data visualization and mining are developed and implemented. Moreover, the potential of Bayesian hierarchical modeling is explored with respect to the issue of finding differential expression.

After an overview of the Thesis in Chapter 1, Chapter 2 gives a brief description of the biological and technological background. In Chapter 3, we describe two different datasets from real experiments, which involve different designs and organisms; through exploratory data analysis we highlight some key features of the data. In Chapter 4, we show how technology-related biases can affect inference and we propose some simple data transformations (normalizations) to deal with them. In Chapter 5, we introduce a novel algorithm for the clustering of RNA-Seq data, which is a non-standard problem given the high dimensionality of the data. In Chapter 6, we propose a hierarchical Bayesian model which simultaneously addresses the normalization procedure and the modeling of the biological variability (over-dispersion), borrowing strength from the ensemble of the genes in the testing for differential expression. Finally, we conclude in Chapter 7 with a general discussion and some possibilities for future work.

1.2 Main Contributions of the Thesis

Given the high dimensionality of the data, even simple tasks such as data visualization and cluster analysis become challenging both from a computational and a statistical point of view. The "digital" nature of the assay leads to a multivariate count distribution, with the number of variables far exceeding the number of observations (e.g. the "MAQC-2" dataset – introduced in Chapter 3 – has 12,340 variables and 14 observations).

With respect to the current literature, the contributions of this Thesis may be summarized as follows.

- Introduction of a simple and effective way to summarize and visualize RNA-Seq data.
- Implementation of the R package *EDASeq* for the exploratory data analysis and normalization of RNA-Seq data (Risso and Dudoit, 2011), freely available within the Bioconductor project (Gentleman *et al.*, 2004) at `http://bioconductor.org`.
- Definition of a novel algorithm for the clustering of RNA-Seq data, based on a variance stabilizing transformation, feature selection, and nonparametric density estimation, extending the algorithm in De Bin and Risso (2011).
- Definition of novel and simple normalization strategies to deal with technology-related biases (e.g., GC-content bias) that can strongly impact differential expression analysis (Risso *et al.*, 2011).
- Introduction of a hierarchical Bayesian approach to simultaneously model gene expressions, allowing for overdispersion, and dependence on technology-related biases. The resulting model borrows strength from the ensemble of the genes, gaining power on the testing for differential expression.

Chapter 2

Background

In this Chapter, we provide an introduction to the molecular biological details, relevant for the understanding of the problem that we will statistically tackle throughout this Thesis. In Section 2.1, we will introduce the concept of gene expression and of differential expression between two or more conditions. In Section 2.2, we will overview the so-called next-generation sequencing technology and its application to the sequencing of RNA (RNA-Seq). Finally, in Section 2.3, we will introduce the notation employed in the rest of the Thesis.

2.1 Gene Expression

The main biological concept that drives the experiments analyzed in this Thesis is known as the "central dogma of molecular biology" (Crick, 1970). Essentially, the central dogma states that DNA is transcribed into RNA and subsequently RNA is translated to protein (Alberts *et al.*, 2007). The set of proteins synthesized by the cell determines the phenotype, and it is therefore of primary interest for molecular biologists. The direct measurement of protein abundance can be difficult and the measure of RNA content in a cell can be viewed as a proxy for this quantity. Under the assumption of translation proportional to RNA quantity, RNA abundance estimates can provide a representative measure of the eventual protein product (Bullard, 2009).

2.1.1 Differential Expression

Ideally, the goal of molecular biology is to completely characterize a cell's *transcriptome* (i.e., the set of all RNA in a cell). However, practical limitations make it difficult to directly measure the *absolute* abundance of individual genes. For this reason, gene expression literature has focused on *differential expression* (DE), i.e., the *relative* expression of each gene between samples. Typically, one considers two or more conditions of interest (e.g., cancer vs. normal tissues) to determine direction and magnitude of the change in RNA abundance (Bullard, 2009).

2.2 RNA Sequencing

Since the 1990s, microarrays have been the technology of choice for high-throughput gene expression studies. Although they allowed researcher to have a global picture of the cell at a molecular level, unfeasible with previously available technologies, microarrays have several limitations: (i) background levels of hybridization limit the accuracy of expression measurements; (ii) they supply a relative measure of expression rather than an absolute estimate of the abundance of a transcript; (iii) probe design limits the researchers to study only known transcripts.

More recently, several companies have commercialized machines for the generation of high-throughput sequencing data, e.g., Applied Biosystems' SOLiD, Helicos BioSciences' HeliScope, Illumina's Genome Analyzer, and Roche's 454 Life Sciences sequencing systems. In the last few years, these technologies have been replacing microarrays as the assays of choice for measuring genome-wide transcription levels, in so-called *RNA-Seq* (Nagalakshmi *et al.*, 2008; Wang *et al.*, 2009), as well as DNA copy number (*DNA-Seq*), protein-nucleic acid interactions (*ChIP-Seq*), and DNA methylation (*methyl-Seq* and *RRBS*).

RNA-Seq have gain popularity since it allows the researcher to conduct studies unfeasible with microarrays platforms, e.g., the study of previously unknown transcripts or the study of non-model organisms for which the genome has not been sequenced yet (Wang *et al.*, 2009). Moreover, several authors have shown significant improvements over microarrays in terms of greater dynamic range and more accurate expression estimation (Bullard *et al.*, 2010a; Marioni *et al.*, 2008; Mortazavi *et al.*, 2008).

In this Thesis, we will focus on RNA-Seq, using data from Illumina's Genome Analyzer II (Bentley *et al.*, 2008). However, since the other sequencers differ only by how they generate the sequences and the "raw" data are small RNA sequences for all the technologies, we believe that most of the results will be easily generalizable to other assays.

We will not go much into the details of the technology. See Bentley *et al.* (2008) for the details on the chemistry behind the sequencer. Briefly, RNA is converted to cDNA fragments which are then sequenced to produce millions of short sequences (typically of 25–100 bases) called *reads*.

In the Genome Analyzer the cDNA fragments are placed on a *flow-cell*, which contains eight *lanes*, each yielding several millions of reads. The process of obtaining the cDNA fragments from the sample of interest is called *library preparation*; each *library* can be sequenced using one or multiple lanes, possibly across multiple flow-cells. Therefore, the statistical unit is the lane and different levels of replication can be performed.

2.2.1 Technical and biological replicates

By *technical replicate lanes*, we refer to lanes assaying libraries that differ only by virtue of the sequencing assay (i.e., library preparation, flow-cell, lane), not in terms of the biology (i.e., growth condition or culture for the Yeast dataset, UHR vs. Brain for the MAQC-2 dataset – see Chapter 3). By *biological replicate lanes*, we refer to lanes assaying libraries that are distinct independently of/prior to the sequencing assay (i.e., libraries Y1, Y2, Y4, and Y7, for different cultures of the same yeast strain under the same condition for the Yeast dataset of Chapter 3).

There are therefore different levels/types of technical replication, depending on which aspect of the assay is varied (i.e., library preparation, flow-cell, lane). In particular, we expect library preparation effects to be greater than flow-cell effects and flow-cell effects to be greater than lane effects. The results in Bullard *et al.* (2010a) are consistent with this expectation. Likewise, there are different levels/types of biological replication. Reassuringly, we have found that biological effects of interest tend to be greater than "nuisance" technical effects. Furthermore, it is possible for biological effects to be confounded with technical effects, as is the case with culture and library preparation effects for the Yeast dataset introduced in Chapter 3.

2.2.2 Mapping and Quantifying Expression

The first step for the quantification of gene expression from RNA-Seq data is to *align* (or *map*) the reads to the reference genome. Several algorithms have been proposed; here, we use Bowtie (Langmead *et al.*, 2009), considering only the reads that map uniquely to the genome with up to two mismatches. For the remainder of the Thesis, we will consider gene-level counts, i.e., the number of reads that fall within a given gene. The read count for a given gene is defined as the number of reads with $5'$-end falling within the corresponding region. We define the gene region, following the union-intersection model of Bullard *et al.* (2010a). Briefly, we define a gene as the union of all the exons that belong exclusively to that gene. See Bullard *et al.* (2010a) for details.

Note that, in complex organisms, each gene can have different *isoforms* and, in general, it would be more precise to measure expression at a transcript level, rather than at a gene level. However, it is not straightforward to obtain transcript-level counts, since different isoforms share the same exons and it is not always clear how to associate a read to an isoform in an unambiguous way. To keep things simple, we decided here to focus on gene-level counts. However, some work has been done to compute transcript-level counts (Trapnell *et al.*, 2010) and future effort will be focused in this direction.

2.3 Notation

We will denote with y_{ij} the expression of gene j ($j = 1, \dots, p$) in sample i ($i = 1, \dots, n$). When we will consider only one sample (gene) we will drop the i (j) index. We will denote the design matrix of a given regression model with X and with gc_j the GC-content (i.e., proportion of G and C nucleotides in the gene sequence) of gene j.

Chapter 3

The Data

The data that we will consider in the remainder of this Thesis are gene-level counts, i.e., the aggregated counts of mapped reads that fall within the genes (see Section 2.2.2 for details on the definition of the genes and on the mapping algorithm). For each dataset, we will briefly describe the experimental design and the sample preparation, and we will explore some key features of the data.

We shall make use of three datasets: the first two datasets, publicly available through the MAQC project (Shi *et al.*, 2006), have become important benchmarking data, since the same samples have been assessed with three different technologies (see Section 3.1), allowing for the distinction between biological signals and technology-related bias. Unfortunately, the MAQC datasets do not contain any biological replicate, making it impossible to assess some important features of the data, such as biological variability. For this reason, we also consider a third dataset designed to answer a biologically interesting question, namely the "Yeast" dataset of Section 3.2.

When dealing with new technologies, Exploratory Data Analysis (EDA) becomes an essential step, as it can discover hidden features of the data and reveal new sources of bias and variation. For this reason, we will present the datasets along with graphic summaries to highlight some key feature of the data.

3.1 MAQC Datasets

Bullard *et al.* (2010a) analyzed RNA-Seq data for two types of biological samples from the MicroArray Quality Control (MAQC) Project (Shi *et al.*, 2006): Ambion's

Table 3.1: *MAQC-2 dataset: Experimental design.*

Lane	Library prep.	Sample	Flow-cell
1	L1	UHR	F2
2	L2	Brain	F2
3	L1	UHR	F2
4	L2	Brain	F2
6	L1	UHR	F2
7	L2	Brain	F2
8	L1	UHR	F2
1	L2	Brain	F3
2	L1	UHR	F3
3	L2	Brain	F3
4	L1	UHR	F3
6	L2	Brain	F3
7	L1	UHR	F3
8	L2	Brain	F3

human brain reference RNA (Brain), pooled from multiple donors and several brain regions, and Stratagene's universal human reference RNA (UHR), a mixture of total RNA extracted from 10 different human cell lines. RNA-Seq was performed using Illumina's Genome Analyzer II high-throughput sequencing system. The data are summarized below; additional details about experimental design, pre-processing, and the associated qRT-PCR and microarray datasets can be found in Bullard *et al.* (2010a).

In dataset "MAQC-2", Brain and UHR RNA were sequenced each using a single library preparation and seven lanes distributed across two flow-cells (i.e., technical replicates, Table 3.1). There are no biological replicates and library preparation effects are confounded with the extreme differential expression that one expects when comparing such different samples as Brain and UHR. Nonetheless, the availability of qRT-PCR measures for a subset of about 1,000 genes makes this a valuable benchmarking dataset.

In dataset "MAQC-3", four different library preparations of UHR RNA were each

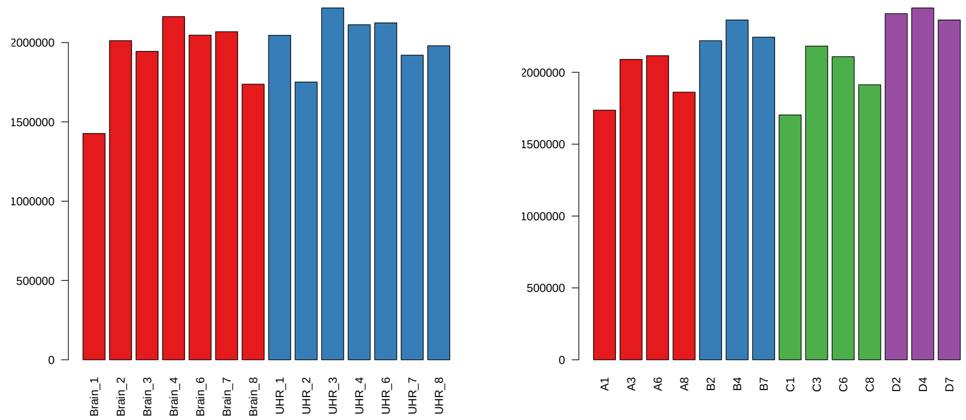
Table 3.2: *MAQC-3 dataset: Experimental design.*

Lane	Library prep.	Sample	Flow-cell
1	A	UHR	F4
2	B	UHR	F4
3	A	UHR	F4
4	B	UHR	F4
6	A	UHR	F4
7	B	UHR	F4
8	A	UHR	F4
1	C	UHR	F5
2	D	UHR	F5
3	C	UHR	F5
4	D	UHR	F5
6	C	UHR	F5
7	D	UHR	F5
8	C	UHR	F5

sequenced using three or four lanes from only one of two flow-cells (Table 3.2). One can use this dataset for examining technical effects such as library preparation and lane effects. However, library preparation effects are nested within flow-cell effects, so that differences between flow-cells are confounded with library preparation effects.

For both the MAQC-2 and MAQC-3 datasets, reads were mapped to the genome (GRCh37 assembly) using Bowtie (Langmead *et al.*, 2009), with unique mapping and up to two mismatches. Gene-level counts were obtained using the *union-intersection* (UI) gene model of Bullard *et al.* (2010a). For the MAQC-2 dataset, genes with an average read count below 10 for both samples were filtered out, i.e., gene j was filtered out if $\max_{k\in\{\text{Brain},\text{UHR}\}} \bar{y}_{j,k} < 10$, where $\bar{y}_{j,k}$ denotes the average read count for gene j in condition k. This procedure retained 12,340 out of 39,359 genes. An analogous procedure was carried out for MAQC-3: genes with an average read count below 10 for each of the four libraries were filtered out, yielding 11,847 (out of 39,359) genes.

Figure 3.1 shows the total number of mapped reads per lane for both MAQC datasets. One can see that the technical replicate lanes can yield a different number



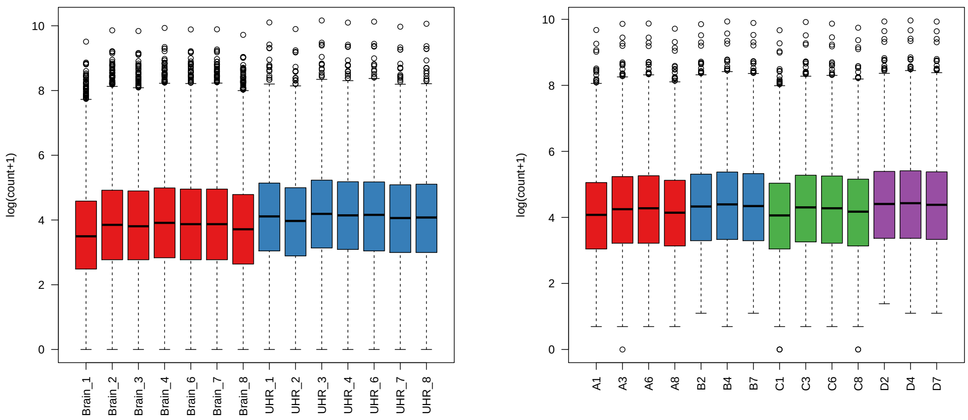
(a) MAQC-2 (b) MAQC-3

Figure 3.1: *MAQC datasets: Number of reads per lane.* Barplots of total number of reads per lane, color-coded by sample (MAQC-2, Panel (a)) or by library (MAQC-3, Panel (b)).

of total reads, due to the different sequencing depth. Figure 3.2 shows the distribution of the gene-level counts for each lane in the datasets. The difference in the sequencing depth yields differences in the distribution of gene-level counts, even for lanes assessing the same RNA. We will need to correct these distributional differences with normalization (Chapter 4) or to take them into account in the model (Chapter 6).

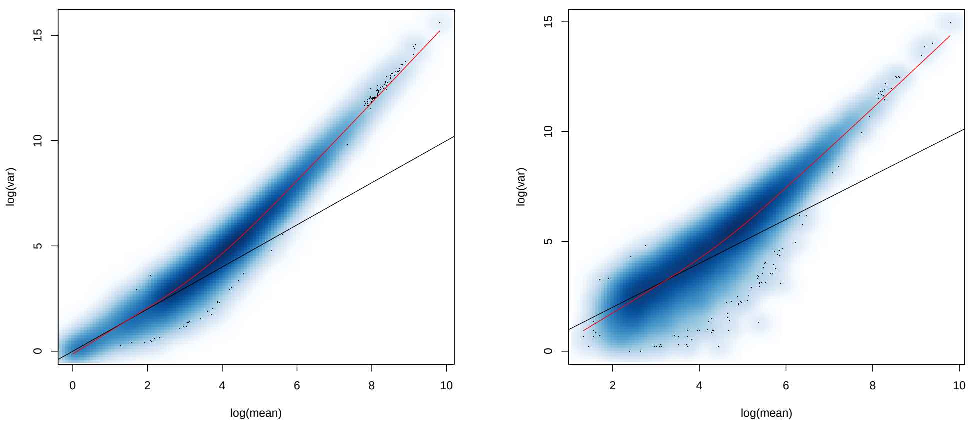
One important feature of count data is the relation between mean and variance. The Poisson model assumes the equi-dispersion of the data, i.e., that the variance is equal to the mean. The presence of over-dispersion is an indication of poor fit of the Poisson model and of the need of more general models, such as the negative binomial. Figure 3.3 shows the mean-variance relation for the MAQC datasets. Even if there are only technical replicates (with no biological variability) there is some overdispersion. We will see in Chapter 4 that the Poisson assumption of equi-dispersion holds after normalization (Figure 4.1, Panel (a)).

In the original MAQC paper (Shi *et al.*, 2006), 997 genes were assayed by qRT-PCR, with four measures (i.e., technical replicates) for each of the Brain and UHR samples. This technology is known to yield accurate estimates of expression levels and it is used here as a gold standard for comparing RNA-Seq methods. Following



(a) MAQC-2 (b) MAQC-3

Figure 3.2: *MAQC datasets: gene-level counts per lane.* Boxplot of gene-level counts per lane, color-coded by sample (MAQC-2, Panel (a)) or by library (MAQC-3, Panel (b)).



(a) MAQC-2 (b) MAQC-3

Figure 3.3: *MAQC datasets: Mean-variance relation.* Scatterplot of log variance vs. log mean for the seven Brain lanes of MAQC-2 (panel (a)) and for the four “library A” lanes of MAQC-3 (panel (b)). Black line: Identity line (Poisson). Red curve: lowess fit.

Bullard *et al.* (2010a), we consider only the genes matching a unique UI gene, are called present in at least three out of the four Brain and UHR runs, and have standard errors across the eight runs not exceeding 0.25. We found 638 genes in common with the RNA-Seq filtered genes and use this subset to compare expression measures between the technologies. The UHR/Brain expression log-fold-change of a gene is estimated by the log-ratio between the average of the four UHR measures and the average of the four Brain measures.

Moreover, as reported in Shi *et al.* (2006), a large number of microarray experiments were conducted on the Brain and UHR samples. As in Bullard *et al.* (2010a), we consider the Affymetrix data from the first site, where each biological sample was assayed using five chips (i.e., technical replicates, GeneChip Human Genome U133 Plus 2.0 Array). We pre-processed the data using RMA (Irizarry *et al.*, 2003), as implemented in the Bioconductor R package *affy*, and obtained *p*-values for UHR vs. Brain differential expression using the *limma* package (Smyth, 2004), with the standard `lmFit` and `eBayes` pipeline. There are 11,081 genes detected by RNA-Seq and present on the Affymetrix chip.

The MAQC data are available in the Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`), under the accession number SRA010153.1.

3.2 Yeast Dataset

Illumina's Genome Analyzer II high-throughput sequencing system was used to sequence RNA from *Saccharomyces cerevisiae* grown in three different media: standard YP Glucose (YPD, a rich medium), Delft Glucose (Del, a minimal medium), and YP Glycerol (Gly, which contains a non-fermentable carbon source in which cells respire rather than ferment). Two different strand specific library preparation protocols were used: "Protocol 1" as described in Maniar and Fire (2011); and "Protocol 2", based on Parkhomchuk *et al.* (2009) as described in Martin *et al.* (2010). See Risso *et al.* (2011) for details on RNA extraction and library preparation.

The experimental design is summarized in Table 3.3. There are four distinct colonies grown in YPD (biological replicates), each yielding its own RNA library, which was then sequenced using two lanes of possibly different flow-cells (technical replicates). The libraries for colonies Y1, Y2, and Y7 were prepared using Protocol 1 and the library for colony Y4 was prepared using Protocol 2. For the Delft medium,

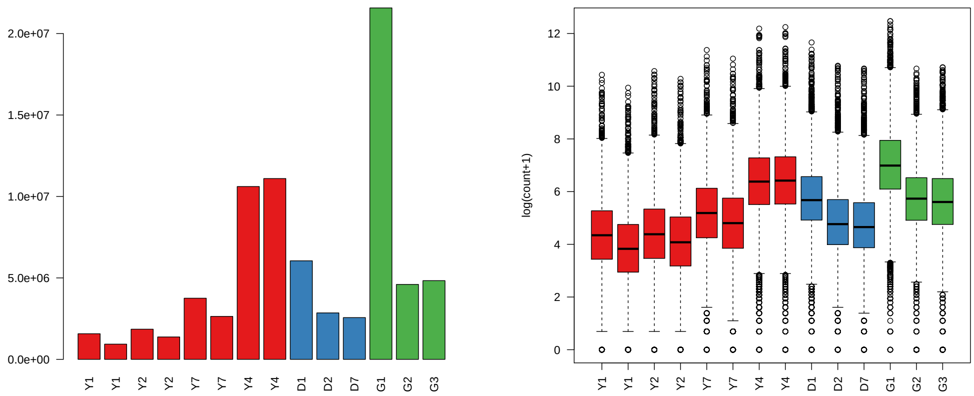
Table 3.3: *Yeast dataset: Experimental design.*

Lane	Colony/ Library	Library prep. protocol	Growth condition	Flow-cell
1	Y1	Protocol 1	YPD	428R1
1	Y1	Protocol 1	YPD	4328B
2	Y2	Protocol 1	YPD	428R1
2	Y2	Protocol 1	YPD	4328B
3	Y7	Protocol 1	YPD	428R1
3	Y7	Protocol 1	YPD	4328B
8	Y4	Protocol 2	YPD	61MKN
7	Y4	Protocol 2	YPD	61MKN
4	D1	Protocol 1	Del	428R1
5	D2	Protocol 1	Del	428R1
6	D7	Protocol 1	Del	428R1
7	G1	Protocol 2	Gly	6247L
1	G2	Protocol 1	Gly	62OAY
2	G3	Protocol 1	Gly	62OAY

there are three biological replicates, each sequenced using Protocol 1 on one lane within the same flow-cell. For the Glycerol medium, there are also three biological replicates; replicate G1 was sequenced in a single lane using Protocol 2, while replicates G2 and G3 were each sequenced using Protocol 1 and one lane of the same flow-cell (distinct from that of G1).

With ten colonies/libraries (i.e., biological replicates), sequenced using two different library preparation protocols and either one or two lanes in a total of five flow-cells, the design allows us to examine both technical effects (e.g., flow-cell, lane) and biological effects (e.g., growth condition, colony). Note, however, that library preparation effects are confounded with colony effects.

Illumina's standard Genome Analyzer pre-processing pipeline was used to yield 36 bp-long single-end reads. Reads were mapped to the reference genome (Saccharomyces Genome Database, r64) using Bowtie (Langmead *et al.*, 2009), considering only unique mapping and allowing up to two mismatches. Low-count genes



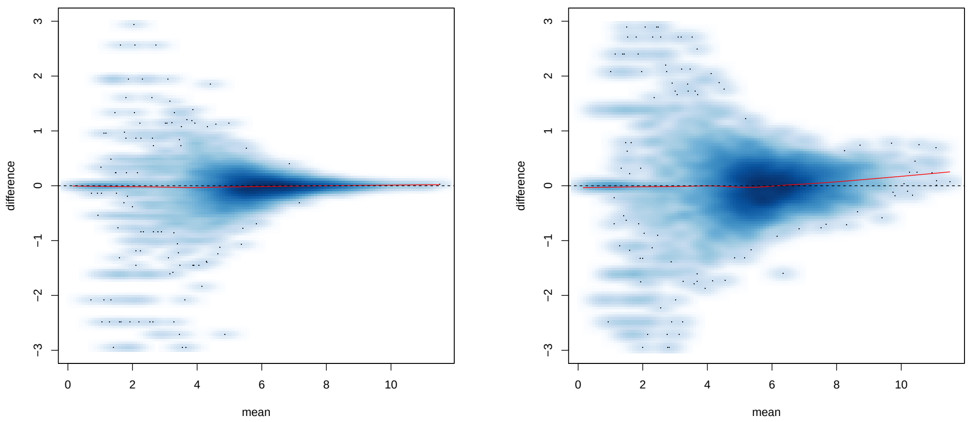
(a) Number of reads

(b) Gene-level counts

Figure 3.4: *Yeast dataset: Number of reads and gene-level counts per lane.* Barplots of total number of reads per lane and boxplot of gene-level counts per lane, color-coded by growth condition.

were filtered out using a procedure analogous to that used for the MAQC datasets. Specifically, genes with an average read count below 10 for each of the three growth conditions were filtered out, yielding 5,690 (out of 6,575) genes.

Figure 3.4 shows the total number of reads per lane and the gene-level counts for the Yeast dataset. Strikingly, the heterogeneity of the counts is much greater than that of the MAQC datasets. Recall that the MAQC samples are commercially available standardized RNAs sequenced in technical replicate lanes of only two flow cells. Things get much more variable in a "real-life" experiment, in which researchers sequence different biological samples with different library preparation protocols and flow-cells. In particular, the biological variability is more than the technical one (Figure 3.5) and the data show a dramatic over-dispersion (Figure 3.6). This highlights the need for new methods to be tested on a biologically meaningful experiment in addition to benchmarking datasets, as they can present features that are not present in controlled and standardized experiments.



(a) Technical Variability (b) Biological Variability

Figure 3.5: *Yeast dataset: Biological vs. Technical variability.* Mean-difference plots of two YPD lanes after scaling by the upper-quartile between-lane normalization of Chapter 4. Panel (a): two lanes assessing colony Y1. Panel (b): first lane of Y1 vs. first lane of Y2.

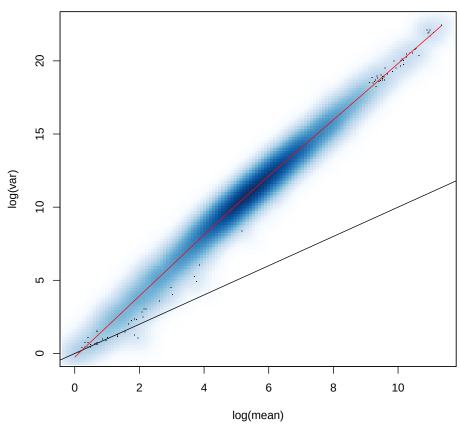
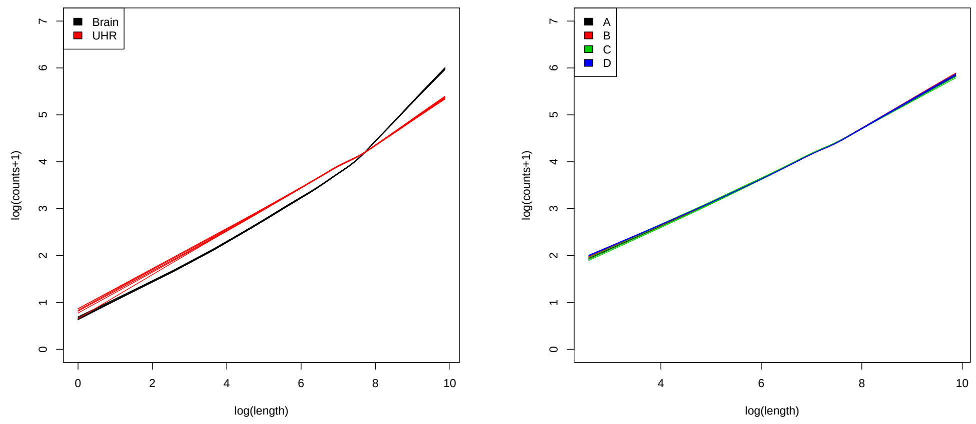


Figure 3.6: *Yeast dataset: Over-dispersion.* Scatterplot of log variance vs. log mean for the eight YPD lanes. Black line: Identity line (Poisson). Red curve: lowess fit.



(a) MAQC-2 (b) MAQC-3

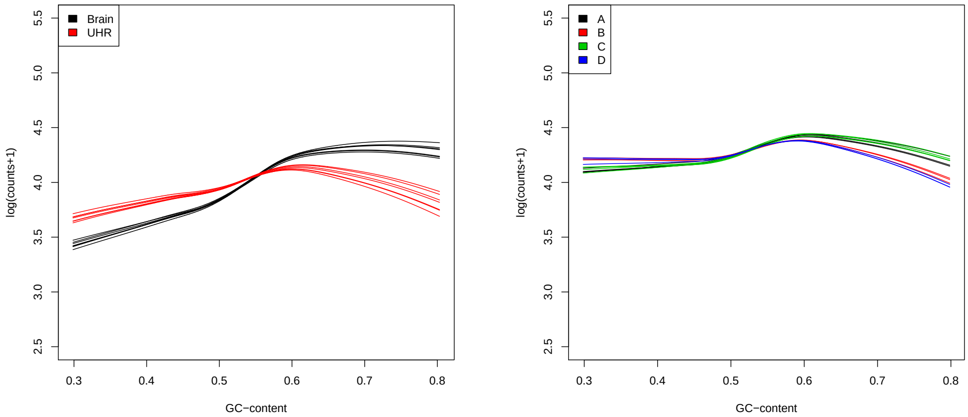
Figure 3.7: *MAQC datasets: Length effect.* Lowess fits of gene-level log(count + 1) vs. log(length), after scaling by the upper quartile (see Chapter 4). Curves are colored according to biological sample for MAQC-2 (Panel (a)) and library preparation for MAQC-3 (Panel (b)).

3.3 Length and GC-content dependence

By definition of gene-level counts (see Section 2.2.2), one expects them to be roughly proportional to both the gene's length and its transcript abundance. This expectation is confirmed in Figures 3.7 and 3.9, Panel (a).

Furthermore, as detailed in the literature review of Chapter 4, previous studies have reported selection biases related to the sequencing efficiency of genomic regions, whereby read counts depend not only on length but also on sequence features such as GC-content and mappability (i.e., uniqueness of a particular sequence compared to the rest of the genome). For instance, GC-rich and GC-poor fragments tend to be under-represented in RNA-Seq, so that, within a lane, read counts are not directly comparable between genes (Figures 3.8 and 3.9, Panel (b)). Unlike length effects, GC-effects tend to be lane-specific (Figure 3.9), so that the read counts for a given gene are not directly comparable between lanes.

Note that for the Yeast dataset the GC-content effect appears much stronger than in the MAQC datasets. This could be due to the different complexity of the two organisms or to the different nature of the experiments (biological vs. technical



(a) MAQC-2

(b) MAQC-3

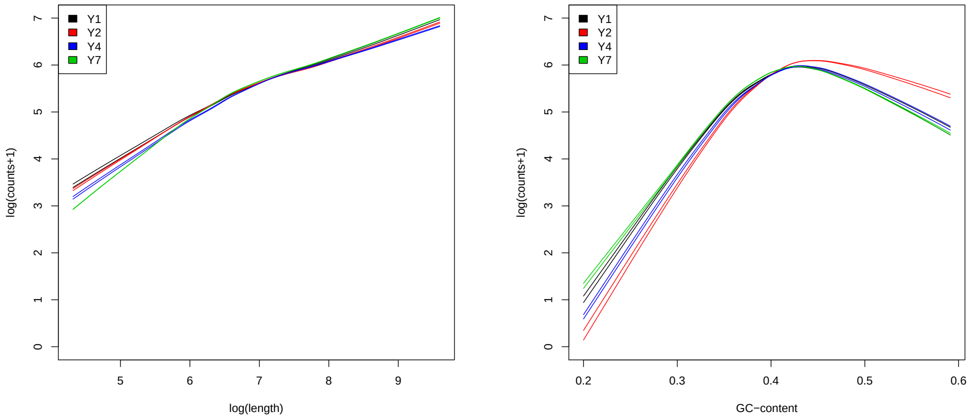
Figure 3.8: *MAQC datasets: GC-content effect.* Lowess fits of gene-level log(count + 1) vs. GC-content, after scaling by the upper quartile (see Chapter 4). Curves are colored according to biological sample for MAQC-2 (Panel (a)) and library preparation for MAQC-3 (Panel (b)).

variability). Previous work has shown strong GC-content effects in complex organisms (Hansen *et al.*, 2011; Pickrell *et al.*, 2010), suggesting that the GC-content dependence is a general concern. Moreover, being a lane-specific effect, even when the dependence is weak (as in the MAQC data), it can strongly bias fold-change estimation (Figure A.2 and A.3 in Appendix A).

3.4 EDASeq: an R package for the exploratory data analysis and normalization of RNA-Seq

Given the high-dimensionality of RNA-Seq data, even EDA is challenging, both from a visualization and a computational point of view. We developed an R package called *EDASeq* (Risso and Dudoit, 2011), publicly available within the Bioconductor project (Gentleman *et al.*, 2004), for the EDA and normalization (see Chapter 4) of RNA-Seq data.

A typical RNA-Seq experiment involves many pre-processing steps before one can obtain a “gene-level” count matrix. The “raw” data consist on several millions



(a) Length effect (b) GC effect

Figure 3.9: *Yeast dataset: Length and GC-content effects.* Lowess fits of gene-level log(count + 1) vs. log(length) (Panel (a)) or GC-content (Panel (b)) for the eight YPD lanes from the Yeast dataset, after scaling by the upper quartile (see Chapter 4). Curves are colored according to biological replicates (i.e., colonies/libraries). The GC-content effect is the same for technical replicate lanes, but can be different for biological replicate lanes.

of short reads (typically 25–100 bases). These reads are then mapped back to a reference genome and the number of reads mapping to a particular gene reflects the abundance of the transcript in the sample of interest.

Before considering the aggregate counts per transcript, it can be of interest to investigate quality metrics of raw and mapped reads, such as percentage of mapped reads, number of unique sequences, nucleotide frequencies. As an illustration of the possible plots that the package can produce, see Figure A.1 in Appendix A.

The standard formats are the FASTQ file format for the raw reads and the SAM/BAM formats for the mapped reads. Given the large size of such files ($\sim$ 1–2 GB) it is unfeasible to use the "in-memory" model of R (i.e., to load in the RAM memory the whole file). A better strategy is to use a "lazy" evaluation, where the data is only loaded when needed.

This is the philosophy of *EDASeq*, which allows the users to perform EDA and quality checks on a standard desktop machine. This is obtained by using the *Fastq-FileList* class of the *ShortRead* package (Morgan *et al.*, 2009) and the *BamFileList* class of the *Rsamtools* package (Morgan and Pagès, 2010).

By leaving the data on disk (in a binary format in the case of BAM) and accessing only the portion of data needed for each step of the analysis, the package allows the researchers to conduct a complete RNA-Seq analysis within the R framework. All the graphic and numerical summaries in this Chapter, as well as the normalizations of Chapter 4, are implemented in the package.

Chapter 4

Normalization

In this Chapter, we focus on biases related to GC-content and other distributional differences that can affect expression level in RNA-Seq and we propose some simple normalization strategies to deal with them. The raw gene-level counts described in Chapter 3 are neither directly comparable between genes within a lane, nor between replicate lanes (i.e., lanes assaying the same library) for a given gene, and normalization of the counts is needed to allow accurate inference of differences in transcription levels.

In Section 4.1 we present an overview of the literature on RNA-Seq normalization. In Section 4.2 we present three simple strategies for GC-content normalization of RNA-Seq. In Section 4.3 we introduce the criteria used to evaluate our proposed methods and in Section 4.4 we applied the normalization methods to the data introduced in Chapter 3. Finally, we conclude the Chapter with a discussion in Section 4.5.

4.1 Literature Overview

Biases related to length and GC-content confound differential expression results as well as downstream analyses, such as those involving Gene Ontology (GO).

As GC-content varies throughout the genome and is often associated with functionality, it may be difficult to infer true expression levels from biased measured read counts. Proper normalization of read counts is therefore crucial to allow accurate inference of differences in expression levels.

Herein, we distinguish between two main types of effects on read counts: (1) within-lane gene-specific (and possibly lane-specific) effects, e.g., related to gene length or GC-content, and (2) effects related to between-lane distributional differences, e.g., sequencing depth. Accordingly, *within-lane* and *between-lane normalization* adjust for the first and second types of effects, respectively.

4.1.1 Within-lane normalization

The most obvious and well-known selection bias in RNA-Seq is due to *gene length*. Bullard *et al.* (2010a) and Oshlack and Wakefield (2009) show that scaling counts by gene length is not sufficient for removing this bias and that the power of common tests of differential expression is positively correlated with both gene length and expression level. Indeed, the longer the gene, the higher the read count for a given expression level (Figure 3.7 and 3.9, Panel (a)); thus, any method for which precision is related to read count will tend to report more significant DE statistics for longer genes, even when considering per-base read counts. Hansen *et al.* (2011) incorporate length effects in a Poisson model for read counts using natural cubic splines and adjust for this effect using robust quantile regression. Young *et al.* (2010) propose a method to account for gene length bias in Gene Ontology analysis.

Another documented source of bias for the Illumina sequencing technology is *GC-content*, i.e., the proportion of G and C nucleotides in a region of interest. Several authors have reported a strong GC-content bias in DNA-Seq (Bentley *et al.*, 2008; Dohm *et al.*, 2008) and ChIP-Seq (Teytelman *et al.*, 2009). Yoon *et al.* (2009) propose a GC-content normalization method for DNA copy number studies, which involves binning the reads in 100-bp windows and scaling the bin-level read counts by the ratio between the overall median and the median for bins with the same GC-content. More recently, Boeva *et al.* (2011) propose a polynomial regression approach, based on binning the reads in non-overlapping windows and regressing bin-level counts on GC-content (with default polynomial degree of three). Still in the context of DNA-Seq, Benjamini and Speed (2011) report that read counts are most affected by the GC-content of the actual DNA fragments from the sequence library (vs. that of the sequence reads themselves) and that the effect of GC-content is sample-specific and unimodal, i.e., both GC-rich and AT-rich fragments are under-represented. They develop a method for estimating and correcting for GC-content bias that works at the base-pair level and accommodates library, strand, and fragment length information, as well as varying bin sizes throughout the genome.

Sequence composition biases have also been observed in RNA-Seq. Hansen *et al.* (2010) report large and reproducible base-specific read biases associated with random hexamer priming in Illumina's standard library preparation protocol. The bias takes the form of patterns in the nucleotide frequencies of the first dozen or so bases of a read. They provide a re-weighting scheme, where each read is assigned a weight based on its nucleotide composition, to mitigate the impact of the bias and improve the uniformity of reads along expressed transcripts.

Roberts *et al.* (2011) also consider the problem of non-uniform cDNA fragment distribution in RNA-Seq and use a likelihood-based approach for correcting for this fragment bias.

When analyzing RNA-Seq data from a yeast diploid hybrid for allele-specific expression (ASE), Bullard *et al.* (2010b) note that read counts from an orthologous pair of genes might overestimate the expression level of the more GC-rich ortholog. To correct for this confounding effect, they develop a resampling-based method where the significance of differences in read counts is assessed by reference to a null distribution that accounts for between-species differences in nucleotide composition.

While there has been general agreement about the need to adjust for GC-content effects when comparing read counts *between genomic regions for a given sample* (as in DNA-Seq and ChIP-Seq) or between orthologs (as in ASE with RNA-Seq), the need to do so was not immediately recognized for standard RNA-Seq DE studies, where one compares read counts *between samples for a given gene.* The common belief was that, for a given gene, the GC-content effect was the same across samples and hence would cancel out when considering DE statistics such as count ratios. Pickrell *et al.* (2010) were the first to note the sample-specificity of the GC-content effect in the context of RNA-Seq and the resulting confounding of expression fold-change estimates. To address this problem, they developed a lane-specific correction procedure which involves binning exons according to GC-content, defining for each GC-bin and each lane a relative read enrichment factor as the proportion of reads in that bin originating from that lane divided by the overall proportion of reads in that lane, and scaling exon-level counts by the spline-smoothed enrichment factors. As noted by Hansen *et al.* (2011), this approach suffers from two main drawbacks. Firstly, as the enrichment factors are computed for each lane relative to all others, the procedure equalizes the GC-content effect across lanes instead of removing it. Secondly, by adding counts across exons and lanes, the method does not account for the fact that regions with higher counts also tend to have higher variances.

Zheng *et al.* (2011) note that base-level read counts from RNA-Seq may not be randomly distributed along the transcriptome and can be affected by local nucleotide composition. They propose an approach based on generalized additive models to simultaneously correct for different sources of bias, such as gene length, GC-content, and dinucleotide frequencies.

In their recent manuscript, Hansen *et al.* (2011) show that GC-content has a strong impact on fold-change estimation and that failure to adjust for this effect can mislead differential expression analysis. They develop a conditional quantile normalization (CQN) procedure, which combines both within and between-lane normalization and is based on a Poisson model for read counts. Lane-specific systematic biases, such as GC-content and length effects, are incorporated as smooth functions using natural cubic splines and estimated using robust quantile regression. In order to account for distributional differences between lanes, a full-quantile normalization procedure is adopted, in the spirit of that considered in Bullard *et al.* (2010a). The main advantage of this approach is that it is lane-specific, i.e., it works independently in each lane, aiming at removing the bias rather than equalizing it across lanes. Modeling simultaneously GC-content and length (and in principle other sources of bias) leads to a flexible normalization model. On the other hand, for some datasets such as the Yeast dataset of Chapter 3, a regression approach may be too weak to completely remove the GC-content effect and other more aggressive normalization strategies may be needed.

4.1.2 Between-lane normalization

The simplest between-lane normalization procedure adjusts for lane sequencing depth by dividing gene-level read counts by the total number of reads per lane (as in multiplicative Poisson model of Marioni *et al.* (2008) and Reads Per Kilobase of exon model per Million mapped reads (RPKM) of Mortazavi *et al.* (2008)). However, this still widely-used approach has proven ineffective and more beneficial procedures have been proposed (Anders and Huber, 2010; Bullard *et al.*, 2010a; Hansen *et al.*, 2011; Robinson and Oshlack, 2010).

In particular, Bullard *et al.* (2010a) consider three main types of between-lane normalization procedures: (1) *global-scaling* procedures, where counts are scaled by a single factor per lane (e.g., total count as in RPKM, count for housekeeping gene, or single quantile of count distribution); (2) *full-quantile* (FQ) normalization procedures, where all quantiles of the count distributions are matched between lanes; and (3) procedures based on *generalized linear models* (GLM). They demonstrate the large impact of normalization on differential expression results; in some contexts, sensitivity varies more between normalization procedures than between DE methods. Standard total-count normalization (RPKM) tends to be heavily affected by a relatively small proportion of highly-expressed genes and can lead to biased DE results, while the upper-quartile (UQ) or full-quantile normalization procedures proposed in Bullard *et al.* (2010a) tend to be more robust and improve sensitivity without loss of specificity.

4.2 Proposed Normalization Strategies

In this Chapter, we propose three different strategies to normalize RNA-Seq data for GC-content following a *within-lane* (i.e., sample-specific) gene-level approach. We examine their performance on both the Yeast dataset and the MAQC datasets (see Chapter 3). For the latter dataset, the gene expression measures from qRT-PCR and Affymetrix chips serve as useful standards for performance assessment of RNA-Seq. We compare our approaches to the state-of-the-art CQN procedure of Hansen *et al.* (2011) (which was shown to outperform competing methods such as that of Pickrell *et al.*, 2010), in terms of bias and mean squared error for expression fold-change estimation and in terms of Type I error and p-value distributions for tests of differential expression. We demonstrate how properly correcting for GC-content bias, as well as for between-lane differences in count distributions, leads to more accurate estimation of gene expression levels and fold-changes, making statistical inference of differential expression less prone to false discoveries.

4.2.1 Within-lane GC-content normalization

We propose three novel within-lane normalization approaches to account for the dependence of read counts on GC-content. The first method is based on the simple idea of regressing gene-level counts on GC-content and is implemented using the loess robust local *regression* procedure; the *global-scaling* and *full-quantile* normalization methods involve stratifying genes in equally-sized *bins* based on GC-content and then "matching" parameters of the count distributions across bins.

For implementation purposes, we consider the logarithms of the gene-level counts as the expression measures. This is convenient for at least two reasons. Firstly, the logarithm is the canonical link for the Poisson (and negative binomial) distribution, hence it seems natural to work on the log-scale when considering regression for count data. Moreover, regression on the log-scale is more robust to the presence of outliers (i.e., extremely high counts) that can bias the fit. Note, however, that quantiles are invariant to monotone transformations, hence the log transformation is not beneficial for global-scaling with quantiles, but it allows us to work with a simpler additive model rather than a multiplicative model.

In what follows, let y_j and gc_j denote, respectively, the logarithm of the read count and the GC-content (i.e., proportion of G and C nucleotides in the gene sequence) for gene $j = 1, \ldots, p$.

Regression normalization

Gene-level read counts (log-scale) y_j are regressed on GC-content gc_j using the loess robust local regression method (Cleveland and Devlin, 1988) and normalized expression measures y'_j are obtained by shifting the residuals to recover the scale of the raw counts, i.e.,

$$y'_j = y_j - \hat{y}_j + T(y_1, \ldots, y_p), \tag{4.1}$$

where $\hat{y}_j$ denote the fitted values and T a summary statistic such as the median.

Global-scaling normalization

Genes are stratified into K equally-sized bins based on GC-content. The normalized expression measures are defined as

$$y'_j = y_j - T(y_{j'} : j' \in k(j)) + T(y_1, \ldots, y_p), \tag{4.2}$$

where $k(j)$ denotes the GC-content stratum to which gene j belongs and T denotes a summary statistic, e.g., median, upper-quartile, or count for control genes. For instance, on the original (unlogged) scale, the normalized count for a particular gene could be its raw count divided by the ratio of the median count in its GC-bin to the overall median count of all genes.

Full-quantile normalization

In full-quantile (FQ) normalization, genes are stratified according to GC-content as for global-scaling normalization. The quantiles of the read count distributions are then matched between GC-bins, by sorting counts within bins and then taking the median of quantiles across bins. This approach is analogous to the microarray between-chip normalization of Irizarry *et al.* (2003) and the RNA-Seq between-lane normalization of Bullard *et al.* (2010a).

4.2.2 Between-lane normalization

GC-content normalization is designed to reduce the dependence of gene-level read counts on sequence composition *within* a lane. However, other technical effects, such as between-lane differences in sequencing depth, can strongly bias differential expression results. For our datasets both a global scaling approach (using for instance the median or the upper quartile) and a full-quantile approach lead to similar results (data not shown). However, since the CQN approach of Hansen *et al.* (2011) involves full-quantile *between-lane* normalization, we settle on this normalization for comparison purposes. We therefore apply the full-quantile *between-lane* normalization procedure of Bullard *et al.* (2010a) after *within-lane* normalization and before differential expression analysis. Such a between-sample normalization approach was used for microarrays in Irizarry *et al.* (2003).

4.3 Evaluation Criteria

Our aim is to evaluate GC-content normalization approaches in terms of their impact on differential expression results. To achieve this, we consider bias and mean squared error in expression fold-change estimation. We also compare normalization methods in terms of their Type I error rates and p-value distributions for likelihood ratio tests of DE based on a negative binomial model for gene-level read counts (Robinson *et al.*, 2010).

The global-scaling and full-quantile within-lane normalization approaches were implemented using $K = 10$ GC-content bins for the MAQC datasets and $K = 50$ bins for the Yeast dataset.

4.3.1 Differential expression analysis

Differential expression analysis is performed using likelihood ratio tests (LRT) based on a negative binomial model for gene-level read counts (Anders and Huber, 2010; Robinson *et al.*, 2010). The negative binomial distribution can be viewed as an extension of the Poisson distribution, which accommodates *over-dispersion* by modeling the variance as a quadratic function of the mean μ, $V(\mu) = \mu + \phi\mu^2$, with dispersion parameter ϕ. For $\phi = 0$, one recovers the Poisson distribution.

We use the Bioconductor R package *edgeR* (Robinson *et al.*, 2010) to fit a negative binomial model to gene-level read counts and perform likelihood ratio tests of DE. A common dispersion parameter is estimated for all genes.

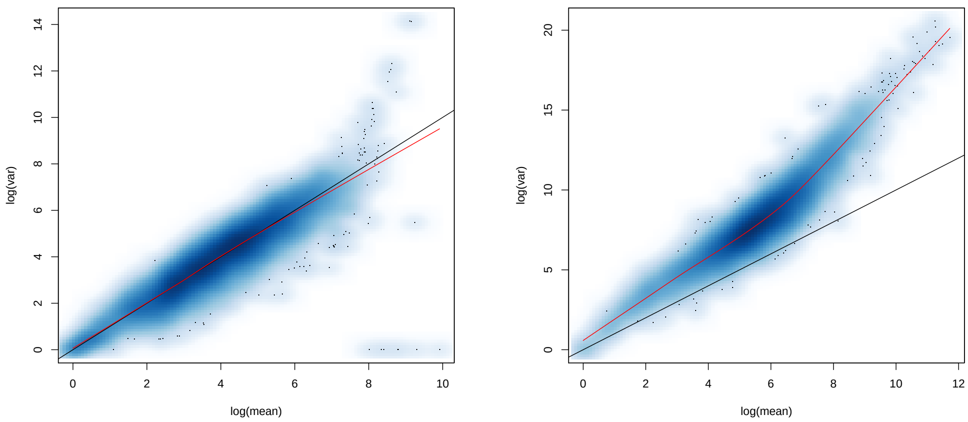
While Bullard *et al.* (2010a) found that the Poisson distribution was appropriate for the MAQC datasets (indeed, the *edgeR* estimates of the dispersion parameters are near zero after between-lane normalization, Figure 4.1, Panel (a)), we have noticed substantial over-dispersion for the Yeast dataset ($\sqrt{\hat{\phi}} = 0.28$ after between-lane normalization, Figure 4.1, Panel (b)).

4.3.2 Bias and mean squared error for expression fold-change estimation

Expression log-fold-changes are estimated by log-ratios of average normalized read counts between two sets of lanes corresponding to the two conditions of interest.

In order to compute bias and mean squared error (MSE), one needs to know the true value of the expression fold-change. For the MAQC-2 dataset, one can use the estimate of the UHR/Brain fold-change from qRT-PCR as the true value, since qRT-PCR is often considered as a gold standard for producing accurate estimates of expression levels. The RNA-Seq estimated fold-change is the ratio of the sum of normalized counts for the seven UHR lanes to the sum of the normalized counts for the seven Brain lanes. For a given gene, bias is then estimated as the difference between the estimated log-fold-changes from the two technologies.

For the Yeast dataset, we consider only the eight YPD lanes (Table 3.3), for which we do not expect any differential expression, and assume that the true log-fold-change



(a) MAQC-2 (b) Yeast

Figure 4.1: *Over-dispersion after between-lane normalization* Scatterplot of log variance vs. log mean for the seven Brain lanes of the MAQC-2 dataset (panel (a)) and for the eight YPD lanes of the Yeast dataset (panel (b)). Black line: Identity line (Poisson). Red curve: lowess fit.

when comparing any combination of such lanes is zero. Specifically, we consider all $\binom{8}{4}/2 = 35$ possible combinations of the eight YPD lanes into two groups of four lanes each. For each such pseudo-dataset, we compute the log-ratio of average normalized read counts between the two groups of four lanes. For a given gene, bias is estimated as the average of these 35 log-ratios and MSE as the average of the square of these 35 log-ratios.

4.3.3 Testing DE based on negative binomial model

To evaluate the impact of normalization on differential expression results, we use the *edgeR* package (Robinson *et al.*, 2010) to perform gene-level likelihood ratio tests of DE, based on a negative binomial model for read counts, with common dispersion parameter.

For the Yeast dataset, we assess Type I error by considering again all 35 YPD pseudo-datasets and by testing for DE between two such groups. Such a situation mimics the null hypothesis of no DE and any gene called DE yields a false positive. For a given pseudo-dataset and nominal Type I error rate α, the actual Type I error rate is defined as the proportion of genes with unadjusted p-values not exceeding α.

It is not possible to assess power with the Yeast dataset, as one cannot identify with certainty the set of all genes that are expected to be DE between growth conditions. Nonetheless, we perform gene-level tests of DE between the three growth conditions using *edgeR* and compare p-value distributions and numbers of genes declared DE between different normalization procedures.

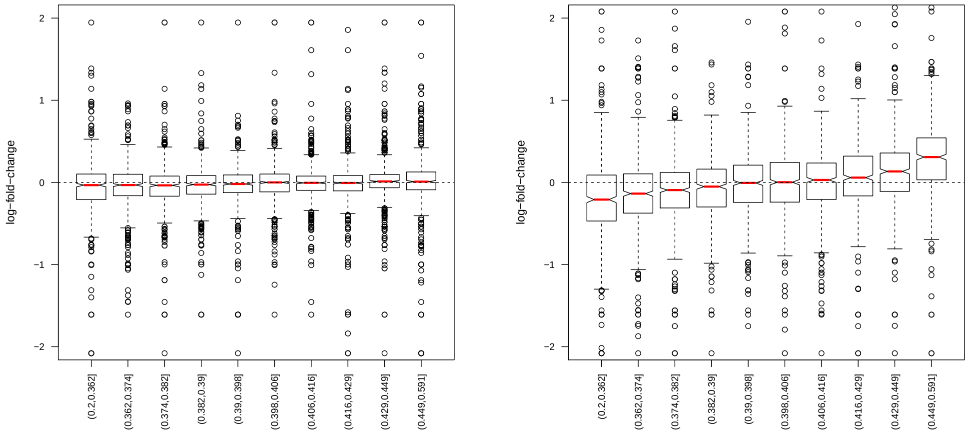
For the MAQC samples, we compare UHR vs. Brain differential expression results based on Illumina RNA-Seq and Affymetrix chip data. For RNA-Seq, we perform tests of DE between the seven Brain and seven UHR lanes using *edgeR*. Tests of DE between the five Brain and five UHR chips are performed using *limma*. We then examine p-value distributions and numbers of genes declared DE for the two technologies.

4.4 Results

4.4.1 GC-content effect

As noted in Pickrell *et al.* (2010), the GC-content bias on read counts is *sample-specific*, meaning that the dependence of gene counts on GC-content may vary between lanes. For the Yeast dataset, Figure 3.9 shows that the relationship between read count and GC-content (after between-lane normalization) is the same for technical replicate lanes, but can be different for biological replicate lanes. This suggests that the GC-content bias is likely to be introduced at the library preparation step (although one should recall that for the Yeast dataset, colony and library preparation effects are confounded; see Table 3.3). The GC-effect is also unimodal, in the sense that read counts first increase, then decrease with GC-content. Figure 3.8 for the MAQC datasets further supports these hypotheses, even though the dependence on GC-content is weaker than for the Yeast dataset. Our findings are in agreement with Benjamini and Speed (2011), who point to PCR as the most important cause for GC-content bias.

The strong impact of GC-content on expression fold-change estimation is illustrated in Figure 4.2, Panel (b), where fold-change increases monotonically with GC-content for YPD biological replicates which are not expected to exhibit any DE. All four normalization procedures considered here reduce the dependence of both



(a) Technical replicates (b) Biological replicates

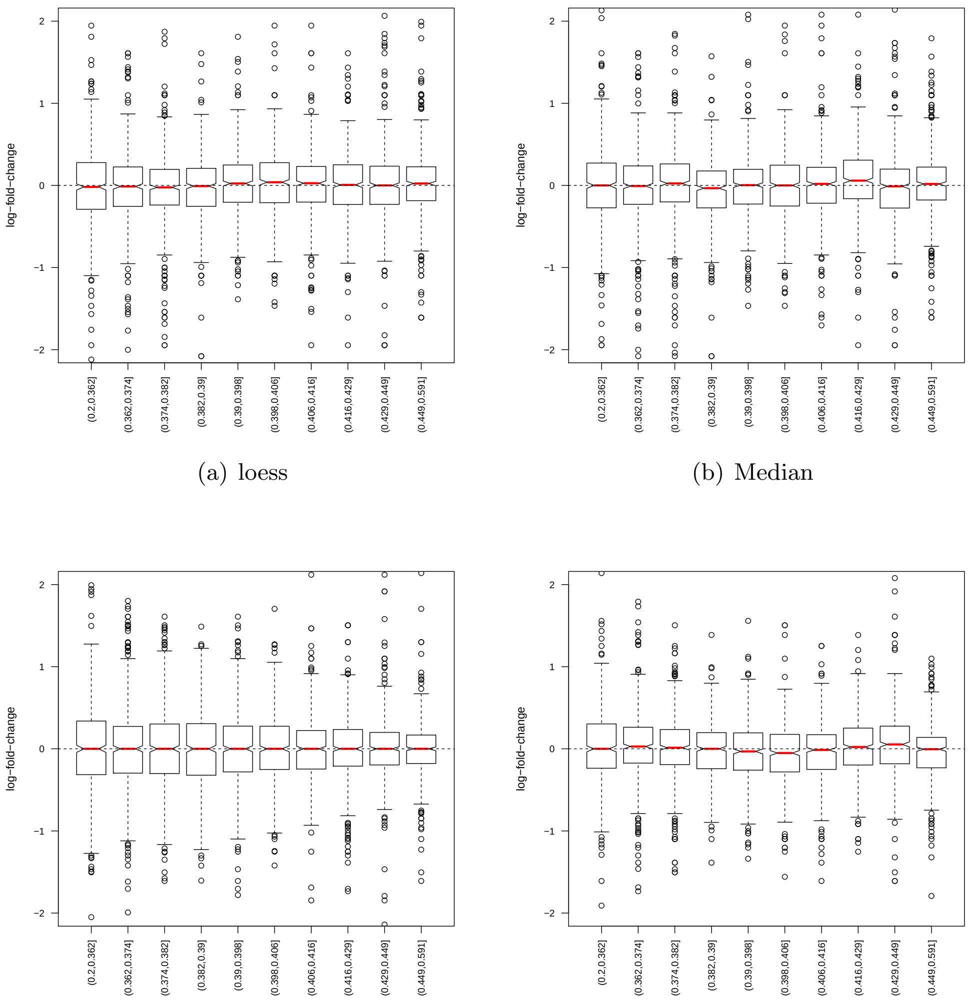
Figure 4.2: *Yeast dataset: Log-fold-change vs. GC-content.* Stratified boxplots of count log-ratios vs. GC-content, after FQ between-lane normalization. Panel (a): Two technical replicate YPD lanes (both lanes for colony Y1). Panel (b): Two biological replicate YPD lanes (Y1 lane vs. Y2 lane from flow-cell 428R1). The GC-content effect appears to be the same for the two technical replicates, so that fold-change estimates do not vary with GC-content. By contrast, the GC-content effect differs between biological replicates and confounds fold-change estimation.

read counts and fold-change estimates on GC-content, with an edge for our proposed full-quantile normalization (Figures 4.3).

Likewise, Figures A.2 and A.3 compare the impact of GC-content on fold-change estimates for technical replicate lanes to that on UHR/Brain fold-change estimates for MAQC-2 and fold-change estimates between two UHR libraries for MAQC-3, respectively.

4.4.2 Bias and mean squared error for expression fold-change estimation

For the MAQC-2 dataset, qRT-PCR may be viewed as a gold standard, so that bias for different RNA-Seq normalization procedures may be assessed based on differences in log-fold-change estimates from the two technologies (Figures 4.4). Figure 4.4, Panel (a), shows that most of the bias is due to differences in sequencing depths between



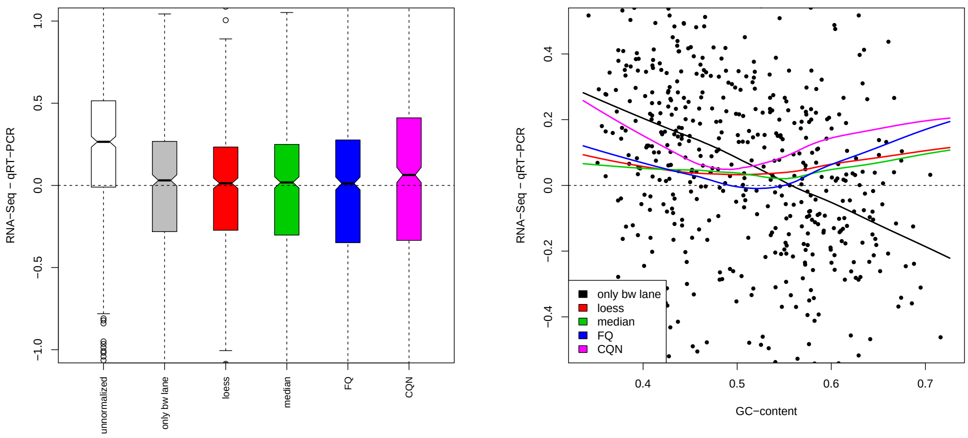
(a) loess

(b) Median

(c) FQ

(d) CQN

Figure 4.3: *Yeast dataset: Log-fold-change vs. GC-content.* Stratified boxplots of count log-ratios vs. GC-content, for the two YPD biological replicate lanes of Figure 4.2, Panel (b), for four within-lane GC-content normalization procedures. Panel (a): Regression normalization using loess. Panel (b): Global-scaling normalization using the median. Panel (c): Full-quantile (FQ) normalization. Panel (d): Conditional quantile normalization (CQN). The first three within-lane procedures were followed by FQ between-lane normalization; CQN includes its own between-lane normalization. All methods seem to effectively reduce the dependence of fold-change on GC-content (compared to Figure 4.2, Panel (b)).



(a) Bias (b) Bias vs. GC-content

Figure 4.4: *MAQC-2 dataset: Bias in fold-change estimation.* Bias in UHR/Brain expression log-fold-change estimation for different RNA-Seq normalization procedures, where bias is defined as the difference between the estimates from RNA-Seq and qRT-PCR for 638 genes assayed by both technologies. Panel (a): Boxplot of bias in log-fold-change estimates. All normalization procedures we propose reduce bias, while CQN tends to overestimate the UHR/Brain fold-change. Panel (b): Dependence of bias on GC-content. The points correspond to bias after only FQ between-lane normalization, the curves are lowess fits of bias vs. GC-content for different within-lane GC-content normalization procedures. There is still substantial dependence of bias on GC-content after CQN.

lanes and that bias is greatly reduced by between-lane normalization. However, the black curve in Panel (b) indicates that with only between-lane normalization, there is a strong dependence of bias on GC-content. All four within-lane GC-content normalization procedures reduce bias and its dependence on GC-content, although CQN still tends to over-estimate fold-changes and is not as effective as the other three approaches in terms of removing the dependence of bias on GC-content.

Similar representations of bias and MSE are provided in Figures A.4 and A.5 for the Yeast YPD pseudo-datasets. All within-lane GC-content normalization methods perform similarly on these artificial data. Note, however, that fold-change estimates can vary greatly between the 35 datasets, perhaps due to colony/library effects (Figure A.6). One would expect the log-fold-changes to be around zero for each dataset.

There is a clear bias for unnormalized counts, with log-fold-change estimates as high as 2. The full-quantile GC-content normalization method seems to be the most coherent in estimating the log-fold-change around zero.

4.4.3 Testing DE based on negative binomial model

Type I error rate

To assess the impact of normalization on DE tests, we first consider Type I error rates. For the Yeast data, the 35 YPD pseudo-datasets simulate an experiment in which all genes satisfy the null hypothesis of equal expression and hence any gene called DE is considered a false positive.

Figure A.7 displays, for each of the 35 pseudo-datasets, the difference between the actual Type I error rate (i.e., the proportion of genes called DE) and the nominal Type I error rate for the negative binomial LRT implemented in the *edgeR* package (Robinson *et al.*, 2010). The figure indicates that Type I error rates vary substantially between pseudo-datasets (perhaps due to colony/library effects, as noted for Figure A.6), although the median actual Type I error rate is close to the nominal value for all within-lane GC-content normalization methods. With only between-lane normalization, DE tests appear anti-conservative, on average.

Figure 4.5 summarizes the Type I error rates for the 35 pseudo-datasets by the area between the curves corresponding to the most conservative and most anti-conservative behavior. Full-quantile GC-content normalization leads to the smallest area, indicating that the DE test is closer to its nominal level than with other procedures.

p-value distribution

To evaluate normalization methods in a biologically meaningful context, we consider the full Yeast dataset (i.e., all fourteen lanes) and perform gene-level LRT of growth condition effects on gene expression using *edgeR*. The stratified boxplots in Figure A.8 reveal a clear dependence of p-values on GC-content for all but the full-quantile GC-content normalization method. Figure 4.6 indicates that the percentage of genes declared differentially expressed increases sharply with GC-content, again for all but full-quantile normalization.

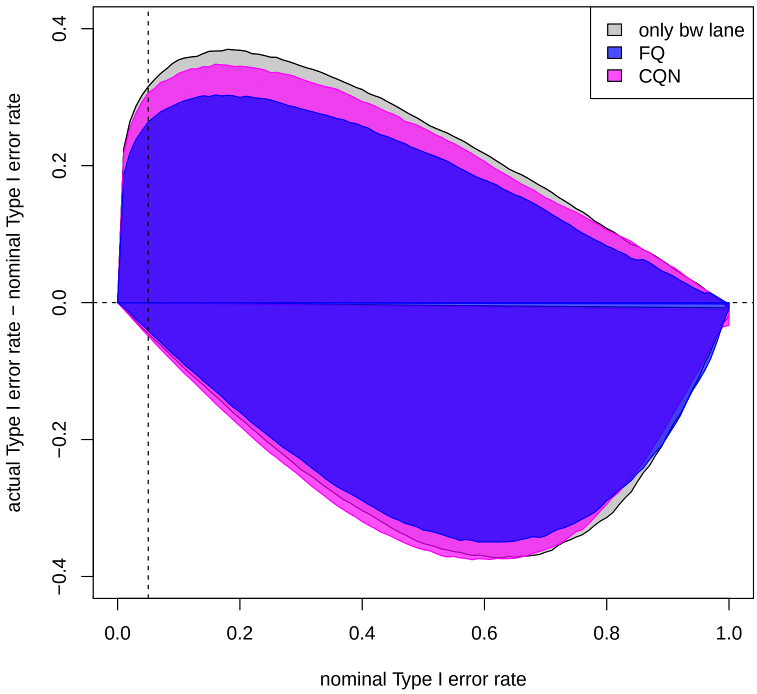


Figure 4.5: *Yeast YPD pseudo-datasets: Type I error.* Difference between actual and nominal Type I error rates vs. nominal Type I error rate, for different normalization procedures. The colored areas correspond to the most conservative and most anti-conservative curves obtained from the 35 YPD pseudo-datasets. The dashed line corresponds to a nominal unadjusted p-value of 0.05. The full-quantile GC-content normalization procedure yields the smallest area, meaning that the actual Type I error rate is closer to the nominal Type I error rate than with the other two procedures.

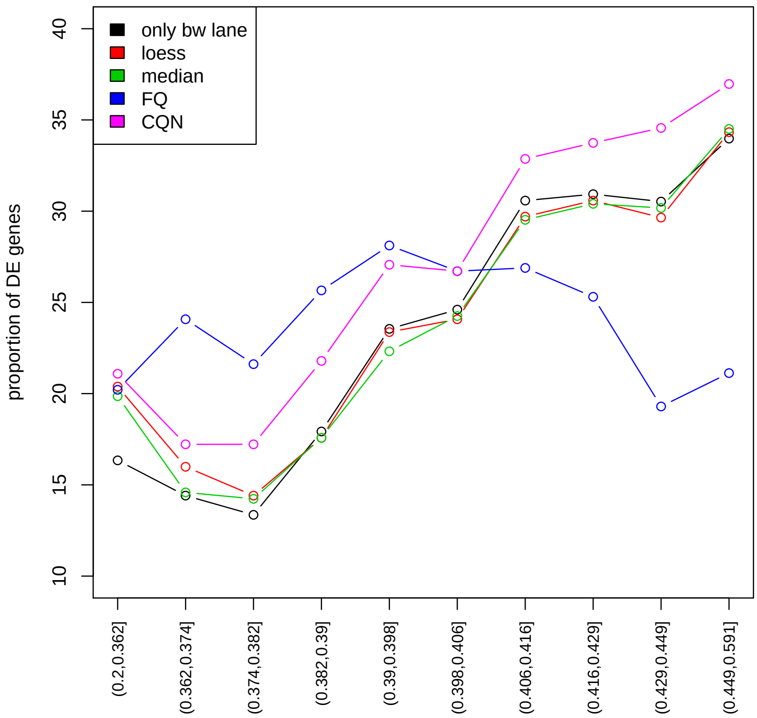


Figure 4.6: *Yeast dataset: Proportion of DE genes vs. GC-content.* Here, a gene is declared DE between the three growth conditions if its nominal unadjusted p-value from the negative binomial LRT is below the threshold of 10^{-5} (corresponding to a nominal Bonferroni family wise error rate of 0.057 and Benjamini and Hochberg (1995) false discovery rate of 4.22×10^{-5}). There is a clear trend towards more detected differential expression at higher GC-content with all normalization procedures but the full-quantile.

For the MAQC-2 dataset, we examine p-value distributions when testing for DE between Brain and UHR using both RNA-Seq and Affymetrix chips. Figure 4.7 shows that the GC-content effect on DE results is technology-specific. Indeed, for microarrays, p-values do not depend on GC-content. By contrast, with only between-lane normalization, RNA-Seq p-values tend to decrease with GC-content. Full-quantile within-lane GC-content normalization removes this dependence. Furthermore, as indicated by the smaller p-values, RNA-Seq appears to yield more power than microarrays in terms of detecting DE.

4.4.4 Tuning parameters

The main tuning parameter in our proposed global-scaling and full-quantile GC-content normalization procedures is the number of GC-content bins. This parameter is analogous to the span in loess robust local regression, thus, the same considerations of bias/variance trade-off should guide its selection. Intuitively, the larger the number of bins, the more adaptive and possibly noisy the normalization. The boxplots of bias and MSE in Figure A.10 indicate that DE results are robust to the number of GC-content bins in FQ normalization. We used $K = 10$ bins for the MAQC datasets and $K = 50$ for the Yeast dataset; a selection which reflects the stronger GC-content effect observed for the latter dataset.

4.5 Discussion

We have compared differential expression results based on our three proposed within-lane GC-content normalization methods and the CQN method of Hansen *et al.* (2011), on the MAQC and Yeast datasets. Only full-quantile GC-content normalization appears to effectively remove the dependence of the proportion of DE genes on GC-content. This could mean either that, for some biological reason, GC-richer genes are more likely to be truly DE (in which case normalization erroneously removes this dependence) or that GC-content bias is so strong that an aggressive normalization method is needed. Since we are not aware of any plausible biological explanation for the dependence of DE results on GC-content, we believe that the MAQC and Yeast data require a full-quantile approach and that merely regressing counts on GC-content is not sufficient to completely remove the bias.

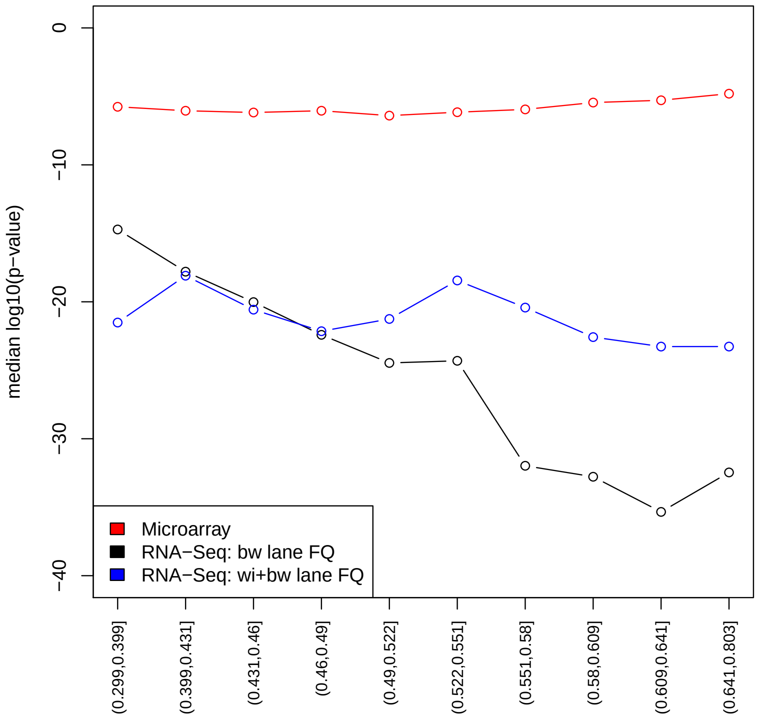


Figure 4.7: *MAQC-2 dataset: p-values vs. GC-content.* Median unadjusted p-value ($\log_{10}$) for each GC-content stratum, for microarray and RNA-Seq UHR vs. Brain DE analysis (11,081 genes detected by RNA-Seq and present on the Affymetrix chip). The figure shows that the GC-content bias is technology-related and that full-quantile within-lane normalization reduces the dependence of RNA-Seq p-values on GC-content. Furthermore, the figure suggests that RNA-Seq has more power for detecting DE genes (smaller p-values) compared to microarrays.

To rule out a biological reason for the dependence of DE on GC-content, we compared UHR vs. Brain DE results based on the MAQC-2 RNA-Seq data to those based on Affymetrix chip data (Shi *et al.*, 2006). Figure 4.7 clearly indicates that the dependence of p-values on GC-content is technology-specific, i.e., unlike RNA-Seq p-values, microarray p-values do not depend on GC-content. Full-quantile within-lane normalization reduces the dependence of p-values on GC-content. Interestingly, and encouragingly, the smaller p-values for the RNA-Seq data suggest that this newer assay is more powerful than microarrays for DE analysis (although it is unclear how to relate numbers of lanes and numbers of chips in terms of sample size).

We identified differentially expressed genes using a likelihood ratio test based on a negative binomial model for read counts. For the MAQC datasets, Bullard *et al.* (2010a) found that it was appropriate to model read counts using the Poisson distribution (negative binomial distribution with null dispersion parameter). For the Yeast dataset, goodness-of-fit analysis suggests that a negative binomial model with common dispersion parameter for the ensemble of genes is sufficient to capture the over-dispersion present in the counts (data not shown). Other possibilities include the estimation of a smooth mean-variance relationship (Anders and Huber, 2010; Robinson *et al.*, 2010). A detailed evaluation of read count models and DE methods is out-of-scope here, since our aim is to compare normalization approaches for a given DE method. It would nonetheless be of interest to determine the extent to which CQN (Hansen *et al.*, 2011) is affected by the over-dispersion of the counts for the Yeast dataset, a violation of its assumption of Poisson-distributed read counts.

Another well-known selection bias in RNA-Seq is due to gene length (Bullard *et al.*, 2010a; Oshlack and Wakefield, 2009). For the Yeast dataset, we noticed only a minor length effect on read counts and DE results (Figure A.9, Panel (a)). In fact, genes with high GC-content tend to be shorter, so there seems to be a compensation effect due to sequence composition (data not shown). Mappability does not appear to affect read counts for this dataset (Figure A.9, Panel (b)). Gene length bias for the MAQC datasets is discussed in Bullard *et al.* (2010a).

In addition to their good performance noted above, our proposed normalization methods offer a number of advantages. They are very simple to implement and lead to DE results that are robust to tuning parameters such as the number of GC-content bins (Figure A.10). They could be applied to other genomic regions (e.g., exons), either "from scratch" or by retaining the scaling from a previous gene-level normalization. They can easily be adapted to incorporate other sequence features such as gene length and mappability. Note, however, that in the process of adjusting for GC-content one may already be adjusting indirectly for other covariates such as length. Controls (e.g., housekeeping genes, spiked-in sequences) could also be included.

Some authors have argued that it is better to leave the count data unchanged to preserve their sampling properties and instead use an offset for normalization purposes in the statistical model for read counts (Anders and Huber, 2010; Robinson and Oshlack, 2010). Our normalization approaches can easily be modified to produce an offset, by considering the difference between normalized and unnormalized counts, in a way similar to Hansen *et al.* (2011).

An alternative approach to normalization is to correct for the GC-content bias after performing DE tests, e.g., in a fashion similar to that proposed in Young *et al.* (2010) for gene length bias in context of Gene Ontology analysis. We find it preferable to adjust for GC-content prior to statistical modeling and DE analysis, given the sample-specificity and possible complexity of the GC-content bias.

As with gene length, GC-content appears to act as a proxy for sample size, in the sense that, as GC-content increases, read counts first increase then decrease, over-dispersion first decreases then increases, residuals first decrease then increase, and evidence in favor of DE increases. One could therefore argue that it is not justified after all to normalize for GC-content. Other approaches which account for GC-content could be based on standardized p-values, i.e., p-values that explicitly account for sample size (Good, 1992). A rule-of-thumb for standardizing a p-value p_n based on a sample size of n to sample size 100 is $\tilde{p}_n = \min\left(\frac{1}{2}, p_n\frac{\sqrt{n}}{10}\right)$. In lieu of the sample size n, one could use gene length or GC-content. However, such an approach does not address the sample-specificity of the GC-content effect and the DE results would still be biased.

GC-content bias is even more of an issue when comparing read counts between species, e.g., allele-specific expression in diploid hybrid of *S. bayanus* and *S. cerevisiae* (Bullard *et al.*, 2010b). We are considering extensions of our methods to address GC-content bias for between-species, within-gene DE analyses.

It would also be interesting to consider adaptations of our methods to other sequencing assays, such as ChIP-Seq and DNA-Seq.

Finally, as with microarrays, positive and negative controls (e.g., housekeeping genes, spiked-in sequences) are essential for conclusive validation and comparison of any inference method, e.g., in terms of bias, variance, Type I error, and power. Controls could also be incorporated as "anchors" within the normalization procedure itself (Yang *et al.*, 2002). The use of controls from the External RNA Control Consortium (ERCC) in the recent article from Jiang *et al.* (2011) is an encouraging step in this direction.

Chapter 5

Cluster Analysis

When dealing with high-throughput data, cluster analysis becomes an important tool both to highlight "batch effects" and anomalous samples, and as an unsupervised classifier, especially in those applications in which a know classification is missing (e.g. cancer studies).

In this Chapter, we propose a novel clustering procedure for RNA-Seq data. In Section 5.1 we review microarray literature on cluster analysis. In Section 5.2 we describe the algorithm (Azzalini and Torelli, 2007) at the base of the proposal. In Section 5.3 we present a variance stabilizing transformation for negative binomial data, that we will use as a first step in our clustering procedure, which is presented in details in Section 5.4. Finally, in Section 5.5 we apply the proposed method to the RNA-Seq datasets presented in Chapter 3.

5.1 The clustering of high-dimensional data

A huge effort has been directed to the identification and evaluation of the best methods for the clustering of microarray data. There are two ways of clustering a gene expression matrix (Kerr *et al.*, 2008; Slonim, 2002): (i) gene functions may be inferred from clusters of genes similarly expressed across samples and (ii) samples can form groups which show similar expression across the genes. Moreover, genes and samples can be clustered simultaneously, with their inter-relationship represented by bi-clusters (Cheng and Church, 2000; Madeira and Oliveira, 2004).

The clustering of genes on the basis of the samples is a standard cluster analysis problem that can be tackled by a variety of algorithms (McLachlan *et al.*, 2002). For a comprehensive review see Kerr *et al.* (2008).

A more challenging problem is the clustering of samples on the basis of the genes; according to Dudoit *et al.* (2002), three main types of statistical problems arise in this context: (a) identification of new classes using gene expression profiles (Alizadeh *et al.*, 2000; Golub *et al.*, 1999); (b) classification of samples into known classes (Dudoit *et al.*, 2002; Tibshirani *et al.*, 2002); (c) identification of "marker" genes which characterize the difference among the classes (Guo *et al.*, 2007). In this Chapter we will address both points (a) and (c). For this kind of problems, the standard clustering techniques do not behave due to dimensionality issues (the number of genes being much greater than the number of samples).

In recent years, computational improvement enabled new clustering techniques, and contributed to the development of previously unfeasible methods. In this context, McLachlan *et al.* (2002) propose a mixture model-based approach to the clustering of microarray expression data. Their scheme accounts for gene selection, through mixtures of t distributions, and dimensionality reduction, through a mixture of factor analyzers. More precisely, they select a gene on the basis of a likelihood ratio statistic for testing one versus two components in the mixture model. In the second step of their algorithm, they group the samples by fitting a two-component mixture of factor analyzers.

Although their approach sounds promising, there are three main limitations. Firstly, the parametric assumptions about clusters' distributions can be restrictive (Li *et al.*, 2007); for example, two Gaussian random variables can result in a single mode (one cluster) or even a two component multivariate Gaussian mixture can lead to more than two modes (Li *et al.*, 2007). Moreover, it needs pre-specification of the number of mixture components; this, from an unsupervised perspective, which assumes that the true number of clusters is unknown, represents a serious limitation. Finally, the number of parameters per component grows as the square of the dimension of the data (Fraley and Raftery, 2002), which is a major shortcoming in high-dimensional data. Therefore, a nonparametric approach seems advantageous.

Azzalini and Torelli (2007) proposed a method, to which we will refer as *pdfCluster*, based on a nonparametric estimate of the underlying density function (the method is briefly reviewed in Section 5.2).

Exploiting the advantages of the nonparametric nature of *pdfCluster*, De Bin and Risso (2011) present a novel strategy, which consists in applying a clustering technique after gene filtering and dimensionality reduction, in order to exploit the most significant dimensions in the definition of the clusters. The procedure can be thought of as a three-step algorithm: (i) gene filtering; (ii) dimensionality reduction; (iii) clustering in the reduced space.

Several authors outlined the importance of a gene filtering step prior to inferential procedures (Bourgon *et al.*, 2010) or cluster analysis (Tritchler *et al.*, 2009). Tritchler *et al.* (2009) empirically showed that principal components and cluster analysis are strongly affected by gene selection, and that filtering out uninformative genes can reduce bias in the clustering of samples. Furthermore, Johnstone and Lu (2009) showed, from a theoretical point of view, that some initial reduction in dimensionality is desirable before applying a principal component analysis, when p is larger than n.

Traditional approaches to gene filtering are based on thresholding the mean or the variance of genes across samples. Bourgon *et al.* (2010) found that gene-by-gene filtering by overall variance increased the power of the subsequent t-test. Tritchler *et al.* (2009) considered the covariance structure of the genes, defining filters that preserve the topology of the network.

Nevertheless, from a clustering point of view, these approaches could be unsafe: a gene should be considered relevant if it is important in the definition of the clusters; therefore, it seems more appropriate to retain those genes whose univariate distribution highlights a clear grouping among the observations rather than the ones with higher variance.

Making use of a nonparametric density estimation for the definition of the clusters, the procedure presented in De Bin and Risso (2011) is specifically aimed at the clustering of a set of continuous variables. In this Chapter, we propose an extension of the algorithm to deal with count data, by assuming a negative binomial distribution for the counts and applying a variance stabilizing transformation.

5.2 The *pdfCluster* algorithm: an overview

In the literature, nonparametric cluster analyses based on mode identification have already been presented. See Banerjee *et al.* (2005); Banfield and Raftery (1993); Fraley and Raftery (2002); Li and Zha (2006); Li *et al.* (2007). The *pdfCluster* algorithm (Azzalini and Torelli, 2007) starts from a quite simple idea, introduced in

Hartigan (1975):

> Clusters may be thought of as regions of high density separated from other such regions by regions of low density.

These regions are achieved by "cutting" the density function computed out of observations by a level c, that varies through the algorithm.

More formally, consider a p-dimensional space, $\mathcal{Y} \subseteq \mathbb{R}^p$. Let $y_1, \ldots, y_n$ be a vector of p-dimensional observations, $y_i \in \mathcal{Y}$, for $i = 1, \ldots, n$. Starting from this vector, using a method of nonparametric density estimation, we can obtain $\hat{f}(y), y \in \mathcal{Y}$, i.e. the empirical version of the density $f(y)$.

There is not a specific method for the nonparametric density estimation related to *pdfCluster*, since the only restriction is that $\hat{f}(y_i) < +\infty$ for all $i = 1, \ldots, n$. This restriction is not limiting, because almost all estimation techniques satisfy it. Following Azzalini and Torelli (2007), we choose a kernel method with Gaussian kernel and constant smoothing parameter $h = (h_1, \ldots, h_p)^\top$, with $h_j = \left(\frac{4}{(p+2)n}\right)^{1/(p+4)} s_j$, $j = 1, \ldots, p$, where s_j is the estimated standard deviation of the j-th variable. This choice is related to the minimization of the asymptotic integrated mean squared error (Azzalini and Torelli, 2007). As suggested by the authors, we slightly shrink h toward zero, using a shrinkage factor of $3/4$.

Cutting the computed $\hat{f}(y)$ at a level $c \in [0; \max \hat{f}]$, they obtain m subspaces $\mathcal{M}_k$, $k = 1, \ldots, m$, of the sample space $\mathcal{Y}$. Dropping the observations not belonging to $\cup_{k=1}^m \mathcal{M}_k$, they select only those observations y_i such that $\hat{f}(y_i) > c$. The observations belonging to the same $\mathcal{M}_k$ are connected by the Delaunay triangulation (see, e.g., de Berg *et al.*, 2008) to form the "cluster cores". Finally, the unallocated observations are allocated by a classification method, based on nonparametric density estimation too: if y_0 is the unallocated observation, the estimated density $\hat{f}_k(y_0)$ based on the data already assigned to group k is computed, and y_0 is assigned to the group with highest ratio $\hat{f}_k(y_0) / \max_{l \neq k} \hat{f}_l(y_0)$. Finally, it is important to notice that *pdfCluster* selects by itself the number of clusters.

We will use the *pdfCluster* algorithm as implemented in the *pdfCluster* R package (Azzalini *et al.*, 2011).

5.3 A variance stabilizing transformation for negative binomial data

Due to their discrete nature, the clustering of RNA-Seq data presents an additional challenge. One should define a suitable distance metric for count data, taking in consideration the relation between mean and variance. One simple alternative is to consider a variance stabilizing transformation (VST) and apply a clustering technique designed for continuous variables (Anders and Huber, 2010). Here, we assume a negative binomial model, with a common dispersion parameter across the ensemble of the genes and we derive a VST for RNA-Seq data. Anders and Huber (2010) derived a similar transformation estimating the mean-variance relation regressing the variance over the mean through a local regression approach.

Denote with Y a negative binomial random variable with mean μ and dispersion ϕ. The variance function is $V(\mu) = \mu + \phi\mu^2$. The parametrization $\psi = g(\mu)$, where $g(\cdot)$ is a differentiable function, is called a variance stabilizing parametrization if

$$V(\mu)\, g'(\mu) = c,$$

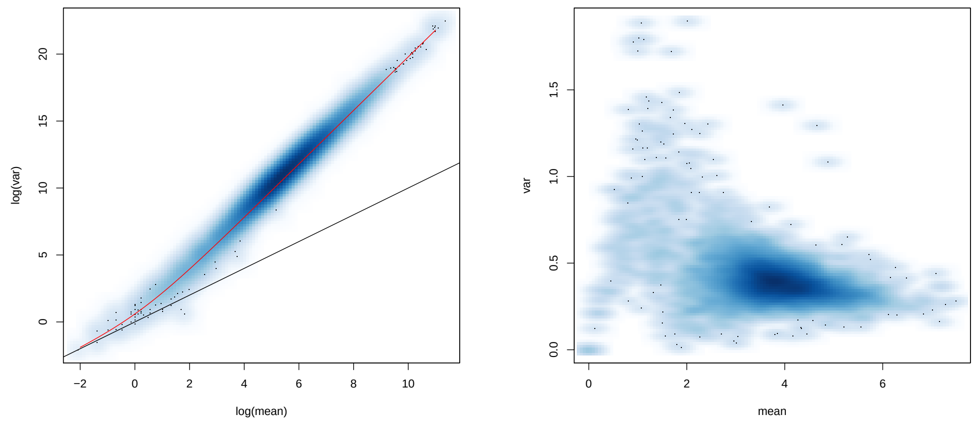
where c is a constant and $g'(\cdot)$ denotes the derivative of $g(\cdot)$ (Anscombe, 1948; Bartlett, 1936). If ϕ is known, it is easy to derive $g(\cdot)$ as

$$\begin{aligned} g(\mu) &= c \int V(t)^{-1} dt \\ &= c \int \frac{1}{\sqrt{t + \phi t^2}}\, dt \\ &= \frac{1}{\sqrt{\phi}} \sinh^{-1}\left(\sqrt{\phi\mu}\right), \end{aligned} \tag{5.1}$$

where $c = \frac{1}{2}$ and $\sinh^{-1}(\cdot)$ denotes the inverse hyperbolic sine.

We estimate a common dispersion parameter ϕ for all the genes, by pooling the expression values of the genes and by maximizing the Cox-Reid adjusted profile likelihood (Cox and Reid, 1987), as implemented in the R package *edgeR* (Robinson *et al.*, 2010). Robinson and Smyth (2008) showed that, for small sample sizes, this is the less biased estimator, when compared to pseudo-likelihood and quasi-likelihood approaches.

By applying the transformation $g(\cdot)$ to the data, we obtain the transformed data on a continuous scale, and with approximately constant variance (Figure 5.1).



(a) Log (b) VST

Figure 5.1: *Yeast dataset: Mean-variance relation before and after VST.* Scatterplot of variance vs. mean for the eight YPD lanes before (log scale, Panel (a)) and after (Panel (b)) variance stabilizing transformation of equation (5.1). In Panel (a): Black line: Identity line (Poisson). Red curve: $\mu + \phi\mu^2$ with $\phi = 0.81$ estimated as common across all the genes. It is striking how a single ϕ value can correctly capture the mean-variance relation.

5.4 A novel algorithm for the clustering of RNA-Seq data

As we said, the clustering of samples using expression data is a challenging statistical problem due to dimensionality issues. Therefore, in this context, it is unfeasible to apply directly a clustering technique to the whole data matrix. Here, we propose a method to tackle this problem exploiting the self-detection of number of clusters feature of *pdfCluster*.

Our general idea can be summarized as follows: (i) apply a variance stabilizing transformation; (ii) consider the clustering of samples using the univariate distribution of each gene and select for the subsequent analyses the p' genes for which the method identifies two or more clusters; (iii) reduce the space dimensionality by selecting the first p'' principal components; (iv) apply a clustering technique in the reduced p''-dimensional space. Clearly, this algorithm is general and, in principle, any clustering technique can be used in steps (ii) and (iv). In this Chapter we use the *pdfCluster* algorithm for both steps.

As for step (ii), i.e., gene filtering, we consider a gene relevant if its values in one category (e.g., Brain samples in the MAQC-2 dataset) are different from the ones in the other category (e.g., UHR) or categories. From another point of view, this means that the observation representing the Brain samples are separated from the UHR ones, or, more simply, the samples are in different clusters. In this way, it seems reasonable to apply a cluster method to each gene, and retain as relevant those genes for which the method identifies distinct clusters. In a nonparametric framework, we can apply *pdfCluster* to each gene, taking advantage of the self-detection number of clusters feature. We select only the genes for which the method detects two or more clusters (see Figure 5.2 for an illustrative example). We consider a clustering technique to define informative genes over the traditional variance-based approaches, because small overall variance does not necessarily imply a single cluster, and high overall variance is not always an indication of two of more clusters. Moreover, variance-based filters depend on the choice of an arbitrary threshold, which can be difficult to choose, since one typically does not know which portion of the genes is responsible for the clustering of the samples.

As for step (iii), i.e., dimensionality reduction, considered if the selected genes are still too many, we propose to keep the first principal components, as in Azzalini

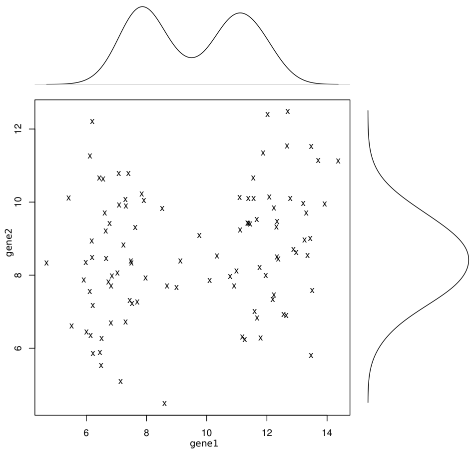


Figure 5.2: *Example of gene selection in the filtering step of our procedure.* Gene1 is crucial in the cluster definition, while gene2 is not. The univariate distributions of the genes reflect this, as one can see from the Gaussian kernel density estimation reported along the axes.

and Torelli (2007). The principal component analysis is a very simple procedure which reduces the dimension of a data set of a large number of interrelated variables, preserving as much as possible of the data set variation. Since it has no requirements about the data distribution, it is consistent with our nonparametric strategy.

5.4.1 Computational issues

The further dimensionality reduction in our step (ii) is necessary since *pdfCluster*, in order to compute the Delaunay triangulation, exploits the *Quickhull* algorithm (Barber *et al.*, 2006). Barber *et al.* (2006) state that the Quickhull algorithm for finding the convex hull of a set of n points in $\mathbb{R}^p$ requires at most $O(n \log n)$ operations if $p \leq 3$, and $O(n^m/m!)$ where $m = \lfloor p/2 \rfloor$ for $p > 3$. Azzalini and Torelli (2007) observed that the computing time increases less than quadratically in n for any fixed p, but it increases more than exponentially in p for fixed n. Our experience is that for $p = 1$, the algorithm is very fast (less than one minute to run 20,000 times with an Intel(R) Core(TM)2 Quad CPU Q9400 @ 2.66 GHz). Therefore, there are no computational problems for the step (i). Moreover, with $p < 10$ it takes a reasonable time (e.g. 11 mins in our machine with $p = 9$) to complete the procedure.

5.4.2 Number of principal components

In order to choose the number of principal components, we carried out a small simulation study (data not shown): we found that, with $n = 100$, *pdfCluster* performs at best, in terms of misclassification error, with 3-4 dimensions, while with $p = 5$ the misclassification error starts to grow. This is probably due to the extreme dispersion of the observations in higher dimensional spaces (the well-known curse of dimensionality). The performances of *pdfCluster* are slightly better with 4 components, but the improvement does not justify the increased computational time (recall that the order of the number of operations needed to compute the Quickhull algorithm massively changes between $p \leq 3$ and $p > 3$). Thus, for the subsequent analyses, we will retain 3 principal components.

5.5 Results

We apply the clustering algorithm to the datasets presented in Chapter 3, properly normalized both for within-lane GC-content effects and for between-lane distributional differences. This allows us to simultaneously evaluate the goodness of the clustering algorithm and the efficacy of the normalization procedure. In fact, if the normalization correctly removes biases without obscuring biologically meaningful differences, the samples will tend to cluster according to biology rather than by technology. On the other hand, if the clustering procedure does not properly work, even perfectly normalized data will not cluster in a meaningful way.

Unsurprisingly, considering the raw counts, the Yeast samples cluster according to the total number of reads. In fact, our proposed algorithm recovers two clusters, clearly representing lanes with high counts and low counts, respectively (Figure 5.3).

For the remainder of the Section we will consider data normalized using a median within-lane GC-content normalization and an upper-quartile between-lane normalization (the other normalizations gave similar results – data not shown). See Chapter 4 for details on the normalizations.

Table 5.1 reports the clustering of the samples in the Yeast dataset after normalization. Our algorithm is able to find three clusters, that have a clear biological meaning, since they correspond exactly to the three growth conditions. Recall that, unlike the k-means algorithm, our procedure is unsupervised also in the number of

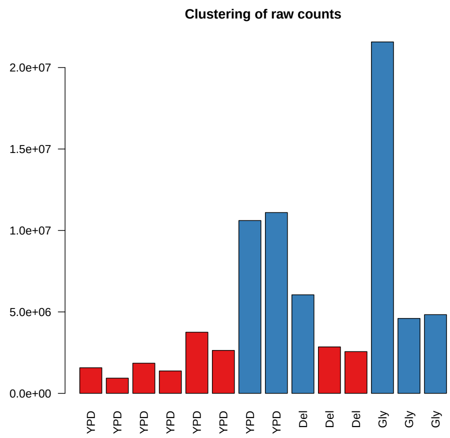


Figure 5.3: *Yeast dataset: clustering based on raw counts.* Barplot of the total number of reads per lane, color coded by cluster membership after a clustering procedure based on raw counts. It can be seen how the samples cluster by total number of reads.

clusters in which to group the samples. In other words, it is not necessary to specify the number of clusters and, at least for our datasets, the algorithm is able to recover the correct number of groups.

Table 5.1: *Yeast dataset:* clustering of lanes.

	1	2	3
YPD	8	0	0
Del	0	3	0
Gly	0	0	3

Moreover, since *pdfCluster* is a hierarchical clustering procedure, we can plot the hierarchy of the clusters, showing that it reflects the biology of the problem (Figure 5.4, recall that YPD and Delft conditions require fermentation while Glycerol needs respiration).

As for the MAQC datasets, we will consider only the MAQC-2 experiment, since in the MAQC-3 dataset there is only one biological sample and the clustering of the lanes is not meaningful.

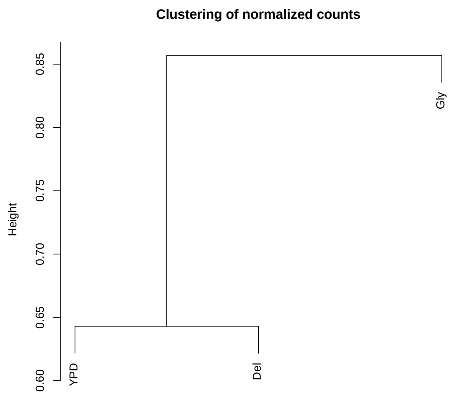


Figure 5.4: *Yeast dataset: clustering based on normalized counts.* Dendrogram of the three clusters produced by our clustering procedure, after normalization.

Table 5.2 shows the clustering of the lanes in the MAQC-2 data. As expected, the lanes representing the same biological sample cluster together and the procedure ends with two clusters, clearly representing the biology.

Table 5.2: *MAQC-2 dataset:* clustering of lanes.

	1	2
Brain	0	7
UHR	7	0

One important characteristic of our proposed algorithm, is the feature selection step. This step is important for the clustering itself: dropping the uninformative genes prior to the dimensionality reduction allows to preserve most of the information relevant for the clustering of the lanes. However, this step can be used as a "stand alone" tool for unsupervised gene filtering.

Feature selection has become important in genomic studies, both as a first step prior to subsequent analyses (Bourgon *et al.*, 2010; Tritchler *et al.*, 2009) and as a way to identify "gene signatures" (especially in cancer studies). For instance, Haury *et al.* (2011) showed how a univariate t-test can produce stable and interpretable signature when comparing different classes of cancer. It would be interesting to see the ability of an unsupervised criterion to find stable signature when no prior labeling of the samples is given.

In this view, it is worth noting that the feature selection step of our proposed procedure yields 281 genes in the Yeast data, which can be interpreted as those genes signifcant in the clustering of samples according to growth condition.

Chapter 6

Statistical Modeling

In this Chapter, we propose a hierarchical Bayesian model for RNA-Seq data, which accounts for overdispersion and for the dependence of the counts on technological biases, such as different sequencing depths and GC-content, without the need for data transformation (normalization).

The Chapter is organized as follows: in Section 6.1 we briefly review the literature on the modeling of RNA-Seq data. In Section 6.2, we propose a model for RNA-Seq data, and, in Section 6.3, we formulate the problem of differential expression within the framework of hierarchical Generalized Linear Models. In Section 6.4 we provide a discussion of the model and a comparison with previously proposed models. In Section 6.5 we perform a simulation study to assess the validity of the model. Finally, in Section 6.6 we show the results on the real datasets presented in Chapter 3, and compare them to the ones of *edgeR* and *DESeq*, two state-of-the-art algorithms for differential expression of RNA-Seq.

6.1 Literature Overview

As we detailed in Chapter 2, sequencing technologies produce short sequences, called "reads", consisting in a small number of DNA bases (usually 25 – 100), that in the context of RNA-Seq are mapped back to a reference genome. Subsequently, one counts the number of reads that fall within some pre-defined regions of interest (e.g., genes), leading to a discrete measure of expression (i.e., counts).

For this reason, the Poisson model has been proposed to find differentially expressed genes, as it is the simplest model to deal with count data (Bullard *et al.*, 2010a; Marioni *et al.*, 2008; Wang *et al.*, 2010). However, several authors have observed overdispersion when the experiment involves biological replicates (i.e., different individuals for each condition), proposing the negative binomial distribution as a more flexible alternative to model RNA-Seq data (Anders and Huber, 2010; Robinson *et al.*, 2010). They consider an exact test for the comparison between two groups as well as a likelihood ratio test for more general designs. The two approaches differ in the estimation of the library size effect and of the dispersion parameter.

Hardcastle and Kelly (2010) present an empirical Bayesian approach to differential expression in two-class and multi-class comparisons. By assuming a negative binomial distribution, they enumerate all the possible patterns of differential expression and retain the pattern which is more likely to generate the data. In a nonparametric setting, Tarazona *et al.* (2011) propose a test for differential expression in two-class comparisons based on the combination of log ratios and absolute differences between the mean expression of the two classes.

Hierarchical Bayesian models are a good alternative in this setting, since they can take advantage of the structure of the problem, allowing to borrow strength from the ensemble of the expression values (Ibrahim *et al.*, 2002). This class of models has been successfully applied to microarray data both from an empirical (Kendziorski *et al.*, 2003; Lönnstedt and Speed, 2002; Smyth, 2004) and a full Bayesian perspective (Baldi and Long, 2001). Moreover, the negative binomial distribution arises as a Gamma-Poisson mixture model (Lawless, 1987); hence one can model overdispersion in a hierarchical Bayes framework, using either a Gamma or a log-normal variable as a prior for the mean parameter.

6.2 Model Genesis

In the context of RNA-Seq, the use of the Poisson distribution naturally arises from the following model. Recall that we denote with $j = 1, \ldots, p$ the genes and with $i = 1, \ldots, n$ the samples (observations). Let us denote with l_j the length (in DNA bases) of the gene j, with c_{ij} the number of copies of the gene j in sample i, and with d_i the total number of reads produced for sample i (sequencing depth). We can model the sequencing process as a simple random sampling, in which every sequence is sampled independently and uniformly from the genome. This means that the gene counts follow a binomial distribution, i.e., $Y_{ij} \sim \mathrm{Bi}(d_i, \pi_{ij})$, where

$$\pi_{ij} = \frac{l_j c_{ij}}{\sum_{j=1}^{p} l_j c_{ij}}.$$

Since d_i is very large ($\sim 10^7$) and π_{ij} is very small, given that p is of the order of 10^4, the Binomial distribution is very well approximated by a Poisson, i.e, $Y_{ij} \sim \mathrm{Poi}(\lambda_{ij})$, where $\lambda_{ij} = d_i \, \pi_{ij}$.

In early publications on RNA-Seq, the Poisson model has been proposed and applied to model differential expression (Bullard *et al.*, 2010a; Marioni *et al.*, 2008; Wang *et al.*, 2010). In particular, Bullard *et al.* (2010a) analyzed the MAQC data of Section 3.1 and found that the Poisson model was appropriate after between-lane normalization. This is confirmed by the mean-variance relation represented in Figure 4.1, Panel (a).

However, the MAQC experiment is rather artificial, involving two standardized samples and technical rather than biological variability (see Section 3.1 for details on the experimental design). When a "real" experiment is considered (e.g. the Yeast data of Section 3.2), biological variability gives rise to overdispersion in the data (Figure 4.1, Panel (b)).

In a Bayesian approach, we can model overdispersion by introducing a hierarchy on the mean parameter of the Poisson. In particular, if one considers the mean parameter of the Poisson to follow a Gamma distribution, one recovers the negative binomial distribution. More formally, consider $Y|\lambda \sim \mathrm{Poi}(\lambda)$ and $\lambda \sim \mathrm{Gamma}(r, \theta)$. Then, by denoting with $p_Y(\cdot)$ and $p_\lambda(\cdot)$ the distribution of Y and λ, respectively, one obtains

$$\begin{aligned}
p_Y(y) &= \int_0^\infty p_{Y|\lambda}(y)\, p_\lambda(\lambda)\, d\lambda, \\
&= \frac{1}{y!\,\Gamma(r)\,\theta^r} \int_0^\infty \lambda^{r+y-1}\, e^{-\lambda(1+1/\theta)}\, d\lambda, \\
&= \frac{1}{\Gamma(y+1)\,\Gamma(r)\,\theta^r}\, \Gamma(r+y) \left(\frac{\theta}{\theta+1}\right)^{r+y}, \\
&= \binom{r+y-1}{y} \left(\frac{1}{\theta+1}\right)^r \left(1 - \frac{1}{\theta+1}\right)^y. \qquad (6.1)
\end{aligned}$$

Equation (6.1) is the probability function of a negative binomial (NB), i.e., $Y \sim NB(r, 1/(\theta+1))$. Usually, Y is parametrized in terms of its mean $\mu = r\theta$ and its variance function is

$$V(\mu) = \mu + \phi\mu^2, \tag{6.2}$$

where $\phi = 1/r$. Note that the Poisson is the limiting distribution of the NB for $\phi \to 0$.

Some real datasets present extremely high counts which are not fully captured by the NB. The Poisson-log-normal distribution (Anscombe, 1950; Bulmer, 1974) can be an alternative in this setting. Log-normal and Gamma mainly differ in skewness and kurtosis (Anscombe, 1950; Kempton and Taylor, 1974). In fact, both distributions are positively skewed, but the log-normal is more skewed and has longer tails. This results in a better fit in RNA-Seq applications in which is not unusual to observe extreme outliers.

By assuming a log-normal distribution on λ, i.e. $\lambda \sim \mathrm{LN}(\eta, \sigma^2)$, equation (6.1) becomes

$$p_Y(y) = \frac{1}{y!}\frac{1}{\sqrt{2\pi\sigma^2}}\int_0^\infty e^{-\lambda}\lambda^{y-1}\exp\left\{-\frac{1}{2\sigma^2}(\log y - \eta)^2\right\} d\lambda. \tag{6.3}$$

Although the integral in (6.3) does not have a closed form, the kth moment has the form

$$\mu_{[k]} = \exp\left\{k\eta + \frac{1}{2}k^2\sigma^2\right\},$$

as shown in Bulmer (1974). Therefore, the mean and variance function of the Poisson-log-normal are

$$\mu = \exp\left\{\eta + \frac{1}{2}\sigma^2\right\},$$
$$V(\mu) = \exp\left\{2\log\mu + \sigma^2\right\}. \tag{6.4}$$

Hence, the scale parameter of the log-normal distribution is linked to the variance of Y, in a similar way of the shape parameter of the Gamma in the NB case. Note however that in the case of the Poisson-log-normal one cannot recover the Poisson as a limiting case, while letting $\sigma^2 \to 0$ one recovers a quadratic variance function.

6.3 Model Specification

Most of gene expression studies are designed to identify genes that are DE between two (or more) conditions. GLMs are a natural class of models to deal with these problems, considering the gene expression level as the response, and an ANOVA type design matrix, that allows multiple class comparison. Note that the GLM approach is able to model other types of studies, in which continuous covariates are involved. Hence, our model is not limited to class comparison.

For each gene j ($j = 1, \dots, p$), we model the expression values as realizations of a Poisson distribution. The expected values are specified in the following way:

$$\log(E[Y_j|X, \beta_j]) = X\beta_j + \alpha, \tag{6.5}$$

where Y_j is the gene expression vector for gene j, X is the design matrix, β_j is a vector of parameters capturing the differences in mean among the classes, and α is a n-size vector of parameters used to model the different sequencing depths (i.e. the different total number of reads per sample).

For the sake of clarity, consider a two-class comparison. The β_j parameter in equation (6.5), has two components: the first component, namely $\beta_{0,j}$, can be interpreted as the log-mean expression of gene j for the first class, while the second component, namely $\beta_{1,j}$, can be interpreted as the log-difference between the mean expression of gene j in the two classes.

Denoting with i ($i = 1, \dots, n$) the samples, we consider the following prior distributions for the parameters.

$$\begin{aligned}
\alpha_i &\sim N(\mu_\alpha, \sigma_\alpha);\\
\mu_\alpha &\sim N(\mu_\mu, \sigma_\mu);\\
1/\sigma_\alpha &\sim Ga(\psi_0, \phi_0);\\
\beta_{0,j} &\sim N(\mu_{\beta_0}, \sigma_{\beta_0});
\end{aligned}$$

where $\mu_\mu = \mu_{\beta_0} = 0$, $\sigma_\mu = \sigma_{\beta_0} = 1$, $\psi_0 = \phi_0 = 10^{-2}$. As for $\beta_{1,j}$, we consider a mixture distribution, which allows us to model a situation in which most of the genes are not DE (i.e., $\beta_{1,j} = 0$),

$$\beta_{1,j} = w_j\, V_j + (1 - w_j)\, Z_j,$$

where

$$
\begin{aligned}
w_j &\sim Bi(1, \pi);\\
Z_j &\sim N(\mu_{Z_j}, \sigma_{Z_j});\\
V_j &\equiv 0;\\
\mu_{Z_j} &\sim N(\mu_{0Z_j}, \sigma_{0Z_j})
\end{aligned}
$$

where $\pi = 0.5$, $\mu_{0Z_j} = 0$, $\sigma_{Z_j} = \sigma_{0Z_j} = 1$ for all j.

Note that the hierarchy on the α_i parameters allows us to model the overdispersion in a global manner, i.e., the relation between mean and variance is assumed to be the same across all the genes. This is justified by the belief that overdispersion arises from biological variability. However, it is easy to have a gene-specific dispersion by considering a hierarchy on the variance of each $\beta_{0,j}$. Simulations gave similar results for both models (data not shown).

6.3.1 GC-content dependence

As described in Section 3.3, the read counts depend on both length and GC-content. In particular, GC-content dependence is library specific and therefore can strongly bias differential expression (as discussed in Chapter 4).

For this reason, we introduce the GC-content dependence in our proposed model in the following way. For each gene j ($j = 1, \ldots, p$), we model the expected expression values as

$$\log(E[Y_j|X, \alpha, \beta_j, \gamma, gc_j]) = X\beta_j + \alpha + \gamma gc_j, \tag{6.6}$$

where γ is a n-size vector of parameters that model the regression of the counts on the GC-content in a lane specific way. Note that, in this model, α cannot be interpreted as a library size effect, but it can be seen as an intercept in the GC-content regression.

The prior distributions for the parameters are the same as in the model in (6.5). As for γ,

$$\gamma_i \sim N(\mu_\gamma, \sigma_\gamma),$$

where $\mu_\gamma = 0$ and $\sigma_\gamma = 1$.

Although graphical inspections of the relation between counts and GC-content suggest a quadratic relation for the Yeast dataset (Figure 3.9, Panel (b)), we found that introducing a quadratic term in the model did not significantly improve the goodness-of-fit. We therefore opted for the more parsimonious choice of modeling a linear dependence of the (log) counts on GC-content.

6.4 Discussion

Robinson *et al.* (2010) and Anders and Huber (2010) propose a negative binomial to model RNA-Seq data in a frequentist setting and they implement the approach in the Bioconductor packages *edgeR* and *DESeq*, respectively. Although the NB is a flexible distribution that allows the modeling of overdispersion, it is known that for small sample sizes the usual maximum likelihood estimator tends to underestimate the dispersion parameter (Robinson and Smyth, 2008). For this reason, Robinson *et al.* (2010) use a weighted conditional log-likelihood approach to shrink the estimates of ϕ towards a common value, mimicking an empirical Bayes solution (Robinson and Smyth, 2007). They also propose to pull together the expression of all the genes and estimate a common dispersion parameter when the sample size is too small (Robinson and Smyth, 2008).

In a similar way, Anders and Huber (2010) assume that the per-gene raw variance is a smooth function of the mean and pool the data from genes with similar expression strength to estimate the variance using a local regression approach.

To deal with the difference in sequencing depth both Robinson *et al.* (2010) and Anders and Huber (2010) incorporate in the model a "size factor" which can be viewed as an offset in the model that accounts for the difference in the library sizes.

A good alternative to frequentist approaches in this setting is represented by hierarchical Bayesian models. Taking advantage of the structure of the problem, they borrow strength from the ensemble of the expression values and more effectively estimate the dispersion parameter. The log-normal prior allows to model experiments in which there is substantial overdispersion and outliers in the form of extremely high counts.

One important feature of the proposed model is that it can simultaneously estimate both biological effects (i.e. differential expression) and technical biases (e.g. GC-content, difference in sequencing depth). In the simpler model of Equation (6.5), the α parameter captures the difference in sequencing depth with no need of an *a priori* step of library size estimation, which is needed in both *edgeR* and *DESeq*. This step is essentially equivalent to a between-lane normalization and can be tricky, as discussed in Anders and Huber (2010) and Bullard *et al.* (2010a). In particular using the sum of the counts to estimate the library size can strongly bias differential expression (Bullard *et al.*, 2010a).

6.4.1 Differential expression

In order for the model to be useful to the researchers, there is the need for a way to declare a gene DE. One solution is to use credibility intervals and declare the gene j DE if the interval for $\beta_{1,j}$ does not include zero. Alternatively, one can use, in an empirical Bayes setting, a frequentist test statistic, using the posterior distribution of β to estimate the variance, in a way similar to Smyth (2004).

Depending on the biological problem and on the sensitivity of the technology, all genes could be considered DE to some extent. In fact, as the technology improves, there is more sensitivity to capture even very small differences in gene expression. For this reason, some authors have argued that using a threshold on the fold-change in combination with controlling for statistical variability is to be preferred to methods based solely on statistical significance (McCarthy and Smyth, 2009; Wu *et al.*, 2010).

Here, we declare the gene j as DE based on the posterior probability of $\beta_{1,j}$. Given a predefined threshold t, we say that j is up-regulated if

$$Pr(\beta_{1,j} > t|Y) \geq 0.9; \tag{6.7}$$

analogously, we say that j is down-regulated if

$$Pr(\beta_{1,j} < -t|Y) \geq 0.9. \tag{6.8}$$

The choice of t has clearly a strong impact on the results; however, often researchers are interested in the "most DE" genes, and they use the DE statistic to rank the genes and select the first genes as the genes of interest, on which to base subsequent analyses, such as pathway analysis or gene set enrichment.

For this reason, if one has in mind a reasonable proportion of genes to expect as DE, one can use an appropriate quantile of the distribution of the posterior means of the $\beta_{1,j}$'s to choose t. For instance, in the simulation study, we use the 95th percentile when considering a scenario in which the 5% of the genes are simulated as DE.

To compute the posterior distribution of the parameter, we consider a Markov Chain Monte Carlo (MCMC) algorithm, implemented using the JAGS software (Plummer, 2003). The JAGS specification of the model is shown in Algorithm 1.

Algorithm 1 Model specification in JAGS.

```
model {
for (i in 1:N) {
  alpha[i] ~ dnorm(mu,tau)
}

for (j in 1:M){
  beta0[j] ~ dnorm(0,1)
  beta1[j] <- w[j] * V[j] + (1 - w[j]) * Z[j]
  w[j] ~ dbern(0.5)
  V[j] <- 0
  Z[j] ~ dnorm(eta[j],1)
  eta[j] ~ dnorm(0,1)
}

for (i in 1:N) {
  for (j in 1:M) {
   lambda[j,i] <- exp(alpha[i] + beta0[j] + (beta1[j] * x[i]))
   y[j,i] ~ dpois(lambda[j,i])
   }
}

tau ~ dgamma(0.01,0.01)
mu ~ dnorm(0,1)
}
```

6.5 Simulation study

In this Section, we simulate the data from our proposed model varying the values of the parameters to mimic different scenarios. We will evaluate our model by considering the true and false positive rates (TPR and FPR, respectively), defining a gene as DE according to the rules in equations (6.7) and (6.8), choosing t as the quantile corresponding to the true proportion of DE genes (e.g., when simulating $\pi = 0.05$ in a balanced setting we will choose t as the 0.975 percentile in both (6.7) and (6.8)).

Moreover, we compare our model to two state-of-the-art approaches, namely *edgeR* and *DESeq*. In both cases, we followed the recommended pipeline, consisting in: (i) estimation of the "size factor" (analogous to our α parameter, or to a between-lane normalization), (ii) estimation of the dispersion parameter, (iii) testing for differential expression via an exact test based on the NB, (iv) definition of the true/false positives by considering a threshold of 0.05 on the (unadjusted) p-values.

6.5.1 Type I error and power considerations

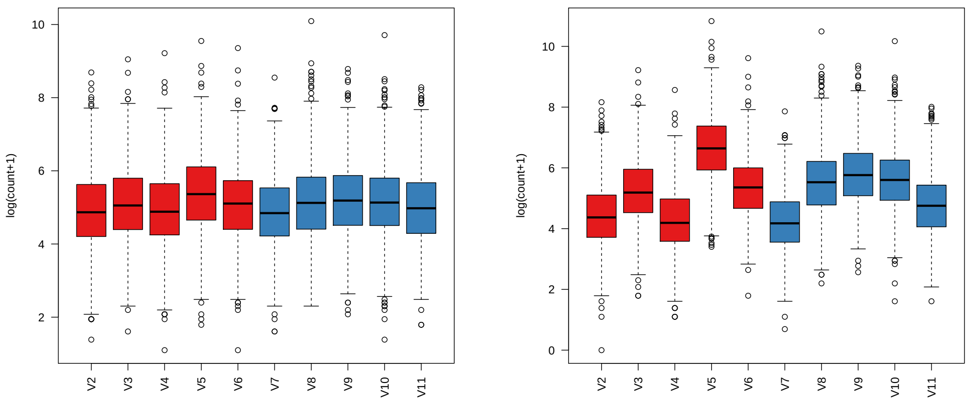
We consider $p = 1{,}000$ genes and $n = 10$ samples in a two-class comparison in which the first five samples belong to class "1" and the others to class "2". We simulate according to the model in equation (6.5), with the following specifications: $\mu_\alpha = 5$, $\sigma_\alpha = \{0.2, 1\}$, $\mu_{\beta_0} = 0$, $\sigma_{\beta_0} = 1$. As for β_1, we fix the values to consider that a subset of genes is DE between the two conditions, with a log-fold-change of 2, either balanced (i.e., 50% up- and 50% down-regulated), moderately imbalanced (i.e., 75% up- and 25% down-regulated) or strongly imbalanced (i.e., 100% up-regulated). We let the proportion of DE genes vary in size, by simulating the hyperparameter $\pi = \{0.05, 0.1, 0.3\}$. The combination of these choices leads to a total of 18 possible scenarios. For each scenario, we simulated $B = 100$ datasets.

The σ_α parameter controls the variability of the sequencing depths in the dataset. We choose these two values to mimic our real datasets. Figure 6.1 shows the distributions of the counts for two simulated datasets: Panel (a) is simulated with $\sigma_\alpha = 0.2$, while Panel (b) with $\sigma_\alpha = 1$.

Table 6.1 reports the results of the simulations for our proposed model. In general, the performance of our model is satisfying. In particular, the small number of false positives (FPR) means that the model is able to control for the type I error. The power of the test (TPR) is not extremely high; this can be due to the small sample

Table 6.1: True positive rate (TPR) and false positive rate (FPR) of our proposed DE test in the 18 scenarios described in the main text. The d parameter is the proportion of down-regulated genes among the DE.

σ_α	π	d	TPR	s.e.	FPR	s.e.
0.2	0.05	0.5	0.548	0.067	0.010	0.004
0.2	0.05	0.25	0.581	0.067	0.010	0.003
0.2	0.05	0	0.648	0.069	0.009	0.003
0.2	0.1	0.5	0.608	0.057	0.015	0.005
0.2	0.1	0.25	0.641	0.051	0.017	0.005
0.2	0.1	0	0.714	0.048	0.015	0.004
0.2	0.3	0.5	0.729	0.039	0.033	0.010
0.2	0.3	0.25	0.746	0.035	0.034	0.009
0.2	0.3	0	0.811	0.034	0.031	0.007
1	0.05	0.5	0.429	0.083	0.013	0.004
1	0.05	0.25	0.464	0.083	0.013	0.004
1	0.05	0	0.541	0.080	0.012	0.004
1	0.1	0.5	0.520	0.066	0.022	0.008
1	0.1	0.25	0.541	0.070	0.023	0.007
1	0.1	0	0.621	0.069	0.020	0.006
1	0.3	0.5	0.689	0.039	0.061	0.017
1	0.3	0.25	0.698	0.042	0.057	0.014
1	0.3	0	0.760	0.043	0.046	0.011



(a) $\sigma_\alpha = 0.2$ (b) $\sigma_\alpha = 1$

Figure 6.1: *Distribution of the simulated counts.* Gene-level counts per lane in two simulated datasets: (a) with $\sigma_\alpha = 0.2$ and (b) with $\sigma_\alpha = 1$.

size (5 samples per class) or to the nature of our DE statistic, and other strategies to call for DE genes can be explored to find an optimal solution.

As expected, as σ_α increases, the model finds it more difficult to fit the data. In fact, when α is more variable, the true expression is more likely to be confounded with the differences in sequencing depth (see Figure 6.1). For instance, the scenario in which $\sigma_\alpha = 1$ and $\pi = 0.3$, corresponding to a large proportion of DE genes in a highly variable experiment in terms of sequencing depth, is the only case which leads to a FPR greater than 0.05 (last three rows of Table 6.1).

Surprisingly, the test has more power for the imbalanced scenarios with respect to the balanced ones. This is an artifact due to the fact that the posterior distribution of $\beta_{1,j}$ tends to be shifted to the right. This is a systematic effect that we are still investigating and can be seen in Figure 6.2. This figure shows the distribution of the posterior means of $\beta_{1,j}$ in one simulated dataset ($\sigma_\alpha = 0.2$, $\pi = 0.05$, $d = 0.5$ — not all the datasets show this shifting). One possible solution to reduce this unwanted effect could be to introduce a constrain on the alpha parameter and future effort will be directed toward this direction.

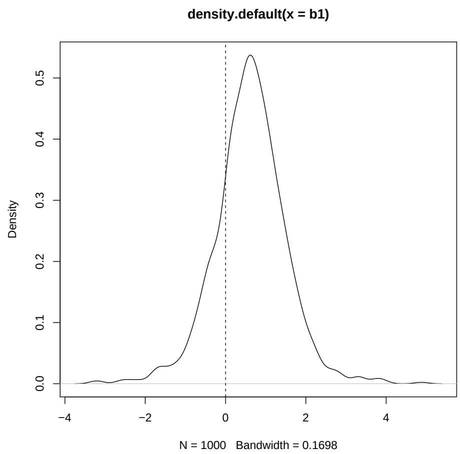


Figure 6.2: *Distribution of the posterior means of* $\beta_{1,j}$. One simulated dataset with $\sigma_\alpha = 0.2$, $\pi = 0.05$ and $d = 0.5$.

6.5.2 Comparison with *edgeR* and *DESeq*

Table 6.2 reports the results for *edgeR*. In all scenarios the test is anti-conservative, failing to control for the type I error.

This algorithm is not sensitive to the higher variability in sequencing depth, being FPR and TPR similar independently of σ_α. However, a strong imbalance between up- and down-regulated DE genes increases the FPR and decreases the TPR, especially when the number of DE genes is large (i.e., $\pi = 0.3$ and $d = 0$).

Table 6.3 shows the results for *DESeq*. Strikingly, things are very different from *edgeR*: *DESeq* is able to control the type I error at the cost of a very low power.

In particular, *DESeq* is very sensitive to the high differences in sequencing depths and this leads to a TPR of less than 0.1 when $\sigma_\alpha = 1$ and $\pi = 0.05$.

Moreover, independently of the other parameters, when the DE genes are imbalanced the power decreases.

It is somewhat surprising that *edgeR* and *DESeq* lead to such different results. Indeed, they are both based on a NB model and they differ only by the way they estimate the size factor and the dispersion parameter.

Our feeling is that the difference is made by how the two algorithms estimate the dispersion: *edgeR* uses a likelihood-based approach (Robinson *et al.*, 2010), while *DESeq* exploits a local regression on the mean-variance scatterplot to estimate the

Table 6.2: True positive rate (TPR) and false positive rate (FPR) of *edgeR* DE test in the 18 scenarios described in the main text. The d parameter is the proportion of down-regulated genes among the DE.

σ_α	π	d	TPR	s.e.	FPR	s.e.
0.2	0.05	0.5	0.842	0.055	0.106	0.011
0.2	0.05	0.25	0.839	0.053	0.107	0.011
0.2	0.05	0	0.819	0.052	0.109	0.011
0.2	0.1	0.5	0.851	0.036	0.107	0.012
0.2	0.1	0.25	0.839	0.035	0.109	0.011
0.2	0.1	0	0.794	0.042	0.116	0.011
0.2	0.3	0.5	0.873	0.022	0.110	0.013
0.2	0.3	0.25	0.808	0.025	0.141	0.013
0.2	0.3	0	0.642	0.031	0.219	0.018
1	0.05	0.5	0.840	0.055	0.105	0.011
1	0.05	0.25	0.841	0.053	0.106	0.011
1	0.05	0	0.820	0.052	0.107	0.011
1	0.1	0.5	0.849	0.036	0.106	0.012
1	0.1	0.25	0.836	0.035	0.108	0.011
1	0.1	0	0.792	0.042	0.116	0.011
1	0.3	0.5	0.872	0.023	0.110	0.013
1	0.3	0.25	0.807	0.027	0.139	0.014
1	0.3	0	0.636	0.033	0.221	0.021

Table 6.3: True positive rate (TPR) and false positive rate (FPR) of *DESeq* DE test in the 18 scenarios described in the main text. The d parameter is the proportion of down-regulated genes among the DE.

σ_α	π	d	TPR	s.e.	FPR	s.e.
0.2	0.05	0.5	0.319	0.114	0.020	0.005
0.2	0.05	0.25	0.315	0.105	0.020	0.005
0.2	0.05	0	0.284	0.100	0.021	0.005
0.2	0.1	0.5	0.405	0.072	0.020	0.005
0.2	0.1	0.25	0.381	0.081	0.021	0.005
0.2	0.1	0	0.313	0.072	0.024	0.006
0.2	0.3	0.5	0.524	0.039	0.022	0.007
0.2	0.3	0.25	0.433	0.038	0.031	0.007
0.2	0.3	0	0.202	0.032	0.058	0.010
1	0.05	0.5	0.080	0.080	0.008	0.005
1	0.05	0.25	0.081	0.076	0.008	0.005
1	0.05	0	0.071	0.077	0.008	0.005
1	0.1	0.5	0.118	0.083	0.008	0.005
1	0.1	0.25	0.108	0.081	0.008	0.005
1	0.1	0	0.080	0.070	0.009	0.006
1	0.3	0.5	0.208	0.099	0.008	0.005
1	0.3	0.25	0.166	0.084	0.013	0.009
1	0.3	0	0.063	0.049	0.027	0.016

dispersion (Anders and Huber, 2010). It could be that this solution is more robust to the misspecification of the model (in terms of type I error) at the cost of less power.

To rule out the size factor estimation as the reason for the difference in the performance of *edgeR* and *DESeq*, we repeated the analyses on the simulated datasets after upper-quartile between-lane normalization (see Chapter 4 and Bullard *et al.*, 2010a). For both models, the results are very similar to those of Tables 6.2 and 6.3 (data not shown), suggesting that the dispersion estimation is responsible for the difference in the performance of the two approaches.

Overall, our proposed model outperforms both *edgeR* and *DESeq* in these simulations. In fact, the high TPR of *edgeR* comes at the cost of the high FPR, meaning that the test fails to control for the type I error (recall that we used a threshold of 0.05 on the p-values). On the other hand, *DESeq* controls for the type I error, but has a very low power. Note that we are simulating from our proposed model, namely a Poisson-log-normal model, while both *edgeR* and *DESeq* are based on a NB. Therefore, we expect that our model will outperform these approaches in this setting, since *edgeR* and *DESeq* assume the overdispersion to be quadratic and we are simulating data with an exponential variance function (see equations (6.2) and (6.4)). However, this simulation study was meant to be a confirmation of the goodness of the model when all the assumptions hold.

To test if our approach is robust to the misspecification of the model, we consider a smaller simulation study, by simulating the data according to a NB model. We consider just one setting, in which we have a proportion of DE genes equal to 0.05, balanced between up- and down-regulated and a dispersion parameter of 0.1, similar to the common dispersion of the Yeast dataset, as estimated by *edgeR* after between-lane normalization. We simulate $B = 100$ datasets.

Table 6.4 reports the FPR and TPR for the three approaches. The results are strikingly good for all the models. In particular, both *edgeR* and *DESeq* are able to control for the type I error (*DESeq* being slightly conservative) and have an extremely high power. Surprisingly, our model is able to achieve almost the same power as the others, having an extremely low FPR. This can be explained by the way we declare a gene DE: besides its statistical significance, the gene must have a high fold-change to be called DE, and this helps in the reduction of false positives, as already noted in McCarthy and Smyth (2009) for microarray data. Even if these are only preliminary results and a more systematic simulation study has to be performed on NB simulated

Table 6.4: True positive rate (TPR) and false positive rate (FPR) of the three DE test in the NB simulated datasets.HB refers to our proposed hierarchical Bayesian model.

model	TPR	s.e.	FPR	s.e.
HB	0.998	0.009	0.000	0.000
edgeR	1.000	0.000	0.053	0.007
DESeq	1.000	0.000	0.035	0.007

data, the results indicate that our proposed model works well even when the data are NB distributed.

6.5.3 Sequencing depth

Finally, we consider the ability of the model to capture the difference between the sequencing depths of the experiments via the α parameter. In particular, we want to test how the simultaneous estimation of α and β reduces the power of detecting DE genes.

For this reason, we simulate the data with $\sigma_\alpha = 0$ (no sequencing depth difference) and estimate the model in equation (6.5) with and without α.

We simulate the data with the following parameters: $\sigma_\alpha = 0$, $\pi = 0.05$ and $d = \{0.5, 0.25, 0\}$. We simulate $B = 100$ datasets for each of the three scenarios. Table 6.5 reports the results in this setting analyzed with the model with and without α.

The general tendency of the model to have more power in the imbalanced settings is confirmed even without the α parameter in the model. The power of the test is greater when α is not present in the model. This is to be expected since the data are generated without sequencing depth difference and if α is present some information has to be used for its estimation. Overall, however, the results are good also when α is present, indicating that the model is able to find the DE genes also in the presence of the additional α parameter.

Table 6.5: True positive rate (TPR) and false positive rate (FPR) of the DE test in our proposed model with and without α

model	d	TPR	s.e.	FPR	s.e.
Without α	0.5	0.637	0.056	0.014	0.003
Without α	0.25	0.646	0.053	0.014	0.003
Without α	0	0.707	0.049	0.012	0.003
With α	0.5	0.552	0.080	0.009	0.003
With α	0.25	0.581	0.065	0.009	0.003
With α	0	0.650	0.062	0.008	0.003

6.6 Results on real data

We apply the model to the MAQC-2 and Yeast datasets introduced in Chapter 3. For the MAQC-2 data we consider the two-class comparison between Brain and UHR samples. This means that the β_j vectors of equation (6.5) have two components and that the second component, namely $\beta_{1,j}$, represents the log-fold-change between the two conditions. As for the Yeast dataset, we consider the three-class comparison among the three growth conditions (i.e., YPD, Delft, Glycerol). Therefore, β_j has three components, and contrasts can be used to provide pairwise comparisons.

6.6.1 Evaluation Criteria

Figures A.11 and A.12 show the convergence of the chains for $\beta_{1,j}$ in the MAQC-2 and Yeast datasets, respectively, for three randomly selected genes.

To evaluate the goodness of our inferential procedure, we consider the gene j to be DE if one between (6.7) and (6.8) holds true, choosing a threshold $t = 2$. To examine if the model can estimate the technical effect and if the resulting estimates of β are meaningful even without normalization, we fit the model (with and without GC-content effect) both to raw and between-lane normalized counts.

We compare the results of our model to those of *edgeR* and *DESeq*, two widely used algorithms for the identification of DE genes. For both algorithms we follow the suggested pipeline of estimating the size factors by considering the TMM normalization (Robinson and Oshlack, 2010) for *edgeR* and the scaling described in Anders and

Huber (2010) for *DESeq*. Since both methods return a list of p-values, we consider a gene to be DE if its Bonferroni adjusted p-value does not exceed 0.01.

For the MAQC-2 data, a subset of approximately 1,000 genes have been assayed by qRT-PCR. As detailed in Chapter 3, we found 638 genes detected by RNA-Seq and assayed by qRT-PCR; we use this subset to compare the two technologies. One can use the estimate of the UHR/Brain fold-change from qRT-PCR as the true value, since qRT-PCR is often considered as a gold standard for producing accurate estimates of expression levels. Following the strategy described in Bullard *et al.* (2010a), we consider three possible sets of genes: "non-DE", "DE", and "no-call", based on whether their qRT-PCR absolute log-fold-change is less than 0.2, greater than 2, or falls within the interval $[0.2, 2]$, respectively. We obtain 208 "DE" and 81 "non-DE" genes. We can then estimate Type I error and power of each test by considering the true positive and false positive rates.

For the Yeast dataset, we do not know which genes are truly DE, hence we can compare the methods only by relating the lists of DE genes. Nonetheless, it is important to examine the behavior of the models on a biologically meaningful experiment in which there is overdispersion and with a more complex experimental design.

6.6.2 MAQC data

First, we consider the correlation between the estimates of $\beta_{1,j}$ and the log-fold-changes obtained by qRT-PCR (Figure 6.3). Strikingly, the estimates are highly correlated with the gold standard, confirming the goodness of the model.

Moreover, when applying the model to raw and normalized counts the estimates of $\beta_{1,j}$ are only slightly different, as one can see from the distribution of the posterior means (Figure 6.4, Panel (a)), confirming that α is a good way to estimate the library size effect. Finally, when adding the parameter for the GC-content effect, the estimates of $\beta_{1,j}$ present only small changes (green and blue curves of Figure 6.4, Panel (a)). This is to be expected since the GC-content effect is weak in the MAQC dataset (Figure 3.8).

Table 6.6 shows the number of up- and down-regulated genes obtained after applying the model with and without the GC-content effect to the raw and normalized data. We found about 700 up-regulated and 1,000 down-regulated genes (out of

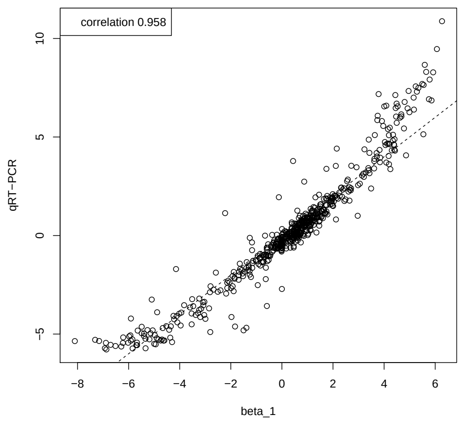
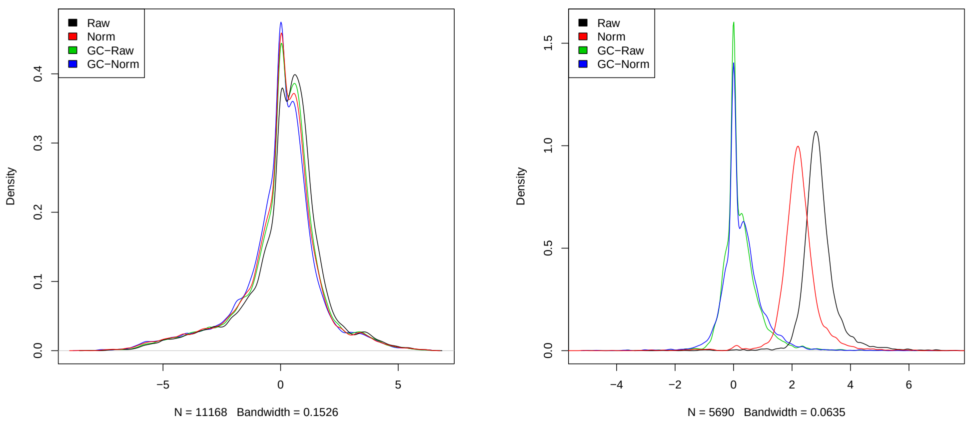


Figure 6.3: *MAQC-2 dataset: correlation with qRT-PCR.* Scatterplot of the UHR/Brain expression log-ratio of qRT-PCR data vs. the posterior mean of β_1. The high correlation (0.96) is a good indication of the goodness of the model.



(a) MAQC (b) Yeast

Figure 6.4: *Distribution of the posterior means of* $\beta_{1,j}$. (a) Posterior means of $\beta_{1,j}$ for the MAQC-2 dataset. Both models applied to both raw and normalized counts lead to a similar distribution. (b) For the Yeast data, if one does not include the GC-content in the model, the distributions of $\beta_{1,j}$ are shifted to the right and normalization can correct only partially this bias. When the GC-content is considered the model estimates correctly β_1 both with raw and normalized counts.

Table 6.6: *MAQC-2 dataset: DE genes.* Number of up- and down-regulated genes for our proposed model with and without GC-content and using raw or between-lane normalized data, *edgeR* and *DESeq*. Both for *edgeR* and *DESeq* we estimated the size factors as suggested by the authors.

	Raw	Normalized	GC-Raw	GC-Norm	Common	edgeR	DESeq
Up	764	684	715	685	640	3957	4480
Down	964	1103	1041	1101	955	3994	4501

Table 6.7: *MAQC-2 dataset: true and false positive rates for the UHR/Brain comparison.* True positive rates (TPR) and false positive rates (FPR) for the UHR/Brain comparison using our proposed model with and without the GC-content and using raw or between-lane normalized data, *edgeR* and *DESeq*. Both for *edgeR* and *DESeq* we calculated the size factors as suggested by the authors.

	Raw	Normalized	GC-Raw	GC-Norm	edgeR	DESeq
TPR	0.856	0.879	0.856	0.879	0.995	0.995
FPR	0.000	0.000	0.000	0.000	0.321	0.506

12,340) in the Brain/UHR comparison, 90% of which are in common between the models. These numbers are much smaller than the DE genes identified by *edgeR* and *DESeq* and this could mean either lack of power of our test or the failure in the control of Type I error for the NB tests carried out by *edgeR* and *DESeq*, or, more likely, a combination of these two effects. Table 6.7 shows that, at least for the subset of genes assayed by qRT-PCR, our model has less power but is able to control for Type I error, while the other two approaches lead to anti-conservative tests.

The results of our model and of *edgeR* in the real data are similar to those obtained in the simulations. Surprisingly, *DESeq* behave very differently in the real data with respect to the simulations.

This is probably due to the library size estimation step that, as shown in Figure 6.5, is biased for both *edgeR* and *DESeq*. In fact, for the MAQC dataset, the difference in sequencing depth is confounded with the biological difference between Brain and

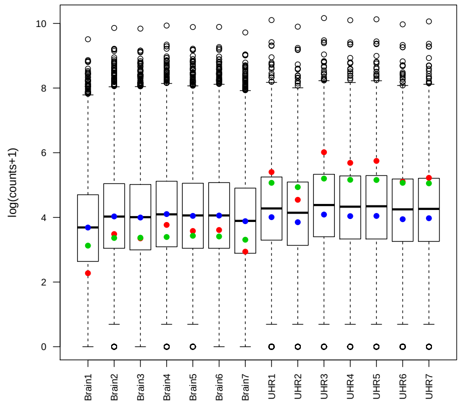


Figure 6.5: *MAQC-2 dataset: sequencing depth estimation.* Gene-level counts per lane for the MAQC dataset. The points represent the library sizes (scaled to scatter around the median) as estimated by: green: *edgeR* ; red: *DESeq* ; blue: our proposed model (α parameter).

UHR, in which many more genes are expressed, being a pool of different tissues. Both *edgeR* and *DESeq* under-estimate the sequencing depth for the Brain samples and over-estimate it for UHR. Our model, simultaneously estimating α and β, is able to account for the limited differences in sequencing depth, while capturing the biological difference of interest between the two samples.

6.6.3 Yeast data

As we said, in the Yeast dataset the GC-content effect is strong and can bias differential expression analysis (Figures 3.9 and 4.2). If one does not consider the GC-content effect, our model is not able to correctly estimate the biological effect of interest, even when considering between-lane normalized counts (Figure 6.4, Panel (b)). When considering the GC-content in the model, using raw or normalized counts leads to similar estimates of β.

Table 6.8 reports the number of DE genes for the three pairwise comparisons between the three growth conditions.

We can see how the conservative nature of our proposed test is confirmed in the Yeast dataset. Note that the higher number of DE genes for the comparison between

Table 6.8: *Yeast dataset: DE genes.* Number of DE genes (out of 5,690) for the pairwise comparisons between the three growth conditions, namely, Del/YPD, Gly/YPD and Gly/Del, for our proposed model with GC-content, *edgeR* and *DESeq*, using raw counts. Both for *edgeR* and *DESeq* we calculated the size factors as suggested by the authors.

	Del vs. YPD	Gly vs. YPD	Gly vs. Del
	$\beta_{1,j}$	$\beta_{2,j}$	$\beta_{2,j} - \beta_{1,j}$
proposed model	116	664	448
edgeR	439	1104	576
DESeq	Total number of DE genes: 1998		

glycerol and the other conditions is to be expected since glycerol requires respiration for yeast to grow, while the other conditions require fermentation. *DESeq* returns only a list of p-values for the global comparison between the model with the growth condition factor and the null model. For this reason we did not attempt to assign DE genes to one of the pairwise comparisons. However, one can manually inspect the estimates of the parameters of the model to infer which was the condition responsible for the differential expression.

Of the 779 (out of 5,690) DE genes found using our hierarchical Bayes model, 83% are found as DE also by *edgeR* and 77% by *DESeq*. Moreover, 72% of them are common to all three methods, meaning that all models can identify the genes that more strongly contribute to the cell biology.

Chapter 7

Conclusions

In this Thesis, we have shed some light on the problems related to the exploration and modeling of RNA-Seq data.

In Chapter 3, we have reported the existence of strong sample-specific GC-content effects on RNA-Seq read counts, which can substantially bias differential expression analysis, and in Chapter 4 we have proposed three simple within-lane gene-level GC-content normalization approaches. The GC-content effect seems to be the same for technical replicates but can vary substantially between biological replicates, suggesting that the bias is likely to be introduced at the library preparation step (as noted in Benjamini and Speed, 2011, for DNA-Seq). We have compared our methods to the state-of-the-art CQN procedure of Hansen *et al.* (2011), which was shown to outperform competing methods such as that of Pickrell *et al.* (2010), on both yeast and human RNA-Seq data, in terms of bias and mean squared error for expression fold-change estimation and in terms of Type I error and p-value distributions for tests of differential expression. Our proposed within-lane procedures, followed by between-lane normalization as in Bullard *et al.* (2010a), reduce GC-content bias and lead to more accurate estimation of expression fold-changes and tests of differential expression.

In Chapter 5, we have discussed a general strategy for the clustering of RNA-Seq expression data, based on variance stabilizing, gene filtering and dimensionality reduction as preliminary steps in the cluster analysis.

We have discussed a nonparametric density estimation-based algorithm within this framework, showing promising results in the real datasets, with surprisingly good computational performances.

Here we have used `pdfCluster` in order to select relevant genes. We underlined the assets of this choice, but it is clear that any unsupervised technique able to discard the irrelevant genes can be used for the gene filtering step. Similarly, the choice of principal component analysis in the dimensionality reduction step is only one among several possible choices. Since there are no guarantees that the first principal components preserve the cluster structure (Menardi and Torelli, 2010), future efforts could be made in trying different approaches, such as the projection pursuit (Friedman and Tukey, 1974; Menardi and Torelli, 2010) or the principal curves (Hastie and Stuetzle, 1989). Nevertheless, in our case the principal component analysis gives good results and provides a low dimensional dataset on which it is feasible to apply a model-based technique such as `pdfCluster`.

As for the clustering in the reduced space, we have exploited the nonparametric approach of `pdfCluster`. To do this, we have to apply a variance stabilizing transformation to the counts since `pdfCluster` needs continuous values to estimate the underlying density function. A possible alternative would be to consider a model-based clustering algorithm specifically designed for count data, e.g., based on a Poisson or a negative binomial model.

In Chapter 6, we have presented a hierarchical Bayesian GLM to model RNA-Seq data. Through an MCMC algorithm, we estimated the posterior distribution of the parameters and used them for inference on differential expression.

The model has the advantage of being very general and can model a wide range of studies, e.g., multi-class comparisons and designs with continuous covariate of interest.

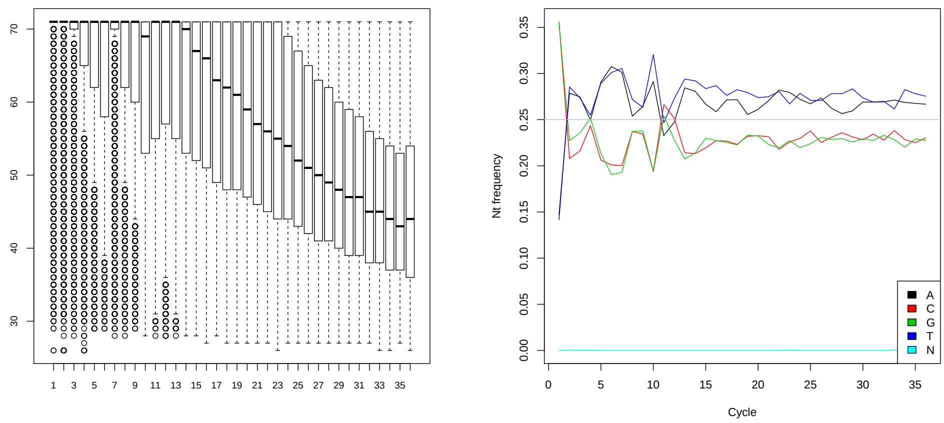
Our model shows promising results when compared to existing approaches, both in simulations and in real data when using qRT-PCR data as a gold standard. In order for the model to be useful to the researchers, there is the need for a way to declare a gene differentially expressed. One can use the posterior credibility intervals and call a gene differentially expressed if the interval does not contain zero or use a threshold on the posterior probability of $\beta_{1,j}$ as we have done here. Nonetheless, future effort could be made on formal specification of the posterior probability of a differential expression statistic, in an empirical Bayes setting.

Moreover, the MCMC algorithm is computationally expensive, and this is a major concern in high-throughput studies, when several thousands of genes are analyzed at once. We will therefore take into consideration the possibility of exploiting the potential offered by the integrated nested Laplace approximations approach to overcome this limitation (Rue *et al.*, 2009).

Finally, we have observed a systematic shifting toward high values in the posterior distribution of $\beta_{1,j}$. This is clearly an unwanted effect and perhaps it can be reduced by introducing a constrain on the α parameter. One possibility is to force the geometric mean of the α's to be one, i.e., $\prod_{i=1}^{n} \alpha_i = 1$.

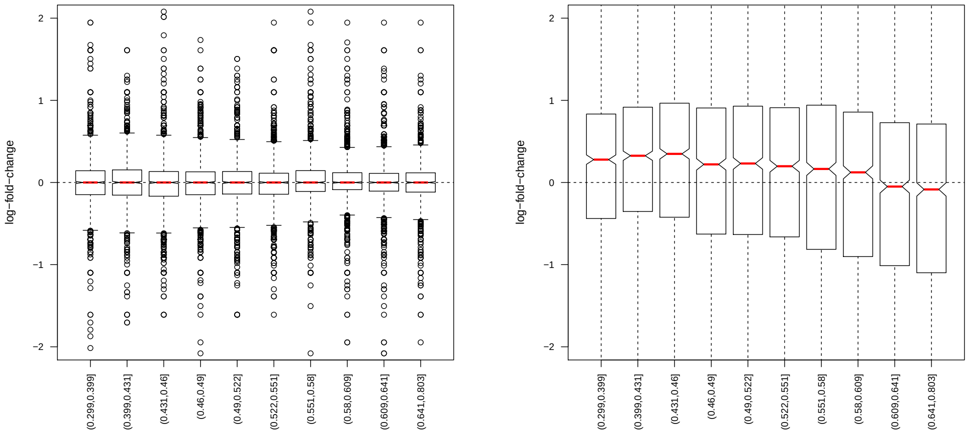
Appendix A

Additional Figures



(a) Quality (b) Nucleotide frequencies

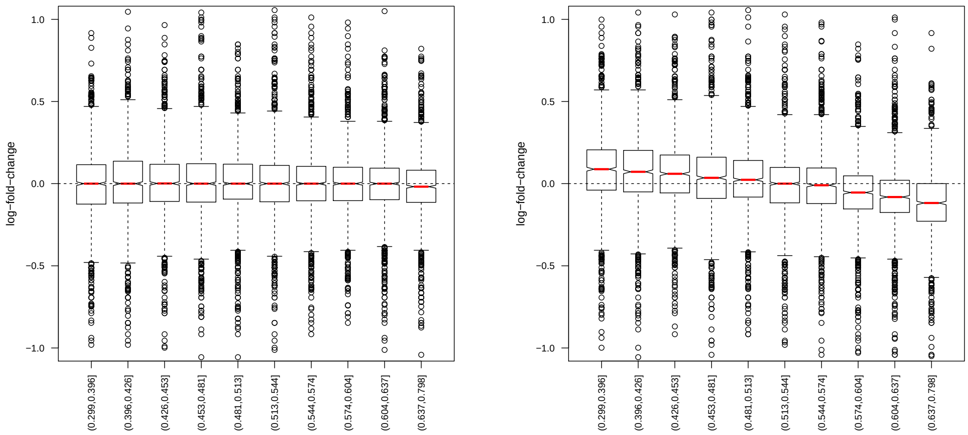
Figure A.1: *Illustration of the EDASeq package.* Panel (a): Per-base quality scores of mapped reads. Panel (b): Per-base nucleotide frequencies of mapped reads.



(a) Replicate lanes

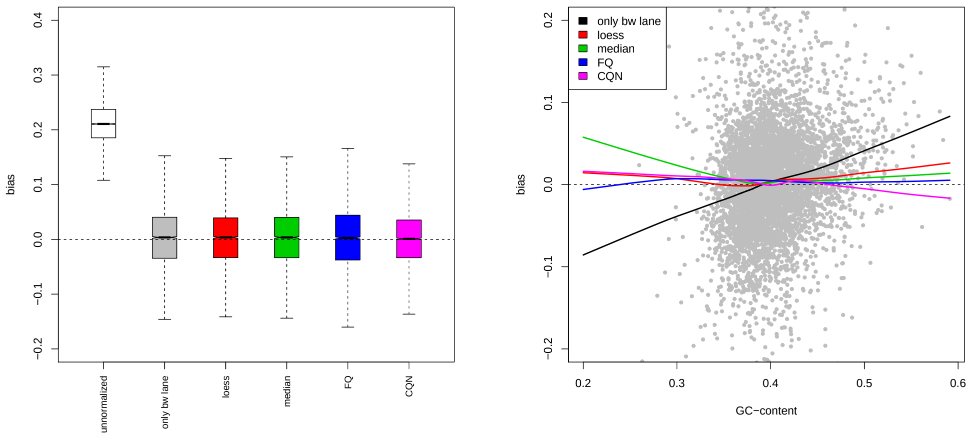
(b) UHR vs. Brain

Figure A.2: *MAQC-2 dataset: Log-fold-change vs. GC-content.* Stratified boxplots of count log-ratios vs. GC-content, after FQ between-lane normalization. Panel (a): UHR lane 1 vs. UHR lane 2 (same flow-cell). Panel (b): UHR lane 1 vs. Brain lane 1 (same flow-cell). The GC-content effect appears to be the same for the two technical replicates, so that fold-change estimates do not vary with GC-content. By contrast, the GC-content effect appears to differ between UHR and Brain libraries and to confound fold-change estimation. Note, however, that here the extreme difference between the two biological samples makes the dependence on GC-content difficult to assess.



(a) Replicate lanes (b) Replicate libraries

Figure A.3: *MAQC-3 dataset: Log-fold-change vs. GC-content.* Stratified boxplots of count log-ratios vs. GC-content, after FQ between-lane normalization. Panel (a): Lane 1 vs. lane 2 of library "A" (same flow-cell). Panel (b): Lane 1 of library "A" vs. lane 1 of library "B" (same flow-cell). The GC-content effect appears to be the same for the two technical replicate lanes, so that fold-change estimates do not vary with GC-content. By contrast, the GC-content effect appears to differ between the two different library preparations and to confound fold-change estimation.



(a) Bias (b) Bias vs. GC-content

Figure A.4: *Yeast YPD pseudo-datasets: Bias in fold-change estimation.* For a given gene, bias is estimated as the average of the 35 log-ratios from the YPD pseudo-datasets. Panel (a): Boxplot of bias in log-fold-change estimates. All normalization procedures reduce bias. Panel (b): Dependence of bias on GC-content. The points correspond to bias after only FQ between-lane normalization, the curves are lowess fits of bias vs. GC-content for different within-lane GC-content normalization procedures.

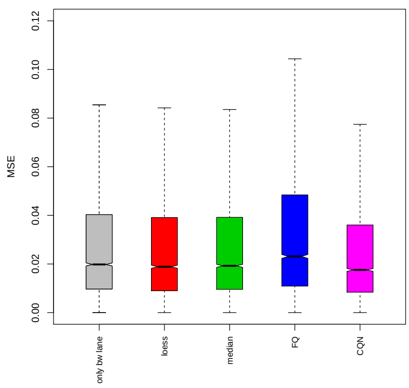
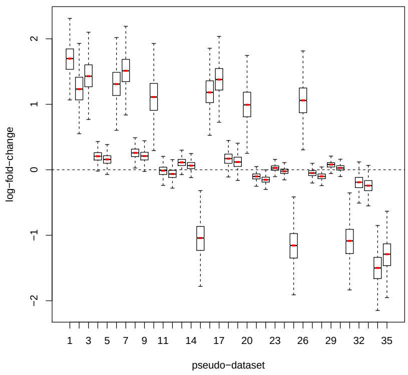
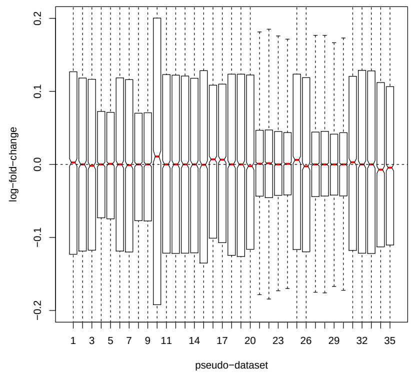


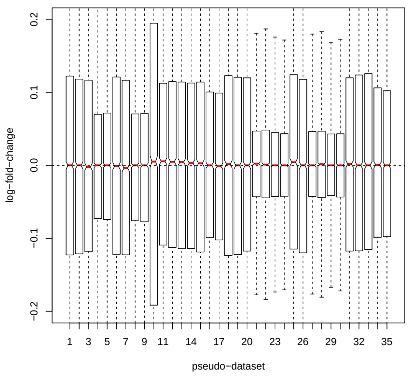
Figure A.5: *Yeast YPD pseudo-datasets: Mean squared error in fold-change estimation.* For a given gene, MSE is estimated as the average of the square of the 35 log-ratios from the YPD pseudo-datasets. Between-lane normalization greatly reduces MSE and within-lane normalization does not appear to further reduce MSE.



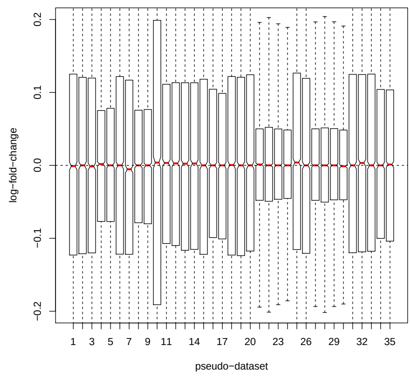
(a) Unnormalized



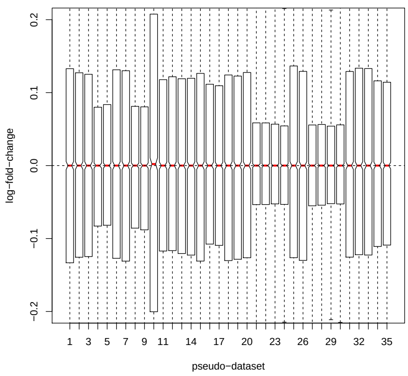
(b) Only between-lane normalization



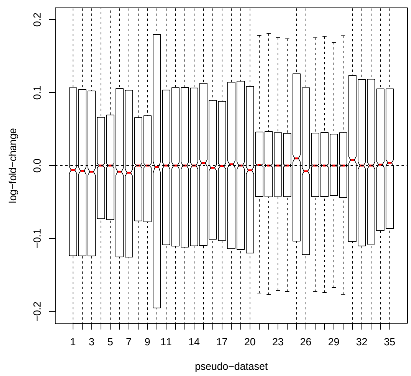
(c) loess



(d) Median

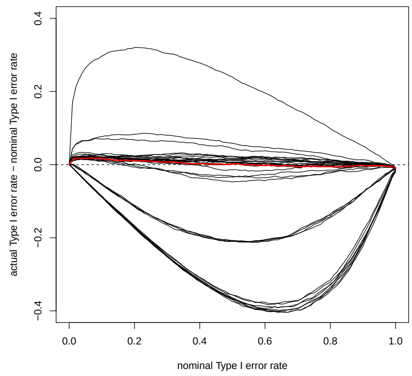


(e) FQ

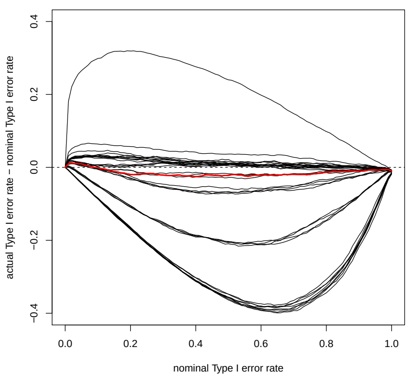


(f) CQN

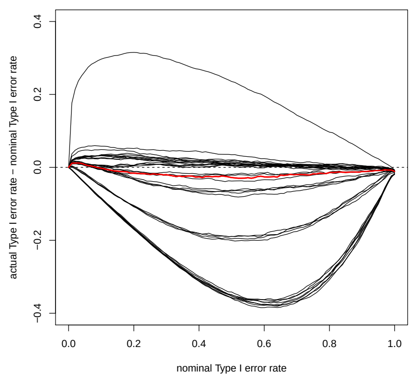
Figure A.6: *Yeast YPD pseudo-datasets: Distribution of log-fold-changes for the 35 YPD pseudo-datasets.* One would expect the log-fold-changes to be around zero for each dataset. There is a clear bias for unnormalized counts, with log-fold-change estimates as high as 2 (note different scale for Panel (a)). The FQ within-lane normalization method seems to be the most coherent in estimating the log-fold-change around zero.



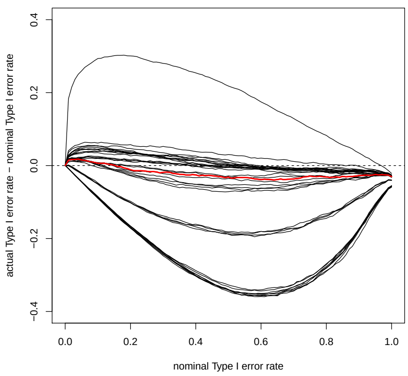
(a) Only between-lane normalization



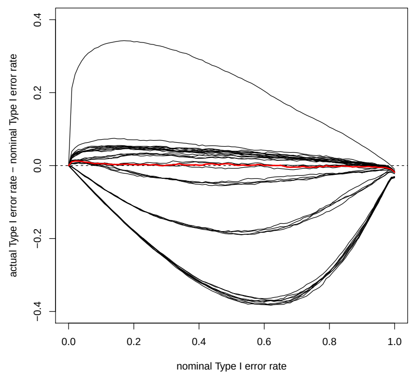
(b) loess



(c) Median

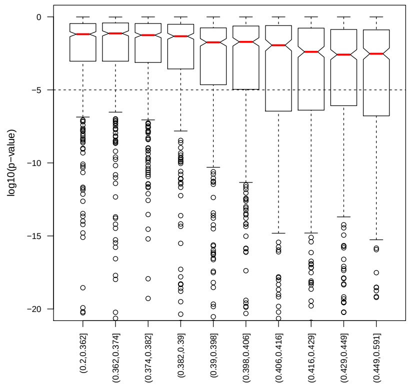


(d) FQ

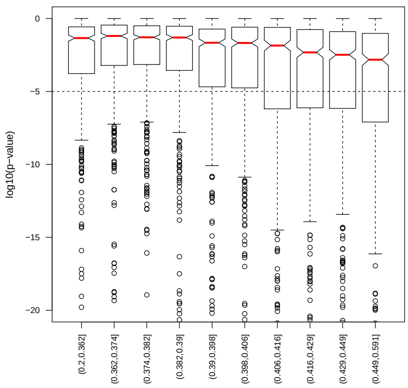


(e) CQN

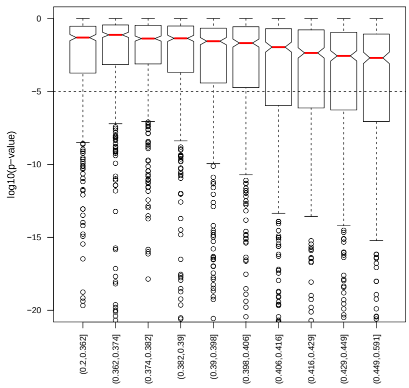
Figure A.7: *Yeast YPD pseudo-datasets: Type I error.* Difference between actual and nominal Type I error rates vs. nominal Type I error rate for each of the 35 YPD pseudo-datasets, for different normalization procedures. In red: median difference in Type I error rates.



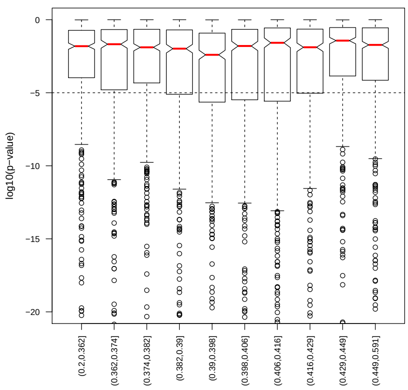
(a) Only between-lane normalization



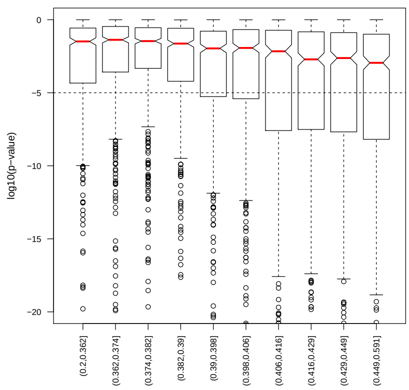
(b) loess



(c) Median

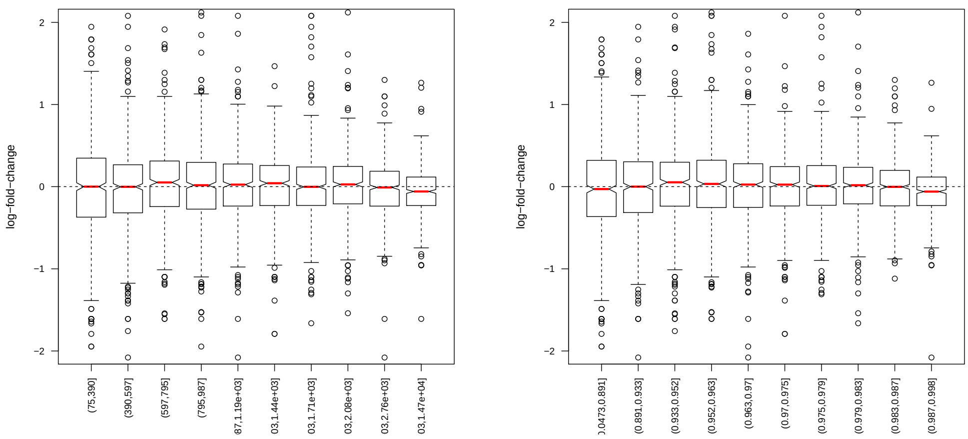


(d) FQ



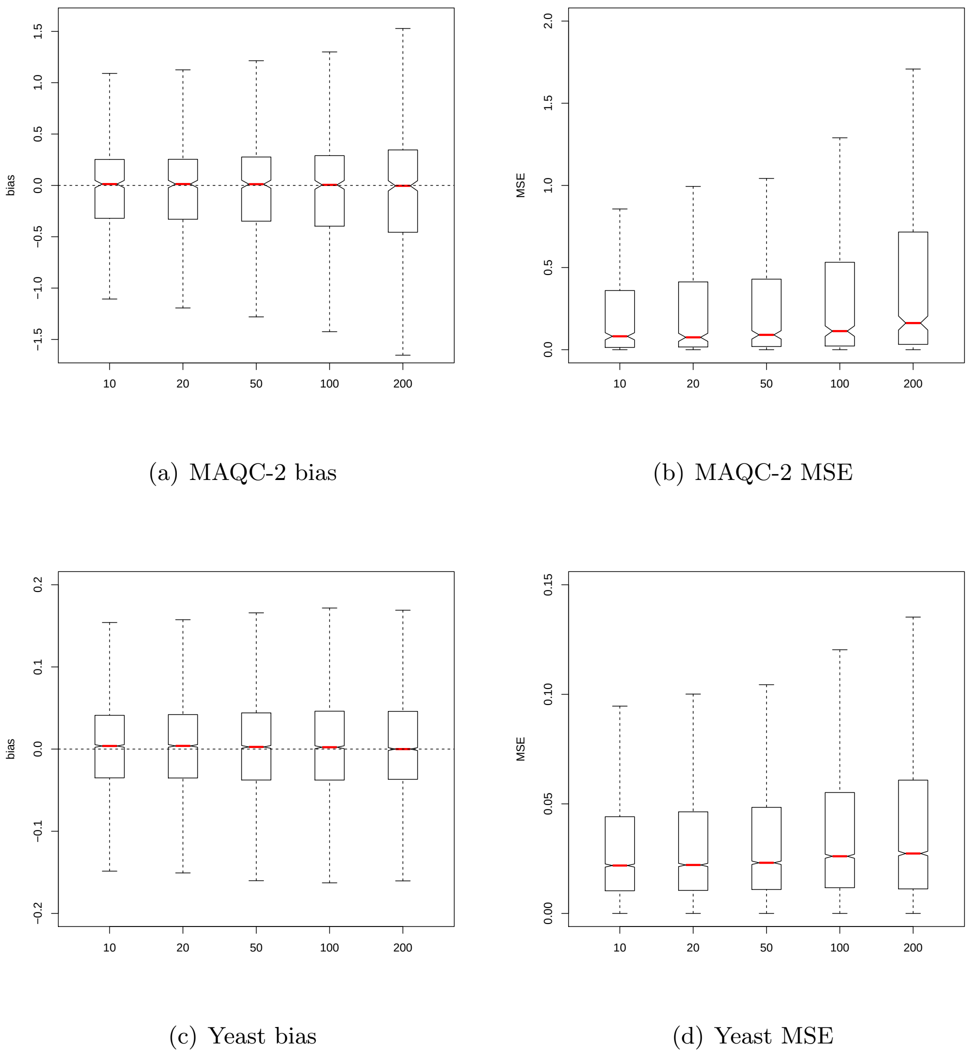
(e) CQN

Figure A.8: *Yeast dataset: p-values vs. GC-content.* Stratified boxplots of unadjusted p-values ($\log_{10}$) for the negative binomial LRT of growth condition effects vs. GC-content, for different normalization procedures. As in Figure 4.4, for every procedure but the full-quantile, the higher the GC-content, the more significant the evidence in favor of DE. The dashed line corresponds to a nominal unadjusted p-value of 10^{-5}.



(a) Gene length (b) Gene mappability

Figure A.9: *Yeast dataset: Log-fold-change vs. length and mappability.* Stratified boxplots of count log-ratios, after FQ between-lane normalization, for two biological replicate YPD lanes (Y1 lane vs. Y2 lane from flow-cell 428R1). Panel (a): Stratified by gene length. Panel (b): Stratified by mappability. Mappability for a given gene is defined as the ratio between the number of mappable bases and the total number of bases. A base is said to be mappable if the sequence of 36 bases starting at that position is unique along the genome. No dependence is observed for the log-fold-change on either length or mappability.



(a) MAQC-2 bias

(b) MAQC-2 MSE

(c) Yeast bias

(d) Yeast MSE

Figure A.10: *Bias and MSE vs. number of GC-content bins in normalization.* Boxplots of bias and MSE for MAQC-2 and Yeast datasets after full-quantile within-lane GC-content normalization with different numbers of GC-content bins. Panel (a): Estimated bias for MAQC-2 dataset, defined as difference between estimated log-fold-changes from RNA-Seq and qRT-PCR. Panel (b): Estimated MSE for MAQC-2 dataset, defined as squared difference between estimated log-fold-changes from RNA-Seq and qRT-PCR. Panel (c): Estimated bias for Yeast YPD dataset, defined as average estimated log-fold-change across the 35 pseudo-datasets. Panel (d): Estimated MSE for Yeast YPD dataset, defined as average of squared estimated log-fold-change across the 35 pseudo-datasets. The FQ normalization procedure appears to be robust to the number of bins.

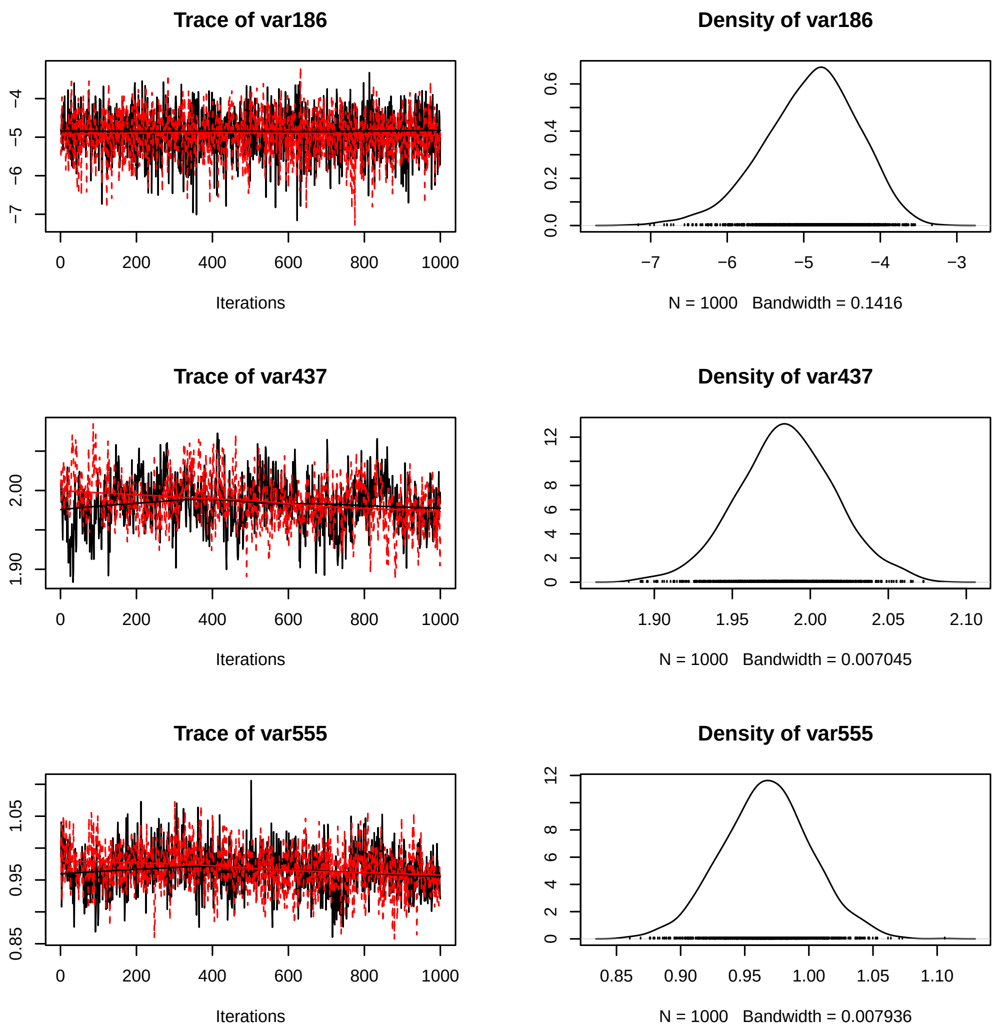


Figure A.11: *MAQC-2 dataset: convergence of the MCMC.* Trace of the two MCMCs and posterior density of $\beta_{1,j}$ for three randomly chosen genes.

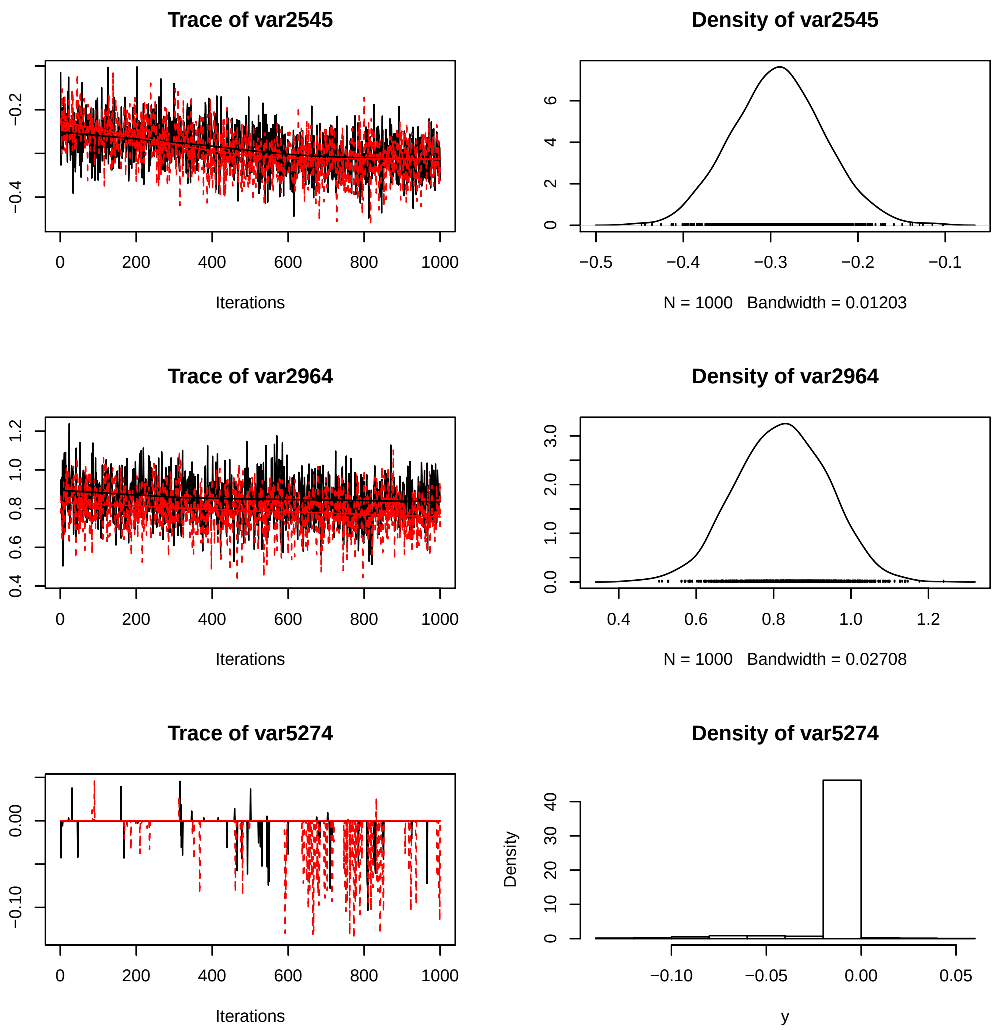


Figure A.12: *Yeast dataset: convergence of the MCMC.* Trace of the two MCMCs and posterior density of $\beta_{1,j}$ for three randomly chosen genes.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*. New York: Garland Science, 5th edition.

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**(3/4), 246–254.

Anscombe, F. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**(3/4), 358–382.

Azzalini, A. and Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing*, **17**, 71–80.

Azzalini, A., Menardi, G., and Rosolin, T. (2011). *R package `pdfCluster`: Cluster analysis via nonparametric density estimation (version 0.1-13)*. Università di Padova, Italia.

Baldi, P. and Long, A. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**(6), 509–519.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, **6**, 1345–1382.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (2006). The Quickhull algorithm for convex hulls. *ACM Transactions of Mathematical Software*, **22**, 469–483.

Bartlett, M. (1936). The square root transformation in analysis of variance. *Supplement to the Journal of the Royal Statistical Society*, **3**(1), 68–78.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Benjamini, Y. and Speed, T. (2011). Estimation and correction for GC-content bias in high throughput sequencing. Technical Report 804, Department of Statistics, University of California, Berkeley.

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**(2), 268.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, **107**(21), 9546.

Bullard, J. (2009). *Statistical Methods and Software for High-throughput Gene Expression Experiments*. Ph.D. thesis, University of California, Berkeley.

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010a). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**(1), 94.

Bullard, J., Mostovoy, Y., Dudoit, S., and Brem, R. (2010b). Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proceedings of the National Academy of Sciences*, **107**(11), 5058.

Bulmer, M. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, **30**(1), 101–110.

Cheng, Y. and Church, G. (2000). Biclustering of gene expression data. In *Proceedings of ISMB*, volume 8, pages 93–103.

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

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–563.

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer.

De Bin, R. and Risso, D. (2011). A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, **12**(1), 49.

Dohm, J., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**(16), e105.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**, 881–890.

Gentleman, R. C., Carey, V. J., Bates, D. M., *et al.* (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Good, I. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, **87**(419), 597–606.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hansen, K., Brenner, S., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, **38**(12), e131.

Hansen, K., Irizarry, R., and Wu, Z. (2011). Removing technical variability in RNA-Seq data using conditional quantile normalization. Technical Report 227, Department of Biostatistics, Johns Hopkins University.

Hardcastle, T. and Kelly, K. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**(1), 422.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, **84**, 502–516.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, **6**(12), e28210.

Ibrahim, J., Chen, M., and Gray, R. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, **97**(457), 88–99.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, **21**(9), 1543–1551.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**, 682–693.

Kempton, R. and Taylor, L. (1974). Log-series and log-normal parameters as diversity discriminants for the lepidoptera. *The Journal of Animal Ecology*, **43**(2), 381–399.

Kendziorski, C., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**, 3899–3914.

Kerr, G., Ruskin, H., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, **38**, 283–293.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**(3), R25.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, **15**(3), 209–225.

Li, J. and Zha, H. (2006). Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis*, **50**, 163–180.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, **8**, 1687–1723.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, **12**(1), 31–46.

Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics*, **1**(1), 24–45.

Maniar, J. M. and Fire, A. Z. (2011). EGO-1, a *C. elegans* RdRP, modulates gene expression via production of mRNA-templated short antisense RNAs. *Current Biology*, **21**(6), 449–459.

Marioni, J., Mason, C., Mane, S., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509.

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, **11**, 663.

McCarthy, D. and Smyth, G. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, **25**(6), 765.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Menardi, G. and Torelli, N. (2010). Preserving the clustering structure by a projection pursuit approach. In F. Palumbo, C. N. Lauro, and M. J. Greenacre, editors, *Data Analysis and Classification*, pages 171–178. Springer.

Morgan, M. and Pagès, H. (2010). *Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import*. R package version 1.5.79.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., and Gentleman, R. (2009). Shortread: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**(7), 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**(5881), 1344.

Oshlack, A. and Wakefield, M. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, **4**(1), 14.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*, **37**(18), e123.

Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., and Pritchard, J. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**(7289), 768–772.

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March*, pages 20–22.

Risso, D. and Dudoit, S. (2011). *EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq*. R package version 1.0.0 `http://www.bioconductor.org/packages/devel/bioc/html/EDASeq.html`.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**(3), R22.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3), R25.

Robinson, M. and Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881.

Robinson, M. and Smyth, G. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**(2), 321.

Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **71**(2), 319–392.

Saccharomyces Genome Database (r64). http://www.yeastgenome.org.

Shi, L., Reid, L., Jones, W., *et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, **24**(9), 1151–1161.

Slonim, D. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature genetics*, **32**, 502–508.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, **3**(1), 3.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Resarch*. Advanced Online Publication.

Teytelman, L., Özaydın, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., and Eisen, M. (2009). Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One*, **4**(8), e6700.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567–6572.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**(5), 511–515.

Tritchler, D., Parkhomenko, E., and Beyene, J. (2009). Filtering genes for cluster and network analysis. *BMC Bioinformatics*, **10**, 193.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**(1), 136.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1), 57–63.

Wu, Z., Jenkins, B., Rynearson, T., Dyhrman, S., Saito, M., Mercier, M., and Whitney, L. (2010). Empirical Bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC Bioinformatics*, **11**, 564.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4), e15.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, **19**(9), 1586.

Young, M., Wakefield, M., Smyth, G., and Oshlack, A. (2010). Gene Ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, **11**(2), R14.

Zheng, W., Chung, L., and Zhao, H. (2011). Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics*, **12**(1), 290.

Davide Risso

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: davide@stat.unipd.it

Current Position

Since January 2009; (expected completion: January 2012)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Simultaneous inference for RNA-Seq data
Supervisor: Prof. Monica Chiogna
Co-supervisors: Prof. Sandrine Dudoit, Dr. Chiara Romualdi

Research Interests

Statistical Modeling of high-throughput data (such as microarray data, deep sequencing).
Development of novel methods for RNA-Seq data analysis.
Normalization, meta-analysis and pathway analysis of gene expression data.
Computationally intensive statistical methods.

Education

October 2005 – March 2008

Master (*laurea specialistica/magistrale*) degree in "Statistics and Informatics"

.

University of Padua, Faculty of Statistical Sciences

Title of dissertation: "Analysis of gene expression data: a comparative study on the impact of normalization on statistical inference"

Supervisor: Dr. Chiara Romualdi

Final mark: 110 (out of 110) with honors.

October 2002 – October 2005

Bachelor degree (*laurea triennale*) in Statistics, Mathematics, and Data Management.

University of Genoa, Faculty of Mathematical, Physical and Natural Sciences

Title of dissertation: "Statistical modeling of fever episodes in neutropenic pediatric patients with cancer"

Supervisor: Prof. Vincenzo Fontana

Final mark: 110 (out of 110) with honors.

Visiting periods

August – December 2010

Graduate Group in Biostatistics,

University of California, Berkeley, USA.

Supervisor: Prof. Sandrine Dudoit.

March – September 2011

Graduate Group in Biostatistics,

University of California, Berkeley, USA.

Supervisor: Prof. Sandrine Dudoit.

Further education

June 2010

Lipari School on BioInformatics and Computational Biology
Organizing Institution: University of Catania
School Directors: Prof. Alfredo Ferro (University of Catania), Prof. Raffaele Giancarlo (University of Palermo), Prof. Concettina Guerra (University of Padova and Georgia Tech.), Prof. Michael Levitt (Stanford University)
Instructors: Dr. Silvio Bicciato (University of Modena and Reggio Emilia), Prof. Charles Lawrence (Brown University), Prof. Dana Pe'er (Columbia University), Prof. Catarina Sismeiro (Imperial College), Prof. Mary Ellen Bock (Purdue University).

Awards and Scholarships

June 2011

BioConductor Annual Conference 2011 Scholarship, Bioconductor Foundation. Registration and housing support for the BioC2011 Conference.

January 2009

PhD Scholarship, University of Padua, Italy.

May – December 2008

Graduate Fellowship, Department of Chemical Process Engineering, University of Padua, Italy.
Project: "Analysis and integration of gene expression data for studying muscle plasticity" supervised by Dr. Silvio Bicciato, in collaboration with Dr. Chiara Romualdi.

Computer Skills

Operating Systems: Unix/Linux, MAC OS X, MS Windows.
Programming languages: C, shell script.
Statistical packages: S-plus/R, SAS, Stata, Matlab.
Other: SQL.

Language Skills

Italian: native; English: fluent.

Publications

Articles in journals

Risso D., Schwartz K., Sherlock G., Dudoit S. (2011). GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, **12**:480.

Martini P., **Risso D.**, Sales G., Romualdi C., Lanfranchi G., Cagnin S. (2011). Statistical Test of Expression Pattern (STEPath): a new strategy to integrate gene expression data with genomic information in individual and meta-analysis studies. *BMC Bioinformatics*, **12**(1):92. *Highly accessed.*

De Bin R., **Risso D.** (2011). A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, **12**:49. *Highly accessed.*

Risso D., Massa M.S., Chiogna M., Romualdi C. (2009). A modified LOESS normalization applied to microRNA arrays: a comparative evaluation. *Bioinformatics*, **25** (20):2685-2691.

Bisognin A., Coppe A., Ferrari F., **Risso D.**, Romualdi C., Bicciato S., Bortoluzzi S. (2009). A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bionformatics*, **10**:201. *Highly accessed.*

Chiogna M., Massa M.S., **Risso D.**, Romualdi C. (2009). A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bionformatics*, **10**:61. *Highly accessed.*

Software

Risso D., Dudoit S. (2011). *EDASeq*: Exploratory Data Analysis and Normalization for RNA-Seq. *R/Bioconductor package.* Version 1.0.0. `http://www.bioconductor.org/packages/release/bioc/html/EDASeq.html`

Conference presentations

Risso D., Sales G., Romualdi C., Chiogna M. (2011). A hierarchical Bayesian model for RNA-Seq data. (contributed talk) *7th Conference on Statistical Computation and Complex Systems – SCo2011*. Padua, Italy, September 19–21, 2011.

Risso D., Dudoit S. (2011). EDASeq: Exploratory Data Analysis and Quality Control for RNA-Seq data. (contributed talk) *BioC 2011*. Seattle, WA, USA, July 28–29, 2011.

Risso D., De Bin R. (2010). Clustering via nonparametric density estimation: an application to microarray data. (poster) *29th Leeds Annual Statistical Research Workshop*. Leeds, UK, July 6–8, 2010.

Risso D., Romualdi C. (2009). Integration methods for microarray data. (poster) *6th Annual Meeting of the Bioinformatics Italian Society*. Genoa, Italy, March 18–20, 2009.

Risso D., Massa MS., Chiogna M., Romualdi C. (2009). A modified LOESS normalization applied to miRNA arrays: a comparative evaluation (poster). *8th European Conference on Computiational Biology*. Stockholm, Sweden, June 27 – July 2, 2009.

Risso D., Chiogna M., Massa MS., Romualdi C. (2008). A comparison on effects of normalizations in the detection of differentially expressed genes (poster). *7th European Conference on Computiational Biology*. Cagliari, Italy, September 22–26, 2008.

Teaching experience

October 2011 – January 2012
Course name: Applied Statistics
Degree: Master degree in Healthcare Biology
Teaching task: computer labs, 25 hours
Institution: University of Padua
Instructor: Prof. Guido Masarotto

2006 – 2009
Junior Tutor at the Faculty of Statistical Sciences
University of Padua.

Other Interests

Since June 2010
Reviewer for *BMC Bioinformatics* and *Computational Statistics and Data Analysis*

References

Monica Chiogna
Associate Professor
Department of Statistical Sciences
University of Padua
via C. Battisti 241, 35121 Padova, Italy
Phone: +39 049 827 4183
e-mail: `monica@stat.unipd.it`

Sandrine Dudoit
Professor
Division of Biostatistics and
Department of Statistics
University of California, Berkeley
101 Haviland Hall, #7358 Berkeley, CA
94720-7358, U.S.A.
Phone: +1 (510) 643-1108
e-mail: `sandrine@stat.berkeley.edu`

Chiara Romualdi
Assistant Professor
Department of Biology
University of Padua
via U. Bassi 58/B, 35121 Padova, Italy
Phone: +39 049 827 7401
e-mail: `chiara.romualdi@unipd.it`

Gavin Sherlock
Associate Professor
Department of Genetics
School of Medicine
Stanford University
300 Pasteur Drive, Stanford, CA
94305-5120, U.S.A.
Phone: +1 (650) 498-6012
e-mail: `gsherloc@stanford.edu`