Data Structures and Algorithms for Analysis of Alternative Splicing with RNA-Seq Data

Marcel H. Schulz

June 2010

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Jens Stoye

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Jens Stoye

Tag der Promotion: 26 August 2010

Preface

The first part of the thesis in Chapter 3 was published in the journal *Science*, as part of a collaborative study that addressed the first application of RNA-Seq experiments to two human cell lines in 2008 [147]. The methods for prediction and quantification of alternative isoforms and their application presented in Chapter 3 and 4 appeared in the journal *Nucleic Acids Research* in 2010 [129]. The last part about the de novo transcriptome assembly method Oases has not been published yet, but a manuscript is in preparation. The successful application of Oases to human, fly, and worm RNA-Seq data will be published in the report about the RGASP competition this year.

My contributions to these papers was the design of statistical methods and the analysis of alternative splicing with junctions reads for the *Science* paper. I designed, implemented, and analyzed the CASI and DASI method. I was involved in the conception and analysis of the POEM method, analysis of the exon array data, as well as primer design for the PCR experiments and their evaluation. I implemented a preliminary R version of the initial steps of the transcriptome assembler, addressing loci and trivial transcript reconstruction. I developed the theory in Section 5.1 and made the analysis with Oases in Sections 5.3 and 5.4. I was involved in the algorithm design in Section 5.3 and error analysis of the Oases software. I did the transcriptome assembly and parts of the downstream analysis for the RGASP submissions.

There are a number of other contributions that are unfortunately not in the thesis. I have implemented linear time algorithms for the construction of variable order Markov chains and the first algorithm for the score distribution computation for ontological similarity searches, presented at the WABI conference in 2008 and 2009 [142, 141]. Also I designed and implemented a linear time truncated suffix tree algorithm [140]. Further, I was involved as a co-author in projects about clinical diagnostics with

ontologies [81], basepair-precise breakpoint detection of human structural variations in resequencing data [27, 170], the influence of highly conserved sequence elements on gene expression [132, 54], and algorithms for frequency pattern mining [164].

Acknowledgements I am grateful to my supervisor Martin Vingron for his ideas, support, initiating of collaborations, and especially for his suggestion to start working with de Bruijn graphs. It is a wonderful atmosphere that he created in his group that I have enjoyed throughout the years. I also want to thank him and the International Max Planck Research School for Computational Biology and Scientific Computing for funding my Phd time and my stay in Cambridge. I also like to thank Hugues Richard, who supervised me for the project about statistical methods with RNA-Seq data (Chapter 3 and 4). My knowledge about statistics, *R*, and many other important subjects in life have grown considerably due to his influence. I acknowledge his writing of the functions for the POEM algorithm, help with its design and analysis, primer design, exon array analysis, and application of POEM on the RGASP data.

I want to thank Daniel R. Zerbino for sharing his expertise about assembly algorithms, which has been an important contribution to the project. In particular, I acknowledge his conception for the treatment of cycles, adaptations for the transitive reduction algorithm of Myers, and running ABySS. I am very grateful to Daniel that he accepted to implement the ideas described in Section 5.2 in the Oases software, which made it possible to have results in time for the RGASP competition. I further want to thank Ewan Birney for financing my stay in Cambridge, hosting me at the EBI, and inspiring and motivating discussions about transcriptome assembly.

I want to thank Marc Sultan for conducting the RNA-Seq, PCR, and exon array experiments, help with primer design, and many interesting discussions about wet lab biology. Additionally, I thank Marie-Laure Yaspo for financing the experiments, as well as, Hans Lehrach, Asja Nürnberger, Sabine Schrinner, Daniela Balzereit, and Emilie Dagand, who conducted or supervised parts of RNA-Seq, PCR, and exon array experiments for Chapter 3 and 4. Further, I would like to thank Stefan Haas for several discussions about alternative splicing and help with EST analysis, paper writing, and primer design. I further thank David Weese for help with setting up a pipeline for the RGASP competition and other interesting projects with did together. Thanks to Axel Rasche for sharing his knowledge about exon array analysis and preprocessing the human exon array data. Thanks to Andreas Klingenhoff, Alon Magen, Dmitri

Parkhomchuk, Matthias Scherf, and Martin Seifert for preprocessing and preparation of data for the Science paper. Thanks to Cole Trapnell, Ali Mortazavi, and Dian Trout for sharing the mouse C2C12 data.

I want to thank Knut Reinert and Peter N. Robinson for constant support and guidance. Thanks to Jens Stoye for reviewing my thesis, Roland Krause for joining my "Verteidigungskommission" on very short notice. I would like to thank Hannes Luz for support throughout my Phd and especially in the last minutes as well as Kirsten Kelleher for support. Also thanks to Anne-Kathrin Emde, Stefan Haas, Marta Luksza, Alena Mysickova, Hugues Richard, Christian Rödelsperger, Marc Sultan, Ewa Szczurek, David Weese, and Daniel R. Zerbino for great comments from proofreading. In addition, I would like to thank Sebastian Bauer, Sebastian Köhler, Jonathan Göke, Tobias Rausch, Christian Rödelsperger, Stefan Roepcke, and Silke Stahlberg and all so far unmentioned members of the Vingron department for help, parties, and other interesting projects I was involved in. I want to thank Markus Bauer, Florian Markowetz, Ole Schulz-Trieglaff as well as the EBI Pre- and Postdocs for sharing many private parties and taking care of me during my stay in Cambridge.

Finally, I want to express my biggest gratitude to my parents and my girlfriend Susi. Their long lasting support and love is the foundation of all my achievements. I love you.

Contents

List of Figures

Chapter 1

Introduction

1.1 DNA, Gene Expression, and Alternative Splicing

DNA

The *deoxyribonucleic acid* (DNA) molecule is the carrier of the genetic information in our cells. Each DNA molecule consists of the four *nucleotides Adenine* (A), *Cytosine* (C), *Guanine* (G), and *Thymine* (T). It is organized as a *double-helix* and the two *strands* of the DNA molecule are complementary to each other. The base-pairing is fixed, where A is complementary to T and G is complementary to C. The formation of the double-helix from two single strand molecules is called *hybridization*, which is an important step of many of the experimental procedures explained later. The complete genome is composed of a set of different DNA molecules which are called *chromosomes*. Eukaryotes, i.e., organisms that have cellular *nuclei*, store the chromosomes in the nuclei of their cells. For example humans have 23 chromosomes which amount to a total of 3 billion nucleotides and each chromosome appears in two copies in each cell .

An important functional unit on a chromosome is the *gene*. The central dogma in molecular biology is that gene regions are *transcribed* to *ribonucleic acid* (RNA), which is then *translated* to proteins. In RNA molecules the nucleotide Thymine is replaced by Uracil (U) and is often single stranded compared to DNA. The half-life of RNA molecules is shorter than that of DNA molecules.

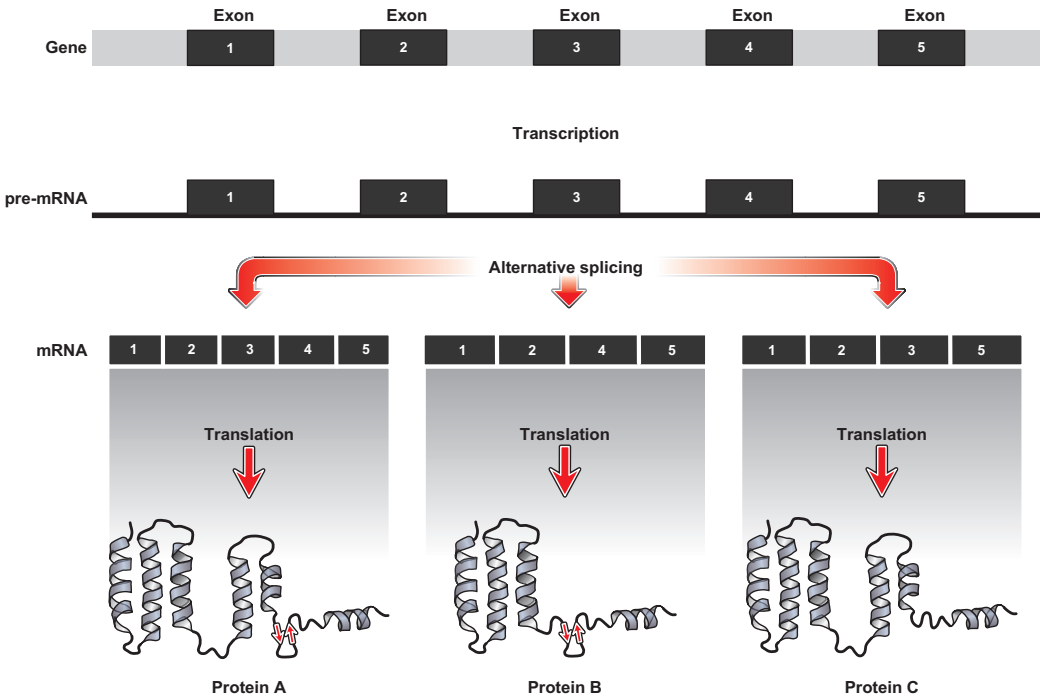

Figure 1.1: All exons of a gene are transcribed into the pre-mRNA. The splicing machinery removes intron sequences and may additionally remove some of the exons through the process of alternative splicing. These mRNAs are then translated into different proteins. Inspired by an illustration in [56].

Gene Expression

A gene is composed of *exons* and *introns*. Exons are the essential parts of the *messenger* RNA (mRNA) that is the template sequence for the protein, whereas introns mostly have a regulatory rule. The complete gene region is first transcribed into a precursor mRNA (pre-mRNA) molecule that consists of exons and introns, see Fig. 1.1. The process of transcription is controlled by proteins, called *transcription factors*, that bind in or near the *promoter* region that resides directly upstream of the gene. A process called *splicing* is inititated by RNA binding proteins called *splicing factors* and the intron sequences of the pre-mRNA are removed. The recognition of intron-exon boundaries is facilitated through the detection of short sequences called *splice sites*. *Polyadenylation* is the addition of a poly-A tail, consisting of a series of Adenine nucleotides, to an RNA molecule. The process of polyadenylation is initiated by binding of proteins to a *polyadenylation site* in an untranslated region of an exon in the pre-mRNA. After splicing and polyadenylation of the pre-mRNA, the final mRNA is obtained. The poly-A tail is important for the transport of the mRNA to

the cell *cytoplasm* and controls further the half-life of the mRNA. The processes of transcription, splicing, and polyadenylation are most likely coupled [80].

Alternative Exon Events

There are three important mechanisms that change the final exon content of a gene's mRNA; *alternative splicing, alternative promoters*, and *alternative polyadenylation*. Altogether, these events are summarized as *alternative exon events* (AEEs). Different mRNAs from the same gene are called *alternative isoforms*. Exons that are involved in at least one AEE of a gene are called *alternative* exons, all other exons of a gene are called *constitutive*.

Alternative splicing (AS) is the mechanism by which a common pre-mRNA produces different mRNA variants, by extending, shortening, skipping, or including exon, or retaining intron sequences. An overview of possible AS events is shown in Fig. 1.2. The combinatorics of such AS events generates a large variability at the post-transcriptional level accounting for an organism's proteome complexity [17, 104], see Fig. 1.1. It has been estimated that 75-92% of all human genes give rise to alternative isoforms [75, 119, 158].

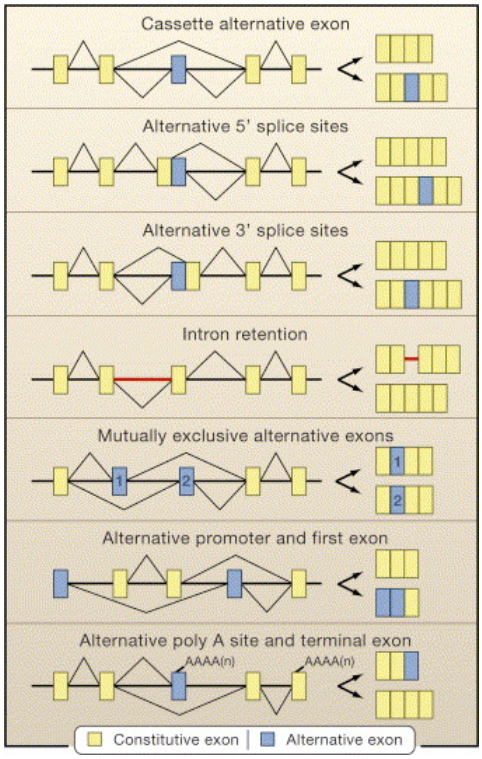

Figure 1.2: Depicted are possible alternative splicing events. Reproduced from [12].

The transcription of a gene can be intitiated from alternative promoters due to regulation of transcription factors, resulting in different pre-mRNAs that perform alternative functions in the cell [40]. Inititation of polyadenylation can similarly be done from different regions in the pre-mRNA through binding of *polyadenylation factors*. These factors recognize the polyadenylation site, a short sequence in untranslated regions of exons. If two

different polyadenylation sites are used in a pre-mRNA, mRNAs with different 3' ends are produced, which is called alternative polyadenylation [174]. There can be a coupling between alternative exon events, for example if transcription is initiated from an alternative promoter there might be additional alternative splicing of some of the exons, see Fig. 1.2. Various gene isoforms generated by AEEs have specific roles in particular cell compartments, tissues, stages of development, etc. In addition, many diseases (e.g. cancer) have been related to alterations in the splicing machinery, highlighting the relevance of AS to therapy [36, 51, 79].

1.2 DNA sequencing

The field of DNA sequencing has a diverse history. In the early 1990s DNA sequencing was conducted dominantly trough application of Sanger sequencing using capillary-based semi-automated sequencers, see Fig. 1.3a. Sanger sequencing was used for sequencing a large fraction of genomes currently used in modern databases, including the human genome [83, 154]. In the past few years a number of sequencing technologies have been developed that are parallelizable and therefore able to create more sequence output compared to convential Sanger sequencing. These are collectively called next-generation sequencing (NGS) approaches. Although these approaches differ in biochemistry they all follow the principle of cyclic-array sequencing, where colonies of immobilized DNA features are sequenced in iterative cycles of enzymatic reactions and imaging-based data detection, see Fig. 1.3b. These technologies have been released as commercial products, e.g., the Solexa Genome Analyzer (marketed by Illumina, San Diego), the SOLiD platform (marketed by Applied Biosystems; Foster City; CA, USA), 454 Genome Sequencers (Roche Applied Science; Basel), and the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA). These technologies create reads of length 25 - 250 bps and with up to 40 million reads per run.

1.3 Methods for Detection of Alternative Splicing

Systematic analysis of alternative isoforms was based on the analysis of expressed sequence tags (ESTs), splicing microarray experiments, or RNA sequencing using

NGS technologies (RNA-Seq). ESTs have been initially used for the detection and prediction of alternative splice forms in different organisms and cell types [17, 55, 88, 169].

Expressed Sequence Tags

mRNA sequences from expressed genes cannot be cloned directly, and are thus reverse transcribed to double-stranded complementary DNA (cDNA). The resultant cDNA is cloned to make cDNA libraries that represent a set of expressed mRNAs of the original cell or tissue. These cDNA clones are sequenced at random from both directions in a single-pass run of the polymerase without validation or sequencing full-length to obtain 5' and 3' expressed sequence tags (ESTs). These ESTs range in length between 100 to 800 bps. The first human gene map was constructed using ESTs [139] and a large collection of ESTs for different species can be found in dbEST [14]. However, EST sequencing showed inherent limitations associated with cloning strategies, non-uniform transcript coverage and low abundance for individual tissues [55, 87].

Splicing Microarrays

The first genome-scale detection methods for the measure of alternative splicing were splicing microarrays. In microarrays small DNA oligonucleotides (called probes) are attached to a solid surface. The probes can be designed to hybridize against complementatry DNA or RNA target samples of, for example, known genes. The strength of the hybridization is quantified by detection of fluorescence labeled targets [136]. Splicing microarrays come in different variants using exon body probes (exon arrays) and/or probes spanning splice junctions (exon junction arrays) [75, 88, 12, 29, 85]. Custom arrays, combining exon body and splice junction probes were designed and used for quantifying transcript expression levels [120]. The standard platform provided by the Affymetrix human exon array allows the monitoring of 10^6 exons derived from 18,000 known genes and approximately 262,000 predicted transcripts [28]. However, several problems inherent to the use of splicing microarrays, such as probe hybridization behaviour, cross hybridization of related probes, and deconvoluting signal-to-noise ratios [87] are difficult to overcome. For instance, for the human

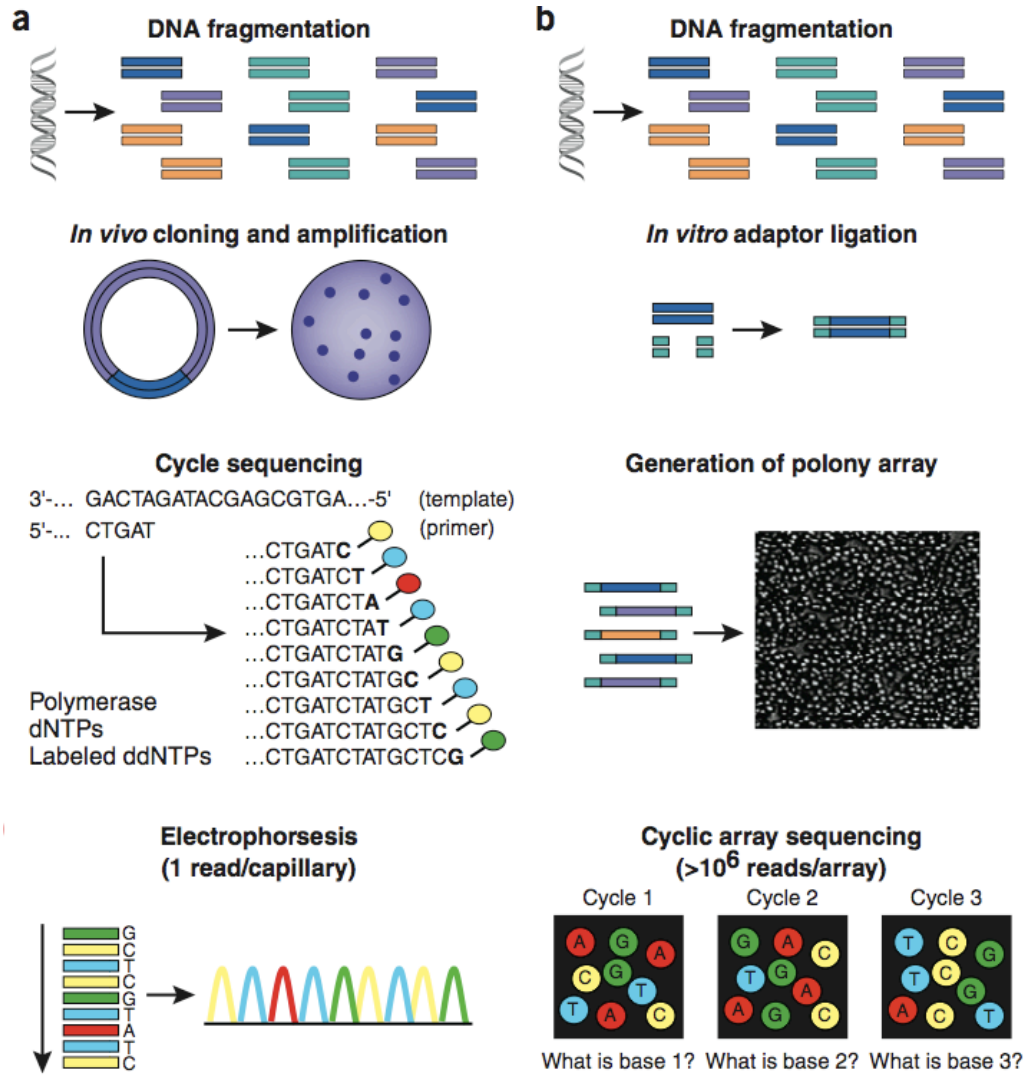

Figure 1.3: Work flow of Sanger sequencing versus next-generation sequencing. (a) With high-throughput shotgun Sanger sequencing, DNA is fragmented and subsequently cloned to a plasmid vector and transformed into E. coli. A single bacterial colony is selected for each sequencing reaction and the DNA is isolated. Each cycle sequencing reaction creates a ladder of dye-labeled products, which are subjected to electrophoretic separation in one run of a sequencing instrument. A detector for fluorescently labeled fragments of discrete sizes in the four-channel emission spectrum facilitates the sequencing trace. (b) In next-generation shotgun sequencing, common adaptors are ligated to fragmented genomic DNA. The DNA is treated to create millions of immobilized PCR colonies, called polonies, each containing copies of a single shotgun library fragment. In cyclic reactions, sequencing and detection of fluorescence labels determines a contigous sequencing read for each polony. Reproduced from [143].

Affymetrix exon arrays, the validation rate ranges from 33% [52] to 86% [28]. Be-

sides, the computational analysis of exon arrays remains a complex task [168, 126].

RNA-Seq

The sequencing of expressed RNAs with NGS approaches is abbreviated as RNA-Seq. In this work it means in particular sequencing of mRNAs. In 2008 a number of papers appeared that applied RNA-Seq to transcriptomes of different organisms [113, 31, 117, 147, 98, 158, 93, 119] and many followed since then. These papers and reviews [32, 13, 161] have established RNA-Seq experiments provide in-depth information on the transcriptional landscape with unprecedented sensitivity and throughput and outperform the previous techniques of EST sequencing and splicing microarrays. RNA-Seq experiments are highly reproducible [107, 147] and have increased sensitivity, therefore more statistical power for the detection of differentially expressed genes, alternative splicing events, and unannotated transcriptional units.

The basic protocol consists of the following steps (i) polyadenylated RNAs in the biological sample are extracted, (ii) these RNAs are converted into more stable cDNA molecules which are randomly sheared, (iii) a size selection on the sheared fragments is done for optimiziation of later steps or paired-end sequencing, (iv) the fragments are amplified and adapters are ligated to the fragments, and finally (v) sequencing of the fragments is carried out using an NGS approach (Figure 1.4). Reads can be obtained from only one end of a fragment (single-end sequencing) or from both ends of a fragment (paired-end sequencing). For paired-end sequencing the approximate fragment size or *insert length* is used for distance estimates between both ends. The reads in the basic protocol loose their orientation and therefore it is not known from which strand of the DNA the mRNA originates.

1.4 Thesis Organization

Sequences and Graphs in Computational Biology Chapter 2 introduces definitions for strings and graphs. A short introduction to sequence alignment is given and the two main approaches for genome assembly are introduced. The Velvet genome assembler is explained in detail. Further, approaches for EST assembly are presented and the splicing graph is introduced.

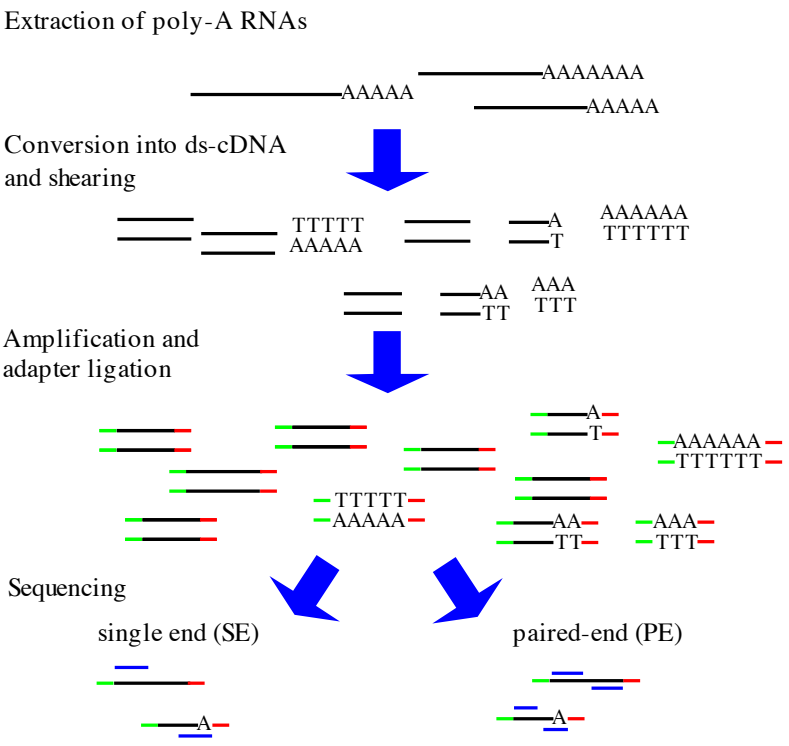

Figure 1.4: Summary of the basic RNA-Seq protocol. As a first step in the protocol only polyadenylated RNAs are extracted. The exact protocol varies mostly in the aspects of shearing the cDNA (e.g. nebulization or sonication), the strategy of amplification (before/after fragmentation), and the type of sequencing (i.e. single end or paired end sequencing). In the basic protocol the orientation of the reads is lost.

Prediction of Alternative Isoforms In Chapter 3 a set of methods is introduced that enable the detection of AEEs within or between conditions using a given gene annotation. All methods are based on a stochastic model of the read distribution along a transcript that is introduced. At first the detection using reads spanning exon-exon junctions is investigated. Secondly, two statistical indices, the Cell type-specific Alternative uSage Index (CASI) for prediction of AEEs within a given condition, e.g. one cell line, and the Differential Alternative uSage Index (DASI) for prediction of AEEs differentiating two conditions are introduced. All methods are applied to a data set from a human embryonic kidney (HEK) and a B cell line. The robustness of the predictions was assessed by bootstrapping. Several thousands of AEEs were predicted and RT-PCR experiments were conducted for validation. In addition, a comparison of splicing prediction by RNA-Seq to predictions made from exon arrays with the same sample is given.

Quantification of Alternative Isoforms Chapter 4 descibes a new method for inferring isoform expression levels from RNA-Seq data. The PrOportion EstiMation (POEM) method enables the relative quantification of known transcript structures within a given condition. Using the Poisson distribution an Expectation-Maximization approach is utilized in the POEM method for maximizing the likelihood of the data and computing the isoform expression levels. The POEM method is applied to RNA-Seq data of the HEK and B cell line and isoform expression levels for sufficiently expressed genes are estimated, after investigating the theoretical power of the method with simulations. Quantitative RT-PCR experiments were used to assess the accuracy of the predictions.

De Novo Assembly of Transcripts considering Alternative Isoforms In Chapter 5 the first method for the *de novo* assembly of an organism's transcriptome from short read RNA-Seq data is introduced. The approach is based on de Bruijn graphs. Similarities to splicing graphs are explored and a theory for *de novo* prediction of AEEs is developed. Based on the graph structure and error correction steps of the Velvet genome assembler, new algorithms are presented that derive transcript clusters, called loci, from short read data and predict full length transcripts for each cluster. A merged assembly approach is devised that improves the results. An application to real data demonstrates the improvement compared to *de novo* genome assemblers that have been utilized so far for RNA-Seq datasets. Further a comparison with a transcriptome assembler that uses the genome is made.

Chapter 2

Sequences and Graphs in Computational Biology

2.1 Definitions

2.1.1 Strings

Let w be a string or sequence over the aphabet Σ. The length of string w is denoted $|w|$ and the size of Σ is denoted $|\Sigma|$. The i^{th} character of a string w is denoted by $w[i]$. If $1 \leq i \leq j \leq |w|$, then $w[i, j]$ denotes the substring beginning at the i^{th} position and ending at the j^{th} position, inclusive. If there exists, i, j such that $v = w[i, j]$, then v is called a *substring* of w. The number of occurrences of a substring v in a string w is denoted as $occ_w(v)$. Let $pos_w(v)$ be the first starting position of substring v in string w. A string of length k is called a k-mer. The k-spectrum(w) is the set of all k-mers that are substrings of w. Analogously, let k-spectrum(v, w) be the set of all k-mers that are substrings of v or w. The reverse complement string of w is denoted $\overleftarrow{w}$. Concatenation of two strings w and v is denoted wv. A string w *overlaps* v if there exists a maximal length non-empty string x which is a prefix of w and a suffix of v.

2.1.2 Graphs

A graph $G = (V, E)$ has nodes V and edges E. Each edge contains a pair of nodes v and w, $v, w \in V$. An edge is *directed* if one endpoint is designated the head and the other the tail. A directed graph denoted as *digraph* has only directed edges. If an edge has no direction it is *undirected*. A directed edge is called *ingoing* at a node if the node is an endpoint for the edge and *outgoing* if the node is the startpoint for the edge. The indegree $indeg(v)$ of node v is the number of ingoing edges and the outdegree $outdeg(v)$ the number of outgoing edges. The degree of node v is denoted $deg(v)$ and is equal to the sum of all ingoing and outgoing edges of v. A *complete* graph is a graph such that every pair of nodes is joined by an edge. The *underlying* graph of a digraph is the graph that results from replacing all directed edges with undirected edges.

A *walk* or *path* from node v_1 to node v_k is a sequence $v_1, e_1, \ldots, e_{k-1}, v_k$, alternating between nodes and edges, such that the endpoints of edge e_i are v_i and v_{i+1}, for $i = 1, \ldots, k-1$. A walk is called *cyclical* if its endpoints v_1 and v_k are the same. A graph is called *acyclic* if it contains no cyclical walk. A graph is said to be *connected* if there exists a walk between every pair of nodes in the underlying graph. A graph that is acyclic and directed is called an acyclic directed graph (DAG).

2.1.3 Sensitivity and Specificity

Sensitivity and *specificity* are measures of the performance of classification tests. Let TP be the number of *true positive* predictions, FP the number of *false positive* predictions, and FN be the number of *false negative* predictions, then

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.1}$$

$$Specificity = \frac{TP}{TP + FP} \, . \tag{2.2}$$

For example, in Chapter 5 the exon sensitivity and exon specificity are used to assess the prediction performance against known transcript annotation. Exon specificity denotes the proportion of correct predictions among all predictions, whereas exon sensitivity denotes the proportion of correctly predicted exons among all annotated exons. A common way to compare the tradeoff between the two measures is the

Receiver Operating Characteristic (ROC) curve, which depicts the prediction performance of a classifier as sensitivity on the x-axis and 1-specificity on the y-axis.

2.2 Sequence Alignment

Probably the most fundamental task in computational biology is *sequence alignment*. A sequence alignment is a way of arranging the letters of DNA, RNA, or protein sequences to identify regions of similarity due to structural, functional, or evolutionary relationships between the sequences. In a *pairwise* alignment two sequences are aligned against each other and in a *multiple* alignment more than two sequences are aligned. An alignment is called *global* if it spans the full length of the sequences and it is a *local* alignment otherwise. A local alignment between sequences is also called a *match*. Pairwise alignments can be solved in optimal $\mathcal{O}(ab)$ time for two sequences of length a and b [162]. However, in practice programs often resort to heuristics that are not guaranteed to return the optimal alignment. In most of the cases a *seed & extend* strategy is utilized, i.e., starting from exact subsequences as seeds the alignment is extended with more accurate but time-intensive algorithms [162]. Due to the short length of sequencing reads from next-generation sequencing machines that have to be aligned against complete genomes, a number of new approaches have been developed [96, 94, 130, 84, 163, 65]. These programs differentiate between read matches that are *unique* and *non-unique*. An alignment or match is unique, if it is the only best scoring alignment, where the score is often just sequence identity but might also include the quality of the sequencing read [94]. If several alignments have the same score it is non-unique. It is a topic of current research how to handle non-unique matches [60, 113].

Spliced Alignment

If cDNA or EST sequences are to be aligned against the genome, the alignment program has to consider the introns that are in the genome. The cDNA is said to be "spliced" against the genome and therefore this type of alignment is called *spliced* alignment. Most of the programs use a predictor for splice sites in order to improve exon boundary prediction and resort to seed & extend strategies. Routinely used

programs include est_genome [114], sim4 [49], Blat [78], and Exonerate [145]. RNA-Seq read spliced alignment is especially difficult and an approach based on known annotations is presented in Section 3.2.1. More recent developments are discussed at the end of Chapter 3.

2.3 Data structures and Algorithms for Genome Assembly

In the following the two most prominent approaches to *de novo* genome assembly are introduced. The first one is based on the Overlap-Layout-Consensus paradigm that proved useful for a number of whole genome assemblies, for example Drosophila [115]. The second one is based on an Eulerian path approach to genome assembly using de Bruijn graphs [123]. A classical measure for the comparison of genome assemblies produced by different programs is the N50. Given a set of sequences of different lengths, the N50 length denotes the length N for which 50% of all bases in the sequences are in a sequence of length $l < N$, that is why it is sometimes called median weighted contig length. Analogously, the N25 and N75 are defined.

2.3.1 Genome Assembly with the Overlap-Layout-Consensus Paradigm

The traditional approach to genome assembly is the Overlap-Layout-Consensus (OLC) paradigm, which consists of the following three phases:

1. **Overlap:** In an all-against-all comparison of reads, pairwise overlaps between read sequences are discovered. The comparison is done using a heuristic seed & extend approach by finding a set of common k-mers between two reads, which are used as seeds for an alignment between them.

2. **Layout:** An *overlap graph*, where the nodes are the reads and edges indicate overlap between two reads, is manipulated and gives an approximate layout of the read sequences. In this phase uniquely assemblable contigs (*unitigs*) are produced by collecting fragments whose layout is uncontested by overlap

of other fragments. Repeat resolution of the intermediate fragments groups unitigs into larger structures called *scaffolds*.

3. **Consensus:** In the final phase, a multiple sequence alignment from the reads determines the exact read layout and the final consensus sequence.

The OLC paradigm was ideal for assemblies based on long Sanger reads and successful softwares include among others the Celera assembler [115] , Arachne [7, 73], and CAP3 [68]. A few works also adapted the OLC paradigm for reads of length a few hundred bases from next-generation sequencing based approaches like 454 [106, 109] and even for short reads from Illumina or SOLiD [64, 66]. The assembly problem is to find a *Hamilton path* that visits each node exactly once in the overlap graph, which is known to be NP-complete.

2.3.2 Genome Assembly Using the Eulerian Path Paradigm

Different from the idea to create an overlap graph between the reads to find the best layout in the graph, Idury and Waterman [70] and later Pevzner, Tang and Waterman [123] suggested to build a de Bruijn graph of the sequencing reads.

Definition 2.3.1. *A de Bruijn graph of* dimension k *has nodes* V *that represent* k*-mers and edges* E *that represent all suffix-to-prefix perfect overlaps between the* k*-mers in* V*. These overlaps have a fixed size of* k*-1.*

In the original definition by de Bruijn [41] V is the k-spectrum(Σ^n), i.e., the set of all possible k-mers of an alphabet Σ. Therefore the graph has $\Sigma^n \cdot |\Sigma|$ many edges. In the paper by Pevzner et al. nodes V represent all k-mers observed in the reads and edges E represent all overlaps between k-mers observed in the reads [123]. Therefore, Pevzner's de Bruijn graph is sometimes called simply word graph or k-mer graph as it is in most of the cases a subgraph of the original de Bruijn graph [103, 110]. In this work, all de Bruijn graphs are built from data and are thus subgraphs of the original de Bruijn graph.

The main idea for the use of de Bruijn graphs for genome assembly is that, in the perfect case where error-free reads have complete genome coverage and no repeats longer than k exist, the genome would form a de Bruijn graph that contains the genome sequence as an *Eulerian path*. An Eulerian path in a graph, is a path that traverses

each edge exactly once, that is why the program was called the Euler assembler [123]. It was hoped that this approach could lead to a polynomial time algorithm. However, the task is to find an Eulerian path that includes all the read paths, as so called *superwalk*. Medvedev *et al.* [108] formulated this as the *De Bruijn Graph Superwalk* problem and showed that it is *NP*-hard for $|\Sigma| \geq 3$ and any positive integer k.

De Bruijn Graphs for Short Read Data

Although the OLC and the de Bruijn graph approach are *NP*-hard, the de Bruijn graph approach is very appealing for genome assembly from a set of millions of short reads. Most importantly, the time consuming overlap computation in the OLC paradigm is substituted by the detection of overlaps of size k-1 as a direct result of the graph construction. The graph construction is efficiently done using hash table based approaches. In addition, because the de Bruijn graph compresses the information of overlapping reads to the extend that reads share k-mers, it leads to a more compact representation of the original reads, especially compared to the overlap graph, where each node is a read.

Unfortunately, the de Bruijn graph approach suffers a number of practical problems with short read data: (i) Each repeat in the data that is longer than k induces a cycle and complicates the genome reconstruction and (ii) the influence of sequencing errors in the read data increases with size k. Further, the data are double barreled and the original orientation of a read is unknown, which complicates the algorithms (this is true for the OLC approach as well).

Many different assembly algorithms have been designed to cope in different ways with the aforementioned problems, often inspired by approaches introduced in the Euler assembler. Among others, these programs include Euler-SR [25, 24], Velvet [172, 173], Allpaths [20, 100], ABySS [144], and more recently SOAPdenovo [97]. The graph construction and error correction procedures of Velvet are explained in detail, because a *de novo* transcriptome assembler based on Velvet is presented in Chapter 5.

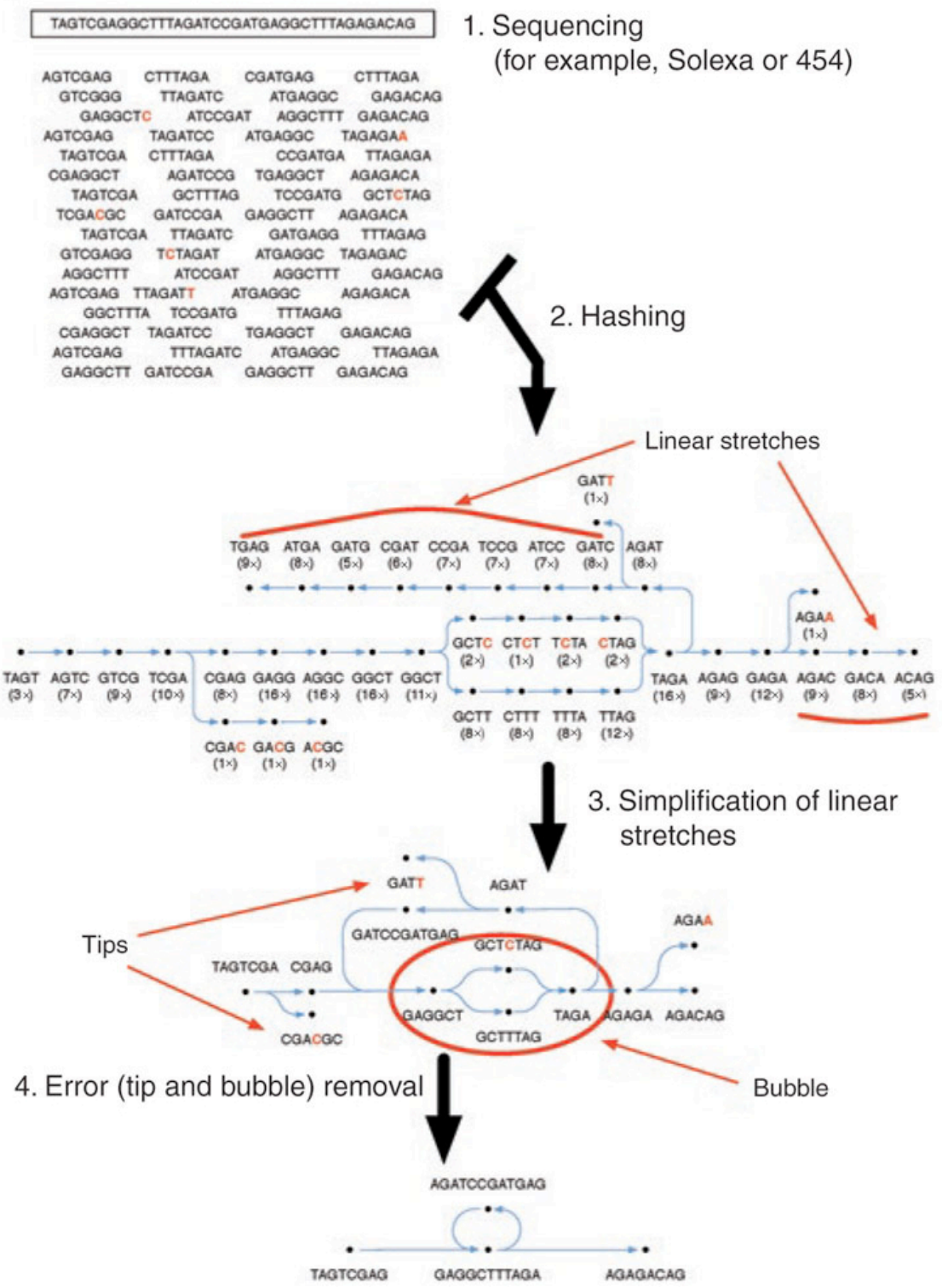

Figure 2.1: The basic steps of Velvet from hashing, simplification of chains, and error correction by tip clipping and bubble removal. Reproduced and modified from [48].

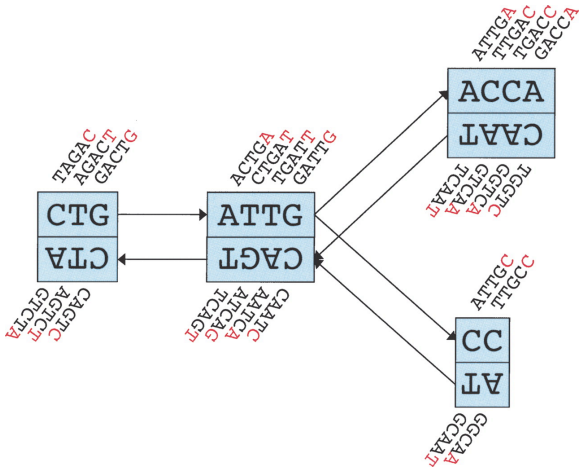

Figure 2.2: An excerpt of a de Bruijn graph in Velvet for k=5. A series of overlapping k-mers (shown above or below) is associated with each node (rectangle) in the graph. The sequence of last nucleotides (in red) of the associated k-mers in a node denotes the sequence of the node. The attached twin node recognizes the reverse series of reverse complementary k-mers. Directed edges are shown as arrows. The last k-mer of an edge's origin overlaps with $k-1$ nucleotides with the first k-mer of its destination. For example the edge from the node labeled "CTG" to the node labeled "ATTG" represents the overlapping string "ACTG" of length $k-1$. Note the opposing direction of edges between two nodes due to the symmetry of twin nodes. Both nodes on the left could be simplified as they represent a chain. Reproduced from [172].

2.3.3 The Velvet Genome Assembler

Velvet first hashes all the reads according to a fixed k-mer length [172], with the program *velveth*. k has to be odd to ensure that a k-mer cannot be the reverse complement of itself. Then a de Bruijn graph from the k-mers in the data is constructed with the program *velvetg*. After construction of the graph chains of nodes are simplified without loss of information, similar to unitigging in overlap graphs, by collapsing them into one node. These steps and the following error correction procedures are depicted in Figure 2.1. Note that a de Bruijn graph can be directly build in its simplified form, thereby reducing peak memory consumption [127].

In order to accommodate data with unknown read orientation a k-mer stored in a node in Velvet has its reverse complement k-mer represented in a so called *twin node*, see Fig. 2.2. As described in the figure legend, the length of a node n in Velvet is the marginal information of each node (red sequence) and is denoted l_n. Any operation on a k-mer node in the graph is symmetrically performed on its twin node.

Error Correction

There are three essential error corection steps for the graph which are depicted in Figure 2.1. The first is tip clipping. A tip is a node that has only one connection and is likely to represent a sequencing error. A tip connected to a node n is removed when its length is smaller than $2k$, to account for two overlapping errors, and when its edge has the smallest number of supporting reads among other edges connected to n. Secondly, small bubbles in the graph are removed by the Tour Bus algorithm, that extracts neighboring paths in a bubble and aligns them against each other. Whether two paths are merged is decided based on three thresholds. (i) Both paths must have less than 200 nodes, (ii) their respective sequences must be shorter than 100 bps, and (iii) the sequence similarity between them must be at least 80%. Lastly, a coverage cutoff for nodes that have not been removed through the previous corrections is applied based on the notion of the k-mer coverage (k-cov) of a node n:

$$k\text{-cov}(n) = \frac{Y_n \cdot (r - k + 1)}{l_n} \ , \tag{2.3}$$

where Y_n is the number of reads in the node and l_n is the length of the node. By default $k\text{-cov}(n) = 3$.

Scaffolding with Paired-end Reads

As most of the short read assemblers, Velvet utilizes paired-end reads to scaffold contigs. The first algorithm, called Breadcrumb [172], was later substituted by the Pebble algorithm [173] that explores distances from *unique* nodes. Unique nodes are flagged by coverage statistics and the shortest path between two unique nodes is found by a heuristic depth-first search. The distance between two nodes is estimated using the maximum likelihood estimator that was used in the greedy-path-merging algorithm in overlap graphs [69]. The estimated distance d between two nodes using paired-end reads is computed as:

$$d = \frac{\sum_{i=1}^{m} \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^{m} \frac{1}{\sigma_i^2}} \ , \tag{2.4}$$

where X_i is the distance according to the i-th read pair and σ_i^2 the variance of the insert length of the i-th pair.

Scaffolding with Long Reads

In addition to paired-end reads, Velvet can make use of long reads, e.g., contigs from previous assemblies. The algorithm, called Rock Band, uses long reads to resolve repeats. Each path between two unique nodes, supported by at least two long reads and not contradicting any other long read, is merged and possibly missing sequence is filled from the long reads [173].

2.4 Data Structures and Algorithms for EST Assembly and Analysis

2.4.1 EST Assembly

In order to assemble EST sequences the first step involves clustering ESTs. The purpose of EST clustering is to collect overlapping ESTs related to the same gene. There are two common approaches. When a reference sequence is available the ESTs can be aligned with a spliced alignment algorithm (see section 2.2) and all ESTs mapping in genomic proximity can be analyzed further [118, 37, 167]. If no reference sequence for mapping is available, the ESTs are often clustered by computing all pairwise alignments between them similar to the overlap phase in genome assembly. Different criteria, like sequence similarity, are defined to form final EST clusters [139, 111, 89, 125].

After EST clusters have been defined, the task emerges to predict consensus sequences that possibly represent full length transcripts of a gene. At first, methods for normal genome assembly, have been used to create single isoforms [118], like the TIGR assembler [89] or CAP3 [68]. However these methods treat the ESTs only as linear sequences. A change in thinking occurred with the proposal of the splicing graph.

2.4.2 Splicing Graphs

The splicing graph was introduced by Heber et al. [62] as follows. Let $\{T_1, \ldots, T_z\}$ be the set of transcript sequences that are given for the gene of interest. Every transcript

sequence T_i is described as a set of genomic positions V_i with $V_i \neq V_j$, for $i \neq j$. The complete set of all *transcribed* positions $V = \cup_{i=1}^{z} V_i$ is defined as the union of all sets V_i [62]. The splicing graph $SG = (V, E)$ is the directed acyclic graph on the set of transcribed positions V that contain an edge from v to w if and only if v and w are consecutive positions in at least one transcript T_j. In such a graph every transcript T_j is represented as a path. Although the SG is the union of z such paths, there might be paths in SG that do not correspond to a transcript T_j.

It is common to simplify the graph by collapsing all nodes v in SG where $indeg(v) = outdeg(v) = 1$, which leads to a more compact representation. Splicing graphs have been used with slight modifications to the original definition of Heber et al. [86, 23, 134]. One important addition to the original definition are *start* and *end* nodes in the graph [45, 134, 133, 23]. The start node is connected to each smallest genomic position of a transcript T_i, i.e., for each T_i there is an edge from the start node pointing to $\min V_i$. Analogously for each T_i there is an outgoing edge from each $\max V_i$ pointing to the end node in a SG. Only when a start and end node are defined, the splicing graph can be used to detect the complete set of alternative exon events defined in section 1.1 [133].

A main feature of splicing graphs is that nodes v with $indeg(v) > 1$ or $outdeg(v) > 1$ are a witness of an alternative exon event. Many algorithms exist for predicting *pairwise* AEEs, i.e., considering each AEE in isolation to other AEEs, from ESTs and cDNA data [86, 23, 134, 45, 57]. Only recently, the classification of ESTs into *complete* AEEs was done by Sammeth [133]. Sammeth introduced an algorithmic framework designed to facilitate the extraction of minimal subgraphs in the splicing graph that explain AEEs between a fixed number of transcripts. In his work he defines a *bubble* as the minimal subgraph that describes the deviation of intersecting paths of transcripts in the splicing graph. The *dimension* of the bubble is the minimal number of paths that disagree inside the bubble. Importantly, this leads to the observation that each AEE is represented in a bubble. For later reference, this fact is formulated for the pairwise case with two transcripts in the next lemma:

Lemma 2.4.1. *Consider a splicing graph $\mathcal{SG}$ build from two transcripts T_i and T_j. Each bubble in $\mathcal{SG}$ describes an AEE between the two transcripts [133].*

A bubble is a cycle in the underlying subgraph. It is straightforward to see that two isoforms have different paths in the splicing graph. The intersecting edges or nodes

in the graph where these paths diverge and converge are created by an AEE between them.

The primary use of splicing graphs is to facilitate the extraction of possible AEEs or possible transcript sequences for a gene. This has been exploited to define the AEE landscape for different species on a genome level [134, 23] and a number of databases exist that have catalogued AEEs for different species using splicing graphs [150, 90, 50].

Note that Heber et al. [62] and later Malde et al. [103] described how to construct a splicing graph from k-mers of the EST data by constructing a de Bruijn graph. This approach is extended to compute the complete transcriptome without reference annotation from RNA-Seq data in Chapter 5.

Chapter 3

Prediction of Alternative Isoforms

3.1 Prediction of Alternative Splicing Events from RNA-Seq Data

Here, we provide a set of methods that enable the detection of alternative exon events (AEEs) within or between conditions using a given gene annotation. First direct detection with short reads that map to splice junction sequences between two exons are investigated. Subsequently the detection of AEEs using read that map to exonic regions of genes are investiaged, The Cell type-specific Alternative uSage Index (CASI) predicts AEEs within a given condition, e.g. one cell line. The Differential Alternative uSage Index (DASI) predicts AEEs differentiating two conditions, e.g. between two cell lines. All methods are based on a stochastic model of the read distribution along a transcript and show high robustness based on simulations. The methods were applied to a new RNA-Seq dataset from HEK and B cell lines.

3.1.1 A General Stochastic Count Model for Transcriptome Analysis

All reads from an RNA-Seq experiment are of the same length r, usually around 25-76 bps. Due to the nature of the RNA-Seq protocol, which involves random shearing of the mRNA molecules (see Section 1.3), it is assumed that the set of sequenced

fragments is picked randomly out of a bag of transcript positions. We assume that the total number of reads T covering a gene is determined by a Poisson process:

$$T \sim \mathcal{P}(\lambda \cdot s \cdot p), \tag{3.1}$$

where s is the total length of the gene, p is the *sample-relative* proportion of the gene compared to all other expressed genes in the sample, and λ is a normalizing factor related to sampling depth or transcript length. The Poisson framework is suited especially for low-coverage datasets, where a normal distribution cannot serve as a good approximation [22]. This model has already been proposed for abundance of EST data [5], as well as SAGE libraries [8]. Marioni *et al.* [107] have further demonstrated that the variation across technical replicates of RNA-Seq experiments can be captured using a Poisson model, as only 0.5% of the genes showed a statistically significant deviation from the model.

For ease of notation, the approach is described for one gene, but all formulas can be extended for a set of genes. Due to the hypothesis that the reads are positioned randomly along every transcript, the number of observed reads within exons

$$Y \;=\; (Y_e)_{e=1\ldots n} \tag{3.2}$$

is drawn according to a multinomial distribution

$$M((p_e)_{e=1}^{n} \cdot T) \;. \tag{3.3}$$

The probability p_e that a read falls in exon e is parameterized for every gene according to the properties of the RNA-Seq experiment. An obvious parameter for p_e is the effective length l_e of an exon. The effective exon length corrects for exonic regions where reads of length r cannot be uniquely mapped due to highly homologous gene families, pseudogenes, repeats or low sequence complexity. Any other information affecting the read coverage – such as GC bias or a bias specific to the protocol used – can be optionally included in the definition of effective exon length. Finally, the normalized expression $\widetilde{y}_e$ of an exon e is defined as the observed exon read count y_e normalized by the exon relative proportion and the gene length:

$$\widetilde{y}_e \;=\; \frac{y_e}{p_e \cdot s} \;. \tag{3.4}$$

In Chapter 4 the model is extended to individual transcripts of a gene.

Computation of Normalized Exon Expression Levels

For each gene, all exons were extracted from the annotated transcript database (Ensembl v.46 [47]) and concatenated, considering the longest exonic form. For the resulting virtual transcripts the sequence of all possible *transcript* reads of length r, in our case 27 bp, was extracted. All *transcript* reads were mapped with the Eland software (Gerald module v.1.27, Illumina) against the human genome and flagged as either unique or non-unique reads. For each exon e, $\phi_e = \{unique\ reads\ in\ e\}$ is recorded and the effective exon length

$$l_e = \frac{|\phi_e|}{s}, \tag{3.5}$$

is computed, assuming a uniform read distribution. This apporach also removes the reads that are shared by duplicated genes or repeat regions within genes. In addition, overlapping gene regions from different strands were excluded from the analysis, as the orientation of the RNA-Seq reads was unknown (see 1.4). This provided the theoretical total number of unique 27-mers representing a given exon. Only 517 genes in the set had exons devoid of any unique read ($\sum_e l_e = 0$) and ca. 1,000 genes were poorly represented ($\sum_e l_e < 0.4$). As suggested above l_e is plugged into the normalized expresson $\widetilde{y}_e$ for an exon (3.4).

3.2 Prediction of Alternative Splicing Events with Exon Junction Read Evidence

3.2.1 Reference based Spliced Alignment of RNA-Seq Reads

First the procedure is explaind that was utilized to map short reads against the genome, as well as to identify known and possibly new exon-exon junctions given transcript annotations. Recall the definition of the splicing graph introduced in Subsection (2.4.2), one obvious approach is to map the reads against the edges in the graph, as all edges correspond to splicing events. The *complete splicing graph* is defined in order to identify previously unobserved exon-exon junctions that agree with a given gene annotation. The complete splicing graph has the same nodes and edges

as the splicing graph, but contains in addition all possible transitive edges, see Fig. 3.1.

A dataset of splice junction sequences from the complete splicing graph was generated for different databases, such that the coverage of detected events was increased. For every annotated gene locus retrieved from the UCSC database (hg18, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/) and from the ElDorado database including EST sequences (Genomatix, release 05/2007) a complete splicing graph was constructed. Splice junctions sequences of length 50-52 bps were retrieved, centered on the junction between the connected exons. In total 2,334,049 and 2,828,506 splice junctions for all gene complete splicing graphs of UCSC and ElDorado were obtained, respectively.

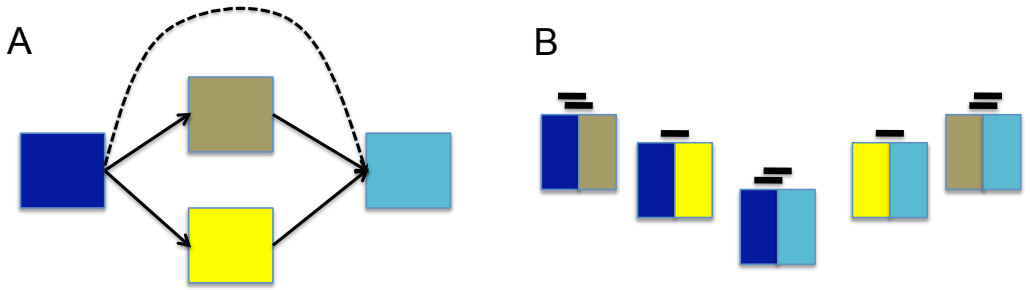

Figure 3.1: A) Depiction of a complete splicing graph for a gene with normal edges (plain) and a new transitive edge (dashed). The dashed edge represents potentially novel splicing events. B) The splice junction sequences adjacent to each edge are extracted and added to the database for read mapping. Reads mapping to the sequences are indicated as black boxes.

Random Model for Junction Hits

The reads obtained from the sequencer were 27 bp long, which brought up the question of what is the probability to obtain random matches on the large sets of extracted splice junctions? A model to study the probability of random hits for reads of length r on splice junctions of length j was considered. Depending on the matching strategy up to σ substitution errors between the read and junction sequence are allowed to occur. Assuming a uniform i.i.d. random model for DNA sequences the probability $P(r, \sigma, j)$ that a read of length r matches a splice junction of size j bps with no more

than σ substitution errors is:

$$P(r, \sigma, j) = (j - r + 1) \sum_{k-0}^{\sigma} \binom{r}{k} \left(\frac{3}{4}\right)^{k} \left(\frac{1}{4}\right)^{r-k}. \tag{3.6}$$

The sum describes the possible sequences of length r that deviate by no more than σ substitution errors [18] and the first factor gives the number of possible matching positions along the junction sequence. The expected number of reads that match the considered splice junctions is calculated by multiplying the number of splice junctions $\mathcal{J}$ and number of considered reads $\mathcal{R}$:

$$E_{r,\sigma,j,\mathcal{J},\mathcal{R}} = P(r, \sigma, j) \cdot \mathcal{J} \cdot \mathcal{R}. \tag{3.7}$$

Expected Number of Splice Junctions per Gene

The expected number of reads hitting splice junctions of a gene by chance is given by:

$$E_J = Y_{ex} \cdot \frac{P_r}{1 - P_r}, \tag{3.8}$$

where Y_{ex} is the number of reads that fall inside exons of the gene, with P_r being the probability that a read from a gene hits a junction and no exonic position:

$$P_r = \frac{(m-1) \cdot (j - r + 1)}{s - r + 1}, \tag{3.9}$$

where s is the gene length, m the number of exons in the gene, and j the length of a splice junction. The complete gene with all its exons was considered as one isoform, but exons for which less than half of the positions corresponded to unique hits were ignored for the estimation.

Identification of Alternative Splicing Events from Exon Junction Reads

After having mapped the reads to the complete splice graph, it is straightforward to predict alternative splicing. An edge E is said to be *expressed*, if at least one read maps uniquely to edge E. Given gene annotation all edges of the complete splicing graph are flagged as either *constitutive* or *alternative*. Whenever an alternative edge is expressed in a complete splicing graph, the gene is considered as alternatively spliced. In the present analysis alternative edges in the graph describing skipped exons, 5', and 3' alternative splice sites were considered, see Section 1.1.

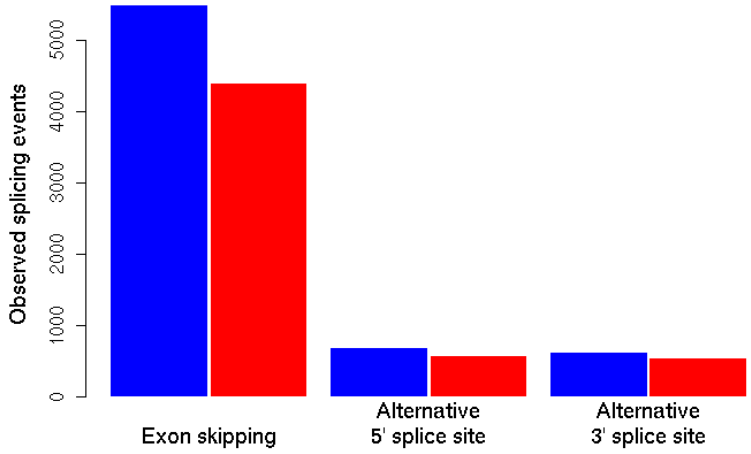

Figure 3.2: Distribution of the three major types of alternative splicing: cassette exons, alternative 5' and 3' splice sites in the HEK (blue) and B (red) cell lines

3.2.2 Application to Human RNA-Seq data

The mRNA content of two human cell lines, a human embryonic kidney (HEK) and a B cell lymphoma cell line were sequenced using single end RNA-Seq. In total 8,638,919 and 7,682,230 reads of length $r = 27$ for HEK and B cells were aligned to the human genome (hg18, NCBI build 36.1) using Eland software (Gerald module v.1.27, Illumina). The mapping criteria imposed by Eland allow up to two mismatches. With this setup, 50% of the reads matched to locations unique in the human genome, whereas 16-18% of the tags mapped to more than one genomic position (Table 3.1).

Reads not mapping to the genome were mapped in a second round to two sets of extracted splice junctions from UCSC and ElDorado again with Eland, see above. 75,662 and 59,889 splice junctions were identified on the UCSC junction set, for HEK and B cells, respectively. Whereas 69,952 and 56,000 splice junctions had reads matching to the Eldorado junction set. The two resulting datasets from Eldorado and UCSC were merged to have one reference data set of splice junctions. For merging the genomic start and end positions of the splice junctions were compared with a tolerance of +3/-3 bps. Redundant junctions were removed leading to a total set of 83,239 and 66,330 identified splice junctions for HEK and B cells, see Table 3.1.

On average 7.2 junctions per gene and a mean density of 3.8 reads per junction were

observed. Although 29,689 junctions in HEK and 24,848 in B cells had only one read, those were considered highly significant as at most 23 reads hitting a junction by chance are expected in the entire dataset as computed using the random model for junction hits. Splice junctions were associated with 81% of the expressed genes. The fact that 2,275 expressed genes in HEK and 2,013 in B cells had no splice junction reads correlats with the fact that those genes contain fewer exons and a lower expression than the average, thus reducing the probability to hit a splice junction.

95% of the splicing events expected in this dataset were observed, given the current sequencing depth (Table 3.1). A set of 4,096 novel splice junctions in 3,106 genes were identified, mostly called by single reads and unique to one cell type. Many of the new junctions were associated with actively transcribed genes exhibiting more exons than average, pointing to rare splicing events. in total 6% of all splice junction reads identified AS events (6,416 junctions in 3,916 genes in HEK and 5,195 junctions in 3,262 genes in B cells, (Supplemental Table S0A-B).

Within a cell type, junction reads identify AS in 30% of the genes expressed genes, where exon skipping was largely over-represented, see Fig. 3.2. An example of alternative splicing is given for the 6th exon of the *NONO* gene (Fig. 3.5 A). It was observed that splice junction reads allow the detection of very complex patterns of AS. For instance, for the gene *EIF4G1*, coding for the eukaryotic translation initiation factor 4 gamma, 12 AS junctions in B cells were found, of which five were novel. While AS is known to regulate the expression of *EIF4G1* [21, 34], such a complex pattern was never described before.

3.3 Prediction of Alternative Isoforms with Exon Expression Levels

In here we turn to exploit another source for the detection of alternative exon events, namely the reads residing in exonic regions. For reads as short as 27 bp this constitutes another rich source of information. First, a test framework to detect AEEs occurring within a given cell type (CASI method) is provided and later a framework to test the presence of different isoform patterns between two cell types or conditions (DASI method). In both cases, a two-step procedure was applied, which (i)

	HEK cells	B cells
Total reads	8,638,919	7,682,230
Low quality reads	234,160	194,999
Reads with non-unique matches	1,546,361	1,324,770
Reads with unique matches	4,640,112	3,895,643
Reads mapping to RNAs (Ensembl + Eldorado)	3,712,476	2,902,387
Ensembl genes with at least 5 reads	12,567	10,668
Ensembl genes with at least 1 read	14,963	13,739
Reads with no match to the genome	2,218,286	2,266,818
Reads aligned to splice junctions	307,904	229,453
Identified junctions (expected)	78,880 (81,302)	62,596 (66,981)
Genes (at least 5 reads) with junctions	10,292	8,655
Genes (at least 1 read) with junctions	10,558	8,910
Genes (at least 1 read) with novel junctions	2,078	1,732
Novel junctions	2,397	1,965
Novel junctions identified by > 1 read	203	182

Table 3.1: Analysis of the reference based mapping statistics for HEK and B cell RNA-Seq data.

detects genes with AEEs based on CASI and DASI p-values and (ii) highlights exons predicted to be alternative according to a z_e-score statistic. The z_e-score statistic is computed for each exon e as:

$$z_e = \frac{R_e - median(R_*)}{MAD(R_*)} , \tag{3.10}$$

where R is defined according to each exon log normalized expression or expression ratio (see below). The *median* and maximum absolute deviation (MAD) were used as robust estimates of mean and standard deviation to avoid a bias for genes with few exons. This statistic assumes that the majority of the exons are constitutive.

3.3.1 Alternative Exon Usage within a Condition

Under the null hypothesis that one transcript includes all the exons of the gene, the counts within exons follow a multinomial distribution of parameters p_e and T. The presence of AEEs within a condition was assessed by using Pearson's chi-square test on Formula (3.3), where the p-value was corrected for multiple testing using the Benjamini-Hochberg procedure [9]. A gene with a small CASI p-value means either that (i) two or more transcripts from one gene are present or (ii) a single isoform is present that expresses only a subpart of the annotated exon. Case (ii) can correspond to events of alternative donor or acceptor sites, where only a part of the exon is expressed. The z^C score (CASI) is computed for each exon according to its log-normalized expression

$$R_e^C = \log(\widetilde{y}_e). \tag{3.11}$$

Exons with less than five counts were not considered for CASI computation. Only genes with at least two expressed exons were tested. The CASI p-value was set to 0.05.

Simulations

In order to assess the theoretical accuracy for CASI, simulation of a single exon skipping event for a template gene model were conducted. The inclusion rate of the exon as well as the length of the skipped exon and the gene expression level varied. Reads were drawn randomly along exons according to the distribution introduced

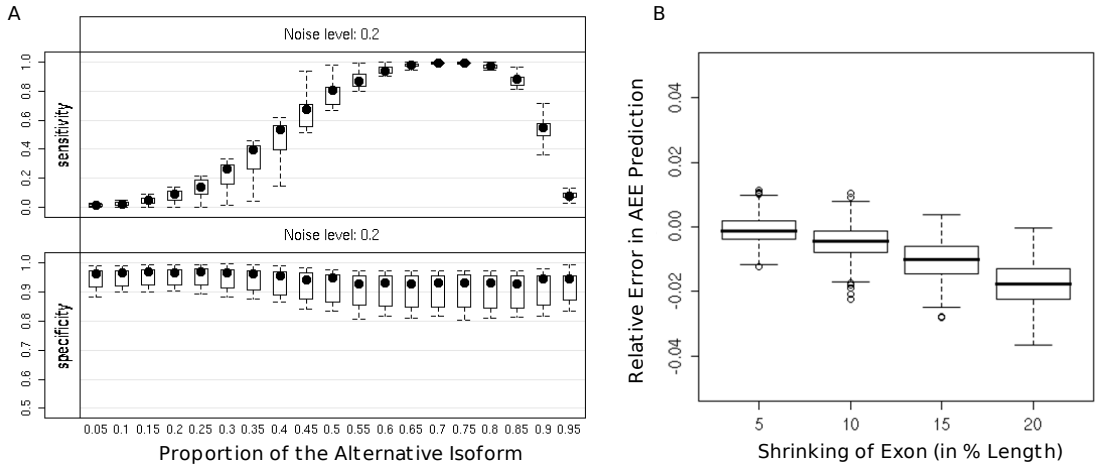

Figure 3.3: (A) Boxplots of sensitivity and specificity (y-axis) for CASI AEE prediction for different alternative isoform proportions (x-axis) based on simulations with 20% noise. (B) Robustness estimation for predictions on HEK data shown as boxplots. The change in predicted number of AEEs is shown relative to the total number of predictions for the whole dataset (y-axis) for 500 bootstrap samples using a $z_e^C \leq -2$. The x-axis shows the reduction in length that was introduced to an exon at random ($p = 0.25$).

previously. Noise was introduced by choosing one exon at random and artificially modifying the proportion of reads mapping to it by 20%.

The simulations for CASI assumed a gene with six exons and a length of 150 bp per exon. Different expression levels for two isoforms (proportion 0.05-0.95) were simulated such that the total read number in all exons was 300. The proportion of genes detected by CASI where the skipped exon was properly flagged as an AEE (sensitivity) was evaluated. Similarly, the proportion of genes detected by the test where only a truly skipped exon was predicted as an AEE (specificity) was evaluated. For different levels of noise, 500 simulations on 1,000 genes were performed.

Figure 3.3A shows the specificity and sensitivity of CASI predictions with noise. The predictions are very robust with >80% specificity. For low expression values of the minor form the test is not able to predict the AEE. The sensitivity increases according to the expression level of the minor isoform and for values between 0.65 and 0.8 it reaches 100% sensitivity.

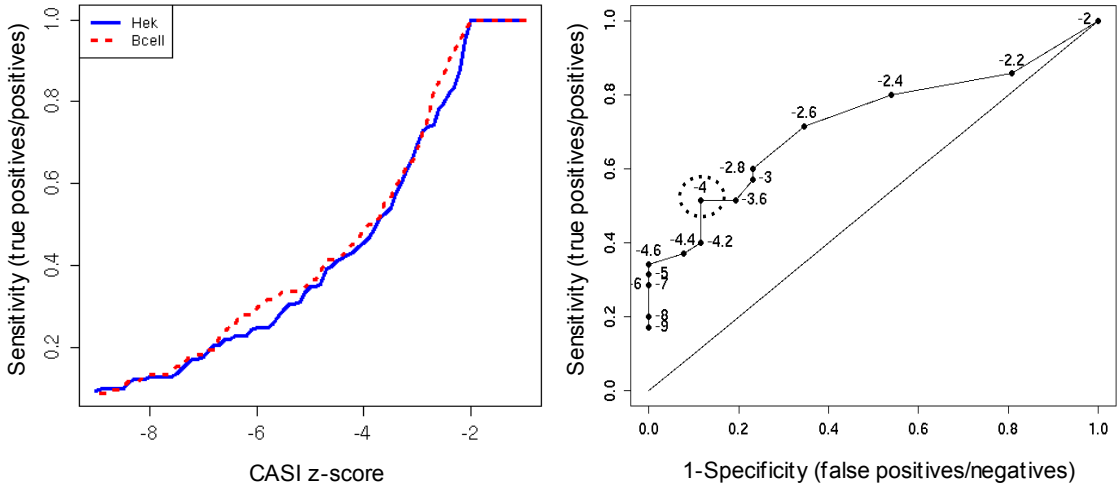

Figure 3.4: Validation of the CASI method with splice junction reads and RT-PCR. Left) Sensitivity of CASI predictions compared to splice junction reads. All AEEs detected by at least 3 splice junction reads are taken as the positive set. The y-axis shows the percentage of CASI AEEs that overlap with the positive set (Sensitivity) for different values of z_e^C on the x-axis. Right) ROC curve of RT-PCR results (positive/negative) testing 61 AEEs predicted by the CASI method. Each exon tested by RT-PCR was associated to its corresponding CASI z_e^C (numbers at each each data point). The best qualifier uses a $z_e^C \leq -4$ (dotted circle) with a specificity of 89% and sensitivity of 51%. Note that the sensitivity for CASI predictions derived by comparison to splice junction reads (left figure) is highly similar with 48% sensitivity for a $z_e^C \leq -4$ in both cell lines.

Application to Human RNA-Seq Data

The CASI test was calculated for all genes expressing at least two exons in a given cell line (12,140 genes in HEK and 10,417 genes in B cells). A total of 7,991 genes in HEK and of 6,837 genes in B cells showed a significant CASI p-value (see above). Data were filtered further by imposing a threshold on the CASI $z_e^C \leq -2$ to yield maximal sensitivity (see below). There remained 4,459 genes in HEK and 3,490 genes in B cells with a significant CASI, for which 6,869 and 5,008 AEEs were predicted, respectively. CASI predicted more than one AEE for 666 and 841 genes in HEK and B cells, respectively. A total of 2,650 AEEs (in 2,428 genes) were shared between HEK and B cells pointing to events common to very diverse cell types (Supplemental Table S1 A-B).

Sensitivity and Bootstrap Analysis

A data-based estimate of sensitivity for CASI predictions was derived for AEEs identified by reads mapping to splice junction sequences (see Subsection 3.2.1). The set of identified AEEs with reads mapping to splice junction sequences was compared to the set of AEEs predicted by CASI for varying z_e^C values, see Fig. 3.4A. At a $z_e^C \leq -2$, all AEEs identified by splice junctions were predicted by CASI, such that the sensitivity reached 100%.

Local heterogeneity of the read distribution along a transcript could lead to false positive predictions. Possible sources for an uneven read distribution along a transcript are preferred break points of the RNA fragments in the sample preparation step or a higher sequencing efficiency for short cDNA fragments with certain sequence characteristics [43, 113, 59, 95]. It was ruled out that such unevenness significantly affects the predictions by performing a bootstrap procedure for each gene.

For each bootstrap sample (total of 500), each exon of a given gene was randomly picked, with a probability of 0.25, and shortened on one end by 5, 10, 15 and 20%. Only exons with more than 80 unique positions were shortened. The read count and the effective exon length l_e for an exon were recomputed for each shortened exon and treated as a new transcript annotation set. The prediction was repeated on every new transcript annotation set. In this context, a highly uneven read distribution will significantly impact the number of predictions. However, the predictions are shown to be very robust with less than 5% relative error even when up to 20% of the exonic region was removed (Fig. 3.3B).

Experimental Validation

In order to optimize the CASI predictions, a subset of predicted AEEs was tested by RT-PCR. Though CASI does not provide indications on the nature of the detected AEEs, the PCR experiments were designed for testing exon-skipping events, as it is the most prevalent form of AS in this cell lines, as shown in Section 3.2.1. A selection of 61 AEEs (50 in HEK and 11 in B cells) was tested, of which more than 50% had CASI as the sole indicator of an alternative isoform (Supplemental Table S2). Thirty-five CASI predictions were validated as true exon skipping events, of which 17 were not supported by junction reads. This emphasizes the power of CASI in identifying

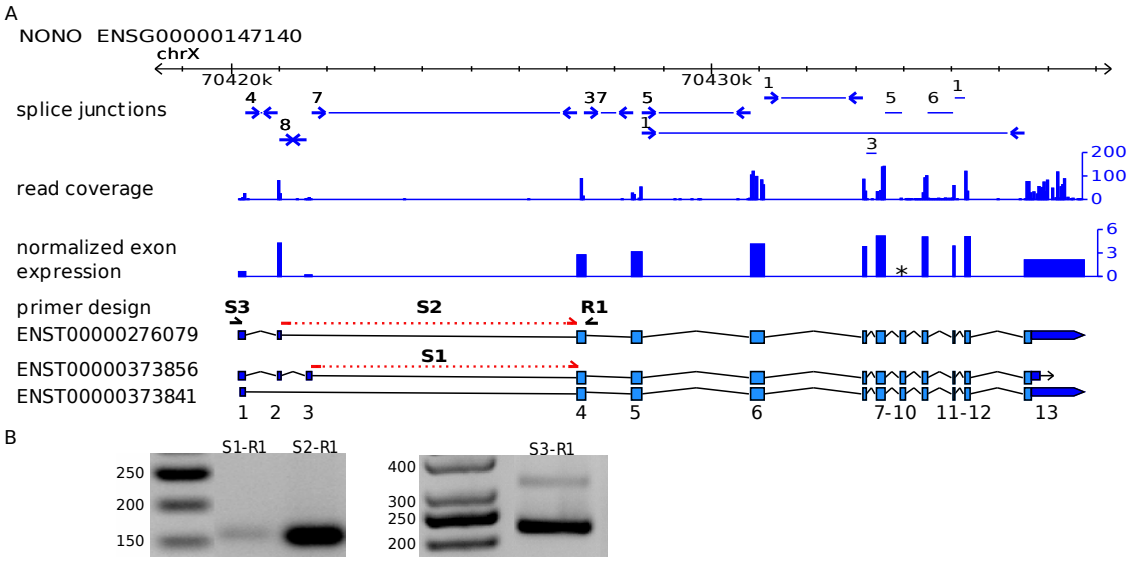

Figure 3.5: (A) RT-PCR validation of a predicted AEE in the *NONO* gene in HEK cells (CASI); it shows the observed exon-exon junction (blue arrows) and the corresponding number of reads (above the arrows) for all exons of the three annotated isoforms (Ensembl v.46). S1 and S2 primers are placed on the splice junctions of the constitutive and the skipped forms, respectively (red dashed line) to uniquely amplify two different splice variants of *NONO*. R1 and S3 primers were designed inside surrounding exons. Exons not considered in CASI analysis are marked by an asterisk. (B) Agarose gels (1.5%) showing the RT-PCR amplification results of S1-R1, S2-R1 and S3-R1 fragments. The observed sizes of the bands correspond to the expected sizes.

AEEs as illustrated for the *NONO* gene (exon three, Fig. 3.5A). Among the 26 AEEs that could not be validated, one likely false negative case was observed, corresponding to a skipped exon in the gene *TCOF1* in HEK supported by only one junction read. As the remaining 25 CASI predictions could, in principle, involve alternative donor or acceptor sites, it was examined whether other sources (e.g. junction reads, ESTs, or annotations in Ensembl) provide clues that could infer these types of AS. Indeed, nine exons were annotated for another type of AEE in at least one source, among which four AEEs were detected by junction reads, such as the usage of an alternative acceptor site in the *DUS1L* gene (Appendix Fig. 6.1). Based on these experimental verifications, the specificity of the CASI predictions was estimated to be close to 60%.

Further, the predictive power of the procedure was estimated by using the receiver operating characteristic curve (see Section 2.1.3), where each exon tested by RT-PCR (negative or positive) was associated with its corresponding z_e^C. Based on these PCR

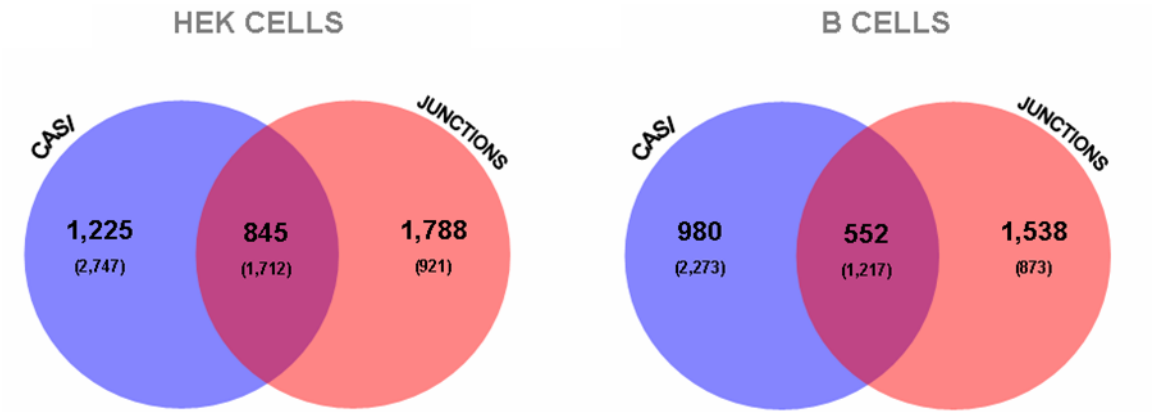

Figure 3.6: Comparison of alternatively spliced events identified with CASI $z_e^C \leq -4$ ($z_e^C \leq -2$) and junction reads: Venn diagrams show the number of genes, for each cell line, with at least one AEE according to CASI (blue) or splice junction reads (red). A gene is selected if any of its exons (3', 5', or internal exons) is flagged as an AEE. The comparison is based on the total set of genes analysed with CASI (12,140 in HEK and 10,417 for B cells).

results, a specificity of 89% and a sensitivity of 51% was obtained for the CASI method ($z_e^C \leq -4$), in line with the genome-wide estimate of sensitivity deduced by splice junction analysis (Fig 3.4B). The number of false positives (1 - specificity) cannot be deduced from the mapping of splice junctions alone, due to the problem of non-unique spliced alignment matches for short reads and the low expression of many alternative transcripts. However, the simulation and bootstrapping results hint to the fact that the number of false positives is not much higher than the 11% observed by PCR experiments on 61 AEEs. By applying the more conservative threshold ($z_e^C \leq -4$), 2,499 AEEs in 2,070 genes for HEK and 1,775 AEEs in 1,532 genes for B cells were predicted, respectively. Of those 712 AEEs in 693 genes were common to both two cell lines.

It is of particular relevance to compare the respective performances of CASI versus prediction with splice junction reads in their abilities to detect genes with AEEs. Out of the 3,858 genes predicted to have an AEE by any of these two methods in HEK cells, only 845 were detected simultaneously by CASI and junction reads, see Fig. 3.6. Moreover, there are notable qualitative differences in the detected AEEs. Splice junction reads revealed a larger number of internal AS exons [147], whereas most of the events detected by CASI targeted terminal exons, particularly the most 3'-exons (Fig. 3.7).

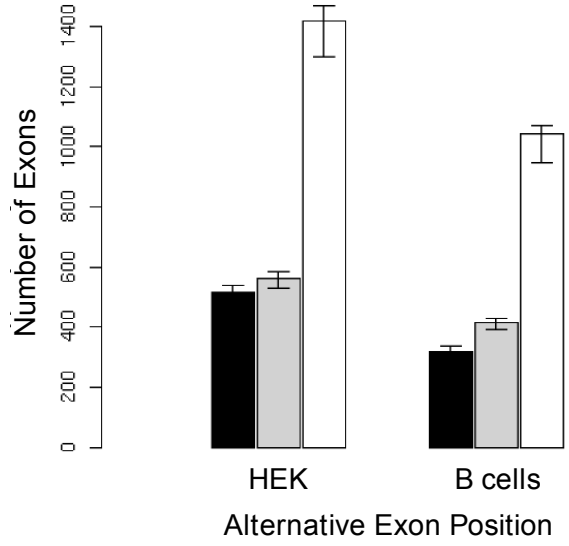

Figure 3.7: Distribution of the number of AEEs predicted by CASI. Bars show the number of 5'- (black), internal (grey) and 3'-exons (white) predicted as AEE with CASI ($z_e^C \leq -4$). The whiskers were obtained by shortening the length of the 5'- and 3'-exons artificially by 20% in order to estimate the error due to the annotation in the 5'- and 3'-end of a gene.

EST-based Validation

The significant expression variation detected in terminal exons might reflect the presence of multiple alternative polyadenylation sites, which are generally poorly annotated in the current databases. As an independent set for validation of the predicted AEEs EST data was consulted. A set of genes with detected alternative polyadenylation sites from EST data from the GeneNest database was generated [58] by screening for putative polyadenylation signals (sequences AATAAA and ATTAAA). To generate a high-confidence set, only EST sequences annotated as 3'-end sequences and aligned in the appropriate orientation were selected. A reliable polyadenylation signal was defined when at least two ESTs carried a putative polyadenylation signal within their 3'-terminus (less than 35 bp) at the same position in the cDNA consensus sequence. Signals not supported by the respective genomic sequence were discarded.

Globally, differential expression involving the 3'-terminal exon was frequently observed in the human dataset, in particular in genes annotated for alternative polyadenylation sites based on the independent EST dataset (B cells: $3.3e^{-244}$, HEK cells:

$1.6e^{-291}$, hypergeometric p-value). This is in line with the observations of Sandberg *et al.* [135], who showed that a large fraction of genes in proliferating cell lines express shortened 3'-UTRs. The gene *HIP2* reported in the publication shows the same behaviour (Appendix Fig. 6.2).

In addition, a set of genes with alternative TSSs from EST consensus sequences that were mapped to the human genome was compiled. For each Ensembl gene, only consensus sequences covering at least two exons and with an exon boundary quality 50 (defined by Gupta *et al.* [55]) were selected. The 5'-termini of mutually exclusive first exons of these consensus sequences were defined as putative TSS. Again, CASI 5'-terminal exons were more frequently found in genes annotated for an alternative TSS, namely 67% in HEK and 74% in B cells ($5.6e^{-17}$ and $4.1e^{-14}$, hypergeometric p-value).

These results illustrate the complementarities between CASI and junction reads for detection of alternative exon usage within one condition. CASI performed better than junction reads for identifying rare splice junctions, whereas junction reads can detect multiple AS events for complex transcript isoforms where CASI performance is poor. In terms of AEEs involving internal exons, only one-fourth of the CASI predictions were corroborated by junction reads. Further, the predicted AEEs ($z_e^C \leq -4$) were compared against a set of 73,948 known AEEs in EST data (Genenest EST database[58], Supplemental Table S3). Data indicated that 22% (126 out of 563) of the predicted internal AEEs in HEK cells and 24% (98 out of 414) in B cells were novel, and that most of these novel AEEs were cell type specific. Taken together, the data indicate that 30% of the genes are expressing alternative isoforms in each cell type. In combination, in these two cell types 49% of the genes express alternative isoforms.

3.3.2 Alternative Exon Usage between two Conditions

So far the expression of alternative isoforms in one condition was analyzed. Another important question, namely to identify AEEs differentiating between two conditions, e.g. control and disease sample, was investigated. Two observed read distributions $y^1 = (y_1^1, \ldots, y_m^1)$ and $y^2 = (y_1^2, \ldots, y_m^2)$ are considered for the same gene in two different experiments. The difference in exon usage pattern between the two conditions

was analysed for every gene conditionally on its expression in both conditions. The presence of differential AEEs was assessed with the FDR-corrected p-value of Fisher's exact test (Benjamini-Hochberg procedure [9]). Every exon e of the gene was assigned a z^D score (DASI) based on the log-ratio of reads between the two experiments

$$R_e^D = \log\left(\frac{y_e^1}{y_e^2}\right) . \tag{3.12}$$

Exons with less than five read counts in both conditions were not considered. A pseudo count of 1 was added to y_e^1 and y_e^2 if its original value was 0. The DASI p-value cutoff was set to 0.05 and the $z_e^D \geq 2$. Genes showing a significant difference between the two biological replicates were removed from the DASI analysis.

Application to Human RNA-Seq Data

This procedure was applied to the 9,242 genes expressed in both HEK and B cells (genes with at least 5 reads), leading to the identification of 613 genes with a significant DASI p-value (5%). After applying $|z_e^D| \geq 2$, it was predicted that 968 exons (in 365 genes) were differentially expressed between the two cell types (Supplemental Table S6), from which the majority (78%) were internal exons. A total of 161 genes had more than one differential AEE between HEK and B cells.

Functional Analysis

In order to check for functional categories that are enriched in the set of DASI genes the Bibliosphere package [137] associating gene ontology categories and PubMed literature mining was utilized with default parameters. Analysis of the functional properties of these 365 genes showed that DASI-predicted genes were enriched for factors whose molecular functions are related to translation and RNA metabolic processes, nucleic acid transport, ribonucleoprotein complex biogenesis and assembly and transcriptional regulation. Three transcription factors (*MEF2B*, *MAZ* and *SMARCB1*) were among the top 20 genes showing the most significant DASI p-values (Appendix Table 6.1). The most striking candidate, *MEF2B*, known to be involved in B cell differentiation [148] showed indeed an alternative TSS in B cells (Appendix Fig. 6.3), suggesting the usage of alternative promoters associated with its specific function.

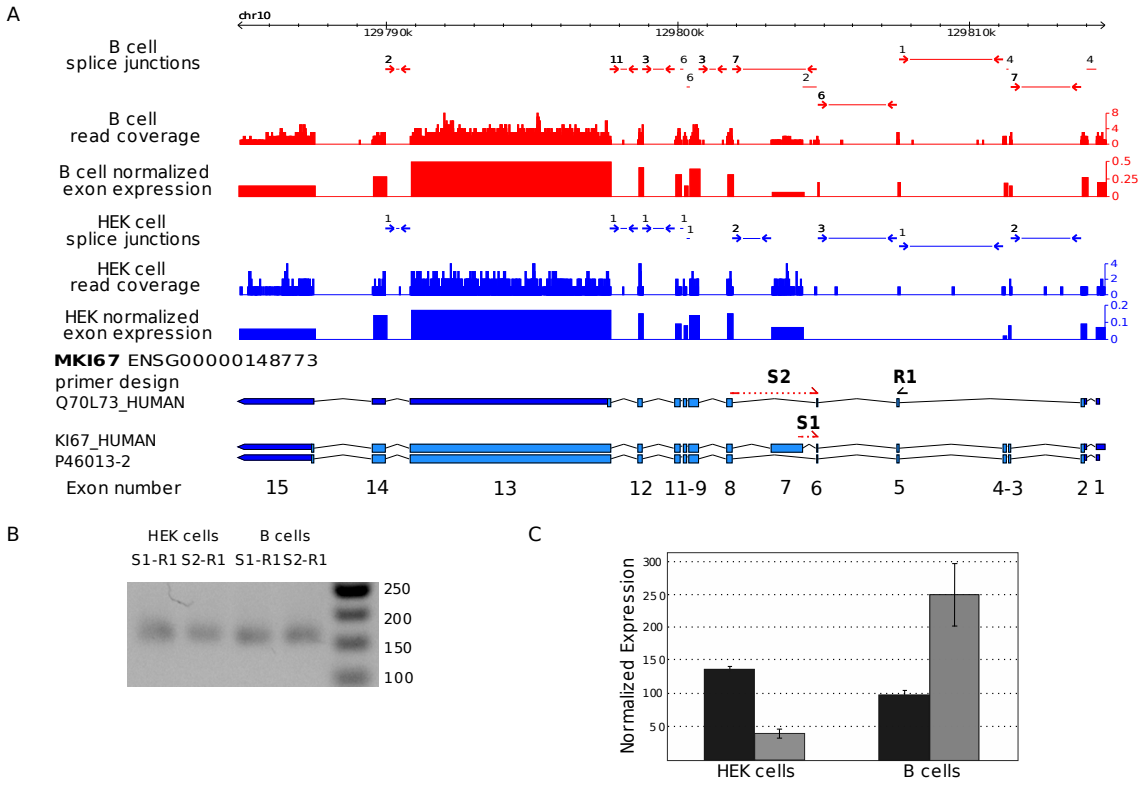

Figure 3.8: qPCR validation of a predicted AEE in *MKI67* between HEK (blue) and B cells (red) (DASI). (A) Screenshot of the *MKI67* gene. The primers were designed to compare the inclusion rate of exon 7 between HEK cells and B cells. (B) RT-PCR results validate the presence of the constitutive and the skipped form in both cell lines. For both S1-R1 (constitutive) and S2-R1 (skipped), a PCR product of length 163 bp is expected if the form is expressed, otherwise no band should be visible. (C) Bar charts representing the normalized expression values for the constitutive form (black) and the skipped form (grey) obtained by qPCR. The results show that the skipped form is more abundant in B cells relative to the constitutive form, as predicted by the DASI method ($z_e^D = 5.2$).

Experimental Validation

As before a subset of 16 high-scoring DASI events was analyzed further by qPCR experiments. Comparison of the expression ratios of the skipped versus constitutive exons between the two cell lines showed that the DASI predictions and the qPCR results were concordant, with a validation rate of 69% (considering a fold change of at least 1.5 for the qPCR) (Supplemental Table S7). An illustrative example is the proliferation marker gene *MKI67*, which is universally expressed in proliferating cells but almost absent in quiescent cells [153]. The *MKI67* mRNA that contains the large exon 7 is equally abundant in B cells and HEK cells, but the skipped form without

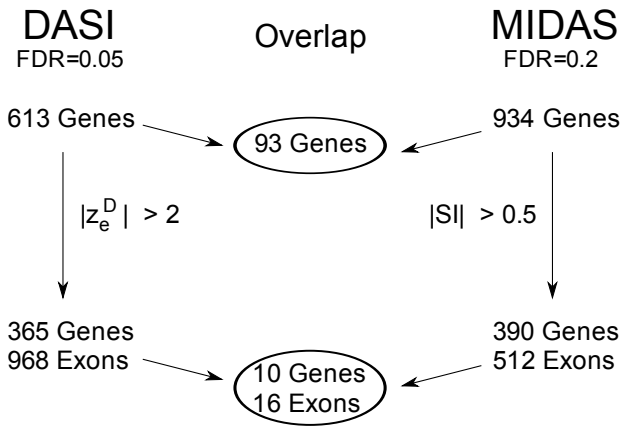

Figure 3.9: Comparison of RNA-Seq (DASI) and exon arrays (MIDAS) for differential exon usage analysis between HEK and B cells. Both methods consist of 2 steps: 1) a corrected p-value identifies genes likely to be alternatively used between HEK and B cell (FDR-corrected p-value ≤ 0.05 and ≤ 0.2 for DASI and MIDAS, respectively); 2) AEEs are scored according to a exon usage index called DASI for RNA-Seq and SI for exon-arrays. The number of predicted AEE genes that passes the threshold criteria at each steps are shown on the figure.

exon number 7 is more highly expressed in B cells than in HEK cells (Fig. 3.8).

Comparison with Exon Arrays

Previous attempts, to systematically decipher AEEs occurring in different conditions or tissues, used of exon arrays alone or in combination with splice junction arrays [51, 28, 52, 39]. For comparative purposes, the human Affymetrix exon arrays 1.0ST were interrogated using the same source of material as well as one biological replicate. A model-based analysis for tiling arrays [76] was applied to perform the intrachip normalization, with the adjustment for exon arrays described by Kapur *et al.* [77]. Quantile normalization was then applied between arrays [71] from the Affy package in BioConductor [53]. Detection call p-values were computed for each probe set with a paired Wilcoxon signed rank test that compares probe intensity to control probes of similar GC content. More precisely, each probe is compared with the 75% quantile of the set of control probes with similar GC content. The detection call p-value of a probe set was calculated using the chip-wise pairing of probe intensities to control intensities. An exon or gene probe set was called present when the corresponding FDR corrected p-value was below 5% (see Gardina *et al.* [52]). The probe-to-exon and probe-to-gene assignment was done using a chip description file

(HsEx10stv2_Hs_ENSE), based on Ensembl v.46, and provided as R package [38]. Exon and gene expression were defined as the mean over probe intensities for both replicates.

For sake of simplicity, the present analysis focused on the probe sets corresponding to all exons annotated in Ensembl, i.e. 149,079 exons in 16,527 genes. A total of 70,627 exons (9,322 genes) in HEK cells and of 57,406 exons (7,823 genes) in B cells were found expressed by both technologies. In terms of detected exons, arrays and RNA-Seq were in agreement, where 90% of the genes detected by exon arrays were also scored by RNA-Seq. As previously reported [147], RNA-Seq is more sensitive than arrays, with 26,300 and 23,866 additional exons detected in HEK and B cells.

For detecting differential AS with exon arrays, the standard MIDAS algorithm of the Affymetrix ExAcT software version 1.8.0 [1] on normalized values was employed. A log ratio for each exon e between condition 1 and 2, called the Splicing Index (SI_e), is introduced:

$$SI_e = \frac{log_2\left(\frac{exp_e^1}{exp_g^1}\right)}{log_2\left(\frac{exp_e^2}{exp_g^2}\right)} \tag{3.13}$$

where exp_e^1 denotes the expression of exon e and exp_g^1 denotes the gene expression for condition 1. MIDAS employs the Splicing Index in an ANOVA model to test the hypothesis that no alternative splicing occurs for a particular exon [52, 1].

The MIDAS p-values were subsequently corrected using the Benjamini-Hochberg procedure [9] and the threshold set to 0.2. This threshold was chosen since only a single gene was found with a corrected p-value <0.05. The threshold for the SI_e ($|SI_e| \geq 0.5$) was set as reported previously [52]. The following filters were further applied: (i) the corresponding gene is expressed in both conditions, (ii) gene expression is higher than the 50% quantile in both conditions, and (iii) the exon is called present in either one of the two conditions.

Comparison of the DASI results with MIDAS showed little agreement in the detection of genes with AEEs between HEK and B cells (10 genes with 16 exons are in common, Fig. 3.9). All genes with predicted AEEs by DASI and MIDAS were among the most highly expressed ones in both cell lines.

In order to investigate the platform differences, the quadratic mean distance for every gene was calculated, between RNA-Seq and exon arrays over exon expression log-

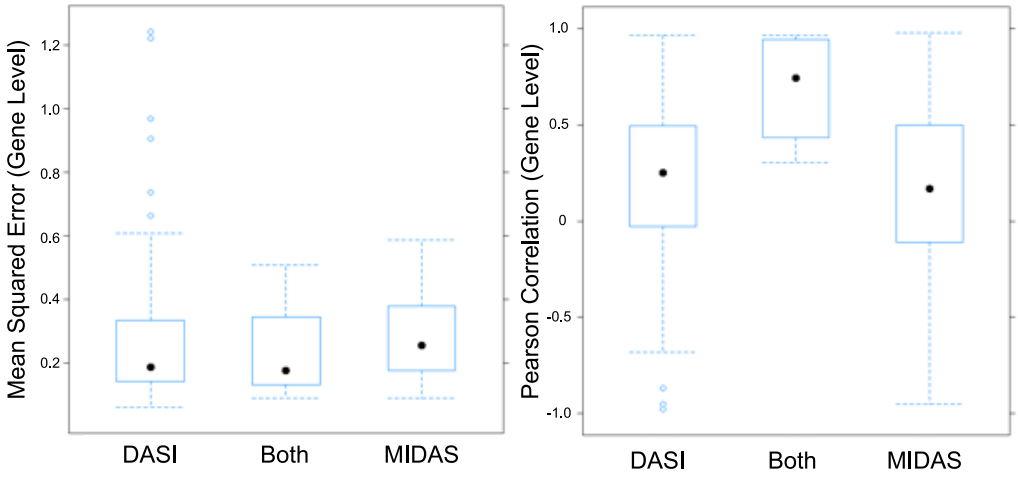

Figure 3.10: Left) Boxplot of the gene-wise mean squared error in exon-expression log-ratios measured with RNA-Seq and exon arrays. The error (y-axis) is shown for 1) genes with predicted AEE only by DASI, 2) genes with predicted exon with DASI and MIDAS (both), and 3) genes with predicted AEE only by MIDAS. Right) Boxplot of gene-wise Pearson correlation of exon-expression log-ratios from RNA-Seq and exon arrays. The Pearson correlation is plotted according to the same three classes as before.

ratios (HEK versus B cells). The quadratic mean distances associated with genes with AEEs predicted by either DASI only, MIDAS only, or by both methods simultaneously did not show major differences (Fig. 3.10). The lack of agreement between the methods could reflect the fact that the analysis of alternative isoforms is very sensitive to subtle variations in expression values that arise both at the individual exon and whole gene expression level, eminent from a difference in correlation at the gene level. In this context, a minor variation of expression between exons is a pre-requisite for pinpointing variable exons with a reasonable specificity. This problem appeared to be less prominent with RNA-Seq, showing clearly a smaller variation of expression values across exons of a given gene (Appendix Fig. 6.4). The *RCC1* gene, for example, was detected by DASI and validated by qPCR, but not detected by MIDAS (Appendix Fig. 6.5). In this case, the alternative exon was below the detection threshold on arrays. Only two of the eight DASI predictions verified by qPCR were also detected by MIDAS (genes *MDC1* and *MKI67*).

Chapter 4

Quantification of Alternative Isoforms

4.1 From Gene to Transcript Expression Levels

In the previous chapter the identification of alternative exon events (AEEs) was described. Another important issue is to estimate the respective proportions of the various transcript isoforms. Given the increased sensitivity and coverage of RNA-Seq data it can be expected that transcript expression measurements can be derived substituting gene expression measurements. As demonstrated below, consideration of the complete gene as the expressed sequence region can lead to false conclusions.

Consider the problem of the construction of a gene regulatory network. Gene expression measurements for genes with alternative isoforms could be imprecise and lead to false positive and false negative associations of regulatory factors. For example, in Fig. 4.1 (top), regulatory sites of TFs and miRNAs are annotated to the complete gene region. However, considering the complete gene hides the fact that binding sites of TF1 and miRNA are not involved in the regulation of the expressed transcript B in the gene (bottom Fig. 4.1), but instead a different factor TF2 is involved in the regulation. Especially for higher eukaryotic organisms, where a large fraction of genes expresses alternative isoforms, network analysis or inference based on gene expression rather than on transcript expression has a limited accuracy.

In this chapter a new PrOportion EstiMation (POEM) Method is explained that estimates the abundance of known isoforms based on a probabilistic model that inte-

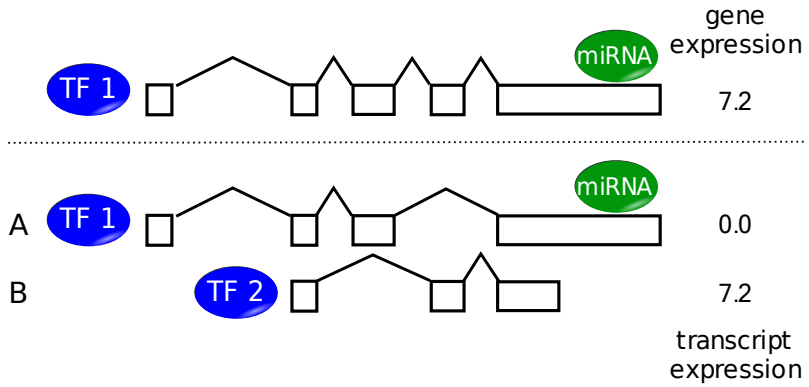

Figure 4.1: A hypothetical gene with 4 exons (boxes) transcribed from left to right with transcripts A and B. If only gene expression measurements exist for the gene and it is known that miRNA is expressed in the condition it seems likely that miRNA has no or no strong influence on the gene expression (top). However, if transcript expression levels are obtained, it can be deduced that the expression of transcript A is effectively shut down by miRNA and/or that TF2 drives the expression of transcript B and therefore gene expression (bottom).

grates the number of reads in exons and the information from transcript annotations. This underlying idea of POEM is shared with previous approaches that were designed for the analysis of splicing arrays combining exon and splice junction probes [159, 3] or EST analysis [169]. However, these approaches have used other assumptions of the underlying model. In here the Poisson framework introduced in Section 3.1.1 is extended for the case of a set of transcripts of a gene.

4.2 Quantification of Transcript Expression Levels

It is assumed that a set of transcripts $S = \{S_1, \ldots, S_k\}$ of a gene is expressed with read counts $T_1, \ldots, T_k$, respectively. The membership of exons $e = 1, \ldots, m$ to a transcript S_j is encoded in the *exon-transcript indicator matrix* $I_{e,j}$ or indicator matrix for short:

$$I_{e,j} = \begin{cases} 1 & \text{, if transcript } j \text{ contains exon } e \\ 0 & \text{, else .} \end{cases} \tag{4.1}$$

See Figure 4.2 for an example with three transcripts.

Naturally, the exon read count Y_e for exon $e = 1, \ldots, m$ is defined as the number of

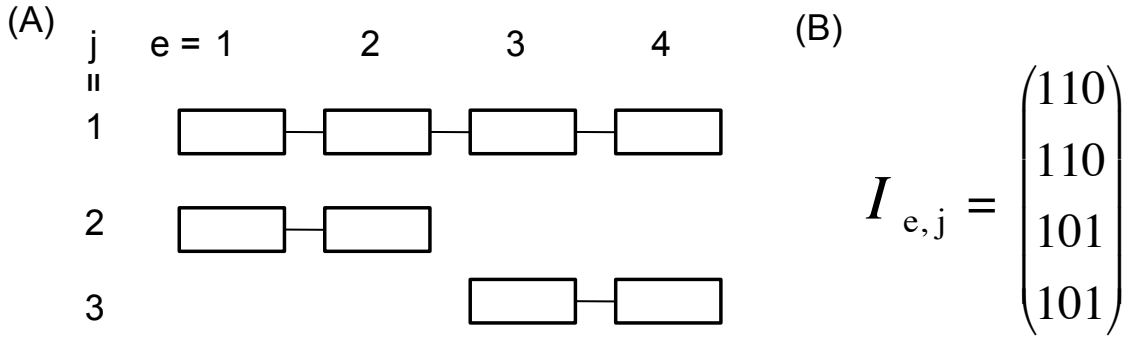

Figure 4.2: A hypothetical gene with three transcripts is depicted in (A). The belonging of exons to the three transcripts is encoded in the exon-transcript indicator matrix $I_{e,j}$ in (B). Note that the matrix $I_{e,j}$ has not full rank because the columns are linearly dependent and therefore the transcripts are not identifiable, see text.

reads that fall within the exon considering each of the k transcripts:

$$Y_e = \sum_{j=1}^{k} \frac{p_e}{\sum_{i=1}^{m} p_i \cdot I_{i,j}} \cdot I_{e,j} \cdot T_j, \tag{4.2}$$

where again p_e is set to the effective exon length l_e in order to accommodate non-unique matches in exonic regions as explained in Section 3.1.1.

Definition 4.2.1. *Given the exon-transcript indicator matrix $I_{e,j}$ and a set of exon read counts Y_e, $e = 1, \ldots, m$, the transcript quantification problem is to infer the transcript expression levels T_j.*

A necessary condition to obtain a unique solution for the transcript quantification problem is the identifiability of the indicator matrix $I_{r,j}$ as shown by Lacroix *et al.* [82].

Lemma 4.2.1. *The transcript quantification problem can be uniquely solved iff the exon-transcript indicator matrix $I_{e,j}$ has full rank.*

Consider Figure 4.2A for an example where the problem has no unique solution, as transcript 1 is a linear combination of transcript 2 and 3. In the application to data in Section 4.2 the expression is only computed for the annotated genes where $I_{e,j}$ has full rank.

There are different ways how to the solve formula (4.2) to estimate the unobserved transcript expression levels T_j. For example, substituting Y_e and T by their expected values, unique mean estimates of the T_j can be obtained by solving a linear system, given that the matrix Y_e is of full rank.

It will be shown that the likelihood function under the Poisson model is easy and an Expectation-Maximization (EM) strategy is suggested for maximizing the likelihood, and to infer the unobserved transcript proportions T_j. The EM formalism is used because it leads to easy recurrences and allows the addition of more complex parameters in the future.

The approach suggested here for solving the transcript quantification problem is to compute the maximum-likelihood solution within the Poisson framework.
Let $q_j = \frac{T_j}{T}$ be the *gene-relative transcript proportion* of transcript j compared to total gene read count T. It follows $\sum_{i=j}^{k} q_j = 1$. Let $q = (q_1, \ldots, q_k)$, be the vector of *gene-relative transcript proportions* for the k transcripts of the gene. The exon read counts $Y_{e,j}$ of isoform j in exon e are fully described by the next two formulas:

$$T_j \;\sim\; \mathcal{P}(\lambda_j) \text{ with } \lambda_j := \; \lambda \cdot \frac{1}{\sum_i p_i \cdot I_{i,j}} \cdot q_j \; . \tag{4.3}$$

Each transcript T_j is distributed according to a Poisson distribution that depends on λ_j as a function of transcript length, relative transcript proportion q_j, and the normalizing factor λ that accounts for other effects, for example sequencing depth. The counts for individual exons in a transcript j are described by a multinomial distribution:

$$P((Y_{1,j}, Y_{2,j}, \ldots, Y_{m,j}) | T_j = t_j) \;\sim\; \mathcal{M}\left(\left(\frac{p_e}{\sum_{i=1}^{m} p_i \cdot I_{i,j}} \cdot I_{e,j}\right)_{e=1}^{m}, t_j\right) \forall j = 1, \ldots, k \; . \tag{4.4}$$

Note that the exons should not overlap, as could be the case with an alternative 5' splicing event for example. In such cases the exon is split into two subexons that do not overlap anymore.

We denote as $y_{e,j}$ the observed read count of the isoform j in exon e and t_j as the total observed read count of isoform j.

Likelihood

From formula (4.3) and (4.4) the complete likelihood of the data can be written as:

$$P(y_{1,1},\ldots,y_{m,k}) = \prod_{j=1}^{k}\left(\frac{\exp^{-\lambda_j}\lambda_j^{t_j}}{t_j!}\cdot\binom{t_j}{y_{1,j},\ldots,y_{n,j}}\cdot\prod_{e=1}^{m}\left(\frac{p_e}{\sum_{i=1}^{m}p_i\cdot I_{i,j}}\cdot I_{e,j}\right)^{y_{e,j}}\right), \tag{4.5}$$

The complete likelihood incorporates the transcript read count T_j which needs to be estimated given only the marginal read counts on the exons

$$Y_e \;=\; \sum_{j=1}^{k} Y_e^j \;. \tag{4.6}$$

The task is to estimate λ and q_j given, $I_{e,j}$, y_i, and p_i, in order to get to the desired transcript read counts T_j. The likelihood of exon read count Y_e under the Poisson distribution is:

$$\mathcal{L}(Y_e|\mu_e) = P(Y_e = y_e|\mu_e) = \frac{e^{-\mu_e}\mu_e^{y_e}}{\mu_e!} \text{ with } \mu_e := \;\lambda\cdot\frac{1}{p_e}\cdot\sum_{j=1}^{k} q_j\cdot I_{e,j}\;. \tag{4.7}$$

Let $\mu = (\mu_1,\ldots,\mu_m)$ and assuming that all the exons are independent, the log-likelihood for all exons becomes:

$$\begin{aligned}
\log(\mathcal{L}(Y_1,\ldots,Y_m|\mu)) &= \log(P(Y_1 = y_1,\ldots,Y_m = y_m|\mu)) && (4.8)\\
&= \log\left(\prod_{e=1}^{m}\frac{e^{-\mu_e}\mu_e^{y_e}}{\mu_e!}\right) && (4.9)\\
&= \sum_{e=1}^{m}\log\left(\frac{e^{-\mu_e}\mu_e^{y_e}}{\mu_e!}\right) && (4.10)\\
&= \sum_{e=1}^{m}\log\left(e^{-\mu_e}\mu_e^{y_e}\right) - \sum_{e=1}^{m}\log(\mu_e!) && (4.11)\\
&= -\sum_{e=1}^{m}\mu_e + \sum_{e=1}^{m} y_e\log\mu_e - \sum_{e=1}^{m}\log(\mu_e!)\;. && (4.12)
\end{aligned}$$

In order to maximize the complete likelihood the iterative Expectation-Maximization (EM) procedure is used [42]. In the E-step the expected transcript read count $\hat{t_j}^{(\upsilon+1)}$ is computed from the previously estimated $\hat{q}_j^{(\upsilon)}$ values that have been maximized in the M-step.

E-step

Assuming current parameters are known, the a posteriori count of isoform j at step υ can be written as:

$$\hat{t}_j^{(\upsilon+1)} = \mathbb{E}_{q^{(\upsilon)}}(T_j|y_1,\ldots,y_m) = \sum_{e=1}^{m} \frac{p_e q_j^{(\upsilon)} I_{e,j}}{\sum_{l=1}^{k} p_e q_l^{(\upsilon)} I_{e,l}} \cdot y_e \ . \tag{4.13}$$

M-step

The following estimator is obtained by maximizing the complete likelihood conditionally on $\hat{t}_j$:

$$\hat{q}_j^{(\upsilon+1)} \quad = \quad \sum_{e=1}^{m} p_e I_{e,j} \cdot \frac{\hat{t}_j^{(\upsilon)}}{T} \ . \tag{4.14}$$

The method is initialized from random estimates and convergence is assumed when the relative increase of the log-likelihood is lower than a threshold value ε. For all experiments $\varepsilon = 10^{-6}$ was fixed. Note that due to the linearity of the likelihood function the EM procedure will always converge to the global maximum.

Note that although the definitions are made for exonic regions, there is no theoretical constraint to include exon-exon junctions as a Y_e. The junctions are by definition not overlapping so the independence is given.

Quality score

As transcript annotation for genes is imperfect, it may happen that a transcript that is expressed in the analyzed RNA-Seq dataset is not yet annotated in the database. In this scenario, the POEM method produces the maximum likelihood solution for

the wrong annotation. Thus, it would be good to have a quality measure of the estimated proportions $(\hat{q}_1, \ldots, \hat{q}_k)$ for transcripts $1, \ldots, k$ that can pinpoint low quality estimations. A test statistic is defined that assesses if the observed and expected counts are significantly different (according to the counts observed on the exons):

$$\chi^2_G = \sum_{e=1}^{m} \frac{(y_e - Y_e^{exp})^2}{Y_e^{exp}} \ , \tag{4.15}$$

where Y_e^{exp} is the expected count of exon e proportionally to the expression of the full transcript. χ^2_G follows a χ^2 distribution with $(n-1)$ degrees of freedom. The quality score is computed for each gene as the $\log_{10}$ of the p-value.

Transcript Database Construction

Annotation of human transcript structures for POEM (i.e. the indicator matrix I) was derived from Ensembl (version 46)[47]. All protein coding transcripts have been downloaded. As mentioned earlier, in order to allow description of any possible isoform (for instance, alternative 5' and alternative 3' sites), exons overlapping with different boundaries across isoforms were further subdivided. Redundant transcripts were filtered out. To this end, two transcripts were recursively clustered when the sequence identity, relative to the mean length of both transcripts, was at least 95%. A representative of each cluster was chosen by taking the union of the corresponding rows in I. From initial 42,635 transcripts in Ensembl, 37,177 remained after clustering and filtering. As stated in lemma 4.2 a necessary condition to compute the transcript expression levels is the independence of all considered transcripts. Only genes with an indicator matrix I of full rank have been used for quantification.

4.2.1 Simulations

In order to assess the accuracy of POEM for different values of transcript expression levels and alternative splicing complexity, two simulations were conducted. Two different error measure were considered. The average error for a set of x simulations is defined as:

$$\text{average}(\hat{q}, q) \ = \ \frac{\sum_{o=1}^{x} |\hat{q}^o - q^o|}{x} \ , \tag{4.16}$$

where $\hat{q}_j^o$ denotes the o-th estimated proportion for a transcript and q^o the correct proportion. Similarly, the max error is defined as:

$$max(\hat{q}, q) \;=\; \max_{o=1,\ldots,x} |\hat{q}^o - q^o| \;. \tag{4.17}$$

First, simulations were performed on a template gene of 1,200 bp and considering one exon-skipping event. The length of the skipped exon (120, 240, and 360 bp) was varied, as well as the exon inclusion rate (20 and 80%) and the gene expression level (100-600 reads) to assess their impact on the estimation error rate. Two thousand simulations were performed for each combination of the parameters. In Figure 4.3A it can be seen that the average error on proportion estimations decreased with gene expression level, as expected. Furthermore, the error of estimation was inversely correlated with the length of the skipped exon (Fig. 4.3A, grey to dark lines). The exon inclusion rate also had an influence on the error, as a 20% inclusion rate (Fig. 4.3A, dashed lines) had constantly higher estimation error than an 80% inclusion rate (Fig. 4.3A, plain lines). A minimum of 300 reads in the gene achieves a reasonable accuracy for POEM.

Secondly, the expected global accuracy of POEM was addressed with a second simulation on all annotated transcripts (Ensembl v.46). The estimation error made was monitored and plotted as a function of gene expression (Fig. 4.3). The number of expressed isoforms was fixed to two, and the relative proportions were incremented in steps of equal size from 16.7% to 83.3% (10,000 runs for each combination). The sampling of the transcripts was done as follows. First, a gene was chosen uniformly among all genes annotated with more than two isoforms in Ensembl. Then, two transcripts were uniformly sampled among the annotated isoforms of the gene. The 90% quantile of errors show that, with a minimum of 300 reads within the gene (Fig. 4.3B, vertical line), the average error is $<$12.6% (maximum error is $<$18.6%).

4.2.2 Proportion Estimation with Junction Reads

For comparison, reads mapping to splice junctions were used to directly quantify AEEs by computing the proportion of reads mapping to the constitutive junction. As a constitutive exon is adjacent to two splice junctions, the proportions deduced from both splice junction read counts identifying the same AEE was averaged. An

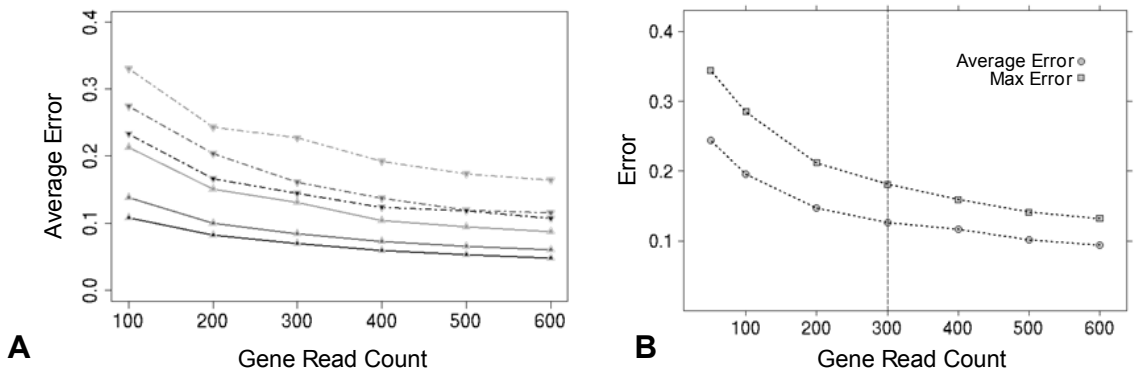

Figure 4.3: POEM simulations:(A) Plot showing the 90% quantile of the average error for proportion estimation by POEM based on simulations for one gene with one exon-skipping event. The average error (y-axis) is calculated for simulations with different total read count in the gene (x-axis) and for various skipped exon lengths: 120 bp (light grey), 240 bp (grey), or 360 bp (black). The average error is shown for a proportion of 20% (dashed lines) and 80% (plain line). (B) This plot shows the 90% quantile of the average (circles) and maximum (squares) error (y-axis) for POEM predictions on all human Ensembl (v.46) transcripts.

illustrative example is shown in Figure 4.4C, where the inclusion rate of exon 5 (0.84) was calculated as the average of (i) counts on junctions 4–5 and 4–6 [8 / (8 + 2) = 0.8] and (ii) counts on junctions 5–6 and 4–6 [14 / (14 + 2) = 0.87].

As stated above the exon-exon junction read counts can also be included in the EM-formalism, but this was not investigated due to the short length of the reads.

4.3 Application to Human RNA-Seq Data

In what follows the POEM method is applied to the data of the human HEK and B cell lines presented in Chapter 3. POEM was applied to all genes in both dataset with at least 300 reads mapping inside the exons of the gene, as suggested by the simulations above. Only genes with at least two isoforms indicative of alternative splicing in internal exons were considered. The CASI analysis showed that large modifications are occurring on the most 3'- or 5'-exons (Subsection 3.3.1). Therefore, POEM estimation focused on information from internal exons, by artificially removing the first and last exon of every transcript before POEM estimation. In this, the relative isoform proportions for 830 and 640 genes in HEK and B cells was estimated, which were annotated with 2,412 and 1,911 transcript variants, respectively (Supplemental Table

S4). From this set, POEM estimated proportions for 1,920 and 1,487 transcripts for HEK and B cells, respectively. These estimations were verified by (i) analysing the number of reads mapping to exon-exon junctions and (ii) experimental validations using qPCR. For instance, Figure 4.4C shows that the inclusion rate of exon 5 in the gene *MPI* could be deduced from the ratio of exon-exon junctions reads, estimated as 84%. The POEM predictions were compared with the information provided by junction reads for 267 constitutive AEEs (149 pairs in HEK and 118 pairs in B cells), showing at least three exon-exon junction read counts. On the whole, POEM agreed well with the estimates deduced from junction read counts, with a correlation coefficient of 0.65 and an estimated proportion difference of $<$20% for 80% of the events (Figure 4.4A).

4.3.1 Experimental Validation

Estimates based on the junction read counts alone are not sufficient, as they were taken from the same RNA-Seq dataset that could probably be biased. In order to get an independent estimate of the isoform proportion for some transcripts, PCR primers have been designed to recognize skipped exon events, similar to Section 3.3.1 The comparison of POEM estimates with qPCR measurements for a total of 47 AEEs in both cell lines (22 exon-skipping events, two mutually exclusive events, Supplemental TableS5) shows a high correlation (PCC = 0.81, Fig. 4.4B). For qPCR and POEM data comparison, POEM estimates were derived for the skipped and constitutive isoforms only, irrespective of other transcripts annotated in Ensembl. The gene *MPI* is an illustrative example (Fig. 4.4C), which is also confirmed by junction reads. Precise inference of a large difference in relative expression levels is hampered if an isoform has a very low expression value. This is illustrated in Figure 4.4B, where 13 events (with qPCR AEE proportion close to 0% or 100%) display an expression level difference of 2-3 orders of magnitude between the constitutive and the skipped isoform.

It is worth mentioning that 38 out of the 47 tested AEEs were supported by junction reads. The comparisons of the estimated proportions derived from junction reads with the estimates from qPCR for these 38 AEEs showed a slightly lower correlation (PCC = 0.74). This is due to the paucity of reads identifying junctions, reducing the significance of ratios associated with low read counts for estimating AEEs. Besides, with

twice 8 million reads sequenced, the junction read depth is still far from saturation, so it is expect to see at most 50% of the expressed junctions. Therefore, exploiting the number of counts in exons offers complementary information in detecting and quantifying AEEs, in particular when the dataset does not reach saturation.

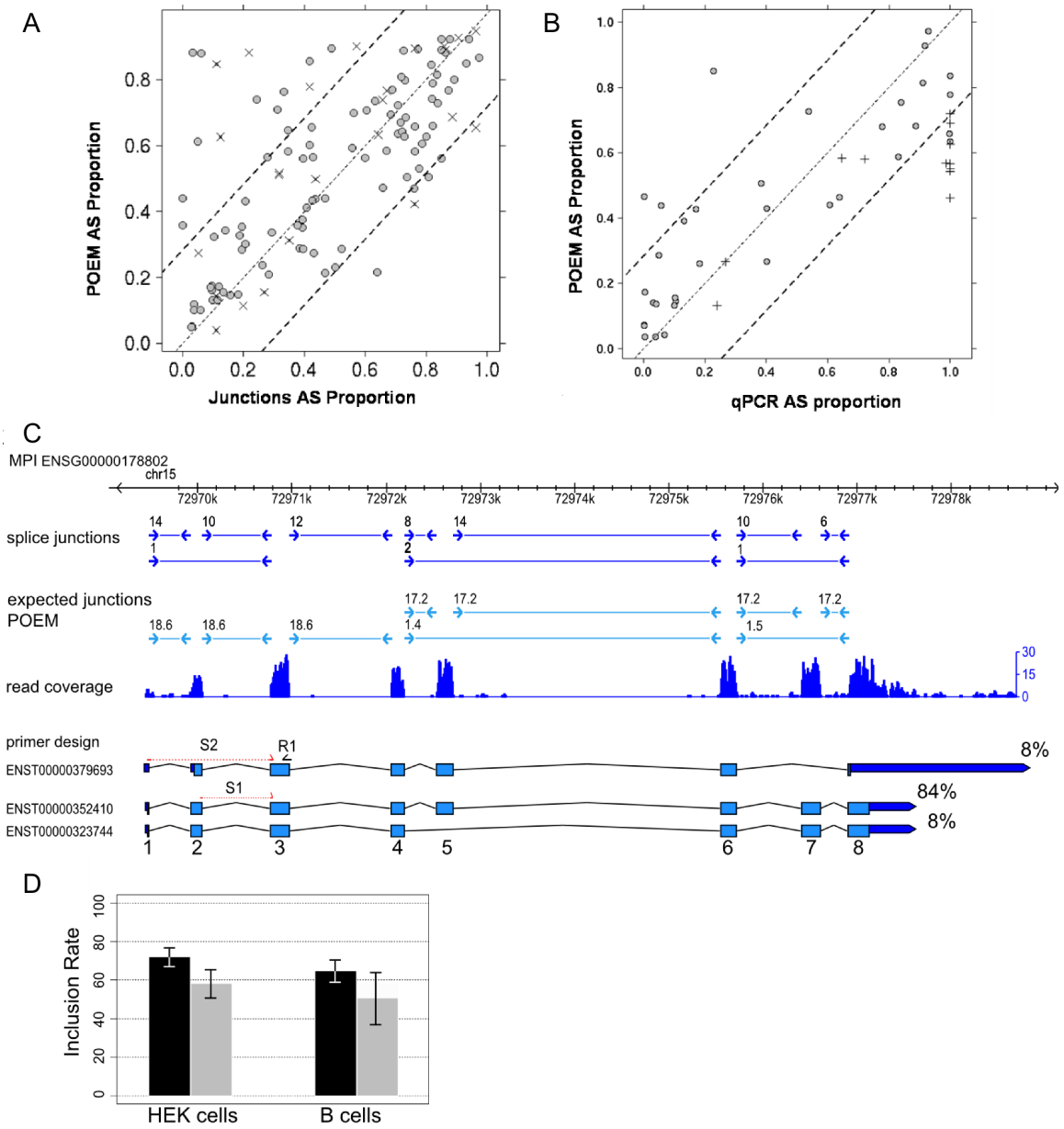

Figure 4.4: POEM validations: (A) Scatter plot of inclusion rates (constitutive isoforms) on 123 AEEs derived from exon-exon junction counts (x-axis) and POEM estimations (y-axis) is shown (PCC = 0.65). Cross marks denote AEEs in genes with a quality score $\leq$ -14. Dashed lines represent the 20% error margin and the dotted middle line is just for orientation in (A) and (B). (B) Scatter plot of inclusion rates on 47 AEEs measured by qPCR (x-axis) and estimated by POEM for a single exon-skipping event (y-axis) is shown (PCC = 0.81). Plus marks denote unannotated AEEs in Ensembl v.46. (C) POEM estimation for annotated transcripts of *MPI* in HEK cells. Numbers reported on light blue arrows represent the expected counts on exon-exon junctions according to the estimated proportions with POEM for the three annotated isoforms (ENST000000379693, ENST000000352410, and ENST000000323744). The proportion estimate for each isoform is shown to the right (in %). qPCR primers were designed to estimate the inclusion rate of exon 2. The skipping event of exon 3 was not annotated in Ensembl v.46, but was supported by an observed junction read. (D) The bar chart shows the inclusion rate of exon 2 computed by POEM (grey) and measured by qPCR (black) for HEK and B cells.

Chapter 5

De Novo Assembly of Transcripts considering Alternative Isoforms

5.1 De Novo Assembly of Transcript Sequences

Instead of using annotation to identify alternative exon events (AEEs) or quantify the expression levels of isoforms as presented in the previous chapters, the task is to assemble transcript sequences directly from RNA-Seq data as input. The aim is to reconstruct full length transcripts from short read data *de novo*, i.e., without the use of a genomic reference sequence. Such a method will be necessary to allow the tanscriptome analysis of genomes for which a genomic reference sequence does not exist. Depending on the application it might become feasible to avoid sequencing of the reference genome alltogether. Except for this obvious argument, there are other practical arguments that vote for the *de novo* approach to transcriptome assembly.

There are a number of cases in biology where the existence of a reference genome of one individual of a species does not easily allow the retrieval of expressed transcript sequences:

1. In the transcriptomes of cancer cells one important issue is to identify **fusion transcripts** of genes that have undergone rearrangements in the cancer genome [101, 102, 91]. These fusion transcripts can be the result of a chromosome translocation, fusing two genes from different chromosomes. With short single-end reads such an analysis requires a database of known gene fusions [101]. Paired-end or long reads allow improved detection of fusion transcripts if

alignment algorithms exist that allow mapping on different chromosomes [102]. An additional complication for the detection of these cases could be the incidence of alternative splicing in fusion genes, something which could not be investigated so far due to a limit in resolution.

2. Organisms like the flatworm *Schistosoma mansoni*, the plant *Arabidopsis thaliana*, and also humans have genes that contain stretches of so called **micro-exons**, exons of length ≤ 25 bp [10, 57]. These micro-exons are among the most difficult problems for gene finding and spliced alignment algorithms and specially designed algorithms exist to handle them [156]. In particular short read alignment under consideration of micro-exons is infeasible because of too many random matches in the genome.

3. In parasite transcriptomes like *Trypanosoma brucei* or *Schistosoma mansoni* and few suggested cases in other organisms like humans, a biologically unclear process called **transplicing** fuses the pre-mRNA sequences from two different genes to create a fused mRNA that harbors exons from both genes [128, 6, 63]. On the mRNA level it is similar to the fusion transcripts in cancer cells, although the biological mechanism is different. Transplicing may have unknown regulatory roles which can now be investigated.

In addition, prediction of alternative splicing events is possible directly from the de Bruijn graph over sequences, as shown later. This could lead to a new way of thinking about functional studies in organisms whithout known reference sequence.

However, *de novo* transcriptome assembly is conceptionally more difficult than transcriptome assembly with a reference genome. A few studies reporting results with genomic assemblers applied to RNA-Seq data have been published using Velvet [35, 157], ABySS [11], and a parallel assembler by Jackson et al. [72]. In this chapter the additional difficulties of *de novo* transcriptome assembly are explained. A new set of algorithms for *de novo* transcriptome assembly called Oases is presented. Application to real RNA-Seq data show the improvement compared to genomic *de novo* assemblers and assemblers that use the genome. First the theoretical capacities of de Bruijn graphs built from RNA-Seq reads are investigated.

5.1.1 Problem Statement

Assume a set of m reads $\mathcal{R} = R_1, \ldots, R_m$ each of length r that are the result of an RNA-Seq experiment of a sample. Assume further that a set of transcripts $\mathcal{T} = T_1, \ldots, T_o$ is present in the sample and the set of reads $\mathcal{R}$ is derived from all possible r-mers of these transcripts $R_i \in r\text{-}spectrum(\mathcal{T}), i = 1, \ldots, m$.

From a given $\mathcal{R}$ reconstruct all the transcripts in the sample, what is called the *de novo transcriptome assembly problem*:

Definition 5.1.1. *The* de novo transcriptome assembly problem *is, given a set of reads* $\mathcal{R} = R_1, \ldots, R_m$ *derived from a set of transcripts* $\mathcal{T} = T_1, \ldots, T_o$ *with* $R_i \in$ r-spectrum$(\mathcal{T}), i = 1, \ldots, m$, *reconstruct the set of original transcripts* $\mathcal{T}$.

5.1.2 Transcript de Bruijn Graphs

Transcript de Bruijn graphs (TGs) are de Bruijn graphs constructed from k-mers of transcript sequences. It is assumed that chains of nodes in TGs have been simplified (Section 2.3.3). The TG is related to the simplified splicing graph, where each chain of exon nodes was merged into one large node (see Section 2.4.2). However, sometimes the TG can be far more complex than the simplified splicing graph.

In the work of Heber and co-workers in 2002 [62] it is mentioned for the first time that a splicing graph over EST sequences can be built without EST alignment by constructing a de Bruijn graph over the EST sequences directly. However, they considered a special case of TGs, (i) they build the TG for a cluster of ESTs that mapped to the same genomic locus without contamination of, e.g., ESTs from homologous genes and (ii) the correct orientation of all ESTs was known and thus the use of digraphs was not necessary. Further, they have the genomic position for each EST, and thus for each k-mer of an EST. Therefore, they study alternative splicing in the TG without treatment of induced cycles by repeated k-mers [62]. One way to cope with cycles was described by Malde et al. [103], where running in cycles of the TG was avoided by maintaining a map of genomic positions for the k-mers in ESTs.

TGs as presented here are digraphs and will sometimes be cyclic, thus complicating the analysis. However, under some circumstances a TG is acyclic as shown later. Cycles in TGs are induced by *repeated* k-mers and their reverse complement sequences.

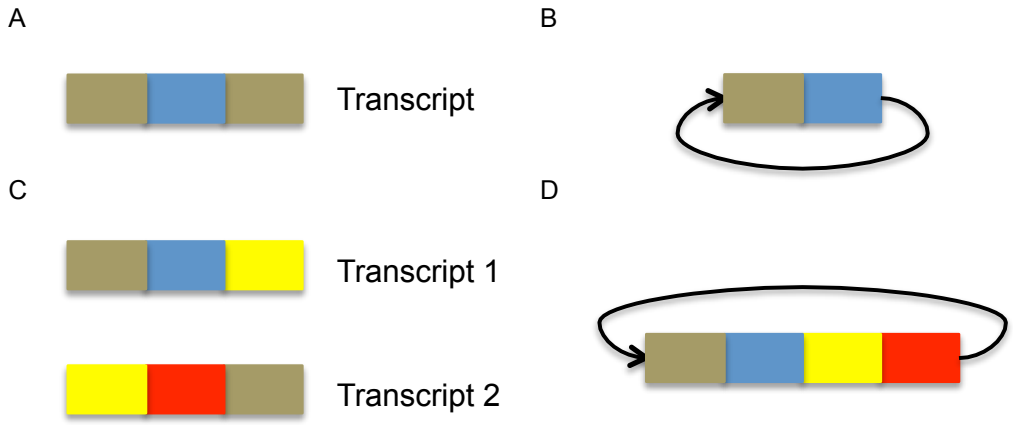

Figure 5.1: Cycles in Transcript de Bruijn Graphs. A) Depicted is a transcript with a repeat (brown block). B) The cyclic de Bruijn graph for transcript in A. C) Depicted are two transcripts that have the same repeat (brown and yellow blocks) but in a different order. D) The cyclic de Bruijn graph for transcripts in C.

Cycles in Transcript de Bruijn Graphs

Definition 5.1.2. *A k-mer ω is said to be* repeated *in a transcript T_i, if $occ_i(\omega) + occ_i(\overleftarrow{\omega}) \geq 2$.*

That means k-mer ω occurs at least twice or together with its reverse complementary k-mer $\overleftarrow{\omega}$ in T_i. The next observation describes two main mechanisms that create cycles in TGs as observed in practice.

Observation 5.1.1. *A Transcript de Bruijn graph $\mathcal{TG}$ of dimension k over a set of transcript sequences $\mathcal{T} = \{T_1, \ldots, T_x\}$ is cyclic if one of the following conditions is true:*

(i) $\exists\ i \in [1, x]_{\mathbb{N}}, \ldots, x, s.t.\ T_i$ *has a repeated* k-mer,

(ii) $\exists\ i, j \in [1, x]_{\mathbb{N}},\ k\text{-spectrum}(T_i, \overleftarrow{T_i}) \cap k\text{-spectrum}(T_j, \overleftarrow{T_j}) = \{\omega, v\}$, *s.t.* $pos_{T_i}(\omega) < pos_{T_i}(v)$ *and* $pos_{T_j}(v) < pos_{T_j}(\omega)$

Proof. In the first condition a k-mer is repeated in a transcript thus inducing a loop in $\mathcal{TG}$, see Fig. 5.1A+B. The second scenario is a special case for two transcripts, where two k-mers ω and v are not repeated in T_i or T_j but their order of occurrence is reversed in both transcripts, again forming a loop in $\mathcal{TG}$, see Fig. 5.1C+D. □

Condition (ii) could in fact be generalized to more than two transcript in such a way that a series of overlapping k-mers, contributed by a number of transcripts, creates a path in $\mathcal{TG}$ starting from and ending with the same k-mer.

It should be mentioned here, that the solution to the *de novo* transcriptome assembly problem using transcript de Bruijn graphs is found by solving the De Bruijn Graph Superwalk problem, which was shown to be ***NP*-hard** for $|\Sigma| \geq 3$ and any positive integer k [108] (see Section 2.3.2). Therefore no polynomial time algorithm for transcript de Bruijn graphs can be designed to find an exact solution for the DNA alphabet.

Nevertheless, acyclic TGs are easy to work with and allow the detection of alternative exon events and the reconstruction of underlying transcripts, see Sections 5.1.3 and 5.2.4 respectively.

Construction of Splicing Graphs from Transcript de Bruijn Graphs

Before it is shown how to construct splicing graphs from Transcript de Bruijn graphs a few necessary definitions are made for a set of transcripts that share a path in a TG. Let G be a gene, G_1 and G_2 two different genes, and if a transcript T_i belongs to a gene G it is denoted $T_i \in G$. Recall that a gene has *constitutive* exons which are not involved in any AEE in the gene, whereas *alternative* exons are involved in at least one AEE. Denote as $const_G(T_i)$ the set of k-mers that are substrings of T_i and that belong to constitutive exons of a gene G. Denote as $alt_G(T_i)$ the set of k-mers that are substrings of T_i and that belong to or overlap with alternative exons of G.

Definition 5.1.3. *Two transcripts $T_i, T_j \in G$ are called* k-alternative*, iff their constitutive exons share a word of length k, $0 < k \leq \min(|T_i|, |T_j|)$ and their alternative exons do not share a word of length k , i.e., $const_G(T_i) \cap const_G(T_j) \neq \emptyset$ and $alt_G(T_i) \cap alt_G(T_j) = \emptyset$.*

This definition is a way of phrasing alternative splicing in k-mer space. Observe that this relation must not hold for arbitrary k for each pair of alternatively spliced transcripts $T_i, T_j \in \mathcal{G}$. For example, if T_i and T_j differ by an alternative last exon, both are not k-alternative for $k = \min(|T_i|, |T_j|)$, namely the size of the smaller transcript. The additional constraint that alternative exons share no k-mer is needed to ensure that no loop in the graph can be created that does not represent an alternative exon event later.

As no genomic positions of the transcripts are known, two transcripts from different genes may share a k-mer as well.

Definition 5.1.4. *Two transcripts $T_i \in G_1, T_j \in G_2$ are called* k-homologous*, if they share a word of length k, $k \leq \min(|T_i|, |T_j|)$, i.e., k-spectrum($T_i$, $\overleftarrow{T_i}$) $\cap$ k-spectrum(T_j, $\overleftarrow{T_j}$) $\neq \emptyset$.*

Note that both the k-spectrum of the transcript and its reverse complement are considered, as the orientation of reads from most of the RNA-Seq experiments today is unknown (see Fig. 1.4). k-homologous transcripts can be either transcripts from genes that share a repeat region or from transcripts of overlapping genes on different strands of the DNA. If k is clear from the context, it will be referred to as alternative and homologous instead of k-alternative and k-homologous.

Obviously the smaller k the higher the chance that two transcripts are k-homologous. That is why one ideally would like to work on the maximum k-spectrum (with $k =$ read length r) to solve the *de novo* transcriptome assembly problem, but in practice such an approach is very prone to sequencing errors as discussed in Section 2.3.3.

A *locus* $\mathcal{L} = \{T_1, \ldots, T_x\}$ is introduced to describe a set of x transcripts. A slight abuse of the word locus is made, defined in biology for the specific location of a gene or DNA sequence, to underline the basic idea of the algorithm later that the transcripts are related in k-mer space. Analogously to the above definitions. A locus $\mathcal{L}$ is called *k-homologous* if at least one pair of transcripts in $\mathcal{L}$ is k-homologous.

Definition 5.1.5. *A locus $\mathcal{L} = \{T_1, \ldots, T_x\} \subseteq G$ is called k-alternative iff*
(i) $\forall i \in [1, x]_{\mathbb{N}}$ $\exists j \in [1, x]_{\mathbb{N}}$ with $i \neq j$, s.t. T_i and T_j are k-alternative, and
(ii) $\forall i, j \in [1, x]_{\mathbb{N}}$ with $i \neq j$, s.t. $alt_G(T_i) \cap alt_G(T_j) = \emptyset$.

Note here that for a locus $\mathcal{L}$ it is enforced that no pair of transcripts shares k-mers in their alternative exons (condition (ii)). The next lemma describes the basic similarity between TGs and splicing graphs.

Lemma 5.1.2. *A simplified transcript de Bruijn graph $\mathcal{TG}$ of dimension k over a locus $\mathcal{L} = \{T_1, \ldots, T_x\}$ can be transformed to a simplified splicing graph over $\mathcal{L}$ iff $\mathcal{TG}$ is acyclic and $\mathcal{L}$ is k-alternative.*

Proof. A constructive proof is given. Each node of $\mathcal{TG}$ is mapped to the genome, assuming error-free spliced alignment, and each node is labeled with its genomic sequence positions. It is ensured that nodes of $\mathcal{TG}$ map to the same gene as $\mathcal{L}$ is

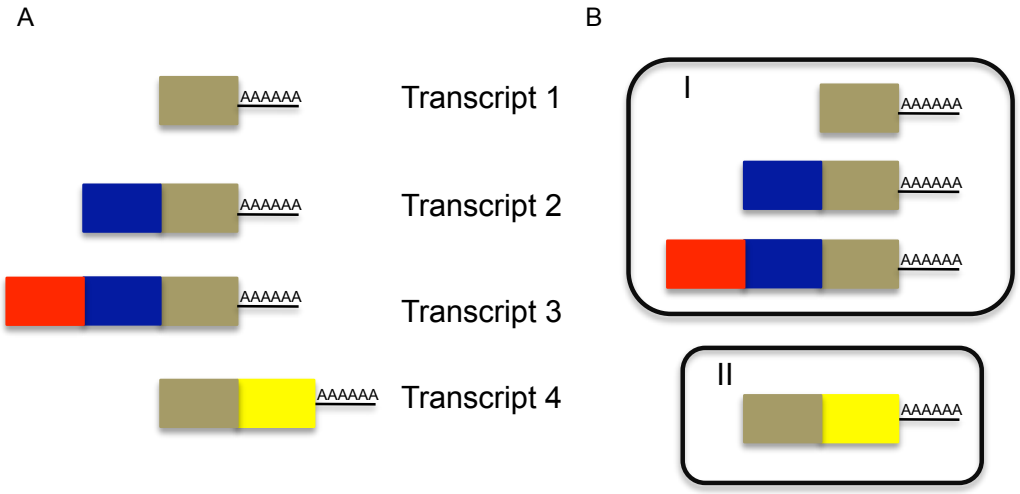

Figure 5.2: A) A gene locus with 4 transcripts is depicted. The colored boxes demarcate exonic regions of the gene with poly-A tail indicated at the 3'-end. B) The four transcripts can be grouped by ambiguity, indicated by ellipse I and II. Transcripts 1-3 differ by alternative promoter events and are ambiguous in the de Bruijn graph. For all transcripts in ellipse I, Transcript 3 would be reported. Transcript 4 can be differentiated from the others due to the poly-A tail that creates an alternate path in the de Bruijn graph.

k-alternative. Further each node must map to a unique genomic position as $\mathcal{TG}$ is acyclic and no alternative exons of the transcripts in $\mathcal{L}$ share k-mers. Small nodes in $\mathcal{TG}$ that only recognize overlaps of two alternatively spliced exons are removed and the two adjacent nodes in $\mathcal{TG}$ are connected by a new edge. Edges between nodes in $\mathcal{TG}$ that describe alternative splicing events can be slightly offset, due to similarity in overlapping k-mers at the junction, which can be corrected with the overlayed sequence position information. □

Note that it might be possible to transform a TG built over a locus that is k-homologous and acyclic into a set of splicing graphs, one for each gene.

Ambiguity of Alternative Promoters and Alternative Polyadenylation Sites in Transcript de Bruijn Graphs

It is impossible to resolve alternative transcription start sites (TSSs) in a gene directly from the TG topology as the transcripts do not have a special signal or sequence at their 5'-end. If a TG over a set of transcripts is built which differ by alternative TSSs, the start point of the shorter form is lost in TGs and the longest form will be

reported, see Fig. 5.2, i.e., the reported form is *maxmimal by inclusion*. With full-length transcripts, artificial start nodes in the TG can be introduced that induce a split in the graph and indicate an alternative start, as is done for splicing graphs [133], see Section 2.4.2. Unless the RNA-Seq protocol is adapted and artificial linkers at the 5'-ends of transcripts are attached before sequencing, the identification of alternative start sites has to be delayed to a postprocessing step of the assembly, for example using variable read coverage as discussed later.

In contrast, the detection of alternative polyadenylation sites (APSs) is feasible given that the mRNAs are sequenced with their poly-A tail attached. In RNA-Seq data each alternative end in an expressed isoform contains reads that span the transition from the last exon into the *poly-A node* in a TG, a highly repeated k-mer containing mostly A nucleotides. In the TG formalism a split at the end of two transcripts with an APSs is induced and can be reported, see Fig. 5.2.

5.1.3 Recognition of Alternative Exon Events in Transcript de Bruijn Graphs

In previous works splicing graphs have been computed after aligning transcript sequences to the genome. Therefore, splice sites in intronic regions of the corresponding gene and ordering of the exons can be used to infer alternative exon events (see Section 2.4.2). Many different algorithms for splicing graphs have been proposed [62, 45, 134, 16]. Because TGs are built *de novo*, the use of sequence information from intronic regions in the genome is impossible. Nevertheless, under some assumptions it can be shown that alternative isoforms in simplified TGs can be classified into distinct alternative exon events given information of the organism's 3' and 5' splice sites. It is assumed that chains of nodes in a TG have been simplified.

Nodes v in a TG that have an $\text{indeg}(v) > 1$ or $\text{outdeg}(v) > 1$ maybe a witness of an alternative exon event. The simple alternative exon events form *bubbles* in the TG. This definition is on purpose restricted to small cycles and not as general as defined by Sammeth [133] for splicing graphs, because only simple events in TGs are considered.

Definition 5.1.6. *Given an acyclic transcript de Bruijn graph $\mathcal{TG}$ over a locus $\mathcal{L}$ that is k-alternative, a cycle of size three or four in the underlying subgraph of $\mathcal{TG}$ is*

called a bubble.

Lemma 5.1.3. *A bubble in a transcript de Bruijn graph $\mathcal{TG}$ of dimension k over a locus $\mathcal{L} = \{T_1, \ldots, T_x\}$ is induced by an alternative exon event between two transcripts T_i and $T_j \in \mathcal{L}$.*

Proof. As shown in Lemma 5.1.2, $\mathcal{TG}$ can be transformed into a splicing graph, because $\mathcal{TG}$ is acyclic and $\mathcal{L}$ is k-alternative. In a splicing graph every bubble induced by two transcripts is caused by an alternative exon event, see Lemma 2.4.1. □

After establishing that simple alternative exon events are to be found in bubbles in TGs, the next step is to show that there are distinct topological attributes and/or sequence motifs associated with the six most abundant alternative exon events: skipped exon, alternative 5', and 3' site, intron retention, alternative polyadenylation, and mutually exclusive exons. These events are by far the most common alternative exon events among higher eukaryotes, although with different species-specific frequencies [134]. These simple events can be considered in the context of two alternatively spliced isoforms denoted as pairwise detection of alternative splicing [134].

In order to make a distinction between the six events, it is useful to differentiate between *source*, *sink*, and *alternative* nodes in a bubble.

Definition 5.1.7. *In a bubble a node n is called source node, if indeg(n)=0 and outdeg(n)=2 .*

Definition 5.1.8. *In a bubble a node n is called sink node, if indeg(n)=2 and outdeg(n)=0.*

Definition 5.1.9. *In a bubble a node n is called alternative node, if indeg(n)=outdeg(n)=1.*

For example consider the bubble in Fig. 5.3 (on the left). Exon A is the source node, exon D is the sink node, and exons B and C are the alternative nodes.

There is a special case of an alternative node in a bubble which is caused by the sequence that describes the overlap between two exons or an exon with the poly-A node only, the *junction node.*

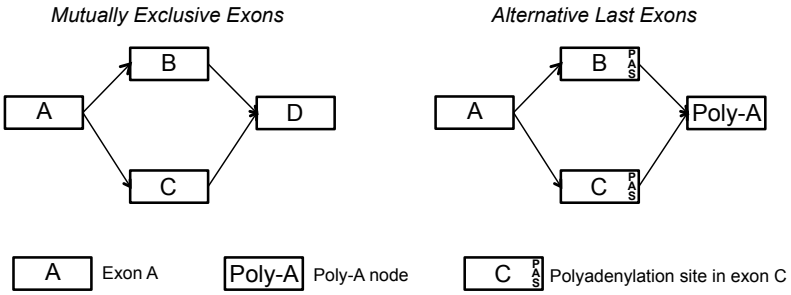

Figure 5.3: Maximum Parsimony Alternative Exon Events in Bubbles without a Splice Junction Node. For alternative last exons both alternative nodes contain a polyadenylation site (PAS) at their 3'-end and the sink node is derived from the poly-A tail of the transcript (right). If both alternative exons are devoid of a PAS the bubble describes a mutually exclusive exon event (left).

Definition 5.1.10. *In a bubble an alternative node n is called a* junction node*, if it recognizes the splice junction of two exons or the junction between an exon and the poly-A.*

It holds that $0 \leq l_n \leq k$-1, for a junction node n. It is straightforward to see that the maximum length of n is k-1, the number of k-mers that describe the overlapping sequence between the two exons. On the other hand, in the worst case none of the k-1 k-mers is unique compared to the other alternative node in the bubble and thus node n is missing. This is the only case when a bubble has size 3 instead of 4.

Before the topology and sequence motifs in bubbles are analyzed a number of assumptions are made. First the principle of maximum parsimony is applied, which means the alternative exon event with the least involved alternative splicing reactions is preferred. For example, if one and two skipped exons are possible events, one exon skipping is predicted, as skipping two exons would require more splicing factors to bind the pre-mRNA. The second assumption is that the orientation of the nodes in the bubble is known. Lastly, it is assumed that splice sites and polyadenylation signals (PAS) can be predicted with 100% accuracy.

If there is no splice junction node in a bubble, there must be a mutually exclusive exon or an alternative last exon event as the next observation shows.

Observation 5.1.4. *Consider a bubble of size 4 in a transcript de Bruijn graph $\mathcal{TG}$ that has no splice junction node. It follows that the alternative exon sequences a and b contribute unique sequence. The most parsimonious alternative exon event underlying*

the bubble is either

(*i*) *an alternative last exon event if a and b contain a PAS, or otherwise*

(*ii*) *it is a mutually exclusive exon event.*

Proof. The only two possible alternative exon events that form a bubble and where both alternative exons have unique sequence with respect to each other are two alternative last exons or two mutually exclusive exons (see Fig. 1.2). The only difference between both events is that alternative last exons contain PASs that are recognized by factors initiating polyadenylation. If no PASs are present the exons must be mutually exclusive, see Fig. 5.3. □

Note that the above observation could be extended for the detection of alternative promoters, if a unique sequence tag would be added to the 5'-ends of transcripts before sequencing as mentioned in Section 5.1.2.

A direct consequence of Definition 5.1.10 is that there must be at least one other alternative node in the bubble that contributes unique exon sequence, which leads to the following observation.

Observation 5.1.5. *Consider a bubble in a transcript de Bruijn graph $\mathcal{TG}$ that is either of size 4 and has a splice junction node or has size 3. Let n be the alternative node that is not a splice junction node. The most parsimonious alternative exon event underlying the bubble is either*

(*i*) *an alternative polyadenylation event if n and the sink node contain a PAS, or*

(*ii*) *an intron retention event if n contains a 5' and a 3' splice site, or*

(*iii*) *an alternative acceptor event if n contains only a 3' splice site, or*

(*iv*) *an alternative donor event if n contains only 5' splice site, or otherwise*

(*v*) *it is a skipped exon exon event.*

Proof. (i) Compared to observation 5.1.4, an alternative polyadenylation event is the shortening of the 3' UTR in an exon and therefore no unique sequence is present. (ii) If both a 5' and a 3' splice sites are present in n, the most parsimonious event is an intron retention event, where the splicing reactions of constitutive splice sites are prevented. (iii,iv) If either a 5' or a 3' splice site is present in n the most parsimonious explanation is an alternative acceptor or alternative donor event, respectively. (v) Finally, if no splice site is present in n, it must be a skipped exon event. □

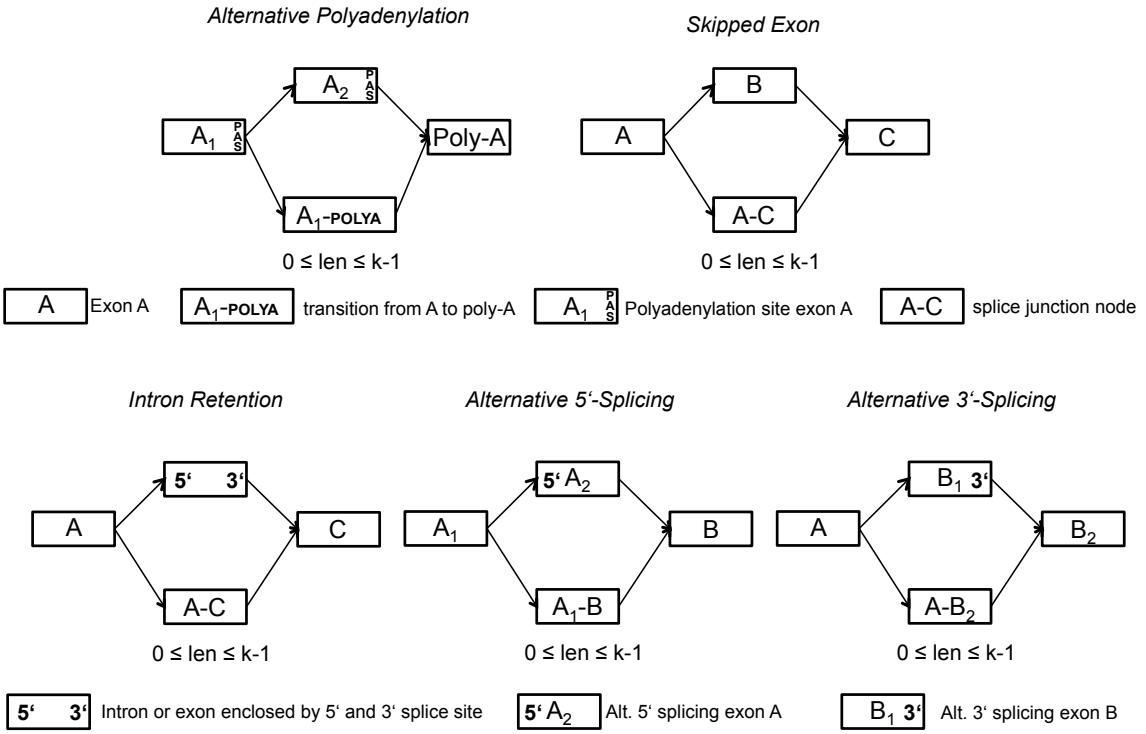

Figure 5.4: Maximum Parsimony Alternative Exon Events in Bubbles with a Splice Junction Node. The occurrence of splice sites or polyadenylations site (PAS) lead to the different depicted alternative exon events, see Observation 5.1.5.

Note that if the assumption of maximum parsimony is dropped, alternative 5' and 3' splicing might be coupled to one or more exon skipping events. Similarly, an intron retention event might be a coupled alternative 5' and 3' splicing event instead.

5.2 Oases: a *de novo* Transcriptome Assembler Based on Transcript de Bruijn Graphs

After the theoretical capabilities of Transcript de Bruijn graphs have been presented in the previous section, a new set of algorithms is explained in this section. These algorithms are based on the graph structure of the Velvet package and are collectively called Oases. The two executables of velvet, *velveth* and *velvetg*, perform the first steps, namely hashing into k-mers, de Bruijn graph construction, simplification ,and finally error correction (Section 2.3.3).

5.2.1 Error Correction and Collapsing

An essential step of any short-read assembler is error correction. RNA-Seq short-read data are exposed to the same errors as resequencing data and therefore these steps of Velvet are not altered. Tip clipping and the Tour Bus algorithm are needed to remove sequencing errors, small polymorphisms due to allele-specific isoforms [112], and substitutions caused by RNA editing [26]. A minimum k-mer coverage cutoff on the remaining nodes is imposed to remove lowly abundant splice forms, which are indistinguishable from sequencing errors (Section 2.3.3). It is assumed that after these corrections no sequencing errors and polymorphisms exist in the node sequences of the graph.

Depending on the organism these steps have to be used with caution. For example, micro-exons [156] form small bubbles that might be removed by the Tour Bus algorithm if the sequence identity cutoff between two nodes is too low. In addition, tip clipping may remove short alternative ends, especially for high k-mer sizes.

Further, a trivial step is to remove chains of nodes in the de Bruijn graph in order to decrease the memory consumption, see Section 2.3.1. In what follows it is assumed that all chains of nodes have been simplified in the de Bruijn graph.

5.2.2 Scaffolding of Loci

As introduced in the previous section, the task is to identify reads from transcripts that are related through alternative splicing. In an ideal scenario all transcripts that are sequenced form groups of transcripts, called *loci*, which are only k-alternative. Therefore the basic idea is to compute the connected components of the transcript de Bruijn graph, each representing one locus. The length of a node is an indicator of its uniqueness [173], nodes are called *short* if their length is smaller than $50+k-1$ bps and they are called *long* otherwise. Nodes that are connected by at least one single read are called *direct*. If a connection is supported only by paired-end reads it is called *indirect*.

In the next part it is outlined which filters on direct and indirect connections are imposed to reduce false positive association of loci.

Construction and filtering of the scaffold

The de Bruijn graph partitions each read into its k-mers and thus the read path in the de Bruijn graph may span over several nodes. If node connections are direct the distance beween nodes can be computed from the read sequence connecting both nodes. Otherwise the distance has to be estimated from the indirect connections, see 2.3.1. There are two measures associated to each connection or edge in the graph. The first is called the *support* of a node, that is the number of direct reads or indirect read-pairs that span the connection. The second is a more finegrained measure. Each edge between node n_i and n_j is assigned a weight w_{ij}. w_{ij} is the sum of the weights w_{ijk} of all read sequences R_k that establish a connection between node n_i and node n_j:

$$w_{ij} = \sum_{\forall R_k : n_i \rightarrow n_j} w_{ijk} \ . \tag{5.1}$$

The individual weights differ for single-end and paired-end reads. For single-end reads the weights are set to $w_{ijk} = 1$, as each read gives direct connection information for this edge. For paired-end reads the situation is slightly different and the insert size between both reads has to be considered. A natural way of setting the edge weights is to relate them according to their probability of occurence which is approximated by the probability of the distance D according to the normal distribution centered around μ, with standard deviation σ. If d is the estimated distance between the two reads of a given pair, then the weight w_{ijk} for the pair is:

$$w_{ijk} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}} \ . \tag{5.2}$$

After calculation of the weights each edge is filtered if its support < 4 (by default) and its weight $w_{ij} \leq 0.1$. These rules apply only for connections between long nodes. A connection between a short and a long node must be direct. Only direct connections between two short nodes are considered, but only if there is no intermediate gap between them.

In order to account for random connections between nodes from highly expressed transcripts the number of expected connections is further computed between long

nodes using a modified version of the statistic presented in [173]. The number of connecting read pairs between nodes n_a and n_b of a gene with a single transcript T_i is estimated to be:

$$E(X) = \rho_a \left[\sigma \left(\Phi(M) - \Phi(N) - M \int_M^N \Phi \right) + l_B \int_N^O \Phi - \sigma \left(\Phi(O) - \Phi(P) - P \int_O^P \Phi \right) \right] \quad (5.3)$$

$$M = \frac{D - \mu}{\sigma} \quad (5.4)$$

$$N = \frac{D + l_a - \mu}{\sigma} \quad (5.5)$$

$$O = \frac{D + l_a - \mu}{\sigma} \quad (5.6)$$

$$P = \frac{D + l_a + l_b - \mu}{\sigma} \ , \quad (5.7)$$

where ρ_a is the mean density of reads in n_a. l_a and l_b denote the length of nodes n_a and n_b, w.l.o.g. assuming that $l_a \geq l_b$. Φ denotes the standard normal probability distribution function. In this model it is assumed that the probability between two read pairs depends only on the read position and the read pair insert length. However, the above formula is only true if T_i is the only transcript expressed in the gene. If an alternative isoform T_j is expressed as well, that does not contain the sequence of n_a or n_b, the value of formula 5.3 overestimates the number of expected connections. Therefore, ρ_a is substituted by the density ρ_a^*, which is adapted for the case that one of the two nodes is part of an alterntive exon:

$$\rho_a^* = \min\{\rho_a, \rho_b\} \ , \quad (5.8)$$

where ρ_b is the density of reads in n_b respectively.

Finally, all indirect connections with support $< E(X) \times 0.1$ are eliminated.

After false positive connections have been removed according to the criteria defined above, the clustering into loci is conducted similar to Butler et al. [20]. First, connected components between long nodes (default 50+k-1 bps) in the graph are constructed, because these long nodes have a higher likelihood of being unique. Second, small nodes that share direct sequence overlap with long nodes are added to form the

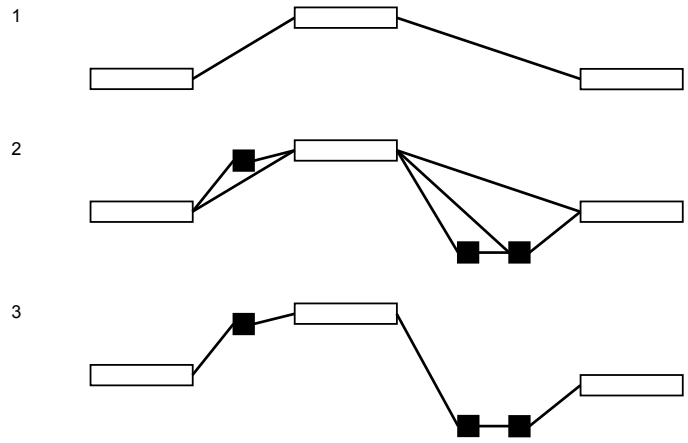

Figure 5.5: (1) The locus is built around a connected component of long nodes (white boxes). (2) To these nodes are added the short nodes (black) which share a direct connection with one of the long nodes. (3) Transitive connections are removed from the locus.

locus, see Figure 5.5 for a simple example. Transitive connections are removed, see below.

Orientation of the loci

As mentioned before, the reads from an RNA-Seq experiment normally come without the orientation of the fragment. However, for running the algorithm explained in Section 5.2.4 it is necessary to determine if a node sequence belongs to the sense or antisense strand of a gene. Oases uses a simple heuristic to process all nodes in a locus in consistent orientation. In a locus that is k-alternative all connections belong to exon parts from the same gene, therefore the orientation is imputed at the locus level. The number of stop codons for each node's sequence and its reverse complement sequence is computed for each of the three possible reading frames. Independent of node size, the orientation of each node is the one with the minimum number of stop codons in all reading frames. Finally, the orientation of a locus is decided by majority vote, weighted by node length. If a locus is in majority antisense, all of its nodes are replaced by their reverse complements, otherwise it is left unmodified.

This is the same problem as predicting the orientation of ESTs, which has been addressed by many researchers and different tools have been designed for finding coding sequences in ESTs and thereby predicting the orientation of the fragment

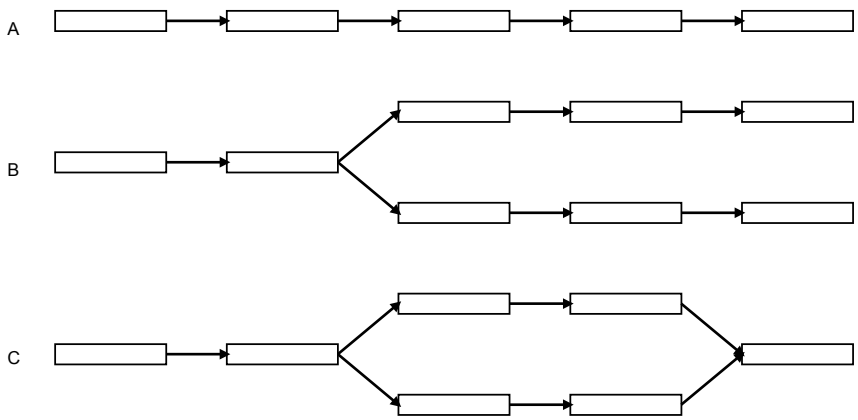

Figure 5.6: The following topologies have only one or two possible transcripts, which can easily be deduced: (A) linear graphs, (B) forks, and (C) bubbles. Boxes denote nodes in the graph with indirect connections that cannot be simplified.

[118, 99]. Most of these tools rely on codon usage composition tables in one way or another. As efficient tools exist to orient partial transcripts, the simple predictions made by Oases can later be corrected in a postprocessing step if codon usage tables for the same or closely related species exist.

Transitive Reduction

Finally, for the following analyses to function properly, it is necessary to remove redundant long distance connections. A connection is considered redundant if it connects two nodes which are connected by a distinct path of connections of comparable total length. An efficient algorithm exists for this task originally designed for string graphs by Myers [116]. This algorithm was adapted to the fact that short nodes can be repeated and induce a cycle in the TG, see Section 5.1.2. Because of this, occasional situations arise where every connection leaving a node can be transitively reduced by another one, thus removing all of them, and breaking the connectivity of the locus. To avoid this, a limit is imposed on the number of removed connections, such that each node keeps at least one in- and one outgoing node. If two connections have the capacity to reduce each other, the shortest one is preserved.

5.2.3 Recognition of Trivial Structures

After each locus has been processed as detailed above, the task is to reconstruct the expressed transcripts. In many cases, the loci present a simple topology which can be trivially and uniquely decomposed into one or two transcripts. The three categories of trivial locus topologies are: *chains*, *forks*, and bubbles (Figure 5.6). These three topologies are easily identifiable thanks to the node degree distribution of the locus, that can be computed with a linear traversal of the graph with V nodes and E edges in $\mathcal{O}(V + E)$ time.

Oases computes for each node the indegree (*indeg*) and the outdegree (*outdeg*), for a total of t nodes in the locus. It is assumed, w.l.o.g., that all nodes in the graph are considered in the same direction. If t-2 nodes have *indeg*=1 and *outdeg*=1, one node only *indeg*=1 and the other node only *outdeg*=1, then the locus graph is necessarily linear. Only one maximal transcript can be produced from this locus. If t-4 nodes have *indeg*=1 and *outdeg*=1, one node has *indeg*=1 and *outdeg*=2, two nodes only *indeg*=1, and the remaining node has only *outdeg*=1, then the graph presents a simple fork. Finally, if t-3 nodes have *indeg*=1 and *outdeg*=1, one node has *indeg*=1 and *outdeg*=2, one node only *outdeg*=1, and the remaining node only *indeg*=2, then the locus graph presents a bubble. In these latter cases, only two maximal transcripts can be extracted.

Note that for trivial cases even if the locus is homologous, the corresponding transcripts would be reconstructed correctly.

5.2.4 Prediction of Full Length Transcript Sequences

The task remains to produce the underlying transcript sequences from a locus graph that has no trivial topology . The idea is that among all possible paths in the graph the one with highest read coverage is likely to represent an actual transcript. A heuristic is applied that utilizes the fact that the sequencing coverage of a transcript is directly linked to its expression (for standard RNA-Seq protocols where the expression levels are not normalized). The assumption is that the isoforms in a locus share the majority of the exons. Further, alternative exon events in a locus are assumed to be statistically independent, unless additional data, like paired end reads, indicate otherwise.

The algorithm of Lee and co-workers, initially suggested for the reconstruction of consensus sequences from ESTs, can be adapted to solve this problem [86, 167]. The algorithm has been proposed for splicing graphs (they call them partial order graphs) built from EST data after alignment to a reference genome. The orientation of RNA-Seq short-read data is unknown (Fig. 1.4), but for the purpose of this algorithm all nodes in the TG will be considered only in sense or antisense direction as explained in Section 5.2.2.

Dynamic Programming Algorithm

The algorithm is based on dynamic programming (DP) and works on the directed TG with node weights s_i and edge weights w_{ij} for two nodes n_i and n_j. The weight of each node is defined as the read density in the node $s_i = \rho_i$. The implementation in Oases differs in that the TG may have cycles which are treated separately and that edge weights from direct and paired-end reads are considered, as defined in Section 5.2.2.

The following recursion is applied. Each s_j is maximized according to the predecessor with maximum edge weight:

$$chooseBestPredecessor(j) = \arg \max_{\forall i,\ n_i = pred(n_j)} (w_{ij}), \tag{5.9}$$

where $pred(n_j)$ denotes a predecessor node of n_j.

Cycle Removal

As the DP-algorithm was originally designed to work on acyclic graphs, that can be partially ordered, it could always be guaranteed that the calculations for all predecessor nodes $pred(n_a)$ of node n_a could be completed. In case of cycles in the graph this is no longer given. There are two cases that may occur which are treated differently in the algorithm. In both cases heuristic rules are applied to remove the cycle but retain precious sequence, inspired by the bulge removal in A-bruijn graphs for repeat analysis [124].

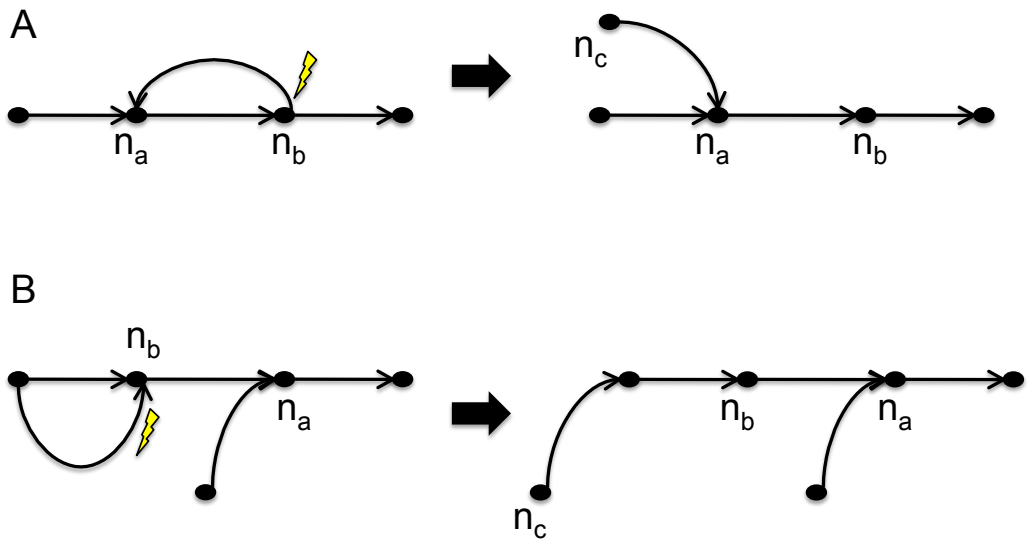

Figure 5.7: Removal of cycles in Transcript de Bruijn graphs that block the DP-recursion from n_a. Nodes are displayed as black dots. (A) The predeccesor node n_b is also a succesor of node n_a. The outgoing cycle edge of n_b is cut and a new node n_c is created. (B) The upstream cycle prevents completion of predeccesor node n_b. The incoming cycle edge of n_b is cut to form a chain starting with n_c.

In the first case, node n_a has another node n_b as its, not necessarily direct, predecessor and as its succesor node at the same time. If this case is detected, the outgoing arc from n_b is detached and a new starting node n_c is created (Figure 5.7A). In the second case, the predeccessor node n_b of n_a lies downstream of a cycle that cannot be resolved. This time the incoming cycle edge to n_b is detached and a chain of nodes starting with the new node n_c is created. In both cases the DP-algorithm is restarted from n_c.

Iterated Traversal

After all nodes in the graph have been processed the highest expressed transcript is found by backtracking from the node in the graph with maximum score $\max_i s_i$. After one iteration of DP, the weights in the graph are adjusted and edges that have not been visited previously are upweighted. The idea being, that long and highly expressed nodes which are not part of any so far predicted transcript from the locus are enforced to be part of the next generated transcript. Unassigned nodes are sorted by expression and the most highly expressed node n_i is selected. Connecting edges

of n_i are upweighted as follows:

$$w_{ij} \;=\; f \cdot w_{ij} \; and \; w_{ji} \;=\; f \cdot w_{ji} \; . \tag{5.10}$$

f is a scaling factor for the weights (f=1000 as in Lee et al. [86]). Note that this couples non-adjacent nodes that have indirect connections. At the end of each iteration all edge weights are first set to their original value, the next unassigned node n_i is selected, the edge weights are upweight and a new transcript is predicted in the next iteration. In Oases each locus can generate up to 10 such transcripts.

The running time of the DP-algorithm is $\mathcal{O}(t(V+E)+(c\cdot E))$, where the first term is the traversal of the locus graph with V nodes and E edges which is done for each of the reconstructed t transcripts. The second term is the contribution from the removal of c cycles each of which involves the traversal of at most E edges.

5.2.5 Merged Assemblies

A common problem in de Bruijn graph assemblers is the optimization of k, see Section 2.3.1. For Transcript de Bruijn graphs this optimization is more subtle as transcript expression levels are distributed over a wide range, see Section 5.4.2. A way to avoid the dependence on the parameter k is to produce a *merged transcriptome assembly* or meta assembly of previously generated transcripts from Oases. Oases is run for a set of $K = \{k_1, \ldots, k_z\}$ values and the produced transcript output is stored. All predicted transcripts from runs in K are then fed into Oases again. The de Bruijn graph is built over the transcripts and all steps are repeated, error correction, loci assembly, and the DP algorithm. The output from this run is considered the merged transcriptome assembly. A clear disadvantage of this strategy is the running time, as the graph needs to be built and analyzed for every single k.

As it is possible to build the de Bruijn graph in Velvet with additional long reads that allow the resolution of long repeats and merging of disconnected regions [173], Oases can also be run with additional input from previous sequencing experiments or known cDNAs that may help to resolve alternative splicing events for example. Therefore predicted transcripts from previous runs are fed into Oases as long reads for the merged transcriptome assembly.

5.2.6 Transcript Confidence Scores

The proposed DP-algorithm produces for each input graph the set of highly expressed transcripts. However, depending on the complexity of the locus the confidence of these predictions may vary. Especially, when a locus is k-homologous with a large number of transcripts from different genes, the initial assumption of the DP-algorithm is violated. Many scenarios can be imagined where two parts of transcripts from distinct genes are fused by the algorithm to create a false positive fusion transcript.

Ideally, one would like to have a confidence value for each reconstructed transcript to discard low quality ones, for example in an application where no closely related reference genome exists. Unfortunately, there is no direct indicator of such a scenario. Nevertheless, the number of nodes in a locus can point to an unusually complex locus. Unless a k-alternative locus is repetitive or has many complex isoforms, the number of nodes rarely exceeds 10, see Section 5.3. Therefore a simple confidence score has been implemented in Oases. Each transcript T_i from a locus $\mathcal{L}$ is attributed a confidence value $\mathcal{C}\ (T_i)$ as follows:

$$\mathcal{C}(T_i) = 1, if\ T_i \text{ is derived from a trivial locus, otherwise} \tag{5.11}$$

$$\mathcal{C}(T_i) = \frac{nodes(T_i)}{nodes(\mathcal{L})}, \tag{5.12}$$

where $nodes(T_i)$ and $nodes(\mathcal{L})$ denote the number of nodes in the transcript T_i and the locus $\mathcal{L}$, respectively.

5.2.7 Prediction of Alternative Exon Events

Transcript de Bruijn graphs offer the base-pair resolution necessary to distinguish between different common alternative exon events, as illustrated in Section 5.1.3. Examining the topology and the sequences around the breakpoints in bubbles, it is possible to infer what type of alternative exon event took place. Oases therefore scans through the breakpoints of the graph and reports all of the identifiable alternative exon events. However, the prediction of splice sites is a difficult problem in bioinformatics and many different methods have been proposed to cope with the high degeneracy and short length of splice sites [171, 131, 105]. Currently Oases uses simple sequence motifs as listed in Table 5.1.

event	motif(s)	comment
poly-A tail	A-rich, $> 95\%$ A's	3' end of mRNAs, ≤ 200 bps
polyadenylation signal	AATAAA, ATTAAA	10-30 bps distance to poly-A
5' splice (donor) site	GT	directly after the exon
3' splice (acceptor) site	AG	directly before the exon

Table 5.1: Simple sequence motifs that are searched in bubbles in transcript de Bruijn graphs.

In order to allow the report of alternative exon events, the nomenclature of Sammeth et al. [134] for splicing graphs is applied to transcript de Bruijn graphs. Instead of annotating the boundaries of splicing graphs, referred to as *sites*, the start and end points of nodes in a locus in the transcript de Bruijn graph are utilized.

5.3 Influence of Repeats, Domains and Paralogs

Before Oases is applied to real data the influence of repetitive sequences, protein domains, and paralogous genes on the composition of loci in transcript de Bruijn graphs in Oases is investigated. The complete transcriptomes of *H. sapiens* (version 57), *M. musculus* (version 57), and *D. melanogaster* (version 56) have been downloaded from the Ensembl database in FASTA format. Further statistics about the gene and transcript content for each species are listed in Table 5.2.

Species	Ensembl Version	#Genes	Protein Coding	Total cDNAs
D. melanogaster	56	15,160	14,076	22,071
M. musculus	57	31,464	23,062	79,168
H. sapiens	57	44,670	22,318	130,240

Table 5.2: General statistics of the number of genes and downloaded cDNAs for Ensembl genomes of *D. melanogaster*, *M. musculus*, and *H. sapiens*.

As mentioned in the introduction and Section 5.1.2, a fundamental problem of de Bruijn graphs is the limit to sequence overlaps of size k-1. In Section 5.2.2 about the scaffolding of loci in Oases a couple of filters have been introduced to assure the assumption that each loci constructed in Oases is k-alternative and does therefore

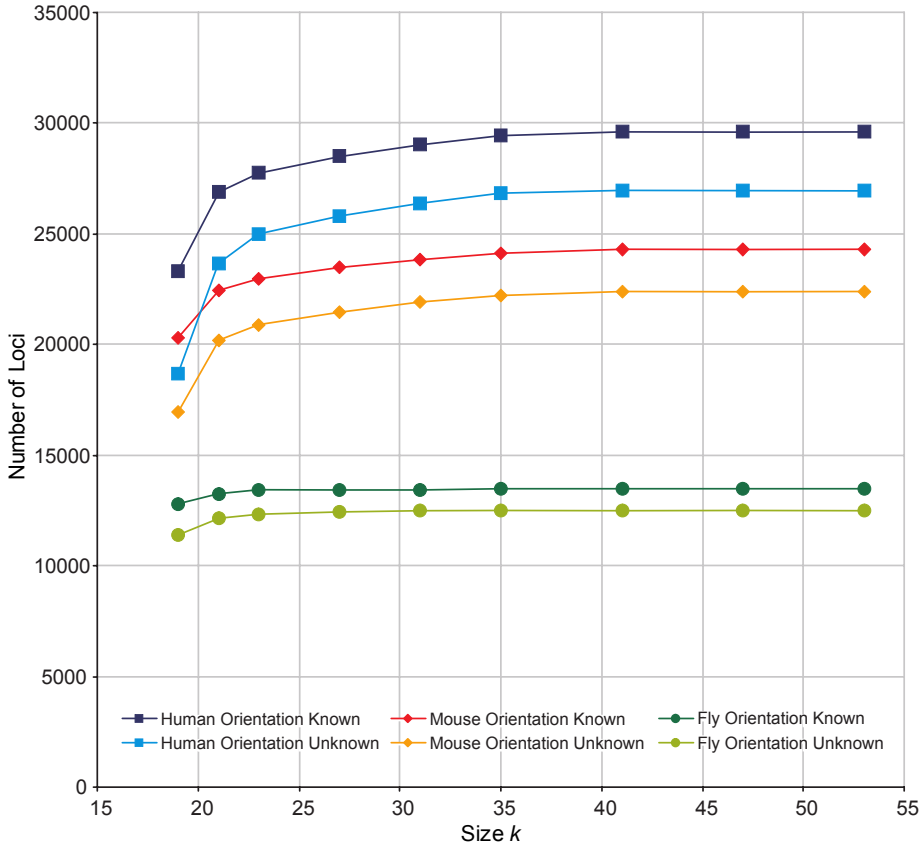

Figure 5.8: Computation of the number of loci assembled by Oases for different k. The numbers are reported for the complete transcriptomes of Human (boxes), Mouse (diamonds), and Fly (circles). In order to understand the difference between strand specific transcripts Oases was run with or without known orientation of the transcripts.

not contain transcripts from different genes. Recall that the initial computation of connected components in the graph is restricted to long nodes (by default (50+k-1) bps) and short nodes are only added to such a component if they have a direct connection to one of the long nodes, see Fig. 5.5. Oases was run for each species with all cDNA sequences provided as long reads (*-long* option). With this option the filtering by support is disabled and loci scaffolding through direct connection of cDNAs in long nodes is investigated. This analysis provides a worst-case upper bound on the complexity of each transcriptome, which simulates the scenario that many different tissues of an organism are sequenced and the data are pooled.

Transcript de Bruijn graphs were built for the transcriptome of each species for a series of $k = 19, \ldots, 53$ and the number of resulting loci are reported in Fig 5.8. In

general, the smaller k the smaller the number of loci, because the chance increases that a node has length smaller than 50+k-1 bps and recognizes overlaps due to (i) repetitive, domain, and paralogous sequences, or (ii) overlapping gene regions from different strands of the chromosome. The separation into loci is very stable and only for $k = 19$ or 21 large jumps are observed. For all three species the number of loci is unchanged if $41 \leq k \leq 53$. If we take $k = 23$ as an example the number of loci reported for human cDNAs is 24,966. That is quite a loss compared to the theoretical number of 44,670 genes in human (Table 5.2) of which, however, 12,308 are pseudogenes that might still share considerable amount of sequence with their template gene.

Overlaps due to (ii) can be removed by treating the de Bruijn graph not as a digraph but treating both strands independently (*-strand_specific yes* option in *velveth*), see Fig. 5.8. In all three species the number of genes overlapping on the opposite strand clearly decreases the number of loci. For example, even for k as large as 53 in mouse and human approximately 2000 genes are clustered into the same locus because they have overlaps on the opposite strand in the genome.

The previous analysis hides the fact that for small k, repetitive sequences often create small nodes with possibly many cycles that complicate the reconstruction of the transcript. Therefore, Drosophila transcripts have been masked for repeats with RepeatMasker version 3.2.8 with default parameters and the *-species Flies* option. Additionally, annotation for Pfam domains [46] and paralogous genes was downloaded for Drosophila transcripts from Ensembl version 56. In Table 5.3 the number of nodes per Oases locus are depicted for different values of k. Although the scaffolding often prevents that k-homologus transcripts are clustered into the same locus, more than 2,500 loci have at least 21 nodes for k=19. The repeat, domain, and paralogous information of each transcript was projected onto its locus in the graph. Taking k=27 as an example, all loci that had no annotation with any of the three categories had ≤ 6 nodes. In contrast, loci that had ≥ 10 nodes, were always annotated by more than one category.

k	$nodes(\mathcal{L}) \leq 10$	$11 \leq nodes(\mathcal{L}) \leq 20$	$20 < nodes(\mathcal{L})$
19	6,682	1,604	2665
23	10,224	902	932
27	11,108	622	446

Table 5.3: Number of loci with 1-10, 11-20, and >20 nodes, when Oases is applied to the complete Drosophila transcriptome. The higher the k the smaller is the influence of repeats on locus size.

5.4 Application to Paired-End RNA-Seq data

After analysis of the properties of perfect long reads with full expression, this section will deal with real data where transcripts differ by expression and sequencing errors complicate their detection. The influence of parameter k for real data is investigated and Oases is compared against the *de novo* genome assembler ABySS [144] and the recently developed transcriptome assembler Cufflinks [152]. In this section the term transfrag, for transcribed sequence fragment, is used for assembly output of one of the programs. Whereas the term transcript refers to a full length transcript given in Ensembl or expressed in the cell.

5.4.1 Data Sets

Two datasets were retrieved from the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra). The first data set was produced in a study by Heap et al., where poly(A)-selected RNAs from human primary CD4(+) T cells were sequenced [61]. Paired-end reads of length 45 bp with an insert size of 200 bps from one human individual (studyID SRX011545) were downloaded. All reads were processed by (i) removing Ns from both ends, (ii) clipping bases with a Sanger quality ≤ 10, and (iii) removing all pairs where one read had more than 6 bases with Sanger quality ≤ 10 after steps (i) and (ii), leading to a total of 30,940,088 reads.

The second data set was taken from the recently published study of Trapnell and co-workers [152]. The authors did a timeseries experiment of C2C12 myoblast mouse cells and sequenced paired-end reads of length 75bp with an insert size of 300bps. Read data from the 60hr timepoint (study id SRX017794) was kindly provided by

the authors due to problems with the Short Read Archive. The authors published the predictions of their transcriptome assembler Cufflinks, which were downloaded for later analyses. The obtained 67,261,680 reads have not been further processed as was done by the authors for running Cufflinks.

For each data set the reads were aligned onto the complete transcriptome of Ensembl with RazerS [163]. The minimum sequence identity required was 92 % and further the following parameters were set for RazerS *-m 20 -dr 2 -pa -mN -of 1 -s 1111011010001110011 -t 3*. Read counts have been summarized on the gene level in order to compute Reads Per Kilobase per Million mapped reads (RPKM) values for each gene [113]

$$RPKM = \frac{10^9 \cdot Y}{N \cdot s}, \tag{5.13}$$

where Y is the observed number of reads in the gene, N the total number of mapped reads in the experiment, and s the gene length. Matches of non-unique reads were equally weighted with $\frac{1}{\delta}$, where δ represents the number of different mapping positions in the transcriptome. Finally, each gene was grouped into one of the following three expression categories for each data set: (i) *low*$= 0 <$ RPKM ≤ 1, (ii) *med* $= 1 <$ RPKM ≤ 20, and (iii) *high* $=20 <$ RPKM.

In all following experiments with Oases the coverage cutoff was set to three and the minimum support for connections to four if not stated otherwise. Predicted transfrags of Oases and ABySS were aligned against the genome using Blat [78] with default parameters and the sensitivity and specificity values against Ensembl annotation were computed with Cuffcompare [152].

5.4.2 Influence of Parameter k

It was shown in the previous section that a large k is generally preferable over a smaller k, in order to simplify the locus topology. But the common problem of de Bruijn graph based sequence assemblers is the susceptibility to sequencing errors. Each sequencing error can destroy up to k k-mers and therefore the influence of sequencing errors has to be balanced against the gain in repeat resolution when choosing k. In RNA-Seq data an additional difficulty is that a large fraction of the expressed transcripts is represented with very few sequence reads as often only 5% of the expressed genes contribute $\geq 50\%$ of all reads [19]. In consequence, a large k will

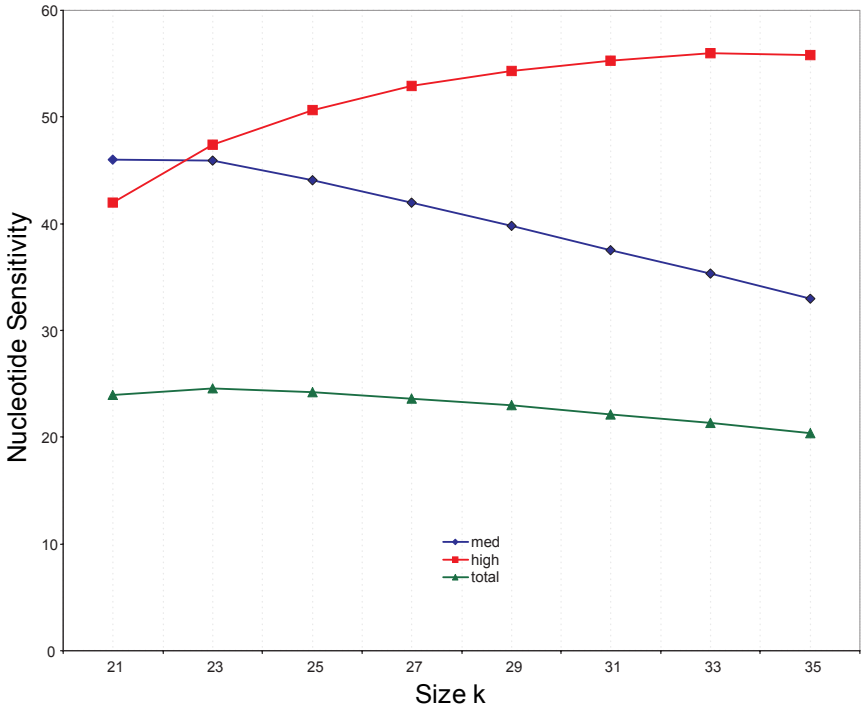

Figure 5.9: Application of Oases to the paired-end mouse C2C12 dataset for different values of k (x-axis). The nucleotide sensitivity for the med and high expression level categories as well as the total set of Ensembl 57 annotations is depicted (y-axis). Highly expressed transcripts are more accurately reconstructed with large k values, where repeat resolution is easier.

additionally favor highly expressed transcripts. In order to observe these oppositional influences Oases was run for $k = 21, \ldots, 35$ on the mouse C2C12 data set and the nucleotide sensitivity for the expression categories low, med, and high was recorded (see above). Categories high and med are shown in Figure 5.9. For $k = 21$ the nucleotide sensitivity in the med category is superior to the high category, due to complex locus topologies for highly expressed transcripts. However, with each stepwise increase of k the sensitivity drops in the med category and rises in the high category until it finally starts to drop in the high category for $k \geq 35$. Nevertheless, if the nucleotide sensitivity is computed against all genes in Ensembl, the best value is attained for k=23 because there are more genes in the med category than in the high category.

The above observations directly show that the computation for one single k is limited, as results are only adequate for a subset of all expressed transcripts. Interestingly, it can be observed in Figure 5.9 that the total sensitivity for the displayed range of k

should increase when the sequence output for multiple k values is combined, due to the large increase in nucleotide sensitivity in the high category for k=33. As mentioned in Section 5.2.5, Oases is able to make use of previously assembled transfrags and can produce a merged assembly for all of them. Application of the merged assembly approach with the considered range of $K = \{21, \ldots, 35\}$ for the mouse data set increases the total nucleotide sensitivity from 24.6 to 27.5 as discussed later for the comparison with Cufflinks.

5.4.3 Comparison with ABySS

As mentioned in the beginning of this chapter, genomic assemblers have been applied to RNA-Seq data [35, 157]. The most comprehensive study was conducted by the group that developed ABySS [144, 11] that applied ABySS to a set of over 150 Mio. human paired-end RNA-Seq reads. However, ABySS was still run as a genomic assembler and the authors analyzed small contigs that described junction nodes. ABySS does not aim at reporting alternative isoforms though. In order to show the advantage of Oases against a genomic assembler, a comparison with ABySS on the human CD4 dataset was conducted. Oases and ABySS were run for $k = 19, \ldots, 35$ and the predicted transfrags where mapped against the genome with Blat [78].

Method	k	transfrags > 100	N25	N50	N75	Total	Mapped
Oases	19	56,858	410	912	1,681	30,628,341	56,737
Oases	19-35	44,021	537	1,287	2,313	29,870,496	43,951
ABySS	23	27,714	365	801	1,557	14,637,999	27,667

Table 5.4: Oases and ABySS Results on 30,940,088 paired-end reads from a human CD4 RNA-Seq data set. The number of transfrags with length > 100 and their N25, N50, N75, as well as the total sequence output are reported. Mapped gives the number of transfrags that mapped to the genome using Blat.

In Table 5.4 the statistics for the best k are reported for Oases and ABySS. The first observation is that Oases reports far more transfrags adding up to about the double amount of total sequence output (Supplemental Data S8A). This is to be expected, as ABySS reports only contigs whereas Oases aims at reporting full length transfrags. For example for the simple fork topology (Fig. 5.6), ABySS would most likely report

Method	k	level	low	med	high	Full-SE	Full-SP
Oases	19	Nucleotide	0.5	27	60.3	15.1	83.7
Oases	19-31	Nucleotide	0.5	27.1	62.1	15	81.8
ABySS	23	Nucleotide	0.3	18.9	53.9	11.5	85.1
Oases	19	Exon	0.2	14.9	38.4	9.3	37
Oases	19-31	Exon	0.3	15.3	41.1	9.5	41.6
ABySS	23	Exon	0.2	10.8	34.2	7.4	44.5
Oases	19	Intron	0.4	26.4	63.4	16.2	82.9
Oases	19-31	Intron	0.4	26.7	65.9	16.1	80.4
ABySS	23	Intron	0.3	19.2	56.8	12.7	84.3

Table 5.5: Sensitivity for nucleotide, exon, and intron level for aligned Oases and ABySS transfrags for human CD4 data subdivided into categories of low, medium, and high expression levels as well as the sensitivity (SE) and specificity (SP) for the full set of Ensembl 57 transcripts (Full).

three transfrags, whereas Oases reports two longer transfrags. The more fragmented and complicated a locus is, the more fragmented will be the transfrags of ABySS. For both programs Blat is able to align more than 98% of the transfrags, pointing to reasonable assemblies. The N25, N50, and N75, especially N50, are classical measures to compare the output of genomic assemblers. A more interesting measure is the coverage of annotated transcripts in the human genome. The classical measures used are Sensititvity and specificity for different levels of granularity, see Section 2.1.3. In Table 5.5 the sensitivity at exon, intron, and nucleotide level is reported additionally divided into the three expression categories, low, med, and high. Oases has higher sensitivity in all categories and most notably in the med and high category. For example in the med category for nucleotide sensitivity Oases has a 43% increase relative to ABySS. The last column of the table gives the specificity against the complete annotation of Ensembl. Although the specificity of ABySS is 2% higher, this is caused by the additional output of Oases that lies outside the reference annotation.

In the second row of Table 5.4 the result of a merged assembly of Oases for all tested k values is reported. The merged assembly is produced at $k = 35$ to have maximal repeat resolution (Supplemental Data S8B). It should be noted here that Oases can in principle also produce a merged assembly of the ABySS transfrags, but this was

Category	Transfrags	% of total	% of length
Match to known isoform	3,558	8.10	17.41
Contained in known isoform	25,312	57.59	40.15
Novel isoform of known gene	2,225	5.06	11.52
Pre-mRNA junction	3,967	9.03	9.15
Intronic	2,397	5.45	2.14
Polymerase run-on	807	1.84	0.89
Intergenic	1,073	2.44	1.17
Other artifacts	4,612	10.49	17.56
Total	43,951	100.00	100.00

Table 5.6: Analysis of Oases transfrags produced from the merged assembly approach for human CD4 RNA-Seq data. The overlap with Ensembl 57 annotation was computed with Cuffcompare [152].

not investigated as the direct output of Oases is more sensitive already. The largest improvement is in the N50 of the transfrags from the merged assembly which is increased by almost 400 bps, without affecting the total assembly length. Although fewer sequences are output the sensitivity is not decreased.

In Table 5.6 the overlap with predicted Oases transfrags and Ensembl 57 mouse annotations is reported. The program Cuffcompare [152] was used to classify all aligned transfrags. In total 8% of the transfrags represent approximately 3,600 isoforms with complete exon-intron borders. The majority (58%) of the transfrags is contained in known isoforms, which are not reconstructed at full length. 2,225 putative new isoforms and 2,397 transfrags residing in introns of known genes are assembled. A further 1,073 transfrags represents putative new intergenic transcripts. The approximately 4,000 transfrags in the category *pre-mRNA junction* represent unspliced fragments that extend into the intronic region of an annotated transcript and might stem from remaining pre-mRNA fragmens in the sample.

5.4.4 Comparison with Cufflinks

In order to explore the difference between a *de novo* approach and a genomic transcriptome assembler, the recently developed Cufflinks method [152] was benchmarked

Method	k	transfrags > 100	N25	N50	N75	Total	Mapped
Oases	23	95,220	399	767	1,378	49,466,400	94,984
Oases	21-35	64,674	898	1,736	2,941	63,279,139	64,518
Cufflinks	-	72,745	956	2,613	4,522	73,084,627	-

Table 5.7: Oases and Cufflinks Results on reads from a mouse C2C12 cell line RNA-Seq data set. The number of transfrags with length > 100 and their N50 is reported. Column Mapped gives the number of Oases transfrags that mapped to the genome using Blat.

against Oases. Cufflinks expects as input RNA-Seq reads that have been mapped with a spliced alignment algorithm. From these set of reads Cufflinks assembles a parsimonious set of transfrags by computing a maximum weighted matching in a weighted bipartite fragment compatibility graph. The C2C12 timepoint RNA-Seq dataset was produced for the same study and Cufflinks predictions with this dataset can be regarded as the possible upper-bound, as Cufflinks parameters were set by the authors.

Four important advantages for Cufflinks compared to the *de novo* setup of Oases should be mentioned: (i) sequencing reads are aligned to the genome such that a higher rate of sequencing errors can be accounted for, (ii) reads from repetitive sequences map to multiple genomic locations and are treated differently, (iii) genomic signals like splice sites are used to extend the exon boundaries, and finally (iv) aligned sequence reads are grouped into exons by genomic proximity.

Table 5.7 shows the number of transfrags for Oases and Cufflinks. First note that the merged assembly of Oases drastically improves upon the single-k assembly, as seen in Fig. 5.9. The N50 is doubled and the total sequence output increased by 14 megabases. As expected, Cufflinks outperforms Oases in all regards. The N50 of Cufflinks is almost 1,000 bps larger. In Table 5.8 the sensitivity and specificity is shown for the three expression categories low, med, and high. Compared to the CD4 data, the merged assembly of Oases improves the sensitivity of nucleotide, exon, and intron level by 3-4% upon the assembly with the best $k = 23$ (Supplemental Data S8C-D). Cufflinks has the greatest improvement over Oases for the low nucleotide sensitivity category, which may be explained by the fact that sequencing errors are more easily accounted for and that nearby reads can be grouped into exons of length at least 100 bps. In the high category the difference between Oases and Cufflinks is

Method	k	level	low	med	high	Full-SE	Full-SP
Oases	23	Nucleotide	0.6	45.9	47.4	24.6	88
Oases	21- 35	Nucleotide	0.7	49.4	62.5	27.5	85.6
Cufflinks	-	Nucleotide	24.1	69.6	71.4	45.3	67.0
Oases	23	Exon	0.1	34.8	28.6	18.5	39.6
Oases	21- 35	Exon	0.2	38.5	46.3	22.4	49
Cufflinks	-	Exon	12.5	52	52.4	33.5	57.1
Oases	23	Intron	0.2	54.7	47.8	29.7	90.8
Oases	21- 35	Intron	0.4	59.3	71.9	34.7	86.7
Cufflinks	-	Intron	19.9	75.7	80.6	50.2	94.8

Table 5.8: Sensitivity for nucleotide, exon, and intron level for aligned Oases and reported Cufflinks transfrags [152] for the mouse C2C12 cell line subdivided into categories of low, medium, and high expression levels as well as the sensitivity (SE) and specificity (SP) for the full set of Ensembl 57 transcripts (Full).

generally smaller than in the low and med category.

Table 5.9 further shows the classification of the transfrags with Cuffcompare. Oases assembles 6,870 isoforms with complete exon-intron borders, which is about 71% achieved for Cufflinks transfrags. Almost three times more transfrags of Oases are flagged as contained in known isoforms compared to Cufflinks, which summarizes the advantage of Cuffllinks that can group nearby fragments into exons. In summary, the *de novo* approach of Oases allows the reconstruction of 60-67% of the transcript fragments - depending on the metric - that are assembled by Cufflinks with the help of the reference genome.

Category	Transfrags	total(%)	length(%)	Cufflinks
Match to known isoform	6,870	10.65	19.20	9,743
Contained in known isoform	33,889	52.53	37.20	13,066
Novel isoform of known gene	7,727	11.98	23.28	6,217
Pre-mRNA junction	5,507	8.54	7.63	3,708
Intronic	1,811	2.81	0.87	10,851
Polymerase run-on	1,708	2.65	1.28	6,224
Intergenic	1,653	2.56	1.26	20,771
Other artifacts	5,353	8.30	9.27	2,165
Total	64,518	100.00	100.00	72,745

Table 5.9: Analysis of Oases transfrags from the merged assembly approach for mouse C2C12 data. The last column gives the number of Cufflinks transfrags for comparison. The overlap with Ensembl 57 annotation was computed with Cuffcompare [152].

Chapter 6

Discussion

Detection of AEEs A similar stategy for the reference based spliced alignment was used by other authors [113, 33]. Other approaches for spliced alignment without reference have been published during this work. QPALMA [15] uses quality scores but is relatively slow, Tophat [151] is much faster. More recently, new algorithms have been developed that take a seed and extend approach to do spliced alignment with paired and single end reads [165, 2, 4, 152].

Here, a simple statistic was used to compute the expected number of random matches on splice junctions. People have extended this simple model for splicing detection incorporating the relative distance to the exon boundary into the statistic [160] and using a classifier that learns discriminative signals between true and artificially generated splice junction sequences [119].

For single genes, it is expected that the length of the variable region between two isoforms will influence the detection power of methods using read coverage (Simulations). For extreme cases affecting only a few bases of one exon -like NAGNAG sites [12] - those methods are likely unable to detect these changes. By design, CASI and DASI have certain biases in detecting splice variants. While CASI requires in most cases the existence of at least two transcripts for a gene, DASI is able to predict variations on single transcripts with only one transcript in each condition. In contrast to DASI, exons with low expression are not taken into account by CASI to avoid, for instance, the influence of potential annotation errors. Consequently, CASI predictions are based on a smaller set of expressed internal exons compared with DASI predictions.

In their principle, the CASI/DASI strategies could be paralleled with the type of analysis performed with exon arrays. It was shown that using digital information derived from only 8 million reads sequenced in each condition, the prediction of AEEs differentiating HEK from B cells exhibited higher sensitivity and specificity than estimations derived from exon arrays with 4 replicates. The comparison further demonstrated that one of the major problems with arrays is that the large variation of expression levels across exons of a given gene adversely affects the detection of AEEs. This problem adds up to the well-known array issues related to probe design, cross-hybridization and detection of specific signals for genes that are poorly expressed. Here, data showed that only a small fraction of the qPCR-verified AEEs were detected by the exon array.

Previous array-based predictions of AEEs reported a specificity of 82-85% and a sensitivity of 4-53% [75, 146]. Here, for the two human datasets, CASI alone reached a specificity of 89% and a sensitivity of 51%. However, it was shown that the inventory of AEEs is drastically improved after integration of splice junction reads. Given that the analysis was conducted with only 4 millions of mapped reads per cell line and allowed to estimate AEEs with largely improved performances as compared with exon array-based analysis, there is no doubt that an exhaustive inventory of alternative transcript isoforms will be made possible via RNA-Seq. It is essential to merge information from junction reads and predictions from CASI/DASI types of analysis. While highly expressed genes are associated with a large number of reads directly identifying the different splice junctions and will therefore identify a larger set of splice junctions in these genes, moderately abundant transcripts will, in many cases, show a sufficient number of exonic tags to allow the prediction of AEEs by CASI, but might not enable the identification of reads at splice junctions. In general, the complexity of AEEs in a given gene might better be addressed by junction or paired-end reads, because for exons affected by multiple variations, the read distribution will be difficult to interpret.

Isoform Quantification Clearly, the model proposed in Chapter 4 can be extended to include junction reads, which should improve the performance, especially as the reads are continously getting longer with updates of NGS sequencers. The introduction of paired-end reads can also be modeled, but this should be done not in the naive way to keep the size of the indicator matrix small. A possible solution might be to do transitive reduction for these connections.

Another estimation algorithm based on the Poisson assumption has been proposed by Jiang and Wong recently [74]. The authors, however, did not maximize the log-likelihood but obtained an estimate of the posterior distribution by using importance sampling. They did not compare their approach with qPCR data but showed that RNA-Seq estimated transcript expression levels correlate with transcript expression levels estimated from custom splicing arrays (PCC ~ 0.6).

As POEM solves the quantification problem for isoforms that share genomic regions, it is straightforward to apply the algorithm to resolve the assignment of non-unique read matches simultaneously with the quantification problem. The indicator matrix just needs to be constructed for all the transcripts that share reads which are mapped to different genomic locations. In a recent work by Li *et al.* [92] and Howard *et al.* [67] such an approach was shown to improve the accuracy of estimations due to increased sample size.

In this work the Poisson distribution was assumed and computed estimations showed good correlation with the qPCR experiments. However, with further understanding of the biases and effects introduced by the sequencers and the protocols it can be expected that POEM, CASI, and DASI can be improved by adapting the p_e. Two interesting approaches that learn a background distribution have been proposed [67, 112] recently.

De novo transcriptome assembly

Establishing strand specific sequencing protocols for RNA-Seq [121, 155] is a very important research direction which will avoid the problem of overlapping or repetitive genes from the opposite strand of a chromosome, see Fig. 5.8. Notably, for transcriptome assembly of RNA-Seq data from cancer cells this could have practical clinical relevance for the differentiation between real and false-positive fusion genes.

Strand specific RNA-Seq data allow the prediction of Alternative Exon Events directly from the de Bruijn graph as shown in Section 5.1.3. If the data is not strand specific, as was the case for the datasets analyzed in this work, the prediction of sequence motifs has to be done on the node sequences and their reverse complement sequences, therefore increasing the risk of false positive predictions. The accurate prediction of splice sites, as assumed in Section 5.1.3, is challenging in practice because

splice sites are short and degenerative sequence motifs [171, 131, 105]. Prediction success may depend on the availability of closely related species data. Moreover, especially alternative splice sites were reported to show deviation from the consensus description of constitutive splice sites [149, 166] complicating the prediction further. It would be interesting to investigate prediction of AEEs in de Bruijn graphs for strand specific data and in combination with efficient splice site predictors, like [131].

One reason why the merged assembly approach works better than the assembly for a single k, is that highly expressed transcripts accumulate far more reads with sequencing errors than lowly expressed transcripts. With higher k more of these errors are removed by the error correction steps of Velvet and the Sensitivity increases. A reasonable extension would be a probabilistic coverage cutoff for nodes in the graph that adapts to the mean expression level of a locus, thus correcting sequencing errors more appropriately for highly expressed transcripts and smaller k.

An alternative approach to merged assembly is to build the de Bruijn graph not for a single k but for a set of sizes $K = \{k_1, \ldots, k_z\}$ and merge the individual de Bruijn graphs for one k into a common data structure. Such an approach was recently suggested by Peng and co-workers for genome assembly, who designed an iterative de Bruijn graph assembler (IDBA) [122]. The IDBA starts with small k, recording and resolving differences in the de Bruijn graph topology for higher k using an intermediate data structure. This approach has the advantages that it is faster than building the graph for each k with subsequent merging and likely more accurate in resolving sequencing errors as compared to the merged assembly approach introduced in Section 5.2.5. An interesting direction for further development would be to design such an intermediate de Bruijn graph data structure tailored for transcriptome assembly.

Along these lines, an improved error correction procedure for short sequencing reads from resequencing studies, devoid of a fixed parameter k as in de Bruijn graph approaches, was proposed by Schroeder et al. [138], who suggested to build a data structure called the *suffix tree* of the reads. In their algorithm overlaps of read subsequences of different sizes are used to detect and correct substitution errors more accurately, which could be extended for RNA-Seq data. Better error correction of reads creates more exact overlaps in the read sequences, which could prove especially useful for de novo assembly of lowly expressed transcripts, where most of the per-

formance is lost compared to approaches that use a reference sequence (*cf.* Table 5.8).

As mentioned in Section 5.1.2, the occurrence of alternative transcription start sites is ambiguous in the de Bruijn graph and the longest of the forms is output by the algorithm. However, as downstream exons are always shared, the coverage increases at the border of downstream exons. In the light of the CASI method described in Chapter 3 a segmentation procedure could be designed that looks for sudden increases in the read coverage, pointing to a putative exon border, given sufficient read coverage.

Throughout this thesis the length of the sequencing reads as well as the total output has steadily increased. For example the Illumina GenomeAnalyzer started with single-end reads of length 25 bps and nowadays 100 bps with paired-end support are routinely achieved. In addition, the output of the machine has increased by a factor 10 from 4 million reads per lane to 40 million. Finally, NGS technologies are almost to be replaced by third-generation sequencing approaches already [30, 44] that will further increase output and read length. With this respect the methods developed here will have to be adapted to the changing demand of the technologies. Especially for *de novo* assembly the longer read length might lead to a "Renaissance" of the overlap graph based assemblers or possibly to hybrid approaches that combine the best of both worlds.

Bibliography

[1] *Affymetrix White Papers:Alternative Transcript Analysis Methods for Exon Arrays v1.1.*, 2005. 24 June 2009, date last accessed.

[2] Adam Ameur, Anna Wetterbom, Lars Feuk, and Ulf Gyllensten. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol*, 11(3):R34, Mar 2010.

[3] Miguel Anton, Dorleta Gorostiaga, Elizabeth Guruceaga, Victor Segura, Pedro Carmona-Saez, Alberto Pascual-Montano, Ruben Pio, Luis Montuenga, and Angel Rubio. SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biol*, 9(2):R46, Feb 2008.

[4] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*, Apr 2010.

[5] S. Audic and J. M. Claverie. The significance of digital gene expression profiles. *Genome Res*, 7(10):986–995, Oct 1997.

[6] N. Bakalara, G. Kendall, P. A. Michels, and F. R. Opperdoes. Trypanosoma brucei glycosomal glyceraldehyde-3-phosphate dehydrogenase genes are stage-regulated at the transcriptional level. *EMBO J*, 10(12):3861–3868, Dec 1991.

[7] Serafim Batzoglou, David B Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P Mesirov, and Eric S Lander. ARACHNE: a whole-genome shotgun assembler. *Genome Res*, 12(1):177–189, Jan 2002.

[8] Tim Beissbarth, Lavinia Hyde, Gordon K Smyth, Chris Job, Wee-Ming Boon, Seong-Seng Tan, Hamish S Scott, and Terence P Speed. Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, 20 Suppl 1:i31–i39, Aug 2004.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[10] Matthew Berriman, Brian J Haas, Philip T LoVerde, R. Alan Wilson, Gary P Dillon, Gustavo C Cerqueira, Susan T Mashiyama, Bissan Al-Lazikani, Luiza F Andrade, Peter D Ashton, Martin A Aslett, Daniella C Bartholomeu, Gaelle Blandin, Conor R Caffrey, Avril Coghlan, Richard Coulson, Tim A Day, Art Delcher, Ricardo DeMarco, Appolinaire Djikeng, Tina Eyre, John A Gamble, Elodie Ghedin, Yong Gu, Christiane Hertz-Fowler, Hirohisha Hirai, Yuriko Hirai, Robin Houston, Alasdair Ivens, David A Johnston, Daniela Lacerda, Camila D Macedo, Paul McVeigh, Zemin Ning, Guilherme Oliveira, John P Overington, Julian Parkhill, Mihaela Pertea, Raymond J Pierce, Anna V Protasio, Michael A Quail, Marie-Adèle Rajandream, Jane Rogers, Mohammed Sajid, Steven L Salzberg, Mario Stanke, Adrian R Tivey, Owen White, David L Williams, Jennifer Wortman, Wenjie Wu, Mostafa Zamanian, Adhemar Zerlotini, Claire M Fraser-Liggett, Barclay G Barrell, and Najib M El-Sayed. The genome of the blood fluke schistosoma mansoni. *Nature*, 460(7253):352–358, Jul 2009.

[11] Inanç Birol, Shaun D Jackman, Cydney B Nielsen, Jenny Q Qian, Richard Varhol, Greg Stazyk, Ryan D Morin, Yongjun Zhao, Martin Hirst, Jacqueline E Schein, Doug E Horsman, Joseph M Connors, Randy D Gascoyne, Marco A Marra, and Steven J M Jones. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872–2877, Nov 2009.

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

[13] Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev*, 23(12):1379–1386, Jun 2009.

[14] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev. dbest–database for "expressed sequence tags". *Nat Genet*, 4(4):332–333, Aug 1993.

[15] Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar

Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180, Aug 2008.

[16] Paola Bonizzoni, Giancarlo Mauri, Graziano Pesole, Ernesto Picardi, Yuri Pirola, and Raffaella Rizzi. Detecting alternative gene structures from spliced ESTs: a computational approach. *J Comput Biol*, 16(1):43–66, Jan 2009.

[17] David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002.

[18] Jeremy Buhler and Martin Tompa. Finding Motifs Using Random Projections. *Journal of Computational Biology*, 9(2):225–242, 2002.

[19] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[20] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res*, 18(5):810–820, May 2008.

[21] Marshall P. Byrd, Miguel Zamora, and Richard E. Lloyd. Translation of Eukaryotic Translation Initiation Factor 4GI (eIF4GI) Proceeds from Multiple mRNAs Containing a Novel Cap-dependent Internal Ribosome Entry Site (IRES) That Is Active during Poliovirus Infection. *Journal of Biological Chemistry*, 280(19):18610–18622, 2005.

[22] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Number 30 in Econometric Society Monograph. Cambridge University Press, 1998.

[23] Elsa Chacko and Shoba Ranganathan. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. *BMC Genomics*, 10(Suppl 1):S5, 2009.

[24] Mark J Chaisson, Dumitru Brinza, and Pavel A Pevzner. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res*, 19(2):336–346, Feb 2009.

[25] Mark J Chaisson and Pavel A Pevzner. Short read fragment assembly of bacterial genomes. *Genome Res*, 18(2):324–330, Feb 2008.

[26] Anne-Laure Chateigner-Boutin and Ian Small. Plant RNA editing. *RNA Biol*, 7(2), Mar 2010.

[27] Wei Chen, Vera Kalscheuer, Andreas Tzschach, Corinna Menzel, Reinhard Ullmann, Marcel Holger Schulz, Fikret Erdogan, Na Li, Zofia Kijas, Ger Arkesteijn, Isidora Lopez Pajares, Margret Goetz-Sothmann, Uwe Heinrich, Imma Rost, Andreas Dufke, Ute Grasshoff, Birgitta Glaeser, Martin Vingron, and H. Hilger Ropers. Mapping translocation breakpoints by next-generation sequencing. *Genome Res*, 18(7):1143–1149, Jul 2008.

[28] Tyson A Clark, Anthony C Schweitzer, Tina X Chen, Michelle K Staples, Gang Lu, Hui Wang, Alan Williams, and John E Blume. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol*, 8(4):R64, 2007.

[29] Tyson A Clark, Charles W Sugnet, and Manuel Ares. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, 296(5569):907–910, May 2002.

[30] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nat Nanotechnol*, 4(4):265–270, Apr 2009.

[31] Nicole Cloonan, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, Alan J Robertson, Andrew C Perkins, Stephen J Bruce, Clarence C Lee, Swati S Ranade, Heather E Peckham, Jonathan M Manning, Kevin J McKernan, and Sean M Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 5(7):613–619, Jul 2008.

[32] Nicole Cloonan and Sean M Grimmond. Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol*, 9(9):234, 2008.

[33] Nicole Cloonan, Qinying Xu, Geoffrey J Faulkner, Darrin F Taylor, Dave T P Tang, Gabriel Kolle, and Sean M Grimmond. RNA-MATE: a recursive

mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics*, 25(19):2615–2616, Oct 2009.

[34] Mark J. Coldwell and Simon J. Morley. Specific Isoforms of Translation Initiation Factor 4GI Show Differences in Translational Activity . *Mol. Cell. Biol.*, 26(22):8448–8460, 2006.

[35] Lesley J Collins, Patrick J Biggs, Claudia Voelckel, and Simon Joly. An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Inform*, 21:3–14, 2008.

[36] David Cork, Thomas Lennard, and Alison Tyson-Capper. Alternative splicing and the progesterone receptor in breast cancer. *Breast Cancer Res*, 10(3):207, May 2008.

[37] Eivind Coward, Stefan A. Haas, and Martin Vingron. Splicenest: visualizing gene structure and alternative splicing based on est clusters. *Trends in Genetics*, 18(1):53 – 55, 2002.

[38] Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, Stanley J Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res*, 33(20):e175, 2005.

[39] Debopriya Das, Tyson A Clark, Anthony Schweitzer, Miki Yamamoto, Henry Marr, Josh Arribere, Simon Minovitsky, Alexander Poliakov, Inna Dubchak, John E Blume, and John G Conboy. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res*, 35(14):4845–57, 2007.

[40] Ramana V Davuluri, Yutaka Suzuki, Sumio Sugano, Christoph Plass, and Tim H-M Huang. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet*, 24(4):167–177, Apr 2008.

[41] Nicolaas Govert de Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946.

[42] A. P. Dempster and D. B. Rubin. Maximum-likelihood from incomplete data via the em-algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[43] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucl. Acids Res.*, 36(16):e105–, 2008.

[44] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[45] Eduardo Eyras, Mario Caccamo, Val Curwen, and Michele Clamp. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res*, 14(5):976–987, May 2004.

[46] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222, Jan 2010.

[47] Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney,

Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M J Searle. Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue):D557–D562, Jan 2010.

[48] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nat Methods*, 6(11 Suppl):S6–S12, Nov 2009.

[49] L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, 8(9):967–974, Sep 1998.

[50] Sylvain Foissac and Michael Sammeth. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res*, 35:W297–W299, 2007.

[51] Pim J French, Justine Peeters, Sebastiaan Horsman, Elza Duijm, Ivar Siccama, Martin J van den Bent, Theo M Luider, Johan M Kros, Peter van der Spek, and Peter A Sillevis Smitt. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res*, 67(12):5635–42, Jun 2007.

[52] Paul J Gardina, Tyson A Clark, Brian Shimada, Michelle K Staples, Qing Yang, James Veitch, Anthony Schweitzer, Tarif Awad, Charles Sugnet, Suzanne Dee, Christopher Davies, Alan Williams, and Yaron Turpaz. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7:325, 2006.

[53] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[54] Gao Guo, Sebastian Bauer, Jochen Hecht, Marcel H Schulz, Andreas Busche, and Peter N Robinson. A short ultraconserved sequence drives transcription

from an alternate FBN1 promoter. *Int J Biochem Cell Biol*, 40(4):638–650, 2008.

[55] S Gupta, D Zink, B Korn, M Vingron, and SA Haas. Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics*, 20(16):2579–85, Nov 2004.

[56] Alan E Guttmacher and Francis S Collins. Genomic medicine–a primer. *N Engl J Med*, 347(19):1512–1520, Nov 2002.

[57] Brian J Haas, Arthur L Delcher, Stephen M Mount, Jennifer R Wortman, Roger K Smith, Linda I Hannick, Rama Maiti, Catherine M Ronning, Douglas B Rusch, Christopher D Town, Steven L Salzberg, and Owen White. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 31(19):5654–5666, Oct 2003.

[58] Stefan A. Haas, Tim Beissbarth, Eric Rivals, Antje Krause, and Martin Vingron. Genenest: automated generation and visualization of gene indices. *Trends in Genetics*, 16(11):521 – 523, 2000.

[59] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucl. Acids Res.*, page gkq224, 2010.

[60] Takehiro Hashimoto, Michiel J L de Hoon, Sean M Grimmond, Carsten O Daub, Yoshihide Hayashizaki, and Geoffrey J Faulkner. Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite.9750195. *Bioinformatics*, 25(19):2613–2614, Oct 2009.

[61] Graham A Heap, Jennie H M Yang, Kate Downes, Barry C Healy, Karen A Hunt, Nicholas Bockett, Lude Franke, Patrick C Dubois, Charles A Mein, Richard J Dobson, Thomas J Albert, Matthew J Rodesch, David G Clayton, John A Todd, David A van Heel, and Vincent Plagnol. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*, 19(1):122–134, Jan 2010.

[62] Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18(Supplement):S181–188, 2002.

[63] Roberto Hirochi Herai and Michel E Beleza Yamagishi. Detection of human interchromosomal trans-splicing in sequence databanks. *Brief Bioinform*, 11(2):198–209, Mar 2010.

[64] David Hernandez, Patrice François, Laurent Farinelli, Magne Osterås, and Jacques Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res*, 18(5):802–809, May 2008.

[65] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, Sep 2009.

[66] Mohammad Sajjad Hossain, Navid Azimi, and Steven Skiena. Crystallizing short-read assemblies around seeds. *BMC Bioinformatics*, 10 Suppl 1:S16, 2009.

[67] Brian Howard and Steffen Heber. Towards reliable isoform quantification using RNA-SEQ data. *BMC Bioinformatics*, 11(Suppl 3):S6, 2010.

[68] X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome Res*, 9(9):868–877, Sep 1999.

[69] Daniel H. Huson, Knut Reinert, and Eugene Myers. The greedy path-merging algorithm for sequence assembly. In *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology*, pages 157–163, New York, NY, USA, 2001. ACM.

[70] R. M. Idury and M. S. Waterman. A new algorithm for DNA sequence assembly. *J Comput Biol*, 2(2):291–306, 1995.

[71] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[72] Benjamin G Jackson, Patrick S Schnable, and Srinivas Aluru. Parallel short sequence assembly of transcriptomes. *BMC Bioinformatics*, 10 Suppl 1:S14, 2009.

[73] David B Jaffe, Jonathan Butler, Sante Gnerre, Evan Mauceli, Kerstin Lindblad-Toh, Jill P Mesirov, Michael C Zody, and Eric S Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*, 13(1):91–96, Jan 2003.

[74] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, Apr 2009.

[75] Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653):2141–4, Dec 2003.

[76] W. Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X. Shirley Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–12462, Aug 2006.

[77] Karen Kapur, Yi Xing, Zhengqing Ouyang, and Wing Hung Wong. Exon arrays provide accurate assessments of gene expression. *Genome Biol*, 8(5):R82, 2007.

[78] W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[79] Roscoe Klinck, Anne Bramard, Lyna Inkel, Geneviève Dufresne-Martin, Julien Gervais-Bird, Richard Madden, Eric R Paquet, ChuShin Koh, Julian P Venables, Panagiotis Prinos, Manuela Jilaveanu-Pelmus, Raymund Wellinger, Claudine Rancourt, Benoit Chabot, and Sherif Abou Elela. Multiple alternative splicing markers for ovarian cancer. *Cancer Res*, 68(3):657–63, Feb 2008.

[80] Alberto R Kornblihtt, Manuel de la Mata, Juan Pablo Fededa, Manuel J Munoz, and Guadalupe Nogues. Multiple links between transcription and splicing. *RNA*, 10(10):1489–1498, Oct 2004.

[81] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85(4):457–464, Oct 2009.

[82] Vincent Lacroix, Michael Sammeth, Roderic Guigo, and Anne Bergeron. Exact Transcriptome Reconstruction from Short Sequence Reads. In *Lecture Notes in Computer Science: Algorithms in Bioinformatics*, volume 5251, pages 50–63, 2008.

[83] E. S. Lander, L. M. Linton, and B. Birren et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[84] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[85] Keith Le, Katherine Mitsouras, Meenakshi Roy, Qi Wang, Qiang Xu, Stanley F Nelson, and Christopher Lee. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*, 32(22):e180, 2004.

[86] Christopher Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19(8):999–1008, May 2003.

[87] Christopher Lee and Meenakshi Roy. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol*, 5(7):231, 2004.

[88] Christopher Lee and Qi Wang. Bioinformatics analysis of alternative splicing. *Brief Bioinform*, 6(1):23–33, Mar 2005.

[89] Y. Lee, J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung, and J. Quackenbush. The tigr gene indices: clustering and assembling est and known genes and integration with eukaryotic genomes. *Nucleic Acids Res*, 33(Database issue):D71–D74, Jan 2005.

[90] Jeremy Leipzig, Pavel Pevzner, and Steffen Heber. The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res*, 32:3977–3983, 2004.

[91] Joshua Z Levin, Michael F Berger, Xian Adiconis, Peter Rogov, Alexandre Melnikov, Timothy Fennell, Chad Nusbaum, Levi A Garraway, and Andreas Gnirke. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol*, 10(10):R115, 2009.

[92] Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, Feb 2010.

[93] Hairi Li, Michael T Lovci, Young-Soo Kwon, Michael G Rosenfeld, Xiang-Dong Fu, and Gene W Yeo. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A*, 105(51):20179–20184, Dec 2008.

[94] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[95] Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biology*, 11(5):R50, 2010.

[96] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, Mar 2008.

[97] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272, Feb 2010.

[98] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A. Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, May 2008.

[99] C. Lottaz, C. Iseli, C. V. Jongeneel, and P. Bucher. Modeling sequencing errors by combining hidden markov models. *Bioinformatics*, 19 Suppl 2:ii103–ii112, Oct 2003.

[100] Iain Maccallum, Dariusz Przybylski, Sante Gnerre, Joshua Burton, Ilya Shlyakhter, Andreas Gnirke, Joel Malek, Kevin McKernan, Swati Ranade, Terrance P Shea, Louise Williams, Sarah Young, Chad Nusbaum, and David B Jaffe. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol*, 10(10):R103, 2009.

[101] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, Mar 2009.

[102] Christopher A Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, Shujun Luo, Irina Khrebtukova, Terrence R Barrette, Catherine Grasso, Jindan Yu, Robert J Lonigro, Gary Schroth, Chandan Kumar-Sinha, and Arul M Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A*, 106(30):12353–12358, Jul 2009.

[103] Ketil Malde, Eivind Coward, and Inge Jonassen. A graph based algorithm for generating est consensus sequences. *Bioinformatics*, 21(8):1371–1375, Apr 2005.

[104] Tom Maniatis and Bosiljka Tasic. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, 418(6894):236–43, Jul 2002.

[105] Sayed-Amir Marashi, Changiz Eslahchi, Hamid Pezeshk, and Mehdi Sadeghi. Impact of rna structure on the prediction of donor and acceptor splice sites. *BMC Bioinformatics*, 7:297, 2006.

[106] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[107] John Marioni, Christopher Mason, Shrikant Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, Jun 2008.

[108] Paul Medvedev, Konstantinos Georgiou, Gene Myers, and Michael Brudno. Computability of models for sequence assembly. In *In WABI*, pages 289–301, 2007.

[109] Jason R Miller, Arthur L Delcher, Sergey Koren, Eli Venter, Brian P Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, Dec 2008.

[110] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, Jun 2010.

[111] R. T. Miller, A. G. Christoffels, C. Gopalakrishnan, J. Burke, A. A. Ptitsyn, T. R. Broveak, and W. A. Hide. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*, 9(11):1143–1155, Nov 1999.

[112] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, Apr 2010.

[113] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, May 2008.

[114] R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci*, 13(4):477–478, Aug 1997.

[115] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C.

Venter. A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204, Mar 2000.

[116] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:ii79–ii85, Sep 2005.

[117] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, Jun 2008.

[118] Shivashankar H Nagaraj, Robin B Gasser, and Shoba Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform*, 8(1):6–21, Jan 2007.

[119] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, Dec 2008.

[120] Qun Pan, Ofer Shai, Christine Misquitta, Wen Zhang, Arneet L Saltzman, Naveed Mohammad, Tomas Babak, Henry Siu, Timothy R Hughes, Quaid D Morris, Brendan J Frey, and Benjamin J Blencowe. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*, 16(6):929–941, Dec 2004.

[121] Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*, 37(18):e123, Oct 2009.

[122] Yu Peng, Henry Leung, S. Yiu, and Francis Chin. IDBA –A Practical Iterative de Bruijn Graph De Novo Assembler. *Research in Computational Molecular Biology*, pages 426–440, 2010.

[123] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*, 98(17):9748–9753, Aug 2001.

[124] Pavel A Pevzner, Paul A Pevzner, Haixu Tang, and Glenn Tesler. De novo repeat classification and fragment assembly. *Genome Res*, 14(9):1786–1796, Sep 2004.

[125] Andrey Ptitsyn and Winston Hide. Clu: a new algorithm for est clustering. *BMC Bioinformatics*, 6 Suppl 2:S3, Jul 2005.

[126] E. Purdom, K. M. Simpson, M. D. Robinson, J. G. Conboy, A. V. Lapuk, and T. P. Speed. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, June 2008.

[127] Josè Augusto Amgarten Quitzau and Jens Stoye. A space efficient representation for sparse de Bruijn subgraphs. Technical report, Technische Fakultät, Abteilung Informationstechnik / Universität Bielefeld, 2008.

[128] A. Rajkovic, R. E. Davis, J. N. Simonsen, and F. M. Rottman. A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. *Proc Natl Acad Sci U S A*, 87(22):8879–8883, Nov 1990.

[129] Hugues Richard, Marcel H Schulz, Marc Sultan, Asja Nürnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Res*, Feb 2010.

[130] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386, May 2009.

[131] Gunnar Rätsch, Sören Sonnenburg, and Christin Schäfer. Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, 7 Suppl 1:S9, 2006.

[132] Christian Rödelsperger, Sebastian Köhler, Marcel H Schulz, Thomas Manke, Sebastian Bauer, and Peter N Robinson. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, 94(5):308–316, Nov 2009.

[133] Michael Sammeth. Complete alternative splicing events are bubbles in splicing graphs. *Journal of Computational Biology*, 16(8):1117–1140, 2009. PMID: 19689216.

[134] Michael Sammeth, Sylvain Foissac, and Roderic Guigó. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, 4(8):e1000147, 08 2008.

[135] Rickard Sandberg, Joel R. Neilson, Arup Sarma, Phillip A. Sharp, and Christopher B. Burge. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, 320(5883):1643–1647, 2008.

[136] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

[137] Matthias Scherf, Anton Epple, and Thomas Werner. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform*, 6(3):287–297, Sep 2005.

[138] Jan Schröder, Heiko Schröder, Simon J Puglisi, Ranjan Sinha, and Bertil Schmidt. SHREC: a short-read error correction method. *Bioinformatics*, 25(17):2157–2163, Sep 2009.

[139] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W. L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. R. Hudson, S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James,

D. Bentley, P. Deloukas, E. S. Lander, and T. J. Hudson. A gene map of the human genome. *Science*, 274(5287):540–546, Oct 1996.

[140] Marcel H Schulz, Sebastian Bauer, and Peter N Robinson. The generalised k-truncated suffix tree for time-and space-efficient searches in multiple DNA or protein sequences. *Int J Bioinform Res Appl*, 4(1):81–95, 2008.

[141] Marcel H. Schulz, Sebastian Köhler, Sebastian Bauer, Martin Vingron, and Peter N. Robinson. Exact score distribution computation for similarity searches in ontologies. In *Lecture Notes in Computer Science: Algorithms in Bioinformatics*, volume 5724. Springer, 2009.

[142] Marcel H. Schulz, David Weese, Tobias Rasch, Andreas Döring, Knut Reinert, and Martin Vingron. Fast and adaptive variable order Markov chain construction. In *Lecture Notes in Computer Science: Algorithms in Bioinformatics*, volume 5251. Springer, 2008.

[143] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26(10):1135–1145, Oct 2008.

[144] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–1123, Jun 2009.

[145] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

[146] Charles W Sugnet, Karpagam Srinivasan, Tyson A Clark, Georgeann O'Brien, Melissa S Cline, Hui Wang, Alan Williams, David Kulp, John E Blume, David Haussler, and Manuel Ares. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol*, 2(1):e4, Jan 2006.

[147] Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, Aug 2008.

[148] B. J. Swanson, H. M. Jäck, and G. E. Lyons. Characterization of myocyte enhancer factor 2 (MEF2) expression in B and T cells: MEF2C is a B cell-restricted transcription factor in lymphocytes. *Mol Immunol*, 35(8):445–458, Jun 1998.

[149] T. A. Thanaraj and F. Clark. Human gc-ag alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res*, 29(12):2581–2593, Jun 2001.

[150] TA Thanaraj, S Stamm, F Clark, JJ Riethoven, TV Le, and J Muilu. ASD: The alternative splicing database. *Nucleic Acids Res*, 34:D64–D69, 2004.

[151] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[152] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[153] Ander Urruticoechea, Ian E Smith, and Mitch Dowsett. Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol*, 23(28):7212–7220, Oct 2005.

[154] J. C. Venter, M. D. Adams, and E. W. Myers et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[155] Ana P Vivancos, Marc Güell, Juliane C Dohm, Luis Serrano, and Heinz Himmelbauer. Strand-specific deep sequencing of the transcriptome. *Genome Res*, Jun 2010.

[156] Natalia Volfovsky, Brian J Haas, and Steven L Salzberg. Computational discovery of internal micro-exons. *Genome Res*, 13(6A):1216–1221, Jun 2003.

[157] Hiroyuki Wakaguri, Yutaka Suzuki, Toshiaki Katayama, Shuichi Kawashima, Eri Kibukawa, Kazushi Hiranuka, Masahide Sasaki, Sumio Sugano, and Junichi Watanabe. Full-Malaria/Parasites and Full-Arthropods: databases of full-length cDNAs of parasites and arthropods, update 2009. *Nucleic Acids Res*, 37(Database issue):D520–D525, Jan 2009.

[158] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[159] Hui Wang, Earl Hubbell, Jing shan Hu, Gangwu Mei, Melissa Cline, Gang Lu, Tyson Clark, Michael A Siani-Rose, Manuel Ares, David C Kulp, and David Haussler. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, 19 Suppl 1:i315–22, 2003.

[160] Liguo Wang, Yuanxin Xi, Jun Yu, Liping Dong, Laising Yen, and Wei Li. A statistical method for the detection of alternative splicing using RNA-seq. *PLoS One*, 5(1):e8529, 2010.

[161] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[162] Michael S. Waterman. *Introduction to computational biology: maps, sequences and genomes*. Chapman and Hall/CRC, June 1, 1995.

[163] David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. RazerS–fast read mapping with sensitivity control. *Genome Res*, 19(9):1646–1654, Sep 2009.

[164] David Weese and Marcel H. Schulz. Efficient string mining under constraints via the deferred frequency index. In *Lecture Notes in Computer Science: Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, volume 5077, pages 374–388, 2008.

[165] Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, Apr 2010.

[166] Huiyu Xia, Jianning Bi, and Yanda Li. Identification of alternative 5'/3' splice sites based on the mechanism of splice site competition. *Nucleic Acids Res*, 34(21):6305–6313, 2006.

[167] Yi Xing, Alissa Resch, and Christopher Lee. The multiassembly problem: reconstructing multiple transcript isoforms from est fragment mixtures. *Genome Res*, 14(3):426–441, Mar 2004.

[168] Yi Xing, Peter Stoilov, Karen Kapur, Areum Han, Hui Jiang, Shihao Shen, Douglas L Black, and Wing Hung Wong. MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, Jun 2008.

[169] Yi Xing, Tianwei Yu, Ying Nian Wu, Meenakshi Roy, Joseph Kim, and Christopher Lee. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*, 34(10):3150–60, 2006.

[170] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov 2009.

[171] Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11(2-3):377–394, 2004.

[172] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829, May 2008.

[173] Daniel R Zerbino, Gayle K McEwen, Elliott H Margulies, and Ewan Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*, 4(12):e8407, 2009.

[174] Haibo Zhang, Ju Youn Lee, and Bin Tian. Biased alternative polyadenylation in human tissues. *Genome Biol*, 6(12):R100, 2005.

Notation and Definitions

Here I list used abbreviations and definitions for quick reference.

Abbreviations

AEE	alternative exon event
APS	alternative polyadenylation site
AS	alternative splicing
CASI	cell type-specific alternative usage index
DASI	dfferential alternative usage index
DNA	deoxyribonucleic acid
EST	expressed sequence tag
FDR	fasle discovery rate
HEK	human embryonic kidney
mRNA	messenger RNA
NGS	next-generation sequencing
PAS	Polyadenylation site
PCC	Pearson's Correlation Coefficient
POEM	proportion estimation method
pre-mRNA	precursor mRNA
qRT-PCR	quantitative real time polymerase chain reaction
RNA	ribonucleic acid
RNA-Seq	RNA sequencing with NGS approaches
ROC	Receiver Operating Characteristic
RPKM	Reads Per Kilobase per Million mapped reads
RT-PCR	real time polymerase chain reaction
SG	splicing graph
TG	transcript de Bruijn graph
TSS	transcription start site

Chapter 3

T	theoretical number of reads in a gene
s	gene length
Y	observed number of reads in a gene
Y_e	observed number of reads in exon e
p	relative proportion of gene in sample
p_e	relative proportion of exon e in sample
$\mathcal{P}$	Poisson distribution
$\mathcal{M}$	Multinomial distribution
r	read length
j	splice junction length
i.i.d.	independently identically distributed
$\mathcal{J}$	total number of splice junction sequences
$\mathcal{R}$	total number of reads from an RNA-Seq experiment
$P(r, \sigma, j)$	probability of random match for read of length r to splice junction of length j with $\leq \sigma$ substitution errors
$E_{r,\sigma,j,\mathcal{J},\mathcal{R}}$	expected value of σ-error random matches for $\mathcal{R}$ reads of length r to $\mathcal{J}$ splice junctions of length j
l_e	effective exon length
ϕ_e	set of unique reads in e
$\widetilde{y}_e$	normalized expression of exon e
λ	sampling depth and gene length normalizing factor
z_e^C	log $\widetilde{y}_e$ based CASI z-score of exon e
z_e^D	DASI z-score of exon e based on the log-ratio between y_e^1 and y_e^2

Chapter 4

T	theoretical number of reads in the gene
T_j	theoretical number of reads in transcript j
t_j	observed number of reads in transcript j
$\hat{t}_j$	estimated number of reads in transcript j
k	number of transcripts of a gene
m	number of exons of a gene
$I_{e,j}$	binary indicator matrix such that $I_{e,j} = 1$ if isoform j uses exon e
Y_e^j	number of reads of isoform j in exon e
y_e^j	observed number of reads of isoform j in exon e
Y_e	number of reads in exon e
y_e	observed number of reads in exon e
q_j	gene-relative proportion of transcript j
$\mathcal{L}$	Likelihood

Chapter 5

k	dimension of the de Bruijn graph
$\mathcal{R}$	set of reads from an RNA-Seq experiment
R_k	read k
$\mathcal{T}$	set of expressed transcripts in an RNA-Seq experiment
T_i	i-th transcript sequence
w_{ij}	sum of edge weights that establish a connection between node i and node j
w_{ijk}	edge weights contributed by read or read pair k that establishes a connection between node i and node j
s_i	node weight of node i
f	scaling factor for the DP-algorithm
l_n	length of de Bruijn graph node n
G	Gene
ρ_n	read density in de Bruijn graph node n
$\mathcal{L}$	locus
nodes($\mathcal{L}$)	number of nodes in a locus $\mathcal{L}$
$\mathcal{TG}$	a transcript de Bruijn graph built from data

Zusammenfassung

Die molekulare biologische Forschung wurde durch die Erfindung der halbautomatisierten Sanger Sequenzierung für DNS in den frühen 1990er Jahren revolutioniert. Sie legte den Grundstein für die Sequenzierung von mehreren Genomen einschließlich des menschlichen Genoms. In den letzten Jahren hat es eine zweite Revolution im Bereich der DNS Sequenzierung gegeben. Die so genannten Next-Generation Sequencing (NGS) Verfahren erlauben die Sequenzierung von Millionen von DNS Fragmenten, in Form von kurzen "Reads", in weniger als einem Tag. Diese NGS Technologien sind noch nicht voll ausgereift und ihre weitere Entwicklung wird eine neue Ära in der DNS Sequenzierung einläuten, in der DNS Sequenzierung preiswert und einfach zu handhaben ist. Diese Entwicklung bedeutet einen entscheidenden Anstieg des Arbeitsaufwand von Bioinformatikern, die Gigabasen von Sequenzdaten bewältigen müssen, was derzeitig den Flaschenhals für wissenschaftliche Analysen mit NGS Daten darstellt.

Diese Dissertation beschäftigt sich mit den Herausforderungen, die sich auf Applikationen von NGS Technologien zum Sequenzieren von exprimierten mRNAs (RNA-Seq) beziehen. Im besonderen wird die Ermittlung von alternativen Exon Ereignissen (AEEs) betrachtet, was zusammenfassend steht für alternatives Spleißen, alternative Promotoren und alternative Polyadenylierungsereignisse. Im folgenden die drei wichtigsten Beiträge.

Zuerst werden Methoden eingeführt, die die Vorhersage von AEEs in einer oder zwischen zwei Konditionen, zum Beispiel krank gegen normal, ermöglichen. Diese Methoden basieren auf bestehender Genannotation und bereits genomisch platzierten RNA-Seq Reads. Alle Methoden basieren auf einem Poisson Modell, welches das zufällige Platzieren der Reads entlang der mRNA beschreibt. Die Methoden werden auf RNA-Seq Datensätze von einer humanen-embryonalen Niere (HEK) und einer humanen B-Zell Zelllinie angewendet. Mehrere Tausend AEEs wurden in diesen Zelllinien

vorhergesagt. Die Robustheit und Genauigkeit der Vorhersagen wurde durch Simulationen, Bootstrapping und RT-PCR Validierungsexperimente bestätigt. Darüber hinaus wurde ein Vergleich von den neuen Methoden für RNA-Seq und bestehenden Methoden für Exon Arrays durchgeführt, der erhöhte Sensitivität und Genauigkeit für die RNA-Seq basierten Vorhersagen offenbart.

Zweitens wird eine neue Methode für die Abschätzung von mRNA Expressions Leveln aus RNA-Seq Daten vorgeschlagen, basierend auf vorhandenen Transkriptannotationen und bereits genomisch platzierten RNA-Seq Reads. Die Methode basiert auf dem Expectation-Maximization Optimierungsverfahren. Die Korrektheit und theoretische Leistung des Ansatzes wird durch Simulationen demonstriert. Anwendung auf HEK und B-Zell RNA-Seq Daten und Vergleich mit Quantifizierung durch quantitative RT-PCR Experimente bestätigen die Genauigkeit der Methode.

Schlussendlich wird die erste Methode vorgestellt die es erlaubt ein *de novo* "Assembly" eines Transkriptoms eines Organismuses ausgehend von RNA-Seq Daten anzufertigen. Dies ist ein wichtiges Problem, welches funktionale Analysen und Genentdeckung für Organismen ermöglicht, von denen das Genom noch nicht sequenziert wurde. Ein Wechsel vom traditionellen Overlap-Layout-Consensus Paradigma zur Anwendung des Eulerpfad Ansatzes für Transkriptom Assembly wird vorgeschlagen, vergleichbar mit der Entwicklung für *de novo* Genom Assembly. Die Gemeinsamkeiten zwischen de Bruijn Graphen und Splicing Graphen wurden erforscht und eine Theorie für *de novo* Vorhersagen von AEEs entwickelt. Ausgehend von de Bruijn Graphen wurden Algorithmen entwickelt, die die Vorhersage von kompletten mRNA Sequenzen, unter Berücksichtigung von AEEs, ermöglichen. Die Anwendung auf reellen RNA-Seq Daten demonstriert die Verbesserung des neuen Ansatzes im Vergleich zur Anwendung von *de novo* Genom Assemblierungsprogrammen, die bis dato für RNA-Seq Datensätze benutzt wurden. Für einen RNA-Seq Datensatz einer Maus Zelllinie mit ungefähr 67 Mio. Reads wurde ein Assembly von insgesamt 63 Megabasen erstellt . Dieses Assembly beinhaltet ungefähr 6,900 mRNAs in vollständiger Länge, welches den Erfolg des Ansatzes untermauert.

Summary

Research in molecular biology was revolutionized by the invention of semi-automated Sanger sequencing for DNA in the early 1990's. It was the foundation for the sequencing of several genomes including the human genome. In the last few years a second revolution in the field of DNA sequencing has occurred that has changed the field. Next-generation sequencing (NGS) approaches suddenly enable the sequencing of millions of DNA fragments leading to short sequencing reads in less than a day. These NGS technologies are still in its infancy and their further development will herald a new era where DNA sequencing is inexpensive and easily manageable. This development has shifted the largest proportion of the workload onto the workbench of the computational biologist that has to cope with gigabases of sequence data, creating a bottleneck for scientific discovery.

This thesis deals with the challenges related to the application of Next-generation sequencing (NGS) technologies to the sequencing of expressed mRNAs (RNA-Seq) and the detection of alternative exon events (AEEs), summarizing alternative splicing, alternative promoter, and alternative polyadenylation events. There are three main contributions.

First, methods are introduced that enable the detection of AEEs within or between conditions, e.g. disease and normal, based on given gene annotation and mapped RNA-Seq reads. All methods are based on a Poisson model that describes the random placement of reads along a transcript. The methods are applied to a dataset from a human embryonic kidney (HEK) and a B cell line. Several thousand AEEs were predicted in these cell lines. The robustness and correctness of the predictions was assessed by simulations, bootstrapping, and RT-PCR validation experiments. In addition, a comparison of splicing prediction by RNA-Seq with prediction from exon arrays shows higher sensitivity and accuracy for RNA-Seq based predictions.

Second, a new method for inferring isoform expression levels from RNA-Seq data is proposed, given annotated isoform structures and mapped read information. The method is based on the Expectation-Maximization framework. The theoretical power of the approach is demonstrated with simulations. Application to HEK and B cell RNA-Seq data and comparison to isoform expression quantification with quantitative RT-PCR experiments show the accuracy of the porposed method.

Finally, the first method that allows the *de novo* assembly of an organism's transcriptome from short read RNA-Seq data is presented, an important problem that enables functional analysis and gene discovery when the genome of an organism was not sequenced yet. A transition from the traditional Overlap-Layout-Consensus paradigm to the Eulerian path approach to transcriptome assembly is made, similar to the development for *de novo* genome assembly. The similarities between de Bruijn graphs and splicing graphs are explored and a theory for *de novo* prediction of AEEs is developed. Further, algorithms for the assembly of full length sequences considering alternative gene isoforms are designed. An application to real data demonstrates the improvement compared to *de novo* genome assemblers that have been utilized so far for RNA-Seq datasets. For a mouse RNA-Seq dataset with approximately 67 Mio. reads a total output of 63 megabases of transcript sequences is assembled of which approximately 6,900 are full-length mRNAs, underlining the success of the approach with few lanes of sequencing.

Software Availability

The statistical estimators for prediction of alternative exon usage within a condition (CASI) and between conditions (DASI) as well as the quantification method for transcript structures (POEM) together with utility functions are made available in the open-source R-package Solas at
`http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/`.

The transcriptome assembler Oases is open source and an addition to the Velvet assembler, availabe at `http://www.ebi.ac.uk/~zerbino/oases/`.

Appendix

Supplemental Table Legends

Supplemental Table S0A: These tables contain all splice junctions with their position mapping on genes (Ensembl v.46) and ESTs (EMBL NSD, release 89) that were found in B cells. Columns "Novel" and "Alternative" indicate whether the identified junction is new and if the junction directly identifies an alternative isoform of the gene, respectively .

Supplemental Table S0B: These tables contain all splice junctions with their position mapping on genes (Ensembl v.46) and ESTs (EMBL NSD, release 89) that were found in Hek cells. Columns "Novel" and "Alternative" indicate whether the identified junction is new and if the junction directly identifies an alternative isoform of the gene, respectively.

Supplemental Table S1A: This table provides all genes with their AEEs for the cell-internal analysis in B cells. It lists the p-value, CASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in B cells the unique junction identifier for junctions is associated to it.

Supplemental Table S1B: This table provides all genes with their AEEs for the cell-internal analysis in Hek cells. It lists the p-value, CASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in Hek cells the unique junction identifier for junctions is associated to it.

Supplemental Table S2: A listing of the primer sequences and results for the 61 CASI exons that were tested for exon skipping with RT-PCR. The table describes which events are validated using the primers S1-R1 and S2-R1.

Supplemental Table S3: This table contains a set of 73,948 reference alternative exons annotated to Ensembl version 46 genes. The set of exons have either evidence from the Ensembl database, as indicated by listing the corresponding Ensembl transcript ids. Or the exons are predicted to be alternatively spliced based on EST data, as indicated by their Unigene Cluster ID.

Supplemental Table S4: This table lists the estimated splice form proportions for 1,487 (640 genes) in B cells and 1,920 transcripts (830 genes) in HEK cells. It lists all transcripts for which proportion estimations were calculated by the means of either the Ensembl transcript ID or the merged transcript ID (with the prefix ENST-M) that was created after merging one or more Ensembl transcripts. For each transcript, the level of expression is listed (relative abundance, estimated splice form proportion) computed with POEM. From this probability and the absolute normalized expression, measured with RNA-Seq, the normalized expression was derived for each transcript.

Supplemental Table S5: A listing of the primer sequences and results for the 24 events that were tested for exon skipping with quantitative RT-PCR for POEM validation. All experiments were conducted using the S1-R1 and S2-R1 primer pairs and inclusion rates of the S1 form computed from the qPCR replicates are listed. In addition, S1 inclusion rates derived from junction read counts and computed from POEM are listed (Junc and POEM).

Supplemental Table S6: This table provides all genes with their AEEs for the differential analysis between Hek and B cells. It lists the p-value, DASI value of the exons and additional gene information. For each exon that is supported by an alternative splice junction in either B or Hek cells the unique junction identifier for junctions is associated to it.

Supplemental Table S7: A listing of the primer sequences and results for the 16 DASI exons that were tested for exon skipping with quantitative RT-PCR. All experiments were conducted using the S1-R1 and S2-R1 primer pairs and inclusion rates computed from the qPCR replicates are listed. The column "validated" depicts the cases which lie inside the 1.5 ratio-difference interval. The DASI value, the MIDAS detection value, and the Splicing Index is given for every exon.

Supplemental Data S8A: Single-k(=19) assembly results of Oases applied to Human CD4 data. The Oases transfrags were aligned with Blat against the reference genome and the matches where transformed into the provided GTF file.

Supplemental Data S8B: Merged assembly results of Oases applied to Human CD4 data. The Oases transfrags were aligned with Blat against the reference genome and the matches where transformed into the provided GTF file.

Supplemental Data S8C: Single-k(=23) assembly results of Oases applied to C2C12 Mouse data. The Oases transfrags were aligned with Blat against the reference genome and the matches where transformed into the provided GTF file.

Supplemental Data S8D: Merged assembly results of Oases applied to C2C12 Mouse data. The Oases transfrags were aligned with Blat against the reference genome and the matches where transformed into the provided GTF file.

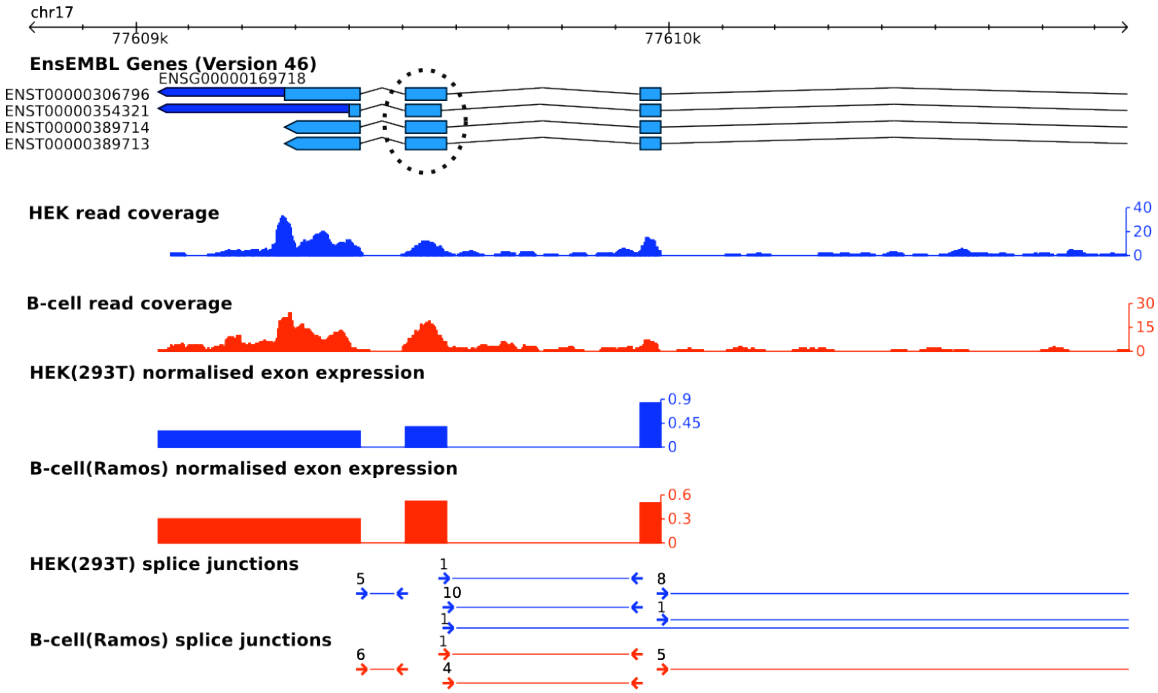

Figure 6.1: Alternative acceptor splice site usage in the *DUS1L* gene. According to the annotation (Ensembl v.46), the circled exon in *DUS1L* shows an alternative acceptor splice site. This exon was predicted as AEE by the CASI method and had further one junction read identifying this event. The view was zoomed in to see the offset of the alternative splice junction.

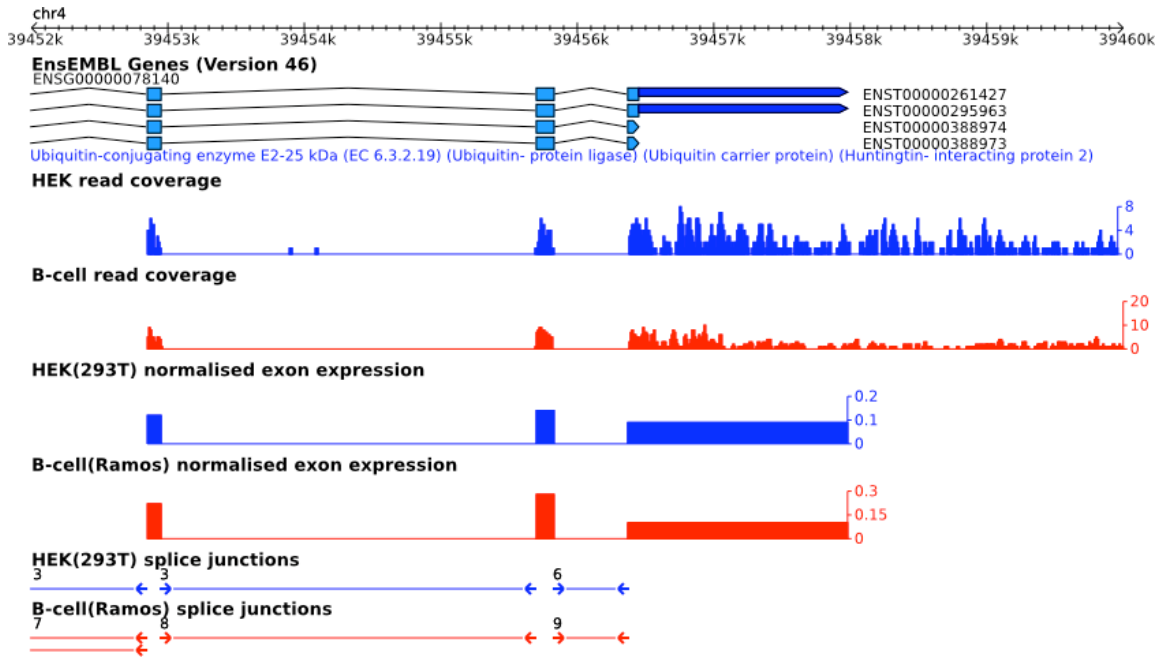

Figure 6.2: Evidence of an alternative polyadenylation site in *HIP2*. The figure shows the 3' end of *HIP2* as annotated in Ensembl v.46 and the read coverage obtained by RNA-Seq for HEK (blue) and B (red) cells. In B cells, the drop in read coverage along the last exon suggests the presence of at least two alternatively polyadenylated forms specific to this cell line. *HIP2* was previously shown to be alternatively polyadenylated in proliferating T cells by Sandberg and colleagues [135].

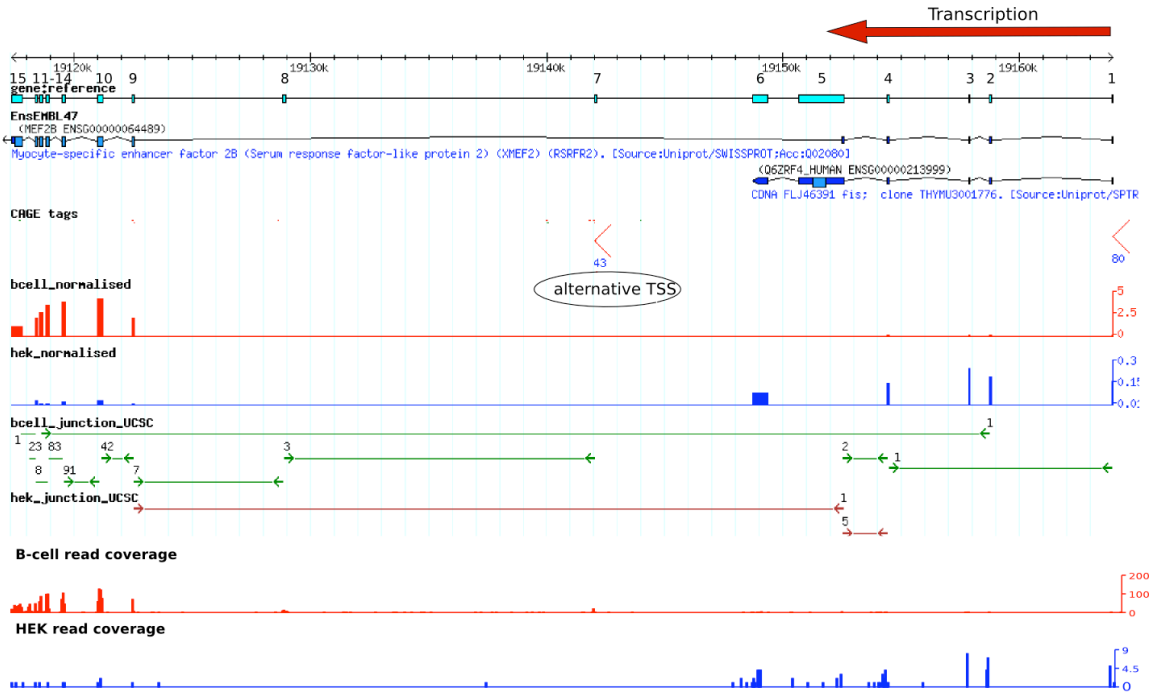

Figure 6.3: This example shows that DASI enables the identification of alternative promoter usage events. The gene structure at the top is derived from EST data (Genenest cluster Hs78881) [55] and includes exon 7 and 8, which were not annotated in the Ensembl database (v.46). The expression profile in both cell lines suggests an alternative transcript starting at exon 7, which is highly expressed in B cells and not in HEK cells. This is supported by CAGE tag evidences nearby exon 7, suggesting the presence of an alternative transcription start site.

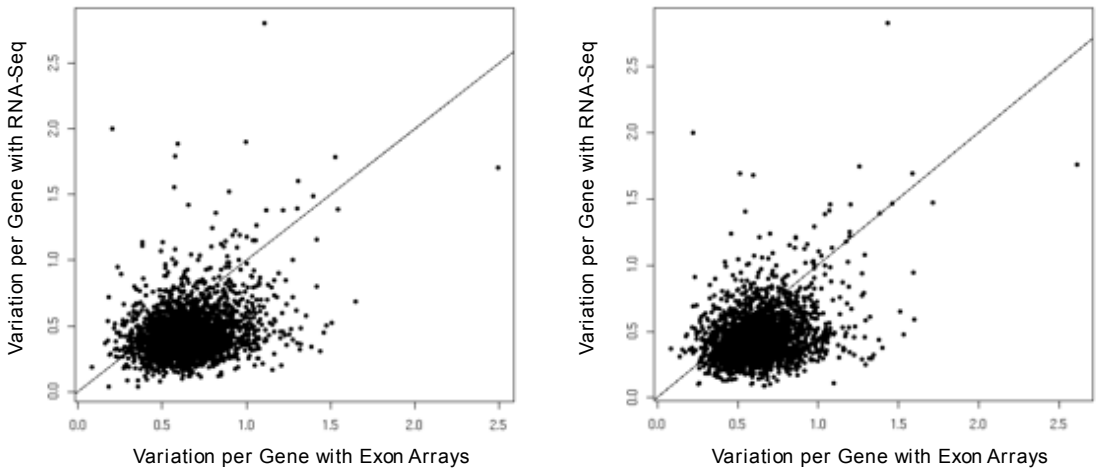

Figure 6.4: Scatterplots of the gene-wise normalized standard deviation (coefficient of variation) of exon expression values in B cells (left) and HEK cells (right) computed for exon arrays (x-axis) and RNA-Seq (y-axis). The coefficient of variation is consistently larger for exon array values in both cell lines (Wilcoxon, p-value<2.2 e-16).

Table 6.1: List of the top 20 genes with alternatively regulated exons as detected by the DASI method between HEK and B cells.

Symbol	Gene Description
MEF2B	Myocyte-specific enhancer factor 2B (Serum response factor-like protein.2)
PTPRCAP	Coronin-1B (Coronin-2)
SEPT9	Septin-9
CTNND1	Catenin delta-1
SMARCB1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily B member 1
RPL37	60S ribosomal protein L37 (G1.16)
RPS2	40S ribosomal protein S2 (S4)
NASP	Nuclear autoantigenic sperm protein
BBS1	Dipeptidyl-peptidase 3 (EC.3.4.14.4)
P2RX5	P2RX5 Tax1-binding protein 3
HLA-G	HLA class I histocompatibility antigen alpha chain G precursor (HLA G antigen)
RPS4X	40S ribosomal protein S4, X isoform
MAZ	Myc-associated zinc finger protein (MAZI) (Purine-binding transcription factor
TREX1	ATR-interacting protein
WNK2	Serine/threonine-protein kinase WNK2 (EC.2.7.11.1)
RPS24	40S ribosoma protein S24
LDHA	Lactate dehydrogenase A chain (EC.1.1.1.27)
SCD	Acyl-CoA desaturase (EC.1.14.19.1) (Stearoyl-CoA desaturase)
CDKN2A	Cyclin-dependent kinase inhibitor 2A, isoform 4 (p14ARF)
GNB1	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta 1

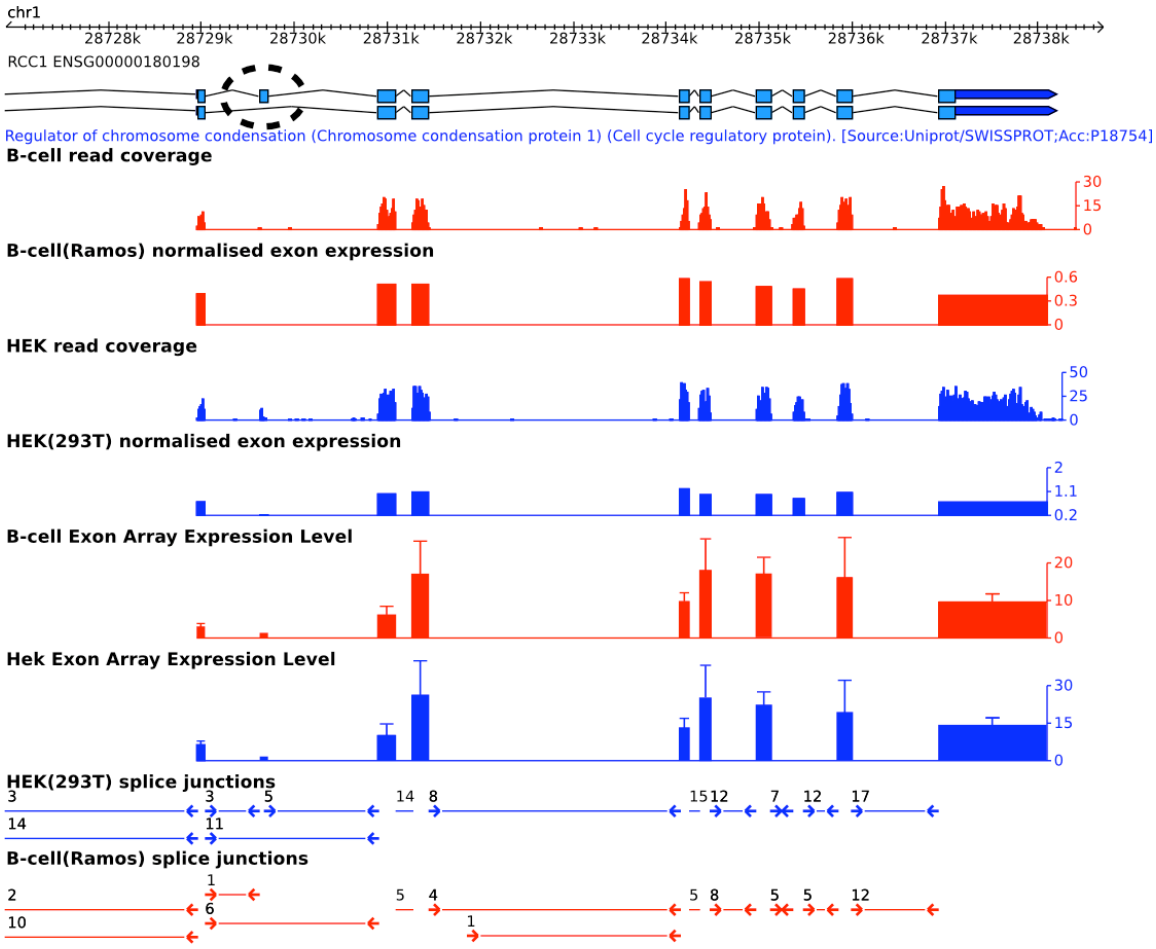

Figure 6.5: The circled exon in *RCC1* is detected as alternatively spliced by the DASI method and verified by qPCR. However, it was not detected by the MIDAS method in exon arrays as the expression of this exon was below the background noise in both cell lines. The whiskers on the "Exon Array Expression Level" tracks denote the standard deviation as measured between the replicates.

Lebenslauf

Marcel Holger Schulz
Gotenstrasse 38
10829 Berlin
Tel (030) 8431 6966
marcel.schulz@molgen.mpg.de

geboren am 31.12.1981
Geburtsort: Berlin
Staatsangehörigkeit: Deutsch
Familienstand: ledig

Akademischer Grad

Bachelor of Science

am	19. August 2005
Betreuer:	Dr. S. Röpcke (MPI-MG Berlin, jetzt Nycomed)
	Prof. Dr. K. Reinert (Freie Universität Berlin)
Thema:	Characterisation of Mitochondrial Ribosomal Protein Gene Promoters

Ausbildung

seit 10/2005	IMPRS-CBSC Doktorand in der Abteilung Computational Molecular Biology am Max-Planck-Institut für molekulare Genetik, Berlin
09/2005-10/2005	IMPRS-CBSC Phd Preparatory Program an der Freien Universität Berlin
10/2002-08/2005	Bachelorstudium Bioinformatik an der Freien Universität Berlin
07/2001-05/2002	Zivildienst Lankwitzer Werkstätten Berlin
06/1995-06/2001	Willi-Graf Gymnasium, Berlin
06/1994-06/1995	5. Gymnasium, Berlin Marzahn

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, Juni 2010