ABSTRACT

Title of Document: RNA-SEQUENCING ANALYSIS: READ ALIGNMENT AND DISCOVERY AND RECONSTRUCTION OF FUSION TRANSCRIPTS

Daehwan Kim, Doctor of Philosophy, 2013

Directed By: Professor Steven L. Salzberg, Department of Computer Science

RNA-sequencing technologies, which sequence the RNA molecules being transcribed in cells, allow us to explore the process of transcription in exquisite detail. One of the primary goals of RNA sequencing analysis is to reconstruct the full set of transcripts (isoforms) of genes that were present in the original cells. In addition to the transcript structures, experimenters need to estimate the expression levels for all transcripts. The first step in the analysis process is to map the RNA-seq reads against the reference genome, which provides the location from which the reads originated. In contrast to DNA sequence alignment, RNA-seq mapping algorithms have two additional challenges. First, any RNA-seq alignment program must be able to handle gapped alignment (or spliced alignment) with very large gaps due to introns, typically from 50-100,000 bases in mammalian genomes. Second, the presence of processed pseudogenes from which introns have been removed may cause many exon-spanning reads to map incorrectly.

In order to cope with these problems effectively, I have developed new alignment algorithms and implemented them in TopHat2, a second version of TopHat (one of the first spliced aligners for RNA-seq reads). The new TopHat2 program can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length insertions and deletions with respect to the reference genome. TopHat2 combines the ability to discover novel splice sites with direct mapping to known transcripts, producing more sensitive and accurate alignments, even for highly repetitive genomes or in the presence of processed pseudogenes. These new capabilities will contribute to improvements in the quality of downstream analysis.

In addition to its splice junction mapping algorithm, I have developed novel algorithms to align reads across fusion break points, which result from the breakage and re-joining of two different chromosomes, or from rearrangements within a chromosome. Based on this new fusion alignment algorithm, I have developed TransFUSE, one of the first systems for reconstruction and quantification of full-length fusion gene transcripts. TransFUSE can be run with or without known gene annotations, and it can discover novel fusion transcripts that are transcribed from known or unknown genes.

RNA SEQUENCING ANALYSIS: READ ALIGNMENT AND DISCOVERY AND RECONSTRUCTION OF FUSION TRANSCRIPTS

By

Daehwan Kim

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Steven L. Salzberg, Chair
Professor Hector Corrada-Bravo
Professor Hal Daume III
Professor Najib El-Sayed
Professor Sridhar Hannenhalli
Professor Stephen M. Mount

Acknowledgements

I would like to express my great appreciation to my thesis advisor, Steven Salzberg, for his patient guidance, constructive suggestions, and sincere support, without which I would not have been able to finish this thesis. Aside from all the knowledge and insight he shared with me in class and during our meetings, he has been a great role model to me as a researcher, a teacher, and an advisor. I try to absorb and incorporate his work ethic, efficiency, and character into becoming a good researcher.

I have had the pleasure of collaborating with such creative and enthusiastic colleagues: specifically, Geo Pertea, Cole Trapnell, Ben Langmead, Ryan Kelley, Harold Pimentel, Adam Roberts, and Lior Pachter. My special thanks go to Geo, who has been an important companion in the development of every aspect of my research. He has been working with me on several projects, co-authoring papers, and helping me prepare for presentations. I owe a huge debt to both Cole and Ben, as my research is substantially based on their original work and they continued to share their research ideas with me. I have also benefited from the discussions I had with former and current members of Dr. Salzberg's lab: Stefan Canzar, Mihaela Pertea, Liliana Florea, Arthur Delcher, Derrick Wood, Daniela Puiu, Tanja Magoc, Todd Treangen, and Li Song. Our projects on RNA-seq analysis have been among the most successful and popular in the field. I believe this is largely because of their open source nature, the active testing and reporting of bugs, and the wealth of new research ideas provided by a myriad of users, which has greatly strengthened the quality of our work.

In addition to my advisor Dr. Salzberg, I especially want to thank the other members of my defense committee: Hector Corrada-Bravo, Hal Daume III, Najib El-Sayed, Sridhar Hannenhali, and Stephen Mount, for their effort and time in carefully reviewing my thesis. I also would like to recognize the additional support from the administrative staff: Jennifer Story, Fatima Bangura, and Denise Cross. The UMIACS staff provided prompt and diligent maintenance of the computing facilities, which enabled me to easily focus on the substance of my research with few technical distractions.

I am fortunate to have friends like Sujal Bista and Jongjun Lee, who began their graduate studies at the University of Maryland the same semester I did in Fall 2008. Sujal introduced me to the world of computational biology when I had trouble deciding what research area to pursue for my Ph.D. He suggested that I take a class taught by Dr. Salzberg in Fall 2009, mentioning that this professor was one of the best in the field. After taking the class and working with Dr. Salzberg for several years as his research assistant, I can easily believe this. Jongjun is an amazing friend who was my roommate for the first two years I lived in America. Together we have shared our many moments of happiness and sadness. During this time he has also been an important source of guidance in my life. Qi Hu is in the same department with me, but we usually find each other more often in the gym than in the CS building, running and swimming together, which made my life enjoyable.

The unconditional love and support from my parents, sister, and brothers made it possible for me to get through. One of my dearest friends is my uncle Jungseo Kim, who, since I was very young, has helped me shape my life perspective by sharing his experiences with me, spending time attentively listening to me, and providing valuable feedback. I feel so fortunate to know someone like him.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background: DNA, RNA, Gene, and Protein

Deoxyribonucleic acid (DNA) encodes and serves as stable storage for the genetic programs for all forms of life. DNA is a very long molecule in which two strands intertwine with each other to form a double helix structure. Each strand consists of four nucleotides or bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Segments of both strands, called genes, are translated into proteins that are directly involved in virtually all aspects of cellular activity. The size of known genomes so far varies from 138 thousand bases (*Candidatus Tremblaya princeps*) [1] to 670 billion bases (*Amoebae dubia*) [2], where the human genome is about 3 billion bases. It is estimated that our genome contains ~21,000 protein-coding genes [3].

Figure 1.1 shows the steps necessary to decode the genetic information, genes*,* to create functioning units, proteins*,* in eukaryotic cells, including human cells. Genes are read by polymerases, which transcribe them into primary RNA transcripts (pre-mRNAs) with both the exons and introns of the genes retained. The introns from the pre-mRNAs are then removed and the remaining exons are stitched together by spliceosomes to produce mature RNA transcripts (mRNAs). During this splicing event, some of the exons can often be skipped, a process called alternative splicing, which produces different RNA transcripts and therefore adds more diversity to their final protein products. From the experiment conducted by Wang et al. [4], most human genes (92~94%) are found to be alternatively spliced. These mRNAs are subsequently exported out of the nucleus into the cytoplasm, where they are translated into proteins by ribosomes. These proteins cooperate with other proteins in a coordinated way to perform all the cell's functions. The rate of gene expression varies significantly in different types of cells. The gene expression within an individual cell will change to meet the cell's needs at any given time. Thus, a further qualitative and quantitative understanding of this fundamental activity will provide invaluable insights into many biological functions. One way we can pursue this with a high level of precision is by sequencing RNA molecules and employing computational approaches in the analysis of the RNA sequencing data.

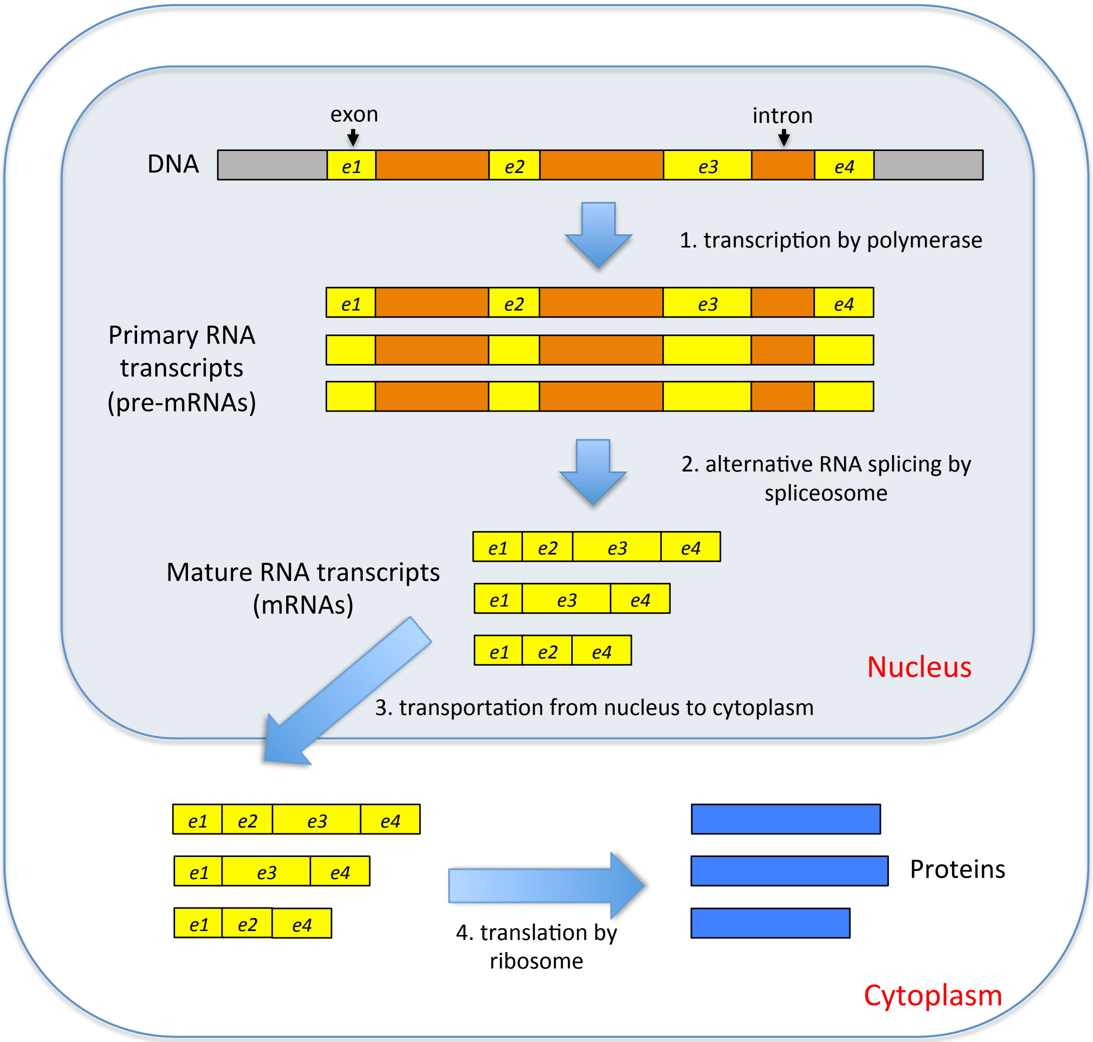

exon
intron
DNA
e1
e2
e3
e4
1. transcription by polymerase
Primary RNA transcripts (pre-mRNAs)
2. alternative RNA splicing by spliceosome
Mature RNA transcripts (mRNAs)
Nucleus
3. transportation from nucleus to cytoplasm
4. translation by ribosome
Proteins
Cytoplasm

Figure 1.1 The path from DNA to proteins in eukaryotic cells
The gene shown above includes four exons (*e1*, *e2*, *e3*, *e4*) separated by three introns. First, the gene is transcribed into primary or precursor messenger RNAs (pre-mRNAs) in which all the exons and introns are retained. Second, these pre-mRNAs are spliced into multiple mature RNA transcripts (mRNAs) in a way that all the introns are removed and some of the exons are selectively excluded. Third, these mRNAs are transferred from the nucleus to the cytoplasm, where finally ribosomes bind to them and then translate them into proteins.

One advantage of RNA-sequencing [5-8] is that, unlike microarray expression techniques [9, 10], it does not rely on pre-existing knowledge of gene content, and therefore it can detect entirely novel genes and novel splice variants of existing genes. Other applications of RNA-sequencing technologies include reconstruction and expression estimation of transcripts [11-13], differential expression analysis [14, 15], identification of transcript start site [16], discovery of fusion genes [17-19], and so on.

Figure 1.2 (upper panel) shows a simplified form of RNA sequencing process showing how mRNAs are sequenced, producing a huge number of reads in a single run: tens of millions to hundreds of millions of reads whose read lengths range from 50 to 400 base pair (bp). The simplified steps of RNA-sequencing are described as follows.

1. mRNAs are extracted from cells of interest.
2. The mRNAs are reverse-transcribed into complementary DNAs (cDNAs).
3. The cDNAs are sheared into smaller fragments, which in turn are size-selected normally from 200 to 500 bp.
4. The resulting fragments from the above step are sequenced from both ends, generating paired-end reads (or from only one end, generating single-end reads).

Many vendors provide RNA-sequencing platforms, notably, Illumina (http://www.illumina.com), Roche 454 (http://www.454.com), and Life Technologies (http://www.lifetechnologies.com).

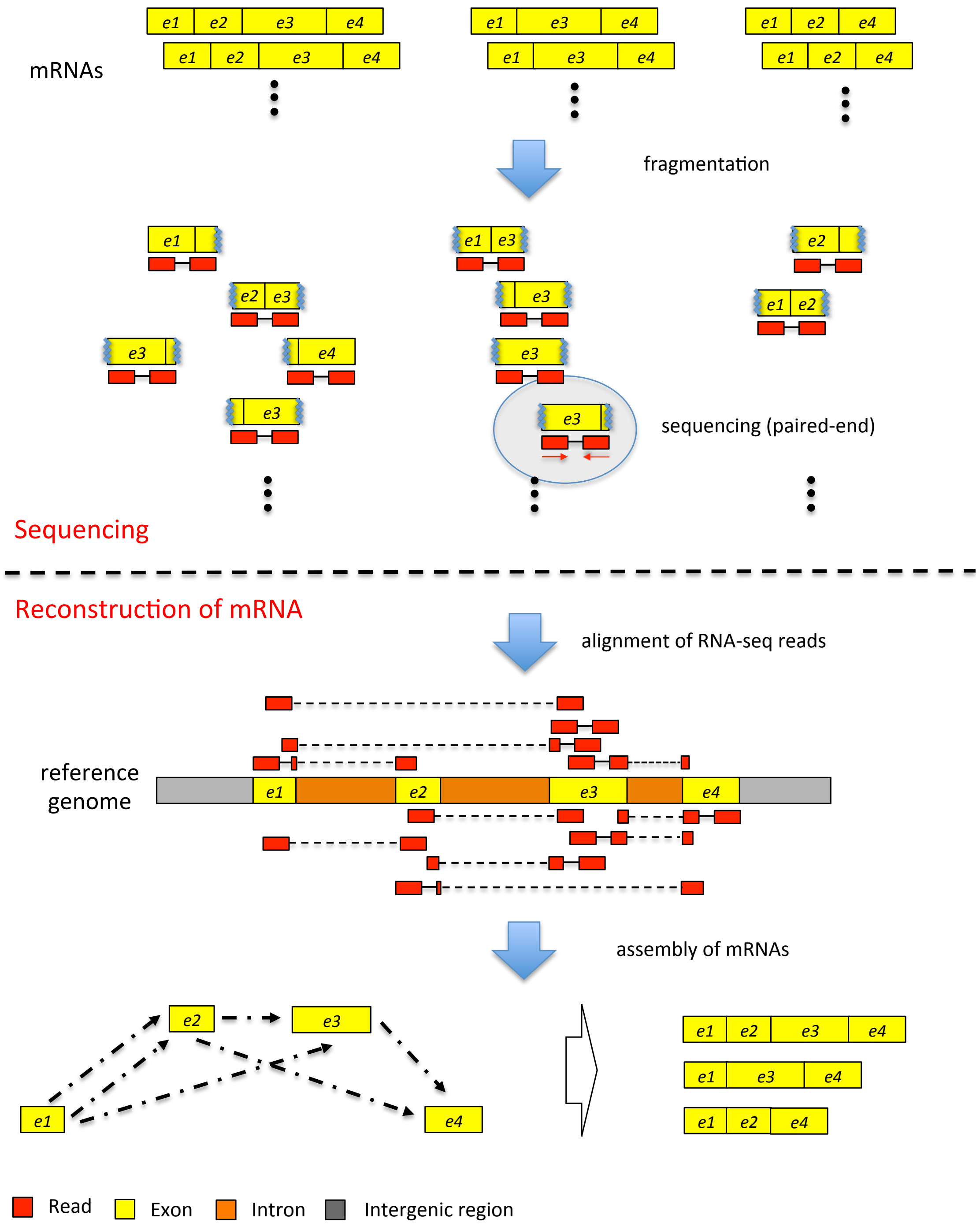


Figure 1.2 Sequencing and reconstruction of mRNAs

The top panel shows sequencing steps, most of which are biochemical processes. First, mRNAs are prepared from cells of interest. Second, these fragments are reverse-transcribed into cDNAs, fragmented, and size-selected. Third, these cDNAs are sequenced producing a huge amount of reads. The bottom panel shows computational steps to reconstruct mRNAs from relatively short RNA-seq reads. First, the reads are aligned against the reference genome to identify where they are likely to originate. Second, the mapped positions of the reads are used to assemble mRNAs. Details are given in the text.

These RNA-seq reads can be used to reconstruct the full set of mRNA transcripts (isoforms) that were present in the original cells using bioinformatics approaches as illustrated in Figure 1.2 (lower panel).

1. Reads are aligned against the reference genome, which provides information about their likely genes of origin. A set of overlapping reads in terms of their mapped positions can be clustered into groups, each group presumably representing isoforms of the same gene.
2. For each group of reads, a graph is built whose vertex v_i represents a read and whose edge (v_i, v_j) represents "compatibility" relationship between two nodes v_i and v_i. This graph can be compressed by combining vertices into "super" vertices without loss of generality, where the new vertices are equivalent to partial or full exons as shown in Figure 1.2 (lower panel).
3. A minimum set of paths covering the graph can be found using several approaches. For instance, the assembly algorithm of Cufflinks first finds mutually incompatible vertices, suggesting there are at least that many different transcripts. It can then reconstruct transcripts by finding paths that go through those vertices [20].

1.2 Spliced alignment for RNA-seq reads

Initially, RNA-seq reads are aligned against the reference genome. The results provide the location from which the reads originated. Assuming that sequencing reads are uniformly distributed along a transcript [21], we would expect 33-38% of 100-bp reads from an RNA-seq experiment to span two or more exons. Note that this proportion increases significantly from 19 to 46% as read length increases from 50 to 150 bp (see Chapter 2 for more details). Based on this observation, the alignment accuracy of those spliced reads can determine the accuracy of downstream steps of the analysis.

This mapping problem for RNA-seq reads turns out to be more challenging compared to that of DNA-seq reads, posing two additional problems. First, because genes in eukaryotic genomes contain introns and because reads sequenced from mature mRNA transcripts do not include these introns, any RNA-seq alignment program must be able to handle gapped alignment (or spliced alignment) with very large gaps. In mammalian genomes, introns span a very wide range of lengths, typically from 50-100,000 bases, which the alignment algorithm must accommodate. Second, the presence of processed pseudogenes from which some or all introns have been removed may cause many exon-spanning reads to map incorrectly. This problem is particularly acute in the case of genomes like the human genome, which contains over 14,000 pseudogenes [22].

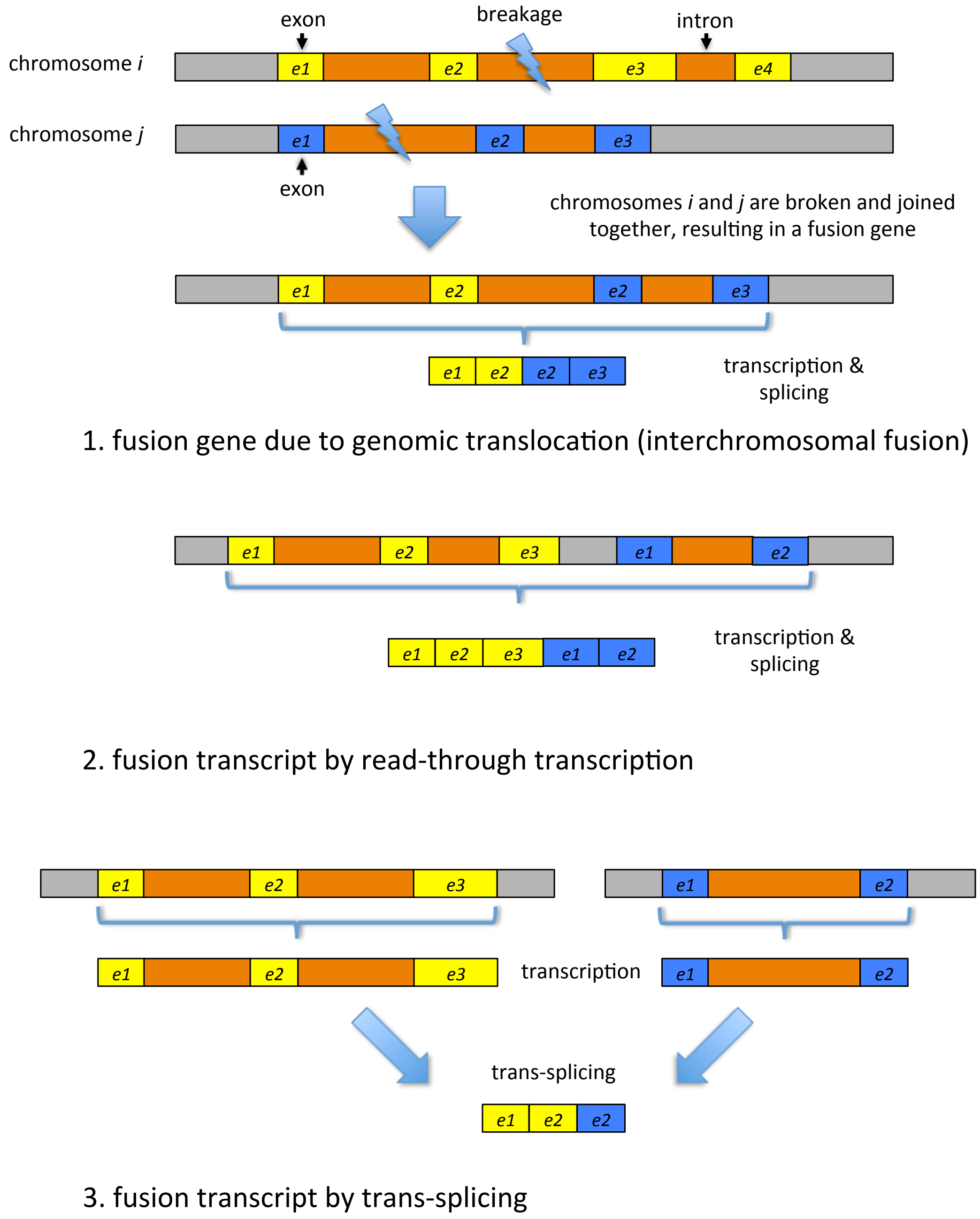
In Chapter 2, we will discuss TopHat2, a new spliced aligner, in an attempt to handle these problems. TopHat2 employs a two-step procedure similar to that of TopHat [23]. First, it detects potential splice sites for introns, but with a much higher precision compared to the original algorithm of TopHat. It then uses these candidate splice sites to align multi-exon spanning reads properly in a subsequent step. In these steps, TopHat2 uses Bowtie as its underlying alignment program. I implemented new procedures that align reads with true insertions and deletions (indels) – a feature critical for studies assessing the impact of genetic mutations on gene and transcript expression. Indels due to sequencing errors will be discovered by TopHat2's underlying mapping engine, Bowtie2 [24], which can detect short indels very efficiently. The new algorithm also makes powerful use of available gene annotations, which prevents it from erroneously mapping reads to pseudogenes and improves its overall alignment accuracy. Annotation also allows TopHat2 to better align reads that cover microexons, noncanonical splice sites, and other unusual features of eukaryotic transcriptomes. These new enhancements will provide major accuracy improvements over previous versions and other RNA-seq mapping tools. Chapter 2 is based on the following work, which is under review.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg

TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions, and gene fusions. *To appear in Genome Biology*

1.3 Discovery of fusion break points using RNA-seq reads

In addition to detection of novel genes, RNA-seq has the potential to discover genes created by complex chromosomal rearrangements. As illustrated in Figure 1.3, “fusion” genes formed by the breakage and re-joining of two different chromosomes have repeatedly been implicated in the development of cancer, notably the BCR/ABL1 gene fusion in chronic myeloid leukemia [17, 25, 26]. Fusion genes can also be created by the breakage and rearrangement of a single chromosome, bringing together transcribed sequences that are normally separate. As of November 2012, the Mitelman database [27] has documented about 62,000 cases of chromosome aberrations and gene fusions in cancer. 1,078 gene fusions have been reported from 1,309 different genes [28]. Most fusion genes are strongly associated with distinct cancerous tumor types, whereas some others are reported even in benign tumor cells or normal cells. As well as from genomic aberrations described above, fusion events can take place during the transcription process in which two adjacent genes are transcribed as a single pre-RNA molecule, and then spliced into a fusion mRNA. Akiva et al. [29] performed a bioinformatics approach using expressed sequence tags (ESTs) and cDNAs downloaded from GenBank [30], showing that about 2% of the human genes are associated with such read-through transcription. Fusion transcripts can also be formed post-transcriptionally when two different pre-RNA transcripts from two genes are spliced together and combined into one single mRNA transcript [31] (see Figure 1.3). This process is called trans-splicing.



1. fusion gene due to genomic translocation (interchromosomal fusion)

2. fusion transcript by read-through transcription

3. fusion transcript by trans-splicing

Figure 1.3 Several pathways leading to the formation of fusion transcripts

Discovering these fusions via RNA-seq has a distinct advantage over whole-genome sequencing. This is due to the fact that in the highly rearranged genomes of some tumor samples, many rearrangements might be present, although only a fraction might alter transcription. RNA-seq identifies only those chromosomal fusion events that produce transcripts. It has the further advantage that it allows one to detect multiple alternative splice variants that might be produced by a fusion event.

In Chapter 3, we will describe a fusion detection algorithm, TopHat-Fusion [19]. TopHat-Fusion directly detects individual reads and paired reads that span a fusion event. Because it does not rely on annotation, TopHat-Fusion can also find events involving novel splice variants and entirely novel genes. TopHat-Fusion's performance was evaluated using RNA-seq reads from four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7). Edgren et al. [18] initially reported 24 novel and 3 known fusion genes in this data sample. Using TopHat-Fusion, 25 of the 27 fusion genes were retrieved, in addition to 51 strong candidates for novel fusion genes. Approximately one year later, Kangaspeska et al. [32] (including Edgren) experimentally verified 9 of those 51 candidates to be genuine fusion genes.

Fusion-finding software currently faces serious problems, including very high false positive rates. FusionSeq [33] and deFuse [34], found 32,644 and 1,670, respectively, for MCF7 cell lines, which harbor three known fusion genes. Almost all of FusionSeq and deFuse's findings are expected to be false positives. In contrast to other fusion-finding software, TopHat-Fusion demonstrates highly accurate and sensitive discovery of fusion transcripts, having reported 3 known fusion genes and only 8 strong candidates.

The following paper in *Genome Biology* is the basis for Chapter 3.

Daehwan Kim and Steven L. Salzberg

TopHat-Fusion: An Algorithm for Discovery of Novel Fusion Transcripts. *Genome biology* 2011, **12:**R72.

1.4 Reconstruction and quantitation of fusion transcripts

As described in the above section, TopHat-Fusion aligns reads across fusion break points and reports the fusion alignments in Sequence Alignment/Map (SAM) format [35]. SAM has rapidly become the most popular format for representing read alignments. Read alignments in the SAM format can be used for reconstruction and quantification of fusion transcripts as well as normal transcripts. More specifically, based on the read alignments from TopHat-Fusion, those that are aligned near or across the break points can be assembled into fusion transcripts, and the expression levels of the transcripts can be quantified based on the number of reads they include.

Two major factors make this problem more difficult: first, eukaryotic genomes are highly repetitive [36, 37], meaning the reads can align to many locations and second, sequencing errors (e.g., random ligation of two cDNAs) may cause chimeric transcripts. The problem of separating genuine fusion transcripts from these spurious fusion-like transcripts, which are much more numerous than true fusions, is a major algorithmic challenge. The problem is made harder by the fact that reads are non-uniformly distributed across transcripts, making low-level transcripts difficult to detect. A sensitive and accurate method for identifying fusions should find as much evidence as possible that can be used as either positive or negative indicators when filtering out potential fusion transcripts.

I have developed TransFUSE to address these problems. TransFUSE is the software system designed to reconstruct and quantify full-length fusion gene transcripts. The newly developed algorithm, using TopHat2 and Cufflinks, can be run with or without gene annotations. As a result, it can detect novel fusion transcripts from known and unknown genes. In Chapter 4, we will discuss more details about TransFUSE. Chapter 4 is based on the following work, which is in preparation for submission.

Daehwan Kim and Steven L. Salzberg
Reconstruction and Estimation of Fusion Transcripts from RNA-Sequencing reads.
In preparation

1.5 Summary

RNA-seq technologies deliver a large amount of data within a short period of time (a few days) at much lower costs. These benefits allow us to quickly and accurately investigate genetic programs and cellular activity. Using these new sequencing technologies, we can examine transcript structures, expression levels of transcripts, and structural variations. However, the sequencing technologies require new computational methods in order to effectively use a large amount of RNA-seq reads they produce.

Most RNA-seq analyses rely on the genomic locations of reads' origins. In order to find the location information, reads may be aligned against the reference genome, if a high-quality reference genome is available and the differences between the reference and the sequenced genome is small, as is the case with the human genome. Mapping accuracy and sensitivity of the alignment determine the quality of the downstream analyses. A significant portion of this thesis is devoted to discussing effective solutions to RNA-seq read alignment problems. As a solution, I have developed several novel algorithms and incorporated them into TopHat2, an RNA-seq alignment system (Chapter 2). In particular, I designed a new algorithm, in collaboration with Cole Trapnell, called “segment-search”, for identifying splice sites with high level of precision and sensitivity (see Chapter 2 for more details). I then implemented this method in the system. I also designed a new algorithm that uses gene annotations to guide transcriptome mapping, a feature that was not part of TopHat before. I collaborated with Harold Pimentel and Geo Pertea on this design, and they also collaborated on the implementation. One problem I discovered in TopHat was that many reads may incorrectly map to processed pseudogenes during the first alignment stage (end-to-end genome alignment). In order to fix this mapping bias, I have came up with the idea of re-aligning many reads in a subsequent step. I added this “realignment” option to TopHat2’s spliced alignment stage, and as a result most of the reads are now correctly aligned. TopHat2 also includes a novel indel alignment algorithm, developed by Ryan Kelley and myself. I made further adjustments to TopHat2 in order to support greater read lengths and to support “colorspace” reads from ABI SOLiD. The colorspace method required substantial changes in many parts of TopHat2. To improve TopHat2's performance, I have parallelized most of the steps in its pipeline. Geo Pertea and I have changed TopHat2 to use compressed files for intermediate files to considerably reduce the disk requirement. All these changes significantly improve the performance of TopHat2 and contribute to the success of TopHat2 in the research community.

RNA-seq also enables us to discover structural variations, including genomic rearrangements. I have developed TopHat-Fusion to detect fusion break points and map reads against them (Chapter 3). In TopHat-Fusion, I have developed novel algorithms for finding fusion break points and align reads across the break points. I have also created sophisticated filtering algorithms to eliminate false fusion transcripts. This filtration is based on several factors such as supporting reads and pairs, sequence similarity, and transcript coverage. TopHat-Fusion's advanced fusion alignment algorithm combined with the filtration step enables efficient and sensitive discovery of fusion transcripts. Furthermore, I have enhanced and modified TopHat-Fusion to allow the assembly and quantification of fusion transcripts. The enhanced pipeline is TransFUSE (Chapter 4). The new assembly and quantification algorithms of TransFUSE are based on Cufflinks, where I have modified almost every aspect of the system in addition to creating a novel algorithm to filter out false fusion transcripts. The new information available from these new algorithms such as transcript structures and expression levels is used to identify genuine fusion transcripts. I have developed a visualization algorithm to display fusion transcripts in html format, which makes it easier to investigate and directly compare fusion transcripts.

However, our software system faces several issues as the sequencing technologies provide longer and more numerous reads. In Chapter 5, we will discuss several related problems that demand some fundamental changes in our pipeline, as well as a proposal for a new pipeline to address these challenges.

Chapter 2: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Since the initial release of TopHat [23], a spliced aligner for sequences from transcriptome sequencing (RNA-seq) experiments, I have made many significant enhancements to the program, contributing to improvements in the quality of downstream analysis. The new TopHat2 program can align reads of various lengths produced by the latest sequencing technologies, including Illumina, 454 pyrosequencing, and ABI "colorspace" reads, while allowing for variable-length insertions and deletions with respect to the reference genome. In addition to its *de novo* splice junction mapping algorithm, TopHat2 incorporates an algorithm to align reads across fusion break points, which occur after genomic translocations or trans-splicing (see Chapter 3 for more details). The new system combines the ability to discover novel splice sites with direct mapping to known transcripts, producing more sensitive and accurate alignments than previously, even for highly repetitive genomes or in the presence of processed pseudogenes. A new re-alignment procedure substantially reduces mis-alignments caused by reads that extend only a few bases into an intronic region. Finally, in order to keep up with dramatically increasing sequencing rates, the TopHat2 algorithm includes new parallel code and other optimized routines that reduce its run time significantly compared to its predecessor.

TopHat2 is free, open-source software available from http://genomics.jhu.edu/software/tophat.

2.1 Background

RNA sequencing technologies [5-7], which sequence the RNA molecules being transcribed in cells, allow us to explore the process of transcription in exquisite detail. One of the primary goals of RNA sequencing analysis software is to reconstruct the full set of transcripts (isoforms) of genes that were present in the original cells. In addition to the transcript structures, experimenters need to estimate the expression levels for all transcripts. The first step in the analysis process is to map the RNA-seq reads against the reference genome, which provides the location from which the reads originated. In contrast to DNA sequence alignment, RNA-seq mapping algorithms have two additional challenges. First, because genes in eukaryotic genomes contain introns and because reads sequenced from mature mRNA transcripts do not include these introns, any RNA-seq alignment program must be able to handle gapped alignment (or spliced alignment) with very large gaps. In mammalian genomes, introns span a very wide range of lengths, typically from 50-100,000 bases, which the alignment algorithm must accommodate. Second, the presence of processed pseudogenes from which some or all introns have been removed may cause many exon-spanning reads to map incorrectly. This probably is particularly acute for the human genome, which contains over 14,000 pseudogenes [22].

In the most recent Ensembl GRCh37 gene annotations, the average length of a mature mRNA transcript in the human genome is 2,227 bp, and the average exon length is 235 bp. The average number of exons per transcript is 9.5. Assuming that sequencing reads are uniformly distributed along a transcript [21], we would expect 33-38% of 100-bp reads from an RNA-seq experiment to span two or more exons. Note that this proportion increases significantly as read length increases from 50 to 150 bp (see Supplementary Material for more details).

More importantly for the alignment problem, ~20% of junction-spanning reads extend 10 bp or less into one of the exons they span. These small "anchors" make it extremely difficult for alignment software to map reads accurately, particularly if the algorithm relies (as most do) on an initial mapping of fixed-length k-mers to the genome. This initial mapping, using exact matches of k-mers, is critical for narrowing down the search space into small local regions where a read is likely to align. If a read only extends a few bases into one of two adjacent exons, then it often happens that the read will align equally well, but incorrectly, with the sequence of the intervening intron. For example, as illustrated in Figure 2.1, suppose that read *r* spans exons e_1 and e_2, extending only 4 bases into e_2. Suppose also that that e_2 begins with GTXX, and the intervening intron also begins with GTXX. Then *r* might align perfectly to e_1 and the first 4 bases of the intron, and the alignment algorithm will fail to find the spliced alignment of *r*.

In order to handle this problem, TopHat2 uses a two-step procedure. First, similar to TopHat1 [23], it detects potential splice sites for introns (detailed further in Methods). It then uses these candidate splice sites to align multi-exon spanning reads properly in a subsequent step. Some RNA-seq aligners, including GSNAP [38], RUM [39], and STAR [40], map reads independently of the alignments of other reads, which may explain their lower sensitivity for these spliced reads (see Results). MapSplice [41] uses a two-step approach similar to TopHat2.

RNA-seq read alignment is further complicated due to the presence of processed pseudogenes in the reference genome. Pseudogenes often have highly similar sequences to functional, intron-containing genes, and in most cases the pseudogene versions are not transcribed [42], though this has recently been disputed [43]. The critical problem for alignment is that reads spanning multiple exons can be mapped perfectly or near-perfectly to the pseudogene version of a functional gene. For example, suppose a read *r* spans two exons of a given gene. If the aligner tries to align the read globally (end-to-end), then it will find an alignment to the pseudogene copy (Figure 2.1). If the spliced alignment phase, which usually occurs later, does not attempt to re-align *r*, then the pseudogene copy will "absorb" all reads spanning splice sites for that gene. TopHat2 can feed *r* into the spliced alignment phase even when *r* has been aligned end-to-end, allowing it to circumvent this problem (see Results and Methods).

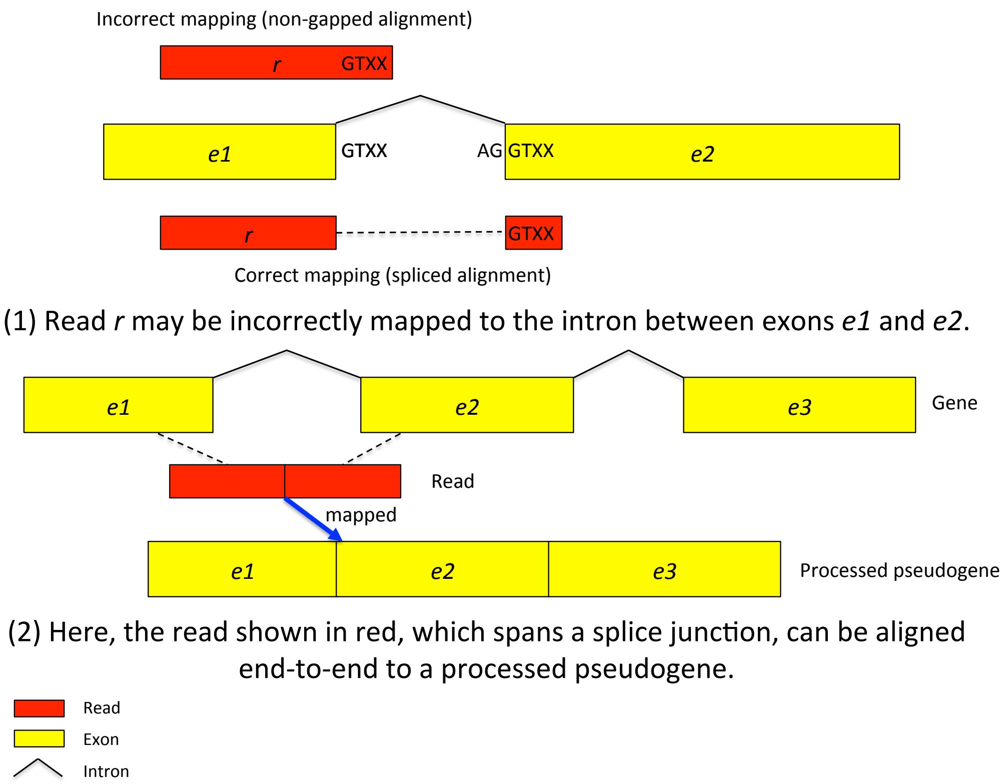

Incorrect mapping (non-gapped alignment)
r
GTXX
e1
GTXX
AG
GTXX
e2
r
GTXX
Correct mapping (spliced alignment)
(1) Read r may be incorrectly mapped to the intron between exons e1 and e2.
e1
e2
e3
Gene
Read
mapped
e1
e2
e3
Processed pseudogene
(2) Here, the read shown in red, which spans a splice junction, can be aligned end-to-end to a processed pseudogene.
Read
Exon
Intron

Figure 2.1 Two possible incorrect alignments of spliced reads
(1) A read extending a few bases into the flanking exon can be aligned to the intron instead of the exon. (2) A read spanning multiple exons from genes with processed pseudogene copies can be aligned to the pseudogene copies instead of the gene from which it originates.

We also note as an aside that, in our analysis of RNA-seq reads from multiple human samples [44, 45], genes with processed pseudogenes seem to be expressed at higher levels than other genes (see Results). Although this observation has not been explored thoroughly, a plausible explanation is that genes with higher levels of expression may, over the course of evolution, have an increased chance of being picked up by transposons and re-integrated into the genome, creating pseudogene copies.

For the human genome, where we have relatively comprehensive annotations of protein-coding genes, we can use the annotations to map reads more accurately, by aligning reads preferentially to real genes rather than pseudogenes. GSNAP [38] and STAR [40] also make use of annotation, although they use it in a more limited fashion to detect splice sites. TopHat2 can use the full-length transcripts defined by annotations during its initial mapping phase, which produces significant gains in sensitivity and accuracy (see Figures 2.3 - 2.6).

Transcripts from a target genome may differ substantially from the reference genome, possibly containing insertions, deletions, and other structural variations [46, 47]. For such regions, previous spliced alignment programs (including the original TopHat) sometimes fail to find a proper alignment. In TopHat2, I implemented new procedures that align reads with true insertions and deletions (indels). Indels due to sequencing errors will be discovered by TopHat2's underlying mapping engine, Bowtie2 [24], which can detect short indels very efficiently. Very large deletions, inversions on the same chromosome, and translocations involving different chromosomes are detected by the TopHat-Fusion algorithms [19], which are now incorporated into TopHat2 and available by a simple command-line switch.

TopHat2 also includes new algorithms to handle more diverse types of sequencing data. This includes the ability to handle reads generated by ABI SOLiD technology using its "color space" representation. To accomplish this, TopHat2 uses a reference genome translated entirely into color space in order to take advantage of the error-correction capability of that format. TopHat2 also handles data sets in which the reads have variable lengths, allowing the experimenter to merge data sets from multiple sequencing runs with different lengths.

2.2 Methods

Given RNA-seq reads as input, TopHat2 begins by mapping reads against the known transcriptome, if an annotation file is provided. This transcriptome mapping improves overall mapping sensitivity and accuracy. It also gives a significant speed boost, owing to the much smaller size of the transcriptome compared to that of the genome (see Figure 2.2).

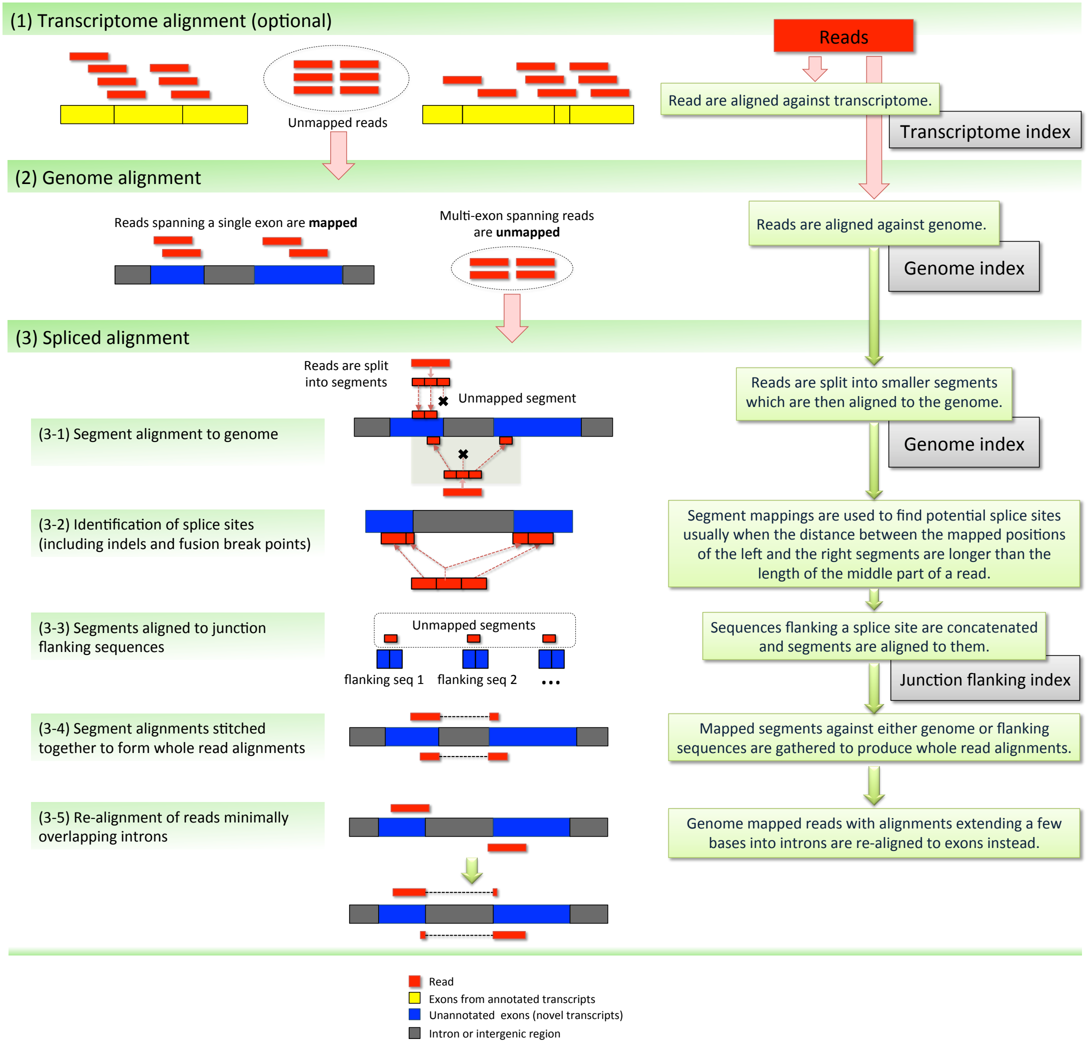

(1) Transcriptome alignment (optional)
Reads
Unmapped reads
Read are aligned against transcriptome.
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are **mapped**
Multi-exon spanning reads are **unmapped**
Reads are aligned against genome.
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
Reads are split into smaller segments which are then aligned to the genome.
Genome index
(3-2) Identification of splice sites (including indels and fusion break points)
Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
…
Sequences flanking a splice site are concatenated and segments are aligned to them.
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.
(3-5) Re-alignment of reads minimally overlapping introns
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.
🟥 Read
🟨 Exons from annotated transcripts
🟦 Unannotated exons (novel transcripts)
⬛ Intron or intergenic region

Figure 2.2 TopHat2 pipeline
Details are given in the main text.

After the transcriptome mapping step, some reads remain unmapped because they are derived from unknown transcripts not present in the annotation, or because they contain many mis-called bases. In addition, there may be poorly aligned reads that have been mapped to the wrong location. In step 2, TopHat2 aligns these unmapped or potentially mis-aligned reads against the genome (Figure 2.2). Any reads contained entirely within exons will be mapped, whereas others spanning introns may not be.

Using unmapped reads from step 2, TopHat2 tries to find novel splice sites that are based on known junction signals (GT-AG, GC-AG, and AT-AC). TopHat2 also provides an option to allow users to remap some of the mapped reads depending on their edit distance values, that is, those reads whose edit distance is greater than or equal to a user-provided threshold will be treated as unmapped reads. To accomplish this, the unmapped reads (and previously mapped reads with low alignment scores) are split into smaller non-overlapping segments (25-bp each by default) which are then aligned against the genome (Figure 2.2, step 3). Tophat2 examines the cases where the left and right segments of the same read are mapped within a user-defined maximum intron size (usually between 50 and 100,000 bp). When this pattern is detected, TopHat2 re-aligns the whole read sequence to that genomic region in order to identify the most likely locations of the splice sites, as shown in Figure 2.2. Indels and fusion break points are also detected in this step using a similar approach.

The genomic sequences flanking these splice sites are concatenated and the resulting spliced sequences are collected as a set of potential transcript fragments.

Any reads not mapped in the previous stages (or mapped very poorly) are then re-aligned with Bowtie2 [24] against this novel transcriptome.

After these steps, some of the reads may have been aligned incorrectly by extending an exonic alignment a few bases into the adjacent intron (see Figure 2.1 and Figure 2.2, step 3-5). TopHat2 checks if such alignments extend into the introns identified in the split alignment phase, and if so, it can re-align these reads to the adjacent exons instead.

In the final stage, TopHat2 divides reads into those with unique alignments and those with multiple alignments. For the multi-mapped reads, TopHat2 gathers statistical information (e.g., the number of supporting reads) about the relevant splice junctions, insertions, and deletions, which it uses to recalculate the alignment score for each read. Based on these new alignment scores, TopHat2 reports the most likely alignment locations for such multi-mapped reads.

For paired-end reads, TopHat2 processes the two reads separately through the same mapping stages described above. In the final stage, the independently aligned reads are analyzed together to produce paired alignments, taking into consideration additional factors including fragment length and orientation.

For the experiments described in this study, the program version numbers were TopHat2 (2.0.8), TopHat1 (1.1.4), GSNAP (2013-01-23), RUM (1.12_01), MapSplice (1.15.2), and STAR (2.3.0e). Specific parameters for each program are given in Table 2.16.

2.3 Results

TopHat2 can use either Bowtie [48] or Bowtie2 [24] as its core read alignment engine. TopHat2 has its own indel-finding algorithm, which enhances Bowtie2's indel-finding ability in the context of spliced alignments. In order to evaluate TopHat2 and compare it other methods, we ran multiple computational experiments using both real and simulated RNA-seq data.

For the simulations, we created multiple sets of 40,000,000 paired-end reads, 100 bp in length, from the entire human genome (release GRCh37). Instead of trying to precisely mimic real RNA-seq experiments, which may not be possible in any practical sense, we generated data with relatively simple settings and expression levels calculated using a model from the Flux Simulator system [49], as follows. For the first test set, we generated reads from the known transcripts on the entire human genome without introducing any mismatches or indels. We then generated additional data sets where we included (a) insertions and deletions into the known transcripts at random locations; and (b) insertions and deletions in the reads themselves to mimic sequencing errors (see Supplementary Material for details).

Each of these types of experimental error was introduced to test different capabilities of TopHat2 and other RNA-seq aligners. Following the simulations, we evaluated the programs using a recent, real RNA-seq data set.

Alignment of simulated reads (error-free)

We generated 40,000,000 paired-end reads and performed two sets of experiments: (1) using 20,000,000 "left" reads from the paired-end data set, shown in

Table 2.1; and (2) using 20,000,000 pairs of reads, shown in Table 2.2. Reads that span multiple exons are called *junction* reads; our single-end data contain 6,862,278 such reads (34.3%). The most challenging alignments are those for which a junction read extends 10 bp or less into one of the exons, which we call *short-anchored* reads; 1,448,022 of the single-end reads (7.2%) fell into this category. We report accuracy separately for junction reads and short-anchored reads in Tables 2.1-2.2.

Program	No. of mapped reads	Correctly mapped reads (%)	Incorrectly mapped reads (%)	Unmapped reads (%)	Correct junction reads (%)	Correct short-anchored reads (%)
TopHat2 +Bowtie1	19,826,638	98.31	0.82	0.87	95.28	93.69
TopHat2 +Bowtie2	19,826,673	98.03	1.10	0.87	94.28	89.67
TopHat1.14	19,616,874	94.64	3.45	1.91	84.44	44.08
GSNAP	19,997,255	94.21	5.77	0.02	83.15	26.01
RUM	19,555,823	88.11	9.67	2.22	65.35	8.59
MapSplice	19,872,372	97.28	2.08	0.64	92.09	75.57
STAR	19,087,508	92.14	3.30	4.56	77.17	3.54

Table 2.1 Performance comparisons on 20 million 100 bp single-end reads
These reads are simulated based on transcripts from the entire human genome. 6,862,278 reads span one or more splice junctions; the alignment accuracy of junction reads refers to this set. 1,448,022 reads extend 10 bp or less into one exon; the alignment accuracy of short-anchored reads is based on these alignments. The last two columns show alignment accuracy for these subsets of the data.

We also tested 20,000,000 read pairs (40,000,000 reads), of which 9,491,394 (47.5%) have at least one read that spans multiple exons. 2,702,624 of these pairs (13.5%) have at least one short-anchored read that extends 10 bp or less into one of its exons. Table 2.2 shows the results of mapping these reads with TopHat2 and other programs.

Program	No. of mapped pairs	Correctly mapped pairs (%)	Incorrectly mapped pairs (%)	Unmapped pairs (%)	Correct junction pairs (%)	Correct short-anchored pairs (%)
TopHat2 +Bowtie1	19,683,426	96.70	1.72	1.58	93.31	90.09

TopHat2 +Bowtie2	19,686,006	96.19	2.24	1.57	92.03	85.88
TopHat1.14	19,219,055	89.57	6.53	3.90	78.36	40.39
GSNAP	19,999,867	88.84	11.16	0.00	76.55	22.87
RUM	19,869,579	79.07	20.28	0.65	56.28	8.42
MapSplice	19,342,087	92.03	4.68	3.29	86.53	72.48
STAR	19,951,620	85.21	14.55	0.24	68.94	3.16

Table 2.2 Performance comparisons on 20 million pairs of 100 bp reads
These paired reads are simulated based on transcripts from the entire human genome. 9,491,394 pairs of reads are junction pairs, and 2,702,624 pairs contain short-anchored reads. The last two columns show alignment accuracy for these subsets of the data.

As shown in Table 2.1, TopHat2 correctly aligns >98% of the reads, more than any of the other methods, whose accuracy ranged from 88–97%. The difference is more pronounced for junction reads, where TopHat2 is able to align >94% while other methods range in accuracy from 65–92%.

GSNAP, RUM, and STAR have particular difficulty aligning short-anchored reads, only aligning 26%, 8.6%, and 3.5%, respectively. MapSplice does considerably better, aligning 75.6% of these reads. By contrast, TopHat2 aligns 93.7% of the short-anchored reads using Bowtie1 as its main aligner (Table 2.1). Both TopHat2 and MapSplice use a two-step algorithm, first detecting potential splice sites, and then using these sites to map reads. This two-step method may explain their superior performance at mapping reads with short anchors.

The results for paired reads (Table 2.2) are similar to those for unpaired reads. TopHat2 aligns the highest percentage of reads, 96.7%, followed by MapSplice (92%) and the other methods (79-88%). The difference widens again for junction reads, with TopHat2 at 93% followed by MapSplice (86%), GSNAP (76%), STAR (69%), and RUM (56%). Most striking of all was the performance on short-anchored reads, which most of the methods had great difficulty aligning correctly. TopHat2 aligned 90% of these, MapSplice aligned 72%, and the other methods aligned only 3–22%.

Figure 2.9 shows alignment rates for reads, spliced reads, and spliced reads with small anchors for a variety of read lengths (50 bp, 100 bp, 150 bp, 200 bp). TopHat2 consistently outperformed all the other aligners for each read length. In Tables 2.6 – 2.9, we compare alignment performance for spliced reads and pairs with a 1-3 mismatches, where TopHat2 and MapSplice show the highest recall rates.

Alignment of simulated reads with short indels (1-3bp)

Next we tested the spliced alignment programs using reads with small indels, using two sets of simulated reads: (1) *true indels*, in which the transcripts were modified by inserting or deleting 1-3 bases at random locations; and (2) *indels due to sequencing errors*, in which indels are randomly inserted into the reads. As before, all transcripts were simulated from known genes from the entire human genome. We used a relatively high rate of indels intentionally, to test the mapping capabilities of the programs in the presence of these types of mutations.

Tables 2.3-2.4 shows the results for these data sets. For single-end reads, RUM, GSNAP, and TopHat2 perform similarly, with 69-82% accuracy (recall) rates for true indels and 62-83% for reads with indel sequencing errors. STAR and MapSplice show relatively lower recall rates for both data sets. Note that when used with the original Bowtie program (a non-gapped aligner), TopHat2 is able to map "true" indel reads using its own indel finding algorithms.

Program	Reads with true indels (1,428,499)		Reads with sequencing-error indels (1,525,657)	
	Accuracy (%)	Accuracy on 351,465 reads with boundary indels (%)	Accuracy (%)	Accuracy on 357,334 reads with boundary indels (%)
TopHat2 +Bowtie1	70.9	16.8	12.1	2.8
TopHat2 +Bowtie2	63.7	25.2	62.6	21.2
GSNAP	82.7	71.9	83.1	71.8
RUM	69.4	43.0	70.3	45.4
MapSplice	27.3	3.7	27.5	3.8
STAR	46.6	16.9	47.7	17.1

Table 2.3 Performance comparisons on single-end reads containing indels
The indels are 1-3bp. The number of reads containing each type of error is indicated in the column header. Boundary indels occur within 25 bp of an exon boundary. Percentages refer only to the reads of each type, not to the entire data set.

Program	Pairs with true indels (2,754,313)		Pairs with sequencing-error indels (2,934,043)	
	Accuracy (%)	Accuracy on 685,937 pairs with boundary indels (%)	Accuracy (%)	Accuracy on 695,771 pairs with boundary indels (%)
TopHat2 +Bowtie1	69.8	16.3	14.0	3.1
TopHat2 +Bowtie2	62.3	24.0	60.8	19.8
GSNAP	77.0	63.8	77.8	64.8
RUM	60.3	34.3	61.3	36.0
MapSplice	25.5	3.4	25.0	3.2
STAR	53.4	19.2	54.9	21.4

Table 2.4 Performance comparisons on paired reads containing indels
The indels are 1-3bp. The number of pairs containing each type of error is indicated in the column header. Boundary indels occur within 25 bp of an exon boundary. Percentages refer only to the pairs of each type, not to the entire data set.

For paired-end reads with indels, GSNAP has the highest rate of correct alignments (77%), followed by TopHat2 (60-69%), RUM (60-61%), and STAR (53-54%). MapSplice shows the lowest accuracy for both single-end and paired-end reads.

We defined *boundary indels* as those within 25 bp of a splice site. We separately computed the accuracy on reads with boundary indels, shown in Tables 2.3-2.4.

Alignment of a large set of real RNA-seq reads

Any test of alignment algorithms should use real data to provide a measure of likely performance in practice. For these experiments, we used a recently released set of RNA-seq reads gathered across a time course experiment reported by Chen et al. [44] (GEO accession number GSM818582). This data includes 130,705,578 million paired-end reads in 65,352,789 pairs. All reads are 101 bp in length.

Because we do not know the true alignments for this RNA-seq data set, we used the following objective criteria to evaluate each program:

1. The cumulative number of alignments with edit distances of 0, 1, 2, and 3 for each read.
2. The cumulative number of spliced alignments that agree with the annotation for the corresponding human genes, taken from the Ensembl GRCh37 release of the human genome.

For each program, we aligned the paired-end reads with and without the known gene annotations, where possible. RUM requires annotations and cannot be run without them, while MapSplice maps strictly without them. We then evaluated the mapping results in terms of the number of read or paired-read mappings.

TopHat2 consists of three mapping steps: (1) transcriptome mapping, used only when annotation is provided; (2) genome mapping; and (3) spliced mapping (see

Methods for details). TopHat2 uses a remapping edit distance threshold *t*, specified by the user, as follows. If a read aligns to the transcriptome in step (1) with an edit distance less than *t*, TopHat2 will not remap the read in subsequent steps. Otherwise, TopHat2 will try to re-align the read in steps (2) and (3), and then depending on the resulting edit distance, it will use the read to detect novel splice sites. A setting of *t*=0 means that TopHat2 will re-align every read in all three steps. When we used *t*=0 ("TopHat2 realignment 0" in Figure 2.3) on the real data, we consistently obtained better mapping results in terms of edit distance and the number of alignments that correspond to known splice sites, as shown in Figures 2.3-2.6 for read and pair alignments, respectively (see also Tables 10-13 for the actual numbers of the alignments).

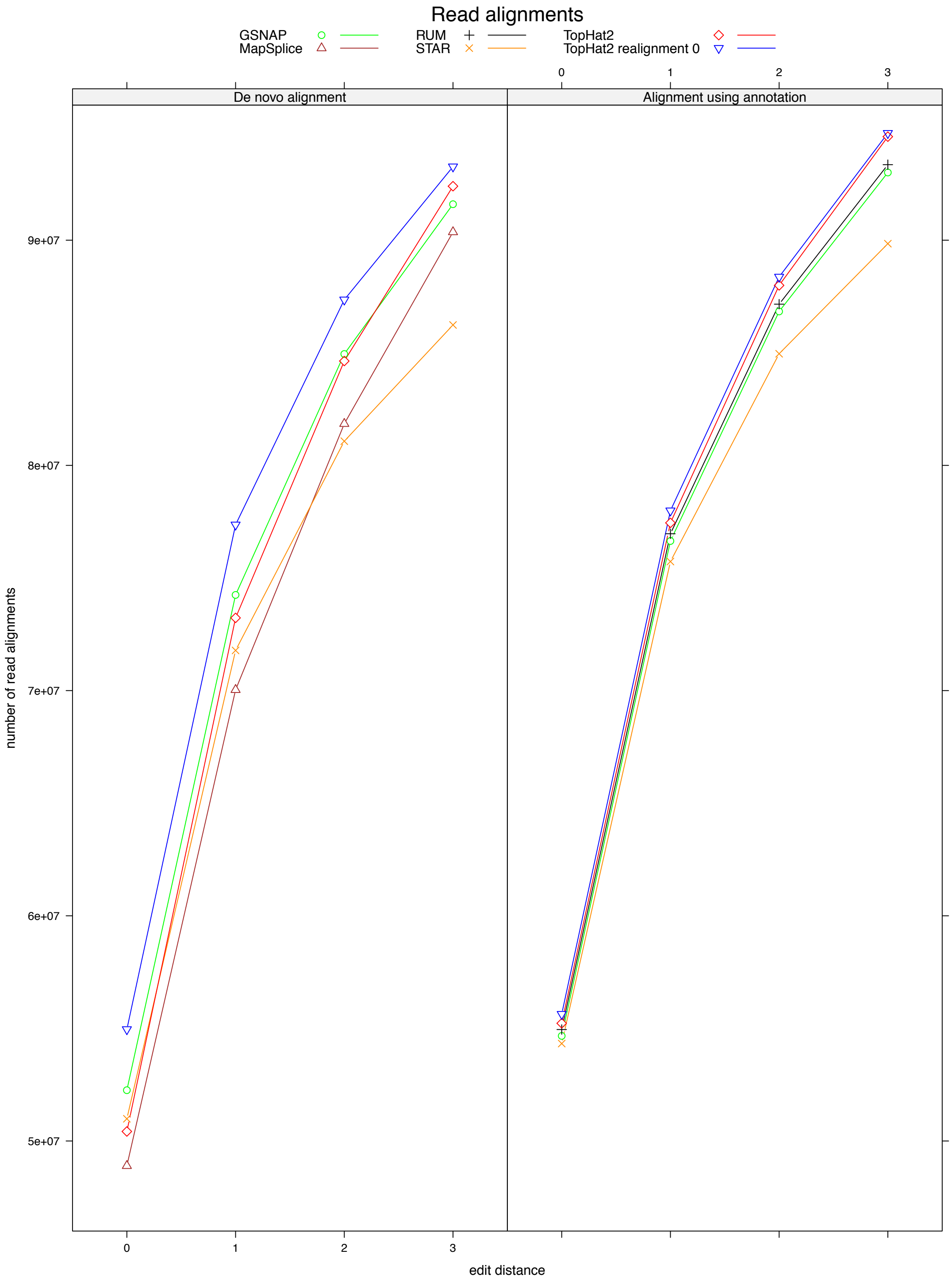


Figure 2.3 The number of read alignments

TopHat2, GSNAP, RUM, MapSplice, STAR are tested for the RNA-seq reads are from Chen et al. [44]. TopHat2 was run without realignment and with realignment (realignment edit distance of 0). TopHat2, GSNAP, and STAR were run in both *de novo* and *gene* mapping modes, while MapSplice and RUM were run only in *de novo* and *gene* mapping modes, respectively. The number of alignments at each edit distance is cumulative. For instance, the number of alignments at edit distance of 2 includes all the alignments with edit distance of 0, 1, 2.

Figure 2.3 shows the alignment performance for each program both with and without using annotations, where all the programs were configured to report alignments with edit distances of up to 3 (and more in some programs). We compared the *de novo* alignments of reads for edit distances of 0, 1, 2, and 3. As expected, all programs find more alignments as the maximum permissible edit distance increases. For edit distance 0 (which only allows perfect matches), TopHat2 without its new realignment function maps noticeably fewer reads than it does with the function. This occurs because TopHat2 first aligns reads end-to-end (with Bowtie2) before trying spliced alignments. Thus if a read is aligned end-to-end with, for example, 1-3 mismatches, then without the realignment function, TopHat2 accepts that alignment and may miss a spliced alignment with fewer mismatches.

On the other hand, TopHat2 with *t*=0 mapped the largest number of reads for all edit distances, followed in most cases by GSNAP. Note that for alignments with an edit distance up to 3, TopHat2 without realignment discovered almost as many alignments as GSNAP.

When alignment methods are run with the assistance of gene annotations (Figure 2.3, right panel), the results are somewhat better than the *de novo* alignments. TopHat2 with or without realignment produced the highest number of mappings, followed by GSNAP, RUM, and STAR. The realignment procedure gives a much small advantage to TopHat2 in these experiments.

One way to estimate the accuracy of mappings is to compare alignments to known splice sites. We compared all aligners on only those reads that required splitting, counting how many known (Figure 2.4, left) and known plus novel (Figure 2.4, right) splice sites they identified. For *de novo* alignment, TopHat2 with realignment has the highest sensitivity, followed by MapSplice. Consistent with our tests on simulated reads, GSNAP and STAR show relatively lower alignment rates. When using annotation, TopHat2 without realignment shows the highest mapping rate, slightly outperforming TopHat2 with realignment. GSNAP and STAR, which do less well, map reads against substrings containing splice sites rather than whole transcripts. Direct mapping against whole transcripts, as done by TopHat2, works well especially when mapping reads spanning small exons, where a single read might span more than two exons.

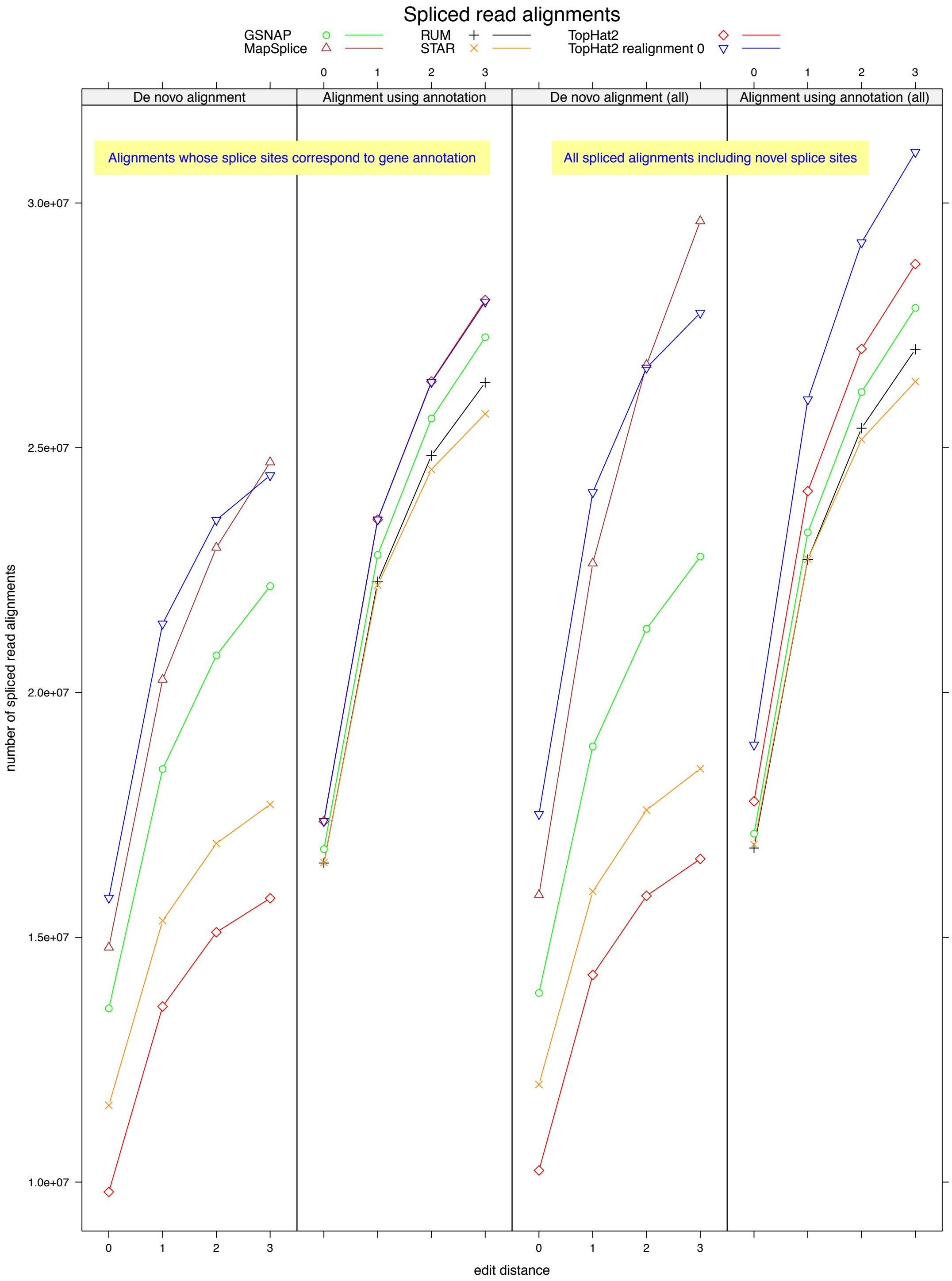


Figure 2.4 The number of spliced read alignments
TopHat2, GSNAP, RUM, MapSplice, STAR are tested for the RNA-seq reads are from Chen et al. [44]. TopHat2, GSNAP, and STAR were run in both *de novo* and *gene* mapping modes. MapSplice and RUM were run in *gene* and *de novo* mapping modes, respectively. For each mapping mode, the left two panels show the number of spliced alignments whose splice sites

are found in the gene annotations and the right two panels show the number of all spliced alignments including novel splice sites.

Based on these results, we would suggest two alternative strategies for alignment with TopHat2. First, if gene annotations are available, as they are for the human genome and some model organisms, then these annotations should be used with TopHat2, even without realignment. Alternatively, if annotations are unavailable or incomplete, then we recommend using TopHat2 with its realignment algorithm to produce the most complete set of alignments.

The runtime and the peak memory usage varied greatly among the programs used in this study. We compared performance on all programs using the Chen et al. data [44], 130 million reads, and results are shown in Table 2.15. Overall, STAR is much faster (32 minutes) than the other programs, which required from 8 to 55 hours. However, STAR requires a large amount of real memory, at least 28 GB, while most other programs required less than 8 GB.

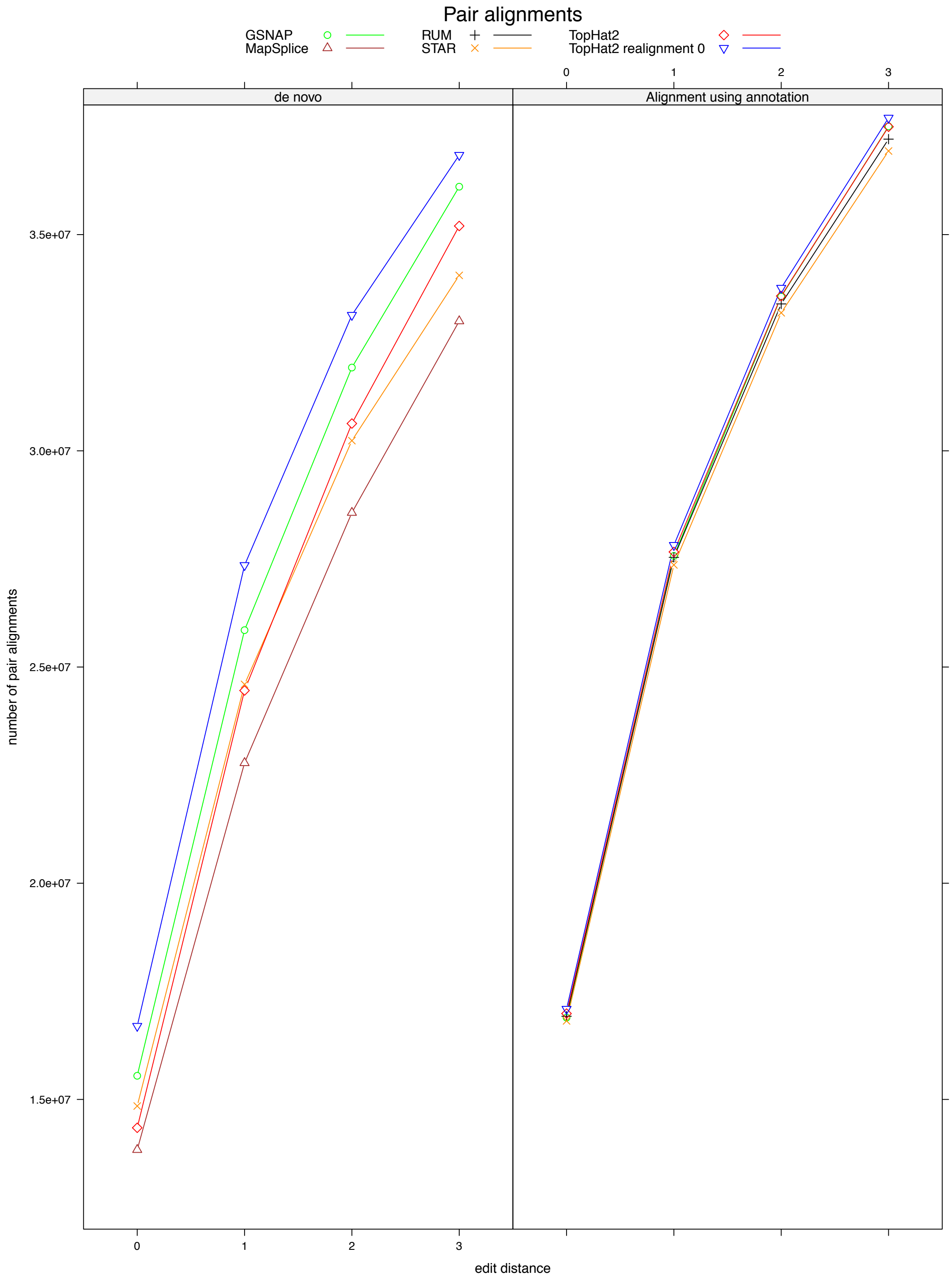


Figure 2.5 The number of pair alignments
TopHat2, GSNAP, RUM, MapSplice, STAR are tested for the RNA-seq reads are from Chen et al. [44].

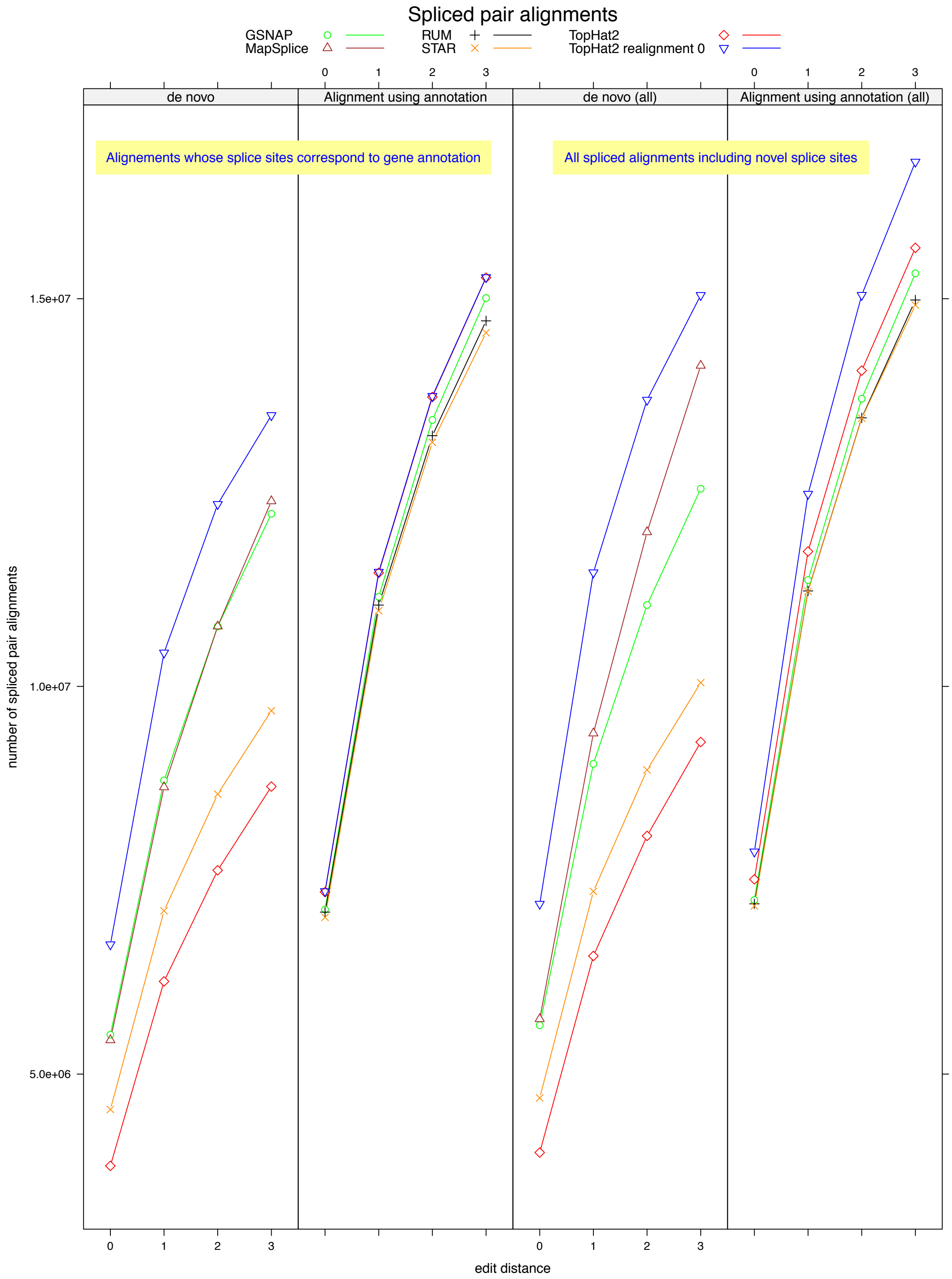


Figure 2.6 The number of spliced pair alignments
TopHat2, GSNAP, RUM, MapSplice, STAR are tested for the RNA-seq reads are from Chen et al. [44].

The impact of pseudogenes on RNA-seq mapping

The Ensembl gene annotations (release 66) contain 32,439 genes, including non-coding RNA genes, and over 14,000 pseudogenes. Of the real genes, we found that 872 (2.7%) genes have pseudogene copies; i.e., at least one transcript (or isoform) can be aligned to a pseudogene with at least 80% identity across the full length of the transcript. Using data from the Chen et al. study [44] and from the Illumina Body Map project [45], we found that genes with pseudogene copies appear to have higher expression levels than those without pseudogene copies. Table 2.5 shows what proportion of reads map to genes with pseudogenes, using both the raw count and a normalized count divided by the length of the transcript. Although only 2.7% of genes have pseudogene copies, these genes account for 22.5% (un-normalized) or 26.9% (normalized) of the RNA-seq reads in the Chen et al. data. In the RNA-seq experiments from the Illumina Body Map (the white blood sample only), we see a 19.1% (normalized) of reads mapping to genes with pseudogenes (Table 2.17). From both RNA-seq experiments, we note that genes with multiple pseudogene copies are more abundantly expressed than those with a single pseudogene copy. We ran a similar analysis looking only at the 20,417 protein-coding genes in Ensembl, with similar results: 22% of read pairs, 26 times more than expected, were mapped to genes with processed pseudogenes (Table 2.18).

Number of pseudogene copies	Gene with pseudogene	Pair Count (%)	Ratio	Normalized count (%)	Normalized ratio
1	553 (1.7%)	6.85	x 4.02	9.37	x 5.49
2	113 (0.4%)	5.15	x 14.79	5.20	x 14.93
3	49 (0.2%)	1.27	x 8.38	1.96	x 12.99

4	27 (0.1%)	2.27	x 27.32	2.28	x 27.35
≥5	130 (0.4%)	6.91	x 17.24	8.08	x 20.16
Total (≥1)	872/32,439 (2.7%)	22.45	x 8.35	26.88	x 10.00

Table 2.5 The expression levels of genes with pseudogene copies
Using Bowtie2, we aligned RNA-seq paired-end reads (Chen et al. [44]) to 32,439 annotated genes. The first column shows the number of pseudogene copies a gene has. The first row shows genes that have just one pseudogene, followed by rows for genes with 2, 3, 4, and at least 5 pseudogene copies. Column 2 ("Gene with pseudogene") is the number of genes with the specified number of pseudogene copies; e.g., 553 genes (1.7% of all genes) have one pseudogene copy. Column 3 shows the percentage of read pairs that were mapped to genes with pseudogene copies. Column 4 contains the ratio of columns 3 and 2. The last two columns are similarly defined using a normalized count, where the number of reads mapping to each gene was normalized to account for gene length.

Figure 2.7 shows various mapping results from TopHat2 with and without realignments at various edit distances. As we allow TopHat2 to realign more reads, it finds the spliced alignments that were otherwise hidden by pseudogene alignments. This in turn substantially increases its mapping rates for known splice sites.

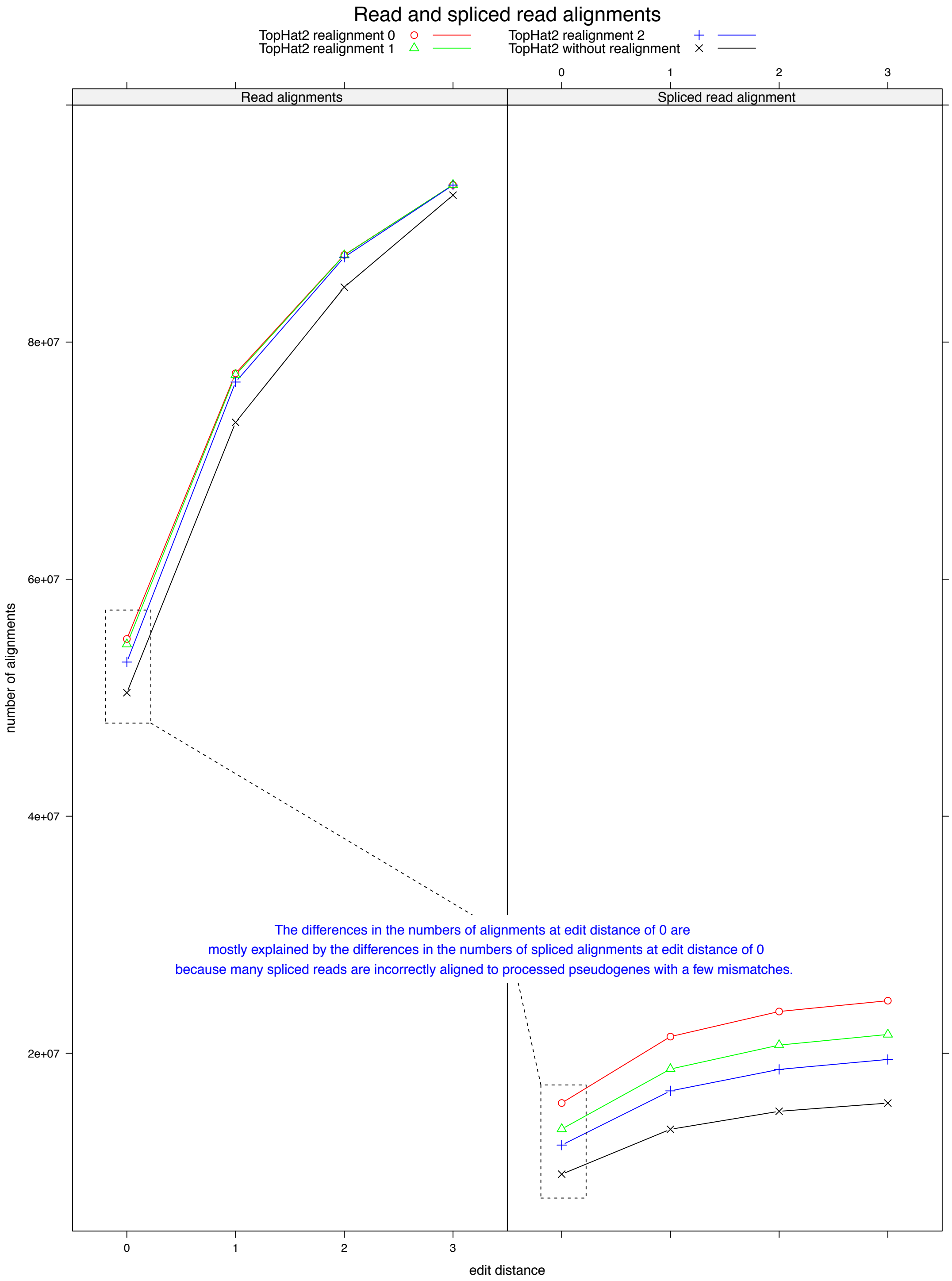


Figure 2.7 The number of read and spliced read alignments

TopHat2 is run using different realignment edit distances of 0, 1, 2 and no-realignment. As TopHat2 allows more realignment from no-realignment to 2 to 1 to 0, the number of read alignments and spliced read alignments increases, where the differences in the numbers of read alignments from TopHat run with different realignment edit distance are mostly explained by the increase in the number of spliced read alignments.

The completeness of human gene annotations

Using the *de novo* mapping mode in TopHat2, GSNAP, MapSplice, and STAR, we looked at how many spliced alignments are found in the Ensembl annotations. As shown in Figure 2.8, the proportions of spliced mappings to known splice sites are 97%, 96%, 88-90%, and 83-93% in GSNAP, STAR, TopHat2, and MapSplice, respectively. Although our analysis only considered RNA-seq data from Chen et al. [44], the TopHat2 result suggests that many additional spliced alignments, up to 12%, might remain to be discovered. Most of the novel splicing events in these alignments are supported by ≥10 reads that extend for ≥50 bases on each side.

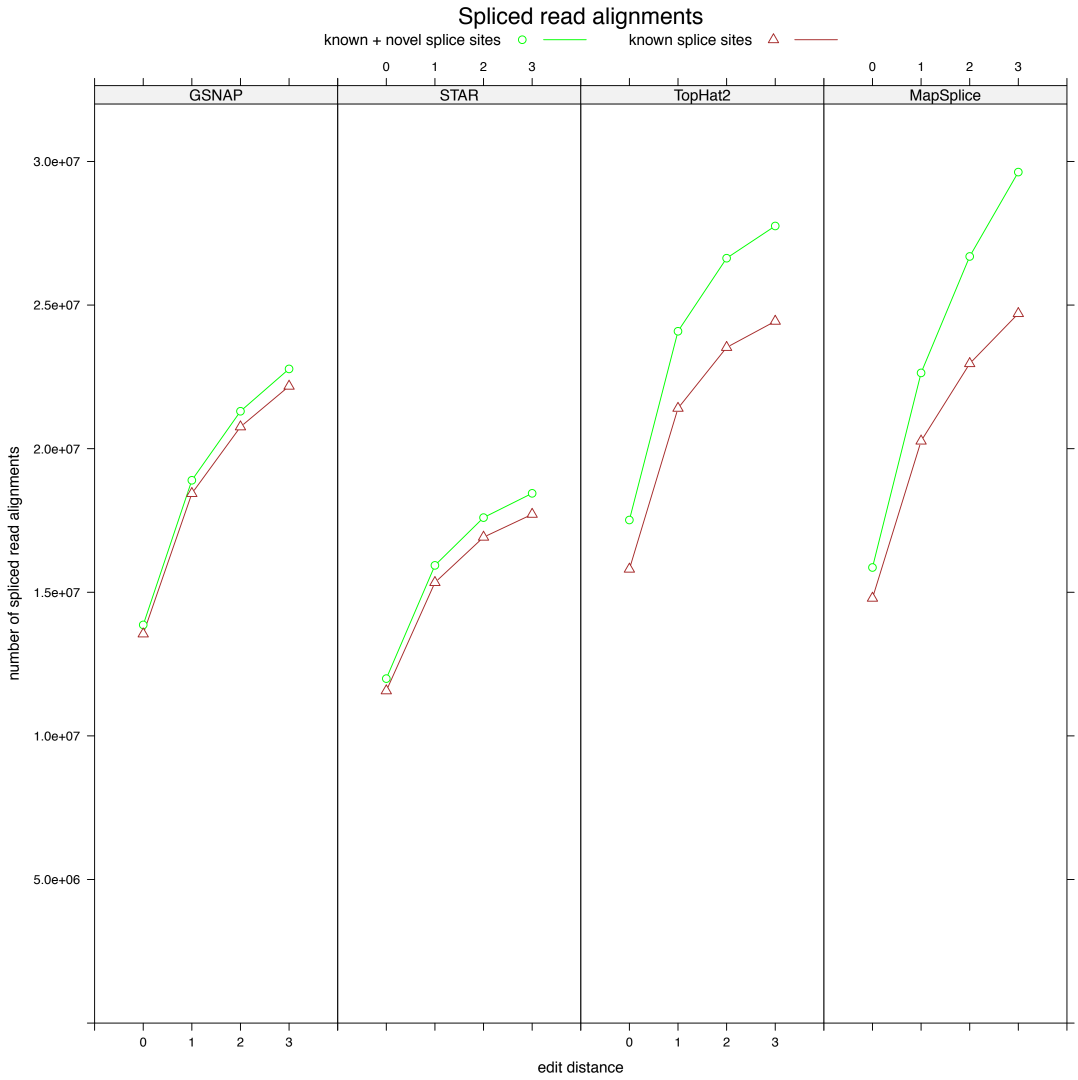


Figure 2.8 The number of spliced read alignments
TopHat2, GSNAP, STAR, and MapSplice are tested without using gene annotation for
The number of read alignments whose splice sites are found in the gene annotations are shown in brown color. The number of all spliced read alignments including novel splice sites are shown in green color.

2.4 Conclusions

Discovery of new genes and transcripts is a major objective in many RNA-seq experiments. Deep RNA-seq experiments continue to uncover previously unseen elements of the transcriptome even in well-studied organisms. Mapping reads to the genome is a core step in such screens, and the accuracy of mapping software can determine the accuracy of downstream steps such as gene and transcript discovery or expression quantitation.

I have described TopHat2, which provides major accuracy improvements over previous versions and other RNA-seq mapping tools. Because TopHat2 is built around Bowtie2, it can now align reads across small indels with high accuracy – a feature critical for studies assessing the impact of genetic mutations on gene and transcript expression. I have engineered TopHat2 to work well with a wide range of RNA-seq experimental designs, and it is optimized for the widely available long, paired-end reads. These reads pose new challenges because they can span multiple splice sites rather than just one or two – we estimate that nearly half of reads 150-bp long would span more than two human exons. The algorithmic improvements in TopHat2 address this challenge, maintaining both accuracy and speed. Other refinements to the algorithm increase accuracy for reads that span a junction with only a small (≤10 bp) overhang, reducing errors in downstream transcript assembly using tools such as Cufflinks. TopHat2 also makes powerful use of available gene annotations, which allow it to avoid erroneously mapping reads to pseudogenes and generally improve its overall alignment accuracy. Annotation also allows TopHat2 to better align reads that cover microexons, noncanonical splice sites, and other "unusual" features of eukaryotic transcriptomes.

TopHat2 has proved to perform well over a wide range of read lengths, making it a good fit for most RNA-seq experimental designs. This scalability suggests that as read lengths grow, TopHat2 will continue to report accurate, sensitive alignment results and allow for robust downstream analysis. We argue that TopHat2 reports more accurate alignments than competing tools using fewer computational resources. RNA-seq experiments are becoming increasingly common and are now routinely used by many biologists. We expect that TopHat2 will provide these scientists with accurate results for use with expression analysis, gene discovery, and many other applications.

2.5 Supplementary Material

Alignments of simulated reads with up to 3 mismatches

We generated single-end and paired-end reads with 0 to 3 mismatches and without indels as shown in Tables 2.6 and 2.8. TopHat2 and MapSplice show the highest mapping sensitivity in read/pair and spliced read/pair alignments for both true mismatches (SNPs) and sequencing-error mismatches (Tables 2.7 and 2.9).

Type	No. of total reads	No. of reads without mismatches (junction)	No. of reads with 1 mismatch (junction)	No. of reads with 2 mismatches (junction)	No. of reads with 3 mismatches (junction)
True mismatches	20,000,000	10,860,864 (4,654,864)	7,579,737 (2,289,006)	1,396,742 (428,136)	162,657 (46,185)
Sequencing-error mismatches	20,000,000	11,258,169 (4,010,662)	7,298,699 (2,610,246)	1,297,051 (462,717)	146,081 (52,481)

Table 2.6 The number of reads and spliced reads with up to 3 mismatches

Program	True mismatches								Sequencing-error mismatches							
	M0	M1	M2	M3	J0	J1	J2	J3	M0	M1	M2	M3	J0	J1	J2	J3
TopHat2 +Bowtie1	98.14	98.71	98.83	97.57	95.81	95.86	96.45	91.52	98.37	98.60	98.79	97.19	95.67	96.23	96.71	92.23
TopHat2 +Bowtie2	97.85	98.70	95.08	86.72	95.00	95.75	84.59	55.21	98.08	98.54	93.87	84.98	94.61	95.95	83.16	58.69
GSNAP	92.85	89.08	83.50	78.33	83.33	77.49	74.19	70.27	93.95	88.19	83.09	77.66	83.03	77.61	74.29	69.35
RUM	85.10	83.45	77.58	73.82	65.25	54.29	45.82	37.93	87.58	81.43	75.37	69.57	65.13	55.16	45.55	36.63
MapSplice	96.77	98.25	97.96	93.94	92.47	94.25	96.98	96.95	96.85	97.77	97.78	94.63	91.16	93.79	95.94	95.82
STAR	90.39	87.84	82.15	78.52	77.65	68.96	61.18	55.10	91.82	86.35	80.68	75.39	77.17	69.07	60.72	53.36

Table 2.7 The recall rates of read and spliced read alignments for true mismatches (SNPs) and sequencing-error mismatches

M0 is the sensitivity of read alignments with zero mismatches. M1 is the sensitivity of alignments with one mismatch. M2 and M3 are similarly defined. J0 is the sensitivity of spliced alignments with no mismatches. J1, J2, and J3 are similarly defined with mismatches of 1, 2, and 3, respectively, for spliced alignments. M0, M1, M2, and M3 also include spliced alignments as well as non-gapped alignments. Note that TopHat2 with Bowtie2 suffers a drop in performance compared to Bowtie1 when a single read has 3 mismatches (column J3). This occurs because TopHat2 splits reads into very short segments, 25 bp, when attempting to align across splice sites. TopHat2 then calls Bowtie1/2 to align these short segments. Bowtie2's default parameters are not designed for such short segments; however these can easily be modified by changing the parameters used to call Bowtie2 within TopHat2.

Type	No. of total pairs	No. of pairs without mismatches (junction)	No. of pairs with 1 mismatch (junction)	No. of pairs with 2 mismatches (junction)	No. of pairs with ≥ 3 mismatches (junction)
True mismatches	20,000,000	5,703,884 (3,809,739)	8,547,831 (4,031,062)	4,345,589 (1,779,403)	1,402,696 (557,825)
Sequencing-error mismatches	20,000,000	5,747,299 (2,818,790)	9,201,311 (4,553,143)	3,897,205 (1,923,580)	1,154,185 (568,559)

Table 2.8 The number of pairs and spliced pairs with mismatches of 0 to 3

The two types of pair reads are simulated: true mismatches (SNPs) and sequencing-error mismatches. Note that each read can contain up to 3 mismatches, it is possible that a pair can have more than 3 mismatches.

Program	True mismatches								Sequencing-error mismatches							
	M0	M1	M2	M≥3	J0	J1	J2	J≥3	M0	M1	M2	M≥3	J0	J1	J2	J≥3
TopHat2 +Bowtie1	95.69	96.96	97.47	97.45	93.03	93.14	93.64	93.12	96.72	96.98	97.19	96.91	93.19	93.78	94.15	93.51
TopHat2 +Bowtie2	95.06	96.77	96.01	91.94	91.90	92.52	89.40	78.18	96.10	96.59	94.92	89.86	91.46	92.60	89.33	78.12
GSNAP	84.03	83.95	79.29	72.72	74.03	69.83	64.84	58.95	88.34	83.20	77.52	70.99	73.84	69.76	64.84	59.17
RUM	69.86	73.97	72.22	67.81	51.85	45.42	39.32	33.22	78.40	72.92	68.09	63.08	52.57	46.43	40.03	33.66
MapSplice	90.53	92.59	93.33	92.47	84.70	84.88	85.59	85.98	91.90	91.77	91.97	91.48	83.57	83.67	84.23	84.73
STAR	79.41	81.17	78.07	60.80	66.65	61.01	55.04	41.48	85.05	80.12	75.03	58.85	66.64	61.25	55.01	41.49

Table 2.9 The recall rates of pair and spliced pair alignments for true mismatches (SNPs) and sequencing-error mismatches

M0 is the sensitivity of read alignments with zero mismatches. M1 is the sensitivity of alignments with one mismatch. M2 and M3 are similarly defined with mismatches of 1, 2, and ≥3, respectively. J0 is the sensitivity of spliced alignments with zero mismatches. J1, J2, and J≥3 are similarly defined for spliced alignments.

Corresponding tables for Figures 2.3-2.7

	Program	0	1	2	3
De novo alignment	TopHat2 realignment	54,956,129	77,364,055	87,355,369	93,265,424
	TopHat	50,422,413	73,228,140	84,633,702	92,396,448
	GSNAP	52,255,865	74,247,781	84,946,229	91,598,102
	MapSplice	48,896,741	70,032,327	81,847,468	90,360,661
	STAR	50,986,666	71,782,717	81,074,505	86,235,516
Alignment using annotation	TopHat2 realignment	55,634,580	77,988,848	88,370,540	94,752,200
	TopHat	55,225,852	77,447,497	87,992,406	94,596,600
	GSNAP	54,666,282	76,642,607	86,835,392	93,005,273
	RUM	54,949,609	76,963,699	87,157,875	93,352,293
	STAR	54,326,036	75,730,313	84,957,399	89,844,775

Table 2.10 Table for Figure 2.3.

Type		Program	0	1	2	3
Alignments whose splice sites correspond to gene annotation	De novo alignment	TopHat2 realignment	15,804,625	21,406,115	23,524,839	24,436,600
		TopHat	9,799,757	13,586,453	15,104,339	15,798,045
		GSNAP	13,549,591	18,438,736	20,759,433	22,175,182
		MapSplice	14,792,707	20,264,394	22,961,083	24,704,514
		STAR	11,568,529	15,338,930	16,918,024	17,714,913
	Alignment using annotation	TopHat2 realignment	17,372,910	23,531,960	26,340,120	27,982,780
		TopHat	17,368,853	23,530,365	26,353,413	28,018,284
		GSNAP	16,801,716	22,812,953	25,598,496	27,259,090
		RUM	16,516,786	22,263,594	24,839,636	26,331,306
		STAR	16,526,673	22,195,936	24,558,091	25,693,885
All spliced alignments including novel splice sites	De novo alignment	TopHat2 realignment	17,516,565	24,088,224	26,632,215	27,754,233
		TopHat	10,238,968	14,232,391	15,847,929	16,601,804
		GSNAP	13,864,319	18,899,654	21,302,999	22,777,308

		MapSplice	15,863,181	22,638,514	26,692,556	29,630,048
		STAR	11,994,236	15,936,866	17,600,134	18,445,153
	Alignment using annotation	TopHat2 realignment	18,932,114	25,985,178	29,191,692	31,039,091
		TopHat	17,779,753	24,112,605	27,019,281	28,752,182
		GSNAP	17,117,374	23,272,081	26,138,915	27,858,112
		RUM	16,823,909	22,716,678	25,399,661	27,009,138
		STAR	16,895,367	22,725,029	25,170,512	26,352,382

Table 2.11 Table for Figure 2.4.

	Program	0	1	2	3
De novo alignment	TopHat2 realignment	16,696,682	27,353,265	33,139,753	36,839,143
	TopHat	14,344,271	24,456,802	30,630,922	35,199,608
	GSNAP	15,546,886	25,853,039	31,925,593	36,108,336
	MapSplice	13,835,185	22,781,288	28,568,799	32,999,167
	STAR	14,847,145	24,598,381	30,235,116	34,057,210
Alignment using annotation	TopHat2 realignment	17,091,131	27,818,953	33,766,156	37,699,996
	TopHat	16,985,383	27,661,740	33,579,775	37,494,323
	GSNAP	16,890,487	27,569,140	33,566,349	37,503,456
	RUM	16,923,302	27,536,281	33,397,563	37,208,206
	STAR	16,815,984	27,361,365	33,191,954	36,933,241

Table 2.12 Table for Figure 2.5.

Type		Program	0	1	2	3
Alignments whose splice sites correspond to gene annotation	De novo alignment	TopHat2 realignment	6,670,997	10,434,104	12,349,897	13,496,341
		TopHat	3,816,460	6,195,116	7,628,348	8,708,508
		GSNAP	5,507,359	8,787,161	10,773,698	12,226,898
		MapSplice	5,438,391	8,701,358	10,775,190	12,389,538
		STAR	4,543,781	7,106,244	8,610,236	9,685,835
	Alignment using annotation	TopHat2 realignment	7,357,496	11,476,154	13,743,868	15,276,373
		TopHat	7,346,821	11,464,534	13,733,001	15,272,369
		GSNAP	7,121,858	11,156,844	13,436,485	15,009,697
		RUM	7,088,842	11,048,936	13,233,486	14,714,367
		STAR	7,021,511	10,975,663	13,147,910	14,561,264

All spliced alignments including novel splice sites	De novo alignment	TopHat2 realignment	7,193,604	11,468,318	13,694,621	15,045,756
		TopHat	3,988,139	6,523,309	8,072,162	9,282,468
		GSNAP	5,630,093	9,002,188	11,049,842	12,550,023
		MapSplice	5,710,435	9,395,612	11,990,324	14,137,678
		STAR	4,692,109	7,355,651	8,922,567	10,048,254
	Alignment using annotation	TopHat2 realignment	7,868,376	12,481,943	15,047,081	16,764,777
		TopHat	7,511,707	11,740,351	14,073,199	15,656,508
		GSNAP	7,245,286	11,371,551	13,710,608	15,328,468
		RUM	7,195,805	11,231,525	13,463,953	14,983,419
		STAR	7,169,487	11,224,944	13,458,212	14,917,855

Table 2.13 Table for Figure 2.6.

	Type	0	1	2	3
Read alignments	Realignment 0	54,956,129	77,364,055	87,355,369	93,265,424
	Realignment 1	54,508,641	77,227,362	87,334,380	93,272,963
	Realignment 2	53,007,141	76,631,857	87,168,673	93,244,130
	No Realignment	50,422,413	73,228,140	84,633,702	92,396,448
Spliced read alignments	Realignment 0	17,516,565	24,088,224	26,632,215	27,754,233
	Realignment 1	14,179,269	19,895,371	22,278,929	23,389,758
	Realignment 2	12,755,976	17,578,938	19,577,384	20,558,593
	No Realignment	10,238,968	14,232,391	15,847,929	16,601,804

Table 2.14 Table for Figure 2.7.

Alignment rates for reads of different lengths (error-free)

In addition to 100 bp simulated reads in the main text we also generated single and paired-end reads of different lengths (50, 150, 200 bp), in order to check how TopHat2 works compared to the other alignment software. We used different fragment lengths 200, 250, 350, 450 bp for read lengths 50, 100, 150, 200 bp, respectively. Figure 2.9 shows TopHat2 performs better than the other programs for different read lengths. TopHat2 also outputs much more accurate alignments for spliced reads and spliced reads with small anchors. These results suggest that TopHat2 may be the better choice for longer reads (≥150 bp) that will likely become prevalent in the near future, as well as for currently available reads (50 ~ 100 bp).

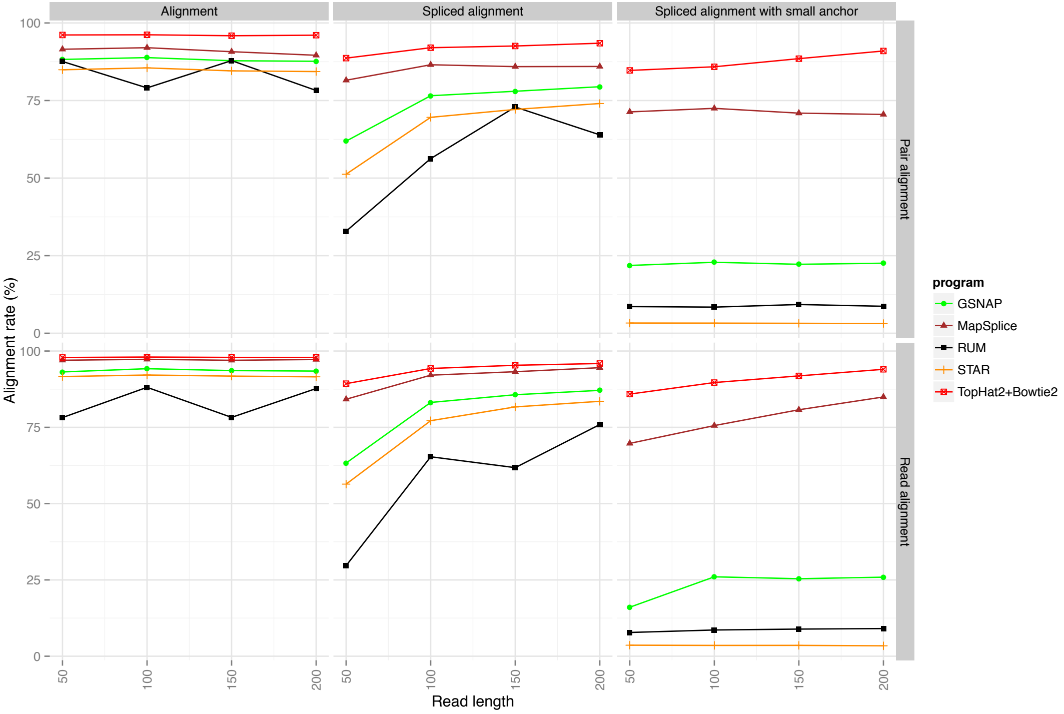


Figure 2.9 Mapping accuracy in different read lengths

Using simulated reads (20 million reads and 20 million pairs), the figure shows the ratio of correctly aligned reads (bottom) or pairs (top) for read alignment (the left column), spliced read alignment (the middle column), and spliced read alignment with small anchors (the right column).

Simulation of reads with indels and mismatches

We used the transcript expression model from the Flux simulator [49] to generate RNA-seq reads from the protein coding genes found in the Ensembl human gene annotation, release 66. First, the transcripts from the protein coding genes are randomly ranked. Then, the expression levels of the transcripts are modeled as follows. The expression level y of a transcript is defined as $y = \left(\frac{x}{x_0}\right)^k e^{-\left(\frac{x}{x_1}\right)-\left(\frac{x}{x_1}\right)^2}$, where x is the rank number of a transcript, $x_0 = 5 \times 10^7$, $x_1 = 9500$, and $k = -0.6$.

Reads are simulated for the purpose of testing the alignment programs instead of trying to precisely mimic real RNA-seq experiments. When generating reads with true indels, we include at most one indel per exon in a way that if the length of an exon L is greater than or equal to 1000 bp, we place either an insertion (50%) or a deletion (50%) into the exon at a random location, otherwise an indel is introduced into a random location of the transcript with the chance of $\frac{L}{1000}$. Reads are generated from these transcripts so that they share the same changes. For reads with true mismatches we change the nucleotides of each transcript in such a way that the average distance between two nearby mismatches is 150.5 bp and the distribution of the distance is uniform (1 to 300). Reads are then generated from these modified transcripts. Reads with either indels or mismatches from sequencing errors are simulated in the same way except the transcript being used is changed every time a read is generated.

Figure 2.10 shows the proportions of reads spanning multiple exons, which increase approximately from 19% to 46% as the length of reads increases from 50 to 150 bp. On the other hand, as we may expect, the fragment length does not affect the proportions of spliced reads.

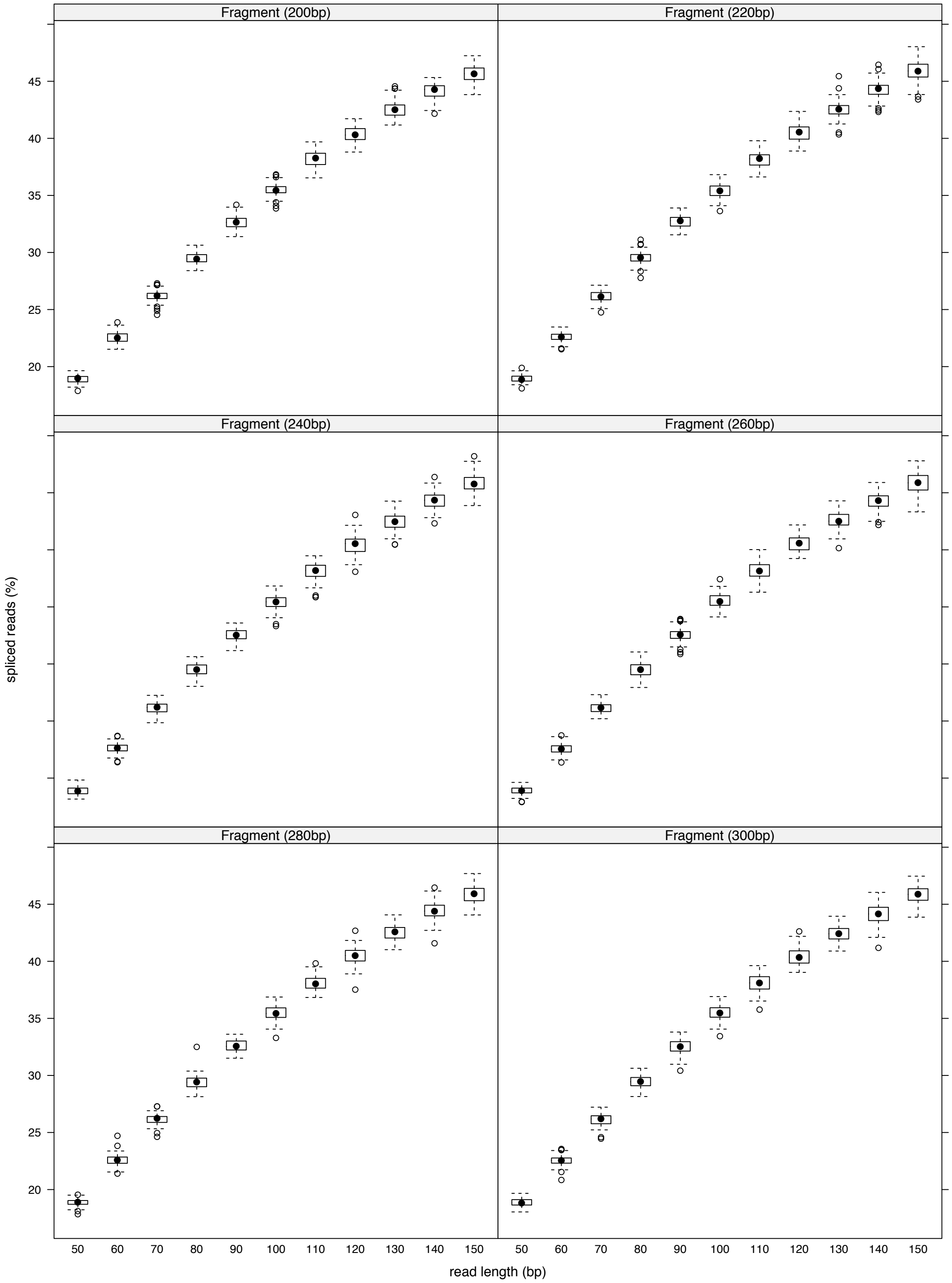


Figure 2.10 Proportions of spliced reads (various read and fragment lengths)
This figure shows proportions of spliced reads from different read lengths (50 to 150 bp) and fragment lengths (200 to 300 bp). For each fragment length (200, 220, 240, 260, 280, 300 bp), a whisker box plot shows 100 simulation results (the percentage of spliced reads) for each read length.

Runtime and memory usage

With ~130 million paired-end reads from Chen et al. [44], we ran each program using 8 threads on a Linux machine with memory of 256GB and 48 AMD processors (2.1GHz). Runtime (or wall time) and peak memory usage were measured using the GNU *time* program as shown in Table 2.15.

Program	Runtime (wall time)	Peak memory (GB)	Parameters
TopHat2 2.0.8 (Transcriptome only mapping)	8h 29m	4.9	-G --transcriptome-only --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Default: genome and spliced mapping)	17h 1m	5.4	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (With transcriptome mapping)	17h 31m	5.2	-G --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Realignment with realignment edit distance of 0)	29h 55m	5.6	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60 --read-realign-edit-dist 0
GSNAP 2013-01-23	55h 26m	7.6	--max-mismatches=3 -N 1
RUM 1.12_01	26h 34m	*36.4	
MapSplice 1.15.2	44h 50m	3.7	min_missed_seg = 0
STAR 2.3.0e	32m	27.8	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3

Table 2.15 Runtime and memory usage of RNA-seq alignment software
Note the last column "Parameters" shows specific parameters for each program to allow a read to be aligned with edit distance of 0, 1, 2, and 3. Parameters for specifying genome, gene annotation, RNA-seq read files, and the number of threads are not shown. The version of each program is shown in blue color in the first column. *Note that RUM uses separate processes, each of which consisted of Bowtie (2394MB) and BLAT (4660MB), requiring a total of 36.4GB memory when using 8 threads.

Specific program parameters

Test	Program	Reference genome	Gene annotation	Specific parameters
Alignments of simulated reads (error-free)	TopHat2 +Bowtie1	Whole human genome	No	--mate-inner-dist 50 --mate-std-dev 40 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40
	TopHat1.1.4			--mate-inner-dist 50 --mate-std-dev 40
	GSNAP			-N 1
	RUM		Yes	
	MapSplice		No	min_missed_seg = 0
	STAR			--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of simulated reads with short indels (1-3 bp)	TopHat2 +Bowtie1			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3
	GSNAP			--max-mismatches=3 --indel-penalty=1 -N 1
	RUM		Yes	

	MapSplice		No	min_missed_seg = 0
	STAR			--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of simulated reads with up to 3 mismatches	TopHat2 +Bowtie1			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3
	TopHat1.1.4			--mate-inner-dist 50 --mate-std-dev 40
	GSNAP			--max-mismatches=3 -N 1
	RUM		Yes	
	MapSplice		No	min_missed_seg = 0
	STAR			--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of a large set of real RNA-seq reads (Chen et al. [44])	TopHat2	Whole human genome	Yes/No	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
	TopHat2 realignment 0		Yes/No	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60 --read-realign-edit-dist 0
	GSNAP		Yes/No	--max-mismatches=3

				-N 1
	RUM		Yes	
	MapSplice		No	min_missed_seg = 0
	STAR		Yes/No	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3

Table 2.16 Specific program parameters
Program parameters to specify genome, gene annotation, and RNA-seq read files are given in the table (the number of threads is not shown). Note that for simulation data set, a TopHat option "--read-realign-edit-dist" can be used to realign reads in the spliced alignment phase that are mapped against either transcriptome or genome.

Number of pseudogene copies	Gene with pseudogene	Pair Count (%)	Ratio	Normalized count (%)	Normalized ratio
1	553 (1.7%)	4.66	x 2.73	7.33	x 4.30
2	113 (0.4%)	3.51	x 10.08	3.97	x 11.39
3	49 (0.2%)	0.62	x 4.13	1.05	x 6.96
4	27 (0.1%)	1.32	x 15.82	1.52	x 18.30
≥5	130 (0.4%)	3.61	x 9.01	5.23	x 13.04
Total (≥1)	872/32,439 (2.7%)	13.72	x 5.11	19.10	x 7.11

Table 2.17 The expression levels of genes with pseudogene copies
llumina Body Map 2.0 data [45] is used. Columns are defined as in Table 2.5.

Number of pseudogene copies	Protein-coding gene with processed pseudogene	Pair Count (%)	Ratio	Normalized count (%)	Normalized ratio
1	267 (1.31%)	6.88	x 5.26	9.55	x 7.30
2	47 (0.23%)	6.31	x 27.42	6.07	x 26.39
3	21 (0.10%)	1.27	x 12.38	1.97	x 19.15
4	16 (0.08%)	0.84	x 10.73	1.02	x 13.02
≥5	40 (0.20%)	6.73	x 34.33	7.92	x 40.45
Total (≥1)	391/20,417 (1.92%)	22.03	x 11.50	26.54	x 13.86

Table 2.18 The expression levels of protein-coding genes with processed pseudogene copies
The RNA-seq data from Chen et al. [44] is used. Columns are defined as in Table 2.5.

Chapter 3: TopHat-Fusion: an algorithm for discovery of novel fusion transcripts

I have developed novel algorithms and them into TopHat-Fusion in order to discover transcripts representing fusion gene products, which result from the breakage and re-joining of two different chromosomes, or from rearrangements within a chromosome. TopHat-Fusion is a part of TopHat2 with the simple command line switch, an efficient program that aligns RNA-seq reads without relying on existing annotation. Because it is independent of gene annotation, TopHat-Fusion can discover fusion products deriving from known genes, unknown genes and unannotated splice variants of known genes. Using RNA-seq data from breast and prostate cancer cell lines, we detected both previously reported and novel fusions with solid supporting evidence. TopHat-Fusion is available at http://genomics.jhu.edu/software/tophat/fusion_index.html.

3.1 Background

Direct sequencing of messenger RNA transcripts using the RNA-seq protocol [5-7] is rapidly becoming the method of choice for detecting and quantifying all the genes being expressed in a cell. One advantage of RNA-seq is that, unlike microarray expression techniques, it does not rely on pre-existing knowledge of gene content, and therefore it can detect entirely novel genes and novel splice variants of existing genes. In order to detect novel genes, however, the software used to analyze RNA-seq experiments must be able to align the transcript sequences anywhere on the genome, without relying on existing annotation. TopHat [23] was one of the first spliced alignment programs able to perform such *ab initio* spliced alignment, and in combination with the Cufflinks program [20], it is part of a software analysis suite that can detect and quantify the complete set of genes captured by an RNA-seq experiment.

In addition to detection of novel genes, RNA-seq has the potential to discover genes created by complex chromosomal rearrangements. 'Fusion' genes formed by the breakage and re-joining of two different chromosomes have repeatedly been implicated in the development of cancer, notably the *BCR*/*ABL1* gene fusion in chronic myeloid leukemia [17, 25, 26]. Fusion genes can also be created by the breakage and rearrangement of a single chromosome, bringing together transcribed sequences that are normally separate. As of November 2012, the Mitelman database [27] documented nearly 62,000 cases of chromosome aberrations and gene fusions in cancer. Discovering these fusions via RNA-seq has a distinct advantage over whole-genome sequencing, due to the fact that in the highly rearranged genomes of some tumor samples, many rearrangements might be present although only a fraction might alter transcription. RNA-seq identifies only those chromosomal fusion events that produce transcripts. It has the further advantage that it allows one to detect multiple alternative splice variants that might be produced by a fusion event. However, if a fusion involves only a non-transcribed promoter element, RNA-seq will not detect it.

In order to detect such fusion events, special purpose software is needed for aligning the relatively short reads from next-generation sequencers. In this chapter, we describe a new method, TopHat-Fusion, designed to capture these events. We demonstrate its effectiveness on six different cancer cell lines, in each of which it found multiple gene fusion events, including both known and novel fusions. Although other algorithms for detecting gene fusions have been described recently [18, 50], these methods use unspliced alignment software (for example, Bowtie [48] and ELAND [51]) and rely on finding paired reads that map to either side of a fusion boundary. They also rely on known annotation, searching known exons for possible fusion boundaries. In contrast, TopHat-Fusion directly detects individual reads (as well as paired reads) that span a fusion event, and because it does not rely on annotation, it finds events involving novel splice variants and entirely novel genes.

Other recent computational methods that have been developed to find fusion genes include SplitSeek [52], a spliced aligner that maps the two non-overlapping ends of a read (using 21 to 24 base anchors) independently to locate fusion events. This is similar to TopHat-Fusion, which splits each read into several pieces, but SplitSeek supports only SOLiD reads. A different strategy is used by Trans-ABySS [53], a *de novo* transcript assembler, which first uses ABySS [54] to assemble RNA-seq reads into full-length transcripts. After the assembly step, it then uses BLAT [55] to map the assembled transcripts to detect any that discordantly map across fusion points. This is a very time-consuming process: it took 350 CPU hours to assemble 147 million reads and >130 hours for the subsequent mapping step. ShortFuse [56] is similar to TopHat in that it first uses Bowtie to map the reads, but like other tools it depends on read pairs that map to discordant positions. FusionSeq [33] uses a different alignment program for its initial alignments, but is similar to TopHat-Fusion in employing a series of sophisticated filters to remove false positives.

TopHat-Fusion is incorporated into TopHat2 with the simple command line switch and the filtering step of TopHat-Fusion is also included in TopHat2 package. The tutorial can be found at http://genomics.jhu.edu/software/tophat/fusion_index.html.

3.2 Methods

The first step in analysis of an RNA-seq data set is to align (map) the reads to the genome, which is complicated by the presence of introns. Because introns can be very long, particularly in mammalian genomes, the alignment program must be capable of aligning a read in two or more pieces that can be widely separated on a chromosome. The size of RNA-seq data sets, numbering in the tens of millions or even hundreds of millions of reads, demands that spliced alignment programs also be very efficient. The TopHat program achieves efficiency primarily through the use of the Bowtie aligner [48], an extremely fast and memory-efficient program for aligning unspliced reads to the genome. TopHat uses Bowtie to find all reads that align entirely within exons, and creates a set of partial exons from these alignments. It then creates hypothetical intron boundaries between the partial exons, and uses Bowtie to re-align the initially unmapped (IUM) reads and find those that define introns.

TopHat-Fusion implements several major changes to the original TopHat algorithm, all designed to enable discovery of fusion transcripts (Figure 3.1). After identifying the set of IUM reads, it splits each read into multiple 25-bp pieces, with the final segment being 25 bp or longer; for example, an 80-bp read will be split into three segments of length 25, 25, and 30 (Figure 3.2).

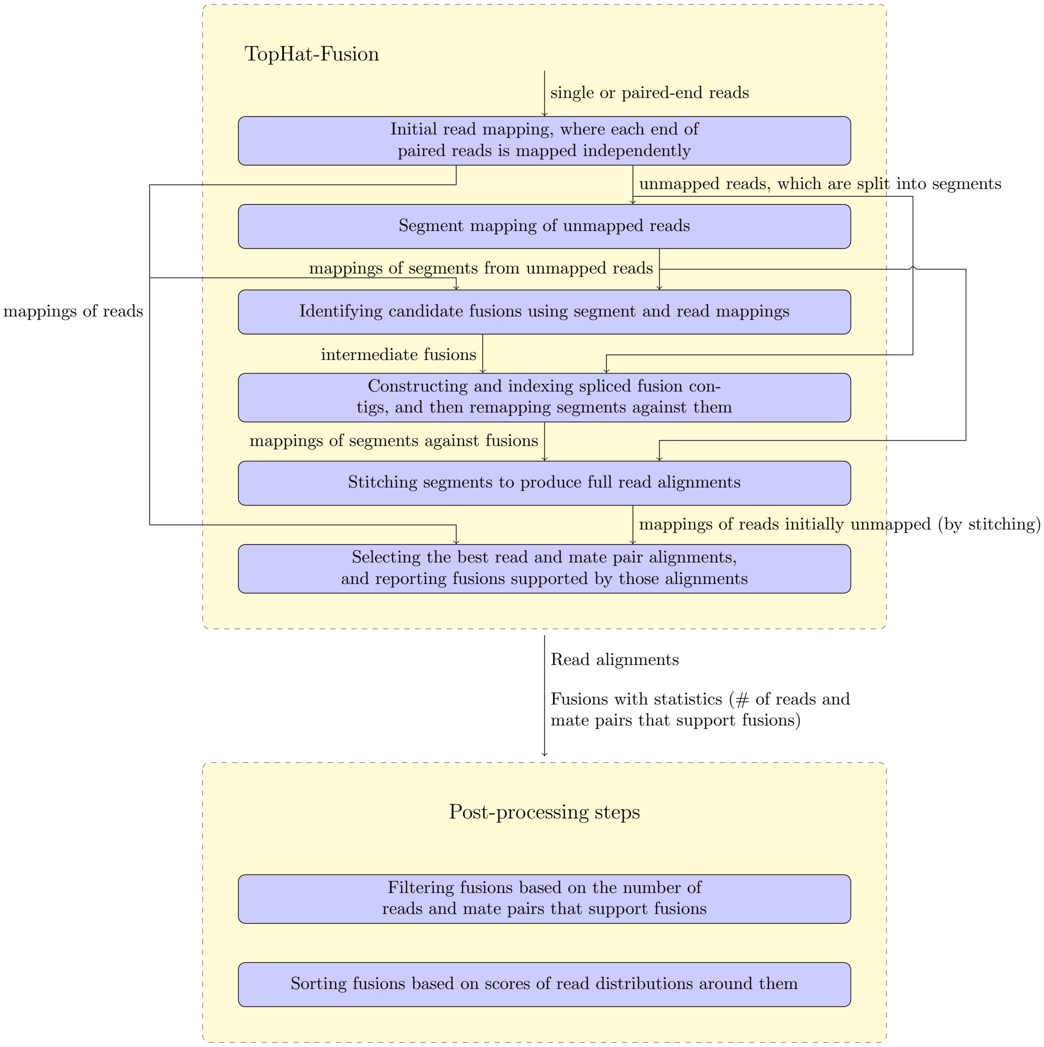
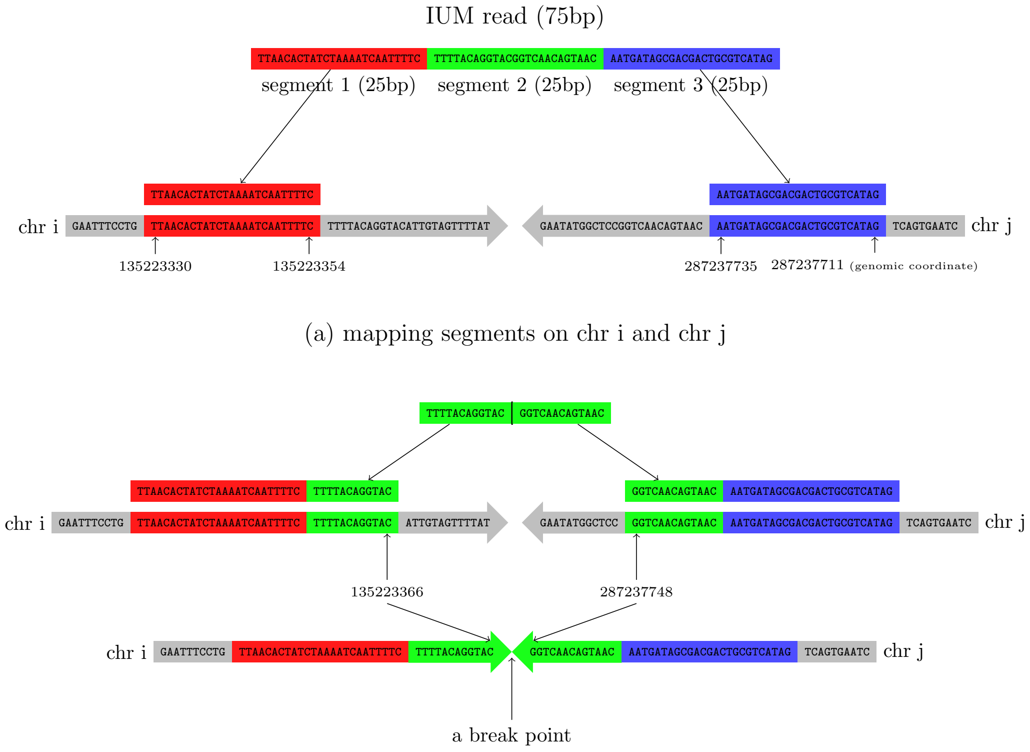

TopHat-Fusion
single or paired-end reads
Initial read mapping, where each end of paired reads is mapped independently
unmapped reads, which are split into segments
Segment mapping of unmapped reads
mappings of segments from unmapped reads
mappings of reads
Identifying candidate fusions using segment and read mappings
intermediate fusions
Constructing and indexing spliced fusion contigs, and then remapping segments against them
mappings of segments against fusions
Stitching segments to produce full read alignments
mappings of reads initially unmapped (by stitching)
Selecting the best read and mate pair alignments, and reporting fusions supported by those alignments
Read alignments
Fusions with statistics (# of reads and mate pairs that support fusions)
Post-processing steps
Filtering fusions based on the number of reads and mate pairs that support fusions
Sorting fusions based on scores of read distributions around them

Figure 3.1 TopHat-Fusion pipeline
TopHat-Fusion consists of two main modules: (1) finding candidate fusions and aligning reads across them; and (2) filtering out false fusions using a series of post-processing routines.

The algorithm then uses Bowtie to map the 25-bp segments to the genome. For normal transcripts, the TopHat algorithm requires that segments must align in a pattern consistent with introns; that is, the segments may be separated by a user-defined maximum intron length, and they must align in the same orientation along the same chromosome. For fusion transcripts, TopHat-Fusion relaxes both these constraints, allowing it to detect fusions across chromosomes as well as fusions caused by inversions.


IUM read (75bp)
TTAACACTATCTAAAATCAATTTTC
TTTTACAGGTACGGTCAACAGTAAC
AATGATAGCGACGACTGCGTCATAG
segment 1 (25bp)
segment 2 (25bp)
segment 3 (25bp)
TTAACACTATCTAAAATCAATTTTC
AATGATAGCGACGACTGCGTCATAG
chr i
GAATTTCCTG
TTAACACTATCTAAAATCAATTTTC
TTTTACAGGTACATTGTAGTTTTAT
GAATATGGCTCCGGTCAACAGTAAC
AATGATAGCGACGACTGCGTCATAG
TCAGTGAATC
chr j
135223330
135223354
287237735
287237711 (genomic coordinate)
TTTTACAGGTAC
GGTCAACAGTAAC
TTAACACTATCTAAAATCAATTTTC
TTTTACAGGTAC
GGTCAACAGTAAC
AATGATAGCGACGACTGCGTCATAG
chr i
GAATTTCCTG
TTAACACTATCTAAAATCAATTTTC
TTTTACAGGTAC
ATTGTAGTTTTAT
GAATATGGCTCC
GGTCAACAGTAAC
AATGATAGCGACGACTGCGTCATAG
TCAGTGAATC
chr j
135223366
287237748
chr i
GAATTTCCTG
TTAACACTATCTAAAATCAATTTTC
TTTTACAGGTAC
GGTCAACAGTAAC
AATGATAGCGACGACTGCGTCATAG
TCAGTGAATC
chr j
a break point

(a) mapping segments on chr i and chr j

(b) finding a break point between chr i and chr j

Figure 3.2 Aligning a read that spans a fusion point
(a) An initially unmapped read of 75 bp is split into three segments of 25 bp, each of which is mapped separately. As shown here, the left (red) and right (blue) segments are mapped to two different chromosomes, i and j. (b) The unmapped green segment is used to find the precise fusion point between i and j. This is done by aligning the green segment to the sequences just to the right of the red segment on chromosome i and just to the left of the blue segment on chromosome j.

Following the mapping step, we filter out candidate fusion events involving multi-copy genes or other repetitive sequences, on the assumption that these sequences cause mapping artifacts. However, some multi-mapped reads (reads that align to multiple locations) might correspond to genuine fusions: for example, in Kinsella *et al*. [56], the known fusion genes *HOMEZ-MYH6* and *KIAA1267-ARL17A* were supported by 2 and 11 multi-mapped read pairs, respectively. Therefore, instead of eliminating all multi-mapped reads, we impose an upper bound *M* (default M = 2) on the number of mappings per read. If a read or a pair of reads has M or fewer multi-mappings, then all mappings for that read are considered. Reads with >M mappings are discarded.

To further reduce the likelihood of false positives, we require that each read mapping across a fusion point have at least 13 bases matching on both sides of the fusion, with no more than two mismatches. We consider alignments to be fusion candidates when the two 'sides' of the event either (a) reside on different chromosomes or (b) reside on the same chromosome and are separated by at least 100,000 bp. The latter are the results of intra-chromosomal rearrangements or possibly read-through transcription events. We chose the 100,000-bp minimum distance as a compromise that allows TopHat-Fusion to detect intra-chromosomal rearrangements while excluding most but not all read-through transcripts. Intra-chromosomal fusions may also include inversions.

As shown in Figure 3.2a, after splitting an IUM read into three segments, the first and last segments might be mapped to two different chromosomes. Once this pattern of alignment is detected, the algorithm uses the three segments from the IUM read to find the fusion point. After finding the precise location, the segments are re-aligned, moving inward from the left and right boundaries of the original DNA fragment. The resulting mappings are combined together to give full read alignments. For this re-mapping step, TopHat-Fusion extracts 22 bp immediately flanking each fusion point and concatenates them to create 44-bp 'spliced fusion contigs' (Figure

3.3). It then creates a Bowtie index (using the bowtie-build program [48]) from the spliced contigs. Using this index, it runs Bowtie to align all the segments of all IUM reads against the spliced fusion contigs. For a 25-bp segment to be mapped to a 44-bp contig, it has to span the fusion point by at least 3 bp.

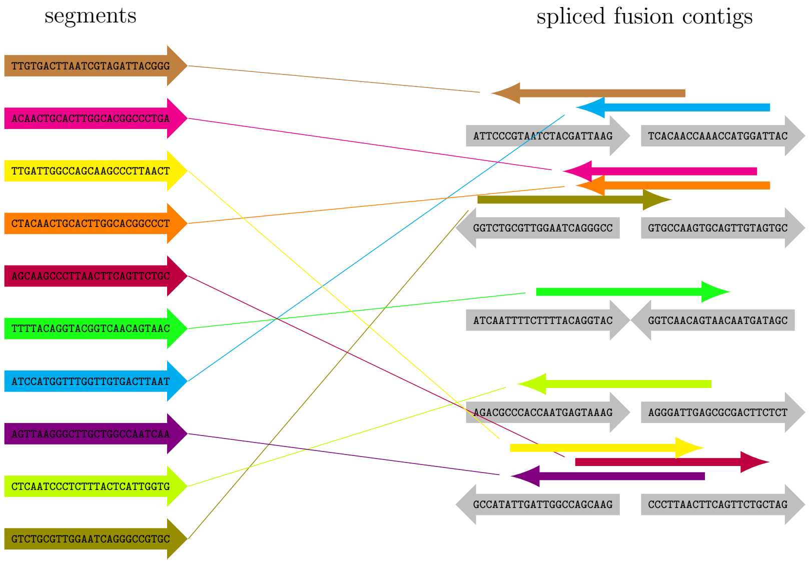
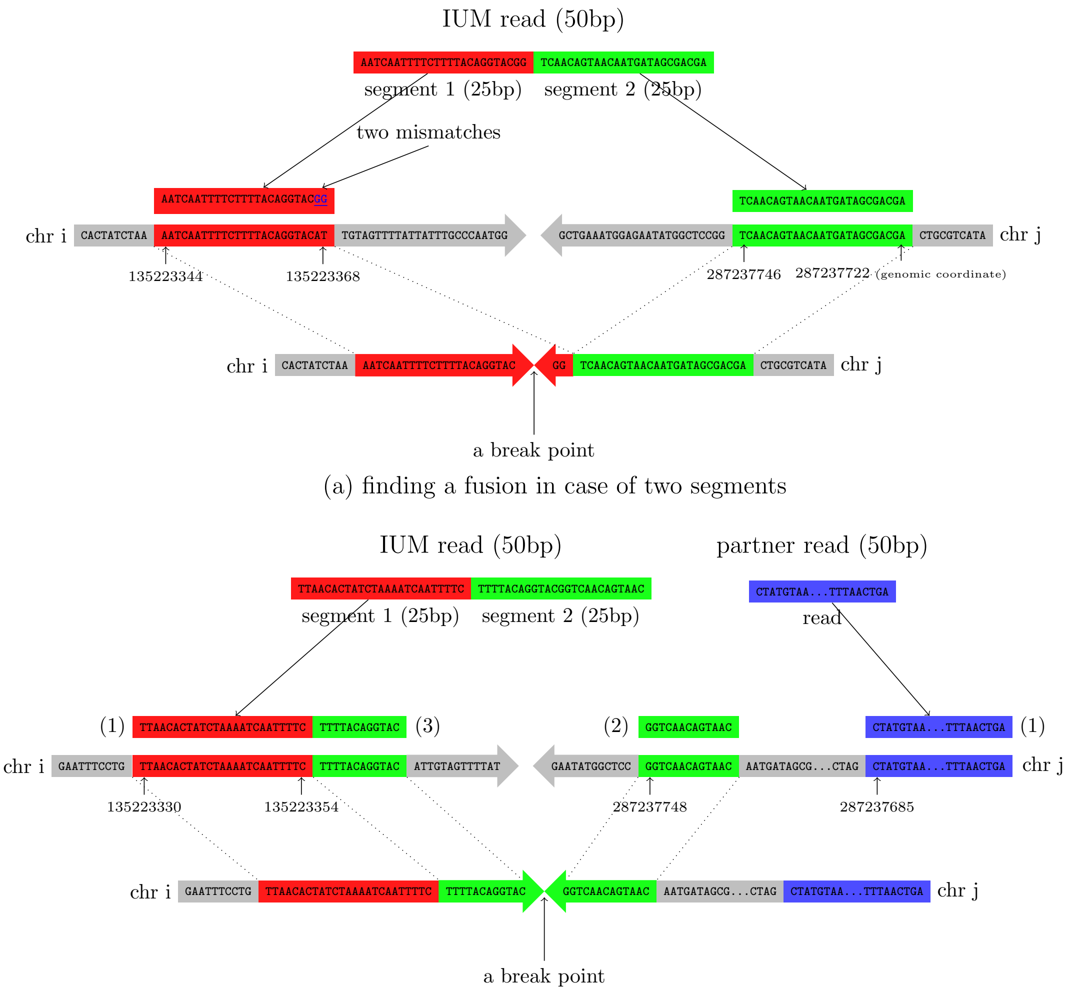

segments
spliced fusion contigs
TTGTGACTTAATCGTAGATTACGGG
ACAACTGCACTTGGCACGGCCCTGA
TTGATTGGCCAGCAAGCCCTTAACT
CTACAACTGCACTTGGCACGGCCCT
AGCAAGCCCTTAACTTCAGTTCTGC
TTTTACAGGTACGGTCAACAGTAAC
ATCCATGGTTTGGTTGTGACTTAAT
AGTTAAGGGCTTGCTGGCCAATCAA
CTCAATCCCTCTTTACTCATTGGTG
GTCTGCGTTGGAATCAGGGCCGTGC
ATTCCCGTAATCTACGATTAAG
TCACAACCAAACCATGGATTAC
GGTCTGCGTTGGAATCAGGGCC
GTGCCAAGTGCAGTTGTAGTGC
ATCAATTTTCTTTTACAGGTAC
GGTCAACAGTAACAATGATAGC
AGACGCCCACCAATGAGTAAAG
AGGGATTGAGCGCGACTTCTCT
GCCATATTGATTGGCCAGCAAG
CCCTTAACTTCAGTTCTGCTAG

Figure 3.3 Mapping against fusion points

Bowtie is used to align all segments from the initially unmapped (IUM) reads against spliced fusion contigs, shown in gray on the right. For example, the brown read on the top left aligns to the first spliced fusion contig on the top right.

In addition to finding fusion points using three (or more) segments as illustrated in Figure 3.2, TopHat-Fusion is able to identify fusions using two segments (the minimum number of segments required), and paired-end alignments are used to make this searching process more sensitive (Figure 3.4). By allowing a few mismatches when TopHat uses Bowtie to map segments from the initially unmapped (IUM) reads, it is possible that a segment will be mapped a few bases past a fusion point. This allows TopHat-Fusion to identify fusions with just two segments by realigning them to two chromosomes, or two different parts of a chromosome. Although this variation on the algorithm is less sensitive than the three (or more) segment approach, which allows middle segments to span a fusion point as well as a few boundary base pairs of the first and third segments, it turns out that this approach is quite effective considering the very deep coverage often available in RNA-seq data sets. As shown in Figure 3.4b, the alignment of a partner read is also used to identify a possible small range in which a fusion point may lie.


IUM read (50bp)
AATCAATTTTCTTTTACAGGTACGG
TCAACAGTAACAATGATAGCGACGA
segment 1 (25bp)
segment 2 (25bp)
two mismatches
AATCAATTTTCTTTTACAGGTACGG
TCAACAGTAACAATGATAGCGACGA
chr i
CACTATCTAA
AATCAATTTTCTTTTACAGGTACAT
TGTAGTTTTATTATTTGCCCAATGG
GCTGAAATGGAGAATATGGCTCCGG
TCAACAGTAACAATGATAGCGACGA
CTGCGTCATA
chr j
135223344
135223368
287237746
287237722
(genomic coordinate)
chr i
CACTATCTAA
AATCAATTTTCTTTTACAGGTAC
GG
TCAACAGTAACAATGATAGCGACGA
CTGCGTCATA
chr j
a break point

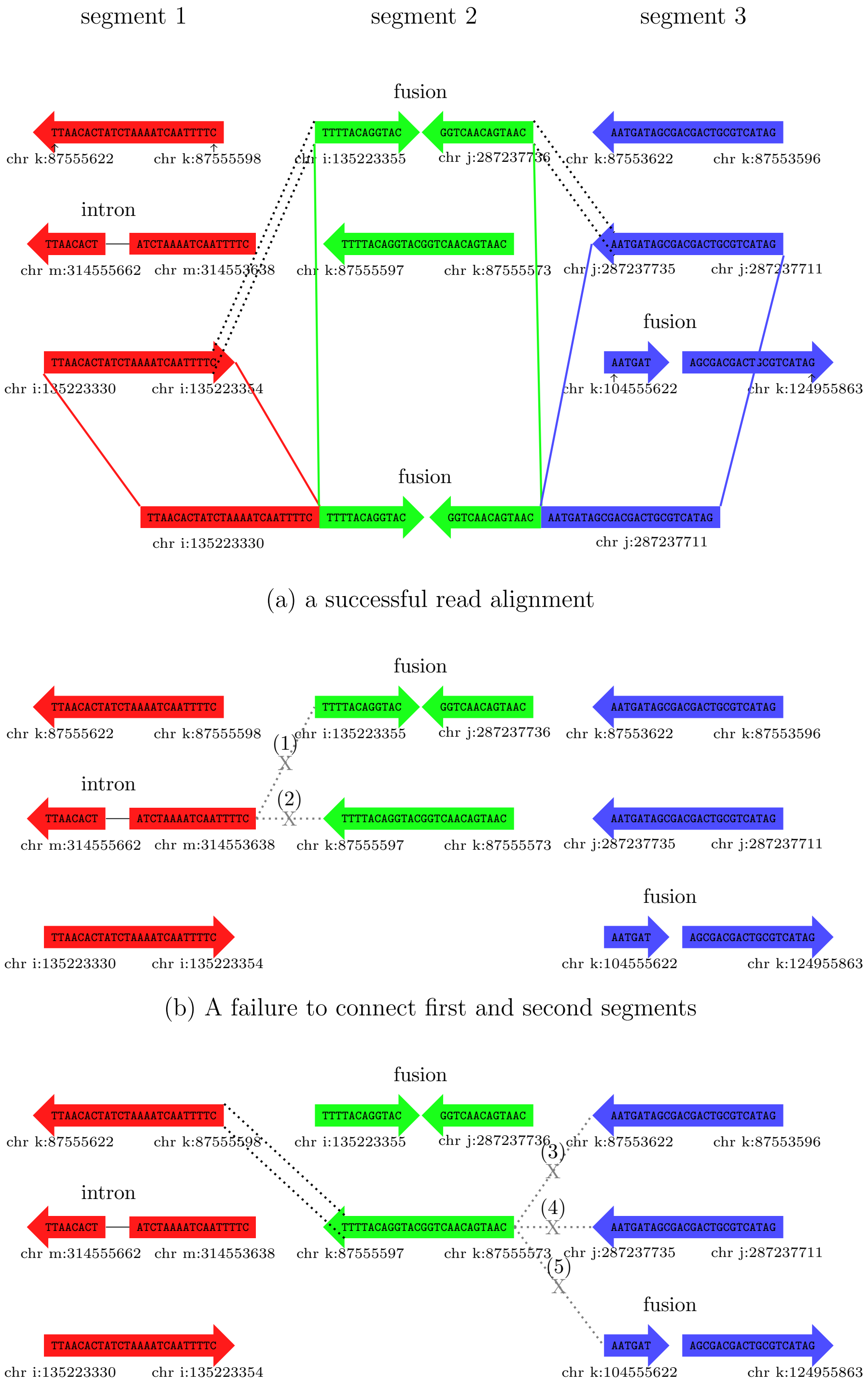
(a) finding a fusion in case of two segments

(b) finding a fusion using paired-end reads

Figure 3.4 Finding fusions using two segments and partner reads
(a) TopHat allows one to three mismatches when mapping segments using Bowtie, which enables segments to be mapped even if a few bases cross a fusion point (the last two bases of the red segment, GG). These two segments, mapped to two different chromosomes, are used to identify a fusion point. (b) For paired-end reads, the mapped position of the partner read is used to narrow down the range of a fusion point. The second segment (shown in green) cannot be mapped because it spans a fusion point. Here, its partner read is mapped and the

fusion point is likely to be located within the inner mate distance ± standard deviation of the left genomic coordinate of the partner read. TopHat-Fusion is able to use this relatively small range to efficiently map the right part of the second segment to the right side of a fusion (case 2). The left part of the second segment is aligned to the right side of the mapped first segment (case 3).

After identifying fusion points in the above step, and mapping segments against such fusions, it is necessary to connect the mapped segments to make a full read alignment, which is one of the most complicated processes in TopHat-Fusion. Given the mappings of the segments comprising a read, TopHat-Fusion stitches them together to produce full-length read alignments according to the following rules (illustrated in Figure 3.5). (1) Two consecutive segments of a read are aligned on the same chromosome with the same orientation, and the right genomic coordinate of a segment corresponds to the left coordinate of its subsequent segment or there is a junction or a deletion to fill the gap between two consecutive subsequences. (2) There is a fusion that connects the segments available. This stitching process is done by depth first search; i.e., given a first segment, TopHat-Fusion examines every second segment to check if any of them can be glued to the first one, and if there is such a second segment, it searches all the third segments. During the search process, an alignment of a segment may be reversed to have the same orientation with its preceding segment.



(a) a successful read alignment

(b) A failure to connect first and second segments

(c) A failure to connect second and third segments

Figure 3.5 Stitching segments to produce a full read alignment

(a) The segment in the third row for segment 1 and the one in the first row for segment 2 are connected because they are on the same chromosome (i) in the forward direction and with adjacent coordinates. These are then matched to the second row in segment 3 and glued together, producing the full-length read alignment at the bottom. (b) TopHat-Fusion tries to connect the segment in the second row for segment 1 with segments in the first and second rows for segment 2, but neither succeeds. Case 1 would require two fusion points in the same read, and case 2 cannot be fused with consistent coordinates. (c) Attempts to connect the segment in the second row for segment 2 with the one in the first row in segment 3: in case 3, there is no intron available, there is no fusion in case 4, and case 5 would require more than one fusion.

After stitching together the segment mappings to produce full alignments, we collect those reads that have at least one alignment spanning the entire read. We then choose the best alignment for each read using a heuristic scoring function, defined below. We assign penalties for alignments that span introns (-2), indels (-4), or fusions (-4). For each potential fusion, we require that spanning reads have at least 13 bp aligned on both sides of the fusion point (this requirement alone eliminates many false positives). After applying the penalties, if a read has more than one alignment with the same minimum penalty score, then the read with the fewest mismatches is selected. For example, in Figure 3.6, IUM read 1 (in blue) is aligned to three different locations: (1) chromosome *i* with no gap, (2) chromosome *j* where it spans an intron, and (3) a fusion contig formed between chromosome *m* and chromosome *n*. Our scoring function prefers (1), followed by (2), and by (3). For IUM read 2 (Figure 3.6, in green), we have two alignments: (1) a fusion formed between chromosome *i* and chromosome *j*, and (2) an alignment to chromosome *k* with a small deletion. These two alignments both incur the same penalty, but we select (1) because it has fewer mismatches.

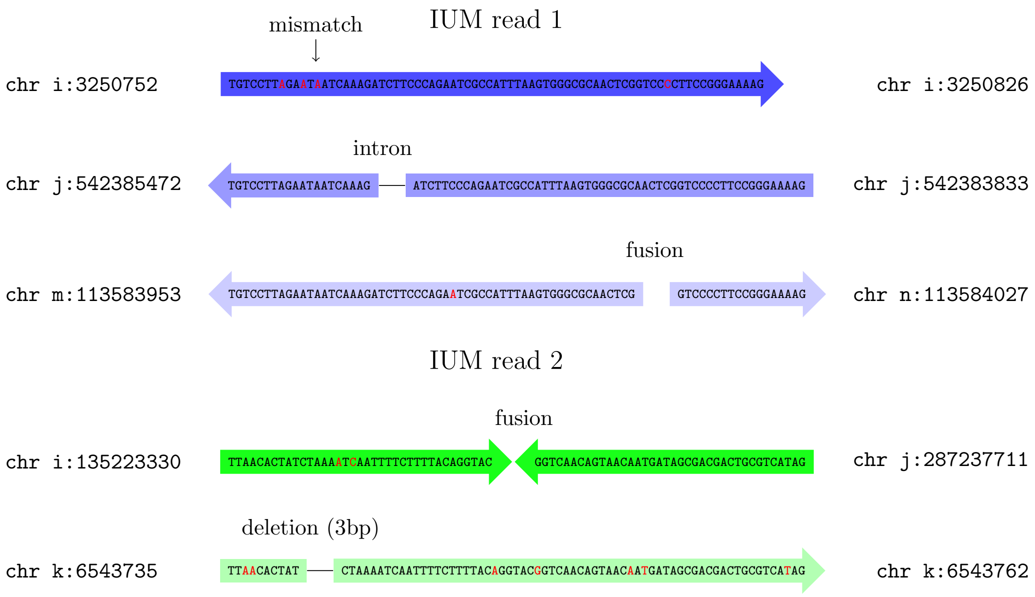
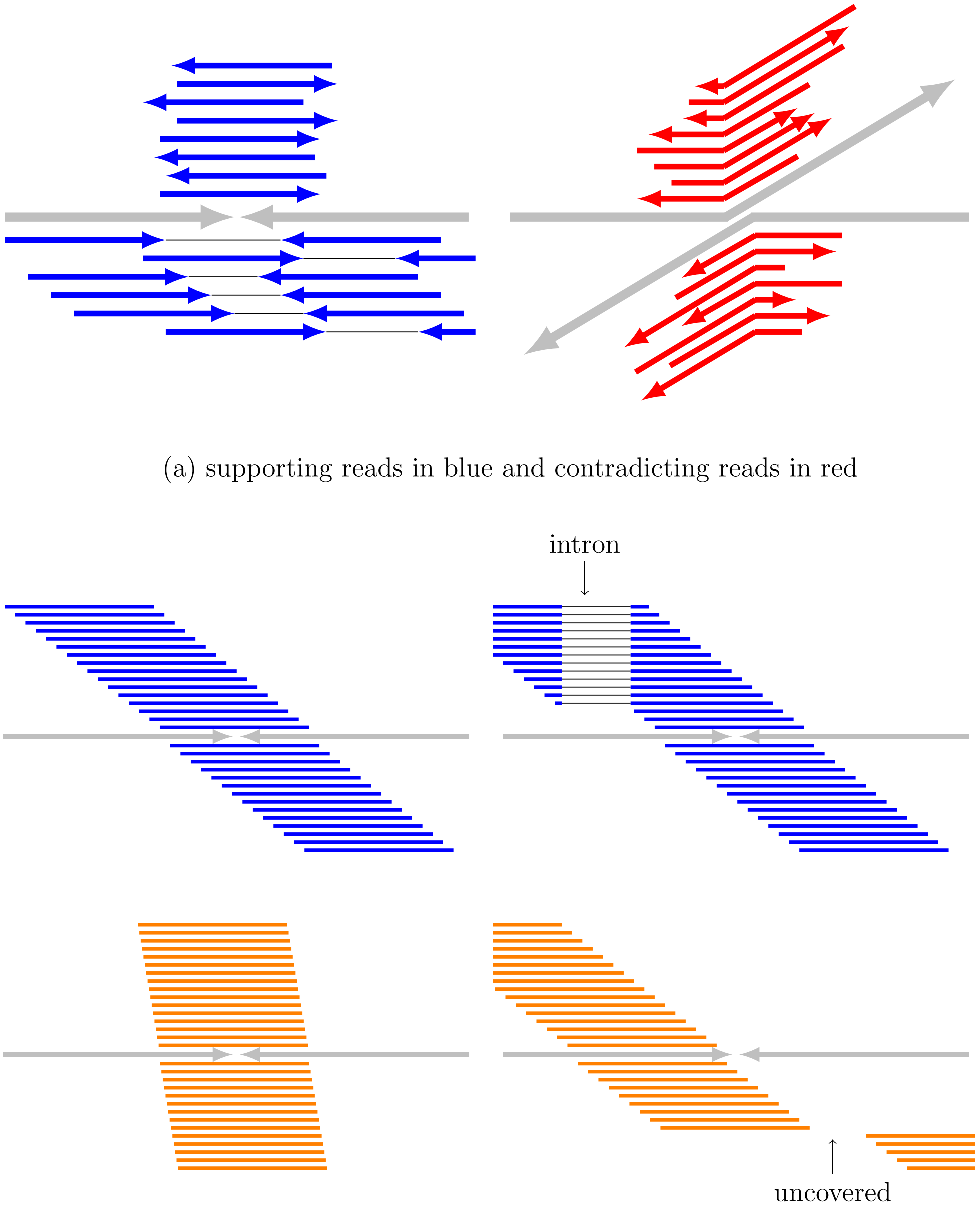

IUM read 1
mismatch
chr i:3250752
TGTCCTTAGAATAATCAAAGATCTTCCCAGAATCGCCATTTAAGTGGGCGCAACTCGGTCCCCTTCCGGGAAAAG
chr i:3250826
intron
chr j:542385472
TGTCCTTAGAATAATCAAAG
ATCTTCCCAGAATCGCCATTTAAGTGGGCGCAACTCGGTCCCCTTCCGGGAAAAG
chr j:542383833
fusion
chr m:113583953
TGTCCTTAGAATAATCAAAGATCTTCCCAGAATCGCCATTTAAGTGGGCGCAACTCG
GTCCCCTTCCGGGAAAAG
chr n:113584027
IUM read 2
fusion
chr i:135223330
TTAACACTATCTAAAATCAATTTTCTTTTACAGGTAC
GGTCAACAGTAACAATGATAGCGACGACTGCGTCATAG
chr j:287237711
deletion (3bp)
chr k:6543735
TTAACACTAT
CTAAAATCAATTTTCTTTTACAGGTACGGTCAACAGTAACAATGATAGCGACGACTGCGTCATAG
chr k:6543762

Figure 3.6 Selecting best read alignments
IUM reads 1 and 2 each have multiple alignments. Read 1 has a gap-free alignment, shown in dark blue, which is preferred over the other two alignments shown in lighter shades of blue. The gap-free alignment with three mismatches is preferred over the fusion alignment with one mismatch. If all alignments have gaps and mismatches, then the algorithm prefers those with fewer mismatches, as shown by the dark green alignment for IUM read 2. Full details of the scoring function that determines these preferences are described in the Materials and methods.

We imposed further filters for each data set: (1) in the breast cancer cell lines (BT474, SKBR3, KPL4, MCF7), we required two supporting pairs and the sum of spanning reads and supporting pairs to be at least 5; (2) in the VCaP paired-end reads, we required the sum of spanning reads and supporting pairs to be at least 10; (3) in the UHR paired-end reads, we required (i) three spanning reads and two supporting pairs or (ii) the sum of spanning reads and supporting pairs to be at least 10; and (4) in the UHR single-end reads, we required two spanning reads. These numbers were determined empirically using known fusions as a quality control. All candidates that fail to satisfy these filters were eliminated.

In order to remove false positive fusions caused by repeats, we extract the two 23-base sequences spanning each fusion point and then map them against the entire human genome. We convert the resulting alignments into a list of pairs (chromosome name, genomic coordinate - for example, chr14:374384). For each 23-mer adjacent to a fusion point, we test to determine if the other 23-mer occurs within 100,000 bp on the same chromosome. If so, then it is likely a repeat and we eliminate the fusion candidate. We further require that at least one side of a fusion contains an annotated gene (based on known genes from RefSeq), otherwise the fusion is filtered out. These steps alone reduced the number of fusion candidates in our experiments from 10^5 to just a few hundred.



(a) supporting reads in blue and contradicting reads in red

(b) read distribution around a fusion

Figure 3.7 Supporting and contradicting evidence for fusion transcripts
(a) Given a fusion point and the chromosomes (gray) spanning it, single-end and paired-end reads (blue) support the fusion. Other reads (red) contradict the fusion by mapping entirely to either of the two chromosomes. (b) TopHat-Fusion prefers reads that uniformly cover a 600-bp window centered in any fusion point. On the upper left, blue reads cover the entire window. On the lower left, red reads cover only a narrow window around the fusion. On the lower right, reads do not cover part of the 600-bp window. The cases shown in orange will be rejected by TopHat-Fusion.

As reported in Edgren *et al*. [18], true fusion transcripts have reads mapping uniformly in a wide window across the fusion point, whereas false positive fusions are narrowly covered. Using this idea, TopHat-Fusion examines a 600-bp window around each fusion (300-bp each side), and rejects fusion candidates for which the reads fail to cover this window (Figure 3.7b). The final process is to sort fusions based on how well-distributed the reads are (Figure 3.8). The scoring scheme prefers alignments that have no gaps (or small gaps) and uniform depth.

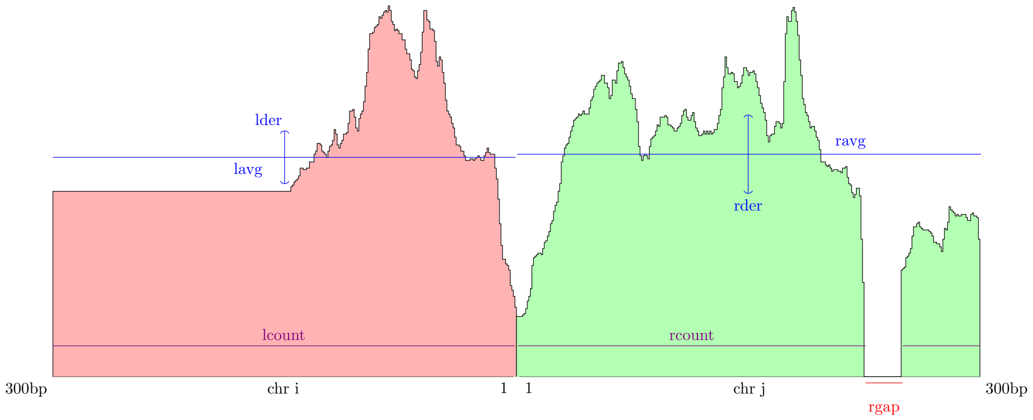


Figure 3.8 TopHat-Fusion's scoring scheme of read distributions
A scoring scheme of how well distributed reads are around a fusion point; these result scores are used to sort the list of candidate fusions. Variables are defined in the main text.

Even with strict parameters for the initial alignment, many of the segments will map to multiple locations, which can make it appear that a read spans two chromosomes. Thus the algorithm may find large numbers of false positives, primarily due to the presence of millions of repetitive sequences in the human genome. Even after filtering to choose the best alignment per read, the experiments reported here yielded initial sets of about 400,000 and 135,000 fusion gene candidates from the breast cancer (BT474, SKBR3, KPL4, MCF7) and prostate cancer (VCaP) cell lines, respectively. The additional filtering steps eliminated the vast majority of these false positives, reducing the output to 76 and 19 fusion candidates, respectively, all of which have strong supporting evidence (Tables 3.2 and 3.3).

The scoring function used to rank fusion candidates uses the number of paired reads in which the reads map on either side of the fusion point in a consistent orientation (Figure 3.7a) as well as the number of reads in conflict with the fusion point. Conflicting reads align entirely to either of the two chromosomes and span the point at which the chromosome break should occur (Figure 3.7b).

The overall fusion score is computed as:

$$\begin{aligned} score = lcount + rcount + \min(max\_avg, lavg) + \min(max\_avg, ravg) \\ - |lcount - rcount| - \min(max\_avg, |lavg - ravg|) \\ - (lgap + rgap) - (lder + rder) * max\ \_avg + rate \\ - \min(1000, dist) \end{aligned}$$

where lcount is the number of bases covered in a 300-bp window on the left (Figure 3.8), lavg is the average read coverage on the left, max_avg is 300, lgap is the length of any gap on the left, rate is the ratio between the number of supporting mate pairs and the number of contradicting reads, |lavg - ravg| is a penalty for expression differences on either side of the fusion, and dist is the sum of distances between each end of a pair and a fusion. For single-end reads, the rate uses spanning reads rather than mate pairs. The variance in coverage lder is:

$lder = \sqrt{\sum_{n=1}^{lwindow} \frac{\left(\frac{lavg - ldepth_n}{lavg}\right)^2}{lwindow}}$, where lwindow is the size of the left window (300 bp).

TopHat-Fusion outputs alignments of singleton reads and paired-end reads mapped across fusion points in SAM format [35], enabling further downstream analyses [57], such as transcript assembly and differential gene expression. The parameters in the filtering steps can be changed as needed for a particular data set.

3.3 Results

We tested TopHat-Fusion on RNA-seq data from two recent studies of fusion genes: (1) four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7) described by Edgren *et al*. [18] and available from the NCBI Sequence Read Archive [SRA:SRP003186]; and (2) the VCaP prostate cancer cell line and the Universal Human Reference (UHR) cell line, both from Maher *et al*. [50]. The data sets contained >240 million reads, including both paired-end and single-end reads (Table 3.1). We mapped all reads to the human genome (UCSC hg19) with TopHat-Fusion, and we identified the genes involved in each fusion using the RefSeq and Ensembl human annotations.

Data source	Sample ID	Read Type	Fragment length	Read length	Number of fragments (or reads)
Edgren *et al.* [18]	BT474	Paired	100, 200	50	21,423,697
Edgren *et al.* [18]	SKBR3	Paired	100, 200	50	18,140,246
Edgren *et al.* [18]	KPL4	Paired	100	50	6,796,443
Edgren *et al.* [18]	MCF7	Paired	100	50	8,409,785
Maher *et al.* [50]	VCaP	Paired	300	50	16,894,522
Maher *et al.* [50]	UHR	Paired	300	50	25,294,164
Maher *et al.* [50]	UHR	Single		100	56,129,471

Table 3.1 RNA-seq data used to test TopHat-Fusion
The data came from two studies, and included four samples from breast cancer cells (BT474, SKBR3, KPL4, MCF7), one prostate cancer cell line (VCaP), and two samples from the Universal Human Reference (UHR) cell line. For paired-end data, two reads were generated from each fragment; thus the total number of reads is twice the number of fragments.

One of the biggest computational challenges in finding fusion gene products is the huge number of false positives that result from a straightforward alignment procedure. This is caused by the numerous repetitive sequences in the genome, which allow many reads to align to multiple locations on the genome. To address this problem, we developed strict filtering routines to eliminate the vast majority of spurious alignments (see Materials and methods). These filters allowed us to reduce the number of fusions reported by the algorithm from >100,000 to just a few dozen, all of which had strong support from multiple reads.

Overall, TopHat-Fusion found 76 fusion genes in the four breast cancer cell lines (Table 3.2; the TopHat-Fusion paper [19], additional file 1) and 19 in the prostate cancer (VCaP) cell line (Table 3.3; the TopHat-Fusion paper [19], additional file 2). In the breast cancer data, TopHat-Fusion found 25 out of the 27 previously reported fusions [18]. Of the two fusions TopHat-Fusion missed (DHX35-ITCH, NFS1-PREX1), DHX35-ITCH was included in the initial output, but was filtered out because it was supported by only one singleton read and one mate pair. The remaining 51 fusion genes were not previously reported. In the VCaP data, TopHat-Fusion found 9 of the 11 fusions reported previously [50] plus 10 novel fusions. One of the missing fusions involved two overlapping genes, ZNF577 and ZNF649 on chromosome 19, which appears to be read-through transcription rather than a true gene fusion.

SAMPLE ID	Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads	Spanning pairs
BT474	TRPC4AP-MRPL45	20-17	33665850	36476499	2	9
BT474	TOB1-SYNRG	17-17	48943418	35880750	26	47
SKBR3	**TATDN1-GSDMB**	8-17	125551264	38066175	311	555
BT474	THRA-SKAP1	17-17	38243102	46384689	28	46
MCF7	**BCAS4-BCAS3**	20-17	49411707	59445685	105	284
BT474	**ACACA-STAC2**	17-17	35479452	37374425	57	59
BT474	STX16-RAE1	20-20	57227142	55929087	6	24
BT474	MED1-ACSF2	17-17	37595419	48548386	10	12
MCF7	ENSG00000254868-FOXA1	14-14	38184710	38061534	2	22
SKBR3	**ANKHD1-PCDH1**	5-5	139825557	141234002	4	15
BT474	**ZMYND8-CEP250**	20-20	45852972	34078459	10	53
BT474	AHCTF1-NAAA	1-4	247094879	76846963	10	42

SKBR3	**SUMF1-LRRFIP2**	3-3	4418012	37170638	3	12
KPL4	**BSG-NFIX**	19-19	580779	13135832	12	27
BT474	**VAPB-IKZF3**	20-17	56964574	37922743	4	14
BT474	DLG2-HFM1	11-1	85195025	91853144	2	10
SKBR3	**CSE1L-ENSG00000236127**	20-20	47688988	47956855	13	31
MCF7	RSBN1-AP4B1	1-1	114354329	114442495	6	7
BT474	MED13-BCAS3	17-17	60129899	59469335	3	14
MCF7	**ARFGEF2-SULF2**	20-20	47538545	46365686	17	20
BT474	HFM1-ENSG00000225630	1-1	91853144	565937	2	43
KPL4	MUC20-ENSG00000249796	3-3	195456606	195352198	13	46
KPL4	MUC20-ENSG00000236833	3-3	195456612	197391649	8	15
MCF7	**RPS6KB1-TMEM49**	17-17	57992061	57917126	4	3
SKBR3	**WDR67-ZNF704**	8-8	124096577	81733851	3	3
BT474	**CPNE1-PI3**	20-20	34243123	43804501	2	6
BT474	ENSG00000229344-RYR2	1-1	568361	237766339	1	19
BT474	**LAMP1-MCF2L**	13-13	113951808	113718616	2	6
MCF7	SULF2-ZNF217	20-20	46415146	52210647	11	32
BT474	WBSCR17-FBXL20	7-17	70958325	37557612	2	8
MCF7	ENSG00000224738-TMEM49	17-17	57184949	57915653	5	6
MCF7	ANKRD30BL-RPS23	2-5	133012791	81574161	2	6
BT474	ENSG00000251948-SLCO5A1	19-8	24184149	70602608	2	6
BT474	**GLB1-CMTM7**	3-3	33055545	32483333	2	6
KPL4	EEF1DP3-FRY	13-13	32520314	32652967	2	4
MCF7	PAPOLA-AK7	14-14	96968936	96904171	3	3
BT474	ZNF185-GABRA3	X-X	152114004	151468336	2	3
KPL4	**PPP1R12A-SEPT10**	12-2	80211173	110343414	3	8
BT474	**SKA2-MYO19**	17-17	57232490	34863349	5	12
MCF7	LRP1B-PLXDC1	2-17	142237963	37265642	2	5
BT474	NDUFB8-TUBD1	10-17	102289117	57962592	1	49
BT474	ENSG00000225630-NOTCH2NL	1-1	565870	145277319	1	18
SKBR3	**CYTH1-EIF3H**	17-8	76778283	117768257	18	37
BT474	PSMD3-ENSG00000237973	17-1	38151673	566925	1	12
BT474	**STARD3-DOK5**	17-20	37793479	53259992	2	10
BT474	**DIDO1-TTI1**	20-20	61569147	36634798	1	10
BT474	**RAB22A-MYO9B**	20-19	56886176	17256205	8	20
KPL4	PCBD2-ENSG00000240967	5-5	134259840	99382129	1	32
SKBR3	**RARA-PKIA**	17-8	38465535	79510590	1	5
BT474	MED1-STXBP4	17-17	37607288	53218672	13	11
KPL4	C1orf151-ENSG00000224237	1-3	19923605	27256479	1	5
SKBR3	RNF6-FOXO1	13-13	26795971	41192773	2	13
SKBR3	BAT1-ENSG00000254406	6-11	31499072	119692419	2	30
BT474	KIAA0825-PCBD2	5-5	93904985	134259811	1	19
SKBR3	PCBD2-ANKRD30BL	5-2	134263179	133012790	1	5
BT474	ENSG00000225630-MTRNR2L8	1-11	565457	10530147	1	35
BT474	PCBD2-ENSG00000251948	5-19	134260431	24184146	2	6
BT474	ANKRD30BL-ENSG00000237973	2-1	133012085	567103	2	8
KPL4	ENSG00000225972-HSP90AB1	1-6	564639	44220780	1	7
BT474	MTIF2-ENSG00000228826	2-1	55470625	121244943	1	11
BT474	ENSG00000224905-PCBD2	21-5	15457432	134263223	2	7
BT474	**RPS6KB1-SNF8**	17-17	57970686	47021335	48	57
BT474	MTRNR2L8-PCBD2	11-5	10530146	134263156	1	6
BT474	RPL23-ENSG00000225630	17-1	37009355	565697	3	19
BT474	MTRNR2L2-PCBD2	5-5	79946288	134259832	1	5
SKBR3	ENSG00000240409-PCBD2	1-5	569005	134260124	2	4

SKBR3	PCBD2-ENSG00000239776	5-12	134263289	127650986	2	3
BT474	ENSG00000239776-MTRNR2L2	12-5	127650981	79946277	2	3
BT474	JAK2-TCF3	9-19	5112849	1610500	1	46
KPL4	**NOTCH1-NUP214**	9-9	139438475	134062675	3	5
BT474	MTRNR2L8-TRBV25OR92	11-9	10530594	33657801	4	4
BT474	MTRNR2L8-AKAP6	11-14	10530179	32953468	1	5
BT474	ENSG00000230916-PCBD2	X-5	125606246	134263219	1	5
MCF7	ENSG00000226505-MRPL36	2-5	70329650	1799907	5	20
SKBR3	**CCDC85C-SETD3**	14-14	100002351	99880270	5	6
BT474	RPL23-ENSG00000230406	17-2	37009955	222457168	109	5

Table 3.2 76 candidate fusions in breast cancer samples

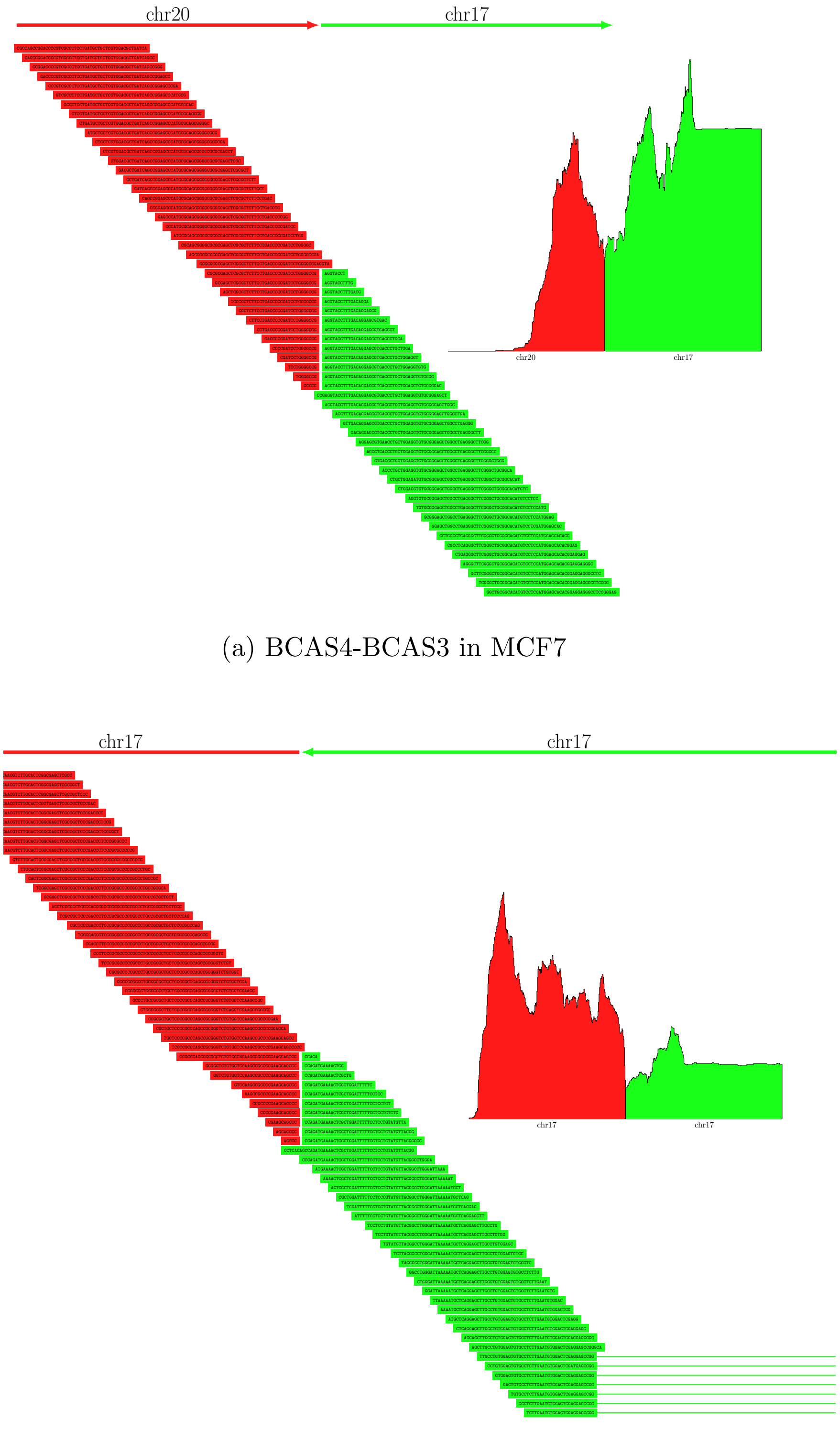
The 76 candidate fusion genes found by TopHat-Fusion in four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7), with previously reported fusions [18] shown in boldface. The remaining 51 fusion genes are novel. The fusions are sorted by the scoring scheme described in Methods.

Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads	Spanning pairs
ZDHHC7-ABCB9	16-12	85023908	123444867	13	69
TMPRSS2-ERG	21-21	42879875	39817542	7	285
HJURP-EIF4E2	2-2	234749254	233421125	3	9
VWA2-PRKCH	10-14	116008521	61909826	1	10
RGS3-PRKAR1B	9-7	116299195	699055	3	11
SPOCK1-TBC1D9B	5-5	136397966	179305324	9	31
LRP4-FBXL20	11-17	46911864	37557613	5	9
INPP4A-HJURP	2-2	99193605	234746297	6	12
C16orf70-C16orf48	16-16	67144140	67700168	2	19
NDUFV2-ENSG00000188699	18-19	9102729	53727808	1	35
NEAT1-ENSG00000229344	11-1	65190281	568419	1	17
ENSG00000011405-TEAD1	11-11	17229396	12883794	7	9
USP10-ZDHHC7	16-16	84733713	85024243	1	22
LMAN2-AP3S1	5-5	176778452	115202366	15	2
WDR45L-ENSG00000224737	17-17	80579516	30439195	1	33
RC3H2-RGS3	9-9	125622198	116299072	3	11
CTNNA1-ENSG00000249026	5-5	138145895	114727795	1	12
IMMTP1-IMMT	21-2	46097128	86389185	1	50
ENSG00000214009-PCNA	X-20	45918367	5098168	1	24

Table 3.3 19 candidate fusions in prostate cancer samples

19 candidate fusions found by TopHat-Fusion in the VCaP prostate cell line, with previously reported fusions [50] indicated in boldface. Fusion genes are sorted according to the scoring scheme described in Methods.

Figure 3.9 illustrates two of the fusion genes identified by TopHat-Fusion. Figure 3.9a shows the reads spanning a fusion between the *BCAS3* (breast carcinoma amplified sequence 3) gene on chromosome 17 (17q23) and the *BCAS4* gene on chromosome 20 (20q13), originally found in the MCF7 cell line in 2002 [58]. As illustrated in the figure, many reads clearly span the boundary of the fusion between chromosomes 20 and 17, illustrating the single-base precision enabled by TopHat-Fusion. Figure 3.9b shows a novel intra-chromosomal fusion product with similarly strong alignment evidence that TopHat-Fusion found in BT474 cells. This fusion merges two genes that are 13 megabases apart on chromosome 17: *TOB1* (transducer of ERBB2, ENSG00000141232) at approximately 48.9 Mb; and *SYNRG* (synergin gamma) at approximately 35.9 Mb.



(a) BCAS4-BCAS3 in MCF7

(b) TOB1-SYNRG in BT474

Figure 3.9 Read distributions around two fusions

(a) 60 reads aligned by TopHat-Fusion that identify a fusion product formed by the BCAS4 gene on chromosome 20 and the BCAS3 gene on chromosome 17. The data contained more reads than shown; they are collapsed to illustrate how well they are distributed. The inset figures show the coverage depth in 600-bp windows around each fusion. (b) TOB1(ENSG00000141232)-SYNRG is a novel fusion gene found by TopHat-Fusion, shown here with 70 reads mapping across the fusion point. Note that some of the reads in green span an intron (indicated by thin horizontal lines extending to the right), a feature that can be detected by TopHat's spliced alignment procedure.

Single versus paired-end reads

Using four known fusion genes (*GAS6-RASA3*, *BCR-ABL1*, *ARFGEF2-SULF2*, and *BCAS4-BCAS3*), we compared TopHat-Fusion's results using single and paired-end reads from the UHR data set (Table 3.4). All four fusions were detected using either type of input data. Although Maher *et al*. [50] reported much greater sensitivity using paired reads, we found that the ability to detect fusions using single-end reads, when used with TopHat-Fusion, was sometimes nearly as good as with paired reads. For example, the reads aligning to the *BCR-ABL1* fusion provided similar support using either single or paired-end data (the TopHat-Fusion paper [19], additional file 3). Among the top 20 fusion genes in the UHR data, 3 had more support from single-end reads and 9 had better support from paired-end reads (the TopHat-Fusion paper [19], additional file 4). Note that longer reads might be more effective for detecting gene fusions from unpaired reads: Zhao *et al*. [59] found 4 inter-chromosomal and 3 intra-chromosomal fusions in a breast cancer cell line (HCC1954), using 510,703 relatively long reads (average 254 bp) sequenced using 454 pyrosequencing technology. Very recently, the FusionMap system [60] was reported to achieve better results, using simulated 75-bp reads, on single-end versus paired-end reads when the inner mate distance is short.

Read type	Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads (RPM)	Spanning pairs
Single	GAS6-RASA3	13-13	114529968	114751268	15 (0.267)	
Paired	GAS6-RASA3	13-13	114529968	114751268	10 (0.198)	43
Single	BCR-ABL1	22-9	23632599	133655755	6 (0.107)	
Single	BCR-ABL1	22-9	23632599	133729450	3 (0.053)	
Paired	BCR-ABL1	22-9	23632599	133655755	2 (0.040)	7
Paired	BCR-ABL1	22-9	23632599	133729450	3 (0.059)	10
Single	ARFGEF2-SULF2	20-20	47538548	46365683	17 (0.302)	
Paired	ARFGEF2-SULF2	20-20	47538545	46365686	10 (0.198)	30
Single	BCAS4-BCAS3	20-17	49411707	59445685	25 (0.445)	
Paired	BCAS4-BCAS3	20-17	49411707	59445685	13 (0.257)	145

Table 3.4 Comparisons: single-end and paired-end reads for finding fusions. Comparisons of single-end and paired-end reads as evidence for gene fusions in the Universal Human Reference (UHR) cell line (a mixture of multiple cancer cell lines), using the known fusions GAS6-RASA3, BCR-ABL1, ARFGEF2-SULF2, and BCAS4-BCAS3. With TopHat-Fusion's ability to align a read across a fusion, the single-end approach is competitive with the paired-end based approach. RPM is the number of reads that span a fusion per millon reads sequenced. For instance, the RPM of single-end reads in GAS6-RASA3 is 0.267, which is slightly better than the RPM for paired-end reads. Single-end reads may show higher RPM values than paired-ends in part because single-end reads are longer (100 bp) than paired-end reads (50 bp) in these data, and therefore they are more likely to span fusions.

Estimate of the false positive rate

In order to estimate the false positive rate of TopHat-Fusion, we ran it on RNA-seq data from normal human tissue, in which fusion transcripts should be absent. Using paired-end RNA-seq reads from two tissue samples (testes and thyroid) from the Illumina Body Map 2.0 data [ENA: ERP000546] (see [45] for the download web page), the system reported just one and nine fusion transcripts in the two samples, respectively. Considering that each sample comprised approximately 163 million reads, and assuming that all reported fusions are false positives, the false positive rate would be approximately 1 per 32 million reads. Some of the reported fusions may in fact be chimeric sequences due to ligation of cDNA fragments [61], which would make the false positive rate even lower. For this experiment, we required five spanning reads and five supporting mate pairs because the number of reads is much higher than those of our other test samples. When the filtering parameters are changed to one read and two mate pairs, TopHat-Fusion predicts 4 and 43 fusion transcripts in the two samples, respectively (the TopHat-Fusion paper [19], additional file 5).

Because it is also a standalone fusion detection system, we ran FusionSeq (0.7.0) [33] on one of our data sets to compare its performance to TopHat-Fusion. FusionSeq consists of two main steps: (1) identifying potential fusions based on paired-end mappings; and (2) filtering out fusions with a sophisticated filtration cascade containing more than ten filters. Using the breast cancer cell line MCF7, in which three true fusions (*BCAS4-BCAS3*, *ARFGEF2-SULF2*, *RPS6KB1-TMEM49*) were previously reported, we ran FusionSeq with mappings from Bowtie that included discordantly mapped mate pairs. Note that FusionSeq was designed to use the commercial ELAND aligner, but we used the open-source Bowtie instead. To do this, we aligned each end of every mate pair separately, allowing them to be aligned to at most two places, and then combined and converted them to the input format required by FusionSeq.

When we required at least two supporting mate pairs for a fusion (the same requirement as for our TopHat-Fusion analysis), FusionSeq missed one true fusion (*RPS6KB1-TMEM49*) because it was supported by only one mate pair. In contrast, TopHat-Fusion found this fusion because it was supported by three mate pairs from TopHat-Fusion's alignment algorithm: one mate pair contains a read that spans a splice junction, and the other contains a read that spans a fusion point. These spliced alignments are not found by Bowtie or ELAND. With this spliced mapping capability, TopHat-Fusion will be expected to have higher sensitivity than those based on non-gapped aligners. When the minimum number of mate pairs is reduced to 1, FusionSeq found all three known fusions at the expense of increased running time (9 hours versus just over 2 hours) and a large increase in the number of candidate fusions reported (32,646 versus 5,649).

Next, we ran all of FusionSeq's filters except two (PCR filter and annotation consistency filter) that would otherwise eliminate two of the true fusions. FusionSeq reported 14,510 gene fusions (the TopHat-Fusion paper [19], additional file 6), compared to just 14 fusions reported by TopHat-Fusion (the TopHat-Fusion paper [19], additional file 7), where both found the three known fusions.

Among those fusions reported by FusionSeq, 13,631 and 276 were classified as inter-chromosomal and intra-chromosomal, respectively. When we used all of FusionSeq's filters, it reported 763 candidate fusions that include only one of the three known fusions.

FusionSeq reports three scores for each transcript: SPER (normalized number of inter-transcript paired-end reads), DASPER (difference between observed and expected SPER), and RESPER (ratio of observed SPER to the average of all SPERs). Because RESPER is proportional to SPER in the same data, we used SPER and DASPER to control the number of fusion candidates: *ARFGEF2-SULF2* (SPER, 1.289452; DASPER, 1.279144), *BCAS4-BCAS3* (0.483544, 0.482379), and *RPS6KB1-TMEM49* (0.161181, 0.133692). First, we used SPER of 0.161181 and DASPER of 0.133692 to find the minimum set of fusion candidates that include the three known gene fusions. This reduced the number of candidates from 14,510 to

11,774. Second, we used the SPER and DASPER values from *ARFGEF2*-*SULF2* and *BCAS4*-*BCAS3*, which resulted in 1,269 and 512 predicted fusions, respectively.

We next compared TopHat-Fusion with deFuse (0.4.2) [34]. deFuse maps read pairs against the genome and against cDNA sequences using Bo wtie, and then uses discordantly mapped mate pairs to find candidate regions where fusion break points may lie. This allows detection of break points at base-pair resolution, similar to TopHat-Fusion. After collecting sequences around fusion points, it maps them against the genome, cDNAs, and expressed sequence tags using BLAT; this step dominates the run time.

Using two data sets - MCF7 and SKBR3 - we ran both TopHat-Fusion and deFuse using the following matched parameters: one minimum spanning read, two supporting mate pairs, and 13 bp as the anchor length. For the MCF7 cell line, both programs found the three known fusion transcripts. For the SKBR3 cell line, both programs found the same seven fusions out of nine previously reported fusion transcripts (one known fusion, *CSE1L*-ENSG00000236127, was not considered because ENSG00000236127 has been removed from the recent Ensembl database). Both programs missed two fusion transcripts: *DHX35*-*ITCH* and *NFS1*-*PREX1*. However, TopHat-Fusion had far fewer false positives: it predicted 42 fusions in total, while deFuse predicted 1,670 (the TopHat-Fusion paper [19], additional files 7, 8 and 9).

Table 3.5 shows the number of spanning reads and supporting pairs detected by TopHat-Fusion and deFuse, respectively, for ten known fusions in SKBR3 and MCF7. The numbers are similar in both programs for the known fusion transcripts.

Considering the fact TopHat-Fusion's mapping step does not use annotations while deFuse does, this result illustrates that TopHat-Fusion can be highly sensitive without relying on annotations. Finally, we noted that TopHat-Fusion was approximately three times faster: for the SKBR3 cell line, it took 7 hours, while deFuse took 22 hours, both using the same eight-core computer.

SAMPLE ID	Fusion genes (left-right)	Chromosomes (left-right)	TopHat-Fusion		deFuse	
			Spanning reads	Spanning pairs	Spanning reads	Spanning pairs
SKBR3	TATDN1-GSDMB	8-17	311	555	322	95
SKBR3	RARA-PKIA	17-8	1	5	1	4
SKBR3	ANKHD1-PCDH1	5-5	4	15	5	11
SKBR3	CCDC85C-SETD3	14-14	5	6	6	3
SKBR3	SUMF1-LRRFIP2	3-3	3	12	5	12
SKBR3	WDR67-ZNF704	8-8	3	3	3	2
SKBR3	CYTH1-EIF3H	17-8	18	37	16	27
MCF7	BCAS4-BCAS3	20-17	105	284	106	105
MCF7	ARFGEF2-SULF2	20-20	17	20	17	12
MCF7	RPS6KB1-TMEM49	17-17	4	3	6	2

Table 3.5 Comparisons of TopHat-Fusion and deFuse

Comparisons of the number of spanning reads and mate pairs reported by TopHat-Fusion and deFuse for 10 previously reported fusion transcripts in the SKBR3 and MCF7 sample data.

Unlike FusionSeq and deFuse (as well as other fusion-finding programs), one of the most powerful features in TopHat-Fusion is its ability to map reads across introns, indels, and fusion points in an efficient way and report the alignments in a modified SAM (Sequence Alignment/Map) format [35].

3.4 Conclusions

Unlike previous approaches based on discordantly mapping paired reads and known gene annotations, TopHat-Fusion can find either individual or paired reads that span gene fusions, and it runs independently of known genes. These capabilities increase its sensitivity and allow it to find fusions that include novel genes and novel splice variants of known genes. In experiments using multiple cell lines from previous studies, TopHat-Fusion identified 34 of 38 previously known fusions. It also found 61 fusion genes not previously reported in those data, each of which had solid support from multiple reads or pairs of reads.

Chapter 4: Reconstruction and Estimation of Fusion Transcripts from RNA-Sequencing reads

Fusion transcripts, in which two distinct genes are fused into a single messenger RNA, can be created by several mechanisms: (1) chromosomal translocations followed by transcription; (2) read-through transcription of two adjacent genes; and (3) trans-splicing of two pre-messenger RNA molecules. One very effective way to detect these fusion events is through the use of RNA sequencing reads, in which fusion breakpoints can be detected by aligning them back to a normal genome. Reads surrounding and spanning the breakpoints can be assembled into fusion transcripts, and the number of reads can be used to estimate expression levels. Two major factors contrive to make this problem more difficult: first, eukaryotic genomes are highly repetitive, meaning the reads can align to many places; and second, sequencing errors (e.g., random ligation of two cDNAs) can give rise to chimeric transcripts. The problem of separating genuine fusion transcripts from these spurious fusion-like transcripts, which are much more numerous than true fusions, is a major algorithmic challenge. The problem is made harder by the fact that reads are non-uniformly distributed across transcripts, making low expression level transcripts difficult to detect. A sensitive and accurate method for identifying authentic fusions should be able to utilize as much evidence as possible that serves as either positive or negative indicators when filtering out potential fusion transcripts. To address these challenges, we have developed TransFUSE, one of the first software systems that can successfully reconstruct and quantify full-length fusion gene transcripts. The newly developed algorithm, which is built on the TopHat2 and Cufflinks systems, can be run with or without known gene annotations, and it can discover novel fusion isoforms that are transcribed from known or unknown genes.

4.1 Background

RNA-sequencing technologies [5, 6, 14, 24] enable us to accurately and precisely assemble and quantify isoforms of genes being expressed in cells. In addition to addressing this fundamental question, RNA-seq data is used to detect fusion genes [17-19]. Fusion genes can be formed at the genomic level when two different chromosomes break and rejoin as illustrated in the first example of Figure 1.3. Fusion genes may also emerge as a result of the breakage and rearrangement of a single chromosome, in which two originally separate sequences are brought together. Most fusion genes have a strong association with distinct types of cancerous tumors, although a few others have been reported in normal cells [25, 26]. As of November 2012, the Mitelman database [27] documented about 62,000 cases of chromosome aberrations and gene fusions in cancer. 1,078 gene fusions have been reported with 1309 participating genes. In addition to these genomic aberrations, fusion events can occur during the transcription process known as read-through transcription, when two adjacent genes are transcribed into a single pre-RNA molecule, which is then spliced into a fusion mRNA. This is illustrated in the second example of Figure 1.3. Akiva et al. [29] applied a bioinformatics approach using expressed sequence tags (ESTs) and cDNAs downloaded from GenBank [30]. Their results showed that about 2% of human genes are associated with such read-through transcription. Fusion transcripts may be formed post-transcriptionally when two different pre-RNA transcripts from two genes are spliced together, forming one single mRNA transcript [31]. This process, called trans-splicing, is shown in the third example of Figure 1.3.

Discovering these fusions via RNA-seq has a distinct advantage over whole-genome sequencing. This is due to the fact that in the highly rearranged genomes of some tumor samples, many rearrangements might be present, although only a fraction might alter transcription. RNA-seq identifies only those chromosomal fusion events that produce transcripts. It has the further advantage that it allows one to detect multiple alternative splice variants that might be produced by a fusion event. Because it does not rely on annotation, it can find events involving novel splice variants and entirely novel genes.

Previously we developed a fusion-finding program, TopHat-Fusion, which is now a part of TopHat2 with a simple command line switch. It discovers fusion break points and it can also align reads across them. After its filtration step, TopHat-Fusion generates highly sensitive and accurate results. Using RNA-seq reads from four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7). Edgren et al. [18] initially reported 24 novel and 3 known fusion genes in this data sample. When we applied TopHat-Fusion to the same data set, 25 of the 27 fusion genes were retrieved, in addition to 51 strong candidates for novel fusion genes. Approximately one year later, Kangaspeska et al. [32] (including Edgren as a coauthor) experimentally verified 13 additional fusion genes. TopHat-Fusion's results already included 9 out of them (see Table 4.4).

In addition to detection of fusion break points by TopHat-Fusion, we have developed Cufflinks-Fusion, a special purpose program of Cufflinks in order to reconstruct and quantify isoforms of a fusion gene. In diploid cells, we have two copies of each gene. For instance, genes *a* and *b* have their homologous copies *a'* and *b'*, respectively. A fusion gene may be formed combining genes *a* and *b*, while genes *a'* and *b'* remain intact (not involved). As a result, several transcripts may be comprised of fusion transcripts from the fusion gene as well as normal transcripts from the intact genes. The splicing patterns of fusion genes and their relative expression levels may be important to understanding underlying causes of some diseases. Expression levels of fusion genes may also be compared with those of normal transcripts from intact genes to provide additional insight. As described previously, TopHat-Fusion provides a list of fusion candidates with high sensitivity and low false positive rates. With more evidence available from Cufflinks-Fusion, including multiple isoforms of fusion genes and their abundance levels, we can put fusion candidates in order, those with more evidence first and those with less evidence after. This will help biologists quickly interpret the data and decide which fusions to address first.

4.2 Methods

We have developed TransFUSE, a new pipeline, in order to discover fusion transcripts using RNA-seq reads. TransFUSE was built based on TopHat2 and Cufflinks-Fusion and consists of three core steps: (1) fusion alignment of reads against the reference genome; (2) assembly and quantification of fusion transcript based on the alignments from step (1); (3) identification of potential fusion transcripts using the evidence collected from the previous steps. Figure 4.1 illustrates these main steps.

Fusion alignment step

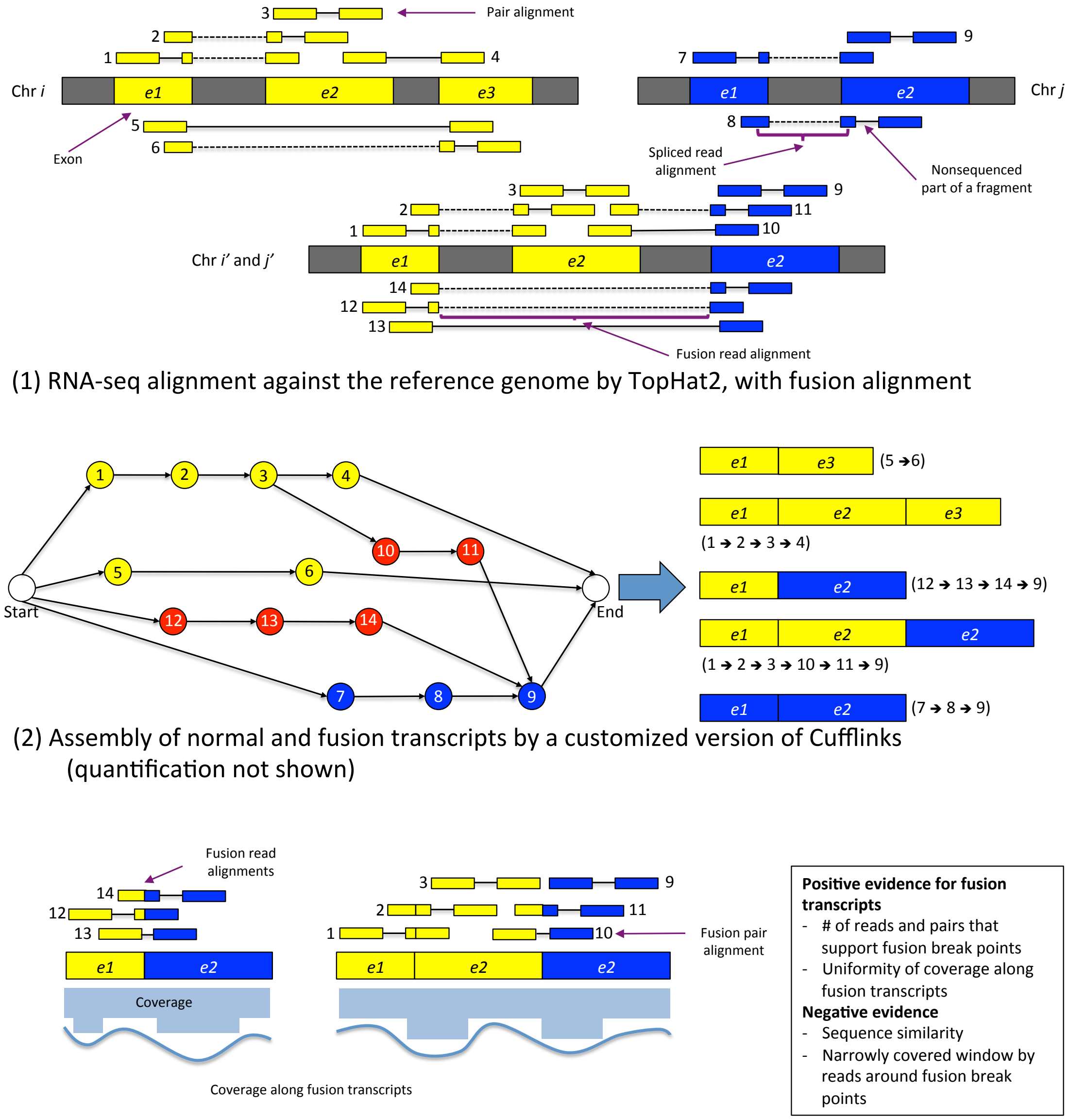
As illustrated in part 1 of Figure 2, RNA-seq reads are aligned against the reference genome, where most of the reads fall into two categories. The reads either lie entirely within an exon or span two or more exons of normal transcripts. TopHat2 effectively handles these cases as follows. Those reads from one exon are aligned by TopHat2's underlying alignment engine Bowtie. However, multi-exon spanning reads need to be aligned across huge gaps due to introns whose length ranges from 50 to 100,000 base pairs in mammalian cells. TopHat2 employs a two-step approach to align these types of reads. First, it identifies splice sites of introns. Second, it stitches together the flanking sequences of introns and then maps reads against the spliced sequences using Bowtie. In contrast to this normal alignment, in RNA-seq samples from abnormal cells, we need to align reads that originate from "fusion" transcripts because fusion transcripts may comprise sequences from two chromosomes, inverted sequences from the same chromosome, or two adjacent sequences that were originally far from each other on the same chromosome. TopHat-Fusion algorithm allows reads to map across fusion break points, which is now incorporated into TopHat2 with a simple command line switch. TopHat2 reports these normal and fusion alignments in Sequence Alignment/Map (SAM) format [35]. SAM is the most popular format that many analysis pipelines use as input. The following steps of TransFUSE also use the SAM file as input.

Assembly and quantification of fusion transcripts

Fusion transcripts can be reconstructed and quantified, using these fusion alignments as input. We have modified Cufflinks [20] to assemble and quantify fusion transcripts based on such mapping information. We will briefly describe the following three steps and then elaborate on the assembly and the quantification steps (see the Cufflinks paper for more details). (i) It goes through the SAM file produced from TopHat2 and identifies groups of overlapping alignments into bundles, where each bundle is likely to represent reads from a normal gene or a fusion gene (see the first part of Figure 4.1). Once it encounters fusion alignments, it stores the fusion break points along with the file offset of the bundle to which the break points belong. (ii) It assembles fusion transcripts using a set of bundles of overlapping reads and fusion break points collected from the above. Because fusions usually involve two distant genomic locations, it is necessary to examine several bundles at the same time and combine them into a fusion bundle group. There may be several conflicting fusion points in the same bundle group. For instance, one fusion break point may involve chromosomes 1 and 7 and another chromosomes 1 and 8. In this case, the fusion point supported by most evidence (e.g. # of fusion reads and # of supporting pairs) is chosen while the others being discarded. This strategy can be used as a filtering step for false positive fusions. (iii) It assesses the abundance level of fusion transcripts based on the number of reads and pairs belonging to them.

(ii) Assembly algorithm: for each bundle, single or paired alignments are sorted based on their mapped locations, which are then represented as vertex in a directed acyclic graph. If two alignments overlap with each other and they are “compatible” (e.g., sharing the same intron or the same break fusion point), then an edge is defined between those two vertices. Based on the graph, Cufflinks finds the minimum number of paths covering the graph. We have extended Cufflinks to handle fusion alignments and define new compatibility relationships between several alignment types such as non-gapped alignment, spliced alignment, and fusion alignment as illustrated in the second part of Figure 4.1.

(iii) Quantification algorithm: Cufflinks-Fusion counts the number of reads or pairs by remapping them against each assembled transcript from the above step. Since some transcripts often share some exons of a gene, it is likely that some reads or pairs are aligned onto several transcripts. In order to disambiguate these conflicting cases, Cufflinks defines the likelihood of observed data (alignments) given abundance of each transcript as parameters. Cufflinks finds the abundance values that maximize the likelihood using EM method. Cufflinks-Fusion (and therefore TransFUSE) reports the expression levels of transcripts using FPKM values. It also provides row counts such as the number of reads and pairs being mapped to each transcript.



(1) RNA-seq alignment against the reference genome by TopHat2, with fusion alignment

(2) Assembly of normal and fusion transcripts by a customized version of Cufflinks (quantification not shown)

Positive evidence for fusion transcripts

- # of reads and pairs that support fusion break points
- Uniformity of coverage along fusion transcripts

Negative evidence

- Sequence similarity
- Narrowly covered window by reads around fusion break points

(3) Filtration of fusion transcripts based on some evidence

Figure 4.1 TransFUSE pipeline

This pipeline comprises of two underlying software: (1) TopHat2 (with fusion alignment option) allows reads to be aligned across fusion break points; (2) Cufflinks-Fusion (an enhanced version based of Cufflinks v1.3.1) allows assembly and quantification of isoforms of a fusion gene and its wild type genes.

Identification of potential fusion transcripts

At the fusion alignment step, we usually observe hundreds of thousands fusion break points for each sample. Almost all of these fusion break points are false positives; they can be attributed to several factors. For instance, the human genome is highly repetitive, with many sequences that are nearly identical to the combined sequences of two distant sequences. This can be problematic because we usually allow a few mismatches in the alignment step to compensate for genomic differences between the reference and the sequenced genomes. This problem can be further complicated when we use short reads, which are likely to be aligned to more locations of the genome. False fusion discovery may also arise due to artifacts in sequencing steps. For instance, accidental ligation of two cDNAs results in chimeric sequences [62].

In order to sort though such an enormous number of fusions, we have defined some positive and negative evidence that can be used to identify fusions. Using the evidence, TransFUSE eliminates most of the false positive fusions and orders the remaining fusions according to the strength of the evidence. Such evidence includes: (1) the number of reads and pairs that support fusion points; (2) sequence similarity; (3) longer transcripts with high and uniform coverage by reads; and (4) alternative splicing around a fusion break point (involving different flanking exons) and different transcript structures on either side or both sides of a fusion gene, which seems to happen even when they share the same flanking exons.

First, our program requires a certain number of reads and pairs that directly support a given fusion break point (e.g., 2 reads and 3 pairs). The more sequenced reads we use, the higher number of reads and pairs can be used to filter out fusions. Second, the flanking sequences around fusion break points are combined, and they are searched against known gene sequences. If matching with high similarity (e.g. >90%) is found, they may not be real, but just instead due to just sequence similarity.

As reported in Edgren et al. [18], true fusion transcripts have reads mapping uniformly in a wide window across the fusion point, whereas false positive fusions are narrowly covered. Our scoring scheme prefers fusions whose window has no gaps (or small gaps), has uniform depth, and is highly covered by reads. These filtration steps usually reduce the number to tens of fusions for a single sample.

During the assembly step of Cufflinks-Fusion, there may be many fusions in the same bundle and conflicting with one another. Instead of trying to assemble every possible fusion, Cufflinks-Fusion chooses just a few of them with the most evidence such as the number of supporting reads and pairs, and assembles the selected fusions. Thus, if there are no assembly results for some fusions, it is mostly the case that there are some other fusions with better supports. Put another way, the very existence of assemblies can be used as positive evidence: therefore we add some positive scores to fusions with assembled transcripts. In addition, some of real fusions may have alternatively spliced transcripts around fusion break points, that is, having different exons flanking the break points. We prefer fusions with such evidence by adding some positive scores. Similarly, if a fusion gene has several transcripts with the same exons flanking the same fusion break point, but different sets of exons on either side or both sides of the break point, we also add some positive scores to those cases. The structural difference among the transcripts may not be direct evidence for true fusions, but it may be worth paying more attention than those without such transcripts.

4.3 Results

In order to evaluate the performance of TransFUSE system, we used two data sets:
(1) Edgren et al. [18] and (2) Seo, Ju, Lee et al. [63] as shown in Table 4.1.

Data source	Sample ID	Read type	Fragment length	Read Length	Number of fragments
Edgren et al.	BT474	Paired	100, 200	50	21,423,697
Edgren et al.	SKBR3	Paired	100, 200	50	18,140,246
Edgren et al.	KPL4	Paired	100	50	6,796,443
Edgren et al.	MCF7	Paired	100	50	8,409,785
Seo, Ju, Lee et al.	LC_S42	Paired	250	101	41045273

Table 4.1 RNA-seq data used to evaluate TransFUSE.
The data came from two studies, and included four samples from breast cancer cells (BT474, SKBR3, KPL4, MCF7) and one sample (LC_S42) from lung cancer cells. For paired-end data, two reads were generated from each fragment; thus the total number of reads is twice the number of fragments.

We mapped all reads to the human genome (UCSC hg19) with TopHat2 (v2.0.7) with fusion alignment enabled. Based on the alignments, we assembled and quantified fusion transcripts using Cuffliks-Fusion. We subsequently applied TopHat-Fusion's filtering algorithm (implemented in "tophat-fusion-post," which is a part of TopHat2) to identify the genes involved in each fusion using the RefSeq and Ensembl human annotations (see Methods for detailed description). 63 fusions from Edgren et al. are present in Table2, 34 of which are true fusions.

SAMPLE ID	Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads	Spanning pairs	Transcripts assembled?
SKBR3	**TATDN1-GSDMB**	8-17	125551265	38066176	324	485	Yes
BT474	**THRA-SKAP1**	17-17	38243105	46384692	25	53	Yes
BT474	**SNF8-RPS6KB1**	17-17	47021336	57970685	57	81	Yes
BT474	**MRPL45-TRPC4AP**	17-20	36476501	33665848	3	11	Yes
MCF7	USP32-PPM1D	17-17	58342772	58679978	2	5	Yes
BT474	**MYO19-SKA2**	17-17	34863350	57232491	5	14	Yes
MCF7	**BCAS3-BCAS4**	17-20	59445687	49411709	103	286	Yes
BT474	**SYNRG-TOB1**	17-17	35880750	48943418	28	87	Yes
BT474	**IKZF3-VAPB**	17-20	37934019	56964572	19	51	Yes
BT474	ENSG00000248527-GNAS	1-20	569609	57484588	1	6	No
KPL4	**NUP214-NOTCH1**	9-9	134062675	139438475	3	8	Yes

BT474	**MED1-ACSF2**	17-17	37595417	48548388	10	20	Yes
SKBR3	**SUMF1-LRRFIP2**	3-3	4418013	37170639	5	7	Yes
SKBR3	SSH2-EFCAB5	17-17	28030079	28256955	1	6	Yes
BT474	ENSG00000248527-PCBD2	1-5	569609	134263267	3	5	No
BT474	**BCAS3-MED13**	17-17	59469337	60129897	3	14	Yes
BT474	**RAE1-STX16**	20-20	55929087	57227142	6	35	Yes
BT474	**ACACA-STAC2**	17-17	35479452	37374425	66	100	Yes
SKBR3	**ZNF704-WDR67**	8-8	81733849	124096579	3	6	Yes
BT474	**CPNE1-PI3**	20-20	34243123	43804501	2	6	Yes
KPL4	**BSG-NFIX**	19-19	580781	13135834	12	39	Yes
BT474	**AHCTF1-NAAA**	1-4	247094879	76846963	12	41	Yes
MCF7	**TMEM49-RPS6KB1**	17-17	57917128	57992063	6	8	Yes
SKBR3	**CSE1L-ENSG00000236127**	20-20	47688989	47956856	12	38	Yes
MCF7	SULF2-ZNF217	20-20	46415148	52210645	12	33	Yes
SKBR3	MRPS28-TPD52	8-8	80831382	80954854	3	4	Yes
SKBR3	**ANKHD1-PCDH1**	5-5	139825559	14123400	5	22	Yes
KPL4	PARP1-ENSG00000227105	1-13	226579911	111589382	1	36	No
MCF7	FOXA1-ENSG00000254868	14-14	38061534	38184710	4	50	Yes
SKBR3	**SETD3-CCDC85C**	14-14	99880270	99880270	5	5	Yes
BT474	**CEP250-ZMYND8**	20-20	34078462	45852969	8	58	Yes
SKBR3	SNTB1-KLHDC2	8-14	121561197	50249311	2	5	Yes
BT474	ENSG00000229344-ERBB2	1-17	568761	37880978	2	21	No
BT474	**TTI1-DIDO1**	20-20	36634798	61569147	1	11	Yes
MCF7	RSBN1-AP4B1	1-1	114354329	114442495	6	9	Yes
BT474	**MCF2L-LAMP1**	13-13	113718617	113951809	2	6	Yes
KPL4	MUC20-ENSG00000236833	3-3	195456609	197391652	7	12	Yes
BT474	GABRA3- ZNF185	X-X	151468339	152114007	1	6	Yes
SKBR3	DIO2-ENSG00000249517	14-14	80669630	80854020	2	4	Yes
MCF7	**SULF2-ARFGEF2**	20-20	46365685	47538546	21	40	Yes
BT474	ENSG00000198744-ATP5B	1-12	569880	57038738	1	13	No
SKBR3	**EIF3H- CYTH1**	8-17	117768257	76778283	19	33	Yes
SKBR3	DHFR-H19	5-11	79946842	2017318	1	6	No
BT474	WBSCR17-FBXL20	7-17	70958326	37557613	2	7	Yes
BT474	ENSG00000225630-HFM1	1-1	570103	91853140	12	42	No
BT474	HFM1-DLG2	1-11	91853144	85195025	2	9	No
KPL4	EEF1DP3-FRY	13-13	32520314	32652967	2	8	Yes
BT474	**CMTM7-GLB1**	3-3	32483331	33055547	2	8	Yes
MCF7	**ENSG00000224738-TMEM49**	17-17	57184951	57915655	4	9	No
MCF7	LRP1B-PLXDC1	2-17	142237963	37265642	2	5	Yes
KPL4	ENSG00000249796-MUC20	3-3	195352201	195456609	7	267	No
MCF7	PRRC2A-ENSG00000224067	6-9	31604384	114565349	2	5	No
KPL4	**SEPT10-PPP1R12A**	2-12	110343414	80211173	4	6	Yes
MCF7	CARM1-SMARCA4	19-19	11015626	11097268	2	4	Yes
BT474	**STARD3-DOK5**	17-20	37793483	53259996	6	10	Yes
BT474	**MYO9B-RAB22A**	19-20	17256206	56886177	8	22	Yes
SKBR3	ENSG00000243185-KRT18	4-12	70296743	53342904	3	14	No
BT474	PPP6R3-SHANK2	11-11	68228294	70803333	4	12	Yes
BT474	**MED1-STXBP4**	17-17	37607290	53218670	13	16	Yes
MCF7	ENSG00000233459- ZNF207	2-17	204499953	30692348	1	49	No
KPL4	BAG4-ENSG00000255107	8-8	38066752	70771975	1	8	No
SKBR3	**PKIA-RARA**	8-17	79510592	38465537	1	6	Yes
BT474	PCBD2-UBB	5-17	134259838	16284410	1	5	No

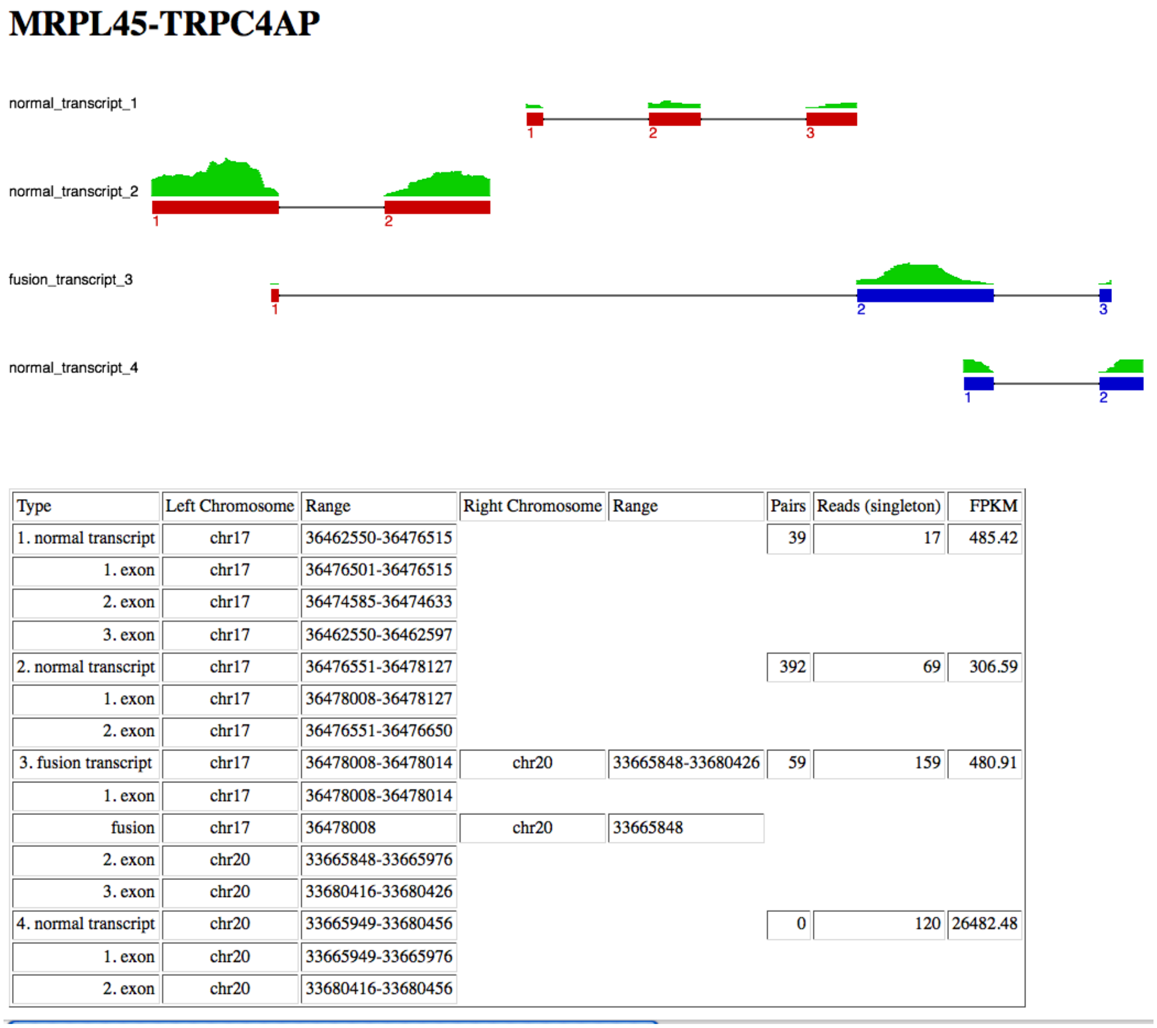
Table 4.2 TransFUSE detected 63 fusions, 34 of which are genuine fusions. Using four breast cancer cell lines (BT474, SKBR3, KPL4, MCF7), Edgren et al. [18] initially discovered 27 true fusion genes. Later, the same research group (Kangaspeska et al. [32]) improved their bioinformatics pipeline, leading to the discovery of an additional 13

fusion genes. The results include 34 out of the 40 true fusion genes. The 25 fusions verified by Edgren et al. are shown in boldface, whereas the 9 fusions by Kangaspeska et al. shown in boldface and red.

Figure 4.2 shows one example from TransFUSE's output, a set of fusion and normal transcripts. These transcripts presumably arise from the known fusion gene MRPL45-TRPC4AP and its wild type genes in the BT474 sample. TransFUSE's output includes depth coverage across the transcripts, the coordinates of the exons in each transcript, and the number of reads and pairs that map to each transcript. TransFUSE also provides a FPKM value, which represents the abundance of each transcript (FPKM stands for Fragments Per Kilobase of transcript per Million mapped fragments). FPKM takes into consideration the length of a transcript in calculating the expression level of the transcript.

Some of the fusion genes appear to be alternatively spliced. For instance, SUMF1-LRRFIP2 from the SKBR3 sample has 3 fusion transcripts as illustrated in the first example of Figure 4.3. The fusion transcripts #3 and #4 appear to have the same exons flanking of the fusion break point. However, they have different transcript structures: #4 has one additional exon, #11, on the right partner gene TRPC4AP. The fusion transcript #2 includes a left flanking exon different from those of #3 and #4. It also involves a different splicing pattern on the right partner gene. It is noteworthy that a fusion gene is expressed, while one or both of the two wild type genes may not be expressed. For instance, the fusion gene IKZF3-VAPB (shown in the second example of Figure 4.3) produces three different fusion transcripts. In contrast to the three normal transcripts expressed from a wild type gene VAPB, the other wild type IKZF3 may be expressed at a very low level, or not expressed at all.

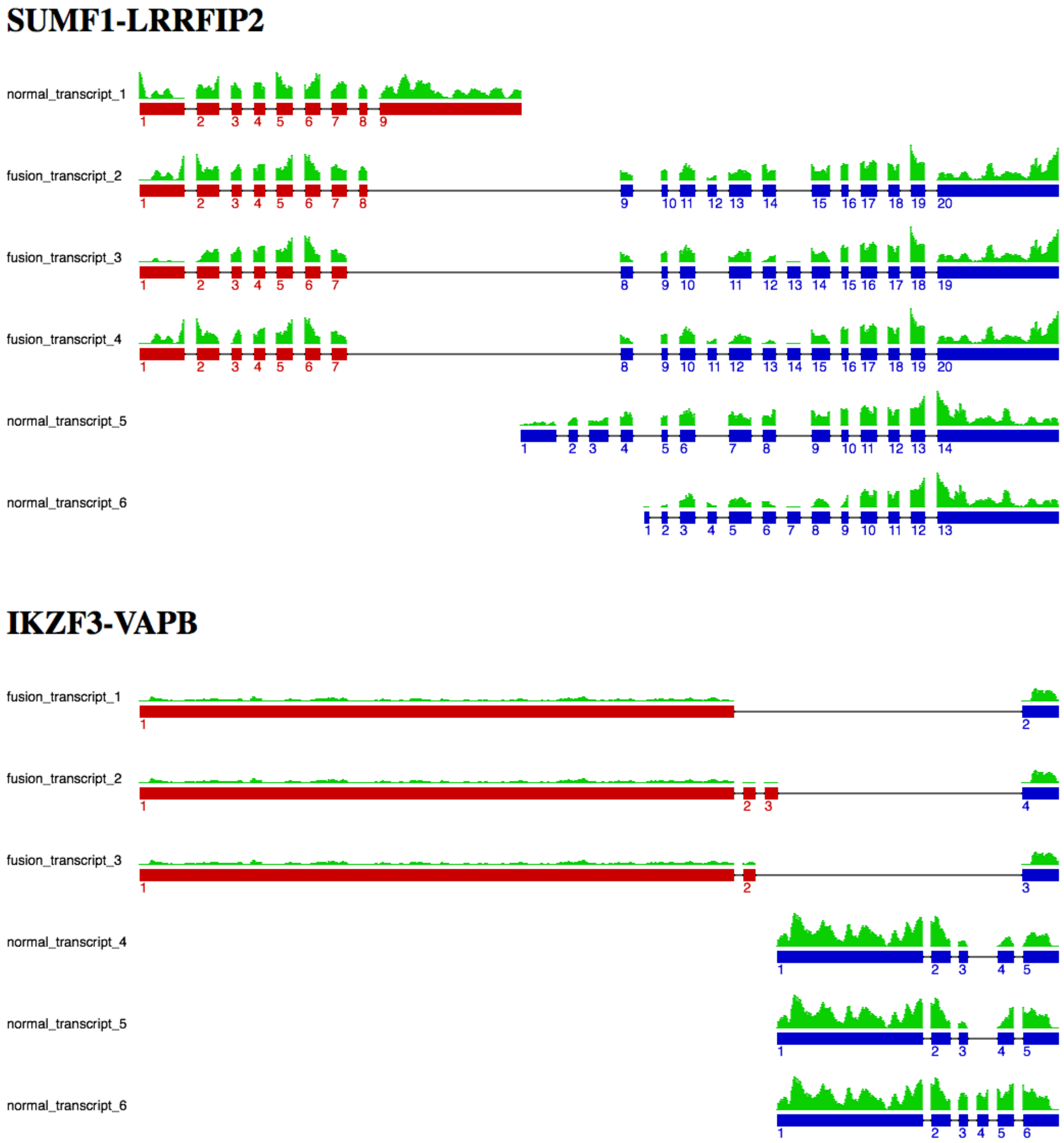
Cufflinks-Fusion is able to assemble 33 fusions of the 34 known fusions in the four breast cancer samples. A total of 17 fusions in the list are not assembled mostly because they do not have enough mapped reads or pairs to reconstruct transcripts. There may be too many other fusion break points nearby that conflict with the genuine one. In this case, the assembly algorithm often chooses just a few of them that may not include the true fusion.



Type	Left Chromosome	Range	Right Chromosome	Range	Pairs	Reads (singleton)	FPKM
1. normal transcript	chr17	36462550-36476515			39	17	485.42
1. exon	chr17	36476501-36476515					
2. exon	chr17	36474585-36474633					
3. exon	chr17	36462550-36462597					
2. normal transcript	chr17	36476551-36478127			392	69	306.59
1. exon	chr17	36478008-36478127					
2. exon	chr17	36476551-36476650					
3. fusion transcript	chr17	36478008-36478014	chr20	33665848-33680426	59	159	480.91
1. exon	chr17	36478008-36478014					
fusion	chr17	36478008	chr20	33665848			
2. exon	chr20	33665848-33665976					
3. exon	chr20	33680416-33680426					
4. normal transcript	chr20	33665949-33680456			0	120	26482.48
1. exon	chr20	33665949-33665976					
2. exon	chr20	33680416-33680456					

Figure 4.2 A set of fusion and normal transcripts from a known fusion gene (MRPL45-TRPC4AP) and its wild type genes, generated by TransFUSE using a breast cancer cell sample (BT474).

TransFUSE generates output in html format. The figure is a part of the output. There are four transcripts; a transcript number is given on the left side of each transcript. The first two of them are normal transcripts, most likely coming from a wild type MRPL45 (shown in red). The third one is a fusion transcript, and the last one is from a wild type TRPC4AP (shown in blue). Red- and blue-colored boxes represent exons from MRPL45 and TRPC4AP, respectively. Introns are indicated by thin black lines. Coverage depths are shown in green. Exons, introns, and coverage depths are scaled to fit into the smaller display of the output. The order of exons in their respective transcripts is indicated by the small numbers below the bottom left corner of the exons. These numbers facilitate reference to the genomic coordinates of the transcripts or exons in the table at the bottom (note that this number is not equal to the exon number). The table also shows the number of pairs and reads that map to each transcript along with FPKM value.



SUMF1-LRRFIP2

IKZF3-VAPB

Figure 4.3 Two known fusion genes SUMF1-LRRFIP2 in SKBR3 sample and IKZF3-VAPB in BT474 sample are shown.
The upper example shows six transcripts. Three of them are normal transcripts, most likely coming from either wild type gene SUMF1 or LRRFIP2. The others are fusion transcripts. The fusion transcripts #3 and #4 appear to have the same flanking exons in common, but it has different transcript structures where #4 transcript has one additional exon, #11. The fusion transcript #1 includes a left flanking exon different from those of #3 and #4. The fusion gene IKZF3-VAPB at the bottom produces three different fusion transcripts. While three normal transcripts are made from a wild type gene VAPB, the other wild type IKZF3 may not be expressed or expressed at low level if at all.

For the lung cancer cell sample (LC_S42), we found 8 fusion candidates in addition to a known fusion gene KIF5B-RET [63] (see Table 4.3).

SAMPLE ID	Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads	Spanning pairs	Transcripts assembled?
LC_S42	HEBP2-VTA1	6-6	138734016	142525201	118	52	Yes
LC_S42	**KIF5B-RET**	10-10	32317355	43612031	57	13	Yes
LC_S42	ENSG00000211653-IGLL5	22-22	22764609	23235959	15	32	Yes
LC_S42	CCT2-LGR5	12-12	69987392	71835366	131	28	Yes
LC_S42	NMBR-CPM	6-12	142400039	69326457	23	14	Yes
LC_S42	HECA-CPM	6-12	139491700	69326457	62	50	Yes
LC_S42	OVCH2-LOC283299	11-11	7726220	7900553	2	4	Yes
LC_S42	INS-COIL	11-17	2153283	55015815	2	2	Yes
LC_S42	KIF5B-RET	10-10	32311963	43610099	6	3	Yes

Table 4.3 Fusions found by TransFUSE using one lung cancer cell sample (LC_S42).
Seo, Ju, Lee et al. [63] previously reported one fusion gene KIF5B-RET (shown in boldface). Note that there is another fusion gene candidate KIF5B-RET at the 9th row. This fusion gene is different from the one at the 2nd row in terms of the location of a fusion break point and in that different strands of the two genes are combined forming the fusion gene.

4.4 Conclusions

TransFUSE augments our previous fusion-finding program, TopHat-Fusion (now a part of TopHat2) with additional functionalities such as assembling and quantifying fusion and normal transcripts that together comprise isoforms of a fusion gene and its wild type genes. Previous results from TopHat-Fusion [19] demonstrated it is highly sensitive and its false positive rate is relatively low. With more evidence available from TransFUSE, such as several isoforms of a fusion gene and the expression levels of transcripts, we can sort fusion candidates in a fashion that fusions with more evidence appear before those with less evidence. This can help biologists quickly interpret the data and decide which fusions to address first. Unlike previous approaches that simply provide a list of candidate fusions (genomic locations of break points), TransFUSE provides detailed information about full-length fusion transcripts. These capabilities enable one to infer the potential function of a fusion gene by examining the participating exons of the transcripts and their splicing patterns and perhaps to identify a basis for the underlying causes of diseases. Expression levels of fusion genes may also be compared with those of normal transcripts from wild type genes to provide additional insight.

4.5 Supplementary Material

SAMPLE ID	Fusion genes (left-right)	Chromosomes (left-right)	5' position	3' position	Spanning reads	Spanning pairs
BT474	**TRPC4AP-MRPL45**	20-17	33665850	36476499	2	9
BT474	**TOB1-SYNRG**	17-17	48943418	35880750	26	47
SKBR3	**TATDN1-GSDMB**	8-17	125551264	38066175	311	555
BT474	**THRA-SKAP1**	17-17	38243102	46384689	28	46
MCF7	**BCAS4-BCAS3**	20-17	49411707	59445685	105	284
BT474	**ACACA-STAC2**	17-17	35479452	37374425	57	59
BT474	**STX16-RAE1**	20-20	57227142	55929087	6	24
BT474	**MED1-ACSF2**	17-17	37595419	48548386	10	12
MCF7	ENSG00000254868-FOXA1	14-14	38184710	38061534	2	22
SKBR3	**ANKHD1-PCDH1**	5-5	139825557	141234002	4	15
BT474	**ZMYND8-CEP250**	20-20	45852972	34078459	10	53
BT474	**AHCTF1-NAAA**	1-4	247094879	76846963	10	42
SKBR3	**SUMF1-LRRFIP2**	3-3	4418012	37170638	3	12
KPL4	**BSG-NFIX**	19-19	580779	13135832	12	27
BT474	**VAPB-IKZF3**	20-17	56964574	37922743	4	14
BT474	DLG2-HFM1	11-1	85195025	91853144	2	10
SKBR3	**CSE1L-ENSG00000236127**	20-20	47688988	47956855	13	31
MCF7	RSBN1-AP4B1	1-1	114354329	114442495	6	7
BT474	**MED13-BCAS3**	17-17	60129899	59469335	3	14
MCF7	**ARFGEF2-SULF2**	20-20	47538545	46365686	17	20
BT474	HFM1-ENSG00000225630	1-1	91853144	565937	2	43
KPL4	MUC20-ENSG00000249796	3-3	195456606	195352198	13	46
KPL4	MUC20-ENSG00000236833	3-3	195456612	197391649	8	15
MCF7	**RPS6KB1-TMEM49**	17-17	57992061	57917126	4	3
SKBR3	**WDR67-ZNF704**	8-8	124096577	81733851	3	3
BT474	**CPNE1-PI3**	20-20	34243123	43804501	2	6
BT474	ENSG00000229344-RYR2	1-1	568361	237766339	1	19
BT474	**LAMP1-MCF2L**	13-13	113951808	113718616	2	6
MCF7	SULF2-ZNF217	20-20	46415146	52210647	11	32
BT474	WBSCR17-FBXL20	7-17	70958325	37557612	2	8
MCF7	**ENSG00000224738-TMEM49**	17-17	57184949	57915653	5	6
MCF7	ANKRD30BL-RPS23	2-5	133012791	81574161	2	6

BT474	ENSG00000251948-SLCO5A1	19-8	24184149	70602608	2	6
BT474	**GLB1-CMTM7**	3-3	33055545	32483333	2	6
KPL4	EEF1DP3-FRY	13-13	32520314	32652967	2	4
MCF7	PAPOLA-AK7	14-14	96968936	96904171	3	3
BT474	ZNF185-GABRA3	X-X	152114004	151468336	2	3
KPL4	**PPP1R12A-SEPT10**	12-2	80211173	110343414	3	8
BT474	**SKA2-MYO19**	17-17	57232490	34863349	5	12
MCF7	LRP1B-PLXDC1	2-17	142237963	37265642	2	5
BT474	NDUFB8-TUBD1	10-17	102289117	57962592	1	49
BT474	ENSG00000225630-NOTCH2NL	1-1	565870	145277319	1	18
SKBR3	**CYTH1-EIF3H**	17-8	76778283	117768257	18	37
BT474	PSMD3-ENSG00000237973	17-1	38151673	566925	1	12
BT474	**STARD3-DOK5**	17-20	37793479	53259992	2	10
BT474	**DIDO1-TTI1**	20-20	61569147	36634798	1	10
BT474	**RAB22A-MYO9B**	20-19	56886176	17256205	8	20
KPL4	PCBD2-ENSG00000240967	5-5	134259840	99382129	1	32
SKBR3	**RARA-PKIA**	17-8	38465535	79510590	1	5
BT474	**MED1-STXBP4**	17-17	37607288	53218672	13	11
KPL4	C1orf151-ENSG00000224237	1-3	19923605	27256479	1	5
SKBR3	RNF6-FOXO1	13-13	26795971	41192773	2	13
SKBR3	BAT1-ENSG00000254406	6-11	31499072	119692419	2	30
BT474	KIAA0825-PCBD2	5-5	93904985	134259811	1	19
SKBR3	PCBD2-ANKRD30BL	5-2	134263179	133012790	1	5
BT474	ENSG00000225630-MTRNR2L8	1-11	565457	10530147	1	35
BT474	PCBD2-ENSG00000251948	5-19	134260431	24184146	2	6
BT474	ANKRD30BL-ENSG00000237973	2-1	133012085	567103	2	8
KPL4	ENSG00000225972-HSP90AB1	1-6	564639	44220780	1	7
BT474	MTIF2-ENSG00000228826	2-1	55470625	121244943	1	11
BT474	ENSG00000224905-PCBD2	21-5	15457432	134263223	2	7
BT474	**RPS6KB1-SNF8**	17-17	57970686	47021335	48	57
BT474	MTRNR2L8-PCBD2	11-5	10530146	134263156	1	6
BT474	RPL23-ENSG00000225630	17-1	37009355	565697	3	19
BT474	MTRNR2L2-PCBD2	5-5	79946288	134259832	1	5
SKBR3	ENSG00000240409-PCBD2	1-5	569005	134260124	2	4
SKBR3	PCBD2-ENSG00000239776	5-12	134263289	127650986	2	3
BT474	ENSG00000239776-MTRNR2L2	12-5	127650981	79946277	2	3
BT474	JAK2-TCF3	9-19	5112849	1610500	1	46
KPL4	**NOTCH1-NUP214**	9-9	139438475	134062675	3	5
BT474	MTRNR2L8-TRBV25OR92	11-9	10530594	33657801	4	4
BT474	MTRNR2L8-AKAP6	11-14	10530179	32953468	1	5
BT474	ENSG00000230916-PCBD2	X-5	125606246	134263219	1	5
MCF7	ENSG00000226505-MRPL36	2-5	70329650	1799907	5	20
SKBR3	**CCDC85C-SETD3**	14-14	100002351	99880270	5	6
BT474	RPL23-ENSG00000230406	17-2	37009955	222457168	109	5

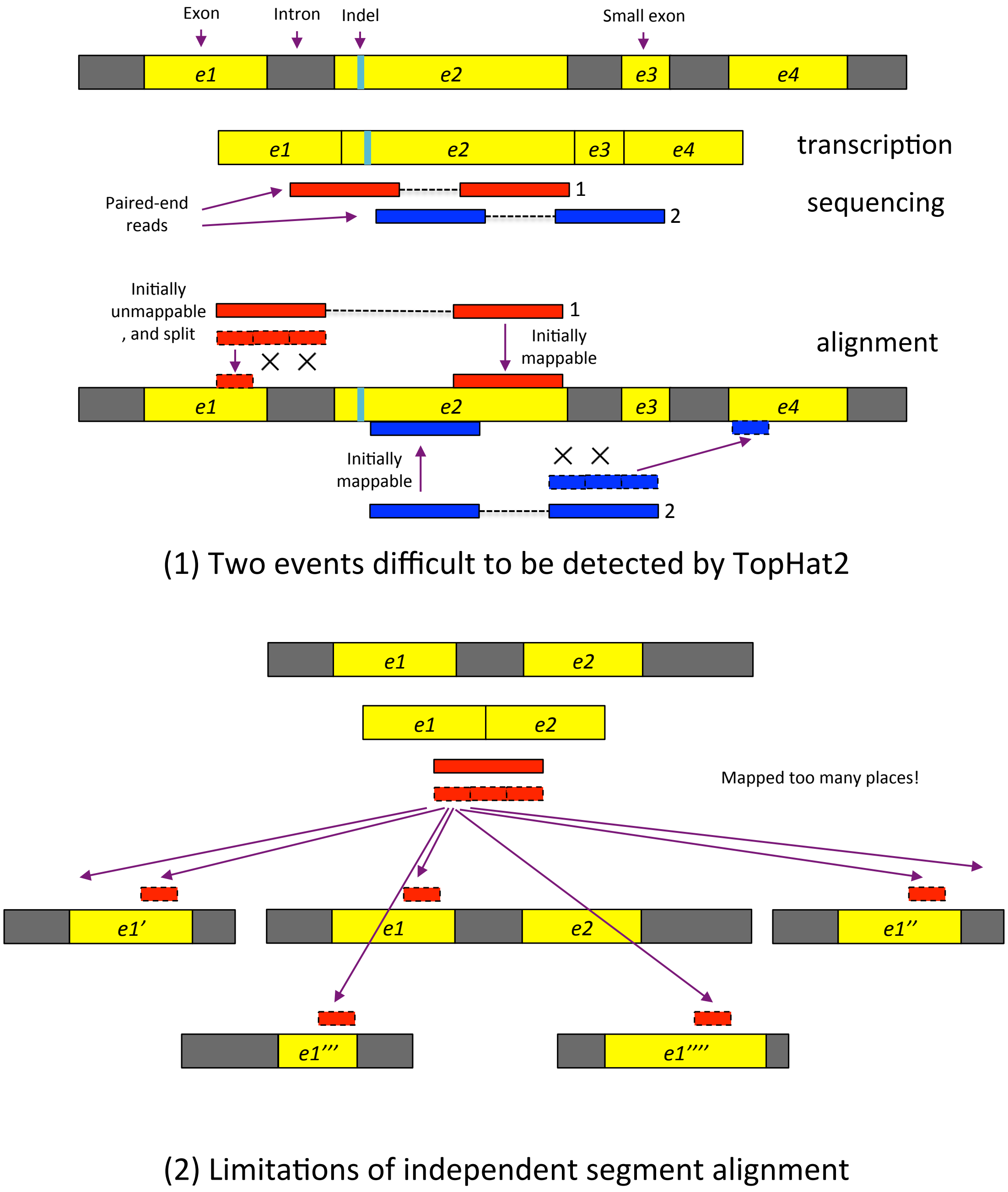
Table 4.4 76 fusions initially identified by TopHat-Fusion.
This table was excerpted from the TopHat-Fusion paper [19], Table 2 and modified as follows. This is a list of 76 of fusion gene previously predicted by TopHat-Fusion at which time 24 of them were known to be true (shown in boldface). Additional 9 new fusion genes validated by Kangaspeska et al. [32] are shown in boldface and red.

Chapter 5: A proposal for a new RNA-seq alignment pipeline

5.1 Limitations of current approaches to the problem of RNA-seq alignment

Chapter 2 covered two popular methods of aligning RNA-seq reads against the reference genome. Many alignment programs employ a one-step approach in which a read is aligned independently of other reads. Most aligners rely on k-mer (usually with k >10 bp) mapping to initially identify potential read origin locations in the genome. Some of the reads are easily aligned with this approach when they have enough bases (>= k) around splicing events or indels. For such reads, we can effectively narrow down the range where events lie, as reads' k-mer mapping allows us to identify the left and right boundaries between which these events fall. However, other reads that have few bases on either side of such events are extremely hard or inefficient to align due to the short anchors. Thus, this approach tends to misalign or fail to align those short-anchored reads. This is a nontrivial issue for RNA-seq alignment, considering a significant portion of reads (e.g., about 20% of 100-bp reads) is estimated to have at most a 10-bp anchor on either side of the introns. In contrast, some other aligners such as TopHat and MapSplice use a two-step approach. First, they find and collect splice sites using reads that have a sufficient amount of bases around them. The sequences flanking the splice sites then are glued together, producing spliced sequences. Second, reads lacking sufficiently long anchors are aligned against the spliced sequences, and then their "transcriptomic" coordinates are converted into the corresponding genomic coordinates. This two-step approach provides a highly sensitive and accurate alignment compared to the one-step method.

However, this is at the cost of much more I/O processing and time due to the two alignment steps: initial alignments of the reads and subsequent alignments of those initially unmapped reads.



(1) Two events difficult to be detected by TopHat2

(2) Limitations of independent segment alignment

Figure 5.1 Limitations of TopHat2 pipeline
Details are given in the text.

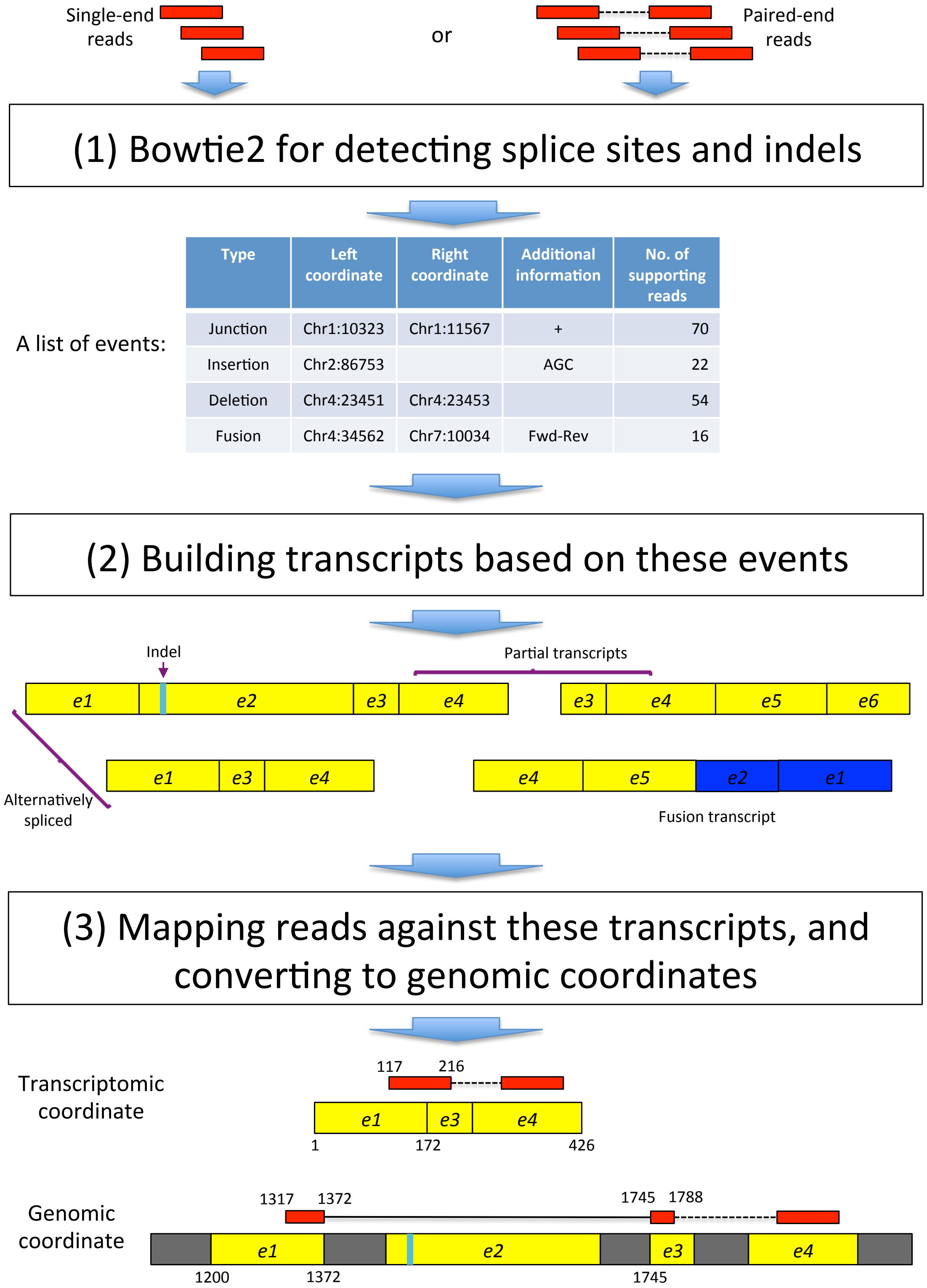
In addition to these inherent limitations of the two-step approach, TopHat2 has its own specific issues this suggests room for further improvement. First, it is hard to detect indels near splice sites. TopHat2 works well for simpler cases in which reads include just one event. With a high coverage of reads, we can expect reads with enough anchors around such an event. TopHat2 splits the reads into several small non-overlapping segments with a default length of 25 bp. Then, by mapping segments, TopHat2 can identify the small range where the event is located. It can also use the unmapped segments of the reads to pinpoint the precise location of the event (see Chapter 2 for more details). However, this is no longer the case when reads span two exons with an indel close to the splice site between them. For the sake of the discussion, it is inconsequential whether the indel is an insertion or a deletion for the sake of the discussion. As mentioned previously, TopHat2 requires at least two segments of reads to be aligned, where one segment is on the left side of the event and the other on the right. However, if an insertion or a deletion is <= 25 bp away from a splice site, it is unlikely in most cases that we have the requisite two left and right segments to be aligned. To provide a more concrete understanding, let us consider a read (shown in read) consisting of three segments (left, middle, and right) in Figure 5.1 (1). This read is initially unmappable because it involves two exons, and the right exon contains an indel. The read is subsequently split into three segments, and only one of them is mapped. This mapping does not satisfy the two-segment mapping condition; therefore, TopHat2 is not able to detect the splicing event or the indel. This prohibits the read from aligning. It is also hard to find small exons, called micro-exons, for the same reasoning. Second, as illustrated in Figure

5.1 (2), TopHat2 aligns segments independently, without using information other segments' mapped locations. Short segments can map to too many locations; in a highly repetitive genome like the human genome, segments can map to hundreds or even thousands of locations. TopHat 2 imposes a certain limit on the number of locations a segment can map to, in order to prevent itself from producing large intermediate files for the segment alignment and consuming too much time searching for all possible alignments. As a result of the limit, reads containing such repetitive segments may not be aligned by TopHat2. On the other hand, the segment mapping would be more efficient if we make use of other segments' alignment location. For instance, if we know some segments of reads are perfectly or nearly perfectly aligned to only a few locations, we may narrow down the search space for the other segments near these locations. This will likely to make it easier and more efficient to find the correct locations for segments and pinpoint the events that reads contain.

5.2 A new pipeline for RNA-seq alignment

In the previous section, we mentioned the limitations of TopHat2; primarily, its segment mapping. First, in case of reads containing more than one event (e.g. one splice site with one insertion close), TopHat2 may not be able to locate the events. Even if two segments are mapped, it may involve a very slow search to identify these events, possibly using dynamic programming algorithm such as Needleman-Wunsch [64] and Smith-Waterman [65]. However, implementing this algorithm is nontrivial. We realized that Bowtie2 includes a very efficient implementation of such an algorithm as part of its engine. It makes use of single-instruction multiple-data

(SIMD) parallel processing, which is significantly faster and is available on modern CPUs [66, 67]. Second, TopHat2 aligns segments independently, without using information other segments' mapped locations. The segment mapping would be more efficient if we make use of other segments' alignment location. For instance, if we know some segments of reads are perfectly or nearly perfectly aligned to only a few locations, we may narrow down the search space for the other segments near these locations.



(1) Bowtie2 for detecting splice sites and indels

A list of events:

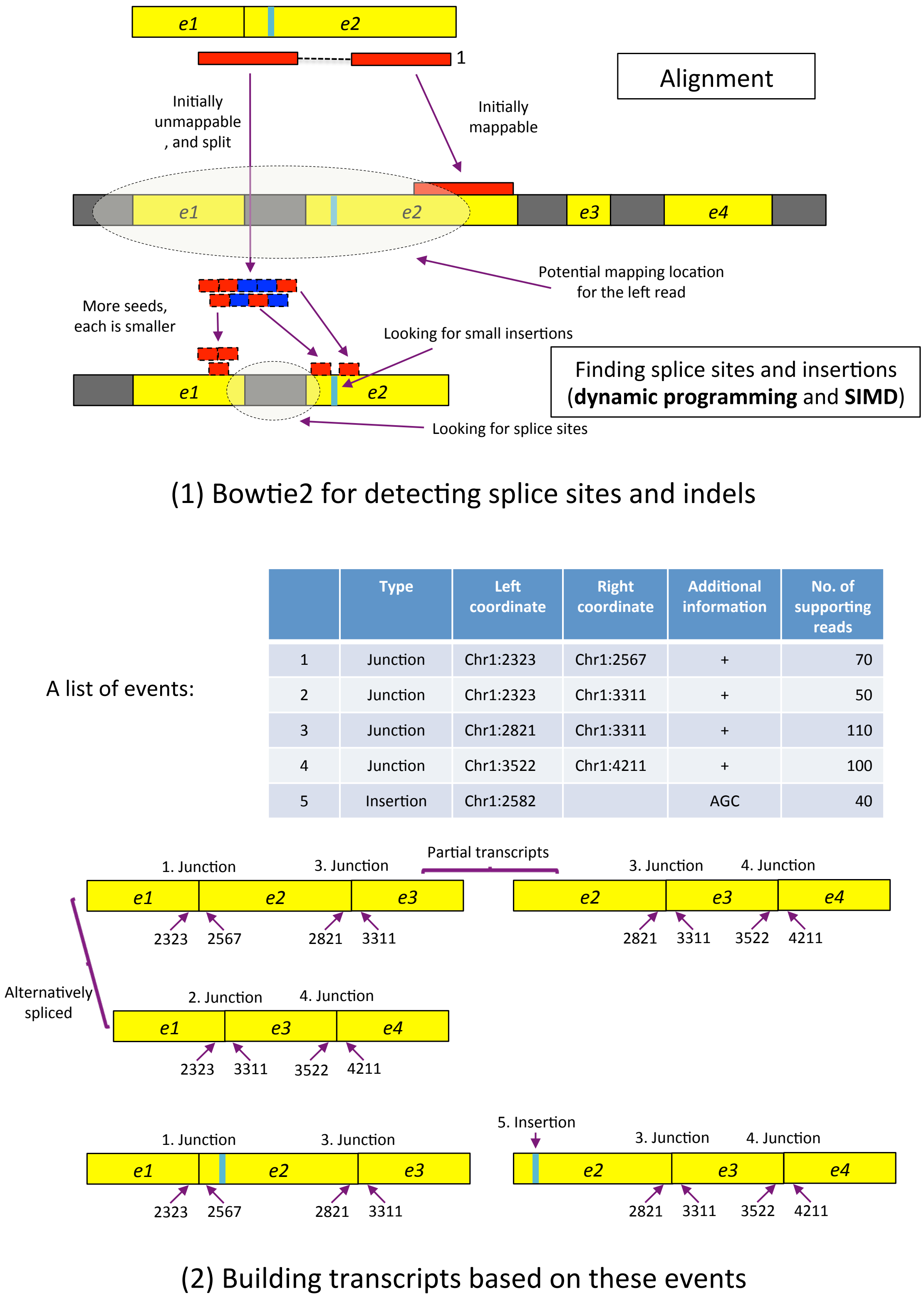
Type	Left coordinate	Right coordinate	Additional information	No. of supporting reads
Junction	Chr1:10323	Chr1:11567	+	70
Insertion	Chr2:86753		AGC	22
Deletion	Chr4:23451	Chr4:23453		54
Fusion	Chr4:34562	Chr7:10034	Fwd-Rev	16

(2) Building transcripts based on these events

(3) Mapping reads against these transcripts, and converting to genomic coordinates

Figure 5.2 Three main steps of a new RNA-seq alignment pipeline

Here, we suggest a new pipeline for RNA-seq alignment, which would incorporate TopHat2's segment alignment and detection of splicing events, indels, and fusions into Bowtie2. As we have discussed previously, Bowtie 2 has already implemented some algorithms that can be modified to efficiently handle these issues. There are three core steps at the heart of the new pipeline, as illustrated in Figure 5.2. As the first step of the new pipeline, Bowtie 2 can be modified and enhanced to identify splice sites and indels. Then, it will report the events with some evidence, such as the number of reads supporting those events (see Figure 5.2 (1)). These events can be used to reconstruct transcripts being observed in samples sequenced. Unlike the problem of the reconstruction of full-length transcripts, we only need to reconstruct partial transcripts as long as a read that is supposed to map to a full-length transcript is mapped to at least one partial transcript that is part of the full-length transcript as illustrated in Figure 5.2 (2). While reconstructing full-length transcripts involves exponential combinations of splicing events, partial transcripts involve dramatically fewer combinations. Similarly, we can create partial transcripts that include indels as well as fusion transcripts. As shown in Figure 5.2 (3), in contrast to the genome alignment, this transcriptome alignment will make Bowtie2 to focus on just base-level mismatches or indels introduced in the sequencing steps. A further advantage is this alignment step is likely very fast because this transcriptome is usually expected to comprise just a small percentage of the whole genome. Finally, the transcriptomic coordinates of read alignments are converted into the corresponding genomic coordinates, and the final alignment is reported in SAM format.



(1) Bowtie2 for detecting splice sites and indels

A list of events:

	Type	Left coordinate	Right coordinate	Additional information	No. of supporting reads
1	Junction	Chr1:2323	Chr1:2567	+	70
2	Junction	Chr1:2323	Chr1:3311	+	50
3	Junction	Chr1:2821	Chr1:3311	+	110
4	Junction	Chr1:3522	Chr1:4211	+	100
5	Insertion	Chr1:2582		AGC	40

(2) Building transcripts based on these events

Figure 5.3 Two core algorithms of the new pipeline
Details are given in the text.

Here, we elaborate on two main ideas: (1) Bowtie2 enhancements to identify splicing events, indels, and fusion break points and (2) reconstruction of partial transcripts using the events. As illustrated in Figure 5.3 (1), the left read of the fragment shown in red involves two events, one splicing event with one indel being close. In order to find these events, Bowtie2 can split the read into segments (or "seeds" in the Bowtie terminology) with shorter length (between 10 and 20 bp), where segments can overlap with some others. Unlike TopHat2's segment mapping (longer segment – 25 bp and non-overlapping segments), this will increase chance to anchor more segments near these events. Once we detect some discrepancies between two segments, that is, the genomic distance between their mapped locations is different from that distance between their positions in the read, we can apply a modified version of Bowtie2's SIMD-accelerated dynamic algorithm using to identify those events. The results from this algorithm are a list of events with some evidence such as the number of reads supporting them. Based on the list, we can reconstruct partial transcripts instead of trying to build full-length ones. We need to ensure that a read that was supposed to map to a full-length transcript is mapped to one partial transcript, which is a part of the full-length transcript. For instance, shown in Figure 5.3 (2), instead of producing a four-exon transcript (*e1*-*e2*-*e3*-*e4*), we can generate two partial transcripts: *e1*-*e2*-*e3* and *e2*-*e3*-*e4*. For the alternative splicing event between *e1* and either *e2* or *e3*, we can produce an additional transcript, *e1*-*e3*-*e4*. For the indel event, two additional transcripts are constructed: *e1*-$e2^{'}$-*e3* and $e2^{'}$-*e3*-*e4*, where $e2^{'}$ is the identical copy of the exon *e2* except the indel. Constructing fusion transcripts can be done in a similar way. While reconstructing full-length transcripts involves exponential combination of splicing events, indels, and fusion break points, building partial transcripts would involve a dramatically less number of such combinations.

Chapter 6: Conclusions

RNA-seq technologies provide us with tremendous opportunities to investigate the structure and abundance of transcripts, differential expression, structural variations, and more. It also delivers high throughput data within just a few days at progressively lower costs. This enables us to investigate genetic programs and cellular activities with precision, accuracy, and speed. However, in order to effectively use RNA-seq reads they generate, the sequencing technologies require new computational methods. In this thesis, I have designed novel algorithms and implemented several software systems to tackle these new challenges.

First, mapping reads to the genome is an essential step in RNA-seq analyses; the accuracy of mapping software can determine the accuracy of downstream steps such as gene and transcript discovery or expression quantification. I have developed TopHat2, which provides major improvements in accuracy over previous versions of TopHat and other RNA-seq mapping tools. In order to find the location information, reads may be aligned against the reference genome. However, RNA-seq reads pose new challenges because they may span multiple splice sites rather than just one or two. We estimate that nearly half of reads 150-bp long would span two or more human exons. The algorithmic improvements in TopHat2 address this challenge, maintaining both accuracy and speed. TopHat2 also avoids erroneously mapping reads to pseudogenes by making effective use of available gene annotations. This improves its overall alignment accuracy.

RNA-seq also enables us to discover structural variations, including genomic rearrangements. I have developed TopHat-Fusion, which detects fusion break points and map reads against them. Unlike previous approaches based on discordantly mapping paired reads and known gene annotations, TopHat-Fusion can find either individual or paired reads that span gene fusions, and it runs independently of known genes. This improves its sensitivity and enables it to find fusions including novel genes and novel splice variants of known genes. I have developed TransFUSE to further expand the analysis of fusion events by allowing the reconstruction and expression estimation of fusion transcripts. TransFUSE makes available more evidence, such as isoforms of fusion genes and estimates of their expression levels. As a result, we can put fusion candidates in order, those with more evidence first and those with less evidence after. This can help biologists quickly interpret the data and decide which fusions to address first. In contrast to previous approaches that simply provide a list of candidate fusions (genomic locations of break points), TransFUSE provides detailed information about full-length fusion transcripts such as exons, introns, and fusion break points. These capabilities enable one to infer the potential function of a fusion gene by examining the participating exons of the transcripts and their splicing patterns. Such analysis will help scientists identify the genetic basis of diseases. Expression levels of fusion genes may also provide additional insight when compared with those of normal transcripts from wild type genes.

I have shown that TopHat2, TopHat-Fusion, and TransFUSE perform well over a wide range of read lengths. This ability makes these programs a good fit for most RNA-seq experimental designs. As RNA-seq experiments are now widely used by many biologists, we expect that such experiments, in conjunction with these software systems, will provide scientists with accurate results for use with expression analysis, gene discovery, and a multitude of other applications.

Bibliography

1. McCutcheon JP, Moran NA: *Extreme genome reduction in symbiotic bacteria.* Nature reviews Microbiology 2012, **10:**13-26.

2. McGrath CL, Katz LA: *Genome diversity in microbial eukaryotes.* Trends in ecology & evolution 2004, **19:**32-38.

3. Pennisi E: *Genomics. ENCODE project writes eulogy for junk DNA.* Science 2012, **337:**1159, 1161.

4. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: *Alternative isoform regulation in human tissue transcriptomes.* Nature 2008, **456:**470-476.

5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods 2008, **5:**621-628.

6. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science 2008, **320:**1344-1349.

7. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell 2008, **133:**523-536.

8. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al: *Stem cell transcriptome profiling via massive-scale mRNA sequencing.* Nature methods 2008, **5:**613-619.

9. Maskos U, Southern EM: *Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ.* Nucleic acids research 1992, **20:**1679-1684.

10. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: *Yeast microarrays for genome wide parallel genetic and gene expression analysis.* Proceedings of the National Academy of Sciences of the United States of America 1997, **94:**13057-13062.

11. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: *Differential gene and transcript expression*

analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 2012, **7:**562-578.

12. Martin JA, Wang Z: *Next-generation transcriptome assembly.* Nature reviews Genetics 2011, **12:**671-682.

13. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al: *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature biotechnology 2011, **29:**644-652.

14. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature biotechnology 2012, **31:**46-53.

15. Anders S, Huber W: *Differential expression analysis for sequence count data.* Genome biology 2010, **11:**R106.

16. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al: *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.* Proceedings of the National Academy of Sciences of the United States of America 2003, **100:**15776-15781.

17. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: *Transcriptome sequencing to detect gene fusions in cancer.* Nature 2009, **458:**97-101.

18. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O: *Identification of fusion genes in breast cancer by paired-end RNA-sequencing.* Genome biology 2011, **12:**R6.

19. Kim D, Salzberg SL: *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome biology 2011, **12:**R72.

20. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology 2010, **28:**511-515.

21. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: *Improving RNA-Seq expression estimates by correcting for fragment bias.* Genome biology 2011, **12:**R22.

22. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al: *The GENCODE pseudogene resource.* Genome biology 2012, **13:**R51.

23. Trapnell C, Pachter L, Salzberg SL: *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics 2009, **25:**1105-1111.

24. Langmead B, Salzberg SL: *Fast gapped-read alignment with Bowtie 2.* Nature methods 2012, **9:**357-359.

25. Rowley JD: *Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.* Nature 1973, **243:**290-293.

26. de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR: *A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia.* Nature 1982, **300:**765-767.

27. Mitelman F JB, Mertens FE: *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.* http://cgapncinihgov/Chromosomes/Mitelman 2012.

28. Mitelman F, Johansson B, Mertens F: *The impact of translocations and gene fusions on cancer causation.* Nature reviews Cancer 2007, **7:**233-245.

29. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: *Transcription-mediated gene fusion in the human genome.* Genome research 2006, **16:**30-36.

30. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: *GenBank: update.* Nucleic acids research 2004, **32:**D23-26.

31. Li X, Zhao L, Jiang H, Wang W: *Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes.* Journal of molecular evolution 2009, **68:**56-65.

32. Kangaspeska S, Hultsch S, Edgren H, Nicorici D, Murumagi A, Kallioniemi O: *Reanalysis of RNA-Sequencing Data Reveals Several Additional Fusion Genes with Multiple Isoforms.* PloS one 2012, **7:**e48745.

33. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, et al: *FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.* Genome biology 2010, **11:**R104.

34. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, et al: *deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data.* PLoS computational biology 2011, **7:**e1001138.

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: *The Sequence Alignment/Map format and SAMtools.* Bioinformatics 2009, **25:**2078-2079.

36. Treangen TJ, Salzberg SL: *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nature reviews Genetics 2012, **13:**36-46.

37. Alkan C, Coe BP, Eichler EE: *Genome structural variation discovery and genotyping.* Nature reviews Genetics 2011, **12:**363-376.

38. Wu TD, Nacu S: *Fast and SNP-tolerant detection of complex variants and splicing in short reads.* Bioinformatics 2010, **26:**873-881.

39. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: *Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM).* Bioinformatics 2011, **27:**2518-2528.

40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics 2013, **29:**15-21.

41. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al: *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic acids research 2010, **38:**e178.

42. Zhang Z, Harrison PM, Liu Y, Gerstein M: *Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.* Genome research 2003, **13:**2541-2558.

43. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, et al: *Expressed pseudogenes in the transcriptional landscape of human cancers.* Cell 2012, **149:**1622-1634.

44. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, et al: *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell 2012, **148:**1293-1307.

45. *The Illumina Body Map 2.0 data* [http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513&expandefo=on]

46. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al: *The diploid genome sequence of an individual human.* PLoS biology 2007, **5:**e254.

47. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB: *Mobile elements create structural variation: analysis of a complete human genome.* Genome research 2009, **19:**1516-1526.

48. Langmead B, Trapnell C, Pop M, Salzberg SL: *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology 2009, **10:**R25.

49. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M: *Modelling and simulating generic RNA-Seq experiments with the flux simulator.* Nucleic acids research 2012.

50. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, et al: *Chimeric transcript discovery by paired-end transcriptome sequencing.* Proceedings of the National Academy of Sciences of the United States of America 2009, **106:**12353-12358.

51. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al: *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature 2008, **456:**53-59.

52. Ameur A, Wetterbom A, Feuk L, Gyllensten U: *Global and unbiased detection of splice junctions from RNA-seq data.* Genome biology 2010, **11:**R34.

53. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al: *De novo assembly and analysis of RNA-seq data.* Nature methods 2010, **7:**909-912.

54. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: *ABySS: a parallel assembler for short read sequence data.* Genome research 2009, **19:**1117-1123.

55. Kent WJ: *BLAT--the BLAST-like alignment tool.* Genome research 2002, **12:**656-664.

56. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V: *Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs.* Bioinformatics 2011, **27:**1068-1075.

57. Oshlack A, Robinson MD, Young MD: *From RNA-seq reads to differential expression results.* Genome biology 2010, **11:**220.

58. Barlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, Kallioniemi A: *Cloning of BCAS3 (17q23) and BCAS4*

(20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes, chromosomes & cancer 2002, **35:**311-317.

59. Zhao Q, Caballero OL, Levy S, Stevenson BJ, Iseli C, de Souza SJ, Galante PA, Busam D, Leversha MA, Chadalavada K, et al: *Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line.* Proceedings of the National Academy of Sciences of the United States of America 2009, **106:**1886-1891.

60. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W: *FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution.* Bioinformatics 2011, **27:**1922-1928.

61. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: *A large genome center's improvements to the Illumina sequencing system.* Nature methods 2008, **5:**1005-1010.

62. Quail MA, Swerdlow H, Turner DJ: *Improved protocols for the illumina genome analyzer sequencing system.* Current protocols in human genetics / editorial board, Jonathan L Haines [et al] 2009, **Chapter 18:**Unit 18 12.

63. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, Lee J, Jung YJ, Kim JO, Yu SB, et al: *The transcriptional landscape and mutational profile of lung adenocarcinoma.* Genome research 2012, **22:**2109-2119.

64. Needleman SB, Wunsch CD: *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of molecular biology 1970, **48:**443-453.

65. Smith TF, Waterman MS: *Identification of common molecular subsequences.* Journal of molecular biology 1981, **147:**195-197.

66. Farrar M: *Striped Smith-Waterman speeds database searches six times over other SIMD implementations.* Bioinformatics 2007, **23:**156-161.

67. Rognes T: *Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation.* BMC bioinformatics 2011, **12:**221.