COMPUTATIONAL DISCOVERY OF SPLICING EVENTS FROM HIGH-THROUGHPUT OMICS DATA

by

Yen-Yi Lin

M.Sc., National Taiwan University, 2008
B.Sc., National Taiwan University, 2003

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Sciences
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2017

Approval

Name: **Yen-Yi Lin**

Degree: **Doctor of Philosophy (Computing Sciences)**

Title: **COMPUTATIONAL DISCOVERY OF SPLICING EVENTS FROM HIGH-THROUGHPUT OMICS DATA**

Examining Committee:

Chair: Faraz Hach
University Research Associate

Cenk Sahinalp
Senior Supervisor
Professor

Martin Ester
Senior Supervisor
Professor

Peter Unrau
Supervisor
Professor
Dept. of Molecular Biology and Biochemistry

Leonid Chindelevitch
Internal Examiner
Assistant Professor

Vineet Bafna
External Examiner
Professor
Computer Science and Engineering
University of California, San Diego

Date Defended: **July 18, 2017**

Abstract

The splicing mechanism, the process of forming mature messenger RNA (mRNA) by only concatenating exons and removing introns, is an essential step in gene expression. It allows a single gene to have multiple RNA isoforms which potentially code different proteins. In addition, aberrant transcripts generated from non-canonical splicing events (e.g. gene fusions) are believed to be potential drivers in many tumor types and human diseases. Thus, identification and quantification of expressed RNAs from RNA-Seq data become fundamental steps in many clinical studies. For that reason, number of methods have been developed. Most popular computational methods designed for these high-throughput omics data start by analyzing the datasets based on existing gene annotations. However, these tools (i) do not detect novel RNA isoforms and low abundance transcripts; (ii) do not incorporate multi-mapping reads in their read counting strategies in quantifications; (iii) are sensitive to sequencing artifacts.

In this thesis, we will address these computational problems for analyzing splicing events from high-throughput omics data. For identification and quantification of expressed RNAs from RNA-Seq data, we introduce CLIIQ, a unified framework to solve these two problems simultaneously. This framework also supports data from multiple samples to improve accuracy. To better incorporate multi-mapping reads into the framework, we design ORMAN, a combinatorial optimization formulation to resolve their mapping ambiguity by assigning single best location for each read. For aberrant transcript detections, we present a computational strategy ProTIE to integratively analyze proteomics and transcriptomic data from the same individual. This strategy provides proteome-level evidence for aberrant transcripts that can be used to eliminate false positives reported solely based on sequencing data.

Keywords: Transcriptome Reconstruction; Alternative Splicing; Genomic Aberrations; RNA-Seq; High-Throughput Sequencing; Proteomics; Proteogenomics; Cancer

Dedicated to my parents, Yu-Hui Tseng and En-Hwa Lin,
and my wife, You-Min Lin

It eluded us then, but that's no matter - tomorrow we will run faster,
stretch out our arms farther.... And one fine morning –
So we beat on, boats against the current, borne back ceaselessly into the past.

– F. Scott Fitzgerald, *The Great Gatsby*

Acknowledgements

First and foremost, I extend my sincerest and utmost gratitude to my senior supervisor, Dr. Cenk Sahinalp, who has supported me throughout my graduate studies with his encouragement, guidance and resourcefulness. In addition to technical skills in computer science, he also taught me the requirements of being a good researcher. I thank him one more time for all his help.

I would like to give my regards and appreciations to Dr. Martin Ester, my co-senior supervisor who inspires me through his questions on fundamental, siginificant, and yet challenging topics. I also thank my supervisor, Dr. Peter Unrau, who keeps reminding me that a healthy and sustainable research career should be guided by passion. I would like to sincerely thank Dr. Leonid Chindelevitch and Dr. Vineet Bafna, who graciously accepted to be my examiner and provided their valuable comments to improve my thesis. I give special thanks to Dr. Faraz Hach, who kindly accepted to be the chair of my examining committee.

I would also like to thank my collaborators Dr. Colin Collins, Dr. Haixu Tang, Dr. Can Alkan, Dr. Phuong Dao, Dr. Ibrahim Numanagic, Dr. Nilgun Donmez, Dr. Pinar Kavak, Dr. Sujun Li, Dr. Milan Radovich, Dr. Fan Mo, and Alex Gawronski. I benefited from these collaborations, and hope to continue working with talented scholars.

In these years I learned a lot from my labmates to whom I owe a great deal: Dr. Fereydoun Hormozdiari, Dr. Iman Hajirasouliha, Raunak Shrestha, Lucas Swanson, Iman Sarrafi, Salem Malikic, Ermin Hodzic, Can Kockan, Ehsan Haghshenas, and Hossein Asghari.

Last, but not least, I am grateful to my family - to mom, dad, my sister, and my wife - for their support through my gradudate studies in all these years. They've given me everything I asked for, although they may never fully understand what this thesis is about.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent advances in high-throughput sequencing (HTS) and mass spectrometry (MS) technologies have expanded our understanding of human genome, and changed our view regarding the complexity of human transcriptome and proteome. This newly discovered information is continuously integrated into the online genome database such as Ensembl [32] and UCSC Genome Browser [148].

Much biomedical research starts by analyzing high-throughput omics data based on gene annotation information, assuming that a reasonable number of sequenced reads or mass spectra can be explained by annotated events. The assumption leads to two potential issues. First, the use of a reference genome, transcriptome, or proteome might introduce biases in detecting events which are not included in the gene annotations. Second, allowance of aberrations in computational methods which rely on gene annotations can easily lead to lots of falsely detected events. To provide accurate molecular profiling for datasets with aberrations, such as tumor samples, it is of great importance to have computational methods which are not limited by gene annotations, and can integrate all available omics datasets to validate potential aberrations.

The study of transcriptomics plays a key role in obtaining molecular profiles for a patient by bridging genomics and proteomics. The classic view of the central dogma of biology indicates that DNA is transcribed into messenger RNA which will be translated into proteins [22]. The splicing mechanism in transcription enables human cells to have vast transcripts and therefore protein diversity from a limited number of genes. In human genes with multiple exons, the intervening regions of introns should be spliced out from the nascent transcripts so that exons can be concatenated into an mRNA. A gene can typically have multiple splicing forms as long as the involving exons concatenate in a way consistent to their relative order in a reference genome, which is known as alternative splicing [9]. Alternative splicing is a ubiquitous mechanism of gene expression which allows a gene to have multiple mRNA isoforms, the set of different mRNA transcripts from the same gene. These mRNA transcripts generated by alternative splicing can differ in either their untranslated

regions or their coding sequence through exon skipping, a choice between mutually exclusive exons, or alternative splicing sites.

In addition to mRNA isoforms generated by alternative splicing, aberrant transcripts from non-canonical RNA splicing events are also observed in a growing number of human diseases [126]. Non-canonical splicing can produce aberrant mRNA transcripts by including non-annotating exons or applying mechanisms that are different from well-defined splicing rules [131]. For the first type, non-canonical splicing events generate transcripts with cryptic splice sites and exons that are not annotated in the current genome databases, microexons that are shorter than 30nt, and recursive splice sites that introduce multiple stage of intron-removal in forming mRNAs. For the second-type, aberrant transcripts are formed due to abnormal splicing efficiency (intron retention, inclusion of exonic introns, or exitrons that are usually considered as exons), unconventional order of exon concatenation (circular RNAs) [139], atypical splice sites, and chimeric RNAs such as gene fusions.

The study of transcriptomics aims to provide accurate profiling for transcriptome, the complete set of all expressed transcripts from both alternative and non-canonical splicing events in a tissue [155]. The main goals of transcriptome profiling are: (1) to identify all species of transcripts from alternative splicing, including mRNAs, non-coding RNAs, and small RNAs,(2) to quantify the expressions of the above splicing isoforms so that we can estimate changing levels for genes under different conditions, (3)to detect aberrant transcripts from non-canonical splicing, especially those disease-relevant events such as gene fusions. Accurate transcriptome profiling is essential for understanding the underlying mechanism of cancers and diseases, including cancer development, progressions, and resistance to therapy. It also provides important information for us to study the corresponding proteome, especially for detecting novel proteins translated from unannotated splicing events.

RNA-Seq technology has become the standard approach for profiling and studying transcriptome. For a specific sample, RNA-Seq protocol first converts all transcripts to a library of their complementary DNA (cDNA) fragments, and then applies DNA sequencer on this library to obtain reads as the RNA-Seq data for the sample. Compared to previous array-based technology, RNA-Seq provides the following advantages: (1) It allows us to detect transcripts for species without genomic sequences or annotations. (2) It allows us to identify transcripts and their transcriptional structure in base-pair resolution. (3) It improves the ability of capture expressed transcripts by providing a larger dynamic range of detectable expressions levels. In the meantime, the following properties of RNA-Seq reads also raise challenges for designing computational methods in transcriptome profiling: (1) Reads are noncontiguous in a genome due to splicing events. This will increase the difficulty of read mappings. (2) Reads can be shared by multiple isoforms in a gene. This increases the difficulty in both identification and quantification of transcripts. (3) Uneven coverage of reads in a transcript, even though it is the only known mRNA is a gene, due to various sources of biases [96, 74]. Since RNA-Seq provides a larger dynamic range of expressed events, the

uneven coverage across a transcript will make it very difficult to distinguish between low-expressed events and sequencing errors. This property is quite different from whole genome sequencing datasets, and it significantly affects the accuracy of recognizing low-abundant transcripts.

Given an RNA-Seq dataset, transcriptome profiling usually starts by either de novo assembly of full-length transcripts [45, 125, 114, 163, 108, 140] or mapping reads back to the reference genome [26, 146, 60, 159]. De novo transcriptome assemblers construct a set of contigs as potential full-length or partial transcripts by analyzing RNA-Seq reads without relying on the reference genome. Most de novo assembler are designed based on de Bruijn graph [25] and using different strategies in selecting single or multiple k-mer lengths for constructing contigs. De novo assembly has its edge when high-quality reference genome is unavailable, or when the analysis targets on novel events which might be missing or highly different from the reference. Its performance can be affected by the uneven coverage within the single transcript and coverage variations across the whole transcriptome.

For species with high-quality reference genome, such as human, transcriptome profiling starting with read mapping usually provides better results in identifying and quantifying isoforms compared to those starting with de novo assembly. The analysis usually starts by mapping RNA-Seq to reference genome using splice-aware mappers with or without information of gene models or gene annotations which describe splice sites and exon boundaries for all genes. The challenges for splice-aware mapping comes from the fact that many RNA-seq reads are formed by concatenating multiple exons which are noncontiguous in the genome. Precise determination of splice sites within a read and their corresponding location on the genome can be a challenging task. In addition, incomplete annotations or unannotated splicing events can both lead to unmapped reads or mappings whose splicing sites are incompatible with gene annotations [30, 11].

Based on RNA-Seq mapping results, the first task in transcriptome profiling is to identify all expressed transcripts. The basic idea of transcript identification is to find a set of transcripts that cover all expressed exons and junctions. Even without considering novel events from non-canonical splicing, the number of potential isoforms is already exponential to the number of exons. Therefore, it is computationally intensive to enumerate all possible solutions for this problem, and in practice all methods either fully stick to known isoforms from gene annotations or will apply some heuristics to reduce the search space. When allowing novel isoforms, two popular models used for constructing potential transcripts are splicing graphs and overlapping graphs. A splicing graph [46] is a graph in which each vertex represents an exon and two vertices are adjacent if there are junction reads between the corresponding exons. Since a junction read usually supports a path containing two or three exons, after extension, a path on a splicing graph can potentially connect to exons which belong to different transcripts. Eventually, identification based on splicing results tends to produce many false positives and also affect isoform quantifications. An overlap-

ping graph [147] represents each mapped read as a vertex and adds an edge between two vertices if the mappings can co-exist in the same transcript. For two overlapping reads which can not exist in the same transcript, the algorithm either needs to ensure that there are different transcripts in the final solution to cover each mapping separately, or discard at least one of them. Therefore, identification results based on overlapping graph can sometimes remove true junction mappings and lead to false negative events.

After obtaining the set of expressed transcripts, we can quantify these transcripts based on the number of reads assigned to them. The challenges of isoform quantification problem mainly come from the following reasons. (1) Many RNA-Seq reads are shared by multiple isoforms having the same exon(s), (2) some RNA-Seq reads can be mapped to multiple genomic locations, and (3) Coverage may vary within the same transcript. There are two popular quantification strategies: EM-Based quantification algorithms quantify isoforms by k-mer profiling [107, 14, 106] or distributing a multi-mapping read into all possible mapping locations according to their relative abundances estimated by the unique mapping [70, 101]. The assigned weights for each transcript will eventually be converted into normalized values. While they resolve multi-mapping issue using a probabilistic way, these normalized values might affect the accuracy of the downstream analysis. Therefore, many pipelines for detecting differentially expressed genes [69, 3, 4, 115, 49] prefer quantification results from counting-based methods [5, 77], which just count the number of reads overlapping with a gene or a transcript. For multi-mapping reads, the counting-based quantification methods can either discard multi-mapping reads or consider all mappings as correct, while each read can only be sequenced from single genomic location.

Results of Isoform identification and quantification affect each other, and it should provide a better solution by considering these two issues simultaneously [137]. A general idea is that, for a specific gene, we first apply some heuristics to select a relatively small set of potential transcripts [35], and then incorporate quantification accuracy to find a solution which minimizes the differences between expected expression of exonic regions and observed expressions from mapping results. While it is general belief that most genes should contain a limited number of expressed isoforms, such formulations can lead to an over-fitting solution which minimizes the objective functions by predicting many low-expression transcripts with single exon. Most computational methods [75, 76, 109, 143, 144, 73, 92] therefore need to consider sparsity in the formulation [141, 21, 21] .

Transcript identification and quantification algorithms mentioned above are mainly designed for alternative spliced isoforms which concatenate exons according to their relative order in a genome. Heavily dependent on read mapping results, these algorithms do not perform well on studying non-canonical splicing events with transcriptional structures highly different from gene annotations. While non-canonical splicing events are relatively rare in the transcriptome, their appearance might indicate interruption of functions for the relevant genes, and therefore signatures of diseases and cancers [89]. For example, TMPRSS2-ERG

gene fusion is the predominant molecular subtype of prostate cancer and observed in almost 50% of prostate cancer patients [145]. The impacts on cancers make gene fusions the most studied non-canonical splicing events, and computational methods for detecting fusions start with either de novo assembly [138, 165] or sequence-mapping results [84, 61, 63, 122, 55, 43], similar to algorithms for detecting alternative splicing isoforms. While various prediction tools are developed for detecting these aberrations, recent reviews [79, 68] show that no single tool consistently outperform others in most experiments. For a specific patient, detection of fusions and aberrations based on a single tool might have lots of false positives, but taking consensus from multiple tools can potentially lose some of the events. To better capture gene fusions and understand their impacts in cancer patients, a possible strategy for validating these aberrations is to check their translation results in proteome level when mass spectrometry data for the same patient is available. While lots of efforts have been made in the development of proteogenomics [100, 119], most of the work focuses on studying correlations between transcripts in gene annotations and their corresponding proteins. Computational methods for validating non-canonical splicing events such fusions are not the main theme in the field of proteogenomics.

1.1 Contribution

In this thesis, we focus on computational methods detecting splicing events based on high-throughput omics datasets, including RNA-Seq and mass spectrometry results for specific patients. The main goal of the study is to design algorithms to improve identification and quantification of alternative splicing isoforms from RNA-Seq datasets, and intergratively detect non-canonical splicing events when RNA-Seq and matching mass spectra datasets are available. More specifically, we present the following contributions:

- We introduce *CLIIQ* [78], a computational tool to simultaneously identify expressed isoforms and estimate their expressions from single or multiple RNA-Seq samples. Given gene annotations and RNA-Seq mapping results, CLIIQ simultaneously identify and quantify transcripts by solving an integer linear program based on the principle of maximum parsimony. CLIIQ can operate in both quantification mode, which are fully based on transcripts in gene annotations, and reconstruction mode, which also supports novel junctions. We design simulation datasets and show that CLIIQ provides better results than both Cufflinks and IsoLasso.
- We introduce *ORMAN* [23], a computational method to determine the single best location for multi-mapping RNA-Seq reads. While in designing CLIIQ, we found that quantification accuracy for isoforms is usually lower than identification accuracy. One of the main reason is that some RNA-Seq reads can be mapped to multiple genomic location, and we might over-count these reads. To resolve this issue, we design ORMAN, which assigns single mapping location for all multi-mapping reads by minimizing local

coverage variations of the relevant regions. ORMAN is done by solving an ILP to minimize local coverage variations in the most likely mapped regions selected by the principle of maximum parsimony. Since ORMAN converts a mapping result containing multi-mapping records into a new mapping file with only a single mapping, it can also be used as a preprocessing tool for counting-based quantification pipelines.

- After designing CLIIQ and ORMAN for studying alternative splicing events, we will introduce *ProTIE* which focuses on detecting non-canonical splicing events using proteogenomics approach. Detection of aberrant transcripts, especially gene fusions, is a significant problem in cancer research, but most pipelines based on sequence mapping results might produce lots of false positives which would take too much time and effort to validate. When RNA-Seq and mass spectrometry datasets for the same patient are both available, we design ProTIE to discover proteome-level signatures which support transcript segment crossing breakpoints for aberrations. Fusions and other aberrations with translation evidence have a higher chance to affect cancer phenotypes of a patient, and should have higher priority in downstream analysis and validation. We apply ProTIE to TCGA/CPTAC breast cancer patients and detect some private events which are not reported in the literature.

In addition to our primary contributions to detection of splicing events using RNA-Seq and proteomics datasets mentioned above, our other contributions to the field of computational biology can be found in [59].

1.2 Organization of The Thesis

The rest of the thesis is organized as follows. We start with an overview of RNA-Seq technology and major computational challenges in studying a specific patient using RNA-Seq datasets in Chapter 2. In Chapter 3, we introduce CLIIQ, our computational algorithm for simultaneous identification and quantification of expressed isoforms based on RNA-Seq mapping results allowing novel junctions. In Chapter 4, we describe ORMAN, a combinatorial optimization formulation which assigns single genomic location for all RNA-Seq multi-mapping reads. The topic of Chapter 5 is about ProTIE, our proteogenomic integration engine which takes RNA-Seq and proteomics datasets for the same patient to detect translated aberrant transcripts. Finally, in Chapter 6 we conclude by a summary of our contribution to splicing detection problems, and potential directions for future work.

Chapter 2

Background and Related Work

This chapter provides preliminaries for transcriptome analysis using RNA-Seq. It starts with an overview of RNA and RNA-Seq. We then describe fundamental computational problems in analyzing RNA-Seq data. Advantages and limitations of different strategies for solving these problems are provided. We conclude by the current status of proteogenomics studies to understand usage of RNA-Seq analysis in cancer research.

2.1 Overview of RNA and RNA-Seq

RNA. *Ribonucleic acid* (RNA) is a biological macromolecule which is essential in all known forms of life. In cells, genetic information transfers as follows: *deoxyribonucleic acid* (DNA) will be transcribed into *messenger RNA* (mRNA), which will be later translated into proteins. The actual process in eukaryotes (e.g., human) is a bit complex than prokaryote (e.g., bacteria) since only *expressed regions* (**exons**) of genes will remain in final mRNAs. The transcription process (i) synthesizes *precursor mRNA* (pre-mRNA) from DNA; (ii) removes regions between exons (**introns**) from the pre-mRNA. The process of forming mRNA by only concatenating exons and removing introns from pre-mRNA is called **splicing**.

Protein coding genes can have multiple splicing forms as long as the participant exons concatenate in a way consistent to their relative order in genome. This **alternative splicing** mechanism allows a single gene to have multiple *RNA isoforms*, the set of expressed transcripts from the same gene, and potentially code different proteins, as shown in figure 2.1.

RNA-Seq. RNA-Seq is the standard approach for studying *transcriptome*, the complete set of expressed transcripts in a tissue, in a high-throughput fashion. For a specific tissue, RNA-Seq applies DNA sequencer to its *complementary DNA* (cDNA)[1] library, and extract

[1]RNA can be viewed as strings over $\{A, C, G, U\}$ while DNA are double-strand strings over $\{A, C, G, U\}$. The complementary DNA of RNA is to duplicate a RNA template by replacing all its U by T, and add the corresponding strand.

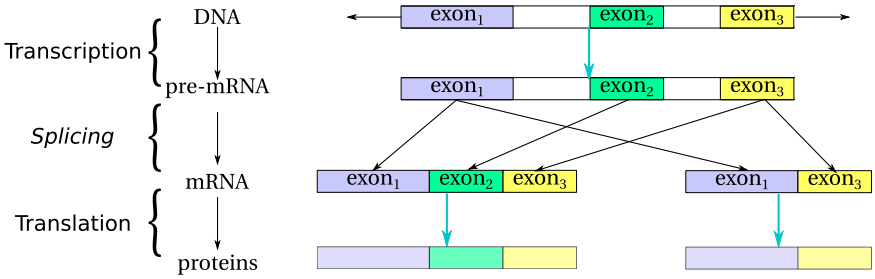


Figure 2.1: Alternative splicing events in a gene containing 3 exons and 2 RNA isoforms. Two introns of the gene will be eventually removed in mRNA during the splicing process.

short reads as the RNA-Seq data for this tissue. The protocol (see Figure 2.2) starts by first converting RNA molecules in a tissue to a library of cDNA fragments with adaptors attached to one or both ends. These cDNA fragments will be sequenced to get short *reads* either from one end (single-end sequencing) or both ends (pair-end sequencing) by DNA sequencers. The fragment length may range between 30âĂŞ500 base pair(bp), depending on the DNA-sequencing technology used.

Given a RNA-Seq dataset, the primary aims in analyzing transcriptome are to identify sequences of expressed transcripts (identification problem) and quantify their expression levels (quantification problem). Aberrant transcript detection is also an important task, especially in cancer research. Note that in practice expression level stands for relative abundance. For example, t_1, t_2, t_3 in Figure 2.2 have 1, 1, and 2 full-length copies separately. After amplifications it is not possible to obtain the original numbers. Therefore, the quantification results are considered accurate as long as the relative abundance of estimated expressions is 1:1:2.

RNA-Seq has the following advantages compared to previous technologies(e.g., microarray methods) [155]. First, we can use RNA-Seq to detect transcripts from species without existing genomic sequence or gene annotations since the data comes from RNA. For example, 454-based RNA-Seq has been used to sequence the transcriptome of the Glanville fritillary butterfly27 [150]. Second, it allows us to detect various events in base-pair resolution. It is especially helpful for species with complex transcriptional structures. Researchers can use RNA-Seq data to augment existing gene annotations [94], discover new genes [15], and even detect variations in the transcribed regions [86]. Third, RNA-Seq approach has a large dynamic range of expression levels (greater than 10000-fold) over which transcripts can be detected [98, 96]. Compared to dynamic range for microarrays which is around few-hundredfold, RNA-Seq improves our ability to capture expressed transcripts.

In contrast, RNA-Seq has the following limitations which lead to computationally challenging problems. First, many of the generated sequences are noncontiguous in genome due to splicing. Determining splicing sites within a read and their locations in genome is a complex issue in mapping. Second, the same read can be shared by multiple isoforms

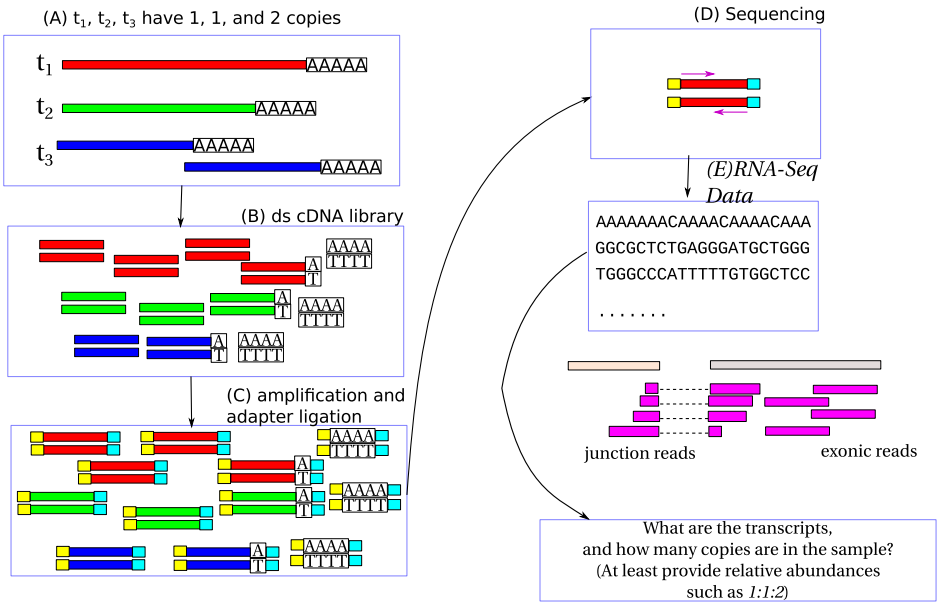


Figure 2.2: High level overview of RNA-Seq protocol. (A) The sample contains 3 different transcripts with 1, 1, and 2 full length copies. (B) These RNA will be converted into double stranded cDNA libraries. (C) The cDNA library will be amplified and have adapter attached. (D) DNA sequencers extract fragments, and generate short reads as the RNA-Seq data as shown in (E).

from the same gene. Assignment of each read to the correct expressed transcript is the key issue for both identification and quantification problem. Third, RNA-Seq usually has uneven coverage within the same isoform due to different biases, while the large dynamic range of RNA-Seq leads to different expression levels across the sample. These are difficult for *de novo* assemblers to distinguish between sequencing errors and reads from low abundance transcript sequences. There are other minor issues which can lead to chimeric sequences, such as read-through, or poly-A tail. We will study how these challenges are being addressed.

2.2 *De novo* Assembly and Mappings in RNA-Seq

For an RNA-Seq dataset, transcriptome analysis usually starts by either (i) *de novo* assembly of full-length transcripts, or (ii) mapping reads back onto reference genome. Most *de novo* assemblers are based on de Bruijn graph [25] and they differ in the strategy dealing with low abundant regions. RNA-Seq mappers differ in their approaches in determining splicing sites for a read, especially those with short anchors at the end.

2.2.1 *De novo* Transcriptome Assembly

De novo assembly: Given a set of RNA-Seq reads R, find a set of contiguous sequence contigs as potential full-length or partial transcripts.

As shown in table 2.1, most popular *de novo* transcriptome assemblers are based on de Bruijn graph, the foundation of multiple popular whole-genome assemblers. Transcriptome assemblers start by collecting all substrings of length k (k-mer) from reads. Different k-mers are merged into a longer contiguous sequence (**contig**) as long as they share substrings of length k-1. The assembly process keeps extending a contig by attaching k-mers overlapping with it of k-1 bp until a contig can no longer be extended. Such a contig will be considered as a transcript, and the assembler iterates to collect all such contigs as the set of potential transcripts.

De novo assemblers do not depend on reference sequences or gene annotations. For species without these information, *de novo* assembly is the only way to start transcriptome analysis. In addition, assembly is also used when transcripts might be highly different from known sequences, such as aberrant transcripts in cancer cells.

The main challenges of *de novo* transcriptome assembly come from (i) uneven coverage of reads in a transcript; (ii) different expression levels across the sample; and (iii) reads shared by multiple RNA isoforms. Transcriptome assemblers can be classified into the following categories based on strategies they use to address these issues: (i) Single k-mer size assembler: the assembler uses k-mers of fixed size to iteratively find all contigs as potential transcripts. Trinity [45] and SOAPdenovo-Trans [163] belong to this category. The drawback of these methods is that they usually miss low-abundance transcripts; (ii) Multiple k-mer size assembler: the assembler first determines a fixed set of k-mer sizes, and for each value of k it builds the corresponding set of contigs. All contigs assembled from different k-mer sizes will be combined as the final set of transcripts based on some consensus methods. Oases-M [125], Trans-ABySS [114], and IDBA-Tran [108] use such a strategy. These methods may not perform well for transcripts of uneven coverage. In addition, combining contigs obtained from multiple k-mer sizes tends to report different regions for a transcript as separate contigs and generate redundant results; and (iii) Dynamic k-mer sizes assembler: the assembler starts by partitioning reads into disjoint clusters. For each cluster, the assembler builds multiple de Bruijn graphs for different k-mer sizes simultaneously and generates final contigs. Bermuda [140] is the only tool using such a strategy.

Trinity

We use Trinity [45], the most popular *de novo* transcriptome assembler, to explain general workflow of the assembly process. Unlike other alternatives built as extensions of underlying whole-genome assemblers, Trinity is designed specifically for transcriptome assembly. As mentioned earlier, Trinity uses single k-mer size in assembly and removes potentially

Tools	de Bruijn assembler	K-mer size for assembly
Trinity [45]	`standalone`	Single k-mer size
Oases-M [125]	*Velvet* [169]	Multiple k-mer sizes
Trans-ABySS [114]	*ABySS* [132]	Multiple k-mer sizes
SOAPdenovo-Trans [163]	*SOAPdenovo2* [80]	Single k-mer size
IDBA-Tran [108]	*IDBA*	Multiple k-mer sizes
Bermuda [140]	`standalone`	Dynamic k-mer sizes

Table 2.1: Summary of *de novo* transcriptome assemblers based on de Bruijn graph

erroneous k-mers of low frequencies. Such design strategy makes Trinity to perform well in highly-expressed transcripts but may fail to capture low-abundance sequences.

Trinity consists of three major steps: *inchworm* for assembling contigs, *chrysalis* for grouping contigs, and *butterfly* for determining transcript sequences. First the *Jellyfish* module constructs the dictionary of all k-mer (k=25) and their frequencies from the dataset. ***Inchworm*** then assembles the initial set of contigs as (sub)sequences of candidate transcripts. Inchworm extends a contig in a greedy fashion by attaching the k-mer with highest frequency among all candidates overlapping the contig with $k-1$ nucleotides. The results usually contain the dominant isoforms in full length and unique regions in alternative isoforms [54].

Chrysalis clusters contigs together if they overlap with $k-1$ nucleotides. Contigs in a cluster suppose to represent transcripts from the same gene or related gene paralogs. Chrysalis then constructs de Bruijn graphs for each cluster, and partitions reads into among these disjoint clusters accordingly.

Butterfly reports final transcript sequences by traversing paths on each de Bruijn graph separately. Butterfly first removes paths with low number of supporting reads on each graph since these paths might be due to sequencing errors or low-abundance variants. Second, Butterfly converts each de Bruijn graph into a compact form by merging k-mers in a non-branching path as a single node. Butterfly eventually determines better paths on each compact graph as candidate transcripts according to their weights obtained from the number of supporting reads and available paired-end information.

2.2.2 Spliced Mappings for RNA-Seq Reads

Spliced Mappings: Given a set of RNA-Seq reads R, a reference genome G, an error threshold e, a scoring function δ, a maximum intron length L, and an optional set of gene annotations M. For each read r in R, find its mapping locations in G such that (i) a mapping can contain multiple junctions of length $\leq L$; (ii) r has the distance $\leq e$ under δ if we consider the cost of junctions as 0; and (iii) when M is provided, those mapping locations compatible with M are considered better than others.

Gene models contain hypothetical transcriptional structures for genes. For a specific gene, its gene models provide exon boundaries and splicing site locations for validated or

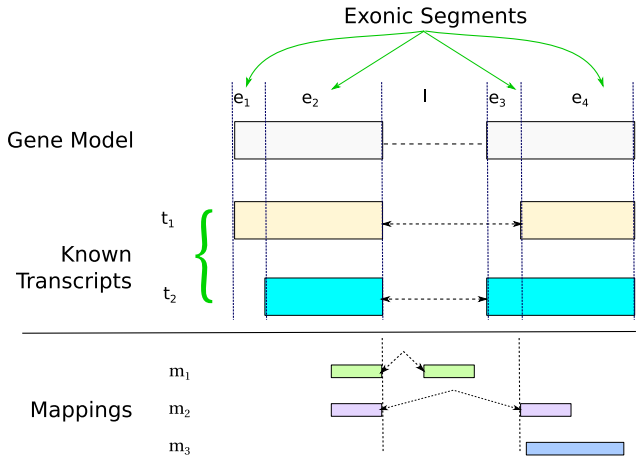


Figure 2.3: Read compatibility based on a gene model for a gene with 2 isoforms and 4 exonic regions. Transcript t_1 contains exonic regions e_1, e_2, and e_4; t_2 is formed by e_2, e_3, and e_4. There are two valid junctions according to this model: $e_2 \rightarrow e_3$ and $e_2 \rightarrow e_4$. The spliced mapping m_1 is not compatible with this gene model since its junction ends in the middle of an intronic region. m_2 is compatible with t_1 but not with t_2. m_3 is compatible with both transcripts.

potential transcripts of this gene. A mapping of r is **compatible** with a gene model M if (i) r can be sequenced from one of the transcript t_i in the model; or (ii) every junction in this mapping starts at some intron starting location in M and ends at some intron ending location in M. For the first case we also call the mapping or the read r is **compatible** with t_i.

Spliced mappers start by mapping reads to the genome with distance $\leq e$ under δ. For an unmapped read, mappers usually anchor the possible mapping locations by either finding longest common substrings between the read and genome, or detecting possible locations for k-mer of the read. They extend from these anchors to complete the spliced mappings based on information of canonical junction site [2], the constraint of intron length L, and splicing sites in a gene model M.

In transcriptome analysis using RNA-Seq data, mapping-based approaches are preferred to assembly-based approaches when reference genome and gene annotations are available [24, 151]. The main reasons include (i) mappings are not affected by coverage issues; (ii) mappings provide higher sensitivity (more analyzed reads) than *de novo* assembly; and (iii) mappings provide direct connections between reads and genes that are better for downstream analysis. More specifically, spliced mappings are more popular than transcriptome mappings (mapping short reads to reference transcriptome built from a gene model) since the latter is not a practical solution in many applications.

[2] An intron usually starts with GU and ends with AG.

The most challenging part for spliced mappings is that many RNA-Seq reads are formed of multiple exonic regions noncontiguous on genome. To report a spliced mapping for a read, mappers need to (i) detect boundaries of these exonic regions in the read; (ii) locate each exonic region on genome; and (iii) select a combination of mapping loci for all these regions when some can be mapped to multiple locations. Spliced mappers can be classified into two types according to their strategies for deciding junction sites for a read: (i) Spliced mapping based on candidate junction-sites: the mapper first detects possible junction sites based on mapped reads. Remaining unmapped reads will be compared to the collected junction site database in order to determine their mapping locations. TopHat [146] uses this approach. (ii) Direct spliced mapping: the mapper maps a read r by iteratively finding the longest common prefix (or suffix) between r and reference genome G. The example includes STAR [26].

Tools	mapping	speed	memory	support
STAR [26]	`standalone`	1	4	developers
TopHat [146]	*Bowtie* [66]	3	1	community
HISAT [60]	*Bowtie* codebase	2	2	community
GSNAP [159]	`standalone`	4	3	developer

Table 2.2: Summary of 4 popular RNA-Seq spliced mappers. The rank 1 specifies either the fastest tool or the tool requiring least amount of memory among these four.

TopHat

TopHat provides spliced mappings by first collecting candidate junction sites from mapped reads. It consists of following steps: (i) Detecting exonic regions: TopHat first detects all possbile exon boundaries. This step is originally done by mapping reads to reference genome using Bowtie, and extracting regions containing mapped reads. For read length $\geq$ 75bp, TopHat maps 25-mer from reads to reference genome for more accurate detections. (ii) Constructing junction database: TopHat then constructs a junction-site database based on all known juctions from the gene model M, and exon boundaries obtained in previous step. (iii) Spliced mapping: Finally TopHat compares unmapped reads in the first step (IUM, *Initially UnMapped reads*) to junction sites in the database, and determines spliced mappings for these reads.

STAR

Spliced **T**ranscripts **A**lignment to a **R**eference (STAR) is a mapper specifically designed for mapping noncontiguous reads back to the reference genome. STAR maps a read by locating the longest common substring between this read and genome sequence. STAR does not depend on any pre-built junction site database to detect spliced mappings, and therefore

provides more accurate results than TopHat in most benchmark studies including the recent one by **R**NA-seq **G**enome **A**nnotation A**S**sessment **P**roject (RGASP) [30].

The mapping of STAR consists of two steps. In the first step, STAR tries to locate seeds for a read by sequentially finding its longest prefix with exact match [3] on genome. STAR store the genomic location for such a prefix as a "seed", and continues the process for the remaining of the read. This step is done by suffix array, and allows STAR to find all possible locations for the seeds. In the second step, the genomic mapping location of a read will be obtained by concatenating the mapping locations of these seeds. Locations of two seeds are combined as a (partial) mapping record if (i) the distance between them is less than the maximum intron length L (default: 10^6 nucleotides); and (ii) seeds are concatenated in a way consistent to their original order in the read without introducing rearrangements. Once the concatenation of seeds does not cover the read completely due to indels or mismatches, a dynamic programming formulation will be used to determine the final mapping result.

2.3 Fundamental Problems in Transcriptome Profiling

RNA-Seq data is widely used to detect alternative splicing events, differentially expressed genes, and complex genomic aberration. These significant applications are based on accurate *transcriptome profiling* that consists of three tasks: (i) transcriptome identification: identifing sequences for expressed transcripts in a tissue; (ii) transcriptome quantification: quantifing expression levels (the number of full length copies) for all identified transcripts; and (iii) aberration detection: detecting aberrant transcripts. These problems can be solved by either assembly-based approach or mapping-based approach. In this section we will focus on mapping-based approach of transcriptome profiling, the standard approach in analyzing human RNA-Seq data.

2.3.1 Transcriptome Identification

Transcriptome Identification: Given a set of RNA-Seq reads R along with its genomic mapping result $\mathcal{M}$, determine the exact set I of expressed transcripts in the tissue.

For each gene or isolated region (no mappings cross the boundaries of this region), an identification tool starts by building a directed acyclic graph $\mathcal{G}$ which represents its mapping results, and any path in $\mathcal{G}$ corresponds to a (partial) transcript.

Splicing graph and *overlapping graph* are two widely-used models, and we will study their properties in the following section. The identification tools determine expressed transcripts for this gene or region by finding a set of paths which cover all nodes and edges in $\mathcal{G}$.

The key issue in transcriptome identification problem is to devise an efficient strategy to select the set of transcripts covering all observed exons and junctions. Consider a gene

[3] It is defined as the Maximal Mappable Prefix (MMP) for a read in original paper.

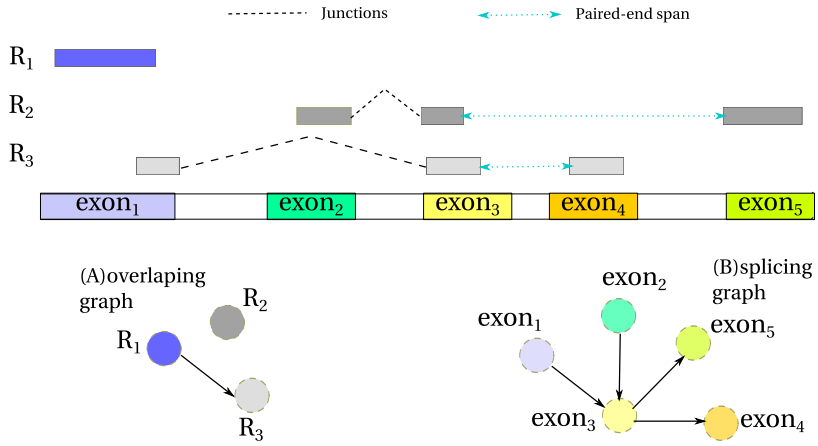


Figure 2.4: The splicing graph and overlapping graph for a gene with 3 mapped paired-end reads. (A) overlapping graph: R_2 and R_3 have to be in different paths. (B) splicing graph: among 4 possible paths, $e_1 \to e_3 \to e_4$ and $e_2 \to e_3 \to e_5$ are correct transcript candidates. $e_1 \to e_3 \to e_5$ and $e_2 \to e_3 \to e_4$ are infeasible.

with k exonic regions: there are $\mathcal{O}(2^k)$ possible transcripts due to alternative splicing, and therefore $\mathcal{O}(2^{2^k})$ different solutions for this problem. It is impractical to explicitly enumerate each possible candidate and check the validity to pick the final solution.

Splicing Graph. A splicing graph $\mathcal{G}$ for a gene g is a directed acyclic graph such that (i) there exists a vertex v_i for every exonic region e_i; and (ii) there exists an edge (v_i, v_j) if there exists a junction between corresponding exonic regions e_i and e_j.

The main issue of splicing graph is that it only provides connectivity information between 2 exons, and a path in the splicing graph does not always correspond to a real transcript in Figure 2.4. As a result, identification tools based on splicing graph can generate false positive candidates. For example, in Figure 2.4, $e_1 \to e_3 \to e_5$ is a path on the splicing graph, but it is not a real transcript since (e_1, e_3) and (e_3, e_5) come from different reads and there is no evidence supporting concatenation of these 3 exons. The same is valid for path $e_2 \to e_3 \to e_4$.

Scripture [46] is the first identification tool based on splicing graph, and it focuses on reporting as many isoforms expressed in a single sample as possible. While Scripture might help to capture low abundance transcripts, it usually reports many incorrect isoforms which seriously affect quantification accuracies.

Overlapping Graph. An overlapping graph $\mathcal{G}$ for an isolated region g is a directed acyclic graph such that (i) there exists a vertex v_i for every mapped read r_i in g; and (ii) there exists an edge (v_i, v_j) if the corresponding reads r_i and r_j can come from the same transcript.

The key issue of overlapping graph is that it requires a proper strategy to detect reads which cannot exist in the same transcript even though they can be partially mapped to the

same exon. Identification tools based on overlapping graph can have false negatives if these reads are removed accidentally. In Figure 2.4, R_2 and R_3 cannot exist in the same transcript due to different splicing structures. An identification tool based on overlapping graph needs to resolve the conflict by either (i) confirming these two reads are in different paths; or (ii) removing at least one of the reads as the error-correction step. Once an identification tool removes R_2 or R_3, it loses essential junction information to identify the corresponding transcript which lead to a false negative.

Cufflinks [147] is one of transcriptome identification tool based on overlapping graph. The goal of Cufflinks is to find the minimal set of isoforms that cover mapping result. Furthermore, Cufflinks sometimes resolves the conflict by removing low abundance reads, and therefore loses the corresponding transcripts in the report.

2.3.2 Transcriptome Quantification

Transcriptome Quantification: Given a set of RNA-Seq reads R along with a fixed set I of transcript sequences. Quantify expression levels (the number of full length copies) for all transcripts in I.

Quantification tools are based on the following observation: for a given transcript t_i, the number of reads (n_i) from t_i is proportional to transcript length (l_i) and the expression value (c_i) of t_i. We can denote the relation by:

$$n_i \propto c_i \times l_i$$

Quantification tools estimate n_i by calculating the number of reads mapped to t_i. Note that a read r mapped to a transcript t_i if r is compatible with t_i as defined in section 2.2.2.

The main challenge of transcriptome quantification comes from reads which can be mapped to multiple locations (*multi-mapping reads*). Quantification tools can be classified into 2 categories based on strategies they used to handle multi-mapping reads: (i) read counting: such strategy simply counts the number of reads mapped to a transcript or a gene to approximate n_i. These tools are used in comparing expression levels of the same gene/transcript between multiple samples. (ii) Expectation-Maximization (EM) algorithm: this quantification strategy estimates expression level for transcripts by distributing multi-mapping reads fractionally into mapped genes or transcripts based on their relative abundances. These tools are used for comparing different gene/transcript within a single sample.

Read Counting: HTSeq-Count and featureCount. Quantification by counting the number of mapped reads is the most popular strategy in cancer research. HTSeq-Count [5] and featureCounts [77] are the two tools based on this approach for counting reads in a SAM/BAM mapping result. This approach cannot deal with multi-mapping reads: HTSeq-

Count consider all mapping records, and featureCounts discard multi-mapping reads. They cannot be used to compare different genes or transcripts in the same tissue since the value is not normalized by transcript length. On the other hand, results of this approach are widely used for comparing expression level of the same gene/transcript across multiple samples. Most differential expression detection tools such as DESeq [3, 4], EdgeR [115], and baySeq [49] normalize expression values before further comparisons. This step is highly affected by normalized values. Therefore these downstream analysis tools perform better on the raw data of read counts.

ExpectationâĂŞMaximization(EM) Algorithm. IsoEM [101] and RSEM [71, 70] are two quantification tools based on EM algorithm. Instead of directly resolving mapping ambiguity, this approach focuses on finding a way of assigning multi-mapping reads fractionally such that extra expression assigned to each gene or transcript is proportional to its original relative abundance. The EM algorithm can be summarized as follows:

- E-step: the algorithm calculates the expected number of n_i (reads from t_i) by assigning multi-reads fractionally to all relevant genes or transcripts based on their relative abundances. Note that initial expression values are estimated based on uniquely mapped reads.
- M-step: the algorithm updates expression levels for t_i based on n_i obtained in previous E-step. The algorithm proceeds to the next iteration if expression levels do not converge compared to previous estimations.

Transcript	Transcript Length	# uniquely-mapped reads	# multi-mapped reads
t_1	200	90	*30*
t_2	100	40	0
t_3	60	40	*30*

Table 2.3: Information for a dataset with 200 reads and 3 expressed transcripts. 30 reads are shared by t_1 and t_3.

Consider the example in Table 2.3 with 200 reads and 3 expressed transcripts. Note that 30 reads are shared by t_1 and t_3.

0. Initialization: the algorithm starts by estimating initial relative abundances f_i^0 for t_i based on uniquely mapped reads: $f_1^0 = \frac{\frac{90}{200}}{\frac{90}{200}+\frac{40}{100}+\frac{40}{60}} = 0.297$, $f_2^0 = \frac{\frac{40}{100}}{\frac{90}{200}+\frac{40}{100}+\frac{40}{60}} = 0.264$, $f_3^0 = \frac{\frac{40}{60}}{\frac{90}{200}+\frac{40}{100}+\frac{40}{60}} = 0.439$.

1. E-step: multi-mapping reads are assigned proportionally to 3 transcripts in order to obtain expected number of reads:

$$\begin{array}{rclcrcl} n_1 & = & 90 + 30 \times \frac{0.297}{0.297+0.439} & = & 90 + 12.11 & = & 102.11 \\ n_2 & = & 40 + 0 & = & 40 & & \\ n_3 & = & 40 + 30 \times \frac{0.439}{0.297+0.439} & = & 40 + 17.89 & = & 57.89 \end{array}$$

2. M-step: the algorithm updates relative abundances for each transcript based on newly obtained n_i:

$$f_1^1 = \frac{\frac{102.11}{200}}{\frac{102.11}{200} + \frac{40}{100} + \frac{57.89}{60}} = 0.272$$

$$f_2^1 = \frac{\frac{40}{100}}{\frac{102.11}{200} + \frac{40}{100} + \frac{57.89}{60}} = 0.213$$

$$f_3^1 = \frac{\frac{57.89}{60}}{\frac{102.11}{200} + \frac{40}{100} + \frac{57.89}{60}} = 0.515$$

   The procedure will continue until some pre-defined convergence threshold is achieved. In fact, **Rescue** [96], one of the earliest methods for transcriptome quantification, can be considered as running single iteration EM which stops after getting f_1^1, f_2^1, f_3^1.

RSEM and IsoEM use different functions to estimate expected number of reads n_i in their E-steps. RSEM uses a comprehensive model which considers read lengths, read orientation, paired-end information, quality score, and the starting position of a read in transcripts. IsoEM only checks if a read is compatible with a transcript (as defined in section 2.2.2). Surprisingly, IsoEM provides similar quantification results to RSEM in much higher speed [56].

Normalized Values: RPKM/FPKM and TPM. RSEM and IsoEM both report normalized expression values. Although these tools provide better solutions for multi-mapping reads than read-counting ones, the normalized values highly affect performance of downstream analysis. Below we will define the normalized value terms that are commonly used. Also, we show how the same transcript can get different normalized values which reduce accuracy of downstream analysis tools. Note that $N = \sum_j n_j$ is the total number of reads generated from the dataset.

- **R**eads/**F**ragments **P**er **K**ilobase per **M**illion mapped reads (RPKM/FPKM) [96]: RPKM (for single-end reads) and FPKM (for paired-end reads) for a transcript t_i are defined as follows:

$$RPKM_i = \frac{n_i}{\frac{l_i}{10^3} \times \frac{N}{10^6}} = \frac{n_i}{l_i N} \times 10^9$$

Dataset	Transcript	Length	Mapped Reads	RPKM
1	t_1	200	100	$0.5 \times \frac{10^9}{160}$
1	t_2	60	60	$\frac{10^9}{160}$
2	t_3	400	200	$0.5 \times \frac{10^9}{260}$
2	t_2	60	60	$\frac{10^9}{260}$

Table 2.4: Comparison of RPKMs of transcripts in datasets with different length distributions. Note t_2 has different RPKMs in two datasets although it has identical number of mapped reads.

Transcript	Transcript length	# uniquely-mapped reads	# multi-mapped reads	# assigned reads	TPM
t_1	200	90	30	**30**	0.36×10^6
t_2	**100**	**40**	0	0	0.24×10^6
t_3	60	40	30	0	0.4×10^6
t_1	200	90	30	0	0.22×10^6
t_2	**100**	**40**	0	0	0.2×10^6
t_3	60	40	30	**30**	0.58×10^6

Table 2.5: Comparison of TPMs for transcripts in the same dataset under different estimation results. Note that t_2 has different TPMs in two quantification solutions although it only contains uniquely mapped reads.

Intuitively, if we scale the sample size to 1 million reads, there are $RPKM_i$ reads on a 1000-bp unit of t_i assuming uniform coverage. Since RPKM is normalized by the sample size N, it is affected by length distribution of expressed transcripts in a tissue. In Table 2.4, t_2 has identical relative abundance in two different datasets but different RPKM values simply due to the length of other expressed transcript. This issue motivates the definition of TPM.

- **T**ranscripts **P**er **M**illion (TPM) [152]: TPM for t_i is defined as follows:

$$TPM_i = \frac{\frac{n_i}{l_i}}{\sum_j \frac{n_j}{l_j}} \times 10^6 = \frac{RPKM_i}{\sum_j RPKM_j} \times 10^6$$

If we scale the RNA-Seq dataset to have 1 million copies of full-length transcripts, TPM_i copies among them are from t_i. TPM is basically a scaled relative abundance, and is therefore affected by estimations of all other transcripts in the tissue. In Table 2.5, t_2 does not contain multi-mapping reads but still gets different TPM values under different strategies of assigning shared reads between t_1 and t_3.

2.3.3 Transcriptome Reconstruction

Transcriptome Reconstruction: Given a set of RNA-Seq reads R and its genomic mapping result $\mathcal{A}$, identify the exact set I of expressed transcripts and quantify their expression levels.

For each gene or isolated region g, a reconstruction tool starts by building its splicing graph $\mathcal{G}$, and adding weight to each node and edge based on the mapping coverage. The goal of reconstruction tools is to find a set of paths with alternative weights based on estimated expressions which cover all nodes and edges in $\mathcal{G}$, and for each node and edge the accumulated weight from all paths is close to its original weight.

The idea of transcriptome reconstruction is to solve identification and quantification problems simultaneously. Transcriptome identification and quantification were initially considered as independent problems [42]. Although Cufflinks [147] and Scripture [46] report expression levels, their quantification steps are based on the fixed set of previously identified transcripts. Results of identification and quantification affect each other, and solving them simultaneously should provide better results than separate solutions [137].

The approach of transcriptome reconstruction can be summarized as follows. For each gene they build a splicing graph $\mathcal{G}$. Note that each node corresponds to an exonic region and each edge represents a junction between two exons. Any path $p \in \mathcal{P}$ of $\mathcal{G}$ is a (partial) transcript. Let $\text{Obs}(v_i)$ denote expressions of exon i and $\text{Obs}(e_{ij})$ be expression of the junction between exon i and exon j based on mapping. A reconstruction solution $\mathcal{S}$ consists of the set $\mathcal{P}' \subset \mathcal{P}$, and estimated expression level $\text{Exp}(p) \geq 0 \ \forall p \in \mathcal{P}'$. Reconstruction tools focus on minimizing summations of their error estimation functions: (i) estimation error for all exons: for each exon v in $\mathcal{G}$, the error is a function f_{exon} of the difference between $\text{Obs}(v)$ and the estimated expression obtained from all transcripts containing v in $\mathcal{P}'$; and (ii) estimation error for all junctions: similar function f_{junc} for all junctions. The objective function can be denoted as follows:

$$\text{minimize}\left(\sum_{v \in V(\mathcal{G})} \overbrace{\left| \text{Obs}(v) - \sum_{p \in \mathcal{P}'; v \in p} \text{Exp}(p) \right|}^{f_{exon}(v)} + \sum_{j \in E(\mathcal{G})} \overbrace{\left| \text{Obs}(j) - \sum_{p \in \mathcal{P}'; j \in p} \text{Exp}(p) \right|}^{f_{junc}(j)} \right)$$

The key issue in transcriptome reconstruction is to find a strategy which maintains balance between accuracies of identifications and quantifications. Many genes have only several expressed isoforms in real datasets, and a *sparse* solution (most RNA isoforms do not express) is usually preferred. It is crucial for a reconstruction tool to design a strategy which prevents overfitting due to the error estimation function. For example, a solution with multiple low abundance isoforms of single exon is believed to be incorrect in practice despite this solution leads to very small estimation error.

Tools	Solutions to multi-reads	Design Principles	Model
Scripture	-	sensitivity	splicing graph
Cufflinks	rescue(-like)	precision	*overlapping graph*
IsoLasso [75]/**CEM** [76]	-	quantification	splicing graph
SLIDE [73]	-	quantification	splicing graph
iReckon [92]	-	quantification	splicing graph
StringTie [109]	-	quantification	splicing graph
Traph [143, 144]	-	quantification	splicing graph

Table 2.6: Summary of RNA-Seq reconstruction tools

IsoLasso

IsoLasso [75] is one of the earliest transcriptome reconstruction tool. It applies an isoform enumeration strategy [4] based on IsoInfer [35], and uses Least Absolute Shrinkage and Selection Operator (LASSO) algorithm [141] to control solution sparsity in its quadratic program formulation.

IsoLasso is based on splicing graph. It first collects all maximal paths from the graph, and applies isoform enumeration strategy from IsoInfer to efficiently filter out incorrect candidate transcripts. In short, for a candidate transcript $\hat{t}$ such as $e_1 \rightarrow e_3 \rightarrow e_5$ in Figure 2.4, IsoLasso takes all reads relevant to $\hat{t}$ (i.e., R_2 and R_3) and checks if $\hat{t}$ can be actually assembled from these reads. IsoLasso will remove incorect $\hat{t}$ and reduce the number of false positive candidates.

Once the incorrect transcripts are discarded, IsoLasso starts the reconstruction based on the following formulation:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{v \in V} \left(\text{Obs}(v) - \sum_{p \in \mathcal{P}'; v \in p} \text{Exp}(p) \right)^2 \\
\text{subject to} \quad & \text{Exp}(p) \geq 0 && , \ p \in \mathcal{P} \\
& \sum_{p \in \mathcal{P}'} \text{Exp}(p) \leq \lambda \\
& \sum_{p \in \mathcal{P}'; v \in p} \text{Exp}(p) \geq \delta && , \text{an exon } v \text{ has mapped reads} \\
& \sum_{p \in \mathcal{P}'; j \in p} \text{Exp}(p) \geq \delta && , \text{a junction } j \text{ has mapped reads}
\end{aligned}
$$

It uses quadratic function as the error estimator f_{exon} for exonic regions. In the formulation of LASSO technique, λ is used to control sparsity where smaller λ leads to fewer

[4] Isoform enumeration strategy is used to list all possible transcripts based on the splicing graph.

expressed transcripts. Finally, IsoLasso uses a small positive number δ to ensure that all expressed exons and junctions will be included in some reported transcripts.

2.3.4 Aberrant Detection

Aberrant Detection: Given a set of RNA-Seq reads R, reference genome G, and gene annotations M, detect the set of expressed aberrations in a tissue.

Aberrant detection focuses on transcript sequences which are different from normal splicing results. The splicing mechanism concatenates multiple exonic sequences in a way consistent with their relative order on genome. For aberrations, however, fusion transcripts can break relative order by concatenating exons from multiple genes or multiple chromosomes, and some aberrations (e.g., inversions and duplications) directly change the sequence contents. It is difficult to detect these aberrations using identification or quantification strategies mentioned earlier. In addition, these aberrations are believed to be potential drivers in many specific tumor types or human diseases [145, 36, 83], and accurate detection of these events is significant in studying these diseases.

Aberrant detection can be done in mapping-based approach or assembly-base approach. In mapping-based approach, RNA-Seq reads are first mapped onto genome (with or without gene annotations). Reads which cannot be mapped properly are believed to be signatures of aberrations. Detection tools will use these reads to determine types and locations of the aberrations. Tools based on this approach include most fusion detection tools such as deFuse [84], TopHat-Fusion [61], FusionMap [43], FusionSeq [122], and SOAPfuse [55]. Assembly-based approach will first *de novo* assemble RNA-Seq reads into possible contigs. Aberrant detection tools will predict events by mapping these contigs back to reference genome. This is not fully independent from reference sequence, but this approach limits biases from gene annotations in its contigs, and can potentially provide us more novel events. These techniques include Barnacle [138] and Dissect [165].

2.4 Proteogenomics

Accurate transcriptome profiling results can be a powerful tool in cancer research once it is combined with *Whole Genome Sequencing* (WGS) and proteomics data from the same patient. In cancer research, it is important to understand the types of variations in different genes that can be potential drivers of specific cancers. Given both WGS and RNA-Seq data from the same patient, we can validate aberrations in transcribed regions. Once we have RNA-Seq and proteomics data, we can provide proteome-level evidence for those aberrations and discover sources for novel proteins. Joint analysis of multiple omics data, or *proteogenomics approach*, can therefore link genotypes to phenotypes and help us to better understand cancer cells. Development of proteogenomics approaches in cancer studies can be roughly divided into three stages.

2.4.1 Initial Stage

In the first stage, the goal of proteogenomics studies was to examine feasibility of combining different omics data from the same patient. The study detected expressed known genes and transcripts from RNA-Seq data, and used mass spectrometry data to verify if peptides from these genes or transcripts are translated. Once a significant correlation between peptide abundances and expressed levels for the corresponding transcripts was observed [104], proteogenomics apparoach was considered as a valid strategy to study potential events for specific species.

2.4.2 Species-Specific Database

In the second stage, mass spectrometry data was used to search against species-specific databases for the following reasons.

Validation for Potential Genes or Transcripts. In online genome databases, especially Ensembl (`http://ensembl.org`), many genes and transcripts were included only based on computational predictions. Mass spectormetry data was used to provide proteome level evidence for these potential genes or transcripts in mouse [103], C. elegans [158], zebrafish [15] and even human [94, 128].

Discovery of Novel Proteins. An application in proteogenomics studies was to detect novel proteins due to aberrations [39]. In this study, novel proteins were detected by searching mass spectrometry data against a pre-build database which contain chimeric transcripts from literature, mostly fusions. Evidence of translation for few aberrant transcripts were found. This strategy was used later to construct ChiTaRS (`http://chitars.bioinfo.cnio.es`), a database containing chimeric transcripts from literature for 6 different species [37, 38]. By searching mass spectrometry data against ChiTaRS, more translated aberrations were observed compared to prior study [39]. Without the the corresponding RNA-Seq dataset, however, analysis through ChiTaRS cannot provide confident genomic sources for novel proteins. These work motivated the third stage of proteogenomics studies: proteogenomics analysis based on patient-specific database.

2.4.3 Patient-Specific Database

In the third stage, the focus of proteogenomics studies moved to analyzing RNA-Seq and proteomics datasets from the same patient, or at least patients from the same group (e.g., the same population, or the same tumor types). Clinical Proteomic Tumor Analysis Consortium (CPTAC) plays a central role in this stage. CPTAC provides mass spectrometry proteomics data for some patients previously sequenced by The Cancer Genome Atlas (TCGA), and studies tumor-specific novel proteins and their genomic origins for multiple types of cancers.

For example, a recent publication from CPTAC focused on identifying novel peptides involving Single Amino Acid Variants (SAAVs) based on human colon and rectal cancer [170]. A later study [16] included novel peptides due to novel splice junctions and (a limited set of user defined) Post-Translational Modifications (PTM) for breast cancer patients. Because of the importance of phosphorylation in cellular activity and cancer treatment [112], this strategy was further expanded to identify novel phosphorylation sites [90].

2.5 Conclusion

In this chapter we introduced basics of RNA-Seq, the standard approach for transcriptome analysis. Applications of transcriptome analysis using RNA-Seq rely on accurate transcriptome profiling: transcriptome identification, transcriptome quantification, and aberrant detections. Methods for solving these problems can be classified as either assembly-based approach or mapping-based approach. *De novo* transcriptome assembly does not depend on reference genome or gene annotations, but its performance is affected by issues of uneven coverage and expression levels. RNA-Seq mapping approaches provide more sensitive results than assembly-based ones. However, it comes with a caveat that transcriptome spliced mapping is a challenging and difficult problem.

We also surveyed advantages and limitations of popular tools for transcriptome profiling. There are sill issues which are not properly addressed in these fundamental problems:

- For transcriptome identification and reconstruction, most tools do not perform well in novel isoforms and low abundance transcripts. Thus, algorithms incorporating much information (e.g., multiple samples from the same populations, co-expression profiles) can be used to improve performances regarding these cases.
- For transcriptome quantification, the most popular read-counting strategy cannot deal with multi-mapping issues. A possible solution is to design algorithms which resolve the mapping ambiguity by assigning single origin for these reads directly in SAM/BAM files.
- For aberrant detections and proteogenomics studies, a better strategy to control false positive rates is necessary for analyzing multiple omics data.

Chapter 3

Transcriptome Reconstruction Using Multiple RNA-Seq Datasets

As mentioned in subsection 2.2, current approaches to transcriptome reconstruction [42] starts by either (i) *de novo* assembly of full-length transcripts, or (ii) mapping reads back onto reference genome. *De novo* assembly-based strategy first tries to assemble novel transcripts solely from RNA-Seq reads using *de novo* assemblers such as Trinity [45] and TransABySS [114]. The assembled transcripts are then mapped back to reference genome for discovering novel splice junctions as well as isoforms provided the assembly quality is high. For mapping-based methods, such as Scripture [46], Cufflinks [147], IsoLasso [75], and SLIDE [73], RNA-Seq reads are first mapped onto the genome by splice-aware mappers (STAR [26], HISAT [60], TopHat [146], SpliceMap [7], MapSplice [153], and segemehl [51]). Although the mapping-based method can incorporate information of genome sequences and even gene annotations, their performance are still effected by many factors such as sequencing errors, multiple mapping locations and short exons.

Currently available mapping-based methods for transcriptome reconstruction typically focus on one of the following objectives: (i) sensitivity of identification, (ii) precision of identification, or (iii) quantification accuracy. The objective of Scripture [46], for example, is to optimize the sensitivity of identification. It focuses on reporting as many isoforms expressed in a single sample as possible. Although such a strategy might help to identify transcripts with low expression levels, it may also report many incorrect isoforms, especially when a gene has many number of exons and complicated splicing events. Cufflinks [147] focuses on the precision in identifying isoforms. Its goal is to find the minimum number of isoforms that explain the mapping results. In other words, it aims to make sure that each read is included in at least one reported isoform. Scripture and Cufflinks primarily aim to detect expressed isoforms, but they also provide means to solve the isoform quantification problem in a followup stage, which is based on maximum-likelihood estimation. Unlike these two methods, IsoLasso [75] tries to solve isoform identification and quantification simultaneously by considering the quantification errors in the formulation. More specifically,

it tries to minize the differences between observed and estimated quantification errors using a quadratic program formulation, and use the LASSO technique [141] to prevent overfitting. This strategy helps to compromise between the quantification errors and sparsity.

The main difficulty of transcriptome reconstruction comes from detections of isoforms which are novel (not contained in gene annotations) or low-abundant. Most available methods do not perform well in detecting these events. The similar issue is also of concern in computational genomics, and in particular structural variations [53]. Rather than analyzing each genome independently, however, joint analyzing multiple related samples have been shown to improve the accuracy in detecting structural variants significantly [53]. In fact, some recent studies [117] demonstrate that joint analysis of RNA-Seq data from multiple samples can improve the estimation of expression levels. However, these studies either focus on quantification only [117], or solely provide consensus transcriptome [102] without reporting transcriptomic profiles for each individual. Thus, it is very tempting to design a computational method for detecting novel isoforms by jointly analyzing RNA-Seq data from many related samples (eg. primary tumor v.s. metastasis v.s. normal samples from the same patient, or samples from the same population).

In this chapter we introduce CLIIQ, a novel computational method for identification and quantification of expressed isoforms from *multiple samples* in a *population*. Motivated by ideas from compressed sensing literature, CLIIQ is based on an integer linear programming formulation for identifying and quantifying "the most parsimonious" set of isoforms. We show through simulations that, on a single sample, CLIIQ provides better results in isoform identification and quantification to alternative popular tools. More importantly, CLIIQ has an option to jointly analyze multiple samples, which significantly outperforms other tools in both isoform identification and quantification.

CLIIQ is available at `https://github.com/sfu-compbio/cliiq`.

3.1 Methods

In this section, we introduce integer linear programming formulations for isoform detection and quantification from one or more RNA-Seq data samples. In Section 3.1.1, we show how to determine the number of isoform candidates from a given set of samples. In Section 3.1.2, we introduce mathematical annotations for exon junctions and isoforms. Finally, in Section 3.1.3, we describe ILP formulations to model and solve the problem.

3.1.1 Enumerating Isoforms

One of the challenges for isoform identification problem is the large search space of potential solutions. For a gene containing m exons, there are $O(2^m)$ possible isoforms and thus $O(2^{2^m})$ possible sets of isoforms which can, in theory, form a solution. In theory, the problem of finding an optimal solution in a search space of $O(2^{2^m})$ is NP-hard. In practice, however,

we can reduce the number of possible isoforms (and hence the solution space) by simple heuristics filtration techniques: an isoform will be only considered "expressed" if each of its exons and exon-exon junctions "attracts" more than a user specified number of reads.

3.1.2 Annotations and Their Relations

From now on, we consider the problem of isoform identification and quantification for a particular gene across many samples. We denote the number of samples as s, the number of exons of a specific gene as m and the number of isoforms as t. Let $E = \{E_1, E_2, \cdots, E_m\}$ be the set of the exons (or exon segments) of the gene considered. A junction $J_{i,j}$ $(1 \leq i < j \leq m)$ is defined as the concatenation of two exons (exon segments) E_i and E_j. We denote by $J = \{J_{i,j} | (1 \leq i < j \leq m)\}$ the set of junctions. Let $I = \{I_j | (1 \leq j \leq t)\}$ stand for the set of isoforms and let $Seg(I_j)$ be the set of exons and junctions in the isoform I_j. For a segment S (an exon or a junction), we denote its length as $len(S)$. Finally we denote the read length with L.

We now establish the relation between the expression level of the isoform I_j and the number of reads mapped its exons and junctions. We denote by the variable $E_l(I_j)$ the estimated expression value of the isoform I_j in the l^{th} sample. Intuitively, this value corresponds to the average number of mapped reads per base. Dohm *et al.* [27] show that there are biases in the number mapped reads among the positions of a segment: the starting and ending positions receive much fewer mapped reads than the middle positions. We denote the observed number of reads that passes through the "middle position" of a segment S in the l^{th} sample as $O_l(S)$. Similarly, we denote by $N_l(S)$, the estimated number of reads that "pass through" the middle position of a segment S in l^{th} sample. $O_l(S)$ is estimated from the experimental data while $N_l(S)$ is estimated through the expression of the isoforms that contain S.

Let $f(S, I_j)$ $(S \in Seg(I_j)$ $)$ be the number of starting locations of the mapped reads that cover the middle position of S. In order to define $f(S, I_j)$, let $Pos(S, I_j)$ be the position of the middle point of the segment in the isoform I_j. Then we define $Le(S, I_j) = \max(0, Pos(S, I_j) - L + 1)$ as the leftmost starting position and $Ri(S, I_j) = \min(Pos(S, I_j) + L - 1, len(I_j)) - L + 1$ as the rightmost starting position of a mapped read that cover the middle point of S in the isoform I_j. Now we can define $f(S, I_j)$ as follows:

$$
\begin{aligned}
& f(S, I_j) \\
= & \begin{cases} Ri(S, I_j) - Le(S, I_j) + 1 & \text{if } Le(S, I_j) = 0 \\ & \text{or } len(I_j) \leq Pos(S, I_j) + L - 1 \\ L - 1 & \text{otherwise.} \end{cases}
\end{aligned}
$$

In short, $f(S, I_j)$ is not equal to $L-1$ when the mapping locations of the reads exceed the left or right boundary of an isoform. Now, the estimated number of reads that cover the middle of a segment S ($N_l(S)$) could be written as:

$$N_l(S) = \sum_{\{j|S \in Seg(I_j)\}} f(S, I_j) \times E_l(I_j)$$

3.1.3 An ILP Solution

Since we would like to minimize the number of isoforms while estimating their expression values as close as possible to the observed ones, we will introduce a two-stage formulation. We first determine the minimum number of expressed isoforms such that the estimated expressions of segments are within a user defined fraction ϵ ($0 < \epsilon \leq 1$) of the observed ones. There could be many feasible optimal solutions. Thus, in the second stage, we try to minimize the difference between the observed and estimated expressions of the exons and junctions among all these optimal solutions.

We will describe the constraints on the expressions of segments (exons, junctions) or isoforms in the following subsections. We describe these constraints for each sample l^{th} ($1 \leq j \leq s$). Finally, we describe the objective function of the ILPs.

Constraints on Exon/Junction Expression Values.

We first estimate the total estimated number of reads that are mapped to exons given the expression of the isoforms:

$$\sum_{\{E_i \in E\}} N_l(E_i)$$

Similarly, the total estimated number of reads that are mapped to junctions could be calculated as follows:

$$\sum_{\{J_{ik} \in J\}} N_l(J_{ik})$$

And we have the total observed numbers of reads that are mapped to exons and junctions: $\sum_{\{E_i \in E\}} O_l(E_i)$ and $\sum_{\{J_{ik} \in J\}} O_l(J_{ik})$. Now we enforce the ratio between the estimated number of mapped reads and the observed number of mapped reads of the exons and junctions should be within the interval $[1-\epsilon, 1+\epsilon]$:

$$(1-\epsilon) \sum_{\{E_i \in E\}} O_l(E_i) \leq \sum_{\{E_i \in E\}} N_l(E_i) \leq (1+\epsilon) \sum_{\{E_i \in E\}} O_l(E_i)$$

$$(1-\epsilon) \sum_{\{J_{ik} \in J\}} O_l(J_{ik}) \leq \sum_{\{J_{ik} \in J\}} N_l(J_{ik}) \leq (1+\epsilon) \sum_{\{J_{ik} \in J\}} O_l(J_{ik})$$

Finally, we require that for each junction segment J_{ik}, its estimated number of mapped reads should be at least the minimum observed number of mapped reads to any junction. We denote $Low(J)$ as the minimum observed number of mapped reads to any junction and $Low(J) = min\{O_l(J_{ik})|(1 \leq i < k \leq m)\}$. We enforce the following constraint for each junction $J_{ik} \in J$:

$$N_l(J_{ik}) \geq Low(J)$$

Constraints on Isoform Expression Values.

The following constraints ensure that the estimated expression of each isoform I_j is no more than $(1 + \epsilon)$ factor of the upper bound on the isoform expression value, which, in the ideal case, would be defined as the minimum expression value of the isoform's exons and junctions. As the minimum expression value of the exons and junctions of an isoform I_j in this ideal case is $min\{O_l(S)/f(S, I_j)|S \in Seg(I_j)\}$, we have the following constraint for each isoform I_j:

$$E_l(I_j) \leq (1 + \epsilon) \times min\{O_l(S)/f(S, I_j)|S \in Seg(I_j)\}$$

In practice, to estimate the upper bound of the expression of an isoform, we may consider to use the median (rather than the minimum) of expression values of its exons and junctions.

Let $Iso(I_j)$ be the indicator variable denoting whether the isoform I_j is expressed in at least one sample. Thus, $Iso(I_j) = 1$ if the isoform I_j is expressed in some sample; 0 otherwise. Thus, the estimated expression $E_l(I_j)$ of an isoform I_j from sample l^{th} is bounded as below:

$$E_l(I_j) \leq B \times Iso(I_j)$$

where B is an upper bound on the expression levels of all the isoforms.

The Objective Functions

In the first stage, with the above constraints, we try to minimize the number of isoforms which are expressed in at least one sample. Here, we give an example: suppose that we have two samples and three isoforms. Sample 1 has expressed isoforms I_1, I_2 and sample 2 has expressed isoforms I_1 and I_3. The number of isoforms expressed in at least one sample can be determined to be 3 (I_1, I_2 and I_3). As a result, the objective function in this stage is to minimize $\sum_{\{I_j \in I\}} Iso(I_j)$.

There could be many feasible optimal solutions by solving the ILP in the first stage. In the second stage, we try to minimize the difference between the observed and estimated expressions of the exons and junctions with the obtained number of expressed isoforms from the first stage. Thus after obtaining the minimum number of isoforms M from the first stage,

we add another constraint to the ILP as $\sum_{\{I_j \in I\}} Iso(I_j) \leq M$. Now we try to minimize the absolute error between the observed number of reads and the estimated number of reads:

$$\sum_{1 \leq l \leq s} \sum_{S \in E \cup J} |N_l(S) - O_l(S)|$$

In order achieve this through ILP, for each absolute value $|a$ above, we add the following constraints to the ILP:

$$-e \leq a \leq e$$

$$0 \leq e$$

and also add e in the objective function of the ILP.

3.2 Experimental Results

We evaluate the performance of CLIIQ based on simulated datasets, and compare results between single sample and multiple sample formulations. We also provide isoform identification and quantification results from two popular tools, namely Cuffilinks and IsoLasso, for comparison purposes. In particular, we compare CLIIQ with Cufflinks version 1.3.0 and IsoLasso version 2.5.2 with the following parameters.

- Cufflinks: we use default settings.
- IsoLasso: we set *min-frac* = 0.05 (the minimum fraction of reported isoforms should be at least 0.05 in a sample), and *minexp*= 0.2 (the minimum expression level of isoforms) to filter isoforms with low or even 0 expression level.

The ILP formulations for CLIIQ are solved by IBM ILOG CPLEX (version 12.2). We monitor the simulation process and generate the corresponding mapping results such that every read is mapped to exactly one location.

Simulated Data.

We use UCSC hg19 human gene annotations and known isoforms to generate 50-bp single-end reads at coverage 30x. Based on the annotations, the simulator generates reads uniformly at random and randomly assigns expression levels to all isoforms such that the ratio between maximum and minimum expression levels from all isoforms of a gene in single sample is less than 10. Here we only consider perfect matching reads, and evaluate performances of CLIIQ and other tools in this case. Note that we only focused on genes for which the coordinates of the exons have no overlap with that of other genes' exons. Our simulations are based on all 770 genes from chromosomes 1, 2, 3 and 4, which satisfy the above condition and contain two to ten isoforms. These genes have a total of 2151 known

	Cufflinks	IsoLasso	CLIIQ (single sample)	CLIIQ (multiple samples)
First Experiment	32 min	12 min	44 min	85 min
Second Experiment	20 min	7 min	34 min	104 min

Table 3.1: Execution time of various methods on isoform identification and quantification.

isoforms.(It is possible to extend CLIIQ formulation to genes with overlapping exons by partitioning each chromosome to disjoint exonic regions; we leave the exploration of this extension to a later study.)

Experiment Design.

We consider two different experiment designs for different samples in a population. For each experiment, we assess the performance of CLIIQ in single sample and multiple sample settings separately. We analyze whether the multiple sample formulation helps to improve performances. In the first experiment, all known isoforms of a gene are expressed in every sample in the population, but with different expression levels. In the second experiment, for each gene we randomly select at most two isoforms such that these isoforms are only expressed in some but not all samples. For example, suppose a gene has three known isoforms I_1, I_2, and I_3. In the first experiment, every sample has these 3 isoforms expressed. In the second experiment, some samples only have I_1 and I_2 expressed, and others contain I_1 and I_3. Although there is a common isoform I_1 for the whole population, the isoforms I_2 and I_3 are not expressed in all the samples. For both experiments, we generate five different samples for each gene. We evaluate the performance of CLIIQ in these two experiments: (1) single sample: we run CLIIQ to identify and quantify isoforms for each sample separately, (2) multiple samples: we formulate all five samples in a population simultaneously.

Running time.

Since the formulation of CLIIQ is based on ILP, it is important to ensure that ILP programs can be solve in reasonable time. We restrict the maximum running time for each ILP program for each gene to 10 mins and there are only 6/914 genes that requires longer running time. Below we report the execution time of all the methods on all the genes selected from chromosome 1, 2, 3 and 4: As seen in the above table, both formulations of CLIIQ on single or multiple samples still have reasonable running time even though it is slower than other methods.

Isoform Identification.

We define precision, recall, and F-score values as follows. Suppose a method reports N transcripts for a sample containing M transcripts. If there are totally C correctly identified

	0.1	0.15	0.2	0.25	0.3
ID Precision	0.8208	0.8325	0.8562	0.8440	0.8509
ID F-score	0.7993	0.8025	0.8021	0.7974	0.8011

Table 3.2: Performance of CLIIQ on isoform identification for test data with different ϵ values

	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.8011	0.7587	0.8351	0.8831	0.8233	0.7825	0.8588	0.8571
Recall	0.6505	0.6739	0.7353	0.7836	0.7213	0.7381	0.8080	0.8788
F-Score	0.7180	0.7138	0.7820	0.8304	0.7690	0.7596	0.8327	0.8678

Table 3.3: Performance of various methods on isoform identification of ϵ=0.2.

isoforms, these three values can be calculated by

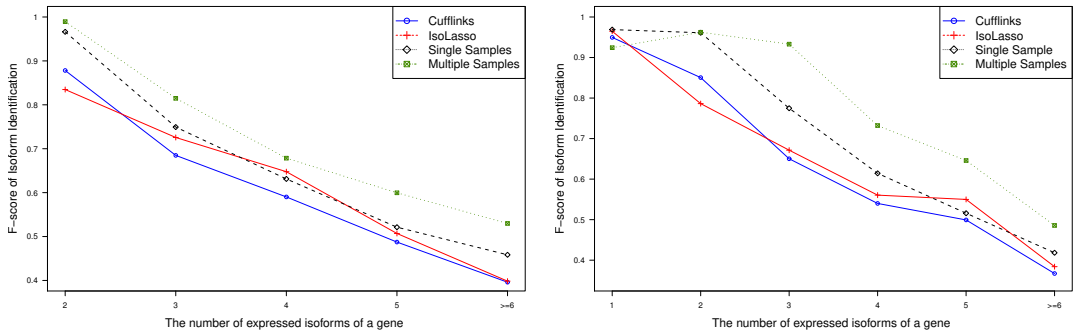
$$\begin{aligned} \text{Precision} &= \frac{C}{N} \\ \text{Recall} &= \frac{C}{M} \\ \text{F-score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

By correct identification between a reported and an isoform derived from an experiment, we mean (1) these two isoforms contain an identical set of exons, and (2) the same exon between these two transcripts can differ at most 2 bases for the boundaries. The second rule tries to eliminate the biases for the first and last exons for an isoform.

Since the performance of CLIIQ relies on a user defined error threshold ϵ, we determine the value for our experiments by running different ϵ of CLIIQ in randomly selected 100 genes. Based on Table 3.2, although ϵ=0.15 can provide best available F-score, we select ϵ = 0.2 which achieves highest precision and the second highest F-score.

The isoform identification results of Cufflinks, IsoLasso, CLIIQ with single sample and multiple samples in two different experiments are shown in Table 3.3.

In single sample formulation, CLIIQ can achieve slightly better identification results than Cufflinks and IsoLasso, and CLIIQ with multiple samples provides more than 10% improvements. More specifically, multiple sample formulation of CLIIQ has similar precision with single sample CLIIQ, but has higher recall rate. It shows that by combining multiple samples, we can retrieve the isoforms effectively. Note that the above results are based on perfect mapping location. For real mapping results using TopHat, multiple sample formulation of CLIIQ also outperforms Cufflinks and IsoLasso as in subsection 3.3.



(a) (b)

Figure 3.1: The F-score of isoform identification of the first experiment(left) and the second experiment(right) with respect to isoform number in a gene.

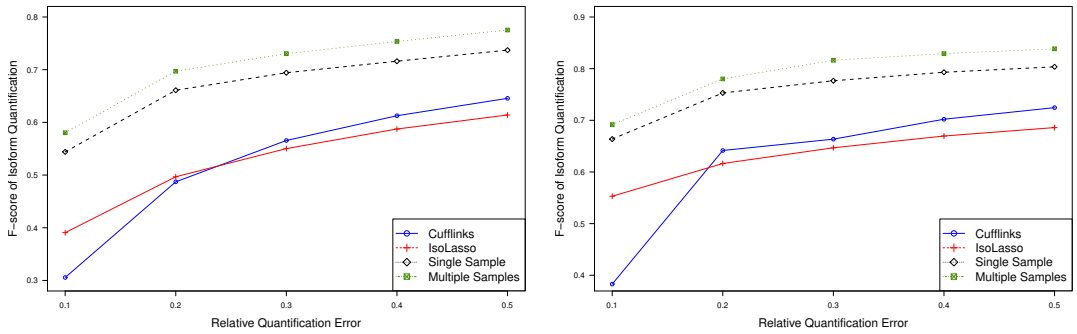
	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.3410	0.4152	0.5810	0.6172	0.4100	0.5697	0.6846	0.6832
Recall	0.2769	0.3688	0.5115	0.5476	0.3592	0.5373	0.4440	0.7004
F-Score	0.3056	0.3907	0.5440	0.5803	0.3829	0.5530	0.6637	0.6917

Table 3.4: Performance of various methods on isoform identification and quantification with error 0.1.

We analyze the performance of isoform identification with respect to the number of expressed isoforms in a gene as in Figure 3.1. Higher number of expressed isoforms is the result of multiple and complicated alternative splicing events. Thus, it is more challenging in identifying and quantifying these expressed isoforms. From Figure 3.1, although the performances of all the methods decrease as the number of expressed isoform grows, multiple sample formulation of CLIIQ still outperforms other tools.

Isoform Quantification.

We follow the definition of precision, recall, and F-score for isoform identification, but we adjust the definition of correct isoforms as follows. A reported isoform is considered as a correctly quantified one if (1) this is a correct identification, and (2) the relative quantification error δ should be less than a threshold. Suppose a reported set of isoforms contains I_1 with expression level 10 and I_2 with 15, and the true answer is I_1, I_2 with expression levels 12, 10 respectively. With the error threshold δ=0.2, we only consider I_1 as a correct quantification ($\frac{|12-10|}{10} \leq 0.2$). I_2 is an incorrect quantification ($\frac{|15-10|}{15} > 0.2$). Table 3.3 show the performance of various methods on isoform identification and quantification. Here we report RPKM values as the abundance estimation results.



(a) (b)

Figure 3.2: The F-score of isoform quantification results (with respect to all reported isoforms) of the first experiment(left) and the second experiment(right) with respect to different error tolerances.

From Table 3.4, we find that in both experiment settings, the quantification performances of single sample CLIIQ are better than Cufflinks and IsoLasso in most metrics given relative quantification error threshold δ= 0.1. Once we combine multiple samples from a population, we can enhance the performance of CLIIQ further, especially for the first experiment in which all samples contain an identical set of expressed isoforms. We also perform the same experiment using real mapping results of TopHat and report in subsection 3.3. And the performance of multiple sample formulation of CLIIQ is better than ones of Cufflinks and IsoLasso.

In Figure 2, we show the performance of isoform quantification of all the methods when δ ranges from 0.1 to 0.5. From Figure 2, we find that the curve of CLIIQ and IsoLasso become smooth since $\delta = 0.2$, but Cufflinks has significant improvement when we increase δ. It shows that CLIIQ and IsoLasso provide more accurate estimation values by considering identification and quantification factors simultaneously.

3.3 Comparisons Based on TopHat Mappings

In previous experiment we assume that all reads can be uniquely mapped to the correct location. In other words, even a read is split into several parts due to junctions, we still provide information in the mapping results such that all methods do not have to handle with problems of ambiguous or disgarded reads. However, in real datasets we may have several problems with split reads: such reads might be missed due to short anchor size, or possess multiple locations of mapping since different exons have similar prefixes/suffixes. These issues, which come from the difficulty of splice junction mapping, will decrease the performance of CLIIQ. To reflect these problems of splice mapping, we use the mapping results generated by TopHat (Version 1.4.1), instead of the perfect mapping results used

	0.3	0.35	0.4	0.5	0.6
ID Precision	0.5758	0.5777	0.5799	0.6001	0.5990
ID F-score	0.6294	0.6285	0.6287	0.6451	0.6419

Table 3.5: Performance of CLIIQ on isoform identification for test data with different ϵ values for real mapping results.

	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.7630	0.6282	0.7331	0.7929	0.7759	0.6203	0.7779	0.7818
Recall	0.6348	0.5903	0.6309	0.6995	0.6996	0.6394	0.7209	0.7642
F-Score	0.6930	0.6087	0.6782	0.7433	0.7358	0.6297	0.7483	0.7729

Table 3.6: Performance of various methods on isoform identification of ϵ=0.5 based on mapping results of TopHat.

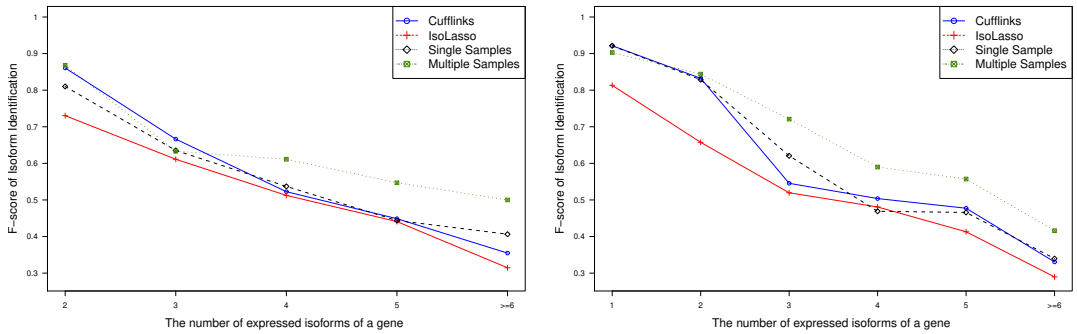
before, and provide the performances of all the methods below. As described in text, we select the error tolerance, ϵ, by running the CLIIQ on 100 randomly selected genes with different ϵ values and select the ϵ value which has the highest precision and F-score. Table 3.5 provides the F-score and precision for different values of ϵ of CLIIQ. We select 0.5 for the remained experiments as CLIIQ has the best F-score and precision.

We first consider the performance of isoform identification given mapping results of TopHat. We provide precision, recall, and F-score of isoform identification in Table 3.6. In both experiments, single sample formulation of CLIIQ basically achieves comparable results to Cufflinks, and better than IsoLasso. Multiple sample formulation of CLIIQ provides better results than Cufflinks.

For isoform quantification, single sample and multiple sample formulation of CLIIQ perform better than other tools in the first experiement. For the second experiment, CLIIQ performs better than Cufflinks, but similar to IsoLasso.

	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.3139	0.3221	0.3672	0.3935	0.3530	0.4371	0.4460	0.4537
Recall	0.2611	0.3026	0.3160	0.3472	0.3182	0.4505	0.4318	0.4435
F-Score	0.2851	0.3121	0.3396	0.3689	0.3347	0.4437	0.4482	0.4485

Table 3.7: Performance of various methods on isoform identification and quantification with error 0.1 based on mapping results of TopHat.



(a) (b)

Figure 3.3: The F-score of isoform identification of the first experiment(left) and the second experiment(right) with respect to isoform number in a gene

We then examine the performance of different methods with respect to the number of expressed isoforms in a sample. In Figure 3.3, we see that in both experiments, single sample formulation of CLIIQ performs similar to Cufflinks. Moreover, as in perfect mapping case, for samples with higher number of expressed isoforms, multiple formulation of CLIIQ outperforms other tools.

3.4 Discussion

The improvement provided by CLIIQ can be attributed to two primary aspects. First, performing identification and quantification simultaneously seems to help considerably. Second, when the RNA-Seq data samples considered contain similar sets of isoforms, it is possible to recover individual isoforms missed in an individually analyzed sample (due to coverage issues, possibility of alternative solutions etc.) through the joint analysis of all samples. We observed that, because of its own "brand" of maximum parsimony objective function, Cufflinks typically reports fewer isoforms than those that are present in the data set. IsoLasso on the other hand tends to generate several false positives CLIIQ provides a better F-score than the alternative tools because it combines these two factors in a single objective function. Moreover, multiple sample formulation of CLIIQ targets to find the most parsimonious set of isoforms for the entire sample set, and not just for one sample.

Chapter 4

Resolution of RNA-Seq Mapping Ambiguity

Owing to the presence of paralogs and homologous regions within a gene, RNA-Seq mappers typically report a fraction of multireads, i.e. reads that map to multiple loci on a reference genome. Based on TopHat mappings on the human reference genome, ∼10% of human RNA-Seq reads are multireads. Similarly, more than 17% of mouse and 50% of some plant RNA-Seq reads are multireads [71]. The presence of multireads complicates the downstream analysis such as determining alternative splicing patterns, gene fusions and other variations.

The common practice for handling multireads is ignoring them in the downstream analysis. This leads to inaccurate estimation of the abundance of expressed transcripts [71, 101]. A simple approach for determining the exact genomic location of a multiread is 'RESCUE' [96]. Here, the initial gene expression values are calculated based on the unique reads that map to them. Each multiread is then assigned to the gene with the fraction equal to the ratio between the gene's initial expression value and the total expression value of all genes that the multiread maps to. A more complex approach based on expected maximization (EM) [105] is designed to handle mapping ambiguity of a read to two or more homologous genes—for the purpose of determining the expression value of each of these genes and not to determine or quantify isoforms. Finally, RSEM [71, 70], IsoEM [101] and iReckon [92] are EM methods based on statistical generative models for sequencing processes to resolve mapping ambiguity.

Many of the above approaches are designed specially for estimating expression values of *known/annotated isoforms*. Their performance is highly dependent on the completeness of the isoform database in use. Furthermore, they cannot handle alternative splicing events such as novel exon skipping, alternative $5'$ donor and $3'$ acceptor sites, intron retention and other structural differences such as insertions or deletions. Some recent computational approaches, in particular IsoLasso [75], CLIIQ [78] and Cufflinks [147], can identify and quantify unknown isoforms and certain types of transcriptomic variations. Unfortunately,

neither IsoLasso nor CLIIQ takes into account multireads, and Cufflinks handles multireads through a simple RESCUE-based approach.

In this chapter, we show how to resolve the multimapping ambiguity in the presence of novel isoforms involving exon skipping, intron retention and small indels towards accurate downstream analysis. To be mathematically precise, we introduce the notion of a partial transcript—a substring of a potential transcript product of a gene, which satisfies certain conditions (a formal definition is provided in the next section).

The objective of our multiread resolution approach, ORMAN [23] (Optimal Resolution of Multimapping Ambiguity of RNA-Seq Reads), is (i) to compute the minimum number of partial transcripts that cover all the multireads and (ii) to assign each multiread to one of these partial transcripts such that each partial transcript is covered according to the estimated local distribution. We achieve the first objective approximately through a reduction to the standard set cover problem. We achieve the second objective through an integer linear programming formulation, which we handle using available integer linear program (ILP) solvers such as CPLEX, or through greedy heuristics we describe in this chapter.

We evaluate ORMAN on both simulated and real human RNA-Seq datasets. For the first experiment, we generate paired-end RNA-Seq reads from a random subset of transcripts from the University of California, Santa Cruz (UCSC) database with the expression distribution modelled after a real human dataset. On this simulated data, we show that the performance of state-of-the-art methods for identifying and quantifying transcripts such as CLIIQ, Cufflinks and IsoLasso is typically improved through the use of our multiread resolution approach. Notably, when combined with IsoLasso or CLIIQ, ORMAN gives the most accurate and comprehensive novel isoform detection and quantification pipeline available.

To evaluate ORMAN in a more 'real world' setting, we also design an experiment using real RNA-Seq data from a cancer patient [67]. For this experiment, we implant artificial genomic repeats into several genes and compare the performance of ORMAN with that of RESCUE in resolving the multireads mapping to these regions. We show that on this dataset, the multiread assignment by ORMAN approximates the original distributions quite well with a maximum relative error of $\leq$0.3.

ORMAN is available at `https://github.com/sfu-compbio/orman`.

4.1 METHODS

Online databases such as the UCSC browser provide known transcripts from specific gene regions. Let $T = \{T_1, T_2, ..., T_p\}$ be the set of known transcripts from a gene region GR. Each transcript T_i is a string that can be partitioned into 'exonic segments' $E(T_i) = \{E_1, E_2, \ldots, E_{|E(T_i)|}\}$. We define the 'gene model' GM implied by the set of transcripts T as an ordered set of alternating substrings of the gene region called canonical exons and

canonical introns. Each exonic segment is a maximal substring of a canonical exon, which is either completely present in a transcript or excluded by that transcript. We refer the readers to Figure 4.1 for an illustration of a gene model derived from known transcripts.

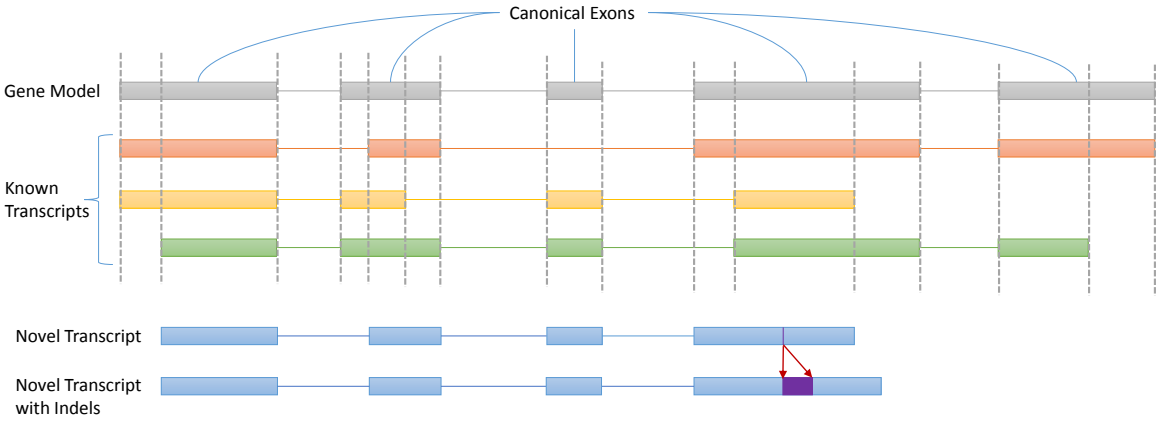


Figure 4.1: A gene model, known transcripts (KT) of the gene model, a novel transcript (NT) derived from known transcripts and a novel transcripts with indels (NTID). Note that the latter may also be derived from known transcripts

Given a read R mapping to a gene region GR, the partial transcript PT supported by R is the shortest substring of the gene model that completely covers R, which starts and ends with an exonic segment (or a canonical intron in the case of intron retention). If there is a small insertion or deletion (our method limits the size of each indel to 15 nt) between the read and the reference sequence, we introduce a modified partial transcript with the corresponding indel. Figure 4.2 illustrates several examples of partial transcripts derived from read mappings to a gene model.

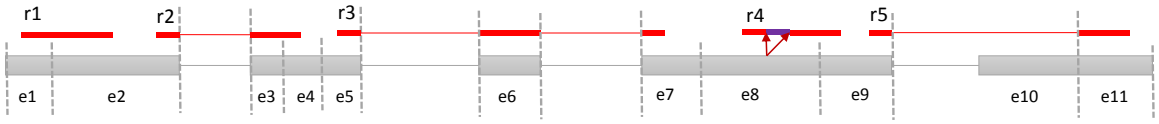


Figure 4.2: Example reads mapping to the gene model of Figure 1. The partial transcripts derived from these reads are as follows: r1:{e1,e2}; r2:{e2,e3,e4}; r3:{e5,e6,e7}; r4:{e8^{ins},e9} and r5:{e9,e11}. Above, e8^{ins} denotes exon 8 with the implied insertion

4.1.1 Combinatorial optimization formulation

The set of partial transcripts present in a sample can be derived from the mappings of RNA-Seq reads to a reference genome with the supply of transcript annotation or to a reference transcriptome. Depending on the given mapping to the reference genome or transcriptome, our objective is to assign each multiread to a single locus on the genome or a transcript. We also need to determine the partial transcript that the multiread should map to. This is done in two phases. In the first phase, we are interested in the **minimum** number of

partial transcripts that could cover **all** the (multi)reads. In the second phase, we try to distribute the (multi)reads to the set of partial transcripts from the first phase such that the distribution of mappings for each partial transcript follows the most likely distribution.

First phase

Let $C = \{PT_1, PT_2, .., PT_p\}$ be the collection of all the partial transcripts derived from mapping results. We also denote by PT_i $(1 \leq i \leq p)$ the set of all reads that support the same partial transcript PT_i. In addition, each PT_i is assigned a positive weight that is proportional to the number of splicing events, i.e. exon skipping and intron retention events with respect to the known transcript it is associated with. For the partial transcripts without any variations with respect to their associated known transcripts, the weight is 1. Each variation adds a user-defined fixed value to this weight. Our default weight contribution of exon skipping is 100 and of indel is 10000. Indel events have high weight due to their significantly low relative frequency [57]. Note that, in the case of paired-end reads, each end of a fragment may be assigned to a different partial transcript. In that case, we assign such a pair to the partial transcript formed by taking the union of the exons from the partial transcripts on both ends. The weight of this new partial transcript PT_i is assigned as the sum of the weights of the two partial transcripts it is composed of. We aim to determine the minimum-weighted set of partial transcripts that can cover all the reads. This problem can be defined as an instance of the minimum-weighted set cover problem, where sets are represented by the partial transcripts, and reads represent set elements. Because the minimum-weighted set cover problem is NP-Hard, we use the standard greedy algorithm which provides a logarithmic factor approximation guarantee [20] to solve this problem and obtain the set of partial transcripts used for the smoothing step.

Second phase

First, we give the formulation of the problem in this phase in terms of an ILP below. We then show the computational complexity of the problem. Finally, we show how to solve the problem in practice.

Let $C' = \{PT_1, PT_2, .., PT_{p'}\}$ be a set of partial transcripts returned from the first phase. For the partial transcript $PT_j \in C'$, we aim to distribute multireads across the partial transcript such that the coverage function of the reads in each partial transcript resembles the most likely distribution. In the case of paired-end reads, we use both ends for the coverage determination. For a read R, let $SPT(R)$ be the set of partial transcripts that R could map to.

Now let len_R be the read length and $len(PT_j)$ be the length of the partial transcript PT_j. Let R_{ij} $(1 \leq i \leq |R|$ and $1 \leq j \leq |SPT(R_i)|)$ be indicator variable, where $R_{ij} = 1$ means that we assign R_i to the partial transcript PT_j; otherwise $R_{ij} = 0$. We enforce that

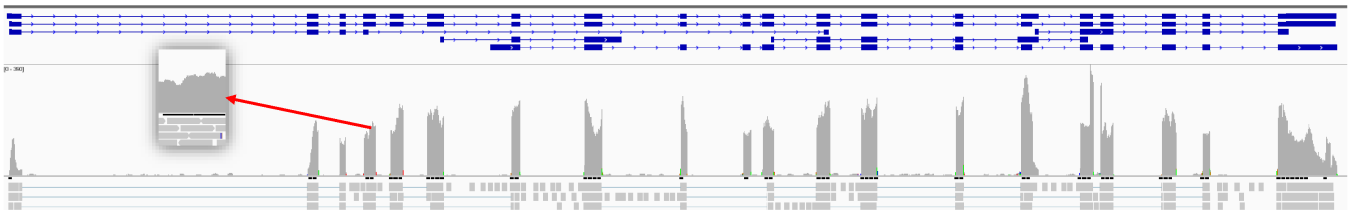


Figure 4.3: The read distribution of gene USP5 taken from a real RNA-Seq dataset (see Section 4.2.2 for details). Although the overall sequence coverage varies significantly along the gene, a small region often coincides well with its neighbourhood

R_i can only be assigned to one partial transcript:

$$\sum_{\{j|PT_j \in SPT(R_i)\}} R_{ij} = 1 \tag{4.1}$$

Let NR_{jk} ($1 \leq j \leq p'$ and $1 \leq k \leq len(PT_j)$) be the number of reads that cover position k in PT_j. Let $Multi(PT_j, k)$ be the set of the multireads that cover the position k in PT_j. In similar manner, we define $Unique(PT_j, k)$ to be the number of reads that are uniquely mapped and cover the position k in PT_j. NR_{jk} could be written as the summation of the number of uniquely mapped reads and multireads that cover the location k:

$$NR_{jk} = Unique(PT_j, k) + \sum_{\{i|R_i \in Multi(PT_j,k)\}} R_{ij} \tag{4.2}$$

Let AV_{jk} be the desired number of reads covering the position k in the partial transcript PT_j. Because we do not know the original distribution of the reads, we approximate AV_{jk} as follows. First, we find the multimapping region M_k of PT_j which encompasses position k. Next, we calculate the average coverage in the left and right neighbourhoods of M_k (the size of each neighbourhood is set to $h \times len_R$ base pairs, where h is a user-defined parameter). We use the calculated average values as defining points for a line l, which approximates the desired function AV. Then, AV_{jk} is calculated as a value on the line l at the position k. The rationale of this approach lies in the observation that coverage level of a small region is often similar to the level of the immediately neighbouring regions, even when the coverage varies significantly along the entire gene (see Figure 4.3).

Let $d_j \geq 0$ denote the maximum difference between the desired number of reads AV_{jk} per position of partial transcript PT_j and the observed number of reads NR_{jk} at any position k. We enforce the following constraints:

$$-d_j \leq AV_{jk} - NR_{jk} \leq d_j \tag{4.3}$$

Our objective is to minimize the total difference:

$$\sum_{1 \leq j \leq p'} d_j \tag{4.4}$$

The problem of smoothing of the distribution of reads along partial transcripts, named SMOOTH, is provably hard. In addition, it is unlikely to have a constant factor approximation algorithm for the SMOOTH problem. The proofs are described below.

NP-completeness of SMOOTH problem We introduce the l-SMOOTH decision problem which asks whether one could distribute reads to partial transcripts such that the value of the objective function is less than or equal to l. We prove that the l-SMOOTH decision problem is NP-Complete by a reduction from the decision problem whether an r-uniform hypergraph ($r \geq 3$) contains a perfect matching, which is known to be NP-Complete [58,].

A hypergraph H is an ordered pair $H = (V, E)$ where V is the set of vertices and E is a collection of distinct non-empty subsets of V. A hypergraph is r-uniform if every edge contains exactly r vertices. A subset $M \subset E$ is a matching if for any two distinct edges $e_1, e_2 \in M$ it implies $e_1 \cap e_2 = \emptyset$. A matching M is called perfect if the union of all the edges in M is V.

Theorem 1. *The 0-SMOOTH decision problem is NP-Complete.*

Proof. The 0-SMOOTH decision problem clearly belongs to NP. Now we only need to provide a mapping from an instance of perfect matching in r-uniform hypergraph ($r \geq 3$) to an instance of the 0-SMOOTH decision problem. Given an r-uniform hypergraph $H = (V, E)$ ($r \geq 3$), we are asked whether H has a perfect matching. We build an instance of 0-SMOOTH decision problem as follows. The set of reads is $R = \{R_1, R_2, ..., R_{|V|}\}$ of size 1, and the set of partial transcripts is $C' = \{PT_1, PT_2, ..., PT_{|E|}\}$. Vertex v_i in H corresponds to read R_i ($1 \leq i \leq |V|$) and similarly edge e_j in H corresponds to partial transcript PT_j ($1 \leq j \leq |E|$). We say that each partial transcript has exactly r positions, at which there is at least one read mapping. For each edge e_j, let $\{v_j^1, v_j^2, ..., v_j^r\}$ be the set of vertices of e_j, and $\{R_j^1, R_j^2, ..., R_j^r\}$ be the corresponding set of reads. We set $Reads(PT_j, k) = \{R_j^1, R_j^2, ..., R_j^r\}$.

Since H is r-uniform, e_j contains exactly r vertices, which means that corresponding PT_j has exactly one read at each position. Suppose that we have a perfect matching in the instance $H = (V, E)$. This implies that $AV_{jk} - NR_{jk} = 0$ for all $1 \leq k \leq r$, $1 \leq j \leq |V|$, where we took $AV_{jk} = 1$ in every point.

On the other side, if $\forall k, j: \ AV_{jk} - NR_{jk} = 0$, then the partial transcript PT_j contains exactly r reads and all reads are covered, which implies that we have a perfect matching of the hypergraph H.

Thus the 0-SMOOTH decision problem is NP-Complete. □

It is unlikely that a constant factor approximation algorithm for the SMOOTH problem exists, because if there were such approximation algorithm with polynomial running time for the SMOOTH problem, then it could be used for solving the 0-SMOOTH decision problem. Thus we obtain the following corollary.

Corollary 1. *There is no k-approximation ($k \geq 1$) algorithm for the SMOOTH problem unless P=NP.*

Practical implementation

The ILP formulation of ORMAN is solved by IBM ILOG CPLEX. In practice, the running time of the proposed ILP depends on the number of integer and non-integer variables and the number of constraints. The number of integer variables of the provided ILP is proportional to the number of mappings of multireads which can be in the order of millions. Here we propose a strategy to decompose the original problem into smaller subproblems such that the solution of each smaller one is independent from each other. We create a graph $G_{PT} = (V_{PT}, E_{PT})$ among the partial transcripts returned from the first phase, i.e. $V_{PT} = C'$. There is an edge between two partial transcripts if there is a multiread r mapping to both of them. It is easy to see that the solution of the ILP corresponding to each connected component in G_{PT} is independent from the solutions of other components. Thus, we can obtain the solution for each component separately using CPLEX. There may still exist some components that could not be solved using CPLEX. In these cases, we propose a heuristic strategy to assign RNA-Seq reads: for each partial transcript in such components, we first calculate on average how many reads would start at a location across the whole partial transcript. If the number of reads starting at a specific location p is higher than the average value, we randomly locate one such read to its alternative mapping location p', as long as p' would not become the location with most reads starting at it in each relevant partial transcripts containing p'. We will continue this greedy procedure until we can not relocate any reads, or until we reach a pre-defined iteration limit for the current component. The strategy can be mathematically described as follows.

Let C be a set of partial transcripts. Define a $S_{PT}(p)$ as a number of reads that start at the position p in a partial transcript PT. Let

$$AS_{PT} = \sum_{p \in \{1 \ldots \text{length}(PT)\}} \frac{S_{PT}(p)}{\text{length}(PT)}$$

. We call the **Smooth** procedure for the set of transcripts C, and we limit the number of iterations by a *threshold* variable as shown in the following pseudocode:
Smooth(C**,** *threshold***)**
while $i <$ *threshold*:

1. set *updated* = false

2. for each partial transcript $PT \in C$:
   (a) $p = \text{argmax}_{p \in \{1 \ldots \text{length}(PT)\}} \{S_{PT}(p) \mid S_{PT}(p) > \max\{AS_{PT}, 0\}\}$
   (b) while $S_{PT}(p) > \max\{AS_{PT}, 0\}$:
      i. select read r from set of reads that start at position p in PT
      ii. if there exists partial transcript PT' such that r belongs to PT' at position p' and $S_{PT'}(p') < \max_{p_i \in \{1 \ldots \text{length}(PT)\}} \{S_{PT'}(p_i)\}$:
         A. assign read r from PT to PT'
         B. update $S_{PT}(p), S_{PT'}(p')$
         C. update $AS(PT), AS(PT')$
         D. set $update = \text{true}$
3. break if *updated* is false
4. set $i = i + 1$

4.2 Experimental Results

We evaluate the performance of ORMAN on both simulated and real datasets. On both types of data, we show that ORMAN resolves mapping ambiguity of multireads accurately and improves the performance of the leading transcript identification and quantification tools.

4.2.1 Transcript identification and expression quantification in simulated data

First, we focus on quantifying how much ORMAN improves downstream analysis tools. We compare the performance of the leading transcript identification and quantification tools by (i) first running each tool without any pre- or postprocessing (ORIGINAL), and (ii) then running each tool after preprocessing the mappings by ORMAN.

Because there are no real world benchmark datasets that provide comprehensive and accurate information on all transcripts and their abundance levels validated by wetlab techniques, we use simulation data for this evaluation. Even though the MAQC project [129] used RNA-Seq technologies to quantify the expression of a limited number of genes, a significant number of these genes have a single isoform and have unique sequence composition [71].

Simulation data

We generated RNA-Seq reads of human transcripts with expression distribution similar to one derived from a real dataset from the GEO database (accession number GSM759513). This dataset comprises paired-end 50-bp RNA-Seq reads of a prostate tissue from Illumina Human BodyMap 2.0 project [127]. The reference transcriptome has 76969 transcripts based

on the UCSC database. We used TopHat version 2.0.7—with the number of mismatches at most 2—to obtain the mappings of the RNA-Seq reads to the reference sequence (version hg19). We ran IsoEM to quantify the expression profile of the UCSC reference transcriptome and determined that 39388 of them are highly expressed.

For the simulations, we assigned one random transcript out of all 76969 transcripts to each one of the expressed transcripts of the prostate dataset. These randomly assigned transcripts represented the expression of 17956 genes. We then set the expression value of each random transcript to that of the prostate dataset transcript it is associated with. We finally selected 10% of this randomly selected set of the simulated transcripts for the production of novel transcripts; for each such transcript, we randomly skip an exon.

To ensure this transcript is novel, we check whether it is highly similar to other known transcripts. We consider a novel transcript to be highly similar to a known transcript if they have the same number of exons and their percentage sequence similarity is >90%. The novel transcript is then assigned the same abundance level as the original transcript.

We generated 80 million paired-end RNA-Seq reads of 75-bp length from the chosen transcripts. The fragment length is determined based on the normal distribution with a mean of 250 bp and a standard deviation of 25 bp. Each transcript received a number of reads proportional to its predetermined expression level, and each read was picked uniformly at random over all possible starting positions of the transcript. We then randomly introduced sequencing errors in the generated reads according to sequencing error model described in [27]. This model places the majority of mismatch errors towards the $3'$-end of the reads. The error percentage per base was set to be 1%. We used TopHat with the above settings to map the generated reads to the reference genome. Approximately 4% of the generated reads had multiple mapping loci.

Performance evaluation

Our performance evaluation is based on three tools: Cufflinks (version 2.0.2), IsoLasso (version 2.6.0) and CLIIQ (version 0.1.0.2). Cufflinks uses a modified rescue strategy to resolve multireads, whereas the latter two are not capable of resolving multimappings. We run CLIIQ in both its standard mode, where it selects the minimum possible number of isoforms, which minimizes quantification errors, and preference mode, where it prefers known isoforms when there are multiple candidate solutions (abbreviated as CLIIQ_pref below). To measure the relative performance of these tools, we provided the complete UCSC gene annotations and disallowed any novel splice sites while allowing novel exon skipping and intron retention events.

The expression values of transcripts are measured in fragments per kilobase per million mapped reads. For each transcript, we define the 'relative quantification error' produced by a given tool as follows. (i) If the known expression value of the transcript is e and the expression value of the transcript reported by the tool is $\hat{e}$, then the relative quantification

error is $|e-\hat{e}|/e$. (ii) If the tool reports a transcript that is not among the simulated expressed transcripts, the relative quantification error is $+\infty$. (iii) If the tool misses a known expressed transcript, the relative quantification error is 1. Following [71, 101], we first investigate the proportion of transcripts whose relative quantification error is above a threshold.

	Cufflinks2	IsoLasso	CLIIQ	CLIIQ_pref
ORIGINAL	1043	1292	1513	1325
ORMAN	1055	1308	1533	1334

Table 4.1: Number of novel isoforms correctly identified by each tool with and without ORMAN.

For each tool, we also compare how ORMAN affects its performance on detecting novel isoforms. The novel isoforms in our simulation generate reads that are incompatible to any known gene annotations. For existing mapping ambiguity resolving tools that require the full list of known transcripts, these reads might be discarded; hence, novel isoforms with multireads may not be detected. On the other hand, ORMAN allows such reads to be used in the solution. In our experiments, all three tools detect more novel isoforms based on ORMAN mappings as can be seen from Table 4.1.

In Figure 4.4, we see that ORMAN improves the performance of IsoLasso and CLIIQ significantly in both modes, which, in comparison with Cufflinks, return fewer incorrectly quantified isoforms for smaller error thresholds. Overall, Figure 4.4 demonstrates that the combination of ORMAN and CLIIQ_pref provides the best results.

We also report the performance of tools on genes that produce a high proportion of multimapping reads separately. Here, we focus on 3784 genes (expressing 7275 transcripts) to which TopHat mapped reads have the top 20% highest mapping multiplicity (see later).

Figure 4.5 shows the proportion of transcripts whose relative quantification error is above a threshold on this subset. As before, ORMAN improves the performance of IsoLasso and CLIIQ significantly in both modes, which are better compared with that of Cufflinks.

Next we consider the performance of each tool in novel isoform detection for those genes that produce multimapping reads. First, we sort all expressed genes according to their mapping multiplicity (i.e. the proportion of the reads that can be mapped to such a gene, which can also be mapped to other genes). Then for genes ranked in the top 10, 20, 30, 40 and 50%, we examine how each tool performs in detecting novel isoforms. Figure 4.5 demonstrates that, in the case of novel isoforms, all tools benefit from ORMAN mappings. In addition, for those genes whose multiplicity is in top 10% in the sample, ORMAN performs particularly well.

4.2.2 Multimapping resolution in real RNA-Seq data

It has been known that real RNA-Seq experiments often suffer from various biases resulting in a rather non-uniform coverage across a gene model [113, 161]. Unfortunately, modelling

of such complex biases in simulations would be cumbersome. To overcome this problem, we design a controlled experiment with real RNA-Seq reads. For this experiment, we use a previously published RNA-Seq dataset with 51-bp Illumina paired-end reads sampled from a human prostate cancer patient [67]. On this dataset, we introduce artificial repeats in 10 genes based on sequences of other genes. By modifying the sequences of the original reads mapping to the artificial repeats, yet keeping everything else intact, we essentially create a multimapping dataset for which the true coverage distribution is known. We then evaluate ORMAN's performance in resolving these multireads.

The following section explains the experimental setup in detail. In the next section, we elaborate on the experiment results.

Experimental setup

First, we map the reads using TopHat (version 1.3.2) to the reference sequence (hg19) and Ensembl annotations (GRCh37.62). Next, we randomly select 10 'decoy' genes according to the following rules:

1. The gene is annotated to have a single transcript based on the Ensembl annotations.
2. The total gene length (i.e. the sum of all canonical exons) is at least 2000 bp.
3. The gene is sufficiently expressed in the sample, having an average coverage >100.
4. The gene is uniquely mappable (i.e. there are no multireads mapping to the gene model).

Similarly, we randomly select 10 'replacement' genes according to the rules 2, 3 and 4 above. Within each replacement gene, we select a 400-bp region to serve as an artificial repeat. This 400-bp sequence is then used to replace the sequence of a region of the same length in the decoy gene. In other words, in each decoy gene, we create an artificial repeat for which the sequence is taken from a randomly chosen replacement gene. The selected genes and the repeat regions are given in Table 4.2.

In the next step, we identify the reads mapping to the coordinates coinciding with the artificial repeat region in each decoy gene. The sequences of these reads are changed according to the new sequence of the decoy gene. All other reads are kept the same. The entire set of reads is then mapped to the new genome reference and the original Ensembl annotations using TopHat with the same parameters.

Evaluation

In this experiment, we compare ORMAN with the modified version of RESCUE as used in Cufflinks [96, 147]. This modified version calculates the initial gene/transcript abundances first by equally distributing the multireads to each gene they map to. In the second phase, each multiread is distributed in proportion to the relative abundance of each gene as computed in the first phase.

Figure 4.6 shows the relative error of coverage in the artificial repeat regions after resolution with ORMAN and RESCUE. This measure is calculated as:

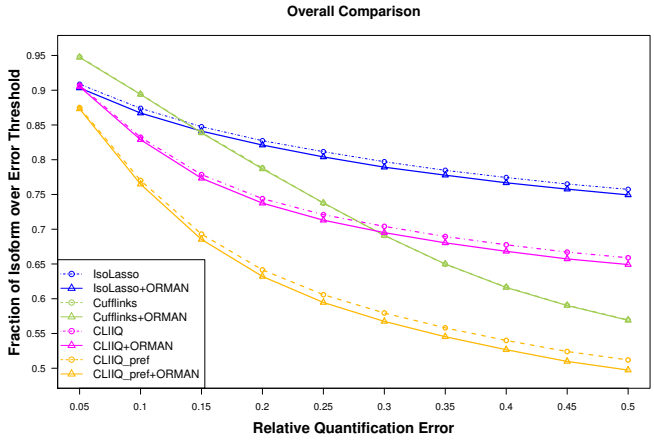
$$\frac{|c_{original} - c_{orman}|}{c_{tophat}} \tag{4.5}$$

where $c_{original}$, c_{tophat} and c_{orman} are the original coverage, raw coverage after the second TopHat mappings and coverage after multiread resolution with ORMAN respectively. The relative error for RESCUE is defined similarly.

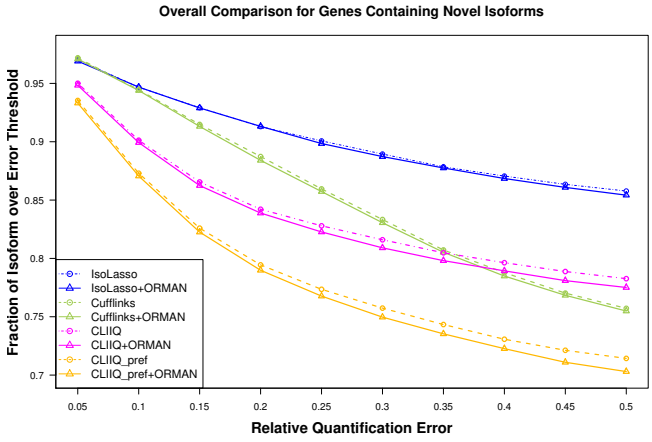
On genes APPBP2, CD164, PPM1H, RCOR1, RYBP, SERPINB6, SSR2, TXNDC16, UQCRC2 and ZBTB42, we see that ORMAN produces lower error values than RESCUE, whereas in the rest of the genes it produces a higher relative error. On the other hand, the relative error of ORMAN never exceeds 0.3. Furthermore, a closer look on some of the genes suggests that ORMAN is still able to reproduce the look of the original distribution quite well despite the fact that RESCUE has a lower relative error. Figure 4.7 illustrates two such genes. Note that although both genes have a high variation in coverage, the coverage distribution in the repeat region is close to the original distribution after processing with ORMAN.

4.3 Discussion

In this chapter, we introduce a combinatorial optimization formulation for resolving mapping ambiguity of RNA-Seq reads. Using a simulated RNA-Seq dataset on humans, we have shown that ORMAN improves the performance of popular computational tools in transcript identification and quantification, especially for genes with novel isoforms. Furthermore, our experiments based on real RNA-Seq reads suggest that the localized approach of ORMAN is able to approximate the original read distribution of the multimapping regions even in genes with highly variable coverage. Although ORMAN's performance was similar to that of RESCUE in our small-scale experiment, we suspect that datasets that suffer from elevated level of sequencing biases such as severe RNA degradation could benefit even more from our approach.

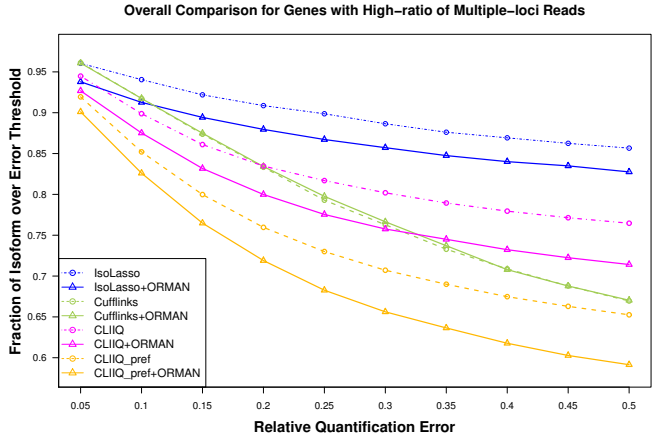


(a)

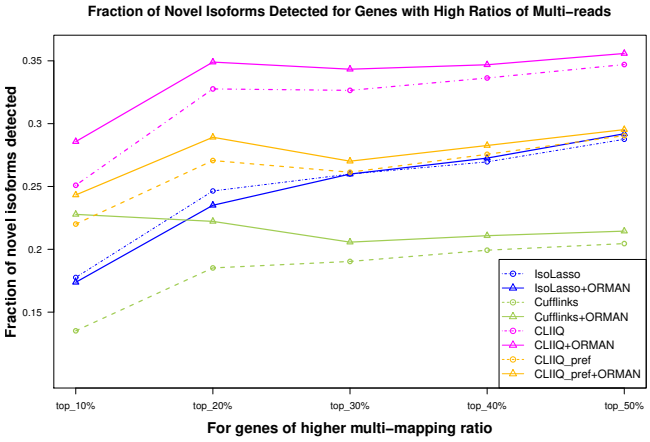


(b)

Figure 4.4: Comparative performance of each tool and its enhanced version with ORMAN measured as the proportion of transcripts whose relative quantification error is above a threshold, as a function of the threshold. We show results of three tools (ORIGINAL) as well as their ORMAN enhanced versions of (a) all 17956 expressed genes (left) and (b) 3148 genes containing novel transcripts (right)



(a)



(b)

Figure 4.5: Comparative performance of each tool and its enhanced version with ORMAN on selected genes that produce multireads, measured as the proportion of transcripts whose relative quantification error is above a threshold, as a function of the threshold. We show results of three tools (ORIGINAL) as well as their ORMAN enhanced versions for 3784 genes containing high ratio of multi-loci reads (left). We also examine the performance of novel isoform detections for gene whose multiread ratio ranked as top 10%–50% in the whole sample (right)

Replacement						Decoy					
Gene	Chr.	Strand	Start	End	# of reads	Gene	Chr.	Strand	Start	End	# of reads
ZBTB42	14	+	105269519	105269918	1026	PPM1H	12	-	63041681	63042080	1839
NFE2L1	17	+	46128078	46128477	3697	UBL3	13	-	30339161	30339560	1949
USP5	12	+	6975253	6975652	931	BCL2L2	14	+	23776979	23777378	708
CD164	6	-	109689719	109690118	9461	TXNDC16	14	-	52898046	52898445	2072
APPBP2	17	-	58522733	58523132	733	RCOR1	14	+	103193777	103194176	902
SERPINB6	6	-	2948403	2948802	6203	UQCRC2	16	+	21994419	21994818	1196
SCAMP2	15	-	75136401	75136800	2386	USP43	17	+	9632438	9632837	838
UBE2K	4	+	39780509	39780908	1724	MUL1	1	-	20827015	20827414	1108
SSR2	1	-	155978849	155979248	7319	RYBP	3	-	72426808	72427207	289
COPG	3	+	128996147	128996546	6110	STK38	6	-	36462615	36463014	935

Table 4.2: The genes and the artificial repeat locations used in the experiments. Note: '# of reads' denote the initial number of reads mapping to the 400-bp region that is used to introduce the artificial repeats.

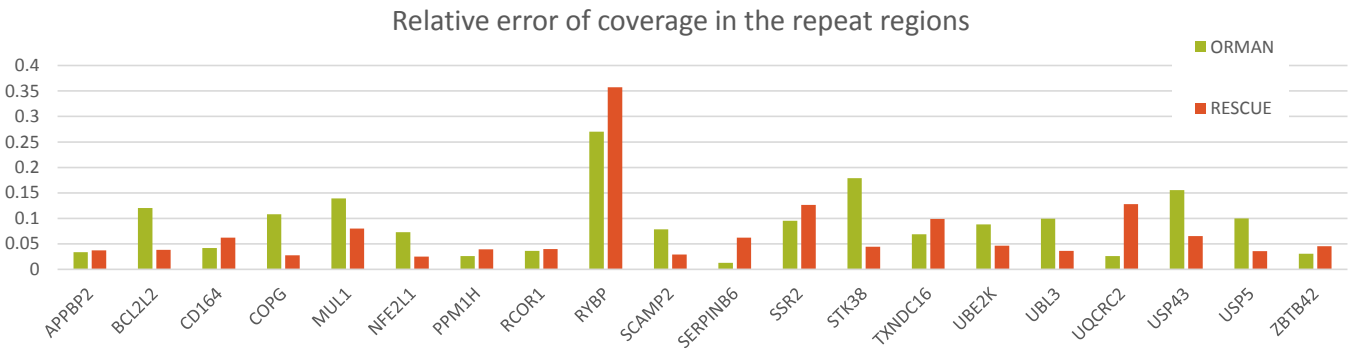


Figure 4.6: The relative error of coverage in the repeat regions after multimapping resolution by ORMAN and RESCUE in the decoy and replacement genes used in the experiments

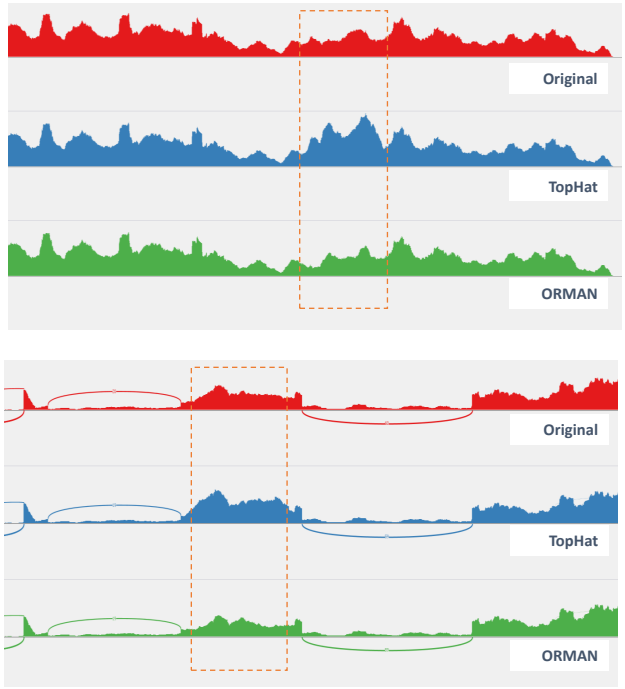


Figure 4.7: Coverage plots for two genes (left: ZBTB42, right: BCL2L2) before and after processing with ORMAN compared with the original mappings. Top track (red) shows the mappings in the unaltered dataset; middle track (blue) shows the TopHat mappings after artificial repeats are introduced and the bottom track (green) shows the mappings after multiread resolution with ORMAN. The boxes outlined with dashed orange lines depict the artificial repeat region

Chapter 5

Proteogenomics

Rapid advances in high throughput sequencing (HTS) and mass spectrometry (MS) technologies has enabled the acquisition of the genomic, transcriptomic and proteomic data from the same tissue sample. The availability of three types of fundamental omics data provide complementary views on the global molecular profile of a tissue under normal and disease conditions [100]. Recently developed computational methods have aimed to integrate two or three of these data types to address important biological questions, such as (i) correlating the abundances of transcription and translation products [104]; (ii) detecting peptides associated with un-annotated genes or splice variants (in mouse [103], C. elegans [158], zebrafish [15] and human samples [94, 128]); (iii) characterizing chimeric proteins by searching unidentified tandem mass spectrometry(MS/MS) data through the use of conventional peptide identification algorithms applied to a pre-assembled database of "known" chimeric transcripts from the literature [39].

In the past year or so, several studies have aimed to identify novel peptides matching patient specific transcripts derived from RNA-Seq data. For example, Zhang et al. [170] focused on identifying novel peptides involving Single Amino Acid Variants (SAAVs) in colorectal cancer. A later study by Cesnik et al. [16] also considered novel splice junctions and (a limited set of user defined) Post-Translational Modifications (PTM) in a number of cell lines. Because of the importance of phosphorylation in cellular activity and cancer treatment [112], this was further expanded to identify novel phosphorylation sites by Mertins et al. [91], on the CPTAC breast cancer data set, which is the subject of our study. However, none of these studies aimed to perform integrative analysis of transcribed and translated genomic structural alterations such as fusions, inversions and duplications in tumor tissues. **Genomic structural variants (SVs)** alter the sequence composition of associated genomic regions in a significant manner. Major SV types include (segmental) deletions, duplications (tandem or interspersed), inversions, translocations and transpositions. SVs observed in exonic regions may lead to aberrant protein products. Many such SVs have been associated with disease conditions and especially cancer. Common SVs associated with cancer include deletions in tumor suppressors such as BRCA1/2 [33] in breast cancer, duplications

in FMS-like tyrosine kinase (FLT3) gene in acute myeloid leukemia (AML) [99] and an inversion causing cyclin D1 overexpression in parathyroid neoplasms [50].

A **gene fusion** occurs when exonic regions of two (or more) distinct genes are concatenated to form a new chimeric gene, as a result of a large scale SV. Gene fusions can disrupt the normal function of one or both partners, for example by up-regulating an oncogene (e.g. TMPRSS2-ERG) or generating a novel or truncated protein (e.g. BCR-ABL1 [36]). They have been demonstrated to play important roles in the development of haematological disorders, childhood sarcomas and in a variety of solid tumors. For example, ETS gene fusions are present in 80% of malignancies of the male genital organs, and as a result these fusions alone are associated with 16% of all cancer morbidity [93]. Others, including the EML4-ALK fusion in non-small-cell lung cancer and the ETV6-NTRK3 fusion in human secretory breast carcinoma occur in much lower frequency [142, 134]. The discovery of such low-recurrence gene fusions may be of significant clinical benefit since they have potential to be used as diagnostic biomarkers or as therapeutic targets - if they encode novel proteins affecting cancer pathways [6, 149, 116].

There are a number of available computational tools for detecting structural variants, each based one or more of the following general strategies. (1) Detection of variants using discordantly mapping paired end reads, more specifically read mappings that either invert one or both of the read ends, or change the expected distance between the read ends. Tools using this approach include Breakdancer [34] and VariationHunter [52]. (2) Detection of variants using split-read mappings - which partition a single end read into two and map them independently to two distant loci - or soft-clipped read mappings - which map only a prefix or suffix of a read. One example employing this approach is Socrates [124]. (3) Detection of variants using an assembly based approach. These tools map assembled contigs for improved precision. Examples include Barnacle [138] and Dissect [165] (both of which happen to be RNA-Seq analysis tools, but can also be used to analyze genomic data). Additional tools employing a combination of these strategies include Pindel [164], Delly [111], GASVPro [133] and HYDRA [110].

One of our focus in this chapter is microSVs (micro structural variants), i.e. events involving genomic sequences shorter than a few hundred bps, especially in exonic regions, since they are more likely to result in a translated protein. Available tools for SV discovery typically fail to capture microSVs, or do so while producing many false positives, thus the problem of robustly discovering microSVs remain open.

In contrast to microSVs, gene fusions can be inferred at a large scale by detecting chimeric transcripts in RNA-Seq data [82]. Currently, there are two general computational approaches to detect gene fusions. (i) The mapping-based approach (e.g. deFuse [84], FusionMap [43], FusionSeq [122], ShortFuse [63], SOAPfuse [55], and TopHat-Fusion [61]) suggests to first map RNA-Seq reads to the reference genome, and then discover fusion transcript candidates by analyzing discordant mappings. More involved methods in this category in-

clude nFuse [86] and Comrad [85], which incorporate WGS (Whole Genome Sequencing) data for more accurate predictions and handling complex fusion patterns that involve three or more genes. (ii) The assembly-based approach such as Barnacle [138] and Dissect [165], on the other hand, suggests to first *de novo* assemble RNA-seq reads into longer contigs by using available transcriptome assemblers (e.g., Trinity [45]), and only then map the assembled contigs back to the reference genome, with the aim of reducing the potential errors introduced by mapping short reads to the reference genome.

Our first contribution in this chapter is a novel algorithmic tool named **MiStrVar** (**Mi**cro **Str**uctural **Var**iant caller), which identifies microSV breakpoints at single-nucleotide resolution by (1) identifying each one-end-anchor (OEA), i.e. a paired-end read where one end maps to the reference genome and the other end cannot be mapped, (2) clustering OEAs based on (i) mapping loci similarity and (ii) the possibility of assembling the unmappable ends into a single contig, and (3) aligning the contig formed by unmappable ends with the reference genome - in the vicinity of the mapped ends - simultaneously detecting putative inversions, duplications, indels or single nucleotide variants (SNVs) through a unified dynamic programming formulation.

MiStrVar approach has several advantages over existing SV discovery tools. Firstly, MiStrVar analyzes many more reads than those considered by the tools using only split-reads or soft clipped reads. Any mapped read which has a hamming distance to the reference greater than four (as a default parameter, which can be user modified) is considered for assembly. This allows for the discovery of inversions or duplications as short as 5bp and inversions with palindromic sequences, improving sensitivity. Secondly, this approach is much less time consuming than assembly based methods, since only the subset of unmappable reads are assembled rather than the entire genome. Finally, MiStrVar uses a unified dynamic programming formulation, superior to tools that identify each type of variant individually, especially because these tools misinterpret certain variants, such as inversions, as a combination of other variants. See Figure 5.1 for a detailed illustration.

Both fusions and microSVs may be independently observed in genomic, transcriptomic, and proteomic data; however, the most impactful aberrations, especially in the context of cancer, are the ones that can be observed in all levels in the same tissue simultaneously. In such cases, integrative analysis of these three omics data types can provide independent evidence for the presence and heritability of aberrations. For example, trans-splicing events, which lead to chimeric transcripts, can only be observed in transcriptomic (but not in genomic) data, and thus can be distinguished from fusion events with genomic breakpoints through simultaneous analysis of genomic and transcriptomic data acquired from the same sample.

The vast majority of large-scale studies of sequence aberrations are based on genomic and transcriptomic data. Most proteogenomics research mainly focuses on detecting single amino acid variants and studying protein abundances affected by single nucleotide variants

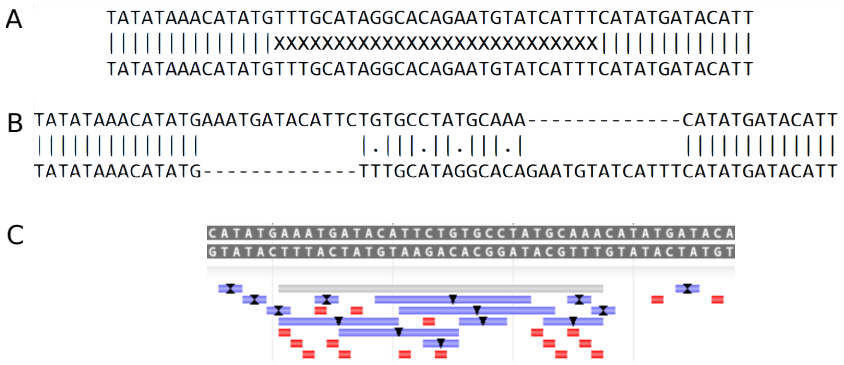


Figure 5.1: Example to illustrate the advantage of a unified algorithm for detecting structural variants. **(A)** Optimal alignment of two sequences using the unified algorithm, in this case a 27bp microinversion in SLC3A1 3'UTR. **(B)** Optimal Smith-Waterman alignment of the same two sequences.**(C)** Cluster of dbSNP entries in the inverted region. Blue lines are indels, red lines are single nucleotide polymorphisms and the grey line is a multiple nucleotide polymorphism. Most, if not all, are incorrect.

[170, 2]. No available large-scale study has been conducted on the detection and validation of aberrant proteins and their genomic and transcriptomic origins. As mentioned earlier, expressed aberrant genome variants can have considerable functional influence on proteins, and as such, they may affect molecular pathway activity or pathogenesis in disease, especially in cancer. Detection of aberrant protein variants provides new insights into diagnostic marker identification and drug development (recurrent protein aberrations can imply potential drug targets) and can help develop novel strategies for therapeutic intervention.

Proteomic technologies have enabled high throughput, sensitive and deep protein analysis for complex disease-associated samples, aiming at discovering potential disease protein biomarkers [97, 44, 162], including low-abundant proteins or protein isoforms, or variants. Moreover, proteomic analyses can provide complementary information to transcriptomic and genomic analysis, as proteomic analyses are carried out by completely different technologies (i.e., mass spectrometry or MS) from DNA sequencing. Furthermore, advancement in MS instrumentation has enabled proteomic analysis to achieve sensitivity on par with RNA-seq in detecting low abundant events of gene expression in complex samples [170]. Therefore, integrating transcriptomic and proteomic data can improve both the sensitivity and confidence in characterizing expressed aberrant variants in complex samples such as tumor tissues.

Our second contribution in this chapter is **ProTIE** (**Pro**Teogenomics **I**ntegration **E**ngine), the first computational framework that integrates high throughput genomic, transcriptomic and proteomic data to identify translated structural aberrations, specifically gene fusions and microSVs, in protein-coding genes. In particular, ProTIE takes sequence aberrations from WGS and RNA-Seq data as its input and validates them on the mass-spectrometry

based proteomics data, while ensuring that each such proteomic signature is unique to the matching sequence aberration. By integrating multiple data sources simultaneously, ProTIE is able to provide a strongly supported set of candidate aberrations from the highly sensitive results of MiStrVar and deFuse. This is particularly helpful for selecting target events or genes for clinical studies.

We ran our computational framework to detect all translated gene fusions in RNA-Seq (low coverage 50bp paired-end) data in the complete set of 105 TCGA (The Cancer Genome Atlas) breast cancer samples for which CPTAC (Clinical Proteomic Tumor Analysis Consortium) mass spectrometry data have been released.[1] These 105 samples include all four of the most common intrinsic subtypes of breast cancer. Among them, 22 samples also have matching WGS data, on which we used our framework to identify exonic microSVs. This resulted in 206,255 fusions and 69,876 microSVs across the 105 samples. 2,215 of these microSVs are also supported by transcriptomic (RNA-Seq) evidence.

All breakpoints from the predicted fusions and microSVs were then analyzed for identifying supporting peptides from mass spectrometry data. This yielded 244 aberrant peptides from 432 possible aberrations. More specifically, 169 novel peptides originate from 295 fusion candidates (many of the fusions are recurrent and thus produce the same novel fusion peptide) and 75 peptides originate from 137 potential microSVs; this is of particular note since many of the genomic microSVs are recurrent, yet the ones that are translated are mostly private. Note that a sequence aberration may give rise to more than one novel peptide in case it results in a frameshift. See Table 5.1 for a summary of results. [2]

5.1 Methods

Our computational framework (see Figure 5.2), is comprised of a number of algorithmic tools that we developed for detecting transcriptomic and genomic aberrations, and searching for expressed protein variants resulting from these aberrant sequences. Given a set of genomic (WGS), transcriptomic (RNA-seq) and proteomic (Mass Spectrometry) data, each collected from the tumor tissue of a patient, our pipeline detects translated *sequence aberrations* in three major steps.

1. Each whole genome sequencing dataset is analyzed with MiStrVar, the microSV discovery tool we introduce in this chapter, to identify microSVs occurring in protein-coding

[1]The primary goal of CPTAC is to characterize protein level expression differences for SNVs/SAAVs. Our focus here is complementary to the goals of CPTAC.

[2]One interesting observation is that among the microSVs discovered, only 4 (specifically 1 microinversions and 3 tandem microduplications) have supporting evidence at all omics levels. This implies that the transcriptomic support for the remaining translated microSVs are too low to be detected, partially due to low abundance of RNA-Seq data made available by TCGA on the breast cancer samples we analyzed. This also suggests that with deeper coverage RNA-Seq data, ProTIE is likely to detect additional translated gene fusions.

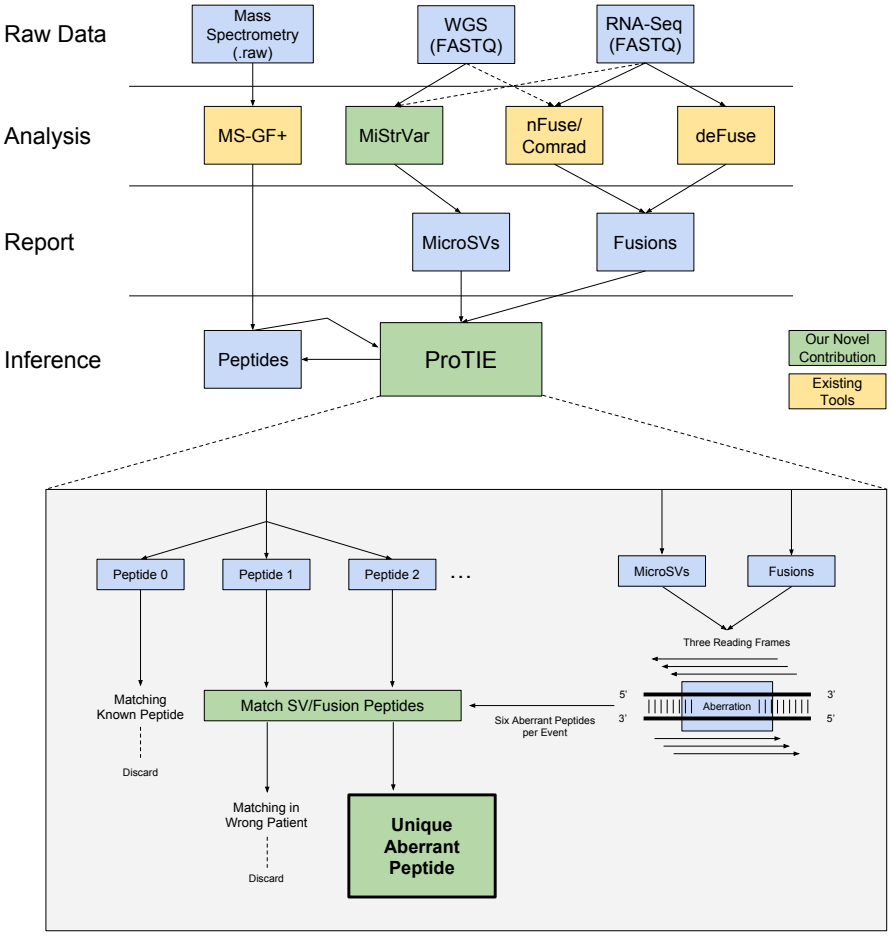


Figure 5.2: Overview of the computational pipeline for identifying translated sequence aberrations. Mass spectrometry data is used to validate fusions detected in the RNA-Seq data and microSVs detected in the WGS data. For tumor samples with matching RNA-Seq and WGS data, our pipeline provides the ability to detect transcribed microSVs and fusions with genomic origins as well. The pipeline introduces MiStrVar, a tool for detecting microSVs from WGS data. It also features our in house developed fusion discovery tool(s) deFuse (as well as nFuse/Comrad), as well as the MS-GF+ mass spectroscopy search engine. The final step is the ProTIE (Proteogenomics integration engine) for sequence and mass spectrometry data: After running deFuse and MiStrVar to respectively identify fusions and microSVs, we generate each possible breakpoint peptide from the 6 distinct reading frames associated with each of these aberrations. For mass spectra from the same tumor sample, we discard those which can be matched to known proteins, and keep only spectra matched to breakpoint peptides identified above. The resulting high quality peptide-spectra matches (PSM) provide proteomics-level evidence for the predicted aberrations.

Cancer Subtype	Total # Patients	# Patients with Aberrant Peptides	# Fusion Peptides	# Inversion Peptides	# Duplication Peptides
Basal-Like	25	22	50	57*	2
HER2-Enriched	18	17	41	3	3
Luminal A	29	26	49	0	2
Luminal B	33	31	78	8	3

Table 5.1: Distribution of 244 detected, high confidence, aberrant peptides over four breast cancer subtypes, across 105 patients. **# Patients with aberrant peptides** indicate the number of patients with either detected fusion peptides or microSV peptides in that subtype. As can be seen, all but one of the patients exhibit at least one translated fusion or microSV. The next three columns respectively indicate the number of peptides detected from fusions, microinversions and microduplications, within specific subtypes. *The high number of microinversion peptides in Basal-Like breast cancer can be attributed to two patients, A0CM, A0J6, whose genomes had gone through substantial reorganization.

genes. (Note that our computational framework provides the option of validating genomic microSVs at the transcriptomic level by identifying RNA-Seq reads associated with each microSV breakpoint.)

2. Each transcriptomic dataset is analyzed by our in-house fusion detection method deFuse [84], which reports potential fusion events between two protein coding genes, and the *fused* transcript sequences spanning the fusion breakpoints. (Note that our computational framework enables the use of our integrative fusion detection methods nFuse [86]/Comrad[85] for corroborating potential fusions observable in WGS and RNA-Seq data.)

3. All omics data is finally integratively analyzed through ProTIE, our novel ProTeogenomics Integration Engine as follows. Each mass spectrometry dataset is searched against a protein sequence database consisting of all human proteins from Ensembl human protein database GRCh37.70 [31], along with a database of proteins generated by fused transcripts and microSVs, by the use of MS-GF+ search engine [62]. Aberrant peptides identified by the procedure with high confidence (e.g., at 1% false discovery rate estimated by using the target-decoy approach [28]) are reported, provided they are also detected in the genomic/transcriptomic dataset from the same tumor tissue sample. (For further validating aberrations identified at multiple omics levels, our computational framework also provides the option of searching for recurrences across multiple tumor samples, possibly representing the same tumor subtype.)

5.1.1 Detection of Fusions and microSVs in WGS and RNA-Seq Data

To detect fusions in RNA-Seq data, we applied deFuse [84] which predicts fusion transcripts based on analyzing discordantly mapped read-pairs and one-end anchors. To detect microSVs in WGS data, we applied our novel micro-structural variant caller, MiStrVar, which works in three major steps (See Figure 5.3 for an overview):

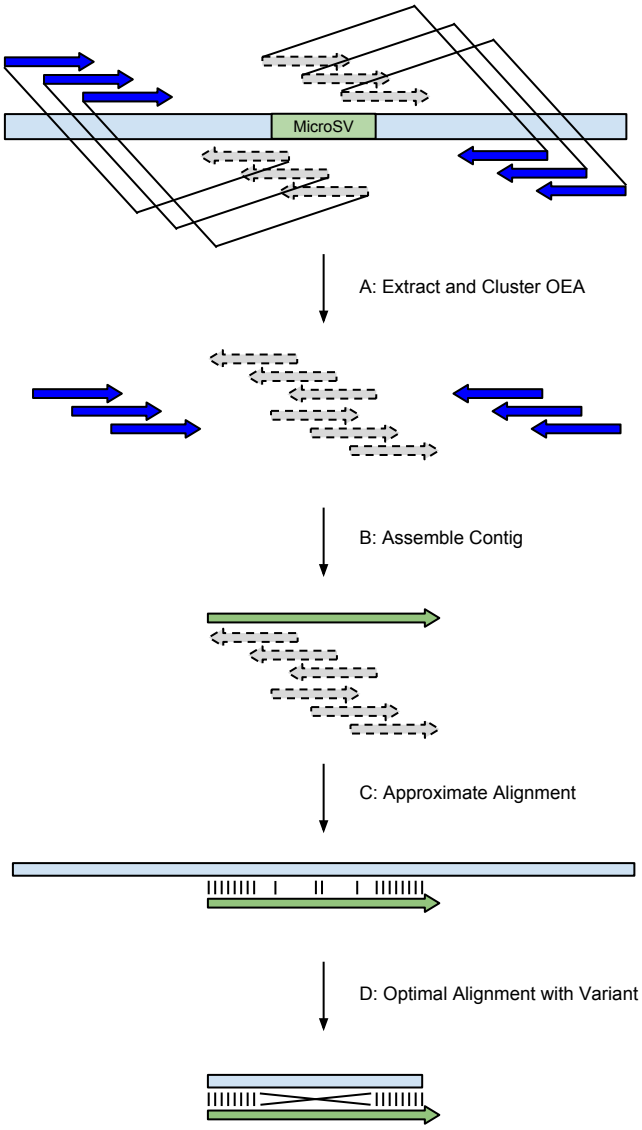


Figure 5.3: A sketch of our computational framework for detecting microSVs in tumor samples. **A.** All one-end-anchors (OEA) are extracted from the mapping file and clustered based on the mapped mates. **B.** The unmapped mates are then assembled into a contig. **C.** The contig is aligned to the reference within 1Kb from the mapped mates. **D.** The reference is clipped and the optimal alignment is found using a dynamic programming formulation allowing for a structural alteration event.

In **step (A)**, MiStrVar identifies all one-end anchors (OEA) in the read data: an OEA is a paired-end-read for which only one end maps to the reference genome within a user defined error threshold. Once all reads are (multiply) mapped to a reference genome using mrsFAST-ultra [47, 48], and all OEAs are extracted, the mapped ends of OEAs are clustered based on the mapping loci. MiStrVar provides the user two options for cluster identification, each satisfying one of the following distinct goals. For applications where sensitivity is of high priority, MiStrVar employs a sweeping algorithm for OEA mapping loci (introduced for VariationHunter [52]). For applications where running time is of high priority, MiStrVar employs an iterative greedy strategy.
In **step (B)**, for each OEA cluster identified in step (A), MiStrVar assembles the unmapped end of the reads to form contigs (of length <400bp in practice) by aiming to solve the NP-hard [41] **dominant superstring (DSS)** problem. MiStrVar employs a greedy strategy similar to that used to compute a constant factor approximation to the shortest superstring problem [12].
In **step (C)**, each contig associated to an OEA cluster is aligned to a region (of length several kilobases long) surrounding the OEA mapping loci, first through a simple *local-to-global* sequence alignment algorithm, that does not consider any structural alteration. (The reverse complement of the contig is also aligned to the same region.) The start and end position of this first, crude alignment is used to determine the approximate locus and length of the potential microSV implied by the contig. The exact microSV breakpoints are obtained in the next step through a more sophisticated alignment that considers structural alterations, which is applied to the portion of the reference genome restricted by the first alignment. The dynamic programming formulation for this alignment is an extension of the Schöniger-Waterman algorithm [123] which was designed to capture inversions in the alignment. Specifically, the extensions enable the user to

1. discover the single best optimal event, rather than an arbitrary number of events,
2. handle gaps extending over breakpoints (in cases of missing contig sequence), and,
3. simultaneously predict duplications, insertions, deletions and SNVs in addition to inversions.

5.1.2 Identification of Translated and Transcribed Sequence Aberrations

ProTIE provides the ability to detect translated aberrations by searching mass spectra against an aberrant peptide database. More specifically, given transcriptomic breakpoints pointing to fusions or microSVs, ProTIE identifies respective aberrant peptides from proteomic data by first generating a peptide database, and then identifying aberrant peptides based on mass spectrometry search results provided by MS-GF+ [62].

Proteomics Search ProTIE provides the ability to detect translated aberrations by searching mass spectra against an aberrant peptide database. More specifically, given tran-

scriptomic breakpoints pointing to fusions or microSVs, ProTIE identifies respective aberrant peptides from proteomic data by first generating a peptide database, and then identifying aberrant peptides based on mass spectrometry search results.

The database is a combination of known (wildtype) human peptides and either the fusion peptides (used for fusion discovery), derived from the fusion breakpoints suggested by deFuse (and/or Comrad/nFuse), or microSVs breakpoint peptides, derived from the breakpoints suggested by MiStrVar. For each fusion or microSV breakpoint, six different reading frames (both forward and backward reading frames) are considered - until a stop codon. Potential peptides (of residue length five or more) resulting from each breakpoint junction, as well as downstream peptides that result from a shift in the reading frame, are included in the peptide database allowing at most one miscleavage site (i.e. consisting of at most two amino acids of K or R for trypsin specificity used in the CPTAC data) as aberrant peptides.[3]

ProTIE uses Ensembl human protein database GRCh37.70 [31] to derive the known (wildtype) human peptides. Note that the Ensembl database includes 104,785 peptide sequences among which 75,994 are annotated as known peptides, 10,449 are annotated as novel peptides and an additional 18,342 are annotated as putative peptides. ProTIE includes only the set of known peptides in the primary peptide database it establishes; however it also provides information for the mass spectra that do not match known or aberrant peptides, but can be matched to novel or putative sequences.[4]

ProTIE conducts peptide identification by searching tandem mass spectra (MS/MS) against the peptide database it sets up, as described above. For that, it first converts raw files into Mascot Generic Format (MGF) and uses MS-GF+ [62] engine to perform the search. We adopted the search parameters recommended by the CPTAC Common Data Analysis Pipeline (CDAP) [118] as follows. (1) The precursor mass tolerance is set to 20ppm. (2) The Fragment method is set as 3 for HCD. (3) Instrument is set as 3 for Q-Exactive. (4) Number of tolerable termini is 1. (5) Maximum length of peptide is 50. (6) Modifications include: Carbamidomythyl is fixed in Cystine, Oxidation is set as variable modification in M, iTRAQ 4plex is fixed at N-termini and any Lysine(K) residue.

To obtain a subset of high confidence matches, ProTIE selects only the spectra where the top 20 peaks in the PSMs have matched fragmentation ions. If fewer than 20 peaks exist, all peaks must match. The major fragmentation ions annotated are: b-, b-neutral loss ions, y-, y-neutral loss ions.

[3]Peptides shorter than five residues are discarded due to mass spectrometry detection range limit.

[4]We establish a single database for the breakpoints identified in all patients, however we maintain the patient for each potential aberrant peptide in order to make sure that mass spectra from a particular patient (or a set of patients) can only match an aberrant peptide from the same patient.

We apply 1% spectrum-level FDR control on the identification as suggested in CDAP. Based on the search result, we keep a spectra in ProTIE if its best PSM (in terms of lowest q-value) can not match any known peptides in Ensembl annotation or decoy sequences (i.e. false positives). Spectra that match novel or putative peptides are still kept with special remarks for further analysis.

Transcriptome Search Our pipeline also provides the user with the additional ability to jointly analyze matching WGS and RNA-Seq data for identifying transcribed genomic (in fact genetic) microSVs. Given a set of genomic microSVs, along with their breakpoints detected by MiStrVar, our pipeline generates corresponding aberrant transcripts. It then maps RNA-Seq reads to the collection of these aberrant transcripts using mrsFAST-ultra (error threshold, 6%). An RNA-Seq read mapping is said to provide evidence for the transcription of a microSV in two ways: either (i) an RNA-Seq read is (uniquely) mapped across a breakpoint - providing evidence for the transcription of the associated microSV (both in the form of inversions and duplications) ; or (ii) a paired end RNA-Seq read is mapped to the reference transcriptome discordantly (due to change of mapping orientation) - again providing evidence for the transcription of the associated microSV (relevant for inversions only). Note that no read that can be mapped concordantly to a known isoform or potential novel spliceform (through the use of the splice-aware mapper STAR [26]) is considered to be a supportive evidence of a transcribed microSV.

5.1.3 Availability

MiStrVar is available for download at `https://bitbucket.org/compbio/mistrvar`, and ProTIE is available at `https://bitbucket.org/compbio/protie`.

5.2 Experimental Results

CPTAC Breast Cancer Dataset. Clinical Proteomic Tumor Analysis Consortium (CPTAC, `http://proteomics.cancer.gov`) [29, 157] aims to provide proteogenomic characterization of specific cancers based on joint analysis of proteomic, transcriptomic, and genomic data acquired from the same group of cancer patients. CPTAC currently focuses on the relationship between protein abundance, somatic mutations and copy number alterations occurring in cancer-related genes [170]. Information about aberrations hidden in unidentified spectra and unmapped sequenced reads have not been revealed in the current CPTAC analysis framework; this happens to be our main focus.

At the time of submission of this thesis, proteomics data for tumor samples from three cancer types had been released by CPTAC: colorectal cancer, breast cancer, and ovarian cancer. In addition, The Cancer Genome Atlas (TCGA, `http://cancergenome.nih.gov/`) has released RNA-Seq and WGS data on both normal and tumor tissues from the same group of patients through Cancer Genomics Hub (CGHub, `https://cghub.ucsc.edu/`).

RNA-Seq data for breast and ovarian cancer patients are in the form of paired-end reads, however, for most of colon and rectal cancer samples only single-end reads were collected. Because we rely on paired-end mappings for detecting fusions and microSVs and since the RNA-Seq data from normal tissues from the ovarian cancer patients had not been released at the time of the submission, our focus in this chapter is the breast cancer dataset. Details about CPTAC samples used in our analysis can be found in Tables 5.2, 5.3.

Table 5.2: Available omics data for TCGA/CPTAC breast cancer samples.

WGS			**RNA-Seq**		**Mass Spec.**	
Solid Tumor	**Blood Normal**	**Solid Normal**	**Solid Tumor**	**Solid Normal**	**Mixture**	**Number of Patients**
✓	✓	✓	✓	✓	✓	**2**
✓	✓		✓	✓	✓	**1**
✓		✓	✓	✓	✓	**3**
✓	✓		✓		✓	**16**
			✓	✓	✓	**10**
			✓		✓	**73**
					Total:	**105**

Breast Cancer Cell Line. In addition to the CPTAC and TCGA datasets, we used the HCC1143 ductal breast cancer cell line (triple negative breast cancer cell line from ATCC) for which we obtained matching tumor/normal Illumina HiSeq WGS, RNA-Seq and mass spectrometry data. The matching normal cell line, HCC1143-BL, is a B lymphoblastoid cell line initiated from peripheral blood lymphocytes from the same patient as HCC1143 by transformation with Epstein-Barr virus (EBV). The WGS data was obtained from NCI Genomic Data Commons (`https://gdc.cancer.gov/`), originally sequenced as part of the Cancer Cell Line Encyclopedia Project [10]. We used this cell line as preliminary validation for our approach before starting full scale analysis.

5.2.1 Gene Fusion Detection by deFuse

Gene Fusions in the HCC1143 Breast Cancer Cell Line. We have run our fusion detection method, deFuse to detect gene fusions on RNA-Seq data from HCC1143 cell line. There are 81.73M paired end reads of 101bp length. Based on concordant mapping results, the average fragment length and standard deviation were 264.2bp and 86.59 bp respectively. deFuse predicts 1,325 fusions from this dataset, out of which 74 are considered high confidence predictions based on the filtering criteria employed by deFuse [84].

Gene Fusions in Breast Cancer Patient RNA-Seq Data. Each RNA-Seq dataset from the CPTAC breast cancer patient cohort was, on average, comprised of 76M paired-end Illumina reads with length 50bp. Based on transcriptome mapping results, the average fragment length and standard deviation were 190.3bp and 65.47bp respectively. In total deFuse detected 206,255 fusions; on average, this amounts to 1,964 predictions per sample. However,

Table 5.3: General information on all 22 TCGA breast cancer patients with both tumor/normal WGS and tumor RNA-Seq data. **(A) Clinical Data.** The PAM50 mRMA cancer subtypes and AJCC stage for each patient. **(B) Data Source.** *Tissue Source* indicates the medical facility the sample and the relevant clinical data originates from; *Sequencing Center* indicates the location of actual sequencing: **WUSM** indicates Washington University School of Medicine, and **HMS** indicates Harvard Medical School. **(C) Number of Reads.** The BAM files corresponding to the majority of the samples contain paired-end reads of length 2x100bp (data from WUSM) or 2x51bp (data from HMS). There are only two exceptions: the solid tumor of patient **A09I** contains additional 206 million paired-end reads with respective lengths of 100bp and 44bp; the solid tumor of patient A0CM contains additional 579 million single end reads. These two inconsistent data sets are not used in our analysis. Note that all RNA-Seq datasets are from UNC (University of North Carolina Medical School), and on average include 76M paired-end reads of length 2x50bp.

Patient	Cancer Subtypes	AJCC Stage	Tissue Source	Sequencing Center	Number of WGS Paired-End Reads (Millions)		
					Solid Tumor	**Blood Normal**	**Solid Normal**
A09I	Basal-like	IIA	Indivumed	WUSM	687.13	584.10	
A0AV	Basal-like	IIIC	U of Pittsburgh	WUSM	954.38	558.24	
A0CE	Basal-like	IIA	Christiana Healthcare	WUSM	628.67	552.64	691.68
A0CM	Basal-like	IIA	Walter Reed	WUSM	784.96	540.58	
A0D0	Basal-like	IIA	Walter Reed	WUSM	788.15	516.83	
A0D1	Basal-like	IIB	Walter Reed	WUSM	1015.35	573.74	
A0D2	Basal-like	IIIA	Walter Reed	WUSM	689.42	646.66	
A0DG	Basal-like	I	U of Pittsburgh	WUSM	893.64		522.71
A0E0	HER2-enriched	IB	U of Pittsburgh	WUSM	686.41	521.02	793.65
A0EY	HER2-enriched	IIA	Walter Reed	WUSM	907.43	579.46	
A0HK	HER2-enriched	II	U of Pittsburgh	HMS	193.03	180.00	
A0J6	HER2-enriched	IIA	MSKCC	WUSM	592.48	661.74	
A0JJ	HER2-enriched	IIIA	MSKCC	HMS	214.75	208.31	
A0JL	HER2-enriched	IIIA	MSKCC	HMS	220.37	217.78	
A0JM	Luminal A	IIB	MSKCC	WUSM	1126.03	671.57	
A0TX	Luminal A	IIB	Mayo	WUSM	1018.72	642.28	
A0YG	Luminal A	IIA	Walter Reed	WUSM	872.51	514.68	
A12L	Luminal B	IIIA	ILSBio	WUSM	1031.04	650.09	
A12Q	Luminal B	IIIC	ILSBio	WUSM	1011.50	640.57	
A130	Luminal B	IIB	ILSBio	WUSM	813.37	654.17	
A18R	Luminal B	IIB	U of Pittsburgh	WUSM	1002.25		594.58
A18U	Luminal B	IIA	U of Pittsburgh	WUSM	906.82		605.45

many of these predictions had low deFuse scores, either due to low sequence similarity or limited read support, and thus were not good fusion candidates. Only 3,907 of these predictions (roughly 2% of all predictions) in total are considered to be high confidence calls by deFuse.

5.2.2 MicroSV Detection by MiStrVar

MicroSV predictions were based on three WGS datasets. The first is a simulation dataset based on the Venter genome developed with the goal of assessing sensitivity and precision of our methods with respect to available tools for SV discovery. These results are summarized in Table 5.4; The second dataset consists of WGS data from the HCC1143 cell line (both tumor and normal), which was used to assess our methods' accuracy on a homogeneous tumor sample. The third dataset is comprised of 22 TCGA/CPTAC breast cancer WGS data, which were used for full scale evaluation of our methods.

Table 5.4: Comparison of precision, recall, false discovery rate (FDR) and false negative rate (FNR) of MiStrVar against other SV discovery tools. All tools were run with default parameters and the calls for each microSV type (we only considered the calls made by each tool for that microSV) were called true or false based on the metrics provided by the tools (quality, identity or support, if they exist). The threshold values for each metric were chosen to maximize the F-score. Only inversions of length $\leq$400bp were considered in the calculations. If a tool does not provide precise breakpoints, breakpoints falling within a provided range are counted as true positives. Known insertion SNPs were filtered for all duplication results.

		5-100 bp				**101-400 bp**			
SV Type	**Tool**	**Precision**	**Recall**	**FDR**	**FNR**	**Precision**	**Recall**	**FDR**	**FNR**
Inversions	**MiStrVar**	91.20%	92.68%	8.80%	7.32%	93.10%	98.78%	6.90%	1.22%
	Breakdancer	66.67%	1.63%	33.33%	98.37%	59.00%	95.00%	41.35%	4.88%
	Delly	67.00%	1.63%	33.00%	98.37%	61.98%	91.46%	38.02%	8.54%
	Pindel	82.64%	81.30%	17.36%	18.70%	88.51%	93.90%	11.49%	6.10%
	SoftSV	0.00%	0.00%	100.00%	100.00%	93.75%	18.29%	6.25%	81.71%
All Duplications	**MiStrVar**	30.85%	53.91%	69.15%	46.09%	N/A	N/A	N/A	N/A
	ITDetector	13.54%	40.87%	86.46%	59.13%	N/A	N/A	N/A	N/A
	Pindel	5.00%	15.65%	95.00%	84.35%	N/A	N/A	N/A	N/A
	SoftSV	16.24%	16.52%	83.76%	83.48%	N/A	N/A	N/A	N/A
Tandem Duplications	**MiStrVar**	100.00%	86.67%	0.00%	13.33%	N/A	N/A	N/A	N/A
	ITDetector	3.17%	80.00%	96.83%	20.00%	N/A	N/A	N/A	N/A
	Pindel	0.00%	0.00%	100.00%	100.00%	N/A	N/A	N/A	N/A
	SoftSV	6.67%	46.67%	93.33%	53.33%	N/A	N/A	N/A	N/A

5.2.3 MicroSVs in the HCC1143 Breast Cancer Cell Line

Before running MiStrVar on the TCGA/CPTAC breast cancer samples, we applied it to the HCC1143 breast cancer cell line. We identified 116 microinversions and 197 microduplications on this sample. Among these, 11 inversions and 12 duplications have both high read coverage and low mapping multiplicity. We focus only on these microSVs for the remainder of the discussion.

Details on the 11 inversion candidates can be found in Table 5.5. All 11 inversions appear in both normal and matching tumor samples indicating that they are germline events. 10 of them occur in intronic regions while one occurs in a 3' UTR.

We experimentally validated these inversions using Sanger sequencing. The primers were constructed by using the inverted sequence flanked by 200-300 bp from the reference genome. Five of the predicted inversions show a clear sequence match between the amplicon (from Sanger sequencing) and predicted inversion, validating these inversion candidates. Four of the remaining inversions had amplicons with some nucleotides matching the reverse genomic strand and some matching the forward strand. This occurred in the amplicons from all four normal samples and two of the tumor samples. To resolve this discrepancy, the chromatogram corresponding to each amplicon was examined, first for the four normal samples, for which each of the inversion locations had either one or two peaks. In locations with two peaks, the bases always matched either the forward or reverse strand, exhibiting a classical case of heterozygous inversion that only occurs on one allele. For the final two inversion predictions, the amplicons for BOK and UBP1 corresponding to the tumor sample, only matched the forward genomic strand, which indicates no inversion at these locations. The amplicon corresponding to the normal sample of UBP1 contained many N bases in the sequence. Not enough information could be drawn from the chromatogram to conclusively say whether the amplicon supports an inversion.

We note here that all the high confidence microinversions, except for the one found in UBP1, have an associated multiple nucleotide polymorphism (MNP) entry in dbSNP. This includes the microinversion in BOK, which was not validated by Sanger sequencing.

In addition to MiStrVar we ran all the SV callers we tested on the HCC1143 cell line data. The parameters for all tools were identical to those used in the simulation. Out of these tools, only Pindel was able to identify any of the inversions. However, Pindel missed 2 of the 9 PCR validated inversion calls (in PFKP and OSBP2), out of 11 tested. The two calls made by MiStrVar that could not be validated were also called by Pindel, providing further evidence that MiStrVar improves Pindel with respect to both precision and recall.

The 12 duplication candidates are summarized in Table 5.5; all were exonic, i.e., fully or partially overlapping with exons. All of these duplications produced amplicons except for the one located in IRAK1BP1. Additionally, two amplicons from the normal sample (on genes ADAMTS19 and CIDEA) yielded a weak signal in the chromatogram so it was impossible

to determine if they support the call or not; furthermore, the corresponding amplicon from the tumor sample showed no evidence of the call. Three of the nine remaining calls, in FAM20C, GTPBP6 and KIAA1009, show a clear match in the tumor sample but not in normal, indicating they are true somatic calls. Two calls, in BAIAP2L2 and RBMXL3, have a clear match in both tumor and normal samples, indicating they are germline calls. The next three showed two peaks at the insertion site and immediately downstream. One of the two peaks support the reference and the other the inserted sequence and the shifted reference, indicating that these calls are heterozygous. This was observed in both normal and tumor samples for GPRIN2 and only in normal for PALM2-AKAP2 and PRSS48. The final amplicon for ADAMTS7 showed only reference sequence at the insertion site, indicating that there is no duplication.

As per the microinversions, we ran all other computational tools mentioned earlier in order to determine if they are able to predict the validated microduplications. None of the tools were able to predict any of the microduplications. (Note that ITDetector was never able to complete execution after more than a month of processing.)

MicroSVs in the Complete Set of TCGA-CPTAC Breast Cancer Samples

We applied MiStrVar and ProTIE to the complete set of matched tumor/normal samples from 22 TCGA breast cancer patients for which matching WGS, RNA-Seq and CPTAC Mass Spectrometry data were all available.Minimal filtering was used on the calls since few calls uniquely matched proteomic signatures. Note that we only focus on exonic microinversion and microduplication calls (either fully or partially overlapping with exons) for further analysis. Although only exonic calls were used for further analysis, the highest confidence calls within intronic and UTR regions, with respective support of > 40 and > 10 (identity $= 100\%$) were also collected .We also provide the highest confidence microduplications without proteomic support (support > 40, identity $= 100\%$) as well as somatic microduplications .

5.2.4 ProTIE Proteogenomics Analysis of CPTAC Breast Cancer Datasets

CPTAC has produced global proteome and phosphor-proteome data for 105 TCGA breast cancer samples using iTRAQ protein quantification method. Samples were selected from all four major breast cancer intrinsic subtypes (Luminal A, Luminal B, Basal-like/triple-negative, HER2-enriched) [64]. Each iTRAQ experiment included three TCGA samples and one common internal reference control sample. The internal reference is comprised of a mixture of 40 TCGA samples (out of the 105 breast cancer samples) with equal representation of the four breast cancer subtypes. Three of the TCGA samples were analyzed in duplicates for quality control purposes.

Our data analysis indicates that a two-dimensional reversed-phase liquid chromatographyâĂŞtandem mass spectrometric (2D-LC/MS/MS) sample comprises of about 0.87

Table 5.5: Sanger sequencing validation of top 11 microinversion and top 12 exonic microduplication (tandem or interspersed) candidates in the breast cancer cell line HCC1143. Entries marked "Yes" indicate a detected amplicon exactly matching the predicted microSV. "1 allele" indicates that two peaks were observed at each position in the chromatogram, only one matching the predicted microSV, and the other matching the reference, implying heterozygosity. For each detected inversion exactly matching an "multiple nucleotide polymorphism" and duplication exactly matching an "insertion" in dbSNP, we provide the dbSNP entry in the last column. As can be seen, all but two of these microSVs have been misclassified as a multiple nucleotide polymorphism or novel insertions in dbSNP. All microduplications are tandem, except for GTPBP6 which is interspersed. "RNA-seq support" denotes the number of reads support the structural variant. Since only tumor RNA-seq data was available, those SVs predicted in the normal sample are marked as "N/A". The gene RBMXL3 is not expressed in this cell line therefore no supporting reads can be expected. Note that all of the microinversions we detected (with minimum support) were intronic and thus had no matching RNA-Seq reads. The duplication in PALM2-AKAP2 was likely missed by Sanger Sequencing in tumor (marked with an asterisk). The breast cancer-related gene FAM20C is marked in green.

								WGS Support		Validated			
Type	**Chr.**	**Location**	**Len.**	**Pali.**	**Gene**	**Region**	**Identity**	**Tumor**	**Normal**	**Tumor**	**Normal**	**RNA-Seq Support**	**dbSNP ID**
Inv.	2	44545739	27	6	SLC3A1	3′UTR	100.00%	66	62	Yes	Yes	-	rs71416108
Inv.	3	170821851	26	3	TNIK	Intron	100.00%	96	76	Yes	Yes	-	rs781523247
Inv.	7	117357036	29	3	CTTNBP2	Intron	100.00%	76	81	Yes	Yes	-	rs386717124
Inv.	10	3173068	24	3	PFKP	Intron	98.82%	62	51	Yes	Yes	-	rs386740061
Inv.	19	56389843	32	2	NLRP4	Intron	98.03%	80	103	Yes	Yes	-	rs386811126
Inv.	19	38062904	29	4	ZNF571/540	Intron	99.33%	26	35	1 allele	1 allele	-	rs386809055
Inv.	22	31291523	23	2	OSBP2	Intron	100%	49	24	1 allele	1 allele	-	rs67147751
Inv.	1	68552108	18	6	GNG12-AS1	Intron	98.74%	37	43	Yes	1 allele	-	rs386632129
Inv.	9	28014540	29	3	LINGO2	Intron	100.00%	18	41	Yes	1 allele	-	rs386733960
Inv.	3	33449797	30	3	UBP1	Intron	100.00%	10	40	Inconclusive	Inconclusive	-	-
Inv.	2	242500549	12	4	BOK	Intron	98.60%	77	117	No	No	-	rs386657165
Dup.	X	229389	6	-	GTPBP6	Exon	100%	28	33	Yes	No	0	-
Dup.	7	286468	34	-	**FAM20C**	Exon	98.03%	12	22	Yes	No	0	rs774848096
Dup.	6	84884494	45	-	KIAA1009	Exon	98.60%	34	0	Yes	No	7	rs539790644
Dup.	22	38483155	9	-	BAIAP2L2	Exon	100.00%	48	33	Yes	Yes	0	rs142739979
Dup.	X	114425181	27	-	RBMXL3	Exon	98.74%	37	41	Yes	Yes	No Expression	rs782097222
Dup.	10	46999591	9	-	GPRIN2	Exon	100.00%	90	69	1 allele	1 allele	36	rs112620425
Dup.	9	112900341	6	-	PALM2-AKAP2	Exon	99.33%	16	46	No*	1 allele	5	rs150402481
Dup.	4	152201018	5	-	PRSS48	Exon	100.00%	0	47	No	1 allele	N/A	rs71901196
Dup.	5	128797315	6	-	ADAMTS19	Exon	98.60%	0	29	No	Inconclusive	N/A	rs142924298
Dup.	18	12254562	16	-	CIDEA	Exon	100.00%	0	24	No	Inconclusive	N/A	rs71369912
Dup.	6	79595167	5	-	IRAK1BP1	Exon	98.82%	79	65	Inconclusive	Inconclusive	0	rs146020132
Dup.	15	79058183	7	-	ADAMTS7	Exon	100.00%	40	33	No	No	0	rs781638345

million MS/MS spectra (per mixture). When we search them against Ensembl Human protein database, about 0.38 million MS/MS spectra in a mixture are matched to at least one peptide under 1% false discovery rate. These spectra lead to 59,387 proteins (42,840 known, 6,250 novel, 10,026 putative) with some peptides being covered by at least one spectra. The remaining 0.49 million spectra ($\approx 56\%$ of the whole set) do not match to any protein in the Ensembl database.

ProTIE obtains the intersection between these (0.49 million) unidentified spectra and the aforementioned set of fusions with missed cleaved polypeptides, to obtain 3,150,502 potential fusion peptides from 105 breast cancer patients [5] (see Figure 5.2). [6] ProTIE uses a similar workflow to identify potential microSV peptides; for this case 635,125 potential microSV peptides were obtained from 22 patients.

Based on the database search strategy mentioned in subsection 5.1.2, in each mixture, our first level analysis resulted in approximately 5,342 spectra (1% FDR) matching to fusion peptide sequences, and about 620 spectra matching to microSV peptide sequences. If a matched peptide is identical to a substring of any known protein in Ensembl database, the corresponding spectra is discarded so as to ensure that the peptide is novel. The remaining results thus consist of all mass spectra in a single mixture supporting novel peptides originating from high confidence sequence aberrations. For a specific mixture, we can extract all the genes and the corresponding patient(s) generating these translated aberrations based on deFuse and MiStrVar calls.

It has been argued in the literature that stringent class-specific peptide-level FDR estimates may be necessary for reporting novel peptides in proteogenomics studies [100]. In order to address this issue, for any search result provided from MS-GF+, we first cluster all peptide-spectra matches into known or novel categories based on their peptide sequences: a PSM is assigned to the known class if the peptide is a known peptide or the decoy sequence of a known peptide; otherwise it will be assigned to the novel class. We then recalibrate FDR for records in the novel class using original E-value from MS-GF+: a peptide p is assigned the best spectral E-value $E(p)$ it can get from any records in the novel class. Given a PSM M with E-value s, we collect all PSMs in the novel class whose E-value $\leq s$, and calculate the ratio of records containing decoy sequences as the new peptide-level FDR for M. In tables 5.6, 5.7, and 5.8, a checkmark in the last column (labeled Str FDR) indicates that the corresponding PSMs pass this more stringent class-specific peptide-level FDR under 1%.

[5] Each breakpoint is associated with six reading frames and thus can result in (one of) six distinct proteins, and each such potential protein can lead to multiple potential peptides according to the number of K/R in the sequence.

[6] Note that a reversed database was also appended here to control the false discovery rate.

ProTIE Inferred Fusion Peptides

Given the proteomics search results for a specific mixture, a peptide will be further evaluated only if the corresponding fusion is also observed in at least one patient within the mixture. Among the remaining 5,579 spectra, 3,185 match to peptides coming from immunoglobulin heavy and light chain fusions. These peptides are not considered any further since highly repeated regions shared between those genes can lead to false positives in both fusion detection and proteomics search stages [19, 13]. Among the peptides remaining, we also discard those associated with a fusion for which no breakpoint crossing peptide is observed (This is due to the difficulty of determining whether such a peptide is a result of a fusion or because of a reading frame shift). At the end of these filtering steps ProTIE returns 807 spectra matching to 169 potential fusion peptides.

Among fusions related to these potential fusion peptides, we summarize special events with either high confidence RNA-Seq level evidence or proteomics support in Table 5.6. The first part of Table 5.6 shows events with better fusion quality based on reports of deFuse (deFuse score ≥ 0.1, cDNA percent identity < 0.1, EST and EST island percent identity < 0.3, no evidence detected for read through). Since 3 of these predicted fusions are between paralogs, specificially CRIP1 and CRIP2, IFITM2 and IFITM3, SRGAP2 and SRGAP2B, they are ignored. Among the remaining fusions, two stand out with respect to peptide-spectrum matching quality, respectively observed in patients A08G and A15A. The PSMs supporting these two fusions generated by pFind Studio [72, 154] are shown in Figures 5.4, 5.5, 5.6.

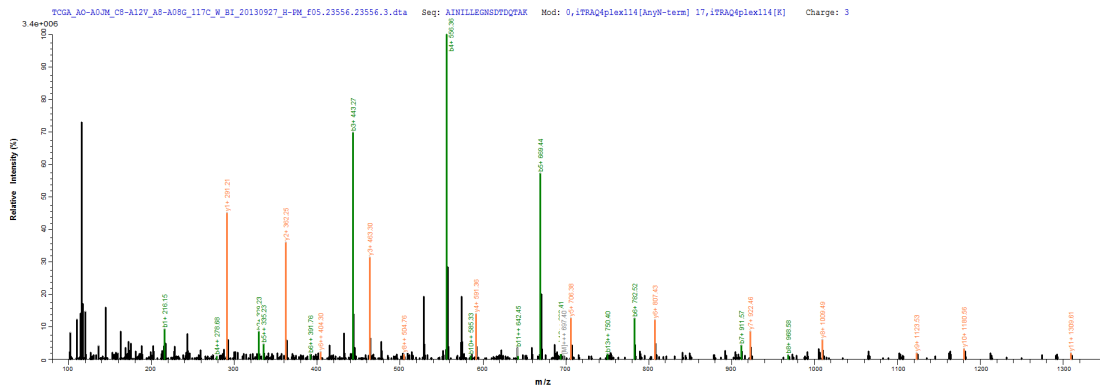


Figure 5.4: A PSM supporting a fusion (See table 5.6 in main text) between genes TEAD1 and UBAP2 in patient A08G (Luminal B, Stage IIA). The fusion breakpoint is located between T and D in the peptide AINILLEGNSDTDQTAK.

We also provide a list of fusions with multiple translation peptides in the second part of Table 5.6. More specifically, four of these fusions have matching peptides located on both at the breakpoint and further downstream. Note that although we only detect a single peptide

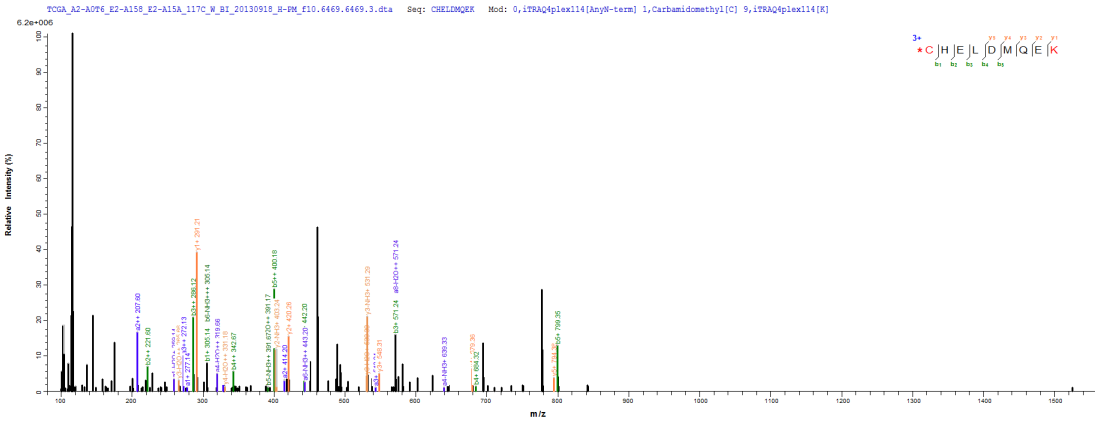


Figure 5.5: A PSM supporting a fusion between genes HOOK3 and CTA-392C11.1 in patient A15A (Luminal B, Stage IIIC). The peptide crosses the fusion breakpoint predicted from RNA-Seq data at amino acid M.

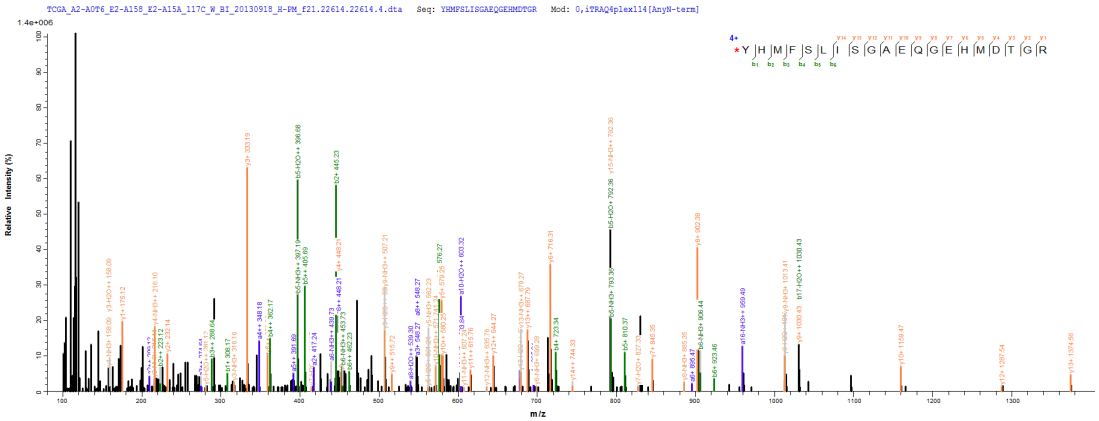


Figure 5.6: Another PSM supporting the fusion in figure 5.5 between genes HOOK3 and CTA-392C11.1 in the patient A15A (Luminal B, Stage IIIC). The peptide is located downstream of the breakpoint.

for some additional fusions, the peptide may be supported by multiple spectra as can be seen in Table 5.7.

ProTIE Inferred MicroSV Peptides

As per ProTIE's translated fusion peptide inference approach, for each mixture, we only consider previously unknown peptides that can be unique byproducts of microSVs detected in at least one patient within the mixture. To ensure that these peptides support microSV (duplication or inversion) calls and not SNVs/SNPs, we only consider potential peptides from an interspersed duplication or inversion with a minimum of two amino acids on each

side of at least one of the two breakpoints associated with that microSV; for tandem duplications we ensure that at least two amino acids are present in the peptide from both sides of the single breakpoint.

Proteomics search of these peptides on 22 patients resulted in 115 spectra potentially supporting microSVs. These spectra support a total of 75 peptides, due to the fact that some of the peptides are supported by more than one spectra. Of these 75 peptides, 7 support microduplications and 68 support microinversions. Incorporating the RNA-Seq results,we obtain 4 microSV calls with support on all omics levels. The resulting peptides with the highest quality spectra support are summarized in Table 5.8. Here the number of spectra supporting these peptides is indicated in the "Spectra" column. Similarly, column "Breakpoint Support" indicates the number and type of the breakpoints supported by spectra for each peptide.

5.3 Discussion

5.3.1 Genomic MicroSVs Detected with MiStrVar

Our simulations show that MiStrVar effectively and accurately identifies all microSVs, specifically, insertions, tandem and interspersed duplications in WGS datasets. In particular, MiStrVar has high sensitivity, as well as high precision - especially for inversions, For duplications, even though its precision may not look as impressive, MiStrVar still outperforms all available alternatives. In addition, the precision values for duplications are likely to have been underestimated, since many of calls labelled as "false positives" could, in fact, be true germline differences between the Venter genome and the reference genome. On a very high coverage dataset (120x) from the Venter genome, with no simulated microSVs, duplications detected by MiStrVar have a large overlap with those it detects in the simulation dataset. Elimination of these calls from the simulation dataset increases MiStrVar's precision to 71% without any additional filtering.

MiStrVar is also very accurate in identifying the exact breakpoint loci of the microSVs. This is particularly important for our proteogenomics analysis where we only consider exact peptide matches. If a breakpoint were off even by only one nucleotide there is a high likelihood the predicted peptide would not match. With the exception of Pindel for inversions, which correctly identified 10% fewer exact breakpoints, no tool was even close to correctly identifying as many single-nucleotide resolution microSV breakpoints as MiStrVar. For inversions, the calls where MiStrVar can not identify the exact breakpoints are often due to the presence of palindromic sequences, resulting in co-optimal breakpoint predictions. More importantly, these cases yield identical peptides and therefore do not affect further analysis results. For duplications, the errors are usually observed in cases where the insertion is into a low complexity region. Again, in many of these cases the resulting peptides would be identical (e.g. consider a duplication that occurs in a polynucleotide tract). Furthermore,

Table 5.6: The list of selected (interesting) fusion events with translated peptides. A check mark in the column BP (BreakPoint) indicates that the peptide crosses the fusion breakpoint, and a check mark in the last column indicates that the peptide satisfies our more stringent FDR criterion. (a) *High confidence fusions*: fusions with high "deFuse Score" are colored purple (these satisfy stringent RNA-Seq level filtration conditions). (b) *Fusions with multiple supporting peptides:* fusion events associated with multiple novel peptides with proteomic support are colored cyan. (c) Among all fusions, one involves *a cancer gene*, TEAD1, and is colored green. (d) Only one fusion peptide is *supported by multiple spectra*: it is associated with the fusion detected in patient A18U, and is colored yellow. Note that peptides with star sign (*) are Single Amino Acid Variants (SAAVs) according to validated peptides in Ensembl GRCh38 protein database.

Patient	Cinical Information	Gene 1	Gene 2	deFuse Score	Breakpoint Location	Peptide Sequence	# of Spectra	BP	Str FDR
Fusions satisfying RNA-Seq level filtration conditions									
A08G	Luminal B, IIA	UBAP2	**TEAD1**	**0.94**	coding, coding	AINILLEGNSDTDQTAK	1	✓	✓
A0AM	Luminal B, IIA	C17orf85	ZMYND15	**0.92**	coding, downstream	AQTPGDQETR	1	✓	
A12E	Luminal A, IIB	C20orf111	FITM2	**0.93**	utr5p, coding	NVLNVVNR	1	✓	
A142	Basal-like, IIB	ACTG1	ACTB	**0.53**	coding, coding	QDATLALGLVTNWDDMEK	1	✓	✓
A159	Basal-like, IIA	ACTL7B	KLF9	**0.54**	coding, utr3p	EAQLPLEALGEAIQLCFLSFLSVR	1	✓	
A15A	Luminal B, IIIC	HOOK3	CTA-392C11.1	**0.43**	coding, intron	CHELDMQEK	1	✓	✓
						YHMFSLISGAEQGEHMDTGR	2		✓
A18U	Luminal B, IIIA	ZNF354A	RP11-383H13.1	**0.48**	coding, intron	DGSGVSSLGVTPESR	2	✓	✓
Additional Fusions with Multiple Supporting Peptides									
A04A	Luminal A, IIIA	ACTG1	ACTB	0.03	coding, coding	QKEALFQPSFLGMESCGIHETTFNSIMK	30	✓	✓
						KEALFQPSFLGMESCGIHETTFNSIMK	1	✓	✓
A06N	Luminal B, IIIB	KRT19	CTD-2165H16.1	0.01	coding, pseudogene	DNPGVLKPGMVVTFAPVNVTTEVK	1	✓	
						NPGVLKPGMVVTFAPVNVTTEVK	13	✓	✓
A0AS	Luminal B, IIIA	ACTG1	ACTG1P2	0.39	coding, pseudogene	(*)DLYTNTVLSGGTTMYPGIADR	5	✓	✓
						(*)LYTNTVLSGGTTMYPGIADR	6	✓	✓
A0AS	Luminal B, IIIA	ACTB	KDM4C	0.01	coding, intron	(*)FCCPEALFQPSFLGMESCGIHETTFNSIMK	1	✓	
						(*)CCPEALFQPSFLGMESCGIHETTFNSIMK	6	✓	
A0D1	HER2-enriched, IIA	RPL8	CTD-2165H16.1	0.01	coding, pseudogene	EAVPIVAAGVGEFEAGISK	1	✓	
						AFVPISGWNGNNMLEPSANMPWFK	2		✓
						KIGYNPDTVAFVPISGWNGNNMLEPSANMPW	1		✓
						KIGYNPDTVAFVPISGWNGNNMLEPSANMPWF	3		✓
						IGYNPDTVAFVPISGWNGNNMLEPSANMPWFK	16		✓
						KIGYNPDTVAFVPISGWNGNNMLEPSANMPWFK	2		✓
A0TT	Luminal B, III	ACTB	ACTG1	0.47	coding, coding	AWSPEALFQPSFLGMESCGIHETTFNSIMK	13	✓	✓
						WSPEALFQPSFLGMESCGIHETTFNSIMK	1	✓	✓
A12Z	HER2-enriched, II	ACTB	FNIP1	0.21	coding, intron	MTQIMFETFNTPAVYMAI	2	✓	✓
						MTQIMFETFNTPAVYMAIQ	1	✓	✓
A158	Basal-like, IIA	EEF1A1P7	EEF1A1P29	0.03	pseudogene, pseudogene	KIGYNPNTVAFVPISGWNGDNMLEPSANMPWFK	1	✓	✓
						DGNASGTILLEALDCILPPTRPTDK	5		✓

Table 5.7: Additional list of selected (interesting) fusion events with translated peptides. A check mark in the BP (BasePair) column indicates that the peptide crosses the fusion breakpoint, and a check mark in the last column indicates that the peptide satisfies our more stringent FDR criterion. (a) *Fusions with multiple supporting spectra:* in addition to fusions in Table 5.6, other fusions have multiple supporting spectra - although all such spectra are associated with the same breakpoint-crossing peptide. These fusions are colored yellow. (b) *Fusions involving cancer genes:* fusions involving cancer-specific genes are colored green. Note that the peptide with a star sign (*) is a Single Amino Acid Variant (SAAV) according to validated peptides in Ensembl GRCh38 protein database.

Patient	Cinical Information	Gene 1	Gene 2	deFuse Score	Breakpoint Location	Peptide Sequence	# of Spectra	BP	Str FDR
Additional Fusions with Multiple Supporting Spectra									
A06Z	Luminal B, IIB	RAB15	TMEM98	0.01	coding, utr5p	QIWDTAGQENR	2	✓	
A0C1	Luminal A, IIIA	RPL14	FAM155A	0.02	coding, coding	ASAAAAAAAAK	2	✓	✓
A0D2	Basal-like, IIB	ACTG1	ACTB	0.52	coding, coding	HHGIVTNWDDMEK	4	✓	
A0E0	Basal-like, IIIC	PEA15	CPEB2	0.05	coding, coding	YPGTLLQDLTNNITLEDLEQLK	2	✓	✓
A0EX	Luminal A, IIB	RAB6B	CFL1	0.02	utr3p, coding	EAGVAVSDGVIK	3	✓	
A0TR	Luminal A, II	ZNF587	TMEM163	0.12	utr3p, intron	QSETLSQNKK	2	✓	
A12D	HER2-enriched, IIA	**RPL19**	**CALR**	0.04	coding, coding	PAGQGVFPASSPGMDGEWEPPVIQNPEYK	5	✓	✓
A12D	HER2-enriched, IIA	SCGB2A2	EEF1A1P5	0.05	coding, pseudogene	ATAFIDQMASSGGLARIYVSNDDNATTNAIDELK	2	✓	
A12D	HER2-enriched, IIA	EIF4A1	ABL2	0.16	utr3p, intron	SLNKCHFLR	3	✓	
A12U	Luminal B, IB	NME1	RP11-111A21.1	0.01	coding, downstream	SVMLGETNPADSKPGTIR	2	✓	✓
A12W	Luminal B, IIIB	CTNNA3	CEP120	0.007	intron, intron	LALDIEIATYKT	2	✓	
A13F	Luminal B, IIIA	RPL14	S100A16	0.17	coding, utr3p	SAAAAAAAAK	2	✓	✓
A142	Basal-like, IIB	HSP90AB1	AC096579.7	0.01	coding, ncRNA	FEINPDHPIVETLR	4	✓	✓
A150	Basal-like, IIA	HSPA8	RP11-537H15.3	0.03	coding, intron	(*)HVAMNPTNTVFDAK	2	✓	✓
A159	Basal-like, IIA	DLG4	**VIM**	0.36	intron, coding	SYVTTSTRR	2	✓	
A15A	Luminal B, IIIC	WASH4P	ABC7-42389800N19.1	0.07	coding, pseudogene	PKSGSGGEGVMEPPR	2	✓	
A18Q	Basal-like, IIB	MGP	EEF1A1P5	0.01	coding, pseudogene	FFFFPQSHLVTFAPVNVTTEVK	5	✓	✓
A18U	Luminal B, IIIA	ZNF354A	RP11-383H13.1	0.47	coding, intron	DGSGVSSLGVTPESR	2	✓	✓
A1AQ	Basal-like, II	**CDKN2A**	LINC00486	0.01	coding, intron	GGGGGGGGCCPR	2	✓	
Additional Cancer-related Genes in Fusions									
A0BZ	Luminal B, IIIA	**MDM2**	ZC4H2	0.69	utr3p, downstream	ISFFLEVLQALFGVDNTSATTK	1	✓	
A0C1	Luminal A, IIIA	USP42	**CD44**	0.19	intron, utr3p	YEKENWSGFFFFFLK	1	✓	
A0EQ	HER2-enriched, II	**ANKRD30A**	BLOC1S6/NEDD4	0.85 0.76	coding, utr3p coding, intron	ISGKLEELEK	1	✓	✓
A09I	Luminal B, IIA	YARS/ZNFX1	**GRB7**	0.08 0.02	intron/utr3p upstream/intron	GQEFKTSLTNMAK	1	✓	
A09I	Luminal B, IIA	**ERBB2**	NME2P1	0.01	coding, pseudogene	IQHYIDLK	1	✓	

Table 5.8: The list of genes containing microSVs with high confidence mass spectra support based on joint analysis of all 22 TCGA breast cancer patients with both tumor/normal WGS and tumor RNA-Seq data. For inversions, associated peptides always span one or two breakpoints (indicated as 1/2 or 2/2), or the inverted sequence between the breakpoints (indicated as "Between"). For duplications (which happen to be all tandem), the associated peptide always spans the single breakpoint and the entire inserted sequence (1/1). Breakpoints in the peptide sequences are marked with "|". Calls marked as "Low" in the RNA-seq column are those from genes with low sequence coverage; similarly calls marked as "N/A" indicate the lack of RNA-seq data for this sample. Note that the microduplication in HSPBP1 is annotated as an insertion, and the microinversion in PLIN4 is annotated as two independent SNPs in dbSNP. Genes colored in green are known to be cancer related, and records colored in yellow have peptides with multiple supporting spectra.

Patient	Cancer Subtype	AJCC Stage	WGS Source	Gene	SV Length	SV Type	RNA-Seq	Breakpoint Support	Spectra	Peptide	dbSNP ID	Str FDR
A0DG	Luminal A	IIA	BOTH	FAM134A	6	DUP	✓	1/1	1	QALDS|EE|EEEEEDVAAK		✓
A0JM	Luminal B	IIB	BOTH	HSPBP1	9	DUP	Low	1/1	2	LPLALPPASQGCSSGGGGG|GG|GGSSAGGSGNSRPPR	rs3040014	✓
A18U	Luminal B	IIIA	BOTH	HSPBP1	9	DUP	Low	1/1	1	LPLALPPASQGCSSGGGGG|GG|GGSSAGGSGNSRPPR		✓
A18R	HER2-enriched	IB	BOTH	NUPL2	12	DUP	✓	1/1	1	QQP|RQQP|QQQPSGNNR	rs200880793	
A12Q	Luminal B	IIIC	BOTH	**RPL14**	9	DUP	✓	1/1	4	GT|AAA|AAAAAAAAAAK	rs369485042	✓
A0DG	Basal-like	I	BOTH	**RPL14**	9	DUP	✓	1/1	3	GT|AAA|AAAAAAAAAAK		✓
A0YG	Luminal A	IIA	BOTH	**RPL14**	15	DUP	✓	1/1	3	GT|AAAAA|AAAAAAAAAAK	rs369485042	✓
A0JM	Luminal A	IIB	BOTH	**RPL14**	15	DUP	✓	1/1	6	GT|AAAAA|AAAAAAAAAAK		✓
A0CE	Basal-like	IIA	NORMAL	**RPL14**	15	DUP	✓	1/1	9	GT|AAAAA|AAAAAAAAAAK		✓
A18R	Luminal B	IIB	NORMAL	**RPL14**	15	DUP	✓	1/1	4	GT|AAAAA|AAAAAAAAAAK		✓
A18U	Luminal B	IIA	BOTH	**RPL14**	27	DUP	✓	1/1	3	GT|AAAAAAAAA|AAAAAAAAAAK		✓
A0D2	Basal-like	IIB	NORMAL	PLIN4	6	INV	N/A	2/2	2	DTVCSGVT|SA|MNVAK	rs12327614, rs56366613	✓
A0AV	Basal-like	IIC	BOTH	PLIN4	6	INV	✓	2/2	4	DTVCSGVT|SA|MNVAK	rs75031432, rs79662071	✓ ✓
A0YG	Luminal A	IIA	NORMAL	**CHD5**	939	INV	N/A	1/2	1	KQVNYNDASQEDQ|GSER		✓
A0J6	HER2-enriched	IIA	TUMOR	C4orf21	73	INV		Between	1	KMTYVVNR		✓
A0EY	Luminal B	IIB	NORMAL	**PTPN4**	794	INV	N/A	Between	1	FINNYIIK		
A0J6	Basal-like	IIA	TUMOR	ZNF415	497	INV	Low	Between	1	QRAEILEK		
A18R	HER2-enriched	IB	TUMOR	ACSM2A	528	INV	Low	Between	1	VSQGNIK		
A0CM	Basal-like	IIA	TUMOR	CC2D2A	886	INV	Low	Between	1	MEHMIQASVT		
A0CM	Basal-like	IIA	TUMOR	ZNF257	732	INV	Low	Between	1	FSHLIAGK		
A0CM	Basal-like	IIA	TUMOR	**RBBP8**	405	INV		Between	1	VEGQGGGK		

even in the worst case, MiStrVar predictions are within 30bp from the real breakpoints, still much better than the available alternatives. It should also be noted here that unlike other tools, MiStrVar provides not only the duplication breakpoint coordinates but also the precise coordinates of the "source" sequence (i.e. the region of the genome that is duplicated). Through this feature it becomes easier for the user to interpret interspersed as well as tandem duplications.

5.3.2 Translated Aberrations Detected with ProTIE

The use of a proteogenomic approach, as described in this study, enables two novel capabilities that are highly relevant to cancer biology and precision medicine. 1) The ability to hone in on potential clinically actionable mutations that are expressed at the protein level. The vast majority of clinical cancer testing focuses only on DNA-level mutations. A gene mutation-drug association is predicated on the assumption that a mutation will translate to the protein level, however, this is often not the case, as genes that contain a mutation may not be expressed in RNA. Moving further to the transcriptome the same paradigm exists, i.e., RNA expression does not always directly translate to protein expression, secondary to a variety of translational control mechanisms. Thus, having protein level evidence to confirm genomic aberrations provides assurance of the functional presence of a mutation. This has wide ranging implications for clinical cancer genomic testing, as well as the development of companion diagnostics for cancer targeted therapies. 2) The ability to observe the presence of protein spectra from fusion transcripts that are predicted to be âĂIJout-of-frameâĂİ. The vast majority of fusion annotation pipelines filter out fusions that are not in-frame secondary to a widely-held reasoning that these protein products would be misfolded and degraded or subject to non-sense mediated decay. Surprisingly, in this study, high quality spectra were observed from out-of-frame fusion spectra. While additional studies will need to be performed, these data suggest these out-of-frame fusion products are stable enough and at a relative abundance to be detected by Mass Spectrometry. Whether these products are stable by chance or confer a gain-of-function capability is yet to be seen, but these data at minimum suggest that out-of-frame fusions should not be eliminated from consideration (as is commonly done), when searching for oncogenic candidates.

Translated Gene Fusions

To better understand the properties of genes with translation evidence for fusions, we analyzed these genes through Ingenuity Pathway Analysis (`https://www.ingenuity.com`). Note that we used all fusion genes detected by deFuse as the background genes in the analysis. The top 3 categories for gene function enrichment are: Cancer (137 genes), Organismal Injury and Abnormalities (150 genes), and Respiratory Disease (39 genes). All 3 sets of genes come with adjusted p-value around 0.0035 (via Benjamini-Hochberg procedure).

Given that fusions are a somatic cancer-specific event, enrichment of cancer related genes provides a validation of our approach.

Many of the fused genes with detected novel peptides (each typically observed in a single patient) are associated with breast cancer. A selection of these fusions are listed in Table 5.6 and 5.7 where cancer-related genes are highlighed. Among them, a fusion of the Ubiquitin Associated Protein (UBAP2) and the transcriptional enhancing factor (TEAD1) is found in the patient A08G and meets our stringent FDR criterion. This fusion retains the DNA binding domain of TEAD1. Interestingly, high TEAD1 expression is associated with poor survival and this fusion may cause hyper-activation of TEAD1 in this patient [17, 1]. Note that the same fusion has also been detected with high confidence in TCGA Fusion gene Data Portal [166].

The remaining fusions associated with highlighted genes in Table 5.7 appear to be novel as they do not appear in the TCGA fusion database. Some of these fusions involve tumor suppressor genes. For example, even though the fusion detected in patient A0BZ does not meet our more stringent FDR criterion, it is interesting that it involves MDM2, a key regulator of the TP53 tumor suppressor pathway [95]. (TP53 is mutated in a large proportion of triple-negative breast cancers.) Another fusion that does not meet our more stringent FDR criteria but still is noteworthy is in patient A1AQ and involves CDKN2A gene, a tumor suppressor that inhibits the cell cycle and is deleted in many cancer samples [87]. The fact that it is fused to a long noncoding RNA, may be a novel mechanism to inactivate CDKN2A, as an alternative to deletion.

In addition to fused tumor suppressors, we also detected peptide evidence for fused oncogenes. The discovered fused oncogenes are: ANKRD30A, also known as NY-BR-1, a breast differentiation antigen observed in many breast cancer cells [8]; GRB7, a breast cancer driver gene which participates in Development ERBB-family signaling pathway [40, 120]; ERBB2, a well known breast cancer oncogene and biomarker [167] as well as the coexpressed gene Ribosomal protein L19 (RPL19); CALR, a gene highly expressed in approximately 5% of breast cancer cells and associated with metastasis [81]; and finally VIM, a protein involved in the epithelial to mesenchymal transition which drives metastasis [88]. The fusions involving ANKRD30A, RPL19 and CALR meet our stringent FDR criteria, while the others do not. In a number of cases, we can not pinpoint its fusion partners based on RNA-Seq data alone. The proteogenomics results help to increase our confidence of these fusions, and reduce the number of fusion partner candidates in the corresponding patients. The ERBB2 fusion is particularly interesting since ERBB2 is amplified in 15% of breast cancers and targeted with a variety of FDA approved drugs, making it a possible target for clinical analysis.

In the final list of 295 candidate fusions, 107 of the involved genes are also reported to be involved in a fusion according to TCGA Fusion gene Data Portal [7]. 58 of these genes have records in breast cancer (BRCA), and among them 19 genes are reported in the breast cancer database alone.

Among the ten cancer-related fusion genes in Table 5.6 and 5.7, nine are also found in TCGA Fusion gene Data Portal, with the exception of the ANKRD30A fusion. Seven of them (excluding VIM and CALR), are involved in fusions specifically in breast cancer patients. As mentioned earlier, UBAP2 is fused with TEAD1 in patient A08G, which matches the Fusion gene Data Portal entry exactly. The remaining six of these genes have different fusion partners in different patients.

Translated MicroSVs

Most of the microinversions with proteomics support are in the 400bp to 1kb length range. Microinversions shorter than 100bp are much less common in exonic regions. However in intronic and UTR regions, microinversions with the best genomic support (in terms of both read coverage and sequence similarity - after the inversion is accounted) are predominately of length less than 100bp; We also observed that shorter microinversions tend to be germline events while longer events tend to be somatic.

All of the microduplication calls with proteomic support (all of these -with the exception of the one in NUPL2- satisfy our more stringent FDR criterion) were predicted to be germline events. Indeed nearly all of these events have corresponding dbSNP entries. The call in FAM134A appears to be a novel germline event. The longest duplication in RPL14 also appears to be novel (rs369485042 includes variants with up to 5 alanines). Deletions, translocations and allele loss at the genomic loci containing this gene has been observed in variety of cancers [130], including breast cancer [18]. This may be the case within patients AOCE, A18R (deletion) and A0JM (LOH). The unusually long case in patient A18U may lead to protein instability, causing the same phenotype as a deletion. Polyalanine tract lengths have been shown to be associated with cancer risk in other genes, such as androgen receptor in prostate cancer [136].

Since we observed relatively few translated microduplications, it is unlikely that this type of microSV plays a major role in breast cancer through translation to aberrant proteins. However we predicted many high confidence microduplications in exonic regions, some with RNA-Seq support, in addition to many in UTRs and introns.It is possible that such exonic duplications lead to truncated or rapidly degraded proteins and the duplications in UTRs and intronic regions may affect gene expression and splicing.

[7]Note that results in this database are based on 10431 calls from 2961 TCGA patients, which contains much broader scope than 105 breast cancer patients selected by CPTAC.

From our list of high confidence microSV calls (Table 5.8), four were found in genes known to be related to cancer (CHD5, RPL14, PTPN4 and RBBP8) and one in drug metabolism (CYP4F11). Among them, CHD5 is a particularly well studied tumor suppressor in neuroblastomas. It is also a known tumor suppressor in breast cancer [160], as well as colon, lung, ovary and prostate cancers [65]. The protein it codes, Chromodomain Helicase DNA binding protein 5, has functions in chromatin remodeling and gene transcription. CHD5 is frequently deleted in breast cancer and in one case a mutation resulted in a truncated, non-functional protein [160]. The microinversion we detected produces a stop codon shortly after the breakpoint which may also lead to the production of a truncated protein. Note that this microinversion satisfies our more stringent FDR criterion. Another interesting example, RBBP8 is a tumor suppressor specifically related to breast cancer. We have observed through inspecting geneMania [156] that RBBP8 is associated with the recombinational repair pathway ($p < 1.27 \times 10^{-9}$) (Figure 5.7). RBBP8 is also known to modulate the important tumor suppressor BRCA1 [135] and act as a tumor suppressor itself through binding with the MRE11-RAD50-NBS1 (MRN) complex [168] or replication protein A (RPA) [121].[8]

Our analysis resulted in 4 microSV calls with support on all omics levels. This includes 3 microduplications (within genes FAM134A, NUPL2 and RPL14) and 1 microinversions (within PLIN4). The microduplications in FAM134A and RPL14 (that with 27bp) appear to be novel events. Additionally, there are several events with both genomic and proteomic support, which possibly lack RNA-Seq support due to low expression of the associated gene or the lack of RNA-Seq data for the sample.

5.4 Conclusion

Integration of genomic, transcriptomic, and proteomic data provides a comprehensive view of the patient's molecular profile. TCGA/CPTAC now offers matching genomic, transcriptomic and proteomic data across several cancer types, with a focus on the impact of Single Amino Acid Variants (SAAVs) and SNVs on protein abundances. In order to complement TCGA/CPTAC study and better establish the relationship between genomic, transcriptomic and proteomic aberrations and the cancer phenotype, we introduce MiStrVar, the first tool to capture multiple types of microSVs in WGS datasets. MiStrVar, and deFuse, a fusion detection tool we developed earlier, form key components of ProTIE, a computational framework we introduce here to automatically and jointly identify translated fusions and microSVs in matching omics datasets. Concurrently, ProTIE also incorporates RNA-

[8] Binding of MRN and RPA occur through a domain at the N-terminus of the RBBP8 protein, which overlaps with the predicted microinversion. We hypothesize that the microinversion in this gene leads to the production of an aberrant peptide which is unable to bind to MRN or RPA, disrupting double stranded break repair and contributing to the cancer.

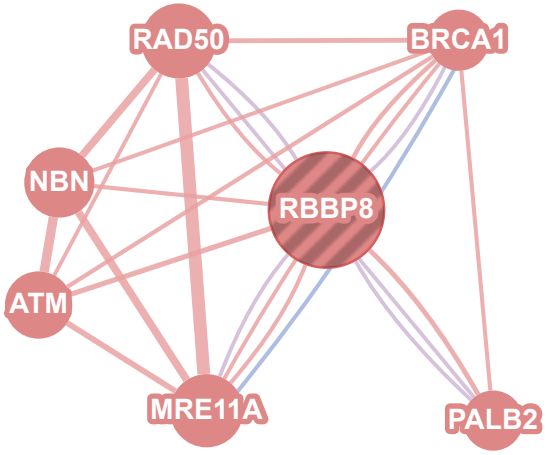


Figure 5.7: Functional analysis graph from GeneMania. Red lines indicate direct physical interaction, purple lines indicate co-expression and blue lines co-localisation. The thickness of the line represents the combined weights of the interaction across all analysed networks of that type. The diameter of the circles is inversely proportional to the rank of the gene in a list sorted by functional relatedness to the striped gene. This graph contains all genes interacting with RBBP8 in the recombinational repair pathway ($p < 1.27 \times 10^{-9}$). RBBP8 is closely associated with BRCA1, an important tumor suppressor gene in breast cancer.

Seq evidence to validate expressed microSVs. Based on both simulation and cell line data, we demonstrate that MiStrVar significantly outperforms available tools for SV detection. Our results on the TCGA/CPTAC breast cancer data sets also suggest the possibility of automatic calibration for some entries in dbSNP, which we believe are misannotated. It is interesting to note that the majority of the translated microSVs and fusions we observed in the breast cancer samples were private events; this prompts a larger and more detailed integrated study of all three omics data types through the use of ProTIE for a comprehensive molecular profiling of breast cancer subtypes.

Chapter 6

Conclusion

High-throughput sequencing (HTS) and mass spectrometry (MS) technologies allow us to reconstruct genome, transcriptome, and proteome for a specific patient and estimate the corresponding expressions of genes and proteins more accurately. Information of analyzed patients enrich the contents of genome databases, and help researchers to construct potential associations between cancers and gene expression profiles or specific aberrations. These efforts accelerate cancer research by profiling transcriptome and proteome of a new patient with unprecedented speed and accuracy.

Most computational methods designed for these high-throughput omics data start with either de novo construct expressed transcripts and proteins, or compare the datasets to existing gene annotations. While the methods based on gene annotations usually provide more accurate results than de novo strategies regarding profiling known transcripts and proteins, these methods do not allow us to capture unannotated events. These unannotated events, while relatively low-abundant, indicate potential interruption of normal functional units, and are essential for better understanding of cancer progression and development. Therefore, it will be of great importance to design computational methods which incorporate existing gene annotation information and still provide flexibility to detect unannotated events.

In this thesis, we present algorithms which improve accuracies in detecting both annotated and unannotated splicing events using high-throughput omics datasets. These algorithms start with RNA-Seq mappings, which provide base-resolution in detecting splicing events, and apply ILP to improve isoform identification and quantification results allowing novel events. We also propose a proteogenomics strategy to validate non-canonical splicing events, especially gene fusions, which are difficult to predict solely based on RNA-Seq mapping results.

We first provide an overview of RNA-Seq technology and the main computational challenges of profiling transcriptome using RNA-Seq datasets. Transcriptome profiling starts with identifying and quantifying expressed transcripts, and we first propose CLIIQ to simultaneously detect expressed transcripts and estimate their expressed levels from the map-

ping results allowing novel splicing sites in a gene. CLIIQ can integratively collect mapping results from multiple samples to improve accuracy in recurrent and low-abundant events. We have shown that CLIIQ provides accurate results in both single and multiple-sample mode in comparison to other popular tools.

We then introduce ORMAN to resolve the mapping ambiguity problem for RNA-Seq reads. To overcome the over-counting issue, which might lead to inaccurate quantification results, we design ORMAN, which first selects expressed regions using a set-cover strategy, and then assigns each read to a single location by minimizing local coverage variation. ORMAN allows novel events by assigning different weights to regions in the first stage. Since ORMAN provides a SAM/BAM file with only singly-mapped reads, it can serve as a general tool to boost performance for read-counting based pipeline.

We consider the problem of detecting non-canonical splicing events, especially gene fusions, using proteogenomics approach. We introduce ProTIE, which integratively takes RNA-Seq and mass spectrometry datasets, and searches for proteome-level signatures corresponding to potential fusions and other aberrations. The detected aberrations have supports in multiple omics datasets, and have higher chances to affect cancer phenotypes. We apply ProTIE in TCGA/CPTAC datasets and discover private events which are not reported before.

6.1 Future Directions

Comprehensive study of RNA-Seq coverage among different sequencing protocols will be essential for obtaining better transcriptome profiling results for specific patients. This will definitely help to improve performances of isoform identification, quantification, and testing for differentially expressed genes, and even lead to better identification of cellular subpopulations based on single-cell RNA sequencing (scRNA-seq) datasets.

The improvement of mass spectrometry-based technology, such as Tandem mass tag, allows us to identify and quantify expressed peptides more accurately. Development of efficient computational methods which incorporate properties from these new protocols can potentially provide estimation of expressions of novel proteins relevant to aberrant transcripts. When we have omics datasets from multiple patients with different cancer subtypes, such information can help us to discover prognosis signatures for identifying cancer subtypes.

Bibliography

[1] The hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37), September 2012.

[2] Rehan Akbani, Patrick Kwok Shing Ng, Henrica M. J. Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, Jun Li, Shiyun Ling, Elena G. Seviour, Prahlad T. Ram, John D. Minna, Lixia Diao, Pan Tong, John V. Heymach, Steven M. Hill, Frank Dondelinger, Nicolas Städler, Lauren A. Byers, Funda Meric-Bernstam, John N. Weinstein, Bradley M. Broom, Roeland G. W. Verhaak, Han Liang, Sach Mukherjee, Yiling Lu, and Gordon B. Mills. A pan-cancer proteomic perspective on the cancer genome atlas. *Nature Communications*, 5, may 2014.

[3] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106+, October 2010.

[4] Simon Anders, Davis J. McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K. Smyth, Wolfgang Huber, and Mark D. Robinson. Count-based differential expression analysis of RNA sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765–86, September 2013.

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

[6] Yan W. Asmann, Brian M. Necela, Krishna R. Kalari, Asif Hossain, Tiffany R. Baker, Jennifer M. Carr, Caroline Davis, Julie E. Getz, Galen Hostetter, Xing Li, Sarah A. McLaughlin, Derek C. Radisky, Gary P. Schroth, Heather E. Cunliffe, Edith A. Perez, and E. Aubrey Thompson. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer research*, 72(8):1921–1928, April 2012.

[7] Kin Fai F. Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung H. Wong. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, 38(14):4570–4578, August 2010.

[8] Dimitrios Balafoutas, Axel zur Hausen, Sebastian Mayer, Marc Hirschfeld, Markus Jaeger, Dominik Denschlag, Gerald Gitsch, Achim Jungbluth, and Elmar Stickeler. Cancer testis antigens and NY-BR-1 expression in primary breast cancer: prognostic and therapeutic implications. *BMC cancer*, 13:271+, June 2013.

[9] Francisco E. Baralle and Jimena Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, advance online publication, May 2017.

[10] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar 2012.

[11] Giacomo Baruzzo, Katharina E. Hayer, Eun J. Kim, Barbara Di Camillo, Garret A. FitzGerald, and Gregory R. Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, 14(2):135–139, December 2016.

[12] Avrim Blum, Tao Jiang, Ming Li, John Tromp, and Mihalis Yannakakis. Linear approximation of shortest superstrings. *J. ACM*, 41(4):630–647, July 1994.

[13] Daniel R. Boutz, Andrew Pitchford P. Horton, Yariv Wine, Jason J. Lavinder, George Georgiou, and Edward M. Marcotte. Proteomic identification of monoclonal antibodies from serum. *Analytical chemistry*, March 2014.

[14] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, April 2016.

[15] Natalie E. Castellana, Zhouxin Shen, Yupeng He, Justin W. Walley, California J. Cassidy, Steven P. Briggs, and Vineet Bafna. An automated proteogenomic method uses mass spectrometry to reveal novel genes in zea mays. *Molecular & Cellular Proteomics*, 13(1):157–167, January 2014.

[16] Anthony J. Cesnik, Michael R. Shortreed, Gloria M. Sheynkman, Brian L. Frey, and Lloyd M. Smith. Human proteomic variation revealed by combining RNA-seq proteogenomics and global Post-Translational modification (G-PTM) search strategy. *J. Proteome Res.*, December 2015.

[17] Cheng Chang, Hira Lal L. Goel, Huijie Gao, Bryan Pursell, Leonard D. Shultz, Dale L. Greiner, Sulev Ingerpuu, Manuel Patarroyo, Shiliang Cao, Elgene Lim, Junhao Mao, Karen Kulju K. McKee, Peter D. Yurchenco, and Arthur M. Mercurio. A laminin 511 matrix is regulated by TAZ and functions as the ligand for the Îś6BÎš1 integrin to sustain breast cancer stem cells. *Genes & development*, 29(1):1–6, January 2015.

[18] L. C. Chen, K. Matsumura, G. Deng, W. Kurisu, B. M. Ljung, M. I. Lerman, F. M. Waldman, and H. S. Smith. Deletion of two separate regions on chromosome 3p in breast cancers. *Cancer Res.*, 54(11):3021–3024, Jun 1994.

[19] Wan C. Cheung, Sean A. Beausoleil, Xiaowu Zhang, Shuji Sato, Sandra M. Schieferl, James S. Wieler, Jason G. Beaudet, Ravi K. Ramenani, Lana Popova, Michael J. Comb, John Rush, and Roberto D. Polakiewicz. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature Biotechnology*, 30(5):447–452, May 2012.

[20] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[21] P. C. Consul and G. C. Jain. A generalization of the poisson distribution. *Technometrics*, 15(4):791–799, 1973.

[22] F. H. C. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970.

[23] Phuong Dao, Ibrahim Numanagić, Yen-Yi Lin, Faraz Hach, Emre Karakoc, Nilgun Donmez, Colin Collins, Evan E. Eichler, and S. Cenk Sahinalp. ORMAN: optimal resolution of ambiguous RNA-seq multimappings in the presence of novel isoforms. *Bioinformatics (Oxford, England)*, 30(5):644–651, March 2014.

[24] Nadia Davidson and Alicia Oshlack. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology*, 15:410+, July 2014.

[25] N. G. de Bruijn. A combinatorial problem. *Koninklijke Nederlandsche Akademie Van Wetenschappen*, 49(6):758–764, June 1946.

[26] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

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

[28] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, March 2007.

[29] Matthew J. Ellis, Michael Gillette, Steven A. Carr, Amanda G. Paulovich, Richard D. Smith, Karin K. Rodland, Reid R. Townsend, Christopher Kinsinger, Mehdi Mesri, Henry Rodriguez, and Daniel C. Liebler. Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. *Cancer Discovery*, 3(10):1108–1112, October 2013.

[30] Pär G. Engström, Tamara Steijger, Botond Sipos, Gregory R. Grant, André Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, Carrie A. Davis, Alexander Dobin, Pär G. Engström, Thomas R. Gingeras, Nick Goldman, Gregory R. Grant, Roderic Guigó, Jennifer Harrow, Tim J. Hubbard, Géraldine Jean, André Kahles, Peter Kosarev, Sheng Li, Jinze Liu, Christopher E. Mason, Vladimir Molodtsov, Zemin Ning, Hannes Ponstingl, Jan F. Prins, Gunnar Rätsch, Paolo Ribeca, Igor Seledtsov, Botond Sipos, Victor Solovyev, Tamara Steijger, Giorgio Valle, Nicola Vitulo, Kai Wang, Thomas D. Wu, Georg Zeller, Gunnar Rätsch, Nick Goldman, Tim J. Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 10(12):1185–1191, November 2013.

[31] Ensembl. Human Protein Sequence. `ftp://ftp.ensembl.org/pub/release-70/fasta/homo_sapiens/pep/Homo_sapiens.GRCh37.70.pep.all.fa.gz`, 2012. [Online; accessed 25-November-2015].

[32] Ensembl. Ensembl Genome Browser. `http://www.ensembl.org/info/data/ftp/index.html`, 2017. [Online; accessed 12-June-2015].

[33] I. P. Ewald, P. L. Ribeiro, E. I. Palmero, S. L. Cossio, R. Giugliani, and P. Ashton-Prolla. Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet. Mol. Biol.*, 32(3):437–446, Jul 2009.

[34] X. Fan, T. E. Abbott, D. Larson, and K. Chen. BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics*, 2014, 2014.

[35] Jianxing Feng, Wei Li, and Tao Jiang. Inference of isoforms from short sequence reads. *Journal of Computational Biology*, 18(3):305–321, March 2011.

[36] J. L. Fernandez-Luna. Bcr-Abl and inhibition of apoptosis in chronic myelogenous leukemia cells. *Apoptosis : an international journal on programmed cell death*, 5(4):315–318, October 2000.

[37] Milana Frenkel-Morgenstern, Alessandro Gorohovski, Vincent Lacroix, Mark Rogers, Kristina Ibanez, Cesar Boullosa, Eduardo A. Leon, Asa Ben-Hur, and Alfonso Valencia. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Research*, 41(D1):D142–D151, January 2013.

[38] Milana Frenkel-Morgenstern, Alessandro Gorohovski, Dunja Vucenovic, Lorena Maestre, and Alfonso Valencia. ChiTaRS 2.1âĂŤan improved database of the chimeric transcripts and RNA-seq data with novel senseâĂŞantisense chimeric RNA transcripts. *Nucleic Acids Research*, pages gku1199+, November 2014.

[39] Milana Frenkel-Morgenstern, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Del Pozo, Michael Tress, Rory Johnson, Roderic Guigo, and Alfonso Valencia. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome research*, 22(7):1231–1242, July 2012.

[40] P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–183, March 2004.

[41] John Gallant, David Maier, and James Astorer. On finding minimal length superstrings. *Journal of Computer and System Sciences*, 20(1):50 – 58, 1980.

[42] Manuel Garber, Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–77, June 2011.

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

[44] Michael A. Gillette and Steven A. Carr. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nature Methods*, 10(1):28–34, December 2012.

[45] Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, July 2011.

[46] Mitchell Guttman, Manuel Garber, Joshua Z. Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J. Koziol, Andreas Gnirke, Chad Nusbaum, John L. Rinn, Eric S. Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, May 2010.

[47] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods*, 7(8):576–577, August 2010.

[48] Faraz Hach, Iman Sarrafi, Farhad Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. mrsFAST-ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Research*, May 2014.

[49] Thomas Hardcastle and Krystyna Kelly. baySeq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422+, 2010.

[50] S. Hemmer, V. M. Wasenius, C. Haglund, Y. Zhu, S. Knuutila, K. Franssila, and H. Joensuu. Deletion of 11q23 and cyclin D1 overexpression are frequent aberrations in parathyroid adenomas. *Am. J. Pathol.*, 158(4):1355–1362, Apr 2001.

[51] 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+, September 2009.

[52] Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, July 2009.

[53] Fereydoun Hormozdiari, Iman Hajirasouliha, Andrew McPherson, Evan E Eichler, and S Cenk Sahinalp. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome research*, 21(12):2203–2212, 2011.

[54] Matthew K. Iyer and Arul M. Chinnaiyan. RNA-seq unleashed. *Nature Biotechnology*, 29(7):599–600, July 2011.

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

[56] Alexander Kanitz, Foivos Gypas, Andreas J. Gruber, Andreas R. Gruber, Georges Martin, and Mihaela Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, 16(1):150+, July 2015.

[57] E. Karakoc, C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson, and E. E. Eichler. Detection of structural variants and indels within exome data. *Nat. Methods*, 9(2):176–178, Feb 2012.

[58] R.M. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, 40(4):85–103, 1972.

[59] Pinar Kavak, Yen-Yi Lin, Ibrahim Numanagić, Hossein Asghari, Tunga Gungor, Can Alkan, and Faraz Hach. Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics (Oxford, England)*, 33(14,15):i161âĂŞi169, July 2017.

[60] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, April 2015.

[61] Daehwan Kim and Steven Salzberg. Tophat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):R72, 2011.

[62] Sangtae Kim and Pavel A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277+, October 2014.

[63] Marcus Kinsella, Olivier Harismendy, Masakazu Nakano, Kelly A. Frazer, and Vineet Bafna. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, 2011.

[64] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Hye-Jung E. Chun, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Erin Pleasance, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, Nam H. Pho, Scott L. Carter, Steven E. Schumacher, Barbara Tabak, Bryan Hernandez, Jeff Gentry, Huy Nguyen, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Gad Getz, Stacey B. Gabriel, Matthew Meyerson, Lynda Chin, Peter J. Park, Raju Kucherlapati, Katherine A. Hoadley, J. Todd Auman, Cheng Fan, Yidi J. Turman, Yan Shi, Ling Li, Michael D. Topal, Xiaping He, Hann-Hsiang Chao, Aleix Prat, Grace O. Silva, Michael D. Iglesia, Wei Zhao, Jerry Usary, Jonathan S. Berg, Michael Adams, Jessica Booker, Junyuan Wu, Anisha Gulabani, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Simeen Malik, Swapna Mahurkar, Hui Shen, Daniel J. Weisenberger, Timothy Triche Jr, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Benjamin P. Berman, David J. Van Den Berg, Stephen B. Baylin, Peter W. Laird, Chad J. Creighton, Lawrence A. Donehower, Gad Getz, Michael Noble, Doug Voet, Gordon Saksena, Nils Gehlenborg, Daniel DiCara, Juinhua Zhang, Hailei Zhang, Chang-Jiun Wu, Spring Yingchun Liu, Michael S. Lawrence, Lihua Zou, Andrey Sivachenko, Pei Lin, Petar Stojanov, Rui Jing, Juok Cho, Raktim Sinha, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Peter J. Park, Lynda Chin, Sheila Reynolds, Richard B. Kreisberg, Brady Bernard, Ryan Bressler, Timo Erkkila, Jake Lin, Vesteinn Thorsson, Wei Zhang, Ilya Shmulevich, Giovanni Ciriello, Nils Weinhold, Nikolaus Schultz, Jianjiong Gao, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Marc Ladanyi, Chris Sander, Pavana Anur, Paul T. Spellman, Yiling Lu, Wenbin Liu, Roel R. G. Verhaak, Gordon B. Mills, Rehan Akbani,

Nianxiang Zhang, Bradley M. Broom, Tod D. Casasent, Chris Wakefield, Anna K. Unruh, Keith Baggerly, Kevin Coombes, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, Jingchun Zhu, Christopher C. Szeto, Gary K. Scott, Christina Yau, Evan O. Paull, Daniel Carlin, Christopher Wong, Artem Sokolov, Janita Thusberg, Sean Mooney, Sam Ng, Theodore C. Goldstein, Kyle Ellrott, Mia Grifford, Christopher Wilks, Singer Ma, Brian Craft, Chunhua Yan, Ying Hu, Daoud Meerzaman, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron D. Black, Robert E. XPATH ERROR: unknown variable "tname"., Peter White, Erik J. Zmuda, Jessica Frick, Tara M. Lichtenberg, Robin Brookens, Myra M. George, Mark A. Gerken, Hollie A. Harper, Kristen M. Leraas, Lisa J. Wise, Teresa R. Tabler, Cynthia McAllister, Thomas Barr, Melissa Hart-Kothari, Katie Tarvin, Charles Saller, George Sandusky, Colleen Mitchell, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Oleg Dolzhansky, Mikhail Abramov, Olga Voronina, Olga Potapova, Jeffrey R. Marks, Wiktoria M. Suchorska, Dawid Murawa, Witold Kycler, Matthew Ibbs, Konstanty Korski, Arkadiusz Spychała, Paweł Murawa, Jacek J. Brzeziński, Hanna Perz, Radosław Łaźniak, Marek Teresiak, Honorata Tatka, Ewa Leporowska, Marta Bogusz-Czerniewicz, Julian Malicki, Andrzej Mackiewicz, Maciej Wiznerowicz, Xuan Van Le, Bernard Kohl, Nguyen Viet Tien, Richard Thorp, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Tran Viet The Phuong, Huynh Quyet Thang, Khurram Zaki Khan, Robert Penny, David Mallery, Erin Curley, Candace Shelton, Peggy Yena, James N. Ingle, Fergus J. Couch, Wilma L. Lingle, Tari A. King, Ana Maria Gonzalez-Angulo, Gordon B. Mills, Mary D. Dyer, Shuying Liu, Xiaolong Meng, Modesto Patangan, Frederic Waldman, Hubert Stöppler, W. Kimryn Rathmell, Leigh Thorne, Mei Huang, Lori Boice, Ashley Hill, Carl Morrison, Carmelo Gaudioso, Wiam Bshara, Kelly Daily, Sophie C. Egea, Mark D. Pegram, Carmen Gomez-Fernandez, Rajiv Dhir, Rohit Bhargava, Adam Brufsky, Craig D. Shriver, Jeffrey A. Hooke, Jamie Leigh Campbell, Richard J. Mural, Hai Hu, Stella Somiari, Caroline Larson, Brenda Deyarmin, Leonid Kvecher, Albert J. Kovatich, Matthew J. Ellis, Tari A. King, Hai Hu, Fergus J. Couch, Richard J. Mural, Thomas Stricker, Kevin White, Olufunmilayo Olopade, James N. Ingle, Chunqing Luo, Yaqin Chen, Jeffrey R. Marks, Frederic Waldman, Maciej Wiznerowicz, Ron Bose, Li-Wei Chang, Andrew H. Beck, Ana Maria Gonzalez-Angulo, Todd Pihl, Mark Jensen, Robert Sfeir, Ari Kahn, Anna Chu, Prachi Kothiyal, Zhining Wang, Eric Snyder, Joan Pontius, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique Berton, Matthew Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter Kigonya, Shelley Alonso, Rashmi Sanbhadti, Sean Barletta, David Pot, Margi Sheth, John A. Demchok, Kenna R. Mills Shaw, Liming Yang, Greg Eley, Martin L. Ferguson, Roy W. Tarnuzzer, Jiashan Zhang, Laura A. L. Dillon, Kenneth Buetow, Peter Fielding, Bradley A. Ozenberger, Mark S. Guyer, Heidi J. Sofia, and Jacqueline D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, sep 2012.

[65] V. Kolla, T. Zhuang, M. Higashi, K. Naraparaju, and G. M. Brodeur. Role of CHD5 in human cancers: 10 years later. *Cancer Res.*, 74(3):652–658, Feb 2014.

[66] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[67] Anna V Lapuk, Chunxiao Wu, Alexander W Wyatt, Andrew McPherson, Brian J McConeghy, Sonal Brahmbhatt, Fan Mo, Amina Zoubeidi, Shawn Anderson, Robert H

Bell, et al. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. *The Journal of pathology*, 227(3):286–297, 2012.

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

[69] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29+, February 2014.

[70] Bo Li and Colin N. Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323+, August 2011.

[71] 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, February 2010.

[72] Dequan Li, Yan Fu, Ruixiang Sun, Charles X. Ling, Yonggang Wei, Hu Zhou, Rong Zeng, Qiang Yang, Simin He, and Wen Gao. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics*, 21(13):3049–3050, July 2005.

[73] Jingyi Jessica J. Li, Ci-Ren R. Jiang, James B. Brown, Haiyan Huang, and Peter J. Bickel. Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):19867–19872, December 2011.

[74] Sheng Li, Pawel P. Labaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, Charles Wang, Danielle Thierry-Mieg, Jean Thierry-Mieg, David P. Kreil, and Christopher E. Mason. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology*, 32(9):888–895, August 2014.

[75] Wei Li, Jianxing Feng, and Tao Jiang. IsoLasso: A LASSO regression approach to RNA-seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707, November 2011.

[76] Wei Li and Tao Jiang. Transcriptome assembly and isoform expression level estimation from biased RNA-seq reads. *Bioinformatics (Oxford, England)*, 28(22):2914–2921, November 2012.

[77] Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.

[78] Yen-Yi Lin, Phuong Dao, Faraz Hach, Marzieh Bakhshi, Fan Mo, Anna Lapuk, Colin Collins, and SÃijleyman Cenk Sahinalp. CLIIQ: Accurate comparative detection and quantification of expressed isoforms in a population. In Benjamin J. Raphael and Jijun Tang, editors, *WABI*, volume 7534 of *Lecture Notes in Computer Science*, pages 178–189. Springer, 2012.

[79] Silvia Liu, Wei-Hsiang H. Tsai, Ying Ding, Rui Chen, Zhou Fang, Zhiguang Huo, SungHwan Kim, Tianzhou Ma, Ting-Yu Y. Chang, Nolan Michael M. Priedigkeit,

Adrian V. Lee, Jianhua Luo, Hsei-Wei W. Wang, I-Fang F. Chung, and George C. Tseng. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Research*, 44(5), March 2016.

[80] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, 2012.

[81] Zin-Mar Lwin, Chunhua Guo, Agus Salim, George W. Yip, Fook-Tim Chew, Jiang Nan, Aye A. Thike, Puay-Hoon Tan, and Boon-Huat Bay. Clinicopathological significance of calreticulin in breast invasive ductal carcinoma. *Modern Pathology*, 23(12):1559–1566, September 2010.

[82] 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, March 2009.

[83] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, May 2005.

[84] Andrew McPherson, Fereydoun Hormozdiari, Abdalnasser Zayed, Ryan Giuliany, Gavin Ha, Mark G. F. Sun, Malachi Griffith, Alireza Heravi Moussavi, Janine Senz, Nataliya Melnyk, Marina Pacheco, Marco A. Marra, Martin Hirst, Torsten O. Nielsen, S. Cenk Sahinalp, David Huntsman, and Sohrab P. Shah. deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*, 7(5):e1001138, 05 2011.

[85] Andrew McPherson, Chunxiao Wu, Iman Hajirasouliha, Fereydoun Hormozdiari, Faraz Hach, Anna Lapuk, Stanislav Volik, Sohrab Shah, Colin Collins, and S. Cenk Sahinalp. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics*, 27(11):1481–1488, 2011.

[86] Andrew McPherson, Chunxiao Wu, Alexander W. Wyatt, Sohrab Shah, Colin Collins, and S. Cenk Sahinalp. nFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Research*, 22(11):2250–2261, November 2012.

[87] Robert R. McWilliams, Eric D. Wieben, Kari G. Rabe, Katrina S. Pedersen, Yanhong Wu, Hugues Sicotte, and Gloria M. Petersen. Prevalence of CDKN2A mutations in pancreatic cancer patients: implications for genetic counseling. *European journal of human genetics : EJHG*, 19(4):472–478, April 2011.

[88] Melissa G. Mendez, Shin-Ichiro Kojima, and Robert D. Goldman. Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. *FASEB journal*, 24(6):1838–1851, June 2010.

[89] Fredrik Mertens, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15(6):371–381, May 2015.

[90] Philipp Mertins, D. R. Mani, Kelly Ruggles, Michael Gillette, Karl Clauser, Pei Wang, Xianlong Wang, Jana Qiao, Song Cao, Francesca Petralia, Filip Mundt, Zhidong Tu, Jonathan Lei, Michael Gatza, Matthew Wilkerson, Charles Perou, Venkata Yellapantula, Kuan-lin Huang, Chenwei Lin, Michael McLellan, Ping Yan, Sherri Davies, Reid Townsend, Steven Skates, Jing Wang, Bing Zhang, Christopher Kinsinger, Mehdi Mesri, Henry Rodriguez, Li Ding, Amanda Paulovich, David Fenyő, Matthew Ellis, Steven Carr, and NCI CPTAC. Abstract IA29: Proteogenomic and phosphoproteomic analysis of breast cancer. *Molecular Cancer Research*, 14(2 Supplement):IA29, February 2016.

[91] Philipp Mertins, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang, Jana W. Qiao, Song Cao, Francesca Petralia, Emily Kawaler, Filip Mundt, Karsten Krug, Zhidong Tu, Jonathan T. Lei, Michael L. Gatza, Matthew Wilkerson, Charles M. Perou, Venkata Yellapantula, Kuan-lin Huang, Chenwei Lin, Michael D. McLellan, Ping Yan, Sherri R. Davies, Reid R. Townsend, Steven J. Skates, Jing Wang, Bing Zhang, Christopher R. Kinsinger, Mehdi Mesri, Henry Rodriguez, Li Ding, Amanda G. Paulovich, David Fenyő, Matthew J. Ellis, Steven A. Carr, and NCI CPTAC. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, June 2016.

[92] Aziz M. Mezlini, Eric J. M. Smith, Marc Fiume, Orion Buske, Gleb Savich, Sohrab Shah, Sam Aparicion, Derek Chiang, Anna Goldenberg, and Michael Brudno. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Research*, 23(3):gr.142232.112–529, November 2012.

[93] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews. Cancer*, 7(4):233–245, April 2007.

[94] Fan Mo, Xu Hong, Feng Gao, Lin Du, Jun Wang, Gilbert S. Omenn, and Biaoyang Lin. A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics*, 9(1):537+, December 2008.

[95] Ute M. Moll and Oleksi Petrenko. The MDM2-p53 interaction. *Molecular Cancer Research*, 1(14):1001–1008, December 2003.

[96] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628, July 2008.

[97] Mehnaz G. Mustafa, John R. Petersen, Hyunsu Ju, Luca Cicalese, Ned Snyder, Sigmund J. Haidacher, Larry Denner, and Cornelis Elferink. Biomarker discovery for early detection of hepatocellular carcinoma in hepatitis câĂŞinfected patients. *Molecular & Cellular Proteomics*, 12(12):3640–3652, December 2013.

[98] Ugrappa Nagalakshmi, Zhong Wang, Chong Shou, Michael Snyder, Debasish Raha, Mark Gerstein, and Karl Waern. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320:1344–1349, June 2008.

[99] M. Nakao, S. Yokota, T. Iwai, H. Kaneko, S. Horiike, K. Kashima, Y. Sonoda, T. Fujimoto, and S. Misawa. Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. *Leukemia*, 10(12):1911–1918, Dec 1996.

[100] Alexey I. Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, November 2014.

[101] Marius Nicolae, Serghei Mangul, Ion I. Măndoiu, and Alex Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-seq data. *Algorithms for molecular biology : AMB*, 6(1):9+, 2011.

[102] Yashar S. Niknafs, Balaji Pandian, Hariharan K. Iyer, Arul M. Chinnaiyan, and Matthew K. Iyer. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nature Methods*, 14(1):68–70, November 2016.

[103] Kang Ning, Damian Fermin, and Alexey I. Nesvizhskii. Comparative analysis of different Label-Free mass spectrometry based protein abundance estimates and their correlation with RNA-seq gene expression data. *J. Proteome Res.*, 11(4):2261–2271, February 2012.

[104] Kang Ning and Alexey Nesvizhskii. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-seq data: a preliminary assessment. *BMC Bioinformatics*, 11(Suppl 11):S14+, 2010.

[105] Bogdan Pasaniuc, Noah Zaitlen, and Eran Halperin. Accurate estimation of expression levels of homologous genes in rna-seq experiments. In *RECOMB*, pages 397–409, 2011.

[106] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, March 2017.

[107] Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, May 2014.

[108] Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Ming-Ju Lv, Xin-Guang Zhu, and Francis Y. L. Chin. IDBA-tran: a more robust de novo de bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29(13):i326–i334, July 2013.

[109] Mihaela Pertea, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng C. Chang, Joshua T. Mendell, and Steven L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, February 2015.

[110] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. C. Mell, and I. M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, 20(5):623–635, May 2010.

[111] T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.

[112] Jüri Reimand, Omar Wagih, and Gary D. Bader. The mutational landscape of phosphorylation signaling in cancer. *Scientific reports*, 3, October 2013.

[113] Adam Roberts, Cole Trapnell, Julie Donaghey, JohnL Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):1–14, 2011.

[114] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D. Jackman, Karen Mungall, Sam Lee, Hisanaga Mark M. Okada, Jenny Q. Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S. Butterfield, Richard Newsome, Simon K. Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna-Liisa L. Prabhu, Angela Tam, YongJun Zhao, Richard A. Moore, Martin Hirst, Marco A. Marra, Steven J. Jones, Pamela A. Hoodless, and Inanc Birol. De novo assembly and analysis of RNA-Seq data. *Nature Methods*, 7(11):909–912, November 2010.

[115] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[116] Janet D. Rowley. Chromosomal translocations: revisited yet again. *Blood*, 112(6):2183–2189, September 2008.

[117] Roye Rozov, Eran Halperin, and Ron Shamir. MGMR: leveraging RNA-seq population data to optimize expression estimation. *BMC Bioinformatics*, 13(Suppl 6):S2+, 2012.

[118] Paul A. Rudnick, Sanford P. Markey, Jeri Roth, Yuri Mirokhin, Xinjian Yan, Dmitrii V. Tchekhovskoi, Nathan J. Edwards, Ratna R. Thangudu, Karen A. Ketchum, Christopher R. Kinsinger, Mehdi Mesri, Henry Rodriguez, and Stephen E. Stein. A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. *J. Proteome Res.*, 15(3):1023–1032, March 2016.

[119] Kelly V. Ruggles, Karsten Krug, Xiaojing Wang, Karl R. Clauser, Jing Wang, Samuel H. Payne, David Fenyö, Bing Zhang, and D. R. Mani. Methods, tools and current perspectives in proteogenomics. *Molecular & Cellular Proteomics*, pages mcp.000024.2017+, April 2017.

[120] Thomas Santarius, Janet Shipley, Daniel Brewer, Michael R. Stratton, and Colin S. Cooper. A census of amplified and overexpressed human cancer genes. *Nature Reviews Cancer*, 10(1):59–64, January 2010.

[121] A. A. Sartori, C. Lukas, J. Coates, M. Mistrik, S. Fu, J. Bartek, R. Baer, J. Lukas, and S. P. Jackson. Human CtIP promotes DNA end resection. *Nature*, 450(7169):509–514, Nov 2007.

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

[123] M. Schöniger and M. S. Waterman. A local algorithm for DNA sequence alignment with inversions. *Bulletin of mathematical biology*, 54(4):521–536, July 1992.

[124] J. Schroder, A. Hsu, S. E. Boyle, G. Macintyre, M. Cmero, R. W. Tothill, R. W. Johnstone, M. Shackleton, and A. T. Papenfuss. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, Jan 2014.

[125] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[126] Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease. *Nature reviews. Genetics*, 17(1):19–32, January 2016.

[127] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, Aug 2012.

[128] Gloria M. Sheynkman, Michael R. Shortreed, Brian L. Frey, and Lloyd M. Smith. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-seq. *Molecular & cellular proteomics : MCP*, 12(8):2341–2353, August 2013.

[129] L. Shi et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24(9):1151–1161, Sep 2006.

[130] S. P. Shriver, M. D. Shriver, D. L. Tirpak, L. M. Bloch, J. D. Hunt, R. E. Ferrell, and J. M. Siegfried. Trinucleotide repeat length variation in the human ribosomal protein L14 gene (RPL14): localization to 3p21.3 and loss of heterozygosity in lung and oral cancers. *Mutat. Res.*, 406(1):9–23, Nov 1998.

[131] Christopher R. Sibley, Lorea Blazquez, and Jernej Ule. Lessons from non-canonical splicing. *Nature reviews. Genetics*, May 2016.

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

[133] S. S. Sindi, S. Onal, L. C. Peng, H. T. Wu, and B. J. Raphael. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, 13(3):R22, 2012.

[134] Manabu Soda, Young Lim L. Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina, Hisashi Hatanaka, Masashi Bando, Shoji Ohno, Yuichi Ishikawa, Hiroyuki Aburatani, Toshiro Niki, Yasunori Sohara, Yukihiko Sugiyama, and Hiroyuki Mano. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153):561–566, August 2007.

[135] I. Soria-Bretones, C. Saez, M. Ruiz-Borrego, M. A. Japon, and P. Huertas. Prognostic value of CtIP/RBBP8 expression in breast cancer. *Cancer Med*, 2(6):774–783, Dec 2013.

[136] J. L. Stanford, J. J. Just, M. Gibbs, K. G. Wicklund, C. L. Neal, B. A. Blumenstein, and E. A. Ostrander. Polymorphic repeats in the androgen receptor gene: molecular markers of prostate cancer risk. *Cancer Res.*, 57(6):1194–1198, Mar 1997.

[137] Tamara Steijger, Josep F. Abril, Par G. Engstrom, Felix Kokocinski, The RGASP Consortium, Tim J. Hubbard, Roderic Guigo, Jennifer Harrow, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184, November 2013.

[138] Lucas Swanson, Gordon Robertson, Karen Mungall, Yaron Butterfield, Readman Chiu, Richard Corbett, T Docking, Donna Hogge, Shaun Jackman, Richard Moore, Andrew Mungall, Ka Nip, Jeremy Parker, Jenny Qian, Anthony Raymond, Sandy Sung, Angela Tam, Nina Thiessen, Richard Varhol, Sherry Wang, Deniz Yorukoglu,

YongJun Zhao, Pamela Hoodless, S Sahinalp, Aly Karsan, and Inanc Birol. Barnacle: detecting and characterizing tandem duplications and fusions in transcriptome assemblies. *BMC Genomics*, 14(1):550, 2013.

[139] Linda Szabo and Julia Salzman. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.*, 17(11):679–692, 10 2016.

[140] Qingming Tang, Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Bermuda: De novo assembly of transcripts with new insights for handling uneven coverage. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, pages 166–175, New York, NY, USA, 2015. ACM.

[141] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[142] Cristina Tognon, Stevan R. Knezevich, David Huntsman, Calvin D. Roskelley, Natalya Melnyk, Joan A. Mathers, Laurence Becker, Fatima Carneiro, Nicol MacPherson, Doug Horsman, Christopher Poremba, and Poul H. Sorensen. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer cell*, 2(5):367–376, November 2002.

[143] Alexandru I. Tomescu, Anna Kuosmanen, Romeo Rizzi, and Veli Mäkinen. A novel min-cost flow method for estimating transcript expression with RNA-seq. *BMC Bioinformatics*, 14(Suppl 5):S15+, April 2013.

[144] AlexandruI Tomescu, Anna Kuosmanen, Romeo Rizzi, and Veli Mäkinen. A novel combinatorial method for estimating transcript expression with RNA-seq: Bounding the number of paths. In Aaron Darling and Jens Stoye, editors, *Algorithms in Bioinformatics*, volume 8126 of *Lecture Notes in Computer Science*, pages 85–98. Springer Berlin Heidelberg, 2013.

[145] Scott A. Tomlins, Bharathi Laxman, Saravana M. Dhanasekaran, Beth E. Helgeson, Xuhong Cao, David S. Morris, Anjana Menon, Xiaojun Jing, Qi Cao, Bo Han, Jindan Yu, Lei Wang, James E. Montie, Mark A. Rubin, Kenneth J. Pienta, Diane Roulston, Rajal B. Shah, Sooryanarayana Varambally, Rohit Mehra, and Arul M. Chinnaiyan. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature*, 448(7153):595–599, August 2007.

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

[147] 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. *Nature Biotechnology*, 28(5):511–515, May 2010.

[148] UCSC. UCSC Genome Browser. `http://hgdownload.cse.ucsc.edu/downloads.html`, 2017. [Online; accessed 12-June-2015].

[149] Katherine E. Varley, Jason Gertz, Brian S. Roberts, Nicholas S. Davis, Kevin M. Bowling, Marie K. Kirby, Amy S. Nesmith, Patsy G. Oliver, William E. Grizzle, Andres Forero, Donald J. Buchsbaum, Albert F. LoBuglio, and Richard M. Myers. Recurrent read-through fusion transcripts in breast cancer. *Breast cancer research and treatment*, 146(2):287–297, July 2014.

[150] J. Cristobal Vera, Christopher W. Wheat, Howard W. Fescemyer, Mikko J. Frilander, Douglas L. Crawford, Ilkka Hanski, and James H. Marden. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular ecology*, 17(7):1636–1647, April 2008.

[151] Nagarjun Vijay, Jelmer W. Poelstra, Axel Künstner, and Jochen B. Wolf. Challenges and strategies in transcriptome assembly and differential gene expression quantification. a comprehensive in silico assessment of RNA-seq experiments. *Molecular ecology*, 22(3):620–634, February 2013.

[152] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, December 2012.

[153] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178, October 2010.

[154] Le-Heng H. Wang, De-Quan Q. Li, Yan Fu, Hai-Peng P. Wang, Jing-Fen F. Zhang, Zuo-Fei F. Yuan, Rui-Xiang X. Sun, Rong Zeng, Si-Min M. He, and Wen Gao. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM*, 21(18):2985–2991, 2007.

[155] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

[156] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server issue):W214–220, Jul 2010.

[157] Jeffrey R. Whiteaker, Goran N. Halusa, Andrew N. Hoofnagle, Vagisha Sharma, Brendan MacLean, Ping Yan, John A. Wrobel, Jacob Kennedy, D. R. Mani, Lisa J. Zimmerman, Matthew R. Meyer, Mehdi Mesri, Henry Rodriguez, Clinical Proteomic Tumor Analysis Consortium (CPTAC), and Amanda G. Paulovich. CPTAC assay portal: a repository of targeted proteomic assays. *Nature Methods*, 11(7):703–704, July 2014.

[158] Sunghee Woo, Seong W. Cha, Gennifer Merrihew, Yupeng He, Natalie Castellana, Clark Guest, Michael MacCoss, and Vineet Bafna. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.*, 13(1):21–28, June 2013.

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

[160] X. Wu, Z. Zhu, W. Li, X. Fu, D. Su, L. Fu, Z. Zhang, A. Luo, X. Sun, L. Fu, and J. T. Dong. Chromodomain helicase DNA binding protein 5 plays a tumor suppressor role in human breast cancer. *Breast Cancer Res.*, 14(3):R73, 2012.

[161] Zhengpeng Wu, Xi Wang, and Xuegong Zhang. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics*, 27(4):502–508, 2011.

[162] Julia D. Wulfkuhle, Lance A. Liotta, and Emanuel F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviews Cancer*, 3(4):267–275, April 2003.

[163] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Tak-Wah W. Lam, Yingrui Li, Xun Xu, Gane Ka-Shu K. Wong, and Jun Wang. SOAPdenovo-trans: de novo transcriptome assembly with short RNA-seq reads. *Bioinformatics (Oxford, England)*, 30(12):1660–1666, June 2014.

[164] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. 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.

[165] Deniz Yorukoglu, Faraz Hach, Lucas Swanson, Colin C. Collins, Inanc Birol, and S. Cenk Sahinalp. Dissect: detection and characterization of novel structural alterations in transcribed sequences. *Bioinformatics*, 28(12):i179–i187, 2012.

[166] K Yoshihara, Q Wang, W Torres-Garcia, S Zheng, R Vegesna, H Kim, and R G W Verhaak. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, 34(37):4845–4854, dec 2014.

[167] Dihua Yu and Mien-Chie Hung. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene*, 19(53):6115–6121, Dec 2000.

[168] J. Yuan and J. Chen. MRE11-RAD50-NBS1 complex dictates DNA repair independent of H2AX. *J. Biol. Chem.*, 285(2):1097–1104, Jan 2010.

[169] Daniel R. Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.

[170] Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, Lisa J. Zimmerman, Kent F. Shaddox, Sangtae Kim, Sherri R. Davies, Sean Wang, Pei Wang, Christopher R. Kinsinger, Robert C. Rivers, Henry Rodriguez, Reid R. Townsend, Matthew J. C. Ellis, Steven A. Carr, David L. Tabb, Robert J. Coffey, Robbert J. C. Slebos, Daniel C. Liebler, and the NCI CPTAC. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, September 2014.