High-throughput RNA sequencing: a step forward in transcriptome analysis

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
vorgelegt von

Konstantin Okonechnikov

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin 2016

Datum des Disputation: ***09.02.2016***

Gutachter:
Prof. Dr. Knut Reinert, *Freie Universität, Deutschland*
Prof. Dr. Steven Salzberg, *Johns Hopkins School of Medicine, U.S.A*
Dr. Fernando García-Alcalde, *Roche Innovation Center, Switzerland*

Abstract

The transcriptome plays an important role in the life of a cell. Detailed analysis of the transcriptome enables interpretation of its structure and functionality. High throughput sequencing technology significantly enhanced the understanding of transcriptome activity. The RNA-sequencing process currently provides the most accurate estimation of gene expression levels. Moreover, RNA-seq allows detection of isoform structure and novel RNA types along with transcription process details such as strand-specificity and much more. The first chapter of this thesis describes the history of transcriptome exploration and effective methods of RNA-seq application.

Nevertheless, all steps of RNA-seq process can produce a number of biases that influence the investigation results. Some typical errors appearing during ligation and amplification procedures might be present in any high throughput sequencing experiment, while other biases occur only in cDNA synthesis or are specific for transcriptome activity. Quality control of sequencing data is important to verify and correct the analysis results. The second chapter of this thesis is devoted to the explanation of these issues and introduces a novel tool, Qualimap2. This instrument computes detailed statistics and presents a number of plots based on RNA-seq alignment and counts data processing. The generated results enable detection of problems that are specific to RNA-seq experiments. Notably, the tool supports analysis of multiple samples in various conditions. Qualimap2 was faithfully compared to other available tools and demonstrated superior functionality in multi-sample quality control.

Importantly, RNA-seq can be applied in a relatively novel research area: detection of chimeric transcripts and fusion genes occurring due to genomic rearrangement. Since fusions are related to cancer, their discovery is important not only for science, but also allows medical use of RNA-seq. The third chapter is devoted to the current status of this approach and illustrates a novel toolkit called InFusion, which provides a number of novelties in chimera discovery from RNA-seq data such as detection of fusions arising from the combination of a gene and an intronic or intergenic region. Moreover, strand-specificity of expressed fusion transcripts can be detected and reported. InFusion was compared in detail to a number of other existing tools based on simulated and real datasets and demonstrated higher precision and recall.

Overall, RNA-sequencing technology goes further and more specialized analysis abilities are becoming available. New applications of RNA sequencing and future directions of research are discussed in the last chapter.

Zusammenfassung

Die Transkription ist ein wichtiger Prozess in biologischen Zellen. Eine genaue Analyse des Transkriptomes eröffnet die Möglichkeit seine Struktur und Funktionen auf neue Weise zu interpretieren. Hochdurchsatzsequenzierunsmethoden haben das Verständnis der Veränderungen im Transkriptom signifikant erhöht. Die RNA-Sequenzierung ist im Moment die akkurateste Methode zur Bestimmung von Genexpressionsniveaus. Weiterhin erlaubt RNA-Seq die Bestimmung von Transkriptisoformen sowie neuen RNA-Formen zusammen mit notwendigen Details, wie unter anderem Strandspezifität. Das erste Kapitel der Dissertation beschreibt die Geschichte der Erforschung des Transkriptoms und effektive Methoden für die Anwendung von RNA-Seq.

In allen Abschnitten des RNA-Seq Prozesses kann es zur Verzerrung der wissenschaftlichen Ergebnisse durch verschiedene Störfaktoren kommen. Einige typische Fehler, z.B. während der Ligation und Amplifizierung sind dabei allen Hochdurchsatzsequenzierungsmethoden gemein, während andere spezifisch bei der Erstellung der RNA-Bibliotheken auftreten oder durch die Eigenschaften des Transkriptoms bedingt sind. Eine entsprechende Qualitätskontrolle ist daher wichtig um Analyseergebnisse zu kontrollieren und zu korrigieren. Das zweite Kapitel dieser Arbeit widmet sich der Beschreibung relevanter Parameter der Qualitätskontrolle und führt als neues Werkzeug Qualimap2 ein. Diese Software berechnet detaillierte Statistiken und generiert eine Anzahl von aussagekräftigen Diagrammen auf der Basis von RNA-Seq Alignments, wodurch für diese Anwendung typische Probleme erkannt werden können. Insbesondere erlaubt das Programm den Vergleich mehrerer Proben aus verschiedenen Bedingungen. Qualimap2 wurde ausgiebig mit ähnlicher Software verglichen und zeigt eine bessere Funktionalität für die Qualitätskontrolle mehrerer Proben.

RNA-Seq kann zur Detektion von bisher unbekannten Transkripten benutzt werden, so z.B. zur Detektion von Transkriptchimären und Fusionsgenen, die bei genomischen Rearrangements entstehen. Da Fusionen häufig in Tumorzellen auftreten, ist ihre Bestimmung nicht nur aus wissenschaftlichen Gründen relevant sondern zeigt auch die medizinische Relevanz von RNA-Seq. Das dritte Kapitel widmet sich der Beschreibung des derzeitigen Kenntnisstands dieses Gebietes und beschreibt mit InFusion ein neue Softwaremethode, die eine Reihe von neuen Ansätzen für die Detektion von chimärischen Transkripten auf der Basis von RNA-Seq Daten wie zum Beispiel die Erkennung von Fusionen mit intronischen und intergenischen Regionen. Weiterhin kann die Strand-Spezifizität der exprimierten Fusionstranskripte erkannt und ausgegeben werden. InFusion wurde mit mehreren existierenden Tools auf der Basis von simulierten und realen Datensätzen verglichen und dabei zeigt eine bessere Präzision und Sensitivität.

Mit dem Fortschritt der RNA-Sequezierungsmethoden werden zunehmend spezialisiertere Analysen möglich. Diese Entwicklungen der RNA-Seq Technologie und neue Forschungsrichtungen werden im letzten Kapitel besprochen.

Acknowledgments

I would like to express my most sincere gratitude to Dr. Fernando García-Alcalde, who was helping me to become a scientist in the world of computational biology. His rich perspective on bioinformatics was supporting me to establish the research goals, fulfill them appropriately and obtain important results.

Bioinformatics cannot exist without biology. Performing work in Max Planck Institute for Infection Biology was necessary for me to realize correctly the valid research objectives. Thanks a lot to Prof. Thomas F. Meyer for providing this opportunity.

Importantly, the world of bionformatics is rather deep and full of technological and algorithmic blocks. It is not possible to get the proper status of existing research directions without perfect perceptive of the whole area. I would like to express my high gratitude to Prof. Knut Reinert for providing the detailed overview of far-reaching aspects in computational biology.

Detailed external view on the performed research work allows to detect undefined issues and improve the results. I would like to thank Prof. Steven Salzberg for his agreement to perform the assessment of my thesis.

The accurate answers to multiple questions in molecular biology are required to perform analysis and contribute correctly to ongoing research. A number of thanks to biologists who I was working with: Frithjof Glowinski, Max Koeppel, Alexander Karlas, Pau Morey. Additionally, when I was obtaining analysis skills and performing development of novel algorithms, serious help in bioinformatics I got from Hilmar Berger and SeqAn developers group. Thanks a lot to them. Importantly, I was working on open-source software projects, and many users were reporting bugs and providing useful suggestions. I would like to express my appreciation to everyone.

Writing a manuscript after accomplishing a lot of research work sounds like an easy task, however correct adaptation of the text is a quite complicated art. I would like to thank Rike Zietlow for performing this duty.

To have a full understanding of research topics it is quite valuable to be able to present them and share the knowledge with the others. Thanks a lot to Nikolay Vyahhi and Ekaterina Chaykina for providing the opportunity to give lectures and seminars during Bioinformatics Summer School events in Russia.

After performing hard work it is quite important to have breaks and clear the mind. For help in this aspect my thanks to Georg Petkau, Piotr Zadora, Maria Del Mar Reines, Ludovico Sepe, Ana Rita da Costa, Danil Chekushin, Laura Martin, Francesco Boccellato, Amina Iftekhar... Additional thanks to Arkady Urkop and Olga Amelkina for external view on biology. Sorry if I missed someone.

And of course, thanks a lot to my parents for supporting me during a hard, but exciting scientific life time period.

”Nature is a harmonious mechanism where all parts, including those appearing to play a secondary role, cooperate in the functional whole. In contemplating this mechanism, shallow men arbitrarily divide its parts into essential and secondary, whereas the insightful thinker is content with classifying them as understood and poorly understood, ignoring for the moment their size and immediately useful properties. No one can predict their importance in the future.”

Santiago Ramón y Cajal, "Advice for a Young Investigator"

CONTENTS

1. Introduction . . . 1

1.1 Overview . . . 1

1.2 Molecular biology and bioinformatics . . . 1

1.2.1 Central dogma of molecular biology . . . 1

1.2.2 Intricacy of transcriptome . . . 3

1.2.3 Analysis of transcriptome . . . 5

1.3 High-throughput sequencing . . . 5

1.3.1 Initial approaches . . . 5

1.3.2 RNA sequencing . . . 6

1.4 Transcriptome analysis using RNA-sequencing . . . 7

1.4.1 Data processing overview . . . 7

1.4.2 Quality control of sequencing data . . . 8

1.4.3 Primary analysis . . . 9

1.4.4 Gene expression analysis . . . 10

1.5 Additional aspects of RNA-seq data analysis . . . 11

1.5.1 Isoform detection . . . 11

1.5.2 Strand-specificity . . . 12

1.5.3 Gene fusion detection . . . 13

1.6 Research goals . . . 13

2. Quality control of RNA-seq alignment and expression data . . . 15

2.1 Overview . . . 15

2.2 Reasons and approaches . . . 15

2.3 Qualimap2: advanced RNA-seq and counts quality control . . . 17

2.3.1 Tool description . . . 17

2.3.2 RNA-seq QC mode . . . 17

2.3.3 Counts QC mode . . . 20

2.3.4 Insight from quality control analysis . . . 25

2.3.5 Comparison to other tools . . . 27

2.3.6 Results and future plans . . . 28

2.4 Summary . . . 28

3. Fusion gene and chimeric transcript detection . . . 31

3.1 Overview . . . 31

3.2 Fusion discovery approaches . . . 31

3.2.1 History . . . 31

3.2.2 Detailed RNA-seq approach explanation . . . 32

3.2.3 Existing tools and their limitations . . . 33

3.3 InFusion: novel fusion discovery toolkit 34
3.3.1 Introduction 34
3.3.2 Pipeline description 35
3.3.3 Comparison to other existing tools: simulation 42
3.3.4 Comparison to other existing tools: public datasets 46
3.3.5 Experimental validation 47
3.3.6 Unanswered questions 51
3.4 Summary 52

4. Discussion 55

A. Fusion discovery tools options 59
A.1 General 59
A.2 Simulated data analysis 59
A.3 Public datasets 61
A.4 In-house datasets 61

B. Fusion discovery supplementary tables 63
B.1 Edgren et al. dataset 63
B.2 Berger et al. dataset 64
B.3 Wu et al. dataset 64
B.4 VCaP and LNCaP cell lines 65

C. List of abbreviations 67

D. Curriculum Vitae 69

E. Declaration 73

Bibliography 74

Chapter 1 Introduction

1.1 Overview

In this chapter an introduction to the theoretical bases of biology and bioinformatics are given. The chapter starts with a description of important molecular biology aspects and explains the need for specific analysis algorithms to produce accurate insights into biological processes. Further, it focuses on transcriptome investigation, defining accomplished blocks of this work and current research procedures. A detailed explanation of high-throughput sequencing technology and RNA-seq analysis procedure are given. Finally, an overview of active research projects and explanation of further chapters are provided.

1.2 Molecular biology and bioinformatics

1.2.1 Central dogma of molecular biology

Biology is a sophisticated science that ranges from the chemical construction of molecules that create biological processes to investigation of complex activity of all living creatures. Organisms have large differences in classification: bacteria, fungi, plants, animals. However, each organism starts with the development from a cell. For example, the human body includes several hundreds of distinct cell types and in total it consists of approximately $3.72 * 10^{13}$ cells [Bianconi *et al.*, 2013]. Even though there are many different types of cells with various activity in the human body, cells are packed in blocks of the same type and work together to perform certain functions of the organism. Detailed investigation of cell structure and activity is a main aspect of molecular biology.

There are special chemical rules that control each cell life cycle. The process starts from deoxyribonucleic acid (DNA), which can be detected in all cells of all organisms. DNA is the "holder" of all information of cell activity. From the DNA the activity instructions are transfered with ribonucleic acid (RNA). Finally RNAs are converted to proteins. Proteins along with several types of RNAs perform all cellular functions. The theory of activity between DNA, RNA and proteins was first introduced by Francis Crick [Crick, 1958] and called *central dogma of molecular biology*. After the initiation of the dogma, further research works allowed to qualify the cell system in more detail.

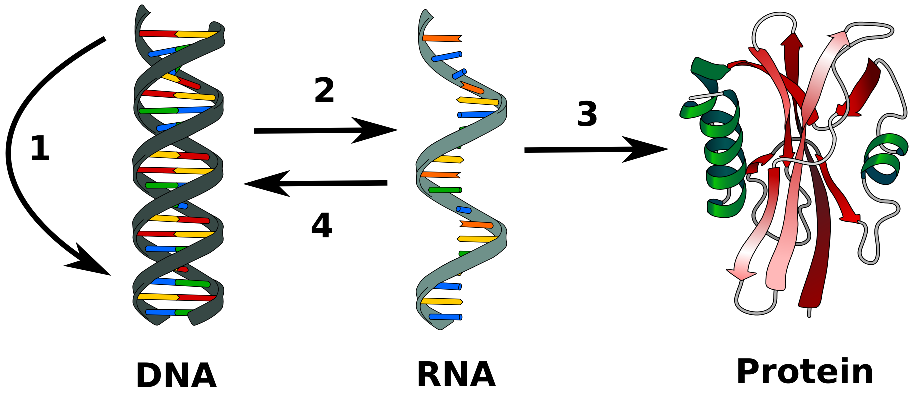

Figure 1.1: *Central dogma of molecular biology*

Main processes of cell activity are the following:

1. DNA maintains all the cell information via the **replication process**, where DNA sequence is copied to produce a nearly identical molecule

2. **Transcription** of specific DNA fragments to RNA is a temporary copy of DNA used to create proteins and also play distinct functional roles in translational apparatus

3. RNAs are **translated** to proteins, which perform all structural functions and play regulatory roles.

4. **Reverse transcription** of RNA is a template for the synthesis of DNA applied for pseudogene replication and other types of transposition.

The first publication that reported identification of DNA structure and activity [Watson *et al.*, 1953] led to an acceleration of research progress in all aspects of the cell system and its activity in connection to organism. The research produced a detailed understanding of various biological processes beginning with DNA. The organization of DNA in cells is currently considered at specific structural levels: nucleotides, sequences, chromosomes and genome. A nucleotide is composed of a monosaccharide sugar called deoxyribose, a 5'phosphate group and a specific nucleobase: cytosine , guanine , adenine or thymine. Each nucleotide (also called base) is connected to a specific partner,forming a double strand: adenine to thymine, cytosine to guanine. The DNA information is collected in a sequence of nucleotides. Chromosomes are physical blocks, containing DNA. The genome is a full collection of all DNA in a cell.

In RNA molecules the thymine nucleobase is replaced with uracil and an oxygen atom is added to the sugar component during the transcription process.

These two small differences have a major impact on the biological role of the molecule. RNA is more chemically active and it usually consists of a single strand of DNA. There are two types of its initial genomic location: forward or reverse strand. An easy way to maintain information about DNA or RNA is with a string of four symbols:

$$\{A, C, G, T\}$$

with left-to-right orientation corresponding to 5' to 3' polarity in case of DNA and correct strand of transcription in case of RNA. Using this technique it is possible to keep information about fragments of DNA chromosomes or expressed RNA sequences to perform further investigations.

From the start molecular biology had a requirement for detailed analysis of cell processes and functions. The first experiments that allowed detection of sequences of DNA and RNA required approaches to understand the translation of DNA to proteins, discovery of similarity between sequences, and much more. To accomplish these tasks, special computational algorithms were required. Additionally, the quantity of novel scientific results was growing and special data collection systems were important. Because of these reasons bioinformatics started its progress. Good examples of the initial computational approaches were conversion of DNA sequences to protein or the Smith-Waterman algorithm for detection of similar DNA/RNA sequences [Smith and Waterman, 1981].

1.2.2 Intricacy of transcriptome

RNA forms a connecting block between DNA and protein in a cell life, however it plays a huge role in the data transfer to perform functional processes of a cell. DNA segments that contain information about cell functions - *genes* - are expressed as RNA molecules - *transcripts*. Additionally there are special RNAs that participate in the translational apparatus and have specific functions. The *transcriptome* is a complete set of transcripts in a cell or tissue. The detailed spectrum of transcriptome functionality regulates the selection of expressed genes in different cell types and changes the activity according to external conditions. [Maniatis and Reed, 2002]. Moreover, transcriptome functions are important for health and incorrect RNA process might lead to disease [Mitelman *et al.*, 2007].

There is a number of RNA types with different functionality forming the transcriptome. Best known are the elements participating in the protein synthesis:

- Messenger RNA (mRNA)

  mRNA is transcribed from genes and used to construct a protein. It carries genetic information specific for the activity of a cell type. Even though the transcripts are crucial, they make up only 5%of the total transcriptome. During the transcription process a *precursor mRNA* molecule is synthesized and a poly-A tail of about 200 bp is added to the 3' end. It is worth noting that mRNA in higher eukaryotes (including Homo sapiens) contains non-coding segments of mRNA called *introns*. These blocks are extracted and

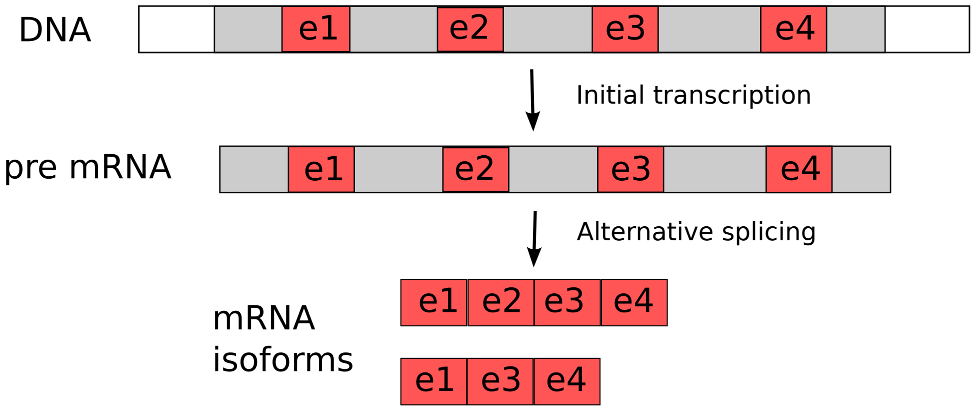

Figure 1.2: *Transcription process in high eukaryotes*

only the remaining segments - *exons* - are applied to construct the protein. The whole process is called splicing (Figure 1.2). Importantly, genes might have several combinations of selected exons (so called alternative splicing) and this leads to several different isoforms of a gene. For example, in human cells there are around six transcripts per protein-coding gene on average [Consortium *et al.*, 2012].

- Ribosomal RNA (rRNA)

  rRNA is located in the cytoplasm of a cell, where ribosomes are found. rRNA directs the translation of mRNA into proteins. This is the largest part of the transcriptome and essential for protein synthesis in all living organisms.

- Transfer RNA (tRNA)

  tRNA is located in the cellular cytoplasm and also involved in protein synthesis. tRNA brings or transfers amino acids to the ribosome that correspond to each of the three-nucleotide codon of rRNA. The amino acids then can be joined together and processed to make polypeptides and proteins.

Additionally, a number of specific RNA types are present in eukaryotes. Even though these RNAs form less than 1% of total RNA, they might perform important cell regulatory operations. Several distinct RNAs such as small nuclear RNA together with ribonuclease participate in post-transcription modification and processing of pre-mRNA [Mamatis, 1987]. Long non-coding RNAs regulate gene transcription [Rinn and Chang, 2012]. Small interfering RNA and microRNA are responsible for regulation of gene expression [Ambros, 2004]. Antisense RNA block mRNA translation and expression [Brantl, 2007].

1.2.3 Analysis of transcriptome

The first experiments to analyze the transcriptome were performed using methods similar to DNA analysis. The extraction of RNA from a cell followed by amplification and conversion of RNA to complementary DNA (cDNA) allowed the application of a standard method called reverse transcription polymerase chain reaction (RT-PCR), that has become one of the fundamental technologies in molecular biology [Mullis *et al.*, 1992]. Such experiments provide confirmation of the presence of a certain expressed transcript or other RNA. The analysis can be performed only if a specific primer pair with sequence blocks appropriate to the target are available. The next important step was the development of the quantitative RT-PCR technique [Becker-Andre and Hahlbrock, 1989], which allows measurement of the gene expression level with high sensitivity. The limitation of this method lay in a restricted number of genes to test and labor-intensive work.

A big advance in DNA and RNA analysis was Sanger sequencing [Sanger *et al.*, 1977]. This method is based on the selective incorporation of chain-terminating nucleotides by the enzyme DNA polymerase during DNA replication. To apply this method for RNA, it also has to be converted to cDNA. Generally, this approach allowed correct detection of a detailed sequence of analyzed transcript. For a long time this technology was regarded as the "gold" standard due to its low error rate.

A certain improvement in transcriptome analysis was the microarray approach [Harrington *et al.*, 2000]. This technique applies a collection of microscopic DNA spots attached to a solid surface and allowed measurement of the expression levels of thousands of genes simultaneously. Due to the speed, RNA microarray analysis has become an essential component of biology and biomedical research.

Even though a number of RNA experimental analysis methods were available, certain limitations were hindering the detailed investigation of transcriptome structure and activity, which required novel approaches.

1.3 High-throughput sequencing

1.3.1 Initial approaches

A huge step forward in transcriptome analysis was taken when Next Generation Sequencing (NGS) technologies entered the field of molecular biology. NGS, also called high-throughput sequencing (HTS) technology, outputs sequences of millions of DNA strands in parallel, providing substantially higher throughput than the Sanger approach in a short time period and minimizing the need for the fragment cloning methods. The first NGS approach called pyrosequencing allowed the creation of *reads*: fragments of DNA providing whole genome segments of a certain size [Ronaghi *et al.*, 1998]. Initial NGS experiments also adapted the cDNA approach to detect the transcriptome activity from a cell [Morin *et al.*, 2008].

The development of HTS quickly went further. Technology was improving in the increased read length and process quality. At the same time, costs were falling. Currently there are three benchtop companies that provide HTS instruments: Illumina (www.illumina.com), Roche 454 (www.454.com) and Life Technologies (www. lifetechnologies.com). Notably, Illumina has presently the highest usage percentage - around 75% of sequencing applications. Importantly, the generated sequencing data has a complex structure and large size. Therefore, to achieve correct results from the data, detailed algorithmic approaches are required. Because of this, the importance of bioinformatics in molecular biology has increased significantly.

1.3.2 RNA sequencing

The initial RNA sequencing process was introduced quite fast [Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008]. Once again, the main change in the sequencing method was extraction of RNA from the cell and conversion to cDNA. The typical mRNA sequencing approach is demonstrated in Figure 1.3. The process starts with extraction of total RNA from a cell. Then, either polyA enrichment or reduction of ribosomal RNA is applied to increase the proportion of mRNA and other elements of transcriptome. Next, the conversion of processed RNA to double-strand cDNA is performed applying random primer hybridization. The generated cDNAs are broken into fragments of a certain size (typically 200-500 bp). These fragments are marked by adapter ligation from one or both ends, resulting in single-end reads or in paired-end reads. The sequencing process is performed after PCR amplification. A single Illumina sequencing run can produce up to hundred of millions of reads with a size of 100 bp. Importantly, each step of the sequencing procedure might introduce biases and errors [van Dijk *et al.*, 2014].

RNA-sequencing has significant advantages in comparison to previous technologies. For example, RNA-seq outperforms microarray approaches with higher accuracy and detection rate in expression analysis [Fu *et al.*, 2009]. The main goals of RNA-seq application are currently:

1. Quantification of gene expression of each transcript based on the cell type

2. Comparison of expression levels of cells between various biological conditions

3. Annotation of all expressed genes including their splice junctions (breaks between introns)

Additionally during past years RNA-seq introduced a number of specific issues that enabled discovery of novel events in the transcriptome. However, RNA-seq data analysis requires a lot of data processing along with specialized algorithms solving these tasks correctly and efficiently. There are typical analysis pipelines

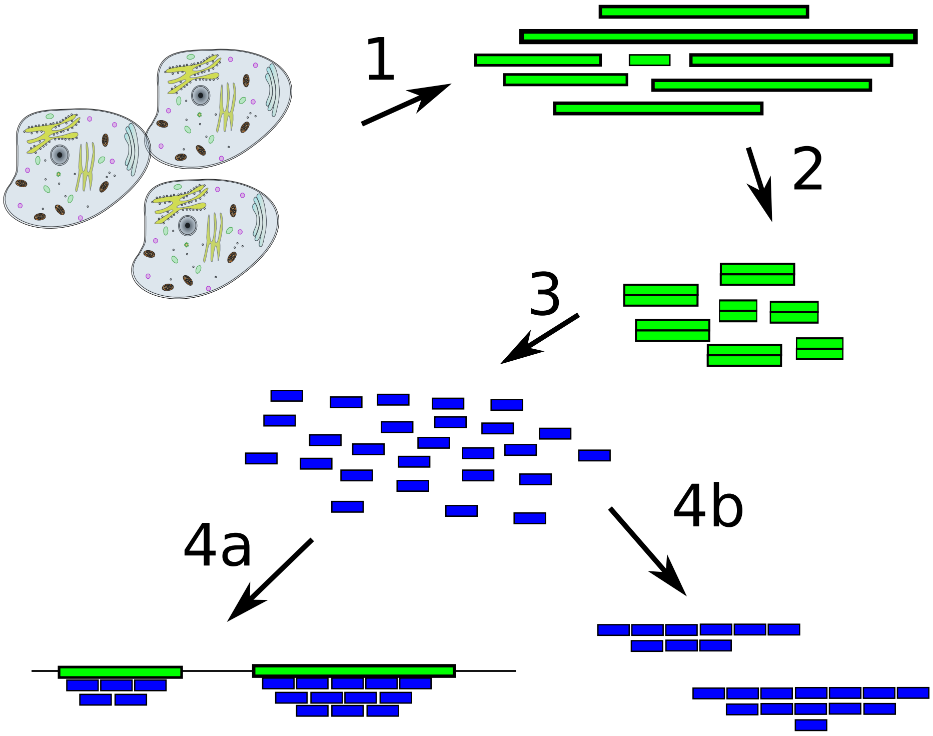

Figure 1.3: *RNA-sequencing process. 1) Total RNA is extracted from cells and separated from rRNA 2) Transcripts are converted to cDNA, multiplied and broken into fragments 3) Reads are generated from cDNA fragment ends 4) Analysis of reads is performed through alignment to reference (a) or assembly (b)*

applied frequently to solve these tasks, even though certain questions still remain unanswered.

1.4 Transcriptome analysis using RNA-sequencing

1.4.1 Data processing overview

Figure 1.4 illustrates the process of RNA-sequencing data analysis, including typical steps such as initial quality control of reads in FASTQ format, further processing techniques (alignment and assembly), and finally gene expression analysis and comparison. Additionally, certain novel approaches of RNA-seq application are usually applied after performing alignment or assembly. Detailed explanations of each step are provided further.

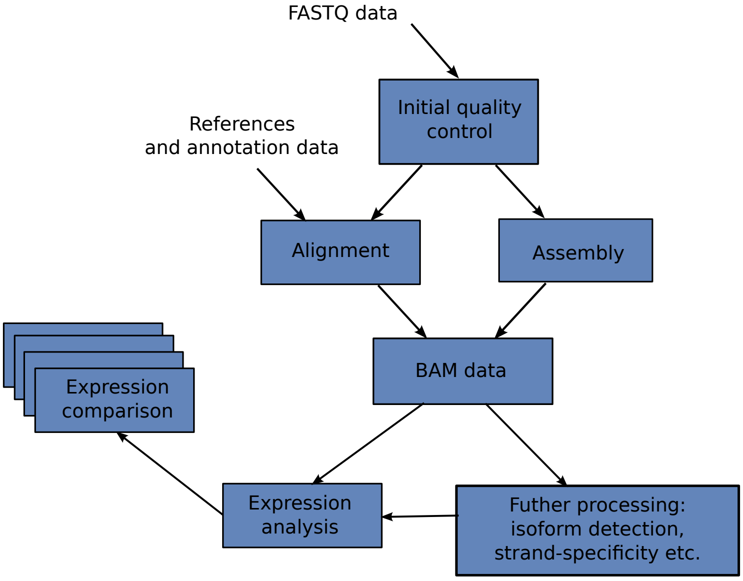

Figure 1.4: *RNA-sequencing data analysis overview*

1.4.2 Quality control of sequencing data

High-throughput sequencing is a powerful technology, however it has a complex structure and processing steps, therefore a careful design of data analysis is required. Moreover, due to the multilevel structure of the full experiment a number of problematic issues might occur. The process can suffer from inconsistent sample and library preparation, specific biases related to the sequencing platform and poor sample quality. Additionally HTS also suffers from inherent difficulties such as PCR-amplification bias and uneven fragment distribution [Ross *et al.*, 2013].Therefore, quality control is one of the requirements during the analysis procedure.

Several tools are available to check the quality of a sequencing experiment.The most widely used tool applied for the initial quality check is FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). It analyzes sequencing data and provides certain statistics, including per-base sequence quality, GC-content, read length distribution, non-detected nucleotide quantity, duplication level and adapter types. Each statistics value allows detection of certain problems in the sequencing experiment. Moreover, warnings are reported based on default expected values.

It is worth noting that certain problems, such as coverage bias or issues with insert size of paired-end reads can be detected only after further RNA-seq analysis steps are performed. Therefore, there are special quality control tools focused on processing of the results of the subsequent analysis operations. For example, we developed one such tool called Qualimap [García-Alcalde *et al.*, 2012], that analyzes alignment data in BAM format and characterizes its quality.

Importantly, there are technical and algorithmic errors specific only to RNA-seq analysis. Despite the smaller size of the transcriptome compared to the genome, RNA-seq data analysis is challenged by the existence of complex regulatory mechanisms like alternative splicing, transcription of processed pseudogenes, dynamic concentration range of isoform expression, etc. Thus, additional rigorous quality control measures are required for processing of RNA-seq data. There are issues, such as 5'-3' bias or over-expression level of a distinct RNA type, which occur only in RNA-seq data. The detailed quality control of RNA-seq will be described carefully in the next chapter. Here, we will continue with the following data processing steps.

1.4.3 Primary analysis

There are two typical analysis types of sequencing reads:

- Alignment

  Most of RNA-seq experiments are performed on organisms whose genomes or even transcriptomes are already known; therefore, alignment of reads can be applied to perform the analysis. Basically, alignment is a computational operation to detect the position where the read is mapped to the reference sequence. However alignment to the transcriptome can be more complicated. Since there are exons and introns in higher eukaryotes, there are additional difficulties with accomplishing the RNA-seq data alignment.

  There are two ways to perform this task:

  1. Alignment is performed to already known transcriptome sequences available in databases such as Ensembl. This method is similar to whole genome sequencing data alignment and there is a number of effective tools available for this task such as Bowtie [Langmead *et al.*, 2009] and BWA [Li and Durbin, 2009].
  2. Alignment is performed directly to the genome, taking into account reads breaking within introns. In this case an additional algorithmic approach is required to align reads that cover exon breaks. To detect these events, reads are separated into small segments that are used to reconstruct the isoform structure. The most frequently applied tools for this task are currently Tophat [Trapnell *et al.*, 2009], GSNAP [Wu and Nacu, 2010] and STAR [Dobin *et al.*, 2013].

  It is worth mentioning that results of alignment procedure are provided in a standard file format called Sequence Alignment/Map Specification (SAM). This format keeps the read alignment positions, along with specific properties such as pair type, mutations, insertions, deletions, duplications, etc. Additionally, data in SAM format can be easily packed to binary archive (BAM), which allows a significant reduce in the file size.

- Assembly

  The alignment process can be performed only if the reference sequence is available. However, in certain organisms this is not possible due to lack of existing information and the sequences must be reconstructed from reads. Popular tools to apply this technique are Trinity [Grabherr *et al.*, 2011] and Oases [Schulz *et al.*, 2012].

Additionally, the exon-intron model highlights the importance of a detailed reconstruction of isoforms. In this case assembly can be combined with alignment to verify the isoform evidence. The widely used approaches that combine these techniques to detect isoforms are described further.

1.4.4 Gene expression analysis

After the alignment or assembly of reads is performed, it is possible to detect the expression of a particular gene based on the computation of *coverage* - the number of aligned reads covering a certain fragment.

It is assumed that if the number of reads mapping to a certain biological feature of interest, such as a gene or a transcript, is sufficient, then it can be used as an estimation of the abundance of that feature in the sample and interpreted as the quantification of the expression level of the corresponding region. To apply this approach the number of reads aligned to a certain gene is counted to measure expression level. There are several methods based on this technology such as edgeR [Robinson *et al.*, 2010a] or DESeq [Anders and Huber, 2010].

Additionally, a more advanced type of analysis performs assembly of reads into full-length transcripts, and coverage estimation is based on the number of reads covering the breakpoints between exons and accurate distribution of reads located in exons. One of the most popular tools based on this approach is Cufflinks [Trapnell *et al.*, 2010].

An important step in computation of gene expression level is the consideration of the impact of noise and the systematic variation between samples. Therefore there are specific normalization techniques available. The most commonly used method normalizes the counts for exon length, assuming that read count distribution is the same in all samples [Mortazavi *et al.*, 2008]. The computed expression computation is reported in Reads Per Kilobase of exon per Million reads sequenced (RPKM) calculated in the following way:

$$RPKM = \frac{R * 10^9}{L * M}$$

In this function R - reads mapped to the gene, L - length of the gene, M - the total number of mapped reads in the experiment. There is also a variation of RPKM value that takes into account paired-end reads called Fragments Per Kilobase of exon per Million mapped fragments [Trapnell *et al.*, 2010]. It is worth noting that the RPKM value might introduce problems, since certain genes have very high expression values, and influence expression comparison results. To fix this issue additional methods have been developed to improve the normalization, such as upper quartile scaling [Bullard *et al.*, 2010] or trimmed mean of M values[Robinson *et al.*, 2010b]. Additionally, there could be a number of non-specific transformations that influence the results, including genetic background, time point and cell type heterogeneity. One such novel method was introduced to validate a variety of biological conditions [Risso *et al.*, 2014] and demonstrated appropriate results in comparison to other normalization methods [Peixoto *et al.*, 2015].

One of the main points of RNA-seq data analysis is the comparison between biological conditions. Detailed observation of gene expression difference is a complicated task due to several reasons: number of genes and samples, statistical significance of the event, systematic variation noise. The normalization, described previously, partly solves the noise problem, however additional estimations should be performed to validate the correctness of the results. Detection of differentially expressed genes is performed by comparing the expression thresholds of samples in different conditions by taking into account statistical modulations such as *p-values*. The toolkits performing computation of gene expression including Cufflinks and DEseq also enable detection of differentially expressed genes, applying various statistical verifications.

Notably, the best practice for gene expression analysis has not been defined yet. Due to complicated sequencing issues such as GC-content, paired-end read size and gene body coverage structure, current tools still have limitations in correct detection of differential gene expression [Rapaport *et al.*, 2013] and more novel methods are being developed to improve the status (for example, Cuffdiff2 [Trapnell *et al.*, 2013], DESeq2 [Love *et al.*, 2014] and PennSeq [Hu *et al.*, 2014]). Interestingly, a recent detailed comparison of current gene expression analysis methods based on simulation data demonstrated that inaccurate results are computed for hundreds of genes in *Homo Sapiens* transcriptome even by most popular and effective tools [Robert and Watson, 2015]. This investigation confirms that gene expression determination from RNA-seq data requires further specific improvements.

1.5 Additional aspects of RNA-seq data analysis

1.5.1 Isoform detection

A lot of genes in higher eukariotic organisms have numerous splice variants, promoters and protein products. Understanding of the isoform structures allows correct measurement of changes in expression of individual transcripts and their influence on cellular processes.

However, detection of isoforms is complex, because in most genes alternative isoforms share large amounts of sequences and differences in isoforms rely on exon extraction or intron region inclusion. Moreover, the read coverage distribution inside transcripts is biased due to polyA selection or rRNA depletion [Lahens *et al.*, 2014]. Additionally, a lot of genes have high sequence similarity and as a result sequencing reads align to a number of genes (so called *multi-mapped reads*). Therefore, counts for gene transcripts must be estimated carefully.

Early computational methods to detect isoform expression performed statistical normalization for isoform-level [Katz *et al.*, 2010]. Later developed methods provided accurate reconstruction of isoforms followed by statistical representation of the coverage [Trapnell *et al.*, 2012; Mezlini *et al.*, 2013] or focused only on exon blocks that perform the construction of various isoforms [Anders *et al.*, 2012]. Interestingly, different existing methods can report completely different results for the same dataset. A recent study compared existing tools for isoform detection on simulated and real datasets [Hayer *et al.*, 2014]. The research demonstrated high error rate for certain tools along with lower recall and precision specific to analysis of real data, concluding that advancement of algorithmic approaches and technology improvements should continue.

1.5.2 Strand-specificity

One of the interesting aspects of transcriptome activity is that certain genes might have both sense and anti-sense transcription. In several situations anti-sense transcription of a gene reduces its own expression level or controls translation [Li *et al.*, 2013]. Standard RNA-seq libraries based on RNA to cDNA conversion did not support strand-specificity detection, because during sequencing process the synthesis of double-stranded cDNA was followed by a random addition of adapters to 5' and 3' ends.

Of course, was possible to identify the strand of a certain gene using algorithmic approaches such as open reading frame information, biases in coverage of 5' and 3' ends or splice orientation. However, the technical validation of the strand would help to discover antisense transcripts, distinguish the strand of non-coding RNAs and correctly resolve expression levels of intersecting genes in different strands.

Therefore certain methods were developed to control strand-specificity during sequencing process [Levin *et al.*, 2010]. The techniques are based on attaching different adapters to 5' and 3' ends (for example, Illumina RNA, NSR or SMART-RNA ligation) or on marking one strand by chemical modification (most common is dUTP). All existing methods are integrated into a standard sequencing procedure. Even though the techniques also introduce specific errors, in general such approaches lead to novel detection of antisense transcripts along with detailed specification of intersecting genes and non-coding RNAs belonging to regions of known mRNAs [Core *et al.*, 2008].

1.5.3 Gene fusion detection

DNA damage events such as mutations, insertions, deletions and others might have dangerous consequences. The most well-known result is the generation of mutated cells that lead to cancer. Quick and correct detection of these events in cells will allow design of medical approaches to disable cancer progression.

One well-known example of such an event is a fusion gene detected in chronic myeloid leukemia (CML). A fusion gene is a combination of blocks of genes due to a break in the genome, such as insertion, deletion or translocation. Transcription and translation of a fusion gene can lead to the creation of a protein with dangerous unexpected activity. Detection of events such as the BCR-ABL fusion in CML led to the development of a treatment for this type of cancer by blocking the novel protein [Lugo *et al.*, 1990].

Studying gene fusions in the transcriptome enables detection of the rearrangements that might be translated into novel functional proteins. One of the most promising approaches is the detection of fusion genes from RNA sequencing data. It was shown that the detection of fusions from RNA-seq data is more convenient than from genome sequencing. For example, based on RNA-seq data analysis around 8000 fusions were detected in 4000 samples from various cancers, leading to potential medical aspects [Yoshihara *et al.*, 2014]. However, the methodology of fusion detection from transcriptome sequencing data is rather complex due to the homology of the genome and sequencing technology limitations. Some aspects of the topic, such as precision and recall properties of available methods, still require additional research. The third chapter of this thesis describes the current status in detail.

1.6 Research goals

RNA-sequencing analysis enables detailed detection of gene expression in a cell. In this aspect it demonstrates higher quality and confidence in comparison to other methods (i.e microarray analysis). Additionally, RNA-sequencing provides a number of important novel analysis approaches to investigate the transcriptome activity.

Certainly, the main goals of RNA-seq data analysis already have quite acceptable solutions. However, the existing methods are not completely correct and there is a number of different approaches to reach improved solutions for such problems as transcriptome assembly, gene expression analysis, isoform and strand-specificity detection.

Moreover, certain aspects of RNA-seq still require novel resolutions. For example, the RNA-seq process can introduce a number of characteristic biases and problems, that can lead to incorrect analysis results. There is a number of elements that should be taken into account before the results of an experiment can be trusted. Therefore, detailed quality control of RNA-seq data is an important task. Chapter 2 focuses on solutions of these problems and describes a novel tool, Qualimap2, which we developed to clarify this task. The manuscript characterizing this tool was published in the journal Bioinformatics [Okonechnikov *et al.*, 2015].

Additionally, the ability to detect certain dramatic events such as fusion genes is an important topic, since it will permit the application of RNA-seq not only in scientific research, but also in clinical care. However, due to the complex structure of the transcriptome in higher eukaryotic organisms, detection of fusions is a complicated task. Moreover, certain aspects, such as for example strand-specificity, are currently not yet taken into account during the fusion detection. Chapter 3 describes a novel toolkit for fusion detection from RNA-seq data called InFusion, which we developed to solve existing problems and advance abilities of transcriptome sequencing application. The paper describing InFusion was submitted to PLOS Computational Biology.

Importantly, RNA sequencing technology continues to improve. Specific aspects, such as increase of read size and RNA processing without cDNA transformation provide appreciable improvements. Fruitful novel techniques such as single cell RNA-sequencing grant a lot of opportunities, but require redesigned analysis approaches. Moreover, the increase in number of samples and experiments in the analysis present additional statistics demands. Chapter 4 provides a discussion about the novelties and focuses on future research targets.

CHAPTER 2

Quality control of RNA-seq alignment and expression data

2.1 Overview

This chapter characterizes the quality control requirements for correct processing of RNA-seq data. It starts with a description of existing technical problems and outlines possible ways to detect and validate them. Further, the design and functionality of a novel approach to accomplish RNA-seq quality control are described and comparison to other existing tools is performed. Finally, future research plans in this area are presented.

2.2 Reasons and approaches

RNA-seq is a powerful technology that provides a full determination of the transcriptome functionality. However, due to a complex technological structure of high-throughput sequencing, it includes a number of biases that infiltrate the analysis process. Moreover, cDNA synthesis, which is currently the most frequent RNA-seq approach root, introduces a number of specific errors, occurring only in RNA sequencing [Hansen *et al.*, 2010].

There is a number of additional sources of a bias. First of all, RNA-seq demonstrates a non-uniform distribution of read coverage in a transcript because of particular errors occurring during RNA extraction [Kim *et al.*, 2012], rRNA depletion [He *et al.*, 2010], adapter ligation [Faulhammer *et al.*, 2000] and PCR amplification [Kozarewa *et al.*, 2009]. Second, both read length and insert size of sequencing fragments limit the detection of a whole scope of the transcriptome [Oshlack *et al.*, 2009]. Third, read mapping and assembly operations might introduce certain limitations and biases due to read errors and algorithmic problems [Li *et al.*, 2010]. Also, the normalization of gene expression levels based on the processing of homologous reads leads to incorrect expression profiling.

Additionally, during the process of reverse transcription, the generated cDNA can dissociate from the correct RNA sequences and connect to a different RNA. This event leads to the generation of chimeric transcripts, that do not exist in reality. Also, the generated second strand of cDNA blocks the detection of strand-specificity of the transcript. Strand-specific libraries that solve this problem can also have some limitations, influencing the computed anti-sense transcription level results [Levin *et al.*, 2010].

Finally, experiments applying RNA sequencing can include a number of samples in different conditions. Even though the technology has the highest level of reproducibility in comparison to microarray approach, the analysis of genes with low expression level is limited and requires replicates. Unfortunately, a large number of samples might provide additional problems when a subgroup of outliers changes the total results of the analysis.

The detection of occurring errors and mistakes is quite important to be sure that the results of the analysis are correct. Therefore, specialized tools are required to perform detailed quality control (QC). A number of problems can be detected by general approaches for analysis of HTS data (for example, FastQC tool), however problems, specific only for RNA-seq data, need additional analysis operations for the generated datasets such as alignment of reads or gene expression counts.

To solve this task several tools were developed. One of the first approaches was RSeQC tool [Wang *et al.*, 2012], which is actually a list of Python scripts that provide statistics computed from the alignment data. The performed analysis results include such aspects as accurate properties of mapped reads, insert size distribution for paired-end reads, gene body coverage, proportion of reads mapping to gene structure (5' UTR, coding, 3' UTR), estimation of sequencing depth in RPKM and junction saturation. The results for each analysis type are generated separately and each script has various options and requirements.

Other similar method is RNA-seq QC [DeLuca *et al.*, 2012], a functional pipeline with a serial number of steps. Except of QC control of BAM files it also produces read counts and performs additional QC operations, including detailed coverage analysis, detection of rRNA reads, expression profile efficiency examination and strand-specificity validation. Moreover, a limited coverage correlation analysis can be performed for a number of samples.

We also developed Qualimap [García-Alcalde *et al.*, 2012] tool, that allowed precise general quality control analysis of BAM files along with counts data. Qualimap was able to detect a number of problems particular to any type of a sequencing experiment including whole-genome sequencing, exome sequencing and RNA-sequencing, while the counts QC mode was providing a solution for detection of QC issues related only to gene expression analysis.

However, additional biases occurring only in RNA-seq experiments were not supported by Qualimap. Also, comprehensive RNA-seq alignment statistics, such as for example, exon coverage proportion and strand-specificity validation, were not available. Moreover, a frequently applied multi-sample RNA-seq analysis can be biased by outliers, thus their detection should be verified in detail. Finally, the performance of the existing tools was not compared so far.

2.3 Qualimap2: advanced RNA-seq and counts quality control

2.3.1 Tool description

The second version of Qualimap tool was developed to handle the described limitations. One of the most important reasons to update the tool was a list of QC requirements specific only to RNA-seq data analysis, which were not supported by the first version. Therefore, Qualimap2 includes a novel mode focused on RNA-seq alignment data quality control. Moreover, the read counts QC mode, introduced in the first version, was redesigned to support global analysis of multiple samples and comparison of sample groups.

In general Qualimap2 is a multiplatform user-friendly application with both graphical user (GUI) and command line interfaces. It includes four analysis modes: **BAM QC**, **Counts QC**, **RNA-seq QC** and **Multi-sample BAM QC**. The latter two modes are first introduced in the second version.

Based on the selected type of the analysis, users provide input data in the form of a BAM/SAM alignment, GTF/GFF/BED annotation and/or read counts table. The results of the QC analysis are presented as an interactive report from GUI, as a static report in HTML or PDF format and as a plain text file suitable for parsing and further processing. Typically, the report contains summary statistics of the dataset, description of the input data, exploratory plots and histograms that visualize multiple properties of the processed data and help to detect potential problems.

The mode **BAM QC** is an initial mode that performs detailed analysis of the alignment data in SAM/BAM format. As it was mentioned previously, it allows detection of a number of problematic issues related to any type of a sequencing experiment. For example, it provides extensive alignment statistics along with coverage plots and histograms. Importantly, certain aspects such as number of marked mutations, insertions, deletions, insert size qualities and other properties that can be detected only from the alignment data are reported.

Multiple BAM QC mode is a novel mode, which is also focused on analysis of BAM files in general. It takes into account and combines results created by **BAM QC** mode to create plots combining statistics from a number of samples. Moreover, principal component analysis (PCA) is applied to detect outliers based on the selected statistics.

The two modes designed for RNA sequencing data QC analysis are further described in detail.

2.3.2 RNA-seq QC mode

A novel mode **RNA-seq QC** reports quality control metrics and bias estimations specific only for the whole transcriptome sequencing, including reads genomic origin, junction analysis, per-transcript coverage, consistency of library protocol and 5'-3' bias estimation. This mode can be applied as a complementary tool together with **BAM QC** mode.

The **RNA-seq QC** mode is designed as an algorithmic pipeline that analyzes alignments from an input BAM file. Each alignment is processed to collect defined statistics such as number of mapped reads, pair construction, etc. The transcriptome annotations are also required. Based on the annotation data, read alignments are analyzed to detect intersections with exon, intron or intergenic regions. During this operation analysis of the coverage of transcripts is performed; read counts and other important statistics are computed.

After the analysis is finished the following results are reported:

- Summary

  The summary contains several sections describing in detail statistics specific for RNA-seq alignment, including :

  - The assignment of read counts per-category

    The total number of mapped reads and the distribution of alignments belonging to a selected type are reported. The types include unique alignments, secondary alignments (duplicates marked by SAM flag), non-unique alignments (SAM format **NH** tag of a read is more than one), reads aligned to genes or without any feature (intronic and intergenic), ambiguous alignments and a number of unmapped reads.

  - Transcript coverage profile

    The ratios between mean coverage at 5' region, 3' region and the whole transcript are reported. To compute this value for each transcript mean coverage along with mean coverage in the first 100 bp (5' region) and the last 100 bp (3'region) are calculated and collected. Afterwards, the collected values are sorted and median is selected from each array to compute the ratios.

  - Reads genomic origin

    The report shows how many alignments fall into exonic, intronic and intergenic regions. Exonic region includes 5'UTR, protein coding region and 3'UTR region. To detect alignment positions, annotations data is used to generate all exonic and intronic intervals. A read alignment is checked if it intersects with an exon or an intron. In case intersection is not detected, it is considered as intergenic.

  - Junction analysis

    The total number of reads with splice junctions and 10 most frequent junction types are reported. The junctions are detected by analyzing SAM format CIGAR field. The N operation detects the skipped region from the reference and represents read alignment covering an exon. Additionally, a pair of nucleotides from left and right side of a skipped region are analyzed to detect the junction rate.

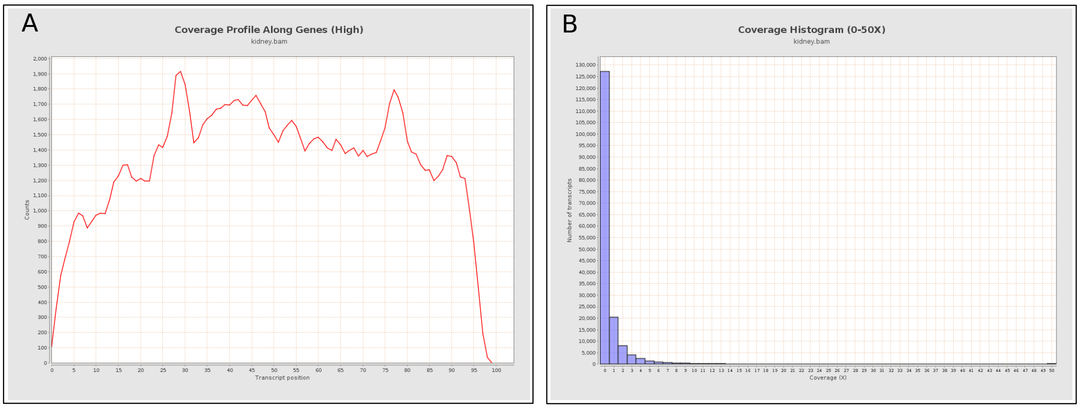

Figure 2.1: *RNA-seq QC plot examples. (A) Coverage profile of highly expressed genes (B) Coverage histogram (from 0 to 50X)*

- Pie chart

  The plot shows how many read alignments fall into exonic, intronic and intergenic regions. Results computed and reported in the Summary are demonstrated in the plot.

- Coverage Profile (Total)

  The plot shows mean coverage profile of the transcripts. All genes with non-zero coverage are used to generate this plot. From each gene only one transcript, having the highest expression, is selected.

- Coverage Profile (High)

  The plot shows mean coverage profile of 500 highest-expressed genes (Figure 2.1A). Transcript selection is similar to the total Coverage Profile.

- Coverage Profile (Low)

  The plot shows mean coverage profile of 500 lowest-expressed genes. Transcript selection is similar to the total Coverage Profile.

- Coverage Histogram

  The histogram demonstrates the coverage of transcripts from 0 to 50X (Figure 2.1B). The genes that have coverage higher than 50X are collected in the last column.

- Junction Analysis

  The pie chart is focused on the types of junction positions in spliced alignments. *Known* category represents percentage of alignments where both junction sides are known from the annotation. *Partly known* value represents alignments where only one junction side is known. All other alignments with unknown junctions are marked as *Novel.*

Importantly, during the **RNA-seq QC** procedure the read counts are also computed. By default, a read alignment is considered supporting the gene and counted, only if it lies exactly inside of exon region. In case of paired-end reads, both pair mates should support the same gene or transcript. Counts computation includes a number of optional parameters. For example, the strand-specificity can be taken into account based on the selected protocol, i.e. if a mapped read doesn't fall the correct strand of a gene, it is not counted. Additionally, if a read is mapped to multiple locations, it is ignored by default. However, there is also an option to count a multi-aligned read as separated proportionally between targets. If this option is activated then, for example, a read mapped to 4 different locations will add 0.25 to the counts at each location. After the analysis is finished, the final counts value is converted to integer.

2.3.3 Counts QC mode

The counts data, generated during the **RNA-seq QC** procedure or computed using some other tool such as HTSeq [Anders *et al.*, 2014], can be utilized to assess differential expression between two or more experimental conditions. However, before performing differential expression analysis, researchers should be aware of some potential limitations of RNA-seq data, as for example the saturation level influence on sequencing depth or feature types detected in the experiment. These and other properties can be analyzed by interpreting the plots generated by **Counts QC** mode.

In the second version of Qualimap **Counts QC** module has been redesigned to work with multiple samples under different conditions. The new functionality is mostly based on NOISeq package [Tarazona *et al.*, 2012], therefore to use **Counts QC** it is required to have R language along with certain packages installed.

To perform the analysis it is also necessary to provide a special table that contains information about input sample datasets. Each sample has a name and a condition, which describes the group. Therefore, biological differences can be easily mentioned by setting group conditions. Additionally, it is possible to perform not only a default analysis of all samples, but also compare the conditions.

In result, after the processing of input data **Counts QC** mode generates three groups of plots.

1. *Global Plots*

   Plots from this group present a global overview of the counts data and include all samples. These plots allow to compare all samples without taking into account experimental conditions.

   - Counts Density

     The plot shows density of counts computed from the histogram of log-transformed counts. In order to avoid infinite values in case of zero counts, the transformation $log_2(expr + 0.5)$ is applied, where $expr$ is

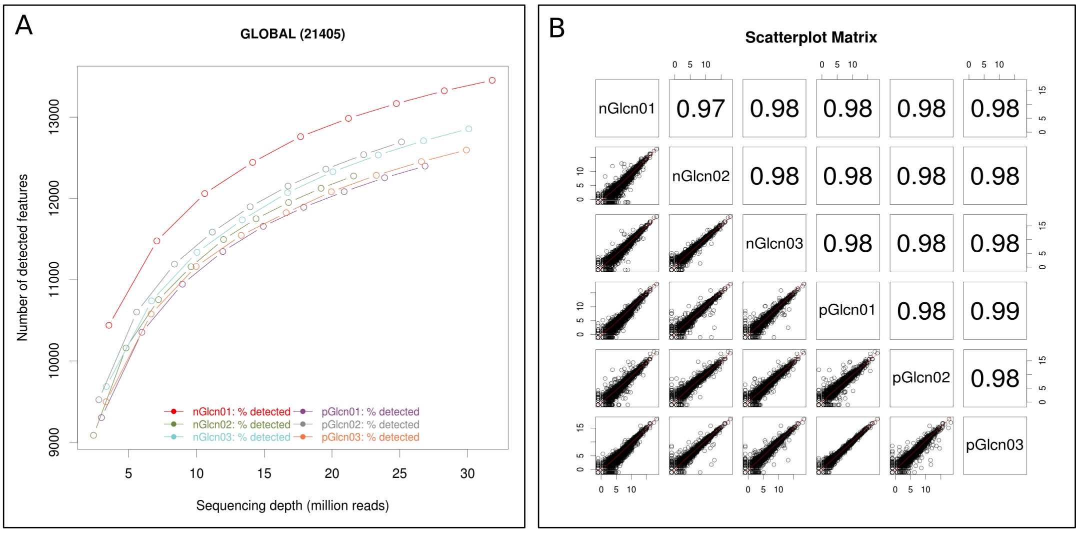

***Figure 2.2:** **Counts QC** Global Plot examples. (A) Saturation of expression (B) Scatterplot matrix*

a number of read counts for a given feature. Only log-transformed counts having value greater than 1 are plotted.

- Saturation

  The plot provides information about the level of saturation in the samples and helps to decide if more sequencing is required (Figure 2.2A). The sequencing depth of the sample is represented at the x-axis. Smaller depths correspond to samples randomly generated from the original sample.The curves are associated to the left y-axis represent the number of detected features at each of the sequencing depths in the x-axis. They show the number of newly detected features, when the sequencing depth increases in one million of reads.

- Scatterplot Matrix

  The panel shows for each pair of samples a scatter plot along with a smoothed line in the lower panel and Pearson correlation coefficients in the upper panel (Figure 2.2B). Plots are generated using log-transformed counts.

- Counts Distribution

  The boxplot shows the global distribution of counts in each sample. For each sample the mean value surrounded by quartiles is demonstrated. Additionally, detected outliers are marked.

- Features With Low Counts

  The plot shows the proportion of features with low counts in the samples. Such features are usually less reliable and could be filtered out.

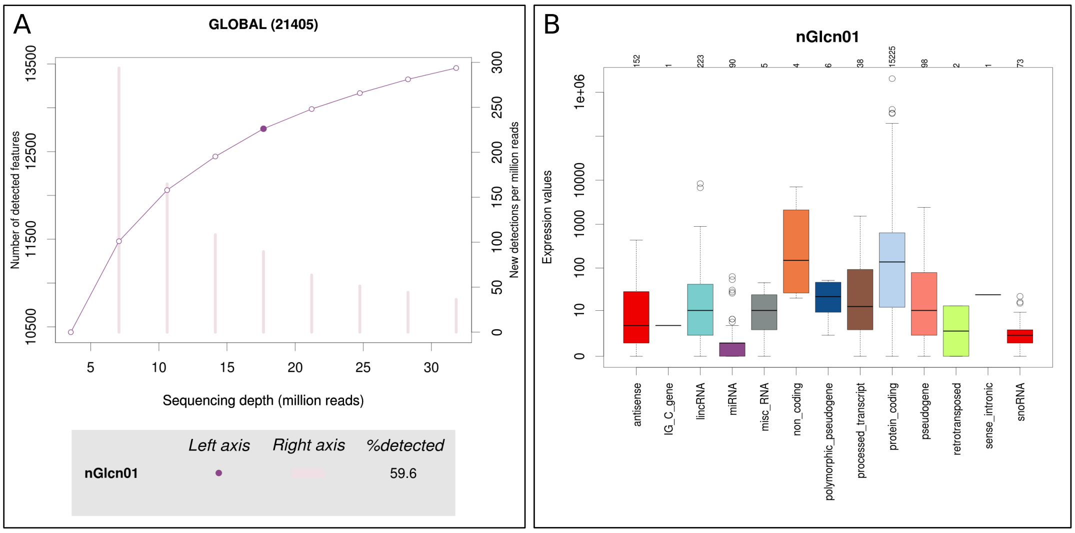

Figure 2.3: ***Counts QC*** *Individual Sample plots examples (A) Saturation with two y-axis demonstrating number of detected fusions along with number of novel detections per million of reads (B) Counts per biotype*

In this plot, the bars show the percentage of features within each sample having more than 0 counts per million (CPM), or more than 1, 2, 5 and 10 CPM. The detection of outliers is possible by comparison of CPM proportions.

2. *Individual Sample Plots*

Apart from the global overview, there are plots generated individually for each sample. When an annotation file describing biotype, length and GC-content of each transcript is provided by the user, additional series of plots is generated.

- Saturation

  For each sample, a saturation plot is generated like the one described in *Global Saturation* (Figure 2.3A). Additionally, the right side of the plot demonstrates the total number of detected features per million of reads.

- Bio Type Detection

  Since RNA-seq experiments might be designed to detect specific RNA types (i.e. microRNA or long non-coding RNAs) it is important to verify the count distribution across bio types. The barplot visualizes features that are detected in the sample. The x-axis shows all the groups provided in the annotations file such as for example protein coding, miRNA, lincRNA, pseudogene etc. The grey bars are the percentage of features of each group within the reference genome (or transcriptome, etc.). The striped color bars are the percentages of features of each group detected in the sample with regard to the genome. The solid color bars are the percentages that each group represents in the total detected features in the sample.

- Counts Per Biotype

  The boxplot per each group describes the counts distribution in the given biotype (Figure 2.3B). For each biotype the mean value surrounded by quartiles is demonstrated, additionally certain outliers are shown. The generated plot allows to compare the expression levels among biotypes and detect possible contamination.

- Length Bias

  The plot describes the relationship between the length of the features and the expression values. The length is divided into bins. Mean expression of features falling into a particular length interval is computed and plotted. A cubic spline regression model is fitted to explain the relation between length and expression. Coefficient of determination R^2 and p-value are shown together with regression curve.

- GC Bias

  The plot describes the relationship between the GC-content of the features and the expression values. The data for the plot is generated similar to Length Bias plot. The GC content divided into beans and then mean expression features corresponding to given GC interval are computed. The relation between GC-content and expression is investigated using a cubic spline regression model.

3. *Comparison Plots*

   When an option to compare conditions is activated, additional plots comparing data in groups of samples having the same biological condition or treatment are generated. The samples belonging to a group are combined and mean values for each comparison are computed. Currently only two types of a group are supported.

   - Counts Distribution

     The plot is similar to the one in the *Global Plots* report. It compares distributions of mean counts across conditions.

   - Features With Low Counts

     The plot is similar to the one in the *Global Plots* report. It compares proportions of features with low counts by computing mean counts across conditions.

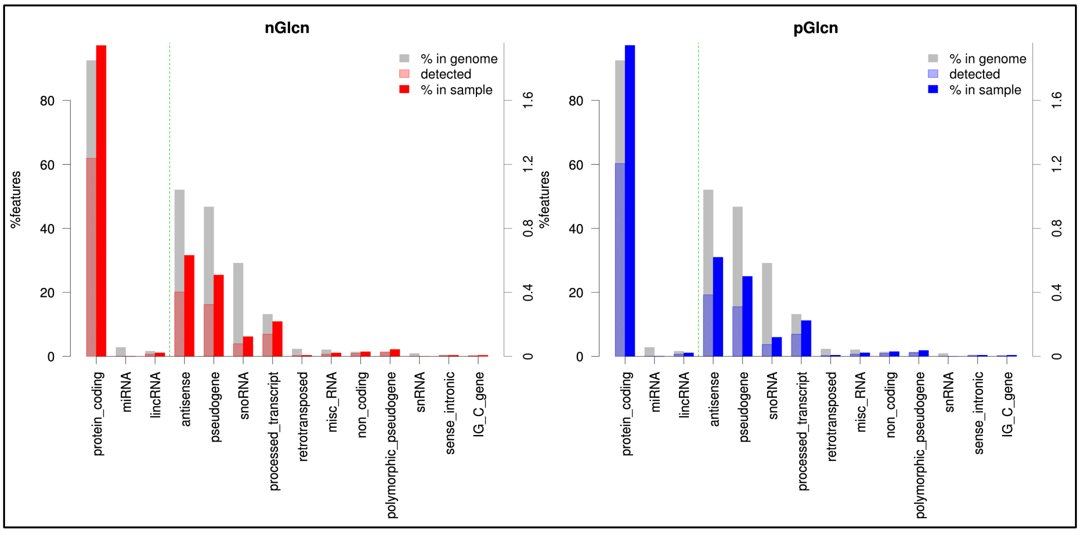

Figure 2.4: ***Counts QC*** *Comparison Plots example. Bio detection plot demonstrates the expression levels of various RNA types among selected groups.*

- Bio Detection

  The plot is similar to the one in the *Individual Sample Plots* report (Figure 2.4C). It compares distribution of the detected features for the given biotype by computing mean counts across conditions.

- Length Bias

  The plot is similar to the one in *Individual Sample Plots* report. It analyzes relation between feature length and expression across conditions.

- GC Bias

  The plot is similar to the one in the *Individual Sample Plots* report. It analyzes relation between GC-content and expression across conditions.

All the described plots are designed to demonstrate biases or problems that can be detected only from the counts distribution. Moreover the *Comparison Plots* group allows to compare two biological conditions. This is a frequent requirement in the transcriptome related research.

It is worth noting that there are additional options that allow to control the counts QC analysis. For example, in order to remove the influence of spurious reads, counts threshold is applied to consider a transcript as detected only if its corresponding number of counts is greater than this threshold. By default, the threshold value is set to 5 counts.

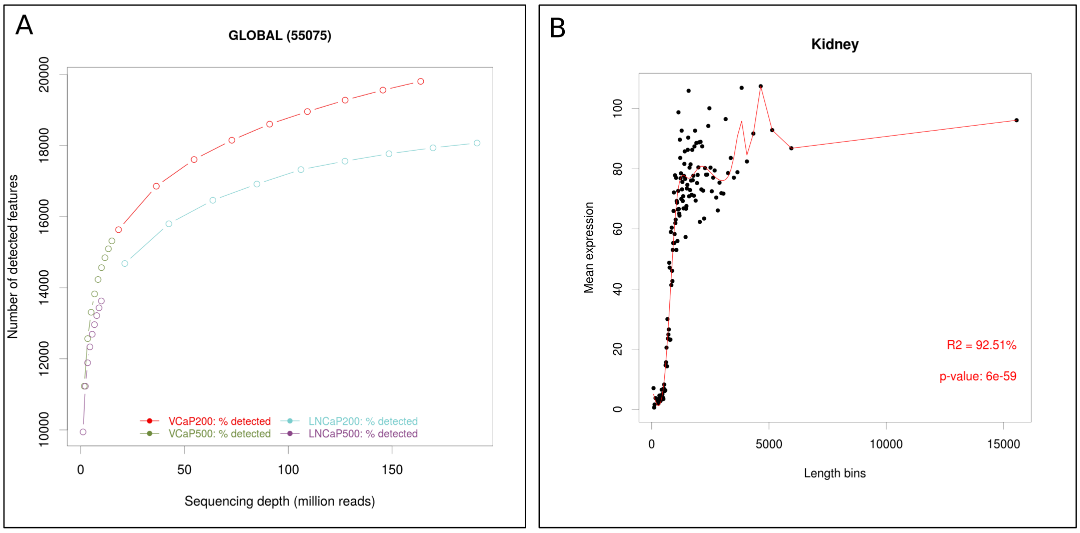

Figure 2.5: *Detection of biases from* ***Counts QC*** *report (A) Global saturation demonstrates low coverage context of samples VCaP500 and LNCaP500 in comparison to samplse VCaP200 and LNCaP200 (B) The influence of the read length on expression: normalization is required*

2.3.4 Insight from quality control analysis

The statistics values computed from RNA-sequencing alignment and counts data allow to detect problematic issues that can not be discovered directly from the reads data. The most of the technological biases, including transcript length and coverage uneven distribution or contamination by non-required transcripts can be detected only from a processed dataset, therefore data investigation performed by Qualimap is required to confirm sufficient quality of the processed data.

The quality control process should start from the examination of **RNA-seq QC** report. The first statistics of the mapped data provides the status of the sequencing process in general. The low proportion of aligned reads reported in the summary can demonstrate ligation and amplification biases. The typical majority of aligned reads should belong to known exonic regions. For example, comparing the proportion of aligned reads falling to exon region in case of well-known organisms such as *Mus Musculus* or *Homo Sapiens* should be up to 90%. Smaller proportion can indicate sequencing biases or some mistakes in the mapping process. The same conclusions can be derived from *"Junction analysis"* plot: known junctions should dominate novel.

Some specific issues occurring during rRNA depletion and polyA selection can lead to biases in 5' region and 3' region. For example, it is quite well-known that polyA selection can lead to high expression in 3' region. The 5'-3' bias allows to detect such events. In correct experiments this bias should be close to one.

The plots focused on transcriptome coverage analysis such as "*Coverage Histogram*" and "*Coverage profile along genes*" give an overview of available level of expressed genes. Importantly, low level coverage in RNA-seq data influences significantly on differential gene expression analysis [Łabaj *et al.*, 2011], therefore demonstration of high and low coverage level is quite useful to detect biases and apply suitable gene expression normalization approach.

After the quality control of the BAM file is performed, the computed counts will allow to investigate further properties of the experiment. Most importantly, **Counts QC** analysis can be performed with taking into account the experimental design, such as biological conditions and the number of samples.

The analysis of expression contamination allows to verify if the total number of reads in the experiment is enough to detect all expressed genes. The plot "*Saturation*" demonstrates the influence of the number of sequencing reads on expression distribution. Basically, the angle of the line plot shows how the increase in number of reads controls the proportion of novel detected genes (Figure 2.5A). If more reads do not lead to the growth of a number of detected genes (line angle is close to zero), additional sequencing is not required. Notably, there is also a specific customizable limitation for a minimum number of read counts required for a transcript to be added to the plot.

The biotype analysis of counts performed for each sample allows to detect the types of expressed features. This is especially important if the RNA-seq experiment is focused on long RNA or other type of RNA. Abnormal contamination can be detected from the plots "*Biotype detection*" and "*Counts per biotype*". Typically, in mRNA-seq experiments protein coding proportion should dominate in counts. Additionally, the plots detect the suitability level of a selected sequencing protocol.

For correct normalization of counts in gene expression analysis it is important to take into account the length and GC content of expressed transcripts. The requirements for such normalization can be checked from "*Length bias*" and "*GC bias*" plots (Figure 2.5B). Cubic spline regression model is applied to detect if length and GC proportion fit the gene expression. Generally, the computed coefficient of determination greater than 70% or a large *p-value* indicates an effect on expression level and importance of normalization [Tarazona *et al.*, 2015].

The computed global plots demonstrating all samples together such as scatterplot matrix, counts density and distribution allow to detect outliers. For example, despite different biological conditions the global expression levels among analyzed samples should match sufficiently. Importantly, different biological conditions should influence on gene expression. Therefore, all the plots related to expression analysis and normalized to a specific condition also allow to detect outliers.

2.3.5 Comparison to other tools

Except of Qualimap there are other tools that are available to perform the quality control task. The most highly-used applications are RSeqQC and RNA-seq QC. We performed a detailed comparison of Qualimap **RNA-seq QC** and **Counts QC** modes to these methods. The results are provided in the table 2.1. It is worth noting that RSeQC and Qualimap also support general sequencing data analysis methods (including GC-content, number of mismatches and indels, insert size, mapping quality etc.). However, in this table these elements are not included, since it is focused on RNA-sequencing data analysis.

Table 2.1: *Comparison of RNA-seq quality control tools.*

Analysis type	**RSeQC**	**RNA-SeQC**	**Qualimap**
Aligned reads statistics types	Non-splice	Total, unique, duplicate and alternative; Vendor Failed Reads.	Total, secondary, aligned to genes, non-unique, no-feature and ambiguous
Read pairs statistics	Pairs aligned, left/right	Pairs aligned, unpaired reads; pair mismatch rate; chimeric pairs	Pairs aligned, left/right
Strand-specificty detection	Available	Available	Available
Alignments location analysis	Exonic (5'UTR, 3'UTR, CDS), intronic	Exonic (5'UTR, 3'UTR, CDS), intronic, TSS	Exonic, intronic, intergenic
Gene coverage analysis	Gene coverage over gene body plot	Coverage gaps (count, length); coverage plots	Coverage profile along genes (total, low, high); coverage histogram
5'- 3' bias analysis	-	Available	Available
Gene expression computation	Available (RPKM)	Available (read counts, RPKM)	Available (read counts, RPKM)
Expression profiling	-	Efficiency (ratio of exon-derived reads to total reads sequenced); rRNA reads	Detailed percentage of expressed exon type; low counts detection
Multisample analysis	Coverage of several samples together	Correlation between each sample pair	Coverage density, scatterplot matrix, saturation, counts distribution
Group comparison	-	-	Counts distribution, bio detection, length bias, GC bias

Comparison of existing RNA-seq quality control tools: RSeQC v2.6, RNA-SeQC v1.1.8 and Qualimap v2.1.

According to this table, quality control tools have some similar functions. However, as it can be seen, Qualimap outperforms other programs in multi-sample analysis and group comparison blocks, providing a number of plots that allow to compare samples and detect outliers. The novelties of Qualimap are mostly focused on comparison of various conditions.

Moreover, Qualimap2 demonstrated superior performance in comparison to other tools. It showed twice work speed increase in typical *H. sapiens* RNA-seq data analysis. Moreover, the analysis performed by RSeQC requires to launch and control a number of different scripts, thus additional work is necessary to design an appropriate pipeline.

2.3.6 Results and future plans

Qualimap2 application performs an important task: detection of errors and limitations specific to RNA sequencing. Overall, the second version of Qualimap was downloaded more than 4000 times since the initial release in September 2014. Importantly, some of the detected QC biases can be fixed during analysis applying normalization techniques, however certain issues can not be improved, therefore novel experiments might be required. There are already examples of publications where Qualimap reports were contributing to relaunch of sequencing experiments [Koeppel *et al.*, 2015].

One interesting subject to mention: Qualimap2 is an open-source tool, which has a public repository. After the second version was released, a number of suggestions to improve required aspects of quality control came from user community. Additionally, specific bugs were reported and fixed by users.

However, some elements of RNA-seq data analysis should be interpreted in more detail. For example, influence of read size and insert size can be processed to verify specific RNA-seq limitations for gene expression analysis. Novel companies such as PacBio (www.pacificbiosciences.com) provide rather long read size resulting in a number of specific errors, which detection is quite important. Recently created single cell RNA-seq process introduces even more biases and issues that influence the analysis results [Stegle *et al.*, 2015]. Moreover, currently experiments might include a number of different conditions, while Qualimap2 supports in *Comparison* mode only two conditions. Therefore, adaption of Qualimap should continue in these directions.

2.4 Summary

As it was described in previous sections, RNA-sequencing might provide various technical biases and errors that can influence the results of the analysis. To detect these events we developed the second version of Qualimap, an application for exploratory analysis and quality control of HTS alignment data written in Java and R. Qualimap2 introduces a novel analysis mode called **RNA-seq QC**.

This mode allows computation of metrics specific for RNA-seq data, including per-transcript coverage, junction sequence distribution and reads genomic localization. Furthermore RNA-seq QC estimates 5'-3' bias and consistency of the library protocol.

The mode **Counts QC**, created in the first version to estimate the quality of the read counts, was completely redesigned to allow processing of multiple samples. Having multiple biological replicates per condition is quite common in RNA-seq experiments; therefore it is beneficial to be able to analyze counts data from all generated datasets simultaneously. Multi-sample analysis allows inspection of grouping of the samples, as well as discovery of outliers and batch effects. Similar to the previous version, the **Counts QC** mode estimates the saturation of sequencing depth, counts density, correlation of samples and distribution of counts among classes of selected features. Additionally there are new plots that explore the relationship between expression values and GC-content or transcript length are available for users. The analysis results include a combined overview of the datasets along with a QC report for each individual sample. Moreover the analyzed samples can include two different conditions, e.g. treated and untreated. In this case, separate plots comparing groups of samples corresponding to a particular condition are generated.

Overall, Qualimap2 has become an important tool for quality control of RNA-seq experiments. The number of downloads and citations of the initial manuscript increased significantly after the release of the second version. It is worth noting that Qualimap has gathered a community of users who frequently report existing problems, suggest new features and contribute their code.

Chapter 3

Fusion gene and chimeric transcript detection

3.1 Overview

This chapter focuses on the detection of fusion genes and chimeric transcripts from RNA-seq data. It describes existing approaches in this area and outlines their limitations. The aim of this chapter: present a novel RNA-seq based toolkit, which introduces a number of improvements. The description of the algorithm, comparison to other tools, experimental validation and possible future research plans are provided.

3.2 Fusion discovery approaches

3.2.1 History

Genomic translocations, insertions and inversions can lead to the appearance of *fusion genes*, which are formed from sequences of several genes. Gene fusions and other genomic translocations are closely related to cancer progression and play a driver role in particular types of cancer [Mitelman *et al.*, 2007]. Examples include TMPRSS2-ERG fusion in a large family of prostate cancers [Tomlins *et al.*, 2005], EML4-ALK in non-small-cell lung cancer [Sheng *et al.*, 2001] and ETV-NTRK3 in several cancers [Rubin *et al.*, 1998; Tognon *et al.*, 2002]

Chimeric transcripts can occur in normal cells due to trans-splicing, cis-backsplicing or errors of the transcription machinery, and in certain cases they have been reported to be active in mammalian genomes [Frenkel-Morgenstern *et al.*, 2012]. Additionally, genes might yield a joined RNA product, which is called a read-through transcript. Both chimeras and read-through transcripts might have specific functions and play a role in a cell process.

The first fusion event was detected due to the search of structural rearrangements in cancer. In 1970 cytogenetic analysis through specialized chemical *banding technique* allowed to find a genomic translocation in chronic myeloid leukemia (CML) [Lugo *et al.*, 1990]. The chromosome banding technique was further applied to discover a number of fusion events in several cancers. The next step in the improvement of fusion discovery was a technology called *Fluorescence in situ hybridization* (FISH). It provided an opportunity to detect fusions by apply-

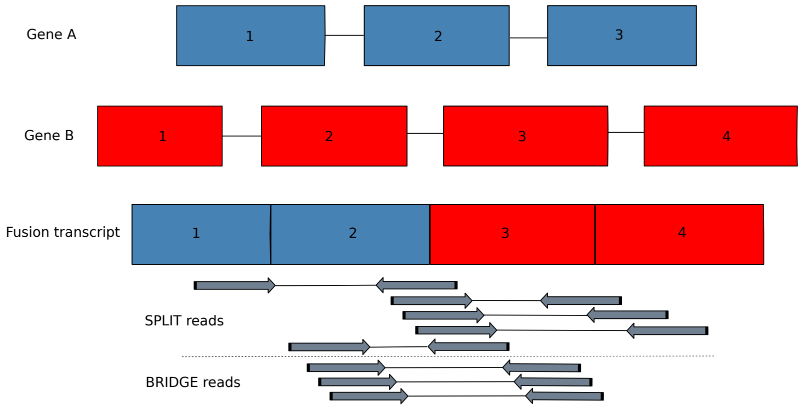

Figure 3.1: *Detection of fusion events via RNA-seq reads*

ing a careful visualization of chromosome blocks based on the color distribution. Further, fusion detection process was improved with *array-based analysis* that allowed detailed gene expression and copy number profiling. The first fusions validated in prostate, lung and other cancers were detected based on this approach.

However, as it was mentioned in the first chapter, RNA sequencing contributed to the progress in fusion discovery: it allowed detailed and global detection of transcribed fusion genes and chimeric transcripts. Maher et al. [Maher *et al.*, 2009] were one of the first groups to apply RNA-seq for gene fusion discovery in several cancer cell lines. They were able to not only confirm previously described fusions and chimeric transcripts, but also detect and validate a number of novel events. In result this was a first step of upcoming frequent and high resulting application of RNA-seq technology for detection of fusions and chimeric transcripts in various organisms.

3.2.2 Detailed RNA-seq approach explanation

The detection of fusion genes and chimeric transcripts from RNA-seq data is possible due to the properties of the sequencing procedure: the breakpoint position and the region between genes forming the fusion can be covered by reads. There are two types of reads that allow to discover fusions: SPLIT and BRIDGE reads.

The SPLIT and BRIDGE read events is easy to demonstrate using a hypothetical fusion of two genes (Figure 3.1). The first event occurs when the read spans the fusion junction, termed a SPLIT read. The second event requires the reads to be paired-end. In this case a pair of reads from the same fragment spans the fusion within the non-sequenced part of the insert, termed a BRIDGE read pair. The detection of SPLIT and BRIDGE reads can be performed by applying an adapted alignment approach.

Additionally, it is possible to apply assembly of reads to reconstruct fusion transcripts. However, to detect fusion events, the assembled transcripts should be also aligned in a SPLIT manner to the known transcriptome. Even though assembly-based methods have several useful abilities and properties, they should be combined with SPLIT- and BRIDGE-read approaches.

3.2.3 Existing tools and their limitations

After the pioneering studies demonstrated the efficiency and positive results, efforts intensified to develop efficient computational methods for the detection of fusion genes from RNA-seq data.

FusionSeq was one of the first published computational pipelines for fusion gene discovery from RNA-seq data [Sboner *et al.*, 2010]. The method is based on the detection of discordantly aligned read pairs (BRIDGE reads), which are used to construct a junction library of possible fused exon pairs. The sequencing reads are then realigned to the constructed library to find the exact fusion junctions. Further methods that adapted and improved this approach were MapSplice [Hu *et al.*, 2010] and ShortFuse [Kinsella *et al.*, 2011].

Although the basis for organizing a pipeline for fusion gene detection was established, a practical application of FusionSeq on a variety of datasets revealed that the approach is not equally sensitive in all cases [Li *et al.*, 2011]. Further developed methods such as TopHat-Fusion [Kim and Salzberg, 2011] and ChimeraScan [Iyer *et al.*, 2011] were based on detecting reads that cover the junction of genes involved in a putative fusion event (SPLIT-reads). However, due to the small size of the sequencing reads and the repetitive nature of the genome, this approach requires intensive filtering to remove the large number of false positives.

The next step in fusion detection from RNA-seq data was focused on the integration of both BRIDGE-read and SPLIT-read approaches. One of the first examples was deFuse [McPherson *et al.*, 2011]. It employs discordantly aligned pairs for initial fusion discovery, followed by the application of the SPLIT-read approach to find the exact fusion breakpoint location. Additionally, it improves specificity of the discovery by utilizing machine learning techniques to better distinguish between true and false positive predictions.

Further developed methods such as SOAPfuse [Jia *et al.*, 2013] and fusionCatcher [Nicorici *et al.*, 2014] were focusing on constructive filtering of false positive results and advanced various aspects of fusion gene discovery, e.g. fusion isoform detection, prediction accuracy, computational resource usage.

Additionally, attempts were made to discover fusions using a reference-guided assembly of chimeric transcripts [Chen *et al.*, 2012; Fernandez-Cuesta *et al.*, 2015]. A major drawback of such approach is that it relies on the detection of possible discordant read alignments and also requires as much filtering as SPLIT-read based methods

3.3 InFusion: novel fusion discovery toolkit

3.3.1 Introduction

Although a number of state-of-the-art fusion discovery methods are currently available, a recent analysis [Carrara *et al.*, 2013] as well as our own investigations showed up differences in results between commonly used methods applied to the same validated datasets. Reliable discovery of the whole spectrum of different fusion events from RNA-seq data is a challenging task [Mertens *et al.*, 2015] and systematic assessment of this feature among available methods has not been performed so far. It is worth noting that antisense transcription can also occur in fusions [Frenkel-Morgenstern *et al.*, 2014]. Up to now, only TopHat-Fusion and ChimeraScan take the strand specificity provided by the library preparation protocol into account, but they do not mark whether a detected fusion was transcribed in sense or antisense direction .

Additionally, there is occasional evidence of functional fusion events with a breakpoint inside an exon [Robinson *et al.*, 2013] or involving non-coding [Shah *et al.*, 2013], intronic [Kangaspeska *et al.*, 2012; Steidl *et al.*, 2011] or even intergenic regions[Kleinman *et al.*, 2014]. Several studies have reported alternatively spliced isoforms of fusion genes [Kangaspeska *et al.*, 2012; Panagopoulos *et al.*, 2013]. A number of existing tools, including TopHat-Fusion, deFuse, SOAPfuse and fusionCatcher, are capable of detecting fusion isoforms. Likewise, deFuse and TopHat-Fusion can also discover fusions that involve intronic regions.

To improve certain aspects of fusion detection from RNA-seq a novel computational method called InFusion was developed. This toolkit for the discovery of chimeric transcripts from RNA-seq data is capable of detecting alternatively spliced chimeric transcripts and fusion genes involving non-coding regions. Specifically, InFusion allows detection of fusions that involve intergenic regions, which to our knowledge has not been addressed previously. The method applies a novel algorithmic approach to cluster and reconstruct fusions from SPLIT and BRIDGE reads. Additionally, it analyzes and filters putative fusion events based on coverage depth, genomic context and strand specificity. We found that InFusion shows improved accuracy on simulated and a number of public datasets. We experimentally validated our method by performing strand-specific deep RNA-sequencing of two well-characterized prostate cancer cell lines. Overall, InFusion discovered more than 400 fusions for each cell line (from 80M RNA-seq reads each), and from 40 tested fusions we confirmed 10 known and 26 previously unreported fusion transcripts using qPCR. Among these validated transcripts are several novel alternatively spliced isoforms of well-known fusions and some that involve fusion of an intergenic region with a coding one.

The InFusion pipeline was developed in C++ and Python. It is capable of working with both single-end and paired-end sequencing reads. It is free for academic use and can be downloaded from *http://bitbucket.org/kokonech/infusion*.

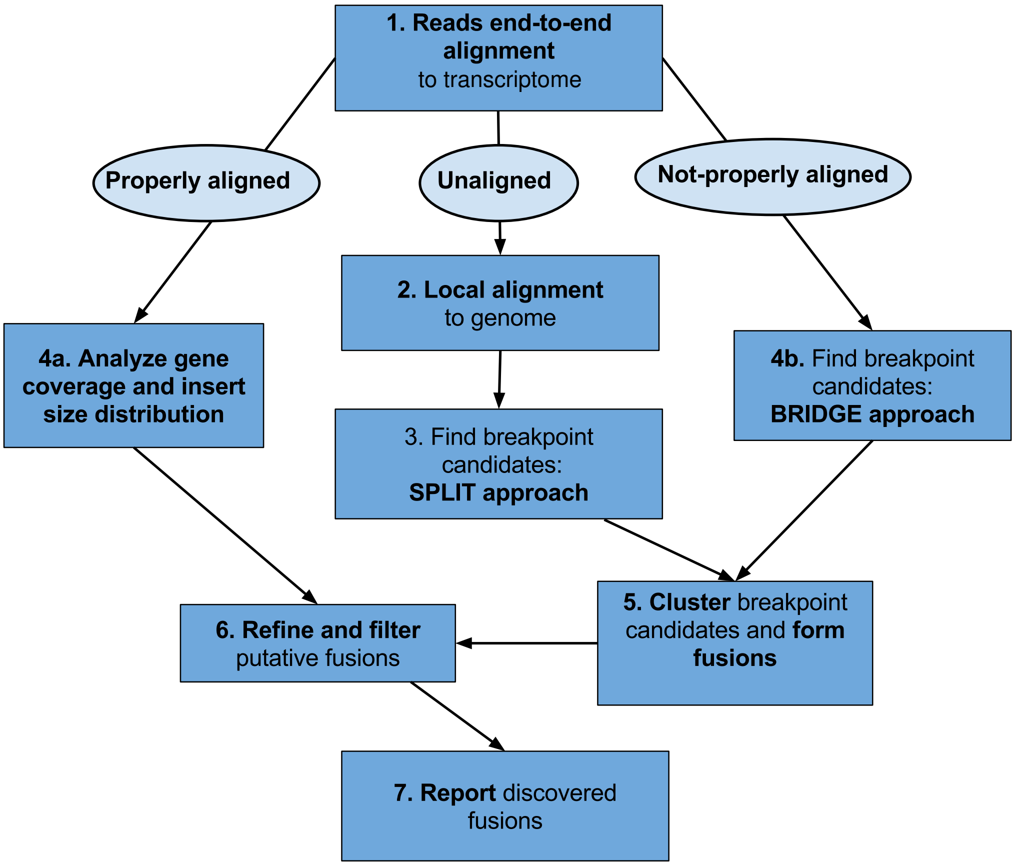
1. Reads end-to-end alignment to transcriptome
Properly aligned
Unaligned
Not-properly aligned
2. Local alignment to genome
4a. Analyze gene coverage and insert size distribution
4b. Find breakpoint candidates: BRIDGE approach
3. Find breakpoint candidates: SPLIT approach
5. Cluster breakpoint candidates and form fusions
6. Refine and filter putative fusions
7. Report discovered fusions

Figure 3.2: *InFusion pipeline structure*

3.3.2 Pipeline description

The InFusion pipeline consists of several independent steps each of which is controlled by a number of configuration parameters. The typical input for the pipeline are the reads from an RNA-seq experiment. The pipeline is capable of working with both single-end and paired-end reads. Most of the pipeline components in C++ are implemented using SeqAn bioinformatics library for efficient computations [Döring *et al.*, 2008]. InFusion relies on the genome and transcriptome information from the organism of interest in an indexed format. The pipeline provides functionality to automatically download required annotation and sequence files and perform indexing of genome and transcriptome sequences. The index has to be constructed only once for the target organism and the index database can be reused for further analysis.

The InFusion analysis pipeline is outlined in Figure 3.2. The pipeline starts with the mapping of reads to the transcriptome and optionally the genome (Step 1) while keeping track of the unmapped reads. In Step 2, the unmapped reads from the previous step are aligned locally to the genome. The resulting local alignments are used to detect potential SPLIT reads in Step 3. Next, the reads aligned to transcriptome and genome are analyzed to collect paired-end read alignment statistics (Step 4a) and detect discordantly mapped read pairs that form BRIDGE read pairs (Step 4b). This step is skipped for single-end sequencing experiments. Next, SPLIT and BRIDGE reads that potentially belong to the same fusion are clustered (Step 5). During this step, the pipeline tries to rescue SPLIT reads that were not detected during initial analysis of local alignments. Finally, putative fusions are analyzed to filter false positive events (Step 6) and discovered chimeric transcripts are reported (Step 7). The first four analysis steps were introduced previously and being used by existing tools such as chimeraScan, deFuse, SOAPfuse and others. However InFusion provides several improvements to these steps and there are a number of novelties in following analysis procedures: clustering of detected SPLIT and BRIDGE reads, advanced filtering and recovery along with special statistics computation for fusion reports. Below, each step of the pipeline is described in detail.

1. Alignment of the short reads

The main task of the read alignment is to find short reads that could reveal a putative fusion event and separate them from those originating from "normal" genes. In addition, the alignment of reads is used to collect statistical information about the sequencing experiment, such as insert size distribution and gene expression levels, which is used later in the analysis and the filtering of putative fusion events. Firstly, the given short-reads are mapped to the transcriptome (Figure 3.2, step 1). To reduce the pipeline running time, reads that were not mapped to the transcriptome can optionally be mapped to the genome. For the alignment of the reads InFusion uses Bowtie2 [Langmead and Salzberg, 2012] by default, however, any modern short read aligner could potentially be used.

2. Local alignment of the short reads

In order to search for possible SPLIT candidates, unmapped reads reported during the initial alignment are further aligned locally to the genome (Figure 3.2, step 2). Local alignment of short reads to the genome is performed using Bowtie2 in *local mode*. Similar to the previous step, InFusion can potentially be used with other aligners that support local mapping. The most crucial parameter of the local alignment is the minimal score to consider a valid mapping. InFusion is designed to accept mappings with a score greater than $Score_{max}/3$, where $Score_{max}$ denotes the maximum alignment score within a given scoring scheme. The following scoring scheme is used by default: 2 for match, -2 for mismatch, -6 for gap open, -3 for gap extension.

3. Analysis of local alignments

All local alignments of a read are analyzed in order to detect if it is possible to form a SPLIT candidate from it. We define the following conditions. Let us assume that there are n local alignments of a read:

$$\{A_1, A_2, .., A_n\},$$

Each alignment A_i is represented by reference sequence (chromosome) c, starting position in the reference sequence coordinates p , starting position q in the read coordinates and length l:

$$A_i : \{c_i, p_i, q_i, l_i\}$$

We consider that 2 local alignments A_i and A_j form a SPLIT read if 2 conditions apply:

1. Either $c_i \neq c_j$ or $p_i > p_j$, $p_i - p_j - l_j > I_{max}$
   The alignments are coming from different chromosomes or the distance between alignments is larger than maximum intron size. Maximum intron size I_{max} depends on the studied genome. The default value refers to the human genome with a maximum intron size of 20000. This value is selected since there are validated chimeras having rather small distance between connected exons [Edgren *et al.*, 2011].

2. $T_{inner} < abs(q_i + l_i - q_j) < T_{outer}$
   The alignments are concordant in the read coordinates based on given thresholds: maximum intersection size T_{inner} and maximum distance between alignments T_{outer} in the coordinates of the read. By default we set these thresholds to 2 and 10 bp respectively.

To increase the sensitivity of discovery we allow multiple local alignments of a single read. The multimappings are resolved in later steps of the pipeline. InFusion accounts only for SPLIT reads which are formed by two local mappings and by default allows up to 20 possible SPLIT configurations to be formed from the same read.

4. Analysis of end-to-end alignments

Paired-end sequencing experiments make it possible to search for BRIDGE reads. To perform this task, we analyze the correctly aligned short reads from Step 1. Firstly, the transcriptomic alignments are converted into genomic coordinates. Two mate alignments of the same read M_1 and M_2 are considered concordant if they are aligned within a distance of the defined maximum intron size I_{max}, otherwise they form a possible BRIDGE read pair. We record all discovered BRIDGE read pairs for further analysis. In addition, we compute the insert size distribution from concordant alignments and estimate expression level for each gene.

5. Clustering and forming putative fusions

A chimeric transcript can be described by a pair of genomic coordinates, which represent the fusion breakpoint. For example, the fusion gene shown in Figure 3.1 has a breakpoint formed by the last base pair of exon 2 of gene A and the first base pair of exon 3 of gene B. We refer to the SPLIT reads and BRIDGE reads detected in previous steps as breakpoint candidates (BPCs). It should be noted that a SPLIT read implies an exact breakpoint, while a BRIDGE pair implies an approximate breakpoint within the corresponding insert size distance. Thus, each BPC is described by 2 alignments, which are the local alignments of the read if the BPC is formed by a SPLIT read, and the end-to-end alignments of a pair of reads if the BPC is formed by a BRIGDE pair.

A clustering procedure groups the alignments forming the BPCs into *clusters* based on their genomic coordinates. Thus, alignments of each BPC are iteratively analyzed: i) the alignment constitutes a new cluster if there is no intersection with other existing clusters; ii) if the alignment intersects with one existing cluster, it is added to this cluster; iii) if the alignment belongs to two or more clusters simultaneously, the clusters are first merged into one single cluster and the alignment is then added to it.

As a result of the initial clustering procedure (Figure 3.3a), clusters can contain multiple possible breakpoints represented by different groups of alignments. Therefore we further separate the clusters based on their directionality: the alignment strand and the order of the alignment in the BPC, defined by location in the read coordinates in case of a SPLIT read or mate identifier in case of a BRIDGE read, predict if the breakpoint position is situated either upstream or downstream of it. (Figure 3.3b)

We next analyze if the coordinates of the breakpoint positions implied by local alignments in a cluster are compatible within a configurable tolerance (10 bp by default). If this is not the case, the cluster is separated into 2 new clusters. (Figure 3.3c) The process continues until there are no more clusters with significant difference in the coordinates of the breakpoint position left.

We further refine the clusters by assigning to them the compatible unused local alignments of reads from Step 2 of the pipeline that have an alignment score greater than 50% of the maximum score and an edit distance less than 2 (both options are configurable). Using an interval tree data structure we intersect the read alignments with the existing clusters. For each found intersecting cluster (which we call host cluster) we make sure that the alignment is concordant with the putative breakpoint position as it is dictated by the directionality and alignments from SPLIT reads in the cluster. We then select the remaining unmapped part of the read sequence and try to realign it to each potential fusion partner of the host cluster, again in concordance with cluster directionality and the fusion breakpoint.

Clusters formed solely from BRIDGE pairs constitute a special case in the rescue procedure described above. For this type of cluster the exact breakpoint

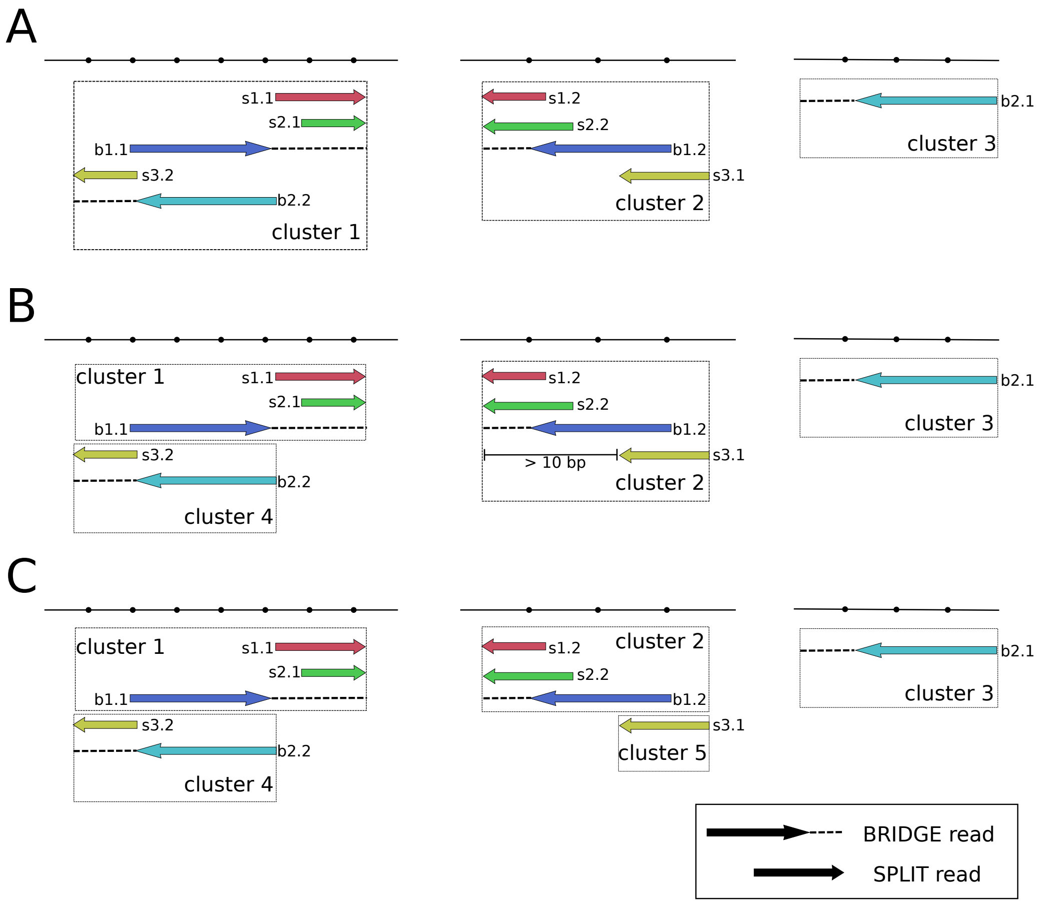

Figure 3.3: *Clustering of fusions. (A) Initial clusters are created from intersecting SPLIT and BRIDGE alignments. (B) Cluster 4 is separated from cluster 1 based on the directionality, which is inferred from the alignment strand and order. (C) Cluster 5 is separated from cluster 2 based on the putative breakpoint position. Alignments belonging to the same breakpoint candidate have the same color. BRIDGE reads are marked with b, SPLIT reads are marked with s. A SPLIT read assumes an exact breakpoint, while a BRIDGE read assumes an approximate breakpoint within allowed insert size distance.*

position cannot be computed, therefore we allow additional tolerance (computed from insert size distribution) in the genomic region upstream to the breakpoint position as dictated by the pair configuration. The rescued reads found in this case undergo an additional cleanup procedure, which selects the most probable breakpoint based on the amount of evidence for a particular position.

Finally we go through the list of BPCs to form putative fusions. For every BPC we check to which cluster each of its alignments belongs to. Then we assign the BPC to a putative fusion event described by two unique cluster identifiers.

During the creation of putative fusions we also take into account the strand-specificity of the sequencing library to reconstruct the correct 5'-3' order of the fusion transcript.

6. Refining and filtering fusions

We make use of the paired-end information by discarding SPLIT reads from putative fusions which do not have their mate-pair located within the maximum intron size. Additionally, clusters which consist uniquely of BRIDGE reads are merged with clusters that are located within maximum intron size in order to avoid reporting two fusions associated with the same event.

We resolve multimapped reads by assigning iteratively each read with several mappings to one putative fusion with the largest score among other fusions, similar to the deFuse algorithm. The fusion score is calculated by counting the number of supporting reads, taking into account their alignment score and the presence of mate pairs for BPCs originating from SPLIT reads.

After resolving multimapped reads we further calculate features associated with each fusion which are used to filter and analyze the fusions. There is a number of filtering algorithms, each of which can be configured with specific options. Default thresholds aim to provide a compromise between recall and precision based on our experience in analyzing human RNA-seq datasets.

The following filters are applied:

- Minimal supporting reads

  The number of supporting BRIDGE and SPLIT reads demonstrate the controlling evidence of a discovered fusion. By default at least four supporting reads of any type and one SPLIT read covering the breakpoint should be detected. Notably, in case of paired-end reads data the SPLIT reads are verified to have a correct pair. Additionally, the number of rescued SPLIT reads can be controlled.

- Unique split read alignment

  We reconstruct in detail the fusion structure and analyze the coverage of the supporting SPLIT reads. For each fusion we compute the proportion of unique alignments based on the their genomic coordinates and estimate if the mean breakpoint position in the read coordinates follows a uniform distribution. Both computed features are used for filtering. Additionally, if the fusion is supported only with SPLIT reads we require at least one SPLIT-read that does not have any multimappings in order to accept the predicted event.

- Homology

  We construct the fusion sequences bounded by the read evidence and align it to the genome and transcriptome. By adapting alignment score, we can filter out false positive predictions arising from parologs or highly homologous genes. To detect false positives the fusion sequences forming fusions are aligned to transcriptome and genome allowing multiple alignments. If both fusion blocks detected belonging to the same transcript or same genome region, then the fusion cluster is marked as homologous.

- Metacluster homogeneity

  By analyzing the genomic intervals forming the putative fusion, it was noticed that clusters of false positive fusions arising from local homology (sequence similarity) are usually found intersecting or close to each other in a small region. By contrast, true fusions are formed from rather unique sequences and dominate over other events in the selected region. We defined intersecting cluster blocks belonging to different fusions as metacluster. Avoiding false positive fusions reported due to homology is possible by filtering SPLIT-read and BRIDGE-read alignments from repeat regions. However, if the repeat regions are not provided for the genome or not detailed enough, then it is important to perform metacluster homogeneity analysis. We apply this observation by calculating the proportion of each fusion cluster in the metacluster and filtering out those fusions that have low weight.

- Insert size

  The reconstructed fusion is used to calculate the insert size for each of the supporting BRIDGE reads. The insert size is considered valid if it lies within 3σ interval as defined by the insert size distribution, computed in Step 4. The ratio of valid insert sizes is calculated and used as filtering parameter.

- Genome coverage

  Additional analysis of coverage in the genes forming the fusion is performed. By default at least one additional read should be supporting each transcript forming the fusion.

The final filtering of fusions is performed by applying configurable thresholds to the computed features. Additionally, InFusion allows repeated filtering of fusions with adjusted thresholds without running the whole pipeline again.

Strand-specific protocols are preferable for fusion discovery since they make it possible to detect antisense transcription in fusions and infer the direction of transcription in fusions involving unannotated and intergenic regions. If the strand specificity is enabled we calculate the proportion of supporting reads which are aligned according to the protocol and annotated strand of the gene. The computed metrics allows analysis of anti-sense transcription in the fusion. If the fusion involves an unannotated segment we infer and report the probable transcription strand of the corresponding fusion part.

7. Reporting fusions

For each predicted fusion event, InFusion reports corresponding genomic regions, the coordinates of the breakpoint, as well as the number of supporting SPLIT and BRIDGE reads together with the features computed during putative fusion analysis and used in the filtering process. Additionally, the genomic regions involved are characterized by annotating implicated transcripts/exons, determining the type of fusion event (e.g. inter-chromosomal, read-through), etc. Optionally, InFusion reports the fusion junction sequence which can be useful for PCR primer design, as well as the original alignments of reads supporting the predicted events in BAM format.

3.3.3 Comparison to other existing tools: simulation

To check the quality of fusion detection we developed an advanced fusion simulation pipeline. The simulation pipeline uses gene annotations and genome as input for constructing fusion gene annotations with given properties.

We defined 5 classes of fusion events:

1. With both fusion partner(s) having a breakpoint close to a known exon boundary;

2. With one or both fusion partner(s) having a breakpodgint inside an exon;

3. With one or both fusion partner(s) forming a novel exon boundary inside intron;

4. With one fusion partner originating from an intergenic region;

5. With several alternatively spliced isoforms having breakpoints close to known exon boundaries.

Using the simulation pipeline, it is possible to create a required number of random fusion gene pairs of each class. After the gene pairs and break positions are created, the fusion transcript sequences are produced. Based on the transcripts the sequencing paired-reads are generated using the Mason software package [Holtgrewe, 2010]. There are certain options to control the simulation datasets: read length, insert size and strand-specificity. The pipeline is designed to make every run reproducible and is available as a part of the InFusion source code package.

For our simulation experiments we generated 50 sets of fusion annotations, each consisting of 100 gene pairs. Each set was including 40 events of the first class, 10 events of the second class, 10 events of the third class, 20 events of the fourth class and 20 events of the fifth class. The second and the third classes were selected to form less fusions since such events have rather low frequency.

Next, for each set of the generated annotations we recreated fusion transcript sequences and randomly assigned coverage ranging from 1X to 60X. For read simulation we used the Illumina error model which includes mismatches, insertions and deletions. The length of the read was chosen as 75 bp with a fragment mean length of 300 bp and a standard deviation of 80 bp.

Overall the generated 50 RNA-seq datasets were containing the evidence for 5,000 fusion genes. Additionally, based on the expression profile from BT474 sample from public RNA-seq data [Edgren *et al.*, 2011] we generated 30 millions of background reads using the same read simulation settings as for the fusion transcripts. These background reads were added to each dataset.

We ran InFusion along with five widely used tools for fusion detection - TopHat-Fusion, deFuse, ChimeraScan, SOAPfuse and fusionCatcher - on the generated RNA-seq datasets. We selected the first three tools for the assessment since they were reported to have the highest sensitivity and specificity in comparison to other tools [Frenkel-Morgenstern *et al.*, 2014], while the last two are quite novel and capable of detecting various classes of chimeras. For each program we measured the number of true positive (TP) and false positive (FP) predictions among all discovered events, as well as recall (TP/(TP+FN)) and precision (TP/(TP + FP)). We considered a prediction as true positive if for each fusion partner the breakpoint position was reported within 20 bp upstream or downstream of the exact junction point defined by the simulation design. The number of false negative predictions was computed by analyzing how many simulated fusions were not detected in the dataset. In order to enable the discovery of larger spectrum of fusion events and apply equal thresholds for fusion filtering, we configured the parameters of each tool accordingly. The results of the analysis are summarized in table 3.1. InFusion demonstrated the best recall and as well as a high level of precision, similar to SOAPfuse.

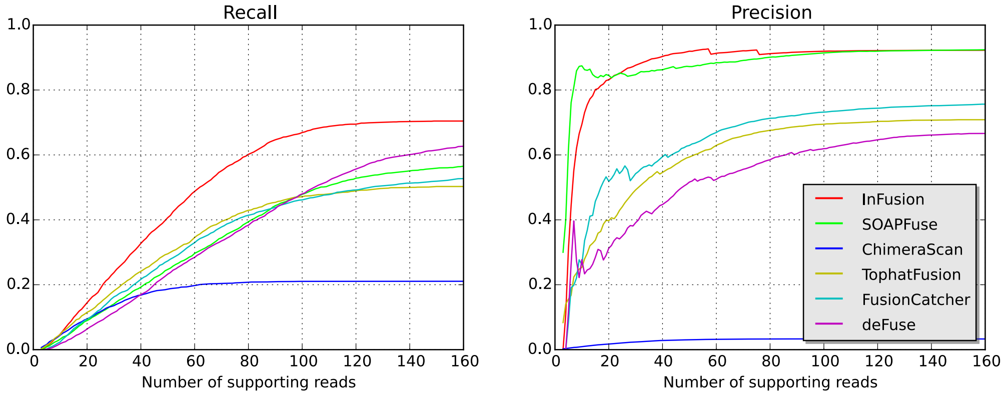

Figure 3.4: *Fusion detection from simulation data.*

Table 3.1: *Comparison of fusion detection tools on simulated data.*

Tool	N	TP	FP	Recall	Precision
InFusion	76 ± 4	70 ± 4	5 ± 1	0.7 ± 0.04	0.92 ± 0.02
SOAPFuse	60 ± 4	55 ± 3	4 ± 2	0.56 ± 0.04	0.92 ± 0.04
ChimeraScan	639 ± 4	21 ± 3	618 ± 6	0.21 ± 0.03	0.03 ± 0.01
TopHat-Fusion	70 ± 5	50 ± 4	20 ± 1	0.50 ± 0.05	0.71 ± 0.02
FusionCatcher	68 ± 6	51 ± 4	17 ± 4	0.52 ± 0.04	0.75 ± 0.05
deFuse	92 ± 7	61 ± 6	30 ± 4	0.61 ± 0.06	0.66 ± 0.04

Based on in silico data from 50 simulated RNA-seq datasets with 100 fusion genes each (5,000 fusions, 10 million read pairs in total). N = total number of predictions reported by each tool, TP = true positives, FP = false positives. Each dataset was analyzed independently, and then mean value and standard deviation were computed for comparison

To enable the discovery of a larger spectrum of fusion events and apply equal thresholds for fusion filtering, we configured the parameters of each tool accordingly. The detailed tool configurations are available in appendix sections A.1 and A.2.

In order to investigate the effect of read coverage we further analyzed the recall and precision based on the number of reads supporting the fusion (Figure 3.4). InFusion demonstrated superior recall and high precision independent of the number of reads supporting the fusion event.

Since the simulated transcripts consisted of five distinct classes, we also investigated the prediction accuracy for each fusion class independently (Figure 3.5). In result we detected that InFusion provided the highest recall and rather high precision in all fusion types except the break inside exon. However, the problem is related the detection of correct breakpoint position. The check of the correctness of breakpoint position in simulation was performed by setting certain thresholds to settings. By default the breakpoint position threshold was set only to 5 bp. However the increase of threshold InFusion demonstrated improvement of sensitivity.

Importantly, InFusion provided the highest recall and precision in detection of fusions within intergenic regions. Among other tools only deFuse demonstrated the ability to detect such events. The structural changes in genome can lead to appearance of novel genes in previously uncovered regions, therefore this subject can play a significant role.

Additionally we performed special simulation experiments to validate the strand-specficity detection by InFusion. For this purpose certain reads simulating fusions were generated using three strand-specific protocols: forward-strand-specific, non-strand-specific and reverse-strand specific. Additionally for each simulation run we provided background read data in non-strand specific mode. Each simulation was performed 3 times. Then we applied InFusion and checked if

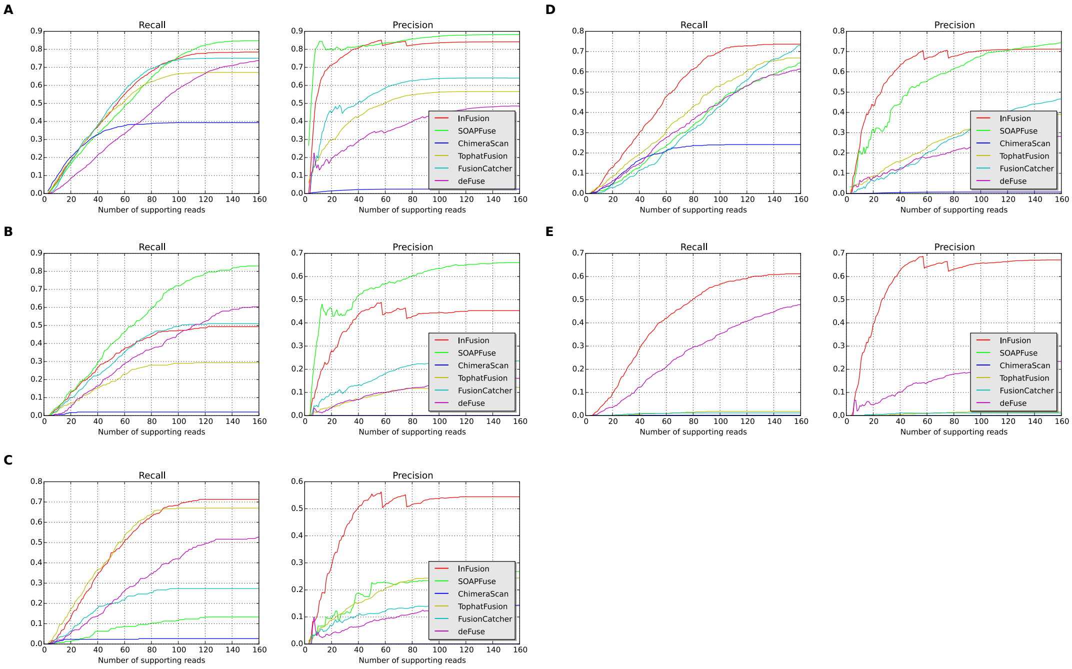

Figure 3.5: *Comparison of recall and precision on simulation data per fusion class. Simulated fusions and corresponding datasets were analyzed per fusion class: (A) with both fusion partner(s) having a breakpoint close to a known exon boundary (B) with one or both fusion partner(s) having a breakpoint inside an exon (C) with one or both fusion partner(s) forming a novel exon boundary inside intron (D) with several alternatively spliced isoforms (E) with one fusion partner originating from an intergenic region.*

the resulting fusion transcript strand-specificity was detected correctly according to the protocol. The results are provided in Table 3.2.

Table 3.2: *Mean strand specificity proportion detected by InFusion*

Simulated fusion strand specificity type	**Computed SSP value without/with background**
Reverse-strand-specific	0.035475 / 0.035477
Non-strand-specific	0.499587 / 0.499436
Forward-strand-specific	0.970641 / 0.970969

InFusion reports strand-specificity proportion (SSP) for each fusion. SSP validates from 0 (reverse-strand speficic) to 1 (forward-strand-specific). For each dataset mean detected SSP among all fusions was computed.

The simulation results demonstrated that InFusion provided correct strand-specificity estimations for detected fusions despite the type of background reads strand-specificity. This ability will allow to check the functionality of chimeras and their activity. As it was demonstrated, in certain cases antisense chimeras might play a significant role [Frenkel-Morgenstern *et al.*, 2014], therefore detection of these events might be important. However, it is worth noting that the main requirement for this analysis is the application of strand-specific sequencing protocol during the experiment.

3.3.4 Comparison to other existing tools: public datasets

The performance of the InFusion pipeline was further tested by analyzing three public RNA-seq datasets from cancer studies: *Edgren et al* describes the transcriptome of cells originating from breast cancer, *Berger et al* from melanoma and *Wu et al* from prostate cancer. Each study provides evidence for fusion genes that were first detected from RNA-seq data and then experimentally validated using RT-PCR or Sanger sequencing. Reanalysis of the Edgren et al dataset performed by *Kangapeska et al* enabled detection and validation of an additional 13 fusions events, which we also included in our test. We assessed the performance of the analyzed tools by comparing the number of rediscovered known fusions and the total number of fusions reported by each algorithm. For this comparison we used gene annotations from Ensembl version 68 and specified default settings for each evaluated tool (more details in appendix section A.3). In most cases no exact genomic locations for fusions were reported in these studies, therefore we considered a fusion event as rediscovered if both fusion partner genes were reported correctly. Results are summarized in Table 3.3.

Table 3.3: *Fusion events detected in public RNA-seq datasets.*

Dataset	Sample	**Num. reads**	**Valid.**	**InFusion**	**Chimera Scan**	**TopHat Fusion**	**SOAP Fuse**	**Fusion Catcher**	**deFuse**
Edgren et al	KPL4	8.41M	3	3|5	3|29	3|11	3|6	3|3	2|8
Edgren et al	MCF7	6.8M	6	5|31	5|71	3|13	4|12	5|8	2|12
Edgren et al	SKBR3	21.43M	10	9|24	9|126	8|27	9|20	6|9	6|32
Edgren et al	BT474	18.15M	21	20|55	16|185	19|60	19|32	18|24	15|39
Berger et al	501Mel	14.86M	4	4|27	4|192	3|7	4|28	0|4	3|43
Berger et al	K562	31.35M	3	3|180	1|535	3|6	2|81	1|1	3|116
Berger et al	M000216	13.87M	1	1|32	1|129	0|1	1|10	0|0	1|20
Berger et al	M000921	14.47M	2	2|46	2|194	0|1	2|5	1|1	1|24
Berger et al	M010403	8.17M	1	1|23	1|84	0|0	1|12	0|0	1|17
Berger et al	M980409	15.77M	1	1|11	1|195	0|1	1|12	0|0	1|33
Berger et al	M990802	16.07M	2	1|20	1|198	0|1	2|11	0|0	1|24
Wu et al	LNCAP	85.39M	11	10|146	6|514	4|13	8|79	5|6	7|274
Wu et al	LTL313H	167.99M	15	11|223	5|733	7|26	5|173	4|88	8|324

"Valid." refers to the number of fusions qPCR-validated in the dataset as reported in the manuscript. For each fusion tool two values are provided, separted by column symbol. First value the number of previously validated fusions from the dataset, while second is the total number of fusions reported by the tool.

Overall, InFusion rediscovered the largest number of fusions reporting 71 events out of 80 previously described. In the dataset by Edgren et al InFusion achieved the highest detection rate, however, it missed three fusion events. One of them was not detected by any other tool, while two other fusions, reported by SOAPfuse and fusionCatcher, did not have enough reads to pass the filtering limits applied by InFusion. In the Berger et al dataset ChimeraScan demonstrated a higher detection rate than InFusion. It detected one additional fusion that was not discovered by the other algorithms. In this case the fusion was filtered out by InFusion due to its lack of reads spanning the corresponding junction. In the Wu et al dataset InFusion rediscovered the largest number of fusions. However InFusion filtered out five events while no unique reads covering the fusion junction were detected. These five fusions were also not detected by any other tool. The detailed list of fusion genes present in the datasets and their detection status can be found in appendix sections B.1,B.2 and B.3.

It is also noteworthy that in the Edgren et al dataset InFusion along with fusionCatcher and TopHat-Fusion, reported isoforms of several known fusions that have been experimentally validated and described in detail [Kangaspeska *et al.*, 2012]. Likewise in the Berger et al dataset InFusion detected the two distinct isoforms of fusion AXL-REC, reported by the authors [Berger *et al.*, 2010]. Both isoforms were also discovered by deFuse and fusionCatcher.

3.3.5 Experimental validation

To further investigate the power of the InFusion pipeline to detect chimeric transcripts from deep sequencing data, we sequenced the mRNA of two well-established prostate cancer cell lines, VCaP and LNCaP. Both cell lines are known to harbour genomic translocations and well-studied fusion genes [Maher *et al.*, 2009]. Using the strand-specific SENSE mRNA library preparation kit (Lexogen GmbH, Vienna, Austria), we constructed for each cell line two libraries with average insert sizes of 176 bp (referred to as VCap200 and LNCaP200) and 457 bp (referred to as VCap500 and LNCaP500). Sample allocation and detailed sequencing statistics are shown in Table 3.4.

Table 3.4: *RNA-Seq sample details.*

Cell line	**Sample**	**Insert size (bp)**	Total read pairs
VCaP	VCaP200	176	71,229,410
VCaP	VCaP500	457	7,653,307
VCaP	LNCaP200	176	74,575,800
VCaP	LNCaP500	457	4,521,899

List of samples with associated number of reads obtained and lane share from the deep sequencing of VCaP and LNCaP cell lines

We ran the InFusion pipeline on the datasets VCaP200 and LNCaP200 and detected 336 and 338 putative fusion events, respectively. The applied toolkit configuration is provided in appendix section A.4. After analysis we computed the exact number of detected fusions of a certain class as described previously, including isoforms and fusions with a breakpoint inside an exon, intron or an intergenic region (Table 3.5).

Table 3.5: *Fusion types detected in VCap and LNCaP cell lines.*

Fusion type	**VCaP200**	**VCaP500**	**LNCaP200**	**LNCaP500**
Total	336	151	338	87
Exon boundary break	36	6	31	4
With isoforms	10	1	1	1
Break inside exon	122	125	113	64
Break inside intron	84	8	99	4
Within Intergenic	94	12	95	15

Total *is the total number of fusions detected.* ***Exon boundary break*** *is the number of fusions where both 5' and 3' fusion breaks are on the exon boundary.* ***With isoforms*** *is the number of fusions of the same type that have several isoforms.* ***Break inside exon*** *is the number of fusions where one or both breaks are inside an exon.* ***Break inside intron*** *is the number of fusions where one or both breaks are inside an intron.* ***Within intergenic*** *is the number of fusions where one of the breaks is inside an intergenic region.*

For the experimental validation, aiming to cover the whole spectrum of fusion events, 21 candidates from VCaP and 19 from LNCaP were selected and subjected to qPCR. Ten out of these 40 were known fusions [Maher *et al.*, 2009] that we used as controls, while the others were novel and to the best of our knowledge have not been reported previously. Four events were selected as indicating anti-sense transcription and nine events as fusions of a coding with an intergenic or intronic region. Four selected events were isoforms of known fusions. The remaining 13 novel events were chosen randomly for validation.

The qPCR confirmed 36 out of the 40 selected chimeric transcripts, including all ten control fusions (Table 3.6). Five out of 26 novel events were validated in both cell lines, while the remaining 21 were specific for one cell line only. Only two fusions from 13 randomly chosen were not validated.

Interestingly, all the events verified in both cell lines appear to be intra-chromosomal with the exception of a single chimeric transcript that involves gene NBEA on chromosome 15 and an intergenic region on chromosome 13. A further four validated predictions (involving genes DIRC2, SPOCK1, SH3D19 and AMZ2) with an intergenic region as a second fusion partner were detected only in one cell line. Fusions INSL6 - JAK2 and two isoforms of ZDHHC7 - UNK present in the VCaP cell line have a breakpoint inside the intron of the 5' fusion gene partner. Notably, four confirmed events (POLR1D - LNX2, CTA-221G9.11 - KIAA1671, CTC-340A15.2 - PPIP5K2, RP11-534G20.3 - SVIL) indicate anti-sense transcription, emphasizing the value of a strand-specific library preparation.

To investigate the effect of the NGS library insert size on fusion detection, we analyzed datasets VCaP500 and LNCap500. These samples received only 9% and 5% of the reads of their VCaP200 and LNCap200 counterparts, respectively.

Table 3.6: *Fusions validated using qPCR from VCap and LNCaP cell lines.*

Name	Chr1	Pos1	Break1	Chr2	Pos2	Break2	Split reads	Bridge reads	Cell line
TMPRSS2 - ERG I1	21	42879876	Exon	21	39817547	Exon	20	0	VCaP
TMPRSS2 - ERG I2	21	42879876	Exon	21	39846047	Exon	11	0	VCaP
TMPRSS2 - ERG I3	21	42880007	Exon	21	39817547	Exon	5	0	VCaP
RC3H2-RGS3 I1	9	125622199	Exon	9	116299073	Exon	32	2	VCaP
RC3H2-RGS3 I2	9	125621318	In exon	9	116299074	Exon	8	1	VCaP
TIA1 - DIRC2	3	122552164	Exon	2	70475537	Exon	12	0	VCaP
LMAN2 - AP3S1	5	176778447	In exon	5	115202365	Exon	2	1	VCaP
HJURP - EIF4E2	2	234749255	Exon	2	233421124	Exon	59	12	VCaP
AAK1 - AC114772.1	2	69732701	Exon	2	69693684	In exon	33	0	Both
INSL6 - JAK2	9	51641179	Exon	9	49992434	Intron	87	3	VCaP
PPIP5K2 - CTC-340A15.2	5	102465407	Exon	5	164598384	Exon	11	26	VCaP
ZNF577 - ZNF841	19	52380532	Exon	19	52570866	Exon	24	1	VCaP
VWA2 - PRKCH	10	116008524	Exon	14	61909827	Exon	109	0	VCaP
CNNM4 - PARD3B	2	97474487	In exon	2	205829873	Exon	11	0	VCaP
ZDHHC7 - UNK I1	16	85029528	Exon	17	73782537	Intron	53	27	VCaP
ZDHHC7 - UNK I2	16	85027706	Intron	17	73782537	Intron	26	0	VCaP
ZDHHC7 - H3F3B	16	85029526	Exon	17	73775267	Exon	45	36	VCaP
SPOCK1 - Intergenic	5	136602698	Exon	5	180144798	Intergenic	15	3	VCaP
Intergenic - NBEA	13	35692611	Intergenic	15	20851765	Exon	44	0	Both
DIRC2 - Intergenic	3	122545912	Exon	2	64697744	Intergenic	3	2	VCaP
HSF1-RERE	8	145515556	Exon	1	8716501	Exon	16	0	VCaP
POLR1D - LNX2	13	28195176	In exon	13	28155942	Exon	18	0	Both
Intergenic - SH3D19	4	152246392	Intergenic	4	152147395	Exon	15	0	VCaP
AC024940.1 - FAM60A	12	31477418	In exon	12	31451159	Exon	9	0	VCaP
MIPOL1 - DGKB	14	37969347	In exon	7	14188860	Exon	16	0	LNCaP
RERE - PIK3CD	1	8482786	Exon	1	9770482	Exon	9	0	LNCaP
SLC45A3 - ELK4 I1	1	205630992	Exon	1	205593020	Exon	41	0	LNCaP
SLC45A3 - ELK4 I2	1	205628618	In exon	1	205593020	Exon	10	1	LNCaP
FAM117B - BMPR2	2	203500510	Exon	2	203329530	Exon	56	0	LNCap
GPS2 - MPP2	17	7218278	Exon	17	419757748	Exon	51	5	LNCaP
SREBF2 - XRCC6	22	42271728	Exon	22	42032115	Exon	587	0	LNCaP
CTA-221G9.11 - KIAA1671	22	25508430	In exon	22	25566787	Exon	22	0	Both
RP11-534G20.3 - SVIL	10	29704341	Exon	10	29746577	Inside exon	20	0	LNCaP
Intergenic - AMZ2	17	66202379	Intergenic	17	66246327	Exon	6	0	LNCap
RP11-180P8.1 - TANC2	17	61044108	In exon	17	61086895	In exon	9	0	Both
CASZ1 - KAZN	1	10820755	Exon	1	15070470	Inside exon	5	0	LNCap

*Fusions in bold were previously reported, others are novel. In column **Cell line** "Both" indicates that the fusion was confirmed both in VCaP and LNCaP.*

Strikingly, InFusion revealed 151 putative fusions in the VCaP500 sample and 87 in the LNCaP500 sample and moreover, we observed that only 15 and 9 fusions, respectively, were shared between the libraries with short and long inserts (Figure 3.6). The majority of fusion predictions are exclusive to either small or large insert size libraries.

Fourteen PCR-validated fusions were also found in datasets with larger insert size and lower coverage , and by qPCR we additionally confirmed one novel isoform of fusion SLC45A3 - ELK4, which was found only in the sample LNCaP500.

Interestingly, we detected and verified several novel splice variants of the known fusions TMPRSS2 - ERG, RC3H2 - RGS3 and SLC45A3 - ELK4. Three isoforms of TMPRSS2 - ERG were also tested by RT-PCR in the non-cancerous prostate cell lines RWPE-1 and PrEC, but there was no evidence of these transcripts in cell lines other than VCaP. Overall, the qPCR measurements of these three isoforms correlate with expression estimated from sequencing data . Sur-

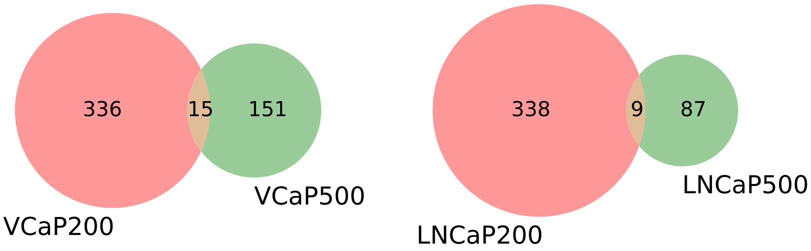

Figure 3.6: *Effect of insert size on fusion discovery. The figure shows comparison of fusion predictions for samples with different insert sizes in the same cell line. The lower number of fusions discovered reflects the lower sequencing depth in VCaP500 (9.2% of VCap200) and LNCaP500 (5.2% of LNCaP200)*

prisingly, isoform 3, which is the best described isoform, has the lowest expression level, as measured by qPCR, and the least support from the sequencing data compared to other isoforms.

Another intriguing observation concerns fusions that involve intergenic regions, which are typically ignored by other fusion detection tools. We discovered that they constitute a large proportion of the predictions. For example 94 out of 336 predictions in VCaP200 sample and 95 out of 338 predictions in LNCaP200 sample have an intergenic region as one part of a fusion. A curios example of such an event is a validated inter-chromosomal fusion that connects the DIRC2 gene and an intergenic region of chromosome 2, a fusion that is supported by a large clustering of reads downstream of the predicted breakpoint.

Additionally, we analyzed the cohort of experimentally validated fusions with TopHat-Fusion, deFuse, ChimeraScan SOAPfuse and fusionCatcher, as we did in previous comparisons (detailed results in appendix section B.4). While a number of novel fusion events was also detected by other tools, 15 novel fusions including several isoforms and events involving intronic and intergenic regions were detected only by InFusion. Certain isoforms of TMPRSS2 - ERG and RC3H2 - RGS3 were also detected by fusionCatcher and SOAPfuse, and two intronic fusions ZDHHC7 - UNK by TopHat-Fusion and deFuse. However, InFusion outperformed these tools in the total detection of these classes.

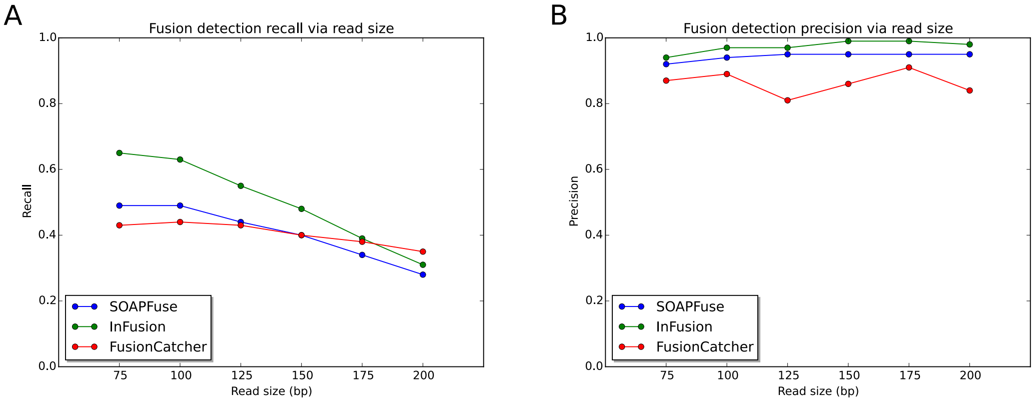

Figure 3.7: *Fusion simulation was performed for read size from 75bp to 200 bp (10 samples each experiment). Then fusion detection was applied by InFusion along with SOAPfuse and fusionCatcher. Recall (A) and precision (B) in fusion detection were computed for the tool results.*

3.3.6 Unanswered questions

Even though such effective detailed fusion discovery tools from RNA-seq as InFusion are available currently, there are several directions for future work.

Firstly, how exactly the read size benefits fusion detection remains to be determined. We performed additional simulation experiments to check the influence of read size (Figure 3.7). We found that InFusion maintains high precision in comparison to existing tools SOAPfuse and fusionCatcher, however, recall decreases with increase of the read size in all tools due to a growing number of breakpoints in SPLIT reads. We plan to improve the recall for larger read size in future versions of the program.

Second, estimating fusion gene expression remains an open problem. InFusion partly solves this issue by allowing output of possible fusion transcript sequences, which can be added to the transcriptome library so that expression can be quantified by applying methods such as RSEM [Li and Dewey, 2011] . However, we believe more robust solutions might be possible.

Third, the discovery of the fusion origin cannot be addressed using RNA-seq short reads alone. To determine if the chimeric transcript originates from a genomic translocation, from trans-splicing or from an experimental artifact, transcriptome sequencing should be combined with other experimental technologies. There are methods that combine long and short RNA-seq reads[Weirather *et al.*, 2015] or whole genome sequencing and transcriptome sequencing data to detect gene fusions [McPherson *et al.*, 2012], but more research work can be performed in this area.

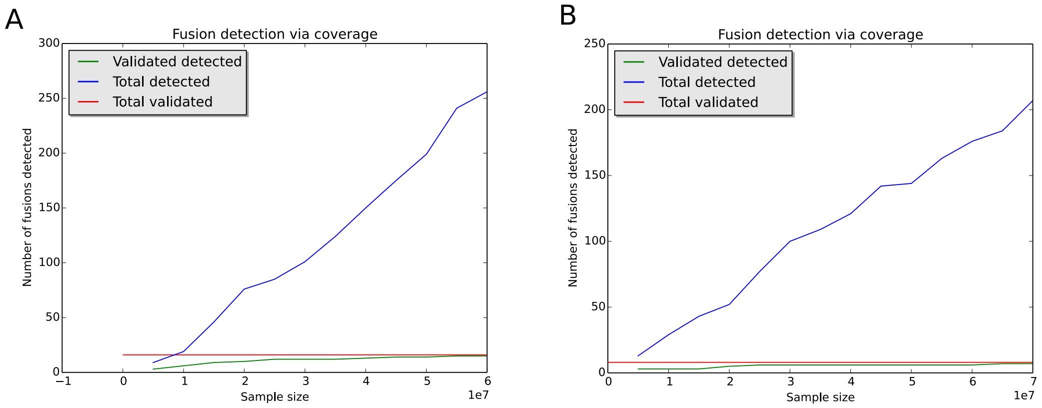

Figure 3.8: *The subsamples are generated from (A) VCaP200 and (B) LNCaP200 RNA-seq datasets. Each subsample is created randomly from the dataset based on the required size. The sample size changes from 5M reads to 70M with a step of 5M.*

3.4 Summary

We have designed and implemented a novel method for chimeric transcript discovery from RNA-seq data. Our method combines and improves ideas proposed by other researchers, such as assignment of reads that map to multiple loci and advanced fusion filtering. Additionally, it introduces several novel algorithmic aspects of chimeric RNA discovery, including intergenic region analysis, fusion cluster homogeneity estimation and consideration of protocol strand specificity. Our comparative analysis demonstrates that InFusion outperforms existing approaches for chimeric transcript discovery in detection accuracy.

Using simulation data we show that InFusion is able to discover a wide spectrum of fusion events that can occur in the transcriptome. Importantly, from our experimental data we discovered in-silico and verified in-vitro alternatively spliced fusion isoforms and chimeric RNAs involving non-exonic regions. In concordance with recent studies [Gonzàlez-Porta *et al.*, 2013] we observed that in most cases of fusion genes one transcript isoform is dominant and highly expressed, while the other isoforms are transcribed at significantly lower levels. However, this expression pattern may be completely different at a different time point or in another cell type, and isoforms might encode for RNAs or proteins of different functionality, which makes isoform detection important for differential gene expression analysis.

Remarkably, in our predictions from cancer data we observed a large number of fusions that involve intergenic regions , four of which were confirmed in vitro. To our knowledge, discovery of such events has not been addressed previously, despite their potential to encode functional proteins or regulate gene transcription.

An important factor influencing the detection of fusions from RNA-seq data is the depth of coverage of the sequencing experiment. Similar to novel transcript discovery and alternative splicing studies, fusion gene discovery benefits from higher coverage depth and longer reads [Sims *et al.*, 2014]. Our analysis shows that highly expressed fusions could be revealed even with relatively low coverage, however a gain in sequencing throughput increases the sensitivity of discovery (Figure 3.8).

Additionally, it is advantageous to use sequencing libraries with various insert sizes, since the fragment length affects the range of detectable fusion events in paired-end sequencing. A notable example of this effect is the additional validated isoform of fusion SLC45A3 - ELK4 detected only in the low-covered sample LNCaP500. Furthermore, strand-specific protocols are preferable for fusion discovery since they allow analysis of antisense transcription [Mills *et al.*, 2013]. InFusion uses information from strand specificity of the library to report antisense transcription in chimeric RNAs and also infers the transcription strand in case of a non-annotated region.

The computational efficiency of InFusion allows it to process large RNA-seq datasets comparatively quickly, e.g. it took approximately 10 h to analyze 74 million 100 bp paired reads on a machine with eight 2.4 GHz CPUs and the memory requirements did not exceed 30 GB.

Correlating fusions with cancer will continue to provide new insights into this disease and inform personalized therapy. Chimeric transcripts, on the other hand, have also been shown to occur in non-cancerous cells due to trans-splicing or transcriptome machinery failure, but the underlying mechanisms have yet to be studied in depth [Mertens *et al.*, 2015]. InFusion may prove to be a useful tool with high software quality in furthering our understanding in this area by detecting the whole scope of possible events

Chapter 4 Discussion

The transcriptome, as the connecting block between the genome and the proteome in a cell, has a number of interesting functions and properties. Deep understanding of transcriptome activity can provide progressive results in many possible research directions. And, as described in the previous chapters, RNA-seq has become an established method for the investigation of detailed transcriptome structure and functionality. Due to the progress in RNA-seq data analysis there are currently established concepts [Griffith *et al.*, 2015] along with detailed existing pipeline descriptions [Trapnell *et al.*, 2012; Anders *et al.*, 2013] available.

Moreover, the substantial novelties of RNA-seq technology opened the doors to explore additional elements of cell structure and activity. The method allowed detailed detection of isoforms, chimeric transcripts, fusion genes, strand-specificity and special types of RNA. Remarkably, correct design of RNA-seq experiment is quite important, since there is a number of parameters that influence experiment results and it is rather difficult to rerun the sequencing. Therefore, once the goal of research is established, the planning process should be performed rather carefully.

Notably, the advanced abilities of RNA sequencing technology continue the growth. Despite the large number of already well-known and established blocks of RNA-seq data analysis, some novelties require additional work. For example, the read length influences the transcriptome analysis significantly. As demonstrated previously, increasing the read size from 75bp to 200bp provides a lower mapping bias, reduces the ambiguity of alignment and advances the detection of splice variants [Cho *et al.*, 2014]. Also, research companies such as PacBio (www.pacificbiosciences.com) and Oxford Nanopore (www.nanoporetech.com) provide a novel sequencing approach in which the reads can have lengths up to 30 Kbp and cover a whole transcript. Long reads allow more precise detection of the transcript isoforms [Bolisetty *et al.*, 2015]. Also, as it was mentioned in the third chapter, application of longer reads might improve the detection of fusion genes and there are already approaches confirming this assumption [Davidson *et al.*, 2015; Weirather *et al.*, 2015]. However, longer reads require an update of algorithms to improve the alignment of fragments covering exon boundaries [Kim *et al.*, 2013]. Additionally, except of the long read sizes, such sequencing methods as PacBio have a rather high error rate (around 15%) and innovative algorithms are required for correct data processing and quality control.

Even though RNA-sequencing has become an advanced and important approach in the analysis of the transcriptome, there are still some poorly known biases that influence the results. Most of the errors can be detected and taken into account, however additional aspects need more investigation. For example, it was shown that the biases occurring during rRNA depletion and polyA selection can result in a 2-fold change in read coverage within over 50% of the expressed transcripts [Lahens *et al.*, 2014]. These biases require novel quality control and normalization approaches to provide correct gene expression analysis results.

Luckily, there are new technologies that are designed to improve the RNA sequencing process. As described earlier, a typical Illumina RNA-seq process is based on the cDNA conversion, which leads to a number of biases and errors, for example template switching and chimeric cDNA creation. Therefore, sequencing technologies that allow direct RNA sequencing (DRS) can become more effective and accurate. The first developed DRS approach was based on the Helicos sequencing platform [Ozsolak *et al.*, 2009]. It applies polyA RNA selection from total RNA and direct further processing without cDNA synthesis. The research in this area continues and DRS technologies will be useful for careful analysis of the transcriptome along with detection of short RNA blocks that cannot be converted to cDNA.

A perfect example of an advantageous RNA-seq innovation is single cell technology (scRNA-seq). Initial sequencing experiments required up to millions of cells, and the computed results such as gene expression levels were average values describing the groups of cells. Unfortunately, specific biological questions remained unanswered. For example, neuron cell groups were difficult to dissect and it was impossible to detect all cell types. Additionally during early embryonic development there is only a small number of cells with different functions, and the measures of gene expression were insufficient to figure out their functionality. To solve these issues, the single-cell method was introduced [Tang *et al.*, 2009]. This technique led to a number of discoveries including identification of novel cell types [Jaitin *et al.*, 2014], highly variable genes [Grün *et al.*, 2014] and global patterns of stochastic (per-cell) gene expression [Deng *et al.*, 2014]. However, to guarantee the correct results of the analysis, particular computational algorithms specified for single-cell data processing were required to be developed. There are currently a number of tasks in this process, for which specific tools and methods are still not available [Stegle *et al.*, 2015]. Especially, the quality control task is rather substantial and requires appropriate solutions.

It's also important to remember that a high number of samples in the experiment improves the reality of research results significantly. It was even shown that array-based analysis can only be trusted if there are enough samples [Marioni *et al.*, 2008]. Previously it was not possible to perform a lot of sequencing experiments, however currently the price of HTS falls down. This results in increase of samples number, and advanced statistics processing is required to understand the errors and collaboration of samples. Certain steps of this approach were already performed (for example, multi-sample analysis in the described Qualimap project), however more should follow.

A number of large scale projects and databases were created to adapt RNA-sequencing. One of these projects - Encyclopedia of the regulatory elements (ENCODE, http://genome.ucsc.edu/ENCODE/). It collected information about dozens of cells of various types. The goal of this project is to construct and validate a list of functional elements in the human cells, most importantly including the transcriptome and the proteome levels along with regulatory elements that control and report cells activity. The discovery of RNA activity is performed using RNA sequencing. Currently around 700 RNA-seq experiments datasets are available and there are more than 50 publications based on this data processing.

Another interesting example is The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov). The aim of this project is to collect and analyze patient's samples from different cancer tumor types in order to understand the underlying mechanisms of malignant transformation and progression. The database was found in 2008 and provided 25000 samples in 50 cancer types. At present it continues the development, focused on the finalization of a cell process in the most common cancer types. Here RNA-seq provides a unique snapshot of the gene expression status. Importantly, it makes possible accurate identification of novel isoforms of transcripts, fusion genes and non-coding RNAs that is quite hard to detect using other technologies.

It is worth noting that the described databases along with a majority of current molecular biology research projects require not only the transcriptome, but also the genome and the epigenome data. The cell structure consists of complicated mechanisms controlled by a number of elements. Therefore, precise comprehension of a cell process is not possible with analyzing only a single part and synergy between various analysis technologies is quite important [Bock, 2014]. Transcriptome investigation should be performed in collaboration with genomic, proteomic and epigenomic analysis. Only careful combination of the signals obtained through the harmonization of the whole cell system will allow to answer important questions [Kolker *et al.*, 2014]. However, of course, these answers can be trusted, only if each member of a full analysis pipeline is correct and detailed. Consequently, accurate and advanced improvements in RNA-sequencing processing along with adaption of the results to other existing methods will remain an important task. Furthermore, there might be specific aspects of transcriptome activity that are still unknown and the RNA-seq approach can help to propel innovative discoveries.

Appendix A Fusion discovery tools options

This section describes parameters applied for the fusion discovery tools in computational experiments.

A.1 General

In our comparison we analyzed output of the following fusion detection tools:

- InFusion v0.7.1: *fusions.txt*
- Chimerascan v.0.4.5 : *chimeras.bedpe*
- deFuse v0.6.0 : *results.filtered.tsv*
- TophatFusion v2.0.6 file: *result.txt*
- SOAPfuse v1.26 : *soap_sample.final.Fusion.specific.for.genes*
- fusionCatcher v0.99.4a,b : *final-list_candidate-fusion-genes.GRCh37.txt*

All described tools require genome and transcriptome sequences as well as gene annotations. For each tool we used hg19 human genome build. We applied Ensembl v.68 gene annotations for InFusion, SOAPfuse and deFuse, while TophatFusion, Chimerascan and fusionCatcher used an internal custom-formatted annotations based on Ensembl and UCSC RefSeq databases.

A.2 Simulated data analysis

In order to increase the spectrum of detected events we applied specific options for analyzed tools. Additionally, we set the same thresholds for the numbers of supporting SPLIT and BRIDGE reads where possible.

InFusion
We enabled the detection of non-coding and intergenic regions:

```
infusion –allow-non-coding –allow-intronic –allow-intergenic $READS_1 -2
$READS_2 /data/fusions-data/infusion.ens68.cfg
```

ChimeraScan

ChimeraScan has high sensitivity by default and does not allow setting minimum threshold for the number of supporting reads therefore we did not modify any options for this tool.

deFuse

We changed the minimum required number of supporting split and span segments in the configuration to 1 and 4 accordingly, which corresponds to the default InFusion values:

```
#
# Configuration file for deFuse
#

span_count_threshold = 4
split_count_threshold = 1
```

TophatFusion

When running TopHat we set the minimum fusion distance 20000, which corresponds to the default value of a similar InFusion parameter max-intron-size. Other parameters have been set as recommended by the authors of the program in the "Getting Started" tutorial:

```
 tophat -o tophat_simulation -p 4 –fusion-search –keep-fasta-order
–bowtie1 –no-coverage-search -r 100 –mate-std-dev 300 –fusion-min-dist 20000
/data/tophatruns/index/hg19 $READS_1 $READS_2
```

In the tophat-fusion-post script we skipped the BLAST filtering step in order to increase sensitivity of the discovery and additionally set the thresholds for supporting reads which correspond to the defaults in InFusion:

```
 tophat-fusion-post -p 4 –num-fusion-both 4 –num-fusion-reads 1
–num-fusion-pairs 0 –skip-blast $TH_DATA/index/hg19
```

SOAPfuse

In main settings of SOAPfuse we did not change any requirements. However the required correct insert size in configuration file for each experiment was provided.

fusionCatcher

We used default configuration parameters for fusionCatcher since it adapts correctly to the provided data without additional settings.

The pipeline depends on a certain number of tools with fixed version. Unfortunately, certain bugs were introduced by STAR aligner as one of the included tools; however automatic restart.sh script was used to rerun analysis of failed experiments. Additionally the novel version of fusionCatcher (0.99.4b) relies on a new version of STAR, and these problems were not seen again.

A.3 Public datasets

For InFusion we applied the following configuration (options are designed based on low coverage, small read size and types of reported fusions in publications):

```
infusion –allow-intronic –allow-non-coding –do-coverage-analysis
–req-homogeneity-weight 0.15 –min-unique-split-reads 0 -1 $READS_1 -2 $READS_2
/data/fusions-data/infusion.ens68.cfg
```

Additionally, due to differences in sizes of the datasets (from 6 millions of reads up to 85 millions) certain options were configured. For example, parameters *–min-bckg-reads* and *–min-unique-split* parameters were adapted in large datasets.

For other tools we used parameters provided in publications (when available) or default parameters.

A.4 In-house datasets

We applied the following configuration for analysis of deep sequencing data from VCap and LNCap cell lines:

```
infusion –library RF –alow-intronic –allow-intergenic –allow-non-coding
–min-split-reads 2 –min-span-pairs 0 –min-fragments 3 –min-unique-split-reads 0
–min-unique-alignment-rate 0.04 -1 $READS_1 -2 $READS_2
/data/fusions-data/infusion.ens68.cfg
```

Appendix B

Fusion discovery supplementary tables

This section provides tables with results of fusion detection in various datasets. *Note: in each table the names of some tools are abbreviated (ChimeraScan = CScan, TopHat-Fusion = TFusion, SOAPfuse = SFuse, fusionCatcher = fCatcher)*

B.1 Edgren et al. dataset

Fusion genes in the Edgren et. al dataset. Additionally the table includes fusions detected and validated by Kangapeska et al. (in bold).

Sample	**Fusion gene**	**InFusion**	**deFuse**	**CScan**	**TFusion**	**SFuse**	**fCatcher**
KPL4	BSG-NFIX	+	+	+	+	+	+
KPL4	PPP1R12A-SEPT10	+	-	+	+	+	+
KPL4	NOTCH1-NUP214	+	+	+	+	+	+
MCF7	BCAS4-BCAS3	+	+	+	+	+	+
MCF7	ARFGEF2-SULF2	+	+	+	+	+	+
MCF7	RPS6KB1-TMEM49	+	-	+	+	+	+
MCF7	**GCN1L1-MSI1**	+	+	+	-	-	-
MCF7	**AC099850.1-VMP1**	+	-	-	-	-	+
MCF7	**SMARCA4-CARM1**	-	-	+	-	+	+
SKBR3	TATDN1-GSDMB	+	+	-	+	+	+
SKBR3	CSE1L-ENSG00000236127	-	-	-	+	-	-
SKBR3	RARA-PKIA	+	+	+	+	+	+
SKBR3	ANKHD1-PCDH1	+	+	+	+	+	+
SKBR3	CCDC85C-SETD3	+	-	-	+	+	+
SKBR3	SUMF1-LRRFIP2	+	+	+	+	+	+
SKBR3	WDR67-ZNF704	+	-	+	+	+	+
SKBR3	CYTH1-EIF3H	+	+	-	+	+	+
SKBR3	DHX35-ITCH	+	-	+	-	+	-
SKBR3	NFS1-PREX1	+	+	+	-	+	-
BT474	ACACA-STAC2	+	+	+	+	+	+
BT474	RPS6KB1-SNF8	+	+	-	+	+	+
BT474	VAPB-IKZF3	+	+	+	+	+	+
BT474	ZMYND8-CEP250	+	+	+	+	+	+
BT474	RAB22A-MYO9B	+	+	+	+	+	+
BT474	SKA2-MYO19	+	+	+	+	+	+
BT474	DIDO1-TTI1	+	+	+	+	+	-
BT474	STARD3-DOK5	+	+	+	+	+	+
BT474	LAMP1-MCF2L	+	+	+	+	+	-
BT474	GLB1-CMTM7	+	+	+	+	+	+
BT474	CPNE1-PI3	+	-	-	+	-	+
BT474	**THRA-AC090627.1**	+	-	-	+	-	+
BT474	**TOB1-SYNRG**	+	-	+	+	+	+
BT474	**AHCTF1-NAAA**	+	+	-	+	+	+
BT474	**MED1-STXBP4**	+	+	+	+	+	+
BT474	**MED13-BCAS3**	+	-	+	+	+	+
BT474	**MED1-ACSF2**	+	+	+	+	+	+
BT474	**TRPC4AP-MRPL45**	+	+	+	+	+	+
BT474	**STX16-RAE1**	+	+	+	+	+	+
BT474	**USP32-MED1**	-	-	+	-	+	+
BT474	**PIP4K2B-RAD51C**	+	-	+	-	+	-

B.2 Berger et al. dataset

Fusion genes in the Berger et. al dataset.

Sample	Fusion gene	InFusion	deFuse	CScan	TFusion	Sfuse	fCatcher
501Mel	CCT3-C1orf61	+	+	+	+	+	-
501Mel	GNA12-SHANK2	+	+	+	+	+	-
501Mel	SLC12A7-C11orf67	+	+	+	-	+	-
501Mel	PARP1-MIXL1	+	-	+	+	+	-
K562	BCR-ABL1	+	+	+	+	+	+
K562	NUP214-XKR3	+	+	+	+	-	-
K562	BAT3-SLC44A4	+	+	+	+	+	-
M000216	KCTD2-ARHGEF12	+	+	+	-	+	-
M000921	TMEM8B-TLN1	+	-	+	-	+	-
M000921	RECK-ALX3	+	+	+	-	+	+
M010403	SCAMP2-WDR72	+	+	+	-	+	-
M980409	GCN1L1-PLA2G1B	+	+	+	-	+	-
M990802	ANKHD1-C5orf32	+	-	+	-	+	-
M990802	RB1-ITM2B	-	+	+	-	+	-

B.3 Wu et al. dataset

Fusion genes in the Wu et. al dataset.

Sample	Fusion gene	InFusion	deFuse	CScan	TFusion	Sfuse	fCatcher
LNCAP	DLEU2-PSPC1	+	-	-	-	+	-
LNCAP	RERE-PIK3CD	+	+	+	+	+	+
LNCAP	MIPOL1-DGKB	+	+	+	+	+	+
LNCAP	MRPS10-HPR	+	-	+	-	+	-
LNCAP	C19orf25-APC2	-	-	-	-	-	-
LNCAP	SLC45A3-ELK4	+	+	-	-	-	+
LNCAP	TFDP1-GRK1	+	+	+	-	+	-
LNCAP	FAM117B-BMPR2	+	+	+	+	+	+
LNCAP	GPS2-MPP2	+	+	+	-	+	-
LNCAP	ITPKC-PPFIA3	+	+	-	+	+	+
LNCAP	CCDC43-YBX2	+	+	+	-	-	-
LTL313H	SLC6A17-CA2	-	-	-	-	-	-
LTL313H	PACS1-PTEN	+	+	+	+	+	-
LTL313H	TMPRSS2-ERG	+	+	+	+	+	+
LTL313H	SSR1-RNF165	+	+	-	+	-	-
LTL313H	SLC35B1-PEMT	+	+	+	+	+	+
LTL313H	USP34-C2orf74	-	-	-	-	-	-
LTL313H	TRAPPC9-SPRYD3	+	-	-	-	-	-
LTL313H	TENC1-TTLL9	+	+	-	-	-	-
LTL313H	KIAA1467-TTLL9	+	-	+	-	-	-
LTL313H	TTLL9-C12orf59	+	-	-	-	-	-
LTL313H	CA2-RUNX1T1	-	-	-	-	-	-
LTL313H	PDRG1-ARF3	-	-	-	+	-	-
LTL313H	PDRG1-RUNX1T1	+	+	-	+	-	-
LTL313H	EEF1D-SDC4	+	-	-	+	+	+
LTL313H	SPRYD3-PTDSS1	+	+	+	-	+	+

B.4 VCaP and LNCaP cell lines

The table indicates if a particular event detected by InFusion and validated using qPCR was also reported by tools deFuse, TophatFusion, ChimeraScan, SOAPfuse and fusionCatcher. Fusions in bold are detected and reported only by InFusion.

Fusion transcript	**deFuse**	**TFusion**	**CScan**	**Sfuse**	**fCatcher**
TMPRSS2 - ERG I1	+	-	+	+	+
TMPRSS2 - ERG I2	-	-	-	-	+
TMPRSS2 - ERG I3	-	-	-	-	+
RC3H2 - RGS3 I1	+	+	+	+	+
RC3H2 - RGS3 I2	-	-	-	+	+
TIA1 - DIRC2	-	-	+	-	-
LMAN2 - AP3S1	-	-	+	-	-
HJURP - EIF4E2	+	+	+	+	+
AAK1 - AC114772.1	-	-	-	-	-
INSL6 - JAK2 (intronic)	-	-	-	-	-
PPIP5K2 - CTC-340A15.2	+	-	-	-	-
ZNF577 - ZNF841	-	-	+	-	-
VWA2 - PRKCH	+	+	+	+	-
CNNM4 - PARD3B	-	-	+	-	-
ZDHHC7 - UNK I1 (intronic)	+	+	-	-	-
ZDHHC7 - UNK I2 (intronic)	+	+	-	-	-
ZDHHC7 - H3F3B	+	+	+	+	-
SPOCK1 - INTERGENIC	+	-	-	-	-
INTERGENIC - NBEA	-	-	-	-	-
DIRC2 - intergenic	-	-	-	-	-
HSF1 - RERE	-	+	+	+	-
POLR1D - LNX2	-	-	-	-	-
INTERGENIC - SH3D19	-	-	-	-	-
AC024940.1 - FAM60A	-	-	-	-	-
MIPOL1 - DGKB	-	-	-	-	-
RERE-PIK3CD	+	-	-	-	-
SLC45A3 - ELK4 I1	+	-	-	+	+
SLC45A3 - ELK4 I2	-	-	-	-	-
FAM117B - BMPR2	+	-	+	+	+
GPS2 - MPP2	+	+	+	+	+
SREBF2 - XRCC6	+	+	+	+	+
CTA-221G9.11 - KIAA1671	-	-	-	-	-
RP11-534G20.3 - SVIL	-	-	-	-	-
INTERGENIC - AMZ2	-	-	-	-	-
RP11-180P8.1 - TANC2	-	-	-	-	-
CASZ1 - KAZN	-	-	-	-	-

Appendix C

List of abbreviations

DNA:	Deoxyribonucleic acid
RNA:	Ribonucleic acid
bp:	base pair, character of the alphabet $\{ A,C,G,T \}$
NGS:	Next Generation Sequencing
HTS:	High-Throughput Sequencing
RPKM:	Reads Per Kilobase per Million mapped reads
FPKM:	Fragments Per Kilobase per Million mapped reads
PCR:	polymerase chain reaction
RT-PCR:	real-time PCR
FISH:	fluorescent in situ hybridization
QC:	quality control

APPENDIX D

Curriculum Vitae

Konstantin Okonechnikov

****
****
***
****
****

Education

2011–2015	**PhD in Bioinformatics**, *Freie University, Berlin, Germany.* The scientific work was performed in Max Planck Institute for Infection Biology, Berlin.
2009–2011	**MSc in Physics**, *Novosibirsk State University, Russia.* Specialization: information systems and networks.Master diploma thesis: "Protein-protein interaction prediction based on primary structure".
2005–2009	**MSc in Physics**, *Novosibirsk State University, Russia.* Bachelor diploma thesis: "Analysis and Visualization of Biological Macromolecular 3D structures" .

Professional experience

2011–2015	**PhD in Bioinformatics**, *Max Planck Insitute For Infection Biology, Berlin, Germany.* Performing data analysis in several biological projects related to investigation of infections and viruses. Examples: RNA-seq data processing of H.pylori infected cells, mutation rates analysis in H1N1 virus, fusion detection in infected and cancer cells. Additionally, development of specific novel tools related to high throughput sequencing data analysis.

2009–2011 **Scientific software developer**, *Unipro.* Participation in development of Unipro UGENE project (open-source bioinformatics toolkit): suffix array based algorithm for DNA short reads alignment, "cloning in silico" tools, cloud computing service, improvements in several other modes.

Teaching experience

Bioinforamtics Summer School 2014, Saint-Petersburg, Russia. Lecture: "Quality control of next generation sequencing data"; seminars: "NGS data analysis: basic QC", "Practical RNA-seq".

Bioinforamtics Summer School 2013, Moscow, Russia. Lecture: "Gentle introduction to RNA-sequencing"; seminars: "Scientific software development: best practices and approaches", "In-lab bioinformatics: making sense of NGS data".

Technical skills

Programming: object oriented design patterns, high-performance computing, parallel and multiplatform programming principles

Programming languages: C++, Python, Java, R, Bash, LaTeX

Bioinformatics tools: sequencing data analysis (Bowtie, BWA, Tophat, Velvet, SAMtools, BamView, Tablet, ANNOWAR,...), sequence homology search (BLAST, SW), multiple alignment (MUSCLE, KAlign), phylogenetics (Phylip, Mr. Bayes)

Publications

Konstantin Okonechnikov, Ana Conesa and Fernando García-Alcalde. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data" Bioinformatics (2015)

Frithjof Glowinski, Fernando Garcia-Alcalde, *Konstantin Okonechnikov*, and Thomas F. Meyer. "Modulation of the host cell RNA splicing program by the gastric pathogen Helicobacter pylori." EMBnet. journal 19, no. A (2013): pp-70.

Fernando García-Alcalde, *Konstantin Okonechnikov*, José Carbonell, Luis M. Cruz, Stefan Götz, Sonia Tarazona, Joaquín Dopazo, Thomas F. Meyer, and Ana Conesa. "Qualimap: evaluating next-generation sequencing alignment data." Bioinformatics 28, no. 20 (2012): 2678-2679.

Konstantin Okonechnikov, Olga Golosova, and Mikhail Fursov. "Unipro UGENE: a unified bioinformatics toolkit." Bioinformatics 28.8 (2012): 1166-1167.

Presentations and talks

QualiMap 2: advanced quality control of high throughput sequencing data. In *Bioinformatics Open Source Conference, ISMB/ECCB 2015, Dublin, Ireland*

QualiMap: quality control of high throughput sequencing data. In *Bioinformatics Open Source Conference, ISMB/ECCB 2013, Berlin, Germany*

Unipro UGENE: a unified bionformatics toolkit. In *Bioinformatics Open Source Conference, ISMB/ECCB 2011, Vienna, Austria*

Languages

Russian	Native
English	Advanced
German	Average

Awards

Winner of the "Win your SENSE Study" award from Lexogen company (2013)

Appendix E
Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Konstantin Okonechnikov
February 23, 2016

BIBLIOGRAPHY

Ambros, V. (2004). The functions of animal micrornas. *Nature*, **431**(7006), pages 350–355.

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

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Genome research*, **22**(10), pages 2008–2017.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, **8**(9), pages 1765–1786.

Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638.

Becker-Andre, M. and Hahlbrock, K. (1989). Absolute mrna quantification using the polymerase chain reaction (pcr). a novel approach by a pcr aided transcipt titration assay (patty). *Nucleic acids research*, **17**(22), pages 9437–9446.

Berger, M. F., Levin, J. Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L. A., Robinson, J., Verhaak, R. G., Sougnez, C., *et al.* (2010). Integrative analysis of the melanoma transcriptome. *Genome research*, **20**(4), pages 413–427.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., *et al.* (2013). An estimation of the number of cells in the human body. *Annals of human biology*, **40**(6), pages 463–471.

Bock, C. (2014). Synergy and competition between cancer genome sequencing and epigenome mapping projects. *Genome medicine*, **6**(5), pages 1–3.

Bolisetty, M., Rajadinakaran, G., and Graveley, B. (2015). Determining exon connectivity in complex mrnas by nanopore sequencing. *bioRxiv*, page 019752.

Brantl, S. (2007). Regulatory mechanisms employed by cis-encoded antisense rnas. *Current opinion in microbiology*, **10**(2), pages 102–109.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, **11**(1), page 94.

Carrara, M., Beccuti, M., Lazzarato, F., Cavallo, F., Cordero, F., Donatelli, S., and Calogero, R. A. (2013). State-of-the-art fusion-finder algorithms sensitivity and specificity. *BioMed research international*, **2013**.

Chen, K., Wallis, J. W., Kandoth, C., Kalicki, J. M., Mungall, K. L., Mungall, A. J., Jones, S. J., Marra, M. A., Ley, T. J., Mardis, E. R., *et al.* (2012). Breakfusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics*, **28**(14), pages 1923–1924.

Cho, H., Davis, J., Li, X., Smith, K. S., Battle, A., and Montgomery, S. B. (2014). High-resolution transcriptome analysis with long-read rna sequencing. *PLOS One*.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), pages 57–74.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**(5909), pages 1845–1848.

Crick, F. H. (1958). On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138.

Davidson, N. M., Majewski, I. J., and Oshlack, A. (2015). Jaffa: High sensitivity transcriptome-focused fusion gene detection. *bioRxiv*, page 013698.

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

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**(6167), pages 193–196.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**(1), pages 15–21.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). Seqan an efficient, generic c++ library for sequence analysis. *BMC bioinformatics*, **9**(1), page 11.

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I. H., Nyberg, S., Wolf, M., Borresen-Dale, A.-L., *et al.* (2011). Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biol*, **12**(1), page R6.

Faulhammer, D., Lipton, R. J., and Landweber, L. F. (2000). Fidelity of enzymatic ligation for dna computing. *Journal of Computational Biology*, **7**(6), pages 839–848.

Fernandez-Cuesta, L., Sun, R., Menon, R., George, J., Lorenz, S., Meza-Zepeda, L. A., Peifer, M., Plenker, D., Heuckmann, J. M., Leenders, F., *et al.* (2015). Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome biology*, **16**(1), page 7.

Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Del Pozo, A., Tress, M., Johnson, R., Guigo, R., *et al.* (2012). Chimeras taking shape: potential functions of proteins encoded by chimeric rna transcripts. *Genome research*, **22**(7), pages 1231–1242.

Frenkel-Morgenstern, M., Gorohovski, A., Vucenovic, D., Maestre, L., and Valencia, A. (2014). Chitars 2.1âĂŤan improved database of the chimeric transcripts and rna-seq data with novel sense–antisense chimeric rna transcripts. *Nucleic acids research*, page gku1199.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., *et al.* (2009). Estimating accuracy of rna-seq and microarrays with proteomics. *BMC genomics*, **10**(1), page 161.

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

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol*, **14**(7), page R70.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, **29**(7), pages 644–652.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., and Griffith, O. L. (2015). Informatics for rna sequencing: A web resource for analysis on the cloud. *PLoS Comput Biol*, **11**(8), page e1004393.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *nature methods*, **11**(6), pages 637–640.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, **38**(12), pages e131–e131.

Harrington, C. A., Rosenow, C., and Retief, J. (2000). Monitoring gene expression using dna microarrays. *Current opinion in Microbiology*, **3**(3), pages 285–291.

Hayer, K., Pizzaro, A., Lahens, N. L., Hogenesch, J. B., and Grant, G. R. (2014). Benchmark analysis of algorithms for determining and quantifying full-length mrna splice forms from rna-seq data. *bioRxiv*, page 007088.

He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., Wang, Z., Chen, F., Lindquist, E. A., Sorek, R., *et al.* (2010). Validation of two ribosomal rna removal methods for microbial metatranscriptomics. *Nature Methods*, **7**(10), pages 807–812.

Holtgrewe, M. (2010). Mason–a read simulator for second generation sequencing data. *Technical Report FU Berlin*.

Hu, Y., Wang, K., He, X., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). A probabilistic framework for aligning paired-end rna-seq data. *Bioinformatics*, **26**(16), pages 1950–1957.

Hu, Y., Liu, Y., Mao, X., Jia, C., Ferguson, J. F., Xue, C., Reilly, M. P., Li, H., and Li, M. (2014). Pennseq: accurate isoform-specific gene expression quantification in rna-seq by modeling non-uniform read distribution. *Nucleic acids research*, **42**(3), pages e20–e20.

Iyer, M. K., Chinnaiyan, A. M., and Maher, C. A. (2011). Chimerascan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**(20), pages 2903–2904.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, **343**(6172), pages 776–779.

Jia, W., Qiu, K., He, M., Song, P., Zhou, Q., Zhou, F., Yu, Y., Zhu, D., Nickerson, M. L., Wan, S., *et al.* (2013). Soapfuse: an algorithm for identifying fusion transcripts from paired-end rna-seq data. *Genome biology*, **14**(2), page R12.

Kangaspeska, S., Hultsch, S., Edgren, H., Nicorici, D., Murumägi, A., and Kallioniemi, O. (2012). Reanalysis of rna-sequencing data reveals several additional fusion genes with multiple isoforms. *PloS one*, **7**(10), page e48745.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, **7**(12), pages 1009–1015.

Kim, D. and Salzberg, S. L. (2011). Tophat-fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*, **12**(8), page R72.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**(4), page R36.

Kim, Y.-K., Yeo, J., Kim, B., Ha, M., and Kim, V. N. (2012). Short structured rnas with low gc content are selectively lost during extraction from a small number of cells. *Molecular cell*, **46**(6), pages 893–895.

Kinsella, M., Harismendy, O., Nakano, M., Frazer, K. A., and Bafna, V. (2011). Sensitive gene fusion detection using ambiguously mapping rna-seq read pairs. *Bioinformatics*, **27**(8), pages 1068–1075.

Kleinman, C. L., Gerges, N., Papillon-Cavanagh, S., Sin-Chan, P., Pramatarova, A., Quang, D.-A. K., Adoue, V., Busche, S., Caron, M., Djambazian, H., *et al.* (2014). Fusion of ttyh1 with the c19mc microrna cluster drives expression of a brain-specific dnmt3b isoform in the embryonal brain tumor etmr. *Nature genetics*, **46**(1), pages 39–44.

Koeppel, M., Garcia-Alcalde, F., Glowinski, F., Schlaermann, P., and Meyer, T. F. (2015). Helicobacter pylori infection causes characteristic dna damage patterns in human cells. *Cell reports*, **11**(11), pages 1703–1713.

Kolker, E., Özdemir, V., Martens, L., Hancock, W., Anderson, G., Anderson, N., Aynacioglu, S., Baranova, A., Campagna, S. R., Chen, R., *et al.* (2014). Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics: a journal of integrative biology*, **18**(1), pages 10–14.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+ c)-biased genomes. *Nature methods*, **6**(4), pages 291–295.

Łabaj, P. P., Leparc, G. G., Linggi, B. E., Markillie, L. M., Wiley, H. S., and Kreil, D. P. (2011). Characterization and improvement of rna-seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**(13), pages i383–i391.

Lahens, N. F., Kavakli, I. H., Zhang, R., Hayer, K., Black, M. B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R. S., *et al.* (2014). Ivt-seq reveals extreme bias in rna sequencing. *Genome Biol*, **15**(6), page R86.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), pages 357–359.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., *et al.* (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biol*, **10**(3), page R25.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nature methods*, **7**(9), pages 709–715.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, **12**(1), page 323.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4), pages 493–500.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**(14), pages 1754–1760.

Li, S., Liberman, L. M., Mukherjee, N., Benfey, P. N., and Ohler, U. (2013). Integrated detection of natural antisense transcripts using strand-specific rna sequencing data. *Genome research*, **23**(10), pages 1730–1739.

Li, Y., Chien, J., Smith, D. I., and Ma, J. (2011). Fusionhunter: identifying fusion transcripts in cancer using paired-end rna-seq. *Bioinformatics*, **27**(12), pages 1708–1710.

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

Lugo, T. G., Pendergast, A.-M., Muller, A. J., and Witte, O. N. (1990). Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science*, **247**(4946), pages 1079–1082.

Maher, C. A., Palanisamy, N., Brenner, J. C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T. R., Grasso, C., Yu, J., *et al.* (2009). Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences*, **106**(30), pages 12353–12358.

Mamatis, T. (1987). The role of small nuclear ribonucleoprotein particles in pre-mrna splicing. *Nature*, **325**, page 673.

Maniatis, T. and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*, **416**(6880), pages 499–506.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), pages 1509–1517.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G., Griffith, M., Moussavi, A. H., Senz, J., Melnyk, N., *et al.* (2011). defuse: an algorithm for gene fusion discovery in tumor rna-seq data. *PLoS computational biology*, **7**(5), page e1001138.

McPherson, A., Wu, C., Wyatt, A. W., Shah, S., Collins, C., and Sahinalp, S. C. (2012). nfuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome research*.

Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, **15**(6), pages 371–381.

Mezlini, A. M., Smith, E. J., Fiume, M., Buske, O., Savich, G. L., Shah, S., Aparicio, S., Chiang, D. Y., Goldenberg, A., and Brudno, M. (2013). ireckon: Simultaneous isoform discovery and abundance estimation from rna-seq data. *Genome research*, **23**(3), pages 519–529.

Mills, J. D., Kawahara, Y., and Janitz, M. (2013). Strand-specific rna-seq provides greater resolution of transcriptome profiling. *Current genomics*, **14**(3), page 173.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, **7**(4), pages 233–245.

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J., and Marra, M. A. (2008). Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, **45**(1), page 81.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**(7), pages 621–628.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1992). Specific enzymatic amplification of dna in vitro: the polymerase chain reaction. *Biotechnology Series*, pages 17–17.

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

Nicorici, D., Satalan, M., Edgren, H., Kangaspeska, S., Murumagi, A., Kallioniemi, O., Virtanen, S., and Kilkku, O. (2014). Fusioncatcher-a tool for finding somatic fusion genes in paired-end rna-sequencing data. *bioRxiv*, page 011650.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, page btv566.

Oshlack, A., Wakefield, M. J., *et al.* (2009). Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, **4**(1), page 14.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct rna sequencing. *Nature*, **461**(7265), pages 814–818.

Panagopoulos, I., Thorsen, J., Gorunova, L., Haugom, L., Bjerkehagen, B., Davidson, B., Heim, S., and Micci, F. (2013). Fusion of the zc3h7b and bcor genes in endometrial stromal sarcomas carrying an x; 22-translocation. *Genes, Chromosomes and Cancer*.

Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A., and Abel, T. (2015). How data analysis affects power, reproducibility and biological insight of rna-seq studies in complex datasets. *Nucleic acids research*, **43**(16), pages 7664–7674.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol*, **14**(9), page R95.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding rnas. *Annual review of biochemistry*, **81**.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, **32**(9), pages 896–902.

Robert, C. and Watson, M. (2015). Errors in rna-seq quantification affect genes of relevance to human disease. *Genome biology*, **16**(1), pages 1–16.

Robinson, D. R., Wu, Y.-M., Kalyana-Sundaram, S., Cao, X., Lonigro, R. J., Sung, Y.-S., Chen, C.-L., Zhang, L., Wang, R., Su, F., *et al.* (2013). Identification of recurrent nab2-stat6 gene fusions in solitary fibrous tumor by integrative sequencing. *Nature genetics*, **45**(2), pages 180–185.

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

Robinson, M. D., Oshlack, A., *et al.* (2010b). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, **11**(3), page R25.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, **281**(5375), pages 363–365.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, **14**(5), page R51.

Rubin, B. P., Chen, C.-J., Morgan, T. W., Xiao, S., Grier, H. E., Kozakewich, H. P., Perez-Atayde, A. R., and Fletcher, J. A. (1998). Congenital mesoblastic nephroma t (12; 15) is associated with< i> etv6-ntrk3</i> gene fusion: Cytogenetic and molecular relationship to congenital (infantile) fibrosarcoma. *The American journal of pathology*, **153**(5), pages 1451–1458.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12), pages 5463–5467.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., *et al.* (2010). Fusionseq: a modular framework for finding gene fusions by analyzing paired-end rna-sequencing data. *Genome Biol*, **11**(10), page R104.

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**(8), pages 1086–1092.

Shah, N., Lankerovich, M., Lee, H., Yoon, J.-G., Schroeder, B., and Foltz, G. (2013). Exploration of the gene fusion landscape of glioblastoma using transcriptome sequencing and copy number data. *BMC genomics*, **14**(1), page 818.

Sheng, W.-Q., Hisaoka, M., Okamoto, S., Tanaka, A., Meis-Kindblom, J. M., Kindblom, L.-G., Ishida, T., Nojima, T., and Hashimoto, H. (2001). Congenital-infantile fibrosarcoma a clinicopathologic study of 10 cases and molecular detection of the etv6-ntrk3 fusion transcripts using paraffin-embedded tissues. *American journal of clinical pathology*, **115**(3), pages 348–355.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**(2), pages 121–132.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1), pages 195–197.

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, **16**(3), pages 133–145.

Steidl, C., Shah, S. P., Woolcock, B. W., Rui, L., Kawahara, M., Farinha, P., Johnson, N. A., Zhao, Y., Telenius, A., Neriah, S. B., *et al.* (2011). Mhc class ii transactivator ciita is a recurrent gene fusion partner in lymphoid cancers. *Nature*, **471**(7338), pages 377–381.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., *et al.* (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, **6**(5), pages 377–382.

Tarazona, S., García, F., Ferrer, A., Dopazo, J., and Conesa, A. (2012). Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, **17**(B), pages pp–18.

Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic acids research*, page gkv711.

Tognon, C., Knezevich, S. R., Huntsman, D., Roskelley, C. D., Melnyk, N., Mathers, J. A., Becker, L., Carneiro, F., MacPherson, N., Horsman, D., *et al.* (2002). Expression of the etv6-ntrk3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer cell*, **2**(5), pages 367–376.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., *et al.* (2005). Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, **310**(5748), pages 644–648.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9), pages 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, **28**(5), pages 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, **7**(3), pages 562–578.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, **31**(1), pages 46–53.

van Dijk, E. L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, **322**(1), pages 12–20.

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

Watson, J. D., Crick, F. H., *et al.* (1953). Molecular structure of nucleic acids. *Nature*, **171**(4356), pages 737–738.

Weirather, J. L., Afshar, P. T., Clark, T. A., Tseng, E., Powers, L. S., Underwood, J., Zabner, J., Korlach, J., Wong, W. H., and Au, K. F. (2015). Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Research*, page 1.

Wu, T. D. and Nacu, S. (2010). Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**(7), pages 873–881.

Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R. (2014). The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*.

LIST OF FIGURES

1.1 Central dogma of molecular biology 2
1.2 Transcription process in high eukaryotes 4
1.3 The RNA-sequencing process structure 7
1.4 RNA-sequencing data analysis overview 8

2.1 RNA-seq QC plot examples 19
2.2 Counts QC global plot examples 21
2.3 Counts QC individual sample plot examples 22
2.4 Counts QC comparison plot example 24
2.5 Detection of biases from Counts QC report 25

3.1 Detection of fusion events via RNA-seq reads 32
3.2 InFusion pipeline structure 35
3.3 Clustering of fusions process. 39
3.4 Fusion detection from simulation data. 43
3.5 Comparison of recall and precision on simulation data per fusion class 45
3.6 Effect of insert size on fusion discovery 50
3.7 Read size influence on recall and precision in fusion detection. . . 51
3.8 Fusion detection performed from sumbsamples 52

LIST OF TABLES

2.1 Comparison of RNA-seq quality control tools. 27

3.1 Comparison of fusion detection tools on simulated data. 44
3.2 Mean strand specificity proportion detected by InFusion 45
3.3 Fusion events detected in public RNA-seq datasets. 46
3.4 RNA-Seq sample details. 47
3.5 Fusion types detected in VCap and LNCaP cell lines. 48
3.6 Fusions validated using qPCR from VCap and LNCaP cell lines. . 49