Novel Methods for the Computational Analysis of RNA-Seq Data with Applications to Alternative Splicing

---

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl. Bioinf. André Kahles
aus Zschopau

Tübingen
2014

Tag der mündlichen Qualifikation:	26.09.2014
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Gunnar Rätsch
2. Berichterstatter:	Prof. Dr. Daniel H. Huson

Abstract

Understanding how genetic information is transformed into a diverse spectrum of complex organisms is one of the longstanding questions of biology. Over the recent years, advancements in sequencing technology have enabled the accurate measurement of the pool of ribonucleic acids (RNAs) contained in a cell at an unprecedented depth. High-throughput RNA-sequencing (RNA-Seq) allows to acquire quantitative measurements of all transcripts in one or more cells and provides qualitative information about isoform structures or sequence alterations. Our goal is to use this information to get a better understanding of RNA-processing and gene regulation with a specific focus on alternative splicing. In this thesis, we present advanced computational methods for the processing of RNA-Seq data, including novel strategies for spliced alignment in the context of genomic variation, accuracy improvements through alignment post-processing and the first high-throughput analysis pipeline for the characterization of alternative splicing events.

Our first contribution is the development and extension of PALMapper, a versatile RNA-Seq alignment method. By using a variation-aware alignment approach, we could markedly improve its alignment sensitivity in cases where reference genome and the source-genome of the measured RNA differ. We also greatly increased its accuracy through an additional re-alignment step for reads that span splice junctions. Due to the high-throughput nature of the data and limited computational resources, most alignment tools only perform an approximate search. To better understand the extent of variability in the alignments results and to identify possible sources of variation, we performed a comprehensive evaluation of alignment algorithms, showing substantial differences between alignment outcomes. Using the insights gained during the evaluation, we developed two powerful alignment post-processing tools that aim at making results more comparable and remove possible false hits from the data: The simple alignment filtering tool (SAFT) optimizes filter criteria on a given training set to increase overall accuracy of the alignment. The tool for multiple-mapper resolution (MMR) disambiguates between several equally good alignment-possibilities of the same read, using an iterative algorithm to minimize the variance of the local read coverage. In order to use RNA-Seq alignments for profiling alternative splicing (AS), we developed SplAdder, a tool that enriches a splicing graph representation of existing genome annotations and extracts AS-events from this augmented graph.

All presented methods were applied in analysis pipelines that align, post-process and then quantitatively analyze RNA-Seq data. We present four biological studies, where the herein presented tools were an integral part of the analysis pipeline. In a study on the mRNA degradation mechanism nonsense-mediated decay (NMD) in *Arabidopsis thaliana*, we analyzed samples mutated in *UPF1* and *UPF3* and thus deficient in NMD to investigate the connection between alternative splicing and transcript degradation and to estimate its pervasiveness. We found that $\approx$ 17% of all protein-coding multiple-exon genes produce isoforms that are subject to NMD and that over 90% of these isoforms share characteristic transcript-features characteristic. In a second study, we investigated the role of polypyrimidine-tract binding proteins (PTB) for alternative splicing regulation in *A. thaliana*. Based on a complementary set of mutant samples with elevated or decreased PTB-expression, we identified 452 events responsive to PTB perturbation with interesting functional implications for flowering and germination. In a third, larger scale study, we focused on the identification of splicing quantitative trait loci (sQTL) and analyzed over 700 RNA-Seq libraries generated from two populations of *A. thaliana*. We identified numerous significant associations proximal and distal to the event site, forming *cis*- and *trans*-sQTL, respectively, and found marked differences between the two populations. In the last study, we set out to identify sQTL in twelve different cancer types in one of the largest available transcriptome datasets. We re-aligned RNA-Seq samples of over 4,000 patients provided through The Cancer Genome Atlas (TCGA). We identified and quantified thousands of novel AS events and could show that many splicing alterations appear to be cancer-type specific. We further used genetic information from whole exome sequencing, to identify numerous *cis*- and *trans*-sQTL, both confirming earlier findings and detecting promising novel associations.

In conclusion, we show that the presented methods are efficient and effectively applicable within a wide range of scenarios. Our work resulted in numerous findings, that could be confirmed through earlier studies or validation experiments but also uncovered exciting new findings for splicing regulation in plants as well as aberrant splicing in cancer. We are confident that our contributions are an excellent basis to spark further improvements and novel methods.

Zusammenfassung

Seit langem ist es eine der zentralen Fragen der biologischen Forschung, wie aus genetischer Information die große Diversität komplexer Organismen entstehen kann. Seit wenigen Jahren haben es verbesserte Sequenziertechnologien ermöglicht, die grosse Menge von Ribonukleinsäuren (RNA) einer Zelle mit bisher ungekannter Genauigkeit zu messen. Das Hochdurchsatz-Verfahren der RNA-Sequenzierung (RNA-Seq) erlaubt es, alle Transkripte einer oder mehrerer Zellen gleichzeitig quantitativ zu erfassen und ermöglicht das Sammeln qualitativer Informationen zu Transkript-Struktur oder Sequenzveränderungen. Unser Ziel ist es, diese Information einzusetzen, um RNA-Prozessierung und Genregulation und vor allem den Prozeß des alternativen Spleißens (AS) besser zu verstehen. Im Rahmen dieser Arbeit präsentieren wir neuartige Verarbeitungsverfahren für RNA-Seq Daten, einschliesslich neuer Strategien zum Alignment gespleißter Sequenzen im Kontext genomischer Variation, Verbesserungen der Alignmentgenauigkeit durch optimale Nachbearbeitung sowie das erste Hochdurchsatz-System zur Charakterisierung von AS-Ereignissen.

Unser erster Beitrag ist die Entwicklung und Erweiterung von PALMapper, einer Methode zum Alignieren von RNA-Seq Daten. Durch die Berücksichtigung genomischer Sequenzvarianten konnten wir eine deutliche Verbesserung der Alignment-Sensitivität auch in solchen Fällen erreichen, in denen Referenzgenom und Quellgenom der RNA-Seq Daten sich unterscheiden. Außerdem konnten wir die Genauigkeit von Alignments über Intron-Grenzen durch ein zusätzliches Rück-Alignment deutlich verbessern. Aufgrund des Hochdurchsatz-Charakters der Daten sowie begrenzter Rechenressourcen beschränken sich die meisten Alignmentprogramme auf eine approximative Suche. Um das Ausmaß der Ergebnisvariabilität besser zu vestehen, haben wir eine umfassende Evaluation verschiedener Programme durchgeführt und ausgesprochen deutliche Unterschiede aufgezeigt. Mithilfe dieser Erkenntnisse, haben wir zwei Programme zur wirkungsvollen Alignment-Nachbearbeitung entwickelt, die die Rate Falsch-Positiver minimieren und die Vergleichbarkeit zwischen den Ergebnissen erhöhen sollen: Das erste Programm, SAFT, berechnet anhand gegebener Trainingsdaten eine optimale Kombination von Filterparametern und erhöht dadurch die Alignment-Genauigkeit. Das zweite Programm, MMR, wählt aus mehreren gleich guten Alignments einer Sequenz das best-passende aus. Dies geschieht mittels eines iterativen Verfahrens bei dem die Varianz der lokalen Alignmentabdeckung minimiert wird. Um aus RNA-Seq Alignments ein Profil alternativen Spleißens (AS) zu erstellen, haben wir SplAdder entwickelt, ein Programm welches einen auf der Genomannotation basierenden Spleißgraphen erweitert und daraus extrahierte AS-Ereignisse quantifiziert.

Alle vorgestellten Methoden wurden im Rahmen mehrstufiger Analyseverfahren eingesetzt, die sowohl das Alignment und dessen Nachbearbeitung als auch die quantitative Datenanalyse umfassen. Wir beschreiben vier biologische Studien, in welchen die entwickelten Programme integraler Bestandteil der Analyse waren. In einer Studie zum mRNA-Abbauweg NMD in *A. thaliana* haben wir NMD-blockierte Pflanzen, mutiert in den Genen *UPF1* und *UPF3*, untersucht, um die Verbindung zwischen Transkriptabbau und AS sowie die Verbreitung von NMD zu erforschen. Wir konnten zeigen, dass $\approx 17\%$ aller Protein-codierenden multi-exonischen Gene mindestens eine Isoform produzieren die von NMD abgebaut wird und dass 90% dieser Isoformen charakteristische Merkmale aufweisen. In einer zweiten Studie untersuchten wir die Rolle von Polypyrimidintraktbindeproteinen (PTB) für die Regulation von AS in *A. thaliana*. Anhand von Mutanten mit erhöhter bzw. verringerter PTB-Produktion konnten wir 452 AS-Ereignisse identifizieren, die sich nach PTB-Perturbation signifikant veränderten und interessante funktionelle Auswirkungen auf das Blüh- und Keimverhalten zeigten. In einer dritten, deutlich umfangreicheren Arbeit lag unser Schwerpunkt auf der Identifikation genetischer Loci deren Spleißen sich quantitativ in Abhängigkeit genetischer Varianten verändert (sQTL). Hierzu analysierten wir 700 RNA-Seq Datensätze aus zwei *A. thaliana*-Populationen und konnten zahlreiche signifikant assoziierte Sequenzvarianten proximal (*cis*-sQTL) und distal (*trans*-sQTL) zum jeweiligen Spleiß-Ereignis identifizieren – mit deutlichen Unterschieden zwischen den Populationen. In einer weiteren Studie nutzten wir einen der größten verfügbaren RNA-Seq Datensätze, um sQTL in zwölf verschiedenen Krebsarten zu finden. Hierzu haben wir Daten von mehr als 4000 Patienten des Krebs Genom-Atlas Projekts (TCGA) analysiert. Dadurch konnten wir tausende neue AS-Ereignisse detektieren und quantifizieren und fanden Hinweise darauf, dass zahlreiche Ereignisse Krebs-spezifisch sind. Weiterhin nutzten wir genetische Information aus TCGA um zahlreiche *cis*- und *trans*-sQTL zu identifizieren, die teilweise durch Studien belegt werden konnten aber auch vielversprechende Neuentdeckungen enthalten.

Wir zeigen die effektive und effiziente Anwendbarkeit der vorgestellten Methoden in einer Vielzahl unterschiedlicher Szenarien. Unsere Arbeit ergab zahlreiche neue Erkenntnisse, die teils durch frühere Studien belegt oder durch Experimente validiert werden konnten, die aber auch spannende Neuentdeckungen zur Spleißregulation in Pflanzen oder fehlerhaftem Spleißen bei Krebs beinhalteten. Wir sind zuversichtlich, dass unsere Beiträge eine sehr gute Basis für die Verbesserung und Entwicklung neuer Methoden bieten.

Acknowledgements

I would like to express my deepest gratitude to my advisor Gunnar Rätsch who has not only been an excellent scientific mentor and teacher but also always cared about personal needs, making the stay in his lab a most pleasant experience.

I am further very grateful to my other advisors Andreas Wachter and Daniel Huson who shared their experience and knowledge and provided me with guidance and critical feedback throughout my time as a PhD student.

During my research in Tübingen and New York, I have met many exceptional people who shared with me their scientific insights, sometimes crazy ideas and, most importantly, a lot of fun time. I would like to thank Jonas Behr for motivating a good work-life balance, Regina Bohnert for showing the red thread through the PhD-jungle, Philipp Drewe for always asking the questions nobody else asked, Kadeem Ho Sang for standing firm as a rock in administrative whitewater, Géraldine Jean for fighting on my side in the RGASP challenge, Theofanis Karaletsos for re-calibrating my view on life, David Kuo for an introduction into American life, Kjong Lehmann for helping to pull a data-elephant through the eye of a needle, Sebastian Schultheiss for his help on the last meters, Vipin Sreedharan for exploring the Galaxy with me, Richard Stein for uncountably many cereal bars, and all the others I had the pleasure to work with: Fabio De Bona, Nico Görnitz, Darya Karelina, Marius Kloft, Xinghua Lou, André Noll, Katherine Redfield Chan, Kana Shimizu, Cheng Soon Ong, Julia Vogt, Christian Widmer, Yun Yan, Georg Zeller and Yi Zhong. Into this list, I would also like to include the students of my year in the Tübingen International PhD Program and all those with whom I worked as PhD representative or in the student council.

None of the projects could have been realized without the data or the biological insights. Therefore, I am very grateful to my collaborators who shared their knowledge, ideas, criticism and data with me. I would like to thank Andreas Wachter, Christina Rühl, Gabriele Drechsel, Lisa Smith, Amelie Baud, Edward J Osborne, Oliver Stegle, Richard Clark and Magnus Nordborg.

Further, I want to thank Philipp Drewe, Kjong Lehmann, Gunnar Rätsch, Anna-Lena Schinke and Richard Stein for critically reading the manuscript of my thesis and providing valuable feedback.

At last, I would like to thank Anna-Lena, for sharing with me that wonderful, exciting and sometimes exhausting experience on two continents – it has been an amazing journey and would have never been possible without you, as well as my parents for their constant support of me and my aims.

Funding was provided by the Max Planck Society, the German Research Foundation (DFG) under RA1894/2-1 and the Memorial Sloan Kettering Cancer Center.

André

Contents

1 Introduction

In this work, we cover a broad range of topics requiring background in both biological and computational fields. This introduction will give a brief overview of the topics most relevant to understand the following chapters. At first, we give an introduction into RNA-biology, including the central dogma of molecular biology, alternative splicing and transcriptional regulation. We also provide some context for common applications. As the studies discussed in Chapter 3 cover samples from both plants and human, we chose agriculture and medicine. Following this, we will discuss sequencing technologies and give a short history of transcriptome analysis to motivate the currently used measurement techniques for alternative splicing, the mechanism that is relevant for most parts of this work. In the subsequent discussion of high-throughput techniques we explain the necessity for large scale analysis and provide an overview of the most common steps in transcriptome analysis pipelines. Lastly, we will review the major strategies for the alignment of high-throughput sequencing reads, the first step in many analysis pipelines and a topic most relevant for the methods discussed in Chapter 2.

1.1 RNA-processing and Alternative Splicing

This section shall provide a general introduction to RNA-processing and regulation with a focus on alternative splicing, discuss its relevance in research for agriculture and medicine and introduce common measurement techniques used for transcriptome analysis.

1.1.1 Biological Background

Different organisms show a broad range of diversity in terms of morphology, behavior, lifespan and many more biological traits, also denoted as *phenotypes*. Most of these differences root in the genetic information carried by each of the cells, the *genotype*. In this biological introduction we will provide an overview on the relations between genotype and phenotype, explain gene expression and translation and put the mechanism of alternative splicing into its biological context. We will also discuss different mechanisms of regulation relevant for these processes.

RNA Processing in the Context of Organism Complexity Following the central dogma of molecular biology suggested by Francis Crick in 1958 [56], the complexity of an organism arises from its genome. All information is encoded in large molecules of *deoxyribonucleic acid* (DNA), that are partially transcribed into molecules of *ribonucleic acid* (RNA) and are finally translated into chains of amino acids that fold and assemble into functional units forming proteins [8] (Figure 1.1). This process can already stop on the level of RNAs that also can act as functional entities, e.g., as part of the ribosomes or as ribozymes.

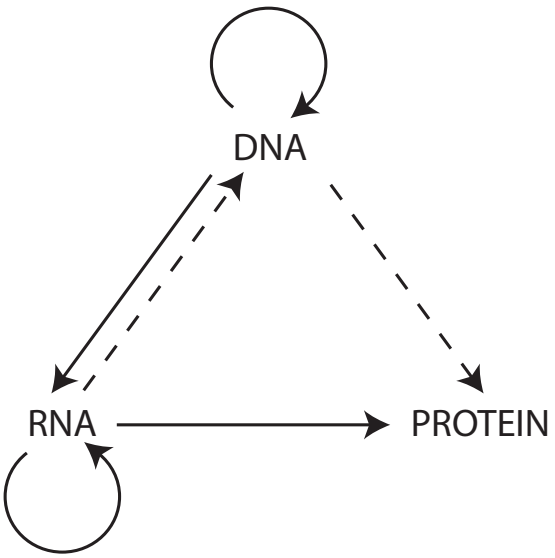


Figure 1.1: Adaptation of the central dogma of molecular biology as suggested in [56] and shown in [57]. Solid arrows show information flow that is probable. Dashed arrows show information flow that is theoretical possible. All other possible arrows are excluded by the central dogma.

A more biological description of the central dogma is shown in Figure 1.2. However, the general principle and the direction of information-flow remain the same. Briefly, the DNA acts as a matrix for RNA-polymerases to produce RNA-molecules complementary to the DNA-matrix. Different polymerases are specialized for different purposes. The *messenger-RNA* (mRNA) as shown in Figure 1.2 is synthesized by RNA-polymerase II. In a maturation process this *precursor-mRNA* (pre-mRNA) then undergoes splicing, a process we will describe in more detail later in this section. Then, to prevent degradation, a 5'-cap structure and a tail of adenine bases (polyA-tail) are added to the RNA molecule, forming the mature mRNA that is ready for export from the nucleus. After export to the cytosol, *ribosomes* attach to the mRNA and initiate translation. In this process the ribosome recruits *transfer-RNAs* (tRNAs) that specifically shuttle amino acids. Each tRNA carries a triplet-code of three nucleotides, denoted *codon*, which is used for specific binding to the mRNA, thus translating the sequence of base-triplets of the mRNA into a chain of amino acids. The relationship between codons and amino acids is also known as the *genetic code* [136]. Not all parts of the transcript are translated into protein, only the *coding sequence* (CDS), the regions before the translation start and after the translation stop remain untranslated and are denoted as *5' untranslated region* (5'-UTR) and *3' untranslated region* (3'-UTR), respectively. Finally, the chain of amino acids folds into a three-dimensional structure and can be functionally modified through the addition of further molecules, e.g., in phosphorylation or glycosylation.

Each of these processes is tightly regulated, thereby integrating information about cell state, cellular context and environmental signals. The principle of the central dogma and the general steps in the processing of DNA into functionally active entities (not only containing proteins but also functional RNA elements) are conserved across the three domains of life. However, on the level of regulation marked differences between organisms can be observed, both within and across different species. Regulation can be categorized into several levels: transcriptional regulation during expression from a gene locus, post-transcriptional regulation on the level of RNAs including splicing regulation and quality control, translational regulation during synthesis of the amino acid sequence of a protein and post-translational

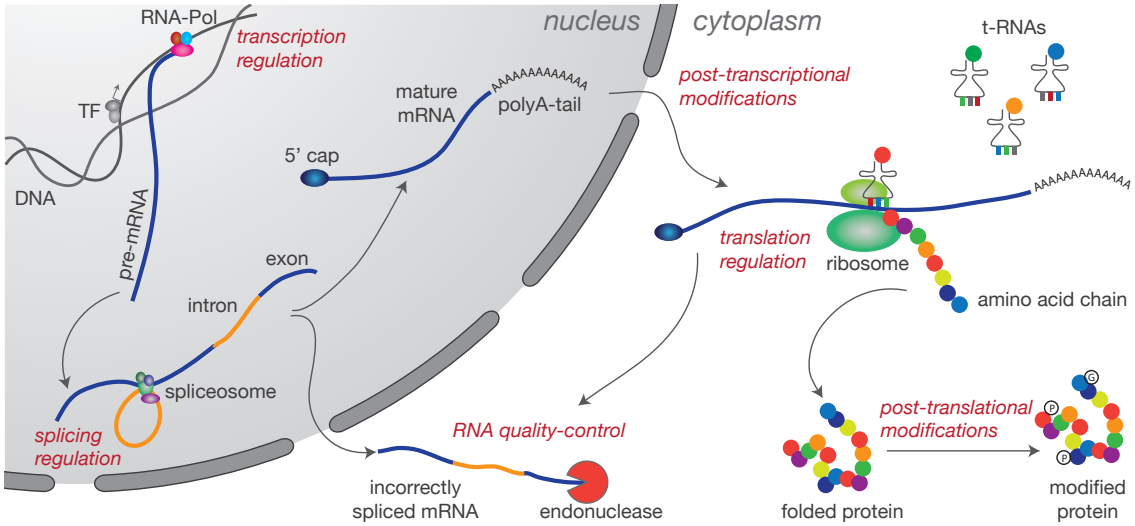


Figure 1.2: Overview on the processes involved in transcription and translation to form a protein from the information encoded in the DNA sequence. All biological entities are labeled in black and all regulatory processes are labeled in red italic.

modifications of single amino acids within proteins. The different levels of regulation are shown as red italic labels in Figure 1.2. In the context of this work we will omit translational and post-translational regulation and only focus on regulation on DNA- and RNA-level.

The overall-similarity of coding gene sequences between species is high. Estimates range to approximately 40% sequence identity between human and mouse genes [230, 308] and range to even 95% identity between human and chimpanzee [201]. Further, the number of genes in higher organisms is within the same order of magnitude, with estimates ranging from 20,000 to 30,000, not providing a good explanation for the drastic variation of phenotypes. This contradiction has also been termed the *G-value paradox* [252]. Over the past decades, numerous studies tried to resolve this disagreement and searched for differences in the genetic architecture of organisms. A first important factor that was identified, was transcriptional regulation [42, 280, 317]. Although the sequences of the genes are highly similar, the temporal and spatial expression patterns differ due to altered or differently used regulatory elements, that are most often located in a region a few kilobases (kb) upstream of the transcription start site, the *promoter region*. Depending on the presence or absence of certain *transcription factors* or other regulatory factors, e.g., hormones, the gene is expressed at different levels or alternative transcriptional starts are chosen.

A second difference that provides a better explanation for the similar number of genes, is the possibility to generate several versions of a gene from the same expression locus. These different versions are generally denoted as *isoforms* and have been shown to contribute to organism complexity [252]. The different isoforms are generated through a variety of different mechanism, which we will discuss in the following.

Alternative Splicing One of the most important processes that help to diversify the transcriptional outcomes of a single gene locus is termed *splicing* [23] and describes the

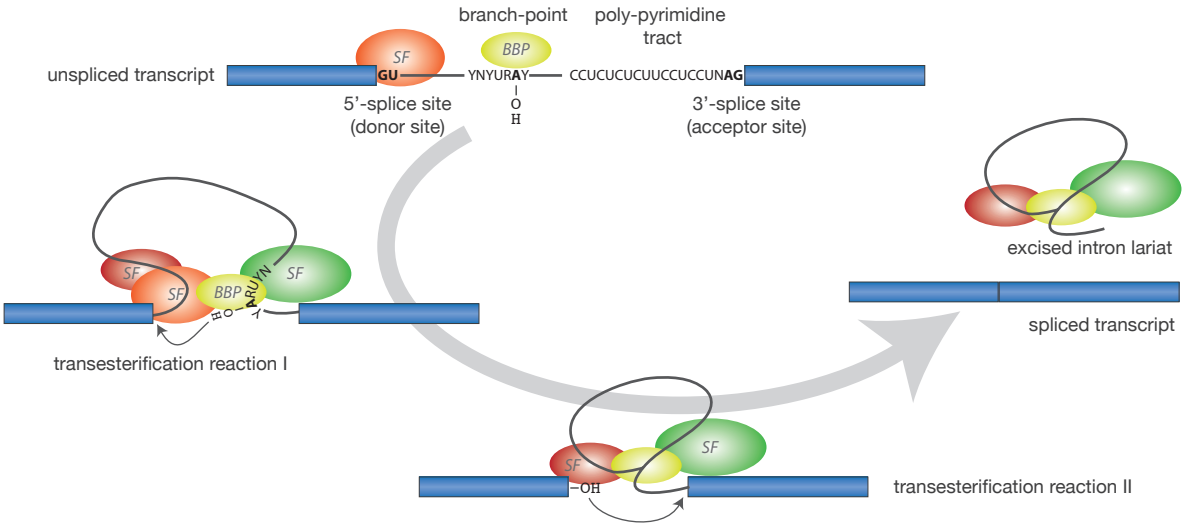


Figure 1.3: Steps of the alternative splicing mechanism. Exons are shown as blue boxes and the intron sequence as solid line or sequence of letters. Various splicing factors (SF) are simplified into a common representation. The branch-point binding protein (BBP) is shown as green oval. The hydroxyl-group is represented by an OH. (This figure was inspired by an illustration in [146].)

combinatorial excision and reconnection of parts of the pre-mRNA into a *spliced* mRNA. During splicing, a substring of the pre-mRNA, the *intron*, is determined and bound by a multi-protein-RNA complex, the *spliceosome*. The intron shows conserved sequence features such as a GU dinucleotide at its 5'-end (termed the *donor site*), an AG dinucleotide at its 3'-end (termed the *acceptor site*) as well as a conserved adenine and a polypyrimidine rich region near the 3'-end (termed *branch-point* and *poly-pyrimidine tract*, respectively). Several more sequence elements with higher or lower grades of conservation exist, that provide binding sites for different sub-complexes of the spliceosome. A transcript can contain multiple introns that are individually subject to splicing.

The illustration in Figure 1.3 shows a step-wise overview of the splicing mechanism. In the first step of the splicing process the adenine in the branch-point motif gets hydroxylated and subsequently, in a first transesterification reaction, gets bound to the donor-site guanine, moving the hydroxyl-group (OH) to the 3'-end of the 5'-exon. In a second transesterification reaction, the intron is cut at its 3'-end and the OH at the 3'-end of the 5'-exon is bound to the loose 5'-end of the downstream exon (cf. Figure 1.3). Whereas this describes the most common form of splicing, other less common mechanisms exist that rely on different acceptor and donor consensus sequences or require a different form of the spliceosome. For an in-depth description of the splicing mechanism, we refer to reviews on this topic [101, 146, 225].

If several isoforms are generated from the same gene and show differential use of exons or introns, the process is termed *alternative splicing*. By selective combination of different exons within a transcript, the possible number of isoforms grows exponentially with the number of exons used. Although most genes have a moderate number of expressed isoforms, there exist cases that use the full range of possible isoforms. The *Dscam1* locus in *D. melanogaster* produces 38,016 different isoforms that could be recently related to self-avoidance in neurite development [203].

Other Mechanisms of Diversification In addition to alternative splicing, several more mechanisms exist that allow for an expression of different isoforms from the same locus. One of them is the use of *alternative transcription start sites* [19]. In this case, several promoter regions exist that harbor binding motifs for specific transcription factors. Depending on intra- or extra-cellular signals, different factors bind, resulting in a differential usage of transcription start sites and thus in different isoforms. Also other molecules such as hormones are capable to bind promoter sequences and to alter gene expression.

Another mechanism to produce different isoforms is *alternative polyadenylation* [65], where different polyadenylation-sites at the transcript-end are chosen during maturation of the pre-mRNA. Although this mechanism does not necessarily alter the coding sequence of a transcript, it can add sequence regions with regulatory potential, such as microRNA-binding sites or structure-forming elements.

The process of *RNA-editing* does not alter the transcript-structure in transcription, but exchanges of modifies single nucleotides in the transcript, thereby altering the information encoded at the respective position. If such changes fall into start- or stop-codons of a transcript, the resulting protein can be drastically different or no protein is produced at all, if non-viable isoforms are created [99].

Alternative Splicing Events The differences between transcript isoforms arising from alternative splicing can be categorized into classes of alternative splicing events:

a) *Intron Retention* An intron is not spliced out and is retained in the sequence.

b) *Exon Skip* Also termed cassette exon. An exon is skipped, if the donor site of the preceding exon is spliced to the acceptor site of the subsequent exon.

c) *Alternative 5' Site* Also termed alternative donor site. For the same intron, different donor sites are used in different isoforms.

d) *Alternative 3' Site* Also termed alternative acceptor site. For the same intron, different acceptor sites are used in different isoforms.

e) *Multiple Exon Skip* Several connected exons are skipped, if the donor site of the exon preceding these exons are spliced to the acceptor site of the exon following these exons.

f) *Mutually Exclusive Exons* For the same pair of preceding and succeeding exons, more than one inner exon varies between isoforms.

A graphical representation of the different categories of alternative splicing events can be found in Figure 1.4. Alternative transcription start or stop sites caused by alternative promoter usage or alternative polyadenylation are generally not considered as alternative splicing events but rather as differences that arise from alternative transcript processing.

Splicing Regulation In addition to the various factors organized in the spliceosome that carry out the splicing process, numerous other elements exist that can influence both choice and efficiency of single splice sites. Based on the location within the transcript and the effect on splicing outcome, they are termed intronic splicing enhancers (ISE), intronic splicing silencers (ISS), exonic splicing enhancers (ESS) or exonic splicing silencers (ESS) [146].

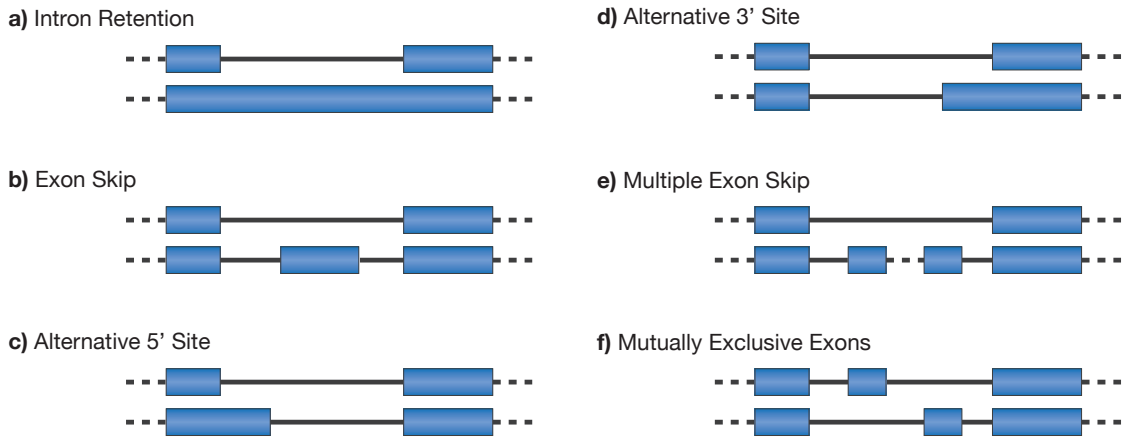


Figure 1.4: List of alternative splicing event types. Blue boxes are exons. Solid black lines are introns that have been spliced out. Dashed black lines can be arbitrary transcript structures in the remaining part of the gene.

These can be either binding sites for proteins that influence splicing behavior or structural motifs that alter the local secondary structure of the mRNA, thus influencing the choice of splice sites. The mechanism of splicing regulation is shown in Figure 1.5. Splicing factors that bind to ESE or ISE can strengthen nearby splice sites and thereby promote their use during the splicing process (Figure 1.5, Panel B). ESS and ISS have the opposite effect and make nearby splice sites less favorable, causing an alternative splicing pattern (Figure 1.5, Panel C).

Nonsense-mediated mRNA-Decay Although splicing is tightly regulated, the process can produce mis-spliced products. As such aberrant isoforms can produce proteins with toxic properties, e.g., aggregation [218], several control mechanisms are in place to remove them. The most prominent one is nonsense-mediated mRNA-decay (NMD), that triggers an immediate degradation of the mRNA molecule. NMD reacts to certain transcript features that are descriptive of aberrant transcripts. For instance, if splicing introduces a premature termination codon (PTC) into an inner exon of the transcript, thus causing a truncated product, in most cases the transcript will be degraded through NMD. Another known feature is the presence of an open reading frame (ORF) in the 5'-UTR [186]. Several studies also suggest a regulatory role for NMD exceeding error-correction, introducing the term *regulated unproductive splicing* [156, 161]. However, there are also studies that report aberrant isoforms that show descriptive features but escape NMD [112].

Although the mechanism is conserved in all eukaryotes, it is currently best understood in mammals. In the current model, the ribosome uses a pioneer round of translation to remove exon-junction-complexes (EJC), protein complexes that remain at the exon–exon junctions after completion of splicing. In case of premature termination the ribosome does not reach all exon–exon-junctions, thus leaving some EJC unremoved, which triggers the degradation reaction. Many factors of the NMD-machinery seem conserved over a wide range of organisms. However, various mechanistic differences have been observed, e.g., in plants, and a broad understanding is lacking. NMD is relevant for the work presented in Section 3.1, where we analyzed this process in the context of alternative splicing for the

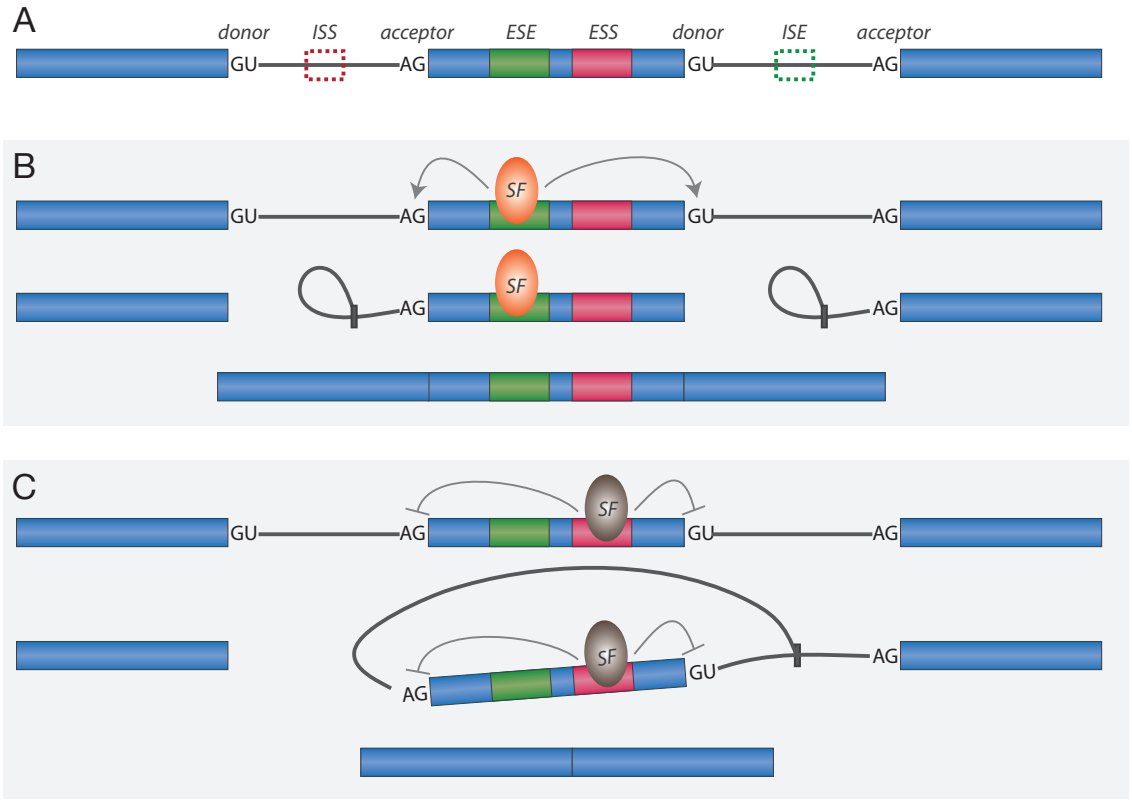


Figure 1.5: Illustration of the regulation of alternative splicing. Regulatory elements are shown as follows: exonic splice enhancer (ESE; solid green box), exonic splice silencer (ESS; solid red box), intronic splice enhancer (ISE; dashed green box), intronic splice silencer (ISS; dashed red box). Exons are shown as blue boxes and introns as solid gray lines. Donor and acceptor sites are marked by GU and AG, respectively. ***A:*** Structure of an unspliced transcript. ***B:*** Splicing factor (SF) bound to ESE, promoting nearby splice sites (gray arrows). The dark gray box in the second step marks the branch-point. ***C:*** Splicing factor (SF) bound to ESS, repressing nearby splice sites (blunt arrows). The dark gray box in the second step marks the branch-point.

model plant *Arabidopsis thaliana*. For further review on the details of the NMD mechanism and its characteristics in different species we refer to [186, 187, 296].

1.1.2 Relevance in Agriculture and Medicine

With its central role in transcriptional and post-transcriptional gene regulation, alternative splicing is key to a better functional understanding of regulatory mechanisms in many organisms. In the following, we will discuss two major fields of application: agriculture and medicine.

Splicing in Agriculture One important context for a better understanding of the role of alternative splicing is agricultural breeding. As plants are immotile organisms that cannot easily change their surroundings, complex regulatory mechanisms have evolved that adapt plant metabolism depending on various signals from the environment. Several studies have shown, that alternative splicing of transcript isoforms is crucial for adaptations. Splicing

has not only been linked to flowering time [74, 232], but also to the reactions caused by environmental stresses such as cold [119, 195], heat [177] or the response to plant pathogens [22]. Especially transcription factors show often alternatively spliced isoforms that either show dominant negative effects through peptide interference and competitive DNA binding [260] or are subject to the mRNA degradation pathway NMD [68, 126]. For a comprehensive discussion of alternative splicing regulation in plants, we refer to reviews in [22, 260]. Better understanding alternative splicing in plants will help to disentangle the complex relationship between environment and transcriptional regulation, to ultimately breed plants that are better adapted to harsh environments or less vulnerable to pathogens and environmental changes.

Splicing in Medicine The second major field of application for research on alternative splicing is medical diagnosis and treatment. For an ever growing list of diseases, causative links to aberrations in splicing are reported, including Alzheimer's disease, muscular dystrophies, cystic fibrosis and Parder-Willi Syndrome [94, 218, 281]. Causative changes can be discriminated in changes of *cis* sequence elements and changes of *trans* acting factors [94]. Causes *in cis* are mostly single nucleotide variants in the acceptor or donor sites or alterations in exonic and intronic enhancer or silencer elements, leading to aberrant transcript isoforms [281]. These isoforms are then either subject to degradation through NMD, resulting in a lack of protein product, or produce aberrant proteins with potentially toxic effects. An example for the latter is the protein Tau, where the mis-balanced expression of different isoforms can cause toxic self-aggregation leading to neurodegenerative disorders [70, 118, 218]. Causes *in trans* act through alteration of splicing factors, leading to expression of physiological isoforms in the wrong temporal or spatial context. Recent estimates suggest that a large fraction of mutations become disease-relevant only through splicing [43, 182]. Very rarely, changes to the core splicing apparatus lead to disease, as reported for retinitis pigmentosa [281, 315]. However, an even more important role of splicing alterations is described for cancer. In line with the complex progression of this disease, alternative splicing can be both key to loss- or gain-of-function mutations driving the cancer progression as well as a mere byproduct of the increasing genetic dysregulation in the affected tissue. Known examples for alterations that drive cancer progression are changes of the splicing factor SF3B1 in chronic lymphocytic leukemia [237, 305], alterations to the splicing factor SF2 in colon cancer [96], or aberrant splicing of tumor suppressor genes in colorectal cancer [278] or breast cancer [176] (excellent review in [268]). However, it is difficult to disentangle whether the splicing aberrations are initial cause or only an amplifier of cancer progression. Interestingly, recent studies show several promising therapeutic possibilities. Bonnal and colleagues [33] discuss several natural compounds originating from bacteria that specifically alter splicing of cancer relevant genes, with direct relevance for apoptosis (cell death) or angiogenesis (formation of blood vessels). An even more versatile approach is the synthesis of artificial oligonucleotides, that can specifically dimerize with mRNA to either enhance or inhibit the formation of a specific splice isoform [135, 143, 251, 307]. As a proof of concept, the technique was successfully applied to treat mice suffering from muscular atrophy [226].

As we discussed above, we are convinced that the research on alternative splicing has a rich field of applications with direct implications for plant breeding and translational medicine. In this work, we present algorithms that directly aim at the identification and characterization of alternative splicing events from sequencing data. We have the goal to gain a comprehensive overview of the alternative splicing state of a sample and to use our methodology to understand the functional and regulatory roles of alternative splicing, helping to identify important targets that can be subject to further analysis in biological and medical research.

1.2 Sequencing Technologies

Here, we give a brief introduction into the history and recent developments of sequencing technologies. We put our main focus on technologies relevant for the measurement of RNA and only mention other applications for completeness. The first part gives a short summary of techniques for whole transcriptome measurements and discusses the differences of hybridization-based and sequencing-based methods. Subsequently, we provide an overview of high-throughput techniques for RNA sequencing, discuss their differences and commonalities and give a short outlook on newly emerging technologies. Lastly, we introduce the different techniques that are commonly used to measure alternative splicing.

1.2.1 Historical Aspects of High-Throughput Sequencing

The major fraction of metabolic functionality in a cell is provided through proteins. However, measuring this whole pool of proteins at once, also denoted as the *proteome*, remains technically challenging, although progress has been made in recent years [21, 216]. Instead, it is much more feasible to measure the state of all RNAs present in the cell at a certain time point, also denoted as the *transcriptome*. Recent studies have shown, that a large fraction of protein diversity can indeed be explained already at RNA level [168]. To better understand recent developments in high-throughput transcriptome sequencing, it is necessary to review some historical aspects and introduce measurement techniques that have influenced the sequencing strategies that are used to today. In general, we can distinguish two man techniques to characterize a pool of RNA- or DNA-sequences: sequencing based methods and hybridization based methods. In the two boxes below, we provide a brief summary of both techniques.

Sequencing Based Techniques
This way of measurement is founded on a principle introduced by F. Sanger in 1977 [249]: sequencing by synthesis. (An alternative method for sequencing by digestion was suggested by Maxam and Gilbert in the same year [196].) Although the technique has been very much refined since then and various adaptations have been made, the core idea remains the same. As nucleotide sequences can be replicated from a single matrix strand, the processes incorporating new bases into the growing sequence can be utilized to produce a specific readout depending on what base is currently added. In his method, Sanger used a fraction of di-deoxy nucleotides to stochastically block synthesis at different positions and employed gels to sort fragments by length, producing a sequence of bands as readout. Techniques developed in the recent years use emission of fluorescence signals during base incorporation to detect the succession of bases added. This technique was first used for the

characterization of longer fragments of expressed sequence tags (ESTs) [29, 95] to resolve gene and transcript structures as well as tag-based quantification of genes, where short sequence fragments of a gene were used as proxy for measuring the expression of whole genes [107, 141, 297]. Soon after the scientific prototypes, commercial solutions became available that combined both approaches and made the quantitative sequencing of whole genomes and transcriptomes feasible. As these techniques are most relevant for the work presented here, we will discuss their implementation within a high-throughout setting separately in Section 1.2.2. Figure 1.6, provides a schematic overview of the sequencing-by-synthesis strategy suggested by Sanger.

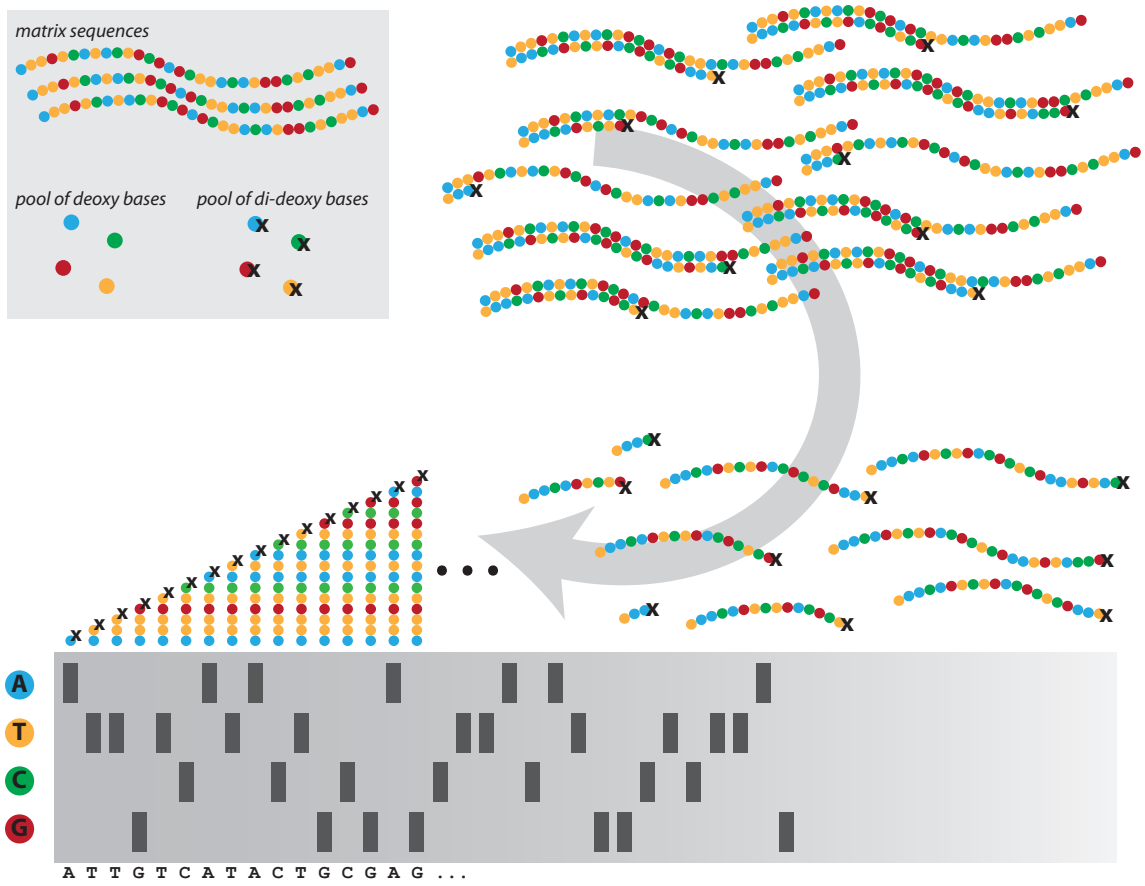


Figure 1.6: Schematic overview of the sequencing by synthesis technique after Sanger. A mixture of deoxy and di-deoxy bases is used to synthesize the complement of a given matrix sequence. Importantly, for one sequencing step all four kinds of deoxy bases are used but only one di-deoxy base (either A, C, T or G). Many parallel synthesis reactions against a multitude of copies of the given matrix happen at the same time. Once a di-deoxy base is incorporated into the sequence, the synthesis reaction stops (marked by an **X** in the schematic). As only a fraction of all bases is di-deoxy, the synthesis reactions stop randomly, creating a pool of sequences with different lengths. These sequences can then be separated according to their length, using gel electrophoresis. Each lane of the gel contains the sequences produced by one of four runs, each with a different di-deoxy base. The combination of gel-bands in the base-specific lanes can then be used to infer the sequence of the given matrix that was used for sequencing. This schematic describes the general principle. Numerous improvements have been made since its introduction, e.g., the separation of sequences by capillary electrophoresis.

Hybridization Based Techniques (Array Based Techniques)
Core principle of these techniques is the hybridization of nucleotide sequences through complementary base pairing. This central property of DNA and RNA that the organic bases adenine and thymine (uracile for RNA) as well as guanine and cytosine form specific base pairings through hydrogen bonds, is fundamental for DNA–DNA and RNA–RNA hybridization. For a more thorough introduction to the basics of nucleic acid sequences we refer to the respective textbooks [8, 139, 148]. The property that two nucleotide sequences with complementary base structure form the energetically most favorable binding, can be used to construct specific sequences as baits to fish for the complementary counterpart. If a short DNA- or RNA-sequence is now immobilized on a surface and then a large number of short different DNA or RNA molecules in solution is presented to that bait sequence, only complementary sequences will bind to the immobilized bait with a relatively high specificity. In DNA- and RNA-microarrays, this principle is applied in a high-throughput manner. That is, oligonucleotides (also oligos, short sequences of approximately 25–50 nt) are immobilized or printed onto a chip in a grid layout such that each sequence can be linked to a coordinate within the grid. After hybridization, the array is washed and only molecules bound to the fixed oligos remain on the array. Fluorescence techniques are then used to measure whether grid positions have molecules bound to them, which provides the read out of which sequences have been present in the sample. This technique was first described to measure the expression of single genes through tags [179] and has then been further improved to assay the whole transcriptome [316] or survey expression on the whole genome through the generation of genome-wide tiling arrays [7, 215, 259]. In Figure 1.7, we provide a schematic overview of this technique.

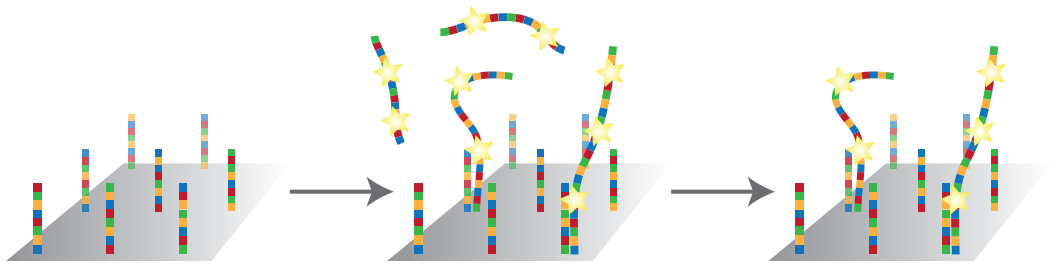


Figure 1.7: Schematic overview of the hybridization-based technique. Bait sequences are immobilized on a surface. Fluorescently labeled target sequences are given over the array. After a washing step, only sequences specifically bound to the array remain, providing a visual readout.

The key difference between the two techniques is the kind of information they are able to provide. Whereas sequencing is able to collect qualitative information and is therefore used to determine the sequence of an unknown DNA- or RNA-sample, hybridization based methods can only collect quantitative information for known sample probes. Only the technological developments of the recent years made it possible for sequencing methods to become quantitative, as we will describe in the following.

High-Throughput Sequencing To overcome the limitation of traditional Sanger sequencing of purely qualitative analysis, a larger throughput of sequences had to be achieved, such that count statistics could be used to infer quantitative information. Historically, the general efficiency of sequencing was limited by two major factors: the manual labor of cloning a sequence fragment into a vector for amplification and the limited number of a

few hundred nucleotides produced during a single experiment [138]. Whereas the latter was difficult to address, the problem of manual labor could be solved by automation. Especially the development of capillary array electrophoresis and automated detection systems increased throughput and made first commercial solutions possible [138]. However, even this increased throughput was still not large enough for quantitative readouts, but enabled large scale qualitative projects such es the shotgun-assembly of the human genome [298]. It was only in the recent years that another solution was found to further boost the throughput, an automatized sequence amplification independent of vector cloning, which allowed for a high grade of automation and parallelization and was able to sequence millions of very short fragments of DNA at a time. This strategy is also known as *high-throughput shotgun sequencing* and the resulting short sequences are commonly denoted as *reads*. Which different implementations of this automated principle have been developed since then and how these high-throughput techniques can be adapted for RNA-profiling will be discussed in the subsequent Section 1.2.2.

High-Throughput Sequencing Compared to Arrays Array based techniques provided the first way to quantitatively asses molecular sequences and were therefore the method of choice for analysis of gene expression. However, the advent of high-throughput sequencing methods that provided both quantitative and qualitative information has caused a paradigm shift that has revolutionized the fields of genomics and transcriptomics over the past years (cf. Section 1.2.2). Since then, sequencing has largely replaced array based techniques, which can be explained by a number of advantages the sequencing techniques provide. Whereas array-based techniques have a limited detection range due to saturation effects of hybridization, sequencing techniques show a much larger dynamic range, as quantification is based on counts generated from the amplified sequence fragments [49, 206]. A second advantage of sequencing is that it still provides qualitative information and is thus able to detect novelties, as no previous knowledge about the sequences to detect is required. This is not only helpful to annotate new organisms but also to identify novel splicing isoforms that are rare or originate as a result of disease, to analyze transcribed regions of the genome that produce non-coding RNAs or to investigate the many different classes of small RNAs, such as miRNAs or piRNAs. Further, it is reported that generally lower amounts of DNA or RNA material are necessary to run the assay [306] and a sequencing approach is generally less laborious [220]. Finally, recent versions of the sequencing-technology show a much larger throughput at tremendously reduced costs.

Both techniques show very different sources of error. Whereas array-based methods suffer from background noise originating from cross-hybridizations [291, 318], physical problems [253], and hybridization bias based on sequence composition [28], sequencing approaches suffer from biases through priming in PCR amplification, non-uniform fragmentation, errors in base calling and artifacts from mapping [67, 106, 193, 209]. For both platforms, these problems have to be tackled computationally by appropriate noise models and filtering techniques. Solutions for sequencing-data will be discussed in detail in the methodological part of this work. For a detailed technical description of similarities and differences of both approaches, we refer to the work of Marioni and colleagues [193].

Adaptation to Other Sequence Sources The two techniques described above can also be used for the quantitative or qualitative assessment of other molecular phenotypes and have been applied in a wide range of different contexts. Hybridization based methods in the form of SNP-arrays are used for the assessment of genomic variation at a given set of positions (single nucleotide polymorphisms, SNPs) and have been successfully used in population genetic studies [293] or to detect copy-number variations in cancer [114]. A hybrid approach of both technologies can be found in whole exome sequencing that is based on targeted exome-capture through array-hybridization [48, 282]. We used data from whole exome sequencing for the work presented in Section 3.4. Another relevant application is whole genome sequencing, that provides measurements of the full DNA sequence. Based on this data numerous projects such as personal genome assemblies and studies for the assessment of genetic variability through re-alignment have been realized [6, 160, 231]. Especially sequencing has found applications in a large number of further contexts that exceeds the scope of this introduction. We will name only some representative examples: SHAPE-Seq [18] to resolve RNA secondary structure, DNase-Seq [55] to assess chromatin structure, 3'-Seq [171] to measure alternative polyadenylation or ChIP-Seq [122] to assess protein–DNA interactions. The implementation of high-throughtput RNA-Sequencing will be discussed in the following.

1.2.2 High-Throughput RNA-Sequencing

In this part, we focus on different methods for high-throughput *RNA-sequencing* (RNA-Seq). As already mentioned before, RNA-Seq is mainly based on the high-throughput sequencing-by-synthesis technologies developed for DNA [263]. However, exceptions to this are single molecule sequencers for RNA, such as PacBio, which we will discuss separately at the end of this section. The high-throughput implementations described below were designed for DNA-sequencing. However, an adaptation from DNA- to RNA-sequencing is straightforward by translating the RNA into DNA at some point in the sequencing protocol. The set of short fragments used for sequencing is denoted as *sequence library*. We will begin by describing the preliminary steps of library preparation, which is shared across all techniques presented in the first half of this part.

Library Preparation After extracting RNA from a population of cells of interest, the first step is to select for RNA molecules of interest. Depending on the research context, either all RNA present in the cell or a specific subgroup such as mRNA or rRNA can be selected for. Usually, the aim is to sequence mRNAs, which requires either an enrichment by poly-A selection, depletion of ribosomal RNA or a combination of both. Various protocols and commercially available toolkits have been developed to accommodate this step [97, 110, 228, 275, 279]. In the next two steps the long molecules are fragmented using chemical hydrolysis or physical force (nebulization) and then translated into complementary DNA (cDNA). These two steps are used in varying order in different protocols. We limit ourselves here to state that this results in a set of short double-stranded cDNA fragments and refer to the literature for further discussion [13, 53, 189, 191]. The short fragments are then ligated to specific adapter sequences that facilitate amplification as well as serve specific needs of the later sequencing method. Subsequently, the fragments are amplified by a polymerase chain reaction (PCR) [207], resulting in an exponentially increased number of fragments.

Amplification is either done with unspecific random hexamer primers or oligo-dT primers that preferentially amplify fragments with a poly-A stretch. Both techniques cause different biases that need to be taken into account for analysis [53, 106]. As most sequencing protocols are limited in the number of bases they can produce, the amplification step is followed by a step for size selection that only retains fragments within a certain length range. Size selection is either done through gel electrophoresis or by using commercially available kits. In most protocols, the information regarding the strand the RNA originated from is lost. However, several adaptations to the library preparation protocol have been developed that allow for preservation of this information [159]. For further review on library preparation, we refer to [13, 53, 222].

454-/Pyro-Sequencing This technique was one of the earliest high-throughput sequencing methods [244] and the first to be commercially available [53, 192]. Initially, this approach produced several hundred thousand reads up to 100 nt in length [53]. Since then, improvements in technology and protocols have increased throughput further to ∼1,000,000 reads of 1,000 nt length (Roche GS FLX Titanium XL+[1]). The basis of 454-sequencing is an emulsion PCR. The single fragments generated in library preparation are transfered into tiny drops within a water–oil emulsion. These small drops contain agarose beads coated with oligonucleotide sequences complementary to the adapter sequence fused to the fragments during library preparation. Stochastically, each drop will contain a single bead that has one sequence fragment bound to it. Fragments are then PCR-amplified within the drop, each new copy binding again to the oligo-covered bead. After amplification the drops are broken up, retaining beads that each contains a large number of identical copies of the same sequence fragment. The beads are then loaded onto a so called PicoTiterPlate, containing thousands of tiny wells that can hold exactly one bead at a time. The wells build a coordinate system that unambiguously identifies each bead. Sequencing is then performed in several rounds of sequencing-by-synthesis. Upon incorporation of each base, a fluorescence signal is emitted, which is recorded by a camera. Stacking the images from each round of base addition and using the information of the 2D well coordinate system, the read sequence for each bead can be reconstructed [189, 192]. Since $\sim 10^6$ copies of the same library fragment are present on the same bead, the same base is incorporated in all sequences at once, thus amplifying the emitted light signal, leading to an improved signal-to-noise ratio.

Illumina Sequencing The Illumina method of sequencing has been introduced as Solexa sequencing in 2006 and is based on a so-called *flow cell*. The technique has since been used for a large number of projects and has become a quasi-standard of high-throughput sequencing. While initial versions produced tens of millions of rather short reads of 32 nt length [53], recent machines can generate up to $3 \cdot 10^9$ reads of 2×150 nt in length (Illumina HiSeq X[2]). The flow cell is a small glass device that contains eight identical flow channels. The bottom surface of each channel is coated with adapter sequences to bind the prepared library fragments. After the fragments bound to random locations in the flow channel, they are amplified via *bridge-amplification*, which works as follows. Each library fragment has two distinct adapters, one fused to each end. The surface of the flow channel contains both

[1] http://454.com/products/gs-flx-system/index.asp
[2] http://res.illumina.com/documents/products/brochures/brochure_sequencing_systems_portfolio.pdf

adapter-complements such that the fragments can bind with both ends forming a bridge-like structure. In the next step, the sequence complementary to the bound fragment is synthesized resulting in a double-stranded molecule. Denaturing the double-stranded fragment results in two copies of the same fragment bound in local proximity on the bottom surface. Repeating bridge-amplification for several rounds, results in local clusters containing up to 10^6 copies of the initial library fragment. As described before, this is sufficient to perform sequencing-by-synthesis. Each sequencing cycle contains four rounds, where in each round one distinct type of labeled nucleotides is incorporated into the sequences. After excitation with a laser, the incorporated nucleotides emit a fluorescence signal that is captured by a camera device. As the cluster positions remain fixed during all cycles, the sequence of images can be used to unambiguously reconstruct the sequence of bases added in each cluster [13, 189, 200].

SOLiD-Sequencing The SOLiD technique was first described in 2005 [264] and has been commercially introduced by Applied Biosystems in 2007 [13, 224]. With an initial throughput of $5 \cdot 10^7$ reads with a length of 35 nt per run [13], several improvements to the technology have resulted in a current throughput of $4 \cdot 10^8$ of 2×50 nt reads per run (Applied Biosystems, 5500xl W[3]). The main principle of this technology, is a modified sequencing-by-synthesis procedure, based on a DNA ligase and cleavage of a structured octamer. Analog to 454-Sequencing, the library fragments are amplified using emulsion PCR but are adapter-ligated to paramagnetic beads instead of agarose beads. In each sequencing cycle, a population of structured, fluorescently labeled octamers is ligated to the template sequences. In this context structured means, that two positions in the octamer strictly correlate with a certain fluorescence label. All possible 16 dimers correspond to four fluorescent dyes, four dimers per dye. Using this system, the octamers bind specifically to sequences that are complementary in this two positions, making them identifiable by the dye. As other systems, SOLiD sequencing works in rounds each containing several cycles: once the octamer bound a sequence, laser excitation triggers the fluorescence signal, that is measured as described before. Then the fluorescent label is removed by cleaving the octamer between positions 5 and 6, which completes one cycle. The next cycle starts with annealing a new octamer, identifying the nucleotide 5 nt downstream of the previous. Thus, in one round each 5th nucleotide starting at position n can be identified. By using a longer initial primer, generating a larger offset in the next rounds, every 5th nucleotide starting at position $n+1$, $n+2$, $n+3$ and $n+4$ can be identified, providing the full sequence. Using dinucleotides for specific binding directly implements an error-correcting code and increases accuracy for base calling, as for each base two measurements are taken. For further review of this technique, we refer to [13, 189, 224, 263].

Paired-end Sequencing A meanwhile common extension to these high-throughput sequencing techniques is the generation of paired-end reads. While in previous protocols only one side of the library fragment was subject to sequencing, the paired-end extension allows for the sequencing of both ends. Already developed for cloning based approaches [88], paired-end sequencing has been successfully adapted for all three sequencing technologies described above. Importantly, the pair-relationship of two reads remains identifiable after

[3] http://tools.lifetechnologies.com/content/sfs/brochures/5500-w-series-spec-sheet.pdf

sequencing and their approximate distance can be computed from the fragment length distribution. This additional information aids the correct alignment to a reference sequence or can act as evidence for long-range dependencies within transcript structures.

Newly Emerging Sequencing Techniques Along with the constant improvements of the techniques described above, many novel technologies have been developed that aim to overcome remaining limitations of the shotgun sequencers or shall provide a more cost efficient alternative. These techniques are also denoted as *single molecule sequencing* as fragment amplification is no longer required.

A system that is based on semiconductor technology is developed by Life Technologies. The Ion Torrent technology uses single hydrogen atoms released during base incorporation to detect whether a base was added to the sequence. As the number of hydrogens freed is proportional to the number of nucleotides added, homopolymers (stretches repeating the same nucleotide) can theoretically be detected at once. Initially very error-prone, the technology has been improved [236] and has been successfully applied to metagenomic samples [124, 314].

The PacBio real-time sequencer developed by Pacific Biosystems uses zero-mode waveguide detectors [158] that are fused to a single DNA polymerase molecule to detect the phospho-labeled single bases during incorporation. A more in-depth review of the method can be found in [200]. PacBio-reads have an average length of 1,000–2,000 nt but suffer from a rather large error rate of up to 15% [236]. However, in combination with shotgun sequencing and error correction, the method has been successfully used for genome assembly [239]. Also other sequencing modes exist, that perform several sequencing runs on a circularized molecule and average over the iterations to improve read quality.

Based on the same idea as PacBio systems to read along the DNA sequence in real time but omitting the synthesis-step, the technology utilizing nanopores identifies each nucleotide while the DNA or RNA molecule is sliding through the pore. Nanopores are essentially tiny holes in either a biological membrane or in synthetic material [20, 312], that are embedded within a bilayer structure, which results in the flow of a low ionic current when low voltage is applied. Different bases sliding through the pore will specifically change the temporal profile of that current, providing a readout for sequencing [255]. Problems of this technique are the speed at which the DNA passes the sensor as well as physical interaction between DNA and the pore. A growing body of work is addressing parameters and algorithms for base calling [234, 286]. Potentially this technique will also be able to detect epigenetic modifications of DNA or modifications of RNA, thus providing an additional layer of information [301].

Quality of Sequencing Whereas sequencing following the Sanger method based on *in vivo* amplified DNA fragments had an error rate as low as 1 in 10,000 for automated capillary sequencers [80, 138], the speed of newer high-throughput sequencing techniques comes at the cost of accuracy. Pyrosequencing has a substitution error rate in a range of 10^{-3}–10^{-4} [138, 192, 238] with a bias towards short insertions and deletions (indels), especially at homopolymers [117, 138, 238]. However, pyrosequencing has still the lowest error rate of the high throughput techniques. The error rate for SOLiD systems is slightly higher, resulting from a higher background error rate due to amplification or ambiguities arising

from the interference of beads. With an average rate of 10^{-2}–10^{-4} [138] (depending on error correction), it has a medium error level. Generating the highest throughput, Illumina sequencing has also the highest error rate, which currently resides in the range of 10^{-2}–10^{-3} and arises mainly from a high background error rate through amplification [54, 67, 138, 235, 236]. However, there have been studies, showing examples of sequence specific, non-random errors in Illumina reads [210], complicating the development of models taking the base calling error into account.

Although many newly emerging technologies address many weaknesses of shotgun sequencing approaches, they also generate new problems that require further research. Especially error rates and measurement biases are not thoroughly studied yet and require a better understanding. First studies report and compare error rates for the new approaches, resulting in a rate of 1.8% for the IonTorrent platform and a range of 12–17% for PacBio [145, 236]. Especially non-random distributions of error patterns are problematic, as they can introduce systematic biases into the measurements. Although PacBio is claimed to produce randomly distributed errors [41, 145, 236], there is still an ongoing debate about possibly undetected systematic errors and first theoretical studies in that direction appear [217].

Sequencing quality is usually computed on a per-nucleotide basis during the base-calling step of sequencing and is expressed as the probability of a wrong call. It has become commonly accepted to use the phred-scale [80], i.e., the negative logarithm of the error probability p, to express the quality value:

$$q := -10 \cdot \log_{10}(p).$$

The quality values for each sequence are commonly represented as a quality-string, consisting of ASCII encodings of the quality values for each base.

1.2.3 Measuring Alternative Splicing

Many techniques covered in this introduction can be applied to measure alternative splicing. For reasons of brevity, we will only discuss the most common ones. As most transcriptome analyses are based on mature mRNA, where introns have already been removed from the sequence during splicing (cf. Section 1.1.1), the different exon combinations have to be considered for the measurement.

Exon Junction Arrays This hybridization-based technique relies on a microarray that is coated with short sequence fragments that span over exon–exon junctions in the mature mRNA transcript. This set can be determined from all junctions present in a given annotation, but can also be a set of junctions that is inferred from annotated exons, allowing for novel combinations of exons that are not observed in any database [123, 262]. However, even if novel combinations are present on the array, only a limited number of novel events can be detected. Events that create exons unseen before, can still not be detected. A further limitation might be the unavailability for non-model organisms. However, in cases where this approach is applicable, it can be a cost efficient alternative to RNA sequencing.

EST Sequencing Sanger-sequencing of expressed sequence tags (ESTs) was historically one of the first methods that enabled a larger scale analysis of alternative splicing. Based on

cDNA, it was first used for gene finding and annotation [5, 81] and later also to determine the structure of transcript isoforms [35, 105, 204]. Due to the laborious cloning part that is necessary to create the cDNA libraries, this method has been mostly replaced by other sequencing techniques. Due to the low throughput of Sanger-sequencing, no quantitative information was available. However, evidence from EST databases has been a valuable resource for splicing research [283].

Deep RNA-Sequencing As introduced in Section 1.2.2, RNA-Seq is based on the highly parallel sequencing of mature mRNA and generates millions of short reads as output. It has numerous advantages over array-based techniques, including the capability to detect novel splice junctions or to detect isoforms expressed at a very low rate. Further, recent RNA-Seq protocols are more cost and material efficient than most other techniques. Various approaches have been suggested, to transform the read set into quantitative and qualitative information for the measurement of alternative splicing. In the following section, we will discuss several common strategies for the analysis of RNA-Seq data.

1.3 High-Throughput Data Analysis

Technological improvements have led to drastically dropping sequencing costs over the recent years. The production of new sequence data has slowly outgrown any present advances in capacity for storage and computation [190, 274, 313] and has largely replaced array based methods for transcriptome analysis. Although this increasing amount of data enables analyses at an unprecedented depth, it comes with challenges on the computational side. In this section, we will introduce the most common steps in RNA-Seq based transcriptome analysis pipelines and show typical applications. We will further discuss difficulties arising from large scale analyses and describe how they are commonly tackled. Lastly, we will introduce genome-wide association studies (GWAS) as another example for an analysis principle that has largely profited from the growing amount of sequencing data and that was used for the studies presented in Sections 3.3 and 3.4.

1.3.1 Qualitative and Quantitative Transcriptome Analysis

Transcriptome analysis pipelines based on RNA-Seq data can be generally sub-divided into three major phases: the alignment of reads to a reference sequence, the identification of transcript isoforms and the quantification of genes and/or transcripts. Data generated from these initial steps is then often used for further downstream analyses such as differential analysis between conditions, the computation of enrichment scores with respect to a given functional annotation or as phenotype data within an association study. In the following, we will briefly introduce each of these analysis steps and provide examples for commonly used tools.

Read Alignment Goal of this initial phase, that is also denoted as *read mapping*, is to identify for each sequencing read the genomic location it most likely originated from. Complicated by short read lengths, low quality of sequencing information or regions of low sequence complexity, for instance, repeats in the target genome, the alignment can result in no found location as well as a long list of equally likely locations. Aligning sequence

reads originating from mRNA to a genome sequence is more difficult than aligning genomic DNA, as the mRNA undergoes a process of maturation that removes certain sequence parts (cf. splicing in Section 1.1.1) and makes it necessary to split up alignments of a read into several segments, also denoted as *spliced alignments*. The alignment data is then used for downstream analysis. Especially the *coverage*, that is the number of reads overlapping a genomic position, is used to infer information regarding expression and transcript structure. Common tools for the mapping of RNA-Seq reads are TopHat [137, 287], STAR [66] or PALMapper [121]. A comparison of RNA-Seq alignment tools can be found in [79]. As the alignment step is central to several methodological contributions presented in this work, we provide a more thorough introduction in Section 1.4, with a focus on alignment of RNA-Seq data in Section 1.4.2.

Transcript Reconstruction If the genome annotation is incomplete or not available at all, it is often necessary to reconstruct the set of transcript isoforms expressed at a gene locus. This can be either achieved using the sequence alignments from the first step or in case no reference genome is available through direct assembly of the sequencing reads. In the first approach, an existing annotation can be augmented with information from read alignments (cf. Section 2.5) or a splicing graph that integrates all exon and intron information can be directly inferred from the alignment data, using the coverage to identify exons and spliced read alignments for introns. A transcript isoform is then represented as a path through that graph. However, due to the large number of possible paths, the identification of expressed transcript isoforms from this graph is a computationally hard problem. Several alignment-based algorithms tackling this problem exist. For instance, Cufflinks [287, 290] tries to identify a parsimonious set of paths through the graph that best explain the observed coverage. Scripture [103] first generates all possible paths that are subsequently evaluated for significance, easily getting infeasible for larger graphs. MiTie [25] employs a mixed integer programming strategy to optimally choose sparse sets of exonic segments that then form transcripts to explain the coverage profile within several provided RNA-Seq samples. The second approach are assembly-based strategies, that use algorithms inspired by genome assembly mostly utilizing De Bruijn graphs [60] and are often extensions to DNA based genome assemblers. Prominent examples are Oases [256], Trans-ABySS [240] and Trinity [100].

Expression Quantification Estimating gene expression from RNA-Seq data is a non-trivial task as well and has been actively discussed within the research community over the past years. The main goal is to estimate how many copies of a transcript isoform were present in a given sample. The term *gene expression* is sometimes misleading in higher eukaryotes, as a gene can produce several isoforms from the same gene locus. Thus the expression of a gene can be a mixture of several isoforms, resulting in two main strategies to infer gene expression: *count-based* estimation and *isoform-based* methods. Count-based methods integrate exons or exonic segments of all isoforms and determine a subset that is used towards estimating gene expression, e.g., by taking the union of exons over all isoforms or by using only exonic segments resulting from an intersection of the isoforms. The gene expression is then determined from the number of reads overlapping to those exons in alignment, leading to possible biases through different isoform lengths or biases in read dis-

tribution over the transcript. Isoform-based strategies, also known as *isoform deconvolution*, try to assign each read to a source-isoform, only counting the reads towards the expression of the isoform it belongs to. These counts are then used to estimate how many copies of an isoform were present in the sample. Gene expression is then computed as the sum of all isoforms. The results of this approach have been shown to be more accurate in certain settings, but are computationally more expensive. The differences between both strategies are discussed in [289] and [304]. Popular count-based methods include HTSeq in the DESeq package [12] and edgeR [242]. On of the first implementations of a method for isoform deconvolution and the first to take processing biases into account was rQuant [30, 31]. In most recent implementations isoform identification and quantification are coupled within the same optimization problem, e.g., in Cufflinks [290], in MiTie [25], or in MISO [131] that uses a bayesian approach.

Downstream Analyses Based on quantification values for expression on gene- or isoform-level, various directions of downstream analysis are commonly established. Methods for the differential analysis between two or more conditions are very often already incorporated into the quantification tools [12, 290]. Approaches for differential testing initially relied on a Poisson model [242] which is more and more replaced by models based on a Negative Binomial distribution that better reflect overdispersion due to biological variability and do not rely on a linear mean–variance relationship thus producing less false positives within the test [12, 69]. Recently, also non-parametric, annotation-free approaches based on the maximum mean discrepancy test have been developed [69].

Results from the differential analysis can then be used for functional enrichment tests. Common strategies utilize the ranked list of genes resulting from differential testing, to either test for enrichment of gene ontology (GO) terms [14, 108] in the top ranks of the list [75] or to compute enrichment scores on pre-defined sets of functionally related genes (gene set enrichment analysis) [276].

Other analyses are not necessarily based on differential testing, such as motif searches in or around transcriptionally active regions to identify binding sites of transcription- or splicing-factors, homology searches to identify closely related entities in other organisms or the integration with other data sources providing conservation scores, epigenetic marks or binding profiles of various factors to put the findings into a functional context.

Most transcriptome analysis pipelines based on RNA-Seq data follow the steps described above more or less closely. For each of the described tasks many more tools exist. We tried to chose examples that are commonly used and represent a certain way of analysis. For a broader review of this topic, including the description of more implementations, we refer to [93, 229].

1.3.2 Genome-Wide Association Studies

As already discussed in the biological introduction in Section 1.1, a central question of biological research is to understand the relationships between the entirety of properties of an organism, its *phenotype*, and its heritable information, the *genotype*. Especially the wide range of diversity or variability within a certain trait has raised the interest of researchers. A principle suggested by Fisher in 1919 [84] was to analyze the variability of a trait within or

across populations and link it to heritable information. Here, we will give a very basic introduction to the techniques used in genome-wide association studies to elucidate relationships between phenotypic and genotypic variation.

Following the central dogma of molecular biology, all information required to build an arbitrarily complex organism is encoded in its genome. Whereas the concept of a genome originally included the sequence of DNA only, notions have since been extended to also contain heritable sequence modifications, such as methylations, forming the epigenome. Further, it has been long established that not all of the phenotypic variation can be explained by the genotype alone and that environmental factors can play a substantial role. Further developing Fisher's ideas, different models have been proposed to interrelate genotype- and phenotype-data. Most common and traditionally used are linear models that follow a regression approach to explain the variance of phenotype Y through a weighted linear combination of fixed genetic effects X or environmental factors plus an additional error-term ϵ assumed to follow a Gaussian distribution:

$$Y = X\beta + \epsilon, \quad \text{with } \epsilon \sim N(0, \sigma_\epsilon^2 I).$$

However, this model does not consider several important factors possibly confounding the analysis, such as the non-random inter-relatedness of individuals, termed *population structure*, or the geographic sub-structure of samples [16, 40, 285]. To also account for such effects, linear mixed models (LMM) have been proposed [327], that augment the model to also account for the non-random genetic similarity within a study population:

$$Y = X\beta + P + \epsilon, \quad \text{with } P \sim N(0, \sigma^2 P_K),$$

where P_K is a kernel matrix containing the pairwise genetic similarity of all individuals. With this general model, it is also possible to account for numerous other contributors to variation, such as batch-effects, other categorizations of input data (sex, smoker/non-smoker, ethnicity) or geographical and environmental factors [175, 299, 325]. In recent years, there is also work on the problem how to incorporate hidden confounders into the model [89, 90, 175]. However, a main limitation of these approaches is that they neglect epistatic effects, that is the non-additive combination of single variants, leading to an unexplained gap in variation, also termed *missing heritability*, that was widely debated in the field [76, 185, 329].

The main idea of *genome-wide association studies* (GWAS) is to measure a certain phenotype, for instance, disease vs. no disease, in a preferably large population and model its relationship to the genotype information, usually the allele at a certain position in the genome, also termed *single nucleotide polymorphism* (SNP). With the null hypothesis that there is no linear relationship between genotype and phenotype, one can then test if the genotype's contribution to the linear model explaining the variance is significantly larger than zero. To identify variant-locations in the genome that show a significant correlation to the phenotype data, all such variants need to be tested, resulting in a large number of test instances, ranging up to 10^7 in a GWAS for the human genome. To accommodate this need for efficient computation, several fast methods have been proposed in the recent years [128, 173, 325, 328]. For the computations presented later in this work, we use the LIMIX package for Python [174].

1.4 Alignment of High-Throughput Sequencing Data

As already discussed, alignment is a central step in most analysis pipelines utilizing high-throughput sequencing data, with the aim to identify for each read its genomic origin. Thus, the alignment problem for RNA-Seq data can be summarized as follows. Given a short query-sequence and a long target-sequence, the task is to identify all locations in the target that are identical or of highly similarity to the query. Many different algorithms have been developed that aim to solve this task. We will first discuss DNA-alignment, the computationally easier variant of the problem, before we review strategies to align RNA-Seq data to a genomic sequence.

1.4.1 Strategies for DNA-Sequencing Alignments

There exist many different definitions of alignment. Here, we will define it in the context of biological sequence analysis. Given two strings $A = a_1a_2 \dots a_m$ and $B = b_1b_2 \dots b_n$, an alignment of A and B is a set of index pairs (i, j) with $i \in \{1, m\}$ and $j \in \{1, n\}$ assigning positions in A to positions in B, where each position can be assigned at most once. We further require, that the order of positions within A and B is conserved. That is, if a_i is aligned to b_j, any position after a_i needs to align to a position after b_j. If we require the assigned positions a_i and b_j to be identical, the alignment is called *exact*. However, for our application it is also necessary to find *approximate* alignments, where a_i and b_j can mismatch. An alignment can further contain gaps. That is, consecutive positions in one string do not necessarily align to consecutive positions in the other string. Generally, two types of alignments can be distinguished: global and local alignments [73]. Whereas global alignments require all positions in A and B to be aligned, local alignments find the optimal alignment of substrings of A and B, leaving certain positions unassigned. In the context of this work, we deal with semi-global alignments, where we require all positions of only one string to be part of the alignment. As the number of possible alignments under this definition is very large, it is necessary to define a scoring scheme that evaluates the quality of an alignment. Most algorithms assign a cost or reward to each pairing used in the alignment and compute a total cost or reward over all positions. The optimal alignment minimizes the cost and maximizes the reward. A standard algorithm to find global alignments using dynamic programming was developed by Needleman and Wunsch [212]. An adaptation to local alignments was introduced by Smith and Waterman [267]. Both algorithms are still central to many bionformatics algorithms and have inspired many further adaptations and improvements, e.g., in space complexity [111] or affine gap penalties [9, 98]. However, as we will see shortly, a direct application of these algorithms to our problem is prohibitive due to the extensive computational cost. For a more thorough introduction to the alignment problem, we refer to [73].

In case of RNA-Seq alignments, assume the following as given: an alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{T}, \mathtt{G}\}$, a very large number ($> 10^7$) of short query strings $S = s_1s_2 \dots s_m$, with $s_i \in \Sigma$ and m usually between 30 and 150, as well as one long target string $G = g_1g_2 \dots g_n$, with $g_i \in \Sigma$ and a total length n of up to $3 \cdot 10^9$. The goal is now to find for each short string S all its approximate occurrences in G allowing up to k mismatches or gap positions, usually ranging from 1 to 10, as a function of m. An optimal search with the Smith–Waterman algorithm has a space and time complexity of $O(n \cdot m)$ for each query sequence and would

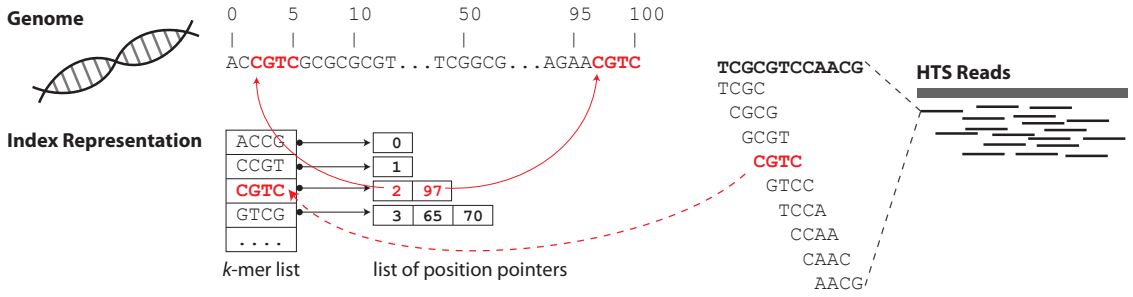


Figure 1.8: Example for the alignment seeding with a k-mer index. All k-mers of the genome ($k = 4$) and their respective positions are represented in an index data structure. To search a query string, each k-mer the query is composed of is looked up in the index database, resulting in a list of positions for each k-mer. These *seed hits* are then used to trigger a full alignment.

.

be computationally infeasible with the given $> 10^7$ queries and a target length of 10^9. Even linear space adaptations [111] do only marginally reduce the cost. For this reasons, heuristic alignment algorithms have been developed, that find almost all approximate matches of S in G. Historically, the size of G was less problematic and the major use case for string alignment was the comparison of one short sequence against a large number of other short sequences, e.g., in homology search of protein or DNA sequences. Thus, the first heuristic algorithms were based on an efficient database representation of many target sequences, that could be used to query single substrings of S against it, only triggering a full alignment against target sequences found through the substring query [10, 227]. In RNA-Seq this task is now reversed. There, a single long target sequence G is offset by a large number of query sequences S. Motivated by this, recent algorithms aim at an efficient representation of G. We can generally distinguish two main classes of alignment strategies to solve the RNA-Seq alignment problem: *seed-and-extend* algorithms and exact matching within a *genome transformation*. We will put our main focus on explaining the first, as this technique is applied in PALMapper, which is described in Section 2.1.

Seed-and-extend Approaches This strategy is inspired by the ideas first applied in the alignment heuristics of BLAST and FASTA [10, 227]. To speed up search, only short substrings from the query, also denoted as *seeds*, are used to initially scan through a target database and only later trigger a full alignment at the respective hit locations. Following this idea, the genome sequence G is represented through an efficiently searchable database, also called a *genome index*. To build this index, each k-mer (substring of length k) of G is stored together with pointers to its positions in G. As efficient search is a central requirement, data structures such as height-balanced search trees can be utilized. Common values for k in a practical setting are 12 to 16, where larger k are better suited for increasing lengths of G. The genome index has to be built only once for each target sequence G. To query a sequence S against the index, each k-mer that is contained in S, is looked up in the genome index, resulting in a list of genome positions (cf. Figure 1.8 for schematic). This list of genome positions can than be used to trigger a full local alignment using a Smith–

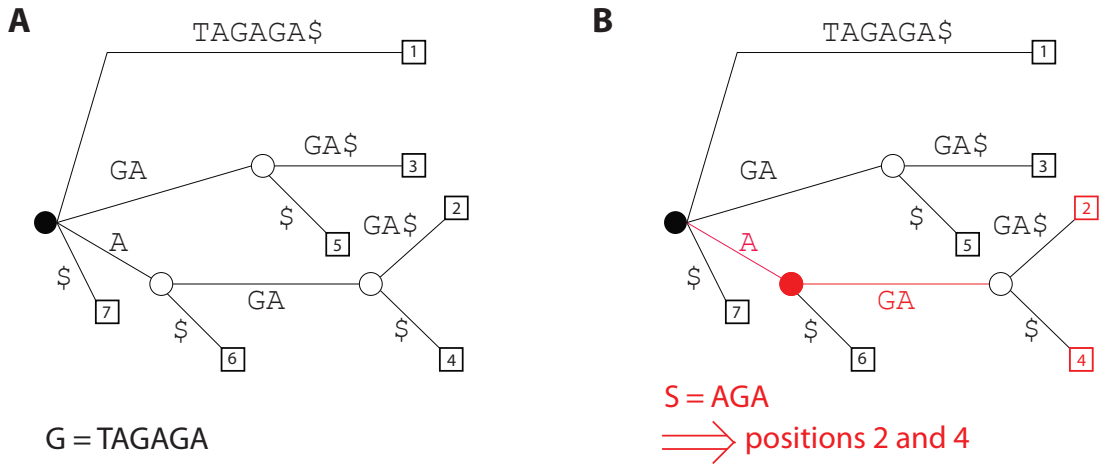


Figure 1.9: Suffix tree representation of genomic sequence. ***A:*** Tree structure for example sequence `TAGAGA`. Positions are indexed from 1 to 7, where `$` is a terminal symbol. ***B:*** Example for query procedure. The query of `AGA` results in two exact matches at positions 2 and 4.

Waterman-like algorithm, either against all genomic regions that show a seed hit or to a reduced subset of regions containing hits from multiple seeds. Seed-and-extend strategies can achieve a very high alignment sensitivity, especially if a larger number of edit operations is allowed. Examples of aligners that implement this strategy are GenomeMapper [254] and MAQ [166]. To increase sensitivity in genomic regions of low complexity, several other indexes have been proposed. Instead of k-mers of a fixed length, they use variable length k-mers, spaced k-mers including gap positions or a sparse set of k-mers. A recent comparison of these techniques can be found in [87].

Genome Transformation Approaches Instead of building an index data structure, these strategies transform G itself into a representation that can be used for efficient search of the full query sequence S. The *suffix tree* is one of the first representations suggested for this [309]. A suffix of $G = g_1 g_2 \dots g_n$ is defined as any substring G' of G with $G' = g_i \dots g_n, 1 \leq i \leq n$. A suffix tree is a tree-structured graph, that has n leaves, each labeled with the start position of one suffix. The concatenated edge labels from the root to a specific leaf represent the full suffix-sequence corresponding to that leaf. Suffixes that are themselves suffixes to other suffixes, share edges in the tree. The structure is best understood from an example (cf. Figure 1.9, Panel A). With the suffix-tree given, all exact matches of the query S in G can be obtained from a simple tree traversal in $O(m)$ time, where m is the length of S, which is a very useful property for short read alignment. For an example, see Figure 1.9, Panel B. Different algorithms have been proposed to efficiently construct suffix trees in $O(n)$ [197, 294].

A different representation has been suggested by Manber and Myers [184], the *suffix array*, which is inherently related to the suffix tree and can be generated as the depth-first pre-order traversal of the suffix tree. The suffix array is the list of all suffix positions in sort-order of the suffixes. It can be more efficiently constructed than a suffix tree for larger alphabet sizes and allows for exact search in $O(m + \log n)$ [184]. It has been shown that any

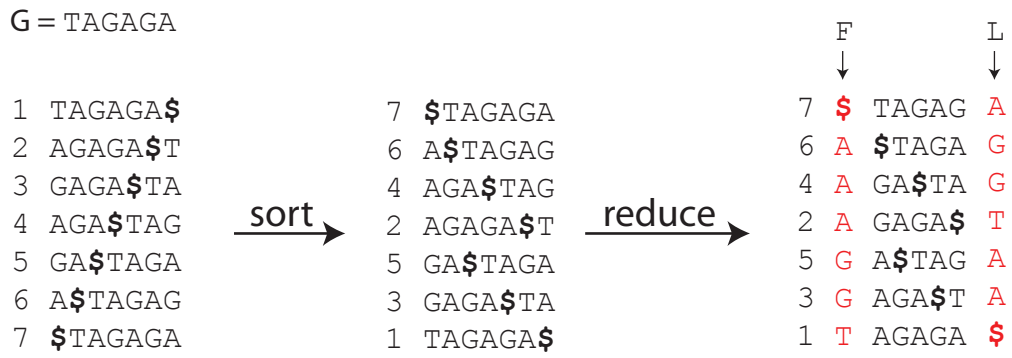


Figure 1.10: Burrows–Wheeler transform of genome string G. The leftmost block shows the array of all rotations of G. Shown in the middle is the lexicographically sorted array of rotations. On the right, the reduction to the BWT L is shown. The first column F, is equivalent to the suffix array of G. The LF-mapping of the FM-index described in [82] is used to implicitly infer F from L which is used for efficient string search.

algorithm using suffix trees can use a suffix array instead, while retaining the same time complexity [3].

The third transformation is based on a technique for string compression originally suggested by Burrows and Wheeler [39]. We define a *rotation* of $G = g_0 \dots g_{n-1}$ by k positions as the string $G_k = g_{f(k,0)} g_{f(k,1)} \dots g_{f(k,n-1)}$, with $f(k, j) = ((n - k + j) \mod n)$. The Burrows–Wheeler-Transform (BWT) L of G is then defined as the sequence of the last positions in the array of all lexicographically sorted rotations of G. For better illustration, we provide an example in Figure 1.10. The transformed sequence L has the property to contain long continuous stretches of the same symbol, which makes it efficiently compressible. Further, it can be shown, that the sequence L as well as the index of G in the array of rotations are sufficient to fully reconstruct G from L [39]. The BWT of G can now be used to construct a compressed index L from G that has the search properties of a suffix array, allowing for search in $O(m + \log n)$. However, due to the compression of L, the data structure requires significantly less space than the original genome sequence G. This representation has been named *Full-text Minute-space index* or *FM-index*. For an in-depth explanation, we refer to the original publication [82].

Numerous algorithms for RNA-Seq alignments use transformation-based genome representations, including Bowtie [154] and BWA [167] that are based on the BWT or STAR [66] and vmatch [2] which utilize suffix arrays. Since the algorithms for pattern matching in a suffix array were designed for exact search, several adaptations have been made, to also allow for alignments containing gaps or mismatches [153, 165].

The numerous different alignment strategies presented in this section were all developed to solve essentially the same task. However, as most of the algorithms are heuristics, their results can differ considerably. We will discuss this issue in depth in Sections 2.2 and 2.3.

1.4.2 Modifications for RNA-Sequencing Alignments

The alignment of RNA-Seq data requires some adaptations to the strategies for DNA-alignments. As the sequenced mRNA is deprived of intron sequences, additional long gaps have to be considered during alignment. This gaps have certain constraints in their start and stop positions, as conserved consensus sequences at donor and acceptor (cf. Section 1.1.1) that can be utilized to identify the correct alignment [59]. Further, a different scoring scheme than for normal gaps should be used, as introns can have a length of up to several hundred kilobases.

Seed-and-extend approaches can easily adapt the local alignment step triggered from seed hits to allow for split alignments over splice junctions. PALMapper [121], the algorithm discussed in this work, uses the clusters of seed hits from a read query to form seed regions, that are then extended in a banded version of the Smith-Waterman alignment that has been presented in [46]. The scoring function that computes the cost of an alignment gap not only takes length but also sequence context or a given set of splice site prediction scores into account [59].

An adaptation of the BWT based aligner Bowtie [154] for the split-alignment of RNA-Seq data is TopHat [287]. This two-step strategy performs a normal un-spliced alignment in the first round resulting in a set of genomic regions that are covered with at least one read. In a second step, these *coverage islands* are tried to be connected, thereby identifying possible splice junctions. The set of reads not aligned in the first step is then aligned against this junction regions by means of a seed-and-extend approach.

Another interesting variety is the software STAR [66], that also employs a two-step approach but does not originate from a DNA aligner. It forms a hybrid approach of suffix array search and a seed-and-extend strategy. In its first step, it finds exact matches for all maximum mappable prefixes (MMP) of the reads using an uncompressed suffix array. These MMPs serve as seeds for the second phase and naturally stop at splice junctions, providing accurate junction information. This includes also non-canonical splice junctions deviating from the common consensus sequences. In the second step, the seed-hits are stitched together using a dynamic programming approach.

In the past years, a large variety of alignment algorithms for RNA-Seq data have been developed and improved [17, 38, 66, 121, 129, 137, 170, 188, 287, 304]. However, many methods differ substantially in their outputs. Especially the detection of novel junctions, the correct placement of reads that can map to multiple regions as well as special cases like reads originating from gene fusions or other genetic re-arrangements are still challenging. We will discuss these differences in detail in Section 2.2.

2 Methods for RNA-Sequencing Data Analysis

This chapter focuses on the methodological contributions we made to the field of computational transcriptome analysis, specifically to the analysis of RNA-Seq data. The sections are ordered according to the flow of data within a typical transcriptome analysis pipeline. We begin with the alignment step as one of the most crucial parts in RNA-Seq data analysis and describe an extension to PALMapper [121] that enables the alignment to several similar reference genomes simultaneously. As PALMapper is only one of many methods for aligning RNA-Seq data, we use the second section to discuss various evaluation metrics and the results of extensive comparisons between different alignment methods. From these evaluations we have learned, that a thorough alignment filtering as post-processing is key to accurate and comparable results from different analysis pipelines. Based on this we developed SAFT, a tool to optimize this post-processing, which we describe in the third section. A second important part of alignment post-processing is the handling of ambiguous read alignments. Our solution to this problem, the tool MMR, is discussed in section four of this chapter. We then use the final section to describe SplAdder, a method using the alignments for the extraction and quantification of alternative splicing events. This and most previously described methods have been used for the research described in Chapter 3.

Author Contributions Several projects described in this work have been carried out as collaborations or in a team of developers. Here, we lay out which parts were genuinely contributed by the author of this work. The alignment software project PALMapper is a long-standing group effort developed by various contributors in the Rätsch laboratory. Especially the computational framework for variation-aware alignments was a collaborative effort with Gunnar Rätsch and Geraldine Jean. The author's contributions to this were the standardized alignment output, most recent improvements in the correct IUPAC encoding of variant strings, the combinatorial remapping of junction combinations for read alignments as well as numerous data-simulation and evaluation routines for the constant assessment and improvement of alignment and variant-alignment performance. The evaluation suite for read alignments was conceived, developed and implemented by the author, with valuable input from Gunnar Rätsch. Ideas for the stratum-wise analysis of ambiguous read mappers were contributed by Paolo Ribeca. SAFT, the tool for optimal alignment filtering was designed, implemented and tested by the author. MMR, the software to resolve ambiguous read mappings, was designed and implemented by the author. The author further developed the simulated test data set and carried out all analyses. The software to detect and quantify alternative splicing events from RNA-Seq data was inspired by a similar approach for EST data implemented by Cheng Soon Ong and Gunnar Rätsch. The author re-implemented the algorithm for high-throughput use on RNA-Seq data, improved sensitivity and specificity of the graph augmentation procedure, implemented the event quantification and visualization

routines and adapted the tool rDiff for directional testing on the count data. For easier portability and improved running time, the author developed a Python version for the software that was previously implemented in Matlab.

2.1 Variation-aware RNA-Seq Alignments

We already discussed that the alignment of RNA-Seq reads to a given reference genome, is one of the first and also most important steps in RNA-Seq based transcriptome analysis pipelines. Biases introduced during the mapping process will affect all subsequent analyses and data lost during this step will be unavailable afterwards. Using an alignment approach with optimal sensitivity and specificity regarding the placement of query reads within a reference genome is therefore key to all proper analysis procedures. To achieve this, algorithms that are both accurate and efficient are needed. As a consequence of the advancements in high-throughput sequencing technologies (Section 1.2.2), also the basic methods for alignment (Sections 1.4.1 and 1.4.2) were subject to numerous improvements. These included optimizations of alignment speed and memory footprint to allow for efficient processing of the ever growing sequencing samples as well as increases in sensitivity while lowering or at least not increasing the rate of false positive alignments. Also PALMapper [121], the alignment tool discussed in this section, quickly evolved with the needs. The main parts of PALMapper were originally published as two independent tools: GenomeMapper [254], for non-gapped sequence alignment, and QPALMA [59], for the spliced alignment of RNA-Seq data. Both tools were integrated into an improved combination: PALMapper [121]. As GenomeMapper before, also PALMapper uses an efficient k-mer index for seeding and an adapted, locally banded Smith–Waterman alignment for the sensitive full alignment of spliced reads (for details see introductory Section 1.4, following Section 2.1.2 and the individual publications). In this section, we describe the most recent improvements of the PALMapper algorithm, to take a set of complex variants into account during alignment to a given reference sequence. We begin by motivating the need for such an extension and then provide a quick overview on the alignment principle behind PALMapper. We then describe how variant sequences from different sources can be collected and merged into a variant set that can then be used for the alignment process. Following this, we outline the graph alignment approach that integrates the variant set into the target sequence of the alignment and discuss how we deal with combinations of overlapping variants. In the subsequent section, we discuss how splice junction combinations are handled in the context of a junction remapping strategy. Lastly, we conclude this section by providing a performance evaluation on simulated as well as real biological data and giving a short description of implementation and software.

2.1.1 Motivation

For human as well as many model organisms, reference genomes have been assembled over the past 15 years (e.g., [4, 152, 298, 308]). Although the number of represented species is ever growing, a central problem remains: even between individuals of the same species there exists a substantial amount of genome-sequence variation. These variations can range from single nucleotide changes, so called SNPs, to long stretches of sequence alterations, such as long inversions, insertions or deletions [83, 144, 247, 292]. Whereas intra-species

differences based on natural evolution are mostly very short and have a rate of $\sim 0.05\%$ between unrelated individuals [1], differences that are based on changes due to diseases can be drastically larger both in size and rate [157]. Especially the genome of cancer patients can be substantially different from the reference genome used for RNA-Seq alignment. If not accounted for properly, this could leave functionally relevant sequence deviations undetected or introduce systematic biases during alignment, possibly lowering the chances for novel therapeutic insights.

Another need for variation-aware alignment is based on the still limited number of available reference genomes. For many species, only a single representative of the genus or the whole family has been sequenced. If an evolutionary distant genome is used as alignment reference instead of the genome of the respective organism, this also can lead to a substantial amount of differences between RNA-Seq dataset and the target genome. How such variation can be detected is described in Section 2.1.3.

In both examples mentioned, the RNA-Seq data originates from a genome that is substantially different from the reference genome used for alignment. This can drastically compromise alignment sensitivity or even lead to false-positive matches. Depending on the degrees of freedom for the alignment, i.e., the number of allowed edit operations, especially regions that contain many unexpected variants can show a lower coverage, as they reduce the number of edit operations available to cover sequencing errors. However, especially these regions are most interesting in both usecases presented above, e.g., to identify transcripts with novel functions caused by a genomic sequence change or to analyze allele-specific expression. To align these variable regions, one can either generally increase the degrees of freedom, i.e., globally allow for more edit operations in the alignment, or take only specific variants into account. This can be realized by allowing additional edit operations only in a pre-defined set of locations, thus not suffering from the burden of false positives through a globally higher level of mismatches. We have modified the PALMapper algorithm to allow additional degrees of freedom at specific variant locations and thus can solve the problem of lower coverage in regions of high sequence variability, while not increasing the number of false positive alignments in general. To our knowledge, no other alignment program is able to take variation into account to an extent exceeding single-nucleotide variants. PALMapper can be used for variants of any length and any grade of complex combinations.

2.1.2 Alignment Principle

PALMapper is a typical seed-end-extend aligner. As described in Section 1.4, the genome is first indexed by storing the location of each k-mer in a data structure for efficient look-up. For alignment, the query read is then also split into k-mers that are queried to the index, resulting in a set of match locations, so called *seed-hits*. These seed-hits are then clustered into long and short hit regions that trigger full local alignments (depending on length and distance of the regions). All available information about the seed-region is integrated to form a pseudochromosome sequence that is used as target for the local alignment (cf. Section 2.1.4 and Figure 2.2).

Table 2.1: List of studies that catalog genetic variation for different organisms. Sample sizes are strains (* denotes individuals).

Study	Organism	Sample Size	SNVs	Indels	Reference
1000 Genome Project	*H. sapiens*	1,092*	38,000,000	1,380,000	[11]
Mouse Genome Project	*M. musculus*	17	129,260,574	21,683,297	[133]
19 Genomes Project	*A. thaliana*	19	3,070,000	1,200,000	[92]
Panzea	*Z. mays*	103	55,061,920	3,200,000	[47]
Million Mutation Project	*C. elegans*	40	630,000	220,000	[284]
Rat Genome Project	*R. norvegicus*	8	7,200,000	633,000	[24]

2.1.3 Variant Detection and Integration

We describe two general strategies for the acquisition of a variant set that can be used for RNA-Seq alignment with PALMapper. The first strategy is to rely on variant sets for a given organism, that have been produced by previous studies. For many organisms of common interest, e.g., human, mouse, or maize, variant sets are publicly available. We have collected a list of examples in Table 2.1. Most studies provide the variants in a standardized format, e.g., the Variant Call Format (VCF) [58]. To be able to use such existing data sources, PALMapper is able to read variant data in most of the common formats and to convert it into an internal binary representation, that can then be used during alignment.

The second strategy is to use PALMapper in detection mode. That is, in a first round of alignment, usually with a high number of degrees of freedom, PALMapper sensitively aligns the reads to the reference genome and records all differences between the reads and the genome used. That is all edit-operations used for the alignment of a read are recorded as possible variant. Depending of user-defined thresholds, different criteria are used to filter confident variants, e.g., variants confirmed by multiple reads. Further, a genome map of covered positions is recorded that can be later used for filtering. One problem arising from this strategy is that biological variation and differences through sequencing noise are intermixed. To allow for further filtering and only retain a set of high-confidence variants, we suggest to repeat the variant-calling for several replicates and to integrate the results, retaining only variants independently found in more than one replicate. A variant that is highly replicable over several samples can usually be considered as a confident call. However, systematic errors reproducing in all replicates could still lead to false calls. We provide the following filtering criteria to minimize the number of false positive detected variants:

- **Minimal number of reads confirming the variant** More evidence from independent read alignments increases the confidence.
- **Minimal number of samples confirming this variant** An independent observation of the variant in multiple samples increases the confidence.
- **Maximal distance to a covered genome position** A variant very far from any expressed genomic region is more likely to be considered noise and might have less relevance for RNA-Seq alignments.
- **Maximal length of the variant** Insertions or deletions exceeding a certain length are very difficult to call accurately on any platform, as they often include repeat-variants of low complexity. We thus consider shorter variants as more confident.

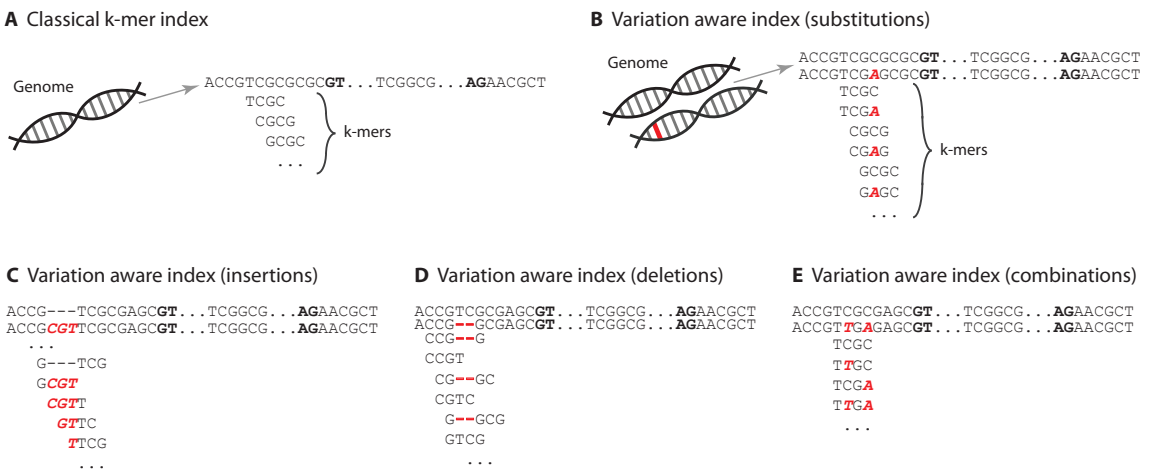


Figure 2.1: Overview on how different variants are considered during index building. Variant sequences shown in red. ***A:*** Classical k-mer index. ***B:*** Index including single nucleotide substitutions. ***C:*** Index including insertions. ***D:*** Index including deletions. ***E:*** Index including combination of variants.

After testing in many applications, we found these criteria to be most effective. The optimal choice depends on factors such as number of replicates, sequencing error rate or expected amount of variation and needs to be adapted for the specific purpose.

2.1.4 Variation Aware Index and Graph Alignment

Following the seed-and-extend alignment paradigm described earlier, we build a k-mer index from the genome, where we use an associative array to efficiently store all genomic locations for each sequence of length k present in the genome (cf. Figure 2.1, Panel A). Using an associative array is a compromise between running time and memory consumption. Even if efficiently stored, the memory footprint of the full index would grow in $O(|\Sigma|^k)$, where $|\Sigma|$ is the size of the alphabet and k the k-mer length. Taking into account both strands of the genome as well as some meta-information would result in an index size of approximately 150 MB for $k = 12$ and 32 GB for $k = 16$. As it is very unlikely that all possible k-mers occur in the genome, we do not store the full index. By using an associative array and a hash-function on the k-mer sequence to efficiently store only k-mers that have been observed, we can reduce index size to several 100 MB (*A. thaliana*) to 12–15 GB (human), depending on genome length and complexity. In the following, we will describe how the indexing step is adapted to take genome-variants into account.

Variation Aware Indexing To allow for additional variation, we extend the index with additional k-mers that are induced by the variants. For instance, given the genomic k-mer $g_i g_{i+1} \dots g_{i+k-1}$, ranging from genomic positions g_i to g_{i+k-1}, the single nucleotide variant g'_{i+j}, with $0 \leq j < k$, induces an additional k-mer $g_i \dots g'_{i+j} \dots g_{i+k-1}$. This affects all k-mers spanning the genomic position $i + j$ (Figure 2.1, Panel B). The same principle is applied to other variants like insertions or deletions. Assume the k-mer to be given as

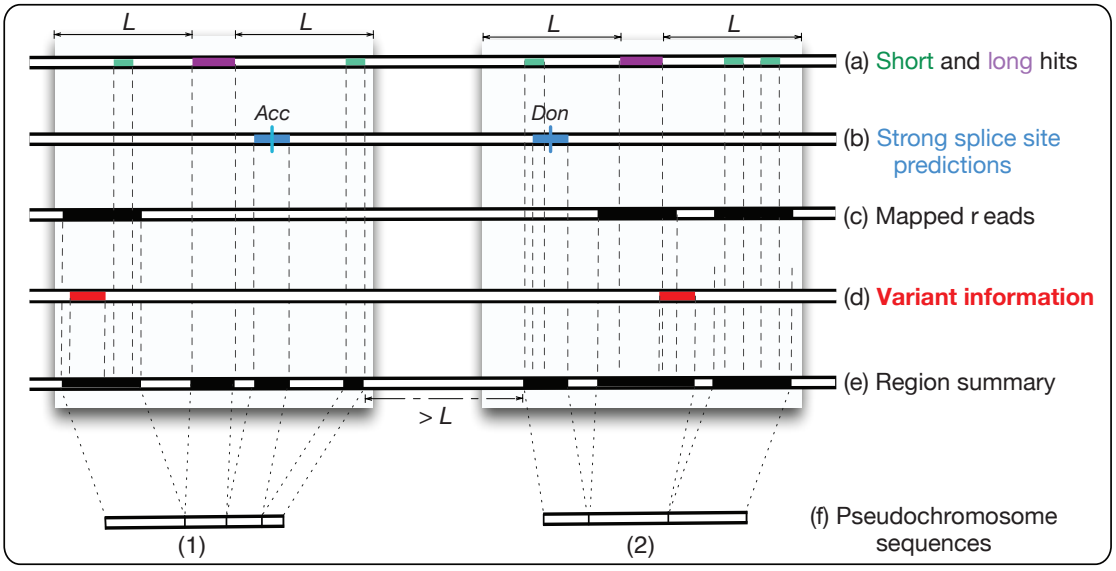


Figure 2.2: Projection to pseudochromosomes. All information available for the region around the seed-hits is projected onto the same coordinate system. This includes (a) the short and long seed-hits triggering the full alignment, (b) genomic splice-site annotations that provide information to place introns, (c) information about already mapped reads that emphasizes expressed genomic regions and (d) variant information augmenting the genomic background sequence. All tracks are collapsed into a region summary (e) and then projected to pseudochromosome sequences (f), which are then used for the full local alignment. The minimal distance between two seed-regions is denoted as L.

before, then the deletion of genomic positions $g_{i+j} \dots g_{i+h}$, with $0 < j < h$, results in the additional k-mer $g_i \dots g_{i+j-1} g_{i+h+1} \dots g_{i+k+h-j}$. The definition for insertions is analog (cf. Figure 2.1, Panels C and D). In case of possible variant combinations, we have to consider all possible subsets of variants within a genomic window of length k (Figure 2.1, Panel E). Since the number of all possible subsets grows exponentially with the number of variants considered, the number of allowed combinations is limited in practical uses.

Graph Alignment to Pseudochromosomes After generating seed-hits with the variation-aware index structure, also the local alignments need to take the given variants into account. As described earlier, several seed-hits that co-localize within a distance L are concatenated to form seed-regions (Figure 2.2, (a)). Triggered by seed-regions of sufficient length, PALMapper builds a pseudochromosome region around the hit region (Figure 2.2). To this end, all available information for that region is projected onto a common coordinate system, forming the pseudochromosome sequence. This information includes the seed-hits themselves, splice-site locations and strengths, evidence of previously mapped reads and the variant information.

The local alignment to the generated pseudochromosome sequences follows a modified Smith–Waterman alignment algorithm [267]. These modifications include an adapted scoring model for substitutions in the context of their error probabilities as well as the proposal

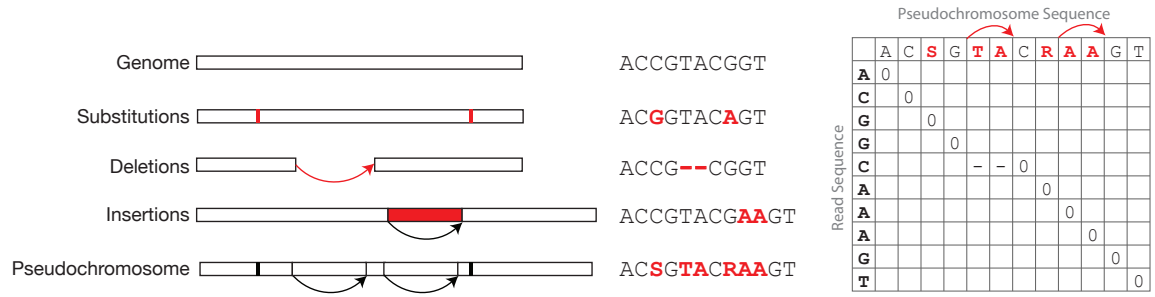


Figure 2.3: Variation-aware local alignment. The left panel shows the four different types of variants recognized by PALMapper and how they are integrated into a pseudochromosome sequence. The middle panel shows the sequence representation for each variant. On the right, the local alignment matrix is shown for aligning a read sequence (leftmost column) to a pseudochromosome sequence (topmost row). The optimal alignment between the two sequences is shown as a sequence of zeros. Variant positions are indicated in bold red (substitutions) or as arrows (insertions/deletions).

for a special treatment of deletions that arise from intronic sequences. All modifications have been described in context of the QPALMA-publication [59] and will not be further discussed in this context. Here, we will focus on the description of the most recent improvements that allow for an efficient consideration of sequence variants. Generally, we distinguish three basic cases of variants, which we will discuss in detail in the following text. A schematic visualization of all cases is provided in Figure 2.3.

First, we describe single-base substitutions. To only allow for certain substitutions at a defined set of genomic positions, the base code at such a variant position in the genome is substituted by the IUPAC representation of ambiguous bases [52], e.g., an ambiguity between A and G is denoted by R. To consider this ambiguity during alignment, we introduce an augmented substitution matrix that does not penalize a substitution of a base if it is contained within the IUPAC-ambiguous set at the genome position. That is, to a given genomic base with variant R both A and G could be aligned at no cost, but not C and T.

Second, for the representation of deletions in the genome, we allow for alignment gaps at the respective genome position at no cost. Conceptually, deletions behave the same as introns that are known in advance. To this end, we add an additional possibility to introduce a gap within the dynamic program of the Smith–Waterman alignment [59], whenever a deletion-variant position occurs. We realize this by adding an additional line to the recurrence defined in section 2.2.3 of [59], that allows certain deletions to be treated differently than a normal gap. This is done as follows. Assume that we store all gaps as a set of pairs G, where each pair consists of start- and end-position of an allowed gap, and that the recurrence $V(i,j)$ describes the cumulative alignment cost up to positions i and j. We can then augment the recurrence to:

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + M(S_E(i), Q_E(i), S_D(j)) \\ V(i-1,j) + M(S_E(i), Q_E(i), \text{'-'}) \\ V(i,j-1) + M(\text{'-'}, \cdot, S_D(j)) \\ W(i,j-1) + \hat{f}_{\text{acc}}(j-1) \\ V(i-1,j-1-k) + M(S_E(i), Q_E(i), S_D(j)), \quad \text{if } (j-1-k, j-1) \in G \end{cases}$$

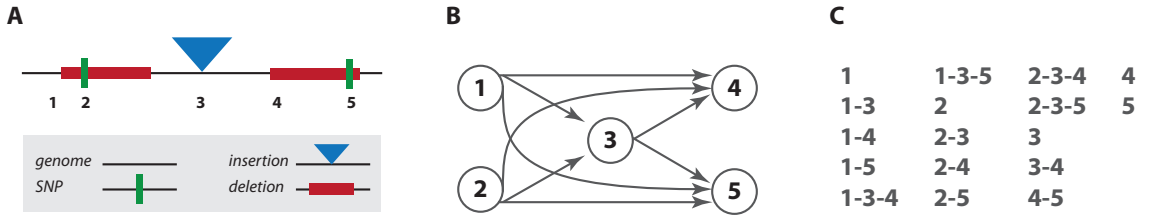


Figure 2.4: Forming all possible variant combinations. ***A:*** Example structure of variants on the genome, where numbers identify single variants. ***B:*** Variant combination graph. Compatible variants are connected by edges. Node labels follow numbering in A. ***C:*** All variant combinations that can be extracted from the graph.

where i and j are positions in the read S_E and the genomic sequence S_D, respectively, Q_E is the sequence of quality values for the read string, $M(\cdot,\cdot,\cdot)$ is the alignment cost function, '-' describes a gap position, $W(i,j)$ is the recurrence matrix for introns between positions i and j and $\hat{f}_{acc}$ is a scoring function for the splice-site acceptor. The alignment cost function maps a tuple consisting of the character at read position i, the read quality at position i and the genome character at position j to a real valued number, describing the alignment cost. For a more in-depth explanation we refer to [59].

Third, insertion-variants are handled as a combination of insertion-operation to the genome and addition of an unpenalized gap. The inserted sequence is fully integrated into the pseudochromosome sequence, whereas the original sequence without insertion can be constructed by allowing for an unpenalized gap (cf. Figure 2.3). In the dynamic program, this gap is handled in the same way as the deletion-case above.

For all cases discussed so far, it was possible to integrate each variant-type into the existing dynamic program through moderate adaptations to the algorithm. However, the most involved part is the combination of several overlapping variants. In this case, we form all possible combinations of compatible variants. Assume the two variants v_1, starting at position i_1 and ending at position j_1, and v_2, starting at position i_2 and ending at position j_2, to be given. We denote v_1 and v_2 as *compatible* if they do not overlap, which is the case if $j_1 < i_2$ or $j_2 < i_1$. To generate the set of possible variant combinations, we take all variants in proximity to the seed-region and represent them as a graph. Each variant forms a node and two nodes are connected by a directed edge from node v_1 to node v_2 if the respective variants are compatible and the start position of v_1 is smaller than the start position of v_2 (cf. Figure 2.4). If a path can begin and end at any node, the paths through this graph form all possible variant combinations. As the number of paths p computes in the worst case as $|p| = \sum_{i=0}^{n-1} 2^i$ for a fully connected acyclic directed graph with n nodes, we limit the number of combined variants to at most 3 in practical applications. An example case is shown in Figure 2.4.

2.1.5 Re-Alignment to Combinations of Known Splice Junctions

A comprehensive evaluation of spliced alignment algorithms, later described in Section 2.2, has shown an increased accuracy of spliced alignments if the introns were confirmed by the alignments of several reads. Motivated by this observation, we added a junction re-

mapping step to the PALMapper workflow to increase the accuracy of spliced alignments. We implemented this by taking a list of splice junctions into account during extension of a seed-hit into a full local alignment. The process works as follows. Assume a list of junctions J is provided, where each junction is represented as a pair of start- and end-position. Given a seed-hit spanning genomic positions $g_S, \ldots, g_E$ and a read r of length k that contains the seed-hit at positions $r_s, \ldots, r_e$, we need to find all junctions in J that could be contained within an alignment of the remaining parts of the read both on the left hand side and the right hand side of the seed, $r_1, \ldots, r_{s-1}$ and $r_{e+1}, \ldots, r_k$, respectively. Once we determined such a combination of junctions, we can assemble the corresponding pseudochromosome sequence taking the junctions into account and can compute a local unspliced alignment between the full read and that sequence. As the algorithm uses the same strategy for handling the left hand and right hand parts, we will use only the left part for explanation. The procedure for the right part follows analogously. At first, the algorithm determines all junctions that end in the genomic region $g_{S-s+1}, \ldots, g_{S-1}$. Each of these junctions forms a valid combination. Additional combinations can now be built by adding more junctions to existing combinations. Assume, we want to augment the combination that contains junction j^3 that spans genomic positions $g_{j^3_s}, \ldots, g_{j^3_e}$, with $g_{S-s} < g_{j^3_e} < g_S$ (naming of junctions follows the example in Figure 2.5, Panel B). If the length of the left remainder of the read was $s-1$, taking the junction into account reduces it to $s - 1 - (g_S - g_{j^3_e} - 1) = s - (g_S - g_{j^3_e})$. For simplicity of notation, we denote this remaining length as m. Only if m is greater than 0, we can add more junctions to this combination. If we can add more junctions, we determine all junctions that end in the genomic region $g_{j^3_e - m+1}, \ldots, g_{j^3_e - 1}$ and form new combinations with j^3, creating combinations containing two junctions each. When all single junctions have been checked for augmentation, the newly added combinations of two junctions are tested to be augmented. This process is repeated until no junction added in the last round can be further extended. This procedure results in the complete set of all junction combinations compatible with the left remaining part of the read. The combinations for the right remainder of the read are computed analogously. In a final step, all possible combinations of junction combinations from the left and the right part are computed, resulting in the complete list of junction combinations around the seed hit. As the number of combinations grows exponentially with the number of junctions, we limit the size of combinations to at most three per side in practice. An example for two possible junction combinations is provided in Figure 2.5, in Panels B and C.

2.1.6 Results and Evaluation

Our evaluation specifically covers the most important addition described in this thesis: the variation-aware alignment. We will not discuss the general performance of PALMapper in comparison to other alignment approaches, but rather focus on improvements within the specific setting that RNA-Seq data and reference genome show substantial sequence differences. For the evaluation, we considered two different datasets: an artificial dataset of simulated data where we had full control over the read-generation as well as a biological dataset, produced from two related subspecies of *A. thaliana*, to evaluate the performance in the context of natural variation. We begin by describing the evaluation on the simulated data and subsequently discuss the biological dataset.

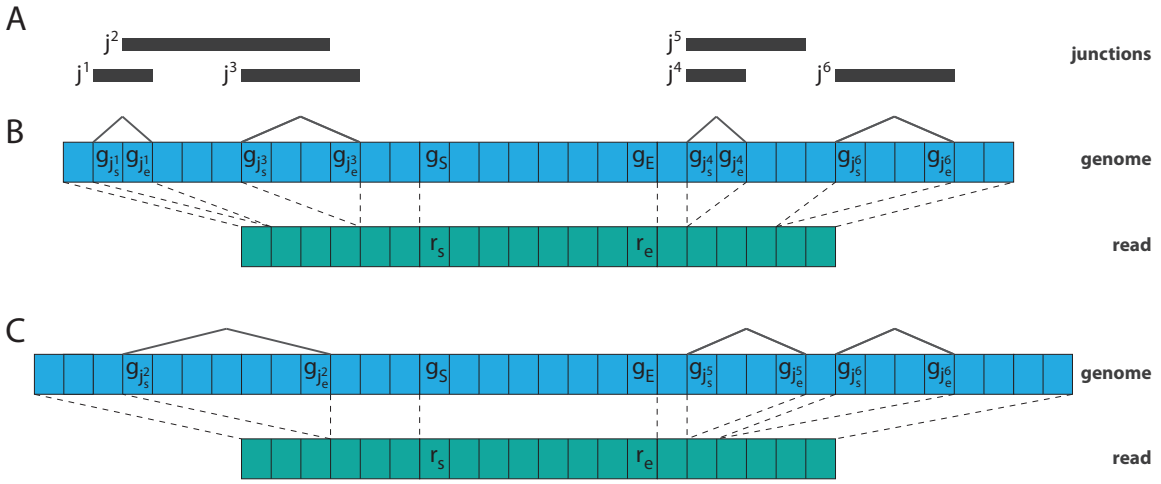


Figure 2.5: Identifying junction combinations for junction remapping. ***A:*** List of available junctions. ***B:*** Possible combination of junctions. ***C:*** Alternative combination of junctions. Genome sequence is blue, reads are green. Junction spans are indicated as dark gray solid lines.

Evaluation on Simulated Data General aim of this evaluation was to measure how much single nucleotide differences between RNA-Seq source and reference sequence influence the alignment performance and to quantify the improvement when variation-aware alignment was used. To answer these questions, we constructed an RNA-Seq dataset originating from a heterozygous genome, generating the same number of reads from each haplotype. In consequence, when the reads are aligned back to the genome in an optimal way, heterozygous positions should show no difference in read coverage. Any measurable deviation for one the two alleles would be due to the alignment procedure. To generate such a set of reads, we randomly chose 5,000 genes from the TAIR10 genome annotation for *A. thaliana* and used the FluxSimulator [102] (version 1.1.1-20121103021450) to sample 10^7 reads of length 76 nt from these genes. We chose the default error model and selected a normal distribution with mean 300 and standard deviation 50 as the insert size distribution. A list of all simulation parameters is provided in Appendix A.1. The read set was then duplicated into two identical read sets, simulating the contribution of two parents. One of the two read sets was then mutated with a uniform mutation rate of 10^{-4} to randomly introduce single base substitutions, thus generating heterozygous positions present in the read set but not in the reference genome. Given an estimated substitution rate of 1 mutation per genome per generation [183] and a generation time of $\approx$ 5 weeks for *Arabidopsis*, the last common ancestor of the two simulated individuals was 500 years ago. In total, we altered 2,951 positions in the sequence of the 5,000 genes. As we mutated the read sets and not the source genome, we expect no biases from statistical fluctuations due to expression model of FluxSimulator.

The two read sets were then merged and used in two different alignment settings. In the first setting, we aligned the reads to the same genome they were originally sampled from. In the second setting, we used the variation-aware alignment to the same genome, taking the list of altered positions into account. As the dataset was artificially constructed from the same parent, only the artificially introduced variant positions were heterozygous, each with the same allele frequency of 0.5.

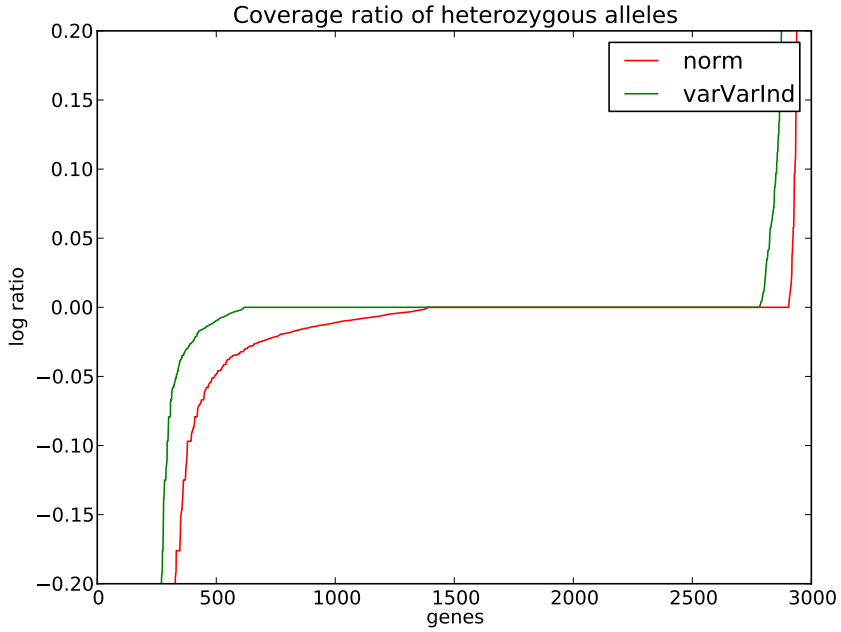


Figure 2.6: Comparison of allele-specific alignment performance at a set of artificial variant locations. Shown is the log-ratio of the two alleles at all simulated heterozygous loci for the alignments with (green, varVarInd) and without (red, norm) the variation-aware extension. The optimal alignment set would show no deviation from zero for any gene.

To assess, how well the alignment was able to reconstruct the allele frequencies at variant positions, we computed the log-ratio of the number of reads carrying one allele over the number of reads carrying the other. This should result in a value of 0, if both alleles occurred at the same frequency and a value above or below zero, if the first or second allele were overrepresented, respectively. The variant-aware alignment showed a substantially larger amount of variant positions that had the same frequency of alleles than the alignment without variant information. A diagram of the results can be found in Figure 2.6.

Evaluation on Biological Data For assessing alignment sensitivity in a biological setting of variation-aware alignment, we used RNA-Seq data that has been published in earlier work [92]. The two ecotypes of *A. thaliana* Col-0 (originating from Columbia, USA) and Can-0 (originating from the Canary Isles, Spain) were two of the evolutionary most-distant sub-species analyzed in [92] and showed a substantial amount of sequence variation between their genomes. To test for the effect of the variation-aware extension on alignment sensitivity, we aligned RNA-Seq reads originating from Can-0 to the Can-0 genome, the Col-0 genome and the Col-0 genome with additional information about the sequence variation. The original data was split into 23 chunks of 250,000 reads each, using the UNIX split command. All chunks were then aligned independently. Even without the variation-aware extension PALMapper shows a higher sensitivity than comparable state of the art tools (TopHat [287]; TH_CA and TH_CO in Figure2.7) and has an even increased performance when using the variation-aware index (Figure 2.7, PM_COvi) and the variation-aware local alignment (PM_COv). The fully variation-aware alignment using improved index and local

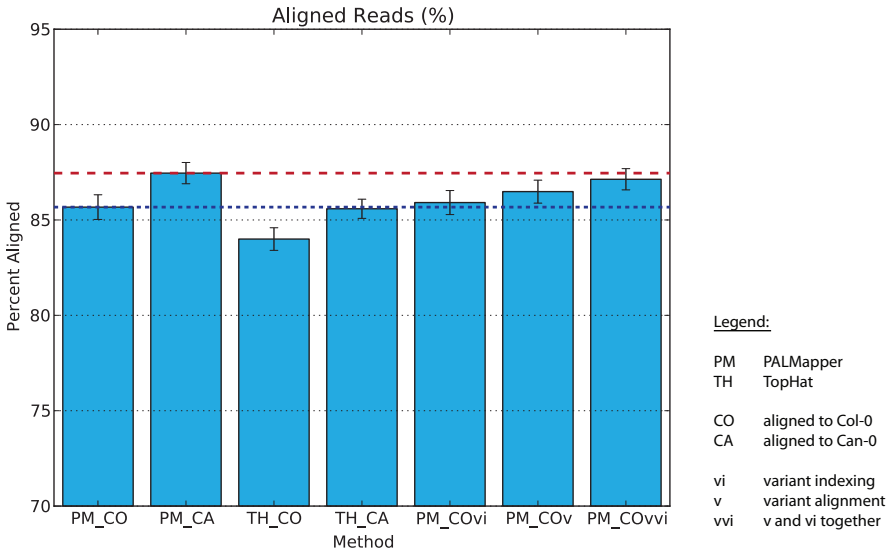


Figure 2.7: Sensitivity of variation-aware alignments on a biological dataset. From left to right, the bars show the percent of aligned Can-0 reads for 7 different alignment settings: PALMapper alignment to Col-0, PALMapper alignment to Can-0, TopHat alignment to Col-0, TopHat alignment to Can-0, PALMapper alignment to Col-0 plus variant aware index, PALMapper variant aware alignment to Col-0, PALMapper variation-aware alignment to Col-0 plus variant aware index. The red dashed line shows that the fully variation-aware alignment (rightmost bar) is almost as sensitive as the alignment to the Can-0 genome (second bar from left). Error-bars indicate the standard error of the mean over replicates of 23 read chunks with 250,000 reads per chunk.

alignment (PM_COvvi), shows almost the same sensitivity as the alignment to the original Can-0 genome (PM_CA). As discussed above, this additional sensitivity is mainly caused through alignments over regions in the genome that show variability in the reference. Although a sensitivity improvement of 2% seems only moderate, it can be essential for the analysis of allele-specific expression or in the context of genome wide association studies, where the link between expression differences and the genetic background is investigated.

2.1.7 Implementation and Software

PALMapper is implemented in C++ (C++11 standard) and uses only standard libraries. The alignment process can be parallelized by using multiple compute threads at the same time that all use a shared memory for representation of the genome index. The source code is publicly available at `https://github.com/ratschlab/palmapper`. The code contains the extensions described in the text above as well as many elements from other contributors. PALMapper is published under the GPL3 license. An overview of the user interface is provided in Appendix A.1.

2.2 Evaluation of RNA-Seq Alignments

Numerous algorithms have been developed to generate alignments of RNA-Seq reads to a reference genome, most of them implementing different flavors of the common alignment

strategies presented in Section 1.4. Although solving the same task, these algorithms are mainly based on heuristic assumptions, resulting only in approximate solutions and thus producing a wide range of different outcomes. To be able to make an informed decision which approach to use and to correctly interpret its results, it is inevitable for users and developers to be able to evaluate and compare the different outcomes. In this section, we describe the development and implementation of a set of RNA-Seq alignment evaluation metrics. The first part discusses the key points making such an analysis relevant. Then, we briefly describe some necessary pre-processing procedures. In the third part, we describe our evaluation metrics and subsequently present different visualizations and discuss possible interpretations. Lastly, we provide a short description of the implementation.

2.2.1 Relevance

The recent improvements in high-throughput techniques for RNA-Seq have revolutionized the fields of genomics and transcriptomics, allowing for analyses of an unprecedented complexity. However, the increased data quantity and the higher throughput of shotgun sequencing methods result in two major computational challenges.

The first problem consists in the size of the input data. Numerous alignment strategies have been developed over the past years with the aim to identify the correct mapping location for each read within the reference genome sequence (see Section 1.4). However, facing the large number of sequencing reads, most alignment strategies use heuristic approaches to only identify the most likely mapping location. Although all strategies aim to solve the same problem, they make different heuristic assumptions, resulting in a wide range of possible results. This range becomes even broader, if alignment parameters and the strategies for post-processing are taken into account.

The second computational challenge is caused by the shotgun-nature of the data. Whereas traditional sequencing strategies like Sanger-sequencing (Section 1.2) produced long reads of up to 400 nt length with an error rate of 1 in 10,000, the newer high-throughput sequencing (HTS) techniques began with 25 nt reads and have only recently evolved to a length of up to 250 nt but still show error rates of up to 1%. These shorter and more noisy reads are especially problematic for RNA-Seq applications, where the sequencing sample is mostly produced from mature mRNA that has been already spliced, causing a possible segmentation of the read during alignment. The shorter the read segments become, the more ambiguity lies within the alignment result.

In the context of the RNA-Seq Genome Annotation Assessment Project (RGASP) organized by the Wellcome Trust Sanger Institute, these alignment differences became first evident to us. Using RNA-Seq data of three different organisms (the nematode *C. elegans*, the fly *D. melanogaster*, as well as *H. sapiens*), the goal of the competition was to produce a gene annotation from RNA-Seq evidence. Motivated by the wide spectrum of results, we suggested a comprehensive evaluation of all available RNA-Seq alignments that the participants had submitted. This data set was especially well suited for such a comparison, as all participants had used the same input data and reference genomes and as it represented a typical scenario for the use of RNA-Seq data.

In the following, we will describe the evaluation metrics that we developed on these datasets (further denoted as *submissions*) and discuss the insights we gained. The metrics are not limited to this evaluation and are generally applicable to any set of alignment files.

Table 2.2: List of all submissions evaluated in the alignment comparison together with the corresponding alignment approaches, including the references to the used method. Two submitters used the same method but with different parameter settings.

Submission Label	Alignment Approach	Reference
ADobin	STAR	Dobin *et al.* [66]
AMortazavi	ERANGE	Mortazavi *et al.* [206]
CIseli	SIBsim4/sim4	Florea *et al.* [86]
GRaetsch	PALMapper	Jean *et al.* [121]
LPachter	TopHat	Trapnell *et al.* [287]
MGerstein	TopHat	Trapnell *et al.* [287]
MStanke	BLAT	Kent [134]
SWhite	Exonerate	Slater and Birney [266]
TAlioto	GEM	Marco-Sola *et al.* [188]
TWu	GMAP	Wu *et al.* [318]

As the submissions were not provided in a standardized format, data conversion was a major part of our efforts discussed below. However, the software implementation described at the end of this section only covers the evaluation and overcomes most of these preprocessing steps by requiring a standardized input.

2.2.2 Input Data and Preprocessing

All input alignments used for our analysis were generated and provided by the submitters to the second round of the RGASP. As mentioned earlier, no formatting convention had been given, which made it necessary to convert all alignments into a common representation. We chose the SAM alignment format (version 0.1.2-draft[1]) [167], as it was best suited for our purposes and provided a compressed binary format representation (BAM). In case our analyses required a genome annotation or reference sequence information, we used the respective versions that were specified for the second round of the RGASP. The full description of input data and formats is provided in Appendix A.2.

The single submissions were labeled by the name of the submitter. Some submitters provided several versions of their alignment sets (filtered and unfiltered). It is further possible that different submitters used the same alignment algorithm. Table 2.2 shows a list of all submitters together with the respective alignment approach they have used.

For SAM and BAM alignment processing we used the SAMtools software package (version 0.1.7a) [167]. *In silico* transcript predictions are based on Scripture (beta) [103] and Cufflinks (version 0.9.2) [288].

As different processing pipelines were used by the submitters, various pre-processing procedures were necessary to harmonize the inputs for evaluation. Differing read-IDs were unified based on the used FASTQ input files. Further, we removed all mate-pair information from the read-ID and integrated it into the alignment flag. Chromosome names follow

[1] available at http://samtools.sourceforge.net/SAM1.pdf

the University of California Santa Cruz (UCSC) genome browser standard (complete list in Table 2.3). Most problematic was, that several submitters had miscounted the edit operations used for the alignment. To allow for a common interpretation of alignment features, we re-evaluated each alignment predicted by the submitters and re-counted all edit operations in a uniform manner, thereby not counting edit operations within the clipped alignment parts. If the alignments were too short and no clipping information was provided, the clipping was inferred during re-alignment. We further removed all unaligned and duplicated reads from the input files and sorted them by read-ID. Due to ambiguous read-IDs for the human data sample, two of the originally five provided lanes had to be excluded from the analysis.

Table 2.3: Chromosome names for the three different organism that were used for evaluation. Names are based on the UCSC naming standard.

Organism	**Chromosome Names**
C. elegans	I, II, III, IV, V, X, MtDNA
D. melanogaster	2L, 2LHet, 2R, 2RHet, 3L, 3LHet, 3R, 3RHet, 4, U, Uextra, X, XHet, YHet, dmel_mitochondrion_genome
Human	1 to 22, X, Y, MT

2.2.3 Metrics

Evaluation of general statistics To get a first overview, we evaluated each submitted alignment set with respect to the following criteria:

- *Distribution of edit operations* All edit operations were distinguished into mismatches, deletions and insertions. For each category, we computed its distribution as the average number of edit operations per position over the length of the read.
- *Distribution of split positions* Split positions are the end positions of the read segments that are implied by spliced alignments. The distribution was computed as number of split positions per position over the read length. Multiple split positions per alignment were counted individually.
- *Alignment error rate* The error rate was computed as the fraction of alignments that showed a mismatch either at an alignment position or for a certain quality value. We computed two distributions. One per read position over the length of the reads and one per quality value over the full quality range.
- *Distribution of quality values* For each alignment position, we determined its average quality value over all alignments.

All information used to compute the statistics were directly inferred from the CIGAR string (a specific alignment representation in the SAM format) or the sequence and quality strings present in the alignment files.

Agreement to the Annotation We used each alignment set to compute the agreement of its predicted intron positions to the given annotation. Agreement was measured by the F-score, which is the harmonic mean of precision (ratio of true positive introns over predicted introns) and recall (ratio of true positive introns over annotated introns). To individually optimize each submission's agreement to the annotation, we determined optimal filter settings for each submission. For this, we performed an exhaustive search over a grid of 700 different filter parameter combinations and computed the corresponding F-scores and only retained the best for comparison. For a detailed list of tested parameters we refer to Appendix A.2.

Evaluation of Ambiguous Mappers Ambiguous mappers, or multimappers, are reads that map to more than one genomic location. To increase sensitivity, we extended this definition and defined a multimapper as a read that maps to more than one genomic location measured over the union of all input submissions. Two genomic locations were considered as the same, if they shared at least one exonic position in the genome. The multimapper evaluation was based on the comparison of *alignment strata*, which are sub-groups of alignments stratified by their respective number of edit operations. To form such a stratum, we joined all alignments of a given read, if they used the same number of edit operations. We call this number the *stratum level*. Such a list of strata was generated for each submission (*submission list*) as well as for the union of all submissions (*union list*). The use of strata enables the comparison of alignment sensitivity for multimappers without a confounding effect of edit distance.

We tried three different strategies to compare the lists of strata. In each strategy, we computed a score between 0 and 1 for each read and stratum, describing its multiple alignment accuracy. The computation of the score differs, depending on how the alignments of the read were split into strata. The total score of a submission was then computed as the score over all reads. The three strategies are defined as follows:

- **Comparison per *mismatch stratum*** defines the stratum score as the fraction of alignments in a stratum of the union list, that can be explained by the alignments of the corresponding stratum in the submission list. Strata are corresponding, if they have the same stratum level. An alignment is counted as explained, when it has at least 90% overlapping exonic positions with an arbitrary alignment in the respective union list stratum.
- **Comparison per *alignment list stratum*** computes for each stratum in a submission list the fraction of alignments in the corresponding stratum of the union list. Here, an alignment counts as explained, if there exists at least one alignment in the submission stratum that overlaps the respective alignment in one of the union strata with a lower or equal level in at least 90% of exonic positions. This fraction is then assigned to the sum of lengths of the union list strata up to the current stratum. Averaged over all alignments, the score reflects how good a single submission can explain the first k alignments of all present multimappers.
- **Comparison per *weighted mismatch stratum*** computes the score similar to the comparison per mismatch stratum but in a simplified manner.

Each stratum in the submission list is scored as the fraction of identical alignments from the same stratum of the union list. Finally, each stratum is weighted with its level plus one, thus assigning strata with more edit operations a lower weight.

Pairwise intron agreement To evaluate the pairwise agreement of spliced alignments, we generated the relative intron agreement of the pairwise submissions. We therefore computed the Jaccard index of the intron agreement (ratio of intersection over union of two submission's intron lists). We further computed for each submission what fraction of its introns is shared with exactly k other submissions. Furthermore, we computed the relative fraction of a submission's intron list shared with each of the other submissions.

Effects on transcript prediction We used two different *in silico* transcript predictors to assess the downstream effects of read alignment on their results: Cufflinks [288] and Scripture [103]. To meet the input specifications of both tools, we sorted all alignments by starting position with SAMtools [167] and inferred strand information for spliced reads if necessary and possible, to provide a valid XS-Flag. For some alignments, insertions and deletions at the alignment boundaries had to be replaced by clippings, otherwise causing runtime errors. If submissions showed alignment qualities generally equal to zero, we replaced them by 255 (no quality measurement available). Due to limited computational resources, all computations were carried out on chromosomes I, 2L, and 1 for worm, fly, and human, respectively. For Scripture we used the option `-upWeightSplices` in all cases. For Cufflinks we limited the intron size to a maximum value of 20,000, 50,000, and 200,000 for worm, fly, and human, respectively. Otherwise, we used the default parameters.

2.2.4 Visualization and Interpretation

General Alignment Statistics To facilitate interpretation of the results, we produced a wide range of different visualizations. A very straightforward assessment of the alignment quality is the distribution of different alignment statistics over the read length (Figure 2.8). In agreement with previous studies [44, 106], we find a strong correlation between read position, the base calling quality of the sequencer and the number of mismatches at this read position. Especially in early versions of the Illumina sequencing machines, a strong 3'- to 5'-bias is evident, often caused by decreasing quality of the sequencing chemistry in later cycles (cf. Section 1.2; Figure 2.8, Panels B and C). Further, certain abnormalities arising during the sequencing process became evident. For instance, a higher error rate replicating at certain positions over several submissions, causing peaks in the mismatch distribution (Figure 2.8, Panel B), is a good indicator for such an abnormality. Other differences in the statistics can be attributed to the alignment algorithms themselves. Especially towards the ends of the reads, differences in the distribution of deletions and insertions (latter not shown) are common, indicating different preferences of the alignment algorithms (Figure 2.8, Panel A).

Accuracy of Intron Prediction In the following, we describe several ways to assess the accuracy of splice junctions that are predicted within an alignment set. First, we

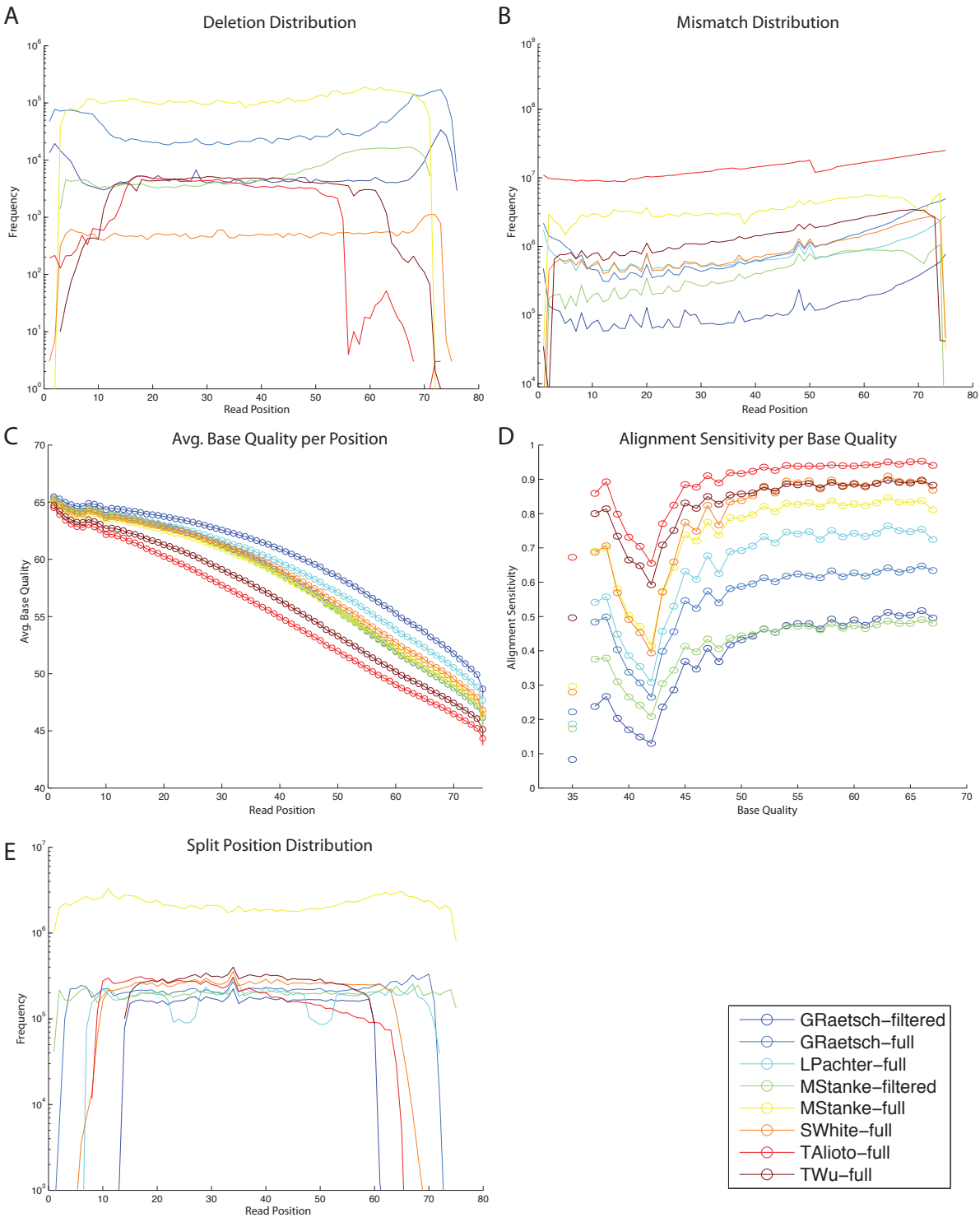


Figure 2.8: Alignment statistics of all submissions for alignments of *C. elegans* samples. ***A:*** Distribution of deletion operations over the read length, showing over- and underrepresentation of deletions towards the read-ends. ***B:*** Distribution of mismatches over read length, showing the quality-dependent 3'-bias that causes more mismatches at the read-end. Peaks indicate sequencing artifacts replicating over several samples. ***C:*** Average base quality per positions, also showing the 3'-quality-bias. ***D:*** Alignment sensitivity per base quality value. ***E:*** Split-position distribution over read-length. Segment-length filters appear as steep drops at the borders. Segment-split artifacts show up as areas with slow split frequency (25 and 50 for submission *LPachter-full*).

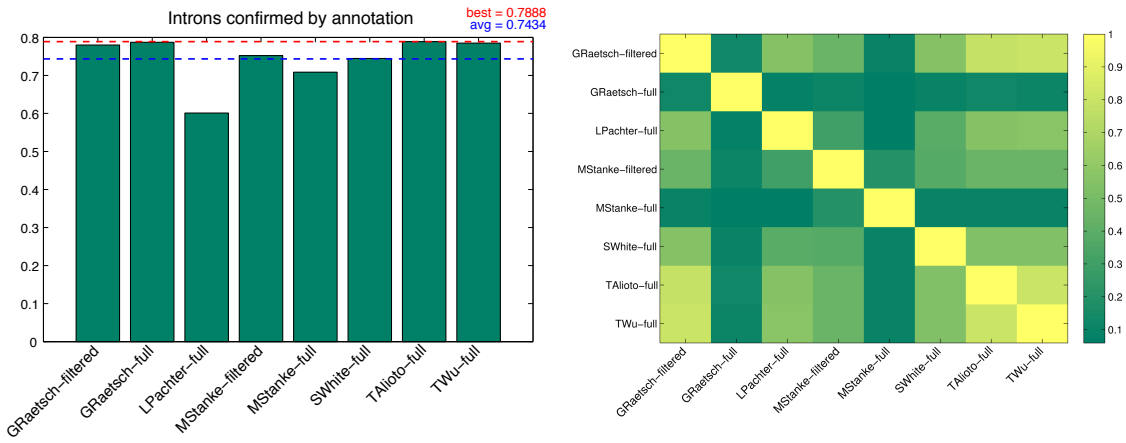


Figure 2.9: Assessment of intron accuracy for different alignment approaches on the data for *C. elegans*. ***Left:*** Intron-level F-Score values for agreement between annotated introns and the predictions by the submissions. Each submission has been individually optimized. Dotted lines show best (red) and average (blue) F-Score. ***Right:*** Relative comparison of the agreement between single submissions. Jaccard-Index between 0 (green) and 1 (yellow).

determined the distribution of split positions in the reads over the read position to get a qualitative understanding of the different spliced alignments. As shown in Figure 2.8, Panel E, some approaches show a higher probability for split positions towards the read ends, which suggests a possible trade-off problem with mismatches, other approaches show lower probabilities for certain positions, e.g., TopHat (*LPachter-full*), where reads are initially segmented into 25-mers, causing a problem to identify split positions close to the segment boundaries.

Second, if a gold standard set is available, e.g., for simulated data or a trustworthy subset of annotated junctions, this can be used to compute an absolute accuracy measure. An example for this is shown in Figure 2.9, left panel. We computed the intron level F-Score of the predictions compared to all annotated introns. As already described earlier, we have optimally filtered each submission before we assessed intron accuracy with respect to the annotation (the full effect of filtering will be described in Section 2.3). We find that all methods show a similar accuracy, with the exception of TopHat (*LPachter-full*) and BLAST (*MStanke-full*) that stand out and show significantly less accurate predictions.

A third way of comparison is to measure the relative agreement between samples, which provides information about strengths and weaknesses of single approaches. The result of the pairwise comparison of all submissions is shown in the right panel of Figure 2.9. We observed a higher agreement between sets that underwent filtering of introns based on read quality and coverage, again emphasizing the importance of alignment post-processing.

Multimappers Ambiguously mapping reads are a general problem for RNA-Seq alignments, as alignment counts are often used as a proxy for gene expression and false alignments can bias these counts. When aligning to the full genome, main sources of ambiguous mappings are repetitive and low complexity regions - especially in non-coding parts of the genome. This could be solved by repeat-masking the respective regions or an alignment to the transcriptome sequence only. However, the first strategy would lead to information

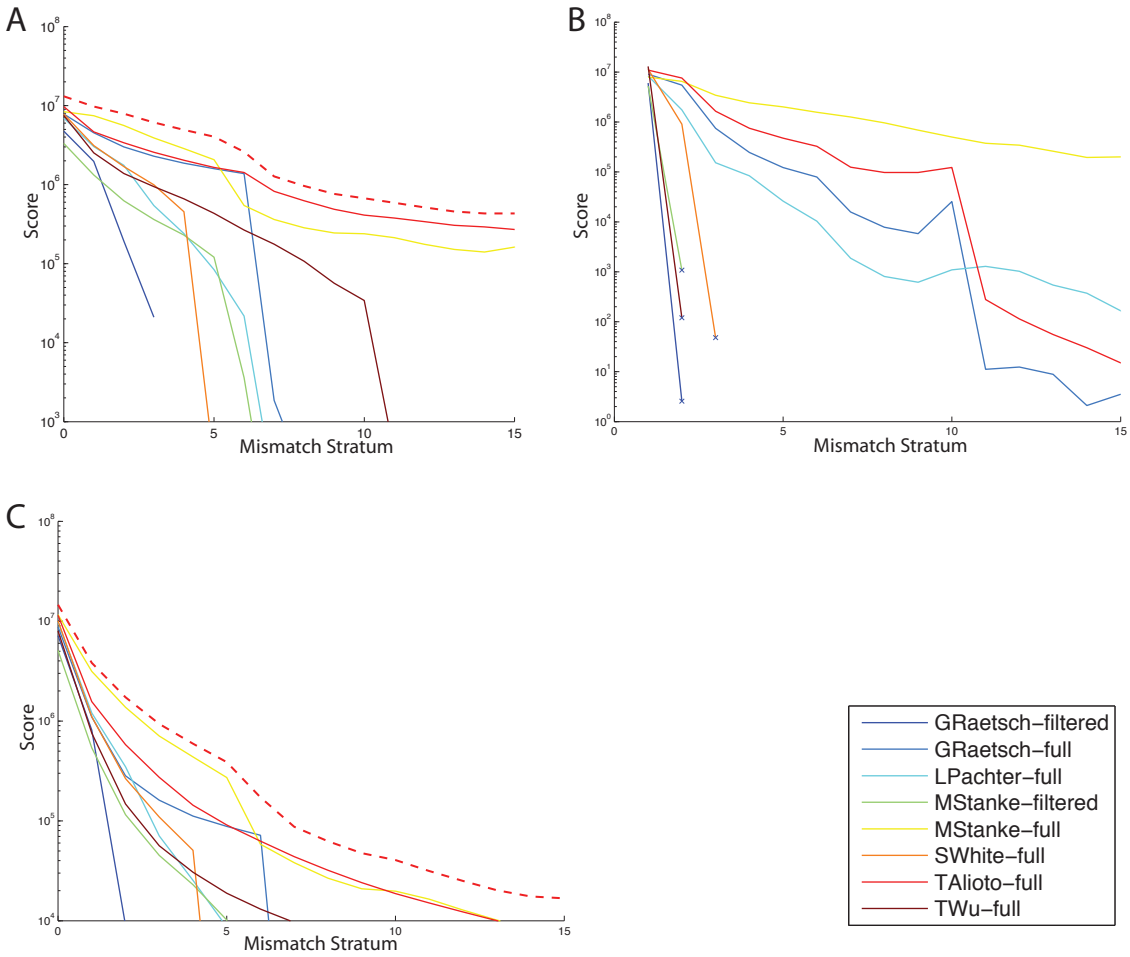


Figure 2.10: Evaluation of ambiguous read-alignments on the *C. elegans* dataset. The three panels correspond to the three different evaluation measures introduced in Section 2.2.3. The dashed red lines show the maximum scores for the respective measures. All scores are shown on a log scale. ***A:*** Comparison per mismatch stratum – alignment accuracy to the global list is stratified by mismatches. ***B:*** Comparison per alignment list stratum – alignment accuracy computed relative to the cumulative number of alignments up to the respective number of mismatches. ***C:*** Comparison per weighted mismatch stratum – alignment accuracy computed relative to the union of possible alignments per stratum weighted with the number of mismatches.

loss and the second would make the detection of novel isoforms impossible. Further, difficulties can still arise from transposable elements, paralogous genes or pseudogenes, that can show a high similarity to expressed genes. The various methods show very different sensitivity and selectivity regarding multiple alignment locations of a read. Therefore, we devised several measures to assess completeness and accuracy of ambiguous mappings (for a detailed explanation see Section 2.2.3). We provide an example overview of each measure in Figure 2.10. We need to note that all three measures only consider sensitivity, as no set of true positive multimappers was available.

Consequences for Downstream Analyses There exist numerous ways how inaccurate read alignments can negatively affect downstream analysis. For instance, the identification

of transcript structures can be strongly impeded by false positive spliced alignments. Especially approaches like Scripture that sample all paths through a splice graph suffer from overly complex graphs containing many false positive edges, but also Cufflinks can produce incorrect isoforms if large numbers of incorrect alignments exist. Also *in silico* quantification of genes or transcripts can be affected. There, ambiguous read alignments can cause both over- and under-estimates of the true quantification value if false positive alignments are taken into account. Even if ambiguous mappings are filtered out, biases can occur, e.g., an artificial underrepresentation of alignments to genes that show high similarity to pseudogenes. We will discuss two practical examples in the context of read filtering in Section 2.3.3, when we show how a reduction of false positive alignments can increase the accuracy of downstream applications.

Further Evaluations Based on our results, another evaluation round was initiated in context of the RGASP competition, to better understand differences in alignment strategies and results. Again, several sets of read data were aligned by the submitters and then compared by a team of evaluators in a comprehensive manner. To also enable an evaluation in absolute terms, two artificial read sets were part of the data. We will not discuss the specific results of that evaluation and refer to the original publication for details [79].

2.2.5 Implementation and Software

The scripts used for the evaluations were implemented in Python and have been merged into a stand-alone tool as part of the RNA-geeq package. Each of the evaluation studies described above can be generated from one or more alignment files provided in SAM format. The source code is available under `https://github.com/ratschlab/RNA-geeq`. A list of dependencies as well as an overview of parameters and the user interface are provided in Appendix A.2.

2.3 Optimal Filtering of RNA-Seq Alignments

In the previous section, we discussed differences and biases caused by alignment algorithms and pointed out how these differences could lead to false positive results in downstream analysis. To overcome this, we suggested to post-process the initial alignments, in order to decrease the number of false positives. In this section, we describe our approach for optimal alignment filtering – in our opinion one of the most essential steps of RNA-Seq alignment post-processing. At the beginning, we motivate the need for appropriate post-alignment filtering strategies and give a description of the different alignment features we use for filtering. In the second part we describe how we chose an optimal set of filter parameters and how they influence each other. In the third part, we show results on the improved accuracy and how optimal filtering corrects the results of downstream analysis. Lastly, we describe the implementation and the software tool resulting from this work.

2.3.1 Motivation and Filter Criteria

In context of the evaluation of RNA-Seq alignment algorithms described in previous Section 2.2, we have indicated that composition and quality of the alignment sets can have

strong influence on the performance of downstream analysis tools, such as transcript identification and quantification [79]. Especially false-positive spliced alignments that introduce evidence for non-existing splice junctions can cause severe problems in subsequent analysis steps. For instance, methods for transcript prediction can have problems if man spurious introns unnecessarily complicate the analysis, such as Scripture [103] and Cufflinks [288]. We propose to use general alignment features such as the number of edit operations, the minimal segment length within a split alignment or the coverage of splice junctions, to determine a subset of high-confidence alignments that help to make downstream analyses more reliable, robust and comparable. Our focus is thereby on spliced alignments. To identify high-confidence alignments, different strategies of filtering can be applied and combined, to produce an optimal outcome. We tested several such strategies, including an approach for training a support vector machine (SVM) to learn a classifier that could identify high-confidence alignments based on alignment features. Interestingly, this rather sophisticated method was only slightly better than a simple alternative. Thus, we lastly devised an algorithm for automatic filtering of alignments that optimizes a set of given filter criteria based on a small set of trusted splice junctions. In the following, we motivate our set of filter criteria:

- **Edit operations** One of the probably most obvious filter criteria is the number of edit operations used in the alignment, as it directly reflects the match-quality between source- and target sequence. Each edit-operation that is allowed during the alignment provides additional degrees of freedom for the mapping and is, thus, a possible source of false-positive alignments. By allowing no or only a small number of edit-operations, the specificity of alignments can be increased.

  We define the criterion `editop` as the number of edit operations of a given alignment.

- **Segment length** For split read-alignments, the position of the intron within the read defines a segmentation of the read. A placement of the introns towards the edges of the read results in at least one very short segment. The shorter a segment becomes, the more difficult it is to identify its correct mapping location. Thus, a decrease in segment-length increases the difficulty to identify its correct mapping location. This problem remains even for reads of increasing length, in cases where the reads become longer than the median exon length (e.g., 160 nt in human [248]) and, thus, many reads contain multiple junctions, again causing short segments. Hence, using the minimal segment length of an alignment as filter, helps to reduce the number of false positive spliced alignments.

  We define the criterion `seg_len` as the shortest continuous segment in an alignment.

- **Junction Coverage** Depending on the depth of sequencing, each position in the genome/ transcriptome is spanned by a certain number of reads on average. As this also translates to split-alignments, we expect several alignments per splice junction. Thus, even for the case that spliced alignments have a short minimal segment-length, the probability of a spurious alignment decreases with the number of independent alignments supporting the used intron.

  We define the criterion `junc_cov` as the number of alignments in the set containing the same splice junction.

Each of the criteria defined above can be used to filter a given alignment set. That is, to select a subset of high-confidence alignments by only retaining alignments that suffice the given criterion. However, we realized that it is often not sufficient to use only one criterion for filtering. Thus, we proposed an algorithm to determine an optimal filter combination.

2.3.2 Search for an Optimal Parameter Combination

As we are most interested in increasing the accuracy of spliced alignments, we only use these to determine a good filter combination. In order to evaluate the quality of a filter, we need a set of true positives that can be used for evaluation. Focusing on the accuracy of spliced alignments, we use a list of true positive junctions as ground-truth. This set can either originate from a set of annotated transcripts, a database for splice junctions or a biological validation experiment. Here, we assume that this list is provided as input.

The general idea of the approach is very simple. We take an alignment set, filter it with respect to one possible combination of the three criteria we defined, that is we remove all alignments that have more than `editops` edit operations and all spliced alignments that have a minimal segment length below `seg_len` or contain an intron that has a coverage below `junc_cov`. We then evaluate the agreement between the list of splice junctions contained in our filtered alignment set and the given set of ground-truth junctions. This is repeated for each combination of values for `editops`, `junc_cov` and `seg_len` that should be tested. The combination that maximizes the agreement can then be used to filter the alignment set.

Different measures can be used to evaluate the agreement to the ground-truth set. The *recall* measures the fraction of correctly predicted true-positives over the number of all predicted junctions. Using this measure maximizes the sensitivity of the alignment set. The *precision* measures the fraction of correctly predicted true-positives over the number of all true-positives in the ground truth. This measure maximizes the specificity of the alignment set. A third measure is the *F-Score*, which is computed as the harmonic mean of precision and recall. Using this measure usually finds a good balance between sensitivity and specificity. Therefore, we use the F-Score for our approach.

The proposed strategy makes it necessary to read the alignment file many times, causing a tremendous overhead. We therefore use a pre-processing step on the alignment files extracting all necessary information. Only $\approx 20\%$ of all alignments are spliced alignments (depending on alignment parameters, organism and read length), thus we reduce the input alignment files into a compressed summary of annotated splice junctions. To this end, we read through the alignment file only once and collect information from each spliced alignment. However, we do not store information per alignment but rather per splice junction that it contains. For example, if a spliced alignment contains an intron from x to y, has three edit operations and a minimal segment length of 12, we increase for intron $x \ldots y$ the count for criteria combination (`seglen`=12, `seglen`=3) by one. Although simplified in the example, each junction can be uniquely identified by chromosome, strand, start- and end-position. After reading all input alignments, we know for each junction, how often which combination of the two filter criteria occurred. The annotated junction list has usually a size of only several MB, such that it can be kept in memory easily and is immediately accessibly. We store this list in an associative data structure such that we can easily query

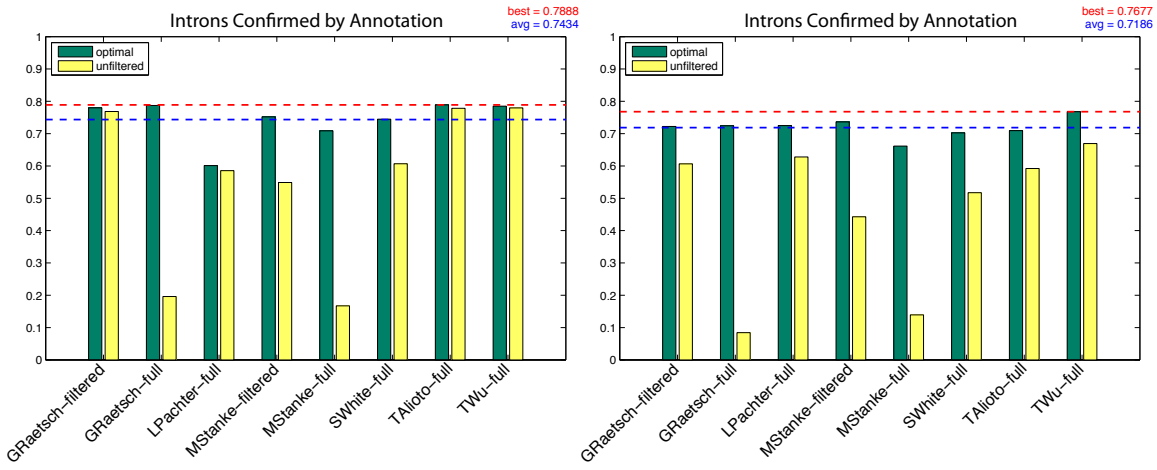


Figure 2.11: Improvements in the F-Score of intron agreement between alignment and annotation. Unfiltered submissions in yellow and optimally filtered submissions in green. Dotted lines show average (blue) and best (red) agreement to the annotation based on the values for the optimally filtered submissions. ***Left:*** Submissions for dataset *C. elegans*. ***Right:*** Submissions for dataset *D. melanogaster*.

it with a combination of filter criteria to request a list of junctions that fulfill the criteria.

We propose two different strategies to identify the best filter combination. The first—much faster—strategy, uses a line search on each criterion to find a good filter set. That is, we iterate over each criterion separately and find the value maximizing the F-Score. The optimal combination is then chosen as the set of combined values. As the different criteria influence each other, the order in which they are optimized strongly influences the result and the optimal solution is possibly not unique. The second strategy is a grid-search over all possible combinations of values for all filter criteria. This strategy is much more costly, as its complexity grows in $O(n^k)$ for k features and n feature steps. However, with $k = 3$ the number of features we use is low and with $n = 15$ the granularity is rather moderate, making even this expensive strategy computationally feasible. For both strategies, further criteria like quality of the read and alignment multiplicity can be easily integrated into the search, although the feasibility to search the whole grid should be considered.

Despite the pre-processing step to extract all junction information, an assessment of the whole alignment-set can be infeasible. In such cases a random subset of the alignment-set can be taken instead, loosing sensitivity for junctions in lowly expressed transcripts. However, for optimizing alignment filter parameters, the highly expressed transcripts should provide a very good proxy.

2.3.3 Results: Effects on Alignment-Accuracy Downstream-Processing

To assess the results of the filtering method introduced above, we relied on the evaluation procedures described in previous Section 2.3.2. We compared the junctions predicted by the submissions to the list of introns available in the respective annotation and computed the F-Score as accuracy measure. As shown in Figure 2.11, all submissions show better agree-

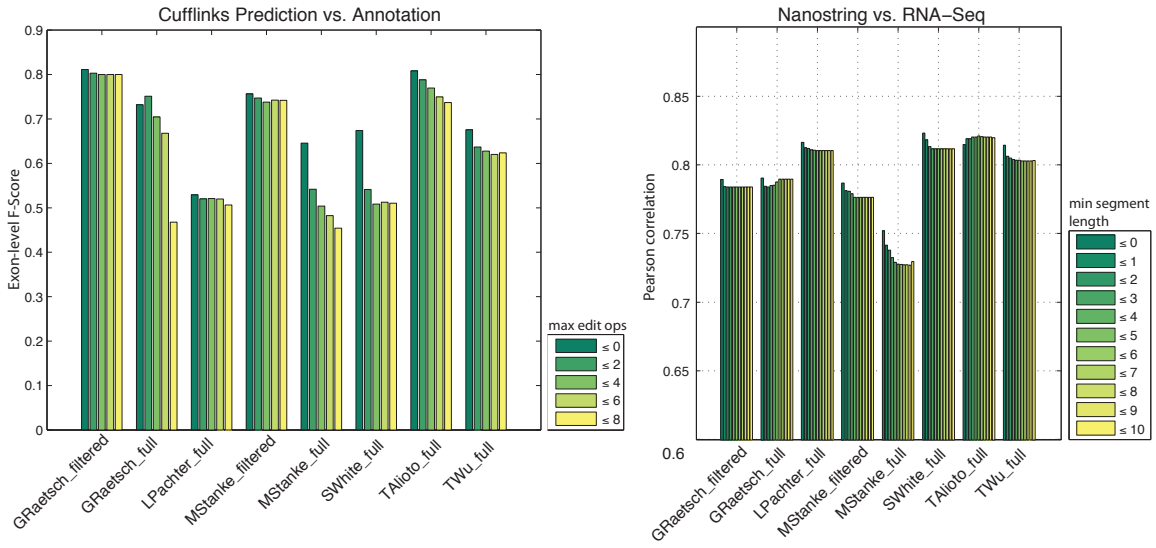


Figure 2.12: Effect of optimal alignment-filtering on the quality of downstream analysis results. ***Left:*** Accuracy of transcript predictions made with Cufflinks on the *C. elegens* alignments, evaluated as exon-level F-Score with the annotated exons as ground truth. Filter criterion were edit operations. ***Right:*** Pearson correlation coefficient of transcript quantifications of rQuant and NanoString-counts. Quantifications were predicted on the *D. melanogaster* alignments. Filter criterion was the minimal segment length.

ment with the annotation after the optimal filter was applied. However, some submissions show only slight improvements (Figure 2.11, submissions *GRaetsch-filtered*, *TAlioto-full* and *TWu-full*). These alignment-sets were pre-filtered by the submitters and no unfiltered set was available for comparison. Notably, the variability between submissions is drastically reduced by filtering and the average agreement is much higher. Single submissions show an improvement of up to 6-fold, e.g., BLAT (Figure 2.11, *MStanke-full*). Even if the submissions were pre-filtered (*MStanke-filtered*), we could further improve on the filter settings that were applied originally.

A second interesting aspect of filtering is its effect on the results of downstream analyses. Short read alignment is usually only the first step within a pipeline for quantitative and qualitative analysis of the transcriptome (cf. Section 1.3.1). From the wide range of possible downstream applications, we chose the *in silico* prediction of transcript isoforms and *in silico* isoform quantification as representatives to measure the effect of alignment filtering on downstream analysis results. For each, we chose a set of true positive results and evaluated the outcomes with respect to that set. Whereas the predicted transcript isoforms were compared to the annotation as a ground truth, the results of the quantification were correlated to NanoString-counts (cf. Section 1.2.1) generated from the same samples. We used two different software packages to predict transcript isoforms from the data: Cufflinks [288] and Scripture [103]. *In silico* quantification was done with rQuant [31]. Two results of this comparison are summarized in Figure 2.12. On the left, the results for transcript predictions by Cufflinks on the alignments for *C. elegans* show a steady improvement of prediction accuracy with stricter filter criteria. Especially previously unfiltered submissions (*MStanke-filtered*) show marked improvements. The right plot shows the accuracy of transcript quantification with rQuant, measured as Pearson correlation coefficient. Again,

most submissions show accuracy improvements with more stringent filtering criteria. Interestingly, one submission (*TAlioto-full*) shows slightly increased performance when filtered moderately, but has a decreased performance if filtered more strictly. We believe that this happens, if the alignments are very fragmented and too many alignments are filtered by the segment-length filter that was applied for this experiment.

In summary, we have presented SAFT, a tool that aims at filtering RNA-Seq alignments to improve comparability between different alignment approaches and increase robustness of downstream analysis. Our evaluations showed that the agreement of different alignment methods was greatly improved after optimal filtering through SAFT (cf. Figure 2.11), boosting the intron accuracy of several compared approaches. We also saw drastic performance-improvements for transcript prediction and quantification, taking the filtered alignments as input. Thus, SAFT is well suited to improve alignment accuracy, if a complete alignment set is provided and a reference set of high-confidence splice junctions is available. However, in cases where alignments have already been pre-filtered or no accurate junction information is available, SAFT is not able to determine a good filter combination. The tool is further restricted to alignments of RNA-Seq data, as DNA-Seq does not provide intron information, which is necessary to evaluate performance.

2.3.4 Implementation and Software

We implemented the filter optimization as a stand-alone tool on Python, called Simple Alignment Filtering Tool (SAFT). The implementation is available as part of the RNA-geeq package and published under BSD license. We made the source code available at `https://github.com/ratschlab/RNA-geeq`. The overview of the Linux command line user interface shows all parameters available to the user and is shown in Appendix A.3. The implementation has the following dependencies: Python (version 2.7 or later), SciPy (version 0.13.0 or later), samtools (version 0.1.12 or later).

2.4 Resolution of Ambiguous Read Mappings

In the previous section we discussed how to reduce the number of false-positive alignments through optimal alignment filtering based on simple features of the alignment. Another source of false positive alignments, already mentioned in Section 2.2, is alignment ambiguity. In this section, we introduce several possible sources of ambiguous read-mappings and discuss their relevance for transcriptome analysis. We then describe the multi-mapper-removal (MMR) algorithm to resolve ambiguity in read-placement and show how this can improve the results of downstream analyses that take these pre-processed alignments as input. Finally, we conclude by describing the implementation of the algorithm and give a short overview on the software.

2.4.1 Motivation

High-throughput shotgun sequencing relies on the highly parallel generation of very short sequence fragments. Although different sequencing platforms show a range of length distributions (cf. Section 1.2.2), the reads are generally short enough to cause ambiguity problems

during alignment. In our experience, the fraction of ambiguous reads ranges from 10–15% for reads of 75–100 nt. We distinguish two main sources of ambiguity: the read-set and the alignment target sequence. We begin by describing several factors that can cause redundancies in the read-set.

- **Paralogous Genes** As a result of gene duplications during the evolutionary history of an organism, genes can have several similar copies within the genome that show a very high sequence identity. This is true for active protein coding genes but also for processed pseudogenes. Reads that are sampled from these genes are likely to be identical.

- **Repetitive Regions** A main problem arising in whole-genome sequencing, are genome regions that show a high fraction of repetitive elements. Main sources for such regions of low complexity are short tandem repeats or longer sequences of transposable elements that spread based on a retrotranscription process [261]. Best known are long and short interspersed elements (LINEs and SINEs, respectively), that account for a substantial fraction of the repetitive part of the human genome, with estimates reaching up to 50% [61]. These regions can be source of a large number of identical reads.

Both mechanisms are possible causes for identical sequencing reads originating from different regions in the genome. Most problematic is the fact, that these reads only occur for specific sequences, causing biases in the read distribution. Interestingly, an effect similar to paralogous genes occurs in metagenomics, where reads can originate from different species with highly similar genomes. However, there the ambiguity is usually resolved by taking the lowest common ancestor as assignment.

The second cause for multiple alignment possibilities of a read does not originate from the sample, but rather from the target sequence of the alignment. Even if no read occurs twice in the sequencing sample, paralogous genes and repetitive regions in the genome can cause multiple alignment possibilities for a subset of reads. This effect can be further boosted through the choice of a relaxed set of alignment parameters which introduce artificial redundancies in the genome. That is, the more edit operations are allowed during alignment, the more possible mapping locations exist in the genome. Although a restriction to perfect matches would resolve the latter issue, this solution is of limited practical use, as sequencing errors (cf. Section 1.2.2) have to be taken into account. Assuming an error rate of 0.1%, on average every tenth read in a set of 100 nt reads has one error and only allowing for perfect matches would discard ≈10% of all reads. Hence, even a conservative setting usually allows for one or two edit operations per 100 nt read-length.

If reads can originate from multiple sources, this generates an uncertainty of read-placement during the alignment process. Each read that can be assigned to several distinct locations in the genome is called *ambiguous mapper* or *multi-mapper*. As implicated earlier, these ambiguities can have strong influence on downstream analyses. For instance, in

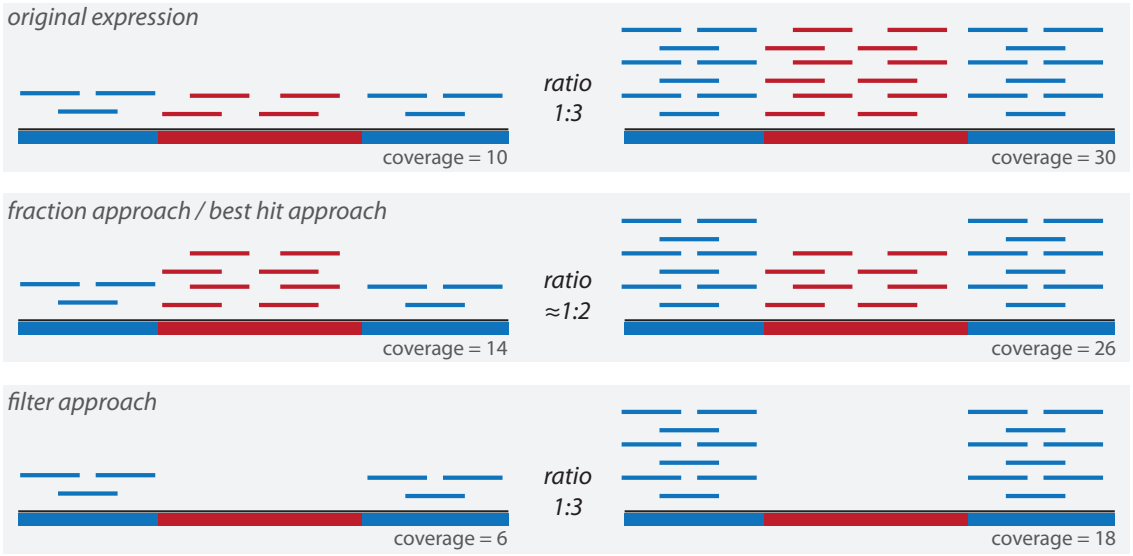


Figure 2.13: Motivation for resolution of ambiguous alignments. On the left and right side are two different genes with identical sequence parts (red boxes) and differing sequence parts (blue boxes). Read coverage is shown as reads (short solid blue and red lines). Top row shows the true expression, the middle row shows the approach assigning same shares of multiple mappers to each possible location and the bottom row shows removal of all ambiguous mappers.

methods estimating quantification, the number of aligned reads is often used as proxy for the expression of a gene or transcript. In this context it is not clear how to correctly count a read that is mapping to more than one gene. We provide a cartoon example in Figure 2.13.

There are several strategies to overcome this problem. One early solution was to assign fractions of the read to all its mapping locations, for instance, in ERANGE [206]. This approach is quite problematic, as it removes signal from the true mapping location, creating at the same time false positive signals at many others (Figure 2.13, middle). It also is not well motivated biologically, as each read can only have exactly one source location. Another strategy is to ignore all reads that show more than one possible alignment. Based on the properties of the underlying transcriptome (complexity, repetitiveness), the sequencing protocol (read-length, error-rate) as well as the alignment approach (degrees of freedom through edit operations), the remaining fraction of uniquely mappable reads can be as low as 70% (empirical estimate). Thus, this approach would ignore a substantial fraction of the input data and likely loose many informative alignments (Figure 2.13, bottom). A second strategy is to retain only the alignment that has the highest alignment score (usually the negative logarithm of a false-positive probability [80]). This decision is based on the assumptions made by the alignment algorithm and is typically arbitrary in cases where several equally likely alignments exist. For ambiguous mappers arising from low complexity regions or paralogous genes, equally likely alignment locations are common. Thus, for these reads this strategy would result in an arbitrary decision which alignment to keep. Placing alignments arbitrarily amongst mapping locations essentially results in the fraction approach described above (Figure 2.13, middle).

In this work, we present an algorithm that resolves mapping ambiguities and assigns a single mapping location to each read, based on local sequence-coverage only. This ap-

proach makes full use of all aligned reads, thereby taking the alignment score into account. It relocates reads to locations in the genome, where uniquely aligned reads provide additional evidence for expression at this locus. Thus, our approach is able to distinguish between expressed and unexpressed genomic locations and can incorporate this into the alignment-choice decision. We are aware of one other approach that also resolves mapping ambiguity [32]. However, this tool uses a re-alignment strategy and is thus less efficient than our suggested approach. It also cannot operate on existing alignment files.

2.4.2 Approach: Local Coverage Minimization

We will begin with an informal description of our algorithm. Its main idea is to use local coverage information to decide where to put a read, if multiple alignment locations exist. The simplest case is to imagine two genomic locations that share both an identical and a non-identical sequence part. Reads aligning to the non-identical part can be used to infer the expected coverage of the gene, which can then be used to infer the amount of alignments to the identical part in each locus. Following the example in Figure 2.13, we would use the level of blue reads to infer the desired number or red reads and use this information to assign the red reads to one of the two locations. An optimal assignment would place each read at a location, such that alternative regions optimally fit into their non-alternative contexts. However, testing all possible combinations of read assignments to all mapping locations is computationally infeasible. Therefore, we suggest an iterative approach to only alter the alignment location of one read at a time, keeping all other reads fixed, and apply this sequentially for all reads. Repeating this process for several iterations, converges to a local optimum in global read coverage.

However, this approach only works, if we assume a uniform coverage over the length of the gene. An idealized sequencing process would sample reads from a source sequence following a uniform distribution. That is, each read can originate from any location with the same probability. However, this is not the case for real sequencing samples. Due to various biases in different parts of the sequencing process, such as priming, amplification or fragmentation, the reads show a non-uniform distribution over the length of a transcript or gene [31, 67, 106, 210]. Therefore, making the assumption that read coverage is uniform over the length of a whole gene is inaccurate. However, most of these biases act on a longer range of several hundred bases or have an effect that is sequence specific and thus locally similar. Hence, within a local window the distribution of reads is much more uniform. Based on this observation, we make the reasonable assumption that the coverage of a given transcript is relatively smooth, that is, the difference of coverage between neighboring positions is small. Stated differently, we assume that the coverage within a small local window is almost uniform. Hence, we apply the procedure described above not on the level of genes but rather on windows around the alignment location. This central assumption of the algorithm is violated, if gene structure and alternative usage of isoforms within a gene influence smoothness even within a local window. To resolve this, the algorithm is able to take known structures into account. We will discuss this in the context of MiTie [25] in Section 2.4.3.

Following the idea described above and using the assumption of locally smooth coverage, the whole set of possible alignments for a given read is evaluated, with the goal to identify the mapping that results in the locally smoothest coverage. In this context we measure smoothness as the empirical variance of the position-wise coverage in a window around the alignment location. The algorithm then minimizes the variance over all possible alignment locations, choosing the alignment with the smoothest coverage as optimal. This works as follows. Given an input of k different alignments for a given read, one alignment is designated as the currently best. Depending on user preference this is either an arbitrary alignment or the mapping with the highest alignment quality. This current best mapping is then compared to each of the remaining mapping possibilities in a pairwise manner. For a single comparison, four variance values are computed. Given two possible alignments $\mathbf{a}_1$ and $\mathbf{a}_2$ to the genomic start locations l_1 and l_2, respectively, the score v_{1+} contains the local variance around genomic location l_1 if $\mathbf{a}_1$ is mapped to that location and v_{1-} if it is mapped somewhere else; v_{2+} and v_{2-} are defined analogously using the alignment $\mathbf{a}_2$ to genomic locus l_2. In each case the score is defined as the empirical variance over the genomic coverage of all window positions

$$v_1 = \frac{1}{k_1 - 1} \sum_{i=0}^{k_1-1} \left(\mathbf{a}_1[l_1 + i] - \frac{1}{k_1} \sum_{j=0}^{k_1-1} \mathbf{a}_1[l_1 + j] \right)^2$$

where k_1 is the number of positions in a window around alignment $\mathbf{a}_1$ and $\mathbf{a}_1[i]$ indicates the coverage at genomic position i. The window length k defaults to 20 nt and can be adapted by the user. If an alignment is present within the window, it influences the coverage and thus the local variance. After computing all four values, $\mathbf{a}_1$ is chosen if

$$v_{1+} + v_{2-} < v_{1-} + v_{2+}$$

is true, otherwise $\mathbf{a}_2$ is chosen. A schematic visualization of the MMR principle is shown in Figure 2.14.

A major complication arising during the computation of v_{1+}, v_{1-}, v_{2+} and v_{2-} is the special case that occurs when the windows of $\mathbf{a}_1$ and $\mathbf{a}_2$ share common positions. In this situation, two different scenarios can occur:

i) the windows share positions but the alignments do not share positions,

ii) the alignments share positions.

As the read is placed at either the one or the other location, in case i) the computation of v_{1-} needs to consider coverage contributed by $\mathbf{a}_2$ as this will be placed instead of $\mathbf{a}_1$ and v_{2-} needs to consider coverage contributed by $\mathbf{a}_1$. Case ii) causes a subset of positions that are shared by $\mathbf{a}_1$ and $\mathbf{a}_2$ to not be altered by the decision. These positions can be masked for analysis and left out in computation, as they contribute to both locations not changing the result.

Our approach can be extended easily to also work for paired-end RNA-Seq alignments. In this case, a preprocessing-step creates all possible valid pairs of alignments of the two

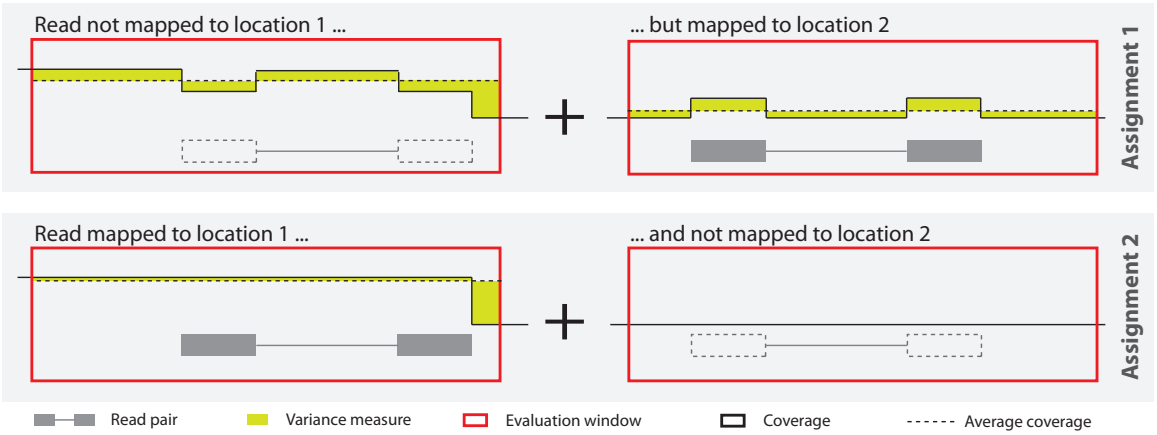


Figure 2.14: Schematic overview of the principle to resolve ambiguous read-mappings. The candidate read-pair in gray has two possible alignments $\mathbf{a}_1$ to location 1 (left) and $\mathbf{a}_2$ to location 2 (right). Variance measures (yellow) are computed for both locations, with and without the read-pair. Variance values from the text have following correspondents in the schema: v_{1-} – location 1 (top), v_{1+} – location 1 (bottom), v_{2+} – location 2 (top), v_{2-} – location 2 (bottom). The evaluation windows are shown in red and the coverage of placed reads as black solid lines.

mates. An alignment pair is valid, if the corresponding alignments do not overlap in a conflicting manner. For instance, a conflict would occur, if the first read-mate is aligned into the intronic portion of the second read-mate, if both reads are aligned in the same direction, if the reads align to different chromosomes, or if both alignments have a distance outside of a user-defined maximum range. After this preprocessing-step, each alignment pair is treated as single alignment possibility $\mathbf{a}_k$ and the algorithm above is applied. As the number of possible pairs is quadratic in the number of alignments in the worst case, the number of allowed pairs can be limited by the user.

2.4.3 Minimization in the Context of Transcript Prediction

One limitation of the strategy described in Section 2.4.2 is that transcript structure is not taken into account. Especially the exon–intron boundaries show steep changes in coverage, but also within exons a change in coverage can often be explained by a mixed signal from several transcript isoforms that superimpose each other. If the underlying transcript structure is known, it can be accounted for during the optimization process. To include structural information into MMR, we devised a strategy that takes transcript structures and quantifications produced in the process of *in silico* transcript prediction into account to resolve read mapping ambiguity. This method can be applied in an iterative scheme. It starts with transcript isoform predictions and isoform quantifications on the alignments using the best hit. Ambiguous alignments can then be re-evaluated based on the transcript structure and the estimated transcript expression. The improved alignments can then be used to generate improved isoform predictions and quantifications. This can be repeated a fixed number of times or until convergence of the predicted quantifications.

How the decision to reposition an ambiguous alignment is made if transcript structures are given, will be explained in the following. We devised a strategy for the iterative application

of MMR and MiTie [25], a tool for the prediction and quantification of transcript isoforms. If the exon boundaries of all transcript isoforms of a gene are projected to genomic coordinates, the gene can be cut into a set of non-overlapping exonic segments. Thus, each isoform can be built from a subset of these segments. Several isoforms can share the same segment. The expression value of a single segment is the sum of the segment's expression values over all transcript-isoforms containing that segment (for a more formal description see [25]). The segments as well as corresponding expression estimates provided by MiTie can be used as input for MMR. These segments imply a segmentation for the whole genome. Each given segment is associated with a predicted expression value. The genomic regions between any segments are implicitly turned into segments with a predicted coverage of 0. Instead of minimizing the local variance, we now minimize the difference between the observed coverage in an exonic segment with and without the alignment of question and the predicted coverage of the segment. To better account for properties inherent to read-count data, we use a log-likelihood loss function L based on a negative binomial distribution. For technical reasons in the optimization of MiTie, we use a piecewise-linear approximation to the log-likelihood loss function, $l \cong L$. For further details cf. [25], Suppl. Section K.

Taking the same two alignments $\mathbf{a}_1$ and $\mathbf{a}_2$ as in Section 2.4.2, for each alignment we can now identify all genomic segments it overlaps with. Let alignment $\mathbf{a}_1$ overlap the m genomic segments $g_1^1, \ldots, g_1^m$. We can then compute two coverage values for each segment. The value c_1^{i+} that contains the coverage of segment g_1^i if alignment $\mathbf{a}_1$ is mapped to segment i and c_1^{i-} that contains the coverage of the same segment if $\mathbf{a}_1$ is mapped to a different location. For each segment we can then compute the difference between observed and predicted coverage, using the expression estimates $e_1^1, \ldots, e_1^m$ corresponding to the respective genomic segments g_1^i and the loss function l described above. Thus, the total loss of $\mathbf{a}_1$ is

$$v_{1-} = \sum_{i=1}^{m} l(e_1^i, c_1^{i-}) \quad \text{and} \quad v_{1+} = \sum_{i=1}^{m} l(e_1^i, c_1^{i+}).$$

Analogously, we define the total loss of alignment $\mathbf{a}_2$ overlapping genomic segments $g_2^1, ..., g_2^n$ with expression estimates $e_2^1, ..., e_2^m$ as

$$v_{2-} = \sum_{i=1}^{n} l(e_2^i, c_2^{i-}) \quad \text{and} \quad v_{2+} = \sum_{i=1}^{m} l(e_2^i, c_2^{i+}).$$

We assume the segments to be independent and thus can sum the log-likelihood losses of the single segments.

Besides the different calculation of v_{1+}, v_{1-}, v_{2+} and v_{2-}, all other steps are identical to the steps described in Section 2.4.2. Although slightly different adaptations need to be made in order to account for overlapping alignment locations, the same general principles apply.

We also implemented a hybrid-strategy that does not rely on predicted expression values for the segments. Instead it uses the minimization of the coverage-variance on a segment-basis and computes the sum over all segments an alignment overlaps, to determine a total variance value. All other other steps are the same as described before.

2.4.4 Results and Evaluation

We evaluated our algorithm on two different simulated datasets for an accurate assessment of its performance. For the first dataset we tiled the complete *A. thaliana* TAIR10 reference genome [151] at each position into a set of overlapping 50-mers, thus generating artificial reads from whole genome sequencing, containing all low-complexity regions of the genome. In this idealized dataset the coverage at each genomic position (except the 50 nt at each end) is exactly 50. We then used PALMapper to realign the first 1,000,000 reads back to the *A. thaliana* genome, allowing for up to 5 edit operations, thus generating a high level of additional ambiguity.

As indicated in Figure 2.15, MMR is able to fully resolve all read-ambiguities in the genomic DNA dataset (peaks and valleys in the upper coverage plot in Panel A) in the genomic DNA dataset, leading to a uniform coverage in the MMR-filtered dataset (lower coverage plot in Panel A). This holds still true, if we evaluate the alignment coverage at all genomic positions. Figure 2.15, Panel B, shows two histograms over all genomic positions covered by at least one alignment. Whereas the alignment relying on the best-hit strategy (left) shows numerous positions with coverage higher or lower than 50, the MMR-filtered alignment (right) almost exclusively shows positions with coverage of 50. The few positions with coverage less then 50 are caused by the boundary conditions of the simulation. The very few positions with a coverage exceeding 50 are likely due to false positive alignments. Notably, the best-hit strategy showed single genome positions with a coverage exceeding 1,700 (these are contained in the last bin of the histogram).

As a second evaluation dataset we simulated RNA-Seq reads. With the aim to generate a dataset that is as realistic as possible, we used the simulation toolbox FluxSimulator [102] that simulates the complete sequencing process and incorporates various biases as well as sequencing errors. We generated $3 \cdot 10^6$ artificial RNA-Seq reads sampled from 5,000 randomly selected genes of the human ENSEMBL annotation [85]. We simulated two different read-lengths of 51 nt and 76 nt, resulting in an average coverage of 18 and 25, respectively. The reads were then mutated with an error-model that was estimated on a publicly available sample of Illumina sequencing reads (Short Read Archive, accession SRX026670) and three different levels of additional random noise, increasing by 1% per level, resulting in a total error-rate ranging from 3.7% to 5.7%. The simulated reads were aligned to the hg19 human reference genome using TopHat (version 2.0.2 [137]) and PALMapper (version 0.5 [121]) allowing up to 6 edit operations without additional annotation information provided. All other parameters were left at the default.

Based on this dataset, we tested the effect of MMR on downstream analyses. For this, we used the unprocessed, the MMR-filtered and the best-hit alignment set to perform *in silico* transcript quantification using both Cufflinks [288] (version 1.3) and rQuant [31], where the best-hit set consisted of those alignments that were ranked highest by the alignment algorithm. For both alignment methods TopHat2 and PALMapper the quantifications based on the MMR-filtered alignments showed a consistently better correlation to the ground truth quantification than both the best-hit and unfiltered alignments sets. The short 51 nt reads processed with MMR (2.15, Panel C) showed higher percent improvements in correlation compared to unfiltered (Cufflinks: 14.8%, rQuant: 16.6%) and best-hit set (Cufflinks: 3.2%, rQuant: 3.0%) than the longer 76 nt reads (2.15, Panel D), that also show consistent but less pronounced percent improvements in comparison to the unfiltered set (Cufflinks: 3.4%,

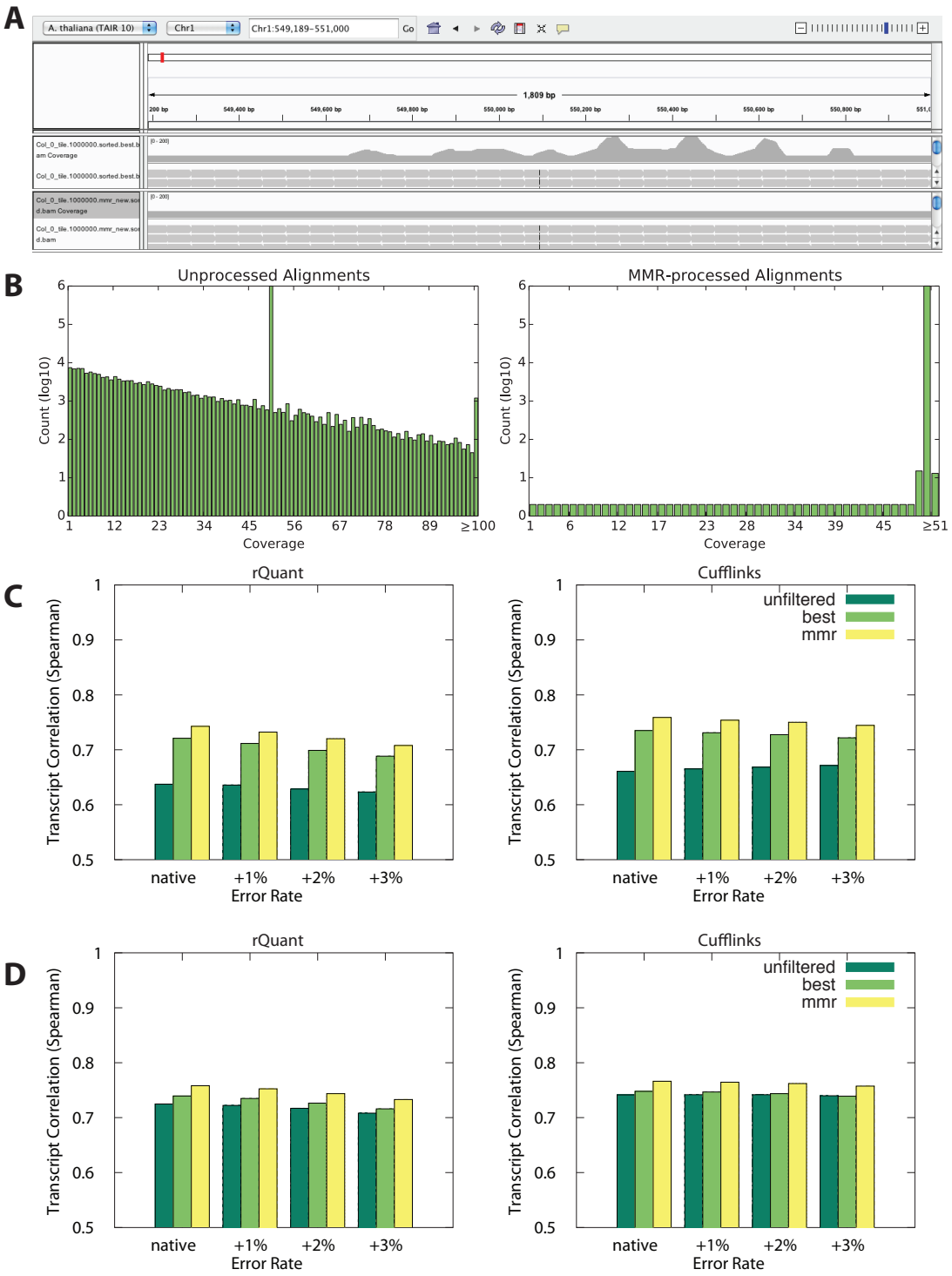


Figure 2.15: Result overview for the resolution of ambiguous read-mappings. ***A:*** IGV [241] snapshot of the best-hit (top) and MMR-processed (bottom) alignments. Ambiguous mappings causing unequal distributions could be fully resolved. ***B:*** Two histograms showing the distribution of position-wise coverage in the aligned sets as log10 counts. The alignment strategy retaining only the best hit is shown on the left and MMR-processed alignments on the right. ***C:*** Results of *in silico* transcript quantification using rQuant (left) and Cufflinks (right) as Spearman rank correlation coefficients between predicted quantifications and ground truth for the simulated set of 51 nt reads. Unfiltered data is shown in dark green, the strategy to take only the best alignment is shown in light green and MMR is shown in yellow. Each block of bars represents a different error-rate. The native error rate is at 2.7%. ***D:*** Same analysis as in Panel C with an altered read-length of 76 nt.

rQuant: 4.6%) and best-hit set (Cufflinks: 2.5%, rQuant: 2.5%).

In summary, the presented the tool MMR that is able to resolve read-ambiguity through taking local coverage information into account. It is able to improve results of downstream analyses, such as transcript quantification and efficiently processes single or multiple input files. Other useful applications could include the assignment of metagenomic reads or to resolve host contamination in samples, e.g., in sequencing probes from mouse xenografts. However, this approach has also several limitations. As it uses an iterative local optimization, it is not guaranteed to find the global optimum. Further, convergence is expected to be very slow in alignments to targets of very low complexity, where many reads have a large number of possible mapping locations.

2.4.5 Implementation and Software

The algorithm is implemented in C++ and is provided with a command line user interface. The implementation uses multi-threading to efficiently parallelize the single local optimizations. Each read can be evaluated independently, as long as no external alterations to the global coverage map are made within the comparison of a read-pair. Appendix A.4 shows the Linux command line user interface of the MMR implementation and provides an overview of the available options. The source code is published under GPL3 license and is available under `https://github.com/ratschlab/mmr`.

2.5 Alternative Splicing Event Detection and Quantification

We describe the tool SplAdder (splicing adder), that we developed to comprehensively capture the alternative splicing state of the measured transcriptome and that is especially well-suited to be applied within a high-throughput setting. At the beginning of this section, we motivate the need for a set of tools and a pipeline that is able to detect different classes of alternative splicing events and to quantify the different states of each event in a given set of samples. The three following subsections, 2.5.2 – 2.5.4, describe the algorithms and data structures that are used to efficiently store all annotated transcript isoforms of a gene, to augment the existing annotation based on RNA-Sequencing data, and to extract alternative splicing events from this structure. In the subsequent two subsections 2.5.5 and 2.5.6, we describe how single events are quantified and how the test for differential event usage between two given sample populations is conducted. In Section 2.5.7, we give a short summary on how multiple input files are handled. We conclude by providing a brief evaluation of this approach as well as a short description of its implementation.

2.5.1 Motivation

Elucidating and understanding the occurrence and regulation of alternative splicing is indispensable for explaining the biological processes that help to turn genetic information into a complex phenotype. Our goal is to gain insights into transcriptional regulation and RNA-processing through the analysis of RNA-Seq data. Depending on the organism, up to 95% of all expressed genes are transcribed in multiple isoforms and undergo alternative splicing [223, 303]. Although these isoforms might never coexist at the same time and place, they

are an important contribution to shaping the complexity of the transcriptome. Their diversity can be essential for cell differentiation, development and signaling processes. To learn more about the various isoforms, it is necessary to generate a picture of a cell's current transcriptome state that is as complete as possible. The various differences in isoform-structure within a gene caused by splicing can be categorized into several classes of alternative splicing events (cf. Section 1.1.1 for a list and detailed review). Whereas numerous approaches exist that aim to predict and quantify whole transcript structures (cf. Section 1.3), only very few focus specifically on single alternative splicing events [36, 77, 243]. Especially in the context of high-throughput applications, a restriction to single alternative splicing events is often computational much more feasible, due to a lower grade of complexity and only local dependencies. Hence, our approach will focus on single events.

The first problem we have to address is completeness. To build a comprehensive catalog of all alternative splicing events occurring in the transcriptome of an organism, it is necessary to take all possible isoform-structures into account. Such a catalog of all genes and their isoform-structures is called *gene annotation* and exists in several versions for numerous organisms, e.g., the main annotations for human are ENSEMBL [85], the RefSeq database [233] and the UCSC Genome Browser database [130]. For human as well as important model organisms, many isoforms have been experimentally validated and can be considered to be quite accurate. For other organisms, either no annotation exists or most of the genes and isoforms are computational predictions based on sequencing patterns or homology search based on whole genome alignments to better annotated species. It is also to note that many early gene prediction tools only predicted the gene locus but had difficulties to identify single isoforms [258, 270]. Although more recent versions aim at the prediction of several transcript isoforms [271], the task remains computationally hard. Also even quite complete annotations very often still lack isoforms that only occur as a result of a certain internal or external signal, are a product of a gene mutation, e.g., in cancer [37, 94, 237], or only occur within a tightly regulated temporal or spatial pattern and have thus remained undetected. Hence, the available sources of transcript annotations have to be considered as incomplete.

SplAdder addresses this problem by filling in missing alternative splicing information into a given annotation based on one ore multiple RNA-Seq samples. This *augmented annotation* is then used to build a comprehensive catalog of alternative splicing events that can be detected from the annotation and the given data, allowing for a much more complete view onto the current state of the transcriptome.

Building a catalog of events is only the first step towards the ultimate goal of elucidating changes in alternative splicing between samples of several conditions. In the analysis steps subsequent to the augmentation, SplAdder quantifies all detected events, filters them by confidence criteria and finally performs differential testing to identify the most significantly changed events.

Several other tools exist that address single steps of this pipeline. The method JuncBase also identifies alternative splicing events and quantifies them from RNA-Seq data, but is not able to augment the annotation from RNA-Seq data and its applicability in a high-throughput setting is not demonstrated. Further, no differential testing is performed. The method has not been published yet, but has been described in context of a biological application [36]. The tool SpliceGrapher [243] augments the annotation based on RNA-Seq or EST data and produces splice graphs. However, it is not a dedicated approach to detect

Figure 2.16: Schematic of the SplAdder pipeline. The downstream analyses are provided as examples. Any other analysis could follow as well.

and quantify alternative splicing events. Further, there exist very early implementations of custom analyses on EST data [205, 322], but these can only handle a few million sequences in a single input sample. To our knowledge, we provide the first complete pipeline that produces a comprehensive view onto the alternative splicing events within the transcriptome of a given RNA-Seq sample set, that is able to quantify and differentially analyze these events, and that can be applied to thousands of RNA-Seq samples within a high-throughput setting. Figure 2.16 shows a schematic of the SplAdder pipeline.

2.5.2 Splicing Graph Augmentation

Definitions and Notation A given gene annotation can be represented as a set of linear directed graphs. Assume gene g is given and has k different isoforms $j_1, \ldots, j_k \in J_g$, where J_g is the set of all isoforms of gene g. As we consider each gene g individually, we will omit the index g wherever possible in order to keep the notation uncluttered. Each isoform consists of a set of exons that are connected by introns. Each exon can be uniquely identified by its start and its end. We thus represent all exons as coordinate pairs of their start end stop position:

$$v = (\text{start}, \text{stop}) = (v_{\text{start}}, v_{\text{stop}}) \in \mathbb{N}^2.$$

Although further coordinate information like chromosome and strand are used in the program implementation, we will limit this description to an identification by start and stop for simplicity. The exons of each isoform j_i can then be represented as a node set $V_i := \{v_{i,1}, \ldots, v_{i,m_i}\}$ with $1 \leq i \leq k$ and m_i as the number of exons in isoform j_i. As transcripts have a direction (the exons within a transcripts follow a strict order), we require, that the index of the nodes reflects the order of the exons in the transcript. As no two exons in a transcript overlap by definition, this order is implied by v_{start} and v_{stop}. We then define the edge set of isoform j_i as

$$E_i := \bigcup_{1 \leq s < m_i} \{(v_{i,s}, v_{i,s+1}) \mid v_{i,s}, v_{i,s+1} \in V_i\} \subset V_i \times V_i$$

with $1 \leq i \leq k$. The pair (V_i, E_i) forms the directed isoform graph of isoform j_i.

Next, we define the set of exons occurring in *any* isoform j_i as V. As the single exons are uniquely identified by their coordinates, we can write $V := \bigcup_{i=1}^{k} V_i$. Hence, we define the set of all edges as

$$E := \bigcup_{i=1}^{k} E_i \subset V \times V.$$

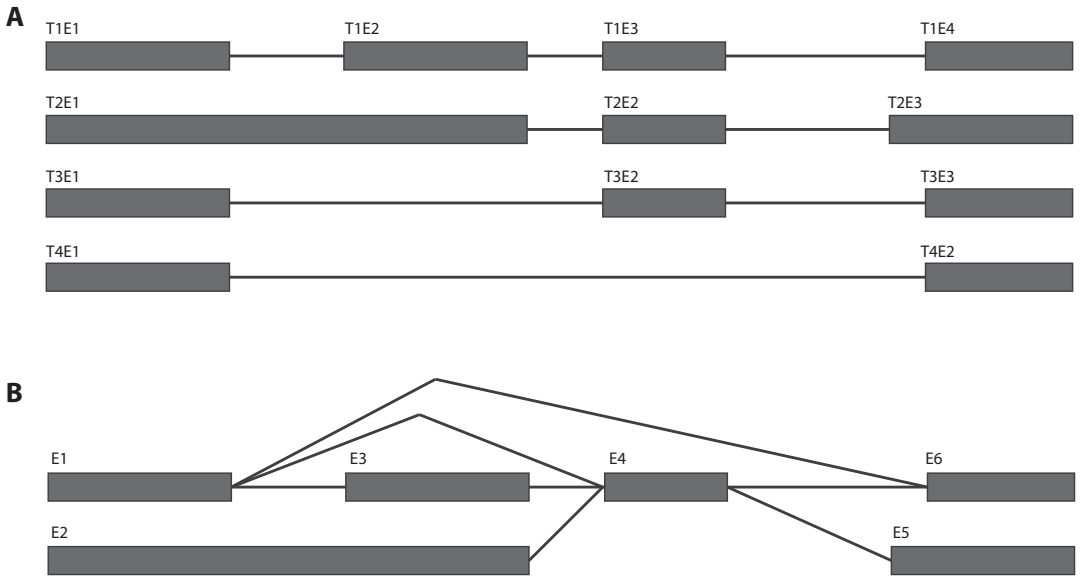


Figure 2.17: Example case for the construction of the splicing graph. ***A:*** Set of four different transcript isoforms. Exons are depicted as gray boxes and introns as solid lines. Labels TiEj denote exon j in transcript i. ***B:*** Splicing graph representation of the same four isoforms. Exons occurring in multiple isoforms are collapsed into a single exon in the graph.

Note that only already existing edges are merged, preserving any existing order of nodes. The pair $G = (V, E)$ is a directed acyclic graph and is called *splicing graph* representation of a gene. Figure 2.17 illustrates how a set of four isoforms is collapsed into a splicing graph.

We define the *in-degree* and the *out-degree* of a node as the number of its incoming and outgoing edges, respectively. We further define a node to be *start-terminal*, if its in-degree is zero and *end-terminal* if its out-degree is zero. Each isoform can now be represented as a path through the splicing graph, beginning at a start-terminal node and ending at an end-terminal node.

Although the splicing graph representation resolves many redundancies and thus can efficiently store large numbers of different but mostly overlapping isoforms, this comes at the cost of information loss. Long range dependencies between single exons are not preserved. An example of this is provided in Figure 2.17, Panel B. Although exon T2E1 exclusively occurs in transcripts that end in exon T2E3, this relationship is lost in the graph, where E2 can connect to both E5 and E6. As will be discussed later, our approach does not suffer from this shortcoming, since we only extract local information about alternative exon- or intron-usage.

The same principle that was applied when collapsing different isoforms that share the same exons into a graph structure, can be applied again to collapse exonic segments that are shared by several exons/nodes of the splicing graph. Following this idea, we divide each exon into non-overlapping segments. Analog to an exon, a segment is uniquely identified by its coordinate pair and the same order as on exons can be applied: $s = (s_{\text{start}}, s_{\text{stop}})$. We say an exon v_i is *composed* from segments $s_{i,q}$ through $s_{i,r}$, if $v_i = s_{i,q} \circ s_{i,r}$, with $q < r$ and

where $\cdot \circ \cdot$ denotes the concatenation of segment positions. Thus, the set of all segments can be defined as

$$S = \bigcup_{v_i \in V} (s_{i,q}, \ldots, s_{i,r} \mid s_{i,q} \circ s_{i,r} = v_i).$$

To explicitly define the set of all segments, at first we define the set V_S of all node-starts in V and the set V_T of all node stops in V. The set of all segments S can then be defined as

$$S = \bigcup_{s_\mathrm{start}, s_\mathrm{stop} \in V_\mathrm{S} \cup V_\mathrm{T}} \{(s_\mathrm{start}, s_\mathrm{stop}) \mid \exists v \in V \colon v_\mathrm{start} \leq s_\mathrm{start} < s_\mathrm{stop} \leq v_\mathrm{stop}\}.$$

The computation of S from V is straightforward. Let P be a sorted array containing all genomic positions that are either start or end of an exon in V. We denote the ith element of the array as $P[i]$. Let L_S and L_E be two binary label-arrays with the same length as P, where $L_S[i]$ is 1 if $P[i]$ is start of an exon in V and 0 otherwise. Analogously, $L_E[i]$ is 1 if $P[i]$ is end of an exon in V and 0 otherwise. Let further C_S and C_E be two arrays with the same length as P, where $C_S[i] = \sum_{j=1}^{i} L_S[i]$ and $C_E = \sum_{j=1}^{i} L_E[i]$ are the cumulative starts and ends up to position i. We can then determine the set of all segments as

$$S = \bigcup_{i=1}^{|P|-1} \{(P[i], P[i+1]) \mid C_S[i] > C_E[i]\}.$$

Analog to the definition of the edges for the splicing graph, we define

$$\begin{aligned} T = \bigcup_{s_u, s_v \in S} \{(s_u, s_v) \mid & \exists v_i \in V, s_r \in S \colon v_i = (s_{r,\mathrm{start}}, s_{u,\mathrm{stop}}) \text{ and} \\ & \exists v_j \in V, s_t \in S \colon v_j = (s_{v,\mathrm{start}}, s_{t,\mathrm{stop}}) \text{ and} \\ & (v_i, v_j) \in E\} \end{aligned}$$

to be the set of segment pairs that are connected by an intron. We then denote the pair $R = (S, T)$ to be the segment graph of a gene. For practical reasons, we store an additional matrix, that relates each node/exon in the splicing graph to the segments it is composed of.

We will use the splicing graph representation to incorporate new information based on RNA-Seq evidence as well as for the extraction of alternative splicing events. However, we will use the segment graph representation for event quantification, as this is computationally much more efficient.

Splicing Graph Augmentation The augmentation of the splicing graph G is a stepwise heuristic. In each step, either a new node or a new edge is added to the graph. If a newly added node shares one boundary with an existing node, the existing edges are inherited by the new node. We will formalize this procedure in the following. We begin by defining the genome $\mathcal{G}$ as a string of consecutive positions $\mathcal{G} = g_1 g_2 \ldots g_n$. Given an RNA-Seq sample and the start g_s and end g_e of a gene, we extract all intron junctions from the alignment, that overlap this region and show sufficient alignment support. Whether an intron junction is sufficiently well supported, is based on a set of given confidence criteria.

These criteria will be discussed later in this section. We define the list of RNA-Seq intron junctions $\mathcal{R}$ as

$$\mathcal{R} = \{(g_i, g_j) \mid s \leq i < j \leq e\},$$

where (g_i, g_j) describes the intron starting at g_i and ending at g_j. An existing node in the splicing graph $v \in V$ will be represented as the tuple of its genomic coordinates $v = (g_x, g_y)$. If we directly access the coordinate tuple of a node, this is denoted by, v_{start} and v_{end}, thus $v_{\text{start}} = g_x$ and $v_{\text{end}} = g_y$. The augmentation process will transform the existing splicing graph $G = (V, E)$ into an augmented version $\hat{G} = (\hat{V}, \hat{E})$. We initialize $\hat{G}$ with G.

Adding Cassette Exons In the first round of augmentation, new cassette exon structures are added to the splicing graph. For this, the algorithm iterates over all non-overlapping pairs of R. For each pair (g_{i_1}, g_{j_1}) and (g_{i_2}, g_{j_2}), the following conditions need to be fulfilled, such that a new cassette exon will be added to the graph:

- $\exists v_i \in \hat{V} : v_{i,\text{end}} = g_{i_1} - 1$ and $\exists v_j \in \hat{V} : v_{j,\text{start}} = g_{j_2} + 1$ and $v_i < v_j$
- $\nexists v_h \in \hat{V} : v_{h,\text{start}} = g_{j_1}$ and $v_{h,\text{end}} = g_{i_2}$

Briefly, both introns need to be attached to existing exons and the cassette exon must not already exist. If all conditions are met, a new node $v_n = (g_{j_1} + 1, g_{i_2} - 1)$ is added to the node set $\hat{V}$ and two new edges (v_i, v_n) and (v_n, v_j) are added to $\hat{E}$. Figure 2.18, Panel A, shows schematically how a cassette exon is added.

Adding Intron Retentions The second augmentation round adds intron retention events to the splicing graph. For each edge $(v_s, v_t) \in \hat{E}$, the algorithm decides if there is enough evidence from the given RNA-Seq sample for expression inside the intron, to consider the intronic sequence as exonic. Again, heuristic confidence criteria are applied that are listed in Appendix A.5. Briefly, the central criteria for adding a new intron retention are the number of sufficiently covered positions within the intron as well as the differences in mean coverage between intronic and exonic part of that regions. In case of sufficient evidence for a retention, a new node $v_n = (v_{s,\text{start}}, v_{t,\text{end}})$ is added to $\hat{V}$. The new node inherits all incoming edges from v_s and all outgoing edges from v_t, thus we get the set of newly added edges

$$E_n = \left\{(x, v_n) \mid \forall x : (x, v_s) \in \hat{E}\right\} \cup \left\{(v_n, x) \mid \forall x : (v_t, x) \in \hat{E}\right\}.$$

Then, the set of edges is updated with $\hat{E} := \hat{E} \cup E_n$. Figure 2.18, Panel B, illustrates this case.

Handle Introns The last augmentation step iterates another time over the list of RNA-Seq supported intron junctions $\mathcal{R}$ that has been generated during the first step. Based on start and end position of the intron, we can test if any existing nodes end or start at these positions, respectively. We have to distinguish four different basic cases: 1) neither start nor stop coincide with any existing node boundary, 2) the intron-start coincides with an existing node end, 3) the intron end coincides with an existing node-start, 4) both the intron-start

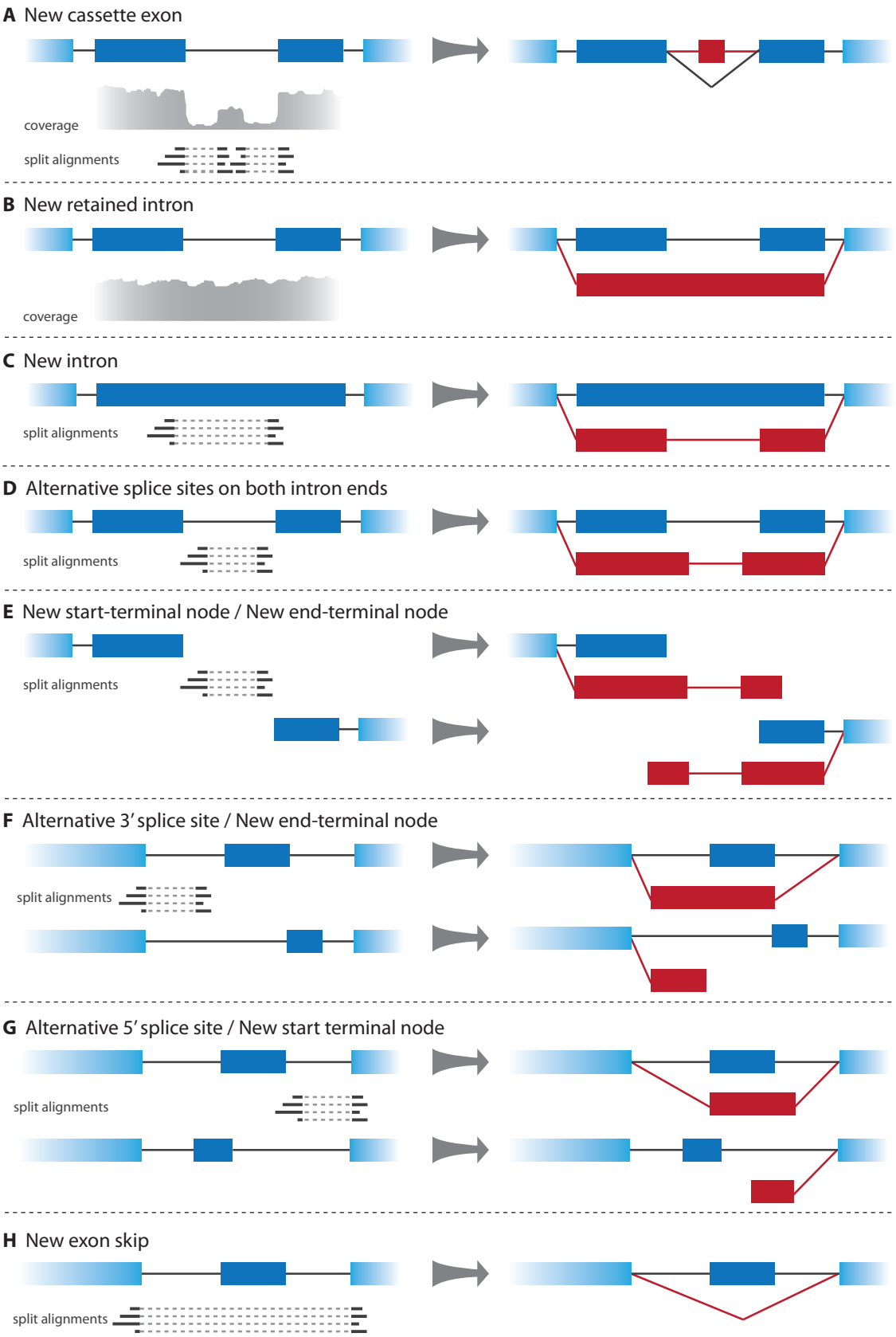


Figure 2.18: Overview of the different classes of splicing graph augmentation. Panels ***A***–***H*** show all possibilities how the splicing graph can be augmented within SplAdder, based on evidence from RNA-Seq alignment data. In cases where no coverage evidence is shown, only junction confirmations by split alignments are used.

coincides with an existing node end and the intron-end coincides with an existing node-start. In the following, we will discuss all four cases in further detail. The four cases and their respective sub-cases are illustrated in Figure 2.18, Panels C–H, which provide a more intuitive explanation and may help the understanding of the following formal definitions.

In the following, we formally define all cases to insert new intron edges into the graph.

1) To handle the first case we can split it into three sub-cases:

   a) If the intron (g_i, g_j) is fully contained within an existing node ($\exists v \in \hat{V} : g_i > v_{\text{start}}$ and $g_j < v_{\text{end}}$), we can insert a new intron into the node, thus creating two new nodes $v_{n_1} = (v_{\text{start}}, g_i - 1)$ and $v_{n_2} = (g_j + 1, v_{\text{end}})$. After adding v_{n_1} and v_{n_2} to $\hat{V}$, we update the edge set to

$$\hat{E} = \hat{E} \cup \{(v_{n_1}, v_{n_2})\} \cup \bigcup_{x \in \hat{V}} \left\{(x, v_{n_1}) \mid (x, v) \in \hat{E}\right\} \cup \bigcup_{x \in \hat{V}} \left\{(v_{n_2}, x) \mid (v, x) \in \hat{E}\right\}$$

   b) If the intron (g_i, g_j) is fully contained within an existing intron, we can connect it to the two nodes v_s and v_t flanking the containing intro, thus introducing two new nodes $v_{n_1} = (v_{s,\text{start}}, g_i - 1)$ and $v_{n_2} = (g_j + 1, v_{t,\text{end}})$ into $\hat{V}$. Again, the new nodes inherit their edges from v_s and v_t providing the following update rule for the edge set:

$$\hat{E} = \hat{E} \cup \{(v_{n_1}, v_{n_2})\} \cup \bigcup_{x \in \hat{V}} \left\{(x, v_{n_1}) \mid (x, v_s) \in \hat{E}\right\} \cup \bigcup_{x \in \hat{V}} \left\{(v_{n_2}, x) \mid (v_t, x) \in \hat{E}\right\}$$

   c) If one of the intron boundaries (g_i, g_j) is in close proximity (we use ≤ 40 nt as a default threshold) to a terminal node, this node is extended to a new node v_{n_1} and a new terminal node v_{n_2} is added to the graph at the other side of the intron. The length k of the new terminal exon is pre-defined to be 200 nt. If the nearby node v is start-terminal, $v_{n_1} = (g_j + 1, v_{\text{end}})$ and $v_{n_2} = (g_i - k - 1, g_i - 1)$ and

$$\hat{E} = \hat{E} \cup \{(v_{n_2}, v_{n_1})\} \cup \bigcup_{x \in \hat{V}} \left\{(v_{n_1}, x) \mid (v, x) \in \hat{E}\right\}.$$

   If the nearby node v is end-terminal, $v_{n_1} = (v_{\text{start}}, g_i - 1)$ and $v_{n_2} = (g_j + 1, g_j + k + 1)$ and

$$\hat{E} = \hat{E} \cup \{(v_{n_1}, v_{n_2})\} \cup \bigcup_{x \in \hat{V}} \left\{(x, v_{n_1}) \mid (x, v) \in \hat{E}\right\}.$$

2) The second case is similar in its handling to case 1c). If the start of intron (g_i, g_j) coincides with the end of an existing node v, we can distinguish two sub-cases.

   a) There exists a node v' in close proximity to intron-end g_j and we can add a new node $v_n = (g_j + 1, v'_{\text{end}})$ and update the edge set to

$$\hat{E} = \hat{E} \cup \{(v, v_n)\} \cup \bigcup_{x \in \hat{V}} \left\{(v_n, x) \mid (v', x) \in \hat{E}\right\}.$$

b) There is no node in close proximity to intron-end g_j, thus we introduce a new end-terminal node $v_n = (g_j + 1, g_j + k + 1)$ and update the edge set to $\hat{E} = \hat{E} \cup \{(v, v_n)\}$.

3) The third case is analog to case 2). If the end of intron (g_i, g_j) coincides with the start of an existing node v in the graph, we again can distinguish two sub-cases.

a) There exists a node v' in close proximity to g_i and we can add a new node $v_n = (v'_{\text{start}}, g_i - 1)$ and update the edge set to

$$\hat{E} = \hat{E} \cup \{(v_n, v)\} \cup \bigcup_{x \in \hat{V}} \left\{ (x, v_n) \mid (x, v') \in \hat{E} \right\}.$$

b) There is no node in close proximity to intron-start g_i, thus we introduce a new start-terminal node $v_n = (g_i - k - 1, g_i - 1)$ and update the edge set to $\hat{E} = \hat{E} \cup \{(v_n, v)\}$.

4) The last case is the most straightforward to handle. If intron (g_i, g_j) coincides with the end of node v and the start of node v', we augment the edge set $\hat{E} = \hat{E} \cup \{(v, v')\}$, if the edge is not already present in $\hat{E}$.

2.5.3 Extraction of Alternative Splicing Events

Starting with the augmented splicing graph $\hat{G} = (\hat{V}, \hat{E})$, we can extract all alternative splicing events as sub-graphs of the splicing graphs:

Exon Skips are all sub-graphs $(V', E') = (\{v_i, v_j, v_k\}, \{(v_i, v_j), (v_j, v_k), (v_i, v_k)\})$ with $V' \subseteq \hat{V}$ and $E' \subseteq \hat{E}$.

Intron Retentions are all sub-graphs $(V', E') = (\{v_i, v_j, v_k\}, \{(v_i, v_j)\})$ with $V' \subseteq \hat{V}$ and $E' \subseteq \hat{E}$ and $v_{i,\text{start}} = v_{k,\text{start}}$ and $v_{j,\text{end}} = v_{k,\text{end}}$.

Alternative 3' Splice Sites are all sub-graphs $(V', E') = (\{v_i, v_j, v_k\}, \{(v_i, v_j), (v_i, v_k)\})$ with $V' \subseteq \hat{V}$ and $E' \subseteq \hat{E}$ and $v_{j,\text{end}} = v_{k,\text{end}}$. This definition assumes the direction of transcription to be positive. For transcripts from the negative strand, the definition for alternative 3' splice site and alternative 5' splice site need to be switched.

Alternative 5' Splice Sites are all sub-graphs $(V', E') = (\{v_i, v_j, v_k\}, \{(v_i, v_k), (v_j, v_k)\})$ with $V' \subseteq \hat{V}$ and $E' \subseteq \hat{E}$ and $v_{i,\text{start}} = v_{j,\text{start}}$. The different strands are handled analogously to alternative 3'-splice sites.

Multiple Exon Skips are all sub-graphs

$$(V', E') = (\{v_i, v_{j_1}, \ldots, v_{j_s}, v_k\}, \{(v_i, v_{j_1}), (v_{j_s}, v_k), (v_i, v_k)\} \cup \bigcup_{l=1}^{s-1} \{(v_{j_l}, v_{j_{l+1}})\})$$

with $V' \subseteq \hat{V}$ and $E' \subseteq \hat{E}$.

The same extraction rules would apply analogously, to extract alternative splicing events from the not augmented graph G. A schematic overview of the extraction process is provided in Figure 2.19.

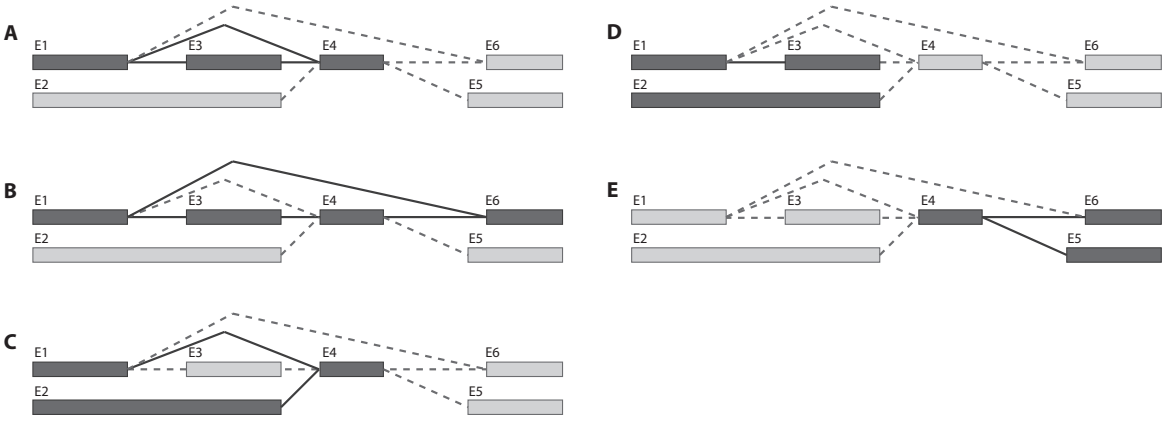


Figure 2.19: Five different types of alternative splicing events are extracted from the splicing graph. The graph structure is given with nodes as gray boxes and edges as solid/dashed lines. Solid/dark parts show the event of interest and light/dashed parts the remainder of the graph structure. ***A:*** Exon skip, ***B:*** Multiple exon skip, ***C:*** Alternative 5' splice site, ***D:*** Intron retention, ***E:*** Alternative 3' splice site.

2.5.4 Event Filtering and Quantification

Alternative splicing events extracted from the graph are filtered at several levels. To remove redundant events, all events are made unique based on their inner event coordinates. The inner event coordinates are defined as the start and end positions of all introns of the event. If two events share the same inner coordinates, they are replaced by a new event with the same inner coordinates but adapted outer coordinates minimizing the total length of the event. An example for this is shown in Figure 2.20. Events in Panel A can be merged, whereas events in Panel B disagree in their inner coordinates and remain separate.

In the next step we use the RNA-Seq data to quantify each of the extracted events. That is, for each intron we count the number of alignments supporting it and compute the mean coverage for each exon. For reasons of computational efficiency, the quantification is performed on the segment graph. As defined above, each segment can be uniquely identified by its genomic coordinates. Thus, we extract for each node its mean coverage and for each edge the number of spliced alignments in the sample confirming this edge. As each exon v_i can be formed through a concatenation of segments $s_q \circ s_r$, we can use the segment-lengths and their average coverage to compute the average coverage of the exon:

$$v_{i,\text{coverage}} = \frac{\sum_{j=q}^{r}(s_{j,\text{stop}} - s_{j,\text{start}} + 1) \cdot s_{j,\text{coverage}}}{\sum_{j=q}^{r}(s_{j,\text{stop}} - s_{j,\text{start}} + 1)},$$

where $s_q \circ s_r$ is the sequence of segments contained in node v_i.

In many applications, the splicing graphs can grow very complex, containing alternative events that are only poorly supported by input data (we provide examples for this in the applications discussed in Section 3.4). Thus, we use the quantifications to further filter the event set and to only retain the most confident events. Each event type has a different set of criteria it has to fulfill in order to become a valid event. A table listing all criteria is

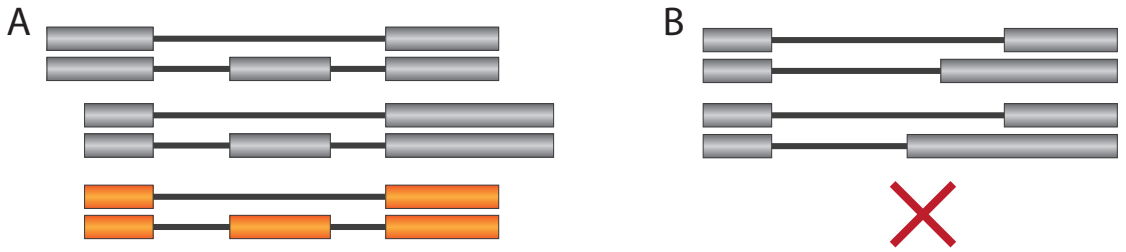


Figure 2.20: Example case when overlapping events can be merged. ***A:*** All inner event coordinates agree and the events can be successfully merged. ***B:*** Both events have only one intron in common, whereas the other introns disagree. The events cannot be merged and remain separate.

provided in Appendix A.5. To determine, if an event is valid, the algorithm checks in which provided RNA-Seq samples which criteria are met. An event is valid, if all criteria are met in at least one sample. To create more stringently filtered sets of events, this threshold can be increased.

2.5.5 Differential Testing

To test for differential usage of alternative events, we use the published tool rDiff [69]. In this context, we treat each alternative event as artificial gene expressing two different isoforms that are defined by the two possible paths through the event-sub-graphs beginning at start-terminal and ending at end-terminal nodes. For instance, the two isoforms of an exon skip event, would be an isoform of three exons, containing the middle exon and an isoform of two exons, skipping the middle exons. The first and the last exon of these two isoforms would be identical. We store all extracted event isoforms in a common event file in GFF3 format that can then be used as input file for rDiff. To account for directionality in the test, that is to identify which of the event's two isoforms was up- or down-regulated, we modified the rDiff output to take the normalized counts of each isoform into account for reporting the final p-value. To this end, we altered the rDiff sourcecode to take the mean expression values of the two tested isoforms into account, when reporting the p-values. We denote an event as *up-regulated*, if the normalized read count of the longer isoform increases between the two tested conditions A and B, and we denote the event as *down-regulated* otherwise. In this context we determine the length of an isoform as the sum of its exonic positions. Depending on the direction of change of the normalized counts, the test p-value is assigned to the respective direction and a value of 1 to the other direction.

2.5.6 Results and Evaluation

We tested SplAdder with both artificial data as well as in application to biological samples. Here, we will focus on the evaluations on simulated data. The performance when applied to several biological datasets and a comparison between augmented and non-augmented annotation is discussed in Chapter 3, Section 3.4.3.

Main goal of this evaluation was to measure how well SplAdder can reconstruct alternative splicing informations from RNA-Seq data if this information is lacking in the annotation. Specifically, we would like to re-construct the same splicing graph where we have access to all isoforms in one case and to only one isoform and additional RNA-Seq data in the other

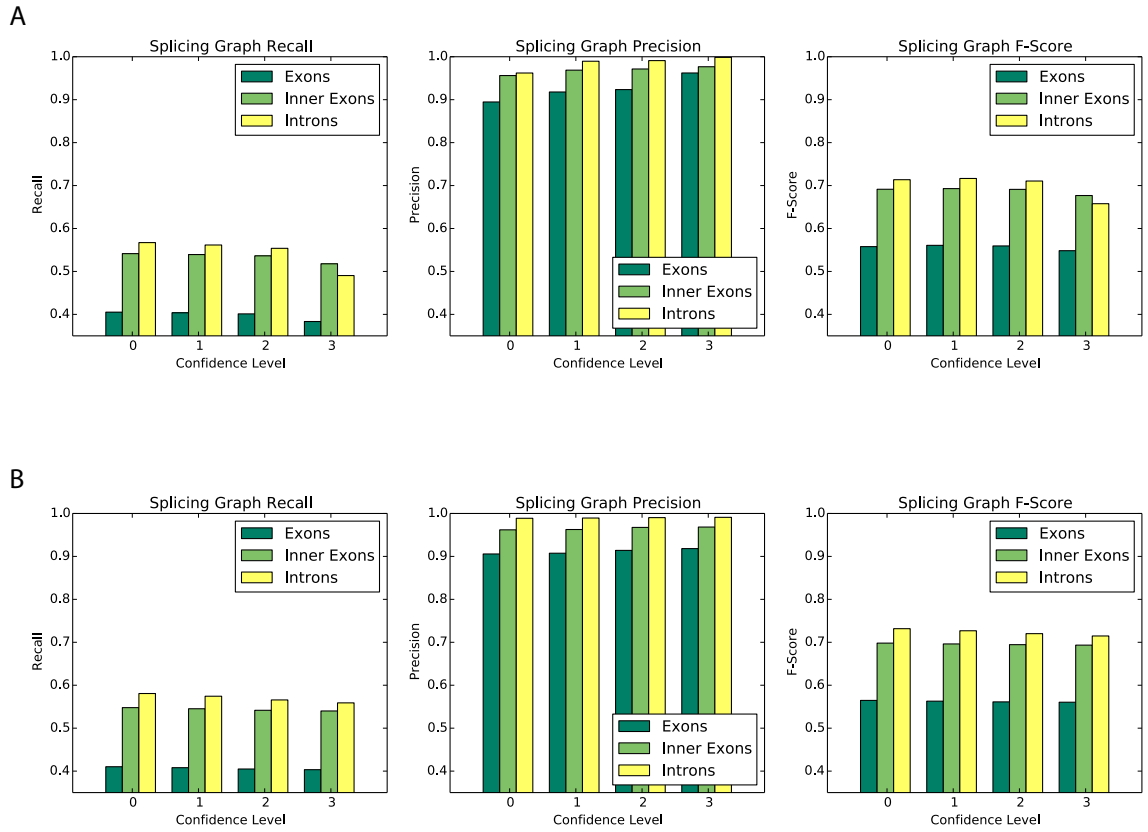


Figure 2.21: Evaluation of the SplAdder performance on artificial data. The two panels show precision, recall and F-score for the prediction of exons, internal exons and introns shown in dark green, light green and yellow, respectively. The predictions are compared to the original annotation the reads were sampled from. ***A:*** Performance measures based on the PALMapper alignments of the reads. ***B:*** Optimal performance based on the ideal alignment set that was created during read simulation.

case. To this end, we took a random set of genes with at least two isoforms and generated RNA-Seq data from all transcripts. We then only kept the first annotated isoform per gene and tried to reconstruct the full splicing graph using SplAdder on the simulated reads. The simulation was implemented as follows.

Pursuing a similar strategy as described in Section 2.4.4, we used FluxSimulator (version 1.1.1-20121103021450) [102] to generate a set of 2×10^6 RNA-Seq reads of 76 nt length. All reads were sampled from a set of 5,000 randomly chosen genes from the ENSEMBL annotation [85]. The reads were then aligned to the hg19 human reference genome with PALMapper, allowing for up to 10 mismatches and at most 2 gaps. To compute an upper performance limit on the given dataset, we also retained the originally sampled reads in BAM format as ideal input data set.

To generate our testing set, we took the set of 5,000 genes the reads were originally sampled from and retained only the first annotated transcript isoform for each gene. This resulted in an annotation without any alternative events. We then used SplAdder to aug-

ment this truncated annotation based on the RNA-Seq read evidence. For several confidence levels, we compared a set of quality measures, describing how similar the splicing graphs generated by SplAdder were to the ones generated from the not truncated original annotation. The evaluation was performed on a subset of 1,491 genes that expressed at least two isoforms. Figure 2.21 shows an overview of the performance evaluation.

We used three different performance measures. The correct augmentation of nodes in the graph is harder than adding single edges. Thus, the exon-level performance measures how many nodes (exons) in the predicted splicing graph match to the graph generated from the not truncated annotation set. Especially terminal exons are difficult to predict, as RNA-Seq alignments are only a poor measure of transcription start- or stop-sites. Hence, the second evaluation measure takes only internal exons into account, as their boundaries can be identified through spliced alignments. The last measure uses the overlap on intron level, which is the easiest task, as the boundaries are completely defined by spliced alignments and existing nodes in the graph.

As shown in Figure 2.21, SplAdder is able to reconstruct almost all intron edges correctly. Notably, with higher confidence levels the precision increases further, which comes at the cost of lower sensitivity, which explains the dropping F-Score measures for more strict filtering.

Although SplAdder detects a large variety of different events, covering a large fraction of the existing variability, there also exist certain limitations to the approach. Especially information regarding transcript starts and ends is difficult to extract from RNA-Seq data due to the coverage slowly running out towards the transcript borders, which makes it difficult to infer specific sites. Also complex events are difficult to evaluate, for instance, in cases where several alternative exons of an exon skip event overlap or in cases of a coordinated retention of introns. For these specific events, approaches that take full transcripts into account might be better suited.

2.5.7 Handling of Multiple Input Files

In all descriptions above, we only discussed how a single sample is used as input. SplAdder is capable of integrating the information of several input files. This is necessary in scenarios with several replicates per sample or if the splicing variation of a whole set of samples should be integrated. We distinguish four different modes to use multiple input files:

1. **Single Graphs:** This mode treats each input file independently and generates the same result as if running SplAdder on each file in a serial manner. That is, all steps are performed on each file, generating one result file per input file.
2. **Merge Graphs:** This mode generates an augmented splice graph for each input file but integrates all these graphs into a common graph representation. It further allows for filtering of the graph, to only retain nodes and edges that are confirmed in a certain fraction of input samples. Events are then detected on the common graph representation but quantified for each input file separately.
3. **Merge Files:** Here, all input files are treated es replicates, merging their information. That is, only a single augmented splice graph is constructed and used for event calling. For the quantification of the splicing events, evidence from all input files is merged.

4. **Merge All:** This is a hybrid mode between merge graphs and merge files. It generates one splice graph per input file as well as a graph constructed from all files at once. All graphs are then integrated into a common representation that is used for event calling. The quantification of events is then performed on the single files again.

These modes can also be combined to create hybrid-strategies, e.g., to create a common splicing graph by merging all input files but quantify each file separately. This can be achieved by providing result files from intermediate steps as input to SplAdder and change the options for the remaining steps.

2.5.8 Implementation and Software

SplAdder has been implemented initially in Matlab/Octave code and was packaged with shell scripts to provide a command line user interface, with no further dependencies than Matlab or Octave. A newer implementation is now also available in Python, resolving the dependency from the Matlab/Octave computing environment. SplAdder relies on standard input formats for the gene annotation (GFF3 format) and the alignments (BAM format). Outputs are provided as plain text files or in HDF5 format. The user interface of the Matlab/Octave implementation showing all available functionality is shown in Appendix A.5. The source code for both implementations is published under GPL license and is publicly available under `https://github.com/ratschlab/spladder`.

3 Applications

In this chapter, we discuss four different projects in which the methods described in the previous chapter have been applied to data from various biological experiments. The first three projects deal with sequencing data from the model plant *Arabidopsis thaliana*, whereas the last project involves data from different human cancer samples. In all *A. thaliana* related projects, we used PALMapper for the alignment and SplAdder for the annotation and quantification of alternative splicing events. The first section discusses the role of alternative splicing in context of the post-transcriptional regulation mechanism of nonsense mediated mRNA-decay (NMD). Based on data from knockdown mutants, we assessed how many alternatively-spliced transcripts are subject to this degradation mechanism [68]. In the second section, we describe our work on *A. thaliana* plants with a mutations in polyrimidine tract binding protein homologs (PTBs) that led to aberrations in splicing patterns and thus revealed functional roles of PTBs for the splicing of flowering regulators [246]. In the subsequent section, we describe the results of a large-scale analysis of two *A. thaliana* populations grown at different temperatures with the aim to identify expression and splicing quantitative trait loci (eQTL and sQTL, respectively) and investigate their effects within different environments. The work described in the last section discusses the analysis of whole transcriptome sequencing samples of more than 4,000 cancer patients. We used SplAdder in a large scale manner to identify and quantify alternative splicing events as phenotypes that were then associated with somatic as well as germline genetic alterations in these patients.

Author Contributions All studies described in this chapter have been conducted in collaborations of either small groups or within larger multi-institutional consortia. Here, we describe which parts where genuinely contributed by the author of this work (AK) and how the remaining work was split. The two studies on NMD and PTB were conducted in collaboration with Andreas Wachter (AW), Gabriele Drechsel (GD), Christina Rühl (CR), Eva Stauffer (ES), Anil K. Kesarwani (AKK), Jonas Behr (JB), Philipp Drewe (PD), Gabriele Wagner (GW) and Gunnar Rätsch (GR). For the research on NMD, AW, GD, AK and GR designed the project, GD, AKK, ES and AW carried out biological experiments and provided the data, AK and GR conceived the computational analysis strategy, AK implemented and designed the computational analysis pipeline, carried out all RNA-Seq alignments, performed alternative event quantification and differential analysis and implemented, performed the NMD feature analysis and carried out functional analyses on the candidate events, PD contributed to the adaptation of rDiff and JB provided predictions of non-coding and intergenic transcripts. For the research on PTB, GR, CR, ES and AW designed the experimental setup, CR, ES, GW, GD and AW performed the biological experiments and provided the data, GR and AK conceived the computational analysis strategy, AK implemented and designed the analysis pipeline, performed all RNA-Seq alignments and data quality controls, characterized alternative splicing events and performed the differential analysis. Both studies resulted in peer-reviewed publications [68, 246].

The presented work on detecting sQTL in two populations of *A. thaliana* is part of an international collaboration with multiple other groups. We will only mention people relevant for the work presented here: Magnus Nordborg (MN), Pei Zhang (PZ), Richard M. Clark (RMC), Robert Greenhalgh (RG), Edward J Osborne (EJO), Bjarni Vilhjalmsson (BV), Oliver Stegle (OS), Philipp Drewe (PD), Yi Zhong (YZ) and Gunnar Rätsch (GR). The data for the CEGS population was provided by MN and PZ, whereas the data for the MAGIC population was provided by RMC, EJO and RG. In both cases this included collection of biological material, preparation and sequencing. MN, RMC, GR and OS conceived the idea of the study. AK and GR designed the alignment pipeline. AK implemented and processed the RNA-Seq alignment, implemented and processed the alternative event detection and quantification, performed read counting and filtering for the expression analysis and carried out the sQTL analyses. YZ helped with a parameter study for the alignment and PD suggested filtering criteria for the expression counting. OS and BV provided code that was used for the linear mixed model analysis. EJO implemented and carried out the eQTL analyses.

The analysis of 12 different cancer types to detect eQTL and sQTL and determine splicing aberrations in cancer was a collaborative effort together with Kjong-Van Lehmann (KL), Gunnar Rätsch (GR), Cyriac Kandoth (CK), William Lee (WL), Nikolaus Schultz (NS), Oliver Stegle (OS) and The Cancer Genome Atlas research network (TCGA). All raw sequencing data was provided by TCGA. KL, GR, OS and AK conceived the study. GR and AK performed alignments and carried out alignment quality control. KL, CK, WL and AK performed variant calling. KL and AK designed and implemented the full data processing and association pipeline. AK implemented the detection and quantification of alternative splicing events, carried out the necessary quality filtering, implemented the pipeline to generate gene expression counts used for the eQTL analysis and performed the analysis of alternative splicing diversity over cancer types. OS provided efficient low-level code for the mixed model analysis used in the sQTL and eQTL analyses. NS provided a comprehensive list of cancer relevant genes.

3.1 Evaluation of Nonsense-mediated mRNA-Decay in *Arabidopsis thaliana*

Regulation of transcription is a complex process that not only involves various protein and RNA factors during mRNA synthesis but also processes that degrade transcriptional products. The most important mRNA degradation mechanism is nonsense mediated mRNA-decay (NMD, cf. Section 1.1), that not only helps to degrade products from pseudogenes [202], transposons [199] or certain non-coding RNAs [149] but also plays an important role in the degradation of physiological transcripts resulting in a major regulatory potential. Numerous factors involved in NMD have been identified over the past years. A small set of proteins was found to be conserved over almost all eukaryotic species: the UP FRAMESHIFT proteins UPF1, UPF2 and UPF3. To investigate the role of NMD in transcriptional regulation and further understand how it can be triggered by alternative splicing, we created *A. thaliana* plants lacking essential NMD factors and studied the effect onto the transcriptome. Although single NMD factors have been knocked out in other organisms [311], neither existed a study in which several NMD factors had been knocked out

nor had a knockout-study with whole-transcriptome assessment been conducted in plants. In the following, we give a detailed description of our study design, the setup of our computational pipeline and the results of our analysis. However, we will focus on the application of the computational methods described earlier, as these are the contributions of the author. For a more detailed introduction into the NMD mechanism and its biological relevance and further details for biological methodology, we refer to our own work [68] as well as to three excellent reviews [45, 51, 186].

3.1.1 Study Design

To elucidate the role of alternative splicing in triggering NMD and how prevalently splicing products undergo degradation in *A. thaliana*, we investigated mutant plants in the following setting. Based on the cross of two existing mutant lines, *low-beta-amylase1* (*lba1* or also *upf1*) [326] and *upf3-1* (*upf3*) [113], we created a new double-mutant line lacking both factors UPF1 and UPF3, further denoted as *upf1upf3*, and compared it to the two single mutant lines. The double-mutants were found to be arrested in early seedling development (cf. Figure 3.1). As NMD is deficient in these mutants, we expected an accumulation of transcript isoforms that usually would undergo degradation. In many eukaryotes, NMD is a translation-dependent process, requiring a pioneer round of translation [186]. For comparison, we measured splicing in wild-type samples treated with the translation inhibitor cycloheximide (CHX) that simultaneously triggers an accumulation of NMD transcript isoforms. A wild-type sample treated with water instead of CHX was added as control. For all samples, we created TruSeq RNA-seq libraries, that subsequently underwent single-end high-throughput sequencing on an Illumina GA II, resulting in approx. 50–60 $\cdot 10^6$ 100 nt reads per library. For a full list of read statistics, see Appendix A.6. Each sample was created in biological duplicates, resulting in a total of 12 sequence libraries with two samples for each wild-type (*wt*), *upf1*, *upf3*, *upf1upf3*, CHX *chx* and the CHX control *mock*, respectively. All read data has been submitted to the gene expression omnibus (GEO) and is available under the accession GSE41432. An overview of genotypes and phenotypes of the mutant lines is provided in Figure 3.1. An initial verification experiment confirmed the accumulation of isoforms that are known to be degraded by NMD in the NMD-deficient mutants (Figure 3.1, Panels C and D).

3.1.2 Analysis Pipeline

In the following, we describe the computational pipeline that we developed for this analysis. An overview of all tools and their dependencies is given in Figure 3.2.

Alignment and post-processing We aligned all reads to the *A. thaliana* TAIR10 reference genome [151] using PALMapper (version 0.5) in non-variant-aware alignment mode, allowing for a moderate number of 6 edit operations and at most 1 gap. Due to remaining adapter sequences in the reads, we trimmed 4 nt from each side of a read for efficient mapping. We additionally allowed local alignments of the read by trimming non-mappable portions down to a minimal read length of 40 nt. We used junctions derived from the TAIR10 genome annotation, but also allowed for the discovery of novel junctions. The full

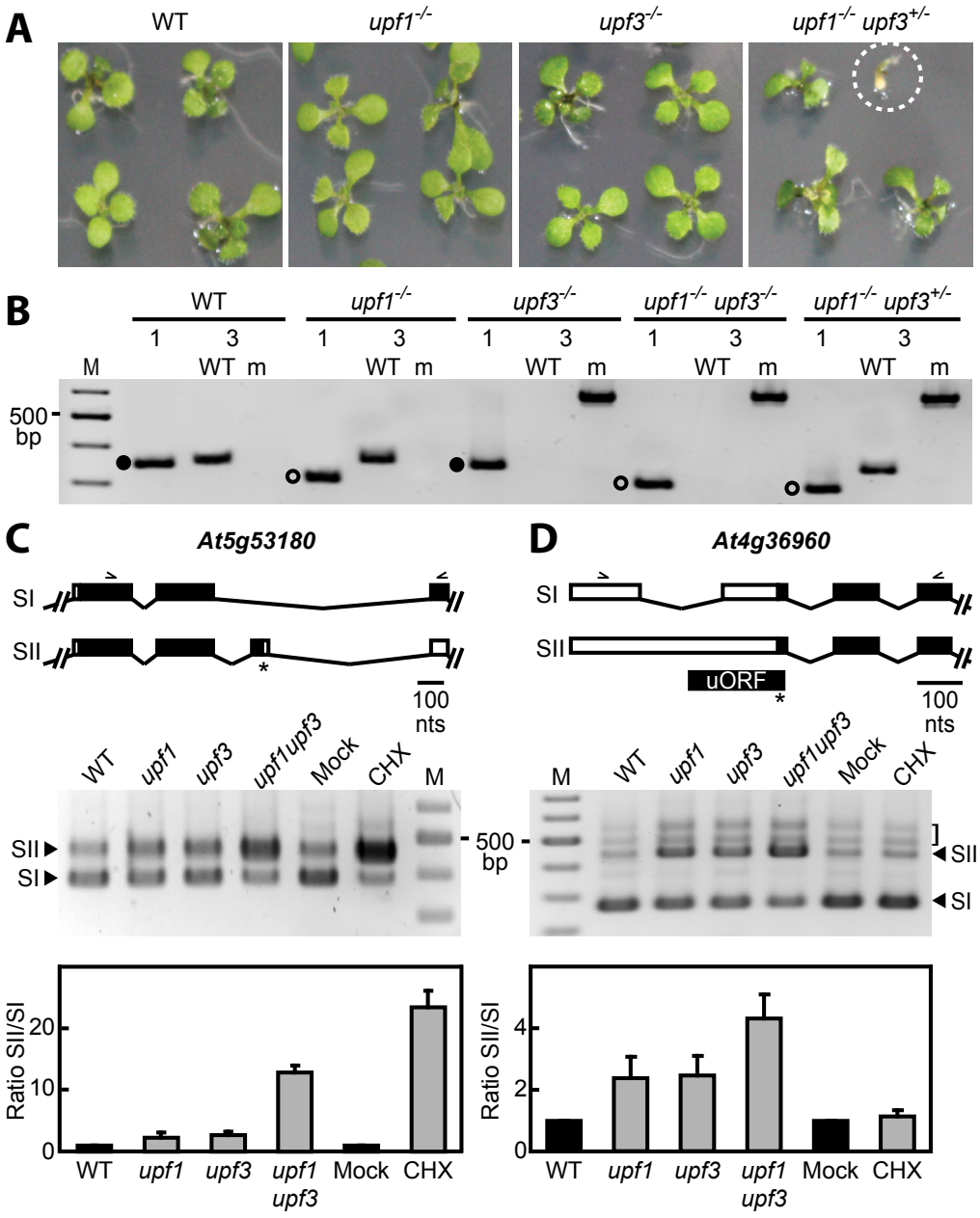


Figure 3.1: Overview of experimental setup and mutant genotypes and phenotypes. ***A:*** Phenotype of wild-type (WT) and mutant samples (*upf1*, *upf3*, *upf1upf3*). Most right: 25% of the cross are homozygous in a loss of *upf3* in a upf^--background, generating the double-mutant (dashed circle). ***B:*** PCR-based genotyping. Filled/open circles mark UPF1 wild-type/mutant alleles in lane 1. In the shared lane 3, WT/m columns mark the wild-type/mutant allele for UPF3. ***C/D:*** Verification of two NMD isoforms found in earlier studies. Top row shows partial gene models. Isoforms SII are NMD targets and accumulate in mutant samples. Bottom row shows RT-qPCR quantification results for all 6 sample types. (Figure has been adapted from [68] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

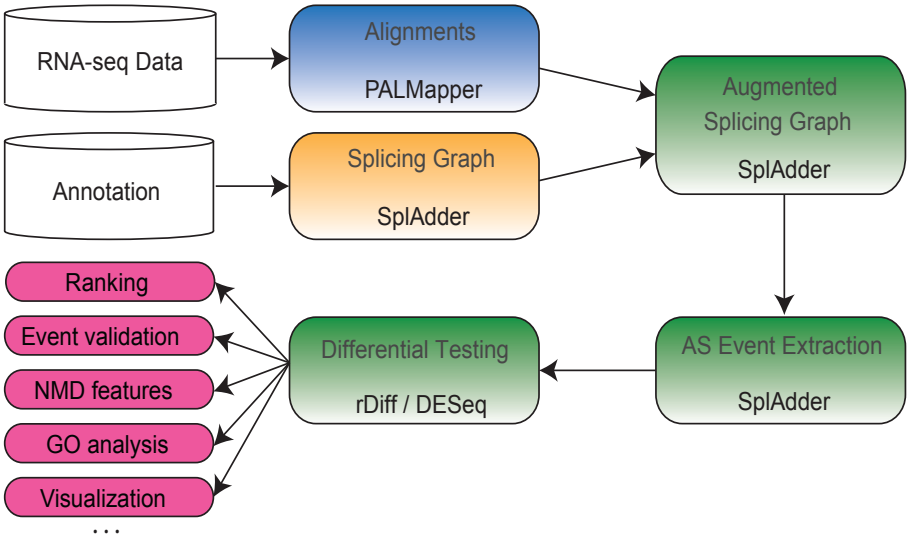


Figure 3.2: Computational pipeline used for the here-described analysis. RNA-Seq data were aligned with PALMapper and subsequently a splicing graph was constructed and augmented using SplAdder. Differential testing was performed with rDiff [69] and DESeq [115]. Only PALMapper and SplAdder are discussed in this work, for all other tools we refer to the respective publications. (Figure has been adapted from [68] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

list of alignment parameters can be found in Appendix A.6. With this strategy, we aligned between 60 and 80% of the reads per sequencing sample, resulting in 30–80 million aligned reads (cf. Appendix A.6 for complete statistics). All alignments were sorted and indexed using SAMtools (version 0.1.12a, [167]). To resolve ambiguous read mappings, all reads were subsequently filtered with MMR, retaining at most one alignment per read, reducing the total number of alignments by up to 25%.

Splicing graph augmentation and event extraction We merged the alignment files for the two replicates of each sample to increase signal strength for alternative event detection and used SplAdder to augment the splicing graph that had been created from the TAIR10 annotation. We used SplAdder with its default parameters, but adapted parameters for a more confident read-filtering. The full list of non-default parameters is provided in Appendix A.6. Running SplAdder in the `merge-graphs` mode, we augmented a splicing graph for each sample independently and subsequently integrated all graphs into a single graph. From this graph, alternative splicing events of four different types were extracted: exon skips, intron retentions, alternative 3'-splice sites and alternative 5'-splice sites. As described in Section 2.5.4, SplAdder automatically cleans the event list of duplicate events and performs initial filtering. Additional filtering of the events was performed based on the confidence-level we chose (3).

Differential analysis All filtered alternative events were then subjected to differential testing using the rDiff toolbox [69]. To accommodate our needs, we adapted the rDiff program as described in Section 2.5.5 to not only return a p-value but also the direction of change. Thus, it was possible to determine, if the expression of an isoform significantly increased or decreased. As rDiff only supports pairwise testing, we tested the following four sample pairs: *wt* vs. *upf1*, *wt* vs. *upf3*, *wt* vs. *upf1upf3* and *mock* vs. *chx*. For the test, the replicates were not merged, but instead used to estimate the overdispersion for the negative binomial test of rDiff. For the resulting p-values, we computed false discovery rates (FDR) following the method of Benjamini and Hochberg [27]. We deemed all events significant that showed significantly different isoform usage between *wt* and *upf1upf3* with an FDR below 0.1. As we were interested in events that showed already a tendency in the single mutants but a much stronger effect in the double-mutant, we further filtered the list of events according to the following criterion. We retained all events that were also identified as significant in one of the two single mutants, but showed a change in the same direction as in the double-mutant. As we were only interested in tendencies, we used the uncorrected p-value of 0.1 as significance threshold in the single mutants, which is correct when only used to further filter the initial list with an FDR below 0.1 in the double-mutant.

NMD feature analysis To assess the effect of the alternative event on the isoform-context, we reintegrated each event into the representative transcript-isoform of the gene the event originated from. The representative isoform is usually the most commonly expressed isoform and is provided with the TAIR10 annotation. An event was only integrated, if the inner event coordinates overlapped in at least one base position of the representative isoform and if information about the codings sequence (CDS) of the transcript was available. Otherwise, the event was discarded. Hence, we could annotate each event with a gene location label, describing the gene part it was altering, either 3'-UTR, CDS or 5'-UTR. Additionally, depending on the direction of change, each of the two isoforms of an event was assigned the label ΔNMD or *control*, if we could observe or could not observe an accumulation upon NMD impairment, respectively. Depending on its assigned gene location label, each event was evaluated according to the following criteria, which are known to be descriptive of NMD targeted isoforms [186, 187]:

5'-UTR

- existence of an upstream open reading frame (uORF)
- a uORF longer than 35 amino acids
- a uORF overlapping an annotated start codon

CDS

- existence of a premature termination codon (PTC) in combination with a 3'-UTR length, larger than the 90-th percentile of *A. thaliana* 3'-UTRs (347 nt)
- existence of a splice junction more than 50 nt downstream of the stop codon

3'-UTR

- existence of a 3'-UTR longer than the *A. thaliana* 90-th percentile (347 nt)
- existence of a splice junction more than 50 nt downstream of the stop codon

PTCs were detected as the first in-frame stop codon when starting at the annotated CDS start. If a stop-codon earlier than the annotated stop was found, we marked it as PTC. For a more detailed description, we refer to the supplementary material of [68].

3.1.3 Results

Running all SplAdder analyses at the highest confidence-level 3, we retained a total of 41,941 alternative splicing events after filtering, containing 10,139 intron retentions, 4,400 exon skips, 18,006 alternative 3'-splice site events and 8,946 alternative 5'-splice site events. After testing and FDR correction, 3,361 events remained that showed significantly different isoform usage between wild-type and double-mutant. The filter taking into account performance in the single mutants as well as directionality, removed 1,743 events from the list, retaining 1,618 events. The number of significant events by event type is listed in Table 3.1. Additionally, we identified 3,238 events that showed significantly different event isoforms between CHX and water-treated control plants. As both sets, the 1,618 events significantly different in the mutants as well as the 3,238 events differing upon CHX-treatment were biologically interesting, we defined the union of these two sets, containing 3,872 events, as *high confidence NMD events* that were used for further NMD feature analysis. As an additional step of validation, we randomly chose a set of 10 events that were predicted to be differentially expressed between wild-type and double-mutant with an FDR < 0.3. Using RT-qPCR, we were able confirm for 9 out of 10 events that one isoform accumulated in the mutant but not in the wild-type. For more details on the validation, we refer to our publication [68].

Table 3.1: Overview of significantly different events. Significant events are shown for two conditions. To fulfill criterion 1, an event had to be significant in the test *wt* vs. *upf1upf3* with an FDR < 0.1. To fulfill criterion 2, in addition to fulfilling criterion 1 an event had to be also significantly different in either *wt* vs. *upf1* or *wt* vs. *upf3* with a p-value < 0.1 and the change in the same direction as in the double-mutant.

Event Type	Number of Events	Criterion 1 fulfilled	Criterion 2 fulfilled
Alt 5'	8,946	685	343
Alt 3'	18,006	1,376	696
Exon Skip	4,400	533	294
Intron Ret.	10,139	767	285
Total	41,941	3,361	1,618

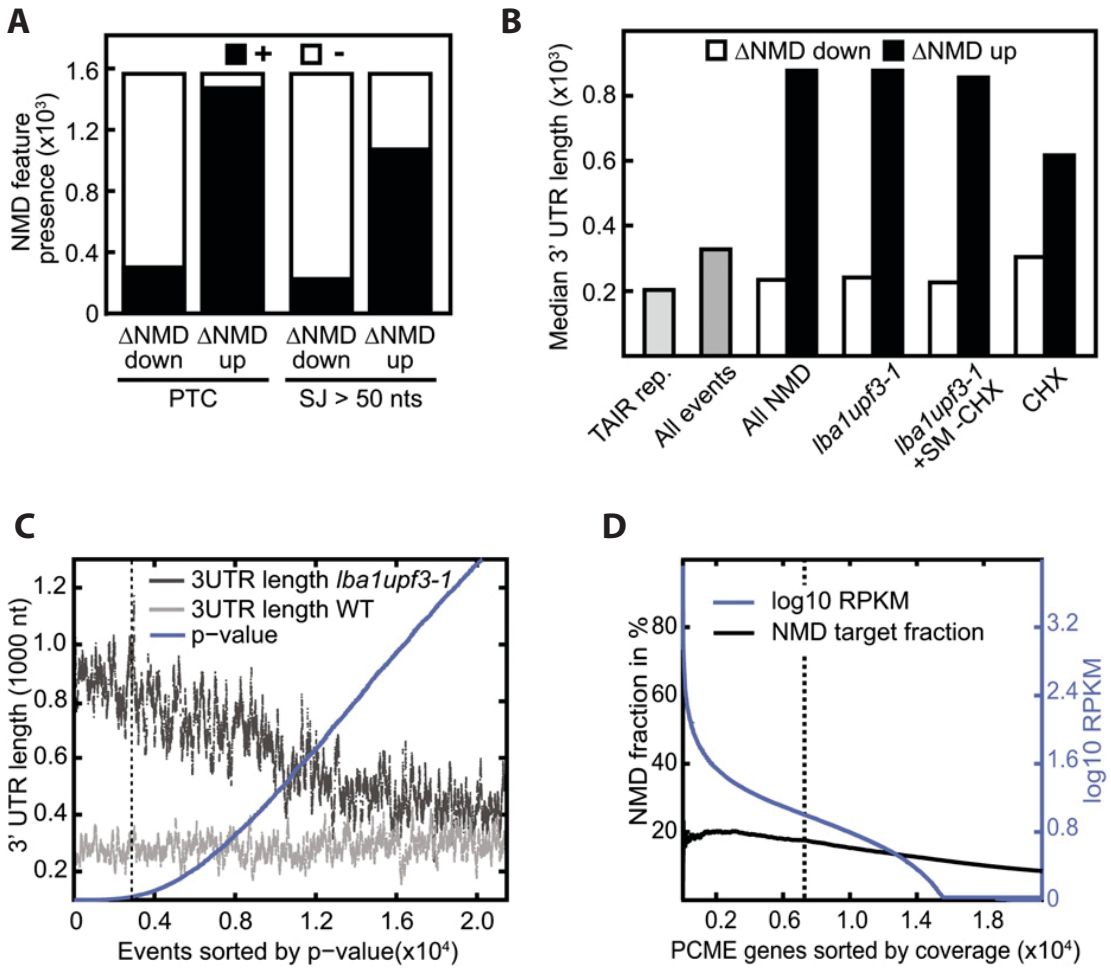


Figure 3.3: Enrichment of NMD features in ΔNMD isoforms. ***A:*** Comparison of isoforms accumulating upon NMD impairment (ΔNMD-up) according to the two NMD features PTC and distant splice junction. ΔNMD-up isoforms are highly enriched for presence of an NMD feature (black, +). Absence of a feature is marked white (–). ***B:*** Median 3'-UTR length for the set of confident NMD events vs. the background sets of all TAIR10 representative isoforms and the set of all events. Accumulating isoforms (ΔNMD-up) show significantly longer 3'-UTRs compared to not accumulating isoforms (ΔNMD-down). ***C:*** Dependency of 3'-UTR length and event FDR. Events with lower FDR show a more pronounced difference in the length of the 3'-UTRs of their isoforms. The dashed line marks an FDR of 0.1. ***D:*** Fraction of protein coding multiple-exon (PCME) genes, that show at least one NMD target isoform (black curve). Events were sorted by expression in reads per kilobase per million mapped reads (RPKM, shown in blue). The dashed line demarcates an expression value of 10 RPKM. (Figure has been adapted from [68] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

Labeling the events by gene location assigned 5.3% of the events to the 5'- or 3'-UTRs and the remainder to the CDS. Strikingly, for almost all of the events located in the CDS the ΔNMD isoform acquired a premature termination codon. Also the event set overlapping to the UTRs showed an enrichment for NMD features in the ΔNMD isoforms. Figure 3.3, Panel A, shows the NMD feature enrichment in CDS events for PTCs and splice junctions more than 50 nt downstream of the stop codon.

We further investigated the enrichment of long 3'-UTRs in ΔNMD isoforms. Generally, we saw long 3'-UTRs overrepresented in accumulating isoforms (cf. Figure 3.3, Panel B). To investigate, if there was a relationship between significance of the NMD event and the UTR-length, we sorted all events by FDR-value and assessed the length of the 3'-UTR in the respective isoform context. The result is shown in Figure 3.3, Panel C, confirming the drastically increased 3'-UTR length for NMD target isoforms and the dependency between FDR and UTR length. Interestingly, even events very far from the significance threshold still showed an increased length of the 3'-UTR, suggesting that our estimates of effected events are quite conservative.

Lastly, we estimated what fraction of all genes contained NMD-targeted transcript isoforms. We based our analysis on the set of all protein coding multiple-exon (PCME) genes in *A. thaliana*. After sorting the set of PCME by gene expression, ranking most-expressed genes the highest, we computed for each position in the list, what fraction of genes up to that point contained at least one high confidence NMD-target event. We found that of all genes with an expression value of at least 10 reads per kilobase per million mapped reads (RPKM) 17.5% contained NMD target isoforms. This fraction is even larger for more highly expressed genes, where our estimate is expected to be more stable. The fraction of NMD targeted PCME genes depending on their expression value is shown in Figure 3.3, Panel D.

3.1.4 Conclusion

In this work on *A. thaliana* mutant plants, deficient in the degradation mechanism nonsense mediated decay, we have shown that a substantial fraction of all expressed genes produced isoforms that were targeted by NMD. We were able to identify numerous alternative splicing events that created isoforms specifically accumulating in the mutant, showing that they are usually subject to degradation. We further found known features of NMD-transcripts to be drastically overrepresented in the accumulating isoforms and found striking differences in 3'-UTR lengths of degraded vs. non-degraded isoforms. Besides the biological insights, we could show that our pipeline is both sensitive and sufficiently specific to accurately identify alternative splicing events linked to NMD. In validation experiments, we could confirm most of the detected changes, emphasizing the high accuracy of SplAdder.

3.2 Analysis of Splicing Alterations in PTB-deficient *Arabidopsis thaliana*

Alternative splicing is a tightly regulated process that involves various protein- and RNA-factors as well as sequence elements both proximal and distal to the splice sites (cf. Section 1.1 for a more detailed introduction). Along with other factors, heterogeneous nuclear ribonucleoproteins (hnRNPs) are important players in this regulatory process that interact with sequence elements in *cis* to influence the splicing outcome. Originally, these were thought to be pure splicing repressors, but recent work suggested a more context dependent behavior [178, 324]. One example for such factors are polypyrimidine-tract binding proteins (PTBs) that bind to mRNA sequence motifs rich in pyrimidines [250, 300]. While PTBs are well-studied in animal systems, our work focuses on the plant homologs. *Arabidopsis thaliana* has three known PTB homologs: *At3g01150* (PTB1), *At5g53180* (PTB2)

and *At1g43190* (PTB3) that have been shown to be auto- and cross-regulated [272, 300] by altering their own isoform pattern into a variant bearing a premature termination codon. This leads to a degradation via the nonsense mediated decay pathway. The aim of this work is to identify other targets in the *A. thaliana* transcriptome that are regulated by one or several of the PTB homologs. For a more detailed discussion of PTB in plants and a broader introduction, we refer to the publication of this study [246]. In the first part of this section, we describe our experimental setup and give some background on the generated data. The second part provides an overview of the computational analysis pipeline, thereby putting a focus on the tools that were described in Chapter 2. Subsequently, we give an overview of the main findings of our analysis and discuss the alternative events we identified and the functional implications of our findings. Finally, we discuss our results and put them into a broader context.

3.2.1 Study Design

Central aim of this study was the transcriptome-wide identification of targets of the three PTB homologs in *A. thaliana*, PTB1, PTB2 and PTB3. To achieve the most pronounced effect for differential analysis, we generated two different kinds of mutant plants that were either producing additional PTB protein through over-expression of the respective gene, or had a lowered PTB production due to a partial gene knockdown. We generated three over-expression mutants *OE1*, *OE2* and *OE3*, by transfecting a construct with the full coding sequence of the respective PTB homolog into the plant cells. We used the coding sequence instead of the full gene sequence to prevent previously reported auto-regulatory feedback, counteracting over-expression via splicing into an NMD targeted isoform [272]. The partial knockdown mutants have been generated through specific artificial microRNA (amiRNA) constructs [221, 257]. Four different amiRNAs were used. The three knockdown mutants *ami1*, *ami2* and *ami3* have been created with amiRNAs specific for the respective PTBs. Further, we created a double knockdown mutant *ami1ami2* using an amiRNA specificically downregulating PTB1 and PTB2 but not PTB3. The genetic background for all mutants was the Columbia-0 *A. thaliana* reference strain that was also used for the wild-type sample, *wt*. Each experiment was performed in biological duplicates. A quantitative analysis of PTB expression in a subset of the mutant replicates is provided in Figure 3.4. For each sample, a sequencing library was prepared that was sequenced on an Illumina Genome Analyzer II, resulting in $40–60 \cdot 10^6$ reads per sample. Sequencing of samples has been spread over several runs. A list of which samples were sequenced together and how many reads per sample were sequenced is provided in Appendix A.7. All read data has been submitted to the gene expression omnibus (GEO) and is available under the accession GSE41433.

3.2.2 Analysis Pipeline

The computational tools used for this analysis are similar to the components of the pipeline discussed in the previous Section 3.1 and are summarized in Figure 3.2. However, several adaptations were made to accommodate the specific needs of this study. As no adapter sequences were present in the reads, we could align all data without additional trimming.

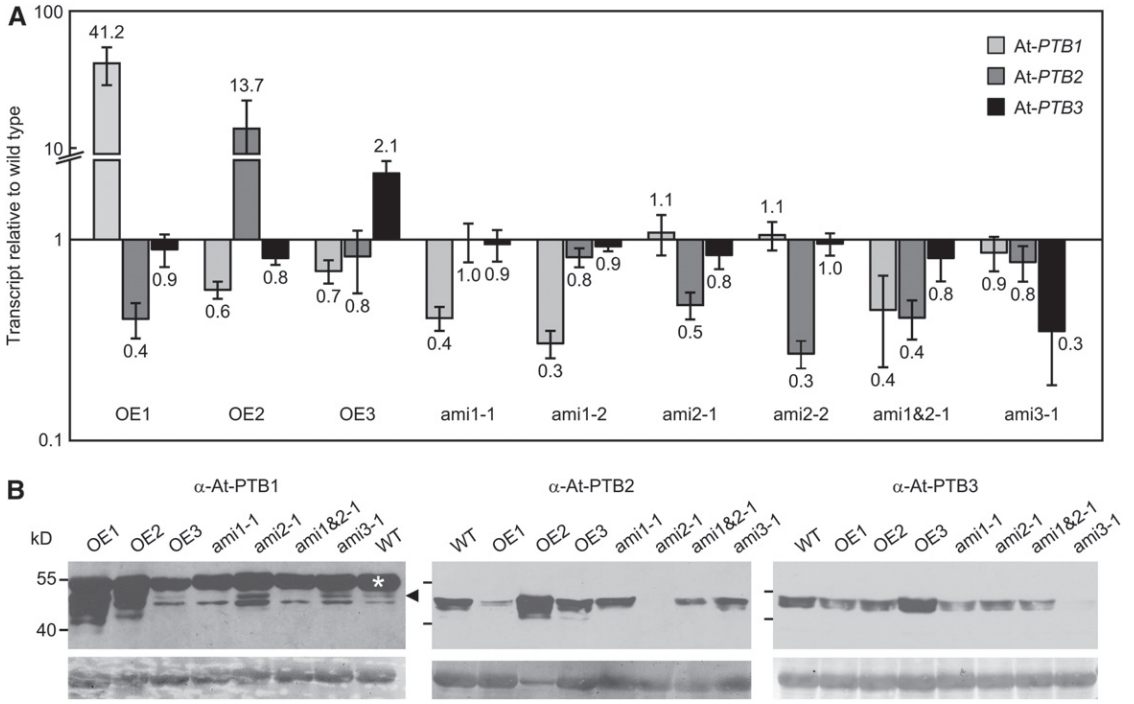


Figure 3.4: Overview of experimental setup and mutant genotypes and phenotypes. ***A:*** Quantification of PTB transcripts in 10-day-old seedlings of the mutant plant by RT-qPCR. Displayed values are mean values from 3 replicates normalized to wild-type and relative to a reference transcript. ***B:*** Immunoblots for protein product quantification with PTB specific antibodies. Leftmost blot for PTB1 shows an unspecific cross-reaction with RBCL protein (white asterisk); PTB1 signal is indicated by a black arrow. (Figure taken from [246] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

Alignment and Post-processing All reads were aligned using PALMapper (version 0.5), allowing for up to 6 edit operations and 1 gap. Additional junction information from the TAIR10 [151] annotation was used to guide spliced alignments. The full list of all parameters is provided in Appendix A.7. In total, we could align 42–60 $\cdot 10^6$ reads. A comprehensive overview of alignment statistics is provided in Appendix A.7. After sorting and indexing with SAMtools (version 0.1.12a) [167], ambiguous read mappings were resolved with MMR that was used with standard settings over three iterations.

Splicing Graph Augmentation We used SplAdder to transform the TAIR10 annotation into a splicing graph and augmented the graph for each sample type individually, merging sample replicates together to generate a stronger signal. SplAdder was used with default settings in confidence level 3. Custom settings are listed in Appendix A.7. All individual splicing graphs were then merged into a common splicing graph that was used for further analysis (using SplAdder-mode `merge-graphs`). From that graph we extracted exon skip, intron retention, alternative 3'-splice site and alternative 5'-splice site events as described in Section 2.5.3. Each event was then quantified in all samples by counting the

support of introns by spliced read alignments and the support for exons as mean read coverage over that exon. We used the SplAdder internal filter at confidence level 3 to reduce the set of events to a high-confidence subset.

Differential Analysis Analog to the analysis described in Section 3.1, we used rDiff [69] for the differential analysis of the events. For each event we performed five different pairwise tests: *ami1ami2* vs. *OE1*, *ami1ami2* vs. *OE2*, *wt* vs. *OE1*, *wt* vs. *OE2*, and *wt* vs. *ami1ami1*. Again, we performed a directed test such that we could detect whether an isoform accumulated significantly or was significantly depleted. Hence, our testing scheme benefited from the opposite effect of up- and down-regulated PTB mutants and could sensitively detect isoforms that showed opposite changes after the respective up- and down-regulation. We called an event *up-regulated*, if the expression of the longer isoform increased in the second condition for a test (condition1 vs. condition2) and otherwise *down-regulated*. The mean-variance relationship to estimate over-dispersion due to biological variance was estimated on the two *ami1ami2* replicates. As the single samples were sequenced in different flow cells, we formed replicate pairs in such a manner that no two replicates of a pair were sequenced in the same flow cell, accommodating possible lane effects. The resulting p-values were corrected for multiple hypothesis testing using the method of Benjamini and Hochberg [27] for computing a false discovery rate (FDR). We combined the results of all five tests to determine a confident set of significantly altered events. Hence, we deemed an event significant, if

- its p-value in the test *ami1ami2* vs. *OE1 (OE2)* was not greater than 0.005 and
- its p-value for *ami1ami2* vs. *OE2 (OE1)* showed the same direction of change and was not greater than 0.6 and
- its p-value for *ami1ami2* vs. *wt* showed opposite direction and was not greater than 0.6 and
- its p-value for *wt* vs. *OE2 (OE1)* in the opposite direction was equal to 1.

For further details, we refer to the Supplemental Methods in [246]. Again the weaker p-value cutoffs of 0.6 were only used to further subset the of significant events from the test of opposite mutants.

NMD Feature Analysis The analysis of transcript features known to be related to NMD was performed as described in 3.1.2. Briefly, both isoforms of each event were integrated individually into the representative isoform of the respective gene, generating two full transcripts. Based on the full transcript, we determined to which gene part the event overlapped (3'-UTR, CDS or 5-'UTR) and which NMD features could be confirmed, e.g., presence of premature termination codons (PTC) or existence of a splice junction more than 50 nt downstream of the stop codon (for a complete list cf. 3.1.2).

3.2.3 Results

Using the pipeline described above, we identified 26,076 alternative splicing events in total that SplAdder was able to confirm as expressed in the tested samples. Table 3.2 provides an

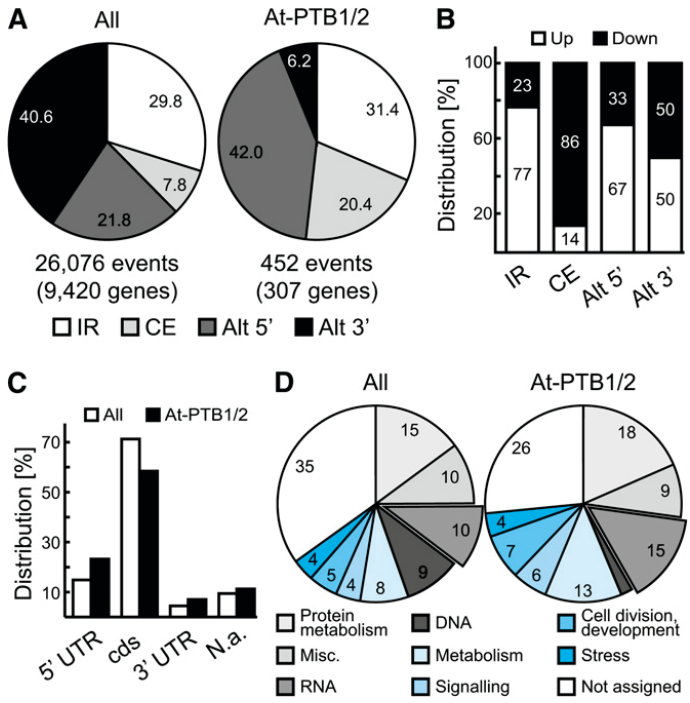


Figure 3.5: Alternative splicing events detected in the PTB analysis. ***A:*** Pie chart of all detected (left) and significantly altered (right) AS events. ***B:*** Isoform specific distribution of significant AS events. Skipped exons (CE) and retained introns (IR) are over-represented. ***C:*** Distribution of event locations within the gene structure. ***D:*** Overview of Gene Ontology categories, assigned to all events (left) and to the 452 significantly altered events (right). (Figure taken from [246] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

overview on the composition by event type. Based on the testing scheme above, we identified a set of 452 events in a total of 307 genes that showed significantly different isoform usage between down-regulated and up-regulated mutants with a consistent direction of isoform expression change (cf. Figure 3.5, Panel A, and Table 3.2). Interestingly, we found the skipping of exons and the retention of introns to be significantly over-represented for the events altered upon PTB misexpression (Figure 3.5, Panel B), with a fraction of 86% skipped exons and 77% retained introns. Analysis of event location showed a slight accumulation

Event Type	**# of Events**	**% of total**	**Signif. altered**	**% of altered**
Alt 3'	10,591	40.6	28	6.2
Alt 5'	5,680	21.8	190	42.0
Exon Skip	2,024	7.8	92	20.4
Intron Retention	7,781	29.8	142	31.4
Total	26,076	100	452	100

Table 3.2: Overview of all detected events and all events significantly altered after PTB expression perturbation, distinguished by event type.

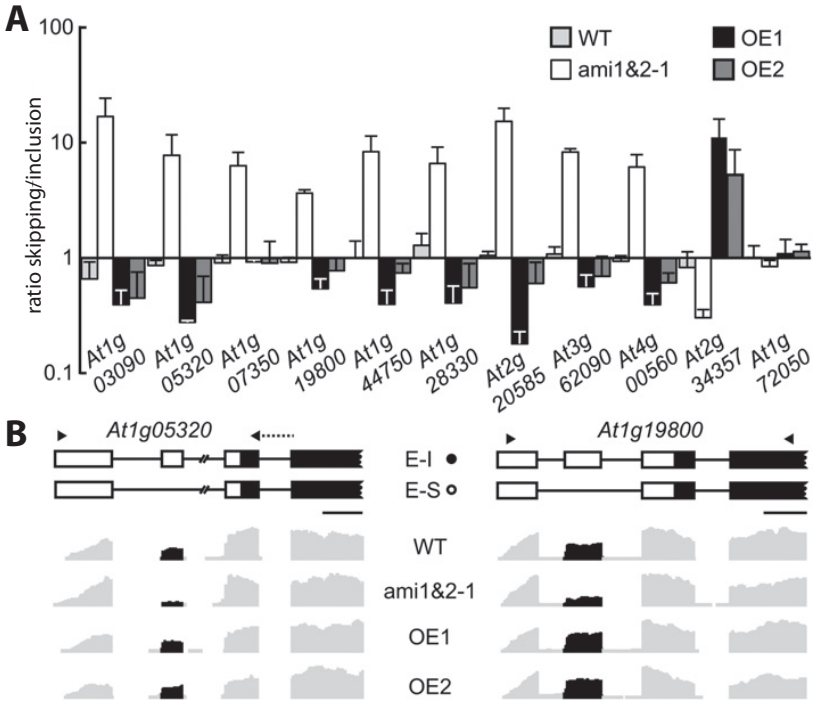


Figure 3.6: Validation of significantly altered exon skip events. ***A:*** Isoform expression quantification by RT-qPCR for 10 exon skip events and 1 control (rightmost) plotted as ratio of exon skipping over inclusion. ***B:*** Coverage profiles for 2 exon skip events along with corresponding event structure show the reciprocal effect of the mutants. Primers are marked by black arrows. (Figure has been adapted from [246] with permission. Copyright American Society of Plant Biologists, www.plantcell.org.)

of events in the 5'-UTR compared to the background distribution of all events (Figure 3.5, Panel C). NMD feature analysis found that 72.3% of all significantly altered events produced transcript isoforms that showed features likely to trigger NMD (cf. Supplemental Table 3 in [246]). An analysis of Gene Ontology (GO) terms showed over-representation of the term *RNA* and the sub-term *RNA processing*, supporting possible functional roles of PTB in the regulation of transcription. An overview on the distribution of all GO terms is provided in Figure 3.5, Panel D.

Experimental validation of ten exon skips that were marked as significantly altered by our analysis, resulted in the confirmation of reciprocal changes in nine events. For the event not showing reciprocal changes (*At1g07350*), we suspect that the PTB effect already saturates in the wild-type and does not experience further changes in the over-expression mutants. As expected, a previously characterized exons skip event that was not marked as differential in our analysis, did not show aberrant expression. Figure 3.6 shows an overview of the validation results.

Further functional studies on genes selected from the list of 307 genes harboring differential AS events revealed PTB regulation in the context of seed germination through differential splicing of *PHYTOCHROME INTERACTING FACTOR6* (*PIF6*) and in the context of flowering through PTB dependent splicing of *FLOWERING LOCUS K* (*FLK*)

and *FLOWERING LOCUS M* (*FLM*). For an in-depth discussion of the functional implications, we refer to our publication [246].

3.2.4 Conclusion

Our work on *A. thaliana* PTB homologs identified numerous potential target genes that produce alternative splicing variants in an PTB1 or PTB2 dependent manner. This work is one of the first comprehensive whole transcriptome analyses of splicing factor targets in *A. thaliana* and has shown that complex splicing patterns arise in a splicing factor dependent manner. We identified PTB dependent alternative splicing in genes relevant for numerous biological processes and showed functional implications for seed germination and flowering. From a computational point of view, we have shown that the analysis pipeline presented in Section 2.5 and applied in Section 3.1 is not tailored to a specific application, but can be used in a broader context. The results were robust and could be confirmed by independent biological validation experiments.

3.3 Identification of Splicing QTL in two *Arabidopsis thaliana* populations

In this section, we discuss the large-scale analysis of two populations of *Arabidopsis thaliana*, to uncover genotype–phenotype relationships in an environment dependent context. This project is part of a collaborative effort between groups from the Gregor Mendel Institute in Vienna, the University of Southern California, Oxford University, the European Bioinformatics Institute, the University of Utah and Memorial Sloan Kettering that is ongoing for the past three years. First, we describe the experimental setup to generate one artificial as well as one natural mapping population. Second, we summarize the pipeline of computational analyses including the variant-aware alignment of the RNA-Seq data, quantification of gene expression, the detection and quantification of alternative splicing events and the association analysis. Although much effort has been spent on data-preprocessing and quality control, we will put more focus on the generation and analysis of expression and splicing phenotypes for the identification of splicing/expression quantitative trait loci (sQTL/eQTL). In the third part, we will provide an overview of the identified alternative splicing events and discuss results of the QTL analyses. Lastly, we summarize our findings and discuss our methodological contributions.

3.3.1 Study Design

We worked with two different *A. thaliana* mapping populations. The first population, further denoted as *CEGS*, consisted of 163 natural accessions that were collected in Sweden, reflecting a broad spectrum of natural variation due to very different environmental conditions over the geographic north-south-spread of the country that is a good proxy for variation found globally [180]. Still, this population showed only little structure caused by inter-relatedness when compared to a set of globally collected accessions. The second population, further denoted as *MAGIC* (multiparent advanced generation inter-cross), was a synthetic mapping population of 203 strains constructed through a multi-parental cross of 19 different founder strains, globally selected from representative populations for a wide

spread of phenotypes [92, 147]. Although such populations show a much more pronounced linkage-structure (coupled inheritance of variants) than a natural population, they provide a controlled framework of parental haplotypes and show much better resolution for genetic mapping than bi-parental crosses due to a higher density of recombination events [147]. Individuals of both populations were reared under controlled growth conditions in long day lighting. To investigate effects caused by gene-environment interactions, the accessions were grown under low and high temperature conditions: 10 °C and 16 °C for CEGS and 20 °C and 30 °C for MAGIC. Non-strand-specific RNA-Seq libraries were generated for each individual in both temperatures of the CEGS population and were then sequenced on an Illumina GAIIx machine, yielding single-end reads of 36 nt length. The read data has been submitted to the gene expression omnibus (GEO) under submission ID GSE54680. For the MAGIC lines non strand-specific RNA-Seq libraries for both temperatures were prepared and sequenced on Illumina machines at the Oxford Genomics Centre, yielding paired-end reads of 2×100 nt length. The MAGIC read data will be made publicly available with publication of the research paper covering this work [219]. Genotype information for the CEGS lines was generated as described in [180], whereas the MAGIC genotypes were imputed by assessing the RNA-Seq data at the variant positions of the founder strains determined in [92]. For an in-depth description of the complete experimental design and all quality control procedures, we refer to the original publication of this work [219].

3.3.2 Analysis Pipeline

With the goal to compare the genetic architecture of the two different mapping populations, it was central to our analysis that the sequencing data from both populations was analyzed in an identical manner. Both alignment strategy and alternative event calling have been uniformly applied to both population. In the following, we provide an overview of our analysis pipeline.

Variant-aware Alignment To generate a sensitive and yet specific alignment set, we compiled information from earlier analyses of the MAGIC founder strains and preliminary alignments of the CEGS data to integrate it into our analysis. As described in Section 2.1.5, spliced alignment accuracy can be improved by junction-remapping, if a list of trusted intron locations is provided as input. To this end, we took all introns that had been identified in the mapping of the 19 founder strains in [92] of the MAGIC population and added them to a list of trusted intron junctions, if the corresponding alignment had at most 3 edit operations, the intron was shorter than 100,000 nt, at least two reads confirmed the junction and the minimal segment length of such reads was at least 6 nt. We compiled one list of trusted junctions for each of the founder strains. For the CEGS alignments, we proceeded similarly. There, we performed an initial round of alignments to detect novel junctions. The parameters for that initial alignment round are provided in Appendix A.8. As the reads of the CEGS samples were only 36 nt in length, splice junctions were only added to a list of trusted junctions if at least 5 reads confirmed the intron with a minimal segment length of 8. The two lists of trustworthy junctions from CEGS and MAGIC alignments were then combined with junctions annotated in the TAIR10 genome annotation [151] and everything was compiled into a combined list that was used for the final alignment run. In addition to the junction information, we collected variant information for all strains from

previous studies. For the MAGIC lines, we took the variant information made available in [92] and for the CEGS lines the variants published in [180]. All variants were integrated into a common list. We then ran PALMapper in its variation-aware alignment mode using the common list of trusted junctions and the common variant file. All samples of both populations were aligned in a uniform manner. The full list of parameters is provided in Appendix A.8. To keep the analyses comparable, the reads of the MAGIC set were trimmed down to the 32 nt length of the CEGS set (4 nt of the 36 nt were adapter-sequence and trimmed during alignment). For the purpose of comparison also a full length alignment set of the MAGIC data was created, which will be further denoted as *MAGIC (untrimmed)*. As an elementary step of alignment post-processing, we used MMR to resolve ambiguous read alignments.

Expression Quantification Based on the alignment data from the previous step, we generated expression counts for various combinations of filter settings. This filtering was introduced to produce an as good as possible expression estimate that was only weakly confounded by splicing structure and differences in genome assembly quality. We counted expression for all genetic elements annotated in the TAIR10 gene annotation enriched with gene structures newly identified in [92], resulting in 65,238 counted entities. Counting was realized with a custom Python script, that counted a read as overlapping to an exon if it overlapped with at least 1 nt to the genomic positions covered by the exon. Depending on the respective filter setting, a read could be excluded from counting if

1. it overlapped to any position that was intronic in all annotated isoforms,
2. it fully fell into a genomic region that had more than one gene annotated,
3. it started at a genomic position that was not present in any of the founder genomes of the MAGIC analysis,
4. it started at a genomic position that was not sufficiently covered for assembly in one of the lines or
5. it fully fell into a genomic repeat region.

We generated single count sets for each of the filter criteria but also sets combining several criteria. For all further analyses described in the following, the most stringently filtered set was used, which combined all five criteria.

Detection and Quantification of Alternative Splicing Events We used the SplAdder pipeline to generate splicing graphs for each of the read libraries of MAGIC and CEGS sets. All single graphs were then integrated into a common graph, representing the total splicing complexity observed in the complete dataset. The confidence of the graph was increased by filtering all edges that were not present in TAIR10 and had spliced read support in less than five libraries. From this confident graph we than called the complete list of splicing events, consisting of exon skips, intron retentions and alternative 3'- and 5'-splice sites. Each event was subsequently quantified in all MAGIC and CEGS strains. If replicates were available for any strain, they were merged to improve the signal. SplAdder was run with confidence level 2 (cf. Appendix A.5). The event list was further filtered, using the SplAdder confidence criteria.

Association using a Linear Mixed Model Using the alternative splicing pattern of single events as phenotypes in connection with the genotypes provided for each of the two populations, enabled us to find associations between the splicing of single events and specific genotypes. We computed the splicing phenotype of an event as the percent spliced in (PSI) value of the two event isoforms. The PSI is computed as the ratio of splice evidence of the longer isoform over the combined splice evidence of both isoforms, resulting in a value between 0 and 1 that describes at what fraction the longer isoform is observed in the mixture of both isoforms (cf. Appendix A.9 for details). As PSIs are computed as ratios, the values become unstable for very low intron counts. Therefore, we only computed PSI values for events that had support of at least 10 spliced counts and assigned NaN otherwise. We further only kept events that had NaN values in less than 80% of the strains in the respective mapping population and each temperature condition. When using the linear mixed model, it is assumed that the phenotype approximately follows a Gaussian distribution. However, this is not the case for the PSI values, that are either mostly 1 or 0 and only a subset of samples deviates. To overcome this problem, we transformed the phenotype using an inverse normal rank standardization. This transformation replaces the PSI values by their ranks in the sorted list of PSI values and uses an Inverse Gaussian distribution on the rank list to generate a normally distributed list of values. As many of the PSI values of a single event were identical for many strains, causing ties in the ranked list, we applied a process known as jittering. That is, we added a very small fraction ($< 10^{-5}$) of random noise to the PSI signal to break up the ties, thus creating a list of continuous ranks. To account for structure within the mapping population, we computed an IBD matrix and used it as cofactor in the association model (cf. Section 1.3.2 for further details on association studies). For efficient computation of the linear mixed model, we used the LIMIX toolbox for Python [174]. We ran one association experiment independently for each combination of mapping population, environmental condition (temperature) and event type, resulting in 16 different test sets.

3.3.3 Results

The alignments with PALMapper were very sensitive for most of the libraries, with a median alignment rate of 92% for both CEGS and untrimmed MAGIC populations and a top alignment rate of 98%. The trimmed MAGIC data aligned even more sensitively, as only the first 32 nt of the reads were kept and the quality usually only drops towards the end. This resulted in a median alignment rate of 98.9% and a top alignment rate of 99.6%.

Although expression counting and subsequent eQTL analysis has been performed on the alignment data, we will focus here on the presentation of the sQTL, as this analysis was contributed by the author. For all further analyses based on the alignments, we refer to the original publication of this work [219].

To facilitate accurate association analysis, we enforced stringent filter criteria on the alternative splicing events detected through SplAdder. We used two different levels of filtering. The first level was the SplAdder internal event confirmation, asserting that each isoform of the event can be validated in at least one sample of the population. For the second level, we required, that at least 10 spliced alignments confirm the two isoforms of the event, such that a stable PSI-value could be computed, otherwise we assigned a NaN-value to the event. We further required that an event has a NaN-value in at most 20% of the samples. Table 3.3 provides an overview on the number of events we detected in the

Table 3.3: Overview of detected alternative splicing events for the two different *A. thaliana* populations, including the alternative alignment version of the MAGIC set. As the splice graphs the events are called from are built from both populations at once, the number of called events for CEGS and MAGIC (trimmed) is identical. However, events are confirmed within the respective populations, leading to a different number of events passing the filters.

CEGS			
Event Type	**Events Detected**	**Events After Filter 1**	**Events After Filter 2**
Alt 3'	5,974	946	233
Alt 5'	3,467	392	92
Exon Skip	1,361	172	41
Intron Ret.	5,451	510	175
Total	10,620	3,209	541

MAGIC (trimmed)			
Event Type	**Events Detected**	**Events After Filter 1**	**Events After Filter 2**
Alt 3'	5,974	1,668	530
Alt 5'	3,467	747	215
Exon Skip	1,361	352	127
Intron Ret.	5,451	827	272
Total	10,620	3,594	1,144

MAGIC (untrimmed)			
Event Type	**Events Detected**	**Events After Filter 1**	**Events After Filter 2**
Alt 3'	9,849	5,604	3,203
Alt 5'	7,879	3,056	1,525
Exon Skip	3,446	1,122	728
Intron Ret.	10,396	5,815	3,221
Total	31,570	15,597	8,677

single populations and how the two filter levels affected this number. Only events passing both filters were used for sQTL analysis.

sQTL Analysis on the Datasets CEGS and MAGIC (trimmed) In total we found 21 and 94 sQTL in the CEGS and MAGIC (trimmed) populations, respectively, to be significantly associated with a genetic variant in low temperature environment and passing the genome-wide significance threshold of 0.05 after Bonferroni-correction. For the high temperature environment we identified 14 and 60 sQTL in CEGS and MAGIC, respectively. Most of these associations were in *cis*, that is the genetic alteration was located within or in close proximity to the alternative splicing event. For the low temperature environment, 6 of the identified sQTL in MAGIC (trimmed) were located in *trans* compared to 4 for the CEGS population. Again, values for the high temperature setting were comparable

with 2 and 5 *trans*-sQTL for CEGS and MAGIC, respectively. Most of them were only weakly significant after Bonferronie-correction and need experimental follow-up for validation. Interestingly, we observed that the associated variants are often in close proximity to transposable elements. A summary of the sQTL identified in the different populations and environments is provided in Table 3.4. We gained higher confidence in our *cis*-association results, when we found that 9 sQTL replicated over both populations, which are 50% of *cis*-sQTL found in CEGS and 9% of *cis*-sQTL found in MAGIC. One example, a *cis*-sQTL where an alternative 3'-splice site within the gene *AT1G31580* is significantly associated with a SNP within or close to the same gene, is shown in Figure 3.7. Due to blocks of variants that are in strong linkage disequilibrium, it is difficult to identify one most probable causative SNP. Variants within such blocks tend to be inherited together and thus a whole set of SNPs shows strong values of significant associated. This effect is even stronger for the MAGIC population, due to its artificial genetic architecture cf.. Figure 3.7, Panels A and B). To evaluate the calibration of our model, we analyzed the distribution of p-values, that should largely follow a uniform distribution and only deviate for the significantly associated variants (cf. Figure 3.7, Panels C and D). Although the computational part of the association study has been completed, functional analysis and possible validations are still ongoing.

sQTL Analysis on the Dataset MAGIC (untrimmed) In the untrimmed dataset for MAGIC, the set of tested events was much larger as more events were detected due to better data quality. For low temperature we identified 921 *cis*- and 32 *trans*-sQTL that were significantly associated with a genetic marker, after Bonferroni-correction. Interestingly, for high temperature, we identified less sQTL in *cis* (672) but more in *trans* (69). An overview of sQTL by event type is shown in Table 3.4. Of the 8,677 tested events, 864 were also detected in the trimmed data. We further found, that of the 94 and 60 sQTL identified with the trimmed data in the low and high temperature environment, respectively, 50 and 29 were also significantly associated in the untrimmed setting. Interestingly, only 1 of the *trans*-sQTL identified with the trimmed dataset was re-discovered in the untrimmed set, suggesting that *trans*-sQTL in both datasets should be critically evaluated as they might be false positives. A two-dimensional overview of event-locations vs. associated variant positions for both trimmed and untrimmed MAGIC data is shown in Figure 3.8. The larger amount of significant associations for the untrimmed data (left) compared to the trimmed data (right) is evident. Whereas *cis*-sQTL appear on the diagonal, *trans*-sQTL are scattered throughout the genome.

sQTL in Different Temperatures Another interesting result is the comparison of detected sQTL between the two different environmental conditions. We assessed how many sQTL were shared between temperatures and how many temperature-specific events we could detect. An overview for all populations is provided in Table 3.5. In almost all cases, we found a lower number of significant sQTL for the high-temperature environment. The only exception were *trans*-sQTL in the untrimmed MAGIC population, that were increased in the higher temperature. We also found that *cis*- and *trans*-sQTL replicate differently. Whereas, *cis*-sQTL where often shared for different environments, we found only two *trans*-sQTL that were significantly associated in both high and low temperature, *AT3G57630* pro-

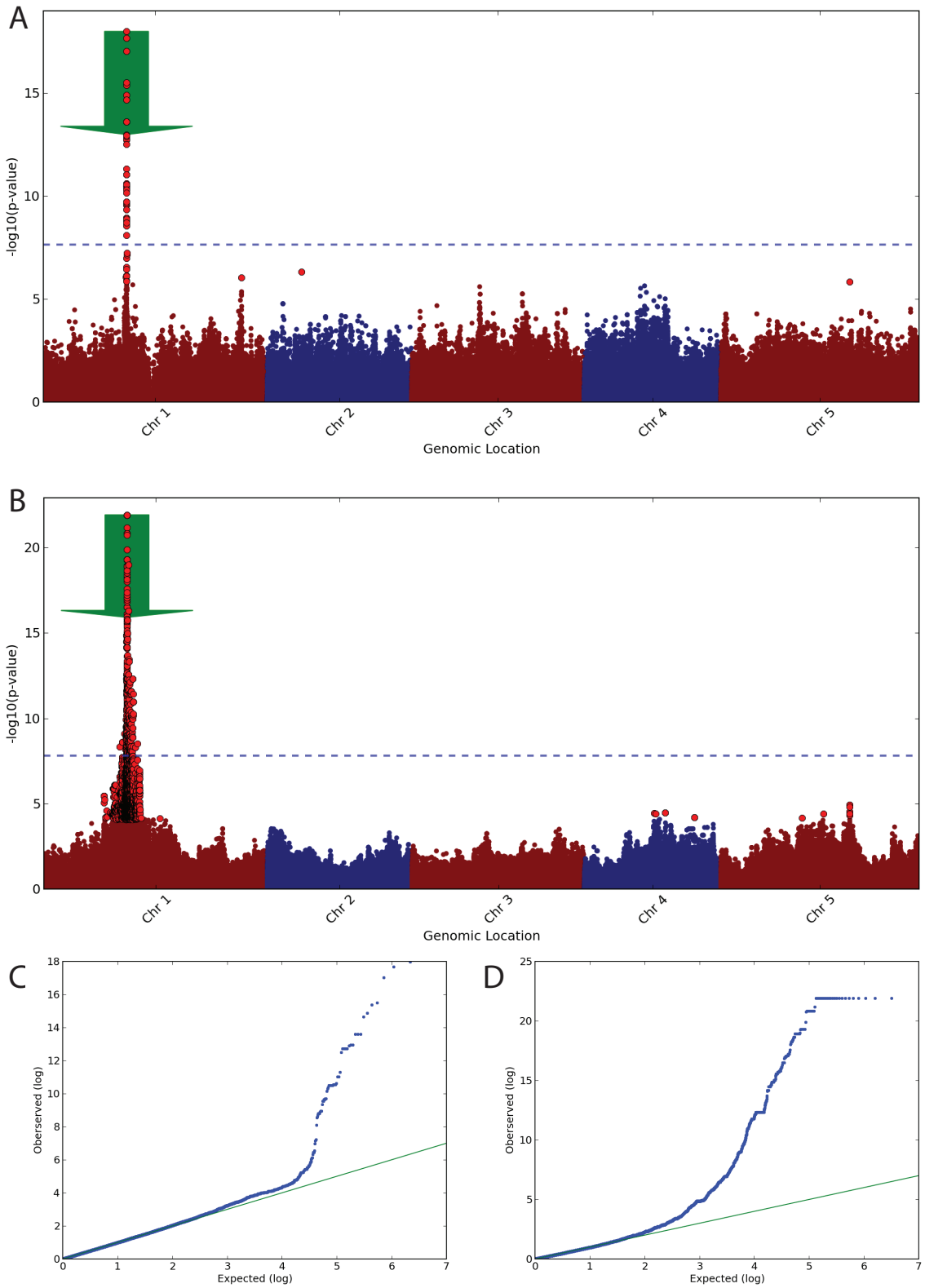


Figure 3.7: Manhattan plots for two *cis*-sQTL replicating in both populations. ***A/B:*** Location of SNPs significantly associated with an alternative 3'-splice site event in *AT1G31580* for CEGS (A) and MAGIC (B) population. The event position is marked with a green arrow. The x-axis shows genomic location and the y-axis the negative $\log_{10}$ p-value. All SNPs that are significant under a 5% FDR are marked in light red. The genome-wide Bonferroni-threshold of a corrected p-value of 0.05 is indicated as dashed line. ***C/D:*** Quantile-quantile plots showing the calibration of the p-value distribution for the same sQTL as above for CEGS (C) and MAGIC (D). The uniform distribution is shown as green solid line.

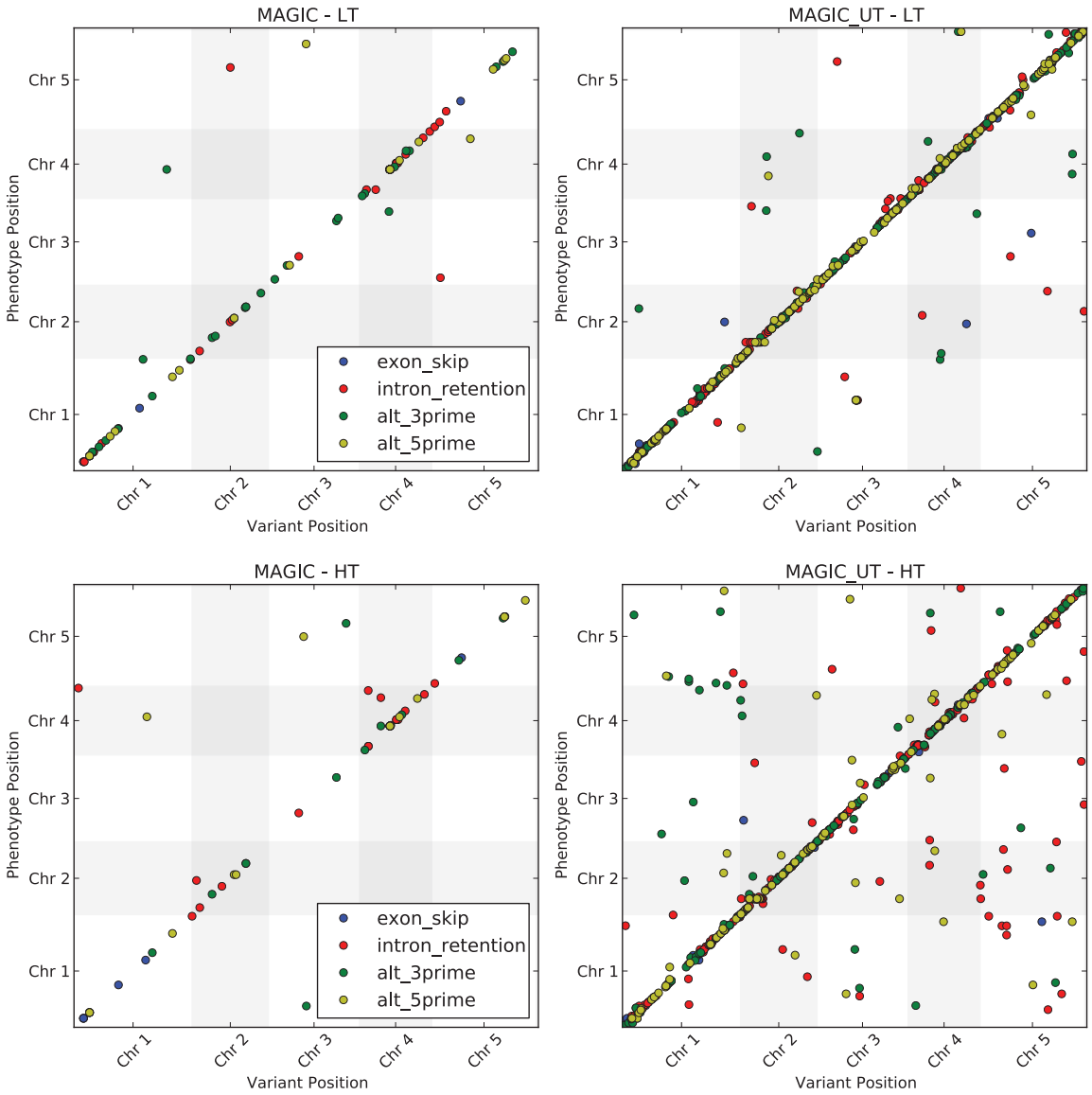


Figure 3.8: Influence of environment and input data size to the number of identified associations. All plots show the variant positions in the genome (x-axis) and the position of the associated event (y-axis) for all AS events passing the Bonferroni-threshold of genome-wide significance. Different event types are shown as different colors. The left plots show the associations on the population MAGIC (trimmed) and the right plots show associations on the untrimmed population. Upper plots show low temperature and lower plots show high temperature. Grey boxes in the background highlight the chromosomes.

ducing a protein of the exostosin family with predicted enzymatic activity and *AT5G66380*, a gene encoding a folate transporter. Although these findings are interesting results as such, they need further biological investigation to confirm any of the suggested links. Especially the low reproducibility of *trans*-sQTL over temperatures needs to be addressed critically before any speculations regarding environment-specific effects can be made. As we observed

Table 3.4: Overview of sQTL detected as significant in the different mapping populations. All sQTL are provided per event type and distinguished into *cis* (cis) and *trans* (trn), The two environments are marked as low temperature (LT) and high temperature (HT). The same sQTL can occur in HT and LT at the same time but not in cis and trn.

	CEGS				**MAGIC (trimmed)**				**MAGIC (untrimmed)**			
	LT		**HT**		**LT**		**HT**		**LT**		**HT**	
Event Type	**cis**	**trn**	**cis**	**trn**	**cis**	**trn**	**cis**	**trn**	**cis**	**trn**	**cis**	**trn**
Alt 3'	4	1	3	2	30	2	15	2	245	13	145	22
Alt 5'	6	1	4	0	19	2	12	2	140	4	98	17
Exon Skip	1	0	0	0	10	0	11	0	77	3	66	2
Intron Retention	6	2	5	0	29	2	17	1	459	12	363	28
	17	4	12	2	88	6	55	5	921	32	672	69
Total	21		14		94		60		953		751	

a similar behavior for the reproducibility of eQTL over temperatures (data not shown), this provides interesting links for further investigation.

Table 3.5: Detected sQTL in *A. thaliana* distinguished by environmental condition. Both *cis*- and *trans*-sQTL are categorized into three classes: sQTL that were detected only in low temperature (LT), in both temperatures (LT/HT) and only in high temperature (HT).

	cis			**trans**		
	LT	**LT/HT**	**HT**	**LT**	**LT/HT**	**HT**
CEGS	6	11	1	4	0	2
MAGIC (trimmed)	45	43	12	6	0	5
MAGIC (untrimmed)	462	459	223	30	2	67

3.3.4 Conclusion

In this study we have shown the applicability of our methods as part of a large-scale analysis pipeline. PALMapper sensitively aligned more than 400 *A. thaliana* full transcriptome RNA-Seq libraries thereby taking a complex set of variants and a list of provided junctions into account. We detected and then quantified several thousand alternative splicing events with SplAdder and used them as phenotypes within a genome-wide association study to identify sQTL. Encouragingly, many sQTL, especially in *cis*, replicated over the two mapping populations or over the two environmental conditions. However, we also found sQTL that were specific to a mapping population or that were only significantly associated for one of the two environmental conditions. We observed a similar behavior in the eQTL analysis, where sQTL *in cis* replicated well opposed to *trans*-sQTL that replicated scarcely. These findings need a thorough technical and biological follow-up analysis to exclude artifacts. After removing false positives, we suggest functional analyses, such as an enrichment-analysis

for functional terms using gene-ontology (GO) annotations or experimental validation of selected, high-confidence sQTL. We are confident that the data we generated can further be used to gain a deeper understanding of how genetic architecture or the quality of the input data influence the findings of genome-wide association studies.

3.4 Identification of Splicing QTL in 12 Cancer Types

In 2005 the National Institutes of Health (NIH) of the USA initiated a large scale data collection effort to comprehensively gather and catalog cancer related molecular phenotypes: The Cancer Genome Atlas (TCGA) [125]. A wide range of different types of molecular data has been collected, including whole genome DNA-sequencing data, whole exome sequencing data, RNA-Seq data, DNA-methylation profiles, reverse phase protein arrays measuring a subset of the proteome, but also clinical information such as risk factors, ethnic group or partial treatment history. Most samples are available for tumor and normal tissue, allowing for comparative analyses. At its beginning starting with only 2 cancer types, TCGA has vastly expanded and currently contains data for over 20 different cancer types[1]. It has already been a great resource for the study of single tumors [26, 104, 127, 140, 208] and first association studies of sub-groups of tumors [169].

To focus on an integrated analysis of several cancer types, in 2012 the Pan-Cancer analysis working group was launched, to comparatively analyze a sub-group of 12 cancer types [310]. As part of this effort, we have performed an association study of splicing phenotypes with a comprehensive set of germline and somatic genotype variants across the full set of 12 cancer types (for a full list cf. Appendix A.10). Whereas germline variants are gained through familial inheritance and can be predispositions for cancer, somatic variants arise during the lifetime of an individuals and are spontaneous mutations that are mostly harmless but can also have dramatic effects towards cancer progression. No previous study has analyzed the variation of alternative splicing with respect to these two mutation classes. In this section, we describe the details of our analysis, provide an overview of first results and discuss both our findings as well as the challenges of such a large scale study comprising thousands of samples.

3.4.1 Study Design

Although TCGA provides already pre-processed data for all cancer types, such as variant calls from exome or whole genome sequencing or gene expression counts from RNA-Seq files, each cancer type has been processed independently, including the usage of different pipelines, tools and parameter settings. To run an association study over all cancer types at once, it is necessary to uniformly process all samples. Otherwise, artifacts originating from differential processing rather than cancer type specific effects will be detected, introducing false-positive associations. Thus, we uniformly re-processed all raw sequencing files using the same analysis pipeline. We processed the following two sequencing data types. For extraction of gene expression counts and characterization and quantification of alternative splicing events, we processed 4,403 RNA-Seq samples, split into 4,073 tumor and 330 normal samples, each comprising between $\sim 5 \cdot 10^6$ and $\sim 450 \cdot 10^6$ reads. To call somatic and

[1] http://cancergenome.nih.gov/cancersselected

germline variation, we re-processed 9,014 whole exome sequencing files, split into 4,313 tumor and 4,701 normal samples, each comprising between $\sim 10 \cdot 10^6$ and $\sim 650 \cdot 10^6$ reads. A full overview of all samples for each cancer type is provided in Appendix A.10. All data has been downloaded from the TCGA data portal through the Cancer Genomics Hub[2] [194] and comprised several hundred terrabyte in total. Only sequencing samples produced with Illumina sequencers were taken.

The main goals of our study were to run both a common variant association study (CVAS) as well as a rare variant association study (RVAS), as described in [330]. Whereas common variants occur with a higher frequency in a population, usually chosen as $> 1\%$, rare variants show a frequency below this threshold. Our goal was to identify splicing quantitative trait loci (sQTL) for each of the two settings. We further aimed at identifying differences and commonalities between cancer types as well as distinguishing between somatic and germline variants associated with splicing alterations.

3.4.2 Analysis Pipeline

Due to the complexity of the study and the large sample sizes, the analysis pipeline consists of many steps. To stay within the scope of this work, we will give a brief summary of each analysis step but put more focus on the parts that are related to the methods presented in this work or are necessary to understand the results discussed in later parts of this section.

Data Acquisition Sequence files in the common FASTQ format were not available through TCGA at the beginning of this study. Instead only processed alignment files in BAM format were provided. As the alignment files contained both aligned and unaligned reads, we downloaded all RNA-Seq and exome-Seq samples in BAM format via the Cancer Genomics Hub (CGHub) and applied a custom script to extract all read information into FASTQ files. We further downloaded all clinical annotations provided for each sample via the TCGA data portal.

Sequence Alignment The re-alignment of all sequence files was done using STAR [66] (version 2.2.0g) for reasons of efficiency. PALMapper would have provided more sensitive alignments but would have also been more costly in terms of running time by a factor of approximately 50. All reads were aligned against the human hg19 genome sequence downloaded from the University of California Santa Cruz (UCSC). For RNA-Seq alignments, the genome index was built with additional splice junction information from the GENCODE annotation [109] (version 14) using the parameter `--sjdbOverhang 75`. An overview of all parameters used for alignment is provided in Appendix A.10. After alignment, reads were converted to sorted and indexed BAM format using SAMtools [167].

Expression Counting For counting expression of all annotated genes, we applied a custom Python script. An aligned read was counted towards the expression of a gene, if any exon in the gene shared at least one genomic position with the alignment. If more than one alignment was available for a read, we took the alignment with the highest alignment score. Two versions of expression estimates were generated. The first variant counts gene expression as the number of all reads overlapping to exons of that gene. The second variant only

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

considers non-alternative exon positions to compute the total expression. Non-alternative positions of a gene are those that are always part of an exon for any transcript isoform. While the first variant is commonly used in gene expression counting, the latter provides a more stable estimate of alternatively spliced genes. The expression estimates were later used as co-factors in the association analysis and as phenotypes for an eQTL study (data not shown).

Identification and Quantification of Alternative Splicing Events Alternative splicing events of four different types were detected using the SplAdder pipeline (cf. Section 2.5): exon skips, intron retention and alternative 3'- and 5'-splicing sites. Again, the GENCODE genome annotation (version 14) was used as a basis for splicing graph augmentation. Only the set of protein coding genes was considered. We generated one augmented splicing graph per sample and subsequently merged all graphs into a common graph. To only keep high-confidence edges in the graph, we applied edge filtering, retaining only edges that were sufficiently supported by at least 10 different samples. We then called splicing events from the graph, using confidence level 3, the highest level available in SplAdder. For a first pilot run of the association pipeline, we further restricted the full gene list to a subset of 2,000 genes that were either part of the cancer gene census [91], were annotated as splicing factor or transcription factor in the ENSEMBL annotation [85] or were an upstream gene of a gene falling in any of the two previous categories in a custom pathway analysis (E. Demir, personal communication).

Variant Identification To identify variants from exome-Seq data, we followed two different strategies. The first strategy was the joint calling of variants in all samples at once using the Unified Genotyper that is part of the Genome Analysis Toolkit (GATK) [64, 198]. This step can identify both common and rare germline variants as well as common somatic variation, thereby sharing evidence over all samples, which helps to make accurate calls at positions lowly covered in single samples. In the second strategy we applied MuTect [50] individually to more than 4,000 tumor-normal sample pairs. MuTect is specialized to call both rare and common somatic variants. By taking also the clonal structure of a tumor into account, it is able to handle more than two alleles per single variant position. After several filtering steps based on the Broad best-practice guidelines [295], we used the GATK variant set exclusively present in the tumor samples for association. The MuTect calls were only used to label single variants as somatic or germline variant for downstream analysis. If variants had missing values in a subset of the samples not exceeding 40% of the total population, we filled up the missing values with alleles randomly chosen from the allele set of the remaining samples. Variants that showed missing values in more than 40% of the samples were excluded from further analysis.

Common Variant Association For the association of common variants, we applied a linear mixed model (LMM) approach (cf. Section 1.3.2). Following a strategy already described in context of the sQTL analysis in *A. thaliana* in Section 3.3, we computed percent spliced in values (PSI, cf. Appendix A.9) for all alternative splicing events, added a small amount of noise in the order of 10^{-5} to break up ties in the ranking and used an Inverse Normal Transformation on the ranks to generate splicing phenotypes following an

approximate Gaussian Distribution. As we were testing for additive effects and used a binary encoding, we restricted the used variants to bi-allelic SNPs only and excluded insertions and deletions. In the common variant association analysis, we filtered for a population-wide minor allele frequency (MAF) of at least 1%. Within the LMM, we considered copy number variation and normalized gene expression as fixed effects. We further integrated several random effects into the model. To stratify structure within the population, we computed two different kinships, describing the pairwise similarity of samples based on a subset of genetic variants. The first kinship should capture germline variation and was computed on all common germline SNPs. The second kinship should reflect structure implied by cancer specific somatic mutation patterns and was computed on all rare somatic SNPs (MAF $< 1\%$). To also account for unknown confounding factors, we estimated a set of 40 confounding factors applying PANAMA [89, 90] to the matrix of expression values and subsequently used them as random effects in the model. As before, all computations regarding the LMM were done using the LIMIX package for Python [174].

Other Association Analyses We performed several other types of association tests. However, we will not present the results in the context of this work and will thus only briefly summarize the ideas of the different analyses. To associate rare variants with a MAF less than 1%, we applied a burden test to group several low frequency variants of a gene that could have severe effects [163]. Instead of running associations on sub-populations split per cancer type, we also performed an association test on the full sample set, including the cancer type as co-factor. As a second strategy to increase testing power, we also implemented a meta-analysis, combining the test results from the individual cancer-types using Fisher's test. In addition to the single alternative splicing events, we also computed a global splicing phenotype, that captures global shifts in the distribution of all events of one type. This phenotype can also be used for association analysis, to identify variants that cause a subtle change in man events of one type.

3.4.3 Results

As none of the methods described in this paper was used for alignment in this study, we will not further discuss the alignment outcome but rather focus on the results describing the diversity of splicing in cancer as well as the outcomes of the different association analyses.

Diversity of Alternative Splicing in Cancer Numerous studies have shown a relationship between tissue type and the expression of alternative splicing isoforms [303, 322] suggesting an active role of splicing in forming cell identities. However, also a close connection between cancer and splicing has been reported [268, 305]. To put the splicing alterations in cancer into a broader context, we compared the number of events we could confirm in the re-processed TCGA samples with various publicly available whole transcriptome RNA-Seq datasets. We used 518 alignment files produced in the context of the ENCODE project [71, 78] and 189 alignment files from the GEUVADIS project [155, 277]. We denoted an event-isoform as confirmed within a sample, if we could observe at least 5 spliced alignments mapping to each junction of the isoform. We denoted an event as confirmed, if each isoform could be confirmed in at least $k\%$ of the samples, for $k \in \{1, 5, 10, 20\}$. Figure 3.9, Panel A, gives an overview of the number of confirmed exon skip events for

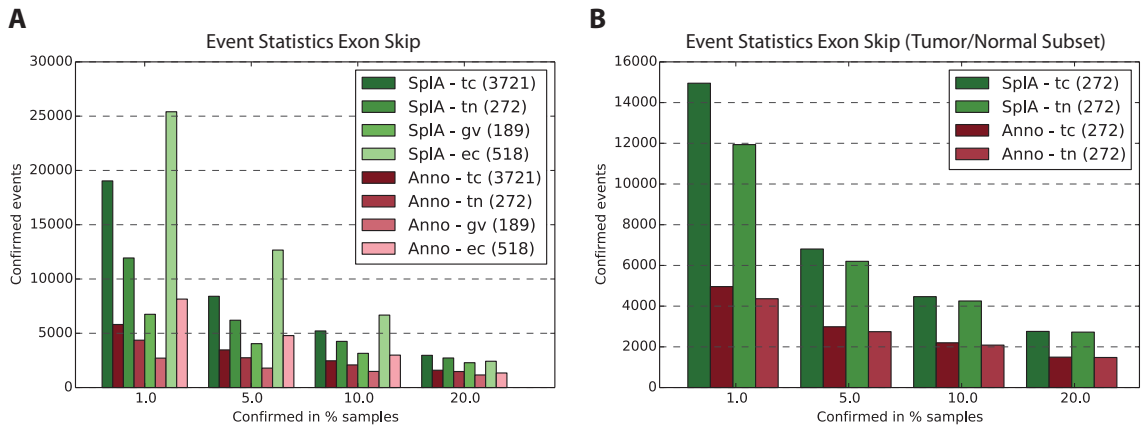


Figure 3.9: Effect of graph augmentation on the detection of splicing events. Based on the whole transcriptome sequencing datasets further described in the text, the number of exon skip events detected in a SplAdder-augmented graph vs. the not augmented annotation is shown. Datasets are named as follows: TCGA normals (tn), TCGA tumor (tc), GEUVADIS (gv), ENCODE (ec). ***A:*** Number of exon skips in each data set confirmed in the annotation (red) and in the SplAdder-augmented graph (green). ***B:*** Confirmed exon skips on a subset of 272 individuals where both tumor and normal samples were available.—Refer to the text for a description when events were counted as confirmed. The number of samples in each dataset is shown in parentheses.

the different choices of k. As expected, we could confirm a much larger number of events in the augmented version of the annotation. Most of the novel events are rather rare but can still be confirmed in a substantial number of samples (1% of all TCGA samples is still a decent number; 40). However, a more conservative setting, requiring confirmation in a large number of samples (20%), shows a much higher concordance between annotation and augmented annotation. Further, we observed more confirmed events in tumor samples than in normal samples. As this could be attributed to the differing sample sizes, we compared tumor and normal sets that were sub-sampled to the same size, confirming our previous observation (Figure 3.9, Panel B). The higher splicing diversity in the ENCODE dataset, resulting in a generally larger number of events (cf. Figure 3.9, Panel A), can be attributed to a higher diversity of tissues, as the cancer samples only represent a subset of the tissue types.

Other interesting observations can be made, when samples are clustered with respect to their splicing complexity. We used the PSI values for all events of a single event type over all individuals to compute a kernel matrix summarizing the observed variation in splicing. Applying principal components analysis (PCA), we identified the main axes of variation in the dataset and color coded them with different sample labels. The result is shown in Figure 3.10. We restricted the analysis to a subset of five cancer types where samples for both tumor and normal were available. As expected, the different cancer types cluster together when plotted over the first main axes of variation, which could be addressed to underlying tissue specific splicing of the tissue source site. A subset of three pairs of the first four principle components (PCs) for exon skips colored by cancer type is shown in Figure 3.10, Panel A. Interestingly, if the same PC structure is colored according to the tumor/normal-state of the corresponding samples, all the normal samples form a cluster, suggesting that the observed structure on the cancer types is rather cancer type specific than caused by any

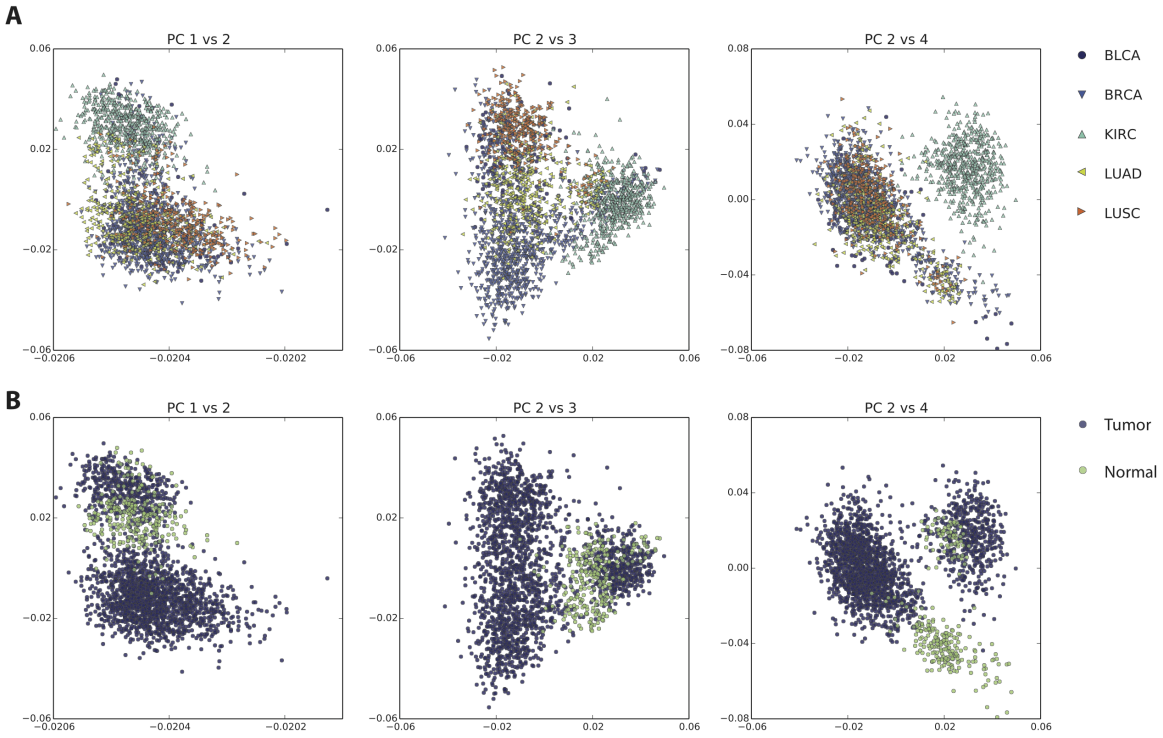


Figure 3.10: Overview of the first three principle components of all exon skip events resulting from PCA on a subset of five cancer types. Both panels show the pairs of principle components plotted against each other but with differently labeled individuals. Surprisingly, the samples cluster by cancer type rather than by tissue, as indicated by the tight clustering of normals. ***A:*** Individual samples are labeled by cancer type. The single types can be distinguished by both marker color and marker shape. ***B:*** Individual samples are labeled according to tumor (blue) or normal (green) state.

underlying tissue-specific splicing (Figure 3.10, Panel B). This result is encouraging, as it suggests that the splicing phenotype contains enough information not only to distinguish cancer from normal samples but also to identify splicing patterns that are characteristic for individual tumor types. This could not only be used for diagnose but also offers interesting possibilities for therapy, e.g., through artificial antisense oligonucleotides [143, 226] or natural compounds [33]. A similar observation had also been made on EST data from cancer sample [321], however, the size of our dataset allows for a much more comprehensive analysis. In the following, we will discuss examples that specifically link the splicing of single events to genetic alterations only found in cancer samples or sometimes even only within samples of a certain tumor type.

Splicing Associations in *cis* As introduced in Section 1.1.1, various regulative mechanisms for alternative splicing exist that involve sequence elements both near and distant to the site of the alternative splicing event. We identified a large number of sQTL that act in *cis*, that is the associated genetic alteration is co-localized with the alternative splicing event. The most common mechanistic explanations for *cis*-events are mutations within the splice acceptor or donor site or mutations in the motifs of splicing enhancers or silencers. Amongst the numerous findings from previous studies we were able to confirm,

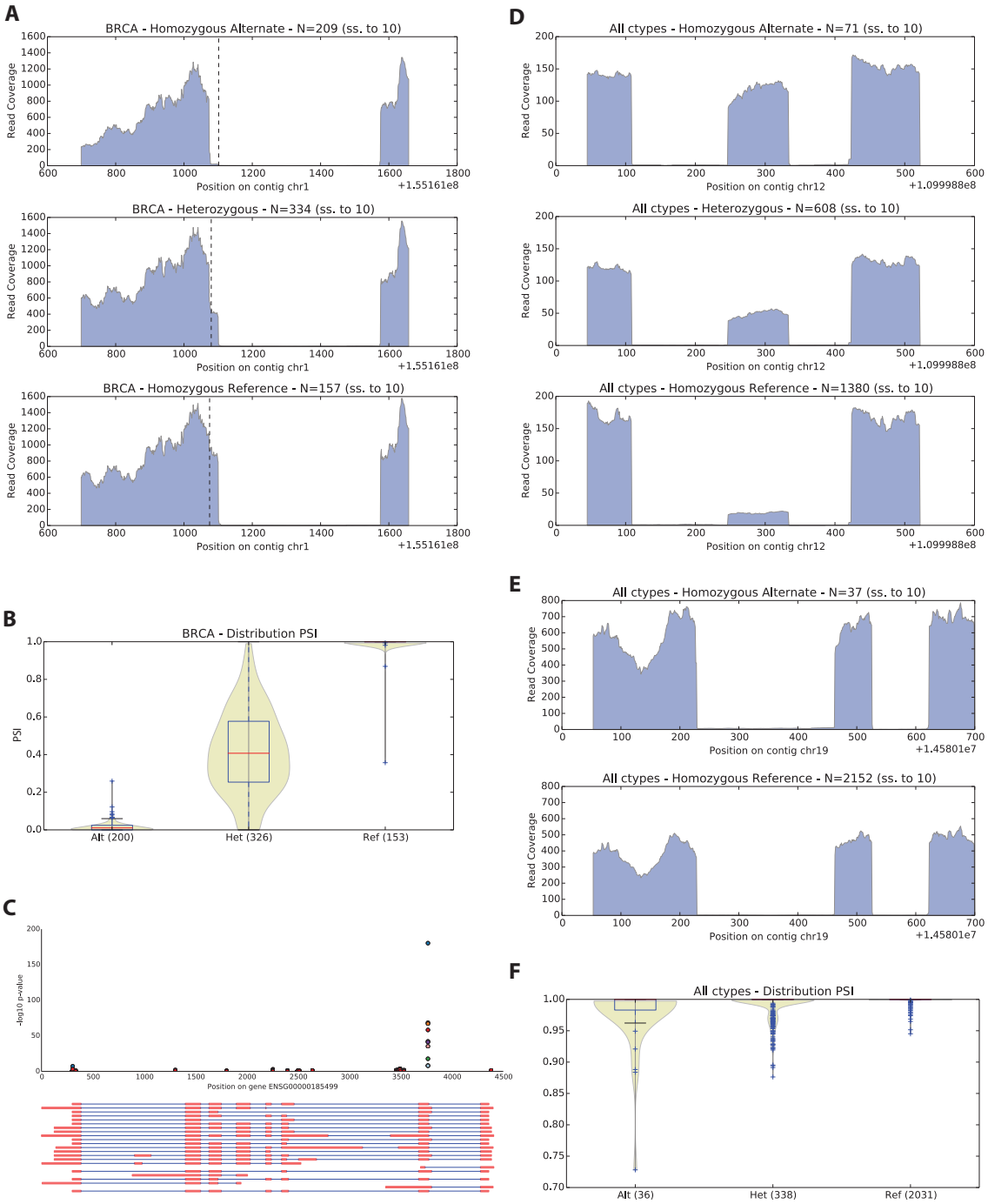


Figure 3.11: Examples for associations in *cis*. ***A:*** Coverage plots for three different allele states of a SNP in the gene *MUC1* showing alternative usage of the 5'-splice site depending on the genotype. The alternative region is highlighted with a dashed black line. ***B:*** Distributions of PSI values for the alternative 3'-event in *MUC1*. ***C:*** SNP locations within the *MUC1* gene and corresponding association p-values in negative log-scale. The different transcript annotations are shown in red. The most significantly associated SNPs co-localize with the affected splice site. ***D:*** Coverage plots for the three different allele states of a SNP significantly associated with an exon-skip in *MMAB*. ***E:*** Coverage plots for two allele states of a SNP significantly associated with an exon skip in *PKN1*. ***F:*** Distribution of PSI values for the three different alleles of the *cis*-association of *PKN1*.

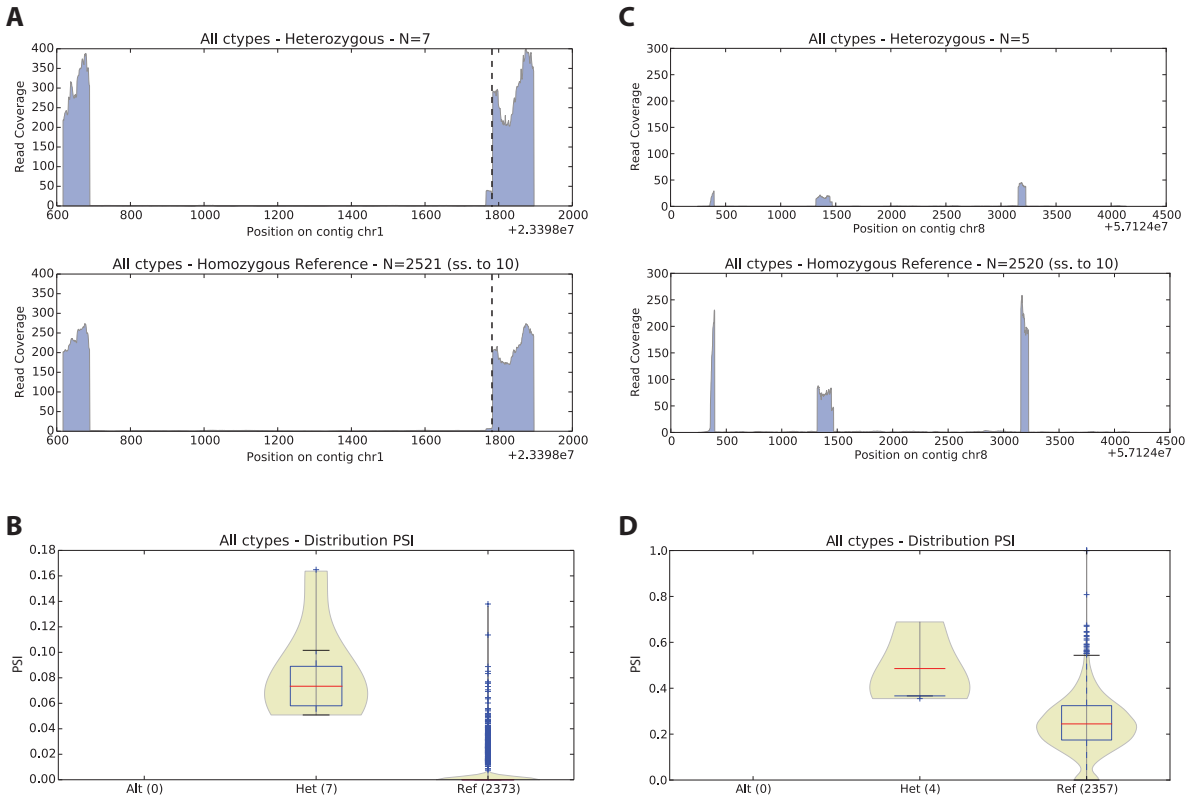


Figure 3.12: Examples for associations in *trans*. ***A:*** Coverage plot for an alternative 3'-splice site in KDM1A that is significantly associated with a SNP in the splicing factor SF3B1 (coverage for the heterozygous genotype shown on top, homozygous reference on bottom). ***B:*** Distribution of PSI values for the event in KDM1A for the two different genotypes. ***C:*** Coverage plot for an exon-skip event in CTNB1 that is significantly associated with a SNP in the splicing factor subunit U2AF1 (heterozygous top, homozygous reference bottom). ***D:*** Splicing index distribution of the event in CTNB1 for the different genotypes.

we picked two examples that shall illustrate these mechanisms (Figure 3.11). We found a *cis*-association in *MUC1*, where a splice site mutation was shown to cause the use of an alternative 3'-splice site [172] (Figure 3.11, Panels A–C). Further, we confirmed an association of a cassette exon in *MMAB1*, where SNPs within a splicing regulatory region were shown to introduce an additional exon [150] (Figure 3.11, Panel D). Additionally, we found numerous associated sQTL that have not been documented so far. One example is the differential exon usage in the gene *PKN1*. Although no clear pattern can be observed in the coverage plot (Figure 3.11, Panel E), a clear shift in the phenotype distribution is evident (Panel F). However, many of the discovered candidates need to be subject to further investigation and biological validation to determine possible functional roles.

Splicing Associations in *trans* Also variants occurring in greater distance from the splice event can have regulatory potential. These so called *trans*-effects are more difficult to detect in larger genomes, as not only variants within a window around the event need to be tested but all possible variant positions in the genome, inflicting a high correction

penalty for multiple testing upon the results, thus requiring a sufficiently large sample set to achieve sufficient statistical power. For the cancer types with many samples, such as BRCA, KIRC and LUAD, the sample sizes are just large enough to identify first *trans*-associations. We could confirm several published sQTL, including a significant association between an alternative 3'-event in CTNB1 and a SNP in the splicing factor subunit U2AF in lung cancer (LUAD) [37] as well as an alternative 3'-event in KDM1A significantly associated with a SNP in the splicing factor SF3B1 [245] (Figure 3.12, Panels A–B). For cancer types with larger sample sizes, as well as through a combination of all sets of cancer types, we were also able to identify several novel *trans*-associations. One example is an exon skip in CHCHD7, a gene previously linked to adenoma [15], that showed significant association to a SNP in the RNA-binding protein RBM39 (Figure 3.12, Panels C–D). These novel findings still need validation but are already a very encouraging result, as they show that we can find links between genetic variation and changes in alternative splicing of distant genes. An increase of sample size and improved methodology will likely find more such connections and help to explain the characteristic splicing changes of cancer cells.

3.4.4 Conclusion

The realization of this study was a major effort in terms of data processing and analysis. We could show, that SplAdder is capable of handling thousands of alignments of whole human transcriptomes and can detect and quantify tens of thousands alternative splicing events in the given transcriptome sets. Through qualitative analyses, we could show that SplAdder substantially increases the sensitivity to detect alternative splicing events and that the observed diversity is largely caused by the cancer state and does not due result from tissue specific splicing. In a genome-wide association study, we could identify numerous splicing QTL, both confirming previous findings and detecting novel associations. Due to the size of our dataset, we were not only able to find associations in *cis* but also found numerous *trans*-hits where genetic variant and affected gene lie in parts of the genome that are distant from each other, e.g., on different chromosomes. Although these findings need to undergo further biological validation, our results are very encouraging. Especially splicing variants are interesting novel targets for therapy and first studies have shown promising results by applying artificial oligonucleotides [143, 226] or shown interesting applications for natural compounds [33]. In this study, we suggest further work to be done on a larger set of both samples and genes. Also the investigation of different techniques to associate rare variants is worthwhile, as especially specific somatic variants occur at a very low frequency in the population but can be recurrent at a higher frequency on a whole gene or pathway level.

4 Discussion

Over the recent years, the introduction of RNA-Sequencing has revolutionized the field of biomedical research. Not only the ever dropping cost to generate sequence data but also the increased sensitivity and accuracy of the measurement techniques enabled larger and more complex studies than ever conducted before. Along with these new possibilities also many new problems arose, especially on the analysis side, making the use of advanced computational approaches necessary. The list of computational challenges is long and comprises numerous problems that are not or only partially solved yet: the accurate alignment of millions of reads to one or several reference sequences, the integrated analysis of thousands of sequencing samples, the efficient representation and storage of thousands of reference genomes, the detection and representation of complex genomic variation, the quantitative and qualitative analysis of RNA-Seq alignment results, or the comparability between data generated from different sequencing platforms and techniques, to name only a few. Although many of these problems were topics of avid research in the past years, most of the challenges remain.

Methodological Contributions

In this work, we presented several methodological contributions aiming to improve the accuracy of RNA-Seq read alignments and to facilitate quantification of alternative splicing events. A key aspect of our efforts is thereby the applicability within a high-throughput setting, allowing for the efficient processing of thousands of samples. In the following, we will discuss our methods in the context of existing research and provide an outlook to possible improvements as well as to interesting new developments in the field.

A central problem of alignment in the context of transcriptome analysis, is the fact that the reference sequence is in the best case still only a good approximation to the genome of the individual the RNA-Seq data is originating from. In many cases only a reference for a closely related species is available, increasing the number of sequence differences. Especially in the context of personalized genomics and recent advances in technology, lowering sequencing cost for a full genome to under $1,000, a solution to represent a set of references rather than a single sequence is strongly needed. With the extension of PALMapper into an approach that can take a set of variants into account during alignment, implicitly representing many possible reference genomes at once, we have made a first important step into that direction. Although other alignment tools have meanwhile begun to allow for additional variation, including TopHat [137] and GSNAP [319, 320], to our knowledge no other approach can presently integrate all variant types PALMapper uses or can form combinations of provided variants. Also the integration of variants into the index structure is a novel contribution. Further, as participants of the largely community-driven RNA-Seq Genome Annotation Assessment Project (RGASP) [273], we were able to contribute early on to

new developments in the alignment field. Hence, along with other participating aligners, PALMapper was one of the first tools that fully supported standardized output in BAM format and implemented junction re-alignment. Both have become standard features in many aligners, which is an important outcome of the RGASP efforts. However, we still see many ways to further improve the current state of alignment approaches. Although PALMapper already provides both an FM-index and a k-mer based index to represent the genome, only the latter can currently be augmented with variants. For future versions, we suggest to also explore other indexing techniques for a variation-aware extension, including suffix arrays and the FM-index. A long-term goal would thereby be, to build the full alignment against a reference graph, not only forming variant combinations on the fly, but rather to develop an expanded representation of the full genome sequence including all variant paths. A further point to address in this context is the speed of complex variant alignment. Approaches like STAR [66] have shown remarkable performance for the non-variant case through utilizing uncompressed suffix arrays. Extending this concept to include variants is a promising strategy to pursue. Therefore, computational solutions to efficiently index a graph structure and local sequence alignment strategies against graphs are needed. First interesting studies that address the indexing of graphs [265] or that describe ideas how to construct a pangenome reference structure to integrate many reference genomes [213] have been published recently.

Although the alignment problem for RNA-Seq data appeared to be solved very early after the technology had been introduced, comparative analyses revealed marked differences between the results produced by the various approaches. This is especially problematic, as alignment usually forms only the first step within a whole analysis pipeline, thus possibly creating strong biases for any downstream results. In the context of the RGASP competition, we suggested to thoroughly analyze the alignment approaches and compare their respective outcomes to further the understanding of differences and possible weaknesses. We developed several analysis metrics comparing not only statistics of quality and edit operations but also included measures capturing the sensitivity of the alignment of ambiguous mappings. The analyses have provided valuable feedback to the research community within RGASP to improve alignment strategies and have been picked up by the RGASP consortium to run a dedicated alignment comparison [79]. We have used the insights gained during the evaluations to both improve the PALMapper software and to suggest filtering strategies as alignment post-processing to increase the accuracy of the final alignments and make them more comparable between different aligners. However, still missing is an accepted gold standard to compare different alignment strategies or to tune parameters for a given problem setting. First developments in this direction are undertaken from large research networks, such as the International Cancer Genome Consortium (ICGC) [116] or the Global Alliance for Genomics and Health, that was founded only recently.

One major challenge one has to solve during alignment also appeared as particularly problematic in our evaluations: the handling of ambiguous read mappings. To resolve such ambiguities, the alignment algorithm can make a decision based on the alignment scoring. However, this is not always possible and mostly too inaccurate, as often several equally good alignment possibilities exist. In such cases only a random choice or the reporting of several hits remain as options. If no decision is made during alignment, the uncertainty is passed down the pipeline. With the Multimapper Resolution tool (MMR) we proposed

an alignment filtering strategy, that uses coverage information to identify the likely correct alignments from a set of given possibilities. We could show improvements for downstream tasks such as read quantification and have successfully used the software in several projects also presented in this work. Ultimately, the decision over the correct alignment should be incorporated into the model for downstream analysis. That is, a quantification approach should optimally explain observed read coverage while at the same time making an optimal choice over possible mapping locations for single reads. We implemented such a combination and suggested an iterative strategy of MMR and MiTie [25]. MMR originally used a squared loss for local coverage optimization to decide on an optimal read assignment. However, a Negative Binomial loss as used by MiTie is much better suited, as it better models overdispersion in the read distribution originating from biological variation. We therefore extended MMR to use a Negative Binomial loss and optimized the same objective as MiTie, altering the read assignments instead of the transcript structures. We could show that this combined strategy improved the accuracy of transcript prediction [25]. We know of only one other approach [32] similar to MMR that is explicitly dedicated to computationally resolve ambiguous mappers from any alignment file and can thus by applied in any context. Whereas this tool requires re-alignment of the reads, MMR can take an existing alignment as input. We further provide a strategy to incorporate prior knowledge into the decision process in form of expectation measures for the coverage. However, several other approaches for isoform prediction and quantification exist that take mapping uncertainty of alignments implicitly into account within their model, such as RSEM [162, 164], IsoEM [214] or eXpress [239]. Nevertheless, especially for complex transcriptomes such as of human and mouse, it remains a hard problem to correctly predict and quantify all transcripts if more than two isoforms are expressed [25].

While the quantification of whole transcript isoforms is still challenging, the quantification of single alternative splicing events is an easier task to tackle. By removing long-range dependencies and only focusing on the event of interest, it becomes computationally feasible to assess alternative splicing in a large set of RNA-Seq samples. We presented SplAdder, a tool that uses RNA-Seq data to augment an existing annotation and employs a splicing graph structure to identify alternative splicing events. By filtering and quantifying the events based on the given samples, it enables differential testing and other quantitative analyses. With the need to process thousands of RNA-Seq samples and detect events in the whole transcriptome, we focused on an efficient implementation that is easily parallelizable. Through applications in large scale studies involving approximately 4,000 human samples and more than 700 samples from *A. thaliana*, we have provided evidence that SplAdder is highly efficient and effective and not restricted to a certain organism. To the best of our knowledge, there exists currently no other tool that can efficiently extract alternative splicing events from thousands of samples and quantify single events for use in downstream analysis. The approach used in [36] is not able to take novel exons into account and seems to be designed for smaller scale projects. As the current implementation employs many heuristic measures, we envision several improvements for future versions. Currently, the features used for graph augmentation are manually optimized based on the large-scale datasets examined in the studies. We suggest to estimate the values for all features that are used to add new events to the graph from a given training set of trusted events. Especially if a large number of samples is available, this approach will be more accurate than the

current strategy. Also an approach based on techniques from machine learning is possible, to learn which edges in the splice graph are correct and which not, using splicing databases, data from available replicates or simulated data as a training set.

All methods presented in this work have been tested in several biological applications and some were integral part of large-scale, multi-institutional studies. Our developments for alignment in context of heterogeneous genomes and the post-processing analyses have largely contributed to improved data quality and the reproducibility of outcomes. Further, the efforts we spent to create SplAdder as an easy to use but yet highly efficient tool have enabled analyses on very large sample populations allowing for tests that would have been statistically underpowered otherwise. Through definition of clean interfaces and the restriction to standardized formats, we ensured that all tools can be used as parts of different pipelines and are applicable in a versatile manner. To make them available to a broad community of users, we did not only publish their source code, but also integrated them into the lab's Galaxy web-server and provide implementations in the framework of Oqtans [269].

Biological Applications

We presented four different biological studies where the tools developed in this work were essential parts of the analysis pipeline. Our approaches grew and improved along with the complexity of the biological questions we aimed to answer. In the following, we will summarize the four different studies, discuss their results in context of the respective field and describe why our tools were integral to the analyses.

We described two projects on mutants of the model plant *Arabidopsis thaliana*, where we used RNA-Seq data to analyze transcriptional and post-transcriptional regulation processes in plants. For the first study, we successfully conceived and implemented a computational analysis pipeline that identified thousands of alternative splicing events leading to transcript isoforms likely targeted by the degradation mechanism nonsense-mediated mRNA decay (NMD). We could show, that there is a direct correlation between the length of the 3'-UTR and premature termination codons introduced through alternative splicing and NMD efficiency. With this finding, we could confirm transcript features likely to trigger NMD in over 90% of significantly enriched event isoforms and have estimated that more than 17% of all protein coding multiple-exon genes can produce transcripts that are targeted by NMD. Given our conservative thresholds and the fact that only a part of the full range of splicing diversity could be observed in the plant seedlings, the fraction of NMD-targeted transcripts is likely to be an underestimate. These findings suggest a much more profound role of NMD in transcriptional processing than expected. As the set of affected genes showed enrichment for biological functions such as stress response and RNA metabolism, we have reason to speculate about a regulatory role for the NMD mechanism that exceeds its function of pure surveillance and error correction. The second study focused on polypyrimidine-tract binding proteins PTB1 and PTB2 and their role as splicing regulators in *A. thaliana*. We identified 452 alternative splicing events originating from 307 genes that were spliced in a PTB dependent manner suggesting that single splicing regulators are linked to many target genes forming a complex regulatory network. Further, we were able to confirm auto- and cross-regulatory potential of the PTBs. Interestingly, several of the target isoforms showed

NMD-eliciting transcript features, suggesting an interesting regulatory link that had not been established in this breadth before. Functional studies on the set of regulated target genes revealed several interesting leads, including *PIF6* that is relevant for seed germination and *FLM*, a regulator of flowering time. In both projects, the alignments were generated with PALMapper, directly applying insights for parameter choice and filtering that were collected during our alignment evaluation study. We further used MMR for alignment disambiguation and SplAdder for graph augmentation and event identification. Both studies have been used to consolidate the analysis pipeline and optimize the implementations. The results were peer reviewed and could be successfully published [68, 246], emphasizing not only the significance of our findings but also approving our analysis pipeline.

Building on the pipeline developed in these two initial projects, we implemented several improvements and followed a similar analysis strategy for two larger-scale studies to map splicing quantitative trait loci (sQTL) in populations of *A. thaliana* and human. Whereas the $\approx$ 700 samples in *A. thaliana* were aligned with the variation-aware PALMapper, the over 4,000 human samples had to be mapped with STAR [66] for performance reasons. As already pointed out before, we see several possible ways to improve running time of the variation aware alignment for future projects.

Central to both projects was the identification and quantification of alternative events as phenotypes for genome-wide association analysis. The extraction of alternative splicing events was successfully completed with SplAdder, showing its easy scalability and portability across organisms. Distributed on several hundred cores of a high performance computing cluster, the identification and quantification using whole transcriptome RNA-Seq data of 4,000 human samples could be completed in less than a week.

In the first study, the comparison of sQTL in the two *A. thaliana* populations CEGS and MAGIC in the context of high- and low-temperature growth conditions revealed hundreds of sQTL in *cis* and also several sQTL in *trans*. We made interesting observations regarding the replicability across populations, the environmental conditions as well as input data of differing quality. We found *cis*-sQTL to be profoundly more replicable than *trans*-sQTL, even across populations, where 50% of all significant *cis*-sQTL in CEGS also appeared as significant in MAGIC. For *trans*-sQTL reproducibility was found to be very weak. This could have several reasons, including a biological cause or a lack of power to detect associations of genome-wide significance. Supporting the latter interpretation, most of the *trans*-associations detected with CEGS and the trimmed MAGIC dataset showed p-values close to the significance threshold. Even when performing the test on the untrimmmed MAGIC data, providing a much better coverage of events, none but two of the *trans*-sQTL found in the trimmed dataset could be rediscovered, although the hits in the untrimmed dataset showed much lower p-values. Thus, before any functional speculations regarding population-specific or environment-specific *trans*-sQTL can be made, reproducibility needs to be further addressed. However, we found encouraging examples for replicating *cis*-sQTL, that show environmental specificity, providing a good basis for functional follow-up.

The second study consisted of the alternative splicing analysis of 12 different cancer types in over 4,000 patient samples provided through TCGA [125]. We presented the first comprehensive assessment of alternative splicing in the transcriptome of multiple cancers. Especially the application of SplAdder was essential for this analysis, as we detected a large fraction of alternative splicing events that could only be observed in cancer samples and

neither in the set of normals nor in representative samples from other studies. Interestingly, through the analysis of principle components we found a tremendous amount of splicing variation in the samples that was explained by cancer type rather than tissue specific splicing. These are very encouraging findings, as this enables us to think about splicing markers for cancer as well as interesting treatment options through natural compounds [33] or artificial antisense olignucleotides [135, 143, 226]. Using splicing as phenotype in a genome-wide association study, we identified numerous *cis*-sQTL confirming previous findings, including *MUC1* [172] and MMAB [150], but could also detect several sQTL *in trans* that had been discovered only recently, e.g., the association of *U2AF* and *CTNB1* in lung cancer [37]. More importantly, in addition to these anecdotal findings, our comprehensive approach was able to detect a whole range of novel sQTL that now need to be subject to further validation. Whereas *cis*-sQTL showed interesting patterns of replication across cancer types, sQTL *in trans* appear to be more cancer type specific. However, a thorough follow-up analysis of these findings is necessary. The sQTL identified in this study are an excellent basis for further research in the direction of splicing therapeutics and could provide interesting leads for drug target detection.

Concluding from the four very different biological applications, we are confident that the methods we presented are generally applicable to a wide range of possible studies, not limited by choice of organisms or number of individuals. We have shown that the suggested pipeline produces meaningful results and is both efficient and effective. Further, several of our findings resulted in peer reviewed publications [68, 246] or confirmed previous findings from the literature. More importantly, our results can provide valuable leads for further functional analysis in plants or therapeutic intervention in cancer. Especially the definition of splicing markers for cancer is an interesting idea that should be pursued further as this could lead to strategies for an early detection of the disease.

Future Directions

Studies as described in this work, including hundreds or thousands of whole transcriptome sequencing samples, would not have been possible without the drastic improvements in sequencing technology and the advancements on the side of computational analysis. However, many further methodological contributions and improvements of the implementations will be necessary as not only more but also new types of sequencing data enter the field. The amount of sequencing samples will further rise as more and more clinical studies begin to collect whole exome, whole genome or whole transcriptome sequencing data [125, 310]. Further, large international consortia such as GEUVADIS [155], GTEx [181], the already mentioned ICGC and the Global Alliance for Genomics and Health have begun to generate or to provide access to sequencing data for thousands of individuals and will have collected hundreds of thousands of samples in the near future, addressing problems ranging from personalized medicine to population genetics. Interestingly, also major health efforts of single countries, like the UK10K project in the United Kingdom [132], include the massive sequencing of individual genomes. This immense amount of data soon to be available will require sophisticated methods for storage, sharing and analysis, facing computational problems such as lossless compression, encryption of sensitive data, or efficient parallelization of a given task via hardware or software solutions. Also many of the analysis challenges

already discussed at the beginning of this chapter remain, including the representation of thousands of reference sequences in a pangenome structure, the efficient alignment to such data structures or the development of online methods enabling data analysis without long-term memory. These needs emphasize once more the value of our early contribution of the first fully variation-aware RNA-Seq aligner. However, also other interesting methodological contributions to address the above mentioned topics have been already made, including the compressed representation of genomes [63, 302], strategies to build a pangenome structure [213] or the efficient matching of haplotypes on large datasets [72].

A second paradigm shift expected in the near future is the quality change of sequencing data. Single molecule sequencing techniques that are able to produce reads of several kilobases in length and are unbiased from amplification have promising applications in transcript identification or genome assembly. Although some sources report unbiased, uniform error distributions [145, 236], the error profiles of such reads are not fully understood yet and the various possible biases have to be investigated to take them properly into account for quantitative approaches. Another interesting emerging technology is single cell sequencing. While most current sequencing approaches aggregate DNA or RNA from a population of cells, thus producing a signal that is only a population average, single cell sequencing techniques specifically aim at sequences contained within an individual cell. First studies applying this technology have already produced interesting insights into allelic expression imbalance [62], the mutational landscape within single tumors [323] or the clonal evolution of cancer cells [211]. In the long term, a large scale application of single cell RNA-Seq to a whole population of cells has the potential to further the understanding of transcriptional regulation and finally establish a model of the relationship between transcriptome and cell identity. First work in that direction already shows promising results [120]. Currently, single cell sequencing techniques based on deep sequencing suffer from artifacts caused by amplification especially for lowly expressed transcripts. Robust models taking this into account need to be developed. A first approach to account for such technical noise is discussed in [34].

In context of the exciting new possibilities provided by more and higher-quality data we also envision new analysis strategies that overcome the existing limitations of purely descriptive models. Especially techniques from machine learning are well suited to use the ever increasing data to learn predictive models that are able to transform the measured molecular phenotypes into a prediction with biological or medical relevance. We are confident that our methods for the high-throughput processing of transcriptome sequencing data are an important but very first step in this direction and provide an excellent basis for further research and improvement.

A Appendix

A.1 Variant-aware Alignments with PALMapper

This appendix contains additional material to the method PALMapper that has been described in Section 2.1, beginning on page 28. We give an overview of available command line parameters and provide the settings that have been used to generate the evaluation data.

PALMapper User Interface

Here, we provide the summary view of PALMapper's user interface as mentioned in its description in Section 2.1.7 on page 38.

```
PALMapper version 0.6   (PALMapper is a fusion of GenomeMapper & QPALMA)
written by Gunnar Raetsch, Geraldine Jean, Andre Kahles, Korbinian Schneeberger, Joerg Hagmann, Fabio De Bona, Stephan Ossowski, and others
Sloan-Kettering Institute, New York City, USA, 2012-2013
Max Planck Institute for Developmental Biology and Friedrich Miescher Laboratory, Tuebingen, Germany, 2008-2010

USAGE: palmapper [options]

mandatory:
 -i STRING                       reference sequence (fasta file and prefix to index files)
 -q STRING[,STRING,..,STRING]    query filename (fasta, fastq, or SHORE flat file)
 -q1 STRING[,STRING,..,STRING]   "left" query filename for paired-end reads (fasta, fastq, or SHORE flat file)
 -q2 STRING[,STRING,..,STRING]   "right" query filename for paired-end reads (fasta, fastq, or SHORE flat file)

optional:
 -stranded STRING                strand specific experiment (left, right, plus, minus)
 -protocol STRING                protocol used to prepared RNA-seq data (first-strand, second-strand, unstranded)
                                 examples: RNA ligation is first and dUTP protocol is second strand
 -f STRING                       output format ("shore", "bed", "bedx", "sam", "bam", "bamp" or "bamn")[sam]
 -samtools STRING                explicit samtools path (used for bam output)
 -ff INT                         bitwise output sam format flag
                                 (0x1: read sequence, 0x2: read quality, 0x4: common sam flags, 0x8: extended same flags)[15]
 -include-unmapped-reads         write directly unmapped reads in sam file
 -o STRING                       output filename [stdout]
 -H STRING                       output filename for spliced hits [no output]
 -u STRING                       output filename for unmapped reads [/dev/null]

 -rlim INT                       limit the number of reads for alignment
 -fromID STRING                  skip the first reads from query file until the readID is identical to the given one
 -from INT                       skip the first <from> reads from query file
 -to   INT                       map only the first <to> reads from query file

 -a                              report all alignments
 -ar INT                         report a limited number of alignments (random subset) [10]
 -z  INT                         report a number of top alignments [5]
 -n  INT                         report a maximal number of best alignments

 -r                              disable alignment on reverse strand [enabled]
 -h                              perform alignment of flanking regions of hits first [whole read alignment]
 -d                              align gaps most right (ignored for spliced alignments) [most left]
 -w                              allow more gaps for best hit (ignored for spliced alignments) [retain gap limit]

 -bwa INT                        use burrows-wheeler index instead of k-mer index (bwa-based) with a given seed length
 -seed-hit-cancel-threshold INT  number of hits of a seed that lead to its ignoration
 -index-extend-threshold INT     number of hits of a seed that lead to a seed-length extension
 -index-extend INT               length of seed-length extension
 -index-precache                 linearly read index file to fill caches
 -l INT                          minimal considered hit length [seed length]
 -c INT                          seed container size [15.000.000]

 -threads INT                    maximal number of threads [1]
 -v                              verbose [silent]
```

```
 -rtrim INT                          shortens the read until a hit is found or the minimal length is reached
 -rtrim-step INT                     rtrim step size
 -polytrim INT                       trims polyA or polyT ends until a hit is found or the minimal length is reached
 -fixtrim INT                        shortens the read to a fixed length
 -fixtrimleft INT                    Removes the given number of first nucleotides of each read
                                     (can be used with -fixtrimright but not -fixtrim)
 -fixtrimright INT                   Removes the given number of last nucleotides of each read
                                     (can be used with -fixtrimleft but not -fixtrim)
 -M INT                              max number of mismatches [auto]
 -G INT                              max number of gaps [auto]
 -E INT                              max edit operations [auto]
 -m DOUBLE                           mismatch penalty [4]
 -g DOUBLE                           gap penalty [5]
 -match-score DOUBLE                match penalty [0]

 -S                                  report spliced alignments (detailed options below)

spliced alignment definitions: (-S required)
 -qpalma STRING                      file name with qpalma parameters (essential)
 -qpalma-use-map-max-len INT         limit the map extension up- and downstream to the given length [10.000]
 -qpalma-prb-offset-fix              automatically fix the quality offset, if necessary

 -acc STRING                         path name to acceptor splice site predictions (essential if -no-ss-pred not provided)
 -don STRING                         path name to donor splice site predictions (essential if -no-ss-pred not provided)
 -acc-consensus STRING               defines consensus sequences for acceptor sites (separated by ",") [AG]
 -don-consensus STRING               defines consensus sequences for donor sites (separated by ",") [GT,GC]
 -no-ss-pred                         indicates that no splice site predictions should be used and only scores positions
                                     corresponding to consensus sequences for acceptors and donors
 -non-consensus-search               switch on spliced alignments with non consensus sequences as plausible splice sites
 -score-annotated-splice-sites STRING[,STRING,..,STRING]  set score of annotated splice sites from gff3 files to 1

 -junction-remapping STRING[,STRING,..,STRING]  enables remapping of unmapped or unspliced reads against the junction list
                                     provided in gff3 files
 -junction-remapping-coverage INT    minimum alignment support to take into account a junction
 -report-junctions STRING            report splice site junctions in gff3 format

 -use-variants STRING                Use variants provided in a sdi, maf, mgf, vcf or samtools file to map reads against
 -use-iupac-snp-variants             Enables the merge of SNPs and DNA base for aligning with variants
                                     (no snps reported in this case)
 -mgf-ref STRING                     Name of the reference genome as it appears in multiple alignments for MGF file given
                                     with -use-variants option
 -discover-variants                  Switch on the discovery of new variant sequences (deletion, insertion, SNP)
 -report-variants STRING             report variants (used and discovered)

 -filter-splice-sites-top-perc FLOAT trigger spliced alignments, if read covers top percentile splice site (between 0 and 1) [0.01]
 -filter-splice-region INT           extension of the read region up- and downstream for triggeringspliced alignments by presence of
                                     splice sites [5]
 -filter-max-edit INT                trigger spliced alignment, if unspliced alignment has at least this many edit operations [0]
 -filter-max-mismatches INT          trigger spliced alignment, if unspliced alignment has at least this many mismatches [0]
 -filter-max-gaps INT                trigger spliced alignment, if unspliced alignment has at least this many gaps [0]
 -log-triggered-reads STRING         log file containing the triggered reads

 -C INT                              min combined length [auto]
 -L INT                              min length of long hit [auto]
 -K INT                              min length of short hit [auto]
 -SA INT                             maximum number of spliced alignments per read [10]
 -NI INT                             maximum number of introns in spliced alignments [auto]
 -CT INT                             distance to tolerate between hit and existing hit cluster [10]
 -QMM INT                            number of matches required for identifying a splice site [auto]
 -I INT                              longest intron length  [auto]
 -MI INT                             shortest intron length  [30]
 -min-spliced-segment-len INT        minimal exon length [auto]
 -report STRING                      file for map reporting
 -report-ro STRING                   file for map reporting (read only)
 -report-rep-seed                    switch on reporting of repetitive seeds
 -report-map-region                  switch on reporting of mapped regions
 -report-map-read                    switch on reporting of mapped reads
 -report-spliced-read                switch on reporting of spliced reads
 -report-splice-sites FLOAT          report splice sites with confidence not less that threshold
 -report-splice-sites-top-perc FLOAT report splice sites with confidence in top percentile (between 0 and 1)
 -report-gff-init STRING             initialize map with exons from GFF file
 -report-coverage-map STRING         report genome coverage in map format
 -report-coverage-wig STRING         report genome coverage in wiggle format
```

Artificial read generation

Reads have been simulated with FluxSimulator (version 1.1.1-20121103021450, [102]). The following parameters were used:

```
EXPRESSION_X0       9500
EXPRESSION_K        -0.6
TSS_MEAN            50
POLYA_SCALE         300
POLYA_SHAPE         2
FRAG_SUBSTRATE      DNA
FRAG_METHOD         NB
FRAG_NB_LAMBDA      500
FILTERING           YES
SIZE_DISTRIBUTION   N-300-50.txt
SIZE_SAMPLING       AC
RTRANSCRIPTION      YES
PCR_PROBABILITY     0.7
RT_PRIMER           PDT
RT_LOSSLESS         YES
RT_MIN              500
RT_MAX              5500
PAIRED_END          YES
FASTA               YES
NB_MOLECULES        40000000
READ_NUMBER         10000000
READ_LENGTH         76
```

The 10,000 data points for the fragment size distribution were sampled randomly with Matlab from a Gaussian distribution with mean 300 and standard deviation 50.

A.2 Evaluation of RNA-Seq Alignments

In this appendix, we summarize information regarding the evaluation of alignment tools described in Section 2.2, beginning on page 38. Input data was provided in context of the RGASP competition as is described below. Information regarding the data was taken from [142]. We further provide an overview on different filter combinations used to optimize the single submissions.

RNA-Seq Data Used For RGASP

```
HUMAN
=====
1. experiment: Homo sapiens polyA+ total RNA, paired reads, HepG2
   lab:        Wold lab, Caltech
   format:     fastq, tar archive with bzipped files
   other details: 75mer sequences, the last base has been removed
                  _1 & _2 are the corresponding pairs
                  includes spike-in sequences for quantification
                  quality scores are Sanger rather than Illumina
                  fragment length is 200bp with a std deviation of 34
WORM
====
1. experiment: Caenorhabditis elegans polyA+ total RNA, paired reads, L3 phase
   lab:        Sternberg lab/Wold lab, Caltech
   format:     fastq, tar archive with bzipped files
   other details: 75mer sequences, the last base has been removed
                  _1 & _2 are the corresponding pairs
                  includes spike-in sequences for quantification
                  quality scores are Sanger rather than Illumina
  fragment length is 165bp with a standard deviation of 28
FLY
===
1. experiment: Drosophila melanogaster polyA+ total RNA, paired reads, L3 stage larvae
   lab:        Celniker lab, Lawrence Berkeley National Laboratory
   format:     fastq, tar archive with gzipped files
   other details: 76mer sequences
                  _1 & _2 are the corresponding pairs
                  produced on an Illumina Genome Analyzer II
                  fragment length is 250-300bp
  low quality reads have been filtered out
```

Genome Versions Used For RGASP

Organism	Version	Info/URL
C. elegans	WS200	http://wiki.wormbase.org/index.php/WS200
D. melanogaster	dmel_r5.20	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/... ...dmel_r5.20_FB2009_07/dna/
H. sapiens	GRCh37	ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/... ...vertebrates_mammals/Homo_sapiens/GRCh37/

Filter Combinations

We tested all 700 combinations of the following filter criteria to produce an optimally filtered set as used for evaluations described in Section 2.2.3, page 41.

Criterion	List of tested values
Min segment length in alignment	2, 4, 6, 8, 10, 12, 15, 20, 25, 30
Max number of mismatches	0, 1, 2, 3, 4, 5, 6
Min number of junction confirmations (split reads only)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

User Interface for Evaluation Tool

```
Usage: gen_alignment_statistics.py [options]

Options:
  -h, --help            show this help message and exit

  REQUIRED:
    -a FILE, --alignment=FILE
                        alignment file in sam format

  OPTIONAL:
    -R FILE, --ignore_multireads=FILE
                        file containing the multireads to ignore
    -g FILE, --genome=FILE
                        genome in fasta or hdf5 format (needs ending .hdf5 for
                        latter)
    -e INT, --min_exon_len=INT
                        minimal exon length [0]
    -X INT, --max_mismatches=INT
                        maximum number of allowed mismathes [-]
    -M INT, --max_intron_len=INT
                        maximal intron length [100000000]
    -I, --ignore_missing_chr
                        ignore chromosomes missing in the annotation
    -s, --shift_start   turn shifting start of softclips to accomodate for old
                        bug OFF - it is usually ON!
    -b, --bam_input     input has BAM format - does not work for STDIN
    -S PATH, --samtools=PATH
                        if SAMtools is not in your PATH, provide the right
                        path here (only neccessary for BAM input)
    -o PATH, --outfile_base=PATH
                        basedir for outfiles written
    -l INT, --lines=INT
                        maximal number of alignment lines to read [-]
    -v, --verbose       verbosity
    -d, --debug         print debugging output
```

A.3 Alignment Filtering

Here, we show the user interface of SAFT, the alignment tool described in Section 2.3, beginning on page 47. Other scripts for preprocessing are part of SAFT. Their user interfaces are similar and will not be shown here.

User Interface

```
Usage: find_optimal_param_set.py [options]

Options:
  -h, --help            show this help message and exit

  REQUIRED:
    -b FILE, --best_score=FILE
                        file to store the best scoring parameters
    -m FILE, --matrix=FILE
                        file to store the full performance matrix
    -f FILE, --features=FILE
                        alignment intron features
    -i FILE, --annotation_introns=FILE
                        annotation intron list

  OPTIONAL:
    -E STRINGLIST, --exclude_introns=STRINGLIST
                        list of comma separated intron files to exclude from
                        submitted features
    -I INT, --max_intron_len=INT
                        maximal intron length [10000000]
    -s, --ignore_strand
                        ignore strand information present in annotation
    -X INT, --max_feat_mismatches=INT
                        max number of mismatches for feat generation [80] (do
                        only change, if you are absolutely sure!)
    -v, --verbose       verbosity
```

A.4 MMR

In this apendix, we provide the user interface of the multi-mapper resolution tool (MMR) that is described in Section 2.4, beginning on page 52. The lower part of options is only relevant in context of the usage in conjunction with MiTie as described in Section 2.4.3.

MMR Output Screen

```
Usage:      ./mmr -o OUTFILE [options] IN_BAM

Available Options:
    Input handling and paralellization:
    -P --parse-complete         parse complete file into memory [off]
    -t --threads                number of threads to use (must be > 2) [1]
    -S --strand-specific        alignments are strand specific [off]
    -C --init-secondary         choose initial alignment also from secondary lines
                                (flag 256) [off]
    Input file filtering:
    -f --pre-filter-off         switch off pre filter for alignments that have F
                                more edit ops than the best [on]
    -F --filter-dist [INT]      filter distance F for pre-filter [1]
    -V --use-variants           use variant alignments for filtering (different edit
                                op count, requires XG and XM Tag in alignment files) [off]
    -L --max-list-length [INT]  max length of alignment list per read (after
                                filtering) [1000]
    Paired alignment handling:
    -p --pair-usage             pre use pair information in the reads [off]
    -i --max-fragment-size      upper limit of GENOMIC fragment length [1 000 000]
    -A --max-pair-list-length [INT] max no of valid pairs before not using pair
                                    modus [10000]
    Output handling:
    -b --best-only              print only best alignment [off]

    Options for using the variance optimization:
    -w --windowsize  [INT]      size of coverage window around read [20]
    -I --iterations  [INT]      number of iterations to smooth the coverage [5]

    Options for using the MiTie objective for smoothing:
    -m --mitie-objective        use objective from MiTie instead of local variance [off]
    -s --segmentfile            MiTie segment file required for MiTie optimization []
    -l --lossfile               MiTie loss parameter file required for MiTie optimization []
    -r --read-len  [INT]        average length of the reads [75]
    -M --mitie-variance         use variance smoothing for regions with no MiTie
                                prediction [off]
    -z --zero-expect-unpred     initializes all covered but not predicted positions with
                                expectation 0.0 [off]
    General:
    -v --verbose                switch on verbose output [off]
    -h --help                   print usage info
```

A.5 Alternative Splicing Event Detection and Quantification

In this appendix, we summarize additional information relevant for the description of SplAdder, which we discussed in Section 2.5, beginning on page 61.

SplAdder User Interface (Matlab/Octave version)

```
Usage: SplAdder [-OPTION VALUE]

Options (default values in [...]):

MANDATORY:

-b  FILE1,FILE2,...     alignment files in BAM format (comma separated list)
-o  DIR                 output directory
-a  FILE                annotation file name (annotation in *.mat format)

OPTIONAL:
-l  FILE                log file name [stdout]
-u  FILE                file with user settings [-]
-F  FILE                use existing SplAdder output file as input (advanced) [-]
-c  INT                 confidence level (0 lowest to 3 highest) [3]
-I  INT                 number of iterations to insert new introns into the graph [5]
-M  <STRAT>             merge strategy, where <STRAT> is one on:
                        merge_bams, merge_graphs, merge_all [merge_graphs]
-n  INT                 read length (used for automatic conf. level settings) [36]
-R  R1,R2,...           replicate structure of files (same number as
                        alignment files) [all R1 - no replicated]
-L  STRING              label for current experiment [-]
-S  STRING              reference strain [-]
-C  y|n                 truncation detection mode [n]
-U  y|n                 count intron coverage [n]
-P  y|n                 only use primary alignments from provided files [n]
-d  y|n                 use debug mode [n]
-p  y|n                 use rproc [n]
-O  y|n                 annotation is in half-open coordinates
-V  y|n                 validate splice graph [n]
-v  y|n                 use verbose output mode [n]
-A  y|n                 curate alt prime events [y]
-x  y|n                 input alignments share the same genome [y]
-i  y|n                 insert intron retentions [y]
-e  y|n                 insert cassette exons [y]
-E  y|n                 insert new intron edges [y]
-r  y|n                 remove short exons [n]
-s  y|n                 re-infer splice graph [n]
-T  y|n                 extract alternative splicing events [y]
-X  y|n                 alignment files are variation aware (XM and XG tags present) [n]
-t  STRING,STRING,...   list of alternative splicing events to extract
                        [exon_skip,intron_retention,alt_3prime,alt_5prime,mult_exon_skip]
```

Confidence levels for graph augmentation

SplAdder has several confidence levels the user can choose from, ranging from 0 (lowest confidence) to 3 (highest confidence). The levels adjust filter parameters to 1) select high confidence alignments and 2) set the criteria for graph augmentation. The parameter r is the length of the reads in the RNA-Seq sample.

Settings for accepted introns

Criterion	**Confidence Level**			
	0	**1**	**2**	**3**
min segment length	$\lceil 0.1 \cdot r \rceil$	$\lceil 0.15 \cdot r \rceil$	$\lceil 0.2 \cdot r \rceil$	$\lceil 0.25 \cdot r \rceil$
max mismatches	$\max\{2, \lfloor 0.03 \cdot r \rfloor\}$	$\max\{1, \lfloor 0.02 \cdot r \rfloor\}$	$\max\{1, \lfloor 0.01 \cdot r \rfloor\}$	0
max intron length	20,000	20,000	20,000	20,000
min junction count	1	2	3	6

Settings for accepted cassette exons

Criterion	**Value**
min exon coverage	5
min fraction of covered positions in exon	0.9
min relative coverage difference to flanking exons	0.05

Settings for accepted intron retentions

Criterion	**confidence level**			
	0	**1**	**2**	**3**
min intron coverage	1	2	5	10
min fraction of covered positions in intron	0.75	0.75	0.9	0.9
min intron coverage relative to flanking exons	0.1	0.1	0.2	0.2
max intron coverage relative to flanking exons	2	1.2	1.2	1.2

Event validation criteria

Each event has different criteria for validation. The table below lists all criteria for the different types of events.

Exon Skips	
Criterion	**Value**
min relative coverage difference to flanking exons	0.05
min intron count confirming the skip	3
min intron count confirming the inclusion	3

Intron Retentions	
Criterion	**Value**
min intron coverage	3
min intron coverage relative to flanking exons	0.05
min fraction of covered positions in the intron	0.75
min intron count confirming the intron	3

Alternative Splice Site Choice	
Criterion	**Value**
min intron count confirming the intron	3
min relative difference of differential exon part to flanking exon	0.05

A.6 Analysis of AS dependent NMD in *A. thaliana*

In this appendix, we provide additional material for the study on NMD in *A. thaliana* that is discussed in Section 3.1, beginning on page 76. Here, we summarize parameter choices of the relevant analysis parts and provide statistics regarding the alignment of the read data.

PALMapper parameters

Full list of command line parameters used for the alignments with PALMapper.

```
-M 6 -G 1 -E 6 -l 15 -L 25 -K 8 -C 35 -I 25000 -NI 2 -SA 100 -CT 50 -a -S
-fixtrimleft 4 -fixtrimright 4
-seed-hit-cancel-threshold 10000
-report-map-read
-report-spliced-read
-report-map-region
-report-splice-sites 0.9
-filter-max-mismatches 0
-filter-max-gaps 0
-filter-splice-region 5
-qpalma-use-map-max-len 1000
-f bamn -threads 2 -polytrim 40
-qpalma-prb-offset-fix
-min-spliced-segment-len 15
-junction-remapping-coverage 5
-junction-remapping-min-spliced-segment-len 15
-junction-remapping <JUNCTION_GFF>
-score-annotated-splice-sites <JUNCTION_GFF>
-qpalma-indel-penalty 1
```

SplAdder parameters

List of non-default parameters used for the SplAdder splicing graph augmentation:

Parameter	**Value**
Maximum intron length	20,000
Minimum segment length for spliced alignments	25
Maximum number of edit operations per alignment	0
Minimum splice junction support	2
Minimum intron retention coverage	10
Minimum relative covered position in intron retention	0.9
Minimum relative coverage in intron retention regions	0.2
Maximum relative coverage in intron retention regions	1.2

Alignment statistics

	tot. reads	aligned reads		unaligned reads		reads w/ spliced aln.		uniquely mapped reads		
	absolute	absolute	% tot.	absolute	% tot.	absolute	% aln.	absolute	% aln.	% tot.
WT R1	59,943,321	38,806,708	65	21,136,613	35	11,940,504	31	38,806,708	100	65
WT R2	48,594,361	43,162,399	89	5,431,962	11	14,371,835	33	43,162,399	100	89
lba1 R1	60,577,841	41,824,267	69	18,753,574	31	12,415,908	30	41,824,267	100	69
lba1 R2	50,870,221	45,183,966	89	5,686,255	11	14,688,425	33	45,183,966	100	89
upf3-1 R1	53,540,941	42,206,547	79	11,334,394	21	12,639,791	30	42,206,547	100	79
upf3-1 R2	55,415,021	48,054,773	87	7,360,248	13	15,698,446	33	48,054,773	100	87
lba1upf3-1 R1	55,490,061	32,871,741	59	22,618,320	41	9,368,074	28	32,871,741	100	59
lba1upf3-1 R2	56,983,201	42,037,471	74	14,945,730	26	11,649,145	28	42,037,471	100	74
lba1upf3-1 R3	51,922,341	45,955,038	89	5,967,303	11	14,449,605	31	45,955,038	100	89
Mock R1	53,347,969	40,896,391	77	12,451,578	23	11,307,635	28	40,896,391	100	77
Mock R2	50,703,357	45,399,929	90	5,303,428	10	13,823,986	30	45,399,929	100	90
CHX R1	56,579,469	40,714,682	72	15,864,787	28	10,069,945	25	40,714,682	100	72
CHX R2	47,569,604	43,099,959	91	4,469,645	9	10,703,309	25	43,099,959	100	91

A.7 Analysis of PTB Dependent Splicing in *A. thaliana*

In this appendix, we provide additional material to the work on PTB-dependent splicing that is discussed in Section 3.2, beginning on page 83. Here, we summarize parameter choices of the relevant analysis parts and provide statistics regarding the alignment of the read data.

PALMapper parameters

Full list of command line parameters used for the alignments with PALMapper.

```
-M 6 -G 1 -E 6 -l 15 -L 25 -K 8 -C 35 -I 25000 -NI 2 -SA 100 -CT 50 -a -S
-seed-hit-cancel-threshold 10000
-report-map-read
-report-spliced-read
-report-map-region
-report-splice-sites 0.9
-filter-max-mismatches 0
-filter-max-gaps 0
-filter-splice-region 5
-qpalma-use-map-max-len 1000
-f bamn -threads 2 -polytrim 40
-qpalma-prb-offset-fix
-min-spliced-segment-len 15
-junction-remapping-coverage 5
-junction-remapping-min-spliced-segment-len 15
-junction-remapping <JUNCTION_GFF>
-score-annotated-splice-sites <JUNCTION_GFF>
-qpalma-indel-penalty 1
```

SplAdder parameters

List of non-default parameters used for the SplAdder splicing graph augmentation:

Parameter	**Value**
Maximum intron length	20,000
Minimum segment length for spliced alignments	25
Maximum number of edit operations per alignment	0
Minimum splice junction support	2
Minimum intron retention coverage	10
Minimum relative covered position in intron retention	0.9
Minimum relative coverage in intron retention regions	0.2
Maximum relative coverage in intron retention regions	1.2

Alignment statistics

	Run	tot. reads	aligned reads		unaligned reads		reads w/spliced aln.		uniquely mapped reads		
		absolute	absolute	% tot.	absolute	% tot.	absolute	% aln.	absolute	% aln.	% tot.
pGPTV	I	49,603,847	42,568,789	88	6,035,058	12	10,917,353	26	42,568,789	100	88
ami1ami2-1		59,709,514	49,261,885	83	10,447,629	17	12,541,291	25	49,261,885	100	83
ami3-1		53,957,222	46,440,585	86	7,516,637	14	11,950,171	26	46,440,585	100	86
WT		57,430,351	48,684,668	85	8,745,683	15	12,629,273	26	48,684,668	100	85
OE2		39,320,273	36,233,684	92	3,086,589	8	9,542,312	26	36,233,684	100	92
OE3		57,351,910	46,735,094	81	10,616,816	19	12,010,470	26	46,735,094	100	81
OE1	II	60,337,363	50,523,695	84	9,813,668	16	16,373,725	32	50,523,695	100	84
WT		52,233,547	46,213,819	88	6,019,728	12	14,379,023	31	46,213,819	100	88
WT	III	52,247,406	42,884,306	82	9,363,100	18	13,619,535	32	42,884,306	100	82
ami1ami2-1		54,754,525	44,377,332	81	10,377,193	19	13,628,798	31	44,377,332	100	81
OE1		58,194,346	45,443,077	78	12,751,269	22	13,653,884	30	45,443,077	100	78
OE2		59,204,960	45,411,890	77	13,793,070	23	14,360,969	32	45,411,890	100	77
OE3		52,790,572	43,222,674	82	9,567,898	18	13,532,156	31	43,222,674	100	82

A.8 Identification of sQTL in Two *A. thaliana* Populations

In this appendix, we provide additional material for the sQTL in two populations of *A. thaliana* that is described in Section 3.3 of this work, beginning on page 89.

PALMapper parameters – CEGS initial alignments

This is an overview of parameters used for the initial alignment run on the CEGS data set to generate a list of trustworthy junctions:

```
-M 4 -G 4 -E 6 -l 12 -L 15 -K 12 -C 12 -I 25000 -NI 1 -SA 10 -CT 50 -a -S
-seed-hit-cancel-threshold 10000
-report-map-read
-report-spliced-read
-report-map-region
-report-splice-sites 0.9
-filter-max-mismatches 1
-filter-max-gaps 0
-filter-splice-region 5
-min-spliced-segment-len 1
-qpalma-indel-penalty 5
-qpalma-use-map-max-len 10000
-f bam
-qpalma-prb-offset-fix
-discover-variants
-report-variants <VARIANT.sdi>
-fixtrimleft 4
```

PALMapper parameters – Final uniform alignment run for CEGS and MAGIC

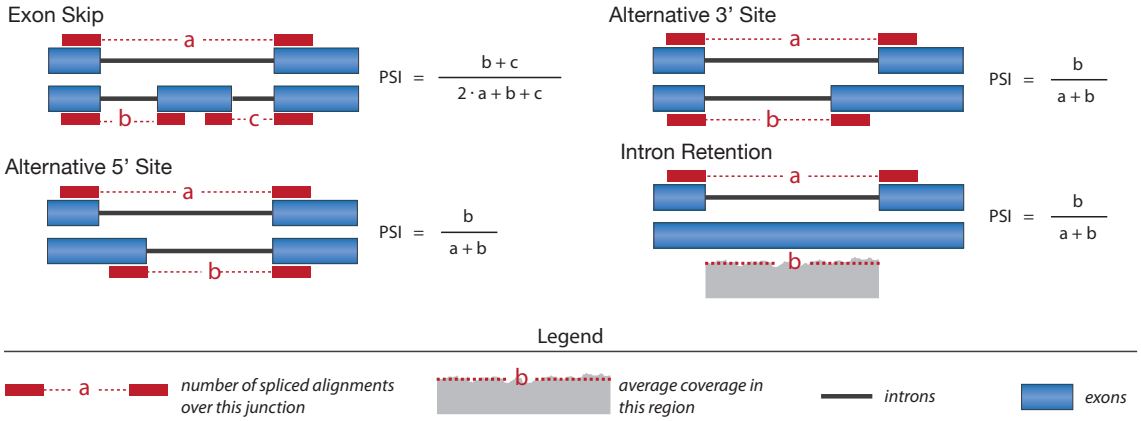
Below are the alignment parameters that were used in the final alignment runs. CEGS and MAGIC datasets have been aligned uniformly using the same parameters. Exceptions are due to adapter trimming in a part of the reads and are marked in the parameter list.

```
-M 3 -G 0 -E 3 -l 12 -L 14 -K 12 -C 14 -I 5000 -NI 1 -SA 5 -UA 50
-CT 50 -JA 15 -JI 1 -z 10 -S
-seed-hit-truncate-threshold 100
-report-map-read
-report-spliced-read
-report-map-region
-report-splice-sites 0.9
-filter-max-mismatches 0
-filter-max-gaps 0
-filter-splice-region 5
-min-spliced-segment-len 1
-qpalma-use-map-max-len 10
-f bam
-qpalma-prb-offset-fix
-junction-remapping <JUNCTIONS.IN>
-score-annotated-splice-sites <JUNCTIONS.IN>
```

```
-max-dp-deletions 2
-use-variants-editop-filter
-use-variants <VARIANTS.IN>
-report-variants <VARIANTS.OUT>
-report-used-variants <VARIANTS_USED.OUR>
-filter-variants-minuse 1
-merge-variant-source-ids
-use-iupac-snp-variants
-report-ro <ALIGNMENTMAP>
-filter-variants-map-window 20
-iupac-genome
-fixtrimleft 4 (ONLY FOR CEGS/TRIMMED)
-fixtrim 32 (ONLY FOR CEGS/TRIMMED)
-filter-variants-maxlen 100
-index-precache
```

A.9 PSI Computation

The computation of the percent spliced in (PSI) values for the single event types as it is used in the sQTL analyses described in Section 3.3, page 89, and Section 3.4, page 98, was done as follows. Each isoform is represented either as spliced alignment evidence or as mean coverage in a region. The ratio of the longer isoform over the sum of both isoforms is taken as the percent spliced in value.



A.10 Splicing QTL in 12 Cancer Types

This section contains supplementary information to data and analyses used for the splicing QTL study on data from The Cancer Genome Atlas (TCGA). Set-up and results of this study are discussed in Section 3.4 of this work, beginning on page 98.

Cancer Types

Cancer Type	Abbreviation
Bladder Urothelial Carcinoma	BLCA
Breast invasive carcinoma	BRCA
Colon adenocarcinoma	COAD
Glioblastoma multiforme	GBM
Head and Neck squamous cell carcinoma	HNSC
Kidney renal clear cell carcinoma	KIRC
Lung adenocarcinoma	LUAD
Lung squamous cell carcinoma	LUSC
Ovarian serous cystadenocarcinoma	OV
Rectum adenocarcinoma	READ
Thyroid carcinoma	THCA
Uterine Corpus Endometrial Carcinoma	UCEC

Overview of RNA-Seq Samples

Cancer Type	Tumor Samples	Normal Samples	Total Samples
BLCA	122	16	138
BRCA	843	105	948
COAD	194	-	194
GBM	168	-	168
HNSC	302	-	302
KIRC	481	71	552
LUAD	355	56	411
LUSC	309	24	333
OV	418	-	418
READ	71	-	71
THCA	493	58	551
UCEC	317	-	317

Overview of Exome-Seq Samples

Cancer Type	Tumor Samples	Normal Samples	Total Samples
BLCA	142	121	263
BRCA	877	908	1,785
COAD	350	395	745
GBM	173	194	367
HNSC	280	342	622
KIRC	332	347	679
LUAD	422	435	857
LUSC	383	426	809
OV	309	414	723
READ	136	163	299
THCA	449	478	927
UCEC	460	478	938

Alignment Parameters for Exome-Seq Data

```
--ReadGroup <TCGAID>
--genomeDir <genome>
--readFilesIn <fastq-left> <fastq-right>
--runThreadN 5
--outFilterMultimapScoreRange 2
--outFilterMultimapNmax 100
--outFilterMismatchNmax 10
--alignMatesGapMax 1000000
--genomeLoad LoadAndKeep
--scoreGap -8
--scoreGapNoncan 0
--scoreGapGCAG 0
--scoreGapATAC 0
--scoreStitchSJshift 0
--alignIntronMax 0
--alignIntronMin 100
```

Alignment Parameters for RNA-Seq Data

```
--genomeDir <genome>
--readFilesIn <fastq-left> <fastq-right>
--runThreadN 5
--outFilterMultimapScoreRange 2
--outFilterMultimapNmax 100
--outFilterMismatchNmax 10
--alignIntronMax 500000
--alignMatesGapMax 1000000
--sjdbFileChrStartEnd <junctionDB>
--sjdbScore 1
--genomeLoad LoadAndKeep
```

B Bibliography

[1] Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

[2] Abouelhoda, M. I., Kurtz, S., and Ohlebusch, E. (2002). The enhanced suffix array and its applications to genome analysis. *WABI*, pages 449–463.

[3] Abouelhoda, M. I., Kurtz, S., and Ohlebusch, E. (2004). Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, **2**(1), 53–86.

[4] Adams, M., Celniker, S., and Holt, R. (2000). The genome sequence of Drosophila melanogaster. *Science*, **287**(5461), 2185–2195.

[5] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., and Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**(5013), 1651–1656.

[6] Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B.-C., Kim, S.-Y., Kim, W.-Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S. *et al.* (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research*, **19**(9), 1622–1629.

[7] Albert, T. J. (2003). Light-directed 5' -> 3' synthesis of complex oligonucleotide microarrays. *Nucleic Acids Research*, **31**(7), 35e–35.

[8] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and And Walter, P. (2008). *Molecular Biology of the Cell*, volume 54. Garland Science.

[9] Altschul, S. F. and Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, **48**(5-6), 603–616.

[10] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

[11] Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O. *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

[12] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

[13] Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, **25**(4), 195–203.

[14] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1), 25–29.

[15] Asp, J., Persson, F., Kost-Alimova, M., and Stenman, G. (2006). CHCHD7-PLAG1 and TCEA1-PLAG1 gene fusions resulting from cryptic, intrachromosomal 8q rearrangements in pleomorphic salivary gland adenomas. *Genes Chromosomes and Cancer*, **45**(9), 820–828.

[16] Astle, W. and Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, **24**(4), 451–471.

[17] Au, K. F., Jiang, H., Lin, L., Xing, Y., and Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, **38**(14), 4570–4578.

[18] Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. a., Luo, S., Schroth, G. P., Doudna, J. a., Arkin, A. P., and Pachter, L. (2011). Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(27), 11069–11074.

[19] Ayoubi, T. a. and Van De Ven, W. J. (1996). Regulation of gene expression by alternative promoters. *FASEB: Official Publication of the Federation of American Societies for Experimental Biology*, **10**(4), 453–460.

[20] Ayub, M. and Bayley, H. (2012). Single Molecule RNA Base Identification with a Biological Nanopore. *Biophysical Journal*, **102**(3), 429a.

[21] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, **389**(4), 1017–1031.

[22] Barbazuk, W. B., Fu, Y., and McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Research*, **18**(9), 1381–1392.

[23] Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T. *et al.* (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**(6114), 1587–1593.

[24] Baud, A., Hermsen, R., Guryev, V., Stridh, P., Graham, D., McBride, M. W., Foroud, T., Calderari, S., Diez, M., Ockinger, J., Beyeen, A. D., Gillett, A., Abdelmagid, N., Guerreiro-Cacais, A. O., Jagodic, M. *et al.* (2013). Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nature Genetics*, **45**(7), 767–775.

[25] Behr, J., Kahles, A., Zhong, Y., Sreedharan, V. T., Drewe, P., and Rätsch, G. (2013). MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, **29**(20), 2529–2538.

[26] Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., Godwin, A. K., Gross, J., Hartmann, L., Huang, M., Huntsman, D. G. *et al.* (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.

[27] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **51**(1), 289 – 300.

[28] Binder, H., Kirsten, T., Loeffler, M., and Stadler, P. F. (2004). Sensitivity of microarray oligonucleotide probes: Variability and effect of base composition. *Journal of Physical Chemistry B*, **108**(46), 18003–18014.

[29] Boguski, M., Tolstoshev, C., and Jr, D. B. (1994). Gene discovery in dbEST. *Science*, **265**(5181), 1993–1994.

[30] Bohnert, R. and Rätsch, G. (2010). rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Research*, **38**(Web Server issue), W348–351.

[31] Bohnert, R., Behr, J., and Rätsch, G. (2009). Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, **10**(Suppl 13), P5.

[32] Bonfert, T., Csaba, G., Zimmer, R., and Friedel, C. C. (2012). A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinformatics*, **13 Suppl 6**(Suppl 6), S9.

[33] Bonnal, S., Vigevani, L., and Valcárcel, J. (2012). The spliceosome as a target of novel antitumour drugs. *Nature Reviews Drug Discovery*, **11**(11), 847–859.

[34] Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. a., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, **10**(11), 1093–5.

[35] Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbrück, S., Krueger, S., Reich, J., and Bork, P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**(1), 83–86.

[36] Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, **21**(2), 193–202.

[37] Brooks, A. N., Choi, P. S., de Waal, L., Sharifnia, T., Imielinski, M., Saksena, G., Pedamallu, C. S., Sivachenko, A., Rosenberg, M., Chmielecki, J., Lawrence, M. S., DeLuca, D. S., Getz, G., and Meyerson, M. (2014). A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PloS One*, **9**(1), e87361.

[38] Burns, P. D., Li, Y., Ma, J., and Borodovsky, M. (2013). UnSplicer: mapping spliced RNA-seq reads in compact genomes and filtering noisy splicing. *Nucleic Acids Research*, **42**(4), e25.

[39] Burrows, M. and Wheeler, D. (1994). A block-sorting lossless data compression algorithm. *Systems Research*, **R**(124), 24.

[40] Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G., and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics*, **37**(8), 868–872.

[41] Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., and DePristo, M. a. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, **13**(1), 375.

[42] Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, **101**(6), 577–580.

[43] Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, **3**(4), 285–298.

[44] Chaisson, M. J., Brinza, D., and Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, **19**(2), 336–346.

[45] Chang, Y., Imam, J., and Wilkinson, M. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry*, **76**, 51–74.

[46] Chao, K. M., Pearson, W. R., and Miller, W. (1992). Aligning two sequences within a specified diagonal band. *Computer Applications in the Biosciences*, **8**(5), 481–487.

[47] Chi, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., Gore, M., Guill, K. E., Holland, J., Hufford, M. B., Lai, J. *et al.* (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, **44**(7), 803–807.

[48] Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(45), 19096–19101.

[49] Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K., and Kreder, D. E. (2002). Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biology*, **3**(1), RESEARCH0005.

[50] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**(3), 213–219.

[51] Conti, E. and Izaurralde, E. (2005). Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Current Opinion in Cell Biology*, **17**(3), 316–325.

[52] Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic Acids Research*, **13**(9), 3021–3030.

[53] Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine & Biotechnology*, **2010**.

[54] Cox, M. P., Peterson, D. a., and Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

[55] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. a., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G. *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, **16**(1), 123–131.

[56] Crick, F. H. (1958). The biological replication of macromolecules. *Symposia of the Society for Experimental Biology*, **12**, 138–163.

[57] Crick, F. H. (1970). Central Dogma of Molecular Biology. *Nature*, **227**(8), 561–563.

[58] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158.

[59] De Bona, F., Ossowski, S., Schneeberger, K., and Rätsch, G. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**(16), i174–180.

[60] De Bruijn, N. G. and Erdos, P. (1946). A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, **49**(49), 758–764.

[61] de Koning, a. P. J., Gu, W., Castoe, T. a., Batzer, M. a., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics*, **7**(12), e1002384.

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

[63] Deorowicz, S., Danek, A., and Grabowski, S. (2013). Genome compression: a novel approach for large collections. *Bioinformatics*, **29**(20), 2572–2578.

[64] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K. *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**(5), 491–498.

[65] Di Giammartino, D. C., Nishida, K., and Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, **43**(6), 853–866.

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

[67] Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**(16), e105.

[68] Drechsel, G., Kahles, A., Kesarwani, A. K., Stauffer, E., Behr, J., Drewe, P., Rätsch, G., and Wachter, A. (2013). Nonsense-Mediated Decay of Alternative Precursor mRNA Splicing Variants Is a Major Determinant of the Arabidopsis Steady State Transcriptome. *The Plant Cell*, **25**(10), 3726–3742.

[69] Drewe, P., Stegle, O., Hartmann, L., Kahles, A., Bohnert, R., Wachter, A., Borgwardt, K., and Rätsch, G. (2013). Accurate detection of differential RNA processing. *Nucleic Acids Research*, **41**(10), 5189–5198.

[70] D'Souza, I. and Schellenberg, G. D. (2005). Regulation of tau isoform expression and dementia. *Biochimica et Biophysica Acta*, **1739**(2), 104–115.

[71] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F. *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

[72] Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*, **30**(9), 1266–1272.

[73] Durbin, R., Eddy, S. R., Krogh, A., and Mitchinson, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

[74] Eckardt, N. A. (2002). Alternative splicing and the control of flowering time. *The Plant Cell*, **14**(4), 743–747.

[75] Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

[76] Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews. Genetics*, **11**(6), 446–450.

[77] Eichner, J., Zeller, G., Laubinger, S., and Rätsch, G. (2011). Support vector machines-based identification of alternative splicing in Arabidopsis thaliana from whole-genome tiling arrays. *BMC Bioinformatics*, **12**(1), 55.

[78] ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**(5696), 636–640.

[79] Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., A Kahles, Rätsch, G., Goldman, N., Hubbard, T. J., Harrow, J., Guigó, R., and Others (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, **10**(12), 1185–1191.

[80] Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**(3), 186–194.

[81] Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, **25**(2), 232–234.

[82] Ferragina, P. and Manzini, G. (2001). An experimental study of a compressed index. *Information Sciences*, **135**(1-2), 13–28.

[83] Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, **7**(2), 85–97.

[84] Fisher, R. A. (1919). XV.–The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

[85] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. *et al.* (2014). Ensembl 2014. *Nucleic Acids Research*, **42**(Database issue), D749–755.

[86] Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). Genomic DNA Sequence A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. *Genome Research*, **8**(9), 967–974.

[87] Frith, M. C. and Noé, L. (2014). Improved search heuristics find 20 000 new alignments between human and mouse genomes. *Nucleic Acids Research*, **42**(7), e59.

[88] Fullwood, M. J., Wei, C.-L., Liu, E. T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, **19**(4), 521–532.

[89] Fusi, N., Stegle, O., and Lawrence, N. (2011). Accurate modeling of confounding variation in eQTL studies leads to a great increase in power to detect trans-regulatory effects. *Nature Precedings*, pages 1–12.

[90] Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, **8**(1), e1002330.

[91] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, **4**(3), 177–183.

[92] Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P. *et al.* (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, **108**(25), 10249–10254.

[93] Garber, M. and Grabherr, M. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, **8**(6), 469–477.

[94] Garcia-Blanco, M. a., Baraniak, A. P., and Lasda, E. L. (2004). Alternative splicing in disease and therapy. *Nature Biotechnology*, **22**(5), 535–546.

[95] Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P., Guyer, M., Peck, A. M., Derge, J. G., Lipman, D., Collins, F. S. *et al.* (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Research*, **14**(10B), 2121–2127.

[96] Ghigna, C., Moroni, M., Porta, C., Riva, S., and Biamonti, G. (1998). Altered Expression of Heterogeneous Nuclear Ribonucleoproteins and SR Factors in Human Colon Adenocarcinomas. *Cancer Research*, **58**(24), 5818–5824.

[97] Giannoukos, G., Ciulla, D. M., Huang, K., Haas, B. J., Izard, J., Levin, J. Z., Livny, J., Earl, A. M., Gevers, D., Ward, D. V., Nusbaum, C., Birren, B. W., and Gnirke, A. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*, **13**(3), R23.

[98] Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**(3), 705–708.

[99] Gott, J. and Emeson, R. (2000). FUNCTIONS AND MECHANISMS OF RNA EDITING. *Annual Review of Genetics*, **34**, 499–531.

[100] Grabherr, M., Haas, B., and Yassour, M. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, **29**(7), 644–652.

[101] Green, M. (1986). Pre-mRNA splicing. *Annual Review of Genetics*, **20**(1), 671–708.

[102] Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, **40**(20), 10073–10083.

[103] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, **28**(5), 503–510.

[104] Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Sougnez, C., Imielinski, M., Helman, E., Hernandez, B. *et al.* (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**(7417), 519–525.

[105] Hanke, J., Gross, S., and Reich, J. (2000). Alternative splicing EST analysis online : WWW tools for detection of SNPs and alternative splice forms. *Trends in Genetics*, **16**(9), 416–418.

[106] Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, **38**(12), e131.

[107] Harbers, M. and Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, **2**(7), 495–502.

[108] Harris, M. a., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. a., Bult, C., Dolan, M. *et al.* (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**(Database issue), D258–261.

[109] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M. *et al.* (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, **22**(9), 1760–1774.

[110] He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., Wang, Z., Chen, F., Lindquist, E. A., Sorek, R., and Hugenholtz, P. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods*, **7**(10), 807–812.

[111] Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, **18**(6), 341–343.

[112] Holbrook, J., Neu-Yilik, G., Hentze, M., and Kulozik, A. (2004). Nonsense-mediated decay approaches the clinic. *Nature Genetics*, **36**(8), 801–808.

[113] Hori, K. and Watanabe, Y. (2005). UPF3 suppresses aberrant spliced mRNA in Arabidopsis. *Plant Journal*, **43**(4), 530–540.

[114] Huang, J., Wei, W., Chen, J., Zhang, J., Liu, G., Di, X., Mei, R., Ishikawa, S., Aburatani, H., Jones, K. W., and Shapero, M. H. (2006). CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*, **7**, 83.

[115] Huber, W. and Anders, S. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

[116] Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D. *et al.* (2010). International network of cancer genome projects. *Nature*, **464**(7291), 993–998.

[117] Huse, S. M., Huber, J. a., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**(7), R143.

[118] Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., Hackett, J., Adamson, J., Lincoln, S., Dickson, D., Davies, P. *et al.* (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, **393**(6686), 702–705.

[119] Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2004). Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Research*, **32**(17), 5096–5103.

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

[121] Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F., and Rätsch, G. (2010). RNA-Seq read alignments with PALMapper. *Current Protocols in Bioinformatics*, **Chapter 11**(December), Unit 11.6.

[122] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**(5830), 1497–1502.

[123] Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**(5653), 2141–2144.

[124] Jünemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., and Harmsen, D. (2012). Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. *PloS One*, **7**(8), e41606.

[125] Kaiser, J. (2005). NIH to Draw Cancer Map. *Science*, **310**(5755), 1751.

[126] Kalyna, M., Simpson, C. G., Syed, N. H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., Dinh, H. Q., Barta, A., and Brown, J. W. S. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Research*, **40**(6), 2454–2469.

[127] Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B. *et al.* (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**(7447), 67–73.

[128] Kang, H., Sul, J., Service, S., and Zaitlen, N. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**(4), 348–354.

[129] Kapustin, Y., Souvorov, A., Tatusova, T., and Lipman, D. (2008). Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct*, **3**, 20.

[130] Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T. *et al.* (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, **42**(D1).

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

[132] Kaye, J., Hurles, M., Griffin, H., Grewal, J., Bobrow, M., Timpson, N., Smee, C., Bolton, P., Durbin, R., Dyke, S., Fitzpatrick, D., Kennedy, K., Kent, A., Muddyman, D., Muntoni, F. *et al.* (2014). Managing clinically significant findings in research: the UK10K example. *European Journal of Human Genetics*, (July 2013), 1–5.

[133] Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellå ker, C., Whitley, H., Cleak, J. *et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**(7364), 289–294.

[134] Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Research*, **12**(4), 656–664.

[135] Khoo, B. and Krainer, A. (2009). Splicing therapeutics in SMN2 and APOB. *Current Opinion in Molecular Therapeutics*, **11**(2), 108–115.

[136] Khorana, H. G. (1968). No Nucleic acid synthesis in the study of the genetic code. *Nobel Lectures: Physiology or Medicine (1963-1970)*, pages 341–369.

[137] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**(4), R36.

[138] Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing–concepts and limitations. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, **32**(6), 524–536.

[139] Klug, W. S., Cummings, M. R., Spencer, C. A., and Palladino, M. A. (2011). *Concepts of Genetics*. Benjamin Cummings.

[140] Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R. *et al.* (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

[141] Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nature Methods*, **3**(3), 211–222.

[142] Kokocinski, F. (2009). Data files for RGASP 2009 round 2. *ftp://ftp.sanger.ac.uk/pub/gencode/rgasp/RGASP2/inputdata/_README.TXT*.

[143] Kole, R., Krainer, A. R., and Altman, S. (2012). RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nature Reviews Drug Discovery*, **11**(2), 125–140.

[144] Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J. *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**(5849), 420–426.

[145] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. a., McCombie, W. R., Jarvis, E. D., and Adam M Phillippy (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**(7), 693–700.

[146] Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, **14**(3), 153–165.

[147] Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, **5**(7), e1000551.

[148] Krebs, J. E., Goldstein, E. S., and Kilpatrick, S. T. (2012). *Lewin's Genes*. Jones & Bartlett Learning.

[149] Kurihara, Y., Matsui, A., Hanada, K., Kawashima, M., Ishida, J., Morosawa, T., Tanaka, M., Kaminuma, E., Mochizuki, Y., Matsushima, A., Toyoda, T., Shinozaki, K., and Seki, M. (2009). Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(7), 2453–2458.

[150] Lalonde, E., Ha, K., and Wang, Z. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, **21**(4), 545–554.

[151] Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S. *et al.* (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(Database issue), D1202–1210.

[152] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

[153] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357–359.

[154] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**(3), R25.

[155] Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J. *et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468), 506–511.

[156] Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., and Brenner, S. E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**(7138), 926–929.

[157] Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, **396**(6712), 643–649.

[158] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**(5607), 682–686.

[159] Levin, J., Yassour, M., and Adiconis, X. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, **7**(9), 709–715.

[160] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B. *et al.* (2007). The diploid genome sequence of an individual human. *PLoS Biology*, **5**(10), e254.

[161] Lewis, B. P., Green, R. E., and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(1), 189–192.

[162] Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

[163] Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, **83**(3), 311–321.

[164] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. a., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4), 493–500.

[165] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5), 589–595.

[166] Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18**(11), 1851–1858.

[167] Li, H., Handsaker, B., Wysoker, A., and Fennell, T. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

[168] Li, J. J., Bickel, P. J., and Biggin, M. D. (2014a). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, **2**, e270.

[169] Li, Q., Stram, A., Chen, C., Kar, S., Gayther, S., Pharoah, P., Haiman, C., Stranger, B., Kraft, P., and Freedman, M. L. (2014b). Expression QTL based analyses reveal candidate causal genes and loci across five tumor types. *Human Molecular Genetics*, page ddu228.

[170] Li, Y., Li-Byarlay, H., Burns, P., Borodovsky, M., Robinson, G. E., and Ma, J. (2013). TrueSight: a new algorithm for splice junction detection using RNA-seq. *Nucleic Acids Research*, **41**(4), e51.

[171] Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, **27**(21), 2380–2396.

[172] Ligtenberg, M. J., Gennissen, A. M., Vos, H. L., and Hilkens, J. (1991). A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA. *Nucleic Acids Research*, **19**(2), 297–301.

[173] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, **8**(10), 833–835.

[174] Lippert, C., Casale, F., Rakitsch, B., and Stegle, O. (2014). LIMIX : genetic analysis of multiple traits. *bioRxiv*, pages 0–26.

[175] Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(38), 16465–16470.

[176] Liu, H. X., Cartegni, L., Zhang, M. Q., and Krainer, A. R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature Genetics*, **27**(1), 55–58.

[177] Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., and Qi, X. (2013). An autoregulatory loop controlling Arabidopsis HsfA2 expression: role of heat shock-induced alternative splicing. *Plant Physiology*, **162**(1), 512–521.

[178] Llorian, M., Schwartz, S., Clark, T. A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A. C., de la Grange, P., Ast, G., and Smith, C. W. J. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Structural & Molecular Biology*, **17**(9), 1114–1123.

[179] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**(13), 1675–1680.

[180] Long, Q., Rabanal, F. a., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. a. *et al.* (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, **45**(8), 884–890.

[181] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., and Ramsey, K. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, **45**(6), 580–585.

[182] Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigo, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, **579**(9), 1900–1903.

[183] Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, **26**(8), 345–352.

[184] Manber, U. and Myers, G. (1993). Suffix arrays: a new method for on-line string searches. *SIAM Journal on Computing*, **22**(5), 935–948.

[185] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

[186] Maquat, L. E. (2004a). Nonsense-Mediated mRNA Decay: A Comparative Analysis of Different Species. *Current Genomics*, **5**(3), 175–190.

[187] Maquat, L. E. (2004b). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Reviews Molecular Cell Biology*, **5**(2), 89–99.

[188] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, **9**(12), 1185–1188.

[189] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

[190] Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, **470**(7333), 198–203.

[191] Marguerat, S. and Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences : CMLS*, **67**(4), 569–579.

[192] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.

[193] Marioni, J., Mason, C., and Mane, S. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**(9), 1509–1517.

[194] Marx, V. (2013). Drilling into big cancer-genome data. *Nature Methods*, **10**(4), 293–297.

[195] Matsukura, S., Mizoi, J., Yoshida, T., Todaka, D., Ito, Y., Maruyama, K., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2010). Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. *Molecular Genetics and Genomics*, **283**(2), 185–196.

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

[197] McCreight, E. M. (1976). A Space-Economical Suffix Tree Construction Algorithm. *Journal of the Association of Computing Machinery*, **23**(2), 262–272.

[198] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

[199] Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F., and Dietz, H. C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nature Genetics*, **36**(10), 1073–1078.

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

[201] Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S.-P., Enard, W., Hellmann, I., Lindblad-Toh, K., Altheide, T. K., Archidiacono, N., Bork, P., Butler, J., Chang, J. L., Cheng, Z. *et al.* (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**(7055), 69–87.

[202] Mitrovich, Q. M. and Anderson, P. (2005). mRNA surveillance of expressed pseudogenes in C. elegans. *Current Biology*, **15**(10), 963–967.

[203] Miura, S., Martins, A., Zhang, K., Graveley, B., and Zipursky, S. (2013). Probabilistic Splicing of Dscam1 Establishes Identity at the Level of Single Neurons. *Cell*, **155**(5), 1166–1177.

[204] Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, **30**(1), 13–19.

[205] Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, **29**(13), 2850–2859.

[206] Mortazavi, A., Williams, B., and McCue, K. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**(7), 621–628.

[207] Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, **51**(1), 263–273.

[208] Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., and Others (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407), 330–337.

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

[210] Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, **39**(13), e90.

[211] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, **472**(7341), 90–94.

[212] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

[213] Nguyen, N., Hickey, G., Zerbino, D. R., Raney, B., Earl, D., Armstrong, J., Haussler, D., and Paten, B. (2014). Building a Pangenome Reference for a Population. *Research in Computational Molecular Biology*, **8394**, 207–221.

[214] Nicolae, M., Mangul, S., Mandoiu, I. I., and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology : AMB*, **6**(1), 9.

[215] Nuwaysir, E. F., Huang, W., Albert, T. J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J. P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L. *et al.* (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Research*, **12**(11), 1749–1755.

[216] Ong, S.-E. and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Biotechnology*, **1**(5), 252–262.

[217] Ono, Y., Asai, K., and Hamada, M. (2013). PBSIM: PacBio reads simulator–toward accurate genome assembly. *Bioinformatics*, **29**(1), 119–121.

[218] Orengo, J. P. and Cooper, T. a. (2007). Alternative splicing in disease. *Advances in Experimental Medicine and Biology*, **623**, 212–223.

[219] Osborne, E. J., Kahles, A., Remigereau, M., Drewe, P., Vilhjalmsson, B., Zhang, P., Parrott, D. L., Greenhalgh, R., Steffen, J., Jean, G., Mott, R., Rätsch, G., Nordborg, M., Stegle, O., and Clark, R. M. (2014). Genetic and environmental influences on gene expression in Arabidopsis. *In preparation.*

[220] Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, **11**(12), 220.

[221] Ossowski, S., Schwab, R., and Weigel, D. (2008). Gene silencing in plants using artificial microRNAs and other small RNAs. *The Plant Journal for Cell and Molecular Biology*, **53**(4), 674–690.

[222] Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**(2), 87–98.

[223] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, **40**(12), 1413–1415.

[224] Pandey, V., Nutter, R. C., and Prediger, E. (2008). Applied Biosystems SOLiD System: Ligation-Based Sequencing. In *Next Generation Genome Sequencing: Towards Personalized Medicine*, pages 29–42. Wiley-VCH Verlag GmbH & Co. KGaA.

[225] Pandya-Jones, A. (2011). Pre-mRNA splicing during transcription in the mammalian system. *Wiley Interdisciplinary Reviews: RNA*, **2**(5), 700–717.

[226] Passini, M. a., Bu, J., Richards, A. M., Kinnecom, C., Sardi, S. P., Stanek, L. M., Hua, Y., Rigo, F., Matson, J., Hung, G., Kaye, E. M., Shihabuddin, L. S., Krainer, A. R., Bennett, C. F., and Cheng, S. H. (2011). Antisense oligonucleotides delivered to the mouse CNS ameliorate symptoms of severe spinal muscular atrophy. *Science Translational Medicine*, **3**(72), 72ra18.

[227] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(8), 2444–2448.

[228] Pease, J. and Sooknanan, R. (2012). A rapid, directional RNA-seq library preparation workflow for Illumina sequencing. *Nature Methods*, **9**(3), i–ii.

[229] Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, **6**(11), S22–32.

[230] Pevzner, P. and Tesler, G. (2003). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, **13**(1), 37–45.

[231] Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A. *et al.* (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**(7278), 191–196.

[232] Posé, D., Verhage, L., Ott, F., Yant, L., Mathieu, J., Angenent, G. C., Immink, R. G. H., and Schmid, M. (2013). Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature*, **503**(7476), 414–417.

[233] Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H. *et al.* (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, **42**(D1).

[234] Purnell, R. F., Mehta, K. K., and Schmidt, J. J. (2008). Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Letters*, **8**(9), 3029–3034.

[235] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, **5**(12), 1005–1010.

[236] Quail, M. a., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**(1), 341.

[237] Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Beà, S., Pinyol, M., Martínez-Trillos, A., López-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M. *et al.* (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics*, **44**(1), 47–52.

[238] Quinlan, A. R., Stewart, D. A., Strömberg, M. P., and Marth, G. T. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**(2), 179–181.

[239] Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, **14**(6), 405.

[240] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S. *et al.* (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, **7**(11), 909–912.

[241] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, **29**(1), 24–26.

[242] Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

[243] Rogers, M. F., Thomas, J., Reddy, A. S., and Ben-Hur, A. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biology*, **13**(1), R4.

[244] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**(1), 84–89.

[245] Rossi, D., Bruscaggin, A., Spina, V., Rasi, S., Khiabanian, H., Messina, M., Fangazio, M., Vaisitti, T., Monti, S., Chiaretti, S., Guarini, A., Del Giudice, I., Cerri, M., Cresta, S., Deambrogi, C. *et al.* (2011). Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood*, **118**(26), 6904–8.

[246] Rühl, C., Stauffer, E., Kahles, A., Wagner, G., Drechsel, G., Rätsch, G., and Wachter., A. (2012). Polypyrimidine Tract Binding Protein Homologs from Arabidopsis Are Key Regulators of Alternative Splicing with Implications in Fundamental Developmental Processes. *The Plant Cell*, **24**(11), 4360–4375.

[247] Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z. *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**(6822), 928–933.

[248] Sakharkar, M. K., Chow, V. T. K., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biology*, **4**(4), 387–393.

[249] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467.

[250] Sawicka, K., Bushell, M., Spriggs, K. A., and Willis, A. E. (2008). Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochemical Society Transactions*, **36**(Pt 4), 641–647.

[251] Sazani, P. and Kole, R. (2003). Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *The Journal of Clinical Investigation*, **112**(4), 481–486.

[252] Schad, E., Tompa, P., and Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, **12**(12), R120.

[253] Schadt, E. E., Li, C., Su, C., and Wong, W. H. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, **80**(2), 192–202.

[254] Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, **10**(9), R98.

[255] Schneider, G. F. and Dekker, C. (2012). DNA sequencing with nanopores. *Nature Biotechnology*, **30**(4), 326–328.

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

[257] Schwab, R. and Ossowski, S. (2006). Highly specific gene silencing by artificial microRNAs in Arabidopsis. *The Plant Cell*, **18**(May), 1121–1133.

[258] Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Ong, C. S., Philips, P., De Bona, F., Hartmann, L., Bohlen, A., Krüger, N., Sonnenburg, S., and Rätsch, G. (2009). mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, **19**(11), 2133–2143.

[259] Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J., and Church, G. M. (2000). RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature Biotechnology*, **18**(12), 1262–1268.

[260] Seo, P. J., Park, M.-J., and Park, C.-M. (2013). Alternative splicing of transcription factors in plant responses to low temperature stress: mechanisms and functions. *Planta*, **237**(6), 1415–1424.

[261] Shapiro, J. a. and von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biological Reviews of the Cambridge Philosophical Society*, **80**(2), 227–250.

[262] Shen, S., Warzecha, C. C., Carstens, R. P., and Xing, Y. (2010). MADS+: discovery of differential splicing events from Affymetrix exon junction array data. *Bioinformatics*, **26**(2), 268–269.

[263] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, **26**(10), 1135–1145.

[264] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741), 1728–1732.

[265] Sirén, J., Valimaki, N., and Makinen, V. (2014). Indexing Graphs for Path Queries withApplications in Genome Research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **11**(2), 375–388.

[266] Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

[267] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

[268] Srebrow, A. and Kornblihtt, A. R. (2006). The connection between splicing and cancer. *Journal of Cell Science*, **119**(Pt 13), 2635–2641.

[269] Sreedharan, V. T., Schultheiss, S. J., Jean, G., Kahles, A., Bohnert, R., Drewe, P., Mudrakarta, P., Görnitz, N., Zeller, G., and Rätsch, G. (2014). Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis. *Bioinformatics*, **30**(9), 1300–1301.

[270] Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, **32**(Web Server issue), W309–312.

[271] Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, **34**(Web Server issue), W435–439.

[272] Stauffer, E., Westermann, A., Wagner, G., and Wachter, A. (2010). Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant Journal*, **64**(2), 243–255.

[273] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S. E., Behr, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, **10**(12), 1177–1184.

[274] Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology*, **11**(5), 207.

[275] Stewart, F. J., Ottesen, E. A., and DeLong, E. F. (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *The ISME Journal*, **4**(7), 896–907.

[276] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., and Ebert, B. L. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550.

[277] 't Hoen, P. a. C., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., Laros, J. F. J., Buermans, H. P. J., Karlberg, O., Brännvall, M., van Ommen, G.-J. B., Estivill, X., Guigó, R., Syvänen, A.-C., Gut, I. G. *et al.* (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology*, (September).

[278] Tanko, Q., Franklin, B., Lynch, H., and Knezetic, J. (2002). A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. *Mutation Research*, **503**(1-2), 37–42.

[279] Tariq, M. a., Kim, H. J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, **39**(18), e120.

[280] Tautz, D. (2000). Evolution of transcriptional regulation. *Current Opinion in Genetics & Development*, **10**(5), 575–579.

[281] Tazi, J., Bakkour, N., and Stamm, S. (2009). Alternative splicing and disease. *Biochimica et Biophysica Acta*, **1792**(1), 14–26.

[282] Teer, J. K. and Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*, **19**(R2), R145–151.

[283] Thanaraj, T. a., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., and Muilu, J. (2004). ASD: the Alternative Splicing Database. *Nucleic Acids Research*, **32**(Database issue), D64–69.

[284] Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., Kieffer, A., Lau, J., Lee, N., Miller, A., Raymant, G. *et al.* (2013). The million mutation project: a new approach to genetics in Caenorhabditis elegans. *Genome Research*, **23**(10), 1749–1762.

[285] Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics*, **17**(R2), R143–150.

[286] Timp, W., Comer, J., and Aksimentiev, A. (2012). DNA base-calling from a nanopore using a Viterbi algorithm. *Biophysical journal*, **102**(10), L37–39.

[287] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–1111.

[288] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511–515.

[289] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**(1), 46–53.

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

[291] Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(22), 14031–14036.

[292] Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., and Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nature Genetics*, **37**(7), 727–732.

[293] Uimari, P., Kontkanen, O., Visscher, P. M., Pirskanen, M., Fuentes, R., and Salonen, J. T. (2005). Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland founder population. *Twin Research and Human Genetics: the Official Journal of the International Society for Twin Studies*, **8**(3), 185–197.

[294] Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica*, **14**(3), 249–260.

[295] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **Chapter 11**(43), UNIT 11.10.

[296] van Hoof, A. (2005). NMD in Plants. In *Nonsense-Mediated mRNA Decay*, pages 167–172. Eurekah.

[297] Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, **270**(5235), 484–487.

[298] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R. *et al.* (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

[299] Vilhjálmsson, B. J. and Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews. Genetics*, **14**(1), 1–2.

[300] Wachter, A., Rühl, C., and Stauffer, E. (2012). The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. *Frontiers in Plant Science*, **3**.

[301] Wallace, E. V. B., Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., Donohoe, T. J., and Bayley, H. (2010). Identification of epigenetic DNA modifications with a protein nanopore. *Chemical Communications*, **46**(43), 8195–8197.

[302] Wandelt, S. and Leser, U. (2012). Adaptive efficient compression of genomes. *Algorithms for Molecular Biology*, **7**(1), 30.

[303] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221), 470–476.

[304] Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. a., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, **38**(18), e178.

[305] Wang, L. and Lawrence, M. (2011). SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, **365**(26), 2497–506.

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

[307] Wang, Z., Jeon, H. Y., Rigo, F., Bennett, C. F., and Krainer, A. R. (2012). Manipulation of PK-M mutually exclusive alternative splicing by antisense oligonucleotides. *Open Biology*, **2**(10), 120133.

[308] Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–562.

[309] Weiner, P. (1973). Linear Pattern Matching Algorithms. *Switching and Automata Theory, 1973. SWAT'08.*, pages 1–11.

[310] Weinstein, J. N., Collisson, E. a., Mills, G. B., Shaw, K. R. M., Ozenberger, B. a., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, **45**(10), 1113–1120.

[311] Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J., Kristiansen, K., Krogh, A., Wang, J., and Porse, B. T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biology*, **13**(5), R35.

[312] Wells, D. B., Belkin, M., Comer, J., and Aksimentiev, A. (2012). Assessing graphene nanopores for sequencing DNA. *Nano Letters*, **12**(8), 4117–4123.

[313] Wetterstrand, K. A. (2014). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

[314] Whiteley, A. S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., Allcock, R., and O'Donnell, A. (2012). Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *Journal of Microbiological Methods*, **91**(1), 80–88.

[315] Wilkie, S. E., Vaclavik, V., Wu, H., Bujakowska, K., Chakarova, C. F., Bhattacharya, S. S., Warren, M. J., and Hunt, D. M. (2008). Disease mechanism for retinitis pigmentosa (RP11) caused by missense mutations in the splicing factor gene PRPF31. *Molecular Vision*, **14**, 683–690.

[316] Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnology*, **15**(13), 1359–1367.

[317] Wray, G. a., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., and Romano, L. a. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**(9), 1377–419.

[318] Wu, C., Carta, R., and Zhang, L. (2005). Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Research*, **33**(9), e84.

[319] Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**(7), 873–781.

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

[321] Xu, Q. and Lee, C. (2003). Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, **31**(19), 5635–5643.

[322] Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, **30**(17), 3754–3766.

[323] Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H. *et al.* (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**(5), 886–895.

[324] Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y., Zhang, C., Yeo, G., Black, D., Sun, H., and Others (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. *Molecular Cell*, **36**, 996–1006.

[325] Yang, J., Zaitlen, N. a., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, **46**(2), 100–106.

[326] Yoine, M., Ohto, M. A., Onai, K., Mita, S., and Nakamura, K. (2006). The lba1 mutation of UPF1 RNA helicase involved in nonsense-mediated mRNA decay causes pleiotropic phenotypic changes and altered sugar signalling in Arabidopsis. *Plant Journal*, **47**(1), 49–62.

[327] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**(2), 203–208.

[328] Zhang, Z., Ersoz, E., Lai, C., and Todhunter, R. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, **42**(4), 355–360.

[329] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(4), 1193–1198.

[330] Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(4), E455–464.

C Curriculum Vitae

■ Personal Data

Name	André Kahles
Date of Birth	March 21, 1984
Place of Birth	Zschopau, Germany
Nationality	German

■ Education

Since 04/2012	Continued Doctoral studies in the group of Gunnar Rätsch at Memorial Sloan-Kettering Cancer Center, New York
09/2011	Machine Learning Summer School 2011 Carcans-Maubuisson, France
09/2010	International School on Mathematics - 1st Workshop on Optimization, Machine Learning and Bioinformatics Erice, Italy
10/2009 - 03/2012	Doctoral studies in the group of Gunnar Rätsch at the Friedrich Miescher Laboratory of the Max Planck Society *Thesis topic:* *Computational Methods for the Analysis of Transcriptome Sequencing Data* *Advisors:* *Prof. Dr. Daniel Huson, University Tübingen* *Prof. Dr. Gunnar Rätsch, SKI New York*
09/2009	*Diplom* (M.Sc.) in Bioinformatics with grade 1.0 at the Friedrich Schiller University Jena *Thesis topic:* *Maximizing Diversity of Statistically Dependent Markers in Mitochondrial DNA* *Advisors:* *Prof. Dr. Sebastian Böcker, University Jena* *Dr. Lars Arvestad, KTH Stockholm*
11/2005	*Vordiplom* (B.Sc.) in Bioinformatics with grade 1.3

10/2003 - 09/2009	Studies of Bioinformatics at the Friedrich-Schiller-University Jena
06/2002	General qualification for university entrance (Abitur) *Advanced courses: mathematics and physics*
1994 - 2002	Secondary school "'Gymnasium Zschopau"' in Zschopau

■ Research interests

High throughput sequence analysis, metagenomics and -transcripomics, alternative splicing regulation, applied machine learning

■ Fellowships and Awards

05/2005 - 03/2009	Scholarship holder of the Stiftung der Deutschen Wirtschaft (Foundation of the German Economy)
10/2009 - 03/2012	PhD Student Fellowship from the Max Planck Society
07/2010	ISMB Student Council Travel Fellowship
07/2014	ISMB Travel Fellowship

■ Professional Skills

Programming	C/C++, Python, Java, Perl, Matlab, R
Applications	Microsoft Office (incl. macros in VBS), Mathematica, Matlab, LaTeX, CMS (Typo3 and Mambo/Joomla), Gimp, Scribus
Operating Systems	Linux, Mac OS, Unix, Microsoft Windows

■ Work Experience

11/2007 - 04/2008	Internship at the Stockholm Bionformatics Center
05/2007 - 10/2007	Student assistant at the Biosystems Analysis Group at the University of Jena with a monthly amount of 40 hours work
08/2002 - 07/2003	Alternative civilian service at the special school for mentally handicapped children "'Johann-Ehrenfried-Wagner"' in Marienberg

■ Extracurricular and Social Activities

10/2010 - 12/2011	PhD student Representative of the Friedrich Miescher Laboratory of the Max Planck Society
09/2006 - 03/2013	Founder member of the non-profit association "Jugend-Unternimmt e.V." (that organizes an economic student competition)
10/2003 - 03/2009	Member of the Bioinformatics student representatives
10/2003 - 10/2006	Deputy member of the council of the Friedrich-Schiller-University Jena
1994 - 2002	Member of the pupil's representatives at school, federal, and national level

■ Talks

PALMapper: Fast, Accurate and Variation-Aware RNA-Seq Alignments.
A. Kahles, G. Jean, D. Kuo, and G. Rätsch
HitSeq 2013, 07/19/2013 - 07/20/2013, Berlin, Germany

Nonsense-Mediated Decay is a Major Transcriptome Regulator.
A. Kahles, G. Drechsel, G. Rätsch and A. Wachter
Genome Informatics, 09/06/2012 - 09/09/2012, Cambridge, UK

■ Publications

1. **Kahles, A** and Ong, C S and Zeller, G and Rätsch, G. "SplAdder: Comprehensiver Alternative Splicing Analysis on RNA-Seq Data.", *In preparation for Bioinformatics.*, 2014.
   We developed a software suite to identify, quantify, and differentially analyzer alternative splicing events based on RNA-Seq data.

2. Jean, G and **Kahles, A** and Sonnenburg, S and De Bona, F and Schneeberger, K and Hagmann, J and Weigel, D and Rätsch., G. "PALMapper: Fast and accurate alignment of RNA-seq reads.", *In preparation for Bioinformatics*, 2014.
   We developed a novel read alignment algorithm that automatically adapts to errors present in the read data and accurately aligns RNA-seq reads to the genome.

3. **Kahles, A** and Behr, J and Rätsch, G. "MMR: Resolving ambiguous Alignment Locations.", *In preparation for Bioinformatics.*, 2014.
   We present a versatile toolkit for alignment optimization and read mapping disambiguation.

4. Lehmann*, K and **Kahles*, A** and Kandoth, C and Lee, W and Network, Cancer Genome Atlas Research and Schultz, N and Stegle, O and Rätsch, G. "Extensive trans and cis-QTLs revealed by large-scale cancer genom analysis.", *In preparation*

for Nature Genetics., 2014.
We conducted a large-scale analysis on RNA-Seq data from 4,000 cancer patients to reveal numerous cis- and trans sQTL

5. Osborne*, E J and **Kahles*, A** and Remigereau, M and Drewe, P and Vilhjalmsson, B and Zhang, P and Parrott, D L and Greenhalgh, R and Steffen, J and Jean, G and Mott, R and Rätsch, G and Nordborg, M and Stegle, O and Clark, R M. "Genetic and environmental influences on gene expression in Arabidopsis", *In preparation.*, 2014.

6. Dubin, M J and Zhang, P and Meng, D and Remigereau, M and Osborne, E J and Casale, F P and Drewe, P and **Kahles, A** and Voronin, V and Song, Q and Long, Q and Rätsch, G and Stegle, O and Clark, R M and Nordborg, M. "DNA methylation variation in Arabidopsis has a genetic basis and appears to be involved in local adaptation", *In preparation.*, 2014.
In this work we investigate DNA methylation variation in Swedish Arabidopsis thaliana accessions.

7. Sreedharan, V T and Schultheiss, S J and Jean, G and **Kahles, A** and Bohnert, R and Drewe, P and Mudrakarta, P and Görnitz, N and Zeller, G and Rätsch, G. "Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis.", *Bioinformatics*, **30**(9), 1300–1301, 2014.

8. Behr, J and **Kahles, A** and Zhong, Y and Sreedharan, V T and Drewe, P and Rätsch, G. "MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples.", *Bioinformatics*, **29**(20), 2529–2538, 2013.

9. Drewe, P and Stegle, O and Hartmann, L and **Kahles, A** and Bohnert, R and Wachter, A and Borgwardt, K and Rätsch, G. "Accurate detection of differential RNA processing.", *Nucleic Acids Research*, **41**(10), 5189–5198, 2013.

10. Engström, P G and Steijger, T and Sipos, B and Grant, G R and A Kahles and Rätsch, G and Goldman, N and Hubbard, T J and Harrow, J and Guigó, R and Others. "Systematic evaluation of spliced alignment programs for RNA-seq data", *Nature Methods*, **10**(12), 1185–1191, 2013.

11. Drechsel*, G and **Kahles*, A** and Kesarwani, A and Stauffer, E and Behr, J and Drewe, P and Rätsch, G and Wachter, A. "Nonsense-Mediated Decay of Alternative Precursor mRNA Splicing Variants Is a Major Determinant of the Arabidopsis Steady State Transcriptome", *The Plant Cell*, **25**(10), 3726–3742, 2013.

12. **Kahles, A** and Sarqume, F and Savolainen, P and Arvestad, L. "Excap: Maximization of Haplotypic Diversity of Linked Markers", *PloS One*, **8**(11), e79012, 2013.

13. Zeller, G and Goernitz, N and **Kahles, A** and Behr, J and Mudrakarta, P and Sonnenburg, S and Rätsch, G. "mTim: Rapid and accurate transcript reconstruction from RNA-Seq data", *arXiv preprint arXiv:1309.5211*, 2013.

14. Smith, L M and Hartmann, L and Drewe, P and Bohnert, R and **Kahles, A** and Lanz, C and Rätsch, G. "Multiple insert size paired-end sequencing for deconvolution of complex transcriptomes", *RNA Biology*, **9**(5), 596–609, 2012.

15. Rühl, C and Stauffer, E and **Kahles, A** and Wagner, G and Drechsel, G and Rätsch, G and Wachter., A. "Polypyrimidine Tract Binding Protein Homologs from Arabidopsis Are Key Regulators of Alternative Splicing with Implications in Fundamental Developmental Processes", *The Plant Cell*, **24**(11), 4360–4375, 2012.

16. Gan, X and Stegle, O and Behr, J and Steffen, J G and Drewe, P and Hildebrand, K L and Lyngsoe, R and Schultheiss, S J and Osborne, E J and Sreedharan, V T and **Kahles, A** and Bohnert, R and Jean, G and Derwent, P and Kersey, P and Belfield, E J and Harberd, N P and Kemen, E and Toomajian, C and Kover, P X and Clark, R M and Rätsch, G and Mott, R. "Multiple reference genomes and transcriptomes for Arabidopsis thaliana", *Nature*, **108**(25), 10249–10254, 2011.

17. Schultheiss, S J and Jean, G and Behr, J and Drewe, P and Görnitz, N and **Kahles, A** and Mudrakarta, P and Sreedharan, V T and Zeller, G and Rätsch., G. "oqtans: a Galaxy-integrated workflow for quantitative transcriptome analysis from NGS Data", *BMC Bioinformatics*, **12**(Suppl 11), A7, 2011.

18. Görnitz, N and Widmer, C and Zeller, G and **Kahles, A** and Sonnenburg, S and Rätsch., G. "Hierarchical Multitask Structured Output Learning for Large-scale Sequence Segmentation", *Advances in Neural Information Processing Systems (NIPS'11)*, 2690–2698, 2011.

19. Jean, G and **Kahles, A** and Sreedharan, V T and De Bona, F and Rätsch, G. "RNA-Seq read alignments with PALMapper.", *Current Protocols in Bioinformatics*, **Chapter 11**(December), Unit 11.6, 2010.

■ Posters

SplAdder: Integrated Quantification, Visualization and Differential Analysis of Alternative Splicing.
A. Kahles, C. S. Ong, G. Zeller, and G. Rätsch
ISMB 2014, 07/11/2014 - 07/15/2014, Boston, USA

PALMapper: Fast, Accurate and Variation-aware RNA-Seq Alignments.
A. Kahles, G, Jean, D. C. Kuo and G. Rätsch
Genome Informatics 2013, 10/30/2013 - 11/02/2013, Cold Spring Harbor, USA

Evaluation and Tuning of RNA-Seq Read Alignments
A. Kahles, R. Bohnert, P. Ribeca, J. Behr and G. Rätsch
Genome Informatics 2011, 11/02/2011 - 11/05/2011, Cold Spring Harbor, USA

RGASP Evaluation of RNA-Seq Read Alignment Algorithms
A. Kahles, R. Bohnert, P. Ribeca, J. Behr and G. Rätsch
Satellite Workshop RECOMB-Seq, 03/26/2011 - 03/27/2011, Vancouver, CA
4th Berlin Summer Meeting, 06/23/2011 - 06/25/2011, Berlin, Germany

An Accuracy Evaluation of Read Alignment Algorithms
A. Kahles, J.Behr, R. Bohnert and G. Rätsch
Genome Informatics 2010, 09/15/2010 - 09/19/2010, Hinxton, UK
ISMB 2010, 07/11/2010 - 07/13/2010, Boston, USA

List of Figures

List of Tables

List of Abbreviations

amiRNA	artificial microRNA
AS	alternative splicing
cDNA	complementary DNA
CDS	coding sequence
CHX	cycloheximide
CVAS	common variant association study
DNA	deoxyribonucleic acid
ENCODE	Encyclopedia of DNA Elements
ESE	exonic splice enhancer
ESS	exonic splice silencer
EST	expressed sequence tag
FDR	false discovery rate
GEO	gene expression omnibus
GO	gene ontology
GWAS	genome-wide association study
HTS	high-throughput sequencing
ICGC	International Cancer Genome Consortium
ISE	intronic splice enhancer
ISS	intronic splice silencer
LMM	linear mixed model
MAF	minor allele frequency
MMP	maximum mappable prefix
MMR	multi-mapper resolution
mRNA	messenger RNA
NMD	nonsense-mediated mRNA decay
PCA	principle components analysis
PCR	polymerase chain reaction
PSI	percent spliced in
PTC	premature termination codon
QTL	quantitative trait loci
RGASP	RNA-Seq Genome Annotation Assessment Project
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RPKM	reads per kilobase per million
RVAS	rare variant association study
SAFT	simple alignment filtering tool

TAIR	The Arabidopsis Information Resource
TCGA	The Cancer Genome Atlas
UCSC	University of California Santa Cruz
UTR	untranslated region
VCF	variant call format
WT	wild-type

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Tübingen, den 29. Juli 2014

André Kahles