Raphael HABLESREITER

Master Thesis

Establishment of an Analysis Pipeline for RNA-seq Datasets

Institute of Computational Biotechnology,
Graz University of Technology
Petersgasse 14, 8010 Graz, Austria
Head: Univ.-Prof. Dipl.-Biol. Dr.*rer.nat.* Christoph Wilhelm Sensen

Supervisors:
DI Julia Feichtinger, PhD
Univ.-Prof. Dipl.-Biol. Dr.*rer.nat.* Christoph Wilhelm Sensen

Graz, March 10th, 2017

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ...
(date) (signature)

Abstract

English

The main goal of my master thesis is the establishment of a RNA-Seq pipeline for transcriptomics datasets with a focus on cancer samples compared to normal controls.

The CancerRNAseq pipeline initially removes successively adapter sequences, low quality bases and RNA sequencing errors in a preprocessing step. Next, the alignment of the preprocessed reads can be generated by TopHat2, STAR or HISAT2. In the final step the CancerRNAseq pipeline calculates the significant differentially expressed genes. The results are documented via graphical output and tables, respectively.

The CancerRNAseq pipeline was applied to four publicly available RNA-Seq cancer datasets (lung, breast, prostate and colorectal cancer). The resulting gene lists were compared to the BioXpress database and candidates for cancer therapy such as TMPRSS4 could be detected.

RNA-Seq is among others an important tool for biomarker screening, which can be used for diagnosis, prognosis, monitoring or finding the correct treatment option.

German

Das Themengebiet meiner Masterarbeit ist die Entwicklung einer RNA-Seq Pipeline mit dem Schwerpunkt auf den Vergleich von krebsartigen und normalen Proben. Die Pipeline besteht grundsätzlich aus drei Teilen: (i) dem Vorverarbeitungs-, (ii) dem Mapping- und (iii) dem Analyse-Teil.

Im ersten Schritt werden Adaptersequenzen von den ungefilterten Sequenzen entfernt. Anschließend die Sequenzen von Basen schlechter Qualität und von Fehlern, die bei der RNA Sequenzierung entstanden sind, gereinigt. Zum Erstellen des Alignments kann TopHat2, STAR oder HISAT2 verwendet werden. Nach der Berechnung der signifikanten differenziell exprimierten Gene erfolgt die Aufbereitung der Resultate und die Speicherung als Grafiken sowie Tabellen.

Die entwickelte Pipeline wurde mit vier verschiedenen Krebs RNA-Seq Datensätzen getestet um die Ergebnisse zu validieren. Die Ergebnisse wurden mit der BioXpress Datenbank verglichen und potenzielle Angriffspunkte für die Krebsbehandlung, wie TMPRSS4, wurden entdeckt.

RNA-Seq kann unter anderem optimal zur Entdeckung von Biomarkern eingesetzt werden und ist daher ein wichtiges Tool in der Diagnose, der Prognose und der Überwachung oder der Auswahl einer angemessenen Krebstherapie.

Contents

Acknowledgements

My thesis would not have been possible without the help and support of many people. I would like to express my gratitude for their assistance, support and encouragement.

First and foremost, I thank my supervisor Julia Feichtinger for her excellent supervision and her continuous advice throughout the course of this thesis. I could not have imagined having a better advisor and mentor for my Masters thesis.

In addition, I would like to thank Christoph Wilhelm Sensen for his insightful comments and encouragement on this thesis.

I would like to acknowledge my colleagues at the Technical University of Graz for stimulating discussions, assistance and friendship.

Last but not the least, I am grateful for the understanding, love and support of my family and friends.

List of Figures

List of Tables

Abbreviations

Abbreviation	Definition
BRCA1/2	Breast cancer 1/2
C18	Malignant neoplasm of colon
C25	Malignant neoplasm of pancreas
C32	Malignant neoplasm of larynx
C33	Malignant neoplasm of trachea
C34	Malignant neoplasm of bronchus and lung
C50	Malignant neoplasm of breast
C61	Malignant neoplasm of prostate
C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue
cDNA	Complementary DNA
cTNM	Clinical staging tumor-node-metastasis
DB	(BioXpress) Database
DE	Differential expression
DNA	Deoxyribonucleic acid
DNase	Desoxyribonuklease
dNTP	Deoxynucleotide triphosphate
dsDNA	Double-stranded DNA
ENA	European Nucleotide Archive
FC	Fold change
FM index	Ferragina-Manzini index
FPKM	Fragments per kilobase of exons per million fragments mapped
lncRNA	Long non-coding RNA
LOH	Loss of heterozygosity
LSCC	Laryngeal squamous cell carcinoma
miRNA	Micro RNA
mRNA	Messenger ribonucleic acid
MSA	Multiple sequence alignment
ncRNA	Non-coding RNA
NGS	Next generation sequencing
NSCLC	Non-small cell lung cancer
nt	Nucleotide
PE	Paired end
PGM	Personal genome machine
pTNM	Pathologic staging tumor-node-metastasis
qRT-PCR	Quantitative reverse transcription polymerase chain reaction

RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RPKM	Reads aligned per kilobase mapped
RT-PCR	Reverse transcription polymerase chain reaction
rRNA	Ribosomal RNA
SBS	Sequencing by synthesis
SCLC	Small cell lung cancer
SE	Single end
smRNA	Small messenger ribonucleic acid
SNP	Single nucleotide polymorphism
SOLiD	Sequencing by oligo ligation detection
ssDNA	Single-stranded DNA
SQCLC	Squamous cell lung cancer
TNM	Tumor-Node-Metastasis
WGS	Whole genome sequencing
BWT	Burrows-Weehler transfoem

Background

1.1 Introduction to RNA-Seq

RNA-Seq is a process to obtain abundance and identity information from cDNA sequences [1]. Furthermore, this process involves experimental as well as computational techniques [1]. The enormous development regarding high-throughput sequencing in the last few years and therefore the advances in transcriptome analysis at the single nucleotide level have increased the demand for alignment and analysis tools for RNA-Seq data [2].

RNA-Seq was initially mentioned by Nagalakshmi *et al.* in 2008 [3]. This method offers a variety of additional features compared to microarray analysis, which is in use since the mid 1990s for differential gene expression analysis [4, 5]. In addition to detecting expression levels of genes, RNA-Seq is capable of finding novel and unannotaded transcripts, transcription boundaries at single nucleotide resolution and sequence variations, such as single nucleotide polymorphisms (SNPs) [5, 6]. RNA-Seq has a wide dynamic range of expression level compared to the hybridization-based methods (>9000-fold compared to a few hundred fold), because the quantification has no upper limit [5]. Microarray and RNA-seq experiments are generally regarded as screening methods. Due to its high sensitivity, (quantitative) reverse transcription polymerase chain reaction (RT-PCR/qRT-PCR) is commonly used for the validation of selected gene candidates from microarray analysis or RNA-Seq experiments [7].

1.2 The RNA-Seq Experiment

In an RNA-Seq experiment (see figure 1.1) RNA is obtained from a cell or tissue sample. RNA that has been converted to DNA is sequenced and the resulting sequencing data is analyzed [1].

1.2.1 Sample and Library Preparation

The main aim of the sample and library preparation is to generate enough copies of DNA fragments (converted RNA fragments) to get a representative and unbiased source of material suitable for the sequencing process [8]. Usually commercially available kits designed specifically for the various sequencing methods are used to derive the RNA molecules [1]. The RNA molecules have to be selected by size and treated with DNase to remove contamination of genomic DNA [1, 8]. The quality control, which involves monitoring of the degradation, purity and quantity of the isolated RNA, has to be accomplished ahead of the library preparation step [1]. To

generate the RNA-Seq library, the cleaned RNA is converted to complementary DNA (cDNA) [9]. Due to the specific sequencing length of each platform, which ranges from 25 to 30,000 nt [1, 9], the high molecular weight DNA has to be separated into fragments [10]. The cDNA fragment is either attached to adapters on one (single-end sequencing) or both ends (paired-end sequencing), which minimizes the bias of the library preparation step, before the conversion [1, 5].

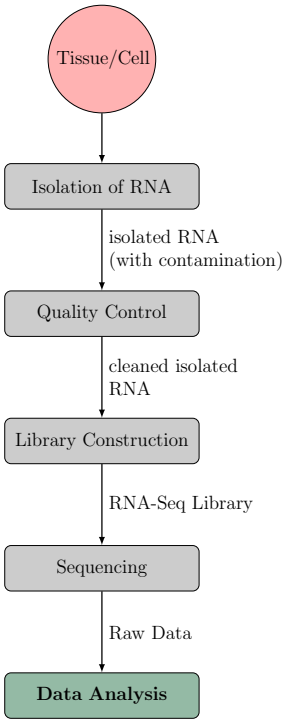


Figure 1.1: ***The RNA-Seq experiment workflow [1].*** *The RNA-Seq experiment initially starts with isolating RNA from a specific cell or tissue. Then the RNA-Seq library, generated from isolated specific RNA, is sequenced. This raw data can be further analyzed to find differentially expressed genes.*

1.2.2 Sequencing

Next-generation sequencing (NGS) technologies have become state-of-the-art in the field of biology since they became commercially available in 2004 [11]. These NGS technologies include Roche 454 [12], Illumina sequencing [13], Life Technologies sequencing by oligo ligation detection (SOLiD) [14] and the Ion torrent personal genome machine (PGM) [15]. In comparison with the automated Sanger method [16], which was the foundation for the first human genome sequence and controlled the sequencing sector for nearly 20 years, the NGS technologies made progress in terms of speed and amount of data generated [1, 8]. High-throughput sequencing technologies can produce millions of short sequence reads and have pushed the research in Genomic Biology a giant leap ahead [9, 11]. With RNA-Seq it is possible to analyze the transcriptome in a quantitative way and additionally gain information on the connection of two exons (observed with short reads) and the connectivity of several exons (analyzed with longer reads) [5]. The paired-end (PE) sequences have the

advantage of increase randomisation of fragments and short fragments might overlap leading to additional information in comparison with single-end (SE) sequences [1].

The different sequencing platforms can be categorized by either their specific library preparation process (*i.e.,* local clonal amplification [17] and no amplification [10, 18, 19]) or by the process on how the identify the nucleotide sequence (*e.g.,* Roche 454 pyrosequencing method [12], Illuminas SBS approach[13] and Life Technologies SOLiD [14]) and Ion torrent PGM [15]. The third-generation of sequencing platforms do not require any amplification of the probes, because the sequencing process is sensitive enough to detect single molecules in the extension process of the template (*e.g.,* Helicos Heliscope [20, 21] and Pacific Biosciences (PacBio) SMRT systems [22]).

The Illumina sequencing process is explained in more detail here, because all of the datasets that were used to test the CancerRNAseq pipeline were generated on this platform (see table 2.1). The Illumina sequencing workflow [23] consists of three steps: (i) library preparation, (ii) cluster amplification and (iii) sequencing [23]. In the library preparation step the fragmented DNA or cDNA, as it is the case in RNA-Seq experiments, is ligated to adapters, which are used to hybridize the denatured double-stranded DNA (dsDNA) on the flow cell [10]. The templates produced for Illumina sequencing [24] are attached or immobilized on a flow cell with one, two or eight separated lanes, depending on the Illumina platform used [8, 10]. Specific oligonucleotides function as a primer to generate the initial copy of the single-stranded DNA (ssDNA) [10]. These oligonucleotides are complementary to the primers of the template DNA/cDNA [10] and the initial strands are duplicated by bridge amplification (see Mardis E. R. [11] for details). In the sequencing step the immobilized strands on the flow cell are read one nucleotide at a time. Fluorescent labeled deoxynucleotide triphosphates (dNTPs) are ligated to the fragments with a terminator, which is cleaved of after reading to allow the next read cycle to start [10]. Each of the four different bases (A, T, C, G) can be identified by the fluorescent signal that is emitted, when the specific fluorescent label of dNTP is cleaved off [10].

1.2.2.1 FASTQ - File Format

The FASTQ file format [25] is commonly used for sequencing data. It is based on FASTA file format, with the capability to add a quality score to the nucleotide sequence (see table 1.1).

Table 1.1: ***FASTQ File Format.*** *The FASTQ file format consist of four lines for each read [25, 26]*

Line Number	Symbol	Description
1	@	title line and optional description
2		sequence line(s)
3	+	optional repeat of title line
4		quality line(s)

Commonly the FASTQ file consists of four lines for each read (table 1.1). The first line prefaced by a "@" character is the title line, which contains usually only a record identifier. The second line contains the sequence information. The third line can optionally be used to repeat the title line and starts with a "+" character. The fourth

line or quality line contains the quality score for the sequence listed in the second line [25].

Because of the fact that there is no formal definition for the FASTQ file format three different types have been developed: (i) the Sanger standard format (fastq-sanger), (ii) the Solexa/early Illumina format (fastq-solexa) and (iii) the Illumina 1.3+ format (fastq-illumina) (for further details see [25]). The Sanger standard format and the Illumina 1.3+ format both use a PHRED [27, 28] quality score but with a different offset [25]. The PHRED quality score is the log-transformed error probability for a base-call, resulting in high values for low error probabilities and *vice versa* [28].

1.3 RNA-Seq Data Analysis Workflow

The CancerRNAseq pipeline established in this project uses the general RNA-Seq workflow (see figure 1.2) as described by Korpelainen *et al.* [1] as a model.

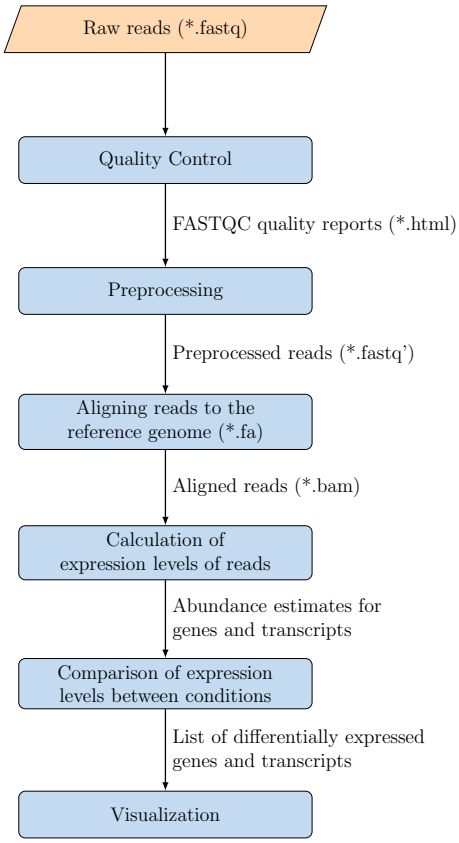


Figure 1.2: ***General RNA-Seq Workflow.*** *Initially a quality control of the input data (raw reads) is performed. The next step is to process the raw reads to improve their quality, which can, for instance, be achieved by an adapter trimming step (preprocessing). After the alignment of the preprocessed reads, the calculation and furthermore a comparison between expression levels can be calculated. The last step in a RNA-Seq workflow is the visualization of the results for a better understanding, as well as better interpretability [1].*

Initially a quality control of the input data (raw reads) should be performed. The

next step is to process the raw reads to improve their quality. This preprocessing step may consist of a different types of trimming steps (*e.g.,* quality trimming, adapter trimming). After the alignment of the preprocessed reads to a reference genome, the expression levels of the reads can be calculated. The subsequent step is to compare the expression levels between the various conditions and to list all the differentially expressed genes/transcripts. The last step in the RNA-Seq workflow is to visualize the calculated data for a better interpretability [1].

In this section tools for preprocessing, mapping and analysis are listed. For a detailed explanation on methodology applied by the tools used in the CancerRNAseq pipeline see appendix 6.1.

1.3.1 Preprocessing Tools

Preprocessing of the raw data can be accomplished by tools such as PRINSEQ [29], SEECER [30], TagCleaner [31], Cutadapt [32], Trimmomatic [33] and rCorrector [34]. The preprocessing step of the CancerRNAseq pipeline consist of three different consecutively used tools (Cutadapt, Trimmomatic and rCorrector) and additionally FASTQC to create quality reports for each input and processed sequence (see figure 3.1), which is the standard tool used for quality reports. HTQC [35], a toolkit to generate quality reports of Illumina sequencing data, could be an alternative for FASTQC.

1.3.2 Mapping Tools

Alignment of RNA-Seq reads can be accomplished by a number of tools such as STAR [2], TopHat [36], HISAT [37] or GSNAP [38]. In the CancerRNAseq pipeline the successor of HISAT, HISAT2 [37], the advanced version of TopHat, TopHat2 [39] and STAR [2] are used to align the preprocessed reads to a reference genome (see figure 3.2). Qualimap [40] or samtools [41] can be used to generate a quality report of the aligned reads. CancerRNAseq uses Qualimap to generate quality reports of each alignment.

1.3.3 Analysis Tools

The calculation of differentially expressed genes can be accomplished by tools such as edgeR [42], Cufflinks pipeline (Cufflinks [43] - Cuffmerge [44] - Cuffdiff [44]), DESeq [45] or limma [46]. Reporting significant DE genes between two conditions in the CancerRNAseq pipeline is achieved with Cufflinks, Cuffmerge and Cuffdiff (see figure 3.3). CummeRbund was used to access and visualize the Cuffdiff output in an efficient manner [47]. Lots of other Rpackages, such as edgeR [42], DESeq2 [48] or RNAseqViewer [49], can be used to visualize RNA-Seq data.

1.4 Introduction to Cancer

Example cancer RNA-Seq datasets were used to test the CancerRNAseq pipeline.

After diseases of the cardiovascular system, cancer was the second most common cause for mortality in Austria in 2014. In numbers, 26.2% of deaths (29.1% of mortality in men and 23,6% of mortality in women) are caused by cancer [50].

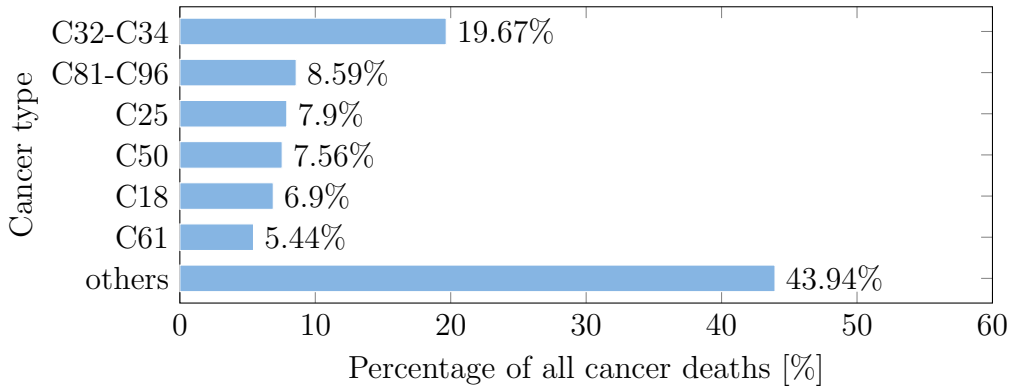


Figure 1.3: ***Frequency of most common cancers accountable for deaths in 2014 [50].*** *[(C18) malignant neoplasm of colon, (C25) of pancreas, (C32) of larynx, (C33) of trachea, (C34) of bronchus and lung, (C50) of breast, (C61) of prostate, (C81-96) of lymphoid, haematopoietic and related tissue]*

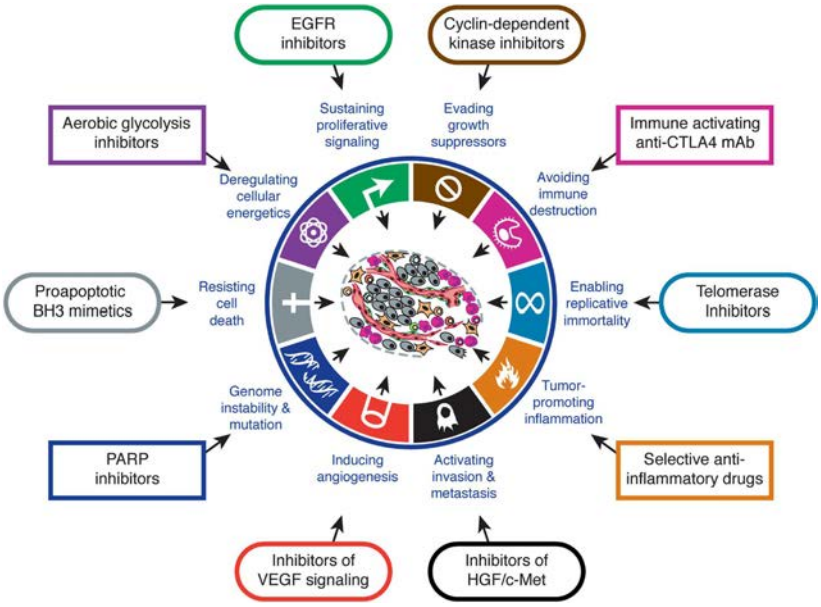


Figure 1.4: ***The 10 hallmarks of cancer and their therapeutic molecular target possibilities according to Hanahan and Weinberg [51].*** *Initially six hallmarks of cancer, which are features of a cancer that enable tumor growth and metastatic dissemination, were proposed in 2000 [(i) sustaining proliferative signaling, (ii) evading growth suppressors, (iii) activating invasion and metastasis, (iv) enabling replicative immortality, (v) inducing angiogenesis and (vi) resisting cell death]. In 2011 they added two emerging hallmarks [(vii) avoiding immune destruction and (viii) deregulating cellular energetics] and two enabling hallmarks [(ix) genome instability and mutation and (x) tumor-promoting inflammation].*

Carcinogenesis or tumorigenesis is a multistep process achieved by two main mechanisms: (i) the activation or epigenetic mutation of oncogenes and (ii) the silencing of tumor suppressors [7, 52]. This multistep process can be caused by various genetic changes, such as single nucleotide polymorphism (SNP), copy number

variations or epigenetic alterations, such as changes in DNA methylation and histone modification, respectively [52]. It has also been found out that non-coding RNA (ncRNA), such as micro RNA (miRNA), and long non-coding RNA (lncRNA) are involved in cancer development [53]. Alterations in genome stability genes can also support carcinogenesis, as they increase the change of alterations in other genes [7].

Hanahan and Weinberg have described six hallmarks of cancer in 2000 [54] and extended these six hallmarks with four additional hallmarks in 2011 [51] (see figure 1.4). Hallmarks can be seen as functional capabilities of cancerous cells that have to be achieved for enabling tumor development, as well as metastatic dissemination. Furthermore, these capabilities provide cancer cells with features needed to survive, proliferate and disseminate [54].

1.4.1 Cancer Types

In 2014, the most mortal cancer types in Austria (figure 1.3), were malignant neoplasms of the respiratory system (*i.e.,* larynx (C32), trachea (C33), bronchus and lung (C34)). Malignant neoplasms of lymphoid, haematopoietic and related tissue (C81-96) were responsible for 8.59% of cancer deaths [50]. The malignant neoplasm of pancreas (C25), of breast (C50), of colon (C18) and of prostate (C61) were responsible for 56.06% of all cancer related deaths in Austria [50].

The tumor-node-metastasis (TNM) classification is used to predict the survival rate of patients and therefore the specific TNM stage is commonly used for finding the necessary treatment method for a individual cancer at the moment, which is in general classified based on anatomical properties before (clinical staging (cTNM)) as well as after the treatment (pathologic staging (pTNM)) [55, 56]. Due to the categorization of cancer patients into TNM stages, where stage 1 means that the tumor could be survived and stage 4 means that the cancer will most probably be lethal, it is possible to treat them worldwide prognostically and therapeutically similar [55].

In this section the cancer types measured in the example datasets, which are used to test CancerRNAseq are explained in more detail. These cancer types have been selected, because they are among the most common cancer types in Austria (figure 1.3).

Prostate cancer is the most common cancer type in men in Austria in 2012 and has a 5-year-survival rate of 92.4% (2005-2009) [50]. Risk factors in favor of prostate cancer growth are age, diet and ethnicity. Inheritable mutations have been described in the *BRCA1* or *BRCA2* gene, respectively. Especially, mutations of the *BRCA2* gene lead to a more lethal and faster progression of prostate cancer [57].

Colorectal cancer is the fifth most common cause of cancer mortality in 2014 (see figure 1.3) [50]. Environmental factors, such as diet, family history, exercise, age and alcohol consumption are the major factors promoting the complex process of colorectal cancer development [58]. Generally, the 5-year-survival rate of colon cancer is 61.8% (2005-2009) [50].

Cancer of the respiratory system was the most common cancer type to die of in 2014 (figure 1.3) [50]. The detection of lung cancers in late stages leads to poor survival rates [59]. The 5-year-survival rate in Austria in 2012 was 18.0% (21.8% women, 15.8% men) [50]. The cause of 90% lung cancer cases can be traced back to smoking or the use of tobbacco products [59]. Lung cancer can be separated into two histological types that grow and spread differently: the small-cell lung cancer (SCLC) and the non-small cell lung cancer (NSCLC) [59].

The malignant neoplasm of breast (C50) was the most newly diagnosed cancer in 2012 in Austria and with 29.2% of all cancer incidences is the most common malignancy in women [50]. This type of cancer is very rare in men, which means that only 241 men were diagnosed with breast cancer from 2005 to 2009 compared to 24.740 women. The 5-year survival rate for breast cancer is 85.2% (85.2% for women, 80.4% for men), which is the second highest 5-year-survival rate after cancer of the thyroid gland cancer in women. Risk factors such as hormone levels (*e.g.,* estrogen), diet, age, pregnancy and family history are involved in breast cancer development [60].

1.5 Aims And Objectives

The first aim of this master thesis was to establish a RNA-Seq pipeline to automate the computation of differentially expressed genes between two different conditions with a focus on cancer samples compared to normal controls. The second aim of the project was the application of the pipeline to a collection of publicly available cancer RNA-Seq datasets and to the visualization of the resulting gene lists. These aims were accomplished by the specific objectives listed below:

Specific Objectives

1. Literature research on RNA-Seq analysis
2. Establishment of a modular RNA-Seq pipeline for two different conditions with a focus on publicly available cancer RNA-Seq datasets
3. Analysis of a publicly available cohort of cancer RNA-Seq datasets using the established pipeline
4. Visualization of the analysis results
5. Comparison of resulting gene lists to the BioXpress [6] database

Methods

The established CancerRNAseq pipeline is publicly available on GitHub[1].

2.1 Data Retrieval

The example cancer datasets (see table 2.1) to test the CancerRNAseq pipeline were selected according to the most mortal cancer types in Austria (see figure 1.3) and downloaded from the European Nucleotide Archive (ENA) [61]. All reads were generated with an Illumina Sequencing platform and all the samples derive from *Homo Sapiens*.

Table 2.1: List of datasets used for testing the CancerRNAseq pipeline

ENA Study Accession	**Organism**	**Cancer Type**	**Samples**	**Sequencing Platform**	**Library Layout**	**Ref.**
PRJEB2449	*Homo Sapiens*	Prostate Cancer	14 Normal 14 Tumor	Illumina HiSeq 2000 (180bp)	PAIRED	[62]
PRJNA218851	*Homo Sapiens*	Primary Colorectal Cancer	18 Normal 18 Tumor	Illumina HiSeq 2000 (100bp)	PAIRED	[63]
PRJNA163279	*Homo Sapiens*	NSCLC (Non-small-cell lung carcinoma)	6 Normal 6 Tumor	Illumina HiSeq 2000 (300bp)	PAIRED	[26]
PRJEB4829	*Homo Sapiens*	Breast Adenocarcinoma	2 Normal 2 Tumor (MCF-7 cell line)	Illumina HiSeq 2000 (100bp)	PAIRED	[64]

For alignment, the *Homo Sapiens* high coverage GRCh38 (GRCh38.p5, Genome Reference Consortium Human Build 38) primary assembly was downloaded from Ensembl [65]. This assembly contains all the sequences flagged as toplevel and is an unmasked genomic DNA sequence. The haplotypes as well as the patches are excluded in this reference genome build.

For annotation, the GRCh38.84 *Homo Sapiens* gene annotation file in GTF (General Transfer Format) format was downloaded from Ensembl [66].

[1] `https://github.com/RaphaelHablesreiter/CancerRNASeqPipeline.git`

2.2 Implementation

The CancerRNAseq pipeline is implemented in Perl[2] (v.5.010), which means that the program file is a simple text-file that is compiled and executed directly with the Perl compiler from shell [67]. Perl packages for the pipeline are available at the Comprehensive Perl Archive Network[3] (CPAN) (for a list of Perl packages used, see appendix 6.2.1).

The CancerRNAseq pipeline uses R (version 3.3.0 (2016-05-03) -- "Supposedly Educational") to generate various figures to visualize the results. The R library cummeRbund (version 2.14.0) [47] from Bioconductor[4] is used to create diagrams and plots of the DE genes as well as diagrams for alignment specific parameters. The R library plyr[5] (version 1.8.4) is used for ordering data frames [68].

The CancerRNAseq has been executed and implemented on a GNU/Linux (`Linux 2.6.32-642.11.1.el6.x86_64 x86_64`) based server (`cbt01.cbt.tugraz.at`), with 64 CPUs (each with 2 threads) and approximately 1.03TB (1025664620kB) RAM.

CancerRNAseq consists of three modules: (i) preprocessing, (ii) mapping and (iii) analysis and visualization. The tools used in the three modules are described in sections 2.4-2.6 (detailed information can be found in the appendix 6.1).

2.3 Configuration Files

CancerRNAseq uses configuration files (see appendix 6.4) to obtain all the necessary information (*e.g.,* names of files, conditions, user and project, storage paths and additional parameters) from the user for a successful pipeline run. Configuration files for the different datasets containing the used parameters to run the analysis are shown in the appendix [PRJEB4829 - breast adenocarcinoma (chapter 6.4.3), PREJB2449 - prostate cancer (chapter 6.4.5), PRJNA163279 - NSCLC (chapter 6.4.4), PRJNA218851 - primary colorectal cancer (chapter 3.3.4)]. Additionally a general configuration file containing paths of the tools used in CancerRNAseq as well as default parameters, which is necessary for the execution of the pipeline.

2.4 Preprocessing

Preprocessing of the raw sequences takes place in three steps: The initial step uses Cutadapt [32] to remove the Illumina adapter sequences from the reads. Then, Trimmomatic [33] performs a quality trimming, as typically the end of the reads are of poor quality, and removes too short reads (read length $<$ 20bp). The last tool of the preprocessing step, rCorrector [34], tries to repair sequencing errors.

Cutadapt (v1.9.1) [32] was used for adapter trimming with default parameters and standard Illumina adapters as input (see Chapter 6.2.2 for more information). The standard Illumina adapters were used with a prefaced `-a` (3' adapter read 1) or `-A` (3' adapter read 2, in case of paired end reads) to trim adapters in both reads.

[2] available at `https://www.Perl.org/get.html`
[3] `http://www.cpan.org`
[4] available at `http://bioconductor.org`
[5] available at `https://cran.r-project.org/`

Trimmomatic (v0.35) [33] was used for quality trimming of the sequences, with a 3 base wide window, and a quality threshold of 20 (`SLIDINGWINDOW:3:20`). Reads that were shorter than 20bp (`MINLEN:20`) were removed.

rCorrector (v1.0.1) [34] was applied with default parameters to correct for random sequencing errors.

FastQC[6] (v0.11.2) was applied to the raw files and was used after each preprocessing step with default parameters to generate a quality report.

2.5 Mapping

The CancerRNAseq can be used with HISAT2 [37], STAR [2] or TopHat2 [39] as alignment tool. All three alignment tools were used for the PRJEB4829 dataset to test their performance (see 2.7), whereas only STAR was used for the datasets PRJNA163279, PRJEB2449 and PRJNA218851.

The FM Index for the HISAT2 alignment tool was built with the HISAT2 (v2.0.4) [37] genome build tool. Sequences were aligned with HISAT2 (v2.0.4) with default parameters and additionally the options `--dta-cufflinks` to provide an appropriate output file for Cufflinks and `--sp 1000,1000` to suppress soft-clipping.

Samtools (v1.2) [41] was used to convert the HISAT2 alignment file into an appropriate BAM file for Cufflinks. Samtools was used with the command `view` and `-bS` to convert the alignment file from SAM format to BAM format. Then, samtools was used with the option `sort` for sorting the converted alignment file by the leftmost coordinates.

The genome index for the STAR alignment was generated with the option `--runMode genomeGenerate` and `-sjdbOverhang 74` as suggested by the developers of STAR. Furthermore, the genome reference file `-genomeFastaFiles` and the gene annotation file `-sjdbGTFfile` were used to generate the STAR index.

Sequences were aligned to the *Homo Sapiens* reference genome with STAR (v2.4.5a) [2] with default parameters and additionally with the options `--outSAMstrandField intronMotif` and `--outFilterIntronMotifs RemoveNoncanonical` to get sorted output alignment file appropriate for Cufflinks, `-alignEndsType EndToEnd` to suppress soft-clipping and `--readFilesCommand zcat` for the uncompression of the gzipped (*.gz) input files.

Samtools (v1.2) [41] was used to convert the STAR alignment file into an appropriate BAM file for Cufflinks.

The index was generated using Bowtie2 with default parameters. Sequences were aligned to the reference genome with TopHat2 (v2.0.14) [39] using default parameters.

QualiMap (v2.2) [40] was used to generate RNA-Seq quality reports, listing a number of specifications of the aligned reads. QualiMap was executed with default parameters and additionally with the options `rnaseq` and `-a uniquely-mapped-reads`.

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

2.6 Analysis

The calculation of differential expression is performed by Cufflinks [43], Cuffmerge [43] and Cuffdiff [44]. The R package CummeRbund [47] is used to process the Cuffdiff output.

Cufflinks (v2.2.1) [43] was used with default parameters on each alignment file. Because of contamination with mitochondrial DNA in case of dataset PRJEB2449, Samtools (v1.2) [41] was used with commands `idxstats` and `view` to delete aligned reads from this chromosome as following:

```
samtools idxstats <input-alignment-file> | cut -f 1 | grep -v MT |
xargs samtools view -b <input-alignment-file> > <output-alignment-file>
```

Cuffmerge (v1.0.0), provided with the Cufflinks package [43], was used with default parameters and the reference genome, reference annotation file and all generated alignment files as input. Cuffdiff (v2.2.1) [44] was used with default parameters and the reference genome, the merged file from the Cuffmerge run and all alignment files as input.

CummeRbund (v2.14.0) [47] was used to read the Cufflinks output into R and to generate plots: (i) dispersion, (ii) density, (iii) box, (iv) MA, (v) volcano, (vi) scatter matrix, (vii) scatter and (viii) bar plots (see appendix). Furthermore, tables from the DE genes were produced (see appendix).

The DE genes were prioritized using a log2-FC with an added constant of 0.1 to each FPKM/RPKM value ($\log_2$(FPKM/RPKM + constant)) (see Warden *et al.* [69]). This has to be used to avoid $\log_2(0)$, which results in infinite $\log_2$-FC values and, therefore, the differential expression is incomparable. The DE genes have been filtered according an abs($\log_2$-FC) $> |1|$ and an adjusted p-value < 0.05 to report significant DE genes.

2.6.1 Circos Plots

Circos [70] is a software package written in Perl for the visualization of data in a circular layout, was used to show the mutual under- and over-expressed genes of the four cancer datasets.

2.7 Comparison of STAR/TopHat2/HISAT2

The three different mapping tools of the pipeline were executed with three different datasets to compare the (ideal) execution time, the amount of mapped and unmapped reads and other alignment specific parameters between them.

Based on the fact that it is hardly possible to provide the same amount of computing power for all alignments, the (ideal) execution time has to be calculated using the system time, the user time and the number of threads. The elapsed wall clock time has not been used for this purpose, because waiting for other users and many other factors can influence it [71]. The system as well as the user time was measured with the GNU tool /usr/bin/time, which was used with the flag –verbose. The system time is the time the job has spent in the system kernel and the user time is the time the job has spent in the user kernel, which means that this is the time used for the system

and user respectively [71]. The ideal execution time was calculated according to the formula 2.1:

$$t_{exec} = \frac{t_{user} + t_{system}}{n_{threads}} \tag{2.1}$$

where:

$t_{exec,ideal}$ = (Ideal) Execution time (s)
t_{user} = User time (s)
t_{system} = System time (s)
$n_{threads}$ = Number of threads

The comparison was done based on QualiMap (v2.2) [40], which was used to generate RNA-Seq quality reports listing all the needed specifications of the aligned reads. QualiMap was applied with default parameters.

Results

3.1 The CancerRNAseq pipeline

CancerRNAseq is a fully modular in-house pipeline using two condition RNA-Seq datasets to calculate DE genes and to visualize the results. I have implemented CancerRNAseq and the pipeline is the main result of this project.

3.1.1 Use of CancerRNAseq

CancerRNAseq consists of three modules to generate reports, graphics and gene lists from raw reads following the workflow described in figure 1.2: (i) preprocessing, (ii) mapping and (iii) analysis and visualization.

CancerRNAseq is a command line tool and, therefore, can be executed by calling the `CancerRNASeq.pl` file. Additional parameters that influence the CancerRNAseq pipeline globally can be directly added to the execution command of the pipeline (*e.g.,* `$ perl CancerRNASeq.pl -config /path/to/file/userconfig.ini -silent -overwrite -threads 20`). The parameters that can be used with two hyphens placed in front of them for the execution of the pipeline are:

- `config <User Configuration File>`
- `help`
- `overwrite (default = FALSE)`
- `silent (default = FALSE)`
- `threads <Number of threads> (default = 1)`

The only input parameter that is required to call the CancerRNAseq pipeline successfully is the path and filename of the user configuration file in INI-format (*e.g.,* `$ perl CancerRNASeq.pl -config /path/to/file/userconfig.ini`), which contains user information such as input file path and parameters (*e.g.,* appendix 6.4.1 for an example). Additionally, a system configuration file (see appendix 6.4.2), which contains the path names for the tools used in the CancerRNAseq pipeline as well as default parameters, is provided within the program folder. The structure of the configuration files is explained in detail in section 2.3.

3.1.2 Preprocessing Module

Figure 3.1 illustrates the preprocessing workflow, which consists of three tools: (i) Cutadapt [32] for removing adapters, (ii) Trimmomatic [33] for quality trimming and (iii) rCorrector [34] for RNA bias correction. Throughout the preprocessing the quality

of the raw reads increases successively. Furthermore, quality reports are generated by FASTQC[1] of the initial raw reads and after each tool to monitor the quality.

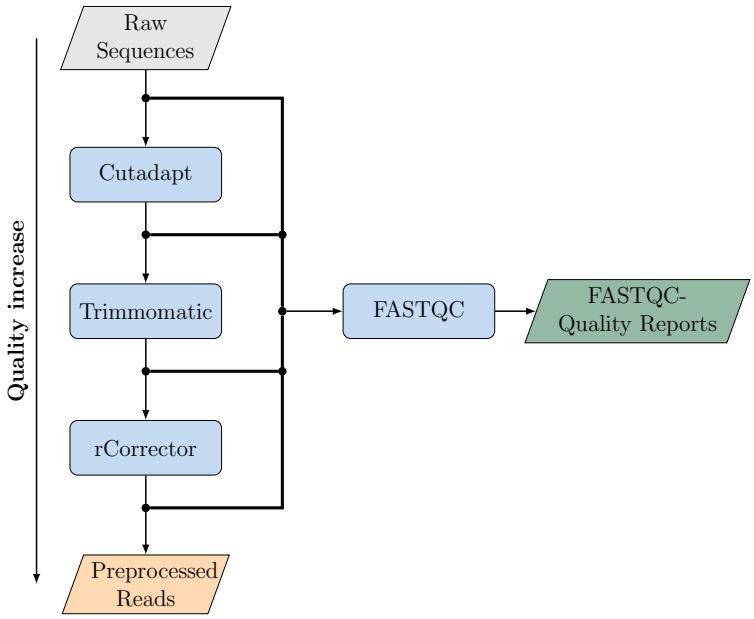


Figure 3.1: ***Preprocessing Workflow.*** *In this figure the blue nodes represent programs, which are used in the pipeline. The grey node illustrates input data and the orange node represents data, which is further used in another component of the pipeline. The green node represents data, which is a result of a section of the pipeline (here quality report).*

3.1.3 Mapping Module

In the mapping step (figure 3.2), one of three available alignment tools can be used for mapping the preprocessed reads to a reference genome: (i) TopHat2 [39] - using Bowtie1/2 as core alignment tool and its own indel-finding algorithm, (ii) HISAT2 [37] - using a global FM (entire genome) and numerous small FM indexes for the alignment and (iii) STAR [2] - aligning the non-contiguous sequences directly to the reference genome in a two step process. The results of the mapping module are the aligned and sorted reads and a quality report of the alignment generated by QualiMap [40].

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

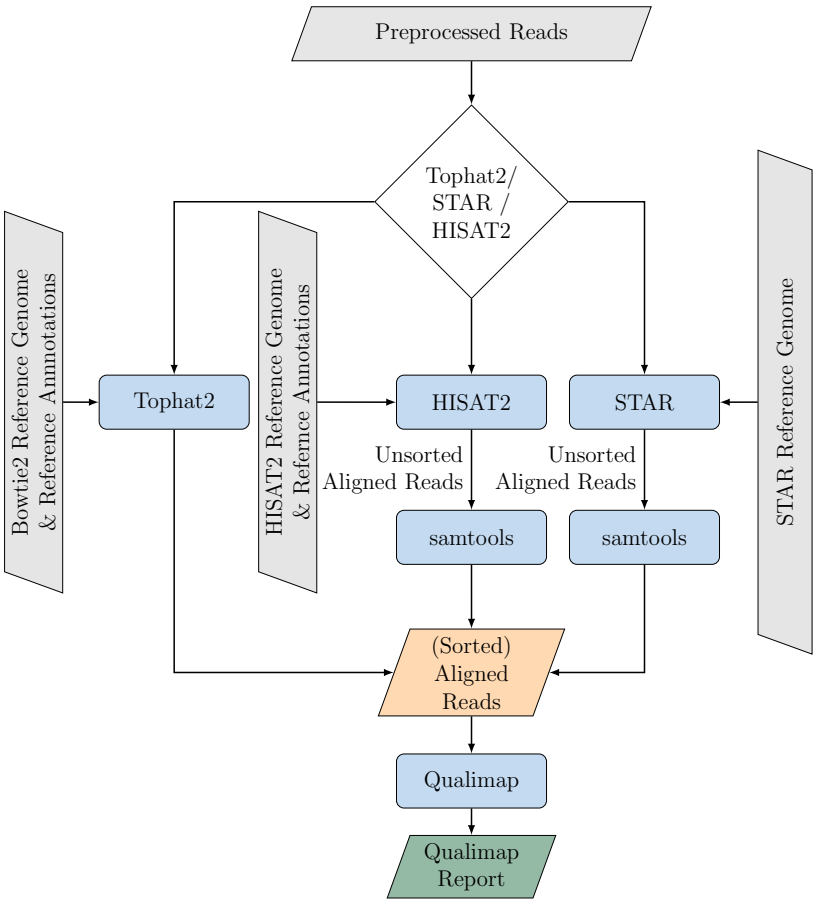


Figure 3.2: ***Mapping Workflow.*** *In this figure the blue nodes represent programs, which are used in the pipeline. The white decision node shows, that the pipeline has the option to decide between the three different mapping methods. The grey nodes illustrate input data and the orange node represents data, which is further used in another component of the pipeline. The green node shows data, which is a result of a section of the pipeline.*

3.1.4 Analysis and Visualization Module

In the analysis and visualization module (see figure 3.3) the calculation of DE genes is accomplished by: (i) Cufflinks [43] - calculates the transcripts and their abundances, (ii) Cuffmerge [43] - merges all Cufflinks files and (iii) Cuffdiff [44] - uses count variances for each transcript in each library for statistical testing to report significant DE genes. The R package CummeRbund [47] is used to create graphics, gene lists and reports from the Cuffdiff output.

3.1.5 Structure of CancerRNAseq

The CancerRNAseq pipeline is called with the `CancerRNAseq.pl` file and, if needed, additional parameters. The script then calls the Perl module `DRIVER.pm`, which controls the stepwise execution of the different tools used to generate the results (see figure 3.4). These tools are separated into three different Perl modules namely `PREPROCESSING.pm`, `MAPPING.pm` and `ANAlYSING.pm`. Two additional Perl modules

are used by the CancerRNAseq pipeline, one for the initial configuration of the necessary parameters (`CONFIGURE.pm`) and one for short functions used several times in the code (`PIPELINETOOLS.pm`).

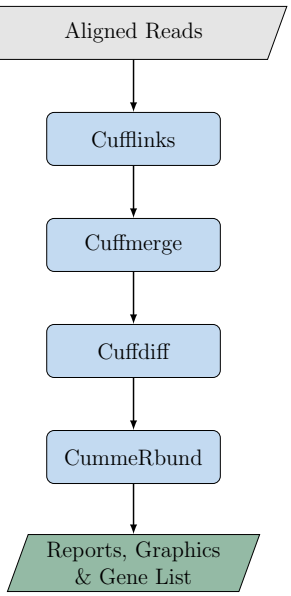


Figure 3.3: ***The analyzing workflow.*** *In this figure the blue nodes represent programs, which are used in the pipeline. The grey node illustrates input data and the green node shows data, which is a result of a section of the pipeline.*

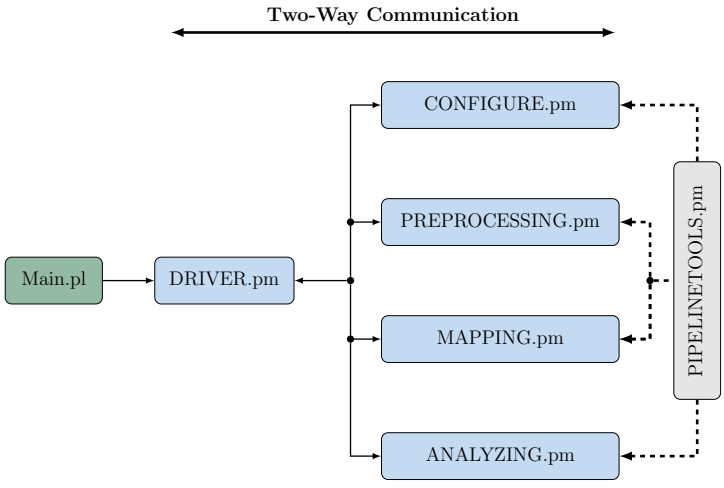


Figure 3.4: ***Fundamental structure of the CancerRNAseq pipeline.*** *The "Main.pl" node illustrates the file, which is used to call the pipeline. The blue nodes represent modules for particular parts of the pipeline and the "PIPELINETOOLS.pm" node is a module, which contains functions used by more than one module.*

The CancerRNAseq pipeline is saved in the Program folder (see figure 3.5). The path to the RawData folder containing all the raw reads, gene annotation files and reference genome files (pink nodes in figure 3.5) have to be provided by the user config

file (see appendix 6.4.3-6.4.6 for details and examples). Furthermore, a path for the results folder has to be provided by the user, but all the subfolders (green nodes) are created automatically. This allows the user to find the latest results and files in an intuitive and quick way.

The structure of CancerRNAseq allows several executions on the same raw data, which can be useful, for instance, to compare the resulting gene lists of different alignment tools. Furthermore, if an index for a reference genome (and annotation file) has already been built, the pipeline does not create the index a second time, which saves execution time.

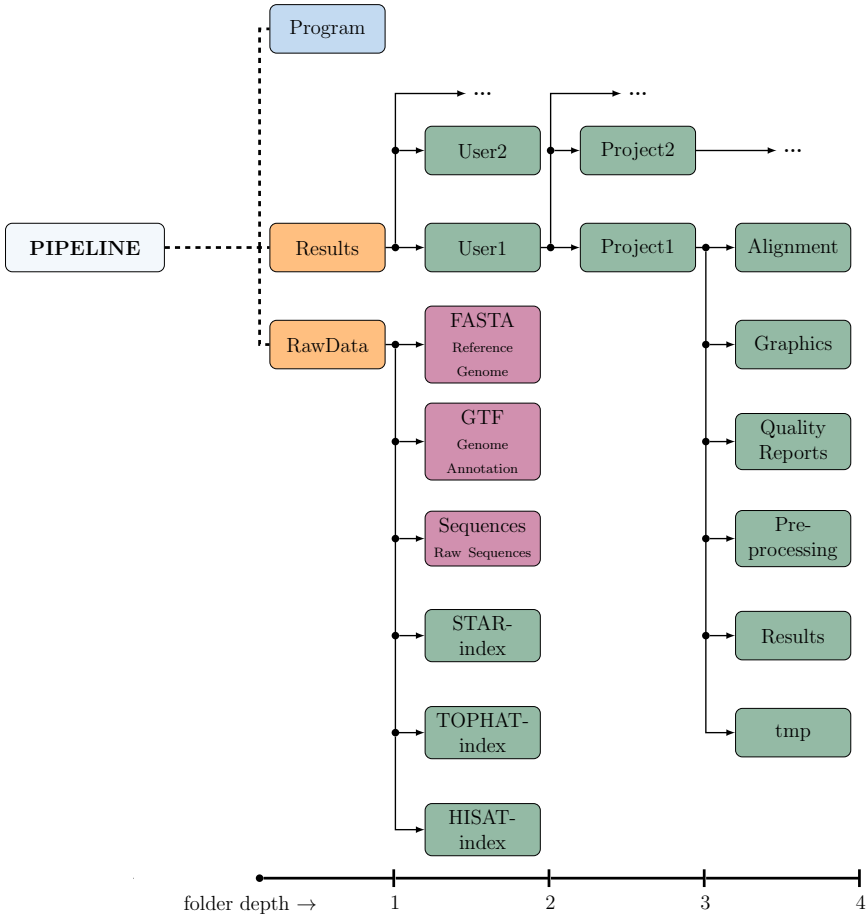


Figure 3.5: ***Directory structure of the CancerRNAseq pipeline.*** *The blue node represents the folder with all the program files of the CancerRNAseq pipeline. The orange nodes illustrate folders of which the path has to be provided by the user and the pink nodes represent folders that have to be within the RawData folder. The green nodes represent folders, which are generated during runtime.*

3.2 Comparison of STAR/TopHat2/HISAT2

To compare mapping STAR, TopHat2 and HISAT2, three datasets were mapped with all three alignment tools are stated. In table 3.1 alignment specific parameters of the three test datasets (simulated RNA-Seq dataset with an overall mismatch rate of 0.5% [37] and two real datasets) for each alignment tool are stated. These parameters have

been derived from the QualiMap [40] report of the alignment generated with the usage of 10 threads. Furthermore, table 3.1 illustrates the alignment specifications for the and for the alignments generated with the real datasets. Figure 3.7b shows the percentage of unmapped reads, whereas figure 3.7a states the percentage of mapped reads for the different datasets. Figure 3.6 illustrates the (ideal) execution time (see formula 2.1) for the mapping process with STAR, HISAT2 and TopHat2 for each dataset. The (ideal) execution time is logarithmically scaled for a better visualization, due to the large differences in execution time between TopHat2 and the two other alignment tools. Figure 3.8 illustrates the distribution of the aligned reads according to their origin (*i.e.,* exonic, intronic or intergenic). Reads aligned to a coding region (exon) are called exonic, reads mapped to a non-coding region (intron) are named intronic and reads aligned between two protein-coding genes are called intergenic [72, 73].

The ideal execution time (see formula 2.1) was used to compare the three mapping tools. As figure 3.6 illustrates, the TopHat2 alignment tool takes up to 30-fold (ERR358485_STAR: 2,518.8s; ERR358485_Tophat: 76,470.4s) the time of STAR and HISAT2 to align the datasets to the reference genome. The smallest difference between the alignment tools can be observed for the execution time of the simulated dataset, but the TopHat2 alignment still takes approximately 11 times longer (SimulatedSequences_STAR: 1,026.9; SimulatedSequences_HISAT2: 1,122.3s; SimulatedSequences_TopHat2: 11,935.4s). Moreover, as shown in figure 3.6, STAR has a slightly better adaption to a higher number of threads than HISAT2. TopHat2 is not able to compete with the two other alignment tools in terms of adaption to a high number of threads, because of the fact that the execution remains approximately the same from upwards of 10 threads.

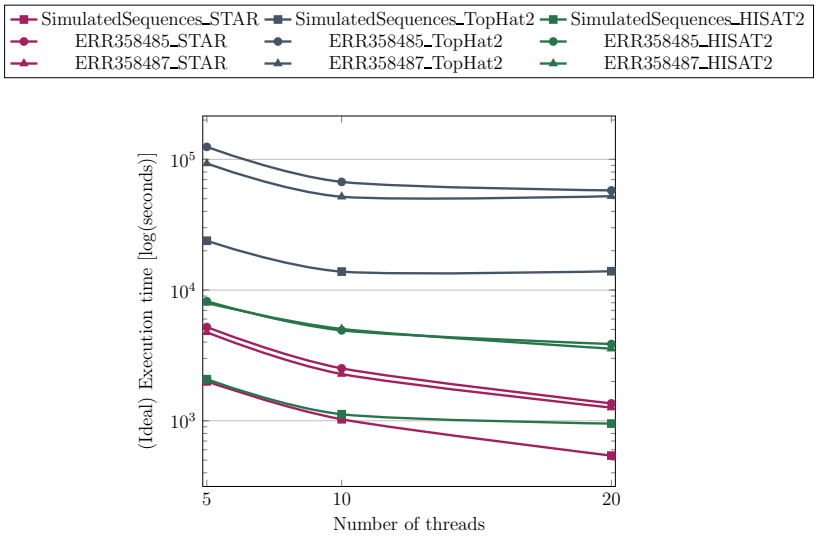
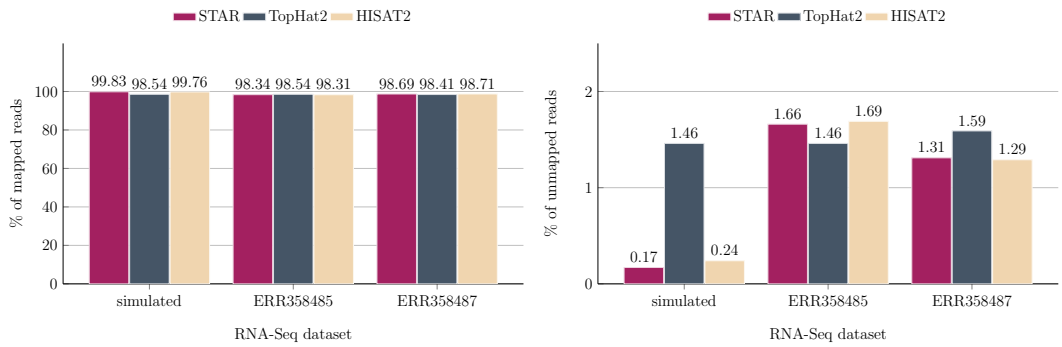


Figure 3.6: Comparison of mapping time between STAR, TopHat2 and HISAT2 for datasets ERR358485, ERR358487 and the simulated sequences

Table 3.1: Information on the alignment of the simulated sequences and datasets ERR358485 and ERR358487 with 10 threads.

	simulated sequences			ERR358485			ERR3584897		
Parameter	**STAR**	**TopHat2**	**HISAT2**	**STAR**	**TopHat2**	**HISAT2**	**STAR**	**TopHat2**	**HISAT2**
# of reads	20,000,000/20,000,000			82,307,086/82,307,086			71,445,283/71,445,283		
# of mapped reads (left/right)	19,965,125/ 19,965,125	19,708,380/ 19,709,430	19,952,605/ 19,951,753	80,922,343/ 80,888,229	80,175,346/ 80,149,546	80,931,570/ 80,894,536	70,518,726/ 70,497,986	70,324,474/ 70,294,754	70,540,234/ 70,507,044
# of aligned pairs (w/o duplicates)	19,965,125	13,662,773	19,930,799	80,735,613	49,784,030	78,824,374	70,420,653	45,970,353	69,268,113
# of alignments	40,759,346	41,348,745	41,078,077	169,946,034	176,610,326	169,243,424	149,133,577	156,181,665	147,916,322
# of secondary alignments	829,096	1,930,935	1,173,719	8,135,462	16,285,434	7,417,318	8,116,865	15,562,437	6,869,044
# of multiple alignments (>2)	1,311,560	2,504,598	1,935,727	13,149,306	21,886,054	12,476,648	12,863,184	20,867,569	11,662,273
# aligned to genes	37,100,118	37,488,310	37,210,126	129,235,495	130,712,825	130,094,448	117,544,083	119,361,471	118,746,891
% Reads exonic origin	96.44%	99.38%	97.5%	84.88%	87.51%	85.47%	89.23%	91.86%	90.16%
% Reads intronic origin	3.42%	0.49%	2.35%	12.31%	10.03%	11.47%	8.55%	6.25%	7.41%
% Reads intergenic origin	0.14%	1.41%	0.15%	2.81%	2.46%	3.06%	2.22%	1.89%	2.9%



(a) Percentage of mapped reads *(b) Percentage of unmapped reads*

Figure 3.7: Percentage of mapped and unmapped reads of the alignments generated by STAR, TopHat2 or HISAT2 for the simulated sequences and the real datasets.

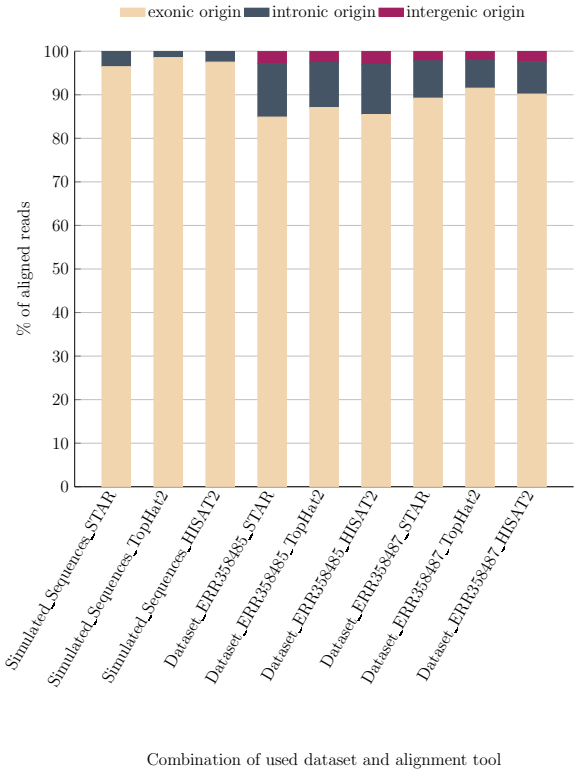


Figure 3.8: Comparison of the percentages of the exonic-, intronic- and intergenic origins of the different datasets and alignment tools.

3.3 Datasets

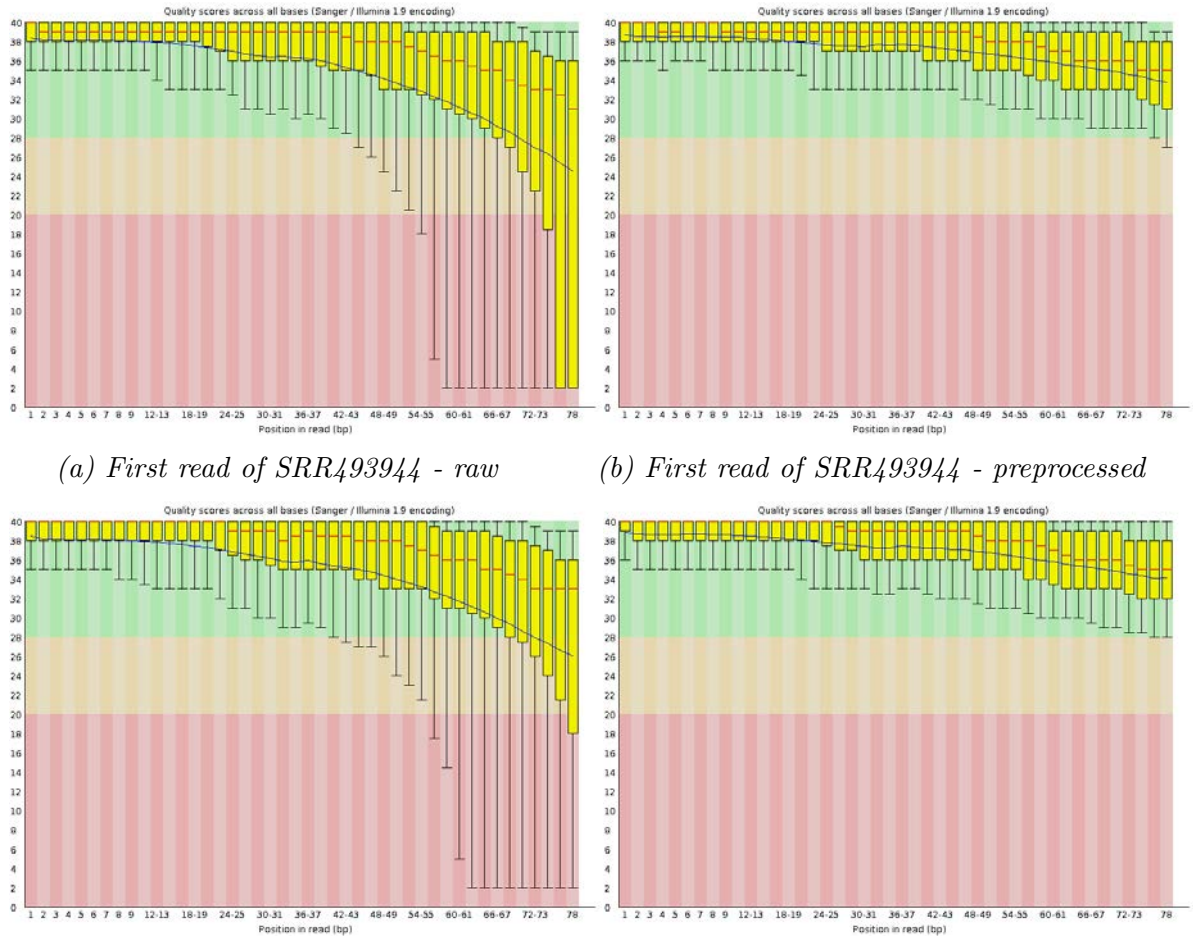
The CancerRNAseq pipeline is the main results of this master thesis. However, to give an insight on how the pipeline works, four example RNA-Seq datasets were used to show how the pipeline can be applied. In section 3.3.1, the results for the first example RNA-Seq dataset (PRJNA163279) are shown in more detail, whereas for the other three example datasets only the essential results are illustrated. Therefore only a fraction of the results the CancerRNAseq pipeline produces in a whole run

is shown. All the generated tables, figures, reports and files from the tested datasets are available on the attached CD. All of the datasets have been generated by using CancerRNAseq with STAR [2] as alignment tool, with the exception that the breast adenocarcinoma dataset has been generated with all of the three available alignment tools (see section 3.3.2).

3.3.1 Dataset: PRJNA163279

The NSCLC example data set (PRJNA163279) for the CancerRNAseq pipeline consists of 6 normal samples *vs.* 6 tumor samples (see table 2.1).

3.3.1.1 Preprocessing



(a) First read of SRR493944 - raw

(b) First read of SRR493944 - preprocessed

(c) Second read of SRR493944 - raw

(d) Second read of SRR493944 - preprocessed

Figure 3.9: FASTQC "per base sequence quality" plot of run SRR493944 from the NSCLC dataset (PRJNA163279) (a,c) before and (b,d) after preprocessing by the CancerRNAseq pipeline

A FASTQC quality report is generated for the raw input files and for each preprocessed file (*i.e.,* CutAdapt, Trimmomatic and rCorrector) to monitor the quality of the input data. The "per base sequence quality" plot (figure 3.9) for sample SRR493944 of the NSCLC dataset (PRJNA163279) before (figure 3.9a & 3.9c) and

after (figure 3.9b & 3.9d) the preprocessing step illustrates that bases at the end of reads tend to have lower quality. The quality trimming step improves the read quality by deleting/trimming sequences with a quality below a certain threshold, leading to a higher quality. This is only an example of one of many quality control plots generated and demonstrates one aspect of improved data quality after the preprocessing step.

For this dataset (6 NSCLC and 6 normal samples) approximately 3.49% (normal: 3.53%, NSCLC: 3.44%) of the sequences from the raw reads were removed during preprocessing. In particular, Cutadapt removed only 0.0044% of sequences for the first run of the normal condition and 0.0032% in case of the first run of the NSCLC condition, whereas 3.96% (normal) and 2.83% (NSCLC) of sequences were removed by Trimmomatic. The last tool in the preprocessing chain, Because rCorrector was developed for correcting errorneous Illumina RNA-Seq reads no reads were deleted during this step. The pipeline was designed to keep as many sequences from the reads as possible to retain the maximum amount of information for the calculation of DE transcripts.

3.3.1.2 Mapping

Based on the mapping, QualiMap creates a plain text-file and a quality report (pdf-format) for each alignment file of the dataset. These files contain alignment-specific parameters, such as the number of aligned reads and the origin of the reads (*i.e.,* exonic, intronic, intergenic and overlapping exons). Figure 3.10 shows the plain text file for the QualiMap report of the aligned run SRR493944 with alignment specific parameters, such as the amount of aligned reads, the genomic origin of the reads and the most common combinations of bases at junctions.

```
RNA-Seq QC report
------------------------------------


>>>>>>> Input

    bam file = Adenocarcinoma_02_STAR_aligned.bam
    gff file = Homo_sapiens.GRCh38.84.gtf
    counting algorithm = uniquely-mapped-reads
    protocol = non-strand-specific


>>>>>>> Reads alignment

    reads aligned = 67,435,451
    total alignments = 81,059,381
    secondary alignments = 13,623,930
    non-unique alignments = 18,213,671
    aligned to genes = 52,044,640
    ambiguous alignments = 2,428,175
    no feature assigned = 8,362,105
    not aligned = 0


>>>>>>> Reads genomic origin

    exonic = 52,044,640 (86.16%)
    intronic = 6,952,116 (11.51%)
    intergenic = 1,409,989 (2.33%)
    overlapping exon = 2,194,552 (3.63%)


>>>>>>> Transcript coverage profile

    5' bias = 0.83
    3' bias = 0.68
    5'-3' bias = 1.16


>>>>>>> Junction analysis

    reads at junctions = 14,246,608

    ACCT : 5.51%
    AGGT : 4.81%
    TCCT : 3.51%
    AGGA : 3.47%
    CCCT : 3.19%
    ATCT : 3.08%
    AGCT : 2.99%
    AGGC : 2.51%
    GCCT : 2.49%
    AGGG : 2.48%
    AGAT : 2.31%
```

Figure 3.10: QualiMap output listing all the alignment specific parameters generated from the aligned run SRR493944

67,435,451 reads, which represented approximately 85.28% of all the the reads, have been aligned in case of the second run of the NSCLC condition to the reference genome

and of these aligned reads 86.16% have an exonic, 11.51% an intronic and 2.33% an intergenic origin (see figure 3.10). The high amount of aligned reads indicates that raw data generation as well as the preprocessing where successful. A low percentage of aligned reads could indicate a problem with the RNA extraction, because the raw data could be, for instance, contaminated with bacterial DNA. Moreover, CancerRNAseq can delete chromosomal DNA contamination in the alignment step. Figure 3.10 shows that the preprocessing had a beneficial impact on the raw input data, because in the junction analysis section there is not an inconsistent 4-base combination. This could be a hint for an adapter sequence that has not been removed from the reads in the preprocessing step.

3.3.1.3 Gene Expression Analysis

Cuffdiff generates a folder containing spreadsheet style files with gene specific values (*e.g.,* counts, FPKM values), but these files are difficult to interpret using a spreadsheet software. Therefore, CummeRbund [47] was used to visualize and filter the Cuffdiff output. All figures generated were created with normalized data to check sample as well as the normalization quality. In this section only six (see figure 3.11) of the nearly 40 plots (including isoforms and replicates) that are generated in the analysis and visualization steps are included (other plots can be found for each dataset on the attached CD).

All annotated transcripts (63,794; where 2,018 (3.16%) of them are significant DE transcripts) were used to generate figures 3.11a-3.11f.

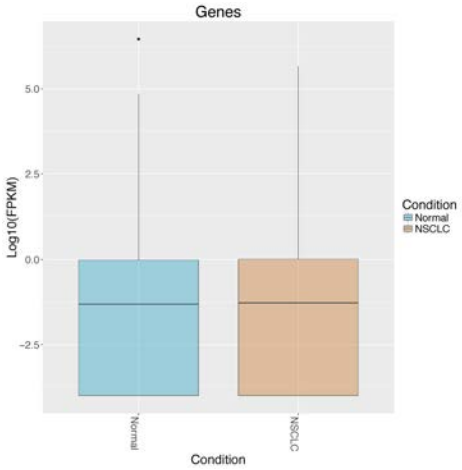
The box plot (figure 3.11a) and the density plot (figure 3.11b) show the distribution of the FPKM values for each condition of the dataset. Figure 3.11a illustrates that the median (of 50% of the transcripts for each condition) of the logarithmic expression level of the NSCLC condition is slightly higher than in the normal condition. This can be also shown in the density plot (figure 3.11b), where the second peak is slightly higher, pointing out that more transcripts of the NSCLC condition are higher expressed than in the normal condition. This type of plot visualizes the density of transcripts for each logarithmic expression value on the y-axis, which means that the integral of the curve for each condition is 1. In general, figure 3.11a as well as figure 3.11b show that the normalization was successful, because both conditions have a similar distribution of the FPKM values.

The full model fit (figure 3.11c) of the Normal and NSCLC samples, where dispersion over count is plotted for each condition, shows only a very small difference between both conditions and for each condition [48]. In this figure variances within the replicates for each condition could be observed, but as figure 3.11c illustrates the dispersion of the transcripts within each condition is relatively small.

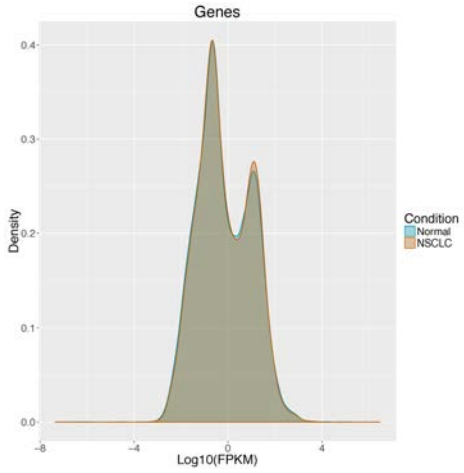
The MA plot (figure 3.11d) shows the relationship between the logarithmic value of the mean (log(A)) and the logarithmic value of the ratios of intensities (log(M)) allowing to see intensity depending trends [74, 75]. The assumption of a good normalization of the dataset can be made, due to the large amount of transcripts located around the y-axis and that they do not drift apart.

The scatter plot (figure 3.11e), illustrating a pairwise comparison of the FPKM values of the Normal and NSCLC samples, shows that most of the transcripts do not have a large difference in expression. The small amount of significant DE transcripts can also be observed with the volcano plot (figure 3.11f), which visualizes the

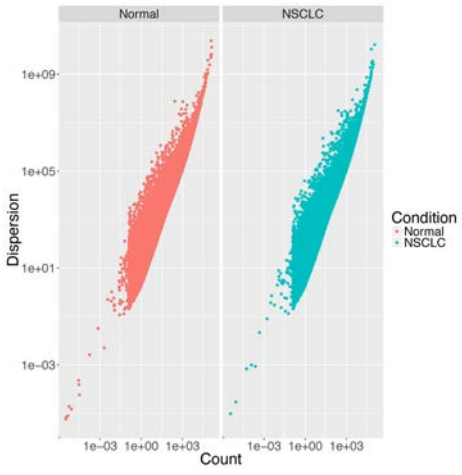
relationship of fold change and p-value of differential expression. Figure 3.11f marks the significant transcripts red and those transcripts can be spotted in the top half of the plot meaning that they have a smaller p-value than the cut-off of 0.05. Furthermore, the volcano plot can be used to spot transcripts that might be interesting for further analysis or, for instance, to check whether the distribution of the transcripts is reasonable. For figure 3.11f the distribution of the transcripts is approximately the same for over-expressed transcripts as well as for under-expressed transcripts, which suggests that the normalization of the dataset was successful. Those significant transcripts are additionally filtered to meet the criteria of an adjusted p-value < 0.05 and a Log2-Fold-Change > 1.



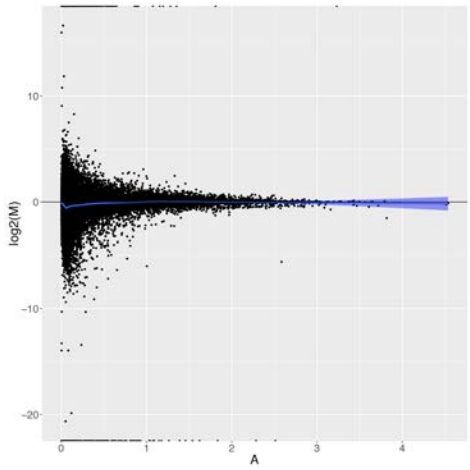
(a) Box plot



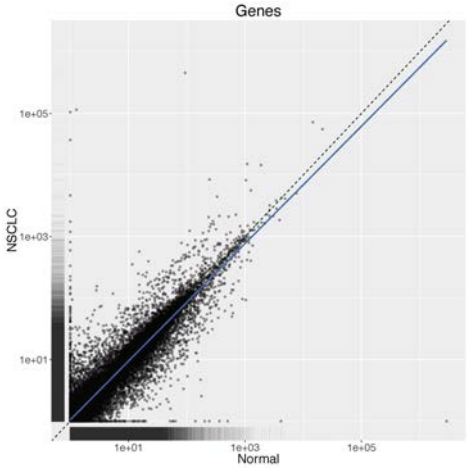
(b) Density Plot



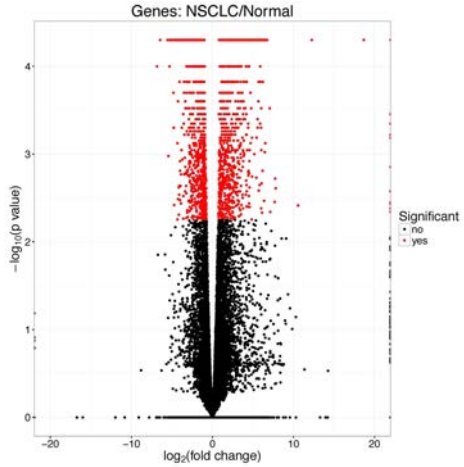
(c) Dispersion Plot



(d) MA Plot



(e) Scatter Plot



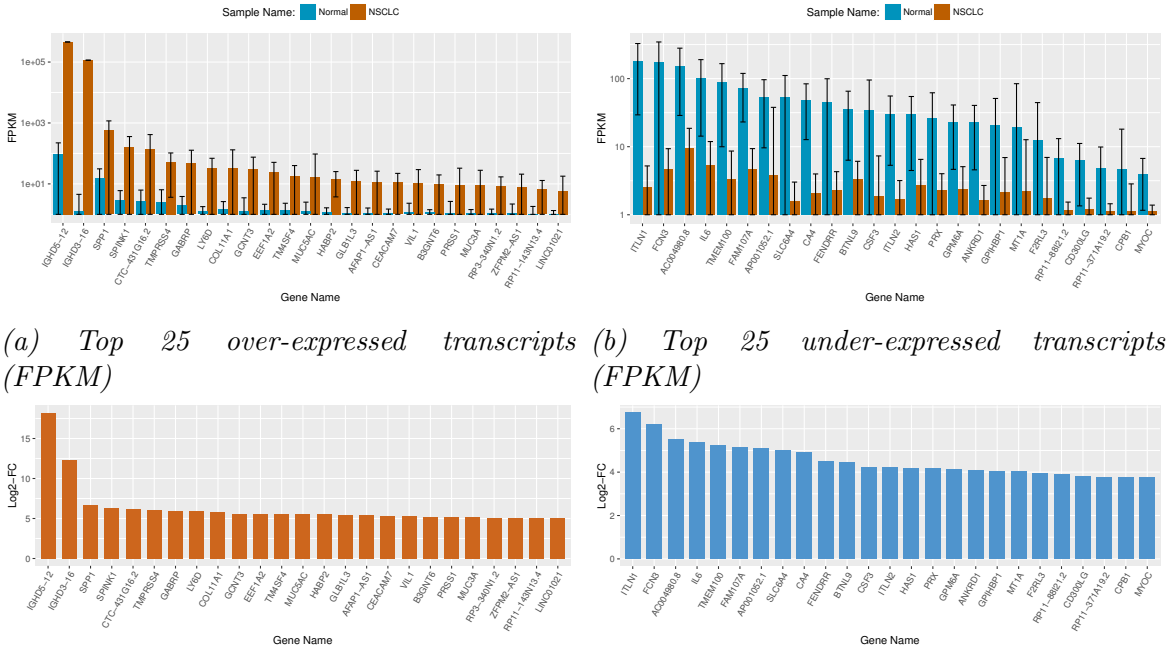
(f) Volcano Plot

Figure 3.11: The (a) box plot and the (b) density plot illustrate the FPKM distribution for the NSCLC and the normal sample of the NSCLC dataset (PRJNA163279). The (c) dispersion plot visualizes the full model fit by plotting the count vs. the dispersion of the two condition. The (d) MA plot shows the relationship between the log of the mean (log(A)) and the log of the ratios of variances (log(M)). The (e) pairwise scatter plot visualizes biases in gene expression by plotting the FPKM values of both conditions. The (f) volcano plot illustrates the relationship between fold change and significance -log_{10}(p-value) of the transcripts in the NSCLC dataset (PRJNA163279).

The top 25 significant DE transcripts were selected according to the highest abs(log2-fold-change-finite). The significant DE are illustrated as bar plots showing either FPKM values (figure 3.12a & 3.12b) or log2-fold-change (figure 3.12c & 3.12d). On the attached CD are bar plots showing FPKM values and log2-fold-change for all tested example datasets (alignment tool: STAR).

In figure 3.12a the top 25 over-expressed transcripts in the NSCLC are shown in descending order according to the FPKM value of the NSCLC samples, whereas figure 3.12b illustrates the FPKM values of the under-expressed transcripts in this cancer type. Figure 3.12d and figure 3.12c show the log2-FC of the top 25 DE expressed transcripts in the same order as figure 3.12a & 3.12b. These figures show that the log2-FC of the top 25 transcripts ranges from approximately 25 (over-expressed) and 6 (under-expressed) to approximately 5. Furthermore, within the top 25 over-expressed transcripts (figure 3.12c and figure 3.12a) transmembrane protease serine 4 (*TMPRSS4*), which is involved in the invasion, metastasis migration and adhesion of cancer cells, can be found (described in detail in section 3.3.5.2) [76, 77].

For a full gene list (including *e.g.,* log2-FC, FPKM values) of the four example cancer datasets see attached CD.



(a) Top 25 over-expressed transcripts (FPKM)

(b) Top 25 under-expressed transcripts (FPKM)

(c) Top 25 over-expressed transcripts (log2-FC)

(d) Top 25 under-expressed transcripts (log2-FC)

Figure 3.12: FPKM and log2-FC values of the 25 highest (abs(Log2-Fold-Change)) over- (a & c) and under-expressed (b & d) transcripts in the NSCLC dataset. The selected transcripts meet the criteria of an adjusted p-value < 0.05 and a Log2-Fold-Change > 1.

3.3.1.4 Expression level of *COL11A1* in the NSCLC dataset

The collagen α-1(XI) chain (COL11A1), which is a procollagen for the minor fibrillar collagen collagen XI [78], is associated with the invasion and metastasis in cancer [6].

Figure 3.13 shows the expression level of the gene *COL11A1*, as an example of an significant DE gene, in the NSCLC dataset. The *COL11A1* gene is approximately 70-fold ($FPKM_{Normal} = 0.433905$, $FPKM_{NSCLC} = 32.2671$) higher expressed in the cancerous condition than in the normal condition. Furthermore, *COL11A1* has been found over-expressed in the primary colorectal cancer example dataset ($FPKM_{Normal} = 0.301622$, $FPKM_{PrimaryColorectalCancer} = 9.14227$), pancreatic ductal adenocarcinomas [78], gastric [79], ovarian [80, 81] and colorectal cancer [80]. *COL11A1* is suggested to be used as a potential biomarker [82, 80] for metastatic NSCLC and as a potential therapeutic target canditate [82].

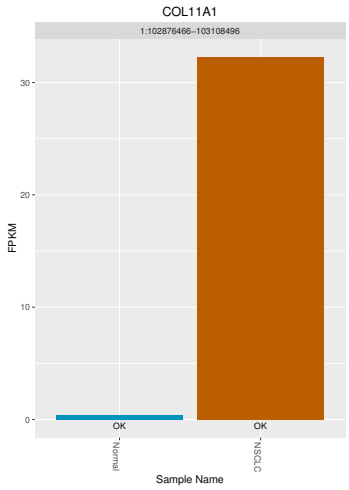


Figure 3.13: Differentially expressed gene COL11A1 of the PRNJA163279 dataset.

3.3.1.5 Comparison to the BioXpress database

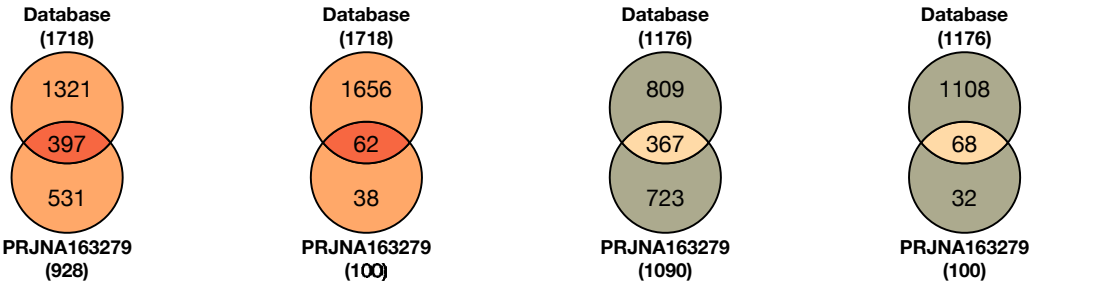
Significant DE transcripts of the *NSCLC* (PRJNA163279) dataset (abs(log2-fold-change) > 1 and an adjusted p-value < 0.05) have been compared with the gene list for this cancer type (adjusted p-value < 0.05) of the BioXpress [6] database.

Table 3.2 lists the total amount of significant DE transcripts as well as the number of over- or under-expressed transcripts. From 6047 significant DE transcripts in the *NSCLC* dataset, 2937 were over-expressed and 3110 were under-expressed in comparison with the normal samples. In figure 3.14 the mutual and unique DE expressed transcripts between the BioXpress database [6] for *NSCLC* and results of the CancerRNAseq pipeline run for the *NSCLC* (PRJNA163279) dataset are shown. The gene list of the BioXpress database [6] for *NSCLC* comprises data of 108 patients. More than half (62 over-expressed & 68 under-expressed) of the top 100 DE transcripts have been found in the BioXpress [6] database as well.

This comparison has been made to show that an overlap between different patients and cancer types exists and, therefore, analysis on large datasets to search for target candidates of cancer therapy as well as cancer diagnosis is promising. The comparison with the database has also been made for the breast cancer dataset (section 3.3.2), the prostate cancer dataset (section 3.3.3) and for the primary colorectal cancer dataset (section 3.3.4).

Table 3.2: Amount of significant DE transcripts and percentage of significant DE transcripts found in the BioXpress [6] database of the non-small cell lung carcinoma (NSCLC) PRJNA163279 dataset with an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05. (Alignment tool: STAR)

Parameter	Number Of transcripts
Significant DE transcripts between NSCLC and normal samples:	2018
Significant over-expressed transcripts in NSCLC samples:	928
Significant under-expressed transcripts in NSCLC samples:	1090
Percentage of over-expressed transcripts found in Database:	42.78%
Percentage of under-expressed transcripts found in Database:	33.67%



(a) ↑DB/NSCLC *(b) ↑DB/NSCLC(Top100)* *(c) ↓DB/NSCLC* *(d) ↓DB/NSCLC(Top100)*

Figure 3.14: Comparison of the significant DE transcripts from the BioXpress database (DB) [6] and the non-small cell lung cancer (NSCLC) (PRJNA163279) dataset generated with the CancerRNAseq pipeline (alignment tool: STAR). Figure (a) & (b) show the mutual transcripts between the DB and the over-expressed transcripts of the NSCLC dataset and figure (c) & (d) show the mutual transcripts between the DB and the under-expressed transcripts of the NSCLC dataset. For figures (b) and (d) 100 transcripts according to the highest abs(log2-fold-change) were selected.

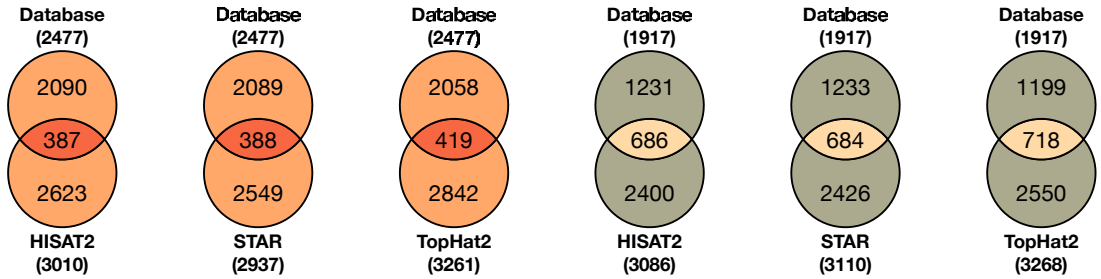
3.3.2 Dataset: PRJEB4829

The breast adenocarcinoma (PRJEB4829) example dataset for the CancerRNAseq pipeline consists of 2 normal samples *vs.* 2 tumor (MCF-7 cell line) samples.

Table 3.3: Amount of significant DE transcripts and percentage of significant DE transcripts found in the BioXpress [6] database with an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05 of the breast adenocarcinoma PRJEB4829 dataset. (Alignment tool: STAR)

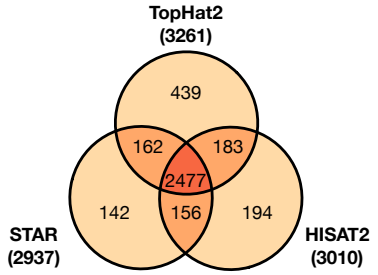
Parameter	Number Of transcripts		
	TopHat2	HISAT2	STAR
Significant DE transcripts between breast adenocarcinoma and normal samples:	6529	6096	6047
Significant over-expressed transcripts in breast adenocarcinoma samples:	3261	3010	2937
Significant under-expressed transcripts in breast adenocarcinoma samples:	3268	3086	3110
Percentage of over-expressed transcripts found in Database:	14.74%	14.75%	15.22%
Percentage of under-expressed transcripts found in Database:	28.16%	28.58%	28.19%

The significant DE transcripts of the breast adenocarcinoma (PRJEB4829) dataset were calculated with alignments generated by each available alignment tool of CancerRNAseq. The selected DE transcripts met the criteria of an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05, which are considered as significant differentially expressed.

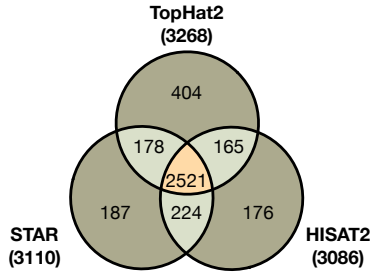


(a) ↑*DB/HISAT2* *(b)* ↑*DB/STAR* *(c)* ↑*DB/TopHat2* *(d)* ↓*DB/HISAT2* *(e)* ↓*DB/STAR* *(f)* ↓*DB/TopHat2*

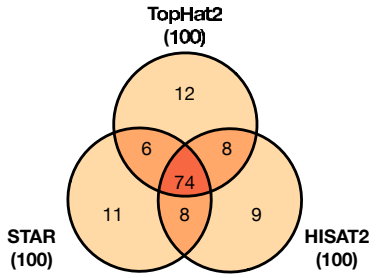
Figure 3.15: Comparison of the significant DE transcripts found in the BioXpress [6] database (DB) and computed by the CancerRNAseq pipeline for the Breast Cancer (PRJEB4829) dataset visualized as Venn diagrams.



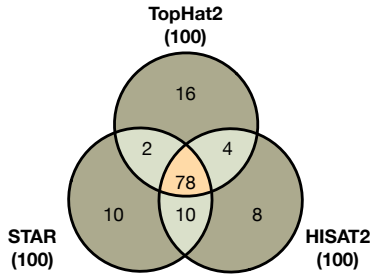
(a) Over-expressed transcripts



(b) Under-expressed transcripts



(c) Top 100 over-expressed



(d) Top 100 under-expressed transcripts

Figure 3.16: Comparison of the significant DE transcripts from the Breast Cancer (PRJEB4829) dataset generated with STAR/TopHat2/HISAT2 as alignment tool. For figure (b) and (d) 100 transcripts according to the highest abs(log2-fold-change) were selected.

Table 3.3 lists the total amount of significant DE transcripts as well as the number of over- or under-expressed transcripts for the breast adenocarcinoma samples in comparison with the normal samples of the PRJEB4829 dataset. In

figure 3.15 the mutual and unique DE expressed transcripts between the BioXpress database [6] for breast adenocarcinoma and results of the CancerRNAseq pipeline run for the PRJEB4829 dataset are shown. The gene list of the BioXpress database [6] for breast cancer comprises data of 114 patients.

In figure 3.15 the overlap between the DE transcripts that have been calculated with each alignment tool of the CancerRNAseq pipeline and the BioXpress database [6] are illustrated as Venn diagrams. The amount of mutual over-expressed transcripts ranges from 387(HISAT2) to 419(TopHat2) (14.75%-15.22%), whereas the overlap of under-expressed transcripts is nearly twice as high and ranges from 684 (STAR) to 718 (TopHat2) (28.16%-28.58%). One aspect of the relatively small overlap between the results and the database might be the small amount of DE transcripts in the database. This difference has to be included in the interpretation of the results, because even the smaller difference in significant DE transcripts in case of the newer STAR and HISAT2 alignment tools is not negligible. One way to improve the bias between alignment tools might be the usage of a larger dataset, because this reduces the difference in significant DE transcripts (see figure 3.15). The difference between the alignment tools may arise from the different approaches of the three alignment tools on how to calculate DE transcripts (for further details see appendix 6.1.0.4-6.1.0.6). Figure 3.16 visualizes that the alignment tools have an impact on the resulting gene list, even if it is relatively low. In fact, 74 (over-expressed) and 78 (under-expressed) transcripts are similar of the 100 highest expressed transcripts of the dataset. DE transcripts based on STAR and HISAT2 alignments (see figure 3.16a-d) resulted in have the highest overlap of DE transcripts for this dataset.

3.3.3 Dataset: PRJEB2449

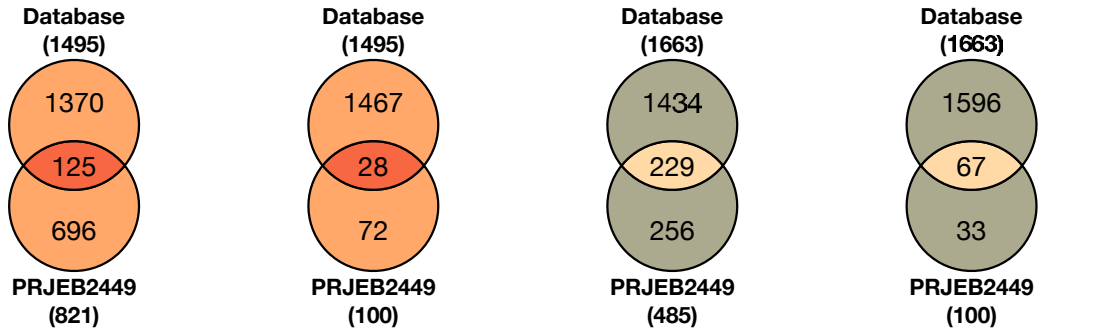
The prostate cancer (PRJEB2449) example dataset for the CancerRNAseq pipeline consists of 14 normal samples *vs.* 14 tumor samples.

The significant DE transcripts of the *Prostate Adenocarcinoma* (PRJEB2449) dataset were calculated with Cuffdiff based on the STAR alignment generated by CancerRNAseq. The selected DE transcripts meet the criteria of an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05 to be considered as significant differentially expressed.

Table 3.4 lists the total amount of significant DE transcripts as well as the number of over- or under-expressed transcripts for the *Prostate Adenocarcinoma* samples in comparison with the normal samples of the PRJEB2449 dataset. Of 1306 significant DE transcripts in the dataset are 821 transcripts over-expressed and 485 transcripts under-expressed.

Table 3.4: Amount of significant DE transcripts and percentage of significant DE transcripts found in the BioXpress [6] database of the prostate adenocarcinoma PRJEB2449 dataset with an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05. (Alignment tool: STAR)

Parameter	Number Of transcripts
Significant DE transcripts between prostate adenocarcinoma and normal samples:	1306
Significant over-expressed transcripts in prostate adenocarcinoma samples:	821
Significant under-expressed transcripts in prostate adenocarcinoma samples:	485
Percentage of over-expressed transcripts found in Database:	15.23%
Percentage of under-expressed transcripts found in Database:	47.22%



(a) ↑*DB/PRAD* *(b)* ↑*DB/PRAD(Top100)* *(c)* ↓*DB/PRAD* *(d)* ↓*DB/PRAD(Top100)*

Figure 3.17: Comparison of the significant DE transcripts from the prostate adenocarcinoma (PRAD, PRJEB2449) dataset generated with the CancerRNAseq pipeline (alignment tool: STAR). For figures (b) and (d), the 100 transcripts with the highest abs(log2-fold-change) were selected.

In figure 3.17 the mutual and unique DE expressed transcripts between the BioXpress database [6] (adjusted p-value < 0.05) for *Prostate Adenocarcinoma* and results of the CancerRNAseq pipeline run for the PRJEB2449 dataset are shown. The gene list of the BioXpress database [6] for prostate cancer comprises data of 51 patients. A total overlap of 125 (15.23%) over-expressed transcripts were found between the dataset and the database. In comparison approximately half (47.22%, 229 transcripts) of the under-expressed transcripts could have been found in the BioXpress database [6] as well. Furthermore, 28 transcripts of the 100 most over-expressed transcripts and 67 transcripts of the 100 most under-expressed transcripts overlap between the CancerRNAseq run and the gene list from the database.

The comparison of the resulting gene list of the prostate cancer dataset with the BioXpress [6] database has been made to show that it might be possible to obtain prostate cancer specific transcripts of large dataset analysis for future cancer diagnosis and therapy.

3.3.4 Dataset: PRJNA218851

The primary colorectal cancer (PRJNA218851) example dataset for the CancerRNAseq pipeline consists of 18 normal samples *vs.* 18 tumor samples.

The significant DE transcripts of the primary colorectal cancer (PRJNA218851) dataset were calculated with the CancerRNAseq pipeline (alignment tool: STAR). The selected DE transcripts meet the criteria of an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05 to be considered as significant differentially expressed.

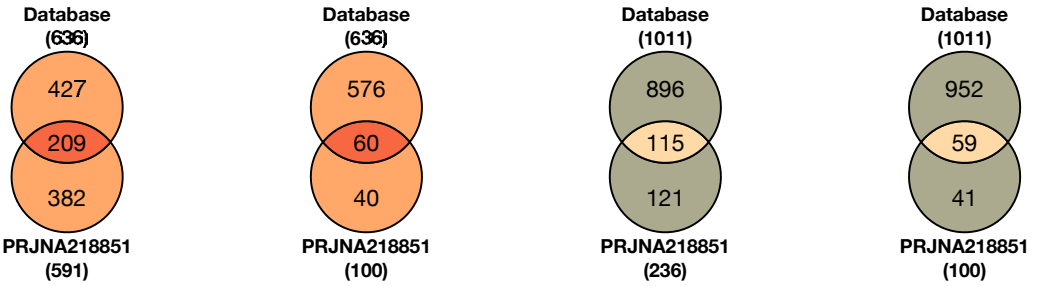
Table 3.5 lists the total amount of significant DE transcripts as well as the number of over- or under-expressed transcripts for the primary colorectal cancer samples in comparison with the normal samples of the PRJNA218851 dataset. In figure 3.18 the mutual and unique DE expressed transcripts between the BioXpress database [6] (adjusted p-value < 0.05) for primary colorectal cancer and results of the CancerRNAseq pipeline run for the PRJNA218851 dataset are shown. The gene list of the BioXpress database [6] for colorectal cancer comprises data of 27 patients.

For this dataset approximately 60% of the most DE transcripts (60 over-expressed transcripts and 59 under-expressed transcripts) could have been found in the database as well. The overlap between all significant DE transcripts of the CancerRNAseq run for the PRJNA218851 dataset and the BioXpress database [6] was 209 in case of over-expressed transcripts and 115 in case of under-expressed transcripts (see figure 3.18).

The comparison of the resulting gene list of the primary colorectal cancer dataset with the BioXpress [6] database has been made to show that it might be possible to obtain primary colorectal cancer specific transcripts, such as *MMP11* that is described in detail in the following section.

Table 3.5: Amount of significant DE transcripts and percentage of significant DE transcripts found in the BioXpress [6] database of the primary colorectal cancer PRJNA218851 dataset with an abs(log2-fold-change) > 1 and an adjusted p-value < 0.05. (Alignment tool: STAR)

Parameter	Number Of transcripts
Significant DE transcripts between primary colorectal cancer and normal samples:	827
Significant over-expressed transcripts in primary colorectal cancer samples:	591
Significant under-expressed transcripts in primary colorectal cancer samples:	236
Percentage of over-expressed transcripts found in Database:	35.36%
Percentage of under-expressed transcripts found in Database:	48.73%



(a) ↑*DB/PCC* *(b)* ↑*DB/PCC(Top100)* *(c)* ↓*DB/PCC* *(d)* ↓*DB/PCC(Top100)*

Figure 3.18: Comparison of the significant DE transcripts from the primary colorectal cancer (PCC, PRJNA218851) dataset generated with the CancerRNAseq pipeline (alignment tool: STAR). For figures (b) and (d), the 100 transcripts with the highest abs(log2-fold-change) were selected.

3.3.4.1 Expression level of *MMP11* in the primary colorectal cancer dataset

Matrix metallopeptidase 11 (MMP11), which is a matrix metalloproteinases and an important part of the tumor stroma, is associated to be an indicator for the development progress of colon cancer [83]. This should be an example gene to give an idea on how CancerRNAseq can be used. Figure 3.19 shows the expression level of the gene *MMP11*, as an example of an significant DE gene, in the primary colorectal cancer dataset. The *MMP11* gene is approximately 10-times higher expressed in the primary colorectal cancer samples than in the normal samples. Furthermore, *MMP11*

is over-expressed in NSCLC and under-expressed in breast cancer of the example datasets (see table 3.6).

Table 3.6: Expression level of (MMP11) in the primary colorectal cancer, NSCLC and breast cancer dataset.

Dataset	$FPKM_{Cancer}$	$FPKM_{Normal}$
Primary colorectal cancer	39.8371	4.08504
NSCLC	71.0304	2.2174
Breast cancer	1.75421	11.5361

MMP11 or stromelysin is considered as a potential target candidate for cancer immunotherapy [83] and tumor marker in serum (see figure 3.19) [84]. Additionally, *MMP11* has been found to be over-expressed in laryngeal squamous cell carcinoma (LSCC) [85], breast cancer [86] and stromal cancer [87]. The expression level of *MMP11* correlates positively with colon cancer progression, which means that the level of expression increases with the progression of the cancer [83]. High expression levels of *MMP11* might correlate with a poor survival for prostatic cancer patients [88] and cholangiocarcinoma patients [89].

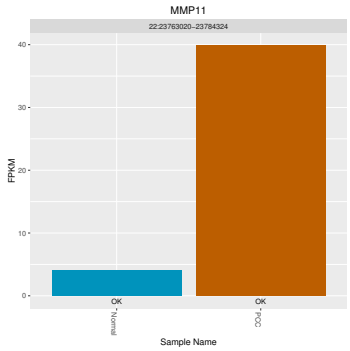


Figure 3.19: Differentially expressed gene MMP11 of the PRJNA218851 dataset.

3.3.5 Comparison between datasets

In this section the analysis results from the four different cancer types (PRJEB4829 - breast adenocarcinoma, PREJB2449 - prostate cancer, PRJNA163279 - NSCLC, PRJNA218851 - primary colorectal cancer) are compared with each other.

3.3.5.1 Mutual DE transcripts of the datasets

In figure 3.20 the mutual transcripts between the four different cancer datasets (alignment tool: STAR) split in over- or under-expressed are illustrated as circos plot (generated with Circos [70] online). In the circos plot the width of the connections are equal to the amount of similar transcripts between two conditions.

The NSCLC and the breast adenocarcinoma (PRJNA163279 over-expressed & PRJEB4829 over-expressed and PRJNA163279 under-expressed & PRJEB4829 under-expressed) have the highest amount of mutual DE transcripts. There are 205 over-expressed transcripts and 511 under-expressed transcripts similar in both datasets. Furthermore, the breast adenocarcinoma and the prostate cancer

Figure 3.20: ***Mutual significant DE transcripts between the four different example cancer datasets.*** *The thickness of the ribbons connecteing the different groups of transcripts can be interpreted as the amount of mutual significant DE transcripts between those connected. For each dataset a group of over- and under-expressed transcripts was established (created with Circos [70] online). PRJEB4829 - breast adenocarcinoma, PRJEB2449 - prostate cancer, PRJNA163279 - NSCLC (non-small cell lung carcinoma), PRJNA218851 - primary colorectal cancer.*

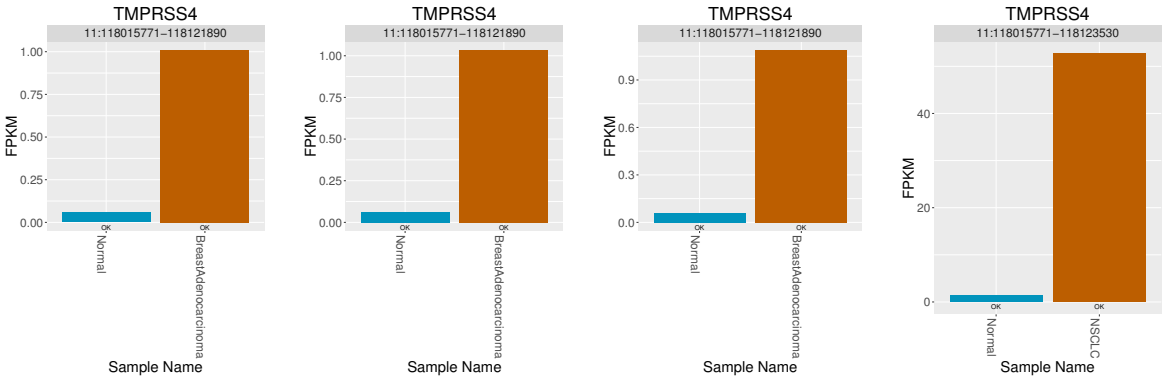
(PRJEB2449) dataset have also a high amount of DE transcripts in common (over-expressed: 81; under-expressed: 188). See table 6.1 (appendix 6.3.0.1) for detailed information on mutual DE transcripts between the tested datasets.

Moreover, the overlap of transcripts between the same direction type of expression (over & over and under & under) of different dataset is higher than the amount of mutual transcripts between different kind of expressions (over-expressed & under-expressed).

3.3.5.2 Expression level of *TMPRSS4* in the datasets

Transmembrane protease serine 4 (TMPRSS4) [90] was determined as over-expressed in various example datasets including breast cancer (PRJEB482), primary colorectal cancer (PRJNA218851) and NSCLC (PRJNA163279) (see figure 3.21). *TMPRSS4* is involved in the invasion, metastasis migration and adhesion of cancer cells [76, 77]. Moreover, *TMPRSS4* correlates with a poor survival rate of patients [91] and has been found over-expressed in prostate cancer [92], colorectal cancer [90], small cell lung cancer (SCLC) [93], NSCLC [93], gallblader cancer [94] and gastric cancer [95] confirming the results of CancerRNAseq. Furthermore, a correlation between the

expression level of *TMPRSS4* and the development stage of colorectal cancer was discovered [76].



(a) PRJEB4829 (HISAT2) *(b) PRJEB4829 (STAR)* *(c) PRJEB4829 (TopHat2)* *(d) PRJNA163279 (STAR)*

Figure 3.21: Expression levels as FPKM values of the over-expressed gene TMPRSS4 (transmembrane protease serine 4) as well as TMPRSS3 (paralog of TMPRSS4) in cancer samples generated with the CancerRNAseq pipeline for the breast cancer dataset (PRJEB4829) based on either (a) HISAT2, (b) STAR or (c) TopHat2 alignment and (d) for the NSCLC (non-small cell lung carcinoma) dataset (PRJNA163279).

Figures 3.21a-3.21d show the various expression levels of transmembrane protease serine 4 (*TMPRSS4*) in the breast adenocarcinoma and the NSCLC [96, 97].

The fold-change of the significant DE gene *TMPRSS4* ranges from 16.75 (figure 3.21b; alignment tool: STAR) to 18.51 (figure 3.21c; alignment tool: TopHat2) in the breast cancer dataset (PRJEB4829). In case of the NSCLC dataset (PRJNA163279) the expression level in the cancerous condition is approximately 37-fold higher than in the normal condition (see figure 3.21d).

Discussion

4.1 CancerRNAseq - Pipeline for RNA-Seq Data Analysis

In the last few years RNA-Seq has become the standard method for gene expression analysis and has nearly replaced the commonly used microarrays [98, 4, 99]. RNA-Seq produces comparable [100, 101, 4] or even better [98, 102] results than microarrays, but has to overcome the higher costs for simple gene expression studies [4].

The gain in popularity lead to an increase of publications and releases of tools for preprocessing, mapping (to a reference genome or for *de novo* assembly) and gene expression estimation of RNA-Seq data [103]. However, most tools that have been released have been developed for only one specific task of the RNA-Seq workflow (*e.g.*,, Cutadapt [32] for removing adapter sequences from the reads) [9]. Therefore, the establishment of an RNA-Seq pipeline is essential so that the tools needed for a whole RNA-Seq workflow (see figures 3.1-3.3) can be automatically executed consecutively. Furthermore, an other important advantage of an analysis pipeline is the small effort that is needed to perform the RNA-Seq analysis of different datasets with the same parameters.

There are only a few tools available that can be used for more than just a specific task, such as the RNA-Seq alignment app from the Illumina, Inc. platform Basespace [104], RAP [105], FX [106] and Myrna [107]. The Basespace platform has implemented TopHat2 or STAR for the mapping and Cufflinks2 for FPKM estimation [104] and has the strict Illumina FASTQ file standard as requirement for input sequences [108]. The input has to be uploaded to the platform, which is a time-consuming aspect in case of the usually large amount of data in RNA-Seq experiments [104, 109]. A disadvantage of such online platforms, is that all the datasets have to be uploaded to the server, which can be an issue in terms of confidential patient data or company projects. The Galaxy platform [110] incorporates more than 650 tools for NGS data. For RNA analysis the Galaxy platform has also HISAT, TopHat, Cufflinks, Cuffmerge, Cuffdiff and CummeRbund included. They also added Trimmomatic as new NGS data QC and manipulation tool in Galaxy Version 0.32.3. More specialized data analysis options are not included in those online platforms, whereas the CancerRNAseq pipeline can and will be integrated in a large meta-pipeline. This large meta-pipeline will include non-standard analysis techniques, which are not available in the pipelines such as allele-specific and antisense analysis approaches. This large meta-pipeline will mainly focus on the analysis of cancer datasets.

An in-house RNA-Seq pipeline allows full control of the version of tools used, of having only a few restrictions on input data types and also allows to reproduce

every single step. Moreover, in case of private RNA-Seq datasets, the user is in full control of where the input data as well as the results are stored and who has access to them. Due to the design of CancerRNAseq it is possible to try out new versions of programs by only changing one parameter in the configuration file (path to the new version of the tool). The pipeline has been developed to be executed on an Unix based machine, because it would be very ineffective to use the pipeline not on a server/cluster, due to the large amount of data as well as CPU-intensive calculations. Another advantage of an in-house pipeline in comparison with a cloud-based alternative is that there are hardly any specific requirements for input data, except that they have to be in fastq-format and use a PHRED33 or PHRED64 quality score. The quality score of PHRED33 ranges from 0 to 93 (ASCII 33 to 126), whereas the quality score of PHRED64 (ASCII 64 to 126) from 0 to 64, leading to a lossy conversion between those two [25]. It is possible to stop after each pipeline step to check intermediate results and, if needed, to adjust the configuration file. The CancerRNAseq pipeline has been established fully modular in terms of preprocessing, mapping and FPKM/RPKM estimation tools. The pipeline is able to handle both single and paired-end reads from RNA-seq datasets, *e.g.,* publicly available cancer RNA-Seq datasets such as the example datasets (see table 2.1). The ability to easily change, update or add tools to the pipeline is a huge advantage. All of the generated plots and reports, except the FASTQC report in html format, are saved in PDF format to be readable platform independent. Moreover, the pipeline outputs csv-files of the DE transcripts that can be easily further analyzed via a spreadsheet software or with R. A comparison between results of several pipeline runs can be easily made for a gene dataset.

To implement the first version of CancerRNAseq certain tools were selected based on the reasons given below. More tools will be added in time. Cutadapt was used for adapter trimming and Trimmomatic was used for quality filtering of the input sequences, because of the high amount of mapped reads after filtering [111]. The only tools that were specifically designed for the correction of sequencing errors for RNA-Seq datasets are SEECER [30] and rCorrector [34], and due to the slightly higher performance [34] and easier usage rCorrector was chosen. TopHat2 was used as an alignment tool for the CancerRNAseq pipeline, because of the low memory requirement of its alignment engine and because it was the workhorse for RNA-Seq analysis for years[1]. The new version of TopHat2, TopHat3, will use HISAT2 as core alignment engine and Trapnell suggests to use HISAT2 instead of TopHat2 - therefore we chose this aligner as well. STAR was chosen as it has a completely different alignment approach (see appendix 6.1.0.4 for more discuss on the alignment tools). Cuffdiff was chosen, because it reduces the bias in differential expression calculation by considering isoforms compared to tools such as DESeq [45] or edgeR [42] [1]. Furthermore, the output of Cuffdiff can be directly visualized with the R-package CummeRbund [1]. Therefore, the whole Cufflinks suite of tools (Cufflinks, Cuffmerge, Cuffdiff and CummeRbund) was used.

Because of the fact that usually RNA-Seq experiments consist of many transcripts with a very low coverage it is possible to get very high log2-FC from the FPKM/RPKM values that are close to zero [69]. For a better DE gene selection an edited log2-FC with a cutoff of 0.1, that is added to the FPKM/RPKM values, is used for gene prioritization. Warden *et al.* found out that a cutoff between 0.01 and 1 can be applied without respect to the number of reads or samples [69]. The DE transcripts met the criteria of an adjusted p-value < 0.05 and a Log2-Fold-Change > 1, because

it has been shown that filters (*e.g.,* p-value, fold-change and small expression levels) applied to the gene lists increases the comparability of gene lists between platforms as well as the accuracy of reported DE transcripts [112].

CancerRNAseq runs of the same dataset with different alignment tools produce a different amount of significant DE transcripts, for the breast cancer dataset the difference was up to 324 (STAR vs. TopHat2) transcripts. The difference in the significant DE transcripts between the three alignment tools (see figure 3.16) may arise, among others, from the different approaches on how to perform a spliced-read alignment, which leads to different mapping accuracy. The different mapping accuracy of alignment tools was also shown by Engström *et al.* [113], who included TopHat and STAR in his evaluation of spliced alignment tools.

Numerous factors can influence an RNA-Seq analysis. Not only the tools used in an RNA-Seq workflow have an influence on the reported DE transcripts, but also the NGS platform used to generate the input reads [114]. Another important factor is the number of replicates for each condition, because Seyednasrollah *et al.* suggests that results generated with fewer than 5 replicates for each condition should be interpreted with caution [115].

4.1.1 Outlook

The CancerRNAseq pipeline can be used to examine the huge amount of publicly available as well as own RNA-Seq datasets. Moreover, a web-interface for a simple execution of the pipeline via the browser will be added in the future, which gives researchers the opportunity to easily obtain DE transcripts of their datasets. Due to the modular configuration it is possible to add *e.g.* TopHat3, which uses the available HISAT2 as the alignment engine, as additional alignment tool to the CancerRNAseq pipeline as well as additional command line tools (*e.g.,* tools for statistical analysis such as limma [46]) [37]. The CancerRNAseq pipeline will also be the first step towards an online meta-analysis tool for publicly available cancer RNA-Seq data, similarly to meta-analysis tools for public EST data [116] and public cancer microarray data [117] available at the Institute. Furthermore, the CancerRNAseq pipeline will be extended to include downstream analysis, such as visualization of the analysis results (*e.g.,* circos plots), and further analysis of the resulting gene lists (*e.g.,* GO analysis). It will also be incorporated in a larger meta-pipeline for specialized comparative transcriptomics and (epi)genomics analyses.

4.2 Example Datasets Results and Comparisson of Results with the BioXpress database

The second aim of my master thesis was the testing of the pipeline, with a comparison between the pipeline results of four different example RNA-Seq datasets (see table 2.1) and results derived from the BioXpress [6] database. This aim has been accomplished with four cancer RNA-Seq datasets for breast, lung, prostate and colon cancer as example datasets. The list of DE transcripts for each dataset has been compared with the over-expressed and under-expressed transcripts for each cancer type from the BioXpress database [6] to determine the overlap. That only approximately 21% of significant DE transcripts are mutual between the dataset and

the BioXpress database may arise from the fact that Wan *et al.* used a lot of datasets to generate the list of significant over- and under-expressed transcripts and that a lot of transcripts of the example datasets could be system-specific DE transcripts. As discussed in the previous section, the resulting gene list will depend on a lot of factors such as the biological samples, sequencing technologies and analysis tools used. Therefore, a relatively small overlap is reasonable. Furthermore, the significant DE transcripts from the CancerRNAseq pipeline meet the criteria of an adjusted p-value < 0.05 and a Log2-Fold-Change > 1, whereas the gene list from the BioXpress database only meet the criteria of an adjusted p-value < 0.05. The gene list from BioXpress does not provide the information on the log2-Fold-Change (since the update to BioXpress v2.0, which has been released after the completion of the analysis, the database includes the log2-Fold-Change), because Wan *et al.* tested between 27 and 114 patients for each cancer type and, therefore, not every gene was DE in each patient [6]. This confirms the relatively small overlap between the dataset and the BioXpress gene list as reasonable.

TMPRSS4, *MMP11* and *COL11A1* have been further analyzed, because they are known to be DE in certain cancer types and are considered as potential target canditates for personalised cancer therapy [118, 83, 80, 82]. This is particularly important as cancer treatments have the lowest success rate for a one-fits-all therapy approach, which means that only one of four patients shows an effective drug response [119].

To compare all tested cancer types a CircosPlot (figure 3.20) showing the amount of mutual transcripts between the datasets has been created. In general, the overlap between DE transcripts of the same direction type of expression (*i.e.,* over-expressed and under-expressed) is the largest. Moreover, the amount of similarity between cancer types might be influenced by the tissue the cancer originates, the cancer pathway itself, the similarity of expression patters of the tissue of origin and if similar pathways are deregulated. Figure 3.20 shows that the breast cancer (PRJEB4829) and the NSCLC (PRJNA163279) dataset have the most over-expressed and under-expressed transcripts in common. Wan *et al.* also found through clustering BioXpress data that breast and lung cancer have similar expression patterns [6]. Furthermore, the comparison of gene expression patterns of various tumor types by Martinez *et al.* identified that breast and lung cancer have similar expression patterns [120].

The same type of cancer can be genetically different, because cancerous cells develop from multiple genetic defects [121]. The conventional treatment for cancer depends on the organ or tissue the tumor is located and on the TNM classification [56, 121]. Genetic differences between cancers in the same organ or tissue show a distinct respond to specific cancer therapy and the main challenge is to identify the appropriate cancer therapy for the individual [121]. With the personalized medicine approach, which considers genetic differences of distinct cancer patients, an appropriate therapy for each patient can be chosen. This approach depends strongly on the discovery and development of biomarkers with the capability to genetically classify a tumor [121]. For example, in 2016 the European Commission and in the United States the Food and Drug Administration (FDA) have approved the first biomarker-driven therapy, called XALKORI® (Crizotinib) from Pfizer, for *ROS1*-positive and anaplastic lymphoma kinase (ALK)-positive advanced NSCLC and additionally for metastatic ROS1-positive NSCLC by the FDA. For this, in approximately 1% of NSCLC occurring difference in the ROS1 gene, hardly any treatment options were available [122].

One way to improve the chance that biomarkers reach clinical use is to use a thorough study design [123], such as McShane *et al.*, who have developed reporting guidelines to improve the quality of tumor marker studies and to provide a list of editorial requirements [124]. In comparison with only a single biomarker a combination of biomarkers leads to an increase in diagnostic specificity as well as sensitivity [125]. Furthermore, five phases for biomarker development have been proposed by Pepe *et al.* [126].

4.3 Alignment Tools Comparison

In comparison to HISAT2 and STAR, TopHat2 lacks in terms of speed (figure 3.6). As it can be seen in table 3.1, the alignment rates of the three alignment tools are comparable between each other. Furthermore, TopHat2 was not able to align as many pairs as STAR and HISAT2, but STAR and HISAT2 produce comparable results (figure 3.6). This could be due to the fact that the inner distance and the standard deviation has to be specified in case of TopHat2 and deviations from true values could result in more broken pairs. Another drawback in case of TopHat2 is that finding the right inner distance requires additional tools. As it can be seen in table 3.1, the TopHat2 alignment tool has the lowest number of aligned pairs, but the highest amount of total alignments for the two real datasets (ERR358485 and ERR358487). The slightly higher rate of aligned reads for the TopHat2 alignments may origin from a higher quantity of secondary and multiple alignments and/or could simply be due to a different parameter choice. Figure 3.8 shows the percentage of reads of either exonic-, intronic- or intergenic origin, which is approximately the same for STAR and HISAT2 alignments compared to TopHat2 alignments, where the amount of reads with exonic origin is slightly higher. Furthermore, the percentage of aligned reads to the reference genome for each mapping tool depends also on the amount of trimming the raw input reads [111].

The STAR aligner outperforms the other two in terms of speed, but at some point the execution time over number of threads will level off asymptotically for all algorithms. However, figure 3.6 show that the STAR algorithm exhibits the best execution time profile, with the steepest gradient and the lowest offset. Hence, the correlation of number of threads and (ideal) execution time is not as good for HISAT2 and grandly inferior for TopHat2, compared to STAR.

Considering all aspects, the STAR algorithm is the best suitable algorithm, especially when taking advantage of a high number of threads. Nevertheless, HISAT2 is also a very fast and good alternative. The TopHat2 algorithm is not as accurate, among others it is too slow and it has been shown that the algorithm is not that effective for using a higher number of threads. This is also confirmed by Kim D. and Salzberg S. [127], they suggested to use HISAT2 instead of TopHat2, because it is outdated and it has entered a low support stage as well as a low maintenance stage.

Conclusion

In the last few years RNA-Seq is steadily replacing microarrays for gene expression analysis. The main goal of this master thesis was the development of an RNA-Seq pipeline. This project was characterized by the two main aims: (i) the establishment of an RNA-seq analysis pipeline to enable the analysis of publicly available cancer RNA-seq datasets (including visualization of the results) and (ii) the application of the pipeline to publicly available cancer RNA-Seq datasets.

The first aim of this master thesis was the establishment of an RNA-seq analysis pipeline. This aim has been achieved with the CancerRNAseq pipeline consisting of three different modules: (i) preprocessing, (ii) mapping and (iii) analysis and visualization.

The second aim of this master thesis was the application of the pipeline to publicly available cancer RNA-Seq datasets. The pipeline was tested with four different cancer RNA-Seq datasets (PRJEB4829 - breast adenocarcinoma, PREJB2449 - prostate cancer, PRJNA163279 - NSCLC, PRJNA218851 - primary colorectal cancer). The results for each dataset and for the reported significant DE transcripts show a high similarity with already published results. The significant DE transcripts of each cancer type contain specific transcripts known to be DE in the analyzed cancer type. Furthermore, the results of CancerRNAseq have approximately 20% of DE transcripts in common with the BioXpress database for the specific cancer type.

In conclusion, the master thesis has met the initial aims as defined in for this project (cf. chapter 1.5).

References

[1] Korpelainen E, Tuimala J *et al.*: *RNA-seq Data Analysis: A Practical Approach*. Taylor & Francis Inc, 2014.

[2] Dobin A, Davis CA *et al.*: **STAR: Ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013. **29**(1): 15–21.

[3] RNA-SEQ: THE FIRST WAVE OF PAPERS. `https://blog.sbgenomics.com/rna-seq-the-first-wave/`, Date accessed: 2016-05-02.

[4] Marioni JC, Mason CE *et al.*: **comparison with gene expression arrays RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays** 2008. 1509–1517.

[5] Wang L, Feng Z *et al.*: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data**. *Bioinformatics* 2010. **26**.

[6] Wan Q, Dingerdissen H *et al.*: **BioXpress: An integrated RNA-seq-derived gene expression database for pan-cancer analysis**. *Database* 2015. **2015**(11): 1–13.

[7] Feichtinger J: *Development of a bioinformatic analytical approach to identify novel human cancer testis Gene candidates*. Ph.D. thesis, Bangor University, 2012.

[8] Metzker ML: **Sequencing technologies - the next generation.** *Nature reviews. Genetics* 2010. **11**(1): 31–46.

[9] Oshlack A, Robinson MD *et al.*: **From RNA-seq reads to differential expression results.** *Genome biology* 2010. **11**(12): 220.

[10] Buermans H and den Dunnen J: **Next generation sequencing technology: Advances and applications**. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2014. **1842**(10): 1932–1941.

[11] Mardis ER: **Next-Generation DNA Sequencing Methods**. *Annual Review of Genomics and Human Genetics* 2008. **9**(1): 387–402.

[12] Margulies M, Egholm M *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005. **437**(7057): 376–380.

[13] Ju J, Kim DH *et al.*: **Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators**. *Proceedings of the National Academy of Sciences* 2006. **103**(52): 19635–19640.

[14] Valouev A, Ichikawa J *et al.*: **A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning**. *Genome Research* 2008. **18**(7): 1051–1063.

[15] Flusberg BA, Webster D *et al.*: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nature methods* 2010. **7**(6): 461–465.

[16] Sanger F, Nicklen S *et al.*: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977. **74**(12): 5463–5467.

[17] Mitra R: **In situ localized amplification and contact replication of many individual DNA molecules**. *Nucleic Acids Research* 1999. **27**(24): 34e–34.

[18] Liu L, Li Y *et al.*: **Comparison of next-generation sequencing systems.** *Journal of biomedicine & biotechnology* 2012. **2012**: 251364.

[19] Van Dijk EL, Auger H *et al.*: **Ten years of next-generation sequencing technology**. *Trends in Genetics* 2014. **30**(9).

[20] Pushkarev D, Neff NF *et al.*: **Single-molecule sequencing of an individual human genome**. *Nat Biotech* 2009. **27**(9): 847–850.

[21] Bowers J, Mitchell J *et al.*: **Virtual terminator nucleotides for next-generation DNA sequencing**. *Nat Meth* 2009. **6**(8): 593–595.

[22] Timp W, Mirsaidov UM *et al.*: **Nanopore sequencing: Electrical measurements of the code of life**. *IEEE Transactions on Nanotechnology* 2010. **9**(3): 281–294.

[23] Illumina images and multimedia for general use. `http://www.illumina.com/company/news-center/multimedia-images.html`, Date accessed: 2016-10-19.

[24] Bentley DR, Balasubramanian S *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008. **456**.

[25] Cock PJA, Fields CJ *et al.*: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Research* 2010. **38**(6): 1767–1771.

[26] Kim SC, Jung Y *et al.*: **A High-Dimensional, Deep-Sequencing Study of Lung Adenocarcinoma in Female Never-Smokers**. *PLoS ONE* 2013. **8**(2).

[27] Ewing B, Hillier L *et al.*: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome research* 1998. **8**(3): 175–85.

[28] Ewing B and Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome research* 1998. **8**(3): 186–94.

[29] Schmieder R and Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011. **27**(6): 863–864.

[30] Le HS, Schulz MH *et al.*: **Probabilistic error correction for RNA sequencing**. *Nucleic Acids Research* 2013. **41**(10): 1–11.

[31] Schmieder R, Lim YW *et al.*: **TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets.** *BMC bioinformatics* 2010. **11**: 341.

[32] Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis* 2011.

[33] Bolger AM, Lohse M *et al.*: **Trimmomatic: A flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014. **30**(15): 2114–2120.

[34] Song L and Florea L: **Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads**. *GigaScience* 2015. **4**(1): 48.

[35] Yang X, Liu D *et al.*: **HTQC: a fast quality control toolkit for Illumina sequencing data.** *BMC bioinformatics* 2013. **14**(1): 33.

[36] Trapnell C, Pachter L *et al.*: **TopHat: Discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009. **25**(9): 1105–1111.

[37] Kim D, Langmead B *et al.*: **HISAT: a fast spliced aligner with low memory requirements**. *Nature Methods* 2015. **12**(4): 357–360.

[38] Wu TD and Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics* 2010. **26**(7): 873–881.

[39] Kim D, Pertea G *et al.*: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome biology* 2013. **14**(4): R36.

[40] Okonechnikov K, Conesa A *et al.*: **Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data**. *Bioinformatics* 2015. **32**(October 2015): btv566.

[41] Li H, Handsaker B *et al.*: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009. **25**: 2078—-2079.

[42] Robinson MD and Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010. **11**.

[43] Trapnell C, Williams Ba *et al.*: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature biotechnology* 2010. **28**(5): 511–515.

[44] Trapnell C, Hendrickson DG *et al.*: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nature biotechnology* 2013. **31**(1): 46–53.

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

[46] Ritchie ME, Phipson B *et al.*: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Research* 2015. **43**(7): e47–e47.

[47] Goff LA, Trapnell C *et al.*: **cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data**. *R Package Version 2.2* 2013.

[48] Love MI, Huber W *et al.*: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biology* 2014. **15**(12): 550.

[49] Rogé X and Zhang X: **RNAseqViewer: Visualization tool for RNA-Seq data**. *Bioinformatics* 2014. **30**(6): 891–892.

[50] Statistik Austria: **Jahrbuch der Gesundheitsstatistik 2014**. Technical Report, 2015.

[51] Hanahan D and Weinberg RA: **Hallmarks of cancer: The next generation**. *Cell* 2011. **144**(5): 646–674.

[52] Zhu X and Wetta H: **Genetics and epigenetics in tumorigenesis: acting separately or linked**. *Austin J Clin Med* 2014. **1**(3): 1–5.

[53] Ling H and Calin GA: **Cancer Genomics**. *Cancer Genomics* 2014. 435–447.

[54] Hanahan D and Weinberg RA: **The Hallmarks of Cancer**. *Cell* 2000. **100**(1): 57–70.

[55] Greene FL and Sobin LH: **The staging of cancer: a retrospective and prospective appraisal.** *CA: a cancer journal for clinicians* 2008. **58**(3): 180–190.

[56] Fraser M, Berlin A *et al.*: **Cancer Genomics**. *Cancer Genomics* 2014. 233–245.

[57] Lavery A, Kirby RS *et al.*: **Prostate cancer**. *Medicine* 2016. **44**(1): 47–51.

[58] Fearon ER: *Colon and Rectal Cancer*, fourth edi edn. 3. Elsevier Inc., 2015.

[59] Lemjabbar-Alaoui H, Hassan OU *et al.*: **Lung cancer: Biology and treatment options.** *Biochimica et biophysica acta* 2015. **1856**(2): 189–210.

[60] Bydoun M and Marcato P: **Cancer Genomics**. *Cancer Genomics* 2014. 213–232.

[61] Leinonen R, Akhtar R *et al.*: **The European nucleotide archive**. *Nucleic Acids Research* 2011. **39**(SUPPL. 1): 44–47.

[62] Ren S, Peng Z *et al.*: **RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings**. *Cell Research* 2012. **22**(5): 806–821.

[63] Kim SK, Kim SY *et al.*: **A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients**. *Molecular Oncology* 2014. **8**(8): 1653–1666.

[64] Lin IH, Chen DT *et al.*: **Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes**. *PLoS ONE* 2015. **10**(2): 1–30.

[65] Homo sapiens high coverage assembly GRCh38, Date accessed: 2016-05-30.

[66] GRCh38.84 Homo Sapiens Gene Annotation, Date accessed: 2016-05-30.

[67] Schwartz RL and Phoenix T: *Learning perl*. O'Reilly & Associates, Inc., 2001.

[68] Wickham H: **The Split-Apply-Combine Strategy for Data Analysis**. *Journal of Statistical Software* 2011. **40**: 1—-29.

[69] Warden CD, Yuan Yc *et al.*: **Optimal Calculation of RNA-Seq Fold-Change Values**. *International Journal of Computational Bioinformatics and In Silico Modeling* 2013. **2**(6): 285–292.

[70] Krzywinski MI, Schein JE *et al.*: **Circos: An information aesthetic for comparative genomics**. *Genome Research* 2009.

[71] Powers S: *Unix Power Tools*. Nutshell handbook. O'Reilly Media, 2003.

[72] Lesk AM: *Introduction to genomics*. Oxford University Press, 2012.

[73] Malik B and Feng F: **Long noncoding RNAs in prostate cancer: overview and clinical implications**. *Asian Journal of Andrology* 2016. **18**(4): 568.

[74] Bolstad BM and Others: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003. **19**(2): 185.

[75] Yang YH, Dudoit S *et al.*: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic acids research* 2002. **30**(4): e15.

[76] Min HJ, Lee MK *et al.*: **TMPRSS4 induces cancer cell invasion through pro-uPA processing**. *Biochemical and Biophysical Research Communications* 2014. **446**(1): 1–7.

[77] Jung H, Lee KP *et al.*: **TMPRSS4 promotes invasion, migration and metastasis of human tumor cells by facilitating an epithelial-mesenchymal transition**. *Oncogene* 2008. **27**(18): 2635–2647.

[78] García-Pravia C, Galván JA *et al.*: **Overexpression of COL11A1 by Cancer-Associated Fibroblasts: Clinical Relevance of a Stromal Marker in Pancreatic Cancer**. *PLoS ONE* 2013. **8**(10): 1–13.

[79] Li A, Li J *et al.*: **COL11A1 is overexpressed in gastric cancer tissues and regulates proliferation, migration and invasion of HGC-27 gastric cancer cells in vitro**. *Oncology Reports* 2017. **37**(1): 333–340.

[80] Kim H, Watkinson J *et al.*: **Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1**. *BMC Medical Genomics* 2010. **3**(1): 51.

[81] Wu YH, Chang TH *et al.*: **COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer**. *Oncogene* 2014. **33**(26): 3432–3440.

[82] Shen L, Yang M *et al.*: **COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance**. *Oncology Reports* 2016.

[83] Peruzzi D, Mori F *et al.*: **MMP11: A novel target antigen for cancer immunotherapy**. *Clinical Cancer Research* 2009. **15**(12): 4104–4113.

[84] Yang YH, Deng H *et al.*: **Identification of matrix metalloproteinase 11 as a predictive tumor marker in serum based on gene expression profiling**. *Clinical Cancer Research* 2008. **14**(1): 74–81.

[85] Li Z, Ding S *et al.*: **Significance of MMP11 and P14 ARF expressions in clinical outcomes of patients with laryngeal cancer**. *International journal of clinical and experimental medicine* 2015. **8**(9): 15581–15590.

[86] Cheng CW, Yu JC *et al.*: **The clinical implications of MMP-11 and CK-20 expression in human breast cancer**. *Clinica Chimica Acta* 2010. **411**(3-4): 234–241.

[87] Tetu B, Trudel D *et al.*: **Proteases by reactive stromal cells in cancer: an attractive therapeutic target**. *Bulletin Du Cancer* 2006. **93**(9): 944–948.

[88] Nonsrijun N, Mitchai J *et al.*: **Overexpression of matrix metalloproteinase 11 in thai prostatic adenocarcinoma is associated with poor survival**. *Asian Pacific Journal of Cancer Prevention* 2013. **14**(5): 3331–3335.

[89] Tongtawee T, Kaewpitoon SJ *et al.*: **High expression of matrix metalloproteinase-11 indicates poor prognosis in human cholangiocarcinoma**. *Asian Pacific Journal of Cancer Prevention* 2015. **16**(9): 3697–3701.

[90] Gray KA, Yates B *et al.*: **Genenames.org: the HGNC resources in 2015.** *Nucleic acids research* 2015. **43**(Database issue): D1079–85.

[91] Zeng P, Zhang P *et al.*: **TMPRSS4 as an emerging potential poor prognostic factor for solid tumors : A systematic review and meta-analysis Demographic characteristics** 2016. **7**(46): 15–17.

[92] Shi G, Yang X *et al.*: **Clinical significance of TMPRSS4 in prostate cancer**. *International Journal of Clinical and Experimental Pathology* 2014. **7**(11): 8053–8058.

[93] Larzabal L, Nguewa PA *et al.*: **Overexpression of TMPRSS4 in non-small cell lung cancer is associated with poor prognosis in patients with squamous histology**. *Br J Cancer* 2011. **105**(10): 1608–1614.

[94] Wu XY, Zhang L *et al.*: **Clinical implication of TMPRSS4 expression in human gallbladder cancer**. *Tumor Biology* 2014. **35**(6): 5481–5486.

[95] Sheng H, Shen W *et al.*: **Prognostic significance of TMPRSS4 in gastric cancer.** *Neoplasma* 2013. **61**(2): 213–217.

[96] The UniProt Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Research* 2014. **43**(D1): D204–212.

[97] Herrero J, Muffato M *et al.*: **Ensembl comparative genomics resources.** *Database : the journal of biological databases and curation* 2016. **2016**(0): bav096–.

[98] 't Hoen PAC, Ariyurek Y *et al.*: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms**. *Nucleic Acids Research* 2008. **36**(21).

[99] Weigelt B, Baehner FL *et al.*: **The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade**. *The Journal of Pathology* 2009. n/a–n/a.

[100] Zhao S, Fung-Leung WP *et al.*: **Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells.** *PloS one* 2014. **9**(1): e78644.

[101] Bradford JR, Hey Y *et al.*: **A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling.** *BMC genomics* 2010. **11**: 282.

[102] Fu X, Fu N *et al.*: **Estimating accuracy of RNA-Seq and microarrays with proteomics.** *BMC genomics* 2009. **10**: 161.

[103] Conesa A, Madrigal P *et al.*: **A survey of best practices for RNA-seq data analysis**. *Genome Biol* 2016. **17**(1): 13.

[104] RNA-Seq Alignment - BaseSpace Sequence Hub. `https://basespace.illumina.com/apps/3086084/RNA-Seq-Alignment?preferredversion`, Date accessed: 2016-08-11.

[105] D'Antonio M, D'Onorio De Meo P *et al.*: **RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application.** *BMC genomics* 2015. **16**(Suppl 6): S3.

[106] Hong D, Rhie A *et al.*: **FX: An RNA-seq analysis tool on the cloud**. *Bioinformatics* 2012. **28**(5): 721–723.

[107] Langmead B, Hansen KD *et al.*: **Cloud-scale RNA-sequencing differential expression analysis with Myrna.** *Genome biology* 2010. **11**(8): R83.

[108] FASTQ upload is now available in BaseSpace. `https://blog.basespace.illumina.com/2014/08/18/fastq-upload-in-now-available-in-basespace/`, Date accessed: 2017-01-24.

[109] Kwon T, Yoo WG *et al.*: **Next-generation sequencing data analysis on cloud computing**. *Genes & Genomics* 2015. (April): 489–501.

[110] Afgan E, Baker D *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.** *Nucleic acids research* 2016. **44**(W1): W3–W10.

[111] Del Fabbro C, Scalabrin S *et al.*: **An extensive evaluation of read trimming effects on illumina NGS data analysis**. *PLoS ONE* 2013. **8**(12): 1–13.

[112] Consortium SI: **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium**. *Nat Biotech* 2014. **32**(9): 903–914.

[113] Engström PG, Steijger T *et al.*: **Systematic evaluation of spliced alignment programs for RNA-seq data.** *Nature methods* 2013. **10**(12): 1185–91.

[114] Xuan J, Yu Y *et al.*: **Next-generation sequencing in the clinic: Promises and challenges**. *Cancer Letters* 2013. **340**(2): 284–295.

[115] Seyednasrollah F, Laiho A *et al.*: **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Briefings in Bioinformatics* 2013. **16**(1): 59–70.

[116] Feichtinger J, McFarlane RJ *et al.*: **CancerEST: a web-based tool for automatic meta-analysis of public EST data.** *Database : the journal of biological databases and curation* 2014. **2014**: 1–8.

[117] Feichtinger J, McFarlane RJ *et al.*: **CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data.** *Database : the journal of biological databases and curation* 2012. **2012**: bas055.

[118] de Aberasturi AL and Calvo A: **TMPRSS4: an emerging potential therapeutic target in cancer**. *Br J Cancer* 2015. **112**(1): 4–8.

[119] Cho SH, Jeon J *et al.*: **Personalized medicine in breast cancer: a systematic review.** *Journal of breast cancer* 2012. **15**(3): 265–72.

[120] Martinez E, Yoshihara K *et al.*: **Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects**. *Oncogene* 2015. **34**(21): 2732–2740.

[121] Verma M: **Personalized Medicine and Cancer**. *Journal of Personalized Medicine* 2012. **2**: 1–14.

[122] XALKORI® (Crizotinib) Receives Approval In European Union For The Treatment Of Patients With ROS1-Positive Advanced Non-Small Cell Lung Cancer | Pfizer: One of the world's premier biopharmaceutical companies. `http://www.pfizer.com/news/press-release/press-release-detail/xalkori_crizotinib_receives_approval_in_european_union_for_the_treatment_of_patients_with_ros1_positive_advanced_non_small_cell_lung_cancer`, Date accessed: 2016-09-22.

[123] Diamandis EP: **The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem?** *BMC Medicine* 2012. **10**(1): 87.

[124] Mcshane LM, Altman DG *et al.*: **Reporting Recommendations for Tumor Marker Prognostic Studies** 2016. **23**(36): 9067–9072.

[125] Kulasingam V and Diamandis EP: **Strategies for discovering novel cancer biomarkers through utilization of emerging technologies.** *Nature clinical practice. Oncology* 2008. **5**(10): 588–99.

[126] Pepe MS, Etzioni R *et al.*: **Phases of Biomarker Development for Early Detection of Cancer**. *JNCI Journal of the National Cancer Institute* 2001. **93**(14): 1054–1061.

[127] TopHat - A spliced read mapper for RNA-Seq. `https://ccb.jhu.edu/software/tophat/index.shtml`, Date accessed: 2016-12-30.

[128] Trapnell C, Roberts A *et al.*: **Differential gene and transcript expressiont analysis of RNA-seq experiments with TopHat and Cufflinks**. *Natures Protocols* 2012. **7**(3): 562–578.

[129] Stajich JE, Block D *et al.*: **The Bioperl Toolkit: Perl Modules for the Life Sciences**. *Genome Research* 2002. **12**(10): 1611–1618.

[130] Illumina: **Illumina adapter sequences**. *Illumina* 2009. (October): 1–4.

[131] Vaingast S: *Beginning Python visualization: crafting visual transformation scripts*. Apress, 2014.

Appendices

6.1 Additional information on RNA-Seq Tools

In this appendix the tools implemented in the CancerRNAseq pipeline are explained in more detail.

6.1.0.1 Cutadapt

Cutadapt developed by Martin M. [32] is an adapter trimming tool for high-throughput sequencing reads. The initial step of the Cutadapt algorithm calculates the optimal alignment between each read and the user provided adapter sequence. Cutadapt incorporates a semi-global (option: -a, -A for second read) and slightly modified semi-global (option: -b, -B for second read) alignment approach for the alignment calculation. The regular semi-global alignment (end-space free alignment) approach does not penalize initial or trailing gaps, which results in the problem that even a short random match at the 5' end of a read leads to the removal of the whole read (see "Option -a" in figure 6.1). The slightly modified semi-global alignment method downgrades initial gaps and therefore is able to handle full or partial adapters at the beginning of a read (see "Option -b" in figure 6.1). Adapters within a read as well as reads running into adapters can be handled from both alignment approaches of Cutadapt (see "Option -a and -b" in figure 6.1).

After all provided adapters have been aligned to the read, the adapter with the highest number of matching characters is used for trimming of the read. The last step in Cutadapt is the calculation of the error rate, which is the number of errors divided by the length of the matching sequence. The read is trimmed if the error rate is below a certain threshold [32].

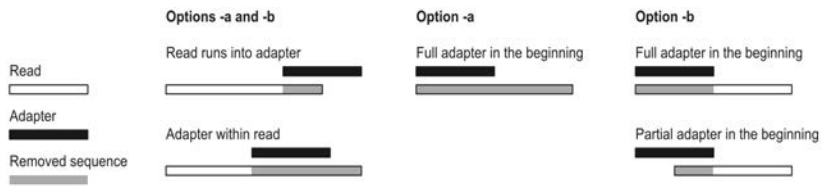


Figure 6.1: ***The different alignment options of CutAdapt between read and adapter sequence [32].*** *Option "-a" illustrates the alignment that uses the regular semi-global approach and "-b" characterises the alignments that used the slightly modified semi-gloabl alignment.*

6.1.0.2 Trimmomatic

Trimmomatic developed by Bolger *et al.* [33] is a flexible trimming and quality filtering tool for Illumina sequence data. Two different approaches for detecting and removing technical sequences are implemented in the Trimmomatic tool. The first approach, called "simple mode", is able to handle all technical sequences, for instance, adapters and PCR primers. The algorithm tries to find the approximate match between the read and a technical sequence with a required minimum overlap. This overlap is needed to prevent from false positive values. A drawback of this approach is the inability to detect short partial adapter sequences at the end of a read. The second approach for deleting technical sequences is the "palindrome mode" designed for the removal of adapter sequences at the end of reads, called adapter "read-through". This method can only be applied on paired-end reads and uses three attributes: (i) both reads have an equal number of bases, (ii) the adapters contamination has to be from an 'opposite' one (iii) the two reads are reverse complements. Adapter identification in a "read-through" manner provides high sensitivity as well as specificity [33].

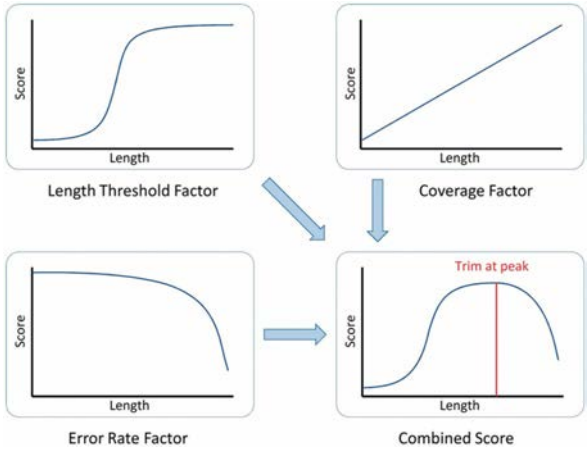


Figure 6.2: ***The maximum information algorithm [33].*** *The maximum information algorithm combines three factors the length threshold, coverage and error rate to choose the ideal trimming point.*

In general, the quality of Illumina reads decreases at the 3' end of the reads and, therefore, the bases toward the 3' end of a defined cut point are removed. Two different methods for quality filtering are implemented in the Trimmomatic tool: (i) sliding window and (ii) maximum information quality filtering. The first approach for quality filtering is the "sliding window quality filtering" method that scans the read with a window of a user defined width from the 5' end to the 3' end. If the average base quality drops below a certain threshold the 3' end will be removed, which means that in case of only one poor quality base the 3' end of the read will remain. The "maximum information" algorithm was introduced with Trimmomatic and determines with three different factors how much of a read should be preserved. First factor is the "length threshold", as the read need a minimal length of the remaining read to be useful for downstream analysis. The second factor, "coverage" is a linear score depending on the remaining read length, because longer reads can be mapped to fewer locations on the reference genome and therefore the information content is higher. The third factor, "error rate" uses the error probabilities of each base to calculate a combined

likelihood of errors over the read, which starts at high value and drops toward the end dramatically (see Error Rate Factor figure 6.2). As it can be seen in figure 6.2, the maximum information algorithm combines these three factors to find the best possible trimming point on the read [33].

6.1.0.3 rCorrector

rCorrector [34] developed by Song L. and Florea L. is a novel k-spectrum based method for correcting errorneous Illumina RNA-Seq reads. Sequencing errors in RNA-Seq, which is nowadays the foundation of the common transcriptome analysis, can significantly influence the analysis results.

In case of whole genome sequencing (WGS) reads the systematic error correction methods can be divided into three different types: (i) k-spectrum, (ii) suffix tree and suffix array and (iii) multiple sequence alignment (MSA) based methods. The most popular error correction methods for WGS reads are the k-spectrum based methods. In the suffix tree and suffix array based method substrings of reads are replaced according to information achieved by the comparison of data structures from the input reads and probabilistic models. The multiple sequence alignment methods construct multiple alignments as well as local proximity measures by clustering reads with stored k-mers, which are used to correct the faulty reads. The main difference between WGS reads and RNA-Seq reads is that WGS data is largely uniform and RNA-Seq data is differentially expressed. Therefore, WGS methods cannot be applied to RNA-Seq data.

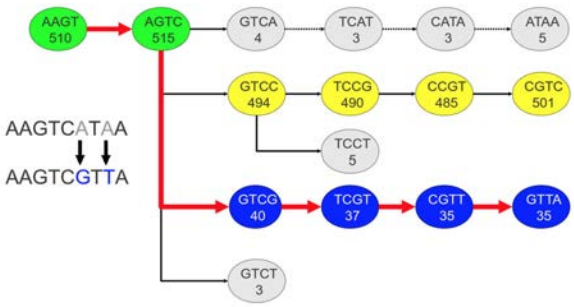


Figure 6.3: ***The path extension algorithm of rCorrector [34].*** *The De Bruijn graph for the read "AAGTCATAA" consists of four different possible proceedings at the AGTC k-mer (k = 4). The first path (top row) is analogous to the original read representation in the graph. The count of "GTCA" falls below the local cutoff and therefore the extension of the path is stopped after the first step. This threshold is calculated based on the maximum k-mer count of the four different successors for "AGTC" in the De Bruijn graph, which is 494 for "GTCC" in the second row. The path in the third (blue) row is used to correct the read, because despite of the lower k-mer count only two corrections instead of four corrections are introduced in comparison to the in the second (yellow) path.*

rCorrector is a novel k-spectrum based method that uses a flexible k-mer count threshold. This is used to consider transcript and expression levels and different thresholds for each read. Furthermore, multiple k-mer choices at any position are allowed and multiple occurring k-mers are stored, which allows the rCorrector to be scalable for large datasets. rCorrector is suited to handle large as well as complex genomes like the human genome.

rCorrector builds a De Bruijn graph of all k-mers and their counts appearing multiple times (>1) in the input reads. The algorithm detects at run time the closest path corresponding to its transcript of origin for each read in the De Bruijn graph. This is used for correcting the initial reads. k-mers can be divided into solid and non-solid ones. If a k-mer passes a certain count threshold it is considered to be a solid k-mer, which means that the k-mer is correct. rCorrector checks at runtime wether the k-mer and its alternatives are solid ones. If the numbers of corrections in one path exceed a given threshold, the algorithm stops searching the current path and starts with a new one. The path with the smallest number of differences is used for the correction of k-mers in the original read [34].

6.1.0.4 STAR

The STAR (Spliced Transcripts Alignment to a Reference) aligner was published by Dobin A. *et al.* in 2013 [2]. The STAR algorithm aligns the non-contigous sequences directly to the genome in a two-step approach:

1. Seed searching step
2. Clustering/stitching/scoring step

The seed searching step can be used to sequentially indentify the MMP (Maximum Mappable Prefix), which is defined as the longest substring of the read sequence matching exactly one or more substrings of the reference genome. Figure 6.4 shows a schematic drawing of the way the STAR algorithm searches for the MMP. Using such an approach detects splice junctions, mismatches as well as tails in a single alignment pass. The second part of the STAR algorithm stitches all the seeds, which were aligned to genome in the previous step, together. The outcome of this step is the alignment of the whole read sequence. An advantage of the STAR algorithm is its ability to allow possible genomic gaps or overlaps in between the mates of paired-end RNA-Seq reads. Another profound advantage is the capability of the STAR algorithm to deal with reads that were generated from third generation sequencing technologies, which means that STAR can handle long reads up to several kilobases in length [2].

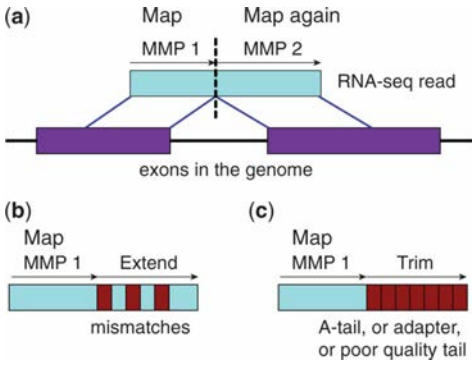


Figure 6.4: Illustration of the MMP (Maximum Mappable Prefix) search of the STAR algorithm for the detection of splice junctions (a), missmatches (b) and tails (c) [2].

6.1.0.5 TopHat2

The TopHat2 alignment tool was published by Kim D. *et al.* in 2013 [39] and it is the successor of the TopHat alignment tool that was published by Trapnell C. *et al.* in 2009 [36]. The TopHat2 algorithm uses either Bowtie or Bowtie2 as the core alignment tool. Bowtie2 is additionally capable of finding short indels for the alignment of the reads compared to Bowtie, which is only a short read alignment tool.

Figure 6.5 illustrates the mapping steps of the tool TopHat, which is the same in case of the successor aligner TopHat2 [43]. The initial step of the TopHat2 pipeline is the mapping of reads spanning a single exon using Bowtie or Bowtie2. The multi-exon spanning reads that could not be mapped in the first attempt are further used in the spliced alignment step. Therefore, these reads are split into smaller fragments, which are realigned. The next step is to detect potential splice sites as well as indels and fusion break points by observing the smaller fragments that have aligned to the genome in the previous step. The sequences on the left and right side of a detected splice junction are concatenated and used to map sequence fragments to them. The aligned sequences are stitched together forming the whole genome read alignment. In the last step TopHat2 is considering reads that are located on a splice junction, but with a few bases aligned to an intronic sequence and tries to re-align them to exons only [39].

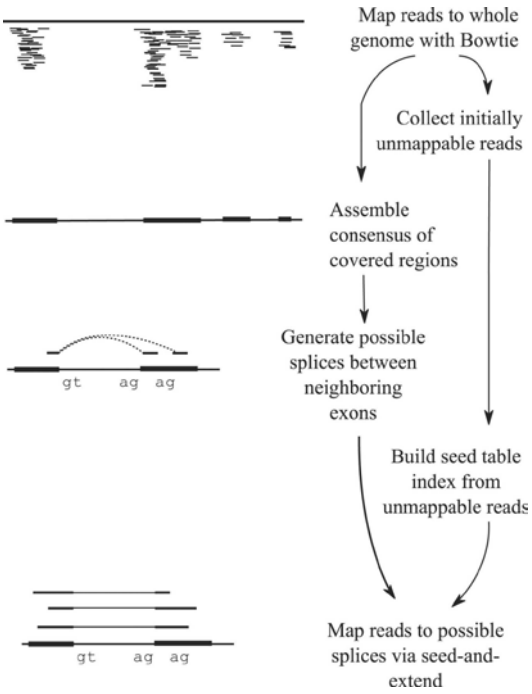


Figure 6.5: Illustration of the two-phase process of TopHat for finding splice junctions by mapping reads to the reference genome [36].

6.1.0.6 HISAT2

The HISAT2 (hierarchical indexing for spliced alignment of transcripts) mapper was published by Kim D. *et al.* in 2015 [37]. HISAT2 was published by the same research group as TopHat2 and should be the core alignment engine for the upcoming successor of TopHat2 named TopHat3. HISAT2 uses two different kinds of indices for

the alignment of reads based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index. For the alignment process, HISAT2 uses a whole-genome FM index, which is used to find the anchor for each sequence, as well as several local FM indices for the extension of these aligned reads. For the initial alignment of each read, the HISAT2 algorithm uses the global FM index for finding the position the read may have originated. In the next step a local index for the complete alignment is chosen out of approximately 48,000 local indicies. In the case of paired-end reads, each read is aligned separately and afterwards combined. When the HISAT2 algorithm does not succeed in aligning both reads of a pair, the read that has been aligned functions as an anchor for the mapping of the initially not alignable read. Like TopHat, the HISAT2 tool uses Bowtie2 for several low-level operations, like the construction of the FM index or searching that kind of index. An advantage of using the BWT and FM indices is that HISAT2 is able to compete with STAR in terms of speed, but with a much lower RAM (Random Access Memory) usage [37].

6.1.0.7 Cufflinks

The Cufflinks algorithm, which was developed by Trapnell C. *et al.* [43], requires a spliced alignment file as input. In case of paired-end data, the Cufflinks algorithm treats mating pairs in the alignment like a single one. The initial step in the algorithm is to assemble an overlap graph from overlapping bundles of the fragment alignment. Next, all the pairs of incompatible fragments (yellow, blue and red in figure 6.6b) are identified, because they derive from miscellaneous spliced isoforms. Now all the compatible fragments can be connected to an overlap graph (see figure 6.6c). This means, each fragment is a node in the overlap graph and is connected via edges with compatible fragments. Isoforms can be generated from this overlap graph. In figure 6.6c the overlap graph can be minimally covered by three different paths, which equals to three different isoforms (yellow, blue and red). Minimally covered means that the origin of all other fragments (grey) can be explained by means of one of three isoforms. The Cufflinks algorithm uses the Dillworth's theorem to proof this minimum in terms of isoforms. The Dillworth's theory says that the number of transcripts of mutually incompatible reads equals to the minimum number of transcripts to explain all fragments.

In the next step the transcript abundance is calculated, with a statistical model that considers the length of the fragment with respect to the isoform it could have originated. At the end the Cufflinks algorithm uses a maximum likelihood abundance to estimate the abundance explaining the fragment in the best possible way (see figure 6.6e) [43].

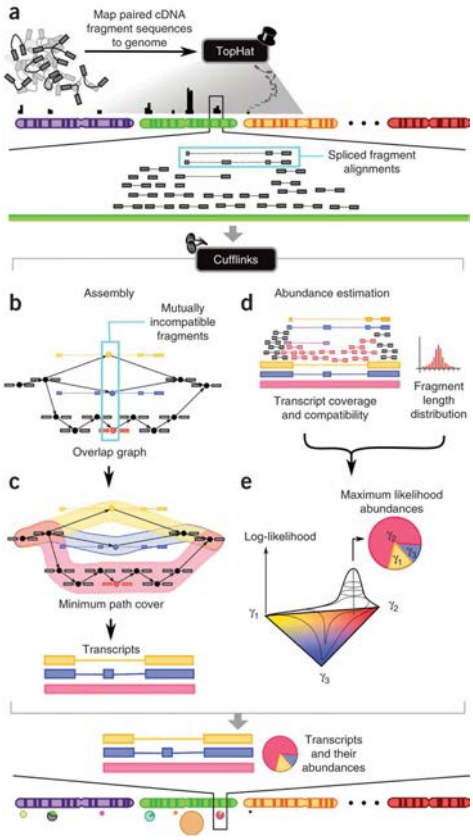


Figure 6.6: ***The Cufflinks algorithm [43].*** *Cufflinks uses cDNA fragments that have been aligned to a reference genome as input (a). The Cufflinks algorithm generates overlapping bundles from the aligned fragments and identifies pairs of incompatible fragments, which have derived from a different spliced isoform (b). An overlap graph from compatible fragments as well as their alignments overlap in the genome (c). The next step is the calculation of the transcript abundance (d). In the last step of Cufflinks the fragments are matched to the transcript they probably have originated and a likelihood for all the possible abundances of the isoforms are assigned to explain the analysed fragments (e).*

6.1.0.8 Cuffmerge

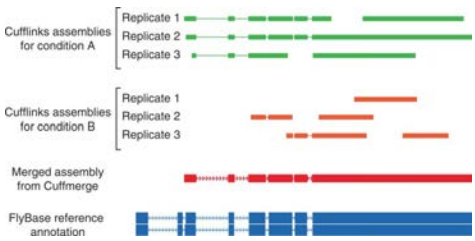


Figure 6.7: ***The Cuffmerge algorithm [128].*** *Merging the two assemblies from Cufflinks (Condition A & B) might recovers the complete gene in many cases, which can be useful in case of poorly expressed genes [128].*

Cuffmerge was developed by Trapnell C. *et al.* [44], because it is essential in case of large datasets with several RNA-Seq samples to assemble them parsimoniously into a

joint set of transcripts. It would be possible to run Cufflinks instead of Cuffmerge on all RNA-Seq samples, but this approach has several disadvantages (*i.e.,* merging can lead to an increase in errors) compared to using Cufflinks in advance and afterwards Cuffmerge. As it can be seen in figure 6.7, with Cuffmerge it is possible to recreate the whole gene even if one condition is less expressed. The DE expressed genes as well as regulated ones can be detected in the final assembly [128].

6.1.0.9 Cuffdiff

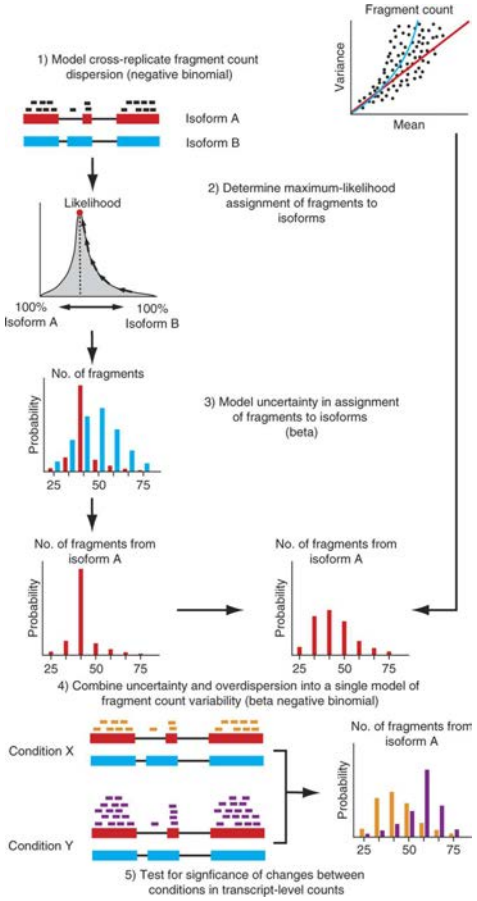


Figure 6.8: ***The Cuffdiff workflow [44].*** *The Cuffdiff algorithm initially models the variability in fragment count for each gene across replicates. The next step is the estimation of fragment count for each isoform in each replicate, with a measure of uncertainty resulting from ambigiously mapped reads. Cuffdiff uses the combined estimates of uncertainty as well as the cross replicate variability to report significant DE genes.*

Cuffdiff was developed by Trapnell C. *et al.* [44] to perform differential analysis on RNA-Seq data to identify differentially expressed transcripts and genes based on the assumption that the expression of transcript in each condition can be measured by the amount of fragments generated by them. A difference in the expression level between conditions can then be detected due to a difference in fragment count.

The Cuffdiff algorithm uses a negative binomial model to identify significant differentially expressed genes and to compute the p-value for alterations in a

transcript's fragment count. The poisson model, which is computationally simple, has major disadvantages compared to the negative binomial model. One of the disadvantages is that the poisson model has problems with the count uncertainty, which is the case in the high amount of ambiguously mapped reads to different transcripts in an RNA-Seq experiment. The variability across replicates (count overdispersion) is in RNA-Seq datasets higher than modeled with a poisson model.

The initial step in the Cuffdiff algorithm models the cross-replicate fragment count dispersion as well as a measure of uncertainty of the different isoforms (see figure 6.8(1)). The next step of the Cuffdiff algorithm models the fragment count for each isoform in each replicate, with a measure of uncertainty from the ambiguously mapped reads (see figure 6.8(2-3)). Then the estimates of uncertainty and the cross-replicate variability are combined under a beta negative binomial model for each transcript in each library. The last step of the algorithm is to statistically test for differentially expressed genes [44].

6.2 Supplemental Methods

6.2.1 Perl packages

The following Perl packages (available at `http://www.cpan.org`) were used for the CancerRNAseq pipeline:

- Backticks (version 1.0.9)
- Bio::SeqIO (BioPerl version 1.6.924) [129]
- Config::Simple (version 4.58)
- Cwd (version 3.62)
- File::chdir (version 0.1010)
- File::Copy (version 2.14)
- File::Find (version 1.14)
- File::Path (version 2.08)
- Getopt::Long (version 2.38)

6.2.2 Illumina Adapters for Cutadapt

Cutadapt was used with standard Illumina adapters (see [130] for additional information on Illumina adapters). The adapters are as following:

- GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG
- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
- ATCTCGTATGCCGTCTTCTGCTTG
- CAAGCAGAAGACGGCATACGA
- GGGGAATGCTGGAGATTGTAATGGGTATGGAGACATATCATATAAGTAAT

6.3 Supplemental Results

6.3.0.1 Mutual DE genes of the datasets

Table shows the similar genes between the different example datasets for over-expressed as well as under-expressed genes.

Table 6.1: Similar genes between over-expressed and under-expressed genes of the tested example datasets. PRJEB4829 - breast adenocarcinoma, PRJEB2449 - prostate cancer, PRJNA163279 - NSCLC (non-small cell lung carcinoma), PRJNA218851 - primary colorectal cancer.

		PRJEB4829		**PRJEB2449**		**PRJNA163279**		**PRJNA218851**	
		↑	↓	↑	↓	↑	↓	↑	↓
PRJEB4829	↑	2928	0	81	44	205	49	122	14
	↓		3101	124	188	190	511	153	72
PRJEB2449	↑			727	0	46	27	28	11
	↓				484	67	83	40	13
PRJNA163279	↑					923	0	102	20
	↓						1082	54	44
PRJNA163279	↑							586	0
	↓								236

6.4 Configuration Files

The system configuration file (6.4.2) was used to provide the pipeline with general information for all four example datasets and for each dataset an additional user configuration file that is specific for each dataset (6.4.3 - 6.4.6 was used to execute the pipeline. All configuration files are written in the INI-format.

6.4.1 INI-Format

The INI-format consists of three different line types. The first type of text used in an INI-file is enclosed by square brackets and marks the start of a section (*i.e.,* `[header]`). The second type, is the parameter line type, which consists of a text left and right of an equal sign (*i.e.,* `parameter1=value1`). The last line type in an INI-file is the comment line or remark line starting with a "#" or ";" as a prefix (*i.e.,* `#Comment` or `;Comment)` (see 6.4.2) [131].

6.4.2 System Configuration File

```
[sdirectories]
;In this category all the necassary directions for a smooth run of the
;pipeline are provided.
Pipeline=/export/home/rhablesreiter_914/Pipeline_Version1/Pipeline/Program
RawData=/export/home/rhablesreiter_914/Pipeline_Version1/Pipeline/RawData
Results=/export/home/rhablesreiter_914/Pipeline_Version1/Pipeline/Results
ReferenceData=

[sprograms]
;In this category the direction as well as the file and some parameters of the
;needed programs are provided.
Trimmomatic=/export/home/rhablesreiter_914/Pipeline/RawData/trimmomatic-0.35.jar
rCorrector=/export/programs/rcorrector/run_rcorrector.pl
BowTie2=bowtie2
BowTie2Build=bowtie2-build
TopHat2=tophat2
STAR=/export/programs/STAR/bin/Linux_x86_64_static/STAR
Cufflinks=cufflinks
Cuffmerge=cuffmerge
Cuffdiff=cuffdiff
Samtools=samtools
QualiMap=java -Djava.awt.headless=true -Xmx20G -XX:MaxPermSize=1024m -classpath /export/home/rhablesreiter_914/QualiMap/qualimap-build-04-04-16/qualimap.jar:/export/home/rhablesrei$
Rscript=Rscript /export/home/rhablesreiter_914/Pipeline_Version1/Pipeline/Program/CummeRbundRun.R
HISAT2=/export/home/rhablesreiter_914/HISAT2/hisat2-2.0.4/hisat2 --dta-cufflinks

[sfiles]
;In this category all the needed default files are
;provided.
AdapterFasta=/export/home/rhablesreiter_914/Pipeline_Version1/Pipeline/Program/default_adapter.fasta

[snames]
;In this category all the default variables are set.
prj_name=Project

[scmd]
STAR=--readFilesCommand zcat --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonical --alignEndsType EndToEnd
Trimmomatic=SLIDINGWINDOW:3:20 MINLEN:20
```

6.4.3 User Configuration File: PRJEB4829

6.4.3.1 Alignment Tool: STAR

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJEB4829

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=BreastAdenocarcinoma

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=STAR

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz


;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33

[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=

[sequences]
;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=ERR358485_1.fastq.gz
sequence1_1_2=ERR358485_2.fastq.gz
sequence1_2_1=ERR358486_1.fastq.gz
sequence1_2_2=ERR358486_2.fastq.gz

sequence2_1_1=ERR358487_1.fastq.gz
sequence2_1_2=ERR358487_2.fastq.gz
sequence2_2_1=ERR358488_1.fastq.gz
sequence2_2_2=ERR358488_2.fastq.gz
```

6.4.3.2 Alignment Tool: TopHat2

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJEB4829_TOPHAT

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=BreastAdenocarcinoma

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=Tophat

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz


;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33

[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=

[sequences]
;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=ERR358485_1.fastq.gz
sequence1_1_2=ERR358485_2.fastq.gz
sequence1_2_1=ERR358486_1.fastq.gz
sequence1_2_2=ERR358486_2.fastq.gz

sequence2_1_1=ERR358487_1.fastq.gz
sequence2_1_2=ERR358487_2.fastq.gz
sequence2_2_1=ERR358488_1.fastq.gz
sequence2_2_2=ERR358488_2.fastq.gz
```

6.4.3.3 Alignment Tool: HISAT2

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJEB4829_HISAT

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=BreastAdenocarcinoma

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=HISAT2

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz


;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33

[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=

[sequences]
;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=ERR358485_1.fastq.gz
sequence1_1_2=ERR358485_2.fastq.gz
sequence1_2_1=ERR358486_1.fastq.gz
sequence1_2_2=ERR358486_2.fastq.gz

sequence2_1_1=ERR358487_1.fastq.gz
sequence2_1_2=ERR358487_2.fastq.gz
sequence2_2_1=ERR358488_1.fastq.gz
sequence2_2_2=ERR358488_2.fastq.gz
```

6.4.4 User Configuration File: PRJNA163279

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJNA163279PE

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=Adenocarcinoma

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=STAR

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz



;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33


[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=

[sequences]
;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=SRR493937_1.fastq.gz
sequence1_1_2=SRR493937_2.fastq.gz
sequence1_2_1=SRR493941_1.fastq.gz
sequence1_2_2=SRR493941_2.fastq.gz
sequence1_3_1=SRR493945_1.fastq.gz
sequence1_3_2=SRR493945_2.fastq.gz
sequence1_4_1=SRR493949_1.fastq.gz
sequence1_4_2=SRR493949_2.fastq.gz
sequence1_5_1=SRR493953_1.fastq.gz
sequence1_5_2=SRR493953_2.fastq.gz
sequence1_6_1=SRR493957_1.fastq.gz
sequence1_6_2=SRR493957_2.fastq.gz

sequence2_1_1=SRR493939_1.fastq.gz
sequence2_1_2=SRR493939_2.fastq.gz
sequence2_2_1=SRR493944_1.fastq.gz
sequence2_2_2=SRR493944_2.fastq.gz
sequence2_3_1=SRR493948_1.fastq.gz
sequence2_3_2=SRR493948_2.fastq.gz
sequence2_4_1=SRR493951_1.fastq.gz
sequence2_4_2=SRR493951_2.fastq.gz
sequence2_5_1=SRR493955_1.fastq.gz
sequence2_5_2=SRR493955_2.fastq.gz
sequence2_6_1=SRR493959_1.fastq.gz
sequence2_6_2=SRR493959_2.fastq.gz
```

6.4.5 User Configuration File: PRJEB2449

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJEB2449

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=ProstateTumor

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=STAR

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz



;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33

[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=yes

[sequences]
;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=ERR031017_1.fastq.gz
sequence1_1_2=ERR031017_2.fastq.gz
sequence1_2_1=ERR031019_1.fastq.gz
sequence1_2_2=ERR031019_2.fastq.gz
sequence1_3_1=ERR031023_1.fastq.gz
sequence1_3_2=ERR031023_2.fastq.gz
sequence1_4_1=ERR031025_1.fastq.gz
sequence1_4_2=ERR031025_2.fastq.gz
sequence1_5_1=ERR031027_1.fastq.gz
sequence1_5_2=ERR031027_2.fastq.gz
sequence1_6_1=ERR031029_1.fastq.gz
sequence1_6_2=ERR031029_2.fastq.gz
sequence1_7_1=ERR031031_1.fastq.gz
sequence1_7_2=ERR031031_2.fastq.gz
sequence1_8_1=ERR031033_1.fastq.gz
sequence1_8_2=ERR031033_2.fastq.gz
sequence1_9_1=ERR031035_1.fastq.gz
sequence1_9_2=ERR031035_2.fastq.gz
sequence1_10_1=ERR031039_1.fastq.gz
sequence1_10_2=ERR031039_2.fastq.gz
sequence1_11_1=ERR031041_1.fastq.gz
sequence1_11_2=ERR031041_2.fastq.gz
sequence1_12_1=ERR031043_1.fastq.gz
sequence1_12_2=ERR031043_2.fastq.gz
sequence1_13_1=ERR299296_1.fastq.gz
sequence1_13_2=ERR299296_2.fastq.gz
sequence1_14_1=ERR299299_1.fastq.gz
sequence1_14_2=ERR299299_2.fastq.gz

sequence2_1_1=ERR031018_1.fastq.gz
sequence2_1_2=ERR031018_2.fastq.gz
sequence2_2_1=ERR031022_1.fastq.gz
sequence2_2_2=ERR031022_2.fastq.gz
sequence2_3_1=ERR031024_1.fastq.gz
sequence2_3_2=ERR031024_2.fastq.gz
sequence2_4_1=ERR031026_1.fastq.gz
sequence2_4_2=ERR031026_2.fastq.gz
sequence2_5_1=ERR031028_1.fastq.gz
sequence2_5_2=ERR031028_2.fastq.gz
sequence2_6_1=ERR031030_1.fastq.gz
sequence2_6_2=ERR031030_2.fastq.gz
sequence2_7_1=ERR031032_1.fastq.gz
sequence2_7_2=ERR031032_2.fastq.gz
sequence2_8_1=ERR031038_1.fastq.gz
sequence2_8_2=ERR031038_2.fastq.gz
sequence2_9_1=ERR031040_1.fastq.gz
sequence2_9_2=ERR031040_2.fastq.gz
sequence2_10_1=ERR031042_1.fastq.gz
sequence2_10_2=ERR031042_2.fastq.gz
sequence2_11_1=ERR031044_1.fastq.gz
sequence2_11_2=ERR031044_2.fastq.gz
sequence2_12_1=ERR299295_1.fastq.gz
sequence2_12_2=ERR299295_2.fastq.gz
sequence2_13_1=ERR299297_1.fastq.gz
sequence2_13_2=ERR299297_2.fastq.gz
sequence2_14_1=ERR299298_1.fastq.gz
sequence2_14_2=ERR299298_2.fastq.gz
```

6.4.6 User Configuration File: PRJNA218851

```
;Raphael Hablesreiter
;The user has to provide the required information.
;The information in the optional section can be provided for a user defined
;run of the pipeline.
;If no optional information is provided the pipeline uses default parameters.

[required]
;Username and Projectname are used to create the folder structure in the "Result" folder
;of the pipeline.
User=RaphaelHablesreiter
ProjectName=PRJNA218851primary

;Name of the two different kinds of sequences
sequence_1=Normal
sequence_2=PrimaryColorectalCancer

;The library type of the sequences[PE/SE] used
seqtype=PE


;Alignment tool used in the mapping section [STAR/Tophat]
MappingType=STAR

;Analyis type of the cufflinks-cuffdiff run
;[pooled/paired/pairedindividual]
analysistype=pooled

;Please provide the name of a gzipped or unzipped version of the reference genome
;(prefered: Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
;as well as the name for the index that will be built if necessary
GenomeFASTA=Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
ReferenceGenome=GRCh38
;Please provide the name of a gzipped or unzipped version of the annotation file
;(prefered: Homo_sapiens.GRCh38.84.gtf.gz)
GenomeGTF=Homo_sapiens.GRCh38.84.gtf.gz



;Savemode is used to adjust the amount of data that is kept after the pipeline run.
;savemode=1 (default) the most important files are kept
;savemode=2 intermediates are also kept
savemode=1

;Quality scores of the sequences[phred33/phred64]
qscore=phred33

[optional]

;If an adapterfile is provided, it will be used in the cutadapt run,
;instead of the default used Illumina adapters.
;(The provided adapter file has to be in the FASTA-Format. The sequence header should
;only contain the character for the trimming style in cutadapt(see cutadpat-manual).)
adapters=

;Additional cutadapt commands can be provided here.
cutadapt_cmds=

;Option for ending the pipeline after a certain section.
;[Cutadapt/Trimmomatic/rCorrector/Mapping/QualiMap/Cufflinks/Cuffmerge/Cuffdiff]
end=

;Delete chromosome MT after mapping [default:NO]
delete_mt=

[sequences]

;List of all sequences
;sequence[1/2]_[NRUMBER]_[SE:1,PE:1/2]=[FILENAME]
sequence1_1_1=SRR975569_1.fastq.gz
sequence1_1_2=SRR975569_1.fastq.gz
sequence1_2_1=SRR975570_1.fastq.gz
sequence1_2_2=SRR975570_2.fastq.gz
sequence1_3_1=SRR975571_1.fastq.gz
sequence1_3_2=SRR975571_2.fastq.gz
sequence1_4_1=SRR975572_1.fastq.gz
sequence1_4_2=SRR975572_2.fastq.gz
sequence1_5_1=SRR975573_1.fastq.gz
sequence1_5_2=SRR975573_2.fastq.gz
sequence1_6_1=SRR975574_1.fastq.gz
sequence1_6_2=SRR975574_2.fastq.gz
sequence1_7_1=SRR975575_1.fastq.gz
sequence1_7_2=SRR975575_2.fastq.gz
sequence1_8_1=SRR975576_1.fastq.gz
sequence1_8_2=SRR975576_2.fastq.gz
sequence1_9_1=SRR975577_1.fastq.gz
sequence1_9_2=SRR975577_2.fastq.gz
sequence1_10_1=SRR975578_1.fastq.gz
sequence1_10_2=SRR975578_2.fastq.gz
sequence1_11_1=SRR975579_1.fastq.gz
sequence1_11_2=SRR975579_2.fastq.gz
sequence1_12_1=SRR975580_1.fastq.gz
sequence1_12_2=SRR975580_2.fastq.gz
sequence1_13_1=SRR975581_1.fastq.gz
sequence1_13_2=SRR975581_2.fastq.gz
sequence1_14_1=SRR975582_1.fastq.gz
sequence1_14_2=SRR975582_2.fastq.gz
sequence1_15_1=SRR975583_1.fastq.gz
sequence1_15_2=SRR975583_2.fastq.gz
sequence1_16_1=SRR975584_1.fastq.gz
sequence1_16_2=SRR975584_2.fastq.gz
sequence1_17_1=SRR975585_1.fastq.gz
sequence1_17_2=SRR975585_2.fastq.gz
sequence1_18_1=SRR975586_1.fastq.gz
sequence1_18_2=SRR975586_2.fastq.gz

sequence2_1_1=SRR975551_1.fastq.gz
sequence2_1_2=SRR975551_1.fastq.gz
sequence2_2_1=SRR975552_1.fastq.gz
sequence2_2_2=SRR975552_2.fastq.gz
sequence2_3_1=SRR975553_1.fastq.gz
sequence2_3_2=SRR975553_2.fastq.gz
sequence2_4_1=SRR975554_1.fastq.gz
sequence2_4_2=SRR975554_2.fastq.gz
sequence2_5_1=SRR975555_1.fastq.gz
sequence2_5_2=SRR975555_2.fastq.gz
sequence2_6_1=SRR975556_1.fastq.gz
sequence2_6_2=SRR975556_2.fastq.gz
sequence2_7_1=SRR975557_1.fastq.gz
sequence2_7_2=SRR975557_2.fastq.gz
sequence2_8_1=SRR975558_1.fastq.gz
sequence2_8_2=SRR975558_2.fastq.gz
sequence2_9_1=SRR975559_1.fastq.gz
sequence2_9_2=SRR975559_2.fastq.gz
sequence2_10_1=SRR975560_1.fastq.gz
sequence2_10_2=SRR975560_2.fastq.gz
sequence2_11_1=SRR975561_1.fastq.gz
sequence2_11_2=SRR975561_2.fastq.gz
sequence2_12_1=SRR975562_1.fastq.gz
sequence2_12_2=SRR975562_2.fastq.gz
sequence2_13_1=SRR975563_1.fastq.gz
sequence2_13_2=SRR975563_2.fastq.gz
sequence2_14_1=SRR975564_1.fastq.gz
sequence2_14_2=SRR975564_2.fastq.gz
sequence2_15_1=SRR975565_1.fastq.gz
sequence2_15_2=SRR975565_2.fastq.gz
sequence2_16_1=SRR975566_1.fastq.gz
sequence2_16_2=SRR975566_2.fastq.gz
sequence2_17_1=SRR975567_1.fastq.gz
sequence2_17_2=SRR975567_2.fastq.gz
sequence2_18_1=SRR975568_1.fastq.gz
sequence2_18_2=SRR975568_2.fastq.gz
```