Detecting anomalies in inferred transcript sequences and expression from RNA-seq

Cong Ma

CMU-CB-20-101

September 2020

Computational Biology Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Cark Kingsford, Chair
Russell Schwartz
Xinghua Lu
Ben Raphael

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research is funded in part by the US National Science Foundation (CCF-1256087, CCF-1319998) and by grant 4100070287 from the Pennsylvania Department of Health. Research reported in this document was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM122935. C.K. received support as an Alfred P. Sloan Research Fellow. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: anomaly detection, RNA-seq, transcriptomic structural variants, expression quantification, unannotated isoforms

To everyone, every pet, every bouquet of flowers, in my life.

Abstract

Anomalies are data points that do not follow established or expected patterns. When measuring gene expression, anomalies in RNA-seq are observations or patterns that cannot be explained by the inferred transcript sequences or expressions. Transcript sequences and expression are key indicators for cell status and are used in many phenotypic and disease analyses. Identifying such unexplainable RNA-seq patterns can inspire improvements in the accuracy of inferred transcript sequences and expression of RNA-seq data and benefit the analyses based on transcripts. We develop computational methods to identify the RNA-seq anomalies that violate inferred sequence variation and expression patterns, and to improve the reconstructed transcripts such that they can explain the anomalies.

The first type of anomaly that we detect is the large-scale sequence variation in transcriptome, or transcriptomic structural variants (TSVs). TSVs are usually induced by genomic structural variants, which can fuse sequences either from a pair of genes or involve intergenic regions. Previous TSV detection methods assume that TSVs only fuse a pair genes and do not consider that some genes are still unknown, thus many RNA-seq reads from the intergenic or intronic regions cannot be explained by gene fusions. We develop a computational method, SQUID, to identify fusions both between a pair of genes and involving non-transcribing regions, thus enlarging the set of explained variants and RNA-seq reads. SQUID is further extended to the MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM, which is able to detect TSVs in the allele heterogeneity context. The second type of anomaly that we identify are coverage anomalies in estimated expression. The number of RNA-seq reads at each position along each transcript follows a distribution determined by the RNA-seq experiment protocol. We develop a method, Salmon Anomaly Detection (SAD), to identify the transcripts with an unexplainable coverage distribution by RNA-seq protocol. We observe that both quantification algorithm mistakes and incomplete reference transcripts cause abnormal coverage patterns. We also develop an adjustment procedure to correct quantification algorithm mistakes indicated by coverage anomalies and improve the accuracy of estimated expression. Our analysis of the coverage anomalies shows that some of the coverage anomalies are indicators of the regulation efficiency of transcription factors and can explain a part of the variability of the target gene expression. The developed methods introduce novel dimensions to more completely explain RNA-seq data, and can be incorporated into RNA-seq analyses to better characterize phenotype-transcript relationships or used to evaluate future transcript reconstruction methods.

Acknowledgments

My Ph.D. is not only a journey of doing research but also an exploration of what it means to do research. It is not grocery shopping that we get something where can be directly used in the dinner tonight. The day may not come when the research during my Ph.D. is useful in some way in making my dinner. It is not a lottery that we do not even know whether there will be something come out of it or not. Doing research is a unique journey to me: being a naughty child and doing whatever I am curious in or my "artistic" taste asks me to do, but still crying for attention and recognition that what I did is correct and important and hoping other people to actually use it some day. And also, try to fit my curiosities my "artistic" taste to the general audience's curiosities and taste when I do not get the candies.

So I really appreciate the help and guidance of my advisor, Carl Kingsford. He encouraged me to keep my own interest and helped me make the ideas better. In my opinion, he is an artist who is brave enough to stick to his artistic sense, in terms of both doing research and actual drawing. The SQUID logo is his masterpiece, and is one of my favorite drawings. His example is one of the inspirations for me to keep doing research and deciding research topics. I received valuable suggestions from him in re-directing research ideas, choosing and designing specific methods in my projects, as well as in terms of coding, presenting, and writing manuscripts. Without his help, I could not have finished the work that I have done.

I would like to thank my co-authors Mingfu Shao, Hongyu Zheng, Yutong Qiu, Han Xie, Adrian Lee, and Chelsea Chen. It is my great pleasure to collaborate with them. I would like to also acknowledge my other labmates, Guillaume Marçais, Dan DeBlasio, Heewook Lee, Hao Wang, Brad Solomon, Natalie Sauerwald, Daniel Bork, Laura Tung, Yihang Shen, Minh Hoang, and Mohsen Ferdosi. It was enjoyable to have random chat with you in the hall way or to have lunch or dinner together and drink beers. I appreciated your feedbacks every time I present my work in group meeting, and I also liked to hear your presentations when I was exposed to new and appealing areas.

I would like to thank Irene Kaplow for giving me advice on postdoc interviews and thank Jose Lugo-Martinez for teaching me Spanish words, and also thank Jenn Williams and my great officemate Michael Kleyman for having tea or coffee with me. I would like to thank Cathy Su, Dora Li, my roommate Fen Pei, and other CPCB students, for being so sweet.

I want to thank my committee members, Carl Kingsford, Russell Schwartz, Xinghua Lu, and Ben Raphael, for their service. The pandemic condition made it hard to have offline meetings. Thank you for attending my presentations and giving suggestions on my thesis work virtually.

Many people accompanied me and helped me during my Ph.D., especially my parents, my piano teacher Lili Cai, and my cat. Thank you for supporting me and having faith in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Anomaly detection

The problem of anomaly detection is to find data points or data patterns that are not generated by an expected process or do not follow an expected behavior. As summarized by Chandola et al. [20], this concept encompasses a variety of data-mining problems in many areas, such as detecting fraud activities in credit card transactions or detecting abnormal heart conditions using human electrocardiagrams. The computational methods for anomaly detection also span a wide range, including machine learning approaches that learn the separate distributions of normal and anomaly data points, rule-based approaches that allow users to define "normal" status, as well as statistical tests that evaluate the likelihood of data points falling in the normal distribution. Because of the versatility, anomaly detection is done with very different approaches in different areas of application and may incorporate specific domain knowledge.

In this dissertation, we introduce the concept of anomaly detection to RNA sequencing (RNA-seq) data. RNA-seq data captures the expressed transcript sequences and expression. A set of transcript sequences along with their expression can be inferred for each biological sample from its RNA-seq measurement. Unexpected RNA-seq patterns indicate the disagreement with the truly expressed transcripts and the inferred transcripts. We developed rule-based anomaly detection approaches to identify the unexpected RNA-seq patterns. Explaining the RNA-seq anomalies improves the inferred transcript sequences and expression.

1.2 Transcripts are indicators of cell function and phenotype

Transcripts are RNA seq molecules that copy subsequences from genes. They either pass the genetic information to proteins or carry the genetic information to perform their own functions. Various amounts of transcripts are synthesized in each cell, which is called expression of transcripts or abundances of transcripts. Transcript expression is related to the amounts of protein that they translate into and is also related to the efficiency in performing their own functions. It is an important task to measure the transcript sequences and expression.

The profile of transcript sequences and expression are key indicators of cell status. The expression of transcripts determine the cell type or tissue type by deciding the proteins that

can be further translated. Therefore, the change of expression profile during cell differentiation attracts many research interests [147, 154, 155]. B cells and T cells encode numerous types of antibodies by VDJ recombination [10, 14, 80]. The expressed transcripts determine the specific antibody each B cell and T cell generates. Transcripts of B cells or T cells are sequenced to reconstruct a repertoire of B cell or T cell receptors.

The transcript sequences and expression can also indicate the disease status of the cells. Many variants in the genome cause diseases through altering the expressed transcripts, which is an evidence to determine pathological variants [82, 125]. For example, copy number variants (CNVs) of oncogenes, specifically copy number gains, usually lead to overexpression of the corresponding transcripts and are associated with many cancer types [93]. Large scale sequence variants (SVs) in the genome are identified to generate fused transcript sequences and protein sequences, which may have altered protein function or abundances. This is another driving force for cancer [32, 104, 148, 152]. Transcript abundances are also affected by genomic single nucleotide variants (SNVs), which are also called expression quantitative trait loci (eQTLs) [26, 109]. Some of the eQTLs are associated with various diseases and other phenotypes according to GWAS studies [46].

It is challenging to accurately infer the transcript sequences and expression because of the large space of sequence and expression variation in transcripts. Transcription starts with the binding between RNA polymerase and promoter sequences in DNA [165] and ends with synthesized RNA (or RNA precursors) of variable abundances. Synthesized RNA or RNA precursors go through post-transcriptional processes such as splicing or polyadenylation and turn into mature RNA. More than half of the transcripts go through splicing and the number of alternatively spliced transcripts is large. Besides alternative splicing, variation in DNA sequence can affect transcript sequences and expression as well, such as genomic mutations and large-scale sequence variants. This makes the space of transcript sequences and expression even more complicated.

1.3 RNA sequencing technique

Next-generation RNA sequencing (or RNA-seq) captures the sequences and abundances of expressed transcripts via sampling short reads from them. It has been over a decade since RNA-seq was first applied [36, 87]. Many current RNA-seq experiments are now carried out on Illumina platform under paired-end sequencing protocol. We briefly summarize the paired-end protocol here since the features and distributions of sequencing reads are used in developing various computational methods regarding RNA-seq. This protocol first fragments RNA molecules extracted from cells, reverse transcribes the RNA fragments into cDNA fragments, PCR-amplifies the cDNA fragments, and finally selects the cDNA fragments around a certain lengths for sequencing. Illunima sequencing technique reads out the nucleotides one by one from the 5' end of both strands of the cDNA fragments, and finally generates a pair of short reads around 50 – 150 bp [144]. The steps in the protocol can be modeled by probabilistic distributions, which have been studied [12, 55, 84] and used in RNA-seq data simulators [39, 50].

RNA-seq has become widely available and the number of RNA-seq datasets has drastically increased in recent years (`https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/`). RNA-seq is broadly adopted in many data cohorts and consortia, such TCGA (`https://www.`

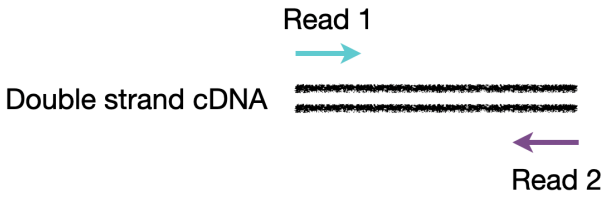

Figure 1.1: An example of an Illumina paired-end RNA-seq read. The two black lines represent double strand cDNA fragments. The fragment length follows the criteria of fragment length selection. A pair of reads will be sequenced from both strands reading from 5' end of cDNA towards 3' end. The length of each read is fixed and is usually shorter than the length of cDNA fragment. The blue and purple arrows represent the paired-end reads sequenced from the two strands of cDNA.

`cancer.gov/tcga`), GTEx [18] (`https://gtexportal.org/home/`), HuBMap [1], and 1000 Genomes Project [24].

Recent experimental breakthroughs have enabled sequencing RNA reads on single cells [56, 136, 150, 155] and incorporating spatial information [105, 164]. However, due to the smaller amount of RNA molecules in each single cell compared to a bulk of cells, the new techniques can only capture a subset of expressed transcripts. We focus on the traditional bulk RNA-seq data for detecting anomalies.

1.4 Current research area to reconstruct transcript information using RNA-seq data

Current methods to reconstruct transcripts usually focus on infer only one aspect between sequences and expression. The most common areas for developing reconstruction methods from RNA-seq data are introduced below.

1.4.1 Transcriptome assembly

Using RNA-seq data to reconstruct transcript sequences is called transcriptome assembly. Sequences are the most fundamental characterization of what composes of gene transcription product. There exist several curated or highly confident databases of transcript sequences, including RefSeq [112], Ensembl [174] or Gencode gene annotations [38]. The gene information of each transcripts is also annotated in these databases. However, these databases do not completely cover all possible transcripts expressed in all cells from all tissues of each individual. Transcriptome assembly methods are designed for inferring the transcript sequences of each RNA-seq sample in a sample-specific manner.

There are two general approaches to assembling transcript sequences: de novo transcriptome assembly and reference-based transcriptome assembly. De novo transcriptome assembly does not depend on a reference genome to which RNA-seq reads are aligned. This type of method follows the same route as genome assembly, and is usually based on constructing a de Bruijn graph

and traversing the paths on the graph as assembled transcripts. Trans-Abyss [130], Trinity [48], Oases [134], and SOAPdenovo-Trans [171] are widely used de novo transcriptome assemblers.

Reference-based transcriptome assembly depends on the alignment of RNA-seq reads to the genome, and identifies where the exon regions, intron regions, and splice junctions of each transcripts locate in the genome. Cufflinks [154] is one of the earliest reference-based transcriptome assemblers, which constructs a compatibility graph among RNA-seq reads to indicate whether the set of implied introns of each RNA-seq read is compatible with that of another RNA-seq read. Cufflinks then parsimoniously partitions the graph into the minimum number of chains, and each chain represents a transcript. Later methods usually represent the RNA-seq read alignment by using splice graphs, which is proposed by Heber et al. [57] for assembly using EST data. IsoLasso [85] and TransComb [89] select a set of graph paths as transcripts such that the consistency of coverages between different transcript regions or with observed RNA-seq read counts is optimized. StringTie [115] and Scallop [137] model the read counts on splice graphs as a network flow and adapt flow decomposition algorithms on the splice graphs. The parsimonious assumption is used in these methods: the number of transcript sequences should be minimal as long as the set of transcripts is able to generate the observed RNA-seq data. Though, different methods impose this assumption at different strictness levels.

These methods usually infer a coverage along with each assembled transcripts. The coverage is an estimate of the expression of transcripts. However, the coverage is derived in coverage consistency optimization or flow decomposition for the purpose of inferring transcript sequences. It is much less accurate compared to the expression estimated using expression quantification methods (Section 1.4.3), especially when sequencing biases are usually considered in expression quantification methods but not transcriptome assemblers.

The accuracy of transcriptome assembly is generally low, especially when compared with the accuracy of many machine learning prediction tasks, such as classifying handwritten digits in the MNIST dataset [75]. Even with the genome sequences, the state-of-art reference-based transcriptome assembly methods usually achieve less than 0.1 AUC when evaluated by the current gene annotation [31]. This is partially because short RNA-seq reads are not able to determine which distant regions co-exist in the same transcript. Long-read RNA-seq, especially full-length transcript sequencing techniques [19], provides a possible direction for improving the precision of reconstructing transcript sequences. But they cannot sensitively capture all expressed transcript sequences [156]. Combining the full-length transcript sequencing and short-read RNA-seq techniques for reconstructing transcript sequences is an approach under exploration [4, 47, 70].

1.4.2 Large-scale sequence variation detection

RNA-seq data is also used to identify large-scale transcript sequence alteration, which is also called transcriptomic structural variants (TSVs). Due to the large-scale disruption of transcript sequences, TSVs usually induce severe disruption in the abundance and function of RNA or translated proteins as well. They have been identified as one of the drivers of certain cancer types [104, 148] and have been used as diagnostic biomarkers [163]. TSVs are usually caused by genomic SVs, the large-scale sequence variation in the genome. Because not all genes are expressed and not all genomic SVs occur in gene regions, only a subset of genomic SVs lead to sequence change in transcripts. And because the controlling mechanism of transcription (for

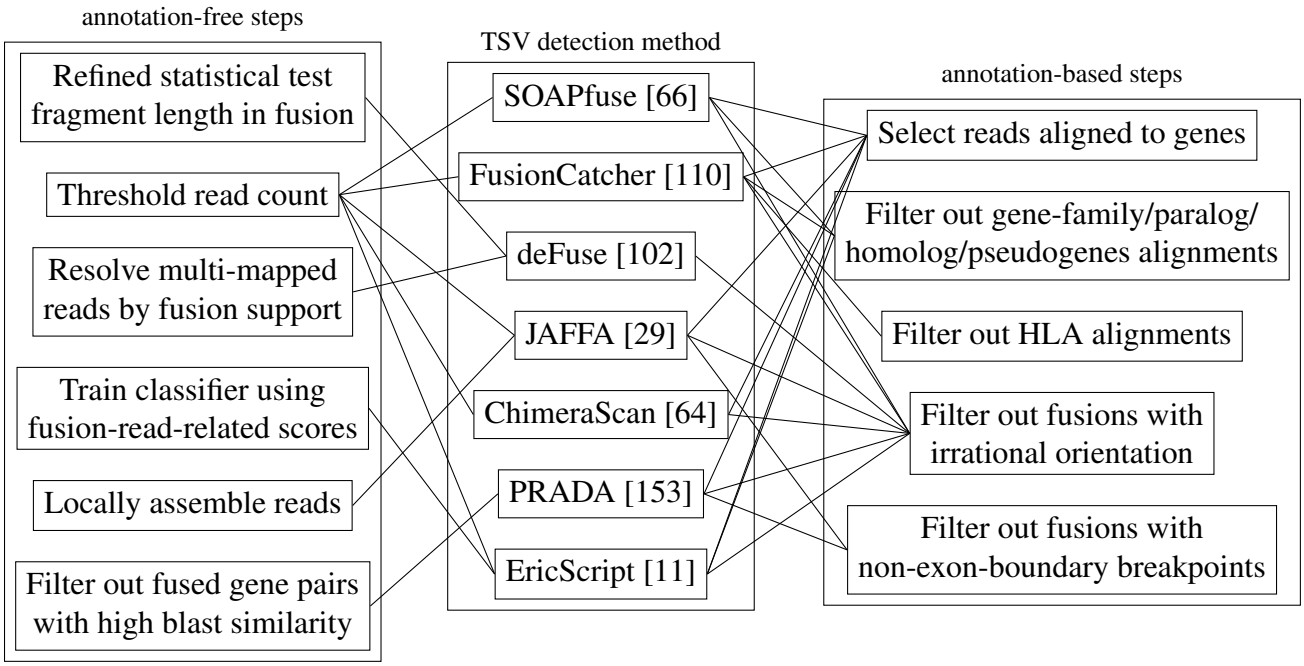

Figure 1.2: Previous TSV detection methods and features used in detection. Middle column is the name of TSV detection software. Left column is the gene-annotation-independent modules are used in TSV detection algorithms. Right column is the gene-annotation-based modules in TSV detection methods. Edges indicate a software contains the corresponding connected modules. These modules are summaries of key modules, and each software has its own additional well-calibrated steps to combine with these key modules.

example, transcript start/end and splice sites) is not fully captured and genomic SVs may break some of the elements, the altered transcript sequences are not exactly predictable from genomic SVs. Therefore, TSVs are usually detected using RNA-seq data.

Many previous methods are designed to detect the TSVs that fuse transcripts from two or more genes together, which are also called gene fusions or fusion genes. The principle of TSV detection is to identify RNA-seq reads that have a discordant alignment compared to how the sequence library is prepared. For example, when both of the paired-end reads are aligned to the same strand of genome or transcriptome, the alignment is discordant. The key challenge in TSV detection lies in distinguishing between TSV-caused discordant alignment and sequencing/alignment errors, which often occur due to alternative splicing as well as sequence similarity between different genes and regions. Different methods use different criteria for distinguishing TSVs and sequencing/alignment errors, which is summarized in Figure 1.2. Applying these criteria usually leads to an increased precision in identifying the true TSVs but also a decrease in sensitivity in the precision-sensitivity trade-off. Some fusion gene detection methods in the early years, for example SOAPfuse [66] and ChimeraScan [64], also include a module targeted for split-aligning RNA-seq reads to multiple regions, which is less used in later detection methods with the help of state-of-art RNA-seq aligners [34, 69]. Currently, there is still no standard pipeline or set of criteria to detect gene fusions with both high precision and high sensitivity

across all RNA-seq datasets. Ensemble approaches that combines the detections from multiple methods have been proposed [91]. Different analyses have different emphases on precision or sensitivity, and the different requirements for fusion gene detection also increase the difficulty for designing a uniformly well-performing fusion gene detection method.

Incomplete gene annotation also poses a challenge in TSV detection: current fusion gene detection methods rely on several criteria based on gene annotation (exon boundaries or orientation) as shown in Figure 1.2, and thus they are not able to identify TSVs that involve previously intergenic or non-transcribing regions. TSVs involving previously non-transcribing regions may be even harder to accurately identify, both because the lack of annotation for applying the annotation-based criteria and because the alignment error in intergenic regions tends to be higher due to the larger sequence search space compared to gene regions.

1.4.3 Expresssion quantification

Table 1.1: Probabilistic models and incorporated RNA-seq biases of expression quantification methods

Method	model parameter	distribution assumption	bias correction
Jiang and Wong [67]	reference transcript abundances	Poisson distribution of read count	None
RSEM [78]	reference/assembled transcript abundances	Multinomial distribution of each read with Dirichlet prior	read start position distribution
PSG [77]	splice graph edge weights	Multinomial distribution of each read with Dirichlet prior	None
eXpress [129]	reference transcript abundances	Multinomial distribution of each read with uniform prior	hexamer priming bias
pRSEM [90]	reference transcript abundances	Multinomial distribution of each read with grouped prior by ChIP-seq	read start position distribution
kallisto [15]	reference transcript abundances	Multinomial distribution of each read	hexamer priming bias
Salmon [113]	reference transcript abundances	Multinomial distribution of each read with Dirichlet prior	hexamer priming, GC, and positional bias

The task of expression quantification is to infer the expression of transcripts that are expressed in a given RNA-seq sample, and the set of transcripts is usually given as input. The estimated

expression of transcripts can be further used in many other analysis, such as differential expression analysis [95, 101, 118, 126], isoform switch detection [51, 86, 111, 161], inferring gene regulatory network [22, 33, 103, 159], and several prediction tasks such as predicting treatment responses [44, 128]. Soneson et al. [142] showed that estimating expression on the transcript level followed by merging transcript expression to genes leads to higher accuracy in expression estimates on the gene-level as well as detection of differentially expressed genes. This result emphasizes the necessity of expression quantification for each transcript.

Expression quantification is not as simple as aligning the RNA-seq reads and counting the number of aligned reads. Instead, a large proportion of RNA-seq reads are multi-mapped to many genes and transcripts because of the sequence similarity among genes and the sharing of exons among alternative splicing isoforms. The ambiguity of origins of the multi-mapped reads confuses the counting approach, since it is not clear which transcripts should absorb the count of a multi-mapped read. Current state-of-art expression quantification algorithms solve the ambiguity of origin problem by optimizing a probabilistic model to describe the probability of generating each sequencing read under transcript expression parameters. The core idea of the probabilistic model is proposed by Xing et al. [172], and first used in RNA-seq data by Jiang and Wong [67]. The expression quantification model by Jiang and Wong [67] assumes that there exists a set of sequences such that each RNA-seq read can be uniquely aligned to and each transcript is a concatenation of a subset of sequences. The ambiguous read counts of transcripts are used as parameters to maximize the probability of observing the unambiguous read counts of the sequences in the set under the Poisson distribution assumption. However, it is difficult to construct such a set of sequences. Further probabilistic models are proposed [15, 78, 113, 129] and they model the probability of generating an RNA-seq read from each transcript by a multinomial distribution and account for the alignment quality. Some steps in RNA-seq protocol introduce biases in the sequenced fragments and result in the phenomenon that the probability of sequencing a fragment is not totally determined by the expression of the sequenced region. Biases can be caused by hexamer priming, RNA degradation (positional bias), and uneven PCR amplification (GC bias). They have been modeled [12, 55, 84, 96] and incorporated in some of the expression quantification algorithms [15, 78, 113]. RNA polymerase II binding information also has been used to calibrate prior distribution of model parameters in expression quantification [90]. Table 1.1 summarizes some of the probabilistic models used in popular expression quantification methods.

One of the disadvantages of these expression quantification methods is that they depend on a set of transcript sequences as input. The current reference transcript annotation is incomplete. Using assembled transcripts from transcriptome assemblers partially solves this problem, but this approach has another disadvantage that transcript assemblers suffer from relatively low accuracy. RSEM [78] seeks to assemble transcript sequences and estimate expression of the assembled transcripts. Transcriptome assemblers also have recommended pipelines to combine reference transcripts and assembled transcripts for expression quantification task (See `https://github.com/Kingsford-Group/scallop` for an example). LeGault and Dewey [77] tackled the incomplete reference transcripts problem from a new angle by directly estimating node (exons) and edge (splice junctions) abundances in splice graphs. This approach, also called graph quantification, assumes that the splice graphs are accurate despite that the known transcripts are incomplete and that all paths in splice graphs can be a transcript. Nevertheless, it is still a challenge to quantify transcript expression under the case of incomplete

reference.

1.4.4 Single nucleotide variation (SNV) detection

There are several works that identify and analyze SNVs in transcripts and infer the haplotypes of alleles [16, 94, 119]. These SNVs and haplotypes are inherited from paternal or maternal genomes. These methods are less used when there exist whole genome sequencing (WGS) data, which is more widely used in SNV detection and haplotype phasing. But once the SNV information is available, it has been incorporated into allele-specific expression quantification [108, 123] to estimate the expression of transcripts corresponding to a specific allele.

1.4.5 The tasks of reconstructing transcript sequences and expression are closely related

Transcriptome assembly, TSV detection, and expression quantification are closely related to each other (Figure 1.3) and, in theory, can be incorporated into each other and improve the accuracy of each task. As discussed in the above description of individual tasks, if transcriptome assembly considers the detected TSVs, they will be able to reconstruct fused transcript sequences. If a more complete and accurate gene annotation can be reconstructed by transcript assembly, the annotation information can be used in TSV detection for more accurate detection of fusion genes. Both fused transcripts and assembled novel splicing variants could be considered in expression quantification to estimate the expression of novel sequences as well as improve the accuracy of expression estimation of known sequences. The refined read generation model in expression quantification is not used in transcriptome assembly or TSV detection, but more accurately modeling the read count may lead to more accurately assembled transcripts or detected TSVs.

However, it is very challenging to incorporate all possible sequences (including fused sequences, novel genes, and splicing variants) and expression into a unified model to describe the observed RNA-seq reads. There is an exponential number of paths in each splice graph, which represents connections between exons in alternative splicing, and the number of paths even increases if splice graphs of fused regions are constructed. Using an exponential number of transcript sequences in expression quantification probabilistic model may lead to an extremely large dimension in the parameter space and further lead to large parameter estimation errors due to the "curse of dimensionality".

1.5 Our contribution

Because there are sequences or sequence variants that are not completely reconstructed as well as expression abundances that are not accurately estimated, the observed RNA-seq dataset may not be fully compatible with the reconstructed transcripts. We focus on identifying the patterns in the observed RNA-seq dataset that are incompatible with or cannot be explained by the inferred transcript sequences and expression. These patterns are called anomalies, and identifying these patterns falls into the framework of anomaly detection.

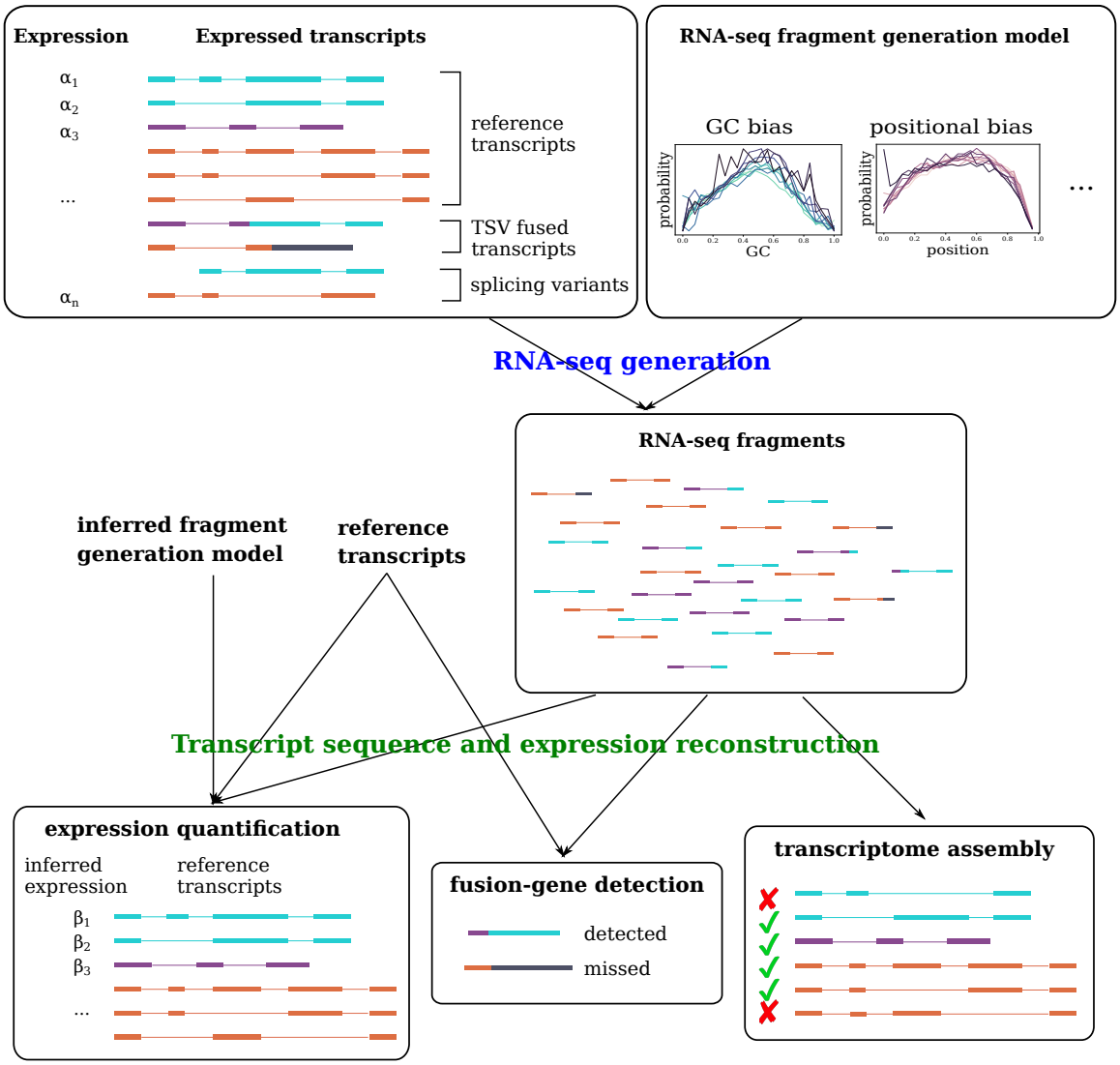

Figure 1.3: Summary for RNA-seq read generation and reconstruction. Cells express different types of transcript sequences with a certain amount of expression, including reference transcripts, TSV-fused transcripts, and novel sequences due to splicing variants. RNA-seq reads are randomly sampled from the expressed transcript sequences under the fragment generation model that describes hexamer priming bias, GC bias, position bias, and potentially other biases. With a given RNA-seq dataset, current computational methods reconstructs TSVs, alternative splicing isoforms, and estimate transcript expression separately under some idealized assumptions or model simplifications. Assuming all TSVs are fusions between a pair of known genes and simplify the fragment generation model to read support thresholds, the fusion gene subset of TSVs are detected (bottom middle). Simplifying the fragment generation model and only considering coverage similarity along each transcript, transcript sequences are reconstructed by transcriptome assembly methods (bottom right). Reference-based transcript assembly methods further assumes that the reference genome is known and genomic SVs are incorporated into the reference genome. Assuming the set of transcript sequences are known, expression quantification methods infer the fragment generation model and further infer the expression of given transcripts by maximizing the likelihood of generating the RNA-seq fragments (bottom left).

Anomaly detection has not been often used in RNA-seq, but it provides a fundamental view of improving transcript reconstruction using RNA-seq data. The current methods in TSV detection, transcriptome assembly, and expression quantification establish what RNA-seq reads or patterns we expect to observe. The unexplained RNA-seq reads or patterns locate the errors of inferred transcript sequences or expression on a fine scale. For example, if the discordant alignments concentrate on one transcript, the pattern indicates that the sequence of this transcript is likely erroneous but the sequences of other transcripts are likely correct. Knowing what the errors are and how they are made is fundamental for method improvements. Anomaly detection in RNA-seq can also be viewed as an evaluation of transcript sequences and expression: it gives assurance to further analyses that are based on the sequences and expression without anomaly patterns.

In this dissertation, we focus on two types of anomalous patterns in RNA-seq: unexplained discordant RNA-seq alignments, and unexplained coverages along each transcript. Seeking to explain the discordant RNA-seq alignments, we expanded the detection of large-scale sequence variants in transcripts, also called transcriptomic structural variants or TSVs. Identifying and explaining the unexpected coverage patterns leads to an improvement on the expression estimates for a subset of transcripts.

For the discordant RNA-seq reads, we developed a method to identify TSVs to explain them. Our detection includes both fusion-gene TSVs and a new type of TSVs, called non-fusion-gene TSVs, that involve previously non-transcribing regions into the sequence merge. We derived a novel problem formulation with specially designed pre-processing steps for distinguishing the discordant alignments that are caused by both types of TSVs and caused by sequencing or alignment errors. Our formulation reduces the dependence on gene annotation but emphasizes the features of graphs that are used to represent RNA-seq alignments. The core problem models RNA-seq alignments as graph edges between the segmented genomic regions and seeks to find a rearrangement of the genomic regions that explains the maximum number of RNA-seq reads. Unifying both concordant and discordant reads contributes to the detection accuracy, especially in the case where there is a lack of annotation information of non-transcribing sequences. We further extended the rearrangement formulation to allow multiple rearrangements to handle the allele heterogeneity scenario. The details of this aim are in Chapter 2.

For the unexplained coverage patterns, we developed a method to identify the transcripts for which the observed read coverage significantly violates the read generation model considering sequence, GC, and positional biases. The abnormal coverage patterns indicate quantification inaccuracy of the corresponding transcript. Focusing on the transcripts with the abnormal coverage pattern, we designed a procedure to re-quantify the subset of transcripts with better consistency between the re-quantified coverage and the read generation model. Using full-length transcript sequencing data, we observed that around 23%–32% of abnormal coverage patterns can be explained by the unannotated transcripts that are absent in the gene annotation. We further analyzed whether the abnormal coverage patterns can be used to predict the functional efficiency of transcription factors. We observed that the abnormal coverages can be used as functional indicators of certain transcription factors can explain a proportion of the expression variance of their regulated genes. See Chapter 3 for details.

Chapter 2

Detecting transcriptomic structural variations

Large-scale transcriptome sequence changes, or transcriptomic structural variants (TSVs), are usually caused by genomic structural variants (SVs) and are known to be associated with certain cancer types [104, 148]. The large-scale sequence changes are represented by the concatenation between two regions under a certain orientation, which is a result in various genomic SVs including deletion, duplication, inversion, and translocation. Sequence change of expressed transcripts may further lead to fused or broken protein sequences or altered protein abundances, thus leading to malfunction of the cell. For example, *BCR-ABL1* is a well-known fusion oncogene for chronic myeloid leukemia [32], and the *TMPRSS2-ERG* fusion product leads to over-expression of ERG and helps triggers prostate cancer [152]. TSVs are used as biomakers for early diagnosis or treatment targets [163]. Accurately identifying TSVs from RNA-seq data benefits such disease studies and biomaker developments.

Genomic SVs are typically detected from whole-genome sequencing (WGS) data and several computational methods are designed for this task [21, 62, 74, 122, 124, 179]. Genomic SVs are closely related to TSVs but the alterations in transcript sequences (such as splice junctions) cannot be determined exactly from genomic SVs. These WGS-based genomic SV detection methods share a fundamental principle with TSV detection from RNA-seq data, which is to identify aligned reads that are inconsistent with the sequencing library preparation. Such alignments are called discordant alignments. However, WGS and RNA-seq data have different features to take into account: RNA-seq is coupled by alternative splicing and expression of transcripts and thus the coverage variance is large; WGS has relatively consistent coverage along the whole genome and thus the breaks in genome and the concatenations need to be considered together.

When the TSV is a concatenation between two transcript sequences, the TSV is also known as a fusion gene. Fusion-gene detection is the focus of previous TSV detection studies [11, 29, 64, 66, 102, 110, 149, 153, 176]. Some of the fusion-gene detection methods depend on de novo transcript assembly [48, 130, 134, 171] followed by transcript-to-genome alignment [71, 169, 175]. In general, these methods rely heavily on current gene annotations: RNA-seq reads are filtered out if they are aligned in intergenic regions, pseudo-genes, paralog gene families, and so on.

However, not all discordant RNA-seq alignments support fusion-genes, and particularly, the

alignments may not be within gene regions. Fusion-gene TSVs are only a subset of TSVs because genomic SVs may not necessarily fuse a pair of genes, and the gene annotation is not guaranteed to be correct and complete. TSVs can also affect genes by causing a previously non-transcribed region to be incorporated into a gene, which we refer to as non-fusion-gene TSVs. This type of TSV can also alter downstream RNA and protein structure in a similar way as fusion-gene TSVs. We aim to explain the discordant RNA-seq reads that cannot be explained by fusion-genes by identifying the fusions involving intergenic or intronic sequences or attributing them to sequencing or alignment errors.

This chapter describes the problem formulation and solution to identify both fusion-gene and non-fusion-gene TSVs. The core TSV detection algorithm is a combinatorial problem of rearranging genome segments to maximally explain RNA-seq reads. The method was first introduced under the allele homogeneity assumption and we called it SQUID. SQUID was published in Genome Biology [98] and the code is available at `https://github.com/Kingsford-Group/squid`. It was a joint work with Mingfu Shao and Carl Kingsford. We explain the model and integer linear programming (ILP) solution in Section 2.1. We then relaxed the assumption and generalize the problem to handle heterogeneous alleles in Section 2.2. This work, accepted in the Workship on Algorithms in Bioinformatics (WABI) in 2019 and later published in Algorithms in Molecular Biology (AMB) [121], was joint with Yutong Qiu, Han Xie, and Carl Kingsford. The code is available at `https://github.com/Kingsford-Group/diploidsquid`.

2.1 SQUID: Transcriptomic structural variation detection from RNA-seq

We propose a method, SQUID, to predict TSVs from RNA-seq alignments to the genome (Figure 2.1 provides an overview). To do this, it seeks to rearrange the reference genome to make as many of the observed alignments consistent with the rearranged genome as possible. Formally, SQUID constructs a graph from the alignments where the nodes represent boundaries of genome segments and the edges represent adjacencies implied by the alignments. These edges represent both concordant and discordant alignments, where concordant alignments are those consistent with the reference genome and discordant alignments are those that are not. SQUID then uses a novel integer linear program (Section 2.2.5) to order and orient the vertices of the graph to make as many edges consistent as possible. Adjacencies that are present in this rearranged genome but not present in the original reference are proposed as predicted TSVs. The identification of concordant and discordant alignments (Section 2.1.3), construction of the genome segments (Section 2.1.4), creation of the graph, and the reordering objective function (Section 2.1.1) are described in the Methods section.

2.1.1 The computational problem: rearrangement of genome segments

We formulate the TSV detection problem as the optimization problem of rearranging genome segments to maximize the number of observed reads that are consistent (termed *concordant*)

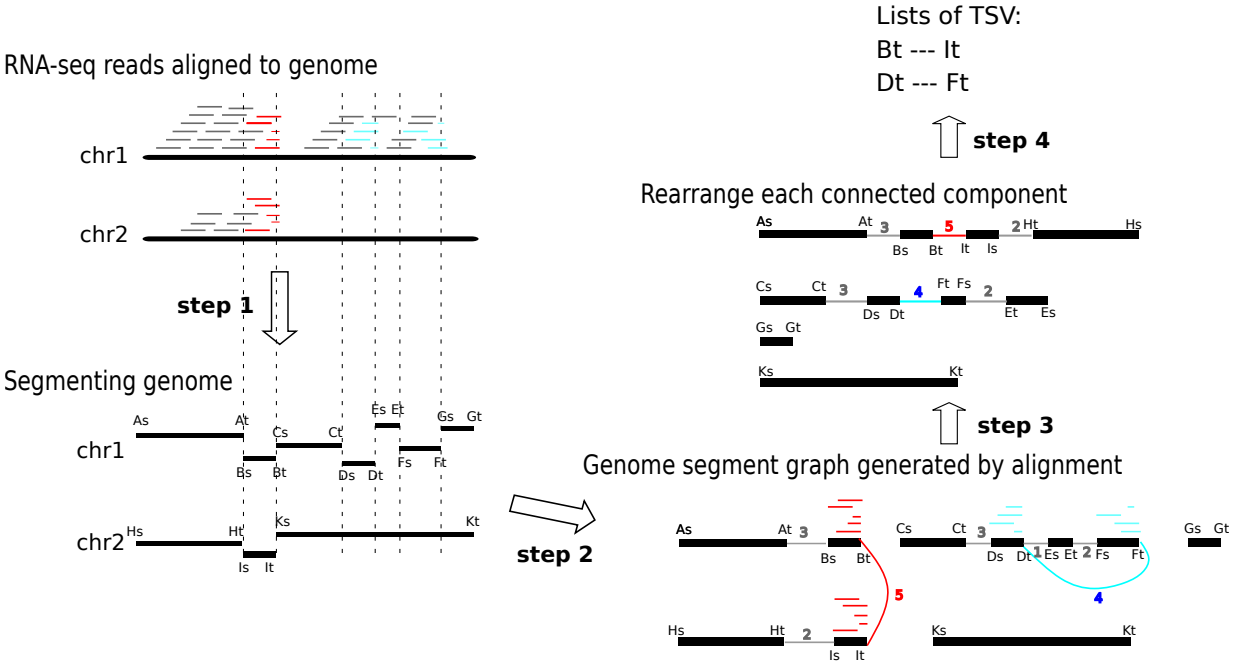

Figure 2.1: Overview of the SQUID algorithm. Based on the alignments of RNA-seq reads to the reference genome, (step 1) SQUID partitions the genome into segments, (step 2) connects the endpoints of the segments to indicate the actual adjacency in transcript, and finally (step 3) reorders the endpoints along the most reliable path. Thick black lines are genome sequences or segments. Grey, red and cyan short lines are read alignments, where grey represents concordant alignment, red and cyan represent discordant alignments of different candidate TSVs. Vertical dashed lines are the separation boundaries between genome segments, and the boundaries are derived based on read alignments. The heads of genome segments are denoted by As, Bs, etc., and the tails are denoted by At, Bt, etc. (step 2) Each read alignment generates one edge between segment endpoints. The edge is added in the following way: when traversing the genome segments along the edge to generate a new sequence, the read can be aligned concordantly onto the new sequence. Multi-edges are collapsed into one weighted edge, where the weight is the number of reads supporting that edge. Red and cyan edges correspond to different candidate TSVs. (step 3) Genome segments are reordered and reoriented to maximize the total number of concordant alignments (concordant edge weights) with respect to the new sequence. (step 4) Discordant edges that are concordant after rearrangement are output as TSVs (in this case, both red edges and cyan edges are output).

with the rearranged genome. This approach requires defining the genome segments that can be independently rearranged. It also requires defining which reads are consistent with a particular arrangement of the segments. We will encode both of these (segments and read consistency) within a *Genome Segment Graph* (GSG). See Figure 2.2 as an example.

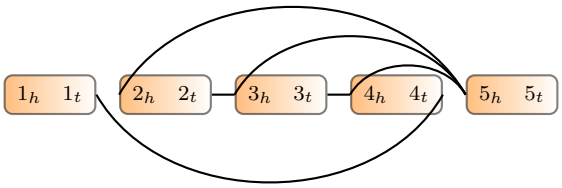

Figure 2.2: Example of genome segment graph. Boxes are genome segments, each of which has two ends subscripted by h and t. The color gradient indicates the orientation from head to tail. Edges connect ends of genome segments.

Definition 1 (Segment)**.** *A* segment *is a pair $s = (s_h, s_t)$, where s represents a continuous sequence in reference genome and s_h represents its head and s_t represents its tail in reference genome coordinates. In practice, segments will be derived from the read locations (Section 2.1.4).*

Definition 2 (Genome Segment Graph (GSG))**.** *A genome segment graph $G = (V, E, w)$ is an undirected weighted graph, where V contains both endpoints of each segment in a set of segments S, i.e., $V = \{s_h : s \in S\} \cup \{s_t : s \in S\}$. Thus, each vertex in the GSG represents a location in the genome. An edge $(u, v) \in E$ indicates that there is evidence that the location u is in fact adjacent to location v. The weight function, $w : E \longrightarrow \mathbb{R}^+$, represents the reliability of an edge. Generally speaking, the weight is the number of read alignments supporting the edge, but we allow a multiplier to calculate edge weight which will be discussed below. In practice, E and w will be derived from split-aligned and paired-end reads (Section 2.1.5).*

Defining vertices by endpoints of segments is required to avoid ambiguity. Only knowing that segment i is connected with segment j is not enough to recover the sequence, since different relative positions of i and j spell out different sequences. Instead, for example, an edge (i_t, j_h) indicates that the tail of segment i is connected head of segment j, and this specifies a unique desired local sequence with only another possibility of the reverse complement (i.e. it could be that the true sequence is $i \cdot j$ or $rev(j) \cdot rev(i)$; here $\cdot$ indicates concatenation and $rev(i)$ is the reverse complement of segment i).

The GSG is similar to the breakpoint graph [7] but with critical differences. A breakpoint graph has edges representing both connections in reference genome and in target genome. While edges in the GSG only represents the target genome, and they can be either concordant or discordant. In addition, the GSG does not require that the degree of every vertex is two, and thus alternative splicing and erroneous edges can exist in the GSG.

Our goal is to reorder and reorient the segments in S so that as many edges in G are compatible with the rearranged genome as possible.

Definition 3 (Permutation)**.** *A permutation π on a set of segments S projects a segment in S to a set of integers from 1 to $|S|$ (the size of S) representing the indices of the segments in an ordering*

of S. In other words, each permutation π defines a new order of segments in S.

Definition 4 (Orientation Function)**.** *An orientation function f maps both ends of segments to 0 or 1:*

$$f : \{s_h : s \in S\} \cup \{s_t : s \in S\} \longrightarrow \{0, 1\}$$

subject to $f(s_h) + f(s_t) = 1$ for all $s = (s_h, s_t) \in S$. An orientation function specifies the orientations of all segments in S. Specifically, $f(s_h) = 1$ means s_h goes first and s_t next, corresponding to the forward strand of the segment, and $f(s_t) = 1$ corresponds to the reverse strand of the segment.

With a permutation π and an orientation function f, the exact and unique sequence of genome is determined. The reference genome also corresponds to a permutation and an orientation function, where the permutation is the identity permutation, and the orientation function maps all s_h to 1 and all s_t to 0.

Definition 5 (Edge Compatibility)**.** *Given a set of segments S, a genome segment graph $G = (V, E, w)$, a permutation π on S, and an orientation function f, an edge $e = (u_i, v_j) \in E$, where $u_i \in \{u_h, u_t\}$ and $v_j \in \{v_h, v_t\}$, is* compatible *with permutation π and orientation f if and only if*

$$1 - f(v_j) = \mathbf{1}[\pi(v) < \pi(u)] = f(u_i) \tag{2.1}$$

where $\mathbf{1}[x]$ is the indicator function that is 1 if x is true and 0 otherwise. Comparison between permuted elements is defined as comparing their index in permutation, that is, $\pi(v) < \pi(u)$ states that segment v is in front of segment u in rearrangement π. We write $e \sim (\pi, f)$ if e is compatible with π and f.

The above two edge compatibility equations (2.1) require that, in order for an edge to be compatible with the rearranged and reoriented sequence determined by π and f, the edge needs to connect the right side of the segment in front to the left side of segment following it. As we will see in Section 2.1.5, edges of GSG are derived from reads alignments. An edge being compatible with π and f is essentially equivalent to the statement that the corresponding read alignments are concordant (Section 2.1.3) with respect to the target genome determined by π and f. When (π, f) is clear, we refer to edges that are compatible as concordant edges, and edges that are incompatible as discordant edges.

With the above definitions, we formulate an optimization problem as follows:

Problem 1. ***Input:*** *A set of segments S and a GSG $G = (V, E, w)$.*
Output: *Permutation π on S and orientation function f that maximizes:*

$$\max_{\pi, f} \sum_{e \in E} w(e) \cdot \mathbf{1}[e \sim (\pi, f)] \tag{2.2}$$

This objective function tries to find a rearrangement of genome segments (π, f), such that when aligning reads to the rearranged sequence, as many reads as possible will be aligned concordantly. This objective function includes both concordant alignments and discordant alignments and sets them in competition, which will be effective in reducing false positives when tumor transcripts out-number normal transcripts. There is the possibility that some rearranged tumor transcripts are out-numbered by normal counterparts. In order to be able to detect TSV in this case, depending on the setting, we may weight discordant read alignments more than concordant

read alignments. Specifically, for each discordant edge e, we multiply the weight $w(e)$ by a constant α, which represents our estimate of the ratio of normal transcripts over tumor counterparts.

The final TSVs are modeled as pairs of breakpoints. Denote the permutation and orientation corresponding to an optimally rearranged genome as (π^*, f^*) and those that correspond to reference genome as (π_0, f_0). An edge e can be predicted as a TSV if $e \sim (\pi^*, f^*)$ and $e \nsim (\pi_0, f_0)$.

2.1.2 Integer linear programming formulation

We use integer linear programming (ILP) to compute an optimal solution (π^*, f^*) of Problem 1. To do this, we introduce the following boolean variables:

- x_e: $x_e = 1$ if edge $e \sim (\pi^*, f^*)$, and $x_e = 0$ if not.
- z_{uv}: $z_{uv} = 1$ if segment u is before v in the permutation π^*, and 0 otherwise.
- y_u: $y_u = 1$ if $f^*(u_h) = 1$ for segment u.

With this representation, the objective function can be rewritten as

$$\max_{x_e, y_u, z_{uv}} w(e) \cdot x_e \tag{2.3}$$

We add constraints to the ILP derived from edge compatibility equations (2.1). Without loss of generality, we first suppose segment u is in front of v in the reference genome, and edge e connects u_t and v_h (which is a tail-head connection). Plugging in u_t, the first equation in (2.1) is equivalent to $1 - \mathbf{1}[\pi(u) > \pi(v)] = 1 - f(u_t)$ and can be rewritten as $\mathbf{1}[\pi(u) < \pi(v)] = f(u_h) = y_u$. Note that $\mathbf{1}[\pi(u) < \pi(v)]$ has the same meaning as z_{uv}; it leads to the constraint $z_{uv} = y_u$. Similarly, the second equation in (2.1) indicates $z_{uv} = y_v$. Therefore, x_e can only reach 1 when $y_u = y_v = z_{uv}$. This is equivalent to the inequalities (2.4) below. Analogously, we can write constraints for other three types of edge connections: tail-tail connections impose inequalities (2.5); head-head connections impose inequalities (2.6); head-tail connections impose inequalities (2.7):

$$\begin{aligned} x_e &\le y_u - y_v + 1 \\ x_e &\le y_v - y_u + 1 \\ x_e &\le y_u - z_{uv} + 1 \\ x_e &\le z_{uv} - y_u + 1 \end{aligned} \tag{2.4}$$

$$\begin{aligned} x_e &\le y_u - (1 - y_v) + 1 \\ x_e &\le (1 - y_v) - y_u + 1 \\ x_e &\le y_u - z_{uv} + 1 \\ x_e &\le z_{uv} - y_u + 1 \end{aligned} \tag{2.5}$$

$$\begin{aligned} x_e &\le (1 - y_u) - y_v + 1 \\ x_e &\le y_v - (1 - y_u) + 1 \\ x_e &\le (1 - y_u) - z_{uv} + 1 \\ x_e &\le z_{uv} - (1 - y_u) + 1 \end{aligned} \tag{2.6}$$

$$\begin{aligned} x_e &\le (1 - y_u) - (1 - y_v) + 1 \\ x_e &\le (1 - y_v) - (1 - y_u) + 1 \\ x_e &\le (1 - y_u) - z_{uv} + 1 \\ x_e &\le z_{uv} - (1 - y_u) + 1 \end{aligned} \tag{2.7}$$

We also add constraints to enforce that z_{uv} forms a valid topological ordering. For each pair of nodes u and v, one must be in front of other, that is $z_{uv} + z_{vu} = 1$. In addition, for each triple of nodes, u, v and w, they cannot be all in front of another; one must be at the beginning of these three and one must be at the end. Therefore we add $1 \le z_{uv} + z_{vw} + z_{wu} \le 2$.

Solving an ILP in theory takes exponential time, but in practice, solving the above ILP to rearrange genome segments is very efficient. The key is that we can solve for each connected component separately. Because the objective maximizes the sum of compatible edge weights, the best rearrangement of one connected component is independent from the rearrangement of another because by definition there are no edges between connected components.

2.1.3 Concordant and discordant alignments

Discordant alignments are alignments of reads that contradict library preparation in sequencing. Here the alignments are with respect to the genome instead of transcriptome. Aligning RNA-seq reads to both known transcribing sequences and non-transcribing sequences allows the detection of the non-fusion-gene TSVs. Concordant alignments are alignments of reads that agree with the library preparation. Take Illumina sequencing as an example. In order for a paired-end read alignment to be concordant, one end should be aligned to the forward strand and the other to the reverse strand, and the forward strand aligning position should be in front of the reverse strand aligning position (Figure 2.3a). Concordant alignment traditionally used in whole genome sequencing (WGS) also requires that a read cannot be split and aligned to different locations. But these requirements are invalid in RNA-seq alignments because alignments of reads can be separated by an intron with unknown length.

We define concordance criteria separately for split-alignment and paired-end alignment. If one end of a paired-end read is split into several parts and each part is aligned to a location, the end has split-alignments. Denote the vector of the split alignments of an end to be $R = [A_1, A_2, \cdots, A_r]$ (r depends on the number of splits). Each alignment $R[i] = A_i$ is comprised of 4 components: chromosome (Chr), alignment starting position (Spos), alignment ending position, and orientation (Ori, with value either $+$ or $-$). We require that the alignments A_i are sorted by their position in the read. A split-aligned end $R = [A_1, A_2, \cdots, A_r]$ is concordant if all the following conditions hold:

$$\begin{aligned} A_i.Chr &= A_j.Chr && \forall i, \forall j \\ A_i.Ori &= A_j.Ori && \forall i, \forall j \\ A_i.Spos &< A_j.Spos && \text{if } A_i.Ori = + \text{ for all } i < j \\ A_i.Spos &> A_j.Spos && \text{if } A_i.Ori = - \text{ for all } i < j \end{aligned} \tag{2.8}$$

If the end is not split, but continuously aligned, the alignment automatically satisfies equation (2.8). Denote the alignments of R's mate as $M = [B_1, B_2, \cdots, B_m]$. An alignment of the paired-end read is concordant if the following conditions all hold:

$$\begin{aligned} A_i.Chr &= B_j.Chr && \forall i, \forall j \\ A_i.Ori &\neq B_j.Ori && \forall i, \forall j \\ A_1.Spos &< B_m.Spos && \text{if } A_1.Ori = + \\ A_m.Spos &> B_1.Spos && \text{if } A_1.Ori = - \end{aligned} \tag{2.9}$$

We only require the left-most split of the forward read R be in front of the left-most split of the reverse read M since the two ends in a read pair may overlap. In order for a paired-end read to be

concordant, each end should satisfy split-read alignment concordance (2.8), and the pair should satisfy paired-end alignment concordance (2.9).

2.1.4 Splitting the genome into segments S

We use a set of breakpoints to partition the genome. The set of breakpoints contains two types of positions: (1) the start position and end position of each interval of overlapping discordant alignments, (2) an arbitrary position in each 0-coverage region.

Ideally, both ends of a discordant read should be located in separate segments, otherwise, a discordant read contained in a single segment will always be discordant no matter how the segments are rearranged. Assuming discordant read alignments of each TSV pile up around the breakpoints and do not overlap with discordant alignments of other TSVs, we set a breakpoint on the start and end positions of each contiguous interval of overlapping discordant alignments.

For each segment that contains discordant read alignments, it may also contain concordant alignments that connect the segment to its adjacent segments. To avoid having all segments in GSG connected to their adjacent segments and thus creating one big connected component, we pick the starting point of each 0-coverage region as a breakpoint. By adding those breakpoints, different genes will be in separate connected components unless some discordant reads support their connection. Overall, the size of each connected component is not very large: the number of nodes generated by each gene is approximately the number of exons located in them and these gene subgraphs are connected only when there is a potential TSV between them.

2.1.5 Defining edges and filtering obvious false positives

In a GSG, an edge is added between two vertices when there are reads supporting the connection. For each read spanning different segments, we build an edge such that when traversing the segments along the edge, the read is concordant with the new sequence (equations (2.8) and (2.9)). Examples of deriving an edge from a read alignment are given in Figure 2.3. In this way, concordance of an alignment and compatibility of an edge with respect to a genome sequence are equivalent.

The weight of a concordant edge is the number of read alignments supporting the connection, while the weight of a discordant edge is the number of supporting alignments multiplied by discordant edge weight coefficient α. The discordant edge weight coefficient α represents the normal/tumor cell ratio (for full table of SQUID parameters, see Appendix Table 2.2). In the case where normal transcripts dominate tumor transcripts, α enlarges the discordant edge weights, and helps to satisfy the discordant edges in the rearrangement of ILP.

We filter out obvious false positive edges in order to reduce both the ILP computation time and the mistakes after the ILP. Edges with very low read support are likely to be a result of alignment error, therefore we filter out edges with weight lower than a threshold θ. Segments with too many connections to other regions are likely to have low mappability, so we also filter out segments connecting to more than γ other segments. The parameters α, θ, and γ are the most important user-defined parameters to SQUID (Appendix Table 2.2, Appendix Figure 2.8, Appendix Figure 2.9). An interleaving structure of exons from different regions (different genes)

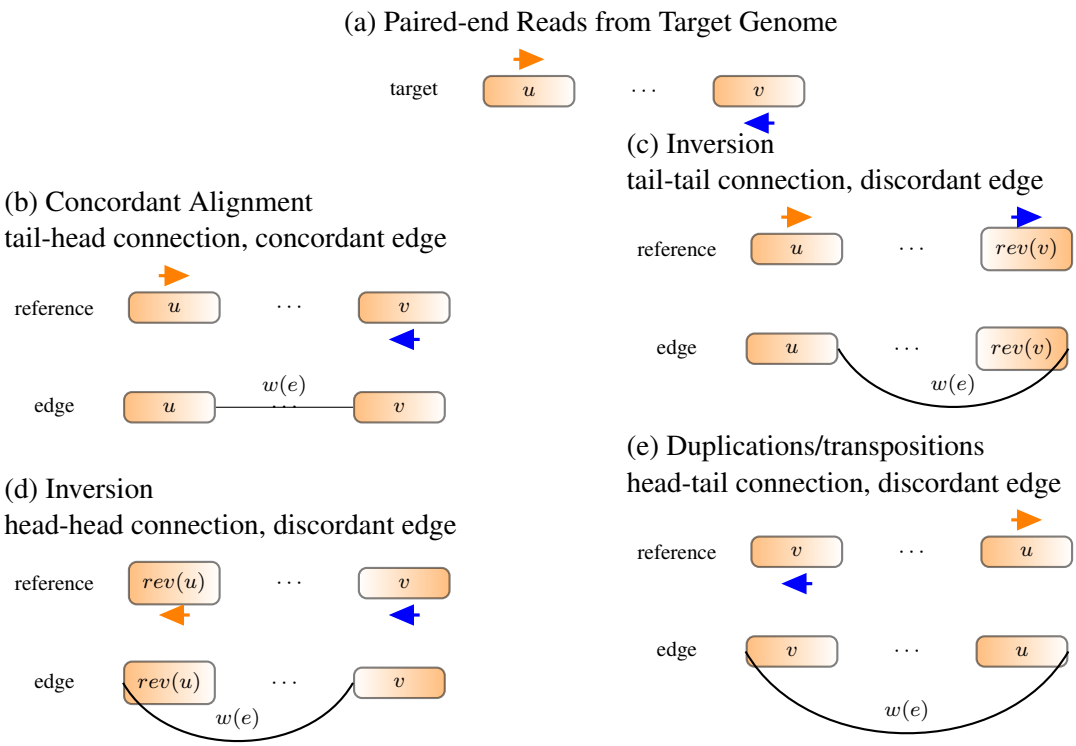

Figure 2.3: Constructing edges from alignment. (A) Read positions and orientations generated from the target genome. (B) If the reference genome does not have rearrangements, the read should be concordantly aligned to reference genome. An edge is added to connect the right end of u to the left end of v. Traversing the two segments along the edge reads out $u \cdot v$, which is the same as reference. (C) Both ends of the read align to forward strand. An edge is added to connect the right end of u to the right end of $rev(v)$. Traversing the segments along the edge reads out sequence $u \cdot rev(rev(v)) = u \cdot v$, which recovers the target sequence and the read can be concordantly aligned to. (D) If both ends align to the reverse strand, an edge is added to connect the left end of front segment to the left end of back segment. (E) If two ends of a read point out of each other, an edge is added to connect the left end of front segment to the right end of back segment.

seems more likely to be a result of sequencing or alignment error rather than structural variation. Thus, we filter the interleaving edges between two such groups of segments.

2.1.6 Identifying TSV breakpoint locations

Edges that are discordant in the reference genome indicate potential rearrangements in transcripts. Among those edges, some are compatible with the permutation and orientation from ILP. These edges are taken to be the predicted TSVs. For each edge that is discordant initially but compatible with the optimal rearrangement found by the ILP, we examine the discordant read alignments to determine the exact breakpoint located within related segments. Specifically, for each end of a discordant alignment, if there are 2 other read alignments that start or end in the same position and support the same edge, then the end of the discordant alignment is predicted to be the exact TSV breakpoint. Otherwise, the boundary of the corresponding segment will be output as the exact TSV breakpoint.

2.1.7 Results: SQUID is accurate on simulated data

We validate the accuracy of the above proposed TSV detection method, SQUID, using both simulated RNA-seq datasets and real-world RNA-seq datasets. The section is structured as follows: We first describe the procedure of RNA-seq data simulation, and then show the comparison between SQUID and SV detection methods (originally designed for whole genome sequencing data) based on the simulated RNA-seq data. We further evaluate the accuracy of SQUID on previous studied breast cancer cell lines and compare with previous fusion-gene detection methods. Finally we apply SQUID on the RNA-seq samples of four cancer types in The Cancer Genome Atlas (TCGA, `https://cancergenome.nih.gov`), and describe the summary of detected TSVs across the four cancer types as well as explain example TSVs related to known cancer genes.

Simulation methodology

Simulations with randomly added structural variations and simulated RNA-seq reads were used to evaluate SQUID's performance in situations with a known correct answer. RSVsim [9] was used to simulate SV on the human genome (Ensembl 87 or hg38) [174]. We use the 5 longest chromosomes for simulation (chromosome 1 to chromosome 5). RSVsim introduces 5 different types of SVs: deletion, inversion, insertion, duplication, and inter-chromosomal translocation. To vary the complexity of the resulting inference problem, we simulated genomes with 200 SVs of each type, 500 SVs of each type, and 800 SVs of each type. We generated 4 replicates for each level of SV complexity (200, 500, 800). For inter-chromosomal translocations, we only simulate 2 events because only 5 chromosomes were used.

In the simulated genome with SVs, the original gene annotations are not applicable, and we cannot simulate gene expression from the rearranged genome. Therefore, for testing purposes, we interchange the role of the reference (hg38) and rearranged genome, and use the new genome as the reference genome for alignment, and hg38 with the original annotated gene positions as

the target genome for sequencing. Flux Simulator [50] was used to simulate RNA-seq reads from the hg38 genome using the Ensembl annotation version 87 [2].

After simulating SVs on genome, we need to transform the SVs into a set of TSVs, because not all SVs affect transcriptome, and thus not all SVs can be detected by RNA-seq. To derive the list of TSVs, we compare the positions of simulated SVs with the gene annotation. If a gene is affected by an SV, some adjacent nucleotides in the corresponding transcript may be located far part in the RSVsim-generated genome. The adjacent nucleotides can be consecutive nucleotides inside an exon if the breakpoint breaks the exon, or the end points of two adjacent exons if the breakpoint hits the intron. So for each SV that hits a gene, we find the pair of nucleotides that are adjacent in transcript and separated by the breakpoints, and convert them into the coordinates of the RSVsim-generated genome, thus deriving the TSV.

We compare SQUID to the pipeline of de novo transcriptome assembly and transcript-to-genome alignment. We also use the same set of simulations to test whether existing WGS-based SV detection methods can be directly applied to RNA-seq data. For the de novo transcriptome assembly and transcript-to-genome alignment pipeline, we use all combinations of the existing software Trinity [48], Trans-ABySS [130], GMAP [169] and MUMmer3 [71]. For WGS-based SV detection methods, we test LUMPY [74] and DELLY2 [124]. We test both STAR [34] and SpeedSeq [23] (which is based on BWA-MEM [81]) to align RNA-seq reads to the genome. LUMPY is only compatible with SpeedSeq output, so we do not test it with STAR alignments.

Validating TSV detection of SQUID in simulated data

Overall, SQUID's predictions of TSVs are far more precise than other approaches at similar sensitivity on simulated data (Section 2.1.7). SQUID achieves 60% to 80% percent precision and about 50% percent sensitivity on simulation data (Figure 2.4A, 2.4B). SQUID's precision is $> 20\%$ higher than several de novo transcriptome assembly and transcript-to-genome alignment pipelines (for details see Section 2.1.12), and the precision of WGS-based SV detection methods on RNA-seq data is even lower. The sensitivity of SQUID is similar to de novo assembly with MUMmer3 [71], but a little lower than DELLY2 [124] and LUMPY [74] with SpeedSeq [23] aligner. The overall sensitivity is not as high as precision, which is probably because there are not enough supporting reads aligned correctly to some TSV breakpoints. The fact that assembly and WGS-based SV detection methods achieve similar sensitivity corroborates the hypothesis that it is the data limiting the achievable sensitivity.

We test SQUID's robustness to various parameter choices of SQUID itself (Section 2.1.5, Appendix Table 2.2). SQUID is robust against different values of the segment degree threshold (Appendix Figure 2.8A – B), which filters edges from segments that are connected to too many other segements. Another parameter, the edge weight threshold, is equivalent to the read support threshold in other structural variation detection software. It controls the precision-sensitivity tradeoff (Appendix Figure 2.8C – D). The discordant edge weight coefficient, which up-weights initially discordant reads to compensate for heterogenous mixtures, does not affect precision or sensitivity in simulation data because simulated reads are homogeneous, and there is no need to adjust for normal/tumor cell ratio (Appendix Figure 2.8E – F).

We also test the robustness of SQUID against different RNA-seq experimental settings. Specifically, we simulate RNA-seq data with read length 51 bp, 76 bp and 100 bp combined

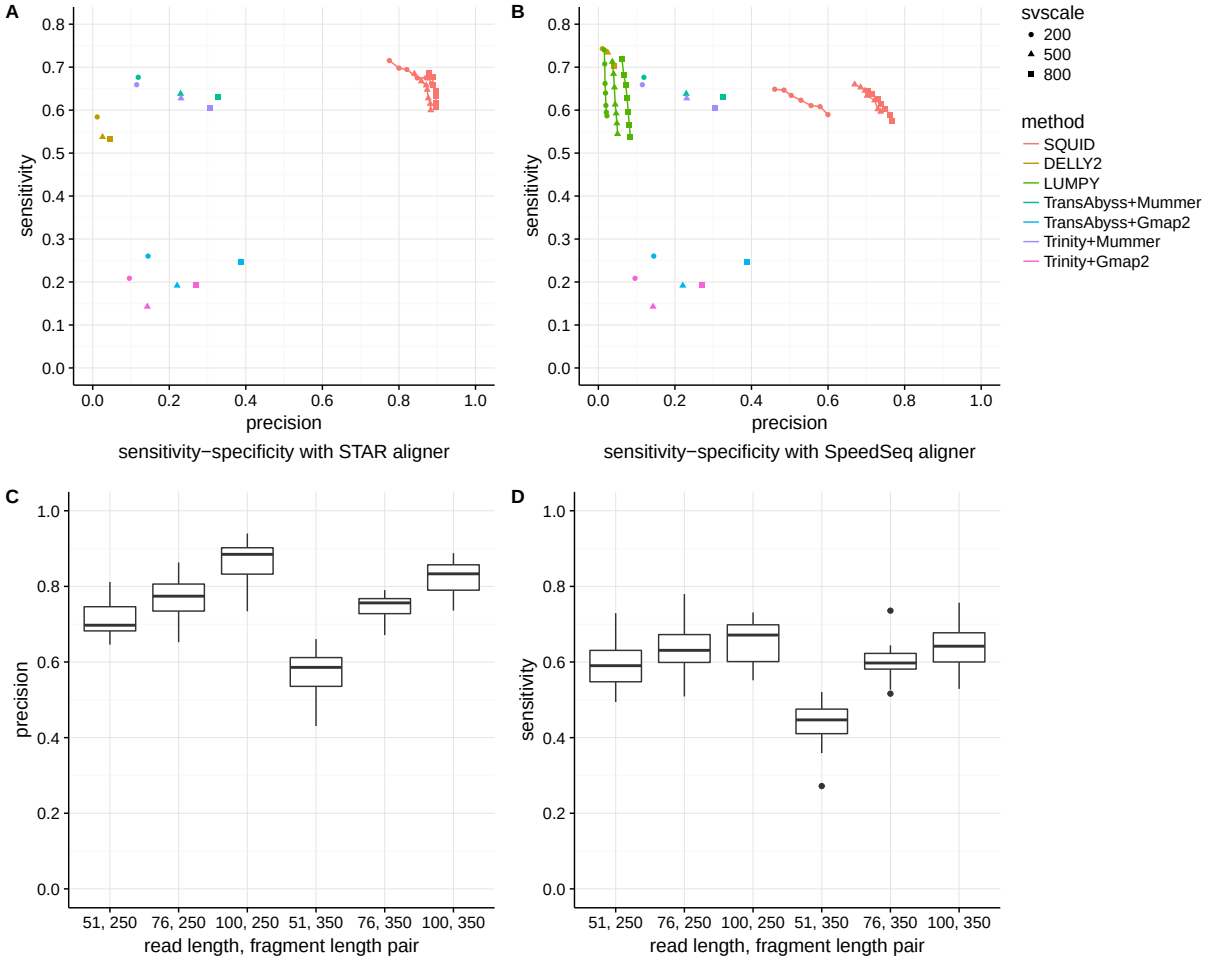

Figure 2.4: Performance of SQUID and other methods on simulation data. (A, B) Different number of SVs (200, 500, 800 SVs) are simulated in each dataset. Each simulated read is aligned with both (A) STAR and (B) SpeedSeq aligner. If the method allows for user-defined minimum read support for prediction, we vary the threshold from 3 to 9, and plot a sensitivity-precision curve (SQUID and LUMPY), otherwise it is shown as a single point. (C, D) Performance of SQUID under different RNA-seq experimental parameter combinations (read length of 51bp, 76bp and 100bp combined with fragment length of 250 and 350). Longer read length increases both precision and sensitivity of SQUID. Longer fragment length slightly decreases SQUID's performance. Short read length with long fragment length leads to the worst precision and sensitivity.

with fragment length 250 bp and 350 bp (Figure 2.4C, 2.4D). Each experimental setting has 4 replicates. With increased read length, SQUID in general performs better in both precision and sensitivity (although there are a few exceptions where randomness of simulation shadows the benefit from longer read length). However, with increased fragment length, SQUID performs slightly worse. In this case, there are fewer reads aligned at the exact breakpoint, possibly due to an increase in split-alignment difficulty for aligners. Short read length (51 bp) with long fragment length (350 bp) leads to the worst precision and sensitivity.

The low precision of the pipeline- and WGS-based methods (Figure 2.4) shows neither of these types of approaches are suitable for TSV detection from RNA-seq data. WGS-based SV detection methods are able to detect TSV signals, but not able to filter out false positives. Assembly-based approaches require solving the transcriptome assembly problem which is a harder and more time-consuming problem, and thus errors are more easily introduced. Further, the performance of assembly pipelines depends heavily on the choice of software — for example, MUMmer3 [71] is better at discordantly aligning transcripts than GMAP [169]. Dissect [175] is another transcript-to-genome alignment method that is designed for the case where SVs exist. (Unfortunately, Dissect did not run to completion on the some of the dataset tested here.) It is possible that different combinations of de novo transcript assembly and transcript-to-genome alignment tools can improve the accuracy of the pipelines, but optimizing the pipeline is out of scope of this work.

SQUID's effectiveness is likely due to its unified model of both concordant reads and discordant reads. Coverage in RNA-seq alignment is generally proportional to the expression level of the transcript, and using one read count threshold for TSV evidence is not appropriate. Instead, the ILP in SQUID puts concordant and discordant alignments into competition and selects the winner as the most reliable TSVs.

2.1.8 Results: SQUID is able to detect non-fusion-gene TSV on two previously studied cell lines

Fusion gene events are a strict subset of TSVs where the two breakpoints are each within a gene region and the fused sequence corresponds to the sense strand of both genes. Fusion genes thus exclude TSV events where a gene region is fused with a intergenic region or an anti-sense strand of another gene. Nevertheless, fusion genes have been implicated (likely because of available methods to detect them) in playing a role in cancer.

To probe SQUID's ability to detect both fusion-gene and non-fusion-gene TSVs from real data, we use two cell lines, HCC1954 and HCC1395, both of which are tumor epithelial cells derived from breast. Previous studies have experimentally validated predicted SVs and fusion gene events for these two cell lines. Specifically, we compile results from Bignell et al. [13], Galante et al. [40], Stephens et al. [146], Zhao et al. [177] and Robinson et al. [131] for HCC1954, and results from Stephens et al. [146] and Zhang et al. [176] for HCC1395. After removing short deletions and overlapping structural variations among different studies, we have 326 validated structural variations for the HCC1954 cell line, of which 245 of them have at least one breakpoint outside a gene region, and the rest (81) have both breakpoints within gene region; we have 256 validated true structural variations for the HCC1395 cell line, of which 94 have at least one

breakpoint outside a gene region, while the rest (162) have both breakpoints within gene. For a predicted structural variation to be true positive, both predicted breakpoints should be within a window of 30kb of true breakpoints and the predicted orientation should agree with the true orientation. We use a relatively large window since the true breakpoints can be located within an intron or other non-transcribed region, while the observed breakpoint from RNA-seq reads will be at a nearby coding or expressed region.

We use publicly available RNA-seq data from the NIH Sequencing Read Archive (SRA accessions: SRR2532344 and SRR925710 for HCC1954, SRR2532336 for HCC1395). Because the data are from a pool of experiments, the sample from which RNA-seq was collected may be different from those used for experimental validation. We align reads to the reference genome using STAR. We compare the result with the top fusion-gene detection tools evaluated in Liu et al. [91] and newer software not evaluated by Liu et al. [91], specifically, SOAPfuse [66], deFuse [102], FusionCatcher [110], JAFFA [29] and INTEGRATE [176]. In addition, we compare to the same pipeline of de novo transcriptome assembly and transcript-to-genome alignment as in previous section (also see Section 2.1.12). Trans-ABySS [130] is chosen for the de novo transcriptome assembly, and MUMMER3 [71] is chosen for transcript-to-genome alignment, because this combination has the best performance in simulation data. Table 2.1 summarizes the total number of predicted TSVs, and the number of TSVs corresponding to previously validated TSVs (hits). The full TSV predictions by SQUID on the two cell lines can be downloaded from `https://doi.org/10.5281/zenodo.4048493`.

After aligning the RNA-seq reads to the genome, the number of chimeric alignments are around twice the number of chimeric alignments when aligned to the transcriptome (2.13 times for HCC1954 sample, and 1.79 times for HCC1395 sample). Thus, many more chimeric RNA-seq alignments are included in TSV detection by considering the non-transcribing regions. Despite that not all chimeric alignments are caused by TSVs, including more chimeric alignments potentially expand the TSVs that can be detected.

When restricted to fusion gene events, SQUID achieves similar precision and sensitivity compared to fusion gene detection tools (Figure 2.5A). These methods have different rankings on the two cell lines. There is no uniformly best method for fusion-gene TSV predictions on both cell lines. SQUID ranks as one of the highest precision and second-to-the-highest sensitivity on HCC1954 cell line and ranks in the middle on the HCC1395 cell line. On both cell lines, the pipeline of de novo transcriptome assembly and transcript-to-genome alignment has very low precision, which suggests that without filtering steps assembly-based methods are not able to distinguish between noise and true TSVs.

It is even harder to predict non-fusion-gene TSVs accurately, since current annotations cannot be used to limit the search space for potential read alignments or TSV events. Only SQUID, deFuse, and the pipeline of de novo transcriptome assembly and transcript-to-genome alignment are able to detect non-fusion-gene events. SQUID has both a higher precision and a higher sensitivity compared to deFuse (Figure 2.5B). The assembly pipeline has a higher sensitivity but very low precision, which again indicates that this pipeline outputs non-fusion-gene TSV signals without distinguishing them from noise. A considerable proportion of validated TSVs are non-fusion-gene TSVs: correctly predicted non-fusion-gene TSVs make up almost half of all correct predictions of SQUID (Figure 2.5C).

We test the robustness of SQUID with respect to different parameter values on the two cell

Table 2.1: Summary of TSV predictions on HCC1954 and HCC1395 cell lines.

Method		SQUID	FusionCatcher	JAFFA	deFuse
HCC1954	fusion-gene predictions	46	54	67	95
	fusion-gene hits	7	5	4	12
	non-fusion-gene predictions	46	0	0	83
	non-fusion-gene hits	7	0	0	5
HCC1395	fusion-gene predictions	44	42	44	110
	fusion-gene hits	11	11	16	15
	non-fusion-gene predictions	57	0	0	121
	non-fusion-gene hits	9	0	0	7

Method		INTEGRATE	SOAPfuse	Pipeline
HCC1954	fusion-gene predictions	67	177	2118
	fusion-gene hits	10	5	4
	non-fusion-gene predictions	0	0	1080
	non-fusion-gene hits	0	0	11
HCC1395	fusion-gene predictions	61	185	2413
	fusion-gene hits	16	19	23
	non-fusion-gene predictions	0	0	1185
	non-fusion-gene hits	0	0	8

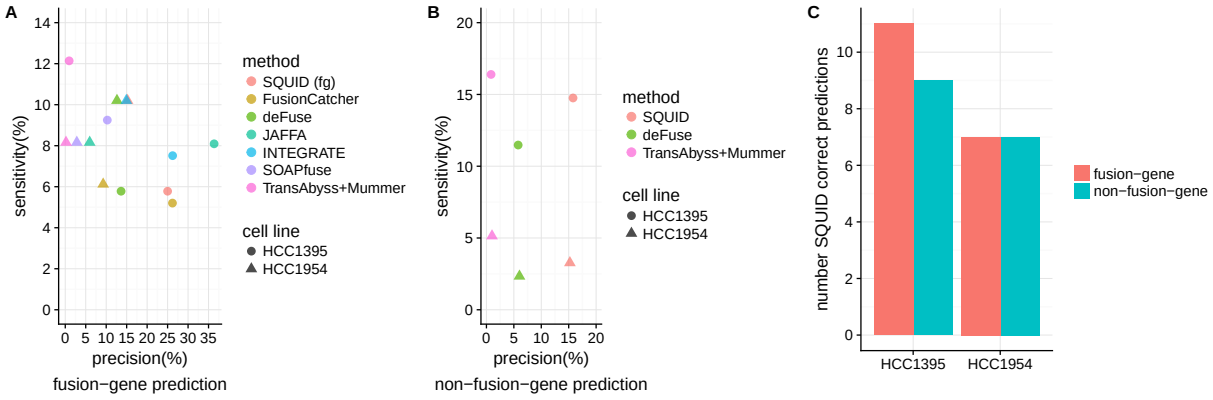

Figure 2.5: Performance of SQUID and fusion gene detection methods on breast cancer cell lines HCC1954 and HCC1395. Predictions are evaluated by previously validated SVs and fusions. (A) Fusion-gene prediction sensitivity-precision curve of different methods. (B) Non-fusion-gene prediction sensitivity-precision curve. Only SQUID, deFuse, and the pipeline of de novo transcriptome assembly and transcript-to-genome alignment are able to predict non-fusion-gene TSVs. (C) Number of correctly predicted fusion-gene TSVs and non-fusion-gene TSVs from SQUID. Non-fusion-gene TSVs makes up a considerable proportion of all TSVs.

lines (Appendix Figure 2.9). We find the same trend regarding the segment degree threshold and the edge weight threshold as with simulated data: the segment degree threshold does not affect either precision or sensitivity much, and the edge weight threshold determines the precision-sensitivity tradeoff. The discordant edge weight coefficient does not change the sensitivity on the HCC1954 cell line, possibly indicating the sequencing data is relatively homogeneous. As this parameter increases, precision on HCC1954 cell line slightly decreases because more TSVs are predicted. In contrast, increase of discordant edge weight coefficient increases both precision and sensitivity of HCC1395 cell line. This implies that for some transcripts, normal reads dominate tumor reads, and increasing this parameter allows us to identify those TSVs.

The sensitivity on both cell lines of all tested methods are relatively low. One explanation for this is the difference between the source of the data used for prediction and validation. In the ground truth, some SVs were first identified using WGS data or BAC end sequencing and then validated experimentally. Not all genes are expressed in the RNA-seq data used here, and lowly expressed genes may not generate reads spanning SV breakpoints due to read sampling randomness. To quantify the feasibility of each SV being detected, we count the number of supporting chimeric reads in RNA-seq alignments. The proportion of ground-truth fusion-gene TSVs with supporting reads is very low for both cell lines: 26.5% for HCC1954 (13 out of 49), and 27.1% for HCC1395 (47 out of 173). The maximum sensitivity of fusion-gene TSV prediction is limited by these numbers, which explains the relatively low sensitivity we observed. For non-fusion-gene TSVs, only 13.0% in HCC1954 (36 out of 277) and 21.7% in HCC1395 (13 out of 83) can possibly be identified.

We use WGS data of the corresponding cell lines to validate the novel TSVs predicted by SQUID (SRA accession number: ERP000265 for HCC1954, SRR892417 and SRR892296 for HCC1395). For each TSV prediction, we extract a 25Kb sequence around both breakpoints and concatenate them according to the predicted TSV orientation. We then map the WGS reads against these junction sequences using SpeedSeq [23]. If a paired-end WGS read can only be mapped concordantly to a junction sequence but not to the reference genome, that paired-end read is marked as supporting the TSV. If at least 3 WGS reads support a TSV, the TSV is considered as validated. With this approach, we are able to validate 40 more TSV predictions in HCC1395 cell line, and 18 more TSV predictions in HCC1954 cell line. In total, the percentage of predicted TSVs that can be validated either by previous studies or by WGS data is 57.7% for HCC1395 cell line and 35.2% for HCC1954 cell line. The WGS validation rate of HCC1954 cell line is much lower than HCC1395 cell line, which can be explained by the relatively low read depth: the read depth for HCC1954 WGS data is 7.6x, and that for HCC1395 WGS data is 22.7x.

2.1.9 Results: characterizing TSVs on four types of TCGA cancer samples

To compare the distributions and characteristics of TSVs among cancer types and between TSV types, we applied SQUID on arbitrarily selected 99 to 101 tumor samples from TCGA for each of four cancer types: breast invasive carcinoma (BRCA), bladder urothelial carcinoma (BLCA), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD). TCGA aliquot barcodes of corresponding samples are can be downloaded from `https://doi.org/10.5281/zenodo.4048493`. For data processing details see Section 2.1.12. Running time of SQUID is less than 3 hours for the majority of the RNA-seq data we selected, and the maximum

memory usage is around 4GB or 8GB (Appendix Figure 2.10).

To estimate the accuracy of SQUID's prediction on selected TCGA samples, we use WGS data of the same patients to validate TSV junctions. There are in total 72 WGS experiments available for the 400 samples (20 BLCA, 10 BRCA, 31 LUAD, 11 PRAD). We use WGS in the same approach to validate SQUID predictions as in the previous section. SQUID's overall validation rate is 88.21%, and this indicates that SQUID is quite accurate and reliable on TCGA data.

We find that most samples have $\approx$ 18–23 TSVs including $\approx$ 3–4 non-fusion-gene TSVs among all four cancer types (Figure 2.6A,B). BRCA has a larger tail of the distribution of TSV counts, where more samples contain a larger number of TSVs. The same trend is observed when restricted to non-fusion-gene TSVs.

Inter-chromosomal TSVs are more prevalent than intra-chromosomal TSVs for all cancer types (Figure 2.6C), although this difference is much more pronounced in bladder and prostate cancer. Non-fusion-gene TSVs are more likely to be intra-chromosomal events than fusion gene TSVs (Figure 2.6D), and in fact in breast and lung cancer, we detect more intra-chromosomal non-fusion-gene TSVs than inter-chromosomal non-fusion-gene TSVs. Prostate cancer is an exception in that, for non-fusion-gene TSVs, inter-chromosomal events are observed much more often than intra-chromosomal events. Nevertheless, it also holds true that non-fusion-gene TSVs are more likely to be intra-chromosomal than fusion-gene TSVs, because the percentage of intra-chromosomal TSVs within non-fusion-gene TSVs is higher than that within all TSVs. We do not know why fusion-gene and non-fusion-gene TSVs have different propensities to be intra- or inter-chromosomal events. Potential hypotheses include that different genomic SVs (such as inversions and translocations) have different intra- and inter-chromosomal occurrences and they induce the two types of TSVs with different rates. And differences in the ability to align RNA-seq reads within one chromosome and between different chromosomes, within transcribing sequences and outside transcribing sequences, may be another potential explanation. Understanding this phenomenon requires further analyses.

For a large proportion of breakpoints occurring multiple times within a cancer type, their partner in the TSV is likely to be fixed and to reoccur every time that breakpoint is used. To quantify this, for each breakpoint that occurred ≥ 3 times, we compute the entropy of its partner promiscuity. Specifically, we derive a discrete, empirical probability distribution of partners for each breakpoint and compute the entropy of this distribution. This measure thus represents the uncertainty of the partner given one breakpoint, with higher entropy corresponding to a less conserved partnering pattern. In Figure 2.6E, we see that there there is a high peak near 0 for all cancer types, which indicates that for a large proportion of recurring breakpoints, we are certain about its rejoined partner once we know the breakpoint. However, there are also promiscuous breakpoints with entropy larger than 0.5.

2.1.10 Results: tumor suppressor genes can undergo TSV and generate altered transcripts

Tumor suppressor genes (TSG) protect cells from becoming cancer cells. Usually their functions involve inhibiting cell cycle, facilitating apoptosis, and so on [138]. Mutations in TSGs may

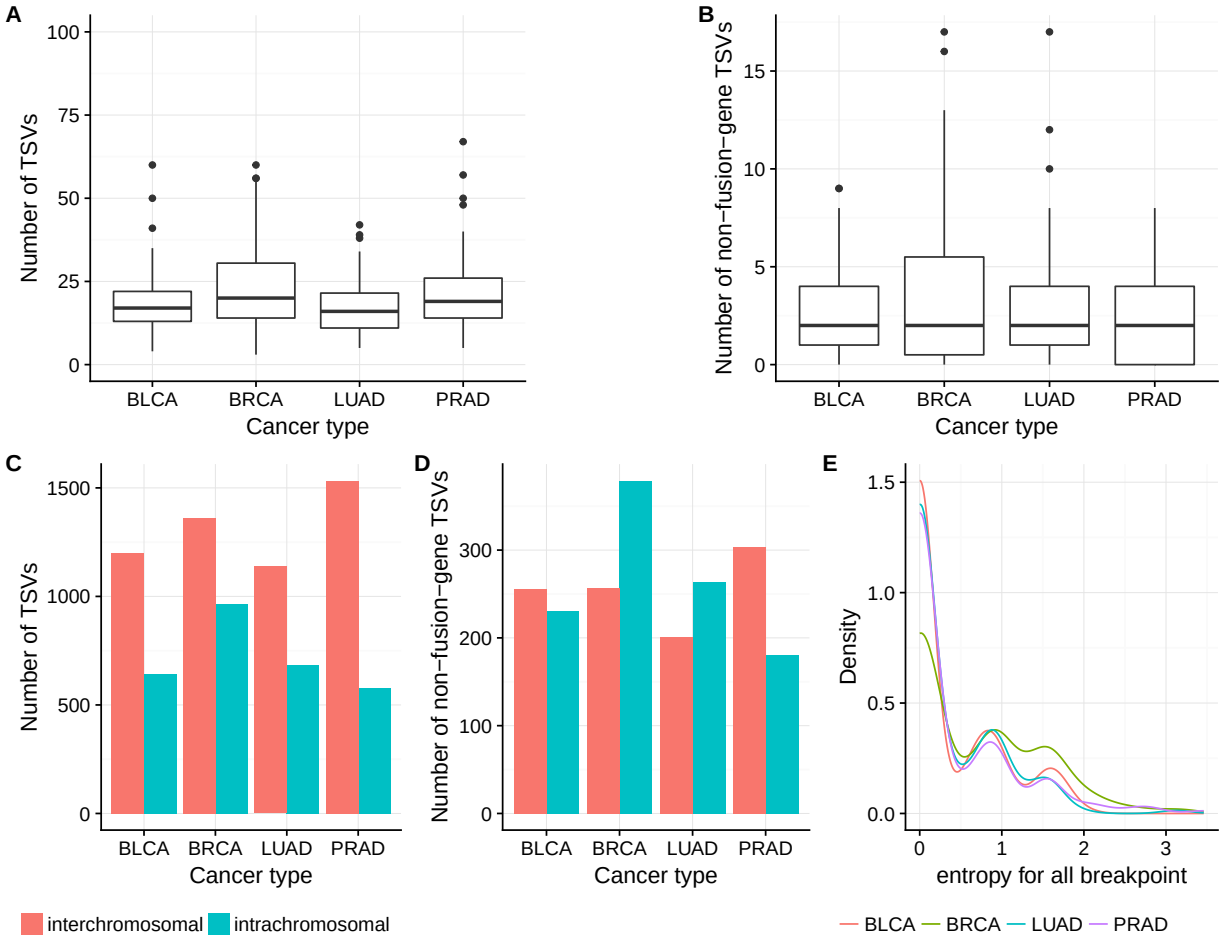

Figure 2.6: (A,B) Number of TSVs and non-fusion-gene TSVs in each sample in different cancer types. BRCA has slightly more samples with larger number of (non-fusion-gene) TSVs, thus showing a longer tail on y axis. (C,D) Number of inter-chromosomal and intra-chromosomal TSVs within all TSVs and within non-fusion-gene TSVs. Non-fusion-gene TSVs contain more intra-chromosomal events than fusion-gene TSVs. (E) For breakpoints occurring more than 3 times in the same cancer type, the distribution of the entropy of its TSV partner. The lower the entropy, the more likely the breakpoint has a fixed partner. The peak near 0 indicates a large portion of breakpoints are likely to be rejoined with the same partner in TSV. However, there are still some breakpoints that have multiple rejoined partners.

lead to loss of function of the corresponding proteins and benefit tumor growth. For example, homozygous loss-of-function mutation in p53 is found in about half of cancer samples across various cancer types [61]. TSVs are likely to cause loss of function of TSGs as well. Indeed, we observe several TSGs that are affected by TSVs, both of the fusion-gene type and the non-fusion-gene type.

The *ZFHX3* gene encodes a transcription factor that transactivates cyclin-dependent kinase inhibitor 1A (aka *CDKN1A*), a cell cycle inhibitor [99]. We find that in one BLCA and one BRCA sample, there are TSVs affecting *ZFHX3*. These two TSVs events are different from each other in terms of the breakpoint partner outside of *ZFHX3*. In the BLCA tumor sample, a intergenic region is inserted after the third exon of *ZFHX3* (See Figure 2.7A. For visualization by Integrative Genomics Viewer (IGV) [151], see Appendix Figure 2.11). The fused transcript stops at the inserted region, causing the *ZFHX3* transcript to lose the rest of its exons. In the BRCA tumor sample, a region of the anti-sense strand of gene *MYLK3* is inserted after the third exon of *ZFHX3* gene (Figure 2.7B, Appendix Figure 2.12). Because codons and splicing sites are not preserved on the anti-sense strand, the transcribed insertion region does not correspond to known exons of *MYLK3* gene, but covers the range of first exon of *MYLK3* and extends to the first intron and 5' intergenic region. Transcription stops within the inserted region, and causes the *ZFHX3* transcript to lose exons after exon 3, which resembles the fusion with intergenic region in BLCA sample.

Another example is given by the *ASXL1* gene, which is essential for activating *CDKN2B* to inhibit tumorgenesis [170]. We observe two distinct TSVs related to *ASXL1* from BLCA and BRCA samples. The first TSV merges the first 11 exons and half of exon 12 of *ASXL1* with a intergenic region on chromosome 4 (Figure 2.7C, Appendix Figure 2.13). Transcription stops at the inserted intergenic region, leaving the rest of exon 12 not transcribed. The breakpoint within the *ASXL1* is before the 3' UTR, so the downstream protein sequence from exon 12 will be affected. The other TSV involving *ASXL1* is a typical fusion-gene TSV where the first three exons of *ASXL1* are fused with the last three exons from the *PDRG1* gene (Figure 2.7D, Appendix Figure 2.14). Protein domains after *ASXL1* exon 4 and before *PDRG1* exon 2 are lost in the fused transcript.

These non-fusion-gene examples are novel predicted TSV events that are not typically detectable via traditional fusion-gene detection methods using RNA-seq data. They suggest that non-fusion-gene events can also be involved in tumorgenesis by causing disruption of tumor suppressor genes.

2.1.11 Discussion

We developed SQUID, the first algorithm for accurate and comprehensive TSV detection that targets both traditional fusion-gene detection and the much broader class of general TSVs. SQUID exhibits higher precision at similar sensitivities compared with WGS-based SV detection methods and pipelines of de novo transcriptome assembly and transcript-to-genome alignment. In addition, it has the ability to detect non-fusion-gene TSVs with similarly high accuracy.

We use SQUID to predict TSVs in TCGA tumor samples. From our prediction, BRCA has a slightly flatter distribution of number of per-sample TSVs than the other cancer types studied. We observe that non-fusion-gene TSVs are more likely to be intra-chromosomal events than

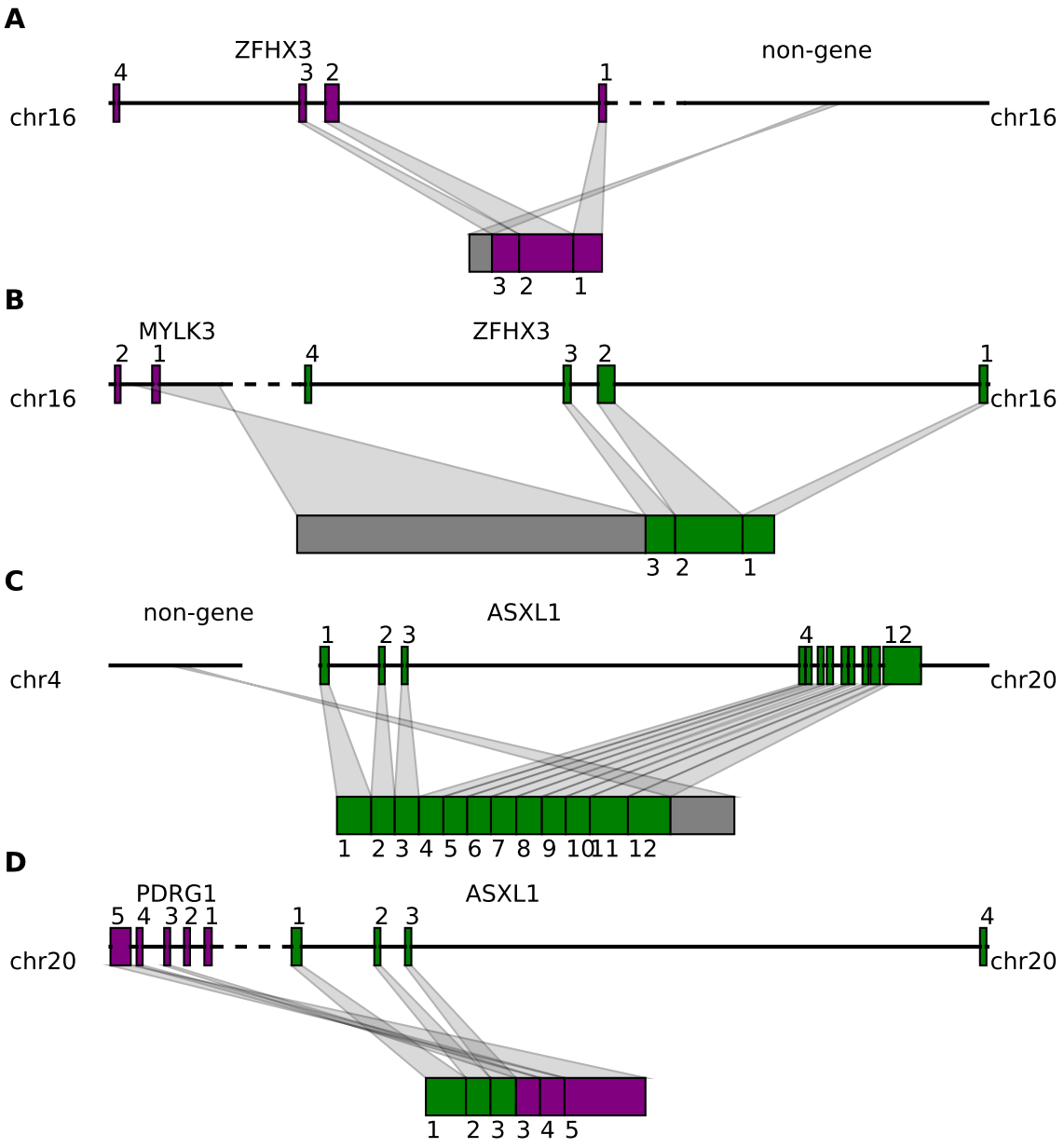

Figure 2.7: Tumor suppressor genes are affected by both fusion-gene and non-fusion-gene TSVs and generate transcripts with various features. (A) *ZFHX3* is fused with a intergenic region after exon 3. The transcript stops at the inserted region, losing the rest of exons. (B) *ZFHX3* is fused with a part of *MYLK3* anti-sense strand after exon 3. Codon and splicing signals are not preserved on anti-sense strand, thus *MYLK3* anti-sense insertion acts the same as intergenic region insertion, and causes transcription stop before reaching the rest of *ZFHX3* exons. (C) *ASXL1* is fused with an intergenic region in the middle of exon 12. The resulting transcript contains a truncated *ASXL1* exon 12 and intergenic sequence. (D) The first 3 exons of *ASXL1* gene are joined with last 3 exons of *PDRG1*, resulting in a fused transcript containing 6 complete exons from both *ASXL1* and *PDRG1*.

fusion-gene TSVs. This is likely due to the different sequence composition features in gene vs. non-gene regions. PRAD also stands out because the percentage of inter-chromosomal TSVs is the largest. Overall, these findings continue to suggest that different cancer types have different preferred patterns of TSVs, although the question remains whether these differences will hold up as more samples are analyzed and whether the different patterns are causal, correlated, or mostly due to non-functional randomness. These findings await experimental validation.

By applying SQUID on TCGA RNA-seq data, we are able to detect TSVs in cancer samples, especially non-fusion-gene TSVs. We identify novel non-fusion-gene TSVs involving known tumor suppressor genes *ZFHX3* and *ASXL1*. Both fusion-gene and non-fusion-gene events detected in TCGA samples are computational predictions and need further experimental validation.

Other important uses and implications for general TSVs have yet to be explored and represent possible directions for future work. TSVs will impact the accuracy of transcriptome assembly and expression quantification, and methodological advancements are needed to correct those downstream analyses for the effect of TSVs. For example, current reference-based transcriptome assemblers are not able to assemble from different chromosomes to handle the case of inter-chromosomal TSVs. In addition, expression levels of TSV-affected transcripts cannot be quantified if they are not present in the transcript database. Incorporating TSVs into transcriptome assembly and expression quantification can potentially improve their accuracy. SQUID's ability to provide a new genome sequence that is as consistent as possible with the observed reads will facilitate its use as a pre-processing step for transcriptome assembly and expression quantification, though optimizing this pipeline remains a task for future work.

Several natural directions exist for extending SQUID. First, SQUID is not able to predict small deletions, instead, it treats the small deletions the same as introns. This is to some extent a limitation of using RNA-seq data: introns and deletions are difficult to distinguish, as both result in concordant split reads or stretched mate pairs. The use of gene annotations could somewhat address this problem. Second, when the RNA-seq reads are derived from a highly heterogeneous sample, SQUID is likely not able to predict all TSVs occurring in the same region if they are conflicting since it seeks a single, consistent genome model. Instead, SQUID will only pick the dominating one that is compatible with other predicted TSVs. One approach to handle this would be to iteratively re-run SQUID, removing reads that are explained at each step. Again, this represents an attractive avenue for future work.

Another future direction is to analyze the possible multiple optimal solutions problem of the genome segment rearrangement problem. There are usually multiple rearrangements of the genome segments that lead to the maximum sum of explained edge weights. This is partially because the transcript sequence information from RNA-seq is not able to uniquely identify the order of segments in the genome. For example, it is not possible to know the order of a pair of mutually exclusive exons in the genome using only transcript sequences. But the set of concordant edges may be common across multiple rearrangements, and the multiple optimal rearrangements may lead to the same set of TSV prediction. However, it remains to be further analyzed whether, in practice, the sets of predicted TSVs from multiple optimal rearrangements are the same as or different from each other. It also remains to be investigated whether other information can be incorporated into the SQUID to alleviate the multiple optimal solutions problem and to select a more accurate set of TSVs among the optimal rearrangements.

SQUID currently leaves out the alignment quality of RNA-seq reads for simplification. In-

corporating the alignment quality is an potential direction for further improving the detection accuracy of SQUID. It may also alleviate the multiple optimal solutions problem. They can be incorporated into the definition of edge weights: instead of defining edge weights as the sum of reads that support the edge, they can be defined as the sum of scores corresponding to alignment quality of the supporting reads. The score can be designed to take into account the sequencing error rate, the fragment length distribution for paired-end alignments, and the breakpoint consistency for split alignments.

With the emerging long read RNA-seq or full-length transcript sequencing techniques, extending SQUID to handle long read RNA-seq data is a valuable future direction. The long sequencing length will benefit TSV detection in various aspects: the co-occurrence relationship (also called phasing relationship) among TSVs in different alleles can be revealed, and the full transcript sequences involved in fusions can be more accurately retrieved along with the breakpoint junctions. While long read sequencing techniques have their unique types of sequencing errors and may suffer from low sensitivity in capturing transcripts with certain transcript lengths. Analyzing the applicability of the rearrangement problem of SQUID on long read RNA-seq data, and combining both long-read and short-read sequencing datasets are potential directions for improving TSV detection methods.

SQUID is open source and available at `http://www.github.com/Kingsford-Group/squid` and the scripts to replicate the computational experiments described here are available at `http://www.github.com/Kingsford-Group/squidtest`.

2.1.12 Appendix

All experiments here are performed with SQUID version 1.3.

Using de novo assembly and transcript to genome alignment to predict TSV

For the pipeline of de novo transcriptome assembly and transcript-to-genome alignment, the direct output is a series of alignment pieces for each assembled transcript. To derive TSV from the pieces of alignment of each transcript, we still need to use the split-read alignment concordance criteria (2.8) and the edge-building approach. In the case of no TSV, equation (2.8) still holds, since a transcript is generated from one strand of one chromosome, without rearrangements but only deletion of introns. Any violation of (2.8) is treated as a TSV. Here TSVs are still able to be represented by edges in GSG, where segments are the intervals of each piece of alignment, and edges are added in the same principle that traversing segments along the edges will result in a concordant alignment of the assembled transcript. The positions of both breakpoints in a TSV are exactly the two positions linked by the discordant edge, and the orientations corresponds to the connection type of the edge.

Processing TCGA RNA-seq data

We use STAR aligner [34] to align TCGA RNA-seq reads to Ensemble genome 87 [174] with the corresponding gene annotation. STAR aligner [34] is set with the option of outputting chimeric alignments with hanging length 15bp. The chimeric alignments generated by STAR [34] are

further filtered out if the paired-end reads can be aligned concordantly by SpeedSeq aligner [23]. SQUID is applied to concordant alignments generated by STAR [34] and the filtered chimeric alignments. The discordant edge weight coefficient α is set to be 1, that is, we require tumor transcripts to dominate normal transcripts (if they are incompatible) in order to predict corresponding TSVs.

A large number of fusions between immunoglobulin genes are predicted by SQUID. However, there is possibility that B cells are in the mixture of sequencing and have very high expression of immunoglobulin genes (Ig). We cannot tell whether Ig rearrangements are generated by tumor cells or B cells. Therefore, we exclude Ig TSVs during post-processing and exclude them from the descriptive statistics. Note that SQUID does not exclude Ig TSVs internally, because Ig expression and VDJ recombination have been observed to exist in tumor cells, and revealing the role of Ig in tumors may be useful. When normal cells are removed from tumor samples, using SQUID to predict Ig TSVs may help study relationship between Ig and cancer.

Additional Tables

Table 2.2: SQUID parameter specification and values in experiments

Symbol	Description	Value
γ	segment degree threshold	4
θ	edge weight threshold	5
α	discordant edge weight coefficient	8 (simulation and HCC cell line), 1 (TCGA)
mq	minimum mapping quality	255 (STAR), 1 (SpeedSeq)
pq	low Phred quality threshold	4 ($p = 10^{-0.4}$)
l	maximum allowed low Phred quality length	10

Note: mq, pq and l are controls for sequencing quality and mapping quality. If mapping quality of a read is less then threshold mq, the read will not be used in edge building. If the read has a low sequencing, in terms of having more than l bases of sequencing quality lower than pq, the read will not be used in edge building.

Additional Figures

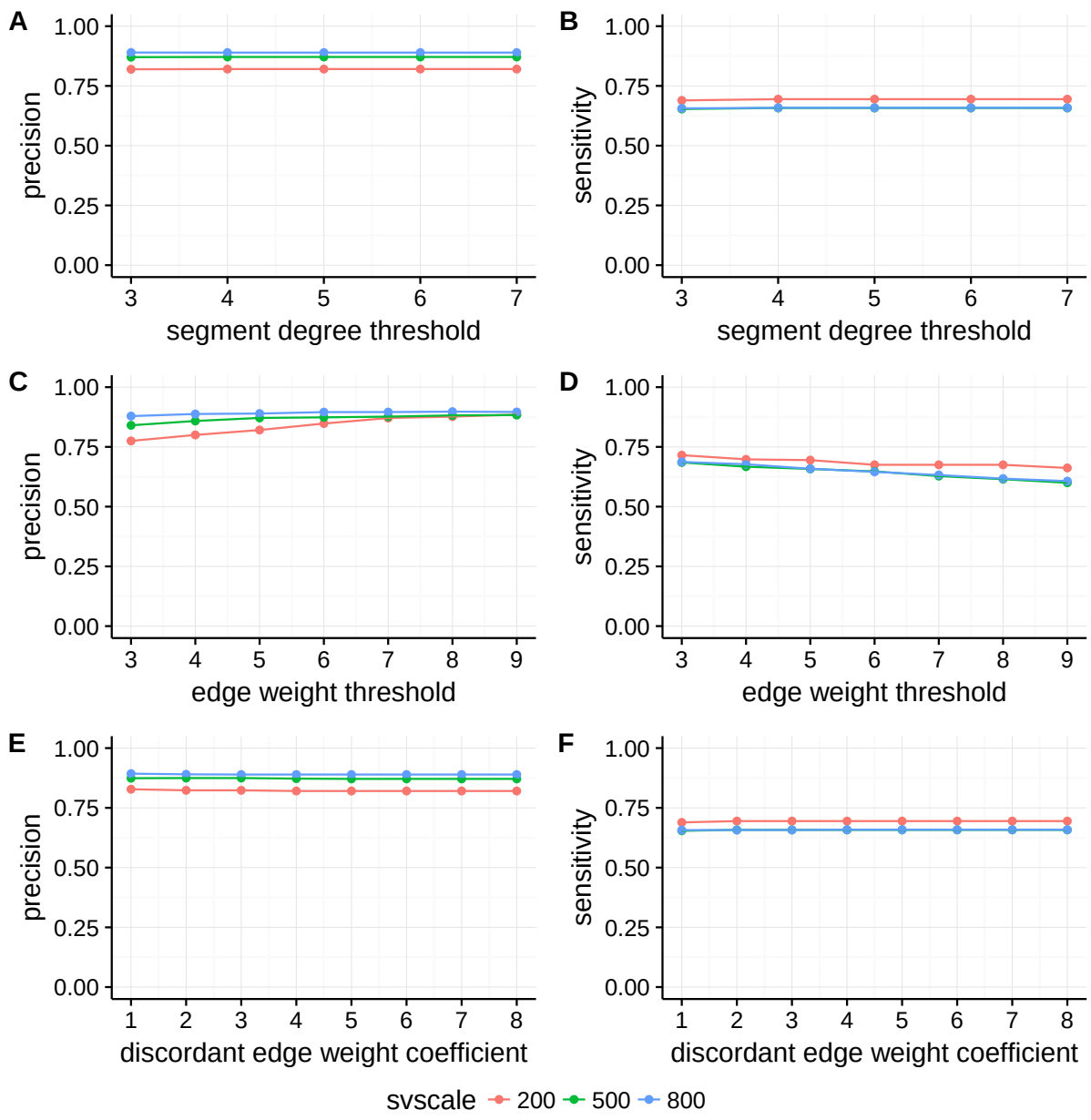

Figure 2.8: Performance of SQUID on simulation data against different parameters values. (A, B) Segment degree threshold γ. Both the precision and sensitivity curves are relatively flat across different values of γ for all numbers of SVs simulated (200, 500, 800). (C, D) Edge weight threshold θ. Increased value of θ leads to increased precision and decreased sensitivity. This parameter determines the natural precision-sensitivity tradeoff and is one of the most important parameters in SQUID. (E, F) Discordant edge weight coefficient α. This parameter adjusts the edge weights according to normal/tumor cell ratio. Since simulation data is homogeneous, varying this parameter does not change the performance of SQUID.

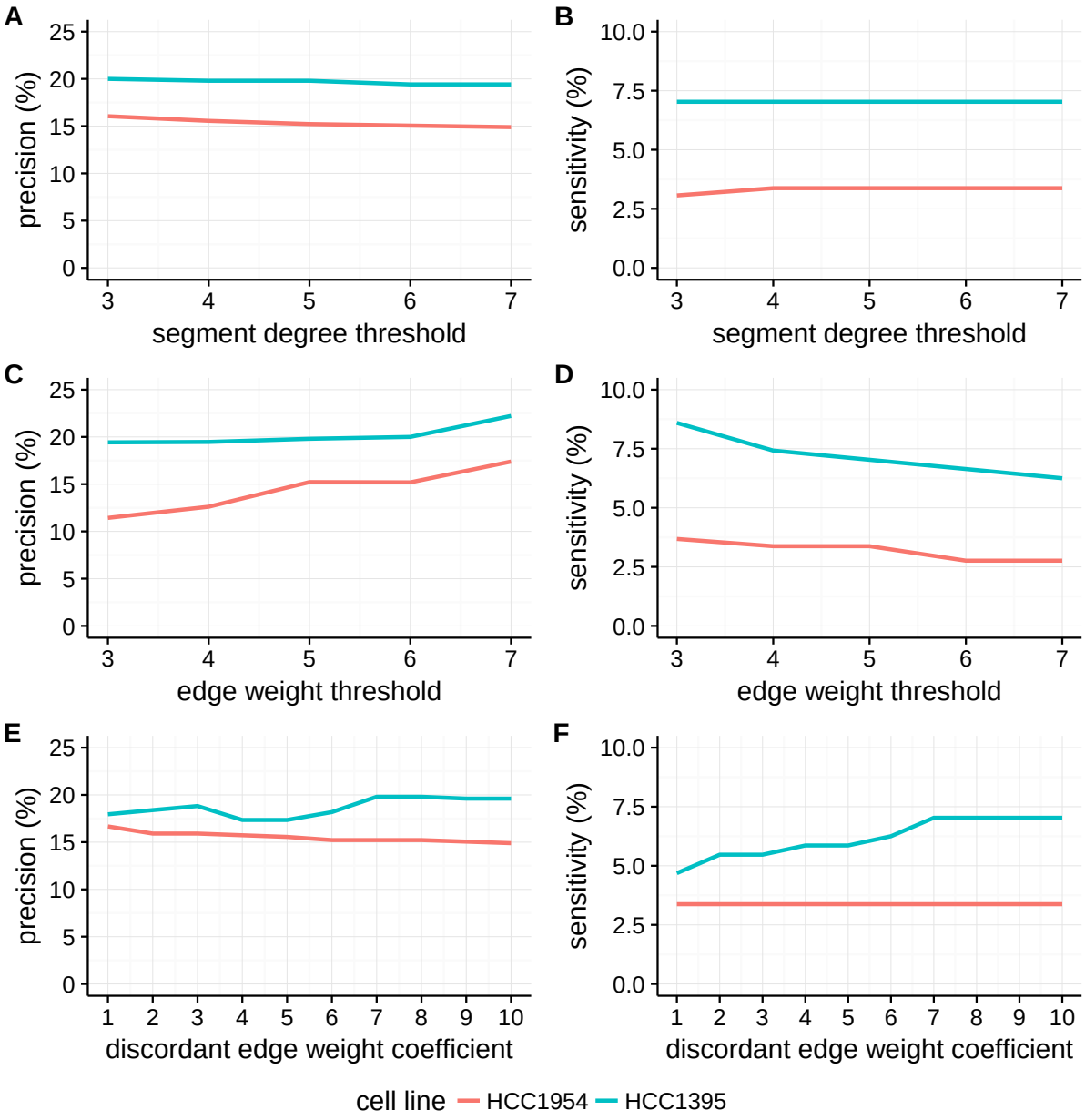

Figure 2.9: Performance of SQUID on real data against different values of parameters. (A, D) Segment degree threshold γ. Both precision and sensitivity are robust against segment degree threshold. (B, D) Edge weight threshold θ. This parameter affects the natural precision-sensitivity tradeoff. For both HCC1954 and HCC1395 cell lines, increasing θ leads to increased precision and decreased sensitivity. (C, F) Discordant edge weight coefficient α. For HCC1954 cell line, sensitivity does not change when increasing α, indicating rearranged tumor transcripts out-number their normal counterparts; while precision decreases slightly because SQUID predicts more TSVs as discordant edge weight coefficient increases. For HCC1395 cell line, sensitivity and precision reach the highest at discordant edge weight coefficient 8 and remain unchanged at 9 and 10. If some normal transcripts out-number the rearranged tumor transcripts, increasing this parameter allows SQUID to capture these TSVs.

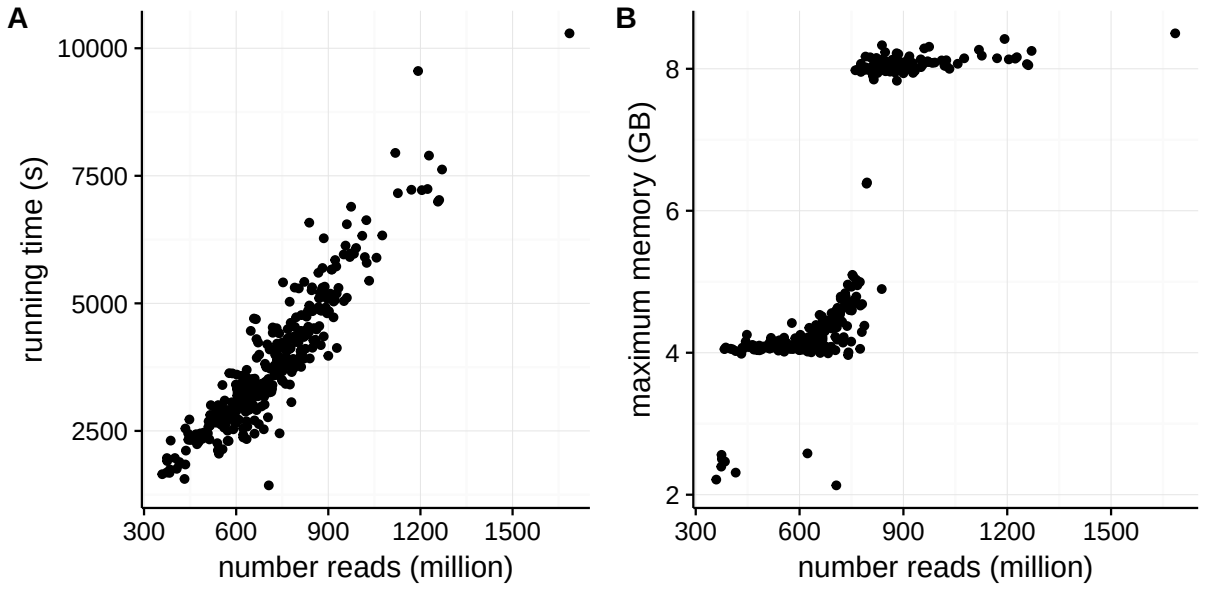

Figure 2.10: (A) Running time and (B) maximum memory usage of SQUID on TCGA RNA-seq datasets with different number of reads. Alignment time and memory are not included. Running majority of TCGA RNA-seq data takes less than 3 hours, and uses around 4 GB or 8 GB memory.

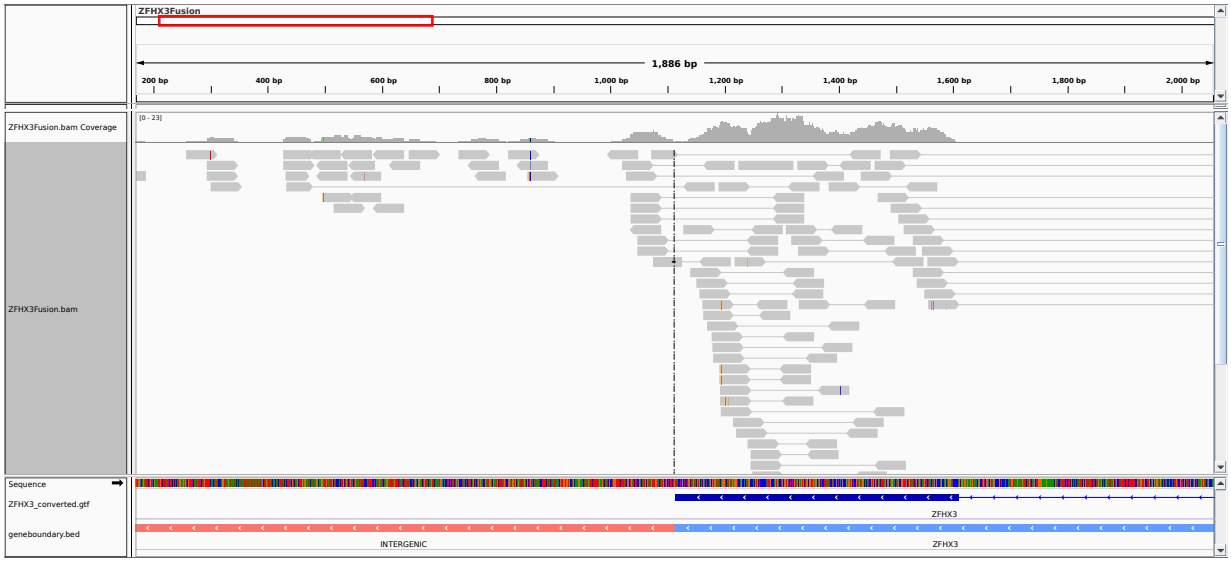

Figure 2.11: IGV visualization of non-fusion-gene TSV involving *ZFHX3* gene. The reference sequence showed by IGV is the junction sequence of TSV. The first track shows the exons of *ZFHX3* gene in the junction sequence. The second track shows the boundaries of the fused genome segments. In the alignment track, read alignments are viewed as pairs (the grey line links two paired-end alignments). Coverage of intergenic segment is less than coverage of *ZFHX3* gene, which indicates the TSV is heterogeneous and appears in a portion of sequencing sample.

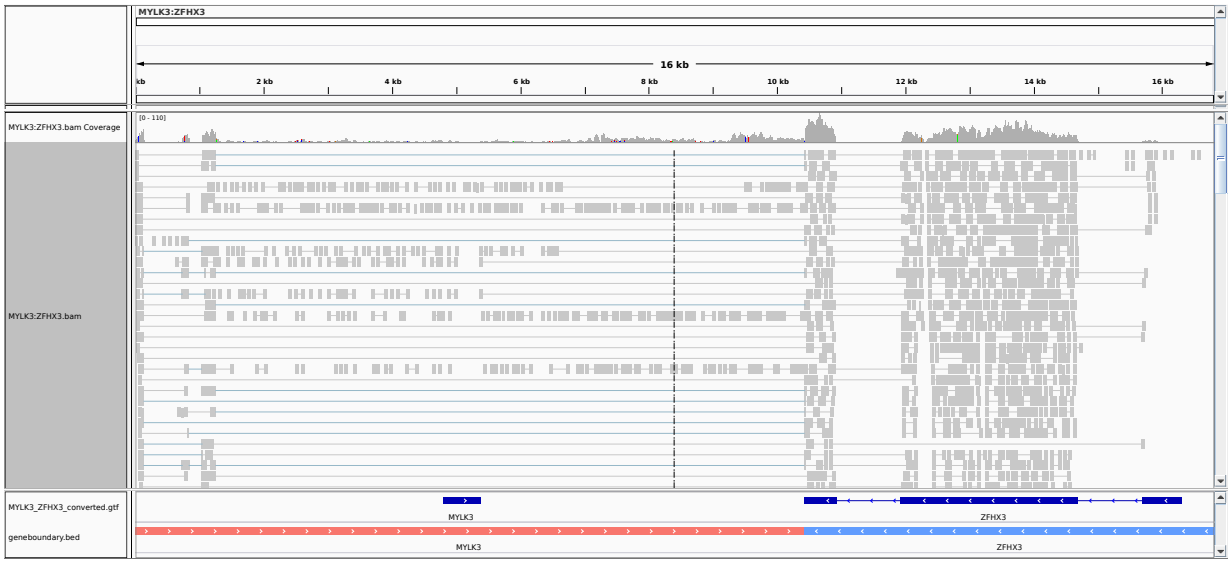

Figure 2.12: IGV visualization of a non-fusion-gene TSV involving the *ZFHX3* gene and the anti-sense strand of *MYLK3* gene. The reference sequence is the junction sequence of TSV. The first track shows the exons of *ZFHX3* gene and the first exon of *MYLK3* gene in the junction sequence. The second track shows the boundaries of the fused genome segments. In the alignment track, read alignments are viewed as pairs (the grey line links two paired-end alignments, and the blue line links split-read alignments). The large coverage difference between *ZFHX3* gene and anti-sense strand of *MYLK3* gene indicates the TSV is heterogeneous. A splicing event in the segment of *MYLK3* anti-sense strand is indicated by the blue lines in alignment track. The splicing sites do not correspond to the exon of *MYLK3* gene because splicing signals are not preserved on the anti-sense strand. Instead, the new splicing junction is the product of the non-fusion-gene TSV.

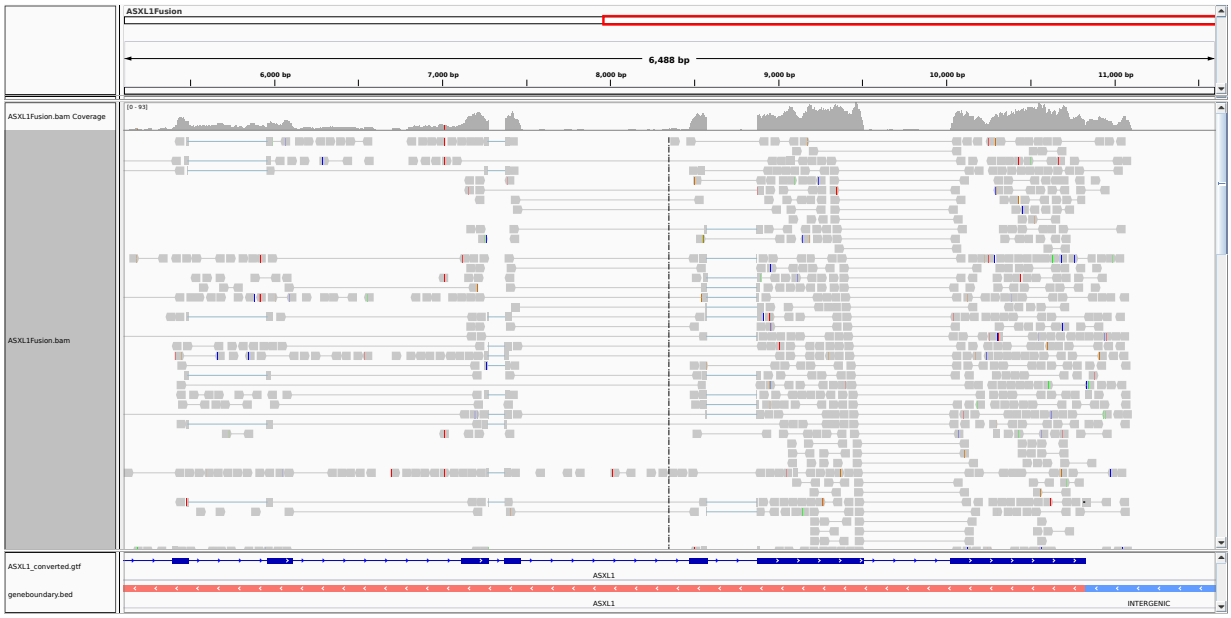

Figure 2.13: IGV visualization of non-fusion-gene TSV involving *ASXL1* gene. The reference sequence is the junction sequence of TSV. The first track shows the exons of *ASXL1* gene in the junction sequence. The second track shows the boundaries of the fused genome segments. In the alignment track, read alignments are viewed as pairs (the grey line links two paired-end alignments). There are many reads spanning the junction point, and the coverage difference between the two segments is small, which implies the TSV is possibly homogeneous.

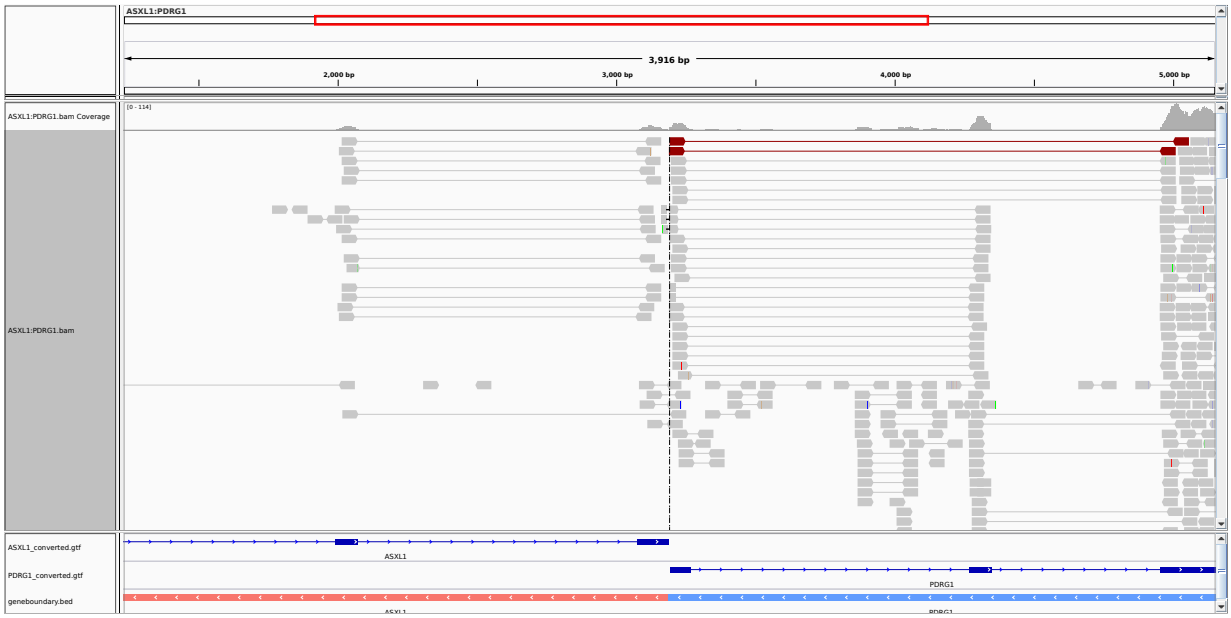

Figure 2.14: IGV visualization of fusion-gene TSV involving *ASXL1* and *PDRG1* genes. The reference sequence is the junction sequence of TSV. The first two annotation tracks show the exons of *ASXL1* and *PDRG1* gene in the junction sequence. The third track shows the boundaries of the fused genome segments. In the alignment track, read alignments are viewed as pairs (the grey line links two paired-end alignments). There are 8 reads spanning the junction point. The coverage of the *ASXL1* gene is much less than that of the *PDRG1* gene, which implies the fusion-gene TSV is heterogeneous.

2.2 Detecting transcriptomic structural variants in heterogeneous contexts via the multiple compatible arrangements problem

SQUID relies on the assumption that the sample is homogeneous, i.e. the original genome contains only one allele that can be represented by a single rearranged string. This assumption is unrealistic in diploid (or high ploidy) organisms. When TSV events occur within the same regions on different alleles, read alignments may suggest multiple conflicting ways of placing a segment. Under the homogeneous assumption, conflicting TSV candidates are regarded as errors. Therefore, this assumption leads to discarding the conflicting TSV candidates that would be compatible on separate alleles and therefore limits the discovery of true TSVs. Conflicting SV candidates are addressed in a few SV detection tools such as VariationHunter-CR [62]. However, VariationHunter-CR assumes a diploid genome, and its model is built for WGS data that lacks ability to handle RNA-seq data.

We provide an extension of SQUID in the heterogeneity context. The heterogeneous allele scenario is counted by seeking multiple (k) rearrangements of genome segments, which turns in the MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM (MCAP). Complexity and approximation algorithms are discussed below. We use "arrangement" to replace "rearrangement" of

genome segments hereafter, while keep using "rearranged genome" to refer to the concatenated sequence according to a specific arrangements of segments.

2.2.1 The MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM (MCAP)

We recapitulate the notations that are used in this section in the following table (Table 2.3). For details of definition and explanation, see Section 2.1.1.

Table 2.3: Notations used in MCAP

notation	meaning
S	A set of genome segments.
$G = (V, E, \mathbf{w})$	Genome segment graph (GSG), where vertex set $V = \{s_h : s \in S\} \cup \{s_t : s \in S\}$ and $\mathbf{w}$ is the weight vector on edges.
$w : \mathcal{P}(E) \to \mathbb{R}$	Weight map that maps a subset of edges in E to the sum of their weights.
$\pi : S \to \{1, \ldots, \lvert S \rvert\}$	Permutation on genome segments.
$f : S \to \{0, 1\}$	Orientation of genome segments.
$A = \{(\pi_i, f_i)\}_i$	A set of arrangements, where the i^{th} arrangement is represented by permutation π_i and orientation f_i.
$e \sim (\pi, f)$	Edge e is concordant with arrangement (π, f)

Problem statement

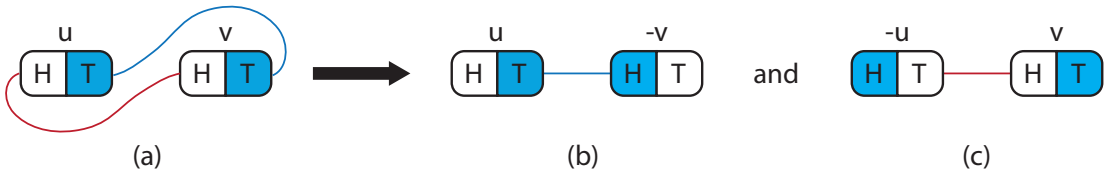

Figure 2.15: MCAP resolves conflicts. The white ends of the segments represent head with respect to the original genome. The blue ends represent tail with respect to the original genome. "H" stands for head and "T" stands for tail in the current arrangement. (a) Two conflicting edges connecting two segments u and v. If the sample is known to be homogeneous ($k = 1$), then the conflict is due to errors. If $k = 2$, MCAP seeks to separate two edges into two compatible arrangements as in (b) and (c). (b) In the first arrangement, segment v is flipped, which makes the blue edge concordant. (c) In the second arrangement, u is flipped to make the red edge concordant.

Given an input GSG $G = (V, E, w)$ and a positive integer k, the MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM seeks a set of k arrangements $A = \{(\pi_i, f_i)\}_{i=1}^{k}$ that are able to

generate the maximum number of sequencing reads:

$$\max_{A} \sum_{e \in E} w(e) \cdot \mathbf{1}\left[e \sim A\right], \tag{2.10}$$

where $\mathbf{1}\left[e \sim A\right]$ is 1 if edge e is concordant in at least one $(\pi_i, f_i) \in A$, and 0 otherwise.

This objective function aims to find an optimal set of k arrangements of segments where the sum of concordant edge weights is maximized in the arranged alleles, where k is the number of alleles and assumed to be known. The objective seeks to maximize the agreement between arranged allelic sequences and observed RNA-seq data. Assuming that the majority of RNA-seq reads are sequenced correctly, the concordant edges with respect to the optimal set of arrangements represent the most confident transcriptomic adjacencies. In heterogeneous samples where $k \neq 1$, MCAP separates the conflicting edges onto k alleles as shown in an example in Figure 2.15.

When $k = 1$, the problem reduces to finding a single arranged genome to maximize the number of concordant reads, which is the problem that SQUID solves in Section 2.1. We refer to the special case when $k = 1$ as SINGLE COMPATIBLE ARRANGEMENT PROBLEM (SCAP).

Predicted TSVs are the concordant edges with respect to any of the arrangements in a solution to MCAP that were either discordant with respect to the reference genome or spanning multiple chromosomes.

2.2.2 NP-completeness of SCAP and MCAP

Theorem 1. *SCAP is NP-complete.*

Proof. We prove the NP-completeness by reducing from the Fragment Orientation Problem (FOP) that has been formulated and studied by Kececioglu et al. [68]. In FOP, for any pair of fragments, there is evidence supporting or against that they have the same orientation. FOP maximizes the agreement with the evidence by assigning the fragment orientation. We rephrase the problem statement as follows.

Input: A set of fragments $\mathcal{F}$ and a score function $S : \mathcal{F} \times \{0,1\} \times \mathcal{F} \times \{0,1\} \to \mathbb{R}_+$ that satisfies the following two conditions:

$$S(F_i, o_i, F_j, o_j) = S(F_j, o_j, F_i, o_i)$$
$$S(F_i, o_i, F_j, o_j) = S(F_i, 1 - o_i, F_j, 1 - o_j)$$

Output: An orientation of fragments $O : \mathcal{F} \to \{0,1\}$.

Objective: Maximize the sum of score according to the orientation,

$$\max_{O} \sum_{F_i, F_j \in \mathcal{F}, F_i \neq F_j} S(F_i, O(F_i), F_j, O(F_j)).$$

Kececioglu et al. [68] defined two symmetric functions and used them to express the objective function in a more specific way:

$$\max_{O} \sum_{F_i, F_j \in \mathcal{F}, F_i \neq F_j} \mathit{same}(F_i, F_j)\mathbf{1}[O(F_i) = O(F_j)] + \mathit{opp}(F_i, F_j)\mathbf{1}[O(F_i) \neq O(F_j)],$$

where $same : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ is defined as $same(F_i, F_j) \triangleq S(F_i, 0, F_j, 0) = S(F_i, 1, F_j, 1)$, and $opp : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ is defined as $opp(F_i, F_j) \triangleq S(F_i, 0, F_j, 1) = S(F_i, 1, F_j, 0)$.

Given any FOP instance, a SCAP instance is constructed in polynomial time by constructing a segment for each fragment in $\mathcal{F}$ and assigning edge weights based on the $same$ and opp function values. Specifically, for fragment F_i, construct a segment s^i. For any pair of segments (s^i, s^j) construct four edges with the following weights: $w(e = (s^i_h, s^j_h)) = opp(F_i, F_j)$, $w(e = (s^i_t, s^j_h)) = same(F_i, F_j)$, $w(e = (s^i_h, s^j_t)) = same(F_i, F_j)$, and $w(e = (s^i_t, s^j_t)) = opp(F_i, F_j)$. Due to the correspondence between segments S and fragments $\mathcal{F}$, they can be viewed as parameter substitution and used in interchangeably in FOP and SCAP.

Because the constructed GSG is a complete graph except that there is no within-segment edges, the maximization of SCAP over permutation π and orientation f can be rewritten as

$$
\begin{aligned}
&\max_{\pi,f} \sum_e w(e)\mathbf{1}[e \sim (\pi, f)] \\
=&\max_{\pi,f} \sum_{s^i,s^j \in S, s^i \neq s^j} w(e = (s^i_h, s^j_h))\mathbf{1}[(s^i_h, s^j_h) \sim (\pi, f)] + w(e = (s^i_t, s^j_h))\mathbf{1}[(s^i_t, s^j_h) \sim (\pi, f)] + \\
&\qquad w(e = (s^i_h, s^j_t))\mathbf{1}[(s^i_h, s^j_t) \sim (\pi, f)] + w(e = (s^i_t, s^j_t))\mathbf{1}[(s^i_t, s^j_t) \sim (\pi, f)] \\
=&\max_{\pi,f} \sum_{s^i,s^j \in S, s^i \neq s^j} opp(s^i, s^j)\mathbf{1}\left\{1 - f(s^i) = \mathbf{1}[\pi(s^i) < \pi(s^j)] = f(s^j)\right\} + \\
&\qquad\qquad same(s^i, s^j)\mathbf{1}\left\{f(s^i) = \mathbf{1}[\pi(s^i) < \pi(s^j)] = f(s^j)\right\} + \\
&\qquad\qquad same(s^i, s^j)\mathbf{1}\left\{1 - f(s^i) = \mathbf{1}[\pi(s^i) < \pi(s^j)] = 1 - f(s^j)\right\} + \\
&\qquad\qquad opp(s^i, s^j)\mathbf{1}\left\{f(s^i) = \mathbf{1}[\pi(s^i) < \pi(s^j)] = 1 - f(s^j)\right\} \\
=&\max_{\pi,f} \sum_{s^i,s^j \in S, s^i \neq s^j} opp(s^i, s^j)\mathbf{1}[1 - f(s^i) = f(s^j)] + same(s^i, s^j)\mathbf{1}[f(s^i) = f(s^j)] \\
=&\max_{f} \sum_{s^i,s^j \in S, s^i \neq s^j} opp(s^i, s^j)\mathbf{1}[1 - f(s^i) = f(s^j)] + same(s^i, s^j)\mathbf{1}[f(s^i) = f(s^j)]
\end{aligned}
$$

In the last step of the above equation, since the objective function does not contain permutation π, we can take π out of the optimization parameter. That means for any permutation the maximum sum of concordant edge weights is the same. Applying reparameterization by changing segment s^* to fragment F_* and changing the segment orientation function f with fragment orientation function O, the above maximization problem is the same as FOP. As a result, the optimal solution of SCAP and FOP can be used interchangeably to maximize the criterion of each other.

Therefore, given any instance of FOP, an instance of SCAP can be constructed in polynomial time whose solution contains an orientation function that maximized FOP instance at the same time. Since FOP is NP-complete, SCAP is also NP-complete. □

Corollary 1.1. *MCAP is NP-complete.*

Proof. SCAP is a special case of MCAP with $k = 1$, so the NP-completeness of MCAP is immediate. □

2.2.3 A $\frac{1}{4}$-approximation algorithm for SCAP

We provide a greedy algorithm for SCAP that achieves at least $\frac{1}{4}$ approximation ratio and takes $O(|V||E|)$ time. The main idea of the greedy algorithm is to place each segment into the current order one by one by choosing the current "best" position. The current "best" position is determined by the concordant edge weights between the segment to be placed and the segments already in the current order.

Algorithm 1: Greedy algorithm for SCAP

Data: Segment set S, genome segment graph $G = (V, E, w)$
Result: An arrangement of the segments and the sum of concordant edge weights

```
order = [];
orientation = [];
for i in 1 : |S| do
    s^i = the i^th segment in S;
    // choose from 4 possible order and orientation options
    options = [(s^i in the beginning of order in forward strand), (s^i in the beginning of
      order in reverse strand), (s^i in the end of order in forward strand), (s^i in the end of
      order in reverse strand)] ;
    for j in 1 : 4 do
        weights[j] =
          w({e ∈ E : e connects s^i with s^k and concordant in options[j], k < i});
    end
    // update the current order and orientation
    opt = argmax_{1≤i≤4, i∈N} weights[i] ;
    order = update segment order as given by options[opt] ;
    orientation = update segment orientation as given by options[opt] ;
end
```

Theorem 2. *Algorithm 1 approximates SCAP with at least $\frac{1}{4}$ approximation ratio.*

Proof. Denote $E' \subset E$ as the concordant edges in the arrangement of Algorithm 1. Let OPT be the optimal value of SCAP. We are to prove $w(E') \geq \frac{1}{4}w(E) \geq \frac{1}{4}OPT$.

For iteration i in the for loop, the edges $E_i = \{e \in E : e \text{ connects } s^i \text{ with } s^j, j < i\}$ are considered when comparing the options. Each of the four options makes a subset of E_i concordant. These subsets are non-overlapping and their union is E_i. Specifically, the concordant edge subset is $\{e = (s_h^i, s_t^j) : j < i\}$ for the first option, $\{e = (s_h^i, s_h^j) : j < i\}$ for the second, $\{e = (s_t^i, s_h^j) : j < i\}$ for the third, and $\{e = (s_t^i, s_t^j) : j < i\}$ for the last.

By the selecting the option with the largest sum of concordant edge weights, the concordant edges E_i' in iteration i satisfies $w(E_i') \geq \frac{1}{4}w(E_i)$. Therefore, the overall concordant edge weights of all iterations in the for loop satisfy

$$\sum_i w(E_i') \geq \frac{1}{4}\sum_i w(E_i) = \frac{1}{4}w\left(\bigcup_i E_i\right).$$

Each edge $e \in E$ must appear in one and only one of E_i, and thus $\bigcup_i E_i = E$. This implies $\sum_i w(E'_i) \geq \frac{1}{4} w(E) \geq \frac{1}{4} OPT$. □

Algorithm 1 can be further improved in practice by considering more order and orientation options when inserting a segment into current order. In the pseudo-code 1, only two possible insertion places are considered: the beginning and the end of the current order. However, a new segment can be inserted in between any pair of adjacent segments in the current order. We provide an extended greedy algorithm to take into account the extra possible inserting positions (Algorithm 2). Algorithm 2 has a time complexity of $O(|V|^2|E|)$, but it may achieve a higher total concordant edge weight in practice.

Algorithm 2: Extended greedy algorithm for SCAP

Data: Segment set S, genome segment graph $G = (V, E, w)$
Result: An arrangement of the segments and the sum of concordant edge weights

$order$ = [];
$orientation$ = [];
for i **in** $1 : |S|$ **do**
    s^i = the i^{th} segment in S;
    `// choose from` $i+1$ `possible order and orientation options`
    $options$ = [(s^i in the beginning of $order$ in forward strand), (s^i in the beginning of $order$ in reverse strand)] ;
    for j **in** $1 : i-1$ **do**
        Append [(s^i right after $order[j]$ in forward strand), (s^i right after $order[j]$ in reverse strand)] to list of $options$;
    end
    for j **in** $1 : 2i$ **do**
        $weights[j]$ = $w(\{e \in E : e$ connects s^i with s^k and concordant in $options[j], k < i\})$;
    end
    `// update the current order and orientation`
    opt = argmax$_{1 \leq i \leq 2i, i \in \mathbf{N}} weights[i]$;
    $order$ = update segment order as given by $options[opt]$;
    $orientation$ = update segment orientation as given by $options[opt]$;
end

2.2.4 A $\frac{3}{4}$-approximation of MCAP with $k = 2$ using a SCAP oracle

If an optimal SCAP solution can be computed, one way to approximate the MCAP's optimal solution is to solve a series of SCAP instances iteratively to obtain multiple arrangements. Here, we prove the solution based on iteratively solving SCAP has an approximation ratio of $\frac{3}{4}$ for the special case of MCAP with $k = 2$.

Algorithm 3: $\frac{3}{4}$-approximation for MCAP with $k = 2$

Data: A genome segment graph $G = (V, E, w)$

Result: a set of two arrangements, sum of weights of edges that are concordant in either arrangement

a_1 = optimal SCAP arrangement on G;
$E' = \{e \in E : e \text{ is discordant in } a_1\}$;
$G' = (V, E', w)$;
a_2 = optimal SCAP arrangement on G';
$\tilde{E} = \{e \in E : e \sim A, A = \{a_1, a_2\}\}$;
$W = \sum_{e \in \tilde{E}} w(e)$;
return $(\{a_1, a_2\}, W)$;

Theorem 3. *Algorithm 3 is a $\frac{3}{4}$-approximation of MCAP with $k = 2$. Denote the optimal objective sum of edge weights in MCAP with $k = 2$ as OPT, and the sum of edge weights in the two iterative SCAP as W, then*

$$W \geq \frac{3}{4} OPT$$

Proof. Denote MCAP with $k = 2$ as 2-MCAP. Let E_1^d and E_2^d be concordant edges in the optimal two arrangements of 2-MCAP. It is always possible to make the concordant edges of the arrangements disjoint by removing the intersection from one of the concordant edge set, that is $E_1^d \cap E_2^d = \emptyset$. Let $E^d = E_1^d \cup E_2^d$. The optimal value is $w(E^d)$.

Denote the optimal set of concordant edges in the first round of Algorithm 3 as E_1^s. The optimal value of SCAP is $w(E_1^s)$. E_1^s can have overlap with the two concordant edge sets of the 2-MCAP optimal solution. Let the intersections be $I_1 = E_1^d \cap E_1^s$ and $I_2 = E_2^d \cap E_1^s$. Let the unique concordant edges be $D_1 = E_1^d - E_1^s$, $D_2 = E_2^d - E_1^s$ and $S = E_1^s - E_1^d - E_2^d$.

After separating the concordant edges in 2-MCAP into the intersections and unique sets, the optimal value of 2-MCAP can be written as $w(E^d) = w(I_1) + w(I_2) + w(D_1) + w(D_2)$, where the four subsets are disjoint. Therefore the smallest weight among the four subsets must be no greater than $\frac{1}{4} w(E^d)$. We prove the approximation ratio under the following two cases and discuss the weight of the second round of SCAP separately:

Case (1): the weight of either D_1 or D_2 is smaller than $\frac{1}{4} w(E^d)$. Because the two arrangements in 2-MCAP are interchangeable, we only prove for the case where $w(D_1) \leq \frac{1}{4} w(E^d)$. A valid arrangement of the second round of SCAP is the second arrangement in 2-MCAP, though it may not be optimal. The maximum concordant edge weights added by the second round of SCAP must be no smaller than $w(D_2)$. Combining the optimal values of two rounds of SCAP, the concordant edge weight is

$$\begin{aligned} W &\geq w(E_1^s) + w(D_2) = w(S) + w(I_1) + w(I_2) + w(D_2) \\ &\geq w(E^d) - w(D_1) \\ &\geq \frac{3}{4} w(E^d). \end{aligned} \tag{2.11}$$

Case (2): both $w(D_1) \geq \frac{1}{4}w(E^d)$ *and* $w(D_2) \geq \frac{1}{4}w(E^d)$. The subset with smallest sum of edge weights is now either I_1 or I_2. Without loss of generality, we assume I_1 has the smallest sum of edge weights and $w(I_1) \leq \frac{1}{4}w(E^d)$. Because the first round SCAP is optimal for the SCAP problem, its objective value should be no smaller than the concordant edge weights of either arrangement in 2-MCAP. Thus

$$w(E_1^s) \geq w(E_2^d) = w(D_2) + w(I_2). \tag{2.12}$$

A valid arrangement for the second round of SCAP can be either of the arrangements in 2-MCAP optimal solution. Picking the first arrangement of 2-MCAP as the possible (but not necessarily optimal) arrangement for the second round of SCAP, the concordant edge weights added by the second round of SCAP must be no smaller than $w(D_1)$. Therefore, the total sum of concordant edge weights of the optimal solutions of both rounds of SCAP is

$$\begin{aligned} W &\geq w(E_1^s) + w(D_1) \\ &\geq w(D_2) + w(I_2) + w(D_1) \\ &= w(E^d) - w(I_1) \\ &\geq \frac{3}{4}w(E^d). \end{aligned} \tag{2.13}$$

□

Corollary 3.1. *An approximation algorithm for MCAP with* $k = 2$ *can be created by using Algorithm 1 as the oracle for SCAP in Algorithm 3. This approximation algorithm runs in* $O(|V||E|)$ *time and achieves at least* $\frac{3}{16}$ *approximation ratio.*

The proof of the corollary is similar to the proof of Theorem 3. By adding a multiplier of $\frac{1}{4}$ to the right of inequalities (2.12) when lower bounding $w(E_1^s)$ by $w(E_2^d)$, the $\frac{3}{16}$ approximation ratio can be derived accordingly.

2.2.5 Integer linear programming formulation for MCAP

MCAP, for general k, can be formulated as an integer linear programming (ILP) to obtain an optimal solution. We rewrite the i-th permutation (π_i), orientation (f_i) and decision ($\mathbf{1}[e \sim (\pi_i, f_i)]$) functions with three boolean variables y_e^i, z_e^i and x_e^i. For $i \in \{1, 2, ..., k\}$ and $e \in E$, we have:

- $x_e^i = 1$ if edge $e \sim (\pi_i, f_i)$ and 0 otherwise.
- $y_u^i = 1$ if $f_i(u) = 1$ for segment u and 0 if $f_i(u) = 0$.
- $z_{uv}^i = 1$ if $\pi_i(u) < \pi_i(v)$, or segment u is in front of v in arrangement i and 0 otherwise.

In order to account for the edges that are concordant in more than one arrangement in the summation in Equation 2.10, we define q_e such that $q_e = 1$ if edge e is concordant in one of the k arrangements and 0 otherwise. The constraints for q_e are as follows:

$$q_e \leq \sum_i^k x_e^i \tag{2.14}$$

$$q_e \leq 1 \tag{2.15}$$

The objective function becomes

$$\max_{x_e^i, y_u^i, z_{uv}^i} \sum_{e \in E} w(e) \cdot q_e \tag{2.16}$$

We then add ordering and orientation constraints. If an edge is a tail-head connection, i.e. concordant to the reference genome, $x_e^i = 1$ if and only if $z_{uv}^i = y_u^i = y_v^i$. If an edge is a tail-tail connection, $x_e^i = 1$ if and only if $z_{uv}^i = 1 - y_v^i = y_u^i$. If an edge is a head-tail connection, $x_e^i = 1$ if and only if $z_{uv}^i = 1 - y_u^i = 1 - y_v^i$. If an edge is a head-head connection, $x_e^i = 1$ if and only if $z_{uv}^i = 1 - y_u^i = y_v^i$. The constraints for a tail-head connection are listed below in Equation 2.17, which enforce the assignment of boolean variables y_e^i, z_e^i and x_e^i:

$$\begin{aligned} x_e^i &\leq y_u^i - y_v^i + 1, \\ x_e^i &\leq y_v^i - y_u^i + 1, \\ x_e^i &\leq y_u^i - z_{uv}^i + 1, \\ x_e^i &\leq z_{uv}^i - y_u^i + 1, \end{aligned} \tag{2.17}$$

These are the same set of constraints as SQUID for each arrangement, and see equation 2.4–2.7 for details of the constraints under other connection types. Additionally, constraints are added so that all segments are put into a total order within each allele. For two segments u, v, segment u will be either precede or follow segment v, i.e. $z_{uv}^i + z_{vu}^i = 1$. For three segments u, v, w, if u precedes v and v precedes w, then u has to precede w: $1 \leq z_{uv}^i + z_{vw}^i + z_{wu}^i \leq 2$.

The total number of constraints as a function of k is $4k|E| + k\binom{|V|}{3} + 2|E| = O(k(|E| + V^3))$. When k increases, the number of constraints grows linearly. When $k = 1$, the ILP formulation reduces to the same formulation as SQUID.

2.2.6 Characterizing the conflict structures that imply heterogeneity

In this section, we ignore edge weights and characterize the graph structures where homogeneous assumption cannot explain all edges. We add a set of segment edges, $\hat{E}$, to the GSG. Each $\hat{e} \in \hat{E}$ connects the two endpoints of each segment, i.e. $\hat{e} = \{s_h, s_t\}$ for $s \in S$. The representation of GSG becomes $G = (E, \hat{E}, V)$.

Definition 6 (Conflict Structures and Compatible Structures)**.** *A* conflict structure*, $CS = (E', \hat{E}', V')$, is a subgraph of a GSG where there exists a set of edges E' that cannot be made concordant using any single arrangement. A* compatible structure *is a subgraph of a GSG where there exists a single arrangement such that all edges can be made concordant in it.*

Definition 7 (Simple cycle in GSG)**.** *A* simple cycle*, $C = (E', \hat{E}', \{v_0, \ldots, v_{n-1}\})$, is a subgraph of a GSG, such that $E' \subseteq E, \hat{E}' \subseteq \hat{E}$ and $v_i \in V$, with $(v_i, v_{(i+1) \mod n}) \in E' \cup \hat{E}'$ and where $v_i \neq v_j$ when $i \neq j$ except $v_{n-1} = v_0$.*

Definition 8 (Degree and special degree of a vertex in subgraphs of GSG)**.** *Given a subgraph of GSG, $G' = (E', \hat{E}', V')$, $deg_{E'}(v)$ refers to the degree of vertex $v \in V'$ that counts only the edges $e \in E'$ that connect to v. $deg(v)$ refers to the number of edges $e \in E' \cup \hat{E}'$ that connect to v.*

Theorem 4. *Any acyclic subgraph of GSG is a compatible structure.*

Proof. We show that any acyclic subgraph with N edges ($|E'| + |\hat{E}'| = N$), $G'_N = (E', \hat{E}', V')$, of GSG is a compatible structure by induction.

When $|E'| + |\hat{E}'| = 1$, G'_1 is a compatible structure because no other edge in G' is in conflict with the only edge $e \in E'$.

Assume the theorem hold for any acyclic subgraph that contains n edges. Let $G'_{n+1} = (E', \hat{E}, V')$ be an acyclic subgraph with $n+1$ edges. Since G'_{n+1} is acyclic, there must be a leaf edge that is incident to a leaf node. Denote the leaf node as v_b and the leaf edge $e = (u_a, v_b) \in E' \cup \hat{E}'$ ($a, b \in \{h, t\}$). By removing edge e and leaf node v_b, the subgraph $G'_n = (E' - \{e\}, \hat{E}' - \{e\}, V' - \{v_b\})$ is also acyclic and contains n edges. According to the assumption, G'_n is a compatible structure and there is an arrangement of the segments in which all edges in $E' \cup \hat{e}' - \{e\}$ is concordant. Because no other edge in $E' \cup \hat{E}'$ except e connects to v_b, it is always possible to place segment v back to the arrangement such that e is concordant. Specifically, one of the four placing options will satisfy edge e: the beginning of the arrangement with orientation 1, the beginning with orientation 0, the end with orientation 1 and the end with orientation 0. Therefore, G'_{n+1} is a compatible structure.

By induction, acyclic subgraph G'_N of GSG with any $|E'|$ is a compatible structure. □

Theorem 5. *A simple cycle $C = (E', \hat{E}', V')$ is a compatible structure if and only if there are exactly two vertices, v_j and v_i such that $deg_{E'}(v_i) = deg_{E'}(v_j) = 2$ and v_i and v_j belongs to different segments.*

Proof. We prove sufficiency and necessity separately in Lemma 1 and Lemma 2. □

Lemma 1. *If C is a compatible structure, there are exactly two vertices, v_i, v_j that belong to different segments, such that $deg_{E'}(v_i) = deg_{E'}(v_j) = 2$*

Proof. We discuss compatiblity in two cases:

***Case (1)**: All edges are concordant in C.* Sort the vertices by genomic locations in ascending order and label the first vertex v_1 and the last v_n, assuming $|V'| = n$. Similarly, sort the set of segments S' in C by the values of their permutation function π and label the first segment s^1 and the last s^m, assuming $|S'| = m$. Since concordant connections can only be tail-head connections (e.g. Figure 2.15 b,c), $v_1 = s^1_t$ and $v_n = s^m_h$. Since C is a simple cycle, all vertices $v \in V'$ have $deg(v) = 2$. Because v_1 and v_n are the first and last vertices in this arrangement, the edges incident to v_1 or v_n must be in E'. It follows that the two edges incident to v_1 connects to s^2_h and s^m_h. Similarly, edges incident to v_n connects to s^1_t and s^{n-1}_t. Therefore, we have $deg_{E'}(v_1) = deg_{E'}(v_n) = 2$. Any other vertex v_i ($1 < i < n$) is connected by one $e \in E'$ and one $\hat{e} \in \hat{E}'$ and thus has $deg_{E'}(v_i) = 1$.

***Case (2)**: Some edges are discordant in C.* If discordant edges exist in cycle C, according to the definition of compatible structure, segments in C can be arranged such that all edges are concordant. This reduces to case (1). □

Lemma 2. *If there are exactly two vertices in V' that belong to different segments, v_i and v_j, such that $deg_{E'}(v_i) = deg_{E'}(v_j) = 2$, then C is a compatible structure.*

Proof. Let v_i and v_j be the one of the end points of segments s^i and $s^j (i \neq j)$, respectively. We can arrange s^i and s^j such that $\pi(s^i) = \min_{s \in S'} \pi(s)$, $\pi(s^j) = \max_{s \in S'} \pi(s)$ and that $v_i = s^i_t$, $v_j = s^j_h$. Rename v_i to v_1 and v_j to v_n. Since C is a simple cycle, we can find two simple paths, P_1 and P_2, between v_1 and v_n and there is no edge between P_1 and P_2. Let P'_1 and P'_2 denote P_1 and P_2 that exclude v_1 and v_n and the edges incident to v_1 and v_n. Since P'_1 and P'_2 as acyclic subgraphs of GSG, according to Theorem 4, P'_1 and P'_2 are compatible structures and therefore segments in P'_1 and P'_2 can be arranged so that all edges are concordant. Denote the first and last vertices in the arranged P'_1 as v_2 and v_3, and the first and last vertices in the arranged P'_2 as v_4 and v_5. Because all the edges are concordant in P'_1, v_2 and v_3 are the head and tail of the first and last segments in P'_1. Because only v_1 and v_n have $deg_{E'} = 2$ in C, v_2 must be connected to v_1 or v_n and v_3 must be connected to v_n or v_1. A similar argument applies to v_4 and v_5. To ensure concordance of edges connected to v_1 and v_n, if v_n is connected to v_2 and v_1 is connected to v_3, we flip all the segments in P'_1. The similar operation is applied to v_4, v_5 and P'_2. Now we have a compatible structure. □

Corollary 5.1. *A necessary condition for a subgraph $(E', \hat{E}', V')$ to be a conflict structure is that it contains cycles. A sufficient condition for a subgraph $(E', \hat{E}', V')$ to be a conflict structure is that it contains a simple cycle which is not a compatible structure.*

The corollary is a direct derivation from Theorem 4 and Theorem 5 when considering general graph structures.

In practice, we determine if a discordant edge, $e = (u, v)$, is involved in a conflict structure by enumerating all simple paths using a modified depth-first search implemented in Networkx [53, 135] between u and v omitting edge e. We add e to each path and form a simple cycle. If the simple cycle satisfies Corollary 5.1, we stop path enumeration and label the e as discordant edge involved in conflict structure. If the running time of path enumeration exceeds 0.5 seconds, we shuffle the order of DFS and repeat the enumeration. If path enumeration for e exceeds 1000 reruns, we label e as undecided.

2.2.7 Results of comparison with SQUID detections and approximation algorithm

To produce an efficient, practical algorithm for TSV detection in diploid organisms, we use the following approach, which we denote as D-SQUID: Run the ILP under the diploid assumption by setting $k = 2$ on every connected component of GSG separately. If the ILP finishes or the running time of the ILP exceeds one hour, output the current arrangements.

D-SQUID identifies more TSVs in TCGA samples than SQUID

We calculate the fraction of discordant edges involved in conflict structures (Figure 2.16a) in 381 TCGA samples from four types of cancers: bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD). Among all samples, we found less than 0.5% undecided edges out of all discordant edges. The distribution of fraction of discordant edges within conflict structures are different among cancer types. The more discordant edges are involved in conflict structures, the more heterogeneous the

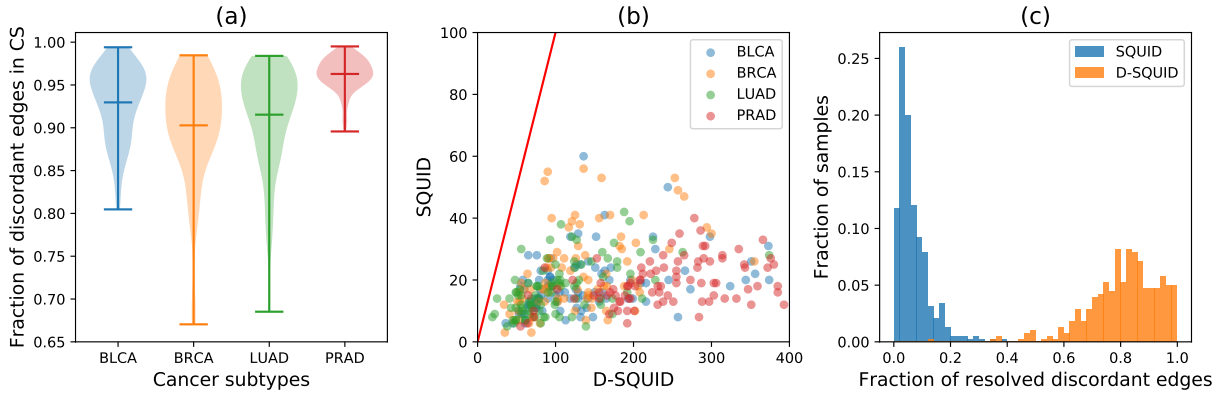

Figure 2.16: Performance of D-SQUID and SQUID on TCGA samples. (a) The distribution of fractions of discordant edges that are involved in each identified conflict structure (CS) in four cancer subtypes. Minima, maxima and means of the distributions are marked by horizontal bars. (b) Number of TSVs identified by SQUID versus D-SQUID. (c) Histogram of fractions of resolved discordant edges by SQUID and D-SQUID.

sample is. Among four cancer types, PRAD samples exhibit the highest extent of heterogeneity and BRCA samples exhibit the lowest. On average, more than 90% of discordant edges are within conflict structures in all samples across four cancer types. This suggests that TCGA samples are usually heterogeneous and may be partially explained by the fact that TCGA samples are usually a mixture of tumor cells and normal cells [5].

We compare the number of TSVs found by D-SQUID and SQUID (Figure 2.16b). In all of our results, all of the TSVs found by SQUID belong to a subset of TSVs found by D-SQUID. D-SQUID identifies many more TSVs than SQUID on all four types of cancers.

A discordant edge is termed resolved if it is made concordant in one of the arrangements. Among all discordant edges in all samples, D-SQUID is able to resolve most of them (Figure 2.16c), while SQUID is only able to resolve fewer than 50% of them. The results demonstrate that D-SQUID is more capable of resolving conflict structures in heterogeneous contexts, such as cancer samples, than SQUID.

D-SQUID identifies more true TSV events than SQUID in cancer cell lines

We compare the ability of D-SQUID and SQUID to detect fusion-gene and non-fusion-gene events on previously studied breast cancer cell lines HCC1395 and HCC1954 [41]. The dataset and ground truth is the same as the ones used in SQUID for performance evaluation in Section 2.1.8. In both cell lines, D-SQUID discovers more TSVs than SQUID. In HCC1954, D-SQUID identifies the same number of known TSVs including fusions of gene (G) regions and intergenic (IG) regions compared with SQUID. In HCC1395, D-SQUID identifies 2 more true TSV events that are fusions of genic regions. We tally the fraction of discordant edges in conflict structures (Figure 2.17c) and find similar fractions between HCC1395 and HCC1954, which indicates that the extent of heterogeneity in two samples are similar. Compared to Figure 2.16a, the fraction in HCC samples is much lower than that in TCGA samples. This matches the fact that

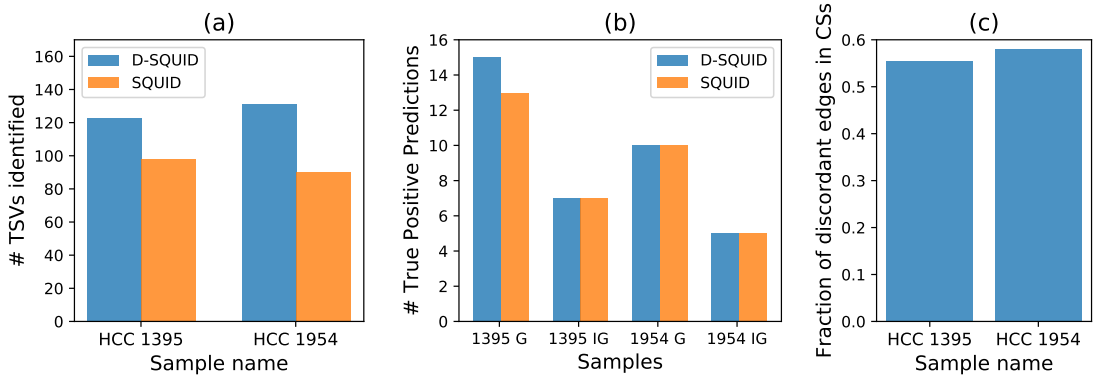

Figure 2.17: Performance of D-SQUID and SQUID on breast cancer cell lines with experimentally verified SV. (a) Total TSVs found. In both cell line samples, D-SQUID discovered more TSVs than SQUID. (b) Number of known fusion-gene and non-fusion-gene events recovered by D-SQUID and SQUID. G denotes TSVs that affect gene regions. IG denotes TSVs that affect intergenic regions. (c) Fraction of discordant edges in conflict structures.

two HCC samples contain the same cell type and are both cell line samples, which are known to be less heterogeneous than TCGA samples.

D-SQUID predicts TSVs in biologically significant genes in cancer cell lines

Figure 2.18 gives two examples of TSVs predicted by D-SQUID but not by SQUID. Such TSVs are involved in conflict structures and can only be resolved by separating discordant edges into different arrangements.

An example of a validated TSV is shown in Figure 2.18(a). The head-tail connection between segment u^1 and u^3 conflicts with the tail-head connections between segments u^1 and u^2 and segments u^2 and u^3. Such a conflict structure is resolved by separating edge (u^1_h, u^3_t) into the second arrangement. Notice that since no discordant edges are made concordant in the first arrangement, no new TSVs are predicted. Therefore, the corresponding gene model for the first arrangement is the same as that of the original arrangement. The affected regions are exons of ERO1A and FERMT2 genes. As predicted by D-SQUID, this TSV involves an insertion of the sixth and the seventh exons of FERMT2 between the sixth and seventh exons of ERO1A.

Among the unvalidated TSVs predicted by D-SQUID, some of them affect genes that are associated with breast cancer. The TSV shown in Figure 2.18(b) involves an insertion of the 3' untranslated region (UTR) of CLPSL1 and the entire CLPS gene between the first and second exons of CLPSL1. It has been reported that CLPSL1 is associated with a prognostic factor of breast cancer [173].

A full list of affected regions in HCC samples can be found in the additional files.

Evaluation of approximation algorithms

We evaluate the approximation algorithms for diploid MCAP ($k = 2$) using two different subroutines described in previous sections. In this subsection, $A1$ refers to using Algorithm 1 with

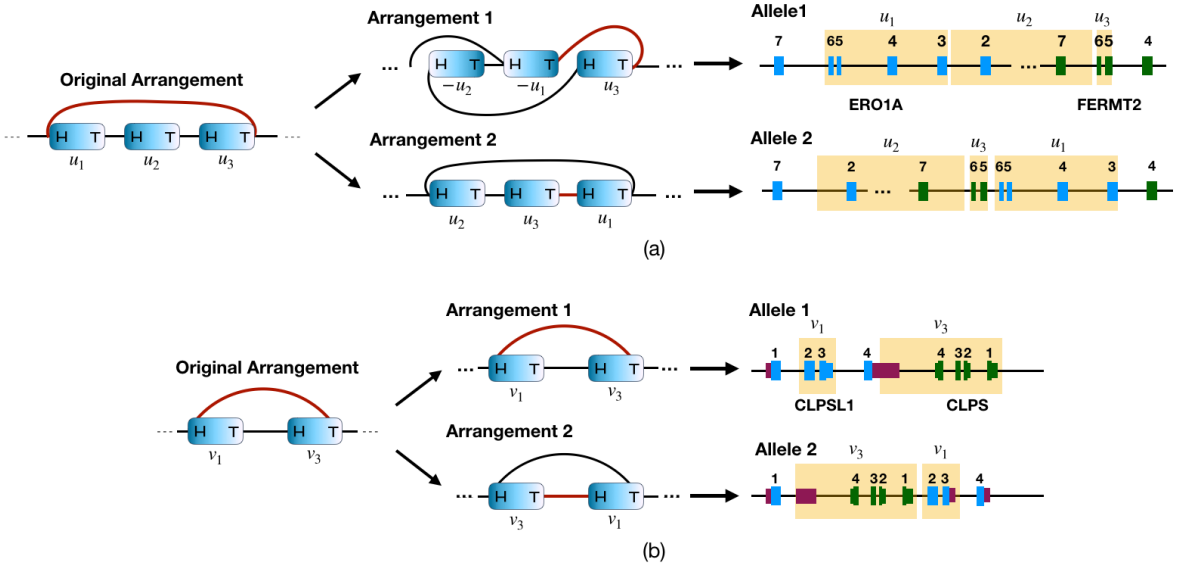

Figure 2.18: Examples on which D-SQUID predicts a validated (a) and an unvalidated (b) TSV event that impacts biologically significant genes. The blue blocks represent segments in the GSG. The red edges mark the discordant edges in the original arrangement. Blue and green blocks mark exons of different genes and dark purple blocks mark UTRs in (b). Regions highlighted in yellow in the gene models mark the corresponding segments in GSG.

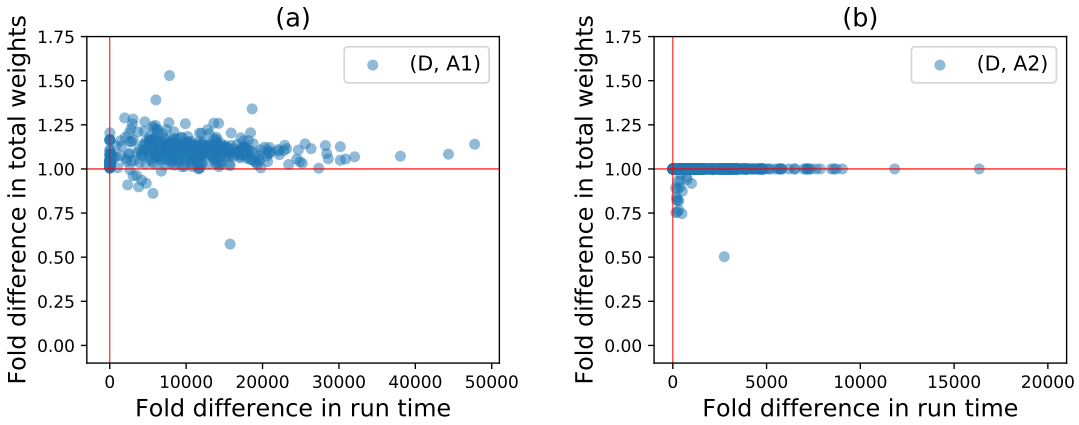

Figure 2.19: Fold differences (ILP/approx) in run time and total weights of concordant edges resolved by D-SQUID, $A1$ and $A2$ on TCGA samples. Horizontal and vertical red lines mark 1.0 on both axes. (a) shows fold differences between D-SQUID and $A1$. (b) shows fold differences between D-SQUID and $A2$.

worst case runtime $O(|V||E|)$ as a subroutine and $A2$ refers to using Algorithm 2 with worst case runtime $O(|V|^2|E|)$ as a subroutine. Both $A1$ and $A2$ solve SCAP by greedily inserting segments into the best position in the current ordering. While $A1$ only looks at the beginning and ending of the ordering, $A2$ looks at all the positions.

In order to compare the performance of approximations to the exact algorithm using ILP, we run D-SQUID, $A1$ and $A2$ on TCGA samples. The algorithms are evaluated on runtime and total weight of concordant edges in the rearranged genomes. "Fold difference" on the axes of Figure 2.19 refers to the ratio of the axis values of D-SQUID over that of $A1$ or $A2$. Both $A1$ and $A2$ output results in a much shorter period of time than D-SQUID. $A2$ achieves better approximation than $A1$, demonstrated by closer-to-one ratio of total concordant edge weight, at a cost of longer run time.

The run time of D-SQUID ILP exceeds one hour on 4.5% of all connected components in all TCGA samples. D-SQUID outputs sub-optimal arrangements in such cases. As a result, approximation algorithms, especially $A2$, appear to resolve more high-weight discordant edges than D-SQUID in some of the samples in Figure 2.19, which is demonstrated by data points that fall below 1 on the y axes. $A1$ resolves more high-weight edges in 10 samples and $A2$ resolves more high-weight edges in 54 samples than D-SQUID.

2.2.8 Conclusions

We present approaches to identify TSVs in heterogeneous samples via the MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM (MCAP). We characterize sample heterogeneity in terms of the fraction of discordant edges involved in conflict structures. In the majority of TCGA samples, the fractions of discordant edges in conflict structures are high compared to HCC samples, which indicates that TCGA samples are more heterogeneous than HCC samples. This matches the fact that bulk tumor samples often contain more heterogeneous genomes than cancer cell lines, which suggests that fraction of conflicting discordant edges is a valid measure of sample heterogeneity.

We show that obtaining exact solutions to MCAP is NP-complete. We derive an integer linear programming (ILP) formulation to solve MCAP exactly. We provide a $\frac{3}{16}$-approximation algorithm for MCAP when the number of arrangements is two ($k = 2$), which runs in time $O(|V||E|)$. It approximates the exact solutions well in TCGA samples.

MCAP addresses this heterogeneity. In 381 TCGA samples, D-SQUID is able to resolve more conflicting discordant edges than SQUID. Since D-SQUID solves MCAP by separating conflicting TSVs onto two alleles, D-SQUID's power to find TSVs generally increases as the extent of heterogeneity increases. In HCC cell lines, D-SQUID achieves better performance than SQUID. Aside from validated TSV events, D-SQUID discovers unvalidated fusion-gene events that impact genes associated with cancer, which requires further investigation.

Several open problems remain. MCAP relies on the number of arrangements (k) to make predictions. It is not trivial to determine the optimal k for any sample. In addition, although MCAP is solved by separating TSVs onto different alleles, there are typically many equivalent phasings. Developing techniques for handling these alternative phasings is an interesting direction for future work. Analyzing the effect of TSVs, especially non-fusion-gene ones, on their impact on cellular functions and diseases is another direction of future work.

Chapter 3

Identifying potential expression estimation inaccuracy by coverage anomaly detection

Transcript expression is another key component of gene transcription status. Expression quantification is used for various analyses, such as differential gene expression [27], co-expression inference [158], disease diagnosis, and various computational prediction tasks [60, 107, 166].

The state-of-art expression quantification methods [15, 59, 67, 78, 79, 90, 113, 129] usually use a generative model to describe the probability of generating the set of RNA-seq reads or fragments. The generative models incorporate a wide range of information, including a set of reference transcript as the most fundamental information as well as base sequencing quality [37] and sequencing biases due to various causes, such as PCR amplification preference and degradation [96]. Even though transcript quantification achieves high accuracy in general, there remain situations where they give erroneous quantifications. For example, most quantifiers rely on a predetermined set of possible transcripts; missing or incorrect transcripts may cause incorrect quantifications. Read mapping mistakes and unexpected sequencing artifacts also lead to misquantifications. Incomplete sequencing bias models can mislead the inferred probability that the reads are generated by each transcript.

To identify the potential misquantification, we introduce an anomaly detection method that detects unexpected coverage patterns in each transcript. RNA-seq reads are generated randomly from expressed transcripts under a probabilistic distribution that describes the experiment protocol, and the coverages along each transcript are expected to agree with the read generation probability distribution. An unexpected coverage pattern, for example, may be that a highly expressed transcript contains an exon with near zero coverage in the middle of the transcript. A possible explanation of this example is that an unannotated transcript without the zero-coverage exon is expressed, in which case the expression unannotated transcript cannot be estimated, and the transcript with the exon is estimated with an erroneous expression. Figure 3.1 shows an illustration of a read generation probability model, an abnormal coverage pattern, and a normal coverage pattern.

When interpreting an expression experiment, particularly when a few specific genes are of interest, the possibility of misquantification must be taken into account before inferences are made from quantification estimates or differential gene expression predictions derived from those quantifications. Statistical techniques such as bootstrapping [3] and Gibbs sampling [45, 79, 157] can

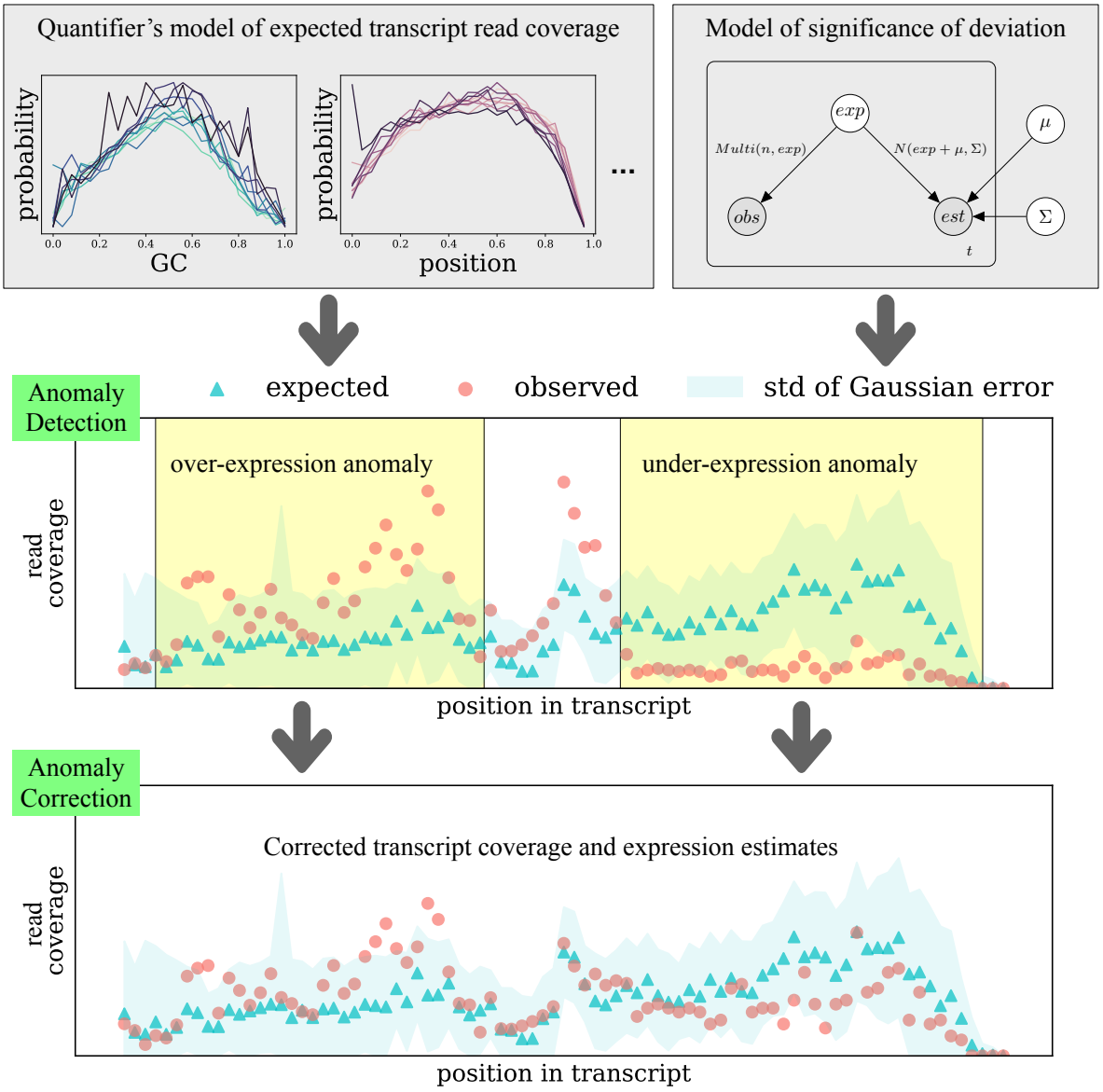

Figure 3.1: The top left panel shows the RNA-seq fragment generation model. Due to biases (such as RNA degradation and PCR amplification preferences), the probability of generating an RNA-seq fragment at different GC percentages and at different positions in a transcript is different. Combining these biases, an fragment generation model can be derived and used to derive an expected coverage along each transcript. The top right panel shows the probabilistic relationship between the true expected coverages, estimated expected coverages, and observed coverages along each transcript. Middle panel shows an example of a transcript containing coverage anomalies. Blue triangles represent the expected probability of generating an RNA-seq fragment at each position, and red circles represent the observed percentage of RNA-seq fragments at each position. The expected and observed coverage largely disagree with each other at the regions marked by yellow boxes. A possible explanation for the coverage disagreement is that an unannotated transcript with an early termination is expressed. The bottom panel shows an example where the expected coverage and observed coverage generally agree with each other.

associate confidence intervals to expression estimates. But confidence interval are not equivalent to quantification error, and it remains a further task to interpret gene expression analyses results using the confidence intervals or incorporate the confidence intervals into expression analyses.

Anomaly detection approaches have been applied in related areas of genomics. Robert and Watson [127] identified uncertainties in gene-level quantification that are due to gene sequence similarity. The quantification uncertainties is related to misquantification, but do not necessarily indicate misquantification. Soneson et al. [143] used a compatibility score between observed and predicted junction coverage to indicate genes with inconsistent splicing junction supports. The idea of identifying the potential errors made in a task regardless the specific algorithm has been applied in other areas in genomics as well. In genome assembly, anomaly detection has been used to detect assembled sequences of low confidence. Genome assembly algorithms seek a set of sequences that can concordantly generate the WGS reads and can be assumed to have near uniform coverage. The assembled sequences that do not fit this assumption can be hypothesized to contain errors and have low reliability [116]. Similarly, anomaly detection in transcriptome assembly identifies unreliable transcript sequences [139]. Low-confidence assembly detection has been used to analyze non-model organisms and incorporated into analysis workflows [17, 43, 180].

Coverage anomalies in expression quantification can be further used to improve the estimated expression. Transcript coverages can be adjusted by changing the proportion of RNA-seq reads assigned to each transcript they can be aligned to, and the estimated expression is changed along with the change of assignment proportion. We developed an RNA-seq re-assignment procedure to maximize the agreement between the assigned transcript coverages and the expected coverages, and an set of adjusted expression estimates is produced accordingly. When it is not possible to obtain concordant coverages with respect to expectation by re-assigning proportions of RNA-seq reads, the coverage anomalies indicate mistakes in other parts of the quantification model, such as incompleteness of the input reference transcript sequences or inaccuracy in the expected coverage. Coverage anomaly detection can be used for evaluating the reconstructed transcript sequences or to guide the development of RNA-seq expression quantification models.

This chapter describes the coverage anomaly detection method, called Salmon Anomaly Detection or SAD, and a case study of using the coverage anomalies to predict functional efficiency of transcription factors. The RNA-seq fragment generation model has a probability assumption that sequencing a fragment from a position of a given transcript follows a multinomial distribution determined by the sequencing biases. The quantifier, Salmon, adopts this assumption and adapts the sequencing bias models depicted by Love et al. [96]. The sequencing bias models describe the coverages we expect to see along each transcript. Large disagreement between the observed coverages and the expected ones indicates that something has gone "wrong" with the quantification for the transcripts. The details of the method is explained in Section 3.1. SAD was published in Cell Systems [97], and the code of SAD is available at `https://github.com/Kingsford-Group/sad`. Anomalies identified by this method are high-level characterizations of violations in the expression generative model regardless of the biological causal events. In Section 3.2, we further analyze the biological implication of coverage anomalies, especially focusing on transcription factors' regulation efficiency. The analysis of coverage anomalies is a joint work with Adrian Lee and Chelsea Chen, and is in preparation for submission.

3.1 Detecting, categorizing, and correcting coverage anomalies of RNA-seq quantification

3.1.1 Overview of anomaly detection and categorization

SAD defines transcripts with anomalous read coverage (Figure 3.2) as those for which the observed coverage distribution contains a significantly over-expressed or under-expressed region compared to the expected coverage. Both the observed and the expected distribution are calculated by the Salmon [113] (or RSEM [78]) quantifier. The observed distribution is the weighted number of reads assigned to each position in the transcript as processed by the quantifier. The expected distribution estimated by the quantifier is the probability of generating a read at each position: Salmon's bias model uses the surrounding GC content, the sequence k-mers, and the read position; RSEM models bias using the read position. The anomaly score can be confounded by either a low expression abundance or an estimation error of the expected distribution. To remove the confounding effect, we model the anomaly score probabilistically and use the empirical p-value to determine whether the observed difference is statistically significant and whether the transcript should be labeled as an anomaly.

To apply the anomaly detection and categorization approaches on other quantification software, the quantification software should output the assignment of each read and the sequencing bias model it learns. Different quantification software may output the read assignment and the biases in different format, and converting their output to vectorized observed and expected coverages that SAD can read is required. The Method section summarizes how to convert the output from Salmon and RSEM to the vector of observed coverage and expected coverage. For the other quantification software, a customized processing script may be needed for the format conversion.

SAD gives rise to two outputs: (1) a list of unadjustable anomalies and (2) the adjusted quantification for the adjustable anomalies. Assuming the expected coverage distributions are correct, the unadjustable anomalies are potentially caused by the incompleteness of reference transcriptome. Given a reference transcriptome or a reference splice junctions, we use "unannotated" to describe an item if it does not appear in the reference. The adjustable anomalies are likely caused by the error in the quantification probabilistic model or optimization algorithm.

Anomaly categorization is done by reassigning the reads across the isoforms using linear programming (LP) and checking whether the anomaly score becomes insignificant after the reassignment. Otherwise, it is labeled an adjustable anomaly. The LP also produces a new set of read assignments for the adjustable anomalies. An adjusted abundance estimation is constructed by combining the new read assignments of the transcripts with adjustable anomalies with the original read assignments of the other transcripts. This combined expression quantification is referred to as SAD-adjusted quantification. If the anomaly score remains significant after the reassignment, the anomaly is labeled an unadjustable anomaly.

3.1.2 An anomaly detection score

Definition 9 (Expected coverage distribution)**.** *Given transcript t with length l, and a fragment f that is sequenced from t, the starting position of f is a random variable with the possible*

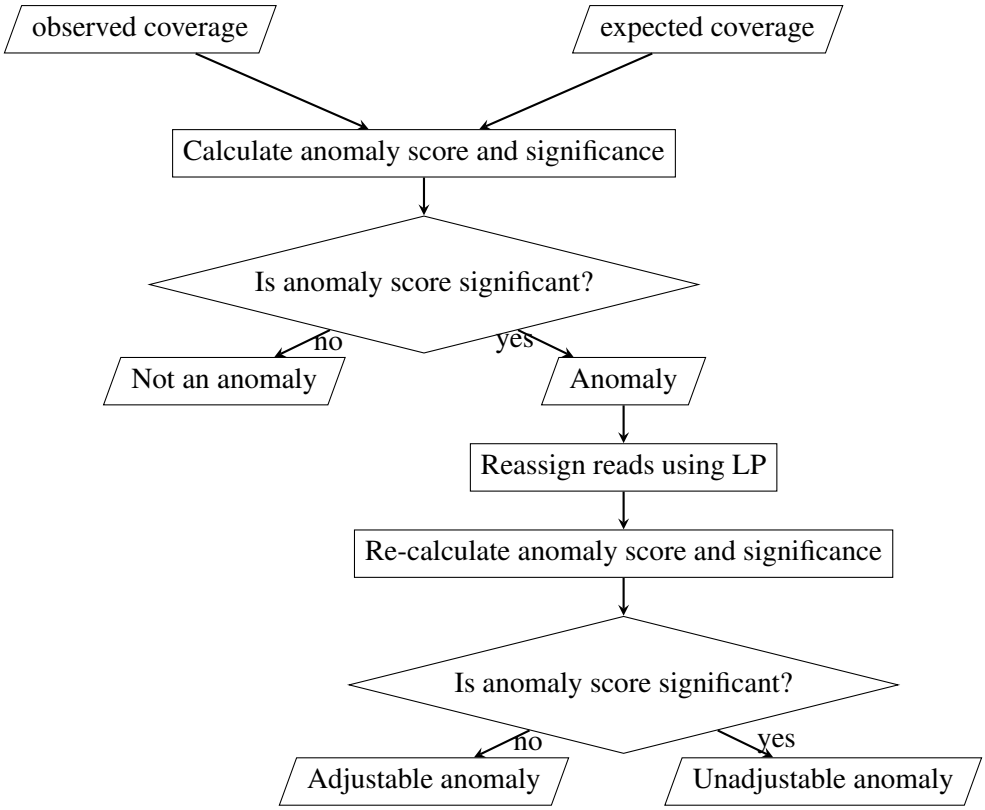

Figure 3.2: Diagram of SAD. SAD detects anomalies by calculating an anomaly score and the significance of its value. To further distinguish the potential cause of the anomalies, it reassigns the reads across isoforms and checks whether the anomaly score becomes insignificant after reassignment. The anomalies whose anomaly scores become insignificant are categorized as adjustable anomalies and considered to be caused by quantification algorithm mistake. The anomalies whose anomaly scores remain significant are categorized as unadjustable anomalies and considered to be caused by external reasons. When the expected coverages are accurate, the external reason is likely the incompleteness of the reference transcriptome.

positions $\{1, 2, 3, \cdots, l\}$ *as its domain. The expected coverage distribution of* t *is the probability distribution of the starting position of any fragment* f*. The expected coverage distribution for each transcript* t *sums to 1.*

With a non-zero fragment length, the viable starting position excludes the last several positions in the transcript. Given a minimum fragment length, it is not possible for a fragment to start at a position within a distance of the minimum fragment length to the end of the transcript. The probability of such positions is set to 0. After aligning and assigning the sequencing reads to transcripts, the number of fragments starting at each position can be counted; this is referred to as the observed coverage. The observed coverage can be converted to distribution by normalizing the coverage to sum to 1. The normalized observed coverage is called the observed coverage distribution, which is comparable to the expected coverage distribution.

We use a slightly different definition of coverage from its classic meaning. We define the coverage of each transcript position to be the number of fragments starting at this position, while the classic definition considers the number of fragments spanning the position. We use the fragment start definition for calculating both the observed and the expected coverage distribution. The observed and the expected coverage are comparable if they are calculated using the same definition. Since the fragment length distribution is often assumed to be a Gaussian distribution with a smaller variance compared to the mean, the coverage distribution under the fragment start definition is approximately the same as the one under the classic definition plus a shift.

Definition 10 (Regional over-(under-)expression score)**.** *Given transcript* t *with length* l*, denote the expected coverage distribution as* exp*, and the observed coverage distribution as* obs*, the over-expression score of region* $[a, b]$ *(*$1 \leq a < b \leq l$*) is*

$$O_t(a, b) = \max\left\{\sum_{a \leq i \leq b} (obs[i] - exp[i]), 0\right\} . \tag{3.1}$$

where index i *denotes the positions in the transcript. The under-expression score of region* $[a, b]$ *is*

$$U_t(a, b) = \max\left\{\sum_{a \leq i \leq b} (exp[i] - obs[i]), 0\right\} . \tag{3.2}$$

The over-expression and under-expression scores are defined as the probability difference between the observed coverage and the expected coverage distribution within region $[a, b]$. The probability difference represents the degree of inconsistency between the two distributions at the given region. The scores indicate the fraction of reads to take away (or add to) from the region in order for the two distributions to match each other.

Definition 11 (Transcript-level anomaly score)**.** *For a transcript* t *with length* l*, the over-expression anomaly of the transcript is defined as*

$$OA_t = \max_{1 \leq a < b \leq l} O_t(a, b) . \tag{3.3}$$

The under-expression anomaly of the transcript is defined as

$$UA_t = \max_{1 \leq a < b \leq l} U_t(a, b) . \tag{3.4}$$

These transcript-level anomaly scores are defined by the largest over- or under-expression score across all continuous regions.

3.1.3 Probabilistic model for coverage distribution

The value of the anomaly score cannot be directly used to indicate an anomaly because its value can be confounded by transcript abundances and the estimation error of the expected distribution. When there are only a few reads sequenced from the transcript, randomness in read sampling can dominate the observed distribution. Because of this, the observed distribution will have large fluctuations along the transcript positions, and thus appear to have a large deviation from the expected distribution. In addition, when the estimation of the expected distribution is inaccurate, the difference between the two distributions can also be large. To address these two confounding factors, we model the relationship between the coverage distributions using a probabilistic framework and calculate the p-value of the anomaly score. With the statistical significance of an anomaly score, we are able to distinguish between true quantification anomalies and randomness from known confounding factors.

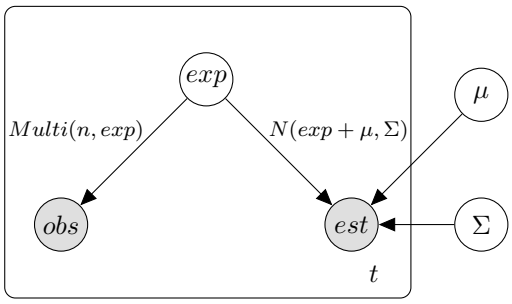

Figure 3.3: The probability model of the expected distribution, the observed distribution, and the estimator of the expected distribution. exp is the expected coverage, obs is the observed coverage, est is the estimation for the expected coverage. Here, exp is a hidden variable, while obs and est are observed. obs follows a multinomial distribution parameterized by the number of reads n and the expected coverage exp. est follows a Gaussian distribution with mean shift μ and covariance matrix Σ. We assume that the estimation errors of the expected coverage have the same pattern for all transcripts, and therefore μ and Σ are shared among all transcripts.

We model the value of the anomaly score probabilistically given the two confounding factors (Figure 3.3). We use the model to indicate the distribution of the anomaly score under the null hypothesis that it is not a true anomaly. For the transcript abundance confounding factor, we assume the observed distribution is generated from the hidden expected distribution through a multinomial distribution parameterized by the given number of reads, n:

$$obs \sim multinomial(n, exp). \tag{3.5}$$

For the estimation error of the expected distribution, we assume the error in the expected distribution is Gaussian with mean μ and covariance Σ. Let est to be estimation of the expected

distribution and let exp be the true hidden expected distribution, the estimation error follows:

$$est - exp \sim N(\mu, \Sigma). \tag{3.6}$$

We further assume that the Gaussian estimation error is generally the same across all transcripts. In practice, transcripts have different lengths and the Gaussian error vectors differ relative to the lengths. We therefore separate positions in each transcript into several bins and transcripts with similar lengths have the same number of bins. A shared mean shift parameter μ and covariance Σ is estimated for the transcripts with the same number of bins.

The variables and parameters of the model (Figure 3.3) can be retrieved or estimated as follows. obs refers to the observed distribution and can be retrieved from the quantification algorithm (Section 3.1.14). est refers to the estimation of the expected distribution, which is processed from the bias correction result of the quantification (Section 3.1.14). exp stands for the expected coverage distribution that is latent. μ and Σ in the probability could be estimated with a Bayesian estimator or maximum a priori (MAP) estimator with a likelihood function. Using subscript t to represent transcripts, the likelihood function is

$$L(\mu, \Sigma) = \prod_t \int_{exp_t : exp_t \geq 0, \sum exp_t = 1} \mathbb{P}(obs_t \mid exp_t)\mathbb{P}(est_t \mid exp_t, \mu, \Sigma)\mathbb{P}(exp_t)d(exp_t)\,. \tag{3.7}$$

However, the above likelihood function does not have a closed form solution and may require using expectation maximization (EM) for optimization. Instead, we estimate μ and Σ using the following approximation: the multinomial distribution for the observed coverage can be approximated by a Gaussian distribution when the number of reads n is large enough:

$$obs \sim Multi\,(n, exp) \overset{n \longrightarrow \infty}{\longrightarrow} N\left(exp, \frac{f(exp)}{n}\right) \tag{3.8}$$

where $f : \mathbb{R}^m \longrightarrow \mathbb{R}^{m \times m}$ maps the m-dimension probability vector of the multinomial distribution into the covariance matrix of the approximating multi-variate Gaussian distribution. Therefore, the difference between obs and est can be approximated by the following Gaussian distribution

$$est - obs \sim N\left(\mu, \Sigma + \frac{f(exp)}{n}\right) \overset{n \longrightarrow \infty}{\longrightarrow} N(\mu, \Sigma)\,. \tag{3.9}$$

We therefore approximate μ and Σ by selecting transcripts with enough reads for each length group, and fit a Gaussian distribution to $est - obs$ of the selected transcripts.

This probabilistic model serves as the null model that assumes the transcript is not an anomaly. That is, the model describes the distribution of the anomaly score under the case where the deviation between the observed and the expected distribution is only due to the two confounding factors: read sampling randomness of sequencing and the estimation inaccuracies of the expected distribution. When the deviation is so large that this null model cannot explain it, we attribute the deviation to an anomaly. To determine whether the deviation is so large that it is unlikely to be observed under the null model, a p-value is calculated. The details of this calculation are explained below.

3.1.4 Statistical significance of the anomaly score

The statistical significance of a value of the anomaly score is the probability of observing an even larger anomaly value given the probabilistic model. Let $O_t(a,b)$ and $U_t(a,b)$ be the random variables of the regional over- and under-expression score of region $[a,b]$, and let $o_t(a,b)$ and $u_t(a,b)$ be the corresponding observed values. Similarly, let OA_t and UA_t be the random variable of transcript-level anomaly score, and oa_t and ua_t be the corresponding observed values. The p-values for a regional over- and under-expression score are

$$\begin{aligned} p\text{-}value \text{ of } O_t(a,b) &= \mathbb{P}(O_t(a,b) > o_t(a,b) \mid exp, n, \mu, \Sigma) \\ p\text{-}value \text{ of } U_t(a,b) &= \mathbb{P}(U_t(a,b) > u_t(a,b) \mid exp, n, \mu, \Sigma) \end{aligned} \tag{3.10}$$

where exp, n, μ and Σ are defined as in Figure 3.3. The p-values for transcript-level over- and under-expression anomaly score are

$$\begin{aligned} p\text{-}value \text{ of } OA_t &= \mathbb{P}\left(OA_t > oa_t \mid exp, n, \mu, \Sigma\right) \\ p\text{-}value \text{ of } UA_t &= \mathbb{P}\left(UA_t > ua_t \mid exp, n, \mu, \Sigma\right) . \end{aligned} \tag{3.11}$$

The statistical testing of the transcript-level anomaly score is more strict to the null hypothesis than that of the regional one, and tends to have a larger p-value. Given transcript t and the largest over-expressed region $[i,j]$, the following inequality between the two p-values holds:

$$\begin{aligned} p\text{-}value \text{ of } OA_t &= \mathbb{P}\left(\max_{1\le a<b\le l} O_t(a,b) > oa_t \mid exp, n, \mu, \Sigma\right) \\ &= \mathbb{P}\left(\max_{1\le a<b\le l} O_t(a,b) > o_t(i,j) \mid exp, n, \mu, \Sigma\right) \\ &\ge \mathbb{P}(O_t(i,j) > o_t(i,j) \mid exp, n, \mu, \Sigma) \\ &= p\text{-}value \text{ of } O_t(i,j) . \end{aligned} \tag{3.12}$$

Conceptually, because the whole transcript contains multiple regions that may have a large over- (under-) expression score, it is easier to observe a large over- (under-) expression score when we look at all possible regions compared to when we focus on only one specific region. From the perspective of statistical testing, the p-value of OA_t and UA_t tend to be larger and less significant than those of $O_t(a,b)$ and $U_t(a,b)$ for any region $[a,b]$. Taking advantage of the different level of strictness about the null model, we use the significance of O_t and U_t for the initial selection of anomalies to adjust read assignment, and use the significance of OA_t and UA_t for the final selection of anomalies within the unadjustable anomaly category.

The p-value of both anomaly scores can be calculated empirically. Specifically, the hidden expected coverage can be sampled from the estimation using multi-variate Gaussian distribution, and the observed coverage can be sampled from the new hidden coverage using multinomial distribution. The null distribution for $O_t(a,b)$, $U_t(a,b)$, OA_t and UA_t can be generated using the sampled observed and hidden expected coverage. The empirical p-value is the portion of times that the anomaly scores exceed the observed valued in the null distribution.

We also derive a numerical approximation for the p-value of regional anomaly score. Empirical p-value calculation requires sampling distributions from a multinomial or multi-variate

Gaussian distribution multiple times, which takes a long time computationally. A numerical approximation without sampling can greatly reduce the calculation time. Denote the region as $[a, b]$ and the current under-expression anomaly score as v. The significance of the over- (under-) expression score under regional null distribution is given by

$$\begin{aligned} p\text{-}value \text{ of } U_t(a,b) &= \mathbb{P}\left(\sum_{i=a}^{b}(exp[i]-obs[i]) > v \mid \sum_{i=a}^{b} est[i]\right) \\ &= \mathbb{P}\left(\sum_{i=a}^{b} obs[i] < \sum_{i=a}^{b} exp[i] - v \mid \sum_{i=a}^{b} est[i]\right) \\ &= \int_x GaussianPDF\left(x \mid \sum_{i=a}^{b} est[i], \mu, \Sigma\right) \mathbb{P}\left(\sum_{i=a}^{b} obs[i] < x - v\right) dx \\ &= \int_x GaussianPDF\left(x \mid \nu, \sigma\right) BinomCDF\left(n * (x-v) \mid n, x\right) dx \end{aligned} \tag{3.13}$$

where $x = \sum_{i=a}^{b} exp[i]$, $\nu = \sum_{i=a}^{b}(est[i] - \mu[i])$, $\sigma = \sum_{i=a}^{b}\sum_{j=a}^{b}\Sigma[i,j]$ and n is the number of reads assigned to the transcript. In the numerical approximation, the function inside the integral is approximated by a step function with small step sizes of x and the integral is approximated by summing up the area under the step function. Since the regional anomaly score focuses on a fixed region, the multinomial distribution can be collapsed into binomial distribution to represent the probability of generating a read from that region. The multi-variate Gaussian distribution can also be collapsed to a single-variate Gaussian distribution to present the expected estimation bias and variance of the region. With all multi-variate distributions collapsed into single-variate distributions, it is feasible to numerically calculate the integral in equation (3.13). In SAD, the p-value of the regional over- (under-) expression score is always calculated using the numerical approximation, while the p-value of the transcript-level anomaly is calculated empirically by sampling.

In practice, we do not calculate the p-value for transcripts with very low abundance. When the randomness of read sampling is very large, we simply assume that the p-value will be dominated by the randomness instead of anomalies. We only calculate a p-value for transcripts with average base pair coverage > 0.01. Using a threshold of 0.01 is equivalent to requiring that on average at least one read is sequenced for every 100 base pairs.

Benjamini-Hochberg correction is used to control the rate of falsely discovered transcripts with regional or transcript-level expression anomaly. A threshold of 0.05 is used in the regional anomaly score. For transcript-level anomalies, 0.01 is used as the threshold. The varied thresholds are set according to their separate purposes: regional anomalies are the initial candidates and do not need to be as precise; after read reassignment, the transcript-level anomalies are the final predictions of unadjustable anomalies and require higher precision.

3.1.5 Categorizing anomalies by read reassignment

We categorize the causes of anomalies into whether or not they are caused by read assignment mistakes of the quantifier's probabilistic model. This is done by seeking an alternative read

assignment for the transcripts with significant regional anomaly score to reduce the inconsistency with the expected coverage.

We use linear programming (LP) to reassign the reads in anomalies. The LP formulation tries to use a linear combination of the expected distributions to explain the aligned reads. By explicitly using the expected coverage to re-distribute the observed number of reads, the deviation between the observed and the expected distribution after the re-distribution is naturally reduced. Accordingly, the anomaly score will decrease and the p-value will increase. We apply LP redistribution separately for each gene since most mis-assignments of reads by the quantifier occur among isoforms of the same gene rather than across genes and gene-level expression estimation is more accurate than isoform-level quantification [28, 142].

The formulation of the LP is

$$\begin{aligned} \min_{\{\alpha_t\,:\,t\in T\}} \quad & \left\| \sum_t \alpha_t exp_t - \sum_t obs_t \right\|_1 + \sum_{j\in J} \left\| \left(\sum_t \alpha_t \delta_t^j exp_t - \sum_t obs_t^j \right) \cdot P^j \right\|_1 \\ s.t. \quad & \alpha_t \geq 0 \quad (\forall t \in T) \end{aligned} \tag{3.14}$$

where t is the index for transcript set T and j is the index of splicing junction set J. Let n be the length of the unique exon positions of the gene. $exp_t \in \mathbb{R}^n$ is the expected coverage distribution (normalized) for transcript t under gene level coordinate. $obs_t \in \mathbb{R}^n$ is the observed coverage (unnormalized) for transcript t under gene level coordinate. $obs_t^j \in \mathbb{R}^n$ is the observed coverage of reads that are assigned to transcript t and spanning junction j. δ_t^j is an indicator that takes value 1 if transcript t has splicing junction j and 0 otherwise. $P^j \in \{0,1\}^n$ indicates which positions are considered close to junction j. Specifically, entries of P^j that represent positions 50 bp to the $5'$ side of the splicing junction position are 1 and the rest are 0. "$\cdot$" is the dot product.

In the LP objective function multiple isoforms of various lengths are included in the same matrix expression. A coordinate conversion is needed to adjust the coverages of multiple isoforms to have the same length. Because each reassignment is performed on isoforms within the same gene, the coverage in transcript coordinates is converted to gene coordinates. In the gene coordinates, each nucleotide is indexed in the sequence of the concatenation of unique exons (or subexons) of the gene. For a given transcript, the coverage is set to 0 for the exons it does not contain.

Let $I_1 = \|\sum_t \alpha_t exp_t - \sum_t obs_t\|_1$ be the first term in the objective function. This is the main minimization goal to reassign reads to isoforms according to their expected coverage distribution. $\sum_t obs_t$ is the aggregated read coverage along the gene. Under the assumption of correct gene-level read assignment but deviated transcript-level read assignment, obs_t may not represent the correct read coverage of transcript t, but $\sum_t obs_t$ represents the correct coverage of the gene. This term seeks to use a linear combination of expected coverage distributions to explain the observed gene coverage.

Let $I_2 = \sum_{j\in J} \left\| \left(\sum_t \alpha_t \delta_t^j exp_t - \sum_t obs_t^j \right) \cdot P^j \right\|_1$ be the second term in the objective function. This term serves as a penalty on the coverage inconsistency around each splicing junction. Because the coverages store only the fragment start positions but not the junction spanning information, a fragment aligning onto a retained intron may have the same starting position of another fragment spanning a splicing junction. Thus an additional penalty is added to control the assignment of junction-spanning reads. The penalty imposed by I_2 encourages that the coverage

of the junction-spanning reads should be explained by a linear combination of the expectation from transcripts with the junction. When transcript t does not contain splicing junction j, we set $\delta_t^j = 0$ to make sure that transcript t has no contribution to the junction coverage. The start positions of the junction-spanning reads are usually near the $5'$ side of the junction. Start positions separated from the junction can contain reads both spanning and not spanning the junction. We specify a 50 bp window to the $5'$ side of the junction to enforce that penalty to be restricted to the most relevant positions to each splicing junction.

Variables α_t stand for the expected number of expressed reads from transcript t. To obtain the actual number of reads reassigned to transcript t at position k, we re-distribution the junction reads and non-junction reads in proportion to α_t. Specifically, the reads starting at position k and spanning junction j are assigned to transcript t with weight $\left(\sum_{t'} obs_{t'}^j[k]\right) \frac{\alpha_t \delta_t^j exp_t[k]}{\sum_{t'} \alpha_{t'} \delta_{t'}^j exp_{t'}[k]}$. Let $n_t[k]$ be the sum of weights assigned to t at position k. The actual total number of reads reassigned to transcript t is $\sum_k n_t[k]$.

After adjusting read assignments using the LP, some transcripts have an insignificant transcript-level anomaly score. These transcripts are labeled "adjustable anomalies" and are considered to have misquantifications due to quantification algorithm mistakes. On the other hand, if the transcript-level anomaly scores are still significant, the corresponding transcripts are labeled "unadjustable anomalies". Assuming the expected distributions are estimated with reasonable accuracy, we suspect the unadjustable anomalies are affected by the expression of unannotated transcripts' expression and indicate incompleteness of the reference transcriptome. Benjamini-Hochberg correction is used to adjust the p-value of transcript-level anomaly score to control for the false positive labeling of anomalies for all transcripts.

3.1.6 Reducing number of transcripts involved in reassignment

In practice, we try to keep the number of transcripts involved in the LP as small as possible. When the quantification of a transcript is good enough, reassigning the reads may lead to a decrease of quantification accuracy. The correctness of the LP reassignment largely depends on the accurate estimation of the expected distribution. However, the accuracy assumption of the expected distribution may not hold for all transcripts. An inaccurate estimation at some positions for one transcript can perturb the reassignment result across all involved isoforms. The perturbation can be large when the coefficient matrix in the LP has a large condition number (called ill-conditioned), which tends to occur more often as the number of isoforms involved in the LP increases. The ill-condition will make the output very sensitive to a small change or error of the input distributions. To reduce this problem in the LP reassignment, we only apply the LP reassignment on a small number of isoforms and reset the other isoforms to the quantifier's read assignments. The choice of isoforms is determined by the following principle: the largest number of transcripts should have insignificant regional anomaly scores across all regions while at the same time minimizing the number of isoforms involved in the LP.

To obtain the largest number of transcripts with insignificant regional anomaly scores, we initially run the LP using all transcripts. Then we exclude each transcript one-by-one from the LP. If excluding a transcript from the LP does not change the set of transcripts with insignificant regional anomaly scores, the transcript is excluded forever from the LP, otherwise, it is kept in

the LP. When excluding any transcript from the LP increases the number of transcripts with significant regional anomaly scores, the iterative process is terminate and the final set of transcripts involved in LP is determined.

3.1.7 Results: examples of detected anomalies

We provide some examples of the detected anomalies found by applying SAD to 30 GEUVADIS [73] and 16 Human Body Map datasets [1]. The 30 GEUVADIS samples are ones used in the work of [113], in which 30 lymphoblastoid cell lines from the Toscani in Italia (TSI) population are sequenced at two different sequencing centers. The Human Body Map project data consists of 16 samples each from a different tissue, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells.

SAD identifies an adjustable anomaly in the gene *TMEM134* in the kidney sample from the Human Body Map dataset. The *TMEM134* gene encodes a trans-membrane protein that is associated with Parkinsons disease [65]. One isoform (ENST00000545682.5) of this gene has an under-expression anomaly after its first splicing junction (Figure 3.4A). See Appendix Figure 3.8 for IGV visualization. This under-expression anomaly can be adjusted by reassigning reads to this isoform from another isoform, ENST00000537601.5 (Figure 3.4B). The expression estimates are changed according to the adjustment: before adjustment, the isoform with the under-expression anomaly has a 1.4 times larger expression than the other isoform, and after adjustment, the ratio of expression is enlarged to 9.0. The two isoforms are different from each other by two splicing junctions (Figure 3.4C). With the quantification more consistent with the read coverage of both isoforms, the analysis on the function and effect of the alternative splicing may benefit.

Another example of an adjustable anomaly is within the *BIRC3* gene in one GEUVADIS sample. This gene is involved in apoptosis inhibition under certain conditions. The second half of the isoform ENST00000532808.5 is under-expressed under Salmon's read assignment (Figure 3.4D). See Appendix Figure 3.9 for IGV visualization. Reassigning the reads between this isoform and another isoform, ENST00000263464.7, removes the under-expression phenomenon (Figure 3.4E) and at the same time alters the expression level of both isoforms. The original expression abundances of the two isoforms were similar to each other, but after SAD adjustment the expression of ENST00000263464.7 is 3 times that of ENST00000532808.5. The two isoforms are different in their starting and ending positions but have the same set of internal exons. The protein domains between the two isoforms are the same according to Pfam [35] annotations (Figure 3.4F) but the 5' and 3' UTR sequences are different.

SAD also reveals unadjustable anomalies in isoforms that have a different set of protein domains from the other isoforms of the same gene. For example, gene *UBE2Q1* and gene *LIMD1* in the heart sample of the Human Body Map dataset contain unadjustable anomalies (Figure 3.5A-B, Appendix Figure 3.10). In both genes, the protein domains in the anomalous isoform are different from those in the other annotated isoforms: ENST00000292211.4 of gene *UBE2Q1* is the only annotated isoform that has ubiquitin-conjugating enzyme domain, and ENST00000273317.4 of gene *LIMD1* contains three zinc-finger domains annotated by Pfam while the other isoforms only contain two or zero. The over-expressed regions of both genes

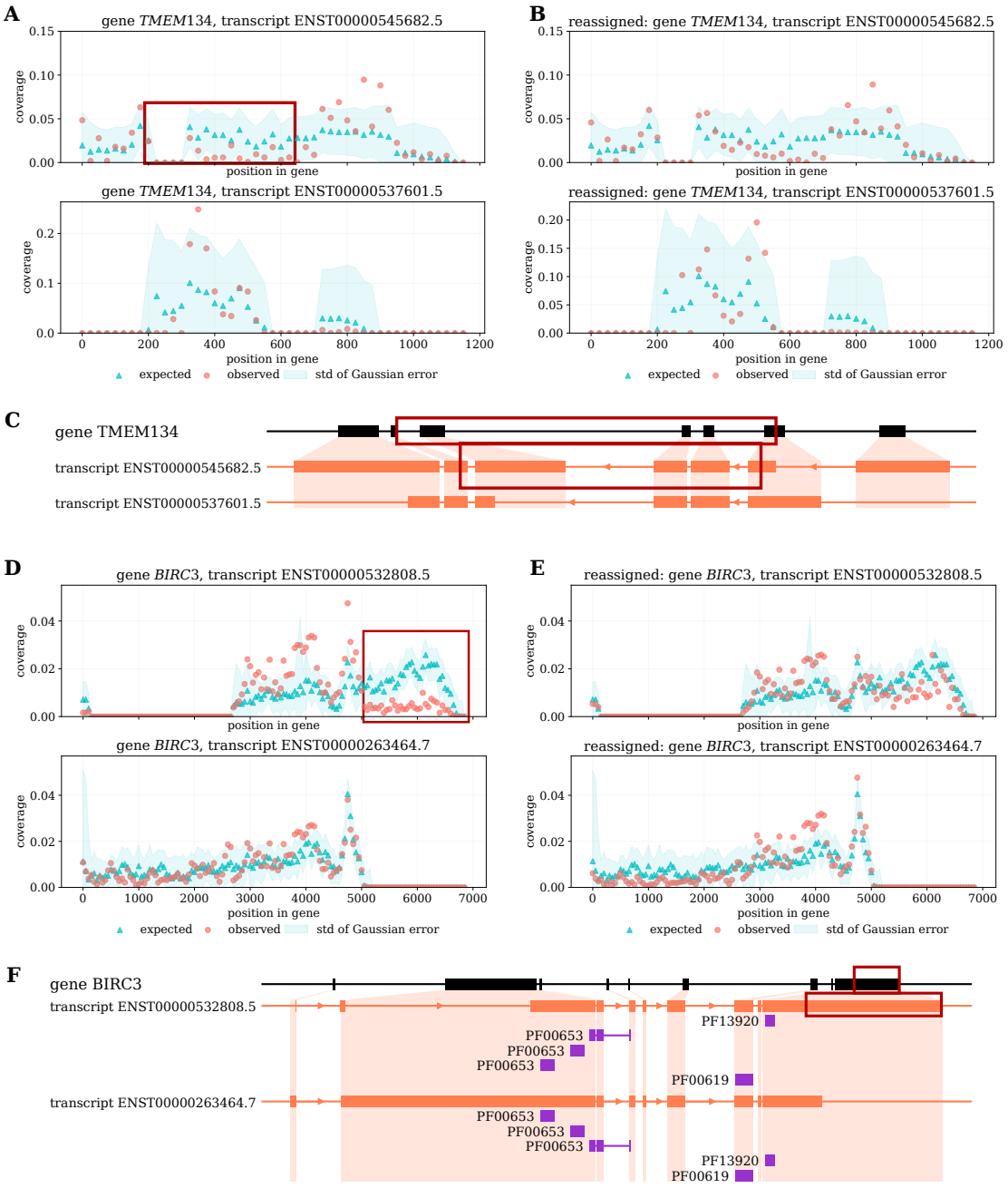

Figure 3.4: Examples of adjustable anomalies. (A)–(C) The kidney sample of the Human Body Map dataset. (A) Red and blue points are the observed and expected coverage distribution before SAD adjustment. The expected distribution is the Salmon-estimated expected distribution subtracted by the mean of Gaussian error. Each point is a 50 bp bint. There is an under-expression in transcript ENST00000545682.5 after its first splicing junction (top), marked by the red box. Another transcript is involved in the adjustment (bottom). (B) The distributions of the same pair of transcripts after SAD adjustment. (C) The protein domain annotation of the two transcripts. Exon regions are expanded and intron regions are reduced for readability purpose. The under-expression anomaly region is marked by the red boxes. (D)–(F) A sample from the GEUVADIS dataset (accession ERR188088). Each panel has the same axes and color coding as previous three panels.

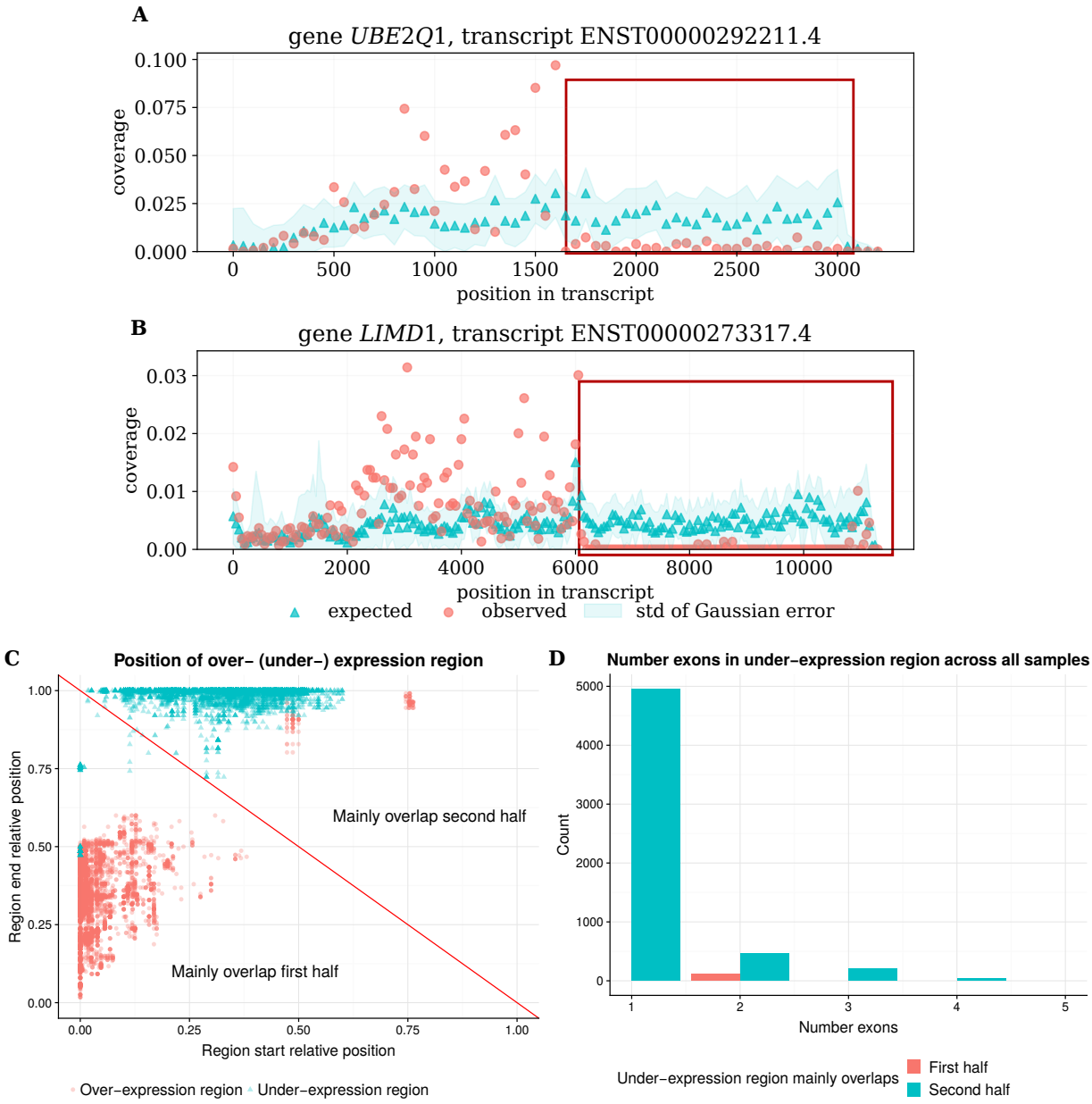

Figure 3.5: (A) An example of unadjustable anomaly of gene *UBE2Q1* (B) An example of unadjustable anomaly of gene *LIMD1*. Both examples are found in the heart sample of the Human Body Map dataset. Red and blue points are the observed and expected coverage distribution of the anomaly transcripts, and the blue shade is the standard deviation of the expected distribution estimation. The red box indicates the under-expression anomaly region. (C) The start and end proportion of the over- (under-) expressed regions of common anomalous transcripts. The red diagonal line separates between anomalies of which the over- (under-) expression regions mainly overlaps with the first half ($5'$ half), and the second half ($3'$ half) of the transcripts. The over-expressed regions mainly overlap with the first half of the anomalous transcript, and the under-expressed regions mainly overlap with the second half. (D) Histogram of the number of exons spanning the under-expressed region of the common anomalies. Y-axis is the count summed over all 46 samples. The under-expressed region usually only contain one or a partial exon.

contain the full set of protein domains, while parts of 3' UTRs are barely expressed for both anomalies. The large unexpressed regions suggest the unadjustable anomalies are unlikely to be explained by the inaccuracy of the expected distribution, instead they imply the existence of unannotated isoforms. The Scallop transcript assembler [137] is able to assemble a unannotated sequence of *LIMD1* without the under-expressed region, thus supporting this detected anomaly. Studies have shown that alternative cleavage can generate isoforms with various $3'$ UTRs in some cells [52] and the length of the $3'$ UTR is correlated with the transcript degradation rate [178]. The detected unadjustable anomalies may be an example of such alternative cleavage or lower degradation rate.

3.1.8 Results: adjustable anomalies give an adjusted quantification that reduces false positive differential expression detections

The adjusted quantification of SAD reduces the number of false positive calls in detecting differentially expressed transcripts. Previously, [113] showed that the 30 TSI samples from GEUVADIS dataset [73] likely do not have differentially expressed transcripts, but quantification mistakes can lead to false positive differential expression (DE) predictions across sequencing center batches. They also showed that a more accurate quantification can reduce the number of false positive detections. We apply SAD to the same samples and compare the number of differentially expressed transcripts detected using Salmon's original quantification and the SAD-adjusted quantification. There are 1938 – 3385 adjustable anomalies within each sample, each of which have SAD-adjusted expression estimates. The estimates for the rest of the transcripts remain the same as Salmon. Differential expression is inferred by DESeq2 [95] on the transcript level. With Salmon expression estimates, 6088 – $13\,555$ transcripts out of $198\,541$ are detected to be differentially expressed across the two sequencing centers under various FDR thresholds. With SAD-adjusted quantification, the relative number of DE transcripts is reduced by 2.29% – 3.84% (Table 3.1). This provides evidence that these anomalies are likely real misquantifications that are correctable using a different read reassignment procedure.

FDR	Salmon	SAD-adjusted	percentage reduced
0.01	6088	5854	3.84%
0.05	10132	9907	2.46%
0.1	13555	13316	2.29%

Table 3.1: The number of DE transcripts detected at a given FDR threshold by using Salmon and SAD-adjusted quantification. Among the 30 samples, there should not be any DE transcripts. With SAD-adjusted expression quantification, the number of false positively detected DE transcripts is reduced.

An isoform of gene *HDAC2* and an isoform of gene *NDUFA13* are two examples of transcripts that have decreased p-value of differential expression after SAD adjustment. Gene *HDAC2* encodes proteins to form histone deacetylases complexes and is important in transcriptional regulation [112]. One of its isoforms, ENST00000519065.5, is differentially expressed

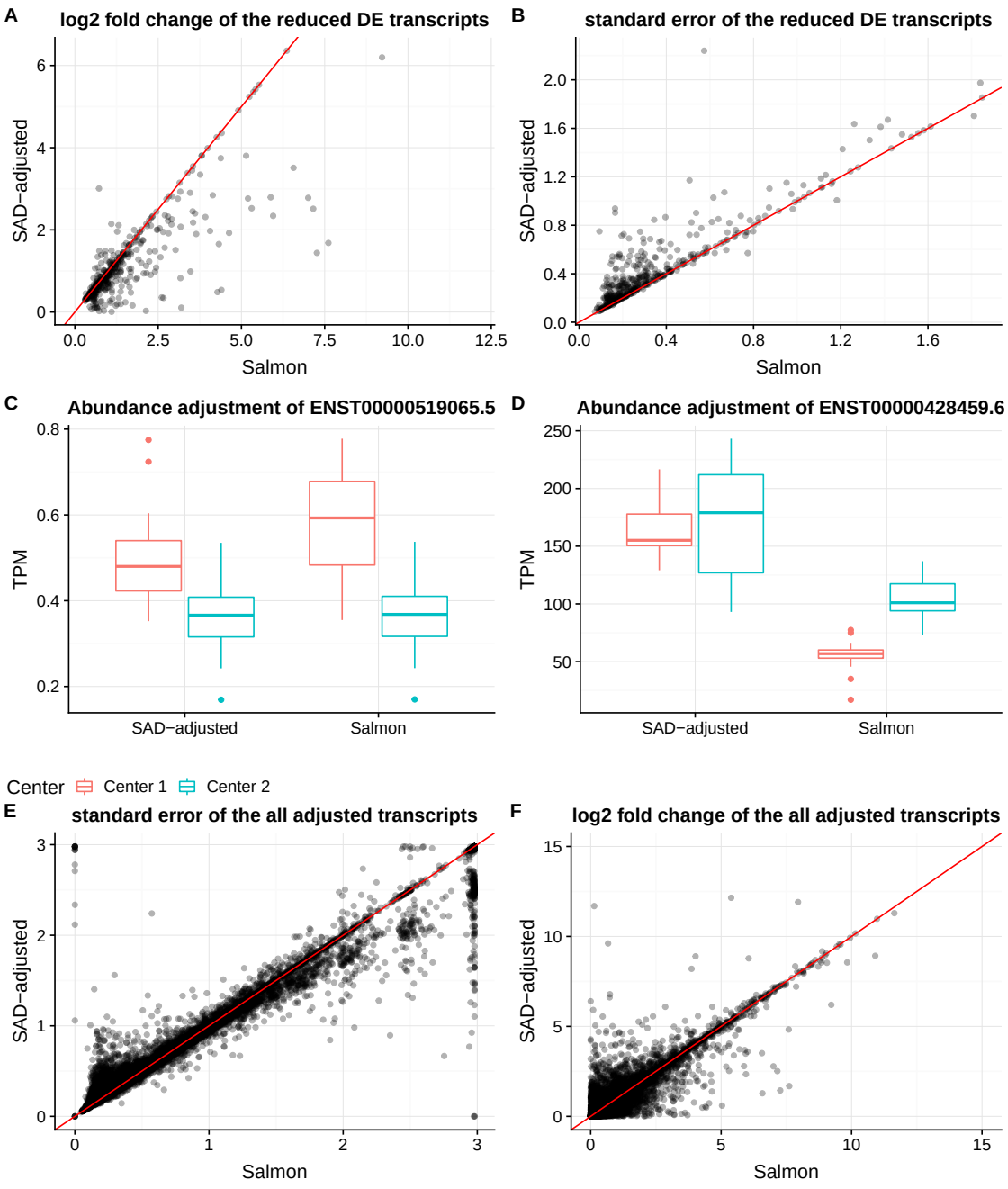

Figure 3.6: Changes in statistics of DE detection by using SAD-adjusted quantification for adjustable anomalies. (A) Absolute $\log_2$-fold change between the two sequencing centers for the transcripts labeled as DE under Salmon but not under SAD-adjusted quantification. (B) Standard error of the $\log_2$-fold change between sequencing centers for the transcripts that are labeled as DE under Salmon but not under SAD-adjusted quantification. (C,D) Two examples of transcript that is detected as DE by Salmon but not detected by SAD-adjusted quantification. Each box indicates the range of estimated expression across RNA-seq samples corresponding to each sequencing center. (E) Standard error of the $\log_2$-fold change for all transcripts under Salmon and SAD-adjusted quantification. (F) Absolute $\log_2$-fold change between the two sequencing centers for all transcripts.

with an adjusted p-value of 0.0008 under Salmon quantification. SAD adjusts its expression by redistributing its reads to the other 16 transcripts of the gene, increasing the p-value of differential expression to 0.075 (Figure 3.6C). With SAD-adjusted quantification, this isoform is not differentially expressed under a p-value threshold of 0.05 or 0.01. The gene *NDUFA13* encodes a subunit of the mitochondrial electron transport chain [112]. Over-expression or under-expression of the gene has been associated with multiple cancer types [100]. Transcript ENST00000428459.6 from this gene was significantly differentially expressed. After SAD reassigns reads from the other transcripts to it, the transcript is no longer differentially expressed (Figure 3.6D).

Whether a transcript is detected to be differentially expressed under SAD-adjusted quantification may be influenced by that only some of the samples undergo the quantification adjustment of the transcript. In the case where the transcript abundances are similar within each condition and are adjusted only in a subset of samples, the within-condition variance may increase, the p-value of DE may increase, and the transcript is less likely to be detected as DE under a given FDR threshold. In this case, DE detection is more conservative by using SAD-adjusted quantification. When the conservation is preferred, especially when trying to avoid uncertain DE calls due to the inconsistencies between the observed and the expected coverage distribution, using SAD-adjusted quantification is helpful. Nevertheless, the influence of the partial adjustment is mild because the majority of the DE predictions under Salmon and SAD-adjusted quantification agree with each other. Many transcripts do not have an increased within-condition variance as what partial adjustment may induce (Figure 3.6E). The switch from DE to not DE after SAD-adjustment is not purely caused by the increase of within-condition variance, but the decrease of across-condition expression differences is also a contributor as well (Figure 3.6A–B). In the case where the abundances are adjusted for all samples in one condition but no samples in the other, it is not predetermined whether the DE detection is more conservative or more aggressive. Whether the transcript is detected as DE depends on whether the adjustment increases or decreases the expression difference between the conditions. Both increase and decrease of expression differences happen with similar frequency empirically (Figure 3.6F).

Occasionally, there are are multiple optimal solutions to the likelihood function of quantification models and the quantifier will output only one of the optimal solutions. The multiple optima scenario is called the non-identifiability problem, and the transcripts with multiple optimal abundances are said to have non-identifiable abundances. However, the majority of anomalies detected by SAD do not suffer from the non-identifiability problem (Appendix Figure 3.11B). Accordingly, the SAD-adjusted quantification is not another optimal solution to Salmon's objective, but rather an assignment under a different model. The quantification improvement using SAD's adjusted anomalies lies in the model of using the expected coverage to explain the observed coverage. See Method Section for how non-identifiable transcripts were detected.

3.1.9 Results: common unadjustable anomalies tend to have an under-expressed region in the 3' exon

Applying SAD reveals 774–1288 unadjustable anomalies per sample on the 30 GEUVADIS samples, and 2029–8269 per sample for the 16 Human Body Map samples. Among the unadjustable anomalies, 88 of them are common in all samples in both datasets (the full list can be downloaded

from https://doi.org/10.5281/zenodo.4048493). The 88 common unadjustable anomalies span 22 chromosomes. The genes they belong to have various numbers of annotated isoforms ranging from 1 to 15. The common unadjustable anomalies generally follow the transcript length distribution of the commonly expressing transcripts (Appendix Figure 3.11A).

For most of the common anomalies, the over-expressed regions tend to mainly overlap with the first half of the transcripts near the $5'$ end (Figure 3.5C). Correspondingly, the under-expressed regions are usually located towards the second half of the transcripts near the $3'$ end. The under-expressed anomaly regions usually only span one exon or a partial exon (Figure 3.5D). Assuming the bias model in Salmon estimates the expected distribution with reasonable accuracy, the unadjustable anomalies are likely to indicate the existence of unannotated transcripts. These unannotated transcripts will share the over-expressed region and exclude the under-expressed region compared to the anomalous transcripts. That is, they will have the same intron chain but different transcript ending locations.

About 40%–60% of the detected unadjustable anomalies have a corresponding unannotated isoform assembled by transcriptome assembly algorithms, specifically StringTie [115] and Scallop [137] (Appendix Figure 3.12A–B). (See Method Section for the details of running transcriptome assembly software.) An assembled isoform corresponds to a predicted unadjustable anomaly if the assembled isoform contains all the splicing junctions within the over-expressed region and excludes at least half of the under-expressed region. Meanwhile, the rest 40%–60% of the unadjustable anomalies do not have a corresponding isoform assembled by transcriptome assemblers. Assuming the expected coverage distribution is modeled correctly, these unadjustable anomalies are likely to indicate true unannotated isoforms that are not able to be detected by transcriptome assemblers. The sensitivity of assembling unannotated transcripts is usually low, which partially explains the difference between the existence of unannotated isoforms indicated by unadjustable anomalies and by transcriptome assembly methods.

While we hypothesize that the unadjustable anomalies are caused by the existence of unannotated transcripts, we cannot rule out the possibility that some of the unadjustable anomalies can be an artifact of inaccurate modeling of the expected coverages or an unsuitable assumption of the Gaussian error of the expected coverages. Neither is it clear whether the unannotated transcripts are natural, well-functioning isoforms, or non-functioning sequences due to errors in transcription, or alternative cleavage and polyadenylation that retain various lengths of UTR [52].

3.1.10 Results: simulation supports the accuracy of SAD for detecting and categorizing anomalies

On simulation data, the predictions of both unadjustable and adjustable anomalies precisely capture the misquantification due to those causes. We created 24 datasets by varying the number of simulated unannotated isoforms, the gene annotations, and the expression matrices. (See Method Section for the details of the simulation procedure.)

Unadjustable anomalies are able to predict the existence of simulated unannotated isoforms that do not contain unannotated splicing locations with 3%–35% higher precision than transcriptome assembly methods (Figure 3.7A–B). Precision is computed as the fraction of "marked" genes that contain simulated transcripts that are unannotated in the reference. For SAD, a gene

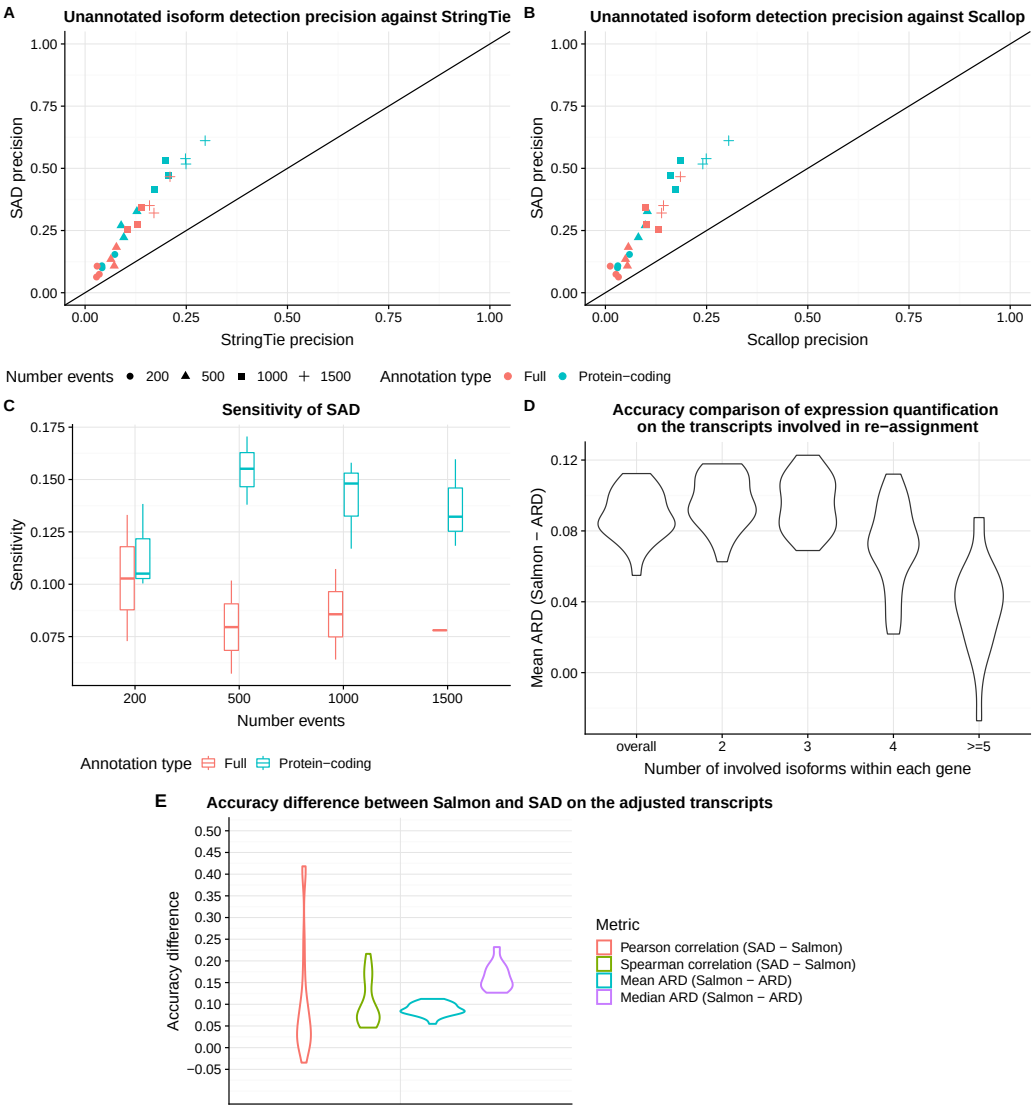

Figure 3.7: Prediction accuracy of transcript expression by SAD-adjusted quantification and of unannotated isoform existence by SAD unadjustable anomalies in simulated data. (A) Precision of unannotated isoform detection using SAD unadjustable anomalies and StringTie assembly. Point color and shape refers to different simulation settings. The simulated unannotated isoforms do not contain unannotated splicing junctions, but only contain unannotated starting / ending sites, or unannotated combinations of known splicing junctions. (B) Precision of unannotated isoform detection using SAD unadjustable anomaly and Scallop. (C) Sensitivity of the unadjustable anomalies of SAD. The boxes and the violins in the next panel indicate the ranges of y-axis values across simulated datasets. (D) Quantification accuracy improvement of SAD compared to original Salmon. Each violin refers to a subset of transcripts where the corresponding genes contain a certain number of isoforms in the adjustment according to the x-axis. "Overall" in the x-axis is the overall mean ARD improvement of all adjusted isoforms. (E) Overall quantification accuracy improvement of SAD compared original Salmon under four metrics. Positive values indicate higher accuracy achieved by SAD.

is marked if it contains a transcript that is detected as an unadjustable anomaly. For the transcript assembler, a gene is marked if it has an assembled transcripts with predicted RPKM $\geq$ a parameter θ, and that transcript either: (1) only uses existing splicing junctions and does not match the intron chains of any existing transcript or (2) matches the intron chain of an existing transcript, but has a starting position or stopping position more than 200 bp away from the matched existing transcript. The parameter θ is chosen so that the transcript assembler marks the same number of genes as SAD does.

Note in this comparison, we compute precision only considering isoforms that use existing splicing junctions in unannotated combinations or with alternative start or termination locations. These are generally the harder transcripts to detect, since for these isoforms, transcript assembly methods can only depend on coverage to assemble transcripts. SAD benefits from using the well modeled expected coverage distribution to identify unadjustable anomalies. On the other hand, the main advantage of SAD is precision, but not sensitivity, because not all unannotated isoforms will significantly alter the coverage of known ones (Figure 3.7C).

In addition, the LP read reassignment is more accurate than the original Salmon quantification [113] on the adjustable anomalies in simulated data (Figure 3.7D–E). The accuracy of quantification is measured by mean ARD (absolute relative difference) [113]. ARD is calculated by taking the absolute difference between the estimation and the true expression and normalizing it by the sum of the estimation and the truth. A smaller value of mean ARD indicates an estimator that is closer to the ground truth. The decrease of ARD on adjustable anomalies is usually more than 0.05. The accuracy improvement of SAD decreases as more isoforms of one gene are involved in the read reassignment. The decrease of improvement is possibly because small estimation errors in the expected distribution are magnified when the LP coefficient matrix used by SAD is large in size and potentially ill-conditioned. When the coefficient matrix is ill-conditioned in the linear system, the output can greatly change even with a small error in the input.

3.1.11 Results: unadjustable anomalies detected based on RSEM have 20% – 50% overlap with those detected based on Salmon

To show the applicability of the anomaly detection method on multiple quantification methods, we apply anomaly detection using the RSEM [78] quantifier and identify unadjustable and adjustable anomalies in the same 30 GEUVADIS samples and 16 Human Body Map samples. See the Method Section for the details of obtaining the expected and observed coverage distributions from RSEM. With these coverages, the anomaly detection and categorization methods we present are able to be directly applied.

About 20%–50% of the RSEM unadjustable anomalies are shared with the ones detected using Salmon [113] (Appendix Figure 3.12C–D). The expected distribution estimated by RSEM only depends on the positional bias and is computed at a coarser resolution than is modeled in Salmon. RSEM does not model sequence-specific or GC content biases. Therefore, it is not surprising that there is a large difference between unadjustable anomalies based on Salmon and those based on RSEM. Indeed, the percentage is much higher than random (hypergeometric test p-value $< 10^{-300}$). These results show that when applied with quantifiers that coarsely esti-

mate the expected distribution, the anomaly detection method can still predict many unadjustable anomalies.

There are 219 – 527 transcripts per GEUVADIS sample and 509 – 1972 per Human Body Map sample that are detected to be unadjustable anomalies only under RSEM quantification. For the ones that are detected as unadjustable anomalies only under Salmon quantification, the number is 258 – 714 per GEUVADIS sample and 1168 – 5657 per Human Body Map sample. The causes of the difference include: (1) Salmon and RSEM estimate different expected coverages but assign similar observed read coverages to the transcript (Appendix Figure 3.13A–B); (2) Salmon and RSEM assign obtain different observed coverages (Appendix Figure 3.13C–F) and the difference remains after SAD read re-assignment; (3) both expected coverages and observed coverages are similar for Salmon and RSEM, but the variances of Gaussian error in the expected distribution estimation are different (Appendix Figure 3.13G–H); (4) a mixture of the above causes. When the cause of different predictions is due to the difference in Gaussian error variances, Salmon tends to predict the transcripts as unadjustable anomalies while RSEM may not. The expected coverage in RSEM is estimated only based on positional bias, which is coarser and usually farther away from the observed read coverage than Salmon's expected coverage. Thus the variance of the Gaussian estimation error is usually larger in RSEM than in Salmon. When the variance of error in expected distribution estimation is larger, the likelihood of observing a large deviation by chance increases and the p-value also increases. The mixture and interplay among the possible causes may be complicated, therefore we do not assign the uniquely detected anomalies to the causes or estimate the weights of the causes.

3.1.12 Results: unadjustable anomalies are supported by long read sequencing data in 1000 Genome samples

To further verify that in real RNA-seq the unadjustable anomalies are likely caused by the incompleteness of the reference transcriptome, we use long-read sequencing evidence to show the existence of unannotated isoforms that are suggested by unadjustable anomalies. In the 1000 Genome [25] samples, 3 trios (9 samples) were sequenced using both short-read RNA-seq [24] and PacBio SMRT technology to obtain expressed full-length transcripts [19]. We apply SAD to the short-read RNA-seq data and compare the detected unadjustable anomalies to sequenced PacBio reads of full-length transcripts. A isoform that is derived from the full transcripts sequences and not included in the reference annotation is considered to correspond to the unadjustable anomaly prediction if it covers 75% of the over-expressed region and excludes 75% of the under-expressed region of the anomaly. Unadjustable anomalies that have corresponding PacBio reads are considered true predictions of the existence of unannotated isoform and are used to calculate precision.

For all 9 samples, the precision of unadjustable anomalies of SAD is within the range of 23% – 32% (Appendix Figure 3.14). A full list of unadjustable anomalies and their correspondence to the long reads can be downloaded from `https://doi.org/10.5281/zenodo.4048493`. The precision is within the range observed in the simulated RNA-seq data. The rest of the unadjustable anomalies are not supported by the long reads. Instead, they may correspond to true unannotated isoforms that are not sequenced by long reads or arise from an inaccurate

estimation of the expected distribution of the anomalous transcripts.

3.1.13 Discussion

We present Salmon Anomaly Detection (SAD), an anomaly detection approach to identify potential misquantification of expression. SAD detects anomalies by comparing the expected and the observed coverage distribution and calculating the significance of the over- or under-expression. SAD also categorizes the anomalies into adjustable anomaly and unadjustable anomaly categories to indicate two possible causes of misquantifications: algorithmic errors and reference transcriptome incompleteness. The categorization is done by reassigning reads across isoforms to minimize the number of significant anomaly scores. We show on simulation data that the detected anomalies and their categorizations are reasonable: the unadjustable anomalies predict the existence of unannotated isoforms (using existing splice junctions) with higher precision than transcriptome assembly methods, and the read reassignment of adjustable anomalies leads to adjusted quantification that is closer to the simulated ground truth compared to the original quantification.

The explanation for LP read assignment leading to a better quantification than Salmon for some transcripts is that the LP focuses only on the base-to-base coverage distribution consistency while Salmon combines multiple aspects into its probabilistic model and also groups reads into equivalent classes. For example, transcript lengths and fragment lengths are considered in its probabilistic model. One equivalence class may include reads starting at various positions, which have various expected coverages. Because Salmon balances these multiple aspects and treats each equivalent class as a unit, it may generate a coverage distribution deviated from the expectation. When this deviation is very large, the quantification results tend to be inaccurate. In the case of a very large deviation, reassigning reads purely based on coverage consistency using the LP leads to a more accurate quantification.

Applying SAD on GEUVADIS and Human Body Map datasets, we are able to identify adjustable and unadjustable anomalies that affect isoforms with different protein domains from other isoforms and isoforms from cell type marker genes. Using the adjusted quantification associated with the adjustable anomalies, the number of false positive predictions of differentially expressed transcripts can be reduced. There are common unadjustable anomalies across all samples. Most of the common unadjustable anomalies have an under-expressed region towards the $3'$ end of the transcript.

SAD is only able to detect the subset of misquantifications that have distorted the observed coverage from the expected one. However, some misquantifications may not alter the shape of the observed coverage distribution. For example, high sequence similarity between a pair of transcripts can also lead to severe misquantification, however, the read coverage can be close to the expectation for both. Alternatively, the coverage distribution of a lowly expressed existing isoform can be affected by a lowly expressed unannotated isoform. In this case, the p-value of the anomaly score may not be significant due to the large fluctuation of the observed coverage due to the low expression. Developing other scores, for example, using transcript similarity or discordant read mapping, could potentially increase the sensitivity and the types of possible misquantification of detection.

Some of the causes of anomalies are not covered by the current anomaly categorization

method. For example, when an anomaly is caused by a mixture of incomplete reference transcriptome and mistakes of the quantification methods, SAD cannot label the cause as the mixture but is only able to attribute to one of the two causes based on the read reassignment outcome. In addition, unadjustable anomalies can be further sub-categorized by whether the corresponding unannotated isoforms are splicing variants or gene-fusions. One contribution of this work is to inspire more systematic investigation of the causes of expression anomalies. Refining the methods to determine the causes of anomalies is a potential direction for future work.

For unannotated isoform detection, only transcript existence is predicted by SAD, not the sequence or exon-intron structure of the unannotated isoforms. Retrieving the exon-intron structure remains a problem. Simply combining the prediction of SAD with the assembled sequences from transcriptome assembly does not solve the problem of reconstructing unannotated isoform sequences. About 40%–60% of SAD's predictions are not assembled by transcriptome assembly methods in the GEUVADIS and the Human Body Map datasets. Incorporating the expected coverage distribution during transcriptome assembly may be a direction to predict the exact exon-intron structure of the unannotated isoforms.

SAD suggests an analysis workflow that contains three steps: quantification, anomaly detection, followed by specialized quantification focusing on the anomalies. The middle step, anomaly detection, and the last step, specialized quantification, can be treated separately and enhancements in either step are needed to improve the accuracy of the adjusted expression estimates. For example, SAD's read reassignment only shuffles the reads across isoforms within the same gene. A better read reassignment across genes can be developed.

An improvement in the accuracy of the approximation of the expected distribution may further increase the accuracy of unannotated isoform prediction and re-quantification by SAD. Currently, the expected distribution is approximated by a bias correction model that uses GC, sequence, and position biases. The sequence bias may also be affected by the secondary structure of cDNA, which is not considered in current modeling of biases.

SAD takes about 8 hours to run on each RNA-seq sample using eight threads on the GEUVADIS samples and about 23 hours on the Human Body Map samples. Empirically the running time scales linearly with the number of sequencing reads as the sequencing depths of Human Body Map samples are about three times those of GEUVADIS samples. The long running time is mainly due to the sampling procedure in the empirical p-value calculation for all transcripts. A derivation of a p-value approximation to avoid sampling could potentially decrease the computational requirements. Implementation engineering can also be applied to reduce the running time, however, this is out of the scope of this work.

Our formulation of anomaly detection is an example of algorithmic introspection: algorithms that can automatically identify where their predictions do not fit the assumptions of the algorithm. This type of algorithmic reasoning is likely to become even more useful as the sophistication of bioinformatics analysis tools increases.

3.1.14 Appendix

Retrieving the expected distribution from Salmon

We processed the auxiliary output from Salmon to obtain the estimated expected distribution. The expected distribution is estimated for each transcript using the bias model from Patro et al. [113]. In the ideal case of sequencing, where the read is sampled randomly without any biases, the expected coverage is uniform along the positions of any transcript. However, in the real sequencing experiments, cDNA fragmentation and PCR amplification have preferences towards certain positional, sequence, and GC patterns, and the coverage is not expected to be uniform. The expected distribution is calculated to represent the probability of sampling a read at a given position of a given transcript. Salmon estimates the positional, sequence, and GC biases by adjusting the uniform distribution based on the read mapping. There could be other biases affecting the expected distribution. However, other biases are not considered in the model, and thus the bias correction model is only an approximation for the true expected distribution.

Retrieving observed coverage from Salmon

The observed coverage is the actual read coverage for each transcript. It is calculated by counting the weighted number of reads at each position at a given transcript after the weights are optimized by Salmon's algorithm [113]. Specifically, when a read is multi-mapped to several transcripts, the weight represents the probability that the read is generated from the transcript.

Retrieving expected and observed coverages from RSEM

The expected coverage can be estimated in RSEM by using the "`--estimate-rspd`" option. RSPD stands for read start position distribution and this models the $3'$ positional bias, which is the only bias considered by RSEM. RSEM discretizes all transcripts into 20 bins (default parameter of RSEM) and estimates a single probability distribution for all transcripts to describe the probability of sequencing a read from each bin. To recover the estimated expected distributions of each transcript, we extend the single probability distribution to the length of each transcript by uniformly distributing the probability to all transcript positions in the corresponding bin. The expected distribution of each transcript will look like a step function with the step size equal to the transcript length divided by the number of bins.

The observed coverage is directly processed from the BAM file output by RSEM, where each alignment record has an additional tag to denote the weight assigned to the corresponding transcripts. Summing up the weight of each alignment starting positions generates the observed coverage of RSEM.

With the expected and observed coverage calculated from RSEM, the anomaly scores and p-values can be calculated in the same way as with Salmon. However, because RSEM has a single, binned expected distribution for all transcripts, the assumption that estimation error of expected distributions follows a Gaussian error may not be true. The estimated error of the expected distributions may not be small enough for the LP read reassignment to achieve an accurate adjusted quantification.

Simulation procedure

To mimic the real scenario where the target transcriptome contains unannotated transcripts outside reference transcriptome, we simulated target and reference transcriptomes as follows. Using the Gencode annotation [38], we randomly selected 200, 500, 1000, 1500 genes, remove one transcript per gene, and use the rest of the transcript sequences as reference transcripts. For the target transcriptome, we simulated 200, 500, 1000, 1500 fusion genes, added them to Gencode transcript sequences, and use the combined full Gencode transcripts and fusion sequences as the target transcriptome that generates RNA-seq data. Each fusion transcript is simulated by randomly choosing a pair of transcripts that have not been involved in other fusion events, randomly choosing breakpoints within the transcript that are at least 20 bp away from the endpoints, and finally concatenating the pair of transcripts at their breakpoints. The 20 bp threshold ensures there is a distinction between indels when aligning or mapping the reads to the reference. In this case, the target transcriptome contains both unannotated isoforms and fusion sequences compared to the reference. We use both the protein-coding-only annotation and full annotation for removal and fusion simulation, to test both polyA RNA-seq and total RNA-seq techniques.

Reads are simulated using the target transcriptome by Polyester [39]. A count matrix is used as input in Polyester to denote the theoretical number of reads to be simulated for each transcript in the transcriptome. The count matrix is generated by quantifying RNA-seq datasets (GEUDAVIS, GM12878, K562) using Salmon [113] and the original Gencode annotation. With the simulation datasets, Salmon version 0.9.1 is used to quantify the reads against the reference transcriptome.

Detecting transcripts with non-identifiable abundances

The software eXpress (version v1.5.1) [129] is used to identify transcripts with non-identifiable abundances. eXpress is a quantification tool that depends on a probabilistic model involving fragment lengths, transcript lengths, and mapping positions variables. It outputs whether the abundance of a transcript can be uniquely maximized, which is the identifiability under its objective. We use the identifiability under eXpress as a proxy for the identifiability under the Salmon quantifier.

Though the quantification model of eXpress is different from that of Salmon, we expect that for many transcripts their identifiability statuses are the same under both models. Identifiability of a probabilistic model is largely determined by the parameter space, objective function and the probability assumptions. Both models maximize the probability of observing the given set of reads and use the same parameter space, which is the abundances vector of all transcripts. The basic model assumption, that the probability of observing a read from a transcript is proportional to the abundance of the transcript, is also shared. Under these input and assumptions, whether the optimal parameter settings are multiple largely depends on the similarity among input transcripts, specifically whether a subset of transcripts can be linearly represented by another subset of transcripts. We therefore expect the identifiability statuses are similar between the two quantification models, despite the difference in their objective functions and their optimal solutions.

STAR version 2.6.0 [34] is used to align RNA-seq reads to Gencode version 26 transcriptome sequences. The alignment is the input to eXpress quantifier. The identifiability is indicated in the

"solvable" column of eXpress output.

Running transcriptome assembly on simulated and real data

RNA-seq reads are aligned to GRCh38 genome [132] using STAR version 2.6.0 [34]. We ran Scallop version v0.9.8 on the alignments of all simulated, GEUVADIS and Human Body Map samples, and set all parameters to their default. We also ran StringTie [115] version 1.3.1c on all these samples, using the option "`-G`" for guiding the transcriptome assembly by the reference transcriptome. When guided by reference transcriptome, the precision of StringTie can be better than Scallop on some samples. We do not guide the Scallop assembly by reference transcriptome since it does not have the option. We use gffcompare [114] to compare the assembled transcripts with the reference transcript.

Additional Figures

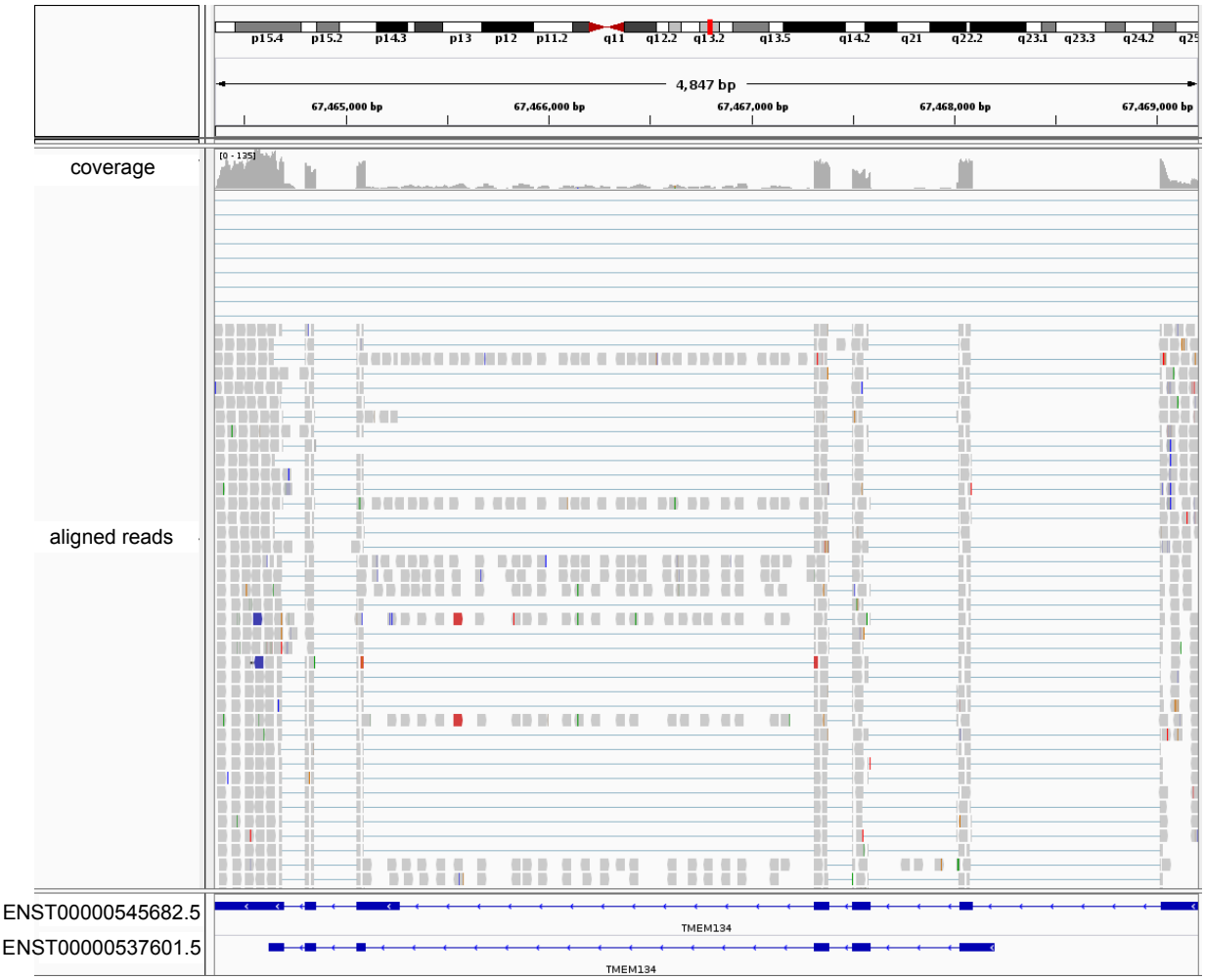

Figure 3.8: IGV visualization of alignments on *TMEM134* of the kidney sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000545682.5; the intron-exon structure of transcript ENST00000537601.5. SAD identifies the region after the first splicing junction of transcript ENST00000545682.5 as an under-expressed region. The anomaly is adjustable by re-shuffling reads with transcript ENST00000537601.5. Before SAD adjustment, expression of ENST00000545682.5 is 1.4 times that of ENST00000537601.5. After SAD adjustment expression of ENST00000545682.5 is 9 times that of ENST00000537601.5. The first splicing junction (right-most junction) of ENST00000545682.5 is highly expressed, which is more consistent with a larger abundance ratio.

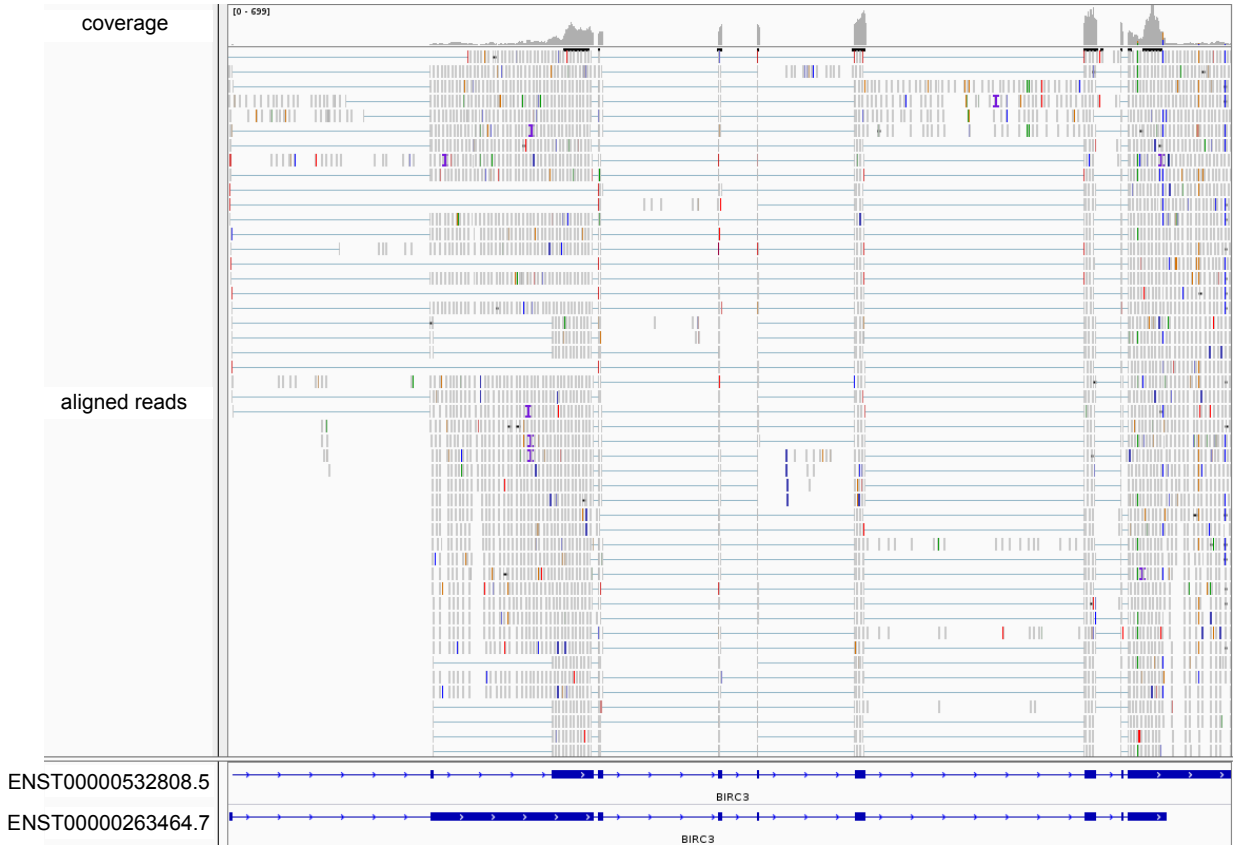

Figure 3.9: IGV visualization of alignments on *BIRC3* of a GEUVADIS sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000532808.5; the intron-exon structure of transcript ENST00000263464.7. SAD identifies the 3' region of ENST00000532808.5 as an under-expressed region. The anomaly is adjustable by re-shuffling reads with transcript ENST00000263464.7. Before SAD adjustment, the two isoforms has similar expression. After SAD adjustment expression of ENST00000263464.7 is 3 times that of ENST00000532808.5. According to the expected coverages, the 3' sides of both transcripts are expected to have higher coverage than 5' sides. That the 3' end of ENST00000532808.5 has low coverage suggests that ENST00000532808.5 may be of low abundance.

Figure 3.10: IGV visualization for the unadjustable anomaly examples. (A). Gene *UBE2Q1* of the heart sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000292211.4. There is an under-expressed region of the transcript at 3' end (left-most region). The anomaly is unadjustable. (B) Gene *LIMD1* of the heart sample. The labeled tracks from the top to bottom are: coverage along genomic positions; the aligned reads onto the genome (using STAR aligned); the intron-exon structure of transcript ENST00000273317.4. There is an under-expressed region of the transcript at 3' end (right-most region). The anomaly is unadjustable.

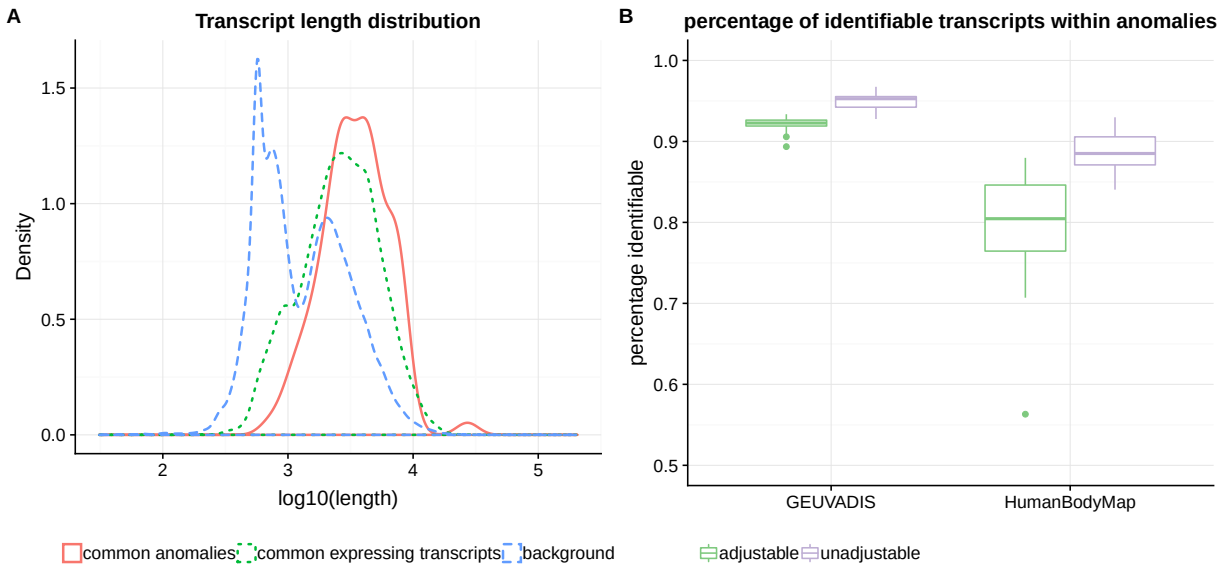

Figure 3.11: Length distribution of unadjustable anomalies and identifiability status. (A) Density curve of length distribution of the common unadjustable anomalies and commonly expressed transcripts across all 46 samples. The length distribution of common unadjustable anomalies generally follows that of the commonly expressed transcripts. Some transcripts are not commonly expressed. When including these transcripts into the background, the length distribution contains a large proportion of transcripts with a shorter length than the commonly expressed ones and the common unadjustable anomalies. (B) Percentage of unadjustable and adjustable anomalies that are identifiable in the quantification model. Each box indicates the range of percentages across the transcripts in the indicated dataset and the indicated anomaly category. Transcripts with identifiable expression indicate the optima is unique and the anomaly is not a result due to the intrinsic uncertainty of quantification optimization objective. Identifiability is determined by eXpress, which uses a different quantification model from Salmon but still reflect the degree of identifiability.

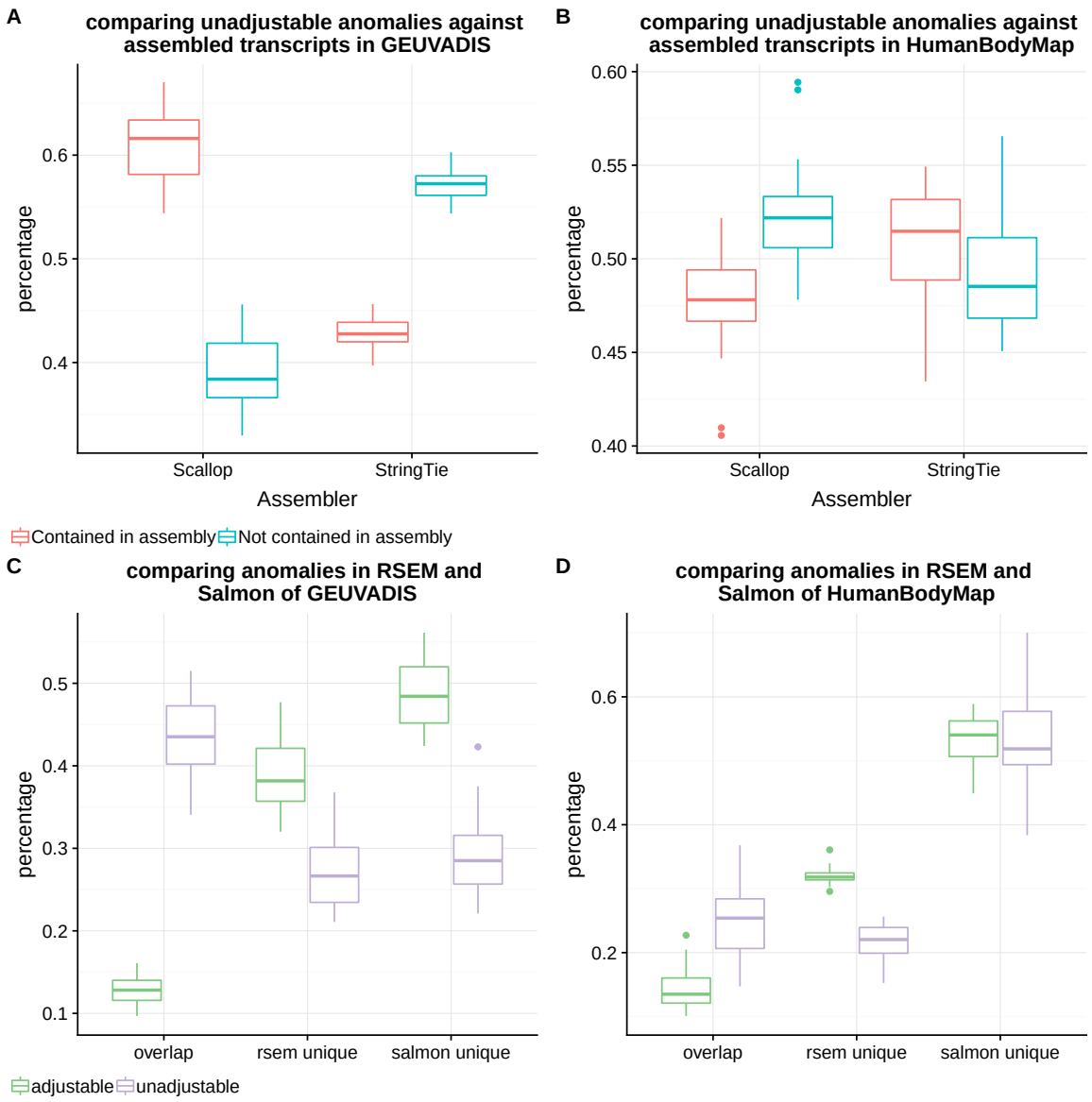

Figure 3.12: Comparing Salmon anomalies with transcriptome assembly and RSEM anomalies. (A–B) Proportion of the unadjustable-anomaly-containing genes that can (cannot) be detected by transcriptome assemblers. Each box indicates the range of percentages across samples in the corresponding dataset. (A) For the GEUVADIS dataset, about 40% of the genes do not have corresponding unannotated isoforms predicted by Scallop, and about 60% of the genes do not have unannotated isoforms predicted by StringTie. (B) For the Human Body Map dataset, about 53% of unadjustable-anomaly-containing genes cannot be detected by Scallop, and about 50% of them cannot be detected by StringTie. The lower percentage of detection from StringTie in GEUVADIS dataset may be an effect of using the "Guided by reference" option. (C–D) Overlapping of unadjustable anomalies predicted based on Salmon and RSEM on (C) GEUVADIS dataset and (D) Human Body Map dataset. Each box indicates the range of percentages across samples in the corresponding dataset. The denominator of the percentage calculation is the number of transcripts that are detected as unadjustable (or adjustable) anomalies under either Salmon or RSEM quantification.

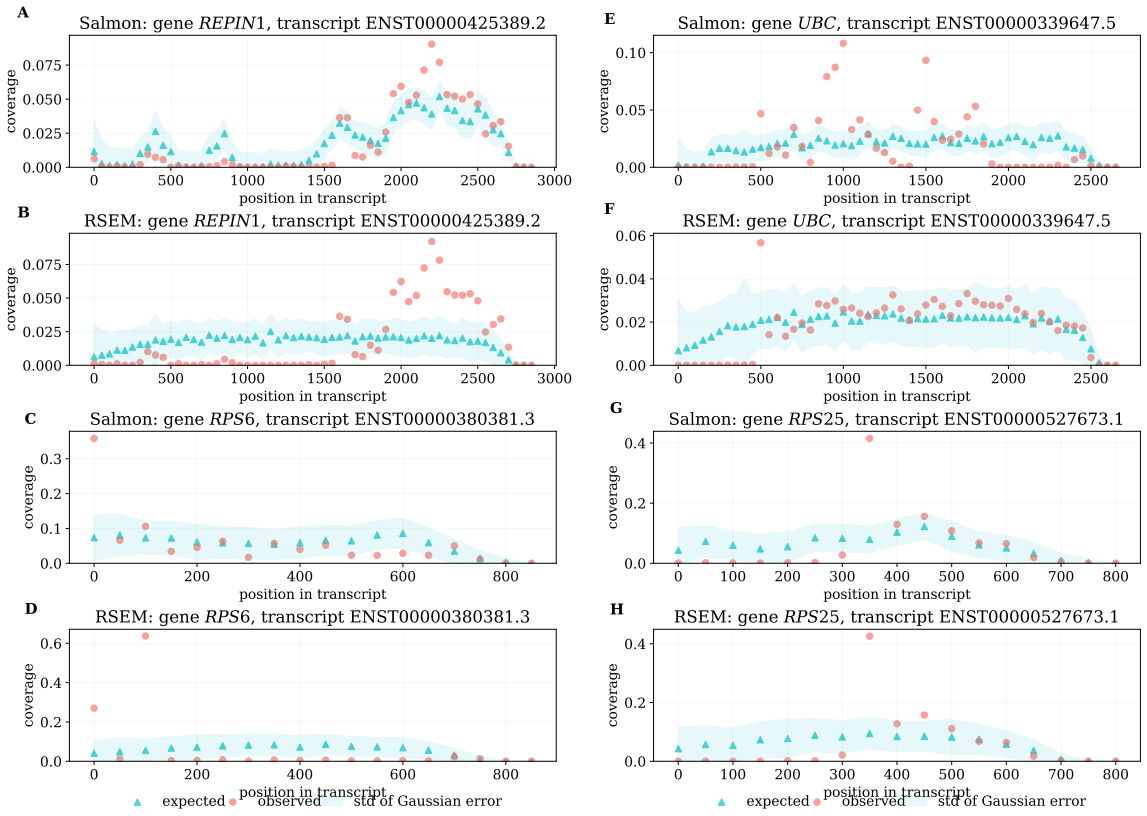

Figure 3.13: Differences between Salmon and RSEM unadjustable anomalies. All examples are from one GEUVADIS sample (accession ERR188265). Red and blue points are the observed and expected coverage distribution separately, and the blue shade is the standard deviation of the expected distribution estimation. Each point is a 50 bp bin along the transcript. (A–B) Expected and observed coverage for transcript ENST00000425389.2 under (A) Salmon and (B) RSEM. The transcript is identified to be unadjustable anomaly under only RSEM. The observed distributions under both quantifiers are similar. The difference in the estimated expected distribution causes the transcript to be detected as an unadjustable anomaly under RSEM but not Salmon. (C–D) Expected and observed coverage for transcript ENST00000380381.3 under (C) Salmon and (D) RSEM quantification. The transcript is identified to be unadjustable anomaly under only RSEM. The observed coverage distributions has large difference between the two quantifiers around position 150, which causes the transcript to be detected as an unadjustable anomaly under RSEM but not Salmon. (E–F) Expected and observed coverage for transcript ENST00000339647.5 under (E) Salmon and (F) RSEM. The transcript is identified to be unadjustable anomaly under only Salmon. The observed coverage distributions has large difference between the two quantifiers. (G–H) Expected and observed coverage for transcript ENST00000527673.1 under (G) Salmon and (H) RSEM. The transcript is identified to be unadjustable anomaly under only Salmon. Both the observed and the expected coverage distribution under the two quantifiers are similar. However, RSEM has a relatively larger variance of Gaussian error in the expected distribution estimation and leads to a insignificant p-value in RSEM.

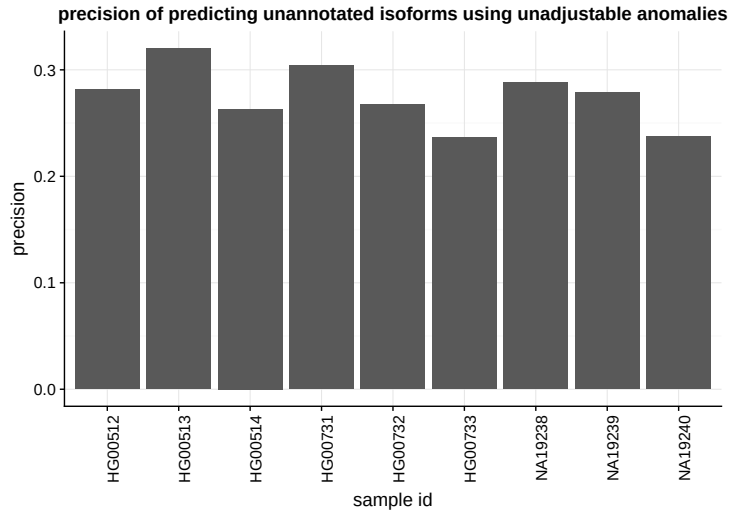

Figure 3.14: Validating unadjustable anomaly prediction using full-length transcript sequencing. Whether an unadjustable anomaly is caused by an unannotated isoform can be validated by PacBio full-length transcript sequencing. When a sequencing reads contain a large proportion of the predicted over-expressed region and exclude a large proportion of the predicted under-expressed region, the unadjustable anomaly is considered to be supported by the long reads and correctly predicted. Y-axis shows the percentage of unadjustable anomalies that have long reads supports, that is, the prediction of unadjustable anomaly prediction. The precision is around 23% – 32% for all 9 samples from 1000 Genome project.

3.2 Coverage anomalies of transcription factors partially explain the expression of target genes in breast cancer

3.2.1 Background

Gene expression varies among individuals, tissues, and cell types. The task of decoding the genetic, epigenetic, and transcriptomic factors that contribute to gene expression variance has attracted researchers for more than a decade and still needs ongoing efforts. Expression of certain genes are predictors of phenotypes, including disease subtypes, treatment responses, and survival [60, 107, 166]. Identifying the factors that explain the expression helps us better understand the association between phenotypes and the cellular systems, including genetic, epigenetic, and transcriptomic status. Several genetic and epigenetic contributors have been identified to regulate gene expression by previous studies. DNA methylation regulates gene expression, and for example, allele-specific methylation of CpG islands on X chromosome inactivates the gene expression of the allele [58, 106]. Some of the single nucleotide variants (SNVs) can either affect expression of nearby genes or distant genes, and such SNVs are also known as expression quantitative trait loci (eQTLs) [26, 42, 46, 83, 109]. The abundance of a specific type genes, transcription factors or TFs, also regulates the expression of the rest of genes. We focus on the regulation effect of TFs and analyze whether the coverage patterns of TF expression can also explain a proportion of the expression variability of their regulated genes.

TFs regulate gene expression by binding to the promotor, enhancer, or silencer regions of DNA, and recruiting RNA polymerase or co-factors to activate gene transcription, or inhibiting the binding of other TFs to reduce gene transcription [165]. Mathematical models have been developed to model the TF regulation process: TF-DNA binding probability is modeled by thermodynamics equilibrium, which further determines RNA synthesis kinetics [6, 30, 92, 133, 140]. TF regulation is involved in a wide range of pathways and cellular processes. Mutations, structural variants, and abnormal expression of TFs can cause a wide range of diseases, including cancer, autoimmune diseases, neurological disorders, and diabetes [76, 160]. Many of the eQTLs that affect the expression of distant genes, called trans-eQTLs, are mutations on TFs. Besides the known genetic variants in TFs, transcriptomic features of TFs are also likely to be related to their regulation efficiency. For example, different 3' UTR cleavage of the TF *SREBF2* is chosen at different stages of spermatogenic cell differentiation [162].

In this study, we investigate whether the coverage imbalance across transcripts' regions in TF expression can explain the expression variance of the target genes (TGs) regulated by the TF. Coverages of transcript expression can be obtained from RNA-seq data by counting the number or the assigned weights of RNA-seq fragments at each transcript position. Expression coverage along a transcript is balanced when the coverage is consistent along all transcript positions. Otherwise, some subsequences of the transcript are over-expressed and the rest of the transcript sequences are under-expressed; we call this a coverage anomaly. If the relative expression coverage along positions of a transcript is known, the coverage anomaly is defined as deviations of the actual expression coverage from the expected coverage, and there are multiple choices of distance metric to measure the deviations. As an example of a coverage anomaly, the pattern that a long subregion of 3' untranslated region (UTR) in a transcript is near zero coverage in a highly expressed transcript is observed in several RNA-seq samples, as shown in the results of previous section. One possible explanation for this coverage anomaly is that an unannotated isoform with alternative transcription termination that leads to a shorter 3' UTR is expressed, which is also known as alternative 3' UTR cleavage. The reads from the unannotated isoform are mapped to the reference isoform and increase the reference isoform coverage at the shared regions, which leads to the coverage anomaly pattern.

It is reasonable to hypothesize that coverage anomalies are candidate indicators of regulation efficacy of TFs as well as explaining factors of the expression variance of the TGs. When coverage anomalies are caused by the expression of unannotated transcripts, the unannotated sequences may contain a unique set of translation regulatory elements and translate into a protein with unique binding property. UTRs in a protein-coding transcript contain regulatory elements that regulate the translation rate, and sequence variation in UTRs possibly alter the protein synthesis rate as well as final abundance [8, 117]. If the unannotated sequence contains a varied coding region, the protein sequence can be altered as well. A protein sequence alteration in both binding domains and intrinsically disordered region can change the binding affinity and specificity [88, 168], which can further lead to altered regulation efficiency and altered expression of TGs. In fact, aside from TFs, coverage anomalies are potential indicators of abundance or functioning efficiency alterations in other proteins with the same reasoning.

Coverage anomalies are one characterization of the transcript expression status besides the expression abundances. They do not distinguish the biological events that lead to the status. Unbalanced regional coverages can be due to the existence of external sequences, or the expression

of unannotated alternative splicing isoforms, or even unknown biological mechanisms. The expression of unannotated alternative splicing isoforms can be further regulated by genomic SNVs or methylation status. Coverage anomalies represent the inconsistency of expression coverage among regions in a transcript regardless of the genetic cause or biological events. Even though the different biological events may not necessarily have the same influence on TF regulation and TG expression, when they do have similar effects, coverage anomalies will capture the shared effects.

With the above hypothesis that coverage anomaly of TFs are a potential factor to explain gene expression variance, the goal of this work is to computationally verify the hypothesis. Specifically, we seek to answer to the following questions: Which TGs' expression variance can be significantly explained by the coverage anomaly status of TFs? To what degree can coverage anomalies explain the expression of TGs? Is the regulation efficiency of TF enhanced or reduced when it contains coverage anomalies?

We only focus on known TFs for investigating the association between their coverage anomaly status and the expression of their TGs. Gene expression control is a complicated process and likely has falsely high associations with many cellular measurements. For example, previous trans-eQTLs are identified with particular caution because there exist many false associations. However, there is a clear chain of biological interpretations on why coverage anomalies have association with gene expression variance: coverage anomalies are status indicators of TFs, and TFs regulate expression of their corresponding TGs under well-studied mechanisms. The known TF-TG regulation relationship increases our confidence of the identified associations between coverage anomalies and expression of TGs.

Using TCGA breast cancer RNA-seq data (`https://www.cancer.gov/tcga`), we use linear models to investigate whether the expression of a gene can be explained by the degree of coverage anomalies of its regulating TFs. Since coverage anomaly and expression abundances are both characterizations of TF expression states, their effects on altering the expression abundances of TGs can be either separate or coupled together. Specifically, when a given degree of coverage anomaly leads to a fixed amount of expression change in the TG regardless of TF expression, the effects of coverage anomalies and of TF expression abundances are separate; when the expression change in TG explained by coverage anomalies increases as TF expression increases, the effects are coupled together. The above hypotheses about the effects of coverage anomalies and expression abundances are also tested in linear models.

We observe that coverage anomaly status of some TFs can indeed significantly explain the expression variance of the TGs, and there are 319 TF-coverage anomaly-TG triples where the coverage anomalies are significant under FDR threshold 0.05. Both separate and coupled effects from coverage anomalies and expression abundances are observed in the 319 significant triples. Both enhancement and reduction of regulation efficiency are observed when TFs contain coverage anomalies, but reduction of regulation is more common. We observe the 69 TFs or TGs in the significant triples are related to cancer according to COSMIC database [141], such as *TNFRSF17* and *ESR1*, however, coverage anomalies in these triples cannot be interpreted as disease-associated. Comparing the explained variance of TGs' expression from coverage anomalies and from TFs' methylation status and known eQTLs, we find that the coverage anomalies contain unique information that is relevant to TGs' expression.

3.2.2 Overview of methods

Coverage anomalies are detected by Salmon Anomaly Detection (SAD) [97]. SAD outputs an indicator of whether each transcript contains a coverage anomaly and an anomaly score for each coverage anomaly. The anomaly score is the difference between observed coverage and expected coverage of the most abnormal region and indicates the degree of abnormality. We construct an anomaly vector for a list of RNA-seq samples to represent the coverage anomaly status of a given transcript: each entry is the anomaly score of the transcript in the corresponding sample when the anomaly indicator suggests a coverage anomaly exists, and is zero when the anomaly indicator suggests no coverage anomaly. Anomaly vectors of TF transcripts are incorporated in linear models to evaluate their explanatory power on TG expression.

The expression of a gene is modeled by a linear model using clinical covariates, methylation and copy number variation (CNV) of the gene, the expression and anomaly status of its regulating TFs. Let y be the log-transformed expression abundance of a TG. Let P be the clinical covariates (age at diagnosis and tumor stage) of the patient corresponding to the RNA-seq sample. Let M be the methylation level of the TG and C be the copy number variation (CNV) of the TG. Let X_g and X_t be log-transformed gene-level and transcript-level expression abundances of the TFs that are known to regulate the TG's expression. We use β_* to represent the coefficients in the linear model. The expression of TG, y, is predicted by the following linear model:

$$y = \beta_0 + \beta_p P + \beta_m M + \beta_c C + \beta_g^T X_g + \beta_t^T X_t + D + \epsilon, \tag{3.15}$$

where term D is a combination of anomaly-related factors, which will be explained in the next paragraph. CNV and DNA methylation of TG are included in the linear model since previous studies have shown that they tend to have strong influence on the gene expression [51, 120]. However, we do not include methylation or CNVs of TFs because expression abundances of TFs include the expression of all copies of TFs, and genetic and epigenetic states of TFs (including methylation) are potential explanations of the occurrence of coverage anomalies. Including two dependent factors in linear models obscures the association between the predicted values and each individual factor. Therefore, to study whether coverage anomalies can explain the expression variance of TG, we do not include other genomic or epigenomic measurements of TFs that may be associated with the occurrence of coverage anomalies.

Considering that both coverage anomaly and abundance characterize the expression status of TFs, we hypothesize that the changes in TG's expression due to coverage anomaly and due to the expression of TFs can either be independent or dependent on each other, or partially dependent. Let A_{ij} be anomaly score of the j^{th} transcript that belongs to the i^{th} TF. D in equation (3.15) may contain a combination of different factors corresponding different hypotheses, listed in Table 3.2. Even though these factors do not directly model the changes in biophysical or biochemical reactions, they can be used to infer the mechanisms and conditions where such a change occurs. The following paragraphs describe the hypotheses and factors in more detail.

The factor A_{ij}, concentration-independent factor, follows the hypothesis that the coverage anomaly and TF's expression change TG's expression separately independently. A given degree of coverage anomaly in a TF leads to a fixed amount of expression alteration in TG regardless of the expression of the TF. One possible scenario when this hypothesis holds is when the coverage anomaly of TF indicates the expression of an unannotated transcripts, which has an altered

Table 3.2: Anomaly-related prediction term D. Under different hypotheses whether the expression variance of TG explained by coverage anomaly and explained by TF's expression are dependent on each other. Columns indicate the dependence between effect of anomaly and effect of TF expression. When the explained expression variance from coverage anomalies is independent of TF's expression, D only contains the concentration-independent factor A_{ij}. When the explained expression variance from anomaly is dependent on TF's expression, D contains the concentration-dependent factor $A_{ij}X_{g,i}$ or $A_{ij}X_{t,j}$. In this case, TG's expression has a larger alteration due to coverage anomaly status if the TF has a larger expression. When the explained expression variance from anomaly is partially dependent on TF's expression, D contains both concentration-independent and concentration-dependent factors. Rows indicate whether the concentration-dependent factor involves gene-level or transcript-level expression of the TF.

	Independent (M_a)	Dependent (M_{int})	Partial dependent (M_{both})
Gene-level TF expression (L_g)	$\beta_a A_{ij}$	$\beta_{g,int} A_{ij} X_{g,i}$	$\beta_a A_{ij} + \beta_{g,int} A_{ij} X_{g,i}$
Transcript-level TF expression (L_t)	-	$\beta_{t,int} A_{ij} X_{t,j}$	$\beta_a A_{ij} + \beta_{g,int} A_{ij} X_{t,j}$
Both levels TF expression (L_{both})	-	$\beta_{g,int} A_{ij} X_{g,i} + \beta_{t,int} A_{ij} X_{t,j}$	$\beta_a A_{ij} + \beta_{g,int} A_{ij} X_{g,i} + \beta_{t,int} A_{ij} X_{t,j}$

biochemical property after binding with a scarce interactant. For example, after binding with the interactant, the protein complex containing the unannotated isoform recruits RNA polymerase more efficiently than the protein complex containing the annotated isoforms. Due to the scarcity of the interactant, the ratio between unannotated and annotated isoforms controls the TG's expression regardless of the raw expression abundance of TFs.

The factors $A_{ij}X_{g,i}$ (gene-level concentration-dependent factor) and $A_{ij}X_{t,j}$ (transcript-level concentration-dependent factor) follow the hypotheses that the changes of TG's expression from coverage anomalies and from TF's expression are dependent on each other. A given degree of coverage anomaly leads to a larger amount of change in TG's expression if the TF has a larger expression. These cases may happen when an unannotated isoform TF has an altered property after binding with an abundant interactant. In this case, the abundance of the interactant is not the bottleneck of TG's transcription process. Instead, increasing the amount of TF protein abundance results in larger amount of protein complex that contains the unannotated TF isoform, which leads to a larger change of TG's expression.

We test the significance of D in predicting TGs' expression by likelihood ratio test. To further increase the confidence of the detected associations between coverage anomalies and TGs' expression, we only perform the statistical on the confident TF-TG pairs where the gene expression of TF has significant prediction power of the TG's expression. See Section 3.2.7 for determining the significance of a TF's expression to predict a TG's expression and Section 3.2.7 for statistically testing the significance of coverage anomalies. When a TG has multiple regulating TFs, we test the coverage anomaly status of each TF transcript one at a time. All forms of D are tested for each TF-coverage anomaly-TG triple, but different D under different hypotheses are adjusted by

Benjamini-Hochberg method separately. When a TF-coverage anomaly-TG triple is significant under multiple forms of D, we select the D form with "maximum relevance minimum redundancy" for the further analysis. See Section 3.2.7 for details of model selection. By comparing the the TF's regulation direction with the changing direction of TG's expression after including the coverage anomaly score, we can infer whether the regulation efficiency of a TF is enhanced or reduced. Determining the regulation direction is detailed in Section 3.2.7.

3.2.3 Coverage anomalies of TFs explain the expression variance of TGs in 319 TF-coverage anomaly-TG triples in breast cancer

After removing the breast cancer samples that fail pre-processing steps of Salmon Anomaly Detection (SAD) and that do not have corresponding clinical, CNV, or methylation data, there are 993 samples remaining. The TF-TG pairs are collected from TRRUST database [54], which contains 781 unique TFs and 9304 TF-TG regulation pairs after intersecting its genes with Gencode genes version 30 [38]. After filtering out the TF-TG pairs where a TF does not have significant regulation effect on the TG's expression in linear models and filtering out the cases where coverage anomaly of a TF occurs in less than 10 samples, there are 1905 TF-TG pairs or 3214 TF-coverage anomaly-TG triples. We test the significance of coverage anomalies in predicting TGs' expression corresponding to the 3214 triples using the 993 breast cancer samples.

	M_a	M_{int}	M_{both}
L_g	39	13	104
L_t	-	19	68
L_{both}	-	35	41

Table 3.3: Number of TF-coverage anomaly-TG triples that coverage anomalies have significant prediction power on the expression variance of TG under each coverage anomaly effect hypothesis.

With FDR threshold 0.05, there are in total 319 TF-coverage anomaly-TG triples where anomaly score has significant prediction power for the TG's expression. Table 3.3 shows the number of significant triples under each hypothesis of coverage anomaly effect under their "maximum relevance minimum redundancy" model. We observe that the significant triples span all tested model hypotheses. The model that contains the largest number of significant triples (around $\frac{1}{3}$ of total number of significant detections) is $D = \beta_a A_{ij} + \beta_{g,int} A_{ij} X_{g,i}$, where TG's expression change from a given coverage anomaly score can be decomposed into a constant part and a variable part which increases as TF's expression increases. But all the other models contain TF-coverage anomaly-TG triples where coverage anomalies explain TG's expression under the corresponding hypothesis. We also observe that the different factors of coverage anomalies sometimes have synergistic prediction effect (Appendix Figure 3.19). That is, the sum of explained variance of TG's expression from each individual anomaly factor separately is smaller than the explained variance when including both factors. The synergy is observed both between

two concentration-dependent effects and between concentration-independent and concentration-dependent effects.

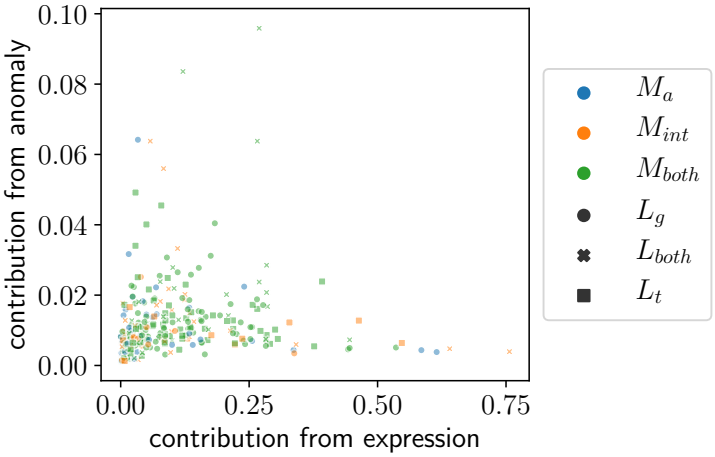

Figure 3.15: Explained TG expression variance from TG expression (x-axis) and from coverage anomalies (y-axis).

Comparing the explanation power between the coverage anomalies and the corresponding TFs' expression, we observe that in more than 90% of the significant triples, the explanation power of coverage anomalies is 3.79% – 80.71% of that of the corresponding gene and transcripts (Figure 3.15). We use increased percentage of TG expression variance (R-squared) that can be explained by adding a feature in the linear model to represent the explanation power. The median ratio between increased R-squared from anomalies and from the corresponding gene among the 319 significant anomalies is 18.87%. Without calculating the ratio, the 90% confidence interval of the extra percentage of explained variance of TGs by considering coverage anomalies ranges from 0.21% to 2.79%.

3.2.4 Both enhancement and reduction of regulation efficiency occur when TF contains coverage anomaly

We investigate the direction of change of regulation efficiency when the TF has a coverage anomaly, specifically, whether the change of TG's expression is along the same or opposite direction of TF regulation. When the linear model contains only one factor related to coverage anomaly, the direction of change of regulation efficiency is determined by whether the sign of coefficient of the anomaly factor in the linear model agrees with the sign of coefficient of the TF expression. When a linear model contains multiple anomaly factors and some of them depend on gene/transcript expression level, there is no single coefficient to represent the changing direction of regulation efficiency. In this case, we fix the TF gene/transcript expression quantile and sum up the coefficients to compare the sign with the coefficient of TF expression.

For more than half of TF-coverage anomaly-TG triples, the TFs' regulation efficiency is reduced when having coverage anomalies under medium to small TF expression levels (Figure 3.16A–C). A possible explanation is that the unbalanced expression of transcript regions

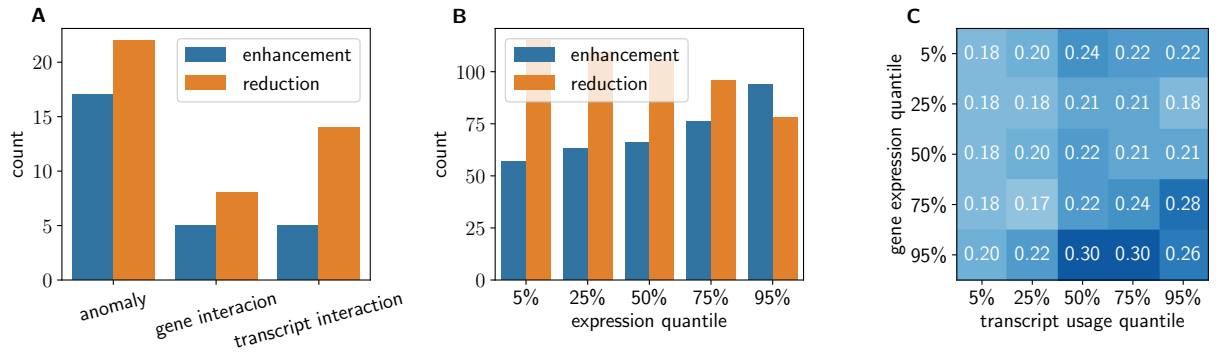

Figure 3.16: Percentage of TF-coverage anomaly-TG triples where enhanced or reduced TF regulation efficiency is observed when containing coverage anomalies. (A) Number of TF-coverage anomaly-TG triples where TF's regulation efficiency is enhanced or reduced when containing coverage anomalies. The triples are restricted to ones where D only contains one anomaly factor. X axis shows anomaly factor in D. (B) Number of TF-coverage anomaly-TG triples where D contains the concentration-independent effect and one of the concentration-dependent effect. X-axis is the quantile of gene/transcript expression of TF under which the overall change of regulation efficiency is computed. (C) Percentage of TF-coverage anomaly-TG triples where TF regulation efficiency is enhanced within the triples where D contains all three anomaly factors. In this panel, transcript usage is the percentage of transcript expression over gene expression (the sum of expression of all isoforms). We switch to use quantiles of transcript usage instead of transcript expression quantile to avoid the impossible expression case where transcript expression is larger than gene expression.

usually lead to reduced function or malfunction of some domains, which further leads to the reduction of overall regulation efficiency.

Despite that the majority of coverage anomalies indicate reduced regulation efficiency of TFs, this trend does not hold when the linear model contains concentration-independent and one of the concentration-dependent factors and the gene/transcript expression of the TF is above 95% quantile of all samples (Figure 3.16B). Under this condition, the number of elevated TF regulation efficiency is larger than the number of reduced TF regulation efficiency. However, further analysis from both computational and experimental aspects is needed to confirm this observation and to rule out the possibility that linear models have low approximation accuracy on extreme values of the input expression.

The absolute coefficients of anomaly-related factors are sometimes larger than the coefficients of TFs' expression (Appendix Figure 3.20), which means a large-enough anomaly score can totally alter the TF regulation direction. However, this phenomenon may also be a result of inaccurate approximation from linear models in extreme anomaly score values, and it needs further experimental validation.

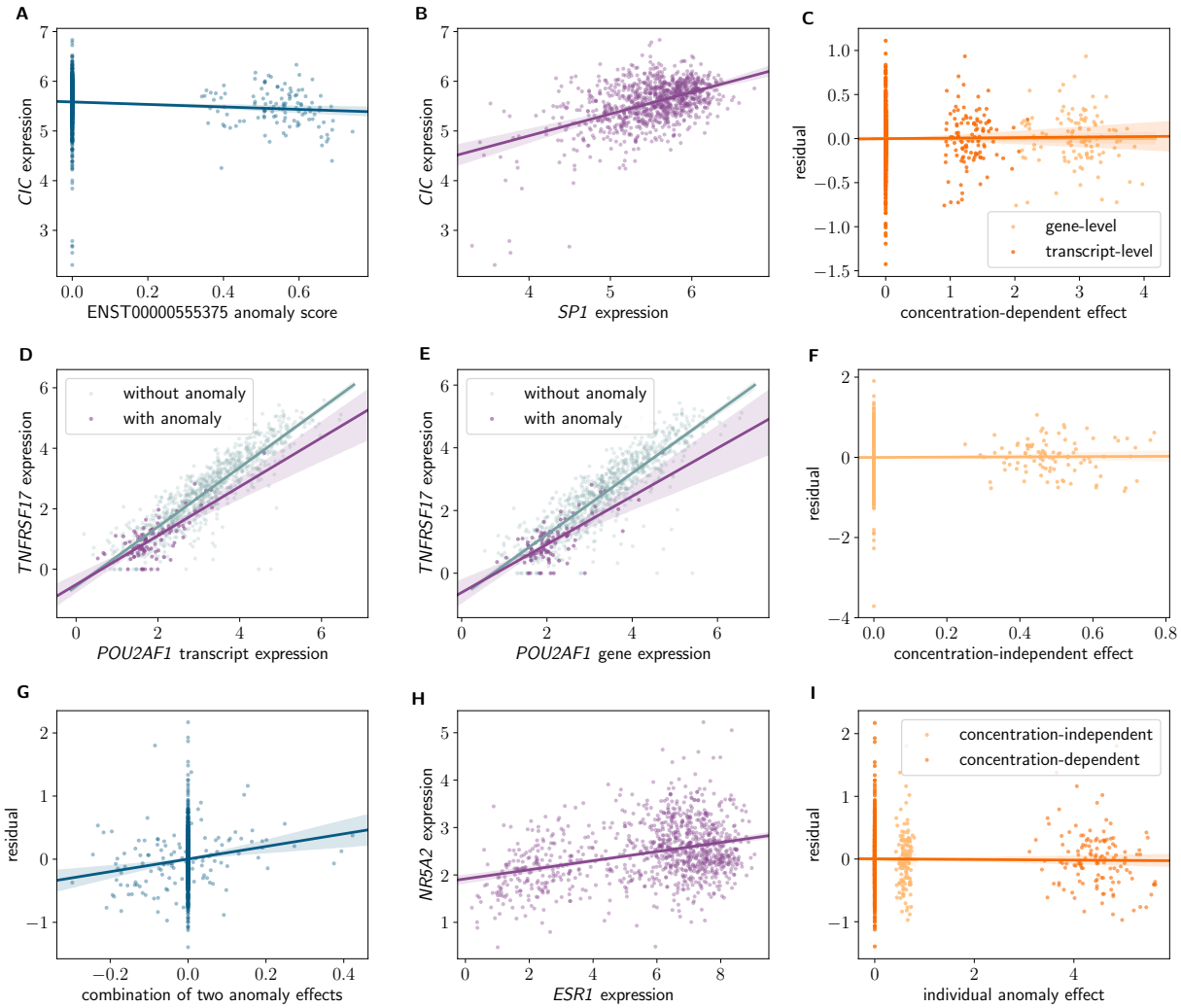

Figure 3.17: Examples of TF-coverage anomaly-TG triples where coverage anomalies are significant in linear prediction. Each row corresponds to a specific triple. (A–C) Example of triple TF *SP1*, anomaly of transcript ENST00000548560, TG *CIC*. Plot between anomaly score and TG expression is in (A); plot between TF-TG expression is in (B); plot between fitting residual and the two concentration-dependent anomaly factors is in (C). (D–E) Example of triple TF *POU2AF1*, anomaly of ENST00000393067, TG *TNFRSF17*. Scatter plot between TF gene-level expression and TG expression when the TF contains/does not contain anomalies is in (D); the scatter plot between TF transcript-level expression and TG expression is in (E); scatter plot between fitting residual and the concentration-independent anomaly factor is in (F). (G–I) Example of triple TF *ESR1*, anomaly of ENST00000482101, TG *NR5A2*. The concentration-independent and gene-level concentration-dependent anomaly factors are significant. Scatter plot between a linear combination of the two anomaly factors and the fitting residual is in (G); the relationship between TF-TG expression is in (H); scatter plot between the fitting residual and each individual anomaly factor is in (I).

3.2.5 Examples of significant anomalies under various hypotheses of coverage anomaly effects

The coverage anomaly in transcript ENST00000548560 of TF *SP1* explains the expression variance of TG *CIC* though concentration-independent effect only. *SP1* is involved in cell differentiation, cell growth, apoptosis, and response to DNA damage [145], and the target gene *CIC* encodes a protein involved in chromatin binding and nuclear localization [145]. Figure 3.17A shows that the anomaly score itself negatively correlates with the expression of *CIC*. The TF expression positively regulates TG expression (Figure 3.17B). After fitting the linear regression model using other covariates and the coverage anomaly score, the residual does not have strong correlation with the concentration-dependent anomaly factors (Figure 3.17C). The over-expressed region locates at the 5' UTR for all samples such that the transcript of the sample contains coverage anomaly, and the under-expressed region spans the coding regions. The mechanism of regulation efficiency alteration in *SP1* remains to be further investigated.

The expression of TG *TNFRSF17* can be explained by the coverage anomaly of transcript ENST00000393067 in TF *POU2AF1*. TF *POU2AF1* is a transcription coactivator required for B-cells to respond to antigens [145], and *TNFRSF17* is involved in B cell development and autoimmune response [145]. The coverage anomaly has both gene-level and transcript-level concentration-dependent effects. In Figure 3.17D and E, the difference between the two fitted lines (alteration of TG's expression due to the coverage anomaly) increases as the TF's gene-level and transcript-level expression increase. We also observe that less amount of *TNFRSF17* is expressed when TF *POU2AF1* contains coverage anomalies, indicating a reduced regulation efficiency of the TF when it contains coverage anomalies. After fitting the linear model using the two concentration-dependent anomaly factors and other covariates, the residual does not show correlation with the concentration-independent anomaly factor (Figure 3.17F). In the coverage anomaly, a part of the 3' UTR region is over-expressed while the 5' UTR and coding sequences are under-expressed.

The last example shows the regulation of TG *NR5A2* from TF *ESR1* is explained by both concentration-independent and gene-level concentration-dependent terms of coverage anomaly of transcript ENST00000482101 (Figure 3.17G). *ESR1* controls cell growth and reproductive function and is one of the key genes in breast cancer to distinguish ER-positive cancer subtype [145], and the target gene *NR5A2* is potentially an important regulator of embryonic development [145]. TF *ESR1* positively regulates *NR5A2* expression (Figure 3.17H), and coverage anomaly reduces the regulation efficiency of TF across 5%–95% quantiles of TF gene expression. The two prediction terms of anomaly separately do not provide a significant explanation of the residual after regressing TG expression using the clinical covariates, CNV and methylation of TG, and TFs' expression, instead combining the concentration-independent and concentration-dependent anomaly terms provides a synergistic explanation power (Figure 3.17I). The transcript ENST00000482101 does not translate into protein according to Gencode annotation, and thus the coverage anomaly might be caused by unannotated transcript and the unannotated transcript has an altered translation and regulation efficiency.

Among these example, *CIC*, *POU2AF1*, *TNFRSF17*, and *ESR1* are known cancer genes of which mutations are causally implied in cancer. In addition, a total number of 69 TFs or TGs are cancer-related according to COSMIC database [141] among the significant triples (See Supple-

mentary File for details). Our analysis shows that coverage anomaly status is an indicator for the expression of cancer-related TGs or the regulation efficiency of cancer-related TFs in a breast cancer cellular environment. But the results do not imply causal relationships between coverage anomalies and cancer.

3.2.6 Explanation power of coverage anomalies does not come from methylation status of TFs or known eQTLs of TGs

We probe whether the prediction power of coverage anomalies can be fully or partially attributed to other genomic and epigenomic features. In this section, we investigate the methylation status of the corresponding TFs and the known eQTLs of TGs and analyze whether these genomic and epigenomic statuses contain overlapping information with coverage anomalies that is relevant to expression prediction of TGs.

Whether the explanation power can be attributed to the two features is analyzed through the extra explained variance of TGs' expression from coverage anomalies. Specifically, we add the coverage anomaly status of a TF to the following two linear models, one with the methylation status of the corresponding TF and the other without. And then we compare the extra percentage of explained variance of the corresponding TG's expression from coverage anomalies in the two linear models. If the extra explained variance in the linear model with methylation status is much smaller than the extra explained variance in the linear model without methylation, the explanation power of the coverage anomalies can be (partially) attributed to methylation status of TFs. Otherwise, if the explained variances from coverage anomalies of the two models are similar, the explanation power of coverage anomalies is irrelevant to that of TFs' methylation status. This approach is also applied to eQTLs features. This approach can be interpreted from an information theory perspective. When methylation status of TFs (or known eQTLs for TGs) and coverage anomaly status of TFs contain the same set of information that is relevant to TGs' expression, adding coverage anomalies does not bring more explanation to TGs' expression when the model contains the methylation status.

We observe that the explanation power of coverage anomalies cannot be attributed to the methylation status of TFs. The explained variance of coverage anomalies does not have big differences between the linear models with and without the methylation status of TFs (Figure 3.18A). A TF usually has multiple sites that can be methylated, and the comparison result is based on a selection of the top 5 sites with the largest prediction power of TGs' expression. This observation suggests that the methylation status of a combination of the top 5 methylation sites does not have large overlapping information with coverage anomalies.

Similar observations are obtained when comparing the extra explained variance from coverage anomaly status when the linear models have/do not have the mutation status of known eQTLs to the TGs (Figure 3.18B). Gong et al. [46] identified eQTLs along with their regulated genes across 33 cancer types in TCGA, and we obtained the list of eQTLs corresponding to each TG in this analysis. We collected both somatic and germline mutation status of the eQTLs for each sample to include in the linear model for comparing percentage of explained variance, where the germline mutation status is computed by Huang et al. [63]. The percentages of expression variance explained by coverage anomalies almost remain unchanged for all TF-coverage

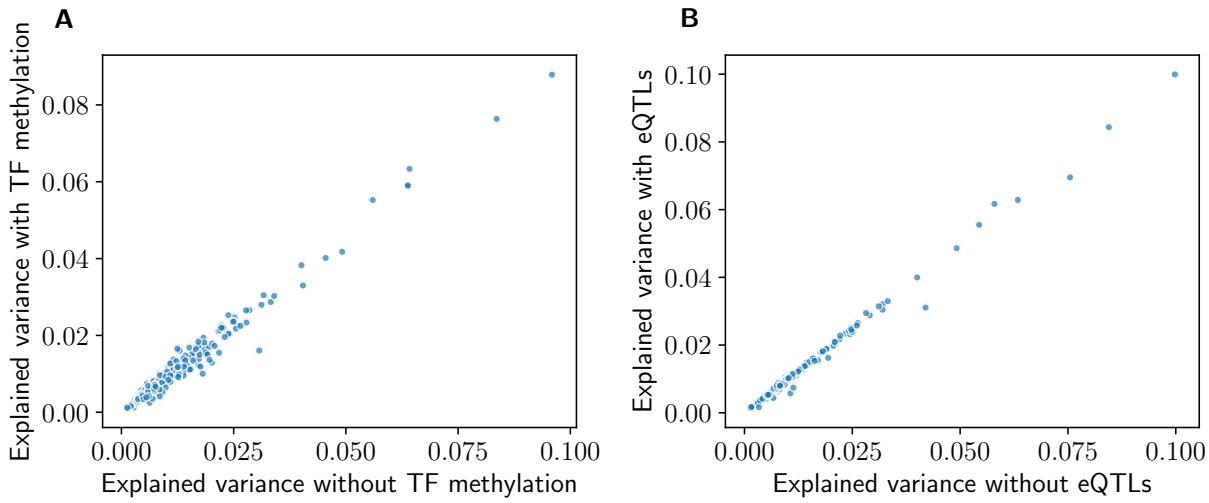

Figure 3.18: (A) The percentage of expression variance of TGs explained by coverage anomaly status when the linear model does not include TFs' methylation states (x-axis) and when the linear model includes TFs' methylation states (y-axis). The points are close to the diagonal line, which indicates that the explained expression variance from coverage anomaly does not have big differences between the two models. (B) The percentage of expression variance of TGs explained by coverage anomaly status when the linear model when the linear model does not include known eQTLs of TGs (x-axis) and when the linear model includes known eQTLs (y-axis).

anomaly-TG triples such that the mutations occur in any eQTLs for the TG. It suggests that the prediction power of coverage anomaly is not originated from the known eQTLs of the TGs.

If the explanation power of coverage anomalies cannot be attributed to the methylation status of the top 5 relevant methylation sites in TFs and known eQTLs that affect TGs' expression, what genetic or epigenetic features are related to or leads to the explanation power? There are many genetic and epigenetic features that we do not investigate, such as unknown eQTLs, SNVs that are not eQTLs, histone modificiation status. Or a combination of many features contribute together to the explanation power of coverage anomalies. If the coverage anomalies are caused by unannotated isoforms, the expression abundances and coverage status of certain splicing factors may contribute to the prediction power of coverage anomalies. Overall, further investigations are needed to answer this question.

The observations indicate that coverage anomalies of TFs contain unique information that is indicative to the expression of TGs compared to the TFs' methylation statuses and eQTLs. Coverage anomalies are indispensable information to predict the expression of TGs at least for some of the TF-coverage anomaly-TG triples.

3.2.7 Details of statistical analysis

Testing the significance of regulation of each individual TF

We only analyze the prediction power of coverage anomalies when the corresponding TF gene expression itself have significant prediction power of TG expression in linear models. To test

the significance of TF gene expression, we use a simplified linear model. Suppose there are n samples, let $y \in \mathbb{R}^n$ be the expression of TG after log transformation. Suppose the target gene is regulated by d TFs, and let $X_g \in \mathbb{R}^{n \times d}$ be the log-transformed expression matrix of the involved TFs. The simplified linear model is:

$$y = \beta_0 + \beta_g^T X_g + \epsilon, \tag{3.16}$$

where error term ϵ is assumed to follow a Gaussian distribution under log-transformed TPM.

Again, this model is an approximation to TF regulation, instead of modeling the exact biological interaction and mechanisms. The exact biological interaction is between TG DNA sequence and TF protein molecules, of which the abundances are not available in the given samples, and the relationship between mRNA abundances and TF protein abundances may be more complex than linear relationships. Nevertheless, linear model between gene expressions has been used in other applications such as reconstructing gene regulatory network from RNA-seq data [49, 72]. The linear approximation between log-transformed expression indicate exponential approximation between the normalized read counts of the TF and TG genes.

The significance of TF expression in regulation TG expression is tested through likelihood ratio test, and the test statistics for the i^{th} TF is

$$\lambda = -2 \ln \frac{\sup_{\beta_0, \beta_g} \mathcal{L}(y \mid \beta_0, \beta_g; X_g)}{\sup_{\beta_0, \beta_{g,-i}} \mathcal{L}(y \mid \beta_0, \beta_{g,-i}; X_{g,-i})}. \tag{3.17}$$

where $\beta_{g,-i}$ and $X_{g,-i}$ are the parameters and feature matrix without the i^{th} feature, and $\mathcal{L}$ denotes the likelihood. Under null hypothesis, λ follows a χ^2 distribution as proved by Wilks [167]. When λ is significantly large under χ^2 distribution, there is strong linear relationship in the regulation of the i^{th} TF in the current dataset. We use p-value threshold of 0.05 to determine the TFs that significantly explain TG's expression.

Testing the significance of anomalies

The prediction power of coverage anomaly to TG's expression is tested also with likelihood ratio test under the linear models of (3.15) and D forms in Table 3.2. For the model containing both concentration-independent term and two concentration-dependent terms, the test statistic is

$$\lambda = -2 \ln \frac{\sup_{\beta_0, \beta_p, \beta_m, \beta_c, \beta_g, \beta_t, \beta_a, \beta_{g,int}, \beta_{t,int}} \mathcal{L}(y)}{\sup_{\beta_0, \beta_p, \beta_m, \beta_c, \beta_g, \beta_t} \mathcal{L}(y)}. \tag{3.18}$$

Under the null model that the parameter vector $(\beta_a, \beta_{g,int}, \beta_{t,int})$ is a zero vector, the above test statistics follows a χ^2 distribution with degrees of freedom equal to 3. For the other forms of D terms, the test statistics can be constructed by removing the coefficient from the numerator of anomaly terms that D does not contain and decreasing the degrees of freedom correspondingly.

Benjamini-Hochberg FDR adjustment is performed within the testings under each type of anomaly contribution (each form of D). With FDR threshold 0.05, the significant detections under each model assumption are combined as the final set of TF-coverage anomaly-TG triples such that the regulation efficiency is altered when TF contains coverage anomalies.

Determining the "maximum relevance minimum redundant model" for each TF-coverage anomaly-TG triple

For some TF-coverage anomaly-TG triples, FDR of the anomaly terms is significant under various linear models corresponding to the anomaly contribution types. It is possible that one of the anomaly terms is crucial, and appending another unimportant anomaly terms still leads to a significant prediction under the appended model. Therefore, we select the minimum redundant model that have the maximum explaining power of the TG expression. The selection is done in the following ways.

We first collect all models such that (1) FDR is significant under this model or the model is the full model (containing all three anomaly terms), and (2) likelihood ratio testing between the this model and the full model is insignificant. In case where the full model does not have a significant FDR but likelihood ratio testing between each significant model with the full model results in significant difference, we still treat all three anomaly terms to be all crucial and directly assign the maximum relevance minimum redundant model to be the full model even though it does not pass FDR threshold. Let the set of the models under these criteria be $\mathcal{M}$. These are the set of models with large and statistically equivalent explanation power on TG expression.

We then select among $\mathcal{M}$ the model with the minimum redundant number of anomaly features. Whether an anomaly term has significantly large and unique information is determined by likelihood ratio testing between the model and the model without the anomaly term. Let $\mathcal{F}_M$ be the set of anomaly terms within $M \in \mathcal{M}$. And the cardinality (or the number of terms) $|\mathcal{F}_M|$ is no greater than 3. The objective of selecting minimum redundant model can be expressed as:

$$\max_{M \in \mathcal{M}} \frac{\sum_{f \in \mathcal{F}_M} \mathbb{1}(Significant(M, M_{-f}))}{|\mathcal{F}_M|} + \lambda \mathcal{LL}(y \mid M), \tag{3.19}$$

where M_{-f} is the model without f term, $Significant()$ function is an indicator function to indicate whether the likelihood ratio testing between the two models reveals significant difference, and $\mathcal{LL}$ is the log-likelihood. The first part of the above objective is the percent of anomaly terms that significantly contribute to the model (or non-redundant), and it is discrete because of the fact that $|\mathcal{F}_M| \leq 3$. Therefore, to avoid non-uniqueness of the models that contain the largest proportion of non-redundant anomaly terms, we add the second term with a very small λ. The second term will select the model leads to largest log-likelihood to explain TG expression, despite that the difference of explanation power among the models in $\mathcal{M}$ is not statistically distinguishable. Model selection based on the above objective leads to the minimum redundant model for each TF-coverage anomaly-TG triple.

Determining the changing direction of TF regulation efficiency when having coverage anomalies

The direction of regulation, positive or negative regulation, is determined by the coefficient in linear regression. Given an anomaly score vector and its corresponding gene/transcript expression of interest, we fit a new linear model with only the expression vector of the corresponding TF and anomaly-related term D under the maximum relevant minimum redundant model. Let y be the TG expression, and suppose the coverage anomaly of the i^{th} TF is to be evaluated, the

new linear model is:

$$y = \tilde{\beta}_0 + \tilde{\beta}_{g,i} X_{g,i} + D + \epsilon \tag{3.20}$$

The sign of fitted coefficient $\tilde{\beta}_{g,i}$ indicates whether the i^{th} TF positively or negatively regulates the expression of TG. When there is one anomaly term, the agreement between the sign of $\tilde{\beta}_{g,i}$ and that of the sign of coefficient of the anomaly term represents the enhancement or reduction of regulation efficiency of TF or of the expression of TG. When there are multiple anomaly terms, and the coefficient of anomaly depends on gene and/or transcript expression of the TF, it is not possible to determine a single coefficient across the whole space of expression abundances. Instead, we select several representative quantiles of gene expression and transcript expression percentages to evaluate the sign agreement between the gene coefficient and anomaly coefficient.

We keep only the relevant terms in linear model fitting, otherwise, the signs of coefficients will be confounded by the potential correlation between the relevant terms and expression abundances of other TF genes and transcripts. For example, if the TG expression is y, the gene-level expression of i^{th} TF is $X_{g,i} = 1.5y$, and there is another TF whose expression is $X_{g,j} = 0.5y$. Both i^{th} TF and j^{th} TF positively regulates and is positively correlated with TG expression. However, one optimal solution for linear regression may set the coefficients as $y = X_{g,i} - X_{g,j}$, and the negative coefficient of the j^{th} TF disagrees with its positive regulation effect. The disagreement is due to including extra features that are not of interest but have correlations with the interested TF expression. Therefore, we remove the expression vectors that are not of interest from the linear model to determine the regulation direction.

3.2.8 Discussion

Our analysis reveals that the regulation efficiency of TFs can be affected by the coverage anomaly status of their transcript expression, and we identify 319 TF-coverage anomaly-TG triples of which the effect is significant in breast cancer. We observe that the effect has various characters: some is independent of TF expression the effect of other triples increases along with TF gene expression or transcript expression. Containing coverage anomalies can either enhance or reduce the regulation efficiency, while the reduction effect is more common under medium to low TF expression. This is the first study to analyze the relationship between transcript coverage status and the TF regulation efficiency in breast cancer.

Linear models are used to approximate the regulation relationship between TFs and their target genes (after log-transformation on expression), however, the true regulation relationship may be more complicated. The direction for further analyses may incorporate a more refined expression regulation model and take into account the competition, cooperation, or coactivation among multiple TFs.

One possible future direction is to compare the effect of coverage anomalies between tumor and normal conditions or among different tumor types. Similar to the effect of eQTLs, the effect of coverage anomaly status may also depend on the cellular environment, which includes in DNA accessibility, binding site occupancy, and abundances of co-factors of TFs. And the environment possibly varies greatly between tumor and normal states or among different tissues. It remains to be answered whether coverage anomalies affect the set of TFs and TGs and whether the directions of effect remain the same.

3.2.9 Appendix

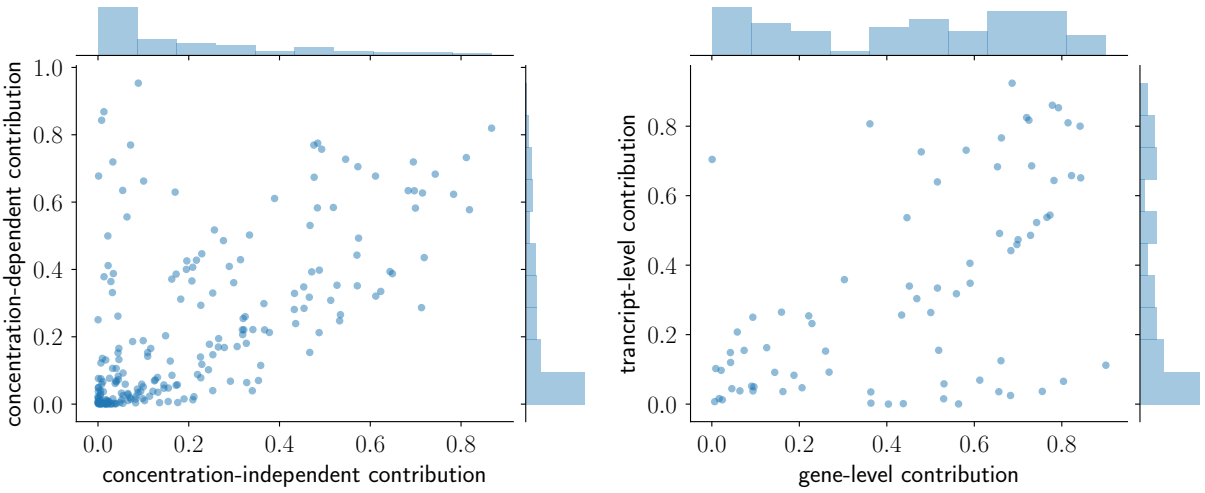

Figure 3.19: (A) Percentage of explained variance from concentration-independent term and concentration-dependent term separately among the total explained variance from both terms. (B) Percentage of explained variance from gene-level and transcript-level concentration-dependent term separately among the total explained variance of both levels of concentration dependent terms.

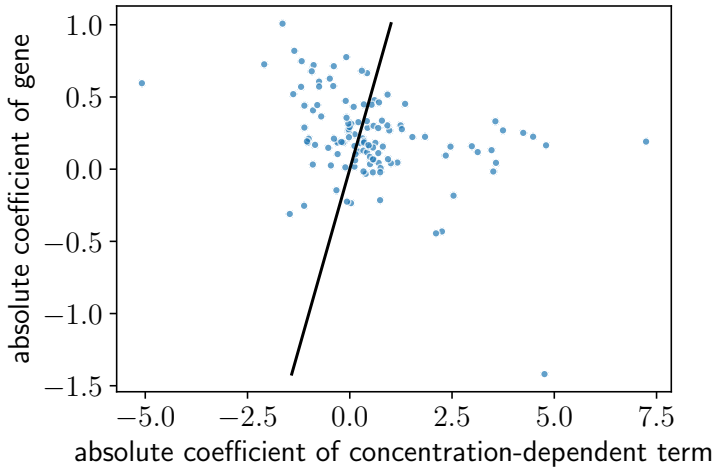

Figure 3.20: Scatterplot of the absolute coefficient of the TF gene expression and of the gene-level concentration-dependent term when the maximum relevance minimum redundant model contain this anomaly term.

Chapter 4

Conclusion and future work

4.1 Summary of contributions

Reconstructing sequences and expression of transcripts has become a key computational task since the development of the RNA-seq technique. The algorithms for identifying sequence variants and estimating expression of transcripts are ever-improving, and the analyses based on reconstructed transcripts using current methods have revealed an unprecedented amount of associations between sequence variants and diseases, between expression and drug responses. Nevertheless, it is still challenging to infer transcript sequences and expression with high accuracy. We study the transcript reconstruction problem from an anomaly detection perspective: identifying the disagreement between RNA-seq observations and the sequence and expression reconstructed under existing methods. This approach allows us to both improve the identified variants and estimated expression for a subset of transcripts, and can be used to inspire and evaluate future transcript reconstruction methods.

This thesis focuses on large-scale transcriptomic sequence variation (TSVs) detection and expression quantification. Many RNA-seq alignments that are discordant with sequencing library preparation cannot be explained by gene fusions. The unexplained discordant RNA-seq reads leads to the detection of non-fusion-gene TSVs. After an expression quantifier assigns a proportion of RNA-seq reads to each transcript they can be mapped to, the coverage along each transcript may not agree with the probability model of RNA-seq fragment generation protocol. The unexplained coverage patterns correspond to the coverage anomalies.

Our method, SQUID, was the first TSV detection method that is optimized for identifying both fusion-genes and non-fusion-gene TSVs. It models the TSVs as rearrangements of genome segments and seeks an rearrangement that makes the most number of RNA-seq reads agree with the RNA-seq library preparation. Experimental validation demonstrates that the rearrangement problem formulation along with the filtering steps in SQUID achieves reasonable accuracy in detecting fusion-gene TSVs and relatively high accuracy in detecting non-fusion-gene TSVs compared to other fusion detection methods. Applying SQUID on TCGA cancer RNA-seq datasets, we observed that a few non-fusion-gene TSVs occur on tumor suppressor genes, which motivates further analysis on these TSVs and their effect on cancer genesis and progression. The set of detectable TSVs is enlarged by including non-fusion-gene TSVs, which enables an alternative

potential mechanism to explain diseases and cellular status.

SQUID was further extended to MULTIPLE COMPATIBLE ARRANGEMENTS PROBLEM (MCAP) for detecting TSVs under the assumption that sequencing samples contains multiple alleles such that different alleles contain different set of TSVs. Theoretical analyses of SQUID and MCAP reveals sufficient and necessary conditions when RNA-seq data cannot be explained a single rearrangement of genome segments. We also proved the NP-completeness of the rearrangement problem in SQUID as well as MCAP, and provided approximation algorithms for MCAP with the number of alleles is 2, corresponding to diploid assumption.

We developed a method, SAD to detect the unexpected coverage patterns of transcript expression quantification. SAD incorporates the state-of-art sequencing bias estimation model of Salmon quantifier to estimate expected coverage and compares the observed coverages with the expected ones. Expression quantification methods have been improved during the last decade, while there have been only a few works identifying potential quantification errors. SAD provides a novel perspective in evaluating the accuracy of expression estimates in absence of ground truth expression. It further categorizes the detected coverage anomalies by whether the anomaly is likely caused by quantification algorithm mistake (adjustable anomalies) or by other causes (unadjustable anomalies, for example due to incomplete reference that cannot be addressed by improving quantification algorithms). SAD also adjusts the expression estimates for transcripts that contain adjustable coverage anomalies, and the adjusted expression estimates have a higher agreement with the expected coverage and can be used in further expression analysis. Unadjustable coverage anomalies deserve further analysis both to discover the biological mechanism that leads to the abnormal coverage pattern and to investigate the biological consequences when genes and transcripts contain the anomalies.

We did the first analysis to probe the biological consequences when genes contain unadjustable coverage anomalies. Specifically, we analyzed whether coverage anomalies of transcription factors contribute in explaining the expression variability of corresponding target genes. Our analysis revealed 319 transcription factor-coverage anomaly-target gene triples in TCGA breast cancer samples where the status of unadjustable anomaly has significant explanation power on the expression variance of the target genes. When a transcription factor contains unadjustable coverage anomalies, the change in target gene's expression can either purely depend on the anomaly score or increase with both anomaly score and the expression of transcription factors. This analysis motivates further investigation on physical and chemical mechanisms for the explanation power of coverage anomalies on expression regulation, as well as the predictability of coverage anomalies on other biological processes.

4.2 Future directions

There are multiple future directions for applying and designing anomaly detection methods for RNA-seq data as well as more the general genomic area of computational biology. Anomalies in RNA-seq include other aspects besides discordant alignments and unexpected coverage patterns. For example, unmapped reads are also unexpected patterns. Having a comprehensive repertoire of RNA-seq anomalies and connecting the ensemble of anomalies to the biological events that cause them will benefit researches on both biological discoveries and method development for

RNA-seq data. Additionally, the anomaly detection methods in this thesis are rule-based approaches that incorporate existing knowledge about RNA-seq technique, such as the existence of sequence fusion and the probabilistic model of sequencing fragment generation. Developing an anomaly detection framework that automatically learns the rules in computational biological area is another research direction for fast adaptation of anomaly detection to the ever-improving biological data types.

There is still space for improving accuracy or speed of SQUID and SAD. For example, current gene annotation can be incorporated in SQUID and the detection criteria for fusion-gene TSV and non-fusion-gene TSV detection can be separated. For another example, deriving approximation for the probabilistic model in SAD will likely speed up the empirical p-value calculation. The details of future directions for method improvements have been discussed in previous chapters separately, and we do not recapitulate in this section.

With the advent of single-cell RNA-seq (scRNA) data, it is important to extend the current transcript reconstruction algorithms to the new experimental technique. Because of the largely reduced amount of RNA molecules in each individual cell and decreased sequencing depth, scRNA usually cannot capture RNA reads for all expressed transcripts, which induces additional challenges for reconstructing transcript sequences, sequence variants and expression. It remains a question what degree of precision and accuracy can be achieved in TSV detection or transcriptome assembly using scRNA data. Nevertheless, scRNA enables a much larger space for biological discovery regarding cellular heterogeneity, inter-cellular communication, and differentiation. Studying the differences of transcript sequences between single cells will potentially benefit the decoding of the function of each transcript and the role in cellular communication and differentiation. Methods for imputing gene expression have been proposed: even though an expressed gene is not captured in scRNA, the expression can be inferred from similar cells where the gene is captured in sequencing. Such imputation methods can be similarly developed for TSV detection and transcriptome assembly, or be used in expression anomaly detection to identify potential misquantification of a group of similar cells.

Another potential is to adapt current methods to develop new methods to detect TSV or coverage anomalies in multiple samples jointly. This is similar to TSV and coverage anomaly in similar single cells. But multiple related bulk RNA-seq datasets does not have the problem of uncaptured genes as in scRNA, while the sample compositions of different samples may emit more different sets of transcripts compared to the transcripts in different single cells. It is unknown whether transcript reconstruction and variation detections methods designed for multiple single cells can be directly applied to multiple bulk RNA-seq samples. Nevertheless, jointly identifying TSV or coverage anomalies for multiple related bulk RNA-seq samples with proper methods will lead to more robust and confident detections, and also ease the comparison of TSV or anomaly among these samples.

With the developed computational methods to identify novel types of TSVs and coverage anomalies, analyzing the TSVs and anomalies in currently available datasets will potentially lead to novel biological discoveries. For example, analyzing whether a novel alternative cleavage of an cancer-related gene suggested by SAD unadjustable anomalies is associated with a certain cancer subtype may reveal novel cancer biomarkers. Investigating these biological questions can unveil the association between cellular states and phenotypes more comprehensively. Studying the biophysical and biochemical effects of TSVs and coverage anomalies in the molecular level

helps understand the mechanism of such associations. For example, a TSV-fused transcript sequence may have altered translation regulatory elements and lead to altered abundances of translated protein, or directly alter protein sequences such that the protein has an altered localization or binding properties. Studying these molecular-level alterations leads to a more complete view of cellular interactions and system regulations.

Bibliography

[1] The Illumina Body Map 2.0 data, 2011. URL `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513`. 1.3, 3.1.7

[2] Bronwen L Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, et al. The Ensembl gene annotation system. *Database*, 2016, 2016. 2.1.7

[3] Sahar Al Seesi, Yvette Temate Tiagueu, Alexander Zelikovsky, and Ion I. Măndoiu. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics*, 15(8):S2, Nov 2014. 3

[4] Dmitry Antipov, Anton Korobeynikov, Jeffrey S McLean, and Pavel A Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 2016. 1.4.1

[5] Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6:8971, 2015. 2.2.7

[6] Ahmet Ay and David N Arnosti. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Reviews in Biochemistry and Molecular Biology*, 46(2): 137–151, 2011. 3.2.1

[7] Vineet Bafna and Pavel A Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272–289, 1996. 2.1.1

[8] Lucy W Barrett, Sue Fletcher, and Steve D Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, 69(21):3613–3634, 2012. 3.2.1

[9] Christoph Bartenhagen and Martin Dugas. RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics*, 29(13):1679–1681, 2013. 2.1.7

[10] Craig H Bassing, Wojciech Swat, and Frederick W Alt. The mechanism and regulation of chromosomal V (D) J recombination. *Cell*, 109(2):S45–S55, 2002. 1.2

[11] Matteo Benelli, Chiara Pescucci, Giuseppina Marseglia, Marco Severgnini, Francesca Torricelli, and Alberto Magi. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics*, 28(24):3232–3239, 2012. **??**, 2

[12] Yuval Benjamini and Terence P Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, 2012. 1.3, 1.4.3

[13] Graham R Bignell, Thomas Santarius, Jessica CM Pole, Adam P Butler, Janet Perry, Erin

Pleasance, Chris Greenman, Andrew Menzies, Sheila Taylor, Sarah Edkins, et al. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Research*, 17(9):1296–1303, 2007. 2.1.8

[14] Dmitriy A Bolotin, Stanislav Poslavsky, Alexey N Davydov, Felix E Frenkel, Lorenzo Fanchi, Olga I Zolotareva, Saskia Hemmers, Ekaterina V Putintseva, Anna S Obraztsova, Mikhail Shugay, et al. Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology*, 35(10):908–911, 2017. 1.2

[15] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016. 1.1, 1.4.3, 3

[16] Jean-Simon Brouard, Flavio Schenkel, Andrew Marete, and Nathalie Bissonnette. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *Journal of animal science and biotechnology*, 10(1):44, 2019. 1.4.4

[17] Cédric Cabau, Frédéric Escudié, Anis Djari, Yann Guiguen, Julien Bobe, and Christophe Klopp. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. *PeerJ*, 5:e2988, 2017. 3

[18] Latarsha J Carithers, Kristin Ardlie, Mary Barcus, Philip A Branton, Angela Britton, Stephen A Buia, Carolyn C Compton, David S DeLuca, Joanne Peter-Demchok, Ellen T Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreservation and Biobanking*, 13(5):311–319, 2015. 1.3

[19] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10, 2019. 1.4.1, 3.1.12

[20] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58, 2009. 1.1

[21] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9):677–681, 2009. 2

[22] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):1–21, 2018. 1.4.3

[23] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966–968, 2015. 2.1.7, 2.1.7, 2.1.8, 2.1.12

[24] Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes

Project data. *Nucleic Acids Research*, 45(D1):D854–D859, 2016. 1.3, 3.1.12

[25] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015. 3.1.12

[26] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017. 1.2, 3.2.1

[27] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE*, 12(12): e0190152, 2017. 3

[28] Matthew Dapas, Manoj Kandpal, Yingtao Bi, and Ramana V Davuluri. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Briefings in Bioinformatics*, 18(2):260–269, 2016. 3.1.5

[29] Nadia M Davidson, Ian J Majewski, and Alicia Oshlack. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Medicine*, 7(1):1–12, 2015. **??**, 2, 2.1.8

[30] Smadar Ben-Tabou de Leon and Eric H Davidson. Modeling the dynamics of transcriptional gene regulatory networks for animal development. *Developmental Biology*, 325(2): 317–328, 2009. 3.2.1

[31] Dan DeBlasio, Kwanho Kim, and Carl Kingsford. More accurate transcript assembly via parameter advising. *Journal of Computational Biology*, 2020. 1.4.1

[32] Michael WN Deininger, John M Goldman, and Junia V Melo. The molecular biology of chronic myeloid leukemia. *Blood, The Journal of the American Society of Hematology*, 96(10):3343–3356, 2000. 1.2, 2

[33] Fernando M Delgado and Francisco Gómez-Vela. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine*, 95:133–145, 2019. 1.4.3

[34] 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*, 29(1):15–21, 2013. 1.4.2, 2.1.7, 2.1.12, 3.1.14, 3.1.14

[35] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurlien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47 (D1):D427–D432, 2018. doi: 10.1093/nar/gky995. URL `http://dx.doi.org/10.1093/nar/gky995`. 3.1.7

[36] Scott J Emrich, W Brad Barbazuk, Li Li, and Patrick S Schnable. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1):69–73, 2007. 1.3

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

[38] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane

Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2018. 1.4.1, 3.1.14, 3.2.3

[39] Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17): 2778–2784, 2015. 1.3, 3.1.14

[40] Pedro AF Galante, Raphael B Parmigiani, Qi Zhao, Otávia L Caballero, Jorge E De Souza, Fábio CP Navarro, Alexandra L Gerber, Marisa F Nicolás, Anna Christina M Salim, Ana Paula M Silva, et al. Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic Acids Research*, 39(14):6056–6068, 2011. 2.1.8

[41] Adi F Gazdar, Venkatesh Kurvari, Arvind Virmani, Lauren Gollahon, Masahiro Sakaguchi, et al. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International Journal of Cancer*, 78(6):766–774, 1998. 2.2.7

[42] Paul Geeleher, Aritro Nath, Fan Wang, Zhenyu Zhang, Alvaro N Barbeira, Jessica Fessler, Robert L Grossman, Cathal Seoighe, and R Stephanie Huang. Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biology*, 19(1):1–14, 2018. 3.2.1

[43] Matthew Geniza and Pankaj Jaiswal. Tools for building de novo transcriptome assembly. *Current Plant Biology*, 11:41–45, 2017. 3

[44] B Michael Ghadimi, Marian Grade, Michael J Difilippantonio, Sudhir Varma, Richard Simon, Cristina Montagna, Laszlo Füzesi, Claus Langer, Heinz Becker, Torsten Liersch, et al. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. *Journal of Clinical Oncology*, 23(9):1826, 2005. 1.4.3

[45] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012. 3

[46] Jing Gong, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, Renyan Liu, Lixia Diao, An-Yuan Guo, et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Research*, 46(D1):D971–D976, 2018. 1.2, 3.2.1, 3.2.6

[47] Narjol Gonzalez-Escalona, George John Kastanis, Ruth Timme, Dwayne Roberson, Maria Balkey, and Sandra M Tallent. Closed genome sequences of 28 foodborne pathogens from the CFSAN verification set, determined by a combination of long and short reads. *Microbiology Resource Announcements*, 9(18), 2020. 1.4.1

[48] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644, 2011. 1.4.1, 2, 2.1.7

[49] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, 2013. 3.2.7

[50] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083, 2012. 1.3, 2.1.7

[51] Wenbin Guo, Cristiane PG Calixto, John WS Brown, and Runxuan Zhang. TSIS: an R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics*, 33(20):3308–3310, 2017. 1.4.3, 3.2.2

[52] Aysegul Guvenek and Bin Tian. Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quantitative Biology*, 6(3): 253–266, 2018. 3.1.7, 3.1.9

[53] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. 2.2.6

[54] Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, et al. TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5:11432, 2015. 3.2.3

[55] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12): e131–e131, 2010. 1.3, 1.4.3

[56] Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, 2(3):666–673, 2012. 1.3

[57] Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, 18(suppl_1):S181–S188, 2002. 1.4.1

[58] Asaf Hellman and Andrew Chess. Gene body-specific methylation on the active x chromosome. *Science*, 315(5815):1141–1143, 2007. 3.2.1

[59] James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, and Magnus Rattray. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, 31(24):3881–3889, 2015. 3

[60] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014. 3, 3.2.1

[61] Monica Hollstein, David Sidransky, Bert Vogelstein, and Curtis C Harris. p53 mutations in human cancers. *Science*, 253(5015):49–53, 1991. 2.1.10

[62] Fereydoun Hormozdiari, Iman Hajirasouliha, Phuong Dao, Faraz Hach, Deniz Yorukoglu, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–

i357, 2010. 2, 2.2

[63] Kuan-lin Huang, R Jay Mashl, Yige Wu, Deborah I Ritter, Jiayin Wang, Clara Oh, Marta Paczkowska, Sheila Reynolds, Matthew A Wyczalkowski, Ninad Oak, et al. Pathogenic germline variants in 10,389 adult cancers. *Cell*, 173(2):355–370, 2018. 3.2.6

[64] Matthew K Iyer, Arul M Chinnaiyan, and Christopher A Maher. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011. **??**, 1.4.2, 2

[65] Iris E Jansen, Hui Ye, Sasja Heetveld, Marie C Lechler, Helen Michels, Renée I Seinstra, Steven J Lubbe, Valérie Drouet, Suzanne Lesage, Elisa Majounie, et al. Discovery and functional prioritization of Parkinsons disease candidate genes from large-scale whole exome sequencing. *Genome Biology*, 18(1):22, 2017. 3.1.7

[66] Wenlong Jia, Kunlong Qiu, Minghui He, Pengfei Song, Quan Zhou, Feng Zhou, Yuan Yu, Dandan Zhu, Michael L Nickerson, Shengqing Wan, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biology*, 14(2): R12, 2013. **??**, 1.4.2, 2, 2.1.8

[67] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009. 1.1, 1.4.3, 3

[68] John D Kececioglu and Eugene W Myers. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13(1-2):7, 1995. 2.2.2

[69] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 2019. 1.4.2

[70] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017. 1.4.1

[71] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, 2004. 2, 2.1.7, 2.1.7, 2.1.7, 2.1.8

[72] Kari Y Lam, Zachary M Westrick, Christian L Müller, Lionel Christiaen, and Richard Bonneau. Fused regression for multi-source gene regulatory network inference. *PLoS Computational Biology*, 12(12):e1005157, 2016. 3.2.7

[73] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter ACt Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013. 3.1.7, 3.1.8

[74] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):1, 2014. 2, 2.1.7, 2.1.7

[75] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1.4.1

[76] Tong Ihn Lee and Richard A Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013. 3.2.1

[77] Laura H LeGault and Colin N Dewey. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. *Bioinformatics*, 29(18):2300–2310, 2013. 1.1, 1.4.3

[78] 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, 2011. 1.1, 1.4.3, 3, 3.1.1, 3.1.11

[79] 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, 2009a. 3, 3

[80] Bo Li, Taiwen Li, Binbin Wang, Ruoxu Dou, Jian Zhang, Jun S Liu, and X Shirley Liu. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. *Nature Genetics*, 49(4):482–483, 2017. 1.2

[81] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. 2.1.7

[82] Marilyn M Li, Michael Datto, Eric J Duncavage, Shashikant Kulkarni, Neal I Lindeman, Somak Roy, Apostolia M Tsimberidou, Cindy L Vnencak-Jones, Daynna J Wolff, Anas Younes, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics*, 19(1):4–23, 2017. 1.2

[83] Qiyuan Li, Ji-Heui Seo, Barbara Stranger, Aaron McKenna, Itsik Peer, Thomas LaFramboise, Myles Brown, Svitlana Tyekucheva, and Matthew L Freedman. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3):633–641, 2013. 3.2.1

[84] Wei Li and Tao Jiang. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics*, 28(22):2914–2921, 2012. 1.3, 1.4.3

[85] 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, 2011. 1.4.1

[86] Yafang Li, Xiayu Rao, William W Mattox, Christopher I Amos, and Bin Liu. RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. *PloS One*, 10(9):e0136653, 2015. 1.4.3

[87] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, 2008. 1.3

[88] Jiangang Liu, Narayanan B Perumal, Christopher J Oldfield, Eric W Su, Vladimir N Uversky, and A Keith Dunker. Intrinsic disorder in transcription factors. *Biochemistry*, 45(22): 6873–6888, 2006. 3.2.1

[89] Juntao Liu, Ting Yu, Tao Jiang, and Guojun Li. TransComb: genome-guided transcrip-

tome assembly via combing junctions in splicing graphs. *Genome Biology*, 17(1):213, 2016. 1.4.1

[90] Peng Liu, Rajendran Sanalkumar, Emery H Bresnick, Sündüz Keleş, and Colin N Dewey. Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq. *Genome research*, 2016. 1.1, 1.4.3, 3

[91] Silvia Liu, Wei-Hsiang Tsai, Ying Ding, Rui Chen, Zhou Fang, Zhiguang Huo, SungHwan Kim, Tianzhou Ma, Ting-Yu Chang, Nolan Michael Priedigkeit, et al. 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):e47–e47, 2016. 1.4.2, 2.1.8

[92] Yi Liu, Kenneth Barr, and John Reinitz. Fully interpretable deep learning model of transcriptional control. *Bioinformatics*, 36(Supplement_1):i499–i507, 2020. 3.2.1

[93] WW Lockwood, R Chari, BP Coe, L Girard, C Macaulay, S Lam, AF Gazdar, JD Minna, and WL Lam. DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. *Oncogene*, 27(33):4615–4624, 2008. 1.2

[94] Hélène Lopez-Maestre, Lilia Brinza, Camille Marchet, Janice Kielbassa, Sylvère Bastien, Mathilde Boutigny, David Monnin, Adil El Filali, Claudia Marcia Carareto, Cristina Vieira, et al. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*, 44(19):e148–e148, 2016. 1.4.4

[95] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. 1.4.3, 3.1.8

[96] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, 34(12):1287–1291, 2016. 1.4.3, 3, 3

[97] Cong Ma and Carl Kingsford. Detecting, categorizing, and correcting coverage anomalies of RNA-Seq quantification. *Cell Systems*, 9(6):589–599, 2019. 3, 3.2.2

[98] Cong Ma, Mingfu Shao, and Carl Kingsford. SQUID: transcriptomic structural variation detection from RNA-seq. *Genome Biology*, 19(1):52, 2018. 2

[99] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(suppl_1):D54–D58, 2005. 2.1.10

[100] Valdemar Máximo, Jorge Lima, Paula Soares, André Silva, Ines Bento, and Manuel Sobrinho-Simoes. GRIM-19 in health and disease. *Advances in Anatomic Pathology*, 15(1):46–53, 2008. 3.1.8

[101] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012. 1.4.3

[102] Andrew McPherson, Fereydoun Hormozdiari, Abdalnasser Zayed, Ryan Giuliany, Gavin

Ha, Mark GF Sun, Malachi Griffith, Alireza Heravi Moussavi, Janine Senz, Nataliya Melnyk, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*, 7(5):e1001138, 2011. **??**, 2, 2.1.8

[103] Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020. 1.4.3

[104] Fredrik Mertens, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15(6):371–381, 2015. 1.2, 1.4.2, 2

[105] Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), 2018. 1.3

[106] Lisa D Moore, Thuc Le, and Guoping Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013. 3.2.1

[107] Ignasi Morán, İldem Akerman, Martijn van de Bunt, Ruiyu Xie, Marion Benazra, Takao Nammo, Luis Arnes, Nikolina Nakić, Javier García-Hurtado, Santiago Rodríguez-Seguí, et al. Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metabolism*, 16 (4):435–448, 2012. 3, 3.2.1

[108] Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yosuke Kawai, and Masao Nagasaki. A Bayesian approach for estimating allele-specific expression from RNA-seq data with diploid genomes. In *BMC Genomics*, volume 17, pages 7–17. BioMed Central, 2016. 1.4.4

[109] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013. 1.2, 3.2.1

[110] Daniel Nicorici, Mihaela Şatalan, Henrik Edgren, Sara Kangaspeska, Astrid Murumägi, Olli Kallioniemi, Sami Virtanen, and Olavi Kilkku. FusionCatcher–a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *BioRxiv*, page 011650, 2014. **??**, 2, 2.1.8

[111] Malgorzata Nowicka and Mark D Robinson. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5, 2016. 1.4.3

[112] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2015. 1.4.1, 3.1.8

[113] 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*, 14 (4):417–419, 2017. 1.1, 1.4.3, 3, 3.1.1, 3.1.7, 3.1.8, 3.1.10, 3.1.11, 3.1.14, 3.1.14, 3.1.14

[114] Geo Pertea. GffCompare. https://ccb.jhu.edu/software/stringtie/gffcompare.shtml, 2018. 3.1.14

[115] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015. 1.4.1, 3.1.9, 3.1.14

[116] Adam M Phillippy, Michael C Schatz, and Mihai Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, 9(3):R55, 2008. 3

[117] Xavier Pichon, Lindsay A Wilson, Mark Stoneley, Amandine Bastide, Helen A King, Joanna Somers, and Anne E Willis. RNA binding protein/RNA element interactions and the control of translation. *Current Protein and Peptide Science*, 13(4):294–304, 2012. 3.2.1

[118] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14 (7):687, 2017. 1.4.3

[119] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. Reliable identification of genomic variants from RNA-seq data. *The American Journal of Human Genetics*, 93(4):641–651, 2013. 1.4.4

[120] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057, 2010. 3.2.2

[121] Yutong Qiu, Cong Ma, Han Xie, and Carl Kingsford. Detecting transcriptomic structural variants in heterogeneous contexts via the Multiple Compatible Arrangements Problem. *Algorithms for Molecular Biology*, 15:1–15, 2020. 2

[122] Aaron R Quinlan, Royden A Clark, Svetlana Sokolova, Mitchell L Leibowitz, Yujun Zhang, Matthew E Hurles, Joshua C Mell, and Ira M Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, 20 (5):623–635, 2010. 2

[123] Narayanan Raghupathy, Kwangbom Choi, Matthew J Vincent, Glen L Beane, Keith S Sheppard, Steven C Munger, Ron Korstanje, Fernando Pardo-Manual de Villena, and Gary A Churchill. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, 34(13):2177–2184, 2018. 1.4.4

[124] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012. 2, 2.1.7, 2.1.7

[125] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423, 2015. 1.2

[126] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and

Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015. 1.4.3

[127] Christelle Robert and Mick Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, 16(1):177, 2015. 3

[128] Jacques Robert, Antoine Vekris, Philippe Pourquier, and Jacques Bonnet. Predicting drug response based on gene expression. *Critical Reviews in Oncology/Hematology*, 51(3): 205–227, 2004. 1.4.3

[129] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71, 2013. 1.1, 1.4.3, 3, 3.1.14

[130] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, et al. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912, 2010. 1.4.1, 2, 2.1.7, 2.1.8

[131] Dan R Robinson, Shanker Kalyana-Sundaram, Yi-Mi Wu, Sunita Shankar, Xuhong Cao, Bushra Ateeq, Irfan A Asangani, Matthew Iyer, Christopher A Maher, Catherine S Grasso, et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nature Medicine*, 17(12):1646–1651, 2011. 2.1.8

[132] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017. 3.1.14

[133] Clarissa Scholes, Angela H DePace, and Álvaro Sánchez. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Systems*, 4(1):97–108, 2017. 3.2.1

[134] 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. 1.4.1, 2

[135] Robert Sedgewick. Algorithms in c, part 5: Graph algorithms, third edition, 2001. 2.2.6

[136] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, 2013. 1.3

[137] Mingfu Shao and Carl Kingsford. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, 35(12):1167–1169, 2017. 1.4.1, 3.1.7, 3.1.9

[138] Charles J Sherr. Principles of tumor suppression. *Cell*, 116(2):235–246, 2004. 2.1.10

[139] Richard Smith-Unna, Chris Boursnell, Rob Patro, Julian M Hibberd, and Steven Kelly. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26(8):1134–1144, 2016. 3

[140] Paul Smolen, Douglas A Baxter, and John H Byrne. Mathematical modeling of gene

networks. *Neuron*, 26(3):567–580, 2000. 3.2.1

[141] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018. 3.2.1, 3.2.5

[142] Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015. 1.4.3, 3.1.5

[143] Charlotte Soneson, Michael I Love, Rob Patro, Shobbir Hussain, Dheeraj Malhotra, and Mark D Robinson. A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Science Alliance*, 2(1), 2019. 3

[144] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019. 1.3

[145] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1):1–30, 2016. 3.2.5

[146] Philip J Stephens, David J McBride, Meng-Lay Lin, Ignacio Varela, Erin D Pleasance, Jared T Simpson, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Laura J Mudie, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010, 2009. 2.1.8

[147] BJ Strober, Reem Elorbany, K Rhodes, Nirmal Krishnan, Karl Tayeb, Alexis Battle, and Yoav Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019. 1.2

[148] A Sveen, S Kilpinen, A Ruusulehto, RA Lothe, and Rolf Inge Skotheim. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*, 35(19):2413–2427, 2016. 1.2, 1.4.2, 2

[149] Lucas Swanson, Gordon Robertson, Karen L Mungall, Yaron S Butterfield, Readman Chiu, Richard D Corbett, T Roderick Docking, Donna Hogge, Shaun D Jackman, Richard A Moore, et al. Barnacle: detecting and characterizing tandem duplications and fusions in transcriptome assemblies. *BMC Genomics*, 14(1):550, 2013. 2

[150] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009. 1.3

[151] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013. 2.1.10

[152] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.

Science, 310(5748):644–648, 2005. 1.2, 2

[153] Wandaliz Torres-García, Siyuan Zheng, Andrey Sivachenko, Rahulsimham Vegesna, Qianghu Wang, Rong Yao, Michael F Berger, John N Weinstein, Gad Getz, and Roel GW Verhaak. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*, 30(15): 2224–2226, 2014. **??**, 2

[154] 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, 2010. 1.2, 1.4.1

[155] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381, 2014. 1.2, 1.3

[156] Laura H Tung, Mingfu Shao, and Carl Kingsford. Quantifying the benefit offered by transcript assembly with Scallop-LR on single-molecule long reads. *Genome Biology*, 20 (1):1–18, 2019. 1.4.1

[157] Ernest Turro, Shu-Yi Su, Ângela Gonçalves, Lachlan JM Coin, Sylvia Richardson, and Alex Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, 12(2):R13, 2011. 3

[158] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2018. 3

[159] Lisa Van den Broeck, Max Gordon, Dirk Inzé, Cranos Williams, and Rosangela Sozzani. Gene regulatory network inference: connecting plant biology and mathematical modeling. *Frontiers in Genetics*, 11, 2020. 1.4.3

[160] Robin van der Lee, Solenne Correard, and Wyeth W Wasserman. Deregulated regulators: Disease-causing cis variants in transcription factor genes. *Trends in Genetics*, 2020. 3.2.1

[161] Kristoffer Vitting-Seerup and Albin Sandelin. The landscape of isoform switches in human cancers. *Molecular Cancer Research*, 15(9):1206–1220, 2017. 1.4.3

[162] Hang Wang, Becky L Sartini, Clarke F Millette, and Daniel L Kilpatrick. A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3'-end formation. *Biology of Reproduction*, 75(3):318–323, 2006. 3.2.1

[163] Jianghua Wang, Yi Cai, Wendong Yu, Chengxi Ren, David M Spencer, and Michael Ittmann. Pleiotropic biological activities of alternatively spliced TMPRSS2/ERG fusion gene transcripts. *Cancer Research*, 68(20):8516–8524, 2008. 1.4.2, 2

[164] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361 (6400), 2018. 1.3

[165] R.F. Weaver. *Molecular Biology*. College Ie Overruns. McGraw-Hill, 2012.

ISBN 9780071316866. URL https://books.google.com/books?id=yg2ucQAACAAJ. 1.2, 3.2.1

[166] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. 3, 3.2.1

[167] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938. 3.2.7

[168] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, 2015. 3.2.1

[169] Thomas D Wu and Colin K Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 2005. 2, 2.1.7, 2.1.7

[170] Xudong Wu, Ida Holst Bekker-Jensen, Jesper Christensen, Kasper Dindler Rasmussen, Simone Sidoli, Yan Qi, Yu Kong, Xi Wang, Yajuan Cui, Zhijian Xiao, et al. Tumor suppressor ASXL1 is essential for the activation of INK4B expression in response to oncogene activity and anti-proliferative signals. *Cell Research*, 25(11):1205–1218, 2015. 2.1.10

[171] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666, 2014. 1.4.1, 2

[172] Yi Xing, Tianwei Yu, Ying Nian Wu, Meenakshi Roy, Joseph Kim, and Christopher Lee. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Research*, 34(10):3150–3160, 2006. 1.4.3

[173] Yingying Xiu, Wei Liu, Tianyi Wang, Yi Liu, and Minwen Ha. Overexpression of ECT2 is a strong poor prognostic factor in ER (+) breast cancer. *Molecular and Clinical Oncology*, 10(5):497–505, 2019. 2.2.7

[174] Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016. 1.4.1, 2.1.7, 2.1.12

[175] 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. 2, 2.1.7

[176] Jin Zhang, Nicole M White, Heather K Schmidt, Robert S Fulton, Chad Tomlinson, Wesley C Warren, Richard K Wilson, and Christopher A Maher. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Research*, 26(1):108–118, 2016. 2, 2.1.8

[177] Qi Zhao, Otavia L Caballero, Samuel Levy, Brian J Stevenson, Christian Iseli, Sandro J De Souza, Pedro A Galante, Dana Busam, Margaret A Leversha, Kalyani Chadalavada,

et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proceedings of the National Academy of Sciences*, 106(6):1886–1891, 2009. 2.1.8

[178] Dinghai Zheng, Ruijia Wang, Qingbao Ding, Tianying Wang, Bingning Xie, Lu Wei, Zhaohua Zhong, and Bin Tian. Cellular stress alters 3 UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nature Communications*, 9(1):2268, 2018. 3.1.7

[179] Jiali Zhuang and Zhiping Weng. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Research*, 43(17): 8146–8156, 2015. 2

[180] Aleksey V Zimin, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, Geo Pertea, Curtis P Van Tassell, Tad S Sonstegard, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology*, 10(4):R42, 2009. 3