# Molecular and Computational Transcriptomics in Prostate Cancer

**Youri Hoogstrate**

# Colofon

ISBN: 978-94-6375-216-9

Youri Hoogstrate
e-mail: y.hoogstrate@gmail.com

The work described in thesis was conducted at the department of Urology, department of Pathology and the former department of Bioinformatics of the Erasmus Medical Center, Rotterdam, The Netherlands. The work was financially supported by:

- CTMM TraIT [05T-401]
- CTMM NGS-ProToCol [03O-40]

Printing and binding by Ridderprint BV, Ridderkerk.



The printing of this work was financially supported by the Stichting Wetenschappenlijk Onderzoek Prostaatkanker (SWOP) and Stichting Urologisch Wetenschappelijk Onderzoek (SUWO).

# MOLECULAR AND COMPUTATIONAL TRANSCRIPTOMICS IN PROSTATE CANCER

MOLECULAIRE EN COMPUTATIONELE TRANSCRIPTOMICS IN PROSTAATKANKER

**P R O E F S C H R I F T**

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof. dr. R.C.M.E. Engels
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
dinsdag 8 januari 2019 om 13:30 uur.

door

**Youri Hoogstrate**
geboren te Goes

Erasmus University Rotterdam

# PROMOTIECOMMISSIE

| | |
|---|---|
| **Promotoren:** | Prof. dr. ir. G.W. Jenster<br>Prof. dr. P.J. van der Spek |
| **Overige leden:** | Prof. dr. J.W.M. Martens<br>Prof. dr. P.A.C. 't Hoen<br>Prof. dr. E.C. Zwarthoff |
| **Copromotoren:** | Dr. A.P. Stubbs<br>Dr. E.S. Martens-Uzunova |

# Contents

# 1 Introduction

## 1.1 General introduction

In today's life on Earth, DNA is almost exclusively the carrier of genetic information although it is believed that RNA molecules were the first carriers of genetic information [1]. In healthy cells, RNA molecules and proteins are produced in amounts that are in balance with the demand. Due to mutations in the DNA, the transcriptome and proteome may change and consequently, cells may not be able to properly maintain themselves and die or survive in a mutated state and behave differently. Cancer finds its origin in changes in the DNA, with mutations in, and dysregulation of RNA as a consequence [2, 3]. Understanding the consequences at the RNA level may be helpful in the understanding of how cancer progresses, which eventually may be useful for precision medicine. Mutations can be detected in DNA and RNA using Next Generation Sequencing (NGS) technologies. NGS DNA analysis can reveal almost all mutations, while analysing RNA gives more information about the type and state of cells.

In the past few centuries, science in biology evolved from observations by eye, to the microscope and to the molecular and digital world. The invention of the microscope by Antoni van Leeuwenhoek has been of great importance in this process as it allowed to study living organisms at a new resolution: the cell became visible. Nowadays it is possible to look at atomic resolution with electron or atomic force microscopes, NMR and X-Ray crystallography. What is remarkable about these techniques is that they do not measure actual light. For example, the electron microscope measures matter waves, which are converted to grey values and are at their turn projected as a photograph. Such conversions make analysis dependent on computer models and require analysis software for data processing. Similarly, for the analysis of RNA and DNA sequencing data a revolution has taken place and currently the vast majority is analysed with computer models. From this perspective the computer can be seen as the modern microscope for DNA and RNA analysis. However, as the sequencing

techniques continue to be improved and more knowledge is gained, new and adapted computer models are needed.

## 1.2 DNA- and RNA sequencing

The discoveries of Franklin, Wilkins, Watson and Crick in the early 1950s led to the discovery of the DNA helix structure [4]. This was the starting point for unraveling the hereditary characteristics encoded in DNA. However, it took until the beginning of the 1970s before the first order of a DNA sequence was determined [5]. Due to the rapid development in technology, the (almost) complete human reference genome has been unraveled and made available in 2001 [6, 7], which is accessible to everybody with a desktop pc and an internet connection today. Nevertheless, due to its static nature, cellular DNA alone is insufficient to explain life, the behaviour of cells and diseases. One of the complex challenges in understanding an organism's DNA and consequently in understanding genetic diseases is to identify all components encoded within a genome [8]. There are multiple open access databases with annotations of genes and transcribed loci, such as UCSC [9] and RefSeq [10]. Using such databases, elements imprinted in a genome and higher order interactions can be studied further.

It was believed that beyond protein coding genes most genomic regions are junk DNA and that most transcripts that do not code for protein sequences are non-functional. Although there is still debate about which RNAs are functional [11], the hypothesis that most transcribed regions of the genome have no function has been revised since RNAs that do not code for protein (ncRNAs) turn out to be highly abundant and involved in a variety of functions in the cell [12].

### 1.2.1 RNA

There are different types of RNA molecules, subdivided based on structure or function (Figure 1.1). In the nucleus, pre-mRNA is transcribed from the DNA whereby the pre-mRNA is immediately capped at the 5′-end. After transcription, the transcript is elongated with a poly-A tail, resulting in mature mRNA. In a process named splicing, certain regions named introns are excised from the pre-mRNA. Splicing can take place both during- (co-transcritional splicing) and after transcription (post-transcriptional splicing), but mainly co-transcritional. The mRNA is transported from the nucleus to the cytoplasm. During the translation process mRNA functions as blueprint for the synthesis of proteins. Beyond mRNA, the transcription process is also responsible for non-coding RNAs (ncRNAs), subdivided in small ($< 200$ nt) and large ncRNAs (lncRNAs; $> 200$ nt) [13, 14].

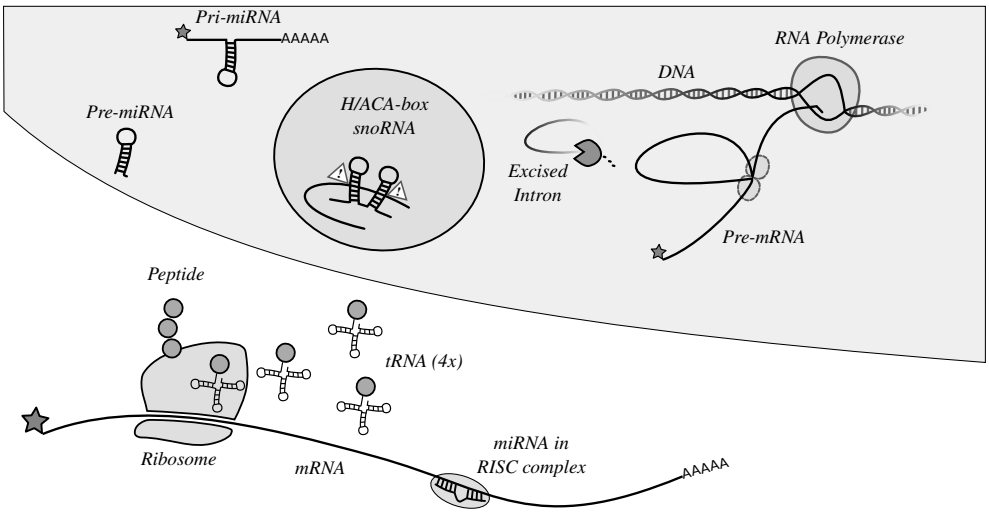



Figure 1.1: Schematic overview of RNA processing. RNA is transcribed in the nucleus. Polymerase first caps the 5′-end and continues transcribing RNA from the DNA. During transcription, introns are excised in a process named splicing. Some introns are further processed into small ncRNAs such as snoRNAs. SnoRNAs are transported into the nucleolus, where it guides rRNA modifications. After polymerase has completed, poly-A polymerase adds a poly-A tail to the transcript. Mature polyadenylated transcripts can be host to both non-coding RNAs (such as pre-miRNAs) and coding mRNA. Mature mRNAs are exported to the cytoplasm. Ribosomes translate peptides from mRNAs, where tRNAs deliver the aminoacids. Mature miRNAs are also exported to the cytoplasm and together with complementary mRNA and specific proteins, the miRNA forms a RISC complex that prevents translation.

The ribosome is a combined protein RNA complex that synthesises proteins. In human cells, the ribosome is composed of two subunits. The large subunit contains 3 RNAs (28S, 5S and 5.8S) whereas the small subunit contains 1 (18S) [15]. About 90% of the total RNA in human cells is rRNA [11].

In the nucleus, pri-miRNAs function as genes for miRNAs and are further processed by `Drosha` into ∼60 nt long hairpin shaped pre-miRNAs. After the pre-miRNA is exported to the cytoplasm, it is processed by `Dicer` into ∼22 nt long microRNAs (miRNAs). Together with specific proteins, miRNAs form the RISC complex that allows binding of a miRNA to a complementary target RNA molecule. This complex may prevent translation of the target RNA or induce its degradation. As a result, miRNAs typically reduce gene activity and function as negative regulators of gene expression. Given their direct influence on translation and their involvement in feedback mechanisms, it is not surprising that certain dysregulated miRNAs are involved in cancer [16, 17, 18].

Small nucleolar RNAs (snoRNAs) are non-coding RNAs that are located in the nucleolus and Cajal bodies [19] and are involved in posttranscriptional modification

of rRNA [20]. H/ACA-box snoRNAs consist of a double hairpin structure and guide pseudouridylation of rRNA, while C/D-box snoRNAs guide 2‘-O-methylation. Small Cajal body-specific RNAs (scaRNAs) are hybrids and guide both types of post-processing. SnoRNAs have been implicated with other roles such as gene silencing and alternative splicing, while dysregulation has been associated with cancer [21].

Transfer RNAs (tRNAs) consist of three hairpin loops folded in a cloverleaf structure. They are involved in protein synthesis by carrying amino acids. The second loop of a tRNA contains a three letter subsequence named the *anticodon*, that due to its complementary binding determines which amino acid gets translated [22].

Recently, small RNAs (<35 nt) derived from many types of ncRNAs, including rRNA [23], have been found in various organisms [24]. Initially, it was believed that these were products of RNA degradation and turnover, but their high abundance and consistent start and end positions contradicts this. The roles of many of these RNAs are not fully understood. Different RNAs derived from tRNAs (tRFs) have been reported, which are classified into two major groups [25]. The tRNA halves are products of tRNA endonucleolytic cleavage near the anticodon resulting in fragments of 30-35 nt. The second group consists of smaller fragments (∼20 nt) which often span one single hairpin of the tRNA. Expression levels of $3'$ tRFs show no correlation with the copy number levels of tRNA isoacceptors [26]. Also, small ncRNAs derived from snoRNAs have been found and roles have been ascribed, including miRNA-like activity and involvement in certain diseases. However, their processing mechanisms and putative function are not fully understood.

Although RNA molecules are typically linear and single stranded, recent studies have reported circular RNAs (circRNAs) [27]. As all nucleotides of a circRNA are covelently bound, it forms a loop with itself. They are more stable than linear RNA because of this structure [28]. The role of circular transcripts is not fully understood. Some circRNAs are found to work as miRNA sponge, while others others are protein coding [29]. Besides circular RNAs, also double stranded RNA have been reported [30].

## 1.3 Fusion Genes

In healthy cells, DNA contains information that ensures a balance in regulation and molecular organisation. When DNA gets damaged this balance may be disrupted, which may lead to changes in cell proliferation and apoptosis. DNA damage can be single base substitutions, insertions, deletions, amplifications, but also more complex rearrangements such as translocations. DNA rearrangements may result in juxtaposition of genes. Such fusion genes are repeatedly found in cancer [31]. The consequence of a fusion gene may be disruption of one or multiple genes, for example by introducing

a frameshift in fused coding sequences. Fused coding sequences can also be in-frame, resulting in chimeric proteins [32, 33]. Rearrangements involving regulatory elements, like gene enhancers or promotors, can alter expression levels and cause changes in regulation. If a DNA rearrangement does not change the cells functioning, it is called a *passenger mutations*. Such mutations are called *passenger mutations* and are expected to be the majority of mutations found in cancer cells. On the contrary, if a mutation does contribute to cancer progression, it is a *driver mutation*.

Oncogenes are genes of which increased or adapted activity results in cancer progression. Activating missense mutations are often driver mutations. For example, in receptor proteins a missense mutations may change the specificity for a ligand or the mutation results in a protein of which the signaling is completely independently of the ligand. Often in fusions involving an oncogene, a regulatory element of a highly active gene becomes adjacent to the oncogene, resulting in increased activity of the oncogene. For such fusion genes it is common that the oncogene stays intact in order to keep fulfilling its function and that the regulatory element does not lose its potential to induce transcription. This means that the DNA breakpoint is a major determinant for the oncogenic potential of the fusion. A well known example is the recurrent fusion gene in prostate cancer `TMPRSS2-ERG`, in which `TMPRSS2` donates its promotor to oncogene `ERG` [34].

Tumor suppressor genes often carry out functions like DNA damage repair, growth control, cell cycle arrest and apoptosis. Consequently, reduced activity or loss of function of such genes is beneficial for progression of cancer. A fusion that involves a tumor suppressor gene may contribute to cancer progression by reducing transcription or by introducing a frameshift resulting in mutated or truncated proteins. It is likely that some of these fusion genes are harder to detect at the RNA level because of reduced expression. Hence, fusion genes can contribute to cancer progression via various mechanisms [31, 35].

One of the hallmarks of cancer, *sustaining proliferative signaling*, is the requirement that cells can adjust their signaling to make them self-determinant with respect to proliferation [36]. This typically involves changes in regulation of growth-factors, receptors and corresponding pathways, which are often tissue specific. Certain fusion genes are more common than others, and can be specific for a certain cancer type. If a specific type of cancer is characterised by a recurrent fusion gene, a test for the fusion gene may be used as indicator for the type of cancer. There are several known recurrent fusion genes used as biomarker, such as `BCR-ABL` [37], `EML4-ALK` [38], `PML-RAR` [39], and `TMPRSS2-ERG` [40].

The `BCR-ABL` fusion gene is recurrently found in chronic myelogenous leukemia (CML) and typically results in transcripts that correspond to thee proteins (p190,

p210 and p230) [41]. Using probes designed against the fusion transcripts, the presence of the fusion gene is tested with RT-PCR [42] and can function as clinical biomarker for CML [43]. `PML-RAR` is an interchromosomal, reciprocal translocation between genes `PML` and `RAR` [44], which causes acute promyelocytic leukemia. Screening for this fusion gene is also performed with RT-PCR [39].

Although several fusion genes are used as biomarker for cancer, it would be ideal use fusion genes as target for specific drugs for targeted treatment. There is ongoing research to develop drugs targeting fusion-genes [45, 46]. For example, imatinib has been approved as drug for Philadelphia-chromosome-positive chronic myeloid leukemia, targeting `BCR-ABL` [47]. Thus, identifying and understanding fusion genes is of high importance for diagnosis as well as therapy of cancer. Fusion genes can not only be detected with DNA-seq but also with RNA-seq, of which the latter may also provide expression levels, splice variants and which gene is donor or acceptor. Unfortunately, the RNA-seq fusion gene detection tools are not highly accurate [48] and therefore more research in this field is needed.

## 1.4 Developments in technology and RNA analysis

Analysing RNA has been helpful in understanding the processes that are taking place in a cell. Traditionally, expression of specific RNA transcripts has been studied by northern blot, qPCR and microarrays. The northern blot is used to quantitatively detect a specific RNA sequence, whereas qPCR is used to quantitatively detect a specific cDNA sequence. The microarray is a chip containing large numbers of predetermined probes (short sequences that are complementary to transcript sequences) and made it possible to analyse expression levels in a high-throughput manner. Hence, due to this scale enlargement, it became possible to look at gene expression of all genes at the same time, without doing wet-lab experiments for each gene separately. Because of the large number of datapoints, further statistical analysis needs to be performed with computer frameworks accordingly. Because the probes are predefined, transcripts for which no probes have been included on the chip are not interrogated.

### 1.4.1 Next Generation Sequencing (NGS)

Nowadays, it is possible to measure vast amounts of DNA sequences simultaneously, with relative high speed [51]. By making a cDNA copy of RNA using reverse transcription, also RNA sequences can be measured with this technology (RNA-seq). A

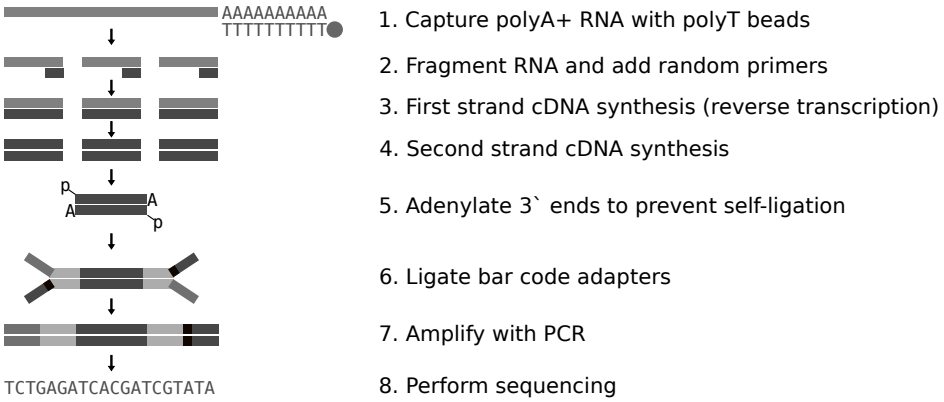



Figure 1.2: Overview of the Illumina TruSeq protocol. mRNA is captured using magnetic beads with poly-T sequences that are complementary to the poly-A tails. RNA is then fragmented and random primers are added to allow cDNA synthesis. The first cDNA strand is synthesised with reverse transcriptase, the second with DNA polymerase. After both cDNA strands have been synthesised, the 3′-ends are adenylated and 5′-ends repaired. Adapter sequences are ligated to the cDNA to allow them to hybridise onto a flow cell. In the sequencer, cDNA molecules are iteratively extended with fluorescent nucleotides and during each iteration these nucleotides are excited and emission at corresponding wavelengths is measured in an imaging step [49, 50]. The corresponding imaging data is transformed into nucleotide sequences.

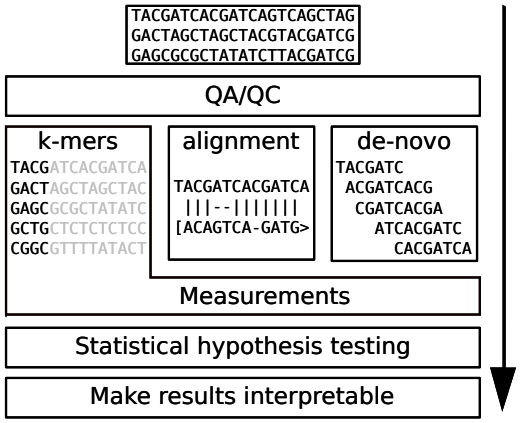



Figure 1.3: Schematic overview of a typical NGS workflow. After sequencing data has been obtained, the quality is assessed and controlled (QA/QC) where possible. During QA/QC it is common to trim low quality bases, remove adapter sequences and discard reads with low complexity sequences. The high quality data is often mapped to an organism's reference genome, into a file refered to as alignment. For certain organisms, the reference genome may be missing, incomplete or the focus could be on complex variants. In these scenarios a reference can be reconstructed using *de-novo* assembly. There are also techniques available that skip both alignment and assembly, but measure only the k-mer (small subsequences) content directly from the reads. Most features, such as expression, polymorphisms and fusion genes are determined from alignments. These features are used to associate phenotypes with differences. Results are presented in tables or figures that facilitate interpretation.

common way to prepare mRNA for sequencing is the Illumina TruSeq protocol[1] (Figure 1.2). The resulting detected sequences in NGS are called *reads*. In the early days of RNA-seq analysis, the preparation protocols produced non-strand-specific reads. The sequence of these reads can correspond to the cDNA or its reverse complement, and therefore loses the information whether the transcript was sense or anti-sense. Nowadays, almost exclusively protocols are used in which information of the strand is preserved (strand-specific).

The first step in computational RNA-seq analysis is usually estimating the quality of the dataset. Artifacts such as adapter sequences and bases of low quality are removed from reads, or reads are removed from the dataset entirely. Although the dataset will become smaller, the overall quality will be higher. From this data, certain features will be extracted (Figure 1.3), such as gene expression levels, splice isoforms, fusion genes or polymorphisms. After feature extraction, associations between these features and certain conditions are examined. In most cases, the reads are first mapped to a reference genome and features are extracted from this alignment. It is possible that a reference genome of an organism does not meet the requirements or it is not available. Under these circumstances it might be worthwile to first use the data to build a reference transcriptome using *de novo assembly*. But using an alignment is not always necessary. For example, expression levels can also be determined from the k-mer content. However, when the reads have been aligned, expression levels are estimated by counting the number of reads aligned to a given locus [52]. Further analysis requires distinct statistical models [53, 54, 55]. Due to technical limitations, the maximum length of a sequence that can be measured by the sequencing machine is limited (50-300 consequent nucleotides per molecule), while the smallest human chromosome has a length of 46,709,983 bp and the longest transcripts exceed 100,000 nt [56, 57]. Sequencing fragments rather than entire transcripts complicates determination of splice isoforms because exon junction spanning sequences are often missed. There are, however, computer programs that take this into account and based on mathematical models make predictions on splice isoforms and corresponding quantities [58]. In the past few years, progress regarding the length issue has been made and techniques have become available that can measure up to 40,000 consecutive nucleotides [59] and even 900,000 [60].

In *paired-end sequencing* a fragment is sequenced from both ends and the link between both sequenced ends is preserved. Combining sequence information from both ends of the read pair with the expected fragment length, can be used to improve alignment to a reference genome. Moreover, if an alignment indicates that the distance

[1] https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqrna/truseq-rna-sample-prep-v2-guide-15026495-f.pdf

between the paired-end reads in the alignment does not fit the range of expected fragment size, this may be evidence for structural variants or splicing. For instance, if the distance between two reads of a pair in a genomic alignment is large (e.g. 10 Mb), while the estimated DNA fragment size was no longer than 1 Kb, it is plausible the reads span a deletion of ∼10 Mb. Reconstruction of transcripts using NGS data relies heavily on computational analysis and the corresponding software largly determines accuracy and processing time.

Besides expression analysis, RNA-seq can be used to look at RNA from various new angles because sequences are measured independent of pre-defined probes. It allows discovery of novel transcripts, gene structures, splice isoforms, circular RNAs, mutations, RNA-editing and fusion genes. These analyses are typically performed using different software modules connected together into computational *pipelines*. A schematic overview of the structure of a typical RNA-seq experiment is presented in Figure 1.3. Each module has usually many different parameters [61] and modules can often be replaced with comparable software modules. Choosing the modules as well as the corresponding parameter settings is an important requirement of a computational environment because this allows adjustments for specific use-cases.

### 1.4.2 Small RNA-seq

Illumina's TruSeq mRNA preparation protocol specifically selects polyadenylated mRNA, but because miRNAs are not polyadenylated, they require a different preparation protocol [62]. After total RNA is extracted, RNA molecules are size seleceted in a range of ∼18-30 nt [63]. Since the molecules are selected to be smaller than the maximum read size, there is no need for further fragmentation. PCR- and sequencing adapters are ligated and the small RNAs are sequenced in a single-end and strand-specific manner. Small RNA-seq can be used to investigate and quantify annotated miRNAs and to detect novel miRNAs [64]. Although miRNAs are annotated as single sequences with exact genomic start and end positions, sequencing data shows that there is variation in these start and end positions [65]. These variants, called *isomiRs*, arise because cleavage is not precise up to the nucleotide. Apart from cleavage, it is also possible that nucleotides are ligated or substituted. This variation complicates determination of miRNAs since the boundaries of a miRNA need to be determined using reads derived from different isomiRs. In small RNA-seq data, expression profiles can be determined and be used to study differential expression. Although small RNA-seq was designed to study miRNAs in particular, this method is applicable to many types of small RNA [24], including PIWI-interacting RNAs (piRNAs) [66] as well as a variety of small non-coding RNAs derived from rRNAs, tRNAs, snoRNAs/scaRNAs

and snRNAs [24]. For miRNAs, their well understood 2D hairpin structure plays a major role in prediction and annotation [67]. Because these models do not apply to the other small non-coding RNAs, annotations of small RNAs other than miRNAs and piRNAs are lacking. The lack of such annotations complicates analysis and therefore a method accurately annotating small RNAs in small RNA-seq data is needed.

### 1.4.3 Fusion genes and RNA-seq

DNA rearrangements may result in fusion genes and if they are expressed, they may result in fusion transcripts. Therefore, fusion genes can not only be observed and detected in DNA-seq but also at the RNA level [68], using RNA-seq. In DNA-seq, the distribution of reads across the genome is almost uniform and expression independent. This facilitates thorough detection and minimises the chance of missing fusion events. In contrast, fusion gene loci must be expressed in order to be observable in RNA-seq. However, using RNA-seq for fusion gene detection can provide several advantages over DNA-seq. It allows to deconvolve the fused splice isoforms and reveal the (fusion) gene structure, including potential novel exons. Also, the direction of transcription and altered expression levels are detectable. Such information helps to understand the role and impact of a fusion gene. There are also fusion transcripts that do not find their origin in genomic rearrangements but only exist at RNA level, including readthrough events and trans-splice isoforms [69], which can not be detected with DNA-seq.

Genomic breakpoints of fusion genes are unique per individual event, while at mRNA level they typically result in a limited number of fusion transcripts. Results found in RNA-seq can therefore more easily be used to develop PCR assays for screening. For example, `BCR-ABL` fusions are typically responsible for transcripts resulting in three distinct proteins. When only the first susceptible intron of both genes is taken into consideration (`BCR` intron 1: 70 Kb and `ABL` intron 1: 120 Kb), there are theoretically ∼8,4 billion possible DNA translocations that would result in the same fusion transcripts.

Although there are many software packages available to detect fusion genes in RNA-seq data [68, 70, 71, 72, 73, 74, 38], there is not a tool that is superior in performance [75]. More strikingly, the overlap of results generated with different tools is limited, which requires the use of multiple tools to find the full repertoire of fusion genes within a sample [76, 77, 75, 78]. A logical next step in RNA-seq fusion gene analysis is to prioritise the identified fusion events, for example by applying filters [48] or by predicting functional impact. The variety in available software is also an advantage. Different tools are developed to solve specific problems and have unique qualities

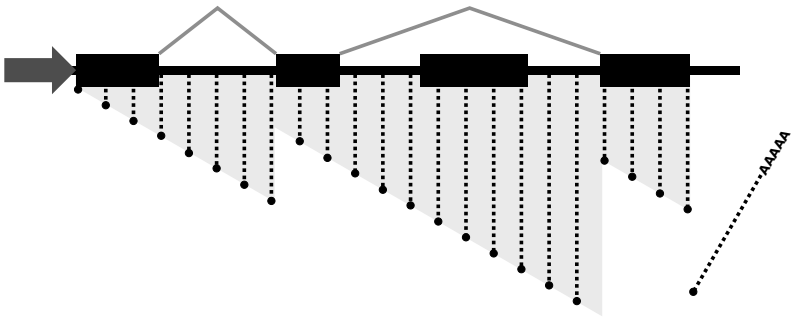



Figure 1.4: At the beginning of the transcription process, an RNA molecule is capped at the 5′-end. RNA polymerase makes the RNA molecule longer and longer. During co-transcriptional splicing, a spliceosome can be formed and the introns (and skipped exons) will be spliced out only after the entire intron and corresponding splice recognition site have been transcribed. This results in truncation of RNA molecules, illustrated with the shorter snapshots at the beginning of exons 2 and 4. Polymerase will continue and the pre-mRNA is further extended until the next spliceosome can be formed. Eventually, only exon sequences remain. The RNA is cut at the poly-A signal and a poly-A tail is attached, resulting in mature mRNA.

that distinguish them from others, which might be a reason why the overlap between tools is limited. For example, JAFFA is designed to cope with relatively long read lengths. SOAPfuse is implemented to return results quickly. INTEGRATE minimises the number of false positives [79], whereas FusionCatcher focuses on data-cleaning by removing viral, bacterial and ribosomal contamination sequences. It has been recommended to choose a detection tool based on its properties in relation to the data, rather than on benchmarking based on artificial data [78]. However, a combination of tools with an appropriate way of aggregating data has the potential to improve the overall results, for example by a majority vote [75, 80, 81]. A complicating factor is that each tool reports its results in its own way. There is no generic file format for describing fusion genes, comparable to what the VCF format means for single nucleotide polymorphisms. Besides a standardised format, there is a clear need to have software that integrates the results of each of these tools and report overlap.

### Fusion genes and pre-mRNA

The majority of RNA-seq fusion gene detection tools focus on detection at the mRNA level. During mRNA processing, introns are spliced out of pre-mRNA. Splicing takes mostly place co-transcriptionally, but can also take place post-transcriptionally [82]. During transcription, pre-mRNA is synthesised from the 5′-end to the 3′-end and protected by a 5′-end cap (Figure 1.4). In co-transcriptional splicing, as soon as the end of an intron is transcribed, a spliceosome will be formed and introns are spliced

out, even before the entire pre-mRNA is transcribed and before the poly-A tail is attached [83]. Because splicing has finished before the poly-A tail has been attached, the poly-adenylated transcripts are free of introns. However, as result of intron-retention and post-transcriptional splicing, it is possible that poly-adenylated transcripts contain introns. In poly-A+ RNA-seq protocols, RNA is extracted based on the poly-A tails and predominantly mature mRNA is sequenced. This means that the data contains only few reads that correspond to introns (intronic reads) and that it specifically targets mature polyadenylated mRNAs. There are multiple pitfalls with respect to fusion gene analysis of such data:

- There are various transcripts that are non-polyadenylated such as circular RNAs and certain ncRNAs. These types of RNA are excluded from the sequencing library.
- Poly-A selection introduces a so called $3'$ coverage bias. Due to this bias there is a significant higher coverage of sequencing reads near the $3'$-end compared to the $5'$-end [84, 85]. For fusion gene detection, the consequence is that breakpoints closer to the $5'$-end are underrepresented with sequencing data and thus harder to predict.
- Genomic breakpoints that fall within intron sequences cannot be detected. Because introns are relatively long, the vast majority of the fusion gene breakpoints are located within introns. Because intronic reads are mostly lacking, corresponding genomic breakpoints cannot be detected.

It is possible to circumvent these pitfalls by using an RNA-seq protocol that is not purifying the library targeting the poly-A tails. This can be done by sequencing total RNA, of which the high abundant rRNA is depleted first. Then random hexamer primers are used as complementary primers to synthesise cDNA [86]. The currently available RNA-seq fusion gene detection tools are not fully designed to cope with ribo-depleted total RNA, as intronic reads are often discarded [38]. There are two exceptions: *gfuse* is particularly designed to analyse FFPE material [87], but the actual software is not publicly available, and *Tophat-Fusion*, which is able to find intronic breaks but is restricted to gene regions [70].

### 1.4.4 Standards in bioinformatics software and genetic data

Software in the field of bioinformatics varies so heavily in quality that scientific reports have been written in which the best-practices of software engineering, development and distribution are highlighted [88, 89]. These include basic rules such as *provide a help option* and *provide warnings on missing input*, but also more sophisticated

development disciplines such as parameter validation, test-driven development and continuous integration. There is debate on open access of source code for at least the peer-review process of software-based publications [90]. There are multiple research projects that investigate and resolve shortcomings in bioinformatics software, for example by defining [91, 92] or implementing [93] standards or providing corresponding code libraries [94, 91, 95, 96]. Such libraries are robust building blocks that make modularity possible, prevent re-inventing the wheel and enforce to respect standards by considering deviant, invalid, files as incompatible. Hence, besides the quality of an algorithm, it is important to write software in a manner that will help other scientists using and improving it.

Most software packges in the field of bioinformatics are written as command line utilities and lack a graphical user interface (GUI), which makes using them rather complicated. There are commercial software packages available that contain all modules needed for a pipeline and are "plug and play" after installation. *CLC-bio* (CLC-bio, Aarhus, Denmark) and *Partek* (Partek Inc., Chesterfield, USA) are examples of such commercial packages and provide a user-friendly graphical interface and a comprehensive list of corresponding modules. They are in particular designed for microarray and sequencing analysis, but for instance *CLC-bio* also has the option to analyse RNA 2D structures. These workbenches have strict and often expensive commercial licenses, provide closed source software and often lack in-depth descriptions of their computational methods. This is a complicating factor for the reproducibility of an experiment, that for instance prevents rerunning analyses by peer reviewers unless they have a commercial license for the same software. It prevents computer scientists to build further on these methodologies as the source code is explicitly kept secret. From the end-user perspective it would be convenient that free (as in freedom) command-line software also becomes available with an easy to use GUI and integrates evenly well as the common commercial workbenches. Although several of these workbenches have been set up [97, 98, 99, 100], like the UEA small RNA Workbench [101], they are not as complete and extensive as commercial ones. Often, incorporated tools are implemented in a non-modular way and the workbenches lack plug-in systems, which requires much more maintenance of the framework in order to keep it up-to-date and complicate adding new tools. In *Galaxy*, many free command-line bioinformatics tools are given a graphical user interfaces [102]. The project tries to make as many tools as possible available via an application store and makes the tools accessible via a web interface [95]. In addition, bioinformatics-specific data formats such as VCF, BAM and BED are integrated within the system. Due to the numerous tools available and the broad scale of applications they comprise, a *Galaxy* server is typically initiated as a bare bone system with a small number of operational tools included, rather than an

over featured system. Consequently, a *Galaxy* system needs to be dressed-up to fit its needs, by installing tools and/or workflows and including corresponding data libraries. There are various pre-selected *Galaxy* workbenches available, that have a predefined set of related tools incorporated from scratch. There are for example pre-configured *Galaxy* workbenches specific for proteomics, metagenomics, imaging and epigenetics.

Modularity is necessary to integrate tools into pipelines and into such workbenches, which requires mutual agreements on the information flow. For example, agreements are required on file formats, transmission protocols, indexing and metadata. In addition, it is important that such agreements are robust enough to be used for future problems. For example, reference genomes are currently stored as FASTA files. Although FASTA files are widely used as consensus for different research applications, they have barely space to describe genomic variation. A FASTA file stores all chromosomes as linear sequences and storing nucleotides only in sequential order is not designed to describe genomic variability. Small variants are then typically stored in separate VCF files, which uses a coordinate system to link the small variants to the FASTA file. In more recent versions of the reference genome, common large variations were added as separate entries isolated from the remainder of their chromosomes. In the very last version of the reference genome, hg38, there are more than 250 alternative loci[2]. It is a logical next step to work out a system to store a reference genome in a way that also allows storage of genomic variation. The first steps into this direction were implemented in *de-novo* assembly algorithms using *de Bruijn Graph* structures, in which variations are represented by branches or nodes in a graph data structure [103]. This, from an evolution point of view more natural representation, seems to be evident for storing and describing genomes. Recently, it was made public that Global Alliance for Genomics and Health (GA4GH) is working on the fundamentals of such reference genomes by outlining standards on corresponding data formats [104]. The VG project[3] (variation graphs) is making progress in using graph data structures as reference genome.

Directly related is the way genetic data itself is accessible. Genome, gene and protein annotations have been available for many years and are still instantly updated. These types of data are typically privacy in-sensitive as they apply to the human population in general. Data of a more personal nature are found in SNP and fusion gene databases such as COSMIC [105], of which the presence of a genetic variant is often associated with a phenotype. Increasingly more raw biomolecular data is becoming available, in particular results of NGS. Sharing such information as *open data* is essential to allow experiments to be reproduced and to further elaborate

[2] `https://www.ncbi.nlm.nih.gov/grc/human/data`
[3] `https://github.com/vgteam/vg`

and improve analysis methods. However, genetic data also allows determination of a personal unique genetic footprint, which may potentially lead to privacy infringement if data is published with too much phenotype information. The European Genome-Phenome Archive [106] (EGA) is a repository of genetic data in which data access is controlled by data committees that review formal data requests and determine whether access is justified.

The Cancer Genome Atlas (TCGA) [107] examplifies the importance of publicly available genetic and combined phenotypic data. This resource, containing NGS and phenotype data of 30 types of tumors from many patients, has been citated more than 1600 times since its publication. Using this database, various recurrent dysregulated and mutated genes for several types of cancer have been identified, (sub-)classifications have been proposed [108] and differences and similarities between different types of cancer have been investigated [109]. In addition, the TCGA database is often used to independently validate findings discovered in other datasets.

Unfortunately, the availability of the increasing amount of NGS data also has a downside. Of the bulky and highly redundant data, a large increase of datasets has resulted in a data explosion which makes storage and transfer an increasing challenge [110]. To overcome this, new data formats such as CRAM are needed. These formats greatly reduce the filesize by also compressing relative to a reference genome [111]. In addition, sequencing techniques that increase the read length are investigated [59, 60, 112], which allow sequencing in a less redundant manner.

## 1.5 Prostate cancer

Prostate cancer (PCa) is one of the most common types of cancer in men [113] and after lung cancer, the most frequent cancer-related cause of death in men in western countries [114]. Although there has been significant research in diagnostic and prognostic biomarkers for prostate cancer, there is a lack of biomarkers that are both sensitive and specific [115]. Common genetic changes in PCa are `TP53` mutations [116], `PTEN` loss [117], changes in `AR` signalling by point mutations, indels and deletions [118], and various structural rearrangements involving genes of the `ETS` gene family, including the `TMPRSS2-ERG` fusion [119]. Although such genetic changes can be detected by RNA-seq or DNA-seq, these are typically not tested for after biopsies or radical prostatectomy for the purpose of targeted treatment because there is no significant improved value. In PCa, approximately 50% of the diagnosed patients have the fusion gene `TMPRSS2-ERG` [120, 121], most often formed by an intrachromosomal deletion of ∼3 Mb on chromosome 21 between `TMPRSS2` and `ERG`. `TMPRSS2` is an androgen-regulated gene and particularly highly expressed in prostate epithelial cells. `ERG` is

a proto-oncogene that encodes a transcription factor which regulates hematopoiesis. Due to the recombination, the promotor of TMPRSS2 causes androgen-regulated elevated expression of ERG [34]. The DNA breaks are most often found in the first two introns of TMPRSS2 and the third and fourth intron of ERG. At a transcript level, the most frequent exon-to-exon boundaries are TMPRSS2-exon 1/ERG-exon 4 and TMPRSS2-exon 1/ERG-exon 5 [122, 123]. The fusion transcripts either encodes an in-frame fusion protein that contains 4 amino acids from exon-2 of TMPRSS2 fused to ERG, or short ERG proteins starting at alternative start codons in ERG exons 3, 4 or 5 [123]. Although this fusion gene is a highly specific marker for PCa [40], it is due to its occurrence rate of 50% not highly sensitive. Developing drugs targeting TMPRSS2-ERG turns out to be complicated as ERG is a transcription factor with many homologues family members. Nevertheless, progress in this direction is being made [124].

## 1.6 Scope of the thesis

Human cells contain different types of RNA, which vary in structure, function and quantities. Different mechanisms such as mutations, RNA-editing, fusion genes and changes in expression can affect the transcriptome of a cell. Alterations of the transcriptiome by these mechanisms have the potential to contribute to cancer progression. This diversity makes RNA more complex, but also more challenging to analyse than DNA. RNA-seq allows to investigate RNA at a sequence level but requires computer programs for analysis. Therefore, RNA seems ideal for biomarker research by means of computational analysis. PCa has an unmet need of diagnostic and prognostic biomarkers, and is therefore relevant as model system for the development of new analysis methods. Computational analysis methods need to meet certain conventions for compatibility, and need to be publicly accessible so that they can be used by other researchers. Therefore, the overall purpose of this thesis is to facilitate finding new PCa markers in RNA-seq data by using (new) software, which ultimately is integrated in a software toolbox specific for RNA analysis.

### 1.6.1 Small RNA-seq

Small RNA-seq data does not only consist of reads derived from miRNAs but also from other small ncRNAs such as fragments of tRNAs, rRNAs and snoRNAs. Software scanning for miRNAs throughout a reference genome as well as corresponding miRNA annotations are publicly available. For most other types of small ncRNAs there is no such software, which complicates high-troughput analysis of such molecules because of lacking annotations. In **chapter 2**, we propose *FlaiMapper*, a method that addresses this issue.

*FlaiMapper* allows to annotate the small ncRNAs content in small RNA-seq data. To assess if biologcally relevant molecules can be found, we are interested from which precursors the small ncRNAs originate and how abundant they are compared to miRNAs. Small ncRNAs derived from snoRNAs had been reported to be upregulated in PCa. A better understanding of how they are processed might help explain their role in cancer. Also, if it becomes clear how their expression relates to the expression of their host, and from which locations they are derived, this might reveal which processing mechanisms are involved. Using the proposed method, many new molecules may be discovered, of which each has the potential to be correlated with cancer. In **chapter 3**, the scientific relevance of *FlaiMapper* is demonstrated by investigating the small ncRNA content of prostate and prostate cancer, focusing in particular on those derived from snoRNAs.

### 1.6.2 Fusion genes

The current generation of tools that allow detection of fusion-genes in RNA-seq are limited in either sensitivity or specificity [76]. Therefore, indicating whether multiple tools have detected a fusion may increase confidence. In order to find fusion genes across samples, for example to find those that are recurrent in a certain disease, it may also be useful to have a report of duplicate fusion gene entries. Therefore, a method capable of making integrative reports of the outcome of such fusion gene detection tools is needed. In **chapter 4**, the issue of integrating such results is addressed.

RNA-seq experiments that focus on fusion gene detection often use of poly-A+ RNA-seq data, in which intronic reads are rare and can consequently not be used to detect genomic breakpoints located in introns. By making use of random hexamer primers in the RT-step in ribo-depleted total RNA, pre-mRNA can be sequenced and corresponding intronic reads provide additional information for fusion gene detection. Currently available tools are not designed for this purpose and can therefore not distinguish genomic breakpoints from exon-to-exon splice junctions and are not designed to investiage intergenic regions [38]. The challenges of fusion gene discovery in ribo-depleted total RNA-seq data taking this into account, are further addressed in **chapter 5**.

### 1.6.3 The toolbox

To let software be a valuable contribution to the scientific community, it is important that it, apart from a scientifically valuable methodology, complies with recently established principles that promote use and reuse [125]. Apart from these principles, the ease of use is important, which in bioinformatics software is often limited. Therefore, a more generic goal is to ensure that the software presented in this thesis follows recently proposed guidelines [89]. In **chapter 6**, we show the added value of integration of RNA related software, including tools used in chapters 2, 4 and 5.

# 2 *FlaiMapper*: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data

pmid: 25338717, doi: 10.1093/bioinformatics/btu696

Youri Hoogstrate, Guido Jenster, and Elena S. Martens-Uzunova

*Bioinformatics*, 31(5):665−673, 2015

[1]*Department of Urology, Erasmus University Medical Center, Be 362a, PO Box 2040, 3000 CA Rotterdam, The Netherlands.*

Supplementary material: https://academic.oup.com/bioinformatics/article/31/5/665/2748143#supplementary-data

## Abstract

**Motivation:** Recent discoveries show that most types of small non-coding RNAs (sncRNAs) such as miRNAs, snoRNAs and tRNAs get further processed into putatively active smaller RNA species. Their roles, genetic profiles and underlying processing mechanisms are only partially understood. To find their quantities and characteristics, a proper annotation is essential. Here, we present FlaiMapper, a method that extracts and annotates the locations of sncRNA-derived RNAs (sncdRNAs). These sncdRNAs are often detected in sequencing data and observed as fragments of their precursor sncRNA. Using small RNA-seq read alignments, FlaiMapper is able to annotate fragments primarily by peak detection on the start and end position densities followed by filtering and a reconstruction process.

**Results:** To assess performance of FlaiMapper, we used independent publicly available small RNA-seq data. We were able to detect fragments representing putative sncdRNAs from nearly all types of sncRNA, including 97.8% of the annotated miRNAs in miRBase that have supporting reads. Comparison of FlaiMapper-predicted boundaries of miRNAs with miRBase entries demonstrated that 89% of the start and 54% of the end positions are identical. Additional benchmarking showed that FlaiMapper is superior in performance compared with existing software. Further analysis indicated a variety of characteristics in the fragments, including sequence motifs and relations with RNA interacting factors. These characteristics set a good basis for further research on sncdRNAs.

**Availability and implementation:** The platform independent GPL licensed Python 2.7 code is available at:
`https://github.com/yhoogstrate/flaimapper`
Corresponding Linux-specific scripts and annotations can be found in the same repository.

## 2.1 Introduction

Sequencing of small non-coding RNAs (sncRNAs) aiming at the quantification and discovery of microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs) and vault RNAs (vtRNAs) has revealed that most types of sncRNAs get processed into smaller RNAs [126]. Initially, it was suggested that these smaller RNAs are degradation products of the turnover of their precursors. Nevertheless, evidence accumulating over the last years demonstrates that some RNA fragments are functional and have specific maturation mechanisms indicating their importance and novelty [126, 127, 128]. Such fragments find their origin in tRNAs, vtRNAs and snoRNAs and are assumed to have a variety of functions. Most importantly, deregulation and involvement of different types of fragments have been demonstrated in different types of cancer [24]. A description of commonly detected fragments and their precursors is given below:

- A *pre-miRNA* is an approximately 75 nt long RNA molecule produced from its primary precursor transcript (pri-miRNA) by `Drosha` [129]. Pre-miRNAs adopt a hairpin structure recognized by `Dicer` that cleaves the terminal loop to release an approximately 22 nt long double-stranded miRNA duplex. One of the strands (miRNA) is loaded into `AGO` to generate the functional miRISC complex [130, 131]. The remaining strand (miRNA*) is usually degraded. Often both strands are found as fragments in small RNA-seq [132].

- Fragments originating from mature tRNAs are commonly classified into two subgroups [127], tRNA halves and tRNA-derived RNA fragments (tRFs):

  - tRNA halves are most probably produced by angiogenin that cleaves the tRNA near its anticodon, resulting into halve tRNAs (∼35 nt). It is believed that some tRNA halves contribute to translational repression and cell stress response [133, 134].
  - The smaller (∼20 nt) tRFs are derived from the tRNAs 5′- and 3′-end and from the pre-tRNAs 3′-end. It is not completely understood which proteins are involved in the production of tRFs, although evidence for associations with both `Dicer` and RNaseZ are reported [126, 127]. Although the putative functions of the majority of tRFs are unclear, evidence suggests that some are involved in RNA interference, with effects on cell proliferation and gene regulation [126].

- snoRNA are (60-250 nt) small RNAs found in the nucleolus. They comprise the subtypes H/ACA-box, C/D-box and small Cajal body-specific RNAs (scaR-

NAs) [135]. Putative functions such as regulation of alternative splicing, post-transcriptional regulation of gene expression and associations with cancer have been proposed for their fragments [128, 136, 137, 138, 139, 140, 141, 142].

Currently, studies on fragments other than miRNA and miRNA* are restricted to (often visual) interpretation of alignments. Consequently, the data are inspected only at a global ncRNA level. Not making use of the annotation of exact fragment coordinates is a shortcoming, since it restricts analysis at the level of individual fragments. Additional benefit of such analysis is the gained statistical power.

Here, we describe Fragment Location Annotation Mapper (*FlaiMapper*) that predicts the locations of sncRNA fragments in small RNA-seq alignments. Prediction is based on the densities of start and end positions of aligned reads. It is important to state that the goal is not to predict any particular subtype of fragment but to annotate data for subsequent quantitative analysis, by making use of sequencing data only. Therefore, FlaiMapper does not use 2D structure prediction or classification based on heuristics of previous discoveries as often is used for the prediction of pre-miRNAs [132].

## 2.2 Methods

Fragments are measured with small RNA-seq, where the corresponding variable-sized sequences, called reads, are aligned back to a reference sequence. The reference sequence is used to determine the reads origin. This reference can be the genome, the transcriptome or specific regions (e.g. miRNA or tRNA databases). The library used for our analysis was manually composed (Supplementary Data). Pre-processing and alignment for each dataset are further discussed in the Supplementary Data.

Analysis was applied to two different publicly available datasets with SRA accession numbers `SRP002175` [64] and `SRP006788` [129]. Dataset `SRP002175` contains 12 small RNA-seq samples, taken from human pigment cells. The reads are 18-23 nt long and processed on the Illumina's Genome Analyzer II platform. Dataset `SRP006788`, processed on the same platform, contains 18-30 nt long reads, taken from six samples from a HeLa cell line. In this dataset, the samples have undergone the following treatments [129]:

- `SRR207111` Total cellular RNA was extracted from HeLa cells.
- `SRR207112` Total cellular RNA was extracted after `RRP40` core subunit depletion; `RRP40` has $3' \rightarrow 5'$ exonucleolytic activity.


Figure 2.1: 1468 reads aligned to `SNORD74` (black line) in dataset `SRP006788` (total RNA). The contours of the aligned reads (grey) form three separate clusters. This may be an indicator for the presence of multiple fragments.

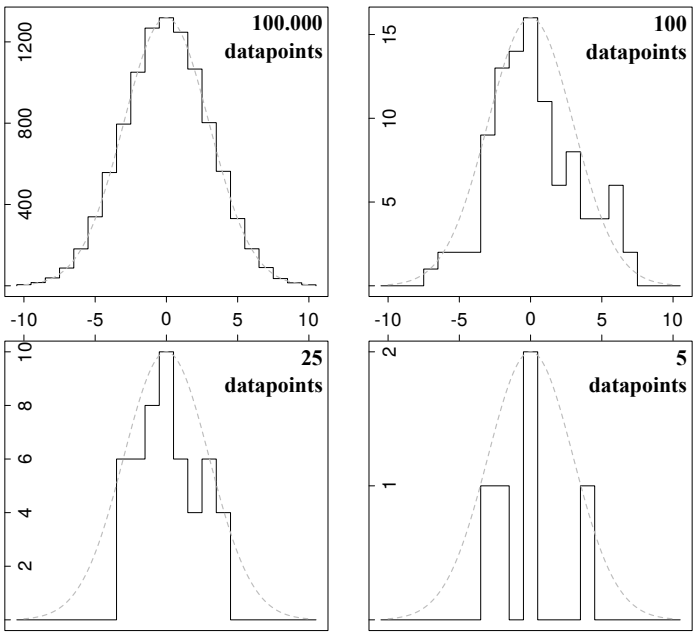



Figure 2.2: Relation between noise and sequencing depth. Under the assumption that the variability of reads near a boundary is normally distributed with a standard deviation of 3, we illustrate effects of noise by binned sampling of this distribution at different resolutions. The horizontal axes give the offset to a fragment's true position and the vertical axes the number of times an (artificial) intensity is sampled from the distribution. The dashed line represents the distribution used for sampling. A sequencing technology with an infinite resolution (**top left**; by approximation) would result into one single peak in the vertical axis. As the sequencing depth decreases, the sampled distribution deviates further from the true distribution and more peaks in the vertical axis may appear by chance (**top right** and **bottom**). Each illustration belongs to the same simulated fragment boundary, derived from the same distribution. Because FlaiMapper expects only one peak per boundary, the remaining peaks, caused by the deviation from their original distribution, are referred to as noise.

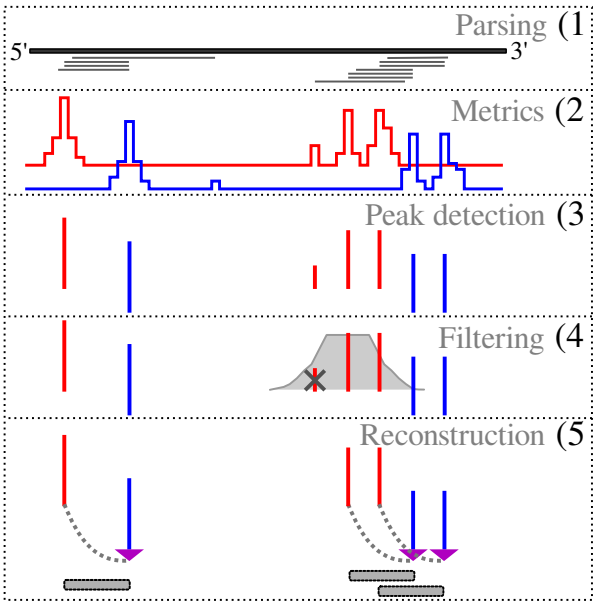



Figure 2.3: Schematic overview of the five steps that FlaiMapper performs per sncRNA. (**i**) *Parsing*: alignment file is parsed; reads (thin lines) are aligned to a sncRNA (bold line). (**ii**) *Metrics*: acquire alignment statistics; for all positions in the sncRNA: *I*: find the number of aligned start (red) and end (blue) positions (referred to as intensity) and *II*: find the average length of mapped reads (not illustrated). (**iii**) *Peak detection*: predict candidate start and end positions (vertical lines) upon the intensity vectors using peak detection. (**iv**) *Filtering*: remove candidate start and end positions expected to be detected due to noise. In the example above, a candidate start position is discarded (grey cross) because it is an artefact of the noise of its neighbour. The remaining positions are considered as actual start and ends. (**v**) *Reconstruction*: reconstruct predicted fragments (grey bars) by finding corresponding start and end positions using a balance (purple triangle) between expected distance and intensity.

- `SRR207113` RNA pool-down obtained from non-treated HeLa cells after `AGO1` and `AGO2` immunoprecipitation.
- `SRR207114` RNA pool-down obtained from `RRP40`-depleted HeLa cells after `AGO1` and `AGO2` immunoprecipitation.
- `SRR207115` Total cellular RNA was extracted after XRN1 and XRN2 depletion; `XRN` has $5' \rightarrow 3'$ exonucleolytic activity.
- `SRR207116` RNA was extracted from the nucleus.

### 2.2.1 Formal problem

For convenience, we use the term *boundary* to describe either a start or an end position of a fragment, without being specific to one of them. If an alignment of a fragment is

inspected in more detail, its boundaries are indicated by the corresponding start and end positions of the aligned reads. Fragment boundaries are variable, as indicated by the aligned reads (Figure 2.1). Read starts and ends are located at variable positions, but close to the boundaries. This results in peaks in the densities of aligned start and end positions, near the boundaries. Therefore, it seems more convenient to estimate fragments using the most common start and end positions instead of the most common read. The number of read starts or ends at a certain position in the sequence (*intensity*) decreases rapidly and symmetrically with respect to the position with the highest intensity. As a result, the peaks have characteristics compatible to a normal distribution, with its expected value being the position with the highest intensity. Because of the variability in the alignments and the limited sequencing depth, the data contain noise (Figure 2.2). In FlaiMapper, a fragment is defined as:

**Definition 1.** The region in a precursor ncRNA in-between the most common start and most common end position, as defined by aligned reads.

Consequently, the problem of finding such fragments is defined as:

**Definition 2.** Given a set of aligned reads to a precursor ncRNA, the challenge is to estimate a fragment by: (**i**) finding the correct candidate start and end positions, (**ii**) taking the optimal proportion of noise into account and (**iii**) relating the corresponding start and end positions that belong to the same fragment back to each other.

### 2.2.2 Algorithm

The FlaiMapper algorithm is divided into five sequential steps (Figure 2.3): (**i**) *parsing*, (**ii**) *metrics*, (**iii**) *peak detection*, (**iv**) *filtering* and (**v**) *reconstruction*.

#### 1. Parsing

For every ncRNA, alignments are parsed from input files. There is no preference towards a specific alignment algorithm as long as its output is in BAM format.

#### 2. Metrics

Given an ncRNA with a length of $n$ nt, the following corresponding vectors are determined:

- **Start and stop position densities**
  1. $\mathbf{p}^{5'} = \left(p_1^{5'}, p_2^{5'}, \ldots, p_n^{5'}\right)$; here, $p_i^{5'}$ is the total number of reads that have their start position (5′-end) aligned to position $i$ of the precursor ncRNA, where $1 \leq i \leq n$.

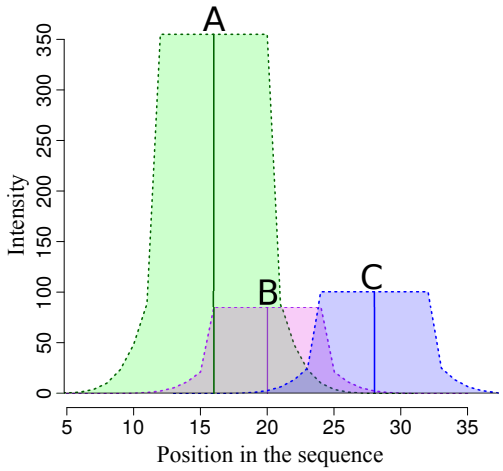



Figure 2.4: Illustration of the filter. Vector $\mathbf{c}^d = \{16, 20, 28\}$ contains three predicted peaks referred to as peak A, B and C. For each peak, the intensity is indicated with vertical solid lines, at positions 16, 20 and 28 in green, purple and blue, respectively. Peak A (350 corresponding reads) has the highest intensity, followed by C (100) and B (85). Using a top-down approach in terms of intensity, the algorithm starts with filtering the noise artefacts that belong to peak A. The borders that separate noise from true fragments are indicated with dashed lines. Peaks within the coloured areas are marked as noise. For peak A, this is the light-green area. Any other peak within this region (peak B; solid purple line) gets discarded. Thus, B is expected to noise of peak A. In the next iteration, the noise of next top peak C is taken into account. Because no other peaks fall in its corresponding blue area, none will be discarded. Since peak B is discarded already, only peak A and C remain.

2. $\mathbf{p}^{3'} = \left(p_1^{3'}, p_2^{3'}, \ldots, p_n^{3'}\right)$; here, $p_i^{3'}$ is the total number of reads that have their end position (3′-end) aligned to position $i$ of the precursor ncRNA, where $1 \leq i \leq n$.

- Read lengths

1. $\mathbf{l}^{5'} = \left(l_1^{5'}, l_2^{5'}, \ldots, l_n^{5'}\right)$; here, $l_i^{5'}$ is the average read length of reads that have their start position (5′-end) aligned to position $i$ of the precursor ncRNA, where $1 \leq i \leq n$. If no reads have their start position aligned to nucleotide $i$, $l_i^{5'} = 0$.

2. $\mathbf{l}^{3'} = \left(l_1^{3'}, l_2^{3'}, \ldots, l_n^{3'}\right)$; here, $l_i^{3'}$ is the average read length of reads that have their end position (3′-end) aligned to position $i$ of the precursor ncRNA, where $1 \leq i \leq n$. If no reads have their end position aligned to nucleotide $i$, $l_i^{3'} = 0$.

**3. Peak detection**

Candidate start and end positions are characterized by peaks in the intensity vectors. Therefore, candidate positions are estimated independently in directions $d = 5'$ (start) and $d = 3'$ (end) on vector $\mathbf{p}^d$. The methodology of independence between start and end positions used by FlaiMapper is different from methods that rely on (i) the most common read or (ii) distributions of read density. Because of this, candidate start and end positions loose their one-to-one relationship. The purpose of peak detection is to find all positions that have an intensity higher than its adjacent positions. Because of noise in the intensities, the difference in intensity with respect to the adjacent values must be above a certain threshold. For direction $d$, the algorithm detects peaks upon corresponding vector $\mathbf{p}^d$ of length $n$. Vector $\mathbf{p}^d$ should be extended with a 0 at the end, to ensure that a peak at the very last position can be called. To avoid confusion about the lengths, we denote $\mathbf{q}^d = \{\mathbf{p}^d, 0\}$ and as consequence its length $n' = n + 1$. For every $i^{th}$ position in the vector, the intensity $q_i^d$ is compared with the previous highest value.

- If the intensity is larger, it becomes the highest value, and is therefore the (new) candidate to become a peak.
- If the intensity is smaller, a drop in intensity is observed. If the drop is more than 90%, the $j^{th}$ peak is called by putting the location in $c_j^d$. Subsequently, the candidate position will be reset and iterator $j$ is increased with 1.

The formal description of peak detection is given in algorithm 1 and per ncRNA, the following vectors are added:

1. $\mathbf{c}^{5'} = \left(c_1^{5'}, c_2^{5'}, \ldots, c_k^{5'}\right)$; for a number of $k$ candidate start positions, the $i^{th}$ start position is located at nucleotide $c_i^{5'}$ of the ncRNA, where $1 \leq i \leq k$ and $1 \leq c_i^{5'} \leq n$.
2. $\mathbf{c}^{3'} = \left(c_1^{3'}, c_2^{3'}, \ldots, c_m^{3'}\right)$; for a number of $m$ candidate end positions, the $i^{th}$ end position is located at nucleotide $c_i^{3'}$ of the ncRNA, where $1 \leq i \leq m$ and $1 \leq c_i^{3'} \leq n$.

**4. Filtering**

Per fragment, multiple candidate start and end positions are frequently found due to noise. A target peak may be derived from the same fragment as surrounding peaks (Figure 2.2). For each target peak at position $i$, a filter tests whether the remaining peaks at $i'$, are indeed noise of the target. The intensity around a boundary has

**Algorithm 1** Peak detection

$\mathbf{q}^d \leftarrow \{\mathbf{p}^d, 0\}$ ▷ Input
$n' = n + 1$
$\alpha \leftarrow 0.1$ ▷ Noise threshold
$val\_previous, val\_max, pos\_max \leftarrow 0$ ▷ Init
$\mathbf{c}^d \leftarrow \{\}$ ▷ Output
$j \leftarrow 1$ ▷ Output iterator
**for** $1 \leq i \leq n'$ **do**
  **if** $q_i^d > val\_previous$ **then**
    **if** $q_i^d > val\_max$ **then**
      $pos\_max \leftarrow i$
      $val\_max \leftarrow q^d_{pos\_max}$
    **end if**
  **else if** $q_i^d < val\_previous$ **then**
    **if** $pos\_max > 0$ **and** $(\alpha \times q_i^d) < val\_max$ **then**
      $c_j^d \leftarrow pos\_max$ ▷ Call peak
      $val\_max \leftarrow 0$ ▷ Reset for next peak
      $j \leftarrow j + 1$
    **end if**
  **end if**
  $val\_previous \leftarrow q_i^d$
**end for**

characteristics of a normal distribution and decreases as the distance to the true start or stop position increases. Peaks caused by noise will have similar characteristics and therefore their intensity is expected to be (I) a function of the distance (between the positions $i$ and $i'$), and (II) proportional to the targets intensity, $p_i^d$. The filter uses these characteristics to separate peaks derived from noise, from peaks derived from other fragments. The distance $\Delta$ (in nt) between a target and noise candidate position is defined in equation 2.1, where $|\ldots|$ is the absolute value operator. $\Delta$ will always be larger than 0 because a target is not compared with itself.

$$\Delta = |i - i'|, \qquad \text{if } i \neq i' \tag{2.1}$$

Because intensities of noise artefacts are proportional to the intensity of the target, a weight matrix is used to define the area border (Figure 2.4). The weights are derived from the probability density function of a normal distribution with a standard deviation of 3, for all integer values $0 \leq x \leq 15$. To rescale densities to weights, the densities were divided through the density for $x = 0$ (0.1329808). To improve performance for peaks with a very low number of corresponding reads, the densities for a $\Delta$ of $1, 2, 3$ and 4 were changed to 1.0. The complete weight matrix $\omega$ is available in the source code. For each target peak, the filter evaluates whether any other peaks fall

within the range that can be expected by noise in both directions ($d = 5'$ or $d = 3'$) as follows (Figure 2.4):

**A** Sort $\mathbf{c}^d$ on corresponding intensities in descending order.

**B** For each $i \in \mathbf{c}^d$ target peak, remove corresponding noise artefacts:

**B.i** For all $i' \in \mathbf{c}^d_{i' \neq i}$ noise candidate peaks, find $\omega_\Delta$ and define whether the candidate is noise or belongs to a separate fragment by evaluating equation 2.2. If the equation is *true*, the candidate peak is considered to be a noise artefact of the target; immediately discard candidate $\mathbf{c}^d_{i'}$. If it is *false*, the candidate peak is not considered to be a noise artefact and must be retained.

$$p^d_{i'} \leq (\omega_\Delta \times p^d_i) \tag{2.2}$$

As a result of the filter, $\mathbf{c}^{5'}$ and $\mathbf{c}^{3'}$ may have shrunk and their respective lengths $k$ and $m$ may have become smaller.

**5. Reconstruction**

The peaks are expected to be the actual boundaries of fragments. Because start and stop positions do not have a direct one-to-one relationship with each other, a trace back is required to reconstruct the fragments. Because the number of predicted start ($k$) and end ($m$) positions is not necessarily equal, it is convenient to start reconstruction from direction $d$ with $min(k, m)$ candidate positions, and find for each position the most likely corresponding position $d'$. Direction $d$ is defined in equation 2.3, and $d'$ is its complement.

$$d = \begin{cases} 5' \textit{ (start positions)} & \text{if } k \leq m \\ 3' \textit{ (end positions)} & \text{if } m > k \end{cases} \tag{2.3}$$

Important information required for reconstruction is the expected length of reads that were used for detecting a peak, given in $\mathbf{l}^{5'}$ and $\mathbf{l}^{3'}$.
Indeed:

- A fragment that starts at position i is expected to have its end $i^*$ close to: $i^* \approx i + l^{5'}_i$.
- A fragment that ends at position i is expected to have its start $i^*$ close to: $i^* \approx i - l^{3'}_i$.

The number of reads that correspond to a start position is expected to be close to the number of reads that define the end position: $p_i^d \approx p_{i'}^{d'}$. Thus, the reconstruction process needs a balance between (I) the expected position and (II) the expected intensity of the counter position. This is achieved by conjoining an associated start and end position into a fragment as follows:

**A** Sort $\mathbf{c}^d$ based on corresponding intensities in descending order.

**B** For all $i \in \mathbf{c}^d$ candidate positions find expected counter position $i^*$

**B.i** For all candidate counter positions $i' \in \mathbf{c}^{d'}$, the goal is to determine the counter position which has the optimal trade-off between a small distance with the expected counter position and a small difference in intensity. This is achieved by solving of equation 2.4. In the equation 0.09 is an arbitrary chosen weight that forms the linear balance between distance and intensity. A predicted fragment is determined with its start position: $\min(i, j)$ and end position: $\max(i, j)$. After reconstruction, positions $i$ and $j$ are discarded from $\mathbf{c}^d$ and $\mathbf{c}^{d'}$, respectively.

$$j = \max_{i'}((1 - 0.09 \times |i^* - i'|) \times p_{i'}^{d'}), \qquad \text{for all } i' \in \mathbf{c}^{d'} \tag{2.4}$$

## 2.3 Results

### 2.3.1 Validation of FlaiMapper performance

#### miRBase

To get an impression of FlaiMapper's performance, its predictions for corresponding miRNAs detected in dataset `SRP002175` were compared with miRNA annotations in miRBase 20 [146]. Because all experiments in this dataset are generated under the same conditions, alignments to the same ncRNA from all 12 experiments were merged, to maximize resolution. Of the 1037 miRNAs annotated in miRBase, 169 lacked supporting reads, and were not included in the quality assessment (because they would influence the outcome negatively without assessing the algorithm itself). Of the remaining 868 miRNAs, FlaiMapper was not able to predict a fragment that overlaps an annotated miRNA only 21 times, with a corresponding sensitivity of $847/868 = 0.98$.

A detailed assessment was performed by measuring the offset between a predicted fragment and a miRNA annotation in miRBase (Figure 2.5, top). We assume that

Table 2.1: Comparison

Table 2.2: Summary of predicted fragments on datasets SRP002175 [64], SRP006788 [129], SRP028959 [143], SRP034013 [144] and SRP041082 [145].

| dataset | type | pre-miR | SNORD | SNORA | tRNA | SCARNA | MISC |
|---|---|---|---|---|---|---|---|
| *total (in ncRNA reference)* | ncRNAs | 1386 | 264 | 106 | 451 | 23 | 39 |
| SRP002175 (Pigment): *RNA extracted* | ncRNAs | 645 | 141 | 51 | 381 | 16 | 29 |
| *Total cellular RNA (12 merged experiments)* | fragments | 947 (37.2%) | 202 (7.9%) | 56 (2.2%) | 1210 (47.5%) | 26 (1.0%) | 107 (4.2%) |
| SRP006788 (HeLa) | ncRNAs | 463 | 92 | 20 | 359 | 8 | 28 |
| *Total cellular RNA* | fragments | 680 (26.1%) | 140 (5.4%) | 24 (0.9%) | 998 (38.3%) | 11 (0.4%) | 755 (28.9%) |
| SRP006788 (HeLa): *Total cellular RNA* | ncRNAs | 455 | 108 | 29 | 367 | 16 | 32 |
| *after RRP40 core subunit depletion* | fragments | 686 (24.2%) | 181 (6.4%) | 34 (1.2%) | 1104 (38.9%) | 24 (0.8%) | 806 (28.4%) |
| SRP006788 (HeLa): *RNA pool-down* | ncRNAs | 415 | 19 | 14 | 151 | 6 | 15 |
| *after AGO immunoprecipitation* | fragments | 560 (39.7%) | 30 (2.1%) | 15 (1.1%) | 208 (14.7%) | 9 (0.6%) | 590 (41.8%) |
| SRP006788 (HeLa): *pool-down from RRP4 core* | ncRNAs | 393 | 30 | 16 | 155 | 4 | 18 |
| *depleted cells after AGO immunoprecipitation* | fragments | 517 (38.6%) | 42 (3.1%) | 17 (1.3%) | 209 (15.6%) | 6 (0.4%) | 550 (41.0%) |
| SRP006788 (HeLa): *RNA pool-down after* | ncRNAs | 513 | 129 | 65 | 301 | 18 | 33 |
| *after XRN immunoprecipitation* | fragments | 738 (25.9%) | 294 (10.3%) | 112 (3.9%) | 760 (26.6%) | 47 (1.6%) | 903 (31.6%) |
| SRP006788 (HeLa) | ncRNAs | 451 | 129 | 57 | 281 | 16 | 33 |
| *RNA was extracted from the nucleus* | fragments | 649 (27.3%) | 280 (11.8%) | 87 (3.7%) | 519 (21.9%) | 32 (1.3%) | 806 (34.0%) |
| SRP028959 (HeLa): *SRR954957* | ncRNAs | 168 | 10 | 0 | 167 | 2 | 4 |
| *Total cell small RNA preparation* | fragments | 203 (30.8%) | 15 (2.3%) | 0 (0%) | 229 (34.7%) | 2 (0.3%) | 210 (31.9%) |
| SRP028959 (HeLa): *SRR954958* | ncRNAs | 146 | 28 | 0 | 104 | 3 | 6 |
| *Nuclear small RNA preparation* | fragments | 169 (33.3%) | 33 (6.5%) | 0 (0%) | 126 (24.8%) | 4 (0.8%) | 176 (34.6%) |
| SRP028959 (HeLa): *SRR954959* | ncRNAs | 162 | 2 | 0 | 183 | 1 | 6 |
| *Cytoplasmic small RNA preparation* | fragments | 197 (29.1%) | 2 (0.3%) | 0 (0.0%) | 274 (40.4%) | 1 (0.1%) | 204 (30.1%) |
| SRP034013 (B cells): | ncRNAs | 1012 | 213 | 99 | 409 | 23 | 39 |
| *Total cellular RNA (3 merged experiments)* | fragments | 2230 (42.4%) | 670 (12.7%) | 384 (7.3%) | 1478 (28.1%) | 105 (2%) | 391 (7.4%) |
| SRP041082 (Prostate): | ncRNAs | 803 | 230 | 90 | 381 | 21 | 38 |
| *Total cellular RNA (2 merged experiments)* | fragments | 1454 (37.9%) | 552 (14.4%) | 201 (5.2%) | 1222 (31.9%) | 99 (2.6%) | 304 (7.9%) |

miRBase provides the "ground truth" in terms of miRNA annotations. The results show that the majority of FlaiMapper predictions are identical to miRBase annotations. Also, the decrease of the offset bars (Figure 2.5, top) is symmetrical, indicating no systematic inconsistency. 89% of the predicted start positions are identical to the reference. When an offset of 1 nt is allowed, the ratio correctly predicted start positions increases to 95%. In contrast, 54% of the end positions are predicted identical to the reference. When an offset of 1 nt is allowed, this increases to 82%. In addition, their offset-bars descend slower. This indicates that estimation of start positions is more precise.

To get an impression of the influence of sequencing depth on accuracy of start and end positions corresponding to miRNA and miRNA* predictions, the number of corresponding reads (*intensity*) was plotted as a function of the offset for dataset `SRP002175`. Figure 2.6 illustrates that with the increase of sequencing depth, the offset for both start and end positions decreases. However, at identical intensity, end positions have higher offset than start positions and require deeper sequencing to achieve the same accuracy.

In addition, we analysed the performance of FlaiMapper on three Supplementary Data generated on other sequencing platforms. Performance was similar to the performance described above.

### Existing software

Previous research reported a comparable method [147]. Its goal is to detect miRNA-offset-RNAs (moRNAs), fragments adjacent to pre-miRNAs. The authors also used the method to demonstrate its ability to discover miRNAs. The algorithm has no restrictions to a certain type of fragment, so its outcome should be comparable to FlaiMapper and therefore their performances can be compared with each other.

In contrast to our method, BlockBuster relies on the overall aligned read density of a fragment and transforms this into a normal distribution. Consequently, the prediction of the start and end positions are dependent on each other since they are derived from the same distribution. This implies that the alignments near a fragment's start and end position should have a symmetrical shape. BlockBuster's performance was tested on dataset `SRP002175` and the alignments of the 12 corresponding experiments were merged and converted into the BED format for compatibility. BlockBuster was used with a variety of parameters where its *scale* parameter of 0.05 and *distance* of 26 were found to be rough estimates for the optimum. Optimum is defined by the lowest amount of root squared error of all predicted miRNAs where error is defined as the offset between a predicted miRNA and its miRBase annotation. The following

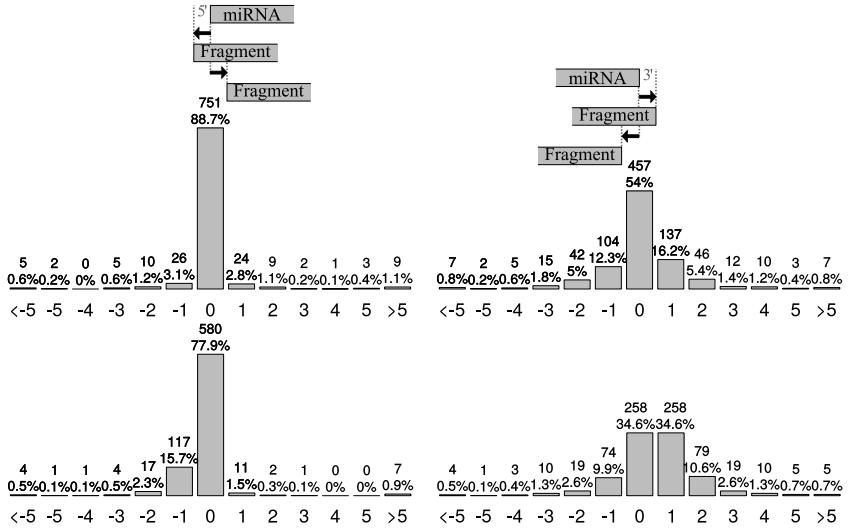



Figure 2.5: Comparison between the predictions of miRBase 20 and FlaiMapper (top) and between miRBase and BlockBuster (bottom) on dataset `SRP002175`, indicating the offset of the start positions (left) and end positions (right). The vertical axes reflect the amount of predictions that correspond to a particular offset. The horizontal axes represent the offset between a predicted fragment and an annotated miRNA; exact matches are located at 0, offsets $< 0$ are predicted upstream the miRNA's boundaries and offsets $> 0$ - downstream.

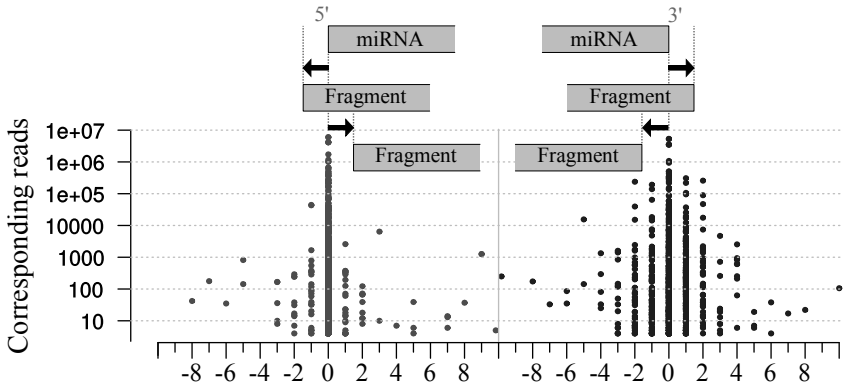



Figure 2.6: Relation between sequencing depth and offset in predicted miRBase annotations for dataset `SRP002175`. The horizontal axis represents the offset for the start (left) and end (right) positions between a miRNA annotation and a predicted fragment. Predictions with exact matches are located at 0, offsets $< 0$ are predicted upstream the miRNA's boundaries and offsets $> 0$ - downstream. The vertical axis represents the number of reads corresponding to a start (left) or end (right) position. The figure indicates that the higher the number of corresponding reads, the lower the offset. Overall, the predicted start positions have a lower offset than the end positions, for the same sequencing depth.

is observed (Figure 2.5, bottom):

- A lower sensitivity: $745/868 = 0.86$ (compared with 0.98 in FlaiMapper).
- A lower accuracy:
  - The number of start positions identical to miRBase is 78% compared with FlaiMapper's 89%. When an offset of 1 nt is allowed, both tools show comparable accuracy of 95%.
  - The number of end positions identical to miRBase is 34% compared with FlaiMapper's 54%. When an offset of 1 nt is allowed, 79% is predicted correctly compared with 82% using FlaiMapper.
- The offset bars of the start position decrease asymmetrically.
- The predictions are shifted; for the start positions, there is an overhang towards the pre-miRNAs 5′-end and for the end positions there is an overhang towards the pre-miRNAs 3′-end, indicating that the predicted fragments are on both sides systematically longer than the miRBase annotations.

### 2.3.2 Fragment analysis

We used FlaiMapper to detect fragments originating from sncRNAs other than pre-miRNAs on datasets `SRP002175`, `SRP006788` and supplementary datasets `SRP028959`, `SRP034013` and `SRP041082`. To maximize resolution, alignments of dataset `SRP002175`, `SRP034013` and `SRP041082` were merged. For `SRP006788` and `SRP028959`, experiments were analysed individually to investigate possible influence of specific RNA processing-related treatments. The numbers of predicted fragments, categorized per type of precursor, are given in Table 2.2. The ratios of predicted fragments per precursor type were used for principal component analysis (Figure 2.7). The largest difference between fragment profiles was observed between datasets `SRP002175`, `SRP034013` and `SRP041082`. Since they are from different tissues and experiments, this is expected. Sub-conditions within dataset `SRP006788` that are taken from `AGO` pool-downs showed nearly identical fragment profiles. In addition, samples of which nuclear RNA was extracted from independent HeLa experiments processed on different sequencers, also show very similar fragment profiles.

Table 2.2 shows that the `AGO`-pool down samples of dataset `SRP006788` have a relatively high proportion of fragments derived from pre-miRNAs compared with the other samples in the dataset. This observation is consistent with the known association of miRNAs with `AGO` proteins [131]. On the same time, it also suggests that fragments derived from other precursor types than pre-miRNAs are not associated

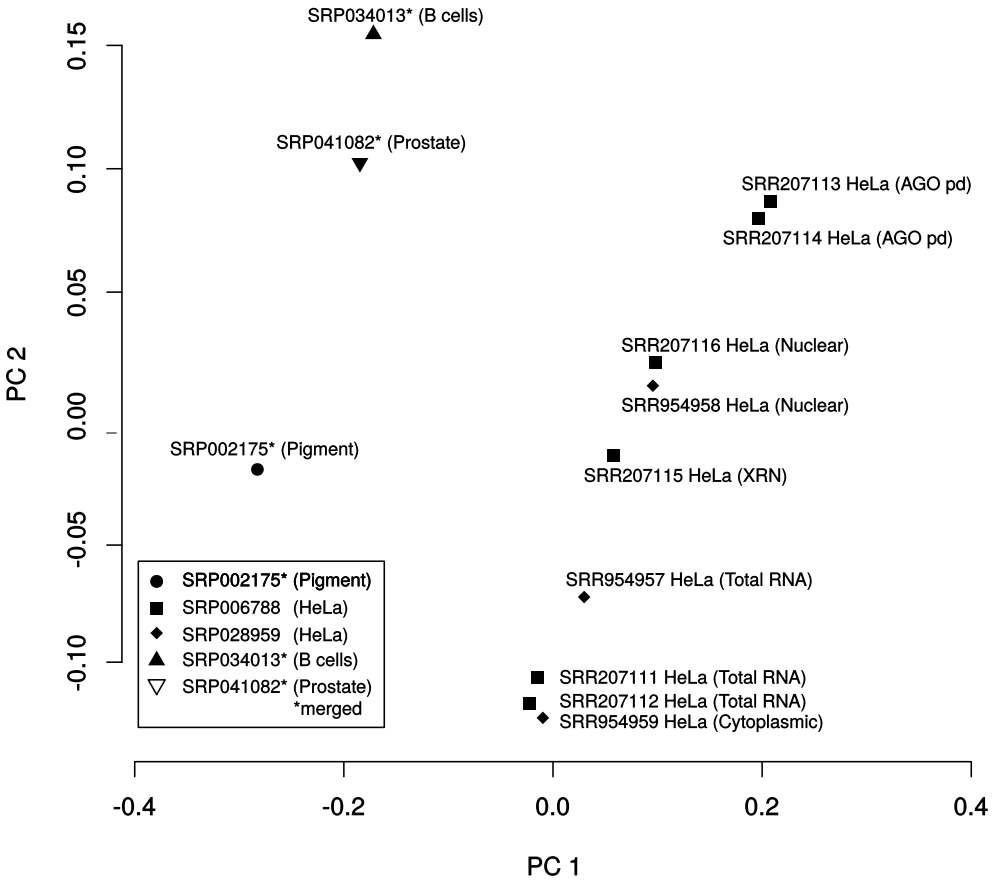



Figure 2.7: The first two components of principal component analysis applied on percentages of predicted fragments (Table 2.2 comprise 92% of the variance. The circle represents the profile of merged dataset `SRP002175` (pigment cells, Illumina GA II); squares individual experiments of dataset `SRP006788` (HeLa cells, Illumina GA II); diamonds individual experiments of `SRP028959` (HeLa cells, Ion Torrent PGM); triangle pointing up dataset `SRP034013` (B cells, Illumina HiSeq2000) and the triangle pointing down dataset `SRP041082` (prostate cells, Illumina HiSeq2000). The last three datasets are addressed in the Supplementary Data. Datasets corresponding to separate tissue types demonstrate unique fragment profiles. Experiments with HeLa cells taken under similar circumstances group together; total RNA samples from independent datasets `SRP006788` and `SRP028959` cluster closely together.

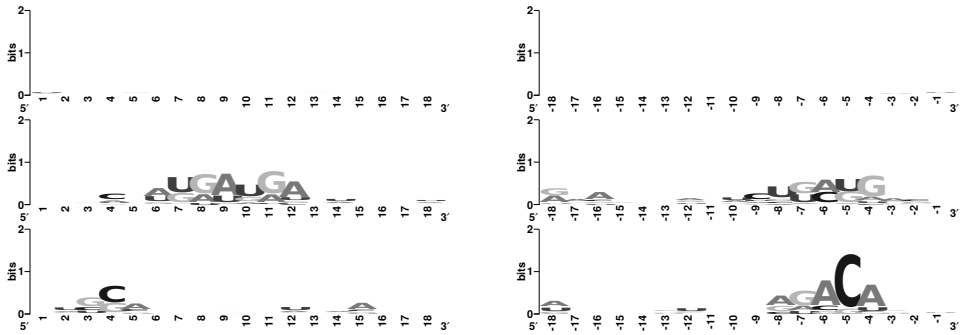


Figure 2.8: Sequence logos [148] of the fragments located on the 5′- (left) and the 3′-end (right) of the precursor ncRNA from the predictions on dataset `SRP002175`. Only fragments with their centre located at $\leq 40\%$ of the precursor are used for prefix analysis and fragments located at $\geq 60\%$ for suffix analysis. (**top**) Pre-miRNA fragments: no common prefix or suffix motif (based on 520 prefixes and 427 suffixes). (**middle**) C/D-box snoRNA fragments: (**middle left**) the prefix motif `UGAUGA` is found more often around the sixth and seventh nucleotide (based on 130 prefixes). (**middle right**) The suffix `UGAUG` is found more often around the −eighth nucleotide (based on 72 suffixes). (**bottom**) H/ACA-box snoRNA fragments: (**bottom left**) the fourth nucleotide of the prefix appears to be `G`/`C` enriched, although the bit score is mild (based on 37 prefixes). (**bottom right**) The suffixes are enriched with the motif `ACA` from the −sixth until the −fourth nucleotide (based on 19 suffixes).

with `AGO` to the same extent. Taken together, this supports the biological context of the FlaiMapper-derived fragment profiles.

### 2.3.3 Sequence logos

To show that the outcome of FlaiMapper can be used to explore characteristics of sncdRNAs like sequence motifs, fragments were analysed for over-represented pre- or suffixes using sequence logo plots [148] (Figure 2.8). The analysis on pre-miRNA-derived fragments did not indicate over-represented motifs. Although it must be stated that the number of predicted fragments derived from C/D-box snoRNAs is lower than for pre-miRNAs, the analysis confirms that the C-box is over-represented [138]. It also shows that sequences of H/ACA-box snoRNA-derived fragments located at the 3′-half of the precursor, most often contain the suffix `ACANNN`, where `ACA` is the precursor's ACA-box and `N` can represent any nucleotide. On the 5′-half of the H/ACA box the fourth is preferentially occupied by a `G`/`C`. However, due to the mild bit score and the low number of used fragments, this observation should be interpreted with caution. Because of the highly conserved sequences and the high number of genomic copies, tRNAs were excluded from motif analysis.

## 2.4 Discussion

We set out a method able to extract and annotate ncRNA fragments, because such annotations can be helpful in further high-throughput research. We designed FlaiMapper, a computer program to predict ncRNA fragments using small RNA-seq alignments. Benchmarking indicated that FlaiMapper is able to predict 97.8% of the miRNAs with corresponding reads. 95% of the miRNAs 5′-end and 82% of the 3′-end predictions were concordant with miRBase annotations. For this analysis, data from the Illumina Genome Analyser II was used. A similar accuracy was observed for sequencing data derived from the Ion Torrent PGM and Illumina HiSeq2000 (Supplementary Data), indicating FlaiMapper can perform well on data from different platforms. We demonstrated that FlaiMapper performs better than existing similar software [147].

FlaiMapper predicts fragments by looking at the most common start and end positions in alignments. It can be argued whether the most common start and end positions should indeed provide the evidence for the prediction of a fragment, since the most common read could be used instead. However, the most common start and end positions should usually be covered by a higher number of reads. This corresponds to a higher resolution, which is especially advantageous for the prediction of fragments with a low read coverage, and should therefore also be more robust towards noise. Together with the demonstrated high performance, this implies that predictions based on start and end position densities provide a more appropriate solution for fragment annotation.

The weights used in the filtering step are based on a normal distribution with an arbitrary chosen $\sigma$. These parameters probably find their optimum in relation with sequencing protocols, 5′/3′-end-specific processing factors or different families of fragments. Therefore, once there is a better understanding of the processing of fragments, it is recommended to spend effort in optimizing these parameters.

Additional analysis indicated that FlaiMapper's performance in miRNA annotation is positively correlated to sequencing depth. Predicted 3′-ends of miRNAs have a larger offset compared with miRBase annotations than the 5′-ends, even for the same sequencing depth. The higher variability of the miRNAs 3′-ends has earlier been reported [146]. In addition, research on the classification of sncRNAs indicated that metrics corresponding to the variability in the alignment are indeed higher for the 3′-end in miRNAs [149]. They were able to indicate that different levels of variability correspond to specific types of sncRNAs. Possible explanations could be RNA post-processing or RNA editing. This means that alignments over the entire fragment can be asymmetrical because of a larger variation observed at miRNAs 3'-ends. Since BlockBuster assumes reads to be symmetrically distributed over a fragment,

this might explain why (i) its accuracy is lower, (ii) its predictions are longer and (iii) shifted with respect to miRBase.

Although it seems counter-intuitive that an ncRNA can produce different fragments that originate from an overlapping region, there are situations where overlapping fragments can be expected. This can be stressed by recalling the not fully understood tRNA processing mechanism(s), where tRNA halves and tRFs spanning similar regions have been reported. Therefore, FlaiMapper has no restriction to the prediction of overlapping fragments, similar to the method of [147].

Sequence logos indicated that the ACA-box, as part of the the `ACANNN` suffix, is over-represented and position specific in fragments derived from the 3′-halve of the H/ACA-box snoRNAs. The analysis also confirmed that the C-box of C/D box snoRNAs is over-represented in corresponding fragments. Yet, this result may be biased by the existence of multiple, highly homologous, genomic copies of certain C/D-box snoRNAs such as `HBII-52` and `HBII-85`.

Fragment characteristics can play an important role in finding associations with their processing mechanism. For example, although the larger variability of the alignments at the 3′-end of miRNAs affects performance, it clearly indicates a difference in the processing of miRNAs ends. Characteristics such as 5′- and 3′-end entropy have been successfully used in the classification of sncRNAs [149, 150]. Using such characteristics on the fragment level, for example for clustering or classification, might provide new insights into the processes of production, functioning or degradation of fragments or indicate a possible sub-grouping. The future in-depth analysis of sncdRNAs will require more comprehensive datasets with higher sequencing depth and more statistical power.

## 2.5 Conclusion

The lack of a sncdRNAs annotation is a short coming in small RNA-seq analysis. To overcome this, we designed the computer program FlaiMapper. FlaiMapper has a high performance in predicting miRNA boundaries, but can be used for the annotation of any type of sncdRNA. Examination of FlaiMapper-predicted sncdRNAs indicated different type specific characteristics: 5′/3′-end-specific variability in miRNAs, associations between `AGO` and relative fragment profiles in dataset `SRP006788` and a position-specific sequence motif in a subset of the H/ACA-box fragments. These characteristics indicate that FlaiMapper is a good starting point for the downstream analysis of small RNA sequencing experiments.

## Acknowledgements

The authors would also like to thank Bas Pigmans for his work on sequencing alignment methodology.

## Funding

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n$^o$201438 and from the research programme ALW-VENI Grant 863.12.014 financed by the Netherlands Organisation for Scientific Research (NWO).

# 3 C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer

pmid: 26041889, doi: 10.18632/oncotarget.4172

Elena S. Martens-Uzunova [1], Youri Hoogstrate [1], Anton Kalsbeek [1], Bas Pigmans [1], Mirella Vredenbregt-van den Berg [1], Natasja Dits [1], Søren Jensby Nielsen [2,3], Adam Bake [2,4], Tapio Visakorpi [5], Chris Bangma [1] and Guido Jenster [1]

*Oncotarget*, 6(19):17430−17444, jul 2015

[1] *Department of Urology, Erasmus University Medical Center, Be 362a, PO Box 2040, 3000 CA Rotterdam, The Netherlands* [2] *Exiqon A/S, Vedbaek, Denmark* [3] *Nuevolution A/S, Copenhagen, Denmark* [4] *Chr. Hansen A/S, Hørsholm, Denmark* [5] *Institute of Biosciences and Medical Technology - BioMediTech, University of Tampere and Tampere University Hospital, Tampere, Finland*

Supplementary material: http://www.oncotarget.com/index.php?journal=oncotarget&page=rt&op=suppFiles&path=4172

## Abstract

Small nucleolar RNAs (snoRNAs) are dynamically regulated in different tissues and affected in disease. SnoRNAs are processed further to stable smaller RNAs. We sequenced the small RNA transcriptome of prostate cancer (PCa) at different PCa stages and generated a quantified catalogue of 3927 small non-coding RNAs (sncRNAs) detected in normal and malignant prostate tissue. From these, only 1524 are microRNAs. The remaining 2401 sncRNAs represent stable sncRNAs species that originate from snoRNA, tRNA and other sncRNAs. We show that snoRNA-derived RNAs (sdRNAs) display stronger differential expression than microRNAs and are massively upregulated in PCa. SdRNAs account for at least one third of all small RNAs with expression changes in tumor compared to normal adjacent tissue. Multiple sdRNAs can be produced from one snoRNA in a manner related to the conservation of structural snoRNA motifs. Q-PCR analysis in an independent patient cohort (n=106) confirmed the processing patterns of selected snoRNAs (`SNORD44`, `SNORD78`, `SNORD74` and `SNORD81`) and the cancer-associated up-regulation of their sdRNAs observed in sequencing data. Importantly, expression of `SNORD78` and its sdRNA is significantly higher in a subset of patients that developed metastatic disease, demonstrating that snoRNA and sdRNAs may present as novel diagnostic and/or prognostic biomarkers for PCa.


**Keywords:** `GAS5`; `SNORD78`; prostate cancer; sdRNA; snoRNA

## 3.1 Introduction

Malignant transformation and cancer progression cause changes in the expression and function of microRNAs (miRNAs) [17, 16]. However, the effects of these processes on other small non-coding RNAs (sncRNAs) are less understood. Recently, we demonstrated the abundance and differential expression of small nucleolar RNA-derived RNAs (sdRNAs) in the small transcriptome of prostate cancer (PCa) [151]. It is generally accepted that small nucleolar RNAs (snoRNAs) are housekeeping, non-protein coding molecules that associate with specific sets of proteins to maintain proper ribosomal maturation in the nucleolus.

Still, several reports show that snoRNAs have tissue-specific expression [152, 153], and may present as novel cancer biomarkers. For example, the H/ACA-box snoRNA `SNORA42` is commonly overexpressed in non-small cell lung cancer (NSCLC) and its expression is significantly inversely correlated with survival [154, 140]. Similarly, the levels of C/D-box snoRNAs `SNORD33`, `SNORD66` and `SNORD76` are significantly elevated in plasma from NSCLC patients compared with cancer-free controls and can provide potential biomarkers for early detection [155]. In chronic lymphocytic leukemia (CLL), heterogeneous snoRNA expression patterns discriminate major CLL subgroups and can stratify patients in different prognostic groups [156], while in multiple myeloma snoRNA expression patterns are associated with distinct molecular subtypes of the disease [157].

Furthermore, resent research demonstrates that the molecular alterations of snoRNA are functionally linked to basic cellular processes associated with cancer proposing either tumor suppressor or oncogene role for different snoRNAs. In NSCLC, `SNORA42` acts as a putative oncogene. Its overexpression enhances cell proliferation and growth in bronchial epithelium and cancer cells, while its knockdown in NSCLC cells inhibits colony forming [140]. In acute promyelocytic leukemia, the `SNORD112-114` is specifically activated in a subset of patients and may influence cell growth through a negative regulation of the cell cycle and the Rb pathway [158]. On the contrary, in peripheral T-cell lymphoma, over-expression of the candidate prognostic marker `SNORD71` (`HBII-239`) is associated with favorable outcome [159]. The C/D box snoRNA `SNORD50`, a translocation partner of BCL6 in B-cell lymphoma [160], is a candidate tumor suppressor significantly associated with clinically relevant prostate [161] and breast [141] cancer. In hepatocellular carcinoma (HCC), `SNORD113-1` has been identified as a tumor suppressor [162]. Down-regulation of this snoRNA is associated with decreased survival of HCC patients, while reconstitution of its expression suppresses tumorigenesis *in vitro* and *in vivo*. In glioblastoma, decreased expression of the `GAS5` encoded `SNORD76` is associated with an aggressive phenotype [163]. Ectopic

expression of this tumor suppressor snoRNA inhibits tumorigenicity by arresting cancer cells in S phase *in vitro* and inhibits orthotopic tumor growth *in vivo*. In breast cancer and head and neck squamous cell carcinoma the low expression of another GAS5 encoded snoRNA, SNORD44, correlates with markers of aggressive pathology and poor prognosis [164, 165].

At present, little is known about the pathways of snoRNA turnover. Apparently, snoRNAs are further processed to sdRNAs in a vast variety of organisms [24]. It is yet unclear whether sdRNAs are novel functional entities or footprint-products of snoRNA downstream processing shielded from degradation by snoRNA-interacting proteins. A miRNA-like activity has been proposed for sdRNAs derived from H/ACA-box snoRNAs (H/ACA-sdRNAs) based on their apparent size of 20-24 nt equivalent to miRNAs, the ability to promote repression of complementary targets *in vitro*, and the association with DICER and AGO complexes [166, 167, 138, 168, 139, 137, 169]. In contrast, a bimodal size distribution of 17-20 nt and 27-30 nt has been reported for sdRNAs derived from C/D-box snoRNAs (C/D-sdRNAs) [151, 167, 138]. C/D-sdRNAs are not efficiently incorporated in AGO2 suggesting a different function for this type of sdRNAs [170]. In addition, it has been reported that the highly abundant in brain 'orphan' snoRNAs, SNORD115 and SNORD116, are processed into larger sdRNAs (34-73 nt) that complex with spliceosomal proteins and may regulate the alternative splicing of target mRNAs [136, 171]. Association of C/D-box snoRNAs with novel RNPs and involvement in alternative splicing has been previously observed in mice for the brain specific MBII-52 [172]. Interestingly, both MBII-52 and its human ortholog SNORD115 produce larger sdRNAs (34-73 nt). Similar observation has also been made for sdRNA regions of SNORD88C, which can influence the alternative splicing of FGFR3 pre-mRNA [128]. At the same time, studies in *Drosophila sp.* and in human cells show that snoRNAs are strongly enriched in the nuclear fractions of chromatin-associated RNA and possibly involved in the maintenance of open chromatin structure [173].

Here, we report the deep sequencing of patient-derived samples from normal prostate, and PCa in different disease stages, which reveals sdRNA production from the vast majority of known human snoRNAs. At least 78 of the detected sdRNAs demonstrate strong differential expression in cancer. Furthermore, the expression of some sdRNAs and their precursors is associated with clinical progression and metastatic occurrence.

## 3.2 Results

### 3.2.1 Library preparation and sequencing

We generated 10 sncRNA libraries from normal adjacent prostate (NAP), benign prostate hyperplasia (BPH), different stages of PCa, and metastatic lymph node (LN) prepared from fresh-frozen patient material (FF) (Supplementary Table 1). To estimate the influence of sample storage on sncRNA abundance and stability, we prepared a replicate library from formalin-fixed, paraffin-embedded tissue (FFPE) from tumor samples used for one of the fresh-frozen libraries (group 3). All sequencing reactions yielded approximately 14 million raw reads, each (13,468,284 to 15,393,670) with the FFPE library producing the highest raw read number (Figure 3.1a).

### 3.2.2 Annotation of the sncRNA transcriptome

The correct mapping of sncRNA reads is challenged by the fact that predominant isoforms of miRNAs and other sncRNAs such as snoRNAs may vary from the mature sequences annotated in public databases. Differences can be caused by alternative 3′-end modifications [174] or alternative 5′-/3′-end positions of the detected sncRNA. Additionally, the length of mature sncRNA transcripts can be ambiguously annotated in different public databases. To map as many sequence reads as possible, we constructed a custom small non-coding RNA database (sncRNAdb) that consists of 2271 unique small non-coding RNA species corresponding to 2356 unique genomic loci (Supplementary Figure 1 and Supplementary File 1).

Mapping to sncRNAdb resulted in the detection of a total of 1637 unique sncRNAs expressed across any of the 11 libraries with an average of 1229 per library. 70% to 84% of the reads generated from fresh-frozen samples and only 52% of the reads generated from FFPE could be annotated by sncRNAdb (Figure 3.1a and Supplementary Table 2). The majority of annotated reads mapped to 873 pre-miRNAs (85.5-95.6%), 385 tRNAs (1.89-7.4%), 228 C/D-box snoRNAs (0.3-1.9%), and 91 H/ACA-box snoRNAs (0.0-0.1%) (Figure 3.1b, 3.1c and Supplementary Tables 2 and 3).

Interestingly, in PCa samples we detected up to 27% more C/D-box and up to 52% more H/ACA-box snoRNAs compared to NAP or BPH. Furthermore, total snoRNA read-counts were increased at least two-fold, indicating possible activation of snoRNA-gene expression in response to malignant transformation. In contrast, the number of detected miRNAs remained relatively stable and the total miRNA read-counts changed by no more than 19% (min. 9,202,300, max. 11,367,682) (Figure 3.1c, Supplementary Figure 2 and Supplementary Tables 3 and 4).

Table 3.1: Number of differentially expressed sncdRNAs.

| **Comparison** | Total number of sncdRNAs with significantly changed expression | | | | Type of sncRNA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Corrected p-value[1] (≤0.01)** | **Fold change (≥ ±4.0)** | **Expression in cancer ($2^{nd}$ group)** | **tRF** | **miRNA** | **sdRNA SNORD** | **sdRNA SNORA** | **Other** |
| **NAP vs. PCa 6 cured (grl vs. gr 3)** | **200** | **68** | Up 66 | 25 | 9 | 23 | 5 | 4 |
| | | | Down 2 | 0 | 2 | 0 | 0 | 0 |
| **PCa 6 cured vs. PCa 6 recurrent (gr3 vs. gr4)** | **155** | **14** | Up 0 | 0 | 0 | 0 | 0 | 0 |
| | | | Down 14 | 10 | 3 | 1 | 0 | 0 |
| **PCa 6 recurrent vs. PCa 6 recurrent (gr4 vs. gr10)** | **156** | **16** | Up 8 | 0 | 4 | 1 | 1 | 2 |
| | | | Down 8 | 0 | 5 | 3 | 0 | 0 |
| **PCa 6 cured vs. PCa 6 recurrent (gr3 vs. grill)** | **141** | **18** | Up 5 | 0 | 3 | 0 | 1 | 1 |
| | | | Down 13 | 3 | 7 | 3 | 0 | 0 |
| **NAP vs. PCa 7 recurrent (grl vs. gr 5)** | **168** | **34** | Up 18 | 2 | 3 | 9 | 3 | 1 |
| | | | Down 16 | 5 | 8 | 1 | 0 | 2 |
| **NAP vs. PCa 8 recurrent (grl vs. gr6)** | **238** | **105** | Up 87 | 27 | 15 | 29 | 5 | 11 |
| | | | Down 18 | 7 | 8 | 1 | 0 | 2 |
| **NAP vs. metastatic LN (grl vs. gr 8)** | **302** | **157** | Up 115 | 36 | 28 | 35 | 5 | 11 |
| | | | Down 42 | 7 | 35 | 0 | 0 | 0 |
| **NAP vs. TURP HR (grl vs. gr 7)** | **254** | **104** | Up 87 | 24 | 38 | 18 | 1 | 4 |
| | | | Down 17 | 3 | 11 | 1 | 0 | 2 |
| **BPH vs. TURF HR (gr2 vs. gr7)** | **352** | **202** | Up 92 | 21 | 42 | 21 | 2 | 6 |
| | | | Down 110 | 94 | 15 | 0 | 0 | 1 |
| **NAP vs. BPH (grl vs. gr2)** | **250** | **107** | Up 99 | 93 | 5 | 0 | 0 | 1 |
| | | | Down 8 | 0 | 5 | 1 | 0 | 2 |
| **FF vs. FFPE (gr3 vs. gal)** | **689** | **540** | Up 462 | 261 | 6 | 80 | 56 | 59 |
| | | | Down 78 | 2 | 73 | 2 | 1 | 0 |

[1]Z-test, Bonferroni corrected p-value

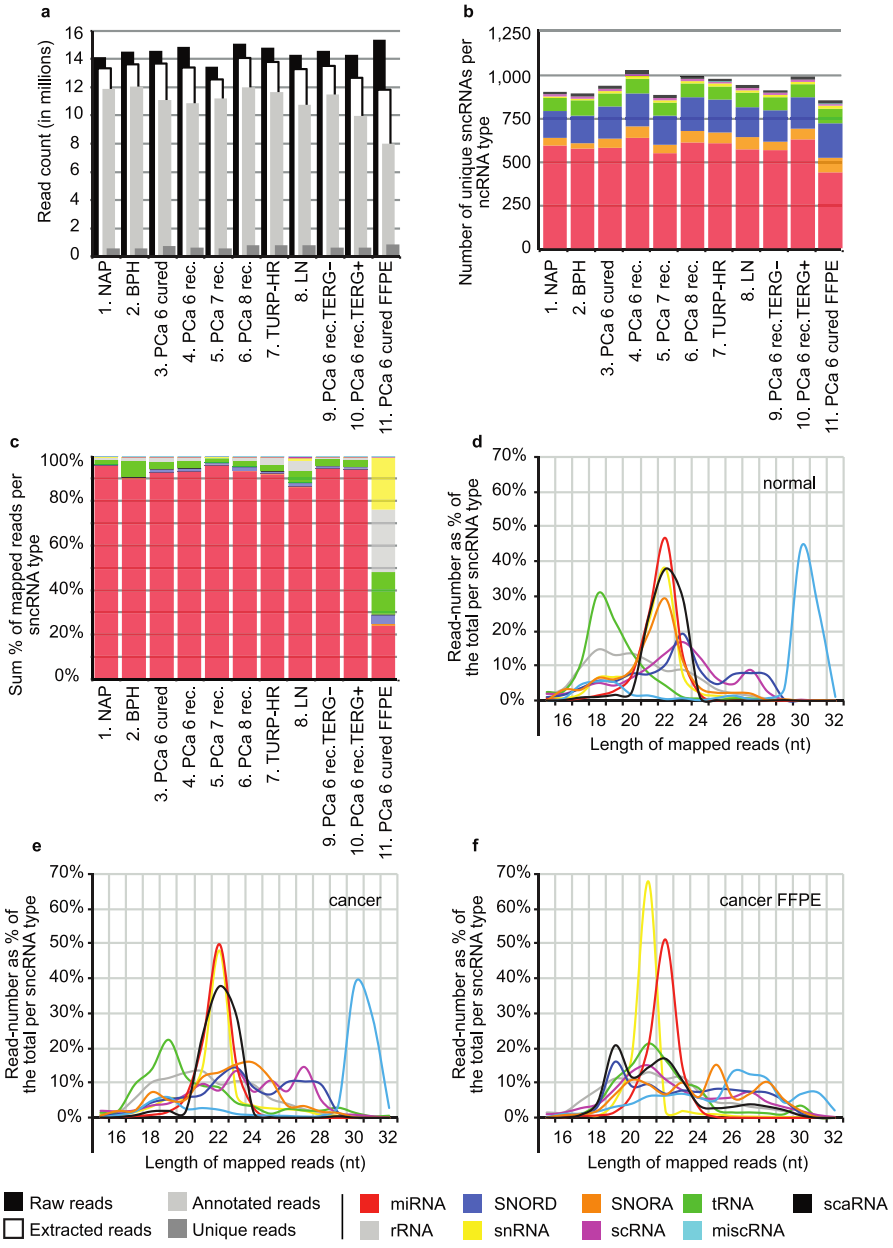



Figure 3.1: **Summary of sncRNA sequencing data from PCa patient samples.** (**a**) Number of retrieved raw, extracted, annotated, and unique reads generated for each one of the sequencing libraries. (**b**) Number of detected sncRNA-species per library. (**c**) Relative abundance of different sncRNA-types per library. Read-length distribution in normal (**d**) and cancer libraries (**e**) derived from fresh-frozen, (**f**) and FFPE material. Each sncRNA type is represented by different color: miRNA (red), SNORD (dark blue), SNORA (orange), tRNA (green), scaRNA (black), rRNA (gray), snRNA (yellow), scRNA (magenta), other miscellaneous RNAs (light blue).

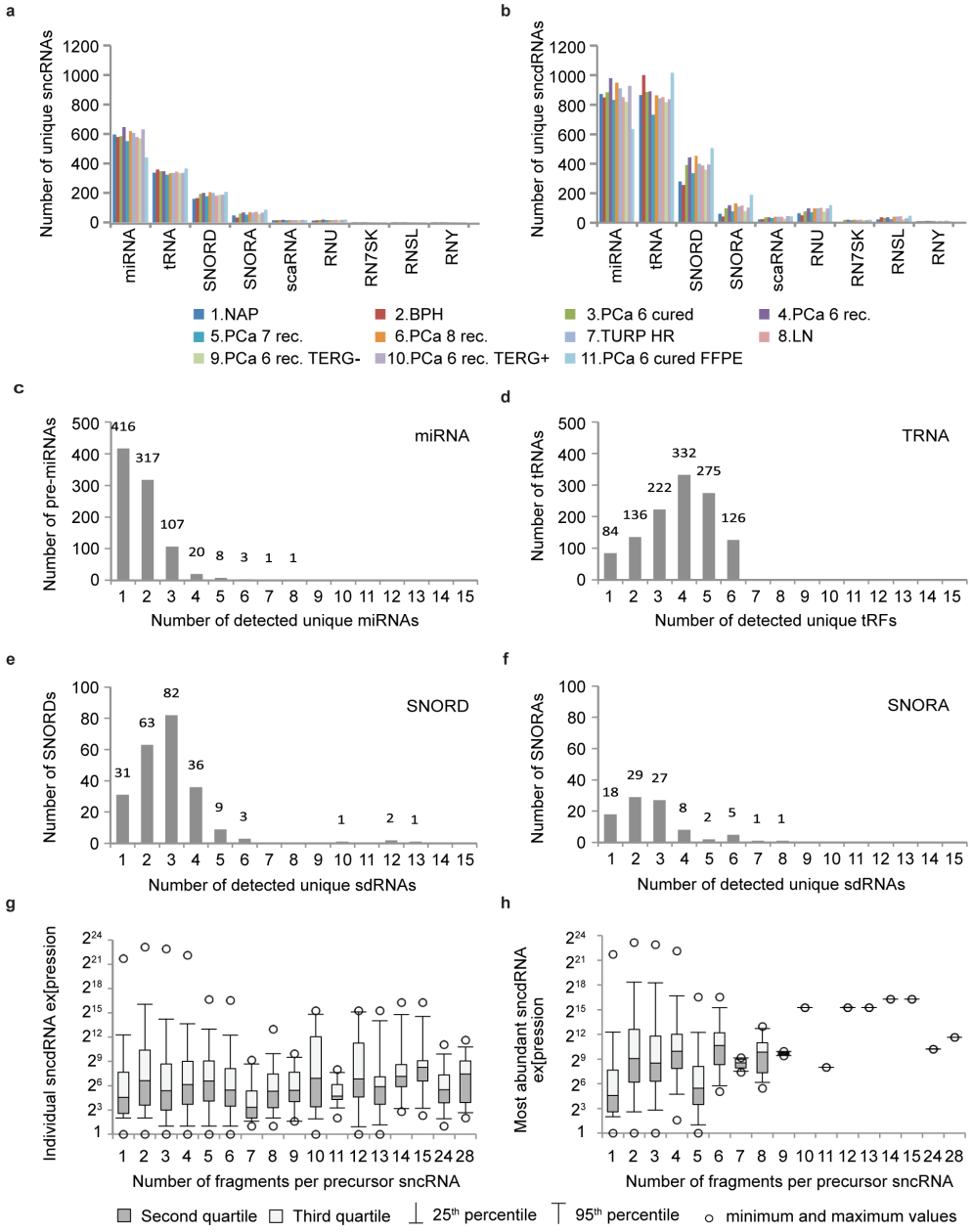



Figure 3.2: **FlaiMapper results.** (**a**) Total number of detected sncRNA precursors per RNA type and sequencing library. (**b**) Total number of sncdRNAs per precursor type and sequencing library. (**c, d, e, f**) Different types of sncRNAs produce different number of fragments. (**g**) Relation between the number of fragments produced per precursor RNA and the expression levels of individual fragments. (**h**) Relation between the number of fragments produced per precursor RNA and the expression levels of the most abundant fragment per precursor.

We also examined the read-length associated with different types of sncRNAs. As expected, miRNA reads had a narrow size distribution between 21 and 23 nt in all libraries. Similar size range was observed for snRNA- and scaRNA-derived RNAs in fresh-frozen libraries. In concordance with our previous results [151], we detected a size peak at 23 nt and a plateau between 26-28 nt for reads mapping to C/D-box snoRNAs. Interestingly, reads mapping to H/ACA snoRNAs and tRNAs demonstrated a shift in size distribution between normal and malignant samples (Figure 3.1d, 3.1e and Supplementary Figure 3) suggesting cancer-associated alterations in sncRNA processing.

Comparison of the sncRNA composition of the FFPE library with its fresh-frozen counterpart demonstrated that the relative miRNA read-content in FFPE decreased 3.9-fold from 92% to 24% of the total annotated reads. On the contrary, the number of reads mapping to other sncRNA species was strongly elevated i.e. sequence read-counts were increased 152-fold for snRNAs, 12.7-fold for H/ACA-box snoRNAs, 5.6-fold for tRNAs, and 2.7-fold for C/D-box snoRNAs ($\chi^2$ test, $p < 0.0001$ for all tested groups) (Figure 3.1c, Supplementary Figure 2 and Supplementary Table 4). The size distribution of read-length in FFPE material was also strongly affected for all examined ncRNA groups except for miRNAs. (Figure 3.1e, 3.1f and Supplementary Figure 3). These observations can be explained with the higher level of RNA degradation in FFPE for transcripts longer than miRNA [175, 176].

### 3.2.3 Mapping and annotation of sncRNA-derived RNAs (sncdRNAs)

The majority of miRNA reads in small RNA sequencing data map to the specific location on their pre-miRNA corresponding to the mature miRNA. Similarly, reads mapping to other sncRNAs, originate from specific positions on their precursor rather than being randomly derived and can represent specific, biologically functional, smaller RNA species, e.g. sdRNAs or tRNA fragments (tRFs) [177]. Nevertheless, the assignment of RNA-seq sequence-reads to specific sdRNAs or tRFs for quantitation purposes is hampered by the lack of proper annotation. Furthermore, many sncRNAs produce multiple fragments [136, 171] that may overlap each other, which further complicates the exact determination of their origin loci and a subsequent quantitative analysis.

To correctly determine the boundaries of sdRNAs, tRFs and other sncRNA-derived RNAs (sncdRNAs) in our dataset and annotate their specific location on the precursor sequence, we applied the computational algorithm Fragment Location Annotation and Identification Mapper (*FlaiMapper*) and evaluated its performance in this data set as described [178]. Shortly, FlaiMapper predicted 5′- and 3′-miRNA ends were

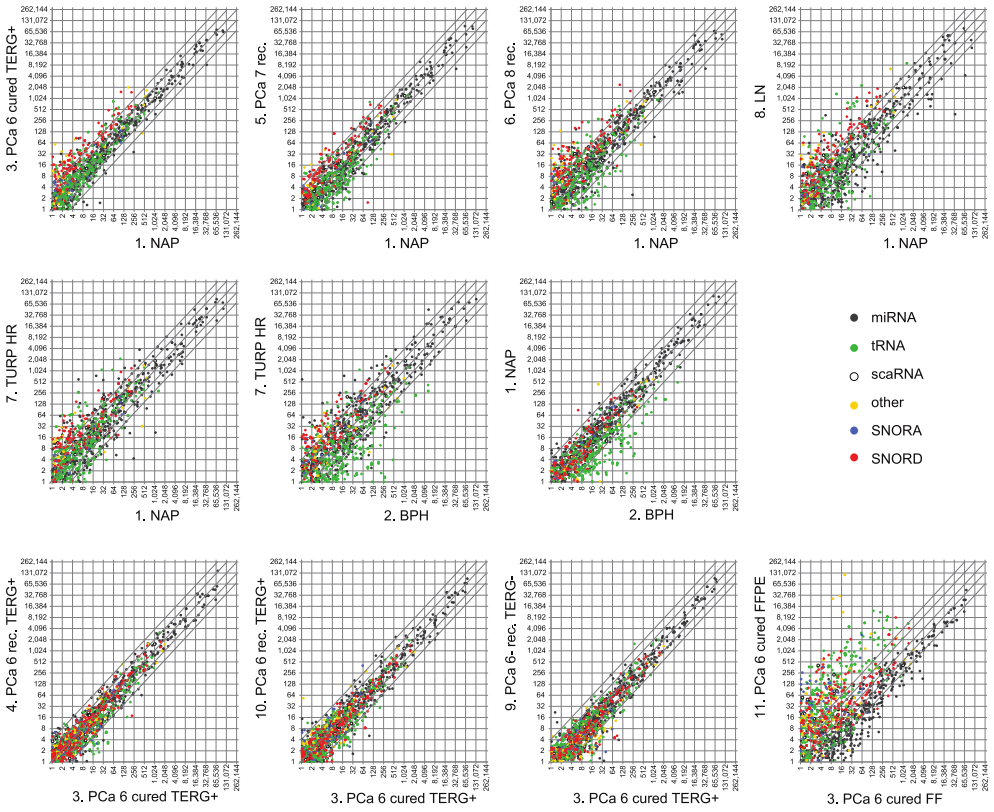



Figure 3.3: **Global expression changes of sncdRNAs in normal and malignant prostate tissue.** Upper and middle panels present scatterplots comparing the normalized expression values of individual sncdRNA (dots) in each prostate cancer library (PCa) during progressing disease to these in the library prepared from normal adjacent prostate tissue (NAP). The expression of sncdRNAs in the hormone-refractory, transurethral resection of the prostate (TURP HR) library is also compared to the benign prostate hyperplasia (BPH) library since the latest represents the normal counterpart of malignant transurethral resection of the prostate material. Differences in sncdRNA expression between biological replicates of Gleason 6 cancers (PCa 6) as well as comparison of a fresh-frozen library (FF) with its formalin-fixed, paraffin-embedded (FFPE) counterpart derived from the same patients are presented in the lower panels. Each sncdRNA type is presented by a different color. Diagonal lines across each scatterplot represent fold change difference in expression. Middle line, crossing the horizontal and vertical axes at 0, no expression change; lines crossing the vertical and horizontal axes at 2, twofold expression change; lines crossing the vertical and horizontal axes at 4, four-fold expression change. Cured, no disease relapse after radical prostatectomy; rec., recurrent disease, biochemical or metastatic relapse after surgery; LN, metastatic lymph node sample; TERG+, `TMPRSS2-ERG` fusion gene event; TERG-, no `TMPRSS2-ERG` fusion event; Numbers (6, 7 or 8) after PCa indicate the pathological Gleason score of the tumors in the respective group.

compared with the 5′- and 3′-end boundaries of corresponding mature miRNAs in miRBase, v17 [146]. 82% of the detected miRNAs had a correctly determined 5′-end exactly matching miRBase annotation. An additional 11% had an offset of 1 nt. In agreement with previous observations [146], 3′-ends of mature miRNAs had higher variability and matched miRBase annotations for 45%. From the investigated miRNAs additional 33% had 1 nt offset, and 14%, 2 nt offset (Supplementary Figure 4). Given the high confidence with which FlaiMapper identified 5′- and 3′-end boundaries of *bona fide* miRNAs, we performed annotation of all sncdRNAs in our fresh-frozen libraries. We detected 3927 unique sncdRNAs derived from different precursor classes. From these, 1524 originated from miRNAs, 1175 from tRNAs, 657 from C/D-box snoRNAs, 244 from H/ACA-box snoRNAs, and 327 from other sncRNA species (Supplementary Table 5 and Supplementary File 2). The total number of detected unique sncdRNAs was higher than the number of detected unique precursor species, showing that individual sncRNA precursors produce more than one sncdRNA (Figure 3.2a, 3.2b and Supplementary Figure 5). For example, the majority of pre-miRNAs produced one or two miRNAs corresponding to the guide and passenger strand. For C/D box snoRNAs we detected between 1 and 6 sdRNAs originating from the same precursor, with the exception of the unusually long `SNORD3A`, `SNORD3B` and `SNORD3C`, which give rise to 10 to 13 C/D-sdRNAs. Most H/ACA-box snoRNAs produced between 1 and 3 sdRNAs, while for tRNAs we detected between 1 and 6 tRFs per precursor (Figure 3.2d-f). Other examined sncRNAs in our libraries produced a varying number of fragments ranging from 3 for the telomerase RNA component to 28 for the small nuclear 7SK RNA (Figure 3.2c; Supplementary Figure 6 and 7). We next argued that the expression level of the sncRNA-precursor might positively influence the number of sncdRNAs detected per sncRNA. We examined the distribution of expression values of individual sncdRNAs in relation to the number of sncdRNAs derived per sncRNA and could not observe a strong dependency between the median expression levels of sncdRNAs and the total number of sncdRNAs produced per sncRNA. We obtained similar results when the expression level of the most abundant sncdRNA per precursor was used as a surrogate measure of the expression of the precursor RNA (Figure 3.2g-h). Based on these results, we can conclude that multiple sncdRNAs originating from the same sncRNA can be detected independently of their (low) expression level or the expression level of their precursor. Vice versa, different precursor RNAs can produce only one sncdRNAs with very high abundance. Hence, it can be assumed that the number and quantity of different sncdRNAs do not directly reflect the abundance of their precursor but, like miRNAs, are probably also influenced by additional aspects of cellular metabolism, e.g. association with protein complexes and/or turnover rates.

The size of unique sdRNAs ranged between 15 and 29 nt (Supplementary Figure

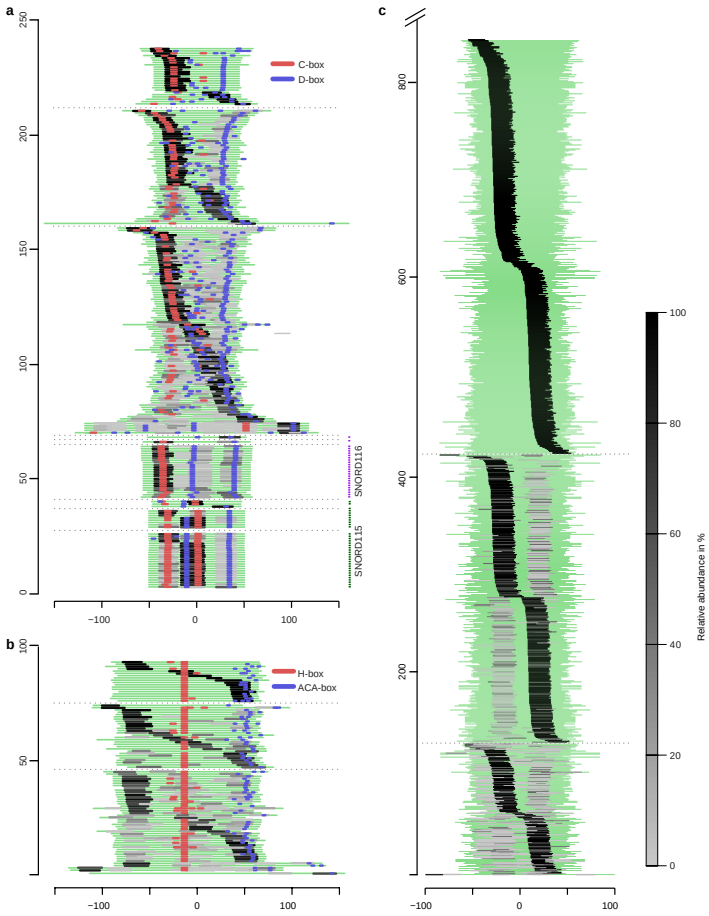



Figure 3.4: **Global processing patterns and relative abundance of sdRNAs and miRNAs expressed in prostate (cancer) tissues.** (**a**) Full-length C/D-box snoRNAs are aligned relative to the middle nucleotide of each sequence. (**b**) H/ACA-box snoRNAs are aligned based on the position of the H-box. (**c**) Pre-miRNAs are aligned relative to the middle nucleotide of each sequence. A green line represents each full-length sncRNA. Sequences are extended 10 nt at each end to avoid mapping ambiguity caused by incorect annotation. Positions of detected conserved H/ACA-boxes or C/D-boxes are shown in blue and red. Light and dark grey lines indicate the positional origin of sdRNAs, miRNAs and miRNA*s. The color intensity corresponds to the relative abundance of sncdRNAs originating from the same precursor (read-count as a percentage of the total read-count per precursor), *e.g.* if only one sdRNA per snoRNA-precursor is detected it is assigned 100% abundance, if two or more sdRNAs originate from the same snoRNA the sdRNA with the highest read-count is given the darkest color and the sdRNA with the lowest read-count, the lightest. Thin dashed lines separate each panel into three subgroups where sncRNAs producing only one sncdRNA are on top, sncRNAs producing two sncdRNAs are in the middle and those producing there or more sncdRNAs are on the bottom. The highly sequentially conserved, multiple gene-copy C/D-box snoRNAs from the `SNORD116` (`HBII-85`) and `SNORD115` (`HBII-52`) families are grouped together below other C/D-box snoRNAs. The X-axis indicates the position of sncdRNAs relative to the center of their precursor sequence. The Y-axis depicts the number of full-length sncRNA precursors.

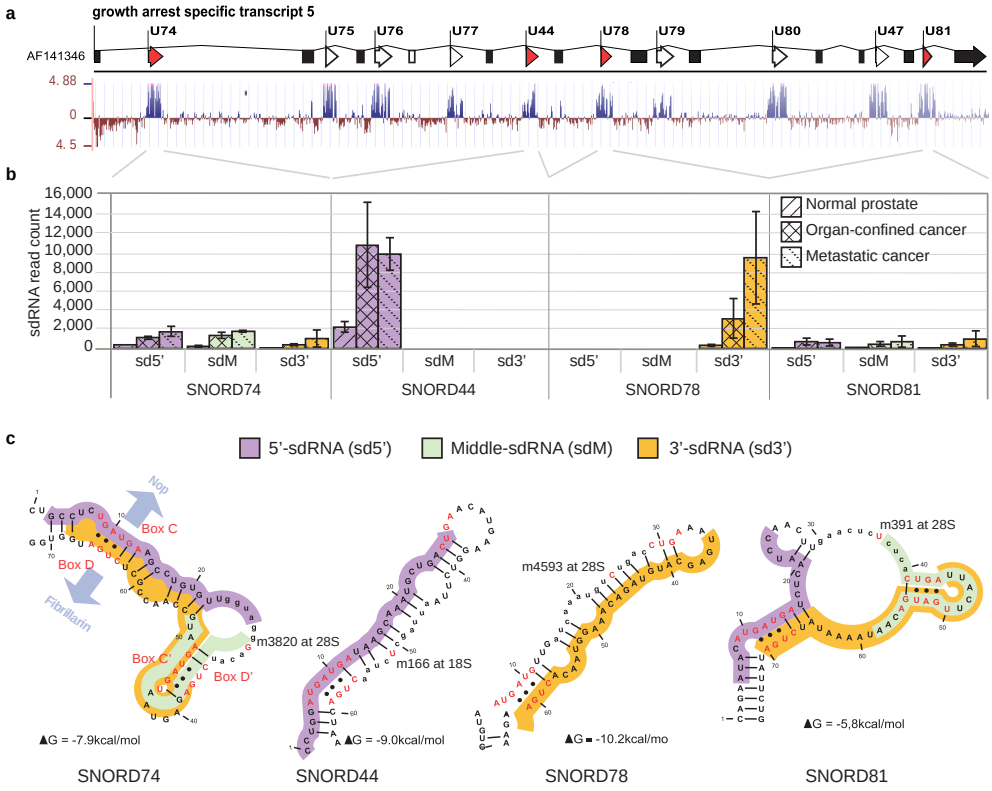



Figure 3.5: **Genomic organization, conservation, secondary folding, fragmentation pattern and expression of SNORD44, SNORD74, SNORD78, and SNORD81.** (**a**) SNORD44, SNORD74, SNORD78, and SNORD81 are transcribed simultaneously from the conserved intronic regions of the protein-non-coding GAS5 gene. Spliced exons (boxes); snoRNA loci (arrows). (**b**) sdRNAs from SNORD44, SNORD74, SNORD78, and SNORD81 are up-regulated in PCa. Expression is comparable to PCa-relevant microRNA (not shown). SNORD74 and SNORD81 produce equally expressed 5′- (sd5′), middle- (sdM), and 3′-sdRNAs (sd3′) with overlapping sdM and sd3′. SdRNAs originating from the middle regions extend towards the antisense box and their 5′-ends map exactly adjacent to the nucleotide complementary to the targeted ribosomal residue. (**c**) SNORD44 and SNORD78 produce predominantly one sdRNA either from the 5′- or 3′-arm of the snoRNA. The position of core snoRNP-proteins NOP58/56 and FIBRILLARIN (indicated at SNORD74) is dependent on the kink-turns formed by non-complementary base-pairing (dots) of the conserved external sequence boxes C and D and/or the internal boxes C'/D'. The rRNA-complementary antisense-box (lower case) is exposed and contains the nucleotide targeted for modification (red), positioned exactly 5 nt upstream of the D or D' box.

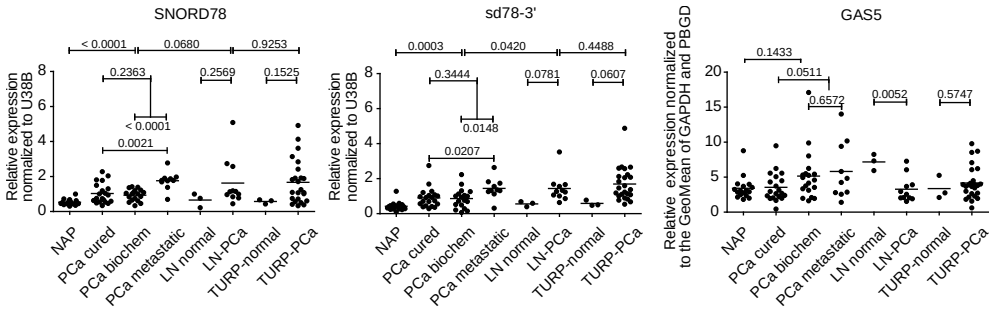



Figure 3.6: **Q-PCR validation of snoRNA and sdRNA expression in an independent cohort of patient samples. NAP**, normal adjacent prostate (n=17); **PCa-cured**, radical prostatectomy sample, no disease relapse after radical prostatectomy (n=20); **PCa-biochem**, radical prostatectomy sample, patients manifested biochemical disease relapse after surgery (n=18); **PCa-met**, radical prostatectomy sample, metastatic disease progression after surgery (n=10); **LN-normal**, normal lymph node (n=3); **LN-PCa**, metastatic lymph node (n=11); **TURP-normal**, transurethral resection of the prostate sample that does not contain cancer cells (n=3); **TURP-PCa**, transurethral resection of the prostate sample that contains cancer cells (n=24). Horizontal line marks the mean of each group. Patient number in each group is indicated in brackets. P-values from unpaired two-tailed t-tests (alpha level 0.05) are indicated above each comparison.

6). However, when the expression of individual sdRNAs of the same length were accounted, we observed a predominant size of 23 nt for the majority H/ACA-sdRNAs and a binominal size distribution for C/D-sdRNAs with two predominant sizes of 22-23 nt and 28 nt (Supplementary Figure 7), which is in agreement with our previous findings and other reports [151, 167, 138, 149]. C/D-sdRNAs demonstrated a broader size distribution, which however could be a reflection of the broader size range of their precursors.

### 3.2.4 sdRNAs are differentially expressed in prostate cancer

Previously, we observed differential expression of sdRNAs between PCa specimens [151]. To examine if such changes are a cancer-specific event we compared the expression of FlaiMapper defined sncdRNAs between normal (NAP and BPH) and malignant tissues of progressing disease (PCa, LN, TURP). We detected between 34 and 202 sncdRNAs with significant differential expression (Table 3.1, Figure 3.3, and Supplementary File 3). Approximately one third of the differentially expressed RNAs in each comparison comprised C/D-sdRNAs upregulated in cancer (Figure 3.3). In contrast, only one sdRNA was differentially expressed between non-malignant samples (NAP and BPH) and only five, between biological replicate samples (PCa, Gleason 6, groups 3, 4, and 10). This suggests that the accumulation of C/D-sdRNAs is primarily driven by malignant transformation.

To examine the effect of sample storage on fragment abundance, we compared the expression of sncdRNAs between the FFPE sample and its fresh-frozen (FF) counterpart. We limited comparison analyses to sncRNAs detected in any of the FF libraries. MiRNAs had decreased expression in FFPE compared with sdRNAs, tRFs and other sncdRNAs (Table 3.1, 3.3 and Supplementary Figure 8). Nevertheless, the reduction of miRNA expression in FFPE appears to be the result of a global decrease in miRNA read-counts compared to read-counts of other sncdRNAs (Figure 3.1c) since the relative expression of miRNAs correlated strongly between both conditions (Pearson $\rho = 0.9289$) (Supplementary Figure 8). This was not observed for sdRNAs (Pearson $\rho = 0.6557$ for C/D-sdRNAs and 0.3895 for H/ACA-sdRNAs) or other sncdRNAs, which have longer precursors and may be more susceptible to degradation in FFPE material.



### 3.2.5 SdRNAs demonstrate specific global processing patterns in prostate tissue

Given the discrete size and specific expression of sdRNAs, we examined detected snoRNAs for the presence of a common processing pattern. To be able to compare with miRNAs, we aligned all snoRNA and pre-miRNA sequences and visualized the position and relative abundance of the corresponding sdRNAs and miRNAs (Figure 3.4, Supplementary file 4 and 5). The majority of sdRNAs originated from equivalent locations of their precursors. Often, one predominant sdRNA was observed per precursor. The position of these predominant sdRNAs was not dependent on the total number of smaller species detected per precursor sequence, showing a rather uniform fragmentation pattern consistent with the precursor-type. This is in agreement with previously suggested specific snoRNAs processing and accumulation of smaller RNAs observed in cell lines [170, 128].

In our patient samples, predominant H/ACA-sdRNAs originate from either the 5′-arm of the first H/ACA-snoRNA hairpin (38.5%) or the 3′-arm of the second hairpin including the region of the ACA-box (31%) (Figure 3.4b). C/D-box snoRNA produce twice as many predominant sdRNAs originating from the 5′-terminus that contain a C-box (60.1%) compared to 3′-terminal sdRNAs that contain a D-box (30.1%) (Figure 3.4a).

Interestingly, individual C/D-sdRNAs with highly similar sequences demonstrate almost identical fragmentation pattern, which is also dependent on the conservation of snoRNA structural features. For example, snoRNAs from the highly conserved, multiple gene-copy `SNORD116` family (`HBII-85`), which have a degenerated C'-box (`UGAGUGA`) produce four sdRNAs where the most abundant one maps to the 5′-region

covering the C-box. SnoRNAs from the SNORD115 (HBII-52) family with conserved C'/D'-boxes produce three sdRNAs, with the most abundant ones mapping to the middle-region and covering the entire K-loop including the C'/D'-box. In contrast, the larger snoRNAs from the SNORD3 family, which lack a conserved C-box, produce between 10 and 13 overlapping sdRNAs with the most predominant mapping to the 3′-end.

SNORD115 and SNORD116 sdRNAs differ in size and position from the previously reported highly abundant psnoRNAs processed from the orthologous MBII-52 and MBII-85 detected by RNase protection assays [136, 171, 179]. This discrepancy could be explained by the implicit methodology differences between sncRNA sequencing and RNase protection assays. However, these differences could be also caused by tissue-specific sdRNA accumulation as previously described for sdRNAs originating from SNORD88C (HBII-180C) [128] or by the dependence of processing mechanisms on the structural conservation of C/D-box snoRNAs. Of note, SNORD115, SNORD116, or SNORD88C-originating sdRNAs were detected at low abundance in our samples.

### 3.2.6 Processing and expression of sdRNAs originating from GAS5 encoded C/D-box snoRNAs is related to the conservation of structural C'/D'-boxes

We investigated whether the fragmentation pattern of other C/D-box snoRNA is also dependent on structural feature conservation. For this we analyzed the positional origin of a highly abundant sdRNA produced from the 3′-end of SNORD78 [151] and other sdRNAs from the same locus. SNORD78 is intronically encoded by the Growth Arrest Specific 5 gene (GAS5) together with 9 other C/D-box snoRNAs [180]. All 10 SNORDs are presumably simultaneously transcribed as a GAS5 precursor-transcript, which undergoes intron removal and posttranscriptional processing. We could detect sdRNAs from all 10 GAS5-encoded snoRNAs. However, only four (SNORD44, SNORD78, SNORD74 and SNORD81) snoRNAs produced abundant sdRNAs (Figure 3.5 and Supplementary Figure 9).

Interestingly, SNORD74 and SNORD81 produced three abundant sdRNAs with similar, relatively low expression levels that mapped to the 5′-, 3′-, and middle region of the snoRNAs. The 3′- and middle sdRNAs overlapped each other and covered the K-loop and the conserved canonical C'/D'-box (Figure 3.5). In contrast, SNORD78 and SNORD44, which lack the canonical C'/D'-box, produced predominantly one 28 nt long sdRNA each, mapping to the 3′-arm for SNORD78 (sd78-3′) or the 5′-arm of SNORD44 (sd44-5′). Sd78-3′ and sd44-5′ were strongly upregulated in samples prepared from malignant tissue compared to normal or benign, while middle- and opposite arm-derived

sdRNAs were present only at very low read-counts in all libraries (Supplementary Figure 9a and 10).



### 3.2.7 SNORD78 and sd78-3′ expression is associated with metastatic PCa

To validate our sequencing data, we tested the expression of SNORD44, SNORD78, SNORD74, SNORD81, and their derivate sdRNAs, in an independent patient cohort of 106 fresh-frozen clinical samples by quantitative real-time PCR (Q-PCR). To evaluate whether increased sdRNA expression is a result of a general activation of the GAS5 locus, we also measured the expression of the spliced GAS5 transcript (Figure 3.6 and Supplementary Figure 9b). All tested snoRNAs and sdRNAs were upregulated in organ-confined PCa compared to normal adjacent controls. This was not related to an elevation of the spliced GAS5 transcript, which did not demonstrate pronounced expression changes between NAP and PCa. Interestingly, overlapping sdRNAs originating from the same snoRNA as well as full-length snoRNAs were simultaneously detectable by Q-PCR suggesting the existence of multiple conformational states of these snoRNAs.

Sd78-3′, SNORD78 and GAS5 expression was also detectable in different normal basal prostate epithelium cell lines (PNT2C2, RWPE) prostate cancer cell lines (PC346C, LAPC4, VCAP, LNCAP, 22RV1, PC3, and DU145N) as well as in hepatocellular carcinoma (HEP3B) and colon adenocarcinoma (COLO205) cells demonstrating that SNORD78 processing to sd78-3′ is not restricted to prostate tissue or cells. Similar to patient data, the expression levels of sd78-3′ and SNORD78 were not correlated to the expression of the GAS5 host gene (Supplementary Figure 11).

Consistent with our previous results [151], sd78-3′ was upregulated in the LN library generated in this study, suggesting association of this sdRNA with aggressive disease. Therefore, in the validation cohort we stratified patients with organ-confined disease at the time of radical prostatectomy into three groups: cured after radical prostatectomy, biochemical disease reoccurrence, and progression to metastatic disease after surgery. Strikingly, the expression of sd78-3′ and its precursor SNORD78 in the third group was significantly higher already at the time of surgery, suggesting an early involvement in PCa progression and possible prognostic marker potential for these sncRNAs.

## 3.3 Discussion

SncRNAs and in particular miRNAs emerged as novel modulators of gene expression and regulators of fundamental cellular processes often disturbed in cancer. At the same time, long-known "housekeeping" RNAs such as snoRNAs appeared to have tissue-specific expression altered in solid tumors and hematological malignances [155, 176, 181]. Furthermore, several studies discussed above demonstrate that similarly to miRNA, snoRNAs carry diagnostic and/or prognostic biomarker potential in different cancer types [154, 140, 155, 156, 157, 161, 141].

Improved detection and screening over the last decade led to large increase in prostate cancer detection. However, the majority of presently diagnosed patients carry clinically insignificant tumors, which would never progress to a life threatening disease. Without the presence of better prognostic markers, many patients undergo unnecessary invasive surgical treatment.

Prompted by our previous findings on elevated levels of snoRNA fragments in metastatic PCa [151] and by accumulating evidence from sequencing data that demonstrates processing of snoRNA to stable smaller sdRNAs [167, 128, 177], we combined RNA sequencing of human prostate (cancer) tissue with tailored computational analysis. This resulted in a methodologically quantitated catalog of 3927 sncdRNAs originating from 1637 unique sncRNAs and allowed us to follow for changes in their expression during malignant transformation and cancer progression.

To investigate possible effects of sample storage conditions we compared the sncRNA transcriptome of fresh-frozen tissue with its FFPE-stored counterpart. We saw large changes in the accumulation of sncdRNAs, particularly sdRNAs and tRFs, when we compared fresh-frozen with FFPE material. This was not the case for miRNAs were we observed only relative down-regulation, most possibly caused by the additional buildup of degradation products of mRNAs and long ncRNAs due to sample preparation and storage [175]. In addition, while miRNA read-length in FFPE tissue remained unchanged, reads from other sncRNAs had changed length distribution indicating that FFPE-preserved tissue is less suitable for the analysis of sncRNAs other than miRNA.

Previously we detected differential expression of sdRNAs between organ-confined PCa and lymph node metastases [151]. The expression analysis presented here indicates that the major accumulation of sdRNAs is associated with malignant transformation and can be described by an increased global production and/or accumulation of sdRNAs already in the early cancer stages but it is not directly associated with the expression levels of precursor snoRNAs. Biological replicate analysis among three libraries (PCa, Gleason score 6) confirms the reproducibility of sequencing experiments

on fresh-frozen tissue as less than 20 sncRNAs show significantly changed expression levels. We did not observe a direct association between the number of sncdRNAs arising from one precursor and its quantity, suggesting that sncdRNA accumulation is not the direct result of increased sncRNA turnover in malignant cells. Q-PCR analysis confirmed the expression changes detected by sequencing and also identified the simultaneous existence of full-length snoRNAs and their derivate sdRNAs from the GAS5 locus. The high levels of SNORD78 and sd78-3′ in a subset of patients, which progressed to metastatic disease, identify these two sncRNAs as possible novel prognostic biomarkers for the further stratification of PCa patients at high risk of developing aggressive disease.

It has been shown that the majority of C/D-sdRNAs are derived from the termini of their precursor and may remain attached to the core snoRNP shielded from further degradation [170]. A large part of the sdRNAs detected in our libraries is also terminally derived. Nevertheless, the processing patterns of snoRNAs that we observe, and the accumulation of specific sdRNAs appear to be dependent on the conservation of structural snoRNA features and do not always correspond to snoRNA termini protected by the snoRNP. Furthermore, the overlap and discrete origin-position of multiple sdRNAs produced from the same precursor exemplified by sdRNAs produced from SNORD44, SNORD78, SNORD74 and SNORD81 suggest rather specific nucleolytic cleavage that requires different conformational states for C/D-box snoRNAs [182] possibly assisted by structural interaction with the core snoRNPs or yet unidentified proteins. Of note, the highly abundant sd78-3′ is derived from the opposite part of SNORD78 and does not overlap with the previously reported snoRNP footprint of SNORD78 [170]. It has been proposed that the specificity of sdRNA processing patterns detected in human cell lines is conserved between different cell types while the accumulation of individual sdRNAs is cell type specific implying the existence of dedicated processing mechanisms [167, 138, 128, 183].

It remains to be established how sdRNAs and other sncdRNAs are produced in the cell and to what extend this process is deregulated in cancer. The miRNA processing RNaseIII, DICER was suggested in the biogenesis of H/ACA-box-originating sdRNAs that have an apparent size of 20-24 nt. However, C/D box-sdRNAs identified by us and others [167, 138] have a bimodal size distribution which deviates from that of DICER products, suggesting the involvement of other nuclease(s) in the generation of sdRNAs. Another protein from the miRNA biogenesis pathway that could be involved in the generation of SNORD-sdRNAs is AGO2. It has been shown that AGO2 is responsible for the maturation of pre-miRNA-451 which is too short to undergo DICER processing. The AGO2 cleaved miR-451 product is a fragment of 30 nt that is processed further to the mature 23 nt long miR-451 by unknown exonucleases [184]. Nonetheless, AGO2-

derived mature miR-451 is predominantly uridilated at is 3′-end, while most of the C/D box-sdRNAs in our libraries are not. Furthermore, recent analyses of AGO2 PAR-CLIP libraries demonstrate that despite their cellular abundance, C/D box snoRNAs-originating sdRNAs are not efficiently incorporated in AGO2 [170].

The small transcriptome is a mix of turnover products and functional entities, where a proportion of the cellular sdRNA pool most probably represents stable degradation products shielded by effector proteins. Nevertheless, the mechanisms of sdRNA generation and their putative functional role in normal and malignant cells should be investigated further alongside with their biomarker potential in prostate and other cancers.

## 3.4 Materials and methods

### 3.4.1 Patient samples and cell lines

Snap-frozen, liquid nitrogen stored and FFPE clinical samples (Supplementary Table 1) were from the tissue bank of the Erasmus University Medical Center, Rotterdam, The Netherlands and from Tampere University Hospital (TAUH), Tampere, Finland. Collection and use of patient material was performed according to the national legislations concerning ethical requirements and approved by the Erasmus MC Medical Ethics Committee, Medical Research Involving Human Subjects Act (MEC-2004-261), and the Ethical Committee of the Tampere University Hospital. Prostate and lymph node tissues were from radical prostatectomy. BPH samples were obtained from cystoprostatectomies and found not to contain any prostate cancer cells. PCa-TURP samples were collected by transurethral resection of the prostate. Histological evaluation of analyzed material was described previously [151].

### 3.4.2 RNA isolation

Total RNA from frozen tissue was isolated using RNABee reagent (Campro Scientific, GmbH, Berlin, Germany) according to manufacturer's protocols. Total RNA isolation from FFPE material was described previously [177].

### 3.4.3 Sequencing

Total RNA sample pools of four individual patient samples each, were outsourced for sequencing (BGI, Shenzhen, China). Library preparations were performed according to the "Small RNA Sample Preparation Guide, Part #1004239", (Illumina Inc.).

Shortly, total RNA pools were separated on 15% Tris/Borate/EDTA urea polyacrylamide electrophoresis gel, and the sncRNA fraction in the size range of 15-35 nt was extracted and purified. After 5′- and 3′-adapter ligation, cDNA was generated by reverse transcription with SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) followed by 15 cycles of PCR by Phusion DNA Polymerase (Finnzymes Oy, Espoo, Finland). The strand-specific single-end reads were sequenced with a read-length of 35 bp.

### 3.4.4 Small non-coding RNA database (sncRNAdb)

Official small non-coding RNA nomenclature lists and NCBI RefSeq identifier numbers for microRNA precursors (pre-miRNAs), small nucleolar RNAs (snoRNAs), small cytoplasmic RNA (scRNAs), small nuclear RNA (snRNAs), and small miscellaneous RNAs (miscRNAs) were retrieved from the HUGO Gene Nomenclature Committee (HGNC) [185]. Genome locations corresponding to the RefSeq entries were further extended with 10 nt at the 5′- and 3′-end to ensure correct mapping of reads derived from ambiguously annotated ncRNAs and mapped against the hg19 assembly at the University of California Santa Cruz (UCSC) [186]. UCSC Genome Browser uses miRBase 15; therefore all miRNAs entries were manually curated to match miRBase 17. Since HGNC does not provide RefSeq identifiers for tRNAs, tRNA data was retrieved from the UCSC dedicated Genomic tRNA Database [187]. The number of mapped reads was positively influenced by the addition of "`CCA`" triplet to the 3′-end of genomic tRNA sequences and intron removal. Therefore, tRNA entries represent the mature tRNA form and are not extended. Sequences, genomic loci and database identifiers of all ncRNAdb entries are given in Supplementary File 1.

### 3.4.5 Computational analysis of sequencing data

Initial mapping of sequencing reads to sncRNAdb was done in CLC-Bio Genomics Workbench v4.9 following the "Small RNA Analysis" workflow. Read-summarizing and adapter-removal parameters from the "Extract and Count" tool were applied: Minimum sampling count was 4; Minimum and maximum number of nucleotides in reads was 15 and 35 nt, respectively; no 3′- or 5′- terminal nucleotide removal was performed. Each read was screened with "no fixed adapter length" for the (partial) presence of Illumina small RNA adapter: `CAAGCAGAAGACGGCATACGA` on the minus strand with alignment mismatches and gaps allowance at a mismatch cost of 3, and a gap cost of 5, minimum score: ns, minimum score end: 3. If adapter was not found reads were discarded from further analysis. Filtered sequence reads were mapped to sncRNAdb with a maximum of 2 mismatches allowed using the "Annotate and Merge"

tool.

### 3.4.6 Location, annotation and quantitation of sncdRNAs



Annotation of sncdRNAs was done using FlaiMapper as described [178]. Only sequence reads from libraries derived from fresh-frozen material were used as an input for the calculation of 5′- and 3′-ends of sncdRNAs. Quantitation and expression analysis of sncdRNAs was performed in a second round of mapping to FlaiMapper annotated sncRNAdb using 'Small RNA Analysis' workflow in CLC-Bio Genomics Workbench v4.9. "Expression values" that equal the sum of all reads mapping to a FlaiMapper annotated sncdRNA were used. Expression data was normalized with the "Reads per Million" algorithm. Differentially expressed sncdRNAs were detected using Kal's Z-test on proportions [188] with two-sided p-value, followed by Bonferroni correction with a corrected p-value cut-off of 0.01.

### 3.4.7 Quantitative real time PCR (qPCR)

snoRNA and sdRNA expression levels were evaluated by qPCR using miRCURY LNA$^{TM}$ Universal RT microRNA PCR, Polyadenylation and cDNA synthesis and SYBR Green kits (Exiqon, Copenhagen, Denmark) and custom LNA$^{TM}$ primers according to the manufacturer's instructions. Custom LNA primers for qPCR analysis of snoRNAs and sdRNAs were designed by Exoqon A/S, Copenhagen, Denmark. Target sequences used for primer design as well as design IDs are listed in Supplementary methods table M1. `SNORD38B` expression was measured with Reference gene primer set 20391 (Exiquon, Vedbaek, Denmark) and used to normalize raw Ct values by the delta delta Ct Method. `GAS5` expression was assessed by the Promega Reverse Transcription System (Promega Benelux, The Netherlands) and SybrGreen qPCR System (Roche, The Netherlands) according to manufacturer protocols. Primers used were `GAS5` FW: `CAAGGACTCAGAATTCATGAT` and `GAS5` REV: `AGTGGTCTTTGTAGACTGCC`. Raw expression values were normalized against the geometrical mean of `GAPDH` and `PBGD` by the delta delta Ct Method.

### 3.4.8 Statistical analysis

Significance of sncRNA composition and read-numbers were assessed with chi-square test for independence without Yates' correction. Two-sided p-values were calculated at alpha level of 0.05. Differences between groups in qPCR experiments were tested with unpaired two-tailed t-test at alpha level 0.05. Pearson correlation coefficients were assessed at an alpha level of 0.05 using GraphPad Prism 5.

## Acknowledgements

We thank GJ van Leenders and TH van der Kwast for the pathological examination of patient material.

## Funding

This work received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n$^o$201438 and NWO-ALW VENI grant 863.12.014.

# 4 | *FuMa*: reporting overlap in RNA-seq detected fusion genes

pmid: 26656567, doi: 10.1093/bioinformatics/btv721



Youri Hoogstrate [1,2], René Böttcher [1], Saskia Hiltemann[1,2], Peter van der Spek[2], Guido Jenster[1] and Andrew P Stubbs[2]

*Bioinformatics*, 32(8):1226−1228, Apr 2016

[1] *Department of Urology,* [2] *Department of Bioinformatics, Erasmus University Medical Center*

Supplementary material: Section 4.5

## Abstract

**Motivation:** A new generation of tools that identify fusion genes in RNA-seq data is limited in either sensitivity and or specificity. To allow further downstream analysis and to estimate performance, predicted fusion genes from different tools have to be compared. However, the transcriptomic context complicates genomic location-based matching. *FusionMatcher* (FuMa) is a program that reports identical fusion genes based on gene-name annotations. FuMa automatically compares and summarizes all combinations of two or more datasets in a single run, without additional programming necessary. FuMa uses one gene annotation, avoiding mismatches caused by tool specific gene annotations. FuMa matches 10% more fusion genes compared to exact gene matching (EGM) due to overlapping genes and accepts intermediate output files that allow a step wise analysis of corresponding tools.

**Availability and Implementation:** The code is available at:
`https://github.com/ErasmusMC-Bioinformatics/fuma`
and available for Galaxy in the tool sheds and directly accessible at:
`https://bioinf-galaxian.erasmusmc.nl/galaxy/`

## 4.1 Introduction

A new generation of bioinformatics tools has been released that aims to detect fusion genes within RNA-seq data; however, the current tools are limited in either sensitivity or specificity, making their results impractical for downstream analysis and subsequent validation [76]. As shown in other domains of high-throughput sequencing analysis, using a consensus of tools may improve performance by compensating for individual tool error profiles [189, 190]. Since no single tool shows superior detection performance, a consensus-based fusion gene detection in RNA-seq can improve both downstream analysis as well as overall performance. Moreover, to identify limiting factors and promote improvement of current algorithms, an accurate estimation of sensitivity and specificity is required, for which an accurate fusion gene comparison is crucial. Therefore, comparing validated and *in silico* predicted fusion genes in an automated fashion and summarizing identical fusion genes in two or more datasets in an easy and accessible way is a desirable feature.

Ideally, sensitivity/specificity estimation should be based upon the identified genomic breakpoints provided as two chromosomal locations. However, the nature of RNA-seq complicates this strategy, as reads may span exon junctions and breakpoints are more likely to be expected in introns because of their relative large size, introducing additional uncertainty when trying to pinpoint the exact genomic position of the DNA breakpoint. For instance, using exact position-based matching (EPM) results in poor overlap between tools (Supplementary Section 4.5.13). A less conservative strategy used in DNA-seq analysis involves comparing genomic intervals [81], which can be defined by adding flanking regions to each breakpoint to increase the likelihood of matching fusion events called by multiple programs. In RNA-seq analysis this strategy is also not sufficient, because the intervals are not related to the organization of the transcriptome. In addition, transcriptome annotations differ substantially between sources which frequently result in inconsistent results in RNA-seq analysis [191], intron sizes are differing and lastly, a substantial number of genes do overlap with each other [192] due to opposite strand positioning. Therefore, we designed a new method, FuMa, which boosts the current fusion gene matching functionality of the Chimera package [48], to address the challenges outlined above.

## 4.2 Methods

Here, we present FuMa, a computer program that reports identical fusion genes detected in RNA-seq, where matching is based on a user provided gene-name annotation. For two or more datasets, FuMa enlists all possible combinations of datasets

that can be compared with each other. The iterative procedure starts by comparing all 2-dataset combinations. Every such comparison results in a new virtual dataset, containing only the matching fusion genes. Consequently, for comparisons with a larger number of datasets, input datasets and merged datasets will be compared with each other such that all possible combinations are tested, by comparing only two virtual datasets at a time.

Because several factors complicate matching using genomic positions, our solution is based on gene-name comparisons. Each breakpoint of a fusion gene consists of two genomic locations. We define the genomic locations of a breakpoint as *left* and *right*, where *left* < *right*, while sorting is applied first on chromosome name and, in case of equality, on genomic position. When the left and right locations are identical to the lexicographical order, we denote the acceptor-donor order as *forward*, otherwise as *reverse*. For each fusion gene a list of genes overlapping each associated genomic location from the user provided gene annotation (BED format) is added with the HTSeq library [52]. This step ensures that all fusion genes are annotated with consistent genomic identifiers rather than those provided by the detection tools themselves. Since genes frequently overlap and multiple genes may be annotated upon one location, we add genes as a list to use them for set-theory-based matching rather than exact gene matching (EGM). FuMa has two matching methods, *subset*-based matching (FuMa-s, default) and *overlap*-based matching (FuMa-o) further explained in Supplementary Sections 4.5 and 4.5.2. Using FuMa-s, matching within any two datasets (both input-versus input- and input- versus merged/virtual dataset) is applied as follows:

- For each fusion gene in both datasets, remove entries that do not have gene annotations associated to both locations.
- Per dataset, merge duplicates such that a dataset contains only unique fusion genes. Two fusion genes are considered a duplicate by the *match*-function, which will later be explained as criterion used for matching fusion genes.
- Iterate over all fusion genes in both datasets such that all fusion genes of the first dataset are compared with all fusion genes of the second dataset. Assessing whether two fusion genes are identical is done by the *match*-function where two fusion genes are considered identical if: *one of the left gene lists is a subset of the other left gene list AND one of the right gene lists is a subset of the other* (Supplementary Table S4.1). Depending on the chosen parameters, the order of the genomic locations (forward or reverse) or the strands of the breakpoint may be taken into account as additional constraints. The comparison of any two virtual datasets will produce a 'merged' dataset that only includes matched fusion genes present in both input datasets. When two fusion genes match, the

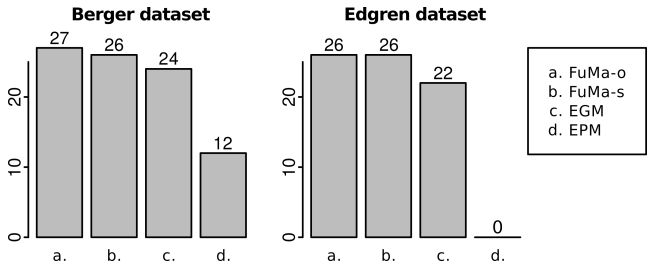



Figure 4.1: Differences between the matching approaches in the Berger (left) and Edgren (right) dataset. Each bar represents the number of fusion genes found in 2 or more samples. (Supplementary section S6). For this analysis a RefSeq gene annotation was used.

left and right gene sets of a matched fusion gene will be the intersect of the left or right gene sets of the input fusion genes. The intersect is chosen over the union to prevent gene lists from 'growing' after multiple iterations of matching (Supplementary Section 4.5.2).

## 4.3 Results

FuMa was tested on publicly available data [193, 194] and results provided as part of the Chimera package [48] (Supplementary Section 4.5.13). Concordance of the matching methods was assessed using RefSeq gene annotation. While EPM reported less than half of the overlaps of FuMa, EGM performed better but was still outmatched by FuMa. Specifically, EGM missed 11.1-15.4% of the fusion genes due to not accounting for overlapping genes (Figure 4.1) and importantly, five of the missed fusions had been validated. FuMa-o also reported a matching intergenic fusion event in a large gene that likely represents a false positive.

## 4.4 Discussion & Conclusion

Accurate comparisons of identical fusion genes between different algorithms are desirable to increase confidence in *in silico* predictions and to allow performance analyses as well as in-depth evaluations of the algorithms used. Therefore, we developed FuMa, a software package that makes use of a gene name and set-theory-based strategy, taking into account the transcriptome to reduce uncertainty and produce a human and computer understandable output (Supplementary Section 4.5.12). FuMa is publicly available, available for Galaxy [195] and available as R package compatible with Chimera [48]. FuMa focuses on comparing breakpoints within annotated genes and is more sensitive compared with EPM and EGM. In addition, FuMa can handle intermediate results of several detection tools and thereby allows an evaluation of the interim steps of an algorithm. Last, we find limited overlap between ChimeraScan, Defuse and

FusionMap [72, 71, 196] in the Edgren dataset (Supplementary Figure S4.5), which is in line with earlier reports [48] indicating that further improvements in detecting fusion genes in RNA-seq data are needed.

## Funding

This study was performed within the framework of the Center for Translational Molecular Medicine (CTMM), TraIT project (grant 05T-401).

## 4.5 Appendix

# Introduction

This is the manual of FuMa which is part of the Supplementary Material that belongs to the manuscript: *FuMa: reporting overlap in RNA-seq detected fusion genes*. FuMa (Fusion Matcher) matches predicted fusion genes (both genomic and transcriptomic) according to chromosomal location and corresponding annotated genes. The organisation of the transcriptome (provided by the user as BED file) forms the basis for FuMa to consider fusion genes to be identical or not. The provided gene annotation can be adjusted to define the biological question. For example, when it is desired to only consider fusion events that occur within exons, FuMa can be provided a list of exon regions instead of entire genes. Currently FuMa supports input files from:

- Chimera [48]
- ChimeraScan [72]
- CompleteGenomics [197]
- DeFuse [71]
- FusionCatcher [74]
- FusionMap [196]
- GMAP [198]
- STAR [199]
- STAR Fusion [200]
- TopHat-Fusion [70]

### 4.5.1 Technical implementation

Matching fusion genes based on the genomic location shows limited accuracy. Therefore it is more convenient to use the gene names overlapping the breakpoints instead. Since ∼10% of the annotated human genes are overlapping [192], and more genes and transcripts are being discovered by the RNA-seq technology, breakpoints frequently span multiple genes. This complicates matching based on gene names and to account for that, matching two fusion genes in FuMa is achieved using set-theory based matching (overlap or subset). First, both genomic partners of a fusion event are annotated

with overlapping gene(s). The overlap- and subset matching approach have the advantage over the more stringent exact gene matching (EGM) approach that a certain level of overlapping genes are considered as acceptable. They behave quite similar but have features that require a more detailed explanation.



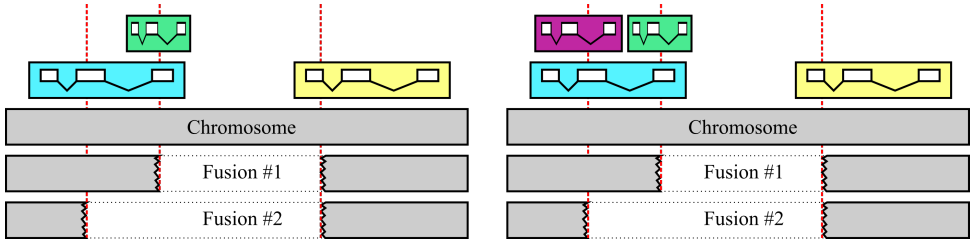



Figure S4.2: **Example of the subset matching methodology** Both scenario's (left and right) illustrate two predicted fusion genes. Both fusion genes have the same right location (red dashed line), located only in the yellow gene. Fusion #1 has two annotated genes on its left location: the green and the blue gene. In the right scenario, Fusion #2 is located in the blue and purple gene while in the left scenario it is only located within the blue gene. In the left scenario, the two fusion genes are considered identical because the left gene set of Fusion #2 (blue) is a subset of the left gene set of Fusion #1 (blue and green) and the right gene sets (yellow and yellow) too. In the right scenario, the left gene sets (purple, blue) and (green, blue) are no subsets of each other and the fusion genes are therefore considered to be distinct. The corresponding truth table of FuMa's subset based matching strategy is given in Table S4.1.

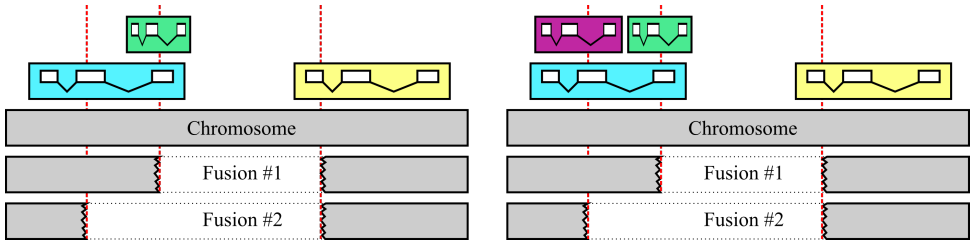



Figure S4.3: **Example of the overlap matching methodology** Both scenario's (left and right) illustrate two predicted fusion genes, Fusion #1 and Fusion #2. Both have the same right location (red dashed line through the yellow gene), located in one single gene annotation, the yellow gene. Fusion #1 has two annotated genes on its left location: the green and the blue gene. In the right scenario, Fusion #2 is located in the blue and purple gene while in the left scenario it is only located within the blue gene. In the left scenario, the two fusions are considered identical because the left gene set of Fusion #2 (*blue*) overlaps the left gene set of Fusion #1 (*blue*, *green*). Also in the right scenario, the left gene sets (*purple*, *blue*) and (*green*, *blue*) are overlapping and the fusion genes are therefore considered to be identical, but the set is reduced to (*blue*) since that's the part that overlaps.

### Subset-based matching

The subset matching approach, which is FuMA's default, considers two fusion genes identical if one of the left gene sets is a subset of the other left gene set, and one of the right gene sets is a subset of the other right gene set. Additional constrains with respect to acceptor-donor gene order and breakpoint strands are dependent on the chosen parameters. When two fusion genes match, for both the left and the right gene set the intersect (subset) will be returned as the gene sets of the matched fusion gene. The subset methodology is illustrated in Figure S4.2 and the corresponding truth table is further outlined in Table S4.1.

### Overlap-based matching

Overlap based matching considers two fusion genes identical if both the genes sets, the left and the right, have at least one overlapping gene in common with the other left and right gene set. We provide a more detailed description (Figure S4.3) and a corresponding truth table is given in Table S4.2. The overlap approach is less stringent than subset based matching and has a few noteworthy characteristics:

- **Long genes.** Long genes may span more other genes by chance. Therefore, two distant fusion genes that, by chance, also fall in the same long gene, may be matched only because they both overlap this same long gene. (See section *Example 1: long genes*)

- **Set shrinkage and expansion.** When two (input) fusion genes match, the matched fusion gene gets annotated genes from the gene sets of the two (input) fusion genes. For the overlap approach, two sets can be returned; the intersect (all genes that must be present in both fusion genes) or the union (all genes, that must be present in at least one of them). Using the union introduces a problem referred to as *set expansion*, which will result in an outcome that is dependent on the order of matching and on the iteration depth. This is very undesirable behaviour and therefore FuMa returns the intersect instead. But the intersect of two gene sets may result in a gene set smaller than the initial gene sets. We refer to this as *set shrinkage*. For example, if set (*green*, *blue*) is being matched with (*blue*, *red*), the set of overlapping genes will be (*blue*). This is different from the subset method, because there the smallest initial gene set is being returned, since that's the set shared by both fusion genes. Therefore the gene sets in the subset method will never become smaller than the smallest input gene set, while for the overlap based method the matched subset is not

| Fusion #1 | | Fusion #2 | | Returning Match | | |
|---|---|---|---|---|---|---|
| Left | Right | Left | Right | Match | Left | Right |
| `Blue` | `Yellow` | `Blue, Green` | `Yellow` | True | `Blue` | `Yellow` |
| `Blue, Purple` | `Yellow` | `Blue, Green` | `Yellow` | False | | |

Table S4.1: **Subset-based truth table** Truth table of FuMa's matching strategy using the examples from Figure S4.2. Depending on the genes spanning the breakpoints (first four columns), FuMa determines whether the fusion genes match (fifth column). The first four columns represent the gene sets (delimited with a comma) spanning the left and right locations. These gene names correspond to the colors used in Figure S4.2. The 5th column indicates whether FuMa considers the two fusions a match or not. The 6th and 7th columns represent the gene sets of the merged fusion gene as result of matching. The top example matches because (blue) is a subset of (blue, green). Although in the bottom example both fusion genes have their left location annotated within the blue gene, they are not considered a match because they are no subsets of each other; they are mutually exclusivily annotated within the purple and green gene.

| Fusion #1 | | Fusion #2 | | Returning Match | | |
|---|---|---|---|---|---|---|
| Left | Right | Left | Right | Match | Left | Right |
| `Blue` | `Yellow` | `Blue, Green` | `Yellow` | True | `Blue` | `Yellow` |
| `Blue, Purple` | `Yellow` | `Blue, Green` | `Yellow` | True | `Blue` | `Yellow` |

Table S4.2: **Overlap-based truth table** Depending on the genes spanning the breakpoints (first four columns), FuMa determines whether a fusion genes matches (fifth column). The first four columns represent the gene sets (delimited with a comma) spanning the left and right locations. These gene names correspond to the colors used in Figure S4.3. The 5th column indicates whether FuMa considers the two fusions a match or not. The 6th and 7th columns represent the gene sets of the merged fusion gene as result of matching Fusion #1 and #2. The top examples matches because (*blue*) overlaps (*blue*, *green*). In contrast to the subset method, the bottom example does match because (*blue*, *purple*) and (*blue*, *green*) have *blue* in common.

necessarily equal to any set observed at the breakpoints. More details are given in section *Example 2: set expansion and shrinkage.*

### 4.5.2 Exact gene set matching (EGM)

EGM consider fusion genes to be identical if their left and right gene sets are exactly identical. This is the most stringent matching scheme implemented in FuMa. EGM considers two gene sets identical if both the left and right gene sets contain exactly the same gene names as the other left and right gene sets.

# Differences between matching types

The matching schemes have different noteworthy characteristics outlined in the following sections.

## 4.5.3 Example 1: long genes

```
    b1                      b2
    |                       |
[ gene-A ]              [ gene-B ]
[---------- long gene ----------]
```

The example above illustrates two breakpoints *b1* and *b2*. Assume they are both part of a fusion gene of which the other genomic locations are identical. Breakpoint *b1* falls within *gene-A* and *long gene* and breakpoint *b2* falls within *long gene* and *gene-B*. When we match these breakpoints with the overlap approach, the breakpoints will be considered to be identical, since they have *long gene* in common. The longer *long gene* is, the more other genes it will span. Therefore, any fusion gene annotated within *long gene* will in the overlap based matching be considered a match with any other fusion gene annotated within *long gene*, even if one breakpoint is overlapping *gene-A* and the other *gene-B*. When subset matching is being used, the breakpoints in the example would not have been considered a match, since (*gene-A*, *long gene*) is not a subset of (*gene-B*, *long gene*). In the unit tests we have confirmed this behaviour and in the analysis Section 4.5.13, a long gene artefact was observed only using the overlap method.

Although the overlap based matching is more sensitive to long genes than subset and EGM based matching, each of these gene name based matching methods are sensitive to long genes because they may overlap more other genes. Therefore it is recommended to treat results containing large genes carefully.

## 4.5.4 Example 2: set expansion and shrinkage

When overlap based matching is used and considers two fusion genes a match, a consensus left- and right gene set is returned for the merged fusion gene. When the intersect is being returned, the merged fusion gene may be affected by shrinkage, while the union may be affected by set growth.

### Set shrinkage

Set shrinkage occurs when the returning gene set is the intersect of the two sets and the input sets are no subsets of each other. Consider two example breakpoints that have the following gene sets:

```
Breakpoint1: GeneA, GeneB, GeneC
                      |      |
Breakpoint2:        GeneB, GeneC, GeneD, GeneE
```

The breakpoints are considered to be a match because *GeneB* and *GeneC* are found in both. The intersect of the gene sets of *Breakpoint1* and *Breakpoint2* is thus (*GeneB*, *GeneC*). Hence, genes *GeneA*, *GeneD* and *GeneE* are not part of the annotation of the merged fusion. When we continue matching with e.g. *Breakpoint3*, we see only one shared gene:

```
Breakpoint1,2*:     GeneB, GeneC
                      |
Breakpoint3:        GeneB,        GeneD
```

Both fusion genes have only *GeneB* in common, and the merged fusion gene will thus only contain *GeneB*. So *GeneC* is taken away from the merged fusion gene, although it was present in *Breakpoint1* and *Breakpoint2*. *GeneB* is the only gene shared in all three breakpoints, although it may be important to know that *GeneC* was shared by the other two breakpoints. This information is lost because of the nature of the overlap matching approach in combination with returning the intersect. We refer to this as the set shrinkage issue. Returning the intersect is the implemented method for overlap based matching. When the subset approach was used instead, *Breakpoint3* would not have been considered a match with merged breakpoint *Breakpoint1,2∗*.

### Set expansion

– *This section illustrates a methodology that is* ***not implemented*** *in FuMa* – When a merged fusion gene would contain the union of the gene sets, the problem referred to as set expansion could be encountered. This will introduce order dependent results and matching may become less stringent after each iteration. To illustrate the problem of set expansion, imagine the following breakpoints:

1. b1 = `(A,A')`
2. b2 = `(A,A'')`
3. b3 = `(A'',B)`

Such situation would look similar to this:

```
       b1       b2        b3
       |        |         |
[---A'---]      |         |
     [-----A-----]       |
             [-----A''-----]
                      [---B---]
```

We denote the following possible orders of matching:

1. `(b1 & b2) & b3`
2. `(b1 & b3) & b2`
3. `(b2 & b3) & b1`

When matching is applied in **order 1**, the following is observed:

1. Iteration 1:

   - `(A,A') & (A,A'') -> (b1 & b2) = (A,A',A'')*`

     The gene sets match and the merged set contains 3 genes

2. Iteration 2:

   - `(A,A',A'')* & (A'',B) -> (b1 & b2 & b3) = (A,A',A'',B)`

     The gene sets match and the merged set contains all 4 genes

When matching is applied in **order 2**, the following is observed:

1. Iteration 1:

   - `(A,A') & (A'',B) -> (b1 & b3) =`

     no match; *b1* and *b3* are not considered identical

When matching is applied in **order 3**, the following is observed:

1. Iteration 1:

   ```
   - (A,A'') & (A'',B) -> (b2 & b3) = (A,A'',B)*
   ```

   The gene sets match and the merged set contains 3 genes

2. Iteration 2:

   ```
   - (A,A'',B)* & (A,A') -> (b1 & b2 & b3) = (A,A',A'',B)
   ```

   The gene sets match and the merged set contains all 4 genes

This shows that *b1* and *b3* are considered identical in *order 1* and *order 3*, but not in *order 2*. It also shows that the gene sets have become larger than the initial gene sets. Before matching, all gene sets had a size of 2 genes, after the first iteration 3 genes and after the second iteration the size of the genes sets have become 4 genes. Therefore, the merged breakpoint can be matched with more other breakpoints than each of the input fusion genes individually, because it will match if a breakpoint is annotated upon any of these 4 genes. Hence, using the union as merged gene set is not convenient.

# Installation

### 4.5.5 Debian, Ubuntu and derivatives

FuMa requires Python 2.7, depends on HTSeq and can be obtained via git. We recommand the following commands to install FuMa (on Ubuntu and Debian derivate systems):

```
sudo apt-get install build-essential python-dev git python-pip
sudo pip uninstall fuma

git clone https://github.com/yhoogstrate/fuma.git

cd fuma

python setup.py build
python setup.py test
sudo python setup.py install

fuma --version
```



### 4.5.6 Galaxy

Because usage of FuMa via the command line can be experienced as complicated, we also provide FuMa as Galaxy [195, 201, 93] tool. The toolshed repository in which FuMa is available is:

https://toolshed.g2.bx.psu.edu/view/yhoogstrate/fuma

To install FuMa via Galaxy, the user has to make sure to have the main tool shed (https://toolshed.g2.bx.psu.edu/) configured in galaxy's `tool_sheds_conf.xml`. To install FuMa within galaxy, follow the procedure via the galaxy admin panel. FuMa in galaxy has been made publicly available at the following galaxy instance:

https://bioinf-galaxian.erasmusmc.nl/galaxy/

The data used in the analysis is available as shared data library at the following url:

https://bioinf-galaxian.erasmusmc.nl/galaxy/library/list#folders/F313c46a90355d6dd

### 4.5.7 R

For compatibility with the Chimera package, FuMa has been embedded in R. This allows an R user to use the Chimera data structures and tests overlap using FuMa. The R package is available under the name FuMaR and at the following URL:

https://github.com/yhoogstrate/FuMaR

The Chimera R package is a prerequisite to get FuMaR working:

http://bioconductor.org/packages/release/bioc/html/chimera.html

The FuMaR package can be installed as follows:

```
git clone https://github.com/yhoogstrate/FuMaR.git FuMaR
R CMD INSTALL FuMaR
```

### 4.5.8 Usage

### 4.5.9 Command line

To run FuMa via the command line, each input dataset should be given as a separate file. The corresponding gene annotation has to be linked to each dataset and the file format has to be linked. This is a rather complex information structure and therefore, unfortunately, the command line arguments may be experienced as complicated. The command line usage of FuMa is:

```
usage: fuma [-h] [-V] [--formats] [-m {overlap,subset,egm}]
            [--strand-specific-matching] [--no-strand-specific-matching]
            [--acceptor-donor-order-specific-matching]
            [--no-acceptor-donor-order-specific-matching] [--verbose]
            [-a [ADD_GENE_ANNOTATION [ADD_GENE_ANNOTATION ...]]] -s ADD_SAMPLE
            [ADD_SAMPLE ...]
            [-l [LINK_SAMPLE_TO_ANNOTATION [LINK_SAMPLE_TO_ANNOTATION ...]]]
            [-f {summary,list,extensive}] [-g LONG_GENE_SIZE] [-o OUTPUT]
optional arguments:
  -h, --help            show this help message and exit
  -V, --version         show program's version number and exit
  --formats             show accepted dataset formats
  -m {overlap,subset,egm}, --matching-method {overlap,subset,egm}
                        The used method to match two gene sets. Overlap
                        matches when two gene set have one or more genes
                        overlapping. Subset matches when one gene set is a
                        subset of the other. EGM is exact gene matching; all
                        genes in both sets need to be identical to match.
  --strand-specific-matching
                        Consider fusion genes distinct when the breakpoints
                        have different strands: (A<-,B<-) != (->A,B<-);
                        default
  --no-strand-specific-matching
                        Consider fusion genes identical when the breakpoints
                        have different strands: (A<-,B<-) == (->A,B<-)
```

```
--acceptor-donor-order-specific-matching
                      Consider fusion genes distinct when the donor and
                      acceptor sites are swapped: (A,B) != (B,A)
--no-acceptor-donor-order-specific-matching
                      Consider fusion genes identical when the donor and
                      acceptor sites are swapped: (A,B) == (B,A); default
--verbose             increase output verbosity
-a [ADD_GENE_ANNOTATION [ADD_GENE_ANNOTATION ...]],
--add-gene-annotation [ADD_GENE_ANNOTATION[ADD_GENE_ANNOTATION ...]]
                      annotation_alias:filename * file in BED format
-s ADD_SAMPLE [ADD_SAMPLE ...], --add-sample ADD_SAMPLE [ADD_SAMPLE ...]
                      sample_alias:format:filename (available formats: fuma
                      --formats)
-l [LINK_SAMPLE_TO_ANNOTATION [LINK_SAMPLE_TO_ANNOTATION ...]],
--link-sample-to-annotation [LINK_SAMPLE_TO_ANNOTATION [LINK_SAMPLE_TO_ANNOTATION ...]]
                      sample_alias:annotation_alias
-f {summary,list,extensive}, --format {summary,list,extensive}
                      Output-format
-g LONG_GENE_SIZE, --long-gene-size LONG_GENE_SIZE
                      Gene-name based matching is more sensitive to long
                      genes. This is the gene size used to mark fusion genes
                      spanning a 'long gene' as reported the output. Use 0
                      to disable this feature.
-o OUTPUT, --output OUTPUT
                      output filename; '-' for stdout
```

### -a ADD_GENE_ANNOTATION

The first command line argument we describe is `-a`. Gene annotations can be provided as a tab-delimited file, with the first column containing the genes' chromosome, the second and the third column the (1-based) start and end position, and the fourth column the (unique) gene identifier or name, as shown in the example below:

```
chr1   100000000  120000000  GeneNameA
chr2   100000000  120000000  GeneNameB
chr21  100000000  120000000  GeneNameC
chr22  100000000  120000000  GeneNameD
chrX   140000000  160000000  GeneNameX
chrY   140000000  160000000  GeneNameY
```

This format is compatible with the BED format[1], but requires that the 4th column is present and requires it to contain gene names. Additional columns are allowed, but are not taken into account. **Do not provide BED files that describe one exon**

[1] https://genome.ucsc.edu/FAQ/FAQformat.html#format1

**per line** because this will exclude the introns, but provide BED files that describe one gene per line instead.

The reason it is not feasible to use BED files with one exon per line and merge these exons automatically into single gene entries: merging exons that belong to the same gene name into single gene entires comes with the assumption that there are no duplicates of a gene, which is not necessarily true. If FuMa would merge based on gene names, duplicates on the same chromosome (that span long distances) effectively create very long genes. If a user only wants to match in exon regions, you can use BED files with one exon per line. In that case is advised to provide non-unique gene names, like the following example:

```
chr1  100000000  100001000  GeneNameA
chr1  100002000  100003000  GeneNameA
chr1  100005000  100006000  GeneNameA
chr2  100000000  100100000  GeneNameB
chr2  100101000  100103000  GeneNameB
```



The gene annotation argument is provided as unique alias followed by the filename, separated with a colon:

```
-a "hg19:somefile.bed"
```

In this case the alias of the BED-file, hg19, will later be used to link it to datasets. In case the user wants to use multiple references, for example to match across different genome builds, multiple arguments can be provided delimited with white-spaces:

```
-a "hg18:somefile_hg18.bed" "hg19:somefile_hg19.bed"
```

### -s ADD_SAMPLE

The `-s` argument can be used to provide a fusion gene detection experiment, which should have the following syntax:

```
sample_alias:format:filename
```

The `sample_alias` will be used for two things: *(1)* as alias (column header) in the final output and *(2)* to link the references to the samples. The `format` is the file format in which the fusion genes are described. The formats are described in more detail in Section 4.5.9. Some tools have multiple formats, often the interim results.

### -l LINK_SAMPLE_TO_ANNOTATION

Each dataset must be annotated with only one gene annotation. This can be achieved using the following argument syntax:

```
sample_alias:annotation_alias
```

When you have sample *s* and a reference named *ref*, you can link *s* to *ref* as follows:

```
-l "s:ref"
```

In case you have two samples, one on hg18 and one on hg19, you can provide it as follows:

```
-l "defuse_hg18:hg18" "chimerascan_hg19:hg19"
```

### -l MATCHING_METHOD

FuMa has the option to use three methods to match fusion genes; '*overlap*', '*subset*' and '*egm*'. The method can be selected with the `-m` or `--matching-method`, argument as follows:

```
fuma -m egm [ ... ]
```

```
fuma --matching-method subset [ ... ]
```

### -f OUTPUT_FORMAT

FuMa has the built-in option for several output formats. The default output format is '*list*' which contains, per unique fusion gene, matched or not matched, for each matching tool, the genomic locations and identifier(s) or an empty column if a tool did not pick it up. There is an additional column to indicate whether one of the annotated genes are large. The following example shows three fusion genes; one detected by TopHat fusion, one by STAR and one by both. The corresponding output in '*list*' format would look like:

| Left Genes | Right Genes | Involves large gene(s) | STAR | TopHat Fusion |
|---|---|---|---|---|
| FOO1 | BAR1 | FALSE | UID_A=chr1:12-34 | |
| FOO2 | BAR2 | FALSE | | TID_A=chr4:66-77 |
| DOX1 | BOX5 | FALSE | UID_B=chr5:85-95 | TID_B=chr5:88-99 |

Tools may predict multiple fusion events within the same left- and right genes, which FuMa will consider as duplicates. In case a duplicate is observed, the output contains the identifiers of all duplicates into one cell delimited with a comma. This allows to trace duplicate entries back in the output. An example of a duplicate entry is given below:

| Left Genes | Right Genes | Involves large gene(s) | FusionMap |
|---|---|---|---|
| FOO1 | BAR1 | FALSE | UID_A=chr1:12-34,UID_B=chr1:12-34 |



When a genomic location is annotated with multiple genes, the genes are delimited with a colon in one cell. An example of a genomic location that spans genes *FOO1* and *FOO2* is given below:

| Left Genes | Right Genes | Involves large gene(s) | FusionMap |
|---|---|---|---|
| FOO1:FOO2 | BAR1 | FALSE | UID_A=chr1:12-34 |

The FuMa package contains an additional tool `fuma-list-to-boolean-list` to replace cells to TRUE or FALSE depending on whether a match was found or not.

The output format '`extensive`' is file format similar to the format Complete Genomics provides and that only contains the matched fusion genes. This format is in particular useful if the output of a run needs to be the input for another run.

The output format '`summary`' is a set of tables that contain the numbers of detected matches per combination of datasets, useful for creating Venn diagrams.

**-g LONG_GENE_SIZE**

Because gene name based matching may be affected by large genes spanning fusion genes by chance, the output of type 'list' has a column indicating whether any of the genes the fusion genes annotated upon the breakpoint are large. Whether a gene is large is defined as having a genomic size larger than the `-g` parameter. When this value is set to 0, genes will not be marked as long gene.

**–strand-specific-matching and –no-strand-specific-matching**

FuMa has the built-in option to match fusion genes by taking the strands of the breakpoints into account. In the following example fusion genes #1 and #2 with

exactly the same breakpoints are shown, but the strands of the second breakpoints are in the opposite direction:

```
#1:
        b1 (+) ->          <- (-) b2
        |                         |
[ --- Gene A --- ]        [ --- Gene B --- ]

#2:
        b1 (+) ->                 b2 (+) ->
        |                         |
[ --- Gene A --- ]        [ --- Gene B --- ]
```

To let FuMa consider them as distinct fusion genes because of the different strands, the user should enable strand specific matching using the `--strand-specific-matching` argument as indicated below:

```
fuma \
   --strand-specific-matching \
    -a  "hg19:genes_hg19.bed" \
    \
    -s  "chimerascan:chimerascan:FOO_chimerascan/chimeras.bedpe" \
        "defuse:defuse:FOO_defuse/results.tsv" \
    -l  "chimerascan:hg19" \
        "defuse:hg19" \
    -f  "list" \
    -o  "chimerascan_defuse_overlap.txt"
```

By default, this option is enabled. If the user wants to disable this feature, the `--strand -specific-matching` argument has to be provided.

### Acceptor/donor (a)specific matching

For most file formats, the order in which the acceptor and donor gene are denoted should correspond to the order in which they appear in the transcript. This information may be crucial to explain the function and biological role of a fusion gene. For example, `TMPRSS2-ERG`, a fusion gene found in about 50% of all screened prostate cancers, uses regulatory elements from the androgen driven gene `TMPRSS2`, fused to

gene ERG that has an oncogenic role in human prostate cancer [202]. These principles would not apply if the order the transcript would be vice versa. To account for this, FuMa has the built-in option to separate fusion genes based on the order of the denotation of the acceptor and donor. In the following example fusion genes #1 and #2 are shown with a different order of donor and acceptor gene:

```
#1:
        break1     ---------->     break2
        |                          |
[ --- Gene A --- ]        [ --- Gene B --- ]

#2:
        break1     ---------->     break2
        |                          |
[ --- Gene B --- ]        [ --- Gene A --- ]
```

To let FuMa consider them as distinct fusion genes because of the different order of the donor and acceptor gene, the user should enable acceptor donor specific matching by including the `-acceptor-donor-order-specific-matching` argument:

```
fuma \
   --acceptor-donor-order-specific-matching \
    -a  "hg19:genes_hg19.bed" \
    \
    -s  "chimerascan:chimerascan:FOO_chimerascan/chimeras.bedpe" \
        "defuse:defuse:FOO_defuse/results.tsv" \
    -l  "chimerascan:hg19" \
        "defuse:hg19" \
    -f  "list" \
    -o  "chimerascan_defuse_overlap.txt"
```

By default this option is disabled. **Some file formats (in particular interim output files and discordant reads) do not take this information into account** and for these file formats this functionality is disabled.

**Input formats**

FuMa supports the following file formats as input:

| Tools | File | Argument at command line |
|---|---|---|
| Chimera | prettyPrint() output | chimera |
| ChimeraScan | chimeras.bedpe | chimerascan |
| Complete Genomics | highConfidenceJu*.tsv | complete-genomics |
| Complete Genomics | allJunctionsBeta*.tsv | complete-genomics |
| DeFuse | results.txt | defuse |
| DeFuse | results.classify.txt | defuse |
| DeFuse | results.filtered.txt | defuse |
| Fusion Catcher | final-list_cand*.txt | fusion-catcher_final |
| FusionMap | | fusionmap |
| Trinity + GMAP | | trinity-gmap |
| OncoFuse | | oncofuse |
| RNA STAR | Chimeric.out.junction | rna-star_chimeric |
| STAR Fusion | _candidates.final | star-fusion_final |
| TopHat Fusion pre | fusions.out | tophat-fusion_pre |
| TopHat Fusion post | potential_fusion.txt | tophat-fusion_post_potential_fusion |
| TopHat Fusion post | result.txt | tophat-fusion_post_result |

To get an overview of the formats available in the installed instance the user can run the following command:

```
fuma --formats
```

### 4.5.10 Galaxy

In Galaxy, after having FuMa installed via the toolshed, it can be opened by typing '*fuma*' in the '*search tools*' field on the left panel. When it has opened, the interface should be similar to Figure S4.4. The main input of the Galaxy wrapper is a set of datasets. The user can as add many datasets as the server can handle in terms of resources. For each dataset the user needs to specify *(1)* the history item in galaxy that contains the output file of the fusion gene detection experiment, *(2)* the corresponding file format of the history item and *(3)* a corresponding gene annotation file (in BED format). Last, the user can select the following advanced settings:

- The desired output format.

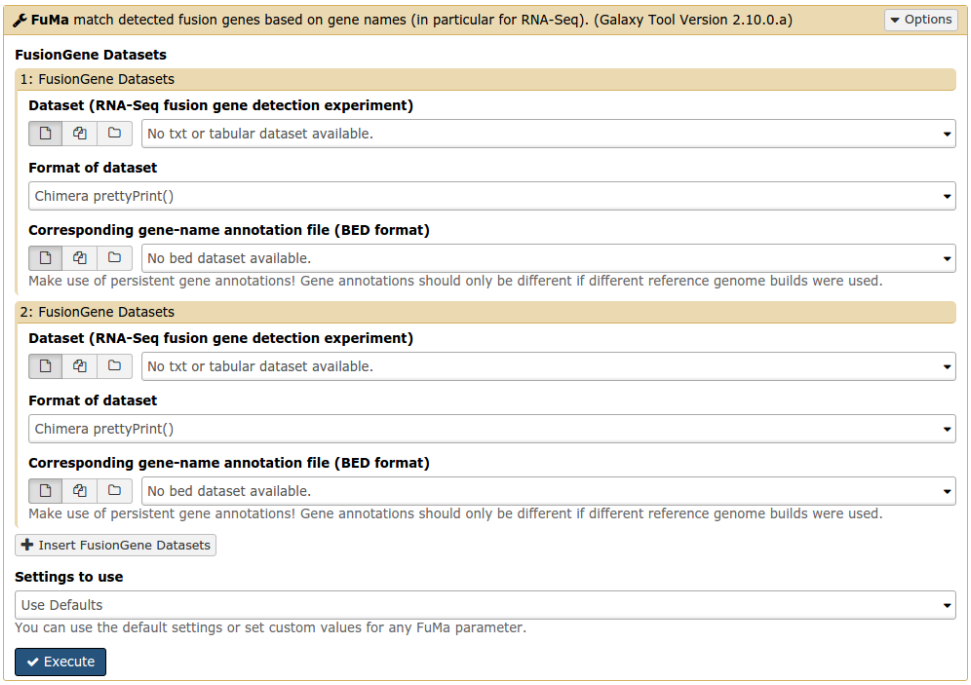



Figure S4.4: FuMa in Galaxy.

- For the 'list'-output, replace the columns to TRUE or FALSE depending on whether a match was found or not.
- Strand specific matching.
- Acceptor and donor order specific matching.

### 4.5.11 R

The R wrapper of FuMa, called FuMaR, depends on BioConductor package Chimera [48]. To run FuMa in R, the packages *FuMaR* and *chimera* should be loaded. To run FuMa, a FuMa object that contains the settings and allows to control FuMa as background process, should be created. Datasets have to be provided to the object as lists of *fSet* objects, the default data type of data importers in Chimera. These datasets can be added to the FuMa object using the *add_data(dataset, name)* function, which requires the data object and a unique name. When the data has been added, the *fuma(FuMa_object, BED_file)* command will export the datasets with Chimera's prettyPrint() to disk and FuMa will be executed in the background. An example with three datasets (DeFuse, ChimeraScan and FusionMap) using Chimera

is given below:

```
> library(FuMaR)
> library(chimera)
> df.e <- importFusionData("defuse",paste(find.package(package="chimera"),
"/examples/Edgren_df.tsv",sep=""))
> cs.e <- importFusionData("chimerascan",paste(find.package(package="chimera"),
"/examples/Edgren_cs.txt",sep=""),org="hs")
> fm.e <- importFusionData("fusionmap",paste(find.package(package="chimera"),
"/examples/Edgreen_fm.txt",sep=""),org="hs")
> f1 <- FuMa(matching="subset")
> f2 <- add_dataset(f1,df.e,"defuse")
> f3 <- add_dataset(f2,cs.e,"chimerascan")
> f4 <- add_dataset(f3,fm.e,"fusionmap")
> f  <- fuma(f4,"refseq_genes.bed")
```



### 4.5.12 Examples

**Example 01: one sample, two tools**

Imagine sample FOO was analysed with Defuse and ChimeraScan on the same reference genome (hg19). The corresponding gene annotation is `genes_hg19.bed` and the output should be stored in `chimerascan_defuse_overlap.txt`. The command line argument to run this analysis would be:

```
fuma \
    -a  "hg19:genes_hg19.bed" \
    -s  "chimerascan:chimerascan:FOO_chimerascan/chimeras.bedpe" \
        "defuse:defuse:FOO_defuse/results.tsv" \
    -l  "chimerascan:hg19" \
        "defuse:hg19" \
    -f  "list" \
    -o  "chimerascan_defuse_overlap.txt"
```

**Example 02: one sample, one tool, different reference genomes**

When it is desired to compare the differences between runs on different genome builds, the user can add each run and define a different gene annotation for each run. Imagine a sample with TopHat-Fusion on reference genomes hg18 and hg19, it can be analysed with FuMa as follows:

```
fuma \
    -a  "hg18:genes_hg18.bed" \
        "hg19:genes_hg19.bed" \
    -s  "thf_hg18:tophat-fusion_post_result:thf_hg18/result.txt" \
        "thf_hg19:tophat-fusion_post_result:thf_hg19/result.txt" \
    -l  "thf_hg18:hg18" \
        "thf_hg19:hg19" \
    -f  "list" \
    -o  "thf_hg18_hg19_overlap.txt"
```



It is important that the gene annotations `genes_hg18.bed` and `genes_hg19.bed` contain similar gene names, since matching is based on these names. Therefore, it is recommanded to remove gene names that are specific per annotation; the latest genes only available in hg19 shall not match with hg18 simply because they do not exist in hg18.

**Example 03: Edgren dataset**

The publicly available data from the Edgren dataset analysed by FusionMap, ChimeraScan and DeFuse was used in the manual of the Chimera package [194, 48]. To obtain these results the user should run the following command:

```
$ wget http://www.bioconductor.org/packages/release/bioc/src/contrib/
  chimera_1.10.0.tar.gz
$ tar -xzf chimera_1.10.0.tar.gz
$ find . -type f | grep -i -E "Edgr[e]{1,2}n"
```

Please check whether the output is similar to:

```
./chimera/inst/examples/Edgreen_fm.txt
./chimera/inst/examples/edgren.stat.detection.txt
./chimera/inst/examples/Edgren_df.tsv
./chimera/inst/examples/Edgren_cs.txt
./chimera/inst/examples/Edgren_true.positives.txt
```

To get a gene reference and the true positivies with genomic coordinates, run at the command line:

```
$ wget https://toolshed.g2.bx.psu.edu/repos/yhoogstrate/fuma/
  raw-file/tip/test-data/refseq_genes_hg19.bed
$ wget https://toolshed.g2.bx.psu.edu/repos/yhoogstrate/fuma/
  raw-file/tip/test-data/edgren_true_positives.txt
```

To proceed with the FuMa analysis, run:

```
edir="./chimera/inst/examples/"
fuma \
   --strand-specific-matching \
   --acceptor-donor-order-specific-matching \
    -m  "subset" \
    -a  "hg19:refseq_genes_hg19.bed" \
    -s  "chimerascan:chimerascan:"\$edir"Edgren_cs.txt" \
        "defuse:defuse:"$edir"Edgren_df.tsv" \
        "fusionmap:fusionmap:"$edir"Edgreen_fm.txt" \
        "edgren_TP:fusionmap:edgren_tp.txt" \
    -l  "fusionmap:hg19" \
        "defuse:hg19" \
        "chimerascan:hg19" \
        "edgren_TP:hg19" \
    -f  "list" \
    -o  "edgren_fuma_list_specific.txt"
```

To convert the columns to boolean values, proceed with:

```
fuma-list-to-boolean-list \
-o "edgren_fuma_booleanlist_specific.txt" \
   "edgren_fuma_list_specific.txt"
```

To find all fusion genes that did match, thus present in 2 or more datasets, proceed with the command:

```
grep -E "TRUE.*?TRUE" edgren_fuma_booleanlist_specific.txt
```

The output will be a list of 26 fusion genes (Table S4.3).

Table S4.3: **FuMa subset-based matching results on the Edgren dataset)** The results of matching the Edgren dataset using a RefSeq gene annotation with *strand specific* and *donor acceptor order specific* matching. The first two columns represent the genes spanning the breakpoints followed by a column indicating whether any of these genes is large (>200,000 bp). The following columns represent the datasets of which *Edgren TP* contains the wet-lab validated fusion genes of the Edgren dataset. Afterwards, the columns have been replaced with TRUE or FALSE using *fuma-list-to-boolean-list*. The table only contains the fusion genes that did match (Figure S4.5, left). For gene sets that were too large, the list was truncated and indicated with an asterisk.

| Left-genes | Right-genes | Large genes | Chimerascan | Defuse | FusionMap | Edgren TP |
|---|---|---|---|---|---|---|
| NOS1AP | C1orf226 | TRUE | TRUE | FALSE | TRUE | FALSE |
| INIP | C5 | FALSE | TRUE | FALSE | TRUE | FALSE |
| FAM208B | FAM208B | FALSE | TRUE | FALSE | TRUE | FALSE |
| MED1 | ACSF2:NM_001288968* | FALSE | TRUE | FALSE | TRUE | FALSE |
| PPP3R1 | VPS35 | FALSE | TRUE | FALSE | TRUE | FALSE |
| PPP1R12A | SEPT10 | FALSE | TRUE | FALSE | FALSE | TRUE |
| *NM_001281784:ZMYND8 | CEP250 | FALSE | TRUE | FALSE | FALSE | TRUE |
| ARFGEF2 | SULF2 | FALSE | TRUE | FALSE | FALSE | TRUE |
| DIDO1 | NM_001303457:TTI1 | FALSE | TRUE | FALSE | FALSE | TRUE |
| DHX35 | ITCH | FALSE | TRUE | FALSE | FALSE | TRUE |
| CPNE1:NM_001198863:NR_037188 | PI3 | FALSE | TRUE | FALSE | FALSE | TRUE |
| CCDC85C | SETD3 | FALSE | TRUE | FALSE | FALSE | TRUE |
| CYTH1 | EIF3H | FALSE | TRUE | FALSE | FALSE | TRUE |
| RARA | PKIA | FALSE | TRUE | FALSE | FALSE | TRUE |
| BCAS4 | BCAS3 | TRUE | TRUE | FALSE | FALSE | TRUE |
| NR_036633:VAPB | IKZF3:NM_001284514* | FALSE | TRUE | FALSE | FALSE | TRUE |
| NM_001272060:RPS6KB1 | SNF8 | FALSE | TRUE | FALSE | FALSE | TRUE |
| SKA2 | MYO19 | FALSE | TRUE | FALSE | FALSE | TRUE |
| BSG | NFIX | FALSE | TRUE | FALSE | FALSE | TRUE |
| RAB22A | MYO9B | FALSE | TRUE | FALSE | FALSE | TRUE |
| GLB1 | CMTM7 | FALSE | TRUE | FALSE | FALSE | TRUE |
| NOTCH1 | NUP214 | FALSE | TRUE | FALSE | FALSE | TRUE |
| WDR67 | ZNF704 | TRUE | TRUE | FALSE | FALSE | TRUE |
| ANKHD1:ANKHD1-EIF4EBP3 | PCDH1 | FALSE | FALSE | FALSE | TRUE | TRUE |
| TATDN1 | GSDMB | FALSE | TRUE | FALSE | TRUE | TRUE |
| ACACA | STAC2 | TRUE | TRUE | FALSE | TRUE | TRUE |



98

FusionMatcher

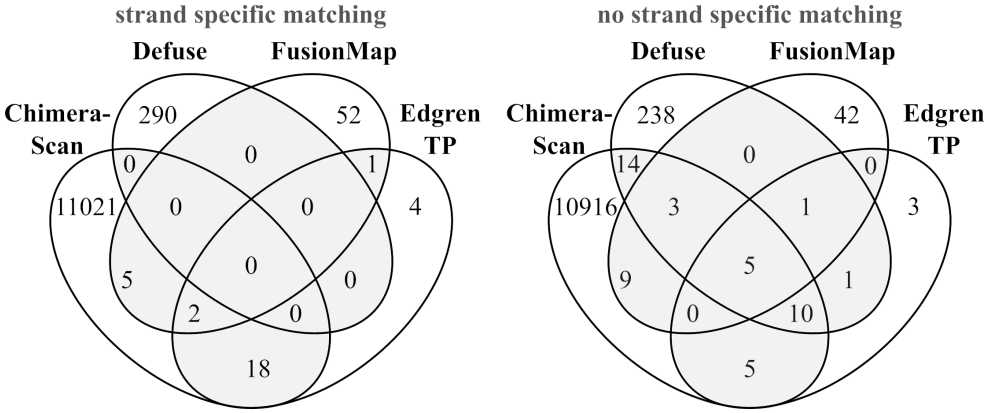



Figure S4.5: **Venn diagram of subset based matching on the Edgren dataset** The matched fusion genes found by FuMa on the Edgren dataset are given in a Venn diagram. In each diagram, *Edgren TP* are the true positives from the Edgren dataset. The gray areas represent the fusion genes that match. *Left*: FuMa with strand specific mode enabled. *Right*: FuMa with strand specific mode disabled. The overlap for all datasets is limited.



### 4.5.13 Analysis

To highlight differences between subset and overlap based matching compared to EGM and EPM, we ran FuMa on the Edgren and Berger datasets [194, 193].

The Edgren dataset has 27 validated fusion genes fusion genes and the Berger dataset contains 25 validated fusion genes, with genomic locations reported over several papers [73, 203]. In a previous study, the Edgren dataset was analysed with ChimeraScan, DeFuse and FusionMap and the results have been made publicly available [48]. Together with a list of corresponding validated fusion genes, these data were analysed with FuMa using a RefSeq gene annotation (Table S4.3). This was repeated with gene annotations from UCSC and Ensembl to indicate potential differences due to different gene annotations. For this analyis, FuMa was set with `--strand-specific-matching` and `--acceptor-donor-order-specific-matching`.

The Berger dataset was analysed with several tools (ChimeraScan v0.4.5, DeFuse 0.6.2, Fusioncatcher v0.99.3e, STAR-2.4.0g1 followed by STAR-Fusion and Tophat-Fusion v2.0.9). For the majority of the validated fusion genes, the corresponding strands were not reported. Therefore, the analysis on this dataset could not be done in strand-specific mode.

For both datasets, only a limited number of predicted fusion genes have been wet-lab validated. Hence, any other predicted fusion gene can be a true or false positive, which makes it difficult to estimate sensitivity and specificity [76]. To estimate potential false positives, additional sample SRR064437 from the Edgren dataset was

| Names | | | unique fusion genes | | | matches | | | Difference in matches | |
|---|---|---|---|---|---|---|---|---|---|---|
| dataset | geneset | samples | $E$ | $S$ | $O$ | $E$ | $S$ | $O$ | $O-E$ | $O-S$ |
| Edgren | refseq | 4 | 11436 | 11393 | 11374 | 22 | 26* | 26 | 4 | 0 |
| Edgren | ucsc | 4 | 13549 | 13508 | 13421 | 25 | 28 | 28 | 3 | 0 |
| Edgren | ensembl | 4 | 11919 | 11897 | 11844 | 20 | 24 | 25 | 5 | 1 |
| SRR064437 (control) | refseq | 4 | 43 | 41 | 41 | 0 | 0 | 0 | 0 | 0 |

Table S4.4: Summary of the results using strand specific and donor acceptor specific matching mode. Abbrevations: $E$=egm, $S$=subset and $O$=overlap. The first column contains the names of the datasets, followed by the gene annotation in the 2nd column. Column 3 contains the number of samples. For Edgren these are ChimeraScan, Defuse, FusionMap and the true positives from the chimera package and for the control sample these are results from Chimerascan, DeFuse, FusionCatcher and Tophat Fusion. Columns 4, 5 and 6 contain the unique fusion genes with annotated genes, and columns 7, 8 and 9 the number of fusion genes found in 2 or more datasets. The 10th column contains the number of fusion genes matched with the overlap method, but missed by EGM. The 10th column contains the matches found by the overlap method, but missed by the subset method. These results are visualised in Figure S4.5 (left).



analysed. This is a sample from healthy breast tissue, that should not contain any fusion gene. Therefore, it should also not match fusion genes in the subsequent results of the prediction tools. The results of the FuMa analyses were summarized (Tables S4.4 and S4.5) and compared with each other to indicate which matched fusions were missed by EGM (Table S4.7), indicating that:

- In the Edgren dataset, 26 fusion genes were considered identical in 2 or more datasets using a RefSeq gene annotation and the overlap approach (Figure S4.5), including 21 of 25 true positives with annotated genes (of which `CSE1L-ENSG00000236127` and `SUMF1-LRRFIP2` did not have annotated genes on both breakpoints). Of the 26 fusion genes matched in 2 or more datasets, EGM missed 4 (Table S4.7), indicating that 15.3% of the fusion genes in this dataset were not matched by EGM due to overlapping genes. Each of these 4 fusion genes were also matched by the subset method but more importantly each of them, `CPNE1-PI3`, `VAPB-IKZF3`, `SKA2-MYO19` and `ANKHD1-PCDH1`, are validated fusion genes.
    - When an Ensembl gene annotation was used, only the overlap method matches an additional true positive: `BCAS4-BCAS3`. This fusion gene was annotated as `BCAS4-BCAS3:RP11-264B14.1` and the ChimeraScan prediction was annotated as `BCAS4-BCAS3:RP11-332H18.5`.
    - Exact position based matching revealed only 6 matches: `SULF2-ARFGEF2`, `ANKHD1-PCDH1`, `GSDMB-TATDN1`, `RARA-PKIA`, `ACACA-STAC2`, `MYO19-SKA2`. These matches were only between DeFuse and FusionMap with strand and

| Names | | | Unique fusion genes | | | matches | | | Difference in matches | |
|---|---|---|---|---|---|---|---|---|---|---|
| dataset | geneset | samples | egm | subset | overlap | egm | subset | overlap | overlap−egm | overlap−subset |
| Edgren | refseq | 4 | 11317 | 11247 | 11229 | 47 | 48* | 48 | 4 | 0 |
| Edgren | ucsc | 4 | 13406 | 13304 | 13190 | 49 | 51 | 51 | 3 | 0 |
| Edgren | ensembl | 4 | 11837 | 11807 | 11740 | 44 | 45 | 45 | 5 | 1 |
| Berger (total) | refseq | 6 | 832 | 817 | 803 | 24 | 26 | 27 | 4 | 1 |
| SRR018259 M0000216 | refseq | 6 | 85 | 85 | 84 | 1 | 1 | 1 | 0 | 0 |
| SRR018260 M990802 | refseq | 6 | 99 | 98 | 96 | 2 | 2 | 2 | 0 | 0 |
| SRR018261 M980409 | refseq | 6 | 109 | 107 | 106 | 4 | 5 | 5 | 1 | 0 |
| SRR018265 M010403 | refseq | 6 | 57 | 57 | 55 | 1 | 1 | 1 | 0 | 0 |
| SRR018266 501-MEL | refseq | 6 | 137 | 136 | 135 | 7 | 7 | 7 | 1 | 0 |
| SRR018267 M000921 | refseq | 6 | 148 | 140 | 136 | 3 | 3 | 3 | 0 | 0 |
| SRR018269 K-562-3-CML | refseq | 6 | 197 | 194 | 191 | 6 | 7 | 8 | 2 | 1 |
| SRR064437 (control) | refseq | 4 | 42 | 39 | 39 | 0 | 0 | 0 | 0 | 0 |

Table S4.5: **Summary of the results using no additional constraints** The first column contains the names of the datasets, the second the used gene annotation and the third column the number of samples. Column 4, 5 and 6 contain the unique fusion genes and 7, 8 and 9 the number of fusion genes found in 2 or more datasets. The 10th column contains the number of fusion genes matched with the overlap method, but missed by EGM. The 10th column contains the matches found by the overlap method, but missed by the subset method. The reason why 501-MEL has an identical number of matches (7) in each matching scheme whilst it contains differences between EGM and overlap, is because `SHANK2-GNA12` was found in more than 2 datasets. This analysis is visualised in Figure S4.5 (right).

acceptor donor specific mode disabled, and 0 matches with these settings enabled.

- In the Berger dataset, EGM misses 4 fusion genes compared with the overlap based method: GNG5-CTBS, SHANK-GNA12, KANSL1-ARL17A and intergenic fusion LOC100288142-LOC100288142 of which the latter was not matched by subset based matching. Of these fusion genes only SHANK2-GNA12 was reported to be a true positive. The three non validated fusion genes, all predicted by two tools, were investigated for being artefacts of large gene annotations.
  - GNG5-CTBS was not considered a match by EGM because the prediction by ChimeraScan has the left breakpoint within GNG5 as well as RPF1. In the used RefSeq annotation RPF1 is a 19.114bp long gene, larger than GNG4 with 8.257bp. Because GNG4, the shortest gene, is the gene shared by the two predictions, and because both genes are not large, this fusion is not likely to be matched because they share a large gene.
  - KANSL1-ARL17A was not considered a match by EGM because the prediction because ChimeraScan's predicted breakpoint is within the genes ARLA17A, NM_001103154 and LRRC37A whereas DeFuse's breakpoint is only within ARLA17A and NM_001103154. NM_001103154 (62.222bp) is an alias for both ARL17B and ARL17A and is entirely embedded in ARL17A. Similarly, the 42.664bp long gene LRRC37A that prevents EGM from matching, is entirely embedded in ARL17A. These three genes are overlapping for the majority of their length (Figure S4.6) and therefore the match is most likely due to overlapping gene annotations rather than because ARL17A is exceptionally long and overlaps these genes by chance. Off note, the overlap approach marked another ChimeraScan prediction as a duplicate: KANSL1-ARLA17A, LRRC37A4P, NM_001288811, NM_001288812 & NM_001288813 and this ARLA17A is a paralog of the ARLA17A used above. Therefore, the multiple genomic copies of ARLA17A allow the overlap method to consider these events as duplicates, which is not the case for the subset method. Because paralogues behave like long genes in the sense that they may be present at genomic distant locations, we consider this duplicate as a large gene type artefact.
  - LOC100288142-LOC100288142 is an intergenic fusion, not matched by EGM and neither by the subset approach, but being matched in ChimeraScan and Tophat Fusion by the overlap approach. The 2.320.934bp long gene is relatively large, lays within a gene rich region and spans several genes. Because of the large size and the number of genes it overlaps, this match

seems to be caused because `LOC100288142` is spanning multiple predicted fusion genes. The more conservative subset matching did not match these fusion genes.

- Using exact position matching in the non specific matching mode, we find:
  - 0 matches in SRR018259_M0000216
  - 2 matches in SRR018260_M990802
  - 1 match in SRR018261_M980409
  - 0 matches in SRR018265_M010403
  - 5 matches in SRR018266_501-MEL
  - 1 match for SRR018267_M000921
  - 3 matches for SRR018269_K-562-3-CML

  With a total of 12 of the 27 found by the overlap method, this misses more than half of the fusion genes.

- For the Edgren dataset, which was analysed with and without strand specific mode, the number of total matches reduced to almost 50% by using strand specific mode, while the true positives missed by the EGM approach were preserved.

- The dataset that should not contain any fusion gene, did not contain any matches in any of the methods (overlap, subset nor EGM). This does not allow to estimate specificity.

The results indicate that using a RefSeq gene annotation, EGM matches a considerable lower proportion (4/26 = 15.4% for the Edgren dataset and 4/27 = 14.8% (overlap) and 3/27 = 11.1% (subset) for the Berger dataset) of the fusion genes than the overlap and subset method, only because overlapping genes are not taken

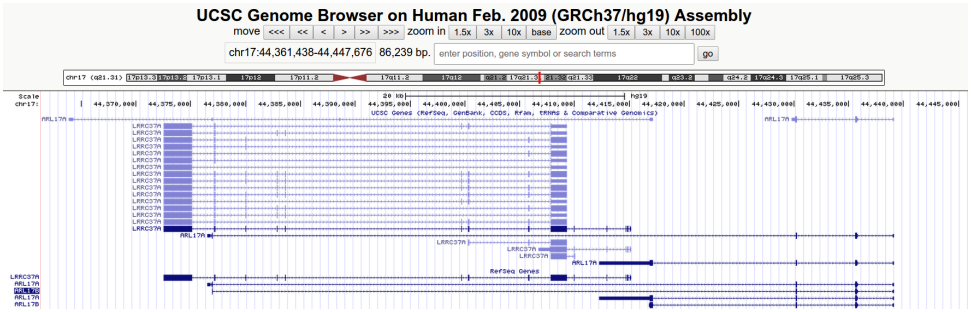



Figure S4.6: Genomic locations of overlapping genes `ARLA17A`, `LRRC37A` and `NM_001103154` (`ARLA17B`) on hg19

| method | Missed | Matched fusion genes | total |
|---|---|---|---|
| **EGM** | 7 | 45 | 52 |
| **FuMa-s** | 0 | 52 | 52 |
| **total** | 7 | 92 | 104 |

Table S4.6: **Contingency table that belongs to the Fisher's exact test.** To find to what extent EGM matches fewer fusion genes than the subset method (FuMA-s), we have used this contingency table for a Fisher's exact test. The total number of matched fusion genes missed by EGM in both datasets is 7 from a total of 52 matched by FuMa.



into account. Using Fisher's exact test on contingency Table S4.6 we find that EGM matches significantly less fusion genes than the subset method, with a P-value that equals 0.0126. Given increasingly more genes are discovered with RNA-seq, this problem will most likely increase over time.

The results do not clarify whether the overlap or subset matching method is superior. This is partially because the results are rather similar and partially because the number of available validated fusion genes is limited. On the one hand, in the Edgren dataset when an Ensembl gene annotation was used, the overlap method matches one more true positive, `BCAS4-BCAS3`, while on the other hand it matches `LOC100288142-LOC100288142` in the Berger dataset, likely to be a large gene artefact. The overlap method is more sensitive, because two sets that are subsets of each other are by definition overlapping, but not vice versa. This implies that any match by the subset method must be match in the overlap method, and thus not vice versa. As result, the overlap based method is more sensitive to large genes, as explained in Section 4.5.3.

The Edgren dataset was analysed with and without strand and donor-acceptor specific matching. This showed that although the number of matched true positives did not decrease, the number of total matches was reduced with ∼50% whith specific matching enabled. Therefore, when data is stranded, strand-specific matching is recommended.

| Edgren | | | | | | |
|---|---|---|---|---|---|---|
| **method** | **left-genes** | **right-genes** | **chimerascan** | **defuse** | **fusionmap** | **true positives** |
| overlap | CPNE1:$*^{1}$ | PI3 | **TRUE** | FALSE | FALSE | **TRUE** |
| egm | CPNE1:$*^{1}$:RBM12 | PI3 | TRUE | FALSE | FALSE | **FALSE** |
| egm | CPNE1:$*^{1}$ | PI3 | **FALSE** | FALSE | FALSE | TRUE |
| overlap | NR_036633:VAPB | IKZF3:$*^{2}$ | **TRUE** | FALSE | FALSE | **TRUE** |
| egm | NR_036633:VAPB | IKZF3:$*^{2}$:NM_001284516 | TRUE | FALSE | FALSE | **FALSE** |
| egm | NR_036633:VAPB | IKZF3:$*^{2}$ | **FALSE** | FALSE | FALSE | TRUE |
| overlap | SKA2 | MYO19 | **TRUE** | FALSE | FALSE | **TRUE** |
| egm | SKA2 | MYO19:$*^{3}$:ZNHIT3 | TRUE | FALSE | FALSE | **FALSE** |
| egm | SKA2 | MYO19 | **FALSE** | FALSE | FALSE | TRUE |
| overlap | ANKHD1:ANKHD1-EIF4EBP3 | PCDH1 | FALSE | FALSE | **TRUE** | **TRUE** |
| egm | ANKHD1:ANKHD1-EIF4EBP3 | NM_001278613:PCDH1 | FALSE | FALSE | **FALSE** | TRUE |
| egm | ANKHD1:ANKHD1-EIF4EBP3 | PCDH1 | FALSE | FALSE | TRUE | **FALSE** |

| Berger (*SRR018261 M980409*) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **method** | **left-genes** | **right-genes** | **chimerascan** | **defuse** | **fusioncatcher** | **star-fusion** | **tophat fusion** | **true positives** |
| overlap | GNG5 | CTBS | **TRUE** | **TRUE** | FALSE | FALSE | FALSE | FALSE |
| egm | GNG5:RPF1 | CTBS | TRUE | **FALSE** | FALSE | FALSE | FALSE | FALSE |
| egm | GNG5 | CTBS | **FALSE** | TRUE | FALSE | FALSE | FALSE | FALSE |
| Berger (*SRR018266 501-MEL*) | | | | | | | | |
| **method** | **left-genes** | **right-genes** | **chimerascan** | **defuse** | **fusioncatcher** | **star-fusion** | **tophat fusion** | **true positives** |
| overlap | SHANK2 | GNA12:$*^{5}$ | **TRUE** | **TRUE** | FALSE | **TRUE** | **TRUE** | **TRUE** |
| egm | SHANK2 | GNA12:$*^{5}$ | **FALSE** | TRUE | FALSE | TRUE | TRUE | TRUE |
| egm | NR_110766:SHANK2 | GNA12:$*^{5}$ | TRUE | **FALSE** | FALSE | **FALSE** | **FALSE** | **FALSE** |
| Berger (*SRR018269 K-562-3-CML*) | | | | | | | | |
| **method** | **left-genes** | **right-genes** | **chimerascan** | **defuse** | **fusioncatcher** | **star-fusion** | **tophat fusion** | **true positives** |
| overlap | KANSL1 | ARL17A | **TRUE** | **TRUE** | FALSE | FALSE | FALSE | FALSE |
| subset | KANSL1 | ARL17A:$*^{5a}$ | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE |
| subset | KANSL1 | ARL17A:$*^{5c}$ | TRUE | **FALSE** | FALSE | FALSE | FALSE | FALSE |
| egm | KANSL1 | ARL17A:$*^{5a}$ | **FALSE** | TRUE | FALSE | FALSE | FALSE | FALSE |
| egm | KANSL1 | ARL17A:$*^{5b}$ | TRUE | **FALSE** | FALSE | FALSE | FALSE | FALSE |
| egm | KANSL1 | ARL17A:$*^{5c}$ | TRUE | **FALSE** | FALSE | FALSE | FALSE | FALSE |
| overlap | LOC100288142 | LOC100288142 | **TRUE** | FALSE | FALSE | FALSE | **TRUE** | FALSE |
| egm | LOC100288142:$*^{6a}$ | LOC100288142:$*^{6b}$ | TRUE | FALSE | FALSE | FALSE | **FALSE** | FALSE |
| egm | LOC100288142:$*^{6b}$ | LOC100288142:$*^{6c}$ | TRUE | FALSE | FALSE | FALSE | **FALSE** | FALSE |
| egm | LOC100288142:$*^{6d}$ | LOC100288142:$*^{6e}$ | **FALSE** | FALSE | FALSE | FALSE | TRUE | FALSE |

Table S4.7: **Differences in fusion genes found in 2 or more datasets between overlap and subset based matching and EGM.** The listed datasets were analysed using a RefSeq gene annotation using subset, overlap and EGM based matching. The **top** table contains the differences in the Edgren dataset, which was analysed with strand and acceptor donor specific matching. The **bottom** table contains the Berger dataset, analysed without these additional constraints. The individual datasets of the Berger dataset are separated. Within each dataset the different fusion genes are separated with a solid line and the results of per matching approach separated with a dashed line. In all cases the subset results are identical to the overlap results and are therefore not shown, except for `KANSL1-ARLA17A` and integenic fusion `LOC100288143`. The output was truncated because of the larger gene sets. The substitutions are listed below:

- $*^{1}$ NM_001198863:NR_037188
- $*^{2}$ NM_001284514:NM_001284515
- $*^{3}$ NM_001281432:NM_001281433:NM_001281434:NR_104009:NR_104010:NR_104011
- $*^{4}$ NM_001282441:NM_001293092
- $*^{5a}$ NM_001103154
- $*^{5b}$ LRRC37A:NM_001103154
- $*^{5c}$ LRRC37A4P:NM_001288811:NM_001288812:NM_001288813
- $*^{6a}$ NBPF9
- $*^{6b}$ NBPF9:PDE4DIP
- $*^{6c}$ NBPF10:NM_001302371
- $*^{6d}$ NBPF10:NM_001302371:PIAS3
- $*^{6e}$ NBPF10:NM_001302371:NUDT17

# 5 *Dr. Disco*: prediction of TMPRSS2-ERG DNA breakpoints in random-primed RNA-Seq data

*unpublished*



Youri Hoogstrate [1,2], Natasja Dits [1], Elena S Martens-Uzunova [1], Malgorzata A Komor [3,4], Adam van Adrichem [2], Christian Rausch [3], David van der Meer [5], Bart Janssen [5], Wilbert van Workum [5], Mark de Jong [5], Job van Riet [1,6], Harmen van de Werken [1,6], Chris H Bangma [1], Geert JLH van Leenders [2], Gerrit A Meijer [3], Peter van der Spek [2], Andrew P Stubbs [2], Remond JA Fijneman [3] and Guido Jenster [1]

[1] *Department of Urology, Erasmus University Medical Center, Rotterdam, The Netherlands* [2] *Department of Pathology, Erasmus University Medical Center, Rotterdam, The Netherlands* [3] *Translational Gastrointestinal Oncology, Department of Pathology, Netherlands Cancer Institute, Amsterdam, the Netherlands* [4] *Oncoproteomics Laboratory, Department of Medical Oncology, VU University Medical Center, Amsterdam, the Netherlands* [5] *ServiceXS, Leiden, The Netherlands* [6] *Cancer Computational Biology Center, Erasmus University Medical Center, Rotterdam, The Netherlands*

Supplementary material: Section 5.6

## Abstract

Fusion genes are often driver mutations in different types of cancer. They can be detected with DNA-sequencing (DNA-seq) and, as fusion-transcript, with RNA-sequencing (RNA-seq). In RNA-seq experiments, typically mRNA is extracted by targeting poly-A tails, or reverse transcribed using oligo-dT primers. The location of the majority of the DNA breaks cannot be detected because they are rarely located in exons, while RNA-seq data typically, in majority, consists of exon derived sequences. By reverse transcribing ribo-depleted total RNA using random hexamer primers, the RNA library will also include pre-mRNA, containing sequences derived from introns. Here, we introduce a computational method for detecting inferenced DNA breaks of fusion genes, on top of exon-to-exon fusion splice junctions, using RNA-seq data only. To detect genomic breakpoints, intronic and exonic data are kept separated in a graph data structure. The graph analysis works with paired-end sequencing data and is capable of determining multiple exon-to-exon boundaries per fusion. Unlike most RNA-seq fusion gene detection software, there is no restriction to gene or exon annotations, allowing detection of novel splice junctions, fusions to non-gene regions and fusions of non-polyadenylated transcripts. The software makes use of community standard file formats to allow integration with workflow management systems, and for compatibility with genome browsers. In this study its relevance is demonstrated by generating an overview of `TMPRSS2-ERG` breakpoints in a cohort of 51 prostate cancer samples. The results confirm similar hotspot regions detected in indepedent DNA-seq analysis of different patients. We also reveal novel exons involved in the fusion. Beyond the `TMPRSS2-ERG` DNA breakpoints, additional deletions in `TMPRSS2` were found in fusion positive samples. Thus, by analysing the entire genome for fusions using random primed ribo-depleted total RNA-seq data, the vastly increased search space allows detection of novel, cancer-specific, RNA molecules.

**Availability and Implementation:** The source code is available at:
`https://github.com/yhoogstrate/dr-disco`

## 5.1 Introduction

Fusion genes are often driver mutations in cancer and have a great potential to function as biomarker for diagnosis and therapy selection [204, 205]. For example, fusion genes of the `ETS` gene family are detected in about 50-70% of the prostate cancer (PCa) patients [206, 207], including `TMPRSS2-ERG` with an estimated incidence of ∼50%. COSMIC[1] is a curated database that gathers information on cancer-related fusions and contains almost 300 annotated fusion genes, linked to almost 10,000 samples [105]. Similarly, the TCGA Fusion Gene Data Portal[2] is a database containing more than 10,000 fusion transcripts [208].

Fusion genes detected on DNA level can be further interrogated in a corresponding browser [209], for example to predict possible fusion proteins. Detecting fusion genes in RNA-seq compared with DNA-seq has several advantages such as determination of altered expression levels, revealing the fusion genes' splice variants [70] and it can more easily be used to predict possible chimeric proteins. In addition, stranding, splice junctions and gene structure may indicate which gene acts as donor and acceptor and may explain whether it functions as oncogene or tumor suppressor. A disadvantage of using RNA-seq is that not or lowly expressed rearrangements will not be detected.

Detecting fusion genes in DNA as well as RNA sequencing is challenging for a variety of reasons. During alignment to a reference genome there is a realistic chance on reads mapping to multiple (homologous) locations, since relatively small fragments are being sequenced rather than entire chromosomes or transcripts. In RNA-seq, the sequencing depth needed to determine a breakpoint with high confidence is higher compared with the sequencing depth needed for expression analysis. The approximate location of a fusion is typically revealed using *spanning reads*; paired-end reads that have both their mates aligned into different fusion partners [71]. The precise location of a fusion is then determined with *split reads*, reads that are split exactly over the junction.

Due to the imperfect and stochastic nature of size selection, the distance between mapped paired-end reads is variable. Terms that are related to the distance between paired-end reads are described in Figure 5.1. The variability in insert size complicates determining whether reads originate from wild-type sequences or rearrangements. RNA-seq data has the additional complexity of RNA processing leading to non-continuous alignments due to for instance splicing, read-throughs and circRNAs. As a result, RNA-seq aligners suffer from longer processing times or higher hardware requirements than for DNA-seq. RNA specific aligners typically align spliced reads by

[1] `https://cancer.sanger.ac.uk/cosmic/fusion`
[2] `http://54.84.12.177/PanCanFusV2/`

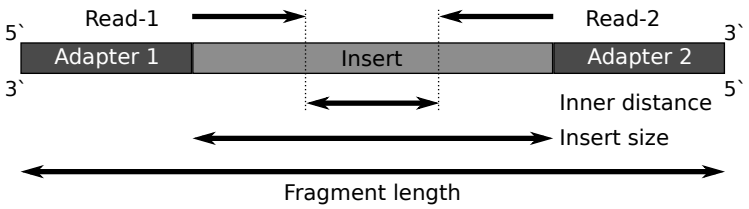



Figure 5.1: The length of the cDNA fragment includes the insert and the adapters. The insert size is the length of the insert, whereas the inner distance is the length of the insert minus the length of both reads. Reads can have a negative inner distance when they are overlapping.



preferring gapped alignment in close proximity over long distances, while for fusion gene detection it is equally important to look elsewhere in the genome for fusion partners. This requires aligners to use different settings for fusion gene analysis, which increases the search space and the number of optima, resulting in a higher number of multi-mapping reads and alignment artefacts.

During co-transcriptional splicing, pre-mRNA is first capped and introns are spliced out immediately when the splice donor and acceptor sequences are transcribed. As a last step, when the poly-A signal is recognised, the transcript will be poly-adenylated and the mature mRNA is released. RNA-seq libraries are typically prepared by targetting the poly-A tails (poly-A+ RNA-seq). Because splicing takes place mostly before poly-adenylation, corresponding sequencing typically consist of exon-derived sequences, while sequences that originate from introns are rare. Although rare, intron sequences are found in poly-A+ RNA-seq and can originate from post-transcriptional splicing and intron retention.

Genomic breaks resulting in fusion genes are most often located in introns. Because reads derived from intronic regions are rare in poly-A+ RNA-seq data, fusion transcripts are detected as exon-to-exon splice junctions [196]. If a fusion gene consists of multiple splice isoforms, multiple exon-to-exon boundaries may be derived from the same fusion gene (Figure 5.2). In contrast, when random hexamer primers are used in the reverse transcription step of ribo-depleted total RNA, also non-polyadenylated transcripts will be a template for RNA-seq library construction. This type of data will further be referred to as *random primed RNA-seq*. Random primed RNA-seq data does not only include mature mRNA, but also (l)ncRNAs, rRNA and pre-mRNA being transcribed. Because random primed RNA-seq data also contains intron spanning reads from pre-mRNA, corresponding data should allow detection of genomic breakpoints [87], as further explained in Figure 5.2.

In this study, we have developed a computational method for the detection of

fusion genes in random primed RNA-seq data, including the ability to detect and distinguish between DNA breaks on top of detecting exon-to-exon fusion splice junctions. Additionally, we provide an improved solution for the interactive visualisation of rearrangements in RNA-seq data.

## 5.2 Methods

For the *CTMM NGS-ProToCol* study, 51 prostate cancer (PCa) and 41 normal adjacent prostate (NAP) samples were sequenced, of which the total RNA was prepared with random hexamer primers for cDNA synthesis. RNA was extracted with a NEBNext Ultra Directional RNA PREP Kit and ribosomal RNA was reduced using a RNase-H based method. Additional clinical and molecular information is given in supplementary Tables S5.1 and S5.2. The 2x126 nt stranded paired-end reads were sequenced on a Illumina HiSeq 2500, with an average sequencing depth of ∼70 million paired-end reads per sample.



### 5.2.1 Computational Analysis

#### Data preparation

Trimmomatic [210] was used to improve the overall base quality, using settings given in section S5.6.1. The trimmed reads are aligned with STAR [199], using similar settings as used by STAR-Fusion [200], given in supplementary section S5.6.2. Within the results are discordant reads, reads that are either split up or have their mate inverted or aligned with an inner distance that does not fit a canonical gene. The analysis of these discordant reads in the application, *Dr. Disco*, is comprised of 5 steps: (`I`) transformation of alignment data into a graph, (`II`) merging edges of split and spanning reads that correspond to the same event, (`III`) extract edges that correspond to different splice isoforms of the same fusion, (`IV`) filtering and (`V`) determination of SV-type (Figure 5.3). It was noticed that after running STAR, certain reads are not directly compatible with the *split view* in Integrative Genomics Viewer [211] or for subsequent analysis because *SA:i:*-tags are not in place. Therefore *Dr. Disco* extends these files, which is explained in more detail in supplementary section S5.6.3.

#### `I`. Transformation into graph

An aligned read is considered *discordant* when it cannot be derived from a transcript of a classical gene because of inconsistent orientation, distance to its mate or an introduced split. Discordant fusion reads are either *singleton* or *paired-end*. Singletons

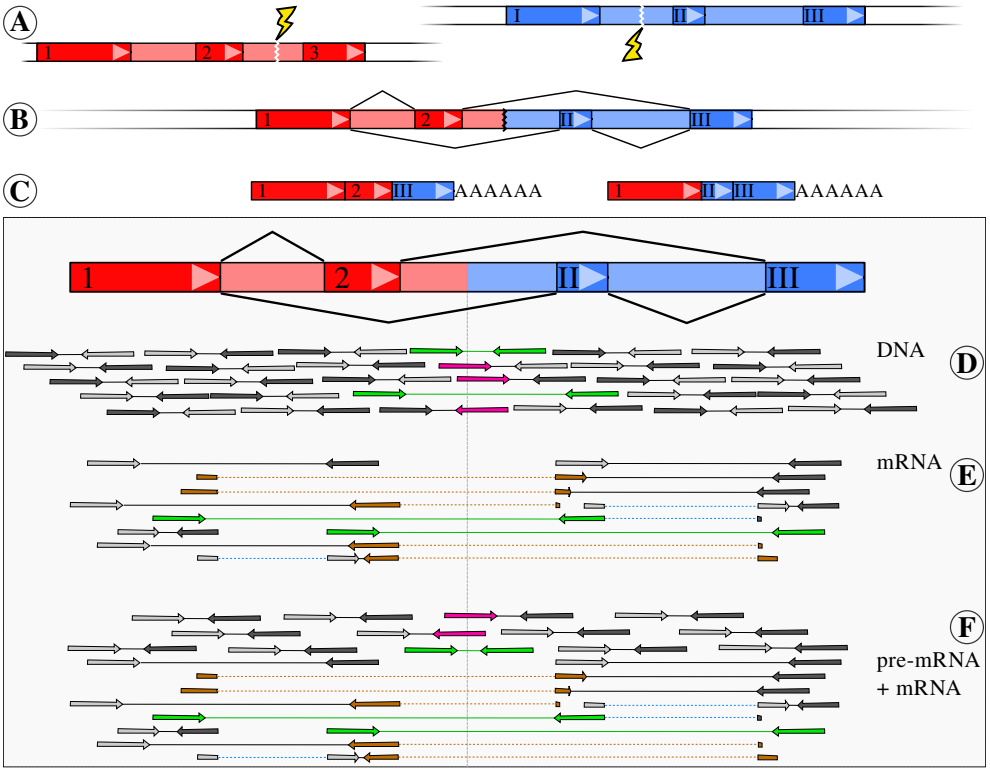
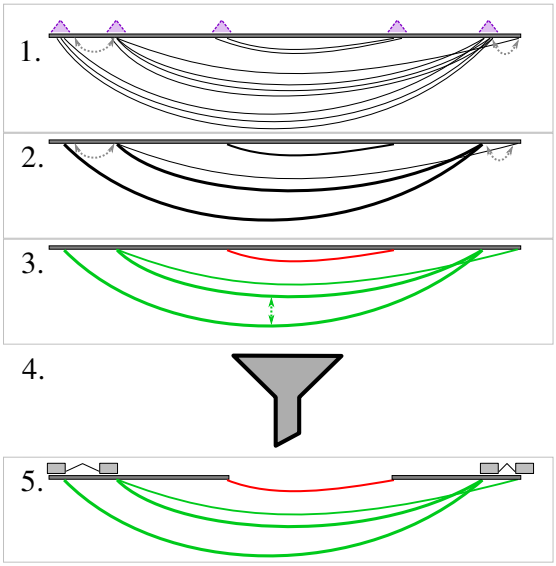



Figure 5.2: Differences between DNA-seq, random primed and poly-A+ RNA-seq data with respect to fusion genes. (**A**) Illustration of a deletion between two genes (red and blue). Both genes consist of three exons (red: *1*, *2* and *3*; blue: `I`, `II` and `III`) and two introns. Note that this is a schematic representation and that introns are typically much larger than exons. (**B**) A DNA recombination that leads to a fusion gene consisting of exons of both partners. Because the genomic breaks are in introns *2*-*3* and `I`-`II`, the fusion gene contains exons *1*, *2*, `II` and `III`. (**C**) The transcribed fusion will be alternatively spliced resulting in different mRNA fusion transcripts. (**D**) DNA-seq data covers the entire genome uniformly, including the fusion gene, and spans the genomic breakpoints. The fusion gene structure and splice isoforms cannot be unambiguously deduced from DNA sequencing data only. The data only contains discordant reads spanning the genomic breakpoint (split reads: pink, spanning reads: green). (**E**) In poly-A+ RNA-seq, data covers mostly exons, in quantities that correspond to splice isoform expression levels. Using sequencing data only, it is possible to predict the strand and splice isoforms of the fusion transcript, without being able to detect the actual genomic breakpoints. Both spanning and split reads are indicated over the exon-exon boundaries (in green and brown, respectively). (**F**) Random primed RNA-seq data contains both mRNA and pre-mRNA derived reads. Similar to DNA-seq data, pre-mRNA-derived reads include both intron and exon spanning sequences and as a result, the entire fusion gene is covered with reads. Whereas DNA-seq data covers an entire genome, pre-mRNA covers expressed regions and therefore also allows detection of genomic breakpoints. Indicated are reads spanning the exon-to-exon boundary as well as the genomic breakpoint (green), split reads covering the genomic breakpoint (pink) and (spliced) split reads covering the exon-to-exon boundary (brown).




Discordant reads are transformed into breakpoints and junctions which are inserted into a graph.

Merge edges with similar genomic locations

Extract edges that originate from different splice isoforms, based on distances at mRNA level

Filter out false positives

Determine type of break: intronic or exonic

Figure 5.3: Flowchart of *Dr. Disco*. (**1**) Edges and nodes are parsed from a discordant alignment and put in a graph. On top, a chromosome is indicated with a horizontal bar, edges with black curved lines and a splice junction representing edge with a dashed light gray line. Nodes are not indicated, but can be seen as the locations where the edges attach to the chromosome. Patterns of coherent edges emerge at more or less the same positions. These edges will be merged to bundle evidence and increase confidence per fusion gene. In paired-end sequencing only the very endings are sequenced. This results in spanning reads, which do not exactly represent the junction but fall in close proximity, limited to a certain distance. This distance should not exceed the inner distance, indicated with purple triangles. (**2**) To bundle edges that originate from the same junction, it is desired to merge split- and (usually slightly shifted) spanning reads. This will reduce the size of the graph and make the weight of the remaining edges, as indicated with thicker lines, heavier. (**3**) The distance between edges at mRNA level is calculated, and indicated with a green arrow. Edges are recursively extracted from the graph based on the distance at mRNA level. This results in subgraphs that consists of all junctions from the same fusion gene, including different splice isoforms, but does not merge corresponding DNA breaks. Each resulting subgraph is indicated with a unique color. (**4**) Sub-graphs are filtered on several variables such as the number of reads, ratio split and spanning reads, the shape of the alignment (triangular or rectangular [194]). (**5**) The extracted sub-graphs are only a set of edges. If the edges within the sub-graph are all in close proximity of splice junctions, they are likely derived from mature mRNA. These subgraphs are classified as either *intronic* or *exonic* by finding the distance the the closest splice junction. The red junction, which is on both sides not close to an exon boundary, is considered intronic. The green junctions, all close to exon boundaries, are considered exon-to-exon fusion events.

are reads of which the mate was excluded from alignment and paired-end are those of which both mates were aligned. Discordant reads are classified in two major types: *split reads* and *spanning reads* [71]. Singletons are always split reads while paird-end reads can be both split and spanning. If paired-end reads have a mate that was split and the remaining mate was not split, the remaining mate is a *silent mate* which is neglected in fusion analysis. Spanning reads have no mates that are split but have a too large or too small inner distance or are inverted with respect to each other. Note that any discordant read (singleton or paired-end, split or spanning) can also be a spliced read.

The main data structure of the computational analysis is a *graph* representing the junctions provided by discordant reads. Each discordant read will be assigned a subtype, defined by whether it was split or spanning and singleton or paired-end. For every read, the gap between the genomic positions is estimated according to the subtype. From the graph's perspective, this gap can be seen as an *edge* between two genomic locations and the genomic locations can be seen as nodes. For the transformation of a split read into an edge, the genomic locations of the split point will represent the genomic location of the corresponding nodes. For the transformation of spanning reads into an edge, the *R1*-read's last aligned base and the *R2*-read's first aligned base will represent the genomic location of the corresponding nodes. If during insertion of an edge into the graph no identical edge was found, it will be inserted and labelled with the corresponding subtype. Otherwise, the (subtype specific) weight is increased by one. Splice junctions will be inserted as edges into a separate graph. An example of how a graph data structure looks like, is presented in Figure 5.4 (top and middle).

However, the graph data structure alone is not efficient for quickly searching within genomic regions, to find closely adjacent edges. For quick access based on genomic coordinates, a reference to all edges is added to a genomic interval data structure [212], implemented with the HTSeq library [52].

### II. Merging edges derived from same event

Typically, split reads are located exactly on the breakpoint, while spanning reads are slightly shifted. A spanning read is shifted no more than its inner distance. Based on insert size statistics (Figure S5.7), we have set the maximum inner distance, a parameter for the algorithm, to 450 bp. From the graph's perspective, split and spanning reads of the same fusion correspond to edges that are on both sides in close proximity. To bundle evidence from the same junction, corresponding edges are merged as illustrated in Figure 5.4. Merging split and spanning reads derived from the same junction

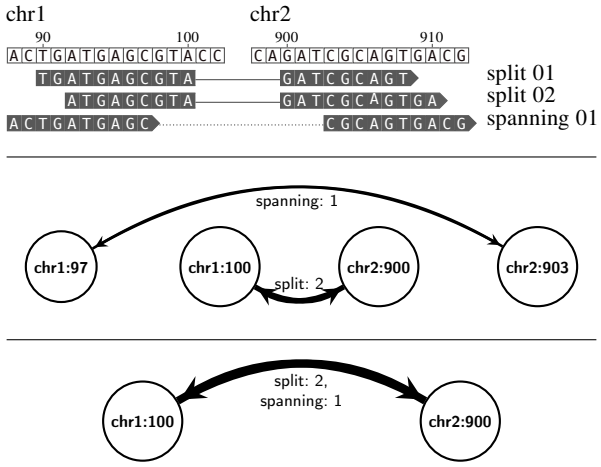



Figure 5.4: Merging closely adjacent edges. (**top**) An alignment consisting of two split reads and a spanning mate pair that represents two distinct gaps (chr1:100-chr2:900 & chr1:97-chr2:903). (**middle**) After transforming this alignment into a graph, one edge represents two split reads and the other represents one discordant mate pair. The weight of the edge that represents two split reads is two and therefore indicated with a thicker line. Both nodes at `chr1` and `chr2` are only 3 bp away from each other. (**bottom**) As both edges are in close proximity (3 bp + 3 bp), they are likely to be derived from the same event (either exon-to-exon junction or DNA break). After merging, only one edge remains of which the weight is the sum of the weight of the edges before merging.



starts with the edge with the largest weight. Because split reads are more powerful in determining exact breakpoints, they give edges a +50% increase in weight over spanning reads only during estimation of the heaviest edge. *Dr. Disco* then searches for other edges that have both their nodes no more than the maximum inner distance away. During merging, the total weight of the graph stays identical, while the number of unique edges decreases and the graph becomes smaller. This is illustrated in Figure 5.4 (middle and bottom); the number of edges decreases from 2 to 1, while the weight stays 3 (2 split, 1 discordant).

#### III. Extract edges from different splice isoforms

In the previous step, edges that belong to the same junction (either exon-to-exon, or DNA break) were merged, but exon-to-exon junctions from different splice isoforms of the same fusion gene are still separated since introns are typically much larger than the maximum inner distance. Edges that correspond to different splice isoforms of the same fusion gene are extracted in a two step approach. First, the shortest possible distance between two nodes at mRNA level is calculated. Second, edges that have a distance at the mRNA level of less than the maximum inner distance, are extracted

from the graph as they are likely to correspond to different splice isoforms of a fusion gene.

Estimation of the distance between two nodes at mRNA level is demonstrated using an example in Figure 5.5 (top). The gene on `chr1` contains two nodes, `chr1:1000` and `chr1:2500`, separated from each other with a genomic distance of 1500 bp. Between these nodes there is a splice junction (`chr1:1003-chr1:2490`). Remark that splice junctions are determined by the NGS data and not by gene annotations. For each node, there will be searched for the closest splice junction within a genomic distance of 450 bp. Starting with node `chr1:1000`, there exists one such splice junction at `chr1:1003`, with a distance of 3 bp. Similarly, for node `chr1:2500`, there exists one such splice junction, with a distance of 10 bp (to node `chr1:2490`). The distance at mRNA level between nodes `chr1:1000` to `chr1:2500` is $3 + 10 = 13$ bp and consequently the nodes shall be connected by inserting an edge that represents the distance at mRNA level, further referred to as *s-link*.

Extracting edges that correspond to different splice isoforms starts with the edge with the highest weight. Considering the example in Figure 5.5 (bottom) that contains two edges, the heaviest edge is `chr1:2500-chr2:900` with 2 split reads. The s-links will be traversed separately and recursively, until a maximum cumulative mRNA distance of 450 bp is reached and until all nodes found during traversal are stored. Selection starts with node `chr1:2500`, stored into a set of nodes that fall within the acceptable mRNA distance. This node has only one s-link, with a mRNA distance of 13 bp to node `chr1:1000`. Because this distance is smaller than 450 bp, set `{chr1:2500}` gets extended with `chr1:1000` into `{chr1:1000,chr1:2500}`. Although there are still 437 bases left for further iterations, no other s-links are connected to `chr1:1000` and the recursion ends. Because there are no s-links conneceted to node `chr2:900`, it will return a set only containing itself: `{chr2:900}`. All edges between `{chr1:1000,chr1:2500}` and `{chr2:900}`, which are `chr1:1000-chr2:900` and `chr1:2500-chr2:900`, will be extracted as they are likely to belong to the same (fusion) gene structure.

Because splice junctions are relatively far apart due to typically large introns, it was decided that such edges will not be merged into a single edge but will be extracted as a subgraph. These subgraphs preserve multiple exon-to-exon junctions and thus the fusion transcript structure. All edges present in a subgraph are removed from the main graph so they can only participate in one subgraph, corresponding to only one fusion. This process continues until all edges have been extracted and the main graph is empty.

It does occur that splice sites are not covered by reads, for example when the sequencing depth is low. As a result, when spliced reads are absent, extraction of subgraphs based on (splice junction derived) s-links cannot take place. Edges that

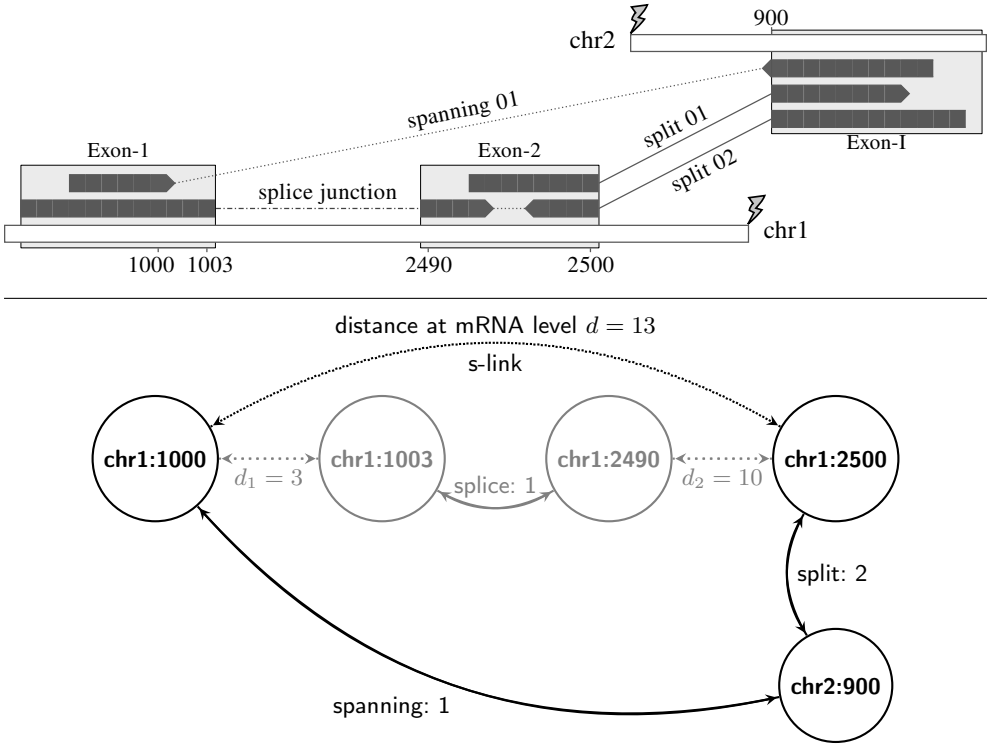



Figure 5.5: Extracting edges from different splice isoforms. (**top**) This schematic view of a fusion gene illustrates how edges corresponding to different splice isoforms are extracted. Exons are indicated as light gray blocks and the junctions within the discordant reads are indicated as lines between the exons. Between the two exons at chr1 and one at chr2 are two split reads (exon-2 − exon-I) and a spanning read (exon-1 − exon-I). The presence of a splice junction at `chr1:1003-2490` indicates these reads belong to the same fusion. Note that no intronic breaks are indicated. (**bottom**) The total genomic distance between edges `chr1:1000-chr2:900` and `chr1:2500-chr2:900` is $(2500-1000)+(900-900) = 1500$ bp. Distances at mRNA level are calculated for each pair of nodes on the same chromosome. These are calculated as sum of the two shortest genomic distances to the nodes that correspond to the splice junction. The distances to the closest splice junctions are $d_1 = 1003-1000 = 3$ and $d_2 = 2500 - 2490 = 10$ bp. The distance at the mRNA level is then calculated as the sum of both distances to the closest splice junction $(3+10 = 13$ bp). If there are two nodes of which the genomic distance minus the splice junction distance is smaller than the maximum inner distance, an s-link is inserted. Although the genomic distance in the example is 1500 bp, the distance at the mRNA level is only 13 bp. Since 13 bp is smaller than the maximum inner distance, 450 bp, the s-link will be included. As result, this allows edges `chr1:1000-chr2:900` and `chr1:2500-chr2:900`, from different isoforms, to be merged.

are located at exons that lack splice junctions will stay separated in the graph, even though they belong to the same (fusion) gene. An example in which this behaviour was repeatedly found is alternative exon-0 of `TMPRSS2`. This is an exon upstream of `TMPRSS2` [213], not present in most gene annotations such as RefSeq [10] and UCSC [214]. It was repeatedly observed that spliced reads in this exon were absent and consequently fusion transcripts involving this exon did not get extracted with the rest of the fusion, despite the fact that they are derived from splice isoforms of a `TMPRSS2-ERG` fusion. To overcome this technical issue, an additional step was implemented in which subgraphs are also extracted based on genomic distance (further explained in section S5.6.4). This fix allows inclusion of `TMPRSS2` exon-0 in `TMPRSS2-ERG` results, which is important since `TMPRSS2` exon-0 is a favourable prognostic marker [213].



### IV. Filtering

Discordant reads do not necessarily originate from fusion genes or rearrangements but could be mapping artefacts, sequencing artefacts, non-human contamination or related to population wide variation, read-throughs, and circRNAs [74]. To reduce the positives, different filters are applied. When there is a high number of identical copies of sequencing reads, the alignment has a rectangular shape. In contrast, when the alignment consists mostly of unique reads, the alignment has on both sides of the junction a more triangular shape (Figure S5.11). To which extent the shape of the alignment near the breakpoints is triangular, has earlier been reported to be useful in determining whether a candidate break is false positive [194], and is therefore used for filtering. Also, the ratio of split and spanning reads, the total number of reads and the alignment mismatch ratio are used to filter out false positives.

### V. Determination of SV-type

Each returned sub-graph is only a set of edges. These can be derived from a genomic breaks and from splice junctions of fusion genes. Near the junctions from those that are derived from spliced mRNA are most often spliced reads. The splice junction graph is used to extract the splice junctions found in these spliced reads. Based on the genomic distance to the closest splice junctions, the SV is classified as either intronic or exonic.

## 5.3 Results

The cohort consists of 41 NAP and 51 PCa samples, of which 40 PCa samples were RT-PCR tested for TMPRSS2-ERG (23 positive & 17 negative, Table S5.1). To determine for all PCa samples which are TMPRSS2-ERG positive, ERG expression levels were investigated, with 32/51 showing elevated expression (read count > 1000). The fusion was confirmed in these 32 samples using FusionCatcher [74].

To investigate TMPRSS2-ERG and related rearrangements, only discordant reads that at least partially map to ERG or TMPRSS2 (chr21:39,737,183-40,035,618, chr21:42, 834,478-42,882,085, hg19) were used for analysis with *Dr. Disco*. TMPRSS2-ERG was found in 32 of the 51 PCa samples, precisely those that have overexpression of ERG and were confirmed by FusionCatcher. Of these 32 samples, 23 had been RT-PCR tested for TMPRSS2-ERG of which all 23 samples were found to be positive. Of the remaining 17 RT-PCR tested samples, all 17 were found negative on both platforms. In three samples, *s027*, *s050* and *s053*, the genomic break of the fusion was not detected while exon-to-exon fusion junctions were. Of the 41 NAP samples, one (*s031*) was predicted to have TMPRSS2-ERG, which was not detected by FusionCatcher. The predicted junctions are given in Table S5.2.

Detected TMPRSS2-ERG DNA breakpoints and exon-to-exon junctions are shown in Figure 5.6. The DNA breakpoints in ERG are located in a hotspot region that spans the last half of intron 3, and were located more or less equally far apart from each other. The difference between the ratio of breakpoints per base in the hotspot region compared with the region at the beginning of intron 3 (chr21:39,898,001-39,947,586), is significantly different ($p < 0.01$, $\tilde{\chi}^2$-test).

Five samples have their genomic break outside the hotspot region in ERG (*s043*, *s031*, *s075*, *s054* and *s048*), of which the last three are relatively close to each other, before ERG starts (Figure 5.6). In two samples of which the ERG break was located within intron 3 but outside of the hotspot region, no other rearrangements were found. In the three samples of which the ERG break was located before ERG, additional rearrangements were present:

- In *s075* there are two small (91 bp & 201 bp) intronic amplifications (chr21:39,929,736-39,929,827 & chr21:40,063,681-40,063,882; hg19), not detected by *Dr. Disco* because of the small size (≤450 bp). These intronic amplifications do not seem to have effect on splicing as no known splice sites are affected.

- The read depth in sample *054* indicates a deletion that erases exon 3 of ERG. This junction is not supported by discordant read from STAR and could consequently not be detected by *Dr. Disco*. Further inspection revealed the presence of reads

harbouring sequences of both sides of the junction, while STAR did not split them but soft-clipped them instead. These reads are characterised by multiple mismatches, possibly caused by an insert-sequence or small mutations.

- The read depth in sample *048* indicates that there is a deletion of half the size of `ERG`, removing exons 2, 3 and *1. This junction is also not supported by discordant reads from STAR, while further inspection indicated concordant reads spanning both sides of the junction, which were softclipped instead of being split. These reads were also characterised by multiple mismatches, possibly as result of an insert-sequence or small mutations.

Each of the three fusions that break before `ERG` make use of two cryptic intergenic exons (`chr21:40,064,445-40,064,721`; `AC[CA..GC]CT` & `chr21:40,073,517-40,073,871`; `AC[CC..CT]CT`; hg19). The corresponding alignments of the fusions and corresponding additional rearrangements are indicated in Figures S5.9 and S5.10.

In `TMPRSS2` also a preferred SV region is apparent (`chr21:42,866,505-42,877,200`) encompassing introns 1 and 2 entirely. Ordering the samples on genomic breakpoints in `TMPRSS2` does not show a trend in the genomic breakpoints in `ERG` (Figure S5.8). This suggests that both DNA breakpoints of a `TMPRSS2-ERG` fusion are independently random. On top of the `TMPRSS2-ERG` fusions, two intronic deletions were identified in `TMPRSS2` (*s055*, *s064*; Table S5.2). These deletions are located within a single intron and therefore do not seem to affect mRNA or protein coding sequences.

The detected exon-to-exon junctions are located on the outside of the DNA breaks, which fits with that the region within the DNA breaks is deleted. Exon-to-exon junction involving exons that are not present in current gene annotions were found. Some of these exons are intergenic, including exon-0 [213] of `TMPRSS2`. Two recurrent cryptic new exons were found near exon(s)-1 of `TMPRSS2`, denoted as exons `*1` and `*2`, which are shown in Figure 5.6. The corresponding number of discordant reads from these exons are: exon-*2: 796, exon-*1: 582, exon-1a: 3306, exon-1b: 7442 and exon-0: 3307 reads, indicating these new exons are expressed to a lower extent than common exons 1a and 1b.

## 5.4 Discussion

STAR-Fusion [200], in particular designed to estimate exon-to-exon junctions of fusions involving annotated genes, makes use of aligner STAR [199]. In another study STAR was used for fusion gene detection in a similar setup [216]. Such implementations re-use publicly available, thoroughly tested and curated software in a modular manner, which gives more control over a pipeline compared with fully embedded

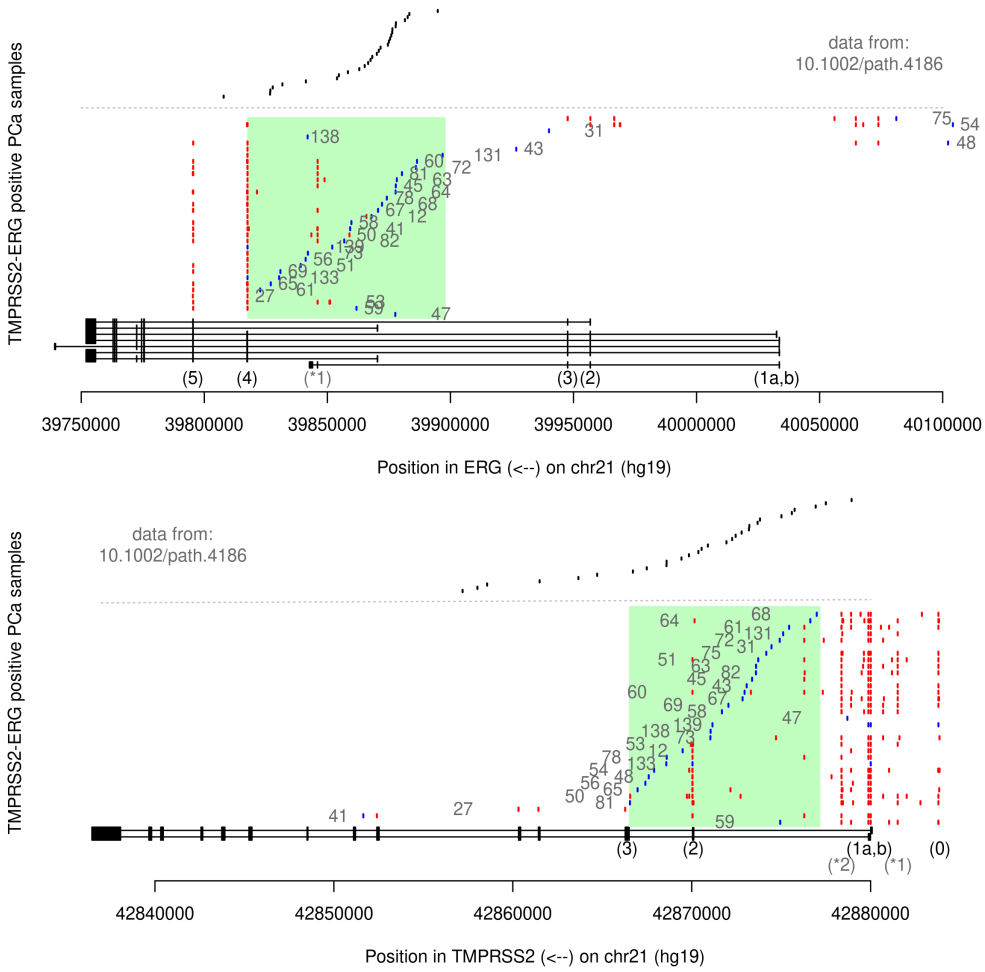



Figure 5.6: Genomic regions of `ERG` (**top**) and `TMPRSS2` (**bottom**). Annotated transcripts are given at the bottom of each plotted gene as thin black lines and the corresponding exons as thicker black boxes. The annotated transcripts of `ERG` are: `NM_001243429`, `NM_001136155`, `NM_001243428`, `NM_001136154`, `NM_001243432`, `NM_004449`, `NM_182918` & `NR_111949` (putative ncRNA), and of `TMPRSS2`: `NM_001135099` & `NM_005656`. Common alternative exons found in exon-to-exon junctions are indicated in gray, as (*1) in `ERG` and as (*1) and (*2) in `TMPRSS2`. Consensus exon numbers are indicated under the transcript annotations in parenthesis. The results of the RNA-seq analysis are separated per sample on the vertical axis. The numbers indicated in gray are the sample IDs. The DNA breakpoints of the RNA-seq analysis are marked in blue and the exon-to-exon junctions in red. The samples are ordered based on the position of the predicted DNA breakpoint. Detected breakpoints found in independent DNA-seq analysis [215], converted from hg18 to hg19 with UCSC liftOver, are plotted on top of both figures. Transcription starts at `TMPRSS2` (minus strand) and breaks at the genomic breakpoint, indicated in blue, and continues at the breakpoint in `ERG` (blue), also in the negative direction. In both genes the hotspot regions of the DNA breaks detected in the RNA-seq data are indicated with a green box.

pipelines. This makes it ideal for compatibility with workflow management systems such as Galaxy [217].

The input of *Dr. Disco* are *discordant* reads are in BAM or SAM alignment format. The SAM alignment format was chosen over the STAR *junctions* files, because BAM/SAM is the *de facto* standard file format for alignment data and is therefore compatible with software such as pysam, pybam, samtools, and HTSlib and with genome browsers.

Random primed RNA-seq is not specifically targeting polyadenylated mRNA but also includes non-polyadenylated RNA such as pre-mRNA. Because DNA breakpoints of fusion genes are most often located in introns, being able to sequence pre-mRNA should allow detection of genomic breakpoints of such fusions, while determination of exon-to-exon junctions remains possible. Combined identification of fusion genes at a genomic and transcriptomic level helps to better understand the relationship between the rearrangement at DNA and RNA level. This could for instance explain why certain exons or splice variants are present or absent. For certain fusions, the presence of a DNA break may rule out the possibility of a read-through event. Also, detection at both levels may increase confidence of a detected fusion.



Software package *Dr. Disco* was developed to also detect DNA rearrangements in RNA-seq data. The method puts all putative fusion gene evidence into a graph data structure, which allows separation of intronic from exonic data. Using 51 PCa samples, we show that the combination of random primed RNA-seq data with *Dr. Disco* allows detection of RNA molecules derived from `TMPRSS2-ERG` fusions, with junctions located in intergenic and intronic regions. It predicted `TMPRSS2-ERG` in the 32 samples that have elevated `ERG` expression and of which FusionCatcher had confirmed the fusion, including 23 RT-PCR `TMPRSS2-ERG` positive samples. The exon-to-exon junctions were typically surrounding the DNA breakpoints and not located within the expected 3 Mb deletion, which fits the expected fusion gene structure. In addition, the predicted DNA breakpoints fall in the same hotspot regions described in independent DNA-seq analysis [215].

Exons 4 and 5 were indeed the most common first `ERG` exons of `TMPRSS2-ERG` fusion transcripts surrounding the break [123]. This also fits with the detected genomic breakpoints, which were not found after exon 5 (`ERG`) and rarely after exon 3 (`TMPRSS2`). Breakpoints located in `ERG` were significantly more often found in the last half of intron 3 compared with the first half. These findings are in concordance with earlier reports [215]. Hence, the detected genomic breakpoints fit with previous reports and the results demonstrate that *Dr. Disco*'s model works as expected.

The mechanism behind `TMPRSS2-ERG` is explained by whether an `ERG` protein coding transcript (without frameshift) can occur [123]. The preference for breakpoints in

introns 1 and 2 of TMPRSS2 fits this critereon. However, the lack of DNA breakpoints in the first half of ERG intron 3 (chr21:39,898,000-39,947,587; hg19) suggests that selection is not only driven by intron/exon structure, because fusions to the region without breakpoints are expected to splice in a similar way and would result in similar proteins.

To make use of the full potential of random primed RNA, our method is not restricted to annotated genes or exons but is capable of revealing rearrangements between any two genomic locations. As a result, we found TMPRSS2-ERG transcripts containing new exons (Figure 5.6). In TMPRSS2, several new exons involved in TMPRSS2-ERG fusion transcripts were detected, including intergenic exon-0 [213].

In three samples (*s064*, *s051*, *s060*), we found exon-to-exon boundaries upstream of the genomic breakpoint of TMPRSS2, thus in the region expected to be deleted. Although there are reports indicating multiple, independent, TMPRSS2-ERG fusion genes in different subregions of the prostate in the same patient [123] as well as multiple variants within the same genome [218], we did not find any further evidence supporting multiple breaks and assume these are cross-sample contaminations, that possibly arose as de-multiplexing errors of the RNA-seq reads. The assumption that these reads are cross-sample contaminations is supported by the low amount of corresponding evidence (*s064*: 1 read, *s051*: 2 reads, *s060*: 5 reads).

In the three samples in which the DNA break was not found while exon-to-exon junctions were, sufficient DNA break spanning discordant reads were present. In sample *s053*, the genomic break in TMPRSS2 is 57 bp away from an exon, and in *s050* 114 bp. This led in both cases to merging the evidence of intronic and exonic data during the *merge* step of the graph analysis. A next step in improving the algorithm could be to find a solution for this. This could for instance by done by using the splice junctions found in the concordant alignment or by using a gene annotation and improving the rules used in the merge steps. In *s027*, the DNA break was detected (chr21:39,945,261-42,856,315; hg19), but the filter had classified it as invalid because of the lack of split reads.

To make use of the full potential of random primed RNA-seq data for fusion gene analysis, also intronic and intergenic genomic regions need to be interrogated for the presence of possible fusion transcripts. In this study we have focused beyond exon regions, and the results confirm that a full genome strategy is indeed capable of revealing different new cancer specific transcripts. This underlines that it can therefore detect more fusion transcripts than conventional fusion detection tools that are restricted to annotated genes and exons. Yet, a disadvantage of this full genome approach is the increased search space and computational complexity, compared with methods leaving this out [75].

## 5.5 Conclusion

We have developed a method, *Dr. Disco*, that can detect fusion genes in RNA-seq data across the entire genome, which allows detection of intergenic and intronic splice variants and DNA breaks. This method is particularly suitable for the analysis of random primed RNA because such data also contains relatively high proportions of pre-mRNA derived intron sequences. The corresponding results of 51 PCa samples have shown that the predictions of `TMPRSS2-ERG` are in concordance with earlier findings, in terms of predicted genomic locations and in terms of fusion gene structure. In addition, it revealed additional intronic deletions in `TMPRSS2` which cannot be detected in poly-A+ RNA-seq data. The results indicate that detecting exon-to-exon in combination with genomic breakpoints may be advantageous in understanding diseases involving large structural variants and could further help understanding mRNA products and fusion events. That the full genome approach of *Dr. Disco* is capable of detecting fusions involving intergenic regions is demonstrated by confirming the presence of `TMPRSS2` exon-0 and two recurrent alternative exons. Thus, a sequencing library enriched with non-polyadenylated transcripts in combination with analysis using *Dr. Disco* allowed discovery of novel cancer specific fusion transcripts, both at mRNA and pre-mRNA level, that could not be detected in a hightroughput manner before. The modular implementation of *Dr. Disco* makes use of community standard file formats and is therefore compatible with genome browsers to allow data visualisation. The software is available in Bio-Conda [3], the Galaxy platform [4] and included in The Galaxy RNA Workbench [219] [5].



## Acknowledgements and Funding

This study was financially supported by the framework of the Center for Translational Molecular Medicine (CTMM), NGS-ProToCol (grant 03O-40) and PCMM (grant 03O-203-1). Support for the Cancer Computational Biology Center was provided by the Daniel den Hoed Foundation.

[3] `https://bioconda.github.io/recipes/dr-disco/README.html`
[4] `https://toolshed.g2.bx.psu.edu/view/erasmus-medical-center/dr_disco`
[5] `http://bgruening.github.io/galaxy-rna-workbench/`

## 5.6 Supplementary materials

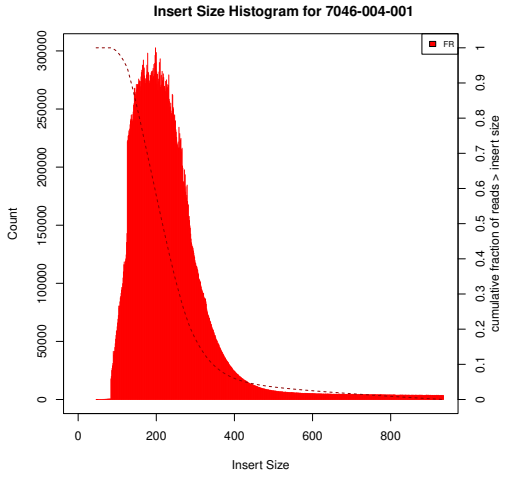



Figure S5.7: Results of the tool *Picard CollectInsertSizeMetrics* on sample *s001*.

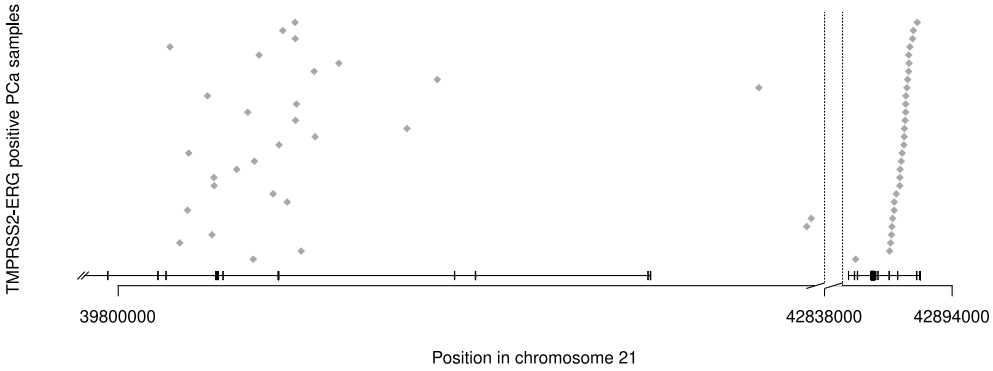



Figure S5.8: Map of predicted `TMPRSS2-ERG` breakpoints. Genomic breakpoints of `ERG` (left) and `TMPRSS2` (right) are indicated with gray diamonds. Gene annotations (hg19) are indicated at the bottom. Samples are ordered on the position of the genomic breakpoint in `TMPRSS2` whereas the order of breakpoints in `ERG` do not show noticible correlation.

Table S5.1: Clinical and molecular sample characteristics

| | Cancer | % in cancer | Normal | |
|---|---|---|---|---|
| Samples | 51 | | 41 | $\sum$=92 |
| `TMPRSS2-ERG` + | | | | |
| FusionCatcher | 32 | 62.7% | 3 | |
| PSA | 11 (0.3-64.3) | | - | |
| GS | | | | |
| 3+3 | 17 | 33.3% | - | |
| 3+4 | 24 | 47.0% | - | |
| 4+3 | 5 | 9.8% | - | |
| 8 | 2 | 3.9% | - | |
| 9-10 | 2 | 3.9% | - | |
| ? | 1 | 2.0% | - | |
| pT stage | | | | |
| 2 | 17 | 33.3% | - | |
| 3 | 20 | 39.2% | - | |
| 4 | 13 | 25.5% | - | |
| x | 1 | 2.0% | - | |

| | PCR tested | % of PCR tested | not PCR tested | |
|---|---|---|---|---|
| Cancer samples | 40 | | 11 | $\sum$=51 |
| `TMPRSS2-ERG` + | | | | |
| RT-PCR | 23 | 57.5% | - | |
| FusionCatcher | 23 | 57.5% | 9 | |
| `TMPRSS2-ERG` − | | | | |
| RT-PCR | 17 | 42.5% | - | |
| FusionCatcher | 17 | 42.5% | 2 | |

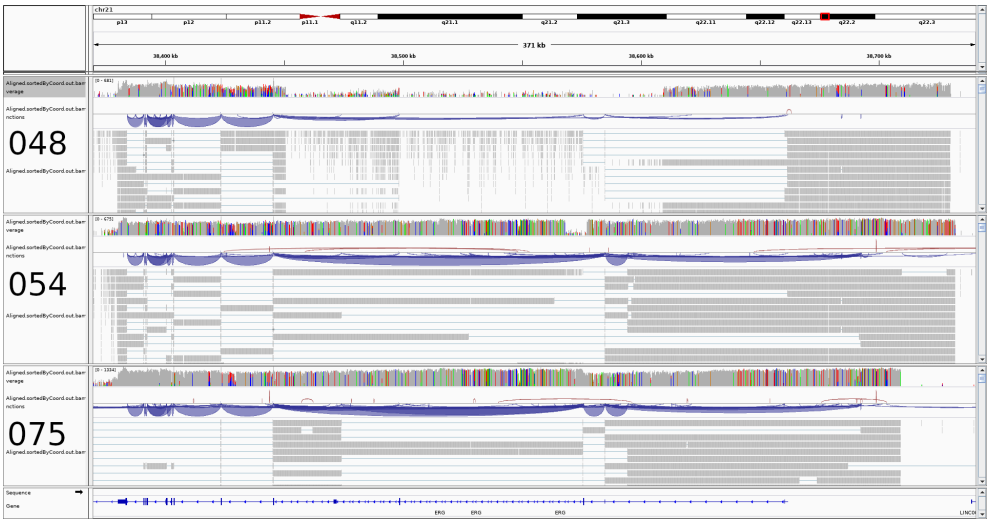



Figure S5.9: Concordant reads of samples with DNA break before `ERG`. This is an IGV view with the density tracks on 'log-scale', showing the concordant alignments (samples from top to bottom: *s048*, *s054*, *s075*) of which the DNA breakpoint was found before `ERG`. In sample *s048*, a region is visible of which the read density is low, caused by a deletion. Inside that region, on the left side, a small region with a lower density is visible. It is possible that there are differences in copynumber or clonality and that those reads correspond to a different deletion. In sample *054*, a drop in the read density can be seen around exon 3, caused by a deletion. The read depth in *s075* is not interrupted, indicating no deletions are present. These alignments are on hg38.

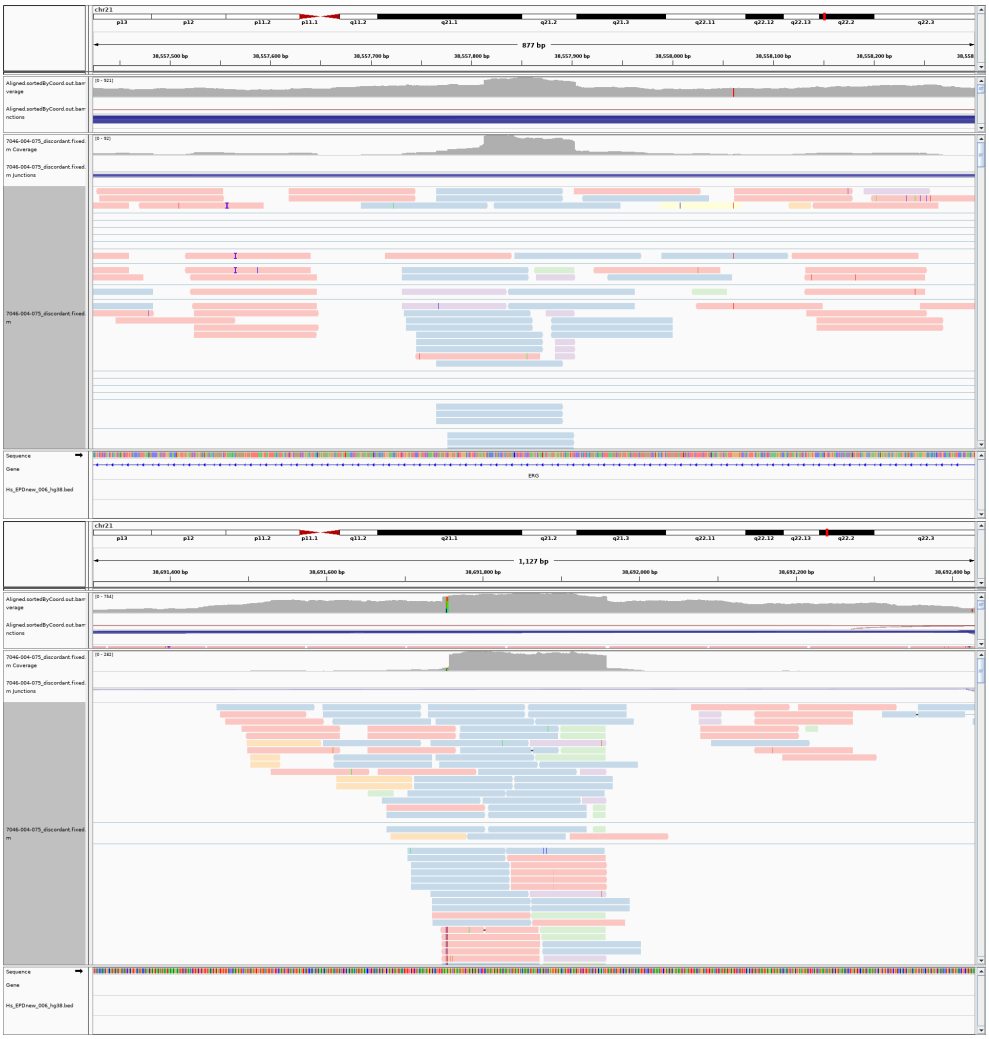


Figure S5.10: Alignments of small amplifications in sample *s075*. The alignment of two `TMPRSS2-ERG` related amplifications in sample *s075* in IGV. In both views, the density of reads from the concordant alignment is shown on top and the alignment of discordant reads at the bottom. These alignments are on hg38.

### 5.6.1 Trimmomatic settings

The following adapters were used for Trimmomatic [210] 0.33:

```
PrefixPE/1: TACACTCTTTCCCTACACGACGCTCTTCCGATCT
PrefixPE/2: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```

It was used with the following commandline arguments:

```
java -jar $TRIMMOMATIC_JAR PE \
    -threads 48 \
    -phred33 \
    $prefix_R1.fastq.gz $prefix_R2.fastq.gz \
    $prefix"_R1_paired.fastq.gz $prefix"_R1_unpaired.fastq.gz \
    $prefix"_R2_paired.fastq.gz $prefix"_R2_unpaired.fastq.gz \
    ILLUMINACLIP:adapter_fa:2:30:10 \
    LEADING:26 \
    TRAILING:26 \
    AVGQUAL:20 \
    SLIDINGWINDOW:4:24 \
    MINLEN:36
```

### 5.6.2 STAR settings

The following settings for STAR-v2.4.2a [199] were used:

```
STAR --runThreadN 9
     --genomeDir ... \
     --readFilesIn ... ... \
     --outFileNamePrefix ... \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMstrandField intronMotif \
     --outFilterIntronMotifs None \
     --alignIntronMax 200000 \
     --alignMatesGapMax 200000 \
     --alignSJDBoverhangMin 10 \
     --alignEndsType Local \
     --chimSegmentMin 12 \
     --chimJunctionOverhangMin 12 \
     --sjdbGTFfile gencode.v19.annotation.gtf \
     --sjdbOverhang 100 \
     --quantMode GeneCounts \
     --twopass1readsN -1 \
     --twopassMode Basic
```

### 5.6.3 Fixing discordant alignments

Discordant alignments of STAR have minor limitations that prevent a split view in Integrative Genomics Viewer. This is partially due to limitations in the BAM file format with respect to chimeric reads. According to the BAM/SAM specification[6] it is not possible to describe an interchromosomal junction as one single alignment. Instead, chimeric alignments are supposed to be split over multiple alignment entries. To correct for sticky ends, remaining parts of the aligned reads that belong elsewhere are made invisible (soft clipping). It is, however, not specified how to deal with PNEXT and RNEXT flags, which are supposed to link to the next aligned piece. Consider an example with two mates, R1 and R2, of which R2 is a discordant split read. This will result in alignment R2-*a* and R2-*b*. STAR links R1 to R2-*a* with the PNEXT and RNEXT flags, while both R2-*a* and R2-*b* link back to only R1. Consequently, during visualisation, no link between R2-*a* and R2-*b* is found and does therefore not allow a split view of split reads. This is fixed by linking R1 to R2-*a*, R2-*a* to R2-*b* and R1, with the remark that this is still not ideal as R2-*a* should link to both R1 as well as R2-*b*. In addition, `SA:Z:`-tags were added in order to reference the other chimeric alignments of the reads. An example of a proper split view of discorant reads that correspond to a `TMPRSS2-ERG` DNA break is given in Figure S5.11.

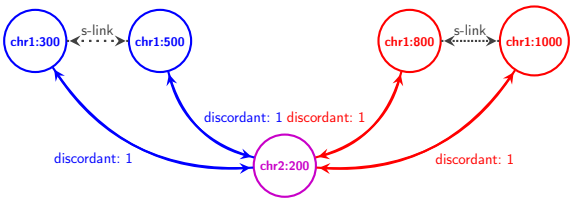

Figure S5.11: Split view of `TMPRSS2-ERG` in Integrative Genomics Viewer. An IGV split view of the reads that correspond to the DNA break of `TMPRSS2-ERG` of sample *s041*'s *fixed* (`dr-disco fix`) alignment. Reads are coloured per subtype where the *red* reads are spanning reads, *green* reads are silent mates and all the remaining reads are split reads.

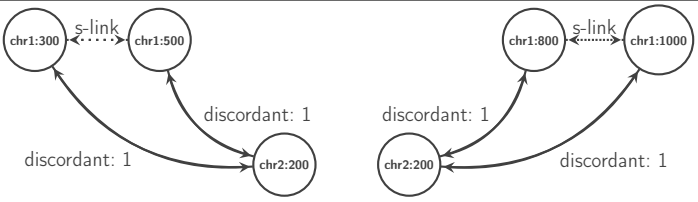
[6] `https://samtools.github.io/hts-specs/SAMv1.pdf` (revision *2a802cd*)

### 5.6.4 Extracting edges from different splice isoforms (without splice junctions)

To also extract edges that were left separated due to a lack of splice junctions, the last step in the extraction phase starts with obtaining all corresponding nodes per subgraph. Given that each edge contains two nodes, we denote for every subgraph two vectors. Each node in an edge is considered *left* or *right*, where a *left* node has a genomic location lexicographically smaller than the *right* node. Then for every two subgraphs, the smallest genomic distances to a node in the other vector is calculated for both the left and right set of nodes, which is illustrated in Figure S5.12. If the number of nodes in both sets is not identical, the distance to the nodes in the shortest vector are used to determine the distance. Note that genomic distances are used since splice junctions are missing. For both vectors a root mean square (*RMS*) is calculated. Using these two vectors (*left* and *right*) of minimal genomic distances and the RMS values, it is determined whether two subgraphs shall be merged by evaluating:

- If in both vectors there are distances smaller than the maximum inner distance size.
- If in one vector:
  - 100% of the distances are smaller than the maximum inner distance and of the other vector the *RMS* is less than 15000 bp.
  - 70% of the distances are smaller than the maximum inner distance and of the other vector *RMS* is less than 1000 bp.
  - 30% of the distances are smaller than the maximum inner distance and of the other vector *RMS* is less than 5000 bp.




$$\text{subnet 1 (left,right)} = \begin{bmatrix} \texttt{chr1:300} \\ \texttt{chr1:500} \end{bmatrix}, \begin{bmatrix} \texttt{chr2:200} \end{bmatrix} \tag{5.1}$$

$$\text{subnet 2 (left,right)} = \begin{bmatrix} \texttt{chr1:800} \\ \texttt{chr1:1000} \end{bmatrix}, \begin{bmatrix} \texttt{chr2:200} \end{bmatrix} \tag{5.2}$$

$$\text{dist (left)} = \begin{bmatrix} \min(\texttt{chr1:300} - \texttt{chr1:800} = 500, \texttt{chr1:300} - \texttt{chr1:1000} = 700) = -500 \\ \min(\texttt{chr1:500} - \texttt{chr1:800} = 300, \texttt{chr1:500} - \texttt{chr1:1000} = 500) = -300 \end{bmatrix} \tag{5.3}$$

$$\text{dist (right)} = \begin{bmatrix} \min(\texttt{chr2:200} - \texttt{chr2:200} = 0) = 0 \end{bmatrix} \tag{5.4}$$

$$RMS\ \text{(left)} = \sqrt{\frac{-500^2 + -300^2}{2}} = 412.3 \tag{5.5}$$

$$RMS\ \text{(right)} = \sqrt{\frac{0^2}{1}} = 0 \tag{5.6}$$

Figure S5.12: Extracting edges from different splice isoforms. (**top**) Two subgraphs (blue and red) that share a common exon are not merged because no splice junction exists between the red and blue nodes. (**upper**) The two subgraphs completely separated. (**mid**) To estimate whether both subgraphs can be merged based on genomic distance, vectors containing the corresponding *left* and *right* genomic locations are estimated. For subnet 1, the *left* locations are `{chr1:300, chr1:500}` and the *right* locations are `{chr2:200}`. For subnet 2, the *left* locations are `{chr1:800, chr1:1000}` and the *right* locations are `{chr2:200}`. (**lower**) Then the minimum distances between both *left* and *right* vectors are estimated. In case the number of nodes in two vectors is not identical, the shortest distances relative to the smallest vector are used and because of this the function becomes symetrical ($d(a,b) == d(b,a)$). The estimated minimal distances are: (*left*): `{-500, -300}` and (*right*): `{0}`. (**bottom**) Root mean square values of these distances are calculated and are used in combination with the distance vectors to make a decision. Because the calculated RMS values are 412.3 (*left*) and 0 (*right*) and both vectors contain distances smaller than 450, these subgraphs can be put together as they are likely to come from the same fusion gene despite no splice junctions were found.

Table S5.2: Results of `TMPSS2-ERG` analysis in NGS-ProToCol data (*hg19*). `ERG*` with asterisk means that the actual breakpoint is close to but not located inside `ERG`.

| Sample | `TMPRSS2-ERG` | Fusion | Breakpoint-1 | Breakpoint-2 | Type |
|---|---|---|---|---|---|
| *s*001-n | - | | | | |
| *s*003-n | - | | | | |
| *s*004-n | - | | | | |
| *s*005-n | - | | | | |
| *s*006-n | - | | | | |
| *s*007-n | - | | | | |
| *s*008-n | - | | | | |
| *s*011-n | - | | | | |
| *s*012-c | + | `TMPRSS2-ERG` | `chr21:39817544(-)` | `chr21:42870045(+)` | exonic |
| | | `TMPRSS2-ERG` | `chr21:39867955(-)` | `chr21:42869493(+)` | intronic |
| *s*013-n | - | | | | |
| *s*014-n | - | | | | |
| *s*015-n | - | | | | |
| *s*016-n | - | | | | |
| *s*019-n | - | | | | |
| *s*020-n | - | | | | |
| *s*021-n | - | | | | |
| *s*022-n | - | | | | |
| *s*024-n | - | | | | |
| *s*025-n | - | | | | |
| *s*026-n | - | | | | |
| *s*027-c | + | `TMPRSS2-ERG` | `chr21:39817544(-)` | `chr21:42860320(+)` | exonic |
| *s*031-n | - | `TMPRSS2-ERG` | `chr21:39940055(-)` | `chr21:42874460(+)` | intronic |
| *s*032-n | - | | | | |
| *s*033-n | - | | | | |
| *s*035-n | - | | | | |
| *s*036-n | - | | | | |
| *s*037-n | - | | | | |
| *s*038-n | - | | | | |
| *s*039-n | - | | | | |
| *s*040-n | - | | | | |

| Sample | TMPRSS2-ERG | Fusion | Breakpoint-1 | Breakpoint-2 | Type |
|---|---|---|---|---|---|
| *s*041-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| | | TMPRSS2-ERG | chr21:39859273(-) | chr21:42851646(+) | intronic |
| *s*042-c | – | | | | |
| *s*043-c | + | TMPRSS2-ERG | chr21:39926744(-) | chr21:42873074(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*044-c | – | | | | |
| *s*045-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39877811(-) | chr21:42873374(+) | intronic |
| *s*047-c | + | TMPRSS2-ERG | chr21:39877602(-) | chr21:42878701(+) | intronic |
| | | TMPRSS2-RERE | chr1:8414073(+) | chr21:42878790(-) | intronic |
| | | TMPRSS2-RERE | chr1:8414473(-) | chr21:42871549(-) | intronic |
| *s*048-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG* | chr21:40102202(-) | chr21:42867598(+) | intronic |
| | | TMPRSS2-ERG* | chr21:40073871(-) | chr21:42870045(+) | exonic |
| *s*049-c | – | | | | |
| *s*050-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| | | TMPRSS2-ERG | chr21:39858981(-) | chr21:42866563(+) | exonic |
| *s*051-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39839166(-) | chr21:42873719(+) | intronic |
| *s*052-c | – | | | | |
| *s*053-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| *s*054-c | + | TMPRSS2-ERG | chr21:39956869(-) | chr21:42883789(+) | exonic |
| | | TMPRSS2-ERG* | chr21:40104140(-) | chr21:42867900(+) | intronic |
| | | TMPRSS2-ERG* | chr21:40073871(-) | chr21:42870045(+) | exonic |
| *s*055-c | – | TMPRSS2-TBX3 | chr12:115112640(-) | chr21:42866282(+) | exonic |
| | | TMPRSS2-TBX3 | chr12:115113110(-) | chr21:42864779(+) | intronic |
| | | TMPRSS2-TMPRSS2 | chr21:42858784(+) | chr21:42871555(+) | intronic |
| | | TMPRSS2-TMPRSS2 | chr21:42862260(-) | chr21:42871707(-) | intronic |
| | | TMPRSS2-TBX3 | chr12:115114117(+) | chr21:42840465(-) | exonic |
| *s*056-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| | | TMPRSS2-ERG | chr21:39841171(-) | chr21:42867411(+) | intronic |
| *s*057-c | – | | | | |
| *s*058-c | + | TMPRSS2-ERG | chr21:39859776(-) | chr21:42871690(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*059-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42883789(+) | exonic |
| | | TMPRSS2-ERG | chr21:39861836(-) | chr21:42874947(+) | intronic |
| | | TMPRSS2-intergenic | chr2:30904387(+) | chr21:42847992(+) | intronic |
| *s*060-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42883789(+) | exonic |
| | | TMPRSS2-ERG | chr21:39886422(-) | chr21:42872955(+) | intronic |

| Sample | TMPRSS2-ERG | Fusion | Breakpoint-1 | Breakpoint-2 | Type |
|---|---|---|---|---|---|
| *s*061-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39822719(-) | chr21:42875445(+) | intronic |
| *s*062-c | - | | | | |
| *s*063-c | + | TMPRSS2-ERG | chr21:39878221(-) | chr21:42873595(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*064-c | + | TMPRSS2-TMPRSS2 | chr21:42876044(+) | chr21:42876752(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39877714(-) | chr21:42876630(-) | intronic |
| *s*065-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| | | TMPRSS2-ERG | chr21:39826992(-) | chr21:42866975(+) | intronic |
| *s*067-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42883789(+) | exonic |
| | | TMPRSS2-ERG | chr21:39870591(-) | chr21:42872844(+) | intronic |
| *s*068-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39872200(-) | chr21:42876988(+) | intronic |
| *s*069-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| | | TMPRSS2-ERG | chr21:39830932(-) | chr21:42872053(+) | intronic |
| *s*070-c | - | | | | |
| *s*071-c | - | TMPRSS2-PADI4 | chr1:17666182(+) | chr21:42866282(+) | exonic |
| *s*072-c | + | TMPRSS2-ERG | chr21:39885972(-) | chr21:42874918(-) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*073-c | + | TMPRSS2-ERG | chr21:39842092(-) | chr21:42871045(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*074-c | - | | | | |
| *s*075-c | + | TMPRSS2-ERG* | chr21:40081166(-) | chr21:42874169(+) | intronic |
| | | TMPRSS2-ERG | chr21:39956869(-) | chr21:42880007(+) | exonic |
| *s*076-c | - | | | | |
| *s*077-c | - | | | | |
| *s*078-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| | | TMPRSS2-ERG | chr21:39874170(-) | chr21:42868603(+) | intronic |
| *s*079-c | - | | | | |
| *s*081-c | + | TMPRSS2-ERG | chr21:39880312(-) | chr21:42866548(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | exonic |
| *s*082-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42883789(+) | exonic |
| | | TMPRSS2-ERG | chr21:39856860(-) | chr21:42873590(+) | intronic |

| Sample | TMPRSS2-ERG | Fusion | Breakpoint-1 | Breakpoint-2 | Type |
|---|---|---|---|---|---|
| *s*125-n | - | | | | |
| *s*126-n | - | | | | |
| *s*130-c | - | | | | |
| *s*131-c | + | TMPRSS2-ERG | chr21:39896846(-) | chr21:42875112(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | exonic |
| *s*132-c | - | | | | |
| *s*133-c | + | TMPRSS2-ERG | chr21:39817544(-) | chr21:42870045(+) | intronic |
| | | TMPRSS2-ERG | chr21:39830403(+) | chr21:42868581(+) | intronic |
| *s*134-c | - | | | | |
| *s*135-c | - | | | | |
| *s*136-n | - | | | | |
| *s*137-c | - | | | | |
| *s*138-c | + | TMPRSS2-ERG | chr21:39841976(-) | chr21:42871057(+) | intronic |
| *s*139-c | + | TMPRSS2-ERG | chr21:39851982(-) | chr21:42871161(+) | intronic |
| | | TMPRSS2-ERG | chr21:39817544(-) | chr21:42880007(+) | intronic |
| | | TMPRSS2-MGA | chr15:42042634(-) | chr21:42871284(-) | intronic |
| *s*140-n | - | | | | |
| *s*141-n | - | | | | |
| *s*143-n | - | | | | |
| *s*144-n | - | | | | |
| *s*145-n | - | | | | |
| *s*147-n | - | | | | |
| *s*148-n | - | | | | |
| *s*149-n | - | | | | |
| *s*150-c | - | | | | |
| *s*151-n | - | | | | |
| *s*152-n | - | | | | |

# 6 The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy

pmid: 28582575, doi: 10.1093/nar/gkx409

Björn A. Grüning [1,2,*], Jörg Fallmann [3], Dilmurat Yusuf [4], Sebastian Will [5], Anika Erxleben [1], Florian Eggenhofer [1], Torsten Houwaart [1], Bérénice Batut [1], Pavankumar Videm,[1], Andrea Bagnacani [9], Markus Wolfien [9], Steffen C. Lott [12], Youri Hoogstrate [10], Wolfgang R. Hess [12], Olaf Wolkenhauer [9], Steve Hoffmann [3], Altuna Akalin [4], Uwe Ohler [4,11], Peter F. Stadler [3,5,6,7], Rolf Backofen [1,2,8,]

*Nucleic Acids Research*, 45(W1):W560−W566, 2017

[1] *Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany* [2] *Center for Biological Systems Analysis (ZBSA), University of Freiburg, Germany* [3] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany* [4] *Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany* [5] *Institute for Theoretical Chemistry, University of Vienna, Austria* [6] *Max Planck Institute for Mathematics in the Sciences, Germany* [7] *Santa Fe Institute, Santa Fe, United States* [8] *BIOSS Centre for Biological Signaling Studies, University of Freiburg, Germany* [9] *Department of Systems Biology and Bioinformatics, University of Rostock, Germany* [10] *Department of Urology, Erasmus University Medical Center, Netherlands* [11] *Departments of Biology and Computer Science, Humboldt University, Germany* [12] *Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Germany*

Supplementary material: *none*

## Abstract

RNA-based regulation has become a major research topic in molecular biology. The analysis of epigenetic and expression data is therefore incomplete if RNA-based regulation is not taken into account. Thus, it is increasingly important but not yet standard to combine RNA-centric data and analysis tools with other types of experimental data such as RNA-seq or ChIP-seq. Here, we present the RNA workbench, a comprehensive set of analysis tools and consolidated workflows that enable the researcher to combine these two worlds. Based on the Galaxy framework the workbench guarantees simple access, easy extension, flexible adaption to personal and security needs, and sophisticated analyses that are independent of command-line knowledge. Currently, it includes more than 50 bioinformatics tools that are dedicated to different research areas of RNA biology including RNA structure analysis, RNA alignment, RNA annotation, RNA-protein interaction, ribosome profiling, RNA-seq analysis and RNA target prediction. The workbench is developed and maintained by experts in RNA bioinformatics and the Galaxy framework. Together with the growing community evolving around this workbench, we are committed to keep the workbench up-to-date for future standards and needs, providing researchers with a reliable and robust framework for RNA data analysis.

**Availability:** The RNA workbench is available at:
`https://github.com/bgruening/galaxy-rna-workbench`.

## 6.1 Introduction

Since recent advances in high-throughput sequencing (HTS) emphasized the importance and versatile role of (non-coding) RNAs, there is high demand for integrated computational analyses investigating RNA-mediated regulation. Previously existing workbenches (such as *miARma-Seq* [100] *RAP* [99] and the *UEA Small RNA Workbench* [101]) were focused on providing tools for the analysis of RNA deep sequencing data and do not contain RNA centric tools.

We addressed these needs by developing the RNA workbench. Based on the Galaxy framework [102] it combines a comprehensive set of tools for the analysis of RNA structures, RNA alignments, RNA–RNA and RNA–protein interactions, RNA sequencing, ribosome profiling, genome annotation and many more. So far, we integrated more than 50 RNA-related tools, including suites like the ViennaRNA package, covering this broad variety of use-cases (a complete list of tools can be found on GitHub). Every available tool works as a single building-block that can be connected with other tools to create computational pipelines. Datasets can be incorporated in a similar manner, facilitating an intersection of diverse data sources such as DNA methylation with RNA-seq experiments. Input and output datasets can be defined by the user, and can be as diverse as the adapted set of tools. Established data types for sequence and/or structure information are accepted as input. Output data types follow the same principle, can be converted to different formats, or ultimately used to draw plots and create figures. The workbench provides tools for visualizations of RNA structure datasets, such as dot-bracket strings, and RNA 2D or 3D structures. The workbench also covers a broad range of RNA secondary structure prediction and analysis tools such as RNAfold [220] or LocARNA [221, 222].



## 6.2 Goals of the RNA workbench

The main driving force behind the development of the RNA workbench is the goal to establish a central, redistributable workbench for scientists and programmers working with RNA-related data, and build a sustainable community around it. This platform is unique in combining available tools, workflows and training material, as well as providing easy access for experimentalists. Simultaneously, it serves as a central hub for programmers, which can easily integrate and deploy their existing or novel tools and workflows. The RNA workbench is based on three pillars: (i) a comprehensive set of RNA-bioinformatics tools, (ii) easy and stable dissemination via Galaxy and Docker and (iii) a set of predefined workflows and associated descriptions/training material. The latter is needed for two reasons: first, it facilitates the use of the RNA

workbench for researchers with limited bioinformatics experience, and second, it allows to integrate the workbench in the daily lab work by combining RNA-related analysis tasks with workflows for RNA-seq analysis.

### 6.2.1 Building on the shoulders of giants

In order to achieve long-term sustainability, we provide the essentials of our work on *BioConda*[1] and *BioContainers*[2] [223] for reproducible deployments of tools into Galaxy. Using easy-to-distribute packages for all tool dependencies also enables automatic continuous integration tests for all developed tools and the workbench. After a tool passes the tests and gets accepted it will be made available via an automatic deployment into the Galaxy ToolShed[3] [95]. From the ToolShed, Galaxy administrators can easily install desired tools and workflows.

### 6.2.2 Easily accessible and reproducible analysis platform



For the fast dissemination of the RNA workbench, as well as for an easy integration with other HTS analysis tasks, we implemented the RNA workbench within the Galaxy framework. A major advantage of relying on Galaxy as the core framework is that it is possible to leverage its scalability, which enables the RNA workbench to run on single CPU installations as well as on large multi-node high performance computing environments. Furthermore, Galaxy provides researchers with means to reproduce their own workflow analyses, enabling them to rerun entire pipelines, or publish and share them with others. The RNA workbench is containerized, *i.e.*, administrators can deploy it via *Docker*. That makes it possible to have all tool installation dependencies already resolved, while still keeping maintenance tasks to a minimum. The provided layer of virtualization also allows the handling of user-defined input data in a secure and compartmentalized way, a key requirement for researchers working on sensitive data (*e.g.* patient data in clinics). Running the containerized RNA workbench simply requires installing *Docker* and starting the Galaxy RNA workbench image. Furthermore, containerizing Galaxy enables a customized Galaxy instance with a selected subset of tools dedicated to specific data analysis tasks, while keeping deployment and installation simple.

[1] `https://bioconda.github.io`
[2] `https://biocontainers.pro`
[3] `https://toolshed.g2.bx.psu.edu`

## 6.3 RNA-Bioinformatics tools

In its current state, the RNA workbench includes more than 50 tools covering all aspects of RNA research. In a community effort, these tools will be kept up-to-date and adapted to future needs. New tools and new ways to visualize data provided to the user will also be integrated. A current overview of tools available in the RNA workbench can be found at: `https://bgruening.github.io/galaxy-rna-workbench/`.

In the following, we will highlight a few of the integrated tools.

The ***ViennaRNA*** package [220] consists of a suite of tools centered around the prediction of secondary structures of RNAs based on the thermodynamic Turner energy model. Thus, it covers prediction of optimal and suboptimal structures from single sequences as well as alignments, prediction of ensemble base pair probabilities, accessibility of sequences, and RNA–RNA interaction prediction. Importantly, predictions can be flexibly controlled by hard and soft structure constraints; the latter enables the inclusion of structure probing data.

***AREsite2*** [224] is a resource for the investigation of `AU`, `GU` and `U`-rich elements (ARE, GRE, URE) in human and model organisms. It provides information on genomic location, genomic context, RNA secondary structure context and conservation of annotated motifs in the whole gene body including introns. It is integrated into the RNA workbench via its REST interface, which provides search results directly in Galaxy for further analysis.

***LocARNA*** [221, 222] provides a comparative analysis of multiple (unaligned) RNAs by simultaneous folding and alignment, implementing a fast variant of the Sankoff algorithm. Beyond pairwise and multiple alignments, it computes reliabilities of alignment columns and provides very fast analysis by simultaneous folding and matching. Finally, *LocARNA* supports anchor and structure constraints, which improve its applicability in practice.

***doRiNA*** [225] is a database of RNA interactions in post-transcriptional regulation. The combined action of RNA-binding proteins (RBPs) and microRNAs (miRNAs) is believed to form the backbone of post-transcriptional regulation. *doRiNA* is implemented as data source tool inside the RNA workbench. This means that the Galaxy user is redirected to the post-transcriptional interaction database and can make selections using the optimized doRiNA interface. Once the selection is done, the data is streamed directly to Galaxy and can be freely analyzed with other tools.

The ***Infernal*** [226] tool suite can construct probabilistic models, also called covariance models (CM), that represent the sequence and structure of an RNA family from a multiple sequence alignment with consensus secondary structure. The covariance model can be used to find more members of this RNA family via homology

search.

***PARalyzer*** [227] generates a high resolution map of interaction sites between RNA-binding proteins and their targets. The algorithm utilizes the deep sequencing reads generated by the PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) protocol. The use of photoactivatable nucleotides in the PAR-CLIP protocol results in more efficient crosslinking between the RNA-binding protein and its target relative to other CLIP methods; in addition a nucleotide substitution occurs at the site of crosslinking, providing for single-nucleotide resolution binding information. *PARalyzer* utilizes this nucleotide substitution in a kernel density estimate classifier to generate the high resolution set of protein-RNA interaction sites.

***FuMa*** [228] can generate an integration report on predicted fusion genes from most RNA-seq fusion gene detection software. It automatically orders the result based on the frequencies of the fusion genes such that frequently predicted fusion genes can be extracted.

## 6.4 Workflows



One of the core concepts of the RNA workbench is the definition of standard workflows as a minimal set of building blocks around which a researcher can compose and tailor specific pipelines. For example, a researcher wants to analyze the effects of an RNA-binding protein (RBP) in regard to expression levels in wild-type compared to knockout or knockdown of the RBP of interest. In this case, one needs to combine the detection of differentially expressed genes in the two conditions with the information of publicly available CLIP-data, as provided for example by the *doRiNA* [225] database, to differentiate between direct and indirect targets. Workflows for the analysis of differentially expressed genes are part of the RNA workbench, as well as an interface to *doRiNA*, such that it becomes an easy task to design a new workflow combining these analysis steps.

In Galaxy, workflows are typically created in two different ways: (i) from an existing history, which stores all tools applied in a previous analysis together with all pertinent parameters, or (ii) from scratch, using a graphical editor via drag-and-drop of tools from the tool panel into the workflow editor. Within workflows, tools can be freely combined to ensure a maximum of flexibility in their usage and connectivity between different analysis steps, e.g. RNA structure analysis tools and RNA-seq data analysis. Various format converters embedded in Galaxy allow combining diverse analysis outputs. Easy sharing of workflows with other Galaxy users guarantees highly reproducible and transparent research. In other words, the workflows ensure that all

analysis steps, tools and parameters of an experiment are documented and visible to researchers, readers and reviewers. Workflows can also be submitted to the Galaxy ToolShed or *myexperiment.org* [229] for further distribution. The RNA workbench currently includes publicly available standard workflows for RNA data analysis, e.g. for RNA-seq. These workflows contain all required steps such as quality control, mapping, differential expression analysis, and visualization of results. Provided workflows can easily be extended or modified, e.g. to use other read mappers available in Galaxy.

In the following, we will describe two sample workflows, one closely related to the detection of ncRNAs, which is a common task in RNA-related research. The other workflow is related to the analysis of RNA-seq data and is often needed as a subworkflow for more complex analysis tasks. These workflows are well annotated and described in the RNA workbench and extended by Galaxy *Interactive Tours*.

### 6.4.1 Analysis of (unaligned) non-coding RNAs

An important task is to test for the existence of a functional structure in a non-coding RNA. However, the secondary structure of structured non-coding RNAs is not significantly more stable compared to random sequences [230]. Thus, putative functional structures can only be detected using information about conservation. Our workflow for non-coding RNAs performs the typical analysis steps required to detect conserved secondary structures, given a set of unaligned RNA sequences. It computes a sequence and a structure-based alignment by *MAFFT* [231] and *LocARNA*, respectively, and analyzes them with *RNAcode* [232] and *RNAz* [233] with appropriate parameter settings. *RNAz* and *RNAcode* both work on a given alignment. *RNAz* tests whether a consensus secondary structure is significantly conserved, whereas *RNAcode* differentiates coding from non-coding RNAs. Together these tools provide information, whether the RNAs are related and conserve a common secondary structure. In addition, a covariance model is built from the *LocARNA* alignment and subsequently used to search the given sequence database for RNAs with similar sequence- and structure-conservation. This workflow resembles the core of *RNAlien* [234], which is based on the same tools and is integrated into the RNA workbench. Going beyond the presented workflow, *RNAlien* automatically gathers sequences via homology search starting from a single sequence and constructs RNA family models in an iterative process.

To give an other example, in the context of $\mu$ORFs detection, RNA-seq analysis, the identification of non-coding RNAs with *RNAcode* and *RNAz* and the detection of transcription start sites can be used to determine new, short transcripts that are expressed and do not exhibit secondary structure conservation (i.e. are likely not

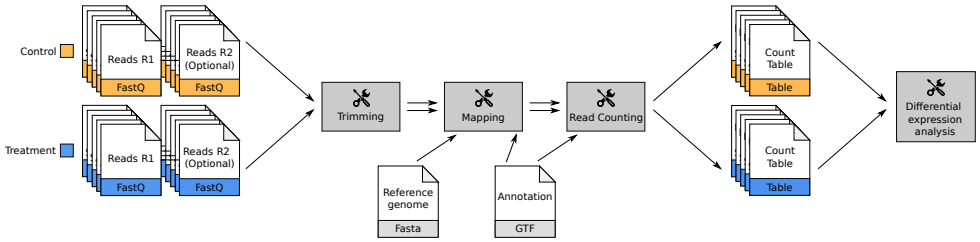



Figure 6.1: The workflow for analyzing RNA-seq data. The workflow tolerates single-end and paired-end reads derived from different conditions. It employs *TopHat2* for mapping and *htseq-count* to create the read counts. The final outputs contain read count per annotated gene for each condition and for each sequencing type.

functional ncRNAs). Subsequent analysis of Ribo-seq data can then provide additional evidence for a new transcript that may code for a small protein. For all these tasks, partial workflows and required tools are already integrated in our RNA workbench, which implies that it is easy to set up a new workflow for a more complex task.



### 6.4.2 RNA-seq analysis: trimming, mapping and read count

As mentioned before, the analysis of RNA-centric data like CLIP-seq requires the combination with other type of data, and very often RNA-seq. For that reason, we provide a standard RNA-seq workflow that can easily be combined with other workflows. The RNA-seq workflow (as shown in Figure 6.1) takes a list of RNA-seq datasets as input and successively executes a series of analysis steps - adapter & quality trimming, mapping to a reference genome and read count per annotated gene. The input allows two conditions, e.g. treatment versus control and it also accepts single-end and paired-end reads for each condition. At the trimming step, the workflow employs *Trim Galore!* [235, 236] to perform adapter trimming. Then, *TopHat2* [237] is used to map the trimmed reads against the reference sequences, which should be provided by the user. As last step, the workflow executes *HTSeq-count* [52] to generate read counts per annotated gene for each condition and for each sequencing type. A reference annotation in Gene Transfer Format (GTF), *e.g.* provided by Ensembl [238], is required at this step. The final read counts can be used for the downstream assessment of differential expression using tools like *DESeq2* [55]. The current workflow can serve as a template that can be modified by the user according to different needs, for instance, replacement of tools or modification of the wrapping strategy.

## 6.5 Implementation

The workbench is implemented as portable virtualized container based on Galaxy. The Galaxy framework allows for reproducible and transparent scientific research which makes it easy to access, deploy and scale−conceptualized as a web service. The foundation of the workbench container is a generic Galaxy *Docker* instance (`https://bgruening.github.io/docker-galaxy-stable/`). On-top of this, pre-configured Galaxy tools can be automatically installed from the Galaxy ToolShed using the Galaxy API *BioBlend* [239]. In Galaxy, tool dependencies are automatically resolved via *BioConda*, which is the bioinformatics channel for the *Conda* package manager. *BioConda* facilitates software packaging and enables installation at a user level, keeping track of different versions of the same software in virtual environments. These features are in line with the scope of Galaxy; maintaining large numbers of dependencies in a reproducible way. Therefore, all available tools within the RNA workbench are also distributed as *BioConda* packages and *BioContainers*, which are persistent, frozen, containerized versions of *Conda* packages. The RNA workbench ships with a variety of tools, tours, documentation, workflows and data that have been added as additional layers on top of the generic *Docker* instance. During development, the software has been tested extensively in a continuous integration setup (CI) at different levels: Galaxy itself, tool integration in Galaxy (IUC, galaxytools channels), dependencies (*BioConda*) and at the workbench level. Together with a strict version management on all levels, this contributes to a high degree of error-control and reproducibility. The RNA workbench started in January 2015 - with constant development over 2 years, and extensive testing in local and public Galaxy instances, such as the Freiburg Galaxy instance, the MDC instance in Berlin and Erasmus MC's Galaxian. More than 500 users accessed the RNA tools during the last two years and the virtualized *Docker* instance was already downloaded >500 times. Moreover, due to an open and transparent development process, there is a growing community that contributes to our workbench, which guarantees the sustainability of the RNA workbench project and maintenance of the underlying *Docker/rkt* images.



## 6.6 Using the RNA workbench

Installation: The RNA workbench can be installed under OSX and Windows using the graphical tool Kitematic (`https://kitematic.com`), or with the following Linux command:

```
docker run -d -p 8080:80 bgruening/galaxy-rna-workbench
```

This installation is production-ready and can be configured to use external computer clusters or cloud environments. Due to the very modular system, it is also possible to install all or only a few tools of the RNA workbench on available Galaxy servers. Just get in contact with your local Galaxy administrator. When using the RNA workbench *Docker* image, the user has full administration rights, which enables customization independent of potential user restrictions.

### 6.6.1 Training

For self-empowering the user, documentation and training of the RNA workbench are important. We included an extensive set of documentation in traditional formats, *e.g.* tool descriptions and 'README' files.

We also provide training sessions around HTS data analyses and RNA-seq data analysis. The training materials ranging from the introduction to Galaxy, to usage and maintenance of Galaxy and the RNA workbench are freely accessible for self-paced studies at the Galaxyproject Github repository (`https://galaxyproject.github.io/training-material`). This training material is constantly improved and extended in an international community effort, including ELIXIR and EMBL. For HTS data analyses we provide training as a specific introduction to the topic with self-explanatory presentation slides, a hands-on training documentation describing the analysis workflow, all necessary input files ready-to-use via *Zenodo*, a Galaxy *Interactive Tour*, and a tailor-made Galaxy *Docker* image for the corresponding data analysis.

To provide an even more intense training experience within the RNA workbench, we also included interactive training such as the Galaxy *Interactive Tours*. Such tours guide users through an entire analysis in an interactive and explorative way. It combines advantages from training videos and detailed protocols. Production of training videos is very time-consuming and tend to become outdated very soon, due to tool version changes or renewed workflows. In contrast to conventional screencasts, a Galaxy *Interactive Tour* can be easily updated and improved to guide the Galaxy user step-by-step, e.g. through a whole HTS analysis starting from uploading the data to using complex analysis tools. Exemplary, the RNA workbench currently integrates two Galaxy *Interactive Tours*. The first one introduces a new user to the Galaxy interface and its usage with an RNA-seq example dataset. The second one illustrates secondary structure prediction of RNA molecules using parts of the *ViennaRNA* package. To show how Galaxy *Interactive Tours* can interactively guide users through the necessary steps of HTS analyses, the tours are also provided as online screencasts.

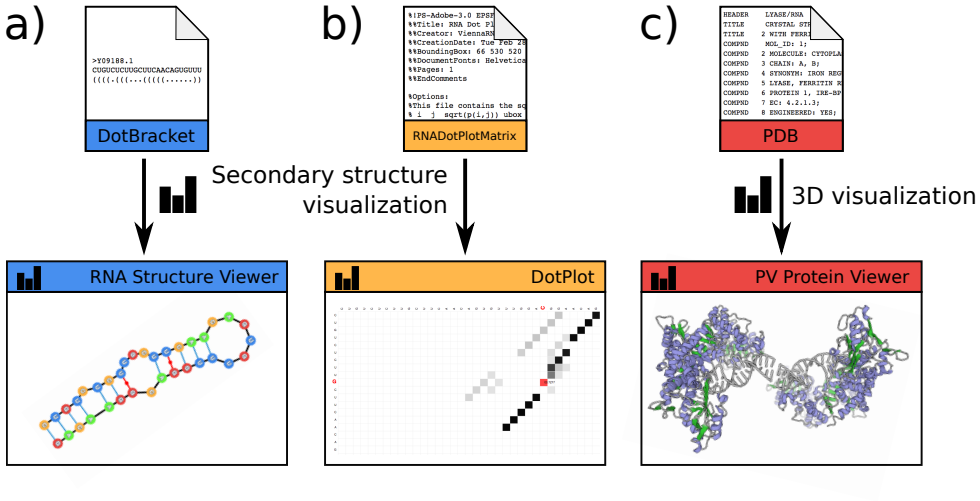



Figure 6.2: RNA structure visualization: the figure shows visualization for an *IRE1* RNA sequence, retrieved from Rfam database [240], via different backends integrated into the toolbox. (A) Secondary structure encoded in dot-bracket notation, can be displayed by the RNA structure viewer. (B) Base pairing probabilities are visualized as DotPlot. (C) Tertiary/Quaternary structure information encoded in protein-database format is rendered via Protein Viewer

### 6.6.2 Visualization

Following data reduction as a key element of explorative research, there is a need for meaningful figures and visualizations that summarize results. The RNA workbench includes standard interactive plotting tools to draw bar charts and scatter plots from all kinds of tabular data and allows for connections to *Integrated Genome Browser* [241] and *UCSC* [242] like any other Galaxy instance. On top of this, we included three visualizations specific to RNA research. An interactive DotPlot visualization for secondary structures in EPS format (Figure 6.2c), a 2D visualization for the common dot-bracket format (Figure 6.2a) and a 3D visualization capable of visualizing PDB, SDF and MOL files containing three-dimensional coordinates (Figure 6.2c).

## 6.7 Community

The RNA workbench project is an open source project that strives to create a community interested in accessible and reproducible RNA-related research. Knowing that real sustainability can only come true with a strong community we are aiming at more open participation, reward, and inclusion. We are working together with Galaxy, *BioConda*, *BioContainers* and *BioJS* and coordinating efforts to not reinvent the wheel

but joining forces to create the new generation of bioinformatics infrastructure together. In the RNA workbench community, we practice the organizations on GitHub, IRC, and Gitter and welcome everyone to contribute on every level to improve the entire stack from documentation to tools and scientific workflows. Support will be provided through the same channels.

## 6.8 Discussion

In this work, we present the RNA workbench, maintained and developed by a constantly growing community. The presented workbench is unique as it allows to easily combine RNA-centric analysis with other types of experiments. It provides a set of tools, each one being available as *BioConda* package as well as a Docker/rkt container (*BioContainers*). Based on the Galaxy *Docker* project, the proposed web server is more than the sum of its parts. It offers a comprehensive virtualized RNA workbench that can be deployed on every standard Linux, Windows and OSX computer, but can at the same time employ high-performance- or cloud-computing infrastructure.



Major advantages of our approach to deliver a dockerized workbench for RNA centric analysis are the ease of installation, the high number of pre-included tools, the flexibility in regard to extension with other tools and workflows and the high reproducibility and transparency of workflows. All tools that are available on the Galaxy *Toolshed* can be installed along with their automatically resolved dependencies with a single click in the Galaxy interface. Best practice pipelines for the analysis of RNA-seq data are provided with the *Docker* image and can easily be modified, extended or combined with other analysis pipelines via Galaxy's workflow editor GUI.

The RNA workbench was designed as a community project, and as such it is easy for users to contribute to the workbench with workflows, new tools and training material, keeping the workbench up-to-date and valuable for research. Moreover, all components such as tools, workflows, visualizations, interactive tours and training material can be easily integrated into any available Galaxy instance for teaching, learning or exploratory purposes.

The main difference to existing solutions such as *miARma-Seq* [100], *RAP* [99] and the *UEA Small RNA Workbench* [101] is that our RNA workbench combines the realm of RNA-centric analysis on sequence and structure level with modern high-throughput sequence analysis. In this regard we provide well established tools for RNA structure prediction, analysis and visualization together with read mappers and expression analysis tools for HTS analysis.

## Acknowledgements

We thank the de.NBI and ELIXIR projects for supporting bioinformatics infrastructure. Thanks also to the Galaxy community, especially to the Freiburg Galaxy Team, for developing, maintaining and supporting this great framework. We also like to acknowledge the *BioConda* and *BioContainers* community for setting new standards in reproducible software deployments. Thanks also to the BioJS community for great discussions about scientific visualizations and how we can make them more accessible. Moreover, the authors acknowledge the support of many upstream developers that helped us to integrate their tools into the RNA workbench and accepted patches.

## Funding

Collaborative Research Center 992 Medical Epigenetics [DFG grant SFB 992/1 2012]; German Federal Ministry of Education and Research [BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)]; Center for Translational Molecular Medicine (CTMM), TraIT project [05T-401 to Y.H.]. Funding for open access charge: German Government.

# 7 Discussion

We set out to find prostate cancer (PCa)-associated RNA molecules in RNA-sequencing (RNA-seq) data, that can potentially be used as biomarker, by making use of new computational methods. Research on RNA is nowadays mostly performed with RNA-seq. What makes RNA-seq interesting is its capability to analyse RNA sequences independent of probes, allowing detection of new transcripts, mutations and revealing gene structures [58]. Corresponding data analysis requires custom software and therefore, software is a determinant for the outcome and plays a key role in the entire analysis. We have used small RNA-seq data to detect small non-coding RNAs and used paired-end RNA-seq prepared with random hexamer primers to detect fusion genes including corresponding genomic breakpoints.

## 7.1 Small RNA-seq in prostate cancer

One of the goals in this thesis was automated annotation and quantification of small non-coding RNAs (sncRNAs) and search for associations with presence and aggressiveness of PCa. Therefore, we developed a new computational analysis method, *FlaiMapper*. In **chapter 2: FlaiMapper**, it is demonstrated that *FlaiMapper* predicts the miRNA 5′ and 3′-ends extremely well when small RNA-seq was performed. In **chapter 3: sdRNAs**, *FlaiMapper* was used to analyse the small RNA content in prostate and PCa samples. Global deregulation of small RNA processing was observed in PCa. It revealed the presence of C/D-box snoRNA-derived small RNAs, significantly overexpressed in PCa. It was also found that abundance of specific C/D-box snoRNA-derived RNAs correlates positively with aggressiveness of cancer, suggesting they may also be useful as prognostic markers. That C/D-box snoRNA-derived RNAs are upregulated in PCa sheds new light on small ncRNAs and PCa, and highlights that the methodology can discover novel molecules that are associated with cancer.

*FlaiMapper* was also used to analyse the composition and expression of tRNA-derived fragments (tRFs). This work revealed that several tRFs are differentially ex-

pressed in PCa [25]. Three of these tRFs were validated with qPCR and confirmed that the predicted molecules exist and that the computer model provides accurate results.

The algorithmic challenge of *FlaiMapper* is detecting start and end positions of small ncRNAs in sequence alignments. The start and end positions of sequenced small RNAs have some degree of variability, meaning they may be a bit longer, shorter or shifted. Detection of these positions needs to be lenient enough to take this variability into account. The variability of the start and end positions that together determine a small RNA, follow a certain distribution. We found no basis on which we can assume that a start position of a small RNA determines the end position or *vice versa*. Therefore, in contrary to the earlier reported method *blockbuster* [147], we explicitly did not use a statistical distribution that models a ncRNA, or its alignment, as a whole. Instead, start and end positions are determined separately and linked together afterwards. The ends of a miRNA are known to be processed by different nucleases, `DROSHA` and `DICER`. The results of our analysis show that the variablility of miRNAs is higher at the 3′-ends than the 5′-ends. Therefore, separate computational determination of start and end positions is a more natural approach, as it reflects the variability of start and end positions as result of different nucleases better.

A new direction in which *FlaiMapper* could be useful is the analysis of small RNAs after knockdown or immunoprecipitation of protiens involved in RNA processing [129]. In such data, *FlaiMapper* could also be used to determine if a small RNA is (post-)processed differently in the absence or presence of certain proteins.

It was found that some RNAs derived from the same host C/D-box snoRNA are overlapping each other, which indicates that they can not both be derived from the same molecule. One explanation could be that their host snoRNA may exist in different conformations, which gives rise to different fragments. Given that specific C/D-box snoRNAs molecules are upregulated in PCa, it may be important to find if there are different conformations and whether they are associated with different functions. A possible direction to look into could be the 2D and 3D structure of C/D-box snoRNAs. For example, using *in silico* RNA folding prediction methods it could be investigated whether C/D-box snoRNAs contain multiple energetic (sub)optima or whether the overlapping small RNAs are found more frequently within certain secondary structure elements (such as hairpin loops, bulge loops or stems). C/D-box snoRNAs contain a secondary structure element named the kink-turn (K-turn), consisting of multiple non-canonical basepairs in a kinked tertiary structure. K-turns cannot be predicted by classical minimum free energy RNA folding algorithms because of the multiple, energy-dependent, non-canonical base pairs. This prevents in-depth analysis of the 2D structures of C/D-box snoRNAs. In an earlier report ([243]; unpublished), I pro-

posed an adaptation to a classical RNA folding algorithm, that makes prediction of 2D structures including energy-dependent multiple-bond sub-structures such as K-turns and loop-E-motifs possible. This method was limited to *in silico* predicted energy values that correspond to the K-turns, as experimentally determined values were not available. Recently, progress was made in the experimental determination of such values [244], which allows further research in the direction of 2D structure prediction of C/D-box snoRNAs.

To find evidence for possible functions of C/D-box snoRNA-derived small RNAs, we have investigated whether the detected small RNAs are the highest conserved regions of the host C/D-box snoRNA ([245]; *unpublished*). The results indicated that they are not in general the highest conserved regions of the host C/D-box snoRNAs.

The accuracy of *FlaiMapper* was estimated by evaluating how accurate it predicts *miRBase* annotations. Since the purpose of *FlaiMapper* was to define in particular small RNAs other than miRNAs, *FlaiMapper* should ideally be benchmarked also on other small ncRNAs. For certain types of small ncRNAs, such as piRNAs and tRFs, annotations are available (in *piRNABank* [246] and *tRFdb* [247] respectively) and can be used for additional benchmarking of *FlaiMapper*.

The filtering procedure of *FlaiMapper* makes use of parameters that follow a probabilistic distribution related to the variability of the small ncRNAs' start and end positions. The parameters of *FlaiMapper* have been estimated by investingating several examples, mostly miRNAs and C/D-box snoRNA-dervied RNAs, and may therefore be biassed towards these RNA types. A currently developed tool named *STARPA*[1], embeds *FlaiMapper* in its pipeline and makes use of adapted filter parameters. By investigating the aligments of valided small RNAs, it is possible to estimate optimal parameters computationally. Therefore, computational parameter optimization using validated small ncRNAs, prompts future work. Ideally, such validation makes use of as many different types of small RNAs as possible, including data from *piRNABank* and *tRFdb*, to minimise a bias towards certain small RNAs types.

Although *FlaiMapper* was designed for the analysis of small RNAs in particular, it should in principle also allow detection of other types of RNAs. The only requirements for the algorithm to work appropriately are that (i) the analysed RNAs represent complete (not fragmented) molecules and (ii) that they are single-end sequenced. Given that the length of sequencing reads keeps increasing, it is plausible that the software may become useful for the detection of larger RNAs.

Small RNAs detected and quantified with *FlaiMapper* have the potential to be used as biomarkers for PCa. For example, expression levels of `SNORD44`-5′-fragment were higher in malignant than in benign and normal tissue. Expression levels of

[1] `https://github.com/luidale/starpa`

SNORD78-3′-fragment were observed in a subset of patients that progressed into metastatic disease. The correlation of SNORD78-3′-fragment expression with prognosis makes it potentially useful as prognostic biomarker. Given the high number of differentially expressed small RNAs, it could be beneficial to investigate whether there are small RNAs that are only differentially expressed in mutually exclusive subsets of cancer samples. They could then be used to create combined expression profiles with improved statistical power.

## 7.2 Fusion genes in prostate cancer

Fusion genes are found in different types of cancer. For instsance, TMPRSS2-ERG is found in approximately ∼50% of the PCa samples. A fusion gene analysis pipeline for RNA-seq data typically reports the predicted fusions for only a single sample, but lacks an aggregation step in order to find recurrent fusion genes across samples. Thus, as a final step, multiple output files need to be aggregated in order to find recurrent fusion genes. Recurrent fusion genes can be detected with the R-package Chimera [48], but it operates so slow that it is practically not applicable for a high number of samples. It also requires users to write additional R scripts.

There are quite a large number of tools available for detection of fusion genes in RNA-seq data, but there is limited overlap in their outcome and none is superior in all aspects [75]. Reporting fusion genes that are detected by at least a number of tools can be helpful to increase confidence. In **chapter 4: FuMa**, we propose a method that reports overlap in the results of publicly available RNA-seq fusion gene detection tools. Such reports are helpful for the scenarios described earlier: (i) it can be used to find recurrent fusion genes in a cohort of samples and (ii) it can be used to aggregate at the sample level such that a combination of results from different tools can improve and prioritise the outcome.

RNA-seq data from ribo-depleted total RNA prepared with random hexamer primers for cDNA synthesis (random primed RNA-seq) is rich in both polyadenylated mature mRNA and non-polyadenylated transcripts, such as pre-mRNA. In contrast to most mRNA, pre-mRNA contains both introns and exons. Therefore, when a genomic breakpoint of an expressed fusion gene is located within an intron, the corresponding pre-mRNA can be used to detect the DNA breakpoint, in RNA-seq data. To achieve this, we designed *Dr. Disco* (**chapter 5: Dr. Disco**), an analysis tool that looks at the entire genome from a graph perspective. Rather than restricting itself to genes or even exons, it is capable of determining expressed rearrangements between any loci in the reference genome. The method was applied on a cohort of 51 PCa samples, and revealed the genomic breakpoint of TMPRSS2-ERG in 29 of 32 TMPRSS2-ERG positive

samples.

In the (fusion) gene structure of TMPRSS2-ERG, the first two exons of TMPRSS2 and exons 4 and 5 of ERG are typically included in the fusion transcripts [123, 248]. The predicted exon-to-exon junctions detected with *Dr. Disco* fit with the proposed structure and, in addition, the corresponding genomic breakpoints detected with *Dr. Disco* are in line with their corrsponding exon-to-exon junctions. DNA-seq analysis revealed a hotspot of TMPRSS2-ERG breaks in ERG intron 3 [215]. The detected genomic breaks in our RNA-seq analyses show the same hotspot, indicating the validity of the results. It was assessed whether there is a correlation in the location of the TMPRSS2 and ERG breakpoints, but no relation between the break in TMPRSS2 and the break in ERG or *vice versa* was found, suggesting these events take place independently random.

The first step in the *Dr. Disco* pipeline is alignment to a reference genome. Although both reads of a read pair can perfectly map a reference genome, they may still span a large genomic distance, be mapped to different chromosomes or be mapped in opposite strands. It may also happen that one read of a pair gets split into two pieces, which are both perfectly aligned. Although such reads align without without any mismatches to the reference genome, they cannot originate from a canonical gene and are therefore marked as *discordant read*. Because fusion genes cause such reads, *Dr. Disco* uses discordant reads (provided by aligner STAR) as input, similar to STAR-Fusion [200]. However, discordant reads do not only originate from fusion genes but can also originate from other transcripts or technical artefacts (Table 7.1).

Discordant reads that originate from reasons provided in Table 7.1, may result in more false positives and need to be separated from those caused by actual fusion events, in order to obtain a high accuracy. A part of this can be addressed in the alignment step and therefore it is trivial to ensure the most optimal alignment settings are used. Fusion genes that are not detected due to low read depth, alignment artefacts or misclassification, prompts future work. What further could be investigated is up to which read length the software is sufficiently accurate. The smaller the reads are, the larger the chance they map to arbitrary locations on the genome, which will result in more discordant reads and consequently more false-positives. In particular with the lenient fusion settings of STAR, the search space increases and the chances on alignment artefacts increases. Testing these limits could for example be performed by using a dataset with long reads and known true-positive fusion genes and by systematically truncating the reads and investigating the differences.

*Dr. Disco* can investigate the entire reference genome and is capable of identifing intronic and intergenic breakpoints. A disadvantage of using reference genome-based alignments as input, is that fusion genes within highly mutated regions result in many alignment mismatches, which is what might have happened with the unde-

Table 7.1: Sources of discordant reads other than fusion genes

| Source | Description |
|---|---|
| Low entropy, satellite and small repeat sequences and homopolymers | The reference genome contains many low entropy sequences like poly-A, poly-T and poly-CA repeats. Due to their low complexity, they have (close to) identical genomic copies. |
| Repeat regions, pseudogenes, transposable elements, retrotransposons and homologs | There are various sources of highly similar sequences with an entropy higher than the small repeats stated above. For example, pseudogenes are near identical copies, which may result in mates mapping to different loci. |
| Alternative loci | Reference genome hg38 has many alternative loci representing subpopulation specific genomic variations. These loci are annotated as distinct chromosomes in the reference genome. Often, reads near the start and end point of the alternative locus have one mate on the alternative locus while the other mate is on the major chromosome. They will be marked as discordant since the aligner interprets this as an interchromosomal junction. |
| Small fragments ($<$ 2x read length) | RNAs with insert sizes smaller than twice the length of the sequenced read, result in pairs with overlapping aligned mates, and will be marked as discordant by the aligner. |
| Circular RNAs | Circular RNAs are formed by a covalent bond, often between an exon at the end of a gene to an exon at the beginning of a gene. As a result, reads are split in an opposite orientation and are marked as discordant by the aligner. |
| Read-throughs | Read-throughs or genes that are incorrectly annotated as separate genes may result in discordant reads. |
| Immunoglobulin recombinations | The V(D)J recombination is a natural system of genome recombination and, if expressed, it may result in discordant reads. |
| Splicing of not annotated genes or exons | Spliced reads that belong to new genes and new exons with non-canonical splice junctions or spanning a large genomic distance can be marked as discordant reads. |

tected intronic deletions (chapter 5). Investigating alignment settings to solve this issue prompts future work.

*Dr. Disco* can determine genomic breakpoints of fusion genes provided that they are expressed and sufficient pre-mRNA is sequenced. To get a better understanding of the number of fusion genes that are indeed expressed, it is part of our ongoing work to integrate RNA-seq with DNA-seq results. We expect to resolve how many arrangements involve intergenic regions and whether they are indeed spliced, even though they are located outside annotated genes. In addition, fusions that only take place at the RNA level, such as trans-splice isoforms, circRNAs and read-throughs, will be further investigated.

The analysis in chapter 5 was limited to the genomic regions of `TMPRSS2` and `ERG`, which span only a fraction of the entire genome. Although the analysis in this region of the genome was performed in a reasonable time, it is clear that for a full genome analysis this is not the case. The algorithm is quite complex, meaning that a substantial number of computations are needed to obtain the results. During development, several improvements to the calculation time have been made but it is plausible that the optimisation is getting close to its limits. Further optimisation of *Dr. Disco* will require more development time as well as computational resources. There are two preferred directions for further speed performance optimisations. The first is inclusion of a gene annotation (e.g. GTF file) at the start of *Dr. Disco*, which may allow reducing or removing the '*merge_overlapping_subnets*' functionality. The second is an implementation in a programming language that gives more control over memory and data structures such as C, C++ or Java. This could for example be achieved by integrating the ported code directly into aligner STAR.

Using *Dr. Disco*, additional fusions such as intronic deletions and fusions to non-gene regions can be found. Detected DNA breakpoints may provide information that explain how a junction arose and how the fused gene structure is composed. Detection of DNA breaks and fusions (or splice isoforms) involving intergenic regions are a new asset to RNA-seq analysis, but only possible when the RNA-seq library includes total RNA, for using random primed RNA-seq. This limits the use of poly-A-purified RNA and oligodT-primed RNA-seq data and we therefore recommend not to use such mRNA RNA-seq protocols for fusion gene analysis.

## 7.3 Challenges in computational methods

Besides implementing new computational methods to address individual research questions, it is important that other scientists get access to the software tools to use them on their own data and to improve them. This way, like publications, also

the software tools serve as building blocks of the foundation of science, upon which science can build further on. To do so, software and data should ideally follow the FAIR principles (findable, accessible, interoperable and re-usable) [125]. With this in mind, we have built a toolbox for RNA related research (**chapter 6: Galaxy RNA Workbench**). In *the Galaxy RNA workbench* the tools presented in this thesis, *Dr. Disco*, *FlaiMapper* and *FuMa*, have been integrated together with closely related tools such as STAR.

Ideally, scientific software is open source and written in a modular way, which is beneficial for designing pipelines and platform integration. For modularity it is important that there are clear agreements about input and output, about data formats and standardised code libraries that enforce making use of correct implementations of such standards where possible. In bioinformatics, companies have often provided specifications that are typically of commercial interest, but lack flexiblity outside the scope of the company's interest. It also happens that specifications turn out not to be future-proof. It would therefore be convenient to have an international independent authority that either proposes new conventions or organises contests in which they review submitted proposals, to generate a better IT fundament for computational research. This may provide a much stronger fundament for bioinformatics resulting in more robust research oriented software and will also benefit further development of the Galaxy RNA workbench.

## 7.4 Future perspectives

The outlined methods to identify novel small RNAs and fusion events, have been successfully applied on PCa data. To learn more about small RNAs and fusion genes, it would be of interest to setup a pan-cancer study, at a scale comparable with TCGA [107, 108] or ICGC [249]. Since the data used in chapter 3 has been generated, a large number of new small RNA-seq datasets have been published [66]. Analysing these new data to generate pan-cancer atlasses could reveal possible fusions or small ncRNAs that are specific for certain types of cancer or tissue or are common across different types of cancer or tissue. Results on small ncRNAs would fit ideally in YM500, a database for small RNA sequencing in human cancer research [66], and the fusion genes and transcripts in the COSMIC database [105]. The methods themselves have proven to be of value for research and have been integrated in an RNA oriented workbench. Before such a workbench can be used in a clinical setting, a risk assessment must be performed and must comply with the ISO-27001 standard.

In this thesis, we have demonstrated that RNA can be analysed with sophisticated techniques that allow to discover new and cancer specific RNAs. However, the results

presented are not yet directly applicable in the clinic. Certain C/D-box snoRNA-derived ncRNAs and tRFs [25] have been found to have a diagnostic and/or prognostic value. But before these potential biomarkers can be used as actual biomarker, further research on their predictive power and reproduciblility of the results is necessary. The RNA-seq data generated and processed in the presented work, are all taken from tissue RNA. Whereas these tissue samples are taken from radical prostatectomy material, biomarkers are tested on tissue samples taken from biopsies. Because biopsies may cause severe complications, it would be more convenient to have liquid biopsies from urine, blood serum or blood plasma instead. Since RNAs are present in such biofluids [250], it is important to translate the tissue based RNA biomarker assays to urine or blood-based assays [251].

Although RNA is the point of focus in all chapters, the proposed computational methods make use of diverse techniques such as graphs (*Dr. Disco*), set theory (*FuMa*) and peak detection (*FlaiMapper*). This indicates that RNA-seq data is versatile and that corresponding analysis requires to look at it from different perspectives in order to make new findings. After having successfully shed light on small RNAs and fusion genes, there are still certain topics that remain underexposed. Previous RNA-seq research has in particular been focusing on gene expression and splicing. Topics that deserve more attention and probably require new computational methods are circular and possibly double stranded RNAs. Another interesting direction is to investigate RNAs that we cannot unambiguously map to the genome. These might contain RNAs that originate from other (micro) organisms or types of RNA we are not (yet) familiar with. In addition, looking into the 2D and 3D structure, RNA-editing and RNA modifications may reveal new insights in RNA.

Last but not least, RNA analysis requires a gearbox containing sophisticated tools that (i) offer solutions for state-of-the-art techniques, (ii) can keep up with the colossality of the data and (iii) have a high level of user-friendliness with the main purpose to improve productivity. A typical group of tools where I do not have the belief that they meet with these criteria are genome browsers, while they should be the home portal of genomic research, where all information should come together. They are typically slow, contain bugs or even crash, cannot handle large datasets, are barely customizable (no 'favorite genomic locations', no way of creating project structures), are graphically primitive, cannot view branchpoints modelling structural variants, are sometimes implemented in webbrowsers despite the requirement to cope with very large datasets, require unnecessary complicated configurations or require manual installation of reference data. I would be honored to contribute to these shortcomings as one of my following missions.

# 8 Summary

## 8.1: Summary

## 8.2: Samenvatting

## 8.1 Summary

Cancer finds its origin in DNA changes, which have consequences at the RNA and protein level. With the next generation sequencing technology, changes in both DNA (DNA-seq) and RNA (RNA-seq) can be analysed on a large scale. In contrast to DNA, RNA molecules are not only static information carriers but fulfill different functions in the cell and therefore have the potential to serve as a good biomarker. By using RNA-seq, new RNA molecules can be detected, potentially including new biomarkers. However, because the data is so colossal, analysis requires the use of different computer programs. For various applications, such as detection of fusion genes and small ncRNAs, current tools are not sufficient and new or better solutions are needed. Because there is a lack of diagnostic / prognostic biomarkers for prostate cancer, one of the most common types of cancer in men, prostate cancer is an ideal model system for the development of these new analysis methods. In this thesis we describe new computational methods for the analysis of RNA-seq data and demonstrate how it can be used to find potential new biomarkers in prostate cancer samples.

In chapter 1, general information about DNA, different types of RNA, cancer and related mutations including fusion genes is provided. Also, more information about prostate cancer and the recurrent fusion gene `TMPRSS2-ERG` is provided. In addition, various recent technological developments regarding next generation sequencing and computational analysis are explained. This is followed by an ideological explanation of software modularity and the integration of bioinformatics software and the added value of corresponding data standards.

Besides protein coding mRNA, cells also contain many different non-protein coding RNA molecules (ncRNAs). This includes several small ncRNAs, which can be detected by small RNA-seq. For the analysis of small ncRNAs of which typically no annotation is available, no corresponding detection methods were available. In chapter 2, it is explained how such small ncRNAs can be discovered using the proposed new method *FlaiMapper*. Subsequenctly, in chapter 3 it is described how this method was applied to samples of different stages of prostate cancer. Several new RNAs, in particular C/D-box snoRNA-derived RNAs, were associated with prostate cancer and correlated with prognosis. This association also demonstrates the relevance and precision of the analysis method.

Fusion genes are often found in prostate cancer. Detection of fusion genes by means of RNA-seq is possible and several programs have been developed for this purpose. The overlap between these programs is limited and it has been recommended to use multiple tools and integrate these results to increase confidence. In chapter 4 we present the method *FuMa* that integrates and reports the overlap of different RNA-seq

fusion gene detection tools.

In most protocols used to prepare RNA for sequencing, mRNAs are selected based on the presence of poly-A tails. An increasingly popular alternative is to make use of random hexamer primers in the reverse transcriptase step. In data produced in this manner, non-polyadenylated RNAs such as lncRNAs, circRNAs and pre-mRNAs will also be sequenced. pre-mRNA molecules do not only consists of exons but also of introns. DNA breaks of fusion genes are mainly found in introns. Therefore, in random hexamer-primed ribo-depleted total RNA-seq data, it should be possible to determine a DNA break of a fusion gene by analysing pre-mRNA-derived sequencing reads. Computer programs able to take this extra layer of information into account did not exist. In chapter 5, we describe *Dr. Disco*, a computational method that takes the presence of pre-mRNA into account and is able to reveal DNA breaks. We demonstrate this on the basis of 51 prostate cancer samples in which we find hotspot regions of DNA breaks in `TMPRSS2-ERG` fusions, regions that are in agreement with literature.

In chapter 6, we emphasise that well written and thoroughly tested software is important for the integration in larger systems. Availability of such public domain software allowed, as a community effort, to integrate many RNA-related analysis tools into one, free and open-source, platform, the *Galaxy RNA Workbench*. This provides easy access to many RNA-specific visualisations and analysis tools, all provided with a generic graphical interface.

In summary, we have devised several new methods to discover new, cancer-related, RNAs that have the potential to be used as biomarkers. In addition, all proposed methods were ultimately integrated into the *Galaxy RNA workbench* that has been made available to other researchers, free of charge.

## 8.2 Samenvatting

Kanker vindt zijn oorsprong in veranderingen in het DNA, welke vervolgens gevolgen hebben op het RNA en eiwit niveau. Met de next generation sequencing technologie kunnen veranderingen in zowel het DNA (DNA-seq) als RNA (RNA-seq) op grote schaal geanalyseerd worden. In tegenstelling tot DNA, zijn RNA moleculen niet uitsluitend statische informatie dragers maar vervullen verschillende functies in de cel en hebben mede daardoor de potentie om als goede biomarker te dienen. Door gebruik te maken van RNA-seq kunnen nieuwe RNA moleculen gedetecteerd worden, waaronder potentiële nieuwe biomarkers. Echter, doordat de data zo kolossaal is, vereist analyse ervan het gebruik van verschillende computer programma's. Voor verschillende toepassingen, zoals detectie van fusie-genen en kleine RNAs, zijn de huidige computerprogramma's niet toereikend en zijn nieuwe of betere oplossingen nodig. Omdat voor prostaatkanker, een veel voorkomend type kanker bij mannen, een gebrek is aan diagnostische/prognostische biomarkers, is dit een ideaal modelsysteem voor de ontwikkeling van deze nieuwe analyse methoden. In deze scriptie beschrijven we nieuwe computer methoden voor de analyse van RNA-seq data en demonstreren we in prostaatkanker hoe we hiermee potentieel nieuwe biomarkers kunnen vinden.

In hoofdstuk 1 wordt algemene informatie gegeven over DNA, verschillende typen RNA, kanker en bijbehorende mutaties waaronder fusie-genen. Ook wordt meer uitleg gegeven over prostaatkanker en het daarbij vaak voorkomende fusie-gen `TMPRSS2-ERG`. Tevens worden verschillende recente technologische ontwikkelingen met betrekking tot next generation sequencing en computationele analyse toegelicht. Aansluitend daarop volgt een ideologische uiteenzetting over modulariteit en integratie van bio-informatica software en de meerwaarde van bijbehorende data standaarden.

Naast eiwit coderend mRNA bevatten cellen ook vele verschillende niet eiwit coderende RNA moleculen (ncRNAs). Hieronder vallen verschillende kleine ncRNAs, welke middels small RNA-seq gedetecteerd kunnen worden. Voor analyse van kleine RNAs waarvan (nog) geen annotatie bekend is, waren geen gerichte analyses beschikbaar. In hoofdstuk 2 wordt uitgelegd hoe met een nieuwe methode, *FlaiMapper*, kleine ncRNAs ontdekt kunnen worden. Vervolgens passen we deze methode in hoofdstuk 3 toe op samples van verschillende gradaties prostaatkanker. Hiermee vinden we verschillende nieuwe RNAs, met name afkomstig van C/D-box snoRNAs, die geassocieerd zijn met prostaatkanker en die gecorreleerd zijn met prognose. Deze associatie toont bovendien de relevantie en preciesie van de methode aan.

In prostaatkanker worden vaak fusie-genen gevonden. Detectie van fusie-genen middels RNA-seq is mogelijk en hiervoor zijn reeds verschillende programma's ontwikkeld. Echter, de overlap tussen deze programma's is gering en er wordt gead-

viseerd meerdere computerprogramma's te gebruiken en deze resultaten vervolgens te integreren. In hoofdstuk 4 presenteren wij een methode, *FuMa*, die in staat is de overlap van analyses door verschillende programma's te integreren en te rapporteren.

In de meeste protocollen die gebruikt worden om RNA klaar te maken voor het sequencen wordt uitsluitend mRNA geselecteerd op basis van de aanwezigheid van poly-A staarten. Een steeds vaker voorkomend alternatief is het gebruik maken van random hexamer primers in de reverse transcriptase stap. In data die op deze manier vervaardigd is, zullen ook niet-gepolyadenyleerde RNAs zitten, zoals lncRNAs, circRNAs en pre-mRNAs. Pre-mRNA moleculen kunnen sequenties uit het gehele gen bevatten met daarin niet alleen exonen maar ook intronen. DNA breuken van fusiegenen liggen voornamelijk in intronen. Daarom zou het theoretisch mogelijk moeten zijn het DNA breekpunt vast te stellen aan de hand van het pre-mRNA in random hexamer-primed RNA-seq data. Er was echter geen computer-programma beschikbaar die met deze extra laag van informatie rekening hield en dit kon oplossen. In hoofdstuk 5 introduceren wij een computer methode, *Dr. Disco*, die rekening houdt met de aanwezigheid van pre-mRNA en daarmee in staat is DNA breuken te vinden. We demonstreren dit aan de hand van 51 prostaatkanker weefsels waarin we hotspot regio's van DNA breuken in `TMPRSS2-ERG` vinden die overeenkomen met bevindingen uit een andere studie.

In hoofdstuk 6 benadrukken we dat goed geschreven en grondig geteste software belangrijk is voor integratie in grotere systemen. Beschikbaarheid van zulke software in het publieke domein stelde ons, als bijdrage door een grote gemeenschap, in staat vele van deze computerprogramma's te integreren in een vrij en open-source platform, de *Galaxy RNA Workbench*. Dit zorgt voor gemakkelijke toegang tot vele RNA-specifieke visualisaties en computerprogramma's, voorzien van een generieke grafische interface.

Samengevat hebben we verschillende nieuwe methoden bedacht om nieuwe, kanker gerelateerde, RNAs te ontdekken, die de potentie hebben als biomarker te kunnen dienen. Daarnaast zijn alle beschreven methoden tenslotte beschikbaar gemaakt in de *Galaxy RNA workbench*, zodat ze op deze manier vrijelijk en gratis beschikbaar zijn voor andere onderzoekers.

# Bibliography


[1] M. P. Robertson and G. F. Joyce. The Origins of the RNA World. *Cold Spring Harbor Perspectives in Biology*, 4(5):a003608–a003608, apr 2010.

[2] Alfred G. Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1(2):157–162, nov 2001.

[3] Carlo M. Croce. Oncogenes and Cancer. *New England Journal of Medicine*, 358(5):502–511, jan 2008.

[4] J.D. D. Watson and F.H.C. H. C. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

[5] Ray Wu and Ellen Taylor. Nucleotide sequence analysis of DNA. *Journal of Molecular Biology*, 57(3):491–511, 1971.

[6] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, et al. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, 2001.

[7] International Human Genome Sequencing. Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[8] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, oct 2004.

[9] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database. *Nucleic Acids Research*, 31(1):51–54, 2003.

[10] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, Nov 2015.

[11] Alexander F. Palazzo and Eliza S. Lee. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6:2, jan 2015.

[12] John S. Mattick and Igor V. Makunin. Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1):R17–R29, apr 2006.

[13] Yiwen Fang and Melissa J. Fullwood. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics*, 14(1):42–54, Feb 2016.

[14] Guoku Hu, Fang Niu, Bree A. Humburg, Ke Liao, Venkata Sunil Bendi, Shannon Callen, Howard S. Fox, and Shilpa Buch. Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. *Oncotarget*, 9(26):18648–18663, Jan 2018.

[15] Heena Khatter, Alexander G. Myasnikov, S. Kundhavai Natchiar, and Bruno P. Klaholz. Structure of the human 80S ribosome. *Nature*, 520:640–645, Apr 2015.

[16] James W F Catto, Antonio Alcaraz, Anders S. Bjartell, Ralph De Vere White, Christopher P. Evans, Susanne Fussel, Freddie C. Hamdy, Olli Kallioniemi, Lourdes Mengual, Thorsten Schlomm, and Tapio Visakorpi. MicroRNA in prostate, bladder, and kidney cancer: A systematic review. *European Urology*, 59(5):671–681, 2011.

[17] Marilena V. Iorio and Carlo M. Croce. MicroRNA dysregulation in cancer: Diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine*, 4(3):143–159, Feb 2012.

[18] Yong Peng and Carlo M Croce. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1(1):15004, jan 2016.

[19] Giorgio Dieci, Milena Preti, and Barbara Montanini. Eukaryotic snoRNAs: A paradigm for gene expression flexibility. *Genomics*, 94(2):83–88, 2009.

[20] Tamás Kiss. Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions. *Cell*, 109(2):145–148, 2002.

[21] Hadi Jorjani, Stephanie Kehr, Dominik J. Jedlinski, Rafal Gumienny, Jana Hertel, Peter F. Stadler, Mihaela Zavolan, and Andreas R. Gruber. An updated human snoRNAome. *Nucleic Acids Research*, 44(11):5068–5082, May 2016.

[22] Stephen Jefferson Sharp, Jerone Schaack, Lyan Cooley, Debroh Johnson Burke, and Dieter Soil. Structure and Transcription of Eukaryotic tRNA Gene. *Critical Reviews in Biochemistry*, 19(2):107–144, jan 1985.

[23] Ulrike Lambertz, Mariana E Oviedo Ovando, Elton Vasconcelos, Peter J Unrau, Peter J Myler, and Neil E Reiner. Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. *BMC Genomics*, 16(1):151, 2015.

[24] Elena S. Martens-Uzunova, Michael Olvedy, and Guido Jenster. Beyond microRNA – Novel RNAs derived from small non-coding RNA and their implication in cancer . *Cancer Letters*, 340(2):201–211, 2013.

[25] Michael Olvedy, Mauro Scaravilli, Youri Hoogstrate, Tapio Visakorpi, Guido Jenster, and Elena Martens-Uzunova. A comprehensive repertoire of tRNA-derived fragments in prostate cancer. *Oncotarget*, 7(17):24766–24777, 2016.

[26] S. Mahlab, T. Tuller, and M. Linial. Conservation of the relative tRNA composition in healthy and cancerous tissues. *RNA*, 18(4):640–652, feb 2012.

[27] Julia Salzman, Charles Gawad, Peter Lincoln Wang, Norman Lacayo, and Patrick O. Brown. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. *PLoS ONE*, 7(2):e30733, feb 2012.

[28] Steven P. Barrett and Julia Salzman. Circular RNAs: analysis, expression and potential functions. *Development*, 143(11):1838–1847, may 2016.

[29] Shujuan Meng, Hecheng Zhou, Ziyang Feng, Zihao Xu, Ying Tang, Peiyao Li, and Minghua Wu. CircRNA: functions and properties of a novel potential biomarker for cancer. *Molecular Cancer*, 16(1):94, may 2017.

[30] Maximiliano M Portal, Valeria Pavet, Cathie Erb, and Hinrich Gronemeyer. Human cells contain natural double-stranded RNAs with potential regulatory functions. *Nature Structural & Molecular Biology*, 22(1):89–97, dec 2014.

[31] Natasha S. Latysheva and M. Madan Babu. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research*, 44(10):4487–4503, 2016.

[32] Milana Frenkel-morgenstern, Vincent Lacroix, Iakes Ezkurdia, Milana Frenkel-morgenstern, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Pozo, Michael Tress, Rory Johnson, Roderic

Guigo, and Alfonso Valencia. Chimeras taking shape : Potential functions of proteins encoded by chimeric RNA transcripts Chimeras taking shape : Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Research*, 22(7):1231–1242, 2012.

[33] T H Rabbitts. Chromosomal translocations in human cancer. *Nature*, 372(6502):143–149, 1994.

[34] Jindan Yu, Jianjun Yu, Ram-Shankar Mani, Qi Cao, Chad J. Brenner, Xuhong Cao, Xiaoju Wang, Longtao Wu, James Li, Ming Hu, Yusong Gong, Hong Cheng, Bharathi Laxman, Adaikkalam Vellaichamy, Sunita Shankar, Yong Li, Saravana M. Dhanasekaran, Roger Morey, Terrence Barrette, Robert J. Lonigro, Scott A. Tomlins, Sooryanarayana Varambally, Zhaohui S. Qin, and Arul M. Chinnaiyan. An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression. *Cancer Cell*, 17(5):443–454, may 2010.

[35] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B. Alexandrov, Sancha Martin, David C. Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B. Brinkman, Sandro Morganella, Miriam R. Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E. Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A. Foekens, Moritz Gerstung, Gerrit K. J. Hooijer, Se Jin Jang, David R. Jones, Hyung-Yong Kim, Tari A. King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A. Purdie, Keiran Raine, Kamna Ramakrishnan, F. Germán Rodríguez-González, Gilles Romieu, Anieta M. Sieuwerts, Peter T. Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A. Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G. Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T. Ueno, Christos Sotiriou, Alain Viari, P. Andrew Futreal, Peter J. Campbell, Paul N. Span, Steven Van Laere, Sunil R. Lakhani, Jorunn E. Eyfjord, Alastair M. Thompson, Ewan Birney, Hendrik G. Stunnenberg, Marc J. van de Vijver, John W. M. Martens, Anne-Lise Børresen-Dale, Andrea L. Richardson, Gu Kong, Gilles Thomas, and Michael R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, may 2016.

[36] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, mar 2011.

[37] David Marin, Amr R. Ibrahim, Claire Lucas, Gareth Gerrard, Lihui Wang, Richard M. Szydlo, Richard E. Clark, Jane F. Apperley, Dragana Milojkovic, Marco Bua, Jiri Pavlu, Christos Paliompeis, Alistair Reid, Katayoun Rezvani, John M. Goldman, and Letizia Foroni. Assessment of BCR-ABL1 Transcript Levels at 3 Months Is the Only Requirement for Predicting Outcome for Patients With Chronic Myeloid Leukemia Treated With Tyrosine Kinase Inhibitors. *Journal of Clinical Oncology*, 30(3):232–238, 2012.

[38] Nadia M Davidson, Ian J Majewski, and Alicia Oshlack. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Medicine*, 7(1):43, 2015.

[39] Daniela Diverio, Roberta Riccioni, Franco Mandelli, and Francesco Lo Coco. The PML/RAR alpha fusion gene in the diagnosis and monitoring of acute promyelocytic leukemia. *Haematologica*, 80(2):155–160, 1995.

[40] Zheng Yang, Lu Yu, and Zhe Wang. PCA3 and TMPRSS2-ERG gene fusions as diagnostic biomarkers for prostate cancer. *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu*, 28(1):65–71, 2016.

[41] E. Fainstein, C. Marcelle, A. Rosner, E. Canaani, R. P. Gale, O. Dreazen, S. D. Smith, and C. M. Croce. A new fused transcript in Philadelphia chromosome positive acute lymphocytic leukaemia. *Nature*, 330(6146):386–388, nov 1987.

[42] Wendy Stock, Daohai Yu, Ted Karrison, Dorie Sher, Richard Stone, Richard Larson, and Clara Bloomfield. Quantitative real-time RT-PCR monitoring of BCR-ABL in chronic myelogenous leukemia shows lack of agreement in blood and bone marrow samples. *International Journal of Oncology*, 28(5):1099–1103, may 2006.

[43] Jacques Chasseriau, Jérôme Rivet, Frédéric Bilan, Jean-Claude Chomel, François Guilhot, Nicolas Bourmeyster, and Alain Kitzis. Characterization of the Different BCR-ABL Transcripts with a Single Multiplex RT-PCR. *The Journal of Molecular Diagnostics*, 6(4):343–347, nov 2004.

[44] P Rousselot, Bhushan Hardas, A Patel, Fabien Guidez, Joop Gaken, S Castaigne, A Dejean, H de Thé, L Degos, and Farzin Farzaneh. The PML-RAR alpha gene product of the t(15;17) translocation inhibits retinoic acid-induced granulocytic differentiation and mediated transactivation in human myeloid cells. *Oncogene*, 9(2):545–551, Mar 1994.

[45] Brittany C. Parker, Manon Engels, Matti Annala, and Wei Zhang. Emergence of FGFR family gene fusions as therapeutic targets in a wide spectrum of solid tumours. *Journal of Pathology*, 232(1):4–15, 2014.

[46] Felix Y. Feng, J. Chad Brenner, Maha Hussain, and Arul M. Chinnaiyan. Molecular pathways: targeting ETS gene fusions in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 20(17):4442–4448, 2014.

[47] Brittany C. Parker and Wei Zhang. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chinese Journal of Cancer*, 32(11):594–603, nov 2013.

[48] Marco Beccuti, Matteo Carrara, Francesca Cordero, Fulvio Lazzarato, Susanna Donatelli, Francesca Nadalin, Alberto Policriti, and Raffaele A Calogero. Chimera: a Bioconductor package for secondary analysis of fusion products. *Bioinformatics (Oxford, England)*, 30(24):3556–3557, December 2014.

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

[50] Bo Wang, Lin Wan, Anqi Wang, and Lei M. Li. An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters. *Scientific Reports*, 7(1):e4134, Feb 2017.

[51] S Goodwin, J D McPherson, and W R McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–351, 2016.

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

[53] Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, jan 2004.

[54] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, nov 2009.

[55] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, Dec 2014.

[56] John Trinick and Larissa Tskhovrebova. Titin: A molecular control freak. *Trends in Cell Biology*, 9(10):377–380, 1999.

[57] M L Bang, T Centner, F Fornoff, a J Geach, M Gotthardt, M McNabb, C C Witt, D Labeit, C C Gregorio, H Granzier, and S Labeit. The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circulation research*, 89(11):1065–72, 2001.

[58] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, feb 2015.

[59] Hiroaki Sakai, Ken Naito, Eri Ogiso-Tanaka, Yu Takahashi, Kohtaro Iseki, Chiaki Muto, Kazuhito Satou, Kuniko Teruya, Akino Shiroma, Makiko Shimoji, Takashi Hirano, Takeshi Itoh, Akito Kaga, and Norihiko Tomooka. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Scientific Reports*, 5(1), nov 2015.

[60] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, Jan 2018.

[61] Rafael Ferreira da Silva, Rosa Filgueira, Ilia Pietri, Ming Jiang, Rizos Sakellariou, and Ewa Deelman. A characterization of workflow management systems for extreme-scale applications. *Future Generation Computer Systems*, 75:228–238, oct 2017.

[62] Markus Hafner, Pablo Landgraf, Janos Ludwig, Amanda Rice, Tolulope Ojo, Carolina Lin, Daniel Holoch, Cindy Lim, and Thomas Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44(1):3–12, 2008.

[63] R C Lee, V Ambros, V. A. Erdmann, R. C. Lee, R. L. Feinbaum, V. Ambros, B. Reinhart, B. Wightman, I. Ha, G. Ruvkun, A. E. Pasquinelli, A. Grishok, G. Hutvagner, P. A. Sharp, W. J. Kent, A. M. Zahler, D. H. Mathews, J. Sabina, M. Zuker, D. H. Turner, S. F. Altschul, N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294(5543):862–4, 2001.

[64] Mitchell S. Stark, Sonika Tyagi, Derek J. Nancarrow, Glen M. Boyle, Anthony L. Cook, David C. Whiteman, Peter G. Parsons, Christopher Schmidt, Richard A. Sturm, and Nicholas K. Hayward. Characterization of the Melanoma miRNAome by Deep Sequencing. *PLoS ONE*, 5(3):e9685, Mar 2010.

[65] Ryan D. Morin, Michael D. O'Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, Connie J. Eaves, and Marco A. Marra. Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem cells. *Genome Res*, 18(4):610–621, Apr 2008.

[66] I-Fang Chung, Shing-Jyh Chang, Chen-Yang Chen, Shu-Hsuan Liu, Chia-Yang Li, Chia-Hao Chan, Chuan-Chi Shih, and Wei-Chung Cheng. YM500v3: a database for small RNA sequencing in human cancer research. *Nucleic acids research*, 45(D1):D925–D931, Nov 2016.

[67] Sébastien Tempel and Fariza Tahi. A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Research*, 40(11):e80–e80, feb 2012.

[68] Andrea Sboner, Lukas Habegger, Dorothee Pflueger, Stephane Terry, David Z Chen, Joel S Rozowsky, Ashutosh K Tewari, Naoki Kitabayashi, Benjamin J Moss, Mark S Chee, Francesca Demichelis, Mark A Rubin, and Mark B Gerstein. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome biology*, 11(10):R104, 2010.

[69] Yan He, Chengfu Yuan, Lichan Chen, Mingjuan Lei, Lucas Zellmer, Hai Huang, and Dezhong Liao. Transcriptional-Readthrough RNAs Reflect the Phenomenon of "A Gene Contains Gene(s)" or "Gene(s) within a Gene" in the Human Genome, and Thus Are Not Chimeric RNAs. *Genes*, 9(1):40, jan 2018.

[70] Daehwan Kim and Steven L Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology*, 12(8):R72, January 2011.

[71] Andrew McPherson, Fereydoun Hormozdiari, Abdalnasser Zayed, Ryan Giuliany, Gavin Ha, Mark G F Sun, Malachi Griffith, Alireza Moussavi, Janine Senz, Nataliya Melnyk, Marina Pacheco, Marco A. Marra, Martin Hirst, Torsten O. Nielsen, S. Cenk Sahinalp, David Huntsman, and Sohrab P. Shah. Defuse: An algorithm for gene fusion discovery in tumor RNA-seq data. *PLoS Computational Biology*, 7(5):e1001138, May 2011.

[72] Matthew K. Iyer, Arul M. Chinnaiyan, and Christopher A. Maher. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011.

[73] Wenlong Jia, Kunlong Qiu, Minghui He, Pengfei Song, Quan Zhou, Feng Zhou, Yuan Yu, Dandan Zhu, Michael L Nickerson, Shengqing Wan, Xiangke Liao, Xiaoqian Zhu, Shaoliang Peng, Yingrui Li, Jun Wang, and Guangwu Guo. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome biology*, 14(2):R12, 2013.

[74] Daniel Nicorici, Mihaela Satalan, Henrik Edgren, Sara Kangaspeska, Astrid Murumagi, Olli Kallioniemi, Sami Virtanen, and Olavi Kilkku. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*, Nov 2014. [This article is a preprint and has not been peer-reviewed].

[75] Silvia Liu, Wei Hsiang Tsai, Ying Ding, Rui Chen, Zhou Fang, Zhiguang Huo, Sunghwan Kim, Tianzhou Ma, Ting Yu Chang, Nolan Michael Priedigkeit, Adrian V. Lee, Jianhua Luo, Hsei Wei Wang, I. Fang Chung, and George C. Tseng. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Research*, 44(5):e47, Mar 2015.

[76] Matteo Carrara, Marco Beccuti, Federica Cavallo, Susanna Donatelli, Fulvio Lazzarato, Francesca Cordero, and Raffaele a Calogero. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC bioinformatics*, 14(7):S2, 2013.

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

[78] Shailesh Kumar, Angie Duy Vo, Fujun Qin, and Hui Li. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Scientific reports*, 6:21597, 2016.

[79] Jin Zhang, Nicole M. White, Heather K. Schmidt, Robert S. Fulton, Chad Tomlinson, Wesley C. Warren, Richard K. Wilson, and Christopher A. Maher. INTEGRATE: Gene fusion discovery using whole genome and transcriptome data. *Genome Research*, 26(1):108–118, 2016.

[80] Giulia Paciello and Elisa Ficarra. FuGePrior: A novel gene fusion prioritization algorithm based on accurate fusion structure analysis in cancer RNA-seq samples. *BMC Bioinformatics*, 18(1):58, Jan 2017.

[81] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: A probabilistic framework for structural variant discovery. *Genome biology*, 15(6):R84, 2014.

[82] Mattia Brugiolo, Lydia Herzel, and Karla M. Neugebauer. Counting on co-transcriptional splicing. *F1000Prime Rep*, 5:9–9, Apr 2013.

[83] Evan C Merkhofer, Peter Hu, and Tracy L Johnson. Methods and Protocols. In *Introduction to Cotranscriptional RNA Splicing*, page 83–96. Humana Press, 2014.

[84] Muhammad A. Tariq, Hyunsung J. Kim, Olufisayo Jejelowo, and Nader Pourmand. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, 39(18):1–10, 2011.

[85] Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods*, 7(9):709–15, 2010.

[86] Nivedita Pathak and BimalKumar Das. Polymerase chain reaction as a diagnostic tool in human viral myocarditis. *Journal of the Practice of Cardiovascular Sciences*, 1(2):168, 2015.

[87] Yan Ma, Ranjana Ambannavar, James Stephans, Jennie Jeong, Andrew Dei Rossi, Mei-Lan Liu, Adam J. Friedman, Jason J. Londry, Richard Abramson, Ellen M. Beasley, Joffre Baker, Samuel Levy, and Kunbin Qu. Fusion Transcript Discovery in Formalin-Fixed Paraffin-Embedded Human Breast Cancer Tissues Reveals a Link to Tumor Progression. *PLoS ONE*, 9(4):e94202, apr 2014.

[88] Torsten Seemann. Ten recommendations for creating usable bioinformatics command line software. *GigaScience*, 2:15, 2013.

[89] Silva Luis Bastiao, Jimenez Rafael C, Blomberg Niklas, and Luis Oliveira José. General guidelines for biomedical software development. *F1000Research*, 6:273, July 2017.

[90] Felipe da Veiga Leprevost, Valmir C. Barbosa, Eduardo L. Francisco, Yasset Perez-Riverol, and Paulo C. Carvalho. On best practices in the development of bioinformatics software. *Frontiers in Genetics*, 5:199, jul 2014.

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

[92] Jon Ison, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice. EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332, 2013.

[93] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb C Miller, W James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455, sep 2005.

[94] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, mar 2009.

[95] Daniel Blankenberg, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, Galaxy Team, James Taylor, and Anton Nekrutenko. Dissemination of scientific software with Galaxy ToolShed. *Genome Biology*, 15(2):403, 2014.

[96] Po-E Li, Chien-Chi Lo, Joseph J. Anderson, Karen W. Davenport, Kimberly A. Bishop-Lilly, Yan Xu, Sanaa Ahmed, Shihai Feng, Vishwesh P. Mokashi, and Patrick S.G. Chain. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Research*, 45(1):67–80, jan 2017.

[97] Danny Challis, Jin Yu, Uday S Evani, Andrew R Jackson, Sameer Paithankar, Cristian Coarfa, Aleksandar Milosavljevic, Richard A Gibbs, and Fuli Yu. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*, 13(1):8, 2012.

[98] Carol Soderlund, William Nelson, Mark Willer, and David R. Gang. TCW: Transcriptome computational workbench. *PLoS ONE*, 8(7):e69401, jul 2013.

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

[100] Eduardo Andrés-León, Rocío Núñez-Torres, and Ana M Rojas. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Scientific reports*, 6, 2016.

[101] M. B. Stocks, S. Moxon, D. Mapleson, H. C. Woolfenden, I. Mohorianu, L. Folkes, F. Schwach, T. Dalmay, and V. Moulton. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15):2059–2061, Aug 2012.

[102] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1):W3–W10, May 2016.

[103] Birte Kehr, Kathrin Trappe, Manuel Holtgrewe, and Knut Reinert. Genome alignment with graph data structures: a comparison. *BMC bioinformatics*, 15(1):99, 2014.

[104] Deanna M. Church, Valerie A. Schneider, Karyn Meltz Steinberg, Michael C. Schatz, Aaron R. Quinlan, Chen-Shan Chin, Paul A. Kitts, Bronwen Aken, Gabor T. Marth, Michael M. Hoffman, Javier Herrero, M Lisandra Zepeda Mendoza, Richard Durbin, and Paul Flicek. Extending reference assembly models. *Genome Biology*, 16(1):13, 2015.

[105] Simon A. Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott, and Peter J. Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, oct 2014.

[106] Ilkka Lappalainen, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif Ur-Rehman, Gary Saunders, Jag Kandasamy, Mario Caccamo, Rasko Leinonen, Brendan Vaughan, Thomas Laurent, Francis Rowland, Pablo Marin-Garcia, Jonathan Barker, Petteri Jokinen, Angel Carreño Torres, Jordi Rambla de Argila, Oscar Martinez Llobet, Ignacio Medina, Marc Sitges Puy, Mario Alberich, Sabela de la Torre, Arcadi Navarro, Justin Paschall, and Paul Flicek. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.*, 47(7):692–695, July 2015.

[107] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G. Van Meir, Daniel J. Brat, Gena M. Mastrogianakis, Jeffrey J. Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Sep 2008.

[108] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkologia*, 1A:68–77, 2015.

[109] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, Sep 2013.

[110] Louis Papageorgiou, Picasi Eleni, Sofia Raftopoulou, Meropi Mantaiou, Vasileios Megalooikonomou, and Dimitrios Vlachakis. Genomic big data hitting the storage bottleneck. *EMBnet.journal*, 24(0):910, Apr 2018.

[111] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5):734–740, Jan 2011.

[112] Sebastian Palluk, Daniel H Arlow, Tristan de Rond, Sebastian Barthel, Justine S Kang, Rathin Bector, Hratch M Baghdassarian, Alisa N Truong, Peter W Kim,

Anup K Singh, et al. De novo DNA synthesis using polymerase-nucleotide conjugates. *Nature Biotechnology*, 36(7):645–650, Jun 2018.

[113] Nicolas Mottet, Joaquim Bellmunt, Michel Bolla, Erik Briers, Marcus G. Cumberbatch, Maria De Santis, Nicola Fossati, Tobias Gross, Ann M. Henry, Steven Joniau, Thomas B. Lam, Malcolm D. Mason, Vsevolod B. Matveev, Paul C. Moldovan, Roderick C.N. van den Bergh, Thomas Van den Broeck, Henk G. van der Poel, Theo H. van der Kwast, Olivier Rouvière, Ivo G. Schoots, Thomas Wiegel, and Philip Cornford. EAU-ESTRO-SIOG guidelines on prostate cancer. part 1: Screening, diagnosis, and local treatment with curative intent. *European Urology*, 71(4):618–629, apr 2017.

[114] Irene V. Bijnsdorp, Martin E. van Royen, Gerald W. Verhaegh, and Elena S. Martens-Uzunova. The Non-Coding Transcriptome of Prostate Cancer: Implications for Clinical Practice. *Molecular Diagnosis & Therapy*, 21(4):385–400, Mar 2017.

[115] Srinivas Pentyala, Terry Whyard, Sahana Pentyala, John Muller, John Pfail, Sunjit Parmar, Carlos G Helguero, and Sardar Khan. Prostate cancer markers: An update. *Biomedical reports*, 4(3):263–268, 2016.

[116] Thorsten H. Ecke, Horst H. Schlechte, Katrin Schiemenz, Markus D. Sachs, Severin V. Lenk, Birgit D. Rudolph, and Stefan A. Loening. TP53 Gene Mutations in Prostate Cancer Progression. *Anticancer Research*, 30(5):1579–1586, 2010.

[117] Milan S. Geybels, Min Fang, Jonathan L. Wright, Xiaoyu Qu, Marina Bibikova, Brandy Klotzle, Jian-Bing Fan, Ziding Feng, Elaine A. Ostrander, Peter S. Nelson, et al. PTEN loss is associated with prostate cancer recurrence and alterations in tumor DNA methylation profiles. *Oncotarget*, 8(48):84338–84348, Sep 2017.

[118] Kurtis Eisermann, Dan Wang, Yifeng Jing, Laura E. Pascal, and Zhou Wang. Androgen receptor gene mutation, rearrangement, polymorphism. *Translational Andrology and Urology*, 2(3):137–147, 2013.

[119] Jeremy P Clark and Colin S Cooper. ETS gene fusions in prostate cancer. *Nature reviews. Urology*, 6(8):429–439, August 2009.

[120] Mark A. Rubin, Christopher A. Maher, and Arul M. Chinnaiyan. Common Gene Rearrangements in Prostate Cancer. *Journal of Clinical Oncology*, 29(27):3659–3668, sep 2011.

[121] Scott A Tomlins, Bharathi Laxman, Sooryanarayana Varambally, Xuhong Cao, Jindan Yu, Beth E Helgeson, Qi Cao, John R Prensner, Mark A Rubin, Rajal B Shah, Rohit Mehra, and Arul M Chinnaiyan. Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia*, 10(2):177–188, February 2008.

[122] Jacques Lapointe, Young H Kim, Melinda A Miller, Chunde Li, Gulsah Kaygusuz, Matt van de Rijn, David G Huntsman, James D Brooks, and Jonathan R Pollack. A variant TMPRSS2 isoform and ERG fusion product in prostate cancer with implications for molecular diagnosis. *Modern Pathology*, 20(4):467–473, mar 2007.

[123] J Clark, S Merson, S Jhavar, P Flohr, S Edwards, C S Foster, R Eeles, F L Martin, D H Phillips, M Crundwell, T Christmas, A Thompson, C Fisher, G Kovacs, and C S Cooper. Diversity of TMPRSS2-ERG fusion transcripts in the human prostate. *Oncogene*, 26(18):2667–2673, oct 2006.

[124] Xiaoju Wang, Yuanyuan Qiao, Irfan A. Asangani, Bushra Ateeq, Anton Poliakov, Marcin Cieślik, Sethuramasundaram Pitchiaya, Balabhadrapatruni V.S.K. Chakravarthi, Xuhong Cao, Xiaojun Jing, Cynthia X. Wang, Ingrid J. Apel, Rui Wang, Jean Ching-Yi Tien, Kristin M. Juckette, Wei Yan, Hui Jiang, Shaomeng Wang, Sooryanarayana Varambally, and Arul M. Chinnaiyan. Development of Peptidomimetic Inhibitors of the ERG Gene Fusion Product in Prostate Cancer. *Cancer Cell*, 31(4):532–548.e7, apr 2017.

[125] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter a.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris a. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:e160018, 2016.

[126] Chong-Jian Chen and Edith Heard. Small RNAs derived from structural noncoding RNAs. *Methods*, 63(1):76–84, 2013.

[127] Andrew Sobala and Gyorgy Hutvagner. Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdisciplinary Reviews: RNA*, 2(6):853–862, 2011.

[128] Michelle S. Scott, Motoharu Ono, Kayo Yamada, Akinori Endo, Geoffrey J. Barton, and Angus I. Lamond. Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Research*, 40(8):3676–3688, 2012.

[129] Eivind Valen, Pascal Preker, Peter Refsing Andersen, Xiaobei Zhao, Yun Chen, Christine Ender, Anne Dueck, Gunter Meister, Albin Sandelin, and Torben Heick Jensen. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat Struct Mol Biol*, 18(9):1075–1082, Sep 2011.

[130] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531, 2004.

[131] Marc R. Fabian and Nahum Sonenberg. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat Struct Mol Biol*, 19(6):586–593, Jun 2012.

[132] Marc R. Friedlander, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotech*, 26(4):407–415, Apr 2008.

[133] Satoshi Yamasaki, Pavel Ivanov, Guo-fu Hu, and Paul Anderson. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *The journal of Cell Biology*, 185(1):35–42, 2009.

[134] Debrah M. Thompson and Roy Parker. Stressing Out over tRNA Cleavage . *Cell*, 138(2):215–219, 2009.

[135] Anthony K Henras, Christophe Dez, and Yves Henry. RNA structure and function in C/D and H/ACA s(no)RNPs. *Current Opinion in Structural Biology*, 14(3):335–343, jun 2004.

[136] Shivendra Kishore, Amit Khanna, Zhaiyi Zhang, Jingyi Hui, Piotr J. Balwierz, Mihaela Stefan, Carol Beach, Robert D. Nicholls, Mihaela Zavolan, and Stefan Stamm. The snoRNA MBII-52 (SNORD115) is processed into smaller RNAs and regulates alternative splicing. *Human Molecular Genetics*, 19(7):1153–1164, 2010.

[137] Christine Ender, Azra Krek, Marc R. Friedländer, Michaela Beitzinger, Lasse Weinmann, Wei Chen, Sébastien Pfeffer, Nikolaus Rajewsky, and Gunter Meister. A Human snoRNA with MicroRNA-Like Functions. *Molecular Cell*, 32(4):519–528, 2008.

[138] Markus Brameier, Astrid Herwig, Richard Reinhardt, Lutz Walter, and Jens Gruber. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic Acids Research*, 39(2):675–686, 2011.

[139] Motoharu Ono, Michelle S. Scott, Kayo Yamada, Fabio Avolio, Geoffrey J. Barton, and Angus I. Lamond. Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Research*, 39(9):3879–3891, 2011.

[140] Y.-P. Mei, J.-P. Liao, J. Shen, L. Yu, B.-L. Liu, L. Liu, R.-Y. Li, L. Ji, S. G. Dorsey, Z.-R. Jiang, R. L. Katz, J.-Y. Wang, and F. Jiang. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene*, 31(22):2794–2804, May 2012.

[141] Xue Yuan Dong, Peng Guo, Jeff Boyd, Xiaodong Sun, Qunna Li, Wei Zhou, and Jin Tang Dong. Implication of snoRNA U50 in human breast cancer. *journal of Genetics and Genomics*, 36(8):447–454, 2009.

[142] Marjan E. Askarian-Amiri, Joanna Crawford, Juliet D. French, Chanel E. Smart, Martin A. Smith, Michael B. Clark, Kelin Ru, Tim R. Mercer, Ella R. Thompson, Sunil R. Lakhani, Ana C. Vargas, Ian G. Campbell, Melissa A. Brown, Marcel E. Dinger, and John S. Mattick. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, 17(5):878–891, 2011.

[143] Baoyan Bai, Hester Liu, and Marikki Laiho. Small RNA expression and deep sequencing analyses of the nucleolus reveal the presence of nucleolus-associated microRNAs. *FEBS Open Bio*, 4(0):441–449, 2014.

[144] Maud Contrant, Aurélie Fender, Béatrice Chane-Woon-Ming, Ramy Randrianjafy, Valérie Vivet-Boudou, Delphine Richer, and Sébastien Pfeffer. Importance of the RNA secondary structure for the relative accumulation of clustered viral microRNAs. *Nucleic Acids Research*, 42(12):7981–7996, 2014.

[145] Luke A Selth, Matthew J Roberts, Clement W K Chow, Villis R Marshall, Suhail A R Doi, Andrew D Vincent, Lisa M Butler, Martin F Lavin, Wayne D Tilley, and Robert A Gardiner. Human seminal fluid as a source of prostate cancer-specific microRNA biomarkers. *Endocrine-Related Cancer*, 21(4):L17–L21, 2014.

[146] Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl_1):D152–D157, 2010.

[147] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovich, and Peter F. Stadler. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, 25(18):2298–2301, 2009.

[148] T. D. Schneider and R. M. Stephens. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Research*, 18:6097–6100, 1990.

[149] Yuk Yee Leung, Paul Ryvkin, Lyle H. Ungar, Brian D. Gregory, and Li-San Wang. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Research*, 2013.

[150] Cheng Yuan and Yanni Sun. RNA-CODE: A Noncoding RNA Classification Tool for Short Reads in NGS Data Lacking Reference Genomes. *PLoS ONE*, 8(10):e77596, Oct 2013.

[151] E S Martens-Uzunova, S E Jalava, N F Dits, G J L H van Leenders, S Møller, J Trapman, C H Bangma, T Litman, T Visakorpi, and G Jenster. Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*, 31(8):978–991, 2012.

[152] John C. Castle, Christopher D. Armour, Martin Löwer, David Haynor, Matthew Biery, Heather Bouzek, Ronghua Chen, Stuart Jackson, Jason M. Johnson, Carol A. Rohl, and Christopher K. Raymond. Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polya-neutral amplification. *PLoS ONE*, 5(7), 2010.

[153] Yuuichi Soeno, Kazuya Fujita, Tomoo Kudo, Masataka Asagiri, Shigeru Kakuta, Yuji Taya, Yoshihito Shimazu, Kaori Sato, Ritsuko Tanaka-Fujita, Sachiko Kubo, Yoichiro Iwakura, Yoshikazu Nakamura, Shigeo Mori, and Takaaki Aoba. Generation of a Mouse Model with Down-Regulated U50 snoRNA (SNORD50) Expression and Its Organ-Specific Phenotypic Modulation. *PLoS ONE*, 8(8), 2013.

[154] A Goeze, K Schluns, G Wolf, Z Thasler, S Petersen, and I Petersen. Chromosomal imbalances of primary and metastatic lung adenocarcinomas. *J Pathol*, 196(1):8–16, 2002.

[155] Jipei Liao, Lei Yu, Yuping Mei, Maria Guarnera, Jun Shen, Ruiyun Li, Zhenqiu Liu, and Feng Jiang. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. *Molecular cancer*, 9:198, 2010.

[156] D. Ronchetti, L. Mosca, G. Cutrona, G. Tuana, M. Gentile, S. Fabris, L. Agnelli, G. Ciceri, S. Matis, C. Massucco, M. Colombo, D. Reverberi, A.G. Recchia, S. Bossio, M. Negrini, P. Tassone, F. Morabito, M. Ferrarini, and A. Neri. Small nucleolar RNAs as new biomarkers in chronic lymphocytic leukemia. *BMC Medical Genomics*, 6(1):27, 2013.

[157] D Ronchetti, K Todoerti, G Tuana, L Agnelli, L Mosca, M Lionetti, S Fabris, P Colapietro, M Miozzo, M Ferrarini, P Tassone, and a Neri. The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. *Blood cancer journal*, 2(11):e96, 2012.

[158] W Valleron, E Laprevotte, E F Gautier, C Quelen, C Demur, E Delabesse, X Agirre, F Prosper, T Kiss, and P Brousset. Specific small nucleolar RNA expression profiles in acute leukemia. *Leukemia*, 26(9):2052–2060, 2012.

[159] Wilfried Valleron, Loic Ysebaert, Laure Berquet, Virginie Fataccioli, Cathy Quelen, Antoine Martin, Marie Parrens, Laurence Lamant, Laurence De Leval, Christian Gisselbrecht, Philippe Gaulard, and Pierre Brousset. Small nucleolar RNA expression profiling identifies potential prognostic markers in peripheral T-cell lymphoma. *Blood*, 120(19):3997–4005, 2012.

[160] Ritsuko Tanaka, Hitoshi Satoh, Masatsugu Moriyama, Kasumi Satoh, Yasuyuki Morishita, Syouko Yoshida, Toshiki Watanabe, Yoshikazu Nakamura, and Shigeo Mori. Intronic U50 small-nucleolar-RNA (snoRNA) host gene of no protein-coding potential is mapped at the chromosome breakpoint t(3;6)(q27;q15) of human B- cell lymphoma. *Genes to Cells*, 5(4):277–287, 2000.

[161] Xue-Yuan Dong, Carmen Rodriguez, Peng Guo, Xiaodong Sun, Jeffrey T. Talbot, Wei Zhou, John Petros, Qunna Li, Robert L. Vessella, Adam S. Kibel, Victoria L. Stevens, Eugenia E. Calle, and Jin-Tang Dong. SnoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. *Human Molecular Genetics*, 17(7):1031–1042, 2008.

[162] Gang Xu, Fang Yang, Cui-Ling Ding, Lan-Juan Zhao, Hao Ren, Ping Zhao, Wen Wang, and Zhong-Tian Qi. Small nucleolar RNA 113-1 suppresses tumorigenesis in hepatocellular carcinoma. *Molecular cancer*, 13:216, 2014.

[163] Luyue Chen, Lei Han, Jianwei Wei, Kailiang Zhang, Zhendong Shi, Ran Duan, Shouwei Li, Xuan Zhou, Peiyu Pu, Jianning Zhang, and Chunsheng Kang. SNORD76, a box C/D snoRNA, acts as a tumor suppressor in glioblastoma. *Scientific reports*, 5:e8588, 2015.

[164] H E Gee, F M Buffa, C Camps, A Ramachandran, R Leek, M Taylor, M Patil, H Sheldon, G Betts, J Homer, C West, J Ragoussis, and A L Harris. The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis. *British journal of Cancer*, 104(7):1168–1177, 2011.

[165] M Mourtada-Maarabouni, Mr Pickard, Vl Hedge, F Farzaneh, and Gt Williams. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene*, 28(2):195–208, Oct 2009.

[166] Ashesh A. Saraiya and Ching C. Wang. snoRNA, a Novel Precursor of microRNA in Giardia lamblia. *PLoS Pathog*, 4(11):e1000224, 11 2008.

[167] Ryan J. Taft, Evgeny A. Glazov, Timo Lassmann, Yoshihide Hayashizaki, Piero Carninci, and John S. Mattick. Small RNAs derived from snoRNAs. *RNA (New York, N.Y.)*, 15(7):1233–40, 2009.

[168] Alexander Maxwell Burroughs, Yoshinari Ando, Michiel Jan Laurens de Hoon, Yasuhiro Tomaru, Harukazu Suzuki, Yoshihide Hayashizaki, and Carsten Olivier Daub. Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA biology*, 8(1):158–77, 2011.

[169] Michelle S. Scott, Fabio Avolio, Motoharu Ono, Angus I. Lamond, and Geoffrey J. Barton. Human miRNA Precursors with Box H/ACA snoRNA Features. *PLoS Comput Biol*, 5(9):e1000507, 09 2009.

[170] Shivendra Kishore, Andreas R Gruber, Dominik J Jedlinski, Afzal P Syed, Hadi Jorjani, and Mihaela Zavolan. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome biology*, 14(5):R45, 2013.

[171] Manli Shen, Eduardo Eyras, Jie Wu, Amit Khanna, Serene Josiah, Mathieu Rederstorff, Michael Q. Zhang, and Stefan Stamm. Direct cloning of double-stranded RNAs from RNase protection analysis reveals processing patterns of C/D box snoRNAs and provides evidence for widespread antisense transcript expression. *Nucleic Acids Research*, 39(22):9720–9730, 2011.

[172] Yuuichi Soeno, Yuji Taya, Taras Stasyk, Lukas a Huber, Takaaki Aoba, and Alexander Hüttenhofer. Identification of novel ribonucleo-protein complexes from the brain-specific snoRNA MBII-52. *RNA (New York, N.Y.)*, 16(7):1293–1300, 2010.

[173] Thomas Schubert, Miriam Caroline Pusch, Sarah Diermeier, Vladimir Benes, Elisabeth Kremmer, Axel Imhof, and Gernot Längst. Df31 Protein and snoRNAs Maintain Accessible Higher-Order Structures of Chromatin. *Molecular Cell*, 48(3):434–444, 2012.

[174] Danijela Koppers-Lalic, Michael Hackenberg, Irene V. Bijnsdorp, Monique A J van Eijndhoven, Payman Sadek, Daud Sie, Nicoletta Zini, Jaap M. Middeldorp, Bauke Ylstra, Renee X. de Menezes, Thomas Wurdinger, Gerrit A. Meijer, and

D. Michiel Pegtel. Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes. *Cell Reports*, 8(6):1649–1658, 2014.

[175] Silke von Ahlfen, Andreas Missel, Klaus Bendrat, and Martin Schlumpberger. Determinants of RNA quality from FFPE samples. *PLoS ONE*, 2(12), 2007.

[176] A Liu, M T Tetzlaff, P Vanbelle, D Elder, M Feldman, J W Tobias, A R Sepulveda, and X Xu. MicroRNA Expression Profiling Outperforms mRNA Expression Profiling in Formalin-fixed Paraffin-embedded Tissues. *Int J Clin Exp Pathol*, 2(6):519–527, 2009.

[177] Olivia Larne, Elena Martens-Uzunova, Zandra Hagman, Anders Edsjö, Giuseppe Lippolis, Mirella S Vredenbregt Van Den Berg, Anders Bjartell, Guido Jenster, and Yvonne Ceder. MiQ - A novel microRNA based diagnostic and prognostic tool for prostate cancer. *International journal of Cancer*, 132(12):2867–2875, 2013.

[178] Youri Hoogstrate, Guido Jenster, and Elena S. Martens-Uzunova. FlaiMapper: Computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data. *Bioinformatics*, 31(5):665–673, 2015.

[179] Marie Line Bortolin-Cavaillé and Jérôme Cavaillé. The SNORD115 (H/MBII-52) and SNORD116 (H/MBII-85) gene clusters at the imprinted Prader-Willi locus generate canonical box C/D snoRNAs. *Nucleic Acids Research*, 40(14):6800–6807, 2012.

[180] C M Smith and J A Steitz. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Molecular and cellular biology*, 18(12):6897–6909, 1998.

[181] Zsuzsanna Kiss-László, Yves Henry, and Tamás Kiss. Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO journal*, 17(3):797–807, 1998.

[182] Audrone Lapinaite, Bernd Simon, Lars Skjaerven, Magdalena Rakwalska-Bange, Frank Gabel, and Teresa Carlomagno. The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature*, 502(7472):519–523, 2013.

[183] Zhihua Li, Christine Ender, Gunter Meister, Patrick S. Moore, Yuan Chang, and Bino John. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Research*, 40(14):6787–6799, 2012.

[184] Daniel Cifuentes, Huiling Xue, David W Taylor, Heather Patnode, Yuichiro Mishima, Sihem Cheloufi, Enbo Ma, Shrikant Mane, Gregory J Hannon, Nathan D Lawson, Scot A Wolfe, and Antonio J Giraldez. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science (New York, N.Y.)*, 328(5986):1694–1698, 2010.

[185] Mathew W Wright and Elspeth a Bruford. Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Human genomics*, 5(2):90–98, 2011.

[186] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[187] Todd M. Lowe and Sean R. Eddy. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1996.

[188] A J Kal, A J van Zonneveld, V Benes, M van den Berg, M G Koerkamp, K Albermann, N Strack, J M Ruijter, A Richter, B Dujon, W Ansorge, and H F Tabak. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular biology of the cell*, 10(6):1859–1872, 1999.

[189] David Goode, Sally Hunter, Maria Doyle, Tao Ma, Simone Rowley, David Choong, Georgina Ryland, and Ian Campbell. A simple consensus approach improves somatic mutation prediction accuracy. *Genome Medicine*, 5(9):90, 2013.

[190] Adam D. Ewing, Kathleen E. Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N. Yamaguchi, J. Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y. Sabelnykova, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Meth*, 12(7):623–630, Jul 2015.

[191] Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1):97, 2015.

[192] Chaitanya R Sanna, Wen-Hsiung Li, and Liqing Zhang. Overlapping genes in the human and mouse genomes. *BMC genomics*, 9:169, 2008.

[193] Michael F. Berger, Joshua Z. Levin, Krishna Vijayendran, Andrey Sivachenko, Xian Adiconis, Jared Maguire, Laura A. Johnson, James Robinson, Roel G. Verhaak, Carrie Sougnez, et al. Integrative analysis of the melanoma transcriptome. *Genome Research*, 20(4):413–427, 2010.

[194] Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, Anne-Lise Borresen-Dale, and Olli Kallioniemi. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome biology*, 12(1):R6, January 2011.

[195] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[196] Huanying Ge, Kejun Liu, Todd Juan, Fang Fang, Matthew Newman, and Wolfgang Hoeck. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, 27(14):1922–1928, 2011.

[197] Paolo Carnevali, Jonathan Baccash, Aaron L Halpern, Igor Nazarenko, Geoffrey B Nilsen, Krishna P Pant, Jessica C Ebert, Anushka Brownley, Matt Morenzoni, Vitali Karpinchyk, et al. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. *Journal of Computational Biology*, 19(3):279–292, 2012.

[198] Thomas D. Wu and Colin K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, 21(9):1859–75, May 2005.

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

[200] Brian Haas, Alexander Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas Doak, Natalie Pochet, Jing Sun, Catherine Wu, Thomas Gingeras, and Aviv Regev. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*, page e120295, mar 2017.

[201] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, 89(1):19.10.1–19.10.21, 2010.

[202] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, Charles Lee, James E Montie, Rajal B Shah, Kenneth J Pienta, Mark A Rubin, and Arul M Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)*, 310(5748):644–8, October 2005.

[203] Yang Li, Jeremy Chien, David I. Smith, and Jian Ma. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, 27(12):1708–1710, 2011.

[204] Mridula Nambiar, Vijayalakshmi Kari, and Sathees C. Raghavan. Chromosomal translocations in cancer. *Biochimica et biophysica acta*, 1786(2):139–52, December 2008.

[205] Frederic J Kaye. Mutation-associated fusion cancer genes in solid tumors. *Molecular cancer therapeutics*, 8(6):1399–408, June 2009.

[206] Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):1011–1025, nov 2015.

[207] Scott A. Tomlins, Anders Bjartell, Arul M. Chinnaiyan, Guido Jenster, Robert K. Nam, Mark A. Rubin, and Jack A. Schalken. ETS gene fusions in prostate cancer: From discovery to daily clinical practice. *European Urology*, 56(2):275–286, aug 2009.

[208] K Yoshihara, Q Wang, W Torres-Garcia, S Zheng, R Vegesna, H Kim, and R G W Verhaak. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, 34(37):4845–4854, dec 2014.

[209] Saskia Hiltemann, Elizabeth A McClellan, Jos van Nijnatten, Sebastiaan Horsman, Ivo Palli, Ines Teles Alves, Thomas Hartjes, Jan Trapman, Peter van der Spek, Guido Jenster, et al. iFUSE: integrated fusion gene explorer. *Bioinformatics*, 29(13):1700–1701, 2013.

[210] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, apr 2014.

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

[212] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118, aug 2013.

[213] K. G. Hermans, J. L. Boormans, D. Gasi, G. J.H.L. van Leenders, G. Jenster, P. C.M.S. Verhagen, and J. Trapman. Overexpression of Prostate-Specific TMPRSS2(exon 0)-ERG Fusion Transcripts Corresponds with Favorable Prognosis of Prostate Cancer. *Clinical Cancer Research*, 15(20):6398–6403, oct 2009.

[214] Cath Tyner, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M. Fischer, David Gibson, Jairo Navarro Gonzalez, Luvina Guruvadoo, Maximilian Haeussler, Steve Heitner, Angie S. Hinrichs, Donna Karolchik, Brian T. Lee, Christopher M. Lee, Parisa Nejad, Brian J. Raney, Kate R. Rosenbloom, Matthew L. Speir, Chris Villarreal, John Vivian, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, 45(D1):D626–D634, 2017.

[215] Christopher Weier, Michael C Haffner, Timothy Mosbruger, David M Esopi, Jessica Hicks, Qizhi Zheng, Helen Fedor, William B Isaacs, Angelo M. De Marzo, William G Nelson, and Srinivasan Yegnasubramanian. Nucleotide resolution analysis of TMPRSS2 and ERG rearrangements in prostate cancer. *The Journal of Pathology*, 230(2):174–183, may 2013.

[216] Nicolas Stransky, Ethan Cerami, Stefanie Schalm, Joseph L. Kim, and Christoph Lengauer. The landscape of kinase fusions in cancer. *Nature Communications*, 5:e4846, sep 2014.

[217] Youri Hoogstrate, Chao Zhang, Alexander Senf, Jochem Bijlard, Saskia Hiltemann, David van Enckevort, Susanna Repo, Jaap Heringa, Guido Jenster, Remond J.A. Fijneman, Jan-Willem Boiten, Gerrit A. Meijer, Andrew Stubbs, Jordi Rambla, Dylan Spalding, Sanne Abeln, Youri Hoogstrate, Chao Zhang, Alexander Senf, Jochem Bijlard, Saskia Hiltemann, David van Enckevort, Susanna Repo, Jaap Heringa, Guido Jenster, Remond J.A. Fijneman, Jan-Willem Boiten, Gerrit A. Meijer, Andrew Stubbs, Jordi Rambla, Dylan Spalding, and Sanne Abeln. Integration of EGA secure data access into Galaxy. *F1000Research*, 5(0):1–9, 2016.

[218] Inês Teles Alves, Saskia Hiltemann, Thomas Hartjes, Peter van der Spek, Andrew Stubbs, Jan Trapman, and Guido Jenster. Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. *Human genetics*, 132(6):709–713, June 2013.

[219] Björn A. Grüning, Jörg Fallmann, Dilmurat Yusuf, Sebastian Will, Anika Erxleben, Florian Eggenhofer, Torsten Houwaart, Bérénice Batut, Pavankumar Videm, Andrea Bagnacani, Markus Wolfien, Steffen C. Lott, Youri Hoogstrate, Wolfgang R. Hess, Olaf Wolkenhauer, Steve Hoffmann, Altuna Akalin, Uwe Ohler, Peter F. Stadler, and Rolf Backofen. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. *Nucleic Acids Research*, 45(W1):W560–W566, 2017.

[220] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[221] Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, 3(4):e65, 2007.

[222] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, May 2012.

[223] Felipe Veiga da Leprevost, Björn A. Grüning, Saulo Alves Aflitos, Hannes L. Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I. Nesvizhskii, and Yasset Perez-Riverol. BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics*, mar 2017.

[224] J. Fallmann, V. Sedlyarov, A. Tanzer, P. Kovarik, and I. L. Hofacker. AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Research*, 44(D1):D90–95, Jan 2016.

[225] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin. DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Research*, 43(Database issue):D160–167, Jan 2015.

[226] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, Nov 2013.

[227] D. L. Corcoran, S. Georgiev, N. Mukherjee, E. Gottwein, R. L. Skalsky, J. D. Keene, and U. Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, 12(8):R79, Aug 2011.

[228] Y. Hoogstrate, R. Bottcher, S. Hiltemann, P. J. van der Spek, G. Jenster, and A. P. Stubbs. FuMa: reporting overlap in RNA-seq detected fusion genes. *Bioinformatics*, 32(8):1226–1228, Apr 2016.

[229] Carole A Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danius Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 38(Web Server issue):W677–82, July 2010.

[230] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

[231] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, Apr 2013.

[232] S. Washietl, S. Findeiss, S. A. Muller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, and N. Goldman. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, 17(4):578–594, Apr 2011.

[233] A. R. Gruber, R. Neubock, I. L. Hofacker, and S. Washietl. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Research*, 35(Web Server issue):W335–338, Jul 2007.

[234] F. Eggenhofer, I. L. Hofacker, and C. Honer Zu Siederdissen. RNAlien - Unsupervised RNA family model construction. *Nucleic Acids Research*, 44(17):8433–8441, Sep 2016.

[235] F. Krueger. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. [`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`; unpublished].

[236] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011.

[237] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, Apr 2013.

[238] B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa, M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2017. *Nucleic Acids Research*, 45(D1):D635–D642, Jan 2017.

[239] C. Sloggett, N. Goonasekera, and E. Afgan. BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, 29(13):1685–1686, Jul 2013.

[240] E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, 43(Database issue):D130–137, Jan 2015.

[241] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, 14(2):178–192, Mar 2013.

[242] C. Tyner, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, C. Eisenhart, C. M. Fischer, D. Gibson, J. N. Gonzalez, L. Guruvadoo, M. Haeussler, S. Heitner, A. S. Hinrichs, D. Karolchik, B. T. Lee, C. M. Lee, P. Nejad, B. J. Raney, K. R. Rosenbloom, M. L. Speir, C. Villarreal, J. Vivian, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, 45(D1):D626–D634, Jan 2017.

[243] Youri Hoogstrate. An algorithm for prediction RNA 2D structures including K-turns. Technical report, Technical University of Delft, 2012. [MSc internship report; unpublished].

[244] Saira Ashraf, Lin Huang, and David M.J. Lilley. Sequence determinants of the folding properties of box C/D kink-turns in RNA. *RNA*, 23(12):1927–1935, 2017.

[245] Youri Hoogstrate. ncRNA sequencing fragmenten onder de loep met ncRNA Mapper. Technical report, ErasmusMC, 2011. [BSc thesis; unpublished].

[246] S. Sai Lakshmi and Shipra Agrawal. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Research*, 36(suppl_1):D173–D177, Sep 2007.

[247] Pankaj Kumar, Suresh B. Mudunuri, Jordan Anaya, and Anindya Dutta. tRFdb: a database for transfer RNA fragments. *Nucleic Acids Research*, 43(D1):D141–D145, 2015.

[248] Sameer Jhavar, Alison Reid, Jeremy Clark, Zsofia Kote-Jarai, Timothy Christmas, Alan Thompson, Christopher Woodhouse, Christopher Ogden, Cyril Fisher, Cathy Corbishley, Johann De-Bono, Rosalind Eeles, Daniel Brewer, and Colin Cooper. Detection of TMPRSS2-ERG translocations in human prostate cancer by expression profiling using GeneChip Human Exon 1.0 ST arrays. *The Journal of Molecular Diagnostics*, 10(1):50–57, January 2008.

[249] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk. International Cancer Genome Consortium Data Portal: a one-stop shop for cancer genomics data. *Database*, 2011(1):26, sep 2011.

[250] Carla Zijlstra and Willem Stoorvogel. Prostasomes as a source of diagnostic biomarkers for prostate cancer. *Journal of Clinical Investigation*, 126(4):1144–1151, apr 2016.

[251] V R Minciacchi, A Zijlstra, M A Rubin, and D Di Vizio. Extracellular vesicles for liquid biopsy in prostate cancer: where are we and where are we headed? *Prostate Cancer and Prostatic Diseases*, 20(3):251–258, apr 2017.

[252] Saskia Hiltemann, Youri Hoogstrate, Peter van der Spek, Guido Jenster, and Andrew Stubbs. iReport: a generalised Galaxy solution for integrated experimental reporting. *GigaScience*, 3(1):1–8, 2014.

[253] Elena S. Martens-Uzunova, Youri Hoogstrate, Anton Kalsbeek, Bas Pigmans, Mirella Vredenbregt-van den Berg, Natasja Dits, Søren Jensby Nielsen, Adam Baker, Tapio Visakorpi, Chris Bangma, and Guido Jenster. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget*, 6(19):17430–17444, jul 2015.

[254] Lale Erdem-Eraslan, Martin J. van den Bent, Youri Hoogstrate, Hina Naz-Khan, Andrew Stubbs, Peter van der Spek, René Böttcher, Ya Gao, Maurice de Wit, Walter Taal, Hendrika M. Oosterkamp, Annemiek Walenkamp, Laurens V. Beerepoot, Monique C.J. Hanse, Jan Buter, Aafke H. Honkoop, Bronno van der Holt, René M. Vernhout, Peter A.E. Sillevis Smitt, Johan M. Kros, and Pim J. French. Identification of Patients with Recurrent Glioblastoma Who May Benefit from Combined Bevacizumab and CCNU Therapy: A Report from the BELOB Trial. *Cancer Research*, 76(3):525–534, 2016.

[255] C Zhang, J Bijlard, C Staiger, S Scollen, D van Enckevort, Y Hoogstrate, A Senf, S Hiltemann, S Repo, W Pipping, M Bierkens, S Payralbe, B Stringer, J Heringa, A Stubbs, LO Bonino Da Silva Santos, J Belien, W Weistra, R Azevedo, K van Bochove, G Meijer, JW Boiten, J Rambla, R Fijneman, JD Spalding, and S Abeln. Systematically linking tranSMART, Galaxy and EGA for reusing human translational research data. *F1000Research*, 6, 2017.

[256] Bérénice Batut, Saskia Hiltemann, Andrea Bagnacani, Dannon Baker, Vivek Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Brillet-Guéguen, Martin Čech, John Chilton, Dave Clements, Olivia Doppelt-Azeroual, Anika

Erxleben, Mallory Ann Freeberg, Simon Gladman, Youri Hoogstrate, Hans-Rudolf Hotz, Torsten Houwaart, Pratik Jagtap, Delphine Larivière, Gildas Le Corguillé, Thomas Manke, Fabien Mareuil, Fidel Ramírez, Devon Ryan, Florian Christoph Sigloch, Nicola Soranzo, Joachim Wolff, Pavankumar Videm, Markus Wolfien, Aisanjiang Wubuli, Dilmurat Yusuf, James Taylor, Rolf Backofen, Anton Nekrutenko, and Björn Grüning. Community-Driven Data Analysis Training for Biology. *Cell Systems*, 6(6):752–758.e1, 2018.

[257] Malgorzata A Komor, Linda JW Bosch, Gergana Bounova, Anne S Bolijn, Pien Delis van Diemen, Christian Rausch, Youri Hoogstrate, Andrew P Stubbs, Mark de Jong, Guido Jenster, et al. Consensus molecular subtypes classification of colorectal adenomas. *The Journal of Pathology*, 246(3):266–276, Aug 2018.

[258] Sujun Chen, Vincent Huang, Xin Xu, Julie Livingstone, Fraser Soares, Jouhyun Jeon, Yong Zeng, Fouad Yousif, Yuzhe Zhang, Nilgun Donmez, Musaddeque Ahmed, Haiyang Guo, Stas Volik, Anna Lapuk, Jessica Petricca, Melvin L.K. Chua, Lawrence E. Heisler, Natalie S. Fox, Michael Fraser, Vinayak Bhandari, Yu-Jia Shiah, Michele Orain, Valerie Picard, Helene Hovington, Alain Bergeron, Louis Lacombe, Yves Fradet, Bernard Tetu, Stanley Liu, Felix Feng, Malgorzata A. Komor, Cenk Sahinalp, Colin Collins, Youri Hoogstrate, Mark de Jong, Remond J.A. Fijneman, Teng Fei, Guido Jenster, Theodorus van der Kwast, Robert G. Bristow, Paul C. Boutros, and Housheng Hansen He. Widespread and functional rna circularization in localized prostate cancer. *Cell*, 2018. [Accepted].

# 9 | Appendices

**9.1: Curriculum Vitae**

**9.2: PhD Portfolio**

**9.3: List of Publications**

## 9.1 Curriculum Vitae

Youri Hoogstrate was born on the $7^{th}$ of September 1987, Goes, The Netherlands. In 2005 he finished secondary school (HAVO) at Het Goese Lyceum, Goes. He then followed the bioinformatics (BSc) program at Hogeschool Leiden and graduated in 2011. The thesis was about small RNA-seq analysis and conservation of ncRNAs, which was awarded with the second prize *thesis of the year 2011* at Hogeschool Leiden. Thereafter, he started the MSc program Computer Science with the bioinformatics specialisation track at Leiden University and the Technical University of Delft and graduated in 2013. The research assignment during the MSc program was about RNA folding including K-turns, in collaboration with the bioinformatics department of TU-Delft. The MSc thesis was about understanding characteristics between biological and artificial protein sequences, in collaboration with the human genetics department of Leiden University Medical Center. During the MSc program he was employed at the Erasmus Medical Center, Rotterdam, as bioinformatician, mainly focusing on RNA-seq analysis. After the MSc program he started a PhD program on RNA-seq analysis in the Erasmus Medical Center as collaboration between the departments of Bioinformatics and Urology. During the PhD program he also contributed to serveral open-source projects that were not part of the thesis, such as *fastafs*, *CrossMap*, *RNA-STAR*, *vg* (genome variation graph), *bio-conda*, *HTSeq* and *galaxy*.

## 9.2 PhD portfolio

| Year | Attended courses | ECTS |
|---|---|---|
| 2015 | MolMed Course Basic and Translational Oncology<br>ErasmusMC | 1.8 |

| Year | Lectured Courses |
|---|---|
| 2014 | Galaxy Community Conference GCC2014<br>Baltimore, USA<br>*lectured:* RNA-seq courses |
| 2015 | TraIT Galaxy NGS course<br>ErasmusMC<br>*lectured:* RNA-seq practical |
| 2015 | Galaxy Community Conference GCC2015<br>The Sainsbury Laboratory, Norwich, UK<br>*lectured:* RNA-seq courses |
| 2015 | Workshop Galaxy, BSc nanobiology course Delft<br>TU-Delft, Delft<br>*lectured:* Galaxy and RNA-seq |
| 2015 | BioSB RNA-seq $5^{th}$course<br>LUMC<br>*lectured:* small RNA-seq and Galaxy courses |
| 2014 | MolMed Galaxy course<br>ErasmusMC<br>*lectured:* RNA-seq tuxedo and Enhanced RNA-seq |
| 2016 | NGS Data Analysis in Galaxy (CTMM-TraIT course)<br>VUmc, Amsterdam<br>*lectured:* entire course |
| 2016 | MolMed Galaxy course<br>ErasmusMC<br>*lectured:* RNA-seq |

| | |
|---|---|
| 2016 | BioSB RNA-seq $6^{th}$course |
| | LUMC |
| | *lectured:* small RNA-seq and Galaxy courses |
| 2017 | Galaxy for NGS |
| | ErasmusMC |
| | *lectured:* MolMed Galaxy course |
| 2017 | ELIXIR European Galaxy Developer Workshop |
| | Strassbourg, FR |
| | *lectured:* Visualisations and tool development |

| **Year** | **Attended Conferences** | |
|---|---|---|
| 2014 | Galaxy Community Conference GCC2014 | Baltimore, MD, USA |
| 2015 | Galaxy Community Conference GCC2015 | Norwich, UK |
| 2016 | BioSB 2016 | Lunteren, NL |
| 2016 | ECCB2016 | Den Haag, NL |

| **Year** | **Supervised Students** |
|---|---|
| 2014 | Hina Naz-Khan (MSc thesis) |
| 2014 | Adam van Adrichem (MSc Research Assignment) |

**Contributions to scientific open-source software**

bioconda recipes: *build-recipes for bioconda package manager*
CrossMap: *Modern implementation of liftOver*
featureCounts: *Tool that quickly estimates read-counts in BAM files*
fusionCatcher: *RNA-seq fusion gene detection tool*
Galaxy: *Web-portal for running scientific analysis tools*
RNA-STAR: *RNA-seq aligner*
VG: *graph based genome*

**Maintainer of scientific open-source software**

bam-lorenz-coverage: *generate Lorenz plots and Coverage plots directly from BAM files*
dr-disco: *detecting genomic breakpoints of fusion transcripts in random primed RNA-seq data*
fastafs: *fuse layer between compressed and true FASTA files*
FlaiMapper: *detection of small ncRNA derived fragments*
FuMa: *reporting overlap in RNA-seq detected fusion genes*
Galaxy IUC-tools: *Main Galaxy Tool shed repository*
HTSeq: *Tool and library for gene counting and genomic indices*
ncRNA Mapper: *Investigation of conservation and folding of ncRNA derived fragments*
segmentation-fold: *Modification to classical RNA folding algorithm to include K-turns and loop-E-motifs*

## 9.3 List of publications


1. Youri Hoogstrate, Guido Jenster, and Elena S. Martens-Uzunova. FlaiMapper: Computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data. *Bioinformatics*, 31(5):665–673, 2015

2. Saskia Hiltemann, Youri Hoogstrate, Peter van der Spek, Guido Jenster, and Andrew Stubbs. iReport: a generalised Galaxy solution for integrated experimental reporting. *GigaScience*, 3(1):1–8, 2014

3. Elena S. Martens-Uzunova, Youri Hoogstrate, Anton Kalsbeek, Bas Pigmans, Mirella Vredenbregt-van den Berg, Natasja Dits, Søren Jensby Nielsen, Adam Baker, Tapio Visakorpi, Chris Bangma, and Guido Jenster. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget*, 6(19):17430–17444, jul 2015

4. Y. Hoogstrate, R. Bottcher, S. Hiltemann, P. J. van der Spek, G. Jenster, and A. P. Stubbs. FuMa: reporting overlap in RNA-seq detected fusion genes. *Bioinformatics*, 32(8):1226–1228, Apr 2016

5. Lale Erdem-Eraslan, Martin J. van den Bent, Youri Hoogstrate, Hina Naz-Khan, Andrew Stubbs, Peter van der Spek, René Böttcher, Ya Gao, Maurice de Wit, Walter Taal, Hendrika M. Oosterkamp, Annemiek Walenkamp, Laurens V. Beerepoot, Monique C.J. Hanse, Jan Buter, Aafke H. Honkoop, Bronno van der Holt, René M. Vernhout, Peter A.E. Sillevis Smitt, Johan M. Kros, and Pim J. French. Identification of Patients with Recurrent Glioblastoma Who May Benefit from Combined Bevacizumab and CCNU Therapy: A Report from the BELOB Trial. *Cancer Research*, 76(3):525–534, 2016

6. Michael Olvedy, Mauro Scaravilli, Youri Hoogstrate, Tapio Visakorpi, Guido Jenster, and Elena Martens-Uzunova. A comprehensive repertoire of tRNA-derived fragments in prostate cancer. *Oncotarget*, 7(17):24766–24777, 2016

7. Youri Hoogstrate, Chao Zhang, Alexander Senf, Jochem Bijlard, Saskia Hiltemann, David van Enckevort, Susanna Repo, Jaap Heringa, Guido Jenster, Remond J.A. Fijneman, Jan-Willem Boiten, Gerrit A. Meijer, Andrew Stubbs, Jordi Rambla, Dylan Spalding, Sanne Abeln, Youri Hoogstrate, Chao Zhang, Alexander Senf, Jochem Bijlard, Saskia Hiltemann, David van Enckevort, Susanna Repo, Jaap Heringa, Guido Jenster, Remond J.A. Fijneman, Jan-Willem Boiten, Gerrit A. Meijer, Andrew Stubbs, Jordi Rambla, Dylan Spalding, and Sanne Abeln. Integration of EGA secure data access into Galaxy. *F1000Research*, 5(0):1–9, 2016

8. Björn A. Grüning, Jörg Fallmann, Dilmurat Yusuf, Sebastian Will, Anika Erxleben, Florian Eggenhofer, Torsten Houwaart, Bérénice Batut, Pavankumar Videm, Andrea Bagnacani, Markus Wolfien, Steffen C. Lott, Youri Hoogstrate, Wolfgang R. Hess, Olaf Wolkenhauer, Steve Hoffmann, Altuna Akalin, Uwe

Ohler, Peter F. Stadler, and Rolf Backofen. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. *Nucleic Acids Research*, 45(W1):W560–W566, 2017

9. C Zhang, J Bijlard, C Staiger, S Scollen, D van Enckevort, Y Hoogstrate, A Senf, S Hiltemann, S Repo, W Pipping, M Bierkens, S Payralbe, B Stringer, J Heringa, A Stubbs, LO Bonino Da Silva Santos, J Belien, W Weistra, R Azevedo, K van Bochove, G Meijer, JW Boiten, J Rambla, R Fijneman, JD Spalding, and S Abeln. Systematically linking tranSMART, Galaxy and EGA for reusing human translational research data. *F1000Research*, 6, 2017

10. Bérénice Batut, Saskia Hiltemann, Andrea Bagnacani, Dannon Baker, Vivek Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Brillet-Guéguen, Martin Čech, John Chilton, Dave Clements, Olivia Doppelt-Azeroual, Anika Erxleben, Mallory Ann Freeberg, Simon Gladman, Youri Hoogstrate, Hans-Rudolf Hotz, Torsten Houwaart, Pratik Jagtap, Delphine Larivière, Gildas Le Corguillé, Thomas Manke, Fabien Mareuil, Fidel Ramírez, Devon Ryan, Florian Christoph Sigloch, Nicola Soranzo, Joachim Wolff, Pavankumar Videm, Markus Wolfien, Aisanjiang Wubuli, Dilmurat Yusuf, James Taylor, Rolf Backofen, Anton Nekrutenko, and Björn Grüning. Community-Driven Data Analysis Training for Biology. *Cell Systems*, 6(6):752–758.e1, 2018

11. Malgorzata A Komor, Linda JW Bosch, Gergana Bounova, Anne S Bolijn, Pien Delis van Diemen, Christian Rausch, Youri Hoogstrate, Andrew P Stubbs, Mark de Jong, Guido Jenster, et al. Consensus molecular subtypes classification of colorectal adenomas. *The Journal of Pathology*, 246(3):266–276, Aug 2018

12. Sujun Chen, Vincent Huang, Xin Xu, Julie Livingstone, Fraser Soares, Jouhyun Jeon, Yong Zeng, Fouad Yousif, Yuzhe Zhang, Nilgun Donmez, Musaddeque Ahmed, Haiyang Guo, Stas Volik, Anna Lapuk, Jessica Petricca, Melvin L.K. Chua, Lawrence E. Heisler, Natalie S. Fox, Michael Fraser, Vinayak Bhandari, Yu-Jia Shiah, Michele Orain, Valerie Picard, Helene Hovington, Alain Bergeron, Louis Lacombe, Yves Fradet, Bernard Tetu, Stanley Liu, Felix Feng, Malgorzata A. Komor, Cenk Sahinalp, Colin Collins, Youri Hoogstrate, Mark de Jong, Remond J.A. Fijneman, Teng Fei, Guido Jenster, Theodorus van der Kwast, Robert G. Bristow, Paul C. Boutros, and Housheng Hansen He. Widespread and functional rna circularization in localized prostate cancer. *Cell*, 2018. [Accepted]

## 9.4 Dankwoord

Ik wil in de eerste plaats Prof. dr. ir. Guido Jenster bedanken voor alle support en mogelijkheden die hij mij heeft geboden gedurende de afgelopen 8 jaar. Dit betreft uiteraard, maar niet uitsluitend, de mogelijkheid tot het doen van een promotieonderzoek evenals de begeleiding tijdens mijn bachelor stage en het aanbieden van een baan als bio-informaticus tijdens het doen van een masterstudie. Daarnaast wil ik ook alle collega's en co-auteurs bedanken die het mede mogelijk gemaakt hebben de eindstreep te halen, met in het bijzonder Dr. René Böttcher, Dr. Remond J.A. Fijneman, Dr. Pim French, Dr. Björn Grüning, Saskia Hiltemann, Dr. Elena S. Martens-Uzunova, Prof. Dr. Peter van der Spek en Dr. Andrew P. Stubbs. Voor dit promotie onderzoek is bijna uitsluitend vrije open-source software gebruikt. Ik wil de betrokken ontwikkelaars die met deze mindset mijn werk medemogelijk hebben gemaakt ook graag bedanken. Uiteraard wil ik ook mijn familie en vrienden bedanken, met in het bijzonder mijn ouders en mijn aanstaande Desirée de Lege.