universität wien

DISSERTATION

Titel der Dissertation

„Exploring the transcriptome.

Innovative methods for analyzing RNA-Seq data.“

Verfasserin

Dipl.-Ing. Stefanie Tauber

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 094 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreut von:	Univ.-Prof. Dr. Arndt von Haeseler

Abstract

The fate of a cell is determined by the set of expressed proteins which governs its phenotype and all metabolic processes. Since quantification of the present protein spectrum turned out to be rather difficult, the transcriptome, the set of all expressed genes, is monitored instead.

RNA Sequencing (RNA-Seq) constitutes the state-of-the-art technology for large-scale gene expression screens. The output from the sequencing machine are short sequence tags, so-called reads, currently in the order of 10^8. By identifying the gene from which each read originates and by counting the number of reads per gene, we obtain an estimate for the underlying gene abundance. In contrast to former technologies, interrogation of gene expression is not restricted to already known genes. Thus, an uninformed view on the present transcriptome is achieved. Moreover, the obtained resolution is at the finest possible level - at the base pair level.

RNA-Seq arose at about eight years ago, since then the library preparation protocols in the wet-lab as well as the subsequent analysis workflows have been continuously improved. Yet, the unprecedented amount of data as produced by RNA-Seq is its boon and bane. In-depth analysis is often hindered by the overwhelming mass of data. In fact it frequently happens that the data is merely wrapped up into summary statistics to be able to handle it at all. Thus, while the standard workflow of RNA-Seq is already well established, a detailed analysis remains a challenging task as a consequence of a shortcoming of methods to do so.

Here we deliberately enter the wealth of data and stress the point of not neglecting the valuable resolution of RNA-Seq which is reflected in the per-base coverage, the

number of reads per position of a given gene. While typically all RNA-Seq analysis is centered on the read counts, we argue not be oblivious of the information contained in the coverage patterns. We contribute a method how to evaluate these patterns by consulting a classical measure, namely the Fractal Dimension. We link the roughness of a coverage graph to its reliability and are thus able to pinpoint suspicious coverage patterns and, as a consequence, unravel its causes being pitfalls while library preparation or analysis.

Additionally, we address the question of the necessary as well as sufficient sequencing depth in order to detect all expressed genes. While the typical aim of the vast majority of global gene expression studies lies in the gene-wise inference of differential expression we propose a global perspective upon the sequencing data. We consider the data as a sampling process (number of reads per gene\position) which may be modeled by sampling formulas originating in the field of population genetics. These sampling formulas allow us to realistically capture the distribution of reads within and between genes which is of immediate benefit for simulation tools. Moreover, we are even in the position of making valid predictions about the expected number of newly detected genes given a certain amount of sequencing reads. Carrying this question to the extreme results in the exploration of the boundaries of the respective underlying transcriptome. Since the repertoire of expressed genes is far from being static and depends on the specific biological set-up such as organism, tissue, cell type and developmental state this question is of particular interest.

Parts of this thesis have been published in the following article:

i) S. Tauber and A. von Haeseler (2013) Exploring the Sampling Universe of RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, **12**(2), 175–188.

In preparation:

i) S. Tauber and A. von Haeseler
FractalQC: A Bioconductor Package for Quality Control of RNA-Seq Coverage Patterns by Means of the Fractal Dimension.

Zusammenfassung

Das Schicksal einer Zelle wird von der Menge aller exprimierten Proteine bestimmt, die wiederum für den Phänotyp der Zelle und für alle metabolischen Prozesse verantwortlich ist. Da sich die Quantifizierung aller exprimierten Proteine als außerordentlich schwierig erwiesen hat, wird stattdessen das Transkriptom, die Menge aller exprimierten Gene, zur Untersuchung herangezogen.

RNA Sequenzierung (RNA-Seq) stellt die modernste Technologie für großangelegte Untersuchungen von Genexpression dar. Der Sequenzierer gibt kurze Sequenzstücke aus, sogenannte 'reads', momentan circa 10^8 reads pro Sequenzierung. Indem man das Gen, von dem der jeweilige read stammt, identifiziert und dann die Anzahl der reads pro Gen aufsummiert, erhält man eine Approximation der zugrunde liegenden Genexpression. Im Vergleich zu älteren Technologien können nicht nur schon bekannte Gene abgefragt werden. Im Gegenteil, RNA-Seq benötigt keinerlei Vorwissen über die Struktur der jeweils vorliegenden Gene. Weiters ist die realisierte Auflösung von RNA-Seq bestmöglich - am Basenpaar-Level.

Die Anfänge der RNA-Seq Technologie liegen circa acht Jahre zurück, seitdem haben sich sowohl die Protokolle im Nasslabor als auch die Analysen kontinuierlich weiterentwickelt und verbessert. Die Menge an Daten, die von RNA-Seq tagtäglich produziert wird, ist ohne Präzedenzfall und sowohl von Vor- als auch von Nachteil. Detaillierte Analysen werden oft durch die überwältigende Masse an Daten erschwert. Tatsächlich werden die Daten häufig in statistischen Maßzahlen zusammengefasst, um sie überhaupt handhaben zu können. Während die Standard-Analyse von RNA-Seq Daten schon sehr gut etabliert ist, bleibt eine tiefgehende Analyse eine Herausforderung, da es noch an passenden Methoden fehlt.

Wir möchten dezidiert die Masse der Daten ausnutzen und betonen daher die Wichtigkeit, die von RNA-Seq gebotene wertvolle Auflösung nicht zu ignorieren. Diese Auflösung spiegelt sich wieder in der sogenannten 'per-base coverage', der Anzahl der reads pro Basenpaar. Während eine gewöhnliche RNA-Seq Analyse auf der Anzahl der reads pro Gen aufbaut, betonen wir, wie wichtig es ist, die Information des 'Coverage Pattern' nicht zu vernachlässigen. Wir haben eine Methode entwickelt, mit der solche Coverage Pattern bewertet werden können und zwar anhand der Fraktalen Dimension. Wir zeigen, dass es eine Verbindung zwischen dem Graphen des Coverage Pattern und seiner Vertrauenswürdigkeit gibt. Infolgedessen sind wir in der Lage, fragwürdige Coverage Pattern und mögliche Gründe, die sowohl im Nasslabor also auch in der Analyse liegen können, zu identifizieren.

Weiters widmen wir uns der Frage der notwendigen als auch hinreichenden Sequenziertiefe, um alle exprimierten Gene zu detektieren. Die meisten Genexpressions-Studien sind an der Quantifizierung der differenziellen Genexpression interessiert. Im Vergleich dazu schlagen wir eine globale Sichtweise der Dinge vor. Wir fassen die Daten als eine Stichprobe auf (Anzahl der reads pro Gen\Position) und charakterisieren den zugrunde liegenden Prozess mittels Formeln, die aus der Populationsgenetik kommen. Dies ermöglicht uns realistisch die Verteilung von reads innerhalb von und auch zwischen Genen zu modellieren. Der Nutzen dieser Methode für Simulationszwecke ist sofort ersichtlich. Darüber hinaus sind wir sogar in der Lage, Prognosen über die Anzahl der zu erwartenden, neu detektierten Gene zu machen, gegeben einer bestimmten Menge an reads. Treibt man diese Fragestellung zum Äußersten, dann führt das zur Erforschung der Grenzen des jeweils vorliegenden Transkriptoms. Das Repertoire an exprimierten Genen ist sicherlich nicht statisch und hängt von den spezifischen biologischen Gegebenheiten wie zum Beispiel Organismus, Gewebe, Zelltyp und Entwicklungsstatus ab. Daher ist diese Fragestellung von besonderem Interesse.

Teile dieser Arbeit wurden in dem folgenden Artikel publiziert:

i) S. Tauber and A. von Haeseler (2013) Exploring the Sampling Universe of RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, **12**(2), 175–188.

In Vorbereitung:

i) S. Tauber and A. von Haeseler
FractalQC: A Bioconductor Package for Quality Control of RNA-Seq Coverage Patterns by Means of the Fractal Dimension.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

All organisms on earth[1] carry their genetic information in the form of Deoxyribonucleic acid (DNA). Its structure, the double helix, arises from two complementary strands consisting of four nucleotides: adenine, thymine, guanine and cytosine. Hydrogen bonds link the matching nucleotides, adenines always pairs with thymine and guanine with cytosine. Any kind of catalytic activity can only take place unidirectional, from the 5'-end to the 3'-end. Information is passed on from the DNA by (i) transcription into messenger Ribonucleic acid (mRNA) and, subsequently, (ii) translation into proteins. This process, DNA $\rightarrow$ RNA $\rightarrow$ proteins, is known as being one part of the Central Dogma of Molecular Biology (Crick, 1970).

However, not the entire DNA is transcribed. Genes constitute that part of the DNA containing genetic information responsible for a specific trait or characteristic of the respective cell. Transcription starts with synthesizing the mRNA from a gene present on the DNA. The resulting pre-mRNA consists of one or more exons possibly interrupted by noncoding introns. In the following the pre-mRNA undergoes multiple processing steps amongst which noncoding introns are excised. Yet, multiple options exist regarding the inclusion or exclusion of exons and\or introns. The splicing apparatus controls the admissible exon compositions of mRNAs which may arise from the very same gene, so-called isoforms. Additionally, a nucleotide

[1] besides one exception, RNA-viruses carry their genetic information in the form of RNA

stretch of about 200 adenines is added at the 3'-end ('poly-A tail') for stability reasons. Eventually the mature mRNA may be translated into its corresponding protein. The set of expressed proteins together with all functional RNAs in a cell governs all metabolic processes and drives the phenotype of the cell (Gardner et al., 1991).

As quantification of the present protein spectrum of a given cell remains to be difficult, the repertoire of expressed genes is monitored instead. The qualitative and quantitative comparison of the mRNA levels between multiple biological sources is the main objective of gene expression studies. Before the advent of high throughput technologies, this question could only be addressed gene-wise by means of quantitative real-time polymerase chain reaction (PCR). Microarrays, and more recently Next Generation Sequencing (NGS) technologies made genome-wide expression screens possible. NGS permits multifaceted applications such as complete genome-resequencing, tracking protein-nucleic acid interactions by means of ChIP-Seq or sequencing of metagenomic samples. See Shendure and Ji (2008) for an overview of applications of NGS technologies. Here we will concentrate on RNA sequencing (RNA-Seq) which aims to quantify all expressed mRNAs in a given biological sample.

1.2 Micoarrays vs RNA-Seq

Both Microarrays as well as RNA-Seq are high-throughput technologies meaning that the expression of thousands of genes can be measured at the same time. While in the last decade of the past century Microarrays initiated the paradigm shift from gene-wise to global expression screens, RNA-Seq took over as state-of-the-art technology beginning in about 2005.
In fact, Microarrays and RNA-Seq differ substantially regarding the underlying technology yet we would like to lay the focus on the most important differences in the resulting data. By doing so the superiority of RNA-Seq over Microarrays is immediately comprehensible. First, Microarrays only yield relative measures of the expression levels while RNA-Seq returns absolute values in the form of count

data. Secondly, Microarrays are restricted to the interrogation of the set of already known genes. In contrast, RNA-Seq permits an uninformed view of the RNA expression landscape. Both technologies output an unprecedented amount of data. The production of large and many data is even more accelerated for RNA-Seq since this technology gets more inexpensive and, simultaneously, increases its throughput substantially every year.

1.3 NGS Technologies

Several NGS technologies are available, the most prominent include SOLiD from Life Technologies\Applied Biosystems, Roche's 454 technology, the Genome Analyzer and the HiSeq System from Illumina and Ion Torrent from Life Technologies (Metzker, 2010). While these technologies pursue their specific proprietary library preparation and sequencing protocol, the resulting data structure is similar. Thus methods developed on data originating from one technology can be easily transferred. Here, we mainly work with data from the Illumina technology (Bentley et al., 2009) being one of the most popular technologies.

1.3.1 Library Preparation for RNA-Seq

Library preparation refers to the protocol in the wet-lab, starting with the extraction of the RNA from the biological sample to the submission of the sample to the sequencing machine. Here we will briefly describe the individual steps of the standard Illumina Protocol (TruSeq Illumina, 2012). This knowledge is immanent to understand possibly occurring biases in the data due to the library preparation.

Extraction RNA, typically the poly-A mRNA, is extracted. The restriction to poly-A mRNA is not severe since nearly all protein coding genes[2] in eukaryotes exhibit a poly-A tail (Proudfoot et al., 2002).

[2] besides some histone genes

Fragmentation Since the average length of a gene exceeds the capacity of the sequencing machine, the mRNA is sheared into smaller fragments. The default fragmentation method of Illumina is using divalent cations under elevated temperatures. Other fragmentation methods include sonication (breaking of the RNA by means of the vibrations of ultrasonic waves), nebulization (compressed air or nitrogen forces the RNA to break) or random enzymatic digestion (the mRNA is cut by the collaboration of two enzymes, one marking the position where then the other enzyme cuts).

First and second strand synthesis Since mRNA is highly unstable by nature conversion to DNA is necessary. The start site of the synthesis is determined by a short nucleotide sequence, a so-called primer. In this protocol priming of the RNA fragments is done with random hexamers (nucleotide sequences of length 6), finally double-stranded complementary DNA (dscDNA) is obtained.

Adapter ligation Subsequent sequencing requires the immobilization of the sequencing templates on the so-called flow cell, a planar surface with a dense lawn of oligonucleotides. In order to capacitate the dscDNA to bind onto the flow cell, short DNA sequences (adapters) are ligated to both ends of the dscDNA.

Enrichment Amplification via PCR of all admissible fragments (fragments with attached adapters) in order to increase the amount of DNA in the library in general.

Preparation for sequencing Finally the dscDNA is denatured and diluted before the actual sequencing starts.

We refrain from giving a detailed description of the sequencing process itself since this knowledge is not required for the understanding of the presented work here. Metzker (2010) gives an excellent overview on this topic.

Up to 8 different biological samples can be sequenced simultaneously since a flow cell contains 8 compartments ('lanes'). The typical sequencing yield per lane is about $100 - 200 \times 10^6$ single-end reads per lane, the read length may vary between $50 - 150$bp. Each read refers to one fragment of a mRNA in the original biological sample. In the case of single-end sequencing each fragment is just sequenced from one end whereas paired-end sequencing delivers two reads per fragment, one from each end. Since RNA-Seq still results in substantial costs efforts have been made to exploit the sequencing yield more efficient. This has been achieved by multiplexing which permits sequencing of multiple biological samples in one lane of the flow cell. Each biological sample is tagged with a so-called barcode which allows subsequent distinction. The seemingly disadvantage of a lower obtained sequencing yield per biological sample frequently vanishes as not all studies require ultra deep sequencing.
Besides from the raw sequence data quality scores for each sequenced base are available indicating the reliability of the base calling. More precisely, the quality values give information on how probable the specific nucleotide has been identified correctly.

1.3.2 Sampling Process

The sequencing process can actually be regarded as sampling process. The original biological sample contains a certain number of mRNAs present at various expression levels. Each mRNA is then subjected to multiple processing steps such as fragmentation, amplification or priming in the course of the library preparation protocol. The final outcome of the sequencing machine consists of sequencing reads whereas each read refers to one fragment. Typically the number of obtained sequencing reads is limited by the throughput of the underlying sequencing technology and not due to exhaustion of the biological sample. Thus the reads represent a subsample of the original amount of input mRNA.
In order to provide some intuition for this sampling process and its orders of magnitude we track the absolute number of molecules at different stages of the library preparation protocol. Since the amount of input mRNA is known and by assuming

an average mRNA length of 1200 bp (*Mus Musculus*, numbers extracted from Ensembl (Flicek et al., 2013)) we can compute the number of molecules in the initial biological sample. Calculation works by converting the amount of mRNA given in grams into moles, being the standard measurement unit. This is achieved by dividing the amount of mRNA by the molecular weight of each nucleotide ($\sim$ 330 pg/pmol) times the expected length of the RNA, and accounting for scaling differences. The absolute number of molecules is easily extracted from the number of moles by multiplication with Avogadro's number ($\sim 6.022 \times 10^{23}$). See Appendix A for a detailed derivation.
Moreover, the concentration of the sample is measured multiple times throughout the library preparation, enabling us to calculate the number of molecules at these stages as well (see Table 1.1). Given a final sequencing yield of about 100×10^6 reads, only 0.06% of the original input amount or, respectively, 0.002% of the sample after library preparation is sequenced. The decrease in sequencing percentage (from 0.06% to 0.002%) can be explained by the fact that the mRNAs are fragmented and subsequently amplified in the course of the library preparation. Thus the original mRNA content is 'blown up'. Additionally, we have to note that is more appropriate to compare the sequencing yield to the number of molecules after library preparation since the mRNA is already fragmented at this stage. Thus the size of the molecules should be comparable.

Given these numbers it is quite surprising that sequencing delivers a representative picture of the underlying set of expressed genes most of the time. The reliability of our calculation is supported by McIntyre et al. (2011) who conducted a similar study resulting in comparable orders of magnitude for the number of molecules. Yet, comparison to Microarrays as well to quantitative real-time PCR confirms the convincing performance of NGS (Bullard et al., 2010, Mortazavi et al., 2008, Marioni et al., 2008).

stage of library preparation protocol	number of molecules $[10^{12}]$
input amount	$1.5 \times 10^{11}[0.15]$
after library preparation	$3.6 \times 10^{12}[3.6]$
flow cell load	$5.1 \times 10^{8}[0.00051]$
sequencing yield	$100 \times 10^{6}[0.0001]$

Table 1.1: Number of molecules (fragments of mRNA) throughout the library preparation. To facilitate the comparison of orders of magnitude numbers in square brackets are on the scale of 10^{12}.

1.4 RNA-Seq Analysis Workflow

The overall goal of RNA-Seq experiments is the comparison of expression levels between genes and conditions. Thus, once the raw sequencing reads are obtained from the sequencing machine several preprocessing steps are necessary to provide the data in appropriate form for subsequent analyses.

Mapping The sequencing reads are mapped to the genome and/or to the transcriptome in order to recover their origin. A variety of mapping software is available, see Lindner and Friedel (2012) for a recent review focusing on RNA-Seq data. Software development happens at rapid pace resulting in faster alignments of more and more reads. Yet, two decisions remain to be taken by the user: How many differences such as substitutions, insertions and deletions are to be sensibly allowed. The number of admissible differences depends on the error rate of the underlying sequencing technology and on the expected genomic divergence of the underlying biological sample to the reference. Secondly, how to deal with reads that cannot be placed unambiguously. No general rule is yet established.

Besides mapping the reads to a reference sequence, assembly of the reads depicts another possibility for transcriptome reconstruction (Garber et al., 2011). In general, RNA-Seq assembly is clearly preferable as it allows an uninformed view on the mRNA spectrum. However, Schliesky et al. (2012) emphasize that current RNA-Seq assemblies not yet achieve satisfying results

in terms of accurate representation of the transcriptome. Therefore and since all data processed in this Thesis relies on reference based mapping we refrain from expanding upon transcriptome assembly.

Summarization Determination of the expression levels depends on (i) the entity per which expression shall be calculated such as exon, isoform or gene, (ii) the chosen summary statistic to represent the respective expression. Typically the sum of reads per entity is chosen. Summarization constitutes by no means a trivial part in the analysis workflow. Cases such as reads mapping to exons shared by multiple isoforms, or reads mapping to genes annotated at the same genomic coordinates but opposite strands are treated inconsistently in different studies. In any case, the summarization step results in a count table which constitutes the fundamental basis for all subsequent analyses. Typically different experimental conditions are displayed in separate columns whereas the entity over which summarization has taken place (e.g. genes) is listed in the rows. While the impact of the summarization method on the count table is immense it was observed that methods development neglects the summarization step (Oshlack et al., 2010).

Normalization Within-lane as well as between-lane normalization is necessary in order to enable valid comparison of expression levels between genes, and, respectively, between conditions. Given two equally expressed genes, the longer gene always yields more sequencing reads as as consequence of the library preparation protocol. Thus normalization for gene length is required. The need for between-lane normalization results from the fact that lanes of a flow cell typically differ in their sequencing yield (Bullard et al., 2010). As a consequence, it may happen that equally expressed genes have different read counts when originating from lanes with a substantial discrepancy in the sequencing depth.

Differential expression inference Finally, having the summarized, normalized expression values per entity of interest at hand, the null hypothesis of no differ-

ential expression is tested. The Poisson or the Negative Binomial distribution is typically adopted to model RNA-Seq count data. Multiple methods are available amongst which baySeq (Hardcastle and Kelly, 2010), edgeR(Robinson et al., 2009) and DESeq (Anders and Huber, 2010) are the most prominent (Kvam et al., 2012). The major challenge lies here in the fact that typically - if at all - a very low number of replicates per condition is available. Additionally the nature of the experimental design is massively parallel since the null hypothesis of no differential expression is tested for each gene.

1.5 Bias Overview

Here we would like to give a concise description of potential occurring biases leading to a distorted distribution of reads within and between genes. In contrast to former beliefs that a technology based on actual counting of the mRNA fragments does not exhibit a need for 'sophisticated normalization' (Wang et al., 2009) it is now evident that RNA-Seq data suffers from a conglomerate of different biases, see Ross et al. (2013) for a most recent review.

We distinguish two main bias sources - either the library preparation protocol or the subsequent analysis. Library preparation protocols as well as the following analyses are not static. Best practice guides for the library preparation protocols are available, similarly automated analysis pipelines are already offered. Yet, each decision comes along with its consequences and thus it is indispensable to know about the causal relationship between individual processing steps and subsequent patterns in the data. Figure 1.1 gives an overview of the RNA-Seq workflow, possibly occurring biases and where these biases enter the analysis workflow.

Biases originating from the library preparation protocol:

Amplification and GC content bias It is well known that the enrichment step is the primary source of base-composition bias (Aird et al., 2011, Oyola et al., 2012). Specifically, the underrepresentation of GC rich and poor regions is

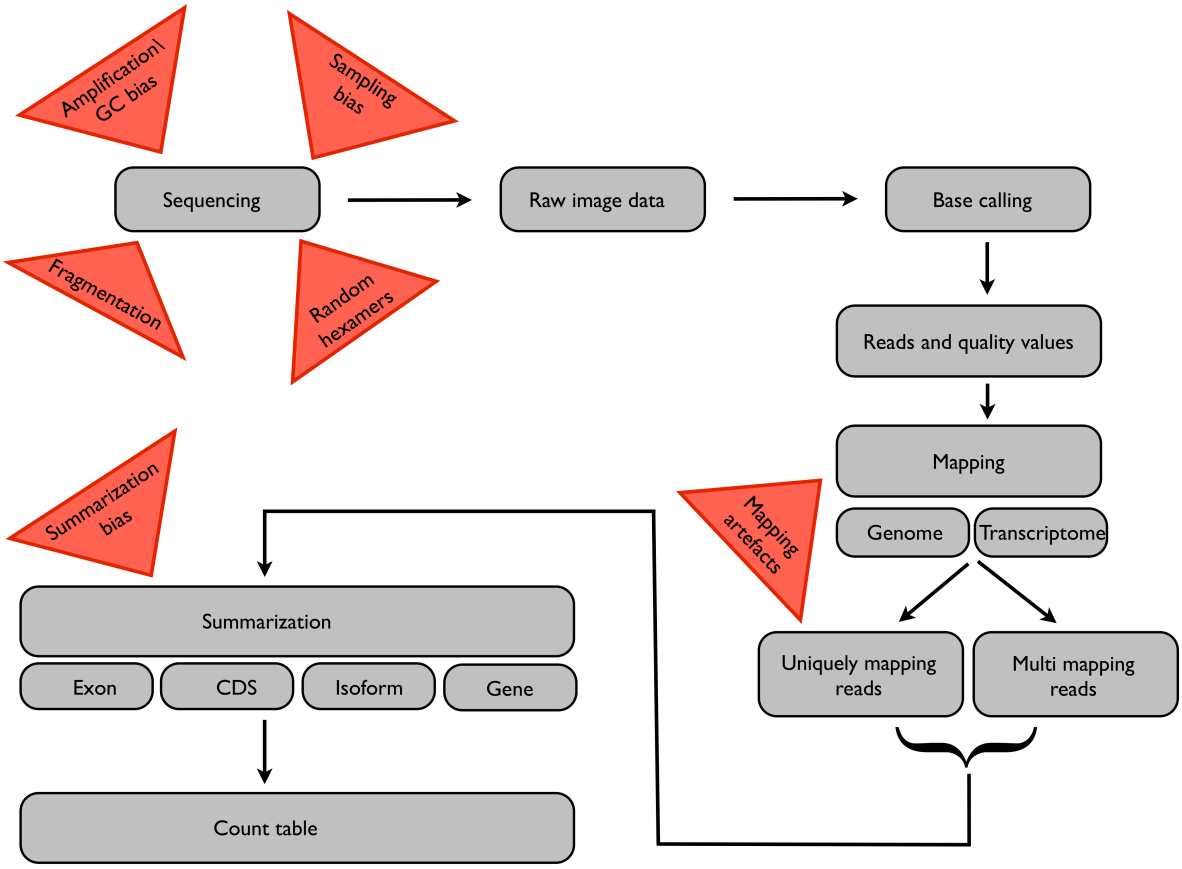

Figure 1.1: RNA-Seq workflow: Multiple biases may distort the composition of the biological sample before even entering the sequencing machine. Reads and corresponding quality values are obtained by application of the so-called base-caller from the raw image data. Subsequently mapping of the reads, summarization per entity of interest, normalization and, eventually, testing for differential expression takes place.

most likely linked to the PCR (Benjamini et al., 2012).

Priming method Typically first strand synthesis is primed with random hexamers. Yet, it has been shown that random hexamers, despite the name, tend to prefer certain sequences and thus lead to a distorted nucleotide composition in the resulting reads (Hansen et al., 2010).

Fragmentation method The default fragmentation method of Illumina is by using divalent cations under elevated temperatures. Other prominent fragmentation methods include nebulization and sonication which are known to perform differently on the body of the gene in contrast to the 5' or 3' end (Wang et al., 2009). This finding is supported by the work from Tauber and von Haeseler (2013) who compared three fragmentation methods, namely nebulization, sonication and random enzymatic digestion, in respect to the resulting coverage. They came to the conclusion that all methods are far from producing uniformly distributed fragments along the gene.

Sampling bias It may happen that very few genes collect the majority of the reads. These so-called key players then hinder the detection of the remaining, lowly expressed genes, by their dominance in the sequencing sample (Bullard et al., 2010).

Biases originating from the analysis workflow:

Summarization method Given two genes annotated at the same genomic coordinates but opposite strands and a read mapping to this location, it may occur that this read cannot be assigned unambiguously unless it brings along strand information. Similarly, reads that map to exons shared by multiple isoforms can be summarized on the isoform-level only by the aid of model-based methods. Thus, the summarization method may lead to wrongly assigned or, even discarded reads.

Mapping strategy Reads may contain sequencing errors and, additionally, may exhibit a certain genomic divergence to the reference. Therefore and since all alignment algorithms incorporate some amount of heuristics, it may happen that reads are assigned to the wrong position on the genome.

Chapter 2

Main Contributions of this Thesis

We have introduced the state-of-the-art technology for global gene expression screens, namely RNA-Seq together with its workflow in the wet-lab as well as the subsequent analysis pipeline. Additionally, we have elaborated on the different bias sources distorting the relationship of the genuine mRNA abundances and measured expression levels.
A wealth of methods already exists for each step of the standard RNA-Seq analysis workflow (Garber et al., 2011, Kvam et al., 2012, Lindner and Friedel, 2012) including bias correction (Ross et al., 2013). Due to the overwhelming amount of data produced by RNA-Seq speed is one of the top criteria for evaluating mapping software as well as subsequent processing methods. However, we believe that new knowledge may only arise from an in-depth analysis. Taken decisions within the analysis workflow must be challenged in the light of biological sensibility and statistical validity.

Thus we deliberately refrain from contributing another method to the already well filled software pool of mapping programs, normalization methods or models for differential expression inference. In this Thesis, we aim for breaking new ground by enlightening aspects of the analysis pipeline which have been neglected so far and by introducing a completely different point of view upon the sequencing data.

2.1 Exploration and Quality Control of Coverage Patterns

RNA-Seq analysis is centered around the count table. Mapping of the reads as well as summarization are merely a means to obtain this count table. The initial research question of the majority of global gene expression screens is to detect differential expression, a hypothesis which again bases on the count table. Yet, we note that the count table is actually a summary statistics, the sum of reads per entity of interest such as gene or exon.

Here we advocate not to neglect the unprecedented resolution of RNA-Seq which is reflected in the per-base coverage, the number of reads mapping to each base. By restricting the analysis to the count table rather than integrating the underlying coverage pattern valuable information is discarded. Since visual comparison of thousands of pattern is not feasible, we contribute a method to evaluate these patterns by means of the Fractal Dimension (FD).
While the FD has a long tradition in the context of self-similar objects we mainly focus on its property as measurement for the roughness of the underlying graph (Gneiting et al., 2012). We link the reliability of a coverage graph to its roughness which in turn can be evaluated by the FD. By doing so we unravel pitfalls while the library preparation and the subsequent analysis. We developed an R package (R-Project, 2013), namely *FractalQC*, to make this method publicly available.

2.2 The Sampling Process of RNA-Seq

Being aware of the conglomerate of biases RNA-Seq data is exposed to, rather than trying to account for each bias individually, we adopt a more pragmatic and global point of view.
We take the data as it comes along and regard it as sampling process which either takes place on the gene level (number of reads starting at each position of a gene) or on the transcriptome level (number of reads per gene). It turns out that these

sampling processes can be very well characterized by means of sampling formulas derived in the field of population genetics. Yet, they naturally fit the needs of current RNA-Seq data. Specifically, we use the Pitman Sampling Formula (PSF) (Pitman, 1995) which is a generalization of the well-known Ewens Sampling Formula (Ewens, 1972).
By means of the PSF we investigate whether the theoretical expectation of uniform coverage holds and how any deviation from uniformity can be evaluated. Additionally, by exploiting the urn representation of the PSF, the so-called Hoppe Urn, we can realistically simulate the distribution of reads within and between genes.

Notably, the PSF even enables us to quantify the absolute number of expressed - and yet undetected - genes. This is of particular interest since the set of expressed genes is highly variable and changes in dependence on the underlying organism, tissue and developmental state. More precisely, given a pilot sequencing experiment yielding a certain number of expressed genes, we address the question how many more expressed genes will be detected when sequencing another sample. Carrying this question to the extreme results in exploration of the boundaries of the respective transcriptome which are driven by the properties of the underlying biological sample. The benefit is of immediate practical value since knowledge about the necessary as well as sufficient sequencing depth in order to detect all expressed genes is crucial for a realistic experimental design.

Chapter 3

FractalQC: Exploration and Quality Control of Coverage Patterns

3.1 Introduction

The advent of next generation sequencing (NGS) technologies (Metzker, 2010) has turned RNA-Seq into the state-of-the-art protocol for global gene expression profiling. The initial research question of the vast majority of RNA-Seq experiments is comparison of RNA expression levels of two or more biological sources. Therefore the standard workflow of a RNA-Seq experiment involves mapping of the reads, summarization of the reads per entity of interest (e.g. gene, exon or coding region), normalization and, eventually, testing for differential expression. Thus, all statistical analyses for differential expression inference base on the count table containing the number of reads per entity. By doing so, important aspects of NGS technologies, namely the unprecedented and discrete resolution which is partially reflected in the per base coverage and, respectively, in the overall coverage pattern, is neglected. While numerous tools exist for explorative quality control of the coverage pattern per gene (DeLuca et al., 2012, Wang et al., 2012, García-Alcalde et al., 2012), no methodical approach is available to extract the information contained in the coverage pattern. Typically, the per base coverage only enters the differential expression analysis in terms of a kind of summary statistic - as read counts.

Similarly do other NGS analysis methods such as structural variation discovery, copy number variation detection (Alkan et al., 2011) or SNP calling (Nielsen et al., 2011) just rely on the per base coverage and do not integrate the knowledge of the overall coverage pattern in the analysis. Recently Lindner et al. (2013) demonstrated how valuable it may be to exploit the overall coverage more thoroughly. They fit mixtures of probability distributions to the so-called genomic coverage profile which is essentially a histogram over all per-base coverages. By doing so they are able to distinguish different contributors in a metagenomic sample.

In this work, we argue not be oblivious of the information contained in the coverage pattern of each gene, or more precisely, each isoform. In theory one would expect uniform coverage along the isoform resulting in a rectangular shape of the coverage regardless of the expression level of the underlying isoform. Yet, the observed coverage patterns may deviate greatly from this ideal due to the fact that the sequencing procedure is typically not exhaustive, meaning that not all mRNAs present in the biological sample get sequenced. Secondly the sequencing protocol itself may introduce several biases. Well known examples of induced biases include that random hexamer priming leads to a distorted nucleotide composition of the reads (Hansen et al., 2010) or that either GC-poor as well as GC-rich genomic regions tend to have underrepresented read counts (Benjamini et al., 2012). Besides non-uniform coverage originating from pitfalls of the sequencing protocol, the individual analysis steps may also introduce peculiar coverage patterns. Isolated stacks of reads might be caused by the mapping procedure (e.g. many reads are trimmed and as a consequence collectively assigned to the wrong position on the genome). Outdated annotation might lead to a sharp drop in the coverage e.g. due to a previously unknown alternative splicing event.
We contribute a method to explore and evaluate these patterns by means of the fractal dimension (FD). The FD has a long tradition with respect to the analysis of time series or transect data (Mandelbrot, 1982). Besides being a popular metric to quantify the roughness of a graph, the FD can even be linked to the variogram function if the underlying process is Gaussian (Gneiting et al., 2012). In this work, we ex-

ploit the fact that the FD can distinguish different shapes of the coverage graph and therefore enables us to identify all kinds of peculiar coverage patterns.

There are few other tools which also work on coverage patterns in the context of RNA-Seq. *rnaSeqMap* (Okoniewski et al., 2011) inspects coverage profiles in order to detect differential expression given comparable read counts. In order to do so, they introduce several normalization and difference measures for pair-wise comparison of coverage shapes. *rnaSeqMap* suffers from the fact that the performance of all presented difference measures depends on the underlying shape of the coverage. Furthermore the difference measures do not have a predictable range of values and thus are difficult to evaluate and interpret. *ReadSpy* (Hower et al., 2012) addresses the question whether reads are uniformly distributed within genes. By representation of the mapped reads as combination of their left-end mapping position and the corresponding fragment length, the sequencing process can be regarded as a two-dimensional Poisson process. Within this framework testing for uniform distribution of the reads is then realized by a χ^2-statistic. While ReadSpy is theoretically appealing, it is of limited practical use as typically even 'good coverage' patterns are far from being truly uniform. Additionally, *ReadSpy* needs paired end information and is therefore restricted to such data.

We introduce *FractalQC*, an application aiming for identification of peculiar coverage patterns. In contrast to the above-mentioned methods *FractalQC* does neither only ask the binary question of uniform vs not uniform coverage nor does it concentrate on the identification of alternative splicing events. *FractalQC* processes mapped reads (bam files) and generates an extensive report pointing out isoforms with interesting coverage patterns. Besides calculation of the FD and the area under the curve (AUC) for detection of those, *FractalQC* offers an extensive visualization of the coverage patterns garnished with meta-information like exon-intron structure and links to the corresponding entries in the annotation databases. Not only the entire set of known isoforms can be processed, also restriction to a subset such as genes e.g. involved in a specific pathway may be processed. Since the FD always ranges between 1 and 2 for one-dimensional data FD values are directly comparable

between experiments.
Thus *FractalQC* makes the analysis more transparent by generating explanations for peculiar coverage patterns originating from mapping artefacts, outdated annotation or issues of the summarization method. In addition, the performance of sequencing protocols in terms of evenness of coverage can be assessed by comparing the FD values for all isoforms across samples.

3.2 Material and Methods

The Fractal Dimension

The topological dimension of euclidean objects such as lines, surfaces and spheres is intuitively assigned as 1, 2 and 3. In this context, the dimension can be understood as the number of necessary coordinate axes to describe the underlying object. Yet, in nature, such ideal platonic objects are typically not observed but more complex and messy shapes are encountered. Mandelbrot (1982) introduced the FD as means to characterize irregular shapes where the integer-valued topological dimension fails to capture the complexity of the underlying shape.
Figure 3.1, containing simulated time series data, exemplifies this situation. Both panels show a Gaussian sample path but differ in the respective exponent of the covariance function. We observe that a smooth sample path results in a small FD (1.08 in Figure 3.1a) whereas a rougher sample path yields a higher FD (1.88 in Figure 3.1b). Additional intuition may be provided by the Gaussian Matérn process. This process is particular suited to demonstrate the relation of the FD to the appearance of the graph since the FD is here increasing linearly from 1 to 2 (Figure 3.2).
Figure 3.1 and 3.2 illustrate that graphs of one-dimensional profiles may differ substantially, possible graph structures range from smooth to space-filling curves. These properties of the graph are missed by the topological dimension that equals to one for all shown sample paths. In contrast, the FD evaluates the roughness of the graph and, given one-dimensional profiles, always ranges between 1 and 2. The larger the FD the rougher, the more space-filling the graph. While historically the

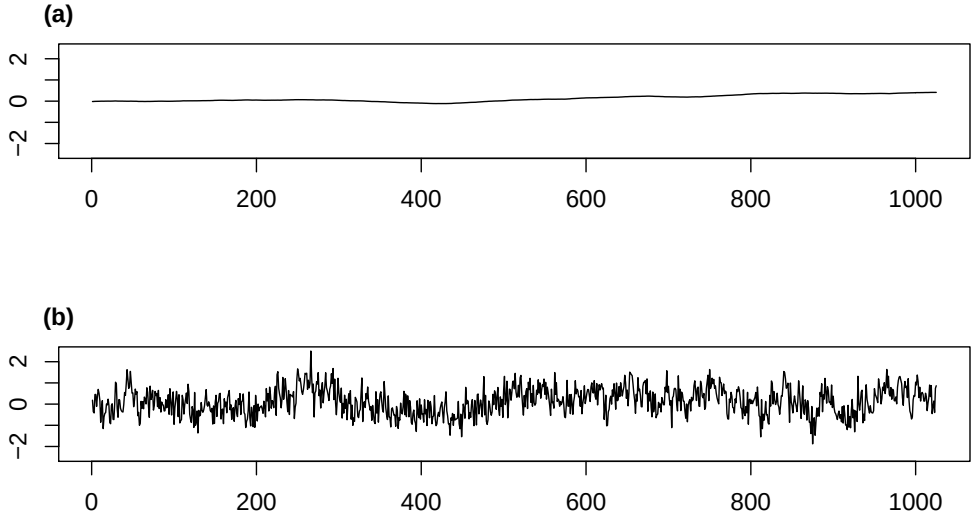

Figure 3.1: Gaussian sample paths of length 1000 from the powered exponential family. The covariance function is given by $\sigma(t) = exp(-t)^{\alpha}$ whereas $\alpha \in (0,2]$. The value of α determines the extent of space-filling of the curve, $\alpha = 1.9$ in (a) and $\alpha = 0.2$ in (b).

FD has been applied most prominently in the context of self-similar objects, we follow Gneiting et al. (2012) and emphasize that the FD can also be merely regarded as means to assess the roughness of a graph.

Calculation of the Fractal Dimension

Here we concentrate on data of the following form

$$X = \{(t, X_t) \in \mathbb{R} \times \mathbb{R} : t \in T \subset \mathbb{R}\} \subset \mathbb{R}^2$$

where X is the graph of a one-dimensional profile, such as a time series, or, in our context, the per base coverage pattern. The observation take place at $T \subset \mathbb{R}$ and are typically equally spaced. In the case of coverage data $t = 1, 2, \ldots, n$ whereas n is the length of the isoform.

The FD is usually defined as the Hausdorff dimension (Falconer, 1990). As this definition is rather abstract we refrain from presenting it here, instead we provide

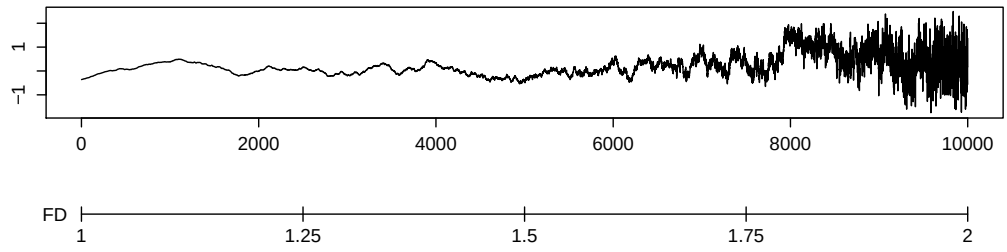

Figure 3.2: Sample path from a Gaussian Matérn process, FD varies linearly from 1 to 2 throughout time.

some intuition. When measuring the length of a one-dimensional object, the size of the ruler determines the actual length. This fact is best understood when thinking of a coast line whose length is typically not straightforward to determine. Therefore there is an implicit relationship between scale (size of the ruler) and length. Let $N(\varepsilon)$ be the number of squares with edge length ε necessary to cover the graph of a straight line. If the edge length is halved the number of squares necessary to cover the graph will double (see Figure 3.3a for an illustration). Hence the number of squares $N(\varepsilon)$ is related to the edge length ε by the following power-law: $N(\varepsilon) \propto \frac{1}{\varepsilon^1}$. The relationship is here linear because we have chosen a trivial example - a straight line living in the euclidean space. The FD is now exactly this exponent when the edge length ε goes to 0.

$$FD = \lim_{\varepsilon \to 0} -\frac{\log N(\varepsilon)}{\log \varepsilon}, \quad FD \in [1,2)$$

Figure 3.3b shows an example where the exponent is greater than 1.
A smooth, differentiable curve has topological and fractal dimension of one (Gneiting et al., 2012). However, the FD may also exceed the topological dimension. The rougher and the more space-filling the graph the greater the FD. See Figure 3.1 for two extreme cases whereas the FD equals to 1.08 in (a) and to 1.88 in (b).

Multiple methods exist for calculating the FD, see Gneiting et al. (2012) for a recent

Method	*Property*	*Scale*	*Scaling Law*	*Regime*
Boxcount	number of boxes: $N(\varepsilon)$	box width: ε	$N(\varepsilon) \propto \varepsilon^{-FD}$	$\varepsilon \to 0$
Madogram	variation estimator: $\hat{V}(l)$	lag: l	$\hat{V}(l) \propto l^{2-FD}$	$l \to 0$

Table 3.1: Methods for calculating the FD

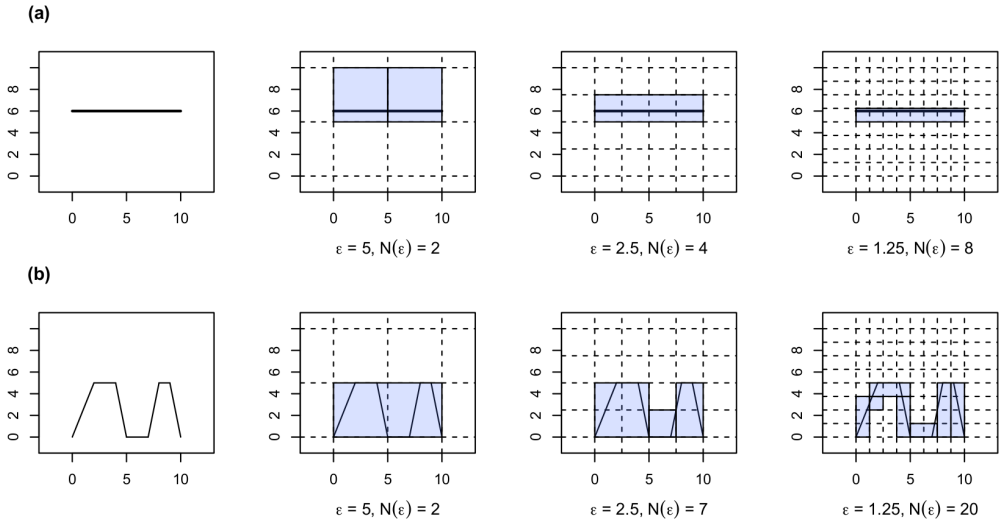

Figure 3.3: Covering with squares of decreasing edge length ε of (a) a straight line and (b) a rougher profile. Whereas $N(\varepsilon) \propto \varepsilon^{-1}$ in (a), the number of squares necessary to cover the graph increases significantly in (b) for $\varepsilon \to 0$.

review. All methods coincide with picking up the relationship of scale and length while both scale and length may also be interpreted in a broader context. While box-counting is probably the best known method and works as mentioned above we focus on the 'madogram' method. It has been shown that this method is the most suitable choice when the underlying process is not Gaussian. Table 3.1 gives an comparison of the box-count and the madogram method. In contrast to box-counting, the madogram method uses a variation estimator $\hat{V}(l)$ which basically sums up increments of lag l:

$$\hat{V}(l) = \frac{1}{2(n-l)} \sum_{i=l+1}^{n} |X_i - X_{i-l}|$$

We follow Gneiting et al. (2012) and just consider lags equal to 1 and 2. It has been shown that this choice minimizes the bias of the variation estimator given a Gaussian process. Additionally, it makes intuitively sense to regard the smallest scales as most important as the power-laws hold in the limit. The FD is then obtained by calculating the slope of the log-log plot of $\hat{V}(l)$ and l. See Figure 3.4 for a demonstration.

Here, we have to note that these power laws hold in the limit - for ε or, respectively, $l \to 0$. Yet, given real data, we have to deal with a limited resolution. In the case of genomic data the finest resolution possible is on the per base level.

Application of the FD to RNA-Seq data

In this work, we aim for identification of interesting coverage patterns. Since RNA-Seq does not yield only read counts per isoform but also resolution on the per-base level, the individual coverage pattern per isoform can be easily computed from the alignment of the reads. The coverage pattern of an isoform is basically an integer vector containing the number of reads mapping to each base of the isoform. The FD is calculated from the coverage patterns and allows immediate evaluation of its reliability. If the sequencing process and the used sequencing protocols were technically impeccable reads should be distributed uniformly within isoforms and

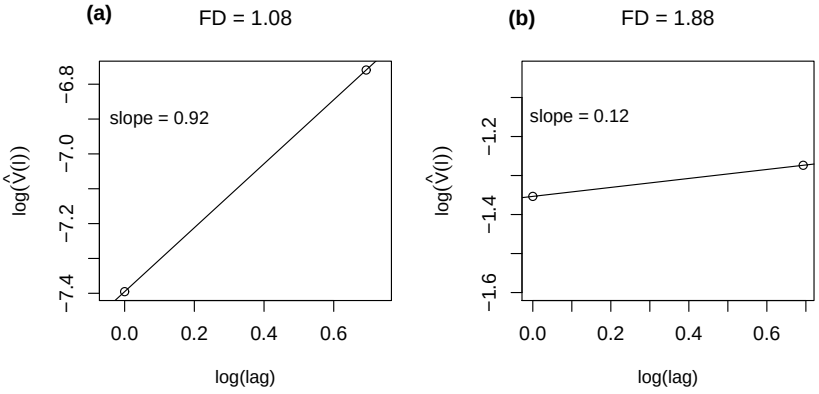

Figure 3.4: Estimation of the FD via the madogram method of the data shown in Figure 3.1a in (a) and, respectively, Figure 3.1b in (b). The considered lags are 1 and 2. The FD is calculated by 2 minus the slope of the linear regression.

coverage patterns would only differ in terms of expression level. Yet, given the constraints of current sequencing technologies, coverage patterns may deviate substantially from this ideal. However, exactly those non-uniform coverage patterns carry important information about the sequencing process and the subsequent analysis and may reveal possible pitfalls of the sequencing process or consequences of individual steps of the statistical data analysis.

Figure 3.5a shows a schematic coverage pattern that is typically regarded as reliable. The isoform is covered by reads at about the same level throughout the entire length. Increase and decrease of the coverage at both 5' and 3' end is as expected and induced by the sequencing protocol. In contrast, Figure 3.5b and Figure 3.5c depict coverage patterns calling for attention by their peculiar shape. Figure 3.5b might stem from a previously unknown splice event while Figure 3.5c indicates a severe mapping artefact. We recall that the FD equals 1 for a smooth, differentiable curve and is greater than 1 the rougher the graph. Since a 'good' coverage pattern is reflected in a rough graph it will be rated with a high FD. In contrast, any kind of sharp change in the coverage or regions with a coverage of 0 will lead to a more discrete and, as a consequence, smoother coverage profile, resulting in a low FD.

Thus the FD naturally pinpoints those patterns that are too smooth than expected given current state-of-the-art sequencing technologies.

While being unlikely, it may happen that coverage patterns such as in Figure 3.5d are observed. This typically occurs when the underlying isoform collects a very pronounced amount of reads and thus accomplishes such a smooth coverage. In this case the FD fails by rating this coverage pattern, being a perfectly smooth curve, with 1. Therefore we introduce, besides the FD, another score, namely the area under the curve (AUC). The AUC calculates the percentage of area covered by the underlying coverage pattern in relation to what could be covered given the data. More precisely, the AUC is the area of the coverage divided by the length of the isoform times the maximum of the observed coverage.
Finally, isoforms are ranked according to a score, namely S_{FDA}, that is composed as follows:

$$S_{FDA} = 3 \times rk(\text{FD}) + 1 \times rk(\text{AUC})$$

whereas *rk* is the rank (the ordinal number of each value when sorted in increasing order). The weights of the FD and, respectively, the AUC part have been optimized empirically. A low score indicates an interesting pattern. Isoforms with a low FD get a low score. Possible 'false positives' such as Figure 3.5d are downweighted by the AUC part (see Figure 3.5d and 3.5e for an illustration).

Features of FractalQC

FractalQC processes aligned reads in bam format (Li et al., 2009). Single as well as paired-end libraries are admissible. *FractalQC* is launched with one single command in the R console specifying the bam file of interest and all necessary arguments such as organism, single\paired-end libraries, strand-specificity, number of bins, number of figures per bin and potential restriction to a subset of isoforms (all options are explained in the following paragraphs). Eventually, a HTML report is created containing visual representation of all putative interesting coverage patterns as indicated by the score S_{FDA}. See Figure 3.6 for an overview of the *FractalQC* application.

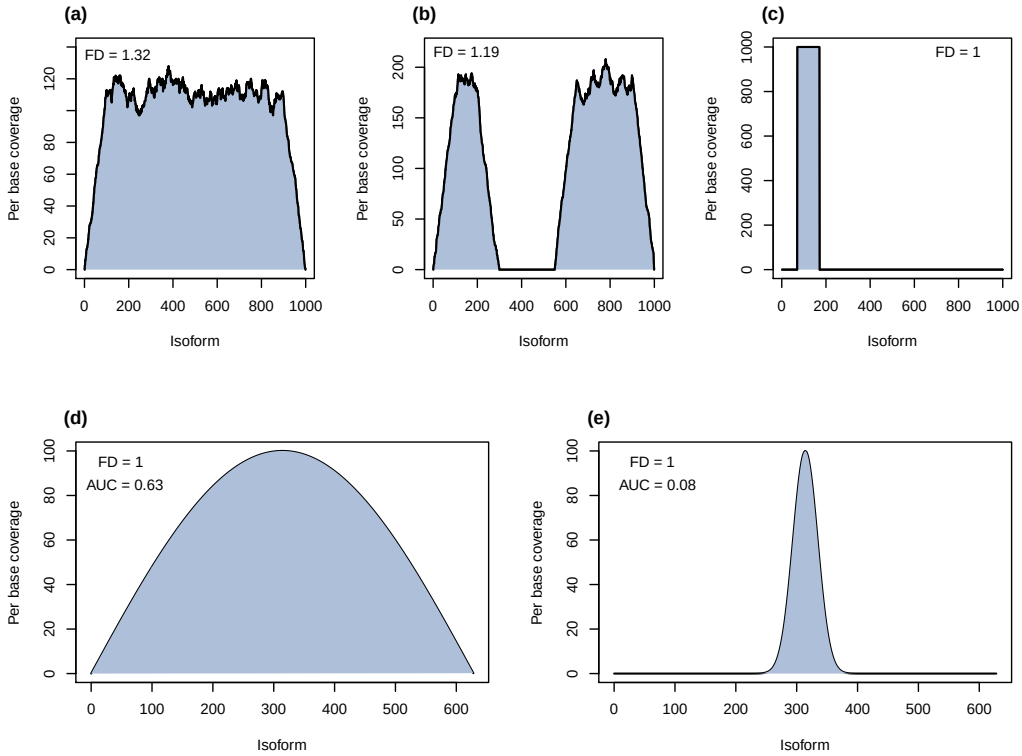

Figure 3.5: Schematic coverage patterns: 1000 reads of length 100 are (a) uniformly distributed along an isoform of length 1000. (b) reflects an alternative splicing event whereas (c) indicates a mapping artefact. The FD is similar in (d) and (e), such cases are filtered out by taking the area under the curve into account.

Choosing the Reference

FractalQC allows alignments to either genome or transcriptome. When mapping against the genome, any currently available mapping software may be used to output the bam file. However, mapping to the transcriptome is not that straight forward due to the difficulty where to correctly place reads that align to exons shared by multiple isoforms. Here, we advocate the use of *eXpress* (Roberts and Pachter, 2012), a software for probabilistic assignment of ambiguously mapping reads. The input for *eXpress* is a bam file containing all admissible alignments for all known isoforms of the organisms of interest. *eXpress* then outputs, besides many other features, a single alignment for each read sampled at random based on the alignment likelihoods calculated by *eXpress*.
Both reference choices are not satisfying. When mapping against the genome reads that can be allocated unambiguously to a certain exon might get excluded later on in the summarization step when this exon is shared by multiple isoforms. In contrast, alignment to the transcriptome via *eXpress* always includes a modeling step and thus the exported alignments are just as inferred by the model.
In general, *FractalQC* is currently restricted to organisms whose annotation is available in Ensembl (Flicek et al., 2013).
The FD is calculated on basis of the coding sequence of each isoform rather than over the entire genomic range. Including introns would add another, non-necessary, layer of complexity and hinder quick interpretations of the coverage pattern. Therefore we opt for using the coding sequence.

Strand Specificity

FractalQC can handle non strand-specific data as well as strand-specific data. In the former case reads contribute to the coverage pattern of the isoform they have been assigned to according to the mapping coordinates - regardless if the strand matches. Given a strand-specific library and alignment to the genome, *FractalQC* applies the following procedure as typically it is not known which strand has been sequenced. *FractalQC* checks if more than 80% of the reads are captured in known

gene models. If not, the strand tag is flipped for all reads. Again, the number of reads captured in known gene models is calculated. If this number is greater than 80% it is taken as granted that the strand of the reads should be flipped and the strand tag is thus treated accordingly. If the number of reads captured is still low, a warning is issued and decision upon strand flipping is based on which option delivers more captured reads.
If a strand-specific library is mapped to the transcriptome the procedure is slightly different. As mRNA sequences are typically stored in 5' to 3' direction in the databases all reads should have the same orientation - as either the coding or the non-coding strand has been sequenced. Thus *FractalQC* considers the strand of the majority of the reads as the correct strand and omits all other reads. *eXpress* also offers to take care of this issue by setting the appropriate option.

Selection of Isoforms

FractalQC uses all known isoforms by default. To facilitate the interpretation of the results isoforms are binned according to their read counts. Equal-sized quantile bins with a step size of 10% are chosen. Thus the 10th bin contains those isoform with read counts greater than the 90% quantile of the overall read count distribution. See Figure B.2 for a visualization. This allows to go through interesting patterns conditioned on the read counts. The incentive for binning is to make those isoforms with high reads counts and low FD more accessible as exactly those key players are prone to lead to wrong interpretations or hypotheses. For each read count bin isoforms are then ranked according to the score S_{FDA}.
However, the analysis can also be performed on a subset of isoforms such as all isoforms involved in a specific pathway. In this case no binning takes place, all isoforms are ranked according to the score S_{FDA}.

Output

FractalQC automatically creates a directory in the current path containing an extensive HTML report, a csv file and a parameter file. The HTML report comprises three sections. First several summary statistics are presented such as a histogram of

the read counts and a figure showing the percentage of isoforms versus percentage of captured reads. Secondly, as the read counts are processed in equal-sized quantile bins, the bins are visualized within the overall read count distribution to give an intuition for the absolute read count values. Finally, coverage patterns for each read count bins are presented, ordered according to the score S_{FDA}. Coverage patterns are always displayed over the coding region, regardless of the respective mapping procedure. The user can determine which bins and how many figures for each bin are presented. Meta-information such as exon boundaries, associated gene name and chromosome location is incorporated into each figure. Besides, each coverage pattern figure is garnished with a link to the corresponding entry in the ENSEMBL database.
Additionally, a csv file is available containing the read counts and corresponding FD values. Finally, a parameter-file contains status and warning messages created during the analysis. Screenshots of the HTML report are provided in Appendix B.

Availability

FractalQC is available as R package (R-Project, 2013) from the author and submitted soon to Bioconductor (Gentleman et al., 2004).

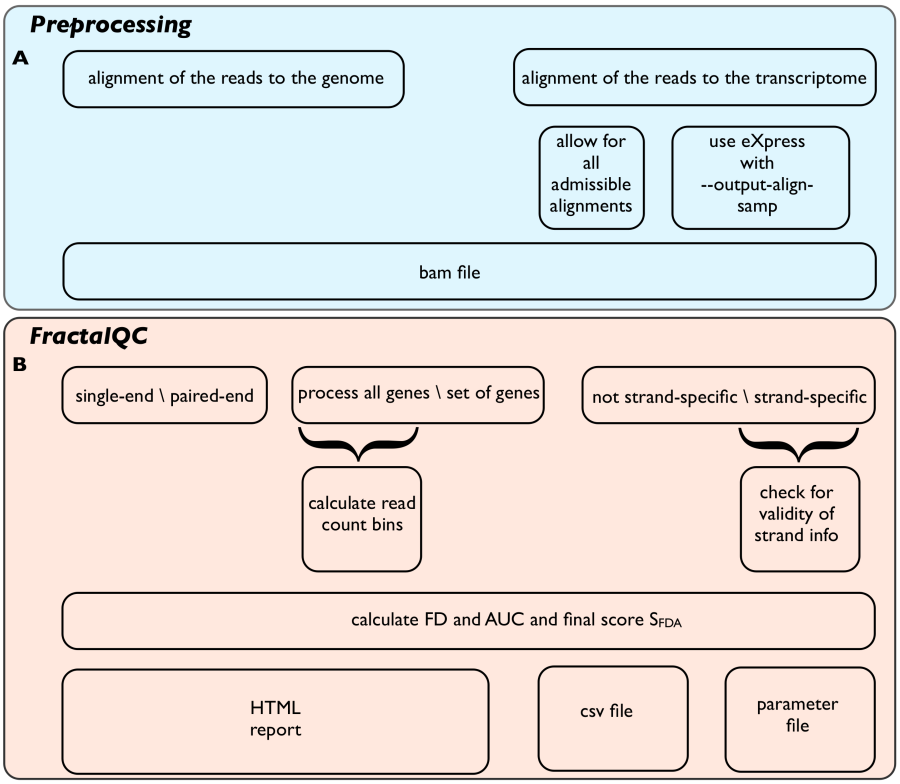

Figure 3.6: Overview of the workflow of *FractalQC* and the required preprocessing. A: Preparing the RNA-Seq data for *FractalQC*: Mapping of the reads to either genome or transcriptome. In the latter case *FractalQC* advocates the use of *eXpress* which needs as input a bam file containing the alignment of all reads versus all known isoforms. The option '–output-align-samp' urges *eXpress* to output one single alignment for each read proportional to the likelihoods as calculated by the method. This bam files serves then as input for *FractalQC*. B: Workflow of *FractalQC*. Bam files are read in, read counts and coverage patterns for all isoforms present in the current annotation are computed. The analysis may be restricted to a predefined set of genes. If strand-specific data is submitted strand information of the mapped reads is cross-checked with the annotation. Eventually, FD, AUC and the final score S_{FDA} is calculated for all isoforms. Isoforms are ranked according to increasing S_{FDA}. A HTML report is created in the working directory. Additionally, a csv files containing isoform IDs, FD, AUC and S_{FDA} is exported facilitating further processing of the scores. The parameter file lists all occurring warning and status messages of the analysis.

3.3 Results

Case study 1: Duplication rates correlate with overall FD levels

To demonstrate the functionality of *FractalQC* we process a subset of data from Risso et al. (2011). Strand-specific RNA-Seq libraries were prepared from *Saccharomyces cerevisiae* growing in three different media. We select two biological replicates for each medium, Table 3.2 gives an overview of the experimental design together with some mapping statistics. Reads were mapped with Bowtie 0.12.9 (Langmead et al., 2009), only uniquely mapping reads were allowed. Isoforms were retained where (i) read counts are greater than the 10% quantile of the read count distribution of the respective lane, and, (ii) when the isoform may be covered throughout its entire length at least theoretically. More precisely, read counts times read length must be greater than the isoform length. Subsequently read counts and FD were calculated for all isoforms.

Figure 3.7 shows the overall FD as well as the read count distribution for the six libraries. The read counts reflect the number of mapped reads. Since G2 and G3 have the overall highest read counts it is legitimate to assume that these two libraries also come along with more balanced coverage patterns. The rational is as follows: the more reads the less probable it is to observe gaps in the per base coverage pattern unless these gaps are due to some systematic error. Thus we expect that G2 and G3 exhibit rougher coverage patterns (compare to Figure 3.5a) resulting in a higher FD. Yet, we do observe a different picture as shown in Figure 3.7b. When inspecting the data more closely it turns out that all samples have a high duplication rate, in particular G2 and G3. Note that the term 'duplication' may be misleading since it is actually referring to not only doubling but any level of replication. Yet, this term has been established (DeLuca et al., 2012, FastQC, 2013). Picard (Picard-Tools, 2013) was used to quantify the level of duplication. To be precise, duplication is here understood in respect to the 5' mapping coordinates of the reads together with its orientation. In both Y1 and Y2 74% of all reads are marked as duplicates, 71, respectively 70% in D2 and D7. In contrast, G2 and G3 contain 94 and 95% of

Library Name	Medium	# reads $[10^6]$	# mapped reads $[10^6]$
Y1	rich	3.6	1.8
Y2	rich	4.1	2.1
D2	minimal	15.4	3.3
D7	minimal	15	2.9
G2	Glycerol	8.4	5.7
G3	Glycerol	8.2	5.7

Table 3.2: Overview of yeast dataset from Risso et al. (2011)

duplicated reads. A certain level of replication is expected in RNA-Seq data either due to low-complexity input samples or, due to the fact that typically very few genes collect the majority of the reads. These big players will then naturally cause a certain level of duplication. Still, such high percentages as in G2 and G3 are remarkable. Therefore we suspect that the overall lower FD levels in G2 and G3 are caused by the high duplication rates of these libraries leading to a more discrete coverage pattern.

In summary, we conclude that the FD is able to indicate quality issues of sequencing samples by comparing read counts and FD values. One major advantage of the FD lies in the fact that it does not rely on ad-hoc thresholds calibrated on the individual data set. As a consequence multiple samples are easily compared as the FD always ranges between 1 and 2 for one-dimensional profiles.

Identification of interesting coverage patterns

Here we want to exemplify the use of the FD as means to uncover pitfalls while sequencing or processing of the data. TDH3 is identified as interesting candidate by filtering for low variance read counts and low variance FD over all six libraries. We observe that the coverage patterns look very similar for all investigated samples, see Figure 3.8. In particular, zero coverage regions are consistent between samples. This indicates any kind of systematic error occurring either while the se-

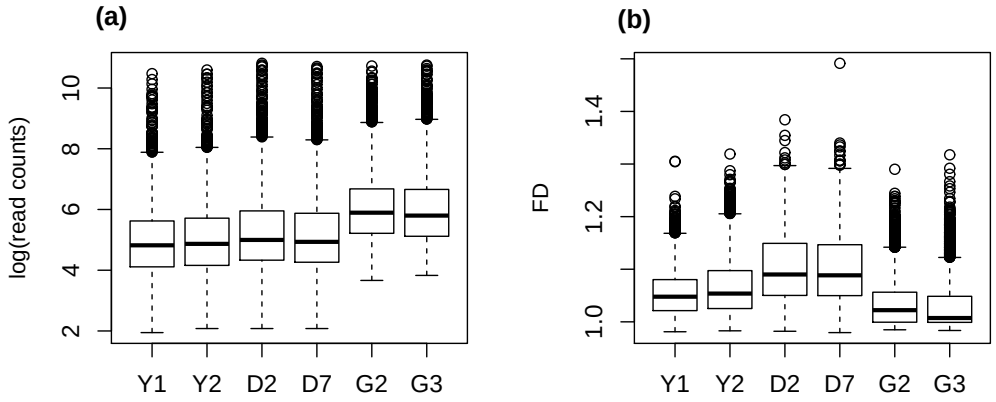

Figure 3.7: Read count distribution (a) and FD distribution (b) of two biological replicates for each medium. The higher the FD the more space-filling, the better the coverage graph.

quencing process or in the analysis afterwards. When checking the annotation of TDH3 we realize that there is another gene, namely TDH2, with nearly identical nucleotide sequence (96% identity). Therefore we reason that these zero coverage regions are a consequence of the mapping strategy - ambiguously mapping reads are discarded. Thus, we conclude that the read counts for TDH3 as well as TDH2 are heavily under-estimated. Depending on the underlying research question, it may be advisable to, rather than mapping the reads to the genome, use the transcriptome as reference. Software such as *eXpress* (Roberts and Pachter, 2012) may then be used to infer the most probable origin of reads mapping to multiple region on the genome.

Another example is presented in Figure 3.9. RRP42 is selected by filtering for low variance read counts and, in contrast to the former example, for highly variable FD. RRP42 seems to be weakly but consistently expressed over all samples. Yet, we do observe a striking difference in the distribution of the reads over the length of the isoform. Given the expression level and the read length of 36bp coverage patterns look reasonably good for the first four libraries. However, G2 and G3 exhibit a much more discrete coverage graph. This might be explained by the higher duplication rates of these libraries.

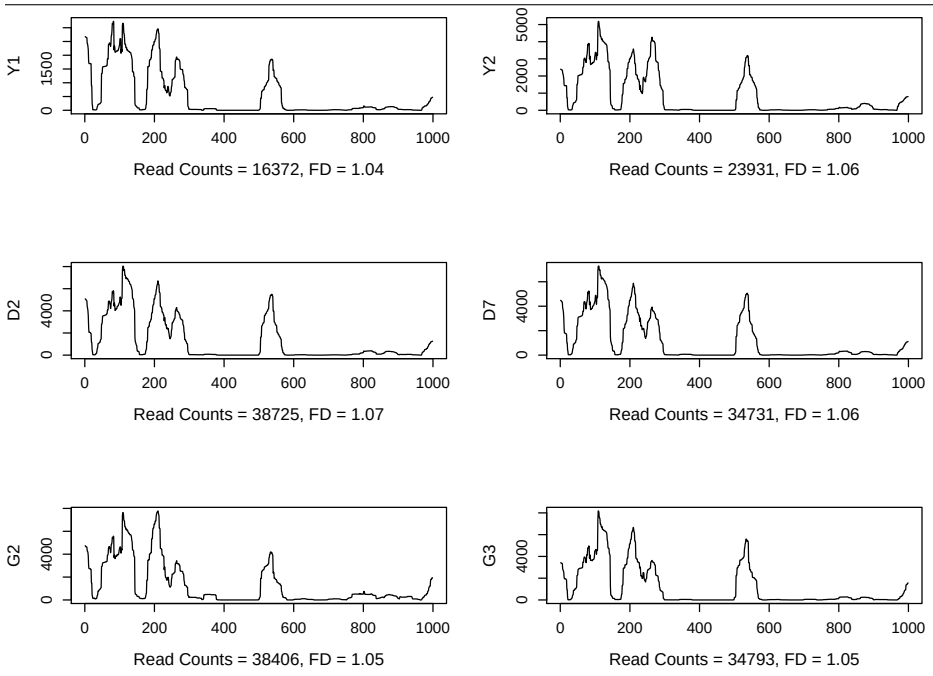

Figure 3.8: TDH3: comparable read counts and FD.

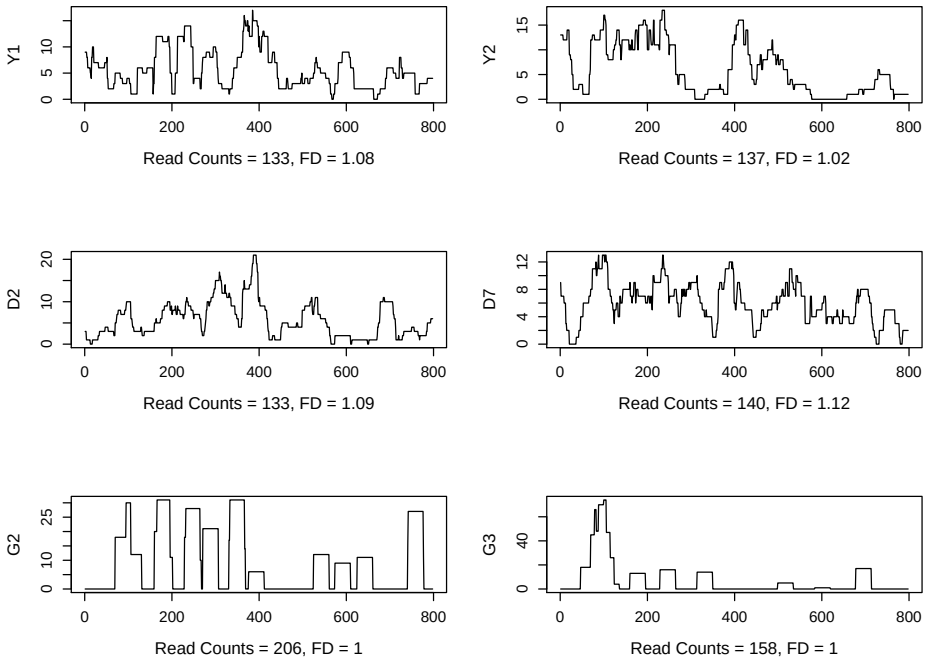

Figure 3.9: RRP42: comparable read counts but varying FD. While the absolute variance of the FD values may not appear particularly high, the overall range of the FD must be taken into account as shown in Figure 3.7b.

Case study 2: Comparison to *ReadSpy*

ReadSpy (Hower et al., 2012) is a tool that aims for quantification of the uniformity of mapped reads. Given paired-end RNA-Seq data, mapped reads are represented by their left-end mapping coordinate and the length of the fragments in $\mathbb{R}^2$. This point set then forms a two-dimensional Poisson Process. The independence property of the Poisson Process is exploited as follows: Testing for uniformity is conducted stepwise by first splitting the two-dimensional data set into horizontal strips. Each horizontal strip is then partitioned into several subintervals. A χ^2-statistic is used to test the null hypothesis of uniformity by calculating it for all subintervals within each horizontal strip and subsequently taking the sum over all horizontal stripes. Naturally the width of horizontal strips and the subintervals drive the test-statistic. Hower et al. (2012) recommend the following choice: each horizontal strip shall contain at least 200 data points which is then partitioned into 20 subintervals. Finally a p-value is computed indicating whether reads are uniformly distributed.

Hower et al. (2012) benchmarked their method on data from Levin et al. (2010). This data set consists of several strand-specific yeast RNA-Seq libraries differing in the respective library preparation. In the original paper Levin et al. (2010) aimed for a comprehensive comparison of these protocols in terms of several criteria such as library complexity, strand specificity and evenness and continuity of coverage among others. One method, namely the dUTP protocol, gave the overall best performance, also for the criteria of evenness and continuity of coverage.

Hower and colleagues re-evaluated this dataset, focusing on genes with more than 200 reads. *eXpress* (Roberts and Pachter, 2012) was used to map reads against the yeast transcriptome. dUTP was again confirmed as being the superior method as the overall p-value range is elevated in dUTP in comparison to the other library preparations (Table 3.3). Yet, when inspecting the p-values more closely, it turns out that the null hypothesis of uniformity of reads is not rejected (p-value ≥ 0.05) for only 2% or even less of all investigated genes. While this result is absolutely plausible it indicates that such a rigorous mathematical approach might not be the

	Quantile	
	50%	95%
Control	8.1×10^{-15}	1.4×10^{-3}
Control dT	7.4×10^{-20}	3.1×10^{-4}
dUTP	1.1×10^{-12}	6.2×10^{-3}
dUTP dt	1.8×10^{-24}	2.4×10^{-5}
Hybrid	7.4×10^{-75}	8.8×10^{-15}
NNSR	2.7×10^{-211}	5.4×10^{-42}
NNSR noactD	1.4×10^{-108}	4.6×10^{-24}

Table 3.3: Overview of the p-value distribution (based on $\sim$ 1500 genes) as output from *ReadSpy*.

most sensible way to look at such data.

We used our method, *FractalQC*, to re-evaluate the data once more. The FD is calculated for all genes as selected and preprocessed by *ReadSpy* in order to stay as consistent as possible. Figure 3.10 shows a comparative boxplot of the FD levels of the used libraries. We observe that dUTP together with the Control library exhibit the highest FD values. Therefore, as a high FD stands for a rough, non-discrete coverage pattern, we agree with the previous drawn conclusion of the good performance of dUTP.

While *ReadSpy* as well as *FractalQC* share the same overall conclusion Figure 3.11 illustrates the difference in outcomes and interpretation when looking at individual coverage patterns. Since both genes get assigned a p-value much smaller than 0.05 the null hypothesis of uniformity of reads is rejected. Yet, we do observe a striking difference in the coverage patterns. While reads are not uniformly distributed from a mathematical point of view, the coverage pattern displayed in Figure 3.11a is still satisfying from a practical perspective. In contrast, Figure 3.11b indeed indicates some kind of pitfall while sequencing or processing.

In summary, we have to emphasize that *ReadSpy* and *FractalQC* are not exactly comparable since *ReadSpy* concentrates on the quantification of uniformity of reads. Instead, *FractalQC* aims for a broader quality control of coverage profiles. Even though both methods come to the same overall conclusion, *FractalQC* addresses

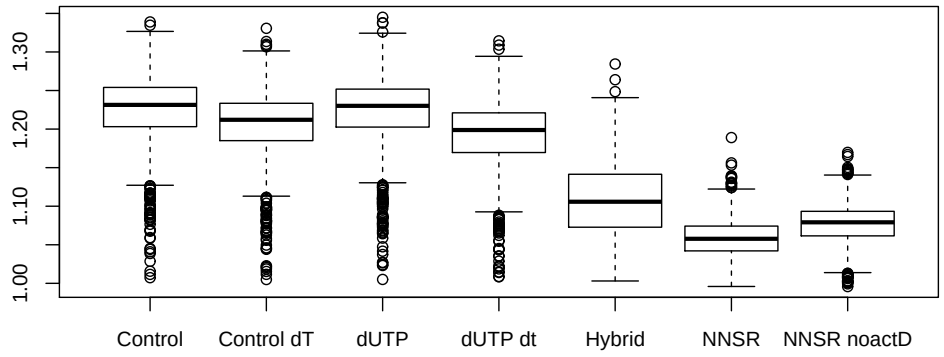

Figure 3.10: FD values for paired-end libraries from Levin et al. (2010). Evaluation is based upon 1464 genes as selected and preprocessed in *ReadSpy*.

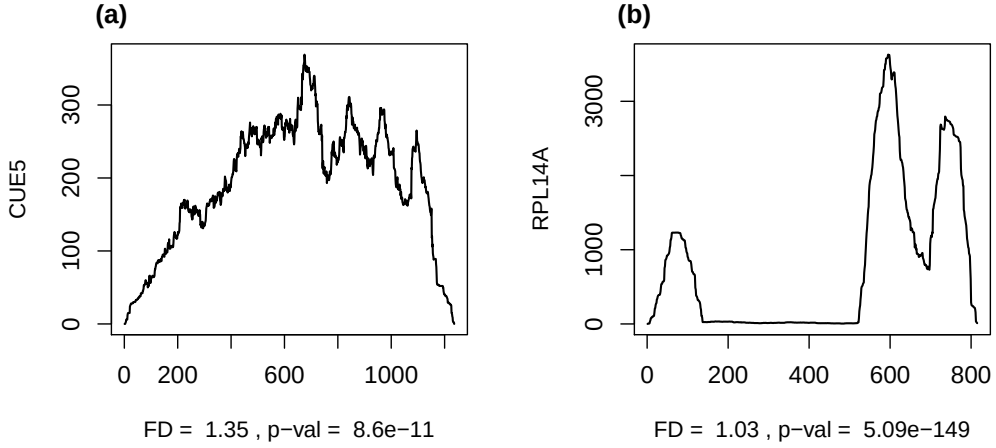

Figure 3.11: Two coverage patterns from the dUTP library. When computing the scores S_{FDA} for this library, (a) can be found on rank 1418 (from 1464 genes), (b) is on rank 1. In contrast, *ReadSpy* is unable to detect these differences, both coverage patterns have a p-value < 0.05 implying rejection of the null hypothesis of uniform coverage.

Protocol	Reference	Strand specific	# reads $[10^6]$	# mapped reads $[10^6]$
RNA Ligation	Genome	yes	24.5	13.2
RNA Ligation	Transcriptome	yes	24.5	15.12

Table 3.4: Overview of the strand-specific RNA Ligation yeast library from Levin et al. (2010)

the needs of RNA-Seq coverage data while less exact more realistic.

Case study 3: Further applications of *FractalQC*

In order to illustrate the versatility of *FractalQC* in more detail, one single-end library from Levin et al. (2010) is chosen, namely RNA Ligation, and processed. Reads are mapped with Bowtie 0.12.9 (Langmead et al., 2009), only uniquely mapping reads are considered. This restriction to unambiguously mapping reads is a popular strategy when mapping data to the genome. Therefore we follow this approach in order to mimic standard analyses. See Table 3.4 for an overview of the data.

Choosing the appropriate reference

First, we compare coverage profiles when mapping the data to the genome or to the transcriptome. In the latter case *eXpress* (Roberts and Pachter, 2012) takes care of handling multi-mapping reads. RPL7A, a protein component of the large 60S ribosomal subunit, is located on chromosome VII with a coding length of 735bp distributed among three exons of length 11, 94 and 630. This particular isoform, RPL7A, calls our interest due to its low FD and rather high read counts.
We observe a prominent peak in the second exon when mapping the data to the genome (Figure 3.12a). The zero coverage stretch is most probable due to the fact that RPL7A is nearly identical to RPL7B (97% sequence identity). Difference in nucleotide content between the two isoforms can only be found at the very beginning and end of the isoform. Since we discard ambiguously mapping reads regions

with highly similar sequences between isoforms will obtain zero coverage.
Given this data we are not able to tell whether the second exon is indeed more expressed than the other two exons or, alternatively, if this peak will actually disappear when reads mapping to multiple isoforms are considered. In order to check this hypothesis we map the data to the transcriptome and rely on the model of *eXpress* to distribute multi-mapping reads. Figure 3.12b shows the corresponding coverage pattern. We observe that while the zero coverage regions gets filled up as anticipated the high coverage peak of the second exon remains present. These results leave room for multiple interpretations: (i) it may be that the second exon is indeed higher expressed indicating some kind of previously unknown alternative splicing event, (ii) even *eXpress* does not manage to handle the situation correctly when having to deal with two isoforms exhibiting such similar sequence content, (iii) the peak may also arise from a bias such as induced by the fragmentation method. Typically the RNA is fragmented into smaller pieces before being subjected to sequencing. Commonly used fragmentation methods include enzymatic digestion, nebulization, sonication or fragmentation using divalent cations under elevated temperatures. The latter is applied in the RNA Ligation library we are dealing with. Yet, it has been reported that this fragmentation method tends to over-fragmentation due to its fast reaction rate (Kumar et al., 2013). In general, each method comes along with its specific preferences for breaking points within the RNA. Hence, the region around the start of the second exon may appear particular attractive for the applied fragmentation method and is therefore over-represented in the sequencing sample.
Any kind of definite answer will require in-depth check-up on the used sequencing protocol and\or knowledge about the splicing apparatus of *Saccharomyces cerevisiae*.

The merit of strand-specific libraries

Here we compare coverage profiles when mapping the data to the genome and then, while the summarization step, treat the data as being either strand-specific or ignorant of strand information. We exemplify the advantages of strand-specific libraries by means of the SSB1 isoform. SSB1 is a cytoplasmic ATPase, located on the

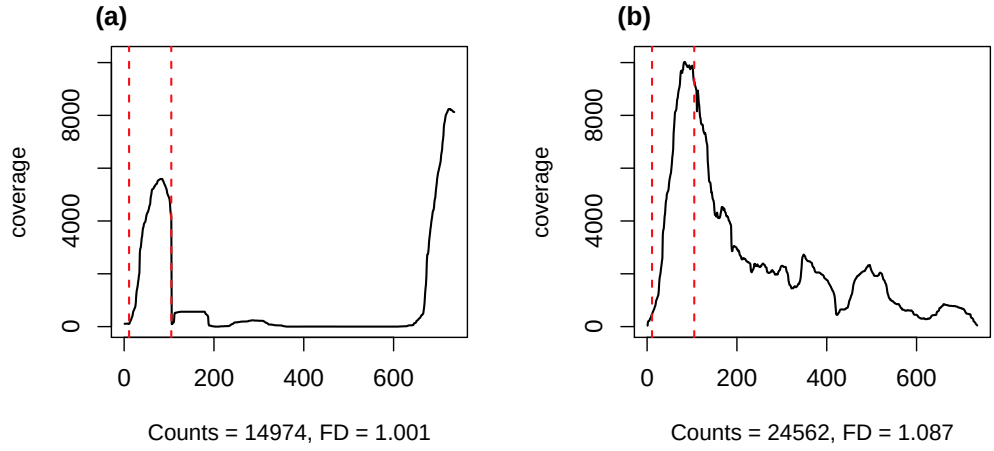

Figure 3.12: Assessing the coverage profile of RPL7A when mapping the data (a) to the genome or, respectively, (b) to the transcriptome. Red dashed lines indicate exon boundaries.

forward strand of chromosome IV, the coding length is 1842bp. The corresponding strand-specific coverage pattern is shown in Figure 3.13a. While the read counts are pretty high zero coverage regions indicate some mapping issue. Indeed, when doing some check-up on SSB1, we note that SSB2 exhibits 98% sequence similarity. As we pursue a mapping strategy of retaining only those reads that map unambiguously, all reads mapping to these identical regions will get lost.

Once the reads are mapped a summarization step follows in order to quantify the expression level per entity of interest such as exon, isoform or gene. Typically the counting of the number of reads per gene relies on the mapping coordinates and the strand information. When no strand information is available assignment is restricted to the information contained in the mapping coordinates. It might occur that genes have the same or very similar genomic coordinates but only differ in their orientation, one gene is located on the forward strand while the other is present on the reverse strand. In this case, given a library which is not strand-specific, reads that map to these regions do not get considered in the summarization step, as it is not clear to which gene they belong to. In contrast, having a strand-specific library at hand, reads can be unambiguously assigned to the correct gene. See Figure 3.14

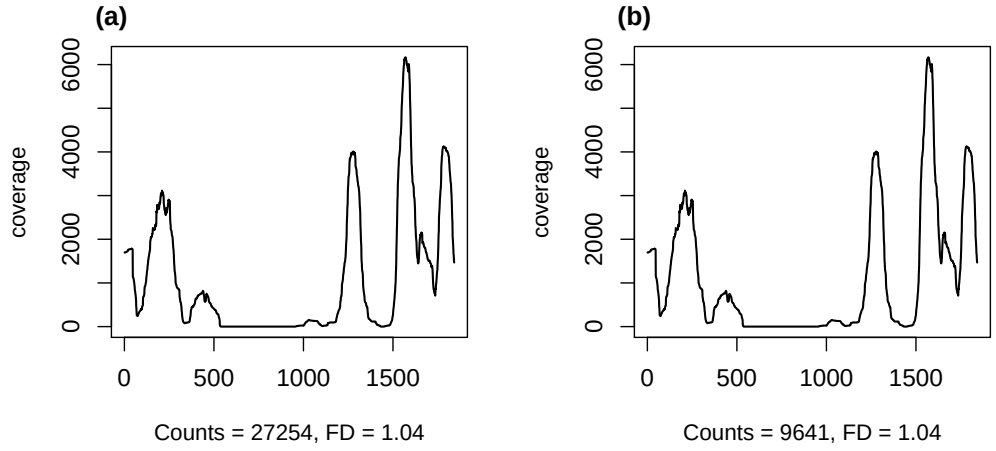

Figure 3.13: Coverage pattern of SSB1 when the library is regarded as (a) strand-specific or, respectively, (b) not strand-specific.

for an visualization.

Coming back to our example, we observe that SSB1 has the same coverage pattern but much less read counts when treating the data as not strand-specific (see Figure 3.13b). It is not surprising that the coverage patterns are equal since the mapping procedure is identical. Yet, there is a subtle differences between the two graphs: in Figure 3.13a the per-base coverage of the forward strand is depicted whereas Figure 3.13b shows the per-base coverage pattern of both strands collapsed. The difference in read counts is explained by the fact that the gene YDL228C is annotated on the opposite strand of SSB1. 98% of the coding region of YDL2882 is overlapped by SSB1 which then continues by about 1200bp. YDL2882 is described as 'dubious open reading frame' in Ensembl indicating a questionable annotation. Thus, if the underlying sequencing library is not strand-specific, reads that map to the genomic region shared by SSB1 and YDL228C will be lost.

In summary, we can conclude that the advantage of strand-specific libraries is striking. *FractalQC* facilitates the evaluation of the effect of different analysis steps such as mapping or summarization.

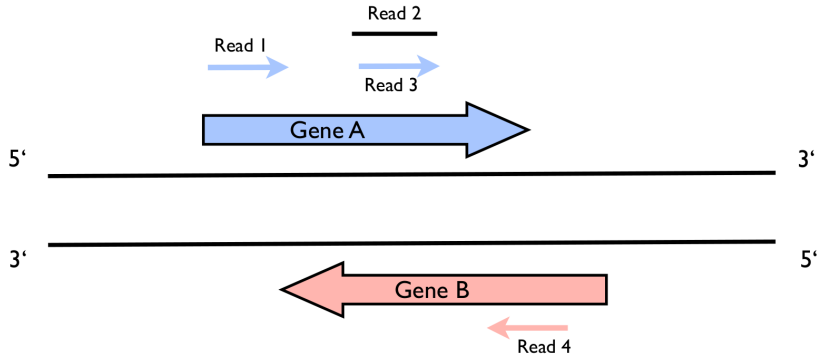

Figure 3.14: Gene A and B are sharing overlapping genomic coordinates and are placed on opposite strands. Arrows stand for different reads whereas the arrowheads indicate the orientation of the reads. Read 1, 3 and 4 are derived from strand-specific libraries and can thus be assigned unambiguously. In contrast, Read 2 does not bring along strand information and could therefore belong to either of the genes. As no definite assignment is possible Read 2 is discarded.

Sequence quality and mapping artefacts

The last use-case concerns the effect of the sequence quality on the mapping. When sequencing a biological sample, the sequencing machine, respectively the base-caller, provides a confidence measure for each called base. These sequence quality values (Q) give information about the accuracy of the base-caller. The range of Q depends on the underlying sequencing technology and even on the version of the used sequencing pipeline. In the case of our data, Q may vary from 0 to 40 (Illumina , 2013). $Q = -10\log 10(e)$ holds whereas e stands for the inferred error probability of the base caller. Thus a Q value of 10 corresponds to a base calling accuracy of 90%, a Q value of 20 to 99%, a Q value of 30 to 99.9% and, eventually, a Q value of 40 to 99.99% accuracy.

ADH2 provided the incentive for closer inspection as it has a low FD paired with high read counts. Motivated by the zero coverage stretches we immediately include ADH1 in the analysis as well. Both genes have a coding region of length 1047bp and exhibit 89% of sequence identity. The left panel in Figure 3.15a shows the cor-

responding coverage patterns when mapping the data to the genome and, respectively, to the transcriptome (Figure 3.15b). We observe that ADH1 collects a high amount of reads while ADH2 is much less expressed. Additionally, the expression of ADH2 is concentrated on two peaks (at 200 and 400 bp) while the remaining coverage is close or equal to zero. When mapping the data to the transcriptome the coverage patterns look basically the same.

As a consequence the question regarding the reliability of the coverage pattern and the read counts of ADH2 arises. Multiple hypotheses are available: (i) while being unlikely, it might be that this peculiar coverage pattern is derived from a previously unknown splicing event, (ii) alternatively, ADH2 is just expressed at a very low level and all reads contributing to these two peaks are wrongly assigned. The latter assumption is supported by the fact that 70% of the coding region of ADH2 has a coverage of at least 1. Additionally, when inspecting the alignment statistics of ADH2 we realize that 54% of the reads map with more than 3 mismatches. In contrast, when computing this alignment statistic over all mapped reads in our sequencing sample, only 25% exhibit more than 3 mismatches. According to the alignment policy of Bowtie (Langmead et al., 2009), a high number of mismatches is only to be explained with overall very low quality values. This is due to the fact that Bowtie limits the number of mismatches in the so-called seed - the first 28 bases - to 2 by default. Additional mismatches are only admissible when the sum of the quality values of the mismatches remains smaller than 70. This goes hand in hand with the observation that the overall quality values drop significantly towards the end of the reads. Specifically, the median of the quality values ranges from 33 at the 5' end of the reads to 4 at the 3' end of the reads. Thus we conclude that these peaks of ADH2 are not a genuine reflection of the underlying expression level. Instead we suspect that the majority of the reads mapping to ADH2 is placed wrongly. Yet, the height of the peaks can not be fully explained by the mapping statistics. There is still a substantial amount of reads, 46% , mapping with a decent number (≤ 2) of mismatches to ADH2. Therefore we suspect that another kind of bias such as amplification bias plays a role. In summary, we hypothesize that ADH2 is indeed expressed but at a very low level.

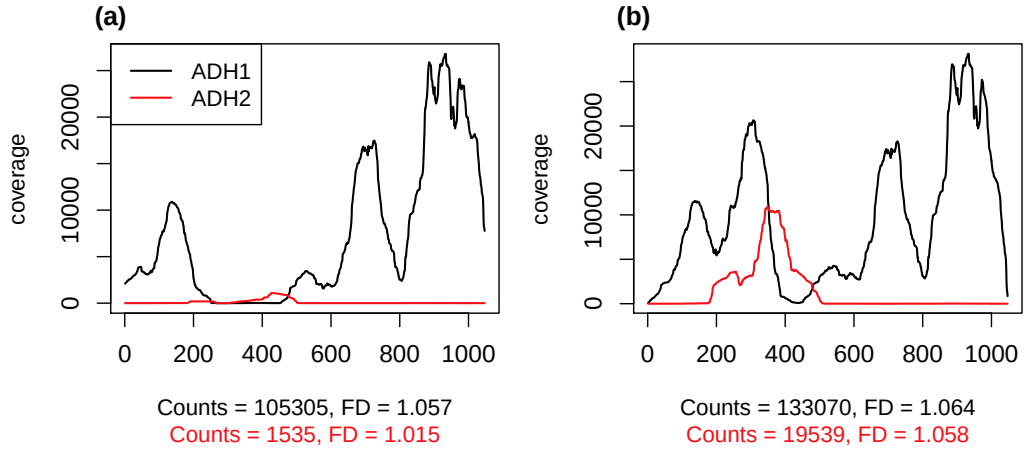

Figure 3.15: Coverage patterns of ADH1 and ADH2 when mapping (a) to the genome or, respectively, (b) to the transcriptome. Please note that the presumably zero coverage stretches of ADH2 are an effect of the scaling, typically the coverage ranges between 0 and 5 in these regions.

3.4 Discussion

Even though RNA-Seq is the state-of-the-art technology for large scale expression profiling the amount of produced data is while already anticipated still overwhelming. Automated pipelines for data-analysis are implemented by large sequencing centers and bioinformatics service facilities in order to speed up the analysis workflow. Yet, we believe that an in-depth analysis of RNA-Seq data still requires a substantial hands-on part. As a standard analysis will not generate new knowledge it is crucial to establish a deeper understanding of the data and, in particular, how certain decisions within the analysis influence the outcome of the experiment.

We developed a Bioconductor package, *FractalQC*, dedicated to fill this gap between large scale data and detailed analysis. The amount of information that is missed when considering read counts being only a summary statistic of the aligned reads is immense. Anscombe's quartet (Anscombe, 1973) nicely pinpoints this problem. While all four data sets have the same summary statistics such as mean or variance the underlying data structure is completely different as shown in Fig-

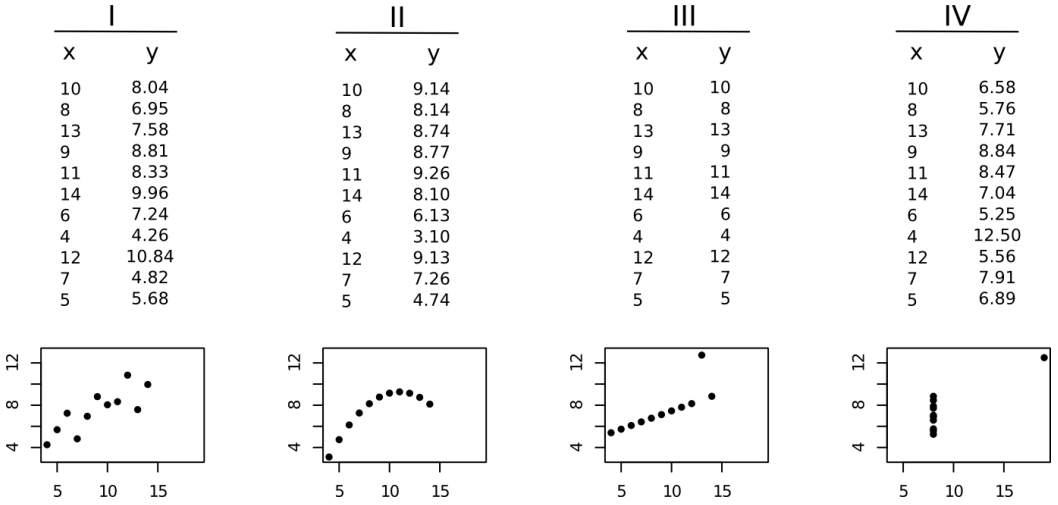

Figure 3.16: Anscombe's quartet. The mean is equal to 7.5, the variance equals 4.1 for all four data sets. These summary statistics do not fully capture the properties of the individual data sets as demonstrated in the different graphs.

ure 3.16. The analogy to RNA-Seq data can be immediately drawn. Even if reads counts are exactly equal, the underlying coverage pattern may be completely different and thus even lead to a re-interpretation of the respective read counts. As a visual assessment of thousands of coverage patterns is not feasible, *FractalQC* offers a way to evaluate the shape of these coverage patterns by means of the Fractal Dimension.

Given a perfect world scenario we have a theoretical expectation of uniform coverage. Any deviation from this uniformity indicates the presence of some bias. Biases may stem from either the library preparation protocol or from the analysis thereafter. Developing such a protocol is a non-trivial task and thus typically one contents himself with currently available library preparation protocols and its inherent biases such as random hexamer bias or fragmentation method bias. Albeit, the data analysis part provides plenty of easily realized options to circumvent biases or at least minimize its effects.

FractalQC pinpoints those coverage patterns calling for attention due to their too discrete shape. By comparing *FractalQC* reports from several analysis strategies such as different references while the mapping step or different summarization

methods the reasons for suspicious patterns may be uncovered. While *FractalQC* lists interesting candidates it can not take over the subsequent 'detective' work. Since the discrete resolution is one of the major advantages of RNA-Seq over Microarrays we are convinced that it is worth to invest some time to decipher these coverage patterns. Genes with a very pronounced signal, the so-called key players, can already be identified by microarrys or even by eye. In contrast, RNA-Seq offers the detection of finer nuances of gene expression and *FractalQC* supports this approach.

Besides quality control of individual samples also multiple samples can be easily compared. The FD always ranges between 1 and 2 for coverage profiles and is therefore directly comparable between experiments. This property makes it also particular attractive for benchmarking experiments such as in the study from Levin et al. (2010) where several strand-specific protocols are compared in terms of the resulting coverage.

FractalQC is offered as a package for the Bioconductor Suite (Gentleman et al., 2004) which is a popular collection of functionality for genomic analyses. The complete analysis workflow of RNA-Seq data can be performed within Bioconductor. In particular, the two most prominent models for inferring differential expression, edgeR (Robinson et al., 2009) and DEseq (Anders and Huber, 2010), are also part of the Bioconductor Suite. Since *FractalQC* is placed between mapping and differential expression inference in the analysis workflow we feel that *FractalQC* fits well within this environment.

Plans for further development include the use of the FD as weights in the model when testing for differential expression. By incorporating the reliability of the coverage pattern into the differential expression inference we expect to decrease the Type 1 error when testing the null-hypothesis of no differential expression. Additionally, we plan to extend the standard output of *FractalQC* in order to allow for more flexible post-processing.

Overall, we are convinced that *FractalQC* fills an important gap in current data analysis of RNA-Seq data. By making the information contained in the coverage patterns more accessible and by offering an intuitive metric for evaluation of those, we believe that *FractalQC* makes the entire analysis much more transparent and thus creates new insights how a optimal analysis workflow might look like.

Chapter 4

Exploring the Sampling Universe of RNA-Seq

4.1 Introduction

RNA-sequencing presents itself meanwhile as a well established technology and depicts a standard choice for investigation of the transcriptomic landscape. In principle the entire analysis workflow is already known and readily available (Garber et al., 2011) while as usual there is no agreement upon the 'analysis gold standard' (Oshlack et al., 2010, Ozsolak and Milos, 2011). A typical analysis pipeline comprises at least the following steps: mapping of the reads, summarization of the reads per adopted gene model, normalization and, eventually, testing for differential expression. Yet, the effective sampling universe, meaning the precise number and structure of genes present in the specific sample, remains unknown. Multiple methods exist to infer the most probable set of expressed genes or isoforms - either already known or newly discovered - given the alignment of the reads (Richard et al., 2010, Trapnell et al., 2010).

Here we are not interested in transcriptome reconstruction in terms of identifying all present known or novel splice forms of a given gene. Instead we want to know how much of the transcriptomic landscape is captured by the respective sequencing experiment. Tarazona et al. (2011) asked a similar question: How many reads are necessary to detach the power of detecting differential expression from the achieved sequencing depth. Blencowe et al. (2009) investigated how many reads would be

necessary to catch approximately 95% of all expressed transcripts in a human cell line (700 million reads). Both papers target the fundamental issue of defining the necessary as well as sufficient sequencing depth for the accurate detection of all expressed genes.
Back in 1988 Lander and Waterman already investigated this topic focusing on genomic mapping by fingerprinting data. They presented several still widely used formulas for the expected number of contigs given a certain number of fragments.

Making statements about the required sequencing depth requires a thorough modeling of the RNA-Seq process itself. We suggest the use of sampling formulas originating in the field of population genetics and which can be easily understood within the concept of species sampling. Given a random sample of animals out of a (potentially unknown) number of distinct species we obtain a frequency table which indicates how often each species is drawn. The well-known Ewens Sampling Formula (Ewens, 1972) gives the probability of observing a certain frequency pattern. In the sequencing context animals can be regarded as reads while species can either be interpreted as genes or positions of an individual gene. In the former case the sampling process of interest is the number of reads per gene model while in the latter case the sampling process takes place on the gene level - number of reads starting at each position of a given gene.

We will characterize these sampling processes by the Pitman Sampling Formula (Pitman, 1995), a generalization of the well-known Ewens Sampling Formula. By doing so we have two parameters at hand which capture the sampling process thoroughly and therefore allow us to realistically simulate RNA-Seq data. Moreover, the parameters of the PSF can be readily used in the context of benchmarking sequencing protocols (e.g., comparison of fragmentation methods, evenness of coverage).
As the set of expressed genes typically depends on organism, tissue type, cell type and developmental status the transcriptomic landscape varies from sample to sample. We contribute a method to explore the boundaries of the respective gene universe by providing an estimate for the size of the underlying gene universe. Based on these findings we evaluate an estimator for the number of newly detected genes

in an additional sequencing sample.
One main advantage of our method is that it merely relies on the observed read counts. While the central dogma of RNA-Seq is that the number of reads per gene reflect the true gene abundances, multiple factors may distort this relationship. Genes come along with different lengths and expression levels. Additionally physical properties like sequence decomposition or CG content attracts different kind of biases like PCR amplification or fragmentation bias. However, in this work we do not want to disentangle the individual effects of each bias on the sampling (Griebel et al., 2012) but exploit the observed frequency pattern as it comes.

4.2 Methodology

Note that from now on the term gene is used as synonym for any kind of expression entity like transcript, coding region or exon.

4.2.1 Motivation

The typical use-case for RNA-Seq is differential expression inference. The RNA-fraction of interest (e.g., mRNA or microRNAs) is extracted from multiple biological samples which differ in some kind of treatment or condition. Subsequent steps of a standard sequencing protocol include fragmentation, reverse transcription into double-stranded cDNA, amplification and, eventually, sequencing (Shendure and Ji, 2008). The resulting sequencing reads are mapped to a reference and summarized per gene (Garber et al., 2011). By doing so, a count table is obtained which contains the number of reads mapped to each gene for each condition. To account for different library sizes and gene length a normalization step is necessary (Bullard et al., 2010). Eventually the null hypothesis of no differential expression is tested gene-wise. The major challenges within differential expression inference include finding dispersion estimates as unbiased as possible and choosing an appropriate test-statistic given the commonly used small number of replicates (Smyth and Robinson, 2008).

Here, we adopt a more global perspective. We investigate the overall observed count frequency spectrum which is either given by the number of reads per gene or, one level beneath, by the number of reads starting at each position of a gene (not to be mistaken with the per base coverage). By doing so we aim to answer the following questions:

i) Can we realistically simulate the distribution of reads within or between genes, respectively?
ii) Are reads uniformly distributed within genes and how can any deviation from this uniformity be evaluated?
iii) Assuming that we capture the sampling process, can we extrapolate to additional future sequencing experiments? How many more genes will be detected if more reads are sequenced?

4.2.2 Hoppe Urn

The allocation of reads to either (i) positions of a gene or (ii) genes in general can be intuitively described by an urn model (Hoppe, 1984). Consider an urn which contains in the beginning just one black ball, the mutator, with weight θ. The urn rules are as follows (Zabell, 1992): If the color of the drawn ball is

black	then the black ball is put back into the urn together with two additional balls: one black ball with weight σ and one ball with a new color with weight $1-\sigma$.
colored	then the colored ball is put back into the urn and exactly one ball of the same color with unit weight is added.

Thus the probability of observing a new color (drawing the black ball) equals to $\frac{\theta+k\sigma}{\theta+n}$. k specifies the number of observed distinct colors when n balls have already been drawn. The probability of re-observing the j-th color equals to $\frac{n_j-\sigma}{\theta+n}$ whereas n_j is the number of balls with color j.

Hoppe Urn in the RNA-Seq context

Each ball can be regarded as a sequencing read. The color of the ball codes the origin of the read, either a specific gene or the starting position of the read within a given gene. The black balls are not counted, they are merely a means to generate new instances. An urn with n colored balls corresponds to a sequencing experiment yielding n reads.

So far nothing has been said about the size of the urn. Of course, in theory, one could imagine an infinite sampling universe while in reality the number of sequencing reads is restricted by the underlying sequencing technology. Depending if the sampling takes place from a finite or infinite sampling universe, two cases for the domains of θ and σ can be distinguished (Pitman, 2006).

Infinite number of colors $0 \leq \sigma < 1$, $\theta > -\sigma$

Finite number of colors $\sigma < 0$, $\theta = m|\sigma|$ for some $m \in \mathbb{N}$. In this case the probability of observing a new color $\frac{\theta+k\sigma}{\theta+n}$ tends to zero for $k \to m$. Hence m can be understood as the maximum number of colors that is available in the sampling pool.

In the RNA-Seq context, m can be interpreted as (i) the length of the gene as this is the maximum number of positions that is available as starting point for reads mapping to this gene or (ii) the total number of genes that could be theoretically expressed in the sample (e.g., all known protein coding genes, all known microRNAs).
The urn parameters θ and σ allow for a sensible interpretation:

θ is responsible for creating new instances: reads that belong either to a previously unobserved gene or to a position of a gene where so far no read has started. The higher θ the more genes or starting positions within a gene will be observed. Therefore we call θ the *innovation rate*.

σ is slightly more subtle to interpret. If σ is very large in comparison to n_j - the number of balls with color j - then n_j becomes negligible. In this case all observed colors are about equally likely to be drawn again. For this reason we call σ the *equilibrium rate*.

We recall that $\sigma < 0$ for an finite sampling universe. In this case σ provides additional information which can be illustrated as follows. Let us consider the observed read counts per gene. Here we are actually sampling from a multinomial distribution with unknown probability vector. The probability vector represents the abundance of each gene in respect to the sample that has been sequenced and naturally sums up to 1. If one gene would account for 50% of the reads then the associated probability of this gene would be 0.5. The probability vector is distributed as sampled from a symmetric Dirichlet distribution with m parameters equal to $|\sigma|$ (Pitman, 2006).

The symmetric Dirichlet distribution offers an intuitive interpretation of σ and is illustrated in Figure 4.1. In this toy example reads are sampled from a sampling universe of size 10. Hence, m, the maximum number of genes that can be drawn equals 10. If $|\sigma|$ is much smaller than one few genes capture most of the read counts. In the specific example shown in Figure 4.1 (left panel) only 7 genes are detected ($k = 7$). Gene number $4, 7$ and 8 are not represented by any read while being theoretically available in the sampling universe. The greater $|\sigma|$ the more balanced are the counts among the genes.

If the frequency vector of interest is the number of reads starting at each position of a given gene, σ can be used to assess to which extent the assumption of uniform coverage holds. In this case Figure 4.1 can be understood as the contiguous positions of a given gene. The height of the bars indicates how many reads start at each position of the gene.

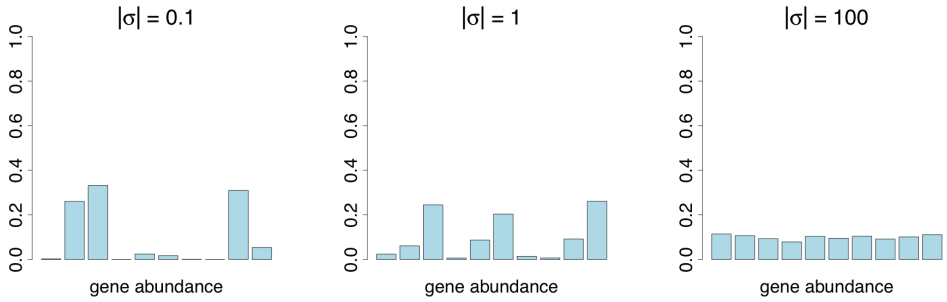

Figure 4.1: Random samples from a symmetric m-dimensional Dirichlet distribution with different values for the equilibrium rate σ. Each bar represents the abundance of a gene, 10 genes are available in the sampling pool.

4.2.3 Pitman Sampling Formula (PSF)

The Hoppe Urn is in fact just the urn representation of the PSF. Consider a random exchangeable partition which is defined as follows:
A partition of a positive integer n is an unordered collection of positive numbers with sum n (Pitman, 1995). The partition can be coded either by its

frequency vector n_j: whereas n_j is the number of balls with color j, $\sum n_j = n$

or

occupancy vector a_j: whereas a_j is the number of colors with j drawn balls, $\sum j a_j = n$.

The partition is exchangeable in the sense that its probability does not depend on the colors itself but just on the induced frequency or occupancy vector.

The PSF (Pitman, 1995) gives the probability distribution for observing a certain partition of n determined by its occupancy vector $\mathbf{A_n}$ and the number of distinctly observed colors K_n.

$$Pr[K_n = k, \mathbf{A}_n = \mathbf{a}_n] = n! \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{n} \frac{(1-\sigma)_{j-1}^{a_j}}{j!^{a_j} a_j!} \tag{4.1}$$

where $\mathbf{a}_n = (a_1, \ldots, a_n)$ and $(x)_l = x(x+1)\ldots(x+l-1)$ is the ascending factorial with $(x)_0 \equiv 1$.
Estimators for θ and σ can be obtained by empirical Maximum Likelihood of Equation 1.

PSF in the RNA-Seq context

Here the frequency vector is either given by the number of reads starting at each position of a gene or, by the number of reads per gene. Consequently the occupancy vector is the number of positions where j reads start or the number of genes with j counts ($j = 1, 2, ...$). The higher θ, the innovation rate, the more positions of a gene will be used as starting points or the more genes will be covered by at least one read.
As we assume a random, exchangeable partition we loose the identifiability of the reads. This basically means that we cannot name which position\gene is how abundant. We just know the overall count frequency spectrum.

Number of Expected Genes in An Additional Sample

The PSF can be obtained by two different approaches. If a random exchangeable partition is constructed according to the Hoppe Urn rules, the PSF gives the probability of observing a certain partition (Pitman, 1995). Alternatively we may adopt a Bayesian point of view. The sequencing reads form an exchangeable sequence $(X_t)_{t\geq 1}$ that can be modeled by a hierarchical model according to de Finetti's representation theorem (Lijoi et al., 2008). Here the X_t's state a random sample from an unknown discrete probability distribution $\tilde{P}$. The Poisson-Dirichlet process $PD(\sigma, \theta)$ is adopted as a suitable prior distribution.

$$\begin{aligned} X_i|\tilde{P} &\overset{iid}{\sim} \tilde{P} \\ \tilde{P}|(\theta,\sigma) &\sim PD(\sigma,\theta) \end{aligned} \tag{4.2}$$

By doing so the induced probability distribution by Equation 2 for a certain partition equals the PSF.

Let us consider a pilot sequencing experiment with n reads and K_n detected genes. Favaro et al. (2009) derived an estimator for $E[K_s^{(n)}|K_n = k]$, the expected number of newly detected genes in an additional sequencing sample of size s.
The parameters θ and σ are estimated by maximizing the PSF based on the observed data (empirical Maximum Likelihood). Then the expected number of additionally discovered genes when sequencing an additional sample of s reads equals to

$$E[K_s^{(n)}|K_n = k] = (k + \frac{\theta}{\sigma}) \left\{ \frac{(\theta + n + \sigma)_s}{(\theta + n)_s} - 1 \right\} \tag{4.3}$$

We have to note that the above-mentioned formula only holds for the infinite sampling universe ($0 \leq \sigma < 1$), in the finite sampling universe it can only be understood as approximation.

4.2.4 Availability

R-Code (R-Project, 2013) for all above-mentioned methods is available at `https://github.com/StefanieTauber/PSF_Hoppe` and is partly based on code from Durden and Dong (2009).

4.3 Applications

4.3.1 Distribution of Reads within Genes

It is well known that reads are not uniformly distributed over the length of a gene due to a mixture of positional and sequence specific biases (Wang et al., 2009, Roberts et al., 2011, Hansen et al., 2010). The allocation of reads is either simulated as uniformly distributed (Huang, 2012, McElroy et al., 2012) or more precisely modeled within a complex model framework (Griebel et al., 2012, Li et al.,

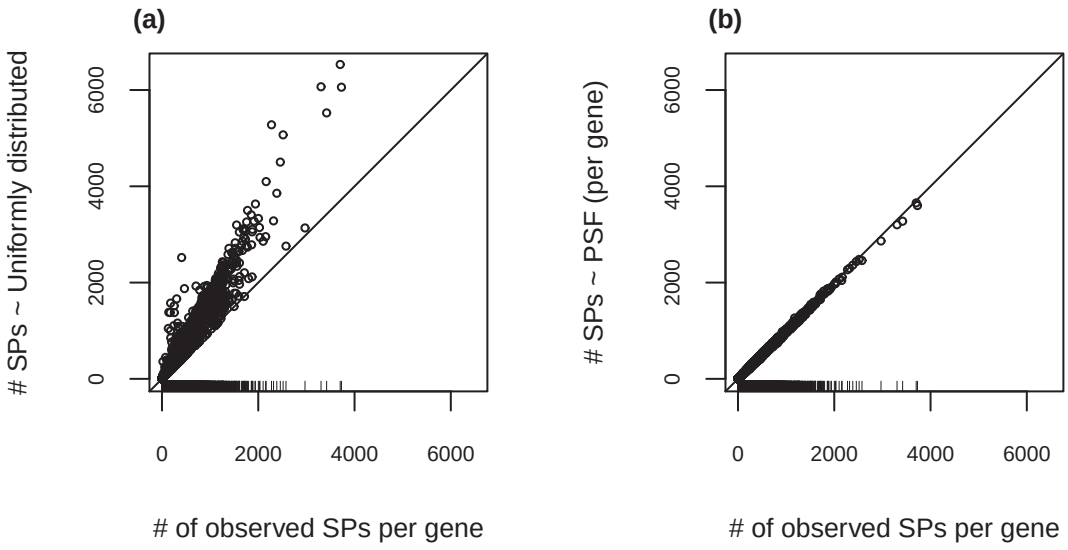

Figure 4.2: Relationship between number of starting points of real and simulated data assuming a uniform distribution (a) or the PSF (b). Each circle represents a gene. The rug plot on the x-axis indicates the density of the data.

2010). For instance, Griebel et al. (2012) apply models motivated by the biochemical properties of the respective fragmentation method (e.g. position weight matrices are used to model the sequence selectivity for the enzymatic digestion method).

The PSF allows us to capture this sampling process accurately by means of just two parameters, θ and σ, while avoiding to take into account each bias individually. Once estimates of the parameters have been obtained via Maximum Likelihood a Hoppe Urn can be started to realistically simulate the occupancy pattern of a gene. As we are sampling from a finite universe and therefore $\theta = m|\sigma|$ holds, just one parameter, e.g. θ, needs to be estimated. m is here the respective gene length. The Hoppe Urn returns, given n reads to place, (i) the number of distinct SPs and (ii) the occupancy of each position by SPs.

We demonstrate the use of the Hoppe Urn on yeast data from Levin et al. (2010) (Accession number: SRR059162). Reads were mapped with BWA 0.6.1. (Li and Durbin, 2009) against the genome (Ensembl, Release 66). Only uniquely mapping

reads were retained. For each gene we count the number of positions used as SPs. Subsequently we simulate for each gene the number of SPs if reads were uniformly distributed along the gene, and, respectively, when distributed according to the PSF. Figure 4.2 shows that the higher the read counts the more the uniformly distributed data deviates from the observed data, whereas the Hoppe Urn performs well throughout the entire range of read counts. Even when θ is estimated over all genes the deviation from the observed values remains moderate (data not shown). The worse performance of the uniformly distributed data is due to the inherent characteristics of different sequencing protocols which impede real uniform coverage. These biases regardless how they are exactly composed, are captured by the Hoppe Urn. Different mapping programs result in similar parameter estimates (data not shown). However, different sequencing protocols and technologies will lead to diverse estimates for the innovation rate.

The merit of this simulation is two-fold: First we now have a method available, the Hoppe Urn together with the PSF, which can be integrated into current RNA-Seq simulation pipelines to improve the quality of the simulation. Once typical parameter estimates are known for the experimental design of interest (organism, sequencing technology, mapping program) the number of distinct SPs and their occupancy can be realistically simulated by the Hoppe Urn. However, due to the exchangeability property of the PSF we do not know how these SPs are distributed along the gene. Secondly, and even more important, it is shown that the PSF accurately captures the sequencing process without bringing along a multitude of assumptions and parameters to estimate.

4.3.2 Comparison of Fragmentation Methods

Knierim et al. (2011) have recently evaluated three fragmentation methods (nebulization, sonication and random enzymatic digestion) for sequencing data in terms of coverage, resulting fragment length and sequence quality. Both nebulization and sonication state mechanical fragmentation methods, in the first case the DNA is broken by the force of compressed air or nitrogen whereas sonication relies on the

vibrations of ultrasonic waves. They used the 454 platform to do DNA sequencing of the LPRR4 gene which comprises roughly 40kb on chromosome 1. 8 overlapping fragments of sizes ranging from 1 to 9 kb are equimolarly pooled and sequenced, 3 technical replicates for each fragmentation method. The overall conclusion was that no method was superior over the others regarding the up-mentioned criteria.

We are now revisiting this data to have a closer look at the fragmentation behaviour of each method. Bam files were obtained directly from the authors. Again, as we are sampling from a finite universe, just one parameter remains to be estimated. m, being the size of the sampling universe, is in this case the length of the individual fragments. Longer fragments will naturally have higher read counts and therefore also a higher innovation rate. However as all three fragmentation methods have about the same sequencing yield θ remains comparable between protocols. When contrasting the read counts for each fragment and replicate with the corresponding estimated θ we clearly see that the random enzymatic digestion method has the overall highest innovation rate (Figure 4.3). Although nebulization and sonication constitute prominent mechanical fragmentation methods the random enzymatic digestion method does deliver more random breaking points. It is indeed known that nebulization and sonication might perform different on the body of the gene in contrast to the 5' or 3' end (Wang et al., 2009).

The PSF provides additional information: the equilibrium rate, $|\sigma|$, is between 0 and 1 for all samples but for the random enzymatic digestion method most closely to 1 (median of $|\sigma|$ over all fragments and replicates: 0.52 (enzymatic digestion), 0.4 (sonication) and 0.38 (nebulization)). This can be interpreted the following way: All three methods are still far from using all available positions. However, in comparison, the random enzymatic digestion method shows the most balanced behaviour (compare to Figure 4.1, center panel).

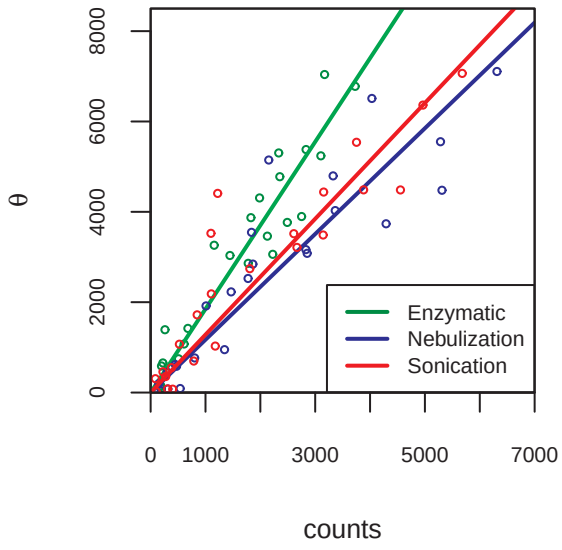

Figure 4.3: Linear regression for estimates of the innovation rate for three fragmentation methods. Each circle represents one fragment (8 fragments with 3 technical replicates each).

4.3.3 Comparison of Strand Specific Protocols

Levin et al. (2010) assessed several strand specific protocols on yeast data and compared them in terms of six criteria: library complexity, strand specificity, evenness and continuity of coverage at annotated transcripts, performance at 5' and 3' ends, and eventually, performance in expression profiling. In order to avoid library size effects the same amount of reads was sampled for each protocol and up-mentioned criteria were only evaluated on these subsets.
Here, we want to reconsider the third criterion, evenness and continuity of coverage from the PSF point of view.

Levin et al. (2010) calculated the average coefficient of variation of gene coverage for the top 50% expressed genes to assess evenness of coverage. The continuity of coverage was checked in two ways: First, they counted the number of stretches for each transcript that are covered by reads whereas a stretch is discontinued when at least 5 contiguous bases have a coverage of 0. Subsequently the average of these

Protocol	Accession Number	Protocol	Accession Number
RNA Ligation	SRR059162	3' split adaptor	SRR014386
Illumina RNA Ligation	SRR059163 SRR059164	SMART	SRR059167 SRR059168
Hybrid	SRR059169 SRR059170	published dUTP	SRR014850
		NNSR	SRR059171
dUTP	SRR059176	Control	SRR059178

Table 4.1: Accession numbers for the used libraries from Levin et al. (2010)

numbers over all transcripts was computed and weighted by the respective expression of the transcript. Secondly, the previously sampled subsets of all libraries are pooled, the number of bases per transcript with a coverage of 0 was calculated for each protocol and contrasted to the fraction of total reads. Two protocols, namely the 3'split adaptor and the published dUTP protocol perform best within this criterion of evenness and continuity of coverage whereas the latter one, the dUTP protocol, even turned out as the protocol performing consistently well throughout all criteria.

We note that several ad-hoc thresholds were introduced in order to evaluate the quality of the coverage. These thresholds range from the sampling from the reads to avoid library size effects, the decision to just use the top 50% expressed genes to check for the evenness of the coverage to the definition of a break in the coverage. These experiment specific thresholds make it hard to compare results between studies as each study comes along with its own set of suitable ad-hoc thresholds. Therefore we suggest to use the PSF which enables us to assess the data as it comes along.

To this end, the innovation rate is estimated for each gene and for each library. Accession numbers are as specified in Table 4.1, reads were mapped with BWA 0.6.1. (Li and Durbin, 2009) against the genome (Ensembl, Release 66). Only uniquely mapping reads were retained.

Here we are especially interested how evenly positions of a gene are occupied by reads. As we are sampling from a finite universe (each gene has a finite length) the

equilibrium rate $|\sigma|$ provides exactly this information. All data is used, no thresholds on expression levels are imposed. The range of $|\sigma|$ for different protocols is depicted in Figure 4.4. Protocols are ordered increasingly according to the number of reads captured in gene models. If the equilibrium rate is approximately 1 all bases of a gene show a balanced occupation behaviour. The greater $|\sigma|$ the more saturated is the occupation. Our analysis shows that the published dUTP and the 3'split adaptor protocol succeed. Importantly, no correlation between the number of counts and the equilibrium rate can be noted. Thus, we do not need any subsampling in order to asses the evenness of coverage by the PSF.

All together, the PSF allows us to assess the quality of the coverage in a very intuitive way. No subsetting or thresholding is required, comparison between different studies is possible.

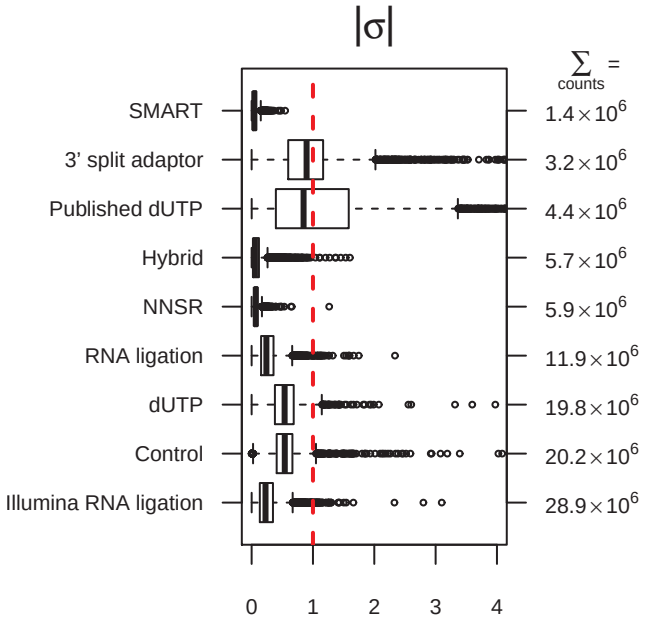

Figure 4.4: Estimation of the equilibrium rate σ for different sequencing protocols. The sum of counts specifies the number of mapped reads. The red line marks the case $|\sigma| = 1$ where the occupation probabilities for all positions are uniformly distributed. Compare to Figure 4.1 for better understanding of the equilibrium rate.

4.3.4 Size of Gene Universe

Considering a sampling process on the gene level, two questions are of special interest: (i) How many more genes will be detected if more reads are sequenced? (ii) What is the extent of the transcriptomic landscape present, yet maybe undetected, in the respective sequencing sample?

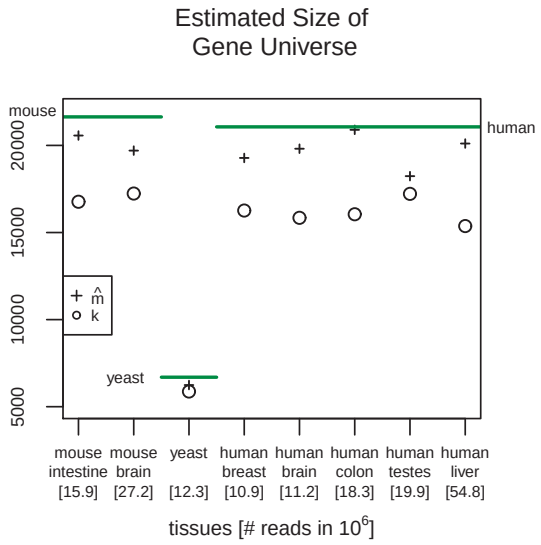

Figure 4.5: Estimated number of transcribed genes in different biological samples. The solid lines specify the number of known protein coding genes for mouse, yeast and human. In contrast, $\hat{m}$ is the estimated size of the gene universe $(+)$. k $(\circ)$ specifies the number of detected genes in the underlying sample.

The PSF is parameterised by the innovation rate θ and the equilibrium rate σ which gives information how balanced reads are distributed among genes. As we are sampling from a finite sampling universe, $\theta = m|\sigma|$ holds where m is the unknown number of genes present in the sequencing sample. The number of protein coding genes is a conservative upper bound for m (assuming a well-known organism whose transcriptome has been sufficiently described). Alternatively we can estimate m from the data by parameterising the PSF not by θ and σ but instead by θ and m. The advantage of not estimating θ and σ but instead θ and m is given by the fact that in the latter case we do not need to plug in an estimate for m.

We assess our estimate for the extent of the gene universe, $\hat{m}$, for several publicly

available data sets comprising different organisms and tissues (Table 4.2, data pre-processing as in section 3.3).
Figure 4.5 shows the corresponding estimates together with k - the number of genes detected in the sequencing sample. We observe that - as expected - different tissues come along with different $\hat{m}$. We note that more reads do not automatically lead to a higher number of detected genes and do also not blow up $\hat{m}$.

There are now two estimates available for m, the number of transcribed genes in the sequencing sample: (i) the number of protein coding genes and (ii) $\hat{m}$ estimated from the data using the re-parametrised PSF. These estimates can now be compared and evaluated in correspondence with the expected number of newly detected genes in an additional sequencing sample. Favaro et al. (2009) have derived this closed estimate (Eq. 3) under the two-parameter Poisson-Dirichlet model. They focus on an infinite sampling universe and demonstrate the usability of their findings on EST data. Here, we will use Equation 3 to assess the performance of $\hat{m}$ (Figure 4.6a).
The human brain data set is taken for illustration. Random samples are drawn repetitively from the original set of reads. Sample sizes range from 5 to 95% of the original sample. These subsamples mimic pilot sequencing experiments whereas the remaining rest of the reads is consequently a second, follow-up sequencing experiment. The innovation rate is estimated for each pilot experiment and the equilibrium rate is computed from $\theta = m|\sigma|$.
Equation 3 delivers $\hat{K}^{new}_{cons}$ - the number of expected newly detected genes when sampling the remaining rest of the original sample. Note that $\hat{K}^{new}_{cons}$ uses the number of protein coding genes as approximation for m (therefore 'cons' for conservative) whereas $\hat{K}^{new}_{\hat{m}}$ is derived by using $\hat{m}$. These two estimates are then compared to K_{obs} - the observed number of newly detected genes. We see that $\hat{K}^{new}_{\hat{m}}$ is always closer to K_{obs} than $\hat{K}^{new}_{cons}$. Interestingly, both $\hat{K}^{new}_{cons}$ and $\hat{K}^{new}_{\hat{m}}$ overestimate the true value. This is due to the fact that in the finite sampling universe case Equation 3 is mainly driven by m. The more m is overestimated the more $\hat{K}^{new}$ is overestimated. Unfortunately the count distribution is very often rather sparse meaning that few genes collect the majority of the reads. This hinders an optimal estimation of m.

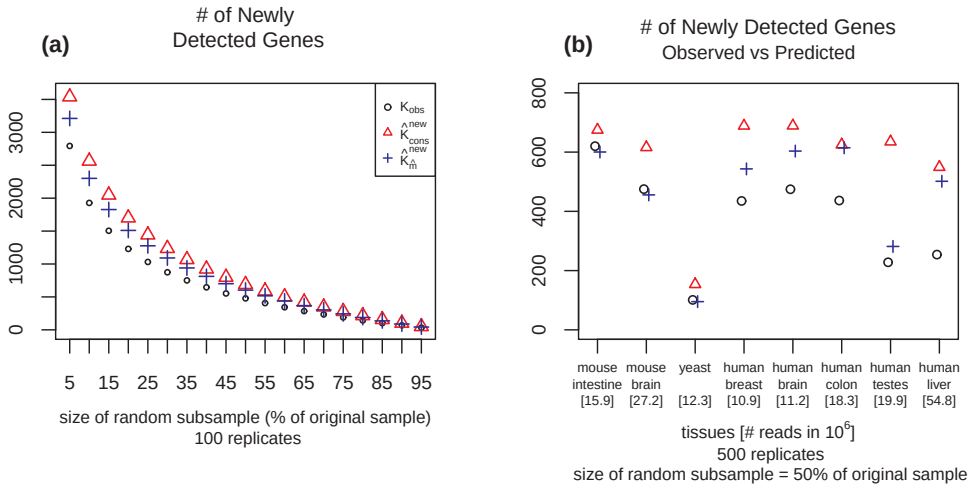

Figure 4.6: Predicted number of newly detected genes in a follow up experiment (a) as a function of the proportion of reads obtained in the pilot experiment compared to the number of reads in pilot and follow up experiment. Plotting symbols depict the median of 100 random samples. (b) for different tissues, where the follow up experiment has the same sample size as the pilot. Plotting symbols depict the median of 500 random samples. K_{obs} is the observed number of newly detected genes when sampling the remaining rest of the sequencing sample. $\hat{K}^{new}_{cons}$ is the number of newly detected genes when using the number of known protein coding genes as the size of the sampling universe. $\hat{K}^{new}_{\hat{m}}$ uses $\hat{m}$ as size of the sampling universe.

Additionally, one has to keep in mind that each random subsample has its own k, number of detected genes, and its specific frequency vector. Depending on the underlying sample some genes might be over- or underestimated in their abundance. This, of course, will also influence the performance of the estimator. To summarize, even for very small sizes of the pilot experiment, the prediction of newly detected genes deviates only moderately from the observed values. This information should be enough to take a considerate decision if further sequencing is necessary and sensible.

Finally we present $\hat{K}^{new}_{cons}$ and $\hat{K}^{new}_{\hat{m}}$for a selection of datasets comprising different organisms and tissues (Figure 4.6b, Table 4.2). Again we divide the original set of reads into two parts in order to be have a 'true' value for the number of newly detected genes at hand. Subsequently θ is estimated for a random subsample com-

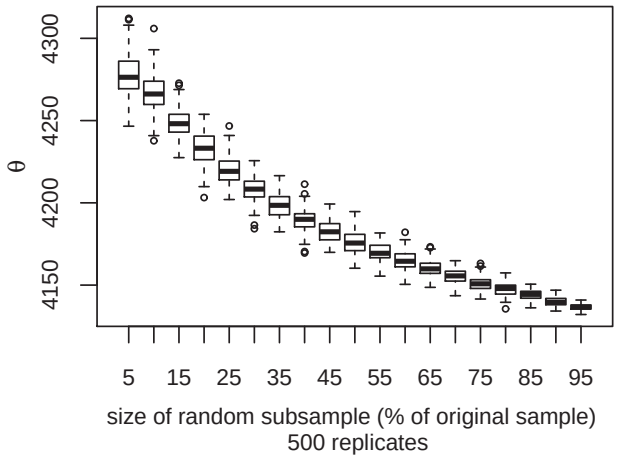

Figure 4.7: Adaption of the innovation rate θ to the sample size.

prising 50% of the original data, this procedure is conducted 500 times. In the following $\hat{K}^{new}_{cons}$ and $\hat{K}^{new}_{\hat{m}}$ is computed for each random subsample. We observe that $\hat{K}^{new}_{\hat{m}}$ is always closer to the true observed value than $\hat{K}^{new}_{cons}$. We further note that the improvement of $\hat{K}^{new}_{\hat{m}}$ over $\hat{K}^{new}_{cons}$ as well as the overall performance of both estimates varies from dataset to dataset. The respective underlying count distribution heavily influences the quality of the estimates. The more sparse the count distribution the more difficult it is to get reliable estimates for θ and m.

To facilitate the intuition for the interpretation of θ Figure 4.7 shows the range of the innovation rate for different sample sizes (same data as in Figure 4.6a). We observe that θ gets smaller the larger the size of the random subsample. This shall not be mistaken with the impression that θ gets closer to the "true" θ the larger the sample size. θ, being the innovation rate, must get smaller for an increased sample size because the probability to detect a new gene declines. Hence there is no fixed true innovation rate for all sample sizes, rather θ adapts to the respective sample size.

Paper	Organism	Tissue	Accession Number
Levin et al. (2010)	*Saccharomyces cerevisiae*	-	SRR059162
Mortazavi et al. (2008)	*Mus musculus*	brain	SRR006489 SRR001356 SRR001357
Shen et al. (2012)	*Mus musculus*	intestine	SRX113073
Wang et al. (2008)	*Homo sapiens*	breast brain colon testes	SRX003922 SRX003920 SRX003931 SRX003933
Human BodyMap Data (2011)	*Homo sapiens*	liver	ERR030895

Table 4.2: Accession numbers for the used sequencing libraries

4.4 Discussion

4.4.1 Characteristics of RNA-Seq Data

While RNA-Seq data basically constitutes count data and thus statistical methods to address such data are widely available and well established, the analysis of sequencing data remains challenging. RNA-Seq data is the result of a complex manufacturing process bringing along a strong need for data preprocessing. Numerous papers (Hansen et al., 2010, Robinson and Oshlack, 2010, Schwartz et al., 2011) are devoted to these issues and propose multifaceted methods for normalization for library size, GC bias and fragment length bias. Other papers concentrate on what kind of underlying distribution is most suitable for testing for differential expression (Robinson and Smyth, 2007, Wang et al., 2010, Anders and Huber, 2010). Here, in this work, we approach RNA-Seq data from a different angle. We do not aim to deliver another method for normalization or bias correction. Being aware that RNA-Seq data is the result of a conglomerate of experimental conditions we suggest the following working-attitude: given the data as it comes along (i) how can

we characterize the data (ii) what can we tell about future experiments under the same experimental conditions.

4.4.2 Characterization of RNA-Seq Data

We propagate the use of sampling formulas which are well known in the field of population genetics, namely the Pitman Sampling Formula which states a generalization of the Ewens Sampling Formula. The PSF operates on the most basis properties of RNA-Seq data: how many reads have been sequenced, how many genes have been detected and how even reads are distributed among genes. On the basis of these features the PSF provides two parameters that characterize the sampling process. θ can be regarded as the innovation rate whereas σ indicates how balanced reads are distributed among genes. The sampling formula cannot only be applied on the transcriptome level which summarizes reads per gene but also on the gene level. In this case the number of reads starting at each position of a gene are counted. In both cases the PSF characterizes the occupancy pattern of genes, and respectively, positions.

4.4.3 Benefits of PSF

The Pitman Sampling Formula offers several valuable applications.

- Once typical parameter values are obtained a Hoppe Urn can be integrated in existing simulation pipelines. The Hoppe Urn realistically simulates occupancy pattern, either on the gene or on the transcriptome level. In the former case the Hoppe Urn replaces the unsatisfying uniform distribution of reads along a gene whereas in the latter case typical occupancy pattern of the respective gene universe can be reproduced.
- The parameters of the PSF together with their practical interpretation enable direct comparison and evaluation of fragmentation methods or sequencing protocols in general with respect to coverage behaviour.

- Most importantly, the PSF provides an estimate for the size of the underlying gene universe as well as an estimate for the expected number of newly detected genes when sequencing an additional sample.

We advertise the fact that we do not put any thresholds on the data. Yet, we admit that we cannot tell what exactly one count per gene effectively means. Being deprived of a general valid threshold value we prefer to process all data and allowing thus the user to apply is own criteria in the very end of the analysis.

The PSF operates on raw, absolute counts. Therefore we have focused so far on summarization on gene level. Obviously it would be very interesting to extend the application of the PSF to the level of splice variants. The main issue here lies in the fact that the absolute assignment of reads to one specific splice variant is by no means trivial. Yet, Roberts and Pachter (2012) recently introduced the software eXpress which outputs hard assignments of counts on isoform level based on likelihoods. Application of the PSF on the splice variant level is subject to current research.
To sum up, we report the application of well known sampling formulas to state-of-art sequencing data. To us it is especially intriguing that these sampling formula have been around for decades and yet naturally fit the needs and characteristics of sequencing data. However, we also note that further work is necessary to further support the suitability of the PSF. We have not yet addressed any goodness of fit statistics, that are required to recognize the limitations of our approach. Such topics will be addressed in future work.

Chapter 5

Summary and Outlook

Since RNA-Seq is a relatively new technology for global gene expression screens major challenges are still present within the analysis workflow. Here, we contribute methods targeting parts of the workflow which have been neglected so far and, respectively, suggest an alternative view on the sequencing data itself.

First, we point out that any RNA-Seq analysis concentrating on reads counts being only a summary statistics of the underlying coverage pattern results in a loss of information. We show that coverage patterns give evidence about possibly occurring pitfalls during library preparation or subsequent analyses. In order to identify suspicious coverage patterns we develop as score, S_{FDA}, integrating the Fractal Dimension as well as the area under the curve of a coverage pattern. Coverage patterns are ranked according to their S_{FDA} and a low score urges for closer inspection. The developed method is realized in *fractalQC*, an R package. In summary, by making the coverage information more accessible and by offering a method to evaluate these patterns, we achieve an increased awareness of the consequences of individual decisions and, respectively, even reconsideration of certain analysis strategies.

Besides the FD, other measures to explore the shape of coverage patterns are conceivable. The Hurst Exponent (HE), widely known in the context of financial markets, is one possibility. The HE is used to determine long range dependencies in time series analysis (Taqqu, 1995). Thus a graph may be either classified as persistent (an increase is prone to be followed by an increase and similarly for a decrease),

antipersistent (an increase is likely to be followed by a decrease and vice versa) or, no specific trend can be observed. Similarly to the FD, a variety of methods is available to compute the Hurst Exponent. The most popular and classical method is the R/S (rescaled adjusted range) statistic (Taqqu, 1995):

Let $X = \{X_i, i \geq 1\}$ be a time series with partial sum $Y(n) = \sum_{i=1}^{n} X_i$ and sample variance $S^2(n) := (1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. The *R/S* statistic is then given by:

$$\frac{R}{S}(n) := \frac{1}{S(n)}\left[\max_{0\leq t\leq n}\left(Y(t) - \frac{t}{n}Y(n)\right) - \min_{0\leq t\leq n}\left(Y(t) - \frac{t}{n}Y(n)\right)\right]$$

As the FD the HE follows a power-law: $\frac{R}{S}(n) \propto n^{HE}$ for $n \to \infty$ (Taqqu, 1995). The HE ranges between 0 and 1. A HE of about 0.5 indicates random behavior while a HE greater than 0.5 points towards a persistent trend. Finally, a HE smaller than 0.5 identifies an antipersistent pattern. See Figure 5.1 for two coverage patterns with similar FD but different HE.

Thus, depending on the value of the Hurst Exponent trends in the coverage pattern may be determined. Notably, the FD and the HE may be linked as follows: $FD + HE = d + 1$ where d is the topological dimension of the underlying object. This depicts an association which also makes intuitively sense: a space-filling curve goes hand in hand with an antipersistent trend whereas a smooth curve reflects a persistent trend. Yet, this relationship only holds for self-affine processes such as fractional Gaussian Noise or fractional Brownian Motion. Gneiting and Schlather (2004) show that FD and HE are in fact decoupled for a variety of stochastic models. Therefore, calculation of the HE in addition to the FD may provide additional knowledge.

Possible applications include benchmarks of different library preparation protocols in respect to the resulting coverage. Preliminary results show that while the HE works well as means to discover trends the FD is easier to link to subsequent meaningful interpretation. However, further studies are required to examine the benefit of the HE within the context of RNA-Seq analysis.

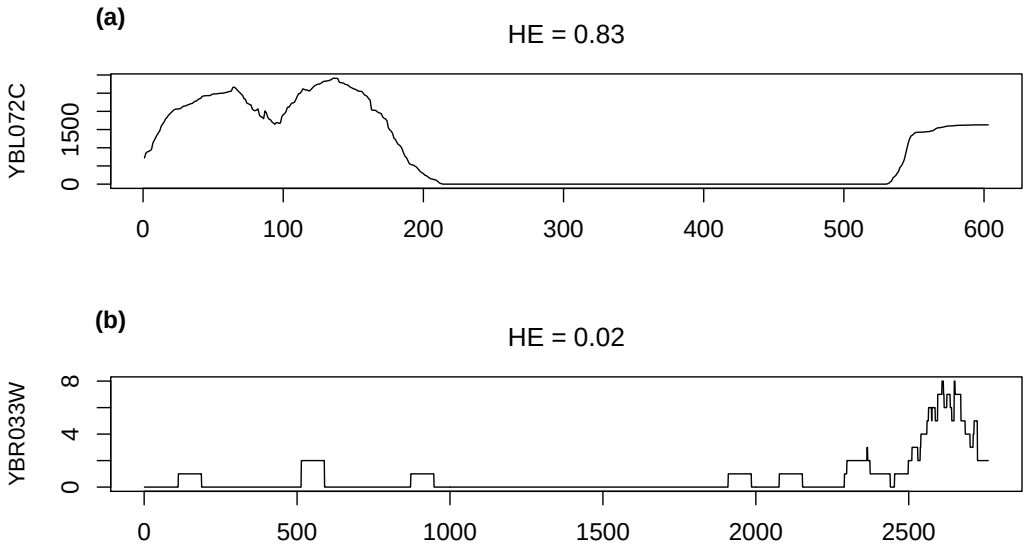

Figure 5.1: Two coverage patterns with different HE indicating a persistent trend in (a) and antipersistent in (b). In both cases the FD is about 1.03, the read counts equal to 7094 in (a) and to 26 in (b). Data is taken from Levin et al. (2010).

Additionally, motivated by Microarray analysis, we plan to use the FD as weights in the RNA-Seq analysis. The Microarray technology is based on hybridization of fluorescently labeled nucleotide probes to oligonucleotides fixed on a planar surface. Each probe on the Microarray serves as a representative for a gene. Gene expression abundance is then proportional to the fluorescent intensity when imaging the Microarray with a scanner (Schena, 1998). Morphological properties such as diameter, area or ratio of foreground to background intensities are recorded for each probe on the Microarrray (Smyth and Speed, 2003). An uneven area or a small foreground to background ratio indicates quality issues of the respective probe. This knowledge is exploited by turning these morphological properties into weights whereas unreliable morphological features are reflected in a small weight. In the following, these weights may be incorporated in normalization methods (Smyth and Speed, 2003) or even in the differential expression analysis to individually downweight unreliable genes (Smyth, 2004). Along these lines, we are convinced that incorporating the FD as weights in the RNA-Seq analysis is promising. In particular, the FD may be

exploited as a scaling factor in the normalization step. In order to ultimately verify this approach comparison to benchmark data such as quantitative real-time PCR data will be necessary.

Apart from exploiting the information contained in the coverage patterns we propose a fresh view upon sequencing data. Rather than adopting a gene-wise perspective we opt for a global point of view on sequencing data. We take the data as it comes along, genuine mRNA abundances distorted by a conglomerate of biases and model it. The advantage of this pragmatic approach is that we are able to make valid predictions for data suffering from the very same technological constraints. This is of particular interest since the majority of RNA-Seq simulation programs bases on unrealistic assumptions such as uniform distribution of reads within genes.

We differ between two sampling processes given either by the number of reads starting at each position of a gene or by the number of reads per gene. Both processes can be captured by the Pitman Sampling Formula.

We have demonstrated the use of the PSF as means to evaluate the evenness of coverage. Additionally, distribution of reads within genes may be realistically simulated by the Hoppe Urn. This is of particular use for benchmarking mapping programs.

While the PSF works for a large variety of examples, a still open question is whether one can determine a relationship between characteristic features of a gene and its innovation rate θ. We suspect that certain genes bring along a certain θ due to their properties e.g. GC content or overall nucleotide composition. If this were true we could use θ to categorize genes which would impact future analyses. This is subject to current research.

Additionally, since summarization on the isoform level remains a challenging task method development for differential expression inference acted accordingly resulting in a model on the exon level (Anders et al., 2012). Thus, it may be of particular

interest to extend the application of the PSF to the exon-level, so modeling the number of reads per exon.

Eventually, the PSF may be applicable in the course of the analysis of non-model organisms where the size of the transcriptome is yet not known.

Acknowledgments

First of all I would like to thank my parents who have managed - from the very beginning - to turn the 'must' into a 'may' when thinking of learning and acquiring knowledge. Thanks to my sisters for providing a valuable change of perspective from time to time by means of their ever growing families.

I would like to thank my supervisor, Arndt, for providing a great infrastructure and especially for the opportunities to attend numerous conferences. Additionally, I really appreciate the relaxing working atmosphere and most of all, the freedom to pursue my own interests and to work independently.

Of course, a large thanks to all the CIBIV members for many enjoyable chitchats and the occasional rounds of self-pity about our so seemingly hard time as PhD students. Thanks to all my friends at the Goethe University who definitely made the last year a highlight.

Thanks to Ingo for being there, at - and even more important - outside work. In particular, thanks for many encouraging conversations and for enduring the few moments of slight distress while writing this thesis.

Curriculum Vitae

Stefanie Tauber

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Dr. Bohr Gasse 9
A-1030 Vienna, Austria
Phone: ++43 +1 / 4277 24024
Fax: ++43 +1 / 4277 24098
Email: stefanie.tauber(AT)univie.ac.at, stefanie.tauber(AT)gmail.com
Homepage: www.cibiv.at/~stefanie

Date of birth April 2, 1984
Place of Birth Baden, Austria
Nationality Austria

Education

JUNE 2008	Diplom-Ingenieurin, **Technische Universität Wien**
	Major: *Mathematics in Science*
	Thesis: Quality Assessment, Normalisation and Mixed Effects Models for Microarray Data: Application to a Sweet Potato Study
JUNE 2002	Matura, **Bundesgymnasium Berndorf**

Research Experience

Current

PhD Student at the *Center for Integrative Bioinformatics Vienna (CIBIV)*, Austria (thesis submisson: july 2013)

Research keywords: analysis of microarray and NGS data, methods development for quality control of RNA-Seq data, modeling the sampling process of sequencing

July 2008-Dec 2009

Researcher at the *DNA-Microarray Facility - Ludwig Boltzmann Institute for clinical & experimental Oncology*, Vienna and at the *Skin and Endothelium Laboratory*, Medical University Vienna, Austria

Research keywords: statistical analysis of numerous microarray experiments, integretation of clinical and microarray data; deciphering melanoma heterogeneity, the molecular basis of melanoma metastasis to sentinel nodes,Wnt-signalling in melanoma-induced angiogenesis

March 2006-March 2007

Researcher at the *Austrian Research Centers Seibersdorf (ARC)*, Austria

Research keywords: evaluation of different statistical methods for the analysis of a sweet-potato microarray-experiment, statistical consulting with respect to microarray data, experimental design, analysis of cross species chips

May 2005-July 2005

Researcher at the *International Maize and Wheat Improvement Center (CIMMYT)*, Mexico

Research keywords: analysis of micorarray experiments, mixed linear models for two-color microarrays

March 2005-Sept 2005

Researcher at the *Austrian Research Centers Seibersdorf (ARC)*, Austria

Research keywords: analysis and statistical consulting with respect to microarray data

List of Publications (peer reviewed)

2013 **Exploring the Sampling Universe of RNA-Seq.**

S. Tauber and A. von Haeseler, **Stat Appl Genet Mol Biol**; 12(2), 175-188.

Focal segmental glomerulosclerosis is induced by microRNA-193a and its downregulation of WT1.

C.A. Gebeshuber, C. Kornauth, L. Dong, R. Sierig, J. Seibler, M. Reiss, S. Tauber, M. Bilban, S. Wang, R. Kain, G.A. Böhmig, M.J. Moeller, H.J. Gröne, C. Englert, J. Martinez* and D. Kerjaschki*, **Nat Med**; 19, 481-487.

*shared senior authorship

Lipoprotein Lipase in Chronic Lymphocytic Leukemia - Strong Biomarker with Lack of Functional Significance.

E. Porpaczy, S. Tauber, M. Bilban, G. Kostner, M. Gruber, S. Eder, D. Heintel, T. Le, K. Fleiss, C. Skrabs, M. Shehata, U. Jäger and K. Vanura, **Leukemia Res**; 37(6), 631-636.

2012 **Epidermal growth factor facilitates melanoma lymph node metastasis by influencing tumor lymphangiogenesis.**

A. Bracher*, A. Soler*, S. Tauber, A.M. Fink, A. Steiner, H. Pehamberger, H. Niederleithner, P. Petzelbauer, M. Groeger and R. Loewe, **J Invest Dermatol**; 133, 230-238.

*equal contribution

MET expression in melanoma correlates with a lymphangiogenic phenotype.

A. Swoboda, O. Schanab, S. Tauber, M. Bilban, W. Berger, P. Petzelbauer and M. Mikula, **Hum Mol Genet**; 21(15), 3387-3396.

Wnt1 is anti-lymphangiogenic in a melanoma mouse model.

H. Niederleithner, M. Heinz, S. Tauber, M. Bilban, H. Pehamberger, S. Sonderegger, M. Knöfler, A. Bracher, W. Berger, R. Loewe and P. Petzelbauer, **J Invest Dermatol**; 132, 2235-2244.

Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs.

C. Vesely*, S. Tauber*, F. Sedlazeck, A. von Haeseler and M.F. Jantsch, **Genome Res**; 22, 1468-1476.

*equal contribution

2010 **Trophoblast invasion: Assessment of cellular models using gene expression signatures.**

M. Bilban, S. Tauber, P. Haslinger, J. Pollheimer, L. Saleh, H. Pehamberger, O. Wagner and M. Knöfler, **Placenta**; 31(11), 989-996.

Transcriptome analysis of human cancer reveals a functional role of Heme Oxygenase-1 in tumor cell adhesion.

S. Tauber*, A. Jais*, M. Jeitler, S. Haider, J. Husa, J. Lindroos, M. Knöfler, M. Mayerhofer, H. Pehamberger, O. Wagner and M. Bilban, **Mol Cancer**; 9, 200.

*equal contribution

Signal Transducer and Activator of Transcription 3 Protects From Liver Injury and Fibrosis in a Mouse Model of Sclerosing Cholangitis.

M. Mair, G. Zollner, D. Schneller, M. Musteanu, P. Fickert, J. Gumhold, C. Schuster, A. Fuchsbichler, M. Bilban, S. Tauber, H. Esterbauer, L. Kenner, V. Poli, L. Blaas, J. W. Kornfeld, E. Casanova, W. Mikulits, M. Trauner and R. Eferl, **Gastroenterology**; 138(7), 2499-2508.

Antiinflammatory effects of tumor necrosis factor on hematopoietic cells in a murine model of erosive arthritis.

S. Blüml, N. B. Binder, B. Niederreiter, K. Polzer, S. Hayer, S. Tauber, G. Schett, C. Scheinecker, G. Kollias, E. Selzer, M. Bilban, J. S. Smolen, G. Superti-Furga, and K. Redlich, **Arthritis & Rheumatism**; 62(6), 1608-1619.

2009 **Stat3 Is a Negative Regulator of Intestinal Tumor Progression in ApcMin Mice.**

M. Musteanu, L. Blaas, M. Mair, M. Schlederer, M. Bilban, S. Tauber, H. Esterbauer, M. Mueller, E. Casanova, L. Kenner, V. Poli, and R. Eferl, **Gastroenterology**; 138(3), 1003-1011.

2008 **The nuclear CoRepressor, NCoR, regulates thyroid hormone action in vivo.**

I. Astapova, L.J. Lee, C. Morales, S. Tauber, M. Bilban, and A.N. Hollenberg, **Proc Natl Acad Sci USA**; 105(49), 19544-19549.

Transcriptomic changes in wind-exposed poplar leaves are dependent on developmental stage.

S. Fluch, C. C. Olmo, S. Tauber, M.Stierschneider, D. Kopecky, T. G. Reichenauer, and I. Matusíková, **Planta**; 28(5), 757-764.

Bibliography

Aird,D., Ross,M.G., Chen, W-S., Danielsson,M., Fennel,T., Russ,C. Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

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

Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Anscombe,F.J. (1973) Graphs in Statistical Analysis. *Am. Statistician*, 27(1), 17–21.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**(10), e72.

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

Blencowe,B.J., Ahmad,S. and Lee,L.J. (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.*, **23**, 1379–1386.

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

Crick, F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561-563.

DeLuca,D.S., Levin,J.Z., Sivachenko,A., Fennell,T., Nazaire,M., Williams,C., Reich,M., Winckler,W. and Getz,G. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**(11), 1530–1532.

Durden,C. and Dong,Q. (2009) RICHEST–a web server for richness estimation in biological data. *Bioinformation*, **3**, 296–2988.

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

Falconer,K. (1990) *Fractal Geometry: Mathematical Foundations and Applications.* Wiley, Chichester.

FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed on 27 June, 2013.

Favaro,S., Lijoi,A., Mena,R.H. and Prünster,I. (2009) Bayesian non-parametric inference for species variety with a two-parameter Poisson Dirichlet process prior. *J. Royal Statistical Soc.*, **71**, 993–1008.

Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**(D1), D48–D55.

Garber,M., Grabherr,M.G., Guttman,M. and Trapnell,C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 468–477.

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

Gardner,E.J., Simmons,M.J. and Snustad,D.P. (1991) *Principles of Genetics, 8th edition.* John Wiley & Sons, New York.

Gentleman,R.C, Carey,V.J, Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., Hornik,K., Hothorn,T., Huber,W., Iacus,S., Irizarry,R., Leisch,F., Li,C., Maechler,M., Rossini,A.J., Sawitzki,G., Smith,C., Smyth,G., Tierney,L., Yang,J.Y. and Zhang,J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gneiting,T. and Schlather,M. (2004) Separate Fractal Dimension and the Hurst Effect. *SIAM Rev.*, **46**(2), 269–282.

Gneiting,T., Ševčíková,H. and Percival,D.B. (2012) Estimators of Fractal Dimension: Assessing the Roughness of Time Series and Spatial Data. *Statist. Sci.*, **27**(2), 247–277.

Griebel,T., Zacher,B., Ribeca,P., Raineri,E., Lacroix,V., Guigó,R. and Sammeth,M. (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, **40**, 10073–10083.

Hansen,K.D, Brenner,S.E. and Dudoit,S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.

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

Hoppe,F.M. (1984) Pólya-like urns and the Ewen's sampling formula. *J. Math. Biol.*, **20**, 91–94.

Hower,V., Starfield,R., Roberts,A. and Pachter,L. (2012) Quantifying uniformity of mapped reads. *Bioinformatics*, **28**(20), 2680–2682.

Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Human BodyMap 2.0 data from Illumina (2011) `http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/`. Accessed on 3 August, 2012.

Illumina Quality Scores Overview (2013) `http://www.illumina.com/truseq/quality_101/quality_scores.ilmn`. Accessed on 18 June, 2013.

Illumina (2012) TruSeq RNA Sample Preparation v2 Guide. `http://support.illumina.com/sequencing/sequencing_kits/truseq_rna_sample_prep_kit_v2.ilmn`. Accessed on 24 June, 2013.

Knierim,E., Lucke,B., Schwarz,J.M., Schuelke,M. and Seelow,D. (2011) Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One*, **6**, e28240.

Kumar,R., Ichihashi,Y., Kimura,S., Chitwood,D.H., Headland,L.R., Peng,J., Maloof,J.N. and Sinha, N.R. (2013) A High-Throughput Method for Illumina RNA-Seq Library Preparation. *Front Plant Sci.*, **3**, 202.

Kvan,V.M., Liu,P. and Si,Y. (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.*, **99**, 248–256.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

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

Levin,J.Z., Yassour,M., Adiconis,X., Nusbaum,C., Thompson,D.A., Friedman,N., Gnirke,A. and Regev,A. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**, 709–715.

Li,H., Handsaker,B. Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Lijoi,A., Mena,R.H. and Prünster,I. (2008) A Bayesian Nonparametric Approach for Comparing Clustering Structures in EST Libraries. *J. Comput. Biol.*, **15**, 1315–1327.

A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PloS One*, **7**(12), e52403.

Lindner,M.S., Kollock,M., Zickmann,F. and Renard,B.Y. (2013) Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics*, **2**(10), 1260-1267.

Mandelbrot,B.B. (1982) *The fractal geometry of nature.* W.H. Freeman and Co., San Fransisco,CA.

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

McElroy,K.E., Luciani,F. and Thomas,T. (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, 74.

McIntyre,L.M., Lopiano,K.K., Morse,A.M., Amin,V., Oberg,A.L., Young,L.J. and Nuzhdin,S.V. (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 293.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

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

Nielsen,R., Paul,J.S., Albrechtsen,A. and Song,Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.

Okoniewski,M.J., Leśniewska,A., Szabelska,A., Zyprych-Walczak,J., Ryan,M., Wachtel,M., Morzy,T., Schäfer,B. and Schlapbach,R. (2011) Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage. *Nucleic Acids Res.*, **40**(9), e63.

Oshlack,A., Robinson,M.D. and Young,M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biol.*, **11**, 220.

Oyola,S.O., Otto,T.D., Gu,Y., Maslen,G., Manske,M., Campino,S., Turner,D.J., MacInnis,B., Kwiatkowski,D.P., Swerdlow,H.P. and Quail, M.A. (2012) Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics*, **13**, 1.

Ozsolak,F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Picard-Tools 1.92. (2013) `http://picard.sourceforge.net/`.

Pitman,J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields*, **102**, 145–158.

Pitman,J. (2006) *Combinatorial Stochastic Processes*. Springer, Berlin, Germany.

Proudfoot,N.J., Furger,A. and Dye,M.J. (2002) Integrating mRNA Processing with Transcription. *Cell*, **108**, 501–512.

R-Development-Core-Team (2013) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria.

Richard,H., Schulz,M.H., Sultan,M., Nürnberger,A., Schrinner,S., Balzereit,D., Dagand,E., Rasche,A., Lehrach,H., Vingron,M., Haas, S.A. and Yaspo,M. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.*, **38**, e112.

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

Roberts,A. and Pachter,L. (2012) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, **10**, 71–73.

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

Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2012) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

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

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

Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.

Schena,M., Heller,R.A., Theriault,T.P, Konrad,K., Lachenmeier,E. and Davis,R.W. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends in Biotechnol.*, **16**(7), 301–306.

Schliesky,S., Gowik,U., Weber,A.P.M. and Bräutigam,A. (2012) RNA-Seq Assembly - Are We There Yet? *Front. Plant Sci.*, **3**, 220F.

Schwartz,S., Oren,R. and Ast,G. (2011) Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads. *PLoS One*, **6**, e16685.

Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. and Ren,B. (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Smyth,G.K. and Speed,T. (2003) Normalization of cDNA Microarray Data. *Methods*, **31**(4), 265–273.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**(1), 3.

Smyth,G.K. and Robinson, M.D. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.

Taqqu,M.S., Teverovsky,V. and Willinger,W. (1995) Estimators for long-range dependence: an empirical study. *Fractals*, **3**(4), 785–788.

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

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

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

Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

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

Wang,L., Wang,S. and Li,W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**(16), 2184–2185.

Zabell,S.L. (1992) Predicting the Unpredictable. *Synthese*, **90**, 205–232.

Appendix A

Supplementary Material to Chapter 1

The input amount of mRNA is typically 100ng in about 50μL. This equals to 1×10^5pg in 50μL. Since the average molecular weight per nucleotide is about 330pg/pmol and the average mRNA length is about 1200bp, $1 \times 10^5/(1200 * 330) = 0.25$ pmoles are in 50μL. Multiplication with Avogadro's number ($\sim 6.022 \times 10^{23}$) that is the number of molecules per mole yields the final number of molecules ($0.25 \times 10^{-12} \times 6.022 \times 10^{23} = 1.5 \times 10^{11}$).

A typical library concentration is about 30μL of a 200nM solution. Note that the Molar concentration M is always specified per Liter. Therefore the number of pmoles is about 6: 200nM $= 2 \times 10^5$pM, and accounting for the volume of 30μL: 2×10^5 pmoles/L $= 2 \times 10^5$ pmoles/$10^6 \mu$L, and thus $(2 \times 10^5/10^6) \times 30 = 6$ pmoles. Multiplication with Avogadro's number yields 3.6×10^{12} ($6 \times 10^{-12} \times 6.022 \times 10^{23}$).
The same calculation holds for the concentration measured when loading the sample onto the flow cell (120μL of a 7pM solution).

Appendix B

Supplementary Figures to Chapter 3

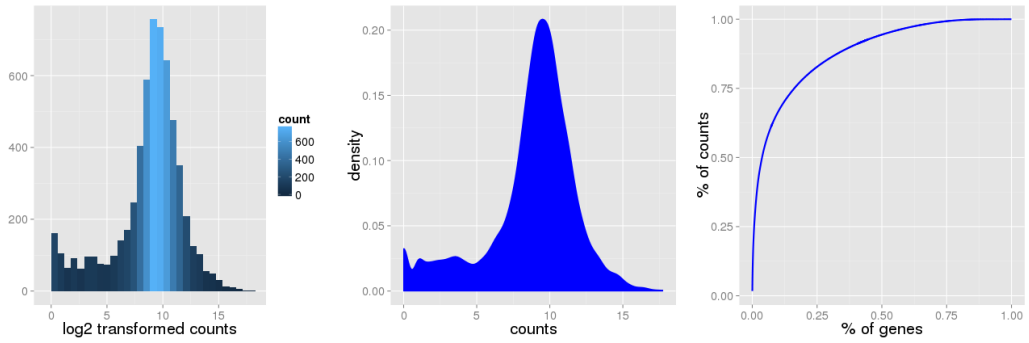

Figure B.1: Summary statistics of the aligned read data. The left most panel shows an histogram of the log 2 transformed read counts. A smooth density estimation is depicted in the middle. Finally, the third figure serves for providing some intuition how evenly reads are distributed between isoforms. Typically few genes collect the majority of the reads in RNA-Seq data.

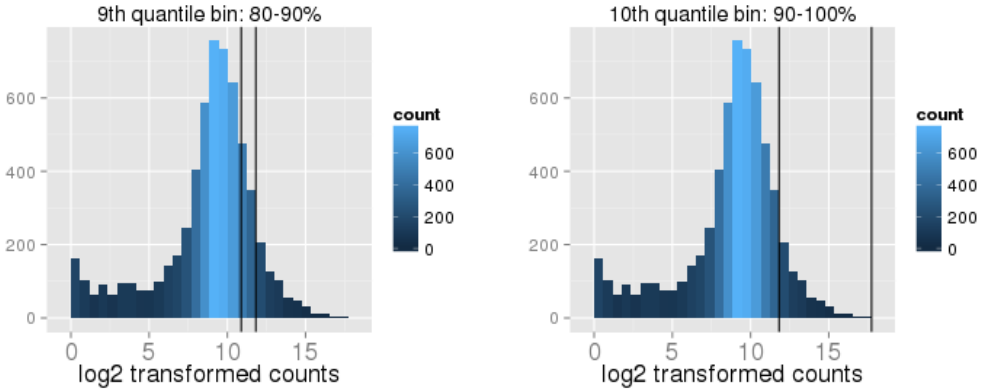

Figure B.2: Isoforms are split into equal sized read count bins. In order to facilitate the interpretation of the absolute read count values, the bins are displayed within the overall read count distribution.

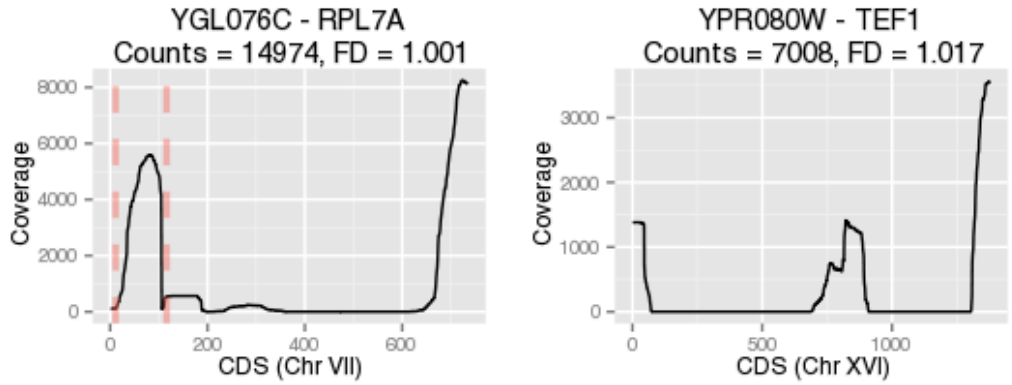

Figure B.3: Typical coverage graph in the HTML report of *FractalQC*. The length of the isoform is displayed on the x-axis while the per-base coverage pattern is shown on the y-axis. Read count information is given in the title. Upon clicking on the gene ID the corresponding entry in the ENSEMBL database is opened in the web browser (see Figure B.4). Red dashed lines indicate exon boundaries as annotated in the databases.

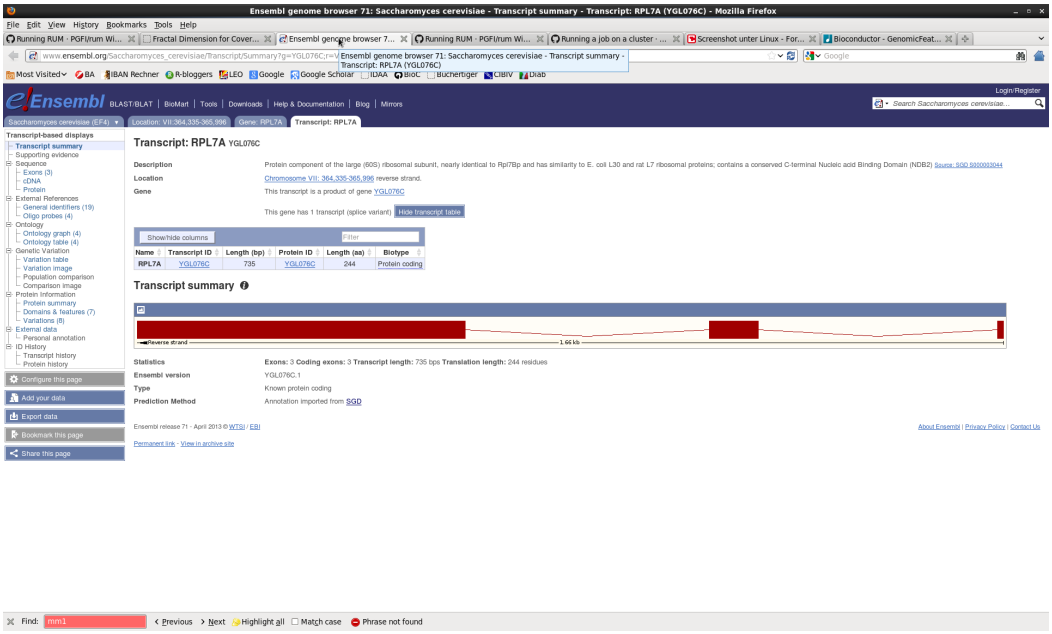

Figure B.4: Database entry for a specific isoform.