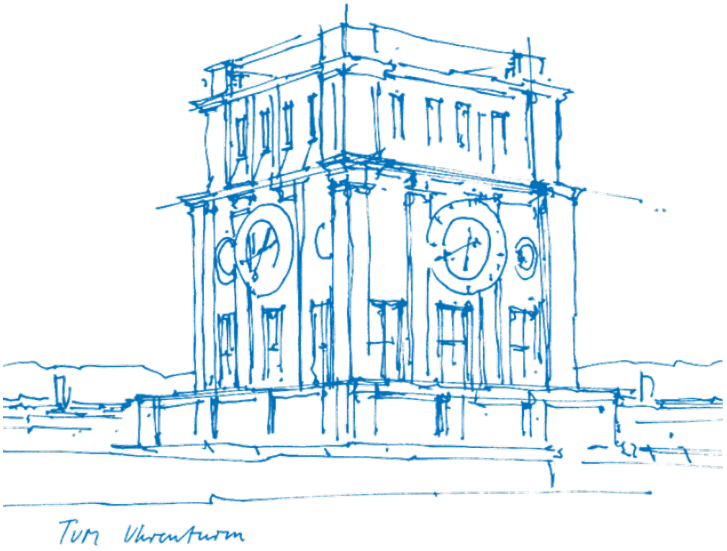
TUM School of Life Sciences
Technical University of Munich

Towards Reproducibility in Alternative Splicing Analysis

Amit M. Fenn



TUM School of Life Sciences
Technische Universität München

Towards Reproducibility in Alternative Splicing Analysis

Amit M. Fenn

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):
Prof. Dr. Mathias Wilhelm

Prüfer der Dissertation:
1. Prof. Dr. Dmitrij Frischmann
2. Prof. Dr. Jan Baumbach

Die Dissertation wurde am 11.11.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 10.01.2023 angenommen.

To my family, for cultivating the spirit of scientific debate over dinners.

Abstract

Alternative splicing (AS) is a type of ribonucleic acid (RNA) editing that occurs naturally within cells and gives rise to a greater protein diversity from a set of genes. Transcriptomic analysis is limited often to the gene level, as it is challenging to confirm experimentally. A lack of ground truth further brings a lack of consensus in reported AS events by the computational methods used in Ribonucleic Acid Sequencing (RNA-Seq) approaches. Benchmarks of the computational methods are often updated by tool developers who have a vested interest in publishing better reports of AS events by their tool than that of their peers. A growing standard of reported metrics within AS event detection tools also challenge a consensus from forming by reporting diverging metrics such as isoform expression, exon expression or events in AS. Even when reporting just AS events, reports may use splice graphs to describe observed exons or report events that use overcomplicated or redundant terminology. These challenges often leave analysts unsure of how to report AS as observed within a transcriptome.

Gold standard transcriptomic datasets were simulated by a simulator we developed called ASimulatoR. The simulator created datasets with modulated quantities of events spanning each type of AS. ASimulatoR also explicitly reported the simulated AS events, so that the dataset may be used as ground truth to analyse the performance of mapping and AS event detection tools.

To address the lack of consensus within reported AS events, we developed DICAST, A Docker Integrated Comparison of AS Tools (available at: https://github.com/CGAT-Group/DICAST). We benchmarked 11 mapping and 8 AS event detection algorithms used to report AS with simulated data with increasing complexity of AS and analysed them with these tools. We also compared these results with a simulated dataset where the distribution of AS events follows that of a real dataset obtained from the Study of Health in Pomerania (SHIP) cohort.

DICAST is an independent workflow for benchmarking for tool developers. Much too often, benchmarks developed by tool developers are confronted by conflict of interest due to the pressure in publishing tools with better reported AS events than current standards. With DICAST, integrating a new tool becomes a modular task that does not affect other tools and maintains an unbiased approach to benchmarking AS events. DICAST also reports what events are found in common with other tools and which ones are unique. When using simulated datasets, DICAST also reports precision and recall plots per event type. This allows one to determine which tools work best in identifying AS events for any organism in study.

Diverging metrics of AS were used in concert to the objective of studying the effect of splicing by protein binding in a web server named DASiRe or Direct Alternative Splicing Regulator predictor (available at: https://github.com/marisolsalb/dasire). Functional studies of splicing factors that include experiments with matching RNA-Seq and Chromatin Immunoprecipitation Sequencing (ChIP-Seq) experiments could be analysed in DASiRe to identify which proteins are bound to genomic regions that exhibit splicing.

We used DASiRe to analyse knockout experiments of candidate protein YBX-1 which is shown to be involved in the mechanism and regulation of AS. DASiRe uses metrics such as isoform usage, exon usage and AS events in concert to detect AS and compares by genomic loci to binding of proteins in study. DASiRe reports Fisher's enrichment results to ask if the genomic regions that show binding by a protein of interest also show more splicing, compared to other genes. DASiRe is implemented as a web tool for visualisations and a pipeline for distributed pre-processing RNA-Seq data. DASiRe can be used by a bioinformatic analyst with basic knowledge of computing platforms. With the development of ASimulatoR and DICAST, benchmarking of AS event tools remains independent. A unified format within DICAST is used to report AS events across tools, reporting AS events by genomic position, to make reports more robust. Reproducible RNA-Seq analysis pipelines were used to report diverging metrics of AS and visualised in a web server, DASiRe. This approach demonstrates the use of genomic positions of spliced genes to observe a dependent event, the binding of potential splicing factors. This thesis therefore builds reproducible elements that unify diverging standards of reporting AS within the study of AS.

Zusammenfassung

Alternatives Spleißen (AS) ist eine Art von Ribonukleinsäure (RNA)-Editierung, die natürlicherweise in Zellen vorkommt und zu einer größeren Proteinvielfalt aus einer Reihe von Genen führt. Transkriptomische Analysen beschränken sich häufig auf die Genebene, da sie experimentell schwer zu bestätigen sind. Das Fehlen einer Basiswahrheit führt außerdem zu einem Mangel an Konsens bei den gemeldeten AS-Ereignissen durch die in RNA-Seq-Ansätzen verwendeten Berechnungsmethoden. Die Benchmarks der Berechnungsmethoden werden häufig von den Entwicklern der Tools aktualisiert, die ein Interesse daran haben, dass die von ihrem Tool gemeldeten AS-Ereignisse besser ausfallen als die ihrer Kollegen. Ein zunehmender Standard der gemeldeten Metriken innerhalb der Tools zur Erkennung von AS-Ereignissen stellt ebenfalls eine Herausforderung für die Konsensbildung dar, da abweichende Metriken wie Isoform-Expression, Exon-Expression oder AS-Ereignisse gemeldet werden. Selbst wenn nur AS-Ereignisse gemeldet werden, können Berichte Spleißdiagramme verwenden, um beobachtete Exons zu beschreiben, oder Ereignisse melden, die eine überkomplizierte oder redundante Terminologie verwenden. Diese Herausforderungen lassen Analysten oft unsicher zurück, wie sie die in einem Transkriptom beobachteten AS berichten sollen.

Goldstandard-Transkriptomdatensätze wurden mit einem von uns entwickelten Simulator namens ASimulatoR simuliert. Der Simulator erzeugte Datensätze mit modulierten Mengen von Ereignissen, die jede Art von AS abdeckten. ASimulatoR meldete die simulierten AS-Ereignisse auch explizit, so dass der Datensatz als Basiswahrheit für die Analyse der Leistung von Mapping- und AS-Ereigniserkennungsprogrammen verwendet werden kann.

Um den Mangel an Konsens innerhalb der gemeldeten AS-Ereignisse zu beheben, entwickelten wir DICAST, A Docker Integrated Comparison of AS Tools (verfügbar unter: https://github.com/CGAT-Group/DICAST). Wir haben 11 Mapping- und 8 AS-Ereigniserkennungsalgorithmen, die zur Meldung von AS verwendet werden, mit simulierten Daten mit zunehmender Komplexität von AS verglichen und sie mit diesen Tools analysiert. Wir haben diese Ergebnisse auch mit einem simulierten Datensatz verglichen, bei dem die Verteilung der AS-Ereignisse der eines realen Datensatzes aus der SHIP-Kohorte entspricht.

DICAST ist ein unabhängiger Arbeitsablauf für Benchmarking für Tool-Entwickler. Viel zu oft sind die von den Entwicklern entwickelten Benchmarks mit Interessenkonflikten konfrontiert, da sie unter dem Druck stehen, Tools zu veröffentlichen, deren gemeldete AS-Ereignisse besser sind als die aktuellen Standards. Mit DICAST wird die Integration eines neuen Tools zu einer modularen Aufgabe, die andere Tools nicht beeinträchtigt und einen unvoreingenommenen Ansatz für das Benchmarking von AS-Ereignissen beibehält. DICAST meldet auch, welche Ereignisse mit anderen Tools übereinstimmen und welche einzigartig sind. Bei der Verwendung von simulierten Datensätzen gibt DICAST auch Präzisions- und Recall-Diagramme pro Ereignistyp aus. Auf diese Weise lässt sich feststellen, welche Tools bei der Identifizierung von AS-Ereignissen für jeden untersuchten Organismus am besten funktionieren.

Unterschiedliche AS-Metriken wurden in einem Webserver namens DASiRe oder Direct Alternative Splicing Regulator Predictor (verfügbar unter: https://github.com/marisolsalb/dasire) zusammen mit dem Ziel verwendet, die Auswirkungen des Spleißens durch Proteinbindung zu untersuchen. Funktionelle Studien von Spleißfaktoren, die Experimente mit übereinstimmenden RNA-Seq- und ChIP-Seq-Experimenten umfassen, könnten in DASiRe analysiert werden, um zu ermitteln, welche Proteine an Genomregionen gebunden sind, die Spleißen aufweisen.

Wir haben DASiRe verwendet, um Knockout-Experimente des Kandidatenproteins YBX-1 zu analysieren, von dem gezeigt wurde, dass es am Mechanismus und der Regulierung von AS beteiligt ist. DASiRe verwendet Metriken wie Isoformnutzung, Exon-Nutzung und AS-Ereignisse, um AS zu erkennen und vergleicht genomische Loci mit der Bindung der untersuchten Proteine. DASiRe berichtet Fisher's Anreicherungsergebnisse, um herauszufinden, ob die genomischen Regionen, die eine Bindung durch ein Protein von Interesse zeigen, auch mehr Spleißungen aufweisen, verglichen mit anderen Genen. DASiRe ist als Webtool für Visualisierungen und eine Pipeline für die verteilte Vorverarbeitung von RNA-Seq-Daten implementiert. DASiRe kann von einem Bioinformatik-Analysten mit Grundkenntnissen über Computerplattformen verwendet werden. Mit der Entwicklung von ASimulatoR und DICAST bleibt das Benchmarking von AS-Event-Tools unabhängig. Ein einheitliches Format innerhalb von DICAST wird verwendet, um AS-Ereignisse für alle Tools zu melden, wobei AS-Ereignisse nach genomischer Position gemeldet werden, um die Berichte robuster zu machen. Reproduzierbare RNA-Seq-Analysepipelines wurden verwendet, um divergierende Metriken von AS zu melden und in einem Webserver, DASiRe, zu visualisieren. Dieser Ansatz demonstriert die Verwendung genomischer Positionen von gespleißten Genen zur Beobachtung eines abhängigen Ereignisses, der Bindung potenzieller Spleißfaktoren. In dieser Arbeit werden daher reproduzierbare Elemente geschaffen, die unterschiedliche Standards für die Berichterstattung über AS innerhalb der AS-Studie vereinheitlichen.

Contents

1 Introduction

1.1 Introduction to Alternative Splicing

The discovery of Deoxyribonucleic Acid (DNA) as the hereditary material [7] of all life as we know it, put DNA in the forefront of the scientific race in molecular biology, mid 20th century. This eventually led to the elucidation of the structure of DNA by Rosalin Franklin [75]. Albrecht Kossel had, by the start of the 20th century, already contributed to discovery that DNA is a polymer made up of units of nucleic acids, adenine (A), cytosine (C), guanine (G) and thymine (T) (and also uracil (U) in ribonucleic acid (RNA)) [64]. The sequence of these nucleic acids within the polymer form the basis of extracting meaning from biological samples in the high throughput sequencing revolution we are in today.

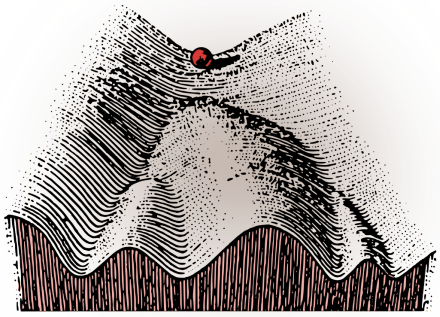


Figure 1.1 Canalization as analogised by Waddington [163].
Suggests that like a ball going downhill and into a canal, so do cellular programs lead to different cell-types being formed right next to each other.

The central dogma of molecular biology as proposed by Sir Francis Crick, who is also accredited for the elucidation of the structure of DNA, describes the transfer of these sequences from DNA to protein products [23]. This makes DNA the inherited blue-print and the protein composition of the cell the final product of what is referred to as the translation machinery. At the time, however, it was a mystery how the genetic information translated into proteins. The discovery of the unstable molecule, the messenger ribonucleic acid (mRNA), came later from studies on bacteria and bacteriophages [62, 22]. mRNA is an intermediate smaller polymer that contains the sequences of nucleotides A, U, G and C, corresponding to a small region of the DNA within chromosomes. These messages have to find their way from the eukaryotic nucleus of a cell to its cytoplasm to be translated to proteins.

The word ‘gene’ was coined by W. Johannsen, who at the time did not refer to its physical properties, but was abstracting concepts that were already laid out by Gregor Mendel, the father of genetics [132, 63, 111]. Gregor Mendel was referring to heritable traits of pea plants and set the field on the very early grounds of trying to separate traits he observed. While Gregor Mendel was not aware of the mechanism of this separation, Robin Holiday described the mechanism of this separation in yeast half a century later

[58]. However, Mendel's concept of traits being linked, was setting the stage for the theory of collinearity proposed by Dounce and Gamow [32, 46]. After the structure of DNA was understood, this theory compared the linear nature of DNA to the linear nature of proteins, proposing that these elements shared grammar as if they were long words and that through the translation process, the sequences maintain a flow of information.

Multicellular organisms tend to consist of specialised cells. This means that for the same DNA sequence, cells within the organism tend to have unique profiles of protein compositions. CH Waddington, most known for his work in epigenetics and developmental biology, compares cells undergoing differentiation, to a marble rolling down a hill guided by the valleys and canals, to its final destination as the ball finally settles (Fig:1.1). He also describes the smaller variations in the state of the cells leading to a determined fate, like water guided within canals flowing down a hill. 'Canalization' was used to describe the tendency of each cell's developmental genetic programs and their consequences as clear distinctions of cell types that arise within tissues. This is how we get from on top of the hill, as generic embryonic stem cells, through intermediates to final cell types that describe adult tissues [163]. Theoretically, every state of a differentiating cell is a position on the landscape or canals of epigenetics and each final state of a differentiated cell to be unique. Therefore describing the elements that regulate gene expression could also help understand the tendencies for a cell's fate, or which elements of metabolism can a cell respond with.Unique protein compositions within the cell are modulated by transcription, the process of converting DNA elements to RNA messages and translation, the process of converting RNA messages to protein sequences.

RNA forms the template for protein translation by ribosomes. Genes within the genome form a stable polymer of DNA which acts as a blueprint for all the protein sequences within a cell. The transcription cycle consists of 3 phases: initiation, elongation and termination [140]. Promoter sequence elements such as the TATA box are required for a gene to recruit the pre-transcription initiation complexes [66, 148]. In initiation, the promoter region of the DNA separates into two strands with the help of single strand DNA binding proteins and enzymes such as helicases, topoisomerases, and the template strand then slips into RNA Polymerase II, the enzyme that synthesis RNA polymers in most eukaryotic organisms [124]. The elongation phase for transcription begins with the polymerase leaving the promoter regions and maturing its structure to be locked on the template DNA strand. This state of the polymerase can express an entire gene until it reaches the termination [131]. RNA-Polymerisation rates have been shown to vary three fold, with many protein associations to the polymerase such as small nuclear ribonucleoproteins (snRNP) or chromatin elements that DNA is wrapped around [160]. Transcription termination occurs when the RNA-Polymerase II enzyme reaches the end of the open reading frame and changes its conformation, releasing the polymerized RNA-particle [140]. The unwinding of double stranded DNA and the conformational changes within what's called the transcription bubble dictates that only a small region of a genome undergoes transcription.

How the DNA is organised physically within the nucleus, determines which subset of genes are expressed within a certain cell type. The double helix DNA strand is tightly wound around octameric histone protein complexes, made up of 2 of histones H2A, H2B, H3 and H4 each [43]. Histones are positively charged proteins that are attracted to the negatively charged DNA backbone. They can be dynamically unbound and rebound anywhere across the genome. This is how a DNA strand, that is typically a few me-

tres long, is organised and compacted within the cell. This complex of nucleic acid and proteins is called the unit of the nucleosome model. 147 base pairs of DNA wrapped helically and tightly over a histone core are kept in place with a histone H1. Unbound DNA then extends to the next nucleosome. When not required, big sections of the DNA are tightly wound to the histones to form their heterochromatin state. When unwound, the regions free of histone can assemble the transcription machinery and express their genes. These regions are euchromatin. Together these fibres are referred to as the chromatin filament. This influence of proteins on the selection of active genes is referred to as epi-genetics, to account for the elements of regulation that are not within the set of genes described in the genome. This contrasts the central dogma of molecular biology where information flows from DNA to RNA to proteins.

Translation is the process where a messenger RNA gets translated to proteins. The site of translation is in the cytoplasm for eukaryotic organisms and here nascent protein sequences are seen to arise from strands of mRNA that are bound to ribosomes. mRNA particles have a non-uniform affinity to translation, where mRNA that are more prone to translation are seen to have many ribosomes attached to them, such complexes are polysomes [17]. Ribosomes are ribonucleic particles that have about two major components, the 80S and 60S component, named after their molecular weights in the Svedberg unit. The large subunit is also made up of 3 ribosomal RNA (rRNA) and around 50 proteins. The smaller subunit is made of one rRNA and around 35 proteins [70]. At the ribosomes, monomers of proteins, amino acids polymerise together guided by transfer RNA (tRNA). tRNA need to match to the template mRNA, briefly within the ribosomes to enable a sequential transfer of sequences of amino acids that follow the mRNA sequence. The matching sequence, referred to as a triplet codon varies across the different kingdoms and even within a cell, when considering the codon usage within the mitochondria [34, 119, 108]. This way, triplet sequences of DNA elements go through an RNA intermediate to build proteins with a corresponding sequence. Both transcription and translation together form the central dogma of molecular biology, a paradigm where sequence fidelity is maintained collinearly.

Early RNA metabolism studies tried to inquire about the fate of the already existing RNA. Having developed the capacity to stop the steady flow of transcription, and with pulse radiolabelling the nucleotides, Harris H was able to show that mRNA within the nucleus degraded faster and did not leave the nucleus, and the mRNA from the cytoplasm had a different composition of nucleotides [45, 54, 55, 168]. Darnell JE identified that the mRNA is translated into proteins in the cytoplasm [47]. With the discovery of a polyA tail and a methyl cap defining the boundaries of a both nuclear and cytoplasmic mRNA, and knowing that they protect the mRNA from rapid degradation, Darnell's group, along with others were puzzled by the depletion of RNA almost 10 fold for the largest mRNA from the nucleus vs the cytoplasm, given that the mRNA's 3' and 5' ends were protected [24, 139]. This means that regions of the gene between the 3' and 5' ends were split before they reached the cytoplasm.

The discovery of split genes by Philip Sharp and Richard Roberts was awarded the Nobel Prize in 1993 for their studies in the expression of genes within adenovirus [154]. Philip Sharp's experiment used electron microscopy of hybridised mature mRNA from viral particles and the viral genome to show three loops of DNA, where the DNA-RNA sequences mismatched, suggesting that mature gene was composed of different segments of the gene these RNA products were derived from. They concluded from their experiments that within higher organisms, protein products within the DNA were organised in discrete sections, separated by DNA that was back then thought to be irrelevant. By selecting which regions of

the mRNA was excluded or included, sequences within the mRNA diverged to give proteins with different combinations of amino acid sequences. This was the first time that the idea of collinearity was challenged.

AS is reconfiguration of sections of genes to give rise to a much more diverse set of protein products greatly surpassing the number of genes in the DNA of an organism. Regions that are usually executed by the translational machinery to proteins are exons. Regions that were excluded giving a template mRNA strand, are called introns. AS reconfigures exons to make unique combinations. The resulting multiple mRNA are referred to as transcripts. Each protein is the product of a unique transcript. The resulting multiple proteins are referred to as isoforms.

1.2 Mechanisms of Splicing

The mechanism of AS is a complex, multi-stage process that involves a collection of small nuclear snRNP. snRNP are complexes made of RNA elements and protein elements that work in concert. Here the uridine-rich RNA elements are bound by weak bonds in many places to positively charged protein scaffolds. RNA elements also host a 3D structure by themselves which act primarily as scaffolds themselves and host some catalytic sites. The protein elements diversify the functional elements of the snRNP by bringing more catalytic sites to the complex.

Regions where splicing cut the mRNA strand are referred to as splice sites and every intron that is spliced out has two splice sites, one at the 3' end of the intron and one at the 5' end of the intron. These regions are referred to as 3' and 5' relative to the carbon within the ring of ribose sugar that along with phosphate groups make the phosphate backbone that makes the spine of these polymers. Splice sites tend to follow the GT-AG rule: the first nucleotides of intron (5' or the donor splice site) is GT, and the last two nucleotides of intron are AG (3' or the acceptor splice site) in the human genome. The profile of the 5' splice site tends to look like nAGGTRAGt [172]. Cleavage at the 5' site depends not only on the 5' splice site sequence, but the sequence of the branchpoint and the sequence roughly 30 nucleotides (nt) after the branchpoint.

Splicing is a two step process [81]: The first stage is a lariat formation and the second step is an exon ligation. Each of these stages are catalysed by a unique collection of proteins and RNA that make up the various spliceosomal complexes in the splicing process.

The formation of a lariat requires a ribose sugar bound to the phosphate backbone in three of the carbon molecules that make the ribose sugar [138]. This base is referred to as the branch point and has a yUnAy motif around it, with the ribose base of Adenine serving as the branch point.

U1 and U2 small nuclear RNA (snRNA) within their respective snRNP initiate the formation of splicing complexes by base-pairing with the 5' splice site and branch points, respectively (see Fig:1.2B). U1 and U2 snRNP can bind co-transcriptionally, or while the gene is being transcribed from DNA to RNA. U1 and U2 snRNP usually binds to the transcribing mRNA and act as guides for further recruitment of the spliceosome. snRNA U4, U5 and U6 tri-snRNP found in the catalytic centre of the spliceosome, structurally and functionally resemble the earliest and simplest form of split genes, the self splicing introns. This similarity, which suggests that the catalytic core of the spliceosome perhaps had its evolutionary roots within the RNA elements of the spliceosome [42, 135]. The nucleophilic reactions that partition the RNA-

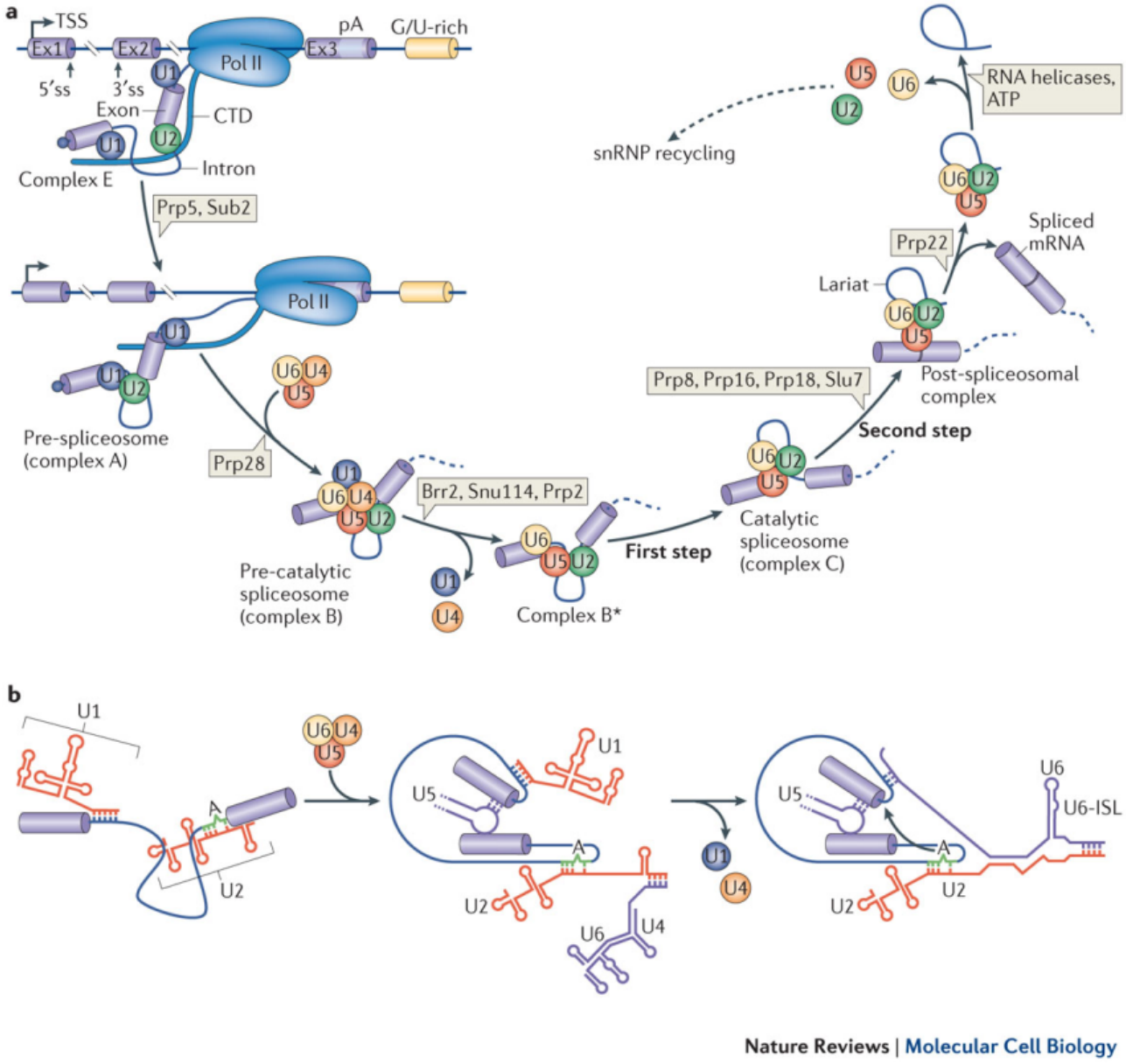


Figure 1.2 Assembly of a spliceosome [105].

A.Splicing requires snRNP for it's mechanism. It is a two step process first the formation of the lariat and second the ligation of the two bordering exons. **B**. The splice site is first recognising by small nuclear ribonucleoacids (snRNA), implying that you need both snRNA and snRNP for splicing.

phosphate backbone are facilitated with the help of magnesium ions that are harboured within the single catalytic site of the U6-U2 snRNP complex [81].

The minor splicing pathway, with the U12 spliceosome accounts for less than 0.4% of human introns being spliced [36]. The minor splicing pathway seems to be as old as the major splicing pathway and have very conserved host proteins and products. One such product, srsf10, acts as an element of the major pathway that could in turn regulate the major splicing pathway as well [110]. The minor spliceosome shares a lot of features with the U2 spliceosome of the major splicing pathway and therefore only the U2 spliceosome will be discussed in this thesis.

1.3 Regulation of Splicing

In humans, as in vertebrates, exons are separated by intronic regions that are on average 3,365 base pairs (bp) long. Intronic sequence features may thus span too far a distance for protein interactions of individual splicing factors within the spliceosome. Most exons are on average 145bp long, and are occupied by nucleosomes. The splicing machinery uses features such as sequences that flank exonic regions to recognise splice sites rather than intronic sequence features. This recognition of splice sites based on exonic features for the assembly of the pre-spliceosomal complex at the acceptor splice sites is termed as exon definition [60, 169, 151]. In the evolutionary perspective, exon definition explains how mutations in a cis regulatory element such as a splice site in the scenario of “large intron - small exon” usually results in inclusion or exclusion of an exon, rather than retention of the affected intron. As a consequence, exon definition leads to a more viable transcript set.

The following sections describe how various factors involved in gene expression have also been shown to play a role in splicing. These mechanisms can be broadly generalised as essential processes in the exon definition paradigm.

1.3.1 Influence of RNA Polymerase

Assembly of splicing snRNP begins co-transcriptionally, however complete excision of introns may take longer[95]. Especially, for longer genes with large introns, splicing has been observed to occur post-transcriptionally, and are often influenced by the speed of RNA Polymerase [25].

The speed of RNA Polymerase enzyme has been observed to have a strong effect on splicing. Genes transcribed using a slow acting polymerase have been shown to express different exon sets as compared to RNA-Polymerase II [104, 33]. The slower polymerase is also shown to be involved in exon inclusion, the choice of alternative 5’ splice site but a generic model for these mechanisms is currently unclear[121].

The Carboxyl-terminal domain (CTD) of RNA-Polymerase II has also been found associated to splicing factors such as U1 snRNP of the pre-spliceosomal complex and has shown to be required for co-transcriptional splicing [10].

Systematic studies of nascent RNA with RNA-Seq (described in Section:1.5.1) show depletions in read coverage from 5’-to-3’ within introns, forming a decreasing slope which sharp increase at the exons showing that splicing occurs co-transcriptionally[5].

1.3.2 Influence of the nucleosome

Nucleosomes wrap 147 bp of DNA wrapped around histone cores. This genomic window is correlated to the size of exons commonly found in the human genome, suggesting that nucleosomes play a selective role in exon inclusion [60].

Histones are shown to preferentially bind to exonic regions of genes, rather than intronic regions. They are also observed to be rare in intronic regions next to splice sites [122]. Even outside the exon definition paradigm, polypyrimidine sequences found near the 3' splice sites hinder binding of histones[142].

Chromatin complexes are often associated with the pre-spliceosomal complex. To give an example, the STAGA complex hosts a histone acetyltransferase, which interacts with the U2 snRNP. Histone acetyltransferases euchromatinise genes to initiate transcription[103].

1.3.3 Influence of Chromatin Modifications

Aside from nucleosome positioning and chromatin complexes, specific modifications on histone tails have shown to regulate splicing. For example, chromatin remodelling enzyme CHD1, interactions with H3K4me3(tri-methylation at the K4 position on histone 3) has also been shown to functionally interact with U2 snRNP [146], engaging elements of spliceosomes for more efficient splicing.

H3K27me2 (Dimethylation at K9 and trimethylation at K27 on histone 3) has been shown to recruit heterochromatin associated protein HP1-α, which could slow down transcription by RNA polymerase II [3].

As an example of polypyrimidine tract binding (PTB) dependent exon inclusion events: H3K36me3(tri-methylation at the K36 position on histone 3) is usually observed to be enriched on actively transcribing gene bodies. H2K26me3(tri-methylation at the K26 position on histone 2) has shown to play a role in the selection of alternative splicing sites by recruiting splicing regulators such as SR proteins or the U2 snRNP. The correlation between H3K36me3 modifications and PTB dependent exon usage have been shown to be independent of cell type [96]. Depletion of H3K4me3 is observed on genes that have H3K36me2 modifications when a histone tail-binding protein Mortality Factor(MORF)-related gene 15, MRG15, was overexpressed. This led to PTB-dependent exons being excluded.

1.4 Types of Alternative Splicing

Exon skipping is the most common event in vertebrates. Usually this event affects one exon but other scenarios are possible. Multiple exon skipping, describes where more than one consecutive exon is spliced out. Mutually exclusive exons, another exon skipping pattern, describes the event where the evidence of a spliced exon is anticorrelated with a usually adjacent exon being retained. Also referred to as alternative exon usage, alternative 5' splice sites and alternative 3' splice sites occur when an exon boundary is not always clear. Alternative first exon skipping, which gives rise to proteins with different N terminals and alternative last exon skipping, which gives rise to proteins with different C terminals (see Fig: 1.3.A).

To initiate a splicing process, a spliceosome recognizes special motifs close to the exon-intron boundary: splice sites, branch point, and polypyrimidine tract. The cleavage is also influenced by the presence of a polypyrimidine tract after the branch point.

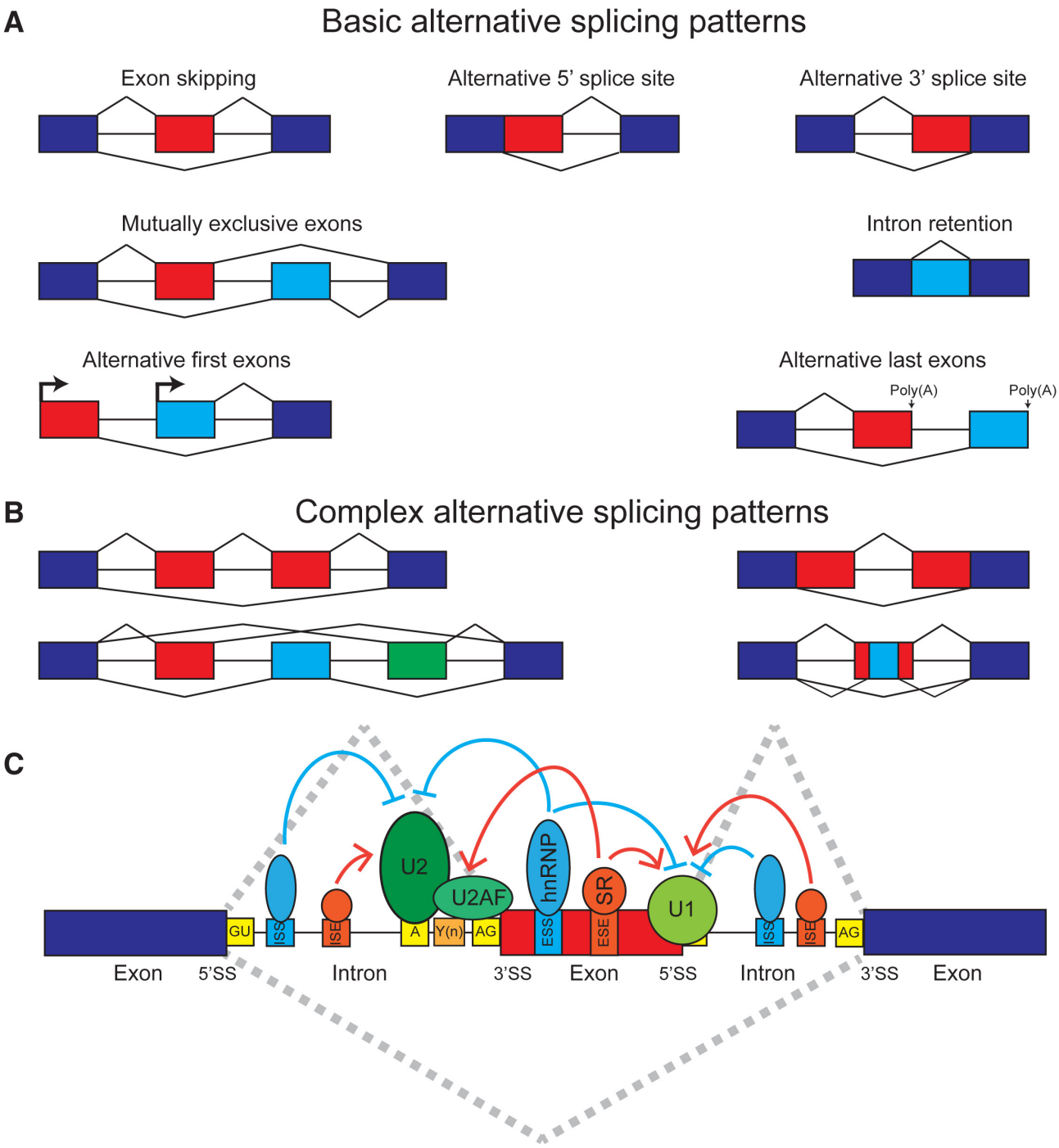


Figure 1.3 Types of AS [128].

A: Basic AS patterns, **B**: Complex AS patterns, **C**: Splicing factors identify sequence features on exons or introns for binding and formation of pre-spliceosomal complex.

The donor or 5' splice sites are recognized by the SF2/ASF splicing factors which later recruit the U1 component of a spliceosome. Some exons have multiple potential splice sites and SF2/ASF is responsible for 5'-splice site selection [37, 2]. The use of different 5' splice sites within an exon is referred to as alternative 5' splicing. (shown in Fig: 1.3.C)

The 3' splice site is usually the first AG dinucleotide downstream of the polypyrimidine tract which is usually marked by (C/U/A)AG and is first recognised by U5 snRNP [19]. 3' intronic splice sites, unfortunately, aren't easy to recognise for the rest of the spliceosome, especially if the intron is longer than 250 nucleotides, which is the more frequently seen gene model in vertebrates. Therefore, the recognition of the 3' splice site is disconnected from both the recognition of the 5' splice region as well as the formation of the major spliceosome [149]. Proteins found in the heterogeneous nuclear protein complexes bound to transcribing mRNA or heterogenous ribonucleoproteins (hnRNP) influence which candidate 3' splice sites are selected for further splicing [87]. The use of alternative positions of attachment to the spliced exon to the downstream exon with respect to different 3' splice sites is referred to as alternative 3' splicing.

Besides alternative 5'- and 3'-splice sites, two more types of AS are usually defined. When an intron is retained after splicing, it's referred to as intron retention. Intron retention is commonly found along with these sequential features are weaker, such as the use of cryptic splice sites [116]. Retained introns are often also GC rich, which is a characteristic feature of exons. GC rich regions often form secondary structures and are protected from the binding of splicing factors, thereby impeding splicing. Another key feature of retained introns is that they're usually much smaller (<200 bp). Serine-arginine rich protein binding sites are also found to be enriched in retained introns. Genes with potential Intron retention are also considered to host a higher density of putative mRNA binding sites, suggesting that they play a role in the regulation of the gene.

Intron retention can occur in the context of exon definition, because of the mutations in the splice-sites that lead to an incomplete assembly of the splicing machinery [9]. In situations of higher GC content and shorter introns initially tend to be recognised as an intron definition scenario are more prone to intron retention [35, 90, 14]. In an example of weaker splice sites being associated with intron retention, a study on erythroid differentiation showed a negative correlation between retained introns and splice site strength [130]. However, intron retention could also be caused by an incomplete assembly of the spliceosome. In these cases, protein elements bound on mRNA transcripts could act as a plug, preventing the export of the mRNA through the nuclear pore [152]. This has been demonstrated on the genes in response to PTBP1 where 5' splice sites are repressed by the stabilisation of U1 snRNP by the PTB1 protein element [144]. A similar mechanism has been shown in the gene of a poly A binding protein, PABPN1, where the binding of this protein to the 3' untranslated region (UTR) region of its own gene regulates a homeostatic feedback loop [8]. Retained introns can also be used as a mechanism for a late signal induced splicing, for example: for introns that are spliced upon the CLK-Kinase signalling, Serine and Arginine-rich (SR) proteins involved in splice site recognition are hyper-phosphorylated prematurely before it associates with the spliceosome. The reduction of CLK activity or an increased phosphatase activity leads to these introns being spliced and the export of these mRNA for translation [12].

Intron retention has so many mechanisms of regulation, but the fate of these mRNA are still widely speculative. mRNA with retained introns are suspected to leave the nucleus at much slower pace, often subject to a delayed splicing mechanism, owing to steric hindrance brought by binding of snRNP that pre-

vent it from leaving the nuclear pore [50]. These mRNA are susceptible to the cellular recycling system for RNA particles that are not translated, known as nonsense mediated decay within and outside the nucleus. Furthermore, should the mRNA be translated in the cytoplasm, it could inhibit or occupy translational machinery. This is referred to as unproductive splicing and leads to a downregulation of translation as a whole within the cell. Retained introns could also host premature termination codons (PTC) in translation. But occasionally, intron retention could give rise to new internal protein domains, bringing protein diversity to the encoded gene (see Fig: 1.3, [61])

1.5 High-throughput Sequencing Methods

Insights into the sequence features of DNA and RNA elements would not be possible without high throughput sequencing approaches and careful experimentation by modifying sequence features observed in genes. This demand to read the sequence features of DNA and RNA elements lead to highly parallel sequence identification approaches.

Modern sequencing approaches were developed from Polymerase chain reaction (PCR) developed by Kary Mullis [118]. PCR is an in-vitro approach to generic DNA replication/amplification by using polymerase enzymes, as cells do, along with deoxynucleotide triphosphates (dNTP). dNTP are the monomers that make up DNA. They are nucleosides bound to three phosphate groups. The phosphate groups act as energy sources needed to polymerise DNA and also one of the three phosphates form the phosphate backbone of the DNA. DNA polymerisation typically extends from 5' to 3' direction, with the phosphate groups being linked to 3rd and the 5th carbon atom's hydroxyl groups in the ribose-sugar base of nucleotides. Most sequencing approaches today, use a polymerase and assemble the replicating strand in-vitro with monomers of the nucleotide triphosphates.

Microarrays were first developed as the first high throughput approach in 1995 [141]. Microarrays use PCR to develop glass slides with 3' regions of a gene bound on them called probes. Hybridisation of probes with fluorescently labelled nucleic acids were used quantitatively to study expression, with the position on the array marking the identity of these genes. These probes were initially developed gene wise with each probe represented by their 3' region and then eventually developed to host multiple probes per gene, representing the exons within a gene [91, 21]. To study the identity of mRNA transcripts, probes were designed specific to exons and splice junctions [21, 127]. This led to some of the initial developments in the systematic study of AS. However, since these probes were engineered to be complementary to strands of the gene, they were limited by what transcripts were already known in terms of their genetic sequence.

Probes required for microarray analysis needed established genetic sequences from gene bodies in order for them to be studied, thereby limiting the discoverability of new genes. This challenge exists in PCR reactions which are guided by carefully designed primers, leading the better the study of sequence features for the genes that already have sequence features. The need for transcriptome wide sequence identification which was developed with a more generic adapter ligation approach. Stable sequence features of mature mRNA transcripts, specifically the poly(A) tail further enriched for protein translating mRNA transcripts more robustly [174]. Microarrays were developed as a sequencing approach that got outclassed by Illumina's RNA-Seq, but they're still used today for quick DNA-screening, owing to its ease of industrialization and quick read out time.

The first sequencing approach is a 'chain termination method' called Sanger sequencing after Frederick Sanger who with colleagues developed this approach in 1977. This approach involves interrupting a strand of DNA, replicating by polymerase based in-vitro protocols, with dideoxynucleotide triphosphates (ddNTP) [39, 155]. ddNTP lacks a 3' hydroxyl group, preventing the extension of the phosphate back-bone. To each reaction, only one type of ddNTP is added. With a concentration of ddNTP much lower than a concentration of dNTP, polymerase based reactions could be interrupted at various lengths of the replicating DNA strand (see Fig: 1.4A-B). This leads to long fragments of DNA but of different lengths, which could be separated by southern blotting and read manually(see Fig: 1.4C). The improved capillary sequencing, where electrophoresis through gel was replaced with an acrylic capillary instead, was developed for the Human Genome Project (HGP) [65]. While this sequencing approach is still considered the gold standard for sequencing DNA elements, this approach is very tedious. Analysis of fragments that are 500 to 700 bps long can take 3 hours [117], is costly and does not scale well.

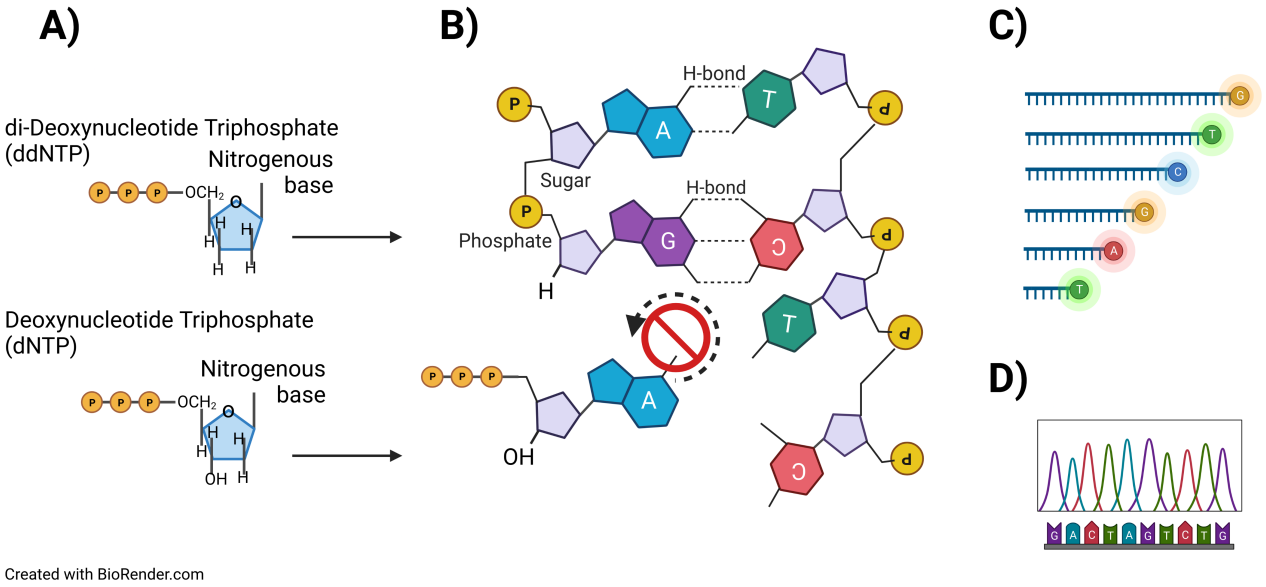


Figure 1.4 Basic Principle of RNA-Seq as Illustrated with Sanger Sequencing.
A: The structural difference of di-Deoxynucleotide Triphosphate (ddNTP) compared Deoxynucleotide triphosphates (dNTP) is a missing oxygen atom at the 3' carbon atom of the Ribose sugar **B**: ddNTP are used to interrupt a polymerising DNA-strand, which otherwise are use dNTP to polymerise. **C**:ddNTP are attached with a fluorescent signal which marks the end of DNA strand with a unique signal for each base. These strands of DNA are then sorted by size either with a gel or through a capillary. **D**: Fluorescent dyes are excited and the signals are read by a light sensor electronically.
(Image copyright: Original image created with BioRender.com)

1.5.1 RNA-Sequencing

Arguably the best way of describing the current state of a cell is by its protein composition, but methods such as mass spectrometry pose its own challenges as the readout of mass spectrometry are fragments of protein particles rather than uniquely identifiable proteins, while still being prohibitively expensive. The

current standard is describing the current state of a cell by describing the composition of mRNA particles within a cell.

The incentive for the HGP to develop massively parallel sequencing techniques led to the development of two dimensional optical imaging of the nucleotides that were integrating on an *in situ* replication of DNA over time [78, 101] (see Fig:1.5.2). Compared to the Sanger approach to sequencing used in the HGP [68], this revolution from Illumina cut down the projected budget and timeline for the HGP. The speed of the high throughput revolution far outpaced Moore's law that governed the time taken for innovations in the semiconductor race and the current short read sequencing approaches are referred to as Next Generation Sequencing (NGS) [107]. Today high-throughput approaches have incorporated much higher parallelization and application specific integrated circuits, which handle much faster compute than general purpose Central Processing Units (CPUs) or Graphic Processing Units (GPUs). Also new approaches in sequencing technologies allow for much larger DNA fragments to be sequenced. While with the Sanger approach, one was limited by the dimensions of an electrophoretic gel, Illumina's short-reads, span up to 150 bp, with each flow-cell measuring up to 48 genomes with 30x coverage in 44 hours [59]. Illumina short read sequencing is an RNA-Seq protocol which involves the following steps. First, a cellular membrane is dissolved and the RNA components of a cell are isolated. The next step is enriching for poly-A tails from the composition of RNA within the cells, to extract only mature mRNA. The oligonucleotides with a known sequence - adapters are ligated to the mRNA transcript - and mRNA is converted to complimentary-DNA (cDNA) by reverse transcription. Primers amplify all cDNA strands with PCR reactions that replicate adapter to adapter. Quite often, the adapters on RNA-fragments have sample identifiers; this is called multiplexing. These are sequences of DNA that act as barcodes to help identify which RNA-particle comes from which sample and which arise from the same molecules. The resulting cDNA now has flanking regions that are complementary to probes found on an Illumina flow cell. The flow cell is the final site of the cDNA particles. Before it can give us a signal, however, the cDNA particles are put through PCR cycles, this way the light emitted by the replicating cDNA strand can be amplified for a robust reading from each cluster of bridge amplified sequences and the sequences are read via a camera. This amplification process is referred to as bridge amplification sometimes, because of how the cDNA strand bends over to meet the primers on either end, which are both connected to the flow cell.

Fluorescently labelled nucleotide triphosphate particles release their fluorescent tags, when the nucleotide polymerises with the synthesising strand of the cDNA. This is called sequencing by pyrosequencing. Cameras identify photo signals that arise from each assimilating nucleotides with the help of filters, which filters out the signals of other nucleotides, separating each nucleotide sequence by Red Green Blue channels. This information is assimilated into FASTQ files electronically by a software called a basecaller. The basecallers are also assessing how well the fluorescent particles synchronise with the bridge amplified DNA particles. This and other key measures are collated to a single quality score that describes how frequently one can expect errors. Resulting FASTQ files contain both the sequence and the quality score. While Illumina's NGS approach works very well and is very popular for short reads, the clusters of cDNA formed with bridge amplification and the polymerase, tend to assimilate new nucleotides in sync until around 150 bp. Beyond this point, asynchronous signals of binding nucleotides drop quality scores of the reads. This brings the limitation on read length for NGS methods.

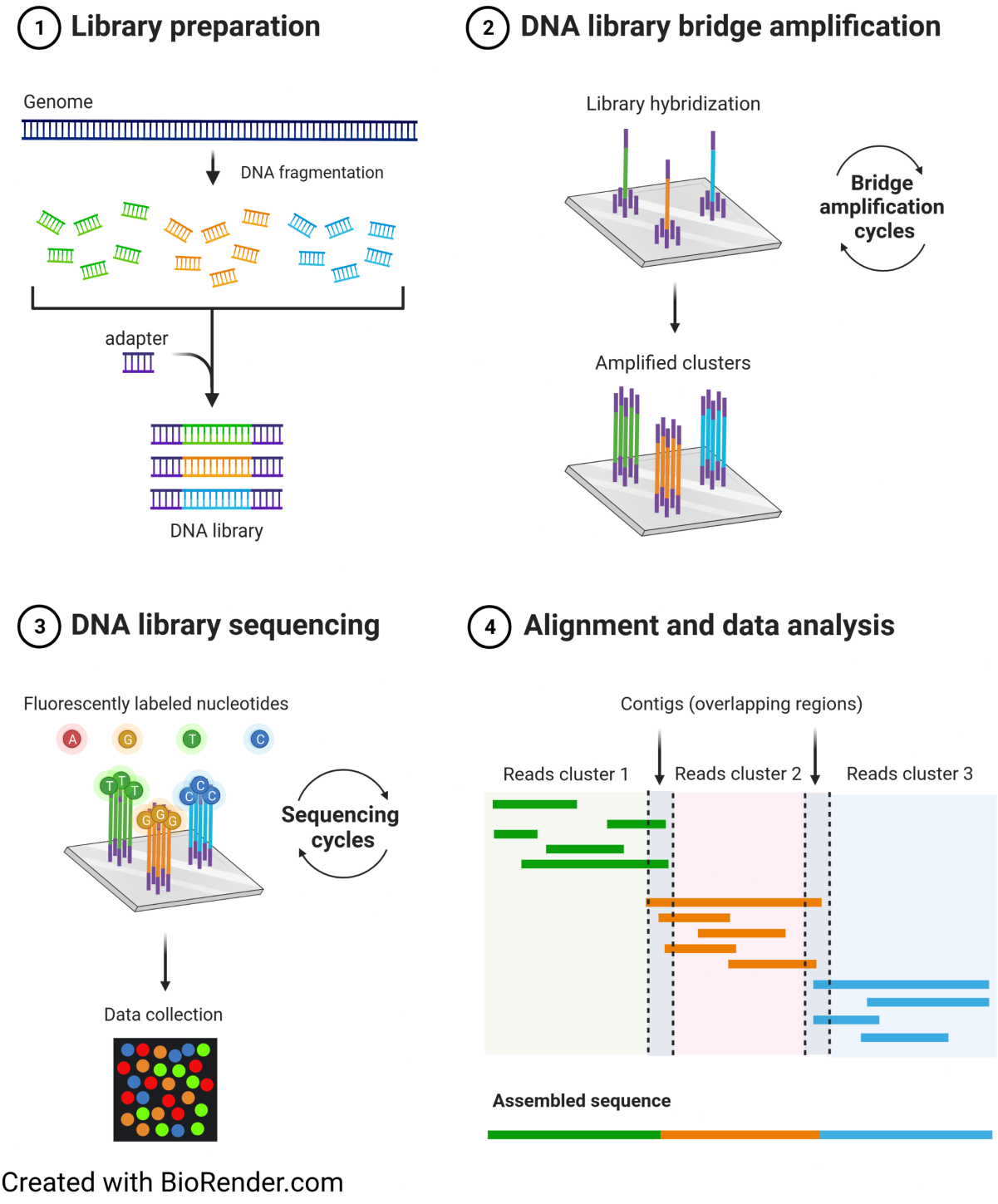


Figure 1.5 Working principle of Illumina Sequencing.

1:Library preparation is where adaptors are ligated to DNA-fragments to allow for primer binding during PCR Steps and for binding on the flow-cell **2**: Spots on a flow cell bind the DNA fragments and bridge amplification makes copies of identical DNA strands, which can give rise to a more robust signal when sequencing. **3**: By changing the wash solution, nucleotides with fluorescent tag release a photo-signal, when polymerising, signals reveal the sequences on DNA strands. **4**: DNA-strands sequenced can be analysed by computational steps for alignment or assembly.

(Image copyright: Original image created with BioRender.com)

Long read sequencing in comparison can detect AS better than NGS approaches within a transcriptome due to the fact that most genes span longer than 150 bps, the standard read size in NGS approaches. This leads to an uneven coverage of all the exons within a gene, leading to false positives in splicing detection [113]. Long read sequences from Nanopore and Pacbio show great potential to be able to quantify RNA isoforms at single cell resolution [16, 53]. Pacbio's long read sequencing approach still utilises pyrosequencing, but within small wells on the flow cell [134]. Pacbio also uses adapters to convert double stranded DNA to circular DNA, giving the capacity for a polymerase to return to the same point without letting go of the template strand. Collated reads give greater confidence for the sequences identified, while still observing just one polymerase throughout its life.

Nanopore sequencing uses a specialised membrane, bound with polymerase, that separates two volumes with an electric potential [13]. DNA strands that go through this pore generate fluctuations within the membrane for a unique 5 sequence window within the polymerase's channel. A deep learning neural net acts as a basecaller, decoding unique 5 window signals to sequences. However, since short read NGS technology has been shown effective for many diagnostic purposes and since they have built a competitive pricing edge, they still dominate the sequencing market of transcriptomic studies.

NGS technologies were also becoming useful in many other applications than genome sequencing. With around 93% of genetic associations in mutations being linked with non-protein coding sequences, NGS sequencing has been finding use in resolving structures that have epigenetic cues [106]. Applications of NGS in Chromatin Immunoprecipitation-Sequencing (ChIP-Seq) for example identified regions of the DNA where protein binding was occurring. These technologies diverged as offshoots of the advances on the HGP.

1.5.2 Chromatin Immunoprecipitation-Sequencing

Chromatin elements and protein modifications on the histone tails have shown to influence the position of nucleosomes, and modulate the speed of RNA-Polymerase (see Section: 1.3). ChIP-Seq is a targeted approach to identify the footprints of DNA-binding proteins. Chromatin immunoprecipitation sequencing (ChIP-Seq) was developed to assess the chromatin states of DNA [126]. However ChIP-Seq provides evidence for all DNA-Protein interactions, and has been crucial to identify the binding sites of transcription factors as well as splicing factors such as the small nuclear ribonucleoprotein (snRNP) and heterogenous ribonucleoproteins (hnRNP), when a diverse set of experiments are generalised to genome wide design [114]. The main technological concept underlying ChIP-Seq is the concept of crosslinking (see Fig:1.6). Crosslinking is the formation of a covalent stable bond between a methyl group found within a nucleotide's nitrogenous base to an amino acid such as Cysteine, Histidine, Seriene, Lysine. This binds proteins that are already bound to DNA permanently until a chloroform phenol extraction phase, after the DNA-protein complexes are extracted. The solution is then sonicated for an optimal period of time to elucidate 200 bp DNA fragments. These complexes are purified by co-immunoprecipitation by binding to specific antibodies, designed to pull down the protein of interest. It is crucial to have antibodies that have a strong and specific binding to the protein of interest, as this determines how many proteins are actually captured this way. Co-immunoprecipitation is followed by protein digestion and centrifugation. These fragments of DNA are put through a PCR reaction, to maintain the integrity of the sequences, even if the DNA fragments are

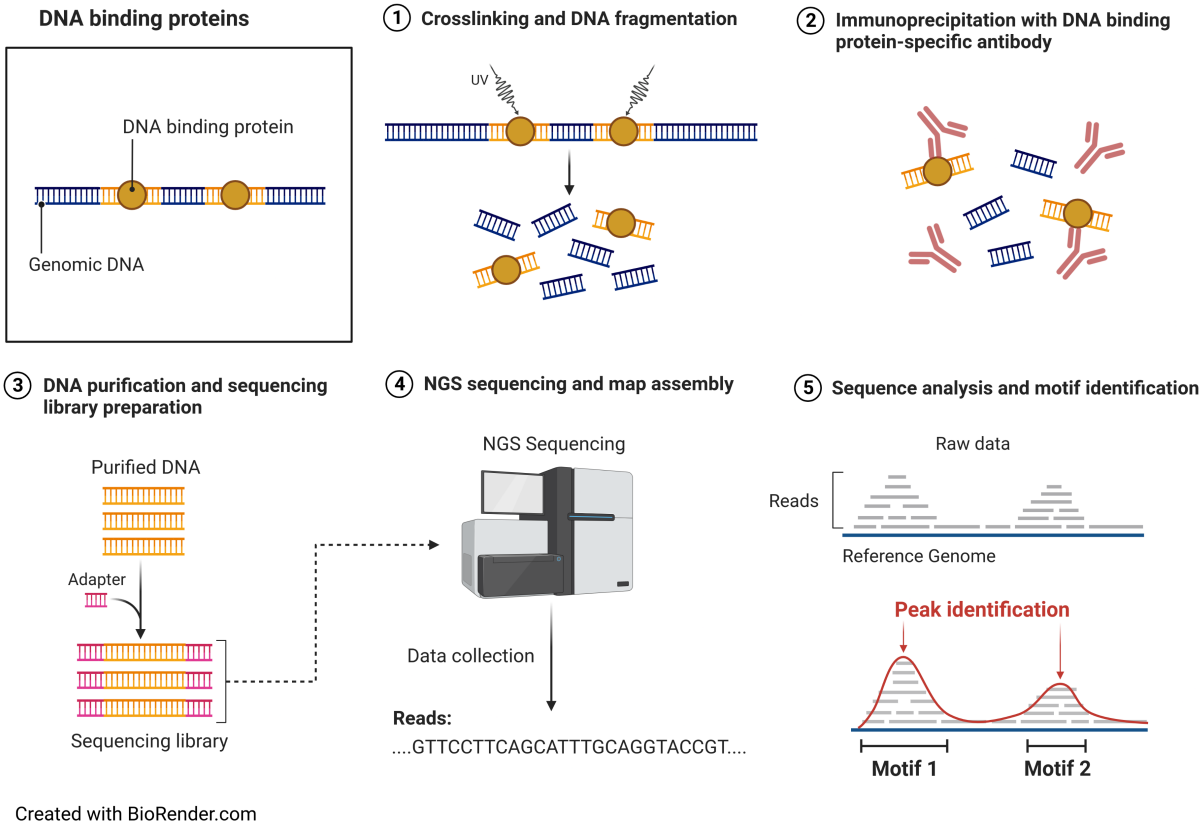


Figure 1.6 Working principle of ChIP-Seq.

1: Cross-linking of Proteins combine protein and DNA elements. **2**: Highly specific antibodies that bind to the protein of interest is is used to pull down the protein of interest and any DNA elements it bound to. **3**: The cross-linking is then reversed and Illumina adaptors are ligated on the DNA fragments. **4**: Amplified DNA fragments are then subject to sequencing by NGS technology. **5**: Reads of DNA fragments are aligned and analysed along with peak callers such as MACS. Reads from ChIP-Seq usually are bi-modal in nature. (Image copyright: Original image created with BioRender.com)

less stable. The DNA fragments are then measured with a nanodrop to identify the concentration of DNA elements in solution and sequenced similar to the way cDNA within RNA-Seq is sequenced.

The shorter read fragments within ChIP-Seq are aligned to a genome and with the help of a tool such as Model-based Analysis for ChIP-Seq (MACS) [173] to identify a sparse alignment of small windows of the genome that have a sequence shift. The cause of such a shift is that fragments of DNA that bind to a protein are always offset, strand specifically, due to the position of the forward and reverse strands of the DNA which interact with the protein. A symmetric shift in the forward and reverse strands suggest that the DNA-elements in this region were fragmented while being bound to a protein. MACS models this sequence shift and identifies binding sites of the proteins in study as narrow peaks. The output of ChIP-Seq is a file in a Browser Extensible Data (BED) format. The first two columns of a BED file describe the boundaries of the peaks on a reference genome, where the binding site has been observed, the other columns describe the strand information, a score and a label for the feature being described.

1.6 Computational Analysis of Transcriptomic Data

Data collected from NGS methods such as RNA-Seq is usually in FASTQ formats, containing units called reads with the sequences found on the mature mRNA transcripts and a quality score. This data is usually put through various pre-processing measures such as: trimming reads of poor quality scores; trimming of adapters used to fix cDNA elements to the sequencing flow cells; pooling technical replicates, etc.

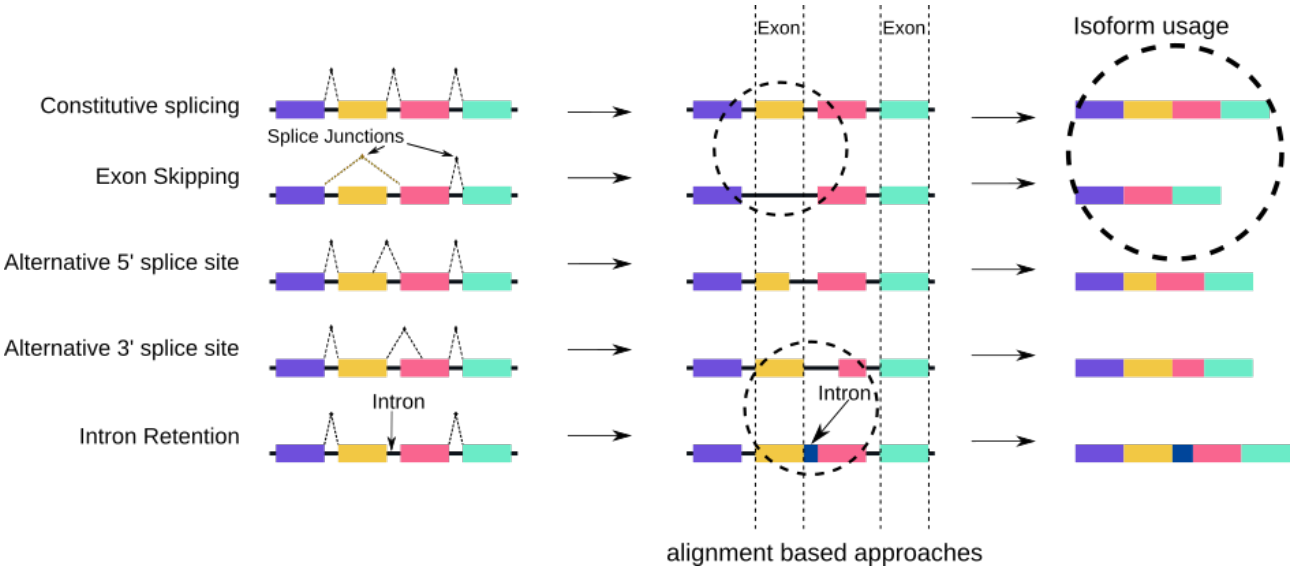


Figure 1.7 AS is quantified in RNA-Seq data by: i) Alignment based approaches, ii) Exon usage approach and iii) Isoform usage approach.
(Image copyright: Original image, license: CC-BY-NC-ND 4.0)

Depending on the research scope, a variety of tools exist for further downstream analysis. The main scope of the thesis is alternative splicing discovery. For this task some tools use FASTQ files directly to detect isoforms or events, while others require the results of the mapping of the RNA-Seq reads to genomes(see Fig:1.7). However these approaches can be broadly categorised to the following approaches.

1.6.1 Approaches to Quantify Alternative Splicing Events

There are three main ways to count alternative splicing (AS) events: Event Based, Exon Usage Based, and Isoform Quantification. Here's a summary of each approach:

Event Based Approach

This method uses tools that count RNA sequencing (RNA-Seq) reads by aligning them to a genome. Mapping tools such as STAR, HISAT2, or MapSplice2 [71, 28, 164] generate mapped read files which quantify the placement of RNA-Seq reads on a genome relative to genomic features (see Fig:1.8). By comparing reads to features such as exons, alternative splice sites can be detected and reported when a reference annotation or transcripts are known. When annotations or transcripts

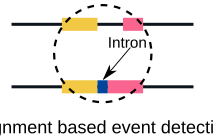


Figure 1.8 Event or alignment based approach to quantify AS events.

are not available, reads are instead compared to other alignments within the same dataset. This is also the approach used to identify and validate genome and gene annotations.

Exon Usage Based Approach

This method involves quantifying reads mapped to a predefined set of exons for each gene after using a splice-aware aligner (see Fig:1.9). The use of exon ids to describe AS limits the ability to quantify intron retention, the use of novel splice sites, or sub-exon features such as alternative 5' or 3' splice sites. However, RNA-Seq reads that align to each exon can provide a clearer picture of nascent RNA strands, as they tend to discreetly enrich exons near the transcription start site[5]. Variations in exon reads within a gene could suggest a population of genes that are in different stages of splicing. Additionally, as these approaches focus on the protein coding regions of a gene, they may also, in principle at least, correlate better to AS events observed in the proteome.

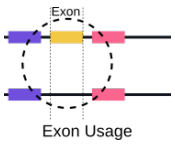


Figure 1.9 Exon usage based approach to quantify AS events.

Isoform Quantification Approach

This method tries to optimise the mapping algorithm by mapping reads to transcriptomes instead of genomes (see Fig:1.10). Tools like Salmon and Kallisto use transcriptome De Bruijn graphs to optimise the mapping step. However in this approach, read quantification becomes very dependent on your choice of transcriptome or how you decide apriori what transcripts to expect. This approach is ideal if you're trying to limit observations of splicing into discrete variables such as isoforms. Describing AS as mRNA isoforms gives a list of products, which, when considered equally, allows for the use of population statistics such as the Gini index or Shannon's entropy. Population statistics summaries can be assigned to each gene as a score for splicing, allowing a gene-wide metric for splicing.

Figure 1.10 Isoform usage based approach to quantify AS events.

1.6.2 Approaches to Interpret Alternative Splicing Events

Alternative Splicing Event Detection Approaches

When describing events that occur within a sample if the reference to the annotation is the steady and not relative to control on an experimental design, then what we learn about the way a gene behaves across all cell types and conditions can be generalised to the gene. This approach helps study AS, rather than a disease. Further downstream analyses could have their focus on AS, describing events that occur across each gene, to understand how the gene is being regulated or to investigate a novel splice variant. Tools that describe AS events in relation to a reference genome or transcriptome are referred to as AS event detection tools. Optionally a reference annotation can be used to identify and report these events better.

Differential Splicing Approaches

When further downstream analysis is coupled with an experiment design in the classic case vs control approach, tools that can quantify the differences observed across conditions are referred to as differential splicing tools. Differential splicing benefits from being able to describe the same variable, be it an exon or a transcript, across both conditions. Differential splicing tools are useful for discovering new putative roles of known transcripts across tissue types or certain conditions. However their strength to interrogate novel splicing variants are limited when considering rare events that are not represented across both conditions.

Mapping Tools

Mapping of RNA-Seq reads is often an upstream step needed for the enumeration of AS events in any dataset. Therefore 11 mapping tools were incorporated into DICAST (Table 4.1), the benchmark for AS event detection tools.

Mapping tools use many strategies for identifying the genomic position where a RNA-Seq read could potentially originate from. This process begins from segmenting the RNA-Seq read into seeds, segments, or kmers to optimise a local search. Optimising the exact location of these splits is often done considering supporting reads and their coverage for alignments at a specific genome position originating from multiple reads.

The mapping algorithms diversify their approach for these tasks by using strategies such as: Identifying breaks in kmers (CRAC); seed-chain-align (Minimap2) seed and vote (Subjunc); seed and extend (STAR); suffix trees (Segemehl); hash tables (GSNAP); Burrows-Wheeler Transform/Alignmer (BWT/A) (DART); context specific re-alignment and inclusion (ContextMap2); triplet scoring and translation (BBMap); or graph based alignment (Hisat2). All splice-aware mapping tools require a FASTQ file with RNA-Seq reads and a fasta file with a genome sequence. The main difference observed in mapping tools, is whether the tool utilises a genome annotation to support the mapping process or not. Hisat2, MapSplice2 and STAR are the only mappers that require annotation files, while the other mapping tools try to identify the genomic position of reads based on only the reference genome sequence.

Similarities among best candidates for mapping tools also lie within their index structures: STAR uses a suffix array based index; ContextMap2 and MapSplice2 uses the BWT from a non-gapped mapper, bowtie's indices; and HISAT creates prefix sorted graphs in their index. MapSplice2 and STAR are perhaps further similar in how they identify good alignments: MapSplice2 was designed to identify splice junctions and therefore use sequence tags (smaller windows of RNA-Seq reads) to anchor a window for a second stage of split alignments. While STAR employs a complex seed and extend algorithm, taking into account the splice junctions from annotations.

Alternative Splicing Event Detection Tools

To confirm the support from mapping tools, AS event detection tools often have a remapping step, where reads from RNA-Seq files are realigned to splicing features. These splicing features are a combination of observations via annotations or via the alignments. The main differences in AS event tools is the use of a splice-graph versus the use of binned alignments. Only IRFinder and ASpli use a binned coverage approach to the analysis. Splice graphs are the main feature of most of the AS-event detection tools.

Splice graphs are built based on the reference annotation (Whippet, EventPointer, Spladder) or splice graphs can be generated from the splice junction identified during alignment (SGSeq, MAJIQ). RNA-Seq reads are then mapped to the splice-graphs when possible and novel splice-events are then re-aligned and identified by various methods. ASGAL defines the splice graph by gap alignments which it then compares to a reference annotation. This additional comparison step can be optimised for a local alignment, based on the reads that only map within unique windows of genomic positions.

The 8 AS event tools listed in Table 4.1 differ in the strategy they employ to minimise the search space for local re-alignment of reads, such as the detection of splice junctions (MAJIQ), optimising for adequate coverage of retained intron (IRFinder), or using binned coverage of exons (ASpli). Another key difference lies in dependencies of input files: ASGAL runs on FASTQ files directly, while SGSeq, EventPointer, SplAdder work with aligned binary alignment/map (BAM) files. Whippet and IRFinder is able to handle both FASTQ files as well as aligned BAM files.

1.6.3 Docker

Docker (https://www.docker.com/) is an open source platform to containerize applications. This technology allows developers to create the same environment across many devices running different operating systems [30]. By containerizing applications with docker, one can test, deploy and ship applications in a reproducible manner. Docker usually hosts a linux kernel, which it borrows from the host if the host is a Linux operated machine. On Mac and Windows platforms, Docker usually hosts a Linux kernel to provide the runtime platform required for applications.

To use Docker one starts with Docker images. Docker images are versioned environments that can be easily adapted, duplicated, shipped and run on any device that hosts the Docker runtime daemon (see Fig:1.11). Docker images contain within them all the files needed to host distributions of Linux operating systems, such as: the popular Fedora, Ubuntu environments to lightweight container distributions such as busybox, coreOS or Alpine Linux. Docker images are shared across repositories by docker registries such as the publicly available at Dockerhub (https://hub.docker.com/).

To create a Docker image one needs to create a Dockerfile. A Dockerfile hosts configurations that describe the image to build upon and commands to set up docker containers - reproducible working environments. These commands help install the dependencies needed for the application, within the container and eventually start the entrypoint script, which is the primary process within a Docker container. Containers are lightweight environments that host all the dependencies for an application to run, without needing the system to host libraries or any further dependencies. Within a container, user groups, file systems and the entire user space is emulated. This allows for lightweight container systems to provide the same reproducibility as virtual machines, without dedicating hardware to the application or running heavy hypervisors, which taxes heavily on performance. Docker hosts a file-system within the container, which can be mounted on the host file system. This allows for an exchange of files across the container and the host.

A container runs only as long as the process within it is running. Once the process halts, the container is stopped as well. The process within a container begins with the entrypoint script being executed. A container is a live instance of an image, and every image hosts an entrypoint script that starts the application and operates it within the container. Processes in linux and within the container communicate error message handling via the Portable Operating System Interface (POSIX). POSIX hosts signals for

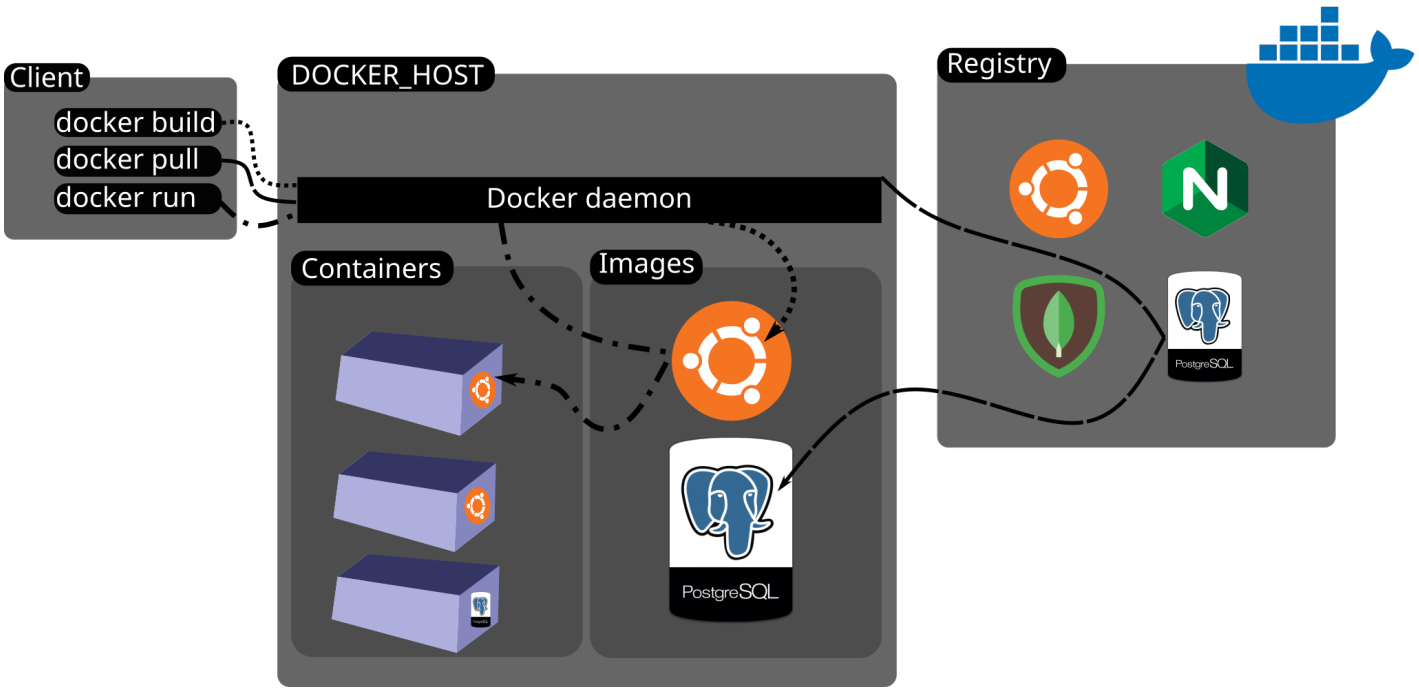


Figure 1.11 Docker architecture depicting from the left the docker api, docker host and registry. Docker images are static files with compiled binaries, they are distributed at registries, the docker host hosts a registry within the local machine. DockerHub is a central registry built for public access. Docker images for DICAST is hosted at DockerHub. A docker container is a live iteration of a docker image, docker containers may host data temporarily and run your applications in their live version.

a successful execution of commands or error messages to the system and eventually the user. POSIX compliance also allows for a sequential execution of commands in a script and can be used to interrupt code, when there's an error that needs handling. Therefore the container knows when the process within it finishes, by receiving a POSIX signal from the entrypoint script.

1.6.4 Snakemake

Snakemake is a workflow management system that is popular among bioinformaticians and widely employed [115]. The sequence of commands ('rules') in a pipeline is described in a python based language. Rules within Snakemake flows have dependencies that are determined on a directed acyclic graph. Each rule is sequentially executed only after all of its dependencies are met. Each Snakemake installation is configurable also to the computing infrastructure allowing users to take optimal advantage of their infrastructure. Snakemake installations can be configured specifically to the high performance clusters that are widely used within bioinformatic communities. This allows Snakemake to take the focus away from repetitive tasks such as configuring specific jobs to be submitted to compute cluster schedulers such as Simple Linux Utility for Resource Management (SLURM) or Sun Grid Engine. Snakemake also uses management systems such as Conda and container platforms such as Docker to build reproducible environments for applications in the workflows. This allows users to write code that frames an analysis without declaring configurations for their specific infrastructure or tying the analysis to specific datasets.

Snakemake then allows the user to focus on the manifests such as the Snakefiles, which specifies sequential steps, or rules,for their analysis along with the dependencies. An example could be the use of snakemake to transform RNA-Seq reads through pre-processing steps such as adapter or quality based trimming of RNA-Seq reads before mapping these reads to a genome. Each rule results in an output that is a dependency for another rule. FASTQ files may be a required input for the mapping rule, however the output of this rule would be a BAM file, which in turn is a declared input for another rule, like AS event detection. This approach allows for a directed acyclic graph to determine which rules must be executed before others. This also allows independent steps to be run in parallel, for better efficiency.

Snakemake also handles POSIX compliance across containers, thereby allowing sequential execution of rules within the Snakefile. This means that a rule could describe the mapping steps, allowing developers to parallelize this task across their infrastructure, and after running the AS event tools, all the while, utilising their clusters optimally with their compute infrastructure. In addition, the Snakemake workflow has also been applied to unify the format of AS tools to compare it to the ground truth in precision and recall plots; or to identify commonly found AS events reported across the different tools as well, as seen in Fig: 5.2. Snakemake also allows one to build convenient easy configurations for their workflows, allowing DICAST to run on any RNA-Seq datasets or to begin the DICAST workflow with ASimulatoR, allowing a new user to benchmark the tools listed in Table 4.1.

1.7 Current Challenges in the Study of Alternative Splicing

1.7.1 Developing Standards for Alternative Splicing Event Detection

Researchers in the field of AS develop a tool for reporting AS events every year, building a steady need for benchmarking these tools. However the benchmarks for these tools that currently exist are made by tool developers who have a vested interest in showing that their tool outperforms current state-of-the art. Furthermore benchmarks constantly argue about how to report the performance of various AS events. This is primarily because a gold standard dataset is hard to achieve.

Current RNA-Seq simulators used frequently for benchmarking generate read level data for short read sequencers. These simulators are Flux simulator [52], RSEM [83], Beers [51] and Polyester [44]. RSEM and Polyester generate read level data for gene level analysis where transcripts are used with a fixed model. Flux simulators modulate the intron level within the gene to replicate the delay of intron splicing. BEERS incorporates a curated list of gene models within its simulations to represent 11 transcript annotations, however when simulating the RNA-Seq reads these gene models are picked at random. Therefore the datasets differ often by a wide range of factors such as the read length and sequence coverage of the 2nd paired read for Flux simulator or the number of alternative forms of a gene for BEERS, but not types or distributions of types of AS events RNA-Seq simulators that generate datasets with modulated AS events had not been developed.

With new tools being developed for quantification of AS frequently, these publications are perhaps the best source of benchmarks for AS tools [4]. However in order to publish your tool in a peer reviewed journal, it should be better than state of the art tools out there, therein lies a conflict of interest. Furthermore benchmarks often report commonly found features of a dataset by these tools such as; splice junctions, polyadenylated sites, differentially spliced genes or further downstream analysis [109, 143, 28]. However,

when comparing AS events by their type, it requires gold standard datasets or an RNA-Seq simulator where specific AS events can be modulated.

Finally, how tools choose to report AS events are vastly different. While some tools use standard formats such as sam or bed, others chart their own path with outputs specific to each tool. This therefore complicated downstream analysis, often restricting users to a certain set of tools that were co-developed with specific tools designed upstream of their analysis, with respect to mapping and AS event detection.

1.7.2 Systematic Analysis of Proteins Involved in Alternative Splicing

Another challenge is integrating different layers of omic data, in the quest for identifying proteins that are involved in AS: Genomic intersections and correlations of ChIP-Seq and RNA-Seq have been used to identify genomic function of transcription factors from matched ChIP-Seq and gene expression data [167]. RNA-Splice junctions have also been shown to correlate with nucleosome positioning as observed in ChIP-Seq.

RNA-binding proteins and their occupancy are identified by high throughput Cross Linking Immunoprecipitation or CLIP-Seq, which is a specialised approach to ChIP-Seq. It was developed to learn more about RNA-Processing such as splicing. RNA-Seq and CLIP-Seq data have also been used to investigate the role of RNA-binding proteins involved in RNA processing [41, 156]. Be it DNA binding or RNA binding elements, while many tools exist for motif extraction [166, 56], RNA-Seq analysis has yet to see a systematic approach to splicing analysis.

When comparing transcriptomic analysis of AS to another omic layer, such as ChIP-Seq, regions of influence for the protein in interest varies from promoter regions for transcription factors, transcription start sites for RNA-binding proteins [145] and even to intron-exon junctions for nucleosomes [122]. For a systematic evaluation of AS across a transcriptome and ChIP-Seq analysis, currently no tool exists that can be used for so many types of nucleic binding proteins.

2 Outline

This thesis presents tools developed toward the aims described in the next section.

- To generate in silico gold standard RNA-Seq datasets I present ASimulatoR [100];
- To evaluate splice-aware alignment tools as a basis for AS analysis and benchmark AS-event detection tools, I present a modular framework for benchmarking with DICAST [40];
- To identify proteins that are implicated to be involved in AS from integrated RNA-Seq and ChIP-Seq datasets, I present DASiRe (manuscript in preparation).

In the chapter 'Methods' (Page:27), I describe packages and datasets used to develop the tools presented in this thesis. A chapter on Implementation describes how these tools were put together and key aspects of the algorithms within the tools presented in this thesis.

The chapter 'Results' (Page:41) pertains to the outcome of the benchmarks of mapping and AS-event detection tools. It also includes findings from having built the web-application DASiRe and describes the example dataset stored within it.

Finally the chapter 'Discussion' (Page:59) highlights key takeaways from this thesis which includes: the potential biases in simulated datasets from ASimulatoR, the benchmark of AS-event detection tools, recommendations on tools used for AS analysis, developing common elements toward a unified format and finally, the potential for integrating RNA-Seq and ChIP-Seq data toward implicating proteins that influence AS.

ASimulatoR is mostly described in chapters Methods and Implementation, as it forms the basis for the benchmarks within this thesis.

The mapping and AS event detection tools within the benchmarks are introduced collectively in the introduction already, however a lot more details for them are presented in the results section or when discussing the results of the benchmark.

Integration of RNA-Seq and ChIP-Seq datasets are algorithmically detailed in the section on implementation, with screenshots for the visualisations from the web-server presented in the section on results.

3 Aim of the Thesis

Within this thesis, I address the current challenges listed in studying AS:

1. **Aim:** Creating an in silico gold standard for AS through simulation.

   When gold standard data is hard to achieve, working with simulations help us assemble our current collective paradigms in the form of simulation models. Currently the state-of-the art RNA-Seq simulators leave out AS, which leaves unaddressed a new paradigm of research in transcriptomics. One where certain genes could potentially share different functions based on splicing patterns observed within the gene. In order to benchmark tools that work with AS, I would address this limitation by discussing work done on building a RNA-Seq simulator, ASimulatoR, that also addresses AS by modifying distributions of specific splicing events.

2. **Aim:** Evaluate splice-aware alignment tools as the basis for AS analysis.

   As is often the case, AS event detection tools require splice-aware alignment as a prerequisite. This means that a good evaluation of AS event detection tools also needs to be benchmark splice-aware alignment tools. Benchmarks of splice-aware alignment tools consider what fraction of reads were mapped and the precision of these tools. They also evaluate which splice-aware mapping tools can work well with each of the AS event detection tools.

3. **Aim:** Evaluate AS event detection tools.

   Current benchmarks are limited to benchmarks with conflicts of interest. Through this thesis, I would address these limitations by discussing work done on building a reproducible framework for independent benchmarking of AS event detection tools, DICAST. AS event detection is benchmarked on Precision and Recall for each of these tools.

   a) **Contrast reference-based and reference-free tools.** AS event detection also often relies on reference annotations. This thesis will evaluate the use of reference annotations as a feature and how tools compare to reference free AS event detection tools.

   b) **Offer users guidelines about how to use AS event detection tools.** When considering any observable trade-off between precision and recall. This thesis will offer guidelines on when and how to use AS event detection tools.

   c) **Offer a unifying standard for reporting of AS events.** Unlike alignment tools which output their results in an established BAM format, AS event detection reports have yet to consolidate on a format for reporting AS events. This thesis develops a unifying standard to report AS events highlighting the most common and useful features reported by the AS event detection tools.

4. **Aim:** Develop a web application to assist the discovery of DNA-binding elements that are involved in splicing.

   Finally, we take alignment based, exon usage based and isoform usage based tools to identify AS events in an ensemble implementation to a specific goal, to keep the analytical strength for the events each kind of tools identify. Here a co-ordinated approach to reporting AS events provides evidence at various scales within the scope of identifying AS genes which also show evidence of splicing factor binding in each of their ChIP-Seq data. We therefore study the role of splicing factors in AS and also visually present the information for each experiment on a web-application, DASiRe.

4 Methods

4.1 ASimulatoR

ASimulatoR is built in R (v3.6) and is a publically available package at the github page: https://github.com/biomedbigdata/ASimulatoR. ASimulatoR is also packaged as a docker container and is hosted at docker hub (https://hub.docker.com/r/biomedbigdata/asimulator). ASimulatoR was built on a customised version of Polyester that incorporates PCR bias and adapter contaminations.

4.2 SHIP Cohort: RNA-Sequencing Dataset

To estimate the number of alternative splicing events in a biological dataset, we analysed RNA-Seq data from real biological samples: the SHIP (The Study of Health in Pomerania) Cohort [162]. The SHIP cohort is an RNA-Seq dataset from whole blood tissue of 117 healthy individuals, sequenced on Illumina HiSeq 4000 platforms. 500 ng of RNA was extracted from whole blood with a mean RNA integrity value at 8.5, and a library was prepared using the TruSeq standard mRNA kits with 24 barcodes and sequenced at a depth of 40 M clusters per sample. The patient biopsies were sequenced at the Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald and FASTQ files were sent via physical storage to Technical University of Munich. The data from this cohort is available on a web-based request at http:// ship.community-medicine.de.

The patient data were all mapped with STAR, as STAR proved to be the best mapping tool (see Section: 6.1.2) in the first iterations of the benchmark, performed with ASimulatoR on the human genome versioned hg38 and Ensembl genome annotation versioned 99. MAJIQ was used to analyse the SHIP cohort as it too seemed to perform well in the first iterations of the benchmark with ASimulatoR. Versions of STAR and MAJIQ used are the same as in DICAST and listed in the Table: 4.1.

4.3 DICAST

DICAST (Docker Integrated Comparison of Alternative Splicing Tools) is a benchmarking suite that aggregates tools hosted within Docker containers (introduced in Section 1.6.3). These containers are orchestrated by Snakemake (introduced in Section 1.6.4), the workflow manager. Together with ASimulatoR, DICAST can benchmark mapping and AS event tools with simulated datasets. DICAST can be installed with only Docker and Conda as dependencies. All the code is made available and free to use via github at the repository: https://github.com/CGAT-Group/DICAST. However, if you do not want to run the full benchmark, but just want the tools outputting the unified format outputs (introduced in Section 5.3), then the docker images are easy to pull from the docker-hub repository at https://hub.docker.com/r/dicastproj/dicast/tags.

DICAST consists of ASimulatoR, 11 splice-aware mapping tools and 8 alternative splicing detection tools (Table 4.1).

4.3.1 Docker Images

In the context of DICAST, the entrypoint script hosts the algorithm that finds the input files, runs the alignment or AS event detection tool within the containerized environment and produces outputs in the output directories. Therefore each container hosts all the code required to emulate and operate each of the RNA-Seq mapping or AS event detection tools. This makes DICAST a very modular pipeline, with dependencies such as libraries and their versions being tied to the environment that hosts the tool and is independent of other libraries from other tools or the host. DICAST images are available across all devices via Docker Hub (https://hub.docker.com/r/dicastproj/dicast/tags).

4.3.2 Benchmarked Tools

The Table 4.1 contains tools that have met the criteria described in Section: 6.1.1 and are benchmarked within DICAST. All listed tools have their own docker container at dockerhub (https://hub.docker.com/r/dicastproj/dicast/tags)

4.4 Compute Platform

All tools were run on Intel Xeon gold 6148 Processors with 27.5M Cache, running at 2.4 Giga-hetrz. Indexing is usually the most time-consuming step in mapping algorithms, but it usually also a step that needs to be run only once, when mapping different experiments to the same genome and annotation.

Containerisation of each of the tools allowed the collection of resource metrics for each container with a tool. Prometheus and cAdvisor were used to collect container metrics, and as each of the tools are given similar environments, when possible, this reduces differences in latency when context switching, between userspace and their kernel runtimes. All mapping tools were run at the same depth of 10 M reads each, with the same reference genomes and annotations as DICAST usually does.

4.5 DASiRe

DASiRe is split into two parts, a preprocessing docker pipeline and a webserver:

The preprocessing docker includes the following tools DEXSeq (v1.40.0) [6], MAJIQ (v2.1-c3da3ce) [159], IsofomSwitchAnalyzeR (v1.16.0) [161], DESeq2 (v1.34.0) [94] , Subread (v2.0.1) [85], STAR (v2.7.10a) [28], Kallisto (v0.48.0) [15]. The entrypoint script runs DESeq2, DEXSeq, and Subread in R (v4.1).

The web server uses Rocker [125] to host the webpage on a R-Shiny Server. The DASiRe web server is served with Genomic Ranges (v1.46.1) [80] to intersect regions with AS or binding sites from ChIP-Seq.

Datasets incorporated within DASiRe are listed in table:4.2

A. Splice-aware mapping	tools		
Name (version)	**Dependencies**	**Reference**	**Genome Annotation***
BBMap (38.94)	Java 7+	[102]	-
ContextMap2 (2.7.9) In the current study used with bowtie 2 [79]	Java, BWA, bowtie 1, bowtie2	[11]	+
CRAC (2.5.2)	Perl, htslib	[129]	-
DART (1.4.6)	GCC,GNU make, libboost-all-dev, libbz2-dev, and liblzma-dev	[88]	-
GSNAP (2020-03-12)	GCC,GNU make, Perl	[170]	+
HISAT2 (2.2.1)	GCC, GNU make, MSYS, zlib	[72]	+
MapSplice2 (2.2.1)	GCC 4.3.3+, GNU make, python 2+	[165]	+
Minimap2 (2.17)	None (precompiled binaries) or GCC, GNU make, zlib	[84]	-
segemehl (0.3.4)	GCC, GNU make, htslib	[57]	-
STAR (2.7.5)	GCC, GNU make	[29]	+
Subjunc (2.0.0)	None (precompiled binaries) or GCC, GNU make	[85]	+
B. Alternative splicing event	**detection tools**		
Name (version)	**Dependencies**	**Reference**	**Supported events**
ASGAL (1.1.6)	python3.6+, biopython, pysam, gffutils, pandas, cmake, samtools, zlib	[27]	ES, IR, A5, A3
ASpli (1.12.0)	R, BiocManager	[99]	ES, IR, A5 A3
EventPointer (2.4.0)	R, BiocManager	[136]	ES, IR, A5, A3, MEE, MES
IRFinder (1.3.1)	GLIBC 2.14+, GCC 4.9.0+, Perl 5+, STAR 2.4.0+, samtools 1.4+, bedtools 2.4+	[112]	IR
MAJIQ (2.3)	htsilb, python3, python packages	[158]	ES, IR, A5, A3
SGSeq (1.24.0)	R, BiocManager	[49]	ES, IR, A5, A3SS, AFE, ALE, AF, AL, MES (with two skipped exons)
splAdder (2.4.3)	python3, python packages	[67]	ES, IR, A5, A3, MEE, MES
Whippet (0.11.1)	julia	[150]	New junctions could be added from alignment as an option ES, A3SS, A5SS, IR, AFE, ALE, tandem transcription start, tandem alternative polyadenylation, circular back splicing

* '-': does not use; '+' - can use as an option

ES - exon skipping; IR - intron retention; A3 - alternative 3'-splice site, A5 - alternative 5'-splice site, MES - multiple exon skipping, MEE - mutually exclusive exons, AFE - alternative first exon, ALE - alternative last exon

Table 4.1 Tools and version numbers incorporated in DICAST. References for the articles that describe the Tool version in DICAST (v0.3 release).

Experiment	Target protein	File format	Accession	Reference
RNA-Seq	non-targeting CRISPR	fastq	ENCFF239POR ENCFF962DHI	Brenton Graveley, UConn
RNA-Seq	non-targeting CRISPR	fastq	ENCFF250AJS ENCFF319WJN	Brenton Graveley, UConn
RNA-Seq	YBX1	fastq	ENCFF803NZJ ENCFF573WJT	Brenton Graveley, UConn
RNA-Seq	YBX1	fastq	ENCFF301JRH ENCFF193LXU	Brenton Graveley, UConn
ChIP-Seq	TARDBP	bed narrowPeak	ENCFF641AXD	Richard Myers, HAIB
ChIP-Seq	HNRNPH1	bed narrowPeak	ENCFF844QFF	ENCODE Processing Pipeline
ChIP-Seq	HNRNPLL	bed narrowPeak	ENCFF662WPN	ENCODE Processing Pipeline
ChIP-Seq	TARDBP	bed narrowPeak	ENCFF448YOS	ENCODE Processing Pipeline
ChIP-Seq	KHSRP	bed narrowPeak	ENCFF317QKH	ENCODE Processing Pipeline
ChIP-Seq	RBM25	bed narrowPeak	ENCFF102XVH	ENCODE Processing Pipeline
ChIP-Seq	TARDBP	bed narrowPeak	ENCFF909RMQ	ENCODE Processing Pipeline
ChIP-Seq	HNRNPL	bed narrowPeak	ENCFF984ESZ	ENCODE Processing Pipeline
ChIP-Seq	FUS	bed narrowPeak	ENCFF688ARM	ENCODE Processing Pipeline
ChIP-Seq	RBFOX2	bed narrowPeak	ENCFF232ASB	ENCODE Processing Pipeline
ChIP-Seq	PTBP1	bed narrowPeak	ENCFF917HXV	ENCODE Processing Pipeline
ChIP-Seq	HNRNPUL1	bed narrowPeak	ENCFF991ZSC	ENCODE Processing Pipeline
ChIP-Seq	YBX1	bed narrowPeak	ENCFF520DIY	ENCODE Processing Pipeline
ChIP-Seq	PCBP2	bed narrowPeak	ENCFF941XZW	ENCODE Processing Pipeline
ChIP-Seq	HNRNPK	bed narrowPeak	ENCFF984QUV	ENCODE Processing Pipeline
ChIP-Seq	PCBP1	bed narrowPeak	ENCFF467RYH	ENCODE Processing Pipeline

Table 4.2 Datasets incorporated within DASiRe

5 Implementation

5.1 ASimulatoR: RNA-Sequencing Simulator

The gold standard quantitative experimental data for alternative splicing discovery is hard to acquire. Thus, simulations of RNA-Seq data with known number and types of alternative splicing events could be the best way to construct an in silico ground truth. Considering the limitations of the existing tools for producing RNA-Seq data with modulated distributions of AS, I helped design the new simulator, one that can be used to benchmark AS event detection tools, called ASimulatoR. I was also involved in extensive testing of the datasets created by ASimulatoR.

ASimulatoR incorporates a modified version of an RNA-Seq simulator called Polyester which takes a genome sequence file in FASTA format and a genome annotation file in a Gene transfer format (GTF) to create RNA-Seq reads in FASTQ format [44]. ASimulatoR constructs novel RNA-Seq read sequences by modifying the genome annotations to include transcripts with predefined number of alternative splicing events.

The steps taken by ASimulatoR is as follows (see Fig: 5.1):
A1. The first step is to identify gene features in a genome annotation file and to create sets of gene features per gene that are called the exon supersets. This step creates an unspliced template for each gene.
A2. The exon supersets are then scrutinised for genes and the compatible AS events that can be imposed on them. Genes that have only two exons, for example, do not host exon skipping events, but may host intron retention.
B. In the second step an artificial set of annotations are created, with AS events imposed on the transcripts. In this step, ASimulatoR's configuration files are considered and the distribution of AS events is reflected on the annotation set as transcript variants. In this step, ASimulatoR also outputs the GTF file that would be useful for AS event tools downstream. To study the capacity of AS event tools to discover novel splice events, ASimulatoR can be configured at this stage to output a GTF file with a user-selected fraction of transcript variants to be skipped in the output GTF file.
C. The final step involves simulating reads with the modified Polyester package. ASimulatoR at this stage supports parameters of Polyester, such as sequencing error rate or sequencing depth. However Polyester was developed to also host additional features such as RNA-Seq technical biases such as adapter contamination and PCR duplicates.

ASimulatoR can also be run to simulate an experiment generating the two sets of FASTQ files and count tables, representing biological groups of samples that have differentially used AS events. This approach can be used to benchmark differential AS event detection tools. The count tables are the source of ground truth that could be used to benchmark AS event tools.

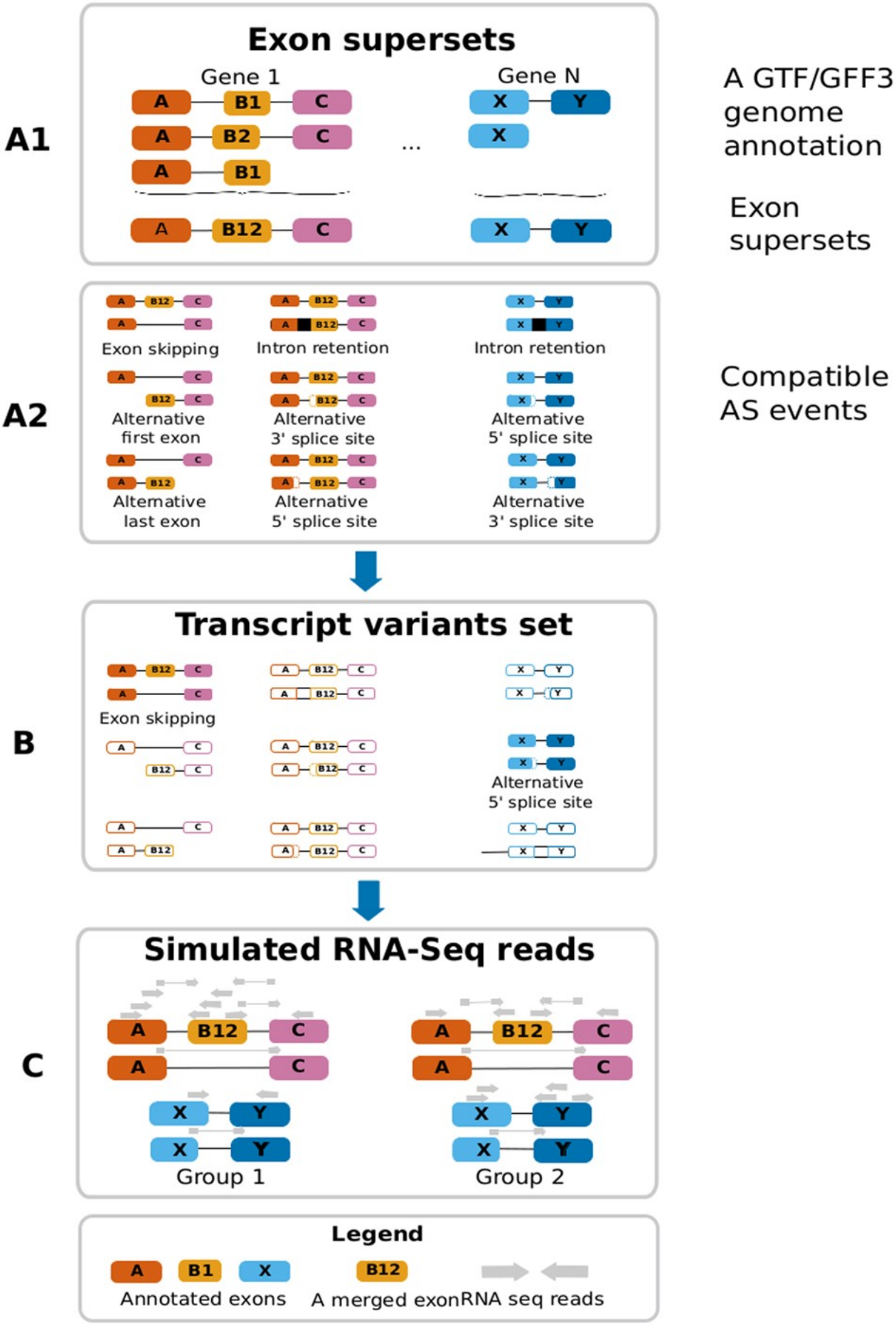


Figure 5.1 Algorithmic overview of ASimulatoR. [100].

(Image copyright: License no.:5393760310463, see Appendix at section:8 for license)

ASimulatoR is built in R (v3.6) and is a publically available package at the github page: https://github.com/biomedbigdata/ASimulatoR. ASimulatoR is also packaged as a docker container and is hosted at docker hub (https://hub.docker.com/r/biomedbigdata/asimulator).

5.2 DICAST Pipeline

Benchmark made by DICAST begins with simulated RNA-Seq files with configured distributions of AS from ASimulatoR. The distributions of AS events as seen in the RNA-Seq simulation is rendered as an output of ASimulatoR and can be compared later, using the unified output from DICAST. After the simulation, the FASTQ files can serve as an input to mapping tools or to AS event detection tools that do not require mapping step. The resulting alignment files in Sequence Alignment Map(SAM) or it's binary-compressed format(BAM) format serve as an input to AS event detection tools that require prior mapping. Finally the AS event detection tool outputs are converted to the unified format using custom scripts that are present within each docker container. The files in a unified format are used for a downstream analysis such as precision/recall and comparison of detected events via UpSet plots. A user can always choose a list of tools to run. The Figure 5.2 illustrates the workflow's overview.

Each of the tools are built within a docker container, in order to give it a reproducible environment. Snakemake orchestrates the containers with a rule based workflow that sequentially runs all mapping tools on the FASTQ file and then, the AS event detection tools. However, DICAST can also be run on real datasets. In this scenario, the benchmark does not cover precision and recall plots (see Fig: 6.6) but it reports which events were found in common across all the AS event tools, via the upset plots (see Fig: 6.8).

5.2.1 Author Contributions

My contributions to DICAST as a shared first authorship in short include design; coordination of a team of students; maintaining version control repositories; writing the first modules of DICAST with dockerfiles and entrypoint scripts; re-writing the Snakemake; managing docker images and repositories; and debugging.

5.2.2 Performance Metrics Implemented for Benchmarking

The advantage of starting from a simulated dataset, such as generated by ASimulatoR, the performance of each of the mapping and AS event detection tools can be measured based on Precision and Recall.

With respect to AS events in the simulated gold standard dataset, precision is a fraction of accurate AS events identified upon all the reported AS events. Recall is the number of accurately identified AS events upon AS events simulated in the dataset.

However, when it comes to measuring the performance of mapping tools, instead of recall is plotted the fraction of unmapped reads, compared to the total read depth as simulated by ASimulatoR. Precision is mapped as the number of accurately mapped reads and junctions, compared to the total AS events generated by ASimulatoR.

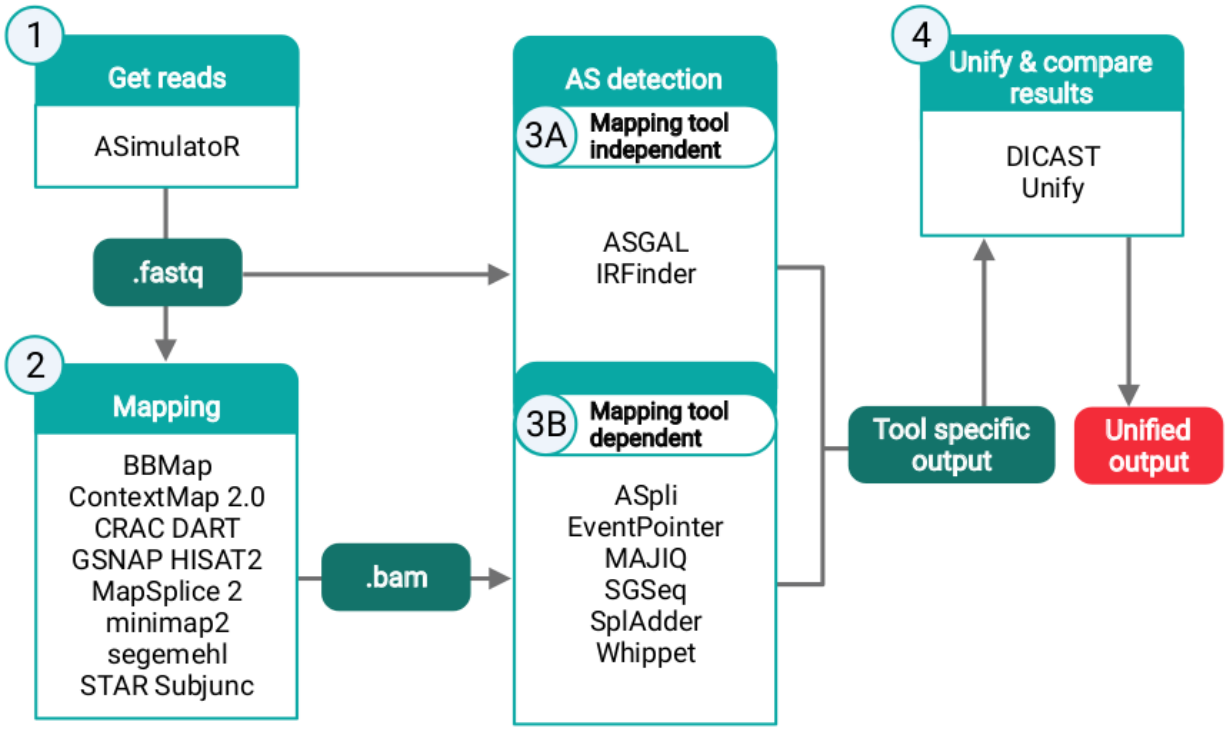


Figure 5.2 DICAST workflow.

1.) The benchmark begins with simulated data from ASimulatoR, simulating alternative splicing(AS) events with configurable distributions (optional). 2.) Fastq files are mapped with 11 mapping tools to output BAM files, 3.a.) AS event detection tools that do not require BAM files run on fastq files along with 3.b.) AS event detection tools that require bam files take them from the mapping tools that were run. 4.) AS event detection tool outputs have tool specific outputs that are converted to DICAST's unified output format. (Image copyright: Creative Commons CC-BY-NC-ND license.
Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

5.3 Unified Common Format

Unified common formats represent the assimilation of information standards that are relevant for a specific scientific paradigm. Just like fixing the BAM/SAM format for mapping tools, allows for more downstream analysis for mapping in AS events, so would a unified common format assist with the further downstream analysis of AS event tools.

The proposed format unifies common elements of most of the tools, but also shows the vast difference between formats such as the Local Splice Variation (LSV) format proposed by MAJIQ [159], which refers to intersections of splice-graphs and evidence of mapping of RNA-Seq reads to the reference splice-graph. By defining a common standard across the tools, we can now compare their accuracy, by comparing the genomic positions described as startcoordinates and endcoordinates (see Table: 5.1). The common unified format for DICAST is incorporated within every docker container, making the unified output an additional obligatory output to running each tool's docker container. This adds an estimated 9 seconds to

outputs for RNA-Seq reads with 200 M reads, and scales sublinearly. It also presents the opportunity for developing further downstream analysis to each tool.

Column	Input
chr	Symbol of chromosome for this event
gene	Gene name for this event
id	Unique identifier for this event
strand	+ or -
event_type	One of the following types: ES, IR, A3, A5, ALE, AFE, MEE, MES
count	Default = 1; can be used for tools like MAJIQ, which report multiple events for one ID to keep track of the number of events; count in combination with id has to be unique
star_coordinates	One or more start coordinates (for each exon)
end_coordinates	One or more end coordinates (for each exon)

Table 5.1 Allowed inputs for each column in the unified output file

5.4 DASiRe

DASiRe pipeline identifies regions in the genome where genes are found to have RNA-Seq reads expressing a diverse set of alternatively spliced transcripts, and regions that also have binding sites for proteins of interest. DASiRe is a web server application and consists of two parts, the preprocessing pipeline and the visualisation platform on the website (https://exbio.wzw.tum.de/dasire/). By distributing the heavy compute required to analyse new RNA-Seq dataset to a local machine, the web server can be employed more efficiently and quickly to run enrichment tests and to visualise the output.

5.4.1 Author Contributions

My contributions to DASiRe as a first author include, the writing of the preprocessing pipeline along with the corresponding docker container and entrypoint script; managing collaborative version control repositories; building the docker container for the web server and deploying it along with assisting nginx configuration for access to the public web and managing the docker repositories for the project.

5.4.2 The Preprocessing Pipeline

The preprocessing pipeline begins by importing metadata, on samples and conditions. Metadata is a simple tab separated text file with a column for a prefix for the sample filename and a column marking conditions as ‘control’ or ‘case’ . The condition ‘case’ could also be ‘treated’, or any other word that is not control, as long as a second condition named ‘control’ exists. A preprocessing step then trims adapter sequences within the FASTQ files, if given (see Fig:5.4.2). Then, indices are built for Kallisto (v0.48.0) [15] and STAR (v2.7.10a) from before it aligns RNA-Seq reads to the user provided genome and transcriptome. These indices are reused and are especially useful to reduce preprocessing runtime, if the user has a

specific use case or works with the same organism and with the same version of the reference genome or annotation. RNA-Seq files are then mapped with STAR and count matrices from STAR are extracted with Subread (v2.0.1) to count tables as 'csv' files. Parallelly, RNA-Seq is also aligned with Kallisto for transcript counts.

The preprocessing pipeline analyses RNA-Seq data on 4 different levels: gene level, AS event level, exon usage level and isoform usage levels (see Section 1.6.1). DEXSeq is a differential splicing tool that focuses on exon based approaches to quantify AS [133]. DEXSeq (v1.40.0) reports differentially used exons, across conditions listed in the metadata, from alignments made by STAR. MAJIQ (v2.1-c3da3ce) differentially quantifies reads that occur across junctions also from alignments made by STAR. Finally, IsofomSwitchAnalyzeR (v1.16.0) uses Kallisto's pseudocounts to detect differential isoform usage. These tools together represent a comprehensive view of AS within an RNA-Seq dataset. Additionally, differential gene expression analysis is performed by DESeq2 (v1.34.0).

Input files such as FASTQ files from RNA-Seq datasets and reference genomes, transcriptomes and their annotations can be need placed in the input directory that mounts to the docker container. As an optional input, adapter sequences can also be placed in the input directory and mentioned in the configuration file. The Git also shares an example of the directory structure for clarity. ChIP-Seq data needs to be processed separately and suggestions for it are described in the Section 5.4.3. The output data includes reports of alignments from STAR and Kallisto along with differential gene expression. Furthermore, AS event detection by DEXSeq, MAJIQ and IsoformSwitchAnalyzer are available for the user as soon as the pre-processing pipeline is finished. However, a select collection of outputs are placed in a directory for upload to the DASiRe web server for visualisation.

By using Docker, the preprocessing pipeline enables creating reproducible environments with version controlled libraries (see Section on 1.6.3). This image can be pulled from Docker Hub (https://hub.docker.com/r/marisalb/dasire) to any device. GNU parallel [153] is adopted to parallelize each task by sample, when possible.The entrypoint scripts within the container run with bash strict mode [1] in order to comply entirely with POSIX signals.

5.4.3 The Web Server

The DASiRe web server builds on rocker/shiny-verse (v3.6.1). The web server requests users to upload the output directory from the directory that the preprocessing pipeline outputs. Also, ChIP-Seq results are provided as a reference, from K562 cell lines, which were useful for the example dataset. These references may not show the same binding sites for proteins in a different cell line or tissue.

The server hosts a publically available example RNA-Seq dataset which was obtained from a K562 cell line, with gene Y-box binding protein (YBX-1) knocked out with Clustered Regularly Interspaced Short Palindromic Repeats and associated proteins (CRISPR/CAS), along with a non-targeting CRISPR/CAS-system as a negative control with two biological samples each. The dataset is available from ENCODE listed in table:4.2.

Key visualisations of DASiRe's RNA-Seq data analysis are presented in Section: 6.2 and described below: RNA-Seq quality control page reports mapping quality scores from MultiQC [38] of STAR and Kallisto runs from the preprocessing pipeline. Differential gene expression results are shown on a principal component analysis (PCA) plot, and a heatmap shows the in-between sample distance for the user to

check if the samples are grouped based on conditions. Normalised gene counts are visualised per gene as dot plots for each condition. DESeq2 visualisations are offered via a volcano plot, where the user could set the log fold change cut off. Splicing factor differential expression shows the results of DESeq2 on a limited list of splicing factors acquired from the SpliceAid-F database [48]. Results from DEXSeq are represented in the tab 'Differential Exon', where the user sets thresholds for p-value and for log fold change. Once set, the results are filtered with the threshold and visualised with expression value of the exon on the y-axis and on the x axis, exons within a gene. Isoform switch results are summarised in tabs on 'genome-wide isoform splicing page'; 'Gene switch plots'. 'Genome-wide isoform splicing' summarised the types of AS detected across the genome in numbers with bar plots and isoform usage in violin plots, with significance calculations. The gene switch plots are typical outputs of IsoformSwitchAnalyzeR, with a page for each gene with representations of isoforms discovered in the dataset. Gene expression, isoform expression and isoform usage plots are shown with significance calculations. MAJIQ's results are summarised in the 'splice events' tab first as a pie chart and also with a bar plot exhibiting the number of genes involved in intersections within types of AS identified in complex events. Finally the comparative analysis presents a summary of all evidence of splicing identified in a Venn diagram.

DASiRe relies on user provided ChIP-Seq peaks in the form of BED files. As a recommendation for processing the ChIP-Seq data, we refer to ENCODE 's uniform pipeline for ChIP-Seq analysis: https://github.com/ENCODE-DCC/chip-seq-pipeline. This pipeline uses MACS [173], for peak calling of chromatin elements and the SPP pipeline [69] for peak calling of transcription factors as recommended by ENCODE (see Fig: 5.4). Further resources such as Galaxy CLIP-Explorer [56] are also recommended for RNA binding proteins.

DASiRe can also load ENCODE ChIP-Seq data from the following cell lines K562, GM12878, MCF-7, HepG2, HEK293T are incorporated within DASiRe for a quickstart analysis. The following pages are then available for the user: Quality control: On this page, one can select the ChIP-Seq experiments they would like to load, and a reference genome for the track. BED files are read into DASiRe and ChIP-Seq peaks for each protein of interest. Regions of their target genes are visualised where binding is observed. ChIP-Seq peaks are also visualised per gene on a gene track, similar to genome browser. Peak Enrichment: ChIP-Seq files and reference genome is as set on the quality control page. A Fisher's exact, two sided test is deployed to check non-random associations between regions with ChIP-Seq peaks vs no ChIP-Seq peaks and regions with AS events vs no observed effect of AS. Associations are described on a heatmap on gene level and on promoter regions of genes.

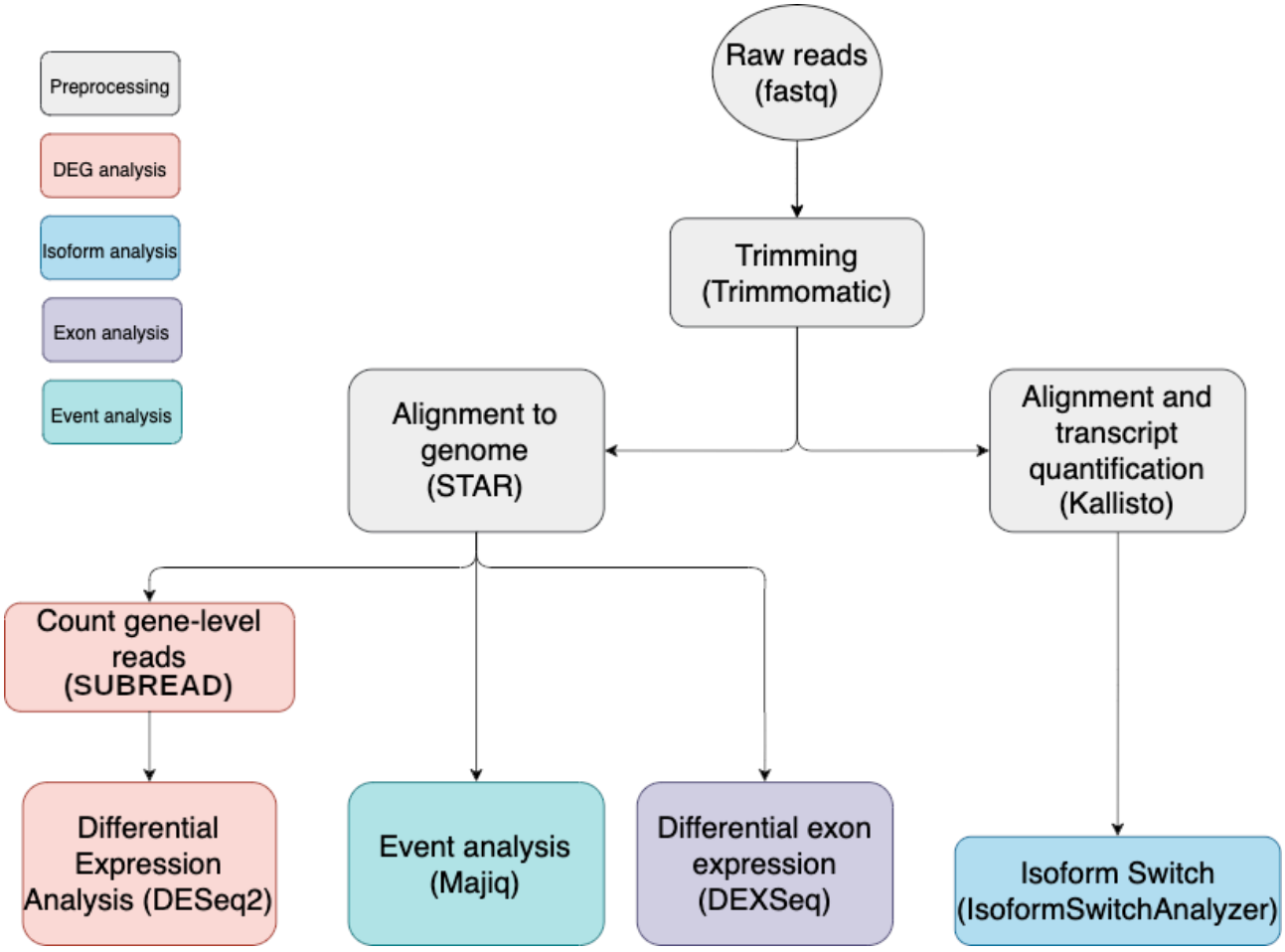


Figure 5.3 DASiRe's preprocessing pipeline differentially quantifies RNA reads on gene level with DESeq2, event level with MAJIQ, exon level with DEXSeq and isoform level with IsoformSwitchAnalyzer.

The pipeline trims adapters with Trimmomatic (optional), then maps these reads to a transcriptome with Kallisto, to quantify isoform expression for isoform switch analysis. RNA-Seq reads are also mapped to a reference genome with STAR. STAR alignments are extracted to count matrices with Subread for use with DESeq2. STAR alignments are also used for MAJIQ and DEXSeq.(Image copyright: Original image, license: CC-BY-NC-ND 4.0)

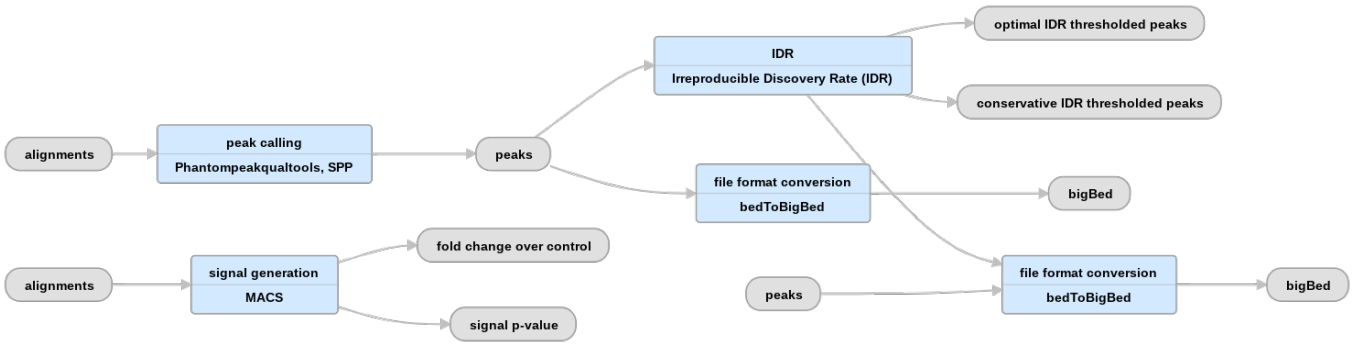


Figure 5.4 ENCODE Processing pipeline for uniform ChIP-Seq peak calling: Peak calling is done by MACS for chromatin elements and SPP for transcription factors.
(Image copyright: Creative Commons CC license.
Image source: encodeproject.org/pipelines/ENCPL138KID/)

6 Results

The aims mentioned in Chapter 3 are addressed in this section.

In silico gold standard for RNA-Seq simulations can now be made with the ASimulatoR described in Section: 5.1. It is also an essential part of the benchmarking workflow of mapping and AS-event detection described in Section: 6.1.2.

Eleven mapping tools aligned synthetic reads of with AS incorporated within it, the results of which will be discussed in the Section: 6.1.4.

Eight AS-event detection tools were then evaluated to: a.) contrast reference-based and reference-free tools in their capacity to identify *de novo* events (see Section: 6.1.5); b) offer guidelines for users to learn how to use AS event detection tools (see Section: 6.1.5 and Section: 6.1.7); c) identify a common unified format that describes commonly found AS-events across all the AS-event detection tools (see Section: 5.3).

Finally, a web-application was developed to identify AS-events in a differential analysis of RNA-Seq datasets, integrating it with ChIP-Seq datasets to identify protein elements that bind to DNA and RNA to induce splicing. The results of which are described in Section: 6.2

6.1 DICAST Benchmark Overview

In the field of alternative splicing (AS) discovery, a new tool is developed every year to discern AS events from RNA-Seq datasets. This growing repertoire of tools needs a comparison and proper benchmark. The benchmark should not just prove that a newly developed tool is better than its current competitors, but also evaluate the current state of performance for each state-of-the-art tool used in the field.

Here we present DICAST, a third party independent benchmark framework for Mapping and AS event detection tools. DICAST is on an open source licence (GNU Public License(GPL) v3) and is free to fork and develop on further, making it a confluence point for new developing tools detecting AS events. These new tools can also easily be incorporated into the modular framework of DICAST for further independent benchmarks. AS-event detection tools quite often require an upstream step of splice-aware alignment. DICAST benchmarks these mapping tools and offers them for the computational workflow required for the evaluation of AS-event detection tools as well.

This section describes the benchmarks carried out by DICAST and the performance of these tools on datasets with increasing sequencing error and sequencing depths, on different AS event types (see Section: 1.4 for more details) and on the *de novo* identification of AS events. For a detailed view on the pipeline workflow, please see Section: 5.2. When not benchmarking, DICAST can also run on real RNA-Seq datasets to describe the fractions of AS events commonly found by tools with UpSet plots (see Section: 6.1.6).

6.1.1 Criteria

Our stringent criteria for mapping/AS-event detection tool selection was that they should be: i.) still be available to download and install; ii) documented; iii) open-source; iv) use standard RNA-Seq analysis software such as FASTQ/FASTA, GFF3/GTF and BAM/SAM files; v) the tool must be used in peer reviewed publications other than the one that describes the tool itself; vi) available as stand-alone software; vii) should be able to work with custom GTF files, for example as built by ASimulatoR.

Within DICAST, while there is the opportunity to tune parameters, we argue that the tools would frequently be used with default parameters. Therefore, for the tool to be widely accepted, the default parameters should be optimal for many scenarios. Furthermore, tuning different parameters from different tools would not be easily comparable across these tools.

Criteria for tools benchmarked for AS event detection included that they: were actively used by the community, maintained, that they used the standard RNA-Seq input files, ran as stand-alone software, and also could use custom annotations. Custom annotations were made part of the criteria as ASimulatoR also outputs a GTF file that represents the AS events contained within the dataset. Splice-aware mapping was frequently an upstream analysis required for AS event detection and therefore the same criteria were applied to curate the list of mapping tools.

6.1.2 Benchmark Workflow

DICAST was run on different datasets simulated by the ASimulatoR. The datasets start off as simple and get more complex (S1-S4) with respect to the kinds and frequency of AS events incorporated within the dataset. Table: 6.1 shown here depicts how complexity of AS events observed in the dataset was incrementally added. S1 starts with one event per transcript. In this dataset, each gene has only one main and one alternative transcript. S1 hosts AS events such as exon skipping (ES), intron retention (IR), alternative 5'-exon usage (A5), alternative 3'-exon usage (A3), multiple exon skipping (MES), alternative last exon (ALE) usage and alternative first exon (AFE) usage and the sequencing error is set to 0%. S2 is a similar dataset but a sequencing error is set to 0.1%. This value is typical for Illumina-based sequencing (citation). S3 allows two or more alternative transcripts per gene. S4 also allows that the same exon is involved in several AS events: e.g., exon skipping and A5'-splice site.

Simulated Dataset	Events per transcript	Transcripts per gene	Events per exon	Event types							Seq error rate
				ES	IR	A5	A3	MES	ALE	AFE	
S1	1	1	1	x	x	x	x	x	x	x	0%
S2	1	1	1	x	x	x	x	x	x	x	0.1%
S3	2	>=1	1	x	x	x	x	x	x	x	0.1%
S4	2	>=1	>=1	x	x	x	x	x	x	x	0.1%
S5	1-4	>=1	>=1	x	x	x	x				0.1%

Table 6.1 Simulated datasets, using ASimulatoR, for the benchmark: Simple AS events in RNA-Seq datasets S1-S4, grow in complexity. Complex dataset S5 simulated, representing AS events observed in real data from SHIP-cohort

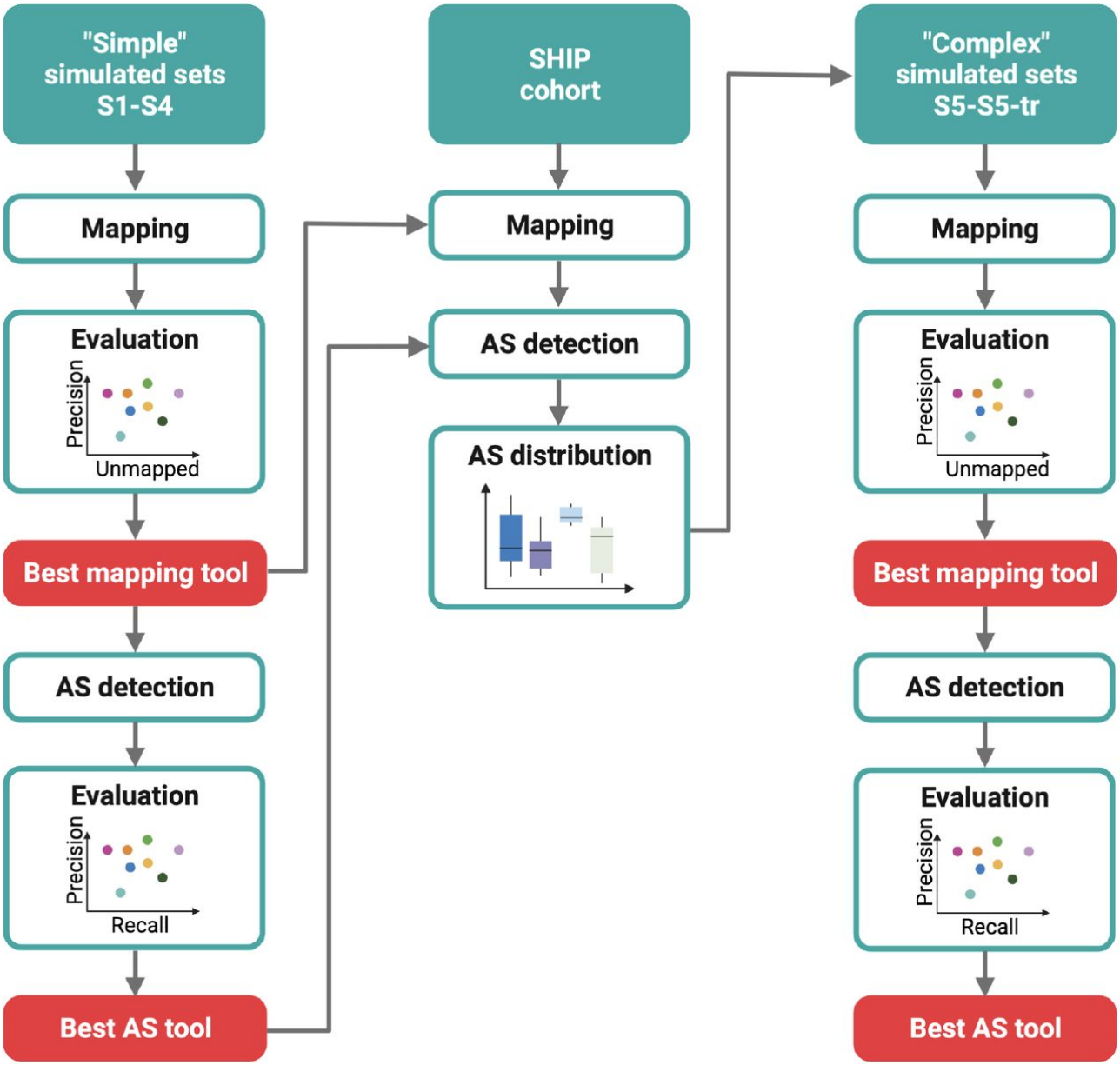


Figure 6.1 Simple simulated datasets S1-S4 created with ASimulator with increasing complexity. The best mapping tools and AS event detection tools, benchmarked in these scenarios are used to detect AS events observed in the SHIP-cohort. Complex simulated datasets S5 are created using the AS distributions learnt from the SHIP-cohort. Dataset S5-tr is also generated with the AS distributions from the SHIP-cohort. S5-tr datasets generate GTF files for only some of the AS event observed in the dataset for the *de novo* event detection experiment. The most informative benchmark is the one performed on realistic dataset S5. (Image copyright: Creative Commons CC-BY-NC-ND license. Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

Using the datasets S1-S4 we evaluated 11 mapping tools and 8 AS detection tools (see Fig:6.1). We identified the tools with the optimal performance.

The datasets S1-S4 have equal probability for all AS events to be incorporated. Usually, it is not the case: e.g., exon skipping is the most frequent event type in mammalians [73]. To account for this unequal probability, we constructed the dataset S5.

The dataset S5 simulates AS events as observed in a real dataset. To estimate the frequency of AS events types in a real dataset, we analysed RNA-Seq reads from SHIP-cohort. The reads were mapped using STAR as STAR showed the best performance with respect to mapping on simple simulations (discussed in Section: 4.2 and Section 6.4). AS events were observed in the SHIP-cohort with MAJIQ, as MAJIQ showed high precision for exon skipping (discussed in Section 6.1.5), the most common type of AS event found in human tissues. The estimated frequencies were used to simulate the dataset S5. MAJIQ detects only four AS event types, thus, it simulates S5 only with these four event types (Table:6.1). S5 also contains 1-4 events per transcript, the same exon can be involved in more than one event, there might be more than one alternative transcript per gene, and sequencing error rate is set to 0.1%.

The entire workflow for the benchmark is depicted in Fig:6.1.

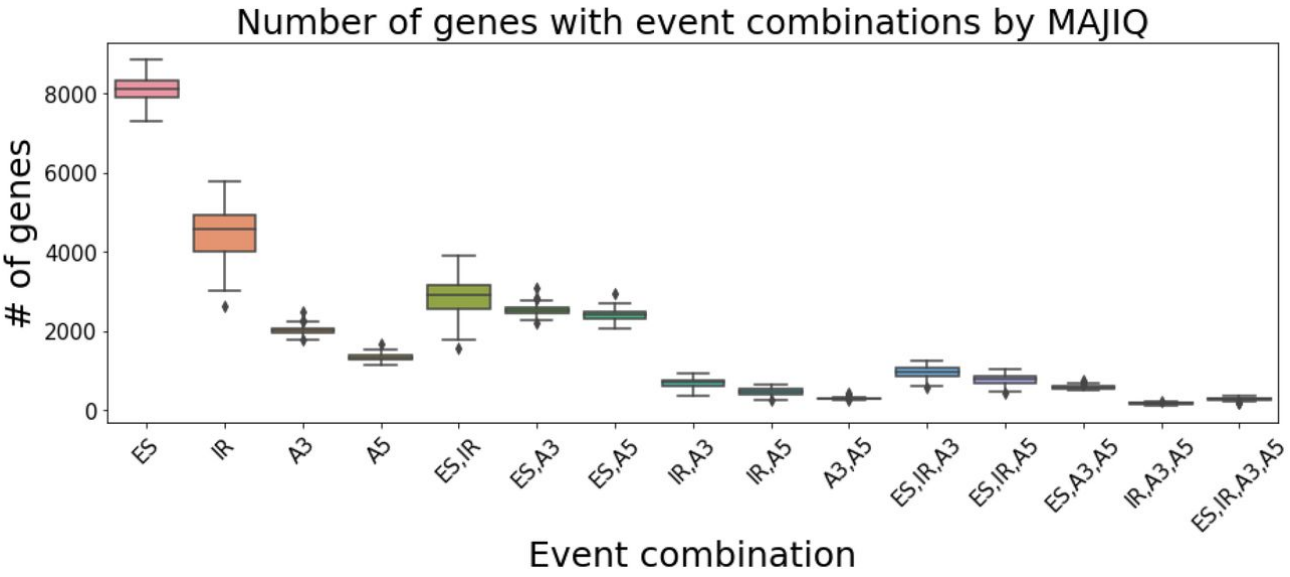


Figure 6.2 AS events as observed in SHIP data were used to inform the distribution of AS events observed in simulated dataset S5.

(Image copyright: Creative Commons CC-BY-NC-ND license.
Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

6.1.3 Runtime of Tools in DICAST

ContextMap2 was designed to map to multiple reference genomes without annotations. It approximates reads that could align to a region and describes them as the region's context first. It then sorts through reads within a context for a better resolution of local alignments. Finally, a global resolution of alignment between contexts is made. This algorithm therefore has a pretty noticeable runtime (see Fig:6.3). Furthermore, ContextMap2 reports having a bad runtime when coupled with bowtie2, but improvements can be made by using bowtie. Minimap2 is the fastest mapping tool available in the repertoire, as it's designed

to map long read sequences, primarily. It operates with the seed-chain-align algorithm and a base-level alignment, which operates as a dynamic program to extend from ends of chains to close regions between anchors which act as seeds. This algorithm was designed for a semi-global alignment.

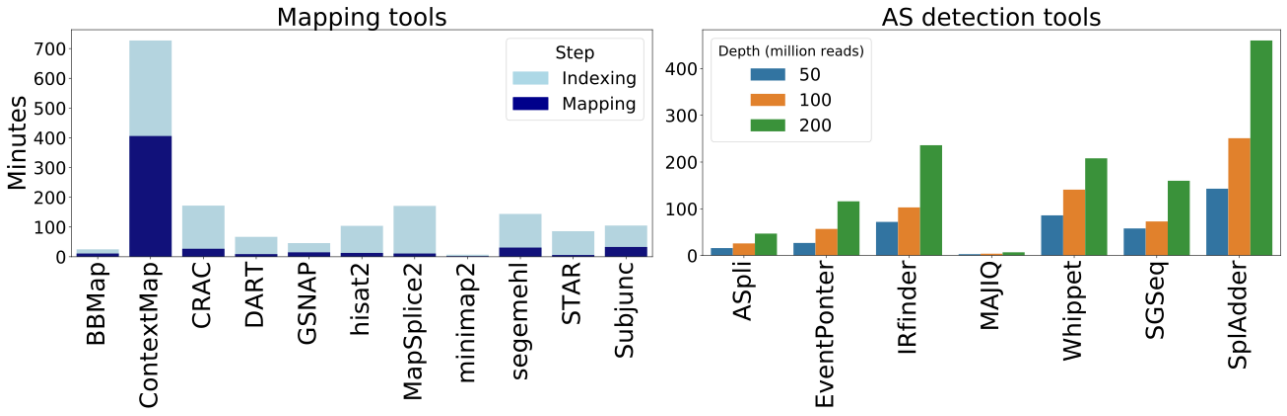


Figure 6.3 Runtimes of Mapping and AS detection tools.
Mapping tools were compared to the same read depth of 10 M reads, while AS tools were diversified to 50, 100 and 200 M reads to see how their runtime performances scale. Stacked bar plots for mapping tools represent the time needed for indexing and mapping. (Image copyright: Creative Commons CC-BY-NC-ND license. Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

AS event detection tools were run at 50, 100 and 200 M reads to understand how well they scale. MAJIQ has the shortest runtimes, with most of it's core algorithm ported to C++. At 200 M reads MAJIQ's run is comparable to the runtime of its nearest contender ASpli at 50 million reads. ASGAL was not included owing to its long runtime (3 days for 50 M reads), because ASGAL takes raw FASTQ files as input, skipping the prerequisite for alignment as many tools do. SplAdder has the next longest runtime. ASGAL and SplAdder are all tools written in python, which could explain why they are slow. IRFinder also takes a long time to run, but an advantage IRFinder has is that it does not depend on an intermediate mapping step for AS event detection. Whippet, one of the winners of the benchmark on AS events is written in Julia, lauded for its concurrency and capacity to parallelise. However, Whippet was developed to run on a single core, leading to relatively long runtimes too.

6.1.4 Mapping Tools

Key features of mapping tools were described in Section 1.6.2, however specific differences between tools are discussed further in this section. In the precision to unmapped reads plots, the best place for a mapping tool to be is where precision is high, but the ratio of mapped reads to unmapped reads is low. Therefore the ideal place to be is in the top left corner. The best performing mapping tools are STAR, HISAT2, MapSplice2 and ContextMap2.

We benchmarked mapping tools on datasets described in Section: 6.1.2 and listed in Table: 6.1. These datasets of growing complexity helps identify the largest influences within a dataset that changes the performance of each of the AS-event detection tools. Overall though, the ranking with respect to both precision and recall of mapping tools seem to not change with these introduced influences. However, DART

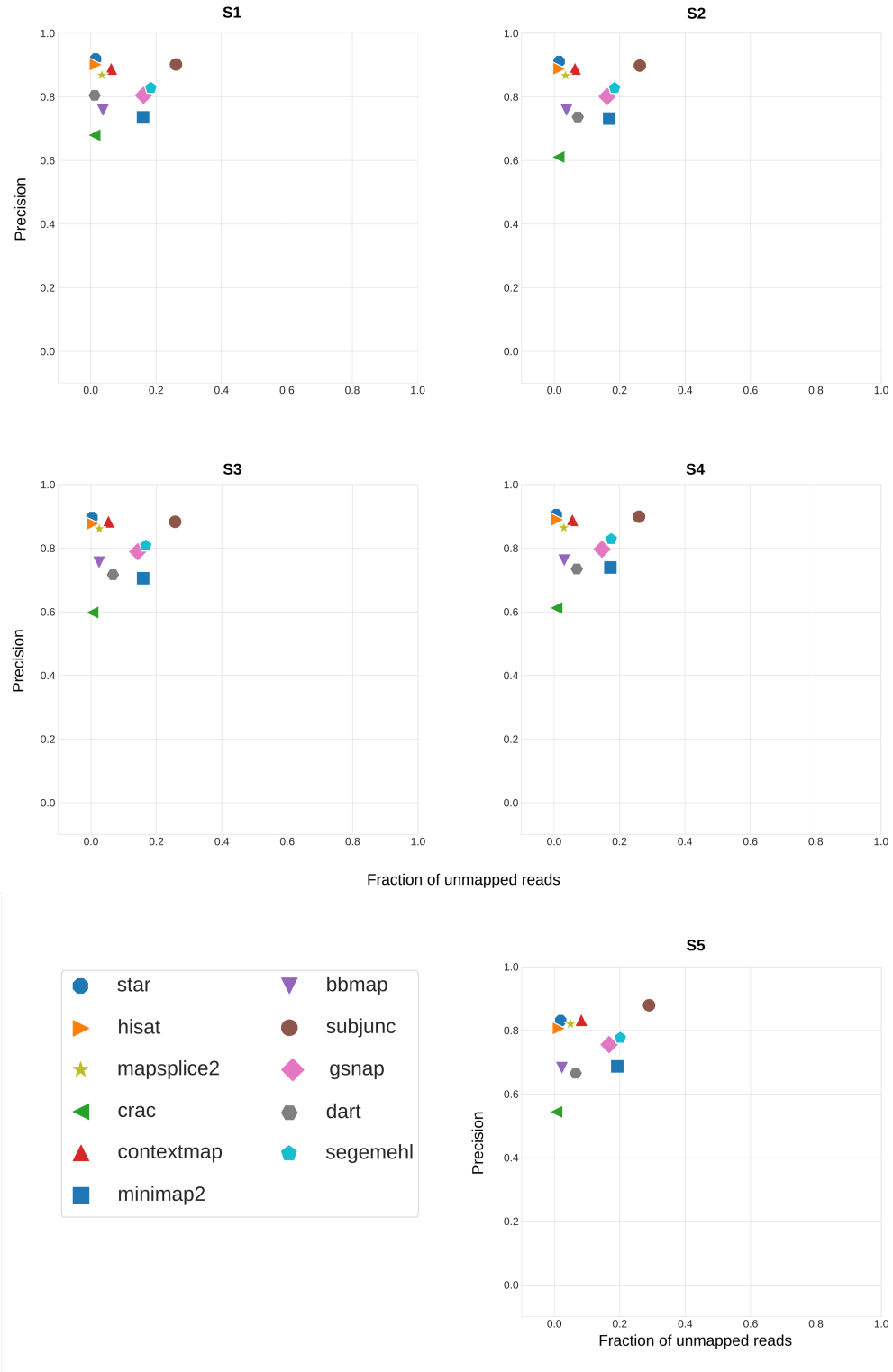


Figure 6.4 Precision vs Fraction of unmapped reads plots of Mapping tools.

Datasets S1-S4 are simulated datasets with increasing complexity in AS events (see table:6.1). S5 dataset represents the AS events distribution as seen in SHIP-cohort (see Fig: 6.1 for an overview on how S5 was made). Ideally, mapping tools should have high precision and a low fraction of unmapped reads.

shows a drop in ranking with respect to precision between datasets S1-S2, suggesting that sequencing error is not well tolerated by DART.

MapSplice2, STAR and HISAT2 use annotated splicing features to make up three of the best candidates for mapping tools. Therefore the use of annotations is the main similarity among the best candidates for mapping tools. Although notable differences in the mapping algorithms still exist, such as the use of a graph based alignment with HISAT, which uses a graph based index to account for small indels in the annotations.

The plot for a complex dataset S5, shows that in realistic datasets, mapping tools such as subjunc can have greater precision than the other mapping tools. While Subjunc usually has a high fraction of unmapped reads, the seed and vote algorithm is more precise than the seed and extend algorithm observed in STAR in the realistic dataset S5. Additionally, Subjunc does not require an annotation, making it a good candidate for aligning RNA-Seq reads to build annotations.

6.1.5 Alternative Splicing Event Tools

A general description of AS-event detection tools is in Section: 1.6.2. This section refers to the performance for each of the AS-event detection tools in key areas. In the precision and recall plots for each of these tools, the ideal place to be is where precision and recall are high, the top right corner.

Overall, most recall values are low for AS-event detection tools. However, the best AS-event tools in this review with good annotations for most use cases are SGSeq, ASpli, Whippet and MAJIQ, in that order. Due to low recall values, Eventpointer and SGSeq (without annotations) cannot be recommended for the detection of AS-events.

Benchmarks of Alternative Splicing Tools per Event Type

As a result of our analysis, we observed that tools tend to have different performances on detecting different AS event types. IRFinder, as the name suggests, works primarily for intron retention. Exon skipping and Intron retention were observed to be the most abundant AS events observed in the SHIP-cohort (see Fig: 6.2). Therefore tools that specialise on these AS events tend to be more popular among researchers who work with human and vertebrate AS. Yet tools that have high precision and recall for exon skipping do not always find introns as well (see Fig: 6.5).

Overall, exon skipping seems to have the highest precision rates across all tools, which could be the focus of AS tool developers. Exon skipping is typically the best way to diversify the protein coding functions within the transcriptome. Alternative splice site challenges appear very similar across the 5' or 3' end. Detecting partial exons seems to be hard, resulting in low recall values for many tools. For partial exons, Eventpointer also has low precision values . Intron retention seems to be hard for many tools. micro RNA (miRNA) and short interfering RNA (siRNA) that are encoded in introns, make this an especially hard task. To consider if such small RNA are expressed within the datasets, tools often have to develop a filter for detecting coverage across the entire intron, just to make sure that the intron is fully present within the dataset.

With annotations, SGSeq and ASpli perform the best across all AS types. ASpli maintains its precision close to 1, across all AS types. SGSeq performs with the best recall rates across the benchmark, however

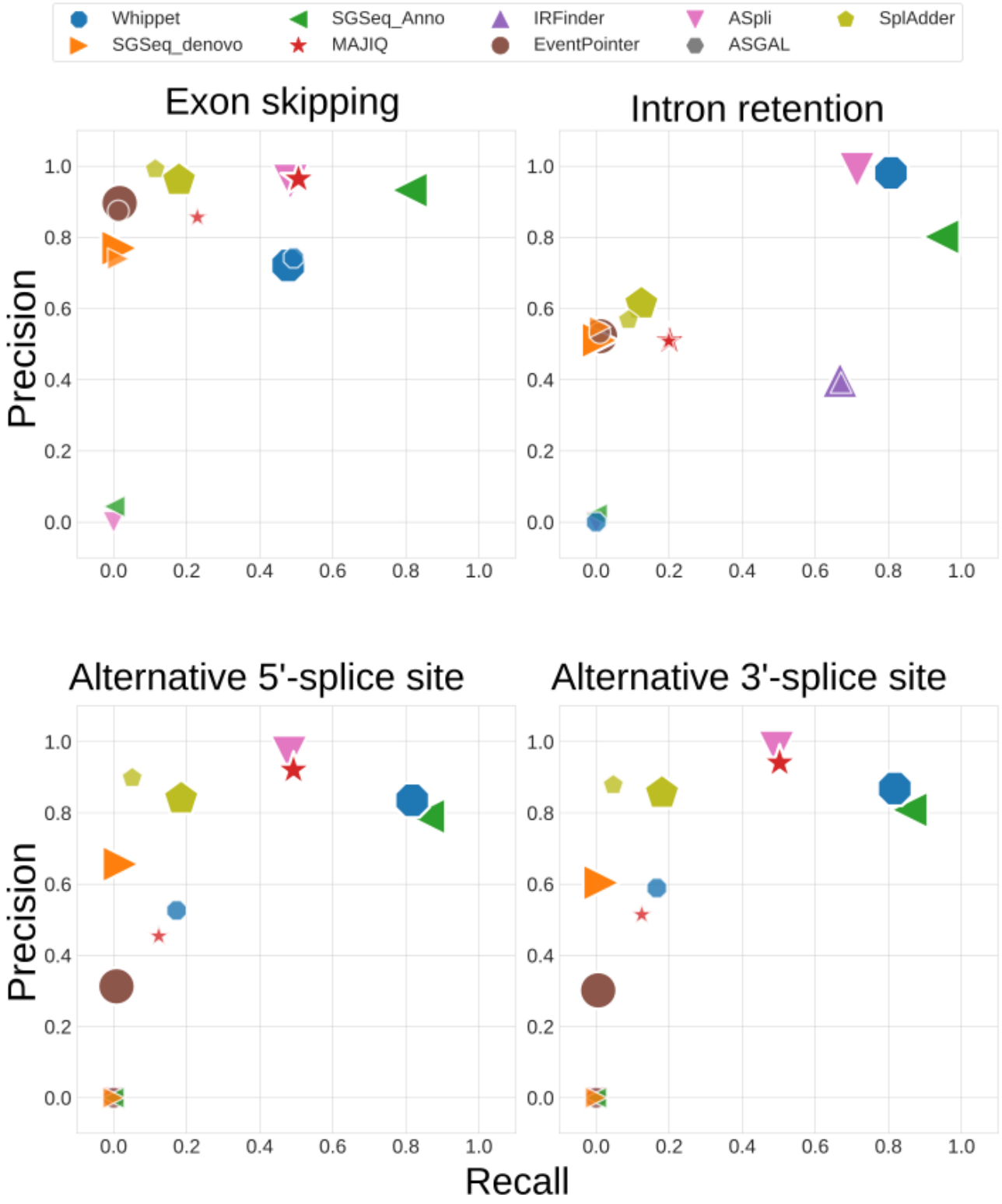


Figure 6.5 Tools benchmarked on S5 dataset across different types of AS.
Smaller symbols show the tool run with truncated annotation files hosting only one transcript per gene, while the dataset expresses multiple transcripts for each gene. This sets the tool up for a de-novo AS event detection experiment. Larger symbols have the support of all the AS events expressed within the dataset. (Image copyright: Creative Commons CC-BY-NC-ND license.
Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

some of the events it reports could have some false positives, resulting in lower precision than ASpli. MAJIQ has very high precision, however it fails to capture nearly half of all events. Its performance in intron retention, however, drops to 0.5 with respect to precision. Whippet has high precision and recall values for alternative splice sites and intron retention, when it has annotations. However whippet's performance in exon skipping, with or without annotations, has low recall.

Benchmarking *De novo* Alternative Splicing Event Finding Feature

De novo AS event finding features of AS event tools are very valuable when annotations for the organism in study are not of high confidence. They can also be beneficial as supportive evidence, when annotations are built by overlaying predicted open reading frames with RNA-Seq reads. I evaluated the performance of the tool to detect AS events *de novo* by using truncated annotation files with only one transcript per gene, while the dataset shows expression of many transcripts per gene. For this evaluation step, I used the most complex dataset S5 with 1-4 transcript per gene. These results are shown as smaller symbols for each tool in the Figure: 6.5.

The results vary greatly for each AS event tool. The recall for each of these tools usually improves with the full annotation as compared to the truncated annotation files. ASpli and SGSeq (anno) have the biggest differences between when it was supplied with full annotations. SGSeq can be run in two modes: anno (when supplied with annotation files) and *de novo* (running without annotation support). There are some tools that do not perform a lot better, even if supplied with full annotations, but only with some types of AS events.

For exon skipping, EventPointer, SGSeq in *de novo* mode, SplAdder and Whippet, perform similarly with and without annotation files. For intron retention, EventPointer, SGSeq in *de novo* mode, SplAdder, MAJIQ and IRFinder all have similar performance with and without annotations.

For alternative exon usage, EventPointer and SGSeq in *de novo* mode exhibit a substantial difference in performance with annotations. Similarly Whippet underperforms drastically without good annotations in detecting intron retention.

Influence of Sequencing Depth

Number of reads generated in the RNA-Seq experiment (sequencing depth or library size) might have an impact on the performance of AS detection especially for genes with low expression level. We investigated the impact of sequencing depth increasing it from the standard 50 M reads used for gene level analysis to 200 M reads. AS tools were fairly consistent in their performance across read depth for the simple data sets S1-S4(see Fig:6.7). Therefore the only observable contrast lies between datasets S1 and S5 (see Fig:6.6). ASGAL however, was not run on 100 M and 200 M reads, as the runtime needed for ASGAL was not scalable and took 3 days for 50 M reads. Furthermore, at 50 M reads, ASGAL was still not a top contender for AS event detection.

Datasets S1-S5 (described in Section: 6.1.2 and listed in Table: 6.1) show modulated differences in AS events, with growing complexity. They have been used to understand key influences of identifying AS events in real data. The biggest differences identifiable across datasets S1 and the realistic S5 are (see

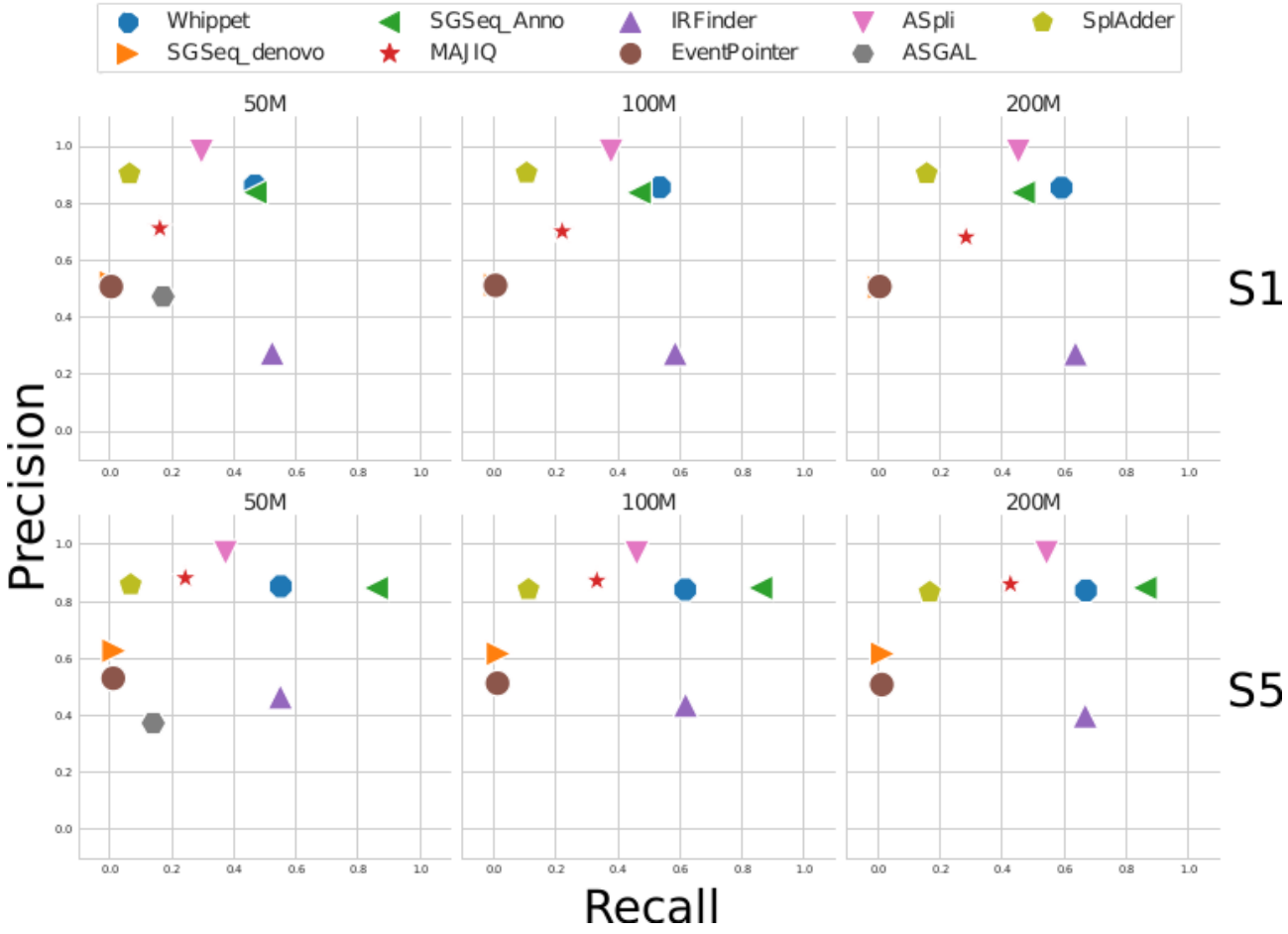


Figure 6.6 Precision and recall plots for each AS tool within DICAST benchmarked at 50 M, 100 M and 200 M read depths for both simulated dataset, the simple S1 and the complex realistic S5 datasets.

The ideal place for an AS event detection tool to be in such a plot, is where the precision and recall values are both high, therefore the top right corner. Additionally a tool is trustworthy when precision values are high, which suggests that the AS events that are reported by each tool are almost entirely composed of accurately identified AS events.

Figure 6.7 Precision and recall plots for each AS tool within DICAST benchmarked at 50 M, 100 M and 200 M read depths for datasets S2-S4. See Table: 6.1

Fig: 6.6): an improved recall value for SGSeq (anno) and MAJIQ; improved precision values for SGSeq (*de novo*) and MAJIQ.

These observed improvements in recall are most prominent in dataset S5, suggesting that they are correlated with the amount of AS within the dataset, as S5 was generated to resemble realistic datasets. The observed improvements to precision, however, arise between datasets S3 and S4, suggesting that SGSeq (*de novo*) and MAJIQ are more precise tools to handle multiple AS events per exon.

6.1.6 UpSet Plots

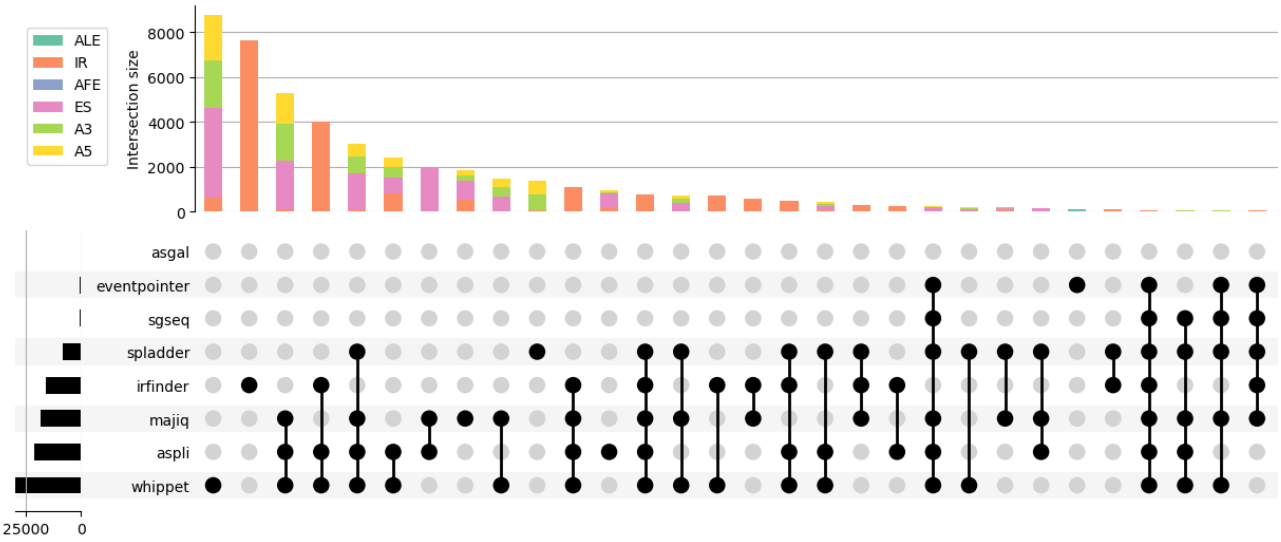


Figure 6.8 UpSet Plot, generated by DICAST in every run.
Columns indicate intersections of reported events. The stacked bar plots show the ratios of AS events reported within the intersection. The bar plots on the rows show the length of the outputs produced by each tool. (Image copyright: Original image, license: CC-BY-NC-ND 4.0)

Identified AS events across the different tools are reported together in UpSet plots. UpSet plots are generated by DICAST and are built on the unified output translated by DICAST (see Fig:6.8). UpSet plots also inform users of how many AS events are represented by each of the tools. This allows users to pick similar or greatly varying tools for combinations. It allows users to see the intersections of the events described by tools that have high precision scores. The stacked bar plots represent ratios of AS events by type. We developed a customised version of the UpSetPlot python with stacked bar plots as a feature, which was contributed upstream, in the spirit of open-source (https://github.com/jnothman/UpSetPlot/pull/137). Finally the UpSet plot allows users to focus on AS events that they are interested in to identify the tools they could use to report AS events better.

In general we observe that the tools that have the highest recall values are the tools that have the most AS events reported. Their intersections tend to follow. An intersection between Whippet, ASpli and IRFinder as the tools with the highest recall with respect to intron-retention as the only event usually comes next. However it's concerning to see that even among the tools with the best recall, intersections, or AS events found in common is more rare than events reported by any one of the tools.

6.1.7 Using Alternative Splicing Tools in Combination

Across AS event detection tools more unique events have been reported than events that are found in common within the same dataset. Having developed a unified output for each of the tools (see Section: 5.1), gives us the opportunity to use these AS event tools in a combinatorial approach. Therefore we combined the output of tools to make precision and recall plots for each AS type (see Fig:6.9). For the combination analysis, only those tools that can detect AS events de novo and demonstrated reasonable recall/precision on S1-S5 datasets were considered. The choice of tools here could possibly be further improved to select tools that have highest precision or recall, from the benchmarks.

Two approaches for combining tools were used: 'AND' approach means that an event is correct if it is found by all of the combined tools; 'OR' approach means that an event is correct if it is found by at least one of the combined tools.

AS event tools integrated within DICAST were taken in combinations of AND and OR. When AS events were identified with two tools with the AND logic, the results of two tools were intersected and plotted on precision and recall. AS events were identified with two tools with the OR logic, when the union of the results of two tools were plotted on precision and recall.

When tools were combined with the AND approach, they have a greater precision than the initially used tools, by themselves. However, this approach also resulted in lower recall.

When tools were combined with the OR approach, they were observed to have a lower precision than the separate tool within the combination. However, recall of the combination of tools resulted in comparable recall performance.

A combinatorial effect on recall does not perform much worse than the tool with the greater recall values. However, the combinatorial approach could be used to achieve greater precision than any of the tools used by themselves. This further demonstrates the value of the unified format for reporting for AS event tools.

6.2 DASiRe

DASiRe is a web tool and a preprocessing pipeline aimed to assist users with integrating RNA-Seq and ChIP-Seq datasets for an enrichment test to search potential splicing factors or DNA binding proteins [77] that regulate alternative splicing. The approach used to quantify splicing in RNA-Seq experiments include the alignment-based, isoform usage based and exon usage based approaches. Differential expression of splicing factor genes give an overview of expressed splicing factors in the sample. While most of the analysis runs on a user's local machine, the results are visualised in the web server. To demonstrate navigation of the web page and to show a use case, an example dataset is incorporated within the web server.

This chapter discusses results as is seen on the online example dataset. Here we used DASiRe to analysed a YBX-1 CRISPR/Cas-System knockout K562 cell line and compare it to the peaks observed in the Encyclopedia of DNA Elements (ENCODE)'s publicly available ChIP-Seq experiments targeted at splicing factors also with the K562 cell lines.

This was a dataset chosen to show an ideal system where the differences in splicing could be attributed to a specific protein involved in splicing. Associations of splicing with their splicing factors could also be explored with DASiRe on other experimental designs such as two tissues from the same biological sample.

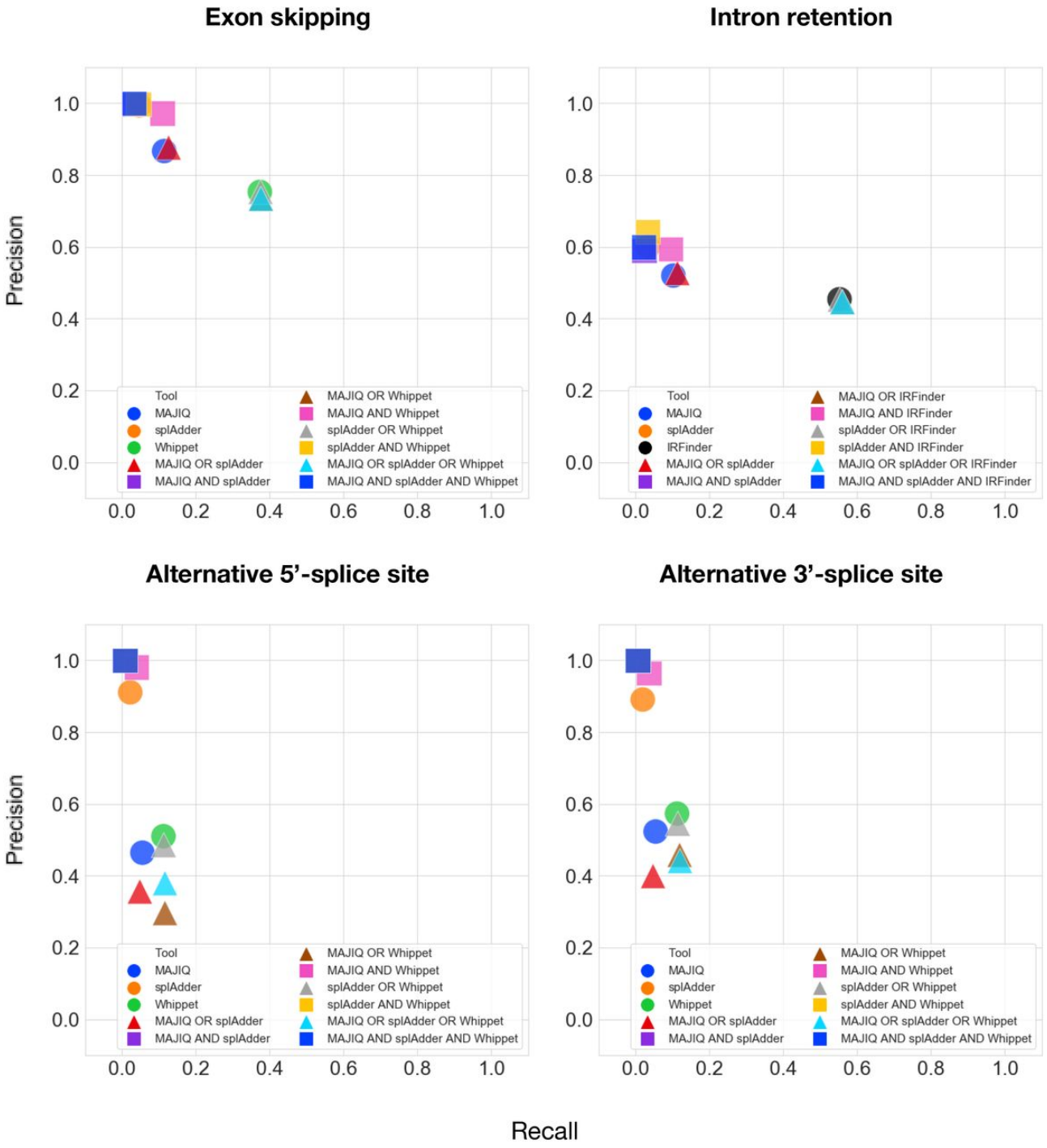


Figure 6.9 Precision and recall plots for tools in combination of AS-event detection tools, run on dataset S5.

The precision and recall for each AS event type was separated in a plot for each event type.

(Image copyright: Creative Commons CC-BY-NC-ND license.

Image source: biorxiv.org/content/10.1101/2022.01.05.475067v1)

However since the regulation of splicing involves other factors such as chromatin accessibility or stress, datasets with proteins involved in splicing knocked out are better to study the role of splicing for that specific protein.YBX-1 was chosen as it binds to the Y-box motif observed in many genes [31]. The Y-box appears to be involved in transcription as the sequences share common elements also found in Kozak sequences associated with transcription initiation. YBX-1 is known to interact with the transcriptional machinery and shown to assist with mRNA stability [97].

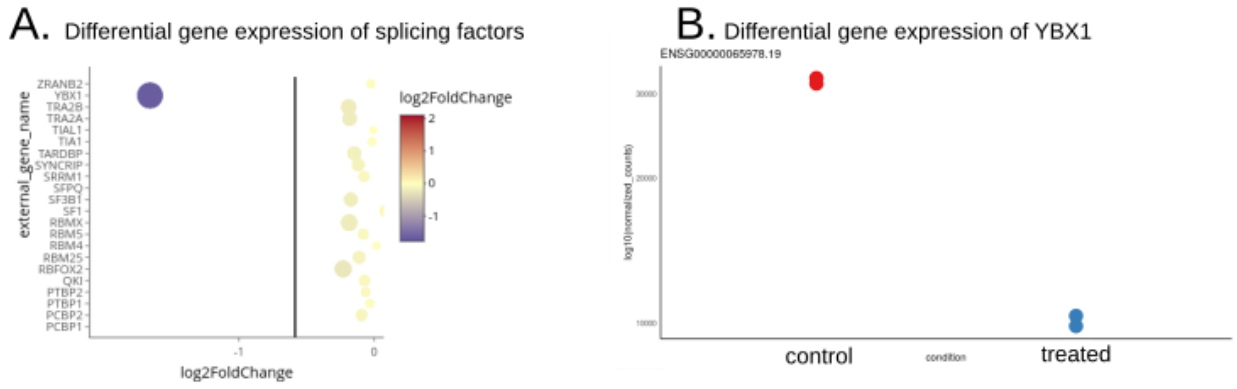


Figure 6.10 A. Differential gene expression of splicing factors; B. Differential expression of gene YBX1 confirms knockdown in RNA-Seq experiment

6.2.1 Confirmation of Knockout Experiment in RNA-Sequencing

As described in Section 5.4.3, key visualisations of DASiRe's RNA-Seq data analysis shows us plots of normalised gene counts across the contrasted condition. Querying for the gene that was the target of the CRISPR/Cas-System for gene knockouts, YBX-1, we observe in Fig:dasire-results.B, that in the targeted CRISPR/Cas knockout, YBX-1 is repressed. DASiRe visualises differential expression plots of any gene within the dataset, here we use it to confirm knockout experiments with simple differential expression plots.

Since YBX-1 is also a splicing factor [97], it is represented within the genes queried as differentially expressed splicing factors. We can also observe it in the plot for differential expression of splicing factors in Fig:6.10.A. Here, the log2foldChange value shows that the expression of YBX-1 is indeed significantly repressed.

6.2.2 Investigation of Gene Targets Binding by YBX-1

In order to explore targets of splicing induction by YBX-1, we observe an example target gene TANGO2. TANGO2 has 3 ChIP-Seq peaks for YBX-1 within it's genomic positions and was therefore identified as a target gene. TANGO2 was also significantly differentially spliced (see exons marked in pink Fig: 6.11.B) Fig: 6.11.C shows in blue genomic positions where ChIP-Seq peaks showing binding of YBX-1 within the genomic positions of the TANGO2. These genomic positions are usually retrieved from the BED files uploaded to DASiRe.

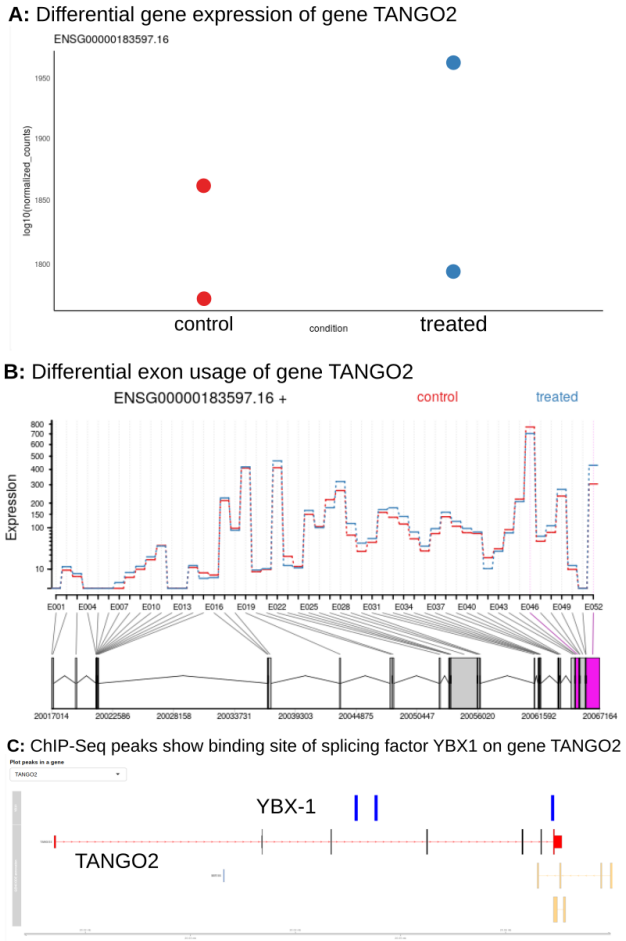


Figure 6.11 A. Differential expression of a target gene TANGO2; B. Differential exon usage of target gene TANGO2; C. ChIP-Seq peaks show evidence of binding of YBX1 at the genomic position of TANGO2

TANGO is a transport and Golgi organisational protein that, when mutated, is shown to reduce the lifespan of patients and cause severe arrhythmia of the heart.

Differential gene and exon expression of TANGO2 can be observed in Fig: 6.11.B-C. DASiRe identifies the last exons of TANGO2 to be differentially spliced, however the gene is not differentially expressed. ChIP-Seq peak overlap indicates a region that could be influenced by binding of YBX-1 (see Fig: 6.11.C).

6.2.3 Enrichment Analysis of ChIP-Sequencing Peaks in Spliced Genes

In order to learn more about the proteins involved in splicing, we utilise datasets of ChIP-Seq experiments of splicing factors from the ENCODE ChIP-Seq repository. These are available as gene tracks you can load on DASiRe. However, users may upload further files for experiments with ChIP-Seq performed elsewhere to DASiRe as well. The columns in the visualisation of enrichment analysis for ChIP-Seq (see Fig: 6.12) would depend on the datasets in use. In the case of the example dataset, we observe that the regions that are associated with the binding of HNRNPH1, PCBP2, SFPQ do not observe significant splicing patterns. The dataset with the knocked-down gene YBX-1, shows very little splicing in the regions overlapping ChIP-Seq peaks presented in the example YBX-1 ChIP-Seq data within DASiRe.

Furthermore, DASiRe shows potential splicing factors that could be involved in the splicing observed within the knockdown experiment (see Fig: 6.12). It identifies other splicing factors within the presented ChIP-Seq data to identify potential binding sites within splicing genes. A good candidate for a splicing factor responsible for splicing within the RNA-Seq dataset represented here is RBFOX2. This splicing factor shows significant enrichment with the colour of the tile, where the odd ratio is estimated to be 1.3, for both differential exon usage and isoform switch events.

6.2.4 Splicing Factors Involved in Splicing in a Knockdown Experiment of YBX-1

Publically available datasets downloaded from the ENCODE website were used to show an example use case of DASiRe. In this dataset, the gene YBX-1 was knocked out using CRISPR/Cas-Systems. YBX-1 is

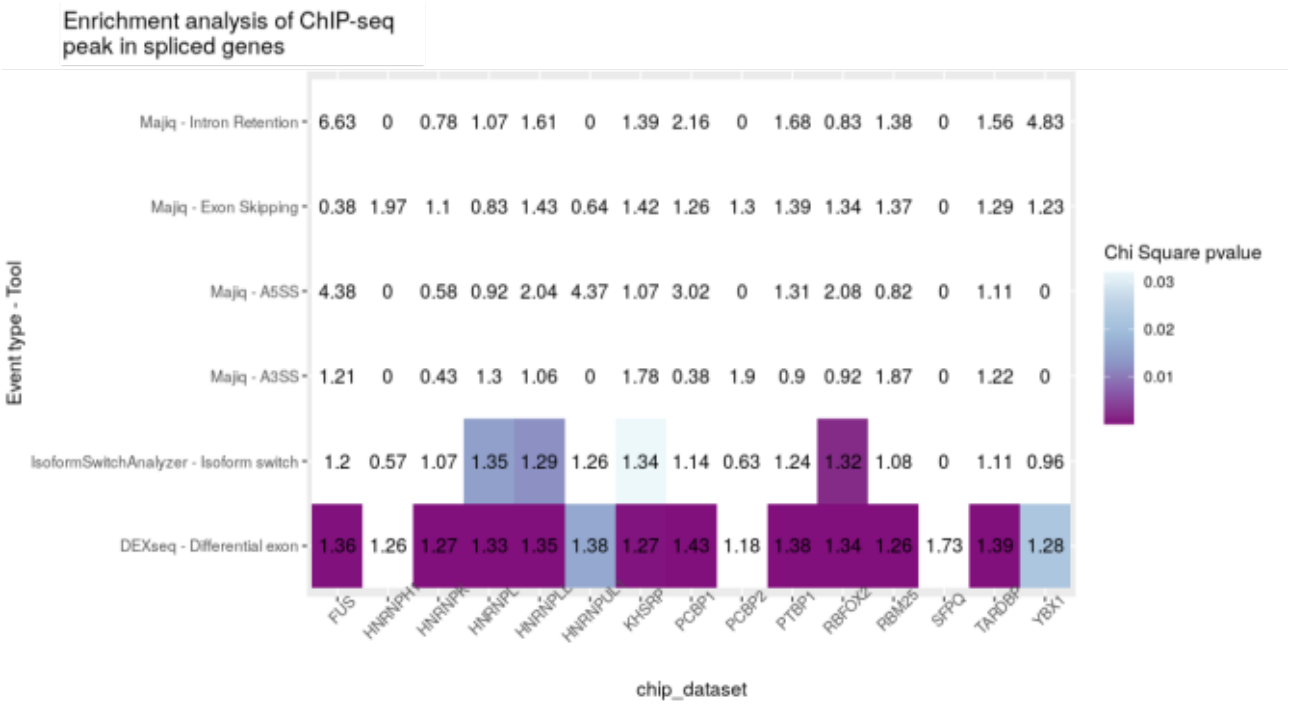


Figure 6.12 Enrichment analysis of ChIP-Seq peak in spliced genes
(Image copyright: Original image, license: CC-BY-NC-ND 4.0)

known to interact with the translational machinery and shown to assist with mRNA stability. The effect of knocking out YBX-1 was hypothesised to have a large amount of differential effects in splicing.

DASiRe was described in Section: 6.2 to confirm that the gene YBX-1 was indeed knocked out. This was achieved in the panel on differential expression of splicing factors, and in the plots for normalised count value for the gene YBX-1. Binding target of YBX-1, TANGO2, shows an example of the gene that was influenced by AS. Exon usage values show differential usage of the last exon in gene TANGO2. ChIP-Seq binding of YBX-1 shows that the last exons are accompanied with binding sites for the splicing factors, without which the mRNA stability could have been compromised.

The ChIP-Seq peak enrichment heatmap (see Fig: 6.12) show that YBX-1 in a knockout cell line, shows no evidence of enriched splicing in the genes that bind to YBX-1, DASiRe also identifies RBFOX2 as a candidate protein for further knockout studies, whose binding causes splicing within the cell lines.

7 Discussion

In this study, we addressed three main challenges:

1. Datasets from currently available RNA-Seq simulators may host different isoforms from each gene, but these reads are generated at random [52]. Many AS event detection tools therefore focus on identifying exons within a read to identify transcripts, but cannot identify AS events such as intron-retention or alternative splice sites as well.

2. Currently available benchmarks of AS event detection tools are conducted by tool developers, who focus on precision of their tools in specific tasks, but the measure of precision alone fails to address weaknesses of their tool. These vulnerabilities can only be highlighted by a third party benchmark due to conflict of interest.

3. Integration of ChIP-Seq and RNA-Seq tools focus on identifying motifs of binding sites and functionally interpreting transcription factors or nucleosomes. With emerging protocols such as CLIP-Seq being developed for RNA-binding proteins, computational analysis of ChIP/CLIP-Seq data and RNA-Seq are being combined to identify splicing factors [41, 156]. However a systematic investigation of the functional role of RNA-binding proteins through this integration is still missing for the study of AS.

As a source of ground truth needed to calculate performance metrics such as precision and recall for the algorithms in the study of AS, we developed ASimulatoR, an RNA-Seq simulation tool for a benchmark of splice-aware mappers and AS events detection tools. ASimulatoR builds on the current state-of-the-art RNA-Seq simulator polyester and additionally allows users to modulate the amount and distribution of AS events within an RNA-Seq dataset. Thus ASimulatoR can be used to create simple datasets with one event per gene, or complex datasets, where one can decide which kinds of AS events and at which frequencies can be observed. ASimulatoR was published in the journal Bioinformatics [100].

We benchmarked 11 splice-aware mappers and 8 AS event detection tools with an independent modular framework DICAST (Docker Integrated Comparison of Alternative Splicing Tools) made with Snakemake and with isolated docker environments. DICAST can work with real data and evaluates AS events found in common or run on simulated datasets from ASimulatoR, to offer a benchmark of AS event tools for a modulated distribution of AS. DICAST is already published on a preprint server at BioRxiv [40] and is currently under review at a peer-reviewed international journal.

Splicing is regulated by proteins that are primarily RNA binding proteins, while some show DNA binding capacity as well [76]. By comparing the effect of binding of these proteins by ChIP/CLIP-Seq (both simplified as ChIP-Seq for readability), to the AS events identified in a matching RNA-Seq sample, we could explore the functional role of these proteins in AS. Proteins involved in the regulation of AS (see Section: 1.3), have been shown to affect alternative splice-sites, exon skipping, intron retention and alternative last and

first exons, while current analysis of ChIP-Seq focuses only on genomic regions such as the whole gene or promoter regions. In our novel approach, implemented in DASiRe (https://exbio.wzw.tum.de/dasire/) (manuscript in preparation), we combined AS events quantified on three levels. a) event-based b) exon-based, and c) isoform-based levels, with state-of-the-art tools. Users can upload RNA-Seq datasets and evidence of binding sites from ChIP-Seq to the web application and evaluate splicing factors associated with AS in the uploaded dataset.

In this chapter, I discuss key findings of the benchmark and the identification of splicing factors.

7.1 Potential for Biases in Simulated RNA-Sequencing Data with Modulated Alternative Splicing, Using ASimulatoR

Gold standard datasets required for this benchmark were simulated with ASimulatoR. With ASimulatoR, simulated datasets can be modulated to give one transcript with one AS event per gene or multiple AS events per gene leading to multiple transcripts. The use of reference annotations for an organism, allows ASimulatoR to use splice-junctions as previously observed to create new artificial transcripts from intermediate exon-supersets for each gene.

However by creating exon-supersets, ASimulatoR ignores splicing features of intronic regions beyond that, which could still host splicing sequences that could be used by AS event detection tools to identify retained introns. It may therefore prove limited when using new computational approaches to identify AS event detection that use sequence features found in junctions or poly-A sequences in introns as features in deep learning applications. However, while sequences could determine the strength of splice sites, regulation of splicing is shown to be implicated by epigenetic factors [171]. Therefore identification of splicing should not need to rely on these sequence features much.

The regulation of complex exon skipping, such as multiple exon skipping or alternative exon usage is influenced by proteins (see section: 1.3). This suggests that sequence features that could play a weak role in complex AS events also may have sequence elements that interact with proteins. ASimulatoR for example, currently doesn't replicate the mechanism of mutually exclusive exons. However, since these events usually go beyond the scope of short read sequencing ASimulatoR stays relevant to AS event detection tools that work with short read alignments. ASimulatoR may prove to be limited in generating realistic long read sequences.

Within the benchmark, ASimulatoR was used to create 4 simple simulated datasets (see Table: 6.1) with increasing complexity of AS events within the datasets. Dataset S1 represented the simplest scenario of 1 event per transcript, 1 transcripts per gene events per exon and 0% sequencing error. A dataset was then generated to understand the influence of each factor: increasing the complexity of sequencing error to 0.1% (S2), having more transcripts per gene (S3), more events per exon (S4). ASimulatoR was also used with prior knowledge of a real dataset, the SHIP-cohort, to create a realistic dataset S5, which simulates the number and distribution of AS events observed in the real dataset. These datasets enable testing of tools like the AS event detection and splice-aware mapping tools under various splicing conditions such as datasets with evidence of small amounts of splicing to splicing as seen in realistic datasets.

While each step appears to contribute a share of the complexity, each condition was tested in growing complexity. This is only reproducible if the order of comparing the influence across the datasets is main-

tained. To study the complexity in depth, as observed in realistic datasets and simulated in dataset S5, more scenarios of isolated influences must be generated and tested upon. Then a combination of two of the influences and so on until all the combinations we test are in one dataset: S4.

7.2 Independent Modular Benchmarking of Alternative Splicing Event Detection Tools and Mapping Tools

Comparing each AS event tool can be challenging. Especially when considering that each tool requires its own programming languages, versions of programming languages, different library versions for each programming language and finally even compatible packages. Therefore due to the steady need for benchmarking of new AS event tools developed each year, DICAST requires a modular approach. Maintaining reproducibility of a running system requires freezing the compute environment along with the versions while needing to stay agnostic of the compute platform for a flexible use case. To this end DICAST uses docker to isolate compute environments, making each tool behave like an executable module. Enabling parallelization in a workflow system, where the order of tool execution needs to be maintained, is done by a POSIX compliant workflow manager, Snakemake. Tool developers who wish to participate in DICAST for an independent evaluation, would be required to make a docker module for their tool, along with reporting in the unified format. All outputs in the unified format can then be tested against datasets from ASimulatoR for precision and recall values in comparison with other tools in DICAST. A copy-left licence, GNU Public License (GPL), brings a central github for collaborative code-sharing, allowing DICAST to integrate new tools for evaluation of AS events, collectively furthering the benchmark.

DICAST could also be run with real datasets, as benchmarks for AS have previously done [109]. Intersections in UpSet plots (see Section: 6.8) show the consistency of an AS event being detected by multiple tools. However, as the plots also show, most events that are reported by AS event detection tools are either unique or shared by at least three tools. We can only see clear trade offs between the tools and recall values for false negatives, when running DICAST on synthetic data. Evaluation of splice-aware alignment tools as the basis for AS analysis Splice aware mapping tools that fit the criteria referenced in Section: 6.1.1, were benchmarked and were included in DICAST (see Table: 4.1). Key strategies and differences among the splice-aware mapping tools were discussed in Results in section: 6.1.4. These mapping tools were benchmarked against datasets S1-S5 (see Section: 6.1) and the results were summarised as precision and fraction of unmapped read plots (see Fig: 6.4). Dataset S1, the simple dataset has 1 AS event for 1 transcript per gene with 0% sequencing error rate. This is an elementary dataset to map to a genome.

The worst performing mapping tool with respect to precision is CRAC, which also, incidentally stopped being supported since the development of DICAST but runs effectively within the docker module. Minimap2 notably is also among the worst performers, despite being very popular to align long read sequences. Owing to the intended design for long read sequence alignment, Minimap2 does not try to optimise local alignments, but rather focuses on global alignments. Further evaluation of Minimap2 with a long-read RNA-Seq simulator is required to see if it can stay competitive. However, when aligning short-read RNA-Seq, Minimap2 is not a contender for a good mapping tool. The main difference between the mapping tools were identified as their dependence on a reference annotation.

Three of the four recommended splice-aware mappers rely on a reference genome annotation: HISAT2, STAR and MapSplice2. HISAT2 identifies splice sites to be extracted for its splicing graph based alignment. STAR reportedly does not require a reference annotation, and yet uses it when available to extract information about splicing junctions. MapSplice2 uses mapped alignment segments, which are fractions of reads, that anchor to exons. It therefore uses annotations to determine which alignments are anchor segments and runs a spliced alignment for segments between the anchors. The use of splicing features such as exons or splice-junctions allow these tools to perform really well compared to all the other mappers that are within the benchmark.

ContextMap compares well to these reference annotation based tools, however does not require an annotation. With the exception of ContextMap, most other reference-free tools do not compare to the annotation aware software for mapping in both precision and fraction of unmapped reads. The exception is Subjunc which tends to perform more precisely than STAR does with annotations, when subject to realistic distributions of AS within a dataset such as seen in the dataset S5.

Both approaches have their merits. When considering organisms with a good reference annotation such as human or murine models. However, operating on an organism that is poorly annotated, annotation-free aligners are preferred because annotating a genome includes identifying the splice-junctions and the exons within a gene.

7.2.1 Benchmark of Alternative Splicing Event Detection Tools

AS event detection tools such as the ones listed in Table: 4.1 have met the same criteria as mapping tools referred to Section: 6.1.1. These tools were incorporated within DICAST and benchmarked against the datasets referred to in Table: 6.1. The datasets are derived from simulations of simple transcriptomes with low number of AS events per gene S1-S4, and of more realistic transcriptomes in datasets S5 and S5-truncated. Realistic datasets are simulated incorporating AS events as frequently as seen in real data, as observed in the SHIP-cohort. The benchmark is summarised in the figures:6.6-6.7 and 6.1.5.

One of the most striking results is the low recall values for almost all AS detection tools. While precision values may be high, low recall values suggest that AS event detection tools are still challenged when it comes to recovering all the events of AS within the simulated datasets. One reason for the low recall value might be the low expression of transcript variants. In the Section 6.6, we demonstrated that sequencing depth affects recall, but has a low impact on precision.

When we increased the complexity of the simulated datasets, we did not observe a difference in ranking of the tools between S1 - S4. However, we noted the change in performance that was observed between the datasets S1 and S5. The largest difference between simple datasets S1-S4 and the complex dataset S5 is the number of AS events simulated. Another difference is the number of AS events simulated per exon in dataset S4 and S5 as compared to S1-S3. These parameters of RNA-Seq simulations seem to have the largest influence for performance for each tool. Exon skipping was also identified as the most abundant AS event type identified in the SHIP-cohort (see Fig: 6.2). Notably the ranking of Whippet and SGSeq (with annotations) switched between datasets S5 and S1-S4 with SGSeq with annotations outperforming Whippet in S5. This coupled with lower recall values for Whippet compared to SGSeq in every AS event type observed, especially exon skipping, suggests that differences were exaggerated by the amount of AS events simulated and especially the amount of exon skipping in the S5 dataset.

However, the change in ranking observed by SGSeq, run in *de novo* mode and EventPointer, between simplistic dataset S1 and realistic dataset S5 could be explained by the increase in certain event types such as Alternative 3' and 5' splice-sites, where SGSEq is more precise than EventPointer. This highlights what AS event tools already seem to be good at, exon skipping, being the biggest factor in identifying all AS events in a dataset.

AS event tools were also tested for their *de novo* event detection capacity. AS tools were given the dataset with a truncated version of the annotation file. The annotation file described each gene with just one isoform, while the dataset showed evidence for the expression of different isoforms per gene. If a tool is able to identify a new splice variant, then it does so without the help of annotations. The results of this study were summarised in the plot Fig: 6.5, with precision and recall values of AS tools when run with a truncated annotation file represented as the smaller version of the symbol for each tool. Specificity wasn't considered for this evaluation, because ASimulatoR generates short-read sequences for alignment. In a scenario without annotations, short-read RNA-Seq cannot guarantee reads that cover splice junctions, as the read lengths are often shorter than exons or introns. Also when looking for alternative splice sites, short-read RNA-Seq may not provide sufficient evidence for identifying new splice-sites effectively without annotations. When compared to an annotation, evidence of RNA-Seq alignment can be compared to known splice-sites to collect more robust evidence of exon skipping or the use of alternative splice-junctions. Annotations were found to increase both the precision and recall of plots in most cases and primarily recall in the case of splAdder. The tools also varied in the kinds of AS events that they could detect without annotations and how much of the dataset they could accurately recover.

With the unified format outputted by each of the tools, we combined the AS events that were detected by each tool with both the 'AND' and 'OR' approaches. When AS events detection tools are combined with the 'AND' approach, the reports increase in precision of each of the tools. This method may be beneficial to identify narrow down candidate AS events for experimental validation, for higher confidence reporting of AS events. When the same tools report with the 'OR' combination, the reports gain in recall. This approach may be used exploratively for studying a systemic phenomenon, however, precision values suffer with this technique.

7.2.2 Compatibility of Splice-aware Mapping Tools Serve as Input to Alternative Splicing Detection Tools

The output of splice-aware mapping tools serve as input to AS detection tools. All mappers output SAM format, which seem to be standardised, but splice-aware mappers have an additional set of standards for reporting alignments. This is because AS event detection tools also map the reads, a more optimised local alignment. The AS event detection tool therefore tries to maximise the information it receives from mapping tools. For a splice-aware mapping tool to have wide scale adoption, it should try and tend to the needs of every AS event detection tool. This option of presenting more information could be used further for new AS event detection tools.

As a reference, the attributes from STAR [28] are used to describe these crucial attributes needed for all of the AS event tools listed in Table: 4.1 to work smoothly. This is because STAR pushed the standards of SAM alignments as set by the SAM/BAM Format Specification Working Group. These attributes of SAM files refer to the following:

1. HI Query hit index
2. XS alignment stand (now offered as flag:TS by the SAM/BAM Format Specification Working Group)
3. AS Alignment score generated by aligner
4. NM Number of mismatches in each mate
5. NH Number of reported alignments that contain the query in the current record
6. nM is the number of mismatches per (paired) alignment

A full compatibility table for which splice-aware mappers work with which AS event tools, please refer to the table in the DICAST documentation at the following link: https://dicast.readthedocs.io/en/master/tools/tools.html.

7.2.3 Recommendations for Using Splice-aware Mappers and Alternative Splicing Event Detection Tools

While there was not a clear winner of the benchmark for AS event detection, these results show the strengths each tool has to offer. Recommendations therefore depend on the use case.

Annotations also play a massive role when it comes to the performance of splice-aware mappers and AS event detection tools. Therefore the state of annotations could affect how well one could observe AS events. AS tools may behave differently in other organisms, especially when AS events themselves look different [73]. While the primary AS events in human transcriptomes are exon skipping, plant transcriptomes have a majority of intron retention [123]. The intron definition approach to regulation of AS events is common in invertebrates. DICAST could have a different result for benchmarking for every new genome or even with new annotations [26].

With recall being harder to achieve, a benchmark aimed at getting the best precision for each AS event type could help inform user's choice of a simulated dataset with ASimulatoR. The tools with the best precision and recall values *de novo* depend upon the kind of AS event users are interested in: exon skipping is best detected with Whippet; intron retention with IRFinder; alternative 5' and alternative 3' splice sites are best recognised with splAdder.

These tools combined in the OR approach could improve recall and help users describe the distributions of AS events to simulate a realistic dataset with ASimulatoR, in order to have a ground truth that resembles the challenge. DICAST could make a new report of precision and recall plots for each organism, giving its users studying splicing an updated third-party benchmark with their genome of interest.

When working with RNA-Seq datasets from human and murine samples, users can expect complete annotations and can work with STAR, and SGSeq with annotations.

When working with RNA-Seq datasets from unannotated organisms, Subjunc's seed and vote approach worked out to have higher precision than even STAR's seed and extend approach did with annotations. A lower fraction of unmapped reads could be achieved without annotations if one uses ContextMap. However, the choice between a very precise mapper that is very precise, but doesn't align all the reads, versus an aligner that maps all the reads observed with lower precision depends on the use-case.

7.2.4 Reporting Alternative Splicing Events: Metrics and Formats

Different tools report results differently. Reports of AS events don't agree on types of AS events observed, for example alternative last/first exon, mutually exclusive exons or multiple exon skipping can all be described as exon skipping events. Each tool also has a challenge with respect to the formats the report AS events in. While an attempt was made by ASGAL to return to the format of alignment such as SAM/BAM formats, other attempts such as the LSV from MAJIQ, demonstrate the challenge of reporting AS events by reporting them with sections of splice graphs as observed in a read. Furthermore analysis downstream of AS event detection becomes linked to a new format described by one of the many AS event detection tools that are available [147].

Most AS event detection tools have low recall. We show that tools, when used in AND or OR combinations, can report AS events with greater recall (see Section: 6.1.7). Comparing reports of AS can be made easier by a common unified format for reporting AS events. DICAST further supports docker modules that report AS events in this format for 8 AS event detection tools.

The unified format combined commonly found features of reported AS events to this goal. The unified format is described in table: 5.1 and it reports chromosome, gene, strand, event type, count and coordinates for start and end of the splice event, along with an id for the event. These were among all the common elements described in each of the AS event tool's outputs and were valuable for the downstream analysis within DICAST. While the unified format described here works towards a goal for a unified format across the study of alternative splicing, there is much room for growth.

Better reports can be prepared, if the same format also includes 'percent spliced in' (PSI) values reported by the majority of the AS event tools. The PSI value is a percentage score that describes AS events from a genomic alignment window. If it's exon skipping, PSI describes how many of the RNA-Seq reads show evidence of mapping to the exon in the form of a percentage. PSI value can be used to define an exon usage quantification or a section of a splice graph for event level tools (see Section: 1.6). It cannot apply to approaches that quantify AS such as isoform usage, because it lacks genomic positions to describe a window

Other metrics need to be more widely accepted among AS event detection tools as well. Whippet offers a measure of entropy that could quantify AS on a gene level. Quantifying splicing is usually an orthogonal measure for a gene, when compared to its gene expression. Especially because AS events are reported on an event level which is smaller than gene level. Furthermore, genes that are under influence of a changing landscape of splicing factors stoichiometrically, could express many isoforms. Genes could also stably express multiple isoforms and have multiple gene products for some conditions. Entropy offers a gene level measure of abundance in isoforms or the different types of AS events. The value 0 describes a uniform expression of all the splice variants within a gene. Value 1 represents a skewed distribution of some isoforms being highly expressed, compared to all the splice variants a gene has.

Altogether this describes a growing collection of metrics involved in splicing, developing further metrics shouldn't come at a compromise in the approaches already established. The unified format mentioned here bolsters the use of commonly existing metrics and should be developed collaboratively, rather than having one per tool developed.

7.3 Identifying Proteins Involved in Splicing with DASiRe

Splicing is influenced and regulated by proteins such as histones that form nucleosomes, the transcription complex, the spliceosome and many transcription factors. They bind both to DNA and RNA to interact with the splicing machinery to carry out splicing effectively.

Regulation of splicing by proteins have been shown to affect alternative splice-site usage, exon-skipping and intron retention (see Section: 1.3). DASiRe combines state-of-the-art tools to bring together a unique way to identify the genes that are undergoing splicing through the various approaches as described in Section: 1.6. This approach addresses the challenge of splicing being regulated at multiple regions that are smaller than the gene or even the choice of transcript used with an isoform-switch detection.

A growing repertoire of adaptations of ChIP-Seq, initially designed for DNA binding proteins, have been developed for RNA-binding proteins as well, such as CLIP-Seq [86], individual-nucleotide resolution CLIP-Seq (iCLIP-Seq) [137] and enhanced CLIP-Seq (eCLIP-Seq) [157]. Studies of splicing factors have previously used RNA-Seq and ChIP-Seq data from matched samples to identify functional roles for splicing factors [167, 56, 171]. For readability, we will summarise these protocols as ChIP-Seq. While many tools exist for motif extraction from ChIP-Seq experiments [166, 56], a combined RNA-Seq analysis has yet to see a systematic approach to the analysis of proteins involved in splicing. The DASiRe web server was developed for biologists to gain insights about the proteins involved in AS regulation.

DASiRe aims to identify genomic regions that have both evidence of AS and binding sites of splicing factors from ChIP-Seq. Binding sites of splicing factors from the ENCODE database are incorporated within DASiRe. A feature to include uploadable BED files to describe binding sites from the user's ChIP-Seq experiments is currently in development. To quantify AS within the transcriptome, DASiRe pre-processes the RNA-Seq dataset by exon usage, AS events and isoform usage on the user's local machine. These results can then be analysed on the DASiRe's web server. As described in Section: 6.2, DASiRe outputs can be used to plot results such as: differential gene expression of all genes or just splicing factors, differential exon usage, isoform switch detection, AS event detection, ChIP-Seq binding on a gene track and ChIP-Seq peak enrichments.

A Fisher's enrichment test for each of the splicing factors helps identify which regions are associated with both the binding of splicing factors and with splicing. A Fisher's enrichment is then run on genomic regions such as observed binding sites (ChIP-binding), observed AS event detected (AS-event), which is then compared to the rest of the genes reported by the tool, without a significant AS event being detected (non-AS-Event) and regions of non-AS-Event genes without ChIP-binding (non-ChIP-binding). Therefore leaving out regions of ChIP-Seq binding causing AS in a neighbouring exon/intron within a gene.

Currently as an example dataset on DASiRe, we use publicly available datasets from ENCODE. To demonstrate the purpose of our web-tool, a well-studied protein involved in AS has been chosen for investigation [145]. YBX-1 has been shown to be involved in the sorting and transport of tRNA and miRNA which are typically hosted within introns of other genes. YBX-1 transcription factor with many targets and is shown to bind to commonly found sequences in promoters. YBX-1 gene is an oncogene and all proteins it interacts with are also oncogenes, making YBX-1 a common target for drugs [120]. And finally YBX-1 is shown to bind both to DNA and RNA elements and therefore was also a good candidate for ChIP-Seq analysis [97].

In our analysis, DASiRe shows that when YBX-1 is knocked out, in K562 cell lines, then it is no longer involved in splicing. However, we should consider a condition with the K562 cell line as case and the K562 cell line with YBX-1 knocked-out as control. This could build a better background model to identify AS events that are associated with YBX-1 binding. Differential analysis requires a contrast to describe change in gene expression while ChIP-Seq files describe binding sites within one of the conditions. This builds a dependency that no longer describes a contrast, but rather a static view, consequently ChIP-Seq binding sites should be describing the condition: case. A good background model should best describe a contrast in functional binding, such as a kock-out experiment. A differential analysis contrasting tissue types could also show how splicing is associated with the different splice factors in DASiRe in one of those tissues.

7.3.1 Further Development

A robust analysis requires stringent filters, low expressed genes are quite often noisy and DASiRe's pre-processing pipelines require further development to incorporate a filter for low expressed genes. Further steps of development include: a q-value filter for DEXSeq in order to correct for multiple testing hypotheses, when comparing exons within a gene.

DASiRe is currently still under development and therefore does not yet accept user inputted ChIP-peak files. Also further improvement is needed for the ChIP-Seq peak visualisation in quality control with a track for each of the splicing factors within the ENCODE ChIP-Seq data in DASiRe. Further minor improvements are still pending, while the manuscript is currently under internal review.

7.4 Containers Assist Reproducibility of Bioinformatics Analysis

Container technologies since docker have brought a robust and yet lightweight solution to reproducibility with respect to library management and runtime. Containers do not have as many resource requirements as more robust solutions such as virtual machines do. Container technologies such as docker can also be used to preserve application data and separate user level processes, crucial for multi-user environments such as the high-performance computing (HPC) environments used in bioinformatic faculties.

Docker established the field of software container technology, by bringing in the first stable automated programming interfaces and daemons required to run many containers within the same computing environment. Docker still has the largest infrastructure in terms of an image registry and set the ground with the initial standards that were useful for the open container initiative by the Linux Foundation. These open-source standards for container technology then permeated beyond docker to other container technology that is specialised in niche environments: IBM offers Podman for the fedora ecosystem that specialises in security; Sylabs offer Singularity, for HPC systems; Canonical offers Linux containers (LXC) for lightweight Ubuntu based containers. The open-container standards that apply to all containers allow for portability between the container technologies, especially with docker as a starting point. Therefore containers such as the ones that are within DICAST could also be ported to other HPC friendly environments such as singularity or podman.

DICAST and DASiRe both use docker environments to keep the methods of these analyses reproducible. DASiRe is separated into two different containers, one for the web-server and one for pre-processing

scripts on a user's local machine. This allows for reproducible and distributed computing. DASiRe's pre-processing docker could also be pulled from docker's common public image repository, Dockerhub. DI-CAST as well, hosts each of the mapping tools and the AS event detection tools within docker containers that are available for public use and can be downloaded from Dockerhub as well. This approach allows for the versioning of libraries and binaries to be frozen in container images, such that production environments still work as intended much after tools within them are no longer supported.

However Docker as a choice of container technology has its limitations. Docker is widely unpopular in HPC environments, where anyone who has permissions to use the docker Automated Programming Interface also has the access to change their user id, due to c-group isolations of user level processes. Docker containers also need to replicate libraries commonly used in linux distributions and therefore are considered too bulky by many users. Despite these drawbacks, Docker remains popular in web server implementations, for its user and user process level isolation and competitive elements of scalability.

8 Outlook

This thesis describes 3 tools offered for the study of AS: ASimulatoR, DICAST and DASiRe.

By simulating short-read RNA-Seq experiments with modulated distributions of AS, ASimulatoR builds gold standard datasets for benchmarking. In combination with DICAST, AS event detection can be optimised to a specific reference genome and annotation. This can perhaps be used to test out AS event detection tools in various organisms with different mechanisms of AS. In plants like Arabidopsis thaliana, for example, intron retention is the most frequently observed AS event and therefore a different set of AS event detection tools would have higher precision or recall values. Similarly invertebrates, such as Drosophila melanogaster, follow an intron definition mechanism of AS and it is likely to have a different set of AS event detection tools as ideal. Therefore ASimulatoR and DICAST could make a report of AS event detection across the model organisms.

Docker containers in DICAST currently aren't popular container technology for wide-scale adoption. DICAST would benefit from reduced image sizes by basing images on lightweight container technology and porting the images to us in singularity or a podman, for more HPC friendly compute environments.

DICAST was initially designed to work with differential AS tools as well. This would further develop the common unified format to include PSI values, which is crucial to the study of AS. Differential AS tools could therefore be easily integrated within DICAST's workflow structures. In exchange for a third party benchmark, tool developers offering the unified output format, DICAST could use its GPL licensed code to become a central point for further collaborative development of AS tools within the benchmark. This brings an additional need to focus on developing the best metrics and format to report AS.

DASiRe the preprocessing pipeline and the web-server can be used for quick integration of RNA-Seq data along with ChIP/CLIP-Seq binding sites to identify splicing factors involved in splicing. DASiRe can be used to systematically compare the splicing factor differences across a differentiating cell line, to learn about which splicing factors are involved in the development of cell types. This requires extensive experimental validation, but currently observed ChIP-Seq data with matching RNA-Seq 'case' samples could help validate the role of histone modifications or splicing factors in various tissues.

A systematic look at AS within my doctorate also includes co-authorships not mentioned in this thesis, they include the bioinformatic tools named SPyCONE [89] and NEASE [93]. SPyCONE is a time series splicing analyser that uses protein-protein interaction networks to identify similar splicing patterns in an RNA-Seq dataset. NEASE combines structural information with pathway analysis within a protein-protein interaction (PPI) network towards a functional enrichment. Enrichments are calculated based on the edges of PPI that cover a pathway of genes compared to edge based perturbations showing splicing structural changes.

Other tools that were developed in collaboration also carry the study of AS further. DIGGER [92], a tool developed for confirming PPI networks with domain-domain interaction networks from co-crystallography

subject to X-ray diffusion studies confirm high-confidence edges in the PPI. Domino [82] is an active module identification tool that eliminates the false positives that PPI networks bring. The study of AS is growing as the field of transcriptomics develops as well. New avenues in spatial transcriptomics [74, 98], may soon show how splicing programs in different cell types within the same tissue splice the same proteins differently. Long read sequencing is also getting more robust and cheaper, identifying new splice variants that have never been recorded before [18, 20].

The study of alternative splicing is developing further and will become more relevant as our approaches to resolve high throughput data improves. I hope this thesis affirms that reproducibility in this field of research is still a vital goal.

Publications

Publications as First Author

1. **Amit Fenn,** Olga Tsoy, Tim Faro, Fanny Rössler, Alexander Dietrich, Johannes Kersting, Zakaria Louadi, Chit Tong Lio, Uwe Völker, Jan Baumbach, Tim Kacprowski, Markus List, Alternative splicing analysis benchmark with DICAST, *bioRxiv* 2022.01.05.475067; https://doi.org/10.1101/2022.01.05.475067, *Under review*

2. Olga Lazareva, Manuela Lautizi, **Amit Fenn,** Markus List, Tim Kacprowski and Jan Baumbach, Volume 1, 2021, Pages 224-233. Multi-Omics Analysis in a Network Context. *Systems Medicine.* https://doi.org/10.1016/B978-0-12-801238-3.11647-2

Publications as Co-Author

1. Quirin Manz, Olga Tsoy, **Amit Fenn,** Jan Baumbach, Uwe Völker, Markus List, Tim Kacprowski, ASimulatoR: splice-aware RNA-Seq data simulation, *Bioinformatics,* Volume 37, Issue 18, 15 September 2021, Pages 3008–3010, https://doi.org/10.1093/bioinformatics/btab142

2. Zakaria Louadi, Maria L. Elkjaer, Melissa Klug, Chit Tong Lio, **Amit Fenn,** Zsolt Illes, Dario Bongiovanni, Jan Baumbach, Tim Kacprowski, Markus List and Olga Tsoy, Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases. *Genome Biology* 22, 327 (2021). https://doi.org/10.1186/s13059-021-02538-1

3. Chit Tong Lio, Zakaria Louadi, **Amit Fenn,** Jan Baumbach, Tim Kacprowski, Markus List and Olga Tsoy, Systematic analysis of alternative splicing in time course data using Spycone, *bioRxiv* 2022.04.28.489857; doi: https://doi.org/10.1101/2022.04.28.489857, Preprint under review

Conference Abstracts

1. **Amit Fenn,** Olga Tsoy, Tim Faro, Fanny Rössler, Alexander Dietrich, Johannes Kersting, Zakaria Louadi, Chit Tong Lio, Uwe Völker, Jan Baumbach, Tim Kacprowski and Markus List, Alternative splicing analysis benchmark with DICAST, *German Conference on Bioinformatics,* September 6-8, 2022, Germany, Halle, Poster Presentation

2. **Amit Fenn,** Olga Tsoy, Tim Faro, Fanny Rössler, Alexander Dietrich, Johannes Kersting, Zakaria Louadi, Chit Tong Lio, Uwe Völker, Jan Baumbach, Tim Kacprowski and Markus List, Alternative Splicing Analysis Benchmark with DICAST. *e:MED Meeting on Systems Medicine, BMBF Online Conference,* September 2021, Virtual Digitalposter Presentation

Manuscript in Preparation, as First Author

1. **Amit Fenn,** Chit Tong Lio, Marisol Salgado-Albarran, Rodrigo González-Barrios, Ernesto Soto-Reyes, Olga Tsoy and Jan Baumbach, DASiRe: Exploring the relationship between DNA-binding proteins and alternative splicing regulation

Abbreviations

General abbreviations

3D	Three Dimensional
A	Adenine
A3	Alternative 3'-splice site / Alternative 3'-exon usage
A5	Alternative 5'-splice site / Alternative 5'-exon usage
AFE	Alternative First Exon
ALE	Alternative Last Exon
API	Automated Programming Interface
AS	Alternative Splicing
ATP	Adenosine Triphosphate
BAM	Binary Alignment/Map
BED	Browser Extensible Data
bp	Base Pair
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transform
C	Cytosine
C terminal	Carboxyl-terminal region of a protein
cDNA	Complementary Deoxyribonucleic Acid
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
CLIP-Seq	Cross Linking Immunoprecipitation Sequencing
CPU	Central Processing Unit
CRISPR/Cas	Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-Associated Protein
CTD	Carboxyl-terminal Domain
DASiRe	Direct Alternative Splicing Regulator Predictor
ddNTP	Dideoxynucleotide Triphosphate
DICAST	Docker Integrated Comparison of Alternative Splicing Tools
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
eCLIP	enhanced CrossLinking and ImmunoPrecipitation
ENCODE	Encyclopedia of DNA Elements
ES	Exon Skipping
G	Guanine
GCC	GNU C Compiler
GFF3	General Feature Format Three

GHZ	Gigahertz
GLIBC	GNU C Library
GNU	GNU's Not UNIX
GPLv3	GNU General Public License version 3.0
GPU	Graphics Processing Unit
GTF	Gene Transfer Format
HGP	Human Genome Project
hnRNP	Heterogenous Nuclear Ribonucleoprotein
HPC	High-Performance Computing
IBM	International Business Machines
iCLIP	Individual-nucleotide resolution CrossLinking and ImmunoPrecipitation
IR	Intron Retention
LSV	Local Splice Variation
LXC	Linux Containers
M	Million
MACs	Model-based Analysis for Chromatin Immunoprecipitation Sequencing
MCF-7	Michigan Cancer Foundation-7
MEE	Mutually Exclusive Exons
MES	Multiple Exon Skipping
mRNA	Messenger Ribonucleic Acid
miRNA	micro Ribo Nucleic Acids
NEASE	Network-based Enrichment method for AS Events
N terminal	Amino-terminal region of a protein
NGS	Next Generation Sequencing
nt	Nucleotide
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
POSIX	Portable Operating System Interface
PPI	Protein-Protein Interaction
PSI	Percent Spliced In
PTB	Polypyrimidine Tract Binding
PTC	Premature Termination Codon
qPCR	Quantitative Polymerase Chain Reaction
RGB	Red Green Blue
RNA	Ribonucleic Acid
RNA-Seq	Ribonucleic Acid Sequencing
mRNA	Messenger Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
tRNA	Transfer Ribonucleic Acid
S	Svedberg unit
SAM	Sequence Alignment Map

SHIP	Study of Health in Pomerania
siRNA	short interfering Ribo Nucleic Acids
SLURM	Simple Linux Utility for Resource Management
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonucleoprotein
SPyCONE	SPlicing-aware time-COurse Network Enricher
STAR	Spliced transcripts alignments to a reference
T	Thymine
tRNA	Transfer Ribonucleic Acid
U	Uracil
UTR	Untranslated Region

Protein names

CARM1	Coactivator-Associated Arginine Methyltransferase 1
CHD1	Chromodomain-Helicase DNA-binding 1
CLK	Cdc2-like Kinase
H2A	Histone 2A
H2B	Histone 2B
H3	Histone 3
H3K27me2	Histone 3 dimethyl Lysine27
H3K36me2	Histone 3 dimethyl Lysine36
H3K36me3	Histone 3 trimethyl Lysine36
H3K4me3	Histone 3 trimethyl Lyines4
H4	Histone 4
HNRNPH1	Heterogeneous nuclear ribonucleoprotein H1
HNRNPK	Heterogeneous Nuclear Ribonucleoprotein K
HNRNPL	Heterogeneous Nuclear Ribonucleoprotein L
MORF	Multiple organellar RNA editing factor
MRG15	MORF-related gene 15
PABPN1	Poly A binding protein 1
PCBP2	Poly(rC)-binding protein 2
PTBP1	Polypyrimidine tract-binding protein 1
RBFOX2	RNA Binding Fox-1 Homolog 2
RNA-Pol II	Ribonucleic Acid Polymerase Two
SF2/ASF	Serine Arginine rich splicing factor
SFPQ	Splicing factor proline and glutamine rich
SR	Serine and Arginine-rich
STAGA	SPT3-TAF(II)31-GCN5L acetylase
TANGO2	Transport And Golgi Organization 2 Homolog
U1	Uridine-rich small nuclear RNA component of a spliceosome
U12	Uridine-rich small nuclear RNA component of a spliceosome

U2	Uridine-rich small nuclear RNA component of a spliceosome
U4	Uridine-rich small nuclear RNA component of a spliceosome
U5	Uridine-rich small nuclear RNA component of a spliceosome
U6	Uridine-rich small nuclear RNA component of a spliceosome
YBX-1	Y-box binding protein 1

List of Figures

List of Tables

Bibliography

[1] Aaron maxwell. *Bash Strict Mode*. URL: http://redsymbol.net/articles/unofficial-bash-strict-mode/#expect-nonzero-exit-status (visited on 10/02/2022).

[2] M. Aebi and C. Weissman. “Precision and orderliness in splicing”. In: *Trends in Genetics* 3 (Jan. 1, 1987), pp. 102–107. ISSN: 0168-9525. DOI: 10.1016/0168-9525(87)90193-4. URL: https://www.sciencedirect.com/science/article/pii/0168952587901934 (visited on 09/14/2022).

[3] Mariano Alló et al. “Control of alternative splicing through siRNA-mediated transcriptional gene silencing”. In: *Nature Structural & Molecular Biology* 16.7 (July 2009). Number: 7 Publisher: Nature Publishing Group, pp. 717–724. ISSN: 1545-9985. DOI: 10.1038/nsmb.1620. URL: https://www.nature.com/articles/nsmb.1620 (visited on 09/14/2022).

[4] Mohammed Alser et al. “Technology dictates algorithms: recent developments in read alignment”. In: *Genome Biology* 22.1 (Aug. 26, 2021), p. 249. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02443-7. URL: https://doi.org/10.1186/s13059-021-02443-7 (visited on 09/19/2022).

[5] Adam Ameur et al. “Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain”. In: *Nature Structural & Molecular Biology* 18.12 (Dec. 2011). Number: 12 Publisher: Nature Publishing Group, pp. 1435–1440. ISSN: 1545-9985. DOI: 10.1038/nsmb.2143. URL: https://www.nature.com/articles/nsmb.2143 (visited on 11/01/2022).

[6] Simon Anders, Alejandro Reyes, and Wolfgang Huber. “Detecting differential usage of exons from RNA-seq data”. In: *Genome Research* 22.10 (Oct. 1, 2012). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 2008–2017. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.133744.111. URL: https://genome.cshlp.org/content/22/10/2008 (visited on 09/18/2022).

[7] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III”. In: *Journal of Experimental Medicine* 79.2 (Feb. 1, 1944), pp. 137–158. ISSN: 0022-1007. DOI: 10.1084/jem.79.2.137. URL: https://doi.org/10.1084/jem.79.2.137 (visited on 08/17/2022).

[8] Danny Bergeron et al. “Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to PABPN1 Autoregulation”. In: *Molecular and Cellular Biology* 35.14 (July 15, 2015). Publisher: American Society for Microbiology, pp. 2503–2517. DOI: 10.1128/MCB.00070-15. URL: https://journals.asm.org/doi/full/10.1128/MCB.00070-15 (visited on 10/30/2022).

[9] Susan M. Berget. “Exon Recognition in Vertebrate Splicing ()”. In: *Journal of Biological Chemistry* 270.6 (Feb. 10, 1995). Publisher: Elsevier, pp. 2411–2414. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.270.6.2411. URL: https://www.jbc.org/article/S0021-9258(17)30674-9/abstract (visited on 10/30/2022).

[10] Gregory Bird, Diego A. R. Zorio, and David L. Bentley. “RNA Polymerase II Carboxy-Terminal Domain Phosphorylation Is Required for Cotranscriptional Pre-mRNA Splicing and 3-End Formation”. In: *Molecular and Cellular Biology* 24.20 (Oct. 2004), pp. 8963–8969. ISSN: 0270-7306. DOI: 10.1128/MCB.24.20.8963-8969.2004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC517882/ (visited on 11/01/2022).

[11] Thomas Bonfert et al. “ContextMap 2: fast and accurate context-based RNA-seq mapping”. In: *BMC bioinformatics* 16.1 (2015), pp. 1–15.

[12] Paul L. Boutz, Arjun Bhutkar, and Phillip A. Sharp. “Detained introns are a novel, widespread class of post-transcriptionally spliced introns”. In: *Genes & Development* 29.1 (Jan. 1, 2015). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 63–80. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.247361.114. URL: http://genesdev.cshlp.org/content/29/1/63 (visited on 10/30/2022).

[13] Daniel Branton et al. “The potential and challenges of nanopore sequencing”. In: *Nature Biotechnology* 26.10 (Oct. 2008). Number: 10 Publisher: Nature Publishing Group, pp. 1146–1153. ISSN: 1546-1696. DOI: 10.1038/nbt.1495. URL: https://www.nature.com/articles/nbt.1495 (visited on 11/01/2022).

[14] Ulrich Braunschweig et al. “Widespread intron retention in mammals functionally tunes transcriptomes”. In: *Genome Research* 24.11 (Nov. 1, 2014). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1774–1786. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.177790.114. URL: https://genome.cshlp.org/content/24/11/1774 (visited on 10/30/2022).

[15] Nicolas L. Bray et al. “Near-optimal probabilistic RNA-seq quantification”. In: *Nature Biotechnology* 34.5 (May 2016). Number: 5 Publisher: Nature Publishing Group, pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: https://www.nature.com/articles/nbt.3519 (visited on 10/02/2022).

[16] Ashley Byrne et al. “Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells”. In: *Nature Communications* 8.1 (July 19, 2017). Number: 1 Publisher: Nature Publishing Group, p. 16027. ISSN: 2041-1723. DOI: 10.1038/

ncomms16027. URL: https://www.nature.com/articles/ncomms16027 (visited on 10/27/2022).

[17] Héloïse Chassé et al. "Analysis of translation using polysome profiling". In: *Nucleic Acids Research* 45.3 (Feb. 17, 2017), e15. ISSN: 0305-1048. DOI: 10.1093/nar/gkw907. URL: https://doi.org/10.1093/nar/gkw907 (visited on 10/30/2022).

[18] Hui Chen et al. "Long-Read RNA Sequencing Identifies Alternative Splice Variants in Hepatocellular Carcinoma and Tumor-Specific Isoforms". In: *Hepatology (Baltimore, Md.)* 70.3 (Sept. 2019), pp. 1011–1025. ISSN: 1527-3350. DOI: 10.1002/hep.30500.

[19] Maria Dolores Chiara et al. "Evidence that U5 snRNP recognizes the 3 splice site for catalytic step II in mammals". In: *The EMBO Journal* 16.15 (Aug. 1997). Publisher: John Wiley & Sons, Ltd, pp. 4746–4759. ISSN: 0261-4189. DOI: 10.1093/emboj/16.15.4746. URL: https://www.embopress.org/doi/full/10.1093/emboj/16.15.4746 (visited on 09/16/2022).

[20] Michael B. Clark et al. "Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain". In: *Molecular Psychiatry* 25.1 (Jan. 2020). Number: 1 Publisher: Nature Publishing Group, pp. 37–47. ISSN: 1476-5578. DOI: 10.1038/s41380-019-0583-1. URL: https://www.nature.com/articles/s41380-019-0583-1 (visited on 11/07/2022).

[21] Tyson A. Clark, Charles W. Sugnet, and Manuel Ares. "Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays". In: *Science* 296.5569 (May 3, 2002). Publisher: American Association for the Advancement of Science, pp. 907–910. DOI: 10.1126/science.1069415. URL: https://www.science.org/doi/full/10.1126/science.1069415 (visited on 10/27/2022).

[22] Matthew Cobb. "Who discovered messenger RNA?" In: *Current Biology* 25.13 (June 29, 2015). Publisher: Elsevier, R526–R532. ISSN: 0960-9822. DOI: 10.1016/j.cub.2015.05.032. URL: https://www.cell.com/current-biology/abstract/S0960-9822(15)00606-5 (visited on 08/17/2022).

[23] Francis Crick. "Central Dogma of Molecular Biology". In: *Nature* 227.5258 (Aug. 1970). Number: 5258 Publisher: Nature Publishing Group, pp. 561–563. ISSN: 1476-4687. DOI: 10.1038/227561a0. URL: https://www.nature.com/articles/227561a0 (visited on 08/17/2022).

[24] James E. Darnell. "E.B. Wilson Lecture, 1998". In: *Molecular Biology of the Cell* 10.6 (June 1999). Publisher: American Society for Cell Biology (mboc), pp. 1685–1692. ISSN: 1059-1524. DOI: 10.1091/mbc.10.6.1685. URL: https://www.molbiolcell.org/doi/10.1091/mbc.10.6.1685 (visited on 08/19/2022).

[25] James E. Darnell. "Reflections on the history of pre-mRNA processing and highlights of current knowledge: A unified picture". In: *RNA* 19.4 (Apr. 1, 2013). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 443–460. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.038596.113. URL: http://rnajournal.cshlp.org/content/19/4/443 (visited on 09/03/2022).

[26] Laura De Conti, Marco Baralle, and Emanuele Buratti. “Exon and intron definition in pre-mRNA splicing”. In: *WIREs RNA* 4.1 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.1140, pp. 49–60. ISSN: 1757-7012. DOI: 10.1002/wrna.1140. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1140 (visited on 11/05/2022).

[27] Luca Denti et al. “ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events”. In: *BMC bioinformatics* 19.1 (2018), pp. 1–21.

[28] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (Jan. 1, 2013), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: https://doi.org/10.1093/bioinformatics/bts635 (visited on 09/18/2022).

[29] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[30] *Docker overview*. Docker Documentation. Sept. 29, 2022. URL: https://docs.docker.com/get-started/overview/ (visited on 10/01/2022).

[31] D. Dolfini and R. Mantovani. “Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y?” In: *Cell Death & Differentiation* 20.5 (May 2013). Number: 5 Publisher: Nature Publishing Group, pp. 676–685. ISSN: 1476-5403. DOI: 10.1038/cdd.2013.13. URL: https://www.nature.com/articles/cdd201313 (visited on 10/12/2022).

[32] A L DOUNCE. “[Duplicating mechanism for peptide chain and nucleic acid synthesis]”. In: *Enzymologia* 15.5 (Sept. 1, 1952), pp. 251–258. ISSN: 0013-9424.

[33] Gwendal Dujardin et al. “How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping”. In: *Molecular Cell* 54.4 (May 22, 2014), pp. 683–690. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2014.03.044. URL: https://www.sciencedirect.com/science/article/pii/S1097276514002779 (visited on 11/01/2022).

[34] Laurent Duret. “Evolution of synonymous codon usage in metazoans”. In: *Current Opinion in Genetics & Development* 12.6 (Dec. 1, 2002), pp. 640–649. ISSN: 0959-437X. DOI: 10.1016/S0959-437X(02)00353-2. URL: https://www.sciencedirect.com/science/article/pii/S0959437X02003532 (visited on 10/30/2022).

[35] Heidi Dvinge and Robert K. Bradley. “Widespread intron retention diversifies most cancer transcriptomes”. In: *Genome Medicine* 7.1 (May 15, 2015), p. 45. ISSN: 1756-994X. DOI: 10.1186/s13073-015-0168-9. URL: https://doi.org/10.1186/s13073-015-0168-9 (visited on 10/30/2022).

[36] Ettaib El Marabti, Joel Malek, and Ihab Younis. “Minor Intron Splicing from Basic Science to Disease”. In: *International Journal of Molecular Sciences* 22.11 (Jan. 2021). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 6062. ISSN: 1422-0067. DOI: 10.3390/ijms22116062. URL: https://www.mdpi.com/1422-0067/22/11/6062 (visited on 09/01/2022).

[37] Ian C. Eperon et al. “Selection of Alternative 5 Splice Sites: Role of U1 snRNP and Models for the Antagonistic Effects of SF2/ASF and hnRNP A1”. In: *Molecular and Cellular Biology* 20.22 (Nov. 15, 2000). Publisher: American Society for Microbiology, pp. 8303–8318. DOI: 10.1128/MCB.20.22.8303-8318.2000. URL: https://journals.asm.org/doi/abs/10.1128/mcb.20.22.8303-8318.2000 (visited on 09/16/2022).

[38] Philip Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (Oct. 1, 2016), pp. 3047–3048. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw354. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039924/ (visited on 11/08/2022).

[39] F. Sanger„ S. Nicklen, A. R. Coulson. *DNA sequencing with chain-terminating inhibitors | PNAS*. Dec. 1, 1977. URL: https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463 (visited on 08/22/2022).

[40] Amit Fenn et al. “Alternative splicing analysis benchmark with DICAST”. In: *bioRxiv* (2022). Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2022/01/06/2022.01.05.475067.full.pdf. DOI: 10.1101/2022.01.05.475067. URL: https://www.biorxiv.org/content/early/2022/01/06/2022.01.05.475067.

[41] Roberto Ferrarese et al. “Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression”. In: *The Journal of Clinical Investigation* 124.7 (July 1, 2014). Publisher: American Society for Clinical Investigation, pp. 2861–2876. ISSN: 0021-9738. DOI: 10.1172/JCI68836. URL: https://www.jci.org/articles/view/68836 (visited on 11/03/2022).

[42] Sebastian M. Fica et al. “RNA catalyses nuclear pre-mRNA splicing”. In: *Nature* 503.7475 (Nov. 2013). Number: 7475 Publisher: Nature Publishing Group, pp. 229–234. ISSN: 1476-4687. DOI: 10.1038/nature12734. URL: https://www.nature.com/articles/nature12734 (visited on 09/01/2022).

[43] J. T. Finch et al. “Structure of nucleosome core particles of chromatin”. In: *Nature* 269.5623 (Sept. 1, 1977), pp. 29–36. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/269029a0. URL: http://www.nature.com/articles/269029a0 (visited on 08/21/2022).

[44] Alyssa C. Frazee et al. “Polyester: simulating RNA-seq datasets with differential transcript expression”. In: *Bioinformatics* 31.17 (Sept. 1, 2015), pp. 2778–2784. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv272. URL: https://doi.org/10.1093/bioinformatics/btv272 (visited on 09/23/2022).

[45] Michael Fry. “The Surprising Discovery of Split Genes and of RNA Splicing”. In: *Landmark Experiments in Molecular Biology*. Elsevier, 2016, pp. 481–521. ISBN: 978-0-12-802074-6. DOI: 10.1016/B978-0-12-802074-6.00011-4. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128020746000114 (visited on 08/18/2022).

[46] G. Gamow. “Possible Relation between Deoxyribonucleic Acid and Protein Structures”. In: *Nature* 173.4398 (Feb. 1954). Number: 4398 Publisher: Nature Publishing Group, pp. 318–318. ISSN: 1476-4687. DOI: 10.1038/173318a0. URL: https://www.nature.com/articles/173318a0 (visited on 08/18/2022).

[47] M. Girard et al. “Entrance of newly formed messenger RNA and ribosomes into HeLa cell cytoplasm”. In: *Journal of Molecular Biology* 11.2 (Feb. 1, 1965), pp. 187–201. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(65)80050-X. URL: https://www.sciencedirect.com/science/article/pii/S002228366580050X (visited on 08/19/2022).

[48] Matteo Giulietti et al. “SpliceAid-F: a database of human splicing factors and their RNA-binding sites”. In: *Nucleic Acids Research* 41 (Database issue Jan. 2013), pp. D125–131. ISSN: 1362-4962. DOI: 10.1093/nar/gks997.

[49] Leonard D Goldstein et al. “Prediction and quantification of splice events from RNA-seq data”. In: *PloS one* 11.5 (2016), e0156132.

[50] David F. Grabski et al. “Intron retention and its impact on gene expression and protein diversity: A review and a practical guide”. In: *WIREs RNA* 12.1 (Jan. 2021). ISSN: 1757-7004, 1757-7012. DOI: 10.1002/wrna.1631. URL: https://onlinelibrary.wiley.com/doi/10.1002/wrna.1631 (visited on 09/17/2022).

[51] Gregory R. Grant et al. “Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)”. In: *Bioinformatics* 27.18 (Sept. 15, 2011), pp. 2518–2528. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr427. URL: https://doi.org/10.1093/bioinformatics/btr427 (visited on 10/18/2022).

[52] Thasso Griebel et al. “Modelling and simulating generic RNA-Seq experiments with the flux simulator”. In: *Nucleic Acids Research* 40.20 (Nov. 1, 2012), pp. 10073–10083. ISSN: 1362-4962. DOI: 10.1093/nar/gks666.

[53] Michael Hagemann-Jensen et al. “Single-cell RNA counting at allele and isoform resolution using Smart-seq3”. In: *Nature Biotechnology* 38.6 (June 2020). Number: 6 Publisher: Nature Publishing Group, pp. 708–714. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0497-0. URL: https://www.nature.com/articles/s41587-020-0497-0/ (visited on 10/27/2022).

[54] H. Harris. “Turnover of nuclear and cytoplasmic ribonucleic acid in two types of animal cell, with some further observations on the nucleolus”. In: *Biochemical Journal* 73.2 (Oct. 1, 1959), pp. 362–369. ISSN: 0006-2936. DOI: 10.1042/bj0730362. URL: https://doi.org/10.1042/bj0730362 (visited on 08/19/2022).

[55] H. Harris et al. “An examination of the ribonucleic acids in the HeLa cell with special reference to current theory about the transfer of information from nucleus to cytoplasm”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 157.967 (Feb. 26, 1963). Publisher: Royal Society, pp. 177–198. DOI: 10.1098/rspb.1963.0004. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1963.0004 (visited on 08/19/2022).

[56] Florian Heyl et al. “Galaxy CLIP-Explorer: a web server for CLIP-Seq data analysis”. In: *GigaScience* 9.11 (Nov. 10, 2020), giaa108. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa108. URL: https://doi.org/10.1093/gigascience/giaa108 (visited on 11/03/2022).

[57] Steve Hoffmann et al. “Fast mapping of short sequences with mismatches, insertions and deletions using index structures”. In: *PLoS computational biology* 5.9 (2009), e1000502.

[58] Robin Holliday. “The Induction of Mitotic Recombination by Mitomycin C in Ustilago and Saccharomyces”. In: *Genetics* 50.3 (Sept. 1964), pp. 323–335. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210654/ (visited on 08/18/2022).

[59] Illumina. *Immense discovery power for deeper insights*. URL: https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html (visited on 10/27/2022).

[60] International Human Genome Sequencing Consortium et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (Feb. 15, 2001), pp. 860–921. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35057062. URL: https://www.nature.com/articles/35057062 (visited on 09/14/2022).

[61] Aishwarya G. Jacob and Christopher W. J. Smith. “Intron retention as a component of regulated gene expression programs”. In: *Human Genetics* 136.9 (2017), pp. 1043–1057. ISSN: 0340-6717. DOI: 10.1007/s00439-017-1791-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5602073/ (visited on 10/30/2022).

[62] François Jacob and Jacques Monod. “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of Molecular Biology* 3.3 (June 1, 1961), pp. 318–356. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(61)80072-7. URL: https://www.sciencedirect.com/science/article/pii/S0022283661800727 (visited on 08/17/2022).

[63] W. Johannsen. “Elemente der exakten Erblichkeitslehre. Mit Grundzügen der biologischen Variationsstatistik”. In: *Zeitschrift für induktive Abstammungs- und Vererbungslehre* 11.1 (Dec. 1, 1914), pp. 200–200. ISSN: 1432-1874. DOI: 10.1007/BF01704312. URL: https://doi.org/10.1007/BF01704312 (visited on 08/18/2022).

[64] Mary Ellen Jones. “Albrecht Kossel, A Biographical Sketch”. In: *The Yale Journal of Biology and Medicine* 26.1 (Sept. 1953), pp. 80–97. ISSN: 0044-0086. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2599350/ (visited on 08/17/2022).

[65] James W. Jorgenson and Krynn DeArman. Lukacs. “Zone electrophoresis in open-tubular glass capillaries”. In: *Analytical Chemistry* 53.8 (July 1, 1981). Publisher: American Chemical Society, pp. 1298–1302. ISSN: 0003-2700. DOI: 10.1021/ac00231a037. URL: https://doi.org/10.1021/ac00231a037 (visited on 10/27/2022).

[66] Tamar Juven-Gershon et al. “The RNA polymerase II core promoter — the gateway to transcription”. In: *Current Opinion in Cell Biology*. Nucleus and gene expression 20.3 (June 1, 2008), pp. 253–259. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2008.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0955067408000355 (visited on 10/28/2022).

[67] André Kahles et al. “SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data”. In: *Bioinformatics* 32.12 (2016), pp. 1840–1847.

[68] Barry L. Karger and András Guttman. “DNA sequencing by CE”. In: *ELECTROPHORESIS* 30 (S1 2009). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/elps.200900218, S196–S202. ISSN: 1522-2683. DOI: 10.1002/elps.200900218. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/elps.200900218 (visited on 10/27/2022).

[69] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. “Design and analysis of ChIP-seq experiments for DNA-binding proteins”. In: *Nature Biotechnology* 26.12 (Dec. 2008). Number: 12 Publisher: Nature Publishing Group, pp. 1351–1359. ISSN: 1546-1696. DOI: 10.1038/nbt.1508. URL: https://www.nature.com/articles/nbt.1508 (visited on 11/08/2022).

[70] Heena Khatter et al. “Structure of the human 80S ribosome”. In: *Nature* 520.7549 (Apr. 2015). Number: 7549 Publisher: Nature Publishing Group, pp. 640–645. ISSN: 1476-4687. DOI: 10.1038/nature14427. URL: https://www.nature.com/articles/nature14427 (visited on 10/30/2022).

[71] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature Biotechnology* 37.8 (Aug. 2019). Number: 8 Publisher: Nature Publishing Group, pp. 907–915. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0201-4. URL: https://www.nature.com/articles/s41587-019-0201-4 (visited on 09/18/2022).

[72] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature biotechnology* 37.8 (2019), pp. 907–915.

[73] Eddo Kim, Alon Magen, and Gil Ast. “Different levels of alternative splicing among eukaryotes”. In: *Nucleic Acids Research* 35.1 (Jan. 1, 2007), pp. 125–131. ISSN: 0305-1048. DOI: 10.1093/nar/gkl924. URL: https://doi.org/10.1093/nar/gkl924 (visited on 11/05/2022).

[74] Iivari Kleino et al. “Computational solutions for spatial transcriptomics”. In: *Computational and Structural Biotechnology Journal* 20 (Jan. 1, 2022), pp. 4870–4884. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2022.08.043. URL: https://www.sciencedirect.com/science/article/pii/S2001037022003786 (visited on 11/07/2022).

[75] A. Klug. “Rosalind Franklin and the Discovery of the Structure of DNA”. In: *Nature* 219.5156 (Aug. 1968). Number: 5156 Publisher: Nature Publishing Group, pp. 808–810. ISSN: 1476-4687. DOI: 10.1038/219808a0. URL: https://www.nature.com/articles/219808a0 (visited on 08/17/2022).

[76] Kimitoshi Kohno et al. “The pleiotropic functions of the Y-box-binding protein, YB-1”. In: *BioEssays* 25.7 (2003). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.10300, pp. 691–698. ISSN: 1521-1878. DOI: 10.1002/bies.10300. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.10300 (visited on 11/03/2022).

[77] Alberto R. Kornblihtt. “CTCF: from insulators to alternative splicing regulation”. In: *Cell Research* 22.3 (Mar. 2012). Number: 3 Publisher: Nature Publishing Group, pp. 450–452. ISSN: 1748-7838. DOI: 10.1038/cr.2012.22. URL: https://www.nature.com/articles/cr201222 (visited on 11/09/2022).

[78] Eric S. Lander. “Initial impact of the sequencing of the human genome”. In: *Nature* 470.7333 (Feb. 2011). Number: 7333 Publisher: Nature Publishing Group, pp. 187–197. ISSN: 1476-4687. DOI: 10.1038/nature09792. URL: https://www.nature.com/articles/nature09792 (visited on 08/22/2022).

[79] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359.

[80] Michael Lawrence et al. “Software for Computing and Annotating Genomic Ranges”. In: *PLOS Computational Biology* 9.8 (Aug. 8, 2013). Publisher: Public Library of Science, e1003118. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003118. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118 (visited on 11/07/2022).

[81] Yeon Lee and Donald C. Rio. “Mechanisms and Regulation of Alternative Pre-mRNA Splicing”. In: *Annual review of biochemistry* 84 (2015), pp. 291–323. ISSN: 0066-4154. DOI: 10.1146/annurev-biochem-060614-034316. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4526142/ (visited on 08/19/2022).

[82] Hagai Levi, Ran Elkon, and Ron Shamir. “DOMINO: a network-based active module identification algorithm with reduced rate of false calls”. In: *Molecular Systems Biology* 17.1 (Jan. 2021). Publisher: John Wiley & Sons, Ltd, e9593. ISSN: 1744-4292. DOI: 10.15252/msb.20209593. URL: https://www.embopress.org/doi/full/10.15252/msb.20209593 (visited on 11/07/2022).

[83] Bo Li and Colin N. Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC Bioinformatics* 12.1 (Aug. 4, 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323 (visited on 11/03/2022).

[84] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (2018), pp. 3094–3100.

[85] Yang Liao, Gordon K. Smyth, and Wei Shi. “The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote”. In: *Nucleic Acids Research* 41.10 (May 1, 2013), e108. ISSN: 0305-1048. DOI: 10.1093/nar/gkt214. URL: https://doi.org/10.1093/nar/gkt214 (visited on 09/29/2022).

[86] Donny D. Licatalosi et al. “HITS-CLIP yields genome-wide insights into brain alternative RNA processing”. In: *Nature* 456.7221 (Nov. 2008). Number: 7221 Publisher: Nature Publishing Group, pp. 464–469. ISSN: 1476-4687. DOI: 10.1038/nature07488. URL: https://www.nature.com/articles/nature07488 (visited on 11/06/2022).

[87] C. H. Lin and J. G. Patton. “Regulation of alternative 3’ splice site selection by constitutive splicing factors.” In: *RNA* 1.3 (May 1, 1995). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 234–245. ISSN: 1355-8382, 1469-9001. URL: http://rnajournal.cshlp.org/content/1/3/234 (visited on 09/14/2022).

[88] Hsin-Nan Lin and Wen-Lian Hsu. “DART: a fast and accurate RNA-seq mapper with a partitioning strategy”. In: *Bioinformatics* 34.2 (2018), pp. 190–197.

[89] Chit Tong Lio et al. “Systematic analysis of alternative splicing in time course data using Spycone”. In: *bioRxiv* (2022). Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2022 DOI: 10.1101/2022.04.28.489857. URL: https://www.biorxiv.org/content/early/2022/04/29/2022.04.28.489857.

[90] Miriam Llorian et al. “The alternative splicing program of differentiated smooth muscle cells involves concerted non-productive splicing of post-transcriptional regulators”. In: *Nucleic Acids Research* 44.18 (Oct. 14, 2016), pp. 8933–8950. ISSN: 0305-1048. DOI: 10.1093/nar/gkw560. URL: https://doi.org/10.1093/nar/gkw560 (visited on 10/30/2022).

[91] David J. Lockhart et al. “Expression monitoring by hybridization to high-density oligonucleotide arrays”. In: *Nature Biotechnology* 14.13 (Dec. 1996). Number: 13 Publisher: Nature Publishing Group, pp. 1675–1680. ISSN: 1546-1696. DOI: 10.1038/nbt1296-1675. URL: https://www.nature.com/articles/nbt1296-1675 (visited on 10/27/2022).

[92] Zakaria Louadi et al. “DIGGER: exploring the functional role of alternative splicing in protein interactions”. In: *Nucleic Acids Research* 49 (D1 Jan. 8, 2021), pp. D309–D318. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa768. URL: https://doi.org/10.1093/nar/gkaa768 (visited on 11/07/2022).

[93] Zakaria Louadi et al. “Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases”. In: *Genome Biology* 22.1 (Dec. 2, 2021), p. 327. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02538-1. URL: https://doi.org/10.1186/s13059-021-02538-1 (visited on 11/07/2022).

[94] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (Dec. 5, 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8 (visited on 11/07/2022).

[95] Reini F. Luco et al. “Epigenetics in Alternative Pre-mRNA Splicing”. In: *Cell* 144.1 (Jan. 7, 2011). Publisher: Elsevier, pp. 16–26. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2010.11.056. URL: https://www.cell.com/cell/abstract/S0092-8674(10)01378-4 (visited on 11/01/2022).

[96] Reini F. Luco et al. “Regulation of Alternative Splicing by Histone Modifications”. In: *Science* 327.5968 (Feb. 19, 2010). Publisher: American Association for the Advancement of Science, pp. 996–1000. DOI: 10.1126/science.1184208. URL: https://www.science.org/doi/10.1126/science.1184208 (visited on 11/01/2022).

[97] Dmitry N. Lyabin, Irina A. Eliseeva, and Lev P. Ovchinnikov. “YB-1 protein: functions and regulation”. In: *Wiley interdisciplinary reviews. RNA* 5.1 (Feb. 2014), pp. 95–110. ISSN: 1757-7012. DOI: 10.1002/wrna.1200.

[98] Clarence K. Mah et al. “Bento: A toolkit for subcellular analysis of spatial transcriptomics data”. In: *bioRxiv* (2022). Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2022/06 DOI: 10.1101/2022.06.10.495510. URL: https://www.biorxiv.org/content/early/2022/06/13/2022.06.10.495510.

[99] Estefania Mancini et al. “ASpli: integrative analysis of splicing landscapes through RNA-Seq assays”. In: *Bioinformatics* 37.17 (2021), pp. 2609–2616.

[100] Quirin Manz et al. “ASimulatoR: splice-aware RNA-Seq data simulation”. In: *Bioinformatics* 37.18 (Sept. 15, 2021), pp. 3008–3010. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab142. URL: https://doi.org/10.1093/bioinformatics/btab142 (visited on 09/22/2022).

[101] Elaine R. Mardis. “A decade’s perspective on DNA sequencing technology”. In: *Nature* 470.7333 (Feb. 2011). Number: 7333 Publisher: Nature Publishing Group, pp. 198–203. ISSN: 1476-4687. DOI: 10.1038/nature09796. URL: https://www.nature.com/articles/nature09796 (visited on 08/22/2022).

[102] Josip Marić. “Long read RNA-seq mapper”. In: *Diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu* (2015).

[103] Ernest Martinez et al. “Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo”. In: *Molecular and cellular biology* 21.20 (2001), pp. 6782–6795.

[104] Manuel de la Mata et al. “A Slow RNA Polymerase II Affects Alternative Splicing In Vivo”. In: *Molecular Cell* 12.2 (Aug. 1, 2003), pp. 525–532. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2003.08.001. URL: https://www.sciencedirect.com/science/article/pii/S1097276503003101 (visited on 11/01/2022).

[105] A. Gregory Matera and Zefeng Wang. “A day in the life of the spliceosome”. In: *Nature Reviews Molecular Cell Biology* 15.2 (Feb. 2014). Number: 2 Publisher: Nature Publishing Group, pp. 108–121. ISSN: 1471-0080. DOI: 10.1038/nrm3742. URL: https://www.nature.com/articles/nrm3742 (visited on 08/26/2022).

[106] Matthew T. Maurano et al. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA”. In: *Science (New York, N.Y.)* 337.6099 (Sept. 7, 2012), pp. 1190–1195. ISSN: 0036-8075. DOI: 10.1126/science.1222794. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771521/ (visited on 10/31/2022).

[107] W. Richard McCombie and John D. McPherson. “Future Promises and Concerns of Ubiquitous Next-Generation Sequencing”. In: *Cold Spring Harbor Perspectives in Medicine* 9.9 (Sept. 1, 2019). Publisher: Cold Spring Harbor Laboratory Press, a025783. ISSN: , 2157-1422. DOI: 10.1101/cshperspect.a025783. URL: http://perspectivesinmedicine.cshlp.org/content/9/9/a025783 (visited on 10/31/2022).

[108] J O McInerney. “GCUA: general codon usage analysis.” In: *Bioinformatics* 14.4 (Jan. 1, 1998), pp. 372–373. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.4.372. URL: https://doi.org/10.1093/bioinformatics/14.4.372 (visited on 10/30/2022).

[109] Arfa Mehmood et al. “Systematic evaluation of differential splicing tools for RNA-seq studies”. In: *Briefings in Bioinformatics* 21.6 (Dec. 1, 2020), pp. 2052–2065. ISSN: 1477-4054. DOI: 10.1093/bib/bbz126. URL: https://doi.org/10.1093/bib/bbz126 (visited on 09/19/2022).

[110] Meinke S. *Srsf10 and the minor spliceosome control tissue-specific and dynamic SR protein expression | eLife*. Apr. 27, 2020. URL: https://elifesciences.org/articles/56075 (visited on 09/01/2022).

[111] Gregor Mendel. *Versuche über Pflanzen-Hybriden*. Accession Number: 33681636 Place: Brunn Series Number: 4. 1865.

[112] Robert Middleton et al. “IRFinder: assessing the impact of intron retention on mammalian gene expression”. In: *Genome biology* 18.1 (2017), pp. 1–11.

[113] Mohit K. Midha, Mengchu Wu, and Kuo-Ping Chiu. “Long-read sequencing in deciphering human genetics to a greater depth”. In: *Human Genetics* 138.11 (Dec. 1, 2019), pp. 1201–1215. ISSN: 1432-1203. DOI: 10.1007/s00439-019-02064-y. URL: https://doi.org/10.1007/s00439-019-02064-y (visited on 10/27/2022).

[114] Tarjei S. Mikkelsen et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells”. In: *Nature* 448.7153 (Aug. 2007). Number: 7153 Publisher: Nature Publishing Group, pp. 553–560. ISSN: 1476-4687. DOI: 10.1038/nature06008. URL: https://www.nature.com/articles/nature06008 (visited on 08/22/2022).

[115] Felix Mölder et al. *Sustainable data analysis with Snakemake [version 2; peer review: 2 approved].* 10:33. Type: article. F1000Research, Apr. 19, 2021. DOI: 10.12688/f1000research.29032.2. URL: https://f1000research.com/articles/10-33 (visited on 10/01/2022).

[116] Geoffray Monteuuis et al. “The changing paradigm of intron retention: regulation, ramifications and recipes”. In: *Nucleic Acids Research* 47.22 (Dec. 16, 2019), pp. 11497–11513. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1068. URL: https://doi.org/10.1093/nar/gkz1068 (visited on 09/17/2022).

[117] Olena Morozova and Marco A. Marra. “Applications of next-generation sequencing technologies in functional genomics”. In: *Genomics* 92.5 (Nov. 2008), pp. 255–264. ISSN: 1089-8646. DOI: 10.1016/j.ygeno.2008.07.001.

[118] Kary B. Mullis. “The Unusual Origin of the Polymerase Chain Reaction”. In: *Scientific American* 262.4 (1990). Publisher: Scientific American, a division of Nature America, Inc., pp. 56–65. ISSN: 0036-8733. URL: https://www.jstor.org/stable/24996713 (visited on 10/27/2022).

[119] Elizabeth E. Murray, Jeff Lotzer, and Mary Eberle. “Codon usage in plant genes”. In: *Nucleic Acids Research* 17.2 (Jan. 25, 1989), pp. 477–498. ISSN: 0305-1048. DOI: 10.1093/nar/17.2.477. URL: https://doi.org/10.1093/nar/17.2.477 (visited on 10/30/2022).

[120] Suriya Narayanan Murugesan et al. “Expression and network analysis of YBX1 interactors for identification of new drug targets in lung adenocarcinoma”. In: *Journal of Genomics* 6 (June 26, 2018), pp. 103–112. ISSN: 1839-9940. DOI: 10.7150/jgen.20581. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030768/ (visited on 11/07/2022).

[121] Shiran Naftelberg et al. “Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure”. In: *Annual Review of Biochemistry* 84.1 (June 2, 2015), pp. 165–198. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev-biochem-060614-034242. URL: https://www.annualreviews.org/doi/10.1146/annurev-biochem-060614-034242 (visited on 11/01/2022).

[122] Satu Nahkuri, Ryan J. Taft, and John S. Mattick. “Nucleosomes are preferentially positioned at exons in somatic and sperm cells”. In: *Cell Cycle* 8.20 (Oct. 15, 2009). Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cc.8.20.9916, pp. 3420–3424. ISSN: 1538-4101. DOI: 10.4161/cc.8.20.9916. URL: https://doi.org/10.4161/cc.8.20.9916 (visited on 11/01/2022).

[123] Hadas Ner-Gaon et al. “Intron retention is a major phenomenon in alternative splicing in Arabidopsis”. In: *The Plant Journal* 39.6 (2004). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2004.02172.x, pp. 877–885. ISSN: 1365-313X. DOI: 10.1111/j.1365-313X.2004.02172.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2004.02172.x (visited on 11/05/2022).

[124] D. B. Nikolov and S. K. Burley. “RNA polymerase II transcription initiation: A structuralview”. In: *Proceedings of the National Academy of Sciences* 94.1 (Jan. 7, 1997). Publisher: Proceedings of the National Academy of Sciences, pp. 15–22. DOI: 10.1073/pnas.94.1.15. URL: https://www.pnas.org/doi/full/10.1073/pnas.94.1.15 (visited on 10/28/2022).

[125] Daniel Nüst et al. “The Rockerverse: Packages and Applications for Containerisation with R”. In: *The R Journal* 12.1 (2020), pp. 437–461. ISSN: 2073-4859. URL: https://journal.r-project.org/archive/2020/RJ-2020-007/index.html (visited on 11/07/2022).

[126] Valerio Orlando. “Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation”. In: *Trends in Biochemical Sciences* 25.3 (Mar. 2000), pp. 99–104. ISSN: 09680004. DOI: 10.1016/S0968-0004(99)01535-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968000499015352 (visited on 10/30/2022).

[127] Qun Pan et al. “Revealing Global Regulatory Features of Mammalian Alternative Splicing Using a Quantitative Microarray Platform”. In: *Molecular Cell* 16.6 (Dec. 22, 2004), pp. 929–941. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2004.12.004. URL: https://www.sciencedirect.com/science/article/pii/S1097276504007610 (visited on 10/31/2022).

[128] Eddie Park et al. “The Expanding Landscape of Alternative Splicing Variation in Human Populations”. In: *The American Journal of Human Genetics* 102.1 (Jan. 4, 2018). Publisher: Elsevier, pp. 11–26. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2017.11.002. URL: https://www.cell.com/ajhg/abstract/S0002-9297(17)30454-8 (visited on 11/07/2022).

[129] Nicolas Philippe et al. “CRAC: an integrated approach to the analysis of RNA-seq reads”. In: *Genome biology* 14.3 (2013), pp. 1–16.

[130] Harold Pimentel et al. “A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis”. In: *Nucleic Acids Research* 44.2 (Jan. 29, 2016), pp. 838–851. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1168. URL: https://doi.org/10.1093/nar/gkv1168 (visited on 10/30/2022).

[131] C. Plaschka et al. “Transcription initiation complex structures elucidate DNA opening”. In: *Nature* 533.7603 (May 2016). Number: 7603 Publisher: Nature Publishing Group, pp. 353–358. ISSN: 1476-4687. DOI: 10.1038/nature17990. URL: https://www.nature.com/articles/nature17990 (visited on 10/28/2022).

[132] Petter Portin and Adam Wilkins. “The Evolving Definition of the Term “Gene””. In: *Genetics* 205.4 (Apr. 1, 2017), pp. 1353–1364. ISSN: 1943-2631. DOI: 10.1534/genetics.116.196956. URL: https://doi.org/10.1534/genetics.116.196956 (visited on 08/18/2022).

[133] Alejandro Reyes et al. “Drift and conservation of differential exon usage across tissues in primate species”. In: *Proceedings of the National Academy of Sciences* 110.38 (Sept. 17, 2013). Publisher: Proceedings of the National Academy of Sciences, pp. 15377–15382. DOI: 10.1073/pnas.1307202110. URL: https://www.pnas.org/doi/full/10.1073/pnas.1307202110 (visited on 10/02/2022).

[134] Anthony Rhoads and Kin Fai Au. “PacBio Sequencing and Its Applications”. In: *Genomics, Proteomics & Bioinformatics*. SI: Metagenomics of Marine Environments 13.5 (Oct. 1, 2015), pp. 278–289. ISSN: 1672-0229. DOI: 10.1016/j.gpb.2015.08.002. URL: https://www.sciencedirect.com/science/article/pii/S1672022915001345 (visited on 11/01/2022).

[135] Aaron R. Robart et al. “Crystal structure of a eukaryotic group II intron lariat”. In: *Nature* 514.7521 (Oct. 2014). Number: 7521 Publisher: Nature Publishing Group, pp. 193–197. ISSN: 1476-4687. DOI: 10.1038/nature13790. URL: https://www.nature.com/articles/nature13790 (visited on 09/22/2022).

[136] Juan P Romero et al. “EventPointer: an effective identification of alternative splicing events using junction arrays”. In: *BMC genomics* 17.1 (2016), pp. 1–18.

[137] Oliver Rossbach et al. “Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L”. In: *RNA Biology* 11.2 (Feb. 1, 2014). Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/rna.27991, pp. 146–155. ISSN: 1547-6286. DOI: 10.4161/rna.27991. URL: https://doi.org/10.4161/rna.27991 (visited on 11/06/2022).

[138] Brian C. Rymond and Michael Rosbash. “Cleavage of 5 splice site and lariat formation are independent of 3 splice site in yeast mRNA splicing”. In: *Nature* 317.6039 (Oct. 1985). Number: 6039 Publisher: Nature Publishing Group, pp. 735–737. ISSN: 1476-4687. DOI: 10.1038/317735a0. URL: https://www.nature.com/articles/317735a0 (visited on 10/30/2022).

[139] M. Salditt-Georgieff et al. “Methyl labeling of HeLa cell hnRNA: a comparison with mRNA”. In: *Cell* 7.2 (Feb. 1, 1976), pp. 227–237. ISSN: 0092-8674. DOI: 10.1016/0092-8674(76)90022-2. URL: https://www.sciencedirect.com/science/article/pii/0092867476900222 (visited on 08/19/2022).

[140] Abbie Saunders, Leighton J. Core, and John T. Lis. “Breaking barriers to transcription elongation”. In: *Nature Reviews Molecular Cell Biology* 7.8 (Aug. 2006). Number: 8 Publisher: Nature Publishing Group, pp. 557–567. ISSN: 1471-0080. DOI: 10.1038/nrm1981. URL: https://www.nature.com/articles/nrm1981 (visited on 10/28/2022).

[141] Mark Schena et al. “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray”. In: *Science* 270.5235 (Oct. 20, 1995). Publisher: American Association for the Advancement of Science, pp. 467–470. DOI: 10.1126/science.270.5235.467. URL: https://www.science.org/doi/10.1126/science.270.5235.467 (visited on 10/31/2022).

[142] Schraga Schwartz, Eran Meshorer, and Gil Ast. “Chromatin organization marks exon-intron structure”. In: *Nature Structural & Molecular Biology* 16.9 (Sept. 2009). Number: 9 Publisher: Nature Publishing Group, pp. 990–995. ISSN: 1545-9985. DOI: 10.1038/nsmb.1659. URL: https://www.nature.com/articles/nsmb.1659 (visited on 11/01/2022).

[143] Ankeeta Shah et al. “Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation”. In: *Genome Biology* 22.1 (Oct. 14, 2021), p. 291. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02502-z. URL: https://doi.org/10.1186/s13059-021-02502-z (visited on 09/19/2022).

[144] Shalini Sharma et al. “U1 snRNA Directly Interacts with Polypyrimidine Tract-Binding Protein during Splicing Repression”. In: *Molecular Cell* 41.5 (Mar. 4, 2011), pp. 579–588. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2011.02.012. URL: https://www.sciencedirect.com/science/article/pii/S1097276511000955 (visited on 10/30/2022).

[145] Matthew J. Shurtleff et al. “Broad role for YBX1 in defining the small noncoding RNA composition of exosomes”. In: *Proceedings of the National Academy of Sciences* 114.43 (Oct. 24, 2017). Publisher: Proceedings of the National Academy of Sciences, E8987–E8995. DOI: 10.1073/pnas.1712108114. URL: https://www.pnas.org/doi/10.1073/pnas.1712108114 (visited on 11/03/2022).

[146] Robert J. Sims et al. “Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing”. In: *Molecular Cell* 28.4 (Nov. 30, 2007), pp. 665–676. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2007.11.010.

[147] Barry Slaff et al. “MOCCASIN: a method for correcting for known and unknown confounders in RNA splicing analysis”. In: *Nature Communications* 12.1 (June 7, 2021). Number: 1 Publisher: Nature Publishing Group, p. 3353. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23608-9. URL: https://www.nature.com/articles/s41467-021-23608-9 (visited on 11/06/2022).

[148] Stephen T. Smale and James T. Kadonaga. “The RNA Polymerase II Core Promoter”. In: *Annual Review of Biochemistry* 72.1 (June 2003), pp. 449–479. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.biochem.72.121801.161520. URL: https://www.annualreviews.org/doi/10.1146/annurev.biochem.72.121801.161520 (visited on 10/28/2022).

[149] Christopher W. J. Smith et al. “Scanning from an independently specified branch point defines the 3 splice site of mammalian introns”. In: *Nature* 342.6247 (Nov. 1989). Number: 6247 Publisher: Nature Publishing Group, pp. 243–247. ISSN: 1476-4687. DOI: 10.1038/342243a0. URL: https://www.nature.com/articles/342243a0 (visited on 09/16/2022).

[150] Timothy Sterne-Weiler et al. “Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop”. In: *Molecular cell* 72.1 (2018), pp. 187–200.

[151] Deborah A. Sterner, Troy Carlo, and Susan M. Berget. “Architectural limits on splitgenes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.26 (Dec. 24, 1996), pp. 15081–15085. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26359/ (visited on 09/16/2022).

[152] Reiko Takemura et al. “Multiple factors in the early splicing complex are involved in the nuclear retention of pre-mRNAs in mammalian cells”. In: *Genes to Cells* 16.10 (2011). _eprint: https://onlinelibrary.wiley.co 2443.2011.01548.x, pp. 1035–1049. ISSN: 1365-2443. DOI: 10.1111/j.1365-2443.2011.01548.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2443.2011.01548.x (visited on 10/30/2022).

[153] Ole Tange. *GNU Parallel 2018*. Ole Tange, Apr. 27, 2018. ISBN: 978-1-387-50988-1. DOI: 10.5281/zenodo.1146014. URL: https://zenodo.org/record/1146014 (visited on 10/02/2022).

[154] *The Nobel Prize in Physiology or Medicine 1993*. NobelPrize.org. 1993. URL: https://www.nobelprize.org/prizes/medicine/1993/press-release/ (visited on 08/19/2022).

[155] Thermofischer. *What is Sanger sequencing? - DE*. URL: https://www.thermofisher.com/de/de/home/life-science/sequencing/sequencing-learning-center/capillary-electrophoresis-information/what-is-sanger-sequencing.html (visited on 08/22/2022).

[156] Michael Uhl et al. “Computational analysis of CLIP-seq data”. In: *Methods* 118-119 (Apr. 2017), pp. 60–72. ISSN: 10462023. DOI: 10.1016/j.ymeth.2017.02.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S1046202317300828 (visited on 11/03/2022).

[157] Eric L. Van Nostrand et al. “Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)”. In: *Nature Methods* 13.6 (June 2016). Number: 6 Publisher: Nature Publishing Group, pp. 508–514. ISSN: 1548-7105. DOI: 10.1038/nmeth.3810. URL: https://www.nature.com/articles/nmeth.3810 (visited on 11/06/2022).

[158] Jorge Vaquero-Garcia et al. “A new view of transcriptome complexity and regulation through the lens of local splicing variations”. In: *elife* 5 (2016), e11752.

[159] Jorge Vaquero-Garcia et al. “A new view of transcriptome complexity and regulation through the lens of local splicing variations”. In: *eLife* 5 (Feb. 1, 2016), e11752. ISSN: 2050-084X. DOI: 10.7554/eLife.11752.

[160] Artur Veloso et al. “Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications”. In: *Genome Research* 24.6 (June 2014), pp. 896–905. ISSN: 1549-5469. DOI: 10.1101/gr.171405.113.

[161] Kristoffer Vitting-Seerup and Albin Sandelin. “IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences”. In: *Bioinformatics* 35.21 (Nov. 1, 2019), pp. 4469–4471. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz247. URL: https://doi.org/10.1093/bioinformatics/btz247 (visited on 11/07/2022).

[162] Henry Völzke et al. “Cohort Profile: The Study of Health in Pomerania”. In: *International Journal of Epidemiology* 40.2 (Apr. 1, 2011), pp. 294–307. ISSN: 0300-5771. DOI: 10.1093/ije/dyp394. URL: https://doi.org/10.1093/ije/dyp394 (visited on 09/26/2022).

[163] C. H. Waddington. “Canalization of Development and the Inheritance of Acquired Characters”. In: *Nature* 150.3811 (Nov. 1942). Number: 3811 Publisher: Nature Publishing Group, pp. 563–565. ISSN: 1476-4687. DOI: 10.1038/150563a0. URL: https://www.nature.com/articles/150563a0 (visited on 08/18/2022).

[164] Kai Wang et al. “MapSplice: accurate mapping of RNA-seq reads for splice junction discovery”. In: *Nucleic Acids Research* 38.18 (Oct. 2010), e178. ISSN: 1362-4962. DOI: 10.1093/nar/gkq622.

[165] Kai Wang et al. “MapSplice: accurate mapping of RNA-seq reads for splice junction discovery”. In: *Nucleic acids research* 38.18 (2010), e178–e178.

[166] Su Wang et al. “Target analysis by integration of transcriptome and ChIP-seq data with BETA”. In: *Nature protocols* 8.12 (Dec. 2013), pp. 2502–2515. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.150. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4135175/ (visited on 11/03/2022).

[167] Nisar Wani and Khalid Raza. “Raw Sequence to Target Gene Prediction: An Integrated Inference Pipeline for ChIP-Seq and RNA-Seq Datasets”. In: *Applications of Artificial Intelligence Techniques in Engineering*. Ed. by Hasmat Malik et al. Advances in Intelligent Systems and Computing. Singapore: Springer, 2019, pp. 557–568. ISBN: 9789811318221. DOI: 10.1007/978-981-13-1822-1_52.

[168] Jonathan R. Warner et al. “Rapidly labeled HeLa cell nuclear RNA: I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA”. In: *Journal of Molecular Biology* 19.2 (Aug. 1, 1966), pp. 349–361. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(66)80009-8. URL: https://www.sciencedirect.com/science/article/pii/S0022283666800098 (visited on 08/19/2022).

[169] Charlotte J. Wright, Christopher W. J. Smith, and Chris D. Jiggins. “Alternative splicing as a source of phenotypic diversity”. In: *Nature Reviews Genetics* (July 12, 2022). Publisher: Nature Publishing Group, pp. 1–14. ISSN: 1471-0064. DOI: 10.1038/s41576-022-00514-4. URL: https://www.nature.com/articles/s41576-022-00514-4 (visited on 09/14/2022).

[170] Thomas D Wu and Serban Nacu. “Fast and SNP-tolerant detection of complex variants and splicing in short reads”. In: *Bioinformatics* 26.7 (2010), pp. 873–881.

[171] Yungang Xu et al. “Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision”. In: *Nucleic Acids Research* 45.21 (Dec. 1, 2017), pp. 12100–12112. ISSN: 0305-1048. DOI: 10.1093/nar/gkx870. URL: https://doi.org/10.1093/nar/gkx870 (visited on 11/05/2022).

[172] M. Q. Zhang. “Statistical Features of Human Exons and Their Flanking Regions”. In: *Human Molecular Genetics* 7.5 (May 1, 1998), pp. 919–932. ISSN: 0964-6906. DOI: 10.1093/hmg/7.5.919. URL: https://doi.org/10.1093/hmg/7.5.919 (visited on 09/14/2022).

[173] Yong Zhang et al. “Model-based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9 (Sept. 17, 2008), R137. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-9-r137. URL: https://doi.org/10.1186/gb-2008-9-9-r137 (visited on 08/26/2022).

[174] Wei Zhao et al. “Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling”. In: *BMC Genomics* 15.1 (June 2, 2014), p. 419. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-419. URL: https://doi.org/10.1186/1471-2164-15-419 (visited on 10/31/2022).

Appendix

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Aug 26, 2022

This Agreement between Amit Fenn ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5376571255315
License date	Aug 26, 2022
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	A day in the life of the spliceosome
Licensed Content Author	A. Gregory Matera et al
Licensed Content Date	Jan 23, 2014
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no

Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	PhD student
Institution name	TUM
Expected presentation date	Jan 2023
Portions	Figure 4.a
Requestor Location	Amit Fenn Pelkovenstraße 98A, Munich Munich, 80992 Germany Attn: Technical University of Munich
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. Grant of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

1. 2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Sep 21, 2022

This Agreement between Amit Fenn ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	5393760310463
License date	Sep 21, 2022
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	ASimulatoR: splice-aware RNA-Seq data simulation
Licensed content author	Manz, Quirin; Tsoy, Olga
Licensed content date	Feb 27, 2021
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	PhD student
Publisher of your work	TUM
Expected publication date	Jan 2023
Permissions cost	0.00 USD
Value added tax	0.00 USD

Total	0.00 USD
Title	PhD student
Institution name	TUM
Expected presentation date	Jan 2023
Order reference number	1
Portions	Figure 1
Requestor Location	Amit Fenn Pelkovenstraße 98A, Munich Munich, 80992 Germany Attn: Technical University of Munich
Publisher Tax ID	GB125506730
Total	0.00 USD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.