



TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

JANNE SEPPÄLÄ

**NOVELLETTE: AN RNA-SEQUENCING DATA ANALYSIS PIPE-
LINE FOR DETECTING NOVEL TRANSCRIPTS**

Master of Science Thesis

Examiner: Prof. Olli Yli-Harja
Tarkastaja ja aihe hyväksytty
Examiner and subject accepted by
the department council 7.9.2011

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO

Biotekniikan koulutusohjelma

SEPPÄLÄ, JANNE: Novellette: an RNA-sequencing data analysis pipeline for detecting novel transcripts

Diplomityö, 53 sivua, 2 liitesivua

Marraskuu 2013

Pääaine: Laskennallinen systeemibiologia

Tarkastaja: Olli Yli-Harja

Avainsanat: RNA, DNA, sekvensointi, novelli transkripti, hypoteesin testaus

Proteiinit ovat avaintekijöitä jokaisen elävän organismin solujen biologisissa prosesseissa ja rakenteissa. Proteiineja muodostetaan proteiinisynteesiksi nimitetyssä prosessissa, jossa DNA:n sisältämä geneettinen koodi muutetaan ensin RNA:ksi transkriptiossa ja lopulta proteiiniksi translaatiossa. Geenit ovat lyhyitä pätkiä DNA-molekyylistä, jotka koodaavat yhden tai useamman tietyn proteiinin synteesiä. Tällä hetkellä yksi yleisimmistä menetelmistä geenien ilmentymisen tutkimisessa soluissa on RNA-sekvensointi, joka sitoo solusta tietyllä ajanhetkellä siellä olevan RNA:n. Vaikka RNA-sekvensoinnilla yleensä mitataankin tunnettujen geenien ilmentymistä, sitä voidaan hyödyntää myös uusien geenien eli novellien transkriptien etsimisessä.

Tähän mennessä ei ole vielä julkaistu standardia työkalua, joka tehokkaasti ja kattavasti tunnistaisi novelleja transkripteja RNA-sekvensointidatasta. Tässä työssä on kehitetty työkalu - nimeltään Novellette - ratkaisemaan edellä esitetty ongelma. Novellette pyrkii tunnistamaan ilmentymiseltään merkittäviä alueita genomissa, jotka ovat erillään kaikista tunnetuista geneistä, ja suorittaa sitten löydetyille alueille kattavan geenirakennepiirteanalyysin. Tässä analyysissä hyödynnetään sekä RNA-sekvensoinnilla tuotettua dataa että tutkittavan organismin tunnettua DNA-sekvenssiä proteiinia koodaavien geenien tunnusomaisia rakennepiirteitä etsittäessä löydetyistä novellista transkriptiehdokkaista. Sen jälkeen lödyt piirteet pisteytetään ja lopulliset novellit transkriptiehdokkaat järjestetään näiden pistearvojen perusteella. RNA-sekvensoinnin data-analyysityökalun kehittämisen lisäksi tässä työssä esitellään RNA-sekvensoinnin data-analyysiin olennaisesti liittyvää tilastomatematiikkaa sekä matemaattisia menetelmiä, ja tutkitaan RNA-sekvensointidatan normaalisuutta julkisesti saatavilla olevan RNA-sekvensointidatan avulla.

Erilaisilla aineistoilla suoritettujen testien perusteella Novellette pystyy luotettavasti tunnistamaan novelleja transkripteja ja erottamaan toisistaan proteiinia koodaavat ja koodaamattomat alueet genomissa työssä kehitetyllä pisteytysmenetelmällä. Lisäksi näytetään, että RNA-sekvensointidata noudattaa heikosti normaalijakaumaa ja siten korostaa sellaisten tilastollisten hypoteesin testausmenetelmien tärkeyttä, jotka eivät perustu datan normaaliudelle. Yhteenvedona todettakoon, että tässä työssä kehitetty työkalu, Novellette, on osoittautunut hyödylliseksi ja toimivaksi, ja sillä on potentiaalia kehittyä standardiksi novellien transkriptien analysointimenetelmäksi.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY

Master's Degree Programme in Biotechnology

SEPPÄLÄ, JANNE: Novellette: an RNA-sequencing data analysis pipeline for detecting novel transcripts

Master of Science Thesis, 53 pages, 2 Appendix pages

November 2013

Major: Computational Systems Biology

Examiner: Olli Yli-Harja

Keywords: RNA, DNA, sequencing, novel transcript, hypothesis testing

Proteins are the key factors in every living organism and they contribute to almost every biological process and structure in a cell. Proteins are formed through the process of protein synthesis, in which the genetic code of DNA in genes is first transcribed into RNA and then finally translated into a protein. Each gene in a cell is a short part of a longer DNA molecule, and each gene encodes the synthesis of a certain protein or proteins. Currently the state-of-the-art tool to evaluate which genes are expressed in a given biological sample is RNA-sequencing, which captures the RNA content of the cells at a specific time point. Although RNA-sequencing is often used to measure the expression of known genes in the genome, it can also be utilized to search for new genes, or novel transcripts.

To date, no standard tool has been published that effectively and thoroughly identifies novel transcripts from RNA-sequencing data. In this work, a tool aiming to solve this issue – denoted Novellette – is presented. Novellette attempts to identify differentially expressed regions in the genome that do not overlap with any known genes, and then performs a full gene structure analysis to the regions. For this process, information both from processed RNA-sequencing data and the known DNA sequence of the studied organism is utilized when searching for features in the novel transcript candidates that are common for protein-coding genes. The features are then scored and the final novel transcript candidates are ranked based on their score values. In addition to developing an RNA-sequencing tool in this work, the basics of statistical testing and other mathematical methods related to RNA-sequencing data analysis are introduced and the normality of count based RNA-sequencing data is assessed with publically available data.

The results from analyses performed with various input data show that Novellette is able to reliably detect novel transcripts and distinguish protein-coding regions from non-coding regions in the genome with the proposed scoring approach. In addition, the count based RNA-sequencing data is shown to very poorly follow the normal distribution, hence pinpointing the importance of statistical hypothesis testing methods that do not assume data normality. In conclusion, a functional and useful bioinformatics tool has been developed in this work that has the potential to become a standard method for novel transcript identification.

PREFACE

This Master of Science thesis was carried out in cooperation with both the Department of Mathematics and the Department of Signal Processing in Tampere University of Technology.

The examiner prof. Olli Yli-Harja is acknowledged for granting me the opportunity to work in the Computational Systems Biology group for many years, and the supervisors prof. Keijo Ruohonen (mathematics) and prof. Matti Nykter (signal processing) are acknowledged for their advice in the thesis structure and for their patience while I was unable to progress with the writing process. My colleagues Matti Annala and Liisa-Ida Sorsa are also greatly thanked for their advice in the mathematics theory, and special thanks are given to Matti for always showing how a good scientist would implement software tools in a different way than how I am doing. Finally, the family - Laura Pekkarinen and the two lovely golden retrievers (Cola and Lyca) - are given the most grateful acknowledgements for cheering me up while I was in the deepest depression state during the whole thesis process.

Tampere, October 31st, 2013

Janne Seppälä

TABLE OF CONTENTS

Abstract	iii
Terms and abbreviations	vii
1 Introduction	2
2 Background.....	4
2.1 Biological background	4
2.1.1 The central dogma of molecular biology.....	4
2.1.2 Gene structure and expression.....	6
2.1.3 Cancer and the role of coding and non-coding RNA	7
2.2 Technical background	8
2.2.1 DNA- and RNA-Sequencing.....	8
2.2.2 Preliminary data analysis.....	9
2.2.3 Gene prediction	11
2.3 RNA-seq data normalization.....	12
2.3.1 RPM normalization	12
2.3.2 RPKM normalization.....	12
2.3.3 Quantile normalization	13
2.4 Statistical testing	13
2.4.1 Z-test.....	14
2.4.2 t-test	16
2.4.3 U-test	18
2.5 Assessing data similarity and normality.....	19
2.5.1 Graphical methods.....	19
2.5.2 Kolmogorov-Smirnov test.....	22
2.5.3 Shapiro-Wilk test.....	25
2.6 Survival analysis	26
2.6.1 Kaplan-Meier estimate and plot	27
2.6.2 Logrank test.....	29
3 Methods	32
3.1 Motivation	32
3.2 Overview	32
3.3 Detection of novel transcript candidates	33
3.3.1 Calculating normalized read counts	34
3.3.2 Two-class analysis.....	34
3.3.3 One-class analysis	36
3.3.4 Filtering and merging	36
3.4 Gene identification	37

3.4.1	Defining and scoring the exon structure.....	37
3.4.2	Promoter score.....	39
3.4.3	3'UTR.....	42
3.4.4	Open reading frames.....	44
3.4.5	Result printing	45
4	Results	48
4.1	One-class analysis with GBM data	48
4.1.1	Preliminary analysis and outlier expression	48
4.1.2	Gene identification	49
4.1.3	Survival analysis.....	49
4.2	Two-class analysis: CRPC vs. PC.....	51
4.3	Performance of the scoring approach.....	52
5	Discussion.....	53
	References	54
	Appendix A	60
	Appendix B	61

TERMS AND ABBREVIATIONS

3' end	The phosphate terminus of a DNA molecule.
5' end	The sugar terminus of a DNA molecule.
Bp	Base pair, the basic unit of a dsDNA molecule. Also a unit used to express the length of a DNA molecule.
cDNA	Complementary DNA.
CDF	Cumulative distribution function.
CDS	Coding sequence.
Chromosome	A structure in the cell consisting of a large DNA molecule (millions of bps in length). A human cell has 24 different kind of chromosomes.
Codon	A trinucleotide in mRNA that encodes for a certain amino acid in translation.
Contig	Contiguous sequence.
CRPC	Castration resistant prostate cancer.
CpG island	A region with a relatively high fraction of CG-dinucleotides.
DNA	Deoxyribonucleic acid.
DNA-seq	DNA-sequencing.
dsDNA	Double-stranded DNA.
eCDF	Empirical cumulative distribution function.
Exon	A part of a gene, which remains in the mature mRNA and hence affects the protein formed in translation.
Gene	Part of a DNA molecule, which encodes for a certain protein or proteins.
Genome	The full DNA sequence of an organism, containing all genes.
Genome browser	A widely used group of tools for data visualization in bioinformatics studies.
GBM	Glioblastoma multiforme.
HTS	High-throughput sequencing, see MPS.
Hypothesis testing	A statistical method to aid in drawing conclusions of a given dataset.
Intron	A part of a gene, which is removed from the mature mRNA prior to translation.
K-M plot	Kaplan-Meier plot, a graphical representation of the survival estimate of a given population, which is often used in survival analyses.
mRNA	(Mature) messenger RNA, the end product of transcription containing only the untranslated regions and exons.

Motif	A specific DNA sequence that a certain protein is able to bind to.
MPS	Massively parallel sequencing.
ncRNA	Non-coding RNA.
Normal distribution	The most common mathematical model of data distribution.
Nucleotide	The basic unit of a DNA molecule, which consists of a sugar group, a phosphate group and a base.
ORF	Open reading frame.
PC	Prostate cancer.
pre-mRNA	The first product of transcription, which contains both introns and exons.
Promoter	A region near the 5' end of a gene, which contains motifs that transcription factors can bind to and enable RNA polymerase to initiate transcription.
p-value	An indicator of statistical (in)significance in hypothesis testing.
Reference genome	An average genomic DNA sequence formed from several individual genomes of the same species.
RNA	Ribonucleic acid.
RNA polymerase	The enzyme responsible for transcribing DNA into RNA.
RNAi	RNA interference, the process in which a two ssRNA molecules form a double-stranded RNA molecule, which then gets cleaved into small fragments by enzymes.
RNA-seq	RNA-sequencing.
RPKM	Reads Per Kilobase of transcript length per Million mapped reads, a normalization method in sequencing data analysis.
RPM	Reads Per Million, a normalization method in sequencing data analysis.
Risk level	A threshold set for the p-value in statistical hypothesis testing.
Sequence	The order of nucleotides in a DNA or RNA molecule.
Sequencing	The process in which the sequence of a DNA or RNA molecule is investigated.
ssDNA	Single-stranded DNA.
ssRNA	Single-stranded RNA.
TCGA	The Cancer Genome Atlas.
Test statistic	A random variable, the value of which is calculated when a hypothesis test is performed.
TP53	Tumor protein 53.
Translation	Protein synthesis, the process in which a protein molecule is produced based on the RNA sequence of a mRNA.

Transcript	A common name for any type of RNA molecule formed in transcription.
Transcription	RNA synthesis, the process in which an RNA molecule is produced based on the DNA sequence of a gene.
Transcriptome	The full, transcribed RNA sequence content of an organism.
UTR	Untranslated region.
Wetlab	Biological laboratory.

1 INTRODUCTION

Genes and DNA form the basis of life and heritage. Molecular genomics is one of the most widely studied fields of biology, which focuses on the molecular events in a living cell, including RNA and protein synthesis, gene regulation and epigenetic mechanisms such as DNA methylation. As the measurement technologies have evolved rapidly in the past few decades, it is now possible to measure the expression of all genes in a cell simultaneously. One of these methods is RNA-sequencing, which enables accurate quantification of RNA expression levels in a cell, including the expression of genes.

In a typical RNA-sequencing experiment, the expression of known genes is examined. Although the human genome has been fully sequenced and the function and genomic locations of most of the genes have been determined, there are still partly unknown regions in the genome that might encode for a functional protein. Some of these regions arise only in some specific states of the cell, such as in disease or cancer. Therefore it is important that one does not limit to studying only the expression of known protein-coding genes. However, there is currently no standard method to discover novel gene transcripts from RNA-sequencing data and to determine whether these transcripts exhibit appropriate gene structure in order to encode functional proteins.

In this work, a data analysis pipeline to identify novel transcripts from RNA-sequencing data and to determine their gene structure – Novellette – is presented. The pipeline features the analysis of both single- and multi-class sample sets followed by the prediction of gene structure by utilizing RNA-sequencing measurements processed with a splice-junction mapper (Kim et al. 2013; Dobin et al. 2013) in addition to using the known reference genome sequence. Instead of making binary calls whether or not an identified novel transcript is protein-coding, each transcript is given a score value $S \in [0,1]$ that describes its protein coding potential (higher score corresponding to a higher protein coding potential). The score consists of several different components, each of which accounts for a certain feature that is common in a protein-coding gene (such as a valid exon-intron structure, promoter motifs etc.). As an output, two different files are produced: a summary table with detailed information of each of the identified transcripts and their structures, and another file that can be opened in a genome browser for graphical illustration of the transcripts and their structures.

As a typical RNA-sequencing data analysis consists of a wide range of mathematical models and statistical hypothesis tests, the basics of statistics related to the analysis performed by Novellette are also discussed in this work. The topics covered in this work include statistical hypothesis testing, data normality and similarity assessment and survival analysis. In addition, the normality of RNA-sequencing data is evaluated by using

publically available data from The Cancer Genome Atlas (TCGA) glioblastoma project (Brennan et al. 2013).

In Chapter 2, the basics of molecular biology including the process of forming proteins based on the DNA sequence of genes, the structure of genes and the basics of cancer biology are discussed, and the DNA- and RNA-sequencing techniques with their typical preliminary data analysis steps are introduced. In addition, the mathematical theory related to this work is also covered in this chapter. In Chapter 3, the pipeline developed in this work is presented in detail and data normality is evaluated with the glioblastoma RNA-sequencing data to justify the choice of statistical testing method in Novellette. Chapter 4 presents the results of several test runs with the developed pipeline and, finally, the results and data normality in an RNA-sequencing project are discussed in Chapter 5. The main functions and tools covering the pipeline of Novellette are listed in Appendix A, along with a download link to the full source codes.

2 BACKGROUND

In this chapter, the biological and technical background related to this work is discussed. The first section focuses on molecular biology elements of the cell, and the second and third sections focus on the techniques used to measure, analyze and normalize (raw) sequencing data. Finally, the last two sections cover the mathematical theory in statistical testing and survival analysis, which are standard downstream analyses in cancer studies.

2.1 Biological background

In this section, the basics of molecular biology related to DNA, genes and proteins are introduced and their role in the development of cancer and other diseases is discussed. As the biological motivation in this work is cancer and disease related, the molecular components and mechanisms are inspected from a human biology point of view.

2.1.1 The central dogma of molecular biology

According to the well-known central dogma of molecular biology, the amino acid sequence of a certain protein is encoded in the DNA sequence of a gene, which is first transcribed into an RNA molecule and then translated into the protein (Figure 2.1). This process involves many different molecules in the cell in addition to the DNA and RNA molecules and the protein, which is the end product of this operation. The transcription process can also be reversed in order to produce DNA based on the RNA sequence, in which case the end product is cDNA (*complementary DNA*).

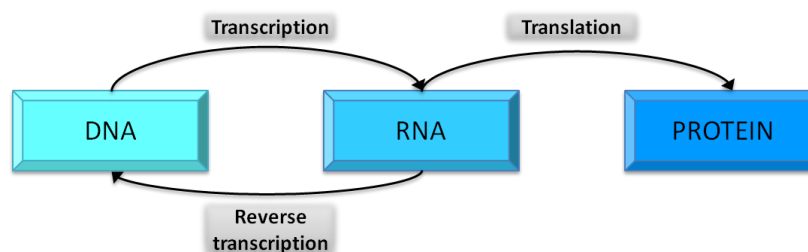


Figure 2.1. *The central dogma of molecular biology.*

DNA (*deoxyribonucleic acid*) is a large biomolecule that contains all of the necessary information to produce a protein in the cell. The basic components of DNA are called nucleotides, which consist of a sugar-phosphate backbone and one of four different possible bases (A, C, G and T, or *adenine*, *cytosine*, *guanine* and *thymine*, respectively).

The bases of consecutive nucleotides determine the sequence of a DNA molecule: e.g. a DNA molecule with the sequence ‘AGCAAT’ is a six nucleotides long molecule, which contain the bases adenine, guanine, cytosine, adenine, adenine and thymine, in this order. RNA (*ribonucleic acid*) is almost identical to DNA in its structure, with a small change in the sugar molecule of the sugar-phosphate backbone and the replacement of thymine with another base, U (*uracil*).

DNA molecules in the cell are very rarely in the simple, single-stranded form of sugar-phosphate backbone with attached bases. Instead, a single-stranded DNA (*ssDNA*) is typically paired with another ssDNA molecule via hydrogen bonds, resulting in double-stranded DNA (*dsDNA*). The binding takes place between the bases of the two ssDNA molecules in a complementary fashion: A binds to T and C binds to G. On the other hand, most of the RNA in a human cell is single-stranded. In fact, if a single-stranded RNA (*ssRNA*) would hybridize with another RNA molecule, it would get cleaved into small fragments by enzymes in a process called RNA interference (*RNAi*) [Macrae et al. 2006]. In Figure 2.2, the structure of both ssDNA and dsDNA are illustrated.

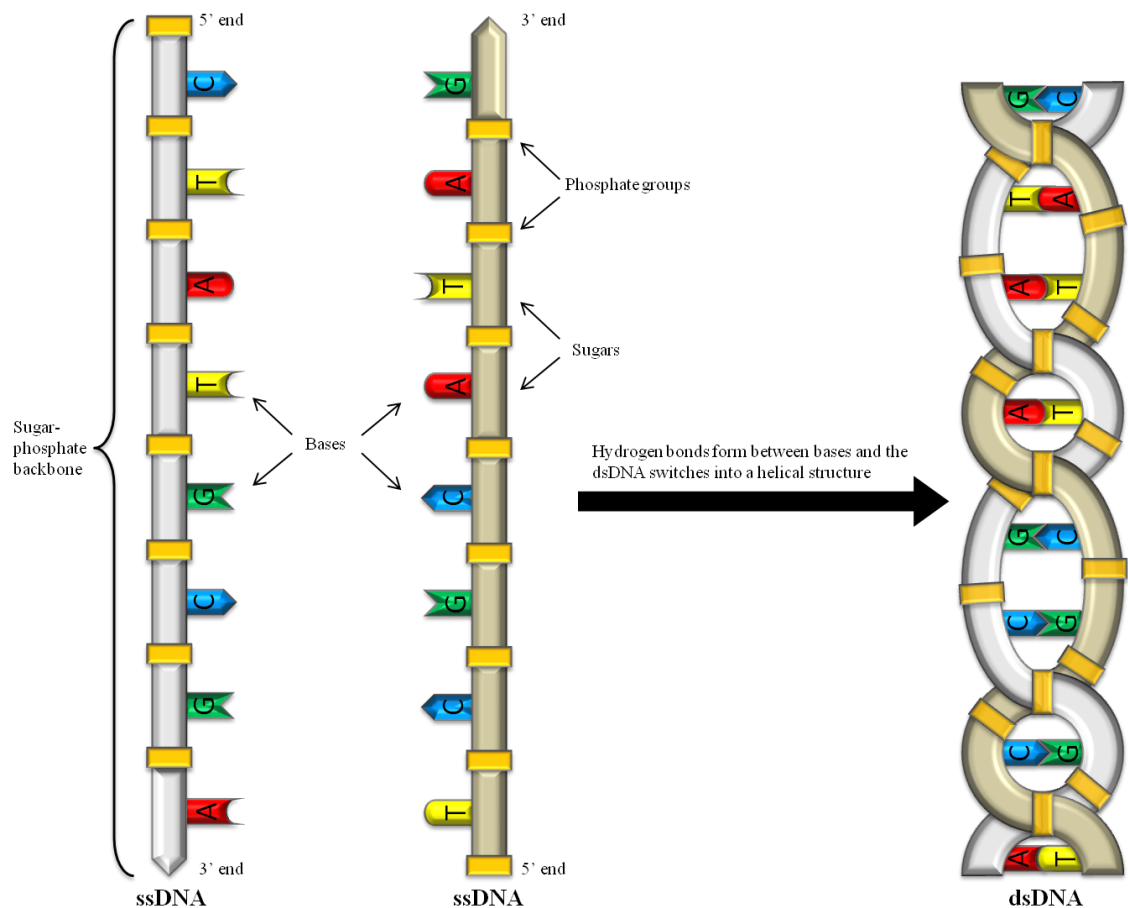


Figure 2.2. The structure of ssDNA and double-stranded, helical DNA. The phosphate terminus of DNA is often denoted 3' end and the sugar terminus 5' end.

In humans and most other eukaryotic organisms, the DNA of a cell is divided into chromosomes. A human cell has 24 different chromosomes: chr1, chr2, ..., chr22, chrX

and chrY. The mitochondria also contain DNA and can hence also be considered as distinct chromosomes (chrM). Most human cells are diploid, meaning that they have an extra copy of chromosomes 1 to 22 but still only a single copy of the sex chromosomes X and Y. Gametes (egg and sperm) are haploid, however, carrying only a single copy of each chromosome. Different chromosomes are of different size in terms of the number of nucleotides, and the total haploid, single-stranded DNA in a human cell consists of approximately 3 billion nucleotides. Since DNA is usually double-stranded and the pairing between two ssDNAs occurs between the bases, sequence lengths are often expressed as base pairs (*bp*) instead of nucleotides.

Genes are specific regions in the DNA on either strand of the dsDNA, spanning from a thousand to a few million nucleotides in length, which encode for a certain protein. The genetic code is stored in the DNA sequence of the gene. In the first step of the central dogma of molecular biology (see Figure 2.1), the DNA of a gene is transcribed into a pre-messenger RNA (*pre-mRNA*) molecule, with a sequence complementary to the DNA sequence of the gene. After several steps of modifications (see Section 2.1.2) to the pre-mRNA molecule, a mature mRNA molecule is formed, which is then used to determine the amino acid sequence of the protein formed in translation.

2.1.2 Gene structure and expression

A typical protein-coding gene consists of a promoter region followed by the transcription start site (*TSS*) and several exons with introns between them (Figure 2.3a). The promoter region contains binding sites for various transcription factors, which are proteins necessary to enable RNA polymerase, another protein, to bind on DNA. In general, the DNA sequence which a certain protein identifies and is able to bind on to, is called a *motif*. In addition, the promoter often also contains regions with a relatively high number of CG-dinucleotides (*CpG islands*). At the *TSS*, the RNA polymerase binds on the DNA and starts transcribing the DNA sequence into RNA. The mechanism of transcription termination in humans and other eukaryotic organisms is not yet fully understood, but it involves the polyadenylation (*poly-A*) signal of mRNA. When RNA polymerase encounters the poly-A signal (sequence AAUAAA), transcription ends and the pre-mRNA molecule is finished.

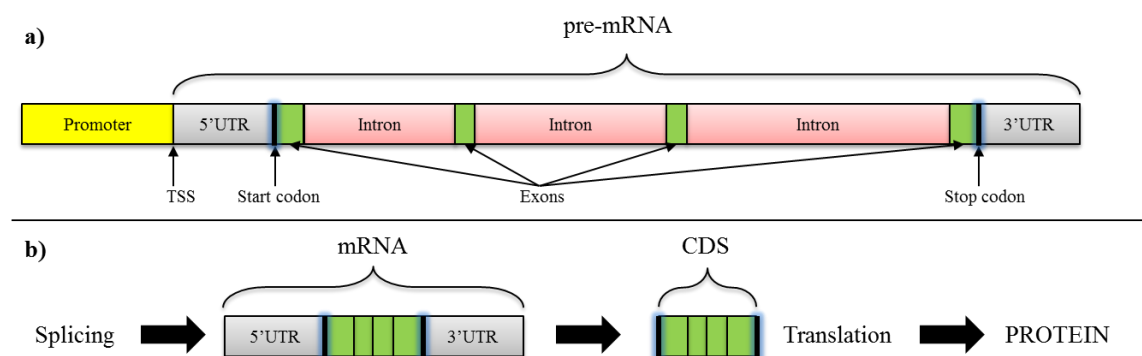


Figure 2.3. The basic components of a gene (a) and the mechanism of gene expression (b).

After forming the pre-mRNA molecule, introns are removed from it in a process called splicing. Splicing is carried out by proteins that recognize the dinucleotides GU and AG in the 5' and 3' ends of each intron, respectively. As a result, a mature mRNA with only the 5' and 3' untranslated regions (*UTR*) and exons is formed. The mRNA sequence is then processed in sets of three consecutive nucleotides (*codon*), beginning from the start codon (sequence AUG) and ending in the stop codon (sequence UAA, UGA or UAG). The region in mRNA between the start and stop codon, usually including all of the exons, is denoted the coding sequence (*CDS*) of the gene. The final protein is then produced in translation based on the CDS (Figure 2.3b). In general, a region in an RNA sequence between a start and stop codon is denoted an open reading frame (*ORF*), which is a common name for regions in RNA that could potentially be translated into a protein. Due to nonsense-mediated decay (Chang et al. 2007), the last exon must be contained in the ORF of a protein-coding gene or else the mature mRNA molecule will be degraded.

2.1.3 Cancer and the role of coding and non-coding RNA

In 2008, cancer accounted for approximately 13 % of all deaths worldwide and was hence the most common cause of death in western countries, and the second most common cause in developing countries (WHO Fact sheet 2013). Cancer forms when cells start growing and dividing in an uncontrollable manner. This may occur due to several different reasons, many of which are still unknown. However, in most cases the formation of cancer can be explained with alterations in oncogenes, tumor-suppressor genes or microRNA genes (for a review, see (Croce 2008)). Oncogenes are genes that may not be dangerous in normal conditions, but when their expression is altered, they can cause cancer. Tumor-suppressors on the other hand are genes that demote cell growth and division in normal conditions, e.g. by contributing to the controlled cell death (*apoptosis*), but cause cancer if their expression is turned off e.g. due to mutations. Finally, microRNAs are a group of small RNA molecules that cause RNA interference by binding to mRNA, hence preventing a gene from being translated into a protein.

One of the most important tumor-suppressor genes is the gene TP53 (*tumor protein 53*), which has been found to be mutated in many different human cancers (for a review, see (Hainaut & Hollstein 2000)). In addition to TP53, hundreds of other genes have been shown to correlate with the progression of various cancer types. However, as some cancer cases cannot be fully explained with any of the known gene markers, recent cancer studies have shifted the focus from known genes to previously unknown areas of the genome. One such study reported 121 novel, non-coding RNA (*ncRNA*) transcripts in prostate cancer, one of which (PCAT-1) was shown to have very high potential for being a cancer subtype marker in prostate cancer (Prensner et al. 2011). Therefore it is important not to neglect regions in the genome that contain no known genes, as they may yield information on cancer progression and subtype distinction.

2.2 Technical background

This section covers the background in sequencing technologies and in the preliminary data analysis steps performed in a typical high-throughput sequencing process. In addition, the basics of gene prediction and related algorithms are introduced.

2.2.1 DNA- and RNA-Sequencing

DNA-sequencing has been a widely used method in biotechnology to determine the DNA sequence of a gene or even a whole bacterial organism for decades. The first sequencing technologies were published in 1970s (Sanger & Coulson 1975; Maxam & Gilbert 1977), but they were able to sequence only short, 300 – 1000 nucleotide long DNA molecules. Within the last ten years, however, sequencing technologies have evolved rapidly, allowing the sequencing of the whole human genome at once. This is enabled by massively parallel sequencing (*MPS*), also denoted high-throughput sequencing (*HTS*), which can process millions of short DNA fragments simultaneously.

In DNA-sequencing (*DNA-seq*), the DNA is first extracted from target cells and purified from other cellular components and molecules. After that the DNA is cleaved into small fragments of approximately 200 – 500 bps e.g. with sonication, and adapter sequences, specific to the sequencing platform used, are added to the fragments. Finally, this set of fragments (often denoted *DNA library*) is sent to a sequencer for high-throughput sequencing. Single-end and paired-end sequencing are currently the most commonly used protocols in HTS. In single-end sequencing, each DNA fragment is sequenced only from one end of the fragment. In paired-end sequencing, both ends of the fragment are sequenced instead. In either case, the fragment is usually only partly sequenced. The output of the sequencer is one large text file per sample, containing the sequences of the fragment ends (single-end sequencing), or two files per sample in the case of paired-end sequencing: both files contain the sequences of the same fragments, but from different ends. These short output sequences are often denoted *reads* or *tags*.

RNA-sequencing (*RNA-seq*) does not fundamentally differ from DNA-seq apart from the library preparation process. In RNA-seq library preparation, mature mRNA in the cell is first isolated and purified from other RNA content by utilizing its poly-A tail as a primer site for reverse transcription (Mortazavi et al. 2008). The mRNA is then reverse-transcribed into cDNA, which is further processed in a similar way as in the DNA-seq protocol. When this cDNA library is finally sequenced, the output sequences should consist of the cDNA of actively expressed genes in the cell. A summary of the RNA-seq protocol is illustrated in Figure 2.4.

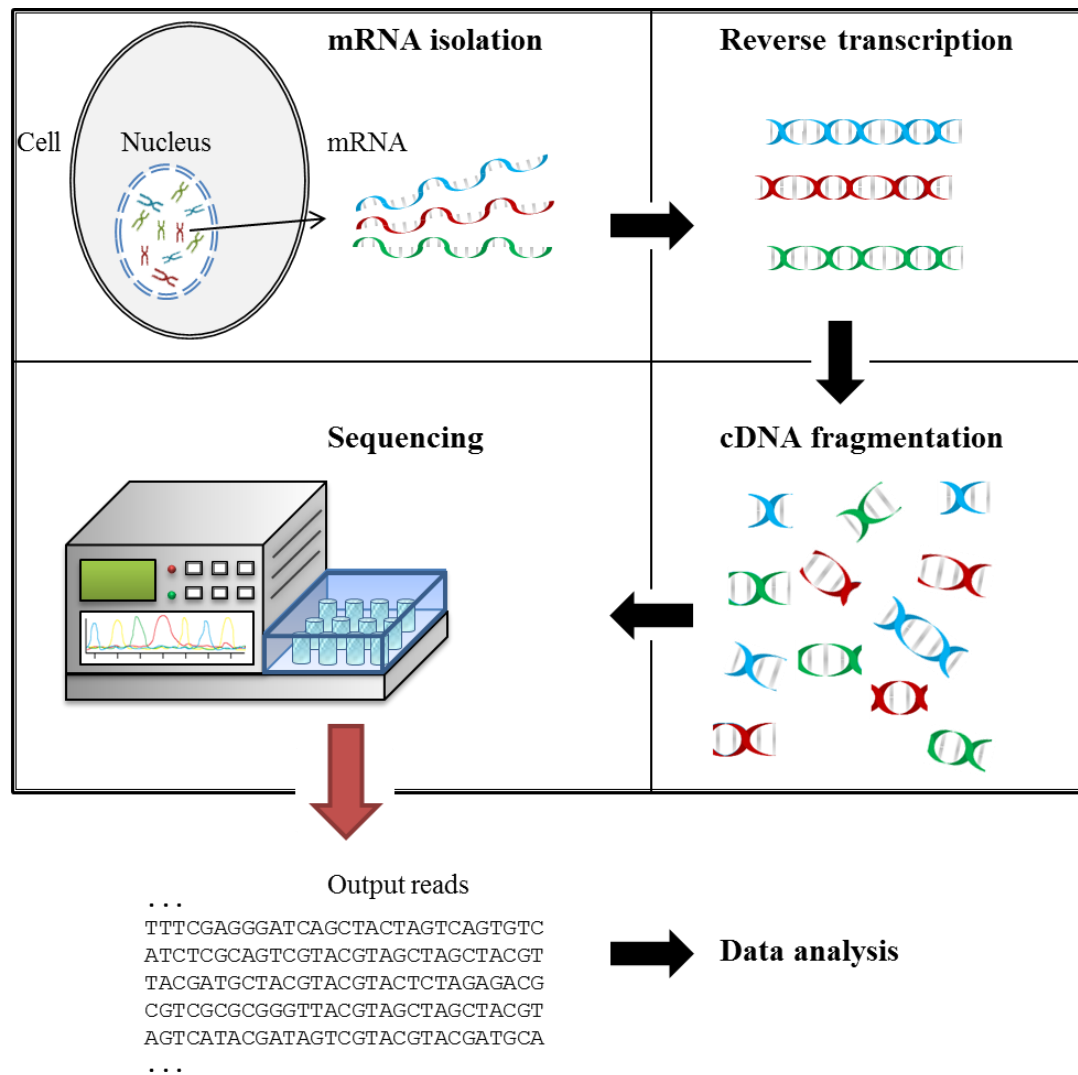


Figure 2.4. The workflow of a typical RNA-seq experiment. This chart covers only the main phases, although the library preparation (from mRNA isolation to cDNA fragmentation) often contains many additional steps before the cDNA can be sequenced.

2.2.2 Preliminary data analysis

This subsection covers some of the main data analysis paths after getting the output reads from RNA-seq. The amount and type of output reads depends on the sequencer used, but in most cases the read length is constant and the output file consists of the cDNA sequences and quality values in FASTQ format (Cock et al. 2010), the latter of which describes the certainty at which each nucleotide in the cDNA has been detected by the sequencer. In an RNA-seq project, the aim can be either to study the expression of known genes of a known organism, to build the *transcriptome* (the set of transcribed regions of the genome) of a certain organism or to study the expression of ncRNA or other unannotated regions of the genome.

When the gene expression levels of a known organism (such as human) are examined with RNA-seq, the origin of each read output by the sequencer should first be determined by comparing its sequence with the known genomic DNA or RNA sequence in

a process called *alignment*. In a DNA sequencing project the reads are aligned against a *reference genome*, after which the chromosome and specific location within the chromosome (*chromosomal coordinates*) can be determined for each read that yielded a unique match with the reference genome sequence (Figure 2.5a). The reference genome is built from the whole genome sequences of a pool of individual organisms of the same species, and often only the sequence of one strand is reported, even if the genome usually exists as dsDNA. However, since the DNA sequences of different individuals of the same organism differ significantly from each other, a uniform reference sequence that would match the DNA of every other organism of the same species cannot be built. Therefore the alignment process is always prone to bias.

In RNA-seq, the cDNA reads are often aligned against a *reference transcriptome*, which is a set of sequences where one continuous sequence corresponds to the sequence of the corresponding mRNA of a certain gene (Figure 2.5b). With this approach the number of matching reads for each gene can be directly calculated from the alignment process output. This number is then considered to correlate (positively) with the expression level of the gene. When aligning against reference transcriptome, however, only the expression of known genes can be measured, while all unknown regions - that could potentially contain a protein-coding or non-coding transcript - are ignored. Therefore aligning against a reference genome is a more thorough approach in RNA-seq, although making the gene expression value calculation less straightforward.

A third approach to processing the raw read data in RNA-seq is to assemble a *de novo transcriptome*, i.e. to compare the read sequences with each other instead of a known reference transcriptome (Figure 2.5c). Each contiguous sequence (*contig*), established by connecting consecutive, overlapping reads, corresponds to a transcribed region (*transcript*) in the genome. Transcriptome assembly is especially useful when studying an organism whose transcriptome is poorly known or even completely unknown and it is an accurate way to map the expressed regions of the genome since it does not rely on any prior knowledge about the genome or transcriptome.

As the amount of output data in a sequencing project is enormous and the task of comparing millions of short sequences with a 3 billion bps long reference sequence is computationally intensive, numerous algorithms have been developed for aligning DNA-seq (Langmead et al. 2009; Li & Durbin 2009) and RNA-seq (Kim et al. 2013; Dobin et al. 2013) reads as well as for de novo assembly of both genomes (Simpson et al. 2009, Zerbino & Birney 2008) and transcriptomes (Haas et al. 2013; Trapnell et al. 2010). To reduce compatibility issues between aligners and downstream analysis tools, a standard format for storing alignment results as well as a toolkit to process them has been published (Li et al. 2009). In addition, for visualizing alignment data as well as several different types of processed data among with known genome features (such as genes), several tools called genome browsers have been developed (Kuhn et al. 2013; Flicek et al. 2013; Thorvaldsdottir et al. 2013).

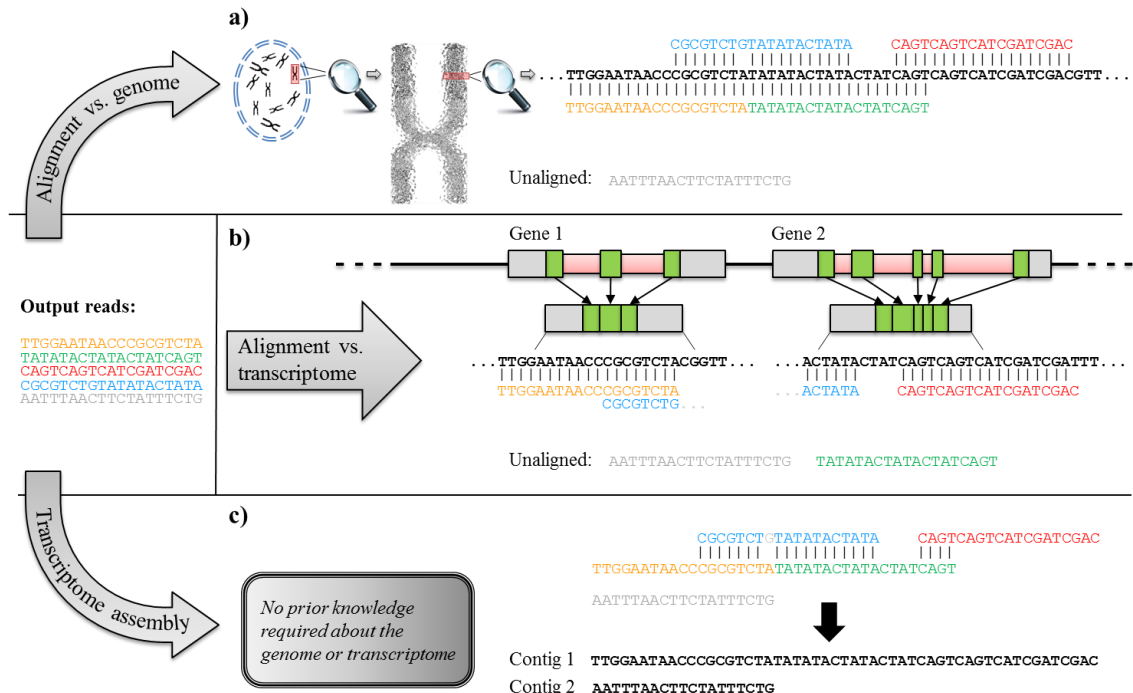


Figure 2.5. Summarization of the three principal data analysis paths described in this subsection. As input to the data analyses, five short reads output by the sequencer are given in this example case. When aligning against a reference genome (a), four out of five reads have a perfect or near-perfect match with a certain chromosome in the nucleus. In the transcriptome alignment case (b), two reads have a complete or near-complete match to the mRNA sequence of a gene, while one read has a partial match for both of the genes and two reads have no match at all. In the last case (c), sets of overlapping reads are used to establish full-length contigs. As one of the reads has poor sequence match with the other reads, it is presented as a separate contig.

2.2.3 Gene prediction

In gene prediction, also known as gene identification or gene finding, protein-coding sequences are searched from a given input DNA sequence. In its simplest form, only open reading frames are used in the identification (Stanke et al. 2004), but more advanced approaches include also other features from a protein-coding gene, such as transcription factor binding motifs and CpG islands on the promoter region, poly-A tail motif at the 3'UTR and splice junction motifs in the case of eukaryotic organisms (Burge & Karlin 1997). As prokaryotes (e.g. bacteria) lack introns and their promoter motifs are more conserved than in eukaryotes, gene prediction is much more straightforward for prokaryotes.

Most gene prediction algorithms apply a Hidden Markov Model (*HMM*) to combine the many different types of signals (promoter motifs, poly-A tail motif, CpG islands etc.) into a single prediction whether a region could encode for a protein or not. As input, most tools only take the DNA sequence in which the gene is to be identified, but some also utilize the homology between the given sequence and known protein-coding

sequences in other organisms (Alexandersson et al. 2003). However, an algorithm that would use aligned RNA-seq reads in the standard SAM format and combine them with the identification of sequence-based features, does not yet exist.

2.3 RNA-seq data normalization

In RNA-seq, the processed alignment data is often expressed as read counts per transcript (transcriptome alignment, see Figure 2.5b) or read counts within bins throughout the whole genome (genome alignment, see Figure 2.5a). As the number of reads per sample and the length of transcripts used in the alignment process vary significantly, data normalization is required to make different samples and transcripts or genes comparable with each other. Regardless of the popularity of RNA-seq in cancer and genome-wide association studies, a standard method for read count data normalization still has not been established. In this subsection, three widely used approaches are introduced.

2.3.1 RPM normalization

The simplest way to account for biases that arise from different number of reads produced by each sequencing lane is to use *reads per million (RPM)* normalization, which is processed in a sample-wise manner. RPM normalized expression for region i in sample s , $N_{s,i}$, is calculated as

$$N_{s,i} = \frac{E_{s,i}}{R_s/10^6}, \quad (1)$$

where $E_{s,i}$ is the raw read count for region i in sample s and R_s is the total number of aligned reads in sample s . In other words, the RPM normalized expression describes the percentage of all reads that aligned within a certain region, multiplied by one million. As RPM normalized values are basically fractions, samples normalized this way result in the same scale and are hence comparable with each other.

2.3.2 RPKM normalization

Since the length of mRNA of different genes is not constant, it is expected that longer mRNAs would yield more sequenced reads than short mRNAs by random chance. Therefore the raw read counts of different transcripts within the same sample are not comparable. The *reads per kilobase of transcript length per million mapped reads (RPKM)* normalization method (Mortazavi et al. 2008) takes this bias into account by further dividing the normalized expression value presented in Equation 1 by transcript length T in kilobases:

$$N_{s,i} = \frac{E_{s,i}}{T \cdot R_s / 10^6}. \quad (2)$$

RPKM normalization assumes a linear dependency between mRNA length and the number of sequenced reads originating from the mRNA, although a recent study (Bullard et al. 2010) has shown that this assumption may not actually hold for longer transcripts.

2.3.3 Quantile normalization

When quantile normalization is applied to a sample set, each sample will result in an identical distribution of data values. This is achieved by performing the following operations for a data matrix:

1. Sort each column in the data matrix and store the ranks of the original values
2. Calculate the arithmetic mean for each row
3. Replace the greatest value in each column of the original data matrix with the greatest mean. Then replace the second greatest value with the second greatest mean etc. and continue this process until each data point in the original matrix has been replaced with a mean value.

As an example, a 5×3 matrix containing random integers between 0 and 10 is quantile-normalized:

$$\begin{bmatrix} 1 & 8 & 2 \\ 5 & 5 & 3 \\ 3 & 2 & 4 \\ 4 & 4 & 9 \\ 7 & 7 & 6 \end{bmatrix} \xrightarrow{\text{1. sort}} \begin{bmatrix} 1 & 2 & 2 \\ 3 & 4 & 3 \\ 4 & 5 & 4 \\ 5 & 7 & 6 \\ 7 & 8 & 9 \end{bmatrix} \xrightarrow{\text{2. mean}} \begin{bmatrix} 1.67 \\ 3.33 \\ 4.33 \\ 6 \\ 8 \end{bmatrix} \xrightarrow{\text{3. replace}} \begin{bmatrix} 1.67 & 8 & 1.67 \\ 6 & 4.33 & 3.33 \\ 3.33 & 1.67 & 4.33 \\ 4.33 & 3.33 & 8 \\ 8 & 6 & 6 \end{bmatrix}$$

Quantile normalization is a robust normalization method when applied on samples that originate from the same distribution and is hence appropriate for RNA-seq data normalization. Since the quantile-normalized expression values in each sample have identical distributions, sample-wise comparisons are valid. However, as quantile normalization does not take into account transcript lengths, a mixture of normalization methods (e.g. RPKM + quantile) is often useful when determining normalized gene expression values.

2.4 Statistical testing

Statistical testing is a widely used method to determine the significance of a certain statistic calculated from input data. In statistical (hypothesis) testing, two different hypotheses are formed: the null hypothesis H_0 and the alternative hypothesis H_1 . Commonly the null hypothesis states an equality or inequality between a certain parameter derived from two different distributions of the same type. For example, in cancer studies the two distributions could be formed from the gene expression values of a certain gene in two different classes: cancerous cells and healthy cells. Given e.g. the RNA-seq measure-

ments for each sample from these two classes, a statistical test can be established to decide whether or not the mean gene expression values of this gene could have originated from the same distribution:

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_1 : \mu_1 &\neq \mu_2 \end{aligned} \tag{3}$$

The null hypothesis in statistical testing often is that the two tested parameters are from the same distribution, i.e. there is no significant difference between them, and the alternative hypothesis is complementary to the null hypothesis. When the population distribution is known (or assumed to be known), a distribution-specific test statistic can be calculated for both classes. If the computed statistics are highly unlikely to have originated from the same distribution, the null hypothesis is rejected and therefore the alternative hypothesis can be accepted (e.g. in the example case, the mean gene expression values between healthy and cancerous cells differ significantly).

In statistical (hypothesis) testing, the *p-value* is a measure that describes the minimum risk involved in neglecting the null hypothesis (and hence accepting the alternative hypothesis), given the evidence based on the input data. When a p-value has been calculated, the null hypothesis is neglected if the p-value is below a certain threshold called *risk level*, α . In hypothesis testing, a risk level of $\alpha = 0.05$ is often set as the threshold for statistical significance. In biological and especially cancer-related studies, p-values are a crucial indicator of biological significance, when comparing e.g. the gene expression values of each gene between cancerous and healthy patients. Genes with a low p-value (i.e. the genes that differ most between cancerous and healthy patients by expression values) are regarded as candidates for having contributed to the cause of cancer, and they are hence further processed in the biological laboratory (*wetlab*).

In the following subsections, two of the most widely used statistical tests in cancer studies, *t*-test and *U*-test, are introduced and their underlying assumptions, p-value calculation and their validity in RNA-seq studies are discussed. In addition, as *t*-test is a special case of the general tests used for normal distributions (Z-tests), the basics of Z-tests are also covered in this work.

2.4.1 Z-test

Whenever the statistic to be tested follows at least approximately a normal distribution, a Z-test can be used to determine whether or not to neglect the null hypothesis. In a Z-test, the sample mean \bar{x} calculated from input data is compared to a given constant μ_0 (one-sample location test) or the mean of another sample (paired difference test), using a standardized statistic (*standard score*) z . For a one-sample case, the standard score is defined as

$$z = \frac{\bar{x} - \mu_0}{s}, \tag{4}$$

where s is the sample standard deviation, $s^2 = \sigma^2/n$ and n is the sample size. As the z statistic is standardized, it can be compared to the cumulative distribution function (*CDF*) of the normal distribution, Φ_N . The probability $p = \Phi_N(-|z|)$ represents the probability that a random sample taken from a standard normal distribution is as extreme as the calculated sample mean (in the same direction as the sample mean). This probability is often also denoted the *one-tailed p-value*. A *two-tailed p-value*, on the other hand, takes into account both tails of the distribution (positive and negative extremes) and is hence more appropriate when testing for inequality between two statistics. Since normal distribution is symmetric, a two-tailed p-value can be calculated using the one-tailed version:

$$p = 2 \cdot \Phi_N(-|z|). \quad (5)$$

Even though Z-test is the most fundamental statistical test for normal distributions, it has important features one must take into account before using the test. Obviously, the tested statistic must follow a normal distribution as discussed earlier, although according to the *central limit theorem*, the distribution of sample means of a sufficiently large number of random variable samples (with a well-defined mean and variance) will be approximately a normal distribution, even if the single random variables originate from a completely different distribution (see Figure 2.6 for an example). Secondly, the population variance has to be known or approximated as accurately as possible.

In RNA-seq data, neither of these assumptions necessarily holds, because a) RNA-seq based count data may not be normally distributed, as shown in Chapter 3, and b) the studied subset of (cancer) cells is often a very small proportion of the whole population. Due to various technical reasons, wetlab procedures etc., the studied cells are prone to systematical effects on either biological content or the measured RNA-seq data. Therefore it is dangerous to assume that the variance of any measured statistic corresponds to the variance of the whole population.

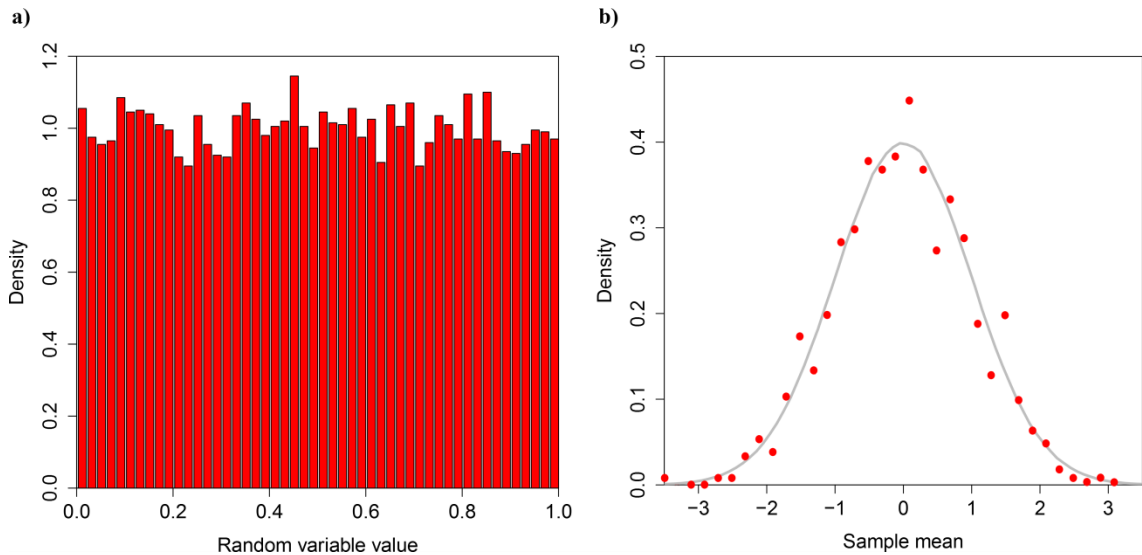


Figure 2.6. The histogram of a random sample of size 10000, drawn from a uniform distribution $U(0,1)$ (a), and the histogram of 1000 standardized sample means of a sample of size 1000 (drawn from a uniform distribution) as point plot (b) are shown. The grey curve represents the probability density function of a standard normal distribution, showing the validity of central limit theorem in this simulated data case.

2.4.2 t-test

When the data is normally distributed and the population variation is known, Z-test can be used to test the difference of means of two samples. However, as discussed in the previous subsection, the population variation is hardly ever known in a biological study, whether or not the data is produced with RNA sequencing. A solution to this problem is the Student's t -test, which assumes that the data is normally distributed but instead of assuming known variance, it models the uncertainty in the true population variance based on the calculated sample variance and the number of samples. The test statistic for a one-sample test is defined in a similar way as for Z:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad (6)$$

which follows a t distribution with $n - 1$ degrees of freedom. The uncertainty of the variance is stored in the degrees of freedom, as the variance follows a $\chi^2(n - 1)$ distribution (*chi-squared* distribution with $n - 1$ degrees of freedom). Unlike for the standardized Z statistic, there is a distinct t distribution for each sample size and as the sample size increases infinitely, t distribution approaches the standard normal distribution.

One of the most popular uses for t -test is to test the difference of two sample means (*two-sample t-test*). In the simplest two-sample t -test, the variances of the two populations where the (independent) samples were drawn from are assumed equal. The test statistic in a two-sample test is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{12} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (7)$$

where S_{12} is the pooled estimate of the common standard deviation:

$$S_{12} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \quad (8)$$

Since the null hypothesis in a two-sample t -test is often that the population means are equal ($\mu_1 = \mu_2$), Equation 7 simplifies into

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{12} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \quad (9)$$

The resulting test statistic follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

If the population variances of two different populations are not equal, they must be estimated separately instead of as a pooled estimate. In *Welch's t-test* this inequality is taken into account and the test statistic is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad (10)$$

As the variances are unequal and unknown, the degrees of freedom, df , cannot be calculated analytically. However, they can be approximated using the Welch-Satterthwaite equation:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2 / (n_2 - 1)} \quad (11)$$

Since t distribution is symmetric as well, the two-tailed p-value in a t -test can be calculated in a similar way to Equation 5, by replacing the normal distribution CDF with the CDF of a t distribution with the specific degrees of freedom.

In biological studies, t -test is by far the most widely used tool for statistical hypothesis testing. Since it allows the population variance to be unknown and is less stringent with the requirement for data normality, it is a better choice especially for small sample sizes than Z -test. Although RNA-seq count data may not be normally distributed as such, some processed statistics can still be approximately normally distributed and

hence a t -test is appropriate for hypothesis testing. The test statistics presented in this subsection all assumed that the samples are independent, but versions of t -test for dependent and paired samples also exist. Additionally, t -tests modified especially for biological data and for very small sample sizes have been developed (Smyth 2004).

2.4.3 U-test

When the distribution of the population is unknown or a mixture of several known distributions (not necessarily normal distributions), t -test may be inaccurate. In these cases a non-parametric test such as the Mann-Whitney U -test (also denoted *Wilcoxon rank-sum test*) is more suitable for hypothesis testing. U -test requires that 1) the two populations have similar, continuous distributions and b) the values in the distribution are ordinal, i.e. their greatness can be compared. In U -test, the null hypothesis states that the medians of two populations are equal:

$$\begin{aligned} H_0 : \tilde{\mu}_1 &= \tilde{\mu}_2 \\ H_1 : \tilde{\mu}_1 &\neq \tilde{\mu}_2 \end{aligned} \quad (12)$$

To calculate the test statistic, u , the values in each sample are combined and sorted in an ascending order into a single sequence. The ranks of each value in the sequence are then summed over the samples 1 and 2, forming the rank sums r_1 and r_2 , respectively. If there are any tied values, each of them will be assigned the same rank (the arithmetic mean of the original ranks of the tied values). Using the rank sums, the test statistic u is defined as

$$u = \min\left(r_1 - \frac{n_1(n_1 + 1)}{2}, r_2 - \frac{n_2(n_2 + 1)}{2}\right), \quad (13)$$

where n_1 and n_2 are the sizes of sample 1 and 2, respectively. Calculating a p-value for u is complicated for small sample sizes and it is often done by using tabulated values. For large sample sizes ($n_1, n_2 > 20$), u is approximately normally distributed based on the central limit theorem, with an expected value defined as

$$\mu_U = \frac{n_1 n_2}{2} \quad (14)$$

and a standard deviation defined as

$$\sigma_U = \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}} \cdot \sqrt{\frac{(n_1 + n_2)^3 - n_1 - n_2}{12} - \sum_{i=1}^k \frac{t_i^3 - t_i}{12}}, \quad (15)$$

where the last summation term takes into account all k tied ranks at the locations denoted by t_i . This correction is required, since by assigning all tied ranks a single rank value, the standard deviation of ranks becomes smaller than without tied ranks. When there are no tied ranks and the summation term equals zero, Equation 15 can be simplified into a much more convenient form:

$$\begin{aligned}
 \sigma_U &= \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}} \cdot \sqrt{\frac{(n_1 + n_2)^3 - n_1 - n_2}{12}} \\
 &= \sqrt{\frac{n_1 n_2 \frac{((n_1 + n_2)^3 - n_1 - n_2)(n_1 + n_2 + 1)}{12(n_1 + n_2 - 1)(n_1 + n_2)(n_1 + n_2 + 1)}}{12}} \\
 &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \cdot \frac{n_1^3 + 3n_1^2 n_2 + 3n_1 n_2^2 + n_2^3 - n_1 - n_2}{n_1^3 + 3n_1^2 n_2 + 3n_1 n_2^2 + n_2^3 - n_1 - n_2}} \\
 &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}
 \end{aligned} \tag{16}$$

Since RNA-seq data (or any other gene expression measurement data) from different patients usually originate from very similar (non-normal) distributions, U -test is often a safe choice for statistical testing. Even if the data is normally distributed, the asymptotical efficiency of U -test compared to t -test is $3/\pi > 0.95$ (Lehmann 1999, p. 176), highlighting its priority as the standard statistical hypothesis testing method for any RNA-seq study.

2.5 Assessing data similarity and normality

As many of the statistical hypothesis tests are based on the normal distribution or the tested datasets originating from the same distribution, and since the normality of RNA-seq data usually does not hold (Marioni et al. 2008), two mathematical methods to assess the normality or similarity of given data are covered in the following subsections: *Kolmogorov-Smirnov test* and *Shapiro-Wilk test*. In addition, two graphical methods to compare any two distributions with each other, quantile-quantile plot ($Q-Q$ -plot) and *boxplot*, are introduced.

2.5.1 Graphical methods

In a quantile-quantile plot, the quantiles of the data from both of the two different distributions or empirical samples are first divided into a set of intervals and then plotted with the first distribution on the x-axis and the other one on the y-axis. For example, a ten-point quantile-quantile plot would consist of each decile (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1) of distribution 1 plotted against the corresponding decile from distri-

bution 2. Each point in the plot is drawn at a threshold value below which one tenth, two tenths, three tenths etc. of the values in each distribution are located.

When two identical distributions are compared in a Q-Q-plot, the points drawn based on the quantiles form a line $y = x$. If the distributions are similar but one of them has been shifted by a constant value, b , a linear dependency is shown again but this time in the form of $y = x + b$. On the other hand, if the other distribution is multiplied by n , the dependency gets the form $y = nx$. If the distributions are different, no linear dependency is seen in the Q-Q-plot. In Figure 2.7, different kinds of Q-Q-plots are shown for data from similar and dissimilar distributions.

Another popular way to describe one or more distributions graphically in statistics is the boxplot. In a boxplot (or a *box and whiskers plot*), the distribution of each sample is illustrated with a box, the height and position of which is determined by the first and third quartiles (0.25 and 0.75) of the data. The second quartile, or the median, is drawn as a horizontal line within the box. In addition, two “whiskers” are connected to the box, the end points of which are not standard. Commonly used values for the whisker end points include the extreme values (maximum and minimum, or percentiles 1.00 and 0.00) and the percentiles 0.91 and 0.09. Values that are not included within the box and whiskers (outliers) are drawn as dots in the plot. An example of a boxplot drawn with the `boxplot` function (using default parameters) in R (R: A language and environment for statistical computing 2013) is shown in in Chapter 4.

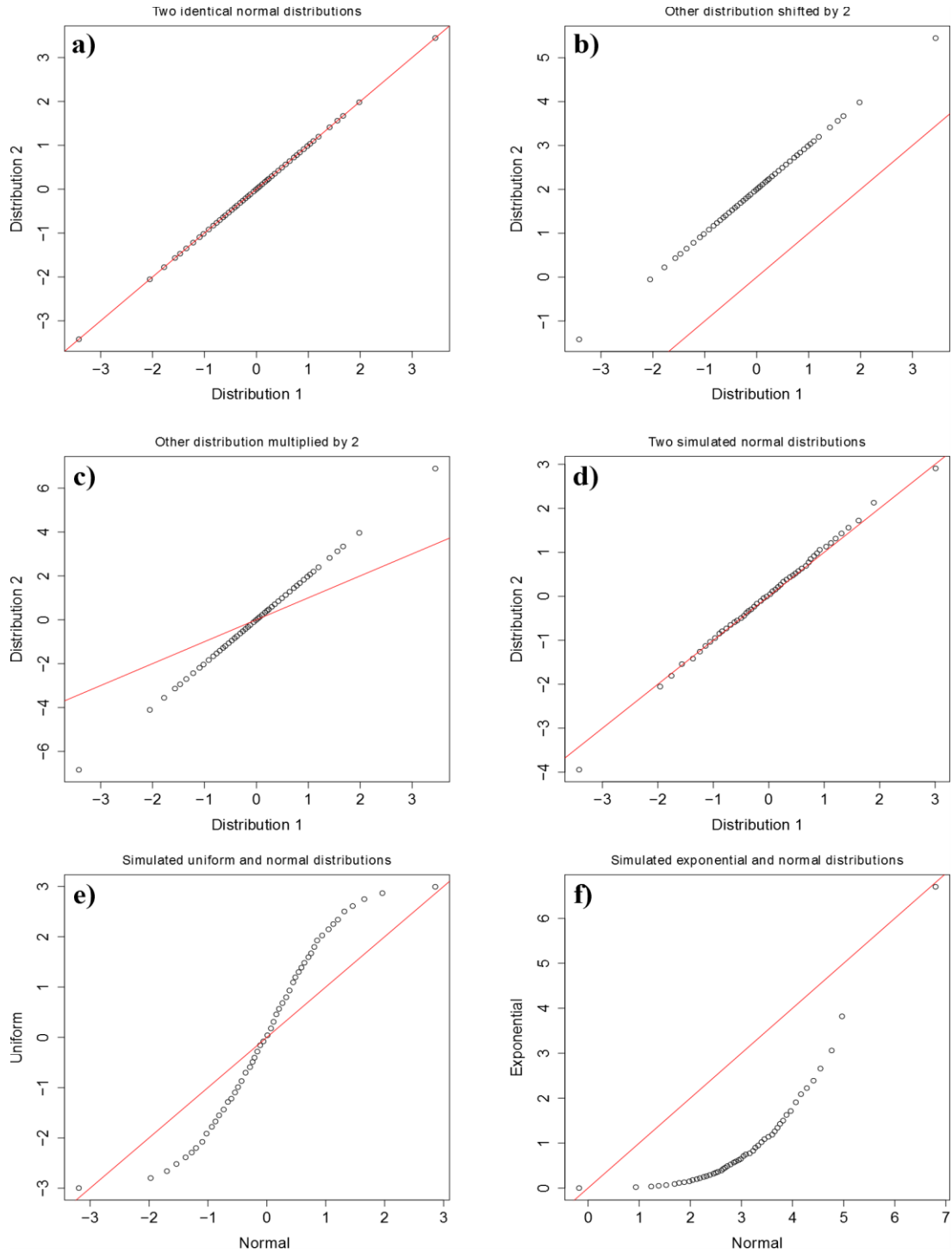


Figure 2.7. Six different Q - Q -plots, each with the quantiles 0.00, 0.02, 0.04, ..., 0.98, 1.00 and the red line ($y = x$) representing an ideal match between the two distributions. In the first case (a), the quantiles of two identical distributions fall on the line $x = y$, while in the second (b) and third (c) cases the other distribution is shifted or multiplied by 2, hence changing the line on which the quantiles are matched. When the samples originate from the same distribution but are not identical (d), the match on the line $x = y$ is no longer perfect especially within the first and last quantiles. Finally, when two completely different distributions are compared in a Q - Q -plot (e, f), no linear dependency is seen.

2.5.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test, named after Andrey Kolmogorov and Nikolai Smirnov, is a nonparametric test to compare a given sample distribution to either a known reference distribution (*one-sample K-S test*) or to another sample distribution (*two-sample K-S test*). In a K-S test, the null hypothesis states that two cumulative distribution functions are equal:

$$\begin{aligned} H_0 : \forall x: F(x) &= F_0(x) \\ H_1 : \exists x: F(x) &\neq F_0(x), \end{aligned} \quad (17)$$

where $F(x)$ is the cumulative distribution function of the sample and $F_0(x)$ either the reference or the CDF of another sample. However, since the true CDF of the original population distribution for the sample is not usually known, it is approximated by an *empirical cumulative distribution function* (eCDF), $F_e(x)$.

In a one-sided, one-sample K-S test, the test statistic is defined by the following two equations (Ruohonen 2002):

$$D_n^+ = \sqrt{n} \sup_x (F_e(x) - F(x)) \quad \text{and} \quad (18)$$

$$D_n^- = \sqrt{n} \sup_x (F(x) - F_e(x)), \quad (19)$$

where D_n^+ is the statistic for the eCDF of the studied distribution being *greater* than the reference CDF, D_n^- for being *smaller* than the reference CDF and n is the sample size. A two-sided, one-sample test statistic can be formed using Equations 18 and 19 simply by choosing the maximum value of the one-sided test statistics: $D_n = \max(D_n^+, D_n^-)$.

The p-value for a one-sample, one-sided K-S test can be calculated using the cumulative distribution function defined by the Birnbaum-Tingey equation (Birnbaum & Tingey 1951):

$$P\left(D_n^+ \leq \frac{x}{\sqrt{n}}\right) = P\left(D_n^- \leq \frac{x}{\sqrt{n}}\right) = \frac{x}{n^n} \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} (k-x)^k (x+n-k)^{n-k-1}. \quad (20)$$

The CDF defined in Equation 20 applies only to the one-sided, one-sample test statistics. To calculate a p-value for the two-sided test, an alternative definition for the CDF of the test statistic can be used (Marsaglia et al. 2003):

$$P\left(D_n \leq \frac{x}{\sqrt{n}}\right) = P(\sqrt{n}D_n \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)}, \quad (21)$$

which yields asymptotically exact p-values for the two-sided, one-sample test statistic.

The K-S test statistic, D_{n_1, n_2} , for testing two eCDFs, $F_{e,1}(x)$ and $F_{e,2}(x)$, with sample sizes n_1 and n_2 , respectively, is defined by

$$D_{n_1, n_2} = \sup_x |F_{e,1}(x) - F_{e,2}(x)|, \quad (22)$$

where the notation $F_{e,i}(x)$ represents the fraction of values in sample i that are smaller than x . In other words, the test statistic D_{n_1, n_2} describes the maximum distance between two CDFs, one or both of which may be empirically derived. Equation 22 represents the two-sided test statistic, i.e. it measures the absolute value of the differences between the two CDFs. For the two-sample K-S test, tabulated values are often used when determining the p-value.

As Equation 20 may contain a computationally challenging component for large sample sizes (the binomial coefficient), Equation 21 is inaccurate for small sample sizes and tabulated p-values do not exist for every sample size and risk level, alternative ways to compute p-values may prove useful. One option is to simulate the hypothesized distribution with a permutation test, i.e. by generating a sufficiently large number of samples (>10000) of size n from the assumed distribution and computing $D_{n,i}$ for each generated sample i . By computing the histogram (the shape of which corresponds to the probability density function) for the distribution of $D_{n,i}$ and comparing the location of the original D_n in the histogram, the corresponding probability can be considered as $(1 - \text{pvalue})$ of the hypothesis test.

Similarly as for the one-sample test, a fairly reliable p-value can be estimated for the two-sample K-S test by using a categorical permutation test to estimate the null distribution of D_{n_1, n_2} . As the underlying distributions of the populations may be of unknown form, the permutation cannot be performed by generating another set of samples as for the one-sample K-S test, but instead the permutation is performed by first pooling each value in samples 1 and 2 together, and then dividing the pooled values randomly to new samples with the same sizes as the original samples. For each permuted sample 1 vs. sample 2 comparison and for the original comparison, the statistic D_{i, n_1, n_2} is calculated using Equation 22 and a histogram of the distribution of D_{i, n_1, n_2} is formed. Finally, the p-value can be estimated similarly as for the one-sample test by comparing the original test statistic to the histogram values.

In Table 2.1, three different samples from the standard normal distribution ($X \sim N(0,1)$) and a uniform distribution ($X \sim U(-3,3)$) are simulated. For this data, a two-sample K-S test was used to determine whether or not a) samples 1 and 2 originate from the same distribution and b) samples 2 and 3 originate from the same distribution. In addition, a one-sample K-S test was used to test if c) sample 1 was drawn from a standard normal distribution and if d) sample 1 was drawn from a uniform distribution, $U(-3,3)$. The p-values are calculated using the permutation approach described above, and as a reference, the function `ks.test` implemented in R is used, which calculates the K-S test p-values using an asymptotic distribution for the test statistic.

Table 2.1. *Samples generated for the permutation tests of the K-S statistic.*

	n	Source	Data					
Sam1	12	U(-3,3)	1.411	1.166	2.452	1.520	-2.666	-2.223
			-2.547	-2.468	-1.829	2.486	-1.457	1.122
Sam2	8	N(0,1)	1.074	-1.204	-0.081	1.328	-1.229	0.807
			-0.041	1.393				
Sam3	10	N(0,1)	-0.693	1.493	-0.881	0.802	1.925	-0.424
			0.173	-0.248	-0.389	-0.243		

The resulting test statistics and corresponding p-values for the comparisons described above are presented in Table 2.2 and an illustration of how they were derived is shown in Figure 2.8. The p-values of `ks.test` in the two-sample test correspond to the two-sample, two-tailed p-values, and in the one-sample test they correspond to the one-tailed ('greater') p-value. As shown in Table 2.2., the p-values calculated with the proposed permutation method are very similar to the results of `ks.test` apart from the test 'Sam2 vs. Sam3'. The discrepancy of these two p-values may have been caused by the inaccuracy related to the asymptotical Kolmogorov distribution in `ks.test`, since the joint sample size is small ($n_1 + n_2 = 18$). However, when considering e.g. a risk level of $\alpha = 0.05$ as the threshold for neglecting the null hypothesis, both methods agree on each of the four tests: only when comparing the simulated data $X \sim U(-3,3)$ to the standard normal distribution, the alternative hypothesis ($X \not\sim N(0,1)$) has strong enough evidence.

Table 2.2. *P-values for four different K-S tests calculated with two different methods.*

Test	Permutation	ks.test
Sam1 vs. Sam2	0.1432	0.1496
Sam2 vs. Sam3	0.1624	0.5423
Sam1 vs. N(0,1)	0.0087	0.0083
Sam1 vs. U(-3,3)	0.2134	0.2078

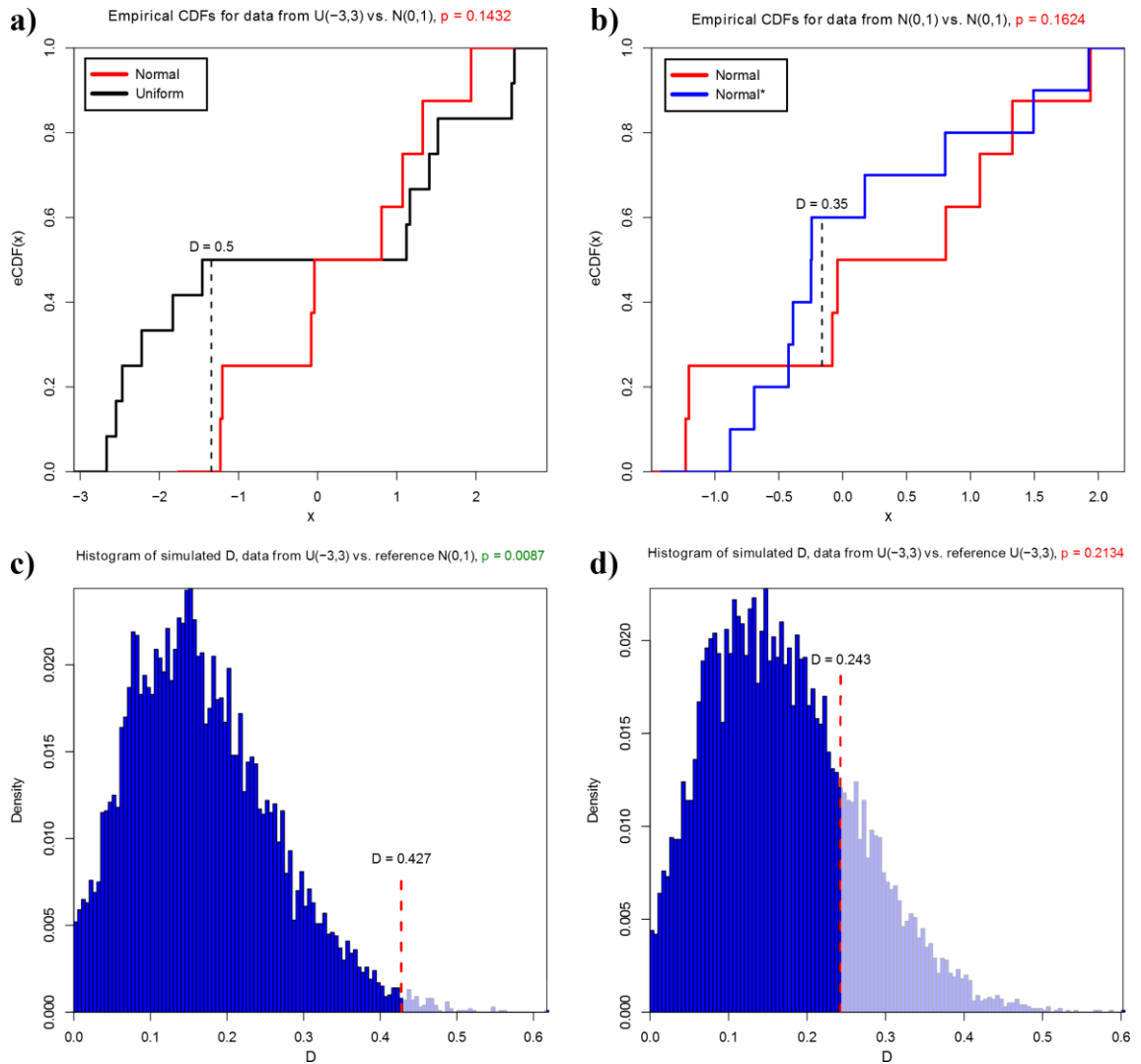


Figure 2.8. An illustration of how the test statistic is determined in a two-sample comparison (a, b) and how the p-value in a one-sample K-S test can be estimated from the simulated distribution of D , when data is sampled from the null-hypothesized distribution (c, d).

One-sample K-S test can be used to determine whether a given sample could have originated from any given distribution, hence providing a useful tool to test for normality e.g. in an RNA-seq study. On the other hand, two-sample K-S tests can be applied to study the similarity of two samples, which is a useful feature for assuring that e.g. the use of U -test is valid given the data.

2.5.3 Shapiro-Wilk test

While the K-S test can be used also to test for data normality, several methods have been developed that are specifically tailored for normality testing and are thus more powerful for that purpose. One of the most popular methods of these is the Shapiro-Wilk test (S - W test) (Shapiro & Wilk 1965), which tests the following hypothesis:

$$\begin{aligned} H_0 &: \exists \mu, \sigma \in \mathbb{R}: X \sim N(\mu, \sigma) \\ H_1 &: \forall \mu, \sigma \in \mathbb{R}: X \not\sim N(\mu, \sigma), \end{aligned} \quad (23)$$

where X is the random variable representing the population whose distribution is to be tested.

To understand how Shapiro-Wilk test is performed and how the test statistic, W , is formed, a few other operations should be carried out. Firstly, the sample $x_i, i = 1, \dots, n$ is sorted into an ascending order, i.e. $x_1 \leq \dots \leq x_n$. Secondly, the expected value of the order statistic of independent and identically distributed (*i.i.d.*) random variables sampled from a standard normal distribution, $\mathbf{m} = [m_1 \ m_2 \ \dots \ m_n]'$, and the covariance matrix V for the sampled order statistics, are calculated. The test statistic W is then defined by

$$W = \frac{(\sum_{i=1}^n a_i x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (24)$$

where \bar{x} is the sample mean and the constant vector $\mathbf{a} = [a_1 \ a_2 \ \dots \ a_n]$ is given by

$$\mathbf{a} = \frac{\mathbf{m}'V}{\sqrt{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}}}. \quad (25)$$

The distribution of W is analytically defined only for the sample size $n = 3$ and thus, for other sample sizes, approximations must be used. Currently the broadest approximation of Shapiro-Wilk test statistic, developed by Rahman & Govindarajulu (1997), is applicable for sample sizes of $3 \leq n \leq 5000$. Since the distribution of W often has to be simulated, an exact p-value cannot be determined (apart from $n = 3$), but a permutation approach as the one discussed in the previous subsection can be used to estimate it.

As the Shapiro-Wilk test statistic takes into account both the mean and variance of the distribution instead of a single and simple metric (such as in the K-S test), it provides more reliable evidence for (or against) the normality of given data. Regardless of its power in normality testing, the limited sample size is an important downside for S-W test, which may be critical in an RNA-seq study where the input data can consist of millions of data points. However, as pointed out in Chapter 3 of this work, the data to be tested in an RNA-seq study does not necessarily consist of the whole input data, but just a small subset of it.

2.6 Survival analysis

A survival analysis, in general, focuses on how many individuals of a population undergo a certain event (death, recurrence of a cancer, mechanical failure etc.) during a defined time frame. In biomedical studies, it is a common way to study the effects of a certain drug in disease or cancer treatment or how the expression of a certain gene marker correlates with patient survival or cancer recurrence. A typical feature of surviv-

al analysis is that some of the individuals under inspection may not remain till the end of the study, which is also known as *censoring*. For censored cases, the last known information (of the patient being alive, machine part being unharmed etc.) is utilized when determining its effect on the survival of the population. If the censored cases are always considered as losses, the survival estimation of the population becomes too pessimistic and if they are always assumed to survive, the estimation becomes too optimistic. However, completely ignoring censored patients is not desirable either, since significant information is lost in that case.

In this section, two fundamental components of a survival analysis that are able to deal with censored data - Kaplan-Meier estimate and logrank test - are introduced. As the focus of this work is on analyzing RNA-seq data from a cancer study, the survival analysis will be discussed from that point of view.

2.6.1 Kaplan-Meier estimate and plot

The Kaplan-Meier estimate (Kaplan & Meier 1958) developed by Edward Kaplan and Paul Meier, is the maximum likelihood estimate of the survival function, $S(t)$, used in a survival analysis. In a cancer study, the survival function could e.g. describe the probability that a patient of a certain population is still alive t days after cancer diagnosis. Assuming that censoring occurs independently of the group and that each patient recruited in the study has an equal survival probability at the time of diagnosis, the Kaplan-Meier estimate of surviving past time t is defined by

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}, \quad (26)$$

where t_0 is the time of diagnosis, n_i is the number of uncensored, alive patients (“patients at risk”) at time t_i and d_i is the number of deaths at time t_i . The index i covers each time point before t when a patient dies or becomes censored.

The most important application of the Kaplan-Meier estimate is a graphical illustration denoted the Kaplan-Meier plot (*K-M plot*). In a K-M plot, a curve with the K-M estimate of the survival function is drawn for each population (e.g. patients treated with drug A, patients treated with drug B, patients without drug treatment etc.) with horizontal steps between each time point when death occurs, followed by a vertical drop to the K-M estimate value at the next event. As an example, artificial survival data has been generated for three classes of patients (with 15 patients in each class) that suffer from a deadly disease:

- Patients treated with a good drug,
- patients treated with a weak drug and
- patients with no drug treatment.

The generated times of death, the last known times to be alive for censored patients and the corresponding K-M estimates are shown in Table 2.3 and a K-M plot for the three different patient categories is shown in Figure 2.9. The time point $t = 0$ equals to the day when the disease was diagnosed and drug treatment given (if it was given at all).

Table 2.3. K-M estimates for simulated data of patients with three different treatment strategies. The event times (in days) with red font represent a loss to censoring, other values are times of death for a certain patient.

Strong drug		Weak drug		No drug	
Time	KM	Time	KM	Time	KM
3	0.93	28	1.00	11	0.93
44	0.93	33	0.93	16	0.87
52	0.86	45	0.93	31	0.80
63	0.86	45	0.85	35	0.73
84	0.78	46	0.77	37	0.67
88	0.78	47	0.70	41	0.60
93	0.70	49	0.70	53	0.53
95	0.61	54	0.61	56	0.47
107	0.52	55	0.61	61	0.40
115	0.44	58	0.51	62	0.33
120	0.44	62	0.41	65	0.33
123	0.44	84	0.30	66	0.25
124	0.29	86	0.20	66	0.17
139	0.29	93	0.10	73	0.08
166	0.00	121	0.00	75	0.08

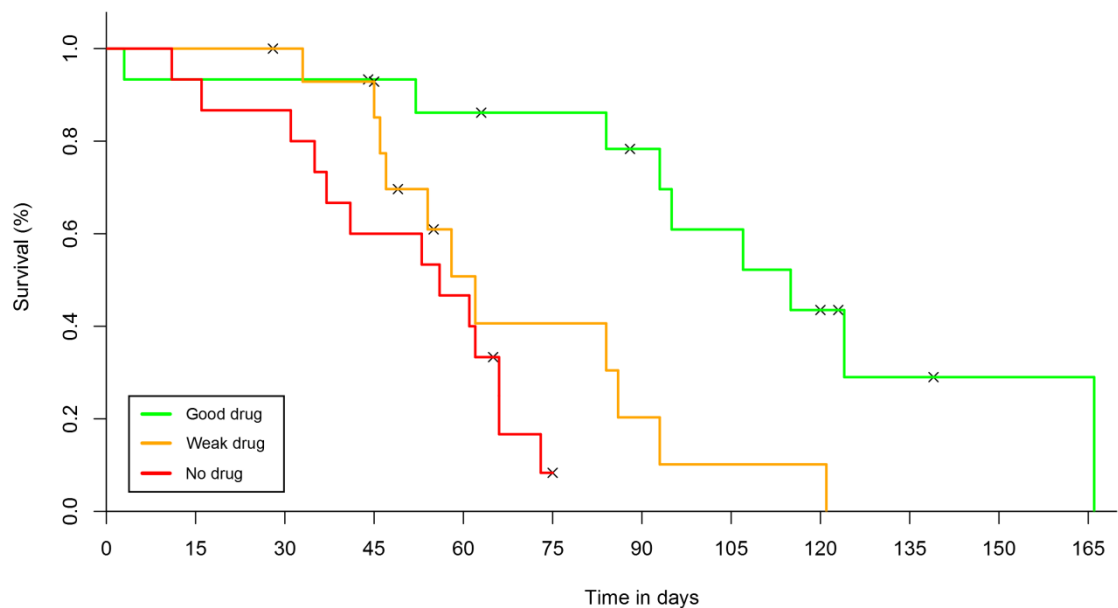


Figure 2.9. Kaplan-Meier estimates for the three different patient groups. Each 'x' in this plot represents a patient becoming censored.

In this example, a clear difference in survival between patients treated with a good drug vs. patients without drug treatment can be seen, which is obviously the outcome of any successful drug development study. Based on the K-M estimates shown in Table 2.3, the estimated probability that a patient of a certain population is alive after 60 days of diagnosis is $p_{good\ drug} = 0.86$, $p_{weak\ drug} = 0.51$ and $p_{no\ drug} = 0.47$, respectively. At this time point, the difference between the weak drug and no treatment is not yet very significant, although at day 80 the difference is noticeable. Since the last surviving patient from the ‘No drug’ group becomes censored at day 75, the K-M estimate never becomes zero for this class. When inspecting the survival beyond day 175, e.g., this would lead to the contradictory result that having no treatment is a better way to survive than taking any drug, since $p_{good\ drug} = 0$ and $p_{weak\ drug} = 0$ but $p_{no\ drug} = 0.08$. This is most likely a false conclusion and points out the noisiness of the right tail of the survival curve, which thus should not be used as significant evidence for differences in survival between two or more groups.

K-M plots are a very common sight in any cancer or medical investigation publication, where two or more types of groups are involved in the study. In this work, the novel transcripts found with the analysis pipeline, Novellette, are also inspected from the survival analysis point of view. Although survival analysis is not the main focus in the development of Novellette, it may give important insight into finding transcripts that may have a significant role in the survival of the patient, and is hence also covered in this work.

2.6.2 Logrank test

To draw conclusions on the statistical significance of the difference between estimated survival functions, a specific test that would take into account censoring in the survival distributions should be used. One of the most widely used tests in this kind of analysis is the *logrank test*, which was originally proposed by Nathan Mantel (Mantel 1966). Under the null hypothesis of a logrank test (with the same assumptions as in the K-M estimate), the *hazard functions* of n different populations are equal:

$$\begin{aligned} H_0 : H_1(t) &= \dots = H_n(t) \\ H_1 : \exists i, j \in \{1, \dots, n\} : H_i(t) &\neq H_j(t), \end{aligned} \quad (27)$$

where $H_i(t)$ are the hazard functions of populations $i = 1, \dots, n$, respectively, and t denotes time. The hazard function, sometimes also denoted the hazard *rate* function, $H(t)$, is defined by

$$H(t) = \frac{f(t)}{S(t)}, \quad (28)$$

where $S(t)$ is the survival function and $f(t)$ is the hazard *density* function, which can be modeled with the probability density function of either exponential distribution or the Weibull distribution.

As the true population hazard functions are again impossible to determine from discrete and finite data, they must be estimated. One way to construct a hazard function estimate is to first combine the data from different groups into a single group (this is a valid operation since, according to the null hypothesis, the survival distributions are same) and then calculate the risk of death for the joint population at each time point, when a death has occurred in any of the original groups. The risk is simply calculated by dividing the number of deaths at a given time point by the number of patients at risk (alive and not yet censored). The expected number of deaths, $E_{j,i}$, is then calculated in each of the original groups at each time point i by multiplying the risk of death with the number of patients at risk in group j . Similarly, the actual observed numbers of death, $O_{i,j}$, for each group and time point are calculated from the survival times. According to Bland & Altman (2004), the test statistic L is then defined by

$$L = \sum_{j=1}^n \frac{\left(\sum_{i=1}^{k_j} O_{j,i} - \sum_{i=1}^{k_j} E_{j,i} \right)^2}{\sum_{i=1}^{k_j} E_{j,i}}, \quad (29)$$

where n is the number of groups to be tested and k_j is the number of patients in group j . Equation 29 is not the only way to define the test statistic, but it is straightforward and easy to interpret. This test statistic follows a χ^2 distribution with $n - 1$ degrees of freedom and hence the p-value can be obtained directly from one minus the cumulative distribution function of the χ^2 distribution, but similarly as for other tests discussed in this work, the p-value can also be estimated with a permutation test.

As an example, logrank test is applied to the simulated data shown in Table 2.3 with four different comparisons: all three groups compared together and three different pairwise comparisons. The p-values of these comparisons calculated with two different methods are summarized in Table 2.4. The p-values are calculated using a) Equation 29 and χ^2 distribution and b) the `survdifff`-function in R. As shown in Table 2.4, the p-values imply significant difference in survival in every comparison apart from weak drug vs. no drug, which also agrees well with the survival curves shown in Figure 2.9.

Table 2.4. Logrank test p-values with two different methods.

Test	Good drug vs. no drug	Good drug vs. weak drug	Weak drug vs. no drug	All vs. All
Eq. 29 + χ^2	7.04e-4	5.98e-3	0.181	1.95e-4
survdifff	1.33e-4	3.10e-3	0.142	1.52e-4

The logrank test has been shown to be a permissive test: it may yield a low p-value even if the data is very inaccurate and therefore alternative and more stringent tests for sur-

vival analysis have been developed (Berty et al. 2010). Logrank test is still the most popular hypothesis test in survival analysis, since it is simple, takes the censoring into account and does not make any assumptions of the underlying survival distributions of the populations. If there is no censoring, however, any other distribution independent and potentially asymptotically more powerful tests (such as U-test) can be used.

3 METHODS

In this chapter, the RNA-sequencing data analysis pipeline developed in this work (Novellette) is introduced and its components along with theoretical rationalization for the methodological choices are presented. To support the theoretical arguments, publicly available human *glioblastoma multiforme* (GBM) RNA-seq data from 169 samples in The Cancer Genome Atlas (TCGA) glioblastoma project (Brennan et al. 2013) is viewed.

3.1 Motivation

As the cost and run-time of whole-genome DNA sequencing and whole-transcriptome RNA sequencing have dropped drastically within the last few years, RNA-seq has become the standard tool to quantitatively and accurately measure the expression levels of both coding and non-coding RNA content in the cell. This has given rise to the development of computationally efficient data processing algorithms (especially for read alignment (Kim et al. 2013; Dobin et al. 2013)) and in storing the increasing amounts of raw and alignment data. The state-of-the-art solution to standardizing data storage and alignment data procession is the *SAM* format (*Sequence Alignment Map*) and the related *SAMtools* package (Li et al. 2009), which contains several command-line tools for binarizing, indexing and fetching of information from aligned read data.

As one of the major downstream analysis options in an RNA-seq study is to detect novel transcripts and assess their protein coding potential, a computationally efficient and SAM format supporting tool to perform this kind of data analysis would be feasible. However, most of the current publicly available tools either use unnecessarily complicated mathematical models, making the processing slow and the output and possible errors harder to interpret, or require very specific configurations and setups on the computer, making the tool difficult or even impossible to install on most platforms. In addition, a tool that would utilize splice-junction aligned reads (in the standard SAM format) in determining exon structures of the novel transcripts in the gene identification analysis, does not yet exist.

3.2 Overview

Novellette is an RNA-sequencing pipeline developed in this work, which takes splice-junction aligned reads in the binary SAM format (BAM) as input, detects novel transcripts that are differentially expressed between two different classes (e.g. cancerous samples vs. healthy samples) or, in a one-class analysis (e.g. only cancerous samples),

detects novel transcripts that are overexpressed in a subset of the samples, and finally performs a full gene identification analysis to all of the detected transcripts. The core of Novellette is implemented in C++, but some of the subtasks (e.g. statistical testing) is performed with R, and the SAMtools package is also utilized for efficient alignment data processing. Apart from SAMtools and R, which are both very standard bioinformatics tools, Novellette does not have any requirements or limitations for installation and use.

In Figure 3.1, an overview of the data analysis steps and different paths in the Novellette pipeline is shown as a chart. The pipeline consists of two fundamental parts: 1) detection of novel transcript candidates and 2) gene structure identification and scoring of the candidates. These two parts and the respective data analysis steps are covered in the following two sections.

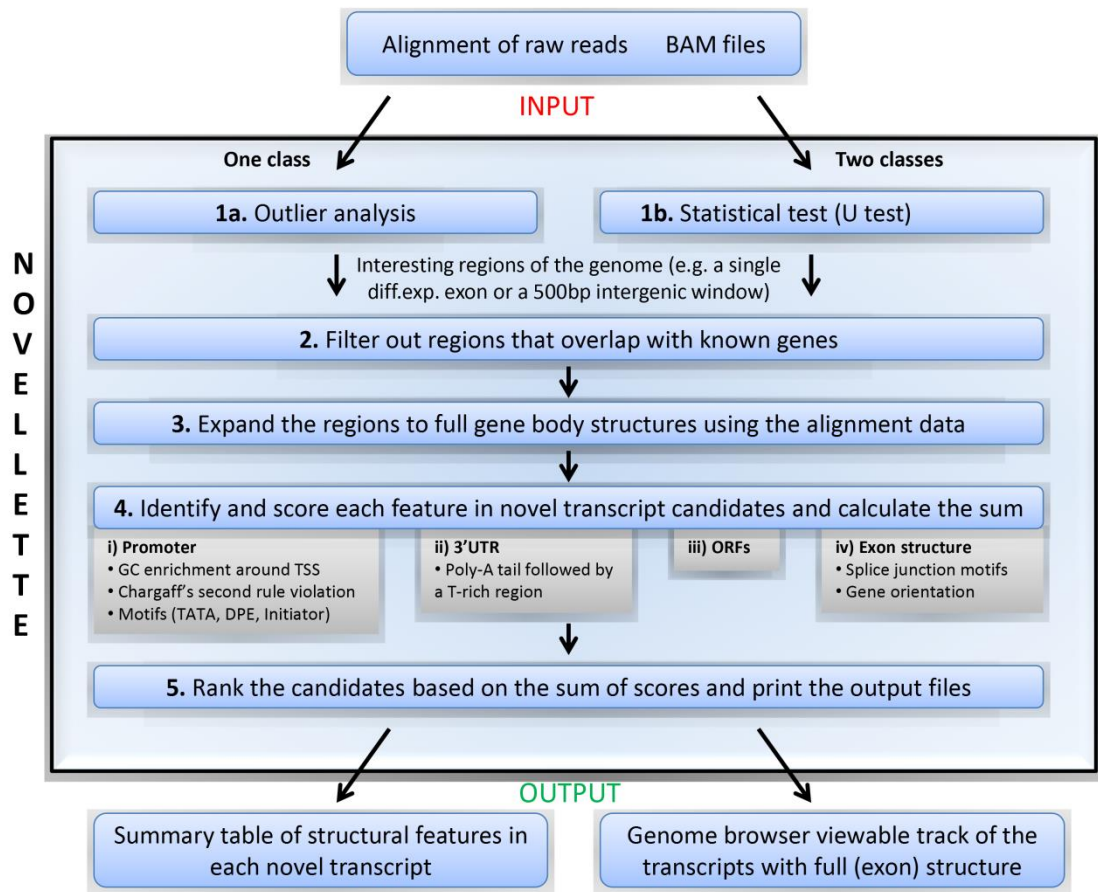


Figure 3.1. An overview of the workflow of Novellette.

3.3 Detection of novel transcript candidates

The analysis in Novellette starts by reading in any number of BAM files, each of which corresponds to the reads of one sample aligned against a reference genome with a splice-junction mapper. The goal of the first analysis part is to calculate the normalized read counts for each sample in windows with pre-defined width across the whole ge-

nome, and then systematically define which of these windows could potentially be a novel transcript or a part of it. The way this is done depends on whether the data consists only of cancerous samples or if reference (healthy) samples are available as well. The following subsections cover the relevant data analysis steps and related theory in forming the putative novel transcript candidates.

3.3.1 Calculating normalized read counts

The first component of Novellette takes any number of BAM or SAM files and a user-defined window length (default: 500 bp) as input. It then calculates the RPM value (Equation 1) in a sliding window throughout the whole genome (one chromosome at a time) for each sample based on the aligned reads in the BAM/SAM file, by sliding the window with half of its length in each step (e.g. 0 – 499, 250 – 749, 500 – 999, 750 – 1249, ...). The RPM value calculated this way corresponds to the read *coverages* in the specified sliding windows in the genome. The resulting RPM values for each processed sample are then stored in a single matrix (*coverage matrix*) in the *.igv* format, which can further be processed into a more convenient and storage-friendly binary file format (*.tdf*) using IGVtools, and viewed with the IGV genome browser (Thorvaldsdottir et al. 2013).

RPM normalization method was chosen at this step of the analysis since 1) the window length is constant and thus no length normalization (e.g. RPKM, Equation 2) is required and 2) the read counts are processed one sample at a time, making sample-wise adjustments (e.g. quantile normalization) impossible. Nevertheless, RPM normalized read counts are comparable each other since the most important source of bias, i.e. the different number of mappable reads in different samples, is taken into account.

3.3.2 Two-class analysis

Most cancer studies focus in finding differences e.g. between cancer cells and adjacent healthy cells, between aggressive/invasive/metastatic cancer cells and less-aggressive cancer cells, or between cells from two different cancers or cancer subtypes. In an RNA-seq study, this usually means the detection of differentially expressed genes, e.g. finding genes that have statistically significantly higher or lower expression in one class versus the other class. In this work, the comparison is performed for each of the sliding windows for which the RPM value is calculated.

Regardless of the assumptions discussed in Section 2.4.2 that do not necessarily hold for RNA-seq data, *t*-test is the most widely used method in any kind of differential expression analysis. To investigate the normality of RPM values calculated by Novellette, a coverage matrix is calculated for the 169 TCGA GBM samples using a 500 bp sliding window. Although no two-class comparison can be performed with this example data, the normality should still hold also for this data as the samples are all from the same population (assuming that *t*-test would generally be valid in an RNA-seq study). As the comparison in Novellette is performed for the window RPM values between dif-

ferent samples, the data vector consists of 169 values. Since the window length used in this analysis is 500 bp, the sliding step is 250 bp and since the length of the whole human genome is approximately 3.14 billion bp long, the number of data vectors is $3.14 \cdot 10^9 / 250 > 12 \cdot 10^6$.

To assess the normality of this many samples simultaneously, K-S and S-W tests are performed for the first 10^6 windows covering the genome and the corresponding p-values for the test statistics are compared to three different risk levels. The p-values are calculated using the functions `ks.test` and `shapiro.test` in R, both of which assume that the given data is normally distributed as the null hypothesis. Therefore a smaller risk level α means a less stringent required evidence for non-normality, and thus more false positive conclusions of the data being normally distributed. As shown in Table 3.1, the RNA-seq data in this work shows poor evidence for normality based on both of the normality tests at the defined risk levels, when compared to simulated data of similar size drawn from a standard normal distribution. In other words, the use of *t*-test is not valid with this data and hence *U*-test will be the primary choice for statistical testing of differential expression in Novellette.

Table 3.1. *The percentages of how many tests supported the hypothesis of the window RPM values in this RNA-seq data set being normally distributed.*

α	K-S	S-W	K-S sim	S-W sim
0.001	18.293 %	15.794 %	99.929 %	99.914 %
0.01	16.726 %	15.792 %	99.225 %	99.149 %
0.1	15.923 %	15.791 %	92.111 %	91.726 %

After calculating the p-value for each window in the two-class analysis of Novellette, a user-defined threshold (default: $p < 0.01$) will be used to determine which windows, or 500 bp long regions in the genome, are considered differentially expressed. As the number of tests performed is large, using only a p-value threshold of $p < 0.01$ without multiple testing correction would lead to having thousands of statistically significant test results just by random chance. In this work, multiple testing correction is not used, but instead this feature is taken into account by setting two additional requirements for a window to be differentially expressed: 1) the absolute difference of the median RPM values between the two classes must be > 1 and 2) the ratio (*fc*, *fold change*) of median RPM values between the two classes must correspond to $|\log_2(fc)| > 1$. These two requirements guarantee that 1) there are enough reads to make a reliable call and 2) the difference in terms of fold change is significant as well (at least two-fold difference). Although using the median values (default option in Novellette) in each window as the single-value representation of a class is fairly robust, this approach may miss the novel transcripts that are e.g. highly expressed only in a small subset of the samples of one class. Therefore the user of Novellette can also choose to use mean instead of median in the two-sample analysis.

3.3.3 One-class analysis

Although cancer studies mostly focus in comparing samples from different classes with each other, one approach in a cancer study is to identify *outliers*: samples in a single class that behave in a different way – e.g. in terms of the expression of certain genes – than other samples in the class. With this approach, it is possible to distinguish subsets of similar samples that may e.g. represent a novel cancer subtype. In an RNA-seq study, both known and unknown genes can be used in this kind of analysis.

In Novellette, if only one class of samples is available, an outlier analysis will be performed. This is carried out by first sorting the RPM values of the samples in each window in an ascending order, and then dividing the samples into two groups based on a user-defined quantile value. By default, Novellette uses the quantile 0.8, i.e. 20 % of the largest values are put into one group and the rest into the other (overexpressed outliers), or the quantile $1 - 0.8 = 0.2$, in which case the smallest 20 % are compared against the rest (underexpressed outliers). Novellette then compares the median expression of these groups with the same criteria as in the two-class analysis apart from the p-value, which in this case is unreasonable to calculate. Finally, windows with a significant difference (in terms of absolute RPM difference and fold change) between the groups are considered as novel transcript candidates for the downstream analyses.

The approach used in the one-class analysis of this work aims to find only such novel transcripts that have high or low expression in a subset of samples. It would also be possible to look for regions that have high expression in all of the samples, but that would not give any information on the differences between the samples, thus preventing any conclusions on potential cancer subtype markers. In addition, if there are no reference samples, it is impossible to determine whether or not a feature that appears in all of the studied cancer samples would also appear in another class of samples.

3.3.4 Filtering and merging

As the final goal in Novellette is to find novel transcripts, the output differentially expressed windows from the previous analysis step should be filtered by using the location information of known genes. This information is available from several different sources (Ensembl (Flicek et al. 2013), Refseq (The Reference Sequence Project 2013), UCSC (Rosenbloom et al. 2013), Gencode (Harrow et al. 2012)), and a reasonable approach is to combine some or all of this information. By default, Novellette uses a file that combines the genome annotations from Ensembl, Refseq, UCSC and Gencode by taking the union of the coordinates that represent a known gene in any of the annotation sources. The differentially expressed windows are then filtered with this annotation file, removing each window that overlaps with any known gene based on any annotation source.

Since the first and last exons of protein-coding genes may be several thousand bp long and the read counts are calculated in a much shorter window (by default), the output of the first part of the analysis may contain several consecutive, overlapping win-

dows. As these most likely represent the same transcript on the RNA level, all windows that are at most 3000 bp (default, can be tuned by user) away from each other, and all regions between them, are merged into a single window. Although this operation is not vital for the functionality of Novellette, it reduces the redundancy in the gene identification analysis and thus moderately reduces the computation time.

3.4 Gene identification

The second major part of Novellette performs a gene identification analysis to the filtered, merged differentially expressed windows. The goal of this analysis is to form the full exon structure of the transcript and evaluate its potential for being a protein-coding gene. In this work, a scoring approach is used to give each novel transcript candidate a numerical representation of its protein-coding potential instead of making binary calls on whether or not a given sequence is a gene. The score consists of four different components: exon structure score, promoter score, 3'UTR score and ORF score. Each component score $S_i \in [0,1]$ is defined by custom rules, with $S_i = 1$ representing good evidence of coding potential and $S_i = 0$ poor or no evidence at all. The final score is then defined as the arithmetic mean of the component scores.

In the gene identification analysis, Novellette uses the aligned reads in the BAM format of $\min(5, n_{\text{samples_total}})$ most highly expressed samples for each differentially expressed window. In addition, a file folder with the whole genome DNA sequence, each chromosome in a separate file, is required as an input parameter. In the next four subsections, forming of the gene structure by utilizing the BAM file and each of the component scores are discussed and, by using the DNA sequences of the genes based on Ensembl v68 annotations, biological rationalization is given for some of the score calculations.

3.4.1 Defining and scoring the exon structure

The first and most vital part of the gene identification analysis is to define the gene boundaries along with exon and intron breakpoints and intervals. To begin with the analysis of expanding a raw novel transcript candidate window into a full gene, each read that was aligned within the window is fetched from the input BAM files by calling SAMtools in Novellette. The following procedures are then carried out:

1. Start from the right end of the window and expand it to the left
 - 1.1. By using the CIGAR string field in the BAM file, exon breakpoint candidates are gathered from split-mapped reads
 - 1.2. As long as there are splice junction candidates that at least 3 reads per sample (on average) support, expand the region to the left end of the splice junction
 - 1.3. When there are no more spliced (split-mapped) reads, keep expanding the region to the left by sliding a 200 bp window, as long as it contains at least 10 reads per sample (on average)

2. Start from the left end of the original window and expand it to the right similarly. Whenever the region is expanded in any of the steps of the procedure described above, SAMtools is called again and new reads from the BAM files are fetched according to the expanded coordinates. Ideally the algorithm described above would terminate the expansion at the transcription start and termination sites and the resulting region would contain all exon-intron breakpoints of the transcript, but this will rarely be the result due to various biological features in the transcription and technical noise.

After the gene boundaries have been defined and a set of exon-intron breakpoints have been gathered, the next step is to determine which breakpoint intervals represent exons and which introns. Due to alternative splicing (i.e. alternative combinations of exons of the same gene) within the samples and noisy reads (i.e. reads that were falsely aligned on an intron of a gene), detected, consecutive intervals do not necessarily represent exons and introns alternately. In this work, Novellette does not attempt to distinguish alternative splice variants from each other, but instead a single consensus transcript with only one specific combination of exons is formed. This is achieved by using expression thresholding: intervals that contain a high amount of reads are considered as exons and intervals with low read count are considered as introns. This approach is based on the fact that RNA-seq targets the mRNA in the cell, which contains only the untranslated regions and exons of a gene, but no introns (see Figure 2.3b). For this purpose, Novellette calculates the number of reads in the interval and divides it with the interval length, and if this value is less than 0.1 (default), the region is considered as an intron. In other words, e.g. a region with length 1000 bp and 90 aligned reads is interpreted as an intron.

As the goal with every component of the gene identification analysis in Novellette is to give a score that describes how well a certain feature in a novel transcript candidate corresponds to the same feature in known, protein-coding genes, an additional aspect is taken into account when determining the goodness of an exon candidate. Since the first two nucleotides (GU and AG) in a splice junction are very well conserved in human cells (Burge & Karlin 1997), each exon candidate is considered valid only if the splice junction motifs in 5' and 3' ends match the dinucleotides GU and AG, respectively. In addition, as most protein-coding human genes have multiple exons, regions in the genome with no evidence for multiple exons (no split-mapped reads) are likely to represent non-coding or repetitive regions (such as LINEs and SINEs). By taking these features in account, the exon score S_{ex} is calculated by

$$S_{ex} = \begin{cases} 0, & n_{ex} = 1 \\ \frac{n_{valid}}{n_{ex}}, & n_{ex} > 1, \end{cases} \quad (30)$$

where n_{ex} is the total number of exons and n_{valid} is the number of exons with valid splice junction motifs.

In addition to calculating the exon score, the directionality of the gene is determined in this part of the analysis. This is done by inspecting the splice junction dinucleotides at each exon: if the dinucleotides are GU and AG in donor and acceptor sites, respectively, the gene lies on the same strand as what was used in the reference genome, based on this exon. In this case, ‘+’ will be assigned as the directionality of the novel transcript candidate. However, as the human genome is double-stranded (most of the time), genes are found also on the reverse complementary strand compared to the reference genome sequence. Therefore the reverse complementary dinucleotides CT and AC in donor and acceptor sites, respectively, found in the splice junctions would imply the opposite directionality for the gene based on this exon, and ‘-’ would be assigned as the orientation. However, due to the fact that genes on the opposite strands in the genome may overlap with each other, the data-driven approach of Novellette may result in assigning the exons of two or more genes on different strands to a single novel transcript candidate, since the reads produced with RNA-seq are not strand-specific. To compensate for this effect, Novellette will give a special flag and reduced exon score to each identified transcript that yield strand-contradictory exons. The score reduction is performed by assigning $n_{valid} = \max(n_{pos}, n_{neg})$ in Equation 30, where n_{pos} and n_{neg} are the number of exons with positive or negative directionality based on valid splice junction motifs, respectively, instead of the total number of valid splice junctions. For single-exon genes the directionality cannot be determined from the splice junction motifs, since there are no splice junctions.

For every subsequent analysis step in Novellette, the directionality information determined in this step is utilized: e.g. if only positive strand splice junctions are found, all subsequent features will be searched in the positive strand as well. On the other hand, if there are contradictory splice junction motifs or there is only a single exon, both directions (positive and negative, i.e. forward and reverse complementary orientation) will be used in the subsequent analyses.

3.4.2 Promoter score

To calculate a score for the promoter region of the novel transcript candidate, the promoter features discussed in Section 2.1.2 and illustrated in Figure 3.2 are taken into account by calculating a separate score for each feature and then combining them by their arithmetic mean. The final promoter score consists altogether of three different components: 1) Promoter motif score, 2) CG enrichment score and 3) Chargaff’s second rule violation score. According to Chargaff’s rules, the fraction of nucleotide A is equal to T and the fraction of C is equal to G in any single- or double-stranded genome (Chargaff et al. 1952). The balance in dsDNA occurs due to complementary base pairing, but much less is known why the same balance holds also for single-stranded DNA. However, this rule no longer holds for the ssDNA sequence of a gene, which is a phenomenon of very little information given in literature, but which can clearly be seen in Figure 3.2b.

The first component of the promoter score is based on three different motifs (TA-TA-box, initiator element and downstream promoter element), which are commonly bound by the general transcription factors required for transcription initiation. Any of the three motifs mentioned above is sufficient for initiating transcription, and therefore the promoter motif score S_{motif} is defined as

$$S_{motif} = \begin{cases} 1, & \text{perfect match found for any of the motifs} \\ 0, & \text{otherwise,} \end{cases} \quad (31)$$

where the match is defined by string comparison of consensus sequences (Figure 3.3).

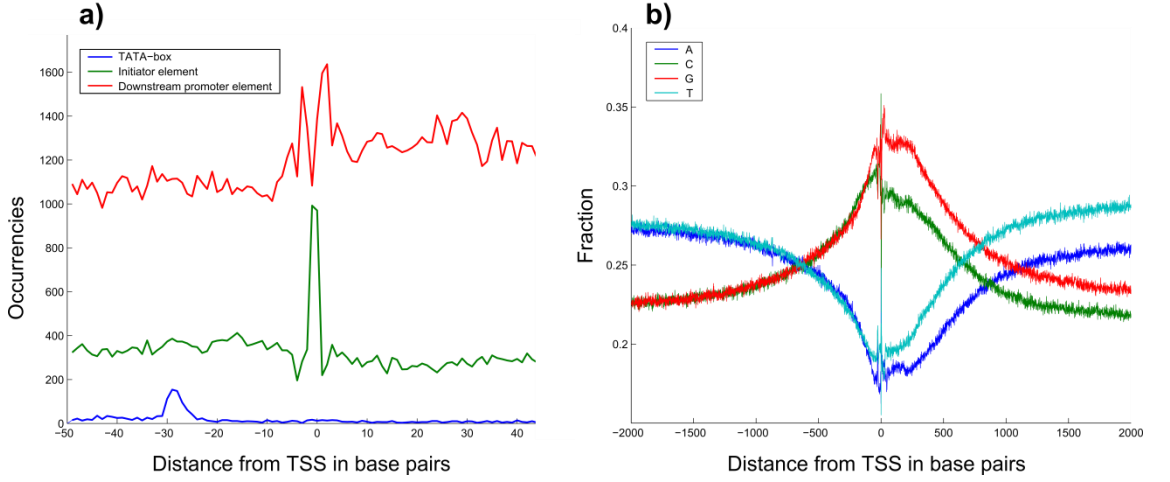


Figure 3.2. Illustrations of the occurrences of three different promoter motifs often bound by transcription factors near the TSS (a) and the nucleotide composition 2 kbp upstream and downstream from the TSS (b). The Chargaff's rule applies upstream from the TSS as the A and T curves (as well as the C and G curves) align perfectly, but when moving downstream from the TSS, the curves diverge. These plots were generated by using the DNA sequence of approximately 21000 protein-coding genes based on Ensembl annotations.

Consensus	Nucleotide
A	A
C	C
G	G
T	T
M	A,C
R	A,G
W	A,T
S	C,G
Y	C,T
K	G,T
V	A,C,G
H	A,C,T
D	A,G,T
B	C,G,T
N	A,C,G,T




Downstream promoter element	
Consensus: RGWYV	
Initiator element	
Consensus: YYANWYY	
TATA-box	
Consensus: TATAWAW	

Figure 3.3. In the consensus sequence representation, a single letter is used to encode for alternative nucleotides at a certain location in a motif. For example, as the consensus sequence of the downstream promoter element begins with ‘R’, the first nucleotide in the motif can be either ‘A’ or ‘G’. The consensus sequences for the three motifs used in this work are based on the most conserved sequences reported in literature (Kadonaga 2002; Shi & Zhou 2006; Xi et al. 2007).

The second and third components of the promoter score are based on the nucleotide composition before and after the TSS of a gene (see Figure 3.2b). Since the fraction of cytosines and guanines of all nucleotides is substantially higher ($\%(C + G) > 0.55$) around the TSS than in the genome on average ($\%(C + G) < 0.4$, not visible in Figure 3.2), this information can be used to distinguish a TSS from its surroundings. In Novellette, the score for GC enrichment is formed by calculating the fraction of G and C nucleotides, $\%(G + C)$, in a sliding window of width 250 bp (sliding step 50 bp), starting from 500 bp upstream of the TSS location predicted in the previous part of the analysis and ending at 500 bp downstream from it. The score S_{GC} is then defined by choosing the largest found fraction and calculating

$$S_{GC} = \begin{cases} 0, & \%(G + C) < 0.4 \\ \frac{\%(G + C) - 0.4}{0.15}, & 0.4 \leq \%(G + C) \leq 0.55 \\ 1, & \%(G + C) > 0.55. \end{cases} \quad (32)$$

A similar approach is used to score the Chargaff’s second rule violation. Since generally $\%(A + C) = \%(G + T) \approx 0.5$, but inside the gene $\%(A + C) \approx 0.45$ and $\%(G + T) \approx 0.55$ (based on visual judgement of Figure 3.2b), this small difference will be utilized when calculating the score. The Chargaff score, S_{Charg} , is finally calculated as

$$S_{Charg} = \begin{cases} 0, & \%(G + T) < 0.5 \\ \frac{\%(G + T) - 0.5}{0.1}, & 0.5 \leq \%(G + T) \leq 0.6 \\ 1, & \%(G + T) > 0.6, \end{cases} \quad (33)$$

where again the largest fraction $\%(G + T)$ is used. The value 0.6 is used instead of 0.55 to assure more stringent evidence for the Chargaff's second rule violation, since the chance for a random DNA sequence of length 500 – 1000 bp to have $\%(G + T) > 0.5$ is fairly large.

The window widths, sliding steps and distances to TSS used in this part of the analysis are all user-tunable, and the values reported above correspond to the default values. By using the three component scores defined above, the final promoter score, S_{prom} , can now be calculated:

$$S_{prom} = \frac{S_{motif} + S_{GC} + S_{Charg}}{3}. \quad (34)$$

In addition to calculating the promoter score, Novellette stores the locations of each found motif (the ones nearest to TSS, if multiple matches are found) and the windows with the highest GC enrichment and Chargaff's second rule violation. This information is not utilized in the current version of Novellette, although it might be useful e.g. for improving the estimation of TSS location.

Although any of the three promoter motifs is enough to enable transcription factor binding, especially the downstream promoter and initiator element motifs are so unspecific that they can be found even in fairly short (< 500 bp), random DNA sequences just by chance. Therefore the motifs alone are a poor indicator of whether a region in the genome could represent the promoter of a protein-coding gene. In addition, as Figure 3.2 shows statistical features of nucleotide compositions of transcription start sites of more than 20000 genes, the nucleotide composition of the promoter of a single gene does not necessarily correspond to these statistical features. This is not a major drawback, however, as no final, binary call of protein-coding potential is made based on the calculated scores.

3.4.3 3'UTR

As the wetlab protocol of RNA-sequencing is based on the detection of the poly-A tail in the mRNA molecule, every full transcript identified by RNA-seq should contain the poly-A signaling motif, AAUAAA. Since both the sequencing and preliminary data analysis processes can yield errors and noise in the data, this cannot always be guaranteed. Therefore the identification of poly-A motif from novel transcript candidates gives both evidence for protein-coding potential and a way to control the quality of detected transcripts, both novel and known.

According to (Retelska et al. 2006), the poly-A tail complex can bind to the motif AAUAAA with a single-nucleotide mismatch in any of the six nucleotides. In addition, the motif is often followed by a U-rich region in the mRNA molecule (or T-rich region in the gene/DNA). This feature is investigated similarly as in Figure 3.2, by using the known 3'UTR DNA sequences of protein-coding genes based on Ensembl annotations. In this case, for each found poly-A motif, the composition of nucleotides surrounding the motif is calculated. As is shown in Figure 3.4, there is a clear enrichment of T's 40 – 60 bp downstream from the motif location.

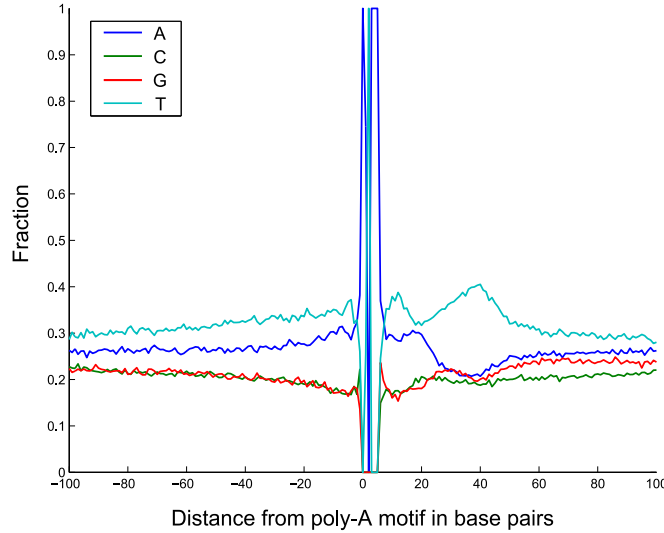


Figure 3.4. The nucleotide composition surrounding detected poly-A motifs, showing the T-rich region, which peaks at 40 bp downstream from the motif with an average fraction of $\%T = 0.4$.

The 3'UTR score consists of two components: poly-A motif score and T-rich region score. To calculate the motif score, poly-A motifs are first searched from the start coordinate of the last exon to the end coordinate of that exon + 50 bp (default) in case of multi-exon genes, or from end coordinate – 500 bp (default) to end coordinate + 50 bp for single-exon genes. The poly-A motif score, is then defined by

$$S_{\text{polyA}} = \begin{cases} 1, & \text{perfect match found} \\ 0.5, & \text{near perfect match found} \\ 0, & \text{otherwise,} \end{cases} \quad (35)$$

where the near perfect match corresponds to a single-nucleotide mismatch between the DNA sequence and the motif.

To take into account the T-rich region often followed by a functional poly-A site, the fraction of T's in up to 80 bp (default) downstream of each location where a poly-A motif was found is calculated. This is performed by using a sliding window of width 20 bp (default) and a sliding step of 5 bp (default). Similarly as for the promoter analysis, the largest fraction $\%T$ is stored, and the T-rich region score is calculated by

$$S_{T-rich} = \begin{cases} 0, & \%T < 0.3 \\ \sqrt{\frac{\%T - 0.3}{0.1}}, & 0.3 \leq \%T \leq 0.4 \\ 1, & \%T > 0.4. \end{cases} \quad (36)$$

Square root is used to make the requirement of T enrichment slightly less stringent, as the evidence of every poly-A site being followed by a T-rich region based on Figure 3.4 is not very strong, implying that it may not be vital to enable the poly-A complex to bind on DNA. Having the motif and T enrichment scores calculated, the final 3'UTR score is then defined as

$$S_{3'UTR} = \frac{S_{polyA} + S_{3'UTR}}{2}. \quad (37)$$

3.4.4 Open reading frames

The last component of the gene structure score is formed by identifying open reading frames within the coding sequence of the gene. This is performed by joining the identified exons together and thereby rebuilding the corresponding DNA sequence of the measured mRNA of each transcript, and then finding each start and stop codon on any of the three possible frames (or six, if the other strand of the gene must be searched as well). Subsequently, each valid start codon - stop codon pair on the same frame is identified and the longest three (per strand), non-overlapping pairs (raw ORFs) are further processed. In Figure 3.5, the raw ORF detection procedure is illustrated.



Figure 3.5. An example DNA sequence and the resulting raw ORF.

In the example of Figure 3.5, one start codon and four stop codons are found in the given DNA sequence. In ORF detection, the start codon frame is denoted with 0 and the following two codons with 1 and 2. Since the mRNA molecule is translated one codon at a time into a protein, each codon in the same frame with the start codon is processed into an amino acid. Therefore the third codon after the start codon ('CGC' in the example) is the next codon in line when translating the mRNA sequence, and it is given the

frame 0. Out of the four different stop codons, only one of them shares the same frame with the start codon and is hence the translation termination codon.

In the next step of the ORF detection analysis of Novellette, for each of the top three raw ORFs, the number of exons and the frames for each exon that belongs to the ORF are calculated. Exon frames are determined by the frame of the first nucleotide in the exon. If the longest raw ORF covers each exon in the transcript, it will represent also the final ORF of the novel transcript candidate and the corresponding start and stop codon coordinates and the exon frames will be stored for result printing. If there are no ORFs that would cover each exon, three longest raw ORFs are reported as the final result.

The score for transcripts with multiple exons is then calculated as

$$S_{\text{ORF}} = \left(\sqrt{\frac{n_{\text{ORF}}}{n_{\text{ex}}}} + p \right) / 1.5, \quad (38)$$

where n_{ORF} is the number of exons that belong to the longest ORF, n_{ex} is the total number of exons and $p = 0.5$, if the last exon is included in the ORF, otherwise $p = 0$. The pseudo-constant p assures that identified ORFs suggesting the transcript to be more prone to nonsense-mediated decay get a significant penalty to score. For single-exon transcripts, the score is set to $S_{\text{ORF}} = 0.25$ if a valid ORF is found, otherwise $S_{\text{ORF}} = 0$. The best score ($S_{\text{ORF}} = 1$) is never given to a single-exon transcript since the expected number of start and stop codons e.g. in a 500 bp long random, continuous DNA sequence are $\left(\frac{1}{4}\right)^3 \cdot 498 = 7.78125$ and $\left(\frac{1}{4}\right)^3 \cdot 3 \cdot 498 = 23.34375$, respectively, and therefore it is also highly likely to find a valid start codon - stop codon pair within the sequence by chance. The probability to find a stop codon is multiplied by three since there are three different possible combinations of nucleotides for stop codons, while there is only one possible trinucleotide that can act as a start codon.

3.4.5 Result printing

As illustrated in Figure 3.1, Novellette produces two different output files. The first one is a genome browser viewable text file in the general transfer format (*.gtf*) [ref], which contains the chromosomal coordinates of each identified novel transcript, both for the full transcript and for each exon and coding sequence region separately. When viewed in a genome browser, the transcript is shown as a line with boxes at the location of each exon. When opening this file in a genome browser, the user can browse through the results graphically and easily inspect the exon structures of the novel transcripts. An example of a *.gtf* file opened in IGV is shown in the next chapter.

The second output file is a table with detailed data of the gene identification analysis, with one row for each identified transcript. In this table, the most relevant information gathered in the gene identification analysis (such as the chromosomal coordinates and the calculated scores for protein-coding potential) is reported. In Table 3.2,

the output table fields are explained and example values for a novel transcript identified in the TCGA GBM dataset are given. In this example, the transcript is found in chromosome 6 and its final gene structure score is 0.9566668, indicating high protein-coding potential. The only feature that did not get a maximum score is the promoter, which may be due to good but not perfect GC enrichment or evidence for Chargaff's second rule violation, although both a downstream promoter element (DPE) and initiator element (ini) are found on the promoter. The total number of exons in this case is 15 and each one of them also belongs to the longest ORF, thus giving a perfect ORF score of 1. In Novellette, the exons are numbered according to their location on the forward strand, and hence the exons of a transcript on the reverse strand are numbered in a reverse order. Therefore the exon with the largest ordinal number is the first one to be translated.

4 RESULTS

In this chapter, the results from two different analysis cases with Novellette are reported and the performance of Novellette is investigated. In the first case, one-class analysis is performed for the 169 TCGA GBM samples. The second case covers a two-class analysis of prostate cancer (PC) samples versus castration-resistant prostate cancer (CRPC) samples from a recent study (Annala et al. 2013). In addition, by running the gene identification analysis with RNA-seq reads from known, protein-coding genes and regions outside any known genes, the capability of the scoring approach in Novellette to distinguish protein-coding sequences from random DNA sequences in the genome is evaluated. Finally, a survival analysis is performed for the one-class analysis case to study whether the outlier expression approach of Novellette is able to discover novel transcripts that may have biological significance.

4.1 One-class analysis with GBM data

In this section, the results for running a full analysis with Novellette with the same 169 glioblastoma multiforme samples that were used in the previous analyses are reported. The full analysis covers the following steps: detection of interesting windows based on outlier overexpression analysis, merging of consecutive windows into longer, overlapping regions, survival analysis based on expression values calculated in these merged regions, and finally the gene identification analysis.

4.1.1 Preliminary analysis and outlier expression

The analysis begins with aligning the 169 raw sequence files with TopHat (using default parameters) against a reference genome (version hg19) and using SAMtools to create sorted and indexed, binary files (BAM files). The RPM normalized read coverage values are then calculated with Novellette, using a 500 bp window and a 250 bp sliding step. Since there are some regions in the genome that contain unmappable sequences (e.g. the telomeric repeat regions) and regions that typically yield no transcribed RNA (intergenic regions), the coverage matrix may contain a significant amount of rows with only zeros. These rows and also windows that overlap with any known genes are removed from the matrix in this analysis before moving to the next step.

With the filtered coverage matrix of RPM values, an outlier overexpression analysis is performed using the quantile $q = 0.8$. This approach aims to detect all 500 bp windows with significantly higher expression in 20 % of the most highly expressed samples compared to the rest. After the RPM values of each 500 bp window with at least one

non-zero value have been divided into two quantiles this way, the default criteria (as described in Section 3.2.2) are used for filtering outliers with significant difference in expression. As a result, 336 differentially expressed windows are detected before and 113 after merging consecutive diff. exp. windows. Based on a graphical inspection with IGV, however, some of these windows appeared to contain millions of reads mapped with a very poor quality. When comparing the sequence of these reads with the PCA adapter sequences used in the wetlab protocol of Illumina Genome Analyzer sequencer (which was used in the TCGA project), a 100 % match is found, implying a failure in purifying the RNA content prior to sequencing. Therefore these windows are systematically removed as technical artifacts, leaving **53 final novel transcript seeds** for the gene identification analysis.

4.1.2 Gene identification

Using the 53 novel transcript candidate regions and default parameters for the gene identification analysis in Novellette, the scores are calculated and the full gene structures are formed, using the top 5 most highly expressed samples for each transcript as input for the data-driven gene identification. The mean and standard deviation of the resulting total score values are 0.5752018 and 0.1439068, respectively. In Figure 4.1, an example of the graphical illustration with IGV is shown for a) the identified novel transcript with the best total score and b) for the transcript with worst score. The best transcript shows a clear exon structure of two exons, and an open reading frame covering both exons is found. The worst transcript, instead, only has a single exon and also lacks the poly-A tail motif in the 3'UTR, suggesting that this transcript may actually be an RNA-seq artifact or a repetitive region in the genome. The full, detailed table with gene structures and scores is shown in Appendix B.

4.1.3 Survival analysis

To evaluate the biological significance of the identified 53 novel transcripts, the expression values in their seed regions (the input given to the gene structure identification analysis) for all 169 GBM samples are calculated. For each novel transcript candidate, the values are then divided into two subgroups (high expression and low expression) using K-means clustering. Finally, the survival times of patients belonging to each group are then used for calculating a p-value with logrank test between the low and high expression groups for each transcript. In Figure 4.2, survival plots for the best four transcripts based on the logrank test ($p < 0.05$ for each) are presented, each of which show a clear difference in survival between samples with high or low expression.

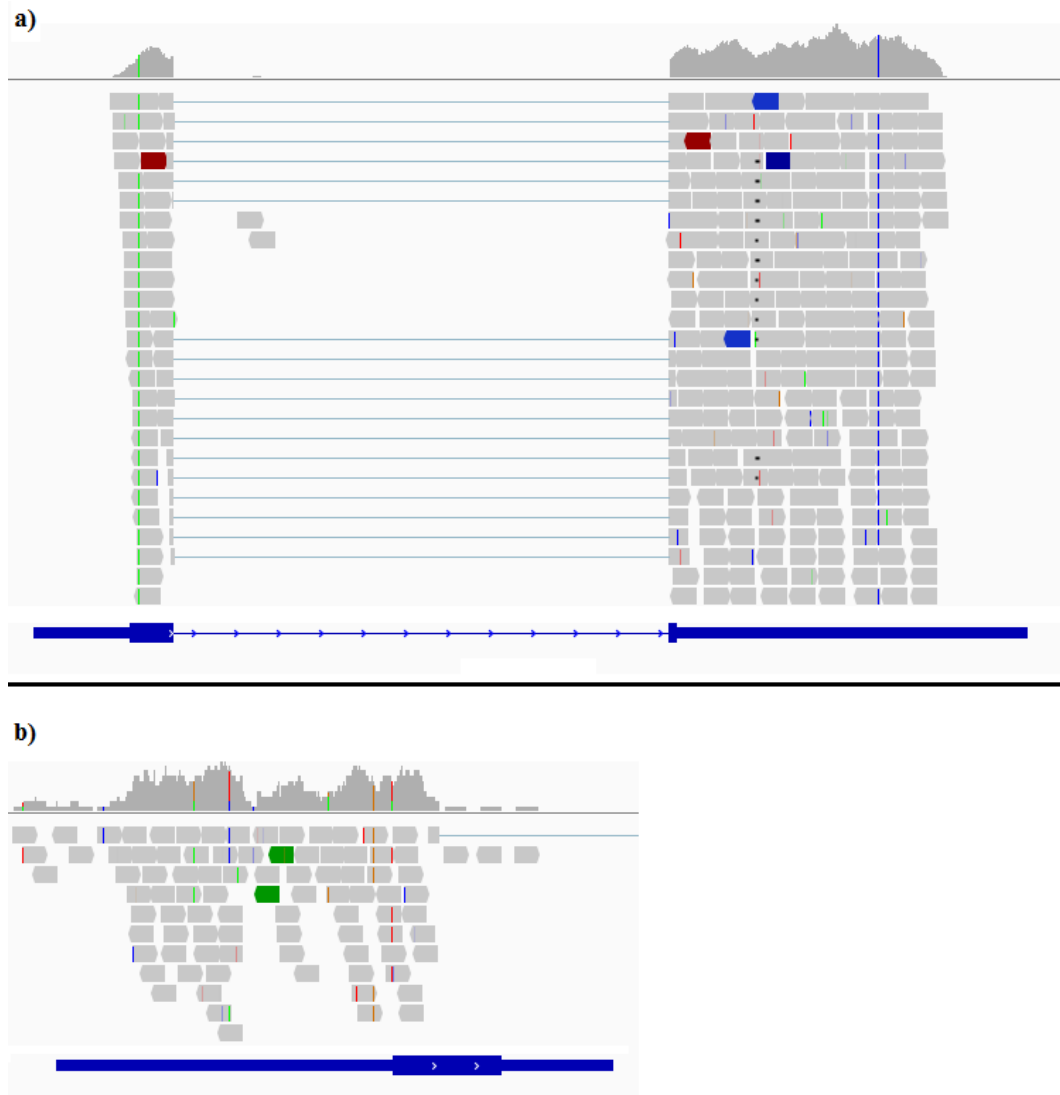


Figure 4.1. The best (a) and worst (b) identified novel transcript based on the scores of the gene identification analysis of Novellette. The grey boxes along with the histogram represent the aligned reads in one of the 169 GBM samples (G17190) and their counts at specific locations, respectively, and the blue connected boxes correspond to the transcript structure predicted by Novellette.

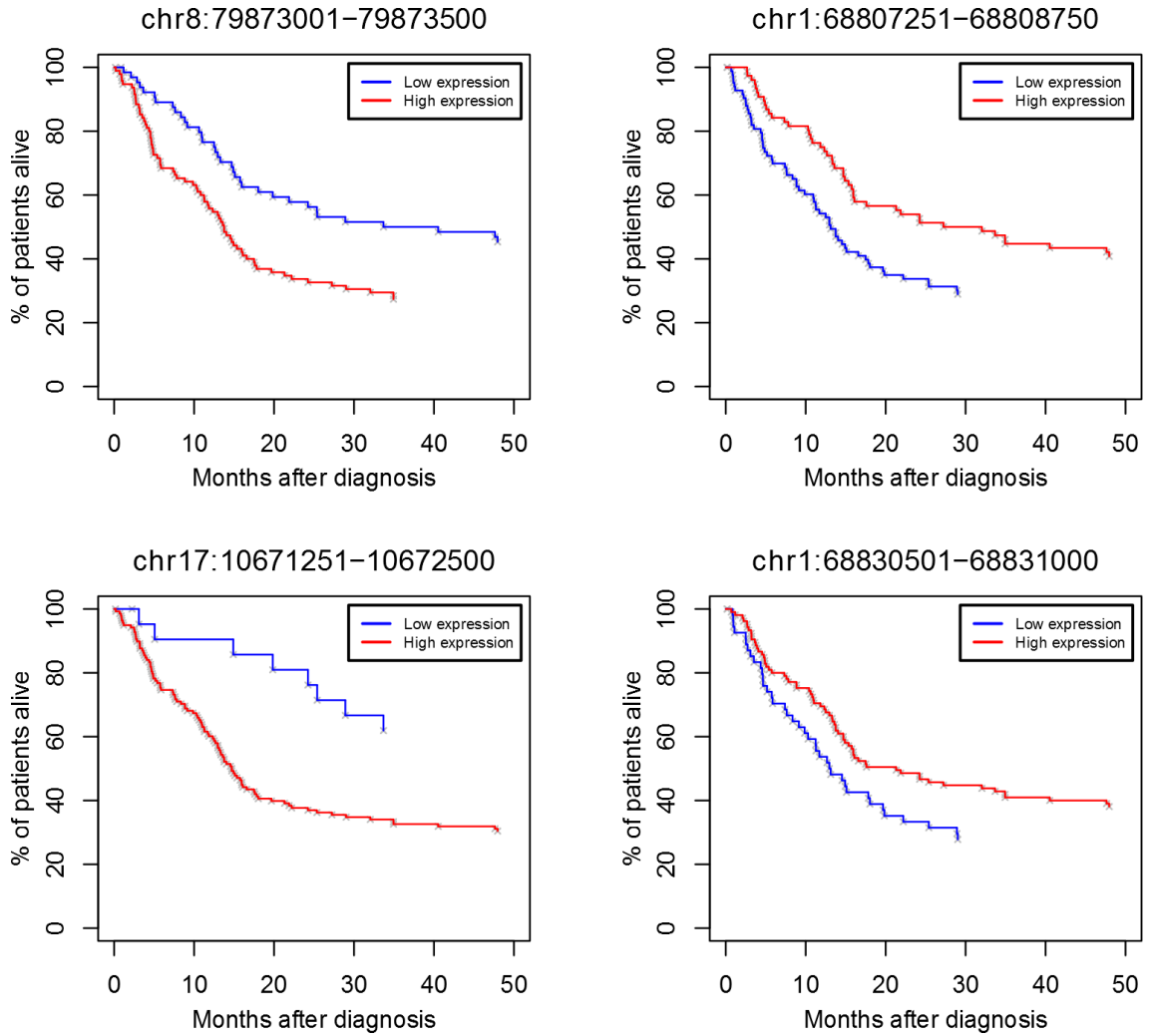


Figure 4.2. Survival plots for the best four novel transcripts, located in chromosomes 1, 8 and 17, based on logrank p -value. The plotted value represents the fraction of patients that are alive or censored, which is an optimistic estimate of the survival compared to a K - M estimate.

4.2 Two-class analysis: CRPC vs. PC

In this subsection, the statistical analysis part of the Novellette pipeline is evaluated by using prostate cancer data from two different subtypes, CRPC (castration resistant prostate cancer) and PC (a mixture of other prostate cancers) from a recent study (Annala et al. 2013). The CRPC class consists of 12 samples and the PC class from 29 samples. First, the coverage matrix is calculated using the default window size and sliding step and regions that overlap with known genes are filtered out. A statistical test (U-test) is then performed for each row in the matrix, comparing the RPM values of CRPC samples with those of PC samples. This analysis resulted in 491 differentially expressed windows with $p < 0.05$, one of which overlaps with a human endogenous retrovirus K (*hERV-K*) known to be highly expressed in several different cancer types (Agoni et al. 2013; Wang-Johanning et al. 2003).

4.3 Performance of the scoring approach

To test the performance of the proposed scoring approach as a method to distinguish protein-coding from non-coding regions in the genome, 25 protein-coding genes known to be differentially expressed in GBM (Brennan et al. 2013) and 181 intergenic regions with sufficient expression ($\text{RPM} > 1.00$), chosen randomly from chromosomes 1 and 2, are investigated in the gene structure analysis. The resulting score distributions for random regions and protein-coding regions are illustrated in Figure 4.3.

A clear difference can be seen in each of the score components and the total score, protein-coding regions yielding higher scores than the intergenic, random regions, which supports the validity of the chosen scoring approach (K-S test p-value for total score: $p < 0.0001$). As was expected, the 3'UTR scores are high for both of the test cases due to the RNA sequencing protocol. The differences in the exon and ORF scores are high due to the fact that protein-coding regions usually have multiple exons, while most intergenic, random regions yielding RNA-seq content are mainly non-coding or repetitive regions. The few outliers in the exon scores for protein-coding genes arise from genes on the opposite strands that either overlap or are very close to each other, which Novellette is unable to separate in the gene structure analysis and thus gives penalty in score due to contradictory strands in splice junction motifs. These also affect the ORF scores, since a continuous ORF covering each exon cannot be assigned to a transcript that consists of parts from genes on opposite strands.

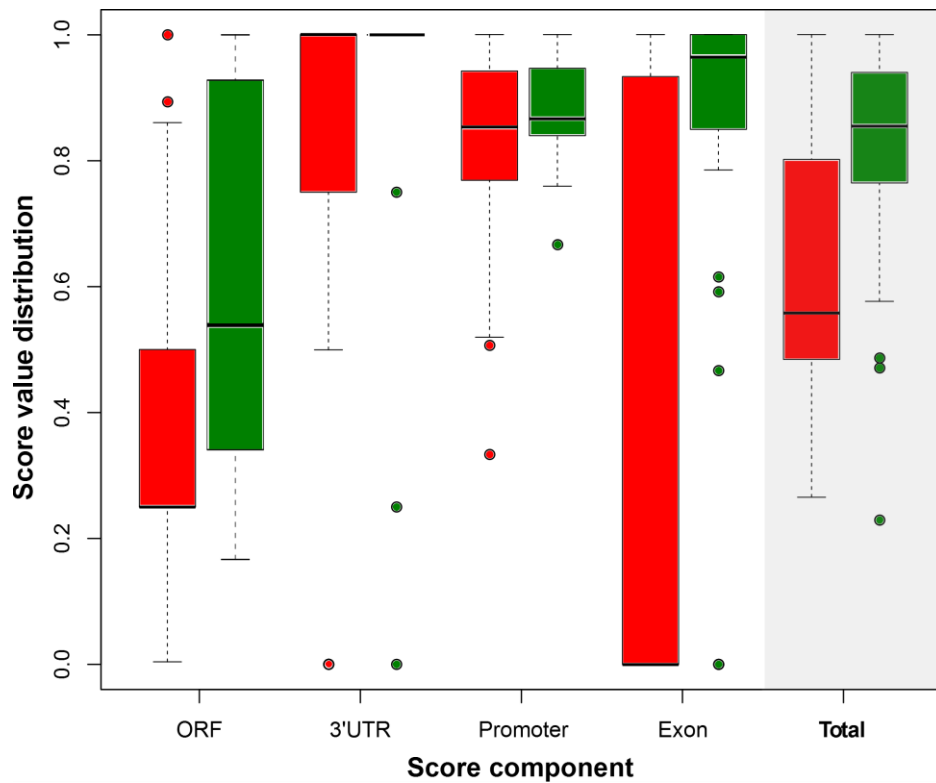


Figure 4.3. Boxplot illustration of the score distributions for each component and the total score, red boxes representing the random regions and green boxes the protein-coding regions.

5 DISCUSSION

In this work, a thorough analysis pipeline for detecting novel transcripts from RNA-seq data, Novellette, has been presented and the basics of statistical hypothesis testing and evaluating data similarity and normality have been introduced. The results of several analyses with different type of data all show that Novellette is able to reliably discover novel transcript candidates with biological significance and to distinguish protein-coding regions from non-coding regions with the scoring approach used in the gene identification analysis. However, improvements still need to be done to the gene identification algorithm as the run time for more than thousand highly expressed, protein-coding regions exceeded two weeks. In addition, the issue of merging two genes on opposite strands in the current version of Novellette is bypassed by assigning a penalty to the exon scores, although the problem could be thoroughly solved with an iterative component to the gene structure analysis, in which potentially separate transcripts would be split into two different transcript candidates. Finally, Novellette is currently available only as a set of separate source code files and binaries, but an integrated and easy-to-use command line tool that covers the whole analysis will be developed in the near future.

Data normality and whether or not it affects the choice of the statistical testing method used in a biological study has been discussed in this work. Although from a mathematical point of view the use of Welch t-test is justified only when a) the data is normally distributed or b) the sample size large enough to make the central limit theorem valid, it is widely used in bioinformatics even when neither of these conditions apply. As was shown in this work, count based RNA-seq data indeed is not normally distributed and therefore t-test is not used in the analysis pipeline of Novellette. However, biological studies utilizing t-test can still be successful even though its use is invalid, since the results are always validated in the laboratory with other methods as well. For gene expression values, e.g., the difference in mean or median expression (both fold change and the absolute difference) between two classes is a more important indicator of biological significance than a single p-value. In addition, since the normal distribution is a theoretical mathematical model that never describes perfectly a data set in a real study (whether or not biological), the t-test is actually never perfectly valid. In conclusion, the t-test is sufficiently accurate for any data set, the distribution of which at least slightly resembles a normal distribution, but if there is no evidence for data normality at all, other hypothesis testing methods such as U-test should be used.

REFERENCES

Agoni, L., Guha, C. & Lenz, J. 2013. Detection of Human Endogenous Retrovirus K (HERV-K) Transcripts in Human Prostate Cancer Cell Lines. *Frontiers in oncology* 3, pp. 180.

Alexandersson, M., Cawley, S. & Pachter, L. 2003. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome research* 13, 3, pp. 496-502.

Annala, M., Kivinummi, K., Ylipää, A., Kartasalo, K., Tuppurainen, K., Latonen, L., Leppänen, S., Kohvakka, A., Karakurt, S., Saramäki, O., Scaravilli, M., Seppälä, J., Rauhala, H.E., Yli-Harja, O., Vessella, R.L., Tammela, T.L.J., Zhang, W., Visakorpi, T. & Nykter, M. 2013. Integrative genomic characterization of untreated and castration resistant prostate cancers reveals novel tumorigenic mechanisms. Manuscript submitted.

Berty, H.P., Shi, H. & Lyons-Weiler, J. 2010. Determining the statistical significance of survivorship prediction models. *Journal of evaluation in clinical practice* 16, 1, pp. 155-165.

Birnbaum, Z.W. & Tingey, F.H. 1951. One-Sided Confidence Contours for Distribution Functions. *Annals of Mathematical Statistics* 22.

Bland, J.M. & Altman, D.G. 2004. The logrank test. *BMJ (Clinical research ed.)* 328, 7447, pp. 1073.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H., Beroukhi, R., Bernard, B., Wu, C.J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S.A., Ciriello, G., Yung, W.K., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D.D., Van Meir, E.G., Prados, M., Sloan, A., Black, K.L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D.W., Guha, A., Iacocca, M., O'Neill, B.P., Foltz, G., Myers, J., Weisenberger, D.J., Penny, R., Kucherlapati, R., Perou, C.M., Hayes, D.N., Gibbs, R., Marra, M., Mills, G.B., Lander, E., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P.W., Haussler, D., Getz, G., Chin, L. & TCGA Research Network 2013. The somatic genomic landscape of glioblastoma. *Cell* 155, 2, pp. 462-477.

Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* 11, pp. 94-2105-11-94.

Burge, C. & Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268, 1, pp. 78-94.

Chang, Y.F., Imam, J.S. & Wilkinson, M.F. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry* 76, pp. 51-74.

Chargaff, E., Lipshitz, R. & Geen, C. 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *The Journal of biological chemistry* 195, 1, pp. 155-160.

Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* 38, 6, pp. 1767-1771.

Croce, C.M. 2008. Oncogenes and cancer. *The New England journal of medicine* 358, 5, pp. 502-511.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29, 1, pp. 15-21.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. & Searle, S.M. 2013. Ensembl 2013. *Nucleic acids research* 41, Database issue, pp. D48-55.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N. & Regev, A. 2013. De novo transcript sequence recon-

struction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8, 8, pp. 1494-1512.

Hainaut, P. & Hollstein, M. 2000. P53 and Human Cancer: the First Ten Thousand Mutations. *Advances in Cancer Research* 77, pp. 81-137.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R. & Hubbard, T.J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 22, 9, pp. 1760-1774.

Kadonaga, J.T. 2002. The DPE, a core promoter element for transcription by RNA polymerase II. *Experimental & molecular medicine* 34, 4, pp. 259-264.

Kaplan, E.L. & Meier, P. 1958. Nonparametric estimation from incomplete observations, *Journal of American Statistical Association* 53, 282, pp. 457-481.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14, 4, pp. R36.

Kuhn, R.M., Haussler, D. & Kent, W.J. 2013. The UCSC genome browser and associated tools. *Briefings in bioinformatics* 14, 2, pp. 144-161.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10, 3, pp. R25-2009-10-3-r25. Epub 2009 Mar 4.

Lehmann, E.L. 1999. *Elements of Large-Sample Theory*. New York, Springer-Verlag. 631 p.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25, 14, pp. 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25, 16, pp. 2078-2079.

- Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D. & Doudna, J.A. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* (New York, N.Y.) 311, 5758, pp. 195-198.
- Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports*.Part 1 50, 3, pp. 163-170.
- Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 18, 9, pp. 1509-1517.
- Marsaglia, G., Tsang, W.W. & Wang, J. 2003. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software* 8, 18.
- Maxam, A.M. & Gilbert, W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74, 2, pp. 560-564.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5, 7, pp. 621-628.
- Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J.T., Robinson, D., Iyer, H.K., Palanisamy, N., Maher, C.A. & Chinnaiyan, A.M. 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* 29, 8, pp. 742-749.
- R: A language and environment for statistical computing, R Development Core Team, 2008 [WWW], [Cited: 31/10/2013], Available at: <http://www.R-project.org>.
- Rahman, M.M. & Govindarajulu, Z. 1997. A Modification of the test of Shapiro and Wilk for Normality. *Journal of Applied Statistics*, 24, pp. 219-235.
- The Reference Sequence (RefSeq) Project [WWW], [Cited: 31/10/2013], Available at: <http://www.ncbi.nlm.nih.gov/books/NBK21091>.
- Retelska, D., Iseli, C., Bucher, P., Jongeneel, C.V. & Naef, F. 2006. Similarities and differences of polyadenylation signals in human and fly. *BMC genomics* 7, pp. 176.

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R.A., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D. & Kent, W.J. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research* 41, Database issue, pp. D56-63.

Ruohonen, K. 2002. "Luotettavuus, käytettävyys, huollettavuus", Lecture notes, Tampere University of Technology.

Sanger, F. & Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94, 3, pp. 441-448.

Shapiro, S.S. & Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4, pp. 591-611.

Shi, W. & Zhou, W. 2006. Frequency distribution of TATA Box and extension sequences on human promoters. *BMC bioinformatics* 7 Suppl 4, pp. S2.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. & Birol, I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome research* 19, 6, pp. 1117-1123.

Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3, pp. Article3.

Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* 32, Web Server issue, pp. W309-12.

Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 14, 2, pp. 178-192.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28, 5, pp. 511-515.

Wang-Johanning, F., Frost, A.R., Jian, B., Epp, L., Lu, D.W. & Johanning, G.L. 2003. Quantitation of HERV-K env gene expression and splicing in human breast cancer. *Oncogene* 22, 10, pp. 1528-1535.

World Health Organization Fact sheet - cancer. [WWW], [Cited: 31/10/2013], Available at: <http://www.who.int/mediacentre/factsheets/fs297/en/>.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A. & Weng, Z. 2007. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome research* 17, 6, pp. 798-806.

Zerbino, D.R. & Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18, 5, pp. 821-829.

APPENDIX A

In this appendix, the most important parts of the analysis pipeline of Novellette are briefly introduced. Each of the source codes listed here are available for download at https://www.dropbox.com/s/zqssssivrs2g30w/thesis_sourcecodes.zip, although the process of developing a more user-friendly, integrative command line tool is still unfinished. In addition, the script for running a statistical testing analysis is currently under revision and hence it will not be available for download yet.

File	Description
copa.R	This R script performs the outlier analysis (Cancer Outlier Profiler Analysis). The quantiles and coverage matrix to be analyzed are given as command line parameters.
seqtools.cc	This file contains all of the functionality of the gene structure analysis pipeline. It is used by the script "gene_structure.cc".
seqtools.hh	The header file for seqtools.cc containing also all the (default) parameter values used in the gene identification analysis.
gene_structure.cc	The main program for running the gene identification analysis. This script will be the one that takes in all the user-defined parameter values for the gene identification (work still in progress).
rna_coverage.cc	This tool takes any number of SAM files, a sliding step size and a window size as input and calculates and prints the RPM normalized expression values in the coverage matrix format (.igv). If the input is in the binary (BAM) format, it can be first transformed into SAM format by using SAMtools.

APPENDIX B

This appendix contains the details of the 53 novel transcripts identified in the GBM dataset. Due to the large size of the table, some fields are omitted.

Chr	Start	End	Strand	TotalScore	ExScore	PromScore	UtrScore	OrfScore	PromInfo	nExons
1	37598751	37602000	+	0.963333	1	0.853333	1	1	ini,DPE,CpG_island	2
17	62461251	62464000	+	0.934167	1	0.986667	0.75	1	ini,DPE,CpG_island,Charg	4
6	92035851	92050750	-	0.869167	1	0.56	1	0.916667	ini,DPE	2
15	46702351	46709851	+	0.864167	1	0.706667	0.75	1	ini,DPE	2
16	47996640	48081601	-	0.84375	1	1	1	0.375	ini,DPE,CpG_island,Charg	2
1	154643960	154652301	+	0.800417	1	0.826667	1	0.375	ini,DPE,CpG_island	2
X	53192381	53209600	-	0.785	1	0.973333	0.25	0.916667	TATA,ini,DPE,CpG_island,Charg	3
7	9118001	9229550	-	0.729583	1	0.626667	1	0.291667	TATA,ini,DPE	4
13	110630447	110709856	-	0.721354	0.666667	1	1	0.21875	ini,DPE,CpG_island,Charg	4
1	68807355	68850000	-	0.699583	1	0.506667	1	0.291667	TATA,ini,DPE	2
17	10575770	10672451	- (+)	0.686797	0.666667	0.88	1	0.200521	ini,DPE,CpG_island	16
6	23307541	23640408	-	0.678822	0.875	0.746667	1	0.0936214	TATA,ini,DPE	9
17	21730501	21731500	+	0.650888	0	1	0.603553	1	ini,DPE,CpG_island,Charg	2
3	153501	157939	+	0.593056	0	0.622222	0.75	1	TATA,ini,DPE	2
2	173099601	173102201		0.5625	0	1	1	0.25	ini,DPE,CpG_island,Charg	1
5	138889751	138891450		0.5625	0	1	1	0.25	ini,DPE,CpG_island,Charg	1
9	67665301	67666101		0.5625	0	1	1	0.25	ini,DPE,CpG_island,Charg	1
17	30454751	30455351		0.5625	0	1	1	0.25	TATA,ini,DPE,CpG_island,Charg	1
9	42608751	42610551		0.5625	0	1	1	0.25	ini,DPE,CpG_island,Charg	1
21	48002751	48007551		0.561389	0	0.995556	1	0.25	TATA,ini,DPE,CpG_island,Charg	1
X	44653751	44654500		0.559167	0	0.986667	1	0.25	ini,DPE,CpG_island,Charg	1
2	112796851	112798251		0.555833	0	0.973333	1	0.25	TATA,ini,DPE,CpG_island,Charg	1
17	72184251	72184750		0.553611	0	0.964444	1	0.25	ini,DPE,CpG_island,Charg	1
9	70631051	70632051		0.5525	0	0.96	1	0.25	TATA,ini,DPE,Charg	1
9	42236301	42237301		0.5525	0	0.96	1	0.25	TATA,ini,DPE,Charg	1
9	68284501	68285250		0.550278	0	0.951111	1	0.25	TATA,ini,DPE,Charg	1
9	70596901	70599000		0.549167	0	0.946667	1	0.25	TATA,ini,DPE,CpG_island	1
1	153560951	153562351		0.548056	0	0.942222	1	0.25	TATA,ini,DPE,Charg	1
13	104880751	104881500		0.546944	0	0.937778	1	0.25	TATA,ini,DPE,Charg	1
2	140649501	140650500		0.543611	0	0.924444	1	0.25	TATA,ini,DPE,CpG_island	1
13	50528801	50530350		0.5425	0	0.92	1	0.25	ini,DPE,CpG_island	1
19	46927751	46932000		0.5325	0	0.88	1	0.25	ini,DPE,CpG_island	1
21	275888001	27589801		0.5325	0	0.88	1	0.25	ini,DPE,CpG_island	1
6	99297401	99299801		0.530278	0	0.871111	1	0.25	ini,DPE	1
9	42202151	42204050		0.529167	0	0.866667	1	0.25	TATA,ini,DPE,CpG_island	1
1	120693251	120697251		0.521389	0	0.835556	1	0.25	ini,DPE,Charg	1
19	47799751	47800351		0.5	0	1	0.75	0.25	ini,DPE,CpG_island,Charg	1
21	46975501	46976501		0.5	0	1	0.75	0.25	ini,DPE,CpG_island,Charg	1
2	104066278	104096501	+	0.499306	0	0.622222	1	0.375	TATA,ini,DPE	2
8	90598051	90600300		0.496944	0	0.737778	1	0.25	ini,DPE,Charg	1
6	153943501	153944250		0.491389	0	0.715556	1	0.25	ini,DPE	1
2	241587551	241589351		0.48	0	0.92	0.75	0.25	ini,DPE,CpG_island	1
13	110076001	110077000		0.4725	0	0.64	1	0.25	TATA,ini,DPE	1
18	64289001	64289750		0.47	0	0.88	0.75	0.25	ini,DPE,Charg	1
8	79872801	79873601		0.454722	0	0.568889	1	0.25	ini,DPE	1
2	104066001	104067900	+	0.45375	0	0.44	1	0.375	TATA,ini,DPE	2
19	31201651	31204451		0.45	0	0.8	0.75	0.25	TATA,ini,DPE,CpG_island	1
7	148199001	148199750		0.45	0	0.8	0.75	0.25	ini,DPE,CpG_island	1
7	54802501	54809750		0.431389	0	0.475556	1	0.25	TATA,ini,DPE	1
22	29574801	29577150	+	0.424777	0	0.595556	0.853553	0.25	TATA,ini,DPE	1
5	63682201	63688401		0.416667	0	0.666667	0.75	0.25	TATA,ini,DPE,Charg	1
5	87063301	87066700	+	0.373333	0	0.493333	0.75	0.25	TATA,ini,DPE	1
1	68807251	68809150	+	0.226944	0	0.657778	0	0.25	TATA,ini,DPE	1