Author: Pulyakhina, Irina
Title: A telescope for the RNA universe : novel bioinformatic approaches to analyze RNA sequencing data
Issue Date: 2016-04-21

A telescope for the RNA universe: novel bioinformatic approaches to analyze RNA sequencing data

by Irina Pulyakhina

ISBN: 978-94-6182-645-9

Cover image: A bioinformatic insight into the world of RNA
Cover design and layout: M. Liem and I. Pulyakhina
Printed by: Offpage, *www.offpage.nl*

Publication of this book was financially supported, in part, by the Department of Human Genetics, LUMC.

A telescope for the RNA universe

novel bioinformatic approaches to analyze RNA sequencing data

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op donderdag 21 april 2016

klokke 10:00 uur

door

Irina Pulyakhina

geboren te Moskou, de Russische Federatie

in 1989

Promotiecommissie

Promotor:	prof.dr. J.T. den Dunnen
Co-promotor:	dr. P.A.C. 't Hoen

Overige leden:	prof.dr. P. de Knijff
	prof.dr. A.M. Aartsma-Rus
	prof.dr. L. Wessels[1,2]
	prof.dr. M.S. Gelfand[3]
	dr. B.T. Heijmans

[1] Nederlands Kanker Instituut (NKI), Amsterdam

[2] Technische Universiteit Delft (TU Delft), Delft

[3] Institute for Information Transmission Problems (IITP), Moscow, Russia

To my parents Natalya Sheremeta and Viktor Pulyakhin.

Contents

Chapter 1

Introduction

Bioinformatics is an interdisciplinary field that provides methods and tools for the interpretation of biological data. Being a bridge between biology and informatics (hence the term *bio-informatics*), this field combines knowledge from computer science, mathematics, statistics, and engineering and applies it on biological data. Unlike biology, bioinformatics itself as a field does not produce data, it performs data analysis and delivers information about given data using given or developed algorithms. Bioinformatics is applied to better understand fundamental cellular and molecular processes, to identify drug targets, to improve disease diagnostics and to support clinical decision making.

Bioinformatics is a rather new field and there is a continuous need for the development of tools to perform various types of analyses (an example of such a tool is provided in **Chapter 2**, and **Chapter 3** contains its further, more extensive biological application). However, the possibilities and opportunities of currently existing methods and tools are rarely explored to their full extent, and important biological and medical questions can already be answered using existing tools (examples of this are given in **Chapter 4** and **Chapter 5**). In this thesis, I focus on the development and application of bioinformatics tools for the analysis of genetic data. The basics of our current understandings of genetics and bioinformatics is summarized below.

1.1 Background information

1.1.1 DNA and RNA

The human genetic code is contained in *DNA*, or *deoxyribonucleic acid*, a double-stranded helical molecule located mainly in the nucleus and additionally in the mitochondria. The helix is formed from a repeated pattern of a monosaccharide sugar deoxyribose and a phosphate group. The bases attached to the backbone create a full DNA molecule. These bases can be four nucleotides: adenine **A**, cytosine **C**, guanine **G**, and thymine **T**. The double helix consists of two complementary antiparallel strands formed in such a way that the nucleotides **A:T** and **C:G** situated on the opposite strands are paired [1].

Genetic information used as a template for proteins is encoded in the regions of DNA called *genes*[1]. The complete DNA sequence including all of its genes is called *genome*, and the genetic makeup of a particular cell or organism under specific conditions is called *genotype.* During *transcription* (Figure 1.1, A) genes and sometimes other regions of DNA are copied into *RNA*[2] by the *RNA polymerase* enzyme.

[1] Note that, however, not all genes code for proteins.

[2] **R**ibo**n**ucleic **a**cid.

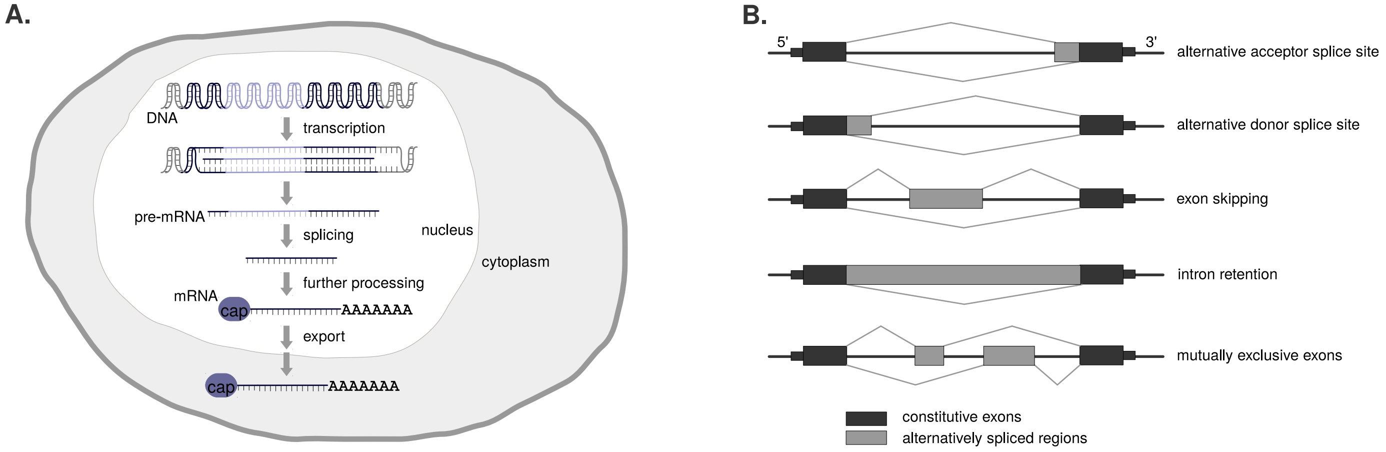


Figure 1.1: **A.** Eukaryotic transcription and mRNA processing. A eukaryotic cell; the light-blue region of the DNA and pre-mRNA molecules indicate introns; dark-blue regions of the DNA and RNA molecules indicate exons. The main steps of mRNA formation and processing – transcription, splicing, further processing (capping and polyadenylation) – are shown as well as mRNA transport to the cytoplasm, where mRNAs are translated into proteins. Note that the processing steps depicted here do not always take place in the indicated order. **B.** Main classes of alternative splicing events.

RNA is a single-stranded molecule similar to DNA in its content. However, instead of thymine **T**, as it is in DNA, RNA contains uracil **U**, and the backbone is formed from ribose instead of deoxyribose.

Different types of RNA have different roles in the cell (Table 1.1). One of the classes of RNA – messenger RNA or *mRNA* – is focused on in this thesis and will be discussed in more detail.

1.1.2 RNA maturation

In humans, RNA undergoes a series of steps before it can be used as a template for proteins. During its life, which begins with transcription, an RNA molecule is processed to mature mRNA, transported from the nucleus to the cytoplasm and then translated.

Table 1.1: Different types of human RNAs and their functions.

Name	Type	Function
mRNA	Messenger RNA	Templates for proteins
rRNA	Ribosomal RNA	Translation
tRNA	Transfer RNA	Translation, amino acid transport
snRNA	Small nuclear RNA	Splicing
snoRNA	Small nuclear RNA	Chemical modifications of RNA
Y RNA	Y RNA	RNA processing, DNA replication
TERC	Telomerase RNA Component	Telomere synthesis
pncRNA	promoter associated non-coding RNA	Transcription regulation
encRNA	enhancer associated non-coding RNA	Transcription regulation
lncRNA	Long non-coding RNA	Gene regulation
miRNA	Micro RNA	Gene regulation
piRNA	Piwi-interacting RNA	Gene regulation
siRNA	Small interfering RNA	Gene regulation

The transcription of mRNA is performed by RNA polymerases. Transcription is a multi-step process, which consists of pre-initiation, initiation, promoter clearance, elongation and termination. During pre- initiation, special proteins called *transcription factors* bind to *promoters* – regions of DNA situated upstream of the start of a gene. This facilitates the binding of RNA polymerases to the DNA, which triggers the next step of transcription – initiation. After the bond between the first two nucleotides has been created, the polymerase releases from the promoter (promoter clearance), and keeps synthesizing RNA using one DNA strand as a template (elongation) to produce the transcript. While RNA polymerase moves from the 3' to the 5' end on the template strand, the transcript is produced in a 5' towards 3' direction. The last step of transcription – termination, when the newly synthesized RNA, or precursor messenger RNA (*pre-mRNA*) is released from the RNA polymerase – is poorly understood.

After (or concurrently with) transcription pre-mRNA undergoes several processing steps: capping, splicing, polyadenylation and editing. During *capping*, which is initiated before the completion of transcription [2], the 5' terminal phosphate group of the nascent RNA molecule is enzymatically removed, GTP is added to the remaining bisphosphate with an unusual 5' to 5' triphosphate bond and is methylated. During *polyadenylation*, which often takes place after transcription, a big protein complex cleaves the 3'-most part of a newly produced RNA and polyadenylates (adds a sequence that consists of dozens or even hundreds of **A**'s) the end produced by this cleavage. During *editing*, which happens during or after transcription, certain nucleotides in the pre-mRNA or mRNA can be chemically modified [3] or deleted [4, 5].

Another processing step happening after or concurrently with transcription is pre-mRNA *splicing*. Unlike transcription, splicing is unique to higher eukaryotes, including humans. Splicing is performed by a big riboprotein complex, the *spliceosome*, which removes *introns* and joins the adjacent *exons* [6, 7]. Spliceosome finds and recognizes two *splice sites* – a donor (5' end of the intron) and an acceptor site (3' end of the intron). More intronic sequences – a *polypyrimidine tract* and a *branchpoint* – are also essential for efficient splicing, as they bind to the spliseosomal proteins. A polypyrimidine tract is a pyrimidine (**C** and **T**) rich area of approximately 20-50 nucleotides upstream the acceptor splice site. A branchpoint is an **A** nucleotide situated upstream from the polypyrimidine tract. After conformational changes within the spliceosome the adjacent exons that were separated by the introns are joined together by a transesterification reaction, and the intron leaves the complex as a lariat, which is usually quickly degraded [8, 9].

Genes can give rise to multiple mRNA transcripts, because pre-mRNA splicing can occur via multiple alternative routes [10, 11, 12, 13]. Depending on the nature of the cell or tissue or as a consequence of signals from the environment, the spliceosome may remove different regions from the same pre-mRNA molecule. This can happen within the same cell. Such variation in splicing is called *alternative splicing*, (Figure 1.1, B), creating a range of different transcripts from one gene. Five major classes of alternative splicing are known – exon skipping (which is the most prominent in humans), mutually exclusive exons (a combination of exon skipping events, where one of two exons is included in one mRNA molecule), intron retention (which is the least prominent in humans) and the usage of alternative donor or acceptor splice sites making exons longer or shorter [14, 15].

Alternative splicing is not the only way to create multiple transcripts from one gene [11]. Alternative polyadenylation or initiation of transcription at alternative start sites [14] also contribute to the variety of transcripts. Alternative transcription initiation, splicing and polyadenylation result in generating over 80,000 transcripts from just over 20,000 genes [16]. The set of all RNA molecules, including mRNA and non-coding RNA like rRNA and tRNA transcribed, is much larger. This set is referred to as the *transcriptome* [12]. However, in this thesis, we define a transcriptome to be only mRNA molecules present in one cell or a population of cells. The amount of identical (not alternative) transcripts generated from identical pre-mRNAs is called *transcript expression* level, and the overall expression of all transcripts from one gene is called *gene expression* level [17].

1.1.3 Principles of Illumina NGS

One way to analyze genomes and transcriptomes is massively parallel sequencing, usually referred to as next generation sequencing, or *NGS* [18, 19, 20]. Unlike previously used sequencing of one amplicon at a time (i.e., Sanger sequencing [21, 22]), NGS involves sequencing large amounts of DNA or RNA at once usually[3] on a whole genome or transcriptome scale. NGS technologies require a pool of DNA or RNA molecules [18] that can be isolated following various protocols depending on the research question. Since the majority of current NGS technologies cannot read beyond 100-250 nucleotides, full-length molecules are usually fragmented and (part of) the sequence of each short fragment is read. Short sequences determined during NGS are called *reads*.

One of the most popular and widely used sequencing technologies at the moment is Illumina [23]. All the analyses in this thesis has been performed on Illumina sequencing data. Sample preparations, protocol details and technical issues and biases described below therefore regard Illumina sequencing. Note that some of them can apply to the data from other sequencers as well.

A basic Illumina protocol for sequencing consists of the following steps (Figure 1.2, A). Double-stranded DNA molecules are fragmented. After fragmentation, artificially synthesized sequences called adapters are ligated to the short, size selected fragments, to facilitate PCR amplification. Denaturated amplified fragments are hybridized to the oligos that are attached to a water tight flowcell. Using bridge amplification, clusters of identical molecules are formed from each fragment. Four types of fluorescently labelled nucleotides[4] and a polymerase are added. The fluorescent label, which is also a block that prevents adding more than one nucleotide, is added per sequence cycle. Incorporation of a nucleotide into a fragment is detected using a CCD camera. The nucleotide is identified based on the signal detected from the fluorescent label. After imaging the fluorescent label is cleaved off, nucleotides are added again and the next nucleotide is incorporated. This process is repeated numerous times, and the number of repetitions defines the length of the fragments that will be sequenced. As every nucleotide has its own color coding, an automated image processing pipeline decodes the signal and produces text files – *fastq* files – containing the sequence and the quality of the individual nucleotide calls, or *Phred* quality score [24].

Phred quality score Q is a value logarithmically related to the base-calling error probability P.

$$Q = \textit{-10 log}_{10}\ P \tag{1.1}$$

E.g., if Phred score of a base is 30, the chance that this base is called incorrectly is 1 in 1000. In case of RNA sequencing, or *RNA-Seq*, cDNA is synthesized from RNA. This step introduces a lot of biases in the data, one of which is the uneven coverage of the molecules due to non-random priming. After RNA has been extracted, the enzyme *reverse transcriptase* is used to create cDNA from the RNA template. Reverse transcriptase is used together with *random hexamers*, synthetic polynucleotide fragments that should in theory anneal to any part of the RNA molecule. However, partly due to the non-uniform sequence of the transcripts expressed in the human genome [25], the use of random hexamer priming for transcriptomics has been shown to result in a non-uniform coverage along the expressed transcripts [26].

After the synthesis and fragmentation [27] of cDNA the protocol for RNA sequencing resembles DNA sequencing.

The quality of Illumina reads is known to be very high (>99.9% probability of calling the correct nucleotide). Quality usually drops towards the end of the read, mainly due to a problem called phasing. As mentioned previously, sequencing works with blocked nucleotides. However, when blocking does not work perfectly, individual molecules in a cluster can get out of phase and have one nucleotide extra (or

[3] Note that NGS can also be targeted and explore variable loci at very high depth.

[4] Note that for Illumina HiSeq, two lasers are used to identify two colours, less and more intense green and red, however, four labels are added.

fewer, when no nucleotide is incorporated). This effect accumulates towards the 3' end of the read and the sequencer cannot correct the phasing anymore.

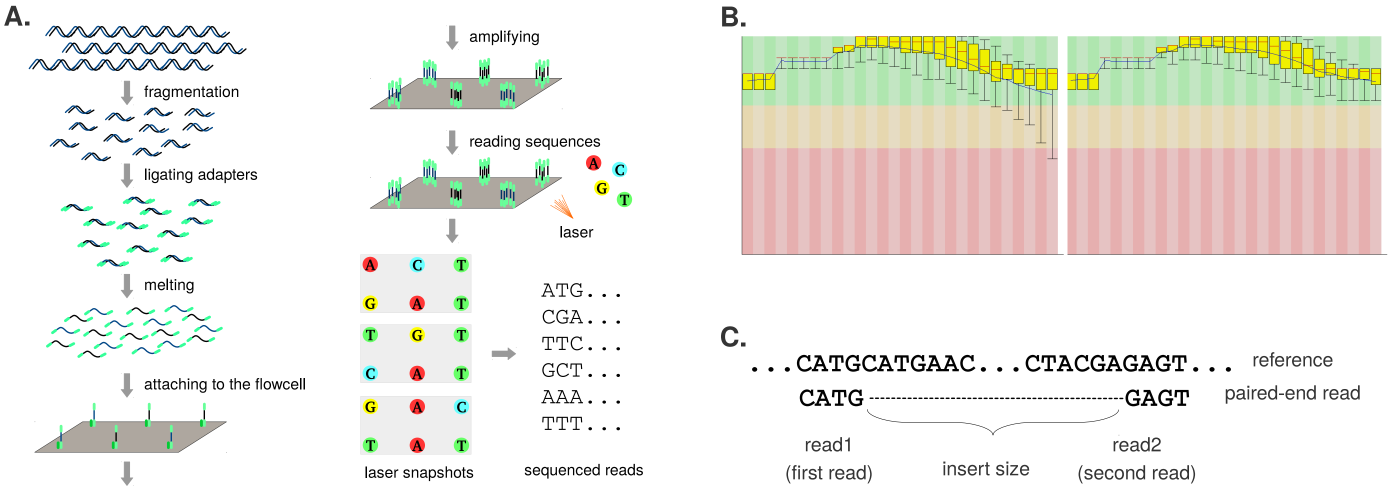


Figure 1.2: **A.** Main steps of Next Generation Sequencing performed by Illumina sequencers. For a detailed description of the process, see Section 1.1.3. **B.** Schematic representation of paired-end reads. Stretches of nucleotides flanked by dots represent two regions of a DNA or RNA molecule. Short stretches of nucleotides separated by a dashed black line (representing the insert size) represent two ends of a read pair. **C.** Example base quality report for an Illumina *fastq* file. Left panel shows the base quality plot of the raw reads received directly from the Illumina sequencer. Right panel shows the base quality report of the reads after clipping and trimming low quality bases. The drop of the base calling quality towards the 3' end of the read can be appreciated from the left panel.

The output of Illumina HiSeq sequencing runs (commonly used for transcriptome analyses) amounts to billions of reads [28, 29]. Maximum available length of one Illumina HiSeq read is currently limited to ˜250 bases. Such small length is a disadvantage in the downstream analysis, as short fragments can be very hard to locate on the genome or transcriptome, especially when mismatches with the reference sequence are allowed (see Section 1.2 for more detail). One way to increase read length is to use *paired-end* sequencing (Figure 1.2, B), when a fragment longer than the maximum Illumina read length is sequenced from both ends [30]. Since the approximate length of the fragment generated is known, the length of the region between two reads (or *ends*) can be inferred (only after the alignment has been generated, see Section 1.2 for more detail). This region is called the *insert size*. Sometimes the terms "inner" or "outer mapping coordinates" are used to describe the insert size. To learn more about how the information about insert sizes can be used during the bioinformatic analysis, see Section 1.2.

1.1.4 RNA-Seq protocols

The first important step in sequencing RNA is to isolate the RNA molecules of interest, i.e., mRNA, miRNA or pre-mRNA, from the rest of the cell. Since this thesis focuses mainly on mRNA, only this type of RNA will be discussed.

Enrichment for mRNA can be achieved using oligo-dT capture or rRNA depletion. The oligo-dT method uses oligo-dT primers immobilized on beads to capture polyA tails of mRNAs (and some long non-coding RNAs with polyA tails, see Section 1.1.1). The method enriches for mRNA quite effectively, and the amount of other types of RNA is minimal. The second widely used way of enriching for mRNA is to deplete rRNA which comprises around 90% of total RNA, by selectively eliminating rRNA. However, since this technique does not specifically target polyA tails of full mRNA molecules, the amount of RNA types other than mRNA is usually higher when using this protocol compared to the oligo-dT protocol.

Since RNA can not be sequenced directly on the Illumina platform (as mentioned in Section 1.1.3), it is necessary to create a DNA copy. One of the most common ways to generate cDNA is to use a reverse transcriptase enzyme and a mixture of random hexamer primers – 6 bases long oligonucleotide synthetic sequences with random nucleotide compositions [31]. These hexamers anneal to random regions on the RNA and work like primers for the reverse transcriptase enzyme, which starts elongating the DNA sequence. The step of synthesizing cDNA is prone to errors occuring in the cDNA product due to various reasons. E.g., the ability of reverse transcriptase to *switch templates* is known to affect cDNA synthesis [32, 33, 34]. During the reverse transcription, the nascent cDNA fragment can periodically dissociate from the RNA template and reanneal to a different stretch of RNA with the sequence similar to the original RNA template. Such template switching might lead to generating artificial cDNA and sequence reads.

Due to the factors mentioned above (and many more), a possibility of sequencing RNA directly, without cDNA synthesis, becomes very appealing and will be highlighted in the **Discussion** section of this thesis.

1.2 Bioinformatic analysis

Pre- and post-alignment data cleaning

Since the quality of reads tends to drop towards the 3' end of the read [35], it is common practice to check the sequencing quality of each base, trim low-quality bases from the end of the read and proceed with trimmed reads (Figure 1.2, C). Another step in the pre-alignment phase is to check for the presence of the adapter sequences, which should be removed from the *fastq* files (introduced in 1.1.3) before running the downstream analysis (such as alignment).

Alignment

Aligning, or *mapping* reads to the reference sequence means finding the most probable location (origin) of the reads in the genome. Alignments can be exact (when all nucleotides from the read match the reference sequence) or contain insertions, deletions and substitutions. RNA-Seq reads can be aligned to the reference transcriptome, as well as to the reference genome. If transcriptome samples are aligned to the genome, some reads need to be split in order to allow mapping across exon-exon junctions and span the introns that are present in the reference genome. Mapping RNA-Seq reads to the transcriptome does not require special aligners, as the reads do not have to be split. A disadvantage of such approach is that the “reference transcriptome” is incomplete. Transcriptomes may differ between different cell types, experimental conditions or treatments. To have a proper alignment to the transcriptome, RNA-Seq reads would ideally be mapped to the transcriptome of the cells they were isolated from. Unfortunately, this information is not always available, therefore it is common practice to align RNA-Seq reads to the genome. Another problem is that a transcriptome contains multiple transcripts coming from the same gene, which makes alignment more challenging. However, when the whole genome sequence is not available and only the transcriptome is known, this alignment strategy may be used.

Alignment strategies

Numerous tools for paired-end RNA-Seq data alignment have been developed. They can be divided into two main categories. Alignment tools from the first category (Figure 1.3, A) use one end as an “anchor” and map it without splitting it [36]. After mapping non-split reads a list of “coverage islands” – regions where the majority of reads mapped to – is generated. Second ends and remaining unmapped reads are initially mapped to the coverage islands. When mapping the second end, the insert size (or *inner* and *outer* mapping coordinates) is taken into account for a more precise mapping. It works as a factor limiting the search for the second read location near the first read and not genome-wide. Additionally, consistent deviations of insert sizes are used to detect deletions, insertions and splicing events.

Another type of alignment tools (Figure 1.3, B) treats reads of one pair as individual reads, splits them and stores the pieces in a table (a *hash table*). A similar table for the reference sequence is created, and the aligner overlaps the two, records the best matches for separate pieces and tries to merge them to give a full alignment of one paired-end read as an output. Tools from both categories can often work with a list of known exon-exon junctions provided by the user. In this case, aligners will try to split reads only or preferably at positions of known junctions. Such an approach usually helps to decrease the number of false-positive alignments, but it introduces a bias towards known splice sites that can incapacitate the discovery of novel exon-exon junctions.

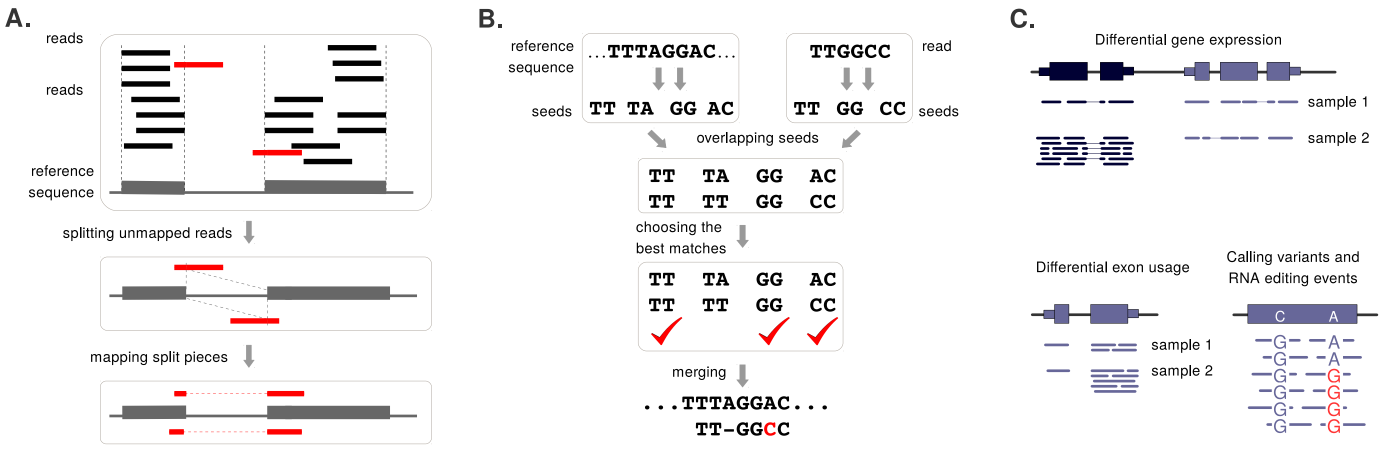


Figure 1.3: **A.** Aligning short reads using “anchor” strategy. Thick black lines represent reads mapped without splitting them. Thick red lines represent reads mapped after being split. Grey blocks represent coverage islands. For a detailed description of the procedure, see Section 1.2. **B.** Aligning short reads using a “hash table” strategy. For a detailed description of the procedure, see Section 1.2. **C.** Post-alignment bioinformatic analysis. In the top panel, differential expression analysis is schematically shown (in this example, differential gene expression analysis), where the gene in black (thick black blocks) is differentially expressed between samples 1 and 2 unlike the gene in blue. Thick lines depict reads and thin lines connecting the thick ones depict the area where the reads were split. In the bottom left panel, differential exon usage analysis is depicted. In the bottom right panel, variant calling on RNA is shown (in this case, identifying RNA editing events), where letters “G” in red depict nucleotides different from the reference sequence (in this case, “A”).

The choice of the aligner should always be guided by the research question. However, the second strategy is becoming more popular, and depending on the application it is used in different modes. For instance, if only known splice sites and quantifying their abundance is the aim of the study, aligners use pre-selected annotation and do not explore the reads mapped elsewhere. On the other hand, if finding novel splice sites, estimating intronic coverage and finding splice sites with low abundance is a priority, the second approach is used in the “annotation-free” mode.

Unique and non-unique alignments

An important parameter of aligners is the allowed number of alignments per read. Since the read length is limited and the reference contains regions that are not unique, there is a possibility of finding the same sequence in the reference genome or transcriptome multiple times [37], especially when allowing for
mismatches. *Unique* mapping mode means that only reads for which one location in the reference with a certain quality (number of mismatches, insertions and deletions) has been found, will be reported. Non-unique or *multiple* mapping mode means that all (or, optionally, a limited number of) alignments per read will be reported. The choice of the best mapping strategy always depends on the research question and will not be discussed here.

1.2.1 Post-alignment processing

Since most Illumina protocols include PCR amplification, there is a possibility of PCR duplicates being sequenced multiple times (to see which stages of sample preparation and sequencing include PCR amplification steps, see Sections 1.1.3 and 1.1.4). To avoid counting PCR duplicates as separate molecules, reads that are considered duplicates can be removed. Whether a paired-end read is a duplicate or not is usually decided based on its mapping coordinates coming from the alignment. When two paired-end reads have the same mapping coordinates at both ends, the read is considered a duplicate, as the probability of the same paired-end read being sequenced twice is rather low. One of the reads can be removed (sometimes regardless the sequence content). Removing such reads from RNA-Seq data is problematic for single end sequencing data [38], as they may also originate from different transcript molecules in case when a transcript is highly expressed. However, such reads occur less in paired-end sequencing because the length of the fragments is not constant.

Regardless of whether reads are mapped to the genome or to the transcriptome, alignment results are reported in a standard *sam* file format (*S*equence *A*lignment *M*ap) [39]. This file contains the identifier, the sequence of the read and the information about its mapping, i.e., mapping coordinates, number of matching and mismatching bases, etc. Another way of showing the alignment results using the information from the *sam* file is using the *coverage* information in *wiggle* files. Each nucleotide of the reference sequence has a certain coverage, which is the number of reads mapped to this position. Web-based (such as the UCSC genome browser [40]) and standalone (such as IGV [41, 42]) tools have been designed to visualize the alignment results for better visibility and easier interpretation.

Biases in NGS data

RNA-Seq is a great source of biological data that can be used for a wide range of applications. However, it is not error-free and suffers from some major biases. One of them is a *5'-3' bias* [43, 44] (see Section 1.1.4 for more details), which means that the 3' end of transcripts are overrepresented because those are the fragments containing the polyA tails [45, 46, 47].

Reference bias in NGS data [48, 49] means that the "perfect" alignment without mismatches is by default prioritized by aligners over an alignment with a mismatch. When a sample and therefore the reads contain a genetic variant, they have a lower probability to map to their original location than a read that does not contain a variant. This bias is also seen with respect to the reference sequence. When a read can be mapped to two locations, one of which is perfect, the aligner usually chooses the perfect alignment, which will filter out the cases when the read is coming from a different region and has a variant. One possibility to tackle this bias is to "mask" the reference sequence [46] – replace certain positions (i.e., know SNPs) of the reference sequence with **N**. However, such approaches tend to increases the number of ambiguous alignments.

Another common bias in NGS data is the *sequencing bias* [26]. One of the sources for that is the specifics of PCR amplification. The PCR amplification step necessary before sequencing is performed at the same temperature for all reads. Some reads might have a higher or lower GC content (the amount of **G** and **C** nucleotides), which can happen due to a number of reasons. PCR conditions favor the amplification of fragments with normal GC percentage and are not optimal for extreme, high GC/high AT fragments [50]. Thus difference in the GC content influences the optimal melting and annealing temperatures for the reads. Since only one condition can be used, this results in a better amplification of reads with an average GC content [51]. Another source of GC bias comes from generating clusters during sequencing, as AT rich molecules tend to make larger clusters that overlap with other clusters, therefore giving a weaker signal.

1.2.2 Post-alignment transcriptome analysis

Researchers analyzing RNA-Seq data may be interested in various features of the transcriptome, each requiring separate analysis strategies (Figure 1.3, C). Transcriptome analyses can be divided into two major categories: annotation-guided analysis – comparison to a reference transcriptome assembly, or an *annotation* – and *de novo* assembly and analyses of the transcripts, without any prior information about the composition of the transcriptome [52, 53, 54].

Annotation-based and annotation-free analyses

When RNA-Seq data comes from an organism that has a known reference sequence, gene, transcript and/or exon annotation – a list of all genes, transcripts or exons identified within a particular organism, tissue or cell culture – is often available as well. It is usually stored in a text file as a list of coordinates of all exons and introns identified in the transcriptome [53]. The analysis of an RNA-Seq sample can be restricted by the annotation, which implies that only known exons, transcripts or genes will be searched for. In other words, the number of reads supporting each exon, transcript or gene will be reported. A less rigorous way of using the annotation – "annotation-guided" – means that the known transcripts will be searched for as well as the novel or unknown ones.

If no annotation is available, transcripts can be assembled either fully in case of longer reads or partly, i.e., at the level of individual exons in case of shorter reads. When the transcripts are assembled, the downstream analysis is not different for an annotation-based or an annotation-guided mode. However, an annotation-free approach is less reliable and requires deeper sequencing and longer reads to provide trustworthy results [55]. Another way to get the transcript information is *de novo* transcript assembly (not to be confused with *de novo* genome assembly [56]). When no reference sequence is available, reads are assembled and the information about transcriptome composition can be obtained [57, 58, 59, 60].

Detecting alternative splicing

Two major approaches to identify alternative splicing events are coverage- and split reads-based approaches. Tools like DEXSeq [61], using the coverage-based approach, require the annotation; the coordinates of a certain exon are selected and the coverage of a region between these coordinates is calculated across all samples. If the region is covered significantly higher or lower in one group of samples compared to the other group(s), such exon points to a potential alternative splicing event and will be reported as *differentially used*. DEXSeq tries to account for possible technical artifacts and considers the variation between samples within a sample group (a group of technical or biological replicates). The purely coverage-based approach models the influence of technical and biological replicates very accurately. However, as it deals with exons and not transcripts, it is often not clear to which transcript a particular differentially used exon belongs.

Another way to detect alternative splicing events is to combine the information from split reads with the coverage information. A popular tool in this category is Cufflinks [55, 62]. Split reads come from two regions that are adjacent in mRNA but not adjacent in DNA (i.e., derived from two exons), thus positions of splits indicate putative splicing events.

As both approaches have their advantages and disadvantages, choosing the best tool to detect alternative splicing depends on the research question [54]. Methods considering both coverage- and split read-based approaches have the highest precision. Information from the coverage, when more reads are considered (this increases the power of statistics), complements the information from split reads, which are only ˜20% of all reads, but nevertheless help to improve the mapping of exon boundaries.

Alternative polyadenylation

A transcript can have multiple polyA sites that can be used under different conditions or in different tissues [63, 64]. This phenomenon, also known as *alternative polyadenylation*, is another mechanism to expand the range of transcripts from one gene (Figure 1.4, A).

Polyadenylation takes place in the 3' UTR, and many RNA-Seq methods dedicated to determine polyA sites are available at the moment [65]. Bioinformatic tools [66, 67] usually look at the coverage in the 3' UTR, identify where the drops in coverage are (i.e., using a *sliding window approach*) and use this to predict the positions of polyA sites [68, 69] (Figure 1.4, B). Sometimes only known polyA sites [70, 71] are considered, and the coverage of these polyA sites or the areas upstream of the polyA sites is compared [72]. Since these are purely coverage-based methods, coverage biases are a significant issue in this type of analysis and should always be taken into account prior to the analysis [73]. It happens partly due to the length of the 3' UTRs, which is often comparable to the gene length.

1.2.3 Gene and transcript quantification

Gene expression is often measured in *RPKM* – *R*eads mapped *P*er *K*ilobase of the targeted region per *M*illion mapped reads. The number of reads C mapped to a certain gene is divided by the total number of nucleotides in this gene L divided by one thousand; the result will be divided by the total number of mapped reads N and multiplied by one million.

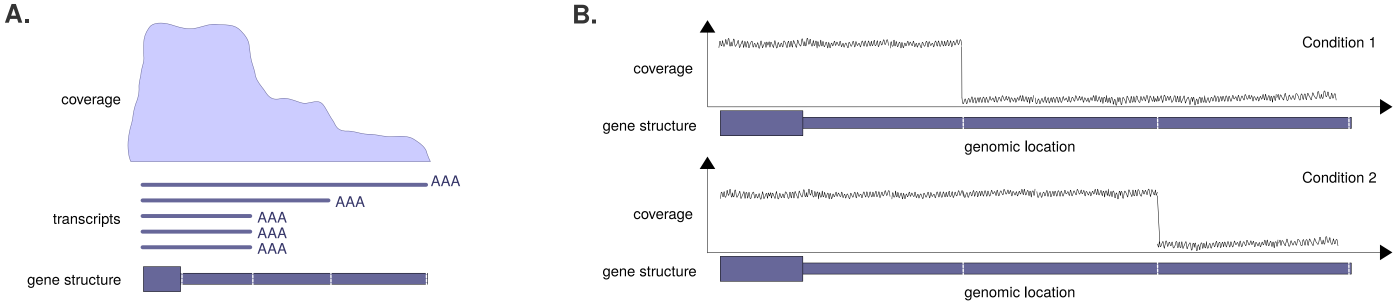


Figure 1.4: **A.** A schematic representation of alternative polyadenylation: gene structure (the bottom panel) with dashed white lines indicating the alternative polyadenylation sites; transcripts (middle panel) and their relative expression (the shortest transcript is expressed three times higher than each of the other two); gene coverage obtained from the sequencing data (the top panel). **B.** The reflection of alternative polyadenylation in RNA-Seq data: the upper panel shows the coverage pattern under condition 1, and the coverage drops at the first polyadenylation site (first dashed line); the lower panel shows the coverage pattern under condition 2, and the coverage drops at the second polyadenylation site (second dashed line).

$$RPKM = \frac{10^6 C}{NL/10^3} = \frac{10^9 C}{NL} \tag{1.2}$$

This measure has been introduced to account for gene length and the number of mapped reads and to make the expression levels comparable across genes and across samples.

RPKM has been revisited for paired-end reads, since paired-end reads are considered as two individual reads in the RPKM formula. Therefore, RPKM would be artificially increased by a factor of two. Another measure has been introduced – *FPKM*, or *F*ragments assigned *P*er *K*ilobase of the targeted region per *M*illion mapped reads. FPKM counts fragments – in case of single-end data this is single reads, and in case of paired-end data this is paired-end reads. Thus, FPKM is used more often nowadays, as it makes the single-end and paired-end RNA-Seq experiments comparable.

Differential expression analysis

Differential gene, transcript or exon expression analysis involves comparing expression levels across (groups of) samples and selecting genes that have significantly different expression between the groups [74, 75]. At first, gene expression per sample is usually normalized to such factors as sequencing depth and transcript length. For that, genes known to have stable expression at the majority of conditions and tissues are used. Next, the intragroup variation (variation within one group of samples, biological or technical replicates) is compared to the intergroup variation (variation between different groups) [52, 76].

When measuring gene expression, or counting reads mapped to a gene, a distribution to describe mapped reads has to be chosen[5] in order to perform statistical analyses. The Poisson distribution is often used to describe RNA-Seq data, as only a limited number of events – molecules that are isolated and sequenced – are selected out of a much bigger pool of events – all RNA molecules present in the cell. The Poisson distribution is generally used on count data and describes the variation that occurs due to the sampling procedure, but does not account for the biological or the technical variation due to the sample preparation. In order to describe the technical and the biological variation, the *overdispersion* parameter (which the Poisson distribution does not contain) is introduced. Overdispersion accounts for the presence of greater variability in a data set than is expected based on a given statistical model. A Poisson mixture model, or the negative binomial distribution, contains this parameter and is therefore used to describe gene expression data.

The estimation of the parameters for the appropriate statistical model is followed by a statistical test to find significant changes in gene expression between conditions. Such tools as edgeR [77] and DESeq [78] use a variant of the log *likelihood* ratio test adapted to work with the data following the negative binomial distribution. Other packages [79] use variations of a T-test to calculate *p-values* or estimate likelihood. All methods also use standard approaches for multiple testing correction (for example, Benjamini-Hochberg) [80].

1.2.4 Post-alignment analysis of variants

A genomic variant is a nucleotide present in DNA of the analyzed sample which differs from the nucleotide at the same position in the reference sequence. Searching for (or *calling*) variants is usually performed on DNA data. However, RNA-Seq data can also be used to detect genetic variation. RNA-Seq contains the information about expression and therefore allows for the study of the expression of the two different alleles in case of heterozygous variants. Differential allelic expression refers to the imbalanced (different from 0.5) expression of the two alleles. However, RNA-based variant calling comes with a series of complications.

Genomic variants and RNA modifications

Since DNA is transcribed directly to RNA, there should be no difference in variants called from DNA or RNA (unless the variant is not expressed). However, post-transcriptional RNA modifications – called *RNA editing* – make variant analysis a challenging task [81, 82]. RNA editing is a chemical change of a nucleotide in an RNA molecule that results in a different nucleotide. Therefore, while finding such a change in a number of reads, this can be mistaken for a genomic variant. Statistical filters for the DNA variant frequency performed to reduce the amount of false positives cannot be applied directly on the RNA-Seq data. Variants identified on DNA are normally present either on both alleles (a homozygous variant, and its frequency is 100%), or on one allele (a heterozygous variant, and its frequency is 50%).

Such filters on the frequency cannot be applied directly on RNA due to the possibility of *allele-specific expression* (ASE). Variant frequencies in RNA can vary from almost 0% (or higher than the

[5] http://arxiv.org/abs/1104.3889

sequencing error rate) to 100%. ASE can happen due to a number of reasons, i.e., pure monoallelic expression (which is an extreme case of allelic imbalance) can be a consequence of genetic imprinting or nonsense-mediated degradation (in case of a variant introducing a premature translation termination signal, or *stop codon*).

Variant calling on RNA

Calling variants on RNA is required for the analysis of ASE. One of the best ways to distinguish between RNA editing events and genomic variants in RNA-Seq data is to have DNA genotypes from the same sample. In this case the variants identified on DNA can be intersected with the variants identified on RNA. The variants identified on RNA which are not present on DNA can be considered as potential RNA editing events (REEs). Specific characteristics of RNA editing might also be used to identify these events more precisely. The most represented type of RNA editing is the hydrolytic deamination of **A** which results in a different nucleotide, inosine **I**. It will be interpreted as a **G** both by the enzymatic machinery of the cell and by the Illumina sequencer. Therefore, a way to exclude potential REEs when calling variants on RNA is to select **A** to **G** conversions that are not known genomic variants. Another possibility to identify REEs is to use the locations of known REEs [83]. However, this is not an optimal way, since such sources are always incomplete, and RNA editing can be tissue-, cell type- or condition specific. Therefore, identifying variants on DNA coming from the same sample as the RNA-Seq data is so far the best way to reliably call variants on RNA.

RNA editing is an interesting and not well-described biological process. It does not always occur on all RNA molecules and in general can target from almost 0% to 100% of the transcripts (although RNA editing usually affects a minority of transcripts and is often position-inspecific [3]). The role of RNA editing has not been elaborately described yet, thus the interpretation of the results of the bioinformatic analysis remains a challenge. This will be extensively discussed in the **Discussion** section of this thesis.

1.3 Outline of the thesis

In this thesis, I focus on the bioinformatic analysis of various types of RNA-Seq data both for medical applications and tried to answer fundamental biological questions. Existing tools, sometimes requiring modification, were used along with newly developed pipelines. **Chapter 2** covers the developing of a pipeline to analyze nuclear pre-mRNA and seek for biological phenomena as non-sequential and multi-step intron splicing. **Chapter 3** continues the research started in Chapter 2 and applies the pipeline (with extensions and more biological interpretation) on the largest and one of the most complicated genes in the human body – the *DMD* gene coding for the dystrophin protein. **Chapter 4** continues exploring splicing in order to better understand a fundamental biological process of human aging and its relation to RNA processing. **Chapter 5** highlights another type of bioinformatic RNA analysis – genomic variants and their differential allelic expression – using RNA from tumors of breast cancer patients (and control samples from the same patients). These chapters aim to cover main directions of the modern RNA-Seq analysis field and show that existing tools often, but not always, provide extensive information about the transcriptome, and that to answer certain biological questions, we had to develop our own tools. Finally, I close this thesis by providing an outlook on ways the field of transcriptomics and RNA-Seq analysis can evolve.

Chapter 2

SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing

I.Pulyakhina[1], I. Gazzoli[1], P.A.C. 't Hoen[1], N.E. Verwey[1], J.T. den Dunnen[1,2], A. Aartsma-Rus[1], J.F.J. Laros[1,2]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
2 Leiden Genome Technology Center, Leiden University Medical Center, Leiden, the Netherlands

Nucleic Acids Research, 2015 Jul 13;43(12):e80. doi: 10.1093/nar/gkv242

2.1 Abstract

Alternative splicing is a powerful mechanism present in eukaryotic cells to obtain a wide range of transcripts and protein isoforms from a relatively small number of genes. The mechanisms regulating (alternative) splicing and the paradigm of consecutive splicing has recently been challenged, especially for genes with a large number of introns. RNA-Seq, a powerful technology using deep sequencing in order to determine transcript structure and expression levels, is usually performed on mature mRNA, therefore not allowing detailed analysis of splicing progression. Sequencing pre-mRNA at different stages of splicing potentially provides insight into mRNA maturation. Although the number of tools that analyze total and cytoplasmic RNA in order to elucidate the transcriptome composition is rapidly growing, there are no tools specifically designed for the analysis of nuclear RNA (which contains mixtures of pre- and mature mRNA). We developed dedicated algorithms to investigate the splicing process. In this paper we present a new classification of RNA-Seq reads based on three major stages of splicing: pre-, intermediate- and post-splicing. Applying this novel classification we demonstrate the possibility to analyze the order of splicing. Furthermore, we uncover the potential to investigate the multi step nature of splicing, assessing various types of recursive splicing events. We provide biological insight into the order of splicing, show that non-sequential splicing of certain introns is reproducible and coinciding in multiple cell lines. We validated our observations with independent experimental technologies and showed the reliability of our method. The pipeline, named *SplicePie*, is freely available online[1]. The example data can also be found online[2].

2.2 Introduction

During messenger RNA (mRNA) splicing, introns are removed and exons are joined to generate a mature mRNA or a transcript. One transcript, usually chosen arbitrarily, to which all other transcripts are compared is called the "reference transcript". The deviations from this standard reference transcript are called alternative transcripts.

Data obtained with recent technologies, such as next generation sequencing (NGS) of a whole transcriptome, revealed that around 90% of human genes undergo alternative splicing [12, 84]. Alternative splicing of a gene comes in different flavours, such as skipping (a part of) an exon [85] or intron retention [86].

Splicing generally occurs cotranscriptionally [87, 88], and alternative splicing can be influenced by the speed of transcription [89]. It has been reported that splicing does not always progress sequentially from the 5' to the 3' end of a gene and with the same speed [90, 91]. Instead, different regions can be spliced rapidly or slowly [92]. A study of pig liver cells showed that for the *pCLEC4G* gene the first intron is spliced simultaneously with several more distal introns, while the second intron is spliced last [93].

Non-sequential intron removal can have unexpected consequences when mutations disrupt splice sites [94]. The splicing of the *COL1A1* gene region between exons 5 and 10 can follow two different routes. Removing introns 5, 6 and 9 is always rapid, while the excision of intron 8 can be before or after intron 7 [95]. Additionally, point mutations in splice signals were shown to cause skipping of exon 8, inclusion of intron 8 or inclusion of both introns 7 and 8. It is not clear yet which factors control the order of splicing. Theoretically, it might be influenced by the presence of intronic and exonic splice suppressors or enhancers, the "strength" of the donor and acceptor splice sites and/or the size of the intron [96, 97].

[1] https://github.com/pulyakhina/splicing_analysis_pipeline

[2] https://barmsijs.lumc.nl/HG/irina/example_data.tar.gz

Recursive splicing is another recently acknowledged feature of the splicing process. The principle of recursive splicing is that an intron might not be spliced in one piece. Instead, the spliceosome can recognize an intra-intronic sequence opposed to the canonical exon-intron border as a splice site, hereby removing the adjacent exon. The rest of such a partly spliced intron will be removed later, at once or again in multiple pieces following the strategy described above. Another type of multi-step intron removal [98] involves the usage of non-canonical donor and non-canonical acceptor splice sites. In this case an inner piece of intron is spliced first and a semi-stable lariat loop structure is formed and will be degraded later [99].

Alternative splicing cannot be fully understood when only mature mRNA is analyzed. When pre-mRNA molecules from different stages of splicing (pre-, intermediate- and post-splicing forms) are captured, the process of splicing can be studied in more detail and previously unknown splicing events can be identified. This type of analysis has been very challenging until recently. However, the development of high-throughput NGS, which involves highly parallelized sequencing of DNA or RNA, enabled the whole transcriptome analysis at high resolution [100, 101, 102]. In contrast to microarray analysis, RNA-Seq is a non-targeted approach, allowing the discovery of novel splicing events. Nevertheless, the analysis of RNA-Seq data is still a challenge, since NGS experiments generally produce millions of relatively short fragments (reads), even after paired-end sequencing came into play [103]. The distance between the two sequenced ends of a read pair ("PE distance") can be calculated and further taken into account during alignment. Standard mRNA analysis of (paired-end) RNA-Seq data includes mapping reads to the reference sequence, assembling the transcriptome and identifying the transcripts (transcript deconvolution), which is often followed by counting transcript levels (transcript quantification).

While mRNA analysis programs and pipelines are suitable for finding novel exons and exon-exon junctions, they do not consider the presence of pre-mRNA. Instead they analyze the end-result of the splicing process, using mainly reads mapped to exons. For this reason these tools are not able to properly deal with mixtures of pre- and mature mRNA, such as found in nuclear RNA extracts.

In this paper we are showing that splicing mechanisms can be analyzed using RNA-Seq data in more detail than previously achieved. We present *SplicePie* – a pipeline which contains new, dedicated method to analyze the order of splicing and pinpoint putative introns undergoing recursive splicing. Applying this method we show that non-sequentially spliced introns can be identified even in a relatively fast spliced gene. We also identify non-sequentially spliced introns in a gene that have never been reported to undergo such splicing scenarios.

2.3 Materials and methods

2.3.1 Captured dataset

Fused myotubes from a healthy human muscle cell line were harvested and the nuclei were separated from the cytoplasm using a sucrose containing lysis buffer, Dounce homogenizer and ultracentrifugation, respectively. Nuclear and total RNA were isolated from the nucleus using the Nucleospin RNAII column from BioKe Kit. DNAse treatment (RNAse-Free DNAse set by Qiagen) was performed to avoid DNA contamination. Three micrograms of each sample were reverse transcribed into cDNA (SuperScript II reverse transcriptase by Invitrogen), fragmented to the range of 100-600 bp by sonicating these samples with two cycles of one minute (Covaris S220, Massachusetts, USA) and purified (QIAquick PCR purification kit by Qiagen).

To capture target sequences we followed the SureSelect XT Target Enrichment System for the Illumina Paired-End Sequencing Library (Agilent Technologies). Illumina adapters were ligated to the fragmented sequences after end repair and A-base tailing (blunting). Further purification steps (performed with the Agencourt AMPure XP beads in 1:1 ratio) eliminated unbound adapters and short fragments (<100 bp).

The library of probes to capture exons, introns and flanking regions of the target genes (*FXR1*, *CKLF* and *ACTB*) was designed with the Agilent Technologies eArray software[3], avoiding areas masked by repeat masker and using partially overlapping probes. The 120 bp length probes were biotinylated four replicates of each probe were designed to reach the required number of baits per library.

The designed library [104] was hybridized with the fragmented cDNA from nuclear RNA and total RNA for 24 hours, followed by a washing step and pull down of the biotinylated cDNA probes using streptavidin- coated magnetic beads. Eluted samples were amplified to allow for a multiplexed Illumina run. The samples were quantified with the Agilent 2100 Bioanalyzer and Agilent HS DNA Chip Kit (Agilent Technologies, USA). The samples were diluted to a concentration of 7 pM and loaded onto an eight-channel flowcell and sequenced with the Illumina HiSeq 2000(Illumina, USA). After sequencing, *fastq* files containing paired-end reads (read length of 100 bp) and the base quality information were generated with CASAVA version 1.1 and used for further analysis.

2.3.2 ENCODE dataset

An RNA-Seq dataset representing a subset of the long RNA-Seq sequencing from ENCODE/Cold Spring Harbor Lab was obtained from the CSHL Long RNA-Seq downloadable files archive (Long RNA-Seq archive from ENCODE/Cold Spring Harbor Lab repository[4]. This dataset ("ENCODE dataset") contains RNA-Seq data from human immortalised myelogenous leukemia cell line, two samples of the chromatin-associated RNA and one sample of total nucleus RNA. These samples were sequenced by the Illumina Genome Analyzer II paired-end sequencing technique with read lengths of 75 and 100 bp.

2.3.3 General information

Alignment was performed with GSNAP version 2012-07-12 using a probabilistic mapping approach (one alignment per read was randomly chosen in case of multiple mappings). Format conversions were done with Samtools version 0.1.18 [39] and in-house scripts. Statistical manipulations and calculations were performed using R version 2.15.1. The Ensembl [105] gene annotation was used for all post-alignment analyses. Only publicly available software was used for the analysis.

2.3.4 Classification

After alignment, reads were classified according to their splicing stage. The first classification step determines the type of region that the reads are mapped to: exon, intron, exon-exon junction or exon-intron boundary (Table 2.1).

Each end of a read pair is given a label (Figure 2.1) and a "mapping distance". Mapping distance can be calculated for mapped read pairs by measuring the inner distance between the two aligned ends of a read pair.

Apart from classifying reads into three main categories, we divide them into more specific subgroups. Read pairs classified as intermediate-splicing reads are used for the analysis of sequentiality if one end is mapped to the exon-exon junction and the second end is mapped to the adjacent upstream or downstream intron. Read pairs having ends that are split across anything but an annotated exon-exon junction are used to identify recursive splicing (ends of a pair are treated separately and the connection

[3] http://earray.chem.agilent.com

[4] http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCshlLongRnaSeq, file identifiers are:
wgEncodeCshlLongRnaSeqGm12878NucleolusTotal,
wgEncodeCshlLongRnaSeqK562ChromatinTotalRep3,
wgEncodeCshlLongRnaSeqK562ChromatinTotalRep4.

Table 2.1: Labels for reads based on the mapping location according to transcript annotation (j ¿ i).

Read start	Read end	Label
intron	intron	"int" (intronic)
exon$_i$	exon$_i$	"ex" (exonic)
exon$_i$	exon$_j$	"ex-ex" (exon-exon junction)
exon	intron	"ex-int" (exon-intron junction)

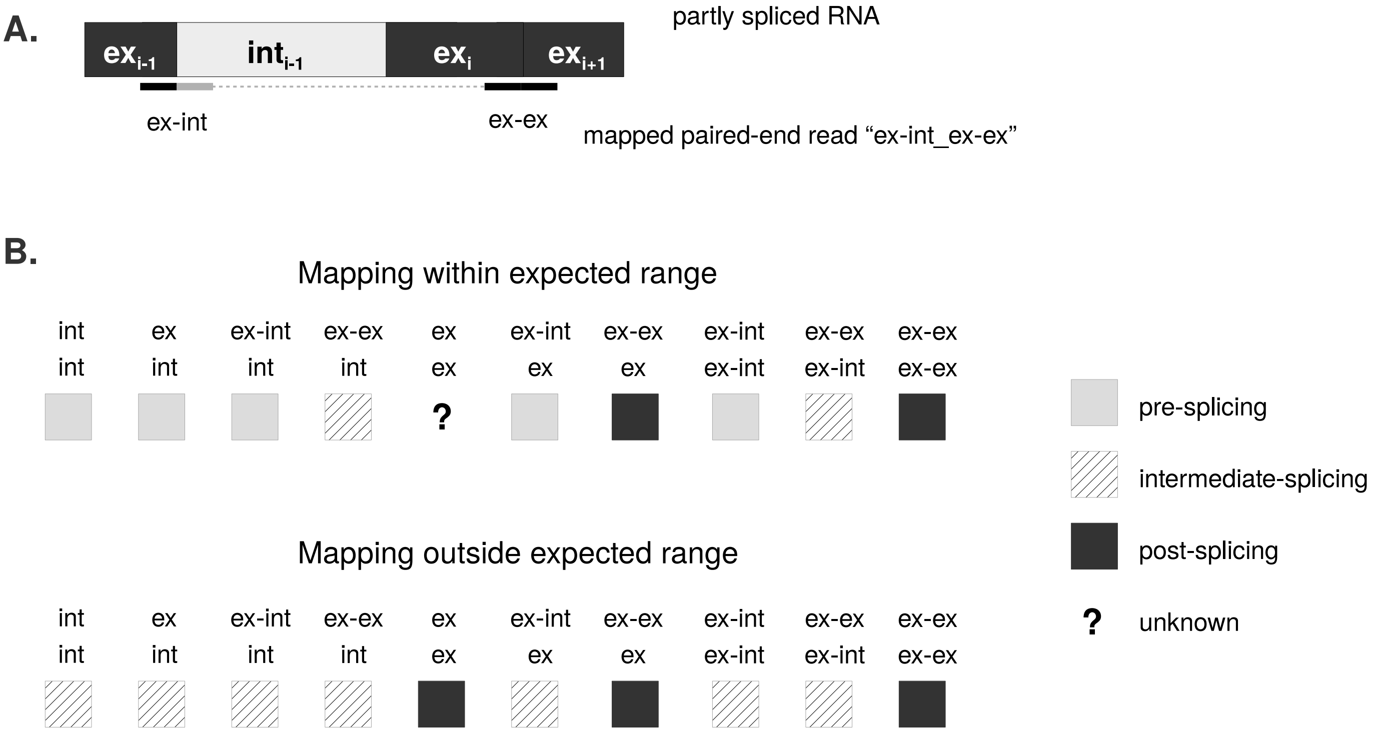


Figure 2.1: Classification of the paired end reads. **A.** "Mapping distance" reflects the inner distance between two read ends according to the genomic coordinates after alignment. **B.** Classification scheme is built on read labels ("ex" stands for exon, "int" for for intron, "ex-int" for intron-exon boundary, "ex-ex" for exon-exon junction) and mapping distance (within/outside expected mapping distance). Reads belonging to pre-, intermediate-, post-splicing and unknown categories are marked with gray, black, striped boxes and a question mark. Example: if one end of the read pair maps to the exon-intron boundary and the other one maps to the exon-exon junction, this read pair will be classified as "intermediate".

between the ends is not considered). Ends of read pairs from the pre-splicing category mapped to the exon-intron boundary and read pairs from the post-splicing category (mapped to the exon-exon junctions) are used to calculate the Splice Site Index.

2.3.5 Detection of non-sequential splicing

Non-sequential splicing is approached from two angles: coverage-based and read-based approaches.

For the coverage-based approach the difference between the median coverage of intron$_{i+1}$ and intron$_i$ is reported (Figure 2.3). We select negative differences and define the cutoff of significance as the first quartile *Q1* (25% percentile). If the difference is bellow *Q1*, this is an indication that intron$_{i+1}$ is non-sequentially spliced before intron$_i$. This is calculated for every input sample. If a certain difference is consistent across a number of samples, the corresponding pair of introns is reported.

The read-based approach addresses read pairs that indicate whether two introns are spliced sequentially or non-sequentially (Figure S1).

For this we defined the *splice-ratio*, a value that reflects the fraction of reads supporting sequential splicing:

$$splice\text{–}ratio = \frac{seq}{seq + non\text{-}seq} \tag{2.1}$$

Here *non-seq* is the number of reads supporting non-sequential splicing and *seq* is the number of reads supporting sequential splicing of two adjacent introns. When introns are spliced sequentially, the splice-ratio will be close to 1. However, when a downstream intron is spliced before an upstream intron, the splice-ratio will be close to 0.

We improve the accuracy of the predictions by assessing how often independent read pairs support the same pair of introns being spliced sequentially. Single events might indicate false positives due to mapping artifacts (Figure 2.3).

2.3.6 Detecting recursive splicing

We hypothesize that if an intron undergoes recursive splicing, split reads will map across intermediate splicing products. Recurring observations of these specific products confirm the existence of partially spliced introns. To extract such reads, we first identify potential hotspots for recursive splicing (Figure S2). For each position we calculate how many times a read has been split over it (in other words, we calculated the "inverted coverage" – coverage of gaps) and then calculate the derivative of this inverted coverage. The derivative indicates where the inverted coverage changes relatively to the previous positions, implying how many reads share a breakpoint at that specific location. Positive values represent a split's start, while negative values indicate a split's end. All other positions will have a derivative value of zero. Each peak is reported in a *wiggle* file. Positions which are start- and endpoints for splits at the same time are excluded from this analysis. Reads spanning exon-exon junctions are also removed, as they indicate annotated exon-exon junctions and do not contribute to the investigation of recursive splicing.

In order to reduce the amount of noise and false positives, we create *wiggle* files for all input samples and evaluate the overlap in the requested number of samples. If the position has positive coverage in a number of samples, the sum of the coverage from all files will be reported. This results in a single file with the most robust positions. Reads covering the positions from the final list are extracted from the initial *bam* file(s) and analyzed to validate the prediction of recursive splicing hot spots from the *wiggle* file and get the connections between the peaks (which are lost in the *wiggle* file).

Additionally, a text file containing a matrix with all discovered junctions and the number of reads supporting each junction is created per gene (Figure 2.4).

2.3.7 Calculating Splice Site Index and processing the coverage

We developed the Splice Site Index (SSI), a value used to detect splicing events, which is calculated in the following way:

$$SSI = \frac{ex\text{-}ex}{ex\text{-}ex + ex\text{-}int} \tag{2.2}$$

Here *SSI* is the splicing index value, *ex-ex* is the number of reads spanning exon-exon junctions and *ex-int* is the number of reads spanning exon-intron junctions.

A similar function, called completed Splicing Index (coSI), has been recently developed [6]. In contrast to coSI, SSI is calculated separately for 5' and 3' splice sites of each intron, allowing for the assessment of a) the relative abundance of each intron and b) whether both ends are spliced simultaneously. SSI, in combination with the median coverage of introns and exons, gives a more complete picture of the alternative splicing events. SSI is calculated as a ratio of different types of reads and does not consider the difference between the absolute values of coverage.

2.3.8 Experimental validation of non-sequential and recursive splicing

Complementary DNA (cDNA) from three different healthy muscle cell lines (also used in the *in silico* analysis) was generated using the reverse transcriptase, 1 μg of pre-mRNA and random primers (following the standard protocol suggested by SuperScript III RT by Invitrogen).

For the validation of non-sequential splicing, forward primers were designed against exon 10 or intron 11. Reverse primers were designed against exon 11 or the junction of exon 11 and exon 12. All combinations of primers were used to detect all splicing intermediates that can be formed in this area of the gene. Each sample was analyzed with three technical replicates and normalized against the *HPRT* gene. QPCR was performed on the LightCycler 480 (Roche Diagnostics Ltd.) using SYBR Green mix. QPCR results were analyzed using LightCycler 480 and LinRegPCR software [106]. Independent amplification, with primers in intron 10 and exon 12, was performed for further Sanger sequencing analysis.

For the validation of recursive splicing events, we amplified the same synthesized cDNA template using a pair of primers located upstream of the predicted donor and downstream of the predicted acceptor splice sites. For the event with one non-annotated splice site the forward primer was located in exon 16 and the reverse primer was located in intron 16. For the event with both non-annotated splice sites both primers were located in intron 16. PCRs were performed with 35 cycles using Phusion High-Fidelity PCR 2X Master Mix with HF Buffer (New England BioLabs Inc).

The amplification products were then separated on a 1.5% agarose gel after an RT-PCR reaction. The dissected bands were extracted and eluted (MinElute gel extraction kit, Qiagen) and sequenced with Sanger sequencing.

2.4 Results

2.4.1 Pipeline overview

SplicePie starts with a standard quality check procedure, removes low quality reads and performs split read alignment to the reference genome. The gene of interest is then extracted from the alignment file and is used to calculate the Splice Site Index (SSI) and classify reads based on their stage of splicing (pre-, intermediate- and post-splicing, see Section 2.3.4 and Table 2.1 for details). Reads from specific categories are used to predict and pinpoint putative non-sequentially or recursively splice introns. An overview of *SplicePie* is shown in Figure 2.2.

Alignment to the reference genome (hg19, GRCh37) is performed using GSNAP [107], an aligner that works with paired-end RNA-Seq data and can split each read end into multiple fragments, thereby coping effectively with a gene's intron/exon structure. We have chosen GSNAP over Tophat [62, 108], PASSion [36], HMMSplicer [109] and MapSplice [96] because, unlike these tools, GSNAP does not give priority to split alignments. This is crucial for pre-mRNA data, as such data contains reads across exon-intron boundaries. Unlike other tools, GSNAP uses the information about canonical and non-canonical splice sites when splitting the reads, which is important for the identification of novel exons. It is also able to split each read of the pair into as many fragments as necessary (in case of multiple adjacent small exons). GSNAP provides the results in the commonly used *sam* format. *SplicePie* generates *bam* and *wiggle* files from these *sam* files. Our analysis approach is mostly suitable for very detailed analysis, therefore it is recommended to analyze one gene of interest at a time.

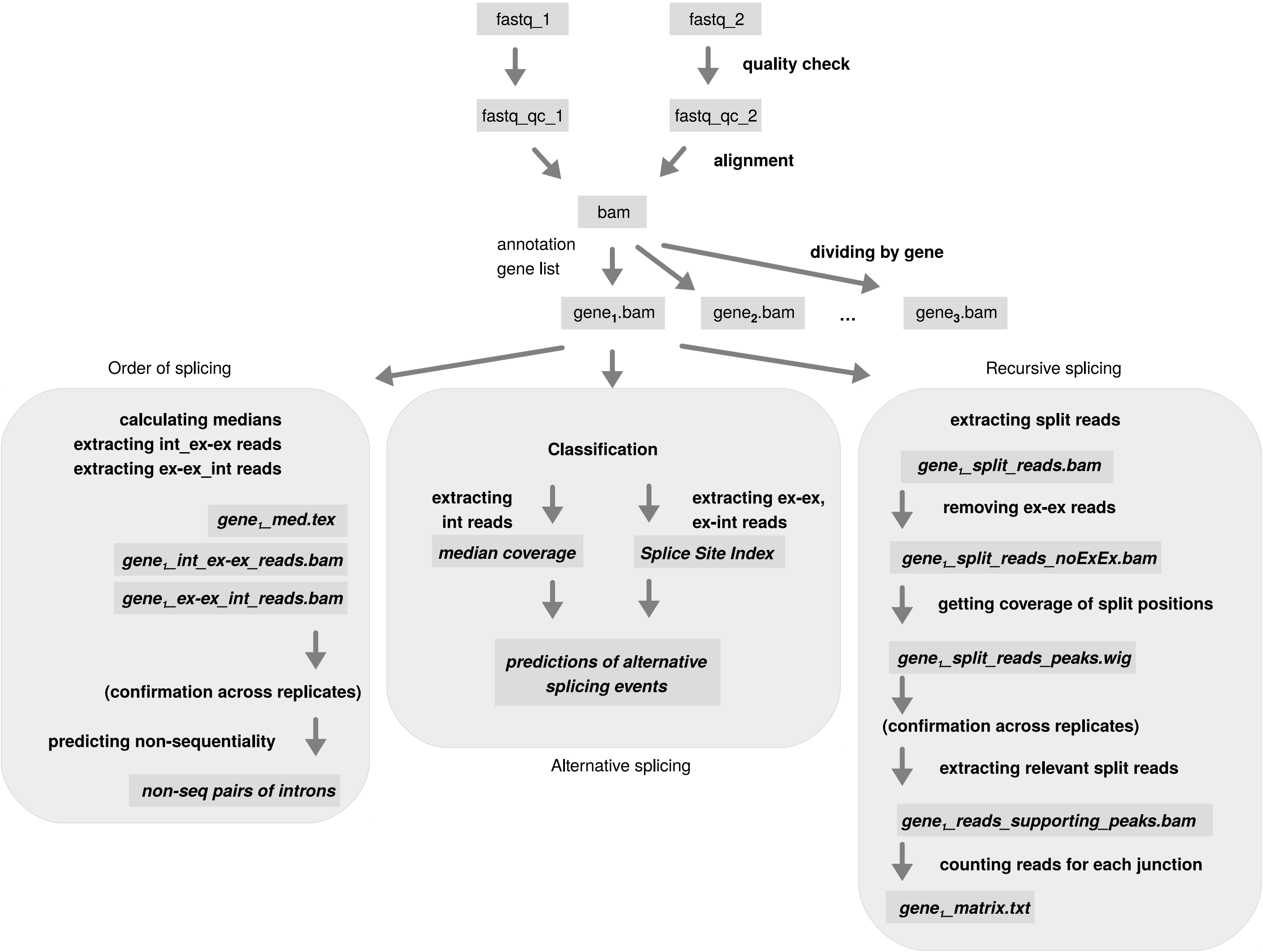


Figure 2.2: Lay-out of *SplicePie*. Light-grey boxes indicate the files required/produced during the mainstream analysis. Labels in **bold** next to the arrows indicate the steps of analysis. Labels in *italic* next to the arrows indicate the additional input files. Label *int_ex-ex* indicates that the file contains read pairs with one end mapped to the intron and the other end mapped to the exon-exon junction (and vice versa for *ex-ex_int*).

However, running *SplicePie* for multiple genes is also supported and all gene annotations provided by the user in a standard *GTF* format will be used to build a list of the genes of interest. *SplicePie* performs the classification of reads as pre-, intermediate- or post-splicing according to their mapping position (see Section 2.3.4 for details). After classification all reads spanning a specific exon-exon junction or an exon-intron boundary are used for the SSI calculation. SSIs will then be calculated for each splice site (see Section 2.3.7 for details). The output is provided as a *text* file containing SSIs for both 5' (SSI^5) and 3' (SSI^3) ends per intron. SSI value is calculated using the reads spanning exon-exon junctions and exon-intron boundaries. This value is similar to the completed splicing index, *coSI* [6], which reflects the amount of RNA molecules containing the exon with spliced adjacent introns. However, SSI is calculated separately per splice site and therefore reflects the difference between 5' and 3', thereby highlighting alternative splicing and/or incomplete splicing (while *coSI* reflects the completion of the splicing of a particular intron).

A parallel branch of *SplicePie* is analyzing the order of splicing and predicts which introns are spliced sequentially and which introns are spliced non-sequentially. Median coverage of each intron is calculated and when the difference between the downstream $intron_{i+1}$ and the upstream $intron_i$ is low, such pair of introns is potentially non-sequentialy spliced. At the moment we define this value as "low" when it is below *Q1*, where *Q1*, or lower quartile, is defined as the middle number between the smallest number and the median of the dataset (Figure 2.3, A).

Another branch of *SplicePie* processes the *wiggle* files and pinpoints introns potentially undergoing recursive splicing. The *wiggle* files contain only the coverage of splice sites. We take all combinations of donors and acceptors and calculate the number of reads supporting them. This creates a summary matrix containing all potential recursive splicing events and their frequency in each sample (Figure 2.4).

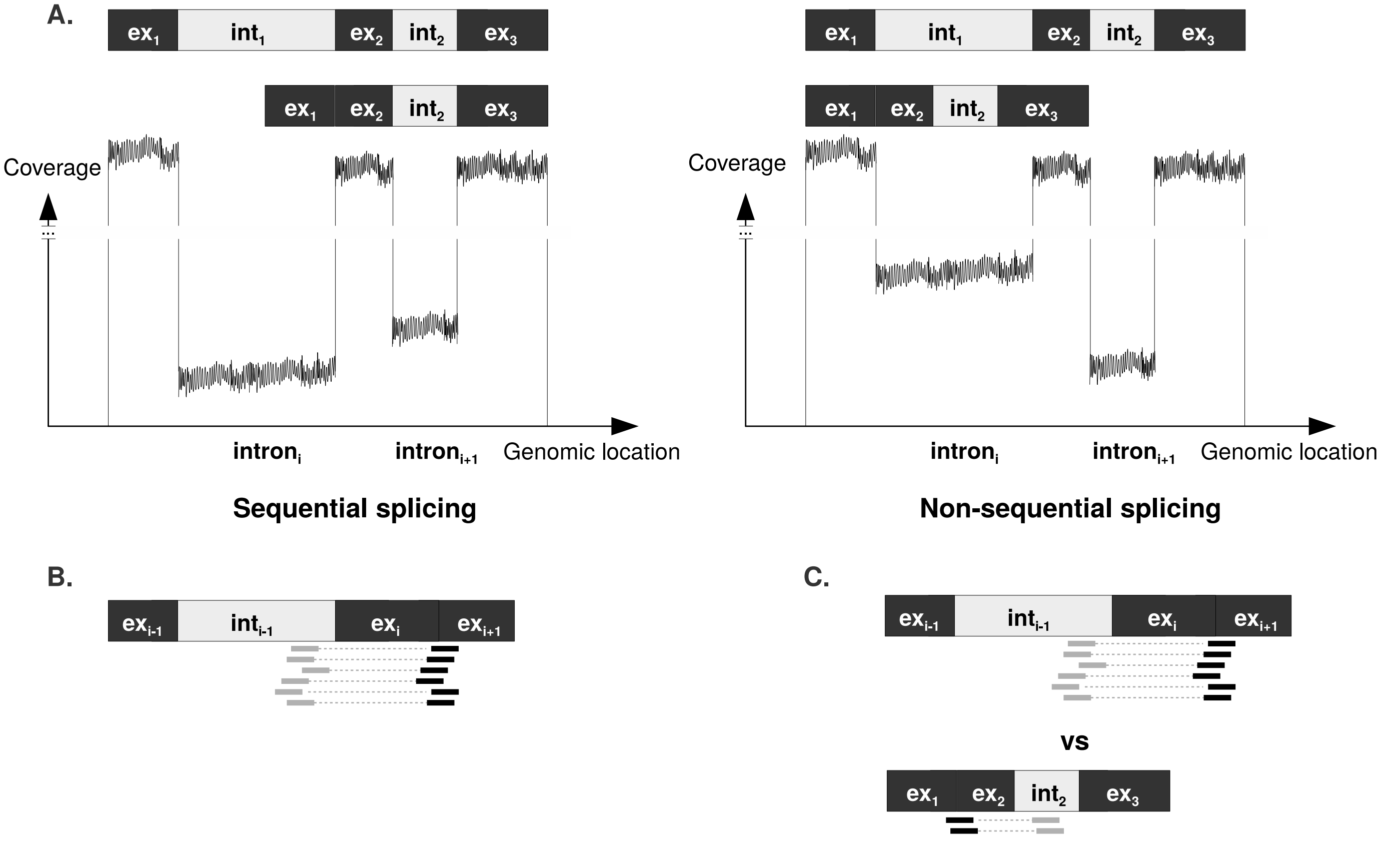
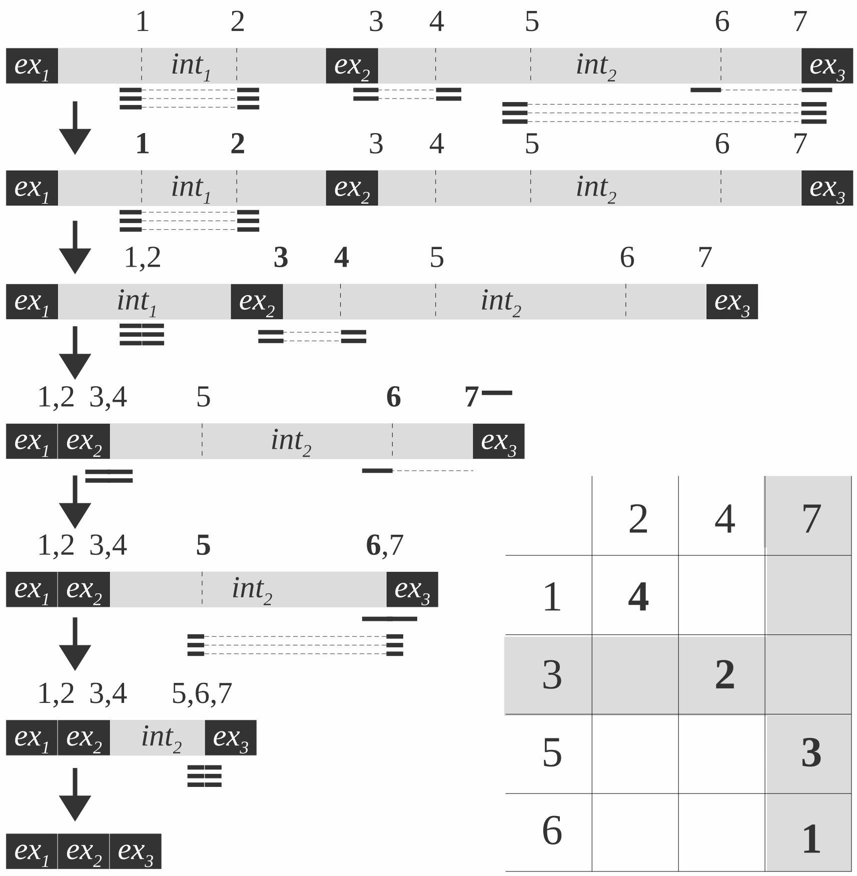


Figure 2.3: Principles of the two approaches to investigate non-sequential splicing. **A.** Coverage-based approach with the underlying assumption that the longer the intron is present in the sample, the higher the coverage will be. In case of non-sequential splicing the coverage of the downstream intron$_{i+1}$ is likely to be lower than the coverage of the upstream intron$_i$. Median of the coverage of each intron is used in this approach. **B.** The underlying assumptions for the read-based approach to detect non-sequential splicing: evidence for non-sequential splicing is obtained from read pairs with one end mapped to the upstream intron$_i$ and the other end mapped to the junction over the downstream intron$_{i+1}$ (exon$_{i+1}$-exon$_{i+2}$ junction). **C.** Method to select the introns with non-sequential splicing. The read pairs supporting the splicing intermediate where intron$_i$ is spliced before intron$_{i+1}$ should be less abundant than the read pairs for the intermediate product where intron$_{i+1}$ is spliced before intron$_i$.

2.4.2 Alternative splicing events in captured dataset

Four nuclear RNA, four total RNA and one DNA sample (Table S1) were sequenced (see Section 2.3.1 for details). The analysis of the target gene – *FXR1* – will be discussed in this section. From the targeted genes, we have studied the pre-mRNA splicing of (70,306 nt long) in detail because it is known to be alternatively spliced, has an intermediate number of exons (18), introns with variable length between 86 and 18,338 nt and a decent intronic coverage (on average above 1,500 for nuclear RNA samples) in the cell lines analyzed (Table S1). Note that the numbers of reads for the categories do not always add up to the total number of reads mapped to the target, since we do not show the number of reads defined as 'unclassified' (see Section 2.3). The classification of reads into three categories supporting pre-, intermediate- and post-splicing events is used to estimate the pre-mRNA content of samples (see Section 2.3.1 for details). Compared to the total RNA samples, all nuclear RNA samples contain a larger fraction of reads coming from the pre-splicing stage (Figure 2.5). This is expected, as total RNA is isolated from the whole cell and consists mostly of mature messenger RNA, while nuclear RNA contains (partly) spliced RNA. The post-splicing category contains a large fraction of reads even in nuclear RNA samples, which might be a consequence of the fast splicing of *FXR1*, which makes it hard to capture the nuclear pre-mRNA of this gene. Reads from the DNA sample were mainly classified as “pre-splicing”, which is expected, since DNA is not supposed to contain any exon-exon junctions.


1 2 3 4 5 6 7
ex_1 int_1 ex_2 int_2 ex_3
1 2 3 4 5 6 7
ex_1 int_1 ex_2 int_2 ex_3
1,2 3 4 5 6 7
ex_1 int_1 ex_2 int_2 ex_3
1,2 3,4 5 6 7
ex_1 ex_2 int_2 ex_3
1,2 3,4 5 6,7
ex_1 ex_2 int_2 ex_3
1,2 3,4 5,6,7
ex_1 ex_2 int_2 ex_3
ex_1 ex_2 ex_3

	2	4	7
1	**4**		
3		**2**	
5			**3**
6			**1**

Figure 2.4: Graphical representation of the analysis of recursive splicing. Black boxes represent exons and grey boxes represent introns, dashed lines across the introns and exon-intron borders represent positions of splits in reads. Numbers “1”, “2”, etc. on top of the gene schema represent the positions of the gene in genomic coordinates. Thick lines connected with dashed lines represent split reads (where the dashed part of the line represents the area across which the read is split). Step-by-step analysis of all split reads is shown and splicing intermediates corresponding to each group of split reads are shown on the schema in left part of the figure. The matrix in the right bottom corner contains donor splice sites (top row) and acceptor splice sites (left column). Each cell in the matrix represents the number of reads supporting such junction. Numbers in the gray cells of the matrix represent reads with one new non-canonical splice site. Numbers in white cells of the matrix represent reads with two new non-canonical splice sites (reads split within an intron).

In order to investigate alternative splicing events we use a combination of SSI values and medians of intronic and exonic coverage.

As the coverage may be influenced by the probe hybridization efficiency, we evaluated the uniformity of the coverage in the DNA sample D1 (Figure S3). The only dips in coverage take place in the Repeat Masker regions, which were excluded from probe design (Figure S4). To confirm the limited influence of probe hybridization on coverage, we calculated the standard deviation of the median coverage per intron, which was only 11.52% of the average.

Median coverage has been calculated per intron, while SSI is calculated per 5’ and 3’ end of each intron separately. SSI can indicate various alternative splicing events, i.e., (partial) intron retention and exon skipping (data shown for sample N2, Figure 2.6).

Decreases in the SSI^5 of intron 2 and in the SSI^3 of intron 1 are indicative of the skipping of exon 2 in a subset of transcripts. This is confirmed by the relatively low coverage of exon 2. The coverage of

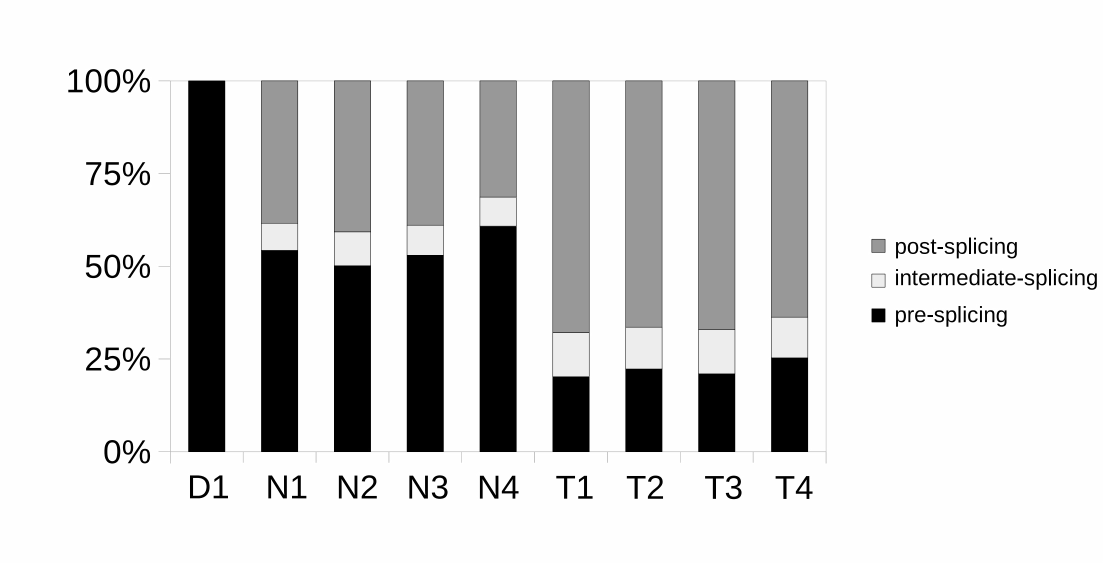


Figure 2.5: Classification of reads from the capture dataset mapped to *FXR1*. For sample identifiers see Table S1. The figure displays the percentage of reads mapped to *FXR1* classified into pre-, intermediate- and post-splicing fractions for pre-mRNA and total RNA samples in the captured dataset.

intron 13 is significantly higher than the average intronic coverage, which can indicate intron retention in both pre- and mature mRNA. This is also supported by low SSI^5 and SSI^3 values of intron 13, which means that the number of reads mapped to the boundary of exon 13 and intron 13 and the boundary of intron 13 and exon 14 are overrepresented. The retention of intron 13 and skipping of exon 2 were experimentally confirmed for sample N2 (Figure S5). Thus, low SSI^5 and SSI^3 on the same intron are the indication of an intron retention, whereas low SSI^3 of an intron in combination with low SSI^5 of the next intron is an indicator of exon skipping.

We also developed a module to rank introns based on their probability to be retained. For retained introns, both SSI^5 and SSI^3 values should be low. To estimate that, we calculate the magnitude of SSIs (how big the difference between SSI^5 and SSI^3 values is) and the likelihood of each magnitude (Appendix). The introns with the highest magnitude and a low p-value are the main candidates for retention (Table S2).

2.4.3 Non-sequential splicing in captured dataset

In the captured dataset, we searched for non-sequentially spliced introns in *FXR1*. Multiple candidate pairs of introns with a difference in median coverage below *Q3* were found: introns 1 and 2; introns 3 and 4; introns 10 and 11; introns 16 and 17. (Here Q3 stands for the third quartile – the middle value between the median and the highest value of the data.) To confirm this, we calculated the ratio of read pairs in support over those not in support of non-sequential splicing. If the splice-ratio is close to *1*, the splicing is most likely sequential, the lower the ratio is, the more non-sequential splicing occurs. We observed a very strong correlation between both methods, supporting the idea that these methods are suitable to identify non-sequentiality spliced introns (Pearson $R^2 \leq 0.86$ and Spearman $R^2 \leq 0.85$ (Figure S7).

We used DNA as a negative control, since DNA does not undergo splicing and the coverage of the introns should not differ significantly. If introns have significantly different coverage on DNA level, this may be due to sequencability, mappability bias or other technical artifacts (not a biological reason). Two pairs of introns predicted to undergo non-sequential splicing in RNA samples (introns 3 and 4; introns 10 and 11) survived this extra control step. They were not classified as non-sequentially spliced in DNA. Introns 10 and 11 have been selected for further the experimental validation and showed to be non-sequentially spliced (Figure 2.7).

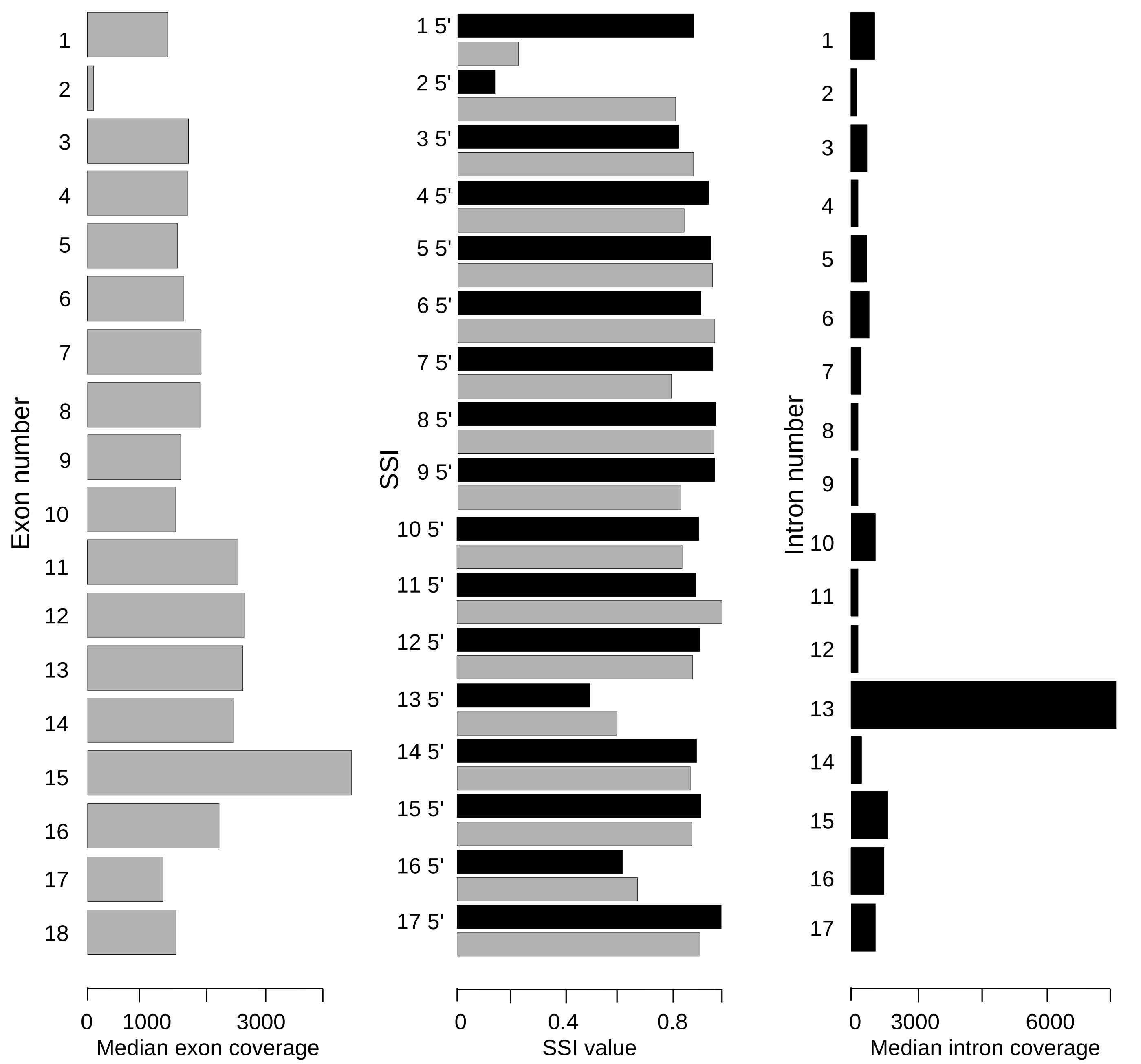


Figure 2.6: Splice Site Index (SSI) and medians of coverage of exons and introns in *FXR1*. Gray bars in the left panel represent the coverage of exons (exon 1 on top). Black bars in the middle panel represent SSI^5 values the introns and gray bars on the middle plot represent SSI^3 values of the introns (intron 1 on top). Black bars in the right plot represent the coverage of introns (intron 1 on top). Data shown for nuclear RNA sample N2.

2.4.4 Recursive splicing in captured dataset

We searched for the evidence of multi-step splicing in our captured dataset (as described in Section 2.3.6). *Wiggle* files containing coverage at positions where the reads were split have been created for four pre-mRNA samples (reads spanning exon-exon junctions were excluded). We considered positions present in all samples with the same sign for further analysis, because they were deemed most reliable.

We assessed the distribution of peak coverage and tried to identify the minimal coverage for the peak to be included in the final list. We first investigated the highest peaks and found out that the coverage of the regions between the peaks is higher compared to the rest of the introns. These observations let us hypothesize that such intronic regions with high coverage and split reads mapped to them and to the

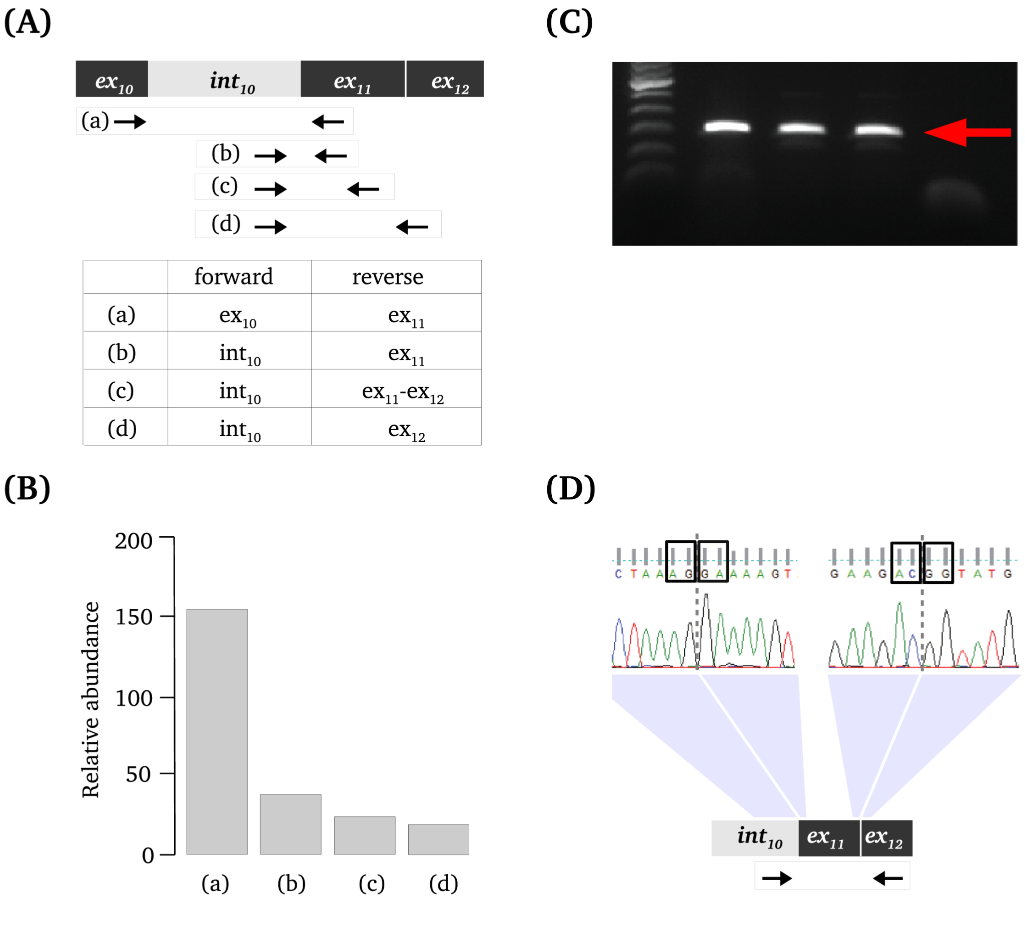


Figure 2.7: Experimental validation of predicted non-sequential splicing of introns 10 and 11 in *FXR1*. **A.** The design of the primers for the validation of non-sequential splicing of intron 10. **B.** The results of quantitative real-time PCR showing the relative abundance (on the Y axis) of splicing intermediates with primer combinations described in **A** (on the X axis). Since three tested cell lines showed similar reproducible results, the figure shows the average. **C.** Results of PCR showing the presence of the fragment of anticipated size. Lane 1 contains the marker, lanes 2, 3 and 4 represent three cell lines, lane 5 shows the negative PCR control. **D.** The results of Sanger sequencing of the band shown on **C**. The top panel shows the output of Sanger sequencing, black box around “AG” depicts the acceptor splice site, the boxes around “GA”, “AC” and “GG” depict ends of the exons. The bottom panel shows the design of the primers for the sequenced amplicon.

adjacent exons might be novel exons. We were able to experimentally validate one of potential novel exons (Figure S6). Our findings thus suggest that the method developed for the detection of recursive splicing events is also suitable for finding novel exons.

After excluding the peaks with the highest coverage we found eight events to be recurrent and consistent across all RNA samples (Table S3). These events were not present in the DNA sample D1, which had only 61 split reads (against hundreds of split reads in RNA samples), none of the split positions was supported by more than one read. The vast majority of split reads in DNA were mapped with a large number of mismatches (unlike split reads in RNA samples).

The most abundant events of recursive splicing occurred on the 5’ end of the intron (when the donor splice site is canonical and the acceptor splice site is situated within the downstream intron, Figure S8). The second biggest class (three cases, or around 40%) were the cases of recursive splicing occurring on the 3’ end of the intron. We also found a pair of events with shared canonical acceptor splice site (the intron 10-exon 11 boundary), which suggests the possibility of multi-step recursive splicing in this area (Figure S9). Furthermore, we were also able to identify one recurrent event of inner recursive splicing with two non-canonical splice sites (Figure S10).

The strength of the newly identified donor or acceptor splice sites associated with potential recursive splicing events was assessed with Human Splicing Finder[5]. All sites were identified as highly probable putative splice sites.

To obtain further evidence that the recursive splicing events identified by *SplicePie* are genuine, we evaluated the sequence motifs flanking the splits. We selected events present in two, three, four or five out of five samples and extracted the splice sites and two nucleotides upstream of the donor and

[5] http://www.umd.be/HSF3/HSF.html

downstream of the acceptor sites. We calculated the percentage of canonical and non-canonical donor and acceptor splice sites in the events with non-annotated donor or acceptor splice site. In this analysis we omit potential recursive splicing events with both non-annotated splice sites, as they are more likely to be novel exons. Such events require further experimental validation, and only experimentally showing there presence in pre- and possibly mature mRNA will distinguish between recursive splicing and novel exons.

Our results (Table S4) show strong enrichment for canonical donor (GT) and acceptor (AG) splice sites for events with one non-annotated splice site present in five samples. The enrichment becomes weaker for the events detected in only a subset of the samples, especially for the events with a non-annotated acceptor site. This indicates that the recurrent events with one non-annotated splice site are the most robust recursive splicing events. Note that the aligner we use does not have any preferences for splice motifs when splitting reads.

Apart from the *in silico* validation, we also used RT-PCR followed by Sanger sequencing to validate potential recursive splicing events. We designed primers flanking two events found in all five samples and analyzed nuclear and cytoplasmic RNA isolated from two of the muscle cell lines. For the first event, (chr3:180,689,975-180,692,201, Table S3), detected a product of the expected length and sequence. However, it was found in both nuclear and cytoplasmic RNA, suggesting that the detected event represents a novel exon and not an intermediate splicing product. The second set of primers captured another recursive splicing event that was not present in the RNA-Seq data.

2.4.5 Performance on non-targeted dataset

To demonstrate the performance of *SplicePie* on regular, non-targeted RNA-Seq data, three samples from the Gencode project (the ENCODE dataset) containing RNA from different nuclear compartments (chromatin- associated and nucleolus RNA) were analyzed following the same procedure as the captured dataset. Around 95% of reads mapping to the *FXR1* gene were classified as pre-splicing for chromatin and nucleolus RNA in contrast with 20 and 60% for the nuclear and total RNA samples, respectively. This is in line with the presence of pre-mRNA in chromatin and nucleolus, while mature mRNA is prevailing in the nucleoplasm.

The values of SSI and medians of exonic and intronic coverage for the non-targeted RNA dataset (Figure S11) coincide with corresponding values calculated for the captured dataset. Low SSI^3 for intron 1 and low SSI^5 for intron 2 indicate the skipping of exon 2 and high coverage of intron 13 together with its low SSI^5 and SSI^3 are indicative of intron 13 retention. However, the pattern of the medians is less consistent for both exonic and intronic coverage. The C1 sample contains chromatin-associated RNA, for which splicing is not known to be in action, the difference between the exonic and intronic coverage is small, therefore the coverage values for chromatin-associated RNA are less informative than those for nucleoplasmic RNA.

We were able to confirm the previous findings (of intron 10 and intron 11 being non-sequentially spliced) using the approach based on median intron coverage (see Section 2.3.5 for details). Other predictions were mostly confirmed in at least two out of three analyzed samples with the coverage-based approach (data not shown). However, the number of reads needed to calculate the splice-ratio was not high enough, so the majority of the ratios equalled zero.

The non-targeted RNA dataset was also analyzed in order to find potential recursive splicing events. Five peaks were present in all three samples. Two of these peaks were found in the list of donor/acceptor splice sites identified for the captured dataset. The number of splice sites identified per sample was approximately 100, while the number of peaks for each sample of the captured dataset was over 450.

In order to demonstrate the performance of *SplicePie* on the whole-transcriptome dataset, we selected *TIA1* for the analysis of non-sequential and recursive splicing.

Based on the difference in median intronic coverage (which was at least three times higher than the upper quartile (75%) in all three samples) and high splice-ratio (average of 0.7 in the three samples), we could predict two introns of *TIA1* to be non-sequentially spliced. Neither intron 2, nor intron 3 are overlapping with any genomic elements that might influence the coverage, such as pseudogenes or non-coding RNAs.

According to both considerable difference in coverage and high splice-ratio, intron 3 is predicted to be spliced before intron 2 (Figure S12). Even more cases of potential non-sequential splicing were found in two out of three samples, however, to claim that the order of splicing of these introns is non-sequential, experimental follow-up is required.

We also investigated multi-step splicing in *TIA1* and were able to detect a number of potential recursive splicing events. Events present in all three samples C1, C2 and U1 with non-annotated donor, non-annotated acceptor and both non-annotated splice sites were detected (Table S5).

Therefore, non-targeted total RNA-Seq data provides sufficient information to analyze recursive splicing for some genes and can be used for the detailed investigation of splicing in action.

2.5 Discussion

Exploring pre-mRNA processing is facilitated by new sequencing technologies reaching higher throughput, hence producing more data. The analysis of both pre- and mature mRNA provides new insights into splicing mechanisms and alternative splicing events. However, current software is focusing on mature mRNA and the identification (and quantification) of transcript variants.

The presented pipeline for the pre-mRNA data analysis, *SplicePie*, offers a number of approaches and solutions to study splicing in more details. The proposed strategy performs well on different sample preparations (sequencing the whole pool of RNAs, or capturing a gene with different relative quantities of pre- and mature mRNA). Our method can detect various genuine alternative splicing events like intron retention, exon skipping and novel exons. Furthermore, it is capable to resolve the order of splicing and recursive splicing events.

The methodology of *SplicePie* significantly differs from existing pipelines, such as Cufflinks, Scripture [110] or MISO [111]. These tools focus their analysis on the end result of splicing and mainly use reads mapped to the exons or exon-exon junctions. Reads mapped to the introns are either treated as putative exons or not addressed at all (in case of annotation-based pipelines). Therefore these pipelines are not able to analyze mixtures of pre- and mature mRNA (as found in nuclear RNA extracts). This is crucial to understand the details of the splicing mechanism. Our pipeline *SplicePie* is specifically geared towards the analysis of the full splicing process in action. In order to do so, it uses all reads mapped to exons, introns, exon-exon junctions or exon-introns boundaries.

For the analysis of alternative splicing events, the SSI and the medians of exonic or intronic coverage methods implemented in *SplicePie* are mutually reinforcing. In case of captured data enriched with partly spliced nuclear RNA, the difference in exonic and intronic coverage makes the patterns of coverage informative for assessing alternative splicing events. In case of “pure” nuclear RNA with lower abundance of spliced fragments, the difference in exonic vs. intronic coverage drops, however, the SSI values become more informative.

The main novelty of the methodology introduced in this paper is the possibility to analyze splicing order and the stepwise nature of splicing. While we are able to judge local splicing order, i.e. one intron is spliced before a neighboring intron, it is not possible to determine the global order of splicing for the entire transcript. This happens due to the co-transcriptional nature of the splicing process and the fact that we capture only one snapshot of the nascent transcript. We show that the local order of splicing for certain introns within *FXR1* is reproducible (in biological replicates) and even consistent across multiple cell lines. Furthermore, this can be confirmed with independent PCR-based technologies.

Although the fact that splicing can be performed in multiple steps has been known for over a decade [112], however, it has never been analyzed bioinformatically. We show that analyzing the intermediate category of reads is possible for both potential novel exons and recursive splicing events. However, focusing on a narrow fraction of reads might result in analyzing random events, which is why we suggest to use as many samples as possible. To improve reliability even further it is advised to select events occurring in a significant number of samples. This strategy also helps reducing biases introduced by PCR duplicates, which are not likely to appear at the same position in replicates. The splice motif analysis of potential recursive splicing events provides evidence that these events are genuine, considering the canonical splice motifs and the occurrence of the events across multiple cell lines. However, as RNA-Seq is more sensitive than PCR, not all detected events can be experimentally confirmed at the moment. Recursive splicing events with both non-canonical splice sites should be treated especially carefully, as for these events it is hard to distinguish between recursive splicing and novel exons without the experimental validation in both nuclear and cytoplasmic RNA.

Detecting background noise is a common problem in bioinformatics and statistics, especially when working with large datasets containing a mix of introns and exons. Our approach was shown to perform well on both high (captured dataset) and low (non-targeted ENCODE dataset) coverage data. Moreover, despite the combination of low coverage and noise, alternative splicing events were detected reliably. This indicates that total RNA sequencing can be used for detecting non-sequential splicing events relying mainly on coverage information. *SplicePie* can be run on any dataset and the main concern is the average coverage of the introns. Even recent total RNA sequencing protocols do not provide enough intronic coverage to perform the analyses as powerful and reliable as the analyses on captured data. Our study shows that even total RNA sequencing of specifically nucleus does not generate enough coverage to detect non-sequential and multi-step splicing as efficiently as captured RNA libraries. Due to low intronic coverage total RNA can be used to analyze splicing and certain events will be detected, however, a lot of events will be missed. Therefore, we would still recommend to do the targeted sequencing of the genes of interest to allow a more in depth analysis.

Using *SplicePie* on different datasets revealed various not yet annotated splicing events. Our work enhances the value of pre-mRNA sequencing data and pioneers the investigation of the mechanisms of (alternative) splicing.

2.6 Acknowledgements

We thank M. van Iterson and S.P. van Leeuwen for their help with statistical analysis, M. Vermaat for his help with software development, M.P. Villerius for the help with running the analysis of the cluster, M. van Galen for his extensive help with this manuscript and W. Arindrarto for suggesting the name of the pipeline. This work was funded by ZonMW (project code NWO VIDI 91710303 to AAR) and by the European Commission 7th Framework Program, Project N. 261123 (GEUVADIS).

2.7 Appendix

Table S1: Characteristics of samples and summary of results for reads mapped to *FXR1*. For each sample the reads were classified as pre-, intermediate- or post-splicing based on the distance between paired ends (<650 being normal and >650 being larger than anticipated). Samples D1-T4 are from the captured dataset, samples C1-U1 from the ENCODE dataset.

sample name	description	reads mapped on target	norm. pre-spl.	norm. int.-spl.	norm. post-spl.	large int.-spl.	large post-spl.
D1	DNA	306103	286501	148	29	209	6
N1	**n**uclear RNA1	317827	139084	11834	64750	6879	33688
N2	**n**uclear RNA2	586827	224231	28520	134229	12336	47954
N3	**n**uclear RNA3	431321	184823	15845	91217	12456	44680
N4	**n**uclear RNA4	1394314	707109	51856	222918	39003	141707
T1	**t**otal RNA1	242979	41232	11347	81826	9456	41048
T2	**t**otal RNA2	181383	28002	9234	62588	7222	31376
T3	**t**otal RNA3	261004	42412	10932	67126	13111	68536
T4	**t**otal RNA4	317123	60915	12528	81132	13901	72375
C1	**c**hromatin RNA1	21632	19445	169	489	77	209
C2	**c**hromatin RNA3	8600	7251	62	189	39	91
U1	n**u**cleolic RNA	20283	16566	173	152	123	158

Table S2: Predicting intron retention events based on the magnitude of the SSIs and the *p*-value. Column "5′ ex-int" contains number of reads mapped to the exon-intron boundary on the 5'-end of an intron (used to calculate SSI^5). Column "3′ ex-int" contains number of reads mapped to the exon-intron boundary on the 3'-end of an intron (used to calculate SSI^3).

intron	5′ ex-int	3′ ex-int	ex-ex	magnitude	*p*-value
1	2463	261	3768	0.03	0
2	679	1465	2100	0.14	6.08e-66
3	1202	1452	6821	0.08	1.32e-06
4	712	1571	10189	0.03	7.54e-74
5	769	1406	9705	0.04	6.09e-43
6	2125	940	17086	0.03	4.19e-104
7	324	1667	5493	0.03	3.13e-217
8	574	409	7064	0.03	1.58e-07
9	1034	1894	12403	0.04	1.46e-57
10	2032	3088	20646	0.05	1.54e-49
11	1840	82	12225	0.01	0
12	880	2660	14853	0.03	1.45e-205
13	**7783**	**9550**	**7525**	**0.34**	**4.29e-41**
14	1935	3046	18715	0.05	3.44e-56
15	8137	4099	30855	0.06	4.23e-297
16	4602	3264	11903	0.12	1.36e-51
17	786	2272	14427	0.03	1.15e-165

Table S3: Representation of recursive splicing events (column “coordinates”) in the captured dataset. All detected events are located on chromosome 3.

Coordinates	Splice site	N1	N2	N3	N4
180,653,019-180,665,633	acceptor	1	1	4	1
180,674,213-180,675,607	donor	2	2	4	2
180,674,835-180,675,607	donor	5	4	2	1
180,680,878-180,681,592	acceptor	5	1	4	2
180,686,042-180,687,934	acceptor	20	73	42	23
180,688,146-180,688,665	acceptor	8	401	85	29
180,689,975-180,692,201	both	5	9	10	3
180,692,935-180,693,101	donor	13	37	26	18

Table S4: Canonical and non-canonical splice sites in potential recursive splicing events.

Number of samples containing an event	Non-annotated acceptor		Non-annotated donor	
	number of events	AGxx	number of events	xxGT
Five out of five	4	75%	3	100%
Four out of five	36	86%	28	100%
Three out of five	53	90%	49	100%
Two out of five	78	85%	95	95%

Table S5: Representation of recursive splicing in *TIA1* detected in the non-targeted dataset. Column “coordinates” contains the coordinates of recursive splicing events in on the reference genome. Column “splice site” indicates which splice site is non-annotated. Columns “C1”, “C2” and “U1” contain the number of reads supporting each recursive splicing event in each RNA sample from the non-targeted dataset.

Coordinates	Splice site	C1	C2	U1
70,443,631-70,443,885	donor	15	3	61
70,451,761-70,452,460	donor	16	5	7
70,451,761-70,452,597	donor	1	1	1
70,452,525-70,454,867	acceptor	15	3	17
70,454,954-70,455,476	donor	23	15	16
70,455,594-70,456,191	acceptor	20	20	13
70,457,986-70,460,773	donor	4	3	6
70,460,894-70,463,211	acceptor	1	1	2
70,463,307-70,465,921	donor	5	2	2
70,469,796-70,469,830	both	5	3	2

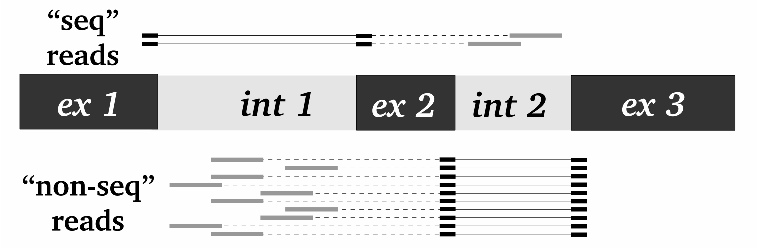


Figure S1: Read pairs supporting sequential ("seq") or non-sequential ("non-seq") splicing. Thick black lines represent ends that were split over a junction (and the thin black line connects the pieces from one end of a read pair). Thick gray lines represent ends mapped to the introns. Dashed line connects two ends of one read pair. In this example, number of "non-seq" read pairs equals 10 and the number of "seq" reads equals 2.

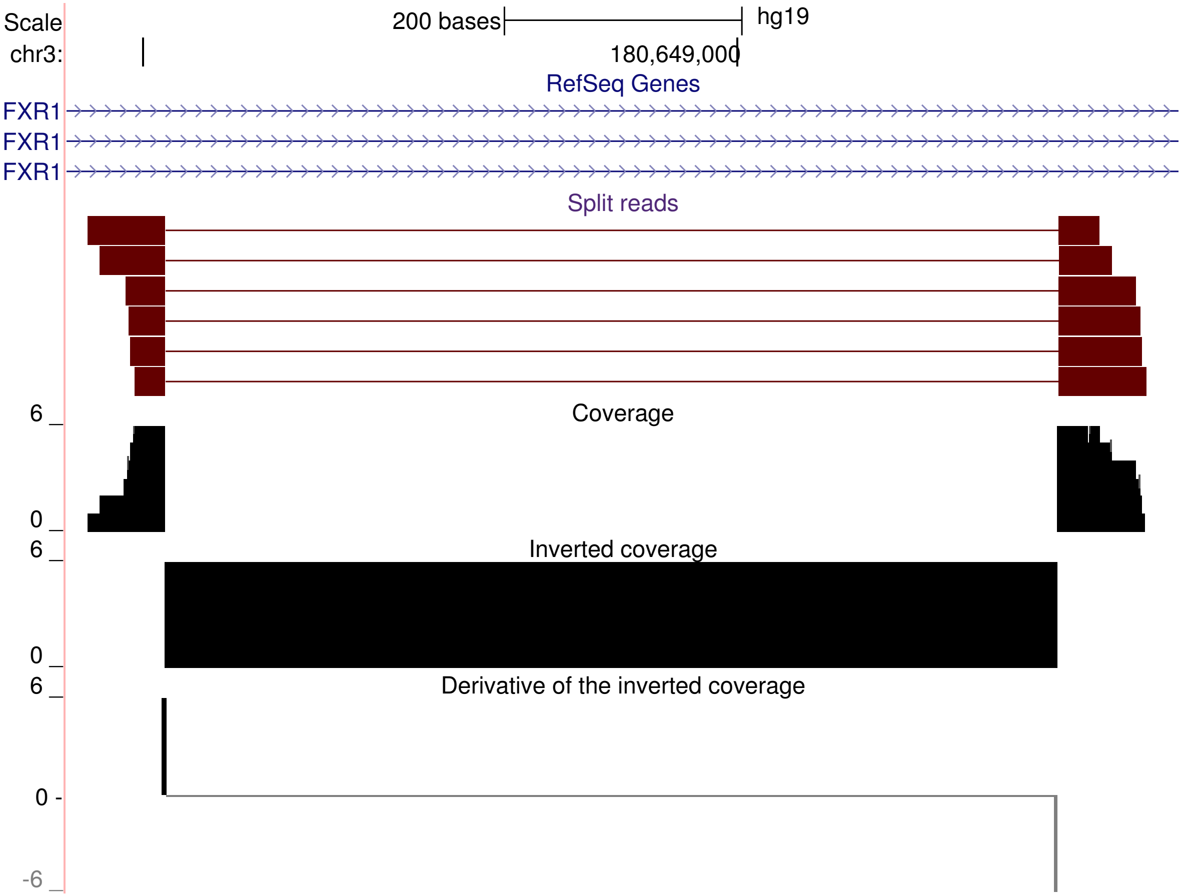


Figure S2: Schematic representation of the recursive splicing analysis. A split read (not mapped to an exon-exon junction) is used to calculate the inverted coverage. The derivative of the inverted coverage is then calculated, producing peaks at the positions where the split starts and drops at the positions where split ends. The size of peaks and drops equals the amount of reads split at this position.

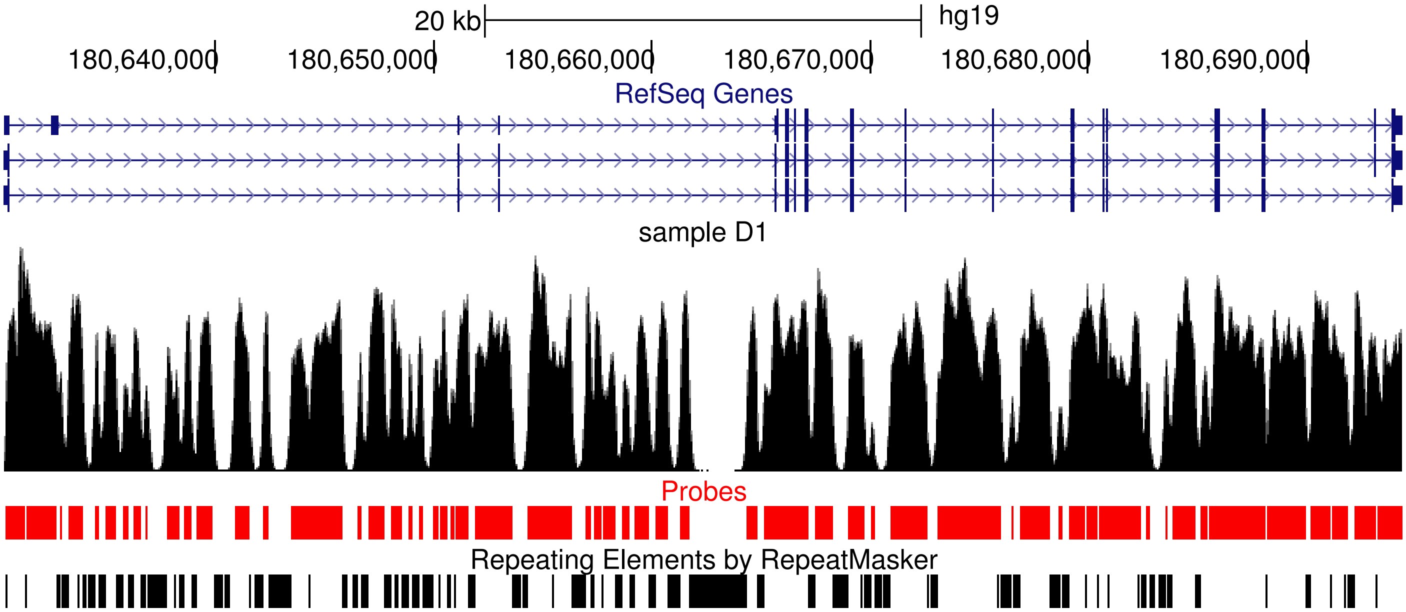


Figure S3: Overview of the coverage across the whole gene being evenly distributed across exons and introns. Coverage of exons and introns in the DNA sample and its correlation with the probes and Repeat Masker regions. Top panel ("RefSeq Genes") indicates the NCBI annotation of the *FXR1* gene used for the analysis, thick blocks depicting exons and thin lines with arrows depicting introns. Second panel ("sample D1") shows the coverage of the DNA sample from the captured dataset (y-axis reflects the coverage, maximum coverage being over 2000). Third panel in red ("Probes") reflects the areas that have been covered by probes (blank areas depict the regions where no probes have been designed). The bottom panel ("Repeating Elements by Repeat Masker") indicate the Repeat Masker track provided by UCSC that has been used to design the probes (black areas depict repetitive elements that were not included in the probes).

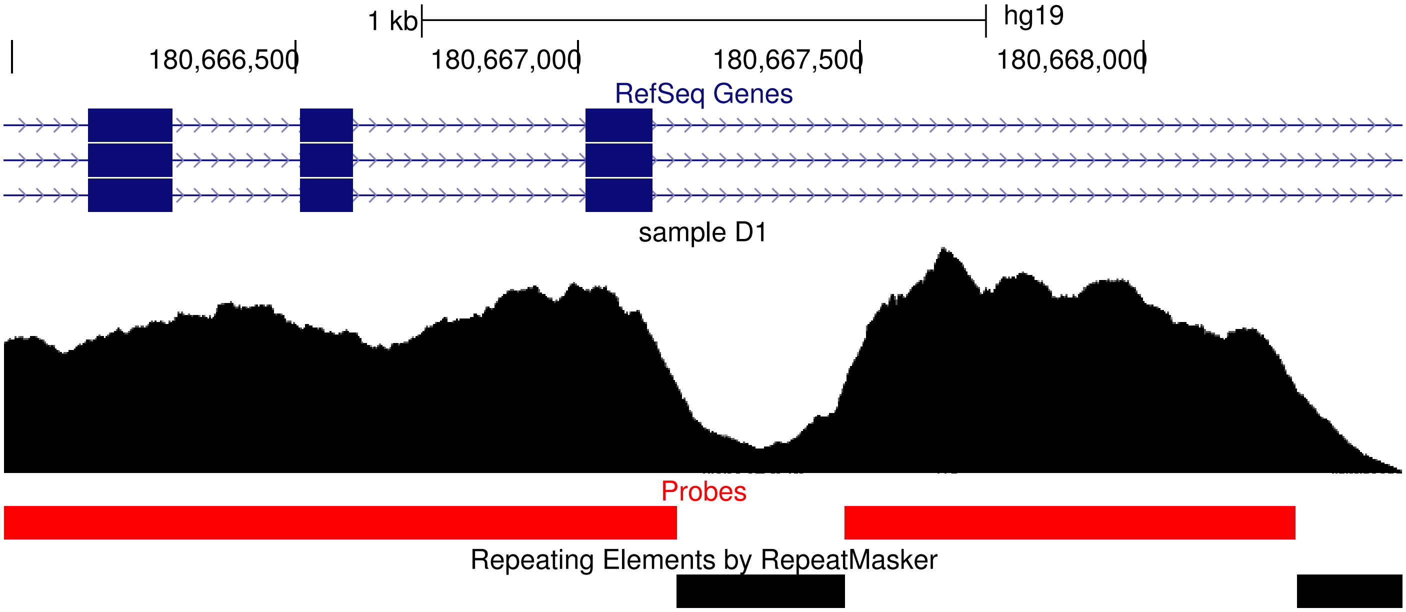


Figure S4: Zoomed-in overview of the coverage, showing no difference between the coverage distribution across an exon and an intron.

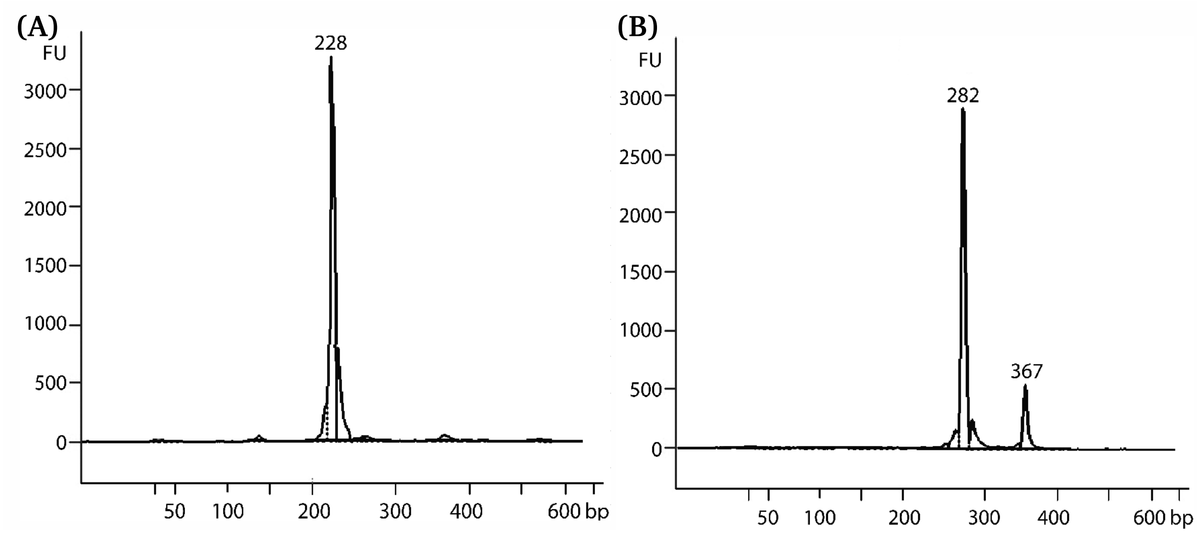


Figure S5: The results of PCR amplification experiments proving a skip of exon 2 and a retention of intron 13 in *FXR1*, as predicted *in silico* by the pipeline. **A.** PCR primers were designed to anneal to exon 1 and exon 5, and this fragment was amplified. The highest peak indicates a fragment of exon 1-exon 5 without exon 2 (228 bp in length). The abundance of transcripts containing exon 2 is very low and the fragment containing exon 2 (571 bp) is not visible. **B.** PCR primers were designed to anneal to exon 12 and exon 15, and the targeted fragments were amplified. The lower peak indicates a fragment with intron 13 inclusion (367 bp in length). The higher peak indicates a fragment without intron 13 (282 bp in length).

Calculating magnitude and likelihood for each intron of a gene in order to estimate its probability to be retained. *"M"* stands for "magnitude" and *"p-value"* stands for the *p*-value of the binomial test for likelihood:

$$M = \frac{min(ex\text{-}int_1,\ ex\text{-}int_2)}{min(ex\text{-}int_1,\ ex\text{-}int_2) + ex\text{-}ex} \tag{2.3}$$

$$p\text{–}value = p\text{-}value_{binom}(ex\text{-}int_1,\ ex\text{-}int_2,\ 0.5) < 0.05 \tag{2.4}$$

```
   1  ATGGCGGAGC TGACGGTGGA GGTTCGCGGC TCTAACGGGG CTTTCTACAA  50
  51  GGGATTTATC AAAGATGTTC ATGAAGACTC CCTTACAGTT GTTTTTGAAA  100
 101  ATAATTGGCA ACCAGAACGC CAGGTTCCAT TTAATGAAGT TAGATTACCA  150
 151  CCACCACCTG ATATAAAAAA AGAAATTAGT GAAGGAGATG AAGTAGAGGT  200
 201  ATATTCAAGA GCAAATGACC AAGAGCCATG TGGGTGGTGG TTGGCTAAAG  250
 251  TTCGGATGAT GAAAGGAGAA TTTTATGTCA TTGAATATGC TGCTTGTGAC  300
 301  GCTACTTACA ATGAAATAGT CACATTTGAA CGACTTCGGC CTGTCAATCA  350
 351  AAATAAAACT GTCAAAAAAA ATACCTTCTT TAAATGCACA GTGGATGTTC  400
 401  CTGAGGATTT GAGAGAGGCG TGTGCTAATG AAAATGCACA TAAAGATTTT  450
 451  AAGAAAGCAG TAGGAGCATG CAGAATTTTT TACCATCCAG AAACAACACA  500
 501  GCTAATGATA CTGTCTGCCA GTGAAGCAAC TGTGAAGAGA GTAAACATCT  550
 551  TAAGTGACAT GCATTTGCGA AGTATTCGTA CGAAGTTGAT GCTTATGTCC  600
 601  AGAAATGAAG AGGCCACTAA GCATTTAGAA TGCACAAAAC AACTTGCAGC  650
 651  AGCTTTTCAT GAGGAATTTG TTGTGAGAGA AGATTTAATG GGCCTGGCAA  700
 701  TAGGAACACA TGGTAGTAAC ATCCAGCAAG CTAGGAAGGT TCCTGGAGTT  750
 751  ACCGCCATTG AGCTAGATGA AGATACTGGA ACATTCAGAA TCTACGGAGA  800
 801  GAGTGCTGAT GCTGTAAAAA AGGCTAGAGG TTTCTTGGAA TTTGTGGAGG  850
 851  ATTTTATTCA GGTTCCTAGG AATCTCGTTG GAAAAGTAAT TGGAAAAAAT  900
 901  GGCAAAGTTA TTCAAGAAAT AGTGGACAAA TCTGGTGTGG TTCGAGTGAG  950
 951  AATTGAAGGG GACAATGAAA ATAAATTACC CAGAGAAGAC GGTATGGTTC  1000
1001  CATTTGTATT TGTTGGCACT AAAGAAAGCA TTGGAAATGT GCAGGTTCTT  1050
1051  CTAGAGTATC ATATTGCCTA TCTAAAGGAA GTAGAACAGC TAAGAATGGA  1100
1101  ACGCCTACAG ATTGATGAAC AGCTGCGACA GATTGGTTCT AGGTCTTATA  1150
1151  GCGGAAGAGG CAGAGGTCGT CGGGGACCTA ATTACACCTC CGGTTATGGT  1200
1201  ACAAATTCTG AGCTGTCTAA CCCCTCTGAA ACGGAATCTG AGCGTAAAGA  1250
1251  CGAGCTGAGT GATTGGTCAT TGGCAGGAGA AGATGATCGA GACAGCCGAC  1300
1301  ATCAGCGTGA CAGCAGGAGA CGCCCAGGAG GAAGAGGCAG AAGTGTTTCA  1350
1351  GGGGGTCGAG GTCGTGGTGG ACCACGTGGT GGCAAATCCT CCATCAGTTC  1400
1401  TGTGCTCAAA GATCCAGACA GCAATCCATA CAGCTTACTT GATAATACAG  1450
1451  AATCAGATCA GACTGCAGAC ACTGATGCCA GCGAATCTCA TCACAGTACT  1500
1501  AACCGTCGTA GGCGGTCTCG TAGACGAAGG ACTGATGAAG ATGCTGTTCT  1550
1551  GATGGATGGA ATGACTGAAT CTGATACAGC TTCAGTTAAT GAAAATGGGC  1600
1601  TAGATGATAG TGAAAAAAAA CCCCAGCGAC GCAATCGTAG CCGCAGGCGT  1650
1651  CGCTTCAGGG GTCAGGCAGA AGATAGACAG CCAGTCACAG TTGCAGATTA  1700
1701  TATTTCTAGA GCTGAGTCTC AGAGCAGACA AAGAAACCTC CCAAGGGAAA  1750
1751  CTTTGGCTAA AAACAAGAAA GAAATGGCAA AAGATGTGAT TGAAGAGCAT  1800
1801  GGTCCTTCAG AAAAGGCAAT AAACGGCCCA ACTAGTGCTT CTGGCGATGA  1850
1851  CATTTCTAAG CTACAGCGTA CTCCAGGAGA AGAAAAGATT AATACCTTAA  1900
1901  AAGAAGAAAA CACTCAAGAA GCAGCAGTCC TGAATGGTGT TTCATAA     1947
```

Figure S6: Novel exon (in red) predicted by the pipeline and its location in the full-length *FXR1* transcript. The exon is located between exons 16 and 17 (according to the annotation used in this paper). The full-length transcript has been experimentally validated by Sanger sequencing.

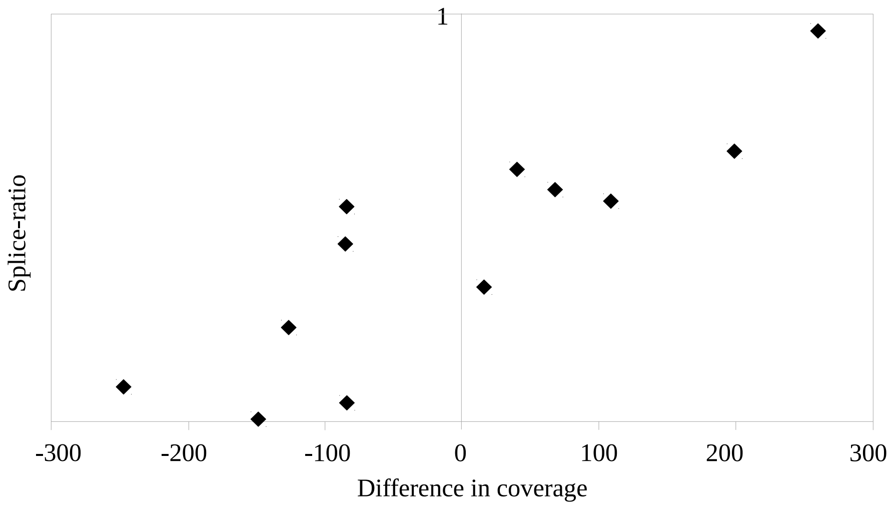


Figure S7: Linear correlation between the difference in median coverage ($\text{intron}_{i+1} - \text{intron}_i$) and the splice ratio. Correlation shown for the pre-mRNA N1 sample. Pearson correlation: -0.86. Spearman correlation: -0.84.

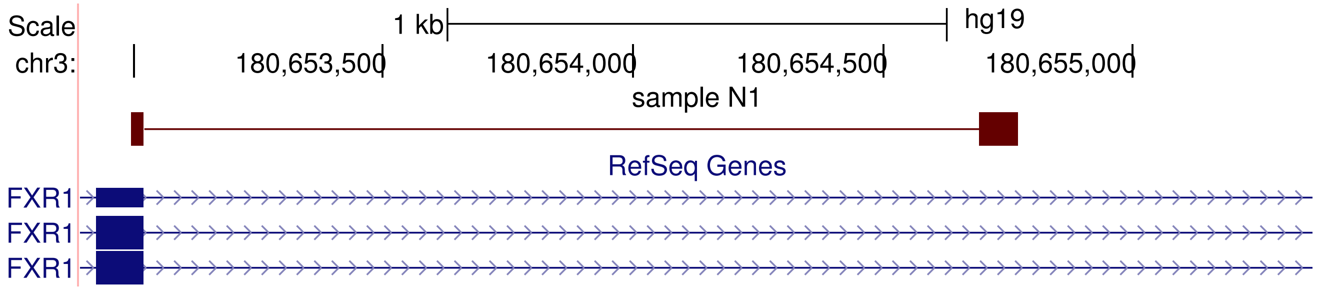


Figure S8: Example of a recursive splicing event happening on the 5' end of intron 14.

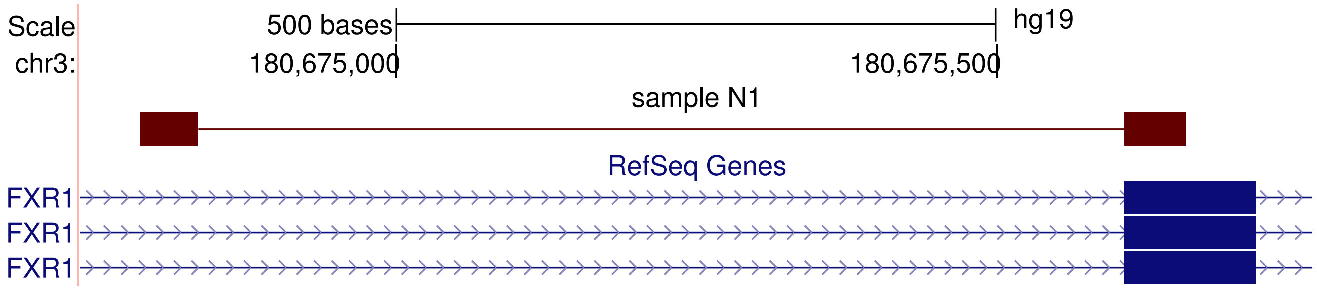


Figure S9: Example of a recursive splicign event happening on the 3' end of intron 10.

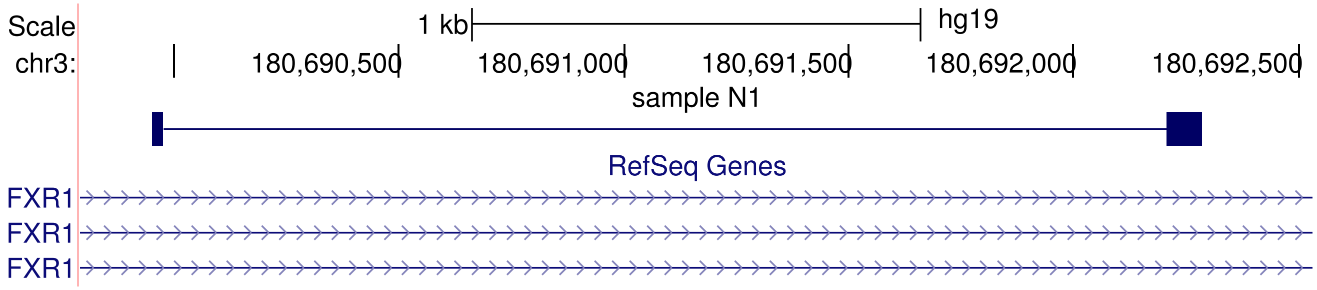


Figure S10: Example of recursive splicing with two non-canonical splice sites in intron 16.

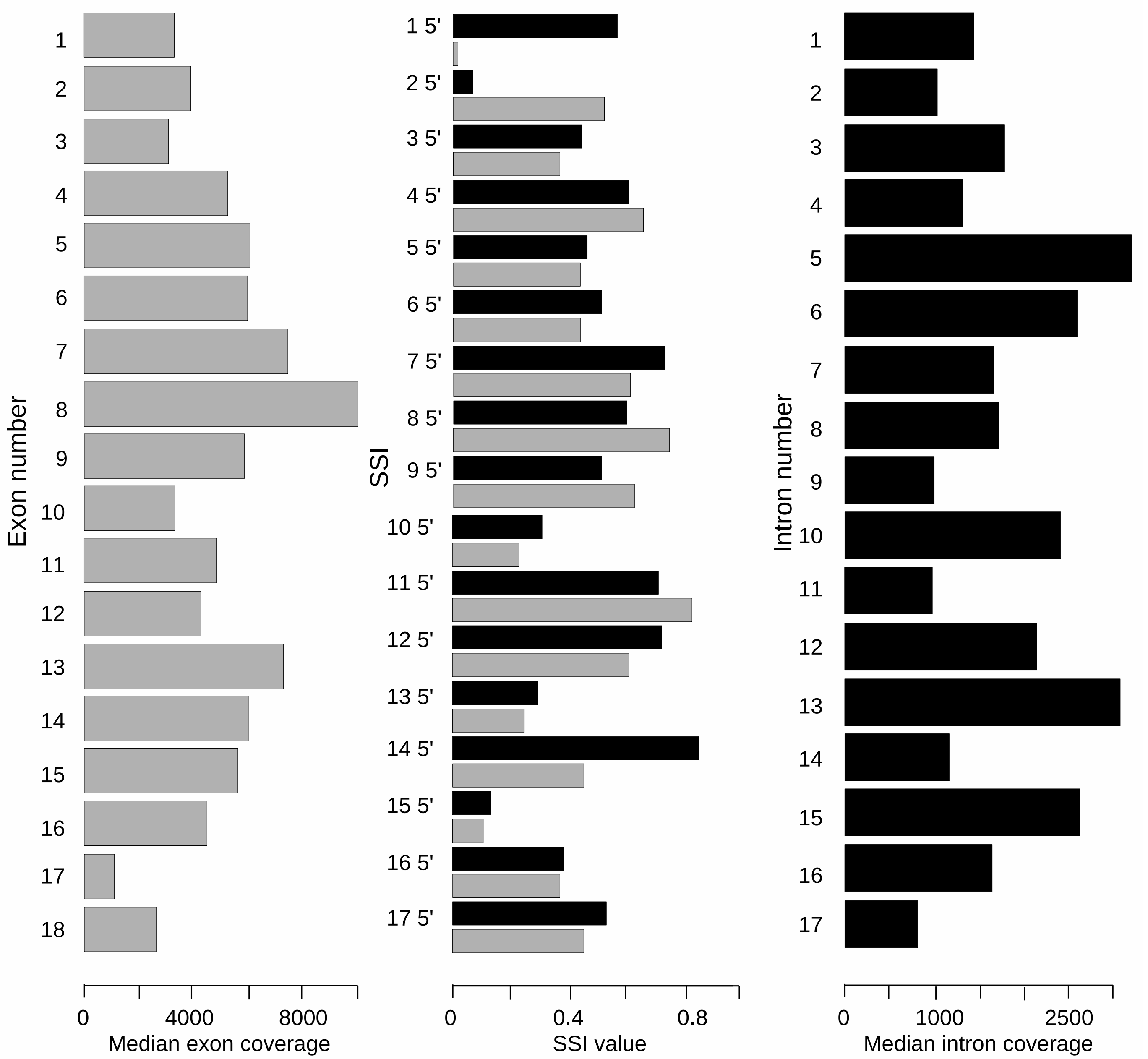


Figure S11: Splice site index (SSI) and medians of coverage of exons and introns in *FXR1* for sample C1 from the ENCODE dataset. Gray bars in the left panel represent the coverage of exons (exon 1 on top). Black bars in the middle panel represent SSI values for the 5' end of the introns and gray bars on the middle panel represent SSI values for the 3' end of the introns (intron 1 on top). Black bars in the right panel represent the coverage of introns (intron 1 on top). Data shown for chromatin RNA sample C2.

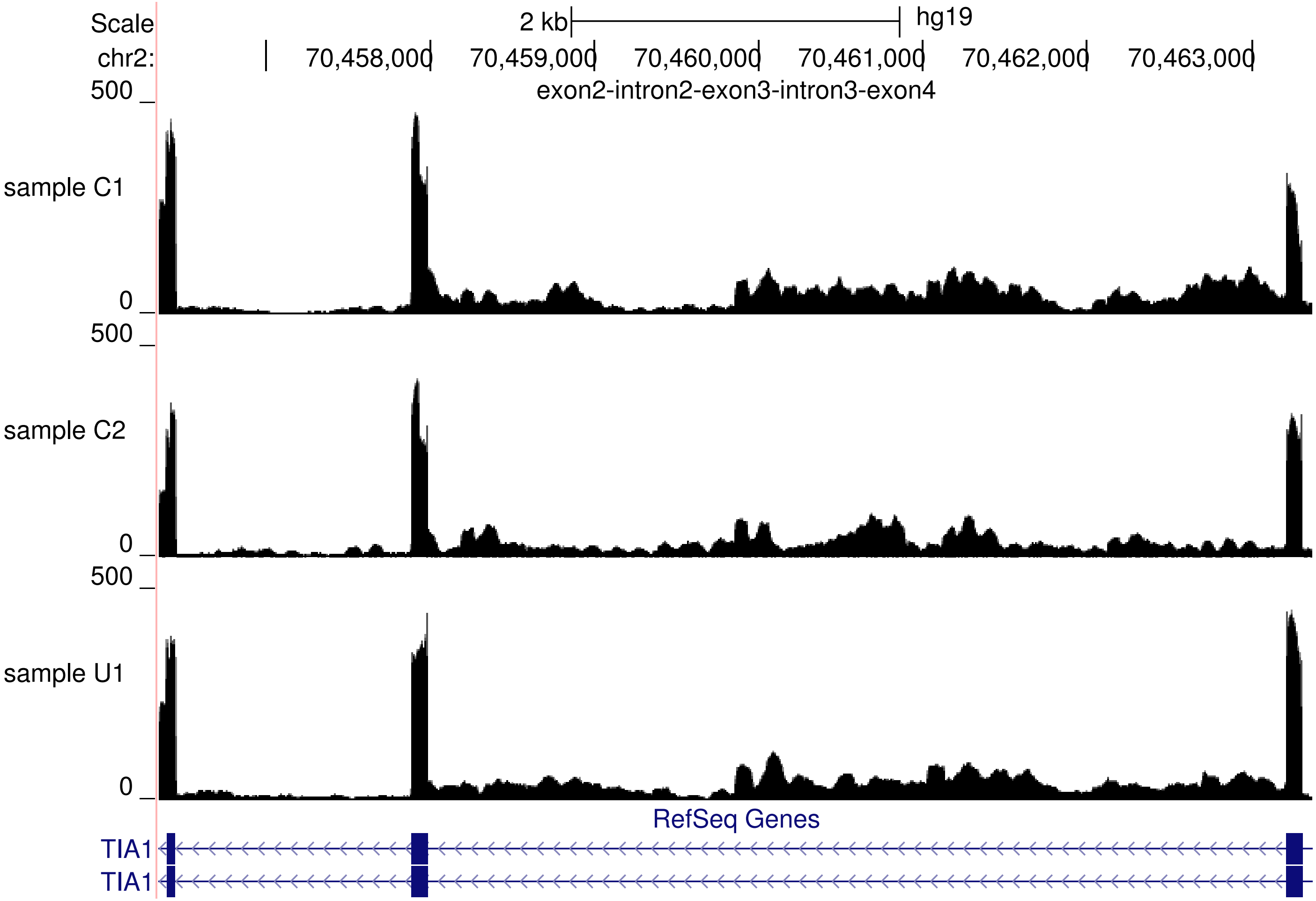


Figure S12: Graphical representation of two potentially non-sequentially spliced introns of *TIA1* – intron 2 is predicted to be spliced after intron 3. Top three panels represent the coverage from samples N1, N2 and U1. Coverage is the value on the y-axis, and the genomic coordinates are the value on the x-axis. The bottom panel represents the annotation of the gene available in the RefSeq database. Thick blocks represent exons, thin lines with arrows represent introns. Note that the gene is transcribed from the reverse strand and on the figure intron 2 is situated downstream (on the right) from intron 3. None of the introns with high coverage are annotated as retained introns, which gives an extra evidence that this is a case of non-sequential splicing and not a case of intron retention.

Chapter 3

Non-sequential and multi-step splicing of the dystrophin transcript

I. Gazzoli[1], I.Pulyakhina[1], N.E. Verwey[1], Y. Ariyurek[2], J.F.J. Laros[1,2], P.A.C. ’t Hoen[1], A. Aartsma-Rus[1]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
2 Leiden Genome Technology Center, Leiden University Medical Center, Leiden, the Netherlands

RNA biology, 2015 Dec 15:0. [Epub ahead of print]

3.1 Abstract

The dystrophin protein encoding *DMD* gene is the longest human gene. The 2.2 Mb long human dystrophin transcript takes 16 hours to be transcribed and is co-transcriptionally spliced. It contains long introns (24 over 10kb long, 5 over 100kb long) and the heterogeneity in intron size makes it an ideal transcript to study different aspects of the human splicing process. Splicing is a complex process and much is unknown regarding the splicing of long introns in human genes. We used ultra-deep transcript sequencing to characterize splicing of the dystrophin transcripts in three different human skeletal muscle cell lines, and explored the order of intron removal and multi-step splicing. Coverage and read pair analyses showed that around 40% of the introns were not always removed sequentially. Additionally, for the first time, we report that non-consecutive intron removal resulted in three or more joined exons which are flanked by unspliced introns and we defined these joined exons as an exon block. Lastly, computational and experimental data revealed that, for the majority of dystrophin introns, multistep splicing events are used to splice out a single intron. Our data show for the first time in a human transcript, that multi-step intron removal is a general feature of mRNA splicing.

3.2 Introduction

Splicing involves hundreds of proteins that coordinate the excision of introns from the pre-mRNA, joining the exons and resulting in mature mRNA transcripts. Multiple alternatively spliced transcripts can be produced from a single pre-mRNA molecule through a highly regulated process, and its disruption contributes to a large number of human genetic disorders that either directly cause disease or increase disease susceptibility [113]. RNA splicing occurs after assembly of the spliceosome on the pre-mRNA, which includes splice site recognition and intron removal steps [114]. Splice site recognition relies on the identification of exon/intron boundaries. This is achieved by five (U1, U2/U12, U4/U6 and U5) small nuclear ribonucleoprotein particles (snRNPs), together with more than 100 auxiliary proteins and trans-acting splicing factors (SR proteins and heterogeneous ribonucleoproteins (hnRNPs)) [115, 116, 117]. The correct recognition is supported by cis-acting splicing signals [118], such as the consensus donor (5') and acceptor (3') splice sites (SS), the branch point sequence (BP) and polypyrimidine tracts (PPT). Additional exonic or intronic splicing enhancers (ESE or ISE) and silencers (ESS or ISS) motifs can influence the inclusion or exclusion of an exon by recruiting trans-acting splicing factors. Intron removal (Figure 3.1, A) is the result of two phosphoryl transfer reactions during the spliceosome assembly formation on the pre-mRNA, and the catalysis can only occur after the intron is transcribed. The precise excision of the intron results in the release of a lariat RNA [119, 120] and in two ligated exons.

It has recently been reported that the chromatin structure, the transcript elongation rate and the pausing of RNA Polymerase (Pol) II can contribute to the regulation of splicing [121, 122, 123]. It has been established that splicing can occur cotranscriptionally, when the nascent transcript is still attached to the DNA through RNA Polymerase II [87, 121, 124, 125, 126], and/or post-transcriptionally, when splicing occurs after transcription has completed and the transcript has been transferred to a different nucleoplasmatic location, the speckles [126]. Additionally, Vargas et al. [126] showed that constitutive introns are mainly co-transcriptionally spliced, while alternative splicing may occur post-transcriptionally [87, 127, 128, 129]. The order of intron removal may confer an important regulatory layer for alternative splicing.

For large introns, the precise excision and the efficiency of splicing may be hampered by the presence of multiple splice site-like sequences. Furthermore, the physical distance between donor and acceptor splice sites offers a challenge. It has been suggested that secondary RNA structure leads to juxtaposition of remote canonical donor and acceptor sites to facilitate identification and joining of splice sites [130], but additional mechanisms to facilitate splicing of long introns have been reported for invertebrates, such as intron removal in multiple steps (Figure 3.1) [112, 131, 132, 133].

Recursive splicing can occur in different ways. In the 5' recursive splicing (RS) (Figure 3.1, B, left panel), a canonical donor splice site is spliced to an internal acceptor site, generating a 5' ratcheting point (5'RP) from the juxtaposed exon and 5' splice site sequences. A similar process can also take place at the 3' splice site, but now an internal donor splice site is spliced to the canonical acceptor, to generate a 3'RP (Figure 3.1, B, right panel). This process can be repeated multiple times, creating 5' or 3' splice sites (SSs) that can be used as donor or acceptor splice sites in the next splicing step. Alternatively, 5' and 3' RP steps can generate an intermediate intronic cassette (intermezzo), which is removed in the last step of splicing (Figure 3.1, B, bottom panel). Finally, intrasplicing or nested splicing has been proposed as a third potential mechanism [98, 133]. Here the intron is first shortened by one or more internal splicing steps using internal donor and acceptor sites, and then in the final step what remains of the intron is spliced out using the canonical 5'and 3' SSs upstream and downstream exons are joined (Figure 3.1, C).

Detailed studies on recursive splicing have been performed in Drosophila [112, 131, 132, 134], but only few analyses, for single intron, have been done for human [98, 135], and vertebrates [130].

The dystrophin protein encoding *DMD* gene is the longest human gene (2.2 Mb). The coding regions represent only 0.7% of the gene, and the gene has exceptionally long introns (average 28 kb, size range 107 bp - 360 kb). In the1990s, evidence for co-transcriptional splicing for dystrophin transcripts was provided [136]. This finding was expected, considering that full transcription of the gene takes an approximately 16 hours at an average elongation rate of 2.4 kb min^{-1}.

The size and complexity of the gene, containing 79 exons, long introns, 7 different promoters, 2 sites of polyadenylation and numerous alternative transcripts, has always hampered characterization of the *DMD* transcriptome and detailed analysis of its processing. Indeed, only recent experimental evidence of an internal lariat of dystrophin intron 7 suggested that this long intron (110 kb) might undergo to nested splicing [98].

In the last few years, the development of next generation sequencing technologies has opened a new horizon for the detailed analysis of transcript processing. The *DMD* gene is an excellent candidate for in depth analysis of the relationship between intron length and the order of intron splicing, as well as the occurrence of splicing of the large introns in multiple steps.

Here we present detailed analysis of dystrophin pre-mRNA intron splicing using targeted paired end sequencing of transcripts, Capture-pre-mRNA-sequencing. We provide evidence that the order at which introns are removed is not consecutive, leading to the formation of blocks of exons flanked by unspliced introns. We further show the occurrence of multi-step splicing in many dystrophin introns, and show for the first time the characterization and validation of recursive splicing in the dystrophin transcript.

3.3 Materials and Methods

3.3.1 Cell culture

All experiments were conducted using three immortalized control muscle cell lines (7304, Km155, 8220) generated by Zhu et al. [137] that were propagated and differentiated as described previously [138]. In short, cells were cultured in Skeletal Muscle cell medium ((PromoCell GmbH, Germany) with 20% Fetal Bovine Serum (FBS), 1 of penicillin/streptomycin (P/S; Gibco-BRL) at 37^{0}C in a humidified atmosphere with 5 CO_2. One hundred million cells were seeded and, as they approached a confluence of 70%, proliferation medium was replaced with differentiation medium (DMEM, 2% horse serum, 1% P/S) to obtain multinucleated myotubes. Cells were allowed to differentiate for 8 or 14 days.

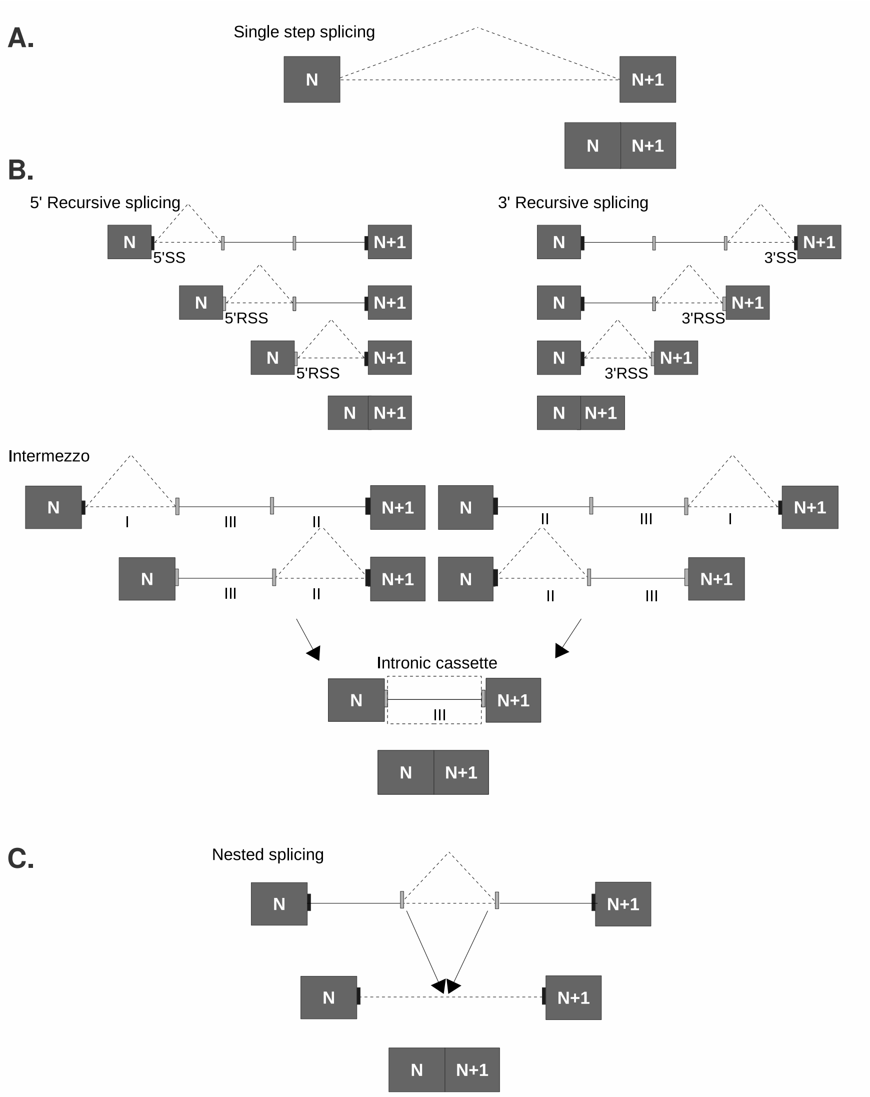


Figure 3.1: Single and multi-step splicing model. **A.** Single step splicing. The intron is fully removed in one step using annotated 5' and 3' splice sites (black rectangles), to join neighboring exons (gray boxes, (N and N+1)). **B.** Recursive splicing. In the case of 5' recursive splicing (5'RS) or 3' recursive splicing (3'RS), the intron is spliced in multiple steps (ratcheting points), each starting from the 5' or 3' splice site (SS), respectively (left and right panel). Each step generates a new 5' or 3' recursive splice site (5'RSS or 3'RSS, white rectangles) respectively. A combination of 5' and 3' recursive splicing or intermezzo can also occur (lower panel). In this intermezzo splicing, parts I and II of the intron are first removed beginning from the 5' or 3' splice sites, leaving part of the intron (III) containing new unannotated 5' and 3' RSS (white rectangles), after which the final part of the intron is spliced. **C.** Nested splicing. The first step(s) of intron splicing consist of removal of (an) internal part(s) of the intron using internal 5' and 3' splice sites (white rectangles). Subsequently, the remaining part of the intron is removed using the regular 5' and 3' splice sites that border the exon-intron boundaries (black rectangles).

3.3.2 Subcellular fractionation and RNA extractions

Nuclear and cytoplasmic fractions were separated as previously described , with minor changes. All steps were carried out on ice.

At eight and fourteen days after initiating differentiation, cells were harvested via trypsinization and centrifuged for 10 min at 2000g. After washing twice with cold PBS, the pellet was resuspended in 2 ml ice-cold sucrose buffer I (0.32 M sucrose, 3 mM $CaCl_2$, 2 mM magnesium acetate, 0.1 mM EDTA, 10 mM Tris-HCl, pH 8.0, 1 mM DTT, 0.5% Triton) and dounced ten times in a cold Dounce homogenizer. The resulting lysate was transferred to a new tube and mixed with 2 ml of sucrose buffer II (2 M sucrose, 5 mM magnesium acetate, 0.1 mM EDTA, 10 mM Tris-HCl, pH 8.0, 1 mM DTT). The sample was carefully layered on 2.2 ml of sucrose buffer II and was balanced using sucrose I buffer on top off the gradient, then centrifuged at 30,000 g for 45 min at 4^0C (SW 40.1 rotor). After centrifugation, the supernatant (cytoplasmic fraction) was carefully removed and treated with proteinase K for 1h at 37^{0}C, whereas the tight pellet consisting of nuclear fraction was dried at room temperature. Precipitation of the cytoplasmic fraction was performed using 0.1 volume of 3M sodium acetate, 2μl paint pellet co-precipitant (Novagen) and 2 volumes of 100% ethanol, followed by 48h at -80^{0}C. After centrifugation at 30,000g for 30 min at 4^{0}C, washing steps with several volumes of 70% (v/v) ethanol were carried out, followed by a second centrifugation with identical conditions. The pellet was stored at -80^{0}C for further RNA isolation. In parallel, the pellet of nuclei was gently rinsed with cold 1mM EDTA in PBS and resuspended with 200μl of ice-cold glycerol storage buffer (50 mM Tris-HCl, pH 8.3, 40% (v/v) glycerol, 5mM $MgCl_2$, 0.1mM EDTA), followed by RNA isolation or storage at -80^{0}C.

RNA from the nuclear and cytoplasmic fractions was isolated using NucleoSpin RNAII (Macherey-Nagel) and eluted into 50μl of water, following the manufacturer's protocol. Additional treatment with DNase-free RNase (Qiagen) was performed for 15 min at 22^{0}C, to completely remove DNA, followed by a precipitation step as previously described. Quality and concentration of isolated RNAs were tested using RNA lab on chip (Agilent's Bioanalyzer 2100) and aliquots were reverse-transcribed with SuperScript^{TM} III (Invitrogen). QPCR was performed with intronic and exonic primers for selected genes in order to test DNA contamination in the samples lacking reverse transcriptase.

3.3.3 cDNA synthesis

Four μg of pre-mRNA was used as template for cDNA synthesis. Reverse transcription was performed with 3μg/μl random hexamers primers (Invitrogen) at 55^{0}C for 1 h, using SuperScript^{TM} III first-strand Synthesis System (Invitrogen), according to the manufacturer's protocol. After the first strand was synthesized, a second-strand synthesis was generated by adding (5X) second strand synthesis buffer (Invitrogen), 25nM dNTPs, RNase H and DNA polymerase I (Invitrogen) for 2.5 h at 16^{0}C. The double stranded cDNA was then cleaned up with the MinElute PCR purification kit (Qiagen) and eluted in 30μl EB buffer.

3.3.4 Custom library design

A customized probe library was generated using the eArray software (Agilent Technologies), as described in the user's guide. The synthetic 120-mer biotinylated oligonucleotide probes (baits) in solution were tiled along targeted intronic and exonic regions of the *DMD* and three different human control genes (*FXR1*, *CKLF* and *ACTB*). The genomic sequences corresponding to the four target genes were based on UCSC hg19-GRCh37[1]. To ensure capturing of intron containing pre-mRNA transcripts and low abundant transcripts, each sequence of the gene (except repeat areas) was covered generally by at least four baits, and the *DMD* promoter regions were covered on average by eight baits. The maximum capacity of the synthesized library was up to 55K baits. The following parameters were chosen: sense strand, 1x capture-probe tiling frequency, layout strategy-centred, 20 bp overlap region between baits and avoid repeated masked regions.

3.3.5 Pre- and post-hybridization sample preparations and Illumina sequencing

We created a library starting from 4 μg of pre-mRNA. Five cDNA capture libraries were generated from three different cultures cell lines: 7304 cells (three independent biological replicates), KM155 and 8220 cell lines. Our method has been slightly modified from the version provided by SureSelect^{XT}

[1] chrX:31,137,336-33,357,726; chr3:180,630,234-180,695,106; chr16:66,586,466-66,600,190; chr7:55,70,372-55,66,779 for *DMD*, *FXR1*, *CKLF* and *ACTB*, respectively

Target Enrichment System for Illumina Paired-End Sequencing Library (based on Agilent Technologies' updated versions). An additional cDNA synthesis step has been integrated to the original procedure Agilent prepped library protocol, which was designed for genomic DNA. Since our customized capture library is highly specific for four genes only, no ribosomal RNA depletion was done.

To define the best sample preparation method, we generated two different cDNA libraries using random primers and pre-mRNA isolated from nuclei of a differentiated healthy muscle cell line. In brief, cDNA, synthesized as previously described, was sheared using a Covaris instrument (Covaris, Inc.) at duty cycle 5, intensity level 5 and 200 cycles for burst (180s). The second method has been tested in parallel, in which pre-mRNA was fragmented before cDNA synthesis by the addition of 5 times fragmentation buffer (Ambion), heating at 70^0C for 15 minutes and 5 minutes on ice. In both methods reverse transcription was applied to generate cDNA, using the same protocol as previously described and followed by purification with MinElute PCR purification kit (Qiagen). The cDNA ends were first repaired to obtain uniform double-stranded fragments with blunt ends, and then additional adenine residues were added to the fragment extremities to increase the efficiency of the following step. Finally, adaptors for Illumina sequencing were ligated in a concentration 2 fold less than provided in the instruction. Between each of the previous modification steps, a clean-up was performed using AMPure XP beads (Agencourt Bioscience Corporation) following the ratio beads/sample suggested by user's guide. With the exception after the adaptor ligation step, where the used ratio was 1:1 (volume). The following minor changes were made to the Agilent Technologies' protocol: the beads/sample thermo-mixture was incubated for 15 min at 22^0C in a thermo mixer (1200 rpm), fresh 80% (v/v) ethanol was used and the elution step was performed at 37^0C in a thermo-mixer.

A pre-hybridization amplification was performed with a limited number of cycles (5), reaching the required 500 ng of sample for the library hybridization step without amplification-induced biases. Primers were removed with 1 volume of AMPure XP beads, following previously described methods. This generated a cDNA library ranging in size from 160 to 660 bp. During the multiple steps of the sample preparations, the library quality was evaluated with an Agilent Bioanalyzer 2100 using HS DNA chips. The sample was concentrated to 3.4 μl and mixed with 2 μl of the customized capture library (Agilent Technologies, Inc. USA). The hybridization was further performed as described in the SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library manual. Post-hybridization amplification (12 cycles) with a different sequence index (barcode) per sample allowed pooling of samples, creating a multiplex libraries.

Amplified material was purified with AMPure XP beads, as early described previously. This step was repeated twice to minimize the amount of unused primers and reduce their sequence read bias. Using SureSelect^{XT} multiplex indexes, several post-capture amplified samples were pooled to a final concentration of 2 nM and with fragments size of 250-650 bp. The resulting pool of libraries was sequenced on an Hiseq PE 2x100 Illumina at a concentration of 7 pM. Output files in *fastq* format of the five Capture-pre-mRNA-seq containing paired-end reads and QC information were generated using CASAVA version 1.1.

3.3.6 Analysis workflow

In order to detect non-sequentially spliced introns, exon blocks and to identify recursive splicing events, we used the pipeline described in [139]. For the classification and motif analysis scripts, we refer the reader to the materials available online[2].

Alignment, post-alignment quality control: The pipeline first maped paired-end RNA-Seq data to a reference genome sequence (hg19, GRCh37) with GSNAP aligner version 2012-07-12 [107]. Only uniquely mapped reads with a maximum of 5 mismatches in each end of a read are reported. All

[2] https://git.lumc.nl/i.pulyakhina/pipeline_paper/tree/master

format conversions were done with SAMtools version 0.1.18 [39]. For downstream analysis we extracted reads mapped to the target genes: *DMD*, *FXR1*, *CKLF* and *ACTB*. The annotations containing the coordinates of each exon and intron for each gene have been extracted from the RefSeq database[3].

To remove samples with low sequencing yields, we included only samples where the number of reads mapped to the *DMD* gene was >500,000. Considering the length of the *DMD* transcript for Dp427m (2,092,329 nt) and the length of paired-end reads (2 times 100 nt) this cut-off means that each position of the *DMD* gene was covered on average around 50 times. The same cut-off was applied to RNA and DNA samples.

Coverage – median coverage of exons and introns: The median value of the coverage (Table S1) of each position within the corresponding intron or exon was calculated using the *median ()* command in *R* (version 2.5.1.1). We excluded such regions as promoters, UTRs and pseudogenes that can potentially influence the coverage and bias the median coverage of introns (Table S2). We only included areas of introns and exons that were covered by our designed probes (Table S3). This calculation reflected an accurate coverage for all introns, except intron 40, where a small area was highly covered even after removal of a known UTR. We considered this intron an outlier. Positions with zero coverage were also included in calculating the median. To make the coverage comparable in the different cell line samples, the coverage for each exon and intron was normalized using the average coverage of the *DMD* exons in all cell lines, and the median of the normalized coverage was used for further analysis.

Coverage – no GC bias: To estimate the influence of GC content on the median coverage, we calculated the length of each exon and intron and the fraction of nucleotides that consisted of **G** and **C** and built a linear regression of the coverage and the GC content for both DNA and RNA samples. Since no significant correlation of GC content with median coverage of introns was found (p-value=0.26 or higher), median coverage values were not normalized for the GC content.

Classification of reads: Reads were classified in three categories, based on the location of the alignment and the distance between the two mapped ends of a read pair (expected insert size is approximately 400 nt). According to the reference alignment the reads were aligned in the exon, intron, exon-intron boundary or exon-exon junction. Following the reads were classified as pre-, intermediate- and post-splicing. Mainly, reads belong to the intermediate-splicing category were used for downstream analysis. If the distance between the two mapped ends of a read pair were higher or lower than the expected insert size, reads were labelled as “large” or “normal”, respectively. Reads fully mapped to an exon were classified into a separate category.

Splicing order analysis: The median coverage of each intron was used to extrapolate an estimated the order of intron removal, based on an assumed correlation between the coverage depth and the relative speed of intron removal: the slower the intron is removed, the longer the target is available and the higher is the coverage. The average of the median normalized intron coverage from the five (Capture-pre-mRNA-seq) libraries was used for downstream analysis. We assessed 26 units of 5 introns, shifting 3 introns for each subsequent unit. Next we compared the normalized coverage of each intron in the unit, defining two recurrent values (90 and 130) as the cut-offs. Introns with an average coverage of less than 90 (low coverage) were considered to be spliced quickly (fast splicing), while introns with an average coverage of more than 130 (high coverage) were considered to be spliced slowly (slow splicing). The remaining introns with a coverage of 90-130 were considered to be spliced at an intermediate rate.

A paired-end split reads-based method was applied for a straightforward analysis to confirm the results of the coverage-based analysis. We counted paired-end reads having one split read spanning an exon-exon (ex-ex) junction and the second read mapped to the intron (in) immediately up- or downstream (Figure S1, A). The identification of this type of fragments is limited by the size of the captured

[3] http://www.ncbi.nlm.nih.gov/refseq, GRCh37.p13 RefSeq gene identifiers are NC_000023.10 (Chromosome X, *DMD*), NC_000003.11 (Chromosome 3, FXR1), NC_000016.9 (Chromosome 16, CKLF) and NC_000007.13 (Chromosome 7, ACTB)

fragments (250-650nt). However, internal *DMD* exons range in size from 32-275nt, and this allows for the detection of splicing intermediates that involve two or more consecutive exon-exon junctions. We used the total number of these split reads in all our 5 libraries, and calculated the splice-ratio for each intron as follows:

$$Splice\text{-}ratio = \frac{S}{S + NS} = \frac{ex_n\text{-}ex_{n+1}...int_{n+1}}{ex_n\text{-}ex_{n+1}...int_{n+1} + int_n...ex_{n+1}\text{-}ex_{n+2}} \tag{3.1}$$

In this formula, "ex_n-ex_{n+1} ... int_{n+1}" or "S" reflects the number of read pairs supporting sequential splicing, where one read of a paired-end spans an exon-exon junction arising from the splicing of the intron, while the other maps to the intron immediately downstream (i.e. for intron 33 this would be the number of read pairs where one read spans the exon 33-34 junction and the other read maps to the intron 34).

"int_n ... ex_{n+1}-ex_{n+2}" or "NS" reflects the number of read pairs supporting non-sequential splicing, where one read of a paired-end maps to an intron, while the other covers the exon-exon junction resulting from the splicing of the intron immediately downstream (i.e. for intron 33, one read pair would map to intron 33, while the other would map to the exon 34-35 junction).

We calculated the splice-ratio for each intron and defined introns with a splice-ratio between 0.5 and 1 as sequentially spliced, as the result of reads supporting sequential splicing (S) are more than non-sequential (NS), while introns with a splice ratio of <0.5 were defined as non-sequentially spliced.

Recursive and nested splicing: Potential recursive and nested splicing events were predicted using split reads belonging to the intermediate-splicing category. The first and second reads of each read pair were analyzed separately as single end reads. Each uniquely mapped read that contained a gapped alignment was selected. Two base pairs at the beginning and at the end of the gap, that were not covered by the reads, were classified as candidate donor and acceptor splice sites. The splice sites were assigned based on the split point of the read, the alignment of the flanking sequences and considering the two nucleotides that had the highest similarity to the splice site sequence. Identified split reads containing two annotated splice sites were discarded. However, when the donor and/or the acceptor were not present in the reference annotation of the gene, the read was selected for downstream analysis. We added the number of reads of the three biological replicates from the same cell line and performed the analysis for the three different cell lines and selected events present in all three datasets.

Predicted recursive splicing events were reported as a matrix containing the upstream and downstream genomic nucleotides flanking the position where the read was split. A matrix where the rows contained positions of the donor and the columns contained positions of the acceptor splice sites was created, and the intersecting cell represented the number of reads for that specific pair of the donor and acceptor site. All the splicing events were also listed in a separate text. We analyzed events happening within one intron. Reads split within the first 50 intronic nucleotides (near the exon-intron boundary) were not considered, as they were thought to represent variation in normal exon-exon splicing (including the well established NAGNAG splice site variations). We classified the events based on (Figure S1, A), as recursive (5'RS, 3'RS) and nested splicing.

Motif analysis: We performed motif analysis on the donor and acceptor splice sites from the predicted recursive and nested splicing events. We extrapolated the sequence of the annotated and non-annotated 5' and 3' splice sites from each event and additionally the two nucleotides upstream from each donor splice site and two positions downstream from the acceptor splice site. Four extracted nucleotides were used to create the sequence logo using Weblogo software[4].

[4] http://weblogo.berkeley.edu/

3.3.7 Experimental validation

For exon block validation, experiments were performed on nuclear RNA from two cell lines and triples were performed independently three times. For recursive splicing validation, experiments were performed on nuclear and cytoplasmic RNA from two cell lines. PCR primers for all targeted *DMD* introns and exons were designed using Genomic refseq ID NG_012232.1 (Table S4). As a template, 1 μg of isolated pre-mRNA was reverse-transcribed using SuperScript III (Invitrogen), following the manufacturer's instructions.

Exon blocks were validated using qPCR. Quantitative RT-PCR was performed in a 8 μl reaction containing 4 μl SYBR Green master mix (ThermoScientific), 0.2 pM each primer, and 2 μl of diluted cDNA template. PCR was performed on the LightCycler 480 (Roche Diagnostics Ltd.). Thermal Cycling conditions were as follows: 50^0C for 2 min, 95^0C for 10 min, 45 cycles of 95^0C for 15 s and 60 ^{0}C for 1 min. Analysis of the raw data and PCR efficiency was performed using the LinRegPCR software [106]. For all combinations of primers a Reverse Transcriptase negative control sample was included to exclude DNA contamination. A pair of primers covering the unspliced intron and immediate downstream exon was used to confirm the ability of the intronic primer to generate a PCR fragment. For primers spanning an exon-exon junction, there was little flexibility for primer design, resulting sometimes in low primer efficiencies. For each exon block, all qPCRs were performed simultaneously. HPRT was used as a reference gene.

PCR followed by Sanger sequence (Figure S1, B) was used to confirm the specific splice junctions in the predicted exon blocks and splicing events. For the exon blocks, we designed primers in the unspliced intron and in the last exon, where for the splicing events, we used specific primers upstream and downstream the split reads (Table S5). cDNA was generated from 100 ng of pre-mRNA using 2x master mix buffer (Ambion) and 1 μl of enzyme in a total volume of 50 μl. PCR reactions were carried out as per manufacturer's instructions. Each assay was performed for the two different cell lines. The PCR was performed using initial denaturation at 98^0C for 2 min followed by 35 cycles of (98^0C for 15s, 55^0C for 30s, 72^0C for 30s) and a final extension of 72^0C for 10min. The PCR products were subsequently analyzed on a 2% agarose gel. After purification with the MinElute PCR purification kit (Qiagen), the amplicons were analyzed using Sanger sequencing. The results were blasted in the UCSC genome browser[5] to confirm the correct sequence and identify intron-exons and exon-exon junctions for each exon block.

3.3.8 Data availability

The fastq and bam files used in this study are accessible online[6] through European Nucleotide Archive.

3.4 Results

To investigate the splicing of the dystrophin transcript in detail, we performed Illumina HiSeq paired end sequencing on pre-mRNA isolated from three differentiated human muscle cell lines. To enrich for pre-mRNA, we isolated RNA from cell nuclei. Since dystrophin is expressed at very low levels, we enriched for dystrophin pre-mRNA using a customized library that consisted of biotinylated probes covering all exons, introns, annotated promoters and UTRs of *DMD* as well as three control genes, excluding repeat masked areas. The captured cDNA was sequenced using Illumina HiSeq 2000 to generate paired-ends reads of 100 nt each producing between 8.5 and 11.5 million of reads for the different samples (Table S1).

[5] http://genome-euro.ucsc.edu/

[6] http://www.ebi.ac.uk/ena/data/view/PRJEB9401

Whereas many next generation sequencing analysis pipelines are available for analyzing mRNA-Seq data, a method for analyzing pre-mRNA has not been reported. To facilitate the analysis of our dataset, we developed a novel pipeline, *SplicePie* [139].

3.4.1 Sample preparation

We first generated and sequenced a library using DNA as input to confirm the specificity of the probes and to rule out any biases in the capture efficiency. Analysis of the DNA sample revealed a 900-fold enrichment with 56% of reads mapped to *DMD* out of 11.5 million of uniquely mapped reads. As expected for DNA, equal coverage of exons and introns was observed with the exception of repeat areas in which no probes were designed.

For the pre-mRNA splicing analysis, PCR analysis confirmed the absence of DNA from the RNA that was isolated from the intact nuclei of differentiated myotubes. Comparison of the fragmented and sonicated cDNA libraries from nuclei of a differentiated healthy muscle cell line (7304) revealed that the number of reads mapping to the *DMD* gene was 240,601 1 (1.1%) and 2,245,758 (12.7%) for the fragmented and the sonicated samples, respectively. Furthermore, for the fragmented sample, we did not observe a uniform coverage profile of the exonic (Figure S2, A) and intronic regions (Figure S2, B), while the coverage was much more uniform for the sonicated sample. Additionally, we generated a cDNA library using reverse transcription with poly (dT) primers to enrich for polyA RNA. However, this resulted in a very low coverage (638,012 (4%) of reads mapped to the *DMD* gene) and reduced representation of 5' ends of the *DMD* transcripts (data not shown). This was expected, as the length of the dystrophin transcript precludes the generation of cDNA from start to finish using oligo dT primers. For the following experiments random primers and sonication post-cDNA synthesis was used as sample preparation.

When analyzing the pre-mRNA, we observed clearly higher coverage in exonic regions compared to intronic regions. This could be due to the presence mixtures of pre-mRNA and co- or posttranscriptionally spliced mRNA in the nucleus. We therefore classified our paired-end reads into three categories: reads originating from post-, intermediate- or pre- splicing phases. The post-splicing category contained paired end reads spanning two different exons (ex˜ex), one exon and one exon-exon junction (ex˜(ex-ex)) or two exon-exon junctions (ex-ex)˜(ex-ex), implying completed splicing events. The intermediate-splicing category included read pairs where one read spanned an exon-intron boundary (ex-in; in-ex) or maps to an intron, while the other spans an exon-exon junction (ex-ex). Additionally, paired end reads mapping to the same intron or different introns (in˜in), but with a mapping distance (between the two reads) exceeding the library insert distance belong to the intermediate group. This category contained reads reflecting the initial and ongoing splicing events within one or multiple introns. The pre-splicing category contained paired end reads where one or both ends were mapped to the intronic sequences or exon-intron junctions (i.e. both reads did not cover exon-exon junctions), thus reflecting unspliced fragments. For the DNA sample, 99.9% of reads mapped to the pre-splicing group, which was expected because the DNA sample of course does not contain exon-exon junction reads. For the five pre-mRNA samples an average of 81% of mapped reads belonged to the pre-splicing group, suggesting pre-mRNA enrichment. Only 1.5-3.8% of read pairs fell in the intermediate category, probably due to the fact that splicing is a relatively fast process. The distribution of the reads over the different categories was similar for all samples.

3.4.2 Reproducibility of the method

We generated libraries from captured pre-mRNA isolated from muscle cell line 7304 after 8 days (biological duplicate) or 14 days of differentiation, and from muscle cell lines KM155 and 8220 after 8 days differentiation using the optimized protocol. Between 749,012 to 6,140,259 reads mapped to the human *DMD* reference sequence (6.9-65.8% of the total number of reads obtained) (Table S1). To allow comparison between different samples, the coverage was normalized by the number of reads mapping to *DMD* exons.

When analyzing the coverage of introns, we removed regions containing annotated promoters, UTRs, and expressed RNAs from the analysis (Table S2), because they can have high coverage unrelated to the splicing process. We did not normalize for the GC content, because we did not observe a correlation between the intron coverage and GC percentage of introns or the GC percentage of the probes. No 5'-3' bias (i.e., a bias in the coverage closer to the 5' or 3' ends) was detected either. This is a first indication for co-transcriptional splicing. When splicing would occur only after completion of transcription, the presence of nascent transcripts would lead to a higher representation of introns at the 5' end of the transcript. High intronic coverage was observed for sample 4 and 5 that were derived from 2 different cell lines (Table S1). For sample 1 and 2 (biological replicates of the third cell line) we observed that the percentage of reads mapping to the *DMD* gene was lower, while sample 3 (same cell line but differentiated for 14 instead of 8 days) had a higher coverage of *DMD* introns. Such biological variation is expected, since dystrophin expression is initiated upon differentiation and depends on the amount of myogenic cells in a culture. We ruled out sample preparation bias, since no large deviation was observed between the percentage of duplicate reads for *DMD* (Table S1) or the control genes (data not shown) between biological replicates or between different cell lines. Out of the reads mapping to the dystrophin transcript for the five capture-pre-mRNA-seq, 13-30% covered exonic sequences while 67-82% covered intronic sequences.

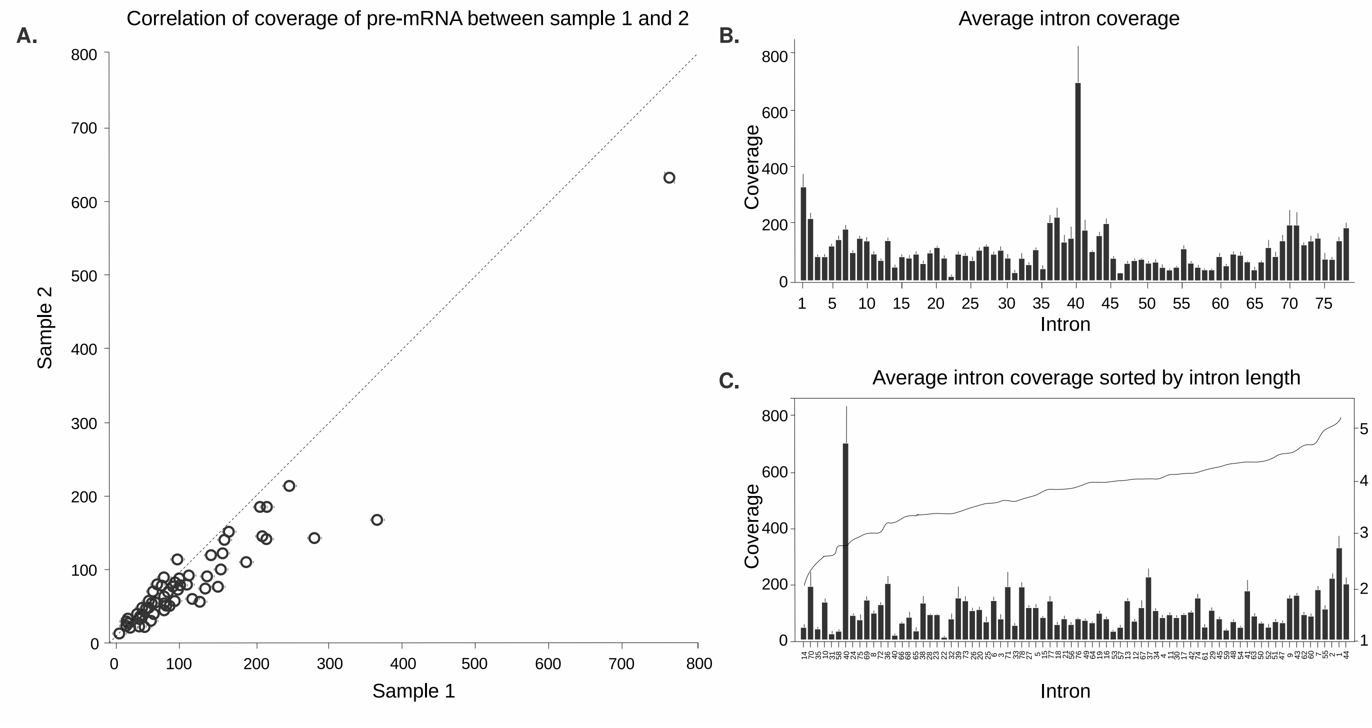


Figure 3.2: Graphical representation of the intronic coverage. **A.** Scatter plot showing a high correlation (r=0.96, p-value <0.0001) of the median normalized intron coverage for two biological replicates (samples 1 and 2) from the same cell line 7304. The dashed diagonal represents identity between the two samples. **B.** Bar graph showing the average normalized coverage of each intron for the three cell lines (error bars reflect the standard error of the mean). **C.** The same bar graph as shown in 2B, but now introns are sorted by length (ascending, length represented by the black dotted line and right y-axis ($\log_{10}$-scale).

To assess the reproducibility of *DMD* capture cDNA seq analysis, we compared the results from two biological replicates, performing two independent experiments with cell line 7304. This (Figure 3.2, A) showed a high correlation (r=0.96, p-value < 0.0001), indicating that the experimental procedure is consistent and reproducible, which was further confirmed by a significant correlation in exonic read distributions (data not shown).

We additionally analyzed two other cell lines (KM155 and 8220). The intronic coverage profiles of the three cell lines along the whole gene showed the same distribution pattern and similar depth (Figure 3.2, B).

3.4.3 Sequential and non-sequential splicing

We reasoned that the intronic coverage would correlate with relative speed of intron removal, i.e., introns that are spliced out quickly are expected to show low coverage, while introns that are spliced out slowly are expected to show higher coverage. We observed a lot of variation between the coverage of the different dystrophin introns, while for each intron the coverage was consistent between the 3 different cell lines (Figure 3.2, B). Since there is a large variation in the length of introns in dystrophin transcript, we first addressed whether the coverage was proportional to the intron length.

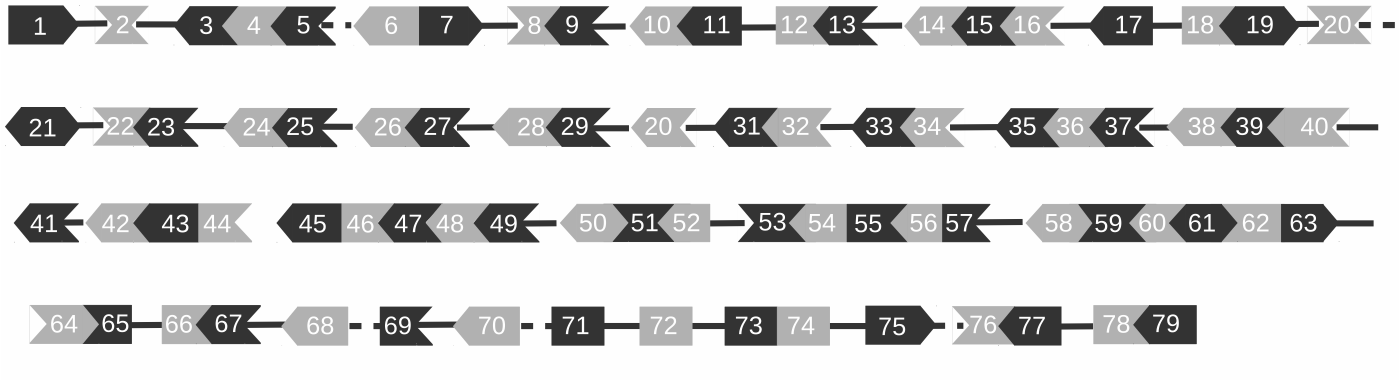


Figure 3.3: Representation of sequential and non-sequential splicing of the dystrophin transcript. Thick and dotted lines between two exons indicate preferential non-sequential splicing (slow and intermediate, respectively), while introns that are mostly sequentially spliced are not shown. The exon shape reflects the phasing of exons (in-frame vs. out-of-frame).

We defined intronic length as the amount of nucleotides covered by the probes and then subtracted sequences containing annotated promoters, UTRs, micro-RNAs for each intron and assessed the read density of the remaining intronic sequences. No significant correlation between intron length and coverage (Figure 3.2, C) indicating that short introns are not spliced before long introns. Rather, these results suggested that the introns are non-sequentially spliced. Therefore, some introns may be removed only after downstream introns have been removed and the splicing does not follow a strict 5'-3' order. Nevertheless, since transcription of the complete dystrophin transcript takes ˜16 hours, it is likely that a very slowly spliced upstream intron is spliced out before a very quickly spliced intron further downstream, simply because the downstream intron is produced hours later than the upstream intron. Therefore, we analyzed the relative order of intron removal in groups of 5 introns, using a sliding window of 3. For every group of 5 introns, each intron was classified as fast, intermediate or slow. Fast introns are represented by a low depth of coverage (normalized coverage <90) due to a quick intron removal compared to the slow ones, while high depth (normalized coverage >130) is an indication of slow splicing. A small group of introns with coverage between 90 and 130 were defined as “intermediate”. The classification of introns was very similar for each of the three cell lines showing a strong indications that several downstream introns were removed before upstream introns, and as a consequence of this, blocks of exons that were flanked by slowly spliced introns were identified.

Figure 3.3 shows a graphical depiction of sequential and non-sequential splicing, (and intermediate stage of few introns), of the dystrophin transcript. We propose the presence of blocks of exons, where 3 or more exons are joined flanked by slowly removed introns.

Sequential and non-sequential splicing events were corroborated by the analysis of paired-end reads from the intermediate-splicing category. To determine the nature of splicing of each intron, we considered intron (n) as a starting point. If intron (n) is spliced sequentially (S), it would be spliced before intron (n+1), leading to read pairs where one end would cover the ex-ex junction (ex_n-ex_{n+1}) and the other read would align to the flanking downstream intron (n+1) (Figure S1, A). Alternatively, a non-sequential (NS) splicing would result in the splicing of intron (n+1) before intron (n). This would be reflected by paired-end reads in intron (n) and in the exon-exon junction of the two exons immediately downstream of intron (n), (ex_{n+1}-ex_{n+2}), implying the presence of an unspliced intron and excluding that reads are derived from excised intron lariats, in which case paired-end reads would both map to the intron.

We defined the splice-ratio for any given intron as the number of reads suggestive of sequential splicing, divided by the sum of the reads suggestive of sequential splicing and those reads suggestive of non-sequential splicing. Intron were classified as being sequentially spliced when the splice-ratio was between 0.5 and 1, while introns with a splice-ratio below 0.5 were classified as non-sequential. For five introns out of 78, splice-ratios were slightly above or below 0.5, and classified as intermediate. Again, there was a good correlation between the 3 cell lines. We also observed a correlation between the intron coverage and the splice-ratio values (Figure 3.4), where introns classified as non-sequential based on the splice-ratio showed higher coverage (indicative of slower splicing) than introns classified as sequential (r=-0.32, p-value=0.0043). The fact that the intron coverage analysis may also have included excised lariats, while the paired-end analysis does not, may have prevented the correlation from being better than it is now.

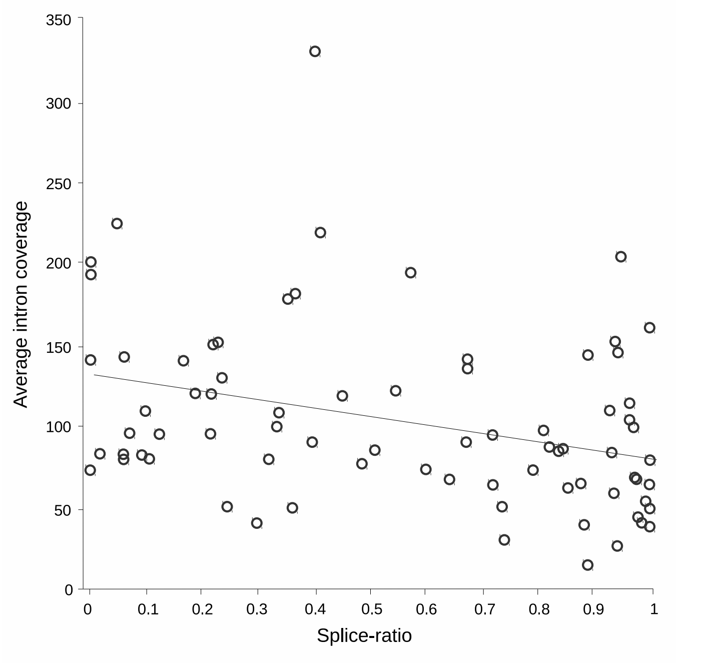
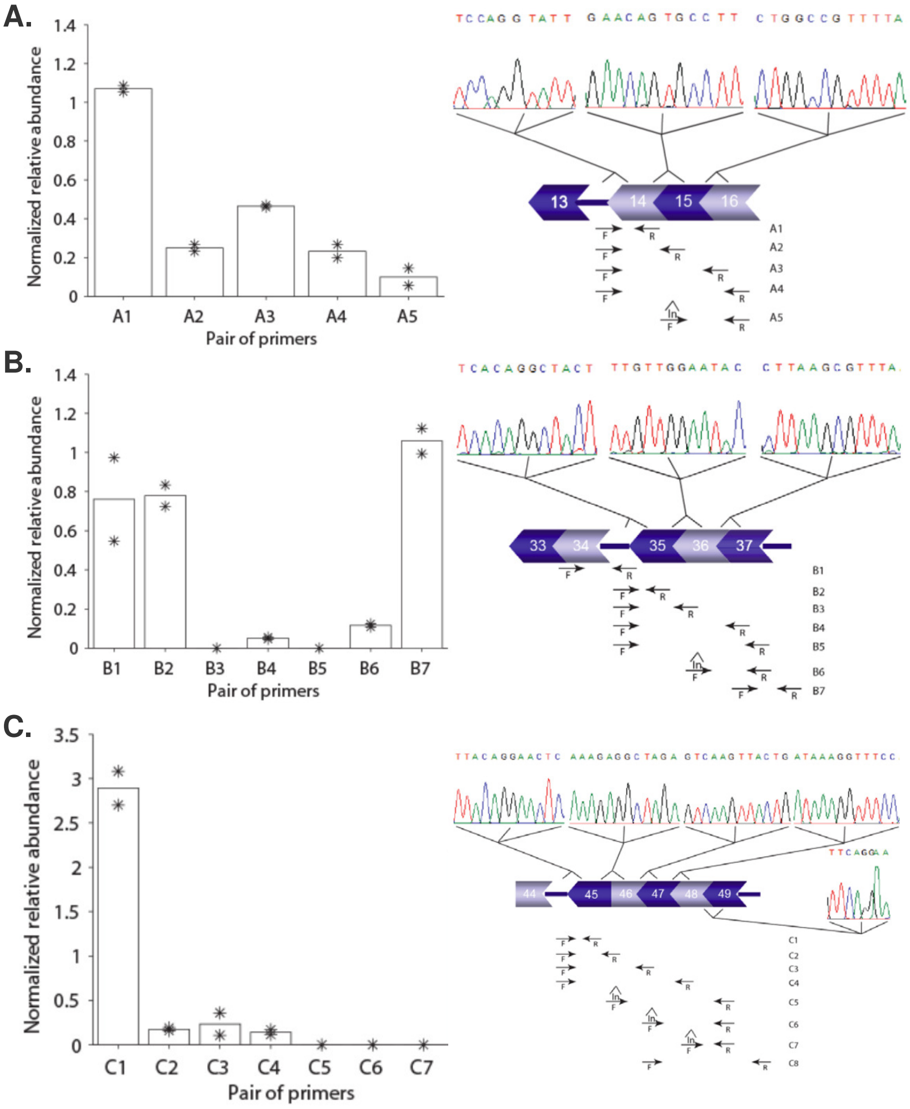


Figure 3.4: Scatter plot of the average intron coverage (y-axis) vs. the splice-ratio (x-axis) of each intron. An inverse correlation between the two methods is observed (r=-0.32, p-value=0.0043): lower coverage (relatively fast splicing) is associated with a higher splice-ratio, indicative of sequential splicing.

We experimentally validated the presence of these blocks using qPCR and Sanger sequencing analysis to confirm the presence of dystrophin pre-mRNA transcripts containing an upstream intron, when downstream exons had already been joined (Figure 3.5). With primers pairs (Table S4) that were designed to cover unspliced introns, exon-exon junctions and using predicted quickly spliced introns as a negative control, we confirmed exon block 14-15-16 (Figure 3.5, A). Using qPCR, we confirmed that intron 15 was spliced before intron 13. Additional evidence of the non-sequential removal of intron 13 was obtained using a forward primer in intron 13 and reverse primers on the exon 14-15 boundary. The relative abundance of the product with the exon 14-15 primer was lower than that obtained with the exon 15-16 primer, suggesting that intron 15 is spliced earlier than intron 14. This finding was supported by the presence of an additional PCR fragment that included intron 14 using primers in intron 13 and exon 16.

Conventional PCR and Sanger sequencing using a combination of primers in intron 13 and exon 16 showed the junctions of the three exons (14-15-16), confirming the predicted non-sequential splicing of intron 13. Likewise, we observed a block of exons 33-34 and exons 35-36-37 (Figure 3.5, B). In this case our data was supportive of non-sequential splicing of intron 34, as we did not detect introns 33, 35 and 36, while intron 34 was still present, albeit with low abundance. Quick removal of intron 35

Figure 3.5: Experimental validation of three of the predicted exon blocks. **A.** Intron13-Exons14-15-16. **B.** Exons 33-34-Intron 34-Exons 35-36-37. **C.** Intron 44-Exons 45-46-47-48-49-Intron 49. The same analysis has been performed for the three predicted cases. For each case the left panel (bar plot) shows (qRT-PCR) results representing relative abundance (to the first primer pair) of the spliced and unspliced introns using primer pairs in an intron and downstream exons, or exon-exon junctions (locations shown in the panel on the right). List of the primer pairs used in the qRT-PCR can be found in Table S4. *HPRT* expression was used for normalization. The qRT-PCR values are based on the average levels of two independent cell lines (individual levels (based on triplicates) are indicated with asterisks). The amplicons A1, B2, C1 were used as internal PCR efficiency controls. The detection of one amplicon (B3) was hampered by very low efficiency of the primer, mainly due to the hairpin and dimer structures. Attempts with an alternative primer did not improve the PCR efficiency. Sequential splicing of introns 14 and 15 is supported by amplicons generated with the pair of primers A2 and A3. Additionally, amplification with the forward primer in intron 14 and the reverse primer in intron 16 (A5), showed partial splicing of intron 14, supported by the difference in the relative abundance between A2 and A3. Sanger sequencing was used to confirm the three predicted exon blocks and the electropherograms (right panel) show the junction sequences for each case (intron-exon, exon-exon or exon-intron boundaries/junctions detected in a single fragment). The schematic illustration on the low side of the electropherogram shows the predicted exon block and the location of the primers used for qRT-PCR and Sanger sequencing PCR.



was validated as well, since we were unable to generate a PCR fragment using a primer pair in intron 35 and exon 37. Sanger sequencing confirmed the presence of a transcript containing introns 34 and 37, but without introns 33, 35 and 36.

Finally, a similar approach was used to test the third exon block including exons 45-46-47-48-49. As shown in (Figure 3.5, C), we were able to detect fragments using the forward primer in the intron 44 and reverse primers in exons 45-46, exons 46-47 and exons 47-48. Using forward primers in introns 45, 46 and 47 combined with reverse primers in exon-exon junction 48-49, no signal was detected, confirming these introns are indeed removed quickly. Sequential splicing of introns 47 and 48 was also shown using primers spanning the junction between exons 46-47 and intron 49. Furthermore, Sanger sequencing confirmed the exon block from exon 45 to 49, between the unspliced introns 44 and 49, validating that intron 45, 46, 47 and 48 can be spliced before intron 44.

In addition, we validated few more predicted sequential and non-sequential events directly by Sanger sequencing (Figure S2, B), showing splicing of intron 8 before intron 7, as well introns 50 and 51 spliced before intron 49.

3.4.4 Recursive and nested splicing

Since *DMD* introns are remarkably long, we hypothesized that multi-step intron removal, such as recursive and nested splicing previously identified in Drosophila, could occur during the splicing of *DMD* transcripts (Figure S2, B, C).

To identify potential recursive and nested splicing events in an unbiased way, we analyzed split reads and first filtered out split reads that aligned to exon-exon junctions or mapping within 50 nucleotides to an exon junction to maintain only split reads representing a splicing event with a non-annotated splice donor and/or acceptor site. We generated a matrix that included coordinates of the two genomic positions for each pair of donor and acceptor sites and the detected number of split reads supporting the combination. As intermediate splicing events are difficult to detect and may be rare, we jointly analyzed all split reads from the biological replicates. We only included events present in all three different cell lines to avoid observations that were a consequence of PCR duplicates and to provide stronger support for the genuine presence of these intermediate splicing events. Using this filter, we identified 414 splicing events (Table S5), 35% of which could be classified as potential recursive splicing, including 5'RP (18%), 3'RP (17%). Splicing events were observed at beginning, in the middle or at the end of the intron, and were independently of intron length. We also found 266 events (65%) indicative of nested splicing. Notably, for 27 introns we identified more than one type of events. This could indicate complementary or independent splicing mechanisms affecting the same or different transcripts, respectively, speculated to be due to RNA secondary structures. Finally, for 31 introns we established single step splicing (Table S6).

We tested 13 predicted events and performed RT-PCR amplifications across the split reads to detect the breakpoints using pre-mRNA from two libraries, followed by Sanger sequencing. We could validate 8 out of 13 events as shown in Figure 3.6. This level of success was higher than anticipated, given that Capture-pre-mRNA-Seq is a much more sensitive method compared to the standard RT-PCR. We chose 5'RP events identified in introns 42, 43 and 53 for the experimental validation. In introns 43 and 53, we confirmed the predicted 5'RP events, generating a spliced sequence of 3095 and 9536 bp, respectively. In both cases, sequencing of the expected PCR products (Figure 3.6, A) showed the junction of the exon 43 or 53 and the 5' ratcheting point.

A similar approach was used for the validation of 3'RP in intron 4, 25, 45, 53, 45. A predicted 3' recursive points in intron 4 was confirmed by Sanger sequencing (Figure 3.6, B). For some selected events, it was not possible to detect the breakpoint sequence. Furthermore, a few of the selected 5' and 3'RP sites were revealed to be intermezzo recursive splicing events, where the 5' and 3' RP sites were used as donor and acceptor splice sites. Intermezzo splicing occurs when upstream and downstream parts of an intron are spliced, leaving the internal area joined to the flanking exons. Theoretically, such an intermezzo intron could also be an alternative exon. Therefore, we amplified cDNA from pre-mRNA and cytoplasmic mRNA, arguing that intermezzo introns should not be present in cytoplasmic RNA. We could validate the intermezzo event for introns 7, 33 and 43 (present in nuclear RNA but not in cytoplasmic RNA) by PCR and Sanger sequencing (Figure 3.6, C). Interestingly, for intron 43 we detected two intermezzo events. No split reads spanning both intermezzo events were found, suggesting that only one intermezzo is used at a time. For the selected nested splicing event in intron 43, we predicted two genomic positions based on split reads. Sanger sequencing of the PCR product (Figure 3.6, D) showed the removal of 58,528 nucleotides from the intron. We observed the retention of nine nucleotides on each side of the predicted breakpoint. However, this retention could be due to misalignment of partial repeat sequences (TCAA) on both sides and the fact that we could pinpoint the removal of 58 kb by RT-PCR confirms this nested splicing event.

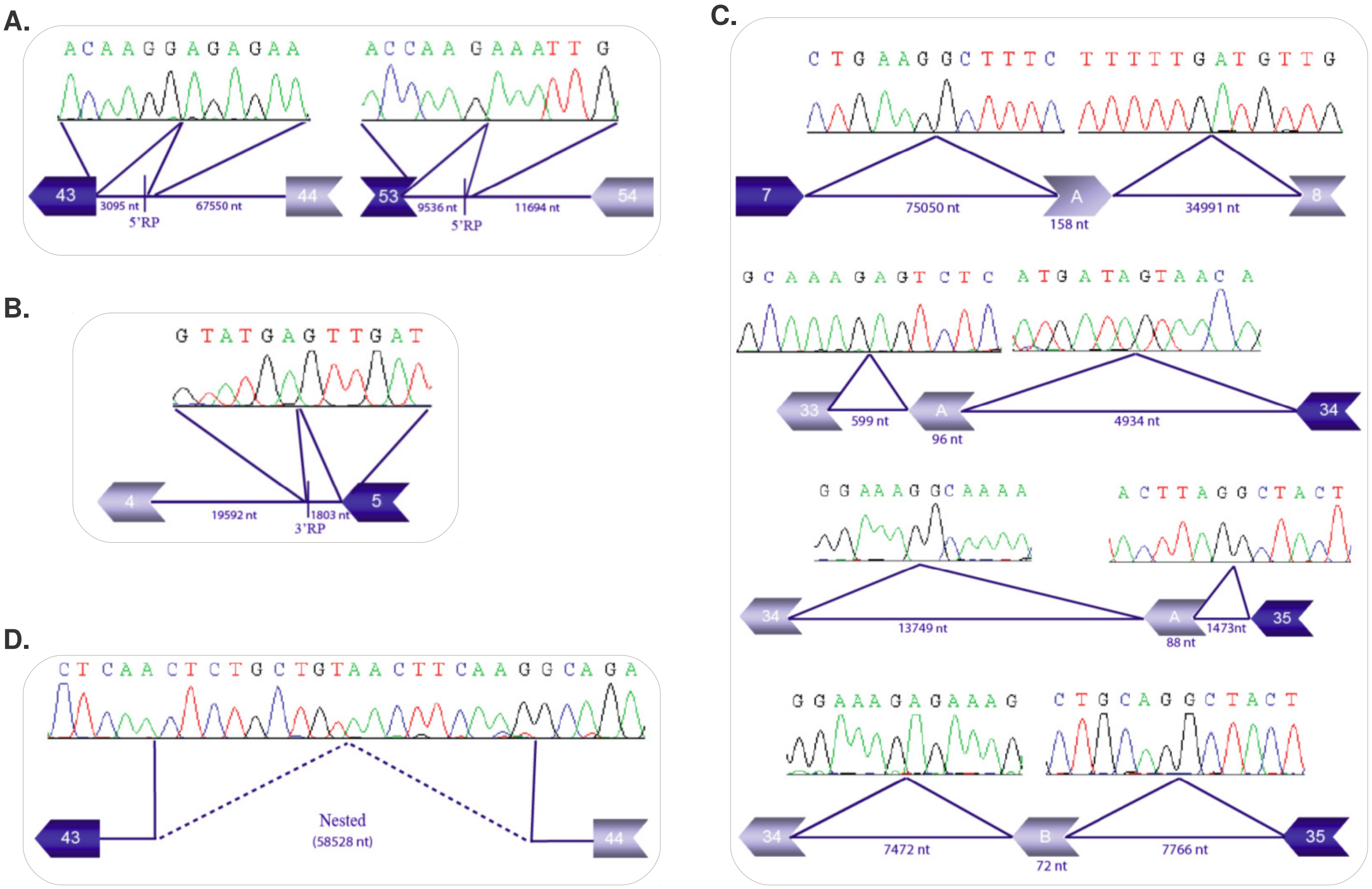


Figure 3.6: Examples of the experimental validation results of different types of intermediate splicing. **A.** Partial splicing of 3095 and 9536 nucleotides (nt) in introns 43 and 53, respectively, using a 5'RP, are reported in the upper panel. For introns 4, partial splicing of intron of 1803 nt used 3'RP. **B.** Each electropherogram shows the last six nucleotides of the exon joined to internal intron sequence as consequence of partial intron removal. **C.** Four preselected intermezzo events have been validated. In intron 7, a sequence of 158 nt (intermezzo, **A.**) was joined to the flanking exons 7 and 8, whereas another intermezzo event involving 96 nt was detected in intron 33. An area of 88 bp for intermezzo A and 72 bp for intermezzo B were identified between exons 34 and 35. **D.** In intron 43, we predicted nested splicing resulting from partial splicing (>58kb) of the intron. Two predicted genomic positions are indicated by blue continuous lines. Retention of nine nucleotides in each side of the split read (identified during the validation experiment) is reported between the continuous and the dashed lines. Joined point of the spliced intron is represented by a dashed line.

3.4.5 Motif analysis

We evaluated the motif of the areas involved in recursive and nested splicing by analyzing sequence conservation of the newly detected donor and acceptor splice sites and the two nucleotides upstream and downstream of these sites, respectively. As shown in the Figure 3.7, for the 5'and 3' recursive splicing (RS) we observed AG and GT as the most frequent motifs for the intra-intronic (non-annotated) acceptor and donor, respectively, showing most 5' RS and 3'RS use canonical splice site motifs (97% of acceptor sites and 95% of donor sites). In case of the nested splicing events, no clear consensus motif could be distilled for the non-annotated donor and acceptor splice sites.

3.5 Discussion

The use of target enrichment in combination with deep sequencing of cDNA offers an opportunity to study rare splicing events [140]. However, the identification of this small portion of intermediately spliced transcripts relies on the accuracy and sensitivity of the analysis and source material. Although RNA-Seq is an appealing approach to study dystrophin transcripts, the use of total mRNA is not

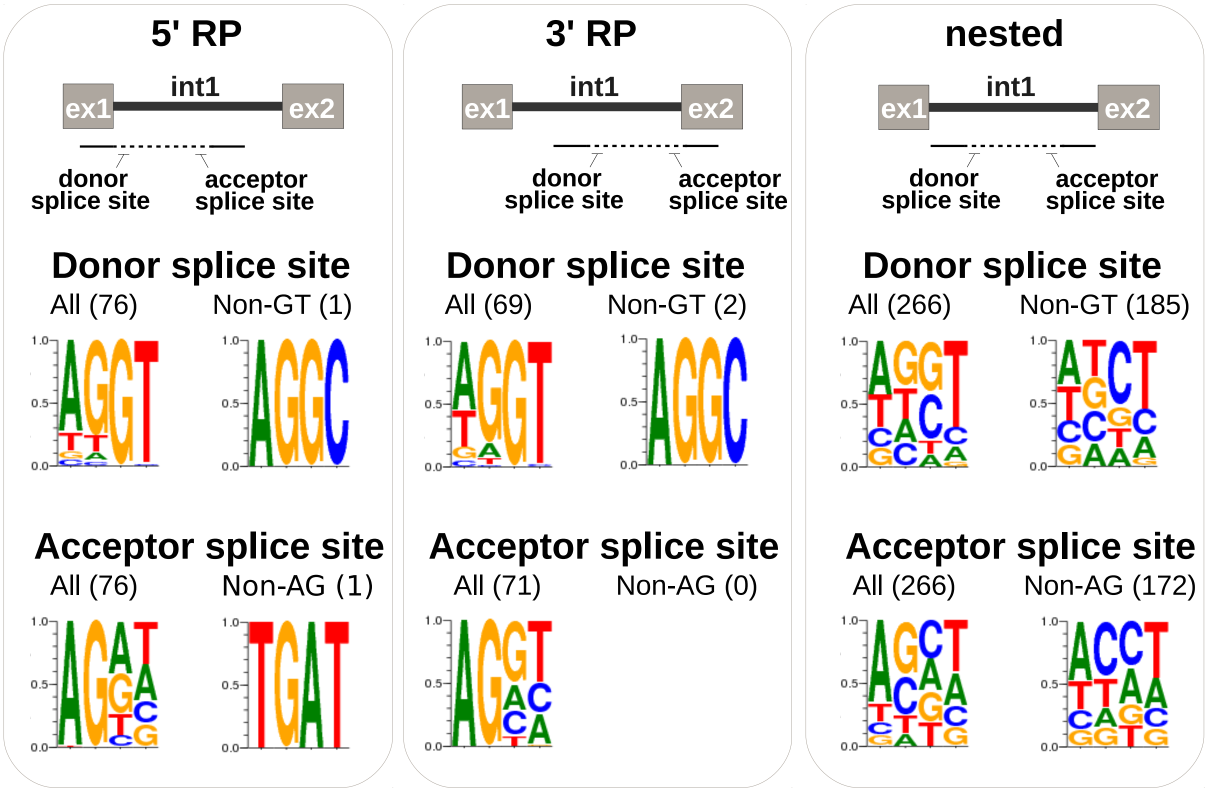


Figure 3.7: Motif analysis: sequence logo of the donor and acceptor splice sites involved in the recursive and nested splicing. On the first and second left panels, 5'and 3 recursive splicing (RS) are represented by the split read spanning the exon (1 or 2, respectively) and the middle part of intron 1. On the right panel (nested splicing) the split read spans part of the internal sequence of intron 1. The beginning and the end of the dotted line show the positions of the donor and acceptor splice sites involved in the splicing step of intron 1. Non-annotated donor and acceptor splice sites are indicated with a black arrow. Four nucleotides, including two from each splice site and two upstream of the donor and two downstream of the acceptor have been used to define the sequence logo. The 5' and 3'RS display a clear preference for the consensus splice site motifs. For nested splicing, sequences of both non-annotated splice sites display no particular consensus motif.

suitable, as the vast majority of sequence reads would reflect spliced transcripts. While this would be useful to identify i.e. alternative splicing or polyadenylation events, it would preclude the analysis of intron removal and transcript processing, because dystrophin is expressed at low levels, and the pre-mRNA transcripts would be in very low abundance. As such, it is unlikely that these transcripts would be picked up during mRNA-Seq analysis. Here, we present a similar approach of deep sequencing of a specific target gene using pre-mRNA isolated from nuclei as input material (Capture-pre-mRNA-seq). This method provided us an unprecedented way of understanding the details and mechanism of the splicing of *DMD* gene. Using subcellular RNA fractions and a solution hybridization library has been engaged before in RNA-Seq analysis for human genes [6, 87, 141], but for the first time the combination of these two methods is applied to a single gene with the aim of dissecting the splicing of large introns. Additionally we have previously developed a computational pipeline, SplicePie, to comprehensively analyse and detect intermediate splicing products [139].

Considering the complexity of the *DMD* gene (Table S7), with co-transcriptional activity varying intron sizes (between 107 bp-360,000 bp), it was hypothesized that the order of intron removal was not sequential. Based on our findings using two independent data analysis methods using SplicePie and experimental validation we confirmed that the order of intron removal does not follow a consecutive 5'-3' direction. Moreover, the relative speed of intron removal is not dependent on intron length, as initially hypothesized. Others have reported that the intron length does not influence the order of intron excision [87, 121]. Additionally, studies in other genes have shown that downstream introns can be spliced before upstream introns [142, 143, 144], and that introns flanking alternative exons are more prone to be spliced slower [87]. The speed and efficiency of intron removal may be regulated by co-transcriptional activity [145, 146, 147]. Since it is now apparent that intron removal does not always follow "first come, first served" model [148] a "first served, first committed" model has been proposed that takes the speed of the RNA Pol II activity into account [129], where the rate of RNA Pol II elongation affects the speed of splicing factor recruitment to different splice sites, facilitating introns excision independent of the co-transcriptional direction and strength of the splice sites. The identification of non-sequential intron removal in *DMD* has been supported by the evidence that exons can be joined to generate what we defined as "exon blocks" . These joined blocks of exons flanked by introns are intermediate steps of the final mature RNA. Notably, our findings reflect a propensity of sequentially or non-sequentially spliced introns.

Mutations in the *DMD* gene underlie a severe muscular dystrophy, Duchenne muscular dystrophy and a milder muscular dystrophy, Becker muscular dystrophy, depending on whether mutations disrupt or maintain the open reading frame, respectively [149]. Antisense oligonucleotide-mediated exon skipping is a therapeutic approach that aims to restore the reading frame for *DMD* transcripts for Duchenne patients to convert a severe phenotype into a milder one [150]. Our findings explain previous findings, where the use of one or two antisense oligonucleotides (AONs) could result in the skipping of multiple exons. Indeed as previously reported [151, 152], all tested AONs targeting exon 8 resulted in skipping of both exon 8 and 9, and here we show experimental validation for this. Another notable example is the exon 45-51 skipping [153]. The exon 45-55 area is a hotspot for *DMD* deletions [154, 155, 156], and skipping these 11 exons would be beneficial for 40% of patients [157]. So far inducing exon 45-55 skipping has been challenging for human *DMD* [153], but successful in the murine *Dmd* gene [158]. Nevertheless, this required a mix of 10 different antisense oligonucleotides, which is untenable for clinical development based on translational and regulatory challenges. Our data on exon blocks however, provides insight in how to induce exon 45-55 skipping with less antisense oligonucleotides, by targeting the blocks rather than individual exons. Based on our data exon 45-55 skipping could be technically challenging since intron 44 is spliced non-sequentially, while intron 55 is not. However, by targeting the exon blocks involving exon 45-49, exon 50-52 and exon 53-57 it might possible to achieve in-frame exon 45-57 skipping.

Additionally for the first time, we found evidence for recursive and nested splicing for different *DMD* introns, employing different ways of multi-step splicing and more likely in the long introns (Figure S2). Previous evidence from another long human gene, *UTRN*, indicated that intron length did not correlate with the time required of splicing [121]. Additionally the authors showed that introns in the range of 1.2kb to 240kb were spliced within 5-10 minutes, suggesting that the physical distance between donor and acceptor splice site is kept small by mechanisms like recursive splicing or alternatively, by the associated 5'splice site to the C-terminal domain of the RNA polymerase II, increasing the efficiency of splicing and reducing the time. Currently, Sibley et al. [159] reported recursive splice sites with high incidence in long introns in all vertebrates and most of the 435 identified in the longest human genes.

Our data showed that the average size of "single step spliced introns" is 6.4 kb (107-38,368 bp), while introns spliced via multi-step splicing introns are on average 40 kb long (650-248,401 bp), suggesting that, as anticipated, multi-step splicing involves generally longer introns. The 31 introns, for which no evidence of multistep intron splicing was found, were on average shorter than introns exhibiting multistep splicing (6.5 kb vs 40 kb), but multistep splicing was also found in short introns (shortest introns: 650 bp for recursive splicing and 7.5 kb for nested splicing). Likewise, we observed that single step introns were primarily spliced sequentially (19/31), while multi-step introns were primarily spliced non-sequentially (24/47). Notably, 72% of introns in the first half of the gene were spliced in multiple steps and non-sequentially, while in the central region (exon 45-55) introns were generally spliced in multiple steps but in a sequential manner, while 65% of introns in the last part of the gene were spliced sequentially in a single step.

We assumed the 31 introns, for which no evidence of multi-step splicing was observed, were spliced in a single step, but since the frequency of reported recursive and nested events was sometimes low, we cannot exclude the possibility that some of these introns are removed in multiple steps. For the remaining 47 introns, evidence for multi-step splicing was found for each of the three tested human skeletal muscle cell lines, including 5' and 3' recursive splicing and nested splicing, which could be validated by RT-PCR analysis. Additionally, during the experimental analysis a few of the predicted 5' and 3'RS turned out to be "non-annotated" donor and acceptor splice sites of intermezzo events. This suggested that some of the other predicted recursive events were also intermezzo splicing events.

Recently, two independent groups reported evidence of recursive splicing in few different human genes [159, 320]. Both works provided experimental validation of intermezzo splicing, where the inclusion or exclusion of a "recursive exon" could be detected in the mRNA or was part the last step of splicing. A previous case of nested splicing was also reported in *DMD* intron 7 [98]. However, this event was not detected in our dataset, even when taking only single cell lines into account. This discrepancy can be due to a different method of identification. Suzuki used PCR primer pairs with downstream forward primers and upstream reversed primers to generate fragments from lariats removed during nested splicing in RNA isolated from a single cell line. We analyzed multiple cell lines with capture-pre-mRNA-seq. It is possible that the events reported by Suzuki et al. [98] occurred in our cell lines, but we were unable to pick them up, or alternatively that they occurred only in the cell line he used.

Although only the results of the *DMD* gene have been reported here, extensive analysis has shown non-sequential and recursive splicing in one of our control genes (*FXR1*) in five capture libraries, which could be validated experimentally [139]. This suggests that recursive splicing may constitute a common mechanism to remove larger introns or introns from transcripts undergoing complex splicing pattern.

Motif analysis of sequences involved in multi-step splicing events for *DMD* revealed that recursive splicing relies primarily on known 5' and 3' consensus splicing motifs. By contrast, no real motif could be identified for nested splicing events. For 63 events we identified conventional GT-AG sequences, this was not the case for the majority of events. This suggests that a different, as yet unknown mechanism is employed for nested splicing.

In conclusion, our work provides splicing analysis of the dystrophin transcript at an unprecedented depth, shows evidence for non-sequential removal of introns, generating exons block, and multi-step intron removal as a common mechanism for dystrophin intron splicing.

3.6 Acknowledgements

This research was financially supported by ZonMW Vidi (project 91710303 awarded to A.A.R.) from the Netherlands Organization for Scientific Research (ZonMw). We would like to thank Prof. Vincent Mouly (Institute de Myologie, Paris, France) for providing patient cell cultures, Peter J. van't Hof for creating ENA accession number, Dr. Arthur R. Houweling for providing technical assistance for preparing figures and Dr. Willeke van Roon for her valuable proofreading of the manuscript. LUMC's Sequence Analysis Support Core is gratefully acknowledged for assistance in the analysis of the next generation sequencing data.

3.7 Appendix

Supplementary Materials are accessible online:
https://git.lumc.nl/i.pulyakhina/thesis/blob/master/Full_thesis/Supplementary_material_DMD_chapter.pdf

Chapter 4

Aging is associated with increased incidence of alternative splicing

I. Pulyakhina[1,*], V. Takhaveev[2,*], M. Vermaat[1], M. van Iterson[3], M.S. Gelfand[2,4], J.F.J. Laros[1,5], P.A.C. 't Hoen[1], BIOS consortium[6], LifeLines[7], Leiden Longevity Study[8]

1 Dept. of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia
3 Dept. of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
4 A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
5 Leiden Genome Technology Center, Leiden, The Netherlands
6 The Biobank-based Integrative Omics Study, part of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL)
7 LifeLines Cohort Study and Biobank (LL), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
8 Leiden Longevity Study (LLS), Leiden University Medical Center, Leiden, The Netherlands

[*] *These authors had an equal contribution to the paper.*

Manuscript in preparation

4.1 Abstract

Human aging has been associated with large-scale changes in cellular composition and gene expression. However, the influence of aging on RNA processing has not been studied extensively yet.

Here, we utilize a large collection of transcriptomes from healthy individuals to study alternative splicing during human aging. Whole blood was collected from 626 individuals between 18 and 81 years of age. The blood samples were subjected to RNA sequencing. Using a range of statistical approaches, we evaluated various characteristics of alternative splicing, taking into account potential confounder effects of phenotypic traits and age related changes in the cellular composition of blood.

We discovered that the rates of exon skipping and intron retention significantly increased with age. Although the identified changes in individual exons were not significant after multiple testing correction, such changes cumulated into significantly increased transcriptome-wide exon skipping and intron retention rates. We consider this a general trend, as the affected genes do not cluster in specific biological pathways and happen throughout the transcriptome. The increased exon skipping and intron retention rate contributed to the observed increase in GC content of the transcriptome with age. Furthermore, more non-canonical donor splice sites were used in older individuals.

Our findings indicate that the splicing machinery undergoes age related changes. These changes lead to increased incidence of alternative splicing events such as intron retention and exon skipping and promote the usage of splice sites with unconventional nucleotide motifs.

4.2 Introduction

Aging is one of the major risk factors for chronic disease in human population. Human aging is characterized by the changes occurring in the individual's appearance as well as at the cellular and molecular levels. Gene expression was found to undergo drastic changes with age in all examined tissues [160, 161, 162, 163, 164, 321] but the extent of these changes differed per tissue [166]. Substantial age-dependent decoupling of mRNA and protein expression was observed after the second decade of life, and the authors speculate that it could be caused by the deregulation of a specific set of RNA binding proteins and microRNAs [167].

Alternative splicing, a mechanism allowing genes to produce multiple transcripts through the use of different splice sites within the same pre-mRNA molecule, has not been studied extensively in the context of aging. Splicing is performed by the spliceosome, a big RNA-protein complex that excises certain regions of pre-mRNA (mainly annotated as introns) and joins the remaining regions (mainly annotated as exons). Alternative splicing is a common process which affects around 95% of human multi-exon genes [12, 84, 168, 169]. The efficiency of splice sites usage and the transcripts produced via splicing vary depending on tissue- and cellular origin, gender [170, 171] and genetic factors [172, 173]. Alternative splicing of individual genes has been shown to contribute to fundamental biological processes, such as synapse formation [174] and cell migration [175]. Alternative splicing was also found to be associated with age related disorders such as amyotrophic lateral sclerosis [176] and Alzheimer's disease [177]. Using microarray technology, Harries et al. were among the first to demonstrate age-associated disruption of the balance between alternatively expressed isoforms from the same gene in blood and suggested that modification of mRNA processing may be a feature of human aging [178]. A more recent study [179] employing RNA sequencing technology indicated that changes in splicing might be occurring over the entire lifespan in the prefrontal cortex and cerebellum of human brain. However, large scale sequencing-based studies focusing on the details of splicing changes with age are currently lacking.

In this study, we explored a cohort of over 600 male and female whole blood RNA-Seq samples with an age distribution between 20 and 80 years. We used another cohort of over 600 samples with the age distribution between 50 and 70 years to validate our initial findings. We performed the first in depth temporal splicing profiling of human blood and addressed changes in alternative slicing events, canonical and non-canonical splicing. We also assessed whether age related changes in alternative splicing accumulated preferentially in certain pathways and functional classes.

4.3 Materials and methods

4.3.1 General information

Whole blood from two Dutch biobanks, "LifeLines" [180] (which will be referred to as *LL*) and "Leiden Longevity Study" [181] (which will be referred to as *LLS*), has been subjected to RNA sequencing. Non-strand- specific polyA+ RNA sequencing of globin-depleted whole blood was performed on Illumina HiSeq2000 sequencers. 2x50bp paired-end RNA-Seq reads were aligned to the reference genome hg19 (NCBI genome build 37, release 104) using the STAR 2.3.0e [182] aligner allowing for at most eight mismatches and a maximum of five alternative mappings. The resulting alignments were stored in *bam* files. Sequencing and the primary analysis (aligning and checking quality) of the data was performed within the Biobank-based Integrative Omics Study (BIOS, part of the Biobanking and Biomolecular Research Infrastructure Netherlands, BBMRI-NL[1]).

We will discuss the findings in LL in Section 4.4 and the replication results in LLS in Section 4.5.

We used the Ensembl v.71 annotation (which corresponds to Gencode v.16) to define exon coordinates, and the expression was further calculated for the annotated exons. Overlapping exons (on the same or opposite strands) were excluded from the alternative splicing analysis. Only exons that have a single tag in the Ensembl annotation (i.e., "skipped exon" or "retained intron") were considered for the analysis of exon skipping and intron retention.

4.3.2 Calculating exon inclusion rate

To evaluate the rate of skipping or retention of a particular exon while accounting for total gene expression level, we assessed the relative inclusion of this exon, or its *inclusion rate*. The inclusion rate R_ℓ of exon l is the ratio between the coverage of l (normalized for its length) and the coverage of the gene (normalized for its length) containing l (the coverage of a region is calculated as the average number of reads covering it):

$$R_\ell = \frac{C_\ell}{C_{g(\ell)}}, C_\ell = \frac{\sum_{i=1}^{d(\ell)} N_i}{d(\ell)}, C_{g(\ell)} = \frac{\sum_{k \in g(\ell)} \sum_{i=1}^{d(k)} N_i}{\sum_{k \in g(\ell)} d(k)}. \tag{4.1}$$

Here, ℓ is the exon; $g(\ell)$ is a gene containing exon ℓ; C_ℓ is the exon coverage; $C_{g(\ell)}$ is the gene coverage; i is the number of a base within the exon; $d(\ell)$ is the length of the exon ℓ; N_i is the amount of reads covering the base with the number i; $k \in g(\ell)$ means an annotated exon situated in the gene $g(\ell)$.

4.3.3 Measures for changes in splicing

To track age related changes in splicing, we looked at several measurements and their changes between samples from individuals of different age. Two groups of measurements, annotation based and annotation free, were used.

Annotation based measurements comprised inclusion rates of individual exons as well as the sum of inclusion rates over the whole transcriptome ("global" measures of splicing, see Section 4.4.2 for more

[1] http://www.bbmri.nl/en-gb/activities/rainbow-projects/268

detail). We assessed the sum of inclusion rates for all exons from expressed genes that were annotated as "retained introns" or "skipped exons". Genes were considered expressed when the average gene coverage in all samples was at least 1 RPKM. This resulted in a list of 6,959 genes.

Annotation free measurements related to the usage of *de novo* identified splice sites (Section 4.3.4). We calculated the fraction of non-canonical donors (donors different from "GT") and non-canonical acceptors (different from "AG") in different samples, as well as the fraction of splice site pairs:

1. canonical donors paired with canonical acceptors;
2. non-canonical donors paired with canonical acceptors;
3. canonical donors paired with non-canonical acceptors;
4. non-canonical donors paired with non-canonical acceptors.

Splice sites found in over 1% of the samples were selected to count the number of canonical and non- canonical ones. The number of U12 introns [183, 184] (introns spliced via the non-canonical U12 splicing [185, 186]) were obtained from U12DB [187] and the linear regression was performed on this number to see whether the frequency of U12 splicing changed with age (for more detail about the linear regression, see Section 4.3.5). We also assessed the average number of donors paired with one acceptor and the average number of acceptors paired with one donor. We explored whether the number of acceptors per donor (and donors per acceptor) increased or decreased with age. Within the annotation free measurements, we also addressed the nucleotide content of used splice sites. A more technical and detailed description of the analysis pipeline, as well as the source code of the scripts, is available online[2].

4.3.4 Identifying used splice sites *de novo*

In order to identify splice sites *de novo* (in an annotation free manner), the following approach was developed (Figure S1). A read was called "split" when at least two pieces were mapped to distant locations (up to 500 kb apart) in the genome. The two positions at which the read was split were considered putative splice sites. To reduce the influence of small and possibly spurious overhangs, we considered only splits where five bases of the read adjacent to the split position on each side had an exact match with the reference sequence. Only when at least one read was split at the same position in at least 1% of the samples, these putative splice sites were selected. Since the RNA-Seq data in this study is not strand specific, it is not known which of the two positions – upstream or downstream – is the donor and which is the acceptor splice site. To predict the strand of the transcript represented by the read, five positions around each breakpoint (two in the exon and three in the intron) were extracted. The reference nucleotide motif of these ten nucleotides around the splice site was extracted from literature [188]. We compared this reference nucleotide motif with the ten nucleotides around each split and their reverse complement. If the edit distance between the extracted sequence and the literature profile was smaller than for its reverse complement, the upstream splice site was considered the donor and the downstream splice site was considered the acceptor, and vice versa. When both the extracted ten bases and their reverse complement were equally similar to the literature profile, the splice site (149 sites in the LL biobank) was excluded from further analysis.

Possible spliceosome slippage events [189, 190, 191] were not considered as independent splice sites. To this end, we clustered tandem donors (and acceptors) when they were situated at a distance of three, six, or nine nucleotides from each other. In this case, only the splice site associated with the highest number of split reads was selected, and the numbers of reads of the adjacent splice sites were added up.

[2] https://git.lumc.nl/vatakhaveev/alternativeprocessingversususage/tree/master/description_of_the_whole_pipeline

4.3.5 Designing statistical models to evaluate age related changes in splicing

To assess age related changes transcriptome wide (*global* measures of splicing), such as sums of inclusion rates, fraction of canonical/non-canonical splice sites and the average number of acceptor/donor splice sites used with one donor/acceptor, we performed the following linear regression:

$$M_s \sim \alpha + \beta \cdot A_s + \gamma \cdot lnSD_s + \vec{\delta}^{\top} \vec{f_s}, s \in S. \tag{4.2}$$

Here S is the set of samples; M_s is a certain measure of splicing in a sample s; α is the intercept; A_s is the age of a sample s; $lnSD_s$ is the logarithm of the sequencing depth of a sample s; β and γ are the coefficients; $\vec{\delta}$ is the vector of coefficients; $\vec{f_s}$ is the vector of covariates, containing measures of phenotypic traits of the sample s (i.e., concentrations of different blood cells). The β coefficient and its p-value reflect age dependence of a measurement. $lnSD$ is essential to normalize for, as more events were measured at higher sequencing depth. We normalized for the natural logarithm of sequencing depth and not the sequencing depth itself, as its natural logarithm has a higher correlation (ρ is 0.91 against 0.89)[3] with our measurements (Figure S2). Note that throughout the manuscript the Spearman's ρ coefficient of correlation is shown as well as the β coefficient, as the correlation coefficient is more common and easier to interpret for the reader.

The full list of phenotypic traits that were accounted for in the linear regression contains the following:

1. Gender (numerical, “0” or male, and “1” or female).
2. Height (numerical, in cm).
3. Weight (numerical, in kg).
4. Smoking status (numerical, “yes” or 2, “gave up smoking” or 1, and “no”or 0).
5. Concentration of HDL (numerical, high density lipoprotein) cholesterol (mMol per L).
6. Concentration of LDL (numerical, low density lipoprotein) cholesterol (mMol per L).
7. Concentration of triglycerides (numerical, mMol per L).
8. Concentration of platelets (numerical, 10^9 cells per L).
9. Concentration of neutrophils (numerical, 10^9 cells per L).
10. Concentration of lymphocytes (numerical, 10^9 cells per L).
11. Concentration of monocytes (numerical, 10^9 cells per L).
12. Concentration of eosinophils (numerical, 10^9 cells per L).
13. Concentration of basophils (numerical, 10^9 cells per L).

[3] Note that throughout the manuscript the rho parameter reflects Spearman's coefficient of correlation (pure Spearman's correlation, and not the result of a statistical model). However, for this particular case, Pearson's coefficient of correlation is used.

To assess age related changes in the relative inclusion of each individual exon ℓ present in the annotation, we calculated its inclusion rate in every sample and performed multi-variable linear regression according to the following model:

$$R_{\ell,s} \sim b + \beta \cdot A_s + \vec{c}^{\top} \vec{f}_s, s \in S. \tag{4.3}$$

Here S is the set of samples; $R_{\ell,s}$ is the relative inclusion of an exon ℓ in the sample s. Use of the other symbols is identical to the use in the formula above.

Analyses stratified for gender gave similar results to joint analyses. Therefore, samples of both genders were analyzed together, and gender was included as one of the covariates. However, data obtained from chromosomes X and Y were ignored to minimize the influence of gender.

4.4 Results

4.4.1 Age distribution and phenotypic traits

The analysis was performed on the whole blood transcriptomes from the "LifeLines" Dutch biobank with a wide age distribution (20–80 years, median around 45, Figure 4.1, A). Several phenotypic characteristics and biochemical parameters were measured for each sample in each biobank alongside age (see Section 4.3.5 for the full list of phenotypic traits). It has previously been shown that blood composition – counts and concentrations of various blood cell types – changes with age [192]. We also observed this in our data. The concentration of lymphocytes decreased with age ($\rho = -0.15$, $p = 2.4 \times 10^{-4}$, Figure 4.1, B). In all downstream splicing analyses, we included cell type composition, gender, other measured phenotypic traits and sequencing depth as covariates in the linear models.

4.4.2 Rates of exon skipping and intron retention increase with age

Using the statistical model described in Section 4.3.5, we assessed the change of alternative exon usage with age. It has recently been shown [166] that the expression profile of certain genes changes with age. In order to normalize for age related differential gene expression while assessing changes in splicing, we used inclusion rates (Section 4.3.2) for the exons available in the Ensembl v.71 annotation.

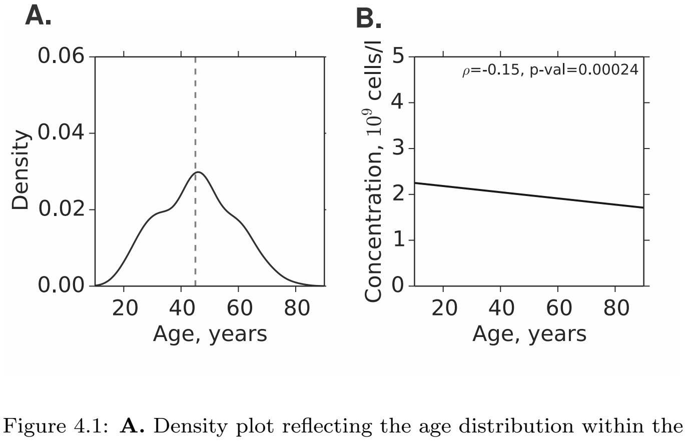


Figure 4.1: **A.** Density plot reflecting the age distribution within the LL biobank. **B.** Scatter plot reflecting age related changes in the counts of lymphocytes. Line represents the fitted linear regression lines for LL data. p-value of the linear regression is shown in the right top corner.

We selected genes expressed in all samples (6,959 genes) and looked at the changes in the inclusion rate of each annotated exon with age. Out of these 6,959 genes expressed in all samples, 4,660 genes contained at least one exon annotated as skipped; 2,823 genes contained at least one exon annotated as retained intron; 2,354 genes contained both exons and introns annotated as skipped and retained, respectively. We refer to the coefficient in the linear regression reflecting age related changes in the ratio as *beta*.

The majority (5,435, or 67.4%) of skipped exons had a negative beta coefficient, which indicated that their relative inclusion decreased with age (Figure 4.2, A, black). However, these changes were mainly non-significant, we observed only 15 significant negative betas (Figure 4.3, A) and two significant positive betas (Figure 4.3, B and C) after adjusting p-values using the False Discovery Rate method of Benjamini and Hochberg, with 5% cutoff (see Supplementary Table **??** for the full lists of genes containing exons with significant betas).

After exploring the changes of individual exons, we investigated the cumulative change in the rate of exon skipping. We calculated the sum of inclusion rate for all exons annotated as skipped, and refer to this as the *exon skipping rate*. We observed that the sum of inclusion rates for skipped exons decreased with age (Figure 4.2, B) ($\rho = -0.15$, $p = 3.3 \times 10^{-4}$; $\beta = -3.12$, $p = 6.7 \times 10^{-4}$). The decreased coverage of this group of exons is suggestive of their overall increased skipping in elderly individuals.

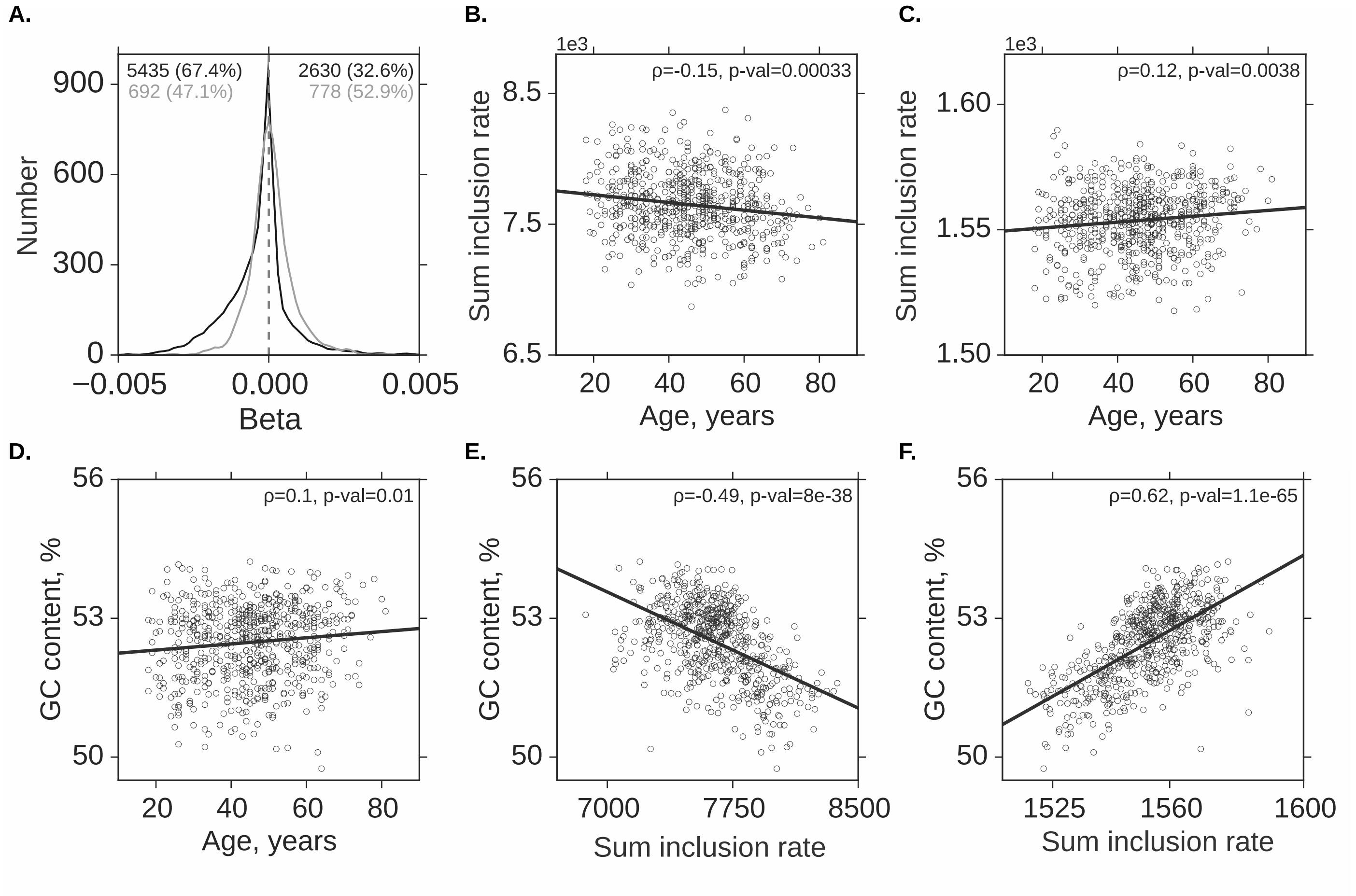


Figure 4.2: **A.** Distribution of betas for exons annotated as skipped exons (black) and retained introns (gray). x-axis contains the beta values, y-axis contains the numbers of events. **B.** Scatter plot showing the correlation between age (x-axis) and the sum inclusion rate of skipped exons (y-axis). **C.** Scatter plot showing the correlation between age (x-axis) and the sum inclusion rate of retained introns (y-axis). **D.** Scatter plot showing the correlation between age (x-axis) and the average GC content (y-axis) calculated based on the sequenced reads. **E.** Scatter plot showing the negative correlation between the sum inclusion rate of skipped exons (x-axis) and the average GC content (y-axis) calculated based on the sequenced reads. **F.** Scatter plot showing the correlation between the sum inclusion rate of retained introns (x-axis) and the average GC content (y-axis) calculated based on the sequenced reads. For the figures B-F each dot represents one sample. The line represents the fitted linear regression line. The data was generated using the statistical model described in Section 4.3.5.

We discovered that the majority (778, or 52.9%) of exons annotated as retained introns (from now they will be referred to as retained introns) showed increased relative inclusion with age. Similarly to skipped exons (an example of an exon with significantly increased skipping rate in the *BLK* gene is shown on Figure 4.3, A), the majority of changes were not significant, only one intron with a positive beta had a significant p-value of 9.1×10^{-5} (after adjusting p-values using the False Discovery Rate method of Benjamini and Hochberg, with 5% cutoff). This alternative splicing event is located in the *DIDO1* gene, death inducer-obliterator 1, and might be a case of alternative acceptor splice site usage (leading to the extension of an exon) rather than an intron retention event (Figure 4.3, B). An example of a real intron retention event, however, non-significant after applying the FDR cutoff, can be found on Figure 4.3, C.

In accordance with this, sums of inclusion rate of all retained introns (Figure 4.2, C) significantly increased with age ($\rho = 0.12$, $p = 3.8 \times 10^{-3}$; $\beta = 0.14$, $p = 4.3 \times 10^{-3}$). This is suggestive of an overall increased rate of intron retention in elderly individuals.

We performed a pathway analysis of exons undergoing changes (a significance threshold of FDR=10% was used). We did not find any pathway or group of genes with a similar function to be enriched for the genes containing exons with age related changes of their relative inclusion, when compared to a control set of genes expressed in our blood samples but without changes in exon inclusion rates.

4.4.3 Increased skipping and retention rates contribute to higher GC content

An important characteristic of an RNA-Seq sample is its average GC content – the fraction of **G** and **C** nucleotides out of all the reads[4]. It should be noted that, unlike GC content of DNA, which can only change due to mutations, GC content of a transcriptome can change due to the changes in expression of GC-rich and GC-poor transcripts and exons.

We observed a higher GC content[5] in older compared to younger samples (Figure 4.2, D, $\rho = 0.1$, $\beta = 0.0085$, $p = 4.7 \times 10^{-3}$). We made an effort to exclude potential technical confounders that may have caused this association. Longer sample storage times were also associated with higher GC content, but there was no confounding between age and storage times (Figure S3). Other experimental factors such as sequencing date or blood collection date were not associated with GC content of the samples (data not shown).

We looked for the biological factors explaining the increase in GC content of the sequencing files from older and younger individuals. For that, we extracted reads mapped to all skipped exons and retained introns and calculated the GC content of there reads (Table 4.1, last four columns). We also calculated the GC content of skipped exons and retained introns based on the reference genome annotation (Table 4.1, second and third columns). We discovered that, compared to the average GC content of all annotated exons, the GC content of exons annotated as skipped was lower (Table 4.1). At the same time retained introns had a higher average GC content compared to all introns.

As it has already been shown (Section 4.4.2), sum inclusion rate of skipped exons decreased with age and skipped exons had lower GC content (calculated on the sequencing files) than average (calculated on the reference genome, Table 4.1). We explored whether the sum inclusion rate of skipped exons correlated with the GC content of a sample. We revealed a significant negative correlation (Figure 4.2, E) between the sum inclusion rate and the GC content of a sample ($\rho = -0.49$, $p = 8 \times 10^{-38}$; in the regression with the GC content as the dependent variable, the sum inclusion rate as a predictor and the covariates used before, the coefficient of the sum inclusion rate equalled -0.0018 with the $p = 8.1 \times 10^{-47}$).

Similarly, a significant positive correlation (Figure 4.2, F) was found between the sum inclusion rate of retained introns – intron retention rate – and the GC content of a sample ($\rho = 0.62$, $p = 1 \times 10^{-65}$; in the analogous regression the coefficient of the sum inclusion rate equalled 0.041 with $p = 1.3 \times 10^{-74}$).

Table 4.1: GC content of exons and introns in the LL biobank.

biobank	all exons	all introns	skipped exons		retained introns	
			pos. β	neg. β	pos. β	neg. β
LL	51.76%	48.13%	49.67%	49.28%	53.97%	53.40%

[4] Note that this measure is calculated on the sequencing files and not on the reference genome.

[5] calculated on the sequencing files.

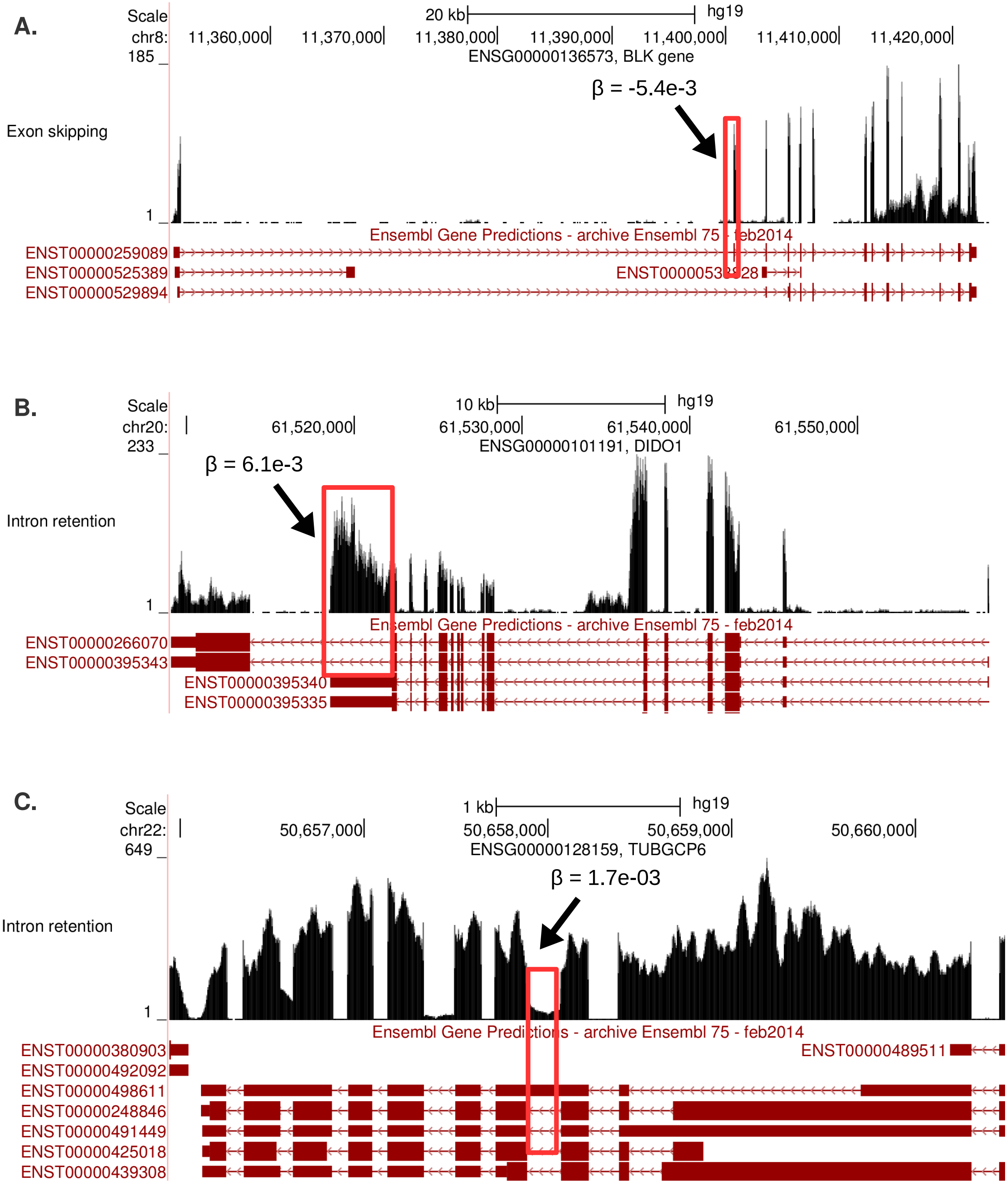


Figure 4.3: Example exons which showed changes in inclusion rate with age. **A.** Exon annotated as skipped (significant changes). **B.** Exon annotated as retained intron, however, as can be appreciated from the figure, this event should be classified as alternative 3' UTR rather than a retained intron (significant changes). **C.** Example for an intron retention event, which was non-significant after the correction for multiple testing (as no significant changes in the betas of exons annotated as retained introns – apart from the example show in the B panel of this figure – has been found). Coverage data was merged from four randomly selected samples of the age of 45. Wider boxes represent coding exons, narrower boxes represent non-coding exons or UTRs, lines represent introns.

4.4.4 Frequency of canonical splicing decreases with age

We explored the nucleotide composition of splice sites used. These splice sites were identified *de novo* as described in Section 4.3.4. Comparing our identification to known sites, we found 371,740 from 1,001,983 annotated sites in our data, which might have happened due to the fact that only a fraction of genes present in the annotation is expressed in blood. For these sites, we predicted the strand and function (donor or acceptor) correctly for the 99.775% of them. We also discovered 614,564 novel splice site pairs. The donor splice site was defined as the two most 5' nucleotides in the intron downstream from the exon and the acceptor splice site as the two most 3' nucleotides in the intron upstream from the next exon. The two splice sites originating from the same read were considered as a pair of splice sites used within the same splicing event.

The pairs of donor "GT" and acceptor "AG" sites were considered canonical, and deviations from this nucleotide pattern were counted as a pair with a non-canonical donor and canonical acceptor, canonical donor and non-canonical acceptor, non-canonical donor or non-canonical acceptor. Over 98% of the identified splice sites were canonical.

Further analysis of the *de novo* identified splice sites revealed a significant decrease ($p = 0.013$) in the number of used canonical splice site pairs (see Table 4.2). Furthermore, using the linear model described in 4.3.5, we discovered an age associated increase of non-canonical donor splice sites and the pairs of non-canonical donor and canonical acceptor ($p = 0.019$). The number of non-canonical acceptors on their own, or paired with any type of donor did not change with age (Table 4.2[6]). We calculated the number of excised U12 introns (introns spliced via the non-canonical U12 splicing mechanism, see Section 4.3.3 for more detail) and whether their splicing rate (identified as described in 4.3.5) changed with age to evaluate the potential contribution of non-canonical splicing performed by the minor spliceosome (also known as U12 splicing [193]). We did not observe any significant changes in the amount of U12 splicing ($p = 0.501$).

Table 4.2: Usage of canonical and non-canonical donor and acceptor splice sites and usage of multiple donor splice sites with one acceptor splice site and vice versa.

Type of splice site	Usage	Beta (p-value)
ND	1.95%	0.00060 (0.017)*
NA	0.58%	0.00028 (0.070)
CD...CA	97.8%	-0.00064 (0.016)*
ND...CA	1.56%	0.00036 (0.024)*
CD...NA	0.19%	0.00004 (0.429)
ND...NA	0.39%	0.00024 (0.076)
(U12_db)	0.18%	0.00008 (0.571)
D_MA	1.179%	0.000005 (0.860)
MD_A	1.180%	0.000022 (0.420)
D/A	1.001%	-0.000014 (0.032)*

We explored the nucleotide content of non-canonical splice sites (the two intronic positions, different from "GT" or "AG" and adjacent to the split). To do this, all splice sites from all samples were used to generate sequence logos. We analyzed non-canonical donors paired with canonical or non-canonical acceptors separately from non-canonical acceptors paired with canonical or non-canonical donors. We observed that the "GC" motif is the dominating motif for non-canonical sites paired with canonical

[6] Abbreviations use in the table:
"A" – acceptor splice site; "D" – donor splice site;
"CD" – canonical donor splice site; "CA" – canonical acceptor splice site;
"ND" – non-canonical donor splice site; "NA" – non-canonical acceptor splice site;
"U12_DB" – pairs of splice sites used by U12 spliceosome known from literature;
"MD_A" – average number of donors used together with one acceptor (multiple donors);
"D_MA" – average number of acceptors used together with one donor (multiple acceptors);
"D/A" – ratio between the numbers of donors and acceptors.
The third column contains age coefficients and their p-values derived from the linear regression of corresponding measures. Asterisks represent significant age-related changes

acceptors (Figure 4.4, A). Splice sites used in the youngest 25% of the samples were also used to create sequence logos. Sequence logos for the splice sites used in the oldest 25% did not show any visual differences. For non-canonical acceptors paired with canonical donors the dominating motif was "AT" (Figure 4.4, B). This suggests the second common splice site motif – "GC-AG" – being used more often with age. Non-canonical splice sites paired with each other had a nucleotide distribution close to random (Figure 4.4, C and D), which suggests that a large proportion of these sites are false positives, for example due to mapping artefacts.

To investigate the amount of novel non-canonical donor splice sites used in older individuals, we overlapped the list of all non-canonical donors used in younger samples (youngest 25% of the samples) and the list of all non-canonical donors used in older samples (oldest 25% of the samples). A splice site was considered used when at least one read was split at this position in at least 1% of the samples. Only splice sites from genes expressed in all samples were considered. We identified 382 non-canonical donor splice sites absent in younger individuals and present in the older ones. They did not cluster in any specific genes and were situated both within known genes (in the exons and in the introns) and in the intergenic regions.

4.4.5 Multiple splice sites paired with one splice site

We investigated whether the number of alternative donors or acceptors changed with age. To do this, we calculated the number of donors used with multiple acceptors and the number of acceptors paired with multiple donors. We then assessed whether the average number of donors or acceptors paired with one acceptor or donor changed with age. Both donors and acceptors were on average used with more than one other splice site (1.1803 acceptors per donor and 1.1793 donors per acceptor). We did not see age related changes in the average number of sites used with a donor or an acceptor (Supplementary Table 4.2).

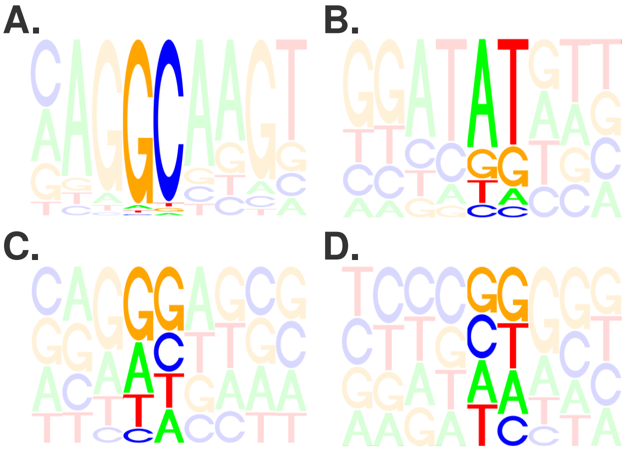


Figure 4.4: Sequence logos with the frequency of nucleotides at and around the splice sites. **A** and **C** show the last three exonic positions (blurred), the first two intronic positions (bright, donor splice site) and the following four intronic positions (blurred). **B** and **D** show the four intronic positions before the splice site (blurred), the last two intronic positions (bright, acceptor splice site) and the first two exonic positions (blurred). **A** – non-canonical donor paired with canonical acceptor splice sites. **B** – non-canonical acceptor paired with canonical donor splice sites. **C** – non-canonical donors paired with non-canonical acceptors. **D** – non-canonical acceptors paired with non-canonical donors.

We applied the linear regression model described in Section 4.3.5 individually to each splice site and investigated whether the number of its partner sites (donors for an acceptor and acceptors for a donor) changes with age. We found 27,757 and 26,412 donors with negative (less acceptors paired with the same donor in older individuals) and positive (more acceptors paired with the same donor) betas, 25,956 and 24,497 acceptors with negative and positive betas respectively (Figure 4.5, A). Only three significant changes were found, and two of them will be highlighted in this section[7].

One of the genes in which the number of alternative donors for a specific acceptor decreased with age significantly from four in younger to two in older individuals (p-value of 2.8×10^{-06} before and 0.097

[7] Note that for the third example – gene *TSPAN* encoding tetraspanin protein – novel splice sites identified in older samples were situated upstream of the gene's 5'UTR on the opposite strand. It is not related to any known transcript, therefore we do not discuss it here and focus on two other examples situated in the known genes.

after multiple testing correction) is the *LTK* gene which encodes leukocyte receptor tyrosine kinase protein (Figure 4.5, B). Two events of alternative splicing uncharacterized before – exon skipping and usage of a novel non-annotated exon – occurred only in younger samples.

An example of a gene where the number of alternative donor splice sites increased with age (p-value of 8.1×10^{-7} and 0.039 after multiple testing correction) is the *GZMH* gene encoding the granzyme H protein (Figure 4.5, C). Where only three alternative donor splice sites were used with a specific acceptor in younger individuals, six donors were used in older samples, all of them representing non-annotated alternative splicing events. Unfortunately, we could not check whether a potential protein can be produced with with any of the new splice sites, as transcript assembly (necessary for that) is not available and our analysis considers only pairs of splice sites. Predicting whether an open reading frame is maintained or an alternative open reading frame is created also becomes complicated, since the second border of the exon is not always known (almost 30% of the splice sites were identified outside the annotated genes).

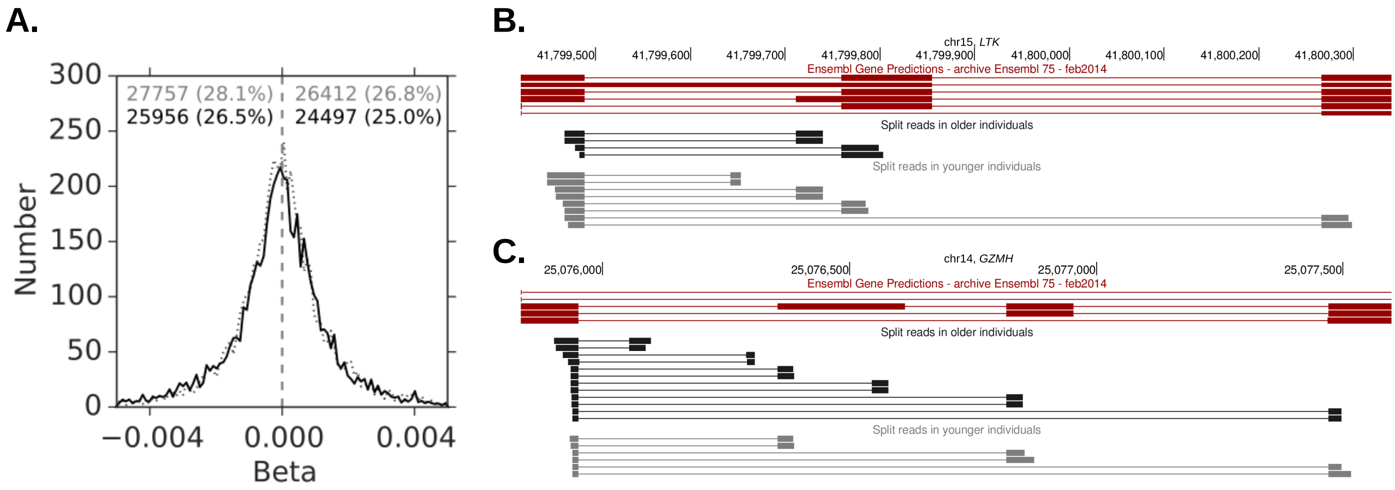


Figure 4.5: **A.** Distribution of betas for donor splice sites paired with multiple acceptors (solid line, label in black) and acceptor splice sites paired with multiple donors (dotted line, label in gray). Linear regression described in 4.3.5 was used for each splice site to assess whether the number of sites it is used with changes with age (positive beta reflects an increasing number, negative beta reflects a decreasing number of splice sites). Absolute values of beta lower than 10^{-10} were discarded, therefore the total percentage of betas signed on the plot does not add up to 100. x-axis contains betas. y-axis contains absolute numbers. **B.** An example of a gene – *LTK* – demonstrating an age dependent decrease in the number of donor splice sites paired with the same acceptor splice site. Top red panel represents the transcripts reported for *LTK* in the Ensembl v.75 database (the gene is situated on the reverse strand, so the leftmost part of the graph is the 3' end of the region). Wider boxes represent coding exons, narrower boxes represent non-coding exons or UTRs, lines represent introns. Middle black panel shows reads mapped to one acceptor and two donor alternative splice sites that were found in older individuals. Bottom gray panel shows reads mapped to one acceptor and four alternative donor splice sites that were found in younger individuals (no reads split at the two of these donor splice sites were found in older individuals). Only representative two reads are shown for each splice site on **B-C**. **C.** An example of a gene – *GZMH* – demonstrating an age dependent increase in the number of donor splice sites paired with the same acceptor splice site.

4.5 Discussion

In this study we performed a transcriptome wide analysis of age related changes in RNA processing. The study was conducted on a large cohort of samples with the widest age distribution collected so far. We statistically evaluated a number of aspects of alternative splicing while accounting for differences in the composition of the major cell types measured for our data – platelets, monocytes, eosinophiles, basophils, lymphocytes and neutrophils – as covariates. Blood contains many more cell types, which may still bias the expression and splicing data. However, given that these cell types are typically rare, we do not expect that they can explain the effects that we observed. We focused on different aspects of splicing. Even though some studies have suggested that age related changes may follow a non-linear behavior and, for example, occur from mid-life only [166, 194], we assumed a linear association with age. In line with this, most of the associations detected were only significant when including samples in the full age range of 20–80 years.

A limitation of current age related expression studies, including ours, is their cross-sectional nature and the large inter-individual variation [166, 179]. Longitudinal studies will have more power to detect molecular changes like the ones described in this paper.

Using the LL biobank as our main data source, we assessed the changes in the relative expression of individual exons and introns across samples with different age. Namely, exons annotated as "skipped exons" or "retained introns" were analyzed, and, as mentioned earlier, we found a restricted list of 15 skipped exons with decreased relative expression in older samples, two skipped exons with increased relative expression and one retained intron with increased relative expression in LL. The list of genes containing these exons was too limited to perform pathway analysis and draw statistically valid conclusions on. Genes involved in various cellular processes were identified (i.e., *ABLIM1* encoding a cytoskeletal LIM protein that binds to actin filaments via a domain that is homologous to erythrocyte dematin; *KDM5A* binding to retinoblastoma protein and regulating cell proliferation). Genes associated with various diseases were also identified (i.e., the deficiency of *AK2* in humans causes hematopoietic defects associated with sensorineural deafness [195, 196]; *ATP2A2* is an ATPase associated with Darier's disease; *MBNL1*, shown to regulate its own splicing, is implicated in Myotonic dystrophy [197]). Unfortunately we could not replicate any of these findings in the LLS dataset; only non-significant changes for the listed exons were found.

The main dataset available for replication was the "Leiden Longevity Study" biobank, also sequenced in the BIOS study. The age range of LLS samples was much more limited (between 50 and 70 years, Figure S4). To study age dependent changes in the Rotterdam Study is even more challenging, because samples from younger individuals were collected earlier than samples from older individuals, and the confounding effects of age and storage time may complicate the interpretation. The number of individuals from the CODAM cohorts was limited and did not provide enough power to detect age related changes. Unfortunately no other large-scale dataset with a similar age range as LL is available at the moment. E.g., the dataset of a similar size covers less than 40 years [172].

The increased rates of exon skipping and intron retention discovered in LL were not found in LLS, likely due to its more restricted age range. Although we observed a similar distribution of betas for skipped exons in LLS (Figure S5, A, black line) and retained introns (Figure S5, A, gray line) with more loci annotated as skipped exons having negative age coefficients (betas), the majority of changes were not significant. Only one exon had a significant negative beta with a p-value of 0.015 and was not significant in LL. We did not observe significant changes in the sum inclusion rate of skipped exons (Figure S5, B, $p = 0.83$) or retained introns (Figure S5, B, $p = 0.95$). We argue that the restricted age range of the LLS biobank does not allow to detect subtle changes in the relative expression of skipped exons and retained introns.

Exploring the increase of exon skipping and intron retention with age, we discovered that the GC content of the transcriptome increased with age (based on the sequenced reads), and that these changes may be partly attributable to alternative splicing events (based on the GC content of skipped exons and retained introns, which was calculated on the reference sequence). Even though we could see a positive correlation between the sum inclusion rate of retained introns and GC content in LLS (Figure S5, F, $\rho = 0.21$, $p = 2.9 \times 10^{-7}$; $\beta = 0.014$, $p = 5.4 \times 10^{-7}$), GC content of the samples from LLS did not increase significantly with age (Supplementary Table S2, Figure S5, D, $\rho = 0.07$, $p = 0.082$; $\beta = 0.0020$, $p = 0.68$). This might once again suggest that the age range in the LLS biobank is too narrow to capture minor changes in splicing.

Assessing used splice sites *de novo*, we showed that the fraction of non-canonical donor splice sites increases with age, which is reflected in the decreasing rate of canonical-canonical splice site pairs in both biobanks (LLS: Table S3, $p = 0.045$). More frequent usage of non-canonical splice sites, in particular donors, might point at new types of splicing which use different donor and a canonical acceptor splice site.

The number of alternative splice sites (both donors and acceptors) did not significantly change with age. However, we found several genes with increased or decreased numbers of alternative splice sites in the elderly. For one of the genes, the *LTK* gene encoding leukocyte receptor tyrosine kinase protein, additionally to the splice sites throughout the gene used in all samples, for one particular acceptor (genomic coordinates chr14:25,075,953-25,075,954) the number of alternative donor splice sites in older samples decreased from four to two. Both splice sites used only in young individuals lead to the shift of the open reading frame (the distance between the two donors and the acceptor is 778 and 154 nucleotides, respectively). Another gene – *GZMH*, encoding a protein that is expressed by cytotoxic immune effector cells, showed an opposite change in the number of splice sites used in younger vs. older individuals. In contrast to *LTK*, in *GZMH* the number of alternative donor splice sites used with age increased from three to six. Two alternative donor splice sites lead to the shift of the open reading frame for this gene, one of which is used only in older individuals, while the other is used in both older and younger samples. Possible outcomes of the usage of a new alternative splice sites and their effect on protein structure and function should, however, be studied only when the whole transcript structure is available. Discovering the number of alternative splice sites both increasing and decreasing with age might once again point at non-specific, non-targeted and a rather random nature of changes happening in splicing with age.

Overall age related changes found in alternative splicing performance are mainly non-significant at an individual event level. In other words, usage of single exons or splice sites on their own do not undergo drastic changes and the contribution of age to the overall variation in inclusion rates is rather limited. However, on a transcriptome wide scale multiple non-significant changes cumulate into significant trends. Unfortunately such subtle changes could not be captured with a narrow age range, and more datasets with wider age distribution (preferably between 20 and 80 years) are needed for the appropriate validation of our findings. We might also underestimate the frequency of some events, i.e. mis-spliced transcripts and transcripts with retained introns, as they may undergo faster degradation. To study the details of splicing in more depth, a targeted approach of extremely deep sequencing might be the best, as it would be more sensitive to very low concentrations of certain transcripts. Reads long enough to make transcriptome assembly possible would also help the age related splicing research, as this would allow to predict potential proteins and open reading frames disrupted or introduced with new splicing events.

4.6 Acknowledgements

This study makes use of data generated by the Biobank-based Integrative Omics Study. A full list of the investigators is available from http://www.bbmri.nl/en-gb/activities/rainbow-projects/bios. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL).

The authors would like to thank S.Y. Anvar, B.T. Heijmans, E.B. van den Akker, M. Beekman, A. Aartsma-Rus for useful discussions, interesting suggestions and extensive advice for the statistical analysis.

4.7 Appendix

Table S1: GC content of exons and introns in the LLS biobank.

biobank	biobank	all exons	all introns
LLS	ENSG00000004455, *AK2*	Exon skipping	-1.2×10^{-3}
LLS	ENSG00000143420, *ENSA*	Exon skipping	-6.9×10^{-4}
LLS	ENSG00000081237, *PTPRC*	Exon skipping	-1.4×10^{-3}
LLS	ENSG00000099204, *ABLIM1*	Exon skipping	-2.9×10^{-3}
LLS	ENSG00000198561, *CTNND1*	Exon skipping	-3×10^{-3}
LLS	ENSG00000073614, *KDM5A*	Exon skipping	-4×10^{-3}
LLS	ENSG00000174437, *ATP2A2*	Exon skipping	-7×10^{-3}
LLS	ENSG00000075399, *VPS9D1*	Exon skipping	-1.2×10^{-3}
LLS	ENSG00000003400, *CASP10*	Exon skipping	-2.9×10^{-3}
LLS	ENSG00000069849, *ATP1B3*	Exon skipping	-4.3×10^{-4}
LLS	ENSG00000152601, *MBNL1*	Exon skipping	-3.9×10^{-3}
LLS	ENSG00000168685, *IL7R*	Exon skipping	-8×10^{-4}
LLS	ENSG00000077809, *GTF2I*	Exon skipping	-1.5×10^{-3}
LLS	ENSG00000136573, *BLK*	Exon skipping	-5.4×10^{-3}
LLS	ENSG00000107099, *DOCK8*	Exon skipping	-1.5×10^{-4}
LLS	ENSG00000168010, *ATG16L2*	Exon skipping	3×10^{-3}
LLS	ENSG00000006114, *SYNRG*	Exon skipping	8.6×10^{-3}
LLS	ENSG00000101191, *DIDO1*	Intron retention	6.1×10^{-3}
LL	ENSG00000106554, *CHCHD3*	Exon skipping	-3.5×10^{-2}

Table S2: GC content of exons and introns in the LLS biobank.

biobank	all exons	all introns	skipped exons		retained introns	
			pos. β	neg. β	pos. β	neg. β
LLS	51.76%	48.13%	49.11%	49.69%	53.18%	54.14%

Table S3: Information about splice site usage and its changes with age in LLS biobank.

Type of splice site	Usage	Beta (*p*-value)
ND	2.04%	1.2×10^{-3} (0.045)*
NA	0.63%	6.1×10^{-4} (0.109)
CD...CA	97.77%	-1.3×10^{-3} (0.051)
ND...CA	1.60%	6.6×10^{-4} (0.077)
CD...NA	0.19%	0.5×10^{-5} (0.602)
ND...NA	0.44%	5.6×10^{-4} (0.093)
(U12_db)	0.19%	-1.9×10^{-5} (0.578)

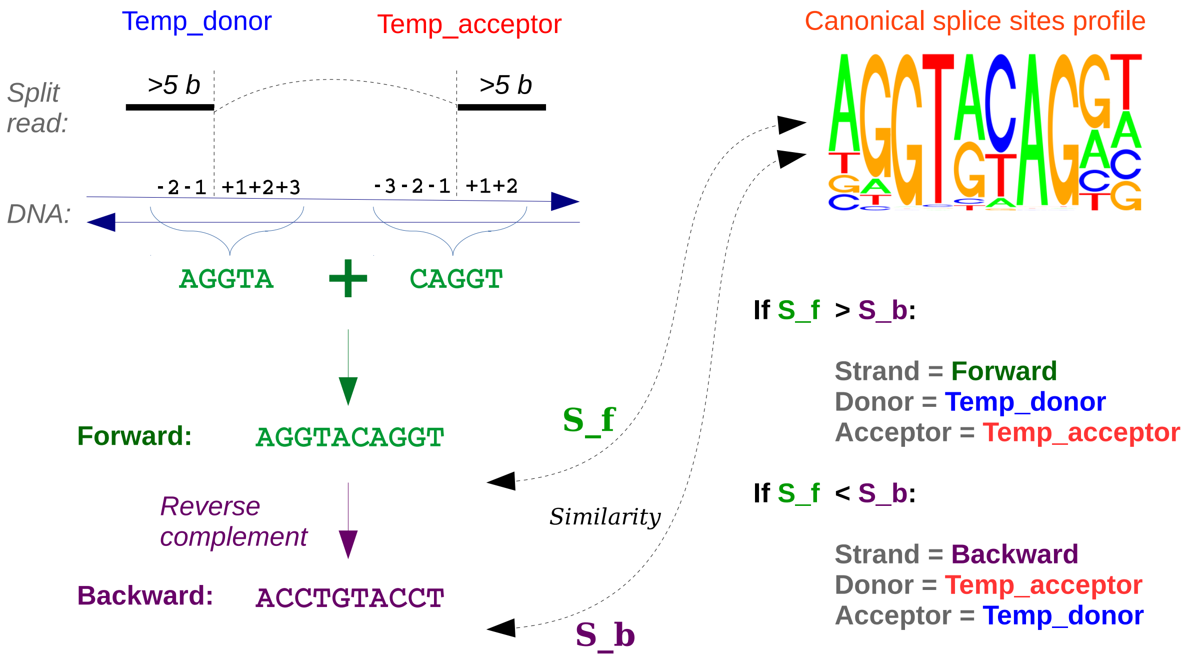


Figure S1: Method to identify putative splice sites and their strand from non-strand-specific data. Split reads are used to detect potential donor and acceptor splice sites. The nucleotides around the splice sites are extracted and compared to the canonical splice site profile known from literature. The reverse complement of the extracted nucleotides is also compared to the canonical profile. Depending on whether the surrounding of a splice site or its reverse complement is closer (based on the *Similarity*, or *edit distance*) to the canonical profile, the read is considered to be mapped on the forward or the reverse strand, respectively.

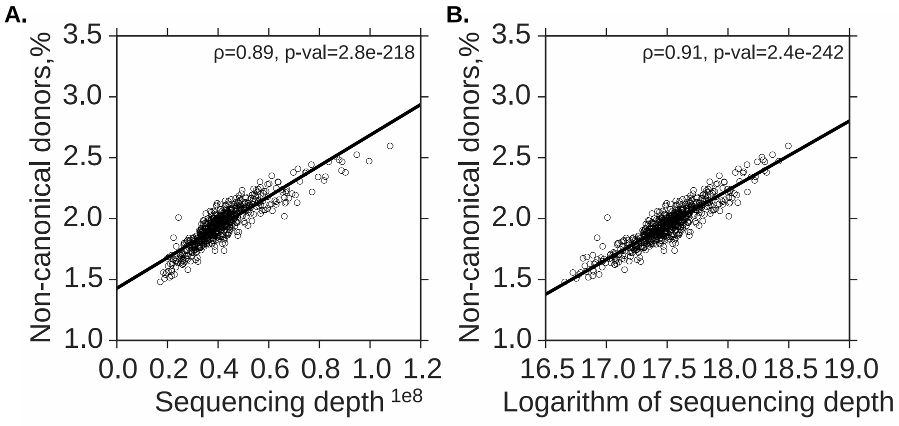


Figure S2: Correlation between the number of canonical splice sites (y-axis) and the sequencing depth (x-axis, **A**) or the natural logarithm (x-axis, **B**) of a sample.

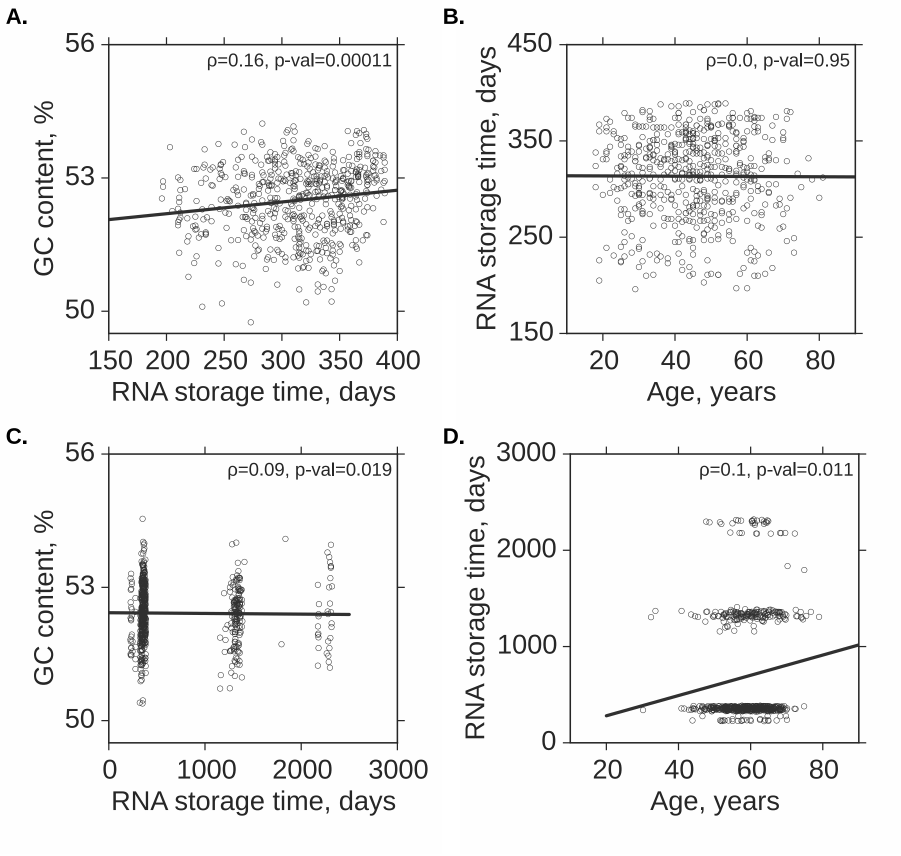


Figure S3: **A.** Scatter plot showing the correlation between the RNA storage time (number of days, x-axis) and the GC content of a sample (y-axis) in LL biobank. RNA storage time was calculated as the number of days from the day of RNA isolation to the day of sequencing. **B.** Scatter plot showing no correlation between the age (x-axis) and the RNA storage time (y-axis) in LL biobank. **C.** Scatter plot showing a subtle correlation between the RNA storage time (number of days, x-axis) and the GC content of a sample (y-axis) in LLS biobank. **D.** Scatter plot showing no correlation between the age (x-axis) and the RNA storage time (y-axis) in LLS biobank.

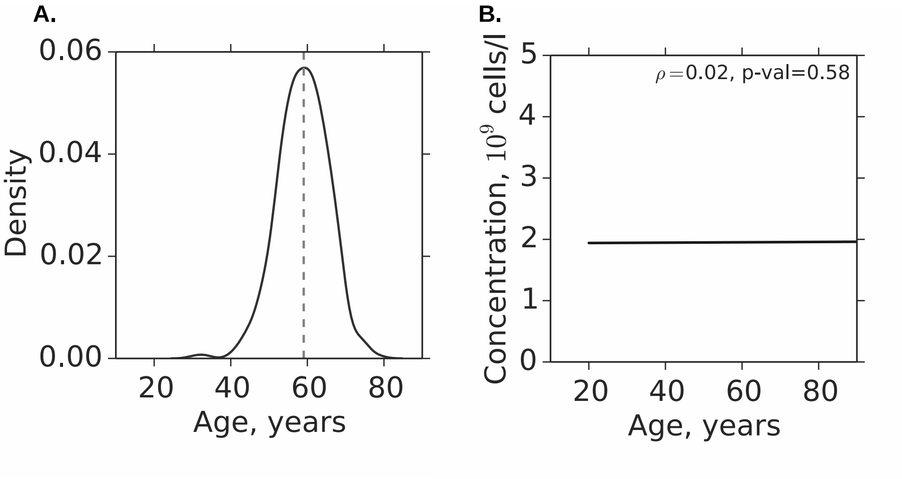


Figure S4: **A.** Density plot reflecting the age distribution within the LLS biobanks. **B.** Scatter plot reflecting age related changes in the counts of lymphocytes. Line represents the fitted linear regression lines for LL data. p-value of the linear regression is shown in the right top corner.

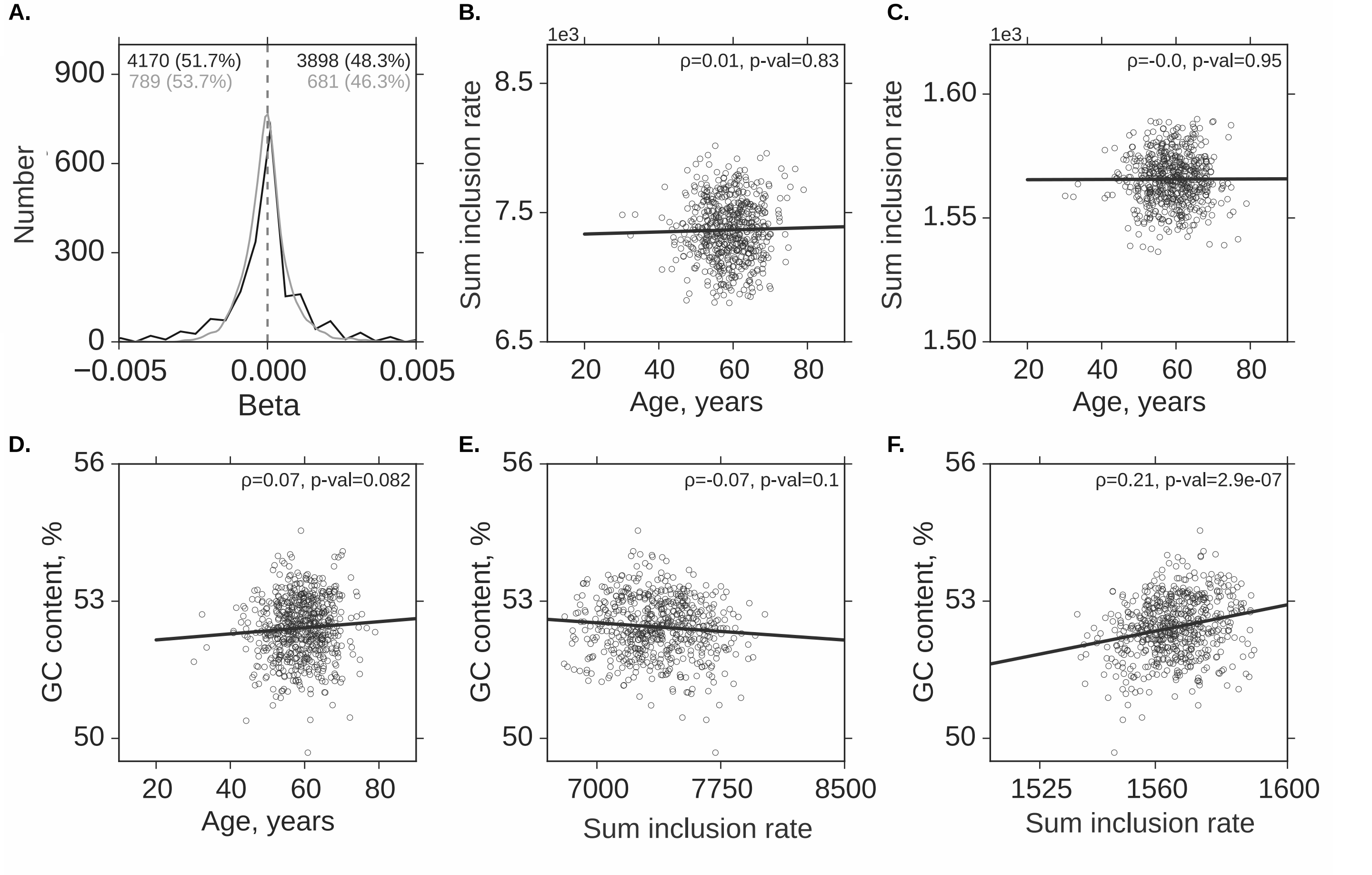


Figure S5: GC content, exon skipping and intron retention in LLS data. For more details on what is represented on the plots see the legend to Figure 4.2 from the main text.

Chapter 5

Allelic imbalance and its potential role in breast cancer pathogenesis

I. Pulyakhina[1], M.P.G. Vreeswijk[1], J.F.J. Laros[1,2], C.M. Meijers[1], P. Devilee[1], P.A.C. 't Hoen[1]

1 Dept. of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Leiden Genome Technology Center, Leiden, The Netherlands

Manuscript in preparation

5.1 Abstract

Differential allelic expression (DAE) refers to differences in expression levels between the two alleles of a gene in a cell. DAE can be caused by genetic variation and is known to affect a considerable fraction of genes in the human genome. DAE has been suggested to play an important role in human phenotypic variability, including complex traits and diseases. We hypothesize that DAE contributes to the pathogenesis of breast cancer and that changes in the DAE patterns in control versus breast tumor tissue can be associated with the disease. Deleting or decreasing the expression of the highest expressed allele of a tumor suppressor gene may have larger consequences than deletion of the lower expressed allele.

We analyzed whole transcriptome sequencing data of paired normal-tumor samples from breast cancer patients provided by the TCGA consortium. The allelic expression was examined by analyzing the coverage of the alleles of each variant. We focused on the heterozygous SNPs showing consistent changes in allelic expression in at least 5% of the patient cohort to identify common regulatory mechanisms in breast tumorigenesis.

We introduced a classification of different types of allelic expression changes, and observed that the majority of SNPs showed differential allelic expression only in tumor but not in normal tissue. We found nine SNPs with significant and consistent changes in allelic expression in normal versus tumor RNA. Evaluation of the allelic ratio in tumor DNA demonstrated that the majority of changes in allelic expression was caused by allele-specific aneuploidy. We also found one SNP situated in the *NBPF9* gene that had imbalanced allelic expression already in normal RNA, and this imbalance increased even further in tumor RNA.

The non-random SNP-behavior in the breast cancer patient cohort suggests that DAE is contributing to the mechanism of breast cancer development. DAE and its quantitative effects on gene expression might become the next milestone in understanding breast cancer pathogenesis after the discovery of LOH (loss of heterozygosity) due to aneuploidy at DNA level.

5.2 Introduction

Breast cancer is one of the most important causes of cancer-related deaths, the most common malignancy worldwide in women and the second leading cause of cancer death after lung cancer [198].

Breast tumors are characterized by genomic instability and gross chromosomal aberrations on a genome wide scale, including loss or duplication of whole chromosome arms [199]. Aneuploidy, a process when (parts of) chromosomes can be lost or duplicated, possibly more than once [200, 201], is often observed in breast cancer [202], as well as other types of cancer [203, 204, 205]. The most common consequence of genomic instability and a hallmark of cancer cells is the presence of an abnormal number of chromosomes, or *aneuploidy* [205]. Aneuploidy has been extensively studied and characterized to better understand the cell's evolutionary pathways towards carcinogenesis. Errors in replication, segregation and telomere function are generally assumed to occur in the absence of properly functioning cell cycle checkpoints [206, 207]. During tumorigenesis, DNA regions that include oncogenes are frequently amplified causing their overexpression and endowing the cell a growth advantage; tumor suppressor genes are often lost during the evolution of cancer allowing cells to escape mitotic arrest and/or cell death. Loss of tumor suppressor genes is often a prerequisite for tumorigenesis [208]. Aneuploidy does not affect every chromosome equally; some chromosomes (e.g., chromosome 17 in breast and colon cancer) are more prone to aberrations [209], giving rise to distinct patterns of aneuploidy in certain tumors [210]. An unresolved issue is why there seems to be chromosomal specificity in the pattern of allelic losses and gains in different types of cancer and also in different subtypes of breast cancer [211].

DNA rearrangements can also occur subchromosomally. These rearrangements may result in distorted number of alleles for individual (or sets of neighboring) genes [212]. A frequent phenomenon in breast cancer is loss of heterozygosity (LOH), signifying chromosome regions in the tumor DNA where either the paternal or maternal copy is missing [213, 214, 215]. Simple copy-number loss – monosomy – will therefore show as LOH, but LOH can also be copy-number neutral, when the final number of alleles equals the initial number of alleles [216]. Copy-number neutral LOH happens when one allele is lost and the other allele is gained at the same time [217]. Deviations from the expected 1:1 ratio (different from 1:0 or 0:1) of parental chromosome copies are called allelic imbalance.

Two examples of genes undergoing LOH in sporadic breast tumors are *BRCA1* and *BRCA2*. Regions of the chromosomes where *BRCA1* and *BRCA2* are located (i.e., 17q21 and 13q12, respectively) undergo allelic imbalance or loss of heterozygosity (LOH) in 30-50% of the cases. It was expected that these events would unmask somatic (i.e., acquired) mutations in these genes, but such mutations have rarely been observed [218, 219, 220, 221, 222]. A number of studies have shown that in many individuals the two parental alleles of a gene are not expressed at equal levels [223, 224]. Both *cis-* and *trans-*regulatory mechanisms are at the basis of this phenomenon, which may affect several thousand genes. Genetically, *BRCA1* and *BRCA2* behave like classical tumor suppressor genes, meaning that almost all breast tumors of gene mutation carriers show loss of the wild-type allele [225, 226, 227]. *BRCA1* and *BRCA2* have been implicated in tumorigenesis of sporadic breast cancer as haploinsufficient tumor suppressor genes through the loss of their expression [228, 229]. Mechanistically, this expression loss has been explained either by promoter hypermethylation or LOH, or a combination of both [230]. However, the proportion of tumors with decreased mRNA expression is much higher than the proportion with promotor hypermethylation, while LOH, although found in 30-50% of breast tumors at the *BRCA1* and *BRCA2* loci, reflects a number of different chromosomal mechanisms, not all of which are predicted to lead to expression loss. LOH could affect the expression level of a gene if one allele was physically lost (copy-number loss). If both alleles were expressed at equal levels, this would be predicted to lead to a 50% decrease in expression, but the effect would be stronger if the two alleles of the gene are not expressed at equal levels and the allele with the highest expression is lost. In the latter situation, even copy-number neutral LOH would affect final expression levels.

The examples of *BRCA1* and *BRCA2* discussed above are not allele specific, meaning that either allele can be gained or lost, and such loss leads to reduced and gain to increased gene and protein expression. However, another type of aneuploidy – allele specific aneuploidy, or *allelic disparity* – has also been described [231, 232]. The wild-type allele of the *APC* gene (located on 5q31, associated with hereditary colorectal cancer) was shown to be lost in 22 out of 23 analyzed colon tumors. However, the mutant allele was shown not to contain any typical, inactivating *APC* mutations, but to be expressed at much lower levels than the wild-type allele, leading to lower APC protein expression levels and predisposition to the disease [233]. Unequal expression of the two alleles of a gene, termed *Differential Allelic Expression* (DAE), has also been shown to be associated with breast cancer [234, 235, 236]. Genes as *BRCA1/2, CCND3, EMSY, GPX1, GPX4, MLH3, MTHFR, NBS1, TP53* and *TRXR2* showed distinctive DAE patterns similar among samples of EBV-transformed lymphoblastoid cells [234]. Array-based analysis of eight breast cancer patient-derived normal mammary epithelial lines revealed genes showing DAE to cluster in one major breast cancer-relevant interaction network, which includes two known cancer causative genes, *ZNF331* and *USP6*, and a breast cancer causative gene, *DMBT1* [236].

Thus, aneuploidy plays an essential role in tumorigenesis and breast cancer. Investigating DNA copy number alterations across a tumor's entire genome was a challenging task until the comparative genomic hybridization (CGH) technology was introduced [237]. Subsequently SNP arrays have been used by numerous researchers to explore chromosomal aberrations in tumor DNA, i.e., by comparing a set of healthy control tissues to a set of tumors [238, 239, 240]. With the arrival of Next Generation Sequencing (NGS) and, more specifically, RNA sequencing (RNA-Seq), it became possible to study DAE genome-wide and in genes that were not primary candidates. RNA-Seq is quantitatively more accurate than SNP arrays in determining mRNA levels, so it should be more sensitive to minor changes in expression. In addition, it facilitates the research on a whole genome level and is not limited to the probe design, as a SNP array is [241]. By comparing genome-wide sequence data (i.e., from normal and tumor RNA of the same patient, or from RNA of healthy subjects and patients with breast tumors), DAE can be investigated for all genes expressed at mRNA level [231, 232, 242, 243, 244].

Some studies have exploited DAE to search for genes in which an allele had been inactivated by protein- truncating mutation, leading to nonsense-mediated mRNA decay. *CHEK2*, a G2 checkpoint kinase 2, showed allelic expression imbalance in 10% of lymphoblastoid cell lines (LCLs) from high-risk breast cancer from whom no mutation in *BRCA1* or *BRCA2* had been identified. All samples with such DAE were carriers of the truncating mutation NM_007194.3:c.1100del [235]. Likewise, in families with pituitary adenoma, a genome-wide scan for genes with DAE in predicted gene carriers led to the discovery of *AIP1* as the culprit underlying gene for this familial cancer syndrome [245]. Individual examples, showing an impact of a certain allelic imbalance event at a patient level, have also been characterized [246]. However, none of the studies so far has attempted to compare DAE of genes with allele-specific patterns of LOH or allelic imbalance in breast tumors, certainly not in a large patient cohort.

In this study, we assessed the phenomenon of allelic imbalance at the DNA level and differential allelic expression at the RNA level. We analyzed a dataset of over 80 breast cancer patients, having data from normal DNA, normal RNA, tumor DNA and tumor RNA from the same patient. The data used in this study was generated within The Cancer Genome Atlas project. We identified coding SNPs in normal DNA to be able to distinguish between the two alleles. We used a combination of statistical tests and developed a statistical framework to assess those SNPs in normal RNA, tumor DNA and tumor RNA from the same patient. We provide a classification of different cases of allele-specific expression using the changes in DAE between normal and tumor RNA. We also address the mechanism causing such DAE patterns using normal and tumor DNA in each patient.

5.3 Materials and methods

5.3.1 TCGA dataset

The analysis was performed on the sequencing data from 89 breast cancer patients having triple-negative breast invasive carcinoma (Figure 5.1, A) available within The Cancer Genome Atlas (TCGA) project. Each patient contributed four samples: normal and tumor DNA samples, and normal and tumor RNA samples. The sample underwent whole exome sequencing. Normal samples were collected from the unaffected part of the patient's breast, while the tumor samples were collected from the breast tumors. DNA/RNA from formalin-fixed, paraffin-embedded tissues were sequenced on the Illumina HiSeq 2000 platform and mapped to the hg19/GRCh37 genome using the BWA aligner [107] for aligning DNA and MapSplice [96] for aligning RNA. The alignment was performed by the TCGA consortium[1]. Level two access data (*bam* files containing aligned reads) were requested from the NCI's Cancer Genomics Hub[2].

5.3.2 Calling and analyzing variants

The workflow of the analysis is represented in Figure 5.1 (panels B and C). Calling variants in patient DNA derived from normal breast tissue has been performed with SAMtools [39] using default parameters. Only heterozygous variants were considered[3]. For that, we selected allelic ratios in normal DNA

[1] http://tcga-data.nci.nih.gov/tcga/

[2] https://cghub.ucsc.edu/

[3] Note that when calling variants in RNA or tumor DNA, heterozygous variants with the unequal coverage of the two alleles are also considered.

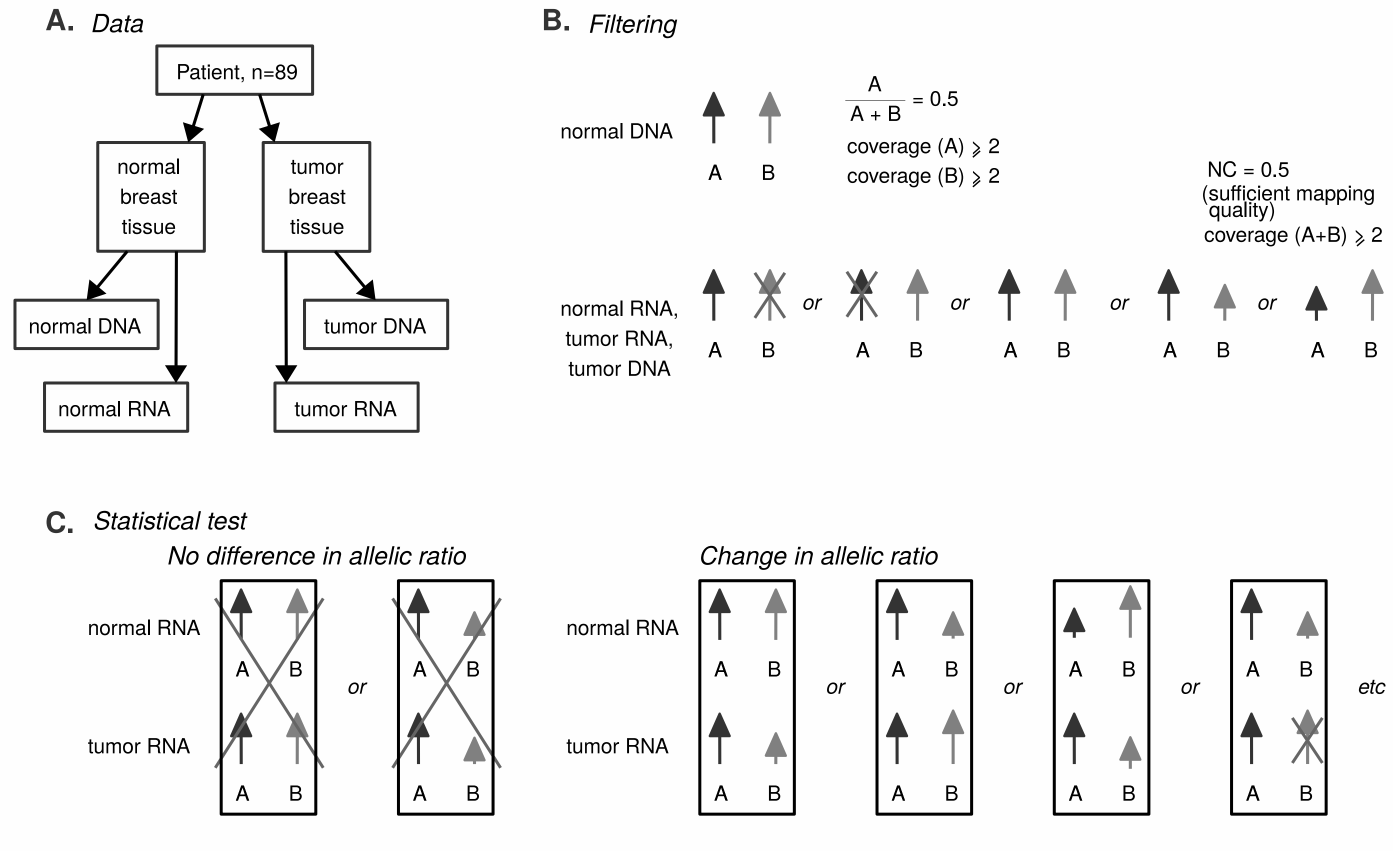


Figure 5.1: **A.** Design of the dataset, in which normal DNA and RNA and tumor DNA and RNA was collected from 89 breast cancer patients. **B.** Filtering steps used during the analysis to select reliable variants. See Section 5.4.2 for more details. **C.** Set-up of the statistical test used to select variants with a change in allelic ratio between normal vs tumor RNA. SNPs showing changes in allelic ratio in normal vs tumor RNA were selected. Note that only recurrent events were selected for the statistical framework to get statistically valid results.

which did not differ significantly from 0.5 (Figure 5.1, B). This was done to exclude variants situated in regions of potential copy number variants (which would result in an allelic ratio different from 0.5), as any change in allelic ratio of these variants in tumors versus normal tissue would be harder to interpret. Each of the alleles also had to be covered at least twice to be selected for further analysis. Calling variants on RNA derived from normal breast tissue has been performed using VarScan [248], reporting homozygous reference, heterozygous and homozygous alternative positions. Only positions where the sum of the coverage of both alleles was above 20 and which had high mapping quality were considered. They were selected using the “NC” flag, variants with “NC=1” (meaning that the position had the mapping quality high enough to be called) were further analyzed. Calling variants on tumor RNA was also performed using VarScan reporting only heterozygous and homozygous alternative variants. The same filtering criteria were used to select high quality variants.

Note that we address variants and SNPs as two separate types of events. We call “a SNP” a nucleotide change on DNA or RNA that has been seen in at least 5% of the cohort, and any other nucleotide change on DNA or RNA is referred to as “a variant”.

5.3.3 Statistical framework

We developed a statistical model to identify changes in allelic expression (the ratio between the number of reads mapped to one allele and all reads mapped to this position), which is consistent across multiple samples (Figure 5.1, C). Paired observations – e.g., allelic counts from a normal and a tumor RNA sample – are required. One SNP at a time is tested. The reference allele is referred to as “allele A”, and the alternative allele is referred to as "allele B".

A generalized linear mixed model is fitted for each SNP considering a negative binomial distribution of the allelic counts c. In the model, the fixed effect X is the status of the tissue (normal/tumor) and individuals are modeled as random effects Z:

A generalized linear mixed model is fitted for each SNP considering a negative binomial distribution of the allelic counts c. In the model, the fixed effect X is the status of the tissue (normal/tumor) and the random effect Z is the read counts (different per individual):

$$c = logN + E, \tag{5.1}$$

where c is the coverage of an allele, N is the linear predictor for the response variable and E contains random errors;

$$N = XB + ZU, \tag{5.2}$$

where X is the matrix containing two indicators – showing the tissue state (normal/tumor) and the type of allele (reference/alternative), both can be 0 or 1; Z are the random effects; B and U are the coefficients.

Since the input data is counts, a negative binomial distribution is used as it allows the variance to differ from the mean via introducing an extra parameter of over-dispersion. In the current model, several parameters are estimated, namely the coverage allele A in normal RNA (intercept), the difference in coverage of allele A between tumor and normal RNA, difference between the coverage of allele A and allele B, the interaction between the tissue status (normal/tumor) and the difference between the coverage of allele A and allele B (A 2-way anova model). Four degrees of freedom are needed to estimate these parameters. Two degrees of freedom for the fixed effects, one degree of freedom for the random effect, and one degree of freedom is needed to estimate the standard error. Variants present in at least five patients (20 complete observations: five for each allele in tumor and normal) were selected. The statistical model was implemented in R and is freely available online[4].

To select changes in the expression ratios for non-recurrent events (see Section 5.4.1 for more detail), Fisher's exact test was used (an FDR of 5% was used as a cutoff).

5.4 Results

We assessed variants showing DAE and/or a change in allelic expression in normal versus tumor RNA. At first, we categorized variants (Figure 5.2, A) based on DAE pattern and after that we identified variants present in multiple breast cancer patients and showing a consistent change in RNA allelic expression between normal and tumor tissue. Lastly, we tried to explain the patterns of allelic expression in tumor RNA using tumor DNA information from the same patients.

5.4.1 Variant categories based on DAE

Analyzing 89 breast cancer patients, we identified 78,539 variants heterozygous in at least one patient in normal DNA and present in normal RNA, tumor DNA and tumor RNA (either heterozygous or homozygous). We ran the exact Fisher's test on normal and tumor RNA on each of the variants in each patient individually (which means that the same variant in multiple patients was counted multiple times) and used the FDR cut-off of 5%. This left 5,446 variants present in at least one patient and

[4] https://git.lumc.nl/i.pulyakhina/ase_in_brc/blob/master/TCGA_analysis/

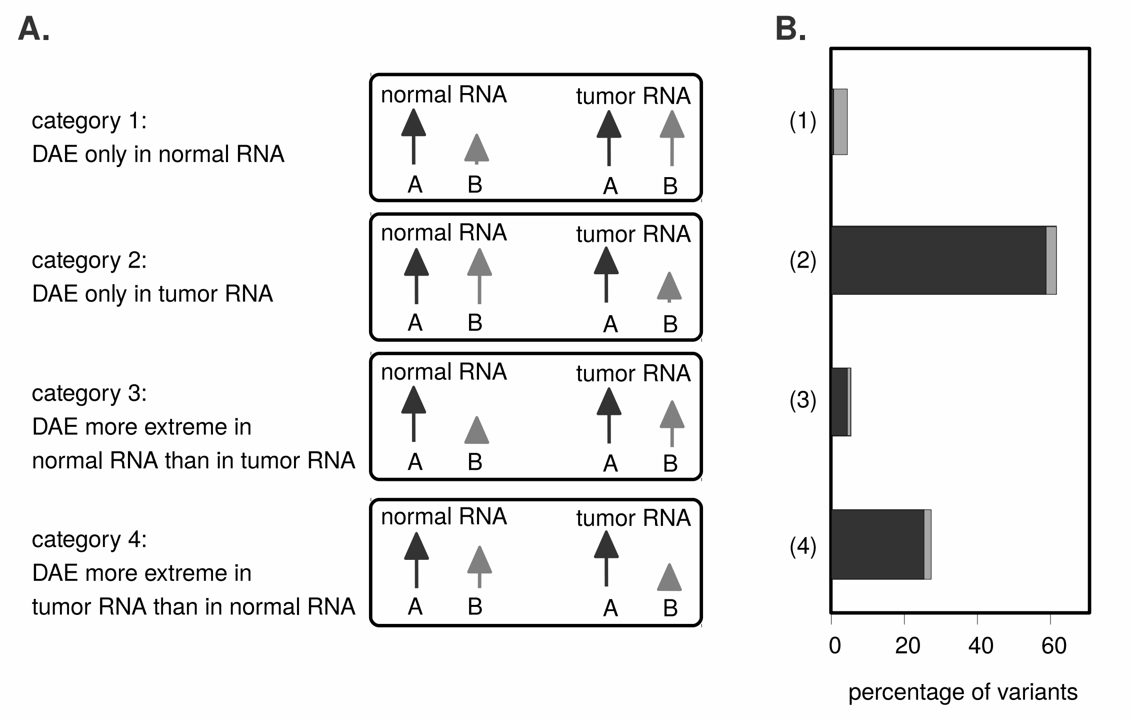


Figure 5.2: Differential DAE in normal or tumor RNA. **A.** Categories of variants based on DAE present in either normal RNA, or tumor RNA, or both. **B.** The percentage of variants present in each category for the variants showing a significant change in the allelic ratio between control and tumor RNA. Note that the fifth category – when both normal and tumor RNA shows DAE to the same extent – is discarded (see Figure 5.1, C) from the analysis and is hence not shown here. Dark grey parts of the bars represent the percentage of variants that showed allelic ratio different from 0.5 in tumor DNA.

showing a change in allelic expression ratio between normal and tumor RNA. We subdivided the variants into the following categories (Figure 5.2, A):

1. DAE only in normal RNA and not in tumor RNA (category 1);
2. DAE only in tumor RNA and not in normal RNA (cat. 2);
3. DAE in both normal and tumor RNA; the imbalance is stronger in normal RNA (cat. 3);
4. DAE in both normal and tumor RNA; the imbalance is stronger in tumor RNA (cat. 4).

Figure 5.2, B depicts the percentage of variants present in each category.

5.4.2 Changes in allelic ratio

Looking at the distribution of the variants across the categories (Figure 5.2), we can appreciate that the vast majority of variants – 89.5% – shows DAE in tumor, i.e., only in tumor – 62% (category 2) – or both in tumor and in normal RNA – 27.5% (category 4). A much smaller percentage of variants shows DAE only in normal RNA (category 1, 4.8% of variants). We can also appreciate that, even when the expression of both alleles is not equal in normal RNA, the imbalance often increases in tumor RNA (category 4) or remains the same (data not shown, as it is not further analyzed, see Figure 5.1, C for details) and rarely decreases (category 3, 5.7% of variants).

We expected aneuploidy to be the major driver of differential allelic expression in tumor RNA, and assessed tumor DNA allelic ratios using two-sided Binomial test against 0.5. When an allelic ratio was significantly different from 0.5 in tumor DNA, we considered that an allele might have been lost or gained. For category 2, 95.7% of the DAE events seen in tumor RNA could be explained by putative aneuploidy in tumor DNA; 97.6% of the variants from category 3 and 97.1% of the variants from category 4. We used category 1 as a control group and expected less allelic imbalance in tumor DNA for the variants from this category, as they did not show any significant imbalance in tumor RNA. As

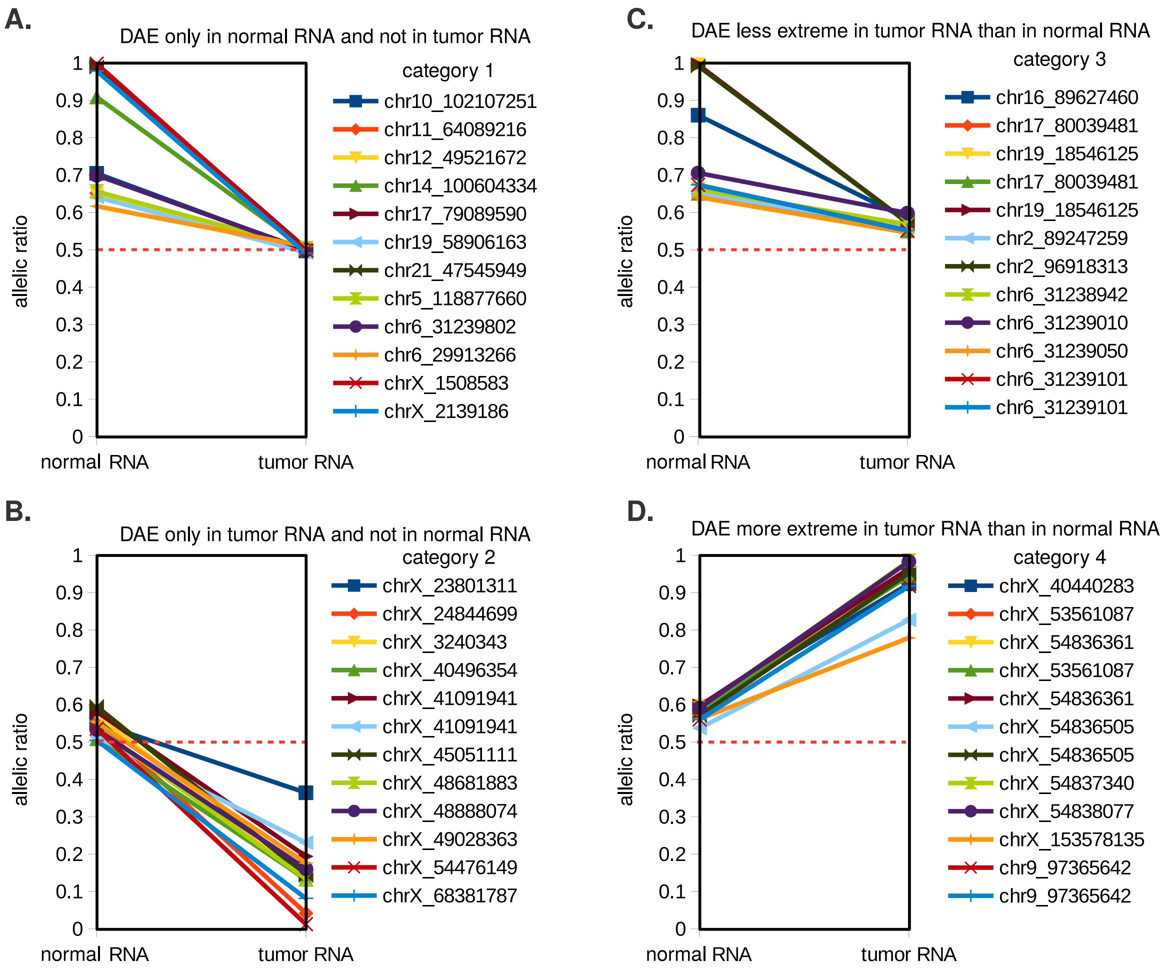


Figure 5.3: Allelic ratio (y axis) and its changes across variant categories introduced in Figure 5.2 in normal vs tumor RNA (x axis). **A.** Variants showing DAE only in normal RNA and not in tumor RNA (category 1 from Figure 5.2). **B.** Variants showing DAE only in tumor RNA and not in normal RNA (category 2 from Figure 5.2). **C.** Variants showing more extreme DAE in normal RNA than in tumor RNA (category 3 from Figure 5.2). **D.** Variants showing more extreme DAE in tumor RNA than in normal RNA (category 4 from Figure 5.2). Each dot represents a variant in one sample, same variant from tumor and normal RNA is connected with a line. Only twelve representative variants are shown per group. Note that we did not distinguish between recurrent and non-recurrent events here. RsIDs are not shown, as not all variants have rsIDs.

expected, only a small amount of variants from category 1 (14.6 %) showed allelic imbalance in tumor DNA (Figure 5.2, B).

We further inspected individual variants present in each category. In Figure 5.3, expression of different alleles (allelic expression ratios) for both normal and tumor RNA are depicted for the different previously introduced classes (Figure 5.2). For Figure 5.3, A, variants show DAE only in normal RNA, which means that the allelic expression ratio can be either <0.5 or >0.5. To make the figure easier to interpret, only variants showing DAE >0.5 in normal RNA are depicted on Figure 5.3. Note that variants showing DAE <0.5 behave similarly to the variants represented in Figure 5.3. For the category of variants where tumor RNA was more imbalanced than normal RNA (category 3 from Figure 5.2; 1,502 variants), the allele ratio was below 0.5 for 627 variants. This indicates that B allele was lost or A allele was gained in tumor compared to normal RNA. For 875 variants the allelic expression ratio was above 0.5 in tumor RNA, which indicates the opposite behavior – the B allele was gained or A allele was lost in tumor compared to normal RNA.

5.4.3 SNPs showing recurrent DAE

After categorizing variants based on DAE, we selected heterozygous SNPs present in multiple patients and consistently showing loss or gain of one of the alleles. We call such SNPs *recurrent*. These SNPs are primary candidates to play a role in the mechanisms of breast tumorigenesis, unlike individual variants which rather explain the cause of the disease in an individual patient [247].

To focus on recurrent SNPs and see how consistent they are across the dataset of 89 breast cancer patients, we selected heterozygous SNPs found in normal DNA of at least five patients (5% of the patient cohort). The statistical framework (described in Section 5.3.3) assigned low p-values to the SNPs that showed significant changes in the allelic expression ratio between tumor and normal RNA, and the change had to be consistent across all patients in which the SNP was expressed. Note that this statistical framework was designed to distinguish between the reference and the alternative allele and to find only cases of allelic disparity. We also performed the analysis to find non-allele-specific aneuploidy events, however, we did not get any significant results different from what is discussed here. From the list of 975 SNPs present in at least 5% of the patient cohort, we detected nine SNPs (Table S1) which showed a recurrent change in allelic expression ratio in tumor versus normal RNA. Figure 5.4 (panel A) contains allelic expression ratios in all patients containing the SNP, and the majority of allelic expression ratios is concentrated around 0.5 in normal RNA and elsewhere in tumor RNA. More information about these nine SNPs can be found in Table S1.

5.4.4 Mechanism behind DAE in tumor RNA

After we identified differences in allelic expression between normal and tumor RNA, we assessed tumor DNA to find an explanation for different allelic expression patterns.

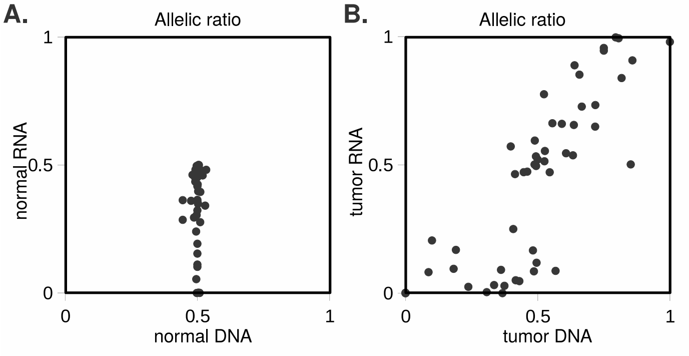


Figure 5.4: **A.** Scatter plot showing the distribution of allelic ratios for the nine SNPs showing significant changes of allelic ratio in normal vs tumor RNA. x axis shows the allelic ratio in normal DNA. y axis show the allelic ratio in normal RNA. **B.** Correlation between the allelic ratio in tumor DNA (x axis) and tumor RNA (y axis). Allelic ratio on plots A and B is calculated as the expression of the alternative allele divided by the total coverage of the allele. Each dot represents one sample, meaning that the same SNP coming from different patients is present multiple times. See Figure 5.5 for the examples of individual SNPs.

We discovered high correlation between tumor DNA and tumor RNA for six out of nine SNPs (average Pearson's correlation coefficient >0.8). For the same six SNPs, allelic ratio in tumor DNA was significantly different from 0.5 (between 0.67 and 0.98). Both observations indicate the presence of aneuploidy in tumor DNA which is likely to explain the allelic expression observed in tumor RNA. Since allelic expression ratio is calculated as the ratio between the B allele (alternative allele) and the total expression of the variant (Figure 5.4, B), allelic expression ratio equal or higher than 0.5 means that the alternative B allele was gained (or the reference A allele was lost) in all patients. The events displayed in Figure 5.5 are examples of allele-specific aneuploidy, previously introduced as allelic disparity – when one specific allele is consistently gained or lost in all the patients. We observed that the same allele that was lower expressed in normal RNA was gained in tumor DNA and became higher expressed in tumor RNA (Figure 5.5, A, three panels show the data for an example SNP, rs946837, *TMTC4*). We also observed one SNP which showed the opposite pattern – the allele that was higher expressed in normal RNA was lost in tumor DNA and became lowew expressed allele in tumor RNA (data not shown, as it resembles Figure 5.5, A, but with a decreasing slope).

We observed a different behavior of allelic expression for the other three SNPs that showed different allelic expression in tumor RNA versus normal RNA (Figure 5.5, B, three panels show the data for an example SNP, rs2798893, *NBPF9*). All three SNPs had coverage above nine in all samples, and for the vast majority of patients the coverage in each sample was above 20. However, at the RNA level we could see that the allele that was lower expressed in normal RNA lost even more expression in tumor RNA (and the allelic expression ratio went down) (for rs7185949, *USP10* and rs2798893, *NBPF9*). A slight (but consistent) increase of the lower expressed allele was observed for rs617073 situated in the *U2AF2* gene.

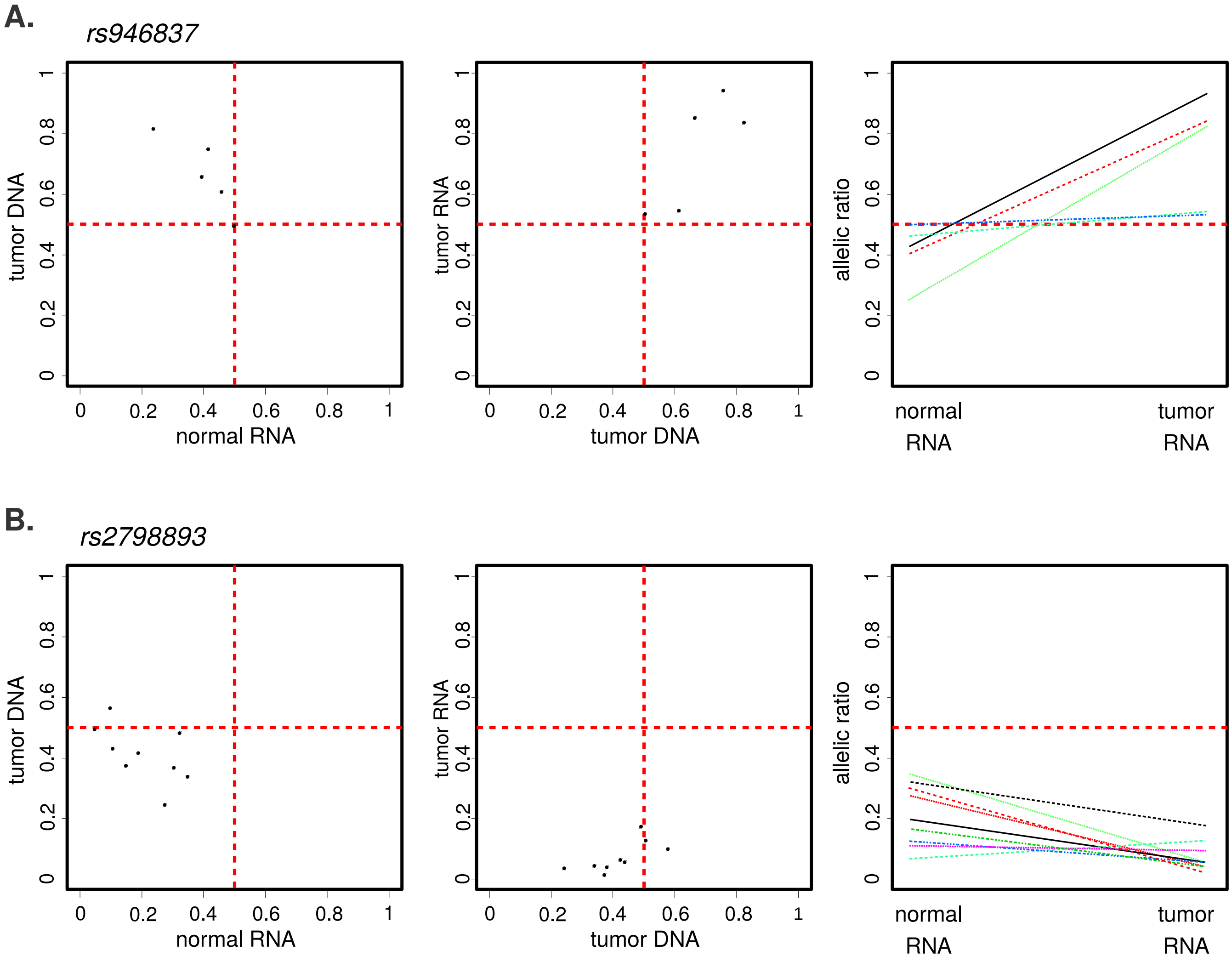


Figure 5.5: **A.** Example SNP representing SNPs with potential allelic disparity. **B.** Example SNP representing the rest of identified SNPs. Left panel – correlation between the allelic ratio in normal RNA (x axis) and tumor DNA (y axis). Middle panel – correlation between the allelic ratio in tumor DNA (x axis) and tumor RNA (y axis). **C.** Changes in the B allele frequency in each pair of samples in normal vs tumor RNA. Each dot represents a variant in one sample, same variant from tumor and normal RNA is connected with a line.

5.4.5 Allelic imbalance in normal RNA and its behavior in tumor RNA

The previous paragraph highlighted SNPs (present in at least five individuals) that showed consistent changes in allelic ratio of RNA in tumor versus normal tissue state. The majority of variants that we identified showed no DAE in normal RNA, meaning that both alleles were equally expressed. Here, we focused on the variants present in at least 5% of the samples, that already have DAE in normal

RNA and show a change in the allelic expression ratio in tumor RNA. We found almost 50% of variants (12,011 of 29,140 variants) imbalanced in normal RNA to be detected only in one sample (Figure 5.5, A and B). Only 15% of the variants (4,348 variants) were found in at least 5 samples with DAE in normal RNA. We assessed whether any recurrent change in the allelic ratio between normal and tumor RNA within these 4,348 variants were found.

Out of these 4,349 SNPs, we discovered only one SNP – already mentioned above, rs2798893 in the *NBPF9* gene – imbalanced in normal RNA to be present in at least 5% of breast cancer patients, which showed a significant change in DAE in tumor vs normal RNA. This SNP was also identified previously without pre-selecting SNPs with DAE in normal RNA before running the statistical framework. Analyzing tumor DNA and RNA profiles, we did not see any correlation between the allelic imbalance in tumor DNA and allelic expression ratio in tumor RNA ($\rho = 0.45$). We observed allelic expression patterns depicted on Figure 5.5 (panel B) – frequencies of the lower expressed allele in tumor DNA were mainly concentrated in the window of [0.4, 0.6]. The majority of nine samples heterozygous at this SNP showed a consistent decrease of the expression of the allele that was lower expressed in normal RNA.

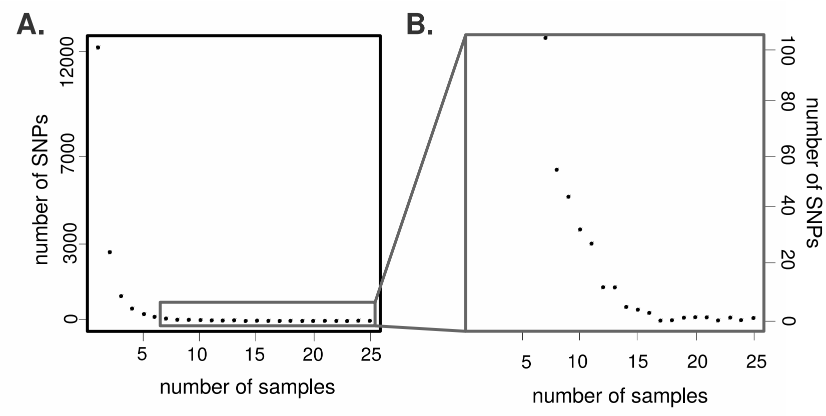


Figure 5.6: Distribution of variants showing DAE in normal RNA across the 89 samples of breast cancer patients. **A.** Number of samples (x axis, i.e., "1" on the x axis means that the variant has to been seen only once and not more) containing a certain number of variants (y axis). **B.** Zoomed-in version of A, y axis is restricted by 100 variants.

5.5 Discussion

Aneuploidy, LOH and their role in breast cancer have extensively been studied over the past decades. Resulting in the unequal expression of the two alleles, aneuploidy was shown to lead to reduced gene expression [230]. Genome-wide research of aneuploidy in breast cancer was conducted mainly using CGH arrays and exome- sequencing with Next Generation Sequencing (NGS) technology and resulted in numerous regions showing genomic instability in multiple or individual breast cancer patients. Earlier considered not to be allele-specific, meaning that either of the two alleles can be gained (or lost) with the same probability, aneuploidy was recently shown to occur in allele-specific manner [220, 242, 243]. One would expect that loss or gain of an allele in tumor DNA would have a direct influence on the gene expression in tumor RNA, which indeed has been shown [220]. However, unequal RNA expression of alleles, or Differential Allelic Expression (DAE), present already in normal RNA – and its changes in breast cancer tumor RNA – has not been extensively studied yet.

In this study we performed a thorough examination of a breast cancer patient cohort provided by The Cancer Genome Atlas project. Leaving the somatic mutations behind, we explored SNPs present in normal DNA and RNA of each patient individually and examined the behavior of these SNPs in the tumor DNA and RNA of the same patient. We observed that the majority (over 50%, data not shown) of variants occurred in normal and tumor DNA and RNA of only one patient and showed equal expression of both alleles in normal RNA and DAE in tumor RNA. We also found variants (33.2%) that showed DAE in both normal and tumor RNA, 27.5% of detected variants showed a stronger allelic imbalance (further away from 0.5) in tumor RNA and only 5.7% of detected variants showed a stronger imbalance in normal RNA.

We assumed that the majority of the SNPs which showed DAE in tumor RNA can partly be explained by aneuploidy, which we tested using tumor DNA of the same patients. We could indeed see that over 95% of variants which showed DAE in tumor RNA also showed allelic instability (allelic ratio different from 0.5) in tumor DNA.

Despite a lot of variation between the patients and the presence of patient-specific variants, we were able to identify 975 heterozygous SNPs present in more than 5% of the patients. We discovered that nine of these SNPs showed a significant and consistent change in allelic expression between normal and tumor RNA. For six out of these nine SNPs, showing allelic expression ratio significantly different from 0.5 in the majority of patients and high correlation between tumor DNA and tumor RNA (average Pearson's correlation coefficient >0.8), suggested that aneuploidy could be one of the potential contributors to DAE in tumor RNA. Moreover, from the values of allelic ratios in tumor DNA and allelic expression in tumor RNA of these SNPs, we derive that one specific allele was gained or lost, which suggests that these SNPs are probable cases of allele-specific aneuploidy (or allelic disparity). Such behavior – the presence of DAE and the absence of aneuploidy – might for instance be explained by allele-specific binding of transcription factor proteins (that would explain DAE in normal RNA), for which the expression is altered in breast tumors (that would explain changes in DAE in tumor versus normal RNA). Further investigation is necessary in order to explain the mechanism driving allelic disparity.

Discovering a very limited number of recurrent events is an expected outcome, as is has already been shown for a number of different cancer types including breast cancer that, i.e., for somatic mutations only a few altered genes are recurrent in at least 10% of the patients, while the vast majority of changes happens in less than 5% of the samples [249, 250]. Another study which focused on the analysis of DAE in a limited set of just eight breast cancer patients was only able to find three genes – *ZNF331, USP6* and *DMBT1* to show a change in DAE in tumor vs normal cDNA. Our analysis assessed individual SNPs (which we used as the markers for different alleles), as we hypothesized that specific gene alterations recurrent in the patient cohort might have a contribution to breast cancer pathogenesis. As the number of identified recurrent SNPs is limited, we suggest that we should rather look to the patterns observed for larger chromosomal regions, as certain parts of chromosomes are expected to show similar aneuploidy pattern. However, a recent study [251] suggests that, as recurrent events are rare due to the nature of cancer, a more meaningful analysis should be conducted on a level higher than gene or locus, which can be done via phasing the genotypes. A pathway or network analysis can help us to effectively uncover biological systems perturbed in tumor cells which is unfortunately still a very challenging analysis and has to be developed and studied more extensively.

Our understanding of aneuploidy and its contribution to breast cancer pathogenesis, as well as LOH and DAE, is still not full despite the decades of analyses on various levels. However, the increasing power of bioinformatic analysis transcriptome- and genome-wide and a potential of extensive research on large patient cohorts make a great promise for understanding and explaining breast cancer development and pathogenesis.

5.6 Acknowledgements

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov.

We would like to thank E.J. de Meijer and H.P.J. Buermans for their valuable advice and useful discussions, A. Morra and L. Eijsackers for the development of the statistical framework, M. van Iterson for his help with using the BiocParallel R package and M.P. Villerius for his help with running the analysis on the cluster. We also would like to acknowledge the TCGA consortium for providing us with the data of the breast cancer cohort.

5.7 Appendix

Table S1: Description of nine identified SNPs described in the main text.

Chrom.	Position	rsID	Gene name	ρ	Gene description
chr1	144,931,392	rs2798893	NBPF9	0.45	neuroblastoma breakpoint family member 9, higher expressed in breast, ovarian and pancreatic cancer
chr1	201,112,981	rs8158	TMEM9	0.84	a positive regulator of neurite outgrowth and its overexpression was suggested to play primary role in prostate carcinogenesis
chr9	96,238,578	rs10821135	FAM120A	0.98	a constitutive coactivator of PPAR-gamma-like protein 1, which has a higher expression in numerous cancers (including breast cancer)
chr12	9,232,268	rs669	A2M	0.72	alpha-2-macroglobulin which was suggested to play a role in inhibiting brain tumors
chr13	101,287,340	rs946837	TMTC4	0.67	transmembrane and tetratricopeptide repeat containing 4 known to be expressed in breast (mammary gland) tumor
chr16	68,857,441	rs1801552	CDH1	0.95	cadherin-1 protein, for which the increased risk of having breast cancer has been shown in case of genetic variations
chr16	711,905	rs2301426	WDR90	0.69	WD repeat-containing protein 90, higher expressed in about 45% of known cancer types (excluding breast cancer)
chr16	84,779,248	rs7185949	USP10	0.21	tumor-associated marker in gastric carcinoma
chr19	56,180,968	rs617073	U2AF2	0.96	U2 small nuclear RNA auxiliary factor 2, has not been associated with any type of cancer before

Chapter 6

General discussion and future developments

As can be appreciated from this thesis, RNA sequencing is used for a variety of applications [252, 253, 254]. Bioinformatic analysis of mRNA sequencing data can reveal the effect of different treatments via the analysis of differential gene or transcript expression (see **Chapter 3** for the examples of mRNA analyses in the context of aging). It can be used to explore alternative splicing, perform functional studies via pathway analyses, detect allele specific expression (see **Chapter 5** for the examples of ASE in context of breast cancer) [255, 256, 257, 258]. Nuclear RNA data analysis makes it possible to study intermediate splicing events and explore intermediate products and their impact on the splicing mechanism (see **Chapter 2** and **Chapter 3** for more details).

However, regardless of the progress in data analyses, one should not forget that the analysis is performed on *biological* systems. The analysis will produce artificial results when the sequencing experiment does not adequately capture the required data and the technical procedure to obtain the data introduces biases [259, 260, 261]. One solution is to improve the analyses, however, a more solid solution is to address and eliminate the steps introducing the biases. Note that technical artifacts will always be present in any dataset, however, it is important to recognize the biases and correct for them, which becomes easier with less biases from other sources.

6.1 Direct RNA sequencing

For RNA sequencing, the main source of biases comes from the cDNA synthesis step. The majority of RNA-Seq techniques does not sequence RNA directly. A possible solution to avoid technical artifacts which are incorporated during the cDNA synthesis step is to skip this step [266] – such techniques are called *direct RNA sequencing*, or *DRS*. Researchers started developing DRS after failing to elucidate numerous biases that cDNA synthesis introduces (see the **Introduction** for more detail).

6.1.1 Possibilities of direct RNA sequencing

The first DRS technique that became available was Helicos True Single Molecule Sequencing, or *tSMS*. Helicos tSMS [263, 264] works via the direct hybridization of RNA to the surface of an ultra clean glass slide containing poly(dT) oligonucleotides covalently attached at their 5' ends [265]. During sequencing, the terminating nucleotides with a fluorescent label are incorporated one per cycle. The terminator prevents incorporating multiple nucleotides in one cycle and the fluorescent label (specific per nucleotide) is used to identify the incorporated base [266]. Helicos tSMS single-end short reads are 30-35 nucleotides long with a relatively high sequencing error rate (the frequency of substitutions is

0.2%, insertions – 1.5%, deletions – 3.0%), which makes it hard to analyze such reads. Another common technical issue that biases the data and complicates the bioinformatic analysis is the "dark" nucleotides. A fraction of nucleotides remains unlabelled and therefore is not detected upon the incorporation. Such dark nucleotides will appear as deletions in the sequenced reads. The problem of unlabeled nucleotides is common for other sequencing techniques, but we only detect them with single molecule sequencers.

Bioinformatic analysis of the Helicos data revealed that, even though DRS has great potential and opens a wide range of opportunities, a new technique comes with new challenges. Helicos tSMS was the first NGS technique to sequence single molecules without the necessary PCR step, which was the main advantage of the technique. Another advantage was that very short sequences could be used as the input material, which would be beneficial when dealing with fragmented DNA, i.e., ancient DNA which is often partly degraded. The main difficulty for the analysis in Helicos tSMS technology is raised by the presence of dark nucleotides. Bioinformatic pipelines that were designed, tested and adjusted on the Illumina data cannot be directly applied on Helicos data, and new tools for the analysis of DRS need to be developed.

6.1.2 Future of direct RNA sequencing

Along with Helicos tSMS, more techniques are currently being developed and are expected to be available soon. One of them – DRS by Oxford Nanopore – will be discussed.

Oxford Nanopore[1] sequencing is a new technology for nucleic acid sequencing [267]. The technique uses a protein nanopore incorporated into a polymer membrane (Figure 6.1). The membrane has a very high electronic resistance and a potential is applied across it. A DNA molecule is sequenced as it passes through the nanopore. Every nucleotide passing through the nanopore gives a disruption of the membrane potential, which can be measured and associated with a base [268]. In addition to DNA sequencing, Oxford Nanopore is currently developing direct RNA sequencing and adapting the nanopore to distinguish RNA nucleotides. Protein sequencing using nanopores is also potentially possible, as proteins have been shown to move through nanopores [269, 270, 271]. The major challenge of unfolding a protein – denaturating the tertiary protein structure and making the amino acids pass through the nanopore – has recently been successfully addressed [272], which promises a wide potential of protein sequencing using nanopores.

As mentioned before, together with the opportunities and expanded applications of direct RNA sequencing come new technical and bioinformatic challenges. One of the technical challenges can be RNA secondary structures. When researchers are interested in full length RNA sequencing, regions of RNA forming semi-stable structures might be more

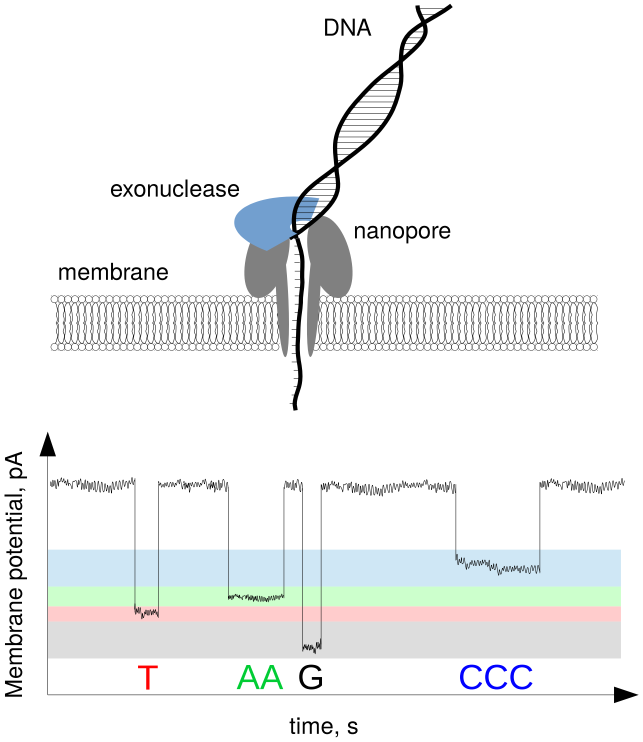


Figure 6.1: Main principles of the Oxford Nanopore sequencing. An exonuclease attached to the protein *nanopore* situated in a hydrophobic lipid layer (*membrane*) cleaves nucleotides on one of the DNA strands (left panel of the figure). The second, untouched DNA strand passes through the nanopore. Every time a new nucleotide going through causes a disruption of membrane potential (right panel of the figure). Based on the intensity and the duration of the potential the number and sequence of passing nucleotides are identified.

[1] https://nanoporetech.com/

challenging to sequence and will potentially have lower coverage. Another challenge comes from RNA modifications (see Section 6.2 for more detail), as we do not know how chemically modified nucleotides will contribute to the process of direct RNA sequencing. For example, using the Oxford Nanopore system, incorporating modified nucleotides will result in a different membrane potential. However, it might also come as an extra complication, as not all types of RNA editing are known at the moment, and not all changes in the membrane potential can be discriminated. This can result in misclassified nucleotides due to unknown changes in the membrane potential. Therefore, current pipelines and tools can rarely be applied directly on the sequencing data generated with novel sequencing techniques.

6.2 Contribution of RNA editing

When RNA-Seq data is generated with as few technical biases as possible, the analyses results can still contain misinterpreted observations due to biological mechanisms that are not well studied or known at all. One of such mechanisms that has a direct influence on the interpretation of certain RNA-Seq analyses and which has been underestimated for a long time is the phenomenon of *RNA editing*. RNA editing is a chemical modification of mature RNA (not necessarily messenger RNA). A number of chemical modifications [273] has been discovered, although the function of the majority of these modifications remains unknown. Some types of RNA editing are introduced enzymatically, while others may occur due to chemical instability, damage or free radical mediated adduct formation [81]. RNA editing events, or *REEs* happen with different frequencies, the most frequent and well studied being editing by deamination and methylation.

6.2.1 Types and potential function of editing

One of the most frequent RNA editing events – adenosine deamination – is performed by *ADAR*, *A*denosine *D*eaminases *A*cting on *R*NA. As the result of this hydrolytic deamination, inosine is created in place of adenine. Inosine is recognized as guanine by the cellular machinery and will be represented as **G** in the output of Illumina sequencing (as the DNA polymerases used in the PCR reactions prior to sequencing will introduce a **C** at all positions complementary to **I**) [82]. **A**-to-**I** editing is known to affect a large amount[2] of adenines [3] and effects mainly (but not only) double-stranded RNA. Since structural RNAs tend to form double-stranded structures, they are extensively undergoing **A**-to-**I** editing. This type of editing can be specific or promiscuous; in some cases a particular position is edited in multiple molecules and in other cases different positions within a certain region are edited in different RNA molecules [274]. Studies on RNA editing in cancer [275, 276, 277, 278] point towards the instability of RNA editing (meaning that the same position will not always be edited in all the transcripts) and go hand in hand with the potential importance of RNA editing.

The second most frequent type of RNA editing event, detected in up to 0.5% of isolated RNA molecules, is RNA methylation [279, 280] performed by the methyltransferase enzyme. Methyltransferases bind to RNA regions with a certain sequence and methylate adenosines [281, 282, 283]. RNA methylation is enriched near stop codons, in 3' UTRs and long internal exons, and methylation was discovered to happen during or after splicing [284].

RNA editing and its functions have not been studied extensively yet. Position-specific RNA editing can be used by cells as a mechanism to recode genomic information and increase functional protein diversity, as editing can take place in exons and alter the coding sequence [285, 286]. Promiscuous RNA editing is considered to play a regulatory function, i.e., occur in miRNA targets or protein binding sites, as it is also known to often take place outside the coding region [287]. Editing happening immediately

[2] Note that the frequency of **A**-to-**I** editing can be low, even too low to be distinguished from sequencing errors, therefore estimating the amount of edited positions is challenging.

after transcription prior to splicing might have an effect on splicing progress, as edited positions can be located around splice sites in the regions that the spliceosomal proteins bind to.

6.2.2 Bioinformatic analysis of RNA editing

RNA editing is a fundamental biological process which deserves thorough exploration on its own. However, as can be appreciated from **Chapter 5**, it is also crucial to detect REEs because they can interfere with RNA-Seq bioinformatic analyses. For instance, **A**-to-**I** REEs can be misinterpreted as **A**-to-**G** genomic variants in samples with unknown genotypes. Since editing does not happen in 100% of the transcripts, it might cause false-positives in the detection of allele-specific expression. There is currently a scarceness of bioinformatics methods to reliably detect REEs [288, 289]. At the moment, all RNA positions different from the reference genome and not present in the list of genomic variants are often considered REEs or sequencing errors and are discarded from the downstream analysis [290]. It is a considerable challenge to differentiate between REEs and sequencing errors. As mentioned before, it is performed by the editing enzymes that introduce the chemical modification at a single RNA molecule level, therefore the frequency of RNA editing varies widely and is mainly very low [278, 291]. There is a strong need in new bioinformatic pipelines to reliably distinguish between genomic variants and REEs [292].

6.3 RNA and proteins

Bioinformatic analysis of mRNA sequencing data is challenged by both technical and biological issues, but when the analyses results are clean and reliable, the next step is to study the consequences of these results at a protein level. Various events that are detected in the RNA-Seq data may have an effect on the sequence and thereby the structure, stability, expression level and even function of proteins. Note that this has limited effect in the experiments, when two samples with similar protein profiles are compared, and has a greater effect on the experiments when no controls or references are used (i.e., one sample or sample group).

6.3.1 RNA and protein expression

Integrative -omics analysis is a class of methods that link RNA and protein expression [293]. RNA-Seq and mass spectrometry data is combined to identify expressed genes and proteins/peptides and to get a reliable list of expressed proteins (Figure 6.2). This is a powerful approach, since not everything that happens with RNA will be translated and will lead to a functioning protein [294, 295, 296], in other words, *RNA ≠ protein*. Not all RNAs will be translated, nor when present, with the same efficiency. It has been shown that the correlation between protein and gene expression is not high, which indicates either biases in the process of sequencing and mass spectrometry [297, 298], or the bioinformatic analysis [299, 300], or the combination of both. One of the main pitfalls in the integrative -omics is that the protein expression is correlated with the *gene* expression, while the real RNA molecule that leads to a protein and should be correlated with protein expression is a transcript [294]. Since more than one transcript is produced from the majority of human genes, and proteins have different stability, correlating gene and protein expression becomes artificial [301, 302].

A logical step to correct for the biases would be to use transcript expression instead of gene expression. Gene expression is measured as the cumulative expression of all transcripts produced from it and expressed in the cell. However, different proteins can be translated from different transcripts of the same gene. Unfortunately, the latter is more difficult to analyze, as to measure transcript expression actual transcripts need to be identified. With currently used Illumina sequencing as one of the main sources of the RNA-Seq data and its relatively short reads of hundreds of bases, reliable detection of full-length transcripts is challenging and often impossible [303]. This task becomes harder when the difference between the transcripts is minor, i.e., in such cases as allele specific gene expression [304]. The structures derived from the two chromosomes may differ due to allele-specific alternative splicing which can be tagged only with one heterozygous SNP [305]. Due to ASE, a certain allele coding a deleterious or a toxic protein (unlike the other allele) might be higher expressed, which might lead to the overexpression of a deleterious protein [306]. However, it is very complex to study its effect on protein level. When two transcripts differ in only one nucleotide, current limitations of Illumina read length (hundreds of nucleotides) will fail to distinguish between the two transcripts on a whole transcript scale. Using sequencing techniques producing longer reads, such as Pacific Biosciences or Oxford Nanopore, might become a pivotal point in the integrative -omics analysis. However, even transcript and protein expression does not always show a good correlation. E.g., as mentioned above, some RNAs can be translated more efficiently than the others. A current study [167] shows a decoupling between transcript and protein expression occurring with age.

6.3.2 RNA and protein structure

Whereas tools for linking RNA and protein expression are developing quite fast, the link between RNA and protein structure remains understudied. Genetic variants and REEs can cause non-synonymous mutations resulting in altered protein sequence, structure and function. A mutation in the human dystrophin gene in the locus Xp21 of the X chromosome can lead to the development of a muscular disorder called Duchenne Muscular Dystrophy. The dystrophin protein (encoded by the dystrophin gene) is responsible for connecting the cytoskeleton of the muscle fiber to the extracellular matrix. Known out-of-frame mutations result in a transcript from which it is not possible to produce dystrophin. This leads to the loss of the dystrophin's function and the development of a muscular dystrophy [307, 308, 309]. A mutation in *LMNA* gene – a gene encoding lamin A, a protein providing structural support to the nucleus – creates a 5' cryptic splice site within exon 11 resulting in an abnormally short mature mRNA transcript [310]. This transcript yields an abnormal isoform of prelamin A – precursor of lamin A – which is not able to provide the necessary support to the nuclear lamina, which leads to reduced cell division and premature aging (also known as progeria, or Hutchinson-Gilford disease). Another example of a mutation which causes the creation of non-functional protein and leads to a disease is a mutation in human *CFTR* gene (*C*ystic *F*ibrosis *TR*ansmembrane conductance Regulator) located in the q31.2 locus of chromosome 7. The proteins encoded by *CFTR* functions as a channel for chloride ions, which can move in and out of cells. The most common mutation, which is a three nucleotide deletion, leads to the loss of an amino acid and a non-functional protein [311, 312].

Apart from single nucleotide changes, alternative splicing events might lead to a whole protein domain removal or incorporation. Studies of protein structures and structural alterations are challenging, as structural biology and exploring protein structure and molecular dynamics involves computationally heavy, intensive tasks. These tasks include protein modelling, molecular dynamics and quantum mechanics calculations (to be precise, short cuts based on knowledge derived from quantum mechanics), i.e., to study the changes in the binding between the substrate and the amino acids in the enzymatic active site, which can be caused by a genomic variant [313]. Apart from the need in extended computational power and resources, the task of linking transcriptomics and structural biology is challenged by the limited amount of available 3D protein structures. This makes the task even more complex, especially when considering linking transcriptomics and structural biology genome wide. Unfortunately a possibility of local modelling of certain protein domains does not give a lot of insight, as even very conserved and closed domains often behave differently in different circumstances [314, 315]. These circumstances can be the surrounding amino acids, the electrostatic potential around them, or the level of the domain's exposure to the surface.

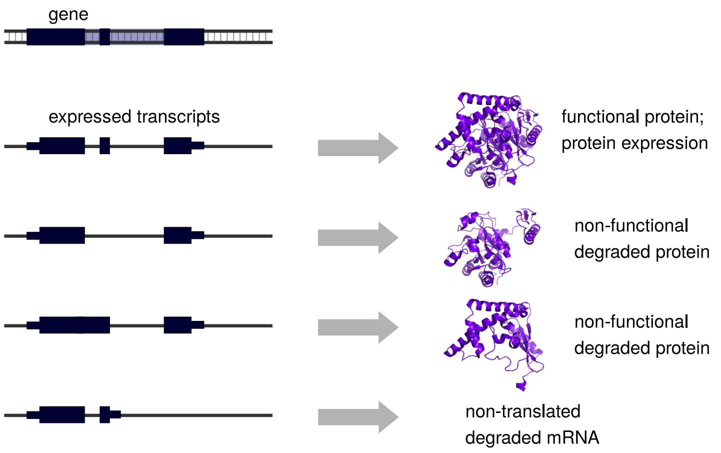


Figure 6.2: The Figure shows that multiple mRNAs can be transcribed from one gene (left panel of the figure). Some of the mRNAs lead to a stable functional protein, while some result in non-functional protein and some are not used as a protein template at all.

A developing field of *structural genomics* might be the bridge connecting transcriptomics and protein structures and functions [316]. The principle of structural genomics is to identify as many protein structures as possible using all currently available resources and techniques. These techniques include *de novo* structure determination using X-ray crystallography and Nuclear Magnetic Resonance, modelling based methods such as *ab initio* modeling, sequence based homology modelling and domain modeling based on the fold similarities rather than the sequence identity [317, 318].

The intermediate goal of the field is to provide as many protein structures as possible, and the end goal is to identify and describe all proteins within the genome of interest. Structural genomics has already successfully been applied on bacteria [319] and extending the initiative to higher organisms, such as humans, is hopefully a matter of time.

The initiative of structural genomics has a great potential, as it will expand our understanding of protein folding, structure conservation among and within species, provide a better and broader understanding of protein function, and make protein stability and function prediction easier and more reliable.

Chapter 7

Summary in English

Bioinformatics, first introduced in 1970 by Paulien Hogeweg and Ben Hesper, has changed and evolved drastically. Bioinformatics was formed as a field to get information about various processes in biological systems (hence the term *bio-informatics*). Initially, bioinformatics focused on analyzing a handful of gene and protein sequences, protein structures and organism evolution based on protein sequences and structural domains. However, with years it shifted to the exploration of whole genomes and transcriptomes and the study of biological processes at a genome- (and transcriptome-) wide scale in model systems, living cells and organisms. After the introduction of one of the greatest scientific and technological breakthroughs of the past two decades – massively parallel Next Generation Sequencing (NGS) – sequencing of genomes and transcriptomes of whole organisms became possible. Bioinformatics faced a new challenge of analyzing high throughput genomic and transcriptomic data. The main challenge introduced by complex NGS data is in its nature – DNA has to be shredded before sequencing, therefore NGS produces millions of very short sequenced DNA fragments (*reads*). Despite the recent increase in read length, finding the origin of reads on the genome is still a challenge, as the length still does not exceed hundreds of nucleotides. Pacific Biosciences' SMRT technology and Oxford Nanopore, recently introduced high-throughtput sequencing technologies, are the only technoloies available on the market capable of sequencing thousands of bases.

An additional challenge in analyzing NGS data emerges when working with transcriptomic material – reads originating from RNA. Multiple RNA transcripts from one gene often share a specific region, and each transcript can be present multiple times. Therefore, the origin of reads – the transcript they are coming from – is hard to find. This introduces numerous biases and limits our understanding of RNA biology and molecular processes happening at the RNA level. As this thesis was written, the analysis of an RNA sequencing sample was limited due to the limited amount of programs for the bioinformatic analysis of RNA. Available RNA analysis tools only focussed on finding differences in the expression of certain genes between different samples (i.e., healthy individuals versus diseased patients), discover alternative splicing events, identify alternative polyadenylation events.

The focus of this thesis was to explore the possibilities of current bioinformatic analysis of transciptomic data and complement such tools by newly developed programs. The latter was particularly important for atypical RNA analyses revealing yet unexplored molecular mechanisms in RNA biogenesis. This is covered in **Chapter 2**, where we analyzed nuclear RNA – not commonly used for RNA-seq. The majority of established RNA-seq data sets are derived from total cellular or cytoplasmic RNA. In the analysis of total RNA, the focus is on the result of transcription and RNA processing. Studying nuclear RNA facilitates a more thorough exploration of ongoing mRNA processing events like splicing.

Splicing is a process during which the regions of a gene called introns are removed and the remaining regions – exons – are joined. As simple as splicing might sound at first, it consists of numerous steps, and each of those steps deserves a thorough exploration. Analyzing reads from nuclear RNA in **Chapters 2** and **3** of this thesis, we discovered that introns in human genes are not always excised sequentially, from the beginning towards the end of the gene. Also, some introns are not removed at once, they are rather excised in multiple steps, piece by piece. In fact, using the example of the gene coding for dystrophin (*DMD*), the longest gene within the human genome, we show that more than half of *DMD* introns are removed in multiple steps.

In **Chapter 4**, we complemented publicly available programs with in-house developed scripts and performed a study of age related changes in RNA processing, namely splicing, in over 600 healthy Dutch individuals. The analyses were conducted on adult individuals with ages spanning more than 60 years. We found alternative splicing events such as exon skipping – when during the process of splicing not only introns but also an exon is excised – occurring more often in older individuals. We noticed the same trend for another type of alternative splicing event, intron retention, an event when an intron is not excised and is included in the nascent transcript. Both of these events were occurring more often in older individuals transcriptome-wide, without any preference for a specific pathway or functional group of genes. These findings suggested that such age related changes might have a fundamental nature and can potentially help to understand how and why we age.

Next to the bioinformatic analyses of RNA sequencing samples from healthy individuals, in **Chapter 5** we analyzed samples from patients with breast cancer. In this study we were particularly interested in allele-specific expression, a biological event potentially important for the pathophysiology of cancer. A healthy human genome is diploid, which means that every chromosome occurs in pairs and that there are two copies of each gene in a cell (the two copies are the two *alleles*). For each position of a gene sequence two different nucleotides can be present (unless we are facing an insertion or a deletion, which results in a missing nucleotide). Chromosomal aberrations – duplication or deletion of (parts of) a chromosome (allele) – are frequent in tumors. Such events may result in an unequal expression of the two alleles. This unequal allelic expression may also be present in healthy individuals and can be directed via biological mechanisms different from DNA chromosomal aberrations (e.g. parental imprinting). In this study we compared the differences between imbalanced allelic expression between healthy and tumor tissue and explored the potential role of allelic imbalance in the pathogenesis of breast cancer. We provided a detailed classification of allele-specific expression cases, compared them in control versus tumor tissues of the same patients and explained each pattern on RNA expression (e.g., a downregulation of an allele's expression) by the underlying DNA alterations.

In summary, this thesis focuses on the uncovering of previously unexplored information in RNA sequencing data. We show that use of unconventional types of RNA for sequencing and new tools for the analysis RNA-seq can reveal novel insights in the transcriptional and post-transcriptional regulation of gene expression, relevant for the understanding of normal physiological processes such as aging and pathophysiological processes such as Duchenne muscular dystrophy and cancer.

Chapter 8

Samenvatting (summary in Dutch)

Bioinformatica werd in 1970 geïntroduceerd door Paulien Hogeweg en Ben Hesper. Sindsdien is deze wetenschap ingrijpend veranderd. De essentie van de bioinformatica ligt in het verkrijgen van informatie over biologische systemen. Vandaar de naam. Aanvankelijk richtte de bioinformatica zich op de analyse van een beperkt aantal eiwit sequenties, eiwit structuren en evolutie studies op basis van eiwit sequenties en structurele domeinen. Met de jaren heeft het zich ontwikkeld tot een wetenschap die hele genomen, transcriptomen en biologische processen op een genoom-wijde schaal bestudeert in model systemen, levende cellen en organismen. Na de introductie van een van de grootste wetenschappelijke en technologische doorbraken van de laatste twee decennia — grootschalige, parallelle next generation sequencing (NGS) – werd het mogelijk om genomen en transcriptomen van hele organismen te bepalen. De bioinformatica werd geconfronteerd met nieuwe uitdagingen bij de analyse van deze genomen en transcriptomen. De grootste uitdaging in de analyse van complexe NGS data ligt verborgen in de benodigde fragmentatie van het DNA voorafgaande aan de sequencing. Hierdoor ontstaan miljoenen korte sequentie fragmenten (*'reads'*). Ondanks de recente toename in de lengte van deze reads, is de gemiddelde read lengte nog steeds beperkt tot ongeveer 100 nuleotiden. Dit maakt het lastig om de oorsprong van de reads te bepalen. Alleen met reads van het Pacific Biosciences platform, die vele duizenden nucleotiden lang zijn, wordt dit probleem verholpen. Het aantal reads per run in dit platform is echter nog beperkt in vergelijking met de platformen met korte read lengtes.

Transcriptoom data (RNA-seq) vormen een extra uitdaging bij de analyse. Verschillende RNA transcripten afkomstig van hetzelfde gen delen vaak een groot gedeelte van hun sequentie. Daarom is het moeilijk om te bepalen van welk transcript de read afkomstig is. Dit resulteert in onevenredige representatie van transcripten (*'bias'*) en belemmert de studie van de biologische processen die bij de vorming en afbraak van RNA een rol spelen. In de tijd dat het onderzoek zoals beschreven in dit proefschrift werd verricht, waren nog niet veel geavanceerde programma's beschikbaar die specifiek voor RNA analyses waren gemaakt. Met de beschikbare software konden slechts verschillen in de expressie van genen worden bepaald (bijvoorbeeld tussen zieke en gezonde personen), en alternatieve splicing en alternatieve poly-adenylerings posities worden gedetecteerd.

De focus van dit proefschrift ligt op het onderzoeken van de mogelijkheden van bestaande bioinformatische analyse programma's voor transcriptoom data en het ontwikkelen van algortimes voor analyses die niet uitgevoerd konden worden met bestaande software. Dit laatste was met name belangrijk voor atypische RNA analyses om nog niet ontdekte mechanismen in RNA biogenese op te helderen. Dit is het geval in **hoofdstuk 2**, waar we RNA uit de celkern hebben geanalyseerd — iets wat niet standaard wordt gedaan omdat men zich over het algemeen richt op de analyse van het totale RNA in de cel of het cytoplasmatische RNA. Waar bij de analyse van totaal RNA de nadruk veelal ligt op het resultaat van transcriptie en RNA processering, opent de analyse van RNA uit de nucleus de mogelijkheid om processen die een rol spelen bij de maturatie van mRNA, zoals splicing, te bestuderen.

Splicing is een proces waarbij bepaalde regionen in een gen, de zogenaamde intronen, worden verwijderd en de overgebleven exonen met elkaar worden verbonden. Hoewel dit simpel klinkt, is splicing een meerstapsproces, waarbij iedere stap nadere bestudering behoeft. Door de analyse van RNA uit de celkern in **hoofdstukken 2** en **3**, ontdekten we dat intronen in humane genen niet altijd in volgorde, vanaf het begin tot het einde van het gen, worden verwijderd. Verder vonden we dat sommige intronen in meerdere stappen worden gespliced. In het *DMD* gen, het langste gen in het humane genome, gold dit laatste voor meer dan de helft van de intronen.

In **hoofdstuk 4** gebruikten we publiek beschikbare software en zelf ontwikkelde scripts en onderzochten leeftijd-gerelateerde verandering in RNA splicing in meer dan 600 gezonde, nederlandse individuen, variërend in leeftijd van 18 tot 80 jaar. We vonden dat een bepaalde vorm van alternatieve splicing, exon skipping, het verwijderen van exonen samen met intronen, meer voorkomt bij oudere dan bij jongere individuen. Eenzelfde trend werd gevonden voor intron retentie, een vorm van alternatieve splicing waarbij een intron terecht komt in het uiteindelijke transcript. Beide vormen van alternatieve splicing kwamen meer voor bij ouderen, maar werden niet aangetroffen in specifieke genen maar kwamen in het gehele transcriptoom voor. Deze bevindingen suggereerden dat deze leeftijds-gerelateerde veranderingen fundamenteel van aard zijn en mogelijk bijdragen tot het verouderingsproces.

Naast de bioinformatische analyse van RNA sequencing monsters van gezonde individuen, onderzochten we in **hoofdstuk 5** ook monsters van patiënten met borst kanker. In deze studie waren we in het bijzonder geïnteresseerd in allel-specifieke expressie, een biologisch fenomeen dat mogelijk een rol speelt in de pathofysiologie van kanker. Een normaal humaan genoom is diploïde. Dit betekent dat er van elk gen twee copieën aanwezig zijn in een cel op twee verschillende chromosomen. Chromosomale afwijkingen — deleties of duplicaties van (gedeeltes van) chromosomen — komen in veel tumoren voor. Dit kan leiden tot een ongebalanceerde expressie van de twee allelen. Echter, ongelijke expressie van de twee allelen kan ook voorkomen in gezond weefsel, door andere biologische mechanismen. In deze studie vergeleken we de ongebalanceerde expressie tussen gezond en tumor weefsel van dezelfde patiënt en probeerden de expressie patronen in de tumor te verklaren uit de DNA veranderingen.

Kortom, dit proefschrift beschrijft verschillende nieuwe manieren om de informatie die verborgen ligt in RNA sequencing data naar boven te halen.

Chapter 10

Biography

Irina Viktorovna Pulyakhina was born in Moscow on the 21st of February 1989. She finished the “Vorobievi Gori” high school in June 2001 and entered the Faculty of Bioengineering and Bioinformatics of Moscow State University in September 2001. After finishing three years of her BSc studies, she was selected for the MoBiLe ’09 summer school – an exchange program between Moscow State University and Leiden University Medical Center (Leiden, the Netherlands), initiated by prof. A.E. Gorbalenya in 2003. During her exchange, Pulyakhina conducted pioneering research in a new field of next generation sequencing data analysis, more specifically human DNA sequencing. She was working under the supervision of prof. Johan den Dunnen at the LGTC – Leiden Genome Technology Center.

Pulyakhina returned to Leiden University Medical Center in 2010 to work on an MSc project in metagenomics under the supervision of dr. Jeroen Laros in the Department of Human Genetics. This nine month collaboration on the comparison of different sequencing platforms for the purpose of metagenome sequencing resulted in a successful grant application that is now used for a fellow PhD student to continue the work that Pulyakhina started.

At the end of her MSc project on metagenomics, Pulyakhina was offered a PhD position by dr. Peter-Bram ’t Hoen, whom she had already met in 2009. Her PhD focuses on answering various biological and medical questions analyzing RNA sequencing data. During the last four years, Pulyakhina has worked on various projects covering the whole spectrum of RNA analysis approaches. This resulted in a number of successful collaborations (including international ones, such as the GEUVADIS project) and a number of papers, which were published in high impact factor peer reviewed journals (such as “Nucleic Acids Research” and “Genome Biology”) or are currently submitted or in preparation.

Pulyakhina concluded her PhD with a thesis entitled “A telescope for the RNA universe: novel bioinformatic methods to analyze RNA sequencing data”, after which she has been appointed as a postdoctoral researcher at the University of Oxford, Great Britain, in the group of prof. Julian Knight. Starting from August 2015, she will focus her research on the genetic background of ankylosing spondylitis.

Chapter 10

Publication list

Allelic imbalance and its role in breast cancer pathogenesis. Pulyakhina I., Vreeswijk M.P.G., Meijers C.M., Laros J.F., den Dunnen J.T. and 't Hoen P.A.C. *in preparation*

Aging is associated with increased incidence of alternative splicing. Pulyakhina I.*, Takhaveev V.*, Vermaat M., Laros J.F., den Dunnen J.T. and 't Hoen P.A.C. *in preparation*

Detailed insight in the splicing of the dystrophin transcript. Gazzoli I., Pulyakhina I., Laros J.F., Verwij N.E., den Dunnen J.T., 't Hoen P.A.C. and Aartsma-Rus A. *RNA biology, 2015 Dec 15:0. [Epub ahead of print]*

A novel analytical approach for the detection of alternative, non-sequential and recursive splicing. Pulyakhina I., Gazzoli I., Verwij N.E., den Dunnen J.T., 't Hoen P.A.C., Aartsma-Rus A. and Laros J.F.J. *Nucleic Acids Research*, 2015 Jul 13;43(12):e80. doi: 10.1093/nar/gkv242

Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant. van Roosmalen W., le Devedec S.E., Golani O., Smid M., Pulyakhina I., Timmermans A.M., Look M.P., Di Z., de Graauw M., Naffar-Abu-Amara S., Kirsanova C., 't Hoen P.A.C., Martens J.W.M., Foekens J.A., Geiger B. and van de Water B. *Journal of Clinical Investigation*, 2015 Apr;125(4):1648-64. doi: 10.1172/JCI74440

Determining the quality and complexity of next-generation sequencing data without a reference genome. Anvar S.Y., Khachatryan L., Vermaat M., van Galen M., Pulyakhina I., Ariyurek Y., Kraaijeveld K., den Dunnen J.T., de Knijff P., 't Hoen P.A.C. and Laros J.F.J. *Genome Biology*, 2014;15(12):555

Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories.
't Hoen P.A.C., Friedlander M.R., Almlof J., Sammeth M., Pulyakhina I., Anvar S.Y., Laros J.F.J., Buermans H.P., Karlberg O., Brannvall M.; GEUVADIS Consortium, den Dunnen J.T., van Ommen G.J., Gut, I.G., Guigo R., Estivill X., Syvanen A.C., Dermitzakis E.T. and Lappalainen T. *Nature Biotechnology*, 2013 Nov;31(11):1015-22. doi: 10.1038/nbt.2702

Transcriptome and genome sequencing uncovers functional variation in humans.
Lappalainen T., Sammeth M., Friedlander M.R., 't Hoen P.A., Monlong J., Rivas M.A., Gonzalez-Porta M., Kurbatova N., Griebel T., Ferreira P.G., Barann M., Wieland T., Greger L., van Iterson M., Almlof J., Ribeca P., Pulyakhina I., Esser D., Giger T., Tikhonov A., Sultan M., Bertier G., MacArthur D.G.,

Lek M., Lizano E., Buermans H.P., Padioleau I., Schwarzmayr T., Karlberg O., Ongen H., Kilpinen H., Beltran S., Gut M., Kahlem K., Amstislavskiy V., Stegle O., Pirinen M., Montgomery S.B., Donnelly P., McCarthy M.I., Flicek P., Strom T.M., Geuvadis Consortium, Lehrach H., Schreiber S., Sudbrak R., Carracedo A., Antonarakis S.E., Hasler R., Syvanen A.C., van Ommen G.J., Brazma A., Meitinger T., Rosenstiel P., Guigo, R., Gut I.G., Estivill X. and Dermitzakis E.T. *Nature*, 2013 Sep 26;501(7468):506-11. doi: 10.1038/nature12531

Bibliography

[1] J. D. Watson et al. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[2] A. J. Shatkin. Capping of eucaryotic mRNAs. *Cell*, 9(4 PT 2):645–653, Dec 1976.

[3] P. Danecek et al. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol.*, 13(4):26, 2012.

[4] J. D. Alfonzo et al. The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. *Nucleic Acids Res.*, 25(19):3751–3759, Oct 1997.

[5] G. J. Arts et al. Mechanism and evolution of RNA editing in kinetoplastida. *Biochim. Biophys. Acta*, 1307(1):39–54, Jun 1996.

[6] H. Tilgner et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, 22(9):1616–1625, Sep 2012.

[7] S. W. Roy et al. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.*, 7(3):211–221, Mar 2006.

[8] A. J. Taggart et al. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.*, 19(7):719–721, Jul 2012.

[9] A. Corvelo et al. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput. Biol.*, 6(11):e1001016, 2010.

[10] A. A. Patel et al. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, 4(12):960–970, Dec 2003.

[11] A. J. Matlin et al. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5):386–398, May 2005.

[12] Q. Pan et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, Dec 2008.

[13] G. Di Segni et al. Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A.*, 105(19):6864–6869, May 2008.

[14] D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, 72:291–336, 2003.

[15] M. Sammeth et al. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4(8):e1000147, 2008.

[16] S. Djebali et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, Sep 2012.

[17] B. Ewing et al. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.*, 25(2):232–234, Jun 2000.

[18] O. Olsvik et al. Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. *J. Clin. Microbiol.*, 31(1):22–25, Jan 1993.

[19] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[20] C. A. Maher et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, Mar 2009.

[21] F. Sanger et al. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, Dec 1977.

[22] F. Sanger et al. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975.

[23] D. R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[24] B. Ewing et al. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, 8(3):186–194, Mar 1998.

[25] A. Mortazavi et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

[26] K. D. Hansen et al. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, 38(12):e131, Jul 2010.

[27] Y. Chu et al. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*, 22(4):271–274, Aug 2012.

[28] J. R. ten Bosch et al. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn*, 10(6):484–492, Nov 2008.

[29] T. Tucker et al. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.*, 85(2):142–154, Aug 2009.

[30] M. J. Fullwood et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, 19(4):521–532, Apr 2009.

[31] R. Morin et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1):81–94, Jul 2008.

[32] S. W. Roy et al. When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *Bioessays*, 30(6):601–605, Jun 2008.

[33] J. Cocquet et al. Reverse transcriptase template switching and false alternative transcripts. *Genomics*, 88(1):127–131, Jul 2006.

[34] R. M. Mader et al. Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J. Lab. Clin. Med.*, 137(6):422–428, Jun 2001.

[35] C. W. Fuller et al. The challenges of sequencing by synthesis. *Nat. Biotechnol.*, 27(11):1013–1023, Nov 2009.

[36] Y. Zhang et al. PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics*, 28(4):479–486, Feb 2012.

[37] M. Hamada et al. Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics*, 27(22):3085–3092, Nov 2011.

[38] G. A. Heap et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, 19(1):122–134, Jan 2010.

[39] H. Li et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[40] W. J. Kent et al. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, Jun 2002.

[41] H. Thorvaldsdottir et al. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, 14(2):178–192, Mar 2013.

[42] J. T. Robinson et al. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26, Jan 2011.

[43] F. Finotello et al. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics*, 15 Suppl 1:S7, 2014.

[44] W. Zheng et al. Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics*, 12:290, 2011.

[45] E. E. Khrameeva et al. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics*, 13 Suppl 6:S4, 2012.

[46] S. Schwartz et al. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE*, 6(1):e16685, 2011.

[47] A. Roberts et al. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, 12(3):R22, 2011.

[48] J. F. Degner et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, Dec 2009.

[49] V. Sousa et al. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.*, 14(6):404–414, Jun 2013.

[50] C. A. Meyer et al. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, 15(11):709–721, Nov 2014.

[51] D. Sims et al. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–132, Feb 2014.

[52] C. Kulahoglu et al. Quantitative transcriptome analysis using RNA-seq. *Methods Mol. Biol.*, 1158:71–91, 2014.

[53] P. Jain et al. Augmenting transcriptome assembly by combining de novo and genome-guided tools. *PeerJ*, 1:e133, 2013.

[54] R. Liu et al. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, 15(1):364, Dec 2014.

[55] A. Roberts et al. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, Sep 2011.

[56] D. Hiller et al. Simultaneous isoform discovery and quantification from RNA-seq. *Stat Biosci*, 5(1):100–118, May 2013.

[57] R. K. Varshney et al. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.*, 27(9):522–530, Sep 2009.

[58] Y. Li et al. RNA-Seq Based De Novo Transcriptome Assembly and Gene Discovery of Cistanche deserticola Fleshy Stem. *PLoS ONE*, 10(5):e0125722, 2015.

[59] D. Powell et al. De novo transcriptome analysis of the banana shrimp (Fenneropenaeus merguiensis) and identification of genes associated with reproduction and development. *Mar Genomics*, Apr 2015.

[60] K. S. Arun-Chinnappa et al. De novo assembly of a genome-wide transcriptome map of Vicia faba (L.) for transfer cell research. *Front Plant Sci*, 6:217, 2015.

[61] S. Anders et al. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017, Oct 2012.

[62] C. Trapnell et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, 31(1):46–53, Jan 2013.

[63] S. Danckwardt et al. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.*, 27(3):482–498, Feb 2008.

[64] B. Tian et al. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33(1):201–212, 2005.

[65] E. de Klerk et al. RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell. Mol. Life Sci.*, 71(18):3537–3551, Sep 2014.

[66] V. Le Texier et al. AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*, 7:169, 2006.

[67] V. D'mello et al. Alternative mRNA polyadenylation can potentially affect detection of gene expression by affymetrix genechip arrays. *Appl. Bioinformatics*, 5(4):249–253, 2006.

[68] G. Ji et al. A classification-based prediction model of messenger RNA polyadenylation sites. *J. Theor. Biol.*, 265(3):287–296, Aug 2010.

[69] Z. Xia et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun*, 5:5274, 2014.

[70] H. Zhang et al. PolyA-DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, 33(Database issue):D116–120, Jan 2005.

[71] J. Y. Lee et al. PolyA-DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, 35(Database issue):D165–168, Jan 2007.

[72] G. Ji et al. Predictive modeling of plant messenger RNA polyadenylation sites. *BMC Bioinformatics*, 8:43, 2007.

[73] P. de la Grange et al. A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. *BMC Bioinformatics*, 8:180, 2007.

[74] Z. H. Zhang et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE*, 9(8):e103207, 2014.

[75] R. Chandramohan et al. Benchmarking RNA-Seq quantification tools. *Conf Proc IEEE Eng Med Biol Soc*, 2013:647–650, 2013.

[76] S. Wesolowski et al. A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach. *Biosensors (Basel)*, 3(3):238–258, Sep 2013.

[77] M. D. Robinson et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

[78] S. Anders et al. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

[79] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[80] F. Rapaport et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, 14(9):R95, 2013.

[81] A. Brennicke et al. RNA editing. *FEMS Microbiol. Rev.*, 23(3):297–316, Jun 1999.

[82] A. A. Su et al. A-to-I and C-to-U editing within transfer RNAs. *Biochemistry Mosc.*, 76(8):932–937, Aug 2011.

[83] A. M. Kiran et al. Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.*, 41(Database issue):D258–261, Jan 2013.

[84] E. T. Wang et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[85] H. J. Mardon et al. A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucleic Acids Res.*, 15(19):7725–7733, Oct 1987.

[86] S. A. Filichkin et al. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.*, 20(1):45–58, Jan 2010.

[87] A. Pandya-Jones et al. Co-transcriptional splicing of constitutive and alternative exons. *RNA*, 15(10):1896–1908, Oct 2009.

[88] D. L. Bentley. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, 15(3):163–175, Mar 2014.

[89] Y. L. Khodor et al. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev.*, 25(23):2502–2512, Dec 2011.

[90] C. Svensson et al. Splicing of adenovirus 2 early region 1A mRNAs is non-sequential. *J. Mol. Biol.*, 165(3):475–495, Apr 1983.

[91] K. M. Lang et al. In vitro splicing pathways of pre-mRNAs containing multiple intervening sequences? *Mol. Cell. Biol.*, 7(10):3428–3437, Oct 1987.

[92] D. Baralle et al. Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.*, 42(10):737–748, Oct 2005.

[93] Y. W. Huang et al. Identification of a porcine DC-SIGN-related C-type lectin, porcine CLEC4G (LSECtin), and its order of intron removal during splicing: comparative genomic analyses of the cluster of genes CD23/CLEC4G/DC-SIGN among mammalian species. *Dev. Comp. Immunol.*, 33(6):747–760, Jun 2009.

[94] U. Schwarze et al. Redefinition of exon 7 in the COL1A1 gene of type I collagen by an intron 8 splice-donor-site mutation in a form of osteogenesis imperfecta: influence of intron splice order on outcome of splice-site mutation. *Am. J. Hum. Genet.*, 65(2):336–344, Aug 1999.

[95] U. Schwarze et al. Splicing defects in the COL3A1 gene: marked preference for 5' (donor) spice-site mutations in patients with exon-skipping mutations and Ehlers-Danlos syndrome type IV. *Am. J. Hum. Genet.*, 61(6):1276–1286, Dec 1997.

[96] K. Wang et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38(18):e178, Oct 2010.

[97] E. Ullu et al. Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. *Mol. Cell. Biol.*, 13(1):720–725, Jan 1993.

[98] H. Suzuki et al. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. *FEBS Lett.*, 587(6):555–561, Mar 2013.

[99] A. Katolik et al. Regiospecific Solid-Phase Synthesis of Branched Oligoribonucleotides That Mimic Intronic Lariat RNA Intermediates. *J. Org. Chem.*, Jan 2014.

[100] S. Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–438, Jun 2004.

[101] Z. Wang et al. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, Jan 2009.

[102] J. C. Marioni et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, Sep 2008.

[103] M. J. Fullwood, C. L. Wei, E. T. Liu, and Y. Ruan. Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. *Genome Res*, 2009.

[104] I. Gazzoli et al. Non-sequential and recursive splicing of the dystrophin transcript. *Manuscript in preparation.*

[105] P. Flicek et al. Ensembl 2011. *Nucleic Acids Res.*, 39(Database issue):D800–806, Jan 2011.

[106] J. M. Ruijter et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.*, 37(6):e45, Apr 2009.

[107] T. D. Wu et al. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, Apr 2010.

[108] C. Trapnell et al. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[109] M. T. Dimon et al. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS ONE*, 5(11):e13875, 2010.

[110] M. Guttman et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat. Biotechnol.*, 28(5):503–510, May 2010.

[111] Y. Katz et al. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat. Biotechnol.*, 7(12):1009–1015, Dec 2010.

[112] A. R. Hatton et al. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Mol. Cell*, 2(6):787–796, Dec 1998.

[113] T. A. Cooper et al. RNA and disease. *Cell*, 136(4):777–793, Feb 2009.

[114] M. B. Coelho et al. Regulation of alternative pre-mRNA splicing. *Methods Mol. Biol.*, 1126:55–82, 2014.

[115] M. C. Wahl et al. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, Feb 2009.

[116] H. Shen et al. RS domain-splicing signal interactions in splicing of U12-type and U2-type introns. *Nat. Struct. Mol. Biol.*, 14(7):597–603, Jul 2007.

[117] G. E. Parada et al. A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.*, 42(16):10564–10578, 2014.

[118] Y. Barash et al. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[119] H. Suzuki et al. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res.*, 34(8):e63, 2006.

[120] T. Takahara et al. Delay in synthesis of the 3' splice site promotes trans-splicing of the preceding 5' splice site. *Mol. Cell*, 18(2):245–251, Apr 2005.

[121] J. Singh et al. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.*, 16(11):1128–1133, Nov 2009.

[122] R. F. Luco et al. Epigenetics in alternative pre-mRNA splicing. *Cell*, 144(1):16–26, Jan 2011.

[123] S. Shukla et al. Co-transcriptional regulation of alternative pre-mRNA splicing. *Biochim. Biophys. Acta*, 1819(7):673–683, Jul 2012.

[124] M. J. Moore et al. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, Feb 2009.

[125] R. Perales et al. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Mol. Cell*, 36(2):178–191, Oct 2009.

[126] D. Y. Vargas et al. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, 147(5):1054–1065, Nov 2011.

[127] J. M. Gudas et al. Ordered splicing of thymidine kinase pre-mRNA during the S phase of the cell cycle. *Mol. Cell. Biol.*, 10(10):5591–5595, Oct 1990.

[128] M. J. Tsai et al. Processing of high molecular weight ovalbumin and ovomucoid precursor RNAs to messenger RNA. *Cell*, 22(1 Pt 1):219–230, Nov 1980.

[129] M. de la Mata et al. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA*, 16(5):904–912, May 2010.

[130] S. Shepard et al. The peculiarities of large intron splicing in animals. *PLoS ONE*, 4(11):e7853, 2009.

[131] J. M. Burnette et al. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*, 170(2):661–674, Jun 2005.

[132] J. F. Conklin et al. Stabilization and analysis of intron lariats in vivo. *Methods*, 37(4):368–375, Dec 2005.

[133] S. Ott et al. Intrasplicing–analysis of long intron sequences. *Pac Symp Biocomput*, pages 339–350, 2003.

[134] N.P. Kandul et al. Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC Genet.*, 10:67, 2009.

[135] T. Kameyama et al. Re-splicing of mature mRNA in cancer cells promotes activation of distant weak alternative splice sites. *Nucleic Acids Res.*, 40(16):7896–7906, Sep 2012.

[136] C. N. Tennyson et al. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat. Genet.*, 9(2):184–190, Feb 1995.

[137] C. H. Zhu et al. Cellular senescence in human myoblasts is overcome by human telomerase reverse transcriptase and cyclin-dependent kinase 4: consequences in aging muscle and therapeutic strategies for muscular dystrophies. *Aging Cell*, 6(4):515–523, Aug 2007.

[138] D. U. Kemaladewi et al. Dual exon skipping in myostatin and dystrophin for Duchenne muscular dystrophy. *BMC Med Genomics*, 4:36, 2011.

[139] I. Pulyakhina et al. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res.*, Mar 2015.

[140] T. R. Mercer et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc*, 9(5):989–1009, May 2014.

[141] A. Gnirke et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, 27(2):182–189, Feb 2009.

[142] F. H. Nasim et al. A Sequential splicing mechanism promotes selection of an optimal exon by repositioning a downstream 5' splice site in preprotachykinin pre-mRNA. *Genes Dev.*, 4(7):1172–1184, Jul 1990.

[143] O. Kessler et al. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Mol. Cell. Biol.*, 13(10):6211–6222, Oct 1993.

[144] M. Yang et al. Splicing of mouse p53 pre-mRNA does not always follow the "first come, first served" principle and may be influenced by cisplatin treatment and serum starvation. *Mol. Biol. Rep.*, 39(9):9247–9256, Sep 2012.

[145] I. A. Swinburne et al. Intron length increases oscillatory periods of gene expression in animal cells. *Genes Dev.*, 22(17):2342–2346, Sep 2008.

[146] K. M. Neugebauer. On the importance of being co-transcriptional. *J. Cell. Sci.*, 115(Pt 20):3865–3871, Oct 2002.

[147] N. J. Proudfoot. Dawdling polymerases allow introns time to splice. *Nat. Struct. Biol.*, 10(11):876–878, Nov 2003.

[148] M. Aebi et al. Precision and orderliness in splicing. *Trends. Genet.*, 3:102–107, Mar 1987.

[149] A. P. Monaco et al. An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. *Genomics*, 2(1):90–95, Jan 1988.

[150] A. Aartsma-Rus. Antisense-mediated modulation of splicing: therapeutic implications for Duchenne muscular dystrophy. *RNA Biol*, 7(4):453–461, 2010.

[151] A. Aartsma-Rus et al. Functional analysis of 114 exon-internal AONs for targeted DMD exon skipping: indication for steric hindrance of SR protein binding sites. *Oligonucleotides*, 15(4):284–297, Dec 2005.

[152] S. D. Wilton et al. Antisense oligonucleotide-induced exon skipping across the human dystrophin gene transcript. *Mol. Ther.*, 15(7):1288–1296, Jul 2007.

[153] A. Aartsma-Rus et al. Antisense-induced multiexon skipping for Duchenne muscular dystrophy makes more sense. *Am. J. Hum. Genet.*, 74(1):83–92, Jan 2004.

[154] M. Koenig et al. The molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion. *Am. J. Hum. Genet.*, 45(4):498–506, Oct 1989.

[155] J. T. Den Dunnen et al. Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. *Am. J. Hum. Genet.*, 45(6):835–847, Dec 1989.

[156] C. Nobile et al. Exon-intron organization of the human dystrophin gene. *Genomics*, 45(2):421–424, Oct 1997.

[157] A. Aartsma-Rus et al. Theoretic applicability of antisense-mediated exon skipping for Duchenne muscular dystrophy mutations. *Hum. Mutat.*, 30(3):293–299, Mar 2009.

[158] Y. Aoki et al. Bodywide skipping of exons 45-55 in dystrophic mdx52 mice by systemic antisense delivery. *Proc. Natl. Acad. Sci. U.S.A.*, 109(34):13763–13768, Aug 2012.

[159] C. R. Sibley et al. Recursive splicing in long vertebrate genes. *Nature*, 521(7552):371–375, May 2015.

[160] T. Lu et al. Gene regulation and DNA damage in the ageing human brain. *Nature*, 429(6994):883–891, Jun 2004.

[161] S. Hekimi et al. Genetics and the specificity of the aging process. *Science*, 299(5611):1351–1354, Feb 2003.

[162] S. A. McCarroll et al. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.*, 36(2):197–204, Feb 2004.

[163] W. M. Passtoors et al. Genomic studies in ageing research: the need to integrate genetic and gene expression approaches. *J. Intern. Med.*, 263(2):153–166, Feb 2008.

[164] M. Jung et al. Longitudinal epigenetic and gene expression profiles analyzed by three-component analysis reveal down-regulation of genes involved in protein translation in human aging. *Nucleic Acids Res.*, May 2015.

[165] F. Licastro et al. Innate immunity and inflammation in ageing: a key for understanding age-related diseases. *Immun Ageing*, 2:8, May 2005.

[166] M. Gheorghe et al. Major aging-associated RNA expressions change at two distinct age-positions. *BMC Genomics*, 15:132, 2014.

[167] Y. N. Wei et al. Transcript and protein expression decoupling reveals RNA binding proteins and miRNAs as potential modulators of human aging. *Genome Biol.*, 16:41, 2015.

[168] J. M. Johnson et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653):2141–2144, Dec 2003.

[169] J. Takeda et al. H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res.*, 38(Database issue):86–90, Jan 2010.

[170] R. Blekhman et al. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, 20(2):180–189, Feb 2010.

[171] J. A. Calarco et al. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.*, 21(22):2963–2975, Nov 2007.

[172] A. Battle et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, 24(1):14–24, Jan 2014.

[173] T. Lappalainen et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[174] J. Ule et al. Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.*, 37(8):844–852, Aug 2005.

[175] B. J. Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, Jul 2006.

[176] B. K. Dredge et al. The splice of life: alternative splicing and neurological disease. *Nat. Rev. Neurosci.*, 2(1):43–50, Jan 2001.

[177] L. Buee et al. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res. Brain Res. Rev.*, 33(1):95–130, Aug 2000.

[178] L. W. Harries et al. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell*, 10(5):868–878, Oct 2011.

[179] P. Mazin et al. Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.*, 9:633, 2013.

[180] S. Scholtens et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*, Dec 2014.

[181] `www.lumc.nl/org/ouderengeneeskunde/research/study-populations/LeidenLongevity-study/`. Accessed: 2015-06-01.

[182] A. Dobin et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[183] H. Konig et al. Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. *Cell*, 131(4):718–729, Nov 2007.

[184] W. Y. Tarn et al. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science*, 273(5283):1824–1832, Sep 1996.

[185] S. L. Hall et al. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear premRNA introns. *Science*, 271(5256):1716–1718, Mar 1996.

[186] W. Y. Tarn et al. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, 84(5):801–811, Mar 1996.

[187] T. S. Alioto. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.*, 35(Database issue):D110–115, Jan 2007.

[188] M. Q. Zhang. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, 7(5):919–932, May 1998.

[189] R. K. Bradley et al. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.*, 10(1):e1001229, Jan 2012.

[190] M. Hiller et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, 36(12):1255–1257, Dec 2004.

[191] A. Busch et al. Extensive regulation of NAGNAG alternative splicing: new tricks for the spliceosome? *Genome Biol.*, 13(2):143, 2012.

[192] A. E. Jaffe et al. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, 15(2):R31, 2014.

[193] J. J. Turunen et al. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*, 4(1):61–76, 2013.

[194] J. Yang et al. Association of DNA methylation in the brain with age in older persons is confounded by common neuropathologies. *Int. J. Biochem. Cell Biol.*, May 2015.

[195] C. Kohler et al. Release of adenylate kinase 2 from the mitochondrial intermembrane space during apoptosis. *FEBS Lett.*, 447(1):10–12, Mar 1999.

[196] G. A. Bruns et al. Adenylate kinase 2, a mitochondrial enzyme. *Biochem. Genet.*, 15(5-6):477–486, Jun 1977.

[197] T. H. Ho et al. Muscleblind proteins regulate alternative splicing. *EMBO J.*, 23(15):3103–3112, Aug 2004.

[198] C. de Martel et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.*, 13(6):607–615, Jun 2012.

[199] I. P. Tomlinson et al. Loss of heterozygosity on chromosome 11 q in breast cancer. *J. Clin. Pathol.*, 48(5):424–428, May 1995.

[200] S. E. Shackney et al. Aneuploidy in breast cancer: a fluorescence in situ hybridization study. *Cytometry*, 22(4):282–291, Dec 1995.

[201] A. A. Owainati et al. Tumour aneuploidy, prognostic parameters and survival in primary breast cancer. *Br. J. Cancer*, 55(4):449–454, Apr 1987.

[202] D. Hanahan et al. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, Mar 2011.

[203] S. Sen. Aneuploidy and cancer. *Curr Opin Oncol*, 12(1):82–88, Jan 2000.

[204] E. Manchado et al. Killing cells by targeting mitosis. *Cell Death Differ.*, 19(3):369–377, Mar 2012.

[205] D. J. Gordon et al. Causes and consequences of aneuploidy in cancer. *Nat. Rev. Genet.*, 13(3):189–203, Mar 2012.

[206] N. C. Levitt et al. Caretaker tumour suppressor genes that defend genome integrity. *Trends Mol Med*, 8(4):179–186, Apr 2002.

[207] K. A. Kwei et al. Genomic instability in breast cancer: pathogenesis and clinical implications. *Mol Oncol*, 4(3):255–266, Jun 2010.

[208] A. Balmain et al. The genetics and genomics of cancer. *Nat. Genet.*, 33 Suppl:238–244, Mar 2003.

[209] B. Orsetti et al. Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes. *Cancer Res.*, 64(18):6453–6460, Sep 2004.

[210] G. L. Ottesen et al. DNA aneuploidy in early breast cancer. *Br. J. Cancer*, 72(4):832–839, Oct 1995.

[211] G. Marx. Possible function found for breast cancer genes. *Science*, 276(5312):531–532, Apr 1997.

[212] E. Rossi et al. High-level detection of gene amplification and chromosome aneuploidy in extracted nuclei from paraffin-embedded tissue of human cancer using FISH: a new approach for retrospective studies. *Eur J Histochem*, 49(1):53–58, 2005.

[213] S. M. Johnson et al. Sporadic breast cancer in young women: prevalence of loss of heterozygosity at p53, BRCA1 and BRCA2. *Int. J. Cancer*, 98(2):205–209, Mar 2002.

[214] B. J. Miller et al. Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides comparative evidence for multiple tumor suppressors and identifies novel candidate regions. *Am. J. Hum. Genet.*, 73(4):748–767, Oct 2003.

[215] H. Schwarzenbach et al. A critical evaluation of loss of heterozygosity detected in tumor tissues, blood serum and bone marrow plasma from patients with breast cancer. *Breast Cancer Res.*, 9(5):R66, 2007.

[216] C. O'Keefe et al. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood*, 115(14):2731–2739, Apr 2010.

[217] M. Tuna et al. Prognostic value of acquired uniparental disomy (aUPD) in primary breast cancer. *Breast Cancer Res. Treat.*, 132(1):189–196, Feb 2012.

[218] P. A. Futreal et al. BRCA1 mutations in primary breast and ovarian carcinomas. *Science*, 266(5182):120–122, Oct 1994.

[219] M. Janatova et al. Novel somatic mutations in the BRCA1 gene in sporadic breast tumors. *Hum. Mutat.*, 25(3):319, Mar 2005.

[220] U. S. Khoo et al. Somatic mutations in the BRCA1 gene in Chinese sporadic breast and ovarian cancer. *Oncogene*, 18(32):4643–4646, Aug 1999.

[221] Z. Haitian et al. Mutation screening of the BRCA1 gene in sporadic breast cancer in southern Chinese populations. *Breast*, 17(6):563–567, Dec 2008.

[222] F. M. Chen et al. High frequency of somatic missense mutation of BRCA2 in female breast cancer from Taiwan. *Cancer Lett.*, 220(2):177–184, Apr 2005.

[223] M. Morley et al. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, Aug 2004.

[224] H. S. Lo et al. Allelic variation in gene expression is common in the human genome. *Genome Res.*, 13(8):1855–1862, Aug 2003.

[225] S. A. Smith et al. Allele losses in the region 17q12-21 in familial breast and ovarian cancer involve the wild-type chromosome. *Nat. Genet.*, 2(2):128–131, Oct 1992.

[226] R. S. Cornelis et al. High allele loss rates at 17q12-q21 in breast and ovarian tumors from BRCAl-linked families. The Breast Cancer Linkage Consortium. *Genes Chromosomes Cancer*, 13(3):203–210, Jul 1995.

[227] A. Osorio et al. Loss of heterozygosity analysis at the BRCA loci in tumor samples from patients with familial breast cancer. *Int. J. Cancer*, 99(2):305–309, May 2002.

[228] M. E. Thompson et al. Decreased expression of BRCA1 accelerates growth and is often present during sporadic breast cancer progression. *Nat. Genet.*, 9(4):444–450, Apr 1995.

[229] C. A. Wilson et al. Localization of human BRCA1 and its loss in high-grade, non-inherited breast carcinomas. *Nat. Genet.*, 21(2):236–240, Feb 1999.

[230] S. Staff et al. Haplo-insufficiency of BRCA1 in sporadic breast cancer. *Cancer Res.*, 63(16):4978–4983, Aug 2003.

[231] P. Van Loo et al. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.*, 107(39):16910–16915, Sep 2010.

[232] P. P. Wegman et al. Biological significance of allele specific loss of the p53 gene in breast carcinomas. *Breast Cancer Res. Treat.*, 118(1):15–20, Nov 2009.

[233] H. Yan et al. Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.*, 30(1):25–26, Jan 2002.

[234] A. T. Maia et al. Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. *Breast Cancer Res.*, 11(6):R88, 2009.

[235] T. Nguyen-Dumont et al. Detecting differential allelic expression using high-resolution melting curve analysis: application to the breast cancer susceptibility gene CHEK2. *BMC Med Genomics*, 4:39, 2011.

[236] C. Gao et al. Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis in mammary epithelial transcriptome. *BMC Genomics*, 13:570, 2012.

[237] A. Kallioniemi et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, Oct 1992.

[238] Y. Li et al. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics*, 30(15):2121–2129, Aug 2014.

[239] B. Li et al. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.*, 15(9):473, 2014.

[240] T. V. Hansen et al. High-density SNP arrays improve detection of HER2 amplification and polyploidy in breast tumors. *BMC Cancer*, 15:35, 2015.

[241] M. Perez-Enciso et al. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet. Sel. Evol.*, 47:43, 2015.

[242] F. Kaveh et al. Allele-specific disparity in breast cancer. *BMC Med Genomics*, 4:85, 2011.

[243] G. Ha et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.*, 22(10):1995–2007, Oct 2012.

[244] Q. Zhao et al. Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. *Genome Biol.*, 11(11):R114, 2010.

[245] O. Vierimaa et al. Pituitary adenoma predisposition caused by germline mutations in the AIP gene. *Science*, 312(5777):1228–1230, May 2006.

[246] K. M. Lee et al. Genetic polymorphisms of selected DNA repair genes, estrogen and progesterone receptor status, and breast cancer risk. *Clin. Cancer Res.*, 11(12):4620–4626, Jun 2005.

[247] H. S. Lo et al. Allelic variation in gene expression is common in the human genome. *Genome Res.*, 13(8):1855–1862, Aug 2003.

[248] D. C. Koboldt et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, Mar 2012.

[249] M. S. Lawrence et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, Jan 2014.

[250] L. A. Garraway et al. Lessons from the cancer genome. *Cell*, 153(1):17–37, Mar 2013.

[251] P. Creixell et al. Pathway and network analysis of cancer genomes. *Nat. Methods*, 12(7):615–621, Jun 2015.

[252] E. Garcion et al. RNA mutagenesis and sporadic prion diseases. *J. Theor. Biol.*, 230(2):271–274, Sep 2004.

[253] O. D. Iancu et al. Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics*, 28(12):1592–1597, Jun 2012.

[254] F. M. Giorgi et al. Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics*, 29(6):717–724, Mar 2013.

[255] N. Calabriso et al. Multiple anti-inflammatory and anti-atherosclerotic properties of red wine polyphenolic extracts: differential role of hydroxycinnamic acids, flavonols and stilbenes on endothelial inflammatory gene expression. *Eur J Nutr*, Feb 2015.

[256] J. Hogan et al. Transcriptional profiles underpin microsatellite status and associated features in colon cancer. *Gene*, Feb 2015.

[257] R. Al-Bahrani et al. Differential SIRT1 Expression in Hepatocellular Carcinomas and Cholangiocarcinoma of the Liver. *Ann. Clin. Lab. Sci.*, 45(1):3–9, Jan 2015.

[258] J. C. Betts et al. Gene expression changes caused by the p38 MAPK inhibitor dilmapimod in COPD patients: analysis of blood and sputum samples from a randomized, placebo-controlled clinical trial. *Pharmacol Res Perspect*, 3(1):e00094, Feb 2015.

[259] E. L. van Dijk et al. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.*, 322(1):12–20, Mar 2014.

[260] J. Durtschi et al. VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC Bioinformatics*, 14 Suppl 13:S2, 2013.

[261] S. M. Teo et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718, Nov 2012.

[262] J. D. Roberts et al. Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol. Cell. Biol.*, 9(2):469–476, Feb 1989.

[263] Z. Su et al. Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev. Mol. Diagn.*, 11(3):333–343, Apr 2011.

[264] I. Braslavsky et al. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 100(7):3960–3964, Apr 2003.

[265] T. Raz et al. RNA sequencing and quantitation using the Helicos Genetic Analysis System. *Methods Mol. Biol.*, 733:37–49, 2011.

[266] F. Ozsolak et al. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA*, 2(4):565–570, 2011.

[267] J. Quick et al. A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. *Gigascience*, 3:22, 2014.

[268] E. A. Manrao et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.*, 30(4):349–353, Apr 2012.

[269] C. Merstorf et al. Wild type, mutant protein unfolding and phase transition detected by single-nanopore recording. *ACS Chem. Biol.*, 7(4):652–658, Apr 2012.

[270] D. S. Talaga et al. Single-molecule protein unfolding in solid state nanopores. *J. Am. Chem. Soc.*, 131(26):9287–9297, Jul 2009.

[271] M. M. Mohammad et al. Controlling a single protein in a nanopore through electrostatic traps. *J. Am. Chem. Soc.*, 130(12):4081–4088, Mar 2008.

[272] J. Nivala et al. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. *Nat. Biotechnol.*, 31(3):247–250, Mar 2013.

[273] Y. Motorin et al. RNA nucleotide methylation. *Wiley Interdiscip Rev RNA*, 2(5):611–631, 2011.

[274] E. Picardi et al. Uncovering RNA Editing Sites in Long Non-Coding RNAs. *Front Bioeng Biotechnol*, 2:64, 2014.

[275] L. A. Crews et al. An RNA editing fingerprint of cancer stem cell reprogramming. *J Transl Med*, 13(1):370, Dec 2015.

[276] M. M. Ahasan et al. APOBEC3A and 3C decrease human papillomavirus 16 pseudovirion infectivity. *Biochem. Biophys. Res. Commun.*, 457(3):295–299, Feb 2015.

[277] X. Hu et al. RNA over-editing of BLCAP contributes to hepatocarcinogenesis identified by whole-genome and transcriptome sequencing. *Cancer Lett.*, 357(2):510–519, Feb 2015.

[278] L. Kang et al. Genome-wide identification of RNA editing in hepatocellular carcinoma. *Genomics*, 105(2):76–82, Feb 2015.

[279] C. M. Wei et al. Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell*, 4(4):379–386, Apr 1975.

[280] F. Rottman et al. Sequences containing methylated nucleotides at the 5' termini of messenger RNAs: possible implications for processing. *Cell*, 3(3):197–199, Nov 1974.

[281] P. Narayan et al. An in vitro system for accurate methylation of internal adenosine residues in messenger RNA. *Science*, 242(4882):1159–1162, Nov 1988.

[282] T. Csepany et al. Sequence specificity of mRNA N6-adenosine methyltransferase. *J. Biol. Chem.*, 265(33):20117–20122, Nov 1990.

[283] P. Narayan et al. Context effects on N6-adenosine methylation sites in prolactin mRNA. *Nucleic Acids Res.*, 22(3):419–426, Feb 1994.

[284] Y. Fu et al. Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, 15(5):293–306, May 2014.

[285] K. M. Lonergan et al. Predicted editing of additional transfer RNAs in Acanthamoeba castellanii mitochondria. *Nucleic Acids Res.*, 21(18):4402, Sep 1993.

[286] L. Agranat et al. The editing enzyme ADAR1 and the mRNA surveillance protein hUpf1 interact in the cell nucleus. *Proc. Natl. Acad. Sci. U.S.A.*, 105(13):5028–5033, Apr 2008.

[287] D. Speijer. Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. *Bioessays*, 33(5):344–349, May 2011.

[288] Y. Shu et al. Predicting A-to-I RNA editing by feature selection and random forest. *PLoS ONE*, 9(10):e110607, 2014.

[289] H. Q. Zhao et al. Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. *Genome Res.*, 25(1):66–75, Jan 2015.

[290] J. M. Toung et al. Detection theory in identification of RNA-DNA sequence differences using RNA-sequencing. *PLoS ONE*, 9(11):e112040, 2014.

[291] M. A. Huntley et al. Dissecting gene expression at the blood-brain barrier. *Front Neurosci*, 8:355, 2014.

[292] J. Gong et al. Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing. *RNA Biol*, 11(11):1375–1385, Nov 2014.

[293] S. Haider et al. Integrated analysis of transcriptomic and proteomic data. *Curr. Genomics*, 14(2):91–110, Apr 2013.

[294] L. Nie et al. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit. Rev. Biotechnol.*, 27(2):63–75, 2007.

[295] M. Hecker et al. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems*, 96(1):86–103, Apr 2009.

[296] S. Rogers. Statistical methods and models for bridging Omics data levels. *Methods Mol. Biol.*, 719:133–151, 2011.

[297] S. P. Gygi et al. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, 19(3):1720–1730, Mar 1999.

[298] G. Chen et al. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell Proteomics*, 1(4):304–313, Apr 2002.

[299] L. E. Pascal et al. Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics*, 9:246, 2008.

[300] A. Ghazalpour et al. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.*, 7(6):e1001393, Jun 2011.

[301] N. T. Ingolia et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, Apr 2009.

[302] C. Gustafsson et al. Codon bias and heterologous protein expression. *Trends Biotechnol.*, 22(7):346–353, Jul 2004.

[303] T. Steijger et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, 10(12):1177–1184, Dec 2013.

[304] H. Wang et al. Detection of allelic imbalance in gene expression using pyrosequencing. *Methods Mol. Biol.*, 373:157–176, 2007.

[305] M. Pirinen et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics*, Mar 2015.

[306] K. R. Kukurba et al. Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.*, 10(5):e1004304, May 2014.

[307] H. Roger et al. [Non-familial Duchenne-Erb myopathy appearing after encephalitis in late childhood; association with mental deficiency and cutaneous dystrophy]. *Bull Mem Soc Med Hop Paris*, 67(7-8):279–281, 1951.

[308] E. A. Cheeseman et al. The sex ratio of mutation rates of sex-linked recessive genes in man with particular reference to Duchenne type muscular dystrophy. *Ann. Hum. Genet.*, 22(3):235–243, May 1958.

[309] C. L. Bladen et al. The TREAT-NMD DMD Global database: Analysis of More Than 7000 Duchenne Muscular Dystrophy Mutations. *Hum. Mutat.*, Jan 2015.

[310] A. De Sandre-Giovannoli et al. Lamin a truncation in Hutchinson-Gilford progeria. *Science*, 300(5628):2055, Jun 2003.

[311] B. P. O'Sullivan et al. Cystic fibrosis. *Lancet*, 373(9678):1891–1904, May 2009.

[312] V. Waters et al. Cystic fibrosis microbiology: Advances in antimicrobial therapy. *J. Cyst. Fibros.*, Feb 2015.

[313] A. Warshel et al. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, May 1976.

[314] M. T. Panteva et al. Multiscale methods for computational RNA enzymology. *Meth. Enzymol.*, 553:335–374, 2015.

[315] D. Xu et al. AIDA: Ab Initio Domain Assembly for Automated Multi-domain Protein Structure Prediction and Domain-Domain Interaction Prediction. *Bioinformatics*, Feb 2015.

[316] J. M. Chandonia et al. The impact of structural genomics: expectations and outcomes. *Science*, 311(5759):347–351, Jan 2006.

[317] D. Baker et al. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, Oct 2001.

[318] S. Goldsmith-Fischman et al. Structural genomics: computational methods for structure analysis. *Protein Sci.*, 12(9):1813–1821, Sep 2003.

[319] S. A. Lesley et al. Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl. Acad. Sci. U.S.A.*, 99(18):11664–11669, Sep 2002.

[320] M. O. Duff et al. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*, 521(7552):376–379, May 2015.

[321] Z. K. Hassan-Smith et al. Gender-specific differences in skeletal muscle 11B-HSD1 expression across healthy aging. *J. Clin. Endocrinol. Metab.*, page jc20151516, May 2015.