UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Computational methods for the analysis of gene expression from RNA sequencing data

Ph.D. candidate
Francesca Finotello

Advisor
Prof. Barbara Di Camillo

School Director
Prof. Matteo Bertocco

Bioengineering Coordinator
Prof. Giovanni Sparacino

Ph.D. School in Information Engineering
Bioengineering section, XXVI Series, 2014

Abstract

In every living organism, the entirety of its hereditary information is encoded, in the form of DNA, through the so-called genome. The genome consists in both genes and non-coding sequences and contains the whole information needed to determine all the properties and functions of each single cell. Cells can access and translate specific instructions of this code through *gene expression*, namely by selectively switching on and off a particular set of genes. Thanks to gene expression, the information encoded into the active genes is transcribed into RNAs. This set of RNAs reflects the current state of a cell and can reveal pathological mechanisms underlying diseases.

In recent years, a novel methodology for RNA sequencing, called *RNA-seq*, is replacing microarrays for the study of gene expression. The sequencing framework of RNA-seq methodology enables to investigate at high resolution all the RNA species present in a sample, characterizing their sequences and quantifying their abundances at the same time. In practice, millions of short sequences, called *reads*, are sequenced from random positions of the input RNAs. These reads can then be computationally mapped on a reference genome to reveal a *transcriptional map*, where the number of reads aligned on each gene, called *counts*, gives a measure of its level of expression. At first glance, this scheme may seem very simple, but the implementation of the whole analysis workflow is in fact complex and not well defined. So far, many computational methods have been proposed to perform the different steps of RNA-seq data analysis, but a unified processing pipeline is still lacking.

The aim of my Ph.D. research project was the implementation of a robust computational pipeline for RNA-seq data analysis, from data pre-processing to differential expression detection. The definition of the different analysis modules was carried out through several steps. First, we drafted a basic analysis framework through the study of RNA-seq data features and the dissection of data models and state-of-the-art algorithmic strategies. Then, we focused on count bias, which is one of the most challenging aspects of RNA-seq data analysis. We demonstrated that some biases affecting counts can be effectively corrected with current normalization methods, while others, like length bias, cannot be completely removed without introducing additional systematic errors. Thus, we defined a novel approach to compute RNA-seq counts, which strongly reduces length bias prior to normalization and is robust to the upstream processing steps. Finally, we defined the complete analysis pipeline considering the best preforming methods and optimized some specific processing steps to enable correct expression estimates even in the presence of high-similarity genomic sequences.

The implemented analysis pipeline was applied to a real case study to identify the genes involved in the pathogenesis of spinal muscular atrophy (SMA) from RNA-seq data of patients and healthy controls. SMA is a degenerative neuromuscular disease that has no cure and represents one of the major genetic causes of infant mortality. We identified a set of genes related to skeletal muscle and connective tissue disorders whose patterns of differential expression correlate with phenotype and may underlie protective mechanisms against SMA progression. Some putative positive targets identified by this analysis are currently under biological validation since they might improve diagnostic screening and therapy.

To pose the basis for future research, which will focus on the optimization of the processing pipeline and to its extension to the analysis of dynamic expression data, we designed two time-series RNA-seq data sets: a real one and a simulated one. The experimental and sequencing design of the real data set, as well as the modelling of the synthetic data, have been an integral part of the Ph.D. activity.

Overall, this thesis considers each step of the RNA-seq data processing and provides some valuable guidelines in a fast-evolving research field that, up to now, has prevented the establishment of a stable and standardized analysis scheme.

Sommario

Il patrimonio genetico di ogni organismo vivente è codificato, sotto forma di DNA, nel *genoma*. Il genoma è costituito da geni e da sequenze non codificanti e racchiude in sé tutte le informazioni necessarie al corretto funzionamento delle cellule dell'organismo. Le cellule possono accedere a specifiche istruzioni di questo codice tramite un processo chiamato *espressione genica*, ovvero attivando o disattivando un particolare set di geni e trascrivendo l'informazione necessaria in RNA. L'insieme degli RNA trascritti caratterizza quindi un preciso stato cellulare e può fornire importanti informazioni sui meccanismi coinvolti nella patogenesi di una malattia.

Recentemente, una metodologia per il sequenziamento dell'RNA, chiamata *RNA-seq*, sta rapidamente sostituendo i microarray nello studio dell'espressione genica. Grazie alle proprietà delle tecnologie di sequenziamento su cui è basato, l'RNA-seq permette di misurare il numero di RNA presenti in un campione e al contempo di "leggerne" l'esatta sequenza. In realtà, il sequenziamento produce milioni di sequenze, chiamate "read", che rappresentano piccole stringhe lette da posizioni random degli RNA in input. Le read devono quindi essere mappate con un algoritmo su un genoma di riferimento, in modo da ricostruire una mappa trascrizionale, in cui il numero di read allineate su ciascun gene dà una misura digitale (chiamata "count") del suo livello di espressione. Sebbene a prima vista questa procedura possa sembrare molto semplice, lo schema di analisi integrale è in realtà molto complesso e non ben definito. In questi anni sono stati sviluppati diversi metodi per ciascuna delle fasi di elaborazione, ma non è stata tuttora definita una pipeline di analisi dei dati RNA-seq standardizzata.

L'obiettivo principale del mio progetto di dottorato è stato lo sviluppo di una pipeline computazionale per l'analisi di dati RNA-seq, dal pre-processing alla misura dell'espressione genica differenziale. I diversi moduli di elaborazione sono stati definiti e implementati tramite una serie di passi successivi. Inizialmente, abbiamo considerato e ridefinito metodi e modelli per la descrizione e l'elaborazione dei dati, in modo da stabilire uno schema di analisi preliminare. In seguito, abbiamo considerato più attentamente uno degli aspetti più problematici dell'analisi dei dati RNA-seq: la correzione dei bias presenti nei count. Abbiamo dimostrato che alcuni di questi bias possono essere corretti in modo efficace tramite le tecniche di normalizzazione correnti, mentre altri, ad esempio il *length bias*, non possono essere completamente rimossi senza introdurre ulteriori errori sistematici. Abbiamo quindi definito e testato un nuovo approccio per il calcolo dei count che minimizza i bias ancora prima di procedere con un'eventuale normalizzazione. Infine, abbiamo implementato la pipeline di analisi completa considerando gli algoritmi più robusti e accurati, selezionati nelle fasi precedenti, e ottimizzato alcun step in modo

da garantire stime dell'espressione genica accurate anche in presenza di geni ad alta similarità.

La pipeline implementata è stata in seguito applicata ad un caso di studio reale, per identificare i geni coinvolti nella patogenesi dell'atrofia muscolare spinale (SMA). La SMA è una malattia neuromuscolare degenerativa che costituisce una delle principali cause genetiche di morte infantile e per la quale non sono ad oggi disponibili né una cura né un trattamento efficace. Con la nostra analisi abbiamo identificato un insieme di geni legati ad altre malattie del tessuto connettivo e muscoloscheletrico i cui pattern di espressione differenziale correlano con il fenotipo, e che quindi potrebbero rappresentare dei meccanismi protettivi in grado di combattere i sintomi della SMA. Alcuni di questi target putativi sono in via di validazione poiché potrebbero portare allo sviluppo di strumenti efficaci per lo screening diagnostico e il trattamento di questa malattia.

Gli obiettivi futuri riguardano l'ottimizzazione della pipeline definita in questa tesi e la sua estensione all'analisi di dati dinamici da *time-series RNA-seq*. A questo scopo, abbiamo definito il design di due data set *time-series*, uno reale e uno simulato. La progettazione del design sperimentale e del sequenziamento del data set reale, nonché la modellazione dei dati simulati, sono stati parte integrante dell'attività di ricerca svolta durante il dottorato.

L'evoluzione rapida e costante che ha caratterizzato i metodi per l'analisi di dati RNA-seq ha impedito fino ad ora la definizione di uno schema di analisi standardizzato e la risoluzione di problematiche legate a diversi aspetti dell'elaborazione, quali ad esempio la normalizzazione. In questo contesto, la pipeline definita in questa tesi e, più in ampiamente, i temi discussi in ciascun capitolo, toccano tutti i diversi aspetti dell'analisi dei dati RNA-seq e forniscono delle linee guida utili a definire un approccio computazionale efficace e robusto.

Contents

1
Introduction

In every organism, DNA encodes all the instruction required to build the RNAs and proteins that are needed to make functioning its cells. Nevertheless, the complexity of an organism and its ability to evolve through diverse developmental stages or to respond to environmental stimuli is not explained by this static set of instructions, but by how and when these instructions are accessed. As pointed out by Alberts *et al.* [1], "*a complete description of the DNA sequence of an organism does not enable us to reconstruct the organism any more than a list of all the english words in a dictionary enables us to reconstruct a play by Shakespeare*". Great part of organisms' complexity and dynamicity is indeed explained by *gene expression*: each cell can selectively activate the set of genes required for executing specific functions or responses to stimuli. Gene expression allows selecting specific instructions from the whole genetic information encoded in DNA and is initiated with their transcription into temporary RNA copies. Although RNA transcription is just the activation of a cascade of processes and control mechanisms that made up the complex gene expression machinery, substantial insights can be drawn from the study of organisms' transcriptional maps. In 2008, the advent of a new methodology called *RNA-seq*, has revolutionized transcriptomics research enabling the simultaneous characterization of the sequences of the transcripts present in a cell and the quantification of their expression levels. The possibility of sequencing transcriptomes at single-base resolution is borrowed from Next-Generation Sequencing platforms, which are the technological framework of the RNA-seq methodology. Indeed, Next-Generation Sequencing technologies produce enormous amount of data, enabling to sequence entire genomes and transcriptomes in a single instrument run at dramatically reduced time and costs. Despite being already widely used, RNA-seq is a very recent methodology that is experiencing a fast and continuous development of both experimental and computational procedures. In particular, the number of available methods for performing each step of RNA-seq data analysis has grown at such a fast pace so to prevent the definition of a unified and standardized computational pipeline.

In this scenario, the research described in this thesis was originally motivated by the need of a definition of a robust computational pipeline for the processing of gene expression in RNA-seq studies, focusing on the least characterized or most critical aspects of data modelling and analysis. In particular, we outlined a comprehensive mathematical description of data generation starting from initial transcript levels and reviewed currently available methods for RNA-seq data analysis. In addition, we assessed and compared state-of-the-art methods for data normalization to identify the best strategy to correct for different systematic biases affecting data. Due to the limits that we found in most of the assessed methods, we defined and implemented a new strategy to compute RNA-seq data, so to directly reduce bias before normalization. Finally, we optimized the quantification strategy by applying this approach and defining a robust scheme for data pre-processing. Overall, this thesis encompassess the definition of a computational framework to detect differential gene expression in human RNA-seq experiments and its application to a real case study of spinal muscular atrophy.

All the contributions described in this thesis, and related to the definition, implementation and application of the analysis pipeline, are the results of the research activity carried out within my Ph.D. program supported by Fondazione CARIPARO: “RNA sequencing for quantitative transcriptomics”. My full Ph.D. research activity has more widely dealt with the implementation of computational methods for the analysis of Next-generation Sequencing data, with application to human health. I was involved in collateral studies regarding: the assembly and characterization of human pathogens from whole-genome shotgun pyrosequencing; the study of the human microbiome in chronic obstructive pulmonary disease and colon cancer through 16s sequencing; the characterization of causative genes in complex diseases through RNA-seq and exome sequencing. These research activities are not described in this thesis, but can be partially outlined by the full list of publications shown in Appendix C (updated to January 28th, 2014).

This chapter provides a short introduction to three prerequisite topics that outline the research context of this thesis: gene expression, Next-Generation Sequencing technologies and RNA-seq methodology. In addition, the aim and the structure of the thesis are presented.

1.1 Gene expression

In every organism, from bacteria to humans, the whole information needed to build and make functioning the cells is stored into DNA, in the so-called *genome*. The DNA contains all “instructions” required to generate the proteins supporting every cell process. DNA, acronym for deoxyribonucleic acid, is a macromolecule made up of four basic “blocks” called *nucleotides*. Each nucleotide consists of a nitrogenous bases (or simply “base”), a deoxyribose sugar and a phosphate. Nucleotides can be of four different kinds, depending on the base that they comprise: adenine (A), guanine (G), cytosine (C) or thymine (T). These four nucleotides are concatenated one to the other forming a strand (Figure 1.1). In particular, a phosphodiester bond joins the 5’ end on one nucleotide to the 3’ end of the previous one, creating a sort of “backbone” of sugars and phosphates. However, DNA is not single-stranded: the nitrogenous bases, which are hanging out of this strand, are connected to the nitrogenous bases of an antiparallel DNA strand (i.e. it is built in 5’-to-3’ direction). The matching bases are joined through hydrogen bonds following a strict rule of complementarity: adenine only matches thymine, while cytosine only binds to guanine. Due to this property of complementarity, one DNA strand univocally determines the sequence of its antiparallel strand. The sequence in which the nucleotides succeed one another in the DNA double-stranded chain encodes the organism’s hereditary information. In particular, DNA comprises *genes*, which are sequences encoding specific proteins. Thus, to accomplish all the necessary processes and functions of the organism that give rise to its physical characteristics (*phenotypes*), the information encoded in the genomic sequence (i.e. the *genotype*) has to be translated into proteins. Proteins are macromolecules consisting of one or more chains of amino acid residues that accomplish several functions within living organisms, such as catalyzing metabolic reactions, responding to stimuli and transporting molecules. Different amino acid sequences fold into different three-dimensional structure, resulting in specific properties and activities. Basically, proteins determine the phenotypes of every living thing. Although several molecules, such as water, minerals and fats, give shape to organisms’ cells, proteins supply the framework for their correct organization and functioning.

But how the genetic information encoded into DNA (genotype) is translated into phenotype? This conversion is accomplished through an intermediary step of transcription in which the genetic information required to generate the final protein is transcribed

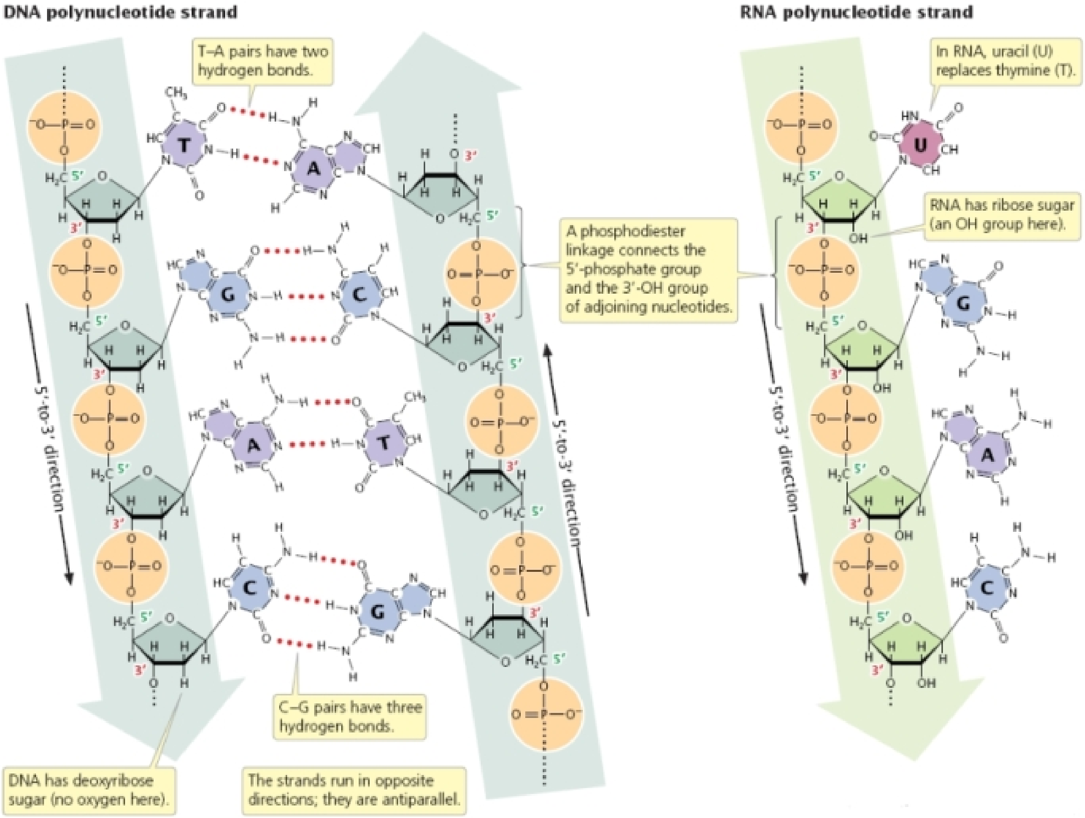

DNA polynucleotide strand
RNA polynucleotide strand
T–A pairs have two hydrogen bonds.
In RNA, uracil (U) replaces thymine (T).
RNA has ribose sugar (an OH group here).
A phosphodiester linkage connects the 5′-phosphate group and the 3′-OH group of adjoining nucleotides.
5′-to-3′ direction
5′-to-3′ direction
5′-to-3′ direction
C–G pairs have three hydrogen bonds.
DNA has deoxyribose sugar (no oxygen here).
The strands run in opposite directions; they are antiparallel.
T
A
G
C
A
T
C
G
U
G
A
C

Figure 1.1: Comparison of DNA and RNA structure. Image taken from [2] and originally proposed in [3].

into a temporary template: an RNA molecule. Ribonucleic acid (RNA), likely DNA, is a nucleic acid made up of a chain of nucleotides (Figure 1.1). Nevertheless, it has some peculiarities with respect to DNA:

- It is single-stranded and folds into characteristic secondary structures;
- It contains ribose sugar instead of deoxyribose, which makes it less stable than DNA;
- It contains uracil in place of thymine.

Once the piece of information needed is transcribed into an RNA molecule by an enzyme called *RNA polymerase*, RNA is transported from cell nucleus to cytoplasm to be used as blueprint for protein synthesis. Transcription keeps safe the whole hereditary information encoded into DNA, while a temporary copy of the message in the form of messenger RNA (mRNA) leaves the nucleus and reaches the cytoplasm. Following the instructions written in the mRNA template, the cell can *translate* DNA nucleotidic code into an amino acid sequence (a more detailed description of DNA transcription and translation can be found in [1]).

In eukaryotic organisms like humans, RNA is not directly exported to cytoplasm but it undergoes some post-transcriptional modifications that transform a pre-mRNA to a mature mRNA. The first modification involves the 5' end of RNA and consists in the addition of a methylated guanine nucleotide, through a process call "capping"; the 5'-methyl cap helps the cell to recognize mRNA from other molecules and protects it from degradation. Also the 3' end of mRNA is modified adding a long tail of adenine bases, called *poly-A* tail. This addition prevents mRNA to be quickly degraded: the longer the poly-A tail is, the longer the mRNA lasts and the more it is translated into proteins. The final step of this modification process is called *splicing* and prepares mRNA for translation: non-coding regions, called *introns* are removed and coding sequences, called *exons*, are concatenated together. Although the order of exons is always preserved, some exons can be removed along with introns, giving rise to different RNAs. This process, called *alternative splicing*, enables to produce different proteins (*isoforms*) starting from the same gene. In humans, alternative splicing allows to produce 90 000 different proteins starting from about 22 000 genes, dramatically increasing the coding potential of the human genome. This mature mRNAs are then transported out of the cell nucleus to ribosomes. Here mRNA, produced by DNA transcription, is decoded and translated by a ribosome complex to produce a specific amino acid chain that will later fold into an active protein.

The terms *transcription* and *translation* specifically identify two different processes, in which the DNA code is transcribed using the same nucleotidic "language" or translated in the different amino acid code of proteins. The propagation of the genetic information in cells from DNA to RNA to proteins is a fundamental process termed the *central dogma of molecular biology*. Despite the recognized universality of the central dogma, the existence of important variations in the genetic information flow, such as the above-mentioned RNA splicing, have been discovered. For instance, not all genes are translated into proteins, but some of them see their final product in RNAs that have structural and catalytic roles in the cell.

Besides the punctual description of the biological activities carried during gene expression, the power of this process can be better realized when we think that all the cells of an organism (with few exceptions) have the same genome. This means that in humans all cells carry the entire set of genes and that, for instance, nervous cells contain also the genetic information required for growing hair. This finding has been experimentally proved just in modern age, while ancient biologists originally thought that genes were selectively lost during cell differentiation specialized. But how is then cell differentiation achieved? How can cells result in so many and various shapes and functions despite having the same genetic blueprint? All these differences are achieved by changes in gene expression, namely by switching on and off different set of genes, thus expressing different proteins. In this way, each different cell type produces specialized proteins that are responsible for its distinctive properties. Besides the basic functions exploited by expressing these *housekeeping* genes, organisms can change gene expression in response to external signals. A nice example of this resilience is represented by the *Lac* operon in *E. coli*. The *Lac* operon comprehends a set of genes required for the transport and metabolism of lactose. The activation of the genes of the *Lac* operon depends substrate available and provides an insightful example of gene expression regulation. Indeed, when lactose is absent, a protein called *Lac repressor* binds the *Lac* operon blocking the access of RNA polymerase and thus preventing transcription. To be fully functioning, *Lac* operon have to interact with an *activator* protein called *CAP*. To be able to bind *Lac* operon, *CAP* has to ligate *cyclic AMP* (*cAMP*), a molecule that regulates several cellular responses. In absence of glucose, which is the preferred bacteria carbon source, intracellular *cAMP* concentration arises and *CAP* can bind to *Lac* operon and activate it, enabling the assimilation of lactose. This logic, depicted in Figure 1.2, integrates two distinct signals so to prevent wasteful activation of the *Lac* operon when lactose is not present or when glucose is available.

The Lac operon also provides an elegant example of how expression can be tightly

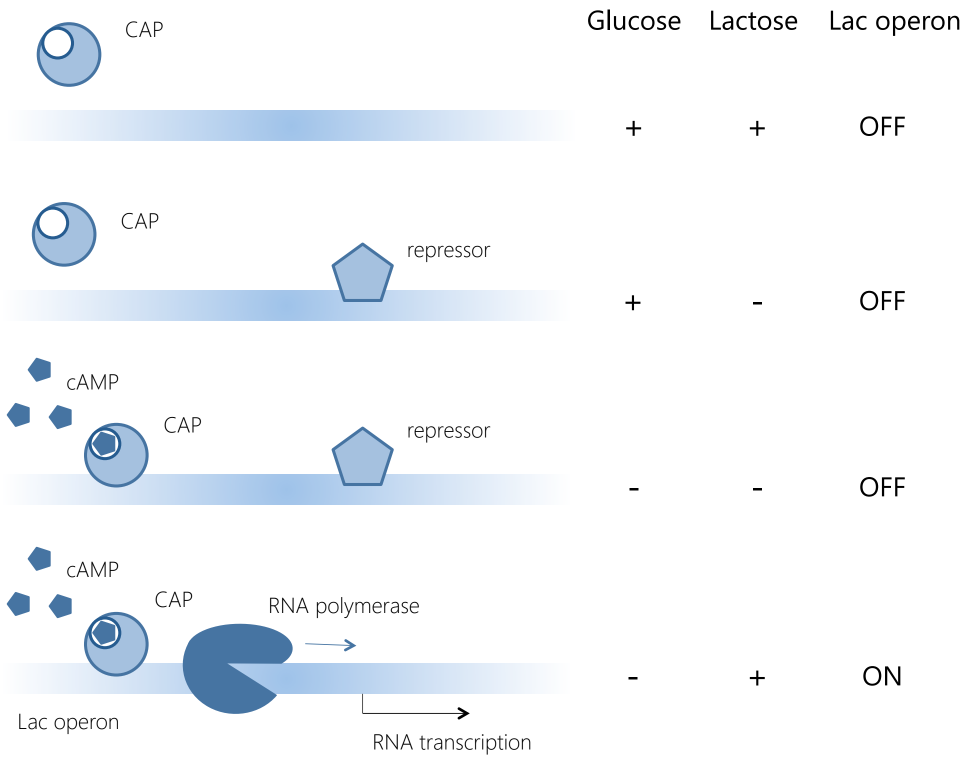


Figure 1.2: The expression of the *Lac* operon, needed for lactose digestion in *E. coli*, is activated by two signals: absence (-) of glucose and presence (+) of lactose.

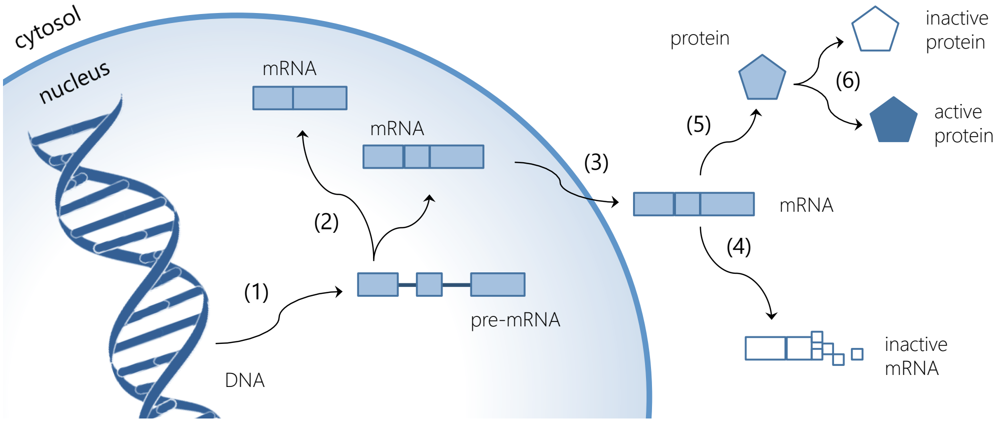


Figure 1.3: Gene expression can be regulated at several steps of the process: (1) gene transcription, (2) alternative splicing, (3) transport from nucleus to cytosol, (4) mRNA degradation, (5) translation into proteins and (6) protein activation and deactivation.

regulated at the transcriptional level by activators and repressor proteins. In eukaryotic organisms, gene expression can be tightly regulated at several levels of the flow from DNA to RNA to protein (Figure 1.3):

(1) Controlling when and how much a certain gene is transcribed;

(2) Controlling how an RNA transcript is (alternatively) spliced;

(3) Selecting which mRNAs are exported from the nucleus to the cytosol;

(4) Degrading specific mRNA molecules;

(5) Selecting which mRNAs are translated by ribosomes into proteins;

(6) Activating or inactivating proteins.

Nevertheless, for most genes, transcription is the primary control because it ensures that no unnecessary intermediates are synthesized (step 1 in Figure 1.3). Thus, the study of transcribed RNAs enables the reconstruction of a “gene transcription map” that reflects which genes are active within the cell in a specific condition and time. Moreover, this map can comprehend also RNA that are not later translated into proteins, thus allowing to investigate their role in other processes, such as regulation of gene expression.

1.2 Next-Generation Sequencing

DNA sequencing consists in the determination of the precise order of nucleotides that constitute a DNA molecule. Frederick Sanger, a British biochemist awarded twice with the Nobel Prize for chemistry, was one the first scientists working on the development of sequencing techniques. In 1975, together with Alan Coulson, he published a sequencing procedure, called “Plus and Minus” technique, in which the *E. coli* DNA polymerase was used to copy single-stranded DNA molecules [4]. Although the low automation of this method allowed the determination of just few hundreds of nucleotides at a time, Sanger’s group was able to exploit it to sequence the first genome: a single-stranded bacteriophage φX174 [5].

The first breakthrough followed shortly after, when Sanger’s group developed the “dideoxy chain-termination” method for sequencing DNA molecules, also known as the “Sanger method” [6]. Thanks to this new methodology, Sanger performed a more rapid and accurate sequencing of bacteriophage φX174 and earned his second Nobel prize in chemistry. A first revolution happened in early 90’s, with the advent of capillary electrophoresis, which eventually led to the development of the first "high-throughput" sequencers in 1998: the MegaBACE 1000 (GE Healthcare Life Sciences) and the ABI Prism 3700 (Applied Biosystems) [7]. These 96-capillary instruments allowed sequencing up to 96 DNA sequences in parallel. Despite several technological improvements, modern capillary-based platforms, such as the ABI Prism 3730 (Applied Biosystems), are still based on the same general scheme adopted to sequence the φX174 genome.

A second revolution took place in the last decade with the development of the so-called "Next-Generation Sequencing" (NGS) technologies, which increased the throughput by a factor of 100-1000 and greatly reduced sequencing costs at the same time [8, 9]. The massive parallelization of the sequencing process that characterizes these new technologies allows sequencing millions of sequences at the same time, reducing the costs due to the reagents needed and drastically increasing the throughput per run. Figure 1.4 shows the costs associated with DNA sequencing projects performed since 2001 for the Genome Sequencing Program of the National Human Genome Research Institute (NHGRI), and compares them to a curve representing the Moore's Law (orange). Moore's law describes a long-term trend in the computer hardware industry that makes the "compute power" doubling every two years. The out-pacing of Moore's Law has an evident start in January 2008, with the advent of NGS technologies.

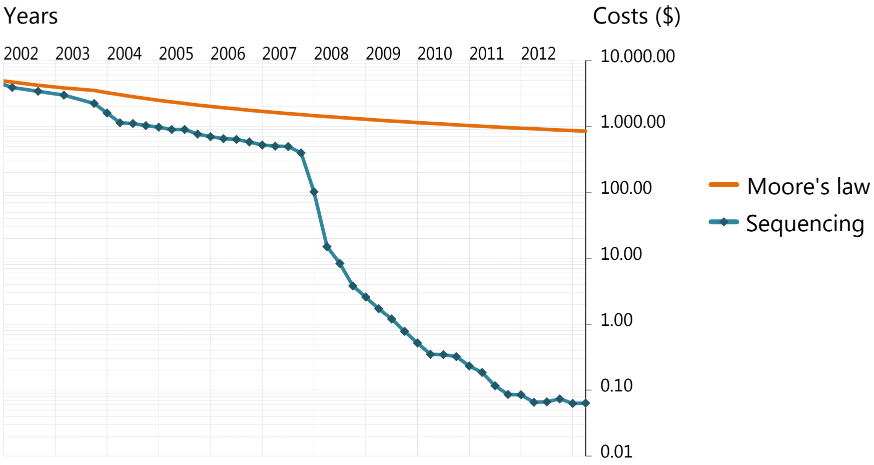


Figure 1.4: Comparison of costs associated with DNA sequencing (in megabase of DNA sequence, log-scale) and Moore's law. Data published by the NHGRI Genome Sequencing Program [10].

The 454 Sequencer (Roche Life Science), the Solexa technology (Illumina) or the SOLiD platform (Life Technologies), were the first NGS platforms developed as a commercial product, and their latest versions are still widely used for most of NGS applications. Despite the recent and rapid spread of NGS technologies, a new generation of single-molecule sequencing technologies is now emerging. Unlike NGS sequencers, single-molecule sequencing technologies directly interrogate single molecules of DNA or RNA, resulting in longer sequences, faster sequencing process and reduced bias due to PCR amplification, which is not needed anymore [11]. Moreover, other technologies, which exploit innovative techniques with respect to NGS approaches, are emerging and complementing single-molecule sequencers. Despite their different features, all togethers these new methodologies promise such an advance with respect to NGS technologies, to be called "Third-Generation Sequencers", while the above mentioned NGS platforms are now seen as representatives of the "Second-Generation". However, further developments are needed to fully affirm these new technologies in genomic and transcriptomics studies, while conventional Sanger sequencing together with Second-Generation platforms are *de facto* the standard for most applications of DNA sequencing. In the following, a review of the main technologies for each type of sequencers is presented.

Sanger sequencing

Sanger sequencing, also called "chain termination method" [6], was the first sequencing approach developed and, as the majority of sequencing protocols, consists in two steps: amplification, to obtain more copies of the DNA of interest, and sequencing. Since it is not possible to directly and continuously read a whole genome from its first base to the last one, the input DNA is randomly shredded into smaller pieces to ensure a uniform representation of all genomic regions. These fragments are then inserted and cloned into a bacterial plasmid to perform amplification, namely to generate multiple identical copies. The modified plasmid, called "recombinant DNA" because it carries genetic material from multiple sources, is then inserted into a host organism, such as *E. coli*. The replication of *E. coli* generates a colony, in which each element carries one ore more copies of the recombinant DNA molecule. Finally, all the clonal copies of the modified plasmid can be picked and the initial DNA fragment, present in multiple copies (also called "amplicons"), can be extracted from plasmids using restriction enzymes. Multiple spatially isolated bacterial colonies have to be created in order to amplify every DNA sequence separately. This *in vivo* technique guarantees a low amplification error, but is slow and barely automated. Once the amplicons are ready, they are sequenced with a technique that employs the DNA polymerase enzyme. In the presence of deoxynucleotides (dNTPs), i.e. the basic "blocks" that constitute DNA, DNA polymerase can synthesize the complementary strand of a single-stranded DNA molecule used as a template. The complementary strand is synthetized in 5'-to-3' direction, by concatenating a new dNTP to the 3'-hydroxyl group of the previous one (Figure 1.5). Indeed, two additional steps are needed to allow the DNA polymerase to initiate DNA synthesis:

- DNA denaturation: DNA polymerase cannot start the synthesis unless DNA is denatured, namely the two strands are separated.

- Primer annealing: DNA polymerases cannot initiate synthesis of a completely new strand, but can only extend an existing DNA strand. To begin synthesis, a short fragment of DNA, called *primer*, must be created and paired with the template DNA strand.

Thus, in the chain-termination method, the original amplicons are first thermally denatured and primer annealed. Then, they are mixed with DNA polymerases and dNTPs. In addition, dideoxy-nucleotides (ddNTP) are also added, despite in quite lower concentrations. ddNTPs (Figure 1.5) are *chain-terminating inhibitors* of DNA polymerase, because they do not have the 3'-hydroxyl group, so once they are added by a DNA polymerase to a newly synthetized DNA strain, no further nucleotides can be added by creating a bond between the 5' end of this new nucleotide and the 3' of the previous ddNTP.

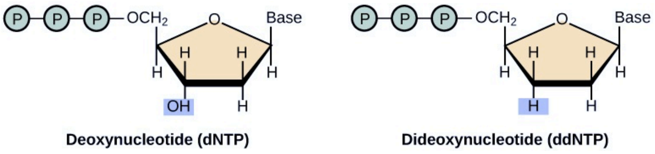


Figure 1.5: Chemical structure of deoxy-nucleotides (dNTP) and dideoxy-nucleotides (ddNTP). ddNTPs do not have the 3'-hydroxyl group.

After DNA denaturation and primer annealing, DNA polymerase can start the elongation of the complementary strand and continue as long as it concatenates dNTPs. Instead, when a ddNTP is incorporated, the elongation terminates. The ddNTPs are also fluorescently labeled, so that the label on the ddNTP terminating the last synthetized fragment univocally determines the nucleotide of the last position read (i.e. A, C, G or T). The dNTP/ddNTP mixture causes random, non-reversible termination of strand extension, creating several molecules complementary to the original template, but having different lengths. These fragments are then denatured and sorted by molecular weight, which corresponds to sequence length and consequently to the position of the last position read. The labels attached to the terminating ddNTPs are identified sequentially, considering fragments of increasing length, and the complemented bases identify the sequence of the template DNA. A schematic representation of this process is given in Figure 1.6. Originally, sorting by molecular weight was performed using gel electrophoresis, but nowadays it is performed through capillary electrophoresis. In both techniques, an electric field is applied so that the DNA fragments, which are negatively charged, migrate from one end to the other. The longer a fragment is, the more it is slowed down by the gel and the later it reaches the opposite end of the electrophoresis apparatus. As fragments of increasing lengths exit the capillary, a laser excites the fluorescent labels, and the four-colors emission spectra are identified by a detector and represented in a sequencing "trace" of fluorescent emissions. Finally, an algorithm translates these traces into DNA sequences, called "reads". During this process of "base-calling", a quality score is also assigned to each read base, to reflect the probability for that base of being called correctly. In particular, the quality Q_p of a read base p is measured using the Phred score:

$$Q_p = -10 \log_{10} E_p, \tag{1.1}$$

where, E_p represents the probability for that base of being wrong. In modern 96-capillary sequencers, up to 96 sequences can be sequenced in parallel in independent capillaries. This system is still used in current research and, due to its high accuracy in base calling, it is considered the gold-standard for genome sequencing. However, the high costs and time related to Sanger sequencing prevent its use for some applications in which a higher throughput is necessary (e.g. whole-genome or whole-exome variant calling).

Next-Generation sequencing

Although Next-Generation sequencers (NGS) implement quite different solutions, all of them share a common scheme: first, DNA fragments are amplified with different versions of the Polymerase Chain Reaction (PCR) technique, so to create localized clusters of amplicons bound to a substrate or array; then, these millions of clonally clustered amplicons are sequenced in parallel, alternating cycles of DNA synthesis with imaging-based data acquisition of the whole array. This framework ensures several advantages over Sanger sequencing [12]:

- *In vitro* PCR amplification overcomes several bottlenecks that limit the parallelism of Sanger sequencing, such as transformation of *E. coli* and colony picking;

- Array-based sequencing enables a much higher degree of parallelism than conventional 96-capillary sequencing: tens to hundreds of millions reads can be sequenced for each instrument run;

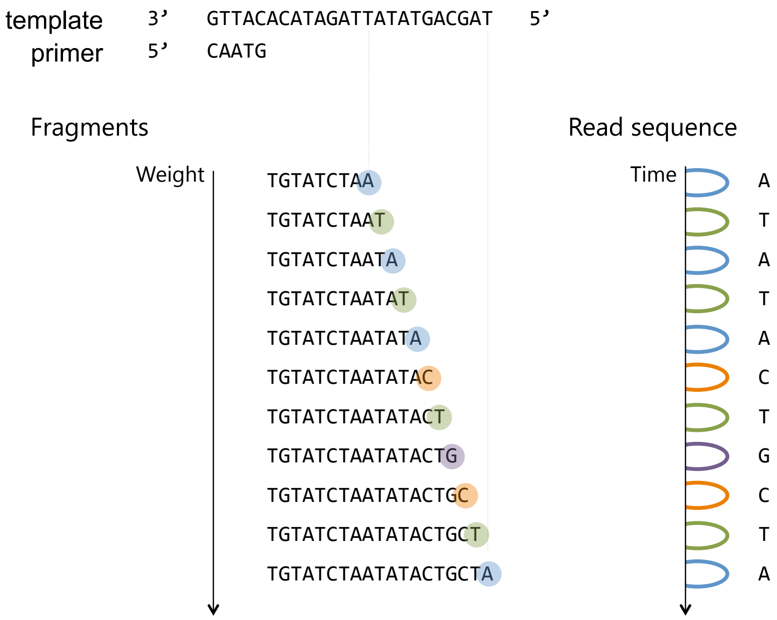


Figure 1.6: Determination of read sequence in Sanger sequencing through electrophoresis: fragment sorting and label identification.

- Since all fragments are immobilized on the same array, a single reagent volume can be used for the whole volume so to drastically reduce costs.

On the other hand, NGS data are characterized by some drawbacks with respect to Sanger sequencing, which include shorter read length and lower base-call accuracy (Table 1.1). However, NGS technologies are experiencing fast technical advances which are leading to more and more apparent reduction of costs and simultaneous rise of read quality and length (compare for example [12] and [9]).

Table 1.1: Comparison of the features of the main NGS technologies and Sanger sequencing (data and description taken from [9]). "Mb" indicates megabases.

Technology	Throughput [Mb/day]	Length [nt]	Quality	Costs [USD/Mb]
Sanger	6	800	10^{-4}-10^{-5}	500
454	750	400	10^{-3}-10^{-4}	20
Illumina	5000	100	10^{-2}-10^{-3}	0.5
SOLiD	5000	50	10^{-2}-10^{-3}	0.5

As it can be noticed from Table 1.1, read length is a limitation factor of current NGS technology with respect to Sanger sequencing. Indeed, sequenced reads have to be later assembled like the pieces of a jigsaw puzzle in order to reconstruct the input genome (or mapped on a reference genome, if available). A reduced read length creates non-trivial issues, especially in the presence of repeated regions (see section 2.2), which cannot be distinguished one from the other if reads do not span their whole sequence. Nevertheless, all NGS platform are experiencing a fast improvement leading to longer and longer reads, with the 454 sequenced promising “Sanger’s-like” read length [13].

A technical advance to ease the challenge of genome assembly was represented by the introduction of the so-called “paired-end sequencing”, now available for all the three NGS technologies above described. Paired-end sequencing allows sequencing both ends of DNA fragments. The information about the expected distance of the reads sequenced from these two ends, estimated from the distribution of DNA fragment lengths, can be exploited to increase mapping or assembly accuracy. They are particularly useful to solve repeats, since they cover a longer genome region, possibly extending into univocally determined regions flanking the repeated ones. Read pairs can be obtained through two different techniques: paired-end or mate-pair sequencing. Although the two approaches have been often confounded in the literature, they refer to different protocols aimed at obtaining read pairs having different distances [7]. In paired-end sequencing, adapter sequences with different priming sites are attached to the ends of a DNA fragment shorter than 1 kb (the exact length depends on the specific protocol). The sequencing process, performed accordingly to the technology adopted (described in the following), is run twice, exploiting the two different adapters, giving rise to one read for each fragment end. In mate-pair sequencing, the fragments are longer than 1 kb (up to 20 kb), and instead of ligating one adapters at each end, the fragment is circularized around a single adapter, with both fragment ends ligated to the adapter ends [14] (Figure 1.7). These circular molecules are then shredded to produce shorter fragments. The sequences containing the two ends of the original fragment and the adapter, which is biotinylated, can be captured using streptavidin magnetic beads. The remaining fragments are instead washed away. The two mates can then be sequenced together in a single run and recognized thanks to the known sequence of the adapter between them. When the aim of a sequencing study is *de novo* genome assembly, mate-pair and paired-end reads can be used together to leverage on their different features to reconstruct the most problematic regions [15]. Moreover, owing the ultra-high throughput characterizing NGS technologies, different samples can be sequenced through *multiplexing*. DNA fragments to be sequenced are attached to short nucleotidic molecules, called “barcodes”, with known sequences. Then, barcoded samples are pooled in a single library and sequenced together. The sequenced barcodes will then allow to computationally separate reads coming from different samples.

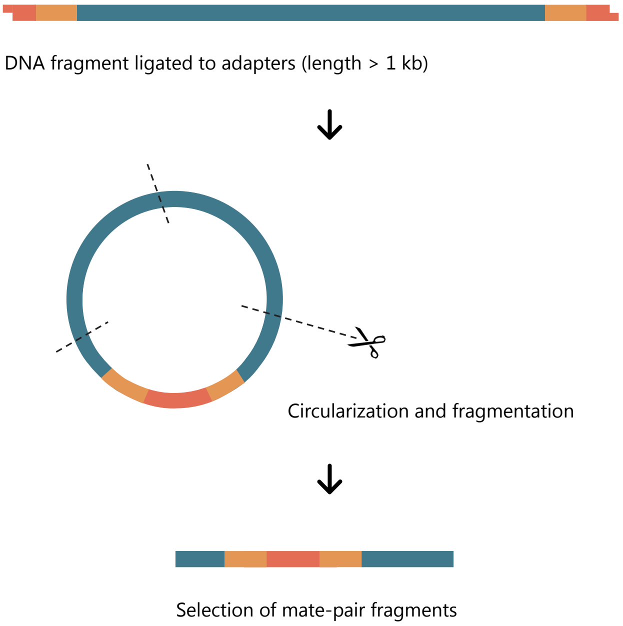

DNA fragment ligated to adapters (length > 1 kb)
Circularization and fragmentation
Selection of mate-pair fragments

Figure 1.7: Mate-pair sequencing: ligation of adapters to DNA fragments, circularization, fragmentation and selection of fragments containing adapters. The final fragments, consisting of two “mate” sequences ligated by an adapter, are subjected to a single sequencing run.

Although sequencing prices base have fallen dramatically with the advent of NGS, high-throughput sequencing still has high acquisition, running and maintenance costs, which are not considered in Table 1.1 [9]. Moreover, once the sequencing is performed, consistent investments in data management and analysis are needed. Thus, smaller research groups may still find prohibitive the costs of the infrastructure needed for storing, handling and analyzing gigabytes or terabytes of data. Even for larger centers, these issues require constant investment in computational infrastructures to keep pace with all the data generated by these technologies and to transform them into biologically meaningful results. However, the impact of NGS technology, with thousand of platforms sold worldwide, has marked a real revolution in the field of genomics and incentivated the understanding of myriads of organisms and microorganisms [16, 17, 18].

454 sequencing

The 454 Sequencer was the first NGS platform released on the market [19] by 454 Life Sciences and later acquired by Roche Diagnostic [20]. In the 454 system, amplification is performed through *emulsion PCR* [21] and followed by *pyrosequencing* [22]. To carry out emulsion PCR, single-stranded DNA fragments are ligated to adapters and bound to 28 μm beads, one fragment per bead. Beads are included in a water-in-oil emulsion so to enclose individual beads in amplification microreactors (Figure 1.8a). Amplification through emulsion PCR creates millions of clonally copies of each library fragment. Once the emulsion is broken, the amplicons remain bound to the same bead of their original template (from which they originated). The beads are then loaded onto a PicoTiterPlate, a flat solid support, containing millions of wells (Figure 1.8b). Each well contains only one amplicon bead and several smaller beads carrying immobilized enzymes required for pyrosequencing (i.e. ATP sulfurylase and luciferase). Pyrosequencing [22] is a *sequencing-by-synthesis* process, in which one class of dNTP (i.e. dATP, dCTP, dGTP or dTTP) at a time is washed over the PicoTiterPlate and incorporated by DNA polymerases in correspondence of complementary bases of the templates (Figure 1.8). When a dNTP is incorporated, one phosphate per nucleotide is released and converted to ATP by ATP sulfurylase. The ATP, in turn, drives a light reaction catalyzed by luciferase, so that each incorporation event is accompanied by a burst of light. At each cycle, several dNTPs of the same species can be incorporated and the light intensity measured during incorporation is proportional to the current homopolymer length (i.e. the number of equal bases read) (Figure 1.8f). One side of the PicoTiterPlate is mounted on a flow cell that wash one type of dNTP per cycle, while on the other side a CCD camera detects light bursts across all the array positions where one or more dNTPs have been incorporated (Figure 1.8).

Since they are not labeled, dNTPs are added in a pre-determined key sequence (e.g. A, G, C, T, A, G, C, T, . . .) and the pattern of detected incorporations (with their relative intensity and array coordinates) determines the sequence of the template represented by each bead. At each cycle, dNTPs are added through the flow cell, an image of the whole array is taken, and the dNTPs left are washed away to start a new cycle. The average substitution rate in 454 sequencing is higher than that of Sanger sequencing, but it is the lowest among all NGS platforms (Table 1.1). Most of the errors observed for this technology are small insertions or deletions (also called *indels*), arising from inaccurate estimate of homopolimers length. Similarly to Sanger sequencing, the error rate increases with the position within the read, namely with the number of cycles performed, due to a reduction of enzymes efficiency [9]. Moreover, when not all molecules are correctly extended in every cycle, the process loses its synchrony (*phasing*) and results in an "echo" of the preceding cycles over the following. With respect to the other NGS technologies, the 454 platform is characterized by a lower throughput, but produces longer reads (Table 1.1).

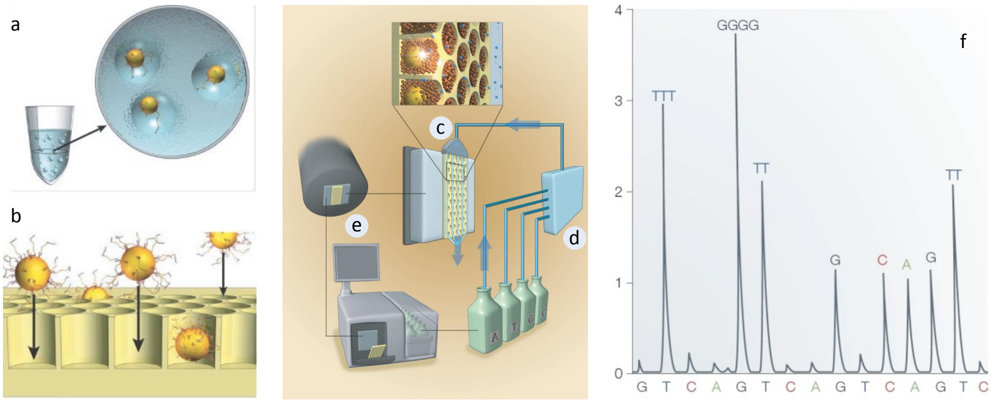


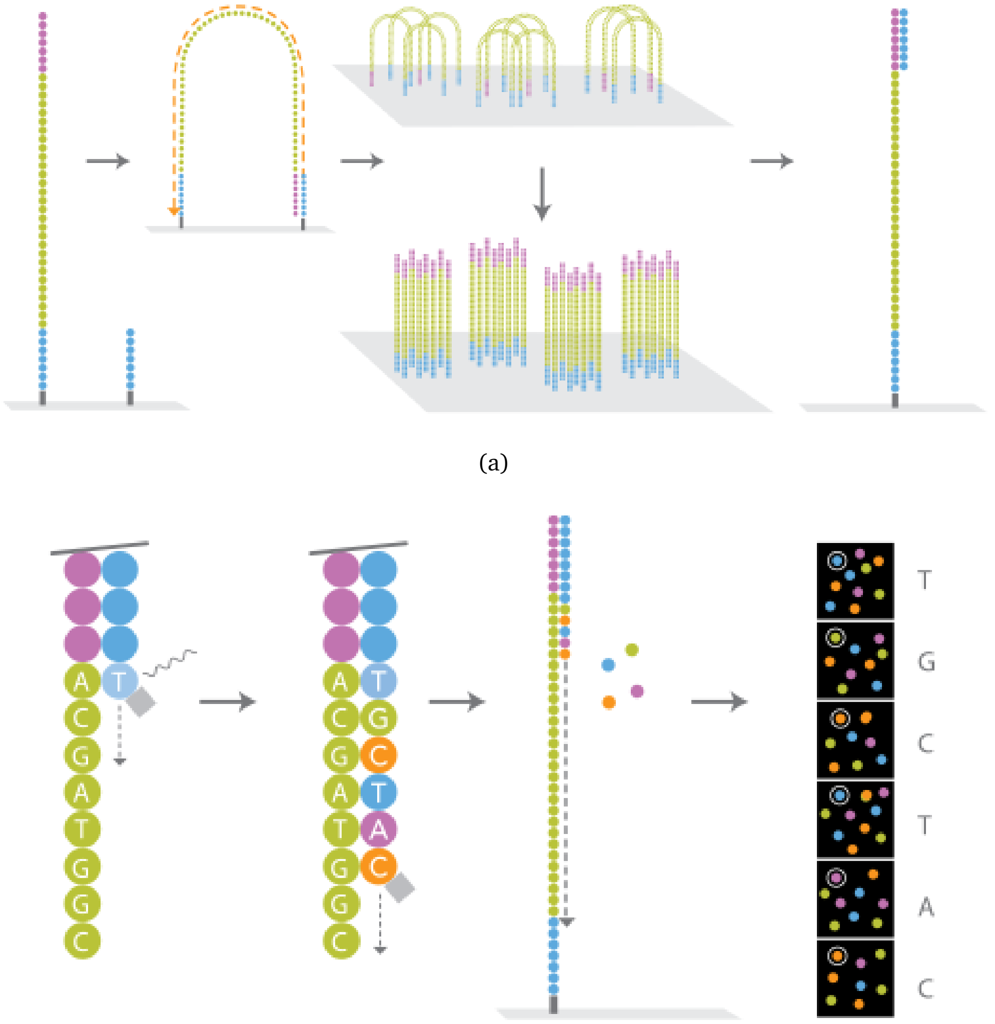
Figure 1.8: The 454 sequencing system: emulsion PCR (a) and deposition of amplicons-carrying beads into PicoTiterPlate's wells (b,c); flow cell (d) and CCD camera (e) mounted on the two sides of the PicoTiterPlate (c); pyrogram of light intensities due to the base incorporated at each cycle for a single well (e). Images taken from [20, 19, 23].

Illumina's technology

From the first "Solexa" to the latest "HiSeq" platform, Illumina technology [24, 25, 26] has always been among the most used solutions for NGS sequencing [27, 28].

Similarly to the 454, adapters are attached to single-strand DNA fragments, so to bound them to a solid substrate and start amplification. In this approach though, a flat array is used. The surface of the array is covered by flexible adapters, which are

Figure 1.9: Illumina sequencing: cluster generation through bridge PCR (a) and sequencing with reversible terminators (b) (images taken from [24]).



(a)

(b)

complementary to the ones linked to DNA fragments. Fragments are thus immobilized on this solid surface and amplification through *bridge PCR* is initiated. At each cycle, fragments are bent to form a “bridge” in which both adapters are tethered to the surface. Fragments are amplified, and both copies of DNA fragment are then separated, letting one adapter detaching from the array. In the next cycles, fragments are again bent and amplified; all fragments remain tethered to the surface, such that all amplicons arising from any original template molecule (all having the same sequence) remain immobilized and clustered to a single physical location onto the array (Figure 1.9a). Several millions of clusters, with about 1000 clonal amplicons each, can be amplified in different locations of the array. Each cluster represents a single template fragment.

The array is further separated into eight “lanes”, so that eight different libraries can be constructed, and later sequenced independently and in parallel, during the same run. After cluster generation, amplicons are single-stranded (*linearization*) and hybridized to a primer to start sequencing-by-synthesis with DNA polymerase. Differently from pyrosequencing, at each sequencing cycle, all dNTPs are washed together on the array. For this reason, dNTP are modified in two ways:

- They have a chemically cleavable moiety (*reversible terminator*) at their 3’ end that prevents the concatenation of multiple nucleotides, so that only one dNTP per cycle is incorporated;

- They are marked with four different fluorescent labels, also chemically cleavable, which correspond to the identity of each nucleotide.

At each cycle, after single-base extension (i.e. single-dNTP incorporation), images of the whole array in four channels are acquired in order to identify the base added in each cluster (Figure 1.9b). After image acquisition, the 3’ ends of the newly synthetized strands are made available again by chemical cleavage of reversible terminators and labels. Unlikely 454 sequencing, the process is synchronous, because the same position within the fragment is interrogated at the same time for all clusters. The final read sequences all have the same length, which corresponds to the number of sequencing cycles performed. Before bridge PCR, the Illumina library preparation includes several *in vitro* amplification steps, cause an error rate increment with respect to 454 error model (Table 1.1). As in the case of the other platforms, the error rate increases with read position. In particular, the problem of phasing affecting this technology can be worsened by errors due to reversible terminators. Indeed, if some dNTPs fail to be correctly terminated, incorporation of multiple nucleotides in the same cycle occurs, resulting in wrong estimates of homopolimers length [9].

SOLiD technology

The SOLiD technology, first described in [29, 30], was developed by Applied Biosystems and later bought by Life Technologies [31]. It also performs amplification trough emulsion PCR but exploits a complete different sequencing technique. In this approach, called *sequencing-by-ligation*, the DNA ligase enzyme is used in place of DNA polymerase for sequence extension. Unlikely 454 technology, the beads used for amplification are paramagnetic. Once the amplification step is finished, each paramagnetic bead carrying millions of amplicons is immobilized to a solid flat substrate to generate a dense and disordered array. After primer annealing, a mixture of octamers (i.e. sequences 8nt long) are washed on the array, and the different octamers compete for binding the template. If an octamer is complementary to the template sequence, it is ligated by the ligase enzyme to the primer end, starting elongation. In the first two bases, the octamers have exactly all the 16 different combinations of the four nucleotides, while the remaining sequence is degenerated (i.e. can have diverse nucleotidic sequences allowing pairing with all possible template sequences). In each octamer, the identity of the first two nucleotides is encoded by a fluorescent label attached to its 5' end. Once the octamer is ligated, the fluorescent label is read and then cleaved out, together with the last three bases of the octamer. Then, a new cycle is initiated: a new complementary octamer is ligated to the previous one, continuing elongation of the complementary strain, the fluorescent label is read and the last three bases are removed. After several cycles carried out with this scheme (about ten [9]), the read bases are in positions 1, 2, 6, 7 and so on (i.e. two bases are read and three are skipped). In order to read the remaining bases, the newly synthetized strand, together with the primer, is washed away and a new primer, shifted one base backward, is annealed. At this point, the sequencing cycles are started again and the bases in positions 2, 3, 7, 8, etc., are read. The whole process is carried out several times, using shifted primers, such that the the whole template is read. However, the color code employed does not allow the univocal identification of each base, because each of the four labels encodes four of the possible dinucleotides (e.g. the red label encodes AT, CG, GC and TA dinucleotides, Figure 1.10).

Even though each base is read twice, this does not suffice to decode the original sequence. For instance, the T at the 6th position in Figure 1.10, is read together with the 5th (step E) and 7th (step D) base, resulting in a green and blue label, respectively. However, this labels can be given by several combinations (e.g. CA-AA, GT-TT, etc.) and cannot be uniquely identified. In order to decipher the encoding, a primer shifted one base back is used, so to read also the last adapter base, which is known. For instance, in Figure 1.10, the last position of the primer, which is a T, is read and associated to

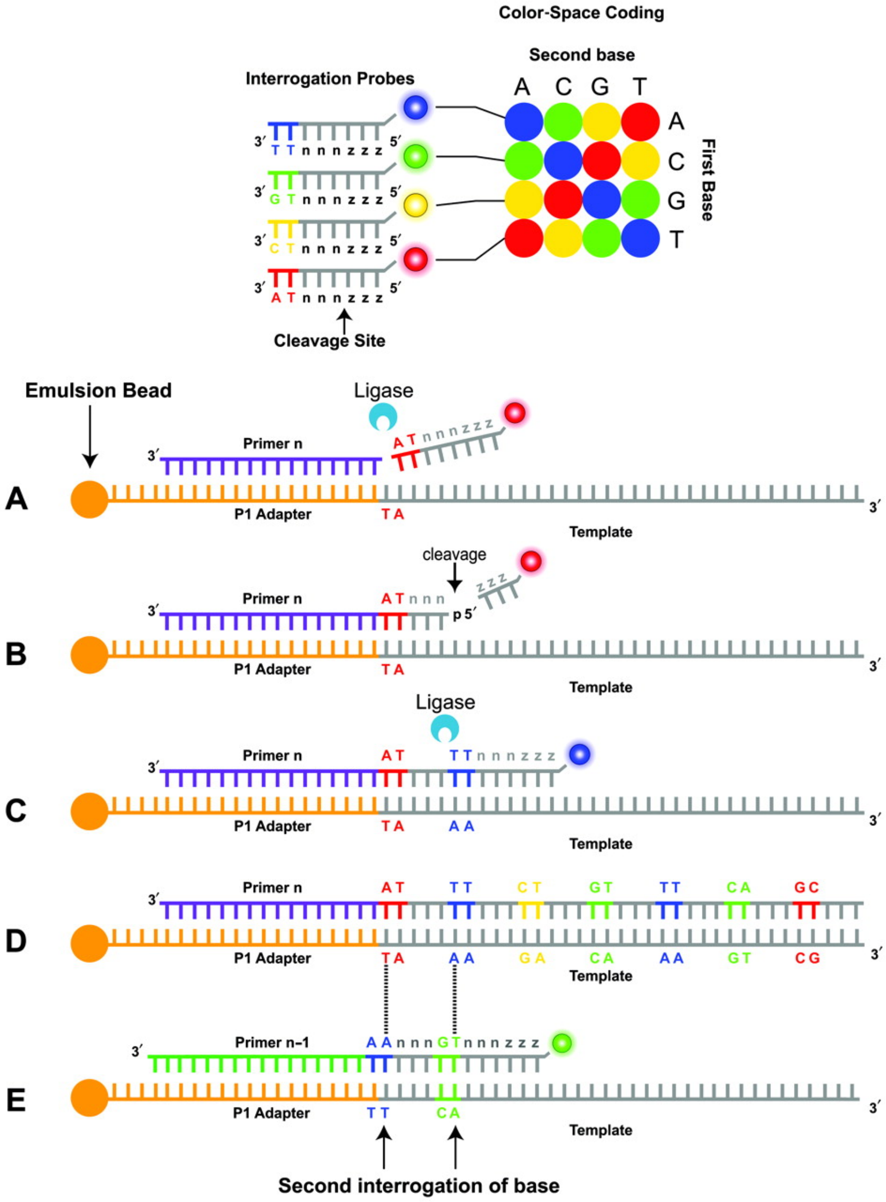


Figure 1.10: SOLiD color-space coding and sequencing by ligation (image taken from [32]); description in the main text.

a blue label (step E), univocally identifying the following base as a T. Then, the label obtained for the following dinucleotide, necessarily indicates its identity: a T, read at the previous step, and an A, because there are no other possible combinations for the red label. In summary, the SOLiD two-base encoding (also called “color-space coding”) can be decoded in the presence of a known base, which is obtained by sequencing the last base of the adapter sequence.

Similarly to Illumina’s technology, SOLiD sequencing enables very-high throughput, but results in shorter reads (Table 1.1). Phasing is not a major issue, but errors can occur if incomplete cleavage is performed [9]. Indels are not as frequent as in 454 reads, but higher substitution rates are present.

Third-Generation sequencing and single-molecule sequencing

Despite the lacking of a consensus definition of what constitute a Third-Generation Sequencing technology, single-molecule sequencing probably represents the most noteworthy breakthrough, leading this second revolution [11]. However, a series of other interesting solutions are complementing single-molecule sequencing technologies.

Among these, Ion Torrent, now property of Life Technologies [31], still employs DNA polymerase, but eliminates the image acquisition step by directly measuring pH changes due to nucleotide incorporation using the proprietary Ion Chip technology [33]. The Ion Chip is a silicon chip designed to detect pH changes within single wells, as nucleotides are incorporated during the sequencing process (similarly to 454 sequencing), since each incorporation is accompanied by the release of a hydrogen ion H^+. The upper side of the Ion Chip functions as a microfluidic cell, delivering reagents needed for the sequencing reaction, while the lower side directly interfaces with a H^+ ion sensor. The sensor measures voltage changes proportional to pH changes; the number of released hydrogen ions is then in turn translated into a measure of the number of nucleotides incorporated. As for 454 sequencer, nucleotides are not labeled, and have to be added in a pre-defined order, one class per cycle. The elimination of the image acquisition step allows reducing time and costs, but read length and throughput remain comparable to that of NGS [11].

Heliscope (Helicos) is another promising technology and the first commercially available instrument for single-molecule sequencing [34, 35]. This technique allows directly scanning DNA fragments without performing PCR amplification. DNA fragments are bound to a solid surface and sequenced using a modified DNA polymerase and special fluorescently labeled nucleotides, called “virtual terminators”, which allow step-wise sequencing. However, since halting is still required in this process, the time needed to sequence a single template base is still high and the final read length limited [11]. Despite the *single-molecule* nature of this approach, the limited improvements and the high costs of the instrument have narrowed the market of this platform to only four machines sold in the first year [9].

The most promising class of single-molecule sequencing technologies is represented by the solutions that, unlikely NGS “scan-and wash” approaches, do not halt the sequencing reaction after each base incorporation, allowing greater sequencing rates, throughput and read lengths [11]. These techniques can be divided into three subclasses, considering the main principle or technology employed:

- Single-molecule real-time (SMRT) monitoring of long DNA or RNA molecules synthetized by polymerase or reverse transcriptase enzymes.
- Nanopore sequencing. A nanopore is immersed in a conducting fluid and subjected to a potential, such that it is crossed by an electric current, which is very sensitive to the size and shape of the nanopore. Single nucleotides from a DNA or RNA molecule can be directly identified as they pass through the nanopore, thanks to the changes in the current intensity they induce.
- Direct imaging of single DNA molecules using advanced microscopy technologies.

Reviewing the whole panel of the emerging technologies is out of the scope of the present thesis, which is focused on the first NGS platforms, but can be found in specific studies (e.g [11]). However, since SMRT sequencing is one the most appealing and mature representative of the Third-Generation sequencing technologies, with an interesting application to gene expression studies, it is briefly described in the following.

Single-molecule real-time (SMRT) sequencing, developed by Pacific Biosciences [36], was the first approach enabling the direct observation of a DNA polymerase synthesizing a strand of DNA [37, 38]. This technique can thus exploit the speed of this enzyme, without halting the process as happens instead in the “scan-and wash” techniques [11]. Given that a single DNA polymerase molecule has a diameter in the order of 10 nm, one of the greatest issues for SMRT is the definition of an observation volume small enough to achieve a sufficient signal-to-noise ratio to perform base calling as nucleotides are incorporated. This problem is solved using the zero-mode waveguide (ZMW) technology (Figure 1.11) [38]. A ZMW is a hole with tens of nanometers in diameter, through a 100 nm metal film deposited on a glass substrate. Visible laser light, which has a wavelength of about 600 nm, cannot traverse the ZMW but exponentially decays along the ZMW. Therefore, by shining laser illumination up through the glass into the ZMW,

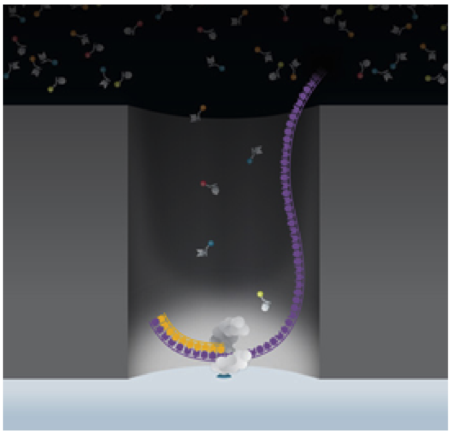


Figure 1.11: Single-molecule real-time sequencing in the ZMW chamber of the Pacific Biosciences system (image taken from [36]).

only the bottom 30 nm of the ZMW are illuminated. A single DNA polymerase molecule is anchored to the glass surface at the bottom of the ZMW through biotin/streptavidin interaction. Labeled nucleotides are flooded above the ZMW array diffuse down into the ZMW and then back through the exit of the hole. As laser light cannot traverse the holes to excite the fluorescent labels, the labeled nucleotides above the ZMW array do not contribute to the measured signals. Only when nucleotides diffuse through the bottom 30 nm of the ZMW, they are excited by the laser. Among these nucleotides, the one complementary to the template being sequenced is detected by the polymerase and it is incorporated into the growing DNA strand. This process takes milliseconds, a time three orders of magnitude longer than simple diffusion, enabling the detection of higher signal intensity for incorporated versus unincorporated nucleotides (i.e. high signal-to-noise ratio). While held by the polymerase, the fluorescent label emits a colored light that corresponds to base identity and is thus detected by the instrument. After incorporation, the signal immediately returns to the baseline and the process repeats, with the DNA polymerase continuing to incorporate multiple nucleotides per second.

The first commercial SMRT platform consisted of an array of about 75 000 ZMWs, thus enabling the detection of about 75 000 single-molecule sequencing reactions in parallel. However, since DNA polymerases and DNA templates are delivered to ZMWs via a random diffusion process, only about a third of the ZMWs of the array are active for a given run [11]. The SMRT sequencing platform requires reduced amounts of reagents and, most of all, the "scan-and-wash" step is avoided, resulting in a dramatic reduction of run time (minutes as opposed to days) [37]. Moreover, PCR amplification is not needed, eliminating systematic amplification biases affecting NGS. Leveraging on the speed and processivity of the DNA polymerase, SMRT sequencing strongly reduces time and increases read lengths, producing read with an average length of 1000 bp and a maximum length of about 10000 bp. Moreover, the possibility to observe the activity of the polymerase enzyme in real time allows the investigation of changes in the dynamics and timing of enzymatic incorporation (i.e. kinetics), which are in turn related to chemical modifications, such as methylation [11]. Beyond DNA sequencing, the flexibility of SMRT sequencing is expected to enable new applications that are still not achievable with current sequencers. For example, using RNA-dependent polymerases and reverse transcriptase, direct RNA-sequencing can be performed.

Despite the many potential advantages of SMRT sequencing, the reduced read throughput and quality still hamper large scale sequencing projects [27, 11]. Moreover, since SMRT sequencing data are different from NGS data, further research is needed to investigate new error models and develop algorithms capable of exploiting the strengths of SMRT reads while minimizing bias. Despite the great expectation about Third-Generation sequencing technologies, further efforts are needed to demonstrate that these sophisticated solutions can be translated into a true advance over NGS, with evident impact on genomics and transcriptomics studies [11].

1.3 RNA-seq: measuring gene expression through Next-Generation Sequencing

The transcriptome is the whole set of mRNAs transcribed from the genes of a cell (see section 1.1). Their relative abundances reflect the level of expression of these genes for a specific developmental stage or physiological condition. Although mRNAs are not the final products of the transcription-translation process, the estimation of transcript levels through gene expression profiling unveils important aspects about the cell state under investigation. More interestingly, differential analysis of gene expression enables the comparison of gene expression profiles of different tissues and conditions, such as treated versus untreated cells or cancer versus normal tissues, to identify the genes that may play a role in the determination of the phenotypic differences.

Hybridization-based approaches such as microarrays, have been the most used solutions for gene expression profiling and differential expression (DE) analysis, thanks to their high throughput and relatively low costs [39]. This type of consists in a set of probes, whose sequences represent particular regions of the genes to be monitored. The probes are bound to a solid array in specific and known coordinates and are present in multiple copies. The sample under investigation is washed over the array, and the transcripts are free to hybridize to the probes with a complementary sequence. A fluorescent is used to label the transcripts, so that image acquisition of the whole array enables the identification of the expressed genes (identified by array coordinates) and their level of expression (given by the signal intensity derived from the labeled transcripts hybridized to the probes). Although widely used in quantitative transcriptomics, these techniques have several limitations [39]:

- Reliance upon prior knowledge about genome sequence to enable probe design;
- High background levels due to cross-hybridization (i.e. imperfect hybridization between quasi-complementary sequences);
- Limited dynamic range of detection, due to both background noise, which hampers low-expressed genes detection, and saturation of signals that happen when the transcripts to be assayed are numerically greater than the complementary probes available;
- Need for sophisticated normalizations to compare data from different arrays.

The advent of sequencing (see section 1.2) brought new sequence-based techniques for gene expression profiling, such as expressed sequence tags (EST) [40, 41], serial analysis of gene expression (SAGE) [42, 43] and massively parallel signature sequencing (MPSS) [44], providing the first *digital* measures of gene expression levels, as opposed to the *analog* signals of microarrays. However, most of these techniques were based on expensive Sanger sequencing, and allowed to sequence only short portions of gene sequences, resulting in the impossibility to assign most of the produced data to a unique gene [39, 45].

A major breakthrough followed the employment of Next-Generation sequencing technologies in transcriptomics, through a methodology called “RNA-Seq” [46], which allows to determine transcript sequences and quantify their abundance at the same time. The standard workflow of an RNA-seq experiment is described in the following. The population of RNAs in the sample of interest are initially fragmented and reverse-transcribed into complementary DNAs (cDNAs), to be suited for deep DNA-sequencing through NGS. In the first protocols, reverse-transcription was performed before fragmentation, but it has been later noted that RNA fragmentation (as opposed to of cDNA fragmentation) ensures less biased estimates of gene expression levels [39]. The obtained cDNAs are then amplified and subjected to NGS. In principle, all NGS technologies can be used for RNA sequencing, even though their features make them more suited for certain applications (e.g. longer 454 reads may facilitate *de novo* transcriptome assembly). The Illumina technology is now the most commonly used NGS platform for RNA-seq [47]. The millions of short reads generated can be then mapped on a reference genome and the number of reads aligned to each gene give a digital measure of gene expression levels in the sample under investigation.

Although RNA-Seq is still a methodology under active development, is now widely used in place of microarrays for measuring and comparing gene transcription levels because it offers several key advantages over the previous technologies [48, 39]:

- It is not limited to the detection of transcripts corresponding to well-annotated genomic sequences, but can be used to sequence non-model organisms or to perform novel transcripts discovery;
- It does not have an upper limit for quantification (i.e. the saturation problem of microarrays signals), thus ensuring a large dynamic range of expression levels over which transcripts can be detected;
- It is characterized by high levels of reproducibility for technical replicates [49, 45, 50];
- The sequencing technology ensures high resolution, such that transcript sequences can be read at single-base level.

The latter property is undoubtedly the most promising one: while microarrays can only assay transcripts corresponding to probes, RNA-seq can, in principle, investigate at a finer level of detail all the transcripts present in a sample, characterizing their sequences and quantifying their abundances at the same time. The possibility of sequencing transcriptomes at single-base resolution has rapidly opened a wide frontier of applications in transcriptomics research, such as: transcriptome profiling of non-model organisms [51, 52], novel transcripts discovery [53], investigation of gene transcriptional structure [54], splicing [55] and RNA editing [56, 57], quantification of allele-specific gene expression [58], and “dual RNA-seq” of pathogen and host [59].

Despite all these newsworthy features and apparently easy scheme of data analysis, RNA-seq studies produce large and complex data sets, whose interpretation is not straightforward [60, 61]. In a recent article published on Nature Methods [61], I. Korf compares *de novo* transcriptome reconstruction to the challenging act of reassembling magazine articles after they have been shredded (Figure 1.12).

Nevertheless, in a presence of a well-annotated reference genome or transcriptome, the analysis scheme can be slightly simplified. For instance, if the aim of an RNA-seq study is detecting gene differential expression (DE), a basic data processing pipeline can

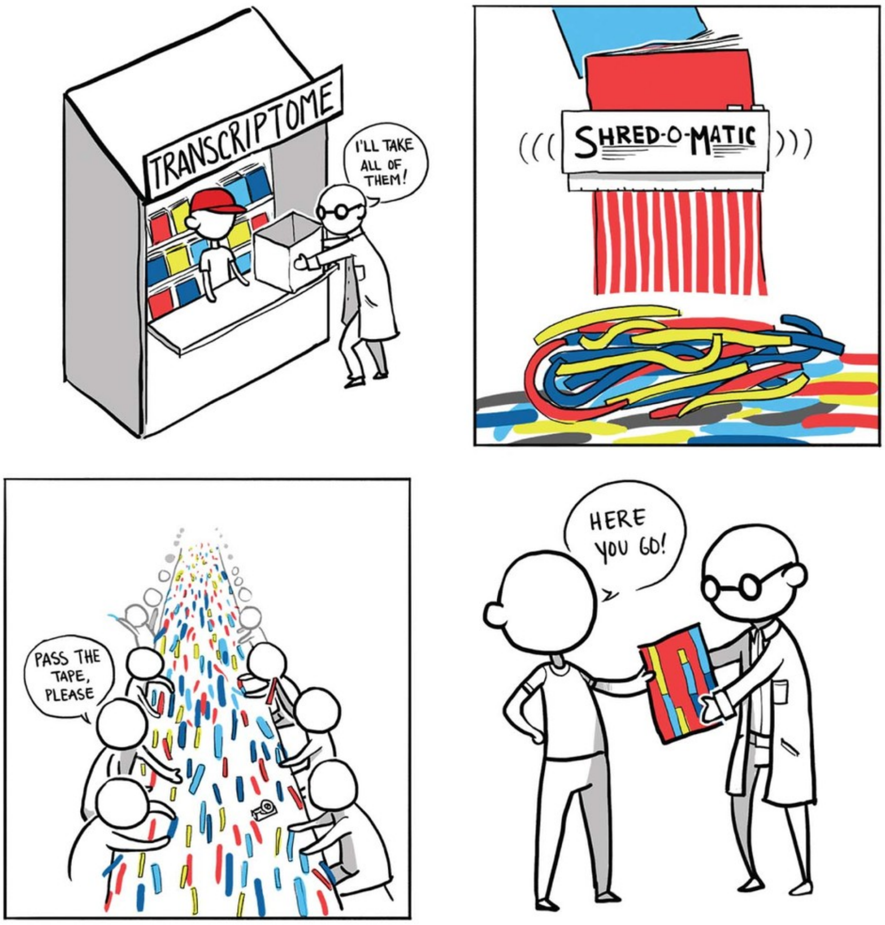


Figure 1.12: According to I. Korf, transcriptome reconstruction is as challenging as reassembling magazine articles after they have been through a paper shredder (story board design by A. Yu) [61].

be outlined as in [60]: read mapping, counts computation, counts normalization and differential expression analysis.

The first analysis step is ***read mapping***: reads are aligned to a reference genome (or transcriptome), identifying gene regions whose sequences match read sequences. In reality, the reads are never a perfect representation of the reference, but can contain polymorphisms, structural variants and sequencing errors. In addition, the presence of repeated regions in the reference and the short length of NGS reads complicate the identification of a single (and possibly correct) mapping position. Thus, RNA-seq reads mapping is not a trivial task, and the wide panel of algorithms available still produces suboptimal solutions [61], making it harder to select a specific tool to be integrated in the analysis pipeline.

Once the reads have be assigned to a genomic location, the next task is to summarize them over some coding units, such as exons, transcripts or genes, to estimate their relative expression levels. The simplest approach is that of counting the number of reads overlapping the exonic bases of a gene, but more sophisticated strategies can be employed. The number of reads aligned to a gene gives a digital measure of its expression level and is called ***counts***.

Before comparing counts between different groups or conditions to detect differential gene expression, within- and between-sample ***normalization*** methods can be used to eliminate possible bias. Within-sample normalization allows a fair comparison of expression levels of each gene relative to other genes in the sample, such that the genes that are more likely to be sequenced do not give rise to inflated counts. Differently, between-sample normalization corrects for differences in the library sizes, i.e. the number of sequenced reads. Unlike some within-sample biases, which may cancel out when comparing samples, between-sample differences have to be corrected before DE analysis to allow a correct comparison, so that only changes due to true varitions in genes expression are detected and not differences due to other confounding factors, such as sequencing depth.

Finally, ***differential expression*** analysis is performed adopting a test statistics that selects the genes for which counts, and consequently underlying expression levels, are significantly different between the compared conditions. For instance, when comparing healthy versus diseased tissues, the identification of DE genes may provide new insights over the genetic variables involved in the pathology.

1.4 Aim and structure of the thesis

Despite being already widely used, RNA-seq methodology is still under active development, and both its experimental and computational methods are changing at a fast pace. In particular, there is not a standard and unified computational pipeline for detecting differential expression analysis from RNA-seq data and several methods for performing each analysis phase are available. This thesis is aimed at defining a robust computational pipeline for RNA-seq data analysis, from data pre-processing to differential expression analysis, which can enable stable and reproducible results. In Chapter 2 and Appendix A, we present a literature review of state-of-the-art methods, for every step of data analysis, and a discussion of data models. To integrate and complement the information gathered by the literature, we carried out three assessments to select the best performers among the available methods, and to develop and test novel strategies to be implemented in the computational pipeline. With this purpose, we considered several data sets, real and simulated, which are described in Chapter 3. The assessment presented in Chapter 4 is focused on the investigation of the bias present in count data and on the comparison of state-of-the-art normalization methods. In Chapter 5 we define and present a novel strategy, called *maxcounts*, to compute counts and directly reduce data bias prior to normalization. A comparison with the standard approaches for count computation and normalization is also presented. In Chapter 6, we optimize the mapping strategy to enable correct expression estimation even in the presence of reads mapping in multiple positions. This approach, along with the methods selected and developed in analyses presented in the previous chapters, is integrated in the final processing pipeline, also presented in Chapter 6. This computational framework is applied to a real case study, described in Chapter 7 and Appendix B, aimed at identifying the genes involved in the pathogenesis of spinal muscular atrophy. Since our future work will be directed at optimizing the current pipeline and extending it to the analysis of time-series RNA-seq data, we designed two data sets, one real and one simulated. These data sets, as well as the data analysis plan, are reported in Chapter 8, while a short introduction about methods for time-series data analysis is integrated in the review presented in Chapter 2. Strengths, limitations and future developments of the present study are discussed in Chapter 9.

2

A review of computational methods for RNA-seq data analysis

The enthusiasm for NGS has paved the way to a fast and wide application of RNA-seq to the study of gene expression. The powerful features of RNA-seq, such as single-base resolution, along with the elimination of many limitations of the previous technologies, have boosted an unprecedented progress of transcriptomics research, producing an impressive amount of data worldwide. To support this exponential growth, several computational tools have been developed and updated at a fast pace to deal with the different steps of data analysis. Nevertheless, the research carried out so far does not enable a complete and up-to-date characterization of their features and the selection of the best performer method for each analysis phase, preventing the definition of a unified analysis pipeline. In this chapter, we review and critique the currently available methodologies, for each step of RNA-seq data analysis. We also describe and discuss a possible probabilistic model of RNA-seq data considering both biological and technical sources of variation. Rather than providing a list of all the available tools, we focus on the underlying mathematical and statistical strategies and present few examples of the most used software.

2.1 Algorithms for read mapping

The advent of the Next-Generation sequencing has led to a significant drop of sequencing costs and boosted an exponential growth of the sequencing capacity worldwide (section 1.2). Less than fifteen years ago, on 23rd November 1999, the Human Genome Project held an impressive celebration to mark the completion of the sequencing of one third of the human genome, about 1 billion bp (`http://www.genome.gov/10002105/`). Nowadays, sequencing 1 billion bp requires few hours of work in any lab equipped with an Illumina or SOLiD sequencer [62]. Despite the revolution that has involved sequencing technologies, making sequencing more accessible, the interpretation of the massive amount of sequence data produced is not straightforward. Indeed, reads generated with NGS technologies are much shorter than conventional Sanger's data, with an error content depending on the specific platform adopted. Thus, new algorithms have to be developed to reconstructthe sequenced genome or transcriptome from short reads. Moreover, NGS sequencers generate hundreds millions of sequences in a single run [63], thus requiring algorithms to be optimized for speed and memory usage. Sequence capacity is growing at such a fast pace that algorithmic speed might become soon a major bottleneck in NGS data analysis [62].

After sequencing, two different approaches can be exploited to reconstruct the original sequence starting from short reads: alignment on a reference sequence or *de novo* assembly. As the latter option does not track the location of each single read, thus requiring a further alignment step, sequence alignment (or *mapping*) can be considered of fundamental importance for all NGS applications. This is especially true for RNA-seq, in which the identification of the correct source position of each read within the transcriptome is essential for expression quantification.

So far, many alignment tools have been proposed [64]. In all cases, the mapping process starts by building an index of the reference genome or the reads, which is then used to quickly retrieve the set of positions in the reference sequence where the reads are more likely to align. Once this subset of possible mapping locations has been identified, alignment with slower and more sensitive algorithms (such as the Smith-Waterman algorithm [65]) is performed in these candidate regions [62, 63]. The available mapping tools can be divided into two main categories, by considering the technique used to build the index: algorithms based on hash tables or on the Burrows-Wheeler transform (BWT).

The ***hash table*** is a common data structure for indexing complex and non-sequential data so to facilitate rapid searching. This feature is particularly suited for DNA sequences, since they are extremely unlikely to contain every possible combination of nucleotides and very likely to contain repeats [62]. Mapping tools can build hash tables either on the set of input reads or on the reference, considering all subsequences of a certain length k (*k-mers*) contained in the reads or in the reference sequence. For instance, when using the reference to build the hash table, the key of each entry is a *k-mer*, while the value is the list of all positions in the reference where the *k-mer* was found. Then, the set of input reads is used to scan the hash table and to find *k-mer* occurrences. The two solutions have different advantages and disadvantages [62, 63]. For instance, building hash tables of the reference requires constant memory (for a given reference and parameter set), regardless of the size of the input read data set. Conversely, building hash tables based on the set of input reads typically requires variable but smaller memory footprint, depending on the number and complexity (i.e. sequence diversity) of the input reads. However, it may require longer processing time to scan the entire reference sequence when searching for hits, even if the input read set is small. Moreover, algorithms based on genome indexing can exploit a possible parallelization resulting in a reduced computational time, while parallelization is not effective when read indexing is used [63]. Examples of hash-based algorithms are: GSNAP [66], Novoalign [67], mrFAST [68], mrsFAST [69], FANGS [70], MAQ [71] and RMAP [72].

Methods based on the ***Burrows-Wheeler transform*** create an efficient index of the reference sequence assembly in a way that facilitates rapid searching in a low-memory footprint. They first employ BWT, a reversible process (i.e. the input sequence can be easily be reconstructed starting from its BWT) that reorders the reference, such that subsequences present multiple times appear together in the data structure. Then, an index of the BTW, called FM-index ("FM" stands for "Full-text index in a Minute space"), is built and later exploited to perform fast *k-mer* searching. An introduction to BWT and FM-index, along with some "didactic examples" illustrating their application to sequence matching, are presented in Appendix A.

The combination of BWT and FM-index ensures both limited memory and disk space requirement. For instance, for mammalian organisms, the FM-index has often the same size, or even less, of the input genome [62]. However, with respect to hash-based algorithms, the employment of BWT significantly increases the computational time needed for index construction [63]. However, the index has to be constructed only once for a given reference, so the required operations only have a minimum impact on computational time. BWT- and hash-based tools perform differently in the mapping step: BWT implementations are much faster than their hash-based approaches, despite with slightly reduced sensitivity [73, 63]. The trend of NGS technologies, with more and more reads produced at increasingly higher quality, is favoring BWT solutions over hash-based algorithms. The most used BWT-based tools include: Bowtie [74], Bowtie2 [75], BWA

[76] and SOAP2 [77].

Besides the underlying algorithm implemented for read mapping, and the possibility to handle either sequence-base reads, color-space reads (i.e. SOLiD) or both, the greatest differences between short-read mapping tools are due to the heuristics adopted to make the mapping problem treatable. Indeed, each tool provides different trade-offs between speed and mapping accuracy, adopting different algorithmic strategies regarding:

- Base quality scores;
- Gapped alignments;
- Mismatches due to sequencing errors or single-nucleotide polymorphisms;
- Paired-end reads mapping;
- Spliced alignment of RNA-seq data ;
- Genome and transcriptome annotations.

As explained in section 1.2, ***base quality scores*** provide a measure of the correctness of each base in the read. Most of the mapping tools use this information to improve mapping accuracy, trusting more the read bases having higher quality score.

Moreover, due to the presence of sequencing errors in NGS data, mapping algorithms must allow imperfect alignments, by tolerating a certain numbers of ***mismatches***. By increasing the number of allowed mismatches, algorithms are able to increase the percentage of mapped reads [63]. However, just a limited number of mismatches should be tolerated, in order not to augment the uncertainty in read mapping. Mapping tools have different default settings for the number of tolerated mismatches, and algorithms with more stringent thresholds (e.g., SOAP and mrsFAST that allow only two mismatches) results in lower percentages of mapped reads [63]. Since the error content increases along read sequences (see section 1.2), most of the algorithms limit the number of tolerated mismatches in the first part of the read, called *seed*. This strategy allows performing fast, near-perfect alignment of seeds to identify candidate alignment regions, which are then possibly extended considering the rest of the read with a slower and more sensitive algorithmic approach. The available tools have very different default settings regarding seed length and number of allowed mismatches (in the whole read and in the seed), but they can be usually modified by the user. Changing these parameters greatly impacts both mapping accuracy and computational performance [63, 78]. The seed strategy, along with the information provided by quality scores, make mapping algorithms suited for capturing sequencing error profiles, with an increasing frequency of mismatches along the read. However, some algorithms report an excess of mismatches in correspondence of read end, whereas other methods avoid this bias by truncating reads [78].

Besides systematic errors, the sequenced organism can present true ***single-nucleotide polymorphisms*** (SNPs), that result in nucleotidic differences between the reads and the reference. The flexibility/stringency given by the mismatch threshold and the information provided by quality scores are thus important to correctly map these reads, since reads having one or more SNPs have a lower probability of being mapped [73]. Higher sensisitivity can be obtained by mapping these reads with algorithms that implement a SNP-aware policy (e.g.[66, 79]).

Reads obtained from a target genome or transcriptome may differ substantially from the reference sequence and, in addition to SNPs, it can contain small insertions or deletions (*indels*). Algorithms that do not perform ***gapped alignment*** sometimes fail to align reads containing indels [80]). The first NGS mapping strategies avoided or limited gaps in the alignment due to the computational complexity of choosing a gap location (which increases with the read length), but more recent sofware versions accommodate gapped alignment (e.g. [74, 75]). If NGS data analysis is aimed at variant discovery, gapped alignment can play a fundamental role. Indeed, when gapped alignment is not implemented, a read containing an indel may still be mapped to the correct genomic location, but with consecutive (false) mismatches near the indel position that might be identified as SNPs [81]. In addition, algorithms that do not perform gapped alignment have been shown to have lower accuracy in mapping RNA-seq data, with a significant reduction of the number of correctly mapped reads in correspondence of regions surrounding indels [82].

Another difference comes from the ability of mapping ***paired-end reads*** (section 1.2). Most of the available tools have adapted their original single-end algorithm to accommodate paired-end reads and to leverage on the confidence provided by the expected distance between read pairs. However, it has been demonstrated that, for many tools, the percentage of mapped reads decreases when using the paired-end algorithm instead of their original single-end version; only BWA, when switched to paired-end mode, is able to maintain almost the same throughput while increasing mapping accuracy [81, 80]. Further research is needed to clarify the algorithmic motivations underlying these results.

In contrast to DNA-sequence alignment, algorithms developed for RNA-seq have to cope with ***splicing*** when aligning reads to a reference genome. Indeed, as explained in section 1.1, genes in eukaryotic genomes contain introns, which are instead removed from mature mRNA transcripts. Thus, mRNA transcripts consisting in concatenated exons, can generate reads that span exon-exon junctions. In order to map these junction reads back to the genome, algorithms for RNA-seq data analysis must handle spliced alignment. Generally, simple gapped alignment is not sufficient to account for introns because they can span a very wide range of lengths; for instance, in mammalian genomes, they can be form 50 to 100,000 bases long [80]. To accommodate junction reads, many tools implement a two-steps procedure: first, reads are mapped to the genome and used to identify putative exons; then, candidate exons are used to build all possible exon-exon junctions, which are considered for mapping junction reads, which failed to map in the first step (this approache is implemented for example in Tophat [55] and RSEM [79], both based on BWT-based alignment with Bowtie [74]).

In order to increase accuracy in transcript reconstruction, several programs can also use available gene and transcript ***annotations*** to guide spliced-reads alignment and improve overall mapping accuracy [80, 78]. For instance, an interesting solution is implemented in the RNASeq Unified Mapper (RUM) [82]. It aligns RNA-seq reads on both the genome and the transcriptome, leveraging on the speed of BWT-mapping. Then, it re-aligns unmapped reads, which may contain insertions, deletions or sequence rearrangements, using BLAT [83], a tool developed for expressed sequence tags, which can perform spliced alignment with high sensitivity. Nevertheless, these annotation-based algorithms can result in high false-positive rates, due to reads wrongly aligned to exon-exon junctions that are not expressed in the sample but are reported in the annotation [78].

Despite attempted in several works, the assessment and comparison of mapping algorithms, especially for RNA-seq reads, is not a trivial task [84, 78, 61]. Ideally, the perfect algorithm would find, for each read, its true genomic source. However, the presence of sequencing errors, repeats, SNPs and other genetic variants, greatly increase uncertainty in read mapping and even challenges the definition of what a correct mapping is [63]. In addition, the need for limited time and memory requirement necessarily force algorithms towards heuristics and suboptimal solutions. Finally, the different features of the input data and the possibility to greatly change the parameter settings, add further variability to the results [63]. In this scenario, it is impossible identifying the best tool, but the top performers have to be selected with respect to the specific application and input data, depending on the biological question under consideration [63, 60].

2.2 Counts: the *digital* measure of gene expression

Once the reads are assigned to some genomic locations, the number of reads aligned to each coding unit, such as exon, transcript or gene, are used as measure of its expression level. This *digital* measure is called “counts”. The most used approach for computing counts considers the total number of reads overlapping the exons of a gene. However, even in well-annotated organisms, a fraction of reads map outside known coding sequences, i.e. outside the boundaries of annotated exons [85]. Thus, an alternative strategy would consider the whole length of a gene, also counting reads from introns. Moreover, if correctly handled in the mapping step (see previous section), junction reads can be used to model the abundance of the different splicing isoforms of a gene [86]. Although the choice of the reads to be considered has the potential to change the gene counts estimates, limited research has been carried out to assess the available approaches [60].

As explained above, quantification of gene expression from RNA-seq data is typically implemented in the analysis pipeline through two computational steps: alignment of reads to a reference genome or transcriptome, and subsequent estimation of gene and isoform abundances based on aligned reads. Unfortunately, the reads generated by the main RNA-Seq technologies (see section 1.2) are generally much shorter than the transcripts from which they are sampled. As a consequence, if the transcripts from which they are derived are characterized by similar sequences, it is not possible to uniquely assign short reads to one specific gene. Indeed, the human genome contains duplicated and paralogous genes, with high sequence similarity (even 100%), and interspersed or tandem repeats that are likely to produce similar or identical short reads [79, 87, 88]. Due to the limited length of NGS reads, repeats challenge the reconstruction of the original input sequence in multiple NGS applications that depend on either read mapping or assembly. Thus, data arising from repeated regions have to be handled properly in order not to bias the results [88, 89]. RNA splicing (see section 1.1) makes transcriptome reconstruction even more challenging, generating alternatively spliced isoforms of the same gene that share a large part of their sequence and can be hardly assigned to one specific isoform. As a consequence, a non-negligible fraction of RNA-seq reads are *multireads*: reads that map with comparable fidelity to multiple positions of the reference genome or transcriptome. The fraction of multireads over the total mapped reads depends on the transcriptome and on read length, varying from 10% to more than 50% [87, 79]. When considering *isoform multireads*, i.e. reads mapping on multiple isoforms of the same gene, this percentage increases dramatically and exceeds 70% [87]. In addition, sequencing errors and true sequence polymorphisms in the sequenced transcriptome cause mismatches in the alignment between the reads and the reference sequence [90, 91, 92]. To handle this variability, mapping algorithms allow mismatches and small indels in reads alignment, resulting in an increased fraction of multireads.

One of the first strategies proposed for handling gene multireads was that of simply discard them, so to estimate gene expression considering only uniquely mapping reads [93, 46]. Due to the uncertainty of multireads mapping, that can introduce further biases in the interpretation of the results, this approach is quite used in the analysis of RNA-seq or NGS data in general [88]. Considering uniquely mappable reads and discarding multireads is a quite common approach but, in some cases, it can produce misleading results (e.g. in regions containing copy number variation). This issue is exacerbated in RNA-seq studies, where the aim is both the reconstruction of transcripts sequences and the quantification of their relative abundances. Discarding multireads necessarily leads to a loss of information and a systematic underestimation of expression levels in correspondence of repetitive regions.

A slightly different scheme uses only uniquely mapping reads to adjust exon coverage by a *mappability index* (i.e. the fraction of exon positions that can generate uniquely mapping reads) [94]. Another strategy "rescues" multireads by proportionally assigning them to genes considering the coverage given by the uniquely mapping reads [45]. With respect to the approaches that use only the uniquely mapping reads, the rescue strategy obtains expression estimates that are in better agreement with microarrays [45]. A more sophisticated approach also takes into account the mismatch profiles between the unique reads and the sequence of the genomic locations they are aligned to [95]. Ji *et al.* propose a method that implements a Bayesian mapping of multireads, called *BM-Map*, to calculate the posterior probability of mapping each multiread to a genomic location. The algorithm estimates multireads mapping probability considering three sources of information: the sequencing error profiles, the likelihood of true polymorphisms and the expression levels of competing genomic locations. Conversely, the proportional method described before only considers the latter information. The mismatch profile is also taken into consideration by MMSEQ [87], which estimates both isoform expression and allelic imbalance (i.e. expression differences between two alleles of the same gene or isoform). It uses a a two-steps alignment procedure to reduce the uncertainty in read mapping. In the first run, mismatch profiles are used to build a sample-specific transcriptome whose genotype can be different from that of the reference sequence. Once the reference transcriptome is updated considering the genotype, reads are re-aligned to estimate isoform expressions and allelic imbalance. More recent methods, such RSEM, define a probabilistic model of RNA-Seq data and calculate maximum likelihood estimates of isoform expression levels using the Expectation-Maximization algorithm [79, 96, 97]. True mappings are identified leveraging on the information provided by the distribution of fragment lengths, read across transcripts and sequencing errors, estimated from the data and modeled as ramdom variables.

2.3 Count bias and normalization

After the first enthusiastic expectations due to RNA-seq advantages over microarrays [39] many works have risen the need for a careful normalization of count data before assessing differential gene expression (DE or DGE) [50, 98, 99, 100, 101, 102], so to correct for different sources of bias.

The first bias to be taken into account is the ***sequencing depth*** of a sample, defined as the total number of sequenced or mapped reads. Let A and B being two RNA-seq experiments with no differentially expressed genes. If experiment A generates twice as much reads as experiment B, it is likely that the counts for experiment A will be doubled too. Hence, a common practice is that of scaling counts in each experiment j by the sequencing depth d_j estimated for that sample. In early works d_j was computed by counting the total number of reads sequenced or mapped in sample j (*global scaling*) [93, 45]. More recent approaches consider count data depending on the state of the whole RNA population of the sequenced sample [103, 104, 105]. For instance, as previously reported in [104], if there is a set of ***highly expressed genes*** in a sample, it will inevitably "consume" the available reads so that the remaining genes will be underestimated. A similar issue derives from the presence of contaminants. According to Bullard *et al.*, global scaling normalization techniques reflect the behavior of a restricted set of high-counts genes [50]. They verified that, in all analyzed samples, about 5% of genes account for 50% of total counts and proposed a quantile normalization similar to that used for microarray pre-processing [106]. They also proposed an altenative global scaling, to adjust counts distributions with respect to their third quartile, so to reduce the effect of high-counts genes. Smyth *et al.* [104] proposed the Trimmed Mean of M-values (TMM) normalization to account for differences in library composition between samples. In order to reduce bias due to high-count genes, TMM is computed removing the lowest and highest 30% of the data, so to exclude those genes that are characterized by extreme M-values (namely, log-fold-changes). This normalization factor is then used to correct for differences in library sizes. Tibshirani *et al.*[105] proposed a novel normalization method that assumes a Poisson model of counts and estimates sequencing depth on a set of genes that are not differentially expressed. A Poisson goodness-of-fit statistic is employed to determine which genes belong to this restricted set. The method reduces to total-count normalization when all genes are considered. Finally, in the R package DESeq [107], the ratios between gene-wise counts in each sample s and the geometric mean of gene-wise counts across all samples are calculated, and the library size is computed as the median of these ratios across genes.

Furthermore, RNA-seq data show a gene ***length bias***: the expected number of reads mapped on a gene is proportional to both the abundance and length of the isoforms transcribed from that gene. Indeed, longer transcripts produce more reads than shorter ones, resulting in higher power for DE detection [50, 60, 108] and biased gene set analyses (GSA) [109]. Thus, the number of reads should be normalized by transcript length to obtain the true gene expression levels. Mortazavi *et al.* [45] proposed to summarize mapped reads as “Reads Per Kilobase of exon model per Million mapped reads” (RPKM), computed dividing the number of reads aligned to gene exons by the total number of mappable reads in the experiment and by the sum of exonic bases, so to correct length bias. Oshlack *et al.* [60] demonstrated that the power of tests statistics for detecting DE genes from RNA-seq data is strongly associated with gene length (calculated as the median length of all transcripts arising from that gene). The authors also showed that scaling counts by gene length does not completely remove this bias. Other studies demonstrated that scaling test statistics by the inverse of the square root of length, without transforming gene counts, improves DE analysis [50, 108].

Recent works reported other evident sequence-dependent sources of bias in NGS data [99, 100, 26, 110]. In particular, many authors documented the presence of a ***GC-content effect*** in RNA-seq data [111, 112, 98]. The results presented by Zengh *et al.* [111] show a strong relationship between different measures of gene expression for RNA-seq data (raw counts, RPKM and FPKM [86]) and sequence specific covariates, such as GC-content, gene length and dinucleotide composition. They used a generalized additive model of log-counts, together with gene length, GC-content and di-nucleotide composition, to remove the effect of these covariates on digital expression estimates. Since GC-content and di-nucleotide composition can be correlated, they selected a lower number of covariates using principal component analysis (PCA). Hansen *et al.* [112] proposed a conditional quantile normalization (CQN) method, which assumes a Poisson model of read counts and assesses the bias due to GC-content and gene length. These covariates are modeled as smooth functions and estimated from data using robust quantile regression on log-counts. CQN does not directly normalize data, but rather provides a normalization offset that can be incorporated in existing methods for DE detection. Risso *et al.* [98] also demonstrated a clear relationship between log-counts and GC-content. They proposed a within-lane normalization method based on loess regression of log-counts on GC-content; regression on the log-scale of counts was used in order to be more robust to the presence of very-high count genes that might bias the fit.

Despite the availability of such a rich panel of methods for data normalization, all of them are based on an initial count of the total number of reads mapping on each transcript [50, 98, 112, 111]. This procedure, in principle robust to random noise, might be error-prone if reads are not uniformly distributed along sequences, as happens indeed due to both sequencing errors and ambiguity in read mapping. ***Non-uniformity of read coverage*** is mainly due to biases associated to the different steps of RNA-seq protocols. For instance, fragmentation methods based on restriction enzymes have recently been reported to be sequence-specific and not random [113]. Reverse-transcription performed with poly-dT oligomers, which bind to poly(A) tails, is strongly biased towards 3' end of transcripts [46, 39]. Conversely, reverse-transcription with random hexamers results in an under-representation of 3' ends [39, 113]. This bias is due to the reduced number of priming positions from which the reverse transcriptase enzyme can start cDNA synthesis. Furthermore, depending on their sequence, RNAs and cDNAs can form secondary structures that alternatively obstruct or facilitate the binding of reverse-transcription primers and sequencing adapters, resulting in different efficiency of the sequencing process [101]. Since the first RNA-seq experiment [46], several changes in library preparations and sequencing protocols have been proposed pursuing the aim of having an unbiased representation of transcript abundances (e.g. postponing reverse transcription after fragmentation), but the non-uniformity of read coverage along transcripts remains an issue of state-of- the-art technologies [114].

2.4 Differential expression analysis and models of RNA-seq data

A fundamental research problem in which RNA-seq has soon found application is the identification of differentially expressed (DE) genes between different conditions or groups, such as healthy and diseased tissues. In recent years, a fervent research has characterized the RNA-seq field and many different tools for DE detection have been developed [115]. Most of them have been implemented in user-friendly R packages [116]. At its simplest, methods for DE detection rely on a test statistic, used to identify which genes are characterized by a *statistical significant* change in gene expression (namely counts in RNA-seq data) in the compared conditions. In principle, non-parametric methods can be used (e.g. [117]). However, due to small number of replicates typically available in RNA-Seq experiments, non-parametric methods do not offer enough detection power and parametric methods are preferred [118, 119]. Each parametric method assumes a particular model to describe the underlying distribution of count data, and seeks to identify those genes whose differences between the tested conditions exceed the variability predicted by the model. The main models considered and implemented in the major analysis tools are the Poisson and the Negative Binomial distribution. In the following, we present a statistical description of the parameterization of RNA-seq count data and a more general summary of state-of-art approaches for DE analysis in RNA-seq studies. However, due to the high number of tools available, here we specifically focus on few interesting data modelling approaches implemented in recently developed methods, but additional details about other approaches and method implementations can be found in the original papers or in comparative studies, such as [120, 118, 121].

Models of RNA-seq count data

Let $f = 1, ..., F$ be the set of transcripts in the sample of interest j. For each transcript f in sample j, let l_f be its length and θ_{fj} the number of copies of f present in the sample. The total number of bases from all the transcripts in sample j can be computed as

$$\sum_{f=1}^{F} \theta_{fj} l_f. \tag{2.1}$$

Therefore, the probability that a read comes from some transcript f in sample j, is given by

$$\pi_{fj} = \frac{\theta_{fj} l_f}{\sum\limits_{f=1}^{F} \theta_{fj} l_f}. \tag{2.2}$$

The formula at the numerator counts all the positions within a transcript that can give rise to a read (i.e. all possible read starts). Thus, the numerator could be more precisely modeled as $\theta_{fj} \cdot (l_f - L + 1)$, where L is read length.

According to [122], we model the sequencing process as a simple random sampling, in which every read is sampled independently and uniformly in sample j. Under this hypothesis, the number of reads arising from transcript f, i.e. the so-called counts, can be modeled as a random variable N_{fj} following a binomial distribution. Indeed, read sampling can be viewed as a Bernoulli's process, a random experiment with only two possible outcomes: *success*, when the read is sequenced from transcript f, and *failure*, when the read is sequenced from another transcript. If R_j is the total number of reads sequenced in sample j, the random variable giving the number of successful events in R_j independent trails is given by the binomial distribution $\mathcal{B}(R_j, \pi_{fj})$, where the *success* event has probability π_{fj} and the *failure* event has probability $1 - \pi_{fj}$. Thus, the probability of having $N_{fj} = r$ reads from transcript f is described as follows:

$$Pr(N_{fj} = r) = \mathcal{B}(R_j, \pi_{fj}) = \binom{R_j}{r} \cdot (\pi_{fj})^r \cdot (1 - \pi_{fj})^{R_j - r} \tag{2.3}$$

Since $R_j \simeq 10^7 \div 10^8$ and $\pi_{fj} << 1$, this distribution can be approximated by a Poisson distribution $\mathcal{P}(\lambda_{fj})$, with parameter $\lambda_{fj} = R_j \cdot \pi_{fj}$:

$$Pr(N_{fj} = r) = \mathcal{P}(\lambda_{fj}) = \frac{(R_j \cdot \pi_{fj})^r}{r!} \cdot e^{-(R_j \cdot \pi_{fj})} \tag{2.4}$$

The parameter λ of the Poisson model corresponds to both the mean μ and the variance of the distribution. The Poisson distribution is commonly used to model RNA-seq count data:

$$N_{fj} \sim \mathcal{P}(\lambda_{fj}). \tag{2.5}$$

It has been demonstrated that the Poisson distribution captures the variability between RNA-Seq technical replicates sequenced in different lanes or flow-cells [93, 50, 123, 124]. However, in the presence of biological replicates, there are two sources of variation that affect RNA-seq counts:

Technical variation representing the measurement error due to the technology.

Biological variation representing the eterogeneity among samples belonging to the same treatment group or condition.

In the presence of biological replicates, the variance is larger than the mean and count data are said to be *over-disperded* [123, 124, 121]. In this case, the Poisson distribution cannot handle this additional variability, and models based on the Negative Binomial (NB) distribution of count data are preferred [123, 124, 107, 121]. Indeed, the number of copies of transcript f is not the same across different biological replicates and the resulting λ is a a random variable, with mean μ and a certain variance $Var(\lambda_{fj})$. If λ is modeled with a Gamma distribution, the marginal probability distribution of counts is Negative Binomial, with mean μ and variance that depends on the chosen parametrization of $Var(\lambda_{fj})$ variance. Indeed, there are different ways to parametrize the Gamma distribution, that lead to different Negative Binomial models [121]. Identifying with $\varepsilon(j)$ all the replicates that belong to the same condition or phenotype, if $N_{fj}|\pi_{fj}$ ~Poisson with mean λ_{fj} and λ_{fj} ~Gamma with mean $\mu_{f\varepsilon(j)}$, the marginal distribution of N_{fj} is NB with mean $\mu_{f\varepsilon(j)}$ and variance that depends on the parametrization of λ_{fj} variance. In particular:

- If $Var(\lambda_{fj}) = \phi\mu_{f\varepsilon(j)}$, then

$$Var(N_{fj}) = \mu_{f\varepsilon(j)}(1 + \phi). \tag{2.6}$$

- If $Var(\lambda_{fj}) = \phi\mu^2_{f\varepsilon(j)}$, then

$$Var(N_{fj}) = \mu_{f\varepsilon(j)}(1 + \phi\mu_{f\varepsilon(j)}). \tag{2.7}$$

- If $Var(\lambda_{fj}) = \phi\mu^{\alpha}_{f\varepsilon(j)}$, then

$$Var(N_{fj}) = \mu_{f\varepsilon(j)}(1 + \phi\mu^{\alpha-1}_{f\varepsilon(j)}). \tag{2.8}$$

RNA-seq counts can be modeled as a NB variable with parameters ϕ and $\mu_{f\varepsilon(j)}$:

$$N_{fj} \sim \mathcal{NB}(\mu_{f\varepsilon(j)}, \phi). \tag{2.9}$$

The *overdispersion* parameter ϕ accounts for the variance that is not explained by the Poisson model. In the case of $\phi = 0$, the NB model reduces to the Poisson distribution. In summary, the NB distribution can be motivated as a Gamma mixture of Poisson distributions: the technical variability is Poisson, but the Poisson means differ between biological replicates according to a Gamma distribution.

Tools for differential expression analysis of RNA-seq data

Given a specific statistical model of RNA-seq count data, all parametric tools for DE analysis consist in two main steps: estimation of model parameters from the data and detection of differential expression with a test statistics. Library normalization (discussed in section 2.3) can also be considered part of the DE analysis [120] since it is required and thus implemented in all DE methods (despite with different approaches). So far, several studies have been focused on DE methods comparison [125, 126, 127, 119, 120, 118, 128, 121], but not all of them can be considered complete (e.g. they use only simulated data) nor present fully concordant results. Cross-comparison of the different insights provided by the studies here considered is further challenged by the fast improvements and updates that are characterizing these tools, with several versions released each year [119, 118]. However, some findings are widely confirmed across several studies, such as the superior performance of NB-based methods over Poisson-based models [125, 126, 123, 127, 120, 121].

The better performance of the NB-based tools is mainly due to their ability of capturing the biological variability. This varibility is due to the stochastic nature of gene expression, which can be gene-specific, causing some genes to have more variable levels of expression than others, and is independent from the specific technology adopted [129]. Thus, it cannot be reduced by increasing the sequencing depth of an RNA-seq experiment, but only by sequencing more biological replicates [119, 130]. In particular, Robles *et al.* demonstrated that increasing the number of biological replicates improves the quality and reliability of DE detection, while higher sequencing depths do not add significant benefits [119]. Given these results and the costs related to RNA-seq, they suggest to sequence more biological replicates deriving from a multiplexing experiment design (see section 1.2) and demonstrated that the gain of multiplexing n biological replicates in the same library is greater than the loss of available reads per sample by $1/n$.

In NB-based models, biological variance is captured by the dispersion parameter ϕ. However, for the reasons discussed above, dispersion is not constant across all genes, but varies depending on the specific gene g. The different ϕ_{gj} parameters are too many to be estimated from data sets with few replicates, as in the case of RNA-seq. edgeR [131] and DESeq [130], which are among the best performers in most of the comparative studies cited above, are both based on the NB model of equation 2.7, but implement different strategies for dispersion estimation. The default strategy implemented in edgeR shrinks gene-wise dispersion estimates towards a common value. Alternatively, the user can select a "trend" approach, requiring edgeR to compute a trend estimate across genes in place of a single value. DESeq considers the variance being a smooth function of the mean μ and uses non-parametric regression to fit the variance as a function of the mean. Another approach, implemented in NBPseq [121], considers instead the model with three parameters described in equation 2.8; ϕ and α parameters are considered constant across genes and estimated jointly. Nevertheless, this approach does not outperform DESeq and edgeR [119]. Apart from being top-ranking methods in several comparisons available in the literature, edgeR and DESeq are widely adopted also thanks to frequently updated sofware, supported by constant research, and well-documented manuals and tutorials that ease their application to different studies.

More recently, Law *et al.* proposed to apply limma [132], a method developed for microarrays and based on normal distribution, to analyze RNA-seq data summarized as log-cpm [126]. The underlying idea is that correctly modelling data mean-variance relationship is far more important that exactly specifying the probabilistic distribution of counts. In their approach, called "Voom" (acronym for "variance modelling at the observational level"), the mean-variance trend is estimated from the data through lowess fit and used to estimate single-gene variances. For each gene, the inverse of the variance is then used as weight in the limma framework. Applied to RNA-seq data, Voom results are comparable to top-ranking NB-based approaches [126, 118]. Even though further assessments are needed to finally select the best approach for differential expression analysis from RNA-seq data, the promising results obtained with this strategy open the possibility to exploit a wide panel of methods developed for microarrays.

2.5 RNA-seq time-series

The analysis of differential gene expression can help investigating which genes are activated in certain conditions or developmental stages. However, since gene expression is a dynamic process that changes over time, the analysis framework presented in the previous chapters may fail in capturing a complete view of gene functions and interactions, as well as the biological implications of different expression dynamics. Conversely, in time-series experiments the process under investigation is assayed several times, so to monitor transient or evolving gene expression changes. This approach is particularly suited for studying developmental stages, cycling processes and response to stimuli. In particular, the latter approach allow investigating if different set of genes respond to the stimulus with different expression dynamics, in terms of both the magnitude and quickness of the induced change. For instance, some genes can be characterized by a long sustained response, with a change in expression that persists for a long time. Differently, other genes may show fast-changing expression patterns more similar to short impulse, in which a new expression level is kept for a limited time and then the original steady state is reached again. In both cases, the changes can represent over-expression or under-expression with respect to the basal expression levels. Gene expression patterns can be generally be described by combinations of these basic expression patterns [133]. All these different kinetic profiles can be captured only by sampling the process at multiple time points (see Figure 2.1a and 2.1b). Nevertheless, the definition of the sampling design is not straightforward and detection of gene expression changes at the time of occurrence may require some prior knowledge about the process under investigation. Moreover, the number of sampled time points is necessarily limited by the experimental costs inherent to the specific technology adopted.

Once the expression time-series data are generated, a possible computational workflow can be defined following [133], independently of the chosen technology.

The first step is ***normalization***, needed for making expression levels comparable between different time points and, if possible, between different genes. Expression data normalization is not a trivial task and has prompted years of research in the microarray field and recent discussions about how to deal with data RNA-seq biases (see section 2.3).

Before real analysis, a second step regards data ***visualization***: expressions levels over time, or their log-fold-changes with respect to the basal level, are graphically plotted to have a first overview of data. One of the most used techniques are heat-maps

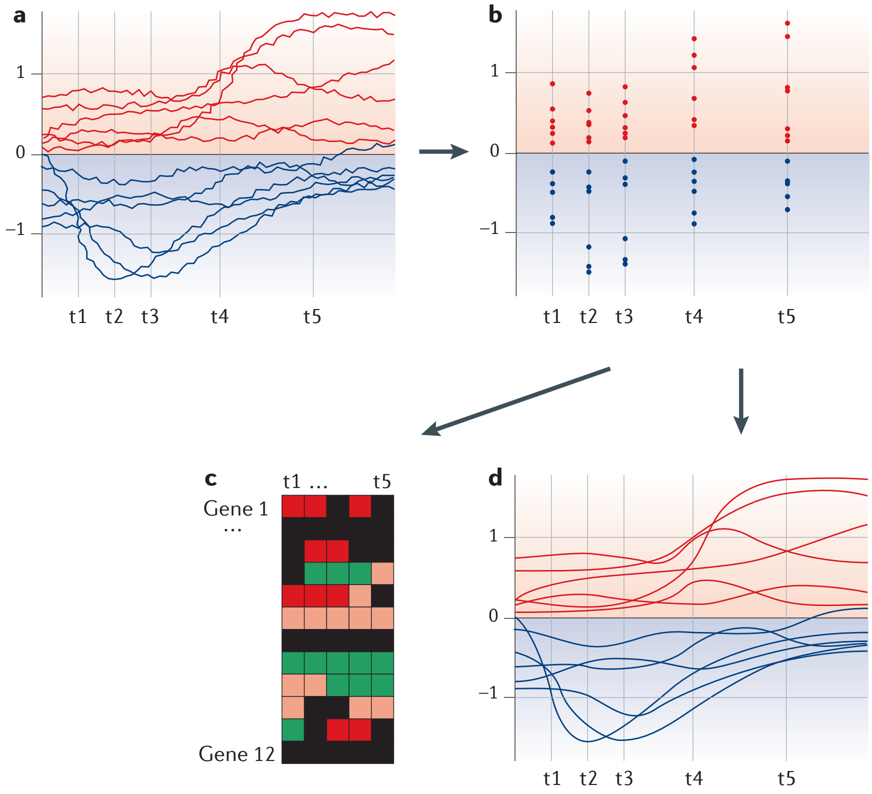


Figure 2.1: Image adapted from [133] showing true gene expression patterns (log-fold-changes) over time (a), values sampled in a time-series experiment (b) graphical data representation through heat-map (c) and spline interpolation (d).

(Figure 2.1c), which encode the expression values measured for each gene and time points, in a 2D matrix of “pixels”. The main limitation is due to the lost of information about duration: the same importance is given to each value, independently from the actual experimental sampling design. More sophisticated methods employ 2D curves of expressions (or fold-changes) referred to correctly spaced time points. Patterns can be then revealed by connecting points through straight lines or fitting data with cubic splines (Figure 2.1d) [134].

Important questions over the investigated process can be then answered identifying ***differentially expressed genes***. One heuristic approach defines a gene as differentially expressed if its log-fold-change, with respect to the basal value, exceeds a chosen threshold in at least two time points. However, the selection of a proper threshold is not straightforward since a certain log-fold-change may be appropriate for some expression levels but not for others. More sophisticated methods consider the whole time-series and may require replicates to estimate an error model so to be robust to stochastic oscillations [135, 132, 135, 136, 137, 138].

Clustering analysis provides the second important contribution to the understanding of gene interaction and co-regulation by grouping expression patterns that share similar kinetics profiles. This step is often coupled with visualization. Although methods developed for static data, such as k-means or hierarchical clustering, have been widely applied also to time series-data, more sophisticated approaches have been specifically developed for dynamic data [139, 140, 141, 142]. Clustered profiles can be then used to infer causality between genes, i.e. how expression changes are propagated through the network of gene interactions.

In the near future, RNA-seq is expected to replace microarrays for time-series expression studies [133]. The analysis framework here described will be probably mantained because the purpose of the analysis remain the same, but the specific methods to be used still have to be defined. Further research is needed to investigate if the available methods, developed for microarray data, can be adopted and adapted for the analysis of RNA-seq time-series data. Moreover, we expect that additional tools will be specifically developed to leverage on RNA-seq single-base resolution, to investigate more deeply all the processes related to gene expression, such as isoforms switching across time.

3
Selection of benchmark data sets

For benchmarking the methods for RNA-seq data analysis and defining the computational pipeline we consider three real RNA-seq data sets with different characteristics and relative gold-standard measures of RNA abundances, such as spike-in RNAs or quantitative Real-Time PCR (qPCR). All RNA-seq data sets were sequenced with the Illumina technology (see section 1.2), which is now the most commonly used NGS platform for RNA-seq [47]. We also considered six simulated data sets of synthetic RNA-seq reads to assess the correctness of read mapping, with particular attention to multireads. Finally, we simulated several synthetic count data sets with different extent of differential expression, percentage and distribution of DE genes across samples, so to investigate the effect of these factors on data normalization and differential expression analysis.

3.1 Real data

We consider three real RNA-seq data sets [143, 50, 144] that are publicly available from the NCBI Sequence Read Archive [145].

The MAQC2 data set [50], generated by the MicroArray quality control (MAQC) project, contains expression data from multiple platforms. Here we consider a data set of 36bp single-end reads [SRA: SRA010153], obtained by sequencing with the Solexa 1G

Genome Analyzer two different biological samples: (i) Ambion's Human Brain Reference RNA ("Brain"), a standard pooled from multiple donors and several brain regions; (ii) Stratagene's Universal Human Reference RNA ("UHR"), a mixture of total RNA extracted from ten different human cell lines. "Brain" and "UHR" samples were subjected to the same library preparation protocol and sequenced in seven lanes of two flow-cells (technical replicates).

In Griffith *et al.* [143], two fluorouracil (5-FU)-resistant ("MIP5FU") and (5-FU)-sensitive ("MIP101") human colorectal cancer cell lines were sequenced on 16 and 23 lanes of a Illumina Genome Analyzer platform, respectively. Here we consider eight "MIP101" and eight "MIP5FU" libraries of 36bp paired-end reads. FASTQ files of RNA-seq reads were kindly provided by Dr. Malachi Griffith. A subset of exons were also assayed with qPCR and made available by the authors [143].

A series of replicates from Jiang's study [144] is also considered, in which paired-end RNA-seq libraries were sequenced after mixing endogenous RNA from a K-562 cell line with spike-in RNA developed by the External RNA Control Consortium (ERCC). ERCC spike-in RNAs are in vitro synthesized transcripts whose nucleotidic sequences and concentrations are known. They can be used to assess whether the final quantification of an RNA-seq experiment correctly represents the composition of the original input. In Jiang *et al.* different Human ENCODE libraries were mixed with the Phase IV test set of ERCC spike-in RNAs. The human samples mixed with ERCC RNAs were sequenced on the Illumina GAIIx platform to generate 2x76bp paired-end reads. Here we consider K-562 RNA-seq data, obtained from RNA extracted from nucleus, cytosol or whole-cell.

3.2 Synthetic data

Simulated reads

Synthetic RNA-seq reads were kindly provided by A. Gatto [146] and were simulated using Flux Simulator [114] (version 1.2). Simulations were based on the GRCh37.p8 assembly of the human genome. Flux Simulator was used to generate 2x76 bp paired-end reads, using a custom error model at 50 bp read length estimated from real RNA-Seq data and adopting default parameters for all the other options. A total of six libraries were produced: three with 8 millions (8M) reads and three with 20 millions (20M) reads.

Simulated counts

Synthetic RNA-seq counts were also simulated following the approach described in [127].

Simulated counts for 10,000 genes were sampled from a Negative Binomial distribution. Experiments with two conditions and five replicates for each condition were generated. Differential expression was simulated multiplying or dividing by $\sqrt{b}$ the mean of the distributions from which they were sampled. The simulation was run varying the extent of differential expression, using b=2, 4, 6, 8, and the percentage of DE genes, set to be 10% and 20% of total genes (DE=0.1, 0.2). Moreover, to investigate the effect of the distribution of over-expressed genes between the two conditions, we set the percentage of over-expressed genes in condition A to be A=0.5 (i.e., equally distributed between the two conditions), 0.75 or 0.9 (i.e., predominantly present in condition A with respect to B).

4

Quantitative assessment of RNA-seq data bias and normalization

In this chapter we assess the biases of RNA-seq data (see a review in section 2.3) before and after normalization. We compare some of the most used methods for library size normalization, transcript length and sequence-specific biases reduction: RPKM [45], the library scaling approaches implemented in the R packages PoissonSeq [105] and DESeq [107], the TMM normalization implemented in edgeR [131], between-lane full-quantile normalization and within-lane full-quantile normalization on GC-content and sequence length, all implemented in EDASeq [98]. CQN normalization is not tested since Risso *et al.* [98] already demonstrated that it is less accurate than within-lane full-quantile normalization. To test and benchmark these methods, we consider real data sets and simulated counts (see Chapter 3). In particular, we employ diagnostic plots to investigate the presence of the above mentioned biases in real data. Moreover, we use the real data sets to identify which methods provide an accurate measure of RNA abundances and differential gene expression, and we simulate synthetic count data to better characterize the data features that represent the greatest challenges to data normalization.

We recognize that a bias may represent an issue for one application but may be harmless for another. For instance, some gene-specific biases might cancel out in differental expression analysis, but can prevent a correct interpretation of time-series data. For this reason, we decided to explore bias at three levels (i.e. counts, RNA quantification and differential expression estimates) to investigate how it propagates and impacts on downstream analyses.

4.1 Materials and methods

Read mapping and data pre-processing

For data downloaded from the SRA [145], FASTQ files were obtained using the function fastq-dump 2.1.12 of the SRA Toolkit [147]. From Jiang's data set (see Chapter 3), we selected five libraries from the "cytosol" group [SRA: SRR317052, SRR317053] and seven libraries from the "nucleus" group [SRA: SRR317042, SRR317043]. Raw reads were aligned to the human genome using TopHat v1.2.0 [55]. Multimappers (i.e. reads that align to multiple positions) were discarded from our computation (TopHat's `-g 1` option). Counts were computed for all Ensembl exons using the function `coverageBed -counts` of bedtools 2.15.0 [148]. To compute transcript counts for spike-in RNAs, reads were directly aligned on ERCC RNAs sequences. Exons and spike-in transcripts with an average counts across replicates lower than 0.5 were discarded from subsequent analyses. Data processed with within-lane full-quantile normalization were further filtered using a more stringent threshold of 10 counts, as suggested by authors [149]. We adopt this strategy because we experienced in previous assessments that reducing the stringency of this filtering phase reduces the effectiveness of within-lane full-quantile normalization, especially the length normalization of high-count sequences (results not shown). Annotations on exon length and GC-content were retrieved using the R package biomart [150] ("ensembl" database, "hsapiens_gene_ensembl" dataset).

Count normalization

RPKMs for each exon i in library j were calculated as follows:

$$RPKM_{ij} = \frac{N_{ij}}{{}^{l_i}\!/_{10^3} \cdot {}^{N_{\cdot j}}\!/_{10^6}} \tag{4.1}$$

where, N_{ij} are counts for exon i in library j , l_i is the length of exon i and $N_{\cdot j} = \sum_i N_{ij}$ is the sum of all counts in library j.

Within-lane full-quantile normalizations on exon length or GC-content were performed using the default parameter settings, except for data sets with more than 1000 exons (after the most stringent filtering phase), for which full-quantile normalization on length was performed using 200 bins. For data to be shown in the "diagnostic plots", the normalization factors provided by TMM were multiplied by library sizes $N_{\cdot j}$ and then used to scale counts in each library.

Spike-in RNA quantification

We considered all the technical replicates from “cytosol” and “nucleus” samples. Counts were summarized at exon level for human genes, while for ERCC spike-in RNAs, which are single-isoform, per-transcript counts were computed. Exons or spike-in RNAs with low counts were discarded from subsequent analyses as described above, eventually considering 74 ERCC sequences over the total 96. Before within-lane full-quantile normalization over exon length or GC-content, we applied a more stringent filter as described above. This further filtering phase reduces the number ERCC RNAs normalized via within-lane full-quantile normalization from 74 to 58. Finally, for all ERCC spike-in RNAs, the mean of counts (raw or normalized) across all technical replicates of the two samples was taken as measure of their abundance. These values were then compared with the true spike-in RNA concentrations reported in [144].

Differential expression analysis

DE analysis was performed with the GLM-based version of edgeR [151] as it can consider a vector or a matrix of normalization factors. In particular, edgeR was provided with raw (i.e. not normalized) counts, along with a vector of normalization factors (`normfactors` parameter in the `DGEList` function), in the case of DEseq or PoissonSeq normalization, or a matrix of “offsets” (`offset` parameter in the `estimateGLMCommonDisp` and `glmFit` functions), in the case of full-quantile or RPKM normalization. TMM normalization factors were directly computed with edgeR, using the `calcNormFactors` function. For Griffith’s data, only those exons whose log-fold-changes, estimated by edgeR, were not greater than 20 in absolute value were considered for comparison to qRT-PCR gold standard. Log-fold-changes computed by edgeR from RNA-seq data, raw or normalized, were then compared with those provided by qPCR [143], here considered as gold standard. To assign each PCR amplicon to the corresponding exon, PCR primers were aligned to human exon sequences using the `ssearch` function of the `fasta-36` package [152]. Human exon sequences were extracted from the whole chromosomes (GRCh37) using BEDTools [148], considering Ensembl annotation.

Sensitivity *vs.* precision curves

For each method, we imposed an iteratively increasing threshold on un-adjusted p-values estimated by edgeR, to select DE genes; the number of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs) was computed, accounting for the direction of differential expression, as explained in Table 4.1.

Table 4.1: Scheme of the rules for determining true/false positives and negatives.

		RNA-seq		
		DE+	DE-	nonDE
	DE+	TP	FP	FN
Gold standard	DE-	FP	TP	FN
	nonDE	FP	FP	TN

Finally, precision and sensitivity indexes were calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Sensitivity = \frac{TP}{DE} \tag{4.3}$$

where DE is the total number of differentially expressed genes detected by the gold-standard.

Precision (Equation 4.2) is computed as usual, while sensitivity (Equation 4.3) is here defined using at the denominator the number of "true" DE genes, in order to account for DE genes that are correctly detected as DE, but with a wrong direction of differential expression (FPs).

4.2 Length bias and GC-content effect

In Figure 4.1, smoothed scatter-plots of log-counts versus exon log-length and GC-content are reported for two replicates from MAQC2 and Griffith's data, to highlight whether counts calculated at exon level are affected by sequence-specific biases. As expected, there is an increasing pattern of log-counts in dependence on exon log-length: longer exons tend to have higher counts than shorter ones. The same patterns are confirmed across all replicates of the considered data sets (results not shown). We investigated sequence specific biases after count normalization considering exon length or GC-content as covariates (Table 4.2 and Figure 4.1). Table 4.2 reports the mean, across all replicates, of correlations between counts and covariates, before and after normalization. Counts normalized only for library size are not inspected here since would present the same features, in terms of dependence on exon length and GC-content, of the raw data. Raw counts show a low, significant, positive correlation of counts with exon lengths (correlation ranges between 0.28 and 0.41, with p-values $< 2.2e-16$), which is removed by full-quantile normalization over exon length. RPKMs are characterized by negative correlations with exon length, meaning that dividing by exon length over-corrects for length bias (Figure 4.1). GC-content effect varies on different data sets, as previously reported in [31], but the bias it introduces is weaker than length bias. Within-lane full-quantile normalization on exon length reduces GC-content bias, although this pattern is completely removed only by full-quantile normalization on GC-content.

Table 4.2: Mean, across replicates, of correlations between counts and covariates, before (*raw*) and after normalization: with RPKM (*rpkm*), within -lane full-quantile normalization on exon length (*fq_length*) or GC-content (*fq_gc*).

Normalization	**MAQC2**		**Griffith**		**Jiang**	
	length	GC%	length	GC%	length	GC%
raw	0.38	0.17	0.41	-0.04	0.28	0.07
rpkm	-0.28	0.25	-0.29	0.00	-0.51	0.12
fq_length	0.00	0.13	0.00	0.02	0.00	0.04
fq_gc	0.34	0.00	0.43	0.00	0.22	0.00

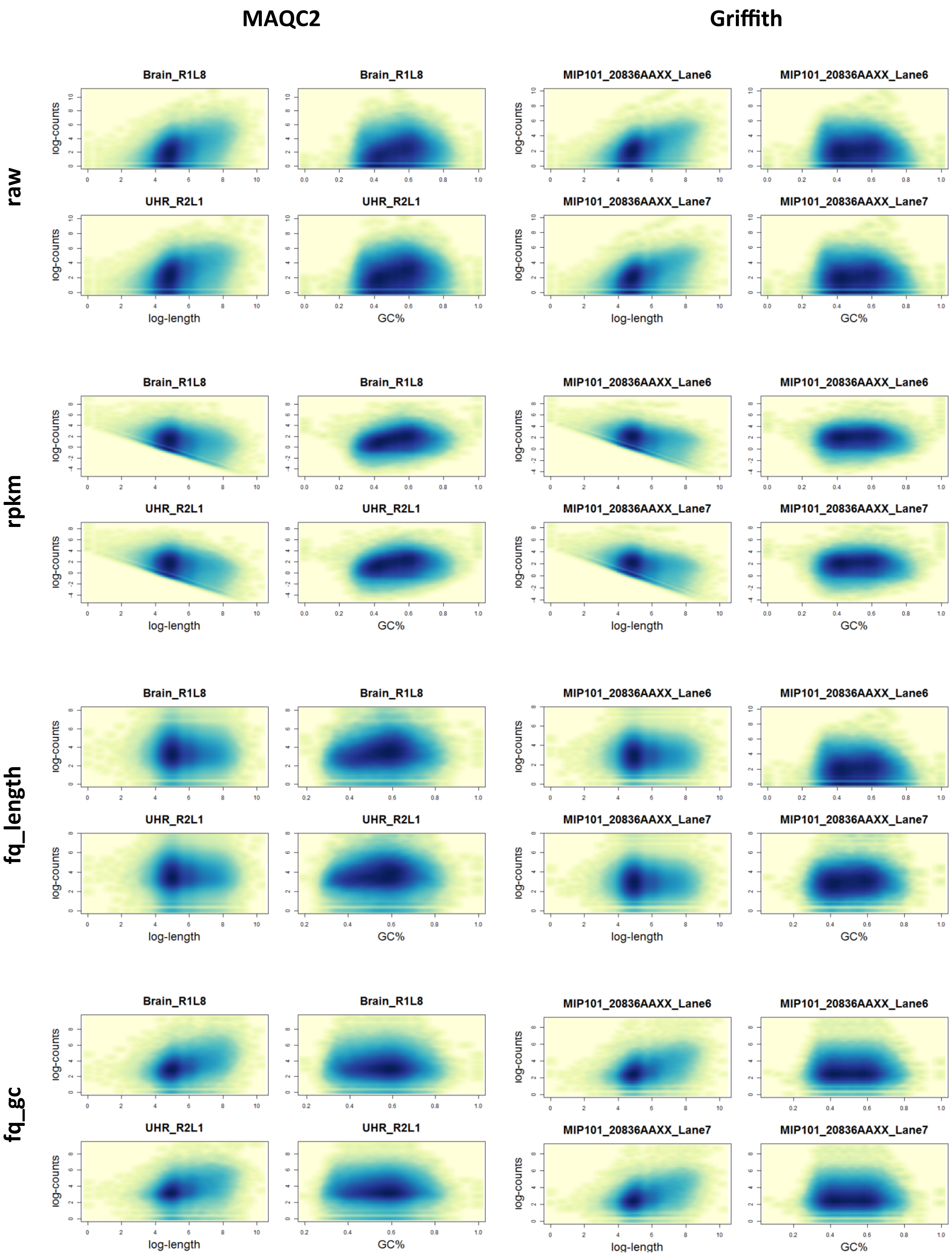


Figure 4.1: Smoothed scatter-plots showing dependence of counts over exon length and GC-content for two libraries of MAQC2 (“Brain_R1L8” and “UHR_R2L1”) and Griffith’s (“MIP101_20836AAXX_Lane6” and “MIP101_20836AAXX_Lane7”) data. Results are shown for counts before (*raw*) and after normalization with RPKM (*rpkm*), full-quantile on exon length (*fq_length*) and full-quantile on GC-content (*fq_gc*).

4.3 Library size normalization

The MA-plot is a diagnostic plot that can display, for each genomic feature such as an exon or gene, the difference of expression levels measured in two different samples (log-fold changes, or M-values) plotted by the average expression (A-values). M and A values for each pair of samples or libraries a, b are computed as follows:

$$M_i = \log(N_{ia}) - \log(N_{ib}) \tag{4.4}$$

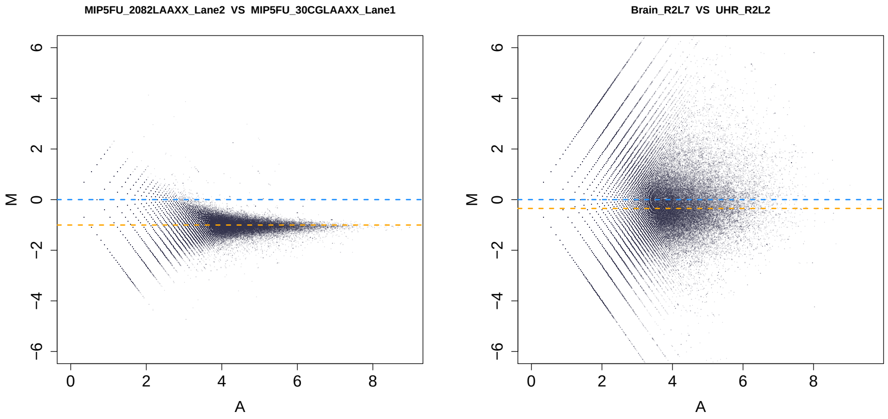
$$A_i = \frac{\log(N_{ia}) + \log(N_{ib})}{2} \tag{4.5}$$

where N_{ij} are the counts for exon i in library $j = a, b$.

Assuming that most of the exons are not DE, MA-plots should be centered at $M = 0$ and should show a relatively low dispersion, representing technical and biological variability, with the (few) points placed far from $M = 0$ representing truly DE exons. In fact, MA-plots of raw counts are more widely dispersed at low expression levels (i.e. low A-values), corresponding to low-abundance exons, which are more easily corrupted by technical noise (Figure 4.2). By inspecting the MA-plot for Griffith's data, it is easy to see the deviance of the median of M-values from zero, revealing the differences in terms of sequencing depths of the considered libraries (see the two technical replicates belonging to "MIP5FU" condition shown in Figure 4.2a). More interestingly, the bias in the MA-plot for MAQC2 data are due to the different composition of the sequenced libraries and depends on the transcriptional profile of the considered samples (see Figure 4.2b). In particular, a subset of exons characterized by higher expressions in "Brain" with respect to "UHR", that can be noticed on the upper part of MA-plot in Figure 4.2a, consumes a part of the available sequencing depth, resulting in a reduction of counts in Brain replicates (MA-plot in Figure 4.2b shifted towards negative values).

Figure 4.3 reports a synthetic representation of the M-values obtained by comparing the two groups considered in MAQC2, Griffith's and Jiang's data. Box-plots show the distributions of the medians of M-values, calculated for all the pairwise comparisons of technical replicates from the two different conditions assessed (replicates belonging to the same condition are not compared in this analysis). The medians of M-values calculated from raw data, which represent the differences between libraries due both to differentially expressed genes and to the sequencing depth, are shifted away from $M = 0$ in all the considered data sets. In particular, for MAQC2 and Jiang's data (Figure 4.3b and 4.3c) they strongly deviate from zero towards negative values. In Griffith's data (Figure 4.3a) M-values show a smaller deviation from zero but a larger variance, meaning that the

Figure 4.2: MA-plots for two libraries of Griffith's (a) and MAQC2 (b) data sets. The median of M-values is represented by the orange line.

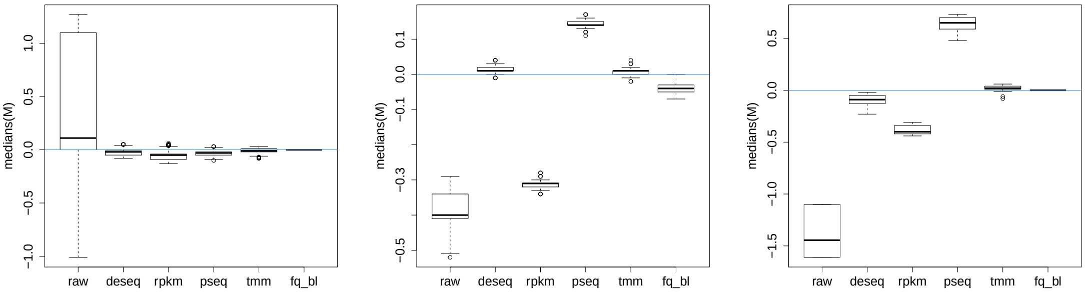


(a) Griffith

(b) MAQC2

differences are not due to a between-condition comparison, as in MAQC2 and Jiang's data, but to a technical bias, i.e. the difference in sequencing depth. These differences are corrected by all methods on Griffith's data: M-values are shifted towards zero and their variance is reduced. The difference in library sizes characterizing the other two data sets are more challenging and the normalizations perform differently. RPKM normalization shifts M-values towards zero, but without reaching $M = 0$. DESeq performs better than RPKM, shifting M-values very close to zero and reducing variance. On two data-sets, PoissonSeq over-corrects for library size, reversing M-values bias. TMM correctly sets M-values approximately to zero and strongly reduces variance. Between-lane full-quantile normalization, always minimizes variance as a result of data quantile normalization and shifts M-values towards zero. They are correctly set to $M = 0$ in all data sets except for MAQC2 data set. For all methods, library normalization is more challenging on MAQC2, probably for the high number of DE exons that can be detected when comparing two different tissues.

Figure 4.3: Box-plots of medians of M-values calculated for raw (*raw*) and normalized data of Griffith's (a), MAQC2 (b) and Jiang's data sets. RPKMs (*rpkm*), DESeq (*deseq*), PoissonSeq (*pseq*), TMM (*tmm*), and between-lane full-quantile (*fq_bl*) normalizations are considered.



(a) Griffith

(b) MAQC2

(c) Jiang

4.4 Effects of normalization on quantification and differential expression analysis

Besides the importance of identifying normalization methods that are able to reduce biases in count data, one of the most desired features of methods for RNA-seq data analysis is to provide a robust measure of RNA abundances and an unbiased estimation of differential gene expression between two or more conditions or groups.

We used Jiang's data set to assess how different normalization methods affect the quantification of spike-in RNAs. For all ERCC spike-in RNAs, the mean of counts (raw or normalized) across all technical replicates of the two samples was taken as measure of their abundance. These values were then compared with the true spike-in RNA concentrations reported in [144] (Figure 4.4A). Both within-lane full-quantile normalization on exon length or GC-content were used along with between-lane full-quantile normalization, to account for differences in sequencing depth. On "nucleus" samples, DESeq, PoissonSeq, TMM and between-lane full-quantile normalization lead to estimates similar to those obtained from raw data. RPKM normalization provides slightly better estimates, demonstrating the importance of accounting for sequence length when assessing transcripts relative abundances. However, the spike-in RNAs data set here considered does not allow to test the methods on very short transcripts, in which this bias is more difficult to correct (Figure 4.1). Within-lane full-quantile normalization over GC-content and sequence length obtains the lowest R^2 and correlation. Similar results are obtained on "cytosol" samples (results not shown).

With the aim of assessing the effect of normalization on DE analysis, we applied

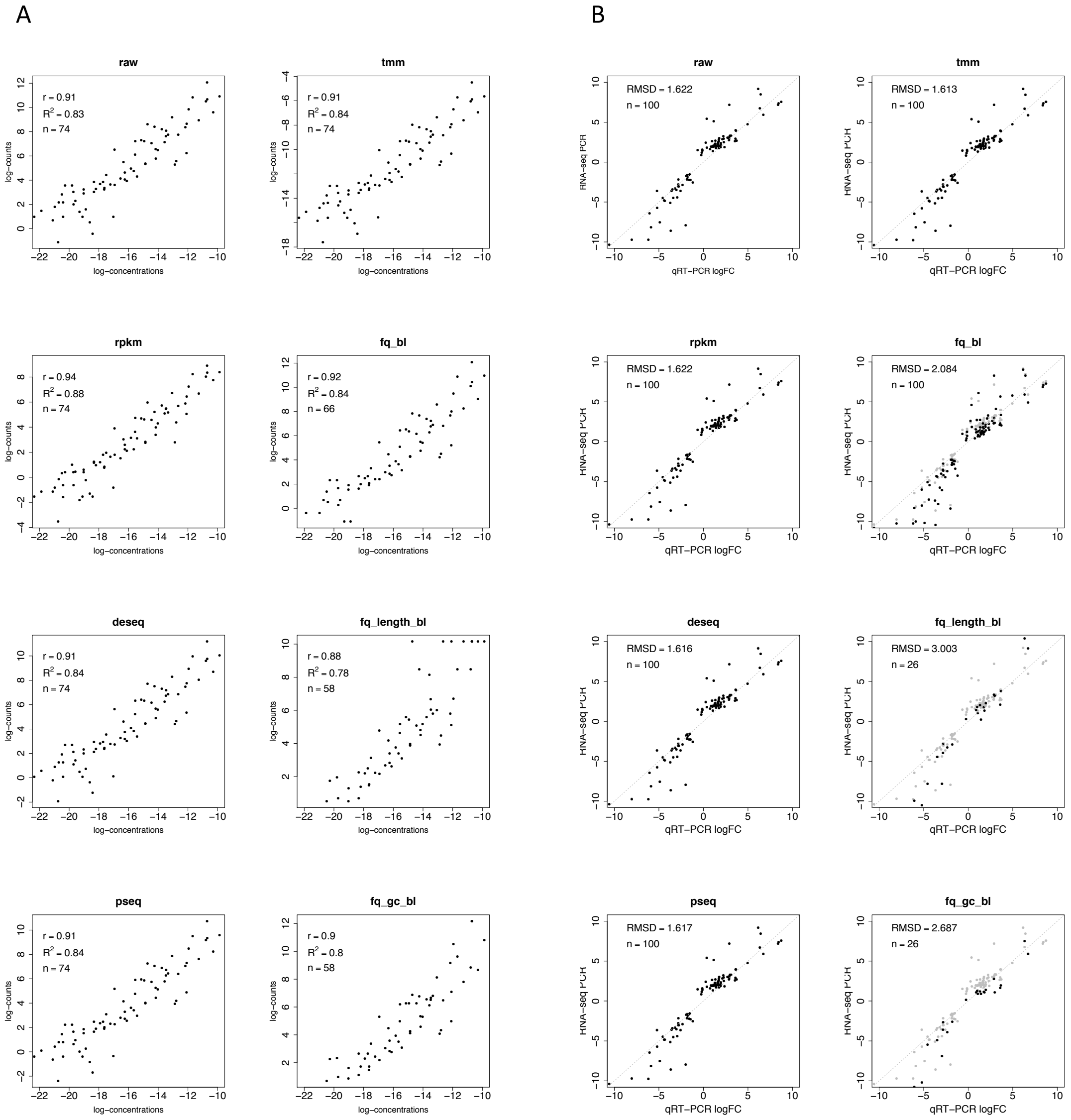


Figure 4.4: Benchmarking on spike-in RNAs quantification and log-fold-changes of exon expression, computed for raw data (*raw*) and counts normalized with different methods: RPKMs (*rpkm*), DESeq (*deseq*), PoissonSeq (*pseq*), TMM (*tmm*), between-lane full-quantile normalization (*fq_bl*), within-lane full-quantile normalization on exon length (*fq_length_bl*) and on GC-content (*fq_gc_bl*). Both within-lane full-quantile normalizations were coupled with between-lane full-quantile normalization. (A) Comparison of spike-in RNA true concentrations (“K562_nucleus” samples from Jiang’s data set) and RNA-seq counts on log-log scale. Coefficient of determination R^2, Pearson’s correlation r and total number of assayed spike-in RNAs n are also reported. (B) Comparison of log-fold-changes (logFC) estimated on Griffith’s data with qRT-PCR and RNA-seq. logFC obtained from raw data are represented in grey. Root-mean-square deviation (RMSD) of log-fold-changes and total number of assayed exons n are also reported.

edgeR to Griffith's data, normalized with the above mentioned approaches; The log-fold-changes estimated by edgeR from the two-groups comparison were then compared to those estimated by the authors in [143] through quantitative Real-Time PCR (qRT-PCR) (Figure 4.4B).

We used root-mean-square deviation (RMSD) of log-fold-changes as a measure of the differences between the values predicted from raw or normalized RNA-seq data and the values actually observed with qRT-PCR, to have an indication of the accuracy of the methods:

$$RSMD = \sqrt{\frac{\sum_{i=1}^{n} \left(logFC_{true}(i) - logFC_{est.}(i)\right)^2}{n}} \tag{4.6}$$

where $i = 1, ..., n$ represent all the assayed exons.

RPKM results are equivalent to that obtained with raw data, with the same RMSD. TMM and between-lane full-quantile normalization show a good agreement with qRT-PCR data, obtaining the lowest RMSD. Full-quantile normalizations lead to higher RMSD, but they cannot be directly compared to the previous ones, since they are computed on a smaller set of exons. DESeq, PoissonSeq and TMM obtain slightly better results than raw data, with TMM showing the lowest RSMD. However, when considering the full set of exons tested for DE with edgeR, it is worth nothing that, in almost all cases, p-values are strongly biased by exon length: longer exons tend to have lower p-values than shorter ones and are more likely to be selected as DE (Figure 4.5). This bias is corrected only by quantile normalizations, but it is not clear if it is due to the normalization itself or to the removal of low-counts genes. On the contrary, p-values do not depend on GC-content (results not shown).

To better identify the data features that challenge library size normalization and to assess the impact of normalization on DE analysis with a known benchmark, we simulated synthetic RNA-seq data sets as in section 3.2. In addition, we investigated whether the distribution of over-expressed genes between pairs of samples affects data normalization, by setting the percentage of over-expressed genes simulated in condition A, rather than in condition B, to be A=0.5 (i.e. equally distributed between the two conditions), 0.75 and 0.9 (i.e. predominantly present condition A). We tested DE expression with edgeR [131], measuring methods performance with sensitivity *vs.* precision curves and their area under curve (AUC). Since gene lengths were not simulated, RPMs (Reads per Million Mapped Reads, i.e. counts normalized by the total counts of each library) were used in place of RPKMs. The inspection of AUC plots (Figures 4.6 and 4.7) reveals an improvement

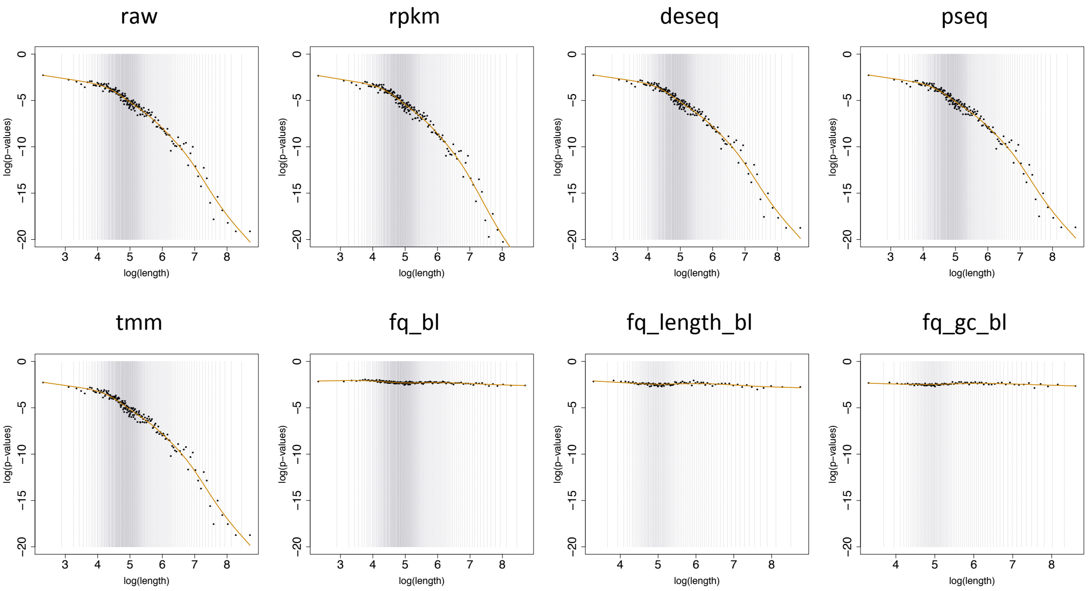


Figure 4.5: Relationship (log-log scale) between p-values obtained by edgeR, on Griffith's data, and exon length. Both raw and normalized data are considered: RPKMs (*rpkm*), DESeq (*deseq*), PoissonSeq (*pseq*), TMM (*tmm*), between-lane full-quantile normalization (*fq_bl*), within-lane full-quantile normalization on exon length (*fq_length_bl*) and on GC-content (*fq_gc_bl*). Exons were divided in equally-sized bins (2000 exons each) considering their log-length, and average p-values (log-scale) were computed for each bin.

of methods performance (i.e. higher AUC) when the percentage of DE genes increase from 10% to 20%, with few exceptions for A=0.75 and A=0.9. When considering raw data or RPMs, having a more balanced distribution of over-expressed genes (i.e. A closer to 0.5) leads to higher performance. On the contrary, the other normalizations obtain comparable or slightly better results as A approaches 0.9. TMM, DEseq and PoissonSeq normalizations are robust to changes in the assignment of over-expressed genes to the two conditions, with AUC values of TMM varying less than 11% in response to variations of A value. On average, TMM, DEseq and PoissonSeq show the best results, while AUC values of RPMs and full-quantile normalization are lower than those of raw data. For all normalization methods, the smallest the difference in DE expression is (lower b), the lower the AUC values are. However, TMM, PoissonSeq and DESeq performance, are less sensitive to this variation and in almost all simulations show better results than those obtained from raw data.

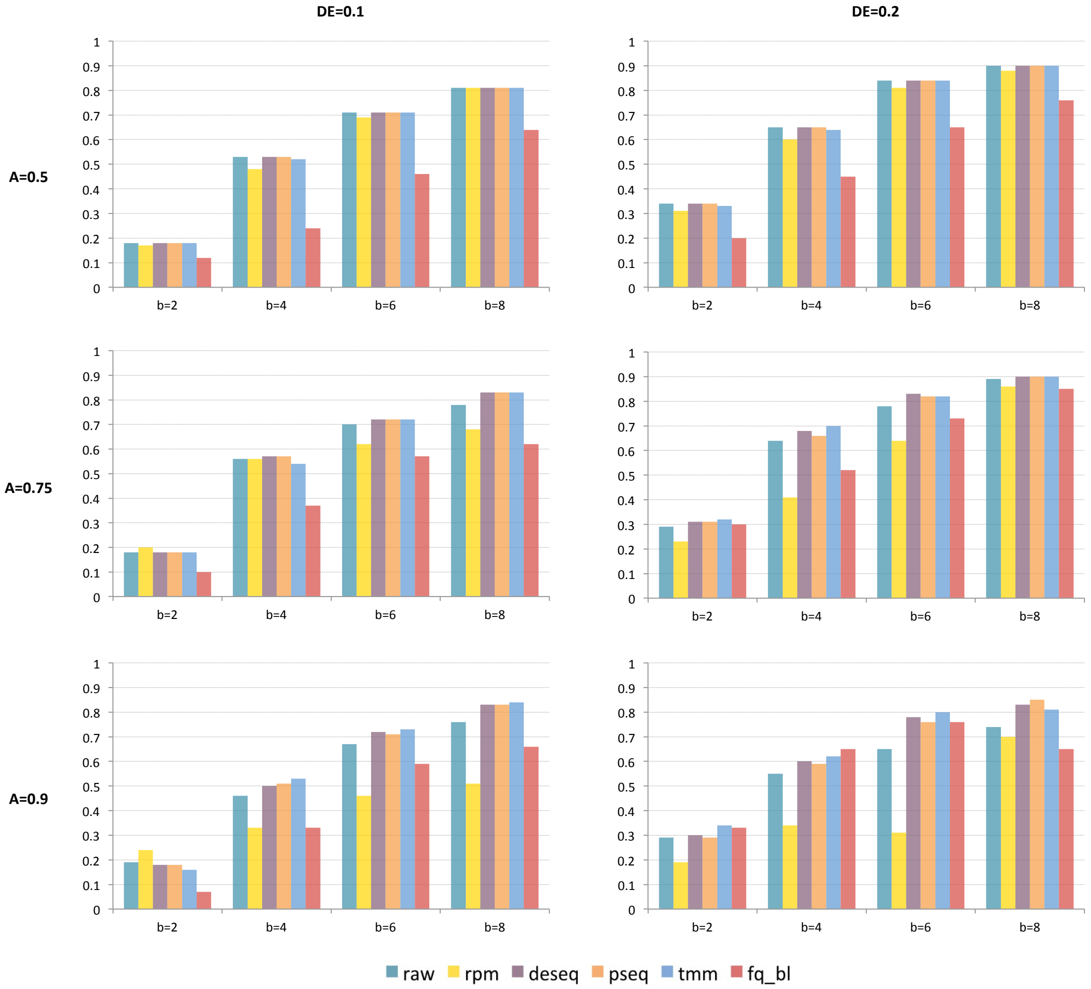
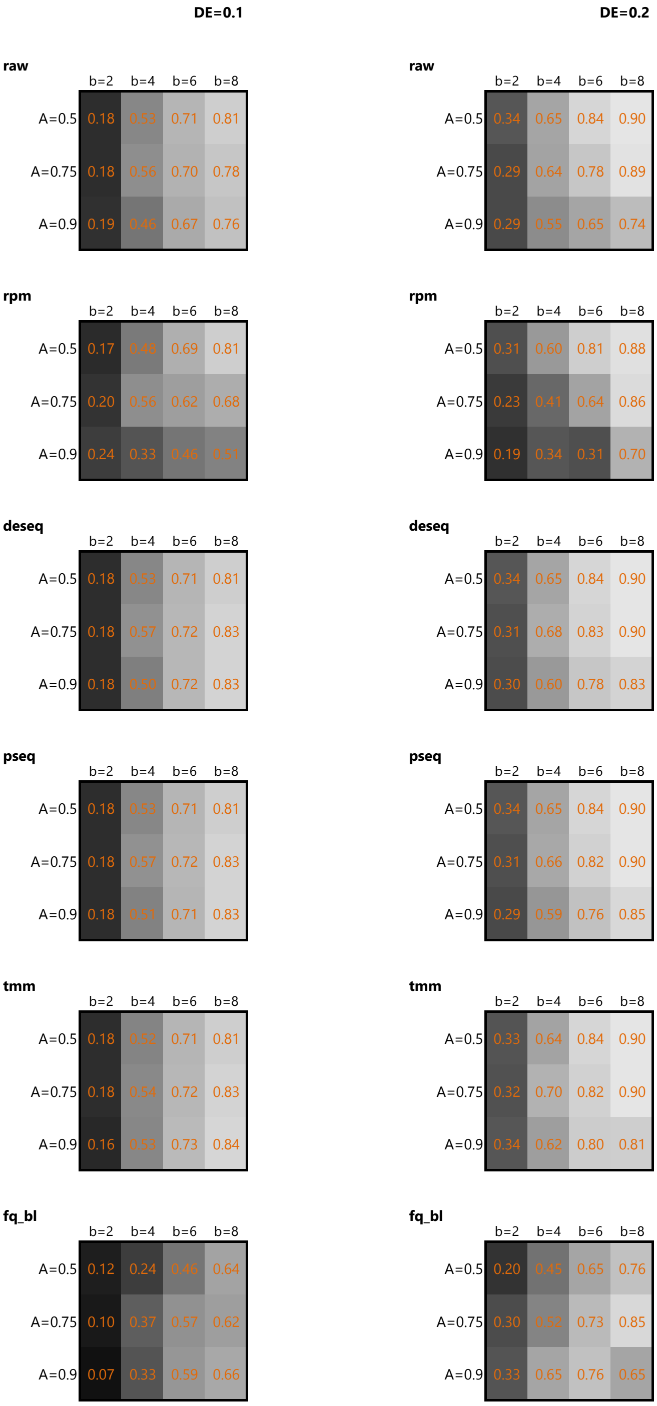


Figure 4.6: Bar-plots of Area Under Curve (AUC) calculated from sensitivity *vs.* precision curves on simulated count data, before (*raw*) and after normalization with RPM (*rpm*), DESeq (*deseq*), PoissonSeq (*pseq*), TMM (*tmm*) or between-lane full-quantile (*fq_bl*). Results are reported in dependence of the features of the simulated data: extent of differential expression (b=2, 4, 6, 8), percentage of DE genes (DE=0.1, 0.2) and percentage of DE genes that are over-expressed in condition A (A=0.5, 0.75, 0.9).



DE=0.1

raw

	b=2	b=4	b=6	b=8
A=0.5	0.18	0.53	0.71	0.81
A=0.75	0.18	0.56	0.70	0.78
A=0.9	0.19	0.46	0.67	0.76

rpm

	b=2	b=4	b=6	b=8
A=0.5	0.17	0.48	0.69	0.81
A=0.75	0.20	0.56	0.62	0.68
A=0.9	0.24	0.33	0.46	0.51

deseq

	b=2	b=4	b=6	b=8
A=0.5	0.18	0.53	0.71	0.81
A=0.75	0.18	0.57	0.72	0.83
A=0.9	0.18	0.50	0.72	0.83

pseq

	b=2	b=4	b=6	b=8
A=0.5	0.18	0.53	0.71	0.81
A=0.75	0.18	0.57	0.72	0.83
A=0.9	0.18	0.51	0.71	0.83

tmm

	b=2	b=4	b=6	b=8
A=0.5	0.18	0.52	0.71	0.81
A=0.75	0.18	0.54	0.72	0.83
A=0.9	0.16	0.53	0.73	0.84

fq_bl

	b=2	b=4	b=6	b=8
A=0.5	0.12	0.24	0.46	0.64
A=0.75	0.10	0.37	0.57	0.62
A=0.9	0.07	0.33	0.59	0.66

DE=0.2

raw

	b=2	b=4	b=6	b=8
A=0.5	0.34	0.65	0.84	0.90
A=0.75	0.29	0.64	0.78	0.89
A=0.9	0.29	0.55	0.65	0.74

rpm

	b=2	b=4	b=6	b=8
A=0.5	0.31	0.60	0.81	0.88
A=0.75	0.23	0.41	0.64	0.86
A=0.9	0.19	0.34	0.31	0.70

deseq

	b=2	b=4	b=6	b=8
A=0.5	0.34	0.65	0.84	0.90
A=0.75	0.31	0.68	0.83	0.90
A=0.9	0.30	0.60	0.78	0.83

pseq

	b=2	b=4	b=6	b=8
A=0.5	0.34	0.65	0.84	0.90
A=0.75	0.31	0.66	0.82	0.90
A=0.9	0.29	0.59	0.76	0.85

tmm

	b=2	b=4	b=6	b=8
A=0.5	0.33	0.64	0.84	0.90
A=0.75	0.32	0.70	0.82	0.90
A=0.9	0.34	0.62	0.80	0.81

fq_bl

	b=2	b=4	b=6	b=8
A=0.5	0.20	0.45	0.65	0.76
A=0.75	0.30	0.52	0.73	0.85
A=0.9	0.33	0.65	0.76	0.65

Figure 4.7: Visualization of Area Under Curve (AUC) calculated from sensitivity *vs.* precision curves on simulated count data, before (*raw*) and after normalization with RPM (*rpm*), DESeq (*deseq*), PoissonSeq (*pseq*), TMM (*tmm*) or between-lane full-quantile (*fq_bl*). For each methods, results are reported in dependence of the features of the simulated data: extent of differential expression (b=2, 4, 6, 8), percentage of DE genes (DE=0.1, 0.2) and percentage of DE genes that are over-expressed in condition A (A=0.5, 0.75, 0.9). AUC value are also visualized as coloured cells, from 0 (black) to 1 (white).

4.5 Remarks on RNA-seq counts normalization

We performed an assessment of the main biases of RNA-seq counts, considering both differences in library sizes and biases due to sequence-specific covariates, such as exon length and GC-content variability. We investigated these biases before and after normalization, considering some of the most used methods for RNA-seq data normalization. In addition, we assessed the impact of data normalization on downstream analyses, such as RNA quantification and differential expression detection. In our comparative analysis, library size normalization with DESeq and TMM leads to the best results in all the assessments. In particular, we found that TMM shows stable results in different real data sets and simulation settings, confirming previous results [153]. In our assessment, GC-content seems to have weak relationship with exon counts. On the contrary, counts show a marked exon length bias, which is also propagated downstream to RNA quantification and DE analysis, where longer exons are more likely to be detected as DE. None of the tested methods effectively removes this bias from counts although they ensure correct estimates of expression or differential expression at the same time. RPKM, which is currently one of the most used normalization approaches, provides accurate estimates of transcript abundances, as reported in recent studies [154]. However, as previously noted in [153, 102, 50], it is ineffective for DE analysis, i.e. when multiple conditions are compared. In particular, Dillie *et. al* suggest to abandon RPKM normalization in the context of differential analysis. This debate could be probably settled by assessing methods in dependence of a single and specific RNA-seq application, such as RNA quantification or differential expression ananlysis. However, the definition of an accurate measure of gene expression from RNA-seq data is essential to ensure a correct interpretation of results, independently from the downstream application. In addition, it represents a fundamental prerequisite for analyzing RNA-seq time-series data, in which expression levels should be correctly compared within- and between-samples.

5
A novel approach to compute counts: *maxcounts*

Our assessment of RNA-seq normalization methods (Chapter 4) confirmed the presence of a length bias in count data, which is not efficiently corrected by any of the considered methods and propagates to downstream analysis of gene expression and differential expression. Moreover, other biases due to highly expressed genes and uneven read coverage characterize RNA-seq counts (see a review in section 2.3).

In this chapter, we propose a novel method for computing counts, called *maxcounts*, with the aim of reducing these biases. Once the reads have been aligned to a feature of interest (exon or single-isoform transcript), we exploit read coverage to obtain counts for every position in its sequence and we quantify its expression as the maximum of its "positional" counts. We assess maxcounts performance in comparison with the standard approach, which considers the total number of reads mapped on an exon (called *totcounts* from here on). With this purpose, we consider the three real human data sets described in Chapter 3. Data were sequenced with single- and paired-end protocols, and have different characteristics, which allow us to test our approach with respect to different features. In particular, the assessment is performed considering some desirable features that a measure of gene expression should have: *(i)* being independent of gene-specific covariates such as transcript length and GC-content; *(ii)* being unbiased towards highly expressed genes; *(iii)* being an accurate estimate of the true gene expression levels; *(iv)* showing low technical variance; *(v)* being robust to possible variations in the quality of alignments. In the following, we present the results of a study that we recently published [155], in which we assessed the properites cited above for both maxcounts and totcounts.

5.1 Definition of the analysis pipeline

FASTQ files of raw reads were obtained as in Chapter 4. From Jiang *et al.* study, we considered a subset of samples from the K-562 cell line, extracted from nucleus or whole-cell. In particular, we considered two libraries from the "cell" group [SRA: SRR307930, SRR307931] and six libraries from the "nucleus" group [SRA: SRR317042, SRR317043].

We defined and implemented a pipeline to pre-process and map reads, and to discard low-similarity alignments and *multireads* (i.e. reads mapping to multiple positions of the reference). The analysis pipeline implemented in this study is depicted by the flowchart of Figure 5.1 (a simplified version of the same pipeline was applied to single-end data).

In the first phase, reads were pre-processed to remove adapter sequences and read ends whose Phred quality was lower than 20, and to discard reads whose length after trimming was less than 33bp. Output FASTQ files were re-formatted to recover the correspondence of paired-end reads, and to store in a separate file the singleton reads, whose mate was discarded during pre-processing.

The pre-processing and re-mating steps were implemented using the FASTX Toolkit [156], version 0.0.13.2, and custom made scripts written in bash and Perl. The effectiveness of the pre-processing phase was assessed using FastQC [157], version v0.10.1.

Pre-processed paired-end and singletons reads were mapped with TopHat v2.0.6 [55], with default parameters settings, in a two-steps procedure. First, paired-end reads were mapped on the reference sequence to generate a BAM file of alignments and a file of junctions. Then, singletons were mapped with TopHat exploiting the information provided by junctions (`-j` option). The expected distance between paired-end reads was estimated by using PASS [158] as in [110]. Alignment files from paired- end and singleton reads were finally merged in a single BAM file using the `merge` utility of samtools [159]. In the last phase of post-processing, a filtered set of alignments was obtained after discarding multireads and reads whose similarity with the reference was lower than 97%. This analysis was performed using SAMsieve, a java alignment-filtering program developed in house based on SAMtools API. It can filter alignments stored in SAM or BAM files based on several criteria, such as number of alignments reported, alignment quality, chromosome, number of mismatches, read coverage, percentage of identity,

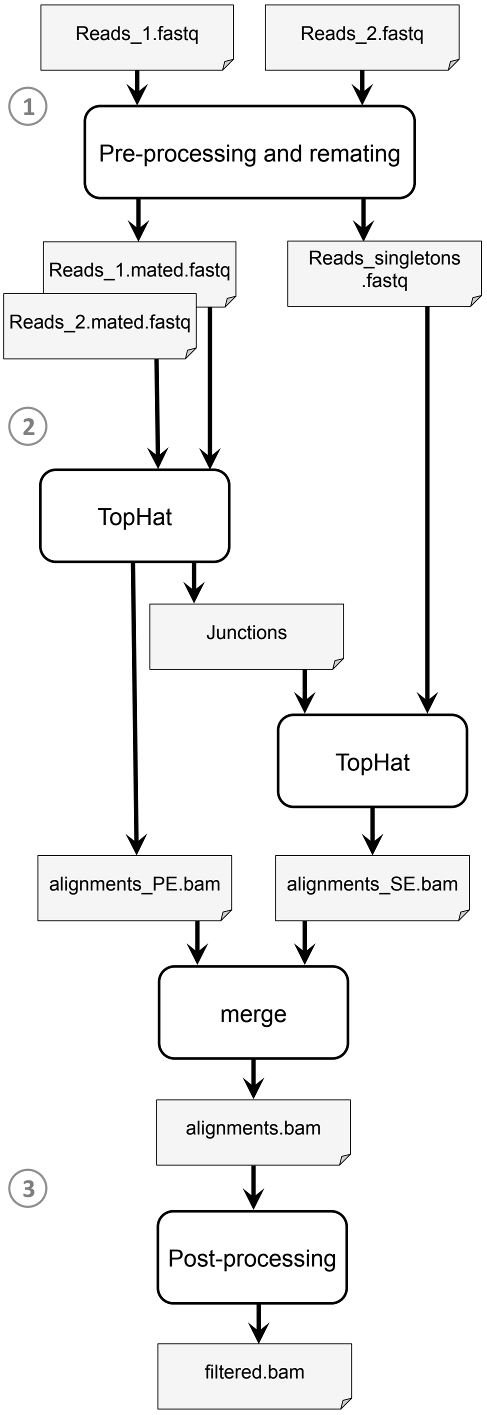


Figure 5.1: Analysis pipeline. 1) Read pre-processing (trimming and filtering) and re-mating of paired-end reads. 2) Separate mapping of paired-end and singleton reads, and merging of alignments. 3) Removal of low similarity alignments and multireads.

etc. In this work, to discard multireads and reads whose similarity with the reference was lower than 97%, the following options were used: `reportedAligns4read<=,1` and `identity>=0.97`.

Computation of counts and normalization

Totcounts were computed using the `coverageBed` utility of bedtools [148], option `-counts`.

For each exon i in library j, maxcounts M_{ij} were computed as

$$M_{ij} = \max(N_{ijsp}) \tag{5.1}$$

where, N_{ijp} is the number of reads covering position p along exon i.

We implemented a new functionality for bedtools' routine `coverageBed` that allows the computation of maxcounts through two new options: `-max` and `-maxm` (the latter allows the user to select the number of exon positions, among those with the highest read coverages, to be used in the computation of maxcounts). This functionality is available as a patch for bedtools that can be downloaded from [160]. The code is distributed along with `formatCounts.sh`, a bash script for creating a matrix of counts starting from multiple files, from different libraries, generated with coverageBed (options: `-counts`, `-max`, `-maxm`). Positional counts along spike-in RNAs were computed using coverageBed, option `-d`.

We considered the GRCh37.p10 human reference genome and the Gencode human exon annotation (version 12), for mapping, computing counts and gathering information about GC-content.

Exons (or spike-in transcripts of Jiang's data set) with average totcounts or maxcounts across replicates lower than 0.5, were discarded from our analysis. For both maxcounts and totcounts, differences in library sizes between replicates were corrected through Trimmed Mean of M-values (TMM) normalization [104] using edgeR [131]. In the following, we will refer to TMM-normalized totcounts and maxcounts simply as "totcounts" and "maxcounts", respectively.

Although providing an assessment of normalization methods is beyond the scope of the present work, we acknowledge that length bias can be corrected through normalization. Thus, to guarantee a fair comparison with current standards, we applied, when necessary, two normalization approaches: RPKM [45], which is widely used in RNA-seq studies, and within-lane full-quantile normalization, using exon length as covariate, since it has been proposed as preferred method in a recent work by Risso *et al.* [98]. RPKMs for each exon

e in library z were calculated as in Chapter 4.

Within-lane full-quantile normalization of counts on exon length was performed using EDASeq [98]. In order to correct for differences in library sizes, this normalization was used together with between-lane full-quantile normalization, also implemented in EDASeq. In this work we consider exons instead of genes or transcripts as we intend to evaluate the different summarization methods described above without biases, possibly introduced by the choice of a transcription model (e.g. how overlapping genes or alternative spliced exons are considered).

5.2 Investigation of count bias

Length bias and GC-content effect

To investigate exon length bias, we used smoothed scatter-plots of counts (averaged across replicates) versus exon length (Figures 5.2-5.7, panel A).

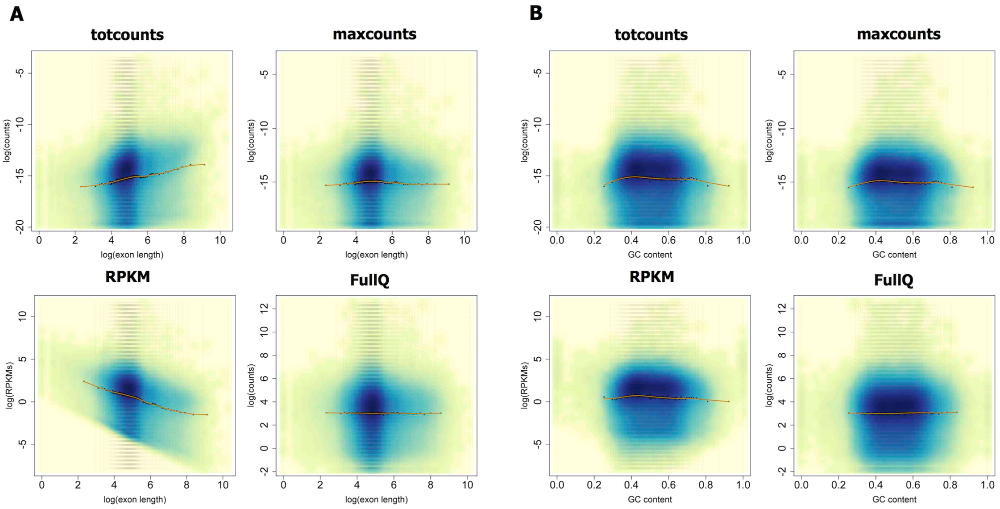


Figure 5.2: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in Jiang's data ("nucleus" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (RPKM) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

In all data sets, plots show an increasing trend of totcounts as exon length increases (see the cubic-spline fit represented by the orange line), revealing that longer exons tend to have higher counts than shorter ones. This bias is reduced, but not completely removed,

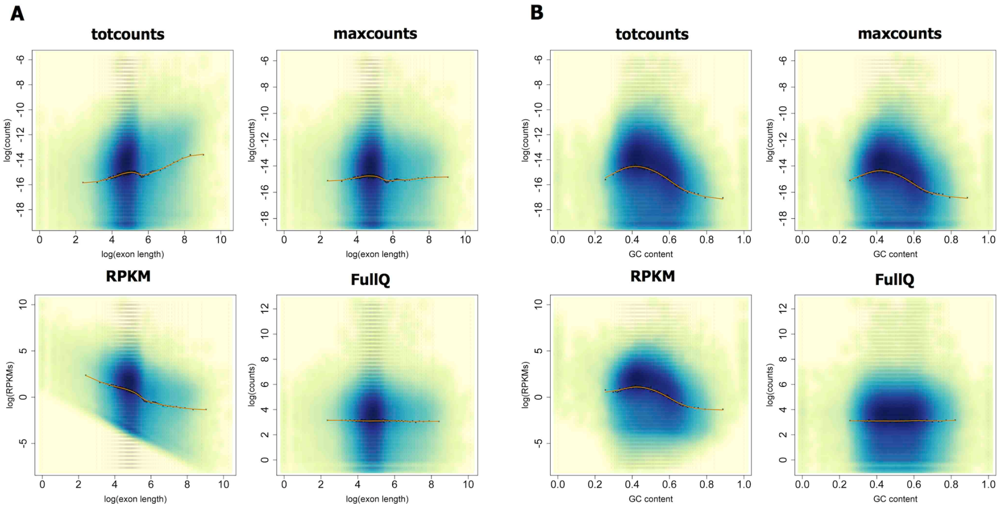


Figure 5.3: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in Jiang's data ("cell" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

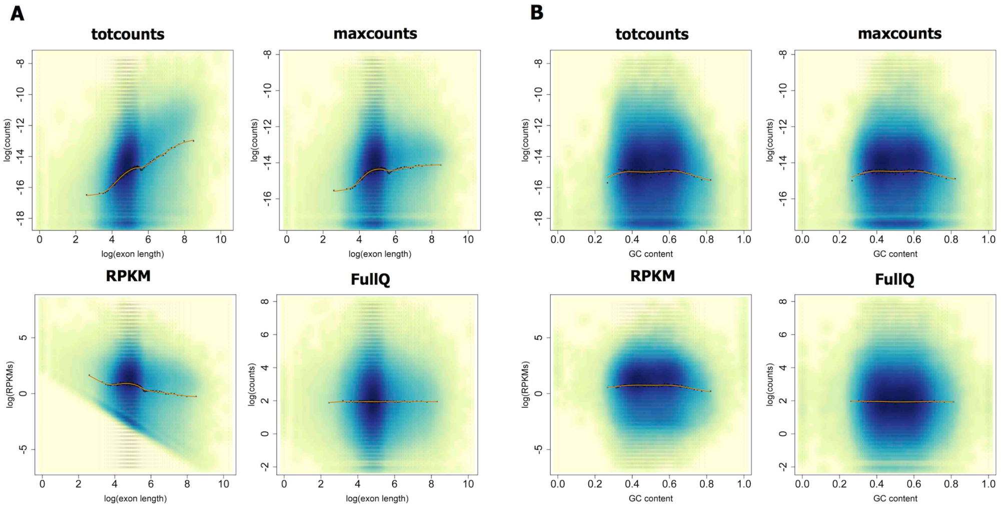


Figure 5.4: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in Griffith's data ("MIP5FU" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

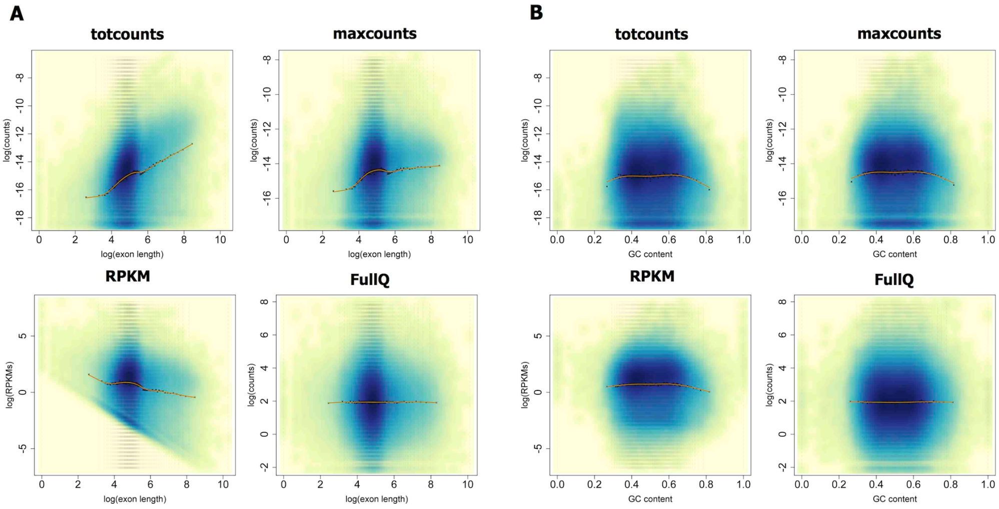


Figure 5.5: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in Griffith's data ("MIP101" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

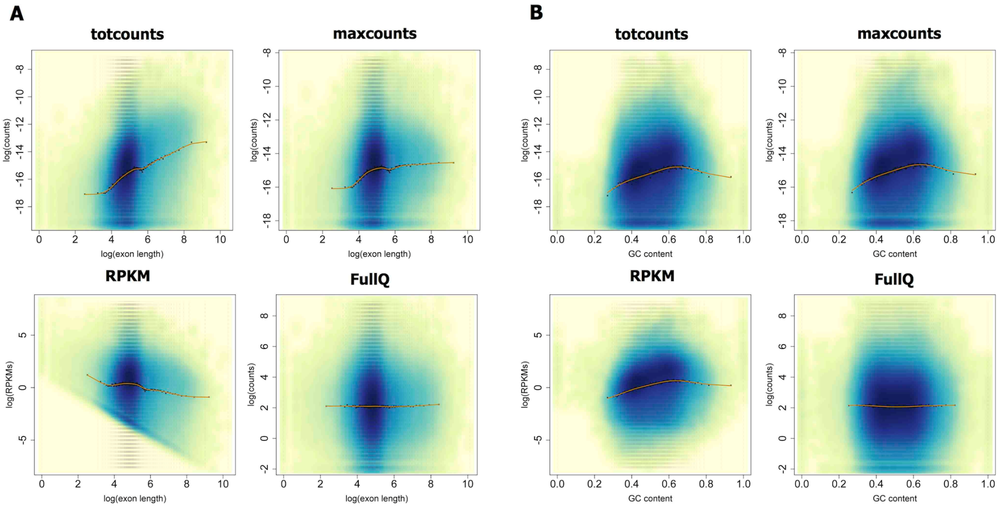


Figure 5.6: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in MAQC data ("UHR" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

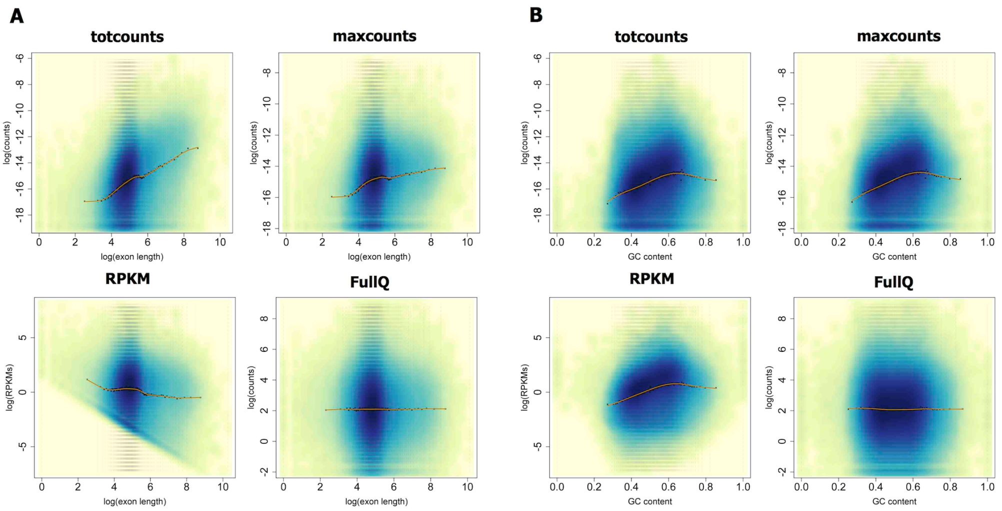


Figure 5.7: Smoothed scatter-plots showing the relationship between log-counts/RPKMs and exon length (log scale, A) or GC-content (B), in MAQC data ("Brain" libraries). The orange curve represents a cubic-spline fit computed on log-counts, averaged in bins of 5000 exons each (black crosses between vertical lines, indicating bin bounds). Counts or RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*), and averaged across libraries.

in maxcounts. Plots for Jiang's data ("nucleus" libraries), depicted in Figure 5.2A, show no dependency of maxcounts on exon length. Conversely, for maxcounts in Griffith's and MAQC2 data sets a slight under-representation of exons shorter than 50bp is still visible. We believe this behavior is explained by the difference in read length among the three data sets and the ability of TopHat to map them on splice junctions. Indeed, we observed that in MAQC2 and Griffith's data sets (36bp reads) only 0.25-0.50% of aligned reads are mapped on splice junctions, as opposed to 2.5-11.5% of reads in Jiang's data set (75bp reads). As a consequence, there is a reduction of counts over exons boundaries, which mainly affects short exons. In all the considered data sets, RPKM-normalized totcounts show a negative relationship with exon length due to an over-correction for length bias on short exons. On the opposite, full-quantile normalization completely removes exon length bias. Similarly, if applied to maxcounts, full-quantile normalization completely removes exon length bias even on short exons (plots not shown).

We used the same approach to investigate GC-content effect, revealing a moderate bias due to GC-composition on almost all data sets (Figures 5.2-5.7, B). As noted in previous studies, GC-content effect is not consistent across data sets [99, 98, 112, 85]. Interestingly, the correction for exon length bias via full-quantile normalization also corrects for GC-content bias all the considered data sets. In the following assessments, we always show raw totcounts and their RPKM- and full-quantile-normalized versions. Given the low length bias characterizing maxcounts, we consider their raw, not-normalized version.

Effect of highly expressed genes

We assessed the distribution of raw counts to detect possible biases due to highly transcribed genes, which may affect detection power of differentially expressed exons [60, 117].

As evident from Table 5.1 and Figure 5.8, we confirm that most of the reads are generated by a small subset of highly expressed genes.

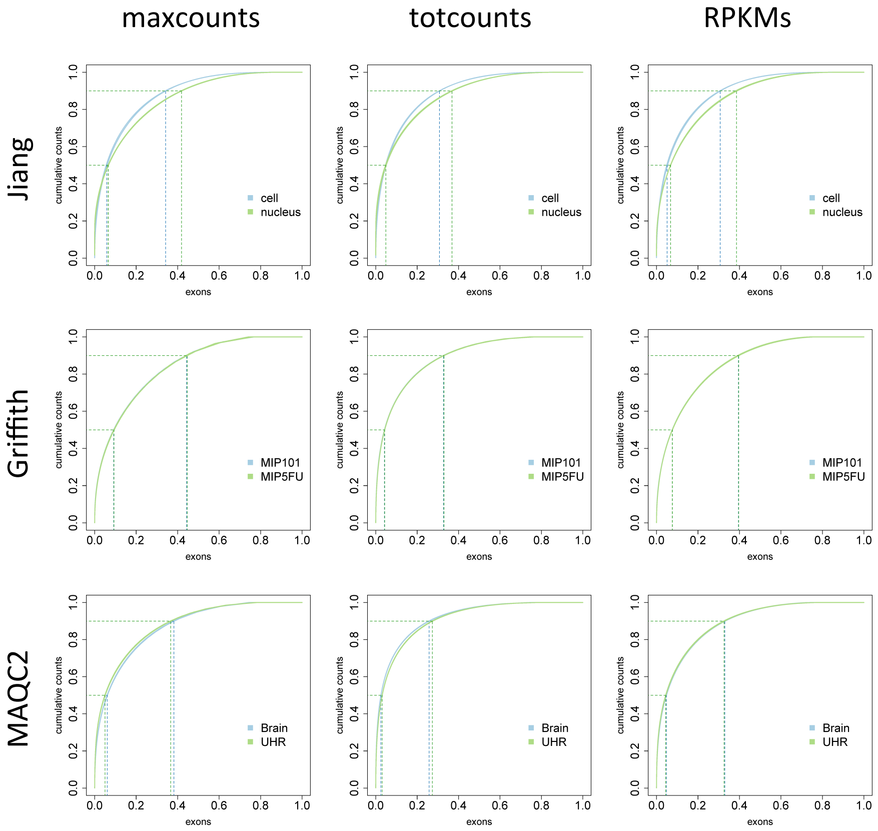


Figure 5.8: Distribution of maxcounts, totcounts and RPKM-corrected totcounts (*RPKM*) across exons, in Jiang's, Griffith's and MAQC2 data sets. Plots represent cumulative counts/RPKMs (y-axis, percentage referred to total counts/RPKMs in a library) assigned to exons (x-axis, percentage referred to the number of exons with more than zero counts/RPKMs). Each curve represents a different library and different colours identify different groups. Dashed lines represent 50% and 90% of total counts/RPKMs and are summarized in Table 5.1.

In particular, Table 5.1 reports the percentage of exons accounting for 50% and 90% of total counts or RPKMs in a sample, highlighting that less than 40% of exons contains more than 90% of all totcounts in a library. RPKM-normalized totcounts are more evenly distributed across exons, but the least biased distribution is that of maxcounts, with a marked improvement on the more biased data sets (see, for example, how this bias is reduced on Griffith's data).

Table 5.1: Summary of the distributions of maxcounts, totcounts and RPKM-corrected totcounts (*RPKM*) across exons in Jiang's, Griffith's and MAQC2 data sets. Table reports the percentage of exons accounting for 50% and 90% of total counts/RPKMs (average values across libraries belonging to the same condition).

Data set	Group	Counts [%]	Exons [%]		
			maxcounts	totcounts	RPKM
Jiang	cell	50	6	5	5
		90	34	31	31
	nucl	50	7	5	7
		90	42	37	39
Griffith	MIP101	50	9	4	8
		90	44	33	40
	MIP5FU	50	9	4	8
		90	45	33	40
MAQC2	Brain	50	6	3	5
		90	38	26	33
	UHR	50	5	3	4
		90	37	27	33

5.3 Quantification of spike-in RNAs

We estimated the abundances of spike-in RNAs on Jiang's data averaging totcounts and maxcounts across all technical replicates within each group (Figure 5.9).

For all measures, plots show higher agreement with the gold-standard on Jiang's "nucleus" data, probably because of the higher number of replicates (six libraries) compared to "cell" data (two libraries). All measures, with the exception of full-quantile-normalized totcounts, obtain high correlation with true concentrations, with RPKM-normalized totcounts and maxcounts having slightly better results than totcounts. Full-quantile normalization performed on totcounts, although eliminates length bias, possibly overcorrects data. Correlations with true concentrations of maxcounts, totcounts and RPKM-normalized totcounts, computed on all libraries of Jiang's data set, do not significantly differ (two-sided t-test, p-value > 0.05). On the contrary, full-quantile-normalized totcounts present the lowest correlation with spike-in RNAs concentrations (two-sided t-test,

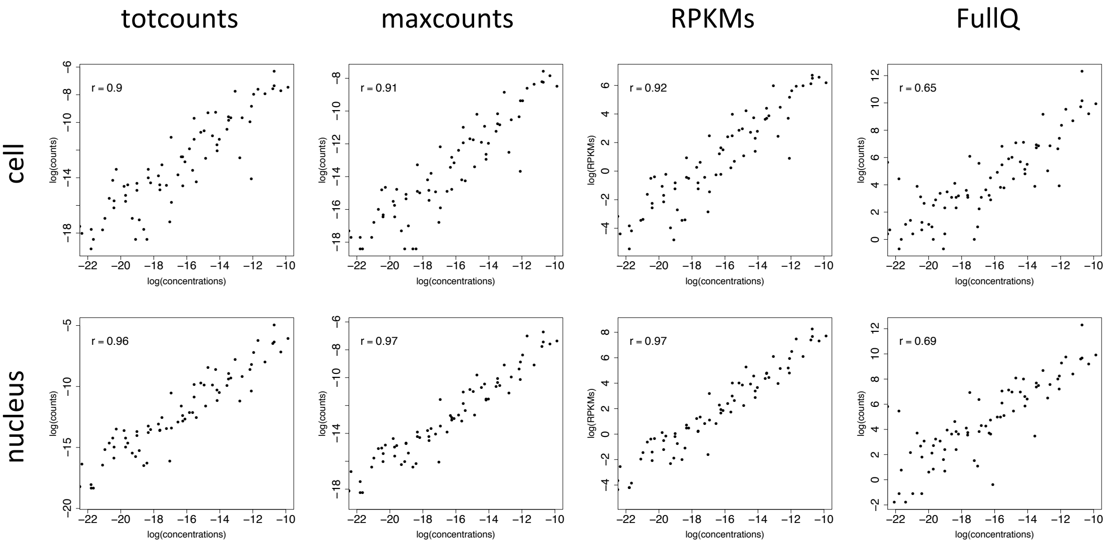


Figure 5.9: Counts/RPKMs obtained for spike-in RNAs from Jiang's data set, "cell" and "nucleus" groups, plotted against true concentrations (log-log scale). Counts/RPKMs are computed using totcounts, maxcounts, RPKM-corrected totcounts (*RPKM*) and totcounts corrected with within-lane full-quantile normalization over exon length (*FullQ*)."*r*" indicates Pearson's correlation (p-values always <1e-11).

p-value < 1e-10). All methods do not depend on transcript abundances, except for full-quantile-normalized totcounts, which are less robust in estimating low-abundance transcripts (results not shown).

Jiang's data set is particularly interesting because it allows the investigation of the non-uniformity of read coverage along spike-in RNAs (Figure 5.10), which was also reported in previous studies [101, 144]. Changes in read coverage are not justified by alternative splicing since spike-in RNAs are single-isoform, and show reproducible patterns on the same transcript sequenced in different libraries and conditions. As previously noted by Li *et al.* [101] reads are not randomly sequenced from transcripts, but some positions present a larger "sequencing preference" and result in higher (positional) counts.

Figure 5.10 highlights differences in read coverage along two transcripts having very similar concentrations, ERCC-00033 (7.06e-07 nmol/μl) and ERCC-00046 (7.08e-07 nmol/μl), with the latter having a more uniform coverage. To have a measure of how much those patterns affect maxcounts and totcounts quantification (for which an overall comparison is given in the previous paragraph), we can compute the variation of maxcounts/totcounts estimates on these two transcripts as:

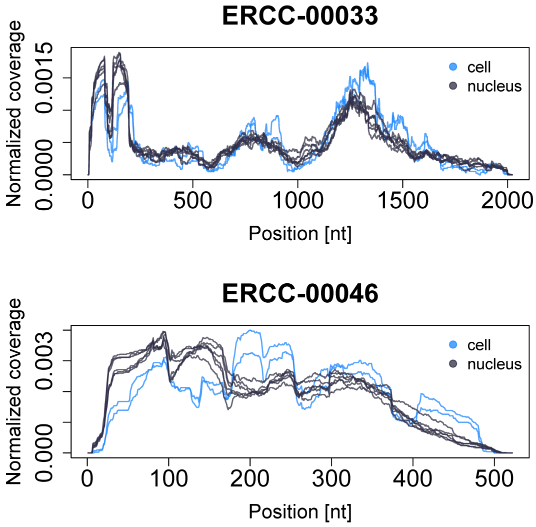


Figure 5.10: Read coverage (or "positional counts") along two spike-in RNAs, ERCC-00033 and ERCC-00046, in Jiang's libraries. "Cell" and "nucleus" replicates are indicated with blue and grey curves, respectively. Read coverage for each library is normalized to its sequencing depths.

$$\Delta = \frac{N_{33} - N_{46}}{N_{33} + N_{46}} \cdot 100 \tag{5.2}$$

where N_i are totcounts or maxcounts, averaged across libraries, for each transcript here considered. Ideally, Δ should be very small, to reflect the closeness of the true concentrations. Whereas totcounts produce a variation of 39%, maxcounts have a much smaller variation of 2%, overcoming read-coverage bias and providing very similar estimates for the transcripts here used as example. It is interesting to note that both transcripts show a reduced read coverage in correspondence to 3' end (Figure 5.10), a bias that is introduced during the reverse-transcription step performed with random hexamers (see Chapter 2). This bias is present in all transcripts of Jiang's data set (results not shown). Maxcounts approach is robust to 3' bias since it considers the bases with the highest read coverage along transcripts.

5.4 Count variance across technical replicates

To compare variance of totcounts (and its normalized versions) versus maxcounts we quantized the estimated average expression intensities in intervals of equal size and, for each interval, we calculated the average intensity and the average variance as explained in [135]. Finally we fitted data using a cubic spline (Figures 5.11, 5.12 and 5.13).

Maxcounts show the lowest variance at low and mean expressions, while totcounts present slightly lower variance at high expressions. In order to account for differences in the range of values, we also considered the coefficient of variation (CV), i.e. the ratio between the standard deviation and the mean. Totcounts and maxcounts obtain comparable CV curves. Totcounts normalized with full-quantile are characterized by larger variance and CV with respect to both maxcounts and totcounts, while totcounts normalized with RPKM-normalized totcounts have the highest variance and CV.

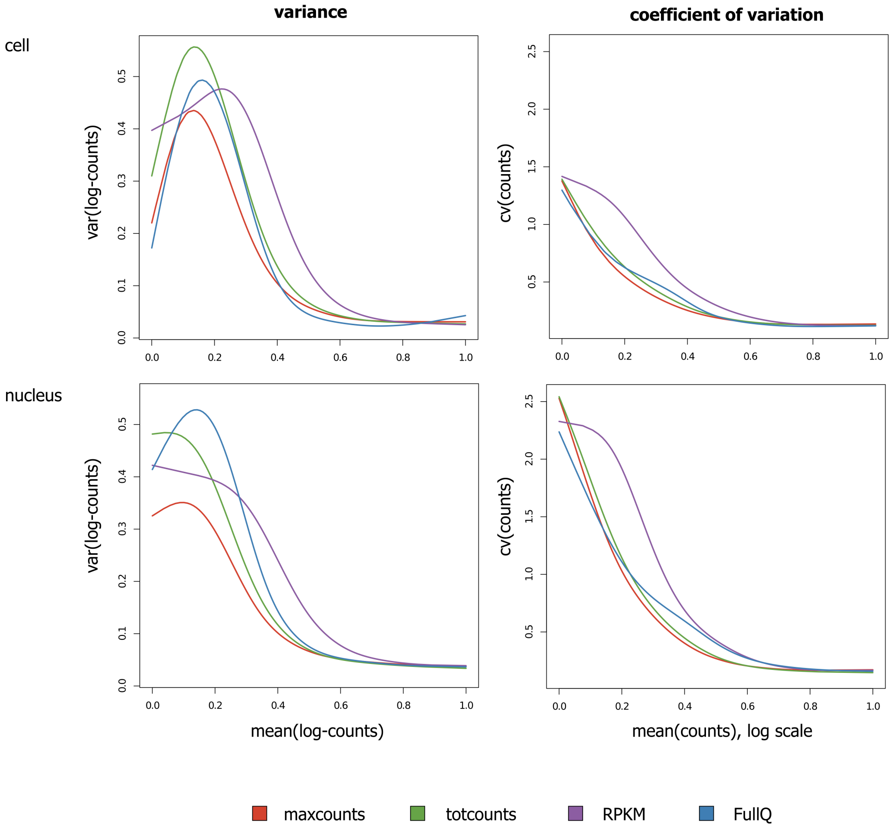


Figure 5.11: Variance and coefficient of variation (CV) of Jiang's data: variance *vs.* mean of log-counts/RPKMs (left plots) and CV vs. log-mean of counts/RPKMs (right plots). Curves represent cubic-spline fits computed on variance/CV, averaged in bins of 5000 exons each. Since maxcounts, totcounts, and totcounts normalized with RPKM (*RPKM*) and within-lane full-quantile normalization over exon length (*FullQ*) approaches are compared, x-values are scaled to cover the range [0, 1] in order to make them comparable.

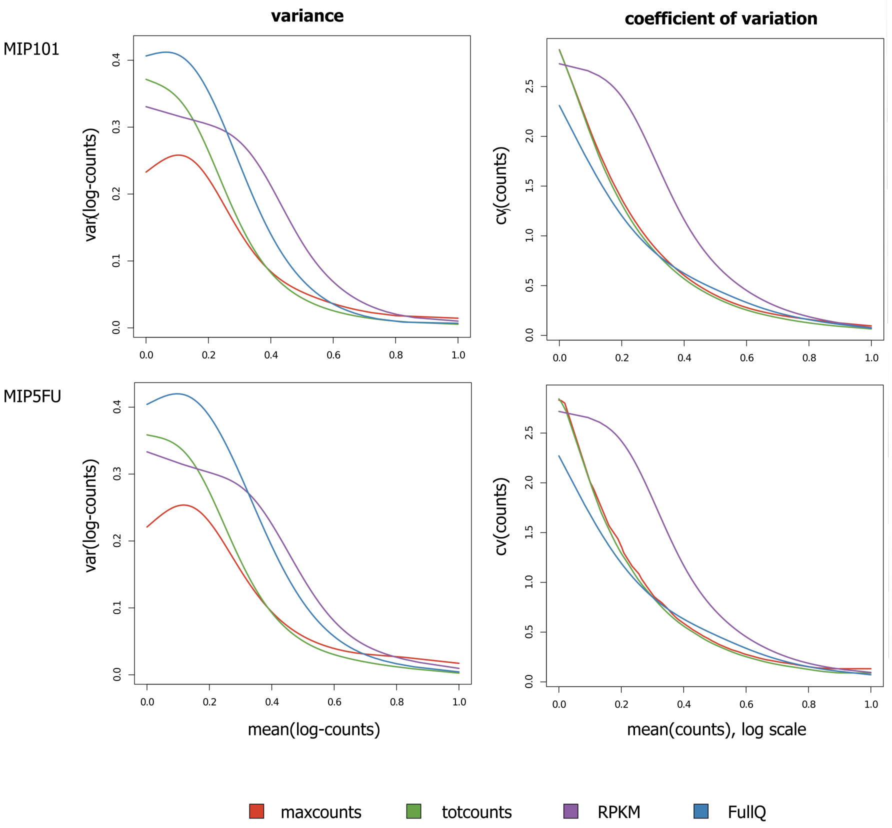


Figure 5.12: Variance and coefficient of variation (CV) of Griffith's data: variance *vs.* mean of log-counts/RPKMs (left plots) and CV vs. log-mean of counts/RPKMs (right plots). Curves represent cubic-spline fits computed on variance/CV, averaged in bins of 5000 exons each. Since maxcounts, totcounts, and totcounts normalized with RPKM (*RPKM*) and within-lane full-quantile normalization over exon length (*FullQ*) approaches are compared, x-values are scaled to cover the range [0, 1] in order to make them comparable.

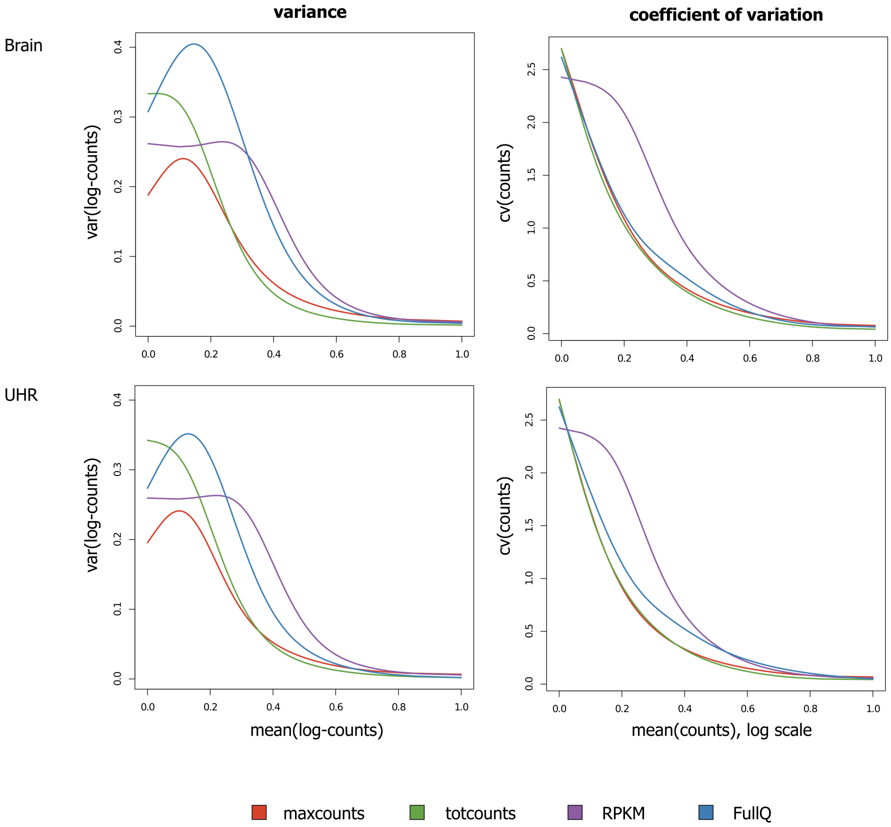


Figure 5.13: Variance and coefficient of variation (CV) of MAQC2 data: variance *vs.* mean of log-counts/RPKMs (left plots) and CV vs. log-mean of counts/RPKMs (right plots). Curves represent cubic-spline fits computed on variance/CV, averaged in bins of 5000 exons each. Since maxcounts, totcounts, and totcounts normalized with RPKM (*RPKM*) and within-lane full-quantile normalization over exon length (*FullQ*) approaches are compared, x-values are scaled to cover the range [0, 1] in order to make them comparable.

5.5 Robustness of count measures to alignment quality

An important criterion for the evaluation of reproducibility is the robustness of totcounts and maxcounts to variations in the quality of alignments. Results presented so far refer to a filtered set of alignments obtained using the analysis pipeline defined for this study, in which multireads and low-similarity alignments were discarded. To investigate how this choice impacts on quantification, for each exon i in each library j, we measured the relative variation $RV(i,j)$ between counts $N(ij)$ obtained from the original set of alignments and from the filtered set, as follows:

$$RV(i,j) = \frac{N_{orig}(i,j) - N_{filt}(i,j)}{N_{orig}(i,j) + 1} \cdot 100 \tag{5.3}$$

where the expression at the denominator is used to avoid possible divisions by zero. Ideally, if a measure is robust to alignment filtering (that depends on the specific analysis pipeline defined by users), relative variation should be 0%. Here we consider raw maxcounts and totcounts, not subjected to any normalization, since we want to assess the direct impact that changes in alignment filtering have on count summarization.

On all data sets, the fraction of exons for which maxcounts have 0% variation is always higher than that of totcounts (one-tailed t-test, p-value = 0.02). In particular, on Griffith's data, more than 80% of exons are not affected by alignment filtering (Figure 5.14A). In addition, histograms of relative variations in Jiang's data show that only a small fraction of exons are affected by medium-high variation (Figure 5.14B). For visualization purpose, exons with null variations are not represented by histograms, since they would result in a very high bar in correspondence of 0%, making it harder to assess variations greater than 0%. Similar results are found on Griffith's and MAQC2 data sets (plots not shown).

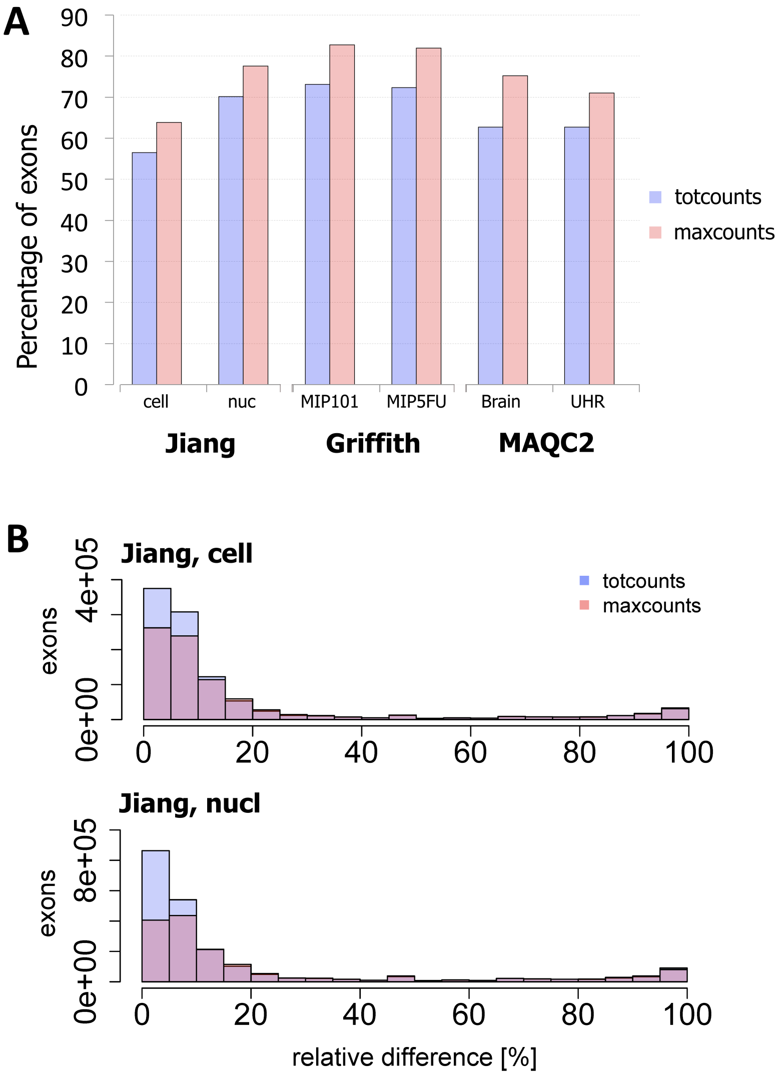


Figure 5.14: Relative variation of non-normalized totcounts (blue) and maxcounts (red) when low-similarity alignments and multireads are discarded: percentage of exons with null variation (A) and superimposed histograms of non-null variations affecting exons in Jiang's data set (B).

5.6 Summary of *maxcounts* performance

In a standard RNA-seq assay, the expression of a coding unit, such as a gene, transcript or exon, is estimated by considering the total number of reads that can be aligned on its sequence (*totcounts*). Despite being widely adopted, this digital measure of expression is not free from biases, and efforts are underway by the scientific community to develop novel methods for data normalization and bias correction. Here we propose an alternative approach for computing RNA-seq counts, called *maxcounts*. Read coverage along an exon is exploited to compute maxcounts as the maximum of its positional counts, i.e. the number of reads covering each base along its sequence.

We characterized and compared totcounts and maxcounts considering the desired features of a measure of expression, irrespectively of downstream applications: no dependence on covariates, such as exon length and GC-content, no over-representation of highly transcribed exons, accurate and precise estimation of true expression levels, low variance and high reproducibility.

Overall, totcounts always need normalization for exon length since they present a strong bias. On the contrary, exon length bias in maxcounts is strongly reduced, so they do not necessarily require normalization. If exon length bias is corrected through within-lane full-quantile normalization, further correction for GC-content is not needed neither for totcounts nor for maxcounts. Moreover, with maxcounts the over-representation of highly expressed exons is reduced with respect to totcounts. When focusing on accuracy and precision of measurements, maxcounts together with RPKM-corrected totcounts reproduce real data in the most accurate way, whereas maxcounts together with totcounts normalized with the full-quantile approach show the lowest variance. Finally, although the quality of alignments has a great impact on both methods, maxcounts approach outperforms totcounts in terms of robustness to variations in alignment filtering. Consequently, we believe that maxcounts approach represents a valuable alternative to totcounts for measuring exon expression from RNA-seq data, since it has comparable or higher performance on all the evaluation criteria.

Although several improvements have been made to understand and correct for possible biases in the RNA-seq experimental protocol, read coverage along transcripts still shows sequence-specific variability and under-representation of specific regions. Maxcounts approach can overcome biases due to the non-uniformity of read coverage, selecting the best represented transcript regions. Nevertheless, RNA-seq is a methodology still under active development, which will experience a fast improvement of experimental protocols and evolution of data characteristics. We made available the code for calculating maxcounts, thus enabling its benchmarking on different data sets.

6 Definition of a computational analysis pipeline for measuring gene expression in human RNA-seq data

The analysis pipeline defined in the previous chapter (see Figure 5.1), despite being appropriate for fair comparison of methods for counts computation (i.e. after read mapping), might present some limitations when applied to complex transcriptomes. Indeed, considering only uniquely mapping reads and discarding multireads can produce misleading results (e.g. in regions containing copy number variation or homologous genes). As seen in section 2.2, discarding multireads necessarily leads to a loss of information and a systematic underestimation of expression levels in correspondence of repetitive regions.

In order to define a pipeline that is robust to the structure of the human transcriptome, we decided to exploit a strategy for performing multiread mapping. Herein we do not intend to employ a method for quantification of gene or isoform expression levels, but to integrate a strategy for handling multireads in our pipeline, between pre-processing and count computation performed with *totcounts* or *maxcounts* approach. So far, several strategies for handling multireads have been proposed, as described in the literature review of section 2.2. However, most of the available methods comparisons focus on the expression levels quantification and do not assess accuracy and precision of multireads alignment, making it harder to identify the best performer on latter task. Among the available methods (reviewed in Chapter 2), we decided to use RSEM [161, 79] since it is one of the most accurate methods, handles also paired-end reads and has a limited memory requirement [79, 97]. Nevertheless, similarly to most of mapping algorithms, RSEM cannot consider paired-end and single end reads in the same run. Thus, we tested it in different algorithmic settings to identify the best configuration to integrate in the final analysis pipeline.

In the following, we present the definition and implementation of RSEM assessment, along with the obtained results. In addition, the integration of the final read mapping strategy into the analysis pipeline is presented and discussed.

6.1 Analysis framework and methods to assess RSEM mapping strategies

The procedure for read pre-processing defined in Chapter 5 (Figure 5.1) generates two sets of high-quality reads: singletons (i.e. single-end reads), whose mate was discarded in the filtering step, and paired-end reads (Figure 6.1).

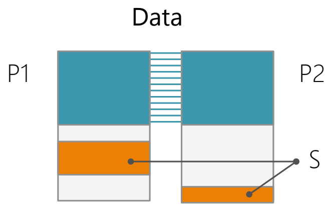


Figure 6.1: Representation of RNA-seq data after pre-processing: some paired-end reads are kept ($P1$ and $P2$ mates, in green), some are discarded (in grey) and others can be kept as singletons, i.e. without their mate (S, in orange).

At the time of the assessment, RSEM mapping was implemented in a script called `rsem-calculate-expression` that did not allow aligning both paired-end and single-end reads in the same run. By default, RSEM aligned reads in *single-end mode* (SE), while *paired-end mode* (PE) could be selected by specifying the `--paired-end` parameter. In the latter case, the algorithm discarded singletons or reads having an unmapped mate. Thus, we decided to explore and compare several strategies to align together singletons and paired-end data with RSEM:

PE map only paired-end reads using the PE mode;

SE map all reads in SE mode;

HPS consider reads mapped in PE mode and add reads mapped only in SE mode;

HPS consider reads mapped in SE mode and add reads mapped only in PE mode.

The HPS and HSP approaches do not necessarily provide the same solutions, since the same read can be mapped on different locations.

Another possible solution is that of avoiding read pre-processing to directly map the original paired-end reads. However, read trimming is important for removing adapters and low-quality sequences that can reduce the number of mapped reads and bias expression results [113, 162]. Thus, we decided to mantain the pre-processing module before mapping in our computational pipeline. Nevertheless, the framework described above does not completely allow a fairly comparison of PE and SE strategies because both consider paired-end reads ($P1$ and $P2$ in Figure 6.1), but only the SE approach also uses singletons (S reads in Figure 6.1). So, in order to separate the effects due to the algorithm or to the data, we tested a further approach, here referred as SE$_\mathrm{p}$: mapping only paired-end reads (i.e. discarding the singletons) with the SE strategy.

In order to benchmark these different strategies we considered six simulated human data sets of RNA-seq reads generated with FluxSimulator [114] and described in section 3. Reads were pre-processed as described in Figure 5.1 and mapped with RSEM v1.1.20 on the human reference genome (GRCh37.p10). We used RSEM only to align reads to the genome, without considering the downstream estimates of gene and isoform expression levels. To this end, RSEM was used together with BamTools v1.0.2 [163] to assign each read to a unique location. In particular, the `rsem-calculate-expression` function was used with the `--sampling-for-bam` parameter to handle multireads. The BAM file output by RSEM (`*.genome.sorted.bam`) was parsed with the `bamtools filter` function, specifying the `-mapQuality 100` parameter, to extract only the best hit for each read. The following RSEM parameters were also specified: `--bowtie-e 60`, `--bowtie-m 30`, `--bowtie-chunkmbs 512`, `--fragment-length-mean 180` and `--fragment-length-sd 50`. Custom Perl and bash scripts were used to merge the alignments output by the SE and PE mode, to obtain the final set of HPS and HSP alignments.

Read counts were computed over human exon regions defined as in [130]. Exon regions were obtained from the Gencode human annotation (version 12) using the `dexseq_prepare_annotation.py` script. Both totcounts (i.e. the total number of mapped reads) and maxcounts (i.e. the maximum read coverage) were computed for each exon region as in Chapter 5, using Bedtools v2.17.0 [148].

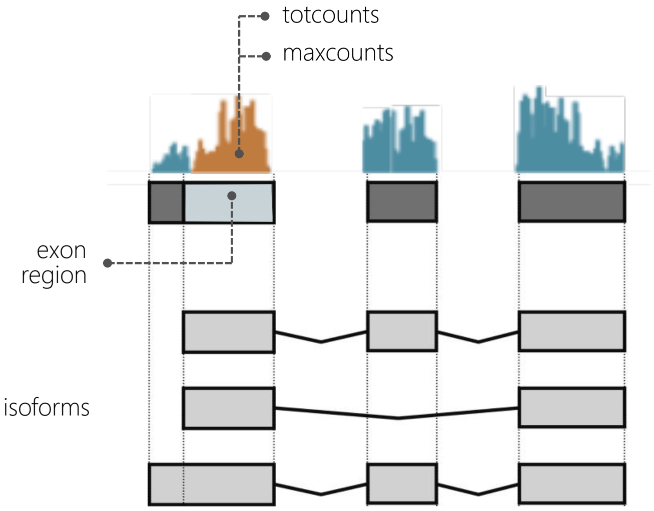


Figure 6.2: Computation of totcounts and maxcounts over exon regions. Exon regions were defined as in [130], splitting in two or more bins the exons having different boundaries in the different gene isoforms.

Considering the true genomic positions from which simulated reads were sampled, we also computed true maxcounts and totcounts. Thus, we were able to compare the estimated counts to the true ones to see if they were correct ("hit"), overestimated or underestimated. In particular, five classes were defined considering if true counts were in the lower half-range or in the higher half-range of expression (see Table 6.1 for clarification).

Table 6.1: Classification of estimated counts, considering the level of expression of the true counts. True counts can be in the high-expression range (if greater than $(\max(counts) - \min(counts))/2$) or in the low-expression range. Estimated counts can be equal to the true counts ("hits"), over-estimated, if greater than the true ones, or under-estimated, otherwise. The combination of this four situation define four classes, while "hits" are defined independently from the range of true counts.

	estimated = true	estimated > true	estimated < true
Low true	hit	low-overest.	low-underest.
High true		high-overest.	high-underest.

6.2 Performance of RSEM mapping strategies

Figure 6.3 shows the results obtained for 8M and 20M reads (average values across libraries). In all cases, for more than 75% of exon regions maxcounts and totcounts estimates are exactly equal to the true ones, indicating that this analysis pipeline provides stable results.

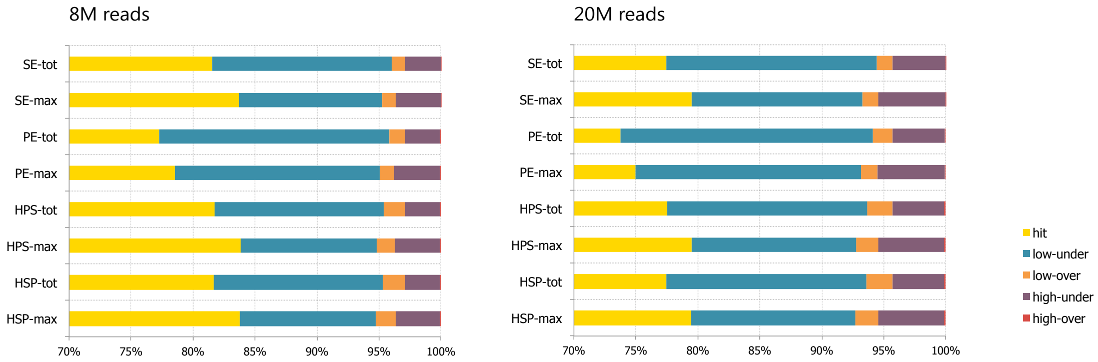


Figure 6.3: Hits and errors in maxcounts and totcounts estimates, in 8M and 20M reads data sets (average across three replicates). For all methods (PE, SE, HPS and HSP), bar-plots start from 70% because hits (in yellow) are never lower than 75%.

Maxcounts approach compared to totcounts, obtains a higher percentage of exact hits. Since we do not know the true measure of gene expression levels, we cannot conclude from these results that maxcounts gives more precise estimates, but only that they are robust to read pre-processing and mapping, confirming the results reported in Chapter 5. Unexpectedly, the SE strategy leads to a higher fraction of exact hits compared to the PE strategy. The performance of the "hybrid" approaches, HSP and HPS, are comparable to those of the SE mode, with a variation smaller than 0.2%. The 20M reads libraries obtain slightly less hits than the 8M reads libraries and are characterized by a higher average error (Figures 6.3 and 6.4).

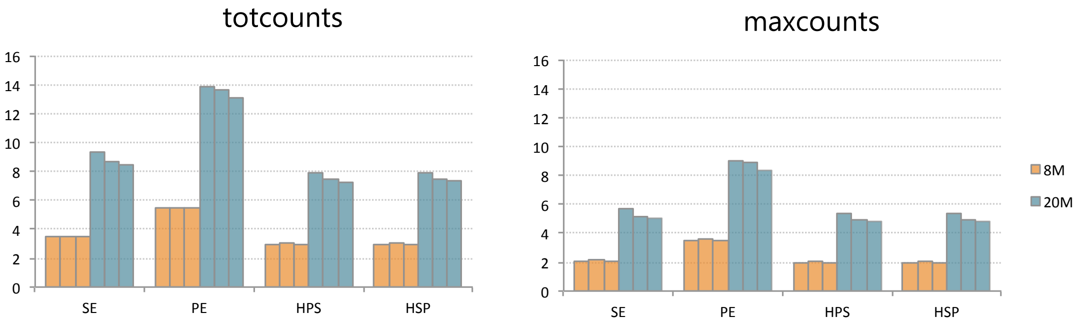
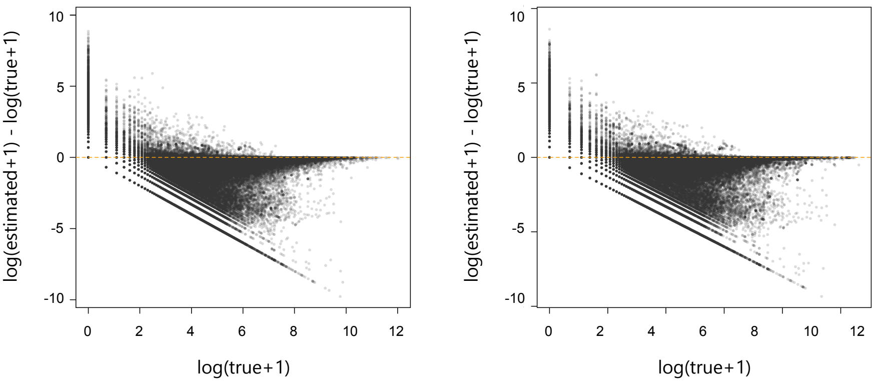


Figure 6.4: Average error (i.e. differences between estimated and true counts) across exon regions. Bar-plots for totcounts and maxcounts estimates in 8M (orange) and 20M (blue) reads libraries.

For all methods, very few exons are over-estimated, with slightly higher percentages in the hybrid strategies. Over-estimation mostly affects low-counts exonic regions while a large fraction of medium expressed exon regions is under-estimated (Figure 6.3 and 6.5). Figure 6.5, representing errors in totcounts and maxcounts estimation after SE mapping, highlights that the largest errors occur at very low and very high expressions; PE, HSP and HPS strategies lead to similar patterns (results not shown). As expected, most of errors are due to under-estimated counts, which might be partly due to reads discarded during pre-processing and mapping. Again, the hybrid modes show similar results compared to the SE strategy, with about -1% under-estimated and +1% over-estimated exon regions Figure 6.3.

Figure 6.5: Errors in totcounts (a) and maxcounts (b) estimated after SE mapping of 20M reads libraries. Scatter-plots of the differences between estimated and true counts versus true counts (log-scale). $log(counts + 1)$ is used to avoid invalid values for null totcounts and each dot corresponds to an exon region estimate for one library.



(a) totcounts (b) maxcounts

Defining a stringent threshold for hits might be a drawback for libraries with higher sequencing depths. Thus, we also assessed the percentage of hits obtained by each approach tolerating from 0 to 10^4 counts of difference between estimated and true counts (Figure 6.6). Nevertheless, the hits percentage for 8M reads libraries is always higher than that of 20M reads libraries, for all tolerated errors.

We also tested the $\mathrm{SE_p}$ approach, by mapping only paired-end reads with the SE strategy, to investigate whether the differences between the PE and SE approaches were due to the algorithm or to the data (Figure 6.7). Compared to SE mapping, the $\mathrm{SE_p}$ approach (same algorithm, less data) obtains fewer hits, as direct consequence of data loss (i.e. singletons reads). Notably, this does not only mean that the SE strategy maps more read, as expected, but that it is capable of assigning them to their exact location,

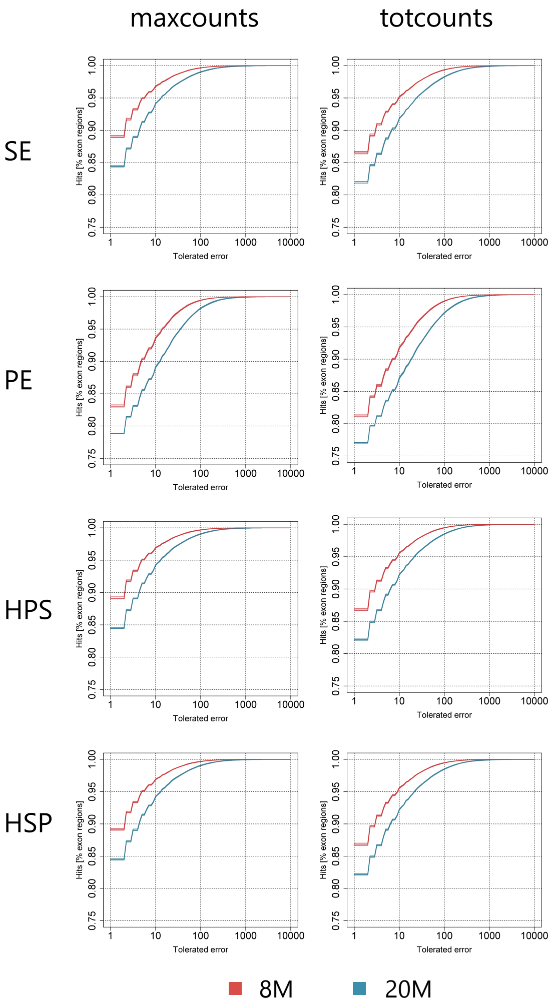


Figure 6.6: Percentage of hits, tolerating from 0 to 10^4 counts of difference between estimated and true totcounts (upper panels) or maxcounts (bottom panels) in 8M (red) and 20M (blue) reads libraries.

resulting in a correct count estimate. On the other hand, the SE_p approach obtains more hits than the PE approach, demonstrating that, if the same set of data is considered, RSEM can better assign reads to their correct position when used in single-end mode.

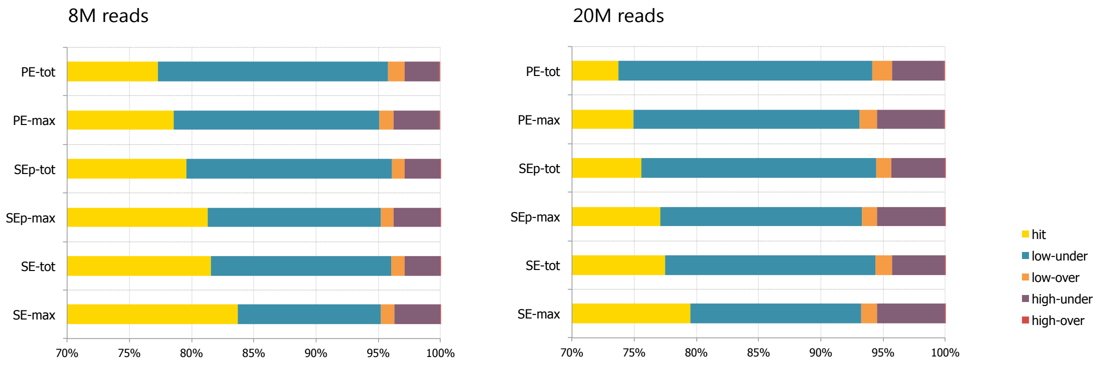


Figure 6.7: Hits and errors in maxcounts and totcounts estimates, in 8M and 20M reads data sets (average across three replicates). For all methods (PE, SE_p and SE), bar-plots start from 70% because hits (in yellow) are never lower than 75%.

6.3 Discussion and definition of the final computational pipeline

Our benchmarking of RSEM on two simulated RNA-seq data sets reveals that SE mapping of reads leads to a higher proportion of perfect hits with respect to PE mapping. The better performance of the SE scheme is partly due to the algorithmic strategy and partly to the data considered. Indeed, while the SE approach allows mapping all the reads output by the filtering step, the PE algorithm considers just the paired-end reads subset. This finding is in agreement with the work of Li *et al.* [79], where the best gene-level abundance estimates was obtained mapping single-end reads with RSEM. From the present results, it is not clear if the slightly worse results on the 20M reads libraries are due to the simulated data or to the settings of the algorithm. A systematic assessment of different data sets, with different features and sequencing depth, is needed to further clarify this aspect. Using SE and PE alignments to build a “hybrid” mode leads to results that are comparable to those of SE mode. However, these approaches are much more computationally intensive, since they require SE mapping, PE mapping and alignments post-processing and merging.

Given these results and considerations, we decided to integrate the single-end mode of RSEM in our analysis pipeline to align all pre-processed reads (see the final pipeline in Figure 6.8). As described above, RSEM is used along with custom made scripts to assign each read to a unique genomic position. Downstream expression quantification is performed at exon-region level, considering the aligned reads and exploiting totcounts and maxcounts strategies. Finally, count normalization (through TMM) and DE analysis and is performed with edgeR, selected for its superior performance over other methods (see sections 2.4 and Chapter 4).

It is worth notice that the pipeline here defined has been tested on human data sets and that its application to other organisms should be preceded by further specific assessments. For instance, a transcriptome that is denser in repeats or not well annotated might challenge more the methods here used and results in less accurate estimates [63].

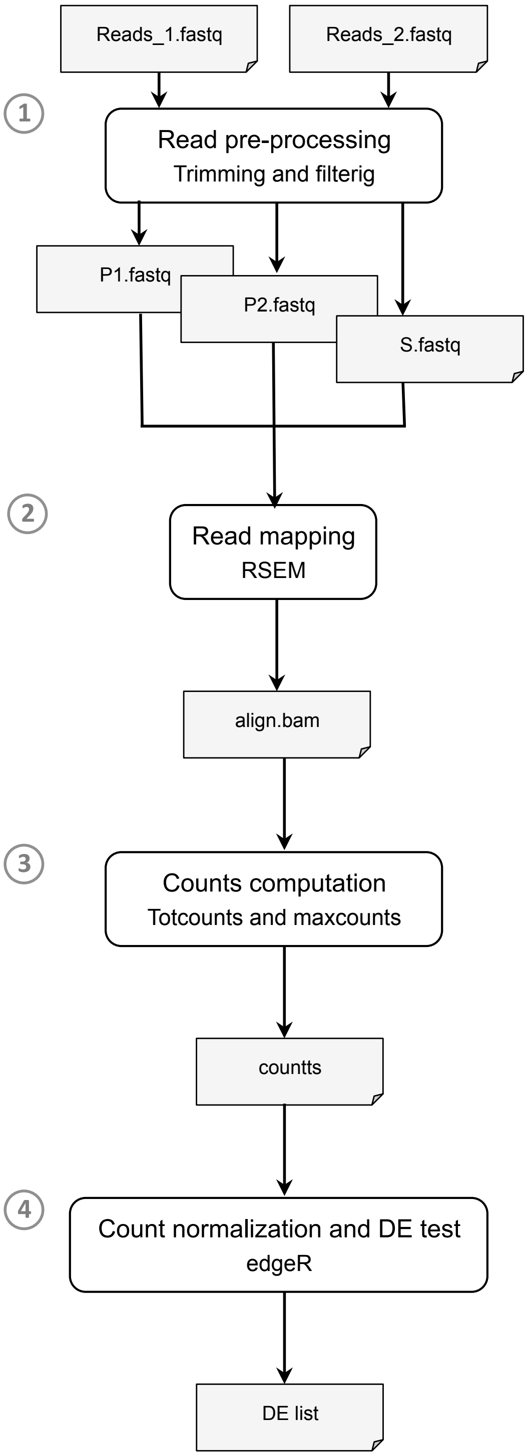


Figure 6.8: Final RNA-seq analysis pipeline. 1) Read pre-processing: trimming of low-quality ends and adapters and filtering of short reads. 2) Read mapping and allocation of multireads with RSEM in single-end mode. 3) Computation of maxcounts and totcounts. 4) Counts normalization and differential-expression analysis.

7

Application of the pipeline to a case study: RNA-seq analysis of patients affected by spinal muscular atrophy and healthy controls

The pipeline described in Chapter 6 was applied to a real case study, to identify the genes involved in the pathogenesis of spinal muscular atrophy (SMA) from RNA-seq data of patients and healthy controls. SMA is a degenerative and mortal neuromuscular disease that has no cure and represents one of the major genetic causes of infant mortality. The aim of the study was to investigate the variables involved in the pathogenesis of SMA, posing particular attention on the genetic factors contributing to phenotypic differences among SMA patients. To this end, a RNA-seq study of patients with mild or severe SMA phenotype, along with their healthy relatives used as controls, was conducted in collaboration with the Centro Nacional de Investigaciones Cardiovasculares (CNIC, Madrid, Spain). RNA-seq technology was chosen because it allows to quantify gene expression and to investigate single-nucleotide polymorphisms at the same time, so it makes possible a deep characterization of the genes involved in SMA pathogenesis. The definition of a pipeline robust to multireads (Chapter 6) is particularly important in this analysis, since the main genes involved in SMA, *SMN1* and *SMN2*, present very high homology and are located in a repetitive region of the human genome. In this chapter, we present the unpublished results of the RNA-seq assay on SMA patients and controls, analyzed with the pipeline defined through the assessments presented in the previous chapters.

7.1 Spinal muscular atrophy

Spinal muscular atrophy (SMA) is an autosomal recessive disorder (Figure 7.1) that causes degeneration of motor neurons in the anterior horn region of the spinal cord (Figure 7.2), which results in progressive muscle weakness and paralysis. SMA is one of the main genetic causes of infant mortality [164, 165] with an estimated prevalence of 1 in 10 000 births [166] and a carrier frequency of 1 in 40 individuals [167].

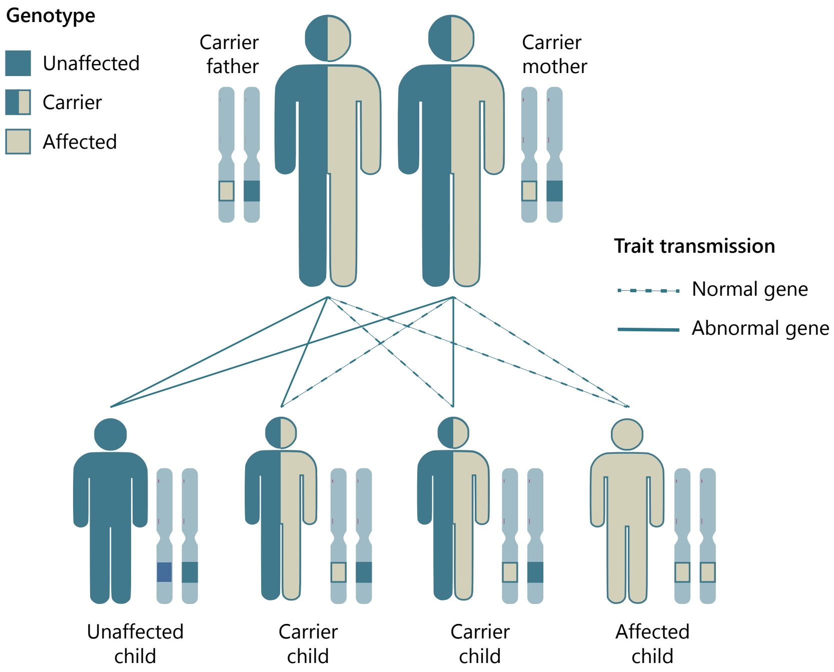


Figure 7.1: In autosomal recessive transmission two copies, that is one copy per chromosome, of the abnormal gene must be present in order for the disease or trait to develop. Heterozygous individuals, which are called "carriers" because they carry one copy of the abnormal (e.g. mutated) gene, do not manifest symptoms of this disorder. In spinal muscular atrophy, patients are characterized by absent or mutated *SMN1* gene on both copies of chromosome 5.

SMA severity is highly variable and the International SMA Consortium has defined four clinical groups depending on the age of onset and the achieved/conserved motor functions [169, 165]:

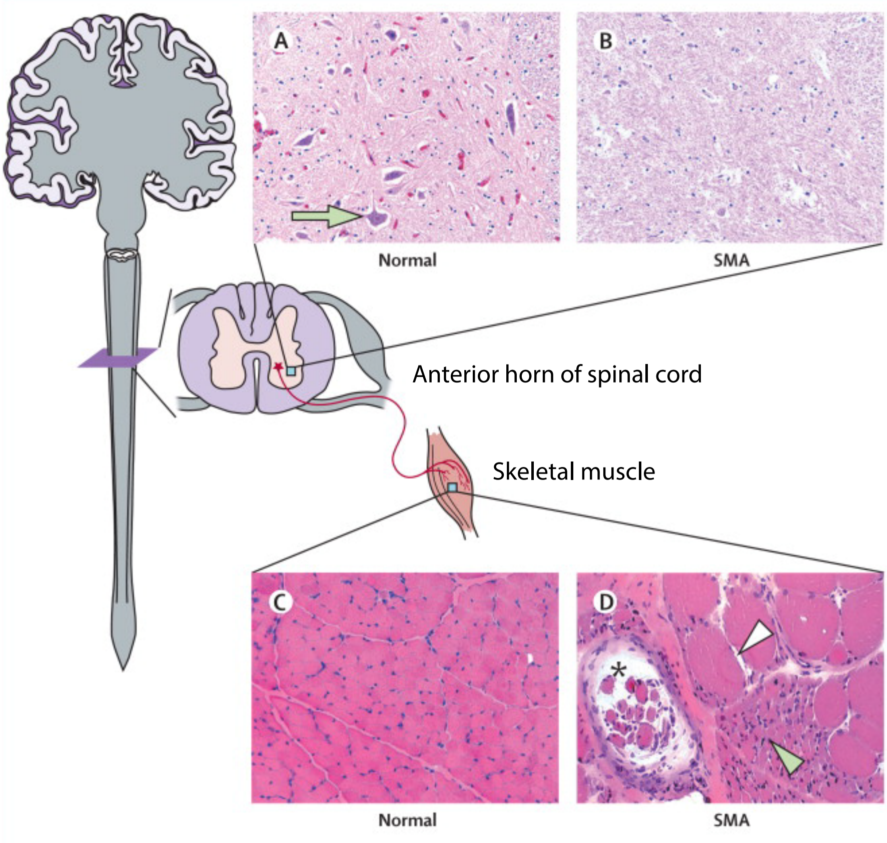


Figure 7.2: Histopathology of spinal muscular atrophy (figure and caption adapted from [168]). In healthy individuals, motor signals generated in the cerebral cortex are transmitted by motor neurons of the spinal cord to the skeletal muscle. The spinal cord anterior horn region in SMA patients (B) shows absence of motor neurons compared to healthy controls (A, green arrow). Compared with the uniform morphology of fibers in healthy muscle (C), skeletal muscle of a SMA patients (D) presents hypertrophic fibers (white arrowhead) surrounded by group atrophy (green arrowhead). Despite muscle fibers atrophy caused by SMA, muscle spindles (black asterisk) are not affected (D).

- Type-I SMA, also called “Werdnig-Hoffmann disease” [170], is the most severe form, usually diagnosed within the first 6 months of life, often at birth. Children are never able to sit or walk and usually die from respiratory failure within the first two years.

- Type-II SMA, also called “Dubowitz disease” [171], manifests is within the first 6 months. Patients are able to sit but cannot walk unaided. Life expectancy is reduced but most of patients live well into adulthood.

- Type-III SMA, also called “Kugelberg-Welander disease”, [172] is a milder form, with onset during infancy or youth. Patients are able to sit and walk and have a nearly normal life span. Disease onset before the age of 3 years is classified as type-IIIa, whereas age of onset beyond 3 years is classified as type-IIIb SMA. Type-IIIa patients experience a faster weakness progression and an earlier loss of ambulatory capacity compared to type-IIIb patients.

- Type-IV SMA [173] is the least severe form of SMA, which is diagnosed in adult age (i.e. after 30 years). Patients are mildly affected and have a normal life expectancy.

SMA is caused by the loss or mutation of the *survival motor neuron 1* gene (*SMN1*), which leads to reduced *SMN* protein levels and a to a selective dysfunction of motor neurons. In particular, 95% of SMA patients are characterized by a homozygous loss of *SMN1* gene while in the remaining cases the disease has been linked to small deletions or mutations in *SMN1* [174]. In humans, there are two nearly identical *SMN* genes, both located in 5q13 region of chromosome 5 and encoding the same open reading frame: the *SMN1* gene (telomeric) and the *SMN2* gene (centromeric) (Figure 7.3). SMA patients have deleted or mutated *SMN1*, but retain one or more copies of *SMN2*. However, the sequence of *SMN2* is characterized by a single-base change in exon 7, from C to T, with respect to *SMN1*. This C/T mutation does not change the amino acid coding but significantly alters the splicing pattern of the *SMN2* pre-mRNA, causing frequent skipping of exon 7 [175]. As a result, *SMN2* predominantly produces a *SMN* isoform lacking exon 7, $SMN\Delta_7$, which results in a protein lacking the last 16 amino acids. This protein is inactive and unstable, and is quickly degraded. Thus, SMA arises because *SMN2* generates a small fraction of full-length transcripts (about 10-15% of total produced RNAs) which cannot fully compensate for the lack of functional *SMN* due to *SMN1* gene deletion [176]. Loss of *SMN2* but not *SMN1* occurs in the human population without consequences [177]. On the contrary, the homozygous loss of both genes has not been reported in literature, probably because it is lethal [177]. Indeed, mice carry a single *Smn* gene and its deletion leads to very early death of embryos [178]. It is not clear yet how a single base substitution can induce exon 7 skipping in *SMN2*, but two hypotheses have been proposed [179, 180, 181]. Cartegni and Krainer suggest that the C/T substitution disrupts an *exonic splicing enhancer* (ESE), which constitute a binding site for *alternative splicing factor 1* (*ASF1*). Differently, Kashima and colleagues suggest that it creates a novel *exonic splicing suppressor* (ESS) site, which the *splicing suppressor hnRNP A1* binds. As Burghes and colleagues point out, these models are indeed compatible and highlight a weak point of the splicing machinery, in that a single nucleotide change can convert an ESE into an ESS, dramatically altering splicing patterns [174].

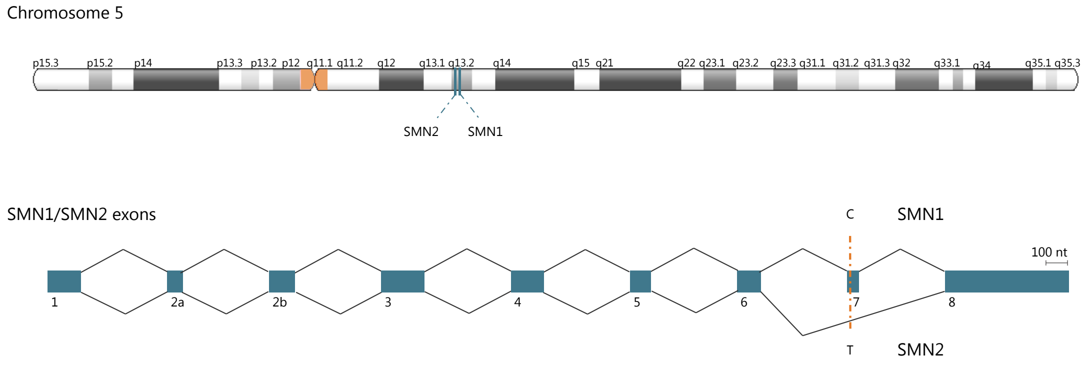


Figure 7.3: Location of *SMN2* and *SMN1* genes on chromosome 5 and splicing pattern due to the C/T mutation in exon 7.

Despite the progresses made during the last decade in the understanding of SMA, many questions about its exact genetic mechanism still have to be answered [182]. First of all, it is not clear how reduced *SMN* levels cause SMA. Two different hypotheses consider either its importance for the transport of mRNA in neurons or its role in the pre-mRNA splicing machinery [174]. Moreover, it is not clear why reduced levels of *SMN*, which is a ubiquitously expressed protein, only affect motor neurons. This issue is of particular interest, because other neurogenic disorders are linked to mutations in ubiquitously expressed genes, such as *Super Oxide Dismutase* (SOD) in amyotrophic lateral sclerosis and *Huntingtin* (HTT) in Huntington's disease [174]. Many of the studies carried out in recent years investigate which other variables, apart from *SMN1* gene, play a role in SMA so to determine such a wide range of clinical severity. The number of copies of *SMN2* gene explains most of the differences in SMA patients' phenotypes: SMA severity inversely correlates with the number of copies of *SMN2* gene [167] and patients with the milder type-II or type-III SMA have more copies of *SMN2* than type-I patients. The link between the number of copies of *SMN2* and SMA severity was also confirmed in mouse models [183, 184]. However, SMA phenotype cannot always be deduced solely from *SMN2* copy number [167] and the heterogeneity in SMA patients phenotypes has in some cases even challenged their prognosis and classification [182]. These limitations have led to the investigation and discovery of new variables playing a role in the disease. Prior and colleagues [176] studied three cases of SMA patients characterized by mild type-IIIb phenotype and having only two copies of *SMN2*. They found a G/C single base substitution in exon 7 of *SMN2* gene and discovered that this change constitutes an EES resulting in increased levels of full-length transcripts, and consequently less severe phenotype. In addition, *plastin 3* (*PLS3*), a gene located on the X chromosome that produces a protein involved in axonogenesis, was suggested as positive modifier by Opera *et al.* They found that, for some rare families, *PLS3* expression was higher in unaffected *SMN1*-deleted females than in SMA-affected males [185].

So far, neither a cure nor an effective treatment for SMA has been developed, but some efforts are underway to develop novel therapies, leveraging on: small molecules targeting *SMN2* that are capable of increasing *SMN* levels, viruses that carry *SMN1* gene and oligonucleotides that can prevent exon 7 skipping in *SMN2* [167]. In this scenario, the identification of protective modifiers is not only important to shred light on the pathogenesis of SMA, but might also led to the discovery of new potential targets for therapy.

7.2 Study design and computational analysis of RNA-seq data

Study design and samples collection

The study was designed by Dr. C. Hernández Chico (Molecular Genetics Unit of Hospital Ramón y Cajal, Madrid, Spain). In total, 201 patients fulfilling the SMA Consortium's diagnostic criteria for proximal muscular atrophy [169, 165] were included in the study (Table 7.1).

Table 7.1: Clinical features of the full cohort of SMA patients recruited for the study.

SMA type	Total	Gender		Family related		Age	
		men	women	yes	no	<16	≥16
I	97	49	48	1	96	97	0
II	68	40	28	1	67	32	36
III	36	17	19	4	32	4	32
Total	97	106	95	6	195	133	68

From the full cohort of patients, a subset of 5 type-II SMA patients, 5 type-II SMA patients and 20 healthy relatives was selected for transcriptome profiling through RNA-seq. SMA patients were selected considering the following criteria: homozygous deletion of *SMN1* and three copies of *SMN2*. Although the primary target of SMA are motor neurons, transcriptome profiling with RNA-seq was performed in peripheral blood. RNA assay in this accessible surrogate tissue allows an easier investigation of markers that can be later employed for diagnosis and therapy. Dr. Hernández Chico's research group recruited subjects and collected blood samples and medical data. Blood samples were collected using PAXgene Blood RNA tubes (PreAnalytiX [186]) in order to reduce in vitro RNA degradation and minimizing gene induction.

Library preparation, sequencing and raw data pre-processing

All the experimental steps form RNA extraction to sequencing were performed by the CNIC Genomics Unit, under the supervision of Dr. A. Dopazo González. Total RNA was quantified by absorbance at 260 nm in a NanoDrop spectrophotometer and its integrity was checked using Agilent Bioanalyzer [187]. 500 ng of total RNA were used with the TruSeq RNA Sample Preparation v2 Kit (Illumina [24]) to construct barcoded cDNA libraries. Libraries were quantified with Nanodrop Spectrophotometer (Nanodrop [188]). Quality and fragment size distribution of the Illumina libraries were determined using the DNA-1000 Kit (Agilent Bioanalyzer). The prepared cDNA libraries were applied to an Illumina flow cell for cluster generation (True Seq SR Cluster Kit V2 cBot) followed by sequence-by-synthesis with the Illumina Genome Analyzer IIx, to generate 2x75bp paired-end reads. The 20 healthy controls were subjected to a single sequencing run. For each of the 10 SMA patients multiple sequencing replicates were performed in order to increase sequencing depth. In total, 59 RNA-seq libraries were sequenced (Table 7.2).

Conversion of raw Illumina BCL files to FASTQ format and demultiplexing where performed by the CNIC Genomics Unit using the `Illumina2bam` function [189] and Picard [190].

Application of the computational pipeline for data analysis

Raw read data consist in 59 FASTQ files from 30 subjects (Table 7.2). From here on, type-II patients, type-III patients and healthy controls will be referred as "SMA2", "SMA3" and "CTRL", respectively. All FASTQ files from to the same subject were concatenated in a single file, obtaining 30 FASTQ files in total. Read pre-processing and mapping were performed following the computational pipeline defined in Chapter 6 (Figure 6.8). The number of original, filtered and mapped reads for each individual is shown in Figure 7.4: 10-17% of original reads were eliminated during the filtering steps and 2-19% resulted

Table 7.2: Samples subjected to RNA sequencing, one per line: subject's ID, sex, SMA type, *SMN2* copy number, family, sequencing lane and run. Multiple sequencing replicates for each SMA patient were performed and are grouped by grey areas. SMA type: '2' and '3' indicate type-II and type-III SMA, 'C' indicates carriers and 'N' indicate unaffected individuals.

Subject	Sex	SMA	SMN2	Family	Run	Lane
CTRL_01_01	M	C	2	father	R1	L1
CTRL_01_02	F	C	2	mother	R1	L1
SMA3_01_03	M	3	3		R1	L7
					R2	L3
					R2	L2
					R2	L3
					R2	L4
CTRL_01_04	M	N	1	brother	R3	L1
CTRL_02_02	F			mother	R3	L4
SMA2_03_03	M	2	3		R1	L6
					R2	L5
					R2	L6
					R2	L7
CTRL_03_04	M	C		brother	R3	L4
CTRL_04_01	M			father	R3	L3
SMA3_04_03	F	3	3		R2	L5
					R2	L6
					R2	L7
					R4	L3
CTRL_04_08	F				R3	L3
CTRL_05_01	M	C			R3	L5
SMA2_05_03	M	2	3		R2	L7
					R2	L6
					R2	L5
					R1	L7
CTRL_05_04	M	C		brother	R3	L5
CTRL_05_01	M	C	2	father	R1	L2
CTRL_05_02	F	C	4	mother	R1	L2
SMA3_05_03	F	3	3		R2	L4
					R4	L2
					R2	L2
CTRL_06_10	F	C		grandmother	R3	L2
CTRL_06_01	M	C		father	R1	L4
CTRL_06_02	F	C		mother	R1	L4
SMA2_06_03	M	2	3		R4	L1
					R2	L2
					R2	L3
					R2	L4
CTRL_07_02	F	C	2		R1	L3
SMA2_07_03	F	2	3		R4	L2
					R2	L2
					R2	L3
CTRL_07_04	M	C	1	brother	R3	L2
CTRL_08_01	M	C	2	father	R1	L3
SMA2_08_03	F	2	3		R2	L2
					R2	L3
					R2	L4
					R4	L1
CTRL_08_07	M	N	1	R3	L1	
CTRL_09_01	M			father	R1	L5
CTRL_09_02	F			mother	R1	L5
SMA3_09_03	F	3	3		R4	L3
					R2	L7
					R2	L5
					R2	L6
SMA3_10_3	F	3	3		R1	L6
					R2	L5
					R2	L6
					R2	L7

unmapped.

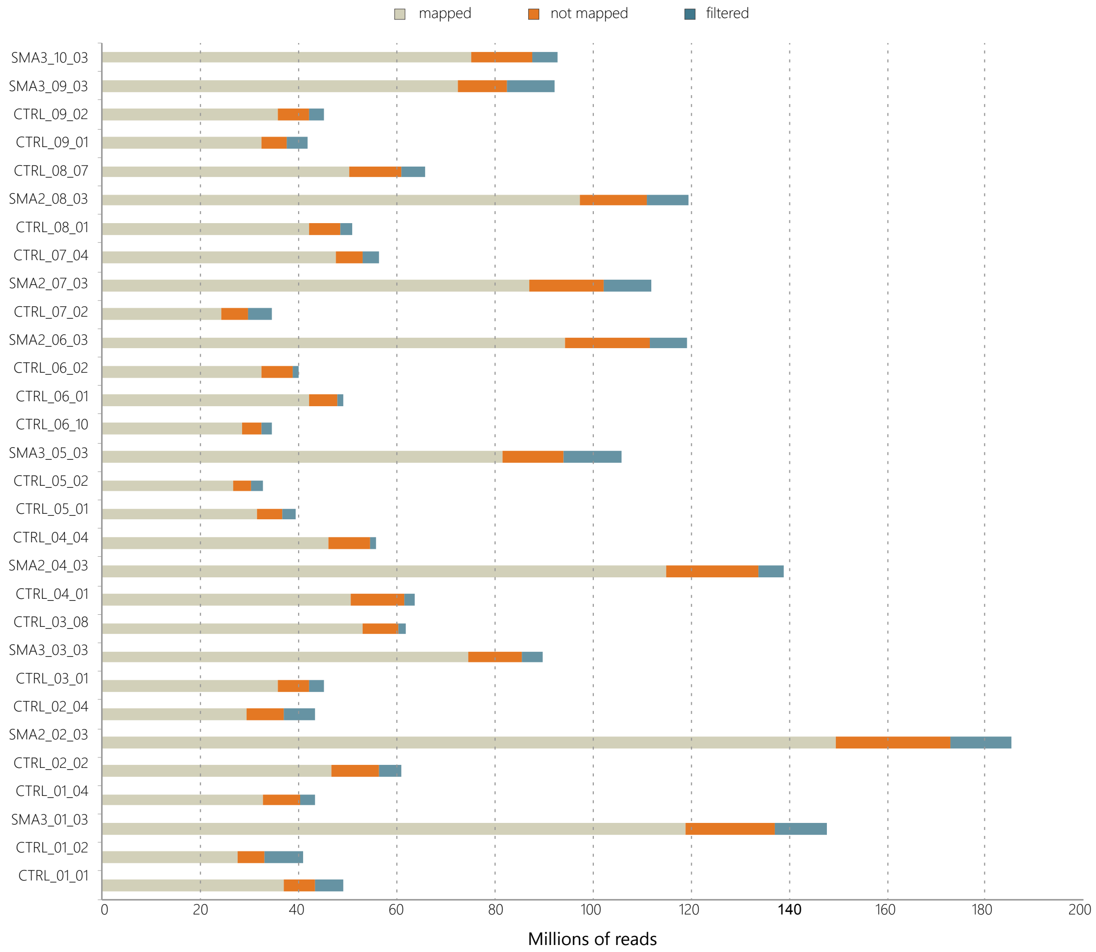


Figure 7.4: Number of reads sequenced for each sample and fraction of mapped (grey), not mapped (orange) and filtered (blue) reads.

Counts were computed as in Chapter 6: both totcounts and maxcounts (see definitions in Chapter 5.) were calculated for each exon region. Totcounts were also computed at gene level using Bedtools v2.17.0 [148]; from here on, we will refer to this approach as *totcounts-genes*. When not differently stated, *counts-per-million* or *cpm* is used in the following paragraphs to indicate totcounts-genes normalized to library size, using the `cpm` function of edgeR package [131].

Analysis of differential expression and genotyping

The count tables were imported in R [191] and exon regions shorter than 30 nt or having less than 5 counts in more than 20% of subjects were eliminated. Finally, totcounts and maxcounts were summarized at gene level considering the median across all exon regions belonging to a gene. Gene counts were finally tested for differential expression after correcting differences in sequencing depth between replicates via TMM normalization [104]; both steps were performed using edgeR package [131]. The "glm" version of edgeR [151] was used to compare gene totcounts/maxcounts between SMA and CTRL, SMA2 and CTRL, and SMA3 versus, correcting for family information. In addition, a reduced matrix consisting only in SMA patients' data was considered, and differences between SMA2 and SMA3 gene maxcounts/totcounts were tested using edgeR and correcting for sex information. p-values were corrected to control false-discovery rate (FDR) in multiple tests (*q-values*; [192]). DE genes were selected imposing a FDR threshold of 5%.

The set of DE genes selected for "SMA versus CTRL" and "SMA2 versus SMA3" comparisons were also subjected to Ingenuity Pathway Analysis (IPA, Ingenuity Systems [193]) to investigate gene networks and molecular functions associated with SMA and with disease severity. IPA system transforms a list of genes into networks, leveraging on the information gathered from the Ingenuity Pathways Knowledge Base (IPKB), an extensive and curated database with annotations about genes, gene products, processes, diseases and drugs. Networks give a graphical representation of molecular relationships between genes. Genes, molecules and complexes are represented by nodes, and biological relationships between nodes are represented by edges and are supported by references stored in the IPKB. IPA also assigns a p-value to each network indicating the likelihood that connections between nodes are due to chance. The list of selected genes, together with q-values and log-ratios estimated by edgeR for maxcounts, were submitted to IPA.

Aligned reads were also processed with `mpileup` function of Samtools [159] to detect SNPs and indels. For each detected variant, the distribution of genotypes of the patients with type-II patients and type-III patients was compared using a Fisher's test.

7.3 Investigation of genes involved in SMA pathogenesis

Selection of differentially expressed genes

Results of the DE analysis performed with edgeR are summarized in Figure 7.5. The intersection between the lists provided by totcounts and maxcounts approach was taken as final lists of DE genes. The overlap with the lists provided by totcounts-genes approach was also checked (Figure 7.5); in all comparisons except “SMA3 versus CTRL”, 77-83% of DE genes are confirmed also by this third method. DE genes confirmed by two or more methods also have the same direction of differential expression (i.e. are over-expressed or under-expressed).

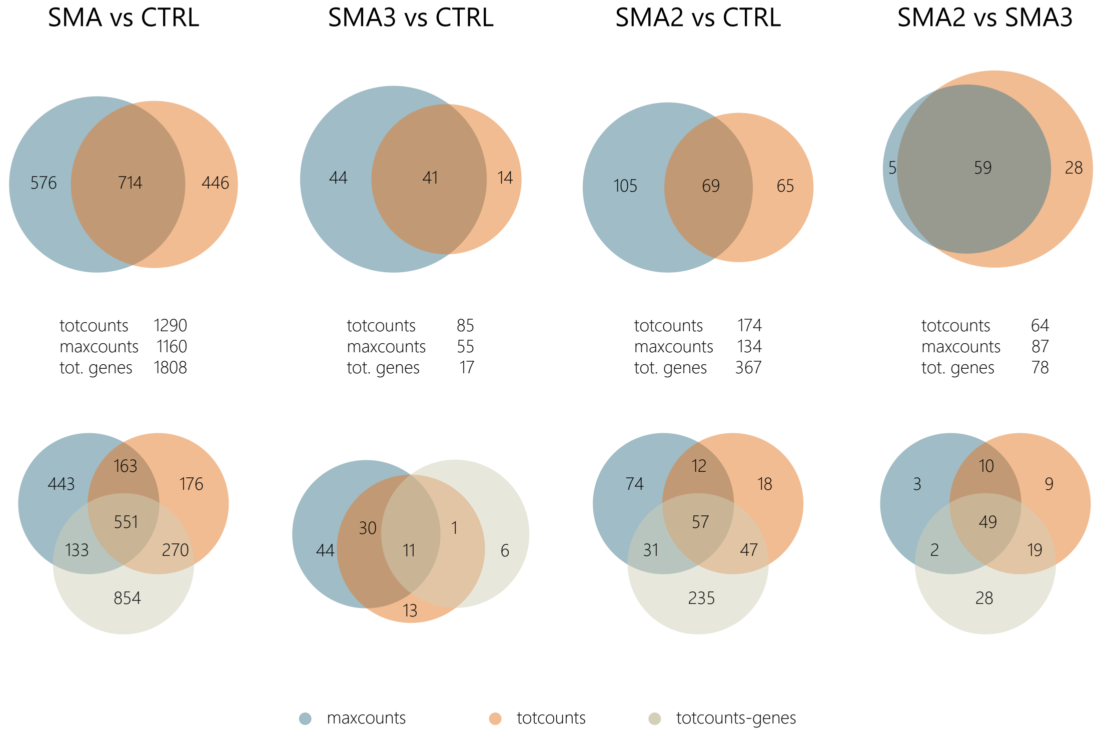
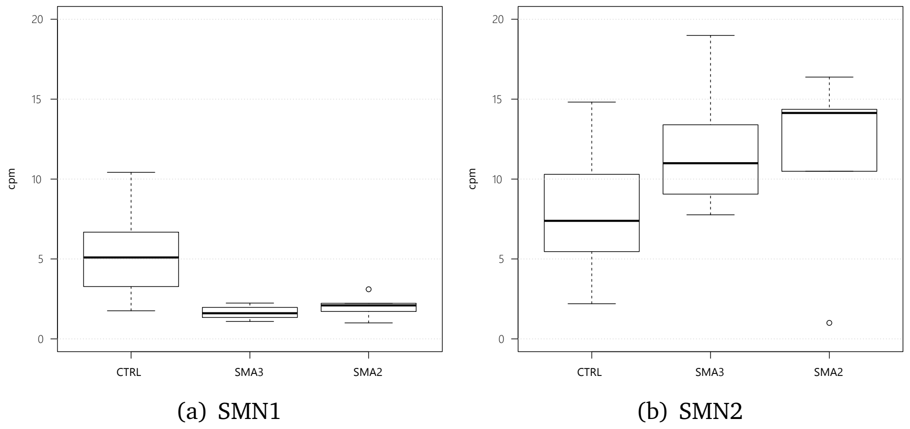


Figure 7.5: Differentially expressed genes: Venn’s diagrams on top show the number of differentially expressed (DE) genes selected with maxcounts (blue) and tocounts (orange) approaches; the final list of DE is given by the intersection. The number of total DE genes detected by maxcounts, totcounts and *totcounts-genes* approaches (grey) is reported in the tables and their intersection is shown in the Venn’s diagrams below.

Some of the DE genes selected (Tables B.1-B.4) are *overlapping-genes*, because their locations onto the human genome overlap, preventing to discriminate completely which genes contribute to the detected changes in gene expression. In summary, 714 DE genes (83 of which overlapping) were detected in the “SMA versus CTRL” comparison, 41 genes (2 overlapping) in “SMA3 versus CTRL”, 69 genes (9 overlapping) in “SMA2 versus CTRL” and 59 in “SMA2 versus SMA3” (6 overlapping). As expected, *SMN1* is among the top-10 ranked DE genes for the all the comparisons of SMA, SMA2 and SMA3 versus CTRL. However, *SMN1* expression in SMA patients is not null (Figure 7.6a), probably due to the impossibility for RSEM to correctly determine the original genome position of reads arising from *SMN*Δ_7 isoform, transcribed from *SMN2* but having 100% similarity with *SMN1*. SMA patients show a slightly higher expression of *SMN2* gene compared to controls (Figure 7.6b) but this gene is not found significantly over-expressed in our analysis. However, DE analysis from maxcounts and totcounts detects an over-expression of *SMN2* gene in SMA2 patients compared to CTRL at FDR of 5.3% and 7.1%, respectively (i.e. slightly over the imposed threshold).

Figure 7.6: Expression of *SMN1* (a) and *SMN2* (b) genes in CTRL, SMA2 and SMA3 individuals: box-plots of counts-per-millions (cpm).



(a) SMN1 (b) SMN2

Ingenuity pathway analysis

For “SMA versus CTRL” comparison, IPA found annotation for 511 over 714 DE genes (Tables 7.3 and B.5) and revealed association with diseases related to inflammatory and antimicrobial response, with developmental disorders and with nervous system development and function (Table 7.3). The main network reconstructed by IPA (Figure 7.7) is associated with cell signaling, cell-to-cell signaling and interaction, and antimicrobial response. It involves 35 molecules, of which 32 in are selected DE genes. Additional IPA networks associate some of the selected DE genes with *SMN* genes or with pathways related to post-transcriptional modifications (Figures 7.8 and 7.9).

For “SMA2 versus SMA3” comparison, IPA found annotation for 35 over 59 DE genes (Tables 7.4 and B.6) and revealed a strong association with infectious, inflammatory and immunological diseases, but also with skeletal, muscular and connective tissue disorders (Table 7.4). DE genes were also related to functions that are fundamental for the maintenance of the neural system, such as cell death and survival or free radical scavenging. The main gene network identified by IPA comprises 35 molecules or complexes, involving

Table 7.3: Summary of IPA results for "SMA vs CTRL" comparison: association of differentially expressed genes with known diseases, molecular functions, physiological systems and functions (top-five ranking). IPA p-values and number of moleclues involved are also reported.

Diseases and biological functions	p-values	molecules
Diseases and disorders		
Antimicrobial response	1.18E-06-1.11E-02	8
Inflammatory response	1.18E-06-2.22E-02	41
Developmental disorder	2.07E-04-2.22E-02	36
Endocrine system disorders	2.07E-04-2.22E-02	10
Gastrointestinal disease	2.07E-04-2.22E-02	14
Molecular and cellular function		
Cell-to-cell signaling and interaction	1.18E-06-2.22E-02	45
Cell death and survival	2.49E-06-2.22E-02	123
Cell morphology	7.41E-05-2.22E-02	41
Cellular compromise	7.41E-05-2.22E-02	17
Cellular movement	7.73E-04-2.22E-02	42
Physiological system development and function		
Embryonic development	1.18E-06-2.22E-02	22
Hair and skin development and function	1.18E-06-2.22E-02	13
Renal and urological system development and function	1.18E-06-2.22E-02	7
Nervous system development and function	2.07E-04-2.22E-02	28
Tissue development	2.07E-04-2.22E-02	31

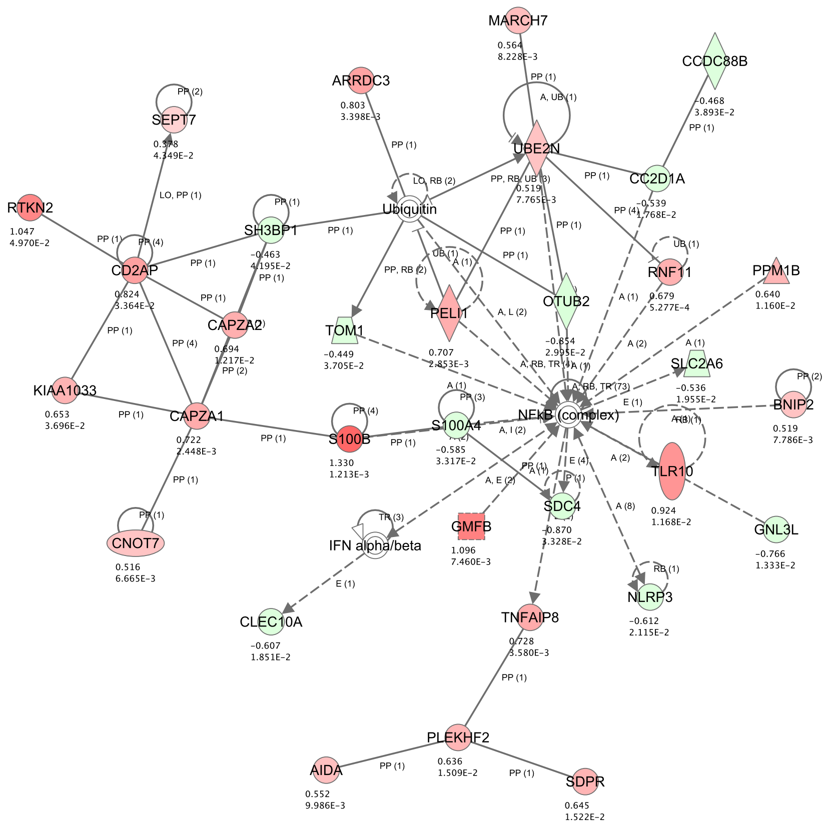


Figure 7.7: Main gene network reconstructed by IPA for "SMA versus CTRL" comparison. Colors identify over-expressed (red) and under-expressed (green) genes in SMA patients.

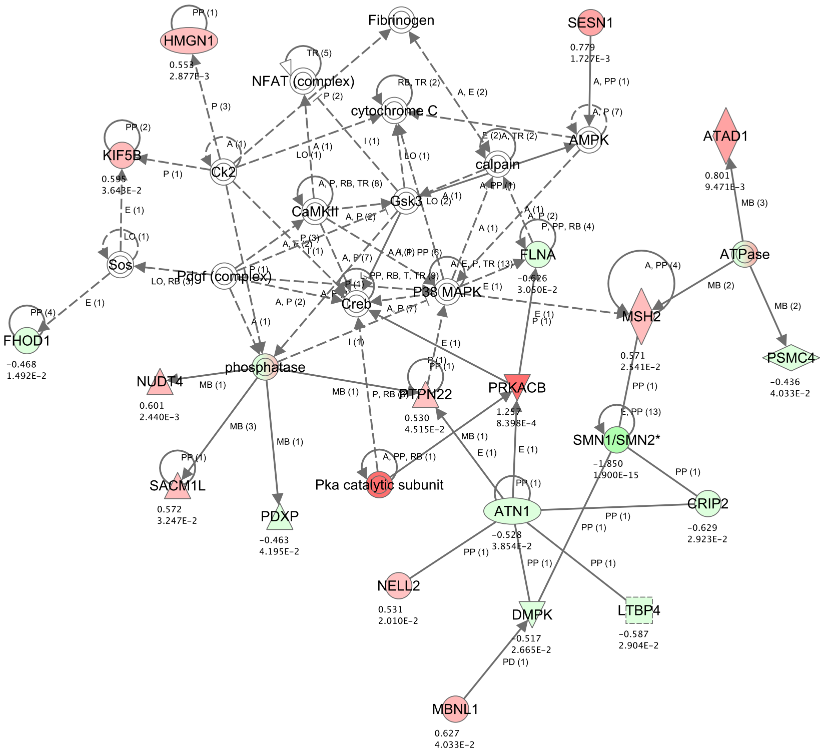


Figure 7.8: Gene network reconstructed by IPA for "SMA versus CTRL"' comparison and associated with post-transcriptional modification. Colors identify over-expressed (red) and under-expressed (green) genes in SMA patients.

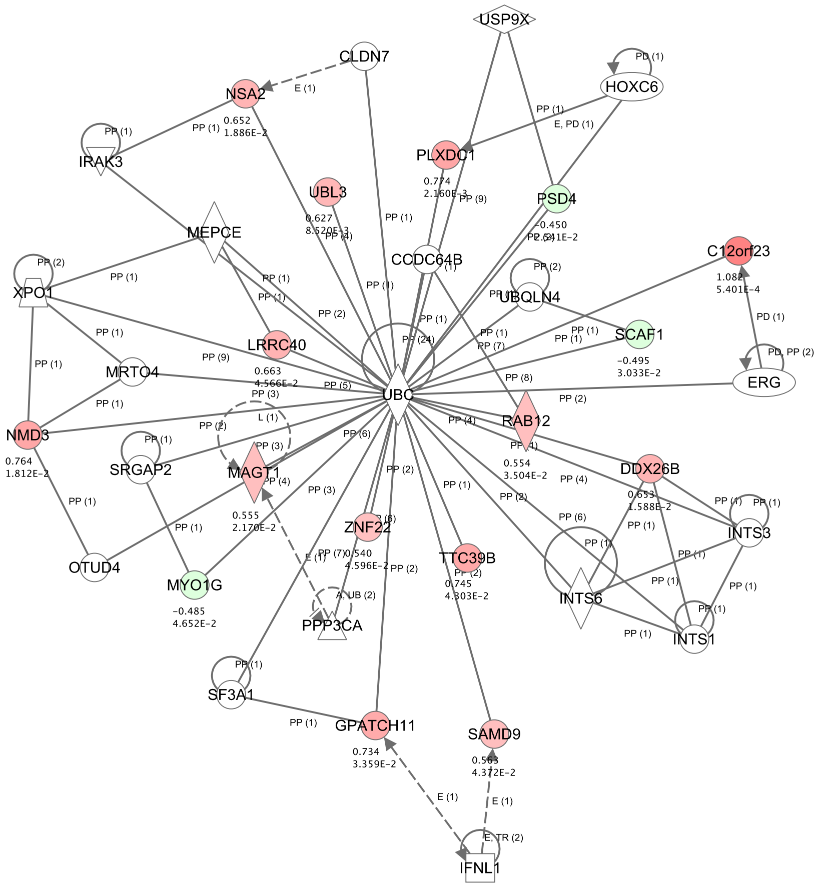


Figure 7.9: Gene network reconstructed by IPA for "SMA versus CTRL" comparison and associated with *SMN* genes. Colors identify over-expressed (red) and under-expressed (green) genes in SMA patients.

21 DE genes (Figure 7.10), and is associated with infectious and immunological diseases, and with connective tissue disorders.

Table 7.4: Summary of IPA results for "SMA2 vs SMA3" comparison: association of differentially expressed genes with known diseases, molecular functions, physiological systems (top-five ranking). IPA p-values and number of genes, molecules or complexes involved are also reported.

Diseases and biological functions	p-values	molecules
Diseases and disorders		
Infectious disease	5.16E-26-4.25E-02	18
Connective tissue disorders	6.96E-18-1.16E-02	17
Immunological disease	6.96E-18-4.73E-02	18
Inflammatory disease	6.96E-18-4.73E-02	16
Skeletal and muscular disorders	6.96E-18-1.16E-02	16
Molecular and cellular function		
Cell death and survival	2.78E-12-4.41E-02	16
Cellular movement	4.50E-12-4.73E-02	14
Cell-to-cell signaling and interaction	3.10E-10-4.73E-02	16
Free radical scavenging	1.89E-06-2.64E-02	8
Cellular growth and proliferation	3.27E-06-4.57E-02	11
Physiological system development and function		
Hematological system development and function	4.50E-12-4.73E-02	18
Immune cell trafficking	4.50E-12-4.73E-02	18
Tissue development	4.18E-09-4.73E-02	11
Tissue morphology	4.18E-09-4.25E-02	12
Organismal development	3.27E-06-4.41E-02	10

Interestingly, most of these genes are associated with connective tissue disorders such as: systemic lupus erythematous, rheumatic disease and rheumatoid arthritis. While "rheumatic disease" is a non-specific term indicating medical problems affecting the joints and the connective tissue, "rheumatoid arthritis" precisely identifies an autoimmune disease that results in a chronic, systemic inflammatory disorder. It affects many tissues and organs, but principally attacks synovial joints, causing substantial loss of functioning and mobility if not adequately treated. Systemic lupus erythematous (*SLE* or *lupus*) is an autoimmune connective tissue disease that can affect any part of the body: the immune system attacks the body's cells and tissue, resulting in inflammation and tissue damage. The expression of these genes, plus three additional DE genes also associated with these disorders, *ARG1*, *CRISP3* and *BPI*, are depicted in Figure 7.11. Eight of these genes are also reported to be free radical scavengers or deactivators: *ARG1*, *CAMP*, *CTSG*, *ELANE*, *LTF*, *OLR1*, *PRTN3*, *RETN*.

Investigation of protective modifiers

Gene expression patterns shown in Figure 7.11 highlight some peculiarities of the SMA2 phenotype: most of the genes have comparable or higher expression in SMA3 patients versus healthy controls, but are down-regulated in SMA2 patients (*KLRC2*, *KLRC1* and *CCL3L3* present inverse patterns). Thus, it is conceivable to hypothesize that this set of genes might underlie protective mechanisms that act against the damaging effect induced

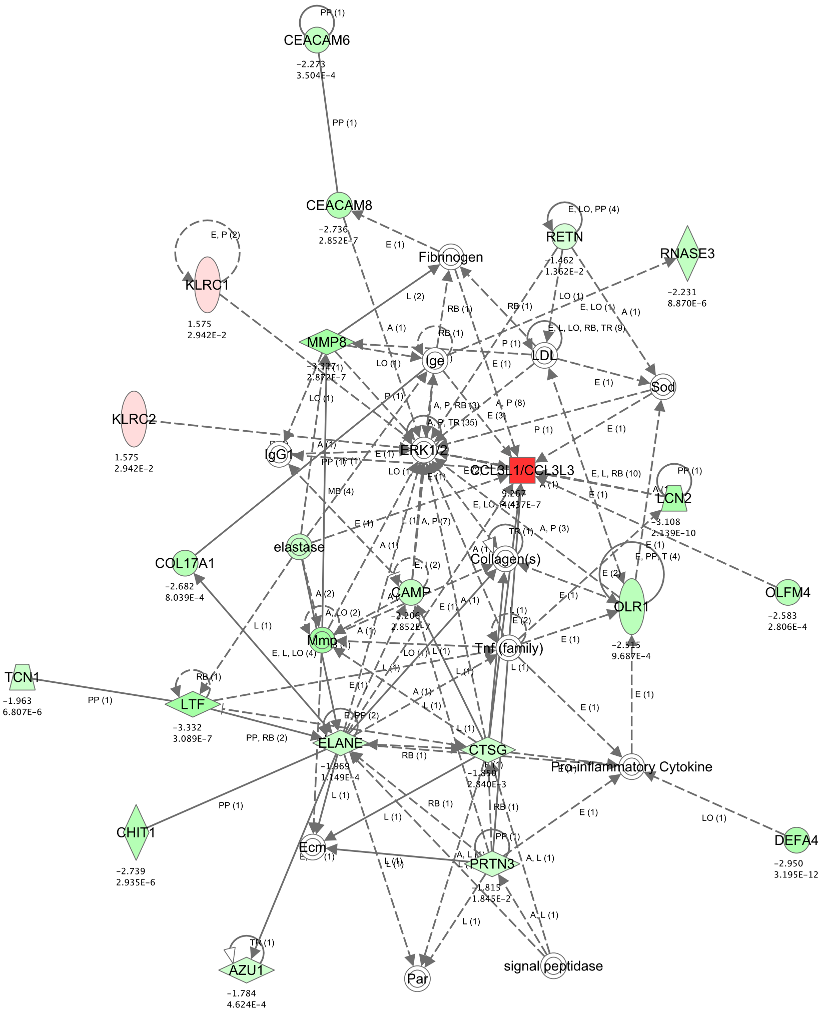


Figure 7.10: Main gene network reconstructed by IPA for “SMA2 versus SMA3” comparison. Colors identify over-expressed (red) and under-expressed (green) genes in SMA2 patients.

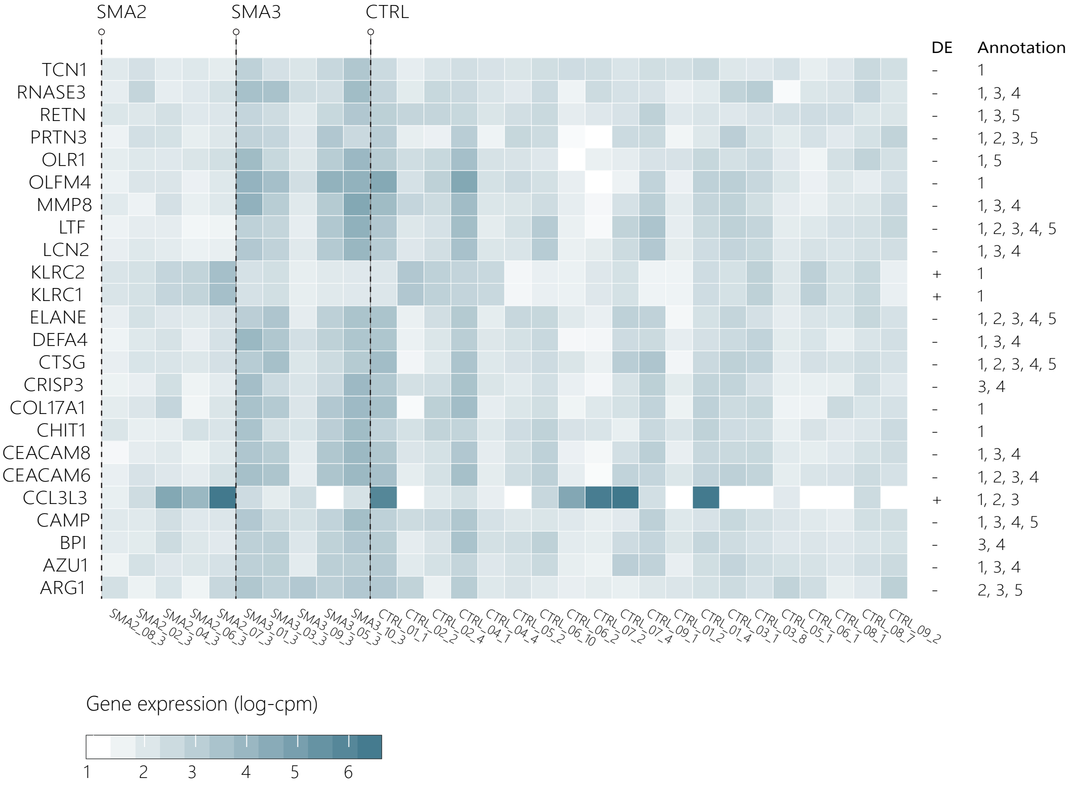


Figure 7.11: Expression patterns, for all individuals, of genes associated with differences in SMA phenotypes, connective tissue disorders and free radical scavenging ($\log(cpm + 1)$ was used as expression measure). For each gene is indicated up-regulation (+) or down-regulation (−) in SMA2 compared to SMA3 patients, and association with: the main IPA network of Figure 7.10, (**1**), systemic lupus erythematous (**2**), rheumatic disease (**3**), rheumatoid arthritis(**4**) and free radical scavenging (**5**).

by SMA. Among these genes, *Lactotransferrin* (*LTF*) presents two interesting features in the present data (Figure 7.12 and 7.13): *(i)* significantly decreased expression in SMA2 patients with respect to both healthy controls and SMA3 patients (Table B.6); *(ii)* presence of two SNPs, *rs9110* and *rs2073495*, located on exon 15 of *LTF* gene. Both SNPs detected in our RNA-seq data through genotyping have been previously validated and annotated in the dbSNP database [194]: the *rs2073495* G/C mutation induces a change in the aminoacid coding from glutammine to aparagine (*missense mutation*); the r*s9110* T/C mutation is silent, i.e. it does not affect protein coding.

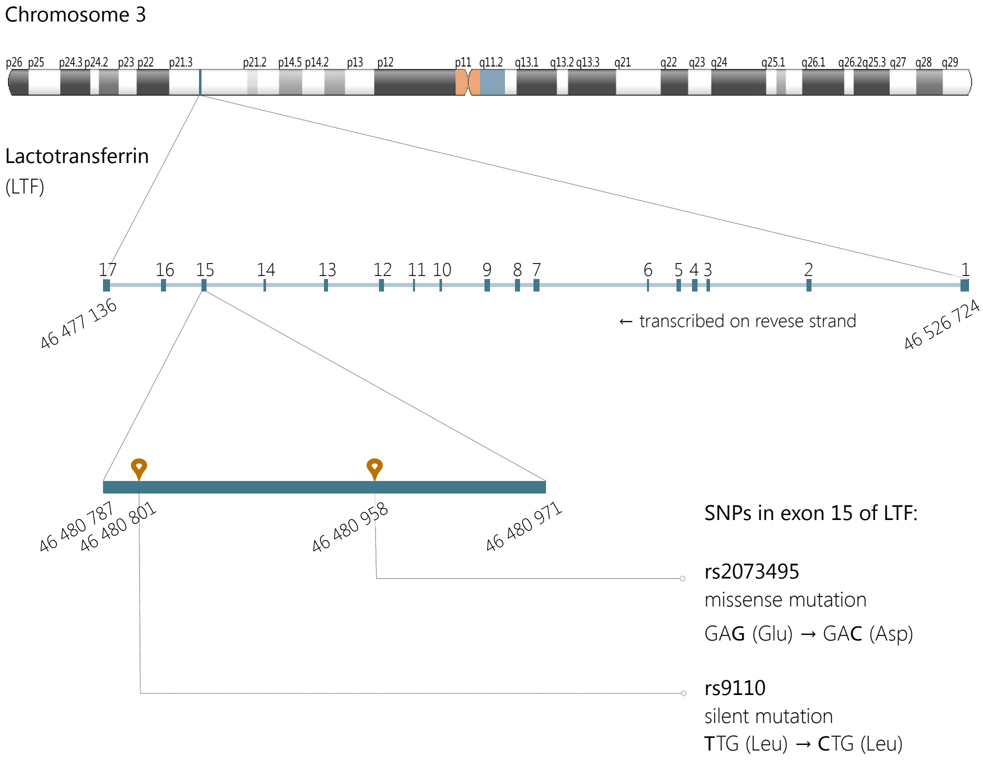
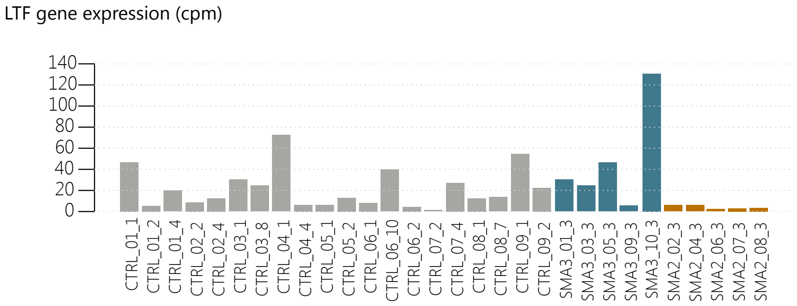


Figure 7.12: Genomic location and structure of the lactotransferrin (*LTF*) gene and localization of *rs9110* and *rs2073495* SNPs on exon 15.

In SMA patients, *rs9110* genotypes (Figure 7.13) show a weak association with *LTF* expression, measured as “cpm” (Anova test, p-value=0.08), and no association with phenotype (two-sided Fisher’s test, p-value=0.5238). On the contrary, *rs2073495* is strongly associated with *LTF* expression (Anova test, p-value=5.402e-05) and with phenotype (Fisher’s test, p-value=7.9e-3). In particular, all SMA2 patients have a decreased expression of *LTF* (confirmed by edgeR DE analysis, Tables B.3 and B.4) and are homozygous for the *rs2073495* reference allele (Figure 7.14). On the contrary, SMA3



Genotypes

rs2073495	SMA2	SMA3	CTRL
GG: homo-ref	5	0	7
GC: hetero	0	4	10
CC: homo-alt	0	1	3

rs9110	SMA2	SMA3	CTRL
TT: homo-ref	1	3	9
TC: hetero	4	2	10
CC: homo-alt	0	0	1

Figure 7.13: Summary of LTF expressions and genotypes. Bar-plots of *LTF* expression levels for CTRL (grey), SMA3 (blue) and SMA2 (orange), measured in cpm. Tables report the number of individuals per group having one of the three possible genotypes for *rs9110* and *rs2073495*.

patients have higher expression and carry also the alternative allele: four over five SMA3 patients are heterozygous and one is homozygous for the alternative allele; the latter patient is also the one having the highest expression among all patients (Figure 7.14).

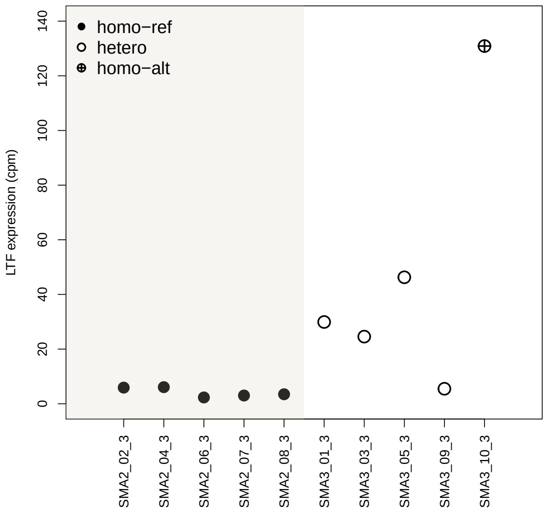


Figure 7.14: Lactotransferrin expression levels and *rs9110* genotypes in SMA patients. Expressions are measured “cpm” and SMA2 patients are highlighted by the grey area.

7.4 Hypotheses on SMA pathogenesis

Our analysis of differential expression between SMA patients and healthy controls led to the selection of genes associated with developmental disorders, diseases related to inflammatory and antimicrobial response and with nervous system development and functions. The main gene network reconstructed is associated with *NF-κB* (*nuclear factor kappa-light-chain-enhancer of activated B cells*) protein complex and *ubiquitin*, which have already been linked to muscle atrophy [195, 196]. IPA analysis also confirms the association of the selected DE genes with *SMN* genes and pathways related to post-transcriptional modifications [175], but further investigation is needful to unveil the biological mechanisms underling these relationships.

The analysis of differential gene expression between SMA2 and SMA3 patients reveals association with pathways affected by skeletal muscle and connective tissue disorders. Interestingly, expression patterns of the genes involved in these networks change between SMA2 and SMA3 patients, and might underlie protective mechanisms against the progression of SMA symptoms. Among these genes, we identified *Lactotransferrin* (*LTF*) as a promising target, due to its power to explain the phenotypic differences in the SMA patients considered in the present study. We found that *LTF* expression is significantly decreased in SMA2 patients, compared to both SMA3 patients and CTRL. Moreover, genotyping for *rs2073495*, a SNP located in *LTF* gene, provides a sharp separation of the two SMA phenotypes (homozigousity for the reference allele is only found in SMA2 patients) and have a marked relationship with *LTF* expression. Zhou et al. [197] tested the association of *LTF* expression with *rs9110* and *rs2073495*, plus two additional SNPs also located on the *LTF* gene: *rs1126477* and *rs116478*. They considered two groups of subjects, with 800 individuals each: patients with nasopharyngeal carcinoma (NPC) and healthy controls. In their study, only *rs9110* and *rs2073495* correlate with the phenotype, with frequencies of alternative alleles higher in NPC patients compared to controls. They also found higher expression of *LTF* gene in NPC patients, assayed at transcript and protein level using qPCR and western blot. Our results are in agreement with this study, confirming a positive relationship between the presence of the *rs2073495* alternative allele and an increase in *LTF* expression.

The potential of *LTF* as protective target is also supported by its central role in health and disease [198, 199, 200, 201]. *Lactotransferrin* (*LTF*), also called *lactoferrin* (LF), is an iron-binding glycoprotein that belongs to the transferrin family. *LTF* is produced by secretory epitheliums and by neutrophils, and has an important role in the host defense, functioning as antimicrobial agent and mediator of inflammatory responses. Moreover, the ability of *LTF* to bind large quantities of iron confers to it bacteriostatic and bactericidal properties [202, 203, 204, 205]. Besides its antimicrobial activity, *LTF* iron-dependent activity may also provide protection against pathogens by enhancing phagocytosis and inducing the release of pro-inflammatory cytokines [198]. *LTF* has an important modulatory action on the immune system (reviewed in [198]), by promoting maturation of T cell precursors and differentiation of immature B cells cells [206]. *LTF* also plays an important role in the *cytokine* cascade, modulating host defense against environmental insults in mammals [207, 208, 209]. Recent studies *in vitro* and *in vivo* propose *LTF* as a potential therapeutic agent for wound healing, since it enhances collagen gel contractile activity in human fibroblasts [210]. The role of *LTF* in host defense responses in infants has also been extensively studied [211, 212]: *LTF* is present at high concentration in human colostrum and is supposed to protect newborns against pathogens and to reduce the risk to be affected by microbe-induced gastroenteritis. Despite its central role in host defense in human, *LTF* level in blood is generally low (0.2-0.6 μg/ml) and increases only transiently in response to environmental insults [213]. Indeed, a high level of *LTF* in plasma has been proposed as predictive indicator of sepsis and has been related to autoimmune and chronic inflammatory disorders, such as Parkinson's and Alzheimer's diseases [214, 215, 216].

Studies also showed the accumulation of *LTF* in the lesions from different neurodegenerative diseases such as Down syndrome, Pick's disease and Alzheimer's disease [217, 218]. However, it is not clear if the increased production of *LTF* in these disorders is not sufficient to stop their progression or does not even activate the pathways defending the host from the harmful effects of the disorders. It is hypothesized that the initial up-regulation of *LTF* during acute inflammation is directed at capturing free iron, which would be a protective response against the damaging function of free radicals [198]. In neurodegenerative diseases, the accumulation of *LTF* around lesions reduces the neurotoxic effects of such lesions or deposits but is not sufficient to completely compensate the growing immunologic dissonance so to stop symptoms worsening. Several studies have reported evidences that *LTF* might modulate the progression of Parkinson's disease (PD), a neurodegenerative disorder characterized by a progressive loss of dopamine neurons. In particular, the analysis of postmortem sections of brain tissue from PD patients revealed that *LTF* was augmented in dopamine neurons resistant to the disease process [219, 220]. Rousseau and colleagues [221] used midbrain cultures and different experimental settings to model the loss of dopamine neurons characterizing PD, in order to study the effects of *LTF* on the progression of the degenerative process. They found that *LTF* has the potential to affect PD-mediated mechanisms of neurodegeneration and suggested that the accumulation of *LTF* in PD patients might be the sign of an attempt by the brain to combat ongoing neuronal insults. These studies report some evidences of the importance of *LTF* as protective agent against neuronal degeneration. The decreased expression in SMA2 patients compared to SMA3 patients detected in our assay might be a causal factor of SMA severity. Previous studies on Parkinson's disease have already found an inverse correlation between *LTF* in plasma and disease severity [215] and proposed to explore new therapies to elevate *LTF* levels in plasma [221].

7.5 Plans for biological validation and future research

With this analysis, we identified a set of genes related to skeletal muscle and connective tissue disorders, whose patterns of differential expression correlate with phenotype and may underlie protective mechanisms against SMA progression. Among these genes, lactotransferrin, which is significantly down-regulated in type-II SMA patients compared to both type-III SMA patients and healthy controls, represents a potential protective modifier. We also found a putative protective SNP located in lactotransferrin gene that correlates with its expression level. We hypothesize that this SNP is harmless in unaffected individuals, but plays an important role in SMA patients, inducing an increased expression of lactotransferrin and a milder phenotype. The protective mechanism which relates lactotransferrin to SMA progression still has to be characterized, but we suppose it to be based on its ability to bind iron and protect neurons from oxidative stress. The investigation of lactotransferrin expression and genotype as protective modifiers might help understanding the pathogenesis of SMA but might also led to the discovery of novel targets for diagnostic screening and therapy. Indeed, studies for the production and delivery of human lactotransferrin for the treatment of neuronal insults are already underway [222, 223, 224].

We plan to assay lactotransferrin expression and *rs2073495* genotype with qPCR in type-I, type-II and type-III SMA patients, considering at least ten subjects per phenotype, to assess their potentiality as phenotype predictors. Samples collection is currently in process at the Molecular Genetics Unit at Hospital Ramón y Cajal and qPCR will be performed by the CNIC Genomics Unit. We also intend to assay healthy mothers of SMA patients because we suppose that lactotransferrin level in the mother might be a protective factor for the embryos, activating a mechanism similar to the one described by Burghes and colleagues [174]. They studied the effects of the levels of maternal *SMN* protein over mice embryos lacking *Smn* gene, and found that death occurs early when the maternal *SMN* is reduced. In the same way, decreased lactotransferrin levels in the mother's blood might disrupt this protective mechanism and lead to manifestation of the most severe SMA phenotypes, which can indeed manifest even before birth [182]. Besides the relevance that reliable markers have in early screening, the investigation of these putative protective modifiers may enable to the development of effective treatments at the natal and pre-natal stage, that are most critical stages of denervation in the severe forms of SMA [225].

8

Extension of the computational pipeline to the analysis of RNA-seq time-series data

The analysis of gene expression and differential gene expression can give novel insights about the genes that are activated in some specific conditions of cells, providing a *snapshot* of the current transcriptional states of the organism under investigation. Anyway, we must remember that most of biological processes, including gene expression, are dynamic. With respect to *static* assays, time-series experiments (see section 2.5) allow investigating transient expression changes, but also characterizing gene expression regulation, coordination and interaction. Although most of time-series gene expression data sets have been generated using microarrays, the appealing features of NGS make RNA-seq a valuable alternative, with increasing number of RNA-seq time-series studies published in recent years [226]. However, further research is needed to clarify if the analysis methods developed for microarray data are suited for RNA-seq time-series experiments [226, 227]. Moreover, normalization issues in within-sample and between-sample comparison must be solved to ensure a correct data interpretation.

Since we intend to extend our computational pipeline to analyze dynamics RNA-seq data from time-series experiments, our research in the near future will be oriented both to the optimization of the current computational pipeline described in Chapter 6 and to its application to dynamic data. The latter objective will require the evaluation, development and integration of methods specifically designed for time-series data analysis. To this end, we designed two RNA-seq time-series data sets, one real and one simulated. We present in the following sections the description of the two data sets and the plan for computational data analysis.

8.1 Real data set: RNA-seq time-series from *sigE*-mutant and wild-type *Mycobacterium tuberculosis*

Background

Mycobacterium tuberculosis (MTB) is a pathogenic bacterial species and the major causative agent of tuberculosis, a common and often lethal infectious disease. Despite an encouraging reduction of new cases in the last decade, tuberculosis remains a major global health problem, with about 8.6 million people infected worldwide in 2012 and 1.3 million related deaths [228]. Lungs are the main target of tuberculosis, but other parts of the body can be also attacked. Tuberculosis can spread in the air, for example through cough or sneeze of infected people. Most infections are asymptomatic and latent, but about one in ten infections eventually progresses to active disease.

When MTB reaches human lungs, it is usually engulfed by the alveolar macrophage cells of the immune system (Figure 8.1). However, MTB has developed mechanisms to survive antimicrobial defense of macrophages, persisting for decades despite the host immune response. As a consequence, infected individuals develop latent tuberculosis: they remain healthy but carry dormant MTB bacteria. When the attack of the immune system diminishes, the pathogens can revive, leading to tuberculosis pulmonary infection. MTB can further infiltrate the bloodstream and spread to other organs.

Several studies have revealed that MTB persistence derives from the activation of a set of genes and metabolic pathways that allow surviving despite oxidative stress and starvation of nutrients in the intraphagosomal environment [229, 230, 231, 232, 233]. Investigation of the genes involved in the MTB response to these stress conditions may give new insights about tuberculosis pathogenesis and treatment. In particular, the intraphagosomal environment is characterized by reduced levels of inorganic phosphate, which is required for many essential cell-related processes, also in the MTB [230]. It is hypothesized that the σ^E factor is involved in the survival of MTB in conditions of phosphate starvation [230, 233]. Bacterial sigma factors are proteins that can interact with the RNA polymerase enzyme determining its binding to specific gene promoters. Bacteria generally have one principal sigma factor, which is required for the transcription of housekeeping genes, and several alternative sigma factors, which are not needed under normal physiologic conditions, but are activated under specific environmental stimuli [226]. The MTB genome encodes 13 sigma factors, 10 of which are involved in virulence and response to stress conditions [234, 235] and are called extracytoplasmic function (ECF) sigma factors [236]. One of the best-characterized ECF sigma factors in the MTB genome is σ^E, which is essential for growth in macrophages. It is induced (i.e. over-expressed) in conditions of phosphate starvation and transcribed in two isoforms [230, 233, 237]. However, the exact mechanism underlying MTB resistance under phosphate starvation has not been characterized yet and several questions remain unanswered (e.g. it is not clear if the two σ^E isoforms play different roles in MTB physiology) [230, 233].

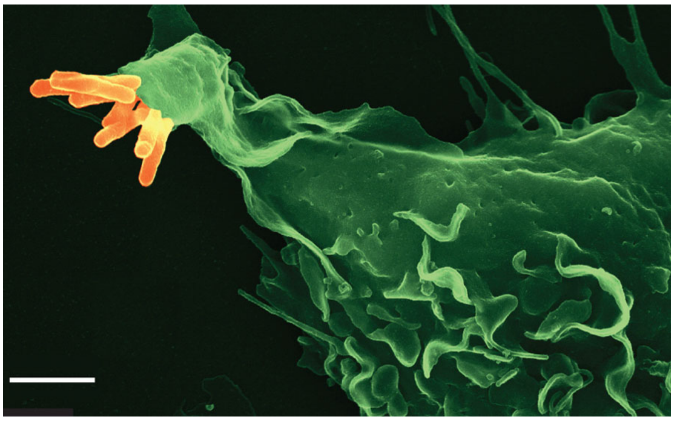
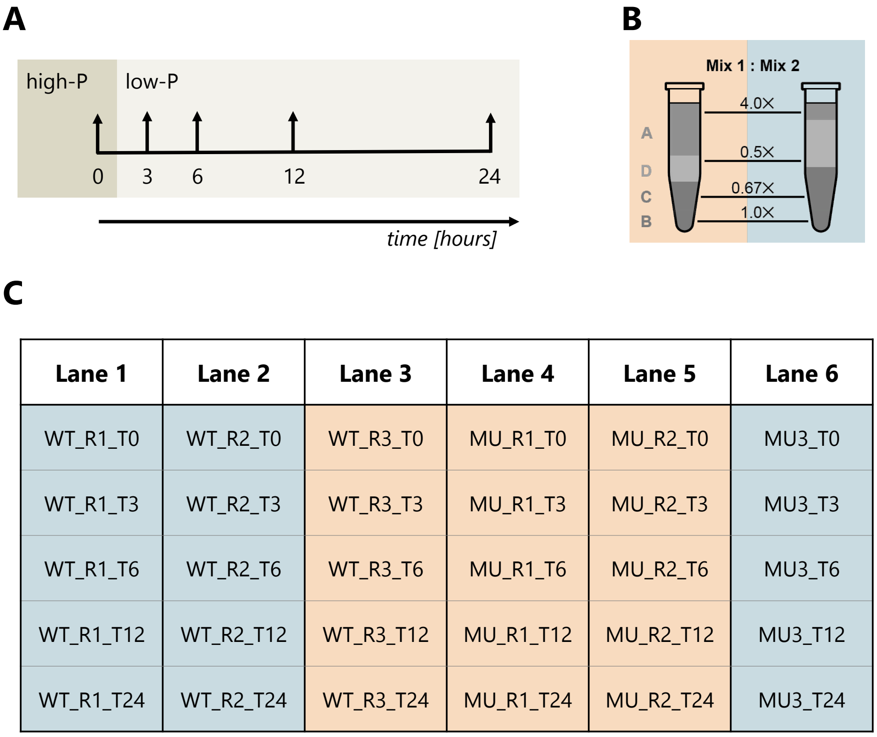


Figure 8.1: A macrophage (green) binds cells of *M. tuberculosis* (orange) through the TLR-2 receptor, to start engulfing. Image Courtesy of Volker Brinkmann, Core Facility Microscopy, Max Planck Institute for Infection Biology, Berlin, taken from [238].

Aim of the study and experimental design

The aim of the study is the characterization of the transcriptional response of MTB in conditions of phosphate (P) starvation, focusing on the identification of the genes regulated by σ_E factor. Particular attention will be directed to the identification of the main patterns of the MTB response and to the investigation of the underlying biological processes. With this purpose, we consider two strains of MTB: the wild-type H37RV and a *sigE*-mutant in which the functional *sigE* gene was deleted as in [237]. MTB cultures have been grown in P-rich substrate. At time $t = 0$, MTB cultures were washed three times in P-free broth and re-suspended in low-P substrate. RNA extraction was performed a time $t = 0$, in P-high conditions, and then in P-low conditions after 3, 6, 12 and 24 hours (Figure 8.2A). Triplicate cultures, for both wild-type and *sigE*-mutant underwent the whole experimental process, to obtain three biological replicates each. The changes of RNAs transcribed at times $t = 3, 6, 12, 24$, with respect to $t = 0$, represent the response of MTB bacteria to phosphate starvation. In particular, the genes that are differentially expressed with respect to $t = 0$ only in the wild-type strain, may underlie pathways controlled by σ^E factor. RNAs changes over time, for each biological replicate of the wild-type and mutant cultures (30 samples in total), have been be assayed through RNA-seq using the Illumina HiSeq sequencer. The experiment has been be conducted in multiplexing, by sequencing five samples per lane and six lanes in total, to obtain 2x100 bp paired-end reads.



Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6
WT_R1_T0	WT_R2_T0	WT_R3_T0	MU_R1_T0	MU_R2_T0	MU3_T0
WT_R1_T3	WT_R2_T3	WT_R3_T3	MU_R1_T3	MU_R2_T3	MU3_T3
WT_R1_T6	WT_R2_T6	WT_R3_T6	MU_R1_T6	MU_R2_T6	MU3_T6
WT_R1_T12	WT_R2_T12	WT_R3_T12	MU_R1_T12	MU_R2_T12	MU3_T12
WT_R1_T24	WT_R2_T24	WT_R3_T24	MU_R1_T24	MU_R2_T24	MU3_T24

Figure 8.2: Experimental and sequencing design. RNA extraction (A) from high-phosphate (*high-P*) cultures at time $t = 0$, and from low-phosphate (*low-P*) cultures, after 3, 6, 12 and 24 hours. Differentially-expressed spike-in RNA mixtures (B) and sequencing design (C).

Before sequencing, two mixtures of differentially expressed spike-in RNAs have been added to each sample, according to the scheme depicted in Figure 8.2C. As explained in Chapter 3, spike-in RNAs are standard transcripts with known sequences and concentrations. In particular, here we consider the Ambion ERCC Spike-In Control Mixes (Life technologies [31]), two mixtures of spike-in RNAs, present at defined “Mix1:Mix2” molar concentration ratios described by four subgroups (Figure 8.2B). Each subgroup contains

23 transcripts spanning a 10^6-fold concentration range. Transcripts have different lengths and GC-contents, with similar distributions within each subgroup.

The experiment and the sequencing design, such as the selection of the optimal trade-off between cost and number of samples, the use of spike-in standards and of paired-end protocol, has been an integral part of the Ph.D. activity. The sequencing of the 30 samples with MTB and spike-in transcripts has just been completed (January 2014) at BMR Genomics (BMR Genomics, Padova, Italy).

8.2 Simulated data set

To simulate time-dependent expression profiles, we used an approach similar to that adopted in [239]. Each profile $P_{\rho\varepsilon}$ represents the log-fold-changes of expression levels over time, with respect to the basal level at $t = 0$, induced by the treatment ε. For a given transcript f, assayed in condition ε (i.e. "treated" or "control"), the time-series reflecting the number of copies present at time t can be modeled, on log-scale, as follows:

$$\log_2\left(\theta_{f\varepsilon}(t)\right) = k_f \cdot P_{\rho\varepsilon}(t) + q_f \tag{8.1}$$

$P_{\rho\varepsilon}(t)$ is the temporal pattern reflecting the changes in transcript levels in response to treatment, while k_f and q_f are transcript-specific parameters. $P_{\rho\varepsilon}(t)$ is non-null only for differentially expressed (DE) transcripts in the "treated" samples. Thus, for non-DE transcripts and transcripts assayed in the "control" condition, equation 8.1 reduces to:

$$\log_2\left(\theta_{f\varepsilon}(t)\right) = q_f. \tag{8.2}$$

Thus, for most of the transcripts, q_f represents the level of expression on the log-scale.

Expressions in the natural scale $\theta_{f\varepsilon}(t)$ can be computed from equation 8.1.

Assuming the same length for all simulated transcripts, the probability that a read comes from some transcript f can be computed, for each time point t, as in equation 2.2:

$$\pi_{f\varepsilon} = \frac{\theta_{f\varepsilon}}{\sum\limits_{f=1}^{F} \theta_{f\varepsilon}} \tag{8.3}$$

$\pi_{f\varepsilon}$ can be in turn used to obtain the final transcripts counts, using a NB distribution $\mathcal{NB}(R \cdot \pi_{f\varepsilon}, \phi)$, where R is the sequencing depth and ϕ is the dispersion parameter.

We performed 100 simulations, generating 100 independent time-series data sets, using the scheme described above. In particular, we simulated six different profiles $P_{\rho\varepsilon}$, with $\rho = 1, \ldots, 6$, considering 13 time points, $t = 0, \ldots, 12$. q_f was sampled from a normal distribution $\mathcal{N}\left(-2.70, (1.44)^2\right)$ according to [124]. k_f was instead sampled from a uniform distribution in the interval $\pm(0.5, 2)$. Counts were simulated for $F = 1000$ transcripts, with 880 non-DE transcripts and 120 DE transcripts. In particular, each single pattern was used to simulate time-series profiles for 20 DE transcripts. The sequencing depth was set to $R = 5 \cdot 10^4$ reads for all replicates and the dispersion parameter was selected to be common to all transcripts and fixed to $\phi = 0.1$, to simulate the biological variability due to different cell cultures (Figure 8.4). Further simulations will be performed considering different values of the dispersion parameter.

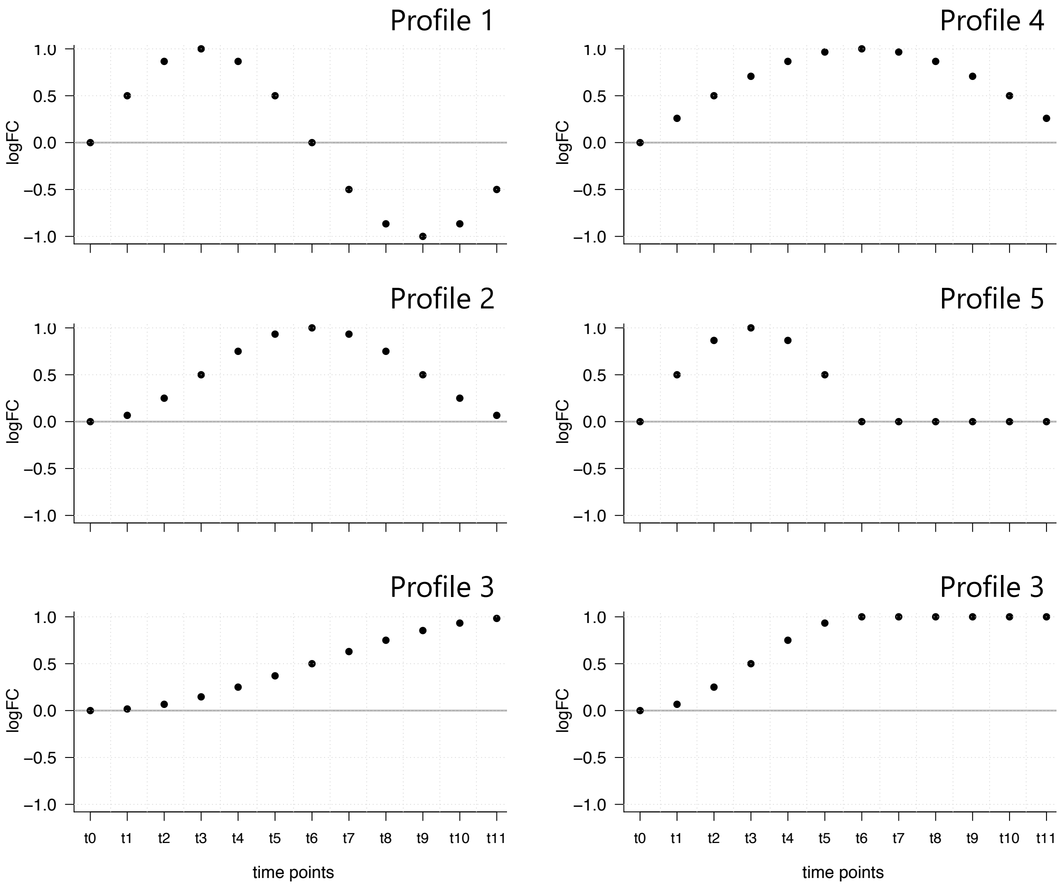
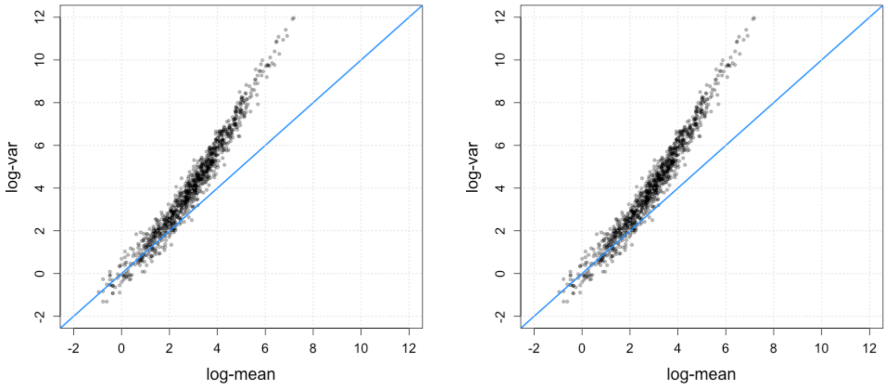


Figure 8.3: Simulated profiles representing log-fold-changes (logFC) of expression levels over time with respect to the steady state at $t = 0$.

Figure 8.4: Example of mean-variance relationship (on log-scale) of transcript counts in one synthetic data set, for "treated" and "control" conditions. The blue line represents the Poisson condition of mean-variance equality. NB distributed data are over-dispersed, i.e. their variance exceeds Poisson variance.



(a) Treated (b) Control

8.3 Data analysis plan

Reads from the real data set will be pre-processed as in Chapter 6, and expression measures will be computed at exon and gene level using totcounts and maxcounts (Chapter 5). The subsequent analyses will be tailored on the specific data features: the normalization strategy will be optimized leveraging on the information provided by spike-in RNAs, while several methods will be compared to define the time-series analysis approach. The latter assessment will be carried out considering both the *M. tuberculosis* data set, the synthetic counts and additional real data sets available in the literature, such as [86, 240]. We are currently optimizing and testing a strategy based on the algorithms defined in [135, 239] for integrating gene selection, clustering and functional annotation of RNA-seq time-series data. In addition, the true concentrations of spike-in RNAs sequenced in the *M. tuberculosis* study will be used as gold-standard to deepen the benchmarking and comparison of maxcounts and totcounts approaches in the presence of biological replicates. Finally, the defined pipeline will be applied to answer the biological questions motivating the *M. tuberculosis* study, to reveal which gene pathways are activated, together with *sigE*, in phosphate starvation conditions and which genes are regulated by σ_E factor. The single-base resolution of RNA-seq will also enable to investigate expression at exon level, and possibly to characterize the function of the two σ^E isoforms.

9
Concluding remarks

Thanks to the advent and progress of NGS technologies, RNA-seq has rapidly become the method of choice for measuring and comparing gene transcription levels. NGS platforms produce millions of short sequences, which are read from the input DNA or RNA and are indeed called *reads*. By mapping reads on a reference genome, the complete "transcriptional map" of the genome under investigation can be revealed. In practice, the expression of a coding unit, such as a gene, transcript or exon, is estimated by *counts*, that is the number of reads that can be aligned on its sequence. Counts can also be compared between different conditions to identify differentially expressed genes. At first glance, this analysis scheme may seem very simple, but its implementation is in fact complex and not well defined. So far, many computational methods have been proposed, but a standard analysis pipeline has not been defined yet.

The main aim of this thesis was the definition of a robust analysis pipeline for measuring and comparing gene expression levels in human studies based on RNA-seq. The definition of a probabilistic model of counts, plus the study of state-of-the art literature allowed us to select, among the top-ranking methods, robust computational solutions for read mapping and differential expression analysis. Given these selected methods, which allowed us to outline a first processing framework, we implemented additional assessments to define a robust strategy to cope with the most challenging aspects of RNA-seq analysis: count bias and multireads. Our results highlight that counts are characterized by a strong length-bias, which cannot be completely removed with current normalization methods without introducing further systematic errors. Thus, we defined a new measure of counts, called *maxcounts*, computed as the maximum read coverage along an exon. We compared our strategy with the standard approach, revealing that it reduces length bias, counts non-uniformity due to highly expressed genes and technical variance, and that it is robust to the upstream pre-processing and mapping steps. This analysis also allowed us to define and implement a pre-processing module in the computational pipeline, in order to both provide the mapping algorithms with high-quality reads, so to reduce data loss (i.e. unmapped reads) and maximize the fraction of correctly mapped reads. This aspect was further refined with our comparative analysis of strategies for multireads handling. Multireads are reads that align to multiple locations of the reference genome, resulting in biased gene expression estimates. Using different strategies based on RSEM algorithm [79], we found that single-end mapping of pre-processed reads leads to accurate alignments while limiting computational load.

Finally, the implemented analysis pipeline was applied to a real case study: thirty samples, from patients with spinal muscular atrophy (SMA) and healthy controls, were analyzed to identify the causative genes involved in SMA pathogenesis. SMA is a degenerative neuromuscular disease that has no cure and represents one of the major genetic causes of infant mortality. To define a list of differentially expressed genes, we used jointly the standard approach for count computation and the maxcounts strategy, so to control false-positive rate. The comparison of the most severe phenotype (type-II SMA) with the milder one (type-III SMA), led to the detection of differentially expressed genes associated to disorders of skeletal muscle and connective tissue. Moreover, the differences in the expression levels shown for healthy controls, type-II and type-III patients reveal common expression patterns across the three phenotypes that may underlie protective mechanisms against SMA progression. Some putative positive targets identified by this analysis, are currently under biological validation since they might improve diagnostic screening and therapy.

The different analyses carried out confirm that the implemented pipeline is stable to RNA-seq bias and variability. In particular, maxcounts approach can overcome the bias due to the non-uniformity of read coverage, selecting the best-represented transcript regions for estimating expression levels. A possible limitation of the current implementation is represented by the use of exons, since the final user might be interested in a having gene or transcript counts. Moreover, the lacking of RNA-seq studies with sufficient sample-size and reliable gold-standard measures prevented a deeper characterization of maxcounts approach and the evaluation of its application to differential expression analysis. For the

SMA study, we adopted a practical strategy to exploit both approaches and to transform exon counts into gene-wise expression estimates. However, our future work will focus on the definition of transcription models that can be used to combine exon maxcounts into an accurate measure of gene or transcript expression. Moreover, the RNA-seq data set with differentially expressed spike-in RNAs that we designed will allow us to deepen the comparison of the maxcounts strategy with the standard approach. Nevertheless, RNA-seq is a methodology still under active development, which will experience fast improvement of experimental protocols and data features. Thus, we made available the codes for calculating maxcounts, implemented in Perl and C++, enabling its benchmarking on different data sets.

Other interesting computational questions have risen from the comparison of single-end and paired-end mapping strategies. The limited extent of our assessment, which was implemented to optimize the mapping module of our analysis framework, does not allow drawing definitive insights. One of the issues emerged from the current assessment that would be worth investigating in the future, is the capability of mapping algorithms to exploit the whole information contained in paired-end data. Developed for *de novo* genome assembly, paired-end reads are appreciated for their power in solving repeats and assembly low-complexity regions, but it is not clear if they are beneficial also for RNA-seq data analysis. In our assessment, we noted a reduced number of reads mapped by the paired-end algorithm with respect to the single-end version, which is confirmed by the literature studies reviewed. This data loss, which could be due to the stringency of the constraints for paired-end mapping, may be harmful in RNA-seq studies since it could result in biased expression estimates. To further investigate this aspect, we intend to perform a systematic comparison of different mapping algorithms, run on paired-end data and tested in single-end and paired end mode and with different parameter settings. More sophisticated hybrid strategies, that blend the precision of the paired-end approach in isoform reconstruction with the sensitivity of the single-end strategy in expression estimate, may also be exploited.

Finally, we intend to extend our computational pipeline to analyze dynamic RNA-seq data from time-series experiments. RNA-seq methodology is expected to be extensively applied to the analysis of time-series analysis in the near future. However, limited research has been carried out to assess which methods can be used to analyze count-based time-series and the challenges posed by normalization must be definitively solved to ensure a correct data interpretation. With this purpose, we designed one real time-series data set, from transcripts extracted from *M. tuberculosis* and mixed with differentially expressed spike-in RNAs (at known concentrations), and a simulated count data set. We defined this specific experimental design also to allow a deeper characterization and benchmarking, thanks to spike-in RNAs, of all the methods considered in this thesis, with particular attention to maxcounts. In the same way, the real experimental data set and the synthetic counts (possibly simulated with different dispersion values) will be used to test different methods for time-series data analysis, in the presence of biological variance. We are currently optimizing and testing a strategy for integrating gene selection, clustering and functional annotation of RNA-seq time-series data based on previous works of Di Camillo *et al*. [135, 239], but other approaches based on models of the count mean-variance relationship will be also considered. The final pipeline will be applied to answer the biological questions motivating the *M. tuberculosis* study, to reveal which pathways are activated to ensure bacterial persistence in condition of phosphate starvation.

A

Burrows-Wheeler transform and FM-index for fast string matching

A.1 The Burrows-Wheeler transform

The Burrows-Wheeler transform (BWT) is a reversible permutation of the characters in a text string, proposed in 1994 by M. Burrows and D. J. Wheeler [241]. It has been originally developed to address the problem of data compression: the permutation makes the original string easy to compress with move-to-front coding algorithms. In the following, the algorithm used to calculate the BWT of a text string (*forward BWT*) and the one used to retrieve the original string starting from its BWT (*reverse BWT)* are described.

Given a string S of N characters $S[1], ..., S[N]$, the BWT algorithm first considers the N rotations of S, sorted lexicographically. Conceptually, all N cyclic shifts of S can be represented by the rows of a matrix M. As an example, consider the string $S =$ 'mississippi', $N = 11$, whose characters are taken from the alphabet $\alpha = \{$'i', 'm', 'p', 's'$\}$. The matrix M is:

row	M
1	imississipp
2	ippimississ
3	issippimiss
4	ississippim
5	mississippi
6	pimississip
7	ppimississi
8	sippimissis
9	sissippimis
10	ssippimissi
11	ssissippimi

At least one of the rows, numbered from zero, contains the original string S. Let I be the first row containing S. In our example, $I = 5$. Let L be the string formed by the characters of the last column of M: $L[1], ..., L[N]$. The final output of the transformation is the pair of (L, I), even though it is often indicated simply as L. In our example, $L =$ 'pssmipissii' and $I = 5$. The transformation described above can be easily reversed by means of an algorithm that uses (L, I) to reconstruct the original string S. The algorithm begins calculating the first column F of the matrix M. As M rows are lexicographically sorted, F is obtained by sorting the characters in L. In our example, $F =$ 'iiiimppssss'. It is worth noting that all columns of M, among which L and F, are permutation of the original string S. To better understand the structure of M, we can consider a matrix M' formed by shifting M rows one char to the right. The L string is now the first column of M' and precedes F. In our example, M and M' are:

row	M	M’
1	imississipp	pimississip
2	ippimississ	sippimissis
3	issippimiss	sissippimis
4	ississippim	mississippi
5	mississippi	imississipp
6	pimississip	ppimississi
7	ppimississi	ippimississ
8	sippimissis	ssippimissi
9	sissippimis	ssissippimi
10	ssippimissi	issippimiss
11	ssissippimi	ississippim

As rows in M' are sorted starting from their second character, if we consider the rows that start with some character c, then they must appear in lexicographical order. Consequently, for any given character c, the rows of M that begin with c appear in the same order as the rows in M' beginning with c. Consider for instance the strings beginning with ‘i’: the rows 1, 2, 3, 4 in M correspond to the rows 5, 7, 10 and 11 in M'. From the previous example it is possible to notice an interesting property of M, called ‘last first (LF) mapping’: the k^{th} occurrence of a character c in the last column L corresponds to the same character as the k^{th} occurrence in the first column F. Exploiting this property, we can calculate the vector T that indicates the correspondence between the rows of the two matrices (and consequently between F and L), so that row j of M' corresponds to row $T[j]$ of M, for each $j = 1, ..., N$. If $L[j]$ is the k^{th} instance of c in L, then $T[j] = i$ where $F[i]$ is the k^{th} instance of c in F. Thus, the correspondence between elements of F and elements of L is given by:

$$F[T[j]] = L[j] \tag{A.1}$$

In our example, using F and L, T is calculated as follows:

j	1	2	3	4	5	6	7	8	9	10	11
F	i	i	i	i	m	p	p	s	s	s	s
L	p	s	s	m	i	p	i	s	s	i	i
T	6	8	9	5	1	7	2	10	11	3	4

For each $i = 1, ..., N$, the characters $L[i]$ and $F[i]$ are the last and the first characters of the row i of M. Since the rows of M are rotations of the original string S, the character $L[i]$ cyclically precedes the character $F[i]$ in S. Consequently, $L[T[j]]$ cyclically precedes $L[j]$ ($L[j] = F[T[j]]$ from Equation A.1) in S. Finally, using L, T and I, the original string S is generated back-to-front. Indeed, given that row I of M is the original string S, $L[I]$ is exactly the last character of S. The predecessors of each character are then calculated for each $i = 1, ..., N$ using T as follows:

$$S[N-i] = L[T^i[I]], \tag{A.2}$$

where $T^1[x] = x$ and $T^{i+1}[x] = T[T^i[x]]$ for $i > 1$.

In our example, we can define $y = T^i[x]$ ($y = x$ for $i = 1$) at each iteration and calculate S back-to-front as follow:

i	y	$T[y]$	$L[T[y]]$
1	I=5	5	i
2	5	1	p
3	1	6	p
4	6	7	i
5	7	2	s
6	2	8	s
7	8	10	i
8	10	3	s
9	3	9	s
10	9	11	i
11	11	4	m

Thus, $S =$ ‘mississippi’.

Both the forward and and the reverse BWT can each be performed in linear time and linear space in the worst case [242].

A.2 The FM-index

In 2000 Ferragina and Manzini proposed a novel data structure called FM-index (“FM” stands for “Full-text index in a Minute space”) [243] to address the issue of compressing and indexing data. Their method combines the BWT compression algorithm with the structural properties of the suffix array. The suffix array A built on the string S, is an array containing all the suffixes of S, lexicographically ordered and represented through pointers to their starting positions. In our example, $S =$ ‘mississippi’ and $A = [11, 8, 5, 2, 1, 10, 9, 7, 4, 6, 3]$. Ferragina and Manzini proposed to apply the BW-transform on S after appending at its end a special character ‘\$’, lexicographically smaller than all characters of the alphabet α. They discovered a strong relation between the matrix M, whose rows are the all lexicographically sorted cyclic shifts of $S\$$, and the suffix array A of $S\$$. The sorted rows of M correspond to the sorted suffixes of A and, consequently, the entry $A[i]$ points to the suffix of $S\$$ occupying a prefix of the i^{th} row of M.

suffix	A	M	row
$	12	$mississippi	1
i$	11	i$mississipp	2
ippi$	8	ippi$mississ	3
issippi$	5	issippi$miss	4
ississippi$	2	ississippi$m	5
mississippi$	1	mississippi$	6
pi$	10	pi$mississip	7
ppi$	9	ppi$mississi	8
sippi$	7	sippi$missis	9
sissippi$	4	sissippi$mis	10
ssippi$	6	ssippi$missi	11
ssissippi$	3	ssissippi$mi	12

The string obtained by applying the BTW to $S\$$, S_{bw}, is then compressed in three steps:

1. Move-to-front encoding of each character c by counting distinct characters seen since its previous occurrence.

2. Run-length encoding of each run of zeroes: the sequence 0^m is replaced by the integer $(m+1)$, written in binary, least significant bit first and discarding the most significant bit.

3. Compression of the resulting string by means of variable-length prefix code, to obtain the final string S_{rlx}.

Ferragina and Manzini also developed a new algorithm, called `Backward Search`, to support fast pattern retrieval on S_{bw}. Although the original version of the algorithm exploits the compressed string S_{rlx}, the applications of FM-index for short-sequence mapping just consider S_{bw} without performing the compression steps. Therefore, in the following section we will discuss the `Backward Search` algorithm performed on S_{bw} while referring to [244] for a detailed description of the compression algorithm.

The *Backward Search* algorithm

The suffix array A has two interesting structural properties:

i. All the suffixes of the string S prefixed by a pattern P of length p, $P[1,p]$, occupy a contiguous portion (subarray) of A;

ii. That subarray has starting position sp and ending position ep, where sp is the lexicographic position of the string P among the ordered sequence of text suffixes.

The `Backward Search` algorithm (see algorithm 1) identifies the positions sp and ep by accessing only S_{bw} (S_{rlx} in the original version of the algorithm) and some auxiliary array-based data structures: $C(\cdot)$ and $O(\cdot, \cdot)$. The array $C(1, ..., |\alpha|)$ stores in $C(c)$ the number of occurrences in $S\$$ of characters that are lexicographically smaller than c. In our example, $S\$$ =‘mississippi$’ and $C('i',' m', p',' s') = [1, 5, 6, 8]$. The matrix $O(c, k)$ reports the number of occurrences of c in $S_{bw[1,k]}$. For example, with $S_{bw} =$ ‘ipssm$pissii’, $O('s', 5) = 2$, $O('s', 12) = 4$ and $O('p', 12) = 2$. $O(\cdot, \cdot)$ and $C(\cdot)$ can be easily precomputed and, together with S_{bw}, form the FM-index of S.

Algorithm 1 `Backward Search`$(P(1, p))$

```
$c = P(p);$
$i = p;$
$sp = C(c) + 1;$
$ep = C(c + 1);$
while $((sp < ep)$ and $(i \geq 2))$ do
    $c = P(i - 1);$
    $sp = C(c) + O(c, sp - 1) + 1;$
    $ep = C(c) + O(c, ep);$
    $i = i - 1;$
end while
if $(ep < sp)$ then
    return "not found"
else
    return "found $(ep - sp + 1)$ occurrences"
end if
```

The `Backward Search` algorithm counts the number of occurrences of $P(1, p)$ in S and consists of p steps; at the i^{th} step sp and ep point to the first and the last row of M prefixed by $P(i, p)$, respectively. When the algorithm completes, rows beginning with $P(1, p)$, the entire query, correspond to the occurrences of the query in S. If the range is empty, S does not contain the query. In Figures A.1 and A.2 are shown two examples of pattern search performed with the `Backward Search` algorithm.

The running time of the `Backward Search` algorithm depends on the cost of the procedure used to calculate $O(\cdot, \cdot)$; the authors described in [243] an algorithm for computing $O(c, k)$ in $O(1)$ time.

Figure A.1: Example of pattern search with the `Backward Search` algorithm. Searching for $P =$ 'ssi' ($p = 3$) in $S =$ 'mississippi'. After two iterations, the algorithm finds $(ep - sp + 1) = 1$ occurrence of 'ssi' in 'mississippi'.

	Initialization	Iteration No. 1	Iteration No. 2
c	'i'	's'	's'
query	'i'	'si'	'ssi'
sp	2	9	11
ep	5	10	12
i	3	2	1

Initialization			Iteration No. 1			Iteration No. 2		
	row	M		row	M		row	M
	1	\$mississippi		1	\$mississippi		1	\$mississippi
$sp \rightarrow$	2	i\$mississipp		2	i\$mississipp		2	i\$mississipp
	3	ippi\$mississ		3	ippi\$mississ		3	ippi\$mississ
	4	issippi\$miss		4	issippi\$miss		4	issippi\$miss
$ep \rightarrow$	5	ississippi\$m		5	ississippi\$m		5	ississippi\$m
	6	mississippi\$		6	mississippi\$		6	mississippi\$
	7	pi\$mississip		7	pi\$mississip		7	pi\$mississip
	8	ppi\$mississi		8	ppi\$mississi		8	ppi\$mississi
	9	sippi\$missis	$sp \rightarrow$	9	sippi\$missis		9	sippi\$missis
	10	sissippi\$mis	$ep \rightarrow$	10	sissippi\$mis		10	sissippi\$mis
	11	ssippi\$missi		11	ssippi\$missi	$sp \rightarrow$	11	ssippi\$missi
	12	ssissippi\$mi		12	ssissippi\$mi	$ep \rightarrow$	12	ssissippi\$mi

Figure A.2: Example of pattern search with the `Backward Search` algorithm. Searching for $P =$ 'psi' ($p = 3$) in $S =$ 'mississippi'. After two iterations, the algorithm finds no occurrences of 'psi' in 'mssissippi'($ep < sp$).

	Initialization	Iteration No. 1	Iteration No. 2
c	'i'	's'	'p'
query	'i'	'si'	'psi'
sp	2	9	9
ep	5	10	8
i	3	2	1

Initialization			Iteration No. 1			Iteration No. 2		
	row	M		row	M		row	M
	1	\$mississippi		1	\$mississippi		1	\$mississippi
$sp \rightarrow$	2	i\$mississipp		2	i\$mississipp		2	i\$mississipp
	3	ippi\$mississ		3	ippi\$mississ		3	ippi\$mississ
	4	issippi\$miss		4	issippi\$miss		4	issippi\$miss
$ep \rightarrow$	5	ississippi\$m		5	ississippi\$m		5	ississippi\$m
	6	mississippi\$		6	mississippi\$		6	mississippi\$
	7	pi\$mississip		7	pi\$mississip		7	pi\$mississip
	8	ppi\$mississi		8	ppi\$mississi	$ep \rightarrow$	8	ppi\$mississi
	9	sippi\$missis	$sp \rightarrow$	9	sippi\$missis	$sp \rightarrow$	9	sippi\$missis
	10	sissippi\$mis	$ep \rightarrow$	10	sissippi\$mis		10	sissippi\$mis
	11	ssippi\$missi		11	ssippi\$missi		11	ssippi\$missi
	12	ssissippi\$mi		12	ssissippi\$mi		12	ssissippi\$mi

B

Differentially expressed genes in spinal muscular atrophy

B.1 Differentially expressed genes

SMA *vs.* CTRL

Table B.1: Differential expressed genes selected for the "SMA versus CTRL" comparison: Ensembl gene IDs, log-ratios and q-values, groups of *overlapping-genes*, labelled with the same group number, and genes confirmed by the *totcounts-genes* approach.

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000172062	-1.85	-2.08	1.90E-15	1.62E-14		yes
ENSG00000198566	-7.73	-10.05	9.03E-10	7.07E-19		yes
ENSG00000171116	-5.26	-4.69	2.60E-08	1.72E-05		yes
ENSG00000198618	-2.23	-2.2	3.40E-08	2.15E-06		yes
ENSG00000228224	-1.34	-1.44	4.05E-07	6.47E-07		yes
ENSG00000212769	-2.34	-2.44	1.34E-05	1.43E-05		yes
ENSG00000162368	0.88	0.74	1.34E-05	1.19E-03		yes
ENSG00000240869	-1.49	-1.58	2.23E-05	2.15E-06		yes
ENSG00000252488	1.3	1.21	2.54E-05	3.03E-04		yes
ENSG00000253676	1.57	1.41	4.22E-05	1.57E-03		yes
ENSG00000251948	-1.06	-1.16	4.39E-05	2.15E-06		yes
ENSG00000232162	1.68	1.57	4.39E-05	8.49E-03		yes
ENSG00000243742	-1	-1	5.10E-05	2.05E-04		no
ENSG00000210144	0.96	0.87	5.38E-05	1.45E-03	1	yes
ENSG00000210140	0.96	0.87	5.38E-05	1.45E-03	1	yes
ENSG00000168685	0.96	0.9	6.11E-05	7.09E-04		yes
ENSG00000221970	1.2	1.14	6.92E-05	2.00E-03		yes
ENSG00000205571	0.87	0.72	1.31E-04	1.07E-02		yes
ENSG00000233251	0.8	0.75	2.55E-04	1.57E-03		yes
ENSG00000231006	1.11	1.04	4.98E-04	1.96E-03		yes
ENSG00000123091	0.68	0.65	5.28E-04	4.81E-04		yes
ENSG00000258988	-2.49	-2.58	5.28E-04	1.46E-03	2	yes
ENSG00000172717	-2.49	-2.58	5.28E-04	1.46E-03	2	yes
ENSG00000259499	1.09	1	5.28E-04	2.76E-03		yes
ENSG00000240223	0.71	0.61	5.28E-04	2.86E-03		yes
ENSG00000158525	1.79	1.86	5.28E-04	4.53E-03		yes
ENSG00000152558	0.88	0.72	5.28E-04	6.38E-03		yes
ENSG00000123179	0.73	0.59	5.28E-04	1.33E-02		no
ENSG00000238179	-0.87	-1.05	5.40E-04	8.40E-05		yes
ENSG00000151135	1.08	1.16	5.40E-04	7.31E-04		yes
ENSG00000230408	0.67	0.57	5.94E-04	7.32E-03		yes
ENSG00000012660	0.73	0.62	6.02E-04	3.65E-03		yes
ENSG00000258511	1.02	1.15	6.51E-04	3.17E-04		yes
ENSG00000201592	0.9	0.81	6.51E-04	4.93E-03		yes
ENSG00000201882	0.9	0.81	6.51E-04	4.93E-03		yes
ENSG00000163519	1.22	1.13	6.51E-04	5.85E-03		yes
ENSG00000202434	0.82	0.73	6.51E-04	5.97E-03		yes
ENSG00000201121	1.16	1.1	6.51E-04	6.67E-03		yes
ENSG00000230272	0.78	0.69	6.51E-04	7.19E-03		yes
ENSG00000135535	0.66	0.5	6.51E-04	1.11E-02		no
ENSG00000035115	0.89	0.65	6.51E-04	2.44E-02		yes
ENSG00000236953	1.1	1.03	6.55E-04	5.54E-03		yes
ENSG00000211619	-7.1	-8.08	6.72E-04	2.57E-03		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000232686	1.06	0.94	8.08E-04	5.51E-03		yes
ENSG00000227907	1.33	1.24	8.08E-04	6.74E-03		no
ENSG00000143742	0.8	0.66	8.08E-04	7.40E-03		yes
ENSG00000238982	0.96	0.88	8.08E-04	7.85E-03		yes
ENSG00000236434	0.77	0.68	8.08E-04	8.49E-03		yes
ENSG00000234374	1.07	0.97	8.08E-04	1.27E-02		yes
ENSG00000142875	1.26	1.38	8.40E-04	2.52E-04		yes
ENSG00000201820	1.01	0.91	8.40E-04	5.66E-03		yes
ENSG00000225300	1.12	1.01	8.40E-04	1.74E-02		no
ENSG00000100528	0.64	0.5	8.40E-04	2.07E-02		yes
ENSG00000206650	0.75	0.65	9.56E-04	6.36E-03		yes
ENSG00000132204	-1.17	-1.23	9.68E-04	1.81E-03		yes
ENSG00000171811	1.78	1.61	9.98E-04	3.06E-02		no
ENSG00000239087	1.32	1.25	1.02E-03	6.40E-03		yes
ENSG00000148700	0.76	0.66	1.02E-03	1.87E-02		yes
ENSG00000234040	0.75	0.65	1.07E-03	6.86E-03		yes
ENSG00000251805	1.06	0.99	1.14E-03	8.49E-03		yes
ENSG00000252620	0.61	0.51	1.15E-03	1.42E-02		yes
ENSG00000151414	1.02	1.05	1.18E-03	2.23E-03		yes
ENSG00000130600	-2.12	-2.52	1.18E-03	5.54E-03	3	yes
ENSG00000211502	-2.12	-2.52	1.18E-03	5.54E-03	3	yes
ENSG00000235945	0.94	0.85	1.18E-03	6.47E-03		yes
ENSG00000233406	0.74	0.64	1.18E-03	1.28E-02		yes
ENSG00000222494	0.9	0.82	1.19E-03	1.20E-02		yes
ENSG00000110675	1.04	0.74	1.19E-03	3.26E-02		yes
ENSG00000160307	1.33	1.43	1.21E-03	1.45E-03		yes
ENSG00000196937	0.83	0.8	1.21E-03	2.00E-03		yes
ENSG00000234925	0.74	0.64	1.21E-03	7.72E-03		yes
ENSG00000198160	0.92	1.01	1.24E-03	2.12E-03		yes
ENSG00000150681	0.88	0.8	1.24E-03	7.91E-03		yes
ENSG00000175147	-1.17	-1.12	1.27E-03	2.19E-05		yes
ENSG00000261655	-0.9	-1.02	1.28E-03	4.74E-04		yes
ENSG00000134294	0.93	0.88	1.29E-03	3.86E-03		yes
ENSG00000015479	0.78	0.69	1.36E-03	7.09E-03		yes
ENSG00000156738	1.02	0.88	1.36E-03	8.21E-03		yes
ENSG00000122026	1.22	1.16	1.44E-03	2.23E-03	4	yes
ENSG00000207500	1.22	1.16	1.44E-03	2.23E-03	4	yes
ENSG00000207051	1.22	1.16	1.44E-03	2.23E-03	4	yes
ENSG00000212829	-1.37	-1.11	1.52E-03	1.35E-02		yes
ENSG00000211589	0.84	0.77	1.67E-03	1.49E-02		yes
ENSG00000252904	1.45	1.41	1.71E-03	1.11E-02		yes
ENSG00000080546	0.78	0.68	1.73E-03	1.39E-02		yes
ENSG00000243005	0.64	0.56	1.76E-03	1.58E-02		yes
ENSG00000134884	0.91	1.05	1.79E-03	1.48E-03		yes
ENSG00000252700	0.87	0.78	1.79E-03	1.20E-02		yes
ENSG00000238924	0.9	0.81	1.82E-03	8.92E-03		yes
ENSG00000153130	1.21	1.37	1.88E-03	1.13E-03		yes
ENSG00000164823	0.71	0.83	1.88E-03	2.00E-03		yes
ENSG00000199545	0.89	0.79	1.88E-03	9.13E-03		yes
ENSG00000160789	-1.05	-1.07	2.01E-03	5.92E-03		yes
ENSG00000179979	-0.72	-0.9	2.16E-03	1.92E-03		yes
ENSG00000224827	-1.1	-1.05	2.16E-03	3.84E-03		yes
ENSG00000161381	0.77	0.76	2.16E-03	7.54E-03		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000133962	0.93	1.02	2.18E-03	1.03E-03	5	yes
ENSG00000165929	0.93	1.02	2.18E-03	1.03E-03	5	yes
ENSG00000205268	0.68	0.6	2.18E-03	1.65E-02		yes
ENSG00000110696	0.66	0.55	2.19E-03	8.49E-03		yes
ENSG00000252197	1.09	1	2.22E-03	8.49E-03		yes
ENSG00000112655	1.04	1.2	2.23E-03	1.27E-03		yes
ENSG00000200397	0.7	0.6	2.23E-03	1.45E-02		yes
ENSG00000251783	0.87	0.8	2.23E-03	1.58E-02		yes
ENSG00000257802	0.8	0.67	2.25E-03	1.58E-02		yes
ENSG00000223551	-1.05	-1.09	2.31E-03	1.30E-03		yes
ENSG00000213639	0.73	0.63	2.31E-03	1.24E-02	6	no
ENSG00000163806	0.73	0.63	2.31E-03	1.24E-02	6	no
ENSG00000199866	1.31	1.29	2.31E-03	1.69E-02		yes
ENSG00000123728	0.69	0.61	2.34E-03	1.20E-02		no
ENSG00000253092	1.6	1.53	2.43E-03	4.56E-02		yes
ENSG00000226937	0.81	0.96	2.44E-03	4.81E-04		yes
ENSG00000135185	0.67	0.57	2.44E-03	9.44E-03		yes
ENSG00000173598	0.6	0.44	2.44E-03	4.56E-02		no
ENSG00000116489	0.72	0.65	2.45E-03	1.00E-02		yes
ENSG00000100575	0.51	0.5	2.45E-03	1.02E-02		yes
ENSG00000207513	1.15	1.08	2.45E-03	1.42E-02		yes
ENSG00000259657	0.75	0.66	2.45E-03	1.58E-02		yes
ENSG00000117335	0.72	0.54	2.45E-03	2.95E-02		no
ENSG00000258486	-1.18	-1.28	2.47E-03	9.36E-04		yes
ENSG00000111269	0.7	0.64	2.49E-03	7.19E-03		yes
ENSG00000226084	1.06	0.97	2.49E-03	1.11E-02		yes
ENSG00000253626	-1.44	-1.48	2.53E-03	2.12E-03		yes
ENSG00000238450	0.69	0.6	2.55E-03	1.75E-02		yes
ENSG00000254208	1.23	1.21	2.64E-03	1.63E-02		yes
ENSG00000240490	-0.74	-0.83	2.75E-03	9.21E-04		yes
ENSG00000201784	0.77	0.68	2.84E-03	8.49E-03		yes
ENSG00000238975	0.74	0.64	2.84E-03	8.79E-03		yes
ENSG00000197329	0.71	0.74	2.85E-03	7.30E-03		yes
ENSG00000241438	0.99	0.91	2.85E-03	1.27E-02		yes
ENSG00000205581	0.55	0.49	2.88E-03	1.19E-02	7	yes
ENSG00000238556	0.55	0.49	2.88E-03	1.19E-02	7	yes
ENSG00000259950	1.14	1.04	2.88E-03	1.25E-02		yes
ENSG00000248785	0.56	0.47	2.88E-03	1.69E-02		yes
ENSG00000223505	0.71	0.61	2.88E-03	1.91E-02		yes
ENSG00000227417	0.8	0.65	2.88E-03	2.38E-02		yes
ENSG00000226976	-1.04	-1.04	2.96E-03	3.84E-03		yes
ENSG00000159714	-0.73	-0.76	2.97E-03	5.92E-03		yes
ENSG00000162594	1.22	1.16	2.98E-03	2.26E-02		yes
ENSG00000223697	1.17	1.09	3.03E-03	1.35E-02		yes
ENSG00000123219	1.81	2.01	3.16E-03	2.13E-03		yes
ENSG00000251892	0.99	0.92	3.17E-03	1.65E-02		yes
ENSG00000259595	0.78	0.76	3.18E-03	8.49E-03		no
ENSG00000225627	0.76	0.67	3.18E-03	1.78E-02		yes
ENSG00000127920	0.7	0.62	3.18E-03	1.88E-02		yes
ENSG00000228366	1.06	0.96	3.18E-03	2.67E-02		yes
ENSG00000164305	0.86	0.95	3.36E-03	2.00E-03		yes
ENSG00000113369	0.8	0.75	3.40E-03	8.49E-03		yes
ENSG00000173338	-0.83	-0.74	3.50E-03	1.55E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000198756	-0.99	-1.07	3.58E-03	3.00E-03		yes
ENSG00000170571	0.68	0.7	3.58E-03	3.73E-03		no
ENSG00000133019	1.21	1.23	3.58E-03	4.93E-03		yes
ENSG00000145779	0.73	0.7	3.58E-03	1.24E-02		yes
ENSG00000223003	1.03	0.95	3.58E-03	2.75E-02		yes
ENSG00000113328	0.61	0.65	3.59E-03	5.07E-03		yes
ENSG00000242732	-1.01	-0.8	3.59E-03	1.77E-02		yes
ENSG00000023445	0.66	0.6	3.87E-03	2.40E-02		yes
ENSG00000111897	0.71	0.7	4.05E-03	9.07E-03		yes
ENSG00000206596	1.28	1.23	4.05E-03	1.57E-02		yes
ENSG00000228187	1.02	0.94	4.05E-03	1.97E-02		yes
ENSG00000105176	0.55	0.46	4.05E-03	2.62E-02		yes
ENSG00000126860	1.03	0.93	4.19E-03	9.48E-03		yes
ENSG00000222086	0.72	0.63	4.66E-03	2.06E-02		yes
ENSG00000197841	1.42	1.65	4.73E-03	4.93E-03		yes
ENSG00000005249	0.94	1.05	5.17E-03	4.55E-03		yes
ENSG00000133740	0.76	0.67	5.17E-03	2.33E-02		yes
ENSG00000147894	0.81	0.86	5.30E-03	1.08E-02		yes
ENSG00000047634	1.37	1.6	5.31E-03	6.78E-03		yes
ENSG00000203386	0.84	0.74	5.50E-03	2.19E-02		yes
ENSG00000183691	0.84	1.13	5.59E-03	4.60E-04		yes
ENSG00000243302	-0.55	-0.64	5.59E-03	1.27E-03		yes
ENSG00000252464	0.92	0.96	5.59E-03	6.75E-03	8	yes
ENSG00000113240	0.92	0.96	5.59E-03	6.75E-03	8	yes
ENSG00000112237	0.77	0.79	5.59E-03	7.51E-03		yes
ENSG00000205147	-0.98	-0.69	5.59E-03	1.19E-02		yes
ENSG00000183160	-1.17	-1.15	5.59E-03	1.55E-02		yes
ENSG00000212259	1.11	1.05	5.59E-03	1.88E-02		yes
ENSG00000231845	0.8	0.71	5.59E-03	1.93E-02		yes
ENSG00000182853	-1.4	-1.61	5.61E-03	3.74E-03		yes
ENSG00000249055	0.66	0.62	5.61E-03	1.64E-02	9	no
ENSG00000151247	0.66	0.62	5.61E-03	1.64E-02	9	no
ENSG00000238449	0.66	0.62	5.61E-03	1.64E-02	9	no
ENSG00000170215	-1.5	-1.42	5.61E-03	1.66E-02		yes
ENSG00000252750	0.7	0.6	5.61E-03	1.67E-02		yes
ENSG00000250473	1.52	1.47	5.61E-03	2.07E-02		yes
ENSG00000232486	0.91	0.83	5.61E-03	3.10E-02		yes
ENSG00000256039	0.84	1.04	5.80E-03	7.09E-04		yes
ENSG00000137492	0.78	0.92	5.80E-03	2.00E-03		yes
ENSG00000228956	0.9	1.12	5.88E-03	1.65E-03		yes
ENSG00000238959	0.89	0.81	5.97E-03	2.41E-02		yes
ENSG00000214548	-1.77	-1.81	6.27E-03	2.59E-02		yes
ENSG00000240342	1.04	1.04	6.30E-03	1.24E-02		yes
ENSG00000235299	0.69	0.6	6.30E-03	1.70E-02		yes
ENSG00000134765	1.75	1.82	6.35E-03	2.28E-02		yes
ENSG00000189266	0.6	0.62	6.44E-03	7.09E-03		yes
ENSG00000125245	0.48	0.41	6.47E-03	3.49E-02		yes
ENSG00000152219	0.71	0.62	6.60E-03	2.15E-02		yes
ENSG00000158555	-0.71	-0.77	6.64E-03	8.49E-03		yes
ENSG00000166770	0.9	0.87	6.64E-03	1.93E-02		yes
ENSG00000207567	-0.8	-0.89	6.67E-03	5.85E-03		yes
ENSG00000102760	0.54	0.54	6.67E-03	9.48E-03		yes
ENSG00000147162	0.75	0.68	6.67E-03	1.69E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000198791	0.52	0.43	6.67E-03	3.33E-02		yes
ENSG00000126261	0.51	0.51	6.92E-03	9.30E-03		yes
ENSG00000258645	1.79	1.76	7.05E-03	1.95E-02		yes
ENSG00000201013	0.76	0.66	7.05E-03	4.78E-02		no
ENSG00000253754	0.82	0.72	7.16E-03	1.49E-02		yes
ENSG00000138795	0.65	0.59	7.16E-03	1.74E-02		yes
ENSG00000127314	0.59	0.48	7.22E-03	1.79E-02		yes
ENSG00000115524	0.62	0.54	7.26E-03	2.54E-02		yes
ENSG00000119616	0.61	0.56	7.29E-03	1.99E-02		yes
ENSG00000202490	1.78	1.75	7.29E-03	2.03E-02		yes
ENSG00000206822	1.78	1.75	7.29E-03	2.03E-02		yes
ENSG00000221214	0.69	0.61	7.29E-03	3.39E-02		no
ENSG00000231128	0.66	0.53	7.29E-03	4.87E-02		no
ENSG00000116918	0.81	0.91	7.34E-03	2.12E-03		yes
ENSG00000260128	-3.85	-4.25	7.34E-03	8.22E-03		no
ENSG00000234219	0.72	0.61	7.34E-03	3.05E-02		yes
ENSG00000254671	0.68	0.59	7.34E-03	4.19E-02		no
ENSG00000104408	0.81	0.76	7.38E-03	1.30E-02		yes
ENSG00000151239	0.92	1.01	7.43E-03	7.28E-03		yes
ENSG00000197045	1.1	1.23	7.46E-03	7.09E-03		yes
ENSG00000215973	1.04	0.94	7.58E-03	3.46E-02		yes
ENSG00000239917	0.6	0.51	7.58E-03	3.90E-02		no
ENSG00000260977	-0.87	-0.61	7.61E-03	1.84E-02		no
ENSG00000177889	0.52	0.41	7.76E-03	4.58E-02		yes
ENSG00000233984	1.41	1.25	7.77E-03	3.03E-02		yes
ENSG00000140299	0.52	0.45	7.79E-03	3.50E-02		no
ENSG00000166317	-0.79	-0.82	8.06E-03	6.47E-03		yes
ENSG00000180776	0.92	0.88	8.14E-03	1.84E-02		yes
ENSG00000101972	0.67	0.83	8.18E-03	2.12E-03		yes
ENSG00000228323	-0.67	-0.76	8.18E-03	4.99E-03		yes
ENSG00000202054	-0.99	-1.09	8.18E-03	1.35E-02		yes
ENSG00000136536	0.56	0.5	8.23E-03	2.71E-02		no
ENSG00000124767	0.5	0.43	8.23E-03	3.26E-02		no
ENSG00000196498	-0.59	-0.63	8.40E-03	1.08E-02		yes
ENSG00000258663	-1.86	-2.14	8.40E-03	1.49E-02		yes
ENSG00000122042	0.63	0.61	8.52E-03	1.63E-02		no
ENSG00000173597	0.69	0.85	8.54E-03	2.12E-03		yes
ENSG00000136111	0.89	0.98	8.54E-03	6.16E-03		yes
ENSG00000170356	0.85	0.87	8.54E-03	1.84E-02	10	no
ENSG00000244479	0.85	0.87	8.54E-03	1.84E-02	10	no
ENSG00000248839	0.81	0.73	8.54E-03	2.37E-02		yes
ENSG00000241300	1.39	1.32	8.54E-03	2.51E-02		yes
ENSG00000154814	0.65	0.63	8.74E-03	1.33E-02		no
ENSG00000123685	-0.85	-1.06	8.84E-03	4.01E-03		yes
ENSG00000169131	0.71	0.74	9.14E-03	1.58E-02		yes
ENSG00000221420	0.58	0.58	9.15E-03	7.45E-03	11	yes
ENSG00000200418	0.58	0.58	9.15E-03	7.45E-03	11	yes
ENSG00000238942	0.58	0.58	9.15E-03	7.45E-03	11	yes
ENSG00000156976	0.58	0.58	9.15E-03	7.45E-03	11	yes
ENSG00000200320	0.58	0.58	9.15E-03	7.45E-03	11	yes
ENSG00000171928	0.94	0.98	9.15E-03	1.25E-02		yes
ENSG00000198796	-1.61	-1.66	9.15E-03	1.64E-02		yes
ENSG00000173542	0.7	0.53	9.15E-03	4.58E-02	12	yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000156136	0.7	0.53	9.15E-03	4.58E-02	12	yes
ENSG00000241376	-1.03	-1.1	9.17E-03	1.67E-02		yes
ENSG00000132485	0.77	0.84	9.38E-03	9.07E-03		yes
ENSG00000138138	0.8	0.99	9.47E-03	4.33E-03		yes
ENSG00000243305	0.92	0.88	9.51E-03	1.93E-02		yes
ENSG00000207034	0.63	0.53	9.51E-03	3.13E-02		yes
ENSG00000262211	1.14	1.19	9.53E-03	8.22E-03		yes
ENSG00000170855	0.73	0.81	9.53E-03	1.01E-02		yes
ENSG00000248265	-1.09	-1.3	9.66E-03	6.74E-03		no
ENSG00000197462	-1.02	-1.11	9.66E-03	1.25E-02		yes
ENSG00000233728	-0.87	-0.79	9.66E-03	1.98E-02		no
ENSG00000254274	0.83	0.74	9.78E-03	3.23E-02		yes
ENSG00000134758	0.67	0.74	9.93E-03	7.91E-03		yes
ENSG00000124193	0.48	0.45	9.93E-03	2.00E-02		yes
ENSG00000186063	0.55	0.64	9.99E-03	3.86E-03		yes
ENSG00000253636	-1.04	-0.72	9.99E-03	4.24E-02		no
ENSG00000221869	0.61	0.76	1.00E-02	2.77E-03		no
ENSG00000214659	-1.61	-1.78	1.00E-02	7.45E-03		yes
ENSG00000106537	0.82	0.74	1.03E-02	3.12E-02		yes
ENSG00000228554	0.58	0.5	1.04E-02	4.38E-02		no
ENSG00000070367	0.77	0.73	1.05E-02	4.87E-02		yes
ENSG00000230625	0.77	0.68	1.06E-02	3.46E-02		no
ENSG00000163412	0.55	0.46	1.06E-02	4.24E-02		no
ENSG00000202237	0.61	0.53	1.07E-02	4.47E-02		no
ENSG00000174579	0.6	0.7	1.08E-02	8.58E-03		yes
ENSG00000227146	0.59	0.55	1.08E-02	1.49E-02		yes
ENSG00000121749	0.71	0.62	1.08E-02	2.51E-02		yes
ENSG00000164830	0.81	0.73	1.08E-02	3.10E-02	13	yes
ENSG00000254615	0.81	0.73	1.08E-02	3.10E-02	13	yes
ENSG00000244275	0.54	0.45	1.08E-02	3.17E-02	14	yes
ENSG00000240369	0.54	0.45	1.08E-02	3.17E-02	14	yes
ENSG00000238916	0.54	0.45	1.08E-02	3.17E-02	14	yes
ENSG00000154330	-1.46	-1.73	1.09E-02	1.93E-02		yes
ENSG00000242681	0.99	0.9	1.09E-02	3.50E-02		yes
ENSG00000237672	0.79	0.7	1.09E-02	3.67E-02		no
ENSG00000144895	0.58	0.6	1.10E-02	1.27E-02		yes
ENSG00000157800	0.66	0.69	1.10E-02	1.64E-02		yes
ENSG00000148154	0.77	0.87	1.11E-02	9.20E-03		yes
ENSG00000200120	1.34	1.25	1.13E-02	2.38E-02		yes
ENSG00000225037	0.81	0.72	1.13E-02	3.65E-02		yes
ENSG00000217527	0.63	0.53	1.15E-02	4.72E-02		no
ENSG00000138032	0.64	0.6	1.16E-02	2.71E-02		yes
ENSG00000164985	0.53	0.47	1.16E-02	3.18E-02		yes
ENSG00000206737	0.94	0.86	1.16E-02	3.94E-02		yes
ENSG00000222383	0.59	0.5	1.16E-02	4.66E-02		no
ENSG00000174123	0.92	0.98	1.17E-02	7.69E-03		yes
ENSG00000185634	1.2	1.12	1.17E-02	4.58E-02		no
ENSG00000115866	0.5	0.47	1.18E-02	2.03E-02		yes
ENSG00000229750	0.97	0.96	1.18E-02	3.00E-02		no
ENSG00000227239	0.62	0.52	1.18E-02	3.06E-02		yes
ENSG00000238268	-1.23	-1.12	1.18E-02	3.77E-02		yes
ENSG00000187109	0.54	0.46	1.18E-02	3.94E-02		no
ENSG00000074201	0.53	0.48	1.18E-02	4.19E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000164109	0.8	1.04	1.19E-02	2.00E-03		yes
ENSG00000198898	0.69	0.7	1.22E-02	1.09E-02		no
ENSG00000262777	0.81	0.72	1.22E-02	4.79E-02		no
ENSG00000233355	0.85	0.99	1.23E-02	5.31E-03		yes
ENSG00000110330	0.71	0.7	1.23E-02	1.79E-02		yes
ENSG00000142892	0.76	0.74	1.23E-02	2.03E-02		yes
ENSG00000120992	0.62	0.56	1.25E-02	2.96E-02		no
ENSG00000124333	0.61	0.57	1.25E-02	2.99E-02		yes
ENSG00000227161	0.7	0.61	1.25E-02	3.84E-02		no
ENSG00000201778	0.57	0.52	1.25E-02	3.95E-02		no
ENSG00000221065	1.03	0.93	1.26E-02	2.97E-02		yes
ENSG00000149656	0.77	0.66	1.26E-02	3.84E-02	15	yes
ENSG00000203880	0.77	0.66	1.26E-02	3.84E-02	15	yes
ENSG00000106560	0.71	0.56	1.26E-02	4.06E-02		yes
ENSG00000139372	0.68	0.75	1.27E-02	1.82E-02		yes
ENSG00000222276	0.94	0.84	1.27E-02	3.87E-02		no
ENSG00000135776	0.69	0.78	1.28E-02	8.49E-03		yes
ENSG00000036054	0.65	0.72	1.28E-02	1.21E-02		yes
ENSG00000116754	0.61	0.64	1.28E-02	1.73E-02		yes
ENSG00000160654	0.7	0.69	1.28E-02	1.88E-02		no
ENSG00000250850	0.62	0.67	1.28E-02	2.32E-02		yes
ENSG00000071994	0.44	0.4	1.28E-02	4.58E-02		yes
ENSG00000238197	0.78	0.92	1.29E-02	7.51E-03		yes
ENSG00000180964	0.58	0.63	1.29E-02	8.22E-03		no
ENSG00000244389	1.16	1.1	1.29E-02	3.95E-02		yes
ENSG00000120742	0.61	0.55	1.29E-02	4.67E-02		no
ENSG00000168913	-0.75	-0.7	1.29E-02	4.84E-02		no
ENSG00000167766	0.87	0.86	1.30E-02	1.87E-02		yes
ENSG00000164209	0.72	0.74	1.30E-02	2.09E-02		yes
ENSG00000232022	-0.82	-0.92	1.30E-02	2.28E-02		yes
ENSG00000200814	0.66	0.56	1.30E-02	4.82E-02		no
ENSG00000185947	0.84	1	1.31E-02	3.88E-03	16	yes
ENSG00000262657	0.84	1	1.31E-02	3.88E-03	16	yes
ENSG00000225792	0.6	0.59	1.31E-02	3.33E-02		no
ENSG00000043093	0.55	0.6	1.32E-02	9.54E-03		no
ENSG00000207491	-0.82	-0.91	1.32E-02	1.64E-02		yes
ENSG00000259781	2.48	2.54	1.32E-02	3.00E-02		yes
ENSG00000091409	0.7	0.66	1.32E-02	3.06E-02		yes
ENSG00000130119	-0.77	-0.87	1.33E-02	1.07E-02		yes
ENSG00000134352	1.05	1.19	1.33E-02	1.29E-02		yes
ENSG00000204882	-0.95	-0.72	1.35E-02	3.77E-02		no
ENSG00000178863	-0.73	-0.69	1.36E-02	2.27E-02		yes
ENSG00000135318	0.76	1.16	1.37E-02	9.28E-05		yes
ENSG00000240163	-0.74	-0.83	1.37E-02	6.35E-03		yes
ENSG00000091129	1.5	1.56	1.37E-02	1.08E-02		yes
ENSG00000155903	0.84	0.85	1.37E-02	1.33E-02		yes
ENSG00000223573	-1.02	-0.78	1.37E-02	1.56E-02		yes
ENSG00000252759	1.16	1.06	1.37E-02	2.94E-02		yes
ENSG00000177144	1.08	1.26	1.39E-02	2.16E-02		yes
ENSG00000067064	0.71	0.73	1.40E-02	1.66E-02		yes
ENSG00000101856	0.49	0.51	1.45E-02	1.42E-02		yes
ENSG00000008323	-1.12	-1.52	1.46E-02	4.74E-04		yes
ENSG00000123358	-0.94	-0.92	1.46E-02	3.39E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000115419	0.63	0.59	1.47E-02	3.10E-02		no
ENSG00000177683	0.67	0.59	1.48E-02	4.42E-02	17	yes
ENSG00000135241	0.67	0.59	1.48E-02	4.42E-02	17	yes
ENSG00000211591	0.57	0.84	1.49E-02	7.31E-04		yes
ENSG00000136535	1.35	1.68	1.49E-02	1.35E-02	18	yes
ENSG00000251621	1.35	1.68	1.49E-02	1.35E-02	18	yes
ENSG00000144290	1.35	1.68	1.49E-02	1.35E-02	18	yes
ENSG00000206713	-0.62	-0.71	1.49E-02	1.66E-02		yes
ENSG00000135723	-0.47	-0.48	1.49E-02	1.82E-02		yes
ENSG00000073849	0.54	0.52	1.49E-02	2.82E-02		no
ENSG00000198856	0.54	0.48	1.49E-02	3.85E-02		yes
ENSG00000207721	0.8	0.7	1.49E-02	4.02E-02		no
ENSG00000091972	0.8	1	1.50E-02	5.31E-03		yes
ENSG00000173372	-1.03	-1.13	1.50E-02	1.33E-02		yes
ENSG00000138078	0.68	0.65	1.50E-02	3.95E-02		yes
ENSG00000175895	0.64	0.77	1.51E-02	2.75E-03		yes
ENSG00000235117	-0.51	-0.49	1.51E-02	2.35E-02		yes
ENSG00000198586	0.6	0.54	1.51E-02	4.24E-02		yes
ENSG00000168497	0.64	0.63	1.52E-02	9.20E-03		yes
ENSG00000172986	0.69	0.74	1.52E-02	1.26E-02	19	yes
ENSG00000163605	0.69	0.74	1.52E-02	1.26E-02	19	yes
ENSG00000178913	0.61	0.57	1.52E-02	3.51E-02	20	yes
ENSG00000255729	0.61	0.57	1.52E-02	3.51E-02	20	yes
ENSG00000162607	0.66	0.66	1.53E-02	2.13E-02		yes
ENSG00000134548	1.09	1	1.53E-02	2.51E-02		yes
ENSG00000089335	1.04	0.97	1.53E-02	2.85E-02		yes
ENSG00000129757	-0.88	-0.82	1.53E-02	3.07E-02		yes
ENSG00000122565	0.6	0.55	1.53E-02	3.48E-02		yes
ENSG00000260808	0.61	0.75	1.54E-02	1.57E-03		no
ENSG00000163600	0.81	0.94	1.54E-02	6.47E-03		yes
ENSG00000149782	-0.64	-0.71	1.54E-02	1.20E-02		yes
ENSG00000138764	0.64	0.69	1.54E-02	1.58E-02		yes
ENSG00000254193	3.59	3.6	1.54E-02	2.33E-02		yes
ENSG00000231995	-4.01	-4.56	1.54E-02	4.02E-02		yes
ENSG00000250267	1.26	1.2	1.54E-02	4.29E-02		yes
ENSG00000134954	0.45	0.42	1.57E-02	3.74E-02		yes
ENSG00000165359	0.65	0.62	1.59E-02	2.55E-02		yes
ENSG00000139921	0.72	0.81	1.61E-02	8.49E-03		no
ENSG00000174106	0.76	0.85	1.61E-02	9.28E-03		yes
ENSG00000177990	1.78	1.89	1.61E-02	1.35E-02		yes
ENSG00000172469	1.18	1.22	1.61E-02	2.01E-02		no
ENSG00000180389	-0.92	-0.89	1.61E-02	2.94E-02		yes
ENSG00000164331	0.66	0.6	1.61E-02	3.89E-02		yes
ENSG00000248977	1.72	1.73	1.61E-02	4.84E-02		yes
ENSG00000213147	-5.49	-6.33	1.62E-02	1.71E-02		yes
ENSG00000253102	-0.53	-0.6	1.64E-02	1.29E-02		yes
ENSG00000108960	0.69	0.69	1.64E-02	2.31E-02		yes
ENSG00000261573	0.56	0.47	1.64E-02	4.67E-02		no
ENSG00000099917	-0.45	-0.53	1.68E-02	1.26E-02		yes
ENSG00000233392	-0.85	-1	1.68E-02	1.45E-02		yes
ENSG00000236287	0.69	0.7	1.72E-02	1.64E-02		yes
ENSG00000085365	0.66	0.77	1.72E-02	1.74E-02		no
ENSG00000253772	-1.71	-2.03	1.72E-02	2.97E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000145623	0.9	0.83	1.73E-02	2.40E-02		yes
ENSG00000199415	1.36	1.27	1.73E-02	3.67E-02		yes
ENSG00000177839	-0.72	-0.74	1.74E-02	2.36E-02		no
ENSG00000108061	0.55	0.69	1.77E-02	5.06E-03		no
ENSG00000132024	-0.54	-0.58	1.77E-02	1.79E-02		yes
ENSG00000163322	0.72	0.76	1.77E-02	2.07E-02		yes
ENSG00000105609	-0.57	-0.65	1.80E-02	1.64E-02		yes
ENSG00000169251	0.76	0.9	1.81E-02	9.07E-03		yes
ENSG00000213881	0.94	0.95	1.81E-02	4.22E-02		no
ENSG00000081320	0.52	0.56	1.84E-02	1.33E-02		no
ENSG00000258800	0.75	1.05	1.85E-02	1.45E-03		yes
ENSG00000132514	-0.61	-0.66	1.85E-02	1.67E-02		yes
ENSG00000171469	0.6	0.57	1.88E-02	4.13E-02		no
ENSG00000260978	3.78	4.38	1.89E-02	8.50E-03		no
ENSG00000149591	-0.58	-0.69	1.89E-02	1.11E-02		yes
ENSG00000164346	0.65	0.67	1.89E-02	1.33E-02		yes
ENSG00000109113	-0.53	-0.61	1.89E-02	1.49E-02		yes
ENSG00000262429	-0.46	-0.51	1.89E-02	1.74E-02		yes
ENSG00000128699	0.56	0.52	1.89E-02	3.77E-02		yes
ENSG00000140030	0.6	0.61	1.89E-02	4.07E-02		yes
ENSG00000125844	-0.53	-0.51	1.91E-02	3.98E-02		yes
ENSG00000113263	0.51	0.48	1.92E-02	4.13E-02		yes
ENSG00000100342	-0.64	-0.65	1.93E-02	3.03E-02		no
ENSG00000160326	-0.54	-0.58	1.95E-02	2.37E-02		yes
ENSG00000125772	0.66	0.66	1.95E-02	2.72E-02		no
ENSG00000166165	-0.77	-0.94	2.00E-02	1.00E-02		yes
ENSG00000230567	-5.58	-6.2	2.01E-02	7.96E-03		yes
ENSG00000066056	-0.8	-0.9	2.01E-02	1.73E-02		no
ENSG00000176390	0.45	0.45	2.01E-02	3.13E-02		no
ENSG00000184613	0.53	0.54	2.01E-02	3.67E-02		yes
ENSG00000155100	0.93	1.13	2.05E-02	1.73E-02		yes
ENSG00000259421	-0.61	-0.65	2.05E-02	1.84E-02		no
ENSG00000163534	0.79	0.93	2.10E-02	5.54E-03		yes
ENSG00000152270	0.63	0.62	2.11E-02	3.39E-02		yes
ENSG00000228265	-0.49	-0.55	2.12E-02	1.27E-02		yes
ENSG00000162711	-0.61	-0.62	2.12E-02	2.54E-02		yes
ENSG00000135002	0.67	0.63	2.12E-02	4.09E-02		no
ENSG00000182141	1.03	1.25	2.13E-02	9.28E-03		yes
ENSG00000102158	0.56	0.53	2.17E-02	2.47E-02		yes
ENSG00000255671	-0.92	-0.99	2.17E-02	4.52E-02		yes
ENSG00000236552	0.76	0.84	2.18E-02	1.51E-02		yes
ENSG00000087338	0.59	0.68	2.18E-02	1.58E-02		no
ENSG00000156564	-1.43	-1.26	2.18E-02	4.06E-02		no
ENSG00000227484	-0.95	-0.94	2.20E-02	3.92E-02		no
ENSG00000146757	1.13	1.32	2.22E-02	1.23E-02		yes
ENSG00000172845	0.67	0.79	2.25E-02	1.58E-02		yes
ENSG00000164938	0.95	1.18	2.26E-02	4.99E-03		yes
ENSG00000253574	-0.69	-0.78	2.26E-02	1.88E-02		yes
ENSG00000163660	0.51	0.52	2.26E-02	4.74E-02		no
ENSG00000152256	0.62	0.77	2.28E-02	5.92E-03		yes
ENSG00000251806	-0.53	-0.62	2.28E-02	6.78E-03		yes
ENSG00000239935	-0.7	-0.8	2.28E-02	7.32E-03		yes
ENSG00000154153	0.77	0.83	2.28E-02	2.44E-02		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000214194	0.96	0.96	2.28E-02	2.62E-02		no
ENSG00000233493	0.57	0.78	2.30E-02	2.05E-03		yes
ENSG00000159713	-0.82	-0.89	2.30E-02	2.53E-02	21	yes
ENSG00000239194	-0.82	-0.89	2.30E-02	2.53E-02	21	yes
ENSG00000215458	-0.61	-0.61	2.31E-02	3.78E-02		yes
ENSG00000176720	-1.54	-1.44	2.31E-02	4.96E-02		no
ENSG00000114062	0.64	0.71	2.32E-02	1.29E-02		yes
ENSG00000124786	0.55	0.57	2.40E-02	3.05E-02		yes
ENSG00000135655	0.6	0.58	2.40E-02	3.17E-02		no
ENSG00000165195	0.96	1.16	2.41E-02	1.00E-02		yes
ENSG00000126254	-0.53	-0.55	2.41E-02	1.88E-02		yes
ENSG00000013441	0.83	0.85	2.42E-02	1.49E-02		yes
ENSG00000152495	0.59	0.57	2.44E-02	3.90E-02		yes
ENSG00000200788	-0.59	-0.68	2.45E-02	1.67E-02		yes
ENSG00000179820	-0.55	-0.69	2.47E-02	8.49E-03		yes
ENSG00000066422	0.61	0.69	2.47E-02	1.28E-02		yes
ENSG00000188725	0.56	0.63	2.47E-02	1.65E-02		no
ENSG00000184730	-0.62	-0.66	2.47E-02	2.09E-02		yes
ENSG00000261744	-0.78	-0.86	2.47E-02	2.47E-02		yes
ENSG00000130479	-0.57	-0.62	2.47E-02	2.75E-02		yes
ENSG00000220842	4.8	4.85	2.47E-02	2.85E-02		yes
ENSG00000133641	0.81	0.89	2.48E-02	2.11E-02		yes
ENSGR0000124333	0.58	0.59	2.51E-02	2.79E-02		yes
ENSG00000125637	-0.45	-0.54	2.54E-02	9.07E-03		yes
ENSG00000250733	-0.69	-0.77	2.54E-02	2.38E-02		no
ENSG00000185761	-0.93	-1.19	2.54E-02	2.46E-02		yes
ENSG00000095002	0.57	0.6	2.54E-02	2.94E-02		yes
ENSG00000256403	-1.08	-1.19	2.54E-02	3.39E-02		yes
ENSG00000100100	0.46	0.42	2.55E-02	4.84E-02		yes
ENSG00000100028	-0.46	-0.5	2.58E-02	1.88E-02	22	no
ENSG00000100031	-0.46	-0.5	2.58E-02	1.88E-02	22	no
ENSG00000130766	-0.5	-0.6	2.61E-02	7.91E-03		yes
ENSG00000261327	-0.76	-0.62	2.61E-02	3.99E-02	23	no
ENSG00000259811	-0.76	-0.62	2.61E-02	3.99E-02	23	no
ENSG00000237336	-0.62	-0.71	2.62E-02	1.95E-02		no
ENSG00000231245	1.27	1.7	2.64E-02	1.58E-02		yes
ENSG00000255423	0.66	0.82	2.66E-02	2.02E-02		yes
ENSG00000104231	0.73	0.72	2.66E-02	3.05E-02		yes
ENSG00000142733	-0.73	-0.68	2.66E-02	4.83E-02		yes
ENSG00000021574	0.6	0.71	2.67E-02	1.58E-02		no
ENSG00000104936	-0.52	-0.58	2.67E-02	1.77E-02		yes
ENSG00000091039	0.77	0.82	2.67E-02	2.72E-02		yes
ENSG00000207523	0.71	0.61	2.67E-02	4.82E-02		yes
ENSG00000235872	0.69	1.03	2.68E-02	2.24E-03		yes
ENSG00000126458	-0.53	-0.6	2.68E-02	2.10E-02		yes
ENSG00000176076	0.57	0.58	2.68E-02	3.00E-02	24	no
ENSG00000068366	0.57	0.58	2.68E-02	3.00E-02	24	no
ENSG00000243736	0.57	0.58	2.68E-02	3.00E-02	24	no
ENSG00000105371	-0.65	-0.7	2.73E-02	1.49E-02		yes
ENSG00000085274	0.67	0.83	2.74E-02	1.20E-02		yes
ENSG00000077458	0.79	0.84	2.77E-02	3.33E-02		no
ENSG00000168300	0.59	0.58	2.77E-02	4.47E-02		yes
ENSG00000177764	0.56	0.83	2.78E-02	1.03E-03	25	no

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000247315	0.56	0.83	2.78E-02	1.03E-03	25	no
ENSG00000149516	0.69	0.74	2.78E-02	2.10E-02		no
ENSG00000106991	-0.62	-0.65	2.78E-02	3.39E-02		yes
ENSG00000224895	-0.81	-0.9	2.80E-02	3.01E-02		no
ENSG00000234106	-0.62	-0.72	2.81E-02	1.73E-02		yes
ENSG00000185246	0.72	0.86	2.81E-02	1.75E-02		yes
ENSG00000008283	-0.57	-0.55	2.84E-02	3.67E-02		yes
ENSG00000107796	-0.74	-0.69	2.85E-02	3.19E-02		no
ENSG00000214688	-0.49	-0.49	2.85E-02	3.38E-02		yes
ENSG00000099204	0.72	0.79	2.87E-02	1.64E-02		yes
ENSG00000168876	0.77	0.81	2.87E-02	1.71E-02		yes
ENSG00000165322	0.77	0.95	2.87E-02	1.79E-02		yes
ENSG00000143382	-0.56	-0.58	2.87E-02	4.29E-02	26	yes
ENSG00000225996	-0.56	-0.58	2.87E-02	4.29E-02	26	yes
ENSG00000090006	-0.59	-0.72	2.90E-02	1.33E-02		yes
ENSG00000103426	-0.39	-0.44	2.90E-02	2.19E-02	27	yes
ENSG00000262246	-0.39	-0.44	2.90E-02	2.19E-02	27	yes
ENSG00000217930	-0.39	-0.44	2.90E-02	2.19E-02	27	yes
ENSG00000110429	0.57	0.61	2.90E-02	3.36E-02		no
ENSG00000228386	-0.5	-0.59	2.92E-02	8.49E-03		yes
ENSG00000182809	-0.63	-0.71	2.92E-02	1.57E-02		yes
ENSG00000067167	0.45	0.47	2.92E-02	2.43E-02		no
ENSG00000248636	-0.55	-0.5	2.92E-02	4.75E-02		no
ENSG00000008311	-0.68	-0.76	2.94E-02	2.63E-02		no
ENSG00000126804	0.66	0.84	2.95E-02	2.92E-03		yes
ENSG00000089723	-0.85	-0.67	2.99E-02	1.84E-02		no
ENSG00000102409	0.5	0.53	2.99E-02	3.37E-02		yes
ENSG00000174780	0.44	0.43	2.99E-02	3.76E-02		yes
ENSG00000213204	0.66	0.66	2.99E-02	4.64E-02	28	no
ENSG00000164414	0.66	0.66	2.99E-02	4.64E-02	28	no
ENSG00000196576	-0.62	-0.72	3.03E-02	1.67E-02		yes
ENSG00000126461	-0.5	-0.55	3.03E-02	2.74E-02		yes
ENSG00000196924	-0.63	-0.69	3.05E-02	1.80E-02		yes
ENSG00000131931	0.53	0.57	3.06E-02	1.75E-02		no
ENSG00000112697	0.6	0.61	3.06E-02	3.07E-02		no
ENSG00000137076	-0.62	-0.63	3.07E-02	3.85E-02		yes
ENSG00000199591	-0.79	-0.88	3.09E-02	3.33E-02		no
ENSG00000228981	1.16	1.42	3.11E-02	7.19E-03		yes
ENSG00000110848	0.86	1.12	3.21E-02	8.22E-03		yes
ENSG00000211456	0.57	0.67	3.25E-02	1.49E-02		yes
ENSG00000100292	-0.58	-0.63	3.25E-02	2.67E-02		yes
ENSG00000129910	-0.94	-1.03	3.28E-02	4.07E-02		no
ENSG00000239975	1.66	1.71	3.28E-02	4.09E-02		yes
ENSG00000214787	-0.57	-0.84	3.29E-02	1.10E-03		yes
ENSG00000132718	-0.53	-0.6	3.29E-02	7.45E-03		yes
ENSG00000162852	0.65	0.77	3.29E-02	1.19E-02		yes
ENSG00000124406	0.87	1.12	3.29E-02	1.49E-02		yes
ENSG00000188229	-0.44	-0.5	3.29E-02	2.16E-02		no
ENSG00000181222	-0.49	-0.54	3.29E-02	2.85E-02		yes
ENSG00000164163	0.55	0.55	3.29E-02	3.28E-02		yes
ENSG00000186265	0.71	0.68	3.29E-02	3.62E-02		yes
ENSG00000005812	0.48	0.68	3.30E-02	2.37E-03		yes
ENSG00000259099	-0.63	-0.73	3.30E-02	2.17E-02		no

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000112851	0.64	0.84	3.31E-02	6.74E-03		yes
ENSG00000154174	0.44	0.51	3.31E-02	1.22E-02		yes
ENSG00000213744	-0.59	-0.7	3.31E-02	2.18E-02		no
ENSG00000196154	-0.58	-0.67	3.32E-02	9.64E-03		no
ENSG00000107317	-1.03	-1.08	3.32E-02	3.06E-02	29	yes
ENSG00000214402	-1.03	-1.08	3.32E-02	3.06E-02	29	yes
ENSG00000229413	-0.61	-0.66	3.32E-02	4.51E-02		yes
ENSG00000124145	-0.87	-0.78	3.33E-02	6.20E-03		yes
ENSG00000078589	0.52	0.6	3.33E-02	3.93E-02		yes
ENSG00000130402	-0.59	-0.65	3.34E-02	2.97E-02		yes
ENSG00000152133	0.73	0.91	3.36E-02	1.99E-02		yes
ENSG00000198087	0.82	0.97	3.36E-02	2.37E-02		yes
ENSG00000109736	-0.47	-0.54	3.37E-02	1.42E-02		yes
ENSG00000141429	0.49	0.55	3.37E-02	1.74E-02		yes
ENSG00000134255	0.49	0.55	3.37E-02	1.81E-02		yes
ENSG00000109943	0.65	0.75	3.38E-02	1.87E-02		yes
ENSG00000155545	1.22	1.55	3.38E-02	2.08E-02		yes
ENSG00000170190	-0.4	-0.46	3.38E-02	2.30E-02		yes
ENSG00000222869	-0.57	-0.67	3.40E-02	2.94E-02		no
ENSG00000160570	-0.42	-0.52	3.41E-02	1.67E-02		no
ENSG00000126821	0.73	1.12	3.44E-02	9.21E-04		yes
ENSG00000255302	0.47	0.45	3.44E-02	4.75E-02	30	no
ENSG00000235883	0.47	0.45	3.44E-02	4.75E-02	30	no
ENSG00000153989	0.5	0.55	3.46E-02	1.58E-02		yes
ENSG00000138468	0.78	0.89	3.46E-02	1.88E-02		yes
ENSG00000185736	-3.74	-4.03	3.46E-02	3.47E-02		yes
ENSG00000071575	0.53	0.53	3.46E-02	4.42E-02		yes
ENSG00000259884	-0.67	-0.69	3.46E-02	4.84E-02		yes
ENSG00000256515	-3.92	-4.05	3.48E-02	3.84E-02		no
ENSG00000166323	0.82	0.79	3.48E-02	4.10E-02		yes
ENSG00000206418	0.55	0.72	3.50E-02	8.83E-03		yes
ENSG00000182578	-0.67	-0.79	3.50E-02	1.33E-02		yes
ENSG00000059758	0.64	0.72	3.50E-02	2.32E-02		yes
ENSG00000101888	0.83	0.96	3.50E-02	3.02E-02		yes
ENSG00000207972	-0.47	-0.56	3.51E-02	2.10E-02		no
ENSG00000241247	0.57	0.55	3.51E-02	4.87E-02	31	yes
ENSG00000105829	0.57	0.55	3.51E-02	4.87E-02	31	yes
ENSG00000211642	0.97	1.08	3.52E-02	2.38E-02		yes
ENSG00000141096	-0.48	-0.55	3.52E-02	2.53E-02		yes
ENSG00000079950	0.56	0.59	3.52E-02	4.29E-02		yes
ENSG00000224032	0.48	0.51	3.53E-02	1.69E-02		no
ENSG00000119714	-0.5	-0.57	3.54E-02	1.29E-02		no
ENSG00000100599	-0.52	-0.65	3.54E-02	1.45E-02		no
ENSG00000201071	-0.43	-0.52	3.55E-02	2.26E-02		no
ENSG00000121753	-0.93	-0.94	3.57E-02	2.43E-02		yes
ENSG00000143554	-0.49	-0.55	3.57E-02	3.03E-02		yes
ENSG00000239932	-0.72	-0.79	3.57E-02	4.72E-02		no
ENSG00000251920	0.63	0.83	3.58E-02	1.88E-02		yes
ENSG00000218980	-0.59	-0.68	3.59E-02	2.59E-02		no
ENSG00000058063	0.54	0.58	3.60E-02	3.91E-02		yes
ENSG00000116095	0.5	0.55	3.63E-02	2.62E-02		no
ENSG00000188223	-0.43	-0.47	3.63E-02	3.36E-02		yes
ENSG00000170759	0.59	0.64	3.64E-02	2.43E-02		no

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000133246	-0.59	-0.66	3.66E-02	1.69E-02		yes
ENSG00000155307	0.49	0.55	3.66E-02	2.46E-02	32	yes
ENSG00000243440	0.49	0.55	3.66E-02	2.46E-02	32	yes
ENSG00000257702	-0.66	-0.59	3.66E-02	4.72E-02		no
ENSG00000232021	0.8	0.95	3.68E-02	2.97E-02		yes
ENSG00000136051	0.65	0.81	3.70E-02	1.33E-02		yes
ENSG00000111358	0.65	0.72	3.70E-02	3.67E-02		yes
ENSG00000140367	0.53	0.53	3.70E-02	3.88E-02		no
ENSG00000100284	-0.45	-0.47	3.70E-02	4.09E-02		yes
ENSG00000230715	-0.42	-0.53	3.75E-02	1.02E-02		yes
ENSG00000228770	-0.59	-0.67	3.75E-02	3.06E-02		no
ENSG00000111790	0.52	0.55	3.76E-02	3.60E-02		yes
ENSG00000147853	0.5	0.68	3.77E-02	4.97E-03		yes
ENSG00000125434	-0.54	-0.51	3.79E-02	2.94E-02		no
ENSG00000106351	-0.43	-0.45	3.79E-02	4.06E-02		yes
ENSG00000134717	0.51	0.51	3.79E-02	4.73E-02		yes
ENSG00000217416	1.06	1.28	3.80E-02	3.47E-02		yes
ENSG00000229679	-0.58	-0.67	3.83E-02	3.10E-02		no
ENSG00000106733	0.46	0.49	3.83E-02	4.62E-02		no
ENSG00000110697	-0.36	-0.43	3.84E-02	2.32E-02		yes
ENSG00000111676	-0.53	-0.6	3.85E-02	3.33E-02		yes
ENSG00000168071	-0.47	-0.64	3.89E-02	9.47E-03		yes
ENSG00000226752	0.67	0.76	3.93E-02	2.46E-02		yes
ENSG00000261691	-0.5	-0.56	3.93E-02	2.78E-02		yes
ENSG00000110881	1.14	1.39	3.95E-02	3.30E-02		yes
ENSG00000125629	0.56	0.7	3.98E-02	1.35E-02		yes
ENSG00000183918	0.74	0.85	3.98E-02	3.13E-02		yes
ENSG00000006025	-0.49	-0.5	3.98E-02	3.95E-02		no
ENSG00000152601	0.63	0.78	4.03E-02	1.65E-02		yes
ENSG00000204469	-0.44	-0.5	4.03E-02	2.50E-02		no
ENSG00000013275	-0.44	-0.49	4.03E-02	2.64E-02		yes
ENSG00000200113	-0.59	-0.69	4.03E-02	3.39E-02		no
ENSG00000162383	-1.04	-1.15	4.03E-02	4.96E-02		no
ENSG00000172766	0.68	0.75	4.05E-02	3.60E-02		yes
ENSG00000090339	-0.49	-0.53	4.08E-02	3.33E-02		no
ENSG00000123700	0.62	0.96	4.11E-02	1.27E-03		no
ENSG00000176406	-0.76	-0.84	4.16E-02	3.05E-02	33	no
ENSG00000239344	-0.76	-0.84	4.16E-02	3.05E-02	33	no
ENSG00000143368	-0.49	-0.54	4.16E-02	4.42E-02		no
ENSG00000255959	-0.47	-0.48	4.17E-02	3.10E-02		no
ENSG00000241360	-0.46	-0.5	4.19E-02	2.63E-02	34	yes
ENSG00000100092	-0.46	-0.5	4.19E-02	2.63E-02	34	yes
ENSG00000207733	-0.55	-0.63	4.20E-02	2.75E-02		no
ENSG00000238766	-0.53	-0.62	4.23E-02	2.99E-02		no
ENSG00000242926	-0.53	-0.62	4.23E-02	2.99E-02		no
ENSG00000205423	0.58	0.66	4.23E-02	4.02E-02		no
ENSG00000205609	-1.47	-1.64	4.23E-02	4.47E-02		no
ENSG00000155158	0.75	1	4.30E-02	1.11E-02		yes
ENSG00000136830	-0.56	-0.65	4.30E-02	3.26E-02		yes
ENSG00000065548	0.58	0.59	4.30E-02	3.68E-02		no
ENSG00000106993	0.65	0.77	4.32E-02	2.40E-02		yes
ENSG00000122545	0.38	0.37	4.35E-02	4.82E-02		no
ENSG00000205413	0.56	0.79	4.37E-02	6.74E-03		yes

Continued on next page

Table B.1 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000185275	0.54	0.6	4.44E-02	2.76E-02		yes
ENSG00000168476	-0.42	-0.46	4.49E-02	3.49E-02		yes
ENSG00000164509	-0.6	-0.68	4.49E-02	4.35E-02		no
ENSG00000172673	0.59	0.71	4.50E-02	2.43E-02		no
ENSG00000110455	-0.62	-0.72	4.50E-02	2.97E-02		no
ENSG00000155367	-0.56	-0.64	4.51E-02	3.39E-02	35	no
ENSG00000155366	-0.56	-0.64	4.51E-02	3.39E-02	35	no
ENSG00000134242	0.53	0.54	4.51E-02	4.66E-02		no
ENSG00000023191	-0.5	-0.56	4.54E-02	1.87E-02		yes
ENSG00000166145	-0.5	-0.55	4.55E-02	4.29E-02		no
ENSG00000066557	0.66	0.79	4.57E-02	2.42E-02		yes
ENSG00000025156	0.52	0.6	4.57E-02	2.75E-02		yes
ENSG00000165512	0.54	0.72	4.60E-02	9.05E-03		yes
ENSG00000234617	0.39	0.47	4.61E-02	1.69E-02		yes
ENSG00000148180	-0.6	-0.75	4.64E-02	1.75E-02	36	yes
ENSG00000244498	-0.6	-0.75	4.64E-02	1.75E-02	36	yes
ENSG00000166200	0.74	0.88	4.64E-02	2.51E-02		yes
ENSG00000255165	-0.75	-0.84	4.64E-02	3.95E-02		no
ENSG00000136286	-0.48	-0.53	4.65E-02	3.12E-02		yes
ENSG00000108064	0.46	0.51	4.66E-02	2.84E-02		yes
ENSG00000104973	-0.45	-0.58	4.67E-02	1.62E-02		yes
ENSG00000196381	1.22	1.57	4.70E-02	1.20E-02		yes
ENSG00000255301	-1.41	-1.92	4.72E-02	1.78E-02		yes
ENSG00000169855	1.21	2.15	4.73E-02	7.31E-04		yes
ENSG00000235865	-0.65	-0.73	4.73E-02	2.29E-02	37	yes
ENSG00000239593	-0.65	-0.73	4.73E-02	2.29E-02	37	yes
ENSG00000151292	0.71	0.85	4.76E-02	2.26E-02		yes
ENSG00000108798	-0.57	-0.63	4.80E-02	3.85E-02		yes
ENSG00000165169	0.66	0.77	4.82E-02	1.68E-02		yes
ENSG00000095370	-0.43	-0.53	4.86E-02	1.88E-02		yes
ENSG00000105953	-0.46	-0.54	4.88E-02	3.48E-02		yes
ENSG00000240925	-0.45	-0.54	4.90E-02	2.85E-02		no
ENSG00000162398	-1.04	-1.33	4.96E-02	2.38E-02		yes
ENSG00000182010	1.05	1.35	4.97E-02	3.28E-02		yes
ENSG00000128654	0.66	0.66	4.97E-02	3.84E-02		yes

SMA3 *vs.* CTRL

Table B.2: Differential expressed genes selected for the “SMA3 versus CTRL” comparison: Ensembl gene IDs, log-ratios and q-values, groups of *overlapping-genes*, labelled with the same group number, and genes confirmed by the *totcounts-genes* approach.

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000212769	-4.51	-4.65	1.37E-07	2.00E-07		yes
ENSG00000228224	-2.29	-2.38	1.37E-07	2.00E-07		yes
ENSG00000198618	-3.61	-3.6	1.81E-07	5.05E-06		yes
ENSG00000253676	2.67	2.47	1.47E-06	1.80E-04		yes
ENSG00000171116	-8.36	-6.52	5.13E-06	4.75E-04		yes
ENSG00000172062	-1.95	-2.12	1.14E-05	4.69E-05		yes
ENSG00000243742	-1.68	-1.62	1.56E-05	1.80E-04		no
ENSG00000252197	2.01	1.93	3.68E-05	2.28E-04		yes
ENSG00000251948	-1.57	-1.65	1.37E-04	2.10E-05		yes
ENSG00000198566	-8.19	-10.44	2.32E-04	9.97E-08		yes
ENSG00000258988	-4.46	-4.51	2.80E-04	2.40E-03	1	no
ENSG00000172717	-4.46	-4.51	2.80E-04	2.40E-03	1	no
ENSG00000230408	1.06	0.94	3.94E-04	1.05E-02		no
ENSG00000223551	-1.8	-1.8	1.13E-03	6.14E-04		yes
ENSG00000132204	-1.92	-1.93	1.58E-03	6.92E-03		no
ENSG00000201121	1.87	1.85	2.09E-03	2.23E-02		no
ENSG00000233251	1.12	1.07	2.65E-03	1.71E-02		no
ENSG00000149516	1.4	1.41	2.92E-03	5.67E-03		no
ENSG00000169397	1.66	1.65	2.92E-03	2.23E-02		no
ENSG00000260128	-7.78	-9.08	4.20E-03	4.75E-04		no
ENSG00000169131	1.31	1.32	4.20E-03	1.83E-02		no
ENSG00000224827	-1.83	-1.62	5.20E-03	2.23E-02		no
ENSG00000203386	1.3	1.23	6.02E-03	3.20E-02		no
ENSG00000231896	2.13	2.04	8.99E-03	3.45E-02		no
ENSG00000158525	2.53	2.8	9.48E-03	3.45E-02		yes
ENSG00000206650	1	0.92	1.44E-02	4.30E-02		no
ENSG00000168913	-1.37	-1.4	1.51E-02	4.40E-02		no
ENSG00000234040	0.99	0.91	1.61E-02	4.79E-02		no
ENSG00000226944	-1.44	-1.53	1.66E-02	1.71E-02		no
ENSG00000160307	1.83	1.96	1.66E-02	2.27E-02		no
ENSG00000174236	-7.24	-8.72	1.66E-02	2.76E-02		no
ENSG00000166317	-1.31	-1.35	1.67E-02	1.71E-02		no
ENSG00000253518	-1.72	-1.86	1.80E-02	2.97E-02		no
ENSG00000221869	0.96	1.21	1.99E-02	3.24E-03		no
ENSG00000185736	-8.1	-9.01	2.10E-02	2.23E-02		no
ENSG00000253626	-2.07	-2.01	2.53E-02	3.51E-02		no
ENSG00000256515	-8.7	-9.35	2.71E-02	2.88E-02		no
ENSG00000226976	-1.53	-1.64	3.03E-02	1.38E-02		no
ENSG00000197329	0.98	1.17	3.03E-02	2.23E-02		no
ENSG00000240490	-0.98	-1.06	3.92E-02	2.23E-02		no
ENSG00000123091	0.78	0.8	4.54E-02	2.40E-02		no

SMA2 *vs.* CTRL

Table B.3: Differential expressed genes selected for the "SMA2 versus CTRL" comparison: Ensembl gene IDs, log-ratios and q-values, groups of *overlapping-genes*, labelled with the same group number, and genes confirmed by the *totcounts-genes* approach.

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000172062	-1.75	-2.05	9.61E-09	3.36E-08		yes
ENSG00000240869	-1.93	-2.03	3.47E-04	3.81E-05		yes
ENSG00000254208	1.83	1.82	3.47E-04	8.04E-03		yes
ENSG00000232162	2.06	1.91	3.63E-04	3.72E-02		yes
ENSG00000242580	1.74	1.51	4.27E-04	1.42E-02		yes
ENSG00000198566	-7.27	-9.66	5.76E-04	2.45E-08		yes
ENSG00000211654	2.13	2.06	1.34E-03	6.34E-03		yes
ENSG00000228956	1.39	1.65	1.44E-03	5.77E-04		yes
ENSG00000168685	1.16	1.13	1.44E-03	6.34E-03		yes
ENSG00000175147	-1.72	-1.39	2.58E-03	5.77E-04		yes
ENSG00000258486	-1.84	-1.92	2.58E-03	7.19E-04		yes
ENSG00000162368	0.96	0.83	2.58E-03	3.01E-02		yes
ENSG00000222494	1.19	1.11	2.77E-03	3.01E-02		yes
ENSG00000180662	1.28	1.23	3.29E-03	4.04E-02		no
ENSG00000156738	1.41	1.28	3.67E-03	1.40E-02		yes
ENSG00000252488	1.4	1.3	3.88E-03	1.61E-02		yes
ENSG00000228495	1.62	2.06	4.56E-03	5.52E-03		no
ENSG00000173372	-1.69	-1.83	6.71E-03	7.71E-03		yes
ENSG00000091972	1.17	1.37	8.27E-03	6.83E-03		yes
ENSG00000196937	0.97	1.02	9.24E-03	5.52E-03		yes
ENSG00000166770	1.25	1.33	9.24E-03	1.63E-02		no
ENSG00000101460	-0.85	-0.88	9.24E-03	3.01E-02		yes
ENSG00000152219	0.99	0.92	9.24E-03	3.01E-02		yes
ENSG00000133740	0.96	0.92	9.24E-03	4.04E-02		yes
ENSG00000023445	0.88	0.87	1.07E-02	4.28E-02		yes
ENSG00000100342	-1.06	-0.98	1.11E-02	4.76E-02		no
ENSG00000261655	-1.07	-1.21	1.37E-02	8.04E-03		yes
ENSG00000159958	1.09	1.11	1.55E-02	2.31E-02		yes
ENSG00000100721	1.03	0.97	1.78E-02	2.61E-02		yes
ENSG00000259781	3.76	3.94	1.78E-02	4.04E-02		yes
ENSG00000161381	0.9	0.96	1.89E-02	3.01E-02		yes
ENSG00000211623	1.42	1.93	2.20E-02	1.83E-03		yes
ENSG00000159714	-0.84	-0.89	2.51E-02	3.91E-02		no
ENSG00000153064	0.77	0.72	2.59E-02	4.90E-02		yes
ENSG00000183691	1.05	1.41	2.79E-02	5.52E-03		yes
ENSG00000250850	0.85	1	2.79E-02	3.01E-02		yes
ENSG00000258663	-2.67	-3.11	2.79E-02	4.04E-02		yes
ENSG00000253754	1.08	0.97	2.79E-02	4.65E-02		yes
ENSG00000102760	0.68	0.7	3.04E-02	3.92E-02		yes
ENSG00000089335	1.43	1.37	3.04E-02	4.65E-02		yes
ENSG00000104408	1.07	1.01	3.04E-02	4.65E-02		yes
ENSG00000126860	1.32	1.23	3.04E-02	4.65E-02		no
ENSG00000214184	0.85	0.82	3.04E-02	4.65E-02		no
ENSG00000163534	1.13	1.33	3.06E-02	6.83E-03		yes
ENSG00000145779	0.85	0.88	3.24E-02	4.66E-02		no
ENSG00000168081	0.66	0.73	3.35E-02	3.01E-02		yes

Continued on next page

Table B.3 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000237683	-0.96	-1.16	3.76E-02	1.42E-02		yes
ENSG00000211978	0.95	1.13	3.94E-02	3.01E-02		yes
ENSG00000221420	0.73	0.73	3.94E-02	3.92E-02	1	yes
ENSG00000200418	0.73	0.73	3.94E-02	3.92E-02	1	yes
ENSG00000238942	0.73	0.73	3.94E-02	3.92E-02	1	yes
ENSG00000156976	0.73	0.73	3.94E-02	3.92E-02	1	yes
ENSG00000200320	0.73	0.73	3.94E-02	3.92E-02	1	yes
ENSG00000182853	-1.56	-1.73	3.94E-02	4.91E-02		yes
ENSG00000012223	-2.09	-2.3	3.96E-02	3.55E-02		yes
ENSG00000120088	0.76	0.99	3.97E-02	8.04E-03	2	no
ENSG00000204650	0.76	0.99	3.97E-02	8.04E-03	2	no
ENSG00000251920	0.92	1.37	4.03E-02	6.83E-03		yes
ENSG00000135185	0.75	0.71	4.03E-02	4.65E-02		no
ENSG00000258511	0.92	1.07	4.04E-02	3.76E-02		yes
ENSG00000142875	1.26	1.46	4.12E-02	1.93E-02		yes
ENSG00000238179	-0.84	-1.03	4.12E-02	2.25E-02		yes
ENSG00000228323	-0.85	-0.95	4.12E-02	3.56E-02		no
ENSG00000226937	0.88	1.02	4.42E-02	3.25E-02		yes
ENSG00000153130	1.26	1.58	4.58E-02	1.93E-02		yes
ENSG00000163520	1.13	1.27	4.90E-02	4.04E-02		yes
ENSG00000249310	-2.28	-2.1	4.90E-02	4.65E-02		yes
ENSG00000133962	0.99	1.12	4.99E-02	3.01E-02	3	yes
ENSG00000165929	0.99	1.12	4.99E-02	3.01E-02	3	yes

SMA2 *vs.* SMA3

Table B.4: Differential expressed genes selected for the “SMA2 versus SMA3” comparison: Ensembl gene IDs, log-ratios and q-values, groups of *overlapping-genes*, labelled with the same group number, and genes confirmed by the *totcounts-genes* approach.

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000164821	-2.95	-2.99	3.19E-12	3.17E-12		yes
ENSG00000249310	-4.17	-3.73	8.66E-12	4.70E-10		yes
ENSG00000148346	-3.11	-2.98	2.14E-10	5.03E-10		yes
ENSG00000235508	8.89	8.22	7.50E-10	5.59E-09		yes
ENSG00000211637	2.11	2.11	1.01E-08	1.09E-09		yes
ENSG00000096006	-3	-3.3	2.90E-08	5.59E-09		yes
ENSG00000124469	-2.74	-3.37	2.85E-07	5.03E-10		yes
ENSG00000164047	-2.21	-2.29	2.85E-07	8.36E-08		yes
ENSG00000118113	-3.33	-3.36	2.87E-07	3.35E-06		yes
ENSG00000012223	-3.33	-3.53	3.09E-07	3.10E-08		yes
ENSG00000197149	4.96	4.47	3.60E-07	3.52E-07		yes
ENSG00000256515	9.27	9.94	4.44E-07	2.30E-07		yes
ENSG00000228695	-8.26	-8.32	1.46E-06	1.19E-05		yes
ENSG00000133063	-2.74	-2.94	2.93E-06	8.96E-06		yes
ENSG00000240342	2.72	2.73	4.37E-06	1.29E-05		yes
ENSG00000134827	-1.96	-1.83	6.81E-06	2.83E-04		yes
ENSG00000169397	-2.23	-2.42	8.87E-06	1.29E-05		yes
ENSG00000223350	1.97	2.2	9.55E-05	3.33E-06		yes
ENSG00000101425	-1.88	-1.99	1.04E-04	1.33E-04		yes
ENSG00000197561	-1.97	-2.29	1.15E-04	6.87E-05		yes
ENSG00000211650	2.61	2.45	1.15E-04	1.27E-04		yes
ENSG00000188056	3.96	4.56	1.15E-04	3.34E-04		yes
ENSG00000181126	-4.75	-4.29	1.17E-04	1.07E-03		no
ENSG00000102837	-2.58	-3.07	2.81E-04	6.87E-05		yes
ENSG00000149516	-1.68	-1.7	3.04E-04	3.19E-04		yes
ENSG00000086548	-2.27	-2.89	3.50E-04	2.25E-06		yes
ENSG00000172232	-1.78	-1.92	4.62E-04	4.15E-04		yes
ENSG00000118520	-1.79	-1.99	5.46E-04	1.67E-04		yes
ENSG00000231896	-1.91	-1.9	7.97E-04	9.55E-03		no
ENSG00000250765	-2.67	-2.7	8.04E-04	2.36E-03		yes
ENSG00000216083	-2.68	-2.85	8.04E-04	5.08E-03	1	yes
ENSG00000065618	-2.68	-2.85	8.04E-04	5.08E-03	1	yes
ENSG00000173391	-2.51	-2.6	9.69E-04	6.67E-03		yes
ENSG00000236650	-4.2	-4.64	1.32E-03	1.58E-03		yes
ENSG00000206249	-1.73	-1.72	2.18E-03	1.07E-03		yes
ENSG00000100448	-1.85	-2.08	2.84E-03	4.36E-04		yes
ENSG00000242580	2.49	2.1	2.84E-03	1.13E-03		yes
ENSG00000211598	2.08	2.1	3.03E-03	1.13E-03		yes
ENSG00000211938	2.11	1.99	3.03E-03	4.25E-03		yes
ENSG00000211665	2.07	1.82	3.87E-03	2.62E-02		no
ENSG00000211654	2.41	2.25	5.78E-03	2.30E-02		no
ENSG00000213147	7.53	9.04	7.00E-03	2.62E-04		yes
ENSG00000211663	1.54	1.48	7.33E-03	1.29E-02		yes
ENSG00000170801	-2.82	-3.3	8.49E-03	3.99E-03		yes
ENSG00000253497	3.68	3.54	1.18E-02	3.32E-02		no
ENSG00000218749	-1.18	-1.22	1.21E-02	1.55E-02		no

Continued on next page

Table B.4 – *Continued from previous page*

Gene	Log-ratio		q-value		Overlap. group ID	Confirmed
	maxcounts	*totcounts*	*maxcounts*	*totcounts*		
ENSG00000104918	-1.46	-1.4	1.36E-02	3.29E-02		yes
ENSG00000243166	-1.16	-1.2	1.69E-02	2.18E-02		no
ENSG00000196415	-1.82	-1.87	1.84E-02	2.30E-02		yes
ENSG00000239862	1.68	1.85	2.42E-02	4.65E-03		yes
ENSG00000231475	1.37	1.51	2.79E-02	1.11E-02		no
ENSG00000255641	1.58	1.57	2.94E-02	3.49E-02	2	yes
ENSG00000205810	1.58	1.57	2.94E-02	3.49E-02	2	yes
ENSG00000205809	1.58	1.57	2.94E-02	3.49E-02	2	yes
ENSG00000134545	1.58	1.57	2.94E-02	3.49E-02	2	yes
ENSG00000212579	-1.36	-1.39	3.33E-02	4.25E-02		no
ENSG00000079393	-2.73	-3.13	3.40E-02	3.50E-02		yes
ENSG00000253239	1.8	1.96	3.65E-02	1.16E-02		yes
ENSG00000244575	1.12	1.13	4.30E-02	3.97E-02		no

B.2 Ingenuity Pathway Analysis of differentially expressed genes

SMA *vs.* CTRL

Table B.5: Features of the genes differentially expressed in "SMA vs CTRL" comparison considered for network analysis: Ensemble gene ID, log-ratio and q-value. For genes annotated in the IPA database, gene symbol and Entrez name are also reported.

Gene	log-ratio	q-value	Symbol	Entrez Name
ENSG00000008311	-0.676	2.94E-02	AASS	aminoadipate-semialdehyde synthase
ENSG00000135776	0.691	1.28E-02	ABCB10	ATP-binding cassette, sub-family B (MDRTAP), member 10
ENSG00000164163	0.55	3.29E-02	ABCE1	ATP-binding cassette, sub-family E (OABP), member 1
ENSG00000108798	-0.573	4.80E-02	ABI3	ABI family, member 3
ENSG00000099204	0.723	2.87E-02	ABLIM1	actin binding LIM protein 1
ENSG00000110455	-0.615	4.50E-02	ACCS	1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)
ENSG00000068366	0.569	2.68E-02	ACSL4	acyl-CoA synthetase long-chain family member 4
ENSG00000107796	-0.738	2.85E-02	ACTA2	actin, alpha 2, smooth muscle, aorta
ENSG00000130402	-0.586	3.34E-02	ACTN4	actinin, alpha 4
ENSG00000143382	-0.559	2.87E-02	ADAMTSL4	ADAMTS-like 4
ENSG00000185761	-0.926	2.54E-02	ADAMTSL5	ADAMTS-like 5
ENSG00000185736	-3.738	3.46E-02	ADARB2	adenosine deaminase, RNA-specific, B2 (non-functional)
ENSG00000148700	0.762	1.02E-03	ADD3	adducin 3 (gamma)
ENSG00000106351	-0.433	3.79E-02	AGFG2	ArfGAP with FG repeats 2
ENSG00000186063	0.552	9.99E-03	AIDA	axin interactor, dorsalization associated
ENSG00000147853	0.497	3.77E-02	AK3	adenylate kinase 3
ENSG00000198796	-1.61	9.15E-03	ALPK2	alpha-kinase 2
ENSG00000164331	0.664	1.61E-02	ANKRA2	ankyrin repeat, family A (RFXANK-like), 2
ENSG00000168876	0.773	2.87E-02	ANKRD49	ankyrin repeat domain 49
ENSG00000184730	-0.616	2.47E-02	APOBR	apolipoprotein B receptor
ENSG00000100342	-0.644	1.93E-02	APOL1	apolipoprotein L, 1
ENSG00000134884	0.91	1.79E-03	ARGLU1	arginine and glutamate rich 1
ENSG00000165322	0.774	2.87E-02	ARHGAP12	Rho GTPase activating protein 12
ENSG00000232686	1.057	8.08E-04	ARID4B-IT1	ARID4B intronic transcript 1
ENSG00000152219	0.706	6.60E-03	ARL14EP	ADP-ribosylation factor-like 14 effector protein
ENSG00000113369	0.803	3.40E-03	ARRDC3	arrestin domain containing 3
ENSG00000110881	1.138	3.95E-02	ASIC1	acid-sensing (proton-gated) ion channel 1
ENSG00000138138	0.801	9.47E-03	ATAD1	ATPase family, AAA domain containing 1
ENSG00000111676	-0.528	3.85E-02	ATN1	atrophin 1
ENSG00000058063	0.539	3.60E-02	ATP11B	ATPase, class VI, type 11B
ENSG00000180389	-0.924	1.61E-02	ATP5EP2	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit pseudogene 2
ENSG00000124406	0.871	3.29E-02	ATP8A1	ATPase, aminophospholipid transporter (APLT), class I, type 8A, member 1
ENSG00000121753	-0.929	3.57E-02	BAI2	brain-specific angiogenesis inhibitor 2
ENSG00000123685	-0.845	8.84E-03	BATF3	basic leucine zipper transcription factor, ATF-like 3
ENSG00000105829	0.572	3.51E-02	BET1	Bet1 golgi vesicular membrane trafficking protein
ENSG00000102409	0.504	2.99E-02	BEX4	brain expressed, X-linked 4
ENSG00000110330	0.707	1.23E-02	BIRC2	baculoviral IAP repeat containing 2
ENSG00000023445	0.659	3.87E-03	BIRC3	baculoviral IAP repeat containing 3
ENSG00000140299	0.519	7.79E-03	BNIP2	BCL2adenovirus E1B 19kDa interacting protein 2
ENSG00000176720	-1.539	2.31E-02	BOK	BCL2-related ovarian killer
ENSG00000134717	0.506	3.79E-02	BTF3L4	basic transcription factor 3-like 4

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000186265	0.706	3.29E-02	BTLA	B and T lymphocyte associated
ENSG00000214688	-0.489	2.85E-02	C10orf105	chromosome 10 open reading frame 105
ENSG00000110696	0.657	2.19E-03	C11orf58	chromosome 11 open reading frame 58
ENSG00000166323	0.815	3.48E-02	C11orf65	chromosome 11 open reading frame 65
ENSG00000151135	1.082	5.40E-04	C12orf23	chromosome 12 open reading frame 23
ENSG00000133641	0.808	2.48E-02	C12orf29	chromosome 12 open reading frame 29
ENSG00000134548	1.091	1.53E-02	C12orf39	chromosome 12 open reading frame 39
ENSG00000162398	-1.04	4.96E-02	C1orf177	chromosome 1 open reading frame 177
ENSG00000173372	-1.028	1.50E-02	C1QA	complement component 1, q subcomponent, A chain
ENSG00000213204	0.663	2.99E-02	C6orf165	chromosome 6 open reading frame 165
ENSG00000147894	0.808	5.30E-03	C9orf72	chromosome 9 open reading frame 72
ENSG00000152495	0.592	2.44E-02	CAMK4	calciumcalmodulin-dependent protein kinase IV
ENSG00000116489	0.722	2.45E-03	CAPZA1	capping protein (actin filament) muscle Z-line, alpha 1
ENSG00000198898	0.694	1.22E-02	CAPZA2	capping protein (actin filament) muscle Z-line, alpha 2
ENSG00000164305	0.864	3.36E-03	CASP3	caspase 3, apoptosis-related cysteine peptidase
ENSG00000133962	0.931	2.18E-03	CATSPERB	catsper channel auxiliary subunit beta
ENSG00000122565	0.602	1.53E-02	CBX3	chromobox homolog 3
ENSG00000132024	-0.539	1.77E-02	CC2D1A	coiled-coil and C2 domain containing 1A
ENSG00000168071	-0.468	3.89E-02	CCDC88B	coiled-coil domain containing 88B
ENSG00000256515	-3.916	3.48E-02	CCL3L1/CCL3L3	chemokine (C-C motif) ligand 3-like 1
ENSG00000112237	0.767	5.59E-03	CCNC	cyclin C
ENSG00000113328	0.613	3.59E-03	CCNG1	cyclin G1
ENSG00000138764	0.636	1.54E-02	CCNG2	cyclin G2
ENSG00000163660	0.509	2.26E-02	CCNL1	cyclin L1
ENSG00000135535	0.665	6.51E-04	CD164	CD164 molecule, sialomucin
ENSG00000091972	0.799	1.50E-02	CD200	CD200 molecule
ENSG00000198087	0.824	3.36E-02	CD2AP	CD2-associated protein
ENSG00000160654	0.7	1.28E-02	CD3G	CD3g molecule, gamma (CD3-TCR complex)
ENSG00000117335	0.716	2.45E-03	CD46	CD46 molecule, complement regulatory protein
ENSG00000110848	0.855	3.21E-02	CD69	CD69 molecule
ENSG00000106993	0.652	4.32E-02	CDC37L1	cell division cycle 37-like 1
ENSG00000129910	-0.938	3.28E-02	CDH15	cadherin 15, type 1, M-cadherin (myotubule)
ENSG00000059758	0.639	3.50E-02	CDK17	cyclin-dependent kinase 17
ENSG00000129757	-0.877	1.53E-02	CDKN1C	cyclin-dependent kinase inhibitor 1C (p57, Kip2)
ENSG00000178863	-0.73	1.36E-02	CEBPA-AS1	CEBPA antisense RNA 1 (head to head)
ENSG00000221869	0.61	1.00E-02	CEBPD	CCAATenhancer binding protein (CEBP), delta
ENSG00000123219	1.815	3.16E-03	CENPK	centromere protein K
ENSG00000134255	0.492	3.37E-02	CEPT1	cholineethanolamine phosphotransferase 1
ENSG00000133019	1.211	3.58E-03	CHRM3	cholinergic receptor, muscarinic 3
ENSG00000233355	0.847	1.23E-02	CHRM3-AS2	CHRM3 antisense RNA 2
ENSG00000166165	-0.767	2.00E-02	CKB	creatine kinase, brain
ENSG00000132514	-0.607	1.85E-02	CLEC10A	C-type lectin domain family 10, member A
ENSG00000013441	0.833	2.42E-02	CLK1	CDC-like kinase 1
ENSG00000113240	0.925	5.59E-03	CLK4	CDC-like kinase 4
ENSG00000074201	0.529	1.18E-02	CLNS1A	chloride channel, nucleotide-sensitive, 1A
ENSG00000162368	0.884	1.34E-05	CMPK1	cytidine monophosphate (UMP-CMP) kinase 1, cytosolic
ENSG00000205423	0.58	4.23E-02	CNEP1R1	CTD nuclear envelope phosphatase 1 regulatory subunit 1
ENSG00000100528	0.636	8.40E-04	CNIH	cornichon homolog (Drosophila)
ENSG00000198791	0.516	6.67E-03	CNOT7	CCR4-NOT transcription complex, subunit 7
ENSG00000162852	0.65	3.29E-02	CNST	consortin, connexin sorting protein
ENSG00000198756	-0.991	3.58E-03	COLGALT2	collagen beta(1-O)galactosyltransferase 2
ENSG00000166200	0.743	4.64E-02	COPS2	COP9 signalosome subunit 2
ENSG00000103426	-0.389	2.90E-02	CORO7/CORO7-PAM16	coronin 7

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000262246	-0.389	2.90E-02	CORO7/CORO7-PAM16	coronin 7
ENSG00000158525	1.788	5.28E-04	CPA5	carboxypeptidase A5
ENSG00000111269	0.704	2.49E-03	CREBL2	cAMP responsive element binding protein-like 2
ENSG00000182809	-0.629	2.92E-02	CRIP2	cysteine-rich protein 2
ENSG00000179979	-0.725	2.16E-03	CRIPAK	cysteine-rich PAK1 inhibitor
ENSG00000176390	0.446	2.01E-02	CRLF3	cytokine receptor-like factor 3
ENSG00000109943	0.649	3.38E-02	CRTAM	cytotoxic and regulatory T cell molecule
ENSG00000182578	-0.666	3.50E-02	CSF1R	colony stimulating factor 1 receptor
ENSG00000151292	0.709	4.76E-02	CSNK1G3	casein kinase 1, gamma 3
ENSG00000008283	-0.566	2.84E-02	CYB561	cytochrome b561
ENSG00000115866	0.497	1.18E-02	DARS	aspartyl-tRNA synthetase
ENSG00000156136	0.699	9.15E-03	DCK	deoxycytidine kinase
ENSG00000043093	0.546	1.32E-02	DCUN1D1	DCN1, defective in cullin neddylation 1, domain containing 1
ENSG00000165359	0.653	1.59E-02	DDX26B	DEADH (Asp-Glu-Ala-AspHis) box polypeptide 26B
ENSG00000160570	-0.423	3.41E-02	DEDD2	death effector domain containing 2
ENSG00000104936	-0.517	2.67E-02	DMPK	dystrophia myotonica-protein kinase
ENSG00000141096	-0.482	3.52E-02	DPEP3	dipeptidase 3
ENSG00000177990	1.776	1.61E-02	DPY19L2	dpy-19-like 2 (C. elegans)
ENSG00000134765	1.75	6.35E-03	DSC1	desmocollin 1
ENSG00000165169	0.66	4.82E-02	DYNLT3	dynein, light chain, Tctex-type 3
ENSG00000133740	0.76	5.17E-03	E2F5	E2F transcription factor 5, p130-binding
ENSG00000255423	0.665	2.66E-02	EBLN2	endogenous Bornavirus-like nucleoprotein 2
ENSG00000123179	0.734	5.28E-04	EBPL	emopamil binding protein-like
ENSG00000255302	0.471	3.44E-02	EID1	EP300 interacting inhibitor of differentiation 1
ENSG00000225037	0.813	1.13E-02	EIF1AX-AS1	EIF1AX antisense RNA 1
ENSG00000144895	0.578	1.10E-02	EIF2A	eukaryotic translation initiation factor 2A, 65kDa
ENSG00000205609	-1.475	4.23E-02	EIF3C/EIF3CL	eukaryotic translation initiation factor 3, subunit C
ENSG00000104408	0.812	7.38E-03	EIF3E	eukaryotic translation initiation factor 3, subunit E
ENSG00000156976	0.575	9.15E-03	EIF4A2	eukaryotic translation initiation factor 4A2
ENSG00000151247	0.657	5.61E-03	EIF4E	eukaryotic translation initiation factor 4E
ENSG00000163412	0.546	1.06E-02	EIF4E3	eukaryotic translation initiation factor 4E family member 3
ENSG00000253626	-1.44	2.53E-03	EIF5AL1	eukaryotic translation initiation factor 5A-like 1
ENSG00000110675	1.039	1.19E-03	ELMOD1	ELMOCED-12 domain containing 1
ENSG00000012660	0.728	6.02E-04	ELOVL5	ELOVL fatty acid elongase 5
ENSG00000170571	0.679	3.58E-03	EMB	embigin
ENSG00000106991	-0.622	2.78E-02	ENG	endoglin
ENSG00000168913	-0.751	1.29E-02	ENHO	energy homeostasis associated
ENSG00000224032	0.484	3.53E-02	EPB41L4A-AS1	EPB41L4A antisense RNA 1
ENSG00000112851	0.639	3.31E-02	ERBB2IP	erbb2 interacting protein
ENSG00000134954	0.448	1.57E-02	ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)
ENSG00000126860	1.028	4.19E-03	EVI2A	ecotropic viral integration site 2A
ENSG00000070367	0.77	1.05E-02	EXOC5	exocyst complex component 5
ENSG00000136830	-0.557	4.30E-02	FAM129B	family with sequence similarity 12,9 member B
ENSG00000154153	0.771	2.28E-02	FAM134B	family with sequence similarity 134, member B
ENSG00000163322	0.716	1.77E-02	FAM175A	family with sequence similarity 175, member A
ENSG00000230567	-5.581	2.01E-02	FAM203A/FAM203B	family with sequence similarity 203, member A
ENSG00000170215	-1.505	5.61E-03	FAM27B	family with sequence similarity 27, member B
ENSG00000196937	0.833	1.21E-03	FAM3C	family with sequence similarity 3, member C
ENSG00000172717	-2.49	5.28E-04	FAM71D	family with sequence similarity 71, member D
ENSG00000077458	0.788	2.77E-02	FAM76B	family with sequence similarity 76, member B
ENSG00000005812	0.481	3.30E-02	FBXL3	F-box and leucine-rich repeat protein 3
ENSG00000110429	0.569	2.90E-02	FBXO3	F-box protein 3
ENSG00000119616	0.607	7.29E-03	FCF1	FCF1 rRNA-processing protein

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000163534	0.791	2.10E-02	FCRL1	Fc receptor-like 1
ENSG00000111790	0.523	3.76E-02	FGFR1OP2	FGFR1 oncogene partner 2
ENSG00000135723	-0.468	1.49E-02	FHOD1	formin homology 2 domain containing 1
ENSG00000248265	-1.092	9.66E-03	FLJ12825	uncharacterized LOC440101
ENSG00000196924	-0.626	3.05E-02	FLNA	filamin A, alpha
ENSG00000141429	0.492	3.37E-02	GALNT1	UDP-N-acetyl-alpha-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase 1 (GalNAc-T1)
ENSG00000158555	-0.714	6.64E-03	GDPD5	glycerophosphodiester phosphodiesterase domain containing 5
ENSG00000100031	-0.457	2.58E-02	GGT1	gamma-glutamyltransferase 1
ENSG00000106560	0.708	1.26E-02	GIMAP2	GTPase, IMAP family member 2
ENSG00000124767	0.497	8.23E-03	GLO1	glyoxalase I
ENSG00000115419	0.626	1.47E-02	GLS	glutaminase
ENSG00000087338	0.59	2.18E-02	GMCL1	germ cell-less, spermatogenesis associated 1
ENSG00000197045	1.096	7.46E-03	GMFB	glia maturation factor, beta
ENSG00000127920	0.695	3.18E-03	GNG11	guanine nucleotide binding protein (G protein), gamma 11
ENSG00000130119	-0.766	1.33E-02	GNL3L	guanine nucleotide binding protein-like 3 (nucleolar)-like
ENSG00000152133	0.734	3.36E-02	GPATCH11	G patch domain containing 11
ENSG00000125772	0.658	1.95E-02	GPCPD1	glycerophosphocholine phosphodiesterase GDE1 homolog (S. cerevisiae)
ENSG00000125245	0.483	6.47E-03	GPR18	G protein-coupled receptor 18
ENSG00000204882	-0.952	1.35E-02	GPR20	G protein-coupled receptor 20
ENSG00000140030	0.596	1.89E-02	GPR65	G protein-coupled receptor 65
ENSG00000119714	-0.501	3.54E-02	GPR68	G protein-coupled receptor 68
ENSG00000148180	-0.599	4.64E-02	GSN	gelsolin
ENSG00000235865	-0.648	4.73E-02	GSN-AS1	GSN antisense RNA 1
ENSG00000111358	0.646	3.70E-02	GTF2H3	general transcription factor IIH, polypeptide 3, 34kDa
ENSG00000172986	0.689	1.52E-02	GXYLT2	glucoside xylosyltransferase 2
ENSG00000130600	-2.116	1.18E-03	H19	H1,9 imprinted maternally expressed transcript
ENSG00000205581	0.553	2.88E-03	HMGN1	high mobility group nucleosome binding domain 1
ENSG00000100292	-0.581	3.25E-02	HMOX1	heme oxygenase (decycling) 1
ENSG00000025156	0.523	4.57E-02	HSF2	heat shock transcription factor 2
ENSG00000171116	-5.263	2.60E-08	HSFX1/HSFX2	heat shock transcription factor family, X linked 1
ENSG00000090339	-0.494	4.08E-02	ICAM1	intercellular adhesion molecule 1
ENSG00000105371	-0.647	2.73E-02	ICAM4	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)
ENSG00000163600	0.814	1.54E-02	ICOS	inducible T-cell co-stimulator
ENSG00000067064	0.711	1.40E-02	IDI1	isopentenyl-diphosphate delta isomerase 1
ENSG00000162594	1.224	2.98E-03	IL23R	interleukin 23 receptor
ENSG00000164509	-0.599	4.49E-02	IL31RA	interleukin 31 receptor A
ENSG00000134352	1.054	1.33E-02	IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor)
ENSG00000168685	0.958	6.11E-05	IL7R	interleukin 7 receptor
ENSG00000125629	0.559	3.98E-02	INSIG2	insulin induced gene 2
ENSG00000091409	0.698	1.32E-02	ITGA6	integrin, alpha 6
ENSG00000113263	0.515	1.92E-02	ITK	IL2-inducible T-cell kinase
ENSG00000176076	0.569	2.68E-02	KCNE1L	KCNE1-like
ENSG00000123700	0.617	4.11E-02	KCNJ2	potassium inwardly-rectifying channel, subfamily J, member 2
ENSG00000173338	-0.825	3.50E-03	KCNK7	potassium channel, subfamily K, member 7
ENSG00000136051	0.653	3.70E-02	KIAA1033	KIAA1033
ENSG00000170759	0.595	3.64E-02	KIF5B	kinesin family member 5B
ENSG00000257702	-0.66	3.66E-02	LBX2-AS1	LBX2 antisense RNA 1
ENSG00000214402	-1.026	3.32E-02	LCNL1	lipocalin-like 1
ENSG00000138795	0.653	7.16E-03	LEF1	lymphoid enhancer-binding factor 1
ENSG00000232021	0.803	3.68E-02	LEF1-AS1	LEF1 antisense RNA 1
ENSG00000174106	0.763	1.61E-02	LEMD3	LEM domain containing 3
ENSG00000105609	-0.57	1.80E-02	LILRB5	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 5

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000149656	0.766	1.26E-02	LINC00266-1	long intergenic non-protein coding RNA 266-1
ENSG00000132204	-1.172	9.68E-04	LINC00470	long intergenic non-protein coding RNA 470
ENSG00000160789	-1.053	2.01E-03	LMNA	lamin AC
ENSG00000233251	0.799	2.55E-04	LOC100129434	uncharacterized LOC100129434
ENSG00000258511	1.019	6.51E-04	LOC100996339	uncharacterized LOC100996339
ENSG00000225996	-0.559	2.87E-02	LOC100996516	uncharacterized LOC100996516
ENSG00000256039	0.844	5.80E-03	LOC101060038	uncharacterized LOC101060038
ENSG00000233392	-0.855	1.68E-02	LOC200772	uncharacterized LOC200772
ENSG00000215458	-0.608	2.31E-02	LOC284837	uncharacterized LOC284837
ENSG00000243440	0.492	3.66E-02	LOC388813	uncharacterized protein ENSP00000383407-like
ENSG00000214194	0.958	2.28E-02	LOC401397	uncharacterized LOC401397
ENSG00000232022	-0.817	1.30E-02	LOC729041	fatty acid amide hydrolase pseudogene
ENSG00000156564	-1.426	2.18E-02	LRFN2	leucine rich repeat and fibronectin type III domain containing 2
ENSG00000066557	0.663	4.57E-02	LRRC40	leucine rich repeat containing 40
ENSG00000253102	-0.533	1.64E-02	LRRC59	leucine rich repeat containing 59
ENSG00000090006	-0.587	2.90E-02	LTBP4	latent transforming growth factor beta binding protein 4
ENSG00000120992	0.618	1.25E-02	LYPLA1	lysophospholipase I
ENSG00000164109	0.803	1.19E-02	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
ENSG00000102158	0.555	2.17E-02	MAGT1	magnesium transporter 1
ENSG00000172469	1.177	1.61E-02	MANEA	mannosidase, endo-alpha
ENSG00000130479	-0.57	2.47E-02	MAP1S	microtubule-associated protein 1S
ENSG00000142733	-0.727	2.66E-02	MAP3K6	mitogen-activated protein kinase kinase kinase 6
ENSG00000136536	0.564	8.23E-03	7-Mar	membrane-associated ring finger (C3HC4) 7, E3 ubiquitin protein ligase
ENSG00000015479	0.778	1.36E-03	MATR3	matrin 3
ENSG00000152601	0.627	4.03E-02	MBNL1	muscleblind-like splicing regulator 1
ENSG00000099917	-0.45	1.68E-02	MED15	mediator complex subunit 15
ENSG00000104973	-0.453	4.67E-02	MED25	mediator complex subunit 25
ENSG00000214548	-1.767	6.27E-03	MEG3	maternally expressed 3 (non-protein coding)
ENSG00000109736	-0.468	3.37E-02	MFSD10	major facilitator superfamily domain containing 10
ENSG00000198160	0.916	1.24E-03	MIER1	mesoderm induction early response 1 homolog (Xenopus laevis)
ENSG00000155545	1.222	3.38E-02	MIER3	mesoderm induction early response 1, family member 3
ENSG00000221065	1.028	1.26E-02	mir-1233	microRNA 1233-1
ENSG00000207567	-0.797	6.67E-03	mir-142	microRNA 142
ENSG00000207721	0.798	1.49E-02	mir-186	microRNA 186
ENSG00000221214	0.69	7.29E-03	mir-548	microRNA 548c
ENSG00000207733	-0.548	4.20E-02	mir-637	microRNA 637
ENSG00000207972	-0.473	3.51E-02	mir-638	microRNA 638
ENSG00000211502	-2.116	1.18E-03	mir-675	microRNA 675
ENSG00000211589	0.842	1.67E-03	mir-744	microRNA 744
ENSG00000211591	0.575	1.49E-02	mir-762	microRNA 762
ENSG00000215973	1.037	7.58E-03	mir-933	microRNA 933
ENSG00000260978	3.78	1.89E-02	MKRN3-AS1	MKRN3 antisense RNA 1
ENSG00000108960	0.692	1.64E-02	MMD	monocyte to macrophage differentiation-associated
ENSG00000173542	0.699	9.15E-03	MOB1B	MOB kinase activator 1B
ENSG00000257802	0.797	2.25E-03	MRS2P2	MRS2 pseudogene 2
ENSG00000156738	1.017	1.36E-03	MS4A1	membrane-spanning 4-domains, subfamily A, member 1
ENSG00000149516	0.685	2.78E-02	MS4A3	membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)
ENSG00000214787	-0.569	3.29E-02	MS4A4E	membrane-spanning 4-domains, subfamily A, member 4E
ENSG00000095002	0.571	2.54E-02	MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
ENSG00000174579	0.603	1.08E-02	MSL2	male-specific lethal 2 homolog (Drosophila)
ENSG00000210140	0.957	5.38E-05	MT-TC	tRNA
ENSG00000210144	0.957	5.38E-05	MT-TY	tRNA

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000128654	0.656	4.97E-02	MTX2	metaxin 2
ENSG00000179820	-0.55	2.47E-02	MYADM	myeloid-associated differentiation marker
ENSG00000085274	0.674	2.74E-02	MYNN	myoneurin
ENSG00000136286	-0.485	4.65E-02	MYO1G	myosin IG
ENSG00000172766	0.677	4.05E-02	NAA16	N(alpha)-acetyltransferase 16, NatA auxiliary subunit
ENSG00000228224	-1.339	4.05E-07	NACAP1	nascent-polypeptide-associated complex alpha polypeptide pseudogene 1
ENSG00000187109	0.544	1.18E-02	NAP1L1	nucleosome assembly protein 1-like 1
ENSG00000196498	-0.592	8.40E-03	NCOR2	nuclear receptor corepressor 2
ENSG00000151414	1.022	1.18E-03	NEK7	NIMA-related kinase 7
ENSG00000184613	0.531	2.01E-02	NELL2	NEL-like 2 (chicken)
ENSG00000162711	-0.612	2.12E-02	NLRP3	NLR family, pyrin domain containing 3
ENSG00000169251	0.764	1.81E-02	NMD3	NMD3 homolog (S. cerevisiae)
ENSG00000106733	0.464	3.83E-02	NMRK1	nicotinamide riboside kinase 1
ENSG00000183691	0.842	5.59E-03	NOG	noggin
ENSG00000123358	-0.937	1.46E-02	NR4A1	nuclear receptor subfamily 4, group A, member 1
ENSG00000091129	1.503	1.37E-02	NRCAM	neuronal cell adhesion molecule
ENSG00000164346	0.652	1.89E-02	NSA2	NSA2 ribosome biogenesis homolog (S. cerevisiae)
ENSG00000135318	0.761	1.37E-02	NT5E	5'-nucleotidase, ecto (CD73)
ENSG00000173598	0.601	2.44E-03	NUDT4	nudix (nucleoside diphosphate linked moiety X)-type motif 4
ENSG00000177144	1.081	1.39E-02	NUDT4P1	nudix (nucleoside diphosphate linked moiety X)-type motif 4 pseudogene 1
ENSG00000153989	0.5	3.46E-02	NUS1	nuclear undecaprenyl pyrophosphate synthase 1 homolog (S. cerevisiae)
ENSG00000101888	0.827	3.50E-02	NXT2	nuclear transport factor 2-like export factor 2
ENSG00000105953	-0.463	4.88E-02	OGDH	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)
ENSG00000147162	0.751	6.67E-03	OGT	O-linked N-acetylglucosamine (GlcNAc) transferase
ENSG00000221970	1.2	6.92E-05	OR2A1/OR2A42	olfactory receptor, family 2, subfamily A, member 1
ENSG00000170356	0.853	8.54E-03	OR2A20P	olfactory receptor, family 2, subfamily A, member 20 pseudogene
ENSG00000128699	0.556	1.89E-02	ORMDL1	ORM1-like 1 (S. cerevisiae)
ENSG00000006025	-0.486	3.98E-02	OSBPL7	oxysterol binding protein-like 7
ENSG00000091039	0.767	2.67E-02	OSBPL8	oxysterol binding protein-like 8
ENSG00000164823	0.71	1.88E-03	OSGIN2	oxidative stress induced growth inhibitor family member 2
ENSG00000145623	0.899	1.73E-02	OSMR	oncostatin M receptor
ENSG00000198856	0.538	1.49E-02	OSTC	oligosaccharyltransferase complex subunit (non-catalytic)
ENSG00000089723	-0.854	2.99E-02	OTUB2	OTU domain, ubiquitin aldehyde binding 2
ENSG00000155100	0.927	2.05E-02	OTUD6B	OTU domain containing 6B
ENSG00000154814	0.647	8.74E-03	OXNAD1	oxidoreductase NAD-binding domain containing 1
ENSG00000164830	0.807	1.08E-02	OXR1	oxidation resistance 1
ENSG00000254615	0.807	1.08E-02	OXR1	oxidation resistance 1
ENSG00000078589	0.525	3.33E-02	P2RY10	purinergic receptor P2Y, G-protein coupled, 10
ENSG00000217930	-0.389	2.90E-02	PAM16	presequence translocase-associated motor 16 homolog (S. cerevisiae)
ENSG00000238197	0.782	1.29E-02	PAXBP1-AS1	PAXBP1 antisense RNA 1
ENSG00000177839	-0.723	1.74E-02	PCDHB9	protocadherin beta 9
ENSG00000168300	0.586	2.77E-02	PCMTD1	protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1
ENSG00000203880	0.766	1.26E-02	PCMTD2	protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2
ENSG00000071994	0.436	1.28E-02	PDCD2	programmed cell death 2
ENSG00000152270	0.626	2.11E-02	PDE3B	phosphodiesterase 3B, cGMP-inhibited
ENSG00000205268	0.683	2.18E-03	PDE7A	phosphodiesterase 7A
ENSG00000152256	0.619	2.28E-02	PDK1	pyruvate dehydrogenase kinase, isozyme 1
ENSG00000241360	-0.463	4.19E-02	PDXP	pyridoxal (pyridoxine, vitamin B6) phosphatase
ENSG00000197329	0.707	2.85E-03	PELI1	pellino E3 ubiquitin protein ligase 1
ENSG00000154330	-1.455	1.09E-02	PGM5	phosphoglucomutase 5
ENSG00000101856	0.487	1.45E-02	PGRMC1	progesterone receptor membrane component 1
ENSG00000165195	0.964	2.41E-02	PIGA	phosphatidylinositol glycan anchor biosynthesis, class A
ENSG00000142892	0.758	1.23E-02	PIGK	phosphatidylinositol glycan anchor biosynthesis, class K

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000100100	0.456	2.55E-02	PIK3IP1	phosphoinositide-3-kinase interacting protein 1
ENSG00000110697	-0.359	3.84E-02	PITPNM1	phosphatidylinositol transfer protein, membrane-associated 1
ENSG00000149782	-0.638	1.54E-02	PLCB3	phospholipase C, beta 3 (phosphatidylinositol-specific)
ENSG00000116095	0.497	3.63E-02	PLEKHA3	pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 3
ENSG00000175895	0.636	1.51E-02	PLEKHF2	pleckstrin homology domain containing, family F (with FYVE domain) member 2
ENSG00000008323	-1.118	1.46E-02	PLEKHG6	pleckstrin homology domain containing, family G (with RhoGef domain) member 6
ENSG00000161381	0.774	2.16E-03	PLXDC1	plexin domain containing 1
ENSG00000196576	-0.624	3.03E-02	PLXNB2	plexin B2
ENSG00000135241	0.673	1.48E-02	PNPLA8	patatin-like phospholipase domain containing 8
ENSG00000189266	0.597	6.44E-03	PNRC2	proline-rich nuclear receptor coactivator 2
ENSG00000181222	-0.49	3.29E-02	POLR2A	polymerase (RNA) II (DNA directed) polypeptide A, 220kDa
ENSG00000138032	0.64	1.16E-02	PPM1B	protein phosphatase, Mg2+Mn2+ dependent, 1B
ENSG00000155367	-0.56	4.51E-02	PPM1J	protein phosphatase, Mg2+Mn2+ dependent, 1J
ENSG00000213639	0.733	2.31E-03	PPP1CB	protein phosphatase 1, catalytic subunit, beta isozyme
ENSG00000163605	0.689	1.52E-02	PPP4R2	protein phosphatase 4, regulatory subunit 2
ENSG00000133246	-0.593	3.66E-02	PRAM1	PML-RARA regulated adaptor molecule 1
ENSG00000138078	0.68	1.50E-02	PREPL	prolyl endopeptidase-like
ENSG00000142875	1.257	8.40E-04	PRKACB	protein kinase, cAMP-dependent, catalytic, beta
ENSG00000005249	0.937	5.17E-03	PRKAR2B	protein kinase, cAMP-dependent, regulatory, type II, beta
ENSG00000137492	0.778	5.80E-03	PRKRIR	protein-kinase, interferon-inducible double stranded RNA dependent inhibitor, repressor of (P58 repressor)
ENSG00000185246	0.725	2.81E-02	PRPF39	PRP39 pre-mRNA processing factor 39 homolog (S. cerevisiae)
ENSG00000204469	-0.444	4.03E-02	PRRC2A	proline-rich coiled-coil 2A
ENSG00000125637	-0.45	2.54E-02	PSD4	pleckstrin and Sec7 domain containing 4
ENSG00000164985	0.53	1.16E-02	PSIP1	PC4 and SFRS1 interacting protein 1
ENSG00000013275	-0.436	4.03E-02	PSMC4	proteasome (prosome, macropain) 26S subunit, ATPase, 4
ENSG00000226752	0.669	3.93E-02	PSMD5-AS1	PSMD5 antisense RNA 1 (head to head)
ENSG00000107317	-1.026	3.32E-02	PTGDS	prostaglandin D2 synthase 21kDa (brain)
ENSG00000112655	1.041	2.23E-03	PTK7	protein tyrosine kinase 7
ENSG00000134242	0.53	4.51E-02	PTPN22	protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
ENSG00000206418	0.554	3.50E-02	RAB12	RAB12, member RAS oncogene family
ENSG00000109113	-0.529	1.89E-02	RAB34	RAB34, member RAS oncogene family
ENSG00000127314	0.59	7.22E-03	RAP1B	RAP1B, member of RAS oncogene family
ENSG00000123728	0.692	2.34E-03	RAP2C	RAP2C, member of RAS oncogene family
ENSG00000155903	0.838	1.37E-02	RASA2	RAS p21 protein activator 2
ENSG00000126254	-0.531	2.41E-02	RBM42	RNA binding motif protein 42
ENSG00000168476	-0.417	4.49E-02	REEP4	receptor accessory protein 4
ENSG00000135002	0.669	2.12E-02	RFK	riboflavin kinase
ENSG00000242732	-1.013	3.59E-03	RGAG4	retrotransposon gag domain containing 4
ENSG00000102760	0.536	6.67E-03	RGCC	regulator of cell cycle
ENSG00000150681	0.879	1.24E-03	RGS18	regulator of G-protein signaling 18
ENSG00000155366	-0.56	4.51E-02	RHOC	ras homolog family member C
ENSG00000176406	-0.759	4.16E-02	RIMS2	regulating synaptic membrane exocytosis 2
ENSG00000100599	-0.523	3.54E-02	RIN3	Ras and Rab interactor 3
ENSG00000258486	-1.184	2.47E-03	RN7SL1	RNA, 7SL, cytoplasmic 1
ENSG00000202054	-0.994	8.18E-03	RNA5SP152	RNA, 5S ribosomal pseudogene 152
ENSG00000223003	1.032	3.58E-03	RNA5SP184	RNA, 5S ribosomal pseudogene 184
ENSG00000199545	0.885	1.88E-03	RNA5SP195	RNA, 5S ribosomal pseudogene 195
ENSG00000222383	0.591	1.16E-02	RNA5SP203	RNA, 5S ribosomal pseudogene 203
ENSG00000251920	0.633	3.58E-02	RNA5SP216	RNA, 5S ribosomal pseudogene 216
ENSG00000123091	0.679	5.28E-04	RNF11	ring finger protein 11

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000134758	0.674	9.93E-03	RNF138	ring finger protein 138, E3 ubiquitin protein ligase
ENSG00000023191	-0.501	4.54E-02	RNH1	ribonucleaseangiogenin inhibitor 1
ENSG00000206596	1.283	4.05E-03	RNU1-27P	RNA, U1 small nuclear 27, pseudogene
ENSG00000207513	1.146	2.45E-03	RNU1-3	RNA, U1 small nuclear 3
ENSG00000202237	0.609	1.07E-02	RNU6-53P	RNA, U6 small nuclear 53, pseudogene
ENSG00000206737	0.939	1.16E-02	RNVU1-18	RNA, variant U1 small nuclear 18
ENSG00000169855	1.209	4.73E-02	ROBO1	roundabout, axon guidance receptor, homolog 1 (Drosophila)
ENSG00000236552	0.756	2.18E-02	RPL13AP5	ribosomal protein L13a pseudogene 5
ENSG00000122026	1.224	1.44E-03	RPL21	ribosomal protein L21
ENSG00000243742	-0.998	5.10E-05	RPLP0P2	ribosomal protein, large, P0 pseudogene 2
ENSG00000217527	0.632	1.15E-02	RPS16P5	ribosomal protein S16 pseudogene 5
ENSG00000126458	-0.532	2.68E-02	RRAS	related RAS viral (r-ras) oncogene homolog
ENSG00000125844	-0.533	1.91E-02	RRBP1	ribosome binding protein 1
ENSG00000182010	1.047	4.97E-02	RTKN2	rhotekin 2
ENSG00000196154	-0.585	3.32E-02	S100A4	S100 calcium binding protein A4
ENSG00000160307	1.33	1.21E-03	S100B	S100 calcium binding protein B
ENSG00000211456	0.572	3.25E-02	SACM1L	SAC1 suppressor of actin mutations 1-like (yeast)
ENSG00000205413	0.563	4.37E-02	SAMD9	sterile alpha motif domain containing 9
ENSG00000155307	0.492	3.66E-02	SAMSN1	SAM domain, SH3 domain and nuclear localization signals 1
ENSG00000126461	-0.495	3.03E-02	SCAF1	SR-related CTD-associated factor 1
ENSG00000085365	0.664	1.72E-02	SCAMP1	secretory carrier membrane protein 1
ENSG00000047634	1.367	5.31E-03	SCML1	sex comb on midleg-like 1 (Drosophila)
ENSG00000153130	1.214	1.88E-03	SCOC	short coiled-coil protein
ENSG00000124145	-0.87	3.33E-02	SDC4	syndecan 4
ENSG00000168497	0.645	1.52E-02	SDPR	serum deprivation response
ENSG00000138468	0.785	3.46E-02	SENP7	SUMO1sentrin specific peptidase 7
ENSG00000122545	0.378	4.35E-02	40062	septin 7
ENSG00000111897	0.715	4.05E-03	SERINC1	serine incorporator 1
ENSG00000120742	0.614	1.29E-02	SERP1	stress-associated endoplasmic reticulum protein 1
ENSG00000080546	0.779	1.73E-03	SESN1	sestrin 1
ENSG00000130766	-0.498	2.61E-02	SESN2	sestrin 2
ENSG00000115524	0.618	7.26E-03	SF3B1	splicing factor 3b, subunit 1, 155kDa
ENSG00000143368	-0.492	4.16E-02	SF3B4	splicing factor 3b, subunit 4, 49kDa
ENSG00000126821	0.732	3.44E-02	SGPP1	sphingosine-1-phosphate phosphatase 1
ENSG00000183918	0.742	3.98E-02	SH2D1A	SH2 domain containing 1A
ENSG00000095370	-0.433	4.86E-02	SH2D3C	SH2 domain containing 3C
ENSG00000100092	-0.463	4.19E-02	SH3BP1	SH3-domain binding protein 1
ENSG00000035115	0.886	6.51E-04	SH3YL1	SH3 domain containing, Ysc84-like 1 (S. cerevisiae)
ENSG00000185634	1.203	1.17E-02	SHC4	SHC (Src homology 2 domain containing) family, member 4
ENSG00000108061	0.555	1.77E-02	SHOC2	soc-2 suppressor of clear homolog (C. elegans)
ENSG00000170190	-0.4	3.38E-02	SLC16A5	solute carrier family 16, member 5 (monocarboxylic acid transporter 6)
ENSG00000162383	-1.042	4.03E-02	SLC1A7	solute carrier family 1 (glutamate transporter), member 7
ENSG00000125434	-0.541	3.79E-02	SLC25A35	solute carrier family 25, member 35
ENSG00000164209	0.723	1.30E-02	SLC25A46	solute carrier family 25, member 46
ENSG00000143554	-0.486	3.57E-02	SLC27A3	solute carrier family 27 (fatty acid transporter), member 3
ENSG00000160326	-0.536	1.95E-02	SLC2A6	solute carrier family 2 (facilitated glucose transporter), member 6
ENSG00000164414	0.663	2.99E-02	SLC35A1	solute carrier family 35 (CMP-sialic acid transporter), member A1
ENSG00000124786	0.552	2.40E-02	SLC35B3	solute carrier family 35, member B3
ENSG00000157800	0.66	1.10E-02	SLC37A3	solute carrier family 37 (glycerol-3-phosphate transporter), member 3
ENSG00000134294	0.932	1.29E-03	SLC38A2	solute carrier family 38, member 2
ENSG00000144290	1.348	1.49E-02	SLC4A10	solute carrier family 4, sodium bicarbonate transporter, member 10
ENSG00000188725	0.556	2.47E-02	SMIM15	small integral membrane protein 15
ENSG00000172062	-1.85	1.90E-15	SMN1/SMN2	survival of motor neuron 1, telomeric

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000205571	0.874	1.31E-04	SMN1/SMN2	survival of motor neuron 1, telomeric
ENSG00000207051	1.224	1.44E-03	SNORA27	small nucleolar RNA, HACA box 27
ENSG00000200320	0.575	9.15E-03	SNORA63	small nucleolar RNA, HACA box 63
ENSG00000207523	0.707	2.67E-02	SNORA66	small nucleolar RNA, HACA box 66
ENSG00000206650	0.749	9.56E-04	SNORA70G	small nucleolar RNA, HACA box 70G
ENSG00000221420	0.575	9.15E-03	SNORA81	small nucleolar RNA, HACA box 81
ENSG00000207500	1.224	1.44E-03	SNORD102	small nucleolar RNA, CD box 102
ENSG00000251806	-0.525	2.28E-02	SNORD119	small nucleolar RNA, CD box 119
ENSG00000201784	0.77	2.84E-03	SNORD14A	small nucleolar RNA, CD box 14A
ENSG00000238942	0.575	9.15E-03	SNORD2	small nucleolar RNA, CD box 2
ENSG00000234617	0.395	4.61E-02	SNRK-AS1	SNRK antisense RNA 1
ENSG00000100028	-0.457	2.58E-02	SNRPD3	small nuclear ribonucleoprotein D3 polypeptide 18kDa
ENSG00000172845	0.667	2.25E-02	SP3	Sp3 transcription factor
ENSG00000021574	0.602	2.67E-02	SPAST	spastin
ENSG00000163806	0.733	2.31E-03	SPDYA	speedyRINGO cell cycle regulator family member A
ENSG00000166145	-0.502	4.55E-02	SPINT1	serine peptidase inhibitor, Kunitz type 1
ENSG00000174780	0.437	2.99E-02	SRP72	signal recognition particle 72kDa
ENSG00000143742	0.8	8.08E-04	SRP9	signal recognition particle 9kDa
ENSG00000116754	0.612	1.28E-02	SRSF11	serinearginine-rich splicing factor 11
ENSG00000124193	0.484	9.93E-03	SRSF6	serinearginine-rich splicing factor 6
ENSG00000073849	0.536	1.49E-02	ST6GAL1	ST6 beta-galactosamide alpha-2,6-sialyltranferase 1
ENSG00000101972	0.668	8.18E-03	STAG2	stromal antigen 2
ENSG00000081320	0.524	1.84E-02	STK17B	serinethreonine kinase 17b
ENSG00000079950	0.559	3.52E-02	STX7	syntaxin 7
ENSG00000173597	0.691	8.54E-03	SULT1B1	sulfotransferase family, cytosolic, 1B, member 1
ENSG00000166317	-0.785	8.06E-03	SYNPO2L	synaptopodin 2-like
ENSG00000132718	-0.529	3.29E-02	SYT11	synaptotagmin XI
ENSG00000178913	0.607	1.52E-02	TAF7	TAF7 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 55kDa
ENSG00000149591	-0.581	1.89E-02	TAGLN	transgelin
ENSG00000121749	0.709	1.08E-02	TBC1D15	TBC1 domain family, member 15
ENSG00000036054	0.649	1.28E-02	TBC1D23	TBC1 domain family, member 23
ENSG00000136111	0.887	8.54E-03	TBC1D4	TBC1 domain family, member 4
ENSG00000136535	1.348	1.49E-02	TBR1	T-box, brain, 1
ENSG00000165929	0.931	2.18E-03	TC2N	tandem C2 domains, nuclear
ENSG00000180964	0.581	1.29E-02	TCEAL8	transcription elongation factor A (SII)-like 8
ENSG00000139372	0.678	1.27E-02	TDG	thymine-DNA glycosylase
ENSG00000108064	0.459	4.66E-02	TFAM	transcription factor A, mitochondrial
ENSG00000131931	0.526	3.06E-02	THAP1	THAP domain containing, apoptosis associated protein 1
ENSG00000177683	0.673	1.48E-02	THAP5	THAP domain containing 5
ENSG00000172673	0.588	4.50E-02	THEMIS	thymocyte selection associated
ENSG00000066056	-0.796	2.01E-02	TIE1	tyrosine kinase with immunoglobulin-like and EGF-like domains 1
ENSG00000100575	0.515	2.45E-03	TIMM9	translocase of inner mitochondrial membrane 9 homolog (yeast)
ENSG00000223573	-1.022	1.37E-02	TINCR	tissue differentiation-inducing non-protein coding RNA
ENSG00000198586	0.599	1.51E-02	TLK1	tousled-like kinase 1
ENSG00000137076	-0.619	3.07E-02	TLN1	talin 1
ENSG00000174123	0.924	1.17E-02	TLR10	toll-like receptor 10
ENSG00000183160	-1.169	5.59E-03	TMEM119	transmembrane protein 119
ENSG00000152558	0.882	5.28E-04	TMEM123	transmembrane protein 123
ENSG00000233493	0.569	2.30E-02	TMEM238	transmembrane protein 238
ENSG00000135185	0.665	2.44E-03	TMEM243	transmembrane protein 243, mitochondrial
ENSG00000112697	0.604	3.06E-02	TMEM30A	transmembrane protein 30A
ENSG00000175147	-1.165	1.27E-03	TMEM51-AS1	TMEM51 antisense RNA 1

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000139921	0.719	1.61E-02	TMX1	thioredoxin-related transmembrane protein 1
ENSG00000145779	0.728	3.58E-03	TNFAIP8	tumor necrosis factor, alpha-induced protein 8
ENSG00000100284	-0.449	3.70E-02	TOM1	target of myb1 (chicken)
ENSG00000154174	0.441	3.31E-02	TOMM70A	translocase of outer mitochondrial membrane 70 homolog A (S. cerevisiae)
ENSG00000164938	0.952	2.26E-02	TP53INP1	tumor protein p53 inducible nuclear protein 1
ENSG00000159713	-0.821	2.30E-02	TPPP3	tubulin polymerization-promoting protein family member 3
ENSG00000067167	0.454	2.92E-02	TRAM1	translocation associated membrane protein 1
ENSG00000163519	1.219	6.51E-04	TRAT1	T cell receptor associated transmembrane adaptor 1
ENSG00000170855	0.729	9.53E-03	TRIAP1	TP53 regulated inhibitor of apoptosis 1
ENSG00000071575	0.531	3.46E-02	TRIB2	tribbles homolog 2 (Drosophila)
ENSG00000116918	0.813	7.34E-03	TSNAX	translin-associated factor X
ENSG00000106537	0.816	1.03E-02	TSPAN13	tetraspanin 13
ENSG00000155158	0.745	4.30E-02	TTC39B	tetratricopeptide repeat domain 39B
ENSG00000171811	1.777	9.98E-04	TTC40	tetratricopeptide repeat domain 40
ENSG00000188229	-0.438	3.29E-02	TUBB4B	tubulin, beta 4B class IVb
ENSG00000171928	0.937	9.15E-03	TVP23B	trans-golgi network vesicle protein 23 homolog B (S. cerevisiae)
ENSG00000151239	0.925	7.43E-03	TWF1	twinfilin actin-binding protein 1
ENSG00000126261	0.509	6.92E-03	UBA2	ubiquitin-like modifier activating enzyme 2
ENSG00000177889	0.519	7.76E-03	UBE2N	ubiquitin-conjugating enzyme E2N
ENSG00000140367	0.53	3.70E-02	UBE2Q2	ubiquitin-conjugating enzyme E2Q family member 2
ENSG00000114062	0.642	2.32E-02	UBE3A	ubiquitin protein ligase E3A
ENSG00000122042	0.627	8.52E-03	UBL3	ubiquitin-like 3
ENSG00000148154	0.774	1.11E-02	UGCG	UDP-glucose ceramide glucosyltransferase
ENSG00000260128	-3.845	7.34E-03	ULK4P2	unc-51-like kinase 4 (C. elegans) pseudogene 2
ENSG00000105176	0.548	4.05E-03	URI1	URI1, prefoldin-like chaperone
ENSG00000162607	0.655	1.53E-02	USP1	ubiquitin specific peptidase 1
ENSG00000232162	1.68	4.39E-05	USP12-AS1	USP12 antisense RNA 1
ENSG00000135655	0.599	2.40E-02	USP15	ubiquitin specific peptidase 15
ENSG00000124333	0.606	1.25E-02	VAMP7	vesicle-associated membrane protein 7
ENSG00000182853	-1.401	5.61E-03	VMO1	vitelline membrane outer layer 1 homolog (chicken)
ENSG00000236287	0.694	1.72E-02	ZBED5	zinc finger, BED-type containing 5
ENSG00000126804	0.661	2.95E-02	ZBTB1	zinc finger and BTB domain containing 1
ENSG00000066422	0.614	2.47E-02	ZBTB11	zinc finger and BTB domain containing 11
ENSG00000065548	0.585	4.30E-02	ZC3H15	zinc finger CCCH-type containing 15
ENSG00000177764	0.555	2.78E-02	ZCCHC3	zinc finger, CCHC domain containing 3
ENSG00000159714	-0.735	2.97E-03	ZDHHC1	zinc finger, DHHC-type containing 1
ENSG00000180776	0.921	8.14E-03	ZDHHC20	zinc finger, DHHC-type containing 20
ENSG00000236953	1.098	6.55E-04	ZDHHC20-IT1	ZDHHC20 intronic transcript 1
ENSG00000104231	0.729	2.66E-02	ZFAND1	zinc finger, AN1-type domain 1
ENSG00000197841	1.425	4.73E-03	ZNF181	zinc finger protein 181
ENSG00000165512	0.54	4.60E-02	ZNF22	zinc finger protein 22
ENSG00000185947	0.84	1.31E-02	ZNF267	zinc finger protein 267
ENSG00000089335	1.044	1.53E-02	ZNF302	zinc finger protein 302
ENSG00000169131	0.708	9.14E-03	ZNF354A	zinc finger protein 354A
ENSG00000171469	0.597	1.88E-02	ZNF561	zinc finger protein 561
ENSG00000198566	-7.734	9.03E-10	ZNF658B	zinc finger protein 658B, pseudogene
ENSG00000166770	0.903	6.64E-03	ZNF667-AS1	ZNF667 antisense RNA 1 (head to head)
ENSG00000182141	1.034	2.13E-02	ZNF708	zinc finger protein 708
ENSG00000196381	1.217	4.70E-02	ZNF781	zinc finger protein 781
ENSG00000167766	0.87	1.30E-02	ZNF83	zinc finger protein 83
ENSG00000146757	1.129	2.22E-02	ZNF92	zinc finger protein 92
ENSG00000132485	0.774	9.38E-03	ZRANB2	zinc finger, RAN-binding domain containing 2
ENSG00000198618	-2.235	3.40E-08		

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000240869	-1.494	2.23E-05		
ENSG00000251948	-1.065	4.39E-05		
ENSG00000212769	-2.34	1.34E-05		
ENSG00000238179	-0.872	5.40E-04		
ENSG00000252488	1.302	2.54E-05		
ENSG00000261655	-0.897	1.28E-03		
ENSG00000226937	0.814	2.44E-03		
ENSG00000240490	-0.74	2.75E-03		
ENSG00000247315	0.555	2.78E-02		
ENSG00000243302	-0.555	5.59E-03		
ENSG00000223551	-1.045	2.31E-03		
ENSG00000258800	0.754	1.85E-02		
ENSG00000258988	-2.49	5.28E-04		
ENSG00000253676	1.57	4.22E-05		
ENSG00000260808	0.612	1.54E-02		
ENSG00000228956	0.904	5.88E-03		
ENSG00000231006	1.108	4.98E-04		
ENSG00000235872	0.688	2.68E-02		
ENSG00000211619	-7.103	6.72E-04		
ENSG00000259499	1.092	5.28E-04		
ENSG00000240223	0.71	5.28E-04		
ENSG00000224827	-1.101	2.16E-03		
ENSG00000226976	-1.039	2.96E-03		
ENSG00000262657	0.84	1.31E-02		
ENSG00000201592	0.901	6.51E-04		
ENSG00000201882	0.901	6.51E-04		
ENSG00000228323	-0.674	8.18E-03		
ENSG00000201820	1.012	8.40E-04		
ENSG00000202434	0.821	6.51E-04		
ENSG00000240163	-0.738	1.37E-02		
ENSG00000239087	1.322	1.02E-03		
ENSG00000235945	0.939	1.18E-03		
ENSG00000201121	1.156	6.51E-04		
ENSG00000227907	1.329	8.08E-04		
ENSG00000252464	0.925	5.59E-03		
ENSG00000234040	0.747	1.07E-03		
ENSG00000228981	1.156	3.11E-02		
ENSG00000230272	0.781	6.51E-04		
ENSG00000230408	0.671	5.94E-04		
ENSG00000239935	-0.703	2.28E-02		
ENSG00000214659	-1.615	1.00E-02		
ENSG00000200418	0.575	9.15E-03		
ENSG00000234925	0.738	1.21E-03		
ENSG00000238982	0.959	8.08E-04		
ENSG00000262211	1.143	9.53E-03		
ENSG00000228386	-0.498	2.92E-02		
ENSG00000236434	0.773	8.08E-04		
ENSG00000251805	1.064	1.14E-03		
ENSG00000252197	1.088	2.22E-03		
ENSG00000259595	0.778	3.18E-03		
ENSG00000238975	0.742	2.84E-03		
ENSG00000238924	0.9	1.82E-03		
ENSG00000230715	-0.417	3.75E-02		

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000252904	1.447	1.71E-03		
ENSG00000226084	1.06	2.49E-03		
ENSG00000205147	-0.985	5.59E-03		
ENSG00000238556	0.553	2.88E-03		
ENSG00000222494	0.902	1.19E-03		
ENSG00000252700	0.87	1.79E-03		
ENSG00000240342	1.039	6.30E-03		
ENSG00000197462	-1.019	9.66E-03		
ENSG00000259950	1.141	2.88E-03		
ENSG00000234374	1.07	8.08E-04		
ENSG00000241438	0.994	2.85E-03		
ENSG00000228265	-0.489	2.12E-02		
ENSG00000233406	0.741	1.18E-03		
ENSG00000251621	1.348	1.49E-02		
ENSG00000223697	1.173	3.03E-03		
ENSG00000212829	-1.368	1.52E-03		
ENSG00000252620	0.606	1.15E-03		
ENSG00000200397	0.695	2.23E-03		
ENSG00000227146	0.587	1.08E-02		
ENSG00000253754	0.817	7.16E-03		
ENSG00000258663	-1.863	8.40E-03		
ENSG00000231245	1.27	2.64E-02		
ENSG00000259657	0.75	2.45E-03		
ENSG00000243005	0.643	1.76E-03		
ENSG00000251783	0.875	2.23E-03		
ENSG00000254208	1.231	2.64E-03		
ENSG00000249055	0.657	5.61E-03		
ENSG00000238449	0.657	5.61E-03		
ENSG00000207491	-0.82	1.32E-02		
ENSG00000251892	0.993	3.17E-03		
ENSG00000206713	-0.617	1.49E-02		
ENSG00000200788	-0.589	2.45E-02		
ENSG00000241376	-1.03	9.17E-03		
ENSG00000252750	0.7	5.61E-03		
ENSG00000199866	1.307	2.31E-03		
ENSG00000248785	0.557	2.88E-03		
ENSG00000235299	0.691	6.30E-03		
ENSG00000213147	-5.489	1.62E-02		
ENSG00000234106	-0.617	2.81E-02		
ENSG00000225300	1.12	8.40E-04		
ENSG00000262429	-0.461	1.89E-02		
ENSG00000244498	-0.599	4.64E-02		
ENSG00000238450	0.694	2.55E-03		
ENSG00000225627	0.755	3.18E-03		
ENSG00000255301	-1.408	4.72E-02		
ENSG00000259421	-0.61	2.05E-02		
ENSG00000260977	-0.869	7.61E-03		
ENSG00000244479	0.853	8.54E-03		
ENSG00000212259	1.108	5.59E-03		
ENSG00000253574	-0.693	2.26E-02		
ENSG00000223505	0.71	2.88E-03		
ENSG00000231845	0.802	5.59E-03		
ENSG00000243305	0.924	9.51E-03		

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000237336	-0.619	2.62E-02		
ENSG00000258645	1.789	7.05E-03		
ENSG00000228187	1.024	4.05E-03		
ENSG00000233728	-0.867	9.66E-03		
ENSG00000202490	1.777	7.29E-03		
ENSG00000206822	1.777	7.29E-03		
ENSG00000222086	0.717	4.66E-03		
ENSG00000250473	1.518	5.61E-03		
ENSG00000259099	-0.628	3.30E-02		
ENSG00000213744	-0.588	3.31E-02		
ENSG00000203386	0.836	5.50E-03		
ENSG00000201071	-0.433	3.55E-02		
ENSG00000239593	-0.648	4.73E-02		
ENSG00000250850	0.618	1.28E-02		
ENSG00000254193	3.59	1.54E-02		
ENSG00000235117	-0.508	1.51E-02		
ENSG00000248839	0.813	8.54E-03		
ENSG00000200120	1.342	1.13E-02		
ENSG00000211642	0.966	3.52E-02		
ENSG00000227417	0.796	2.88E-03		
ENSG00000250733	-0.687	2.54E-02		
ENSG00000238959	0.886	5.97E-03		
ENSG00000261744	-0.777	2.47E-02		
ENSG00000241300	1.387	8.54E-03		
ENSG00000239194	-0.821	2.30E-02		
ENSG00000218980	-0.591	3.59E-02		
ENSG00000228366	1.065	3.18E-03		
ENSG00000185275	0.539	4.44E-02		
ENSG00000261691	-0.497	3.93E-02		
ENSGR0000124333	0.575	2.51E-02		
ENSG00000240925	-0.452	4.90E-02		
ENSG00000220842	4.8	2.47E-02		
ENSG00000222869	-0.575	3.40E-02		
ENSG00000252759	1.164	1.37E-02		
ENSG00000253772	-1.709	1.72E-02		
ENSG00000238766	-0.534	4.23E-02		
ENSG00000242926	-0.534	4.23E-02		
ENSG00000243736	0.569	2.68E-02		
ENSG00000229750	0.974	1.18E-02		
ENSG00000259781	2.481	1.32E-02		
ENSG00000224895	-0.807	2.80E-02		
ENSG00000233984	1.405	7.77E-03		
ENSG00000234219	0.719	7.34E-03		
ENSG00000239344	-0.759	4.16E-02		
ENSG00000227239	0.618	1.18E-02		
ENSG00000228770	-0.587	3.75E-02		
ENSG00000229679	-0.584	3.83E-02		
ENSG00000232486	0.912	5.61E-03		
ENSG00000255959	-0.471	4.17E-02		
ENSG00000207034	0.628	9.51E-03		
ENSG00000244275	0.538	1.08E-02		
ENSG00000240369	0.538	1.08E-02		
ENSG00000238916	0.538	1.08E-02		

Continued on next page

Table B.5 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000254274	0.832	9.78E-03		
ENSG00000225792	0.602	1.31E-02		
ENSG00000199591	-0.792	3.09E-02		
ENSG00000188223	-0.433	3.63E-02		
ENSG00000200113	-0.591	4.03E-02		
ENSG00000256403	-1.084	2.54E-02		
ENSG00000230625	0.767	1.06E-02		
ENSG00000217416	1.058	3.80E-02		
ENSG00000242681	0.985	1.09E-02		
ENSG00000255729	0.607	1.52E-02		
ENSG00000199415	1.358	1.73E-02		
ENSG00000237672	0.788	1.09E-02		
ENSG00000238268	-1.231	1.18E-02		
ENSG00000227161	0.699	1.25E-02		
ENSG00000222276	0.935	1.27E-02		
ENSG00000239917	0.605	7.58E-03		
ENSG00000227484	-0.952	2.20E-02		
ENSG00000201778	0.575	1.25E-02		
ENSG00000244389	1.155	1.29E-02		
ENSG00000255165	-0.751	4.64E-02		
ENSG00000261327	-0.755	2.61E-02		
ENSG00000259811	-0.755	2.61E-02		
ENSG00000231995	-4.009	1.54E-02		
ENSG00000239975	1.661	3.28E-02		
ENSG00000254671	0.678	7.34E-03		
ENSG00000213881	0.941	1.81E-02		
ENSG00000253636	-1.039	9.99E-03		
ENSG00000250267	1.262	1.54E-02		
ENSG00000228554	0.58	1.04E-02		
ENSG00000229413	-0.611	3.32E-02		
ENSG00000255671	-0.923	2.17E-02		
ENSG00000253092	1.603	2.43E-03		
ENSG00000261573	0.565	1.64E-02		
ENSG00000239932	-0.719	3.57E-02		
ENSG00000248636	-0.546	2.92E-02		
ENSG00000235883	0.471	3.44E-02		
ENSG00000201013	0.756	7.05E-03		
ENSG00000262777	0.808	1.22E-02		
ENSG00000200814	0.656	1.30E-02		
ENSG00000248977	1.716	1.61E-02		
ENSG00000259884	-0.671	3.46E-02		
ENSG00000231128	0.657	7.29E-03		
ENSG00000241247	0.572	3.51E-02		

SMA2 *vs.* SMA3

Table B.6: Features of the genes differentially expressed in “SMA2 vs SMA3” comparison considered for network analysis: Ensemble gene ID, log-ratio and q-value. For genes annotated in the IPA database, gene symbol and Entrez name are also reported.

Gene	log-ratio	q-value	Symbol	Entrez Name
ENSG00000249310	-4.171	8.66E-12	APOBEC3B-AS1	APOBEC3B antisense RNA 1
ENSG00000118520	-1.789	5.46E-04	ARG1	arginase, liver
ENSG00000172232	-1.784	4.62E-04	AZU1	azurocidin 1
ENSG00000101425	-1.882	1.04E-04	BPI	bactericidal/permeability-increasing protein
ENSG00000164047	-2.206	2.85E-07	CAMP	cathelicidin antimicrobial peptide
ENSG00000256515	9.267	4.44E-07	CCL3L1/CCL3L3	chemokine (C-C motif) ligand 3-like 1
ENSG00000086548	-2.273	3.50E-04	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
ENSG00000124469	-2.736	2.85E-07	CEACAM8	carcinoembryonic antigen-related cell adhesion molecule 8
ENSG00000228695	-8.261	1.46E-06	CES1P1	carboxylesterase 1 pseudogene 1
ENSG00000133063	-2.739	2.93E-06	CHIT1	chitinase 1 (chitotriosidase)
ENSG00000065618	-2.682	8.04E-04	COL17A1	collagen, type XVII, alpha 1
ENSG00000096006	-3.004	2.90E-08	CRISP3	cysteine-rich secretory protein 3
ENSG00000100448	-1.85	2.84E-03	CTSG	cathepsin G
ENSG00000164821	-2.95	3.19E-12	DEFA4	defensin, alpha 4, corticostatin
ENSG00000079393	-2.727	3.40E-02	DUSP13	dual specificity phosphatase 13
ENSG00000197561	-1.969	1.15E-04	ELANE	elastase, neutrophil expressed
ENSG00000170801	-2.819	8.49E-03	HTRA3	HtrA serine peptidase 3
ENSG00000231475	1.366	2.79E-02	IGHV4-31	immunoglobulin heavy variable 4-31
ENSG00000211598	2.076	3.03E-03	IGKV4-1	immunoglobulin kappa variable 4-1
ENSG00000134545	1.575	2.94E-02	KLRC1	killer cell lectin-like receptor subfamily C, member 1
ENSG00000205809	1.575	2.94E-02	KLRC2	killer cell lectin-like receptor subfamily C, member 2
ENSG00000255641	1.575	2.94E-02	KLRC3	killer cell lectin-like receptor subfamily C, member 3
ENSG00000205810	1.575	2.94E-02	KLRC3	killer cell lectin-like receptor subfamily C, member 3
ENSG00000148346	-3.108	2.14E-10	LCN2	lipocalin 2
ENSG00000012223	-3.332	3.09E-07	LTF	lactotransferrin
ENSG00000216083	-2.682	8.04E-04	mir-936	microRNA 936
ENSG00000118113	-3.327	2.87E-07	MMP8	matrix metallopeptidase 8 (neutrophil collagenase)
ENSG00000149516	-1.675	3.04E-04	MS4A3	membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)
ENSG00000102837	-2.583	2.81E-04	OLFM4	olfactomedin 4
ENSG00000173391	-2.515	9.69E-04	OLR1	oxidized low density lipoprotein (lectin-like) receptor 1
ENSG00000196415	-1.815	1.84E-02	PRTN3	proteinase 3
ENSG00000104918	-1.462	1.36E-02	RETN	resistin
ENSG00000169397	-2.231	8.87E-06	RNASE3	ribonuclease, RNase A family, 3
ENSG00000134827	-1.963	6.81E-06	TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)
ENSG00000188056	3.96	1.15E-04	TREML4	triggering receptor expressed on myeloid cells-like 4
ENSG00000211637	2.106	1.01E-08		
ENSG00000235508	8.885	7.50E-10		
ENSG00000197149	4.96	3.60E-07		
ENSG00000223350	1.971	9.55E-05		
ENSG00000240342	2.718	4.37E-06		
ENSG00000211650	2.609	1.15E-04		
ENSG00000213147	7.532	7.00E-03		
ENSG00000181126	-4.747	1.17E-04		
ENSG00000206249	-1.733	2.18E-03		
ENSG00000242580	2.485	2.84E-03		

Continued on next page

Table B.6 – *Continued from previous page*

Gene	Log-ratio	q-value	Symbol	Entrez Name
ENSG00000236650	-4.198	1.32E-03		
ENSG00000250765	-2.67	8.04E-04		
ENSG00000211938	2.11	3.03E-03		
ENSG00000239862	1.681	2.42E-02		
ENSG00000231896	-1.914	7.97E-04		
ENSG00000253239	1.804	3.65E-02		
ENSG00000211663	1.541	7.33E-03		
ENSG00000218749	-1.176	1.21E-02		
ENSG00000243166	-1.16	1.69E-02		
ENSG00000211654	2.407	5.78E-03		
ENSG00000211665	2.069	3.87E-03		
ENSG00000253497	3.676	1.18E-02		
ENSG00000244575	1.116	4.30E-02		
ENSG00000212579	-1.357	3.33E-02		

C
Full list of publications

C.1 Journal papers

J. Fadista, P. Vikman, E. Ottosson Laakso, I. Mllet, P. Osmark, J. Esguerra, J. Taneera, C. Ladenvall, K. Hansson, F. Finotello, U. Krus, B. Di Camillo, O. Hansson, L. Eliasson, A. Rosengren, E. Renström, C. B. Wollheim, L. Groop. *Global transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism*. Submitted.

F. Finotello, E. Lavezzo, L. Bianco, L. Barzon, P. Mazzon, P. Fontana, S. Toppo, and B. Di Camillo. *Reducing bias in RNA sequencing data: a novel approach to compute counts*. BMC Bioinformatics Suppl 1 (2014): S7.

E. Lavezzo, S. Toppo, E. Franchin, B. Di Camillo, F. Finotello, M. Falda, R. Manganelli, G. Palù, and L. Barzon. *Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitidis outbreak*. BMC Infectious Diseases 13(1):554, 2013. Highly accessed.

F. Finotello, E. Lavezzo, P. Fontana, D. Peruzzo, A. Albiero, L. Barzon, M. Falda, B. Di Camillo, and S. Toppo. *Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data*. Briefings in Bioinformatics, 13(3):269-280, 2012.

E. Lavezzo, S. Toppo, L. Barzon, C. Cobelli, B. Di Camillo, F. Finotello, E. Franchin, D. Peruzzo, G.M. Toffolo, M. Trevisan, et al. *Draft genome sequences of two Neisseria meningitidis serogroup C clinical isolates*. Journal of bacteriology, 192(19):5270-5271, 2010.

C.2 Abstracts ad short papers

T. Sanavia, F. Finotello, and B. Di Camillo. *FunPat: a function-based pattern analysis pipeline for RNA-seq time-series data*. Accepted for BITS 2014, Eleventh Annual Meeting of the Bioinformatics Italian Society. February 26-28, 2014. Rome, Italy.

F. Finotello, E. Lavezzo, L. Barzon, P. Fontana, S. Toppo, and B. Di Camillo. *Characterization and reduction of biases in RNA sequencing data*. In IDAMAP, Intelligent data analysis in biomedicine and pharmacology, November 22, 2012. Pavia, Italy. Oral Communication.

F. Finotello, E. Lavezzo, L. Barzon, P. Mazzon, P. Fontana, S. Toppo, and B. Di Camillo. *A strategy to reduce technical variability and bias in RNA sequencing data*. EMBnet. journal, 18(B):pp-65, 2012. In NETTAB workshop on Integrated Bio-Search. November 14-16, 2012, Como, Italy. Oral Communication.

F. Finotello, E. Lavezzo, L. Barzon, P. Fontana, A. Si-Ammour , S. Toppo, and B. Di Camillo. *RNA sequencing data: biases and normalization*. EMBnet.journal 18 Suppl. A. In BITS, IX Annual Meeting of the Bioinformatics Italian Society Meeting Abstracts. May 2-4, 2012. Catania, Italy.

F. Finotello, E. Lavezzo, L. Barzon, P. Fontana, A. Si-Ammour, S. Toppo, and B. Di Camillo. *Comparison of parametric methods for detecting differential expression in RNA sequencing data*. In Third National Congress of Italian Group of Bioengineering, June 26-29, 2012. Rome, Italy.

F. Finotello, D. Peruzzo, E. Lavezzo, B. Di Camillo, G. M. Toffolo, C. Cobelli, and S. Toppo. *Complete and comparative analysis of algorithms for whole genome shotgun assembly*. In BITS 2010. VII Annual meeting of the Bioinformatics Italian Society. April 14-16, 2010. Bari, Italy. Oral Communication.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 2002.

[2] S Clancy. Chemical structure of RNA. *Nature Education*, 1(1), 2008.

[3] Benjamin A Pierce. *Genetics: A conceptual approach*. Macmillan, 2010.

[4] Frederick Sanger. Determination of nucleotide sequences in DNA. *Bioscience reports*, 1(1):3–18, 1981.

[5] F Sanger, GM Air, BG Barrell, NL Brown, AR Coulson, CA Fiddes, CA Hutchison, PM Slocombe, and M Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, 1977.

[6] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[7] Elaine R Mardis. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, 2013.

[8] Daniel C Koboldt, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

[9] Martin Kircher and Janet Kelso. High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[10] `http://www.genome.gov/sequencingcosts/`.

[11] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human molecular genetics*, 19(R2):R227–R240, 2010.

[12] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[13] `http://454.com/downloads/GSFLXApplicationFlyer_FINALv2.pdf`.

[14] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

[15] Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, Giles Hall, Terrance P Shea, Sean Sykes, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2011.

[16] Duccio Medini, Davide Serruto, Julian Parkhill, David A Relman, Claudio Donati, Richard Moxon, Stanley Falkow, and Rino Rappuoli. Microbiology in the post-genomic era. *Nature Reviews Microbiology*, 6(6):419–430, 2008.

[17] Eric D Green. Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics*, 2(8):573–583, 2001.

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

[19] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[20] `http://www.454.com`.

[21] Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W Kinzler, and Bert Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, 100(15):8817–8822, 2003.

[22] Mostafa Ronaghi, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242(1):84–89, 1996.

[23] J Adams. DNA sequencing technologies. *Nature Education*, 1(1), 2008.

[24] `http://www.illumina.com`.

[25] Gerardo Turcatti, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4):e25–e25, 2008.

[26] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[27] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):341, 2012.

[28] HTStec Limited. Next generation sequencing trends 2012 report. Technical report, 2012.

[29] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, 2005.

[30] Kevin Mckernan, Alan Blanchard, Lev Kotler, and Gina Costa. Reagents, methods, and libraries for bead-based sequencing, December 2 2009. US Patent App. 12/629,858.

[31] `http://www.lifetechnologies.com`.

[32] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641–658, 2009.

[33] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[34] Jayson Bowers, Judith Mitchell, Eric Beer, Philip R Buzby, Marie Causey, J William Efcavitch, Mirna Jarosz, Edyta Krzymanska-Olejnik, Li Kung, Doron Lipson, et al. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature methods*, 6(8):593–595, 2009.

[35] Timothy D Harris, Phillip R Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, Jennifer Colonell, James DiMeo, J William Efcavitch, et al. Single-molecule DNA sequencing of a viral genome. *Science*, 320(5872):106–109, 2008.

[36] `http://www.pacificbiosciences.com`.

[37] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[38] Michael J Levene, Jonas Korlach, Stephen W Turner, Mathieu Foquet, Harold G Craighead, and Watt W Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686, 2003.

[39] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[40] Mark S Boguski, Carolyn M Tolstoshev, and Douglas E Bassett Jr. Gene discovery in dbEST. *Science*, 265(5181):1993–1994, 1994.

[41] Daniela S Gerhard, Lukas Wagner, Elise A Feingold, Carolyn M Shenmen, Lynette H Grouse, Greg Schuler, Steven L Klein, Susan Old, Rebekah Rasooly, Peter Good, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome research*, 14(10B):2121–2127, 2004.

[42] Paul Bertone, Viktor Stolc, Thomas E Royce, Joel S Rozowsky, Alexander E Urban, Xiaowei Zhu, John L Rinn, Waraporn Tongprasit, Manoj Samanta, Sherman Weissman, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–2246, 2004.

[43] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725):1149–1154, 2005.

[44] Sydney Brenner, Maria Johnson, John Bridgham, George Golda, David H Lloyd, Davida Johnson, Shujun Luo, Sarah McCurdy, Michael Foy, Mark Ewan, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–634, 2000.

[45] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

[46] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[47] Marcel C Van Verk, Richard Hickman, Corne MJ Pieterse, and Saskia Van Wees. RNA-Seq: revelation of the messengers. *Trends in plant science*, 2013.

[48] Jay Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7):585–587, 2008.

[49] Nicole Cloonan, Alistair RR Forrest, Gabriel Kolle, Brooke BA Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–619, 2008.

[50] James Bullard, Elizabeth Purdom, Kasper Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[51] Jacob E Crawford, Wamdaogo M Guelbeogo, Antoine Sanou, Alphonse Traoré, Kenneth D Vernick, N'Fale Sagnon, and Brian P Lazzaro. De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. *PLoS one*, 5(12):e14202, 2010.

[52] J Cristobal Vera, Christopher W Wheat, Howard W Fescemyer, Mikko J Frilander, Douglas L Crawford, Ilkka Hanski, and James H Marden. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular ecology*, 17(7):1636–1647, 2008.

[53] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[54] Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[55] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[56] Zhiyu Peng, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, Yu Liang, Xueda Hu, Xuemei Tan, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology*, 30(3):253–260, 2012.

[57] Jae Hoon Bahn, Jae-Hyung Lee, Gang Li, Christopher Greer, Guangdun Peng, and Xinshu Xiao. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research*, 22(1):142–150, 2012.

[58] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1), 2011.

[59] Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, 10(9):618–630, 2012.

[60] Alicia Oshlack, Mark D Robinson, Matthew D Young, et al. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.

[61] Ian Korf. Genomics: the state of the art in RNA-seq analysis. *Nature methods*, 10(12):1165–1166, 2013.

[62] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6:S6–S12, 2009.

[63] Ayat Hatem, Doruk Bozda, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC bioinformatics*, 14(1):184, 2013.

[64] `http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment`.

[65] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[66] Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[67] `http://www.novocraft.com`.

[68] Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–1067, 2009.

[69] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E Eichler, and S Cenk Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods*, 7(8):576–577, 2010.

[70] Sanchit Misra, Ramanathan Narayanan, Simon Lin, and Alok Choudhary. Fangs: High speed sequence mapping for next generation sequencers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1539–1546. ACM, 2010.

[71] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[72] Andrew D Smith, Zhenyu Xuan, and Michael Q Zhang. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics*, 9(1):128, 2008.

[73] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, 8(6):469–477, 2011.

[74] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[75] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[76] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[77] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[78] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12):1185–1191, 2013.

[79] Bo Li and Colin Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[80] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

[81] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.

[82] Gregory R Grant, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[83] W James Kent. BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[84] Vivien Marx. The author file: Paul Bertone. *Nature Methods*, 10(12):1137–1137, 2013.

[85] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.

[86] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[87] Ernest Turro, Shu-Yi Su, Ângela Gonçalves, LJ Coin, Sylvia Richardson, and Alex Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, 12(2):R13, 2011.

[88] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 2011.

[89] Hsueh-Ting Chu, William WL Hsiao, Theresa TH Tsao, D Frank Hsu, Chaur-Chin Chen, Sheng-An Lee, and Cheng-Yan Kao. SeqEntropy: Genome-Wide Assessment of Repeats for Short Read Sequencing. *PloS one*, 8(3):e59484, 2013.

[90] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90–e90, 2011.

[91] Christian Ledergerber and Christophe Dessimoz. Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics*, 12(5):489–497, 2011.

[92] Susanne Balzer, Ketil Malde, and Inge Jonassen. Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, 27(13):i304–i309, 2011.

[93] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[94] Ryan D Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J Pugh, Helen McDonald, Richard Varhol, Steven JM Jones, and Marco A Marra. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45(1):81, 2008.

[95] Yuan Ji, Yanxun Xu, Qiong Zhang, Kam-Wah Tsui, Yuan Yuan, Clift Norris Jr, Shoudan Liang, and Han Liang. BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data. *Biometrics*, 67(4):1215–1224, 2011.

[96] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[97] Marius Nicolae, Serghei Mangul, Ion I Mandoiu, and Alex Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, 6(1):9, 2011.

[98] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC bioinformatics*, 12(1):480, 2011.

[99] Yuval Benjamini and Terence P Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72, 2012.

[100] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16):e105, 2008.

[101] Jun Li, Hui Jiang, and Wing H Wong. Method Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R25, 2010.

[102] Alicia Oshlack and Matthew J Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):14, 2009.

[103] Yong Lin, Jian Li, Hui Shen, Lei Zhang, Christopher J Papasian, et al. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics*, 27(15):2031–2037, 2011.

[104] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[105] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[106] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[107] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[108] Liyan Gao, Zhide Fang, Kui Zhang, Degui Zhi, and Xiangqin Cui. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, 27(5):662–669, 2011.

[109] Matthew Young, Matthew Wakefield, Gordon Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010.

[110] Francesca Finotello, Enrico Lavezzo, Paolo Fontana, Denis Peruzzo, Alessandro Albiero, Luisa Barzon, Marco Falda, Barbara Di Camillo, and Stefano Toppo. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in bioinformatics*, 13(3):269–280, 2012.

[111] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in RNA-Sequencing data. *BMC bioinformatics*, 12(1):290, 2011.

[112] Kasper D Hansen, Rafael A Irizarry, and WU Zhijin. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 2012.

[113] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131, 2010.

[114] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.

[115] `http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools#Quantitative_analysis_and_Differential_Expression`.

[116] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[117] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[118] Fatemeh Seyednasrollah, Asta Laiho, and Laura L Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*, page bbt086, 2013.

[119] José A Robles, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics*, 13(1):484, 2012.

[120] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9):R95, 2013.

[121] Yanming Di, Daniel W Schafer, Jason S Cumbie, and Jeff H Chang. The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1):24, 2011.

[122] Hui Jiang and Wing H Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

[123] Ann L Oberg, Brian M Bot, Diane E Grill, Gregory A Poland, and Terry M Therneau. Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC genomics*, 13(1):304, 2012.

[124] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 2013.

[125] Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany*, 99(2):248–256, 2012.

[126] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom! Precision weights unlock linear model analysis tools for RNA-seq read counts. *Technical report. Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia*, 2013.

[127] Thomas J Hardcastle and Krystyna A Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422, 2010.

[128] Yan Guo, Chung-I Li, Fei Ye, and Yu Shyr. Evaluation of read count based RNAseq analysis methods. *BMC Genomics*, 14(Suppl 8):S2, 2013.

[129] Kasper D Hansen, Zhijin Wu, Rafael A Irizarry, and Jeffrey T Leek. Sequencing technology does not eliminate biological variability. *Nature biotechnology*, 29(7):572–573, 2011.

[130] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome research*, 22(10):2008–2017, 2012.

[131] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[132] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[133] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564, 2012.

[134] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K Gifford, and Tommi S Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences*, 100(18):10146–10151, 2003.

[135] Barbara Di Camillo, Gianna Toffolo, Sreekumaran Nair, Laura Greenlund, and Claudio Cobelli. Significance analysis of microarray transcript levels in time series experiments. *BMC bioinformatics*, 8(Suppl 1):S10, 2007.

[136] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[137] Martin Aryee, Jose Gutierrez-Pabello, Igor Kramnik, Tapabrata Maiti, and John Quackenbush. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC bioinformatics*, 10(1):409, 2009.

[138] Jeffrey T Leek, Eva Monsen, Alan R Dabney, and John D Storey. EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, 22(4):507–508, 2006.

[139] Marco F Ramoni, Paola Sebastiani, and Isaac S Kohane. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, 99(14):9121–9126, 2002.

[140] Alexander Schliep, Alexander Schönhuth, and Christine Steinhoff. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 19(suppl 1):i255–i263, 2003.

[141] Alexander Schliep, Christine Steinhoff, and Alexander Schönhuth. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*, 20(suppl 1):i283–i289, 2004.

[142] Paolo Magni, Fulvia Ferrazzi, Lucia Sacchi, and Riccardo Bellazzi. TimeClust: a clustering tool for gene expression time series. *Bioinformatics*, 24(3):430–432, 2008.

[143] Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, et al. Alternative expression analysis by RNA sequencing. *Nature methods*, 7(10):843–847, 2010.

[144] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, 2011.

[145] `http://www.ncbi.nlm.nih.gov/sra`.

[146] Alberto Gatto, Carlos Torroja-Fungairiño, Francesco Mazzarotto, Stuart A Cook, Paul JR Barton, Fátima Sánchez-Cabo, and Enrique Lara-Pezzi. FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Submitted*.

[147] `http://www.ncbi.nlm.nih.gov/books/NBK47540/`.

[148] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[149] `http://bioconductor.org/packages/devel/bioc/vignettes/EDASeq/inst/doc/EDASeq.pdf`.

[150] Steffen Durinck, Wolfgang Huber, and Sean Davis. biomaRt: Interface to BioMart databases (eg Ensembl, Wormbase and Gramene). *R packageversion 1.16*.

[151] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

[152] William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[153] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 2012.

[154] Zhaonan Sun and Yu Zhu. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics*, 28(20):2584–2591, 2012.

[155] Francesca Finotello, Enrico Lavezzo, Luca Bianco, Luisa Barzon, Paolo Mazzon, Paolo Fontana, Stefano Toppo, and Barbara Di Camillo. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics*, 15(Suppl 1):S7, 2014.

[156] `http://hannonlab.cshl.edu/fastx_toolkit/`.

[157] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[158] Davide Campagna, Alessandro Albiero, Alessandra Bilardi, Elisa Caniato, Claudio Forcato, Svetlin Manavski, Nicola Vitulo, and Giorgio Valle. PASS: a program to align short sequences. *Bioinformatics*, 25(7):967–968, 2009.

[159] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[160] `http://www.dei.unipd.it/~finotell/maxcounts/`.

[161] Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[162] Timo Lassmann, Yoshihide Hayashizaki, and Carsten O Daub. SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, 27(1):130–131, 2011.

[163] Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 2011.

[164] Kevin Talbot and Kay E Davies. Spinal muscular atrophy. In *Seminars in neurology*, volume 21, pages 189–198. Copyright© 2001 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662, 2001.

[165] B Wirth, L Brichta, B Schrank, H Lochmüller, S Blick, A Baasner, and R Heller. Mildly affected patients with spinal muscular atrophy are partially protected by an increased SMN2 copy number. *Human genetics*, 119(4):422–428, 2006.

[166] JOHN Pearn. Incidence, prevalence, and gene frequency studies of chronic childhood spinal muscular atrophy. *Journal of Medical Genetics*, 15(6):409–413, 1978.

[167] Stephen J Kolb and John T Kissel. Spinal muscular atrophy: a timely review. *Archives of neurology*, 68(8):979, 2011.

[168] Mitchell R Lunn and Ching H Wang. Spinal muscular atrophy. *The Lancet*, 371(9630):2120–2133, 2008.

[169] Theodore L Munsat and Kay E Davies. International SMA consortium meeting. *Neuromuscular Disorders*, 2(5):423–428, 1992.

[170] `http://omim.org/entry/253300`.

[171] `http://omim.org/entry/253550`.

[172] `http://omim.org/entry/253400`.

[173] `http://omim.org/entry/271150`.

[174] Arthur HM Burghes and Christine E Beattie. Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick? *Nature Reviews Neuroscience*, 10(8):597–609, 2009.

[175] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, 2009.

[176] Thomas W Prior, Adrian R Krainer, Yimin Hua, Kathryn J Swoboda, Pamela C Snyder, Scott J Bridgeman, Arthur HM Burghes, and John T Kissel. A positive modifier of spinal muscular atrophy in the SMN2 gene. *American journal of human genetics*, 85(3):408, 2009.

[177] AH Burghes. When is a deletion not a deletion? When it is converted. *American journal of human genetics*, 61(1):9, 1997.

[178] Bertold Schrank, Rudolf Götz, Jennifer M Gunnersen, Janice M Ure, Klaus V Toyka, Austin G Smith, and Michael Sendtner. Inactivation of the survival motor neuron gene, a candidate gene for human spinal muscular atrophy, leads to massive cell death in early mouse embryos. *Proceedings of the National Academy of Sciences*, 94(18):9920–9925, 1997.

[179] Luca Cartegni and Adrian R Krainer. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nature genetics*, 30(4):377–384, 2002.

[180] Tsuyoshi Kashima and James L Manley. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature genetics*, 34(4):460–463, 2003.

[181] Tsuyoshi Kashima, Nishta Rao, Charles J David, and James L Manley. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Human molecular genetics*, 16(24):3149–3159, 2007.

[182] Klaus Zerres, Sabine Rudnik-Schöneborn, Eric Forrest, Anna Lusakowska, Janina Borkowska, and Irena Hausmanowa-Petrusewicz. A collaborative study on the natural history of childhood and juvenile onset proximal spinal muscular atrophy (type II and III SMA): 569 patients. *Journal of the neurological sciences*, 146(1):67–72, 1997.

[183] Hsiu Mei Hsieh-Li, Jan-Gowth Chang, Yuh-Jyh Jong, Mei-Hsiang Wu, Nancy M Wang, Chang Hai Tsai, and Hung Li. A mouse model for spinal muscular atrophy. *Nature genetics*, 24(1):66–70, 2000.

[184] Umrao R Monani, Daniel D Coovert, and Arthur HM Burghes. Animal models of spinal muscular atrophy. *Human molecular genetics*, 9(16):2451–2457, 2000.

[185] Gabriela E Oprea, Sandra Kröber, Michelle L McWhorter, Wilfried Rossoll, Stefan Müller, Michael Krawczak, Gary J Bassell, Christine E Beattie, and Brunhilde Wirth. Plastin 3 is a protective modifier of autosomal recessive spinal muscular atrophy. *Science*, 320(5875):524–527, 2008.

[186] `http://www.preanalytix.com`.

[187] `http://www.genomics.agilent.com`.

[188] `http://www.nanodrop.com`.

[189] `https://github.com/wtsi-npg/illumina2bam`.

[190] `http://picard.sourceforge.net`.

[191] R Developement Core Team et al. R: A language and environment for statistical computing, 2005.

[192] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[193] `http://www.ingenuity.com`.

[194] `http://www.ncbi.nlm.nih.gov/SNP/`.

[195] Robert W Jackman and Susan C Kandarian. The molecular basis of skeletal muscle atrophy. *American Journal of Physiology-Cell Physiology*, 287(4):C834–C843, 2004.

[196] David J Glass. Skeletal muscle hypertrophy and atrophy signaling pathways. *The international journal of biochemistry & cell biology*, 37(10):1974–1984, 2005.

[197] Yanhong Zhou, Wei Wang, Danwei Zheng, Shuping Peng, Wei Xiong, Jian Ma, Zhaoyang Zeng, Minghua Wu, Ming Zhou, Juanjuan Xiang, et al. Risk of nasopharyngeal carcinoma associated with polymorphic lactotransferrin haplotypes. *Medical Oncology*, 29(3):1456–1462, 2012.

[198] MARIAN L Kruzel and MICHA Zimecki. Lactoferrin and immunologic dissonance: clinical implications. *ARCHIVUM IMMUNOLOGIAE ET THERAPIAE EXPERIMENTALIS-ENGLISH EDITION-*, 50(6):399–410, 2002.

[199] B Lönnerdal and S Iyer. Lactoferrin: molecular structure and biological function. *Annual review of nutrition*, 15(1):93–110, 1995.

[200] Lourdes Sanchez, Miguel Calvo, and Jeremy H Brock. Biological role of lactoferrin. *Archives of disease in childhood*, 67(5):657, 1992.

[201] YA Suzuki, V Lopez, and B Lönnerdal. Lactoferrin. *Cellular and Molecular Life Sciences*, 62(22):2560–2575, 2005.

[202] W Bellamy, M Takase, H Wakabayashi, K Kawase, and M Tomita. Antibacterial spectrum of lactoferricin b, a potent bactericidal peptide derived from the n-terminal region of bovine lactoferrin. *Journal of Applied Microbiology*, 73(6):472–479, 1992.

[203] Michal Zimecki and Marian L Kruzel. Systemic or local co-administration of lactoferrin with sensitizing dose of antigen enhances delayed type hypersensitivity in mice. *Immunology letters*, 74(3):183–188, 2000.

[204] T Zagulski, P Lipiski, A Zagulska, S Broniek, and Z Jarzabek. Lactoferrin can protect mice against a lethal dose of Escherichia coli in experimental infection in vivo. *British journal of experimental pathology*, 70(6):697, 1989.

[205] TADEUSZ Zagulski, P Lipinski, ALINA Zagulska, and Z Jarzabek. Antibacterial system generated by lactoferrin in mice in vivo is primarily a killing system. *International journal of experimental pathology*, 79(2):117, 1998.

[206] M Zimecki, J Mazurier, G Spik, and JA Kapp. Human lactoferrin induces phenotypic and functional changes in murine splenic B cells. *Immunology*, 86(1):122, 1995.

[207] Marian L Kruzel, Yael Harari, Chung-Ying Chen, and Gilbert A Castro. Lactoferrin protects gut mucosal integrity during endotoxemia induced by lipopolysaccharide in mice. *Inflammation*, 24(1):33–44, 2000.

[208] ML Kruzel, Y Harari, D Mailman, and JK Actor. Differential effects of prophylactic, concurrent and therapeutic lactoferrin treatment on LPS-induced inflammatory responses in mice. *Clinical & Experimental Immunology*, 130(1):25–31, 2002.

[209] ML Kruzel, T Zagulski, M Zimecki, K Shimazaki, H Tsuda, M Tomita, T Kuwata, JP Perraudin, et al. Lactoferrin and insult-induced metabolic imbalance in humans and other animals. In *Lactoferrin: structure, function and applications. Proceedings of the 4th International Conference on Lactoferrin: Structure, Function and Applications, held in Sapporo, Japan 18-22 May 1999.*, pages 301–310. Elsevier Science BV, 2000.

[210] Yoshiharu Takayama and Toshiaki Takezawa. Lactoferrin promotes collagen gel contractile activity of fibroblasts mediated by lipoprotein receptors. *Biochemistry and cell biology*, 84(3):268–274, 2006.

[211] JH Brock. Lactoferrin in human milk: its role in iron absorption and protection against enteric infection in the newborn infant. *Archives of disease in childhood*, 55(6):417, 1980.

[212] Peter W Howie, J Stewart Forsyth, Simon A Ogston, Ann Clark, and CD Florey. Protective effect of breast feeding against infection. *BMJ: British Medical Journal*, 300(6716):11, 1990.

[213] KS Erga, E Peen, O Tenstad, RK Reed, et al. Lactoferrin and anti-lactoferrin antibodies: effects of ironloading of lactoferrin on albumin extravasation in different tissues in rats. *Acta physiologica scandinavica*, 170(1):11–20, 2000.

[214] Roy D Baynes and Werner R Bezwoda. Lactoferrin and the inflammatory response. In *Lactoferrin*, pages 133–141. Springer, 1994.

[215] Armin J Grau, Vera Willig, Wolfgang Fogel, and Egon Werle. Assessment of plasma lactoferrin in Parkinson's disease. *Movement disorders*, 16(1):131–134, 2001.

[216] Zhong Ming Qian and Qin Wang. Expression of iron transport proteins and excessive iron accumulation in the brain in neurodegenerative disorders. *Brain research reviews*, 27(3):257–267, 1998.

[217] Béatrice Leveugle, Geneviève Spik, Daniel P Perl, Constantin Bouras, Howard M Fillit, and Patrick R Hof. The iron-binding protein lactotransferrin is present in pathologic lesions in a variety of neurodegenerative disorders: a comparative immunohistochemical analysis. *Brain research*, 650(1):20–31, 1994.

[218] Yoshifumi Iwamaru, Yoshihisa Shimizu, Morikazu Imamura, Yuichi Murayama, Ryo Endo, Yuichi Tagawa, Yuko Ushiki-Kaku, Takato Takenouchi, Hiroshi Kitani, Shirou Mohri, et al. Lactoferrin induces cell surface retention of prion protein and inhibits prion accumulation. *Journal of neurochemistry*, 107(3):636–646, 2008.

[219] Baptiste A Faucheux, Nathalie Nillesse, Philippe Damier, Genevieve Spik, Annick Mouatt-Prigent, Annick Pierce, Beatrice Leveugle, Nathalie Kubis, Jean-Jacques Hauw, and Yves Agid. Expression of lactoferrin receptors is increased in the mesencephalon of patients with Parkinson disease. *Proceedings of the National Academy of Sciences*, 92(21):9603–9607, 1995.

[220] B Leveugle, BA Faucheux, C Bouras, N Nillesse, G Spik, EC Hirsch, Y Agid, and PR Hof. Cellular distribution of the iron-binding protein lactotransferrin in the mesencephalon of Parkinson's disease cases. *Acta neuropathologica*, 91(6):566–572, 1996.

[221] Erwann Rousseau, Patrick P Michel, and Etienne C Hirsch. The iron-binding protein Lactoferrin protects vulnerable dopamine neurons from degeneration by preserving mitochondrial calcium homeostasis. *Molecular pharmacology*, 84(6):888–898, 2013.

[222] Patrick HC Van Berkel, Mick M Welling, Marlieke Geerts, Harry A van Veen, Bep Ravensbergen, Mourad Salaheddine, Ernest KJ Pauwels, Frank Pieper, Jan H Nuijens, and Peter H Nibbering. Large scale production of recombinant human lactoferrin in the milk of transgenic cows. *Nature Biotechnology*, 20(5):484–487, 2002.

[223] Pauline P Ward, Christopher S Piddington, Grainne A Cunningham, Xiaodong Zhou, Roger D Wyatt, and Orla M Conneely. A system for production of commercial quantities of human lactoferrin: a broad spectrum natural antibiotic. *Nature Biotechnology*, 13(5):498–503, 1995.

[224] Feng-Yun J Huang, Wan-Jou Chen, Wan-Yu Lee, Su-Tang Lo, Te-Wei Lee, and Jem-Mau Lo. In vitro and in vivo evaluation of Lactoferrin-conjugated liposomes as a novel carrier to improve the brain delivery. *International journal of molecular sciences*, 14(2):2862–2874, 2013.

[225] Nathalie Kayadjanian, Arthur Burghes, Richard S Finkel, Eugenio Mercuri, Francoise Rouault, Inge Schwersenz, and Kevin Talbot. Sma-europe workshop report: opportunities and challenges in developing clinical trials for spinal muscular atrophy in europe. *Orphanet journal of rare diseases*, 8(1):44, 2013.

[226] Tanja M Gruber and Carol A Gross. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annual Reviews in Microbiology*, 57(1):441–466, 2003.

[227] Sunghee Oh, Seongho Song, Gregory Grabowski, Hongyu Zhao, and James P Noonan. Time series expression analyses using RNA-seq: A statistical approach. *BioMed research international*, 2013:203681, 2013.

[228] `http://apps.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf`.

[229] John D McKinney, Kerstin Höner zu Bentrup, Ernesto J Muñoz-Elías, Andras Miczak, Bing Chen, Wai-Tsing Chan, Dana Swenson, James C Sacchettini, William R Jacobs, and David G Russell. Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature*, 406(6797):735–738, 2000.

[230] Dalin Rifat, William R Bishai, and Petros C Karakousis. Phosphate depletion: a novel trigger for Mycobacterium tuberculosis persistence. *Journal of Infectious Diseases*, 200(7):1126–1135, 2009.

[231] James E Gomez and John D McKinney. M. tuberculosis persistence, latency, and drug tolerance. *Tuberculosis (Edinburgh, Scotland)(ISSN: 1472-9792)*, 84(1-2):29–44, 2003.

[232] Eugene L Opie and JD Aronson. Tubercle bacilli in latent tuberculous lesions and in lung tissue without tuberculous lesions. *Arch Pathol Lab Med*, 4(1), 1927.

[233] Riccardo Manganelli and Roberta Provvedi. An integrated regulatory network including two positive feedback loops to modulate the activity of σE in mycobacteria. *Molecular microbiology*, 75(3):538–542, 2010.

[234] Riccardo Manganelli, Roberta Proveddi, Sebastien Rodrigue, Jocelyn Beaucher, Luc Gaudreau, and Issar Smith. σ factors and global gene regulation in Mycobacterium tuberculosis. *Journal of bacteriology*, 186(4):895–902, 2004.

[235] Sébastien Rodrigue, Roberta Provvedi, Pierre-Étienne Jacques, Luc Gaudreau, and Riccardo Manganelli. The σ factors of Mycobacterium tuberculosis. *FEMS microbiology reviews*, 30(6):926–941, 2006.

[236] John D Heimann. The extracytoplasmic function (ECF) sigma factors. *Advances in microbial physiology*, 46:47–110, 2002.

[237] Riccardo Manganelli, Martin I Voskuil, Gary K Schoolnik, and Issar Smith. The Mycobacterium tuberculosis ECF sigma factor σE: role in global gene expression and survival in macrophages. *Molecular microbiology*, 41(2):423–437, 2001.

[238] Stefan HE Kaufmann. Tuberculosis: deadly combination. *Nature*, 453(7193):295–296, 2008.

[239] Barbara Di Camillo, Brian A Irving, Jill Schimke, Tiziana Sanavia, Gianna Toffolo, Claudio Cobelli, and K Sreekumaran Nair. Function-based discovery of significant transcriptional temporal patterns in insulin stimulated muscle cells. *PloS one*, 7(3):e32391, 2012.

[240] Alicia D Henn, Shuang Wu, Xing Qiu, Melissa Ruda, Michael Stover, Hongmei Yang, Zhiping Liu, Stephen L Welle, Jeanne Holden-Wiltse, Hulin Wu, et al. High-resolution temporal response patterns to influenza vaccine reveal a distinct human plasma cell gene signature. *Scientific reports*, 3, 2013.

[241] Michael Burrows and David J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, 1994.

[242] Donald Adjeroh, Timothy C Bell, and Amar Mukherjee. *The burrows-wheeler transform: data compression, suffix arrays, and pattern matching*. Springer, 2008.

[243] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

[244] Giovanni Manzini. An analysis of the Burrows-Wheeler transform. In *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 669–677. Society for Industrial and Applied Mathematics, 1999.