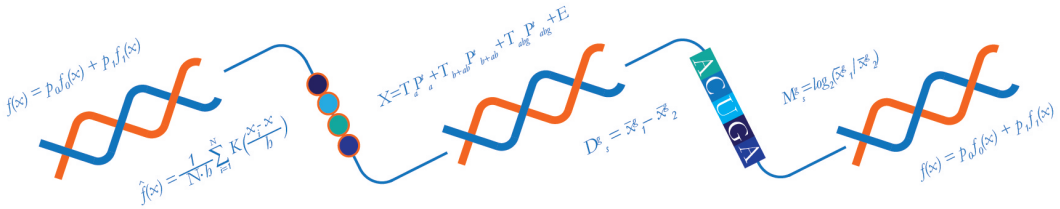
Statistical methods for transcriptomics: From microarrays to RNA-seq

Sonia Tarazona Campos

PhD Thesis

November 2014



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Supervisors:
Dr. Ana Conesa Cegarra
Dr. Alberto J. Ferrer Riquelme

Statistical methods for transcriptomics: From microarrays to RNA-seq

Sonia Tarazona Campos

Supervisors:
Dr. Ana Conesa Cegarra
Dr. Alberto J. Ferrer Riquelme

A doctoral thesis submitted to

Department of Applied Statistics, Operations Research and Quality

November 2014

Graphic design: Charo Sanchis Font

Abstract

One of the most common types of analysis in genome research is the comparison of gene expression profiles (or transcriptomics) to understand the relationship between genes (or genotype) and the phenotype. Transcriptome analysis has been traditionally conducted using microarrays and with increasing frequency since 2008 by RNA sequencing (RNA-seq). A fundamental goal in these types of studies is to identify the genes whose expression changes between different conditions, in other words, to select the most relevant variables (genes) in terms of inter-condition variability. The variable selection problem, usually known in transcriptomics as "differential expression analysis", can be addressed from the univariate or multivariate point of view, but must always take the complexity of the experimental design into account. One of the challenges that biostatisticians face when tackling this problem is the so-called "curse of dimensionality": hundreds or even thousands of variables have to be analyzed with few observations usually available. Therefore, it is essential to provide researchers with efficient statistical tools to perform this task.

A typical approach to variable selection is to test the null hypothesis of equality of average expression levels between two conditions in a gene-wise fashion and to do this repeatedly for all genes in the transcriptomics data set. However, transcriptomics may also involve multifactorial experimental designs (eg multiple treatments, several developmental states, time series...). The first part of this thesis is dedicated to the variable selection problem

in multifactorial designs when multivariate methods are used to model microarray gene expression profiles. In particular, we chose the ASCA-genes multivariate technique as a starting point to propose some strategies to select differentially expressed genes in the multivariate context, that were tested under different biological scenarios.

RNA-seq technology emerged when work for this thesis was first started and now it is widely applied in transcriptomics. RNA-seq data is fundamentally different from microarray data and this has motivated the development of new statistical methods to study differential expression in this technology. Thus, the second part of the thesis is entirely focused on RNA-seq experiments. First, we develop a set of procedures to assess the quality of RNA-seq measurements, to identify the potential biases of the technology and to process the data to reduce the impact of technical noise on statistical results. Secondly, we address the variable selection problem for the two-class comparison case. Given that some controversy exists on the theoretical distribution followed by RNA-seq data, we opted to investigate non-parametric data-driven techniques to overcome the limitations of parametric assumptions and propose a strategy that is efficient in controlling the false positive rate. Two methodologies, NOISeq for technical and NOISeqBIO for biological replicates, were developed and compared to the state of the art methods.

Resumen

Uno de los análisis más comunes en investigación genómica es la comparación de perfiles de expresión génica (o transcriptómica) para entender la relación entre los genes (o genotipo) y el fenotipo. El análisis del transcriptoma se ha llevado a cabo tradicionalmente utilizando *microarrays* y cada vez con mayor frecuencia desde 2008 mediante secuenciación del ARN (RNA-seq). Un objetivo fundamental en este tipo de estudios es identificar aquellos genes cuya expresión cambia entre condiciones, en otras palabras, seleccionar las variables más relevantes (genes) en términos de variabilidad entre-condiciones. El problema de selección de variables, normalmente conocido en transcriptómica como "análisis de expresión diferencial", se puede abordar desde el punto de vista univariante o multivariante, siempre teniendo en cuenta la complejidad del diseño experimental. Uno de los retos a los que los bioestadísticos se enfrentan al tratar de resolver este problema es la llamada "maldición de la dimensión": se tienen que analizar cientos o incluso miles de variables con muy pocas observaciones disponibles normalmente. Por tanto, es esencial proporcionar a los investigadores herramientas estadísticas eficientes para llevar a cabo esta tarea.

Un enfoque típico en selección de variables es contrastar la hipótesis nula de igualdad del nivel de expresión medio entre dos condiciones para un gen particular y hacerlo repetidamente para todos los genes del conjunto de datos transcriptómicos. Sin embargo, la transcriptómica puede conllevar también diseños experimentales

multifactoriales (múltiples tratamientos, varios estados de desarrollo, series temporales...). La primera parte de esta tesis se ha dedicado al problema de selección de variables en diseños multifactoriales cuando se usan métodos multivariantes para modelizar los perfiles de expresión génica en *microarrays*. En particular, elegimos como punto de partida la técnica multivariante ASCA-genes para proponer algunas estrategias de selección de genes diferencialmente expresados en un contexto multivariante, que fueron evaluadas bajo distintos escenarios biológicos.

La tecnología RNA-seq apareció al comienzo de esta tesis y ahora se aplica ampliamente en transcriptómica. Los datos de RNA-seq son en esencia diferentes a los datos de *microarrays* y esto ha motivado el desarrollo de nuevos métodos estadísticos para estudiar la expresión diferencial en esta tecnología. Por tanto, la segunda parte de la tesis se centra exclusivamente en experimentos de RNA-seq. Primero, desarrollamos una colección de procedimientos para determinar la calidad de las medidas de RNA-seq, identificar los sesgos potenciales de la tecnología y procesar los datos para reducir el impacto del ruido técnico en los resultados estadísticos. En segudo lugar, abordamos el problema de selección de variables para el caso de comparación de dos grupos. Dado que existe cierta controversia en la distribución teórica que siguen los datos de RNA-seq, optamos por investigar técnicas no paramétricas dirigidas por los datos para vencer las limitaciones de las hipótesis paramétricas y propusimos una estrategia que es eficiente a la hora de controlar la tasa de falsos positivos. Se desarrollaron dos metodologías, NOISeq para réplicas técnicas y NOISeqBIO para réplicas biológicas, y se compararon con los métodos más punteros.

Resum

Una de les anàlisis més comunes en investigació genòmica és la comparació de perfils d'expressió gènica (o transcriptòmica) per entendre la relació entre els gens (o genotip) i el fenotip. L'anàlisi del transcriptoma s'ha dut a terme tradicionalment utilitzant *microarrays* i cada vegada amb més freqüència des de 2008 mitjançant la seqüenciació de l'ARN (RNA-seq). Un objectiu fonamental en aquest tipus d'estudis és identificar aquells gens l'expressió dels quals canvia entre condicions, en altres paraules, seleccionar les variables més rellevants (gens) en termes de variabilitat entre-condicions. El problema de selecció de variables, normalment conegut en transcriptòmica com a "anàlisi d'expressió diferencial", es pot abordar des del punt de vista univariant o multivariant, sempre tenint en compte la complexitat del disseny experimental. Un dels reptes a què s'enfronten els bioestadístics en tractar de resoldre aquest problema és l'anomenada "maldició de la dimensió": s'han d'analitzar centenars o milers de variables amb molt poques observacions disponibles normalment. Per tant, és essencial proporcionar als investigadors ferramentes estadístiques eficients per a dur a terme aquesta tasca.

Un enfocament típic en selecció de variables és contrastar la hipòtesi nul·la d'igualtat de nivells d'expressió mitjans entre dues condicions per a un gen particular i fer-ho repetidament per a tots els gens del conjunt de dades transcriptòmiques. No obstant això, la transcriptòmica pot comportar també dissenys experimentals

multifactorials (múltiples tractaments, diversos estats de desenvolupament, sèries temporals...). La primera part d'aquesta tesi s'ha dedicat al problema de selecció de variables en dissenys multifactorials quan s'usen mètodes multivariants per a modelitzar els perfils d'expressió gènica en *microarrays*. En particular, triàrem com a punt de partida la tècnica multivariant ASCA-genes per a proposar algunes estratègies de selecció de gens diferencialment expressats en un context multivariant, que van ser avaluades sota diferents escenaris biològics.

La tecnologia RNA-seq va aparèixer al començament d'aquesta tesi i ara s'aplica àmpliament en transcriptòmica. Les dades de RNA-seq són en essència diferents a les dades de *microarrays* i açò ha motivat el desenvolupament de nous mètodes estadístics per a estudiar l'expressió diferencial en aquesta tecnologia. Així doncs, la segona part de la tesi se centra exclusivament en experiments de RNA-seq. En primer lloc, vam desenvolupar una col·lecció de procediments per a determinar la qualitat de les mesures de RNA-seq, identificar biaixos potencials de la tecnologia i processar les dades per a reduir l'impacte del soroll tècnic en els resultats estadístics. En segon lloc, es va abordar el problema de selecció de variables per al cas de comparació de dos grups. Donat que existeix certa controvèrsia en la distribució teòrica que segueixen les dades de RNA-seq, vam optar per investigar tècniques no paramètriques dirigides per les dades per a véncer les limitacions de les hipòtesis paramètriques i vam proposar una estratègia que és eficient a l'hora de controlar la tasa de falsos positius. Es desenvoluparen dues metodologies, NOISeq per a rèpliques tècniques i NOISeqBIO per a rèpliques biològiques, i es compararen amb els mètodes més capdavanters.

A mis padres

A Enric

Agradecimientos

Esta tesis ha sido para mí como una carrera de fondo. Ha habido momentos de agotamiento en los que piensas que no superarás el reto. Por suerte, he podido contar con mis compañeros, familia y amigos que me han ayudado a seguir avanzando. Gracias a vosotros, la experiencia ha resultado gratificante. Y al final del camino me ha gustado descubrir que lo más importante no es haber llegado a la meta, sino que has aprendido a correr.

El mérito de este trabajo lo quiero compartir, en primer lugar, con mis directores Ana y Alberto. Gracias por vuestro tiempo, por vuestra ayuda y por la confianza que siempre habéis depositado en mí. A Ana, con quien he tenido el placer de compartir mi día a día, quiero dedicarle un agradecimiento muy especial porque me siento muy afortunada de haber podido trabajar a su lado. Gracias por contar conmigo desde el primer momento, cuando aterricé en el departamento con mis escasos conocimientos de biología, y por haberme enseñado casi todo lo que sé de ciencia en general y de bioinformática en particular. Gracias por todas las oportunidades que me has dado y por tu forma tan entusiasta de "vender" mi trabajo. Ha sido un lujo tenerte como jefa y espero que esta relación de complicidad se mantenga en la distancia.

Durante este tiempo en el CIPF he tenido la gran suerte de contar con mis queridos bioinfos, tanto los que siguen ahí como los que se fueron por esos mundos. Gracias por echarme una mano siempre que lo he necesitado y por hacer más divertida la vida en el labo. En especial, gracias a mi grupo que siempre me anima tanto y

me aguanta en los momentos de agobio. No sé qué haría sin vosotros.

A las chicas japo, con las que además de compartir sushi he compartido risas y llantos, quiero agradecerles sus ganas de disfrutar de la vida y que estén siempre ahí cuando las necesitas.

También agradecer a Charo y Patri el diseño de esta tesis y los buenos ratos que paso con ellas. Y a Susana su ayuda con las clases y sus buenos consejos.

I am very grateful to my colleagues at the Karolinska Institutet in Stockholm because they made me feel like at home.

Fuera del terreno profesional, me gustaría agradecer la comprensión de mis amigos y familia, cuando a causa del trabajo no he podido dedicarles todo el tiempo que se merecen.

Son muchos los amigos a los que me gustaría nombrar aquí pero no es posible. Quiero dar las gracias a todos y, en particular, a Judit por todos estos años de amistad incondicional y por los que vendrán, a Nelo por descubrirme tantas cosas, y a Cris por su fortaleza en los momentos difíciles, que tanto admiro.

Y por supuesto, un reconocimiento muy especial a toda mi familia. A mis padres, los mejores del mundo, que se han esforzado tanto para que yo llegara hasta aquí. A mis hermanos, que son increíbles y no me fallan nunca. A Rocío y Pili, que saben comprendernos. Y a Laura y Martí, que me sorprenden cada día y lo llenan todo de alegría.

Gràcies també a la família alcoiana, Horte i Miquel, Pilar i Fernando, per estar tan orgullosos de mi.

I per últim, vull agrair a Enric la seua dedicació i comprensió. Gràcies per fer-me sempre costat. Sense tu açò no hauria estat possible.

Contents

This work was funded by the following grants:

From 2009 to 2012. *PathoGenoMics: Transcriptional networks controlling virulence in filamentous fungal pathogens* (TRANSPAT). MICINN, Bio2008-04638-E, ERA-NET 2009.

2012. *Genomics and transcriptomics of detoxification pathways in Drosophila*. MICINN, PIB2010AR-00266.

From 2012 to 2015. *STATegra: User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, Proteomics and Metabolomics data*. European Comission (FP7), Project number 306000.

Chapter 1

Introduction

1.1 What is transcriptomics about?

The cell is the basic structural, functional and biological unit of a living organism. There are two types of cells, eukaryotic cells, which contain a nucleus, and prokaryotic cells, which do not. Prokaryotic cells are usually single-celled organisms, while eukaryotic cells can be either single-celled or part of multicellular organisms. Plants, animals, fungi, slime moulds, protozoa, and algae are all eukaryotic.

The three main components of eukaryotic cells are the membrane, cytoplasm, and nucleus (Figure 1.1). The membrane is a lipid structure that preserves the integrity of the cell and selectively regulates the flux of nutrients and proteins. The nucleus contains the chromosomes, which are made up of DNA (deoxyribonucleic acid). The genes, which are the hereditary units of biological organisms, are localized along the DNA chains. In the cytoplasm, there are many different molecules such as RNA (ribonucleic acid), proteins, carbohydrates, etc. and also organelles that are responsible for various cellular functions. RNA molecules are also known as transcripts and contain information transcribed from DNA, which is translated into proteins by ribosomes. Proteins are constituted by amino acid sequences and are the basic functional elements of cellular physiology and metabolism.

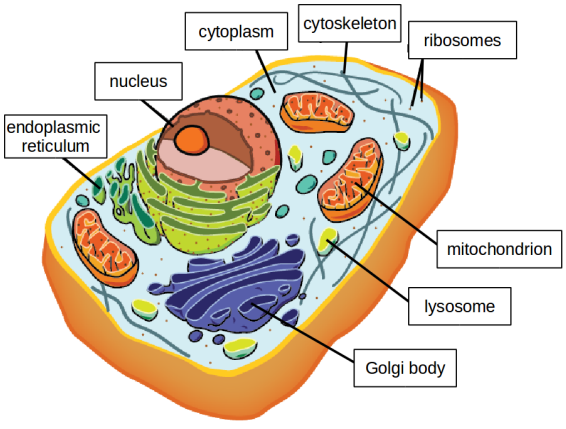


Figure 1.1: Diagram of a eukaryotic cell

DNA molecules (where genes are located) form a double helix: each DNA chain is a linear sequence of nucleotides that are composed of one of the four following bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The genetic information encoded by this sequence of bases determines the order of amino acids in the protein and three DNA bases (a codon) encode one amino acid. However, only a small part of the whole genome (3-8% in humans) is considered "functional", the part that encodes proteins or functional RNA [32]. This means that most of the DNA is "non-functional" and may have other structural or regulatory functions. For many years, it was believed that each gene coded for a single protein because there are genetic diseases in which a single gene mutation causes disease by disrupting the protein it encodes. However, it is now known that a gene can encode more than one protein through the process of "alternative splicing" and that a protein can be encoded by more than one gene. Moreover, proteins may undergo post-translational modifications and therefore, in higher eukaryots one single gene serves as the basis for many possible versions of a particular protein.

The genetic code is the set of rules that defines how the information contained in the sequence of bases in the genes is transfered to the amino acids that form the proteins. But the fact that DNA is in the cell nucleus and that proteins are synthesized by ribosomes located in the cytoplasm implies the existence of a mechanism to transfer this information from the nucleus to the cytoplasm. This process is called transcription (Figure 1.2), the synthesis of DNA-derived intermediary molecules in the nucleus: RNA transcripts. RNA has a similar structure to DNA, but is single chained rather than a double helix and one of the four bases, thymine (T) is replaced by uracil (U). RNA molecules are transported to the cytoplasm and the information they contain (determined by the A, U, C and G base sequence) is translated to a specific sequence of amino acids that constitutes a protein. Alternatively, RNAs may not be translated into proteins and instead perform their cellular function as "non-coding RNA" molecules which have different regulatory and processing roles.

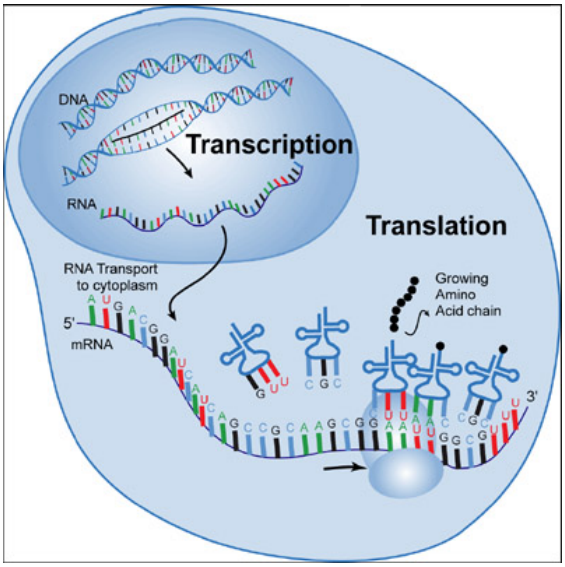


Figure 1.2: Transcription and translation

Thus, the process of producing a biologically functional molecule of either RNA or protein from DNA is called gene expression and this process consists of two basic steps: transcription and translation. During transcription, a gene (DNA) is copied into a transcript (RNA). In the case of protein-coding genes, the translation allows for the RNA to generate a protein; non protein-coding genes go through transcription but are not translated into a protein. The number of transcripts synthesized from a certain gene in a given cellular context is known as the expression level of that gene. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome varies through development and between different tissue types in response to external environmental conditions. The composition of the transcriptome reflects which genes are being actively expressed at any given time.

Therefore transcriptomic analyses essentially compare gene expression profiles between biological samples or experimental conditions to identify differences that could help to infer the function of the genes or understand the ongoing biological processes. For instance, studying the genes that are differentially activated between healthy and diseased people could help to find

candidate genes that may be responsible for causing the disease, and which could therefore become therapeutic targets.

There are many methods for measuring gene expression levels: transcriptomics refers to the situation where all genes are observed simultaneously so high-throughput techniques are needed to estimate the expression of thousands of genes in order to obtain a global picture of cellular activity. The most widely used techniques in this field are DNA microarray technology and RNA-seq which are both described in detail in the next section.

1.2 Measuring gene expression

The cell transcriptome is dynamic: as opposed to the static genome, the transcriptome continually changes. Expression profiling experiments often involve measuring the relative amount of transcripts arising from each gene in two or more experimental conditions. Altered gene expression levels suggest a change in the level of the protein encoded by the gene and may indicate the cause of a disease or the response to perturbations, clinical treatments, or environmental conditions. Thus, gene expression changes may provide important clues to understanding the biological mechanisms underlying differences between phenotypes, i.e. the differences in the observable characteristics or traits of an organism.

Some techniques for measuring gene expression include: microarrays, real-time polymerase chain reaction (RT-PCR), serial analysis of gene expression (SAGE), and RNA sequencing (RNA-seq). In this section, we describe the most widely used high-throughput techniques for measuring gene expression: microarrays and RNA-seq which were used to generate the data analyzed in this work.

1.2.1 Microarrays

DNA microarrays (also called chips or arrays) are used to measure the expression level of a large number of genes simultaneously. The most common

Figure 1.3: Example of a microarray.

types of microarrays for transcriptomics are cDNA arrays and oligonucleotide arrays that are used in combination with one or two dye-labeling strategies.

A microarray is a solid surface which contains a collection of microscopic DNA spots representing short sections of genes (Figure 1.3); these spots are usually known as probes, and each one corresponds to a different gene. In the case of one-color microarrays (one dye-labeling), the RNA of the biological sample being studied is labeled with a fluorescent dye and hybridized to these target probes. Hybridization occurs between complementary nucleic acid sequences because of their propensity to specifically pair with each other by forming hydrogen bonds. Fluorescently labeled target sequences that bind to a probe generate a signal. The intensity of that signal is measured with a scanner and the intensity of each probe is registered. These intensity values are the expression level estimations of the target genes, once they have been corrected for the background noise or other technical effects: the higher the intensity of a given probe the higher the amount of RNA (transcripts) in the biological sample and, therefore, the higher the expression level of that particular gene.

Transcriptomic experiments usually consist of studying several biological samples from different experimental conditions, which requires the processing of several microarrays (one per sample). However, the measurements obtained must be comparable for all samples: the procedure for transforming the data to make meaningful biological comparisons is called normalization

[115]. The gene expression values produced by microarray technologies are continuous measurements and it is generally accepted that, albeit with appropriate transformations on occasions, they follow a Gaussian distribution, and many statistical methods have been adapted or developed to model this kind of data based on this assumption [33, 133].

The advantage of DNA expression microarrays is that they can estimate the relative activity of thousands of genes. However, the target genes must have been previously characterized and it is not possible to identify or measure genes that have not yet been discovered.

1.2.2 RNA-seq

Massive parallel DNA sequencing[1] is termed as next generation sequencing (NGS) or second-generation sequencing, and encompasses several technologies such as DNA-seq, RNA-seq, ChIP-seq, DNase-seq, Methyl-seq, among others. Some of these technologies have been commercially available since 2005. Sequencing platforms can sequence millions of short reads (50-400 bases each) per instrument run and thus genome-wide studies are possible using this technology. The advent of NGS technologies has created unprecedented possibilities for the characterization of genomes and has significantly advanced our understanding of its organization. They can now be used to tackle the de novo sequencing of large genomes [9, 87, 146], report individual genome differences within the same species (DNA-seq or resequencing) [1], characterize the interaction spectrum of DNA-binding proteins (ChIP-seq) [110], map chromatin structure (DNase-seq, Hi-C), and to create genome-wide epigenetic modification profiles (Methyl-seq) [84].

One of the most ground-breaking applications of NGS techniques is the deciphering of the complexity of the transcriptome by means of RNA sequencing technology (RNA-seq). In the last few years the use of RNA-seq has resulted in an incredible amount of new data that has dissected gene

[1] DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule.

isoforms[1], allelic expression[2], and extended 3' UTR regions[3], as well as revealing novel splice junctions, modes of antisense regulation, and intragenic expression [23, 52, 100, 144]. In addition, RNA-seq has made sequence-based expression analysis an increasingly popular alternative to microarrays because any active gene can be measured, as opposed to the pre-characterized set of known genes which are recognized by microarrays. Other advantages of RNA-seq have also been claimed, such as having a wider dynamic range of expression measurements and a lower technical variability.

As previously described [107], in most RNA-seq experiments a sample of purified RNA is taken, sheared, and converted to cDNA (as also done for cDNA arrays). However, in contrast to microarrays, this cDNA is sequenced on a high-throughput platform (e.g. Illumina, SOLiD or Roche 454). For instance, the sequencing chemistry in the Illumina system occurs in the flow-cell, which is a glass slide separated into lanes (usually 8) into which different biological samples can be deposited (Figure 1.4).

The sequencing process generates millions of short reads (25 to 300 bp) taken from one end of the cDNA fragments. When short reads from both ends of each cDNA fragment are generated, they are called "paired-end" reads. The platforms differ substantially in their chemistry and processing steps, but the raw data they produce always consist of a long list of short sequences with associated quality scores.

[1]Isoforms are transcripts produced from the same locus (specific spot on a chromosome), but their transcription start sites (TSSs), protein coding DNA sequences (CDSs), or untranslated regions (UTRs) can differ, potentially altering gene function.

[2]Most multicellular organisms have two sets of homologous chromosomes. An allele is the copy of a gene on each chromosome. The two alleles of a given gene are inherited, one from each parent.

[3]DNA strands have directionality, since double helices are necessarily directional (a strand running 5'-3' pairs with a complementary strand running 3'-5'). DNA replication occurs only in the 5'-3' direction. The 3'-UTR is the section of messenger RNA that immediately follows the translation termination codon, so this part of the gene is not translated into a protein. However, the 3'-UTR often contains post-transcriptional gene expression regulatory regions.

Figure 1.4: Illumina flow-cell.

In contrast to microarrays, where the raw data generated requires little pre-processing to obtain an expression level estimation, a more laborious procedure must be followed to obtain an expression estimation using the RNA-seq strategy (an example of a process pipeline is shown in Figure 1.5). First, sequencing reads are mapped onto the reference genome, i.e. their genomic location of origin must be identified. Some of the reads will not map to the reference genome because of potential sequencing errors, or differences in the sample and reference genome sequences, etc. Second, if the reference genome has previously been studied and characterized (i.e. annotated) and gene positions within each chromosome are known, the mapped reads from each of the biological samples corresponding to each annotated gene can be quantified. The number of sequencing reads mapped to a given gene is an estimation of the expression level of that gene [92]. Finally, as occurs when microarrays are used, these estimated expression levels have to be normalized in order to remove unwanted technical effects and to calibrate the observations (samples).

Consequently, the nature of RNA-seq data is very different to microarrays. Expression estimates from RNA-seq for a given gene are discrete measurements since they are defined as the number of reads mapping to that gene (read counts). Hence, statistical models developed for microarray analysis have had to be revised for use for RNA-seq analysis. It is not easy to find a good formula to transform these count data into a continuous distribution,

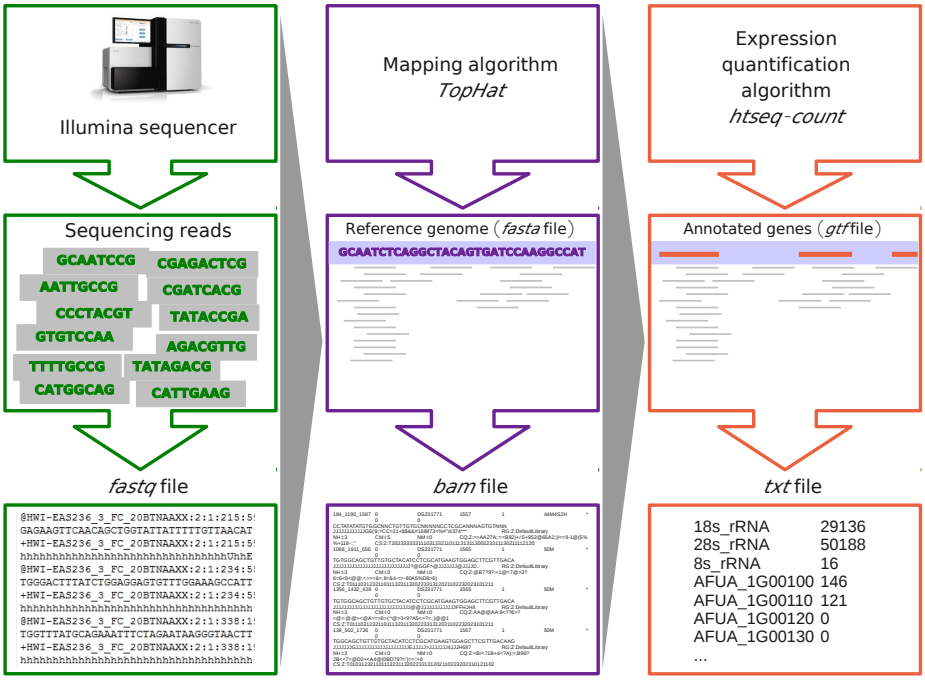


Figure 1.5: Example of an RNA-seq analysis pipeline.

especially in the lower count range and for small samples. Initial analyses using only technical replicates[1] assumed that read counts followed a Poisson distribution [18, 92]. However, soon it became clear that biological variability was not well described by this distribution as over-dispersion was observed among biological replicates. Thus, to take biological variability into account, a negative binomial distribution (which would be equivalent to a Poisson distribution when the mean and variance are equal) is generally accepted as a good option for modeling RNA-seq data [4, 122]. Even so, other specific statistical methods for this new technology following different approaches are still being developed.

[1] In RNA-seq, technical replicates are often considered to be the different sequencing runs performed on the same biological sample. For instance, in Illumina technology, each lane containing the same biological material might be taken as a technical replicate.

1.3 The role of statistics in transcriptomics

Transcriptomics and other omics disciplines that study biological systems from different perspectives all yield a huge amount of data to be stored, retrieved, organized and analyzed. Bioinformatics is an interdisciplinary field that deals with these topics. It covers tasks from the design or adaptation of statistical models and algorithms for analyzing high-throughput data to the implementation of efficient software tools to generate useful biological knowledge. Hence different areas such as computer science, mathematics, engineering and, of course, statistics converge and interact in bioinformatics in order to successfully address these problems. There are several omics data analysis challenges which are discussed in the following paragraphs:

First, the technical noise in the data is often high and thus an efficient pre-processing step is required before starting the statistical analysis. The process of removing technical effects to make samples and observations comparable is called "normalization". These normalization methods aim to ensure that technical biases minimally impact the results of the statistical analyses. There are technical biases specific to each technology: two such technical aspects which characterize microarrays are background noise and the differences between array platforms, although many methods have been suggested to correct these unwanted sources of data variation [150]. The main biases that can alter the expression levels of RNA-seq are the different number of sequencing reads generated for each sample, the gene length, and the guanine and cytosine nucleotide gene content (GC content). These aspects will be looked at in more depth in Chapter 4. There are also other potential biases that are more general, such as "batch effects". These effects may appear when biological samples have been processed at different moments, by different laboratories, on different devices, etc, although if the experiment has been appropriately designed it is possible to remove or at least mitigate this effect.

Another important challenge for statistics is the dimensionality of omics data. Thousands of variables are typically generated in one experiment, while

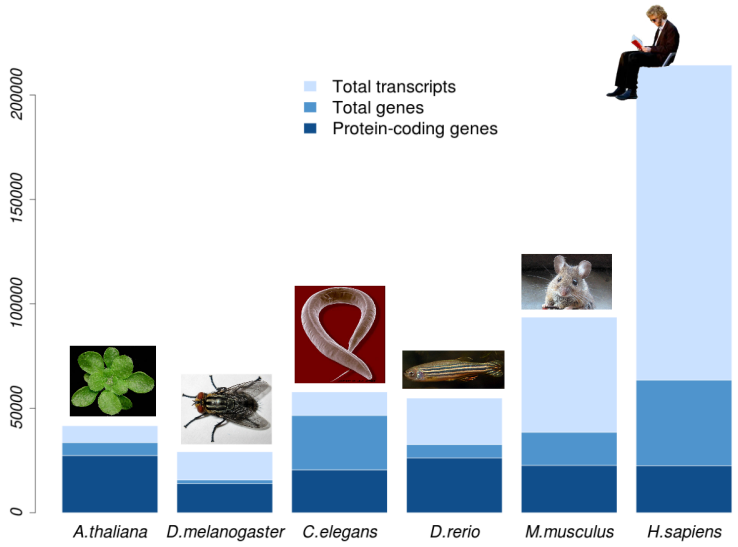


Figure 1.6: Total number of genes and transcripts for different species.

the sample size tends to be relatively small (most studies are not bigger than twenty or thirty observations). In transcriptomics, and in particular in this work, the variables analyzed are gene expression levels and the observations are biological samples where the expression levels have been measured. Figure 1.6 shows the number of genes or transcripts for different species to illustrate the high number of variables in the analysis of transcriptomic data.

When analyzing an experiment in which the expression levels were measured under different experimental conditions (over time, for different treatments, or diseases, etc.), the first question that obviously arises is if an association between the expression of the gene and the experimental conditions (i.e. the covariate) exists. In other words, the aim of the analysis is to identify the genes with expression changes across conditions, a problem which is also known as "differential expression analysis". Many different approaches can be envisaged to address differential expression (DE) studies and these can be classified into two groups: univariate and multivariate methodologies.

Univariate statistical approaches are probably the most popular. In general, they consist of testing a null hypothesis for each one of the variables

(genes). The null hypothesis is equivalent to saying that gene expression is not affected by the covariate that describes the experimental conditions compared [108, 121, 133]. Due to the dimensionality problem mentioned above, specifically the low number of available observations, it is common that methods in this category borrow information from all genes to better estimate the parameters of the statistical model. In addition, the high number of tests performed may lead to a pronounced increase in the false positive rate and so multiple testing corrections must be used to adjust the p-values and reduce the number of false positives [11, 39, 138]. A very well-known procedure to do this is Bonferroni's adjustment which aims to reduce the family-wise error rate (FWER), i.e. the number of rejected true null hypotheses. However, Bonferroni's method is too restrictive and the reduction in false positives it would produce comes at the expense of increasing the number of false negatives, thus losing statistical power. Hence, other more permissive procedures have been proposed with the goal of reducing the false discovery rate (FDR), rather than the FWER. The FDR is defined as the proportion of true null hypotheses with respect to the number of rejected null hypotheses. There are a number of procedures that aim to reduce the FDR, among them the one proposed by Benjamini and Hochberg [11], which is widely used in Bioinformatics.

An alternative to univariate methods is multivariate analysis which takes into account the joint distribution of all variables in the study [20, 78, 103]. In transcriptomics, it is quite common to apply multivariate tools such as clustering or principal component analysis (PCA) to explore the data prior to statistical analysis to check if replicates within the same condition are properly clustered or if there is a batch effect in the data. However, it is more unusual to resort to these kinds of tools to analyze differential expression, maybe due in part to the difficulty of identifying differentially expressed genes (DEG) using these models. Even so, some examples can be found in the literature that tackle the problem of differential expression from a multivariate approach which succeed in selecting the most relevant genes [103, 114]. In addition, these multivariate methodologies can be complemented by incorporating pre-

viously known biological information, such as the biological pathways in which genes are involved or their specific functional role. This class of methods is classified as "pathway analysis" and focuses on analyzing sets of genes which share, for instance, the same biological function [31, 36, 74, 104].

Both univariate and multivariate models can be constructed either under parametric assumptions, i.e. based on data following a given probability distribution, or without distributional assumptions. In transcriptomics (or bioinformatics in general), it is not always straightforward to find a probability distribution that fits the data or even a suitable transformation that fulfills the parametric assumptions. As previously mentioned, gene expression is generally supposed to follow a Gaussian distribution when measured by microarrays and a Poisson or Negative Binomial distribution when RNA-seq is used. Nevertheless, this may not always be a true and model validation (that is required to check the distribution hypothesis), becomes tedious given the huge number of models obtained in univariate analyses. On top of this, the small number of replicates makes it difficult to estimate the model parameters and information must often be borrowed from the rest of the variables [122, 133]. Alternatively, non-parametric methods can instead be used to tackle all these difficulties and resampling procedures are a very popular approach in this field [82, 109].

Once the differentially expressed genes (DEGs) have been identified, their functional profile is usually interrogated to establish if they share biological functions that are not characteristic of the rest of the genes. This type of analysis is named "functional enrichment analysis" or "pathway analysis". There are many genes associated with a given biological function (e.g. "lipid transport"), and a single gene may have many different biological functions. Therefore, for each biological function and a given set of DEGs, a contingency table can be generated to count the number of genes inside and outside this set, and belonging to that functional category or not. To assess if the set of DEGs is enriched in a particular function, a test of independence such as

Fisher's exact test can be performed. Again, many tests must be performed (one per functional category) and multiple test correction is advisable.

In summary, transcriptomic studies do not only require a wide-ranging knowledge of existing statistical methods to extract relevant biological information from the data, but also a lot of inventiveness and flexibility in order to find efficient and practical solutions for researchers. Moreover, biotechnology is permanently evolving, and at high speed: in particular, technologies used for measuring gene expression have undergone a major revolution over the last few years. Microarrays became a very popular technique during the nineties, but the more recent NGS technologies such as RNA-seq are replacing them. Third generation, or single molecule long-read technologies (PacBio, Oxford Nanopore), are also starting to take off which might create new sets of transcriptome data in the future. As previously mentioned, the nature of the data generated by each technology is very different, consequently, models used to describe each type of transcriptomic data should be different and therefore the development of new statistical procedures is constantly required [83].

Chapter 2

Motivation, Aims, and Contributions

2.1 Motivation

One of the most common types of analysis in biological research is the comparison of gene expression profiles. A fundamental goal in these types of genome-wide study is to identify genes whose expression profile changes between conditions, in other words, to select the most relevant variables (genes) in terms of inter-condition variability. The variable selection problem, which is usually known in transcriptomics as "differential expression analysis", can be addressed from the univariate or multivariate point of view, but the complexity of the experimental design must always be taken into account.

When this PhD thesis started in 2009, DNA microarrays were a mature technology used to measure gene expression levels and, due to their affordable cost, there were many experiments available that included many different treatment types, developmental states, time series, etc. A wide variety of statistical procedures had already been designed to identify differentially expressed genes and their functional relationships [108, 115, 133, 139]. Specifically, our group had been working on both univariate [3, 29] and multivariate [3, 30, 103, 104] methodologies to solve these problems for complex experimental designs such as single or multiple time course experiments, and we detected a lack of proper tools to address differential gene expression analysis in these scenarios. This motivated the first part of this thesis (Chapter 3), which is dedicated to the variable selection problem when using multivariate approaches to model microarray gene expression profiles. In particular, we chose the ASCA-genes multivariate technique [103] as a starting point to propose some strategies to select the genes responsible for phenotype changes among different experimental conditions.

However NGS technologies quickly became common place in transcriptome analysis and RNA-seq data were being generated in our projects, making it necessary to develop new strategies to deal with this new kind of data. Therefore, the second part of this work is entirely focused on RNA-seq experiments. Being a novel technology, we first had to address the issue of data quality monitoring to evaluate the accuracy of expression estimation. Chapter

4 of this thesis deals with the quality assessment of expression measurements, the potential sources of bias in the technology and how to process the data to reduce the impact of technical noise on statistical results. In Chapter 5, the variable selection problem for the two-class comparison case (differential expression) is discussed. As stated in the previous chapter, the nature of expression data for RNA-seq is different to microarrays, so specific DE methods were needed for this technology. We opted to develop non-parametric data-driven procedures to overcome the limitations of parametric assumptions, and showed that these were efficient in controlling the false positive rate: two methodologies (NOISeq and NOISeqBIO), for technical and biological replicates respectively, were proposed and compared to the state of the art methodologies.

This thesis has given me the opportunity to examine new and exciting fields such as cell biology and bioinformatics, and specifically transcriptomics. The principles of gene expression, the different ways of measuring it and the statistical approaches for analyzing expression data obtained from different technologies were studied in depth. It has been very challenging and enriching to search for and develop suitable statistical tools to discover the biological stories at the root of each transcriptomic project. Hopefully, these pages will reflect the enthusiasm devoted to it.

2.2 Aims

1) **To develop variable selection strategies for multivariate methods applied to microarray data.**

   When using multivariate methods to model the association of gene expression to covariates describing the experimental conditions under study, a posteriori selection of the genes which meaningfully contributed to the model construction is normally desired. These should be genes with a significant change in expression levels between conditions. Variable selection strategies will be studied, using our group's previous work

[103] as a starting point. In particular, the following tasks will be undertaken:

- Proposal of new variable selection methods or variations of existing ones in the context of multivariate techniques for dimension reduction.
- Simulation of multi-factorial expression data and assessment of variable selection strategies on these synthetic data.
- Application of the best selection strategies to experimental data to identify the genes responsible for human stem cell differentiation under different oxygen concentration conditions.
- Application of these strategies to other analysis scenarios to assess the general validity of the methods when using other multivariate techniques.

2) **To generate tools to control the quality of count data from sequencing experiments in order to discover potential biases and to propose procedures to mitigate their effect.**
Sequencing technologies such as RNA-seq produce count data that might be biased due to technical noise. It is convenient to remove or, at least, reduce these unwanted technical effects before performing further statistical analyses. Regarding this issue, we will focus on the following aspects:

- Design exploratory and diagnostic graphical tools to detect potential technical biases.
- Show the usefulness of these exploratory tools using different RNA-seq experimental data sets.
- Review of the most popular normalization procedures to remove these biases, and apply some of them to experimental data.

- Propose methods to filter out low-count features, which are unreliable and can decrease the power of statistical methods to identify true effects of the experimental factors.

3) To develop differential expression methodologies for RNA-seq data.

There are many parametric methods available to study DE in pair-wise comparisons. These kinds of methods might present some limitations regarding the distributional assumptions but few non-parametric alternatives exist. In this work, we will propose two complementary non-parametric approaches for application on data with technical replicates, no replicates at all or with biological replicates. These are the main issues we will address in this section:

- Development of the NOISeq method which can be applied to data with technical replicates or without replicates. Comparison of NOISeq to other DE methodologies on both simulated and experimental datasets.
- Adaptation of the NOISeq method for use on data with biological replications (NOISeqBIO). Evaluation of several versions of NOISeqBIO on simulated data to determine the best option.
- Comparison of NOISeqBIO to other DE methods in several simulated scenarios to check the method's performance and to assess the influence of biological parameters such as noise, number of features, percentage of differentially expressed genes, etc. on the DE results.

2.3 Main contributions

I first arrived to Bioinformatics Department at the CIPF in 2008 to co-supervise the final year project of one of my students at the Technical University of Valencia ("Functional prediction of novel citrus sequences from gene

expression analysis", awarded by Bancaja Prize 2008). About one year later, I started to work on this thesis. Therefore, my first collaborations [104, 114] taught me the basics of transcriptomics (specifically about microarray data analysis) and allowed me to identify which lines of work in the group I would follow. I became primarily interested in multivariate approaches and, in particular, in the variable selection problem. That is why the first part of this document (Chapter 3) is focused on the study of variable selection strategies in multivariate models, as published in [31, 113, 142]. Some of these strategies were implemented in a web tool for analyzing serial gene expression data named SEA [102].

By that time, NGS technologies emerged and there was a need for the development of ad-hoc statistical methods to analyze this kind of data. RNA-seq was rapidly gaining popularity in gene expression estimation but, due to the still high cost of the technology, the experiments available at that time had very simple designs, usually including only two experimental groups and only a few replicates. Hence, the context for the variable selection problem had slightly changed. First, we established pipelines for RNA-seq data quality assessment and pre-processing to obtain normalized and bias-free data (Chapter 4) and then we focused on the variable selection problem for pairwise comparisons (Chapter 5). All these methodologies were gathered into a Bioconductor R package named NOISeq and summarized in [141] and in another paper (in preparation). The quality control tools were also implemented in Qualimap software [49]. Our expertise in analyzing RNA-seq data helped us to update the maSigPro tool [29] for dealing with RNA-seq time series [105].

2.3.1 Journal papers

1. Nueda MJ, Sebastián P, Tarazona S, García-García F, Dopazo J, Ferrer A, and Conesa A. *Functional assessment of time course microarray data*. **BMC Bioinformatics**, 10(Supp 6):S9, **2009**.

2. Conesa A, Prats-Montalbán JM, Tarazona S, Nueda MJ, and Ferrer A. *A multiway approach to data integration in systems biology based on Tucker3 and N-PLS*. **Chemometrics and Intelligent Laboratory Systems**, 104(1):101-111, **2010**.

3. Prado-López S, Conesa A, Armiñán A, Martínez-Losa M, Escobedo-Lucea C, Gandía C, Tarazona S, Melguizo D, Blesa D, Montaner D, Sanz-González S, Sepúlveda P, Götz S, O'Connor JE, Moreno R, Dopazo J, Burks DJ, and Stojkovic M. *Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium*. **Stem Cells**, 28(3):407-418, **2010**.

4. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. *Differential expression in RNA-seq: A matter of depth*. **Genome Research**, 21:2213-2223, **2011**.

5. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, and Conesa A. *Qualimap: evaluating next-generation sequencing alignment data*. **Bioinformatics**, 28(20):2678-2679, **2012**.

6. Tarazona S, Prado-López S, Dopazo J, Ferrer A, and Conesa A. *Variable selection for multifactorial genomic data*. **Chemometrics and Intelligent Laboratory Systems**, 110:113-122, **2012**.

7. Ponzoni I, Nueda MJ, Tarazona S, Götz S, Montaner D, Dussaut JS, Dopazo J, and Conesa A. *Pathway network inference from gene expression data*. **BMC Systems Biology**, 8(2), 1-17, **2014**.

8. Nueda MJ, Tarazona S and Conesa A. *Next maSigPro: updating maSigPro Bioconductor package for RNA-seq time series*. **Bioinformatics**, 30(14), **2014**.

9. Tarazona S, Furió P, Turrà D, Di Pietro A, Ferrer A, and Conesa A. *NOISeq: An R package for visualization, quality assessment and differential expression for RNA-seq experiments*. (In preparation)

2.3.2 Conferences

- VII Colloquium Chemiometricum Mediterraneum. Granada, Spain. June, 2010. Tarazona S, Prado-López S, Dopazo J, Ferrer A, and Conesa A. "Variable selection for multifactorial genomic data" (Poster).

- XXXII Congreso Nacional de Estadística e Investigación Operativa. A Coruña, Spain. September, 2010. Tarazona S, Prado-López S, Dopazo J, Ferrer A, and Conesa A. "Variable selection for multifactorial genomic data" (Oral presentation by Sonia Tarazona).

- X Symposium on Bioinformatics. Málaga, Spain. October, 2010. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. "Differential Expression in RNA-seq: A Matter of Depth" (Oral presentation by Sonia Tarazona).

- 8^a Reunión Red Valenciana de Genómica y Proteómica. Valencia, Spain. November, 2010. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. "Differential Expression in RNA-seq: A Matter of Depth" (Poster).

- ISMB/ECCB. Viena, Austria. July, 2011. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. "Differential Expression in RNA-seq: A Matter of Depth" (Oral presentation by Ana Conesa).

- XI Symposium on Bioinformatics. Barcelona, Spain. January, 2012. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. "Differential Expression in RNA-seq: A Matter of Depth" (Poster).

- HitSeq 2013. Berlin, Germany. July, 2013. Tarazona S, Furió P, Turrà D, Di Pietro A, Ferrer A, and Conesa A. "Quality-control, experimental design and FDR controlled differential expression of RNA-seq with the NOISeq R package" (Poster).

- XII Symposium on Bioinformatics. Sevilla, Spain. September, 2014. Sebastián A, Pascual-García A, Abascal F, Aguirre J, Andrés-León E,

Bajic D, Bau D, Bueren-Calabuig JA, Cortés-Cabrera A, Dotu I, Fernández JM, Dos Santos HG, García-Jiménez B, Guantes R, Irisarri I, Jiménez-Lozano N, Klett J, Méndez R, Morreale A, Perona A, Stich M, Tarazona S, Yruela I, Zardoya R. "Bioinformática con Ñ v1.0: a collaborative project of young Spanish scientists to write a complete book about Bioinformatics" (Oral presentation by Álvaro Sebastián).

2.3.3 Software

- Tarazona S, Furió P, Ferrer A, and Conesa A.
  NOISeq, Bioconductor R package.
  http://www.bioconductor.org/packages/release/bioc/html/NOISeq.html

- García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, and Conesa A.
  Qualimap, platform-independent application.
  http://qualimap.bioinfo.cipf.es/

2.3.4 Courses

Apart from being a part-time lecturer on Statistics at the Technical University of Valencia, I have had the opportunity during the last years of showing my own work and the state of the art in transcriptomic analyses to the scientific community by teaching in different courses:

- International Course of Massive Data Analysis (Centro de Investigación Príncipe Felipe, Valencia). From 2008 to 2014, lectures on "Basic Statistics applied to Bioinformatics, "Differential Expression in RNA-seq", "Functional Enrichment Analysis" or "DNase-seq".

- Data analysis workshop for massive sequencing data (University of Granada, Granada). 2011.

- Course on RNA-seq and ChIP-seq analysis (IDIBAPS, Barcelona). 2012.

Chapter 3

Variable selection for multifactorial genomic data

Tarazona S, Prado-López S, Dopazo J, Ferrer A, and Conesa A.
Variable selection for multifactorial genomic data
Chemometrics and Intelligent Laboratory Systems, 110:113-122, **2012**

Prado-López S, Conesa A, Armiñán A, Martínez-Losa M, Escobedo-Lucea C, Gandía C, Tarazona S, Melguizo D, Blesa D, Montaner D, Sanz-González S, Sepúlveda P, Götz S, O'Connor JE, Moreno R, Dopazo J, Burks DJ, and Stojkovic M.
Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium
Stem Cells, 28(3):407-418, **2010**

Conesa A, Prats-Montalbán JM, Tarazona S, Nueda MJ, and Ferrer A.
A multiway approach to data integration in systems biology based on Tucker3 and N-PLS
Chemometrics and Intelligent Laboratory Systems, 104(1):101-111, **2010**

Ponzoni I, Nueda MJ, Tarazona S, Götz S, Montaner D, Dussaut JS, Dopazo J, and Conesa A.
Pathway network inference from gene expression data
BMC Systems Biology, 8(2), 1-17, **2014**

3.1 Introduction

High-throughput genomic and transcriptomic experiments generate data for a high amount of variables (e.g. genes) on a much lower number of individuals (samples). Common approaches to explore this kind of data are clustering methods such as hierarchical or KNN clustering [89, 135], and dimensionality reduction techniques such as Principal Component Analysis (PCA). PCA [63, 64, 68, 112] is frequently used in transcriptomic data to group samples, identify associated genes or to spot those genes or samples behaving completely different from the rest [34, 117]. In simple case-control studies, the methodology is able to provide biologically interpretable results. However, more complex experimental designs can also be found in transcriptome research, that include factors such as time effect, treatment, tissue, strain, etc., at different levels, giving rise to high-dimensional multifactorial datasets. For these multifactorial experiments, other dimension reduction techniques exist that tackle the analysis of the data in a more efficient way and achieve a better interpretation of the results. Some examples are Tucker3 [145] or PARAFAC [62], which have been successfully applied to the analysis of genomic data [151]. Another interesting approach is ASCA (ANOVA-Simultaneous Component Analysis) [132], adapted to genomic data in the ASCA-genes software [103]. ASCA-genes is a powerful tool to extract targeted signals from noisy data in complex experimental setups using a combination of ANOVA-like data decomposition and PCA.

In many cases, though, descriptive analysis is not the only goal of the experiment, but also the identification of responsive (or activated) genes, since they give the clue to the molecular biology interpretation of transcriptional regulation. When facing the issue of variable selection within the framework of dimension reduction techniques, there exist some rules of thumb such as considering that a variable is important if its loading absolute values are higher than a certain threshold. However, this is a rather arbitrary way of selecting variables. More sophisticated variable selection methods can be found in the literature, especially for PCA. Jolliffe [69, 70] used the absolute value

of PCA loadings to measure the contribution of the original variables to the model and selected as many of these variables as the number of selected latent variables in order to retain the maximum variance of the data. McCabe [95] recommended four different criteria to select what he called principal variables and then evaluated all possible subsets of original variables to find the one optimizing the pursued criterion. Krzanowski [76] combined PCA with Procrustes analysis to select those variables preserving the multivariate data structure, and used a Procrustes criterion to quantify the similarity of compared structures. Since exploring all the subsets of q variables (q being the number of variables to be selected) might be very computationally expensive, he included a backward procedure to discard variables. Guo *et al.* [55] improved the search of the best subset in the latter method by applying a genetic algorithm to avoid exhaustive searching. Westad *et al.* [149] used Student's *t*-tests based on loadings and their estimated standard uncertainties to calculate the significance on each variable for each component. Principal Feature Analysis [88] is based on taking PCA loadings and clustering them using the K-Means algorithm. The number of clusters must be equal or greater than the number of PCs. In each cluster, the closest variable to the mean of the cluster is selected (principal feature). Finally, variable selection can be carried out by applying Sparse Principal Component Analysis [90]. Sparse PCA generates linear combinations of the data variables explaining a maximum amount of variance in the data while having only a limited number of nonzero coefficients.

The purpose of most of these methods is reducing the number of variables to achieve a better interpretation of the principal components. Several of them are unfeasible in the context of genomic data due to the large number of variables (genes) or inappropriate due to the low signal to noise ratio that characterizes these data. Another drawback is that the majority of these approaches need to set the number of variables to be selected (or removed) in advance, which is generally an unwanted constraint when trying to identify responsive genes. In this chapter, several selection strategies are compiled

that avoid this constraint and are compared by applying them to the analysis of multifactorial genomic data following the work of adapting ASCA [132] to genomic experiments in the ASCA-genes tool [103]. ASCA-genes was shown to be an effective approach for the analysis of complex datasets and the gene selection strategy presented in that work was proven to give good results with signal rich transcriptomic datasets. Here, that study is extended and a vast array of signal to noise conditions will be considered together with different selection strategies to provide a comprehensive understanding of the behavior of complex transcriptomic designs.

In this chapter, two novel approaches are proposed for variable selection in the context of multifactorial gene expression experiments: minAS and Gamma approximation. The ASCA-genes framework is used for treating the multifactorial nature of the data. However, the gene selection strategies proposed rely on the probability distribution of PCA statistics and can be applied together with other dimension reduction techniques. Both minAS and Gamma methods in combination with ASCA-genes have been implemented in the web suite for Serial Gene Expression Analysis: SEA (http://sea.bioinfo.cipf.es/) [102], which is freely available to the scientific community. Moreover, the minAS strategy has been applied to other multivariate variable selection scenarios [31, 113] that will be presented and briefly discussed at the end of this chapter.

3.2 Methods

The methods presented in this chapter were initially designed to be used for analyzing gene expression data measured by microarrays. However, with proper data transformations, they could be also used on data coming from other technologies, such as RNA-seq. In this work, methods have been validated on microarray data and the simulation studies also mimic the behavior of this kind of data.

Let $\mathbf{X}_0$ be the gene expression matrix, with dimensions $M \times N$, where N is the number of variables (e.g. genes) and M is the number of observations

(biological samples). If samples have been taken according to a certain experimental design, including one or more different factors such as treatment, tissue, time, etc. with different levels and different number of replicates in each level, we are dealing with multifactorial datasets. The experimental setup must be taken into account when choosing the appropriate dimension reduction technique in order to better extract the information contained in the data. ASCA model was used because it tackles the problem of complex experimental designs and efficiently separates signal from noise to achieve an optimal interpretation of the results in terms of experimental factors effects [103].

3.2.1 The ASCA-genes framework

To present the ASCA-genes methodology, let us consider the specific case of an experiment with two factors. In the context of genomic experimental designs, one of the factors is usually time (say, for example, factor a). The other factor b indicates the experimental group, such as treatment or tissue. If x_{ijr} is the expression level for a given gene measured at time point i, under experimental condition j and for replicate r, Equation 3.1 shows the ANOVA model definition for that gene, where μ is an offset term, α_i is the model parameter for the time factor on level i, β_j measures the j-th group effect, $(\alpha\beta)_{ij}$ represents the interaction effect between the i-th time and j-th group, and the individual variation is indicated by $(\alpha\beta\gamma)_{ijr}$:

$$x_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijr} \tag{3.1}$$

Estimates of the ANOVA parameters of Equation 3.1 can be obtained for all genes and collected into matrices as in Equation 3.2, where the gene expression matrix $\mathbf{X}_0$ has been mean centered, resulting in matrix $\mathbf{X}$.

$$\mathbf{X} = \mathbf{X}_a + \mathbf{X}_b + \mathbf{X}_{ab} + \mathbf{X}_{abg} \tag{3.2}$$

Each one of the submatrices on the right hand side of Equation 3.2 contains the estimated effects associated with a certain experimental factor, for

example factor a, factor b, the interaction ab between them or the residuals abg. The estimation of these effects depends on the nature of the factors (between or within subjects, random or fixed effects, etc.). In this work, we have considered the most simple case of ANOVA-like decomposition: fixed effect factors between subjects. Independence of measurements holds when expression values over time are independent, as frequently happens in genomics because they correspond to different biological samples.

As the goal in time-course experiments is usually to detect gene expression profile changes between experimental groups (factor b), in this study, the interaction effect has been joined to factor b effect and analyzed in one submodel as it is shown in Equation 3.3:

$$\mathbf{X} = \mathbf{X}_a + \mathbf{X}_{b+ab} + \mathbf{X}_{abg} \tag{3.3}$$

For the remainder of this work, ASCA submodels in Equation 3.3 will be named as "submodel a" and "submodel $b + ab$", respectively.

PCA is applied to each one of the submatrices (Simultaneous Component Analysis) to reveal major expression patterns associated to the experimental factors and to identify relevant experimental conditions. At this point, dimensionality reduction is undertaken by selecting, for each submodel, k_x principal components (for $x{=}a, b + ab, abg$). Many criteria have been proposed to decide the optimal number k_x of principal components to choose. In this work, a component was selected if the percentage of total variability it explained was higher than $C/nmax$, $nmax$ being the maximum number of components (given by the rank of the matrix) and C a constant, set to 1.5 in our case. The resulting ASCA-model is given in Equation 3.4:

$$\mathbf{X} = \mathbf{T}_a\mathbf{P}_a^t + \mathbf{T}_{b+ab}\mathbf{P}_{b+ab}^t + \mathbf{T}_{abg}\mathbf{P}_{abg}^t + \mathbf{E} \tag{3.4}$$

where, the scores of each submodel are given by the $M \times k_x$ matrices indicated by $\mathbf{T}_a$, $\mathbf{T}_{b+ab}$, $\mathbf{T}_{abg}$, and the submodel loadings are given by the $N \times k_x$ matrices $\mathbf{P}_a$, $\mathbf{P}_{b+ab}$, $\mathbf{P}_{abg}$, where $\mathbf{P}_x^t\mathbf{P}_x{=}\mathbf{I}$ for $x{=}a$, $b + ab$ or abg. $\mathbf{E}$ is a matrix in which the residuals of all submodels of ASCA-model are

collected: $\mathbf{E} = \mathbf{E}_a + \mathbf{E}_{b+ab} + \mathbf{E}_{abg}$, where $\mathbf{E}_x = \mathbf{X}_x - \mathbf{T}_x\mathbf{P}_x^T$ for $x = a$, $b+ab$ or abg. The extension of this model to more than two experimental factors is straightforward.

Once the major variability patterns have been identified, and assuming that the model is biologically meaningful, the next step is to select genes whose expression is affected by the experimental factors. When considering the expression of a single gene, this might follow the general model, change according to a different pattern, or simply present a flat profile. Two statistics are proposed to characterize the behavior of genes within each submodel: the leverage and the Squared Prediction Error (SPE).

The leverage measures the importance of a variable (gene) in the PCA model. Leverage values for all the genes in the submodel x can be computed from the loadings matrix according to Equation 3.5 (see [93]):

$$\mathbf{h}_x = diag[\mathbf{P}_x\mathbf{P}_x^t]; \quad x = a, b + ab \tag{3.5}$$

The SPE associated with a particular gene is a measure of the goodness of fit of the model for that specific gene. Genes not following the general structure of the model will have high SPE. SPE values can be computed from the residuals matrix in each submodel ($\mathbf{E}_x = \mathbf{X}_x$ - $\mathbf{T}_x\mathbf{P}_x^t$) according to Equation 3.6:

$$\mathbf{SPE}_x = diag[\mathbf{E}_x^t\mathbf{E}_x]; \quad x = a, b + ab \tag{3.6}$$

By combining the information given by the leverage and the SPE, genes can be classified (as proposed in [103]) in the following groups:

- Genes relevant to the model (following the main trends): high leverage and low SPE.
- Influential but poorly modeled genes: high leverage and high SPE.
- Badly modeled genes which are potential outliers: low leverage and high SPE.

- Non-responsive genes (not affected by the experimental factors): low leverage and low SPE.

Therefore, we are interested in the genes which present a high leverage or high SPE, because these genes may be affected by the experimental conditions. To decide which genes should be classified as "interesting" (responsive), a threshold must be established for both leverage and SPE in such a way that those genes presenting an SPE or leverage higher than this threshold will be selected. Nueda and co-workers calculated the SPE threshold by using Box's approximation [17] for SPE distribution. The leverage threshold was obtained by resampling techniques [41]. However, they observed that these selection strategies presented a good performance when the signal to noise ratio in the dataset was high, but were not so effective for data with low signal to noise ratio. Hence, in this chapter, other selection methods have been introduced and compared to the ones in ASCA-genes under a much wider variety of biological scenarios. Both simulated and real datasets are used to evaluate the performance of the proposed selection methods.

3.2.2 Variable selection strategies

Once the dimension reduction model has been established, the goal is often finding the variables with higher contribution in the model. In our case, the most "regulated" genes. The variable selection strategies we propose here consist of three steps: first, choosing an appropriate statistic to measure the importance of the variables in the model (leverage and SPE in this study); second, estimating the probability distribution of this statistic (in a parametric or non-parametric way) and, finally, establishing the threshold to separate "interesting" from "uninteresting" variables (genes). As in ASCA-genes, our proposals are focused on studying the univariate distribution of both SPE and leverage statistics, although most of the methods we present are valid for other statistics or even other multivariate methods, as it will be shown at the end of this chapter.

It should be noted that SPE and leverage statistics can be computed for each gene in each of the different ASCA submodels a, $b + ab$ and abg. Gene selection is therefore possible for each of these submodels independently. In this work we have chosen to evaluate the gene selection coming from both a and $b+ab$ submodels as these capture the gene expression changes of interest in the proposed scenario, namely, the time associated changes (submodel a) and the time-experimental factor interaction (submodel $b + ab$). Depending on the aim of the experiment, all or only specific submodels might be relevant for the study, and selection will have to be based on the SPE and/or leverage statistics of the corresponding submodels. Thus, interpretation of the gene selection has always to be done on the light of the ASCA submodels considered.

Generally and because of the nature of expression data, most genes present a low SPE or low leverage values. Hence, it is expected that these statistics follow a mixture distribution of, at least, two populations. The biggest population is that of "uninteresting" genes (with statistic values closer to zero). The other population(s) correspond to "interesting" genes (those with higher values for the statistic). As our aim is to separate "interesting" from "uninteresting" genes, the mixture model can be written as in Equation 3.7:

$$f(x) = p_0 f_0(x) + p_1 f_1(x) \tag{3.7}$$

where, x is the value of either SPE or leverage for a particular gene, p_0 is the proportion of "uninteresting" genes (a priori unknown), $f_0(x)$ is the null probability density function (i.e. probability density function for "uninteresting" genes), and p_1 and $f_1(x)$ are, respectively, the proportion of "interesting" genes and their probability density function.

Two different approaches can be used to establish the threshold for SPE or leverage values. The first one consists of estimating the "uninteresting" genes distribution (null distribution) and using a percentile of this estimated distribution as the threshold. The methods compared in this work that follow this first approach are: Box's method [17], Jackson & Mudholkar's method

[66], Gamma method and resampling techniques [41]. In the first three, the null distribution is estimated in a parametric way, while resampling is considered a non-parametric technique. In the second approach, an approximation is obtained for the mixture distribution and the threshold is taken as the value which best separates the two components of the mixture. Many authors have focused on the parametric estimation of the distribution of the mixture components (see, for example, Efron's work at [43] or [42]). But we observed that, due to the huge difference between the sizes of both populations, it was very difficult to parametrically estimate the probability distribution of each component. Therefore, only a non-parametric approach is introduced here, which is called minAS (MINimum Algorithmic Selection).

Box's method

Assuming that errors from a PCA model approximately follow a multivariate normal distribution and given that SPE is a quadratic form of the error associated with a particular variable, Box [17] showed that SPE distribution could be estimated by a weighted χ^2-distribution ($g\chi^2_h$). In ASCA-genes [103], this distribution was used to calculate the (1-α)% confidence SPE threshold for each PCA submodel. Parameters g and h are estimated by the matching moments method and the following expression is obtained for SPE threshold at α level of significance, where m is the sample mean and v is the sample variance:

$$SPE_\alpha = \frac{v}{2m}\chi^2_{\frac{2m^2}{v}}(\alpha) \tag{3.8}$$

Jackson & Mudholkar's method

Jackson and Mudholkar [66] found another approximation for SPE distribution in PCA models, by using the residuals matrix **E**. Therefore, for PCA coming from each ASCA submodel, the SPE threshold at α level of significance can

be computed as follows:

$$SPE_\alpha = \theta_1[1 - \frac{\theta_2 h(1-h)}{\theta_1^2} + \frac{z_\alpha(2\theta_2 h^2)^{1/2}}{\theta_1}]^{1/h} \tag{3.9}$$

where $\mathbf{V}=\frac{\mathbf{E'E}}{N-1}$, N being the number of variables (genes) in the model; θ_i=trace($\mathbf{V}^i$), for i=1,2,3; and h=1-$\frac{2\theta_1\theta_3}{3\theta_2^2}$.

Gamma method

Gnanadesikan and Kettenring [51] proposed the Gamma distribution for the squared residuals from a PCA. Following this idea and because of the flexibility of this distribution to suit many density curves, we used it to approximate both the SPE and leverage null distributions. Given the statistic values for the N genes in each submodel $(x_1, ..., x_N)$, shape (k) and scale (θ) parameters for the gamma distribution can be estimated by maximum likelihood [27]:

$$\hat{\theta} = \frac{\bar{x}}{\hat{k}} \tag{3.10}$$

$$\hat{k} = \frac{3 - s + \sqrt{(s-3)^3 + 24s}}{12s}$$

$$s = \ln(\bar{x}) - \frac{1}{N}\sum_{i=1}^{N} \ln(x_i)$$

The corresponding threshold for the statistic (either SPE or leverage) is then the percentile (1-α)% of the estimated gamma distribution.

Resampling techniques

Resampling methods are non-parametric procedures to determine the statistical significance of a result, sampling repeatedly within the same data. An empirical distribution is generated for an statistic under the null hypothesis by taking the original data, randomly shuffling them numerous times and

computing the statistic value for each of the permuted datasets. The way of permuting the data depends on the null hypothesis to be tested [41].

In ASCA-genes [103], a permutation method was used to define the threshold of leverage. In the present work, we study the performance of permutation techniques to obtain the confidence thresholds not only for leverage but also for SPE. We also compare their permutation strategy with our proposal. Both strategies are described below for the $M \times N$ data matrix **X** and are summarized in Figure 3.1.

Strategy 1.- As implemented in ASCA-genes, K row permutations of matrix **X** are generated, destroying the structure of the experimental design. In this case, the null hypothesis to be tested is that experimental conditions do not affect gene expression, i.e. all genes have a flat profile across conditions. For instance, the null hypothesis to test for submodel a (analogous for submodel $b + ab$) would be:

$$H_0 : (\alpha_1)_k = ... = (\alpha_t)_k = 0, \forall k = 1, ..., N \tag{3.11}$$

Strategy 2.- The null hypothesis to test in this strategy is that all genes are equally responsive. In the case of leverage, for example, it would imply that all genes have equal leverage values.

$$H_0 : (\alpha_i)_1 = ... = (\alpha_i)_N, \forall i \tag{3.12}$$

If this is true, all the genes would have the same contribution in the PCA model and the residual errors would also be similar. Hence, the novel permutation strategy we propose in this work consists of performing K column permutations. Moreover, the permutation of values in the columns is different for each row so that the structure in the data (associations among genes, and among genes and experimental conditions) is totally broken.

In this work, the number of permutations K was set to 1000. Once the permuted matrices have been generated, an ASCA model is fitted to each one of them. SPE and leverage values are then obtained for each gene in each permutation to generate the reference distribution. The threshold can

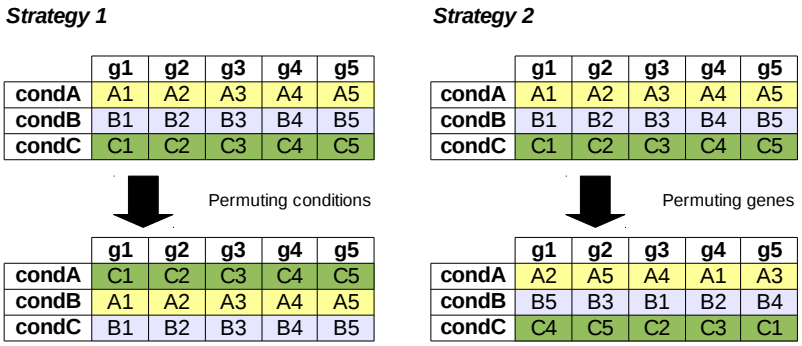


Figure 3.1: Resampling strategies.

Table 3.1: Methods to calculate SPE or leverage threshold by resampling techniques

Method	*Permutation strategy*	*Threshold computation*
1	1 - Permuting conditions	Option (a) - For each gene
2	2 - Permuting genes	Option (a) - For each gene
3	2 - Permuting genes	Option (b) - Globally

be calculated from this reference distribution in two ways:
Option (a).- First, the (1-α)% percentile of the K statistic values for each gene is computed and the threshold is obtained as the (1-α)% percentile of the N gene percentiles. This is the option implemented in ASCA-genes.
Option (b).- We propose using the (1-α)% percentiles of the KxN statistic values obtained from the K permutations and N genes.
The three resampling methods to be compared in this work are combinations of permutation strategies 1 and 2 and options (a) and (b) to compute thresholds. They are described in Table 3.1.

minAS

In this work we introduce minAS, which is a very intuitive data-driven method. This algorithmic approach consists of empirically estimating the mixture den-

sity function for either the SPE or the leverage and then computing the first local minimum closest to the "uninteresting" genes probability density curve. The SPE or leverage value in which this minimum is reached is taken as the threshold that separates both distributions. The minAS strategy assumes that the mixture distribution in Equation 3.7 for SPE or leverage is, at least, bimodal. The intrinsic nature of genomic data makes this assumption hold in general. However, it is not always possible to visualize this bimodality in histograms, due to the large difference in the sizes of both populations.

To estimate the mixture density curve, a non-parametric density estimator was used: the kernel density estimator (KDE) [125]. A KDE is a sophisticated version of histograms that produces smoothed density curves and it is defined in Equation 3.13:

$$\hat{f}(x) = \frac{1}{N \cdot h} \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right) \tag{3.13}$$

where x_i are the observed values, N is the total number of observed values (in this case, the number of genes), h is the bandwidth (the smoothing parameter) and $K(x)$ is the kernel function, that weights each observation depending on the distance to the point for which the density is being estimated. For instance, when choosing a Gaussian kernel to estimate f at x_0, and for a given bandwidth h, the highest weights correspond to the closest observations to x_0, and the weight diminishes as the distance to x_0 increases. However, the KDE goodness of fit relies more on bandwidth h than on the kernel choice. There are different rules of thumb to compute the optimum bandwidth. For instance, Silverman [129] takes into account the dispersion in the data and the sample size to compute bandwidth for KDE with a Gaussian kernel. These are the default options in the R *density* function from the library *stats*, which we used to obtain the KDE within minAS.

In the minAS algorithm, users can choose the kernel and the method to calculate bandwidth (as in the R *density* function), as well as the number of points for which the density is fitted. The smoothing of the KDE is determined by the bandwidth computed by the chosen method. To increase or

decrease this smoothing, the value of the coefficient *adjust* (which defaults to 1) can be increased or decreased, respectively. Several kernel functions or methods to calculate the bandwidth can be chosen and then minAS selects the mixture estimation that best fits the data according to one of the two implemented options: "max" and "mean". As the true density function is unknown, cumulative distribution functions computed from the KDE are compared with the empirical cumulative distribution function derived from SPE or leverage values. In order to compare them, the difference between the empirical distribution and the KDE cumulative distribution is computed for each value. Then, in the case of the "max" option, the maximum of these differences (Kolmogorov-Smirnov distance) is taken. For the "mean" option, the mean of all these differences is obtained. The KDE with the smallest maximum (or mean) difference is selected.

Once the best KDE has been obtained, minAS computes the minima of this curve. By default, the first local minimum after the highest peak is taken as the cutoff value to separate the two populations, i.e. "interesting" from "uninteresting" genes. However, minAS users can also set the maximum number of minima to be computed, calculate all of them or provide the interval where the minimum has to be found. A plot is provided in which all the computed minima are represented over the mixture distribution. Then, if more than one minimum is found, users can decide to reduce the number of selected genes by choosing a more restrictive threshold.

3.2.3 Data simulation

The synthetic data sets used to evaluate the variable selection methods were generated using a simulation algorithm. This algorithm is intended to mimic the behavior of gene expression across time and for different experimental conditions. In order to cover a wide variety of biological scenarios, we varied the values of the input parameters of the algorithm when generating the data. The input parameters are listed in Table 3.2, and some of them are explained in more detail below:

Experimental factors: Factors that are controlled by the experimentalist. The algorithm was designed to work with two experimental factors: the time factor and another one referring to the experimental group such as treatment, tissue, illness, etc.

Signal genes (deg): Also called responsive genes or differentially expressed genes. Genes which are activated at any time point and for any of the experimental groups.

Expression pattern: Gene behavior over time for a certain experimental group. As only short time series have been studied in this paper, expression patterns have been summarized and modeled according to the following functions of time t:

- Continuous induction ($\beta_0+\beta_1 t$; $\beta_1>0$): Gene activity increases linearly as time elapses.
- Continuous repression ($\beta_0+\beta_1 t$; $\beta_1<0$): The gene is initially active and the activity decreases linearly as time elapses.
- Transitory induction ($\beta_0+\beta_1 t+\beta_2 t^2$; $\beta_1>0$, $\beta_2<0$): The gene is initially inactive, it increases its activity and, after a certain time and until the final time point, this activity decreases.
- Transitory repression ($\beta_0+\beta_1 t+\beta_2 t^2$; $\beta_1<0$, $\beta_2>0$): The gene is initially active, it decreases its activity and, after a certain time and until the final time point, this activity increases.
- Plain (for non-responsive genes): The gene remains inactive over time.

Coefficients β_0, β_1 and β_2 have been computed in such a way that the absolute value for gene expression level is not higher than a certain maximum value. In microarray studies, expression levels are obtained as a log-ratio from color intensities, so the expression values for responsive genes may vary between approximately 2 and 5. Thus, the maximum expression value computed

by the algorithm is a random number between 2 and 5, 0.5, 0.3, 0.1 and 0.1 being the probabilities for values 2, 3, 4 and 5, respectively.

Class of genes: A class of genes is a group of genes that follow the same temporal pattern within a certain experimental group (e.g. treatment). The number of classes of genes will determine the variety of different gene behaviors in the experiment.

The criterion for allocating the expression patterns to each class of genes in each experimental group tries to imitate biological behavior as much as possible. Every pattern can be found in at least one class of genes. When genes follow an induction pattern for a certain class, there must be another class with the equivalent repression pattern. Inside every class containing responsive genes, experimental groups for which genes are activated are randomly selected taking into account that:

- Genes in that class must be active for at least one experimental group.
- If genes are expressed for several experimental groups, the expression pattern must be the same for all of these experimental groups and replicates.
- For two complementary gene classes, the genes must be active for the same experimental groups.

When modeling gene behavior, noise must be introduced to get realistic simulated expression data. Two types of noise have been considered: random and structural noise. Random noise is generated by technology and affects all the genes and samples in a similar way. For microarray data, it is common that about 20% of the signal is random noise. Structural noise is related to biological sample handling to obtain signal intensities. Hence, it takes the same value for all the genes in the same microarray. It has been observed that expressed genes are usually more affected by this kind of noise.

Let G_{ij} be the "pure" expression value (without noise) for a certain gene i and sample j.

Table 3.2: Input parameters for simulation algorithm

Parameter	*Description*
N	Number of genes.
SN	Percentage of differentially expressed genes ($\%deg$).
t	Time points array.
f	Number of experimental groups.
r	Number of replicates for each time point and experimental group.
d	Number of classes of signal genes.
rn	Maximum percentage of expression level corresponding to random noise in signal genes.
sd	Standard deviation for structural noise and for random noise in non-signal genes.

Random noise. Let rn be the parameter that determines the amount of noise in the data, which is chosen by the user (by default, it is set to 20%). The simulation algorithm takes a random value δ_{ij} from a uniform distribution between -rn and rn to be used as random noise. Then, the amount of random noise for an expressed gene is equal to δ_{ij} times G_{ij}, because it is expected that the effect of this kind of noise is proportional to the level of expression. For non-expressed genes, with G_{ij}=0, the amount of random noise is a random value λ_{ij} generated from a normal distribution with mean 0 and standard deviation sd (a parameter that can also be set by users, by default sd=0.3). If, by chance, the expression value for a signal gene is 0, the random noise for that gene will also be computed this way.

Structural noise. The algorithm considers that structural noise only affects signal genes. For a given sample j, structural noise is computed as a random value ε_j from a normal distribution with mean m and standard deviation sd and this value is equal for all the genes in the same sample. The parameter sd is the same used for random noise in non-responsive genes and the mean m is calculated as $rn \times msv$, where rn is the random noise parameter defined above and msv is the expected value for gene signal expression. As previously mentioned, signal genes expression level can be 2, 3, 4 or 5 with probabilities 0.5, 0.3, 0.1, and 0.1 respectively. Therefore, msv turns out to be 2.8.

Hence, once the "pure" gene expression value G_{ij} is generated, the "observed" gene expression value G^*_{ij} is computed by the algorithm according to the following equations:

$$\begin{aligned} G^*_{ij} &= G_{ij} + \delta_{ij} \times G_{ij} + \varepsilon_j; \quad \text{if } G_{ij} \neq 0 \\ G^*_{ij} &= G_{ij} + \lambda_{ij}; \qquad\qquad\quad \text{if } G_{ij} = 0 \end{aligned} \tag{3.14}$$

where $\delta_{ij} \sim U[-rn, rn]$; $\lambda_{ij} \sim N(0, sd)$; and $\varepsilon_j \sim N(m = rn \times 2.8, sd)$.

3.2.4 Performance indicators

The variable selection methods previously described are to be compared on simulated data in order to assess their performance. Thus, we need to define some indicators to measure this performance. A variety of indicators to assess the classification of features into two groups (e.g. activated/non-activated, differentially expressed/non-differentially expressed, etc.) have been described in the literature. These two groups will be named as positives and negatives. The confusion matrix in Table 3.3 illustrates the potential classification errors. The performance indicators are then defined according to these successes and failures.

Table 3.3: Confusion matrix for two-classes classification

		Actual values	
		Positives P	Negatives N
Predicted values	Positives P'	True Positives TP	False Positives FP
	Negatives N'	False Negatives FN	True Negatives TN

These are the performance indicators we will use in this chapter:

- **Sensitivity** (SE), which is also known as the True Positive Rate (TPR):
  $SE = \frac{TP}{TP + FN} = \frac{TP}{P}$
- **Specificity** (SP) or True Negative Rate (TNR). The False Positive Rate (FPR), which is the performance indicator used in Receiver Operating Characteristic (ROC) curves, is equal to 1-SP.
  $SP = \frac{TN}{TN + FP} = \frac{TN}{N}$
- **False Discovery Rate** (FDR), which is the percentage of FP over the total number of detections:
  $FDR = \frac{FP}{TP + FP} = \frac{FP}{P'}$

- **Matthews correlation coefficient** (MCC) [94]. Can take values from -1 to 1, where 1 would indicate a perfect match and -1 an inverse prediction. This coefficient takes into account every type of classification error and is especially appropriate when the size of both groups is very different, which usually happens in variable selection problems in transcriptomics.
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{P' \times P \times N' \times N}}$$

3.3 Results

The variable selection methods described in Section 3.2.2 were first evaluated on simulated data in several comparative studies. According to the results of these comparisons, the best strategies were determined and applied to an experimental dataset.

3.3.1 Simulated data

Simulation studies were conducted, on the one hand, to compare the performance of the proposed variable selection methods and, on the other hand, to see which methods are preferred under certain biological scenarios or which ones are less affected by the biological characteristics of the data.

To simplify the interpretation of the results, for each simulated dataset, only two factors (e.g. time and experimental group) were considered: the time factor consisting of three time points, and the number of experimental groups that was also three. Four replicates were generated for each experimental group at each time point. Two different simulation experiments were conducted (see Simulation experiment 1 and Simulation experiment 2). The first experiment was used to compare different options in each selection method and to determine a good range for parameter values. Next, a global comparison of the best combinations for each method was carried out on the second simulation experiment to obtain a more precise selection approach benchmarking.

Simulation experiment 1

The biological scenarios to be simulated for this first experiment were defined by the values of the following parameters:

- Number of genes in the dataset (N): 3000, 15000, or 30000.
- Percentage of differentially expressed genes (responsive or signal genes) with regard to the total number of genes ($\%deg$): 1%, 5% or 15%.
- Number of gene classes ($class$): 5, 10 or 25. Genes in the same class have the same expression time pattern under the same experimental group.
- Level of noise in the data ($noise$): 10% or 30%.

These parameters define 54 different biological scenarios, and 10 datasets were generated for each of them. We made several comparisons on these 540 datasets whose results are described in the following sections.

Comparing resampling strategies

As already mentioned in Section 3.2.1, the variable selection strategies implemented in ASCA-genes were Box's method for SPE, and resampling techniques for leverage. However, these approaches were not efficient in separating "interesting" from "uninteresting" genes in large dataset scenarios. Therefore, a complete study was designed to determine which biological scenarios these selection strategies failed in, and to compare the three different resampling options to calculate the leverage threshold (see Table 3.1) at significance levels of 0.01 and 0.05. Box's method was maintained to compute the SPE threshold, as in the ASCA-genes paper.

Hence, the ASCA model was obtained for each of the 540 simulated datasets and these variable selection strategies were applied. The Matthews Correlation Coefficient (MCC) was obtained in each case and the results were analyzed by means of an ANOVA model with repeated measures [10] to evaluate the effect of the biological factors indicated above, the resampling strategy ("leverage method") and the significance level (α) on MCC

values. An ANOVA with repeated measures was used because the variable selection methods were applied to SPE and leverage values obtained from the same simulated datasets, so the measurements were not independent in this sense. The ANOVA results indicated that factors with a significant effect on MCC (p-value$<$0.002) were: leverage method, significance level, number of signal gene classes ($class$) and percentage of signal genes ($\%deg$). The noise level and the number of genes had no statistically significant influence on MCC (p-value$>$0.6). Post-hoc tests showed that the best MCC results (p-value$<$0.001) were obtained for leverage method 3, i.e. permuting genes and computing threshold as a global percentile; α=0.01; low number of signal genes classes and medium signal genes percentage (5%). See Figure 3.2. We also observed that for α=0.01, the real False Positive Rate (FPR) obtained with any of the resampling methods was similar to the significance level, but when setting α to 0.05, FPR reached 80% in some cases. Classification failures were mainly due to the strategy used to calculate SPE threshold (Box's method).

Comparing selection methods for SPE

In the second study on these simulated datasets, Box's method was compared to the other SPE parametric methods: Jackson & Mudholkar's and Gamma. In this case, twelve different significance levels were evaluated, varying from 0.001 to 0.1. No leverage thresholds were calculated, so gene selection was based only on SPE values. Consequently, MCC results can be used to compare SPE methods, but not to measure the global performance of the methods. As shown in Figure 3.3, when the significance level is around 0.03, all three methods perform similarly. For the rest of the significance levels, Box's method produces much worse results than the other two, which behave similarly.

An ANOVA model with repeated measures showed that the SPE method, significance level and all the biological factors had a statistically significant effect on MCC (p-value$<$0.001), except the number of genes (p-value$>$0.7).

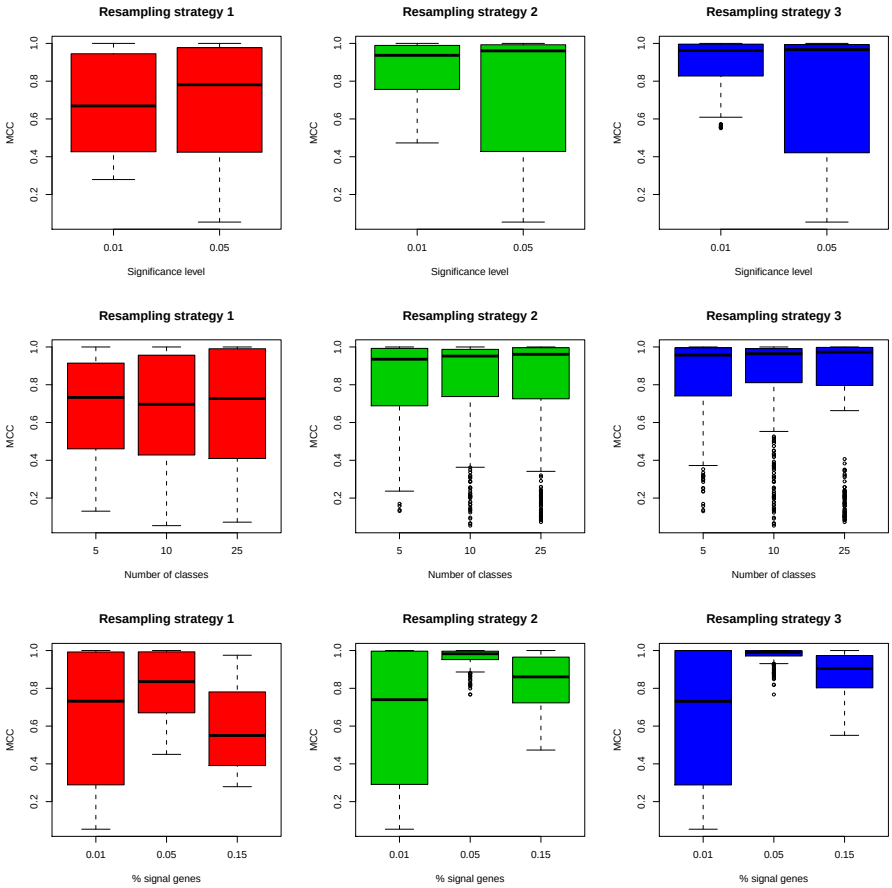


Figure 3.2: Resampling strategy performance (measured by MCC) according to the significance level, the number of signal gene classes and the percentage of signal genes.

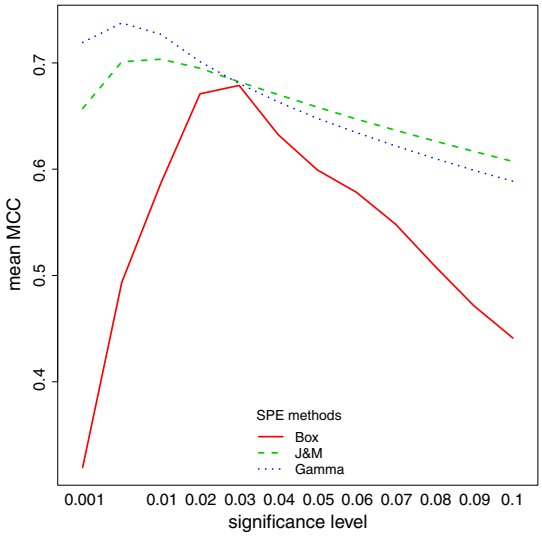


Figure 3.3: SPE selection method performance (measured by MCC) according to significance level.

From post-hoc tests, it was deduced that SPE methods were significantly different (p-value$<$0.008), and that Jackson & Mudholkar and Gamma methods produced the best results. For significance levels between 1% and 3% the best MCC results were obtained (p-value$<$0.001). No statistically significant differences were observed between 5 or 10 signal gene classes (p-value$>$0.3), but significantly better results were obtained when number of classes was 25 (p-value$<$0.001), maybe because when so many different patterns are present in the data, there are more genes badly explained by the model and hence those genes have a high SPE value. The best MCC results were obtained when the percentage of responsive genes was 5%, followed by 15%, and lastly 1% (p-value$<$0.001). Finally, MCC was higher when the noise level was 30% (p-value$<$0.001).

Joining the results of this study, we determined the most convenient significance levels for each method to obtain the best MCC value, despite the signal gene percentage or the number of signal gene classes. Our recommendations were α=0.03 for Box's method, 0.005$<\alpha<$0.02 for Jackson & Mudholkar's

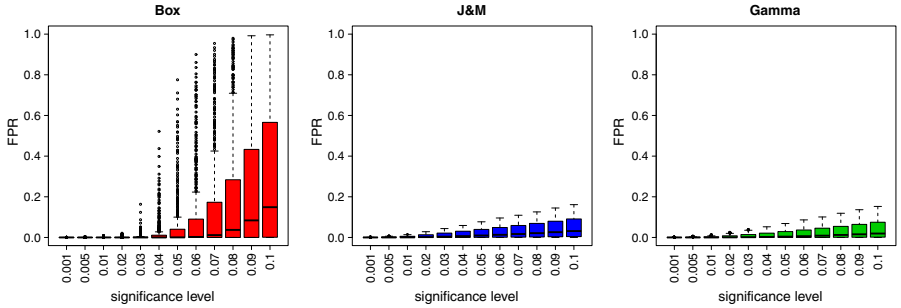


Figure 3.4: The FPR according to the significance level for each one of the SPE methods studied.

method, and α=0.01 for Gamma approximation (see Figure 3.3).

Again, the significance level was compared to the False Positive Rate (FPR) obtained for each method. Figure 3.4 presents these results and shows that in Gamma and Jackson & Mudholkar's methods this relation was preserved, while this did not happen for Box's method.

Comparing minAS to the other methods

Once the methods estimating the null distribution were compared, we included the minAS method in the study (always taking the first local minimum after the highest peak as the threshold for both SPE and leverage). To see if minAS selection was sufficiently satisfactory to continue studying the method in depth, it was compared to the combinations of methods evaluated in the first study (Box's method for SPE and resampling techniques for leverage). In this preliminary comparison, default options in R "density" function (Gaussian kernel and "nrd0" method to compute bandwidth) were used. As shown in Figure 3.5, the MCC obtained from minAS was, in general, higher than the MCC obtained with the other methods.

In addition, using the same simulated datasets, the default options in minAS (Gaussian kernel and "nrd0" bandwidth computing method) were compared to the best estimators according to minAS options "max" and "mean" (see Section 3.2.2). Figure 3.6 shows that minAS resulted in better MCC

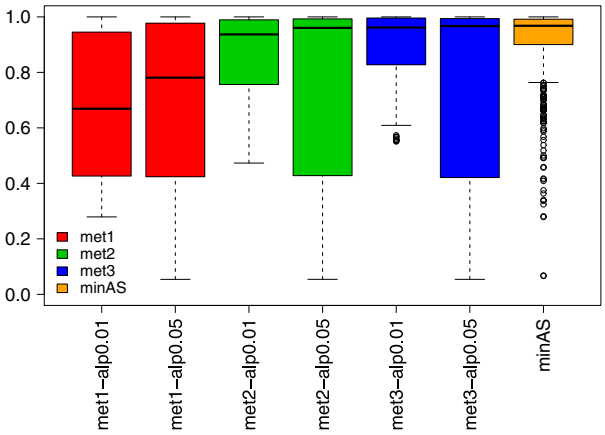


Figure 3.5: MCC obtained by applying the three resampling methods in Table 3.1 with α=0.01 and α=0.05 for leverage and Box's method for SPE, and minAS method for both of them.

scores when using the default KDE than with the KDE producing the minimum maximum or minimum mean distance to the empirical data distribution. The reason for this is that the other kernels or methods to compute bandwidth tended to generate infra-smoothed curves with too many local minima. In these cases, the selection by the first local minimum increased the number of false positives. Therefore, default "density" options were used when applying the minAS procedure hereinafter.

The influence of biological parameters defining the scenarios on MCC results for minAS method was also analyzed using an ANOVA model. All the parameters had a statistically significant effect on MCC (p-value$<$0.01), especially the number of genes, the number of classes and the signal genes percentage, as well as the interactions between them. It was observed that the greater the number of classes and the percentage of signal genes, the better MCC results minAS produced, no matter the number of genes. As the number of genes and signal percentage increased, MCC was less dependent on the number of classes. Boxplots describing these results can be seen in Figure 3.7.

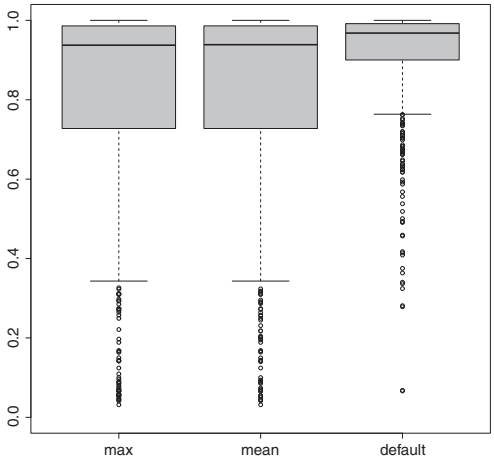


Figure 3.6: MCC obtained by applying minAS for both SPE and leverage on datasets from simulation experiment 1, considering three options: "max" criterion, "mean" criterion, and default options.

Hence, as general guidelines, we recommend using minAS for datasets with a high number of variables because otherwise the goodness of fit of KDE is not guaranteed and the multimodality is more dependent on the value of the smoothing parameter. The method can be applied to datasets with approximately a thousand variables, but results show that the best performance is obtained for more than 15000 variables. A Gaussian kernel and the method "nrd0" to compute bandwidth have been proven to offer the best minAS performance. Furthermore, increasing the parameter "adjust" to get a more smoothed KDE produces even better results (as shown in the following section), although this parameter was not changed in any of the simulation experiments we performed.

Simulation experiment 2

To conclude the evaluation of variable selection methods on simulated data, a new simulation experiment was designed in order to compare simultaneously all of the previously described methods for computing SPE and leverage thresholds. In this last comparison, other biological scenarios were simulated

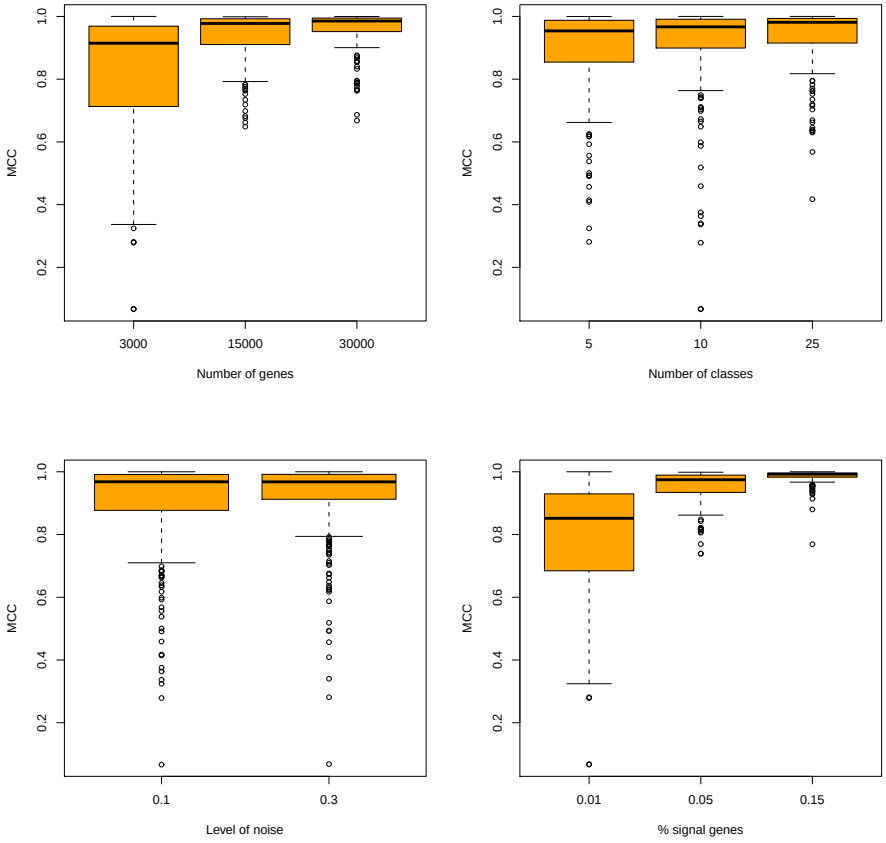


Figure 3.7: Performance of minAS (measured by MCC) according to the number of genes, the level of noise, the number of signal gene classes and the percentage of signal genes.

taking into account the results obtained in the previous studies. The level of noise was not included as a parameter in these simulations because it had, in general, very little influence on MCC results, so it was set to 20%. The values for the rest of biological parameters were:

- Number of genes in the dataset: 5000 or 20000.
- Percentage of responsive genes: 3% or 10%.
- Number of gene classes: 5 or 25.

For each one of the 8 possible scenarios, 10 datasets were again generated. The SPE selection methodologies to be compared in this analysis were Box's method, Jackson & Mudholkar's (J&M), Gamma, minAS, and resampling using permutation strategy 2 (genes permutation) and option (b) to compute threshold by global percentile (Permut2b). Regarding leverage, we compared the resampling method (Permut2b), Gamma approximation and the minAS method. The resulting combinations of all these methods are shown in Table 3.4. The significance level that produced the best results in the previous studies was chosen.

Figure 3.8 shows 95% confidence intervals for the mean MCC produced by each of these methods. The overall good performance of the methods can be deduced from this plot, because all of them obtained a mean MCC higher than 0.9. However, the ANOVA model with repeated measures showed a statistically significant difference between them (p-value$<$0.001). The worst results were obtained for those combinations in which resampling techniques were used to compute the SPE threshold. Box's method for SPE is not recommended for its high standard deviation. The Gamma approximation for leverage worked excellently. Considering both MCC mean and standard deviation, the best combinations were number 6 (J&M+Gamma), number 8 (Gamma+minAS) and number 9 (Gamma+Gamma). The ANOVA model also showed that the number of genes and the number of signal genes classes had no significant effect on mean MCC value (p-value=0.137 and

Table 3.4: Selection methods combinations included in global comparison.

Combination	*SPE method*	*Leverage method*
1	Box - α=0.03	Permut2b - α=0.01
2	Box - α=0.03	minAS
3	Box - α=0.03	Gamma - α=0.01
4	J&M - α=0.01	Permut2b - α=0.01
5	J&M - α=0.01	minAS
6	J&M - α=0.01	Gamma - α=0.01
7	Gamma - α=0.01	Permut2b - α=0.01
8	Gamma - α=0.01	minAS
9	Gamma - α=0.01	Gamma - α=0.01
10	minAS	Permut2b - α=0.01
11	minAS	minAS
12	minAS	Gamma - α=0.01
13	Permut2b - α=0.01	Permut2b - α=0.01
14	Permut2b - α=0.01	minAS
15	Permut2b - α=0.01	Gamma - α=0.01

p-value=0.353, respectively). However, signal genes percentage significantly affected the MCC value (p-value<0.001), as well as the interaction between signal gene percentage, and method combination (p-value<0.02). In general, the higher signal genes percentage, the higher the mean MCC. Combination 9 (Gamma+Gamma) did not result in big differences in mean MCC for the different percentages of signal genes. However, some combinations including minAS, for example numbers 10, 11, and 12, worked much better when the percentage of signal genes was higher.

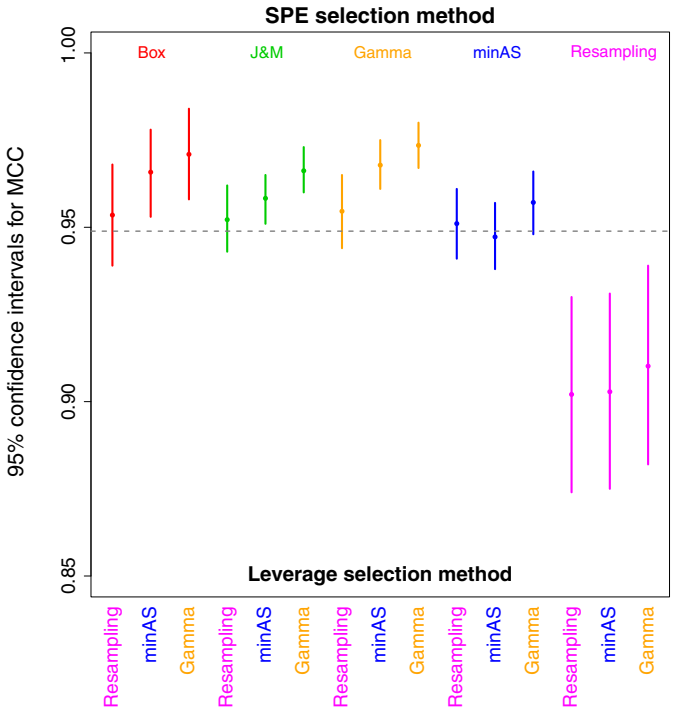


Figure 3.8: 95% confidence intervals for mean MCC according the combination of methods used. Horizontal dashed-line corresponds to overall average MCC.

In all the simulation studies, the bandwidth was computed following Silverman's rule ("nrd0" option). To check what happened if the bandwidth was modified with the "adjust" coefficient, minAS was applied to the 80 simulated datasets in simulation experiment 2, using the default options in "density"

and varying the coefficient "adjust" from 0.5 (i.e., half the bandwidth obtained by the "nrd0" method) to 5 (i.e., 5 times the bandwidth obtained by "nrd0" method). The MCC results for each "adjust" value are displayed in Figure 3.9. Interestingly, minAS performance improves for "adjust" values higher than one, that is, when the estimated density curve is more smoothed. It is expected that for a certain value of "adjust" coefficient not considered in the study, performance gets worse, because the density curve would be so smoothed that the distribution would become unimodal. Moreover, if the bandwidth is decreased, the curve is not smooth enough and may yield false local minima. As shown in the plot, for an "adjust" coefficient of 0.5, there are MCC values close to 0, and the MCC first quartile is less than 0.8. The important conclusion is that the default minAS options provide a good estimation for SPE and leverage density curves. Increasing the bandwidth can improve gene selection, but smaller bandwidths result in over-adjustment and in the occurrence of false local minima, leading to an increase of the number of false positives.

To summarize, minAS and Gamma approximation (with α=0.01) behaved slightly better than the rest of the studied methods. Furthermore, the Gamma method presented fewer differences in MCC values for different signal gene percentages, while minAS had a better performance when this percentage was higher.

3.3.2 Experimental data: Hypoxia

Once the benchmarking with simulated data was completed, the methods producing the best results were applied on an experimental dataset and evaluated for their ability to select genes that led to outstanding biological information. The Hypoxia gene expression data in [114] was used for this biological validation. This dataset collects the transcriptomic profile of human embryonic stem cells cultured under different oxygen concentrations. The oxygen conditions were: normoxia (21% oxygen) and hypoxia (5% or 1% oxygen). Gene expression for 30826 genes was measured at several time points using Agilent

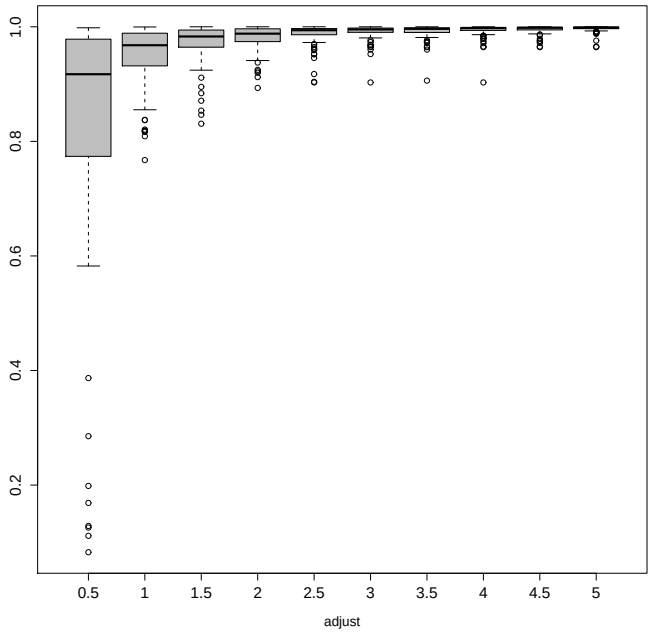


Figure 3.9: Matthew's correlation coefficient obtained in simulated datasets according to different values for the coefficient "adjust" (that multiplies the bandwidth obtained by the "nrd0" default method) when applying the minAS selection strategy for SPE and leverage values from the ASCA model in datasets from simulation experiment 2.

Table 3.5: Number of genes selected by the method combinations studied in the hypoxia dataset.

Combination	*SPE method*	*Leverage method*	*Sub-model* a	*Sub-model* $b+ab$	*Total*
5	J&M	minAS	1347	1827	2668
6	J&M	Gamma	1182	1919	2618
8	Gamma	minAS	1287	1076	2034
9	Gamma	Gamma	1122	1176	1976
11	minAS	minAS	1862	1309	2705
12	minAS	Gamma	1706	1405	2649

microarrays. An ASCA model was fit to the data. Factor a is the time (0 hours, 12 hours, 24 hours, 5 days and 10 days) and factor b was used for the oxygen level (21%, 5% and 1%). The oxygen level and interaction effects were joined together in the model (as in Equation 3.3). Two principal components were selected in each submodel (a and $b+ab$), which explained 83.2% of the variability in submodel a and 71.9% in submodel $b+ab$. Model analysis showed different gene behaviors for each oxygen level, clearly differentiating normoxia from hypoxia conditions, and time points 12-24 hours from 5-10 days (results not shown). In order to compute SPE and leverage thresholds, several combinations of selection methods showing the best performance in the previous simulation studies were used: Jackson & Mudholkar's SPE method (α=0.01), Gamma approximation (α=0.01) and the minAS method. Table 3.5 shows the number of genes selected by each one of these combinations and Figure 3.10 shows the histograms and distributions fitted for SPE and leverage values in each submodel, as well as the thresholds obtained by the selection methods.

To evaluate the validity of the different variable selection methods, selected gene lists were investigated to see whether the biological information they contained was relevant to the study. Hence, for each one of the selected

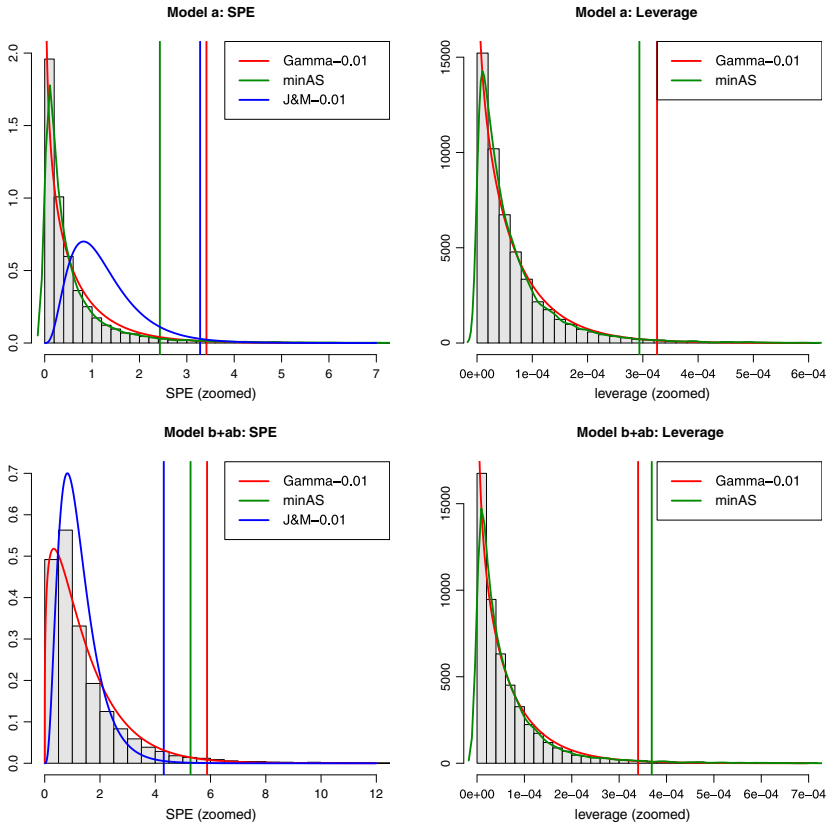


Figure 3.10: Histograms for SPE and leverage in each submodel. Curves represent the distributions fitted by the variable selection methods applied. Vertical lines are the thresholds computed from these distributions. The X-axis has been zoomed for better visualization and therefore they do not show the full range of SPE and leverage values.

gene sets, we carried out a functional enrichment (FE) analysis by means of the FatiGO tool, included in Babelomics suite [2], using the Gene Ontology (GO) gene function annotation to compare selected versus non-selected genes. FE is an established methodology to interpret and evaluate transcriptomic data, that assesses whether specific cellular functions (in this case, GO terms) are overrepresented within the set of significant genes. Significant enriched GO terms for the selected genes sets were visualized with the Blast2GO software [28], that allowed them to be colored, depending on the number of selection methods by which they had been detected. This kind of graph enabled us to evaluate which of the selected gene sets contributed more to the biological interpretation of the experimental results (see an example in Figure 3.11).

In general, all of the tested methodologies generated gene selections enriched in a number of GO terms that represent key general processes of the hypoxia treatment. These were, among others, "developmental process", "metabolic process", "response to stimulus", "transcription factor activity", "chemokine receptor binding", "lipid transport activity", "immune system process", "intrinsic to plasma membrane", "organ morphogenesis", "angiogenesis", "response to wounding" and "humoral immune response". "Organ morphogenesis" and "angiogenesis" refer to the establishment of the circulatory system in mammals, one of the first events during the embryo development [22, 75]; while the metabolism of lipids has also been postulated to play an important role in the embryo differentiation [56]. Also "ectoderm development" and "epidermis development" were functions identified by most of the methods, and are directly related to the differentiation process analyzed in this experiment. Additionally, some specific processes were only revealed by some of the selection methods. For example, combinations 5 and 6 (both using J&M method for SPE selection) highlighted the "central nervous system development" (associated to normoxia) and "sensory organ development" or "chemokine receptor binding" (both related to hypoxia). Combinations 11 and 12 (SPE selection by minAS) discovered metabolic processes such

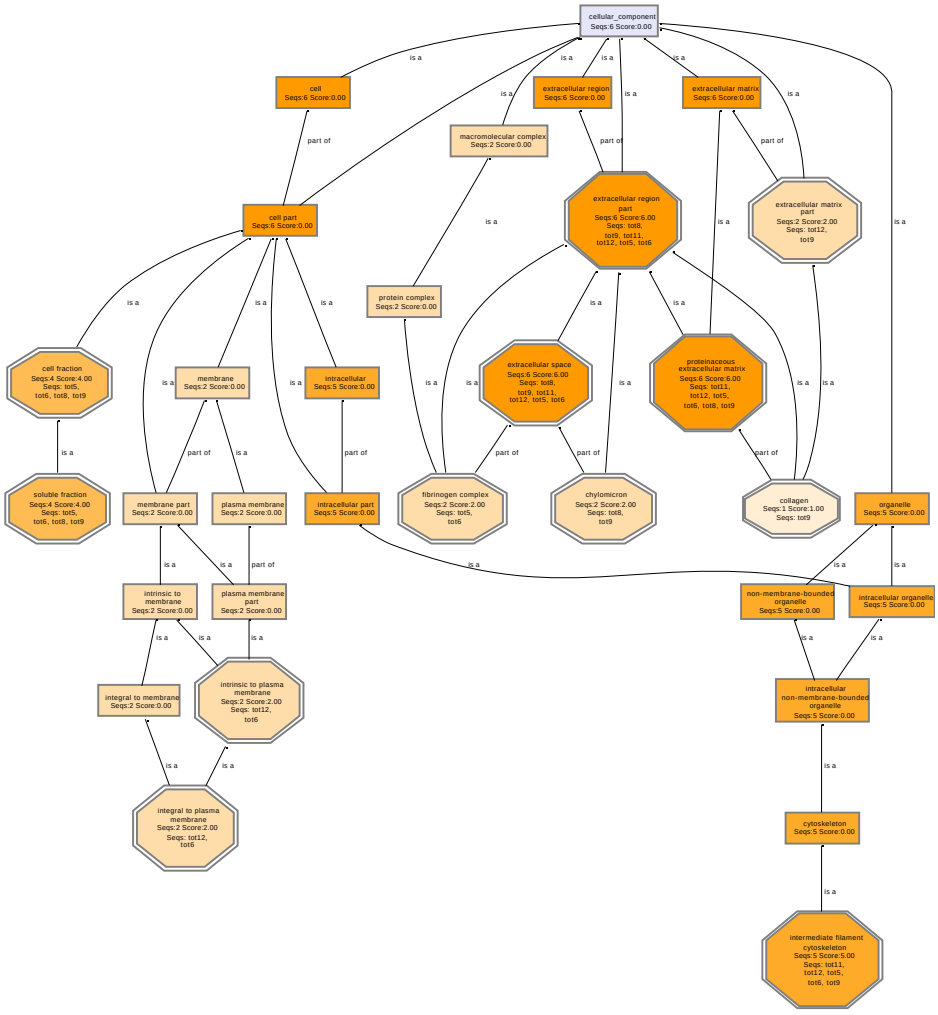


Figure 3.11: Example of a Blast2GO graph displaying enriched GO terms detected by the FatiGO tool (adjusted p-value < 0.05) when comparing genes selected by each of the method combinations against the others.

as "hormone metabolic process", "hexose metabolic process" and "glucose metabolic process". Combinations 9 and 12 (leverage selection by Gamma) found the "extracellular matrix part" GO term, which plays a fundamental role in regulating remodeling processes in embryo development and is also involved in repair processes, inflammation and tumor invasion [65].

In summary, most of the biological information is shared by all the compared methods combinations, but not all of them contribute equally to improving our biological knowledge about the gene products dynamics in this context. Each of the combinations leads to the extraction of some particular biological functions than the rest of the methods do not detect or, at least, not with the same degree of specificity.

3.3.3 Other applications

Some of the variable selection strategies described here have also been applied in other multivariate contexts different from ASCA method.

In [31], multi-way projection methods such as Tucker3 and N-PLS were used for the integrative analysis of a functional genomics dataset where transcriptomics, metabolomics and physiological data were available from an experiment on rats to assess the effect of hepatotoxicant bromobenzene. The most relevant biological features were selected by applying minAS to either the loadings corresponding to one of the model components or the projection of several component loadings into the line that best separated the experimental conditions.

In [113], the PANA computational methodology is introduced. It studies the functional interconnections among the molecular elements of a biological system by using high-throughput genomics measurements and a functional annotation scheme. PCA is applied to extract an activity profile from each functional block -or pathway. Next, machine-learning methods infer the relationships between these functional profiles to obtain an interconnected network of pathways that represents the functional cross-talk within the molecular system. minAS is again used to identify the main gene contributors to

pathway profiles computed from PCA models also taking the loadings as the variable importance measurement. We showed the benefits of the PANA approach to describe the functional transcriptional connections during the yeast cell cycle and to identify pathways that change their connectivity in a disease condition using an Alzheimer example.

3.4 Discussion

In this chapter, we have presented and compared several strategies to select the most relevant genes in multivariate models applied to the analysis of complex genomic data. The starting point of this contribution is the adoption of a multivariate dimension reduction strategy, commonly used in data exploration for the identification of important genes. In comparison to univariate methods that carry out gene-wise analysis, the multivariate approach exploits the coordinated nature of gene expression and avoids the application of multiple testing corrections that seriously diminishes statistical power in genomic research. In these scenarios, two additional factors are also important. Firstly, the high-dimensionality of the feature space, that results in data structures where the number of variables can be two or three orders of magnitude the number of observations. And second, the low signal to noise ratio of the measurements. This implies that traditional multivariate feature selection methods are generally not applicable. The basic contribution in this paper is that the variable selection choices we propose always involve studying the distribution of the statistics used to measure the importance of the variables. Hence, the threshold for these statistics is set according to the shape of the distribution rather than selecting a fixed percentage of the total number of variables, which is a common and rather arbitrary practice in this kind of analysis. Our main concern in this study was to identify methodologies that will generally work well in different scenarios of dataset size, diversity of gene expression signals and levels of noise, since these features are not normally fixed by the experimentalist. We also tried to gather selection methods with an easy implementation and comprehension, as we understand that

variable selection is only a small part of a genomic study and researchers may need quick but consistent solutions. In this work, some methods were taken from the literature (Box, Jackson & Mudholkar) or adapted to be used in this context (resampling), while others are novel proposals (minAS, Gamma approximation). These variable selection methods were first compared on simulated datasets to evaluate which ones presented the best performance and to quantify the influence of some biological data features on the goodness of the selection. In general, Gamma and minAS methods showed the best behavior for both SPE and leverage thresholds computation, as well as Jackson & Mudholkar's method for SPE. It was also seen that the higher the percentage of signal genes or the number of genes are, the better minAS performance is, while Gamma approximation is not significantly affected by these biological parameters, therefore making it a more robust methodology. However, modifying minAS default options (such as increasing the smoothing parameter) improved the performance of this method. The application of these three approaches on a real experimental dataset verified their usefulness for selecting relevant genes. In all cases, relevant biological conclusions could be obtained on the gene selection provided by the different methods, although specific biological functions were differentially uncovered by each approach. Interestingly, the major differences in gene selection and functional enrichment were the result of the method choice for the SPE statistic, while leverage seemed to be more robust for the statistical model applied. This result is interesting as the SPE measures the deviation of each gene from the general multivariate model. Differentially expressed genes that follow a minority expression pattern tend to have high SPE values [103]. Our results indicate that selection on this part of the signal is also biologically relevant.

It should be outlined that the conclusions of this chapter are based on the simulation studies performed and might not be valid outside the biological scenarios analyzed. However, since the simulation algorithm was carefully designed to mimic real datasets and a vast variety of scenarios was considered

(comprising more than 600 datasets), we believe that the results are generally valid for most multifactorial gene expression experiments.

Finally, we focused on multifactorial designs because the variable selection issue has not yet been sufficiently developed for these complex experimental setups. The ASCA-genes framework was chosen to model these data, since it is considered a suitable methodology for the analysis of genomic datasets with such experimental designs. However, as the proposed variable selection methods are based on modeling the distribution of multivariate statistics, they are generally applicable to different dimension reduction techniques and kind of data by changing the statistic measuring the importance of the variables in the model. In fact, we have successfully applied our methods in other contexts, as for example in [31], where minAS was used for selecting variables from genomic and metabolomic data in Tucker3 and N-PLS models, and in [113], where minAS is again used to identify the main gene contributors to pathway profiles computed from PCA models taking the loadings as the variable importance measurement.

The minAS and Gamma variable selection methods applied to the ASCA-genes analysis have been implemented in the web suite for Serial Expression Analysis, SEA (http://sea.bioinfo.cipf.es/) [102], which is freely available to the scientific community.

Chapter 4

RNA-seq data quality control

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A.
Differential expression in RNA-seq: A matter of depth
Genome Research, 21:2213-2223, **2011**

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, and Conesa A.
Qualimap: evaluating next-generation sequencing alignment data
Bioinformatics, 28(20):2678-2679, **2012**

Tarazona S, Furió P, Turrà D, Di Pietro A, Ferrer A, and Conesa A.
NOISeq: An R package for visualization, quality assessment and differential expression for RNA-seq experiments
(In preparation)

4.1 Introduction

RNA-seq is a recently emerged high-throughput sequencing technology which is increasingly being applied to quantify gene expression levels. With RNA-seq, no previous knowledge of the genome is required, and in addition to estimating expression levels, it also allows structural analysis of the transcriptome: alternative splicing, 3' UTR regions, novel splice junctions, antisense regulation, intragenic expression, etc. [91, 124].

At the dawn of RNA-seq applications it was claimed that this technology would produce unbiased, ready-to-analyze gene expression data. However, the reality has turned out to be very different. It is now generally admitted that it collects a number of biases and that accuracy at the low expression level is still limited [77, 99]. RNA-seq technology boasts a generally high level of data reproducibility across lanes and flow-cells, which reduces the need for technical replication within these experiments [92, 99], but neither data processing nor experimental design are straightforward. While microarrays provide gene expression measurements in a format that can be directly analyzed by the researcher, the information generated by sequencing platforms must be carefully processed to obtain these expression estimations. There are many steps between biological sample collection and expression quantification, including RNA isolation, library construction, sequencing, read alignment, etc. Unfortunately, the files generated at each of these processing steps frequently contain biases that are introduced by the sequencing technology [124], during sample preparation [60, 98] and/or by the selected mapping algorithm [45]. Therefore, when analyzing sequencing data performing quality control at each step to get an idea of how reliable the data are, and how well they fit with the expected outcome is a fundamental requirement. However, despite the many procedures that have been developed to reduce the noise at each one of these processing steps and to control the quality of the data generated [8, 49], the technology is still far from being perfect and therefore it is absolutely necessary to be able to detect potential biases once the expression levels (read

counts) have been obtained, and crucially, before proceeding with any further analyses such as differential expression analysis.

An important trait of sequencing technologies that must be taken into account when designing an experiment is the amount of reads to be generated (sequencing depth). In RNA-seq, in particular, the more the target is sequenced, the more transcripts are identified and the higher the value of the expression level. Although most of the existing analysis methods address this issue by including a correction factor related to library size [18, 99], higher sequencing rates will presumably result in a more accurate estimation of the expression level and, concomitantly, the detection of significant changes in expression may be very much determined by the sequencing depth. Inevitably this leads to the question of how many reads should be generated in an RNA-seq experiment to obtain robust results. Some reports suggest that, in a mammalian genome, ~ 700 million reads would be required to obtain accurate quantification of more than 95% of expressed transcripts [14]. Knowledge of the relationship between sequencing depth and feature detection is needed for experimental design purposes and for understanding the characteristics of the analysis results. Hence, the number of replicates and the sequencing depth at which one should sample remains an important question to be answered when designing an RNA-seq experiment [19].

Another issue that must be faced when dealing with analysis of short reads is that the quantification of expression depends on the length of the biological features under study (genes, transcripts, or exons), as longer features generate more reads than shorter ones [106]. In the case of the Illumina RNA-seq platform, which is probably the most widely used, a guanine and cytosine nucleotide content (GC content) bias has also been reported [119]. Genomic sequences with either a high or low GC content are prone to underrepresentation in the sequencing outcome, which means that the corresponding estimated expression levels will be lower [12, 57]. Another drawback is the very nature of the sequencing technology, which is basically a sampling procedure from a population of transcripts, implying that differences in the relative

distributions of transcripts between samples affect the assessment of differential expression [15, 120]. Furthermore, the ability to detect and quantify rare transcripts is obscured by the wide dynamic range of mapped reads and the concentration of a large portion of the sequencing output in a reduced number of highly expressed transcripts. Hence, all these factors interfere in the linear relationship between transcript abundance and the number of mapped reads at a gene locus. The correction of these potential biases to make expression across samples or genes comparable is known as normalization, and it is therefore a substantial step in RNA-seq data processing. Different methods are available to tackle both within-sample and between-sample normalization [18, 38, 119, 154].

In this chapter, we will focus on quality control procedures to be applied on expression data, i.e. read count, which may be useful to explore the data before proceeding with further analyses. This exploratory analysis helps to gain knowledge about the biological characteristics of the detected features and also to choose an appropriate normalization method according to the potential biases observed in the data. We also discuss the normalization procedures to be applied in order to correct these biases and propose some methods to filter out the low count features since they may be inaccurate and may introduce noise into the analysis.

All these procedures have been implemented in the NOISeq R package, which is available in the open-source Bioconductor repository [50]. Some of them were introduced in [141] and are also available in the Qualimap software [49]. The functionalities of the NOISeq package are summarized in Figure 4.1 and were used to generate most of the results in this chapter and in chapter 5.

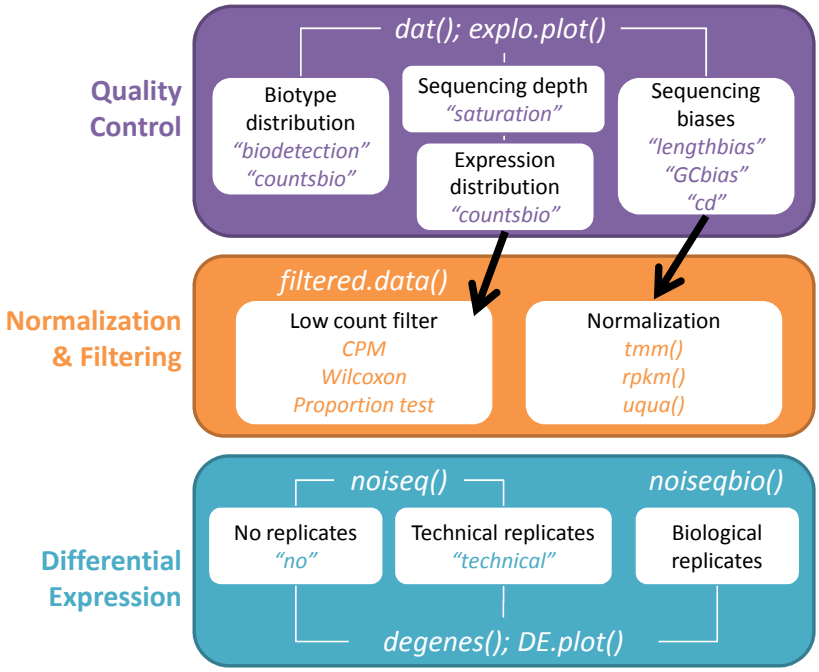


Figure 4.1: Outline of NOISeq package functionalities. Black arrows highlight that some quality control plots are used to take decisions on data processing.

4.2 Data

To illustrate the usefulness of the quality control tools described in this chapter, three experimental RNA-seq data sets were chosen. Two of them will also be used in the next chapter to assess the performance of the methods introduced there.

- **ENCODE data set**
  RNA-seq data from human monocytes (cell line $Cd14$) were obtained by the Cold Spring Harbor Laboratory for the ENCODE project [32]. Two different RNA extraction protocols were applied: PolyA+ extraction method (Pap) and PolyA- selection procedure (Pam). The sequencing

was done with an Illumina GAIIx platform and the reads were mapped to the reference genome (hg19 GRCh37) downloaded from UCSC [61] using TopHat v2.0.8 [71]. Gene expression was quantified using the HTSeq Python package version 0.5.3p3 [6], taking multi-hits into account by using an in-house script. This dataset will be referred to as **ENCODE**.

- **Fusarium oxysporum data set**

  Fusarium oxysporum is a soil-borne fungal pathogen that may affect a broad range of animals, from arthropods to humans, and also more than 100 plant species. It can produce localized skin or corneal infections in immunocompetent humans and frequently lethal infections in immunocompromised patients. This experiment was conducted by one of our collaborators on the "Transcriptional networks controlling virulence in filamentous fungal pathogens" project (TRANSPAT), which co-funded this work. The aim of TRANSPAT was to explore transcriptional networks which enable fungi to survive and proliferate in the mammalian bloodstream, given the increasing incidence of invasive fungal infections in immunocompromised patients, and associated mortality rates as high as 85-90%. Therefore, in order to study the infection mechanisms, the fungus was cultured in either minimal medium (MM) or human whole blood, and RNA was extracted from these samples and sequenced using SOLiD protocols. Two biological replicates were obtained for both blood and MM conditions and these were mapped to the reference genome from the Ensembl Fungi database (release 14) [47] using Lifescope software. CLC Bio tools were used to quantify the gene expression. This dataset will be referred to as **FO**.

- **Human prostate data set**

  These RNA-seq data were taken from the work of Ren *et al.* [118] and were publicly available at the SRA repository. In this work, samples of tumoral and healthy prostate that came from Shanghai Changhai

Hospital patients were sequenced. There were 11 biological replicates for the tumoral prostate condition and 12 replicates for the healthy prostate condition. The sequencing was done with an Illumina HiSeqTM 2000 and the reads were mapped to the *Homo sapiens* reference genome downloaded from Ensembl (release 68) [47] using TopHat 1.4.1 [143]. To quantify the gene expression the HTSeq Python package version 0.5.3p3 [6] was used. This dataset will be referred to as **HS**.

4.3 Quality control analysis

The quality control measures for read count data proposed in this chapter address three issues. First, biotype detection is used to obtain a global view of the type of genes detected in relation to the composition of the reference genome. Second, sequencing depth is analyzed to assess detection and quantification of the expressed genes. Finally, analysis of potential technical biases in the data is performed, such as the influence of gene length or GC content on expression or the differing RNA composition of the samples.

The NOISeq package offers the possibility of easily generating a quality control report with all these plots in PDF format (see an example in Appendix).

4.3.1 Biotype distribution

The Ensembl database [48] provides a biological classification of genes according to the role they play in transcription or translation. For instance, some of the biological groups include "protein-coding" genes (genes that code for proteins), "miRNA" (for microRNAs, which are small transcripts that can degrade messenger RNA in the cytoplasm), or "tRNA" (for transfer RNA, which serves as the physical link between the nucleotide sequence of DNA or RNA and the amino acid sequence of proteins), among others.

RNA-seq experiments may follow different RNA purification protocols to select specific target RNA species (i.e., long mRNAs or microRNAs) that may

be subjected to different levels of technical variation. Also, library preparation choices result either in data being stranded and hence allowing the identification of antisense expression, or ignoring the strand origin of the transcript. A standard protocol for an mRNA library preparation includes poly-A mRNA isolation, RNA fragmentation, and size selection from a gel. Therefore, transcripts should be polyadenylated and larger than the size selection cutoff (typically 200 bp) to be captured by the sequencing procedure. Polyadenylation signals are present in protein-coding genes but have also been identified in long-range, noncoding transcripts [23] and some snoRNAs[1] [54, 81]. The expression of pseudogenes is controversial, but reports indicate that these might be transcribed, giving rise to non-functional messengers in a tissue specific manner [153]. Furthermore, poly-A stretches might be present in retrotransposed pseudogenes[2] that originate from genome insertion events of these transcribed messengers [153]. Poly-A tails are also added to pri-miRNAs, nascent miRNA transcripts that undergo processing to reach the mature miRNA state [72]. Although pri-miRNAs can be long molecules, they are of transient nature and miRNAs are typically not captured by mRNA-seq library preparation protocols; alternatively, miRNAs embedded in the introns of coding genes could still be sequenced from partially processed transcripts. Other RNAs such as tRNAs, snRNAs[3], snoRNAs, and rRNAs[4] may undergo cytoplasmic polyadenylation to targeting them for degradation [7, 131]. Additionally, rRNA depletion usually precedes mRNA preparation and the presence of rRNA is considered as contamination in mRNA-seq experiments. In general, these small RNA species can be considered as non-targeted by the mRNA-seq procedure. Therefore, in RNA-seq experiments involving polyA selection, it is

[1] Small nucleolar RNAs

[2] Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Pseudogenes often result from the accumulation of multiple mutations within a gene whose product is not required for the survival of the organism. Although not protein-coding, the DNA of pseudogenes may be functional. (*Wikipedia*)

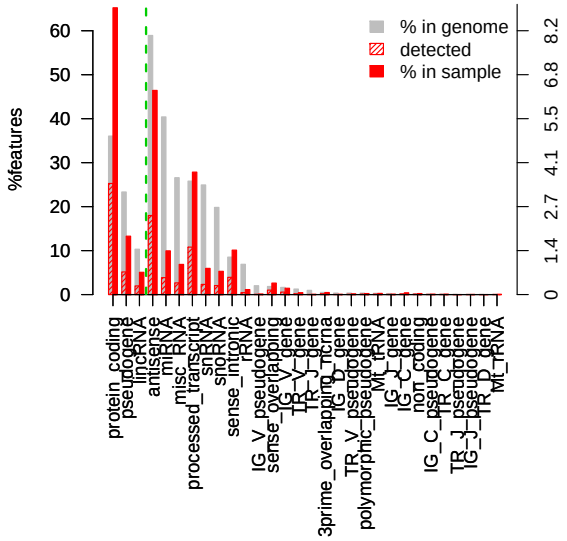
[3] Small nuclear RNAs

[4] Ribosomal RNAs

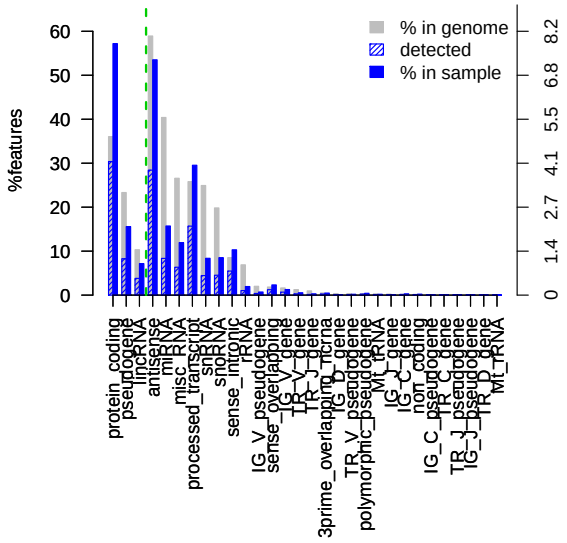
expected that most of the genes with mapped reads belong to the "protein-coding" category. The identification of other biotypes at proportions higher than in the reference genome might indicate inefficient mRNA purification or reveal new discoveries.

Thus, different experimental protocols may result in RNA-seq data having a non-uniform RNA composition that may not be directly comparable. This may be relevant, for example, when trying to combine data from different public sources to perform a joint meta-analysis or when technical variability with specific protocols is high. The biotype plots (Figures 4.2 and 4.3) are useful for determining which biotypes are present in the sample to provide an assessment on the homogeneity of samples within the data set. The "Biotype detection" plots (Figures 4.2a and 4.2b) show the proportion of genes within each biotype in the genome (gray bars), the proportion of them being detected in the biological sample with counts higher than 0 (red stripped bars) and which proportion represents each biotype within the genes detected in the sample (red solid bars). Bars on the left of the green vertical line correspond to the left Y axis, while bars to the right side are associated with the right Y axis (they are less abundant biotypes). Since the plots were generated for one of the conditions, the mean expression values (normalized to counts per million) across all samples for that condition were computed. The "Biotype expression range" box-plots (Figures 4.3a and 4.3b) show the range of expression levels in counts per million reads (CPM) within each biotype and help to decide if the expression quantification for the genes within a given biotype is good enough to perform further analyses on that biotype.

To illustrate the usefulness of these diagnostic plots we used them to compare RNA-seq samples generated with two different purification protocols (**ENCODE** data). By comparing the "Biotype detection" plots of both experimental procedures some differences are readily evident. The Pap protocol identifies a higher relative proportion of protein-coding genes (more than 60%) than the Pam protocol (around 55%). As a consequence, the second

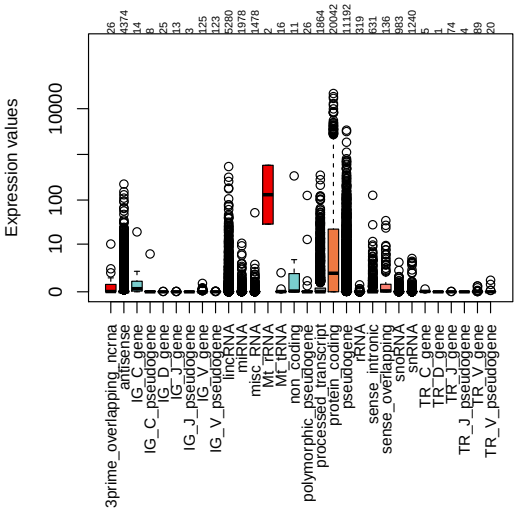


(a) Biotype detection. Pap protocol.

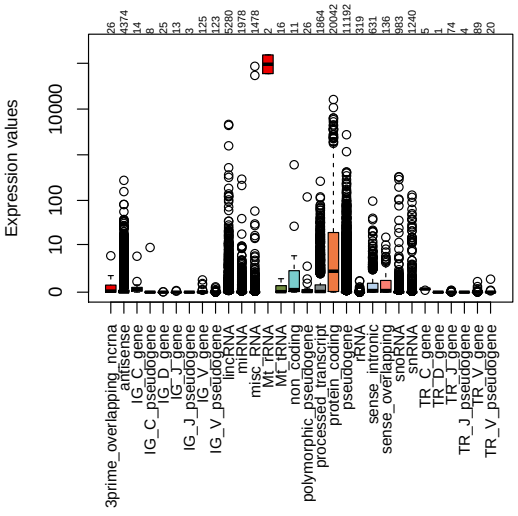


(b) Biotype detection. Pam protocol.

Figure 4.2: Biotype detection. **ENCODE** data. Grey bars represent the percentage of each biotype in the reference genome. Stripped color bars illustrate the proportion of genes in the genome, by biotype, that are detected in the sample. Solid color bars give the percentage of each biotype within genes detected in the sample.



(a) Biotype expression range. Pap protocol.



(b) Biotype expression range. Pam protocol.

Figure 4.3: Biotype expression range. **ENCODE** data. Expression values (Y axis) are given in counts per million of sequencing reads (CPM). Numbers in the upper part of the plot are the number of genes, by biotype, that are represented in each boxplot.

protocol had a relatively higher level of other RNA species such as pseudo-

genes, lincRNA[1] or antisense transcripts (Figures 4.2a and 4.2b). Differences in the relative percentage of detected biotypes also impact the quantification of the different RNA species, as revealed by the "Biotype expression range" plots (Figures 4.3a and 4.3b). The Pap protocol results in a wider dynamic range of expression for protein-coding genes than the Pam one. In contrast, the two Mt_rRNAs[2] accumulate a huge number of reads when using the Pam protocol (around $267,000$ and $64,000$ CPM, respectively) in comparison to Pap protocol (around $1,100$ and 230 CPM, respectively). These differences do not only affect the quantification of each transcript but may also have an effect on the results of any statistical analyses performed on them such as differential expression.

4.3.2 Sequencing depth and expression quantification

A key issue when analyzing RNA-seq data is to assess whether the available sequencing depth (total number of sequencing reads) provides sufficient coverage of expressed transcripts and an accurate quantification of gene expression. Alternatively, one may ask which sequencing depth is required to interrogate the transcriptome with good coverage and precision. It is generally admitted that genes detected by only a few reads are not reliably quantified and should be removed before further statistical analysis. The quality control plots described in this section are targeted to answer questions and provide solutions related to these matters.

The "Saturation" plot (Figures 4.4 and 4.5) indicates the number of detected genes in a biological sample (left axis) at the given sequencing depth (solid dot) and also at simulated higher and lower sequencing depths. The bars (right axis) show the new detections per each additional million sequencing reads. If more sequencing does not lead to a high number of new detections, then the saturation point has been reached and the sequencing will improve

[1] Large intergenic non-coding RNAs
[2] Mitochondrial ribosomal RNAs

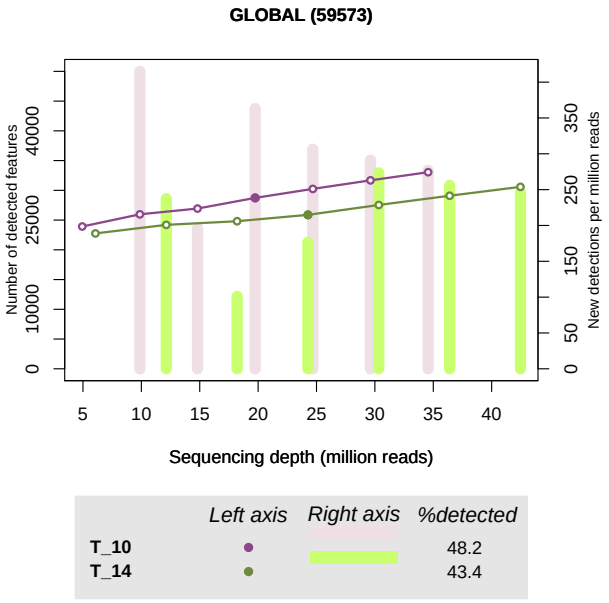
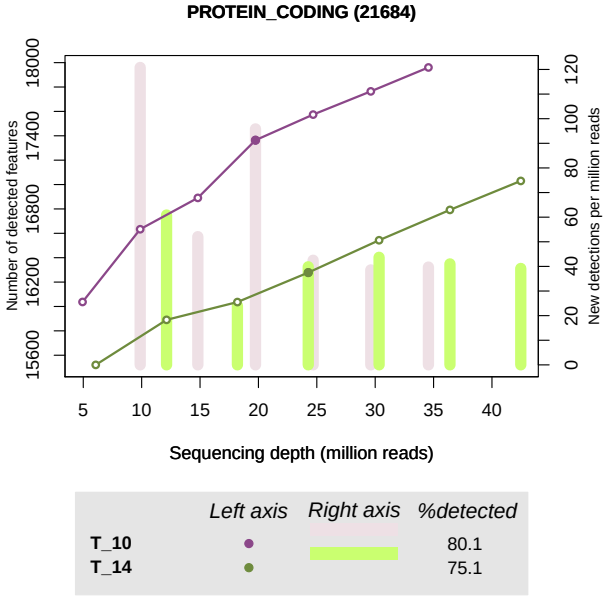


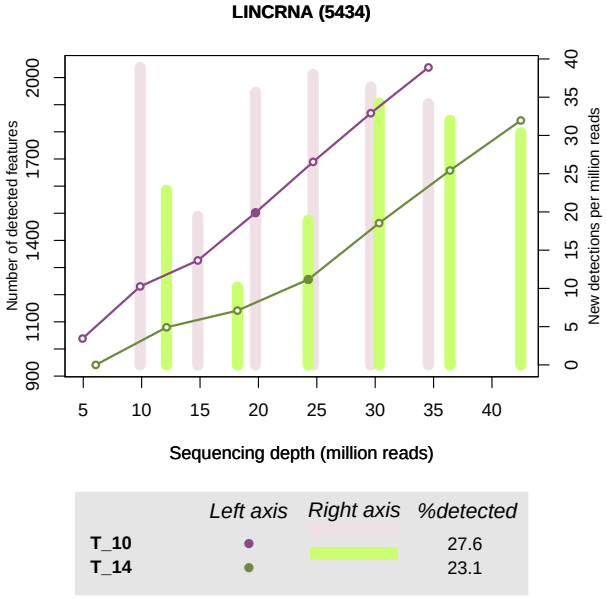
Figure 4.4: Saturation. **HS** data. Two of the tumoral samples are displayed. Genes from all biotypes are considered.

the quantification of the previously detected genes. The data for lower sequencing depths were simulated from a multinomial distribution taking the counts of that sample as the reference probabilities and aggregating the simulated samples to increase the depth. For each case, 10 simulations were performed and the number of detected genes in each simulated sample were averaged. To simulate the higher sequencing depths, the same procedure was applied but, in order to give genes with no counts a chance to appear, we added 0.2 to the original data.

In **HS** data we observed that around 50% of the annotated genes are found at the nominal sequencing depth of between 20 and 25 million reads (Figure 4.4). The "Saturation" plot estimates that in this range of total reads, around 250 additional genes are detected per additional million reads. This implies that increasing the sequencing depth by 10 million reads will increase transcriptome coverage by 10%. While this information alone is informative,

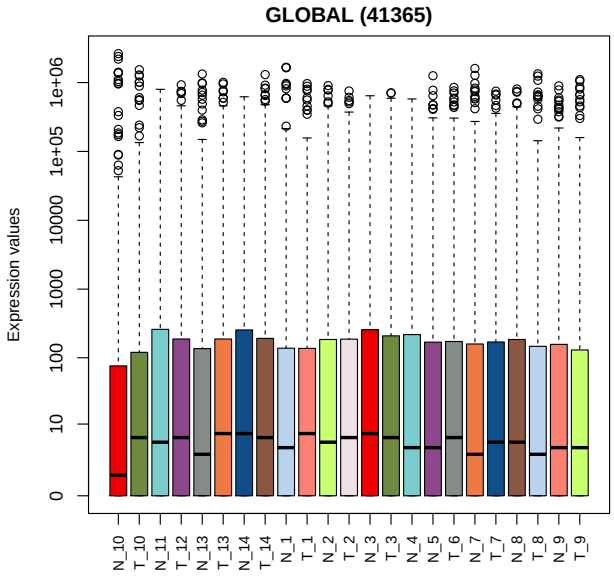


(a) Saturation plot for "protein coding" biotype.

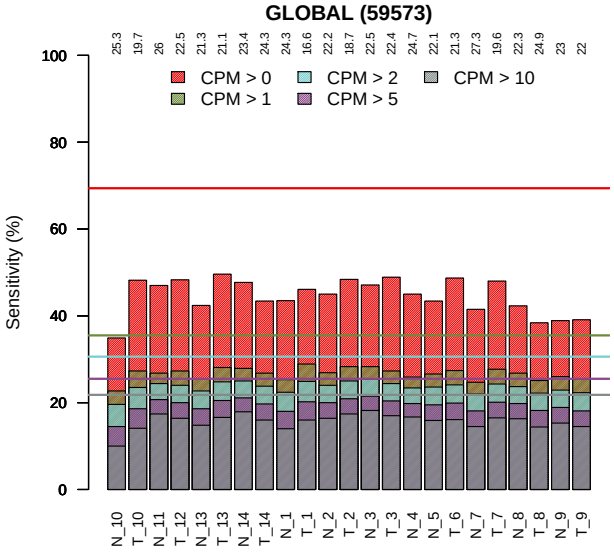


(b) Saturation plot for "lincRNA" biotype.

Figure 4.5: Saturation. **HS** data. Two of the tumoral samples are displayed.



(a) Dynamic range of expression.



(b) Sensitivity plot.

Figure 4.6: Expression quantification. HS data.

it may be more relevant to analyze it in the context of biotype break-down. Figures 4.5a and 4.5b show the "Saturation" plots for the protein-coding and the lncRNA[1] biotypes. Between 16,200 and 17,400 protein coding genes were found in these tumor samples in RNA-seq data sets at a new detection rate of 40. This implies that the feature detection improvement with the addition of 10 million additional reads is estimated to be around 2% for the gene-coding biotype. In contrast, between 1200 and 1400 lncRNAs are found in these samples and their new detection rate stays at 35. This translates into an estimated 25% more lncRNAs with a 10 million read sequencing depth increase. Depending on the goal and scientific questions of the study (i.e. whether lncRNAs are of interest) decisions on the need for additional reads may change.

In the "Dynamic range of expression" plot (Figure 4.6a) , the distribution of the number of read counts per million (CPM) for all the samples in the experiment is compared. Genes with no counts in any of the samples are not used for this plot. Similar to the biotype boxplots, this plot is useful to identify differences in count distributions within the data set. In **HS** data, we observed that the distribution of expression levels for detected genes varies considerably among samples, and suggests that a normalization approach that corrects for these differences would be needed to make the samples comparable. This plot could also be used to reject samples with odd expression level distributions. For example, one could consider removing sample N_10 from the analysis for having a too low median expression level.

The analysis of expression quantification is complemented by the "Sensitivity" plot (Figure 4.6b), that displays the number of genes with more than 0, 1, 2, 5 or 10 CPM for each sample (bars) and in any of the samples (horizontal lines). The sequencing depth (in million sequencing reads) is also provided in the upper side of the plot. This plot reflects the proportion of genes with low expression from the total number of transcripts. In the **HS** example, less than 35% of the genes have more than 1 CPM in any of the

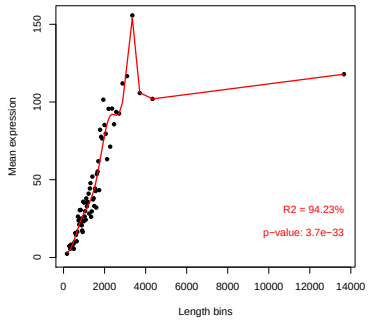
[1]Long non-coding RNA

samples. This means that there are a high number of genes with 1 CPM or less in all of them. The estimation of gene expression is less reliable for low-count genes which represent a source of noise that negatively affects sensitivity and specificity in most statistical analyses [134], and therefore their removal is recommended [5]. Hence, the "Sensitivity" plot provides a graphical representation that helps to make decisions on the CPM threshold because it shows the percentage of features that would be removed at different CPM values.

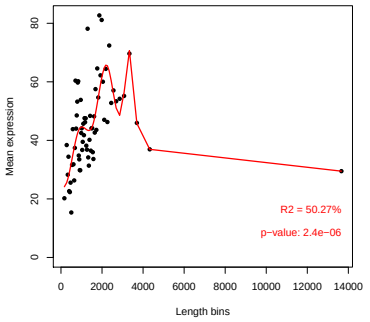
4.3.3 Sequencing biases

Finally, when sequencing artifacts are present in the data, the expression quantification could be biased and lead to misleading conclusions in posterior analysis when comparing samples or genes (e.g. differential expression studies). Therefore a proper and timely detection of these biases is needed to choose an appropriate normalization procedure that corrects data errors and improves downstream statistical analyses: The diagnostic plots in Figure 4.7 are especially designed for this purpose.

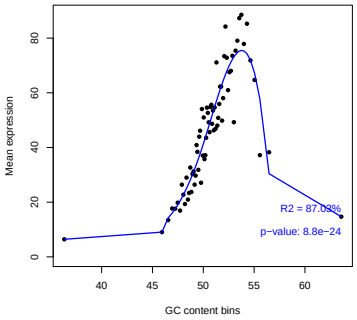
For the "length" and "GC content" plots (Figures 4.7a and 4.7c), bins containing 200 genes were created according to length or GC content. The dots in these plots show the 5% trimmed expression value mean for the 200 genes within each bin. For instance, in the case of length, the first dot on the left corresponds to the 200 shortest genes, the following dot to the 200 next longest genes, and so on. To assess the relationship between the length or GC content and the trimmed gene expression mean, a cubic spline regression model was fitted (red and blue line for length and GC content plots, respectively). When the model p-value is lower than a given significance level (e.g. $\alpha = 0.05$) and R^2 is high enough (e.g. higher than 70%), the fitted curve shows a trend for the relationship between length or GC content and expression and helps to decide whether the bias is strong enough to be corrected. Both feature length and GC content bias are manifest in the **FO** data (Figures 4.7a and 4.7c). The longer the gene, the higher the expression, while



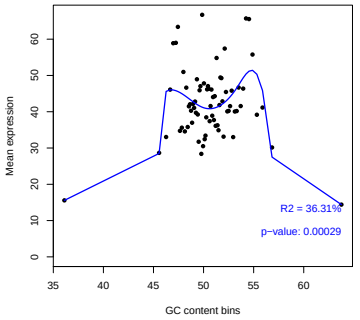
(a) Length bias.



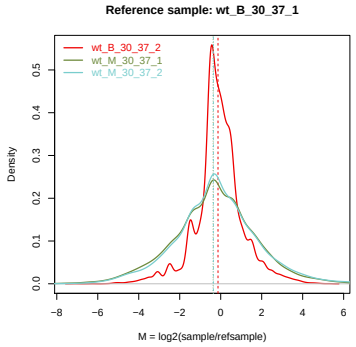
(b) After RPKM normalization.



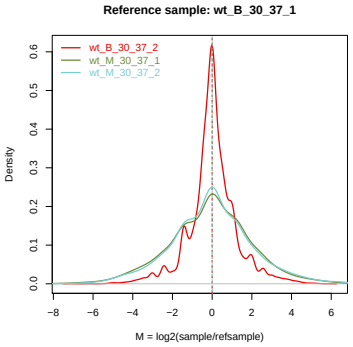
(c) GC content bias.



(d) After EDASeq normalization.



(e) RNA composition bias.



(f) After TMM normalization.

Figure 4.7: Report III. Sequencing biases. FO data (Blood condition).

the curve that relates GC content and expression shows that the expression level diminishes when GC content is very low or very high.

The "RNA composition" plot (Figure 4.7e) indicates if significant differences in the composition of the RNA sample are present. In these plots, M values are computed between each sample s and a reference sample r (which can be arbitrarily chosen) as $M = log_2(x_s/x_r)$, where x_s are the counts in sample s. If no bias is present, the median of M values for each comparison is expected to be 0 [120]. Deviations from this value indicate that expression levels of a fraction of genes in one sample tend to be higher than in the others and mean that the data violate the assumption of a uniform global RNA distribution which is frequently made in genome-wide gene expression experiments. Figure 4.7e shows a deviation from 0 in the M medians for the **FO** data. Confidence intervals for the M median were also computed by resampling. The confidence level was adjusted for multiple testing using Bonferroni's correction. These confidence intervals showed that this deviation was statistically significant.

Confidence Intervals for the median of M values in **FO** *data*

```
Warning: 4197 features with 0 counts in all samples are to be
removed for this analysis.
Reference sample is: wt_B_30_37_1
Confidence intervals for median of M:
                 0.83%                99.17%     Diagnostic Test
wt_B_30_37_2  -0.153815574418935  -0.108393535088251  FAILED
wt_M_30_37_1  -0.370808505381945  -0.370808505381945  FAILED
wt_M_30_37_2  -0.36066807247099   -0.29410617912117   FAILED
Diagnostic test: FAILED.
Normalization is required to correct this bias.
```

Therefore, these diagnostic plots suggest that specific normalization procedures are required to remove these observed biases. In the next section, the most common normalization procedures are briefly described.

4.4 Normalization

Normalization is an important step in RNA-seq analysis since it helps to reduce potential sequencing biases and to make the samples and features comparable. Two different types of normalization may be considered [38, 119]:

- Within-sample normalization procedures try to adjust data for gene specific effects such as gene length or GC content. Longer genes tend to obtain a higher number of read counts and hence a higher estimation of their expression level. A higher or lower GC content in a region of interest may lead to an under-estimation of gene expression [12].

- Between-sample normalization methods aim to correct for systematic differences among samples due to either different sequencing depths or different RNA compositions. The sequencing depth affects expression levels because the more reads that are sequenced for a given sample, the higher the estimation of gene expression will be for that sample. Regarding the RNA composition, this unwanted effect is present when there are highly expressed genes in one of the samples but not in the other samples. These genes have a relatively higher number of reads in that sample while not in the others. If this is the case, genes with a lower number of read counts would not be comparable among samples and could contribute to artificially increasing the biological variability among samples, possibly leading to unreal expression changes between conditions.

Many normalization strategies have been developed to deal with these potential biases that can be detected using the plots described in the previous section. Some of the most widely used normalization methods are briefly described in Table 4.1, where x_i^s is the number of read counts for gene i in sample s, y_i^s is the normalized expression value for gene i in sample s, N^s is the total number of counts in sample s (sequencing depth), l_i is the length of gene i, and Q_{75}^s is the upper quartile of non-zero gene counts in sample s.

Table 4.1: Normalization methods for RNA-seq data

Type	*Method*	*Description*
Within-sample	EDASeq [119]	Four approaches to correct for length or GC content: Regression (loess), Full-quantile, Median or Upper Quartile normalization.
	RNASeqBias [154]	Generalized Additive Model to correct for gene length or GC content.
Between-sample	Upper Quartile [18], also in EDASeq [119]	$y_i^s = x_i^s / Q_{75}^s$
	TMM [120], also in edgeR [123]	Trimmed Mean of M values. Given a reference sample r, $f_0^s = 2^{WM(log_2((x^s/N^s)/(x^r/N^r)))}$, where WM is the weighted mean after removing the highest log-ratios and the most expressed genes. The scaling factor is $f^s = f_0^s / e^{\overline{ln(f_0)}}$. Then: $y_i^s = x_i^s / (f^s \times N^s / \bar{N})$
	Quantile [16], also in EDASeq [119]	The distributions of gene counts are matched across samples.
	DESeq [4]	If $f^s = Me_i(x_i^s / GM(x_i))$, where Me is the median and GM the geometric mean: $y_i^s = x_i^s / f^s$.
Both	RPKM [99]	Reads Per Kilobase per Million mapped reads adjusts for length bias and sequencing depth: $y_i^s = 10^9 \times x_i^s / (N^s \times l_i)$
	CQN [58]	Conditional Quantile Normalization to correct for any within-lane systematic bias and also for between-lane normalization.

In the NOISeq R package, three of these popular normalization methods were included: RPKM [99], Upper Quartile [18] and TMM [120]. However, to perform a differential expression analysis (see Chapter 5), the package also accepts previously normalized data, so any other normalization procedure can be applied instead.

The diagnostic plots described in the previous section (Figure 4.7) revealed the presence of sequencing biases in the data that required a within-sample (length or GC content) and between-sample (RNA composition) normalization. The choice of the type of normalization depends on the ulterior analysis to be performed. For instance, in differential expression studies, it is essential to assure that changes in expression between two samples are indeed due to biological differences and not to sequencing artifacts. Therefore, at least a between-sample normalization needs to be applied.

Different types of normalization procedures conceived to target each specific bias were applied: RPKM [99] (included in the NOISeq package) to correct for length bias, "full" within-sample normalization in the EDASeq package to correct for GC content bias [119] and TMM [120] (also included in the NOISeq package) to correct RNA composition bias. The expression dependence on gene length was reduced after applying RPKM (Figure 4.7b). The same happened for GC content bias after using the EDASeq package (Figure 4.7d). After TMM normalization, the distributions of M values did not shift and the median was approximately 0, which indicated that the RNA composition bias had been mitigated (Figure 4.7f).

4.5 Filtering out low-count features

It has been often argued that, in RNA-seq, expression estimation for low count genes is less reliable because read counts could have been assigned by chance [97, 137]. Thus, excluding features with low counts may improve the results of statistical analyses because the level of noise is reduced. However, the best procedure to filter these low count features has not yet been decided.

To the best of our knowledge, no filtering procedures have been implemented so far in statistical packages for RNA-seq data, but it is a common practice to simply remove genes with total counts for all the samples lower than a certain cutoff, e.g. 10 counts [4, 18, 134]. This approach does not take into account the sequencing depth of the experiment to decide the cutoff, so genes with a relatively high expression in one of the conditions could be ignored. A better method is the procedure described in the edgeR R package User's Guide in the Bioconductor repository. The authors proposal consists of keeping genes with counts per million reads (CPM) above a given threshold in at least as many samples as the number of samples per condition. By setting the cutoff for the CPM instead of the raw counts, it can be assured that no genes with a high relative expression are eliminated. In the NOISeq package, we implemented three different filtering procedures: the CPM method, Wilcoxon test, and Proportion test, which are described in detail in the next section.

4.5.1 CPM method

Let x_g^s be the number of raw counts of gene g in sample s. As in the edgeR proposal, counts for each sample are transformed to counts per million reads (CPM): $CPM_g^s = 10^6 \times \frac{x_g^s}{\sum_g x_g^s}$. A value for CPM under which a feature is considered to have low counts must be previously set (cpm). By default, the CPM method takes a cutoff of $cpm = 1$. If there are S samples in a given experimental condition, the cutoff for that condition would be $cpm \times S$. A gene g is filtered out if the sum of CPM values across all the samples in the same condition is below the condition cutoff ($\sum_s CPM_g^s < cpm \times S$) for all the experimental conditions.

It is also possible to remove genes that present inconsistent expression values in any of the experimental conditions with the CPM method. A cutoff for the coefficient of variation per condition cv has to be set a priori. Then, a gene g will be filtered out either if it has a total CPM value per condition

of less than $cpm \times S$ or a coefficient of variation per condition higher than the cv cutoff for all the conditions.

4.5.2 Wilcoxon test

Although the CPM method takes the experimental design and the variability per condition into account , it has the drawback of having to decide the cutoffs to use for both the CPM and the coefficient of variation. Hence, we propose the Wilcoxon test to identify those genes with a CPM value median per condition that is significantly higher than 0. Thus, the hypothesis to test for each gene and condition is $H_0 : m = 0$ versus $H_1 : m > 0$, where m is the median of the CPM values per condition. To be more conservative, no multiple testing correction was applied in order to retain as many genes as possible. Genes with a p-value higher than 0.05 in all the conditions are filtered out.

By using the Wilcoxon method, genes with inconsistent values across replicates within the same condition or with a low median expression value tend to be removed. However, this non-parametric procedure is only recommended when the number of replicates per condition is at least five.

4.5.3 Proportion test

The proportion test aims to be the alternative to the Wilcoxon test when few replicates per condition are available. This method requires a cutoff to be set for CPM (cpm), but not for the coefficient of variation. It is based on the idea that read counts for a given gene follow a binomial distribution where the number of trials n is the sequencing depth and the probability p is the probability of expression for that gene under a given experimental condition, which is unknown. Thus, in this case, $H_0 : p = p_0$ is tested versus $H_1 : p > p_0$. Since it is not possible to use $p_0 = 0$ in a binomial proportion test, we define $p_0 = cpm/10^6$. If several replicates are available for an experimental condition, we sum across replicates ($x_g = \sum_s x_g^s$) and use

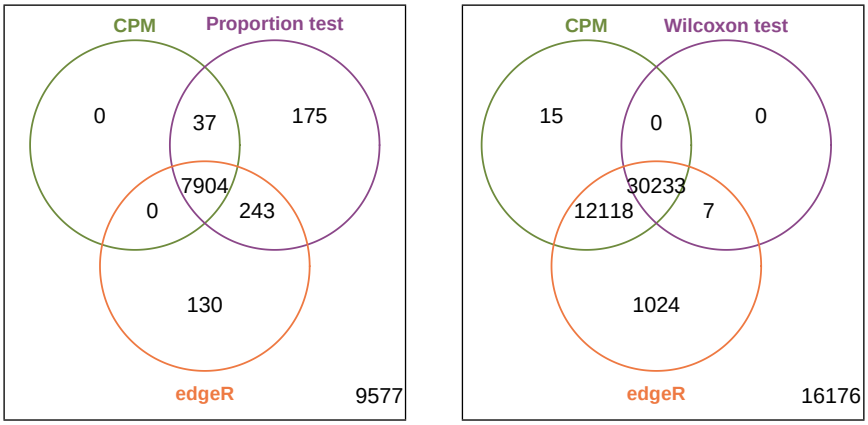
this single value as the observed binomial variable. Then, $n = \sum_g x_g$. Again, to be conservative, the raw p-values are used and genes with a p-value higher than 0.05 in all conditions are filtered out.

4.5.4 Comparing filtering methods

We applied the three NOISeq filtering procedures and edgeR proposal to **FO** (with 2 replicates per condition) and **HS** data (with 11 and 12 replicates per condition) to illustrate the similarities and differences of the methods. We set a cutoff of $cpm = 1$ for the CPM method, Proportion test, and edgeR approach. Because of the number of replicates, the Proportion test was only applied to **FO** and the Wilcoxon test was applied to **HS**. We considered a coefficient of variation of 500 for the CPM method to cancel this filter and make this method more comparable to edgeR approach. According to the number of replicates per condition in each dataset, genes with a CPM higher than 1 in at least 2 or 10 samples for each dataset respectively were retained in the edgeR approach.

Both datasets originally contained 18066 (**FO**) and 59573 (**HS**) genes. Out of these, 9577 and 16176 respectively, were not filtered out by any of the methods (Figure 4.8). Most of the filtered genes (7904 and 30233) were removed by all the methods which indicates that, in general, there were very few differences among them. The greatest difference was found for the Wilcoxon test (**HS**), since there were more than 12000 genes that were removed by CPM and edgeR but not by Wilcoxon.

We studied the characteristics of the removed genes that were not in common for the compared filtering methods by plotting the difference between the mean CPM per condition against the maximum variability between replicates (Figures 4.9 and 4.10). Genes filtered only by edgeR tended to show higher differences in expression between conditions which is obviously not good because genes with potentially significant changes in expression between conditions could be removed from the analysis. Although these genes generally present a high variability among replicates and will probably not be



(a) FO data **(b) HS** data

Figure 4.8: Number of genes filtered out by each method

declared as differentially expressed by statistical methods, it may be preferable to leave the decision about these cases to the statistical method instead of filtering them out of the ulterior analysis.

Finally, to illustrate how low-count filtering affects statistical analyses, for instance, differential expression analysis (see next chapter), we applied the CPM filtering method with a cpm threshold of 1 to **HS** data and obtained differentially expressed genes with two statistical packages (described in the next chapter): NOISeqBIO and edgeR [123]. Figure 4.11 shows the results of the filtering approach. A total of 42366 low-count genes were removed after applying the CPM threshold, of which 292 and 887 had been detected as differentially expressed by NOISeqBIO and edgeR, respectively. In turn, removing these low-count genes resulted in 683 (NOISeqBIO) and 1195 (edgeR) newly detected genes that belonged to a higher expression range. These results highlight the impact of low-count filtering in RNA-seq differential expression analysis and how the resources of the NOISeq package can be used to address this task easily.

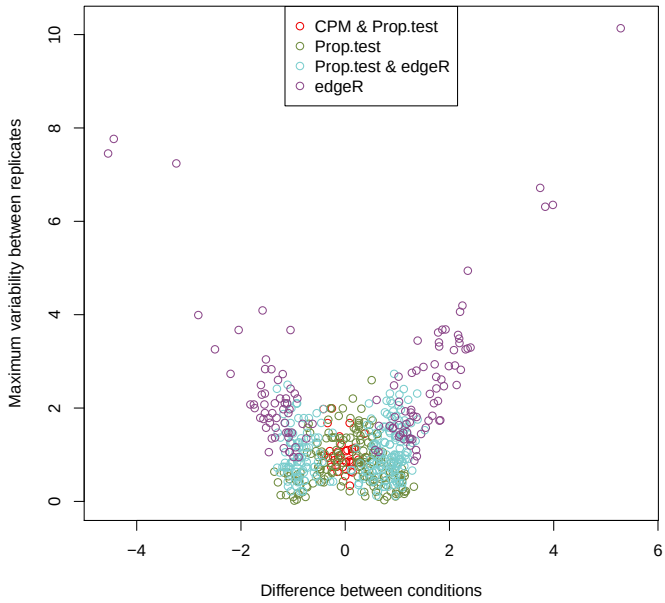
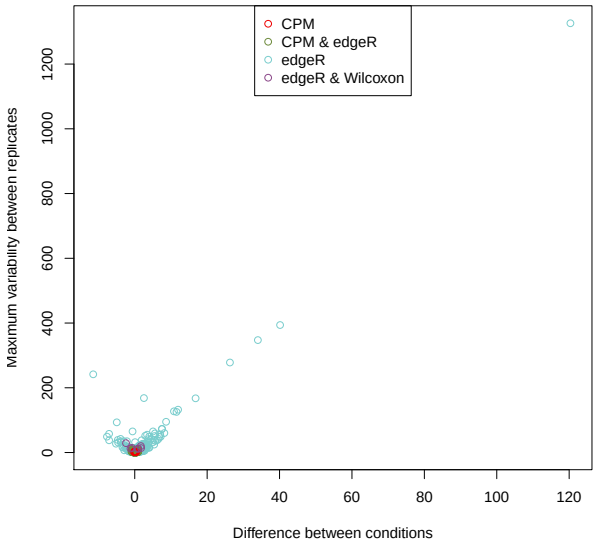


Figure 4.9: Difference between expression mean per condition versus maximum difference between replicates for genes not removed by all methods from **FO** data.

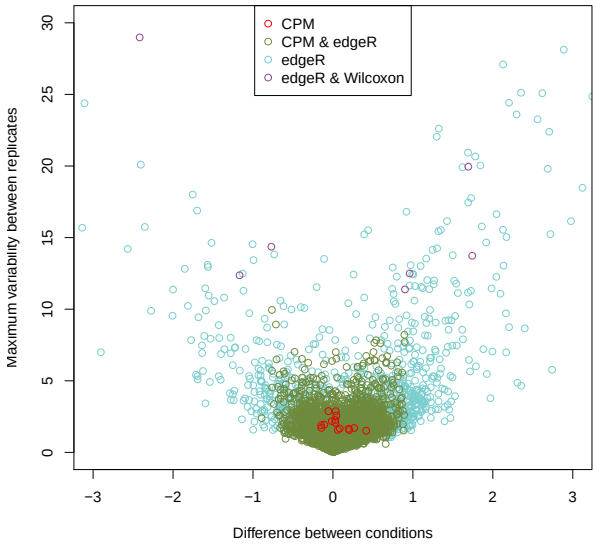
4.6 Discussion

RNA-seq technology reads the population of RNA molecules in a given sample and renders a direct quantification of the abundance of each transcript, mapping ambiguities and sequencing errors issues separately. Although this is fundamentally true, as shown in studies on how RNA-seq data corresponds with microarray and RT-PCR data [18, 53, 92], there is still some work to be done to fully understand the characteristics of RNA-seq data and to properly process them in order to obtain accurate results in further statistical analysis.

Estimation of gene expression levels from sequencing reads seems conceptually simple and was initially seen by many researchers as a very straightforward task. However, it implies the execution of a series of complex computational algorithms that should be chosen and adapted to the characteristics



(a) All data



(b) Zoomed data

Figure 4.10: Difference between expression mean per condition versus maximum difference between replicates for genes not removed by all methods from **HS** data.

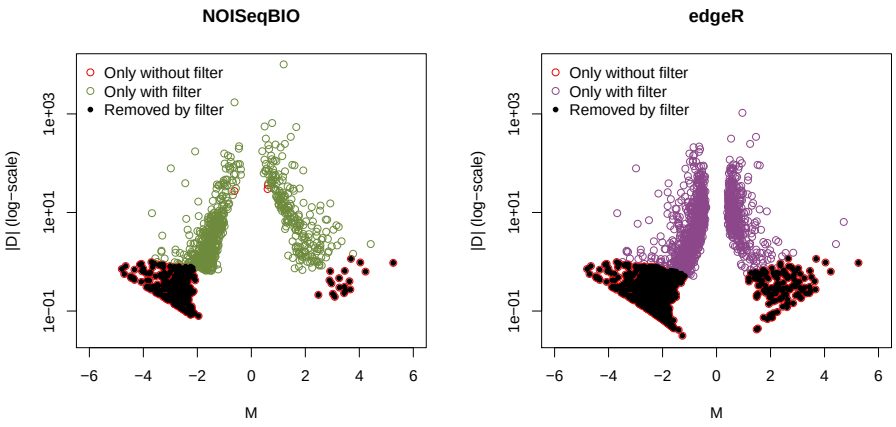


Figure 4.11: Differentially expressed genes not in common when filtering or not filtering low-count genes using CPM method in **HS** data.

of the sequencing platform, sample preparation protocol, organism, etc.

Thus, from the files containing the sequencing reads to the estimation of expression given by the number of reads assigned to each biological feature being studied, a variety of tools to assess the quality of the results at each step of the analysis have been introduced. FastQC [8] is a popular software package for analyzing the data quality and some other aspects such as duplication of reads, GC or N content, etc. of the raw sequencing reads. There are indeed other software packages such as the NGS QC Toolkit [111] that not only generate a quality control report but also include other functionalities such as trimming the sequencing reads to remove low quality bases or adapter sequences before continuing with the mapping step. Once the reads are mapped to the reference genome, tools such as RSeQC [148], RNA-SeQC [37], Qualimap [49], or RNASeqGUI (an R package with a graphical interface) [126] can evaluate the results of the mapping. Some of them also offer other possibilities such as expression quantification or differential expression analysis. If mapping results are satisfactory, the next step would be to obtain the number of reads assigned to each biological feature, i.e. the expression quantification. These read counts will then be used as the input for statistical

models.

However, no matter what quality control procedures were performed during this process, it is essential to determine if there are systematic technical effects that are biasing the expression estimation before going on with further analyses. One of the biases that rapidly becomes evident is the effect of the depth of the sequencing experiment; not only because counts from samples with different sequencing depths are not comparable. Ultra-high throughput sequencing seems advisable to detect transcripts with low expression values because of the large dynamic range of gene expression obtained. However, as more sequencing output is considered, the diversity and quantity of detected off-target RNA species, such as several types of small RNAs, also increases [141]. The extent to which each of these biotypes and transcripts are purification artifacts or have a biological significance warrants a separate study but it does show an important property of RNA-seq data: the effect that sequencing depth has on the distribution of reads among transcripts and the quantification of expression, essentially a percentage in the case of this technology. Robinson *et al.* [120] have already highlighted the implications that different transcript distributions might have on RNA-seq normalization and differential expression. A preliminary exploration of the count data may also be helpful to know the biological and quantitative characteristics of the data, or even to detect any potential contamination. In addition, when there are highly expressed transcripts in some of the samples that accumulate a huge number of reads, an unwanted decrease in the read counts of transcripts with low or medium expression is produced that might distort their comparison across samples.

Other important elements that may affect the expression quantification are the transcript length and the GC content. The nature of the short read procedure makes it inevitable that longer transcripts will be preferentially detected over shorter ones, and this has been shown to have implications on the biological interpretation of the data [106, 152]. Also, an effect of the

GC content of the transcripts has been reported for Illumina sequencing reads [12, 119] which may also bias the data.

In this chapter, we described a whole set of useful graphical and diagnostic tools to assess the quality of the data prior to statistical analysis. We illustrated which kind of information these tools provide by using experimental data. With these quality control plots we obtained a useful description of biological and quantitative traits of the data, and also assessed the effect of sequencing biases in expression quantification in order to choose an appropriate normalization procedure. After reviewing the most popular normalization methods and, according to the results from the exploratory analysis, we selected the normalization procedure which should be applied in each case. In addition, as it has been often claimed, features with low counts may hamper the study of the behavior of the rest of features because they cause the statistical methods used to lose their power. Here we proposed some methods to filter out these noisy low-count genes, compared the differences among these methods, and showed the impact of filtering on the differential expression analysis results.

All these functionalities are included in an open Bioconductor R package called NOISeq, which also offers the possibility of generating a Quality Control Report PDF to facilitate exploration of the user's data. In this package, we also implemented a non-parametric statistical procedure to compare two experimental conditions and to find out which genes present a significant change in expression between conditions (differentially expressed genes) that will be introduced in the next chapter.

Chapter 5

Differential expression in RNA-seq

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A.
Differential expression in RNA-seq: A matter of depth
Genome Research, 21:2213-2223, **2011**

Tarazona S, Furió P, Turrà D, Di Pietro A, Ferrer A, and Conesa A.
Quality-control, experimental design and FDR controlled differential expression of RNA-seq with the NOISeq R package
(In preparation)

5.1 Introduction

When high-throughput technologies are used to quantify the expression of thousands of genes in several experimental conditions, one question rapidly arises: which genes are differentially expressed? In bioinformatics, this means identifying the genes that present a statistically significant change in expression due to an external stimulation, experimental perturbation, or simply changes over time. These differentially expressed genes (DEGs) are likely involved in the biological mechanisms that give rise to different phenotypes between conditions (e.g. diseased versus healthy individuals, treated versus control, etc.).

Methodologies to analyze differential expression (DE) have been thoroughly studied for microarray data [33, 108]: these take a measurement of luminescence to estimate the expression level of the genes. Thus, the estimated expression level is a continuous variable and most of the proposed methods assume a normal distribution for this type of data [133]. RNA-seq expression measurement is totally different to microarrays, since in RNA-seq the number of sequencing reads falling into a gene is considered to be the estimation of the gene expression level. Therefore in this case the expression level is a discrete variable, which has motivated the development of novel differential expression methods specific for the NGS technologies. Methods traditionally used for microarrays have paved the way to other approaches that take into account the discrete nature of the expression quantification by using discrete probability distributions to model the data. The vast majority of the methodologies proposed so far rely on parametric assumptions [4, 59, 92, 105, 123, 136, 140]. For technical replication, they generally use a Poisson distribution to model feature counts, following the rationale of the sampling procedure in RNA sequencing. While in a Poisson distribution the mean and variance have to be equal, the Negative Binomial distribution allows for over (or under) dispersion, as occurs in the case of biological replication, and so the read counts are modeled with the Negative Binomial distribution.

However, the subsequent confirmation of distributional assumptions is important as they might not always hold true [18]. Moreover, there are usually very few replicates, if any, available, which hinders the estimation of model parameters. Non-parametric methods such as SAMseq [82] or NPEBseq [13] have also been proposed and they have the advantage of not requiring such strong assumptions to be fulfilled. However, the low replication problem remains a drawback, because these kinds of methods tend to perform better with a higher number of replicates. Another common alternative would be the adaption of popular microarray methodologies such as Limma [133] for application on transformed counts (e.g. $\log_2$ transformation or *voom* transformation), as described in the literature [79, 116].

In this chapter, we introduce two non-parametric approaches for DE analysis: NOISeq and NOISeqBIO, which have been implemented in the open Bioconductor R NOISeq package (`http://www.bioconductor.org/packages/release/bioc/html/NOISeq.html`). The NOISeq DE method [141] was developed in the infancy of RNA-seq when experiments with only technical replicates or no replicates at all were being produced by researchers (because of the still high cost of the technology) and hence were available from public repositories. Therefore, NOISeq is optimized for the use of technical replicates and can even process experiments without replications. It has been successfully used in several studies [21, 26, 40, 44, 85, 127, 155] and benchmarked against the most popular differential expression methods with good results [13, 73, 101, 134].

Biological replicates are now more common in RNA-seq experiments, although the number of them is still limited. Researchers tend to invest their budget in increasing the sequencing depth of the experiments in order to detect a wider variety of transcripts. However, they should be aware of the fact that having more replicates per condition improves the biological variability estimation which results in a more robust DE analysis that gives more power to the DEG capture. Therefore, the strategy of increasing replication

instead of depth should be recommended when DE analysis is to be performed [86, 130].

To better handle biological variability, we adapted the NOISeq method and named it NOISeqBIO. NOISeqBIO implements an empirical Bayes approach that improves the way of dealing with the biological variability specific to each gene and is very successful in controlling the high false discovery rate in experiments with biological replicates, where other methods have previously been shown to fail [134].

We tested our NOISeq and NOISeqBIO methods on both experimental and simulated datasets, and compared them to other widely used differential expression methods such as Fisher's Exact Test, edgeR [123], baySeq [59], DESeq [4], and SAMseq [82].

5.2 Data

5.2.1 Experimental data

5.2.1.1 Experimental data with technical replicates

Two publicly available human RNA-seq datasets with different sequencing depths were used in the studies with technical replication. In both of them, sequencing was done with Illumina technology.

The **MAQC** dataset [18, 128] was generated for benchmarking purposes on RNA-seq. It consists of two samples: Ambion's human brain reference RNA (Brain) and Stratagene's human universal reference RNA (UHR). Each sample comprises seven lanes, providing a total of 42 and 45 million reads respectively. Each lane was considered as a technical replicate. This project additionally has RT-PCR data for validation of RNA-seq analysis results.

Griffith's dataset [53] contains 96 and 198 million paired-end reads respectively of the transcriptome of two human colorectal cancer cell lines which only differ in the fluorouracil (5-FU) resistance phenotype. The technical replicates corresponded to 13 lanes selected from each condition to equilibrate the

sequencing depth in both conditions to around 100 million reads. RT-PCR data were also available in this data set for a number of genes.

Raw *fastq* files containing the sequencing reads were downloaded from the SRA archive [80] and mapped against the *Homo sapiens* high coverage assembly *Hg19* from Ensembl [46] using Tophat [143], allowing up to 2 mismatches and discarding reads mapping at multiple locations. Counts for each gene were computed by means of the HTSeq Python package [6] using Ensembl genes' annotation (version 60) and only exonic reads.

Regarding the RT-PCR measurements from these two experiments, we identified positive (RT-PCR differentially expressed) and negative (RT-PCR non-differentially expressed) genes following a previously reported procedure [18, 53] and matched them to Ensembl IDs. After discarding replicates and eliminating unmatched genes, a total of 330 and 82 *positive* genes and 83 and 12 *negative* genes for MAQC and Griffith's dataset, respectively, were taken to compute the indicators for the performance plots.

5.2.1.2 Experimental data with biological replicates

We used two experiments with a very different number of biological replicates that were described in detail in the Data section in Chapter 4. Briefly, the **FO** experiment was done with the fungal pathogen *F. oxysporum*, and compares fungal growth in a human blood culture against growth on a minimal medium with the purpose of studying the mechanisms of proliferation of the fungus in humans. It has two replicates per condition. The prostate cancer experiment (**HS**) compares tumoral and healthy tissues from Chinese patients with 11 and 12 replicates per condition respectively.

For the preliminary studies on simulated data with biological replication we also used samples from an experiment carried out within our TRANSPAT project. In this case another opportunistic human pathogen, the fungus *Aspergillus fumigatus*, was cultured in human blood and in minimal medium with the same goal as in the **FO** experiment: comparing gene expression between

these two experimental conditions to elucidate which genes were responsible for the infection process in humans.

5.2.2 Simulated data

5.2.2.1 Simulated data with technical replication

The synthetic data included in the baySeq R package [59], which contains counts for 1000 features evaluated in two experimental conditions with 5 replicates each, were used. The first hundred features are differentially expressed.

5.2.2.2 Simulation algorithm for biological replication

We developed an algorithm to simulate data with biological replicates in order to better assess the performance of the DE methods being compared.

It has been reported [4, 59, 123] that the number of reads mapping to a given gene resembles an over-dispersed Poisson distribution when considering biological replicates and that one way of modeling this over-dispersion is by taking the negative binomial distribution. Thus, our simulation algorithm is based on randomly generating the counts from a negative binomial distribution as done previously in other studies [124, 134]. Figure 5.1 shows the outline of the algorithm.

These are the main steps of the simulation algorithm:

1. An initial number of counts per gene (μ_0^g) is used to simulate the replicates for each condition. This μ_0^g determines the proportion of sequencing reads initially assigned to each gene g. It can be either provided by the user or randomly generated from a power-law distribution: $f(x) \propto x^{-\lambda}$, where x is the gene expression level, $0 \leq x \leq depth/1000$ and $\lambda = 0.5$. Thus, if $ngenes$ is the number of genes in the simulated dataset, $ngenes$ values are randomly generated from this distribution to be used as the initial counts μ_0 when no experimental samples are provided by the user.

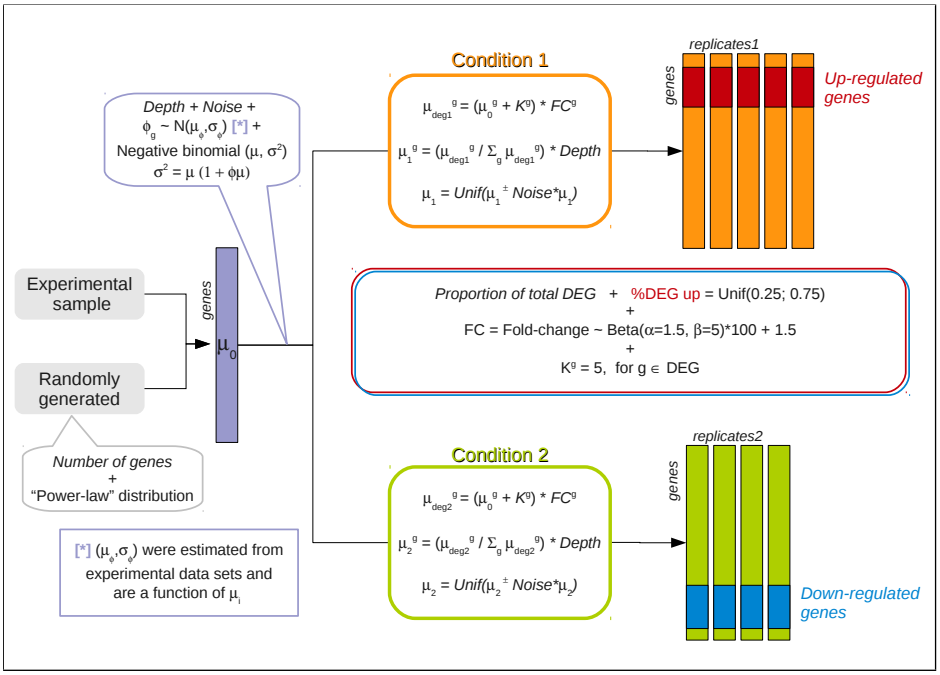


Figure 5.1: Outline of the simulation algorithm used for biological replications

2. The proportion of differentially expressed genes ($propdeg$) is chosen by the user and is used to obtain the number of DEGs. The proportion of DEGs that will be upregulated in condition 1 is generated from the uniform distribution $\mathcal{U}(0.25, 0.75)$, and the rest of the DEGs are down-regulated in this condition. Genes that are up and down regulated are randomly taken from the total set of genes.

3. The number of replicates per condition is given by the parameter $nrepl$. Each biological replicate for a given gene and condition is simulated from a negative binomial distribution with mean μ and variance σ^2. To describe the relationship between the mean and the variance, previously used parametrization [123] was applied: $\sigma^2 = \mu(1 + \phi\mu)$. This is how μ and σ^2 are estimated from the initial counts μ_0:

   - **Estimation of μ**

     For each condition i $(i = 1, 2)$, the mean expression is defined

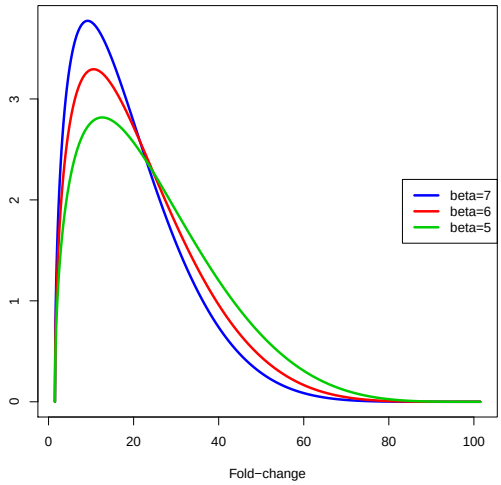


Figure 5.2: The fold-change is generated randomly from a Beta distribution with shape parameters $\alpha = 1.5$ and β (which can be modified, by default $\beta = 6$). The larger the β value, the lower the probability of having high fold-changes.

as $\mu_i^g = (\mu_0^g + K^g) \times FC^g$, for $g \in DEG$, and $\mu_i^g = \mu_0^g$, for $g \notin DEG$. The fold-change FC^g is randomly generated from a Beta distribution: $\frac{FC^g - 1.5}{100} \sim Beta(\alpha, \beta)$, where $\alpha = 1.5$. By default, $\beta = 6$, but it can be modified by the user (see Figure 5.2). The constant K_g is included because, if $\mu_0^g = 0$, no change would be applied to that specific gene. We set $K^g = 5, \forall g \in DEG$. The mean μ_i^g thereby obtained for each gene and condition is adjusted so their sum is equal to the given total number of counts ($depth$). Finally, in order to allow a certain level of noise in the data ($noise$), the final μ_i^g is the maximum between 0.1 and a random value from the uniform distribution $\mathcal{U}(\mu_i - noise \times \mu_i, \mu_i + noise \times \mu_i)$. The reason for taking the maximum is to give any gene with no initial counts some chance to appear.

- **Estimation of ϕ**

  To compute the variance σ^2, we first need to estimate the value of the dispersion parameter ϕ. We evaluated several experimental

data sets with different number of replicates and biological variability to obtain realistic scenarios of either high or low biological variability. We followed the estimation procedure described in [124]. For each dataset, only samples with a total number of counts higher than 10^6 and the genes with a mean expression higher than 1 were chosen. Once this filter was applied, the remaining samples were adjusted so all of them had the same number of counts (depth). With these normalized data, the mean expression of each gene was computed, which is the maximum likelihood estimator (MLE) of μ^g. The MLE of ϕ^g was obtained by maximizing the log-likelihood function. This was done for each experimental dataset and all the pairs (μ^g, ϕ^g) from every dataset were pooled. All μ values were divided into bins containing approximately 1000 values each. Figure 5.3 shows the dependence of ϕ^g on μ^g for the scenario of high biological variability. The higher the value of μ, the lower the median and variability of ϕ. The mid-point of the bin was computed for each bin of μ values, as well as the median and the median absolute deviation (mad) of ϕ values within the bin. Thus, for each condition i, ϕ^g is randomly taken from a normal distribution $N(\mu^g_\phi, \sigma^g_\phi)$, where $\mu^g_\phi = \mu_\phi(\mu^g_i)$ and $\sigma^g_\phi = \sigma_\phi(\mu^g_i)$ are obtained by linear interpolation from μ mid-points and ϕ medians.

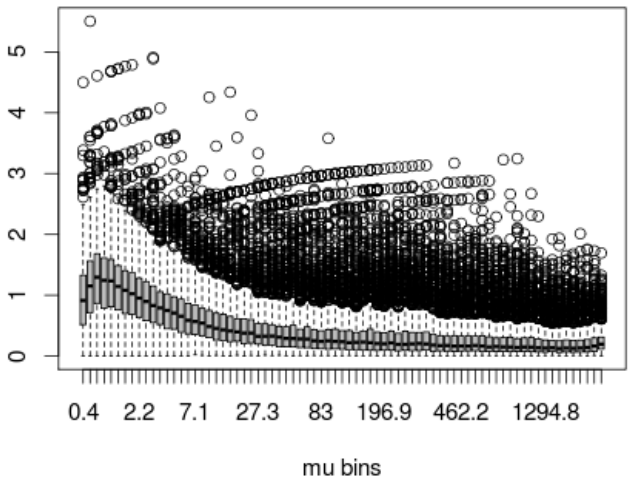


Figure 5.3: Distribution of ϕ values (in log-scale) from experimental datasets within each bin of μ values (containing approximately 1000 values each).

Figure 5.4 shows some examples of simulated datasets for different levels of noise and DEG proportions and with high biological variability.

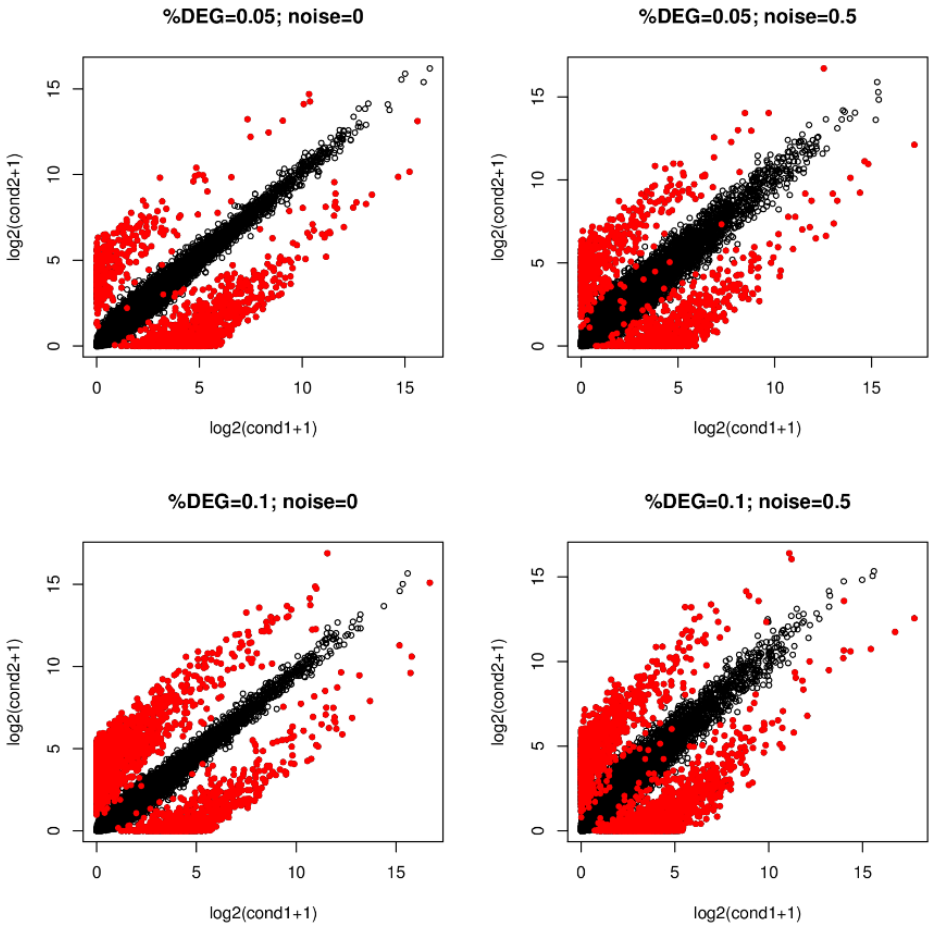


Figure 5.4: Example of simulated data. A sample from the *Fusarium oxysporum* experimental dataset was taken as the initial counts μ_0. The parameters in common for all the simulations were: $nrepl = 5$ in both conditions, $depth = 30$ million and $beta = 6$. The values for parameters $noise$ and $propdeg$ are indicated in each plot. The differentially expressed genes are highlighted in red.

5.3 Methods

5.3.1 NOISeq

NOISeq is a non-parametric approach for the identification of DEGs between two experimental conditions from count data with technical replicates or no replicates at all. The basic idea underlying NOISeq is that a given feature may be considered differentially expressed if its change in expression between the two experimental conditions is greater (or has a higher probability of being greater) than the change observed among replicates within the same condition. Essentially, NOISeq creates a noise distribution of count changes by contrasting fold-change differences (M) and expression differences (D) for all the genes in samples within the same condition. This reference distribution is then used to assess whether the (M, D) values computed between two conditions for a given gene are likely to be part of the noise or represent true differential expression (Figure 5.5). In practice, NOISeq creates the noise distribution by joining (M, D) values from all possible pair-wise comparisons between replicates of either condition.

Let x_{ij}^g be the expression of gene g in condition i $(i = 1, 2)$ and replicate j. To measure the expression level change between two conditions, NOISeq takes two statistics into consideration: the log fold change (M) and the difference (D), that are calculated as in Equations 5.1 and 5.2:

$$M_s^g = log_2(\bar{x}_1^g / \bar{x}_2^g) \tag{5.1}$$

$$D_s^g = \bar{x}_1^g - \bar{x}_2^g \tag{5.2}$$

The reason for using these two statistics is to get more reliable measurements of the change, because the fold change for features with low read counts can be misleading; the same occurs for the difference in expression between two conditions in the range of high counts. To avoid indeterminate values when computing the M statistic, values of 0 are changed to a constant

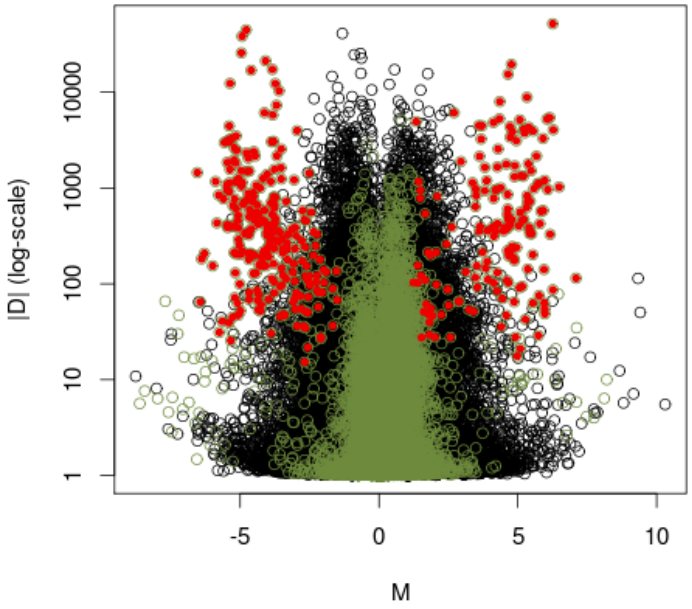


Figure 5.5: $(M, |D|)$ plot for signal distribution (green), noise distribution (black) and genes declared as differentially expressed (red).

value k, where k can be chosen by the user or computed as the midpoint between 0 and the next highest value. The use of normalized data for analyzing differential expression with NOISeq method is highly recommended. NOISeq has the advantage of accepting already normalized expression values instead of gene counts, so it can be used with any normalization procedure. It is also possible to use the raw counts as input and specify which of the normalization methods implemented in the package is to be used.

Note that when technical replicates correspond to different lanes of the flow-cells used to sequence the biological samples, the counts of the replicates are summed instead of averaged.

There are two variants of the method: NOISeq-real and NOISeq-sim. When technical replicates are available, NOISeq-real uses them to compute the noise distribution. NOISeq-sim simulates them in the absence of replication. It should be noted that the NOISeq-sim simulation procedure equates to technical replication and does not reproduce biological variability, which

is necessary for population inferential analysis. However, many RNA-seq experiments still have low replication, thus the ability of statistical methods to work with technical replicates, or no replicates at all, is relevant. It is also very common to generate pilot data before completing a whole experiment, so NOISeq-sim can be applied in these cases to get a preliminary picture of the genes which change between conditions.

5.3.1.1 NOISeq-real

In order to determine the probability of differential expression, the algorithm creates a so called "noise" distribution by pooling the (M_n, D_n) values computed among replicates within the same condition. Thus, these noise values can be calculated as: $M_n = log_2(x^g_{ij}/x^g_{ik})$ and $D_n = x^g_{ij} - x^g_{ik}$, for each pair of replicates j and k ($j \neq k$) and for any condition i or gene g. All these values are pooled together to generate the bivariate noise distribution.

Next, we derive an empirical cumulative distribution function $F(M, D)$ from the absolute value of these "noise" measurements (M_n, D_n). Given a certain gene with (m_s, d_s) values of the "signal" statistics (M_s, D_s), computed from the comparison of both experimental conditions, we suggest that the probability of differential expression be estimated as follows:

$$F(|m_s|, |d_s|) = P(|M_n| \leq |m_s|, |D_n| \leq |d_s|) \tag{5.3}$$

The higher this probability, the higher the change in expression between conditions with regard to "noise", and the more we expect that change between conditions is due to the experimental factor effect and not to chance. We recommended that a threshold around 0.8 is used for this probability. This threshold would be equivalent to an odds of 4:1, which means that the gene is four times more likely to be differentially expressed than non-differentially expressed.

Because the NOISeq method was developed for technical replicates or no replicates at all it was not our intention to compute statistical significance in terms of p-values. However, many NOISeq users have requested equivalence

between the probability of differential expression returned by NOISeq and p-values. The statistic (M, D) used by NOISeq to assess differential expression is a bivariate statistic. We searched in the literature for cases in which a p-value was computed from bivariate statistics, but we found that in multivariate hypothesis testing, the statistic is transformed to a univariate distribution (e.g. Hotelling's T^2 when testing equality of means in multivariate normal populations). This is the approach we followed in NOISeqBIO, as we will see in the next section. However, to preserve the bivariate nature of the NOISeq statistic when testing $H_0 : \mu_1 = \mu_2$, we suggest computing the p-value as in Equation 5.4, taking into account that in NOISeq we consider the absolute values of M and D to obtain the probability of differential expression and that we intend to declare a gene as DEG when both M and D values are higher than in noise.

$$p - value(m_s, d_s) = P(|M_n| > |m_s|, |D_n| > |d_s|) \tag{5.4}$$

Note that the p-value is not equivalent to $1 - F(|m_s|, |d_s|)$. NOISeq performance using this p-value definition has not yet been tested.

5.3.1.2 NOISeq-sim

When there are no technical replicates available in any of the experimental conditions, the NOISeq algorithm can simulate them. The simulation relies on the assumption that read counts follow a multinomial distribution, where the probability for a given class (gene) in the multinomial distribution is the probability of a read to map to that gene. These mapping probabilities are approximated using counts in the only available sample of the corresponding experimental condition. Counts equal to zero are replaced with 0.5, to give all genes some chance to appear.

Given the sequencing depth s_i of the unique available sample in condition i, sequencing depth for the simulated samples is generated randomly from a uniform distribution in the interval [(pnr-v)*s_i), (pnr+v)*s_i]. The parameter pnr is a percentage that determines the number of reads of each simulated

replicate, and the v parameter allows for some variability in the sequencing depths across simulated samples. Both parameters can be chosen by the users, as well as the number of replicates to be simulated (nss). The recommended values for these parameters are: $nss \geq 5$, pnr=0.2 and v=0.02.

Once the replicates have been simulated, the procedure to estimate differential expression is the same as in NOISeq-real.

5.3.2 NOISeqBIO

In RNA-seq, technical replicates have been reported to present low variability [18, 92]. Therefore, although no real inference can be made from technical replicates, the NOISeq method can serve to estimate differential expression under these circumstances. When biological replication is available, it is crucial to consider the biological variability inherent to each gene under each experimental condition in order to design a method able to detect genes with significant statistical changes in expression between conditions.

Starting from the statistical framework developed in the NOISeq method [141], we proposed a variant named NOISeqBIO that takes into account the biological variability and tests the hypothesis $H_0 : \mu_1 = \mu_2$, where μ_i is the average of expression in experimental condition i. NOISeqBIO incorporates the statistical modeling proposed by Efron *et al.* [43] into the NOISeq formulation. These authors used an empirical Bayes approach on microarray data in which they defined a statistic Z to evaluate the change in expression between two conditions. Their method assumes that the genes can be classified into two different populations: genes with invariant expression between two experimental conditions and genes with expression changing between conditions, so the probability distribution of this Z statistic can be described as a mixture of two distributions. The mixture distribution f of Z can be written as in Equation 5.5, where p_0 is the probability of a gene having the same expression in both conditions and $p_1 = 1 - p_0$ is the probability of a gene having different expression between conditions. f_0 and f_1 are the densities of Z for

genes with no change in expression between conditions and for differentially expressed genes respectively.

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \tag{5.5}$$

If one of these two distributions can be estimated, the probability of a gene belonging to one of the two groups can be calculated.

Our adaptation of this strategy to the NOISeq context consists of the following steps:

1. Choose an appropriate differential expression statistic Z.
2. Estimate the values of the Z statistic when there is no change in expression, i.e. the null statistic Z_0.
3. Estimate the probability density functions f and f_0.
4. Obtain the *posterior* probability of differential expression $p_1(z_i)$ for each gene i.

1. **Differential expression statistic** Z

   Let x_{ij}^g be the expression of gene g in condition i and replicate j. As previously described, we measured the expression change in signal (that is, between both conditions) by computing $M_s^g = log_2(\bar{x}_1^g / \bar{x}_2^g)$ and $D_s^g = \bar{x}_1^g - \bar{x}_2^g$. Let M_s and D_s be the vectors containing M_s^g and D_s^g values for all the genes. M_s and D_s statistics must be corrected for the biological variability by dividing them by their standard errors. Let $\hat{\sigma}_M^2$ and $\hat{\sigma}_D^2$ be the vectors containing the standard errors of M_s and D_s for all genes respectively. They were estimated as follows, assuming that $\bar{x}_1$ and $\bar{x}_2$ are independent.

$$\begin{aligned} \hat{\sigma}_M^2 &= Var(log_2(\bar{x}_1/\bar{x}_2)) = Var(log_2(\bar{x}_1) - log_2(\bar{x}_2)) = \\ &= Var(log_2(\bar{x}_1)) + Var(log_2(\bar{x}_2)) \end{aligned} \tag{5.6}$$

We used a Taylor approximation (δ-method) to estimate the variance [24]: $Var(log_2(X)) = \left(\frac{1}{E(X)log(2)}\right)^2 Var(X)$. For each condition i, we estimated $E(\bar{x}_i) = \bar{x}_i$ and $Var(\bar{x}_i) = S_i^2/n_i$. Hence:

$$\hat{\sigma}_M^2 \approx \frac{1}{\bar{x}_1^2 log(2)^2}\frac{S_1^2}{n_1} + \frac{1}{\bar{x}_2^2 log(2)^2}\frac{S_2^2}{n_2} \tag{5.7}$$

$$\hat{\sigma}_D^2 = Var(\bar{x}_1 - \bar{x}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \tag{5.8}$$

Therefore, M_s and D_s statistics are corrected for biological variability as described in Equations 5.9 and 5.10, where the constant a_0 serves to stabilize the denominator of the statistic [109]. This constant may be computed as a given percentile of all the values in $\hat{\sigma}_M$ or $\hat{\sigma}_D$, respectively, as in [43]. The authors suggest the 90th percentile as the best option.

$$M_s^* = \frac{M_s}{a_0 + \hat{\sigma}_M} \tag{5.9}$$

$$D_s^* = \frac{D_s}{a_0 + \hat{\sigma}_D} \tag{5.10}$$

Finally, we computed the Z statistic by combining the M_s and D_s values. Combining both statistics instead of using the bivariate statistic (M_s^*, D_s^*) allows us to derive a procedure to obtain a probability of differential expression which can be considered equivalent to an adjusted p-value [43]. We considered two different possibilities to combine both statistics (Equations 5.11 and 5.12).

$$Z = \frac{M_s^* + D_s^*}{2} \tag{5.11}$$

$$Z = \frac{D_s^*}{|D_s^*|}\sqrt{(M_s^*)^2 + (D_s^*)^2} \tag{5.12}$$

2. **Null scores** Z_0

   Let **X** be the gene expression matrix with G rows (genes) and $n_1 + n_2$ columns, where n_i is the number of biological replicates for condition i. We assume that **X** is normalized (e.g. by converting the raw counts to TMM [120] or RPKM [99] values) and that genes with no expression across all the replicates for both conditions have been removed.

   In order to compute the null density f_0, we first need to estimate the values of the Z-scores for genes with no change between conditions. To estimate these null scores, we permuted the labels of samples (columns) in **X** r times and each time we computed the differential expression statistic Z, as previously described. In this way, we obtained a matrix $\mathbf{Z}_0$ with as many columns as the number of permutations r. The $G \times r$ elements of matrix $\mathbf{Z}_0$ are pooled together and are considered the null scores that will be used to later estimate the null distribution f_0.

   However, when there are few replicates available for each condition, the resulting null distribution is very poor because the number of different permutations is low. Therefore, when the number of replicates per condition is less than 5, it is convenient to borrow information from across genes. The procedure we followed was to cluster all the genes according to their expression values across replicates by using the k-means algorithm. Hence, genes with similar expression values across all replicates were clustered together. For each cluster of genes (k), we considered the expression values of all the genes in the cluster as observations within the corresponding condition (i.e. as replicates). Thus, we shuffled this submatrix $r \times g_k$ times, where g_k is the number of genes within cluster k. For each permutation, we calculated (M, D) values and their corresponding standard errors. In order to reduce the computing time and to get a refined clustering of the greatest clusters, if $g_k \geq 1000$, we re-applied the k-means algorithm to subdivide cluster k into subclusters.

3. **Estimation of densities**

   Once the Z and Z_0 scores have been obtained, the density functions f and f_0 can be estimated. In [43], they directly estimate the ratio f_0/f using nonparametric logistic regression. In NOISeqBIO, we propose to estimate f and f_0 separately using a kernel density estimator (KDE) with a Gaussian kernel (see Section 3.2.2). By default, the smoothing parameter (adj in noiseqbio() function) was set to 1.5, which means that the bandwidth is computed as 1.5 times the optimum bandwidth obtained by "'nrd0" Silverman's rule of thumb [129]. Therefore, the density curves we estimate are smoother than the default curves generated by the R density() function. In Section 5.4 we show that using this KDE option in NOISeqBIO improved the performance of the method.

4. **Probability of differential expression**

   Given a gene with a score z for the Z statistic, let $p_1(z)$ be the probability of that gene being differentially expressed between the two experimental conditions being compared. Therefore, $p_1(z)$ is the conditional probability of differential expression for an observed value of z for a given gene. Thus, this probability can be derived from Bayes Rule as follows:

   $$p_1(z) = \frac{p_1 f_1(z)}{f(z)} = 1 - p_0 \frac{f_0(z)}{f(z)} \tag{5.13}$$

   Moreover, as Efron *et al.* showed [43], the *a posteriori* probability $p_0(z) = 1 - p_1(z)$ we calculate in Equation 5.13 is closely connected to the FDR defined by Benjamini and Hochberg [11], so $p_0(z)$ can be considered equivalent to a multiple testing adjusted p-value.

   Thus, we only need to estimate p_0 in order to calculate $p_1(z)$ because we already estimated f_0 and f. We took an upper bound of p_0 as suggested in [43]. Taking into account that $p_1(z)$ must be nonnegative leads to the restriction $p_0 \leq \min_Z\{f(Z)/f_0(Z)\}$, which can be used as the estimate for p_0.

According to [43], more stable upper bounds can be constructed by integrating over an interval I near $Z = 0$. Then, $p_0 \leq \frac{\int_I f(Z)}{\int_I f_0(Z)}$. The choice $I = [-0.5, 0.5]$ is recommended, particularly when the true p_0 is near 1, which usually happens in differential expression analysis. This upper bound can be directly estimated by $\frac{\gamma_I(Z)/|Z|}{\gamma_I(Z_0)/|Z_0|}$, where $\gamma_I(X)$ is the number of X values inside the interval I. However, when applied to our work, this upper bound for p_0 sometimes resulted in negative probabilities, and so we discarded this option.

5.3.3 Other differential expression methods

NOISeq and NOISeqBIO were compared to other differential expression methods that are summarized in the following paragraphs.

These are the methods NOISeq was compared to. All of them are parametric except Fisher's exact test (FET).

- Fisher's exact test. This procedure is used when no biological replicates are available to assess if the percentage of reads falling in a given gene is significantly different for the two experimental conditions. Thus, a FET is performed on each gene and the resulting p-values are corrected for multiple testing using the Benjamini and Hochberg procedure [11].

- DEGseq [147] is a Bioconductor R package which integrates three existing differential expression methods and another two methods which were developed by the authors. In this work, we chose two of them:

  - DEGseq-LRT: RNA sequencing can be modeled as a random sampling process, in which each read is sampled independently and uniformly [67]. Under this assumption the number of reads coming from a gene follows a binomial distribution (and can be approximated by a Poisson distribution). A likelihood ratio test (LRT) based on a Poisson distribution had previously been proposed to identify differentially expressed genes [15, 92]. This test was implemented inside the DEGseq package.

  - DEGseq-MARS: MA-plot is a widely used tool to detect the dependence of fold-change (M) on intensity (A) for microarray data that can also be extended to RNA-seq data. Let C_i denote the counts of reads mapped to a specific gene in sample i, with $C_i \sim Binomial(n_i, p_i)$, $i = 1, 2$, where n_i is the total number of mapped reads and p_i is the probability of a read coming from that gene. M is defined as $M = log_2 C_1 - log_2 C_2$, and $A = (log_2 C_1 + log_2 C_2)/2$. The authors proved that the conditional distribution of M given that $A = a$ follows an approximate normal distribution. For each gene on the MA-plot, they do the hypothesis test of $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$.

- baySeq [59]: An empirical Bayesian approach to estimate the posterior probabilities of models that define different patterns of differential expression across experimental groups. First, the models are defined by indicating which samples behave similarly to each other, and for which sets of samples there are identifiable differences. Samples behaving similarly to each other should possess the same prior distribution on the underlying parameters of that tuple, while samples behaving differently should possess different prior distributions. The method is based on either a Poisson distribution (PO) or a Negative Binomial distribution (NB) for the tuple data, and derives an empirical distribution in the set of underlying parameters from the whole dataset.

- edgeR [123] models count data using an over-dispersed Poisson model, i.e. a Negative Binomial distribution, and uses parametrization to relate the mean μ and the variance σ^2: $\sigma^2 = \mu + \alpha\mu^2$. In this way, only the α parameter has to be estimated from the data. Gene-wise dispersions are estimated by conditional maximum likelihood, conditioning on the total count for that gene. An empirical Bayes procedure is used to shrink the dispersions towards a consensus value, effectively borrowing information from across genes [121]. Finally, differential expression is assessed for

each gene using an exact test analogous to Fisher's exact test, but adapted for over-dispersed data [122]. In the technical replication study, two variants of this method were considered:

  - CD, which estimates a Common Dispersion for all tags.
  - TWD, which estimates a Tag-Wise dispersion.

- DESeq [4] is a statistical procedure similar to edgeR. The authors extend the edgeR model by allowing more general, data-driven relationships of variance and mean, and they use a different procedure to estimate the dispersion.

NOISeqBIO was compared to the following methods. Because the NOISeqBIO study was more recent, new versions of the methods were available with improved performance.

- edgeR [123] (see description above).
- DESeq2 is a more recent and improved version of DESeq [4].
- SAMseq [82] is a non-parametric procedure that uses resampling to account for the different sequencing depths. The method is based on the Wilcoxon statistic to compare two samples and a permutation approach is applied to approximate the distribution of this statistic.

5.3.4 Data processing

In NOISeq comparisons, a library size correction was applied to compute differential expression whenever the methods included this option. In some cases, gene length correction was also applied, where gene length was computed as the median length of transcripts of each gene. Since NOISeq-sim works with no replicates, counts from different lanes were summed up so there was always a unique replicate but with a different sequencing depth in each case.

In the studies related to NOISeqBIO, we applied the following normalization and low count filtering procedures on both simulated and experimental

data. TMM normalization was used for NOISeqBIO and edgeR differential expression methods. For DESeq2 and SAMseq we used their own normalization algorithms. Prior to normalizing the data, we filtered out genes with an average expression per condition lower than 1 count per million in both conditions. Filtering was done using the CPM method from the NOISeq R package (see Chapter 4 for more details).

5.3.5 Tools for performance assessment

5.3.5.1 Performance indicators

The performance of the DE methods was assessed according to the values of some indicators previously defined in Chapter 3: Sensitivity (SE), False Discovery Rate (FDR) and Matthews Correlation Coefficient (MCC). These indicators were either used for the plots described in the following text or computed for the DE results obtained at a significance level of 5% (which is the most frequently used by researchers).

5.3.5.2 Precision-recall curves and false discovery rate plots

Precision-recall curves (PRC) and False Discovery Rate (FDR) plots were generated for both simulated and RT-PCR datasets. "Recall" is the sensitivity or true positive rate (TPR) and "Precision" is defined as $TP/(TP+FP)$, making it equal to 1-FDR. PRCs are good performance estimators when the number of negatives greatly exceeds the number of positives, as is the case of expression datasets [35].

5.3.5.3 Box plots

In some cases, to compare the performance of DE methods, we used box plots with a notch at both sides of the box. This notch represents a sort of 95% confidence interval for the median computed from Equation 5.14, where IQR

is the interquartile range and n is the number of observations represented in a given box plot.

$$Median \mp 1.58 \times IQR/\sqrt{n} \tag{5.14}$$

This confidence interval was proposed by McGill *et al.* [96] and provides a way to measure the difference between medians. If two of these confidence intervals (for two different methods) do not overlap, it means that the medians are, roughly, significantly different at an approximate confidence level of 95%. The assumptions made in order to construct such confidence intervals are: asymptotic normality for the median and equal (or very similar) sample sizes for the two medians being compared. In principle the procedure should not be sensitive to the underlying distributions of the samples. In this way, we can use these intervals to test the null hypothesis that the true medians are equal. However, when more than two samples are compared, it must be taken into account that no multiple testing correction is applied [25].

5.4 Results

5.4.1 NOISeq performance

We compared NOISeq to a selection of RNAseq differential expression methods, namely edgeR [123], baySeq [59], DEGseq [147], DESeq [4] and FET on data with technical replicates. In contrast to NOISeq which makes no assumptions about the distribution of the M and D statistics, all these methods follow parametric approaches (except FET).

5.4.1.1 Comparison on simulated data

To concentrate the comparison on a subset of representative and but different approaches, we first evaluated the general performance of all of these differential expression methods and compared them to NOISeq using the synthetic dataset included in the baySeq R package [59].

Both PRC and FDR plots separated RNA-seq statistical methodologies into two groups (Figure 5.6). The two options for the edgeR method (CD and TWD), DESeq, baySeq-NB and both versions of NOISeq showed good accuracy and consistent control of false discoveries, whereas MARS, LRT, FET, and baySeq-PO showed poorer performances. Therefore, we selected edgeR-CD, baySeq-NB, DESeq, and FET (the last one was included for its extensive use by researchers), together with NOISeq-sim and NOISeq-real for further analysis.

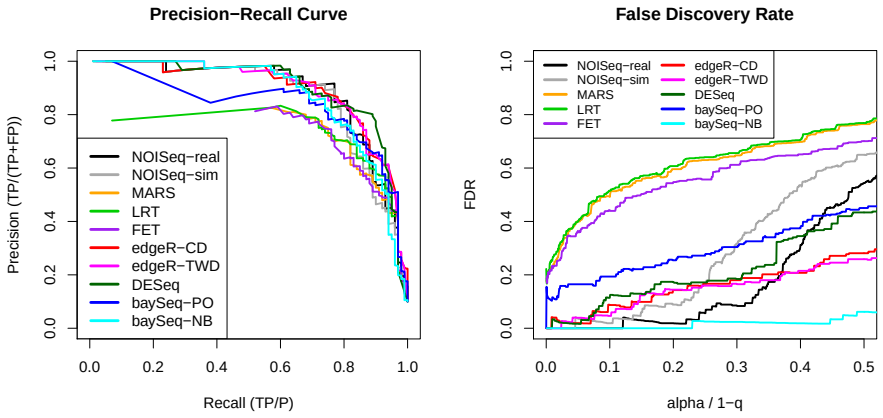


Figure 5.6: Precision-Recall and False Discovery Rate curves for the differential expression methods compared, as applied to the synthetic dataset included in the baySeq R package.

5.4.1.2 Comparison on experimental data

The selected methodologies were applied to the **MAQC** and **Griffith** datasets. We also included the analysis of gene length corrected data when the methods permitted this input. Note that FET was applied on counts normalized by the library size.

On the **MAQC** dataset, two performance indicators, PRC and FDR indicated that NOISeq performed better compared to other methodologies (Figure 5.7). Specifically, false discoveries were higher for edgeR, DESeq and baySeq.

FET had a low FDR regardless of the significance threshold but also showed a poorer precision-recall figure. Interestingly, PRC and FDR were very similar on data with and without length correction. **Griffith** RT-PCR data were more limited but led to the same conclusions (Table 5.1).

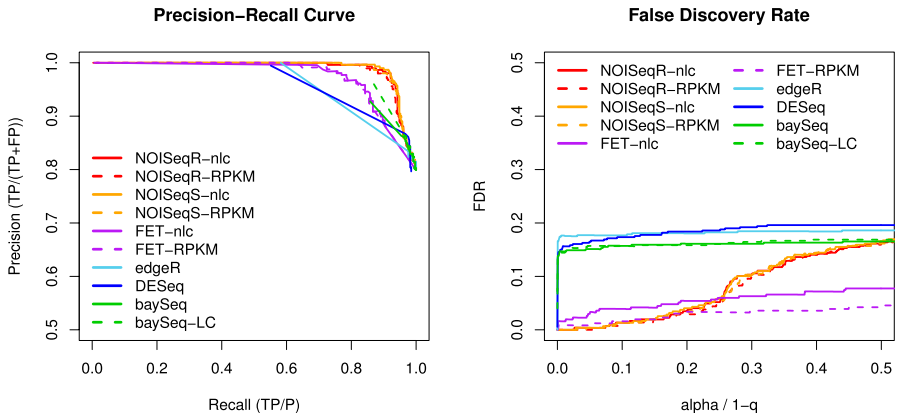


Figure 5.7: Precision-recall curves and false discovery rates for the differential expression methods compared on the **MAQC** dataset using RT-PCR results as a gold-standard.

In summary, our performance analysis highlighted differences between RNA-seq differential expression methods and pointed to NOISeq as a high performing methodology.

We also investigated how the number of technical replicates influences the number of differential expression calls (Figure 5.8), the gene length, the fold-change (M) and the mean expression level of DEGs (Figure 5.9).

Figure 5.8 shows the very pronounced dependency between gene selection and number of replicates (lanes) observed for edgeR, DESeq and baySeq. FET did not show this dependency but did identify a reduced number of significant genes. NOISeq had an intermediate behavior with a moderate number of DEGs in the **MAQC** dataset which increased only slightly with the number of replicates. Results for **Griffith** data were slightly different. While FET and NOISeq identified a small number of significant genes (between 150 and 200), close to the figure reported in the original paper, other methods

Method	Length correction	# TP	TPR	# FP	FPR
NOISeq-real	None	47	57.3%	0	0.0%
NOISeq-real	RPKM	35	42.7%	0	0.0%
NOISeq-sim	None	63	76.8%	2	16.7%
NOISeq-sim	RPKM	61	74.4%	0	0.0%
FET	None	14	17.1%	0	0.0%
FET	RPKM	9	11.0%	0	0.0%
edgeR	None	73	89.0%	5	41.7%
DESeq	None	70	85.4%	4	33.3%
baySeq	None	58	70.7%	3	25.0%
baySeq	Yes	61	74.4%	3	25.0%

Table 5.1: True and false positive rates for differential expression methods applied to the **Griffith** dataset and using RT-PCR as a gold-standard. Genes considered as true positives are those declared as differentially expressed on RT-PCR data (82 genes). Genes considered as true negatives are those not found to be differentially expressed by RT-PCR (12 genes).

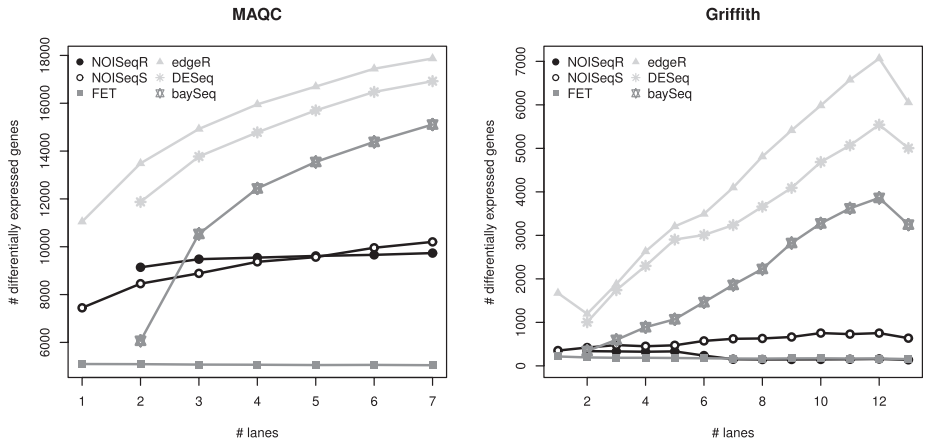


Figure 5.8: The number of differentially expressed genes according to the number of technical replicates for each dataset and method. No gene length correction was applied to the data.

resulted in larger selection sets. Moreover, both FET and NOISeq-real lost significant calls as more lanes were considered, reflecting the high variability of this dataset.

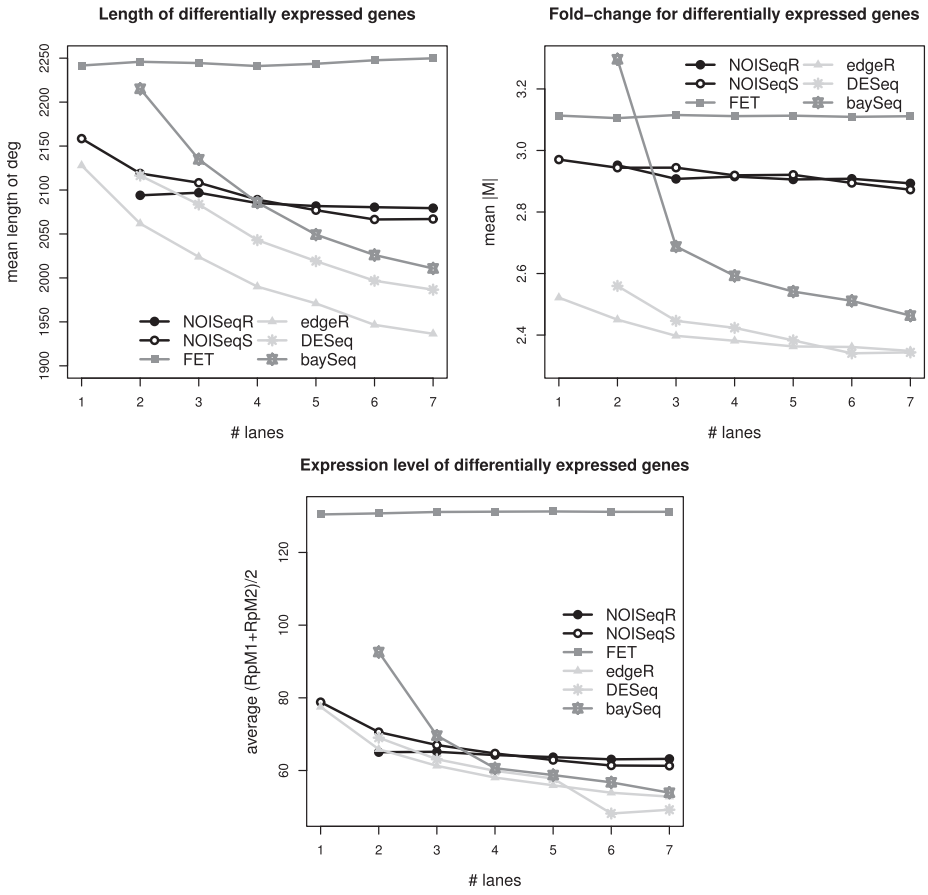


Figure 5.9: Relationship between gene length, fold-change M, expression level of differentially expressed genes, and the number of lanes used, for each method using the **MAQC** data set. No length correction was applied to the data. RpM_i is the number of reads per million reads in condition i.

Regarding gene length, fold-change and mean expression (Figure 5.9), the dependence of the results on the number of replicates was once again stronger for parametric methods. The mean gene length decreased as the number of lanes grew. This length shortening effect was only very moderately present in NOISeq which, at the highest sequencing depths, generally selected larger genes than the other methods. The mean fold-change of the genes detected by parametric methodologies was smaller for a larger number of replicates. On the contrary, NOISeq selected genes with larger count differences and

behaved robustly with changing sequencing depth. Finally, we also observed a strong dependency on the level of expression. Current RNA-seq statistical methods tend to identify genes with a lower relative abundance as the number of available replicates grows. Again here NOISeq, and especially NOISeq-real, gave more constant and intermediate results, selecting genes with lower expression and genes with larger count numbers than parametric RNA-seq methods when a lower or higher number of lanes were used respectively. FET had large and constant values for these three parameters.

Previous results indicated that the number of DEGs identified by parametric approaches strongly increases in number as more technical replicates are used for the analysis. Although this could be explained by an apparent higher accuracy of gene expression estimates in large sampling sizes, the prominent discrepancy with a data-driven methodology such as NOISeq, along with the results from our initial performance analysis, led us to suspect a general failure of these methods in controlling the FDR as the sequencing output increased. To verify this, we analyzed the available **MAQC** RT-PCR data as a function of the number of replicates, looking both at the false (FPR) and true (TPR) positive rates. As suspected, current RNA-seq analysis methods progressively incorporated more false calls as more lanes were used, reaching more than 60% false positives using the edgeR method (Fig. 5.10). In contrast, NOISeq maintained a stable and low FPR even as the number of lanes increased. Only FET had better FPR performance, however at a significant cost of the number of true detections. The TPR obtained from the other methods compared was slightly higher than that of NOISeq, which is logically the consequence of the large number of the DEGs called by these methodologies. Notably, genes selected by both NOISeq and other approaches did contain a functional signature, i.e., they were significantly enriched in many biological functions while those only detected by parametric methods had no specific functional charge (Table 5.2).

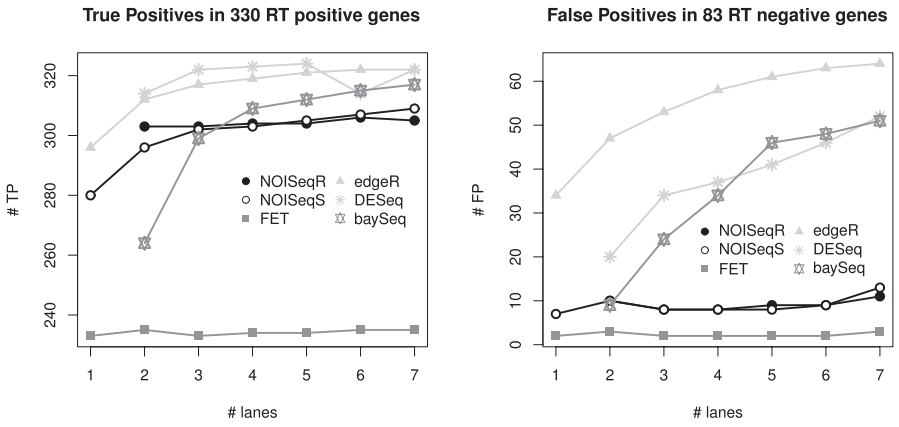


Figure 5.10: The relationship between the number of true positives (TP) and false positives (FP) and the number of technical replicates (lanes). TP and FP were obtained by applying the statistical methods to be compared to the MAQC dataset and then comparing results to RT-PCR positive and negative genes.

5.4.2 NOISeqBIO performance

In this section, we evaluate the performance of the NOISeqBIO method. Prior to the comparison of NOISeqBIO to other DE methodologies, we show the results of some preliminary studies that had two main purposes. Firstly, they served to adjust the various options and parameters of NOISeqBIO that we described in Section 5.3.2. Secondly, and in order to simplify the huge number of biological scenarios that could be defined when simulating data, we identified which parameters in the simulation algorithm had a greater influence on DE results and which were the most informative values for these parameters. This information was then used to generate the final set of simulated datasets to compare the DE methods.

5.4.2.1 Preliminary studies on simulated data

Determining the best options for NOISeqBIO

The purpose of the first study on simulated data with biological replications was to determine the best choice for the parameters and options

	edgeR	DESeq	baySeq
In common between NOISeq and the other method	**9735**	**9693**	**9712**
Up in BRAIN	3468	3431	3457
Up in UHR	6267	6262	6255
# GO terms (Up in BRAIN)	192	178	190
# GO terms (Up in UHR)	486	481	485
Detected by the other method and not by NOISeq	**137**	**7230**	**5398**
Up in BRAIN	2731	3707	1826
Up in UHR	5406	3517	3572
BRAIN = UHR	0	6	0
# GO terms (Up in BRAIN)	0	0	2
# GO terms (Up in UHR)	0	4	1
Detected by NOISeq and not by the other method	0	42	23

Table 5.2: Comparison of the genes declared as differentially expressed by NOISeq and by the other methods, including the number of significantly enriched GO terms for each set of genes.

considered in NOISeqBIO. These optimized options were then used in the remaining analyses.

In this study, we simulated data from an *A. fumigatus* sample for 9862 genes, with 5% DEGs and 0% noise. Data were normalized using the TMM method. A gene was declared as differentially expressed when its probability of having differential expression was higher than 0.95 which is, as we previously stated, equivalent to $FDR = 0.05$ [43].

We compared the performance of the method for datasets with 5 and 10 replicates and simulated 10 different datasets for each number of replicates, checking the following values for the input parameters:

- Differential expression statistic:
    1. Mean of M^* and D^* values ($stat = 1$).
    2. Distance of (M^*, D^*) to the origin with the sign of the difference D^* ($stat = 2$).

- Method to estimate the densities f and f_0:
  1. Kernel density estimators ($dens = 1$).
  2. Logistic regression using natural splines ($dens = 2$).
- Value for a_0:
  1. P_{50} of the values of the standard deviation of M or D, respectively, for all the genes ($a_0 = 0.5$).
  2. P_{90} of the values of the standard deviation of M or D, respectively, for all the genes ($a_0 = 0.9$).
- Number of sample label permutations to generate the noise distribution: 20, 50 and 70.

As shown in Figures 5.11 to 5.14, NOISeqBIO generally performed well in terms of sensitivity (SE) and FDR for simulated data sets with 5 or 10 replicates. When using the mean of M^* and D^* as the differential expression statistic ($stat = 1$), we found no big differences for $a_0 = 0.5$ and $a_0 = 0.9$, although perhaps P_{90} produced slightly better results. KDE ($dens = 1$) improved logistic regression ($dens = 2$) for density estimation, especially in the 5-replicate case. In addition, for the KDE option, the number of permutations does not seem to influence the results.

Therefore, in the following studies we took the mean as the differential expression statistic, KDE to estimate the densities, and used 50 permutations; a more thorough analysis was done to choose the most proper value for a_0.

A second simulation study was undertaken to assess the performance of NOISeqBIO in terms of SE and FDR according to the level of noise in the data (0, 0.2 or 0.4) and the proportion of DEGs (0.01, 0.05,0.10), and also to determine the best value for a_0. We compared the following values of a_0: 0, P_{25}, P_{50}, P_{70}, P_{90} and B, which means that a_0 is 100 times the standard deviation. Again, we simulated data from an *A. fumigatus* sample and took a cutoff of 0.95 for the probability of differential expression.

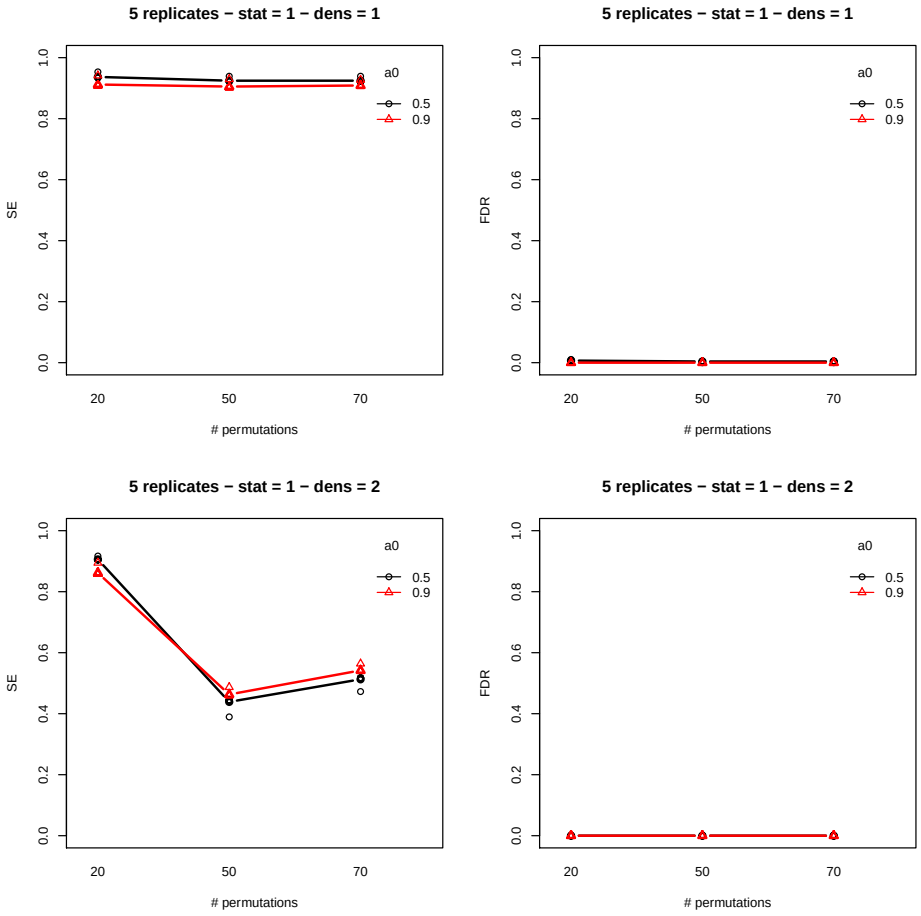


Figure 5.11: Comparison of different options in NOISeqBIO used on simulated datasets with 5 replicates. Differential expression statistic was the mean of (M^*, D^*) values.

When considering 5 or 10 replicates per condition, the method performed best for lower levels of noise and higher proportions of DEGs, as expected. SE was close to 1 in most cases and was always higher than 0.8. However, in general, the FDR drastically rose when there were high levels of noise, and was more dependent on the value of a_0. The best choice for a_0 was once again P_{90}, except in the case of 10 replicates and $noise = 0.4$, where option B clearly outperforms the rest (see Figures 5.15-5.20). Thus, we set a_0 to

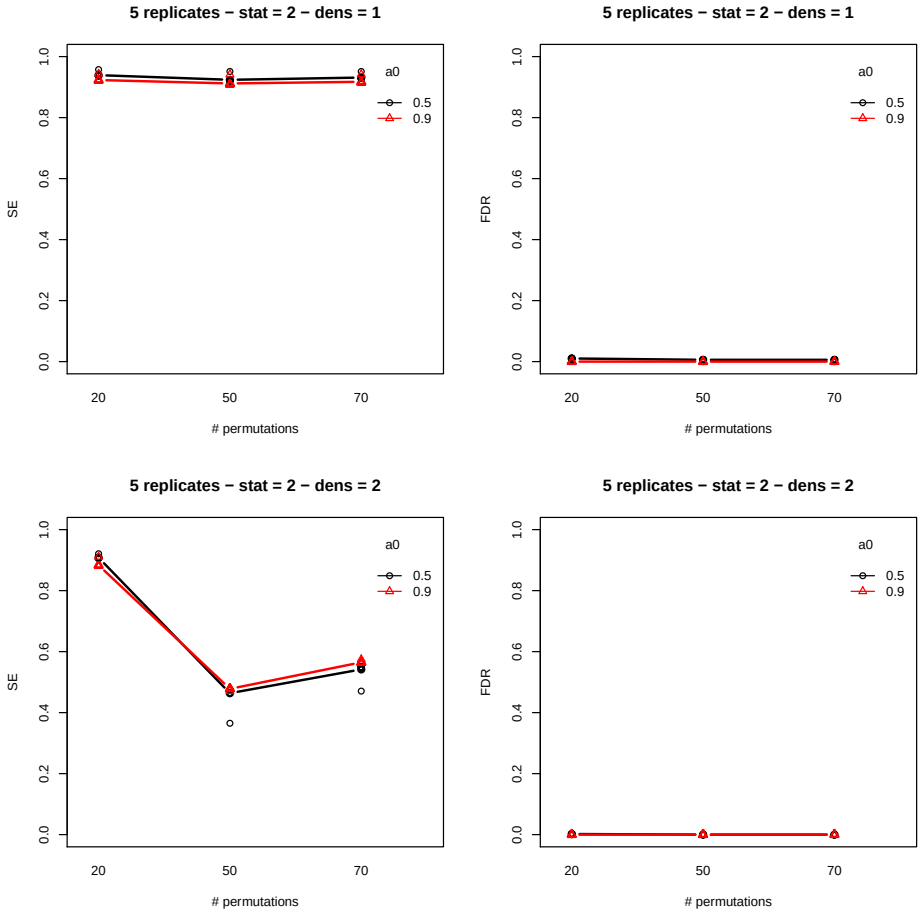


Figure 5.12: Comparison of different options in NOISeqBIO used on simulated datasets with 5 replicates. The differential expression statistic was the distance to the origin of (M^*, D^*) values.

P_{90} for our subsequent studies.

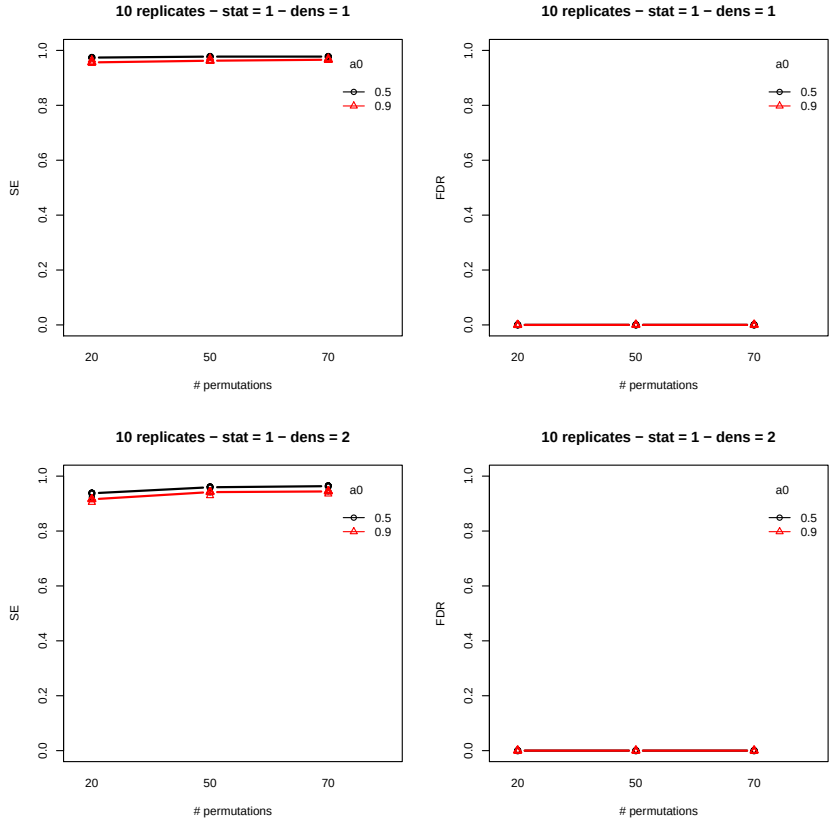


Figure 5.13: Comparison of different options in NOISeqBIO used on simulated datasets with 10 replicates. The differential expression statistic was the mean of (M^*, D^*) values.

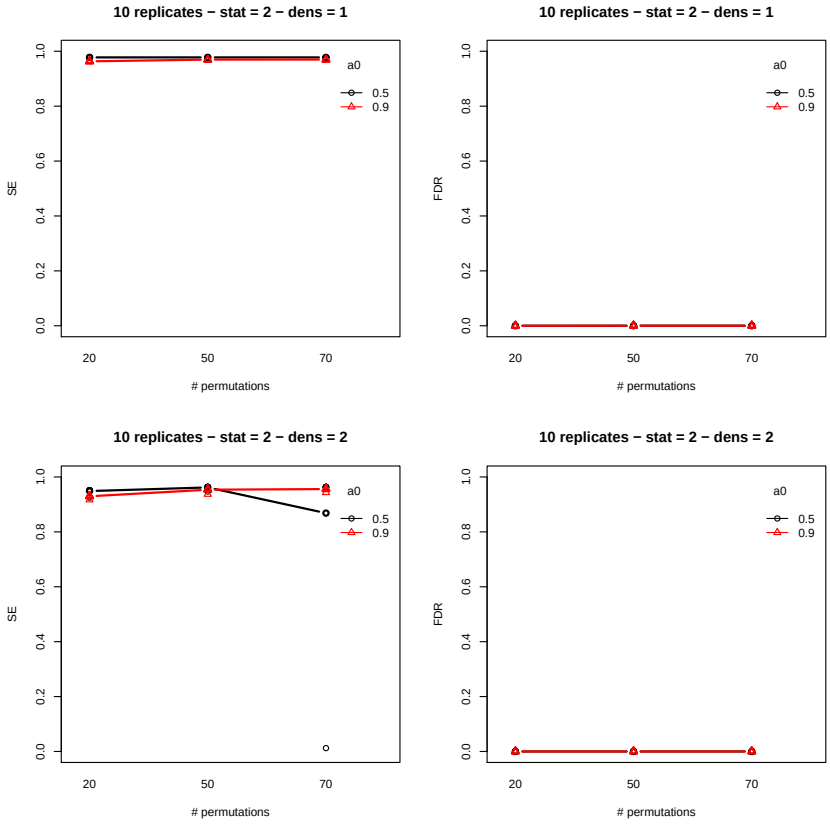


Figure 5.14: Comparison of different options in NOISeqBIO used on simulated datasets with 10 replicates. The differential expression statistic was the distance to the origin of (M^*, D^*) values.

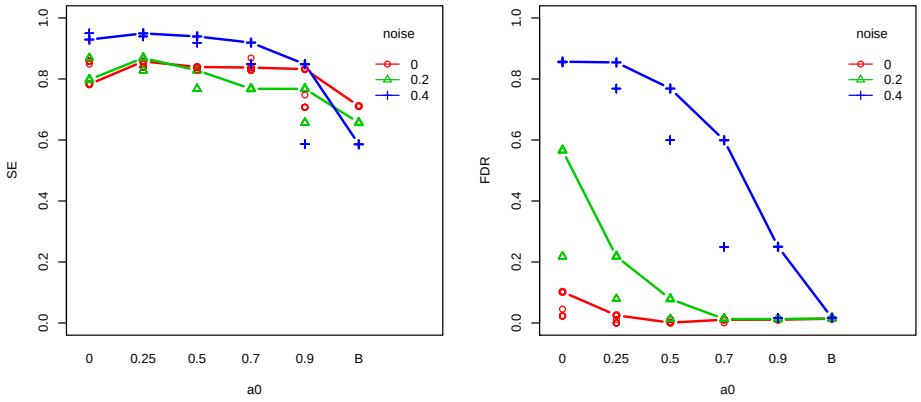


Figure 5.15: Performance of NOISeqBIO applied on simulated data with 5 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 1%.

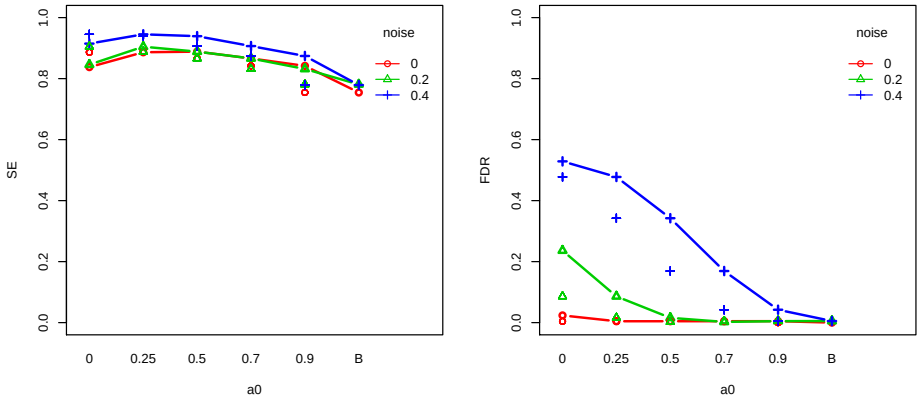


Figure 5.16: Performance of NOISeqBIO applied on simulated data with 5 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 5%.

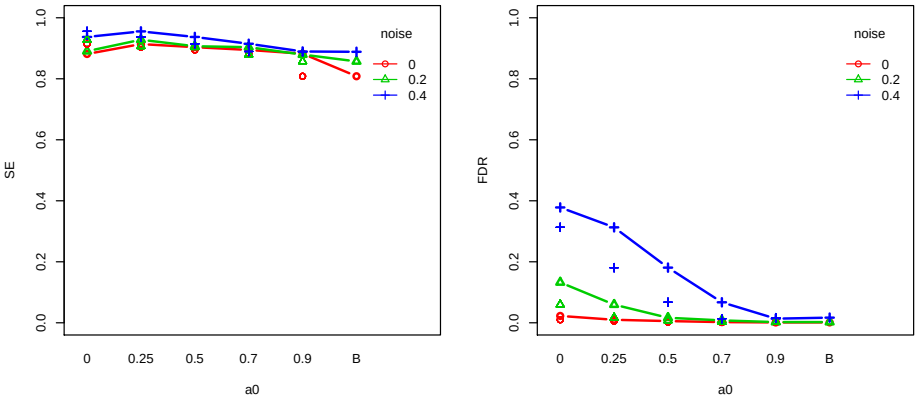


Figure 5.17: Performance of NOISeqBIO applied on simulated data with 5 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 10%.

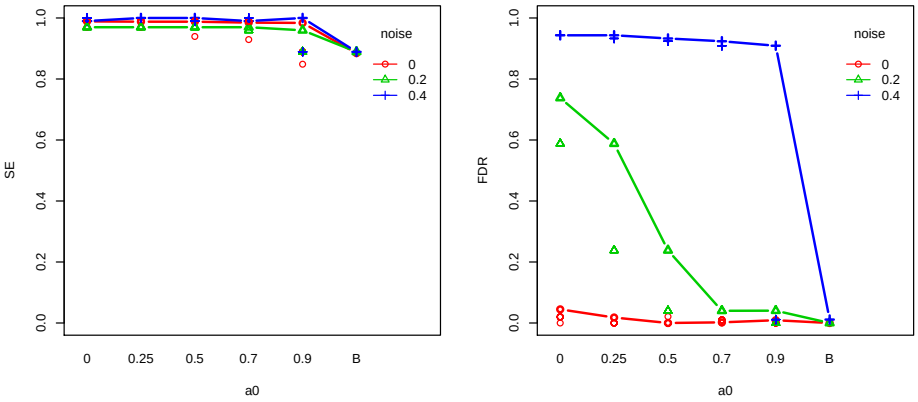


Figure 5.18: Performance of NOISeqBIO applied on simulated data with 10 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 1%.

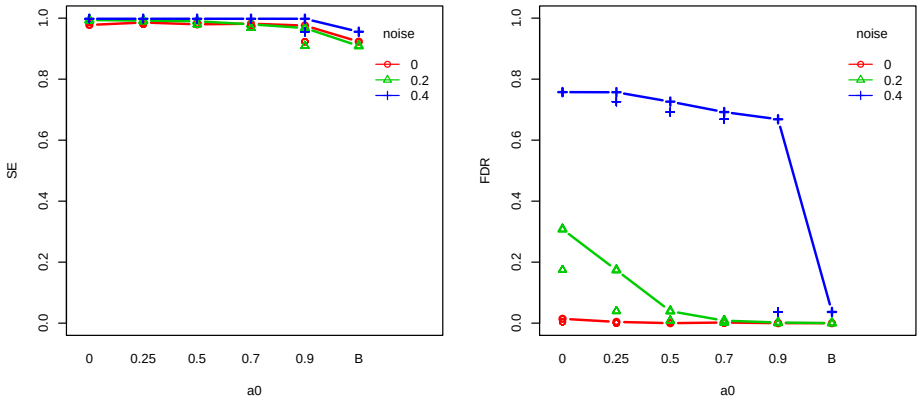


Figure 5.19: Performance of NOISeqBIO applied on simulated data with 10 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 5%.

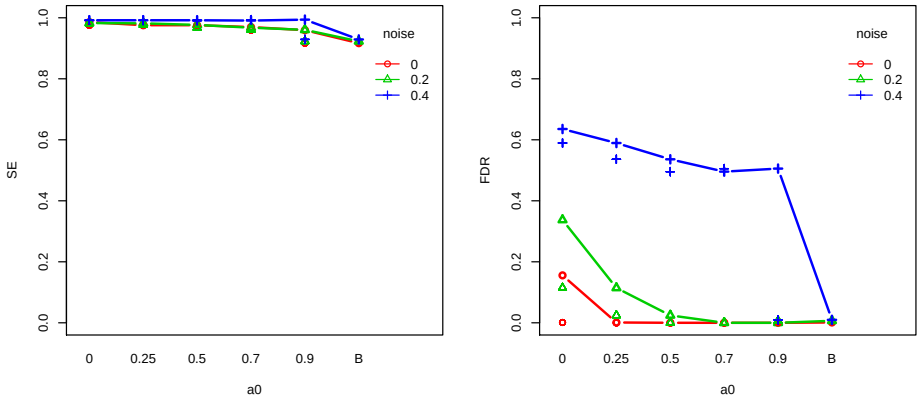


Figure 5.20: Performance of NOISeqBIO applied on simulated data with 10 replicates according to different levels of noise and a_0 values. Percentage of DEGs is 10%.

NOISeqBIO for few replicates

Results derived from the previous simulation studies showed that NOISeqBIO behaved very poorly when the number of replicates per condition was less than 5, no matter which options were chosen, or what parameters were used (results not shown). This is logical because the algorithm is based on permuting the sample labels to generate the noise distribution and so few replicates result in bad estimations of the noise distribution, as occurs in non-parametric methods based on resampling. Thus, we modified the algorithm for cases in which the number of replicates was less than 5. When this happens, genes are clustered according to their expression values and resampling is done within each cluster, considering gene expression values in the same cluster and condition as the replicates of that condition (see Section 5.3.2 for more details).

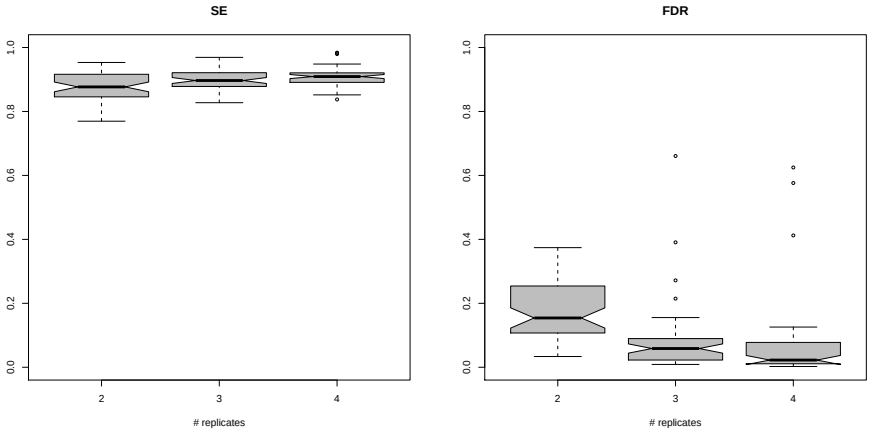


Figure 5.21: Performance of NOISeqBIO using simulated data sets with 2, 3, or 4 replicates. The parameters for the simulations were: $propdeg = 0.05$ and $noise = 0$ or 0.3. Three data sets were generated for each scenario, organism, and number of replicates.

In this section, we evaluated the performance of NOISeqBIO algorithm for the case where there are few replicates. We simulated data from real samples with different number of genes: *A. fumigatus*, *F. oxysporum* and *H. sapiens*, and studied the effect of the number of replicates on SE and FDR.

Figure 5.21 shows the expected increase in SE and decrease in FDR when the number of replicates was increased from 2 to 4.

We also evaluated what the optimum number of clusters (k) was when applying the k-means algorithm (Figure 5.22). The SE and FDR values were computed for $k = 10, 15$ and 20 clusters in different simulated scenarios but the results were not significantly different for 10 or 15 clusters. SE slightly improved for 20 clusters but at the cost of also increasing the FDR. Therefore, we decided to set the number of clusters to $k = 15$ by default, although users can vary this parameter. We also used 15 clusters for further comparisons with NOISeqBIO in this work.

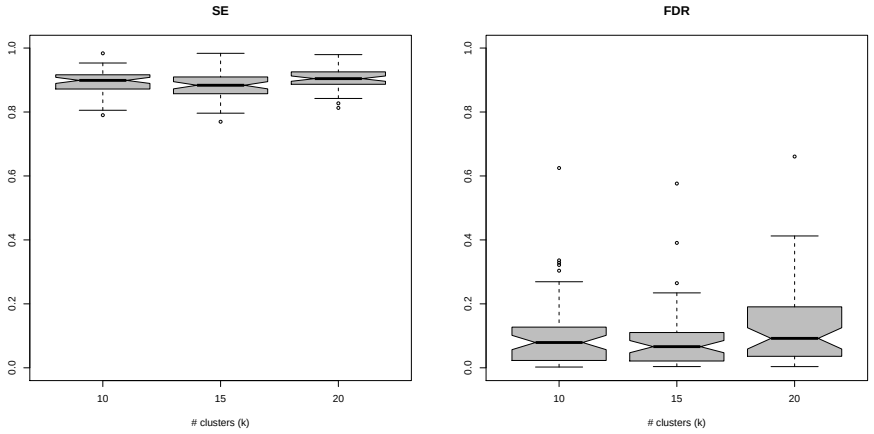


Figure 5.22: Performance of NOISeqBIO using simulated data sets for different numbers of clusters. The parameters for the simulations were: $propdeg = 0.05$, and $noise = 0$ or 0.3. Three data sets were generated for each scenario, organism and number of replicates (2, 3 or 4).

Influence of simulation parameters

Whilst still in the preliminary phases of the study, we analyzed the effects of some simulation parameters on the performance of five differential expression methods: NOISeq, NOISeqBIO, edgeR, DESeq and SAMseq. These are the simulation parameters we studied and their values:

- The β parameter, which determines the magnitude of the change in expression between conditions: $\beta = 5, 6, 7$.

- The proportion of DEGs: 0.01, 0.05 and 0.1.
- Noise, which is the percentage of deviation with regard to the average expression that is allowed: 0, 20% and 40%.

For each combination of parameter values, 10 datasets were generated from a sample of *A. fumigatus* RNA-seq data with 5 replicates per condition and a sequencing depth of about 30 million reads. We used a significance level of 0.05 for the methods that returned p-values, and a threshold of $q = 0.8$ for NOISeq and $q = 0.95$ for NOISeqBIO. Note that we included NOISeq in this preliminary comparison to see if it was outperformed by NOISeqBIO. However, differential expression probabilities returned by NOISeq are not equivalent to p-values so in this case we used the recommended cutoff for the differential expression probability ($q = 0.8$). We evaluated the performance of the methods by measuring the SE, the FDR and the MCC.

We observed no influence of the β parameter on the performance of the methods (see Figure 5.23), so we set $\beta = 6$ for further evaluations. However, the proportion of DEGs, the level of noise, and the method applied all had strong effects as shown in Figure 5.24, which shows MCC values for noise $= 0.2$ and noise $= 0.4$ (results for noise $= 0$ are not shown because they were very similar to noise $= 0.2$). In addition, we corroborated that the performance of NOISeqBIO (Bio4 in Figure 5.24) was better than NOISeq on data with biological replicates.

We also observed that sequencing depth had no influence in the performance of the methods, since data are normalized to correct this effect prior to computing differential expression (data not shown). Therefore, we used a default value of 30 million reads for further studies. As expected, the number of replicates had an important effect on the results so we paid special attention to this parameter in the following comparisons.

5.4.2.2 Comparison on simulated data

For this final comparison of DE methods on data with biological replicates, we also used the simulation algorithm described in Section 5.2.2.2 to emu-

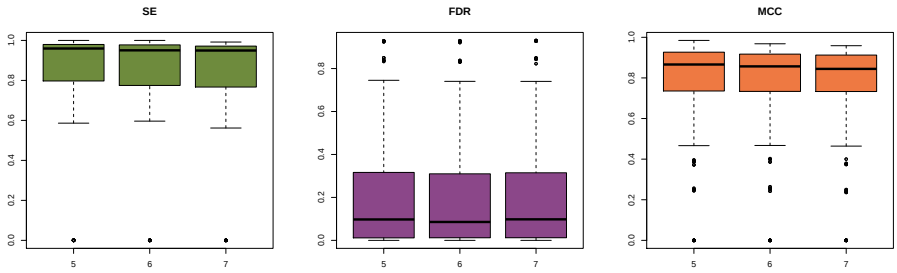


Figure 5.23: The effect of the β parameter on the performance of several methods (NOISeq, NOISeqBIO, edgeR, DESeq, and SAMseq) on datasets simulated from samples in an *A. fumigatus* experiment. The parameters for the simulation were: $ngenes = 9862$, $nrepl = 5$ in both conditions, $depth = 30$ million, $noise = 0, 0.2, 0.4$, and $propdeg = 0.01, 0.05, 0.1$.

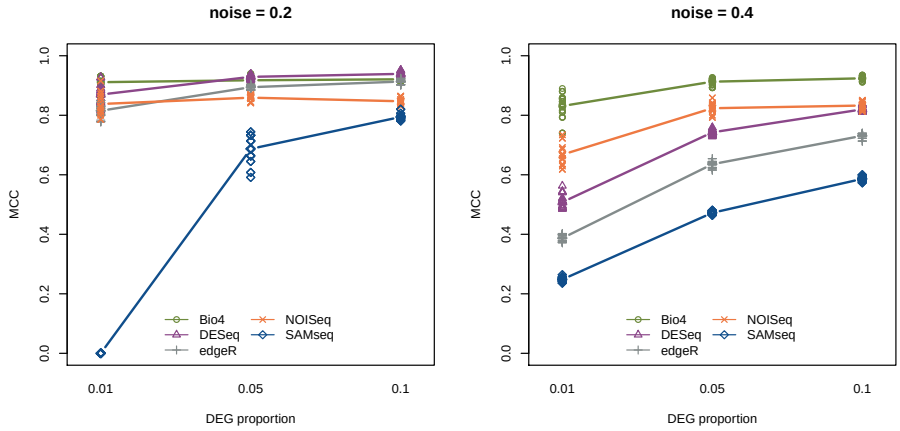


Figure 5.24: Performance measured by MCC of several methods when applied to datasets simulated from samples in an *A. fumigatus* experiment. The parameters for the simulation were: $ngenes = 9862$, $nrepl = 5$ in both conditions, $depth = 30$ million. Data were normalized with the TMM method.

late different biological scenarios of high and low variability. The values of the simulation parameters were defined according to the results from studies described in the previous section:

- **Organism**: The data were simulated from either **FO** or **HS** experi-

mental samples, giving rise to data with different numbers of genes ($ngenes$) and initial count distributions (μ_0).

- **Noise**: We considered no noise (0) and 20% noise (0.2).
- **Replicates**: The number of replicates per condition is decisive for the performance of statistical methods. We considered data with few replicates (2 or 3 replicates), which are still quite common in RNA-seq experiments, and data with 5 and 10 replicates.
- **DEG**: The percentage of differentially expressed genes was set to either 5% or 10%.

We generated 10 datasets per each combination of these parameter values, resulting in a total number of 320 simulated datasets for each scenario of biological variability (high and low). To assess the performance of the methods being compared when setting a given adjusted p-value cutoff (e.g. 0.05), we computed the SE, FDR and MCC.

Figures 5.25 and 5.26 considered all simulated scenarios with high biological variability and showed that SE is, in general, higher than 90% for all methods, except for SAMseq. This method fails to detect any DEGs on low replication experiments (2-3 replicates per condition) (Figure 5.25), although it tends to improve for higher numbers of replicates. Differences in performance for the other three methods can be observed in the boxplots but were also analyzed by an ANOVA model with repeated measures for each of the three indicators (SE, FDR and MCC) on results excluding the SAMseq method. Tukey post-hoc tests revealed that NOISeqBIO significantly outperforms the other two methods in terms of FDR, except in the 2 replicates case, where no significant differences compared to edgeR were observed. In contrast, parametric methods (edgeR and DESeq2) have significantly higher SE than NOISeqBIO (except for data with 10 replicates), and generally produced an SE of higher than 90% for all three methods. However, this higher SE for the low-replicate case comes at the expense of increasing the FDR, which was

higher than 5% in all cases. The SE was close to 100% in high replication experiments for all methods, but the FDR for parametric methods was surprisingly high in many cases, especially in scenarios with a high technical noise level (Figures 5.27 to 5.30). Thus, the MCC results lead us to conclude that for 2 replicates, the performance of NOISeq falls between edgeR and DESeq2 and, for 3 replicates, there was no difference in MCC between NOISeqBIO and edgeR, which both outperformed DESeq2. Finally, for highly-replicated experiments, NOISeqBIO performed significantly better than the other two methods in terms of MCC.

When evaluating the results from the DE methods in more depth in different biological scenarios of high variability (Figures 5.27 to 5.30), we observed that the performance of the methods is very similar for different organisms (with a different number of genes) and for different proportions of DEGs. Again, all methods except SAMseq presented good results in terms of SE, FDR and MCC when applied to non-noisy data (Figures 5.27 and 5.29). However, when considering the more realistic scenario of a noise level of 0.2 (Figures 5.28 and 5.30), bigger differences were found between the methods and it is in this scenario that NOISeqBIO best controlled FDR compared to the other methods.

When analyzing the scenarios of low biological variability (Figures 5.31 and 5.32) we observed that NOISeqBIO performance on data with 5 and 10 replicates was again better than for the other methods, especially in terms of FDR. For the 3 replicates case, NOISeqBIO FDR increases with regard to the high variability scenario but is still lower than for the other methods. However, when only 2 replicates are available, NOISeqBIO FDR is higher than for edgeR but again lower than for DESeq2. SAMseq fails to detect any DEG in data with two replicates but performs better with three replicates when compared to high variability scenarios. If MCC results are analyzed, NOISeqBIO outperforms the rest of the methods in all cases except the 2 replicate case, where edgeR is the best option.

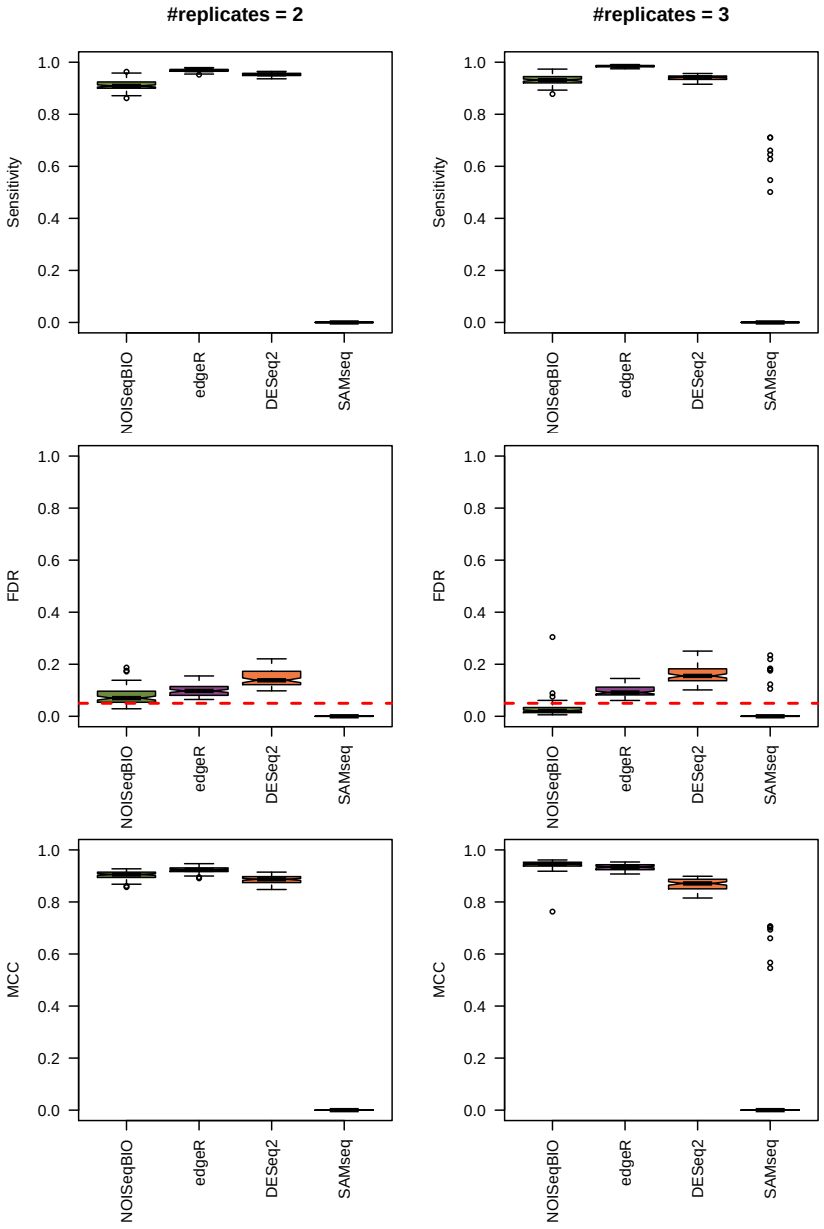


Figure 5.25: HIGH biological variability scenario. SE, FDR, and MCC of differential expression methods for data with a low number of replicates using an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all simulation parameter values were aggregated.

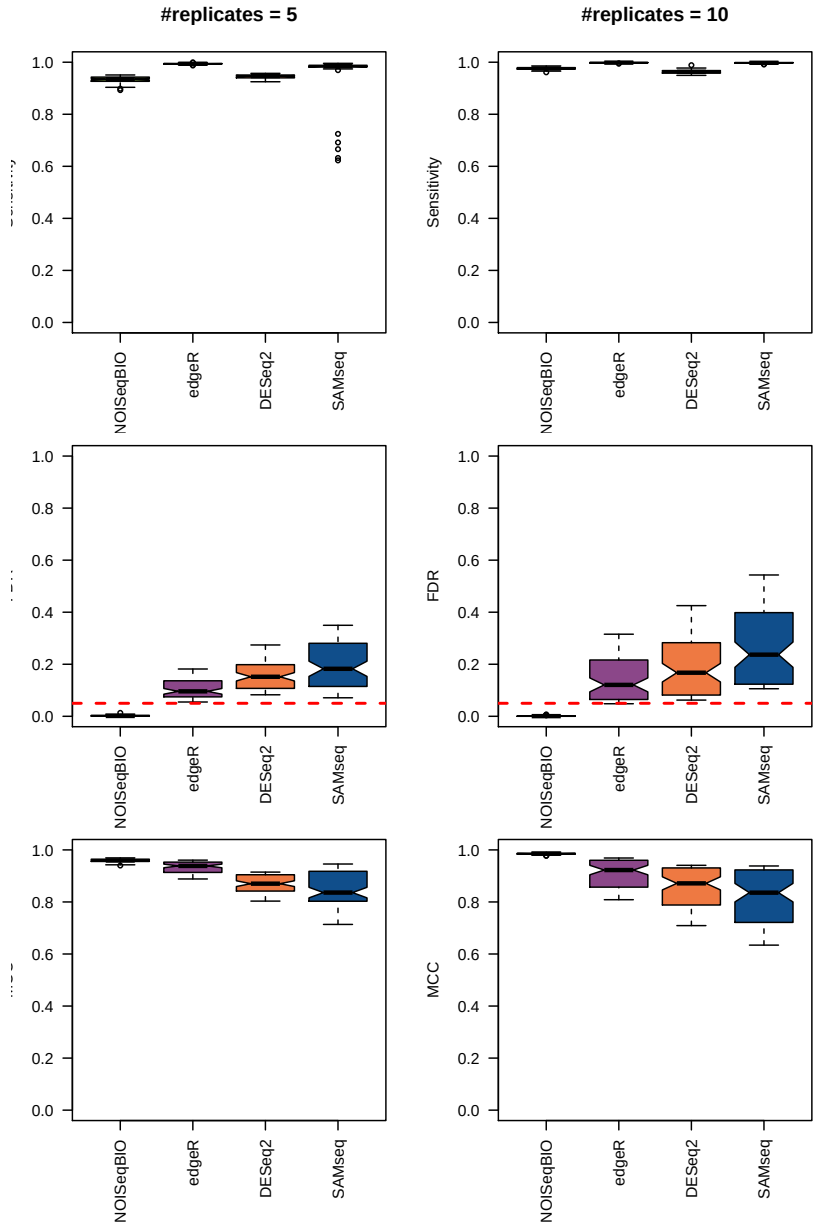


Figure 5.26: HIGH biological variability scenario. SE, FDR, and MCC of differential expression methods for data with a high number of replicates using an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all simulation parameter values were aggregated.

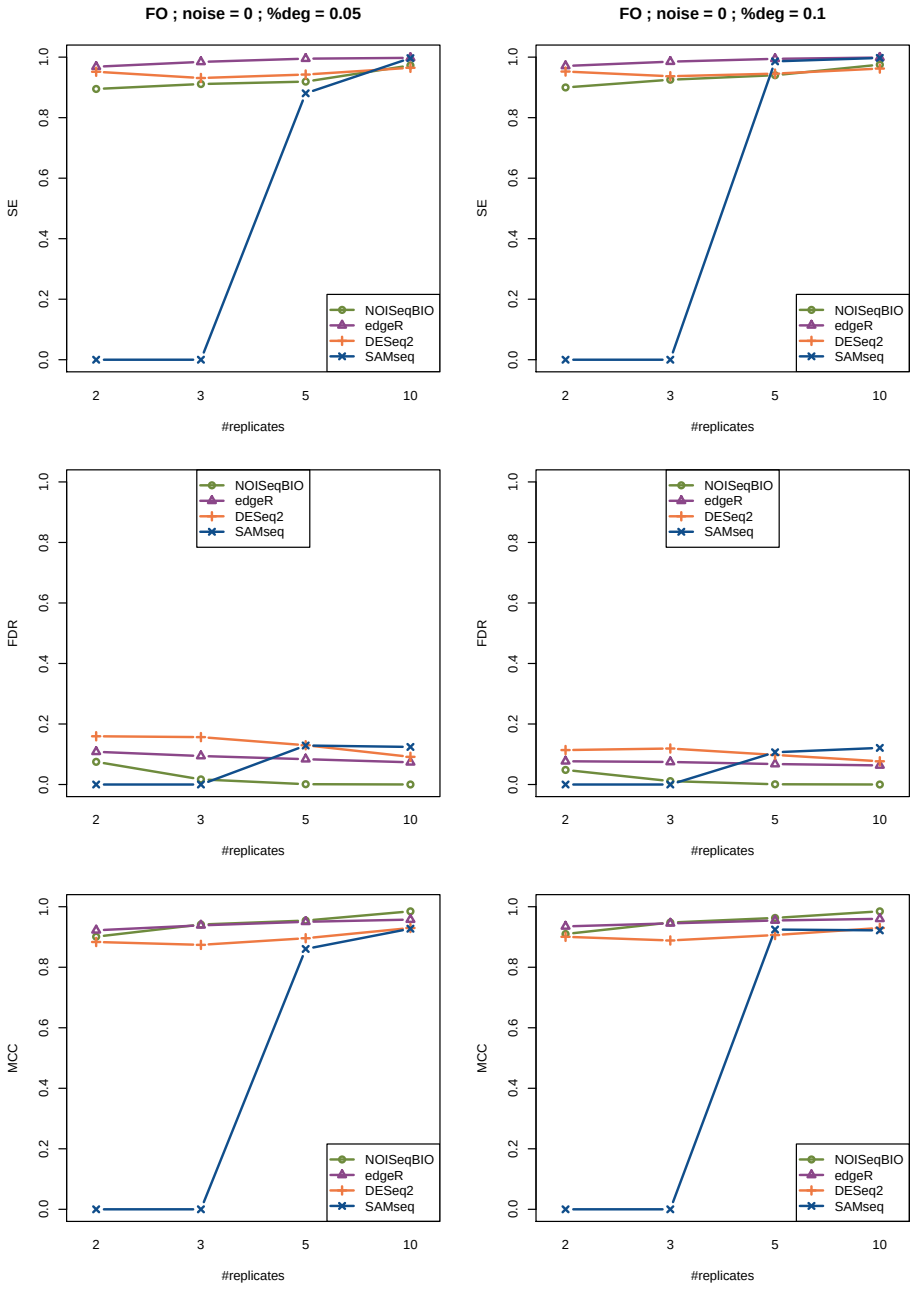


Figure 5.27: Performance of differential expression methods on data simulated from *F. oxysporum* data with $noise = 0$, and a FDR cutoff of 0.05 for all methods.

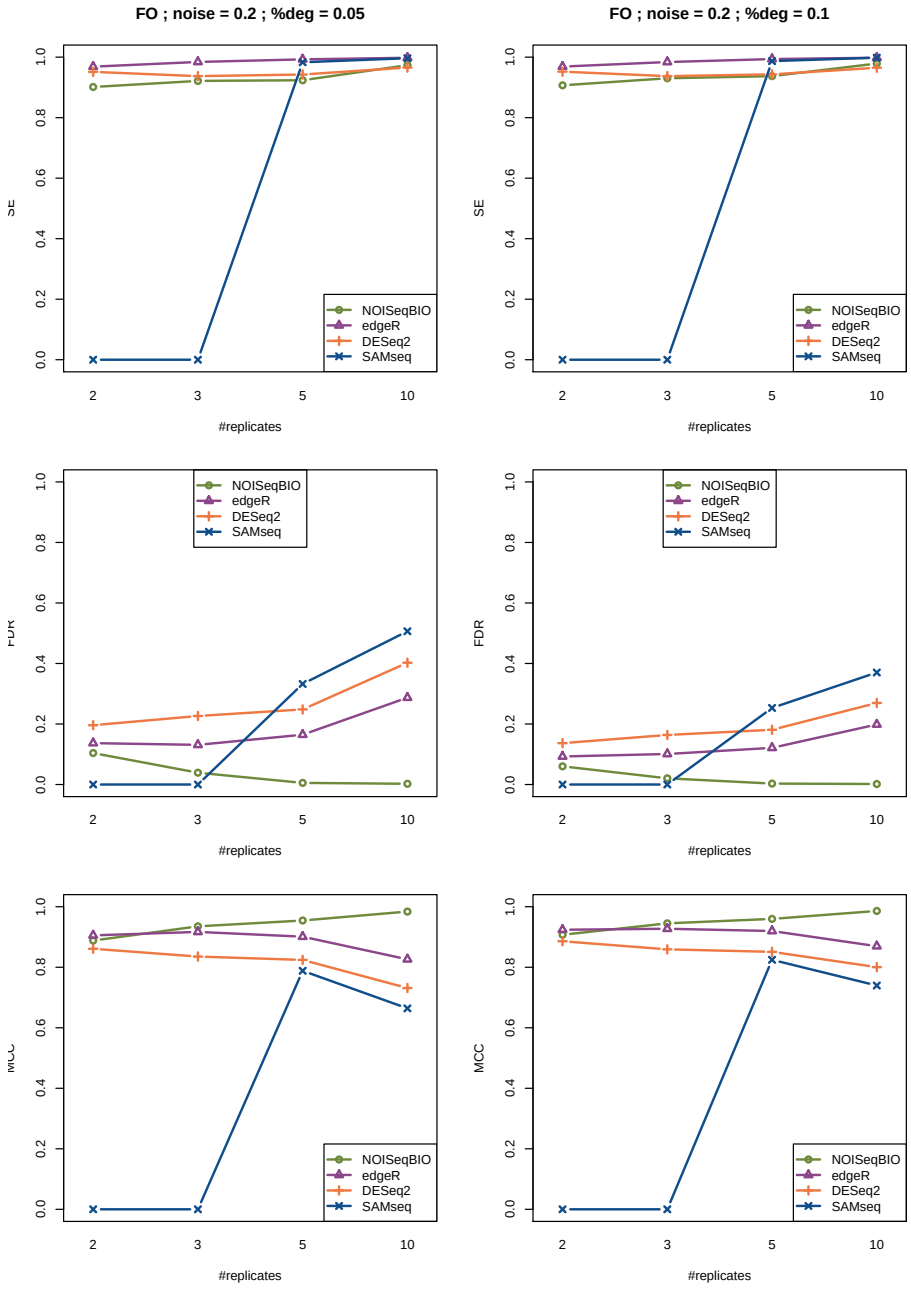


Figure 5.28: Performance of differential expression methods on data simulated from *F. oxysporum* data with $noise = 0.2$, and a FDR cutoff of 0.05 for all methods.

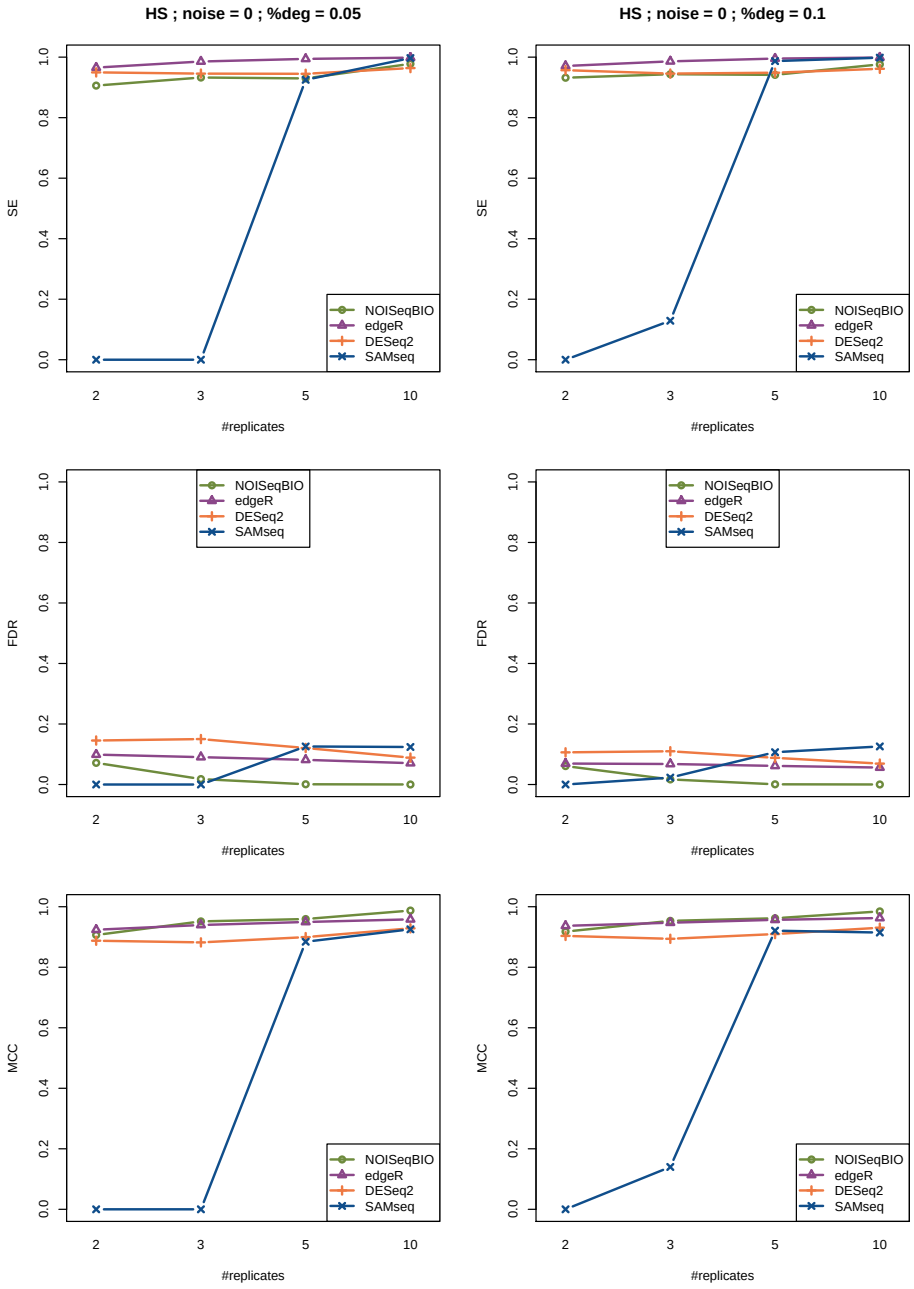


Figure 5.29: Performance of differential expression methods on data simulated from prostate cancer data with $noise = 0$, and a FDR cutoff of 0.05 for all methods.

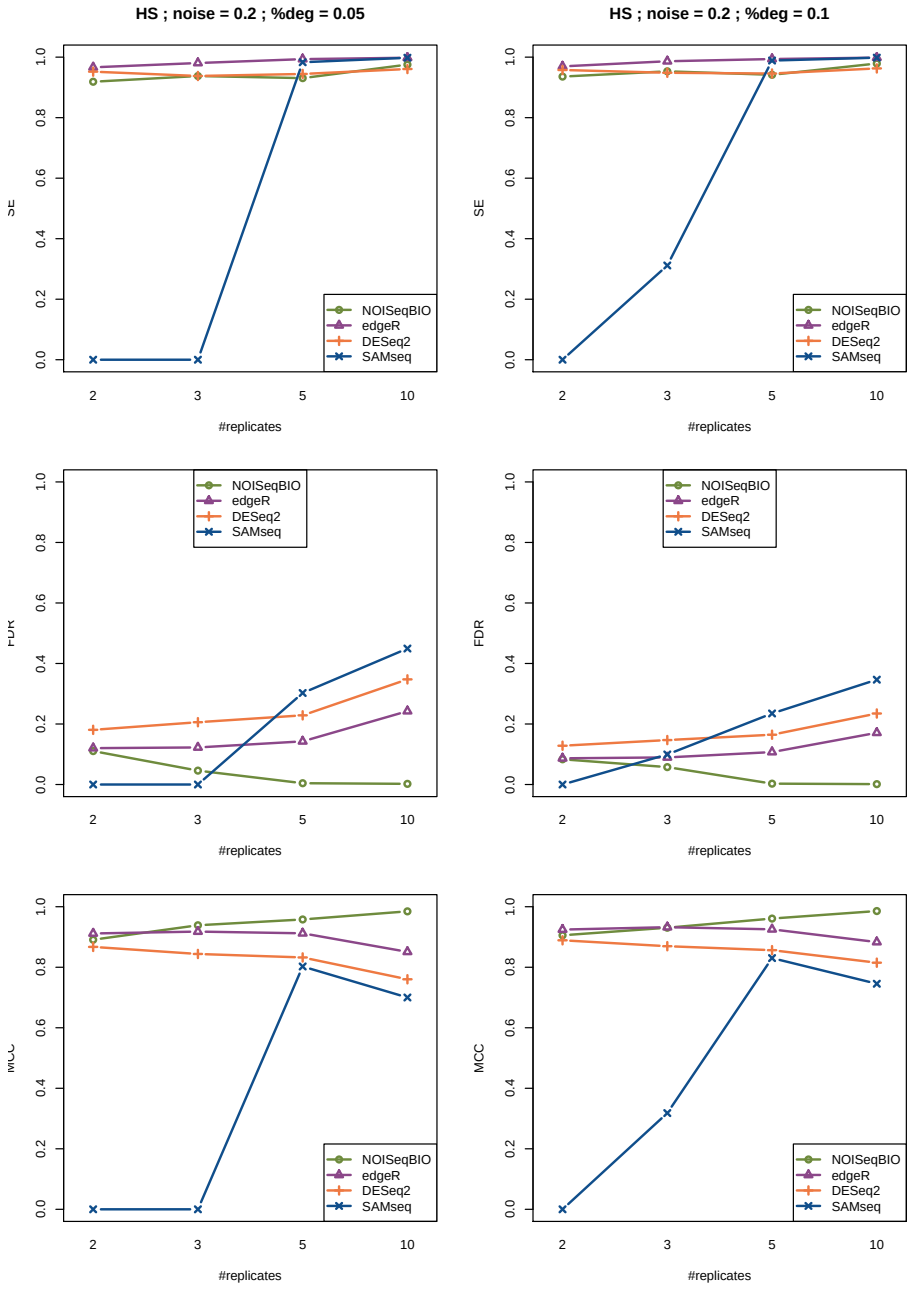


Figure 5.30: Performance of differential expression methods on data simulated from prostate cancer data with $noise = 0.2$, and a FDR cutoff of 0.05 for all methods.

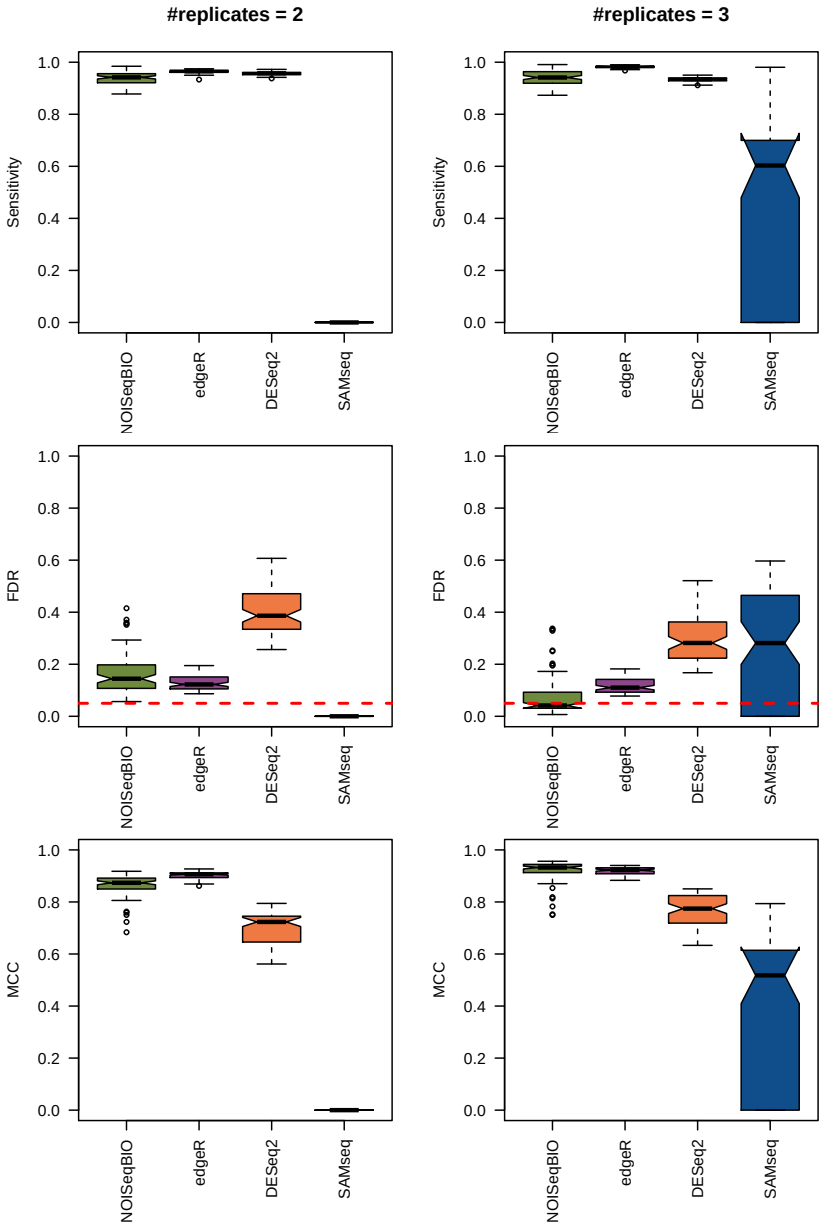


Figure 5.31: LOW biological variability scenario. SE, FDR, and MCC of differential expression methods for data with a low number of replicates using an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all simulation parameter values were aggregated.

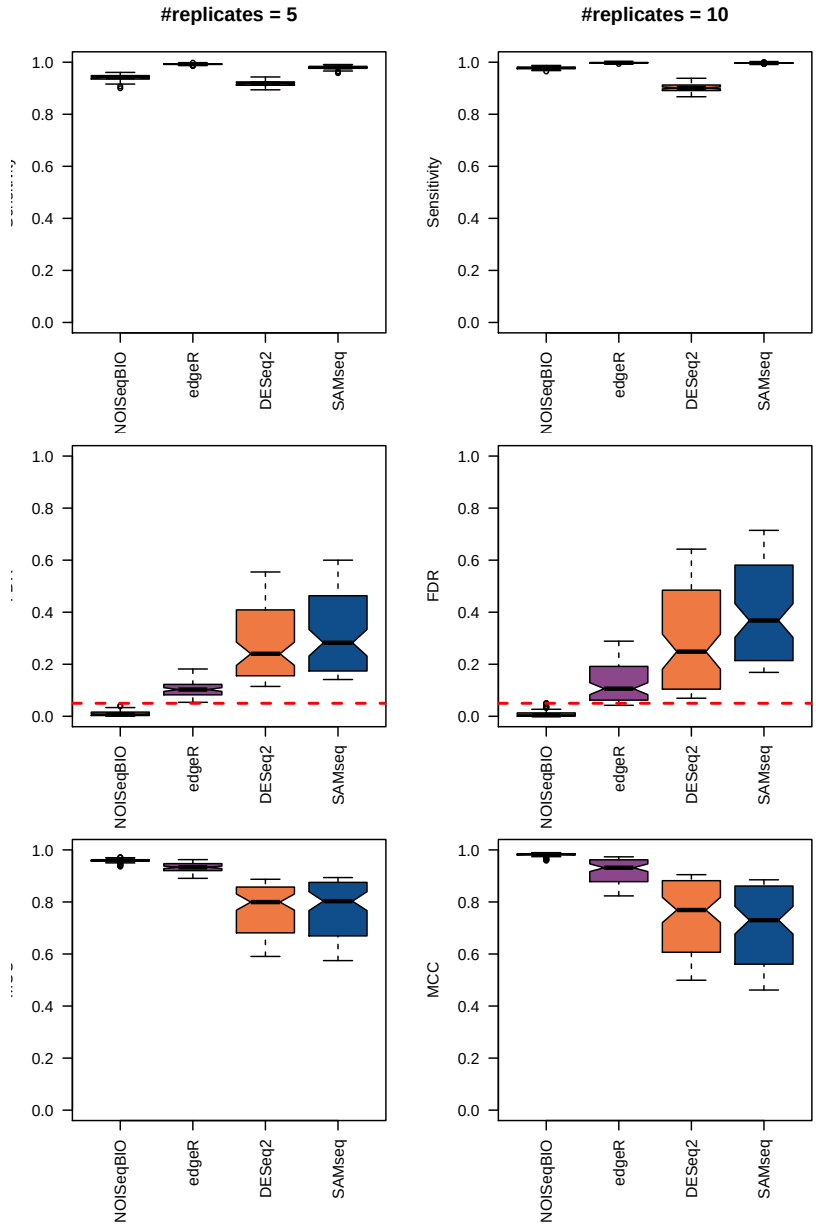


Figure 5.32: LOW biological variability scenario. SE, FDR, and MCC of differential expression methods for data with a high number of replicates using an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all simulation parameter values were aggregated.

5.4.2.3 Results on experimental datasets

Finally, we computed differential expression on **FO** and **HS** experimental datasets taking a cutoff of 0.05 for adjusted p-values, or equivalently, a probability cutoff of 0.95 for NOISeqBIO. The results are summarized in Figure 5.33.

		# genes		CV (%)	% DE genes (over total)			
Data	# replicates	total	after filter	median	NOISeqBIO	edgeR	DESeq2	SAMseq
FO	2	18066	10125	21.2%	31.5%	26.5%	24.5%	39.0%
HS	12 – 11	59573	17207	39.5%	2.9%	6.7%	7.4%	9.6%

Figure 5.33: Characteristics of **FO** and **HS** data sets regarding number of replicates, number of genes, coefficient of variation and %DEG.

Results for the **FO** dataset (Figure 5.34), which has 2 replicates per condition, indicated that the number of DEGs was very high for all methods (from 25% DEG in DESeq2 to 39% in SAMseq), especially for non-parametric approaches (NOISeqBIO and SAMseq). According to the coefficient of variation, this data set could be considered to belong to a high variability scenario so it seems that it is confirmed that NOISeqBIO may be returning a high number of false positives. However, NOISeqBIO results were more similar to edgeR or DESeq2 results than SAMseq results. When considering an experiment with more replication such as **HS** data with 11 and 12 replicates per condition (Figure 5.35), the results change considerably and the proportion of DEGs varies from 3% (NOISeqBIO) to nearly 10% (SAMseq).

Spearman's correlation coefficient between FDR values obtained from all the methods (Figures 5.36 and 5.37) was generally higher than 0.95, showing a good agreement on the gene ranking between methods. The only exception was SAMseq that presented a correlation with the rest of the methods of around 0.6 in **FO** data, and from 0.85 to 0.91 in **HS** data, again highlighting the effect of the number of replicates on its performance. Figures 5.36 and 5.37 also show the number of DEGs in common between each pair of methods.

GOseq [152] functional enrichment analysis was performed on each set of DEGs from the prostate cancer data to try to determine which DE method

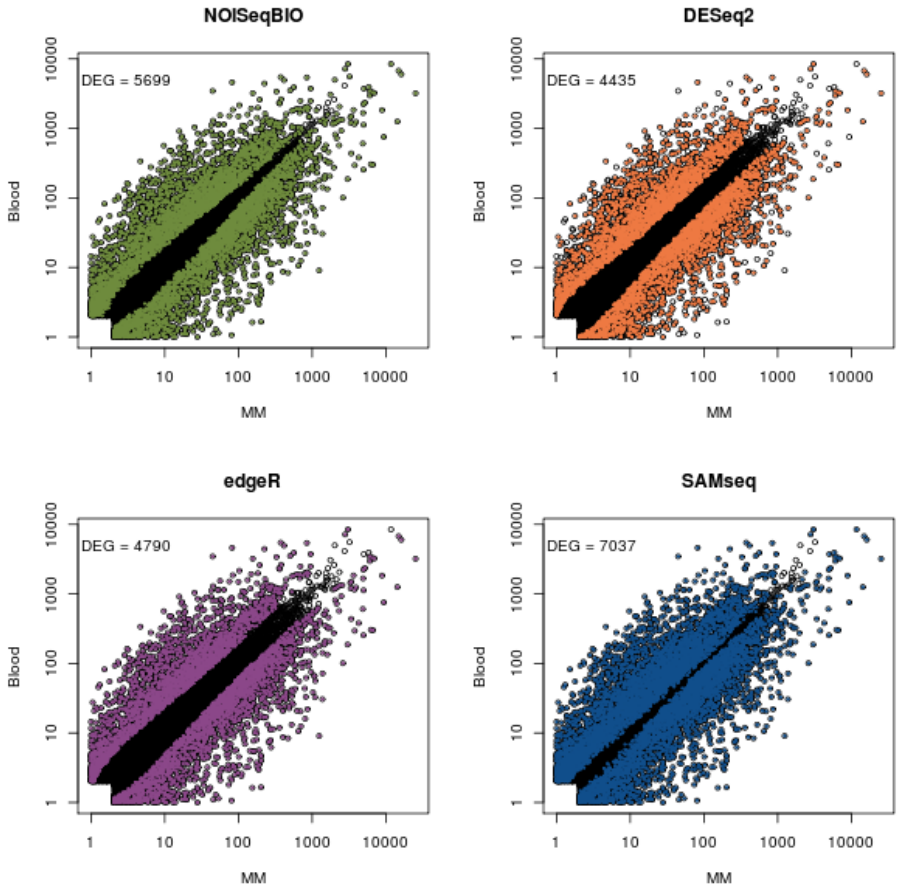


Figure 5.34: Differential expression results from methods compared applied to *F. oxysporum* data. The differentially expressed genes declared by each method are displayed in color.

resulted in a more biologically meaningful set of DEG genes, i.e. to unveil if the DEGs that were not in common contributed to generating better functional characterization of the results from any of the methods. However, this analysis revealed no major functional differences between the DE results for the methods compared (results not shown). In all cases, the enriched Gene Ontology terms included functions related to cancer or to prostate.

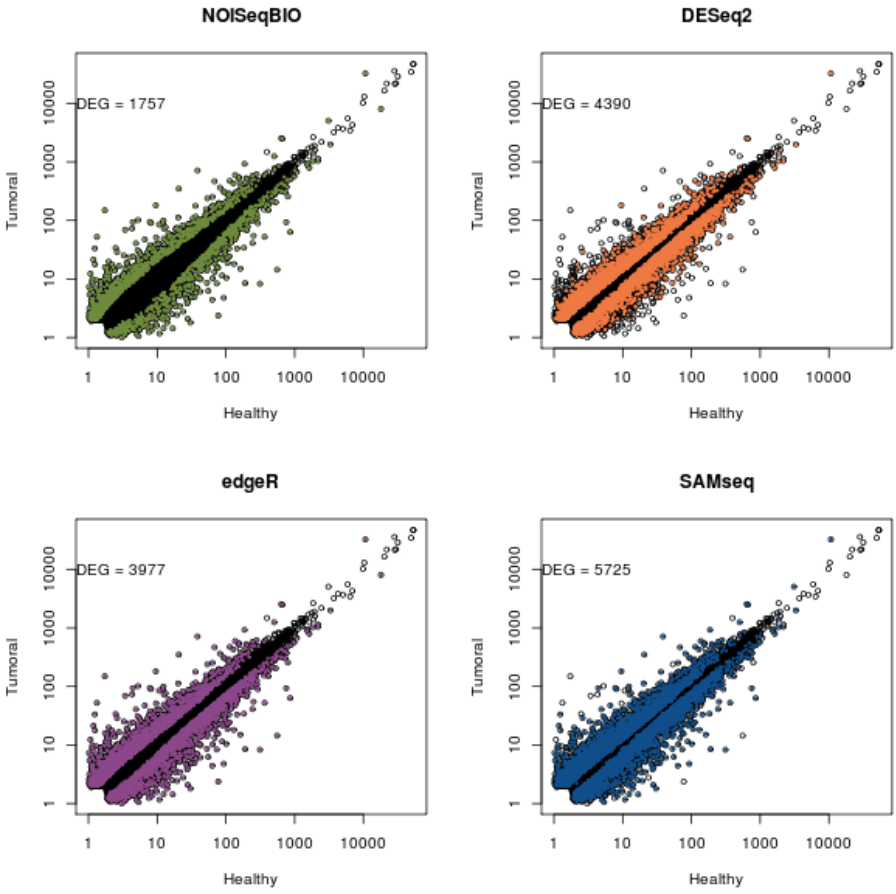


Figure 5.35: Differential expression results from methods compared applied to human prostate data. The differentially expressed genes declared by each method are displayed in color.

	NOISeqBIO	edgeR	DESeq2	SAMseq
NOISeqBIO	5699	4770	4385	5554
edgeR	0.955	4790	4241	4743
DESeq2	0.951	0.992	4435	4428
SAMseq	0.606	0.572	0.579	7037

Common DEG

Spearman's rank correlation

Figure 5.36: Differential expression results from **FO** data. The diagonal contains the number of DEGs for each method. Above the diagonal the number of DEGs in common for each pair of methods is shown. Below the diagonal Spearman's rank correlation coefficient between FDR or 1-probability for each pair of methods is shown.

	NOISeqBIO	edgeR	DESeq2	SAMseq
NOISeqBIO	1757	1733	1434	1642
edgeR	0.979	3977	3554	3669
DESeq2	0.960	0.992	4390	4138
SAMseq	0.848	0.889	0.908	5725

Common DEG

Spearman's rank correlation

Figure 5.37: Differential expression results from **HS** data. The diagonal contains the number of DEGs for each method. Above the diagonal the number of DEGs in common for each pair of methods is shown. Below the diagonal Spearman's rank correlation coefficient between FDR or 1-probability for each pair of methods is shown.

5.5 Discussion

Differential expression analysis is widely used in transcriptomics to identify changes in gene expression. The genes that are differentially expressed under a given experimental condition may be responsible for the change in the phenotype and therefore are candidates for experimental validation or as the basis for proposing new hypotheses.

With the arrival of RNA-seq technology, researchers in bioinformatics realized that the statistical methods previously used for microarrays were no longer valid because of the different nature of the measurements: discrete values (read counts), in contrast to continuous values from microarrays (intensities). Many DE methods have been developed since then, and most of them are based on parametric assumptions [107]. Parametric methods are said to have more power and robustness than non-parametric methods, however, there are some drawbacks that must be taken into account. First, researchers should check that the parametric assumptions are fulfilled, although it is neither easy nor common to find this kind of verifications in bioinformatics because of the huge number of models derived. Thus, in some cases parametric methods could be unsuitable and using them could generate a false positive rate higher than expected [79]. Second, parametric methods need raw counts to be provided since they are based on discrete distributions such as the Poisson or Negative Binomial. However, sometimes the available data have already been normalized by the expression quantification software (e.g. FPKM values coming from Cufflinks [144]) and so raw read counts are not available. Other times the application of transformations to remove technical biases or batch effects may be necessary. In these circumstances, the distributional assumptions may not hold. Hence, the scientific community would benefit from having appropriate non-parametric alternatives for differential expression.

In this chapter, we presented two non-parametric differential expression approaches included in the NOISeq R Bioconductor package: NOISeq and

NOISeqBIO. NOISeq [141] was optimized for experiments with technical replicates or without replication at all: the characteristics of the datasets available at the dawn of the technology. NOISeqBIO is a new non-parametric empirical Bayesian method for use with biological replicates, which are now more common in RNA-seq experiments because of the decreasing cost of the technology, and it has been developed by joining the ideas from our previous work and from that of Efron *et al.* [43].

NOISeq was compared to several RNA-seq differential expression methods: edgeR, DESeq, baySeq, DEGseq and the traditional FET on data with technical replications. All but FET are parametric approaches. NOISeq creates an empirical distribution of count changes (which is adapted to the available data) from which the probability of differential expression for each feature can be derived. In this non-parametric approach, differential expression does not rely on individual transcript measurements, but rather on the joint distribution of (M, D) values for all the features within the dataset.

Comparison of these methods on synthetic data showed that DEGseq-MARS, DEGseq-LRT, FET and baySeq-Poisson performed worse than the other methods. Therefore, we discarded all these methods except FET (because of its extended use) for further study on experimental data. Although it has been claimed that the Poisson distribution is adequate for technical replication [18, 92, 107], these results seem to highlight that the Negative Binomial distribution may also be more suitable with technical replicates.

The experimental data we chose for the evaluation of the selected methods on data with technical replicates (MAQC and Griffith's) also included RT-PCR measurements that could be used as a gold-standard to assess if the DEGs identified by each method were correct. On MAQC data, both the Precision-Recall curves and especially the FDR curves showed that NOISeq performed much better than the parametric methods. FET presented a good FDR but a poorer PRC. Griffith's data, although more limited, confirmed these findings.

We studied the effect of the number of technical replicates (lanes) on the number of differentially expressed genes, their length, fold-change value,

and expression level. The pattern produced by NOISeq and FET was more constant across the different variables analyzed, whereas the other three methods showed a pronounced dependence. The parametric approaches strongly increased the number of significant calls as more sequencing output was included, resulting in a considerable number of false positives (Fig. 5.10). The newly detected genes were shorter, had a lower relative expression, and had smaller fold-change differences than those obtained with less data. False positive genes identified in the analysis of the MAQC data had similar characteristics, suggesting that large library-size datasets analyzed by these parametric approaches incorporate many falsely called significant genes and/or with small fold-change differences at the low expression range. The constant pattern of FET was intrinsically due to a low detection power that identified only highly expressed transcripts. However, NOISeq showed more robustness against these sequencing depth biases while maintaining a high true positive detection rate. We believe that, given the number of lanes sequenced and the specific characteristics of the data analyzed, this approach creates a more realistic estimation of the probability that a given count difference will occur by chance, and also results in the stable control of false positives. The parametric approaches compared do not have this flexibility and tend to render significant small fold-changes as sequencing numbers grow. With regard to the two variants of NOISeq, overall NOISeq-sim and NOISeq-real performed similarly throughout the whole study, although a slightly higher detection rate and dependency on the sequencing depth was observed with NOISeq-sim, and these differences were more pronounced with Griffith's data. These results indicate that the simulation procedure of NOISeq-sim works well to replace technical replicates but may tend to overestimate DEG in data with high variability among replicates.

Regarding differential expression in data with biological replicates, we carried out several studies to assess the performance of the new non-parametric method NOISeqBIO, by comparing it to some of the most widely used differential expression methods such as edgeR [123] and DESeq2 [4] on both simulated

and experimental data. Since NOISeqBIO is a non-parametric method, we also included the non-parametric method SAMseq [82] in these comparisons.

Preliminary studies on simulated data helped us to determine the best variant of NOISeqBIO; this was computation of the differential expression statistic Z as the mean of M and D, taking the 90th percentile of gene standard deviations for the constant a_0 for the variability correction, and using the KDE to estimate densities f and f_0. Since NOISeqBIO clusters genes with similar expression levels when the number of replicates is lower than 5, we also evaluated the effect of the number of clusters in these preliminary studies, finally deciding to leave it to the user's choice, although we chose 15 clusters as the default option. The preliminary studies also served to analyze the influence of the parameters which define the biological scenarios to simulate: the number of replicates per condition, the noise level, and the proportion of DEGs had the strongest effects on the performance of the methods, so a set of scenarios for the final comparison were defined according to this information.

The final simulation study showed the superiority of NOISeqBIO in controlling the FDR while maintaining a sensitivity rate above 90% in most cases, except for the 2-replicate case in low biological variability scenarios. We observed the dependence of the comparison results on the number of available biological replicates and, interestingly, we observed that parametric methods tend to present a higher FDR as the number of replicates increases, especially for noisy data. SAMseq failed to detect DEGs for low replication number data and presented a high FDR for high replication number cases.

Results from experimental datasets indicated that for data with few replicates (**FO**) all the methods tend to declare a high number of genes as differentially expressed, including SAMseq. In contrast, the proportion of detected DEGs was lower for data with many replicates (**HS**), which shows the need to increase the number of replicates to efficiently control the FDR. Correlation of p-values was higher than 0.95 for all the methods except SAMseq. The functional enrichment analysis rendered biologically meaningful results for all methods.

As RNA-seq technology becomes more affordable, experiments with a higher number of replicates are expected and so DE methods which efficiently deal with FDR whilst maintaining high sensitivity rates, will be needed. Hence, NOISeqBIO perfectly fulfills these requirements. Although it has been reported that non-parametric methods tend to require a higher number of replicates to perform well, as we observed for SAMseq, NOISeqBIO performed generally well with a small number of replicates while retaining the advantage of not being based on distributional assumptions.

Therefore, in conclusion, we have proven here that we have successfully designed two non-parametric differential expression methods (NOISeq and NOISeqBIO) for pairwise comparisons that are a good alternative to popular parametric approaches.

Chapter 6

General discussion and Conclusions

6.1 General discussion

This thesis mainly focuses on the analysis of gene expression data. Specifically, the aim was to propose methodologies for variable selection i.e. to identify genes whose expression significantly changes among different experimental conditions.

Variable selection in transcriptomics is complicated for several reasons. On the one hand, the high dimensionality of the data (many variables versus few observations) drastically diminishes the power of statistical methods because of the small sample size and the multiple testing correction when using univariate approaches. On the other hand, data are noisy. It is not always possible to identify the noise source and not all variables are equally noisy (for instance, low-expression genes tend to be more noisy). The variable selection problem is equivalent in this context to the identification of differentially expressed genes, or genes whose change is biologically relevant and the two considerations are not always equivalent. For example, small expression changes may be statistically significant but have little biological impact. However, the effectiveness of methods is frequently evaluated in terms of biological meaning. Univariate statistics can capture the specifics of the behavior of each gene, although they are strongly penalized when applying multiple testing corrections. On the contrary, multivariate or Bayesian approaches may be more robust against these problems but fail to identify changing genes with particular behaviors. The technologies for measuring gene expression are constantly evolving and statistical methods have to evolve to fit these new data structures. Moreover, these new transcriptomics technologies give rise to novel analysis scenarios or novel types of biases so methods need to be revisited and adapted to broader sets of scenarios. The fact that transcriptome analysis is mostly performed by biologists and scientists with a limited statistical background requires that any methods proposed should be easy to adopt and apply by these users, implying that not only methods but also accessible tools should be developed.

In this thesis we have addressed data analysis problems that were relevant to the state of the art technologies in transcriptomics statistical analysis at

the time. We developed variable selection methods for complex microarray experimental designs where solutions were still limited. For this purpose, we opted for a solution that uses multivariate dimension reduction statistics with the assumption that this would be an efficient approach to treating data that contains both many variables and multiple conditions (Chapter 3). Subsequently, and as the RNA-seq technology became a reality, we investigated suitable differential expression methods for analyzing count data (Chapter 5). Although these two approaches may appear to have distinct target technologies, the methods or principles they use could actually be extended to each other or to other analysis scenarios. For example, the minAS method (Chapter 3), which was initially developed for using the SPE and leverage statistics of an ASCA analysis on multifactorial data, was also a very useful strategy for variable selection coupled to other multivariate methods. In the case of N-way statistics, minAS effectively selected both genes and metabolites from multivariate datasets that were modeled as three-dimensional data structures. In the pathway network approach, where PCA is used to reduce pathway-level gene expression data matrices to indexed pathway activities, minAS was able to identify the most relevant (or "driving") genes associated with each pathway. Arguably, these methods based on dimension reduction could also be applied to highly transformed RNA-seq data. In fact, when RNA-seq data undergo extensive normalization and data transformation pre-processing steps (such as RPKM and TMM normalization, GC bias correction, and eventually removal of batch effects), the resulting dataset may have lost the properties of count data and be amenable to treatment with methods normally used for continuous data. We have observed this situation in the analysis of RNA-seq data from the STATegra project (not shown in this work), where ASCA and linear models were applied after several rounds of data pre-processing (including log-transformation). On the other hand, modeling noise in gene expression data using the M and D statistics (Chapter 5) does not specifically require that values are discrete and hence could be applicable to other types of genome wide gene expression measurements.

One of the assumptions of both the minAS and the NOISeqBIO methods is that transcriptomic changes can be modeled as a mixture of two distributions: one corresponding to the genes that do not change their expression between conditions and another to the genes whose expression does change. This implies a kind of on/off situation for gene expression which is biologically justified because of how the transcriptional machinery functions. In practice, the detection of gene expression signals is buffered by post-transcriptional regulation, the broad-magnitude range of gene expression, and the noise produced by transcriptomics technologies, which may mask the bimodal distribution. However, we have found that in many cases bimodal modeling does return meaningful breakpoint thresholds (as in minAS) or results in gene selection with high accuracy (as in NOISeqBIO) which supports the general validity of the bimodal assumption.

Another relevant aspect of this thesis is the choice of non-parametric statistics to analyze gene expression data. Both the minAS and the NOISeq methods are data-driven and do not rely on theoretical distributions imposed on the data. While parametric methods can be powerful for modeling data when the number of observations is limited, they may introduce data analysis inaccuracies if the model assumptions are not fulfilled. This may frequently be the case in transcriptomics, as the technologies that generate the data do not always follow uniform or constant error rates. Conversely, non-parametric methods may fail to identify significant changes when data is insufficient. We have seen that the non-parametric approaches implemented in the NOISeq package work well in terms of sensitivity and, importantly, in controlling false discovery rates better than its parametric counterparts. Good control of the false positives is particularly relevant in transcriptome studies due to the large number of variables present in the datasets and the need to reliably identify biomarkers associated with the phenotype. Alternatively, parametric methods might be a better choice when small changes need to be detected or false calls are not as relevant, for example, when additional filtering or validation steps are applied to the data.

Finally, one of the most-important lessons learned while producing the body of work presented in this thesis was the importance of a good understanding of the nature of the data, and of the potential biases of the measuring technologies, to produce quality analysis. While data pre-processing tends to be considered a minor or prolegomenon statistical task, the reality is that pre-processing takes most of the data analysis time and that it greatly influences the inferential results. One important aspect here is to first identify any potential biases of the technology, second to know how to assess the magnitude of the biases in the data, and third to have tools that can, at least partially, correct them. In this thesis, we devoted a lot of effort to developing diagnostic plots and quality improvement procedures for RNA-seq data. These procedures are now implemented in two bioinformatics resources (Qualimap [49] and NOISeq Bioconductor R package) to make them widely accessible to the transcriptome research community. We believe that, as quality control protocols become more generalized and easy-to-use analysis tools become more available, the overall quality of transcriptome research will greatly improve.

6.2 Conclusions

The conclusions of this thesis are summarized and organized in the following text according to the goals defined in Chapter 2.

1) To develop variable selection strategies for multivariate methods applied to microarray data.

- We proposed new variable selection methods (*minAS* or *Gamma*) or variations of existing ones (*permutation approaches*) for multifactorial gene expression data and compared them to existing methods (*Box* or *Jackson & Mudholkar*).
- The methodologies presented are all based on studying the probability distribution of a statistic which measures the importance of the variables in the model so the selection is data-adaptive.
- We tested the performance of variable selection strategies on simulated multi-factorial expression data and checked that they generally work well in different scenarios which were defined by the dataset size, the diversity of gene expression signals and the levels of noise. The best strategies were *Jackson & Mudholkar*, *minAS* and *Gamma*.
- These three selection strategies were applied to experimental data to identify the genes responsible for human stem cell differentiation under different oxygen concentration conditions, and relevant biological conclusions were obtained in all cases.
- The major differences in gene selection and functional enrichment were because of the method chosen for the SPE statistic, while leverage seemed to be more robust for the statistical model applied.
- These strategies have been successfully applied in cases where other multivariate techniques and "importance" statistics were

used. They are also implemented in SEA, which is a web tool for analyzing time-series gene expression data.

2) To generate tools to control the quality of count data from sequencing experiments in order to discover potential biases and to propose procedures to mitigate their effect.

- A whole set of useful graphical and diagnostic tools was designed to assess the quality of the RNA-seq count data prior to statistical analysis, and the functionality of each plot was illustrated by using experimental data.
- We provided a statistical assessment of several typical RNA-seq biases and offered appropriate normalization tools to correct them.
- All these tools were included in an open Bioconductor R package called NOISeq, which also offers the possibility of generating a Quality Control Report PDF to facilitate data exploration to the users. Some of the plots are also available in the Qualimap web tool.

3) To develop differential expression methodologies for RNA-seq data.

- We presented the NOISeq differential expression method for application on RNA-seq data with technical replicates or without replicates. NOISeq was compared to several differential expression methods: edgeR, DESeq, baySeq, DEGseq and Fisher's exact test. It performed well on both simulated and experimental data and was more robust against the number of technical replicates.
- The NOISeqBIO method was adapted from NOISeq for data with biological replicates following an empirical Bayes approach. NOISeqBIO was successfully compared to edgeR, DESeq2 and SAMseq when applied on both simulated and experimental data.

- The non-parametric NOISeq and NOISeqBIO approaches were able to control the False Discovery Rate well when compared to parametric or other non-parametric methods.
- Both methods have been implemented in the NOISeq Bioconductor R package.

6.3 Reach and relevance

The relevance of this thesis is justified in the following points:

- This thesis was developed within the framework of three international research projects (TRANSPAT, Genomics and transcriptomics of detoxification pathways in Drosophila and STATegra) and therefore the methods developed were used to analyze the data generated by these projects, which has contributed to the dissemination of the results of the thesis.
- The methodologies described here have been implemented as software tools which are freely available to the scientific community: the SEA web tool, the Qualimap suite, and the Bioconductor R package, in order to facilitate their use. The NOISeq method has also been implemented as an analysis option by third party software such as RNASeqGUI [126].
- This work was developed not only to generate efficient statistical methods but also with the final users of these tools (mainly biologists) in mind. This fact is reflected in the type of journals where we have published our research. We intend our tools to be used by as many people as possible and considering that the number of users is proportional to the number of citations, we are satisfied with the impact of our work (for instance our NOISeq paper [141] had 194 citations in Google Scholar by October 4, 2014).

- The usefulness of these methodologies has been demonstrated to the end users at more than ten international courses.

6.4 Future research lines

Our current and future lines of research are defined by the STATegra European project which we are currently working on. The goal of this project is to develop appropriate and accurate statistical procedures to integrate multiple omics data (which have been measured on the same biological system) in order to gain knowledge on this system. Specifically, we are tackling the analysis of several sequencing technologies such as RNA-seq, miRNA-seq, DNase-seq, ChIP-seq, and Methyl-seq as well as both proteomic and metabolomic data. Efficiently integrating this variety of omics data is still a challenge and the solutions may lie in finding the answers to each one of the following critical points:

- Experimental design: We will study how to determine the sequencing depth and/or the optimal number of biological replicates for each data type, taking into account the number of biological features (genes, genomic regions, etc.) and the intrinsic technical noise of each data type.

- Data pre-processing: It is essential to detect and correct unwanted effects in the data such as technical biases, contamination, batch effects, etc. in order to obtain meaningful results from posterior analysis and to increase the power of statistical methods by reducing noise. Therefore, we will set up efficient pre-processing procedures. We will also identify the best approach to impute missing values.

- Variable selection: Due to the huge number of biological features that are measured in most of the data types mentioned above, we believe that integration strategies will benefit from previously selecting the most relevant variables for each data type. In this sense, we are adapting some

approaches such as pair-wise linear models (between two data types) or machine learning methods (e.g. decision trees) and using existing methods such as NOISeq or maSigPro.

- Integration approaches: We will apply structural equation models to infer the connections among the different omics features from all data types and available biological information.

- Results validation: We aim to experimentally validate some of the results generated from the statistical analysis.

- Implementation of visualization methods and user-friendly tools: Software will be developed that allows statistical results to be visualized so that they can be easily understood from a biological point of view. Statistical methods will also be implemented in a Bioconductor R package and in user-friendly commercial software.

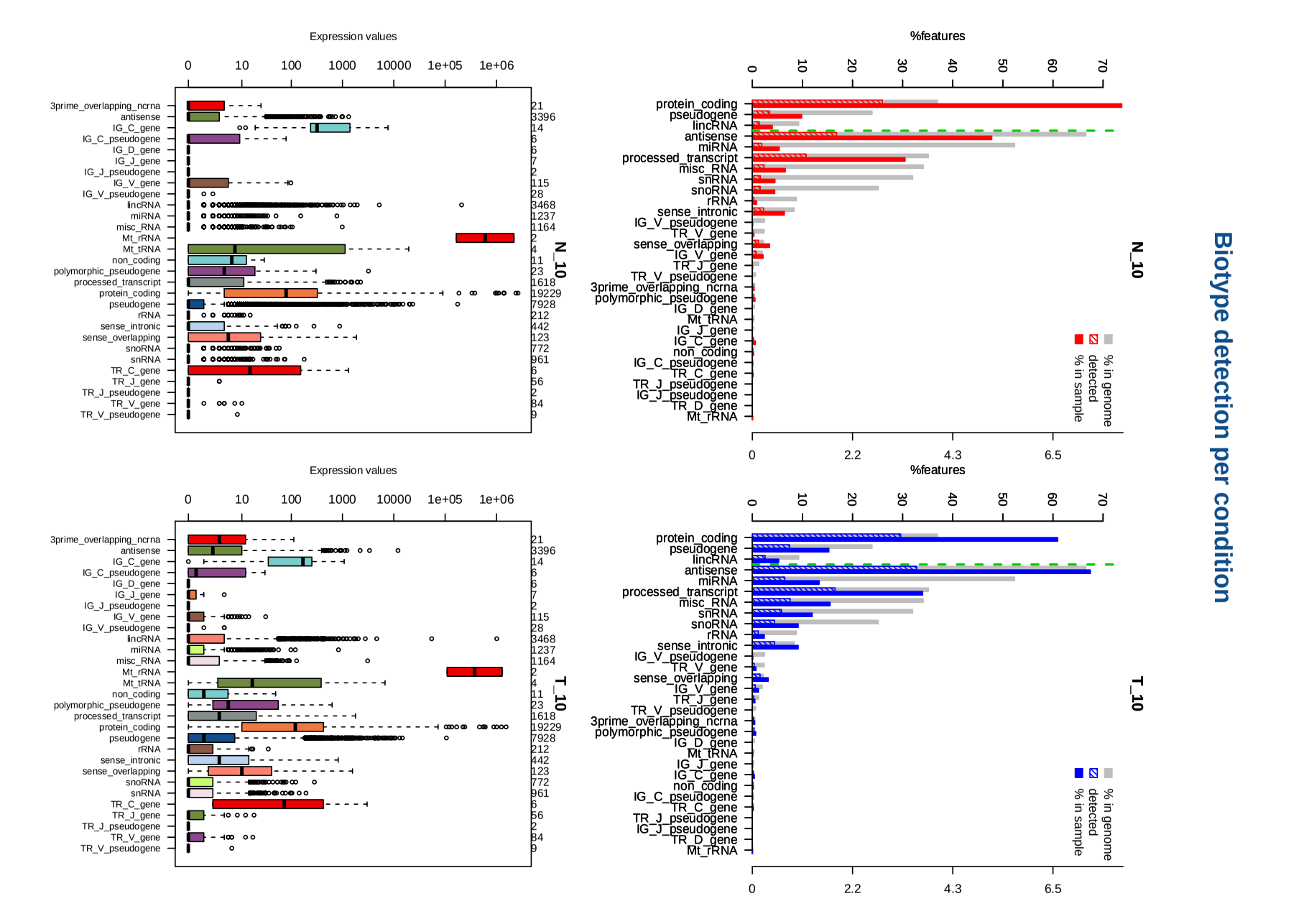
Appendix:
Quality Control Report

Quality Control of Expression Data

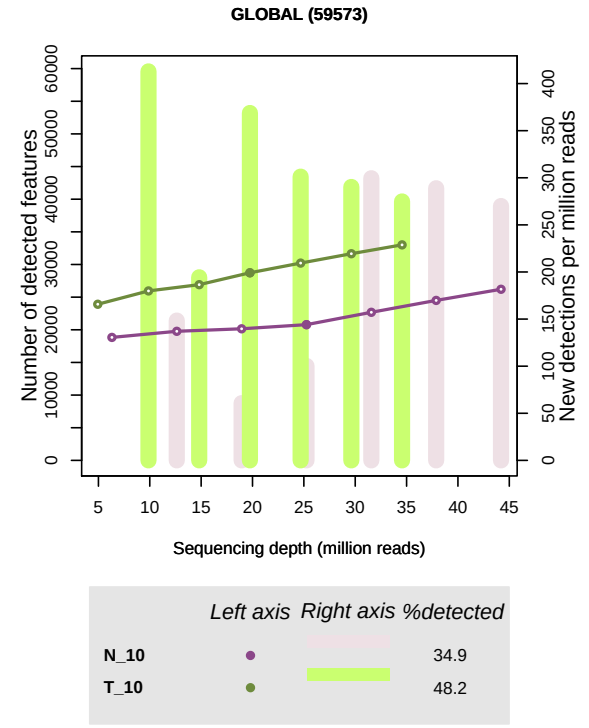
Generated by NOISeq on 21 Jan 2014, 18:43:47

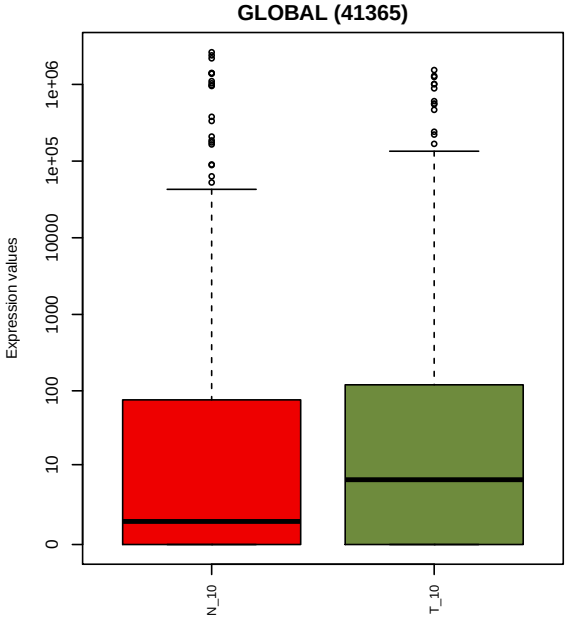
Content

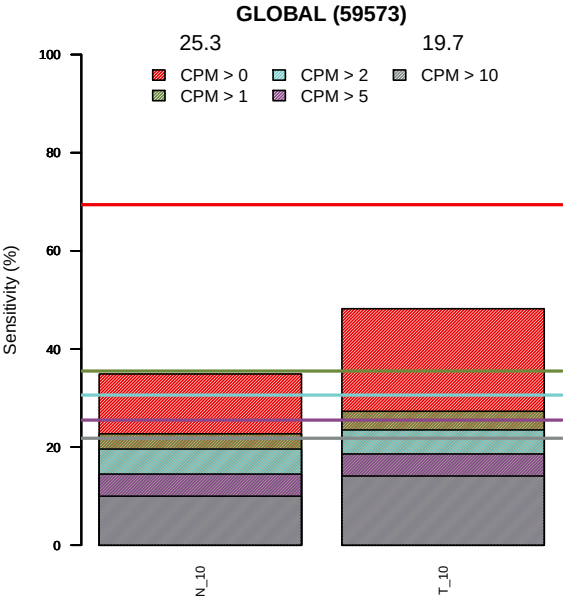
Plot	*Description*
Biotype detection	Number of genes per biotype in the genome, and detected (counts > 0) in the sample/condition.
Biotype expression	Distribution of gene counts per million per biotype in sample/condition (only genes with counts > 0).
Saturation	Number of detected genes (counts > 0) per sample across different sequencing depths
Expression boxplot	Distribution of gene counts per million (all biotypes) in each sample/condition
Expression barplot	Percentage of genes with >0, >1, >2, >5 or >10 counts per million in each sample/condition.
Length bias	Mean gene expression per each length bin. Fitted curve and diagnostic test.
GC content bias	Mean gene expression per each GC content bin. Fitted curve and diagnostic test.
RNA composition bias	Density plots of log fold changes (M) between pairs of samples. Confidence intervals for the median of M values.


Biotype detection per condition
Expression values
0
10
100
1000
10000
1e+05
1e+06
3prime_overlapping_ncrna
antisense
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_J_gene
IG_J_pseudogene
IG_V_gene
IG_V_pseudogene
lincRNA
miRNA
misc_RNA
Mt_rRNA
Mt_tRNA
non_coding
polymorphic_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
TR_C_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
21
3396
14
6
6
7
2
115
28
3468
1237
1164
2
4
11
23
1618
19229
7928
212
442
123
772
961
6
56
2
84
9
N_10
T_10
%features
0
10
20
30
40
50
60
70
protein_coding
pseudogene
lincRNA
antisense
miRNA
processed_transcript
misc_RNA
snRNA
snoRNA
rRNA
sense_intronic
IG_V_pseudogene
TR_V_gene
sense_overlapping
IG_V_gene
TR_J_gene
TR_V_pseudogene
3prime_overlapping_ncrna
polymorphic_pseudogene
IG_D_gene
Mt_tRNA
IG_J_gene
IG_C_gene
non_coding
IG_C_pseudogene
TR_C_gene
TR_J_pseudogene
IG_J_pseudogene
TR_D_gene
Mt_rRNA
0
2.2
4.3
6.5
% in genome
detected
% in sample

Sequencing depth & Expression quantification





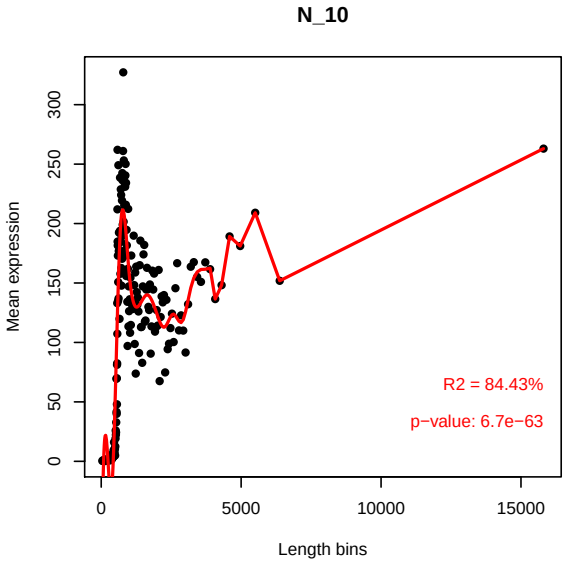


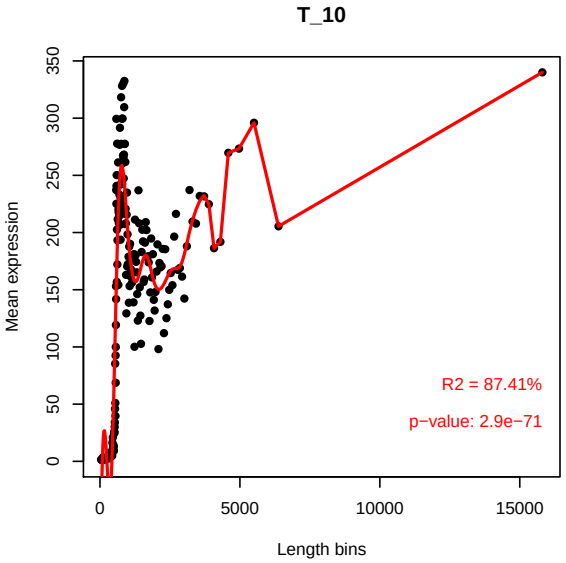
Sequencing bias detection

Diagnostic plot for feature length bias

FAILED. At least one of the model p−values was lower than 0.05 and R2 > 70%.

Normalization for correcting length bias is recommended.

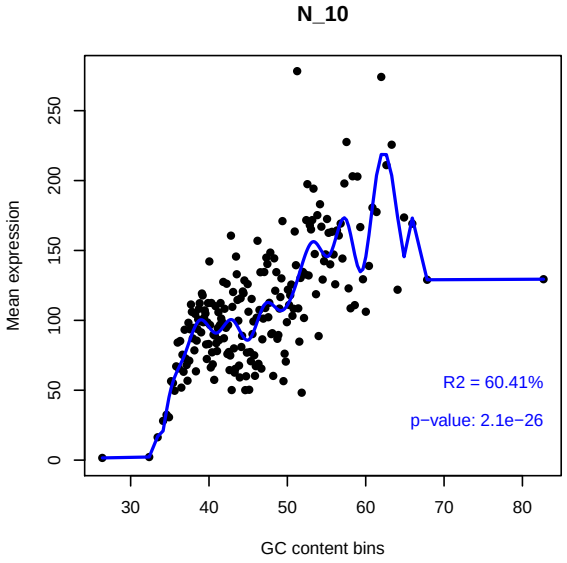


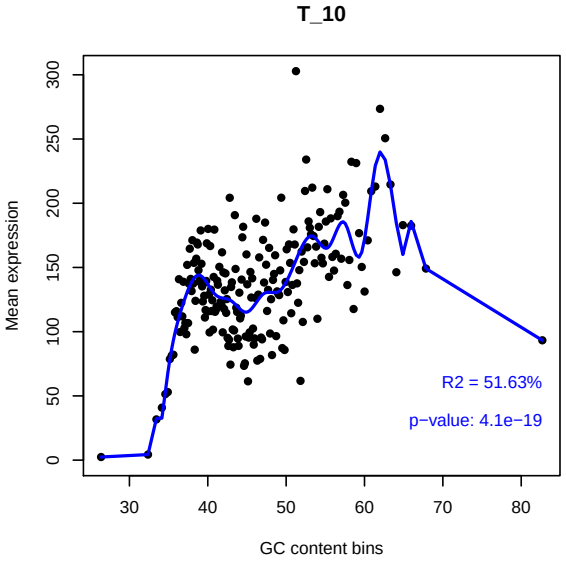


Diagnostic plot for GC content bias

WARNING. At least one of the model p−values was lower than 0.05, but R2 < 70% for at least one condition.

Normalization for correcting GC content bias could be advisable.
Plese check in the plots below the strength of the relationship between GC content and expression.

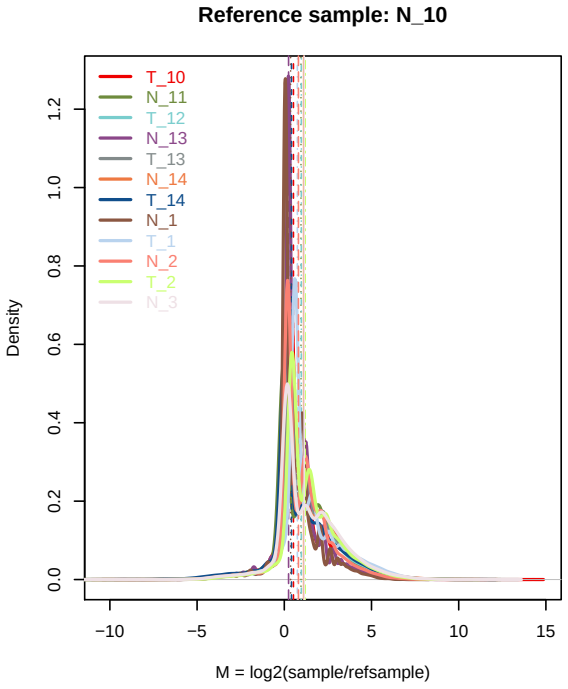




Diagnostic plot for differences in RNA composition

FAILED. There is a pair of samples with significantly different RNA composition

Normalization for correcting this bias is required.



Confidence intervals for median of M values

Sample	0.11%	99.89%	Diagnostic Test
T_10	0.4881	0.566	FAILED
N_11	0.9027	0.9564	FAILED
T_12	0.9211	1.0132	FAILED
N_13	0.2454	0.2906	FAILED
T_13	1.08	1.1778	FAILED
N_14	1.0874	1.1099	FAILED
T_14	0.3523	0.4732	FAILED
N_1	0.3262	0.3966	FAILED
T_1	0.6985	0.7956	FAILED
N_2	0.7666	0.8518	FAILED
T_2	1.1539	1.2401	FAILED
N_3	1.0704	1.1668	FAILED
T_3	0.9633	1.1055	FAILED
N_4	0.5381	0.659	FAILED
N_5	0.5988	0.6994	FAILED
T_6	0.9491	1.0409	FAILED
N_7	−0.1104	−0.0093	FAILED
T_7	1.0063	1.0996	FAILED
N_8	0.4158	0.5271	FAILED
T_8	0.0205	0.0205	FAILED
N_9	0.1355	0.1355	FAILED
T_9	0.1982	0.1982	FAILED

References

[1] Abecasis, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R.A., Hurles, M.E., McVean, G.A., Bentley, D., Chakravarti, A. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073. 10

[2] Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J.M., Conde, L., Blaschke, C., Vera, J. & Dopazo, J. (2006). BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research*, **34**, W472–476. 66

[3] Al-Shahrour, F., Carbonell, J., Minguez, P., Goetz, S., Conesa, A., Tárraga, J., Medina, I., Alloza, E., Montaner, D. & Dopazo, J. (2008). Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Research*, **36**, W341–W346. 21

[4] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106. 13, 94, 96, 107, 109, 111, 128, 130, 166

[5] Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W. & Robinson, M.D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature Protocols*, **8**, 1765–1786. 90

[6] Anders, S., Pyl, P.T. & Huber, W. (2014). HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 79, 80, 110

[7] Anderson, J. (2005). RNA turnover: unexpected consequences of being tailed. *Current biology*, **15**, R635–R638. 81

[8] Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data. *URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc*. 75, 102

[9] Argout, X., Salse, J., Aury, J., Guiltinan, M., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. *et al.* (2010). The genome of Theobroma cacao. *Nature Genetics*, **43**, 101–108. 10

[10] Barcikowski, R.S. & Robey, R.R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *The American Statistician*, **38**, 148–150. 51

[11] Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. 16, 125, 126

[12] Benjamini, Y. & Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, **40**, e72–e72. 76, 93, 104

[13] Bi, Y. & Davuluri, R.V. (2013). NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 262. 108

[14] Blencowe, B.J., Ahmad, S. & Lee, L.J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development*, **23**, 1379–1386. 76

[15] Bloom, J., Khan, Z., Kruglyak, L., Singh, M. & Caudy, A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, **10**, 221+. 77, 126

[16] Bolstad, B.M., Irizarry, R.A., Åstrand, M. & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193. 94

[17] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, **25**, 484–498. 37, 38, 39

[18] Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94+. 13, 76, 77, 94, 95, 96, 100, 108, 109, 110, 121, 165

[19] Cai, G., Li, H., Lu, Y., Huang, X., Lee, J., Müller, P., Ji, Y. & Liang, S. (2012). Accuracy of RNA-Seq and its dependence on sequencing depth. *BMC Bioinformatics*, **13**, S5. 76

[20] Cao, K.A.L., Robert-Granié, C., Rossouw, D. & Besse, P. (2008). A Sparse PLS for Variable Selection when Integrating Omics Data. *Statistical Applications in Genetics and Molecular Biology*. 16

[21] Carcel-Trullols, J., Aguilar-Gallardo, C., Garcia-Alcalde, F., Pardo-Cea, M.A., Dopazo, J., Conesa, A. & Simón, C. (2012). Transdifferentiation of MALME-3M and MCF-7 Cells toward Adipocyte-like Cells is Dependent on Clathrin-mediated Endocytosis. *SpringerPlus*, **1**, 1–12. 108

[22] Carmeliet, P. (2005). Angiogenesis in life, disease and medicine. *Nature*, **438**, 932–936. 66

[23] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005). The transcriptional landscape of the mammalian genome. *Science (New York, NY)*, **309**, 1559. 11, 81

[24] Casella, G. & Berger, R. (2002). Statistical inference. *Pacific Grove, California, USA: Duxbury Press*. 123

[25] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). Graphical methods for data analysis. *Wadsworth&Brooks/Cole, Pacific Grove, CA*. 130

[26] Chen, G., Chen, J., Shi, C., Shi, L., Tong, W. & Shi, T. (2013). Dissecting the Characteristics and Dynamics of Human Protein Complexes at Transcriptome Cascade Using RNA-Seq Data. *PLOS ONE*, **8**, e66521. 108

[27] Choi, S.C. & Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, **11**, 683–690. 40

[28] Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676. 66

[29] Conesa, A., Nueda, M.J., Ferrer, A. & Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102. 21, 25

[30] Conesa, A., Bro, R., García-García, F., Prats, J.M., Götz, S., Kjeldahl, K., Montaner, D. & Dopazo, J. (2008). Direct functional assessment of the composite phenotype through multivariate projection strategies. *Genomics*. 21

[31] Conesa, A., Prats-Montalbán, J., Tarazona, S., Nueda, M. & Ferrer, A. (2010). A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics and Intelligent Laboratory Systems*, **104**, 101–111. 17, 25, 33, 68, 71

[32] Consortium, E.P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**, 57–74. 6, 78

[33] Cui, X. & Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, **4**, 210. 10, 107

[34] Dai, J., Lieu, L. & Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical applications in genetics and molecular biology*, **5**, 6. 31

[35] Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240, ACM. 129

[36] de Tayrac, M., Le, S., Aubry, M., Mosser, J. & Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, **10**, 32+. 17

[37] DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W. & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532. 102

[38] Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M. & Jaffrézic, F. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. 77, 93

[39] Dudoit, S., Shaffer, J.P. & Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 71–103. 16

[40] Durban, J., Pérez, A., Sanz, L., Gómez, A., Bonilla, F., Chacón, D., Sasa, M., Angulo, Y., Gutiérrez, J.M. & Calvete, J.J. (2013). Integrated "omics" profiling indicates that miRNAs are modulators of the ontogenetic venom composition shift in the Central American rattlesnake, Crotalus simus simus. *BMC Genomics*, **14**, 234. 108

[41] Edgington, E. (1980). *Randomization Tests*. 37, 39, 41

[42] Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, **99**, 96–104. 39

[43] Efron, B., Tibshirani, R., Storey, J. & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160. 39, 121, 123, 125, 126, 137, 165

[44] Ferreira, P.G., Patalano, S., Chauhan, R., Ffrench-Constant, R., Gabaldon, T., Guigo, R. & Sumner, S. (2013). Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biology*, **14**, R20. 108

[45] Flicek, P. & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, **6**, S6–S12. 75

[46] Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. *et al.* (2011). Ensembl 2011. *Nucleic Acids Research*, **39**, D800. 110

[47] Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. *et al.* (2012). Ensembl 2012. *Nucleic Acids Research*, **40**, D84–D90. 79, 80

[48] Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S. *et al.* (2013). Ensembl 2013. *Nucleic Acids Research*, **41**, D48–D55. 80

[49] García-Alcalde, F., Okonechnikov, K., Carbonell, J., Ruiz, L.M., Götz, S., Tarazona, S., Meyer, T.F. & Conesa, A. (2012). Qualimap: evaluating next generation sequencing alignment data. *Bioinformatics*, **28**, 2678–2679. 25, 75, 77, 102, 174

[50] Gentleman, R.C., Carey, V.J., Bates, D.M. & others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80. 77

[51] Gnanadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81–124. 40

[52] Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B. *et al.* (2010). The developmental transcriptome of Drosophila melanogaster. *Nature*. 11

[53] Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D., Corbett, R., Tang, M.J., Hou, Y.C., Pugh, T.J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J.K., Chan, S.Y., Li, H.I., McDonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G.B., Jones, S.J.M., Tai, I.T. & Marra, M.A. (2010). Alternative expression analysis by RNA sequencing. *Nature Methods*, **7**, 843–847. 100, 109, 110

[54] Grzechnik, P. & Kufel, J. (2008). Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. *Molecular cell*, **32**, 247–258. 81

[55] Guo, Q., Wu, W., Massart, D.L., Boucon, C. & de Jong, S. (2002). Feature selection in Principal Component Analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, **61**, 123–132. 32

[56] Hannun, Y.A. & Obeid, L.M. (2008). Principles of bioactive lipid signalling: lessons from sphingolipids. *Nature Reviews Molecular Cell Biology*, **9**, 139–150. 66

[57] Hansen, K.D., Brenner, S.E. & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, **38**, e131. 76

[58] Hansen, K.D., Irizarry, R.A. & Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216. 94

[59] Hardcastle, T. & Kelly, K. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422+. 107, 109, 111, 127, 130

[60] Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. *et al.* (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32. 75

[61] Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, **22**, 1760–1774. 79

[62] Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84. 31

[63] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441. 31

[64] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 498–520. 31

[65] Huang, S.C., Sheu, B.C., Chang, W.C., Cheng, C.Y., Wang, P.H. & Lin, S. (2009). Extracellular matrix proteases - cytokine regulation role in cancer and pregnancy. *Frontiers in Bioscience*, **14**, 1571–1588. 68

[66] Jackson, J.E. & Mudholkar, G.S. (1979). Control procedures for residuals associated with Principal Component Analysis. *Technometrics*, **21**, 341–349. 39

[67] Jiang, H. & Wong, W.H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032. 126

[68] Jolliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag. 31

[69] Jolliffe, I.T. (1972). Discarding variables in a Principal Component Analysis. I: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **21**, 160–173. 31

[70] Jolliffe, I.T. (1973). Discarding variables in a Principal Component Analysis. II: Real data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **22**, 21–31. 31

[71] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**, R36. 79

[72] Kim, V., Han, J. & Siomi, M. (2009). Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, **10**, 126–139. 81

[73] Klambauer, G., Unterthiner, T. & Hochreiter, S. (2013). DEXUS: identifying differential expression in RNA-Seq studies with unknown conditions. *Nucleic Acids Research*, **41**, e198–e198. 108

[74] Kong, S.W., Pu, W.T. & Park, P.J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*. 17

[75] Kovacic, J.C., Moore, J., Herbert, A., Ma, D., Boehm, M. & Graham, R.M. (2008). Endothelial progenitor cells, angioblasts, and angiogenesisold terms reconsidered from a current perspective. *Trends in Cardiovascular Medicine*, **18**, 45–51. 66

[76] Krzanowski, W.J. (1987). Selection of variables to preserve multivariate data structure, using Principal Components. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **36**, 22–33. 32

[77] Łabaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S. & Kreil, D.P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**, i383–i391. 75

[78] Landgrebe, J., Wurst, W. & Welzl, G. (2002). Permutation-validated principal components analysis of microarray data. *Genome Biology*. 16

[79] Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15(2):R29**. 108, 164

[80] Leinonen, R., Sugawara, H. & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, **39**, D19. 110

[81] Lemay, J., D'Amours, A., Lemieux, C., Lackner, D., St-Sauveur, V., Bähler, J. & Bachand, F. (2010). The nuclear poly (A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. *Molecular cell*, **37**, 34–45. 81

[82] Li, J. & Tibshirani, R. (2011). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*. 17, 108, 109, 128, 167

[83] Li, J., Witten, D.M., Johnstone, I.M. & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538. 18

[84] Li, N., Ye, M., Li, Y., Yan, Z., Butcher, L., Sun, J., Han, X., Chen, Q. *et al.* (2010). Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, **52**, 203–212. 10

[85] Liu, W.Y., Chang, Y.M., Chen, S.C.C., Lu, C.H., Wu, Y.H., Lu, M.Y.J., Chen, D.R., Shih, A.C.C., Sheue, C.R., Huang, H.C. *et al.* (2013). Anatomical and transcriptional dynamics of maize embryonic leaves during seed germination. *Proceedings of the National Academy of Sciences*, **110**, 3979–3984. 108

[86] Liu, Y., Zhou, J. & White, K.P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304. 109

[87] Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K.D., Fronick, C., Schmidt, H., Fulton, L.A., Fulton, R.S., Nelson, J.O., Magrini, V., Pohl, C., Graves, T.A., Markovic, C., Cree, A., Dinh, H.H., Hume, J., Kovar, C.L., Fowler, G.R., Lunter, G., Meader, S., Heger, A., Ponting, C.P., Marques-Bonet, T., Alkan, C., Chen, L., Cheng, Z., Kidd, J.M., Eichler, E.E., White, S., Searle, S., Vilella, A.J., Chen, Y., Flicek, P., Ma, J., Raney, B., Suh, B., Burhans, R., Herrero, J., Haussler, D., Faria, R., Fernando, O., Darré, F., Farré, D., Gazave, E., Oliva, M., Navarro, A., Roberto, R., Capozzi, O., Archidiacono, N., Valle, G.D., Purgato, S., Rocchi, M., Konkel, M.K., Walker, J.A., Ullmer, B., Batzer, M.A., Smit, A.F.A., Hubley, R., Casola, C., Schrider, D.R., Hahn, M.W., Quesada, V., Puente, X.S., Ordoñez, G.R., López-Otín, C., Vinar, T., Brejova, B., Ratan, A., Harris, R.S., Miller, W., Kosiol, C., Lawson, H.A., Taliwal, V., Martins, A.L., Siepel, A., Roy-Choudhury, A., Ma, X., Degenhardt, J., Bustamante, C.D., Gutenkunst, R.N., Mailund, T., Dutheil, J.Y., Hobolth, A., Schierup, M.H., Ryder, O.A., Yoshinaga, Y., de Jong, P.J., Weinstock, G.M., Rogers, J., Mardis, E.R., Gibbs, R.A. & Wilson, R.K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533. 10

[88] Lu, Y., Cohen, I., Zhou, X.S. & Tian, Q. (2007). Feature selection using Principal Feature Analysis. In *ACM Multimedia Conference*. 32

[89] Lukashin, A.V. & Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414. 31

[90] Luss, R. & d'Aspremont, A. (2008). Clustering and feature selection using Sparse Principal Component Analysis. *Optimization and Engineering*, 1573–2924. 32

[91] Malone, J. & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, **9**, 34. 75

[92] Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517. 12, 13, 75, 100, 107, 121, 126, 165

[93] Martens, H. & Naes, T. (1989). *Multivariate calibration*. John Wiley & Sons, Ltd. Chichester. 36

[94] Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451. 50

[95] McCabe, G.P. (1984). Principal Variables. *Technometrics*, **26**, 137–144. 32

[96] McGill, R., Tukey, J.W. & Larsen, W.A. (1978). Variations of box plots. *The American Statistician*, **32**, 12–16. 130

[97] McIntyre, L., Lopiano, K., Morse, A., Amin, V., Oberg, A., Young, L. & Nuzhdin, S. (2011). RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 293+. 95

[98] Metzker, M.L. (2009). Sequencing technologies-the next generation. *Nature Reviews Genetics*, **11**, 31–46. 75

[99] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628. 75, 76, 94, 95, 124

[100] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**, 1344–1349. 11

[101] Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlén, M. & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*. 108

[102] Nueda, M., Carbonell, J., Medina, I., Dopazo, J. & Conesa, A. (2010). Serial Expression Analysis: a web tool for the analysis of serial gene expression data. *Nucleic Acids Research*, **38**, W239. 25, 33, 71

[103] Nueda, M.J., Conesa, A., Westerhuis, J.A., Hoefsloot, H.C.J., Smilde, A.K., Talón, M. & Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics*, **23**, 1792–1800. 16, 21, 23, 31, 33, 34, 36, 39, 41, 70

[104] Nueda, M.J., Sebastián, P., Tarazona, S., García-García, F., Dopazo, J., Ferrer, A. & Conesa, A. (2009). Functional assessment of time course microarray data. *BMC Bioinformatics*, **10**, S9. 17, 21, 25

[105] Nueda, M.J., Tarazona, S. & Conesa, A. (2014). Next masigpro: updating masigpro bioconductor package for rna-seq time series. *Bioinformatics*, **30**, 14. 25, 107

[106] Oshlack, A. & Wakefield, M. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, **4**, 14+. 76, 103

[107] Oshlack, A., Robinson, M. & Young, M. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, **11**, 220+. 11, 164, 165

[108] Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554. 16, 21, 107

[109] Pan, W., Lin, J. & Le, C.T. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & integrative genomics*, **3**, 117–124. 17, 123

[110] Park, P. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680. 10

[111] Patel, R.K. & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619. 102

[112] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572. 31

[113] Ponzoni, I., Nueda, M.J., Tarazona, S., Götz, S., Montaner, D., Dussaut, J.S., Dopazo, J. & Conesa, A. (2014). Pathway network inference from gene expression data. *BMC Systems Biology*, **8**, 1–17. 25, 33, 68, 71

[114] Prado-López, S., Conesa, A., Armiñán, A., Martínez-Losa, M., Escobedo-Lucea, C., Gandia, C., Tarazona, S., Melguizo, D., Blesa, D., Montaner, D., Sanz-González, S., Sepúlveda, P., Götz, S., O'Connor, J.E., Moreno, R., Dopazo, J., Burks, D.J. & Stojkovic, M. (2010). Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium. *Stem Cells*, **28**, 407–418. 16, 25, 62

[115] Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496–501. 10, 21

[116] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D. & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, **14**, R95. 108

[117] Raychaudhuri, S., Stuart, J.M. & Altman, R.B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput.*, 455–466. 31

[118] Ren, S., Peng, Z., Mao, J.H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W. *et al.* (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Research*, **22**, 806–821. 79

[119] Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, **12**, 480+. 76, 77, 93, 94, 95, 104

[120] Robinson, M.D. & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25+. 77, 92, 94, 95, 103, 124

[121] Robinson, M.D. & Smyth, G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887. 16, 127

[122] Robinson, M.D. & Smyth, G.K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332. 13, 17, 128

[123] Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140. 94, 99, 107, 109, 111, 112, 127, 128, 130, 166

[124] Robles, J.A., Qureshi, S.E., Stephen, S.J., Wilson, S.R., Burden, C.J. & Taylor, J.M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, **13**, 484. 75, 111, 114

[125] Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, **27**, 832–837. 43

[126] Russo, F. & Angelini, C. (2014). Rnaseqgui: A gui for analysing rna-seq data. *Bioinformatics*, btu308. 102, 177

[127] Shearman, J.R., Jantasuriyarat, C., Sangsrakru, D., Yoocha, T., Vannavichit, A., Tragoonrung, S. & Tangphatsornruang, S. (2013). Transcriptome analysis of normal and mantled developing oil palm flower and fruit. *Genomics*, **101**, 306–312. 108

[128] Shi, L., Reid, L., Jones, W., Shippy, R., Warrington, J., Baker, S., Collins, P., De Longueville, F., Kawasaki, E., Lee, K. *et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151–1161. 109

[129] Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC. 43, 125

[130] Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132. 109

[131] Slomovic, S., Laufer, D., Geiger, D. & Schuster, G. (2006). Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Research*, **34**, 2966. 81

[132] Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamers, R.J.A.N., van der Greef, J. & Timmerman, M.E. (2005). ANOVA-Simultaneous Component Analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048. 31, 33

[133] Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, 3. 10, 16, 17, 21, 107, 108

[134] Soneson, C. & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91. 90, 96, 108, 109, 111

[135] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297. 31

[136] Srivastava, S. & Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, **38**, e170. 107

[137] Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P., Consortium, R. *et al.* (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, **10**, 1177–1184. 95

[138] Storey, J.D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445. 16

[139] Suárez, E., Burguete, A. & Mclachlan, G. (2009). Microarray Data Analysis for Differential Expression: a Tutorial. *PRHSJ*. 21

[140] Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. & Yaspo, M. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960. 107

[141] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, **21**, 2213–2223. 25, 77, 103, 108, 121, 165, 177

[142] Tarazona, S., Prado-López, S., Dopazo, J., Ferrer, A. & Conesa, A. (2012). Variable selection for multifactorial genomic data. *Chemometrics and Intelligent Laboratory Systems*, **110**, 113–122. 25

[143] Trapnell, C., Pachter, L. & Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111. 80, 110

[144] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515. 11, 164

[145] Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311. 31

[146] Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagné, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner,

R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F. & Viola, R. (2010). The genome of the domesticated apple (Malus domestica Borkh.). *Nature Genetics*, **42**, 833–839. 10

[147] Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138. 126, 130

[148] Wang, L., Wang, S. & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185. 102

[149] Westad, F., Hersleth, M., Lea, P. & Martens, H. (2003). Variable selection in PCA in sensory descriptive and consumer data. *Food Quality and Preference*, **14**, 463–472. 32

[150] Yang, Y.H., Dudoit, S., Luu, P. & Speed, T.P. (2001). Normalization for cDNA microarry data. In *BiOS 2001 The International Symposium on Biomedical Optics*, 141–152, International Society for Optics and Photonics. 14

[151] Yener, B., Acar, E., Aguis, P., Bennett, K., Vandenberg, S. & Plopper, G. (2008). Multiway modeling and analysis in stem cell systems biology. *BMC Systems Biology*, **2**, 63. 31

[152] Young, M.D., Wakefield, M.J., Smyth, G.K. & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, **11**, R14. 103, 160

[153] Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S., Lu, Y., Denoeud, F., Antonarakis, S., Snyder, M. *et al.* (2007). Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Research*, **17**, 839. 81

[154] Zheng, W., Chung, L.M. & Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, **12**, 290. 77, 94

[155] Zhu, Q.H., Stephen, S., Kazan, K., Jin, G., Fan, L., Taylor, J., Dennis, E.S., Helliwell, C.A. & Wang, M.B. (2013). Characterization of the defense transcriptome responsive to Fusarium oxysporum-infection in Arabidopsis using RNA-seq. *Gene*, **512**, 259–266. 108