universität wien

# DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Understanding the sampling properties of high throughput sequencing technologies“

verfasst von / submitted by

Luis Felipe Paulin Paz

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2018 / Vienna 2018

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | A 794 685 490 |
| Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet: | Molecular Biology |
| Betreut von / Supervisor: | Univ.-Prof. Dr. Arndt von Haeseler |

# Abstract

The function of a cell is determined by which genes are expressed at a given time point, either as protein coding genes, or non-coding RNAs. Accurate quantification of gene expression is an intensely researched field in molecular biology. Nowadays, with the widely adoption of next-generation sequencing technologies, which includes RNA-sequencing, the study of gene expression is ubiquitous in medical and biological sciences. The experimental procedure for RNA-sequencing is at large well known, with commercial kits and options of automation available. However, the true number of genes expressed in a cell remains unknown. Based on the observation that the laboratory experimental procedure of RNA-sequencing consists of a series of sampling events; from extracting the RNA fraction of interest (i.e. mRNA) to taking a small aliquot of the prepared library to sequence, we studied RNA-sequencing experiments in the context of a sampling problem.

First, we present the Pitman Sampling Formula (PSF), a sampling formula derived in the field of population genetics that is general enough to be applied to the study of RNA-sequencing. Then, we systematically evaluated the application of PSF and its derived statistics to RNA-sequencing experiments. We showed that the PSF allows an accurate inference of the number of undetected genes of an RNA-sequencing experiment. In the same scope, we used statistics of the PSF to estimate the number of additionally detected genes when increasing the sequencing depth in order to calculate the cost-benefit of further sequencing experiments.

Second, we used the sampling scheme of the PSF to develop RNACountSim, a new method to simulate RNA-sequencing experiments. Nowadays, simulated data is key for the development and evaluation of bioinformatic tools. For RNA-sequencing, simulation tools are aimed to generate count data where the number of genes showing differential expression is known. Many of the currently available methods use the

same distribution (i.e negative binomial) to generate simulated data and then test for differential gene expression. We instead used the Hoppe urn, an urn model of the PSF, to simulate count matrices where the number of genes showing differential expression is known *a priori*. We used both simulated and experimental replicates a to evaluate the performance of two widely used tools for differential expression: edgeR and DESeq2. We obtained similar results when using simulated and experimental data, thus showing that RNACountSim generates simulated data that resemble RNA-sequencing experiments. Moreover, with RNACountSim, we can simulate RNA-sequencing experiments where the number of differentially expressed genes is known to evaluate current tools that test for differential gene expression and aid in the development of new ones.

Finally, we propose the use of the PSF to evaluate the completion of genome annotation projects. Annotating a genome is a titanic task that arrives with each genome sequencing project. With the increased number of new genomes being sequenced every year, RNA-sequencing is nowadays one of the main methods used to improve genome annotation. Here, we used the PSF to predict the number of genes that remain to be annotated. To test this assertion we used the annotation of the human genome. We selected an older version of the human genome annotation (version 3b, dated 03.09.2009), to predict the number of genes that remain to be annotated. We then compared our predictions to a recent version of the annotation (version 25, dated 19.07.2016), which represents seven years of continuous improvement. We showed that our method accurately predicts the number of genes present in the newer version and thus, showing that the PSF provide good summary statistics to evaluate the state of the annotation in current genome projects.

# Zusammenfassung

Die Funktion einer Zelle wird dadurch bestimmt, welche Gene zu einem bestimmten Zeitpunkt exprimiert werden, entweder als proteincodierende Gene oder als nicht-codierende RNAs. Die genaue Quantifizierung der Genexpression ist ein intensiv erforschtes Feld in der Molekularbiologie. Heutzutage, mit der weiten Verbreitung von Sequenzierungstechnologien der nchsten Generation, zu denen auch die RNA-Sequenzierung gehrt, ist die Erforschung der Genexpression in den medizinischen und biologischen Wissenschaften allgegenwrtig. Das experimentelle Verfahren zur RNA-Sequenzierung ist im Allgemeinen bekannt, wobei kommerzielle Kits und Automatisierungsoptionen zur Verfgung stehen. Die wahre Anzahl der in einer Zelle exprimierten Gene bleibt jedoch unbekannt. Basierend auf der Beobachtung, dass das experimentelle Laborverfahren der RNA-Sequenzierung aus einer Reihe von Probenereignissen besteht; von der Extraktion der RNA-Fraktion von Interesse (d.h. mRNA) bis hin zur Sequenzierung eines kleinen Aliquots der vorbereiteten Bibliothek, haben wir RNA-Sequenzierungsexperimente im Rahmen eines Probenproblems untersucht.

Zuerst stellen wir die Pitman Sampling Formula (PSF) vor, eine Sampling-Formel, die im Bereich der Populationsgenetik abgeleitet wurde und allgemein genug ist, um auf die Studie der RNA-Sequenzierung angewendet zu werden. Anschlieend haben wir die Anwendung von PSF und den daraus abgeleiteten Statistiken auf RNA-Sequenzierungsexperimente systematisch untersucht. Wir zeigten, dass das PSF einen genauen Rckschluss auf die Anzahl der unerkannten Gene eines RNA-Sequenzierungsexperiments ermglicht. Im gleichen Umfang haben wir die Statistik des PSF verwendet, um die Anzahl der zustzlich detektierten Gene bei Erhhung der Sequenziertiefe zu schtzen, um den Kosten-Nutzen weiterer Sequenzierungsexperimente zu berechnen. Zweitens haben wir den Stichprobenplan des PSF verwendet,

um RNACountSim zu entwickeln, eine neue Methode zur Simulation von RNA-Sequenzierungsexperimenten. Simulierte Daten sind heute der Schlssel fr die Entwicklung und Bewertung von Bioinformatikwerkzeugen. Fr die RNA-Sequenzierung werden Simulationswerkzeuge eingesetzt, um Zhldaten zu erzeugen, bei denen die Anzahl der Gene mit differentieller Expression bekannt ist. Viele der derzeit verfgbaren Methoden verwenden die gleiche Verteilung (z.B. negatives Binomial), um simulierte Daten zu erzeugen und dann auf differentielle Genexpression zu testen. Stattdessen haben wir die Hoppe Urne, ein Urnenmodell des PSF, verwendet, um Zhlmatrizen zu simulieren, bei denen die Anzahl der Gene, die eine unterschiedliche Expression zeigen, bekannt ist *a priori*. Wir verwendeten sowohl simulierte als auch experimentelle Replikate, um die Leistung von zwei weit verbreiteten Werkzeugen fr die differentielle Expression zu bewerten: edgeR und DESeq2. Wir haben hnliche Ergebnisse bei der Verwendung von simulierten und experimentellen Daten erzielt, was zeigt, dass RNA-CountSim simulierte Daten erzeugt, die an RNA-Sequenzierungsexperimente erinnern. Darber hinaus knnen wir mit RNACountSim RNA-Sequenzierungsexperimente simulieren, bei denen die Anzahl der differentiell exprimierten Gene bekannt ist, um aktuelle Werkzeuge zu bewerten, die auf differentielle Genexpression testen und bei der Entwicklung neuer Gene helfen. Schlielich schlagen wir vor, die Verwendung des PSF zur Bewertung der Fertigstellung von Genom-Annotationsprojekten zu verwenden. Die Annotation eines Genoms ist eine titanische Aufgabe, die bei jedem Genomsequenzierungsprojekt anfllt. Da jedes Jahr mehr neue Genome sequenziert werden, ist die RNA-Sequenzierung heute eine der wichtigsten Methoden zur Verbesserung der Genomannotation. Hier haben wir mit dem PSF die Anzahl der Gene vorhergesagt, die noch zu kommentieren sind. Um diese Behauptung zu testen, haben wir die Annotation des menschlichen Genoms verwendet. Wir haben eine ltere Version der Annotation des menschlichen Genoms (Version 3b, vom 03.09.2009) ausgewhlt, um die Anzahl der noch zu annotierenden Gene vorherzusagen. Anschlieend verglichen wir unsere Vorhersagen mit einer aktuellen Version der Annotation (Version 25, vom 19.07.2016), die sieben Jahre kontinuierliche Verbesserung darstellt. Wir haben gezeigt, dass unsere Methode die Anzahl der in der neueren Version vorhandenen Gene genau vorhersagt und damit zeigt, dass das PSF gute zusammenfassende Statistiken liefert, um den Zustand der Annotation in aktuellen Genomprojekten zu bewerten.

# Acknowledgments

First, I want to express my gratitude to my supervisor Arndt von Haeseler. Arndt provided me amazing opportunities to learn and grow not only as scientist but also as a person.

I would like to thank all my colleagues at CIBIV, specially Celine and Florian for all those hours of discussion almost every second Wednesday for now more than three years. Thanks you for all your helpful comments and suggestions.

Big thanks to all the members of CIBIV for making my experience here great.

Special thanks to my friend that I met at CIBIV: Peter, Milica, Celine, Olga, Florian and Konstantina, for all the great moments we shared.

También quiero agradecer a mis padres y mi hermano quienes me apoyaron todos estos años. Y en especial, le quiero agradecer a mi esposa Benelli. Gracias por tu infinita paciencia y apoyo durante estos más de cuatro años. Gracias por todo el cariño, dedicación y sacrificio para hacer posible esto posible.

# Contents

**Abstract** i

**Zusammenfassung** iii

**Acknowledgments** v

**Organization of the thesis** 1

**1 Introduction** 3

1.1 Motivation . . . 3

1.2 Next-generation sequencing technologies . . . 4

1.3 RNA Sequencing . . . 4

1.3.1 RNA-sequencing experimental procedure . . . 6

1.3.2 Bioinformatic analysis work-flow . . . 8

1.4 Statistical models for RNA sequencing . . . 9

1.4.1 RNA-sequencing as a sampling process . . . 11

1.5 The Ewens Sampling Formula . . . 11

**2 The Pitman sampling formula** 13

2.1 The Pitman Sampling Formula . . . 13

2.2 The Hoppe Urn model . . . 15

2.2.1 The Hoppe urn sampling algorithm . . . 15

2.3 The Pitman Sampling applied to expression data . . . 16

2.3.1 The Pitman Sampling formula in the study of expressed sequence tags . . . 16

2.3.2 Pitman Sampling formula in the study of RNA-sequencing . . . 17

2.4 Statistics of the Pitman Sampling Formula . . . . . 17
2.4.1 Expected value of the number of detected genes . . . . . 17
2.4.2 Number of undetected genes . . . . . 18
2.4.3 Limiting relative frequencies of the genes . . . . . 18
2.4.4 Size-biased random permutation . . . . . 18
2.5 Conclusions . . . . . 19

**3 Evaluation of the Pitman Sampling Formula applied to RNA-sequencing 21**
3.1 Methods . . . . . 22
3.1.1 Experimental data used in this work . . . . . 22
3.1.2 Estimation of the number of genes expressed in a transcriptome . 24
3.1.3 Experimental design . . . . . 26
3.2 Analysis and Results . . . . . 28
3.2.1 Estimating the number of non-detected genes . . . . . 28
3.2.2 Additional detected genes when increasing the sequencing depth 31
3.2.3 Assessment of the number of shared genes between replicate experiments . . . . . 37
3.3 Discussion . . . . . 39
3.4 Conclusions . . . . . 41

**4 RNACountSim: fast simulation of RNA-Sequencing experiments. 43**
4.1 Introduction . . . . . 43
4.2 Methods . . . . . 45
4.2.1 Input . . . . . 45
4.2.2 Simulating RNA-sequencing experiments . . . . . 45
4.2.3 Measurements of performance . . . . . 50
4.3 Results . . . . . 51
4.3.1 Simulating data-set with no genes showing differential expression 51
4.3.2 Simulating data-set with $D$ genes showing differential expression 54
4.4 Discussion . . . . . 58
4.5 Conclusions . . . . . 60

**5 GeneComplete: a tool for fast assessment of the completion of a genome annotation. 61**

5.1 Methods . . . . . 62

5.1.1 Selection of annotated genes . . . . . 62

5.1.2 Experimental design . . . . . 62

5.2 Results and Discussion . . . . . 63

5.3 Conclusions . . . . . 67

**6 Summary 69**

# Organization of the thesis

In **Chapter 1** we present the basic concepts of next-generation sequencing (NGS) technologies with special emphasis on RNA-sequencing which is an application of NGS. Then, we explain the current statistical models for studying RNA-sequencing. Finally, we detail the Ewens Sampling Formula, a formula developed in the field of ecology, which is a special case of the sampling formula used in this work. In **Chapter 2**, we introduce the Pitman Sampling Formula (PSF), a sampling formula derived in the field of ecology, that is general enough that we applied it to RNA-sequencing data. Here, we describe previous works that used the PSF in the context of expression genetic data. Finally, we describe statistics for the PSF that are relevant for this work. **Chapter 3** details the evaluation of the statistics described in Chapter 2 applied to RNA-sequencing. Here, we assessed the estimation of the number of missed genes, the cost-benefit of follow-up sequencing experiments. Then, we described and evaluated a statistic compute the expected number of shared genes in experimental replicates. We aimed to use the expected number of shared for quality control. **Chapter 4** describes RNACountSim, an algorithm based on an urn model for the PSF, to simulate RNA-sequencing experiments. RNACountSim can simulate large experiments in the sense of number of replicates and sequencing depth in almost negligible time while the user can define the number of genes to show differential expression. **Chapter 5** introduces GeneComplete, an application of a statistic of the PSF to evaluate the completion of genome annotation and give insights towards which tissues or conditions may be worth of further sequencing. Finally, Chapter 6 summarizes the results of the thesis and gives an outlook of future work.

# 1 Introduction

## 1.1 Motivation

Next generation sequencing is now ubiquitous in life sciences and medical research (van Dijk et al., 2014; Renkema et al., 2014) generating every year a vast amount of genetic data. A better understanding of NGS through mathematical and statistical models aids in the development of new computational methods and in the improvement of experimental workflows (Conesa et al., 2016; Zhaoa et al., 2017), which altogether, lead to a broader adoption of the technology.

RNA-sequencing is nowadays the preferred method for the investigation of the transcriptome landscape. Since its inception Mortazavi et al. (2008) stated that "If enough reads are collected from a sample, it should in theory be possible to detect and quantify RNAs from all biologically relevant abundance classes...". This is mainly true when RNA-sequencing is used to test for differential gene expression between different tissues (Blazie et al., 2015), developmental stages or conditions (Wang et al., 2018). However, the amount of sequenced reads necessary to reliably detect and measure low abundant RNAs is still an open question. Finally, with the decrease in sequencing cost, scientist not only worry about "How much it will sequencing cost?", but also for the time and effort invested and if "Is it worth to sequence more?"

Generally, in every sequencing experiment only a fraction of the extracted nucleic acids is used for sequencing. This motivated us to study RNA-sequencing in terms of a sampling problem. We made use of sampling formulas developed in the field of ecology and population genetics that are general enough that can be applied to RNA-sequencing. In this thesis we evaluated one of such formulas with the aim of better understand the sampling properties of RNA-sequencing.

## 1.2 Next-generation sequencing technologies

Since 1977 when Sanger and colleagues (Sanger et al., 1977) developed a method to sequence DNA, DNA sequencing has become an ubiquitous tool in molecular biology and gave rise to the genomics era. For almost 30 years Sanger sequencing became "the method" for DNA sequencing, which led to the sequencing of the first bacterial genome in 1995 (Fleischmann et al., 1995), the first eukaryotic genome in 1997 (Goffeau et al., 1996), the first mammalian genome in 2001 (Waterston et al., 2002) to finally the completion of the first draft of the human genome in 2004 (IHGSC, 2004). However this methodology was slow and technically laborious.

This led to the development of new sequencing methodologies that did not required DNA cloning and could yield a higher amount of DNA sequences in a single experiment. These new methodologies were called next-generation sequencing (NGS) making a reference to Sanger being the first generation. In the middle to late 2000s several NGS technologies competed for most of the market share: 454 pyrosequencing in 2005, both SOLiD and Illumina in 2007 and ionTorrent PGM in 2010 (van Dijk et al., 2014). With these new technologies one could sequence hundred thousands to hundred of millions of small DNA fragments called *reads*. As an example of the fast development of this technologies, in the last 10 years for the Illumina platform the read length has increased from 35 bases (bp) to 300 bp, and the yield of reads per experiment has grown up to 400 million reads. Furthermore, DNA sequencing can be used to study other nucleic acid molecules such as RNA. In all NGS technologies RNA is not sequence itself, but the complementary DNA (cDNA) produced by reverse-transcription of RNA molecules. Similar to genomic projects, where the goal is to have an understanding of the entire genome sequence of an organism, transcriptomics projects aim to study the entire transcription profile of an organism. A more in depth history of the development of RNA-sequencing is detailed in the section to follow.

## 1.3 RNA Sequencing

Historically, RNA molecules were classified into three major classes: ribosomal RNA (rRNA), transfer RNA (tRNA) and messenger RNA (mRNA). The rRNA is a component of the ribosome which is essential for protein synthesis in all living organism and it is

the most abundant RNA molecule in a cell. The tRNAs are small RNA molecules that serve as an intermediary between the mRNA and the amino-acid sequence of proteins. Finally, the mRNA constitutes a large family of RNA molecules that contain the genetic information coded in DNA, which is then processed to produce proteins. mRNAs are the product of gene expression where a section of genomic DNA is copied into an RNA molecule in a process called transcription.

Before NGS, several methods existed to sequence small pieces of RNA molecules. Examples of such methods are expressed sequence tags (EST) (Boguski and Schuler, 1995), serial analysis of gene expression (SAGE) (Velculescu et al., 1995) and cap analysis of gene expression (CAGE) (Shiraki et al., 2003). All of these methods were used to survey the collection of all transcribed RNA molecules, known as *the transcriptome*. With the advent of NGS technologies, in mid-2008, different groups described methods to sequence and survey the transcriptome of mouse (Cloonan et al., 2008; Mortazavi et al., 2008) and yeast (Nagalakshmi et al., 2008; Wilhelm et al., 2008), which they called **RNA-sequencing (RNA-Seq)**. Nagalakshmi et al. (2008) described the experimental and bioinformatic pipeline to study the transcriptional landscape in yeast. Their focus was on establishing the methodology and describing possible uses both as a quantitative (as a method to quantify RNA) and as a qualitative method (to improve genome annotation). Then, Wilhelm et al. (2008) described the whole transcriptome of the fission yeast by performing RNA-sequencing from multiple conditions and compared their results with, at the time, the most used methodology to quantify transcript abundance: microarrays. They found that RNA-sequencing, in comparison to microarrays, showed little to no background noise and where able to detect novel transcripts that are condition-specific. Finally, Mortazavi et al. (2008) described methods for studying the more complex transcriptome of mammals. In their work, Mortazavi and colleagues asked questions that to date are intense areas of research such as ”How relative quantification will be converted to absolute RNA concentrations”.

Since then the number of distinct applications of RNA-sequencing has vastly grown over the years. Different applications such as quantification of transcript levels, differential gene expression and detection of alternative splicing have different challenges and the scientific community has put a great effort in the development of theoretical models and computational methods to handle these type of data (Conesa et al., 2016).

RNA-sequencing is now routinely being used to survey the transcriptome with the aim of catalog and quantify all possible transcripts present in an organism. Different to genomic DNA, mRNA and other transcribed RNA molecules are present in distinct abundances which depend on the developmental stage and environmental conditions the organism is living in. Therefore to catalog all possible transcripts one must survey different tissues, developmental stages and environmental conditions.

Moreover in eukaryotic cells, transcribed mRNA have exons which are the coding sequences and introns which do not encode to protein sequences but are spacers that may also contain other type of information. The removal of introns is called splicing, and the presence-absence of distinct exons from the same gene can derived distinct transcript variants. The study of distinct splice variants of the same transcribed gene is another field of research in which RNA-sequencing is highly used. A detailed review of alternative splicing can be found in Baralle and Giudice (2017).

As we mentioned before, RNA-sequencing is used as a quantitative method to determine the abundance of gene expression. Here, the amount of mRNA serves as proxy for the amount of protein expressed. When studying more than one biological condition, developmental stage or tissue, one can compare the abundance of the same gene in each condition. This widely used methodology is called *differential gene expression (DGE)*, where the amount of reads that map to a specific gene is used as a proxy for its abundance. Up to now we have used the terms transcript, mRNA and gene almost interchangeably. In order to avoid confusion hereafter we will use the term **gene** as a placeholder for a much broader definition that include a wide variety of gene models.

In the following sections we will explain the general experimental and bioinformatic workflow of RNA-sequencing, to then, focus in the statistical models developed for RNA-sequencing.

### 1.3.1 RNA-sequencing experimental procedure

Gene expression is a dynamic, yet tightly regulated process. In contrast to DNA sequencing, when studying RNA molecules, we get a snapshot of the current transcription profile averaged from all the cells. To better understand the connection between RNA-sequencing and a sampling process, first we need to explain the basics of the experimental protocol prior sequencing which is known as **library preparation**.

The first step in the library preparation protocol is the **extraction** of the RNA molecules. Here, we must pay specific attention to the fact the rRNA constitute the most abundant RNA molecule in the cells accounting up to 90% of total cellular RNAs. To deal with this issue, scientist use commercial kits that remove cytoplasmic rRNA (i.e. Ribo-Zero by Illumina Inc.). Additionally, if the RNA molecule of interest is the mRNA, we can use the poly(A) tail, which is a post-transcriptional modification that consist in the addition of multiple adenosine molecules in the 3' end to of the mRNA molecule, to capture the mRNA.

The next step is the **fragmentation** of the extracted RNA molecules as NGS technologies do not allow us to sequence the full-length of most RNA molecules. Examples of protocols to fragment the RNA molecules are **physical fragmentation** by acoustic shearing (Covaris instrument) and sonication (Diagenode Inc.), **enzymatic methods** by the use of an endonuclease that cleaves RNA into small fragments or **chemical fragmentation** by the use of divalent cations (magnesium or zinc) under high temperature (Illumina Inc.).

The following step consist on converting the RNA molecules into DNA molecules. All NGS technologies requires DNA for sequencing. By **reverse-transcription** the RNA molecules are transcribed into complementary DNA (cDNA) molecules. Then, small oligonucleotides of known sequence called *adapters* are attached to the 5' and 3' of the DNA molecules in a step known as **adapter ligation**. The adapters are used as anchors to bind the DNA fragments to the flow cell which is a glass slide containing thousands of millions of nanowells where the sequencing occurs.

The final step before sequencing consist in **enrich** for the molecules that contain the adapters by PCR, to then take the adequate amount of DNA in solution (usually in nanograms per microliter) for sequencing. In most if not all cases, only a small fraction if the library is actually sequenced.

Finally, during the **sequencing** step million of reads of a fix length are produced. The length of the read will depend on the sequencing technology used, but generally they range between 50 and 400 bp. Also, and depending of the sequencing protocol, one can sequence only from one side of the DNA fragment as **single end reads** or from both sides of the fragment as **pair end reads** (figure 1.1).

### 1.3.2 Bioinformatic analysis work-flow

The main goal of RNA-sequencing experiments is either a qualitative description of the expressed genes or the comparison of the gene expression levels between different conditions. The result of a sequencing experiments are millions to hundred of millions of reads. Before the downstream analysis can be performed, a series of processes need to be applied to the *raw sequencing data* to reduce the large amount of reads into summarized information.

The first step is **mapping** the reads to a reference genome. Here, the goal is to assign each read a chromosomal location that represents the most likely position where the read may have originated. Here, an **annotation file** is used to aid in the mapping procedure. An annotation file contains a detailed description of the currently known gene models (genes, transcripts, exons, and so on). The essential parts of the annotation file are the gene name or ID, and the chromosomal position of the gene. During the mapping procedure this information is used for example to split a read that overlaps an exon-exon junction (where an intron was removed). Examples of RNA-sequencing read mappers are TopHat2, STAR and HISAT (Kim et al., 2013; Dobin et al., 2013; Kim et al., 2015).

After mapping a quality filtering step is usually performed to remove the reads were assigned to a chromosomal location with low confidence. This may occur if the read contains low complexity sequences (e.g. `GCGCGCGCGCGC`) which could be mapped to more than one genomic position. The same occurs when the read is mapped to a repetitive region of the genome. Read mappers give a quality score to each of the mapped reads that relates to the confidence that the reads belongs to that particular genomic position. It is important to note that different mappers define the quality score differently.

Finally, for comparison of gene expression, the number of reads that mapped each genes is needed. Here, researchers need to make the decision of which gene model will be used for the summarization. Each read is then assigned to a gene model and the number of reads per gene is summarized into a *count table* or *count matrix*. For downstream analysis, such as differential gene expression, the count table is the preferred input for many tools (Conesa et al., 2016).

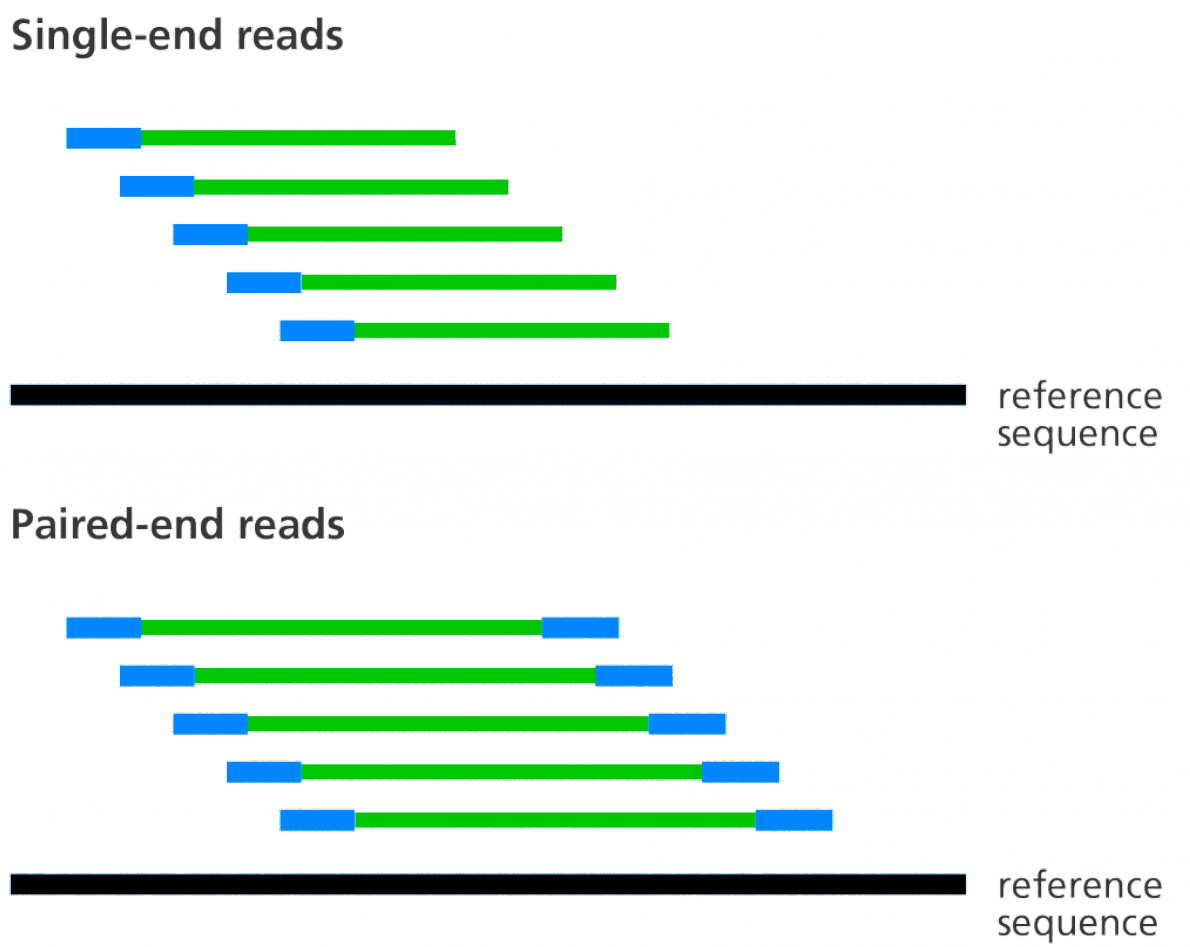



Figure 1.1: Example of single-end and pair-end sequencing. In black is represented the RNA molecule that was sequenced. The blue segments represent the sequenced reads. In the top panel is shown single-end reads where only one side of the fragment is sequenced. The non-sequenced part of the fragments is shown in green. In the bottom panel pair-end reads are shown. Same as before, the green section represent the non-sequenced part of the fragment. Figure adapted from https://www.biostars.org/p/267167/ by user Devon Ryan.

## 1.4 Statistical models for RNA sequencing

An RNA-Sequencing experiment offers a comprehensive survey of the expressed genes in a given sample of interest. From its inception, RNA-sequencing was described

as a quantitative method which could be used to detect the difference in abundance of RNA molecules. As with any process that generates count data, it is important to use the appropriate statistical model to account for the inherent variation that comes from the sampling process itself. One of the first statistical models for RNA-sequencing is RPKM by Mortazavi et al. (2008). They aimed to quantify transcript levels taking into account the molar concentration and transcript length to facilitate the comparison of transcript levels within and between samples.

The wider adoption of RNA-sequencing to quantify expression levels over tiling arrays lead to the development of two big areas of research of statistical model for RNA sequencing: 1. development of normalization methods and 2. development of models to test for differential gene expression from count data. In the later, the Poisson distribution was first proposed to model the read counts distribution for the same gene across replicates. Data that follows the Poisson distribution is integer-valued and thus it is sensible to use for count data. However, it was later shown that the assumption of equal mean and variance of the Poisson distribution was not applicable for RNA-sequencing data and thus, the negative binomial distribution distribution is used instead as a ”overdispersed” Poisson (Anders and Huber, 2010).

Additionally, questions regarding sequence depth and how it affects RNA-sequencing have been addressed (Sims et al., 2014). Many genes are expressed at very low levels, making the analysis of rare transcripts problematic (Kuznetsov et al., 2002). Busby et al. (2013) addressed the question of how to balance between the number of experimental replicates and the sequencing depth per sample to get the most power in differential expression analysis. Lijoi et al. (2007) asked what is proportion of unique genes represented in a given sample?, which translated to ”**How many unique genes can we detect?**”. This questions opened the door to different models for RNA-sequencing where the question to answer is related to how much of the transcriptomic landscape is captured by the respective sequencing experiment (Tauber and von Haeseler, 2013) and the number of undetected genes (Garcia-Ortega and Martinez, 2015).

### 1.4.1 RNA-sequencing as a sampling process

A sampling process is the random selection of a subset of individuals (a sample) from a population. Each individual sampled from the populations is considered an observation. Samples are used to study characteristics of the whole population, as surveying the whole populations is, in most of the cases, intractable.

RNA-sequencing can be viewed as a sampling process. The library preparation prior sequencing can be seen as a series of sampling events. First, we collect sample of the tissue of an organism or a sample of a cell culture. Then, we extract the RNA. The RNA population contains distinct RNA molecules in different abundances. Therefore each RNA molecules has a distinct probability of being detected. Then, during the library preparation the RNA molecules are fragmented and amplified. Here, the fragmentation process will give large RNA molecules a higher probability of being detected as more fragments of a fixed length can be generated. Finally, the amplification by PCR is a stochastic process itself. PCR has been studied as a branching process and it has been shown that it introduces amplification heterogeneity (Best et al., 2015) which means that not all the fragments will amplified equally.

Griebel et al. (2012) developed a model to deal with every step of these process. Instead, we propose the use of sampling formulas that offer a simple, yet powerful method to study RNA-sequencing. By using sampling formulas we do not expect to explain nor study every detail which comprises RNA-sequencing, but rather we will make use of their powerful predictive statistics to provide insights on the benefit of further sequencing experiments.

## 1.5 The Ewens Sampling Formula

In his seminal paper Ewens (1972) developed theory to study the sampling of neutral alleles. Ewens considered an infinite allele model and developed a sampling formula to study the distribution of the number of times different alleles are observed in a sample from the population. This model is known as the Ewens' Sampling Formula (eq. 1.1).

$$Pr(a_1, ..., a_n, \theta) = \frac{\theta}{\theta(\theta+1)...(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!} \tag{1.1}$$

where $\theta$ is the model parameter that represents the mutation rate scaled to the effective population size, $n$ is the total number of alleles taken from a population, which are classified according to the number of times they were detected with $a_1$ representing the number of alleles observed exactly once, $a_2$ being the number of alleles observed twice, and so on.

By studying the inference properties of the model, Ewens was able to compute the mutation rate by estimating the model parameter $\theta$. Finally, Ewens discussed the possibility of developing a test to assess whether the sampled alleles are neutral.

Ewens discusses possible complications due propertied inherent to a population such as linkage and fluctuation in population size. However, Ewens relatively simple model has proved to be quite powerful and has been since used in the field of ecology and population genetics (Anderson et al., 2014; Rodriguez and Quintana, 2015; Griffiths and Tavare, 2018).

# 2 The Pitman sampling formula

The Pitman Sampling Formula (PSF) is a generalization of the Ewens Sampling formula which introduces a second parameter. The PSF was not developed as the generalization of the ESF, but while studying random partitions Pitman derived the two-parameter generalization of Ewens' model.

## 2.1 The Pitman Sampling Formula

Let us consider an exchangeable random partition of $n \in \mathbb{N}$. A partition of $n$ is an unordered collection of positive integers with sum $n$. Pitman (1995) described two common ways to code a partition of $n$:

- by the sequence $n_1, n_2, ..., n_k$ with $\sum n_i = n$ where $n_i$ is the number of times $i$ was observed and $k$ is the number of distinct integers in the sequence (*frequency vector*).
- by the numbers of terms of various sizes, for example $g_j$ describes which numbers were observed exactly $j$ times, with $\sum g_j = k$ and $\sum j g_j = n$ (*occupancy vector*).

If we consider an RNA-sequencing experiment that yielded $n$ reads summarized in a count table, which represents the number of reads assigned to each of the $k$ detected genes, we can use the partition scheme described by Pitman to study of RNA-sequencing. Here, the count table can be seen as the *frequency vector*. Moreover the *occupancy vector* $\mathbf{g}_n = (g_1, g_2, ...g_j)$ will describe the number of genes detected with the same read count, for example $g_1$ is the number of genes detected with exactly one read, $g_2$ are the genes detected with exactly two reads and in general $g_j$ are the genes detected with exactly $j$ reads. With this in mind we will talk about detected genes for genes with read counts $\geq 1$ and non detected genes otherwise.

The **Pitman Sampling Formula** (Pitman, 1995) describes the probability distribution associated with an *occupancy vector* $\mathbf{G}_n$ and the number of distinct observed genes $K_n$:

$$P[K_n = k, \mathbf{G}_n = \mathbf{g}_n] = n! \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{n} \frac{(1-\sigma)_{j-1}^{g_j}}{j!^{g_j} g_j!} \tag{2.1}$$

where $(x)_n = x(x+1)(x+2)...(x+n-1)$ is the ascending factorial with $(x)_0 = 1$ and $\theta$ and $\sigma$ are the two parameters of the PSF. For our case of study the parameters $\theta$ and $\sigma$ are unknown and need to be estimated from data.

Pitman (1995) described that the values of $\theta$ and $\sigma$ must follow one of two conditions in order to satisfy the rules of probability. For the first condition, $0 \leq \sigma < 1$ and $\theta > -\sigma$ determine an infinite sampling universe.Note that the ESF is the special case when $\sigma = 0$. If, $\sigma = -\kappa < 0$ and $\theta = m\kappa$ with $m \in \mathbb{N}$, then there exist a finite number of detectable genes. We estimate the model parameters from data assuming the later condition. Finally, the upper bound of the number of detectable genes is given by $m$.

The PSF follows a sequential sampling scheme (Pitman (1995), Figure 2.1). Let us consider a particular experiment with $n$ mapped reads and $k$ detected genes. Pitman described for the PSF the probability for detecting a new gene in the next read

$$\frac{\theta + k\sigma}{\theta + n} \tag{2.2}$$

and the probability of assigning the next read to an already detected gene

$$\frac{n_i - \sigma}{\theta + n} \tag{2.3}$$

where $n_i$ is the read count assigned to the $i^{th}$ gene. With this in mind, and given that $\sigma$ can only take negative values, when $k \to m$ the probability of detecting new genes goes to zero. Thus, we can then interpret $m$ as the maximum number of genes we can detect.

At this point the reader may ask why sequential sampling is a good analogy for RNA-sequencing if all sequenced reads come at once. Let us consider a follow-up experiment where $n_f$ reads are sequenced. Here, we can treat them as arriving sequentially and compute the probability that the next read, whichever is selected from the $n_f$, is assigned to a previously undetected gene.

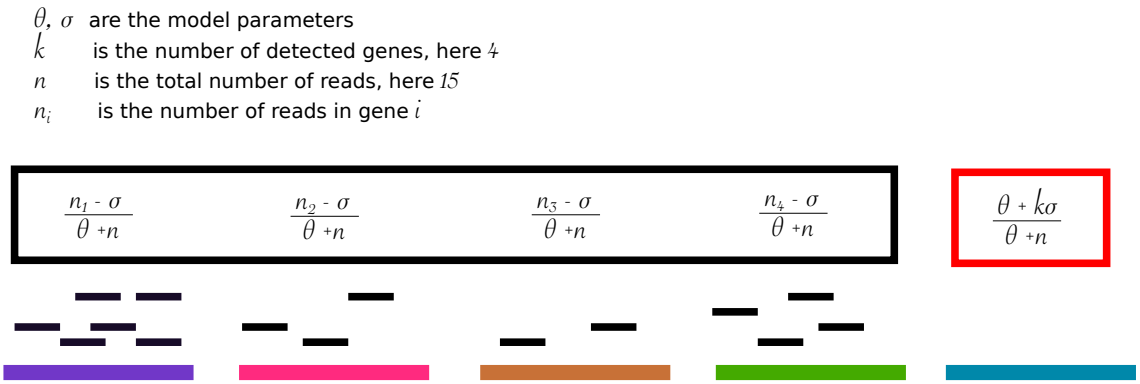



Figure 2.1: Example of the sampling scheme for the Pitman Sampling Formula. Given $n = 15$ mapped reads and $k = 4$ detected genes, we represent in the black box we show the probabilities of assigning the next read to an already detected gene (colors purple, pink, brown and green. The red box displays the probability of detecting a new gene (blue) with the next read.

## 2.2 The Hoppe Urn model

The Hoppe urn (Hoppe, 1984) is a model of sequential sampling developed to generate samples from the Ewens Sampling Formula. The Hoppe urn can be adapted to generate samples from the PSF by using the previously described probability of detecting an new gene in the next read (eq. 2.2) and the probability of detecting an already seen gene (eq. 2.3).

The basic idea of the Hoppe urn is sampling with replacement, with a twist, in which each sampling step will modify the probability distribution. In the following description we will refer to the *weight* of each element in the urn instead of its probability of being selected.

### 2.2.1 The Hoppe urn sampling algorithm

The algorithm to sample from the Hoppe urn is divided into two steps: initialization and sampling. For the initialization step consider an urn containing one black ball and $n \in \mathbb{N}$ colored (non-black) balls. If $n = 0$ then the urn contains only the black ball

which has weight $\theta$. If $n > 0$ then the urn contains the black ball with weight $\theta$ and a $n$ colored balls. Here, all the balls of the same color have an overall weight $n_i - \sigma$ where $n_i$ is the number of balls with the same color $i$.

Then, during the sampling step, when the black ball is selected a new ball of different color is added to the urn with a weight of $1 - \sigma$, and the weight of the black ball is decreased by $|\sigma|$. When a colored ball is drawn, then such ball is returned to the urn, together with an additional ball of the same color.

The sampling procedure ends when either of two conditions is met:

- we have observed a given number of distinct colored balls $k > 0$
- the total number of colored balls in the urn is equal to a given $n > 0$.

Translated to RNA-sequencing, each colored ball represents a read that can be unambiguously assigned to only one gene. In consequence, all the balls of the same color represent the reads that are mapped to the same gene. Then, the black ball is used as an instrument to represent the probability of detecting additional genes.

## 2.3 The Pitman Sampling applied to expression data

### 2.3.1 The Pitman Sampling formula in the study of expressed sequence tags

Lijoi et al. (2007) used a Bayesian non-parametric approach which used the structure partition from the PSF as a prior distribution and applied formulas derived from the field of ecology to expressed sequence tags (EST) data. Lijoi and colleagues aimed to calculate the redundancy of the EST libraries by calculating the proportion of detected genes. In the same scope Lijoi and colleagues estimated the discovery rate for future experiments as the number of new genes to be detected. They showed that their predictions, when compared to frequentist non-parametric methods, were more reliable for any size of the additional sample.

This work inspired the use of available theory developed in other fields to the study of similar problems. In ecology, scientist have wondered the number of unobserved species (for a review see Bunge and Fitzpatrick (1993)), which translated to RNA-sequencing data, we ask for the number of undetected genes.

### 2.3.2 Pitman Sampling formula in the study of RNA-sequencing

In their paper, Tauber and von Haeseler (2013) aimed to evaluate use of the PSF applied to RNA-sequencing. When compared to RNA-sequencing, EST technology was low throughput, meaning that the number of sequenced reads did not surpass the thousands which is closer to the individuals in ecological survey.

Tauber and von Haeseler studied the sampling process of RNA-sequencing on the gene level to ask ”How many genes were missed?” and ”How many more can be detected when more reads are sequenced?”. Here, they propose the use of parameter $m$ of the PSF as the unknown number of genes present in the sequencing sample. Moreover, they tested the estimation of the ”gene universe” by estimating $m$ from public data-sets. Furthermore, they showed that the PSF can be used to accurately estimate the number of additional detected genes with increase of sequence depth. They exemplified the applicability with different data-sets of human, mouse and yeast.

Their results also showed that the PSF is a general enough sampling formula such that it can be applied to high throughput sequencing data. Finally, their work was key in the development of the present thesis. We built upon their results and developed new applications to the study of RNA-sequencing.

## 2.4 Statistics of the Pitman Sampling Formula

Here we present a summary of statistics developed for the Pitman Sampling Formula that we will be using and evaluating in this thesis.

### 2.4.1 Expected value of the number of detected genes

Yamato and Sibuya (2000) developed a formula for the expected value (eq. 2.4) and variance (eq. 2.5) of number $k$ of detected genes for a given read count $n$.

$$\mathbb{E}(K_n) = \frac{\theta}{\sigma}[\frac{(\theta+\sigma)_n}{(\theta)_n} - 1] \tag{2.4}$$

$$\mathbb{V}(K_n) = \frac{\theta(\theta+\sigma)}{\sigma^2}\frac{(\theta+2\sigma)_n}{(\theta)_n} - \frac{\theta^2}{\sigma^2}[\frac{(\theta+\sigma)_n}{(\theta)_n}]^2 - \frac{\theta}{\sigma}\frac{(\theta+\sigma)_n}{(\theta)_n} \tag{2.5}$$

Ewens (1972) showed that for the ESF when $n \to \infty$ the distribution of $k$ approaches a normal distribution and for small $n$ normality may not be reached. The ESF is the

special case of the PSF when $\sigma = 0$, thus we assume that for the PSF $k$ is normally distributed for large enough $n$.

### 2.4.2 Number of undetected genes

Tauber and von Haeseler (2013) showed that we can use the estimation of parameter $m$ ($\hat{m}$) as an approximation of the number of detectable genes. From it we can compute the number of missed genes $g_0$

$$g_0 = \hat{m} - k \tag{2.6}$$

where $k$ is the number of detected genes .

### 2.4.3 Limiting relative frequencies of the genes

In his book "Probability, Statistics, and Truth", Richard von Mises (1939) discusses that under the frequentist interpretation of probability, we assume that for a given experiment with infinitely many trials, the proportion of trials in which a given event occurs will converge to a fixed value known as the limiting relative frequency.

Pitman (1995) described for the PSF the limiting relative frequencies $P_i$ of the genes in order of appearance for the finite condition ($\sigma = -\kappa < 0$ and $\theta = m\kappa$ with $m \in \mathbb{N}$)

$$P_i = (1 - W_1)(1 - W_2)...(1 - W_{i-1})W_i \tag{2.7}$$

where the $W_i$ are beta distributed independent random variables $\beta(1 + \sigma, \theta + i\sigma)$, with the special case of $W_m = 1$ and $W_i$ undefined for $i > m$.

### 2.4.4 Size-biased random permutation

Consider a sequence of positive numbers $\mathbf{x} = (x(1), x(2), ...)$ with finite sum $s = \sum x(i)$. A size-biased random permutation (SBRP) of $\mathbf{x}$ is a reshuffling of its elements $x(\alpha_1), x(\alpha_2)...$ (Pitman and Tran, 2015) where the probability of observing $x(i)$ in position $\alpha_1$ is $\mathbb{P}(\alpha_1 = i) = \frac{x(i)}{s}$ and for $c$ distinct indices $i_1, ...i_c$

$$\mathbb{P}(\alpha_c = i_c | \alpha_1 = i_1, ..., \alpha_{c-1} = i_{c-1}) = \frac{x(i_c)}{s - (x(i_1) + ... + x(i_{c-1})} \tag{2.8}$$

where an index $i$ tend to appear earlier in the permutation if it's "size" $x(i)$ is bigger.

In the context of RNA-sequencing, $x(i)$ represents the read counts for a gene $i$. A SBRP can be applied to the count table to change the order of the genes. Here, the SBRP is used to simulate the order of appearance. A gene with higher read counts is likely to be detected earlier, but not necessarily.

## 2.5 Conclusions

We have presented the Pitman Sampling Formula (PSF). The PSF has been previously applied to expression data (EST and RNA-sequencing). In this chapter we detailed numerous statistics developed for the PSF that we evalueated in Chapter 3 applied to study RNA-sequencing.

# 3 Evaluation of the Pitman Sampling Formula applied to RNA-sequencing

It has been more than 10 years since RNA-sequencing was introduced as a method to survey the entire transcriptome profile of an organism (Wilhelm et al., 2008). Since then, sequencing technologies have evolved and RNA-sequencing is now routinely used as the main method to survey RNA, especially in gene expression assays.

The laboratory procedure of library preparation for RNA-sequencing is well known and ready-to-use commercial products exist (i.e TruSeq RNA Library Prep Kit v2 by Illumina and NEBNext Ultra RNA Library Prep Kit by New England Biolabs). During the sequencing procedure tens to hundreds of million reads are produced, and yet the true number of the expressed genes is still unknown. For this reason, a trade-off between how much to sequence and how much information in terms of the number of detected genes we obtain is still an open question.

In this chapter we explore applications of the Pitman Sampling Formula (PSF) to study RNA-sequencing as a sampling problem, and in the process we aim to give insights to the questions:

- How many genes did we miss?
- How many additional genes do we detect if we sequence more?
- What is the expected number genes to be shared between replicates?

Here, we perform a comparative analysis with current methods to predict the number of missed genes and showed that the PSF performed better in cases where the sequence data is scarce (i.e. a spike-in experiment). Furthermore, the PSF performs

similarly well in cases where sequence data is plenty. The evaluation of statistics of the PSF with a benchmark data-set show its predictive power in the estimation of the number of additional genes in further sequencing experiments and the number of shared genes in replicate experiments.

## 3.1 Methods

We implemented all the statistics described in Chapter 2 as a series of functions and a package for the statistical environment R (R Core Team, 2017). In the following section we describe how the evaluation of such statistics was performed. For the evaluation we used public available benchmark data-sets.

### 3.1.1 Experimental data used in this work

We used two large data-sets to evaluate the PSF to the study of RNA-sequencing. The first data-set is the recount bioconductor package (Collado-Torres et al., 2017). Recount contains the count tables of 9,662 human RNA-sequencing samples from distinct sequencing projects. (Collado-Torres et al., 2017) analyzed all the samples with the same pipeline with the aim of providing to the scientific community a repository that combines many data-sets into one accessible website and R package.

Then, we used a benchmark RNA-sequencing data-set produced by the SEQC/MAQC-III Consortium (Su et al., 2014). From this data-set we used the data generated from two reference samples that were sequenced in six different facilities for the Illumina platform (Table 3.1, figure 3.1) The two reference samples are:

- Sample A: Universal Human Reference RNA which is comprised of a mix of RNA of 10 human cell lines (Agilent Technologies) plus the addition of a known concentration of synthetic RNA sequences from the ERCC (labled sample E)

- Sample B: FirstChoice® Human Brain Reference RNA which is comprised of reverse-transcribed human brain RNA (ThermoFisher Scientific) plus the addition of a known concentration of synthetic RNA sequences from the ERCC (labled sample F)

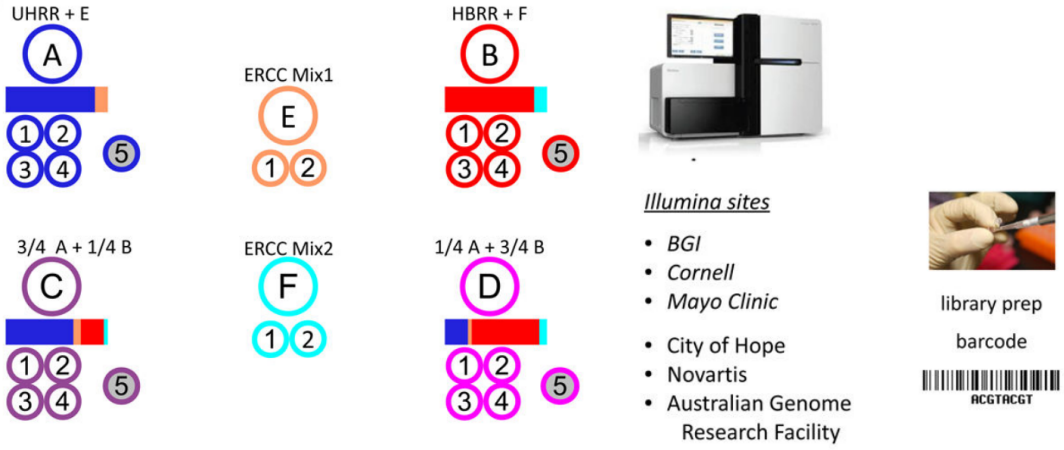



Figure 3.1: Samples sequenced by the SEQC/MAQC-III Consortium. Samples A and B consist of two human RNA reference samples plus a fraction of samples E and F respectively. Samples C and D are mixes of A and B in 3:1 and 1:3 ratios. Finally, samples E and F consist of known synthetic RNA sequences from the ERCC that are used as input control. For samples A, B, C and D four libraries were prepared in each sequencing facility (numbers 1-4). A fifth library was prepared by the vendor, in this case Illumina (number 5). The figure is a modified version from Su et al. (2014)

Table 3.1: The six sequencing facilities that generated sequencing data using the Illumina technology. The count matrices are available in the bioconductor package "seqc".

| Site (abbreviation) | Number of mapped reads (in millions) | |
|---|---|---|
| | Sample A | Sample B |
| Australian Genome Research Facility (ARG) | 464.91 | 476.73 |
| Beijing Genomics Institute (BGI) | 485.24 | 475.10 |
| City of Hope (COH) | 340.83 | 322.17 |
| Weill Cornell Medical College (CNL) | 254.54 | 291.47 |
| Mayo Clinic (MAY) | 234.16 | 266.46 |
| Novartis (NVS) | 277.95 | 273.22 |
| TOTAL | 2,057.63 | 2,105.14 |

The SEQC/MAQC-III Consortium released the count tables along with the raw SRA files. As the Consortium has already summarized all the read counts into frequency tables, instead of sampling reads from the raw data and redo the summarization analysis, we used subSeq (Robinson and Storey, 2014) to sample the counts as if they where reads. subSeq uses a count table $X$ and the proportion $p$ of the subsample in the interval $(0, 1]$. Then, it generates a subsampled count table $Y$ such that $Y \sim B(X, p)$. The assignment of each mapped read to a gene is a deterministic process which solely depends on the annotation file. For this reason, sampling counts with subSeq is equivalent to the common approach of sampling from the read alignment files (Robinson and Storey, 2014). In contrast to sampling from alignment files, subSeq run-time and usage of computing resources is negligible.

### 3.1.2 Estimation of the number of genes expressed in a transcriptome

Inspired by the research in the field of ecology, in which the number of unobserved species is estimated, Garcia-Ortega and Martinez (2015) who aimed to estimate the number of genes expressed in a given RNA-sequencing experiment. Garcia-Ortega and Martinez based their work in non-parametric estimators, specifically in the work of Anne Chao C.H. Chiu and et al. (2014), who developed several estimators for the number of missing or undetected classes (the classes can be genes, alleles or species).

Similar to the approach by Anne Chao, Garcia-Ortega and Martinez use the assumption that the genes with low read counts contains the most information regarding the number of genes with zero reads. With this idea they developed the $h_6$ or harmonic estimator of degree 6 of $g_0$.

$$h_6(g_0) = \frac{6}{10} \frac{g_1^2}{H(g_2, ..., g_6)} \tag{3.1}$$

where $H(g_2, ..., g_6)$ is the harmonic mean of $g_2, ..., g_6$ and $g_i$ represents the number of genes detected with exactly $i$ reads with $g_0$ being the number of undetected genes.

### Computing the expected number of shared genes between replicate experiments

Let us consider two replicate experiments $A$ and $B$ with $n_A$, $n_B$ mapped reads and $k_A$, $k_B$ detected genes respectively.

We assume that if experiments $A$ and $B$ are replicates, they share a "common RNA pool". For example, if we consider experiments $A$ and $B$ as technical replicates, during library preparation each replicate is prepared from a different aliquot of the same biological material. Here, the starting material from which the aliquots are taken is the common RNA pool.

Given the assumption of common RNA pool, two properties arise: 1. the maximum number of possible detectable genes is the same and 2. the limiting relative frequencies of the detectable genes is the same.

Let $g = \{g_1, g_2, ...g_m\}$ be the set of all detectable genes in experiments $A$ and $B$. For each gene $g_i$ with $i = 1, 2, ..., m$, we want to know the probability of it being present in both replicates. To do so, first we need to compute the limiting relative frequencies of the genes. With the assumption that $A$ and $B$ are two independent replicates, we use equation 2.1 to estimate the model parameters $\theta$ and $\sigma$ using both replicates as input

$$P[k_A, k_B, \mathbf{a}_n, \mathbf{b}_n] = \left( n_A! \frac{\prod_{i=1}^{k_A-1}(\theta + i\sigma)}{(\theta+1)_{n_A-1}} \prod_{j=1}^{n_A} \frac{(1-\sigma)_{j-1}^{a_j}}{j!^{a_j} a_j!} \right) \left( n_B! \frac{\prod_{i=1}^{k_B-1}(\theta + i\sigma)}{(\theta+1)_{n_B-1}} \prod_{j=1}^{n_B} \frac{(1-\sigma)_{j-1}^{b_j}}{j!^{b_j} b_j!} \right) \tag{3.2}$$

where $\mathbf{a}_n$ is the occupancy vector of replicate $A$ and $\mathbf{b}_n$ is the occupancy vector of replicate $B$.

Then, with the estimated parameters, we compute the limiting relative frequencies $P_i$ of the genes in order of appearance using equation 2.7. The limiting relative frequencies of the genes as described by Pitman (1995) are for the genes in *order of appearance*. To emulate sequential sampling in experiments $A$ and $B$, we performed a size-biased random permutation (SBRP, eq. 2.8) of the genes. Here, genes with large read counts are likely to appear earlier, but not necessary. With the SBRP, we get the order of appearance of each gene in replicates $A$ and $B$ as we consider them to be independent sampling events. This means that we may a different order of appearance of the same gene in each replicate.

Finally, we estimate the probability of each gene being in both replicates. For exam-

ple, let $g_i$ be a gene with index $\alpha = 1$ as it appeared first in replicate $A$ and index $\beta = 15$ as it appeared $15^{th}$ in replicate $B$. Consequently, $P_{g_i,\alpha}$ is the relative frequency of $g_i$ in replicate $A$ and $P_{g_i,\beta}$ is the same but for replicate $B$. Hence, the probability of $g_i$ being in both replicates can be computed as

$$\mathbb{P}(g_i \in A \cap B) = (1 - (1 - P_{g_i,\alpha})^{n_A}) \times (1 - (1 - P_{g_i,\beta})^{n_B}) \tag{3.3}$$

where $(1 - P_{g_i,\alpha})^{n_A}$ is the probability that $g_i$ is not detected by any of the $n_A$ reads in replicate $A$ in, and $(1 - P_{g_i,\beta})^{n_B}$ is the probability that $g_i$ is not detected in replicate $B$. Consequently, the expected number of shared genes between two replicate experiments can be calculated as the sum of the probability of every detectable gene being in the intersection

$$\mathbb{E}(I_{A,B}) = \sum_{i=1}^{\hat{m}} \mathbb{P}(g_i \in A \cap B) \tag{3.4}$$

where $\hat{m}$ is the estimated number of genes present in the sample.

### 3.1.3 Experimental design

In order to evaluate the PSF and its statistics applied to RNA-sequencing we defined three scenarios of RNA-sequencing experiments. Each experiment differ on the number of read counts to examine the effect of sequencing depth on the PSF.

The first case depicts a spike-in experiment ($E_{spike}$), which is commonly used to assess the viability of the sample. $E_{spike}$ was set as a sample of one million reads. The second case exemplifies experiments with low coverage. To assign the sequencing depth we made use of the recount project. We downloaded the sequencing depth information of 100 RNA-sequencing experiments and set the low-coverage experiment, $E_{low}$, to be $5^{th}$ percentile, that is equivalent to 10 million reads. The third case exemplifies an average RNA-sequencing experiment. Here, as before we use sequencing depth from the recount project and assigned the read count of the average experiment, $E_{av}$, to the $50^{th}$ percentile which corresponds to 35 million reads.

#### Estimating the number of non-detected

We created 100 independent replicates of $E_{spike}$, $E_{low}$ and $E_{av}$ by sampling with subSeq 1, 10 and 35 million reads respectively, from samples A and B of the SEQC/MAQC-III

data-set. Here, we combined the data from the six sequencing facilities. Sample A contain a total of 2,057.63 million reads and sample B 2,105.14 million reads.

Then, for each replicate we estimated the PSF model parameters using equation 2.1 and then used $\hat{m}$ as the estimated number of detectable genes to then calculate the number of missed genes $g_0$ (eq. 2.6). In parallel, for each replicate, we computed the number of undetected genes using equation 3.1 to compare both estimators. The ground truth of the total number of detectable genes is not known. We use the fact that the SEQC/MAQC-III data-set is large enough and used the number of detected genes in sample A (24,718 detected genes) and B (24,590 detected genes) as a proxy of the the total number of detectable to evaluate the $g_0$ and $h_6$ estimators.

### Estimating the number additionally detected genes with increase sequencing depth

**Simulation study.** We evaluated the estimation of the expected number of genes (eq. 2.4) using simulated data. We selected arbitrary values for the PSF parameters ($\theta = 5,000$ and $\sigma = -0.2$) which result in $m = 25,000$ genes. We then simulated experiments with 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000 million reads using the Hoppe urn and counted the number of detected genes, repeating the process 100 times for each sequencing depth. In parallel, for each simulated experiment, we estimated the model parameters using the PSF (equation 2.1) and computed the expected number of genes using equations 2.4 with 95% confidence intervals, under the assumption that number of genes, $k$, is normally distributed (the variance of the expected number of genes was computed using equation 2.5).

**Evaluation with experimental data.** After the simulation study, we evaluated the estimation of the number of additionally detected genes using experimental data. We created 100 independent replicates of $E_{spike}$, $E_{low}$ and $E_{av}$ for each of the six Illumina sequencing sites (table 3.1). For each replicate we estimated the model parameters for the PSF using equation 2.1. Then, to increase the sequencing depth we performed addition steps of **one million reads** (sampled with subSeq) to each replicate until 100 million extra reads were added. At each addition step, the number of additional genes was estimated by computing the expected number of genes with the new sequencing

depth (eq. 2.4). At the same time, we counted the number of additionally detected genes from the additional data.

#### Estimating the number of shared genes between replicate experiments

We evaluated the performance of the expected number of shared genes between two replicate experiments (eq. 3.4). For the evaluation we defined two type of replicates:

- **Proper replicates:** which are replicates from the same tissue or condition. Here, we tested technical replicates, however this category may be also applicable for biological replicates.
- **Improper replicates:** which are replicates that are NOT from the same tissue or condition. This replicates can be caused by human error when a sample is either mislabeled or handled incorrectly. Here, we used as improper replicates samples from different tissues.

For proper replicates we compared 100 pairs of different replicates of sample A (A vs A) or sample B (B vs B) from the SEQC/MAQC-III. Sample A constitutes RNA from 10 human cell lines while sample B is composed of RNA extracted only from neural tissue. As improper replicates we compared 100 pairs of replicates of sample A and B (A vs B). For each pair we computed the expected number of shared genes using equation 3.4 and, at the same time, we counted the number of shared genes observed from the data.

## 3.2 Analysis and Results

### 3.2.1 Estimating the number of non-detected genes

First, we addressed the question "How many genes were missed" for a given sequencing experiment. Tauber and von Haeseler (2013) showed parameter $m$ of the PSF, which is the absolute ratio between the model parameters $\theta$ and $\sigma$ (defined as $\theta = m|\sigma|$), can be used as the number of genes present in a given sample from which $k$ where detected. Here, instead of fixing the value of $m$ to the number of genes present in the annotation file, we estimated $m$ from experimental data ($E_{spike}$, $E_{low}$ and $E_{av}$) together with the model parameters $\theta$ and $\sigma$.

We evaluated the estimated of the number of undetected genes, $g_0$ (eq. 2.6), and the harmonic estimator of degree 6 $h_6$ (eq. 3.1 Garcia-Ortega and Martinez (2015)) using two benchmark samples: sample A - the universal human reference RNA (2,057.63 million reads) and sample B - human brain reference RNA (2,105.14 million reads). With this read count 24,718 genes were detected in sample A and 24,590 in sample B. We used For the evaluation we used 100 replicates of $E_{spike}$, $E_{low}$ and $E_{av}$ as input.

Figure 3.2 shows the results of the evaluation. The $g_0$ estimator overestimated of number of undetected genes in five of the six evaluation cases (figure 3.2, blue box plots). In contrast, the $h_6$ estimator underestimates the number of undetected genes in all six cases (figure 3.2, red box plots). Interestingly, the $g_0$ estimator performed similarly regardless of the difference in sequencing depth, having a deviation of $\sim 1,000$ genes which account to 4% of the total number of detected genes in both sample A and B. This result of of particular interest because it shows the predictive power of the PSF. When the model parameters are accurately estimated, the computing the number of missed genes is not affected by the sequencing depth.

For the case of the $h_6$ estimator, we observed and improvement in the estimation of the number of missed genes with the increase in sequencing depth (deviation of 16% $\rightarrow$ 8.1% $\rightarrow$ 3.3% for $E_{spike}$, $E_{low}$ and $E_{av}$ respectively). This estimator uses the information of the genes detected with exactly one to six reads to compute the number of genes with zero reads, thus with additional sequenced reads the $h_6$ estimator has more data to estimate the missed genes.

For us, the most relevant result comes from the evaluation of $E_{spike}$. A spike-in experiment is generally done are part of the experimental design. In contrast, additional sequencing experiments after an experiment that yielded 35 million reads is not common. The $g_0$ estimator performed better in the with $E_{spike}$ while both the $g_0$ and the $h_6$ estimator performed similarly good with $E_{av}$.

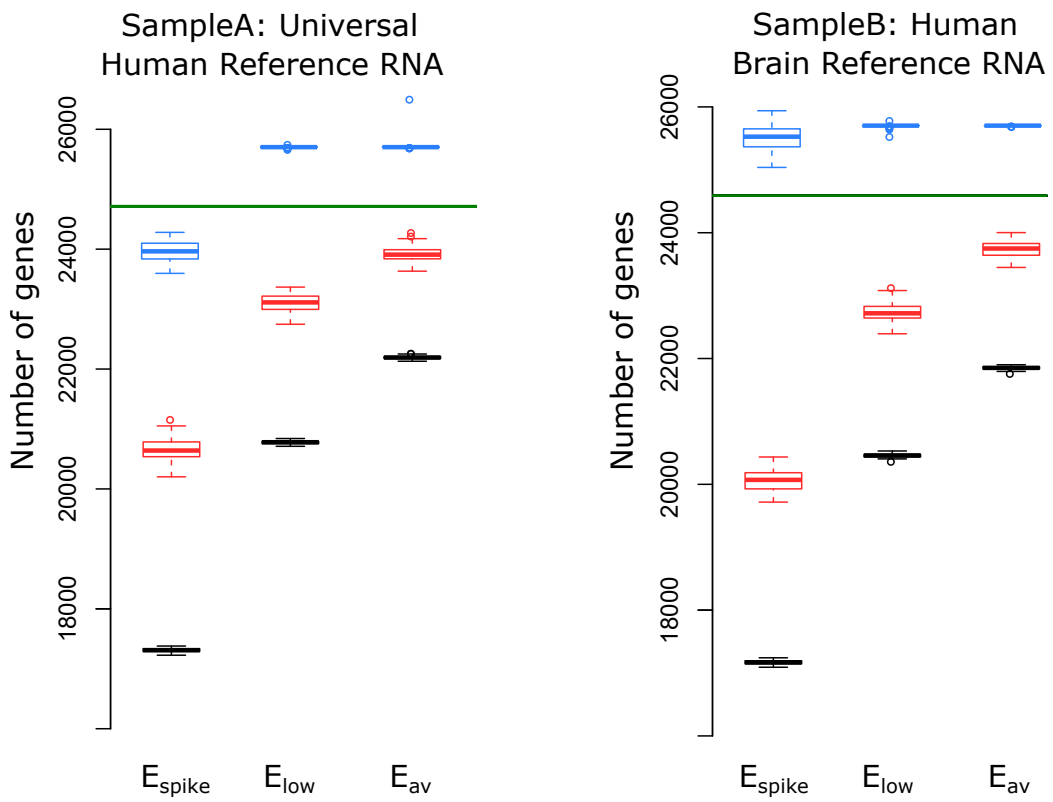



Figure 3.2: Estimation of the number of missed genes for three different experiments. We computed the number of missed genes using the $g_0$ (in blue) and the $h_6$ estimators (in red). We evaluated both estimators using SEQC sample A (left panel) and sample B (right panel). The green line represents the true number of detectable genes that we aim to estimate. In black is shown the number of detected genes in $E_{spike}$, $E_{low}$ and $E_{av}$ to which applied the $g_0$ and $h_6$ estimators. The $g_0$ estimator performs similar regardless of the sequencing depth, having more accurate estimates in four of the six evaluation cases ($E_{spike}$ and $E_{lo}$). In contrast the $h_6$ estimator improved performance with the increase in sequencing depth, with better performance in $E_{av}$

### 3.2.2 Additional detected genes when increasing the sequencing depth

Following the estimation of the number of missed genes, we investigated the the predictive power of the PSF for the estimation of the additionally detected genes when increasing the sequencing depth.

First, we performed a simulation study where the model parameters of the PSF are known. Here, we used arbitrary values for $\theta = 5,000$, $\sigma = -0.2$ and consequently $m = 25,000$. We used the Hoppe urn to simulate experiments of a wide range of sequencing depths ranging from one to one thousand million reads. For each simulated experiment we counted the number of detected genes and compared them to the prediction which will be affected by the estimation of the model parameters. Figure 3.3 show the results of the simulation study. We showed that the estimation of the number of detected genes is accurate regardless of the sequencing depth. Table 3.2 shows the estimated model parameters for each of the evaluated sequencing depths. These results show that when the model parameters are accurately estimated, the predicted number genes to be detected is accurate invariant of the sequencing depth.

Then, we evaluated the predicted number of additionally detected genes using experimental data. Here, the value of the model parameters is unknown to us and we estimate them from the data. We used $E_{spike}$, $E_{low}$ and $E_{av}$ as our starting point and then we added sequentially one million reds until additional 100 million reads were added to the initial sample. Figure 3.4 shows the results of the evaluation of 600 independent replicates (100 per Illumina sequencing site, table 3.1). We can observe that when the increase of sequencing depth is low (e.g. one to ten million reads) the estimated number of genes is quite accurate for all studied cases. After this point, when using $E_{spike}$ we underestimated the number of detected genes for all the analyzed experiments. From $E_{low}$ we can see that the predicted number of genes improves as more data is available for the parameter estimation. This leads to very accurate predictions when using $E_{av}$.

It is important to notice that, for each of the six sequencing facilities we observe similar results. These show the reproducibility of our analysis. Moreover, we can detect that replicates from some sequencing facilities (e.g Australian Genome Research and Mayo Clinic) are strikingly similar no only in the number of detected genes but also in the estimation of the model parameters (table 3.3). These observation lead us

to the idea that the model parameters could be sufficient statistics compare library complexity between replicate samples. Future work can be focused on studying this observed property.

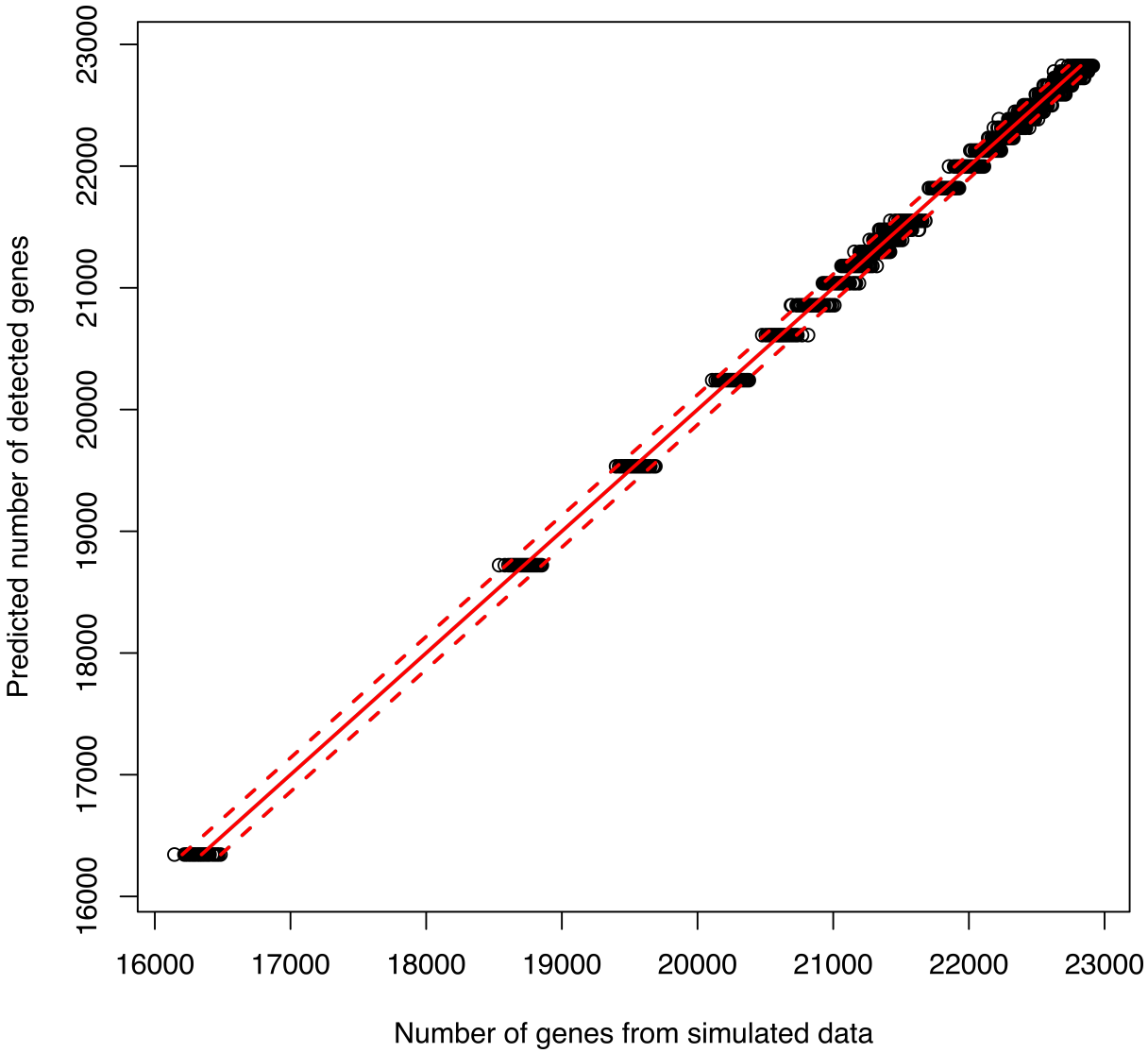



Figure 3.3: Number of detected genes from 100 simulated replicates ($\theta = 5,000$ and $\sigma = -0.2$) for several sequencing depths (points) and the expected number of detected genes with 95% confidence intervals (shown in red). We show that the predicted number detected genes is accurate invariant of the sequencing depth.

Table 3.2: Estimated model parameters from the simulated data. The mean of 100 replicates is shown. The RNA-sequencing simulations were performed with model parameters $\theta = 5,000$ and $\sigma = -0.2$.

| Million reads | $\theta$ | $\sigma$ | Million reads | $\theta$ | $\sigma$ |
|---|---|---|---|---|---|
| 1 | 4,992.017 | -0.1996553 | 200 | 5,013.609 | -0.2006700 |
| 5 | 5,000.511 | -0.1999795 | 250 | 4,998.589 | -0.1999543 |
| 10 | 5,008.335 | -0.2003853 | 300 | 5,007.509 | -0.2003300 |
| 20 | 4,996.311 | -0.1997719 | 350 | 5,003.883 | -0.2001586 |
| 30 | 5,001.812 | -0.2000317 | 400 | 4,996.795 | -0.1998647 |
| 40 | 4,991.684 | -0.1995250 | 450 | 4,994.426 | -0.1996358 |
| 50 | 5,013.709 | -0.2006679 | 500 | 4,994.291 | -0.1996858 |
| 60 | 5,005.281 | -0.2004002 | 600 | 5,004.978 | -0.2001387 |
| 70 | 5,004.432 | -0.2002856 | 700 | 5,000.049 | -0.2000294 |
| 80 | 4,999.760 | -0.1999971 | 800 | 4,996.917 | -0.1998188 |
| 90 | 5,011.669 | -0.2007347 | 900 | 4,993.597 | -0.1997004 |
| 100 | 4,998.454 | -0.1997877 | 1000 | 5,013.019 | -0.2005652 |
| 150 | 4,998.651 | -0.1999796 | | | |

Table 3.3: Average estimation of the model parameters for $E_{spike}$, $E_{low}$ and $E_{av}$

| | | AGR | NVS | MAY | BGI | CNL | COH |
|---|---|---|---|---|---|---|---|
| $E_{spike}$ | $\theta$ | **6,032.251** | 5,970.093 | **6,021.175** | 6,105.827 | 5,662.884 | 6,096.075 |
| | $\sigma$ | **-0.25516** | -0.24905 | **-0.25199** | -0.25929 | -0.22688 | -0.25876 |
| $E_{low}$ | $\theta$ | **5,586.655** | 5,610.474 | **5,571.806** | 5,634.167 | 5,339.665 | 5,616.843 |
| | $\sigma$ | **-0.21623** | -0.21894 | **-0.21373** | -0.21961 | -0.19802 | -0.21722 |
| $E_{av}$ | $\theta$ | **5,466.992** | 5,463.437 | **5,470.614** | 5,506.387 | 5,279.410 | 5,503.880 |
| | $\sigma$ | **-0.20679** | -0.20738 | **-0.20594** | -0.20979 | -0.19279 | -0.20888 |

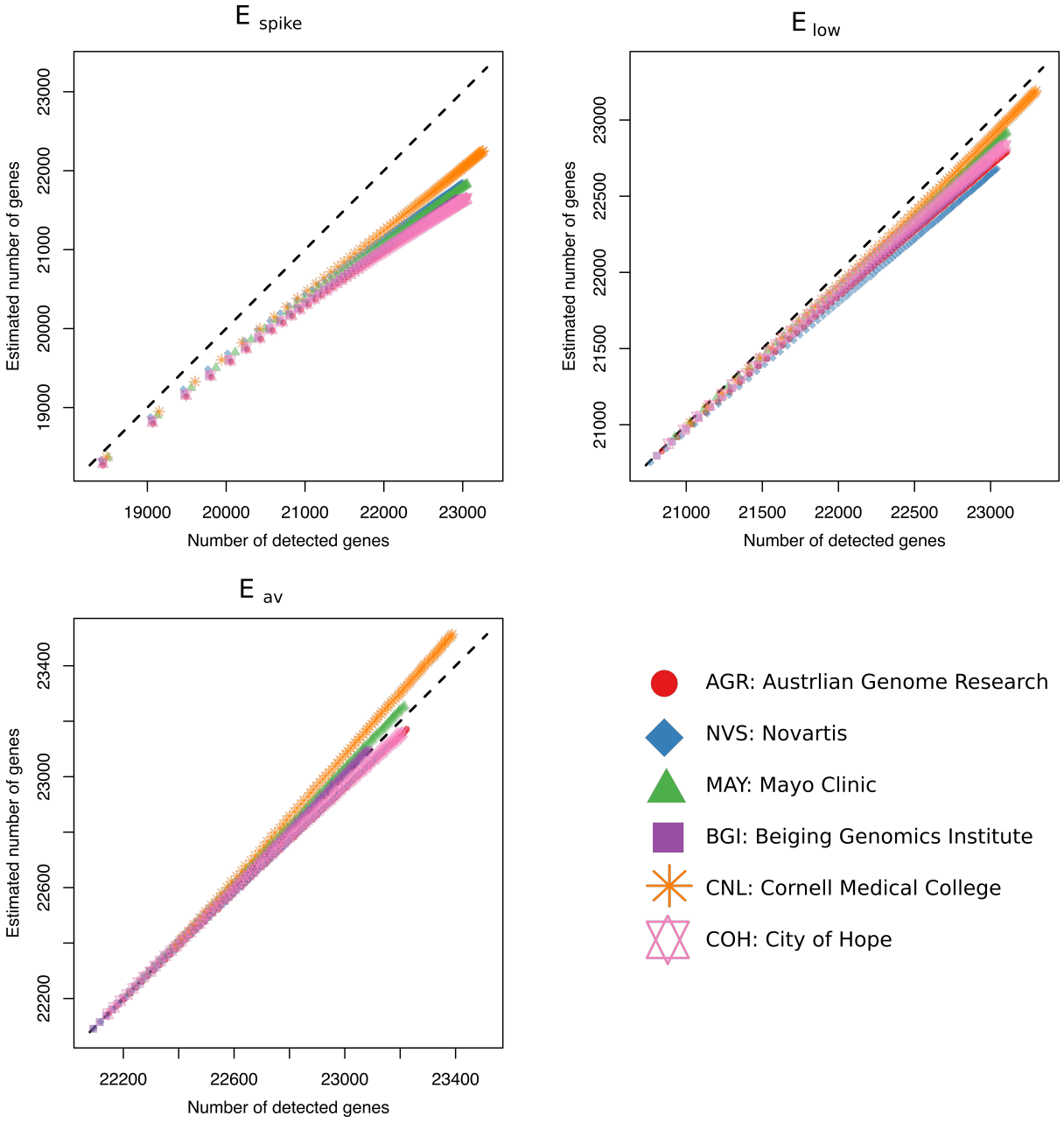



Figure 3.4: We used $E_{spike}$, $E_{low}$ and $E_{av}$ to evaluate estimation of the number of detected genes for increase in sequencing depth. The evaluation was performed using data from six sequencing sites from table 3.1. The average of 100 replicates for each sequencing site is shown. To $E_{spike}$, $E_{low}$ and $E_{av}$ one million reads were sequentially added and the number of genes detected genes was estimated and compared to the data. The dashed line represents the identity line, below the line means we underestimated the number of genes and above the line we overestimated. Note that estimation of the number of detected genes is very similar for different sequencing sites and thus, the points in the graph are overlapped (e.g AGR, BGI and COH in $E_{spike}$ and AGR, NVS and COH in $E_{av}$).

### A trade-off between sequencing depth and gene detection

We used the results form the two previous sections to develop a tool that estimates the sequencing costs of future sequencing experiments. As input, we require the count table $T$ of an RNA-sequencing experiment, the sequencing cost per million reads $CPM$ and proportion of missed genes one intend to detect $NG$.

First, we estimate the PSF model parameters from input count table $T$ and then we estimate the number of undetected genes $g_0$ (eq. 2.6). Then, we compute the number of genes the user intend to detect in a further sequencing experiment $g_{new} = NG \times g_0$ rounded down to the previous integer. Following this, we compute the necessary sequencing depth to detect $g_{new}$ to finally, use the user provided cost per million reads $CPM$ to estimate the sequencing cost. This tool, will provide the user the necessary information to calculate a trade-off between increase sequencing depth and return of investment in the number of newly detected genes.

To test this idea, we used public available sequencing cost from the in-house sequencing facility: The Vienna Biocenter Core Facilities GmbH (VBCF `https://www.vbcf.ac.at/facilities/next-generation-sequencing/`) to estimate the sequencing cost three experiments in which we aim to detect 20%, 50% and 70% of the missed genes. We calculated the sequencing cost for 125bp paired end reads in a HiSeq 2500 to be approximately $CPM \sim 11$ Euros.

Here, we used as input an RNA-sequencing experiment of two million reads. From this experiment 17,460 genes were initially detected. We then calculated the number of undetected genes $g_0 = 7,540$. Figure 3.5 shows the per million cost if we aim to detect 20%, 50% and 70% of the 7,540 missed genes (which correspond to detect 1,508, 3,770 and 5,278 additional genes respectively) compared to the sequencing cost. Here, we can observe that the number of additional genes decreases very fast, which translates to a very high cost when aiming to detect large proportions of the initially undetected genes. This can be seen in our example experiment: in order to detect 1,508 additional genes (20%) we need to sequence five million reads and spend 55 Euros; if we aim to detect 3,770 additional genes (50%) then we need to sequence 63 million reads ans spend over 693 Euros. Finally if we aim detect 5,278 additional genes (70%) we will need to sequence an additional 825 million reads. This is approximately four Illumina HiSeq2500 lanes for a single experiment at a cost of over 9,000 Euros.

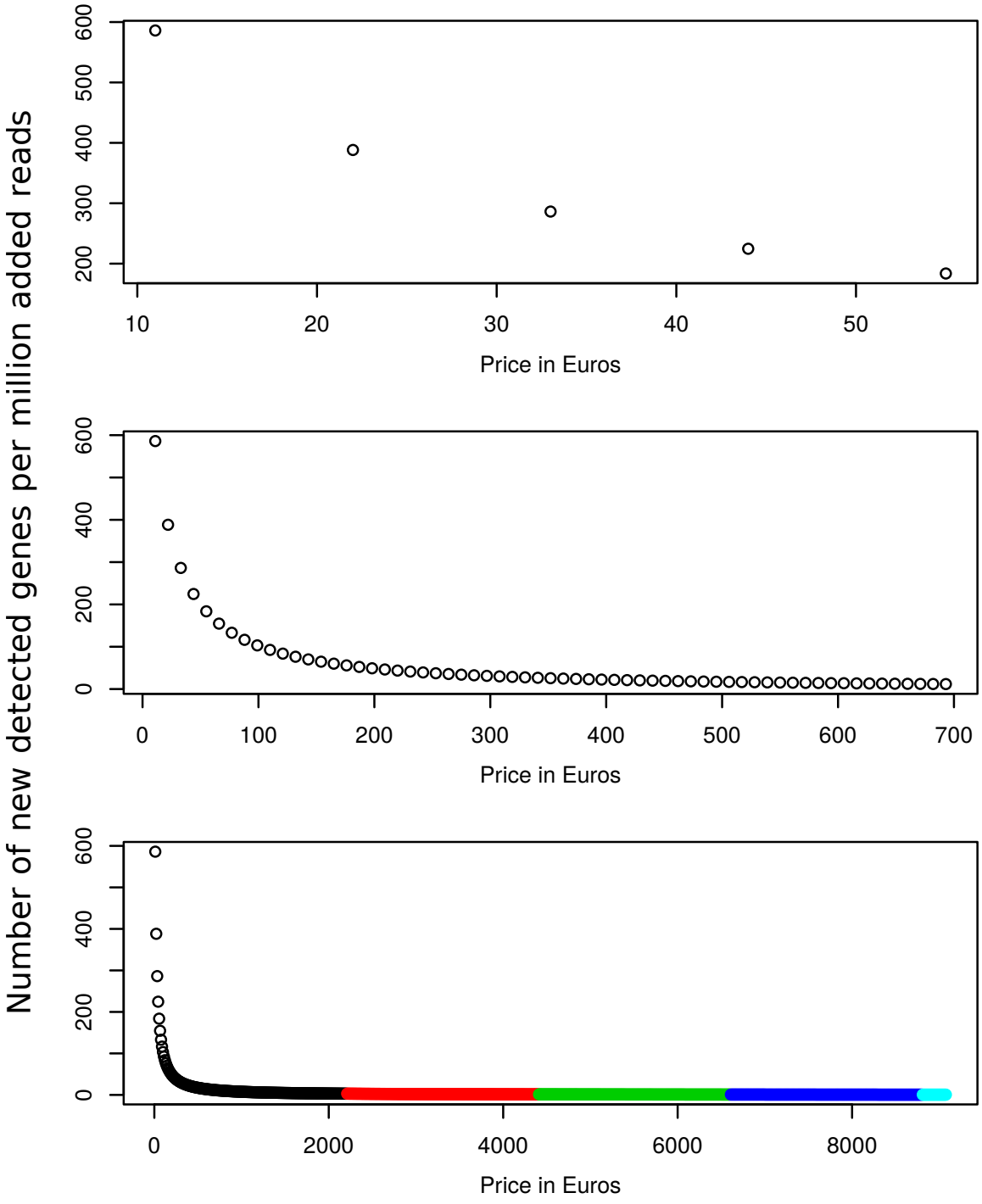



Figure 3.5: **Estimated cost of detecting additional genes**. Based on the price of per million reads for the VBCF, we calculated the price for detecting 20%, 50% and 70% of the 7,540 missed genes for our example experiment (top, middle and lower panel respectively). Here, we show the estimated number of additional detected genes by the addition of one millions sequenced reads. In the lower panel we color coded every additional 200 million sequenced reads.

### 3.2.3 Assessment of the number of shared genes between replicate experiments

Here we present the evaluation of a new statistic that estimates the expected number of shared genes between replicate experiments (eq. 3.4). Here, we used technical replicates for the evaluation however, it may also work with biological replicates.

We evaluated our statistic using the fact that the SEQC/MAQC-III Consortium sequenced the same biological samples in six independent sequencing sites and prepared four independent libraries (Table 3.1). First, we evaluated our statistic by comparing each replicate to itself. The best case scenario here is that the difference between the expected and observed intersection in a self-comparison is small. In our comparisons we computed the relative difference to the expected value ($d$) as

$$d(x, y) = (Obs(x, y) - Exp(x, y))/Exp(x, y) \tag{3.5}$$

where $Obs(x, y)$ is observed the number of shared genes between replicates $x$ and $y$ and $Exp(x, y)$ is the expected number of shared genes for the same replicates. We selected 100 replicates at random from the six sequencing facilities and performed a self-comparison $d(x, x)$. Then, we selected at random 100 pairs of replicated from the six sequencing facilities and estimated here $d(x, y)$ where $x$ and $y$ are proper replicates. Finally, we assessed 100 pairs of improper replicates by comparing replicates from sample A (a mix of 10 human cell lines) to replicates of sample B (RNA from human brain). Figure 3.6 show the results of the three different comparisons.

For the self-replicate comparisons we noted that the relative difference between the expected and observed number of genes $d$ was small and centred near zero. This first result shows that the proposed statistic has the desire property that when comparing self-replicates the expected and observed number of shared genes is very similar. Then, when comparing proper and improper replicates we observed that for each case $d$ was larger than zero meaning that our statistic under-estimate the number of shared genes. Moreover, we observed a discernible difference between proper and improper replicates. Figure 3.6 shows the distribution of $d$ for self-comparisons, proper and improper replicates. It can be observed that our statistic may be of use to detect abnormal replicates.

We tested this case scenario as follows: we selected six replicates (three from sample

A and two from sample B and one a mix A-B in a 3:1 ratio) and computed all the pairwise expected and observed number of shared genes. Then we computed the relative difference $d$ and performed a hierarchical clustering with the aim of distinguish the three groups based in our statistic. Figure 3.7 shows the clustering result. For this test scenario we were able to separate the replicate experiment that originated from different tissues (cancer cell lines and brain). Moreover, the mixed sample clusters with A, as it contains a greater proportion of sample A.

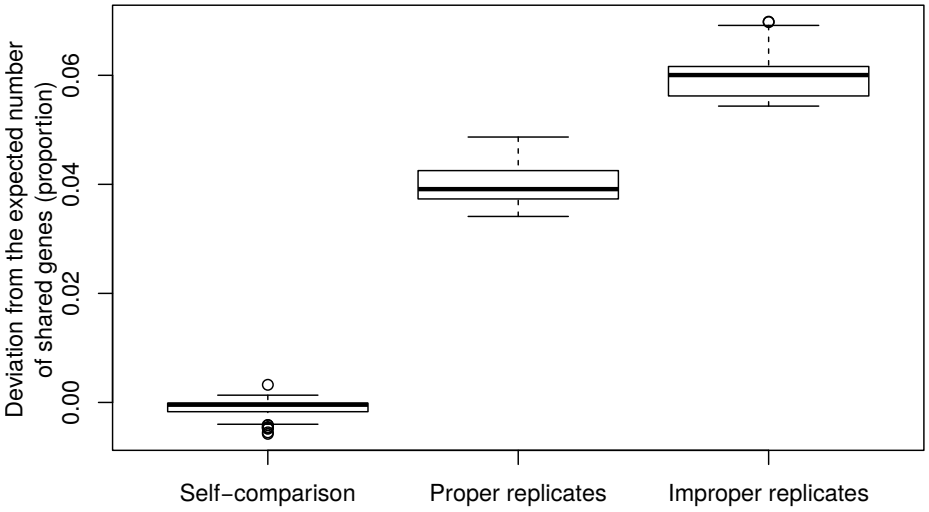
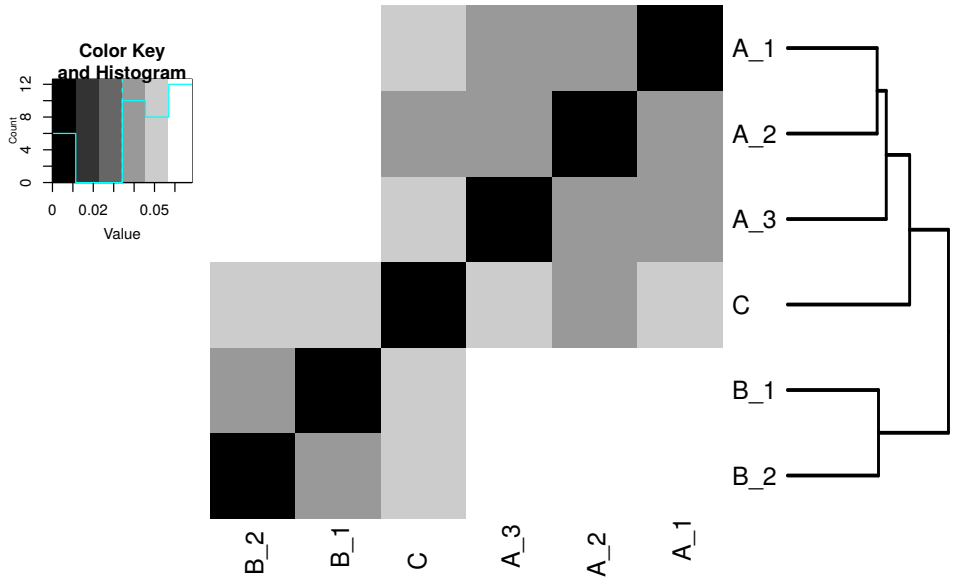



Figure 3.6: Deviation from the expected intersection of two replicate samples. The plot shows the relative difference to the expected value for self-replicate comparison, proper replicates and improper replicates.




Caption for figure 3.7 in the next page.

Figure 3.7: **Clustering of replicate experiments based on** $d$. Six samples were randomly selected from the SEQC/MAQ-III data-set (three from sample A, cancer cells, two from sample B, human brain and one mix AB in a 3:1 ratio), we then used $d$ (eq. 3.5) as the metric used to performed a hierarchical clustering with the aim of distinguishing distinct replicates.

## 3.3 Discussion

RNA-sequencing is a widely used method to study the presence and abundance of RNA molecules. There are several groups who work in the statistics to compare the relative abundance of gene models to establish which genes show differential expression (Anders and Huber, 2010; Love et al., 2014; Conesa et al., 2016).

We, as others (Sims et al., 2014; Busby et al., 2013; Tauber and von Haeseler, 2013; Garcia-Ortega and Martinez, 2015), are interested not in differential expression but in questions related to the sampling properties of RNA-sequencing which lead to questions such as: *"How many genes did we miss?"* and *"What is the discovery rate (as in additionally detected genes) in follow-up experiments?"*.

It is well known that with increasing sequencing depth more genes can be detected. Sequencing experiments similar to Su et al. (2014), who sequenced the same biological sample with thousand of millions of reads are not routinely performed. Nevertheless, scientist are still interested in knowing how much may they detect if they do so.

Here, we propose the use of the Pitman Sampling Formula (PSF), a model derived from the field of population genetics that is general enough that we can apply it to the study of RNA-sequencing. With the PSF, we simplify the sampling procedure of RNA-sequencing by eliminating the necessity to introduce extra parameters and experimental observations to account for library preparation and all its sub-steps (Griebel et al., 2012). The PSF has two parameters to describe the distribution of reads among genes and has previously been used to model expression sequence tags (EST) (Lijoi et al., 2007) and RNA-sequencing (Tauber and von Haeseler, 2013).

In this chapter, we present three distinct applications of the PSF tailored to the study of RNA-sequencing in terms of a sampling problem. First, we evaluated the estimation

of the number of genes that were not detected in an RNA-sequencing experiment. The PSF performed consistently invariant of the sequencing depth. This means that the smaller sampled we use ($E_{spike}$) had enough information for us to estimate the model parameters accurately and in consequence the estimation of the number of missed genes. When compared to current methods, the PSF performs better when the data is scarce and similarly well otherwise. Thus, the PSF can be used during the experimental design to make inform decisions on the appropriate sequencing depth necessary for each particular problem.

These results come in hand with the second evaluation: estimating the number of additionally detected genes when increasing the sequencing depth. Here, we showed the predictive power of the statistics developed for the PSF and computed the cost-benefit of further sequencing experiments. We use the per-million reads sequencing cost to estimate the price of future experiments back-to-back with the amount of information gained as the number of detected genes. Currently, one read is sufficient for us to state that a genes has been detected. In applications such as differential gene expression, arbitrary threshold are imposed on the minimum number of mapped reads (Conesa et al., 2016) and genes below such threshold are discarded. To cope with these requirements we are working on a generalization of this problem where, we estimate the number of detected genes with a minimum number of mapped read.

Finally, we provide a statistic for the expected number of shared genes between replicate experiments. With this statistic we were able to detect abnormal replicates that may occur, for example, due to human error. Detecting this type of replicates in an early stage of the analysis is crucial, as abnormal replicates may lengthen the analysis, disrupt the results and thus, lead to erroneous conclusions. Our method uses as input count matrices which are a standard in differential gene expression and it is fast. The inclusion of out method as an extra step of the analysis pipeline comes at no cost and can save time if a bogus replicate is detected.

With the accelerated drop in sequencing costs, the time spent in the bench to prepare the libraries and the downstream analysis after sequencing can still be a limiting factor. This work offers experimental scientists powerful tools that may play an important role in both experimental design and assessment of quality control.

## 3.4 Conclusions

The Pitman Sampling Formula (PSF) is a general enough sampling model that we applied it to study RNA-sequencing. The statistics developed for the PSF can be useful for experimental biologist, so that they can make educated choices in the matter of how much is worth sequencing. We evaluated the PSF model with benchmark data-sets but it can be applied to any routinary RNA-sequencing experiment. Further investigation is necessary to test the PSF model with newly developed techniques such as single-cell RNA-sequencing. With the advent of newly sequenced genomes we proposed the use of the PSF model to assess the level of completeness of transcriptome assemblies by using the statistics of number of missed genes and discovery rate.

# 4 RNACountSim: fast simulation of RNA-Sequencing experiments.

## 4.1 Introduction

RNA-sequencing is nowadays widely used beyond the genomics community and it is now a standard tool in life sciences research (Conesa et al., 2016), with differential gene expression being one of its main applications. Since 2010, we have seen a surge in the development of tools to assess differential gene expression (Hardcastle and Kelly, 2010; Anders and Huber, 2010; Robinson et al., 2010; Love et al., 2014) and with the advancements in sequencing technologies we expect to see even more.

Simulated data is widely used in bioinformatics during the development of tools (Escalona et al., 2016). For RNA-sequencing, the currently available tools can be divided into two categories: tools that simulate reads from a given sequencing technology (mostly Illumina) and tools that simulate read count matrices that describe the reads counts for each gene.

Examples of the first category are Polyester (Frazee et al., 2015) and the FLUX simulator (Griebel et al., 2012). Polyester uses the negative binomial as the gene expression model for the number of reads per transcript, while the FLUX simulator uses a modified Zipfs Law (Ogasawara et al., 2003; Furusawa and Kaneko, 2003) to randomly assign expression levels to genes models. Both tools simulate steps of the library preparation that may affect the overall analysis (e.g. RNA fragmentation). Moreover, the simulated reads need to be mapped to a reference genome and then summarized into count matrices. The processing can take between minutes to days depending on the number of simulated experiments and computing power at hand. If one is only interested in differential gene expression, the simulation procedure and

processing of the simulated reads becomes a time limiting step.

For the second category, we have examples like DESeq (Anders and Huber, 2010) with its function `DESeq::makeExampleCountDataSet` and compcodeR (Soneson, 2014) an R package dedicated to generate synthetic RNA-seq count matrices based on the negative binomial (Robles et al., 2012). These methods were criticized by Benidt and Nettleton (2015) for using the same distribution to simulate the counts and then test for differential expression. Instead Benidt and Nettleton (2015) proposed simulating RNA-sequencing counts by subsampling columns from a large RNA-seq data-set, to then swap individual read counts within genes adjusted by a correction factor to create differential expression.

In this chapter we present RNACountSim, a tool to simulate RNA-sequencing matrices based on the Hoppe urn. RNACountSim uses as input a single RNA-sequencing count table, from which the distribution of the reads among the genes is infered with the PSF. RNACountSim is extremely fast even when simulating a high number of read counts and replicates. We used RNACountSim to evaluate the two most widely used tools to test for differential gene expression: edgeR and DESeq2 (Anders and Huber, 2010; Love et al., 2014). We found that both tools perform very well in all our evaluations, which include different sequencing depths and number of differentially expressed genes. Moreover, we detected particular behaviors like an increase in the number of false positives with increased sequencing depth in both tools for different data-set.

## 4.2 Methods

### 4.2.1 Input

RNACountSim takes four parameters as input: **1.** a table $T$ of count data (counted reads per gene) of an RNA-sequencing experiment, **2.** the desired number $D \geq 0$ of genes showing differential expression, **3.** the number of replicates $r$ and **4.** the desired sequencing depth $n$. RNACountSim simulates two conditions with $r$ replicates each. The difference between the two conditions is given by $D$ genes showing differential expression.

### 4.2.2 Simulating RNA-sequencing experiments

#### Simulating count matrices with $D = 0$ genes showing differential expression

Figure 4.1 shows the steps used to simulate count matrices using RNACountSim. Below is described the simulation steps when no genes show differential expression.

**Input data.** Let $T$ be the count table of an RNA-sequencing experiment that is used as input. We use all non-zero counts in $T$ to estimate the model parameters of the PSF by maximum likelihood using equation 2.1.

**Initialization vector.** During the simulation procedure one can generate thousands if not millions of possible distinct count tables given the properties of exchangeability (Pitman, 1995) and sequential sampling (Hoppe, 1984) of the Hoppe urn. In order to simulate similar but not identical count tables we initialize the Hoppe urn with an initialization vector. The initialization vector is created only once by sampling from the Hoppe urn until reaching the detection of 70% of estimated number of genes (eq. 2.4) for the given sequencing depth $n$.
This value was selected by analyzing four sequencing projects containig 123 biological replicates (SRP029889, SRP043108 and SRP043108) and 696 technical replicates (SRP025982) for different human tissues (skeletal muscle, brain, liver and cancer cell lines). From each data-set we computed the number of genes shared by all replicates. For the initialization vector we used the average number of shared genes for the four data-sets which is 70%.

**Simulation RNA-sequencing replicates.** RNACountSim simulates RNA-sequencing experiments for two conditions which differ in $D$ genes showing differential expression, by sampling reads from the Hoppe urn. First, the Hoppe urn is initialized with the *initialization vector*. Then, for each condition $r$ independent sampling events are performed by drawing $n$ reads. It is important to notice that for both conditions the Hoppe urn is initialized with the same initialization vector. Table 4.1 shows the first rows of an example of two simulated replicates. We can observe that the read counts for the same gene are similar, and given that the read counts are the same, we expect not to detect differential gene expression. Moreover, the read counts were not simulated using the binomial distribution. For comparison table 4.2 show the read counts of five genes corresponding to six replicates (REP1 to REP6).

Table 4.1: Read counts of two simulated conditions (three replicates per condition, five genes are shown) where no gene show differential expression. The simulation parameters ($\theta = 5680.085$, $\sigma = -0.2209978$) were estimated from a RNA-sequencing experiment of 20 million reads.

| Gene ID | Simulated Condition 1 | | | Simulated Condition 2 | | |
|---|---|---|---|---|---|---|
| 1 | 922 | 1043 | 831 | 1000 | 1056 | 1053 |
| 2 | 3071 | 2745 | 2660 | 3046 | 2802 | 2903 |
| 3 | 218 | 190 | 190 | 182 | 259 | 267 |
| 4 | 4153 | 3860 | 4127 | 3655 | 3702 | 3875 |
| 5 | 290 | 438 | 278 | 331 | 400 | 297 |

Table 4.2: Read counts of five genes corresponding to six replicates (REP1 to REP6) from sample A of the SEQC/MAQC-III data-set.

| Gene ID | REP1 | REP2 | REP3 | REP4 | REP5 | REP6 |
|---|---|---|---|---|---|---|
| 100 | 663 | 749 | 612 | 726 | 657 | 717 |
| 10001 | 166 | 219 | 167 | 231 | 181 | 184 |
| 10006 | 386 | 451 | 422 | 432 | 406 | 470 |
| 100132247 | 1919 | 2075 | 1836 | 2076 | 1928 | 2125 |
| 100462981 | 9625 | 10315 | 9281 | 10354 | 9377 | 10351 |

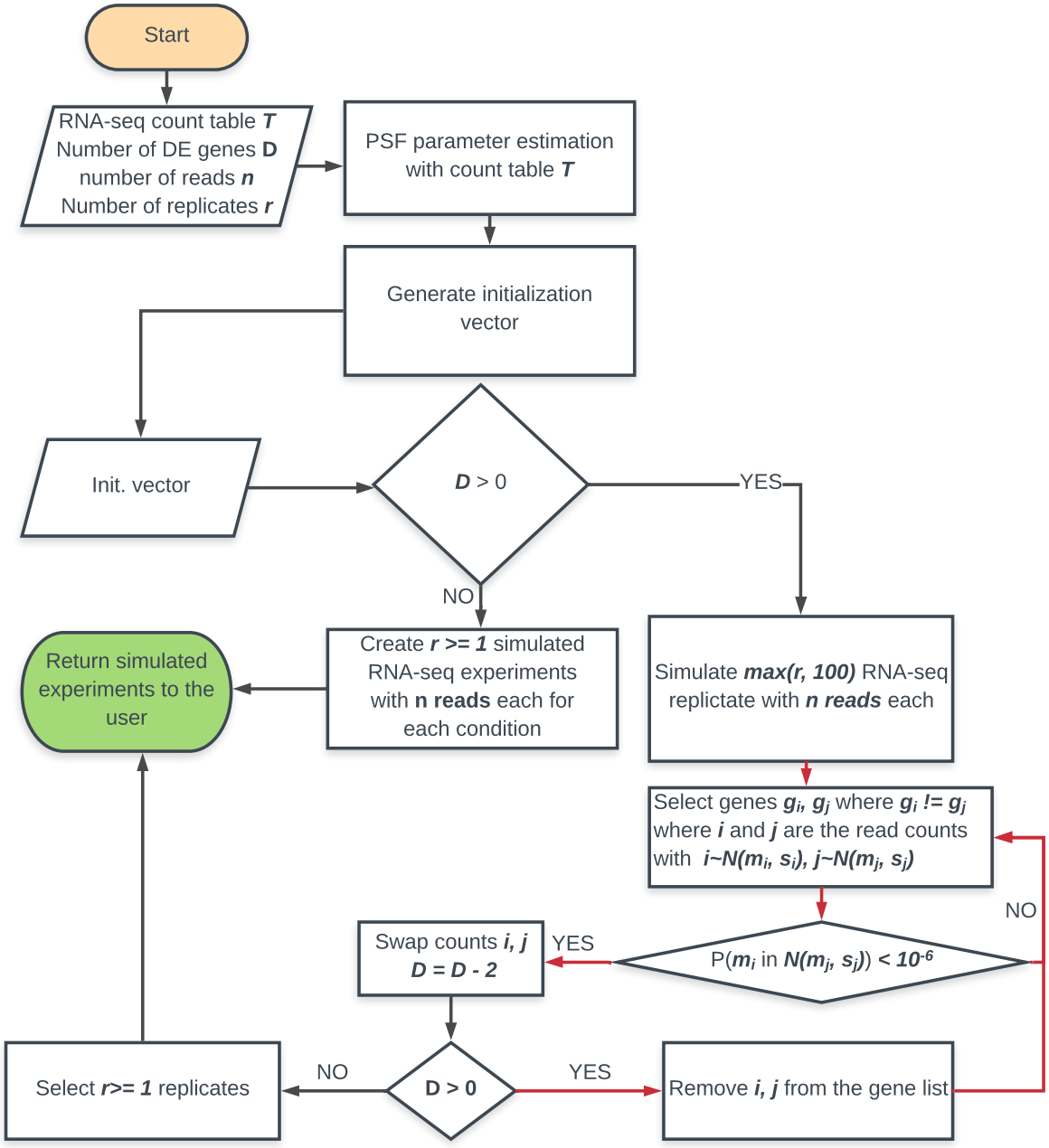



Figure 4.1: Steps followed to simulate RNA-sequencing experiments with RNA-CountSim. DE means differentially expressed.

### Simulating count matrices with $D > 0$ genes showing differential expression

To generate matrices with $D$ genes showing differential expression we add an extra step of swapping read counts (figure 4.1, red arrows). Below is described the additional steps required to select $D$ genes as differentially expressed.

**Selection of differentially expressed genes.** Anders and Huber (2010) stated that with large replicate numbers the normal distribution might provide a good approximation of the between-replicate variability of the read counts. We used our ability to generate a large number of simulated replicates to model the read count distribution of each gene with a normal distribution. Here, we simulated $max(r, 100)$ replicates per condition as described in the preceding section.

Let $i$ and $j$ be the read counts of a pair of genes $g_i$, $g_j$ where $g_i \neq g_j$ and $0 \notin i$ $0 \notin j$. We then estimate the parameters of the normal distribution (mean $\mu$ and standard deviation $s$) from the read counts $i \sim N(\mu i, s_i)$ and $j \sim N(\mu_j, s_j)$. Then, we use the probability density function of calculate the likelihood of $\mu_i$ being drawn from $N(\mu_j, s_j)$. If the likelihood is less-equal to one in a million then the read counts $i$ and $j$ are swapped in condition two to simulate difference in abundance. Then, genes $g_i$ and $g_j$ are tagged to avoid selecting them again (removed from the gene list). Otherwise, we randomly selected another pair $g_i, g_j$. We stop when the number of swapped read counts is equal to $D$, and thus simulating $D$ genes showing differential expression. For the case of odd number of differentially expressed genes, in the last pair only the read counts $j$ are assigned to $i$. Finally, from the $max(r, 100)$ simulated replicates, we return to the user the desired number of replicates $r$. Table 4.3 shows the first rows of an example of two simulated replicates. Here, for example the read counts of **gene 1** and **gene 2** were swapped to simulate differential expression. Similarly, the read counts of gene 5 were swapped. For comparison we show the read counts of five genes corresponding to three replicates of sample A and three replicates of sample from the SEQC/MAQC-III data-set. Here we show genes that were detected as differentially expressed using edgeR.

Table 4.3: Read counts of two simulated conditions (three replicates per condition, five genes are shown) where 10,000 genes were selected to show differential expression. The simulation parameters ( $\theta = 5692.109$, $\sigma = -0.2214656$) were estimated from a RNA-sequencing experiment of 20 million reads.

| Gene ID | Simulated Condition 1 | | | Simulated Condition 2 | | | DE |
|---|---|---|---|---|---|---|---|
| 1 | 9665 | 9927 | 9130 | 752 | 703 | 711 | YES |
| 2 | 616 | 690 | 592 | 9384 | 9962 | 9547 | YES |
| 3 | 3900 | 3715 | 3786 | 3941 | 3381 | 3499 | NO |
| 4 | 2858 | 2682 | 2434 | 3089 | 2629 | 2544 | NO |
| 5 | 11147 | 10936 | 10789 | 412 | 337 | 372 | YES |

Table 4.4: Read counts of five genes from sample A and sample B of the SEQC/MAQC-III data-set. Three replicates for each sample are shown (REPa1 to REPa3 and REPb1 to REPb3). Three genes show differential expression (DE).

| Gene ID | REPa1 | REPa2 | REPa3 | REPb1 | REPb2 | REPb3 | DE |
|---|---|---|---|---|---|---|---|
| 100 | 663 | 749 | 612 | 32 | 35 | 43 | YES |
| 1000 | 542 | 579 | 510 | 576 | 628 | 595 | NO |
| 10006 | 386 | 451 | 422 | 625 | 709 | 653 | NO |
| 100462981 | 9625 | 10315 | 9281 | 34434 | 37168 | 33828 | YES |
| 100506965 | 123 | 125 | 126 | 2446 | 2578 | 2332 | YES |

### 4.2.3 Measurements of performance

**Sensitivity (true positive rate).** The **sensitivity** measures the proportion of positives that were correctly identified as such

$$sensitivity = TP/(TP + FN)$$

where $TP$ are true positives and $FN$ are false negatives (van Stralen et al., 2009).

**Specificity (true negative rate).** Similar to sensitivity, the **specificity** measures the proportion of negatives that were correctly identified as such

$$specificity = TN/(TN + FP)$$

where $TN$ are true negatives and $FP$ are false positives.

**Accuracy (positive predictive value).** The **accuracy** measures the proportion of correctly identified positives in respect to all the identified positives

$$accuracy = TP/(TP + FP)$$

## 4.3 Results

To illustrate RNACountSim, we simulated RNA-sequencing count matrices to test the performance of edgeR version 3.18.1 (Robinson et al., 2010) and DESeq2 version 1.16.1(Love et al., 2014).

### 4.3.1 Simulating data-set with no genes showing differential expression

First, we simulated $1,000$ experiments where no genes show differential expression ($D = 0$). For the simulation procedure we used as input $T$, an RNA-sequencing experiment comprised of 20 million reads from the SEQC/MAQC-III dataset, and $r = 3$ replicates per condition. For the number of reads we selected $n = 20, 25, 30, 35, 40$ million reads which cover the range of the most common sequencing depths based on data from the recount data-base (Collado-Torres et al., 2017). To evaluate our simulated experiments we used the SEQC/MAQC-III data-set to construct experimental replicates of the same sequencing depth using subSeq (Robinson and Storey, 2014).

For both tools a gene was considered to show differential expression if the adjusted p-value of the respective test (depending on the software) was smaller-equal to 0.05. Hence, we expect that 5% of the genes detected as differentially expressed (DE) are actually not. Moreover, given that $D = 0$ the null hypothesis is true for all genes so we only evaluated the number of false positives (**FP**).

Figure 4.2 shows the result of the evaluation using the $1,000$ simulated and experimental data-sets and figures 4.3 and 4.4 show the results for the different sequencing depths. When using data generated by RNACountSim edgeR detected DE genes in 365 of the simulated data-sets with an average of 0.56 DE genes per data-set ($min = 0, max = 6, median = 0$), while DESeq2 detected in 624 with an average of 2.2 DE genes per data-set ($min = 0, max = 19, median = 1$). Here when taking into account the 5% FDR the average number FP DESeq2 per simulated data-set is two.

When using experimental data edgeR detected DE genes in all the $1,000$ data-sets ($mean = 24, min = 2, max = 60, median = 23$) while DESeq2 detected none. When analyzing these results by sequencing depth, we noticed that the number of false positives positively correlates with sequencing depth only when using experimental

data with edgeR 4.3. This observation could be caused by the increment in the number of genes that are analyzed. Table 4.5 shows the average number of genes analyzed by either edgeR and DESeq2. For all but one case (edgeR with RNACountSim) we can observe an increase in the number of analyzed genes.

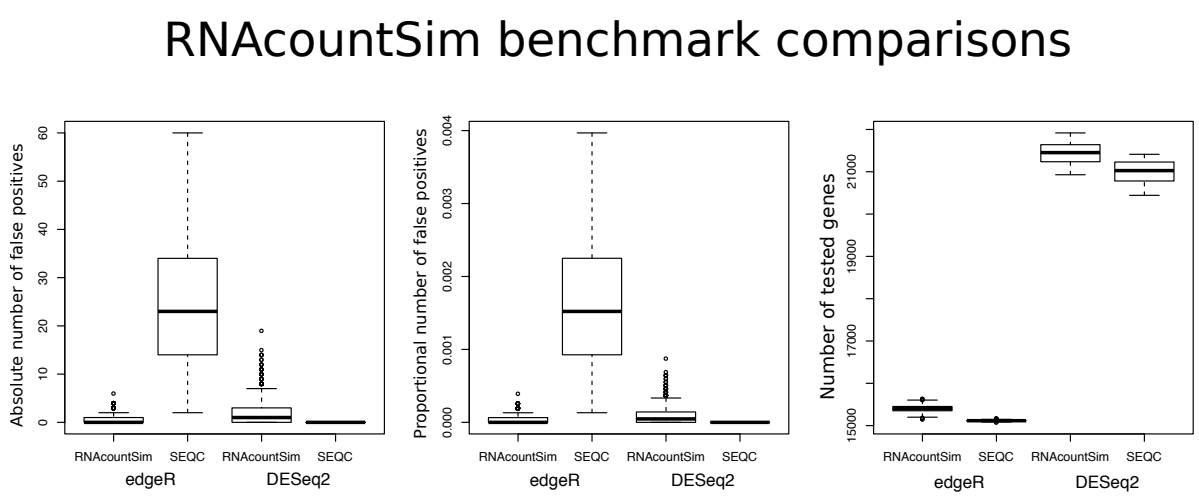



Figure 4.2: Results of the differential gene expression analysis using the 1,000 simulated replicates (RNACountSim) and experimental replicates (SEQC). By design no genes should show differential expression.

Table 4.5: Average number of genes analyzed by edgeR and DESeq2.

| Milllion reads | edgeR | | DESeq2 | |
|---|---|---|---|---|
| | RNACountSim | SEQC | RNACountSim | SEQC |
| 20 | 15,125.25 | 15,367.41 | 21,032.60 | 20,491.04 |
| 25 | 15,119.92 | 15,381.15 | 21,271.33 | 20,792.38 |
| 30 | 15,115.26 | 15,397.16 | 21,454.20 | 21,027.40 |
| 35 | 15,112.10 | 15,418.17 | 21,610.01 | 21,217.71 |
| 40 | 15,110.02 | 15,435.70 | 21,749.34 | 21,376.03 |

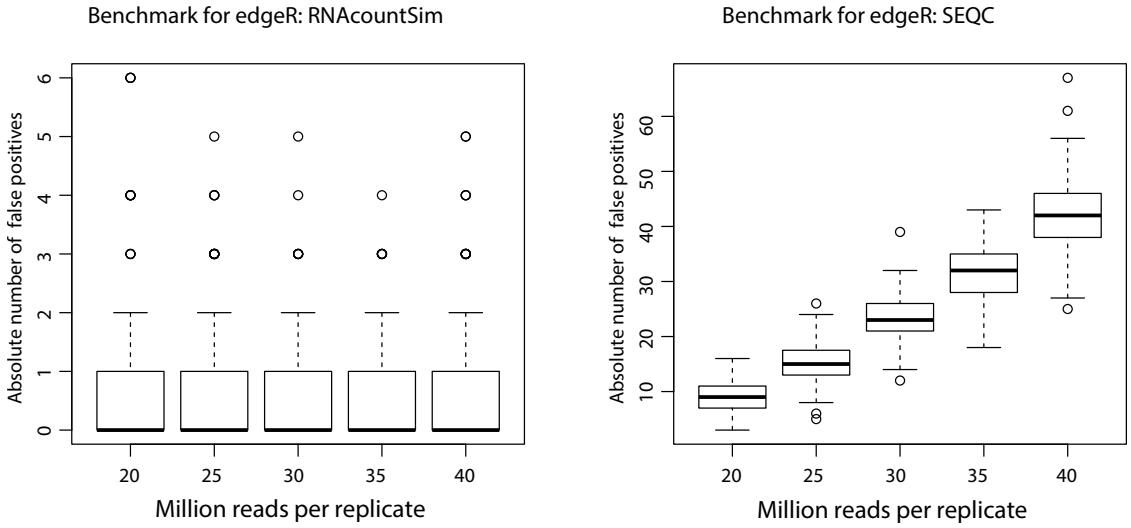



Figure 4.3: Results of the differential gene expression analysis with edgeR. The results of the 1,000 simulated replicates (RNACountSim) and experimental replicates (SEQC) are presented for the different sequencing depths (in millions of reads).

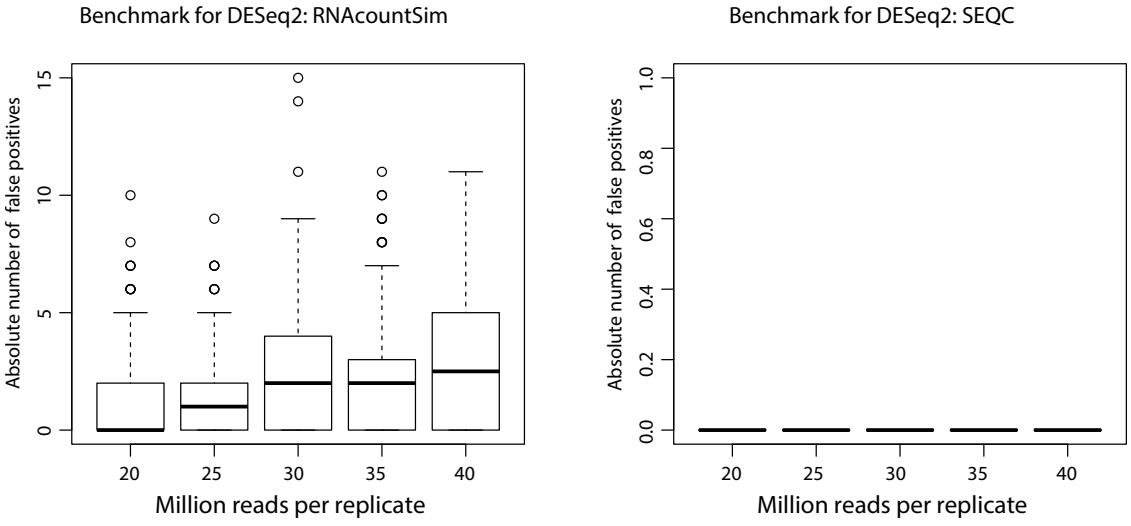



Figure 4.4: Results of testing for differential gene expression with DESeq2. The results of the 1,000 simulated replicates (RNACountSim) and experimental replicates (SEQC) are presented for the different sequencing depths (in millions of reads).

### 4.3.2 Simulating data-set with $D$ genes showing differential expression

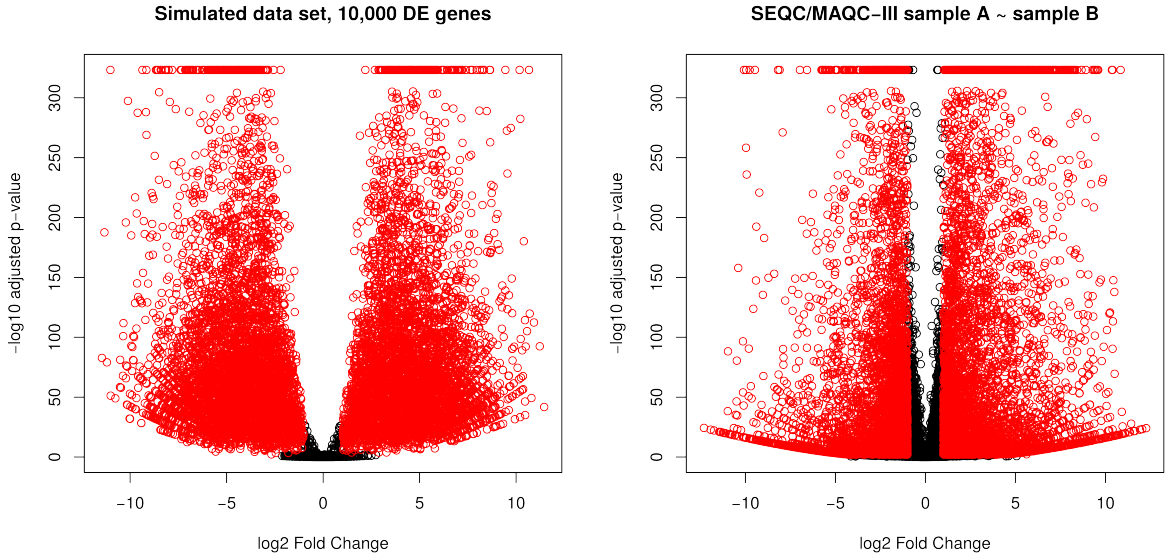
For the simulation procedure we used as input $T$, an RNA-sequencing experiment comprised of 20 million reads from the SEQC/MAQC-III dataset, and $r = 3$ replicates per condition. For the number of reads we selected again $n = 20, 25, 30, 35, 40$ million reads. Finally, we simulated five different values of the number of genes showing differential expression $D = 100$, $1,000$, $2,500$, $5,000$ and $10,000$. Figure 4.5 shows an example of a simulated experiment in which $10,000$ genes where selected to show differential expression. For comparison, we used two distinct human tissues (human brain and human non-brain cancer cell lines) and performed differential gene expression analysis with edgeR and DESeq2 (figure 4.5).

With the simulated experiments we assessed the performance of each tool by computing the sensitivity, specificity and accuracy. Figure 4.6 shows the overall performance of edgeR and DESeq2. The sensitivity and specificity scores of both tools was greater than 0.95. Moreover, non of the tools detected all genes showing differential expression. Table 4.6 show the proportion of TP that each tool detected for the different values of $D$.

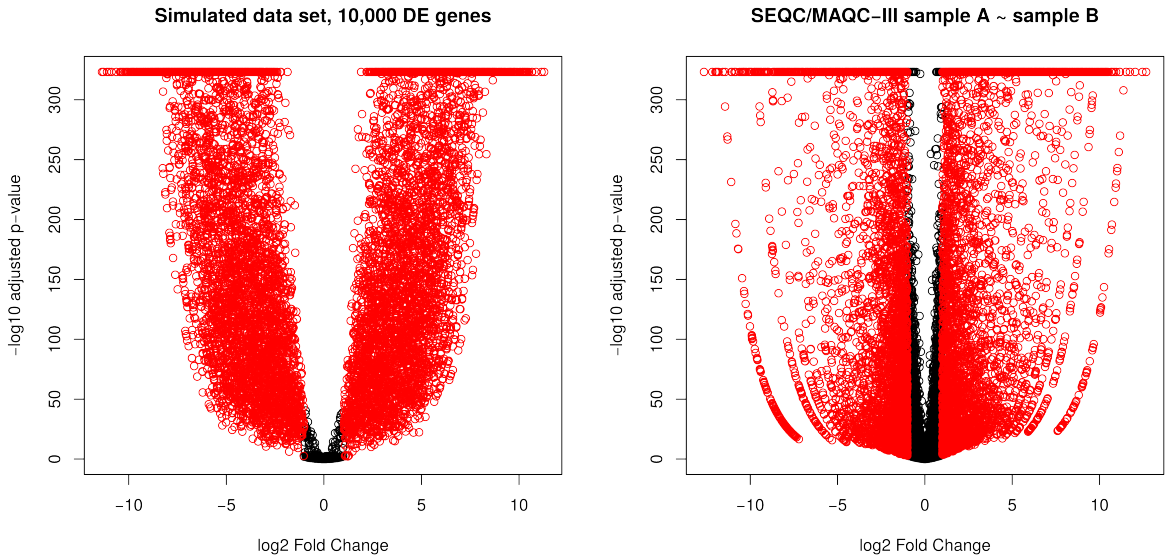
Table 4.6: Proportion of true positives (TP) for the different values of $D$.

| D | edgeR | DESeq2 |
|---|---|---|
| 100 | 1.00 | 1.00 |
| 1,000 | 0.978 | 1.00 |
| 2,500 | 0.951 | 0.999 |
| 5,000 | 0.915 | 0.969 |
| 10,000 | 0.857 | 0.897 |

For the case of the accuracy, the score of DESeq2 drastically decreased as it detected higher number of false positives 4.7. When computing the expected number of FP we noted that DESeq2 fails mostly when the number of genes showing differential expression is low (i.e $D = 100$). Here the number of FP taking into account the FDR was higher than expected (mean number of $FP \sim 12$, expected $FP \sim 5.5$, figure 4.7).




(a) DESeq2




(b) edgeR

Figure 4.5: Example of the analysis of differential gene expression. The left panel shows a simulated data-set with two conditions where 10,000 genes show differential expression. The right panel is shows the comparison of RNA from cancer cell lines to RNA from human brain. The red circles represent the genes that were detected as differentially expressed.

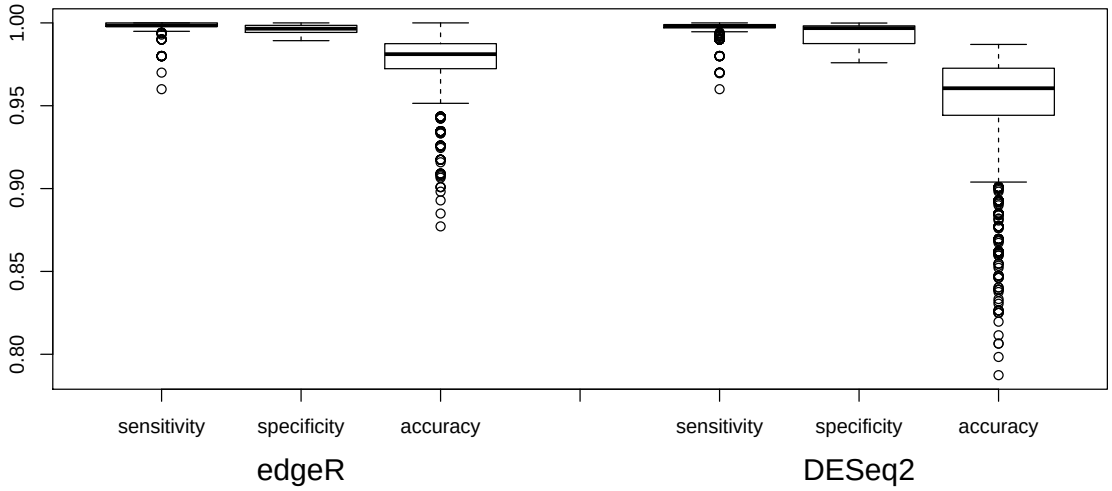
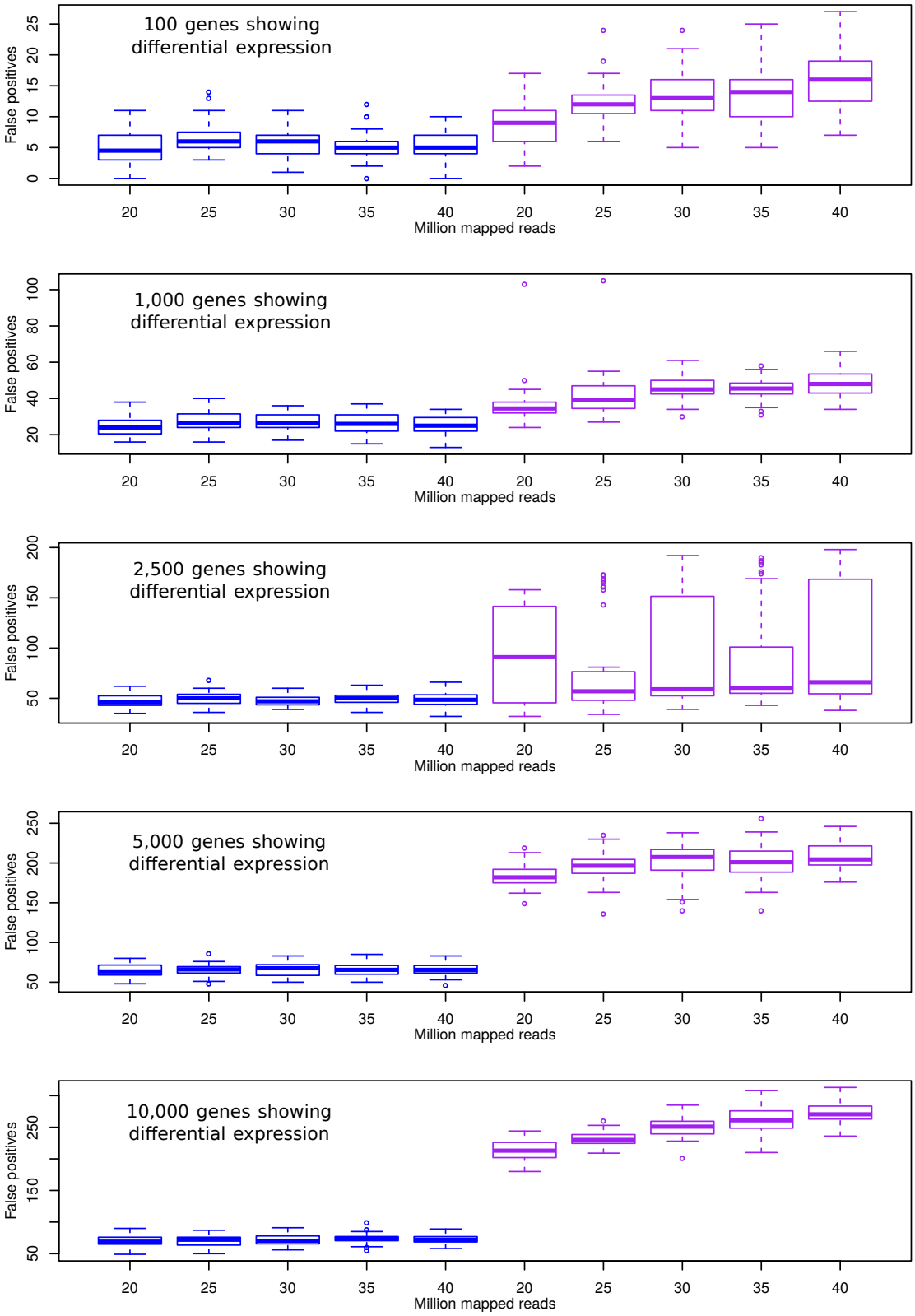



Figure 4.6: Performance of edgeR and DESeq2 from simulated data (1,000 experiments) quantified by sensitivity, specificity and accuracy.

Each tool has its own suggested filtering criteria. edgeR has a stricter criteria and uses two CPM (normalized read count, counts per million reads) which can be translated of each gene having in total for all the replicates between 20 to 40 reads. In contrast DESeq2 suggests to keep only the genes that have at $\geq 10$ reads in total. If this was the case we should observe an increase in the number of false positives when increasing the sequencing depth. Thus, we analyzed each of the 25 tested conditions of different sequencing depth and number of genes set to show differential expression.

Figure 4.7 shows the results of the 25 tested case scenarios (40 replicates each). Here, we can observe that the number of false positives detected by DESeq2 positively correlates with the increase in sequencing depth in *four* of the *five* tested case scenarios for the number of genes showing differential expression. This was particularly interesting in light of previous results, in which edgeR showed a higher number of false positives (with experimental data) while DESeq2 detected none. Here, the number of FP detected by edgeR does not look to be affected by the sequencing depth, but by the number of genes $D$ selected to show differential expression. These results shed a light into possible imporvement, and thus showing the power simulated data generated by RNACountSim for benchmark purposes.




Caption for figure 4.7 in the next page.

Figure 4.7: Performance of edgeR (in blue) and DESeq2 (in purple) from 1,000 simulated experiments. We show the number of false positives for each of the 25 distinct simulated case scenarios. From top to bottom the panels are divided by the number of genes $D = 100, 1,000, 2,500, 5,000$ and $10,000$ showing differential expression.

## 4.4 Discussion

Simulated data is now widely used during the development of tools to analyze genomic and metagenomic data generated by next generation sequencing (NGS) (Escalona et al., 2016). It is essential for the progress of the field that the newly developed methods are compared to existing ones in order to show their improvement. Researchers and developers have used empirical data as it represents real-life scenarios. However with such data-sets it is complicated to correctly assess each tool as the ground-truth is unknown. In contrast, simulated offers a solution and allow us to generate as much data as we need and under controlled scenarios in which the "truth" is known.

For the case of RNA-sequencing each tool has a different scope. Polyester (Frazee et al., 2015) and the FLUX simulator (Griebel et al., 2012) are examples of software that simulate reads that need to be processed before the analysis, which tend to be very time consuming when the main goal is studying and analyzing differential gene expression. However, both tools give insights of the library preparation and sequencing process which may affect the downstream analysis. We are not interested in understanding each and every step of the experimental procedure but, our aim is to generate large quantities of simulated data under known conditions in a short period of time.

Benidt and Nettleton (2015) criticized that most of available tools used to simulate the count matrices are based on the same distribution that the benchmarked software uses to test for differential expression. They instead suggested using source data-sets with sufficiently large number of replicates in order to simulate RNA-sequencing counts by sub-sampling columns the large data-set. Finding such large data-sets is not always easy nor feasible.

Here, we developed RNACountSim, a tool to simulate RNA-sequencing count matrices which is based in an urn model of sequential sampling for th PSF. RNACountSim selects the genes to show differential expression based on the normai distribution and thus, overcoming the criticism presented by Benidt and Nettleton (2015) without the need of large data-sets as input. We showed that RNACountSim is capable of simulating large number of experiments by generating $2,000$ independent experiments comprising of two conditions and three replicates per condition. The run-time for the simulations was less than 30 second per simulated experiment with no genes showing differential expression and less than one minute per simulated experiment when simulating genes showing differential expression.

With RNACountSim we made explicit the number of genes to show differential expression. With such information one can compare the performance of different tools by computing the sensitivity, specificity and accuracy as the truth is known. As a proof of concept, we evaluated two widely used software used to test for differential expression: edgeR and DESeq2. Even when both have in development for the past eight years we detected certain trends like an increase in the number of false positives with increase in sequencing depth. These results can shed a light into potential improvements.

RNACountSim is able to simulates RNA-sequencing in negligible time, such that researchers and developers can focus their time and effort in developing tools for analyzing RNA-sequencing and not in the simulation procedure. At the same time the simulated count matrices resemble experimental data with the additional benefit that the true number and identity of the genes showing differential expression is known.

## 4.5 Conclusions

RNACountSim, which is based on the Hoppe urn, is a tool to simulate count matrices of RNA-sequencing experiments. RNACountSim does not use the negative binomial distribution to simulate read counts and thus avoids the criticism by Benidt and Nettleton (2015), and at the same time is very fast (average run-time was 30 seconds). When compared to experimental data (SEQC) the simulated matrices showed similar behaviour when used to evaluate tools that test for differential gene expression. With simulated data we have the advantage of knowing ”the truth” which is of great use when developing and testing new methods or when comparing existing ones. We showed this by evaluating the most widely used tools for differential gene expression: edgeR and DESeq2. Further investigation is necessary to test our proposed model with newly developed protocols such as single-cell RNA-sequencing and the use of unique molecular identifiers (UMI).

# 5 GeneComplete: a tool for fast assessment of the completion of a genome annotation.

Next-generation sequencing is nowadays the main technology used in genome sequencing projects. After the genome sequencing is completed, the titanic task of annotating all functional elements, which include genes, begins. This task can take years and even decades (IHGSC, 2004) in which scientist improve the annotation over the years. For gene annotation, there exist pipelines that uses information of existing gene models to train gene-finders (Holt and Yandell, 2011; Mudge and Harrow, 2016). This is complemented with the inclusion of RNA-seq data, which aid in the annotation process (Nowoshilow et al., 2018; IWGSC, 2018). It is well known that the expression of many genes is tissue specific, for that matter researchers often collect as many tissues as possible to have a much more comprehensive annotation.

Here we present the use of the estimated number of detectable genes $m$ from the Pitman Sampling Formula (PSF, eq. 2.1) as a tool to assess the how complete a given gene annotation is. We use the read coverage information over the annotated genes to estimate the number of genes missing from the annotation. We evaluated the use of the estimated number of detectable genes $m$ with the annotation of the human genome by comparing the number of genes we estimated to be missing from version 3b of the GENCODE project (GENCODE, 2006) to version 25 which reflects seven years of continuous improvement. We examined 43 RNA-sequencing experiments covering 19 distinct human organs and found that muscle and esophagus have mostly reached saturation, while in thymus, trachea and in males testes, we expect to detect more genes in future annotation releases.

## 5.1 Methods

### 5.1.1 Selection of annotated genes

We downloaded the annotation files for the human genome from the GENCODE ftp server for versions **3b, 5, 10, 15, 20** and **25**. We started with 3b as since then the format has been maintained mostly unchanged. From the annotation files we only considered the feature type "gene". Also, the GENCODE project categorizes each gene in several distinct "biotypes" (e.g pseudogene, rRNA, protein coding, etc.). From this list we filtered out the Immunoglobulin variable chain and T-cell receptor (TcR) genes, all pseudogenes, mitochondrial genes, rRNA genes, To be Experimentally Confirmed, nonsense mediated decay sequences, non stop decay, processed transcripts, ambiguous ORFs, retrotransposed elements, genes in the category "artifact" and genes with a disrupted domain.

We used the annotation version 3b as our starting point to estimate the number of genes to be discovered and annotated in future releases. For annotation version 3b we used all gene status (known, novel and putative) and annotation confidence levels (1, 2 and 3).

We tested our prediction with annotation versions five to 25. In this cases we only included genes with gene status "KNOWN" and annotation confidence level "1: verified loci" and "2: manually annotated loci".

### 5.1.2 Experimental design

From the recount database (Collado-Torres et al., 2017) we downloaded 43 experiments from project SRP047192, which comprise the sequencing of 19 human organs: esophagus, heart, kidney, liver, lung, adipose, bladder, brain, cervix, colon, ovary, prostate, intestine, muscle, spleen, thymus, testes, thyroid, and trachea. The recount project released the base-pair coverage for each gene from which the read coverage can be calculated.

For each version of the annotation, we classified the genes as verified by RNA-sequencing if they were completely covered based on the base pair coverage information. Then, for these genes we calculated the read coverage using the `read_count`

function provided in the recount bioconductor package. We then constructed a table contains the read count information of the genes classified as **verified by RNA-sequencing** from annotation 3b to estimate the model parameters of the PSF. We then used the estimated parameter $\hat{m}$ as the predicted number of genes that can be annotated for that particular tissue. For the rest of the annotation files we counted the number of genes that were also completely covered based on the base pair coverage information in order to evaluate our prediction over seven years of continuous annotation improvement (version 5, 10, 15, 20 and 25).

## 5.2 Results and Discussion

Table 5.1 shows the results of the evaluation of the estimated parameter $\hat{m}$ from the PSF. In only five out of 43 analyzed experiments we overestimated the number of genes to be annotated.

For muscle and esophagus (table 5.1 marked with $\star$) we estimated a larger number of genes to be annotated from these tissues compared to the number of detected genes. In the past seven years the number of genes detected in these particular tissues have incremented, however since annotation version 15 the "discovery rate", meaning the number of new annotated genes, has decreased (figure 5.1 left panel). These results suggest that saturation has most likely been reached.

In contrast for testes, thymus and trachea (table 5.1 marked with †, figure 5.1 right panel) since annotation version 10 more genes have been annotated compared to our prediction. We also observed a decrease in the "discovery rate", however it does not look as pronounced as in the case of muscle and esophagus.

Generally, we observed a substantial increase in the number of annotated genes from version five to 15 (figure 5.1). This increase comes in hand with the description and overall adoption of RNA-sequencing as the preferred method to confirm and improve previously annotated genes, transcripts and exon/intron boundaries (Mortazavi et al., 2008; Nagalakshmi et al., 2010). RNA-sequencing has also served in the annotation of non-coding RNA genes and the classification of new RNA molecule categories. This can be seen in table 5.2. Note that the biotypes with a substantial increase are mostly in the category of RNA genes (antisense, lincRNA, miRNA, snoRNA, snRNA) while

protein coding has a decrease of almost 50% in the number of new annotated genes.

Table 5.1: Evaluation of the number of genes missed in annotation 3b compared to the genes in annotation 25. The evaluations is shown as "gencode - $\hat{m}$" which represents the difference between the number of genes from annotation 25 that were classified as "verified by RNA-sequencing" and the estimated number of genes with $\hat{m}$. Marked are tissues that most likely have reached saturation for the number of detected genes (⋆) and tissues which we may detect more genes in future annotations (†)

| Sample ID | Tissue | gencode - $\hat{m}$ | Sample ID | Tissue | gencode - $\hat{m}$ |
|---|---|---|---|---|---|
| SRR1576140⋆ | esophagus | -285 | SRR1576153 | cervix | 2158 |
| SRR1576141⋆ | esophagus | 210 | SRR1576178 | cervix | 3052 |
| SRR1576170⋆ | esophagus | -3778 | SRR1576154 | colon | 1760 |
| SRR1576142 | heart | 3806 | SRR1576179 | colon | 3209 |
| SRR1576143 | heart | 3810 | SRR1576155 | ovary | 974 |
| SRR1576171 | heart | 1801 | SRR1576180 | ovary | 1887 |
| SRR1576145 | kidney | 2062 | SRR1576157 | prostate | 1584 |
| SRR1576144 | kidney | 2004 | SRR1576182 | prostate | 2457 |
| SRR1576172 | kidney | -478 | SRR1576159 | intestine | 838 |
| SRR1576146 | liver | 2229 | SRR1576184 | intestine | 2184 |
| SRR1576147 | liver | 1973 | SRR1576158⋆ | muscle | -4753 |
| SRR1576173 | liver | 135 | SRR1576183⋆ | muscle | -2773 |
| SRR1576148 | lung | 4030 | SRR1576160 | spleen | 5462 |
| SRR1576149 | lung | 3940 | SRR1576185 | spleen | 5595 |
| SRR1576174 | lung | 1919 | SRR1576162 | thymus | 5313 |
| SRR1576150 | adipose | 3080 | SRR1576186† | thymus | 8950 |
| SRR1576175 | adipose | 3909 | SRR1576161† | testes | 8853 |
| SRR1576151 | bladder | 1040 | SRR1576187 | testes | 5514 |
| SRR1576176 | bladder | 2418 | SRR1576163 | thyroid | 5973 |
| SRR1576152 | brain | 2480 | SRR1576188 | thyroid | 6111 |
| SRR1576177 | brain | 3896 | SRR1576164† | trachea | 6282 |
| | | | SRR1576189† | trachea | 6488 |

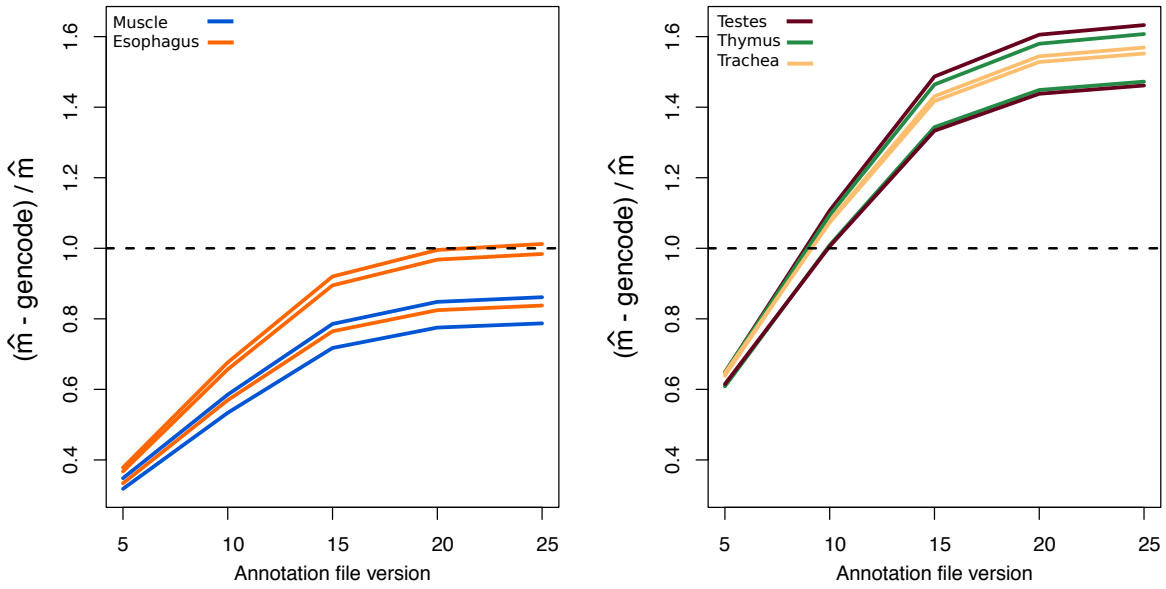



Figure 5.1: Evaluation of the number of genes missed in annotation 3b compared to the genes in annotation 25. The results of muscle and esophagus are shown in the left panel and the results for testes, thymus and trachea are shown in the right panel. Lines of the same color are replicates. In muscle and esophagus we estimated more genes to be annotated compared to the number of annotated genes from the data. In contrast for testes, thymus and trachea we expect to detect more genes in future annotations

Finally, we tested if the removal of certain categories from our evaluation procedure may impose some bias. Figure 5.2 shows hand-to-hand the number of new annotated genes compared to our predictions. Here, the solid lines are as before, while the dot-dashed line show when we use the same gene biotype categories in the genes used to predict and evaluate our proposed metric. Even when we can observe an increase in the number of genes, the number is not substantially different from our previous result. Therefore, we show that our propose metric can be of use to evaluate the status of genome annotation projects. Furthermore, we are able to detect tissues that require further sequencing from those which likely have been ”completed”.

Table 5.2: Gene biotypes that show a substantial increase in the number of annotated genes. In bold are marked those with an overall increase trend, while in protein coding (in italics) have shown a decrease.

| | Annotation improvement | |
|---|---|---|
| | Version 5 to 15 | Version 10 to 15 |
| **antisense** | 64 | 1438 |
| **lincRNA** | 488 | 1825 |
| miRNA | 1756 | 1598 |
| misc RNA | 1187 | 1185 |
| processed transcript | 31 | 28 |
| *protein coding* | 4574 | 2302 |
| **sense intronic** | 5 | 219 |
| snoRNA | 1521 | 1514 |
| snRNA | 1944 | 1923 |

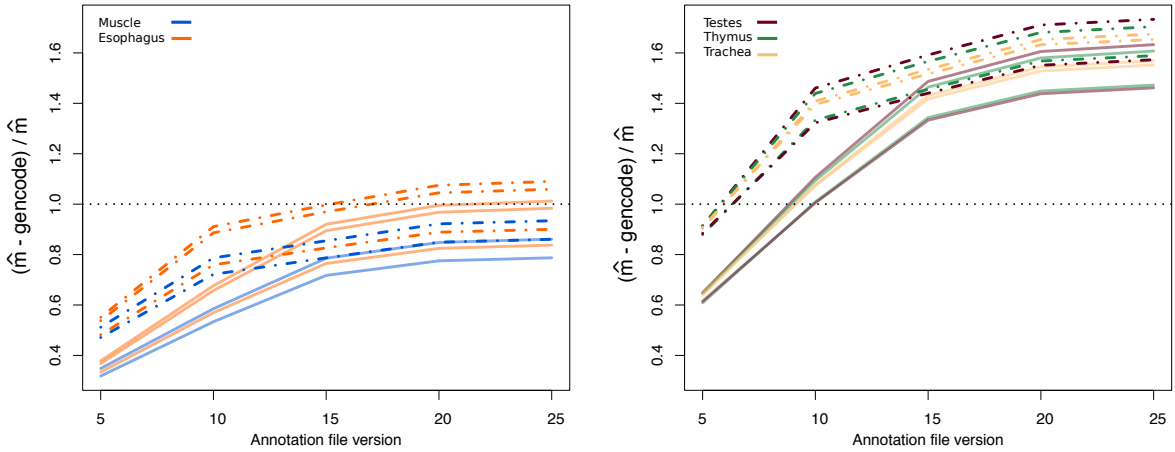



Figure 5.2: Evaluation of the number of genes missed in annotation 3b compared to the genes in annotation 25. Here, we tested if the removal of certain categories from our evaluation procedure. The results do not change in nature, meaning that for muscle and esophagus most likely have reached saturation for the number of detected genes (left panel), while testes, thymus and trachea we may detect more genes in future annotations (right panel)

## 5.3 Conclusions

We proposed the use of a parameter from the PSF as a metric to evaluates the completion of genome annotations. Our proposed metric was able to predict the number of annotated genes in different human tissues spanning seven years of continues annotation improvement. Also, we detected tissues where we are likely to detect new genes in future annotations, and those which most likely have reached saturation and thus, aiding the the selection tissues/conditions worth of further investigation.

# 6 Summary

This thesis presents three main contributions for the study of RNA-sequencing:

**Sampling formulas applied to RNA-sequencing (Chapter 3)**

I proposed and evaluated the use of a sampling formula developed in the field of population genetics, the Pitman Sampling Formula (PSF), to study RNA-sequencing. Here, I showed that statistics of the PSF to estimate the number of missed genes and the number of additional genes in follow-up experiments are accurate. By performing several experiments using a benchmark data-set, I provided evidence that the PSF performs better than the current available methods with experiments of low sequencing depth, and perform similarly good in experiments where the sequencing depth is high. Finally, I provided a way to calculate the return-of-investment of follow-up sequencing experiments in the number of additional detected genes, which can be used by experimental biologist during the planning and design of experiments.

**Simulation of RNA-sequencing experiments (Chapter 4)**

The wide adoption of RNA-sequencing for gene expression assays have lead to the development of several tools to test for differential gene expression (DGE). Many of these tools are tested first using simulated data that is produced with the same distribution they test for DGE. This motivated me to propose the use of an urn model of the PSF, the Hoppe urn, to simulate RNA-sequencing experiments and developed RNACountSim. I showed that the simulated data produced with RNACoutSim resembles experimental data by using both (simulated and experimental data) to evaluate two widely used tools to test for DGE. Finally, RNACoutSim can produce many simulated replicates in negligible time with the desired number of differentially expressed genes.

**Assessment of genome annotation (Chapter 5)**

As the number of complete genomes rapidly increases, the completion of the genome annotation usually takes years to decades. RNA-sequencing is now widely used to aid in the titanic labor that is the genome annotation, in which scientist sequence a wide variety of tissues with the aim of covering the whole transcriptome. Here, I make use of the PSF to evaluate the level of completeness of a current annotation. As proof of concept I used an old version human genome annotation to predict the number of genes in current versions. By using RNA-sequencing data of 19 different tissues I showed that our predictions are accurate. Moreover, I showed that certain tissues have not been saturated which offer scientist a lead to which tissues are worth sequencing more and which are not.

All methods described in this thesis were implemented for the R statistical environment and are planned to be distributed as software packages. Additionally, a manuscript showing the use of the PSF applied to RNA-sequencing is in preparation.

# Bibliography

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

E. Anderson, H. Skaug, and D. Barshis. Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Mol Ecol.*, 23(3):502–512, 2014.

F. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol.*, 18(7):437–451, 2017.

S. Benidt and D. Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics.*, 31(13):2131–2140, 2015.

K. Best, T. Oakes, J. Heather, et al. Computational analysis of stochastic heterogeneity in pcr amplification efficiency revealed by single molecule barcoding. *Sci Rep.*, 5:14629, 2015.

S. Blazie, C. Babb, and H. W. others. Comparative rna-seq analysis reveals pervasive tissue-specific alternative polyadenylation in caenorhabditis elegans intestine and muscles. *BMC Biol.*, 13:4, 2015.

M. Boguski and G. Schuler. Establishing a human transcript map. *Nat. Genet.*, 10(4): 369371, 1995.

J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

M. Busby, C. Stewart, and C. Miller. Scotty: a web tool for designing rna-seq experiments to measure differential gene expression. *Bioinformatics.*, 29(5):656–657, 2013.

Y. W. C.H. Chiu and, B. Walther, et al. An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. *Biometrics.*, 70(3): 671–682, 2014.

N. Cloonan, A. Forrest, G. Kolle, et al. Stem cell transcriptome profiling via massive-scale mrna sequencing. *Nat Methods.*, 5(7):613–619, 2008.

L. Collado-Torres, A. Nellore, K. Kammers, et al. Reproducible rna-seq analysis using recount2. *Nat. Biotech.*, 35:319–321, 2017.

A. Conesa, P. Madrigal, S. Tarazona, et al. A survey of best practices for rna-seq data analysis. *Genome Biol.*, 17:13, 2016.

A. Dobin, C. Davis, F. Schlesinger, et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics.*, 29(1):15–21, 2013.

M. Escalona, S. Rocha, and D. Posada. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat Rev Genet.*, 17(8):459–469, 2016.

W. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3: 87–112, 1972.

R. Fleischmann, M. Adams, O. Whiteand, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science.*, 269(5223):496–512, 1995.

A. Frazee, A. Jaffe, B. Langmead, et al. Polyester: simulating rna-seq datasets with differential transcript expression. *Bioinformatics.*, 31(17):2778–2784, 2015.

C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Phys Rev Lett.*, 90(8):088102, 2003.

L. Garcia-Ortega and O. Martinez. How many genes are expressed in a transcriptome? estimation and results for rna-seq. *PLOS ONE*, 10(6):e0130262, 2015.

GENCODE. Gencode: producing a reference annotation for encode. *Genome biol.*, 7(Suppl 1):S4.1–9, 2006.

A. Goffeau, B. Barrell, H. Bussey, et al. Life with 6000 genes. *Science.*, 274(5287):546567, 1996.

T. Griebel, B. Zacher, P. Ribeca, et al. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic Acids Res.*, 40(20):10073–10083, 2012.

R. Griffiths and S. Tavare. Ancestral inference from haplotypes and mutations. *Theor Popul Biol.*, 122:12–21, 2018.

T. Hardcastle and K. Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics.*, 11:422, 2010.

C. Holt and M. Yandell. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.*, 12:491, 2011.

F. Hoppe. Polya-like urns and the ewens sampling formula. *J. Math. Biol.*, 20:91–94, 1984.

I. H. G. S. C. IHGSC. Finishing the euchromatic sequence of the human genome. *Nature.*, 431(7011):931945, 2004.

I. W. G. S. C. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science.*, 361(6403):eaar7191, 2018.

D. Kim, G. Pertea, C. Trapnell, et al. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, 2013.

D. Kim, B. Langmead, and S. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nat Methods.*, 12(4):357–360, 2015.

V. Kuznetsov, G. Knott, and R. Bonner. General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics.*, 161(3):1321–1332, 2002.

A. Lijoi, R. M. RH, and I. Prnster. A bayesian nonparametric method for prediction in est analysis. *BMC Bioinformatics.*, 8:339, 2007.

M. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.*, 15(12):550, 2014.

A. Mortazavi, B. Williams, K. McCue, et al. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods.*, 5(7):621–628, 2008.

J. Mudge and J. Harrow. The state of play in higher eukaryote gene annotation. *Nat Rev Genet.*, 17(12):758–772, 2016.

U. Nagalakshmi, Z. Wang, K. Waern, et al. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science.*, 320(5881):1344–1349, 2008.

U. Nagalakshmi, K. Waern, and M. Snyder. Rna-seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol.*, 89:4:11.113, 2010.

S. Nowoshilow, S. Schloissnig, J. Fei, and others. The axolotl genome and the evolution of key tissue formation regulators. *Nature.*, 554(7690):50–55, 2018.

O. Ogasawara, S. Kawamoto, and K. Okubo. Zipf's law and human transcriptomes: an explanation with an evolutionary model. *C.R. Biol.*, 326(10-11):1097–1101, 2003.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

J. Pitman and N. Tran. Size-biased permutation of a finite sequence with independent and identically distributed terms. *arXiv.org.*, page arXiv:1210.7856, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

K. Renkema, M. Stokman, R. Giles, and others. Next-generation sequencing for research and diagnostics in kidney disease. *Nat Rev Nephrol.*, 10(8):433–444, 2014.

D. Robinson and J. Storey. subseq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics.*, 30(23):3424–3426, 2014.

M. Robinson, D. McCarthy, and G. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.*, 26(1): 139–140, 2010.

J. Robles, S. Qureshi, S. Stephen, et al. Efficient experimental design and analysis strategies for the detection of differential expression using rna-sequencing. *BMC Genomics*, 13:484:PMC3560154, 2012.

A. Rodriguez and F. Quintana. On species sampling sequences induced by residual allocation models. *J Stat Plan Inference.*, 157-158:108–120, 2015.

F. Sanger, S. N. S, and A. Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA.*, 74(12):5463–5467, 1977.

T. Shiraki et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA.*, 100(26):1577615781, 2003.

D. Sims, I. Sudbery, N. Ilott, et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Rev Genet.*, 15(2):121–132, 2014.

C. Soneson. compcoder–an r package for benchmarking differential expression methods for rna-seq data. *Bioinformatics.*, 30(17):2517–2518, 2014.

Z. Su, P. Åabaj, S. Li, et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.*, 32(9):903–914, 2014.

S. Tauber and A. von Haeseler. Exploring the sampling universe of rna-seq. *Stat Appl Genet Mol Biol.*, 12(2):175–188, 2013.

E. van Dijk, H. Auger, Y. Jaszczyszyn, et al. Ten years of next-generation sequencing technology. *Trends Genet.*, 30(9):418–426, 2014.

K. van Stralen, V. Stel, J. Reitsma, et al. Diagnostic methods i: sensitivity, specificity, and other measures of accuracy. *Kidney Int.*, 75(12):1257–1263, 2009.

V. Velculescu et al. Serial analysis of gene expression. *Science.*, 270(5235):484–487, 1995.

J. Wang, D. Dean, and F. H. others. todo. *Gynecol Oncol.*, S0090-8258(18):31283–31286, 2018.

R. Waterston, K. Lindblad-Toh, E. Birney, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.*, 420(6915):520562, 2002.

B. Wilhelm, S. Marguerat, S. Watt, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature.*, 453(7199):1239–1243, 2008.

H. Yamato and M. Sibuya. Moments of some statistics of pitman sampling formula. *Bull Inform. Cybernet.*, 32(1):1–10, 2000.

M. Zhaoa, D. Liu, and H. Qu. Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. *Brief Funct Genomics.*, 16(3): 121–128, 2017.