TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR INFORMATIK

DEPARTMENT FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Assembly and Analysis of Next-Generation Sequencing Data

Thomas Schwarzmayr

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:
Univ.-Prof. Dr. Nassir Navab

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Burkhard Rost
2. Priv.-Doz. Dr. Tim M. Strom

Die Dissertation wurde am 31.05.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 25.10.2017 angenommen.

Acknowledgements

First and foremost I would like to thank my supervisor Tim Strom for the possibility to work on many interesting projects, the freedom to pursue own ideas, the numerous meetings and conferences I was allowed to attend where I could learn a lot and most notably for his valuable support and guidance.

I am especially grateful to my doctoral advisor Burkhard Rost for supporting my thesis and all the efforts he took to finish this project.

Next I would like to thank Thomas Meitinger for the opportunity to do my PhD project in his institute but also for the countless valuable comments and the possibility to travel to interesting conferences.

Important to mention are also my colleagues at the Institute of Human Genetics who upvalued my time at the Helmholtz Zentrum München and helped me with numerous problems and questions. Special thanks goes to Thomas and Riccardo, who took a lot of work off my shoulders while writing this thesis, but also to all other colleagues for their creative input and words of advice.

Last but not least I would like to thank all my family and friends who gave me indispensable support throughout my whole life and without whom this would not have been possible.

Thank you . . .

Abstract

The advent of the next-generation sequencing technology and RNA sequencing (RNA-seq) in particular facilitates the investigation of the entire transcriptome of an individual. The resulting data comprise plenty of useful information and more than 1,500 RNA-seq samples have been analyzed in the course of this PhD project. In order to process these data and to interpret the gained information sophisticated software is necessary. Although many tools are available for the analysis of RNA-seq data, their application is often complex and time-consuming. Furthermore, large amounts of sequencing data are produced in our local environment per week. Thus, fast and easy yet reliable and tailored data analysis is vital.

The aim of this thesis was to design and to develop an automated RNA-seq data analysis pipeline and to unite it with an already existing and established whole exome sequencing data analysis pipeline. Furthermore, methods and parameters were investigated and implemented for each analysis step in order to generate reliable results. For that purpose, in-house as well as publicly available RNA-seq data were used to benchmark existing strategies and to determine guidelines regarding study design and input data quality. The here presented pipeline is able to use RNA-seq data for different types of analyses like differential expression analysis, gene fusion detection or variant calling and thanks to the flexible and modular architecture new features can be added without great effort. In the course of the analysis process, various results are inserted into a relational database which can be browsed via a convenient and user-friendly web interface. The pipeline constitutes a valuable tool for biologists or clinicians but also for bioinformaticians and could already demonstrate its feasibility and utility in several projects.

Zusammenfassung

Die Entwicklung der Next-Generation Sequencing Technologie und RNA Sequencing (RNA-seq) im Speziellen ermöglicht die Untersuchung des gesamten Transkriptoms eines Individuums. Die daraus resultierenden Daten beinhalten eine Vielzahl von nützlichen Informationen und mehr als 1.500 RNA-seq Proben wurden im Zuge dieses PhD Projekts analysiert. Um diese Daten zu verarbeiten und die gewonnenen Informationen zu interpretieren ist jedoch passende Software nötig. Mittlerweile gibt es eine Vielzahl von Programmen für die Analyse von RNA-seq Daten, jedoch ist deren Nutzung oft komplex und zeitaufwändig. Hinzu kommt, dass in unserer Einrichtung pro Woche große Mengen an Daten produziert und in weiterer Folge analysiert werden müssen. Eine schnelle und einfache, aber auch zuverlässige und den Anforderungen entsprechende Verarbeitung der anfallenden Daten ist daher entscheidend.

Das Ziel dieses PhD Projekts war die Konzeptionierung und Entwicklung einer automatisierten RNA-seq Daten Analyse Pipeline, welche anschließend mit einer bereits bestehenden Exome-Sequenzdaten Analyse Pipeline zusammengeführt werden sollte. Des Weiteren wurden Methoden und Parameter für einzelne Analyseschritte untersucht und eingeführt, um zuverlässige Ergebnisse zu gewährleisten. Dazu wurden sowohl interne als auch öffentlich zugängliche RNA-seq Daten verwendet, um verfügbare Strategien zu vergleichen sowie Richtlinien bezüglich Studiendesign und Qualität des Startmaterials festzulegen. Die hier präsentierte Pipeline ist in der Lage RNA-seq Daten hinsichtlich differenzieller Expression, Fusionstranskripte als auch genetischer Mutationen zu untersuchen und Dank des flexiblen und modularen Aufbaus können weitere Analyseschritte ohne großen Aufwand hinzugefügt werden. Im Zuge der Datenanalyse werden zahlreiche Ergebnisse in eine Datenbank eingefügt und diese Daten können dann mittels einer benutzerfreundlichen Webapplikation durchsucht werden. Die Pipeline unterstützt Biologen und Kliniker aber auch Bioinformatiker gleichermaßen bei der Auswertung von RNA-seq Daten und konnte schon in einigen Projekten erfolgreich eingesetzt werden.

Contents

List of Figures

List of Tables

Part I.

Introduction

1. Introduction

The genome is thought to be the totality of hereditary information of an organism. According to the Stanford Encyclopedia of Philosophy is the term "genome" hard to define and to describe it solely as informational content is not accurate enough since it is not a stable but rather an adaptive entity which is constantly changing in response to the environment[232]. Thus, they suggest to understand it as "[...] a process, a highly complex set of dynamic activities crucial in maintaining the structural and functional stability not only of the organism but also, through its role in reproduction, of the lineage"[232]. Alternatively, two tenable ways of defining the genome is either as "[...] the sequence of nucleotides" or as "[...] a material object, presumably, in most cases, the nuclear chromosomes"[232]. Generally, a chromosome consists of a genomic deoxyribonucleic acid (DNA) molecule as well as proteins. DNA molecules are macromolecules that consist of two strands forming a double helix and each DNA strand is composed of four different units called nucleotides (adenine (A), cytosine (C), guanine (G) and thymine (T)). The specific succession of these four nucleotides encodes the actual genetic information which characterizes the phenotype of a cell and in further consequence of the entire organism. Different regions of the nucleotide sequences fulfill different tasks and a major one is to encode for functional products like proteins or nucleic acids such as transfer ribonucleic acid (tRNA) or ribosomal ribonucleic acid (rRNA). These functional products are the actual transmitter of the genetic information to the cell. DNA regions coding for functional products are called genes[165, p. 11]. On the other hand, there are other sections in the DNA not serving as template for functional products that possess important regulatory functions.

1.1. Gene Expression

DNA itself is unable to release its information to the cell directly, for this purpose, intermediate steps are necessary. The flow of genetic information became known as the central dogma of molecular biology[236, p. 13] and was proposed by Francis Crick [38][39]. According to his dogma and as shown in Figure 1.1, the most common transfer of genetic information is either from DNA to DNA (replication) or from DNA to RNA to protein (gene expression). However, there are special cases where RNA can be replicated or reverse transcribed into DNA and DNA directly translated into polypeptides whereby the latter information transfer could only be performed on cell-free extracts and under laboratory conditions[250].

DNA replication is the process by which an exact copy of the DNA is produced.

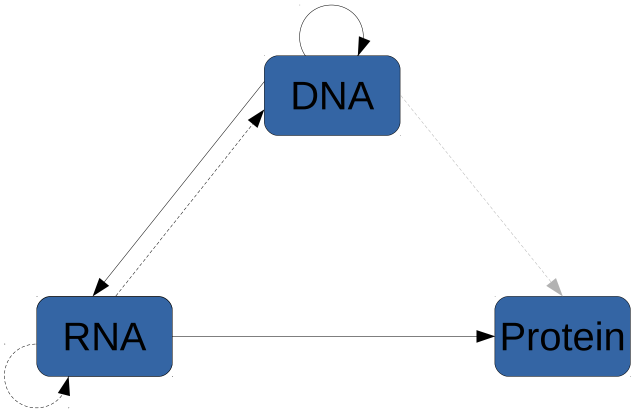


Figure 1.1.: The central dogma as published by Francis Crick in 1970. Solid arrows indicate common transfers of information which can take place in a cell whereas dotted arrows indicate special transfers occurring only in special cases. (Figure adapted from Crick, 1970[39])

The process that uses the genetic information stored in the DNA in order to produce functional products is called gene expression.

Two major steps are part of the process of gene expression, namely transcription and translation. Transcription denotes the production of ribonucleic acid (RNA), itself a macromolecule made up of four nucleic acids (adenine (A), cytosine (C), guanine (G) and uracil (U)). Based on the information stored in genes, RNA is a transcript of that information where again the particular sequence of the four nucleotides encodes the information. In contrast to DNA, RNA is often single stranded and can be assigned to different categories. The major distinction is between RNAs coding for proteins, so called messenger RNAs (mRNAs), and non-coding RNAs (ncRNAs). The latter ones can be subdivided into further categories, e.g. tRNAs or rRNAs to name but a few. These non-coding RNAs will not be further translated into polypeptides but are already functional end products with a wide variety of functions. mRNAs, on the other hand, derive from protein coding genes and thus are capable of being translated into polypeptides. The full range of mRNA molecules within a cell is termed transcriptome[24, p. 70]. Other than DNA or RNA, polypeptides consist of amino acids. In order to translate the encoded information of mRNA into polypeptides, the process of translation makes use of the genetic code. This code defines how the nucleotide sequence is decoded. The underlying rule is that three consecutive nucleotides, named codon, can be unambiguously translated into one amino acid and the particular succession of nucleotides determines the resulting amino acid sequence. Finally, one or more polypeptides make up a functioning protein which is now able to release the genetic information to the cell. The total abundance and composition of all proteins in a cell is termed proteom which, in the end, determines the nature of the biochemical reactions that the cell is able to carry out. Although transcription and translation are important steps of gene expression there are a number of additional processes involved in the regulation of the transcriptome and the proteom.[24, pp. 70ff.][95, pp. 169ff.]

1.1.1. Regulation of Gene Expression

As already mentioned, the way a cell behaves is encoded in the DNA that transduces its information via RNA and consequent proteins to the cell and almost every cell in an individual organism possesses the same genetic material. However, in multicellular organisms the functions and characteristics of different cell types differ. This is because the presence of individual transcripts and in further consequence the repertoire of proteins in a cell are a crucial factor. Thus, depending on a cell's needs it expresses only a distinct proportion of genes.[236, pp. 13ff.] On top of that, not only the pure absence or presence of a gene product is decisive but also its abundance.

Several regulatory processes are involved in the decision of which gene will be expressed to what extent. An overview of the key regulatory processes is listed below[236, pp. 19ff., 276ff.][95, pp. 232ff.]:

- **Transcriptional Regulation of Gene Expression:** In eukaryotic cells the synthesis of mRNA takes place in the nucleus using DNA-dependent RNA polymerase enzymes. The initial regulatory mechanism of gene expression affects the initiation of the transcription of a gene and the processivity of the respective RNA polymerase. Essential influencing factors for this are, on the one hand, regulatory regions in the DNA like promoters, enhancers or silencers and, on the other hand, proteins that are able to recognize those regions. The initiation of the transcription of a gene happens by binding of the RNA polymerase to an according promoter whereby a couple of genes are known to have more than one promoter enabling the cell to produce alternative transcripts from the same gene. However, regulatory regions can be modified by a process named DNA methylation, thus affecting initiation of gene expression. Furthermore, histone modifications can lead to a change in chromatin structure which can also influence the expression of a gene. Moreover, the binding affinity and processivity of the RNA polymerase and as a result the cellular level of the respective transcript are greatly influenced by other proteins that are capable of binding to various proximal regulatory regions. Those transcription factors together with the RNA polymerase and regulatory DNA regions are therefore the first control elements of transcription.

- **Post-transcriptional Regulation of Gene Expression:** Once synthesized most of the RNA molecules undergo further processing steps. One of them is splicing by which several different versions of a transcript can arise out of a precursor RNA. This is possible as a lot of RNAs have regions called exons and introns.

The former ones are regions that are actually serving as templates for the synthesis of functional products whereas the latter ones are not and they are located between the exons. Thus, splicing takes place by removing intronic regions and subsequently joining together the exons. Provided that a precursor RNA comprises several exons the diversity of the resulting product can be achieved by joining together exons in various different ways.
Another processing step of mRNA molecules is 5′ capping where a methylated nucleoside is attached to the 5′ end of a transcript. This cap, in turn, is capable of influencing splicing, the degradation rate and the transport to the designated target of the respective transcript.
On top of that, the majority of mRNAs become polyadenylated. Polyadenylation is defined as the linking of a poly(A) tail to the 3′ end of the transcript. The exact position where this tail is placed enables once more the production of several different versions of the mature mRNA.
RNA editing, a further regulatory mechanism, has a direct impact on the nucleotid sequence of the RNA molecule by inserting, deleting or replacing nucleotides and thereby directly changing the information stored in the RNA. The most frequent type of RNA editing is the conversion of adenosine to inosine (A→I editing). However, the reverse transcriptase interprets the inosine as a guanosine, thus when RNA is sequenced the A→I editing appears as an A→G mutation[171].
In addition to that there are other regulatory mechanisms acting post transcriptional like RNA interference (RNAi) that influence the repertoire of functional products by inducing the degradation of RNA or the repression of its translation[198].

- **Translational Regulation of Gene Expression:** The translation of mature mRNAs into proteins takes place in the cytoplasm of eukaryotic cells and is carried out by ribosomes. Ribosomes are ribonucleoproteins, hence made up of rRNAs and proteins, and consist of two subunits. Although the noncoding introns are already removed, mature mRNAs still possess regions that do not code for proteins. These untranslated regions (UTRs) are located at the 5′ and 3′ ends of the coding region, respectively. Along with the 5′ cap and the 3′ poly(A) tail they are important elements for determining the binding of the ribosome to the according mRNA and thus are able to modulate translation. Depending on that, the initiation of translation takes place by binding of one of the two ribosomal subunits to the 5′ cap. This unit moves along the RNA molecule until it reaches the start codon, the signal for the ribosome to start translation. Once arrived at the start codon the second subunit binds to the first one and the protein synthesis starts with the help of tRNAs which provide different amino acids to the growing polypeptide chain based on the succession of codons of the mRNA molecule. This is done until a stop codon is reached.

- **Post-translational Regulation of Gene Expression:** During or after protein synthesis the originated polypeptides can be modified in order to alter their stability, activity or destination[150]. One possible post-translational modification is the addition of modifying groups to the polypeptide chain. Phosphorylation for example, a common chemical modification[106], is important for the activation or deactivation of enzymes. Another post-translational modification is cleavage by which a long polypeptide results in a smaller one with altered functionality and stability.

1.1.2. Gene Expression Profiling

Since RNAs and proteins and their respective composition within a cell are important determinants for its activity, the investigation of expression levels allow interesting insights into a cell's and in further consequence into an organism's state. In general, it is possible to examine the abundance of RNAs and also proteins. As already discussed in the previous chapter, there are several regulatory mechanisms acting on intermediate gene products influencing not only RNA but also protein abundances. The RNA repertoire of a cell is a major parameter for its properties as changes in a cell's state are often associated with alterations of mRNA levels and therefore the examination of the transcriptome can give valuable insights[191, p. 1].

Initial methods for mRNA abundance measurement include northern blots[105] and qPCR[69]. These techniques are suitable for the investigation of a limited number of genes at a time and as a consequence they can be useful if one is interested in just a few candidate genes. However, they are impractical when the aim is to measure thousands of genes or even the entire transcriptome at one go which is a common scenario in modern research studies. Furthermore, there are several advantages when investigating the entire transcriptome. For example, whole transcriptome analysis makes it possible to compare expression levels of all expressed genes between different conditions, thus it allows the identification of all genes with differing expression patterns across conditions.

More recent technologies like serial analysis of gene expression (SAGE)[251] or the consequent Super-SAGE[156] made it possible to analyze RNA abundances of thousands of genes at once. Another and former commonly used[102] high-throughput gene expression analysis technology are DNA microarrays[220]. Microarrays are hybridization-based[255] which means that, in terms of gene expression profiling, they are based on the hybridization of labeled RNA molecules to pre-defined DNA sequences which are attached to the microarray surface. These DNA sequences usually represent sequences of known genes of the organism of interest. However, microarrays are known to have a number of limitations[76][154][177][255][267]. For example, since the analysis is based on pre-defined sequences only transcripts whose sequence is already known can be analyzed. Thus, new isoforms or previously unknown genes will be missed. Another problem is cross-hybridization leading to high levels of background noise due to unwanted binding of sequences to incorrect targets[175]. Furthermore, targets on the microarray have varying hybridization properties which biases the expression quantification as well and makes comparisons more difficult[154].

In comparison to microarrays, sequencing-based methods like RNA-seq have several advantages and do not possess the abovementioned limitations. RNA-seq is explained and discussed in more detail in Chapter 1.3.

1.2. DNA Sequencing

DNA sequencing is defined as the determination of the actual order of the four nucleotides adenine, cytosine, guanine and thymine of a DNA molecule. In 1977, Allan M. Maxam and Walter Gilbert proposed a method for DNA sequencing which is also known as chemical sequencing[157]. Although used for some time it was mostly replaced be another technology published by Frederick Sanger and colleagues also in 1977[218]. Their method is known as dideoxy method, chain termination method or, named after its inventor, Sanger sequencing. It utilizes dideoxynucleotides (ddNTPs) which in comparison to deoxynucleotides (dNTPs) lacks the 3′-hydroxyl group and thus, other than dNTPs, leads to the termination of DNA chain elongation when incorporated. To perform Sanger sequencing the double stranded DNA molecule that should be sequenced has to be denaturated into single stranded DNA. The resulting single stranded molecule serves then as template for the creation of new complementary strands. In order to create the complement of the template sequence, DNA primer, DNA polymerase and dNTPs (dATP, dCTP, dGTP and DTTP) as well as ddNTPs (ddATP, ddCTP, ddGTP and ddTTP) are needed. For the actual sequencing process the single stranded DNA template is divided into four reactions each of which containing DNA polymerases, all four types of dNTPs but only one type of ddNTPs, i.e. one reaction only containing ddATP, one only ddCTP, one only ddGTP and one only ddTTP, with the concentration of ddNTP in each reaction just amounting to a fraction of that of dNTPs. The DNA polymerase then starts to create the complementary DNA strand based on the template strand by adding the respective dNTPs. Randomly ddNTPs instead of dNTPs are incorporated into the growing sequence which stops the elongation resulting in DNA molecules with varying length. After several rounds of synthesis the resulting molecules are sorted by size and since each fragment can be assigned to one of the four reactions one knows which ddNTP is last in each sequence and hence the complementary sequence of the input template can be reconstructed.

After its invention Sanger sequencing was gradually enhanced which laid the foundation for automated sequencing machines, also referred to as first-generation sequencing technology, and led to a decrease in sequencing costs and increased quality and sequencing length. As a result, it currently allows to sequence reads with a length of up to about 1,000 base pairs (bp) with an accuracy of 99.999%[225]. Due to its strength Sanger sequencing has been the most widely used DNA sequencing technology for nearly three decades and is still used for particular projects and issues, e.g. validation. However, after the completion of the human genome in the course of the Human Genome Project, which was accomplished by means of Sanger sequencing, cost almost three billion US dollar and took nearly 13 years[73][93][161], the interest in sequencing increased greatly which led to the development of even faster and cheaper sequencing technologies named second-generation or next-generation sequencing.

1.2.1. Next-Generation Sequencing

Next-generation sequencing (NGS) is the massively parallel sequencing of DNA producing millions of short fragments simultaneously[161][225]. Its development and the resulting decrease in cost and time revolutionized the field of genomics. In the period between 2005 and 2007 three companies introduced distinct NGS platforms namely Roche (454 sequencing)[153], Illumina (Solexa technology)[18] and LifeTechnologies (ABI SOLiD sequencing)[159]. Although all of them have their strengths and weaknesses the Sequencing by Synthesis (SBS) method proposed by Illumina prevailed and all data discussed in this thesis have been produced by means of Illumina's sequencing technology. Since its invention Illumina gradually refined their technology which is currently able to generate more than 10,000 gigabases (Gb) per week at a price of below 10 US dollar per Gb (Figure 1.2).

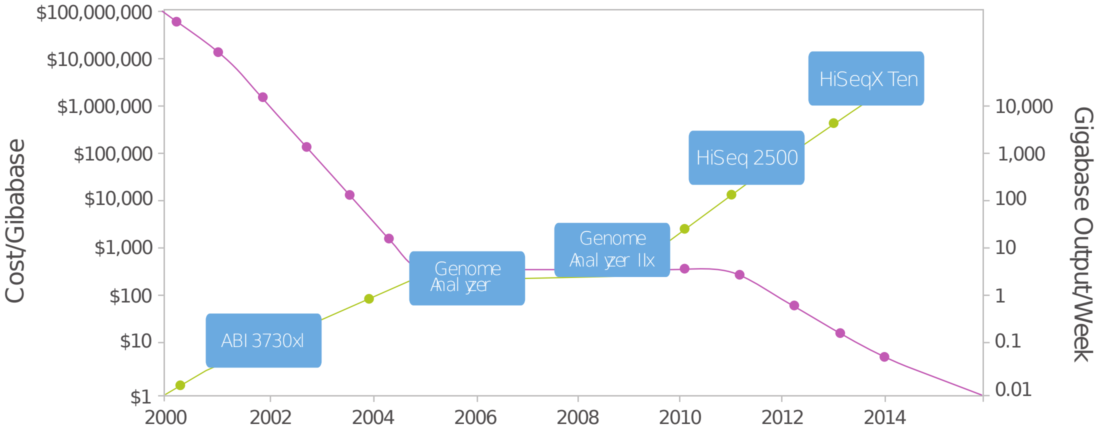


Figure 1.2.: This graph illustrates the trend of price and output of several Illumina instruments between 2000 and beyond 2014 (x-axis). As can be clearly seen there is a remarkable increase in sequencing output while costs are falling. The left y-axis shows the sequencing costs per Gb in logarithmic scale (violet graph) and the right y-axis the weekly output also in logarithmic scale (green graph). (Figure taken from [89])

The basic steps included in Illumina's sequencing technology are as follows:

1. **Library Preparation:** First of all, the sample that should be sequenced has to be prepared. This involves the fragmentation of the DNA into smaller pieces.

As the resulting fragments have random length and neither too long nor too short fragments are convenient for the Illumina sequencing technology only fragments with appropriate size are kept. Finally, specific adapters are ligated to the fragments which enables the attachment of the fragments to the flow cell (Figure 1.3).

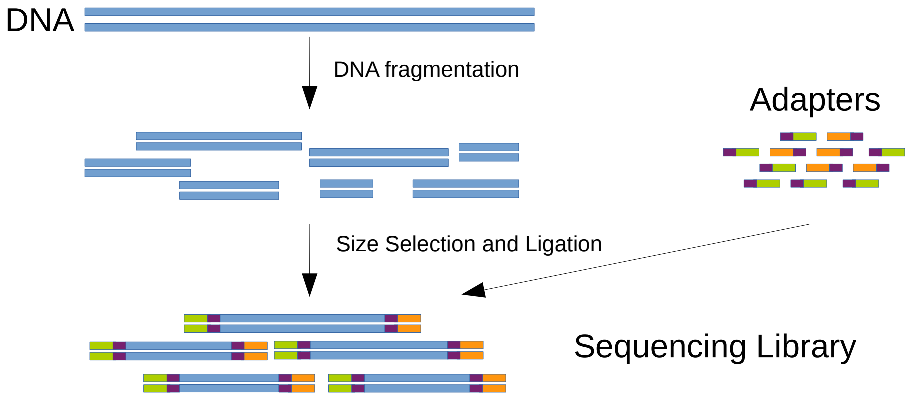


Figure 1.3.: Schematic illustration of the Illumina Library Preparation step. Initially, genomic DNA is fragmented and after that specific adapters are ligated to the resulting, size selected fragments. (Figure adapted from [89])

2. **Attachment to Flow Cell:** A flow cell is a glass slide with eight lanes. A lane is defined as a channel and each channel contains a lawn of oligonucleotides[85]. Using a machine named Cbot the prepared DNA fragments of the samples to sequence can be attached to the flow cell (Figure 1.4). This is possible as the oligonucleotides which are attached to the flow cell are complementary to the specific adapters ligated to the fragments in the library preparation step.

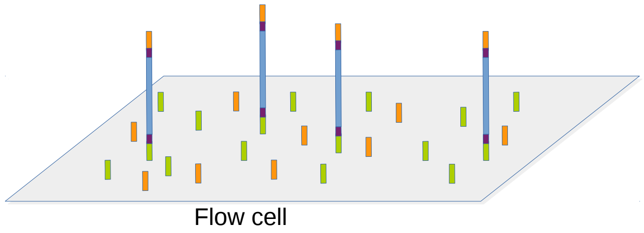


Figure 1.4.: The prepared fragments are attached to the flow cell by randomly binding to the present oligonucleotides. (Figure adapted from [89])

3. **Cluster Generation:** Since the sequencing process is based on the detection of a fluorescent signal that is emitted when a labeled dNTP binds to a fragment and the signal from a single incorporation process would be too weak the fragments

have to be amplified. This is done by a process called bridge amplification (Figure 1.5) by which the single fragments are copied multiple times in order to produce dense clusters constituted of up to 1,000 fragments with identical sequence information.

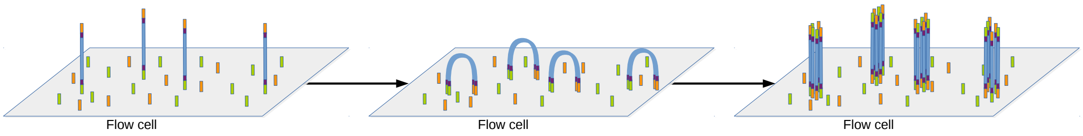


Figure 1.5.: Each fragment attached to the flow cell is amplified in multiple cycles in order to create clusters. (Figure adapted from [89][152][256])

4. **Sequencing:** Once the fragments are amplified the flow cell is ready to be transferred to the sequencer. Here, in a first step, an universal adapter for sequencing is hybridized to the single stranded fragments. Sequencing is then performed in cycles where in each cycle the complementary sequence of the fragment is extended by one base (Figure 1.6). This is possible as in each cycle DNA polymerases and modified dNTPs are washed through the flow cells and the polymerase extends the appropriate dNTPs to the growing sequences. The modification that is made to the dNTPs comprises the use of a reversible terminator with four different removable fluorophores [18], one for each type of dNTP (dATP, dCTP, dGTP and dTTP). This modification ensures that only one dNTP can bind to the growing sequence per cycle and that the type of the incorporated dNTP can be detected. The surplus of polymerases and dNTPs is washed away and the incorporated bases are identified by laser-induced excitation of the fluorophores and imaging of the signal[18]. Subsequently, the terminators and fluorophores are removed and a new cycle can start. Based on the fluorescent signals detected in each cycle the Illumina software assigns the according base to each cluster in a process called base calling. In this way, decoding the actual sequence of each fragment that was loaded on the flow cell.

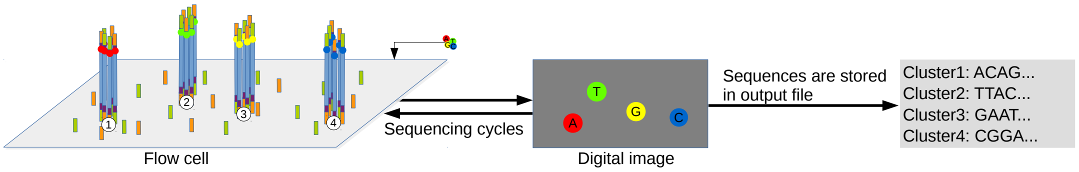


Figure 1.6.: Schematic illustration of the Illumina Sequencing by Synthesis technology. In each cycle an according fluorescently labeled dNTP is incorporated to the growing sequence and the respective emitted fluorescing signal is detected by a camera. (Figure adapted from [89])

In the early days the Illumina technology was able to perform 35 cycles[18], i.e. sequence 35 bp per fragment. The rather small number is mainly due to the fact that possibly not all fragments of a cluster incorporate a dNTP in each cycle leading to a biased signal in the following cycles where mixed fluorescent signals might be detected as not only the actual correct dNTP binds to the cluster but also the previously missed ones. As a result, the more cycles performed, i.e. the longer the sequenced read, the more the quality of the called bases suffers. This phenomenon is called dephasing. However, the chemistry and reagents improved over the years and by the time of writing the Illumina systems are able to achieve read lengths of up to 300 bp[90].

Furthermore, with Illumina sequencing systems it is possible to perform paired-end sequencing (Figure 1.7). This means that the fragments are sequenced from both ends which offers several advantages for data analysis (see Chapter 1.4.1).

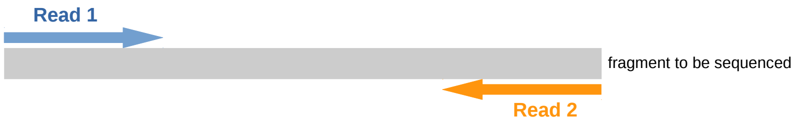


Figure 1.7.: Schematic illustration of the paired-end sequencing method where each fragment is sequenced from both ends. (Figure adapted from [89])

Additionally, not only read length but also the number of fragments that can be sequenced per run increased. This was possible thanks to the improved sequencing chemistry but also due to refined optical systems. Nowadays, Illumina systems are able to produce up to 1,800 Gb of sequence in less than three days[90].

1.2.1.1. Applications

As stated by Grada and Weinbrecht [73] “The applications of NGS seem almost endless [. . .]”. It allows the investigation of the genome, the transcriptome or the epigenome of any organism[89]. Although providing an opportunity to answer a multitude of different questions the sequencing process itself stays the same but the way how the sequencing material is obtained and prepared and the final data analysis make the difference.

First of all NGS can be used for *de novo* assembly, i.e. to reconstruct the genomic sequence of an organism without using a reference genome. In this process the short reads produced by the sequencer are searched for overlaps and assembled into larger fragments, so called contigs, thus trying to trace back the entire genomic sequence of the investigated organism.

Another widely used application of NGS is the detection of disease associated variants. The most sensible way to do this is by sequencing the entire genome, i.e. whole genome sequencing (WGS). Although there is an ongoing decrease in sequencing costs, it is reasonable for several cases not to sequence the entire genome but only specific parts of it. Several targeted sequencing methods are available for this purpose. One of them is amplicon sequencing which is suitable to interrogate rather small regions. Further targeted sequencing methods are gene panels, where only a limited number of genes of interest are captured[15]. This approach is useful, for example, when several known disease genes should be analyzed for a multitude of samples. However, if the number of genes to investigate gets too big it is preferable to use another, more comprehensive method where all genes of an organism can be captured and this method is referred to as whole exome sequencing (WES)[169].

Beside the already mentioned applications NGS can be used in epigenetics as well. Chromatin immunoprecipitation followed by NGS (ChIP–seq)[190] is used to detect protein-DNA or protein-RNA interactions *in vivo*. Therefore, it is a widely used technique to detect binding sites of transcription factors and other DNA-binding proteins which are important for the understanding of regulatory mechanisms. The analysis of cytosine methylation is another application of NGS. The method for this purpose is called bisulfite sequencing.

Finally, NGS cannot only be used to study the genome or epigenome but also the transcriptome. A more detailed discussion regarding the application of NGS in the study of the transcriptome can be found in the following Chapter 1.3.

1.3. RNA Sequencing

RNA sequencing (RNA-seq) is a method that utilizes NGS technology in order to perform transcriptome profiling[164][166][255][258]. It provides a snapshot of the transcriptome at a specific point in time, has several advantages and the potential to overcome the limitations that are associated with microarrays[154] (limitations of microarrays are discussed in Chapter 1.1.2).

By this time, RNA-seq has become a powerful technology with a big variety of applications. First of all, it can be used to quantify gene expression[164] and in further consequence perform differential gene expression analysis, which is the identification of genes whose expression differs among different experimental conditions[177]. Since RNA-seq does not rely on a priori sequence knowledge it can be used to detect novel transcripts as well[77]. In addition to that one can interrogate alternative splicing (AS)[188][238][253] but also gene fusion events[148]. Aside from that, RNA-seq demonstrated its potential to identify genomic variants to study allele specific expression (ASE)[45] or RNA editing[193][215].

Depending on the biological question that should be answered the sample preparation steps involved in each RNA-seq experiment can slightly differ. A typical RNA-seq experiment workflow (Figure 1.8) can be summarized as follows.

First of all, total RNA is extracted from the sample to analyze. Subsequently, the desired RNA subpopulation is enriched. This is done as >90% of a cell's RNA repertoire consists of ribosomal RNA (rRNA)[257]. A common and widely used method for mRNA enrichment is poly(A) capturing by which oligo(dT) beads are used to select solely polyadenylated RNAs. Another method is rRNA depletion where complementary rRNA sequences conjugated to magnetic or biotinylated beads are used to get rid of the abundant rRNA[265]. Once the favored RNA is enriched, the remaining molecules are sheared into smaller fragments and the resulting snippets are converted into cDNA. Both steps are necessary since the Illumina sequencing technology is only able to cope with DNA fragments of a specific length[257]. The resulting cDNA molecules are then amplified and sequenced as described in Chapter 1.2.1.

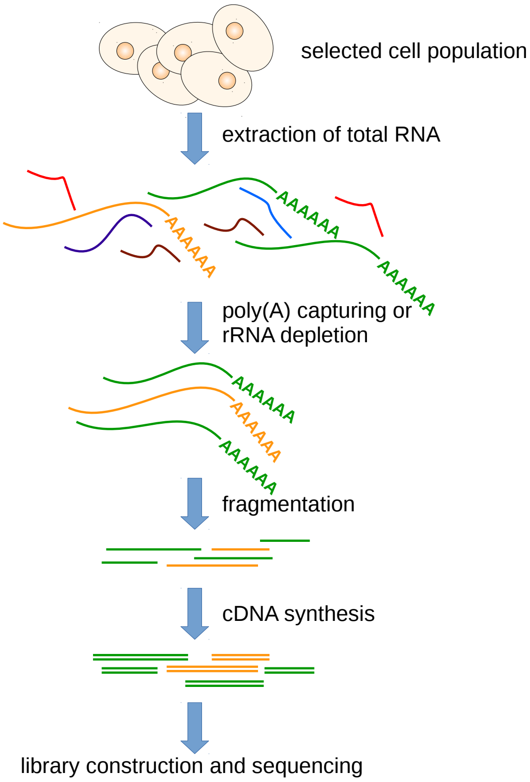


Figure 1.8.: Overview of a typical RNA-seq library preparation workflow. Total RNA is extracted and mRNA is enriched. After that the mRNA molecules are fragmented and converted into cDNA which is then used to prepare a library ready for sequencing. (Figure adapted from [265])

The abovementioned workflow can be applied to answer multiple questions. However, there are special cases where specific steps in the preparation have to be modified or added. One of these special cases, for example, is 4sU-tagging[51] which can be used to interrogate changes in RNA synthesis as well as RNA decay by labeling newly transcribed RNA in living cells with 4-thiouridine (4sU)[160]. The labeling process is performed for a specific amount of time, then total RNA is extracted and after thiol-specific biotinylation the labeled, newly transcribed RNA can be separated from the pre-existing one with the aid of streptavidin-coated magnetic bead[51]. The extracted total RNA but also the labeled as well as unlabeled fraction of RNA are applicable for subsequent RNA sequencing[259].

Another advancement of RNA-seq is the possibility to perform low-input or single-cell RNA sequencing. In common RNA-seq experiments total RNA is taken from a sufficient number of cells and the extracted RNA composite provides a snapshot of the average expression profile of these cells[224]. Although this is suitable for a lot of cases there are special experimental setups where only a limited number of cells is available or where one is interested in the behavior or properties of single cells. For this purpose, common RNA-seq library preparation methods are not suitable since they typically require micrograms of RNA as starting material, in which case thousands of cells are needed[240]. When dealing with very low amounts of RNA factors like contamination, degradation or even sample loss might mess up the quality of the results[168]. Therefore, methods especially designed to handle small quantities of starting material have to be used such as SMART-seq[204] or CEL-Seq[84] which are able to substantially amplify the extracted RNA population while preserving the relative abundances of them[53]. Applications for small input or single cell RNA-seq are, for example, the identification of differences in cells even if they are morphologically indistinguishable[1] or to interrogate expression patterns of rare cell types such as circulating tumor cells[204].

1.4. RNA-seq Data Analysis Workflow

Once sequencing on an Illumina sequencing instrument finished the resulting binary base call files (BCL) can be converted into a more common file format named FASTQ. A FASTQ file is a plain text file that stores the sequence information of each sequenced fragment as well as the corresponding base quality score. An example of the FASTQ file format is shown in the following:

```
@SND00115:164:C86HWANXX:1:1104:1248:2037 1:N:0:CGTACTAG
ATTCCTGATACTCCTGCCTCCAGCTCTGGATTGTAGGCATGCACTACCATA
+
BBBBBFFFFFFFFFFFJFFFFFFFJJJJJJFFFFFFFF<<FFFJJFF##FF
...
```

By definition, each fragment is represented by four line types where the first line starts with an ’@’ character followed by an unique identifier and an optional description, the second line stores the actual nucleotide sequence information and can theoretically be wrapped into multiple lines, the third line type starts with a ’+’ character optionally followed by a copy of the unique identifier of the first line and finally, the fourth line type, which stores the base qualities of each base of the second line[34]. Nowadays, base calling qualities are usually represented by PHRED scaled quality scores which are calculated as $q_i = -10log_{10}(p_i)$ where q_i is the PHRED score for the estimated error probability p_i for base i[59][60]. The collectivity of base quality scores is already an early and crucial quality control parameter to check whether the sequencing process was successful and hence if the reads can be used for further analysis.

Typically, each FASTQ file stores millions of short sequence reads of just a single sample. Moreover, it is not unusual that the sequence reads of one sample are distributed among multiple FASTQ files. This is particularly the case if a sample's sequence library is sequenced on multiple lanes on a flow cell which is common practice to prevent sources of bias like lane effects, for example.

Regardless of how many FASTQ files per sample are produced they constitute the starting point for the downstream data analysis. The structure of a common RNA-seq data analysis workflow is shown in Figure 1.9 and each step involved is described in the following sections.

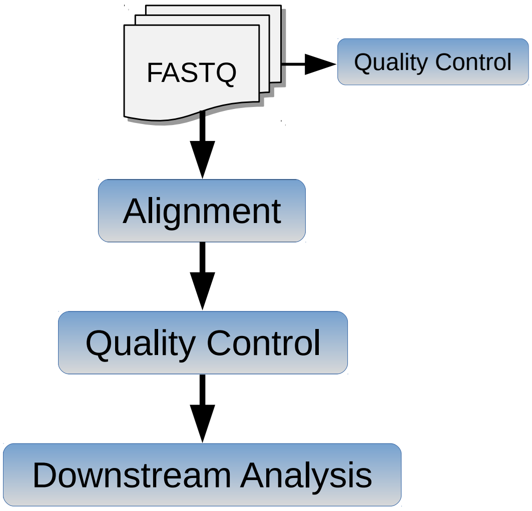


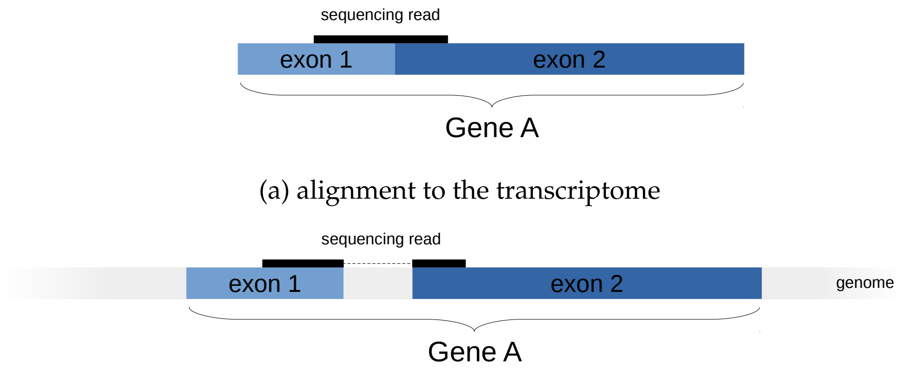
Figure 1.9.: This figure depicts a standard RNA-seq data analysis workflow. Starting point are FASTQ files storing the sequence information for each sample. Usually these files are checked for quality and subsequently aligned. After alignment, quality control is performed once again and good quality alignments are kept for downstream analysis like differential expression analysis, fusion detection or variant calling.

1.4.1. Alignment

The reads produced by an Illumina sequencing instrument have limited length (see Chapter 1.2.1), thus one single read rarely represents the entire RNA molecule of origin. Taking these yet rather uninformative reads and trying to give them a meaning is a fundamental step in RNA-seq data analysis[61]. There are several approaches to do so and which one to choose depends to a large extent on the experimental question and whether there is a profound reference sequence as well as gene annotation available for the investigated sample's organism or not.

One approach is to reconstruct the transcripts by assembling the reads based on their overlaps. If, however, a reference genome is available it is also possible to initially align the reads to the reference and subsequently assemble them into putative transcripts[67]. The resulting assemblies can give valuable insights regarding novel transcripts or alternative splicing[19][72].

For well studied organisms like human or mouse both, a sophisticated reference genome as well as gene annotations are available. If this is the case, another way to reveal the origin of the reads is by just aligning them to the reference genome or transcriptome (Figure 1.10).



(a) alignment to the transcriptome

(b) alignment to the genome

Figure 1.10.: Sequencing reads can be aligned either to the transcriptome (a) or genome (b). For alignments to the genome, RNA-seq reads spanning exon-exon junctions must be treated in a special way since they need to map across introns which is not the case for alignments to the transcriptome.

Reference sequences are usually stored in FASTA format. Like FASTQ files, FASTA files are plain text files and have a rather simple format, e.g.:

```
>chr1
AGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAG
TGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTT
AATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAG
...
>chr2
CGATTAACAGGTACCAAAGGATTACAGGAAATATAGGAAGTTAACCACTA
...
```

where each sequence starts with a '>' character followed by an unique identifier for the particular sequence, which in many cases represents entire chromosomes. The subsequent lines constitute the actual sequence.

For the vast amount of reads produced by NGS instruments, formerly developed alignment algorithms such as BLAT[103], SSAHA[170] or GMAP[263] are not efficient enough[126]. Therefore, new short-read aligner have been implemented that are specifically designed to satisfy the requirements presented by NGS data like dealing with short, possibly paired-end sequenced reads that might possess sequencing errors and perhaps map equally well to different sites in the reference genome[249]. In addition to that, RNA-seq reads pose another challenge since they can span exon junctions[67]. The new alignment algorithms exploit sophisticated methods to meet these challenges in a best possible way. Based on the method used they belong to different categories which will be described in the following section[16].

A first distinction can be made based on whether the aligner is splice-aware, i.e. if it is able to allow a read to span an entire intron, with intron lengths commonly ranging from 50 to 100,000 bp in mammalian genomes[107], or not. Aligner that are not splice-aware are especially suitable when aligning RNA-seq reads to a reference transcriptome. Here, the alignments are restricted to all known transcripts included in the reference and the detection of novel isoforms is impossible, thus reads that originated from unknown transcripts might remain unmapped. For the category of splice-unaware aligner two major approaches emerged using either Burrows-Wheeler transform (BWT)[26] or seed methods[67]. Alignment programs utilizing BWT are, for example, Bowtie[118], SOAP2[130] or the Burrows-Wheeler Alignment tool (BWA)[124]. The idea behind these methods is to use BWT to rearrange the reference genome such that similar sequences occur together and to build an index from the resulting data structure that allows for efficient yet accurate mapping of the reads to the reference[61]. Seed methods, on the other hand, are usually based on another indexing technique, namely hash tables[126]. An early generation of NGS short-read alignment programs such as MAQ[127], SOAP[129] or ELAND[18] belong to this category. A more recent representative of this category is Stampy[143]. The basic principle of the seed methods is to divide the query sequences into subsequences, the so called seeds, and try to perfectly map them to the reference. The mapped seeds are then joined and extended by more sensitive alignment methods, such as the Smith-Waterman algorithm[229], to obtain full alignments[16][67]. Overall, most of the splice-unaware aligner were actually developed for the mapping of reads coming from genomic DNA to the reference genome.

Aligner belonging to the second category, i.e. splice-aware aligner, are able to align RNA-seq reads to the reference genome as they support larger gaps in order to allow for reads to span introns, thus enabling the detection of novel isoforms or chimeric transcripts[50]. For splice-aware aligner also two major approaches emerged, namely exon-first and seed-and-extend methods[67]. Exon-first methods, as their name implies, initially try to completely align the reads to the genome without considering large gaps, thus neglecting introns. This step is performed by means of splice-unaware aligners. Reads that could not be aligned in the first step are split into subsequences which are then aligned to the reference genome as well. The expected read clusters are then used to detect potential spliced connections. A prominent and widely used example of an aligner using this approach is TopHat2[107]. The other approach, seed-and-extend methods, just as the abovementioned seed methods split the reads into parts, the seeds, and try to align these seeds to the genome. Again, candidate alignments are then used to find the proper alignment location for each read through iterative extension and merging of initial seeds with the use of more sensitive alignment algorithms[67]. The ultrafast universal RNA-seq aligner STAR[50] or the BWA-MEM algorithm[123] are both included in this category.

Apart from that, other splice-aware aligner, such as GEM[151], emerged that exploit a hybrid strategy where they use an exon-first method for the initial alignment of unspliced reads and then employ a seed-and-extend method for spliced reads[4].

If paired-end sequencing was performed, the typical procedure of alignment algorithms is to initially process both reads of a pair separately, i.e. try to align them independently and subsequently join them, provided that both of them could be mapped. Paired-end reads constitute an advantage inasmuch as they can provide additional information to the alignment algorithm especially for multi-mapped reads, i.e. reads that fit equally well on multiple locations, as the mapping position of one read can help to determine the correct position of its paired read.

For some of the aforementioned alignment programs it is possible to provide gene annotations in order to guide the initial mapping steps[4]. Usually, gene annotations are stored in the General Feature Format (GFF) or General Transfer Format (GTF) file format[1]. Each line in these files represents one feature where a feature might be any region in the genome but is typically an annotated exon or a coding region belonging to a known transcript. A few example lines of a GTF file are shown in the following:

```
chr1  UCSC  start_codon 11866320  11866322  0 + . gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  CDS         11866320  11866406  0 + 0 gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  exon        11866153  11866406  0 + . gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  CDS         11867188  11867247  0 + 0 gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  exon        11867188  11867247  0 + . gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  CDS         11875903  11876010  0 + 0 gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  stop_codon  11876011  11876013  0 + . gene_id "CLCN6"; transcript_id "uc009vne.2";
chr1  UCSC  exon        11875903  11876844  0 + . gene_id "CLCN6"; transcript_id "uc009vne.2";
```

where the first column represents the name of the reference sequence (e.g. chromosome), the second one the data source, the third the feature name, the fourth and fifth the start and end position of the feature, respectively, the following three columns the score, the strand (forward(+) or reverse(-)) and the frame (for coding sequences only; specifies whether the initial codon starts at the first(0), second(1) or third(2) position of the feature) and the last column contains a list of attributes, usually the gene and transcript name or identifier the feature belongs to.

Nowadays, most of the alignment programs use the same output format which is termed Sequence Alignment/Map (SAM) and the companion Binary Alignment/Map (BAM) format, respectively[125]. The latter is a binary version of the former one and includes the same information while requiring less disk space due to compression. Furthermore, BAM files can be indexed which improves seek times con-

[1] `http://www.ensembl.org/info/website/upload/gff.html` (last accessed 01.10.2016)

siderably. Each line in a SAM file represents the alignment of one read. An example alignment of the two reads of a paired-end sequenced fragment is depicted here:

```
HWI_ST081   99      chr16   11846592    255 97M3404N3M  =   11846613    3524    GGAAAACC...GT   ??@DD?:^...8?
HWI-ST081   147     chr16   11846613    255 76M3404N23M =   11846592    -3524   TCTCGGAG...TG   ?:GFD=|_...^C
```

where each of the 11 mandatory and tab-separated columns store a specific information which is explained in the following[2]:

1. **QNAME** - the unique identifier of the sequenced read

2. **FLAG** - bitwise flag

3. **RNAME** - the name of the reference sequence (e.g. chromosome) the read was aligned to

4. **POS** - the position in the reference sequence the read was aligned to

5. **MAPQ** - mapping quality calculated as $[-10log_{10}(x)]$ where x is the probability that the mapping position is wrong

6. **CIGAR** - the CIGAR string comprises information of how the read could be aligned to the reference, e.g. completely, with gaps or just partly

7. **RNEXT** - if paired-end sequencing was performed this field stores the reference sequence to which the mate pair read aligned ('=' indicates same reference as in RNAME)

8. **PNEXT** - same as RNEXT but holds the alignment position

9. **TLEN** - determined fragment length

10. **SEQ** - the read sequence

11. **QUAL** - the PHRED scaled base qualities of the read sequence

Optionally, additional columns can be attached for further information. These columns have to be tab-separated as well and each of them should be in the `TAG:TYPE:VALUE` format where `TAG` is a unique identifier for the column, `TYPE` specifies the format of the `VALUE` field which itself stores the actual information.

To parse and edit a SAM/BAM file a special software package named SAMtools[125] is available.

[2] based on `https://samtools.github.io/hts-specs/SAMv1.pdf` (last accessed 01.10.2016)

1.4.2. Quality Control

Quality control (QC) is an essential procedure when analyzing RNA-seq data both pre- and post-alignment since the entire data generation process is complex and involves many steps. Errors can theoretically be introduced in any of these steps and detecting them before performing any kind of downstream analysis is vital to guarantee reliable results.

It is known that the RNA-seq technology possesses various limitations and thus making it prone to certain errors, artifacts and biases[83]. As stated by Lahens *et al.*[116], previous studies could reveal several sources of errors in RNA-seq experiments such as GC-content and PCR enrichment[3][17], biases originating from different rRNA depletion methods[2], biases caused by random hexamer priming[80] or errors introduced by the sequencing step itself[167]. Furthermore, problems during sequencing can lead to lane effects[154] or an overall decrease in sequencing output. Especially when comparing different samples with each other outliers and batch effects can lead to flawed conclusions[83]. In addition to that, nontechnical incidents like sample swap and mixup, respectively, can mess up downstream analysis when unidentified. Although many sources of these errors are known and improvements of the technology help to overcome certain issues, chances of introducing biases still exist which makes QC indispensable.

Regarding pre-alignment QC, a number of methods exist to check the quality of RNA-seq samples already in the laboratory, i.e. during the sample preparation and library creation process. Being important in order to identify possible problems in an early stage of the analysis workflow and potentially saving money due to exclusion of faulty samples before sequencing, this procedures can reveal just a limited number of issues. For that reason, QC in the subsequent analysis steps is necessary by all means. When sequencing with an Illumina instrument one has access to an on-board tracking system which provides a multitude of quality metrics, e.g. number of clusters, signal intensities or number of filtered reads due to low quality, giving further insight whether the material loaded on the sequencer was of good quality and moreover if the sequencing process is performing well. Once sequencing finished and the FASTQ files are created they can be used for further investigation. Programs like FastQC[11] are able to use FASTQ files as input and create metrices and plots that help to determine whether the sample's quality is sufficient for further analysis. Widely used data in this scope are, for example, number of optical duplicates, overall, lane-wise or sample-wise alignment yield and base quality of the individual reads.

If no problem could be detected up to this stage, the reads of the sample can be aligned (see previous Chapter 1.4.1). The alignment data in turn comprise further important information that can be investigated with tools like RSeQC[254], RNA-SeQC[47] or QoRTs[83]. All of them produce informative metrices to answer a multitude of quality related questions such as how many reads could be aligned to annotated exons in order to generally check whether the library preparation step worked well or the fraction of reads that apparently originated from rRNA in order to see if the rRNA depletion succeeded. Further popular measures concern issues like mapping quality, mapping rate, duplicate rate or insert size (for paired-end sequencing).

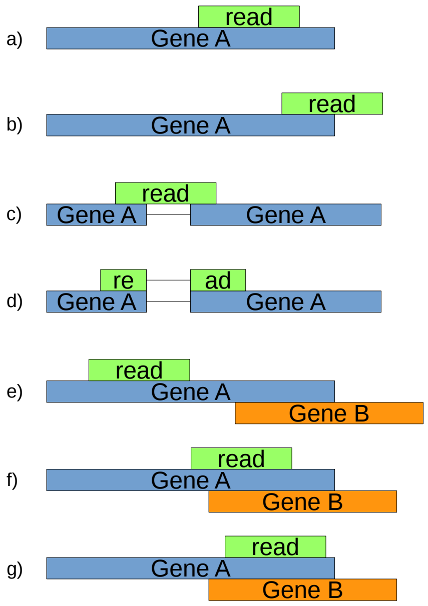
Considering the myriad of possible quality metrices, some sequencing laboratories implemented automated rules in their analysis pipelines which, based on certain cutoff values, decide whether a sample can be used for downstream analysis or has to be resequenced (McSherry, T., Illumina Inc., personal communication, February 02, 2016).

1.4.3. Downstream Analysis

Once the sample is sequenced, aligned and proved to be of good quality all further analysis steps can be performed. Based on the question that should be answered various steps can be performed.

1.4.3.1. Expression Quantification

A common use case of RNA-seq is expression profiling. In order to determine the expression pattern of a sample, original transcript abundances are estimated based on the number of reads that align to respective regions in the genome. In many cases abundances are summarized gene-wise, i.e. all reads mapping within the annotated region of a gene are assigned to that gene. This fairly simple task can be performed using software tools like featureCounts[131] or HTSeq[9], or more specifically its subcomponent htseq-count. These tools make use of the alignment information in SAM/BAM format as well as the annotation in GFF/GTF format to calculate the respective quantitative values. However, summarizing reads in a gene-wise manner is just one option. Theoretically, instead of genes any other type of genomic feature, e.g. exons, introns, UTRs or even self-defined regions, can serve as measuring unit. Although virtually straightforward several issues must be beared in mind when summarizing reads of genomic features. First, reads that cannot be uniquely aligned to the genome, so called multi-mapped reads, must be dealt with. A simple way to handle multi-mapped reads is to either discard them or to count them multiple times for each mapping position. htseq-count uses the former strategy while featureCounts provides the user the option to choose between both of them. Second, genomic features can overlap in their genomic regions. Thus, the quantification algorithm must be able to decide what to do with reads mapping to such overlapping regions. For example, htseq-count allows to choose between three options of how to assign those reads (Figure 1.11). Although gene-wise quantifications are widely used, assigning reads to different isoforms of a gene (if present) can give additional valuable biological insight. However, this is not a trivial task as isoforms often share parts of their sequence and thus reads that align to common regions are difficult to assign uniquely to a single transcript. In order to deal with this issue, algorithms using statistical models have been implemented which try to estimate the expression level of each isoform[4][12][98].



	union	intersection-strict	intersection-nonempty
a)	Gene A	Gene A	Gene A
b)	Gene A	no feature	Gene A
c)	Gene A	no feature	Gene A
d)	Gene A	Gene A	Gene A
e)	Gene A	Gene A	Gene A
f)	ambiguous	Gene A	Gene A
g)	ambiguous	ambiguous	ambiguous

Figure 1.11.: This figure illustrates the behavior of the three htseq-count algorithm modes of how to deal with reads mapping to overlapping features. The three modes are union, intersection-strict and intersection-nonempty. While in many cases the behaviour is the same, example b), c) and f) show the specific differences between the three modes. (Figure adapted from [88])

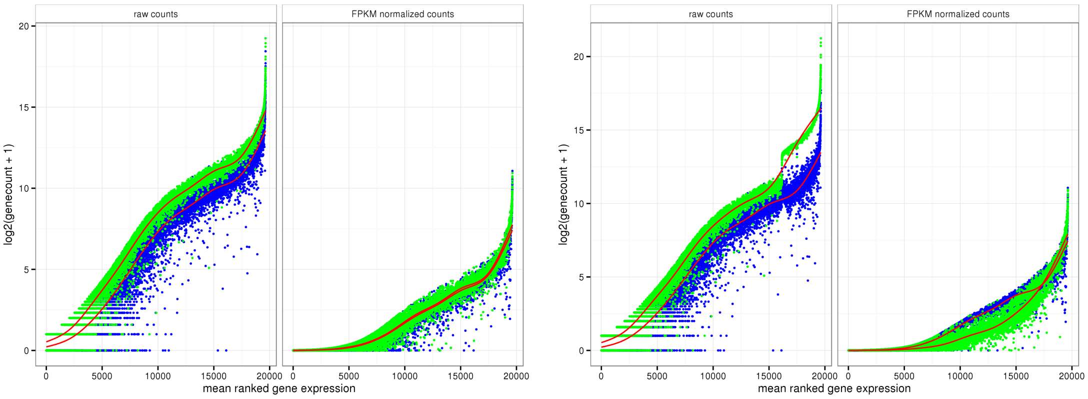
The output of expression quantification tools is usually a file in tabular form where the rows represent the particular feature and the columns the respective samples.

Normalization

The obtained expression values of the previous step are commonly used for comparisons either of different features within a sample (e.g. difference in expression of various genes) or of the same feature across different samples (e.g. differential expression of a gene between samples). However, prior to performing any kind of comparisons, normalization is necessary in order to get meaningful expression values as RNA-seq count data are subject to certain biases[67][164][166][238][255]. Two major categories of biases have been reported: within-sample and between-sample bias[48][210]. The former category must be taken into account especially when performing within-sample comparisons, i.e. comparing different features of the same sample. The main source for within-sample bias is the different length of the respective features[79][195][211][215] since longer features result in more reads due to RNA fragmentation during library construction[67][178]. This means that the obtained expression values are not a direct measure of the expression levels of the features but rather a proportional measure following $N_{ij} \propto \mu_{ij}l_i$ where the obtained read count N_{ij} for feature i in sample j is proportional to the true expression level μ_{ij} and the feature length l_i[206]. On the other hand, a major source of between-sample bias arise from differences in produced sequencing amount, i.e. the total number of sequenced reads, as features with similar expression strength tend to have more reads in samples with higher sequencing depth and vice versa which means that the measured expression also depends on the total amount of produced sequence[154][164]. When testing for differential expression of features between different samples, it is crucial to take this kind of bias into account.

One of the first and widely used normalization procedures is to calculate reads per kilobase per million mapped reads (RPKM)[164]. This method considers both within- and between-sample bias to some extent since observed read counts are divided by feature length as well as total sequencing amount. When paired-end sequencing is performed usually an equivalent adjustment named fragments per kilobase per million mapped fragments (FPKM)[247] is applied that uses fragment counts, i.e. both pairs of a read represent one fragment instead of single read counts as measuring unit. Although aiming to correct for both, within- and between-sample bias, it was shown that using RPKM/FPKM values for within-sample comparisons is appropriate but for between-sample comparisons scaling counts by library size is too simple since the obtained read counts not only depend on feature length and library size but also on the composition, i.e. expression pattern, of the sampled RNA population[25][215]. This means that each FPKM value of a sample depends on the expression level of each gene in this sample since FPKM values represent a proportional expression level of a gene to all other genes[206]. For example, if a substantial number of features is highly or even solely expressed in a sample the normalized expression values of the remaining features are likely underestimated (Figure 1.12).

As already mentioned, normalization is especially important when expression profiles between samples are compared. Bullard *et al.*, 2010[25] demonstrated that the choice of the normalization procedure has a major impact on differential expression detection. Thus, the aforementioned RPKM/FPKM adjustment with its deficiencies is not suitable when investigating differential expression. Several normalization methods have been proposed over the last couple of years with the aim to optimally remove between-sample bias while minimizing the introduction of other noise in order to guarantee best possible differential expression results[70][128][207]. Most of these methods are global procedures which means that they calculate a single factor to scale the raw read counts[48]. For example, Bullard *et al.*, 2010[25] presented a method that uses only feature counts of the upper quantile (UQ) to calculate the scaling factor. Another normalization method is the relative log expression (RLE) proposed by Anders and Huber, 2010[7]. In this method, a pseudo-reference sample is initially created through computing the geometric means of feature counts across all given samples. This pseudo-reference sample is then used to create individual scaling factors for each



(a) (b)

Figure 1.12.: This figure illustrates the results of FPKM normalization for two different scenarios. a) FPKM normalization is illustrated for two human samples which have comparable expression patterns but different library sizes with sample 2 (green) deeper sequenced than sample 1 (blue). The characteristics of the unnormalized raw read counts is shown in the left panel and the respective FPKM normalized counts on the right. After normalization both data sets show good agreement (red fitted regression curve). b) The same two samples as in a), but this time with a subset of genes in sample 2 artificially increased in order to mimic a set of highly differentially expressed genes. After FPKM normalization, resulting values of low expressed genes from sample 2 are skewed. Thus, genes which have the same expression level across samples can appear differentially expressed. (Idea for this plot is based on [71])

sample by calculating all feature ratios between a sample and the reference sample and taking the median of those ratios as scaling coefficient (see Figure 3.19 in Chapter 3.1.4). A further procedure, similar to the RLE method, is the trimmed mean of M values (TMM)[215] which estimates global scaling factors between provided samples with the aid of an empirical strategy where features with very high overall or large differences in expression are removed and the weighted mean of log ratios of the remaining features is calculated.

Differential Expression Analysis

Differential expression analysis is used to identify features, often genes, that show differences in expression levels between experimental conditions. Several methods for differential expression analysis of microarray data are available which are based on their nature of continuous intensity values. RNA-seq, on the other hand, results in discrete read count measurements which require other statistical models[177]. Initially, Poisson distribution was used to model RNA-seq read count data and it could be demonstrated that it provides a sufficient fit for the read distribution across technical replicates[25][120][154]. However, the basic assumption of the Poisson distribution that the variance is equal to the mean is not appropriate for the higher variability in biological replicates[117][216]. As a result, this overdispersion makes the analysis prone to high false positive rates[7][177]. To overcome this problem an alternative, namely the Negative Binomial distribution, was introduced first for SAGE[142][217] and later also for RNA-seq data[214]. This approach is applied in widely used differential expression analysis software packages such as edgeR[214], DESeq[7] and its derivative DESeq2[140]. In contrast to the Poisson distribution, the Negative Binomial distribution does not rely on the assumption of equal mean and variance. Instead, they are distinctly and uniquely determined[7]. Nevertheless, it is common that the number of available replicates in experiments is low which makes the estimation of both parameters for each feature unreliable. Therefore, Robinson *et al.*, 2010[214] proposed an alternative solution for their edgeR package where the variance σ^2 is related to the mean by $\sigma^2 = \mu(1 + \mu\phi)$ with μ representing the mean and ϕ a single dispersion constant estimated from the data that is applicable for the entire experiment. For the DESeq package, Anders and Huber, 2010[7] extended the edgeR model by allowing a more data-driven relationship between mean and variance and claim to reach better fits. More precisely, they assume similar dispersion for features with similar expression strength, thus sharing information between multiple features by using local regression to obtain an estimate for the variance. In DESeq2[140] they advanced the DESeq approach by shrinking the dispersion estimates towards a fitted smooth curve for the purpose of minimizing the impact of individual outliers.

A feature is said to be differentially expressed when the difference in observed expression levels is greater than what would be expected by random variation. To test for this between two conditions A and B, the abovementioned differential expression analysis algorithms assume the null hypothesis H_0 that for each feature i the mean values of both conditions μ_{iA} and μ_{iB} are equal, i.e. $H_0 : \mu_{iA} = \mu_{iB}$[7][217][214]. Under this null hypothesis it is then possible to calculate the probability $p(a, b)$ of observing any two counts a and b in both conditions and the resulting p-value for two observed counts c_A and c_B is calculated as the sum of all probabilities $p(a, b)$ less or equal to $p(c_A, c_B)$ divided by the sum of all probabilities $p(a, b)$ where the variables a and b can have values from 0 to $(c_A + c_B)$[7].

Enrichment Analysis

Identifying single features which show altered expression patterns across different experimental conditions is of big importance but investigating differing expression patterns in sets of features might reveal further valuable biological insights[177]. Features are typically grouped into categories on the basis of common biological properties and a widely used technique to do so is based on Gene Ontology (GO)[244] categories while another technique is to group genes based on their affiliation to a common biological pathway[264]. Subsequently, GO and pathway enrichment analysis, respectively, can be performed by testing whether individual categories are overrepresented among differentially expressed features. This can contribute to the detection of weaker signals which would not have been detected on the individual feature level. Software tools like topGO[5], GAGE[145] or GOseq[145] have been developed for this purpose.

1.4.3.2. Fusion Detection

In addition to the quantification and comparison of expression levels, RNA-seq data can be utilized for the detection of gene fusion products, also known as chimeras[148][149]. They arise from chromosomal rearrangements and have been associated with tumor initiation and progression[162]. For the purpose of identifying potential fusion events in RNA-seq data the usual splice-aware alignment step needs to be adapted to the extent that single reads as well as read pairs (when paired-end sequencing was performed) can have larger gaps in between or even align to different chromosomes. Furthermore, the assumption that successive nucleotides of a read align in the same orientation must be discarded to enable the detection of fusion events caused by inversions. These adjustments make the alignment computationally more expensive, but are necessary in order to detect rearrangements. TopHat-Fusion[108], an extension of the splice-aware alignment software TopHat[246], is especially designed to align reads in the aforementioned fashion by first trying to map reads entirely to the genome and subsequently splitting the initially unmapped reads into smaller segments to align them with these relaxed constraints. If the outer segments of a read map to different genes the interjacent segments are used to identify the exact breakpoint location. The detected fusion candidates are then used to align the initially unmapped reads against them. Putative fusion events are identified by means of supporting reads where reads can either span, i.e. portions of the same read map to different genes, or encompass, i.e. each part of a paired-end read map to different genes, a fusion breakpoint.

A big challenge in the identification of true fusion events is the high number of false positives in consequence of mapping problems caused by repetitive sequences, paralogous genes or antisense RNA[64][35]. In order to get rid of false positives several filtering strategies have been proposed to address various sources of errors[108]. An early filtering step, typically performed before the initial identification of putative fusion products, is the exclusion of multi-mapped reads and reads that span a fusion breakpoint just slightly, i.e. only by a specified number of bases that is below a certain threshold. After that, solely fusions with a sufficient amount of supporting reads, both spanning and encompassing, are considered true. A further approach is to filter fusions on the basis of read distribution which was proposed by Edgren *et al.*, 2011[55] who revealed that correct fusions show an uniform read distribution around the fusion junctions while wrong ones do not[108]. Utilizing the expression levels of involved genes is another option whereby potential fusions are filtered if the junction is supported by more reads as expected according to the expression levels[13].

Further filtering can be performed by ensuring that at least one of the fusion partners constitute an annotated gene, when both involved genes are known paralogs or, more generally, are classified as duplicated genes[179].

1.4.3.3. Variant Detection

Identifying variations in the DNA sequence of an individual is one of the major tasks when performing DNA sequencing since it has the potential to reveal mutations possibly causing diseases or other phenotypic traits. When dealing with RNA-seq data, it is also possible to perform variant detection[31][75]. However, RNA-seq data possess some specific properties that make it more complex to reliably call variants. For example, low expression and thus low coverage can result in incorrect variant calls. Furthermore, possible allelic imbalances at heterozygous sites makes it difficult to detect the variant when the mutated allele is weakly expressed[68]. Another potential source of error concerns splicing. When sequencing mRNA the intronic sequence is normally spliced out and the obtained reads align solely in exonic regions. Nevertheless, some reads might not be properly aligned and incorrectly extend into intronic regions. This behavior can lead to numerous false positive variant calls, yet is relatively easy to overcome by restricting variant calls to exonic regions or just discard the overhang.

Although it is more complex to call variants in RNA-seq than in DNA-seq data, there is still a number of benefits[196]. First of all, if both RNA-seq and DNA-seq data are available the detected variants in the former can be used to validate findings from the latter ones. Second, when only RNA-seq data are on hand, variant calling in RNA-seq data does not produce any additional sequencing costs while possibly already providing useful variant information and thirdly, variant calling in RNA-seq data enables the investigation of RNA editing[189][196].

Sophisticated variant caller such as SAMtools[125] or the Genome Analysis Toolkit (GATK) UnifiedGenotyper[158] use Bayesian models in order to calculate genotype probabilities.The GATK HaplotypeCaller[158] performs a local de-novo assembly of present reads in potentially mutated regions which increases accuracy especially for sites where variant calling is difficult. For each potential region the software builds a graph, with each path representing a distinct identified haplotype. With the aid of the raw reads, the HaplotypeCaller finds the most likely haplotype for the investigated region and this haplotype is then used to obtain potential variants, again, by means of a Bayesian model[22]. VarScan [3][111][112], another variant caller, uses SAMtools' mpileup function to process the input BAM file and subsequently, the resulting pileup file is passed to a heuristic algorithm which uses the coverage, base quality and allele frequency information of each position in order to determine the respective genotype.

Filtering variants is a usual step in order to get rid of false positive ones. For variant calls, common filtering criteria concern base quality, coverage, accumulation of vari-

[3] http://dkoboldt.github.io/varscan (last accessed 01.10.2016)

ants in small regions[68] but also homopolymers and repetitive regions[196]. Finally, variants are annotated with additional information, such as affected gene product, effect on amino acid composition or the allele frequency of the variant in the 1000 Genomes Project, to name but a few.

The output of variant calling programs is typically stored in Variant Call Format (VCF)[43] files. Each VCF file has a header and a data section. The header contains meta- and the data section the actual variant information where each line represents one variant. An example of the data section with two variants is shown in the following:

```
chr1    14599 .   T   A   507.7 PASS  .   GT:AD:DP:GQ:PL  0/1:2,12:14:45:536,0,45
chr1    18503 .   G   A   948.7 PASS  .   GT:AD:DP:GQ:PL  0/1:29,39:68:99:104,0,765
```

where the first and second column stores the genomic position of the variant, the third column is reserved for an optional unique identifier for the variant, the fourth and fifth represent the reference and alternative allele, respectively, the sixth column stores the quality score of the variant call, the seventh whether the variant passed the filtering steps or not, the eighth column is reserved for user specific annotations, the ninth column specifies the format of the following columns where any number of columns can be added and each additional column represents one specific sample.

Part II.

Material and Methods

2. Materials and Methods

Before the implementation of the RNA-seq data analysis pipeline started, a whole exome sequencing data analysis pipeline was already in use[54][256] (Figure 2.1). However, this pipeline was designed to handle solely exome sequencing data and the ever increasing number of RNA-seq data necessitated the development of an RNA-seq data analysis pipeline.

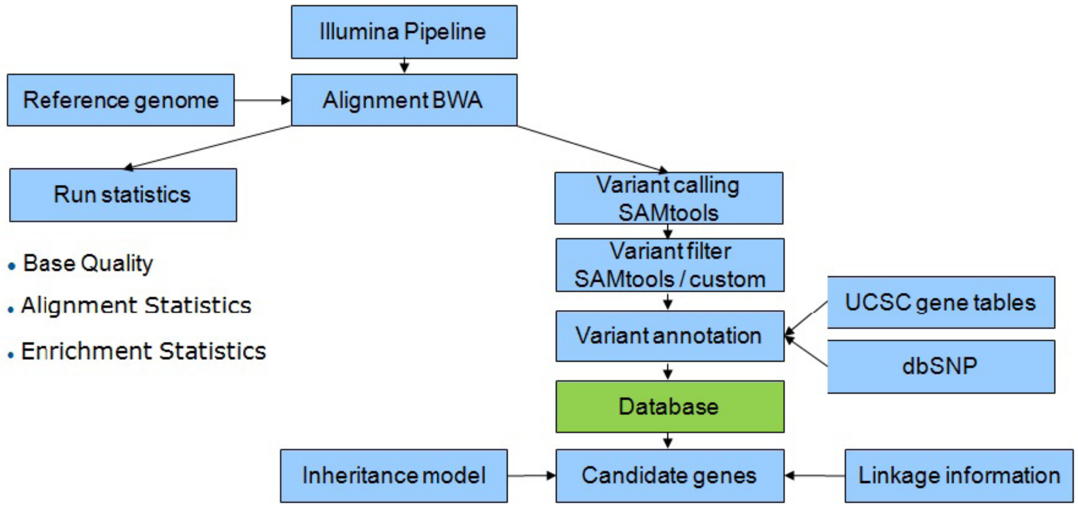


Figure 2.1.: This picture shows an overview of the pre-existing exome sequencing data analysis pipeline as initially developed by Eck, 2014[54] and which was subsequently enhanced by Wieland, 2015[256]. (Figure taken from Eck, 2014[54])

Even so, the pre-existing infrastructure constituted a solid foundation and provided several useful functionalities and information the new pipeline benefited from. For example, a MySQL[1] relational database system was already in use. It comprises a number of databases storing results from exome sequencing data analysis right up to informative meta data for each sample that goes along with an in-house laboratory information management system (LIMS) which provides useful information supporting the automated sample analysis and thus reduces necessary user interactions. Furthermore, the exome sequencing data analysis pipeline was constructed in a way that independent analysis steps can be executed in parallel by using the batch-queuing

[1] `http://www.mysql.com/` (last accessed 24.02.2016)

system Open Grid Scheduler (OGS)[2]. This feature allows to utilize the available resources as efficiently as possible and as a consequence drastically reduces analysis time per sample[256].

It is important to mention that in some cases, the decision which technology, tools, versions or reference provider, e.g. gene annotation file from UCSC, to use was affected by the pre-existing setup. For the other cases, the decision was based on comparisons of competing products which is discussed in more detail in the Results Chapter 3.

2.1. RNA-seq Data

By the time of writing, more than 1,500 RNA-seq samples from 20 different projects and 4 distinct organisms were sequenced on Illumina HiSeq2000 and HiSeq2500 machines, respectively (Figure 2.2). For all of them, sequencing libraries were produced with the Illumina TruSeq RNA Library Preparation Kit v2 and sequencing was performed as 100 bp paired-end runs. After sequencing, every sample was processed with the RNA-seq data analysis pipeline that was implemented in the course of this PhD project. The majority of samples stem from mouse and human with a total number of 760 and 737 samples, respectively. The remaining samples constitute only a minor fraction.

Furthermore, publicly available data were obtained in order to test the performance of individual parts of the pipeline and detailed information about each data set used is provided in the respective chapters.

Depending on the experimental design and biological question to answer, different steps of the pipeline were performed. The most frequent type of analysis was the identification of differentially expressed genes across two experimental conditions. Nevertheless, numerous samples were also subject to other kind of analyses like fusion detection, variant calling or differential exon usage analysis.

2.1.1. Simulated RNA-seq Data

Simulated human RNA-seq read data were generated with the aid of a benchmarking framework named Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEERS)[74]. This framework simulates paired-end reads coming from Illumina sequencing instruments by randomly choosing a specific number of transcripts (by default 30,000) out of a plethora of transcript models from 11 different gene annotation sets. The reason for this is to not bias the simulation towards any single annotation set[74]. The final simulated transcriptome consists of the chosen transcripts and additional alternative splice forms of the initial transcripts where the number of alternative splice forms per transcript can be modified by an input parameter. Based on this transcriptome, pre-defined gene quantifications and intron inclusion

[2] `http://gridscheduler.sourceforge.net/` (last accessed 01.10.2016)

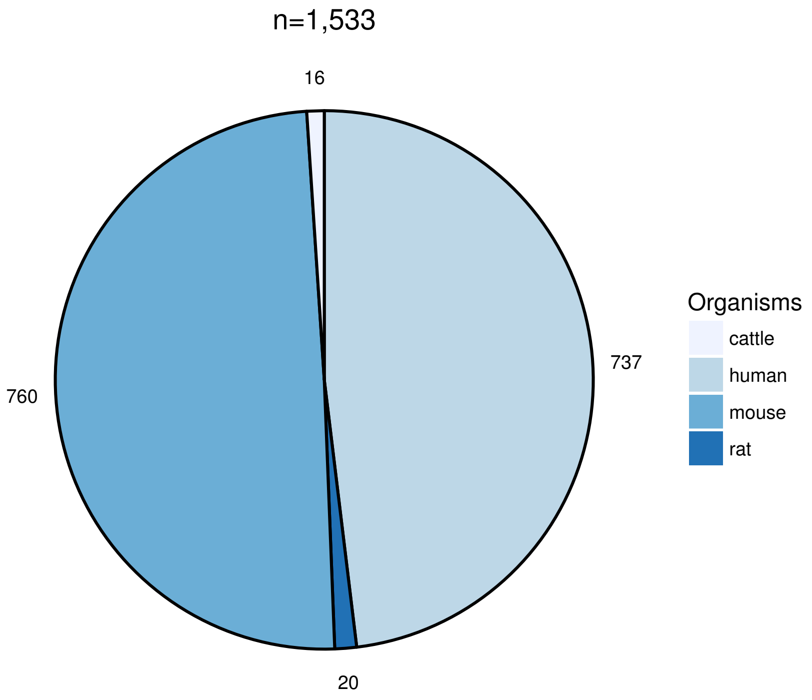


Figure 2.2.: This pie chart depicts the total number of RNA-seq samples that were sequenced and analyzed in the course of this PhD project. In total, 1,533 samples were processed by the analysis pipeline where the majority (about 98%) stem from human or mouse.

probabilities, BEERS chooses RNA fragments of normally distributed length and reports 100 bp paired-end reads by returning the rightmost and leftmost 100 bp of the fragment while introducing polymorphisms (substitutions and indel) to both of them with polymorphism rates again depending on input parameters.

Here, three distinct *in silico* samples were generated. Two (Simulated Read Sample 1 and 2) were simulated using BEERS with standard parameter settings and a third one (Simulated Read Sample 3) with slightly higher polymorphism rates (-error 0.01, -subfreq 0.005, -indelfreq 0.001) and worse read tail qualities (-tpercent 0.001, -tqual 0.9). For each of the three samples 40 million 100 bp paired-end reads were simulated.

2.2. Genome Assembly and Annotation

Both organisms, mouse and human, are well studied and have a mature reference genome as well as gene annotations available. Thus, sequenced reads are not assembled but directly aligned to the reference genome. For that purpose, the GRCh37/hg19 and NCBI37/mm9 assemblies for human and mouse, respectively, were downloaded from the UCSC Genome Browser[104] website and used throughout the entire analysis workflow as needed. Moreover, all further files and results are based on these assemblies.

In order to guide the splice-aware alignment and also for feature counting, gene annotations in the GTF format are used. Concerning this, the pipeline makes use of the UCSC knownGene annotation[87] downloaded with the aid of UCSC's table browser tool[99]. A known issue with GTF annotations downloaded from the UCSC table browser is that transcript based UCSC identifier are used as gene identifier which will flaw gene-wise counting since each transcript will be treated as independent gene. To overcome this problem, UCSC identifier were translated to gene symbols using a custom script and UCSC's kgXref table. The choice of which annotation to use was on the one hand affected by the pre-existing database structure that was built up on the UCSC knownGene annotation and on the other hand based on a comparison of three different annotations (Chapter 3.1.1). The comparison was performed for both organisms, human and mouse and tested annotations were GENCODE (v11 for human; vM1 for mouse), RefSeq Genes (Build 37.2 for human; Build 37.2 for mouse) and UCSC knownGene (Feb. 2009 for human; July 2007 for mouse).

2.3. Data Analysis Pipeline

The semi-automated RNA-seq data analysis pipeline was developed using the highly-capable, feature-rich programming language Perl[3]. In order to fulfill all necessary analysis steps both, publicly available software as well as custom Perl scripts, are combined. All implemented features are shown in Figure 2.3 and will be discussed in more detail in the following sections. Further on, an important issue is the minimization of computing time. The default behavior of the pipeline is to initially check whether specific files for a sample already exist and to automatically perform only those analysis steps whose result files could not be found. However, the user can change this behavior if desired by simply passing the `--co` parameter which forces the pipeline to start from scratch.

2.3.1. Split Read Alignment

So far, the splice-aware aligner GEM v1.7.1[151] is used as the default alignment tool in the pipeline. The main reason for this is that the pipeline was developed

[3] `https://www.perl.org` (last accessed 01.10.2016)

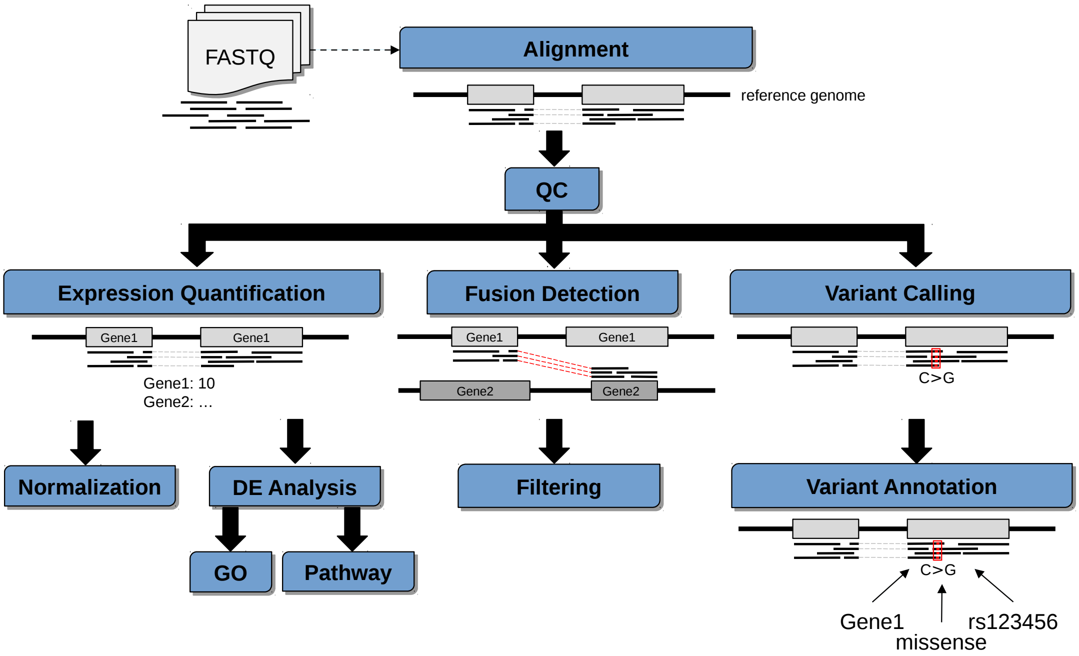


Figure 2.3.: Overview of the steps implemented in the RNA-seq data analysis pipeline.

in the course of the *RNA sequencing project of 1000 Genomes samples*[4] of the Genetic European Variation in Health and Disease (GEUVADIS) Consortium[5] whose results were published in Lappalainen *et al.*, 2013[119]. The GEM alignment tool was the aligner of choice for this project and thus it was implemented in the pipeline. The alignment with GEM is performed with standard parameter settings except for two, i.e. `--mismatches=0.04` and `--min-decoded-strata=2`. Other than in Lappalainen *et al.*, 2013[119] no read trimming is performed. Currently, two additional and widely used aligner, namely STAR v2.3.0[50] and TopHat v2.1.0[107], are incorporated in the pipeline and can be easily used instead of GEM by just changing the default aligner argument `--al gem` to `--al star` and `--al tophat`, respectively. Depending on the sample's organism, all three aligner use the respective reference genome together with the UCSC knownGene annotation as input parameter. Moreover, the sample's FASTQ files which store the actual sequenced read information are provided to the aligner as well. Often sample libraries are pooled together and the consequent pool is sequenced on multiple lanes of the flowcell. As a consequence, multiple FASTQ files, or more specifically one FASTQ file per lane, are generated for each sample. If this is the case, alignment is performed on each FASTQ file sepa-

[4] `http://www.geuvadis.org/web/geuvadis/RNAseq-project` (last accessed 01.10.2016)
[5] `http://www.geuvadis.org` (last accessed 01.10.2016)

rately or all together, depending on the aligner since not all of them are able to handle multiple FASTQ files. The information how many FASTQ files per sample exist is all stored in the LIMS. Furthermore, independent of what aligner chosen and owing to the sample meta information obtained from the in-house database, the pipeline automatically detects whether single-end or paired-end sequencing was performed and starts the alignment process accordingly. The alignment is stored in BAM format and if separate alignment of multiple FASTQ files generated multiple BAM files per sample the BAM files are subsequently merged with the result that a single BAM file per sample is available for all following analysis steps. Finally, by default the pipeline creates another BAM file where duplicated reads are removed by the Picard toolkit v1.139[23].

2.3.2. Quality Control

Quality control is performed both, pre- and post-alignment. The former is not carried out automatically but rather accomplished by using Illumina's Sequencing Analysis Viewer (SAV)[6] and manual inspection of the yield and quality of the sequencing runs. If no noticeable problems are observed the sequencing run is declared valid for the moment and downstream analysis will be performed. In the case of base quality issues especially at the beginning and the end of the reads quality trimming can be performed before alignment.

After the alignment a multitude of QC metrics are computed on a per sample basis and stored in the in-house database. For that purpose, publicly available tools, namely RNA-SeQC v1.1.8.1[47], SAMtools v0.1.19[125] and the Picard toolkit v1.139[23], as well as custom scripts are utilized. All of them run for each sample separately and generate measures regarding sequencing and alignment yield, read distribution, coverage and read properties, among others. Furthermore, the approach that enhances the likelihood of detecting possible sample mix-ups or between sample contamination differs slightly between the exome sequencing and RNA-seq data analysis pipeline. The former one calculates the coverage of the male-specific SRY gene and checks whether it is consistent with the actual sex of the sample[256]. In contrast, the RNA-seq data analysis pipeline checks the female-specific XIST gene expression together with the Y-chromosome gene expression as was done in the GEUVADIS project[119][239].

All generated QC metrics are stored in files. At the end of the QC step all of these files are parsed and the metrics are stored in the in-house database.

2.3.3. Expression Quantification

Expression levels of genes are quantified using the htseq-count module of the HTSeq v0.6.0 package[9]. The software requires aligned reads in SAM/BAM format as well

[6] `https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html` (last accessed 01.10.2016)

as gene annotation in GFF/GTF format in order to count reads per gene. The proper GTF file together with specific parameters are determined based on the characteristics of the RNA-seq sample to analyze which are obtained as meta information from the database. However, several parameters like `--type=exon`, `--idattr=gene_id` and `--mode=intersection-nonempty` are invariable for all samples. If desired, reads can be counted for introns instead of exons as well. For this purpose, htseq-count is executed with `--type=intron`. Nevertheless, with the common GTF file from UCSC this would not be possible since by default it does not contain intronic regions. Therefore, intronic regions were added to the mouse and human GTF files using a custom script. Once finished, htseq-count produces an output file with a count value per gene.

In some cases researchers might be interested in differences in exon usage of particular genes. With this in mind, the pipeline allows for counting reads per exons in addition to accumulating reads of entire genes. For the exon counting step, the pipeline utilizes two Python scripts that come along with the differential exon usage analysis tool DEXSeq[10]. The first script, *dexseq_prepare_annotation.py*, had to be executed only once at the first run and flattened the gene annotation file in a way that it takes a GTF file as input and creates a new file where each exon constitutes an independent feature. If exon boundaries differ between isoforms, the script creates two distinct exon parts which are treated independently in the downstream analysis (Figure 2.4). However, the script was not able to cope with GTF files downloaded from UCSC, thus gene annotation files for mouse and human were downloaded from ENSEMBL[7] and used throughout the entire differential exon usage analysis. The second Python script, *dexseq_count.py*, is a wrapper script for htseq-count and performs the actual counting. Again, an output file with counts is generated but in this case counts are reported for each exon. Finally, resulting counts are inserted in the database.

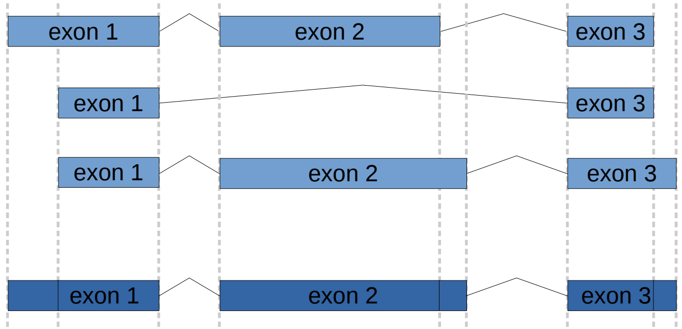


Figure 2.4.: Schematic representation of the gene model flattening as done by the *dexseq_prepare_annotation.py* script. The artifical gene has three transcripts with different exon boundaries. In this case, the script creates six couting bins (dark grey boxes at the bottom). (Figure adapted from [10])

[7] `http://www.ensembl.org/info/data/ftp/index.html` (last accessed 01.10.2016)

2.3.3.1. Normalization

For now, FPKM measures are the only normalized values that are generated by the pipeline. A custom script was implemented for this step and the FPKM value for sample i and gene j is calculated as

$$FPKM_{ij} = \frac{F_{ij} * 10^9}{N_i * L_j} \tag{2.1}$$

where F_{ij} is the raw fragment count of the gene, L_j is the length of the gene in bp and N_i is the total number of fragments produced for the sample. Once calculated, FPKM values are stored in the database along with the raw counts.

2.3.3.2. Differential Expression Analysis

Before differential gene expression analysis can be performed the user has to specify which groups of samples to compare against each other. Once the groups are defined and passed to the pipeline the analysis continues. Subsequently, the required count files are collected and serve as starting point for the downstream differential expression analysis. The language and environment for statistical computing R (currently v3.2.1)[201] is used to fulfill this task and to produce informative plots. Initially, a comprehensive R script was built around the R/Bioconductor package DESeq v1.22.1[7] which not only executes DESeq but also performs all necessary preparatory steps, generates a variety of helpful plots and prepares the results for later inspection. Later, a newer version of DESeq, namely DESeq2 v1.10.1[140], was available and another R script using DESeq2 was implemented. Additionally, a third R script for differential expression analysis utilizing the R/Bioconductor package edgeR v3.12.0[214] was developed mainly for comparison purposes (Chapter 3.1.4.3). All three scripts are integrated into the pipeline and can be used for differential expression analysis.

No matter what script is chosen for this step the results of the differential gene expression analysis are stored in the database but also on the file system in tabular text files.

Beyond testing for differential gene expression the pipeline allows testing for differential exon usage as well. For this purpose, again, another R script was implemented that uses the R/Bioconductor package DEXSeq v1.16.8[10]. This script takes as input the exon-wise counts as mentioned in Chapter 2.3.3 and just like the differential gene expression analysis scripts performs all necessary steps from preprocessing the input data to postprocessing of the results.

2.3.3.3. Gene-Set Enrichment Analysis

After differential gene expression analysis the pipeline automatically performs gene-set enrichment analysis in two different ways. First, it uses the R/Bioconductor package *goseq* v1.22.0[264] to load publicly available Gene Ontologies (GOs)[20] and subsequently test whether any of them are enriched among previously identified significantly differentially expressed genes. Second, pathway enrichment analysis is conducted with the aid of the R/Bioconductor package *gage* v2.20.0[145]. At the beginning of this step, Kyoto Encyclopedia of Genes and Genomes (KEGG)[174] pathways as well as the results from the differential expression analysis are loaded into R and after that *gage* checks for overrepresented pathways. The results are then visualized with the aid of another R/Bioconductor package, namely *pathview* v1.10.1[144].

Additionally, independent of the organism, gene sets are commonly provided with Entrez Gene Identifier while the pipeline works with gene symbols by default. For that reason, first of all gene symbols of the differential expression analysis must be translated into Entrez Gene IDs which is done with the aid of R and the R/Bioconductor package *mygene* v1.6.0[155].

So far, GO and pathway enrichment results as well as pathway plots are solely stored on the file system.

2.3.4. Fusion Detection

In order to minimize computational costs, fusion detection is performed optionally rather than by default for each sample. Thus, the user has to select the fusion detection mode at the start of the pipeline with TopHat-Fusion v2.1.0[108] being the method of choice. This tool requires reads to be aligned with TopHat and parameter settings are employed as suggested in TopHat-Fusion's online tutorial[8]. Following read mapping, TopHat-Fusion calls putative chimeras and after applying a number of integrated filters (Chapter 1.4.3.2) it generates a file containing the detected fusion breakpoints including annotated gene names, if applicable, and several reliability scores in order to assess each fusion.

As for enrichment analysis, detected fusions are not stored in the database yet but rather provided to users in tabular files.

2.3.5. Variant Detection

Variant detection is also an optional feature of the pipeline. GATK v3.5[158] is used for this purpose and the analysis steps were implemented according to the GATK Best Practices workflow for single-nucleotide polymorphism (SNP) and indel calling on RNA-seq data[68] (Figure 2.5) and parameter settings are used as suggested unless indicated otherwise. Briefly, reads are aligned using STAR aligner and duplicate reads are marked using the Picard toolkit. After that, reads extending into intronic regions are clipped with GATK's integrated *SplitNCigarReads* tool since these likely misaligned fragments might otherwise increase the number of false positive

[8] `http://ccb.jhu.edu/software/tophat/fusion_tutorial.shtml` (last accessed 01.10.2016)

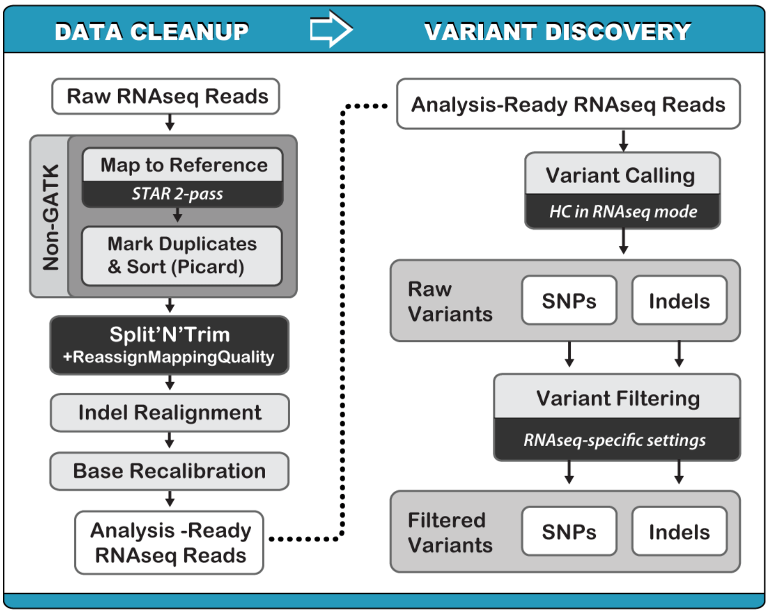


Figure 2.5.: Overview of the variant detection step as suggested by GATK. (Figure adapted from [68])

variant calls. Actual variant calling is then performed with the aid of the GATK HaplotypeCaller and resulting variant calls are filtered with GATK's *VariantFiltration* tool. Finally, detected variants are annotated and inserted in the database using already existing scripts from the exome sequencing data analysis pipeline.

2.4. Database

As already mentioned, the pre-existing exome sequencing data analysis pipeline uses a MySQL relational database system. For practical reasons, the already existing MySQL setup was utilized yet expanded by a table for RNA-seq quality metrics as well as two new relational databases for RNA-seq data analysis results, one for mouse and one for human. A detailed discussion regarding the new tables can be found in the Results section in Chapter 3.3.

2.5. Web Application

Just like the MySQL database system, a web application was already available that allows collaborators to access their analyzed data via the internet[54][256]. It runs on an Apache HTTP Server[9] and uses the Perl Common Gateway Interface (CGI) module.

[9] `https://httpd.apache.org` (last accessed 01.10.2016)

To guarantee data safety, the web application comprises a comprehensive user management along with various security features. Once successfully logged in, the users have access to multiple components such as a general overview of samples associated with their projects right up to several sophisticated ways to browse, investigate and filter their analysis results. Again, a detailed explanation along with screenshots of the web application can be found in the Results section in Chapter 3.4.

Part III.

Results

3. Results

The aim of this PhD project was the design and development of an automated RNA-seq data analysis pipeline along with the investigation and determination of key characteristics and requirements of the input data material. The pipeline provides multifaceted analysis of RNA-seq data and convenient ways to investigate and browse sample information and results. A wide range of competing tools are available for most of the analysis steps involved in the RNA-seq data analysis workflow . Hence, the pipeline is composed of publicly available software as well as custom scripts. The performance of the implemented tools and the assessment of important criteria are discussed in this chapter.

3.1. RNA-seq Data Analysis Pipeline

The pipeline includes several analysis steps and the structure is shown in Chapter 2.3. Prior to the implementation, proper gene annotations as well as analysis tools for each step were selected in order to produce reliable results. For this purpose, competing annotations and tools were compared to each other and the outcomes of these comparisons are explained in the following. Furthermore, parameter settings and filter criteria for selected tools and analysis steps were investigated and implemented if necessary.

3.1.1. Gene Annotation

At the beginning, publicly available gene annotations were compared to each other. This was done to decide which one of them to use for the pipeline as this choice can already have a considerable impact on the downstream analysis results[237]. Since the majority of analyzed samples are from human and mouse the comparison and presented results are restricted to these two organisms. For each of both organisms three common annotations were tested, namely GENCODE, RefSeq and UCSC and annotations were investigated regarding

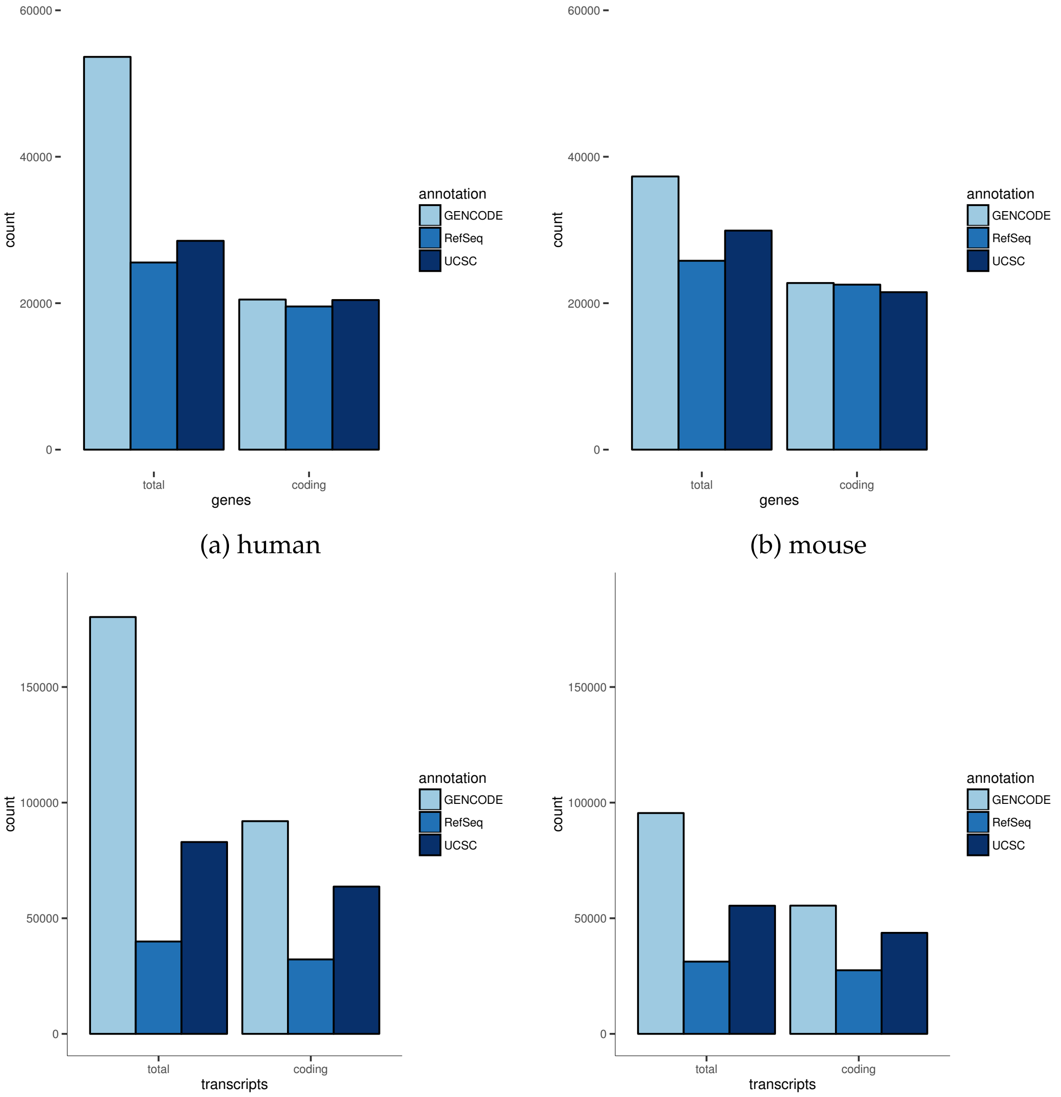
i. number of genes and transcripts

ii. covered region

iii. effect on alignment and expression quantification.

3.1.1.1. Number of Genes and Transcripts

All three annotation sets differ in terms of complexity in both, human and mouse (Figure 3.1). GENCODE is the most comprehensive set and comprises a considerable



(a) human

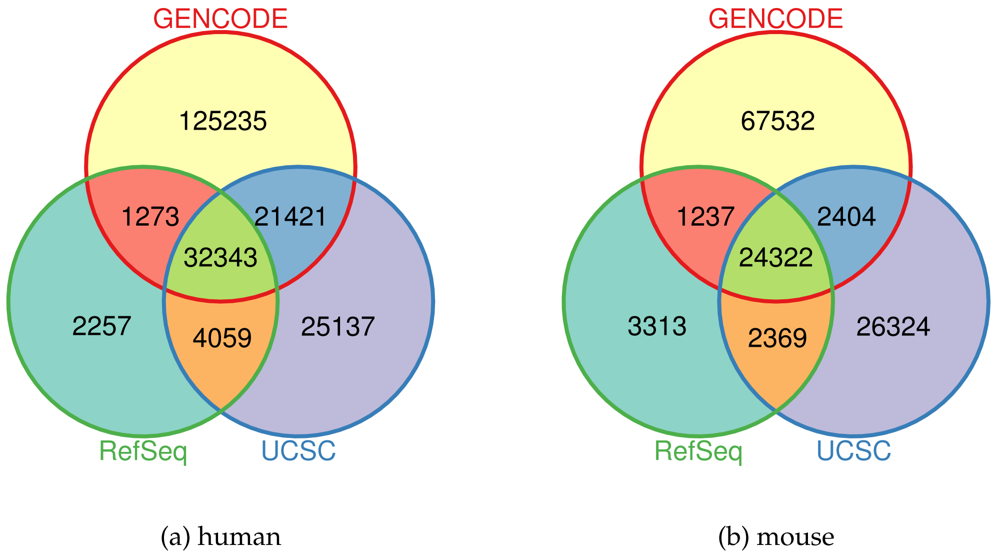
(b) mouse

(c) human

(d) mouse

Figure 3.1.: Barplots showing the total number of genes as well as the number of coding genes in a) human and b) mouse for the three gene annotations, GENCODE, RefSeq and UCSC. c) and d) depict the respective transcript numbers.

amount of non-coding genes. In comparison, RefSeq and UCSC include less annotated genes in total with RefSeq having the smallest proportion of non-coding genes. All three annotations feature a comparable number of coding genes in human as well as mouse. On the transcript level, however, the difference between annotation sets is higher even for the coding part (Figures 3.1c and 3.1d) and the average number of total transcripts per gene for GENCODE, RefSeq and UCSC is 3.7, 1.6 and 2.9 in human and 2.6, 1.2 and 1.9 in mouse, respectively. A comparison of how many transcripts are identical across annotation sets shows that 32,343 and 24,322 transcripts in human and mouse, respectively, are similar in all three sets (Figure 3.2). RefSeq, the least comprehensive annotation set, has the smallest amount of unique transcripts while GENCODE includes the most of all three sets. A comparatively small number of transcripts, namely 53,764 and 26,726 in human and mouse, respectively, are identical across GENCODE and UCSC. One reason for the rather small number of overlaps are the partially minor differences in exon start and end coordinates of basically similar transcripts and in many cases the disagreement involves just one exon of compared transcripts.



(a) human

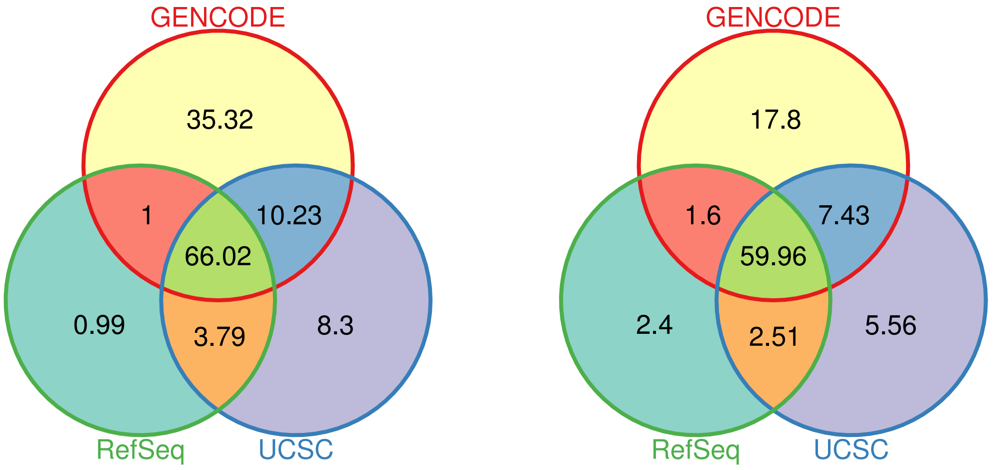
(b) mouse

Figure 3.2.: These venn diagrams show the number of identical as well as unique transcripts in the three investigated gene annotation sets GENCODE, RefSeq and UCSC in a) human and b) mouse.

3.1.1.2. Covered Region

In order to check how well the annotations agree, an additional comparison regarding the covered regions of the genome was conducted. Here, matches are reported in terms of number of overlapping bases between any exons of annotation sets.

As expected, the most comprehensive annotation, GENCODE, covers the biggest portion of the genome (Figure 3.3) with a total of about 113 megabases (Mb) and 87 Mb for human and mouse, respectively, followed by UCSC with 88 Mb and 75 Mb and finally RefSeq with 72 Mb and 66 Mb. In both organisms, overlaps between covered regions are high and constitute the highest portion in each set. The amount of uniquely covered regions is largest for GENCODE in both sets which could be expected with the substantially higher total number of annotated transcripts in these sets compared to the others in mind. Considering coding genes only, all annotations show a high overlap with around 35 Mb of covered region in both, human and mouse.



(a) human

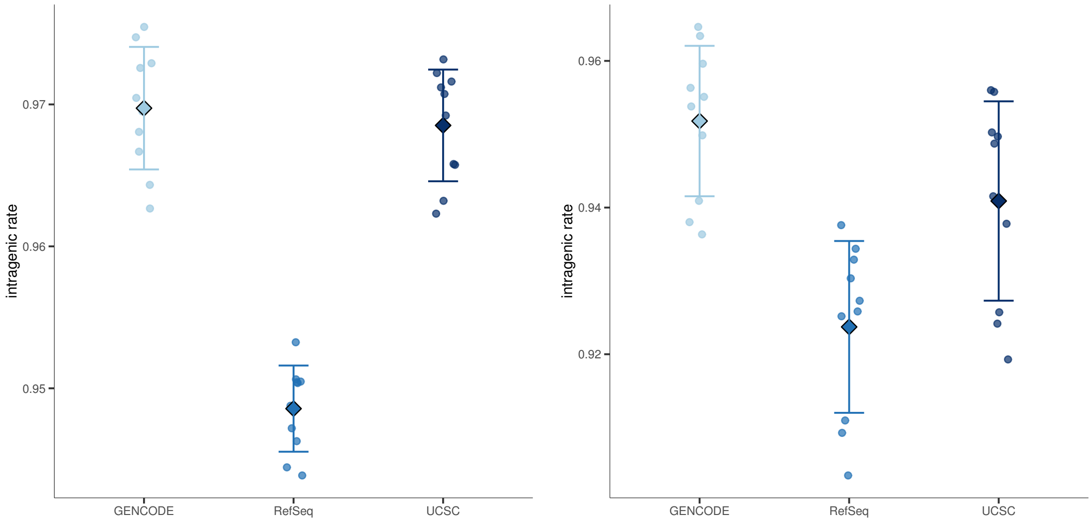
(b) mouse

Figure 3.3.: Venn diagrams depicting the overlaps of covered regions for GENCODE, RefSeq and UCSC annotations in megabases (Mb). Numbers for human annotations are shown in a) and for mouse annotations in b).

3.1.1.3. Effect on Alignment and Expression Quantification

In order to test the impact of each annotation on the alignment, 10 randomly selected, in-house sequenced human as well as mouse poly(A) RNA-seq libraries were aligned to the respective reference genome using the STAR aligner and one of the three annotations at a time. The overall mapping rates were consistently above 97% which is likely due to the fact that STAR initially tries to align the reads to the genome and do not rely as strong as other aligners on a provided reference annotation (Chapter 1.4.1). Nevertheless, considering the intragenic rate, i.e. the fraction of reads mapping within annotated genes, a difference between annotation sets can be observed

(Figure 3.4). For both organisms, human and mouse, the RefSeq annotations reach slightly smaller intragenic rates while GENCODE and UCSC show high agreement. The small difference between these two annotations is not as big as one might expect from the yet considerable higher number of covered bases of the GENCODE annotation compared to the UCSC annotation, suggesting that a lot of transcripts solely covered by GENCODE are rarely present in poly(A) RNA-seq libraries.



(a) human

(b) mouse

Figure 3.4.: Intragenic mapping rates (y-axis) using the same ten a) human and b) mouse samples, respectively. Rhombuses represent the mean intragenic rates and the error bars the standard deviations for each group.

Furthermore, when performing expression quantification using the three annotation sets and comparing the results for each of the 10 samples a high concordance between GENCODE and UCSC (average Spearman correlation coefficient of 0.997) can be observed with genes unique to the GENCODE annotation frequently showing expression values of zero (average of medians of expression values of genes unique to the GENCODE annotation is 0), again suggesting that the UCSC annotation is suitable for the purposes of the pipeline.

3.1.2. Split Read Alignment

Alignment of sequencing reads to the correct genomic position is a fundamental step when analyzing RNA-seq data since all subsequent analysis steps strongly depend on its accuracy. Thus, it is crucial for the overall performance of the pipeline to choose the right alignment tool. The performances of three frequently used split read aligner, namely GEM v1.7.1, STAR v2.3.0 and TopHat v2.1.0, were compared to each other and henceforth referred to as GEM, STAR and TopHat. For comparison, BWA-MEM v0.7.5a-r405[123], an aligner designed for the alignment of genomic sequences, was included as well.

Again, 10 randomly chosen, in-house and 100 bp paired-end sequenced human poly(A) RNA-seq libraries were used and aligned with each of the four aligner. The average amount of sequence of these 10 samples was 7.4 Gb (sd ±0.92 Gb). In addition, three *in silico* samples (see Chapter 2.1.1) were used to ascertain the alignment accuracy of each alignment tool. All four aligners were run with standard parameters except `--mismatches=0.04` and `--min-decoded-strata=2` for GEM and `--twopassMode Basic` for STAR. To validate the performance of the four tools various metrics like

i. runtime

ii. alignment yield and unambiguity

iii. extent of mismatch and indel introduction

iv. extent of soft clipping

v. exonic mapping rate

vi. effect on expression quantification

vii. alignment accuracy

were calculated and are discussed in the following. Some of these metrics were utilized in related form in other studies as well[50][58][74][246].

3.1.2.1. Runtime

In a first step, the runtime of each program was measured (Figure 3.5). The alignment was performed for all 10 samples on an Intel® Xeon® CPU E5-2697 v3 @ 2.60GHz machine allowing only one thread per execution. In terms of runtime, the four alignment tools show major differences with STAR performing best with roughly three CPU hours on average. BWA-MEM and TopHat required about ten times longer to finish while GEM took by far the longest.

3.1.2.2. Alignment Yield and Unambiguity

Here, the overall alignment rate is defined as $1 - \alpha$, where α is the fraction of read pairs that could not be aligned to a single position in the genome. On average, GEM could align the highest proportion (98%) of the sequenced paired-end reads followed by STAR (97%), BWA-MEM (93%) and finally TopHat (84%) (Figure 3.6). Regarding TopHat, another 6% of total read pairs were present in the reported alignments where at least one of the two read pairs could be aligned, resulting in about 10% of read pairs

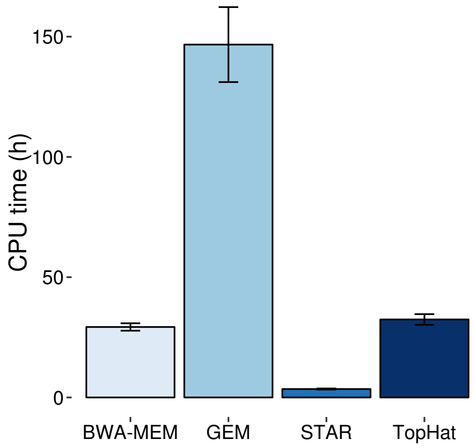


Figure 3.5.: Average per sample alignment runtimes in hours for each of the four aligners based on the alignment of 10 samples. Error bars indicate the standard deviation.

that could not be mapped at all by TopHat. GEM alignments have a high number of fragments (18%) where the position of origin in the genome could not be unambiguously revealed, i.e. multimapped reads. STAR returned the highest fraction (90%) of uniquely aligned read pairs while TopHat and GEM performed comparably in this category with 82% and 80%, respectively. As expected, BWA-MEM did not perform as well as the other three tools when dealing with RNA-seq reads.

It is important to mention that due to the high variability in ambiguous alignments of the different alignment tools only primary alignments (i.e. reads that do not have the "*not primary alignment*" flag set in the BAM file) were used for further metrics calculations in this section.

3.1.2.3. Extent of Mismatch and Indel Introduction

Reliable alignment software should be able to cope with mismatches and indels caused by sequencing errors or true differences between the sample and the reference genome. The quality of the resulting alignment is influenced by the extent of mismatch and indel introduction. This means that true differences should be accounted for but the introduction should not be exaggerated just to align reads at any cost as this might lead to an increased number of misplaced reads. As can be seen in Figure 3.7a, GEM introduced at least one mismatch in nearly 50% of aligned reads. A considerable fraction (about 3%) of reads aligned by GEM have five or more mismatches. On the other hand, BWA-MEM and TopHat reported about 79% and STAR even more than 83% of read alignments with zero mismatches. A similar performance can be observed when inspecting the length and respective rate of introduced indels, with BWA-MEM, STAR

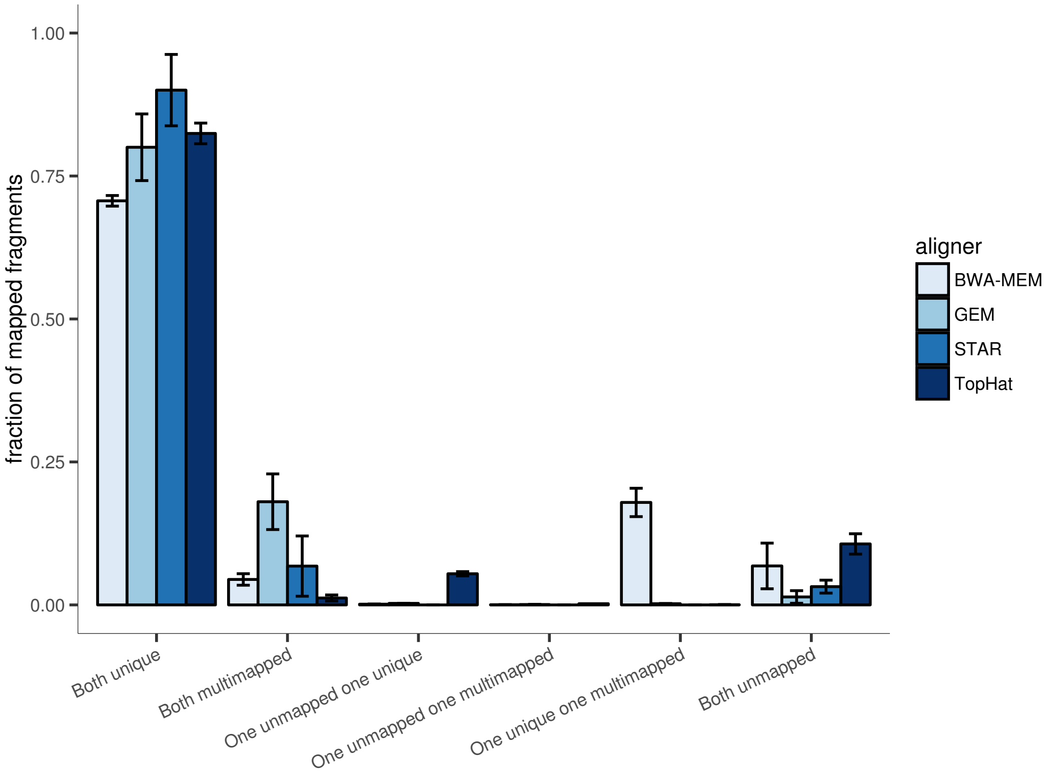
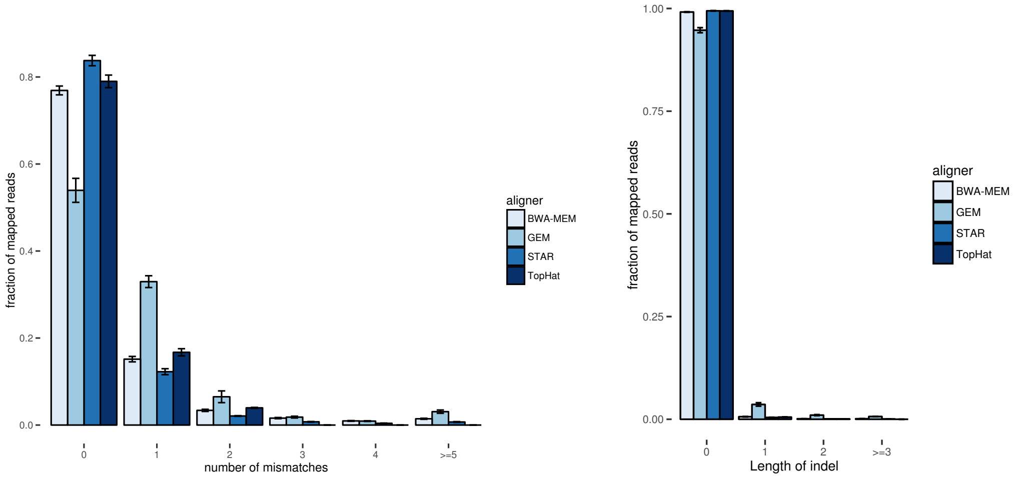


Figure 3.6.: In this plot the alignment performance of each of the four tested aligners is shown in terms of how many read pairs could be uniquely, ambiguously or not at all mapped, how many read pairs with one read uniquely and the other one multimapped and furthermore, how many read pairs show the alignment pattern where one read cannot be mapped while the other one can be uniquely and ambiguously mapped, respectively. Error bars indicate the standard deviation. (Idea for the metrics used in this plot is based on [58])

and TopHat consistently returning more than 99% of aligned reads without any indels while GEM introduced on average at least 1 bp indels in more than 5% of aligned reads (Figure 3.7b).

3.1.2.4. Extent of Soft Clipping

Soft clipping means that a few bases at the beginning and/or the end of the read are ignored by the aligner if they do not match the reference sequence. This can increase alignment sensitivity but also decrease alignment specificity when done too extensively since the more bases are ignored the likelier it becomes that a read is



(a) Single nucleotide mismatches

(b) Indels

Figure 3.7.: a) Average fraction of aligned reads of the 10 tested samples with zero, one, two, three, four and five or more mismatches grouped by aligner. b) Average fraction of aligned reads containing indels of length zero, one, two and three or more bp. Error bars indicate the standard deviation. (Idea for the metrics used in this plot is based on [58])

mapped to a wrong position. On the other hand, soft clipping is useful, for example, when reads partially include adapter sequences. If soft clipping is not supported and the aligner has to map a read with a few unreliable bases at the end it can either introduce several mismatches, align the read somewhere else if possible or leave it unmapped. Figure 3.8 shows the soft clipping behavior of the four alignment tools. STAR truncates at least one base from about 9% and even 10 or more bases from 1.7% of aligned reads while GEM acts rather moderate. TopHat does not perform soft clipping at all by default. Most extensive clipping is performed by BWA-MEM whereas the majority of soft clipped reads in this case result from reads where the clipped parts could have actually been mapped to an adjacent exon but were clipped by BWA-MEM.

3.1.2.5. Exonic Mapping Rate

Another important metric for RNA-seq data derived from mRNA is the fraction of reads that can be mapped to exonic regions in the genome. For the 10 samples involved in this comparison all three RNA-seq split read aligner achieve comparable results (Figure 3.9). STAR performs best with an average exonic rate of about 85% followed by TopHat with 82%, GEM with 79% and BWA-MEM with 61%. The lower exonic rate achieved by BWA-MEM can be mainly explained by a substantially higher intronic rate with, on average, about twice as many reads mapping to intronic regions

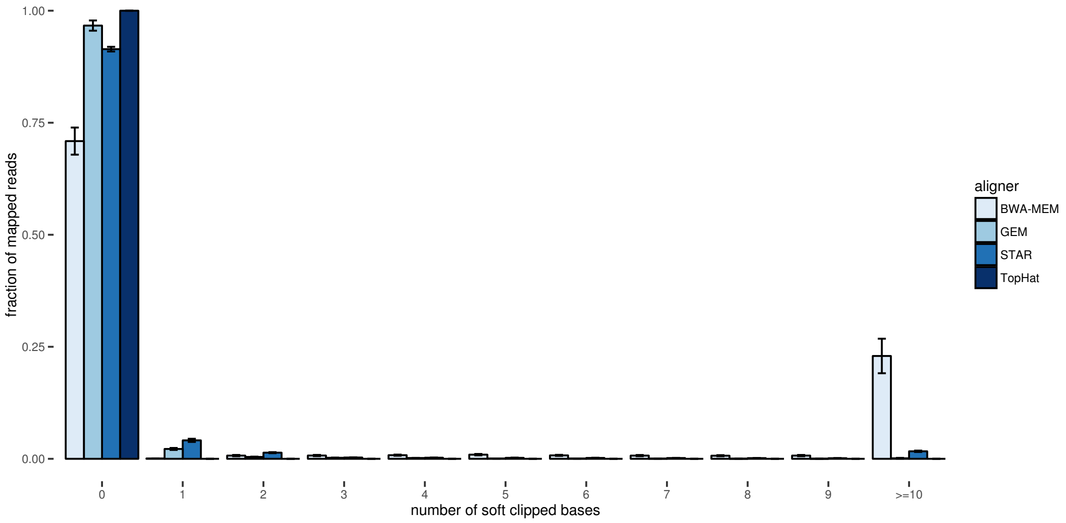


Figure 3.8.: Average proportion of aligned reads grouped by aligner and number of clipped bases with groups of clipped bases ranging from zero to 10 or more. Error bars indicate the standard deviation. (Idea for the metrics used in this plot is based on [58])

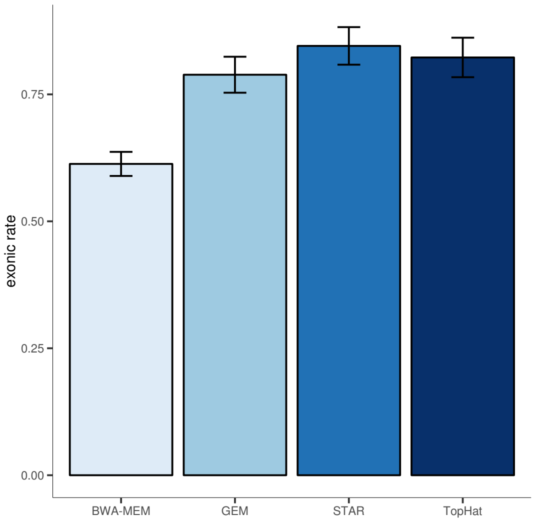


Figure 3.9.: Average achieved exonic rate per aligner. Error bars indicate the standard deviation.

compared to the other three tools.

3.1.2.6. Effect on Expression Quantification

Expression quantification is an important feature of the pipeline which heavily depends on the preceding alignment. To check whether the choice of the aligner has an impact on gene expression quantification, gene expression values were calculated for the 10 samples using htseq-count and then compared to each other. Results are similar across the 10 samples and Figure 3.10 shows the correlation of read counts between the four alignment tools for one of the 10 samples. Overall, all four methods yield rather similar results. BWA-MEM shows the least concordance with the other three methods (Spearman correlation coefficient of 0.986, 0.983 and 0.978 with GEM, STAR and TopHat, respectively) while STAR and TopHat show the highest correlation (Spearman correlation coefficient of 0.997).

3.1.2.7. Alignment Accuracy

Finally, the mapping accuracy of the four alignment tools was tested with the aid of three simulated 100 bp RNA-seq samples (Chapter 2.1.1). For this purpose, four coefficients were calculated, i.e. the fraction of reads where every single base could be mapped correctly (match), the fraction of reads which do not align completely correct but overlaps the correct mapping position by at least one base (partly match), the fraction of reads that map to a different than the true position (no match) and the fraction of reads that could not be aligned at all (not aligned). The performances of the four aligner on each of the three simulated samples are shown in Figure 3.11. In terms of perfectly placed reads, all three split read aligners perform equally well on the standard *in silico* datasets (Simulated Read Sample 1 and 2) with consistently more than 93% of reads falling in this category. For the sample with higher polymorphism rates (Simulated Read Sample 3) the observed differences are more conspicuous. While GEM could align the highest amount (92%) of reads to the correct position STAR and TopHat lag behind and achieve 88% and 80%, respectively. For expression quantifications it is vital that reads can be assigned to the correct gene. Thus, taking together reads falling into both categories, i.e. match and partly match, GEM and STAR yield comparable numbers for each of the three samples. Furthermore, investigation of the distribution of the number of overlapping bases for partly matched alignments revealed that most of them show high agreement with the true mapping position with an average of 93%, 94% and 93% overlapping bases per partly matched read for STAR, GEM and TopHat, respectively (Figure 3.12). BWA-MEM, in contrast, yields by far the lowest fraction of perfectly placed reads but on the other hand returns a high number of partly matched ones. However, on average only 64 bases of them overlap the correct position. TopHat, again, shows the highest fraction of reads that could not be aligned. And eventually, STAR reports least misplaced reads.

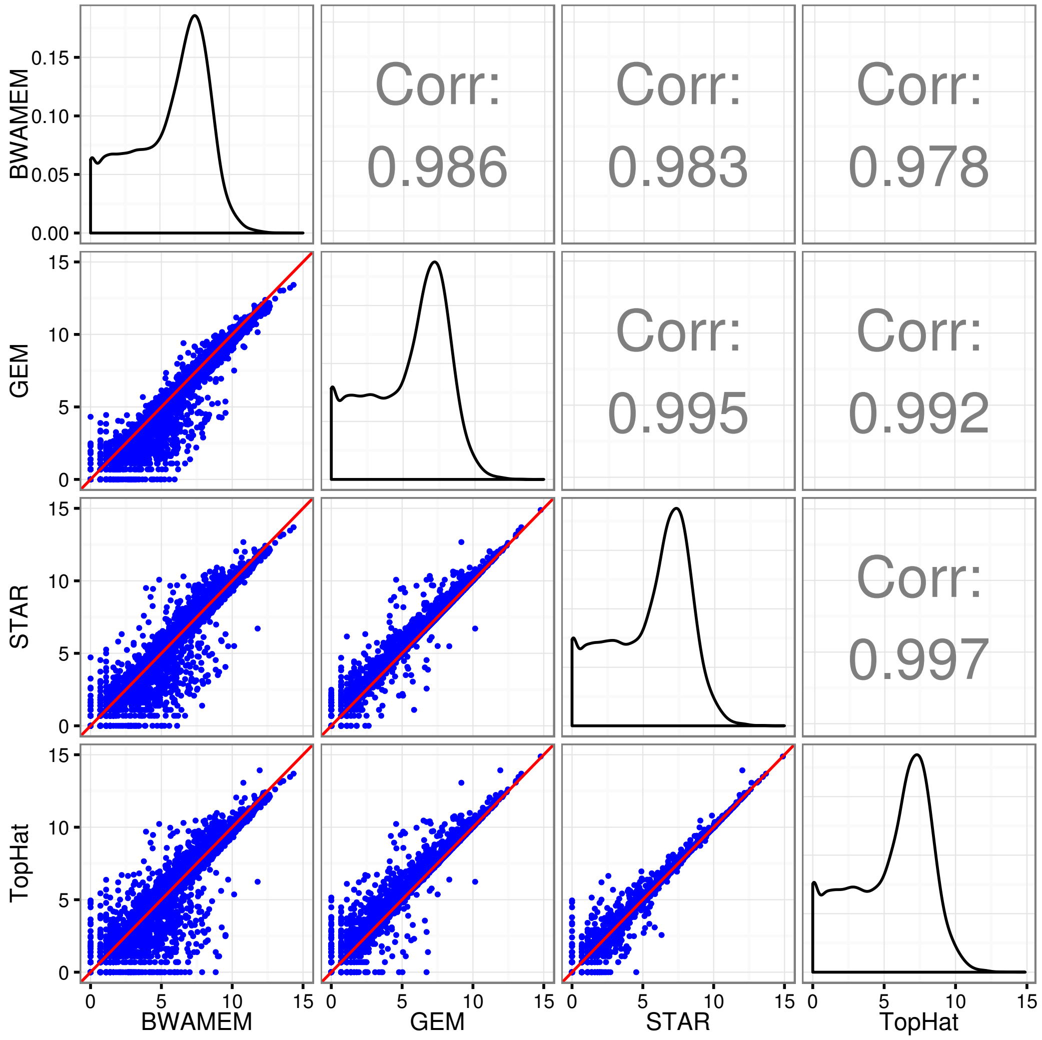
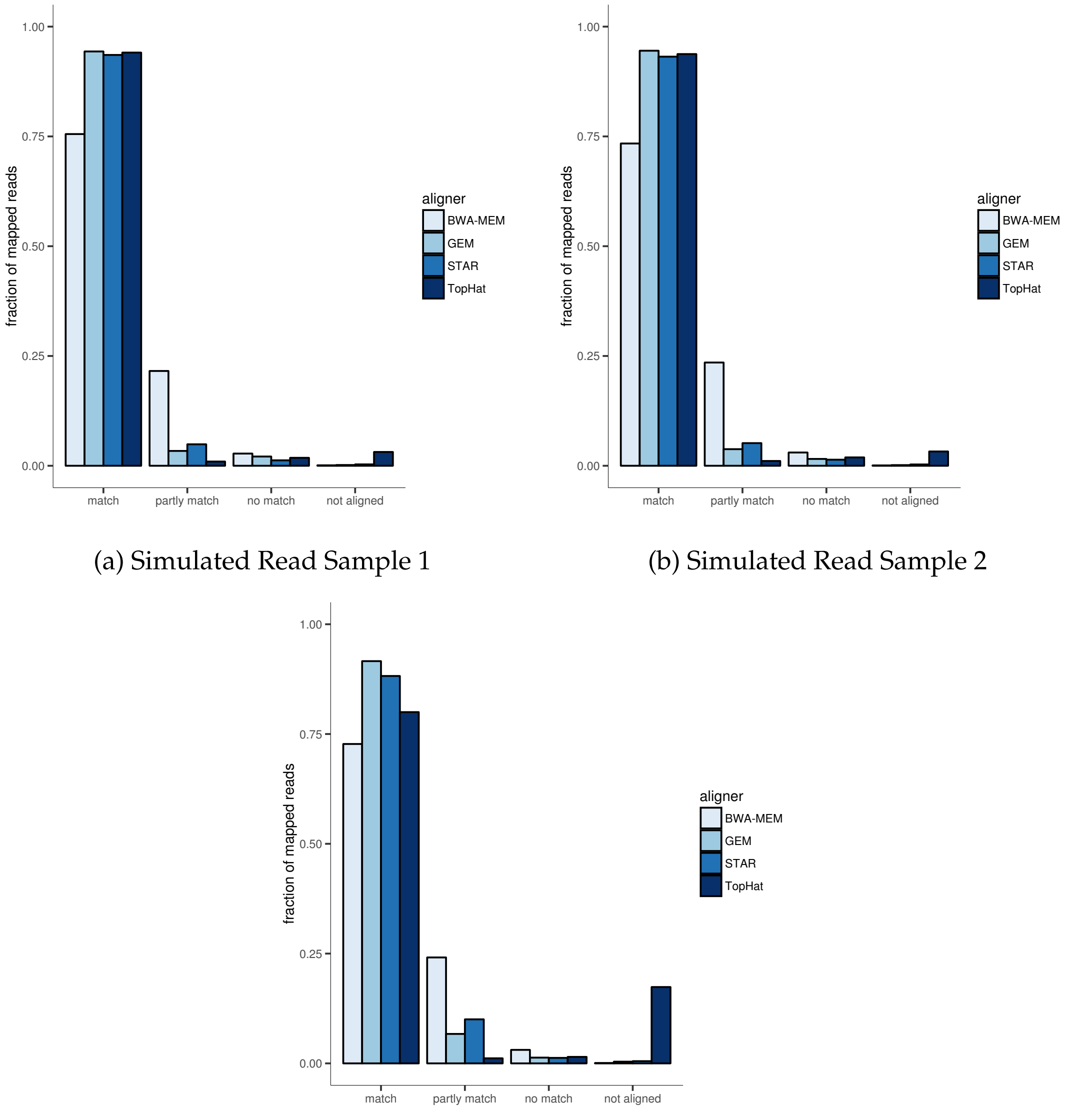


Figure 3.10.: This plot depicts the comparison of gene read counts of one sample that was aligned with all four aligners. In the lower left part of the plot logarithmized read counts are plotted against each other. The upper right part shows the respective Spearman correlation coefficient of two aligner associated read count sets each and the diagonal illustrates the density plots of the logarithmized read count sets of each aligner.

Using the numbers illustrated in Figure 3.11, alignment qualities can be evaluated in terms of precision and recall as was done in Lindner and Friedel, 2012[132] where



(a) Simulated Read Sample 1

(b) Simulated Read Sample 2

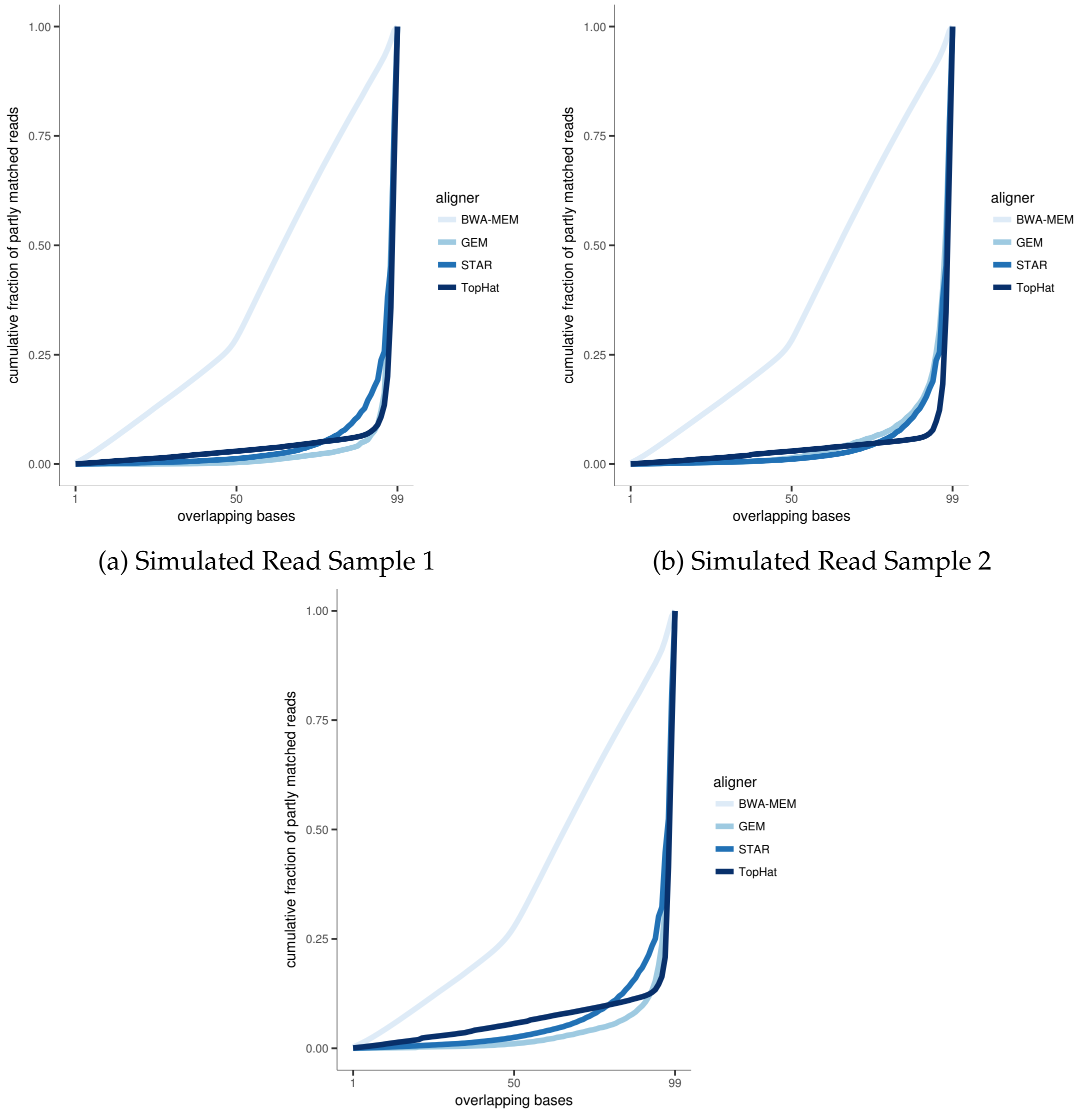
(c) Simulated Read Sample 3

Figure 3.11.: Alignment accuracy of simulated read data of the four tested alignment tools. Reads were grouped into four categories, i.e. match, partly match, no match and not aligned. Results are shown for each data set separately.

precision and recall are calculated as:

$$precision = \frac{TP}{TP + FP} \tag{3.1}$$

$$recall = \frac{TP}{TP + FN} \tag{3.2}$$



(a) Simulated Read Sample 1

(b) Simulated Read Sample 2

(c) Simulated Read Sample 3

Figure 3.12.: Cumulative distribution of the number of overlapping bases across partly matched reads. The y-axis represents the fraction of reads that show not more than x overlapping bases with the correct mapping position (x-axis).

For that purpose, alignments have to be assigned to one of the three categories, namely true positive (TP), false positive (FP) and false negative (FN):

- TP: All alignments that perfectly but also partly match the correct mapping position belong to this category. Partly matched alignments were included, first, because assignment of a read to the correct gene is paramount and second, due to the high agreement of them with the correct mapping position (Figure 3.12).

- FP: Alignments that were mapped to a wrong position are classified as FP
- FN: All alignments that could not be mapped to the correct position, i.e. unaligned reads but also incorrectly aligned ones, are included in this category. Thus, like in Lindner and Friedel, 2012[132], the sum of TP and FN is equal to the total number of analyzed reads.

Resulting precision and recall values are shown in Table 3.1. Overall, each of the four aligners yield relatively high precision as well as recall scores. However, TopHat results in slightly lower recall levels which is due to the higher number of unaligned reads.

	BWA-MEM		GEM		STAR		TopHat	
	precision	recall	precision	recall	precision	recall	precision	recall
Sample 1	0.9722	0.9712	0.9789	0.9772	0.9876	0.9844	0.9813	0.9505
Sample 2	0.9692	0.9685	0.9844	0.9829	0.9861	0.9832	0.9803	0.9484
Sample 3	0.9698	0.969	0.9868	0.9831	0.9874	0.9825	0.9821	0.8114

Table 3.1.: Performance of the tested alignment tools on the three simulated RNA-seq samples. The table shows the respective precision and recall scores for each of the four aligners.

3.1.3. Quality Control

All QC metrics provided by the pipeline are calculated after the alignment step for each sample separately using the alignment information stored in the BAM files (Chapter 2.3.2). Quality metrics for RNA-seq data were evaluated and established in the course of the GEUVADIS RNA-seq project and published in a companion paper by t'Hoen *et al.*, 2013[239]. Some of them as well as other quality metrics are used and implemented in the here presented pipeline. In the following, four of them, namely

i. sequencing depth

ii. exonic rate

iii. rRNA rate

iv. XIST and Y-chromosome gene expression

are discussed in more detail.

3.1.3.1. Sequencing Depth

A fundamental QC measure is the number of reads that result from sequencing since an insufficient amount of reads might lead to an underrepresentation of biological signals and in further consequence to potentially wrong conclusions. Of course, the number of desired reads depends on the purpose of the experiment and the composition and dimension of the transcriptome under investigation[227]. For example, detecting variants and reliably identifying variant allele frequencies in highly expressed genes requires considerably less sequencing depth than intending to do the same for very low expressed ones. The same holds true for the identification of alternative isoforms as well as differential expression analysis. Furthermore, and especially when costs are a limiting factor, there is a tradeoff between sequencing depth and the number of biological replicates[137]. According to the ENCODE Consortium[37], about 30 million paired-end reads, or fragments, are sufficient to compare expression profiles. In the GEUVADIS project the median number of single reads was 58 million[119][239] which is equivalent to 29 million paired-end reads. For the more than 1,500 samples that were processed by the here presented RNA-seq data analysis pipeline the median number of generated reads is 60.6 million, or 30.3 million fragments (Figure 3.13).

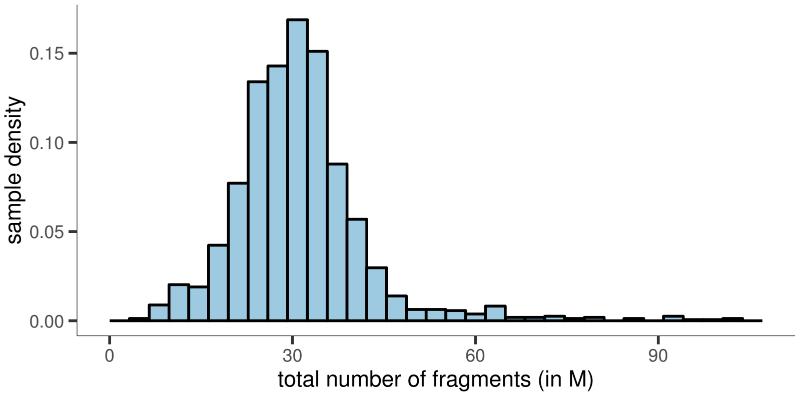


Figure 3.13.: Total number of fragments that were sequenced for each of the about 1,500 human and mouse RNA-seq samples, respectively. Numbers are indicated in million (M).

3.1.3.2. Exonic Rate

To check whether the RNA-seq library preparation worked properly and especially if the mRNA enrichment was sufficient, the fraction of reads that align to annotated exons (exonic rate) constitute an informative score. Furthermore, the exonic rate gives some indication of possible alignment problems or library contamination. The majority of the analyzed samples have more than 70% of reads mapping to annotated exons (Figure 3.14) which is in good agreement with published data[239].

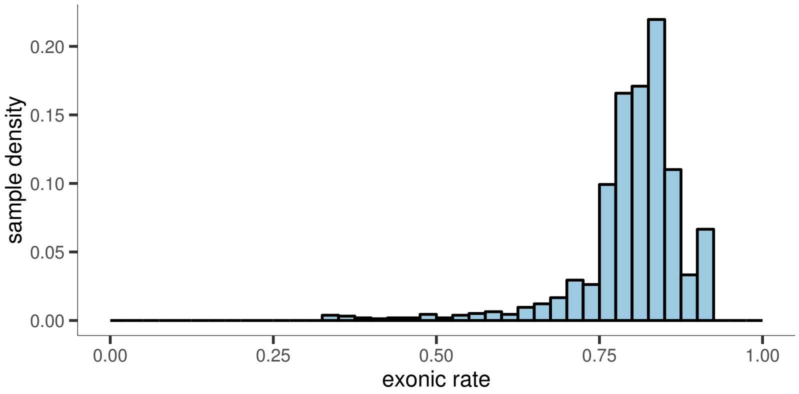


Figure 3.14.: Histogram of exonic rates for each of the analyzed RNA-seq samples.

3.1.3.3. rRNA Rate

One of the major steps of the RNA-seq library preparation is the removal of rRNA (Chapter 1.3) since it accounts for >90% of total RNA in a mammalian cell and commonly other RNA subpopulations, such as mRNA, are desired. In order to check whether rRNA removal worked properly the rRNA rate is calculated for each processed sample. rRNA rate represents the fraction of reads that map to rRNA associated regions and should be low for such samples where poly(A) capturing or rRNA depletion was performed. High rRNA rates might be an indication of problems during library preparation and samples showing this trait should be considered for exclusion from further analysis. Most of the here analyzed samples have rRNA rates below 1% (Figure 3.15) which is again in good agreement with published results[40].

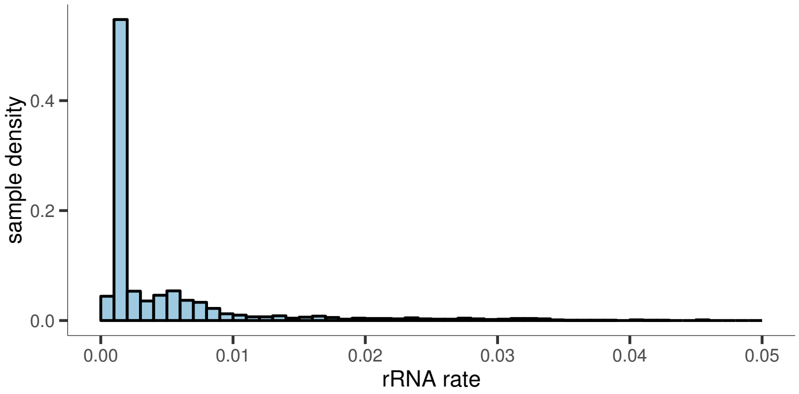
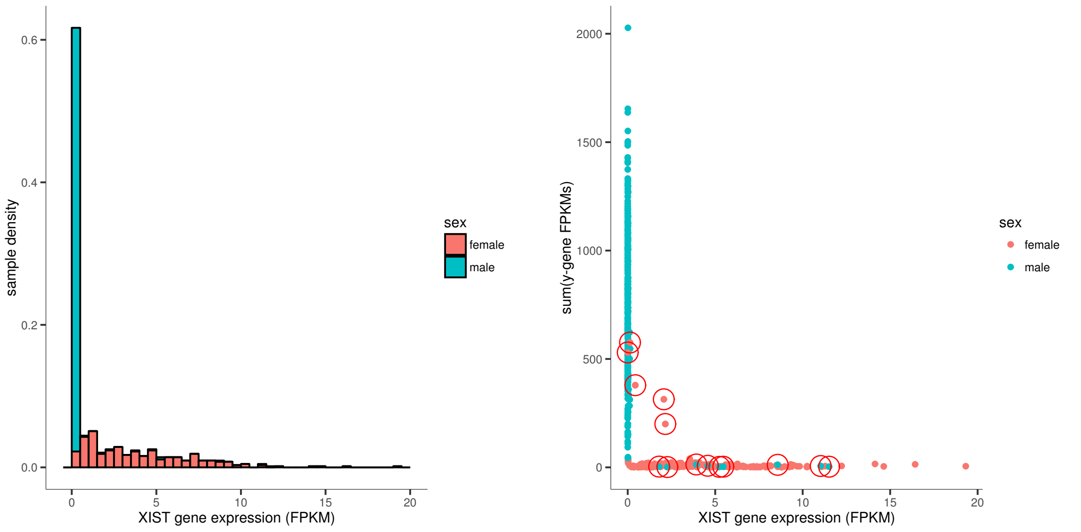


Figure 3.15.: Histogram of rRNA rates for each of the analyzed RNA-seq samples.

3.1.3.4. XIST and Y-chromosome Gene Expression

A further potential source of error is sample mix-up. If undetected this would distort analysis results. One way to check for sample mix-ups in human RNA-seq data, where the gender of the samples are known, is to screen the expression of the XIST gene. This gene is located on the X-chromosome and solely expressed in females[239]. If a sample is tagged to be of female origin and does not show XIST expression, or the other way around for male samples, this would point towards a likely sample swap. The expression strength of the XIST gene varies widely across human female samples (Figure 3.16a) but with the exception of a few male samples which show noticeable XIST expression and a few female samples with zero XIST expression the overall accordance of gender and XIST expression is high. Further evidence can be obtained by additionally checking the expression of genes on the Y-chromosome while excluding those located in the pseudo-autosomal region (Y-genes). Only male samples are expected to show expression for Y-genes. This provides even better resolution and revealed additional female samples in the data set which show expression of the XIST gene but also substantial expression of genes on the Y-chromosome suggesting likely contaminations (Figure 3.16b).



(a)

(b)

Figure 3.16.: Summary of gender-specific gene expression: a) Distribution of normalized XIST gene expression values divided by gender. b) Normalized XIST gene expression values (x-axis) plotted against the sum of normalized expression values of genes on the Y-chromosome (pseudo-autosomal regions excluded).

3.1.4. Differential Expression Analysis

One of the main features of the RNA-seq data analysis pipeline is differential expression analysis on the gene level. Three major steps are necessary for this, namely

i. expression quantification

ii. normalization

iii. differential expression analysis.

and in order to provide state of the art differential gene expression analysis several available tools were tested. After the identification of differentially expressed genes the pipeline automatically performs enrichment analysis and all of these steps are discussed in the following.

3.1.4.1. Expression Quantification

Reliable expression quantification depends on correct read alignment and quantification. Combinations of three alignment (GEM, STAR, TopHat) and two expression quantification tools (htseq-count, featureCounts) were tested. htseq-count offers three different overlap resolution modes (see Chapter 1.4.3.1) and each of them was considered independently. On top of that, the built-in quantification module of the STAR aligner was included in the test which, according to the manual[49], is implemented like htseq-count's union mode quantification algorithm but without any kind of parameter options though.

In order to test the performance of different alignment and gene expression quantification tool combinations the *in silico* sample "Simulated Read Sample 1" from Chapter 3.1.2 was used. For this sample, the true read counts per gene are provided by the simulation software and serve as gold standard.

The correlation of observed and expected values are shown in Figure 3.17. All tested combinations yield comparable results and tend to underestimate expression levels. STAR in combination with htseq-count performs slightly better than the other tested combinations as shown by the correlations coefficient of 0.771.

Runtimes of the quantification tools were tested on an Intel® Xeon® CPU E5-2697 v3 @ 2.60GHz machine. A clear difference between individual quantification tools can be observed in terms of runtimes (Figure 3.18). htseq-count, irrespective of the used mode, takes by far the longest with nearly 60 minutes on average for the about 40 million reads of the simulated sample. In contrast, STAR's quantification module and featureCounts are both fast with runtimes of around 2 and 3 minutes, respectively.

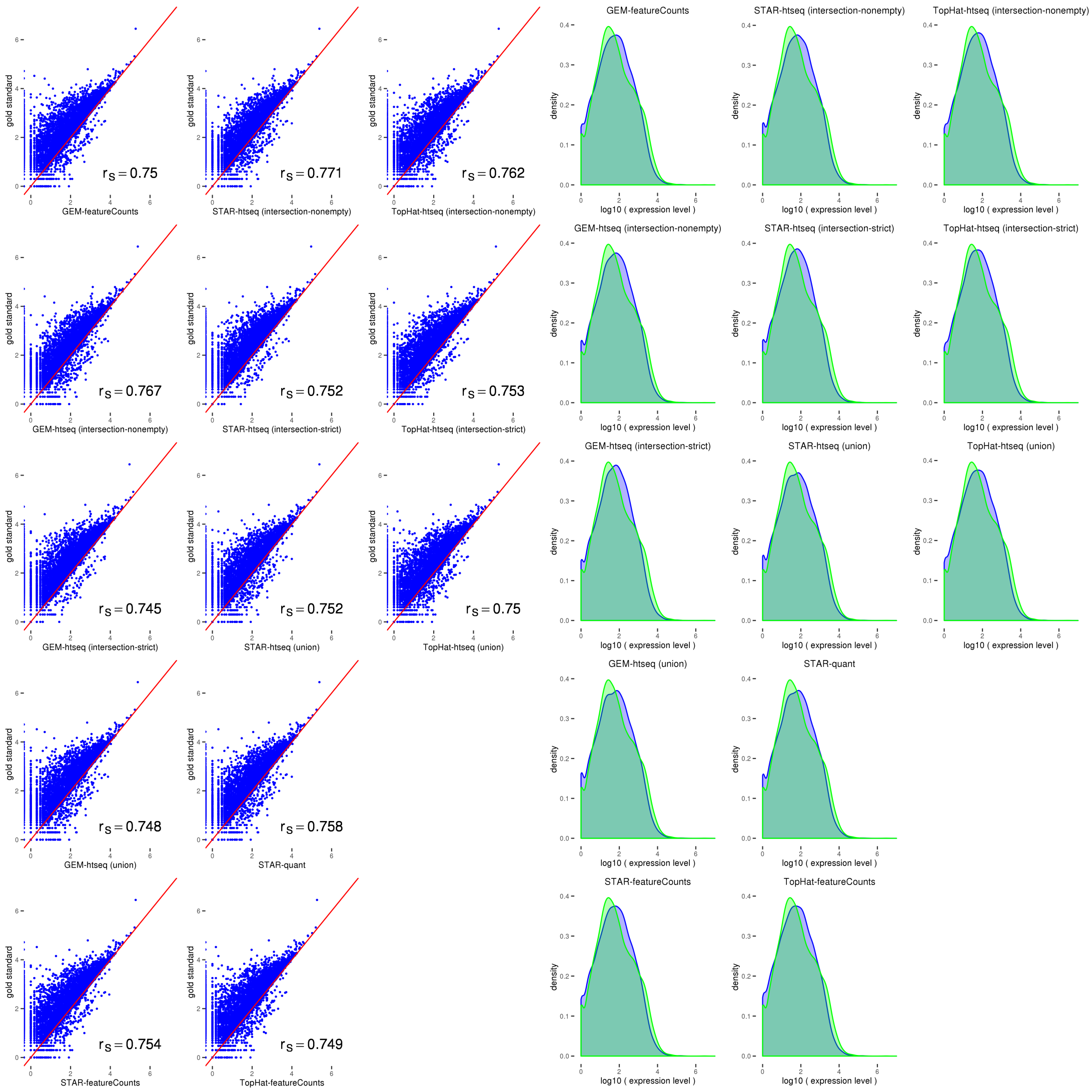


Figure 3.17.: Performance of different alignment and expression quantification tool combinations. Left: Scatter plots showing the correlation between the observed (x-axis) and the expected (y-axis) gene counts for each of the tested tool combinations in log-scale and Spearman correlation coefficients are added. Right: Observed (blue) and expected (green) gene count distributions for each of the tested combinations, again in log scale and in the same order as scatterplots on the left.

3.1.4.2. Normalization

Normalization of expression values is required before different samples are compared to each other. Initially, FPKM values are calculated and stored in the database. The

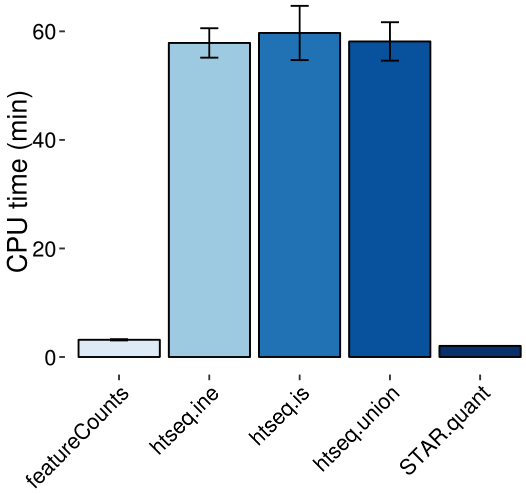
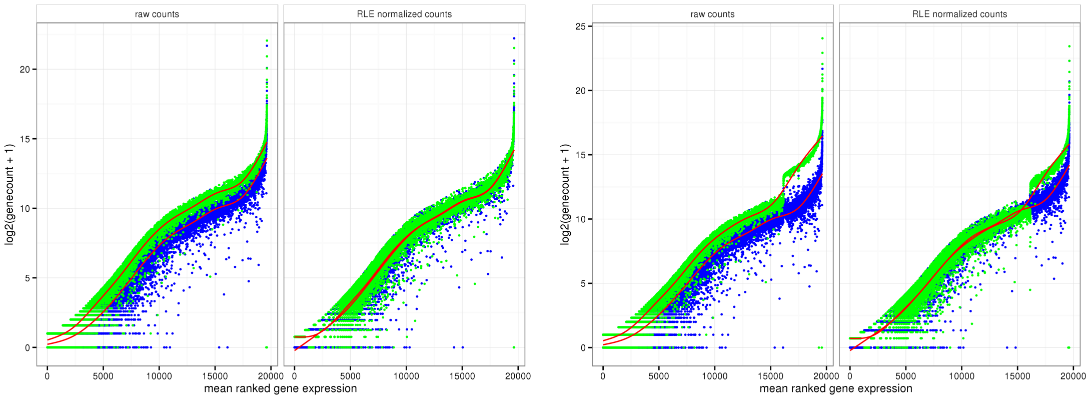


Figure 3.18.: This figure shows the average runtimes of the test expression quantification tools including different options for htseq-count (ine=intersection-nonempty, is=intersection-strict, union). Average runtimes were calculated for each tool based on the three different alignments except for STAR's quantification module where only STAR alignment could be used. Error bars indicate the standard deviation.

normalized FPKM measures serve as a valuable source of information for users who want to get an idea of the overall expression characteristics of a sample or of specific genes of interest. FPKM values together with raw read counts can be easily queried via the provided web interface (Figure 3.47 in Chapter 3.4).

However, FPKM values are not convenient for differential expression analysis. Several studies showed that the FPKM normalization method is prone to underestimate lowly expressed genes especially when a small set of very high expressed genes is present[25][35][48][178][206] (Figure 1.12 in Chapter 1.4.3.1). Thus, more sophisticated between-sample normalization methods have been introduced. DEseq as well as DESeq2 use a Relative Log Expression (RLE) approach for normalization[8][140] which, they state, should not have the abovementioned limitations. To illustrate the difference between FPKM and RLE, the same data used for FPKM normalization in Figure 1.12 were used here in the same manner for RLE normalization (Figure 3.19). Other than FPKM, this method does not mistakenly underestimate lower expressed genes. Instead, the subset of highly expressed genes remains overexpressed while the expression of the remaining genes were adjusted accordingly.



(a)

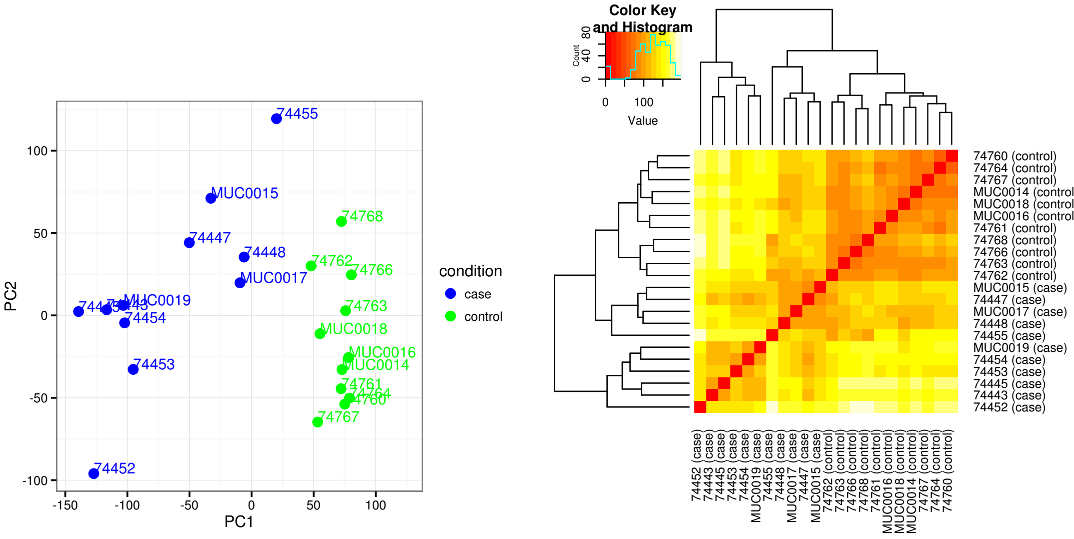
(b)

Figure 3.19.: In this figure, DESeq2's RLE normalization strategy is shown for the same two samples (blue and green) as used for Figure 1.12. a) Normalization of two data sets with different sequencing depths yet comparable gene expression patterns show good agreement (red fitted regression curve) after RLE normalization (which is also the case for FPKM values (see Figure 1.12)). b) RLE normalization of the same two samples but this time with one sample having a subset of highly and differentially expressed genes resulting in proper normalized values for the set of not differentially expressed genes while preserving the information regarding the subset of overexpressed ones. (Idea for this plot is based on [71])

3.1.4.3. Differential Gene Expression Analysis

The aim of differential gene expression analysis is to identify genes that show dissimilar expression levels across different conditions. Each condition should be represented by at least three replicates which is the minimal amount to enable a sufficient estimation of within-group variance and gene expression levels and in further consequence to make inference on the population[35]. The proper amount of replicates depends first, on the technical variability which might be introduced anywhere between sample extraction right up to final gene expression quantification and second, on the biological variability of the biological system to analyze[35]. Technical variability can be reduced when samples are processed in batches. Biological variability, on the other hand, depends strongly on the type of the experiment. For example, expression variability can be expected to be low in model organisms or cell culture experiments but high in population-based studies or samples from individuals affected by diseases like cancer.

When testing for differential expression across conditions one would expect that samples of the same group, i.e. biological replicates, cluster together and if this is not the case it might be indicative for some unwanted bias in the data to analyze. During this PhD project various RNA-seq samples from different projects were analyzed with the pipeline (Chapter 2.1). One of them comprises 11 pairs of liver tumor and matching normal tissue samples which constitutes the project with the highest number of replicates analyzed so far. In order to enable the user to investigate the conformity between groups in a convenient way the pipeline produces several figures including principal component analysis (PCA) plots as well as heatmaps displaying euclidean distances between samples as provided by DESeq2 (Figure 3.20). In this example,



(a) (b)

Figure 3.20.: Cluster analysis for 11 tumor-control sample pairs: a) Principal component analysis plot b) Heatmap showing euclidean distances between samples.

a separation between the two sample groups can be observed by the first principal component of the PCA plot (Figure 3.20a) as well as by the cluster of control samples on the top right of the heatmap plot (Figure 3.20b). The normal tissue samples cluster together in a more consistent way while tumor samples show more heterogeneity.

After checking the integrity of the samples under study the actual differential expression analysis can be performed. By the time of writing, three differential gene expression analysis tools (DESeq, DESeq2 and edgeR) were included in the pipeline from which the user can choose, with DESeq2 being the pipeline's default option. 5,621 significant hits (Benjamini–Hochberg adjusted p-value < 0.01) were reported by DESeq2 when looking for differentially expressed genes across the 11 tumor and control samples. The results of each performed differential expression analysis are stored in the database as well as in a tabular text file including the following information:

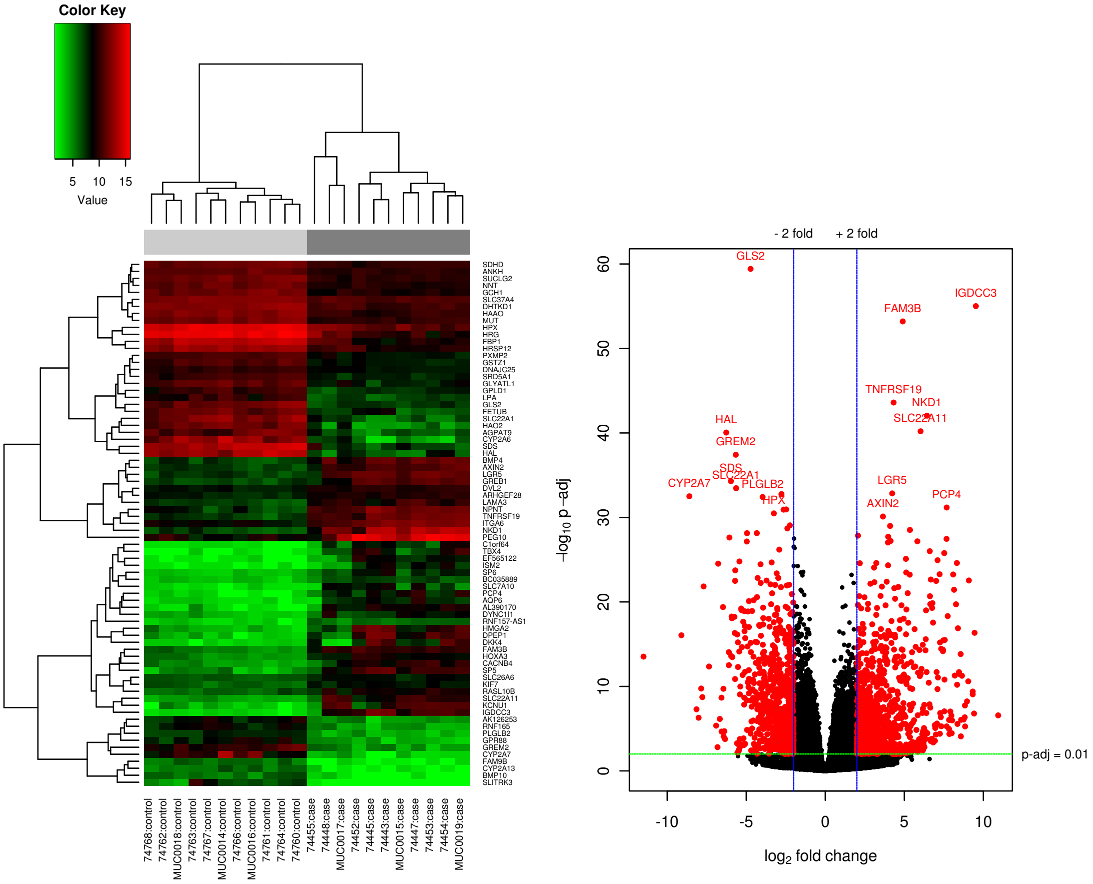
```
gene     meanCount   log2FC   p-value              padj                 regulation
IGDCC3   841.807     8.572    3.20203719520227e-60 4.40116894317215e-56 up
GLS2     1960.95     -4.607   3.52093515453772e-60 4.40116894317215e-56 down
FAM3B    567.04      4.754    4.41583595114812e-54 3.67986329262344e-50 up
TNFRSF19 1560.41     4.201    2.36395609102288e-44 1.4774725568893e-40  up
NKD1     4842.47     6.023    8.67876162955515e-43 4.33938081477758e-39 up
HAL      5558.49     -5.866   3.1831419482173e-41  1.32630914509054e-37 down
```

where these six lines depict six top differentially expressed genes from the aforementioned tumor-control analysis. Column one represents the name of the gene, column two the average read count across all samples involved, column three the binary logarithm of the fold change of the mean read count of all tumor samples compared to the mean read count of all control samples of the respective gene, column four the assigned p-value, column six the Benjamini–Hochberg adjusted p-value and column seven whether the gene is up- or down-regulated in the tumor samples compared to the control samples. Furthermore, the pipeline produces several plots for visualization of the results (Figure 3.21).

Comparison of DEseq2 and edgeR Results

DESeq or DESeq2 as well as edgeR are widely used tools for differential gene expression analysis[8][241]. However, previous studies showed that there are notable differences in the output of these tools, especially between DESeq and edgeR[172][223][230][266]. Yet, this issue can be illustrated for DESeq2 and edgeR as well. For that purpose, the 11 tumor-control samples were analyzed with both tools independently. Figure 3.22a shows the agreement of DESeq2 and edgeR results in terms of significantly detected genes (Benjamini–Hochberg adjusted p-value < 0.01). In comparison to the 5,621 genes reported by DESeq2, edgeR returns roughly 20% more genes (6,649) using the same adjusted p-value threshold. The number of overlapping genes between the result lists is 3,985. Hence, 1,636 and 2,664 genes were solely returned by DESeq2 and edgeR, respectively. Furthermore, the ranking of genes reveals a notable difference between DESeq2 and edgeR (Figure 3.22b), with gene ranks based on first, ascending adjusted p-values and second, ascending p-values. The same fact can be observed from the distribution of p-values and adjusted p-values for the top 5,000 differentially expressed genes as well (Figure 3.22c). While DESeq2 assigns smaller p-values for top hits, edgeR is more conservative for p-values $> 10^{-7}$ and adjusted p-values $> 10^{-5}$, respectively. Based on these results, so far, the pipeline offers the option to use both tools for differential expression analysis.

In order to test the influence of different numbers of replicates on DESeq2 and edgeR, respectively, differential gene expression analysis was performed for the same data set but this time with different numbers of replicates, with numbers of replicates ranging from 2 to 11. All possible $\binom{n}{k}$ combinations were tested for each group where n is the total number of sample pairs, i.e. 11, and k the respective number of replicates in the group. Average total number of reported significant genes (Benjamini–Hochberg adjusted p-value < 0.01) as well as average overlap between DE-

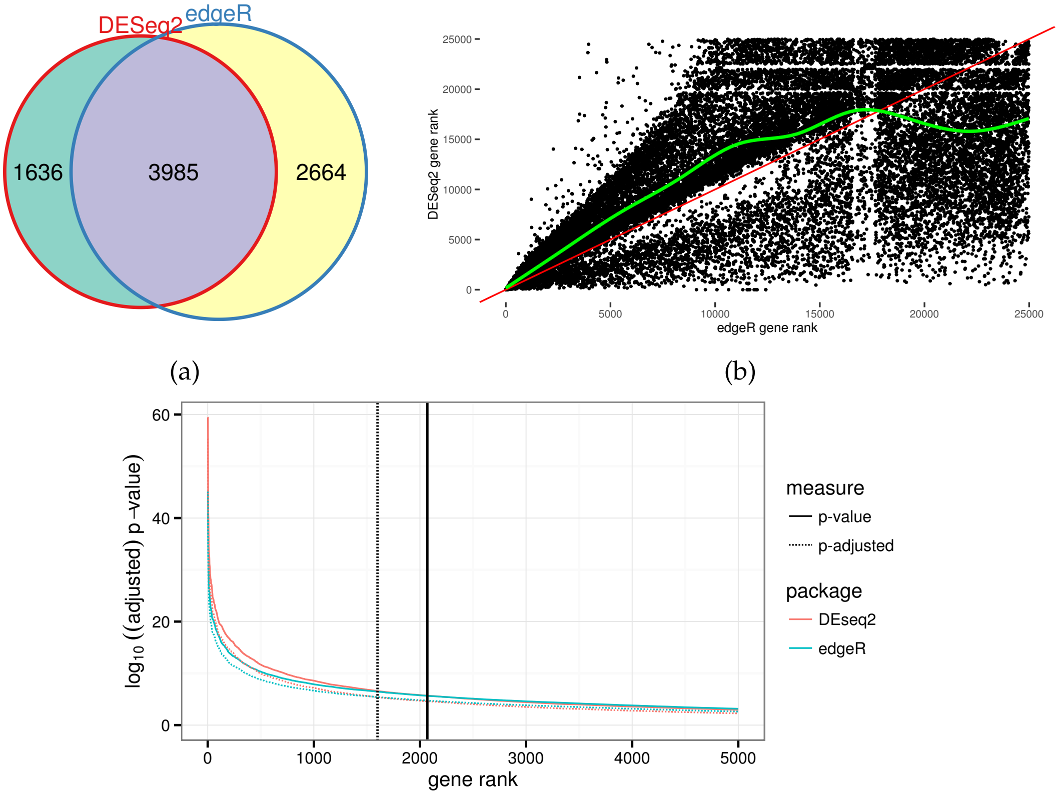


(a) (b)

Figure 3.21.: Subset of differential expression result plots: a) Heatmap showing the top 75 differentially expressed genes (according to adjusted p-values) b) Volcano plot illustrating the overall expression relationship between the two conditions plus top gene hits.

Seq2 and edgeR results for each group were calculated. As can be seen in Figure 3.23, there is a clear relationship between number of replicates and number of reported significant genes. Strikingly, the number of detected genes does not reach a plateau with higher numbers of replicates indicating that for very heterogenous samples, like tumors, even more replicates are necessary for a representative inference.

Enrichment Analysis The pipeline automatically performs enrichment analysis in two different ways. For this purpose, the results of the preceding differential gene expression analysis step are used. First, Gene Ontology (GO) enrichment analysis and second, pathway enrichment analysis are performed on a set of differentially ex-



(a) (b)

(c)

Figure 3.22.: Comparison of DESeq2 and edgeR results based on the analysis of the 11 tumor-control sample pairs: a) Total number and overlap of reported significantly differentially expressed genes (Benjamini–Hochberg adjusted p-value < 0.01). b) Gene ranks based on adjusted p-values and p-values, respectively, for both, DESeq2 and edgeR, and plotted against each other where each dot represents a gene and the particular x-coordinate the rank assigned by edgeR and the y-coordinate the rank assigned by DESeq2. If ranks would totally agree, all dots would lie on the red line and the green fitted regression curve would superimpose it. c) Gene ranks as in b) on the x-axis and the respective (adjusted) p-value on the y-axis.

pressed genes which are selected based on an user-defined adjusted p-value threshold (default: Benjamini–Hochberg adjusted p-value < 0.01).

Using the results of the differential gene expression analysis of the 11 tumor-control samples of the previous chapter, 799 significantly enriched GOs (Benjamini–Hochberg adjusted p-value < 0.01) could be detected. The results of the GO enrichment analysis are stored in tabular text files and the top six significant hits of the analysis are shown in the following:

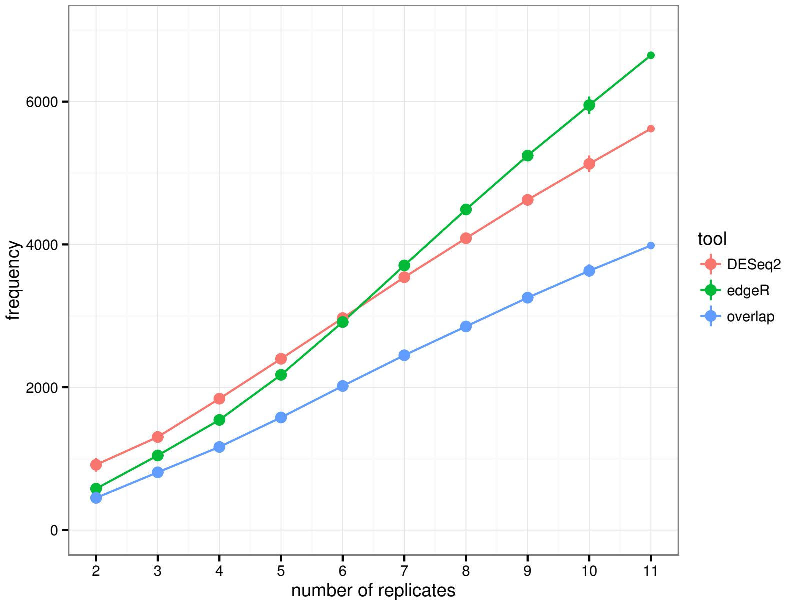


Figure 3.23.: Relationship between number of replicates (x-axis) and number of reported significantly differentially expressed genes (y-axis) for DESeq2 (red) and edgeR (green). Furthermore, the overlap between the reported genes are shown (blue). The average number of genes (dot) and the standard error of the mean (coloured vertical lines) for each replicate quantity group are depicted.

```
category    term                                ontology pvalue        padj
GO:0044281  small molecule metabolic process    BP       3.603045e-79  7.254010e-75
GO:0019752  carboxylic acid metabolic process   BP       1.109189e-70  1.116565e-66
GO:0006082  organic acid metabolic process      BP       1.826085e-64  1.225485e-60
GO:0043436  oxoacid metabolic process           BP       3.464821e-64  1.743931e-60
GO:0055114  oxidation-reduction process         BP       1.984909e-62  7.992435e-59
GO:0016491  oxidoreductase activity             MF       1.020264e-56  3.424396e-53
```

where the first column represents the GO identifier, the second column the description, the third one the domain with possible values being BP (biological process), CC (cellular component) and MF (molecular function), the fourth the assigned p-value and the fifth column the Benjamini–Hochberg adjusted p-value.

Performing pathway enrichment analysis based on the same differential gene expression analysis results, the pipeline reports 37 significantly enriched KEGG pathways (Benjamini–Hochberg adjusted p-value < 0.1). Again, results are stored in tabular text files which have the following format (extract of six resulting pathways):

```
keggid   name                                  p-value       padj
hsa04610 Complement and coagulation cascades   4.365456e-08  2.525102e-06
hsa00071 Fatty acid metabolism                 3.630276e-05  6.534498e-04
hsa04630 Jak-STAT signaling pathway            1.278777e-04  1.726349e-03
hsa04976 Bile secretion                        3.297373e-03  2.428066e-02
hsa04210 Apoptosis                             6.137185e-03  3.823938e-02
hsa03320 PPAR signaling pathway                1.885611e-02  8.485249e-02
```

where the first column represents the KEGG identifier, the second one the name of the pathway, the third the assigned p-value and the fourth the Benjamini–Hochberg adjusted p-value. On top of that, the pipeline creates illustrative figures for each of the identified pathways (Figure 3.24).

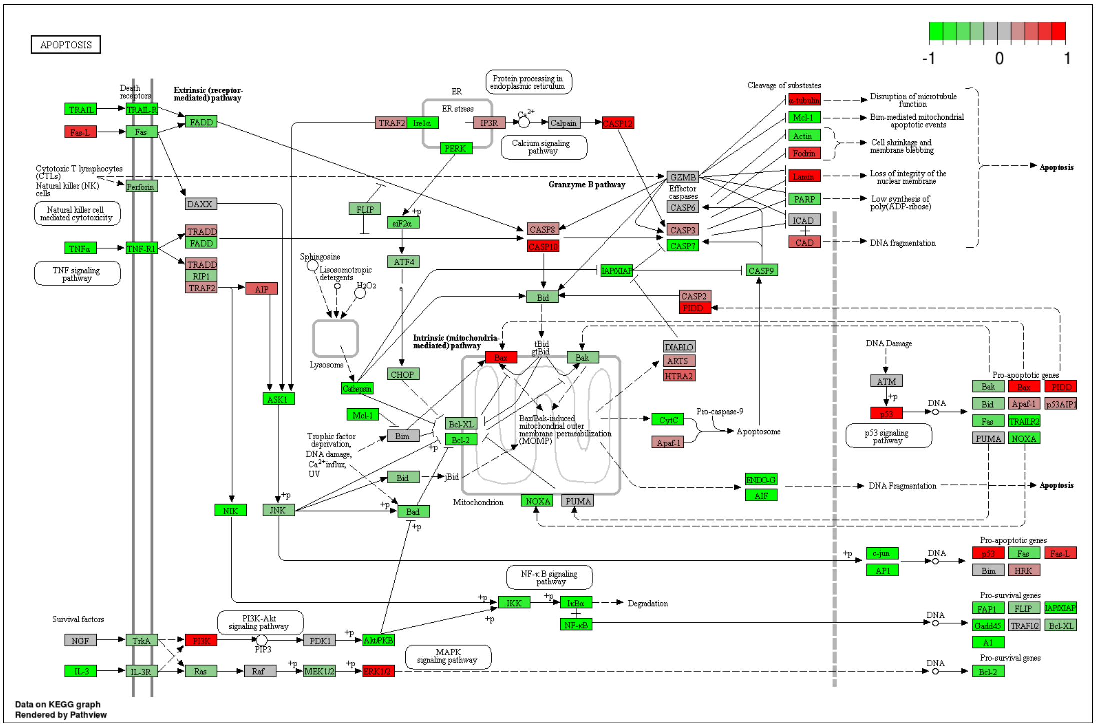


Figure 3.24.: One of the enriched pathways identified in the analysis of the 11 tumor-control sample pairs. Genes or respective proteins are represented by rectangular boxes and the background colour indicates whether the gene is upregulated (red) or downregulated (green) in the tumor samples compared to the normal tissue samples.

3.1.5. Alternative Splicing Analysis

RNA-seq data can be used to analyze splicing as well. One approach is to compare samples of different conditions to each other and check if they show some splicing differences. Another approach is to check for aberrant splicing events. If matching whole genome sequencing data are on hand, one can check whether mutations result in aberrant splicing events in an individual. By the time of writing, the pipeline offers two different ways to analyze alternative splicing, i.e.

i. differential exon usage analysis

ii. Sashimi plots

which could be successfully used in a number of projects and are explained in the following.

3.1.5.1. Differential Exon Usage Analysis

First of all, the pipeline provides the option to detect differential exon usage across different conditions. This functionality was used, for example, in a project with 30 sequenced mouse RNA-seq samples. 15 of these mice were mitochondrial brown fat uncoupling protein 1 deficient which was achieved by targeted inactivation of the Ucp1 gene as described earlier[57]. In short, an essential membrane-spanning domain located in exon 2 and partly exon 3 was deleted, thus disabling the functionality of the resulting protein[57]. The remaining 15 mice were wild-type. To test whether the experiment succeeded, differential exon usage analysis was performed with the pipeline and the result of the test is shown in Figure 3.25. As expected, exon 2 and to a minor degree exon 3 are considerably underrepresented compared to the remaining four exons.

3.1.5.2. Sashimi Plots

The second approach for splicing analysis can be used by loading the BAM files of the samples of interest into IGV. IGV has a built in functionality to produce Sashimi plots which are used to quantitatively visualize alternative isoform expression[100][101]. The step of loading the desired samples into IGV is provided in a convenient way via the web interface. IGV requires split read alignments in BAM file format, as created by common alignment tools such as STAR, and a reference gene annotation which defines the exon boundaries in order to create Sashimi plots. Several gene annotations are already included in IGV and can be used for support. The BAM files, on the other hand, are streamed in a secure way over the Internet to the local IGV of the user where the alignments can be inspected and in further consequence the Sashimi plots created (Figure 3.26).

Exploiting this feature helped to discover atypic isoforms of the gene FLAD1 caused by biallelic frameshift variants which could be linked to Multiple acyl-CoA dehydrogenase deficiencies (MADDs)[176].

Furthermore, another gene encoding the respiratory chain complex I assembly factor TIMMDC1 could be identified and selected for further testing. As a result, TIMMDC1 could be established as a mitochondrial disease-associated gene[114].

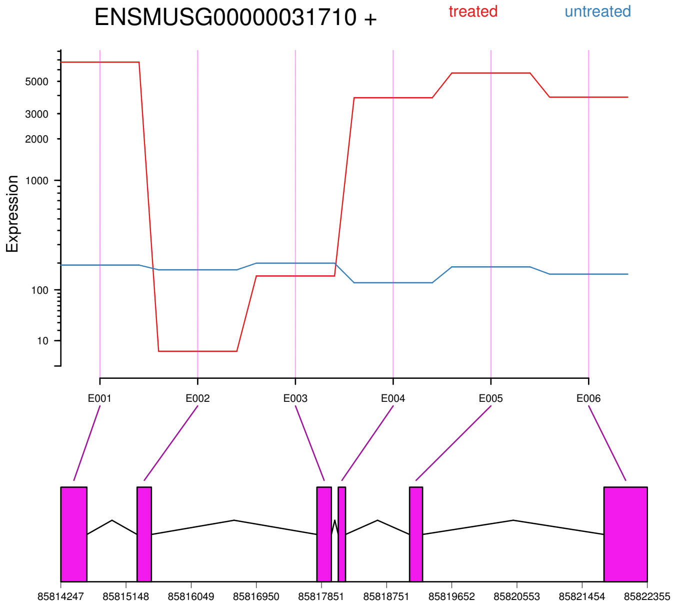


Figure 3.25.: Differential exon usage plot produced with DEXSeq: Normalized average read counts are shown for each exon of the gene encoding for the mitochondrial uncoupling protein and for both groups, UCP deficient (red) and UCP wild type (blue).

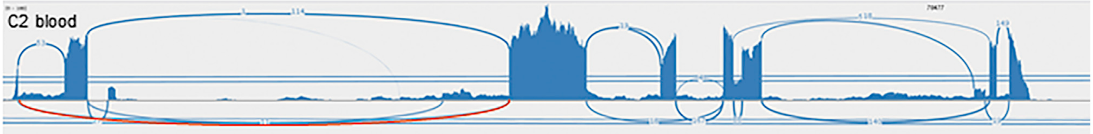


Figure 3.26.: Sashimi plot of the FLAD1 gene as published in [176]. (Figure adapted from [176])

3.1.6. Gene Fusion Analysis

Gene fusion detection is sensitive to various NGS data properties which hampers the reliable identification of fusion break points (Chapter 1.4.3.2). The major limiting factor is the short length of the sequencing reads which makes them prone to incorrect alignment due to polymorphisms and homology, respectively, and in further consequence to missed or mistakenly reported fusion events[35]. Different studies have already compared various fusion detection tools to each other but so far there is no single tool that outperforms all others but rather each of them have their strengths and weaknesses[28][29][115][134]. Here, another tool, the fusion detection module of the STAR aligner, is compared to TopHat-Fusion. Furthermore, several filtering strategies were tested in order to increase the accuracy of fusion detection. For this purpose, four extensively investigated breast cancer cell lines[55] (BT474, SKBR3, KPL4, MCF7) were analyzed and raw data were downloaded from the NCBI Sequence Read Archive (SRA)[1] with accession number SRP003186.

3.1.6.1. Gene Fusion Detection Software Comparison

The two tools used for comparison are TopHat-Fusion v2.1.0[108] and the integrated fusion detection algorithm of the STAR aligner. Using parameter settings as recommended[2,3], TopHat-Fusion and STAR reported 179 and 66 fusions, respectively. The results of both tools were then evaluated in terms of precision (Equation 3.1) and recall (Equation 3.2). Reported fusions were compared with the respective and continually refined fusion result sets of previous studies of the four breast cancer cell lines[55][97][134] in order to obtain the numbers of true positive, false positive and false negative fusions. Table 3.2 illustrates the performance of the two tools. This comparison reveals that overall, TopHat detects more true positive fusions. However, this is accompanied with a higher number of reported false positive ones which underlines the importance of a balanced, i.e. not too stringent nor too loose, filtering strategy.

		BT474	SKBR3	KPL4	MCF7
total	known	21	10	3	6
	STAR	24	21	10	11
	TopHat	94	51	17	17
precision	STAR	0.29	0.24	0.2	0.09
	TopHat	0.2	0.1	0.12	0.24
recall	STAR	0.3	0.5	0.67	0.17
	TopHat	0.9	0.5	0.67	0.67

Table 3.2.: Performance of STAR-Fusion and TopHat-Fusion on four extensively tested breast cancer RNA-seq data sets. “known” indicates the total number of validated fusions for each of the four breast cancer cell lines based on previous studies[55][97][134].

3.1.6.2. Filtering

Since STAR missed more true positive fusions, TopHat-Fusion v2.1.0[108] was implemented in the pipeline for gene fusion detection. In order to try to decrease the

[1] `http://www.ncbi.nlm.nih.gov/sra` (last accessed 01.10.2016)

[2] `http://ccb.jhu.edu/software/tophat/fusion_tutorial.shtml` (last accessed 01.10.2016)

[3] `https://github.com/STAR-Fusion/STAR-Fusion/wiki` (last accessed 01.10.2016)

amount of mistakenly reported fusions, i.e. false positives, several filtering strategies were tested and applied to TopHat-Fusion's results.

The first one checks whether the fusion breakpoints are located within a ±5 bp window around known exons.

The second filter is a rough estimate of the conformity of coverage (COC) at the fusion breakpoints of two fusion partner genes A and B and is calculated as

$$COC_{fusion_{AB}} = \left| log_2 \left(\frac{\frac{r_A}{tr_A}}{\frac{r_B}{tr_B}} \right) \right| \tag{3.3}$$

where r_A is the number of reads directly mapping to the fusion breakpoint of gene A, tr_A the total expression level of gene A and equivalent r_B and tr_B for gene B. As a consequence, fusions exhibiting a value close to zero should be kept while values far from zero can be indicative for questionable fusions. Fusions falling in this category should be treated with caution and possibly require additional evaluation. This formula was implemented in the pipeline and is based on earlier findings that true fusion events are supported by a comparable read distribution at the breakpoints[13][55][108].

The final filter verifies whether the two fusion partners are known ENSEMBL paralogs or known duplicate genes[179].

All of the abovementioned filters have a supportive character, meaning that they do not remove fusions from the final output but just add additional annotations to each of the reported fusions.

Harnessing this complementing information by keeping only those fusions of the TopHat-Fusion results where first, at least one of the two breakpoints is located in close proximity to an annotated exon, second, they receive COC values between 0 and 2, and third, both involved genes are not annotated as ENSEMBL homologs nor duplicate genes, reduced the list of initially 179 detected fusions to 50. Calculating precision and recall values based on the list of filtered fusions shows a clear improvement (Table 3.3). The total number of false positives could be decreased from 149 to 24. However, three putative true positive fusions were filtered since they were present in the duplicate gene list suggesting to be conservative in interpreting fusions flagged with this information.

	BT474	SKBR3	KPL4	MCF7
total	28	10	6	6
precision	0.64	0.3	0.33	0.5
recall	0.86	0.3	0.67	0.6

Table 3.3.: Performance of TopHat-Fusion plus custom filtering steps on the four breast cancer RNA-seq data sets.

Finally, results reported by the pipeline's fusion detection module can be compared to and in further consequence annotated with publicly available fusion data sets in order to detect possible recurrent chimera candidates. The data sets which are used for this kind of comparison were downloaded from the Catalogue Of Somatic Mutations In Cancer (COSMIC)[63] and the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer[163].

3.1.7. Variant Detection

RNA-seq data comprise sequence information, thus they can be used for variant detection as well. This possibility can be valuable in terms of detecting functionally important variants since only expressed sequence information is represented by RNA-seq data[196]. Furthermore, it allows for the examination of RNA editing and can help to validate variant calls from matching DNA-seq experiments, if available[189][196]. However, variant detection in RNA-seq data is not a trivial task due to factors like weak expression, allelic imbalances or alignment problems at splice junctions[68] (Chapter 1.4.3.3). To evaluate the performance of three prominent variant calling tools, benchmark tests were conducted. The three tested tools are GATK v3.5, SAMtools v0.1.19 and VarScan v2.3.9. GATK was executed as described in Chapter 2.3.5 and SAMtools and VarScan with standard parameters except for `-d 999999` for SAMtools in order to allow for variants in highly expressed genes. The performance of the tools was determined on the basis of two tests, i.e.

i. comparison to a gold standard data set

ii. comparison to available and matching whole exome and whole genome sequencing data.

One of these tests was also used by Wieland, 2015[256] in a comparable way to examine the performance of variant calling programs on exome sequencing data. Here, the alignment was performed using STAR aligner and variant calling was performed for single-nucleotide variants (SNVs) and indels. Due to their distinct characteristics, SNVs and indels are treated separately in the following sections unless stated otherwise.

3.1.7.1. Comparison to a Gold Standard Dataset

One way to assess the performance of a variant calling tool is to compare its resulting set of variants with a corresponding high confidence variant call set and calculate statistical performance measures such as precision and recall where precision represents the fraction of reported variants that are actually true and recall the fraction of true variants that could be identified correctly. Both values were calculated as stated in Equation 3.1 and 3.2, respectively.

In order to validate variant calls the Illumina tool hap.py v0.3.1[4] was used as sug-

[4] `https://github.com/Illumina/hap.py` (last accessed 01.10.2016)

gested by the Global Alliance for Genomics and Health (GA4GH) Benchmarking Team[5]. The reason to do so is that especially complex mutations can be stored in different yet correct ways in the VCF format and thus mistakenly treated as wrong calls if not accounted for[269]. hap.py overcomes this issue by not comparing the VCF file entries itself but rather build all possible haplotype sequences for all variants and compare those to each other. Based on this comparison the analyzed variants are classified into one of the three categories TP, FP and FN which are defined as follows:

- TP: In this category are all those variants of the test variant set that are present in the gold standard variant set with matching genotypes
- FP: Variants of the test variant set that are not present in the gold standard variant set or that have differing genotypes
- FN: Variants in the gold standard variant set that are not present in the test variant set or that have differing genotypes

A gold standard variant set is provided by the Genome In a Bottle Consortium[6]. It was first published by Zook *et al.*[269] in 2014, constantly refined ever since and contains high confidence SNP as well as indel calls. The VCF file[7] containing the gold standard variants as well as a BED file[8] that specifies gold standard regions in which confident variant calling should be possible were downloaded and used for the benchmark tests in this chapter. The BED file includes genomic human regions from the 22 autosomes and sex chromosome X and covers about 2.5 Gb. The VCF file comprises nearly 4.3 million variant calls from the HapMap/1000 Genomes CEU genome NA12878 and evolved from sequencing of this genome with different instruments but also analyzing the resulting data with various alignment and variant calling tools. NA12878 stems from the lymphoblastoid cell line GM12878 (Coriell Institute) and RNA-seq raw read data from this cell line can be freely downloaded from the NCBI SRA with accession number SRX082565. This data set comprises about 47 million 75 bp paired-end reads which is in good agreement with the average number of reads per sample the pipeline usually has to deal with.

By aligning these reads with STAR and subsequently calling variants with the three different tools stated above a total of 164,833, 171,592 and 68,615 variants were called by GATK, SAMtools and VarScan, respectively. These numbers include SNVs as well as indels and were obtained without applying any kind of specific filtering or region restrictions, except for VarScan which, by default, does not report any mutations with

[5] `https://github.com/ga4gh/benchmarking-tools` (last accessed 01.10.2016)

[6] `http://jimb.stanford.edu/giab` (last accessed 01.10.2016)

[7] `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2.2/NA12878_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_ALLCHROM_v3.2.2_highconf.vcf.gz` (last accessed 01.10.2016)

[8] `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2.2/NA12878_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_ALLCHROM_v3.2.2_highconf.bed` (last accessed 01.10.2016)

less than 8 supporting reads. In the following, the comparison of the gold standard variant set with the identified variants in the RNA-seq data is restricted to the confidential gold standard variant calling region (2.53 Gb). Furthermore, only variants that are located within annotated exons of the UCSC gene annotation discussed in Chapter 3.1.1 are used for downstream analysis since the gold standard data set comprises variants from the entire human genome and RNA-seq data cover mainly transcribed exons which represent just a fraction of the genome. Using these restrictions, 62,423 of the roughly 4.3 million variants remained in the gold standard variant set. For the RNA-seq variant calling results of GATK, SAMtools and VarScan the resulting numbers are 26,076, 25,987 and 19,001, respectively. The performance of the tools is illustrated in Table 3.4. GATK performs best in terms of precision and recall for both classes of variants, SNV and indel. Nevertheless, these numbers reveal that there is a considerable amount of variants which could not be detected (FN) or which were detected but are not present in the gold standard variant set (FP). For example, 20,807 of the 24,099 SNVs identified by GATK overlap with the gold standard variant set leaving 3,292 SNVs exclusively found in RNA-seq data. On the other hand, 35,676 SNVs were not detected at all. A big fraction of undetected variants can be explained by low- or unexpressed genes but other factors might contribute as well. The following sections discuss several of these factors and SNVs and indels are addressed separately.

	GATK		SAMtools		VarScan	
	SNV	Indel	SNV	Indel	SNV	Indel
total	24,099	1,977	24,343	1,644	17,971	1,030
precision	0.86	0.87	0.84	0.76	0.85	0.62
recall	0.37	0.29	0.36	0.21	0.27	0.11

Table 3.4.: Results of the comparisons between the refined gold standard variant set and each of the three refined variant caller outputs. Total numbers of SNVs as well as indels identified by the three tools are shown together with precision and recall measures. All numbers are based on variants that are located within annotated exons only.

Single-nucleotide Variants

This section focuses on the identification of reasons for the FP and FN SNV calls of GATK to define quality and filter criteria for reliable SNV calls. Especially

i. coverage

ii. RNA editing

iii. genotype discordance

are discussed in detail.

Coverage As already mentioned, low coverage is a serious factor favoring both, FP as well as FN SNVs (Figure 3.27a). In fact, the majority (35,063 out of 35,676) of FN GATK SNVs have a read coverage below 10 and 26,707 of these 35,063 sites

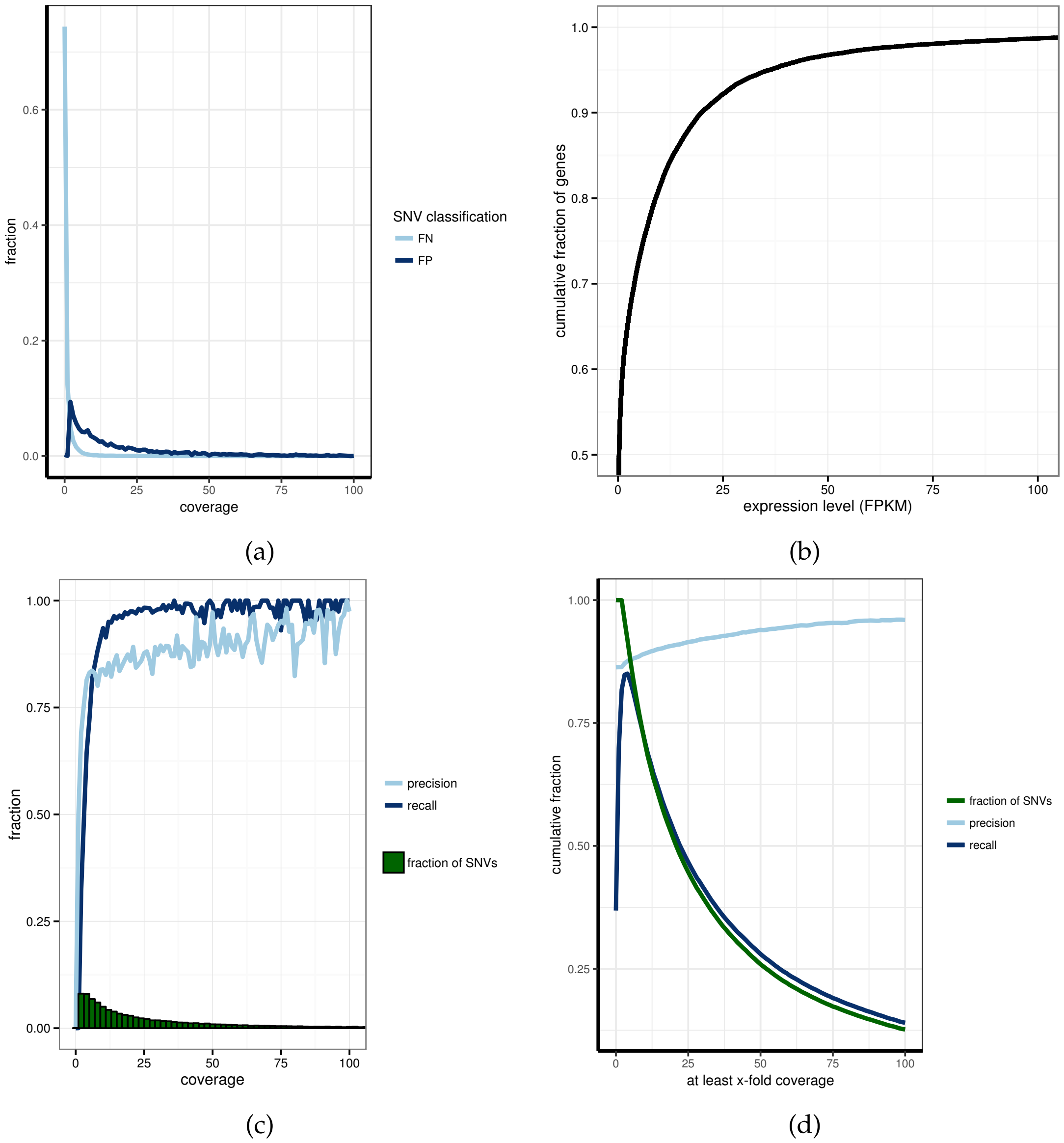


Figure 3.27.: Effect of coverage on SNV calling performance: a) fraction of FP and FN SNVs per read depth b) the cumulative fraction of genes based on expression values (FPKM) c) precision and recall measures as well as number of SNVs per read depth d) precision and recall measures as well as total number of SNVs with at least x-fold coverage

show even zero coverage in the RNA-seq data. The main reason for this is the large number of low- and unexpressed genes (nearly 60% of genes with FPKM $\leqslant 1$; Figure 3.27b). Examining precision and recall values per read depth reveals that both values depend on coverage (Figure 3.27c). Filtering SNVs with read depth lower than 10 improves the recall outcome from 0.27 to 0.71. However, filtering all SNVs with less than 10 supporting reads also removes correct ones (5,557 out of 20,807). Figure 3.27d illustrates the effect of filtering SNVs below a certain coverage in terms of precision and recall and reveals that filtering SNVs with less than 5 supporting reads yields the highest recall measure. On the other hand, the number of FP GATK SNVs does not depend on coverage as much as the FN ones (734 out of 3,292 FP SNVs with coverage < 5; Figure 3.27a) since a high fraction of FP SNVs can be explained by other reasons (see following paragraphs).

RNA editing One cause for the substantial number of FP SNVs is post transcriptional RNA editing. The examination of the mutation profile of the 2,558 FP SNVs with coverage ≥ 5 reveals that more than 80% of the FP SNVs show an adenine to guanine and thymine to cytosine mutation pattern, respectively, which is characteristic for the most common type of RNA editing, A→I[171] (Figure 3.28; Chapter 1.1.1). The number of FP SNVs in the GATK variant set drops from 2,558 to 765 when SNVs at known RNA editing sites are removed. Known sites were downloaded from the Rigorously Annotated Database of A-to-I RNA editing (RADAR)[202]. This database includes about 2.5 million RNA editing sites with about 75% and 20% of them located in intronic and intergenic regions, respectively[202][203]. After removal, there were still 259 remaining SNVs that showed the typical RNA editing mutation pattern which might be indicative that there are still many RNA editing sites in the human genome that have not been discovered yet.

Genotype discordance A further reason for FP SNVs is genotype discordance. Allele-specific expression, meaning that one allele of a gene is preferentially expressed, is common in multiploid organisms and can lead to monoallelic expression where solely the preferred allele is expressed[110]. Hence, this can result in homozygous genotypes in RNA-seq data although heterozygous in the genome. As a consequence, monoallelic expressed genes, where exclusively the alternative allele is expressed, lead to additional FP SNVs. Actually, a substantial amount of the 765 remaining FP SNVs show a high alternative allele frequency (Figure 3.29) and 340 of them are indeed classified as false positive due to differences in the assigned genotypes between gold standard and RNA-seq variant calling data. 313 out of these 340 FP SNVs have a homozygous genotype in the RNA-seq data while heterozygous in the gold standard. In fact, 275 of them show alternative allele frequencies higher than 0.85, which points to true monoallelic expression. The remaining 38 exhibit frequency values between 0.5 and 0.85 but also low coverage ($5 \leq$ coverage ≤ 10) and more than 90% of reads that support the reference allele have low mapping quality with at least one

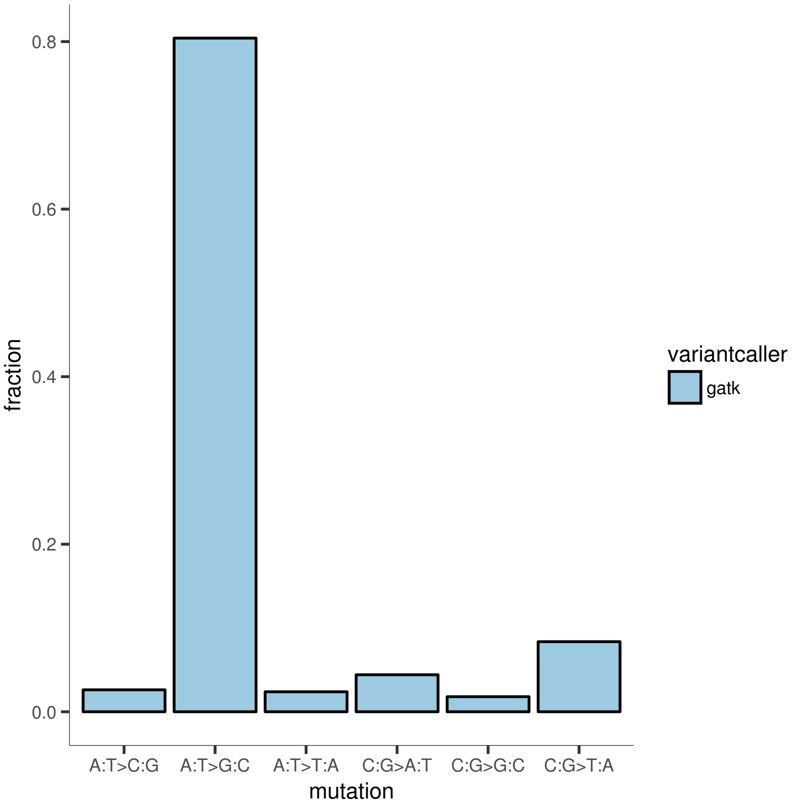
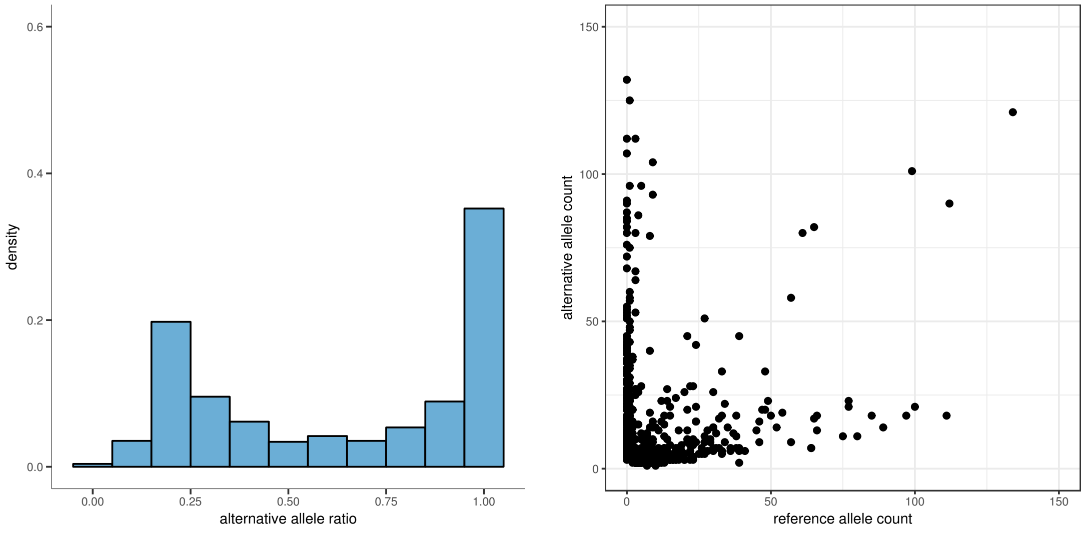


Figure 3.28.: The mutation profile of the subset of false positive variant calls of GATK.

alignment mismatch at another base, thus they are likely true homozygous.

On the other hand, 27 FP SNVs are denoted as homozygous in the gold standard variant set while heterozygous in GATK's RNA-seq variant calling results which is highly likely an indication for incorrectly assigned genotypes in the RNA-seq variant calling results. In fact, 20 of these presumably wrong SNVs show alternative allele frequencies higher than 0.85 and another five have alternative allele frequencies between 0.8 and 0.85, i.e. the vast majority of reads support the alternative allele. One of the remaining two has low coverage (6 supporting reads) and an alternative allele frequency of 0.6̇ where 16% of both of the two reference allele supporting reads are soft clipped even though they are neither located at an exon boundary nor an indel. Using BLAT[103] to look for alternative alignment positions in the human genome for these two reads revealed that the reported alignment positions are the best ones. Nevertheless, these two reads likely did not origin from the reported genomic position, thus probably constitute misalignments and as a consequence hamper the assignment of the correct genotype.

The last one of the misgenotyped SNVs show sufficient coverage (24 supporting reads) but an alternative allele ratio of not more than 0.2. This SNV is located within a



(a)

(b)

Figure 3.29.: Alternative allele ratios of the 765 FP SNVs where known RNA editing sites have already been excluded: a) distribution of alternative allele ratios b) reference versus alternative allele read counts

known Short Interspersed Nuclear Elements (SINE) region in the 5′UTR of the DTD2 gene. Since SINEs are frequently occurring, highly repetitive genomic elements with lengths of less than 500 bp[228] the supporting reads are again likely due to alignment problems and the SNV is probably a true FP in terms of assigned genotype.

Gold Standard Issues and Problematic Regions Downloading whole genome sequencing data of the GM12878 cell line from the European Nucleotide Archive (ENA) with the accession number ERP001229, aligning them and subsequently assessing the allele frequencies at the sites of the residual 284 FP GATK SNVs (coverage$\geq$5; no known RNA editing site and no characteristic RNA editing mutation pattern; no genotype discordance) revealed that for 11 of them there is strong evidence in the whole genome sequencing data (allele frequency $\geqslant 0.2$; coverage $\geqslant 50$) that the variant calls in the RNA-seq data are actually true. For another 52 sites there is at least one read in the WGS data set showing the alternative allele. Moreover, 28 of these 63 SNVs are also present in dbSNP[226] version 142. Additionally, 64 SNVs occur in clusters with at least three SNVs located within a 50 bp window which are indeed likely false positives (Figure 3.30). The residual FP SNVs show low coverage and are mainly caused by misaligned reads in repetitive regions.

Indels

Concerning indel identification, the performance of GATK is comparable to its SNV

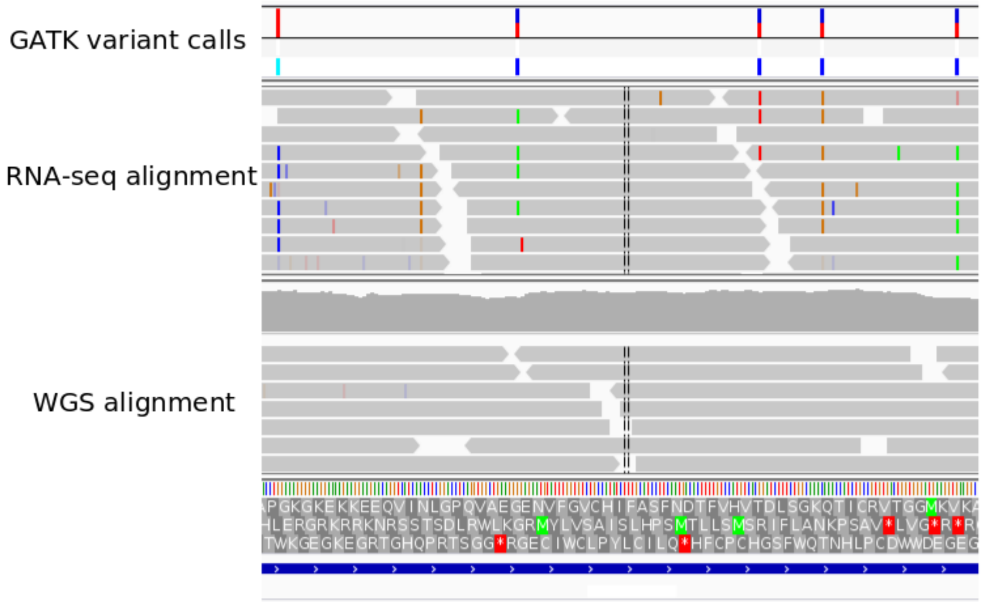


Figure 3.30.: IGV screenshot: Example of likely false positive SNVs that are located in close proximity to each other.

results in terms of precision and recall. Nevertheless, the causes for FP and FN indels are partly different and

i. coverage

ii. mapping problems

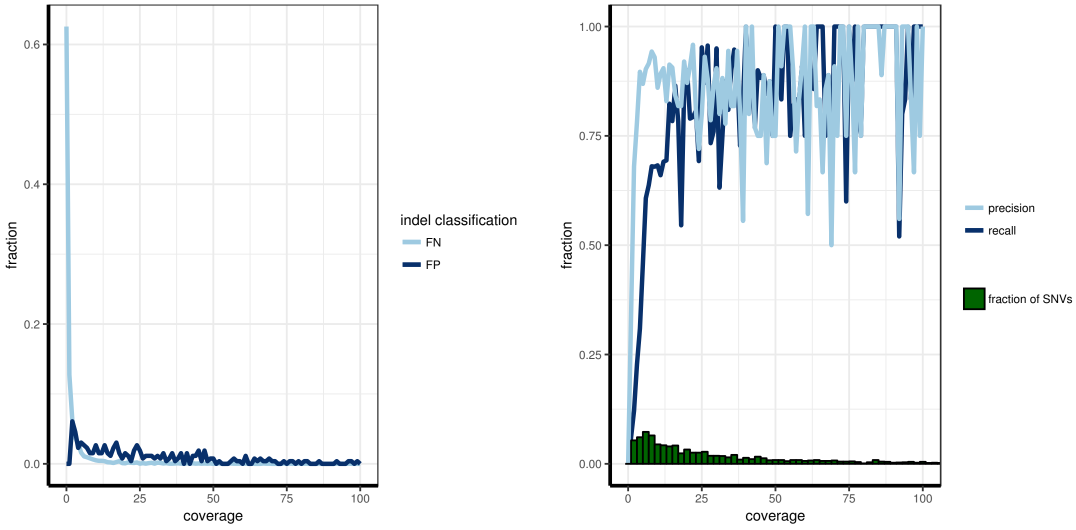
iii. genotype discordance

are discussed in the following, again, with the aim to identify quality and filter criteria for reliable indel calls.

In total, 1,710 of the 5,937 indels that are present in the filtered gold standard variant set could be detected by GATK (TP) while another 267 were called presumably wrong (FP). Altogether, GATK detected 1,038 insertions and 939 deletions where the lengths of them range from 1 to 16 bp in both classes. However, one true FP deletion has a length of 90 bp which is caused by a short intron of exact that size.

Coverage is not only a crucial factor for SNVs but for indels as well. Figure 3.31a illustrates that the majority of FN indels has coverage below 5 and that precision and recall depend both on coverage (Figure 3.31b).

However, not all FN indels can be explained by low coverage. Instead, 244 of the 465 FN indels that exhibit coverage equal to or above 5 can be ascribed to repetitive regions. Another issue concerns indels close to exon boundaries since reads aligned



(a) (b)

Figure 3.31.: Effect of coverage on indel calling: a) fraction of FP and FN indels per read depth b) precision and recall measures as well as total number of indels with at least x-fold coverage

to those positions tend to be soft clipped instead of introducing an actual indel. Furthermore, there are five cases where STAR incorrectly introduces a split read instead of a deletion. 24 FN indels show clear evidence in the RNA-seq data, hence should have been called while 6 do not have any alternative allele supporting reads although sufficient coverage and thus indicating monoallelic expression of the reference allele.

As with SNVs, indels can be involved in monoallelic expression. According to the VCF files, 81 of the 267 FP indels are caused by genotype discordances. Again, some of them, i.e. 26, are homozygous in the gold standard variant set but heterozygous in the RNA-seq results. Verifying these sites with the RNA-seq alignment data illustrates that the majority of reads support the homozygous genotype for all but one of the 26 cases. However, they are all located around repetitive regions. Thus, several reads, especially those which do not span the repeat, mistakenly support the heterozygous genotype. 45 FPs, on the other hand, show a reasonable zygosity pattern in terms of monoallelic expression, that is heterozygous in the gold standard variant set and homozygous in RNA-seq data. Here, 42 could be confirmed as indeed homozygous in the RNA-seq alignment data while another three have only $\leqslant 75\%$ of reads supporting the indel but also low read depth (coverage < 5).

Furthermore, repetitive regions and particularly homopolymers are not only an issue for genotype discordance but also contribute to variant calls that are actually not present in the data. More precisely, 257 of the 267 FP indels detected by GATK overlap with known repetitive regions downloaded from the UCSC table browser[9] or are located within homopolymers $\geqslant 4$ bases, suggesting to be especially careful with the interpretation of indels in such regions.

3.1.7.2. Variant Filtering

Based on the insights of the previous chapter, several criteria were defined for the purpose of obtaining high confident variant calls. Thus, the following filters can be successively applied to the GATK output:

- use only variants located within annotated exons
- removal of known RNA editing sites[10] (except RNA editing is of interest)
- SNVs as well as indels with less then 5-fold coverage
- indels with lengths longer than read length minus 20 bp
- variants located within known repetitive regions[11] or in homopolymers $\geqslant 4$ bases
- all variants where at least three occur within a 50 bp window (adapted from GATK Best Practices workflow for single-nucleotide polymorphism (SNP) and indel calling on RNA-seq data[68])

Applying these filters on the RNA-seq data used in the previous chapter results in higher precision but on the other hand reduced recall measures for SNVs and indels (Table 3.5). The higher precision shows that most of the false positive variants could be removed with the strict filters. On the other hand, numerous true positives are discarded as well. Thus, these filters are optional and can be applied if solely high confident variants are desired.

	SNV	Indel
total	9,636	558
precision	0.99	0.98
recall	0.17	0.09

Table 3.5.: Performance of GATK variant calling plus custom filters based on the comparison with the refined gold standard variant set.

[9] `https://genome.ucsc.edu/cgi-bin/hgTables` (last accessed 01.10.2016)

[10] RADAR[202] dataset downloaded from `http://rnaedit.com` (last accessed 01.10.2016)

[11] downloaded from `https://genome.ucsc.edu/cgi-bin/hgTables` (last accessed 01.10.2016)

3.1.7.3. Comparison to Whole Exome and Whole Genome Sequencing Data

In order to test the behavior of GATK together with the defined filters on in-house sequenced RNA-seq data, RNA-seq variant calling results were compared to whole exome and whole genome sequencing data. Using GATK as described in Chapter 2.3.5 and subsequently applying the strict filters discussed in the previous chapter, variant calling was performed for 74 human fibroblast in-house RNA-seq samples. For all of them matching WES variant calling data were available. These WES libraries were consistently prepared using the same capture kit, i.e. Agilent SureSelect 50Mb v5, in order to prevent any region-related bias in the follow-up analysis. Additionally, 14 of the 74 samples have WGS on top of the WES data available. Together, they were used to assess the results of the implemented variant detection strategy. Again, hap.py was utilized to identify TP, FP and FN variants.

Comparison to Whole Exome Sequencing Data

The limited overlap of covered regions (37 Mb) is a problem when comparing RNA-seq with WES data since the enrichment kit region of the WES data covers a lot of intronic and some intergenic regions while not covering the majority of UTRs. Thus, both kinds of data sets, RNA-seq as well as WES, were restricted to variants that are located first, within annotated exons of the UCSC gene annotation and second, within the Agilent SureSelect 50Mb v5 kit target region. After limiting to this region and applying the filters defined in the previous chapter, on average only 6,016 variants were reported for the 74 RNA-seq samples. The comparison to their matching WES data yields precision values consistently above 0.99 and 0.97 for SNVs and indels, respectively (Table 3.6). Recall measures, on the other hand, are again low. Of course, the strict filtering criteria are limiting factors in terms of recall but also the big fraction of low- or unexpressed genes has a considerable impact (Figure 3.32).

	¬ custom filtering		custom filtering	
	SNV	Indel	SNV	Indel
total	12,993	760	5,944	72
precision	0.93	0.36	0.99	0.97
recall	0.47	0.31	0.25	0.09

Table 3.6.: Average performance of GATK variant calling on 74 RNA-seq samples with and without applied custom filters based on the comparison with matching whole exome sequencing data and restricted to annotated exons and the Agilent SureSelect 50Mb v5 kit target region.

Comparison to Whole Genome Sequencing Data

For the comparison with the 14 WGS data sets no preceding region-based restrictions were applied. In this way, 12,720 SNVs and 700 indels remain on average for the 14

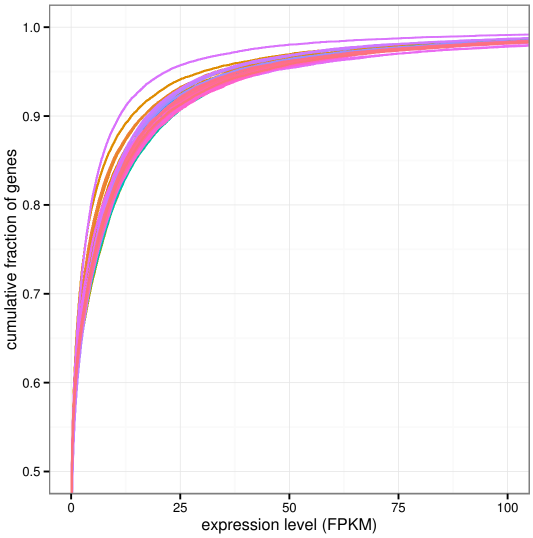


Figure 3.32.: Cumulative fraction of genes based on expression values (FPKM)

matching RNA-seq samples resulting again in high precision measures of 0.99 and 0.95 for SNVs and indels, respectively (Table 3.7). In contrast, recall values are below 0.005. Certainly, the substantial amount of FNs in this case is mainly due to the much larger region covered by WGS data and the consequently much higher number of about 3 million variants per sample. Annotation of the FN variants shows that the majority of them is located in intergenic (57%) and intronic (42%) regions while most of the RNA-seq mutations are detected at UTR (38%), coding (36%) and intronic (19%) loci (Figure 3.33).

	SNV	Indel
total	12,720	700
precision	0.99	0.95
recall	0.004	0.002

Table 3.7.: Average performance of GATK variant calling on 14 RNA-seq samples with applied custom filters based on the comparison with matching whole genome sequencing data.

Variant Characteristics

Examining solely the annotated functional classes of coding variants in both data sets reveals that the average amount of loss-of-function mutations is significantly enriched in WGS data (2.15%) compared to RNA-seq data (0.43%) (χ^2-test, p-value $< 3.9 * 10^{-18}$) (Figure 3.33). Furthermore, also the proportion of missense mutations

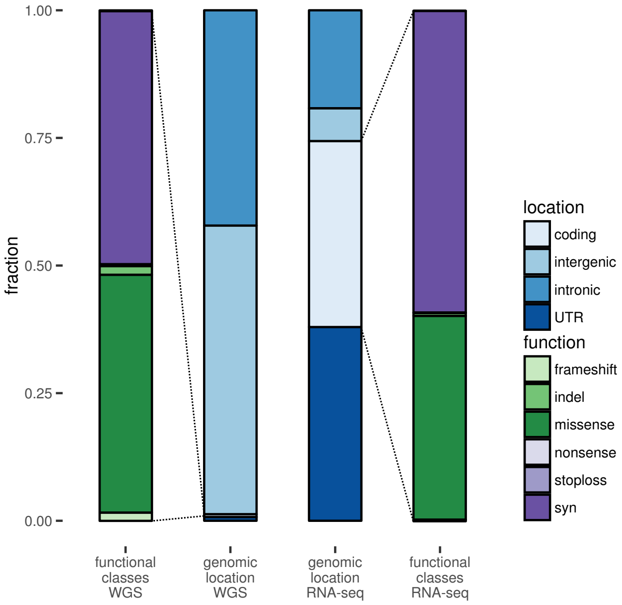
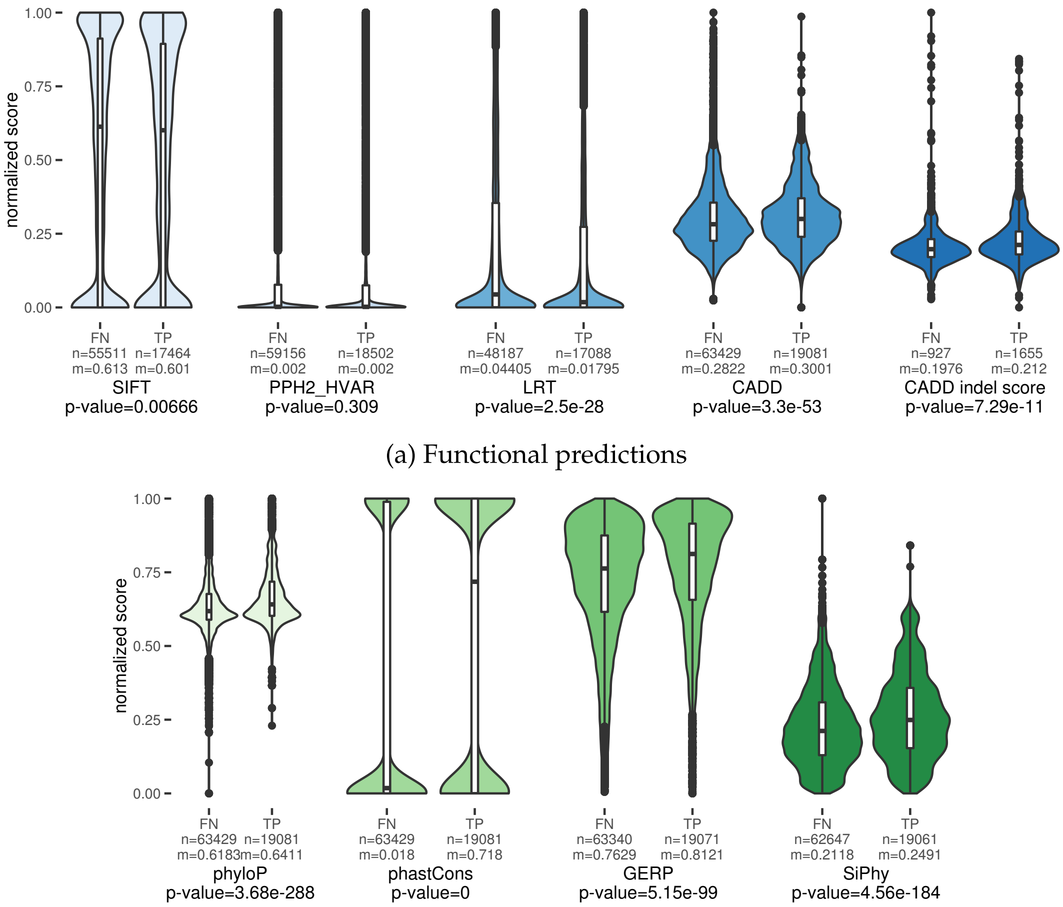


Figure 3.33.: This figure displays the variant characteristics of the high confidence variant calls in terms of where in the genome they are located (inner bars) and which functional class the coding variants are assigned to (outer bars).

is notably higher in the WGS (47%) versus the RNA-seq variant sets (40%) (χ^2-test, p-value $= 1.16 * 10^{-14}$).

Hypothesizing that deleterious mutations might have an adverse effect on gene expression one would expect that such mutations are less frequently detected in RNA-seq data. To check this, variants were annotated with conservation as well as functional prediction scores which were downloaded from the Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations (dbNSFP)[135][136]. In short, the underlying idea of conservation score is that more conserved sites are more important in terms of function than less conserved ones. On top of conservation, functional prediction scores encompass additional information to predict the harmfulness of a mutation at a specific loci. Depending on the algorithm, these information can be population-based allele frequencies or the predicted functional effect on the respective protein[109]. Annotated variants were divided into two groups where group one (TP) comprises all variants which could be detected in both data sets, WGS as well as matching RNA-seq, and group two (FN) the ones that were detected solely in the WGS data. In order to avoid any bias from possibly less important intergenic regions, only variants located within annotated exons were considered. However, no clear trend of enrichment of more deleteriously predicted variants in the TN group could be observed (Figure 3.34a). In fact, there is a gen-



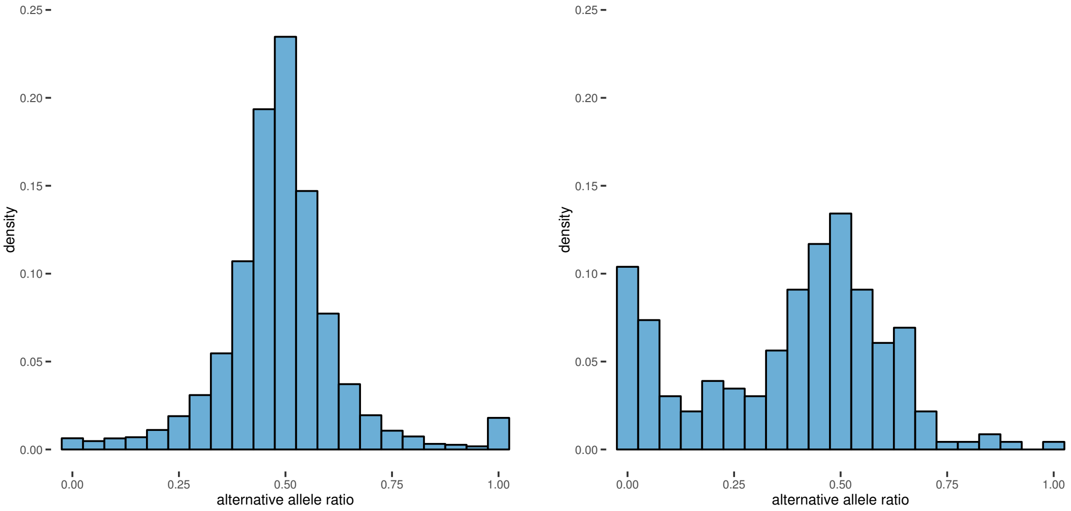
(a) Functional predictions

(b) Conservation scores

Figure 3.34.: Subset of prediction (a) and conservation scores (b) provided by dbNSFP: The depicted results are based on the pair-wise comparison of the 14 RNA-seq - WGS sample pairs and grouped by overlapping (TP) and WGS exclusively (FN) variants located within annotated exons. Indel scores (top right) were calculated with the CADD's online scoring tools[109]. All scores were normalized to fit the common scale between 0 and 1 where 0 is the least and 1 the most extreme value for both kinds of scores. n=total number of variants in the respective group where scores were present; m=median normalized score; p-values were calculated with Mann-Whitney U test (two sided); (Idea for this plot is based on [109] and [256])

eral disagreement between different prediction algorithms which was also perceived by Wieland, 2015[256]. Conservation scores, on the other hand, are more consistent and favor TP mutations (Figure 3.34b). When evaluating the relationship between the genome-wide coverage of each single base of the 14 RNA-seq samples with the corresponding conservation score no real correlation can be observed. In fact, two conservation scores were utilized and resulted in a Spearman correlation coefficient of 0.12 and 0.15 for phyloP and phastCons, respectively.

Apart from that, functional effects can be observed as well. Nonsense mediated decay, where due to nonsense mutations corrupted transcripts are degraded soon after transcription, could be quantified and described in the course of the GEUVADIS project[119]. The same effect can be detected in in-house data too. Comparing alternative allele frequencies between heterozygous synonymous and heterozygous nonsense variants reveals that while alternative allele ratios of heterozygous synonymous variants follow a normal distribution with the mode of the distribution around 0.5, multiple heterozygous nonsense mutations show a notable decrease of the alternative allele. Thus, an increased loss of the variant allele for nonsense mutations can be observed, which is indicative for nonsense mediated decay[119] (Figure 3.35).



(a) heterozygous synonymous variants

(b) heterozygous nonsense variants

Figure 3.35.: Alternative allele ratios for heterozygous synonymous (a) and heterozygous nonsense variants (b) of 74 RNA-seq samples

3.1.8. Merging with Preexisting Architecture

Each of the abovementioned analysis steps is implemented in a modular way meaning that they can be combined, exchanged and in further consequence executed in a dynamic fashion. In order to avoid the maintenance of two parallel analysis systems the RNA-seq analysis modules were integrated in the already mentioned, preexisting in-house exome sequencing data analysis pipeline infrastructure rather than constituting a stand-alone system. This entails several advantages like consistent usability and reduced hands-on time which is vital since the pipeline is used in a core-facility-like environment where more than two terabase of sequence can be produced per week and the analysis of the resulting output should be fast and easy. Thanks to the combined implementation, the analysis pipeline can be started for all samples of an entire flow cell with a single, consistent command, regardless of which libraries, i.e. whole exome sequencing or RNA-seq, were on the flow cell. The pipeline collects all necessary meta-information from the database and starts the analysis for each sample accordingly. If no additional parameters are provided, default analysis is performed for all RNA-seq samples which involves alignment, quality checks, expression quantification and FPKM value calculation. However, the user can modify the pipeline's behavior by passing various parameters to the initiation command.

In order to perform differential gene expression analysis or differential exon usage analysis an intermediate step is necessary where the user has to specify which groups of samples should be compared to each other. Furthermore, a separate command has to be used to issue this kind of analysis.

3.2. Fundamental Considerations for RNA-seq Data Analysis

In order to ensure proper RNA-seq data analysis, the data included in each experiment should feature certain characteristics. Thus, before the start of an RNA-seq experiment particular questions regarding study design and data quality must be addressed and criteria for them were determined using RNA-seq data which were sequenced in the local environment and analyzed as described in the previous chapters. In the following, pivotal considerations regarding

i. sequencing depth

ii. number of replicates

iii. duplicate reads

iv. insert size

are discussed.

3.2.1. Sequencing Depth

An important yet difficult to define quality is how much sequence should be produced for each sample since this number depends on the purpose of the data, i.e. will it be used for differential expression analysis and/or variant detection, etc., and in further consequence on the biological question that should be answered. Furthermore, samples might differ in terms of number of expressed genes and overall expression patterns depending on the tissue the sample originated from and the underlying condition. Figure 3.36 illustrates the number of detectable protein coding genes (FPKM value > 0.5; based on UCSC reference annotation) for 565 human samples that were analyzed with the presented pipeline and where tissue information was provided. On

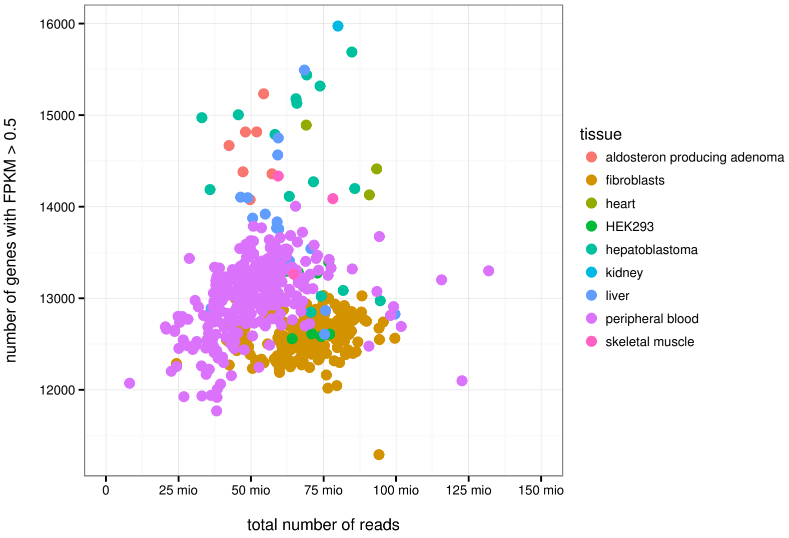
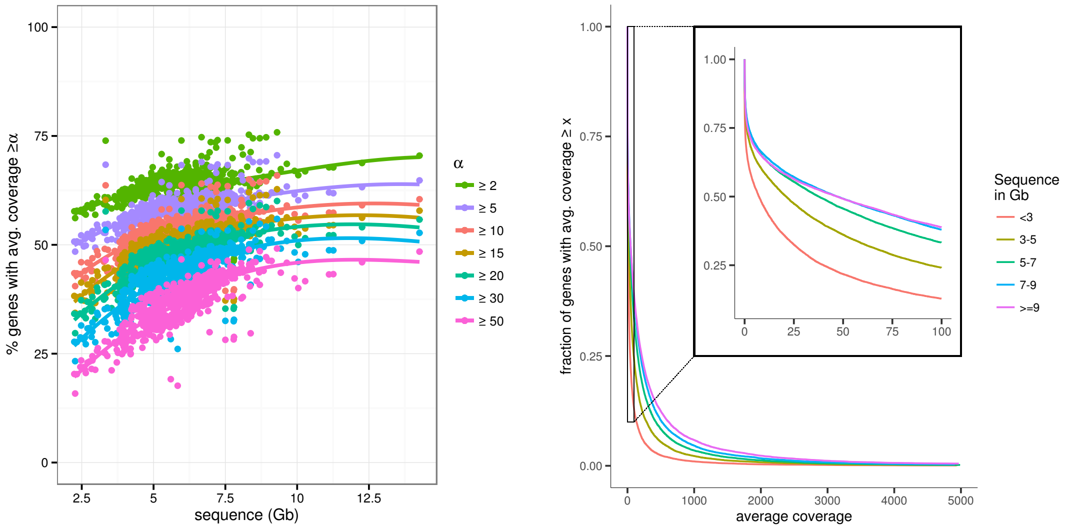


Figure 3.36.: Number of genes per sample with FPKM values > 0.5 as a function of total number of mapped reads and samples are coloured by tissue type.

average, between 50 and 75 million reads were produced per sample which results in around 13,000 traceable genes for each sample. Based on the UCSC gene annotation this corresponds to about 65% of coding genes which is in good agreement with published results[205].

The number of detected genes depends on the amount of produced sequence (Figure 3.37a). However, this is just the case up to a certain threshold. Like for exome sequencing data[256] there is a saturation phenomenon meaning that above a certain amount of produced sequence additional information does not result in an increased number of detected genes. This point is reached by about 7.5 Gb of sequence which corresponds to roughly 75 million 100 bp reads. This circumstance is also shown in Figure 3.37b. For samples with a total amount of sequence of 5 Gb, and especially 7 Gb upwards, only a minor difference can be observed on the distribution of the fraction of genes as a complementary cumulative distribution function of the average gene coverage. Hence, especially for differential gene expression analysis an outcome of 30 to 40 million 100 bp paired-end reads should be sufficient.

While the number of detectable genes is an essential measure for differential gene expression analysis, an especially important one for variant detection is the number of single bases that are covered by an adequate amount of reads. In Chapter 3.1.7 it



(a)

(b)

Figure 3.37.: Correlation between the amount of sequence and average gene coverage: a) How many percent of genes per sample have average coverage $\geqslant \alpha$. Here, average coverage instead of FPKM was used since FPKM values lack the dimension of amount of sequence and average gene coverage was calculated as $GC * RL/GL$ where GC is the number of reads that map to the gene, RL is the length of the sequenced read in bp and GL is the length of the respective gene. b) Distribution of the fraction of genes with at least x-fold average gene coverage and averaged for samples with specific amounts of sequence.

could be shown that the accuracy of variant detection depends on coverage to a high degree. However, it was also shown that a large amount of genes is lowly- or not expressed at all. In terms of base-wise coverage, and based on the human samples analyzed with the pipeline, this means that on average about 70% of exonic bases are covered at least 1-fold while about 60%, 50% and 40% of bases are covered at least 4-, 8- and 20-fold, respectively (Figure 3.38).

3.2.2. Number of Replicates

How many replicates to include in an experiment is a fundamental consideration for differential expression analysis. Two types of replicates can be included, namely technical and biological. Technical replicates are samples that originated from the same material but are treated independently in the downstream analysis and can be used to alleviate possible variability introduced by technical steps ranging from library preparation right up to sequencing. However, nowadays most experiments have

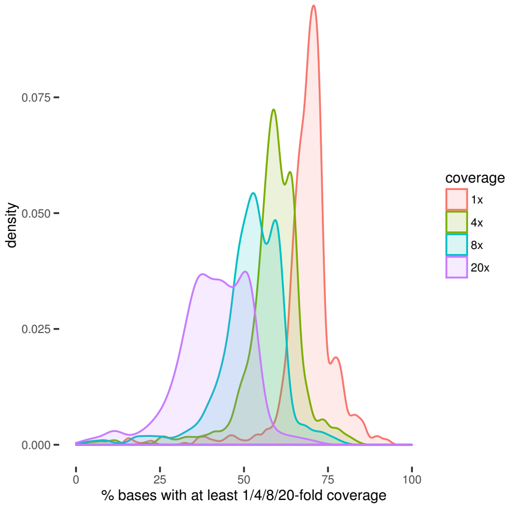


Figure 3.38.: Distributions of percentages of exonic bases per sample that are covered at least 1-, 4-, 8- and 20-fold.

high technical reproducibility which means that technical replicates are dispensable[137][154]. Biological replicates, on the other hand, are samples with the same biological background, e.g. same condition, but from different biological sources. Their purpose is to estimate the variation within each condition group and are neccessary to make generalized inferences about the involved conditions[71][137]. According to the ENCODE Consortium[37] and other published work[35][230][231], at least three biological replicates per condition should be involved in each experiment. Furthermore, they state that the correlation based on gene read count between biological replicates should be high (> 0.9). As could be shown in Chapter 3.1.4.3 (Figure 3.23) the number of replicates heavily influences the number of significantly differentially expressed genes and a clear difference could be observed between the results obtained with three replicates compared to the results yielded with a higher number of replicates. Calculating the $n * (n-1)/2$ intra-group pairwise correlation coefficients for the $n = 11$ tumor and $n = 11$ control samples used in Chapter 3.1.4.3 reveals a high correlation among biological replicates (Figure 3.39). Except for one case, correlation values are consistently above 0.9 with the control group showing higher homogeneity than the tumor group. In fact, the two samples of the pair that result in a Spearman correlation coefficient below 0.9 are responsible for the majority of numbers below 0.93 which is in agreement with the PCA as well as cluster analysis for these samples (Figure 3.20). When excluding these two samples and recalculating the influence of the number of replicates, as was done in Figure 3.23, a slightly increased number of significantly differentially expressed genes (Benjamini–Hochberg adjusted p-value < 0.01) is reported (Figure 3.40). However, similar to the results based on all 11 sample pairs, no saturation in terms of number of differentially ex-

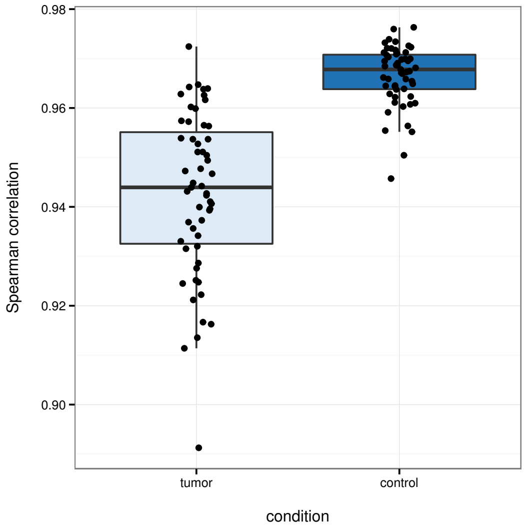


Figure 3.39.: Spearman correlation coefficients of gene expression levels between biological replicates of the 11 tumor-control pairs.

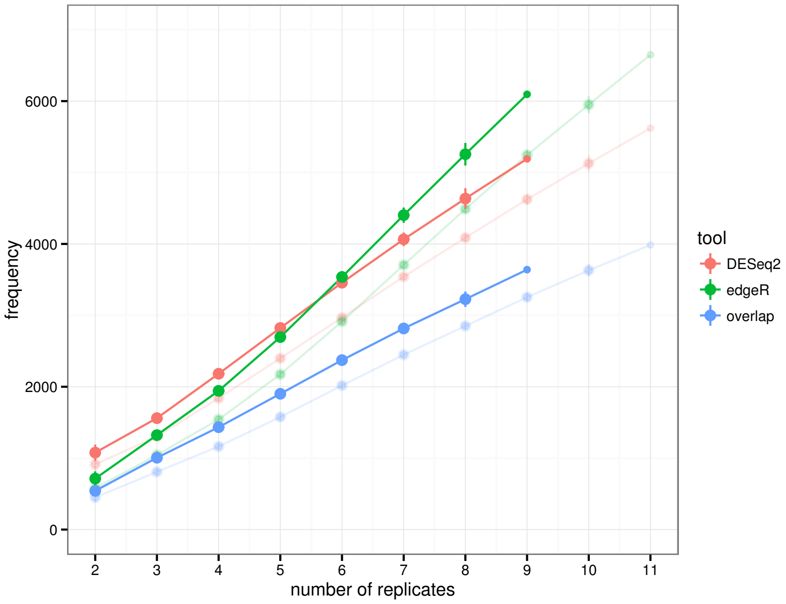


Figure 3.40.: Same graph as in Figure 3.23 (pale colours) but this time with putative outliers excluded (solid colours).

pressed genes can be observed with higher numbers of replicates even if putative outliers are excluded.

3.2.3. Duplicate Reads

The detection of duplicate reads is commonly based on the mapping position of the reads which means that multiple reads that map to exactly the same position are classified as duplicates and if duplicate removal is performed, all but one of them will be discarded. One source of duplicate reads is the PCR amplification step in the course of the library preparation and these reads should be removed as they have the potential to mistakenly amplify a signal which in further consequence might lead to wrong conclusions. However, duplicate reads can also occur by chance if two reads stochastically map to the same position. This can be mitigated by paired-end reads as the information of both pairs is combined for duplicate detection, thus decreasing the probability of identical mapping positions.

For the mouse and human samples analyzed with the pipeline the average duplicate rate is 0.27 (Figure 3.41). This means that if duplicate removal is performed on average about 25% of sequenced reads per sample get discarded.

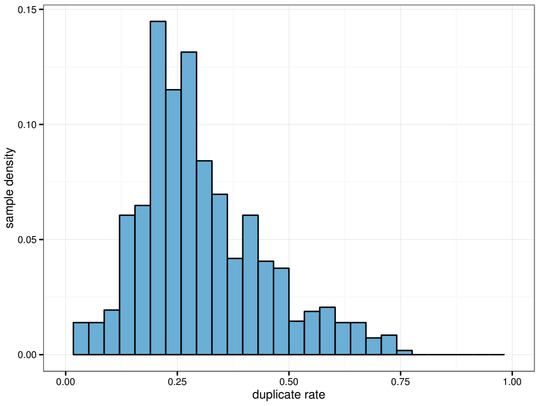
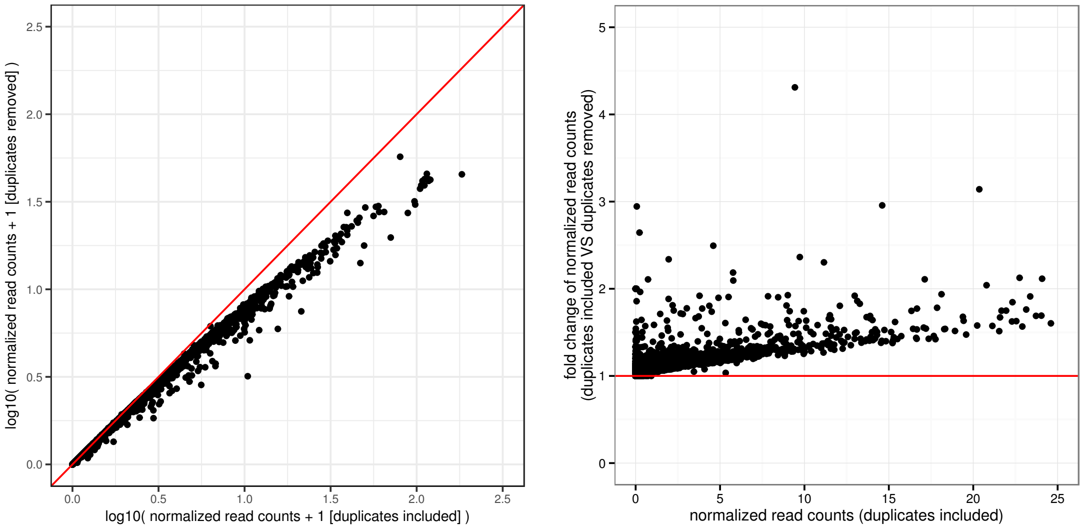


Figure 3.41.: Distribution of duplicate rates among sequenced samples.

For whole exome sequencing data it is common practice to remove duplicate reads before performing any kind of downstream analysis. Doing the same with RNA-seq data preliminary to differential expression analysis is still an ongoing discussion and both strategies were used in different studies[30][119][172][260].

In order to ascertain the best approach for the pipeline gene read counts normalized by respective gene lengths are used since raw read counts do not cover the fact that the total number of unique alignment positions per gene is proportional to the length of the gene. Widely used FPKM values, on the other hand, are additionally normalized by the total number of reads which changes when duplicates are removed. Thus, resulting values might be skewed by very highly expressed genes (see Chapter 1.4.3.1) and genes with little duplicates would receive higher FPKM values in the set with removed duplicates than in the one without. For investigation of the characteristics of duplicate reads, length normalized gene read counts were calculated for an average human RNA-seq sample with nearly 8 Gb of total sequence produced and a duplicate rate of 0.23 for both, gene expression levels with duplicate reads included and excluded. The comparison of them shows that especially higher expressed regions are affected by duplicate reads (Figure 3.42) and a similar pattern can be observed for other samples as well.



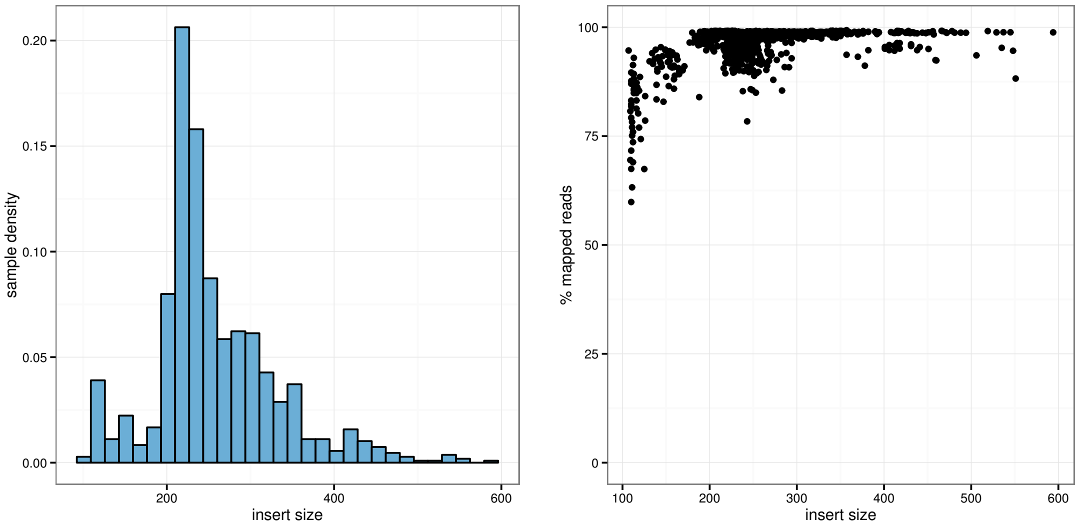
(a) (b)

Figure 3.42.: a) Normalized gene read counts with (x-axis) and without (y-axis) duplicate reads where each dot represents a gene. b) Fold changes per gene based on normalized gene read counts are calculated as d_{ij}/d_{ej} where d_{ij} is the read count with duplicates included for gene j and d_{ej} is the respective read count with duplicates excluded. Subsection from 0 to 25 of normalized and duplicate included read counts are shown.

Based on these findings, the pipeline keeps a BAM file per sample with duplicate reads included but also generates a BAM file with removed duplicates. Thus, for differential expression analysis the file with duplicates is used and for other kinds of analyses, e.g. variant detection, the one without duplicates is usable.

3.2.4. Insert Size

The RNA-seq data analysis pipeline processes almost entirely paired-end reads and the quality of them depends on the length of the sequenced fragments, i.e. the insert size, to a large extent. If the insert size is too small the resulting reads will contain similar read information and, if insert size is very small, adapter sequences might be included as well with both cases likely flawing the resulting alignment. Thus, for 100 bp paired-end reads a sufficient insert size would be above 200 bp[256]. As can be seen in Figure 3.43a the majority of RNA-seq samples processed by the pipeline have insert sizes between 200 and 300 bp. However, a small fraction of samples show lower values and combining insert size with mapping rate reveals that for samples with insert sizes below 200 a substantially lower fraction of reads could be successfully mapped to the reference, underlining the importance of sufficient insert sizes (Figure 3.43b).



(a)

(b)

Figure 3.43.: Insert sizes of sequenced mouse and human samples: a) Distribution of insert sizes across samples b) Correlation of insert size with fraction of mappable reads per sample

3.3. Database

The pre-existing database infrastructure was used and extended by several tables which are depicted in Figure 3.44. The central point of the entire database system is the `sample` table and any generated information, either by the LIMS or subsequently by the analysis pipeline, is linked to that table. This enables on one hand the pipeline to gather all information needed for automated analyses and on the other hand the web interface to display the results in a sample-wise manner.

One of the new tables is `rnaseqcstat`. Here, all generated metrics of the quality control step are stored. Furthermore, two identical databases, one for mouse and one for human, were created, holding nine new tables each. The tables `genebased` and `exonbased` store results produced by the expression quantification steps along with calculated FPKM values. Differential expression analysis findings are inserted into the `deresult` table. Since samples might be used for multiple comparisons

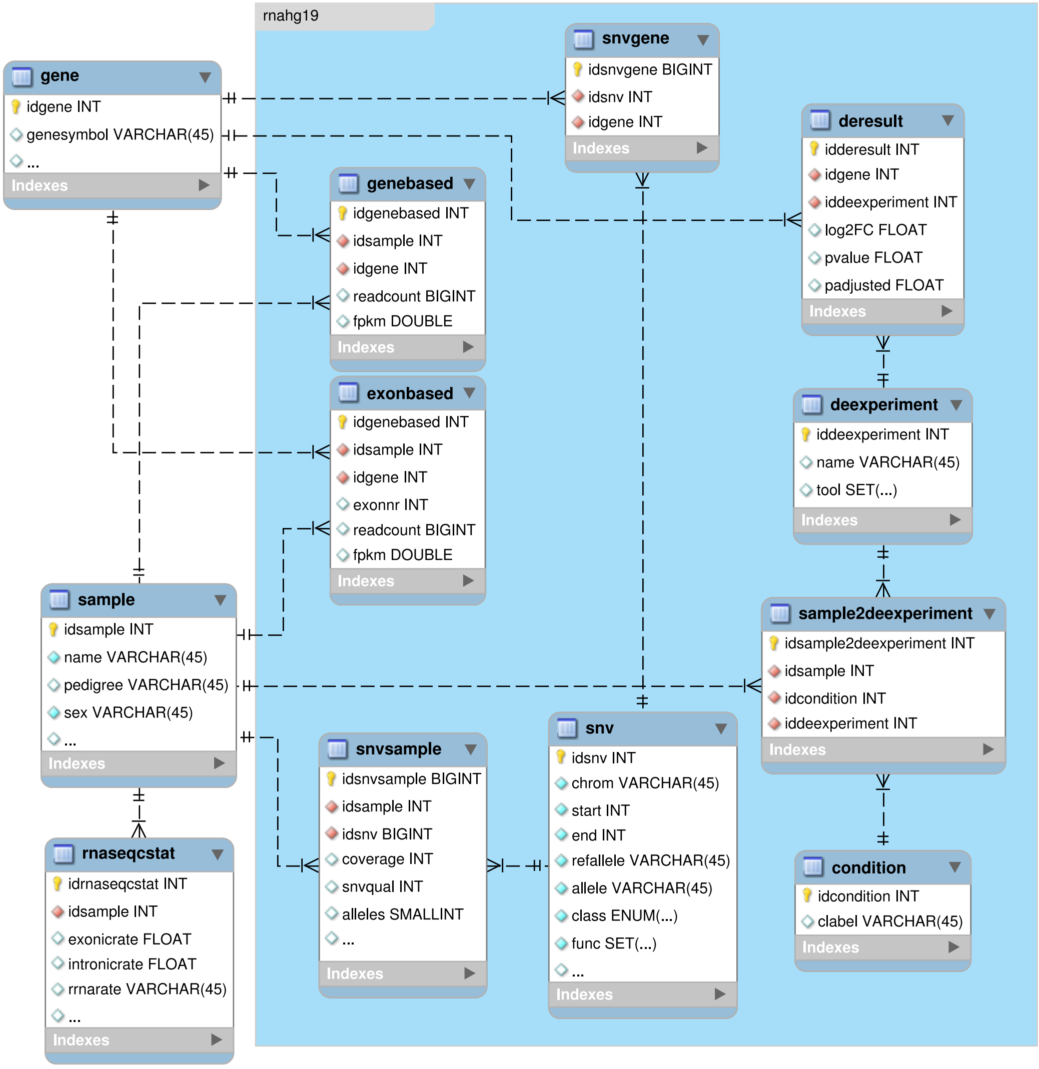
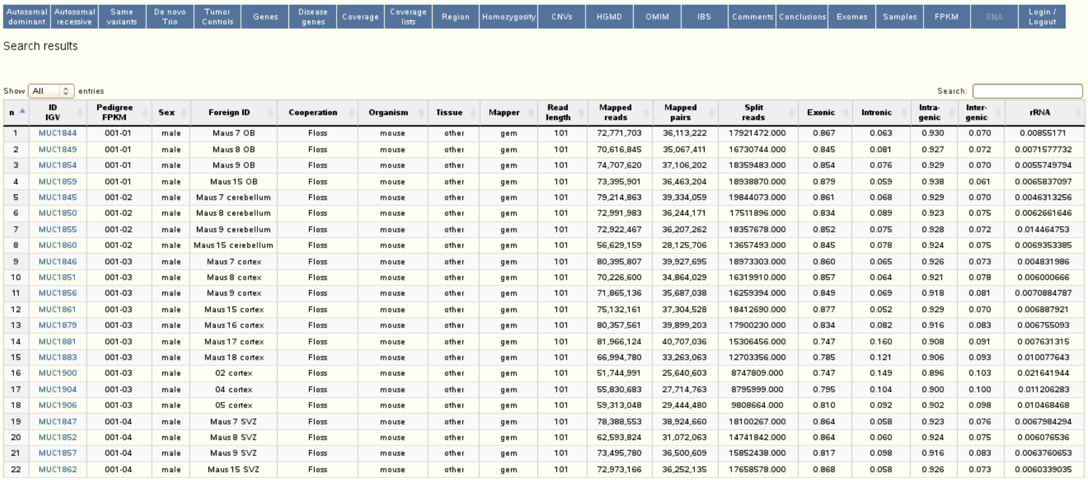


Figure 3.44.: Entity Relationship Diagram of the RNA-seq database for human samples.

the table `deresult` is linked to `deexperiment` which in turn is connected to the cross-reference table `sample2deexperiment`. The latter table assigns samples to conditions and associates them with experiments. Finally, SNVs as well as indels identified in the variant detection step are inserted into the `snv` table where basic information like genomic position, reference and alternative allele, effect of the mutation (e.g. synonymous, missense, frameshift, etc.) or class (e.g. SNV, indel) is stored for each mutation. These entries are then linked with the respective samples via the table `snvsample`. Here, sample specific information like coverage at the variant position, assigned variant quality or zygosity is saved.

3.4. Web Application

In order to enable the users to investigate their RNA-seq data, two new features were added to the web interface so far. First of all, an informative RNA sample overview (Figure 3.45) where not only general sample information like internal and foreign sample ID, organism, tissue or name of the collaborator the sample belongs to, but also important quality metrics such as number of mapped reads, exonic and intronic rate, intra- and intergenic rate or rRNA rate are displayed. Additionally, by clicking on the internal sample ID link the split read alignment of the sample is opened in the Integrative Genomics Viewer (IGV)[213] (Figure 3.46) and can be inspected by the user in a convenient manner.



Search results

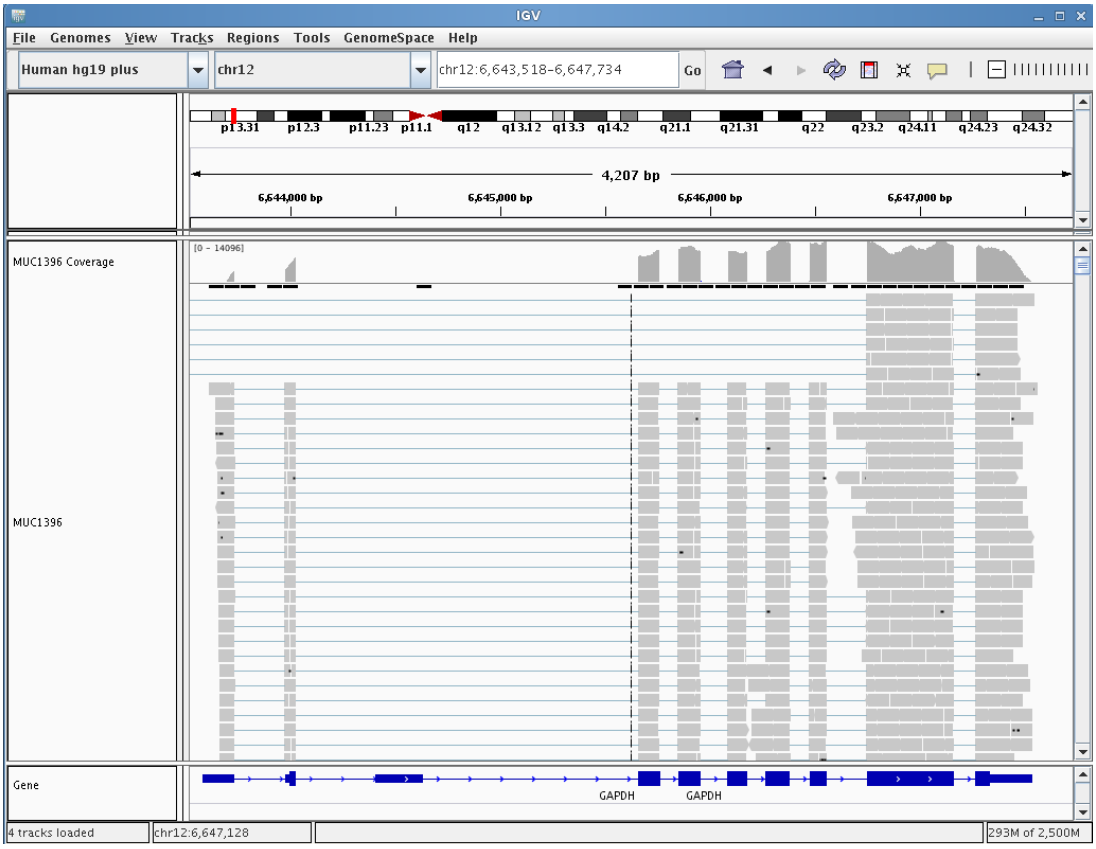
Show All entries

Search:

n	ID IGV	Pedigree FPKM	Sex	Foreign ID	Cooperation	Organism	Tissue	Mapper	Read length	Mapped reads	Mapped pairs	Split reads	Exonic	Intronic	Intra-genic	Inter-genic	rRNA
1	MUC1844	001-01	male	Maus 7 OB	Floss	mouse	other	gem	101	72,771,703	36,113,222	17921472.000	0.867	0.063	0.930	0.070	0.00855171
2	MUC1849	001-01	male	Maus 8 OB	Floss	mouse	other	gem	101	70,616,845	35,067,411	16730744.000	0.845	0.081	0.927	0.072	0.0071577732
3	MUC1854	001-01	male	Maus 9 OB	Floss	mouse	other	gem	101	74,707,620	37,106,202	18359483.000	0.854	0.076	0.929	0.070	0.0055749794
4	MUC1859	001-01	male	Maus 15 OB	Floss	mouse	other	gem	101	73,395,901	36,463,204	18938870.000	0.879	0.059	0.938	0.061	0.0065837097
5	MUC1845	001-02	male	Maus 7 cerebellum	Floss	mouse	other	gem	101	79,214,863	39,334,059	19844073.000	0.861	0.068	0.929	0.070	0.0046313256
6	MUC1850	001-02	male	Maus 8 cerebellum	Floss	mouse	other	gem	101	72,991,983	36,244,171	17511896.000	0.834	0.089	0.923	0.075	0.0062661646
7	MUC1855	001-02	male	Maus 9 cerebellum	Floss	mouse	other	gem	101	72,922,467	36,207,262	18357678.000	0.852	0.075	0.928	0.072	0.014464753
8	MUC1860	001-02	male	Maus 15 cerebellum	Floss	mouse	other	gem	101	56,629,159	28,125,706	13657493.000	0.845	0.078	0.924	0.075	0.0069353385
9	MUC1846	001-03	male	Maus 7 cortex	Floss	mouse	other	gem	101	80,395,807	39,927,695	18973303.000	0.860	0.065	0.926	0.073	0.004831986
10	MUC1851	001-03	male	Maus 8 cortex	Floss	mouse	other	gem	101	70,226,600	34,864,029	16319910.000	0.857	0.064	0.921	0.078	0.006000666
11	MUC1856	001-03	male	Maus 9 cortex	Floss	mouse	other	gem	101	71,865,136	35,687,038	16259394.000	0.849	0.069	0.918	0.081	0.0070884787
12	MUC1861	001-03	male	Maus 15 cortex	Floss	mouse	other	gem	101	75,132,161	37,304,528	18412690.000	0.877	0.052	0.929	0.070	0.006887921
13	MUC1879	001-03	male	Maus 16 cortex	Floss	mouse	other	gem	101	80,357,561	39,899,203	17900230.000	0.834	0.082	0.916	0.083	0.006755093
14	MUC1881	001-03	male	Maus 17 cortex	Floss	mouse	other	gem	101	81,966,124	40,707,036	15306456.000	0.747	0.160	0.908	0.091	0.007631315
15	MUC1883	001-03	male	Maus 18 cortex	Floss	mouse	other	gem	101	66,994,780	33,263,063	12703356.000	0.785	0.121	0.906	0.093	0.010077643
16	MUC1900	001-03	male	02 cortex	Floss	mouse	other	gem	101	51,744,991	25,640,603	8747809.000	0.747	0.149	0.896	0.103	0.021641944
17	MUC1904	001-03	male	04 cortex	Floss	mouse	other	gem	101	55,830,683	27,714,763	8795999.000	0.795	0.104	0.900	0.100	0.011206283
18	MUC1906	001-03	male	05 cortex	Floss	mouse	other	gem	101	59,313,048	29,444,480	9808664.000	0.810	0.092	0.902	0.098	0.010468468
19	MUC1847	001-04	male	Maus 7 SVZ	Floss	mouse	other	gem	101	78,388,553	38,924,660	18100267.000	0.864	0.058	0.923	0.076	0.0067984294
20	MUC1852	001-04	male	Maus 8 SVZ	Floss	mouse	other	gem	101	62,593,824	31,072,063	14741842.000	0.864	0.060	0.924	0.075	0.006076536
21	MUC1857	001-04	male	Maus 9 SVZ	Floss	mouse	other	gem	101	73,495,780	36,500,609	15852438.000	0.817	0.098	0.916	0.083	0.0063760653
22	MUC1862	001-04	male	Maus 15 SVZ	Floss	mouse	other	gem	101	72,973,166	36,252,135	17658578.000	0.868	0.058	0.926	0.073	0.0060339035

Figure 3.45.: A screenshot of the RNA sample overview component of the web interface. Information is presented in tabular form with one sample per row.

The second feature allows users to investigate read counts and FPKM values of all or just particular genes of their samples (Figure 3.47).


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Go
p13.31 p12.3 p11.23 p11.1 q12 q13.12 q13.3 q14.2 q21.1 q21.31 q22 q23.2 q24.11 q24.23 q24.32
4,207 bp
6,644,000 bp
6,645,000 bp
6,646,000 bp
6,647,000 bp
MUC1396 Coverage
[0 - 14096]
MUC1396
Gene
GAPDH
GAPDH

Human hg19 plus

chr12

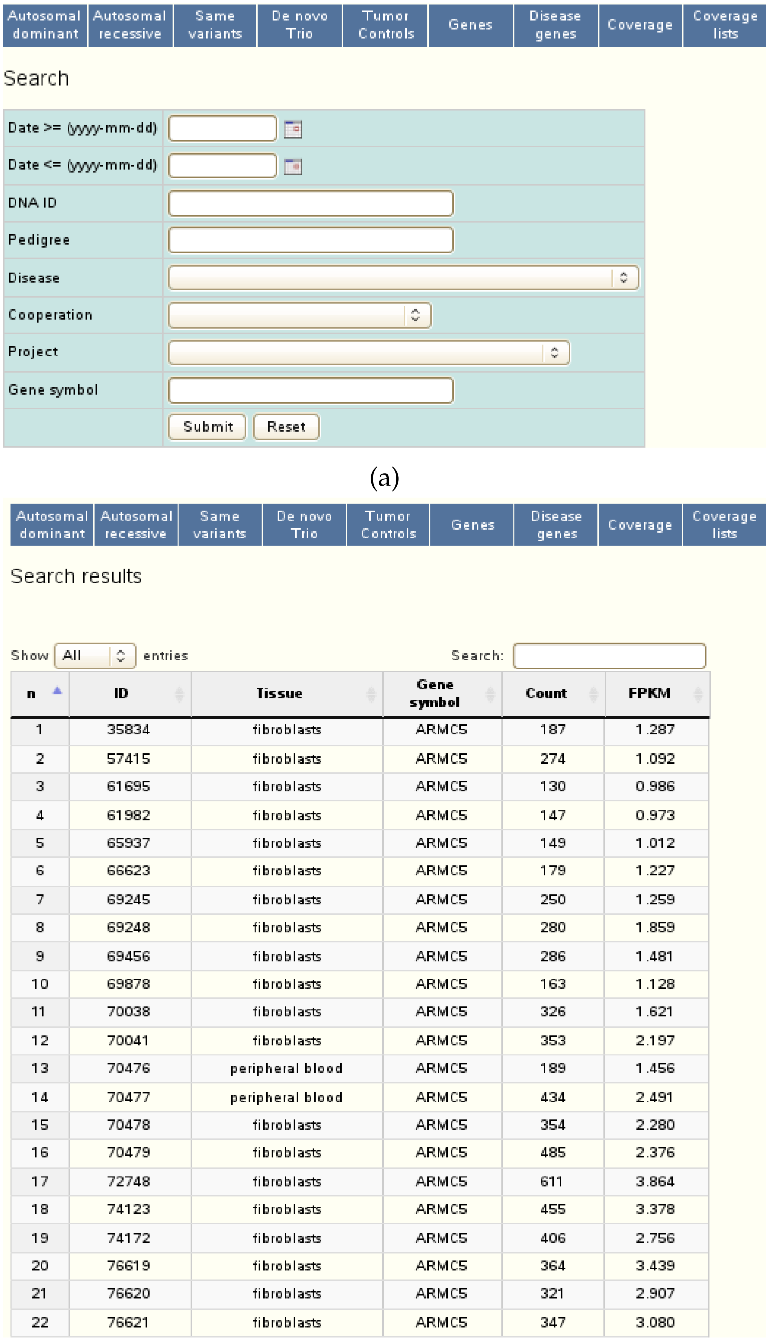
chr12:6,643,518-6,647,734

4 tracks loaded

chr12:6,647,128

293M of 2,500M

Figure 3.46.: A screenshot of the IGV showing read alignments to gene GAPDH. Grey bars represent the actual reads and blue lines in between indicate that the read spans an intron.



Search

Date >= (yyyy-mm-dd)	
Date <= (yyyy-mm-dd)	
DNA ID	
Pedigree	
Disease	
Cooperation	
Project	
Gene symbol	
	Submit Reset

(a)

Search results

Show All entries Search:

n	ID	Tissue	Gene symbol	Count	FPKM
1	35834	fibroblasts	ARMC5	187	1.287
2	57415	fibroblasts	ARMC5	274	1.092
3	61695	fibroblasts	ARMC5	130	0.986
4	61982	fibroblasts	ARMC5	147	0.973
5	65937	fibroblasts	ARMC5	149	1.012
6	66623	fibroblasts	ARMC5	179	1.227
7	69245	fibroblasts	ARMC5	250	1.259
8	69248	fibroblasts	ARMC5	280	1.859
9	69456	fibroblasts	ARMC5	286	1.481
10	69878	fibroblasts	ARMC5	163	1.128
11	70038	fibroblasts	ARMC5	326	1.621
12	70041	fibroblasts	ARMC5	353	2.197
13	70476	peripheral blood	ARMC5	189	1.456
14	70477	peripheral blood	ARMC5	434	2.491
15	70478	fibroblasts	ARMC5	354	2.280
16	70479	fibroblasts	ARMC5	485	2.376
17	72748	fibroblasts	ARMC5	611	3.864
18	74123	fibroblasts	ARMC5	455	3.378
19	74172	fibroblasts	ARMC5	406	2.756
20	76619	fibroblasts	ARMC5	364	3.439
21	76620	fibroblasts	ARMC5	321	2.907
22	76621	fibroblasts	ARMC5	347	3.080

(b)

Figure 3.47.: a) A screenshot of the web interface showing the search formular for the count search. b) The resulting page provides read counts and respective FPKM values, here for 22 different samples of a random project for gene ARMC5. Again, information is presented in tabular form.

Part IV.

Discussion

4. Discussion

RNA-seq has become a widely used technique for studying the transcriptome and offers a wide variety of applications. Thanks to advanced sequencing technology, multiple samples can be processed in parallel and depending on the desired sequencing depth, up to more than 100 RNA-seq samples can be sequenced on a current Illumina instrument within a few days resulting in several hundred gigabyte of data. Processing this kind of data in an efficient and systematic way is vital. In order to meet these requirements, an RNA-seq data analysis pipeline was implemented. Several publicly available pipelines existed before or emerged during this PhD project[41][44][46][65] [78][86][96][138][208][245][261]. However, they were not suitable mainly for the following reasons: First of all, the majority of them constitute a closed system where only some of the needed analysis steps are covered and an adaption, and especially the integration into the existing IT infrastructure of them was not feasible. Second, some published pipelines are web-based, meaning that the data under investigation should have been uploaded to a web server which poses a data security risk. This is particularly an issue since a lot of the 1,500 samples are of human origin. And third, pipelines only executable via a graphical user interface are not suitable as they are difficult or simply impossible to execute in a parallel manner. However, the sequencing instruments produce a lot of data per week and the available computer resources have to be utilized as efficiently as possible, thus parallelization of analysis steps is important. Furthermore, the already existing whole exome sequencing data analysis pipeline and the established infrastructure as a consequence thereof encouraged the design and development of a custom pipeline which is specially tailored to and dovetailed with the pre-existing system and which provides fast and convenient yet reliable and tailored RNA-seq data analysis. To satisfy these requirements, publicly available tools as well as custom scripts are used and the decision which tool to use for each step was based on comparisons of competing software. In order to further improve the performance of the selected tools, methods and filtering criteria for implemented analysis steps were defined. Additionally, default parameter settings of individual tools were changed if proved to produce improved results. Investigation of different properties of the input data material, on the other hand, helped to define key input data quality and study design requirements which are adopted for new RNA-seq projects.

4.1. Components of the RNA-seq Data Analysis Pipeline

Gene annotation The pipeline is designed to process samples from well studied organisms for which a reference genome as well as a gene annotation exist. By the time of writing, two default organisms, i.e. mouse and human, with all necessary reference information are covered with version NCBI37/mm9 and GRCh37/hg19, respectively. These versions are still in use since they constituted the most established genome assemblies for their respective organisms for several years with reliable genome annotations and a comprehensive collection of supportive data. Nevertheless, newer and already widely used versions for both organisms, i.e. GRCm38/mm10 and GRChg38/hg38, are available and an upgrade for the pipeline is planned for the near future. Thanks to the flexible implementation it is easy to add newer versions, entirely new organisms or even custom references. If it is desired to save the results of a newly added organism to the database, all required tables have to be created manually which can be done with available SQL scripts. Furthermore, for mouse and human in particular, there are usually a standard reference genome but several different gene annotations available which differ in terms of comprehensiveness and the choice of which one to use can influence the outcome of the analysis considerably[237][268].

Three prominent gene annotations (GENCODE, RefSeq and UCSC) were compared to each other in order to support the choice of which one to use. To some extent, Harrow *et al.*, 2012[82] performed a comparable analysis when they released the human genome annotation GENCODE version 7. Here, not only a later release (v11; see Chapter 2.2) but also mouse annotations were investigated. The here presented results proved that GENCODE is the most comprehensive one followed by UCSC and finally RefSeq for both, mouse as well as human[81]. However, GENCODE comprises a lot of non-coding transcripts and the agreement between all three sets regarding coding genes is high. Furthermore, alignment of 10 random samples using all three annotations separately revealed that at least GENCODE and UCSC show a comparably high fraction of total reads that could be mapped to annotated regions which is indicative that the additional transcripts covered by GENCODE are dispensable for the requirements of the pipeline. This was also concluded by Wu *et al.*, 2013[262], who solely examined the effect of different human gene annotations on expression quantification results, and stated that the supplementary transcript of more complex annotations belong to lowly- or unexpressed genes. These results, and the fact that the pre-existing exome data analysis pipeline uses it, led to the utilization of the UCSC gene annotation.

Alignment Split read alignment is a fundamental step in RNA-seq data analysis since every single kind of the downstream analysis depends on its results. Thus, the choice of a reliable alignment software is crucial when analyzing RNA-seq data. Initially, GEM was the only aligner implemented in the pipeline since it was used for the GEUVADIS project[119]. Later, two additional and widely used alignment tools were added, namely STAR and TopHat2. In order to find out which one should be the first choice all three of them together with a further development of BWA, BWA-MEM, which is also able to introduce larger gaps to the alignment[123], were compared to each other (Chapter 3.1.2).

In general, all of them have their strengths and weaknesses. STAR, for example, reported the highest number of unambiguously mapped reads which is especially important for expression quantification as different quantification tools have different strategies how to deal with ambiguous reads where they either discard them, distribute them equally or assign them randomly to one of the reported positions. Comparable results could be shown by Engström *et al.*, 2013[58] who also evaluated split read aligner but with shorter reads, a partially different collection of tools and older versions of them, which is also the case for other published aligner comparisons of this type[74][107][132]. Moreover, it was important to assess the performance of up to date tools on the present infrastructure and on data with similar properties to the ones generated by our sequencing facility. A first and particularly essential benchmark was mapping speed. Although Dobin *et al.*, 2013[50] stated that their STAR aligner outperforms other mapping tools, including TopHat2, in terms of mapping speed by more than 50 times, such a big difference could not be observed in the here used setup. However, they used six and 12 threads in their comparison, respectively, compared to just one used here, where STAR was still several times faster than the other three tools. The main reason for the increased speed, according to Dobin *et al.*, 2013[50], is the utilization of an uncompressed suffix array that is loaded in the RAM (random-access memory). As a tradeoff, STAR requires a large amount of memory (nearly 30 GB for the human organism) but with the benefit of decreased runtime[50]. Moreover, when processing multiple samples from the same organism in parallel on the same server, the suffix array is loaded just once and can be shared among all STAR processes, thus keeping the memory footprint within reasonable limits. Another important measure is the overall mapping rate since a high number of unmapped reads constitute lost information, provided that the unaligned reads are not caused by sequencing artefacts. GEM could align the highest proportion of total reads, but this is accompanied by a higher fraction of ambiguously mapped reads and a higher number of introduced mismatches as well as indels. While it should be accounted for true variations between the reads and the reference genome, excessive introduction of them is not desired since inadequately added mismatches or indels might lead to an aggregation of reads at a potentially incorrect position and in further consequence to wrong conclusions. STAR, on the other hand, showed the lowest number of mismatches and indels while performing more soft clipping than the other tools. Engström *et al.*, 2013[58] hypothesized that the number of introduced variations is inversely proportional to the extent of soft clipping which is a likely explanation for the observed results. Furthermore, evaluation of the distribution of the aligned reads on the genome revealed that STAR reaches the highest exonic rate while the other tools map more reads to intronic or intergenic regions. In spite of that, comparison of gene expression quantification values of each of them showed high concordance among all tools, with STAR and TopHat exhibiting the highest correlation suggesting that the choice of the aligner might have only a minor effect at least in terms of subsequent differential gene expression analysis. Precision and recall measures calculated on the basis of alignments of simulated reads back this hypothesis as all aligners received comparably high values when including partly matched reads to the group of true positives. However, regarding simulated reads, results should be interpreted with caution since *in silico* reads are based on artificial error models which may not perfectly represent real RNA-seq reads[50]. Overall, STAR was the most convincing tool particularly with regard to runtime and it also reached high accuracy. In future, when longer reads will be common (see Chapter 5.1), alignment tools able to meet the challenges longer reads entail will be necessary and according to Dobin *et al.*, 2014[50] is STAR already able to align several kilobases long reads with high accuracy.

Quality Control In order to guarantee reliable results it is vital to detect any kind of bias or outlier as there are numerous possibilities to introduce errors in RNA-seq experiments before analysis-ready data are available[83][116]. For this reason, a comprehensive set of quality measures are calculated by the pipeline which should help to unvail different kinds of problems like technical biases or sample swaps. In fact, quality control is performed on multiple stages of sample processing starting in the laboratory and continuing with Illumina's Sequencing Control Software for the sequencing step. Finally, the quality control measures implemented in the pipeline are performed with most of them being calculated based on the BAM file after the alignment but also later based on intermediate analysis results, e.g. inter- and intra-condition-group correlation for differential expression analysis, total number of identified mutations for variant detection or total number of reported fusion.

All together, these metrics constitute a sound basis for the assessment of whether the RNA-seq experiment produced reliable data and in further consequence whether to include or exclude the sample from downstream analysis.

Expression Quantification and Associated Downstream Analysis Although a relatively straight forward task, several aspects have to be taken into account for expression quantification. Most prominent ones are how to deal with multi-mapped reads, reads with low mapping quality and reads that map to overlapping features. While it is common to have a fixed mapping quality threshold there are different approaches of how to deal with the remaining two issues. htseq-count, for example, completely ignores multi-mapped reads and offers three different strategies for reads mapping to overlapping features. Other tools like featureCounts or STAR's quantMode adopt a similar approach to htseq-counts' union-mode. However, Liao *et al.*, 2013[131] wrote that htseq-counts' counting algorithm interprets the necessary gene annotation input wrong to the extent that, different from the GFF file specification, it mistakenly excludes the right-most base of each feature, i.e. $[startpos, endpos[$ rather than $[startpos, endpos]$. Extending each right-most position of the UCSC gene annotation features by one base resolves this issue though. Using simulated reads with known expected read counts for each gene showed that the best performing combination of alignment and quantification tool is STAR with subsequent htseq-count quantification in intersection-nonempty mode, although the resulting Spearman correlation coefficient is only 0.771. Schuirer and Rome, 2016[221] used the same simulated data set as Fonseca *et al.*, 2104[62] in order to evaluate their exon quantification pipeline and reported Spearman correlation of slightly below 0.77 for STAR combined with htseq-count. Robert and Watson, 2015[212] performed a related evaluation where they simulated a fixed number of 1,000 perfect, i.e. zero mismatches, paired-end reads for each of about 20,000 genes and reported slightly higher correlation values overall. However, they stated that more realistic data will likely yield far worse results which is the case in the here presented comparison and multiple issues might be blamable for the rather small correlation. A substantial reason are alignment problems. For example, the overestimation of a few genes can be partially explained with the small fraction of misplaced reads and underestimated genes, on the other hand, with reads that could not be mapped (Figure 3.11a). A further reason are overlapping features, meaning distinct entities with shared genomic regions, with 9% of exons or more specifically 6% of total covered bases overlapping in the annotation file of the simulated set, thus reads mapping to these regions are underrated since discarded by default by htseq-count, featureCounts as well as STAR's quantification module. Concerning UCSC's gene annotation files for mouse and human, which are utilized by the pipeline, these numbers are even higher (19% of exons or 16% of bases for mouse and 15% of exons or 12% of bases for human). One approach to mitigate the issue of overlapping features is the application of stranded RNA-seq library preparation protocols, where the information about which strand a read was transcribed from is preserved, hence the quantification tools are able to unambiguously assign the reads to overlapping features that are located at different strands. The vast majority of analyzed samples so far were prepared with unstranded protocols but this will likely change in future and the pipeline is already able to cope with stranded RNA-seq data. Apart from that and as a result of the rather small correlation of expected and observed counts and the increasing importance of RNA-seq for diagnostics (see Chapter 5.3), Robert and Watson, 2015[212] proposed a list of trustworthy and untrustworthy human genes, respectively, which should support researchers in the interpretation of the results.

In terms of runtime, htseq-count performed worst with around 60 minutes regardless of the mode compared to the 20 to 30 times faster featureCounts and STAR. However, the slightly higher accuracy and the still acceptable duration of one hour endorsed the usage of htseq-count in intersection-nonempty mode.

For read count normalization two different strategies are implemented. After expression quantification, FPKM values are calculated for all processed samples by default. These values possess some shortcoming (see Chapters 1.4.3.1 and 3.1.4), thus should not be used to draw any final conclusions but constitute a helpful source to get a first impression of the data. For differential expression analysis a more adequate normalization is advisable. For that reason, the Relative Log Expression approach, implemented in the DESeq2 package, is used. Even so, additional normalization can be advantageous, especially when expression among samples under study is very heterogeneous with spike-in controls possessing the potential to improve normalization in cases like this[14][141][164][215]. Tools like DESeq2[140] or RUVseq[210] can directly employ the additional dimension of information provided by the spike-ins and incorporate it into the normalization factors. So far, only very few libraries with included spike-ins were processed with the here presented pipeline and spike-in normalization has to be performed manually yet.

Irrespective of the normalization strategy, differential gene expression analysis is performed using DESeq2 by default. Here (Chapter 3.1.4), as well as in other studies where partially other organisms were used, and, except for the DESeq2 publication itself[140] and briefly mentioned in Seyednasrollah, *et al.*, 2013[223], DESeq2 was not included[172][230][266], it could be shown that the results of distinct differential expression analysis tools disagree notably. With the used human tumor-control sample data set edgeR reported considerably more differentially expressed genes than DESeq2 using the same Benjamini–Hochberg adjusted p-value cutoff. On the other hand, edgeR was more conservative with lower numbers of replicates ($\leqslant 6$). A shortcoming of using tumor samples for this kind of comparison is the typically higher degree of heterogeneity of them. Apart from that, this tumor-control data set constituted the one with the highest number of replicates analyzed with the pipeline so far, hence valuable to analyze the effect of different numbers of replicates and moreover complement other comparison studies of differential expression analysis tools. However, more studies will be necessary to find a consensus about which DE analysis tool should be used.

Regarding the step of differential expression analysis, the pipeline is semi-automated in that the user has to define the condition groups and to assign the according sample to them before analysis can be started. This, however, could be easily automated provided that the collaborators specify the type of analysis and sample classification at the beginning of the project. Another possibility would be to offer this functionality via the web application, which would again enhance the usability of the pipeline for the collaborators and, on the other hand, would not be connected with great expenses since all necessary scripts are already in place.

In addition to differential gene expression analysis there are two features implemented in the pipeline that allow for investigation of RNA-seq data at a higher resolution than at the gene level. Both of them have different aims with one of them being able to reveal differences on the exon level between samples of different conditions by utilizing DEXSeq[10], thus allowing for differential alternative splicing detection, and the other one to visualize alternative isoform expression[100][101]. The latter one, i.e. IGV's Sashimi plot feature, is only applicable when one or just a few genes of interest have to be investigated for a few samples at a maximum since examination has to be done visually, thus manually, and cannot be performed in a systematic and automated manner. Aside from that, differential expression analysis on the transcript level can provide additional valuable insights, but this is not a trivial task since reliable transcript expression quantification is challenging due to the fact that transcripts belonging to the same gene often share a substantial amount of their sequence and revealing the transcript of origin, especially for short reads, is difficult[35][243]. Recent advances in this field led to the advent of new tools like Sailfish[192], kallisto[21] or RSEM[122] which could demonstrate improved transcript quantification reliability[243], hence are able to produce reliable results even on the transcript level.

Fusion Detection An important feature particularly for the analysis of RNA-seq data of tumor samples is the possibility to identify chimeras, i.e. gene fusions. A big challenge is still the relatively high number of false positives reported by fusion detection tools (Chapter 3.1.6)[28]. In order to increase the precision of the results additional annotations are created by the pipeline for each of the reported fusions which purpose is not primarily to filter fusions but rather to support the user in making decisions whether or not a fusion appears to be reliable. These annotations were designed and implemented based on the comparison with a set of four well-studied breast cancer cell lines where initial fusion detection reported a substantial amount of false positive but also some false negative chimeras. However, the data sets comprise reads with a length of 50 bp, what is notably shorter than the nowadays common 100 bp or longer reads, and it could be shown recently that longer reads can considerably increase the accuracy of fusion detection[115]. Nevertheless, using the additional annotation information for the fusions detected in the four cell lines could help to reduce the number of false positives (Chapter 3.1.6). On the other hand, using too strict filtering criteria proved to involve the danger of discarding true fusions and more sophisticated algorithms will be needed for adequate fusion detection in future.

Variant Detection Since RNA-seq data include sequence information of the processed individuals they can be used for variant detection as well. In comparison to whole exome or whole genome sequencing data, RNA-seq data possess some special characteristics which complicates variant calling and must be taken into account. In spite of that, these characteristics also hold some additional advantages since RNA-seq data comprise expressed thus functionally important genetic regions and the detection of variants in these regions can provide valuable biological insights[196]. While other studies already investigated variant calling in RNA-seq data[32][173][196][200][252] they focused on SNVs but not indels, did not compare the performance of different tools or, if they did, compared older versions or different collections of tools. Here, three widely used programs were evaluated in terms of performance not only for SNVs but also for indels in order to determine which one should be used by the pipeline for variant calling in RNA-seq data. For that purpose, a published gold standard data set including high confidence SNV and indel calls as well as high-confidence variant calling regions of a well-studied sample was downloaded and used along with respective RNA-seq and whole genome sequencing raw data. STAR was used for the alignment due to the results of Chapter 3.1.2 where it was shown that GEM, for example, introduces disproportionately large amounts of mismatches and indels. Regarding variant calling, the results showed that GATK performed best in terms of precision and recall for both, SNVs and indels. However, results based on this gold standard data tend to euphemize the outcome since they are restricted to high-confidence regions, thus performance might be overestimated here compared to common RNA-seq variant calling results. In spite of that, a substantial amount of false positive as well as false negative mutations were reported even by the best performing tool GATK, which is partly due to the special characteristics of RNA-seq data. Several reasons for that could be determined where read coverage was the most crucial factor. Certainly, variants in unexpressed regions cannot be detected at all, but also for expressed regions it could be shown that variant detection accuracy enhances with increasing read depth, indicating that especially variants detected in low expressed regions, that are of particular interest, need further validation. Another reason is RNA editing and filtering of known RNA editing sites showed that a high number of putatively wrong variants are likely caused by this post-transcriptional mechanism. Moreover, after filtering there were still more than 250 remaining SNVs which showed the RNA editing characteristic A→I mutation pattern[171], which might indicate that numerous RNA editing sites are still unknown. An additional reason for supposedly wrong mutations could be shown to be allelic differences. Apart from SNVs, special emphasis was also put on the detection of indels. Stenson *et al.*, 2014[234] stated that more than 23% of mutations in HGMD are indels which underpins the importance of reliable indel detection. As for SNVs, indel detection resulted in good precision but moderate recall values which is partially due to poor coverage as well but also other indel specific reasons. For example, short introns that caused the alignment program to introduce a deletion instead of the correct intron, i.e. split the read, led to several false positive deletions. This issue could be mitigated by adapting STAR's `--alignIntronMin` parameter that determines the minimum intron size, i.e. all alignment gaps below this value are classified as deletion, which is set to 21 by default[49]. However, defining an appropriate value is difficult since there are 1,589 unique introns which are shorter than 21 bases in the UCSC hg19 gene annotation and even 184 of them have a length of one. Thus, more sophisticated approaches like GATK's Indel Realignment feature, which is commonly used for indel calling in DNA sequencing data but not yet mature enough for RNA-seq according to GATK's GATK Best Practices Guide[68], would be beneficial. Apart from that, several incorrect mutations remain where no biological explanation could be found, hence likely true false positives and false negatives, respectively. Most of them are caused by repetitive regions where the alignment program cannot correctly map the reads, thus introducing SNVs or indels that are not true. Moreover, allelic mapping bias, which means that a read carrying a mutation has a lower probability to be mapped correctly, can lead to wrong genotypes or even missed mutations[30]. Problems like this will likely be solved with longer reads (see

Chapter 5.1) which in turn will increase the accuracy of variant calling and Piskol *et al.*, 2013[196] concluded in their study that correct mapping is the most fundamental factor for reliable variant calling. Nevertheless, for now, sophisticated filtering strategies have to be applied in order to gain reliable results. Based on the findings in Chapter 3.1.7 several additional filter criteria for variants were defined and tested for sequenced in-house RNA-seq samples, where matching whole exome and partially whole genome sequencing data were available. Using these filters resulted in high precision yet low recall values which is largely due to the high fraction of unexpressed regions, but too strict filtering, on the other hand, involves also the danger of discarding true and informative variants. Filtering known RNA editing sites, for example, is only appropriate when this kind of mutations do not matter for the biological question that should be answered. Furthermore, correctly identified variants might have less than 5-fold coverage, hence would be discarded. Consequently, there is a tradeoff between precision and recall and filter criteria should be applied or ignored based on the application. For instance, the high precision values achieved by RNA-seq variant calling and additional filtering indicate that the remaining variants are indeed highly confident, which is especially important if RNA-seq is used as a diagnostic tool (see Chapter 5.3). On the downside, recall values are low, which shows that a lot of true variants are missed. Thus, depending on the scientific question that should be addressed with RNA-seq variant calling, individual filtering criteria can be either relaxed or discarded at all in the pipeline.

4.2. Input Data Quality

Crucial factors for successful RNA-seq data analysis are, on the one hand, the study design and, on the other hand, the quality of the starting data material. Several fundamental considerations regarding design and input data quality were examined and helped to define basic criteria that should be met. For example, generating more sequence does not linearly increase the number of sufficiently covered bases. In fact, more than 7.5 Gb of sequence will only marginally enhance average coverage per expressed gene. This led to the decision that on the HiSeq2500 six and on the HiSeq4000 seven RNA-seq samples per lane are sequenced since the former is able to produce about 400 Gb per flow cell[92] and the latter about 500 Gb per flow cell[91] if run in 100 bp paired-end mode. Read depth is especially essential for variant detection and using the aforementioned criteria for sequencing the resulting samples show on average about 50% of exonic bases that are covered more than 8-fold. For differential expression analysis, however, coverage is not as important as a sufficient number of replicates which could be shown in multiple studies that investigated the impact of both factors[137][206][222]. Liu *et al.*, 2014[137] concluded that everything above 10 million reads per sample results in only slightly enhanced power to detect differentially expressed genes but increasing the number of replicates shows an considerable effect. This could also be shown here in Chapter 3.2.2 where the number of detected differentially expressed genes increased with the number of replicates. Schurch *et al.*, 2016[222] suggested to use at least six biological replicates and even 12 if it is important to identify differentially expressed genes for all fold changes. The main reason for this is that the underlying mathematical models of differential expression analysis tools heavily depend on a sufficient number of replicates in order to be able to properly model the variability in gene expression measurements[206].

Another important yet controversial issue is whether to remove duplicate reads or not. In Chapter 3.2.3 it could be shown that duplicate reads mainly occur in highly expressed genes, indicating that duplicates mainly occur by chance since alignment of a high number of reads to a limited number of bases will inevitably result in reads that map to the same position. These findings were also published in the course of the GEUVADIS project and underlined the conclusion that duplicates are more likely due to a saturation of the read mapping space than technical artifacts and that the removal of duplicates would underestimate the expression of numerous genes and possibly interfere the detection of true variation[119]. Accordingly, duplicate reads are kept for expression quantification. Nevertheless, it is common practice for variant calling to remove duplicates beforehand to not bias the outcome by possibly accumulated artifacts, thus duplicate reads are removed prior to variant calling by the pipeline.

Furthermore, a basic consideration that has to be taken into account already before alignment is whether or not to preprocess sequenced reads. A sound reason for that are possible present adapter sequences in the resulting reads as a consequence of too short insert size. These undesired bases would likely distort the resulting alignment and should be clipped. However, the majority of the sequenced RNA-seq samples showed insert sizes larger than 200 which is sufficient for 100 bp paired-end reads to not suffer from included adapter sequences, hence adapter clipping and quality trimming are not performed by the pipeline by default but can be done if there are any base quality issues.

4.3. Database and Web Application

Various results are inserted into the database in the course of the analysis of an RNA-seq sample. Together with the variant information of whole exome as well as whole genome sequencing data and in near future likely also with ChIP-seq data, this data pool constitutes a valuable source of information and provides the opportunity to combine these data in order to gain deeper understanding of biological processes. All the information stored in the database enables the implementation of database queries and in further consequence of scripts that help to answer biological questions. For this purpose, the Database Interface and MySQL Driver for R (RMySQL)[1] package can be exploited which offers the possibility to load all necessary data into R and process the information there in a convenient way. The scripts, in turn, can be integrated into the analysis pipeline and results can be stored in the database.

[1] `https://cran.r-project.org/package=RMySQL` (last accessed 01.10.2016)

Taken together, the information in the database in combination with the web application constitutes a useful tool where data can be visualized but also analyzed in a user-friendly manner. Several sample-centered queries are provided which can be used to browse general sample information but also detailed, mainly quality related information that helps to determine whether the sample should be used for downstream analysis or not. On top of that, queries featuring analysis purposes are implemented which allows to browse stored RNA-seq data analysis results but also to identify causative mutations based on pre-defined queries[256].

Part V.

Outlook

5. Outlook

RNA sequencing has superseded microarrays for transcriptome analysis and already covers a wide range of applications. Nevertheless, consistent further development helps to improve analysis but also opens up new possibilities. One remaining issue with RNA-seq analysis is the limited read length which complicates analysis and makes the results prone to certain errors but longer sequencing reads will help to overcome these problems. Furthermore, new methods like the sequencing and subsequent analysis of RNA-seq data from very small amounts of starting material will further contribute to a better understanding of biological functions. Finally, enhanced sophistication and reliability of RNA-seq data and analysis results will make RNA-seq not only attractive for scientific reasons but also for clinical decisions. These efforts will be discussed in more detail in the following.

5.1. New Sequencing Technologies

The entire area of genomics made a great leap forward in development since the advent of second- or next-generation sequencing. The decreasing costs together with the massively increasing amount of sequence produced in a short period of time[152] promoted the investigation of the transcriptome as well as genome of many different species. Consequently, this led to a better understanding of underlying mechanisms and, especially in humans, facilitated the identification of many pathogenic processes in the genome[52].

Despite numerous unquestionable advantages of NGS compared to earlier sequencing technologies there are also downsides[121] which complicate data analysis and in further consequence the interpretation of analysis results. The most obvious issue is the limited length of resulting sequence reads. This hampers the correct and unique alignment of reads, respectively, most notably in repetitive regions of the target genome[249]. Misaligned reads, in turn, can lead to mistakenly called variants and multimapped reads, i.e. reads which cannot be mapped uniquely, have the potential to bias expression quantification if not accounted for. Regarding gene fusion detection, mapping problems due to sequencing errors or homologous or polymorph regions can mislead fusion detection algorithms to report fusion events that are not true[35]. Furthermore, another problem of short read length arises when trying to reconstruct transcripts and quantifying their expression levels since limited read length makes it hard to correctly assign a read to a distinct isoform as they usually do not span all splice junctions[35].

New sequencing technologies, also referred to as third generation sequencing (TGS) technologies, can generate much longer reads and thus are able to overcome the abovementioned short read length related issues. Resulting reads can have the length of up to 100,000 bp and beyond which becomes possible since they do not have to deal with amplification- or phasing-induced bias. This, in turn, is due to the fact that they do not rely on amplification nor on cycle-wise sequencing but rather perform uninterrupted sequencing of a single molecule[121][219].

Eventually, the advantages that come with longer reads are accompanied by new challenges regarding data analysis and new tools, which are able to meet those challenges, will be needed.

Two available single-molecule sequencing technologies are discussed in the following two chapters.

5.1.1. Single-Molecule Real-Time Sequencing Technology

Single-molecule real-time (SMRT) sequencing is a technology developed by Pacific Biosciences (PacBio) which was published in 2009[56]. Here, sequencing is performed by replication of target DNA molecules with a polymerase that is immobilized at the bottom of a so called zero-mode waveguide (ZMW) (Figure 5.1), the central sequencing unit of the SMRT technology[209]. SMRTcells, containing up to one million ZMWs[186], are then used by the PacBio sequencing instruments for the actual sequencing process. Beforehand, SMRTbells (Figure 5.2) must be prepared which represents the target DNA molecules in double-stranded form. The SMRTbells are capped with hairpins on both ends containing complementary primer sequences so that they can bind to the polymerase at the bottom of the ZMWs[248]. SMRTbells are put on a SMRTcell and once the prepared target molecules are bound to the polymerases in the ZMWs sequencing can start. For this purpose, flourescently labeled nucleotides are added to the SMRTcells which, in turn, are used by the DNA polymerase to replicate the template DNA. Each time a nucleotide is incorporated a fluorescent signal is emitted which is recorded in real time resulting in a sequence of signals that represent the sequence of bases of the target DNA molecule[209]. In order to ensure that the signal of the incorporated nucleotide is higher than that of the surrounding ones the ZMWs have a diameter which is markedly smaller than the wavelength of the used laser light. As a consequence, the light will only illuminate the very bottom of the ZMW, thus just the nucleotide in this area will emit a signal[219].

At the time of writing, PacBio offers two sequencing instruments, namely the PacBio RS II and the Sequel System. Exploiting the SMRT technology, they can produce average read length of over 10,000 bp with some reads even longer than 60,000 bp[187]. One drawback is that single reads have a high error rate[6]. However, this can be considerably improved since the target DNA molecules are provided in the form of SMRTbells. Thus, sequencing does not have to stop when the sequence is read once but can rather be continued to produce several copies of it, given that the template is short enough that the polymerase is functioning for several cycles[209]. When per-

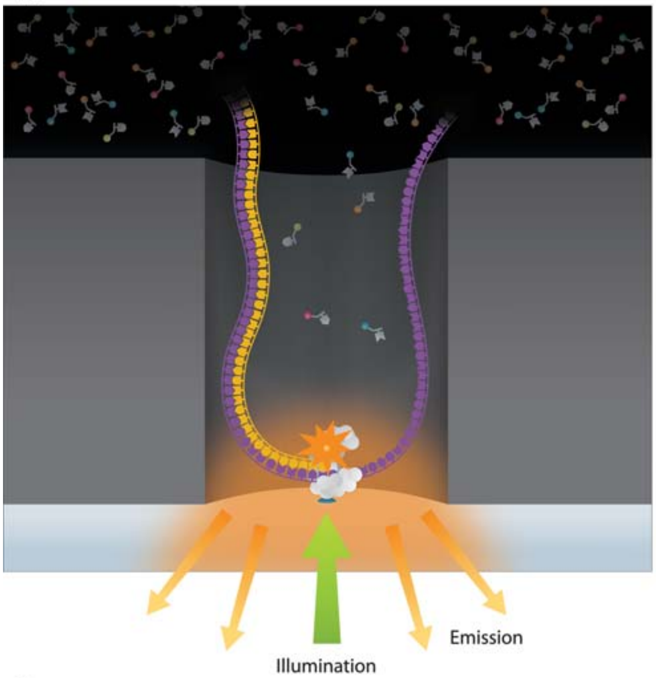


Figure 5.1.: Schematic representation of a zero-mode waveguide (ZMW) with a DNA polymerase (grey) fixed at the bottom and flourescently labeled nucleotides (red, blue, green and yellow, respectively) that are incorporated to the growing sequence by the polymerase where the emitted fluorescent signal can be detected by an optical system in real time. (Image taken from [219])

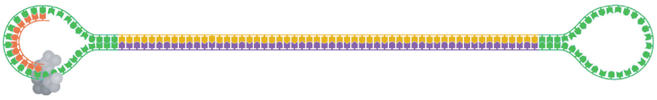


Figure 5.2.: SMRTbell representing the target DNA molecule. Forward as well as reverse strand are present and form a circular molecule through hairpin sequences at both ends. The hairpins include complementary primer sequences where the polymerase (grey) can bind to. (Image taken from [248])

forming sequencing in this manner for at least 30 cycles the resulting consensus sequences can have an accuracy of above 99.999%. Apart from that, both instruments can run for a maximum of 4 hours in the course of which the RS II can produce up to 1 Gb and the Sequel up to 10 Gb[6] of sequence.

Currently, the main limitation of SMRT sequencing are the high costs and relatively small throughput in comparison to second-generation sequencing technologies, yet it has already been used successfully in several studies especially for *de novo* assembly of various organisms[121].

5.1.2. Nanopore Sequencing Technology

In early 2014, Oxford Nanopore Technologies rolled out their first sequencing instrument, the MinION, through the MinION Access Program where early access users could test their pocket-sized portable device. Several months later, in May 2015, the MinION became commercially available[180]. By the time of writing, two additional instruments were in the pipeline of Oxford Nanopore Technologies. The PromethION instrument provides very high throughput real-time analyses and is available through an early access program since the middle of 2015[182]. The second instrument, SmidgION, was announced in early 2016 and is designed to work with low power devices such as smartphones[184].

The central sequencing unit of nanopore sequencing technologies is the nanopore which can be either biological or synthetical[219]. All of the three abovementioned instruments use biological ones which will likely change in future devices[185]. In the current generations, biologically engineered protein-nanopores are placed in an electrically resistant synthetic polymer membrane (Figure 5.3) and sequencing is performed by measuring and recording the characteristic modulation of the flowing current as the single stranded target DNA molecule is driven through the nanopore base by base[235].

Just as the SMRT sequencing technology, nanopore sequencing allows for sequencing of a single molecule in real time. Furthermore, molecules do not have to be modified nor amplified, thus sequencing can be performed quickly with a relative small amount of starting material while producing long reads with constant quality[33][219]. Recently, Oxford Nanopore published results where they directly sequenced entire RNA molecules without the need of preceding reverse transcription and amplification, thus producing full-length, strand-specific RNA sequences[66]. In general, the average read length yielded by the R7 generation of the chemistry is around 6,000 to 8,000 bp[94][147] while the latest version, R9, can create reads with an average length of 9,000 bp, with the longest reads exceeding 130,000 bp[139]. However, despite ongoing improvements the accuracy of the resulting reads is still a limiting factor, where a single read can reach a per-base accuracy of about 90%[133][183]. Using Oxford Nanopore's 2D method, where a hairpin adapter is added to one end of the double stranded DNA template resulting in the continuous sequencing of template as well as complement strand, the accuracy of a read improved to 95% and higher[183].

The key advantages of the Oxford Nanopore instruments are the low costs, the effortless library preparation, the handy size (i.e. of the MinION and likely the SmidgION) and as a consequence thereof the simpleness and flexibility for the user. These features predestinated the MinION to be used even in remote locations and in 2015, for example, helped to monitor the Ebola virus disease epidemic in West Africa[121][199].

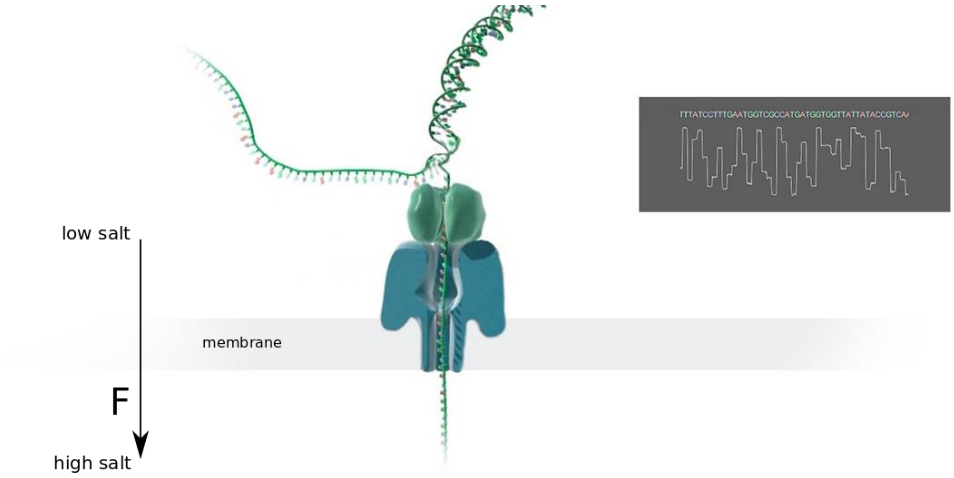


Figure 5.3.: Schematic illustration of the Oxford Nanopore sequencing technology. A protein-nanopore (petrol blue) is inserted into an electrically resistant membrane (light grey) and the different salt concentrations on both sides of the membrane result in a voltage across that membrane which leads to an ionic current flowing solely through the nanopore[33][219]. An enzyme (green) is attached to a single stranded end of the double stranded target DNA molecule and the entire complex is transferred to the nanopore. Subsequently, the enzyme unwinds the double strand and pulls it through the nanopore where sequencing takes place by identifying the characteristic ionic current disruption of each of the four nucleotides as they run through the nanopore (dark grey shaded graph). (Image adapted from [181])

5.2. Single-cell RNA-seq

Advances in library preparation protocols make it possible to create RNA-seq libraries from very small amounts of input materials. This is useful if just a limited amount of biological material is available or insights at the cell level are of particular interest. Especially single-cell analysis is getting more popular as it implicates several benefits and offers new possibilities like the characterization of different cell populations in a specific tissue, gaining deeper understanding of different cell states, analyzing the stochasticity of gene expression or revealing new insights into gene regulation mechanisms[35][113]. Nevertheless, before sequencing on a single-cell level can be performed, cells must be isolated and the tiny amount of resulting RNA must be amplified[113]. For both steps, several methods emerged which could already demonstrate their usability. A single-cell capturing workflow starts with dissecting the tissue, dissociating the cells, suspending them in a buffer and subsequently sorting them using sophisticated tools and methods like the Fluidigm C1[197] or Dropseq[146] where the latter one is able to process up to 10,000 cells at a time[35][113]. After single-cells are isolated sequencing libraries can be created with protocols like Clonetech's SMART-Seq[204][194] or NuGEN's Ovation RNA-seq system[242] and resulting libraries can be sequenced on common instruments. The resulting data, however, possess some special characteristics, mainly due to the minute amount of starting material and different analysis strategies must be pursued for single-cell RNA-seq compared to traditional RNA-seq data. Stegle *et al.*, 2015[233], who discussed computational as well as analytical challenges when processing single-cell RNA-seq data, concluded that apart from existing ones, the development of additional, sophisticated analysis tools, which are tailored to the distinctive properties of single-cell RNA-seq data, will be vital in order to uncover all information these data can reveal. If desired, these single-cell RNA-seq data analysis designated tools can be implemented in a wrapper script which in turn can be incorporated in the here presented pipeline thanks to its modular architecture.

5.3. RNA-seq in Diagnostics

Owing to its increasing accuracy RNA-seq is coming closer to diagnostics and recent studies already demonstrated the potential and importance of the additional information RNA-seq can provide especially when whole exome or whole genome sequencing data do not return any genetic diagnosis which applies to about 50% to 75% of cases for a lot of rare Mendelian diseases[42][114]. In Kremer *et al.*, 2016[114], for example, RNA-seq data of 47 unsolved mitochondrial disease patients were analyzed and a diagnosis for 5 of them could be provided which was not possible with DNA sequencing data alone. Cummings *et al.*, 2016[42] investigated 50 cases with rare neuromuscular disorder which again could not be diagnosed with DNA sequencing data and were able to solve 17 of them with the aid of RNA-seq. Both of the mentioned studies focused on aberrant splicing and allele specific as well as differential expression but additional applications are emerging[27]. Large scale fusion detection, for example, can help to better identify and classify cancer and in further consequence help to determine appropriate treatment based on the detected fusions. However, if RNA-seq should be established in clinical decision making, further efforts concerning general standards and best practices are required in order to guarantee accuracy, reproducibility and precision[27]. The GEUVADIS project was such an effort where the reproducibility of RNA-seq could be proved and certain quality measures established[119][239]. Furthermore, other endeavors like the Genotype–Tissue Expression (GTEx)[36] or the Sequencing Quality Control (SEQC)[237] project also aimed to advance the RNA-seq technology. Revealing the gene expression landscape across different tissue types, as was done by the GTEx pilot analysis, will further help to understand distinct variations in disease relevant tissues which is crucial since disease causing differences are often solely observeable in specific tissues[42].

In conclusion, RNA-seq has already demonstrated that it is a powerful technique with a wide variety of applications and that it has the ability to complement DNA sequencing in clinical diagnostics. However, further quality standards and guidelines to enhance and consequently guarantee reliability, but also specially trained clinical staff which is able to interpret as well as communicate results, will be vital.

Bibliography

[1] Method of the Year 2013. *Nature Methods*, 11(1):2014, 2014.

[2] Xian Adiconis, Diego Borges-Rivera, Rahul Satija, David S DeLuca, Michele A Busby, Aaron M Berlin, Andrey Sivachenko, Dawn Anne Thompson, Alec Wysoker, Timothy Fennell, Andreas Gnirke, Nathalie Pochet, Aviv Regev, and Joshua Z Levin. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature Methods*, 10(7):623–9, 2013.

[3] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011.

[4] Gael P Alamancos, Eneritz Agirre, and Eduardo Eyras. Methods to study splicing from high-throughput RNA sequencing data. *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, pages 357–397, 2014.

[5] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[6] AllSeq, Inc. Pacific Biosciences - Overview. `http://allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/`, 2016. [Web page; last accessed 01.10.2016].

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

[8] Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, 8(9):1765–86, 2013.

[9] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[10] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome research*, 22(10):2008–17, 2012.

[11] Simon Andrews. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010. [Web page; last accessed 01.10.2016].

[12] Claudia Angelini, Daniela De Canditiis, and Italia De Feis. Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics*, 15(1):135, 2014.

[13] M. J. Annala, B. C. Parker, W. Zhang, and M. Nykter. Fusion genes and their discovery using high throughput sequencing. *Cancer Letters*, 340(2):192–200, 2013.

[14] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, and Others. The external RNA controls consortium: a progress report. *Nature methods*, 2(10):731–734, 2005.

[15] Callum J Bell, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science Translational Medicine*, 3(65):65ra4, 2011.

[16] A M Benjamin, M Nichols, T W Burke, G S Ginsburg, and J E Lucas. Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics*, 15(1):570, 2014.

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

[18] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A.

Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie VandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[19] Inanç Birol, Shaun D. Jackman, Cydney B. Nielsen, Jenny Q. Qian, Richard Varhol, Greg Stazyk, Ryan D. Morin, Yongjun Zhao, Martin Hirst, Jacqueline E. Schein, Doug E. Horsman, Joseph M. Connors, Randy D. Gascoyne, Marco A. Marra, and Steven J. M. Jones. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872–2877, 2009.

[20] J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, S. Burgess, T. Buza, C. Gresham, F. McCarthy, L. Pillai, H. Wang, S. Carbon, H. Dietze, S. E. Lewis, C. J. Mungall, M. C. Munoz-Torres, M. Feuermann, P. Gaudet, S. Basu, R. L. Chisholm, R. J. Dodson, P. Fey, H. Mi, P. D. Thomas, A. Muruganujan, S. Poudel, J. C. Hu, S. A. Aleksander, B. K. McIn-

tosh, D. P. Renfro, D. A. Siegele, H. Attrill, N. H. Brown, S. Tweedie, J. Lomax, D. Osumi-Sutherland, H. Parkinson, P. Roncaglia, R. C. Lovering, P. J. Talmud, S. E. Humphries, P. Denny, N. H. Campbell, R. E. Foulger, M. C. Chibucos, M. Gwinn Giglio, H. Y. Chang, R. Finn, M. Fraser, A. Mitchell, G. Nuka, S. Pesseat, A. Sangrador, M. Scheremetjew, S. Y. Young, R. Stephan, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, M. D. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, G. T. Hayman, S. J. Wang, V. Petri, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J. M. Cherry, M. C. Costanzo, J. Demeter, S. S. Dwight, S. R. Engel, B. C. Hitz, D. O. Inglis, P. Lloyd, S. R. Miyasato, K. Paskov, G. Roe, M. Simison, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, D. Li, E. Huala, J. Argasinska, C. Arighi, A. Auchincloss, K. Axelsen, G. Argoud-Puy, A. Bateman, B. Bely, M. C. Blatter, C. Bonilla, L. Bougueleret, E. Boutet, L. Breuza, A. Bridge, R. Britto, C. Casals, E. Cibrian-Uhalte, E. Coudert, I. Cusin, P. Duek-Roggli, A. Estreicher, L. Famiglietti, P. Gane, P. Garmiri, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, C. Hulo, R. Huntley, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, A. Macdougall, M. Magrane, M. Martin, P. Masson, P. Mutowo, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Shypitsyna, A. Stutz, S. Sundaram, M. Tognolli, C. Wu, I. Xenarios, J. Chan, R. Kishore, P. W. Sternberg, K. Van Auken, H. M. Muller, J. Done, Y. Li, D. Howe, and M. Westerfeld. Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[21] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[22] Broad Institute. GATK HaplotypeCaller - Call germline SNPs and indels via local re-assembly of haplotypes. `https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php`, 2016. [Web page; last accessed 01.10.2016].

[23] Broad Institute. Picard. `http://broadinstitute.github.io/picard/`, 2016. [Web page; last accessed 01.10.2016].

[24] Terence A Brown. *Genomes*. 2nd edition. BIOS Scientific, 2002.

[25] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[26] Michael Burrows and David J Wheeler. A block-sorting lossless data compression algorithm. 1994.

[27] Sara A. Byron, Kendall R. Van Keuren-Jensen, David M. Engelthaler, John D. Carpten, and David W. Craig. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17(5):257–271, 2016.

[28] Matteo Carrara, Marco Beccuti, Federica Cavallo, Susanna Donatelli, Fulvio Lazzarato, Francesca Cordero, and Raffaele A Calogero. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC Bioinformatics*, 14(7):1–11, 2013.

[29] Matteo Carrara, Marco Beccuti, Fulvio Lazzarato, Federica Cavallo, Francesca Cordero, Susanna Donatelli, and Raffaele A Calogero. State-of-the-art fusion-finder algorithms sensitivity and specificity. *BioMed Research International*, 2013, 2013.

[30] Stephane E Castel, Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. Tools and best practices for allelic expression analysis. *Genome Biology*, 16(1):195, 2015.

[31] Iouri Chepelev, Gang Wei, Qingsong Tang, and Keji Zhao. Detection of single nucleotide variations in expressed exons of the human genome using RNA-seq. *Nucleic Acids Research*, 37(16):1–8, 2009.

[32] Elizabeth T Cirulli, Abanish Singh, Kevin V Shianna, Dongliang Ge, Jason P Smith, Jessica M Maia, Erin L Heinzen, James J Goedert, and David B Goldstein. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biology*, 11(5):1, 2010.

[33] James Clarke, Hai-chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, 2009.

[34] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[35] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016.

[36] GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[37] The ENCODE Consortium. Standards, guidelines and best practices for rna-seq. `https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf`, June 2011. [Online; last accessed 01.10.2016].

[38] Francis H C Crick. On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[39] Francis H C Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[40] Peng Cui, Qiang Lin, Feng Ding, Chengqi Xin, Wei Gong, Lingfang Zhang, Jianing Geng, Bing Zhang, Xiaomin Yu, Jin Yang, Songnian Hu, and Jun Yu. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 96(5):259–265, 2010.

[41] Jason S. Cumbie, Jeffrey A. Kimbrel, Yanming Di, Daniel W. Schafer, Larry J. Wilhelm, Samuel E. Fox, Christopher M. Sullivan, Aron D. Curzon, James C. Carrington, Todd C. Mockler, and Jeff H. Chang. GENE-counter: A computational pipeline for the analysis of RNA-seq data for gene expression differences. *PLoS ONE*, 6(10), 2011.

[42] Beryl B Cummings, Jamie L Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A Reghan Foley, Veronique Bolduc, Leigh Waddell, Sarah Sandaradura, Gina L O'Grady, and Others. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *bioRxiv*, page 074153, 2016.

[43] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[44] Mattia D'Antonio, Paolo D'Onorio De Meo, Matteo Pallocca, Ernesto Picardi, Anna Maria D'Erchia, Raffaele A Calogero, Tiziana Castrignanò, and Graziano Pesole. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC Genomics*, 16(Suppl 6):S3, 2015.

[45] Jacob F Degner, John C Marioni, Athma A Pai, Joseph K Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–12, 2009.

[46] Nicolas Delhomme, Ismaël Padioleau, Eileen E Furlong, and Lars M Steinmetz. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics*, 28(19):2532–3, 2012.

[47] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–2, 2012.

[48] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–83, 2013.

[49] Alexander Dobin. *STAR manual v2.3.0*.

[50] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[51] Lars Dölken, Zsolt Ruzsics, Bernd Rädle, R G Mages, Reinhard Hoffmann, Paul Dickinson, and Thorsten Forster. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *Rna*, 14(9):1959–1972, 2008.

[52] Richard M. Durbin, David L. Altshuler, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, Francisco M. De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Bartha M. Knoppers, Eric S. Lander, Hans Lehrach, Elaine R. Mardis, Gil A. McVean, Debbie A. Nickerson, Leena Peltonen, Alan J. Schafer, Stephen T. Sherry, Jun Wang, Richard K. Wilson, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J. Fennell, David B. Jaffe, Erica Shefler, Carrie L. Sougnez, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer Stone, Kevin J. McKernan, Gina L. Costa, Jeffry K. Ichikawa, Clarence C. Lee, Ralf Sudbrak, Tatiana A. Borodina, Andreas Dahl, Alexey N. Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V. Soldatov, Bernd Timmermann, Marius Tolzmann, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja,

Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, John Burton, David M. Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P. Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Xiaodong Fang, Xiaosen Guo, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Gabor T. Marth, Erik P. Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R. Quinlan, Chip Stewart, Michael P. Stromberg, Alistair N. Ward, Jiantao Wu, Charles Lee, Ryan E. Mills, Xinghua Shi, Mark J. Daly, Mark A. DePristo, Aaron D. Ball, Eric Banks, Brian L. Browning, Kiran V. Garimella, Sharon R. Grossman, Robert E. Handsaker, Matt Hanna, Chris Hartl, Andrew M. Kernytsky, Joshua M. Korn, Heng Li, Jared R. Maguire, Steven A. McCarroll, Aaron McKenna, James C. Nemesh, Anthony A. Philippakis, Ryan E. Poplin, Alkes Price, Manuel A. Rivas, Pardis C. Sabeti, Stephen F. Schaffner, Ilya A. Shlyakhter, David N. Cooper, Edward V. Ball, Matthew Mort, Andrew D. Phillips, Peter D. Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C. Yoon, Carlos D. Bustamante, Adam Boyko, Jeremiah Degenhardt, Simon Gravel, Ryan N. Gutenkunst, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Richard E. Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O. Korbel, Adrian M. Stütz, Markus Bauer, R. Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Kai Ye, Yutao Fu, Fiona C. L. Hyland, Jonathan M. Manning, Stephen F. McLaughlin, Heather E. Peckham, Onur Sakarya, Yongming A. Sun, Eric F. Tsung, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Ralf Herwig, Dimitri V. Parkhomchuk, Richa Agarwala, Hoda M. Khouri, Aleksandr O. Morgulis, Justin E. Paschall, Lon D. Phan, Kirill E. Rotmistrovsky, Robert D. Sanders, Martin F. Shumway, Chunlin Xiao, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L. Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, James Knight, Roger Winer, David W. Craig, Steve M. Beckstrom-Sternberg, Alexis Christoforides, Ahmet A. Kurdoglu, John V. Pearson, Shripad A. Sinari, Waibhav D. Tembe, David Haussler, Angie S. Hinrichs, Sol J. Katzman, Andrew Kern, Robert M. Kuhn, Molly Przeworski, Ryan D. Hernandez, Bryan Howie, Joanna L. Kelley, S. Cord Melton, Yun Li, Paul Anderson, Tom Blackwell, Wei Chen, William O. Cookson, Jun Ding, Hyun Min Kang, Mark Lathrop, Liming Liang, Miriam F. Moffatt, Paul Scheet, Carlo Sidore, Matthew Snyder, Xiaowei Zhan, Sebastian Zöllner, Philip Awadalla, Ferran Casals, Youssef Idaghdour, John Keebler, Eric A. Stone, Martine Zilversmit, Lynn Jorde, Jinchuan Xing, Evan E. Eichler, Gozde Aksay, Can Alkan, Iman Hajirasouliha, Fereydoun Hormozdiari, Jeffrey M. Kidd, S. Cenk Sahinalp, Peter H. Sudmant, Ken Chen, Asif Chinwalla, Li Ding, Daniel C. Koboldt, Mike D. McLellan, John W. Wallis, Michael C. Wendl, Qunyuan Zhang, Cornelis A. Albers, Qasim Ayub, Senduran Balasubramaniam, Jeffrey C. Barrett, Yuan Chen, Donald F. Conrad, Petr Danecek, Emmanouil T. Dermitzakis, Min Hu, Ni Huang, Matt E. Hurles, Hanjun Jin, Luke Jostins, Thomas M. Keane, Si Quang Le, Sarah Lindsay, Quan Long, Daniel G. MacArthur, Stephen B. Montgomery, Leopold Parts, Chris Tyler-Smith, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Michael Snyder, Alexej Abyzov, Suganthi Balasubramanian, Robert Bjornson, Jiang Du, Fabian Grubert, Lukas Habegger, Rajini Haraksingh, Justin Jee, Ekta Khurana, Hugo Y. K. Lam, Jing Leng, Xinmeng Jasmine Mu, Alexander E. Urban, Zhengdong Zhang, Cristian Coafra, Huyen Dinh, Christie Kovar, Sandy Lee, Lynne Nazareth, Jane Wilkinson, Allison Coffey, Carol Scott, Neda Gharani, Jane S. Kaye, Alastair Kent, Taosha Li, Amy L. McGuire, Pilar N. Ossorio, Charles N. Rotimi, Yeyang Su, Lorraine H. Toji, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Assya Abdallah, Christopher R. Juenger, Nicholas C. Clemm, Audrey Duncanson, Eric D. Green, Mark S. Guyer, and Jane L. Peterson. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[53] James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature Methods*, 11(1):25–27, 2014.

[54] Sebastian H Eck. *Identification of genetic variation using Next-Generation Sequencing*. Phd thesis, Technische Universität München, 2014.

[55] Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, Anne-Lise Borresen-Dale, and Olli Kallioniemi. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology*, 12(1):1, 2011.

[56] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, 2009.

[57] Sven Enerbäck, Anders Jacobsson, Elizabeth M Simpson, Carmen Guerra, Hitoshi Yamashita, Mary-Ellen Harper, and Leslie P Kozak. Mice lacking mitochondrial uncoupling protein are cold-sensitive but not obese. *Nature*, 387(6628):90–94, 1997.

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

[59] Brent Ewing, Brent Ewing, Ladeana Hillier, Ladeana Hillier, Michael C Wendl, Michael C Wendl, Phil Green, and Phil Green. Base-Calling of Automated Sequencer Traces Using. *Genome Research*, (206):175–185, 2005.

[60] Brent Ewing, Brent Ewing, Ladeana Hillier, Ladeana Hillier, Michael C Wendl, Michael C Wendl, Phil Green, and Phil Green. Base-Calling of Automated Sequencer Traces Using. *Genome Research*, (206):175–185, 2005.

[61] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11):6–12, 2009.

[62] Nuno A Fonseca, John Marioni, and Alvis Brazma. RNA-seq gene profiling-a systematic empirical comparison. *PloS ONE*, 9(9):e107026, 2014.

[63] Simon A. Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott, and Peter J. Campbell. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, 2015.

[64] Richard W. Francis, Katherine Thompson-Wicking, Kim W. Carter, Denise Anderson, Ursula R. Kees, and Alex H. Beesley. Fusionfinder: A software tool to identify expressed gene fusion candidates from RNA-seq data. *PLoS ONE*, 7(6), 2012.

[65] Dimos Gaidatzis, Anita Lerch, Florian Hahne, and Michael B Stadler. QuasR: quantification and annotation of short reads in R. *Bioinformatics*, 31(7):1130–1132, 2015.

[66] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Andrew J Heron, Mark Bruce, Anthony Warland, Nadia Pantic, Tigist Admassu, Jonah Ciccone, Sabrina Serra, Samuel Martin, Luke Mcneill, Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Botond Sipos, Stephen Young, Sissel Juul, James Clarke, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*, page 068809, 2016.

[67] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–77, 2011.

[68] GATK Dev Team. Calling variants in RNAseq. https://www.broadinstitute.org/gatk/guide/topic?name=methods#methods3891, 2016. [Web page; last accessed 01.10.2016].

[69] David G. Ginzinger. Gene quantification using real-time quantitative PCR: An emerging technology hits the mainstream. *Experimental Hematology*, 30(6):503–512, 2002.

[70] Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, and Leroy Hood. Optimal scaling of digital transcriptomes. *PloS ONE*, 8(11):e77885, 2013.

[71] Ignacio González. Statistical analysis of RNA-Seq data. http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf, 2014. [Tutorial; last accessed 01.10.2016].

[72] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn a Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–52, 2011.

[73] Ayman Grada and Kate Weinbrecht. Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology*, 133(8), 2013.

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

[75] P A Greif, S H Eck, N P Konstandin, A Benet-Pages, B Ksienzyk, A Dufour, A T Vetter, H D Popp, B Lorenz-Depiereux, T Meitinger, and Others. Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia*, 25(5):821–827, 2011.

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

[77] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, 2010.

[78] Florian Halbritter, Harsh J Vaidya, and Simon R Tomlinson. GeneProf: analysis of high-throughput sequencing experiments. *Nature Methods*, 9(1):7–8, 2011.

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

[80] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):1–7, 2010.

[81] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G R Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis, and Roderic Guigo. GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, 7(1):1, 2006.

[82] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760—-1774, 2012.

[83] Stephen W Hartley and James C Mullikin. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics*, 16(1):224, 2015.

[84] Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673, 2012.

[85] Robert a Holt and Steven J M Jones. The new paradigm of flow cell sequencing. *Genome Research*, 18(6):839–846, 2008.

[86] Dongwan Hong, Arang Rhie, Sung-Soo Park, Jongkeun Lee, Young Seok Ju, Sujung Kim, Saet-Byeol Yu, Thomas Bleazard, Hyun-Seok Park, Hwanseok Rhee, and Others. FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics*, 28(5):721–723, 2012.

[87] Fan Hsu, James W. Kent, Hiram Clawson, Robert M. Kuhn, Mark Diekhans, and David Haussler. The UCSC known genes. *Bioinformatics*, 22(9):1036–1046, 2006.

[88] HTSeq. Counting reads in features with htseq-count. http://www-huber.embl.de/users/anders/HTSeq/doc/count.html, 2016. [Web page; last accessed 01.10.2016].

[89] Illumina, Inc. An Introduction to Next-Generation Sequencing Technology. http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf, 2015. [Web page; last accessed 01.10.2016].

[90] Illumina, Inc. Sequencing Systems. http://www.illumina.com/systems/sequencing.html, 2015. [Web page; last accessed 01.10.2016].

[91] Illumina, Inc. HiSeq 3000/HiSeq 4000 System quality and performance. http://www.illumina.com/systems/hiseq-3000-4000/specifications.html, 2016. [Web page; last accessed 01.10.2016].

[92] Illumina, Inc. Performance specifications for the HiSeq 2500 System. https://www.illumina.com/systems/sequencing-platforms/hiseq-2500/specifications.html, 2016. [Web page; last accessed 01.10.2016].

[93] International Human Genome Sequencing Consortium and others. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[94] Camilla L C Ip, Matthew Loose, John R Tyson, Mariateresa de Cesare, Bonnie L Brown, Miten Jain, Richard M Leggett, David A Eccles, Vadim Zalunin, John M Urban, and Others. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 4, 2015.

[95] W Janning and E Knust. *Genetik: Allgemeine Genetik, Molekulare Genetik, Entwicklungsgenetik*. Thieme, 2004.

[96] M Aleksi Kallio, Jarno T Tuimala, Taavi Hupponen, Petri Klemelä, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Käki, and Eija I Korpelainen. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, 12(1):507, 2011.

[97] Sara Kangaspeska, Susanne Hultsch, Henrik Edgren, Daniel Nicorici, Astrid Murumägi, and Olli Kallioniemi. Reanalysis of RNA-Sequencing Data Reveals Several Additional Fusion Genes with Multiple Isoforms. *PLoS ONE*, 7(10), 2012.

[98] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, 16(1):150, 2015.

[99] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(suppl 1):D493–D496, 2004.

[100] Yarden Katz, Eric T Wang, Jacob Silterra, Schraga Schwartz, Bang Wong, Jill P Mesirov, Edoardo M Airoldi, and Christopher B Burge. Sashimi plots: quantitative visualization of RNA sequencing read alignments. *arXiv preprint arXiv:1306.3466*, 2013.

[101] Yarden Katz, Eric T. Wang, Jacob Silterra, Schraga Schwartz, Bang Wong, Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov, Edoardo M. Airoldi, and Christopher B. Burge. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14):2400–2402, 2015.

[102] Ernest S Kawasaki. The end of the microarray Tower of Babel: will universal standards lead the way? *Journal of Biomolecular Techniques*, 17(3):200–6, 2006.

[103] W James Kent. BLAT – The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, 2002.

[104] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[105] C G Kevil, L Walsh, F S Laroux, T Kalogeris, M B Grisham, and J S Alexander. An improved, rapid Northern protocol. *Biochemical and Biophysical Research Communications*, 238(2):277–279, 1997.

[106] George A Khoury, Richard C Baliban, and Christodoulos A Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports*, 1:1–5, 2011.

[107] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):1, 2013.

[108] Daehwan Kim and Steven L Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):1, 2011.

[109] Martin Kircher, Daniela M Witten, Preti Jain, B J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[110] Julian C Knight. Allele-specific gene expression uncovered. *Trends in Genetics*, 20(3):113–116, 2004.

[111] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 2009.

[112] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D Mclellan, Ling Lin, Christopher a Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[113] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610–620, 2015.

[114] Laura S Kremer, Daniel M Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, and Others. Genetic diagnosis of Mendelian disorders via RNA sequencing. *bioRxiv*, page 66738, 2016.

[115] Shailesh Kumar, Angie Duy Vo, Fujun Qin, and Hui Li. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Scientific Reports*, 6, 2016.

[116] Nicholas F Lahens, Ibrahim Halil Kavakli, Ray Zhang, Katharina Hayer, Michael B Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S Thomas, Gregory R Grant, and John B Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biology*, 15(6):1, 2014.

[117] Ben Langmead, Kasper D Hansen, and Jeffrey T Leek. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology*, 11(8):1, 2010.

[118] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):1, 2009.

[119] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter a C ’t Hoen, Jean Monlong, Manuel a Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy,

Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013.

[120] Albert Lee, Kasper Daniel Hansen, James Bullard, Sandrine Dudoit, and Gavin Sherlock. Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genetics*, 4(12), 2008.

[121] Hayan Lee, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W. Richard McCombie, and Michael Schatz. Third-generation sequencing and the future of genomics. *bioRxiv*, page 048603, 2016.

[122] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, jan 2011.

[123] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

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

[125] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

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

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

[128] Peipei Li, Yongjun Piao, Ho Sun Shon, and Keun Ho Ryu. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16(1):347, 2015.

[129] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

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

[131] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, nov 2013.

[132] Robert Lindner and Caroline C. Friedel. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PLoS ONE*, 7(12):1–10, 2012.

[133] Stuart Lindsay. The promises and challenges of solid-state sequencing. *Nature Nanotechnology*, 11(2):109–111, 2016.

[134] Silvia Liu, Wei-Hsiang Tsai, Ying Ding, Rui Chen, Zhou Fang, Zhiguang Huo, SungHwan Kim, Tianzhou Ma, Ting-Yu Chang, Nolan Michael Priedigkeit, Adrian V Lee, Jianhua Luo, Hsei-Wei Wang, I-Fang Chung, and George C Tseng. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Research*, 44(5), 2015.

[135] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, 32(8):894–899, 2011.

[136] Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbNSFP v3. 0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, 37(3):235–241, 2016.

[137] Yuwen Liu, Jie Zhou, and Kevin P. White. RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

[138] Marc Lohse, Anthony M Bolger, Axel Nagel, Alisdair R Fernie, John E Lunn, Mark Stitt, and Björn Usadel. RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40:622–627, 2012.

[139] Loman Labs. Nanopore R9 rapid run data release. `http://lab.loman.net/2016/07/30/nanopore-r9-data-release/`, 2016. [Web page; last accessed 01.10.2016].

[140] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15(12):1, 2014.

[141] Jakob Lovén, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee, and Richard A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

[142] Jun Lu, John K Tomfohr, and Thomas B Kepler. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165, 2005.

[143] Gerton Lunter and Martin Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011.

[144] Weijun Luo and Cory Brouwer. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[145] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):1, 2009.

[146] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, and Others. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[147] Alberto Magi, Betti Giusti, and Lorenzo Tattini. Characterization of MinION nanopore data for resequencing analyses. *Briefings in Bioinformatics*, page bbw077, 2016.

[148] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, 2009.

[149] Christopher A Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, Shujun Luo, Irina Khrebtukova, Terrence R Barrette, Catherine Grasso, Jindan Yu, Robert J Lonigro, Gary Schroth, Chandan Kumar-Sinha, and Arul M Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12353–12358, 2009.

[150] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261, 2003.

[151] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188, 2012.

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

[153] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C.

Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

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

[155] Adam Mark, Ryan Thompson, and Chunlei Wu. *mygene: Access MyGene.Info_ services*, 2014. R package version 1.6.0.

[156] Hideo Matsumura, Akiko Ito, Hiromasa Saitoh, Peter Winter, Günter Kahl, Monika Reuter, Detlev H. Krüger, and Ryohei Terauchi. SuperSAGE. *Cellular Microbiology*, 7(1):11–18, 2005.

[157] Allan M Maxam and Walter Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, 1977.

[158] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Others. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[159] Kevin J McKernan, Heather E Peckham, Gina L Costa, Stephen F McLaughlin, Yutao Fu, Eric F. Tsung, Christopher R Clouser, Cisyla Duncan, Jeffrey K Ichikawa, Clarence C Lee, Zheng Zhang, Swati S Ranade, Eileen T Dimalanta, Fiona C Hyland, Tanya D Sokolsky, Lei Zhang, Andrew Sheridan, Haoning Fu, Cynthia L Hendrickson, Bin Li, Lev Kotler, Jeremy R Stuart, Joel a. Malek, Jonathan M Manning, Alena A Antipova, Damon S Perez, Michael P Moore, Kathleen C Hayashibara, Michael R Lyons, Robert E Beaudoin, Brittany E Coleman, Michael W Laptewicz, Adam E Sannicandro, Michael D Rhodes, Rajesh K Gottimukkala, Shan Yang, Vineet Bafna, Ali Bashir, Andrew MacBride, Can Alkan, Jeffrey M Kidd, Evan E Eichler, Martin G Reese, Francisco M De La Vega, and Alan P Blanchard. Sequence and structural variation in a human

genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19(9):1527–1541, 2009.

[160] William T. Melvin, Helen B. Milne, Alison A. Slater, Hamish J. Allen, and Hamish M. Keir. Incorporation of 6-thioguanosine and 4-thiouridine into rna. *European Journal of Biochemistry*, 92(2):373–379, 1978.

[161] Michael L Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[162] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245, 2007.

[163] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. `http://cgap.nci.nih.gov/Chromosomes/Mitelman`, 2016. [Web page; last accessed 01.10.2016].

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

[165] Jan Murken. *Humangenetik*. Georg Thieme Verlag, 2006.

[166] U Nagalakshmi, Z Wang, K Waern, C Shou, D Raha, M Gerstein, and M Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[167] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):1–13, 2011.

[168] Tal Nawy. Single-cell sequencing. *Nature Methods*, 11(1):18–18, 2014.

[169] Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Evan E Eichler, Deborah A Nickerson, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Michael Bamshad, and Jay Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.

[170] Zemin Ning, Zemin Ning, Anthony J Cox, Anthony J Cox, James C Mullikin, and James C Mullikin. SSAHA: A Fast Search Method for Large DNA Databases. *Genome Research*, (2):1725–1729, 2001.

[171] Kazuko Nishikura. Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry*, 79:321, 2010.

[172] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, 40(20):10084–10097, 2012.

[173] Timothy D O'Brien, Peilin Jia, Junfeng Xia, Uma Saxena, Hailing Jin, Huy Vuong, Pora Kim, Qingguo Wang, Martin J Aryee, Mari Mino-Kenudson, Jeffrey A Engelman, Long P Le, A. John Iafrate, Rebecca S Heist, William Pao, and Zhongming Zhao. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods*, 83:118–127, 2015.

[174] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[175] Michał J Okoniewski and Crispin J Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7(1):276, 2006.

[176] Rikke K J Olsen, Eliška Koňaříková, Teresa A. Giancaspero, Signe Mosegaard, Veronika Boczonadi, Lavinija Mataković, Alice Veauville-Merllié, Caterina Terrile, Thomas Schwarzmayr, Tobias B. Haack, Mari Auranen, Piero Leone, Michele Galluccio, Apolline Imbard, Purificacion Gutierrez-Rios, Johan Palmfeldt, Elisabeth Graf, Christine Vianey-Saban, Marcus Oppenheim, Manuel Schiff, Samia Pichard, Odile Rigal, Angela Pyle, Patrick F. Chinnery, Vassiliki Konstantopoulou, Dorothea Möslinger, René G. Feichtinger, Beril Talim, Haluk Topaloglu, Turgay Coskun, Safak Gucer, Annalisa Botta, Elena Pegoraro, Adriana Malena, Lodovica Vergani, Daniela Mazzà, Marcella Zollino, Daniele Ghezzi, Cecile Acquaviva, Tiina Tyni, Avihu Boneh, Thomas Meitinger, Tim M. Strom, Niels Gregersen, Johannes A. Mayr, Rita Horvath, Maria Barile, and Holger Prokisch. Riboflavin-Responsive and -Non-responsive Mutations in FAD Synthase Cause Multiple Acyl-CoA Dehydrogenase and Combined Respiratory-Chain Deficiency. *American Journal of Human Genetics*, 98(6):1130–1145, 2016.

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

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

[179] Marion Ouedraogo, Charles Bettembourg, Anthony Bretaudeau, Olivier Sallou, Christian Diot, Olivier Demeure, and Frédéric Lecerf. The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes. *PLoS ONE*, 7(11):1–8, 2012.

[180] Oxford Nanopore Technologies. Company history. `https://nanoporetech.com/about-us/history`, 2016. [Web page; last accessed 01.10.2016].

[181] Oxford Nanopore Technologies. Media resources and contacts. `https://nanoporetech.com/about-us/for-the-media`, 2016. [Web page; last accessed 01.10.2016].

[182] Oxford Nanopore Technologies. PromethION. `https://nanoporetech.com/products/promethion`, 2016. [Web page; last accessed 01.10.2016].

[183] Oxford Nanopore Technologies. Rapid sequencing with MinION. `https://nanoporetech.com/rapidsequencing`, 2016. [Web page; last accessed 01.10.2016].

[184] Oxford Nanopore Technologies. SmidgION. `https://nanoporetech.com/products/smidgion`, 2016. [Web page; last accessed 01.10.2016].

[185] Oxford Nanopore Technologies. Types of nanopores. `https://nanoporetech.com/how-it-works/types-of-nanopores`, 2016. [Web page; last accessed 01.10.2016].

[186] Pacific Biosciences of California, Inc. SMRT CELLS, SEQUENCING REAGENT KITS, AND ACCESSORIES. `http://www.pacb.com/products-and-services/consumables/sequel-consumables/smrt-cells-sequencing-reagent-kits-and-accessories/`, 2016. [Web page; last accessed 01.10.2016].

[187] Pacific Biosciences of California, Inc. SMRT SEQUENCING: READ LENGTHS. `http://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/`, 2016. [Web page; last accessed 01.10.2016].

[188] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008.

[189] Eddie Park, Brian Williams, Barbara J. Wold, and Ali Mortazavi. RNA editing in the human ENCODE RNA-seq data. *Genome Research*, 22(9):1626–1633, 2012.

[190] Peter J Park. Chip-seq advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[191] G Parmigiani, E S Garett, R A Irizarry, and S L Zeger. *The Analysis of Gene Expression Data: Methods and Software*. Statistics for Biology and Health. Springer New York, 2006.

[192] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, 2014.

[193] Zhiyu Peng, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, Yu Liang, Xueda Hu, Xuemei Tan, Jing Guo, Zirui Dong, Yan Liang, Li Bao, and Jun Wang. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature Biotechnology*, 30(3):253–260, 2012.

[194] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181, 2014.

[195] J K Pickrell, J C Marioni, A A Pai, J F Degner, B E Engelhardt, E Nkadori, J B Veyrieras, M Stephens, Y Gilad, and J K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.

[196] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics*, 93(4):641–651, 2013.

[197] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, and Others. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053–1058, 2014.

[198] A. J. Pratt and I. J. MacRae. The RNA-induced Silencing Complex: A Versatile Gene-silencing Machine. *Journal of Biological Chemistry*, 284(27):17897–17901, 2009.

[199] Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H. J. Baum, Beate Becker-Ziaja, Jan Peter Boettcher, Mar Cabeza-Cabrerizo, Álvaro Camino-Sánchez, Lisa L. Carter, Juliane Doerrbecker, Theresa Enkirch, Isabel García-Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigael Kosgey, Eeva Kuisma, Christopher H. Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo, Katja Nitzsche, Elisa Pallasch, Livia Victoria Patrono, Jasmine

Portmann, Johanna Gabriella Repits, Natasha Y. Rickett, Andreas Sachse, Katrin Singethan, Inês Vitoriano, Rahel L. Yemanaberhan, Elsa G. Zekeng, Trina Racine, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'Faly Magassouba, Cecelia V Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Frank Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J. Williams, Facinet Yattara, Kuiama Lewandowski, James Taylor, Phillip Rachwal, Daniel J Turner, Georgios Pollakis, Julian A. Hiscox, David A. Matthews, Matthew K O'Shea, Andrew McD. Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Wölfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A. Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keïta, Andrew Rambaut, Pierre Formenty, Stephan Günther, and Miles W. Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[200] Emma M Quinn, Paul Cormican, Elaine M Kenny, Matthew Hill, Richard Anney, Michael Gill, Aiden P Corvin, and Derek W Morris. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS ONE*, 8(3), 2013.

[201] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[202] Gokul Ramaswami and Jin Billy Li. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, 42(D1):109–113, 2014.

[203] Gokul Ramaswami and Jin Billy Li. Identification of human RNA editing sites: A historical perspective. *Methods*, 107:42–47, 2016.

[204] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory a Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, 2012.

[205] Daniel Ramsköld, Eric T. Wang, Christopher B. Burge, and Rickard Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology*, 5(12):1–11, 2009.

[206] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):1, 2013.

[207] Rahul Reddy. A Comparison of Methods : Normalizing High-Throughput RNA Sequencing Data. *bioRxiv*, page 026062, 2015.

[208] Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. GenePattern 2.0. *Nature Genetics*, 38(5):500–501, 2006.

[209] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289, 2015.

[210] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9), 2014.

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

[212] Christelle Robert and Mick Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, 16(1):177, 2015.

[213] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[214] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, jan 2010.

[215] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1, 2010.

[216] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[217] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

[218] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.

[219] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):227–240, 2010.

[220] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[221] Sven Schuierer and Guglielmo Roma. The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. *Nucleic Acids Research*, 44(16):e132–e132, 2016.

[222] Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, and Others. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, 2016.

[223] Fatemeh Seyednasrollah, Asta Laiho, and Laura L Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70, dec 2013.

[224] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–30, 2013.

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

[226] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[227] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–32, 2014.

[228] Maxine F Singer. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell*, 28(3):433–434, 1982.

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

[230] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):1, 2013.

[231] Daniel Spies and Constance Ciaudo. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Computational and Structural Biotechnology Journal*, 13:469–477, 2015.

[232] Stanford Encyclopedia of Philosophy. Genomics and postgenomics. `https://plato.stanford.edu/entries/genomics/`, October 2016. [Online; last accessed 16.03.2017].

[233] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[234] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Shaw, Andrew D Phillips, and David N Cooper. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9, 2014.

[235] David Stoddart, Andrew J Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19):7702–7, 2009.

[236] T Strachan and A P Read. *Human Molecular Genetics 3*. Garland Science, 2004.

[237] Zhenqiang Su, Paweł P Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P Schroth, Robert a Setterquist, John F Thompson, Wendell D Jones, Wenzhong Xiao, Weihong Xu, Roderick V Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J Oakeley, Zhiyu Peng, Craig a Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K Smyth, Frank Staedtler, Peter Sykacek, Xin-Xing Tan, E Aubrey Thompson, Jo Vandesompele, May D Wang, Jian Wang, Russell D Wolfinger, Jiri Zavadil, Scott S Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H Brilliant, Pierre R Bushel, Weimin Cai, Jennifer G Catalano, Ching-Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu-Ming Chu, Djork-Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li-Wu Guo, James Hadfield, Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan-Zhen Li, Li Li, Peng Li, Yan Li, Haiqing Li, Jianying Li, Shiyong Li, Simon M Lin, Francisco J López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G Perkins, Ryan Peters, John H Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure Sambourg, Susanna-Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T Szkotnicki, Matthew Tinning, Bimeng Tu, Joost van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D Wieben, James C Willey, Po-Yen Wu, Jiekun Xuan,

Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate-Ching Yuan, John Zhang, Ke K Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P Kreil, Christopher E Mason, and Leming Shi. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9), 2014.

[238] Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-laure Yaspo. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptomeof the Human Transcriptome. *Science*, 321(5891):956–960, 2008.

[239] Peter a C 't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen F J Laros, Henk P J Buermans, Olof Karlberg, Mathias Brännvall, Johan T den Dunnen, Gert-Jan B van Ommen, Ivo G Gut, Roderic Guigó, Xavier Estivill, Ann-Christine Syvänen, Emmanouil T Dermitzakis, and Tuuli Lappalainen. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology*, 31(11):1015–1022, 2013.

[240] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

[241] Sonia Tarazona, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, page gkv711, 2015.

[242] Muhammad A Tariq, Hyunsung J Kim, Olufisayo Jejelowo, and Nader Pourmand. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, 39(18):e120–e120, 2011.

[243] Mingxiang Teng, Michael I Love, Carrie A Davis, Sarah Djebali, Alexander Dobin, Brenton R Graveley, Sheng Li, Christopher E Mason, Sara Olson, Dmitri Pervouchine, Cricket A Sloan, Xintao Wei, Lijun Zhan, and Rafael A Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17(1):74, 2016.

[244] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[245] Wandaliz Torres-García, Siyuan Zheng, Andrey Sivachenko, Rahulsimham Vegesna, Qianghu Wang, Rong Yao, Michael F Berger, John N Weinstein, Gad Getz, and G W Roel. PRADA : Pipeline for RNA sequencing Data Analysis. *Bioinformatics*, 30(15):2224–2226, 2014.

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

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

[248] Kevin J Travers, Chen Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, 38(15):e159, 2010.

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

[250] Taketoshi Uzawa, Akihiko Yamagishi, and Tairo Oshima. Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, Thermus thermophilus HB27 and sulfolobus tokodaii strain 7. *Journal of Biochemistry*, 131(6):849–853, 2002.

[251] Victor E Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W Kinzler. Serial Analysis of Gene Expression. *Science*, 270(5235):484–487, 1995.

[252] Gabriel Wajnberg and Fabio Passetti. Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery. *Expert Opinion on Drug Discovery*, 11(3):257–268, 2016.

[253] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[254] Liguo Wang, Shengqin Wang, and Wei Li. RSeQC: Quality Control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012.

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

[256] Thomas Wieland. *Next-Generation Sequencing Data Analysis*. Phd thesis, Technische Universität München, 2015.

[257] Brian T Wilhelm and Josette-Renée Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*, 48(3):249–57, 2009.

[258] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

[259] Lukas Windhager, Thomas Bonfert, Kaspar Burger, Zsolt Ruzsics, Stefan Krebs, Stefanie Kaufmann, Georg Malterer, Anne L'Hernault, Markus Schilhabel, Stefan Schreiber, Philip Rosenstiel, Ralf Zimmer, Dirk Eick, Caroline C Friedel, and Lars Dölken. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research*, 22(10):2031–2042, 2012.

[260] Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources*, 13(4):559–572, 2013.

[261] Markus Wolfien, Christian Rimmbach, Ulf Schmitz, Julia Jeannine Jung, Stefan Krebs, Gustav Steinhoff, Robert David, and Olaf Wolkenhauer. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics*, 17(1):21, 2016.

[262] Po-Yen Wu, John H Phan, and May D Wang. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*, 14(11):1, 2013.

[263] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 2005.

[264] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):1, 2010.

[265] Weihua Zeng and Ali Mortazavi. Technical considerations for functional sequencing assays. *Nature Immunology*, 13(9):802–807, 2012.

[266] Zong Hong Zhang, Dhanisha J. Jhaveri, Vikki M. Marshall, Denis C. Bauer, Janette Edson, Ramesh K. Narayanan, Gregory J. Robinson, Andreas E. Lundberg, Perry F. Bartlett, Naomi R. Wray, and Qiong Yi Zhao. A comparative study of techniques for differential expression analysis on RNA-Seq data comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE*, 9(8):e103207, 2014.

[267] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS ONE*, 9(1):e78644, 2014.

[268] Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):1–14, 2015.

[269] Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3):246–251, 2014.