Computational Tools and Databases for Fusion Transcripts: Therapeutic Targets in Cancer

6

Aditya Narayan, Bhavya Pahwa, and Shailesh Kumar

Abstract

In recent years, a rapid expansion in the field of RNAomics has led to a steep rise in data regarding expressed genes. This expansion in data has necessitated a consequent increase in the breadth and depth of tools which may be used for the study of RNA types. Gene fusions are considered hallmarks of many cancer types and may occur through chromosomal rearrangement or through noncanonical mechanisms in which chimeric RNA forms without rearrangement of the genome. To more effectively identify, validate, and understand the function of these novel RNA molecules, we present this chapter as a resource. In it, we discuss the role of fusion transcripts, identification of fusion transcripts, relevant software packages, and databases.

6.1 Introduction

Gene fusions are often considered to be a common feature present in cancer cells and present with rare cytogenetic signatures which may offer applications for disease identification, characterization, and treatment. Gene fusions are genes that possess DNA sequences from two different parental genes and may be created through

A. Narayan
University of Virginia, MR6 (Carter-Harrison Research Building), Charlottesville, VA, USA
e-mail: aditnara@stanford.edu

B. Pahwa
University College of Medical Sciences and GTB Hospital, New Delhi, Delhi, India

S. Kumar (✉)
Bioinformatics Lab, National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New Delhi, India
e-mail: shailesh@nipgr.ac.in

S. Singh (ed.), *Systems Biomedicine Approaches in Cancer Research*,
https://doi.org/10.1007/978-981-19-1953-4_6

several mechanisms including chromosomal translocations, inversions, deletions, and duplications. This may lead to proteins with domains derived from two genes in a novel fusion protein, a shift in reading frame, rearrangements of gene regulatory elements, and so on.

There has been a wide array of efforts to understand their prevalence, mechanism of creation, and function which have in turn lead to improvements in the ability to study several cancer subtypes. Broadly, well-studied examples of such gene fusions in cancer cells are described below:

- The first identified chromosome abnormality which was found to be strongly associated with cancer is BCR-ABL1, or the Philadelphia chromosome. This fusion of the BCR and ABL1 genes resulting from a reciprocal translocation event leads to the creation of a constitutively active tyrosine kinase (Ren 2005).
- In Burkitt's lymphoma, an aggressive mature B-cell neoplasia, chromosomal rearrangement leads to the creation of the IGH-MYC fusion and subsequent overexpression of the c-myc oncogene, a transcription factor which in turn leads to lymphomagenesis as well as accumulation of double-strand breaks in DNA (Yan et al. 2007).

While recurrent fusion genes are often associated with cancer phenotypes, fusion events are not necessarily limited to oncogenic processes. The formation of fusion genes in normal, noncancerous cells has been identified and has been shown to contribute to the development of more complex, multidomain proteins. This, in turn, contributes to protein evolution over longitudinal time scales.

While fusion genes are defined by the combination of DNA sequences, their precursor, chimeric RNAs, are hybrid RNA transcripts which contain nucleotides from different parental genes. These chimeric RNAs are not necessarily produced through the creation of the fusion of genes at the genomic level, and instead refer more broadly to any hybrid transcript based on gene annotations (Elfman et al. 2020). A critical reason for this distinction is that many means of chimeric RNA production in which there are no changes to the corresponding genome have been elucidated.

- Chimeric RNAs may be produced through the process of intergenic splicing. This most commonly occurs through a read-through of genes which lie in cis to create a hybrid mRNA. This is referred to as cis-splicing of adjacent genes (cis-SAGe) and has been found to be the primary way in which chimeric RNA production forms in noncancer cells (Singh et al. 2020).
- Chimeric RNA may form from parental genes which are found on different chromosomes. This process may be referred to as trans-splicing and is theorized to take place through splicing of precursor mRNAs (Jia et al. 2016).
- Parental genes may be separated by large linear distances on the same chromosomes.
- Trans-splicing of sense and antisense transcripts may occur between sense and antisense transcripts of a single gene.

- "Back-splicing" in which a downstream gene is transcribed prior to an upstream gene, leading to the creation of a circular fusion chimera (Wu et al. 2019).

Broadly, due to their ubiquity, varying mechanisms of formation, and diversity, chimeric RNAs are seen as a way in which the body expands the functional genome. Our understanding of chimer RNA identification and function has been expanded through large-scale analysis of datasets (GTEx, TCGA, Ensembl, etc.). The creation of such datasets has rapidly expanded in recent years given the advent of novel sequencing technologies which have improved researchers' access to critical insights. Despite this veritable explosion of databases and associated software tools, our knowledge of chimeric RNAs remains incomplete. Challenges which arise in the study of chimeric RNA are numerous, but not insurmountable. These are as follows:

- Relatively low levels of chimeric RNA expression may lead to biased statistical analyses leading to over or underestimation of certain sequences.
- The possibility of chimeric RNA developing from template switching events during RT-PCR.
- The only unique sequence in the chimeric RNA lies at the relatively small junction between the two parent sequences.
- Homology between chimeric RNA and parental genes causes bioinformatic predictive tools, biochemical techniques such as sequence-targeted assays, and the like to be particularly challenging.

Despite the barriers to studying this novel class of molecules, chimeric RNA is a fertile field of study for geneticists at all levels. To better support future generations of researchers in exploring this field, this chapter will explore software tools which may be applied to effectively identify and characterize chimeric RNA.

6.1.1 Identification of Chimeric RNA

A wide array of chimeric RNA prediction tools exists to support researchers in their search for potential candidates. These tools employ RNA-Seq datasets as purely genomic DNA datasets do not encompass the full potential for chimeric RNA production. This is since genomic instability is a hallmark of cancer cells. Further, RNA seq captures only the expressed parts of the genome (exome) which are transcriptionally active. This reduces the cost of the entire process to detect fusion transcripts in cancer. Datasets may be obtained from online databases such as TCGA. TCGA offers the GDC Data Transfer Tool to download raw sequencing data and we suggest that readers attempt to apply these tools for themselves. Raw data must first be processed by applying basic software tools to ensure their usage for various detection software.

- First, it is necessary to engage in quality control of raw sequencing reads. Example tools for this purpose are described below:
- ClinQC: This a highly accessible pipeline which may be used for converting between raw data formats, quality control, and trimming of raw sequencing data from both Sanger and NGS sequencing platforms. Data are converted to FASTQ format, which is among the most accepted formats for chimeric RNA prediction software which will be described later. The software prepares a quality control report which further facilitates downstream analysis (Pandey et al. 2016).
- NGSQC Toolkit: The NGSQC toolkit allows for highly efficient processing of NGS data with filtration of high-quality results as well as quality checking. It is an open-source application freely available online and implemented in perl. The toolkit offers high ease of use and is effective for sequencing data sourced from Rorche 454 and Illumina platforms (Patel and Jain 2012).
- FastQC: FastQC is one of the most applied tools to perform quality control screens on raw data from high-throughput sequencing methods. It allows for the import of data as BAM, SAM, or FastQ formats, creates summary graphics to assess data, and allows for export of data as various file types (Andrews n.d.).

Following quality control steps and conversion to file types appropriate for downstream analysis, it is possible to pass the data to fusion transcript prediction tools. There are over 35 software tools that are implicated in the identification of fusion transcripts. Fusion detection occurs in three stages, namely, (1) mapping and filtering, (2) fusion junction detection, and (3) fusion assembly and selection.

These mechanisms establish the categorical basis of division of tools for the identification of fusion transcripts.

1. *Mapping and filtering*: This is the initiation step in the identification of fusion transcripts and much software are based solely on this principle. After mapping is completed, pairs are evaluated for alignment and the irrelevant reads are removed. This is referred to as split mapping. Example software which applies split mapping are FusionMap and TopHat-Fusion. While some software, for example SnowShoes-FTD, utilizes spanning reads in which all mapped reads are preserved without filtration. Further incorrect reads are discarded by filtering techniques, as exemplified by FusionSeq with ten filters to remove illegitimate fusions. One such filter is when the fusion is intrachromosomal, such that the two genes are located on the same chromosome, and they can be recognized as a read-through transcript. This is applied by tools such FusionMap, FusionHunter, ShortFuse, SnowShoes-FTD, and TopHat-Fusion.
2. *Fusion junction detection*: This is the second step in fusion junction detection via "split read" mapping. It involves the independent alignment of first and last segments of the each "split read" that are generated by the discarding of unmapped reads in the previous stage. Alignment patterns are recognized, boundaries of the original fragments are adjusted, and realignment is performed to accurately identify fusion transcripts. Split read mapping is influenced by the size of partitioned segments. Small fragments not only sensitize the process but

are also more likely to provide false-positive results. To combat this, either the read is further split into two segments or a fixed "proposed" segment is utilized. Spanning reads facilitate the detection of fusion breakpoints followed by extraction of candidates by split read.

3. *Fusion assembly and selection*: In this step, mapped reads are referred to as "supporting reads". Owing to the presence of fusion junctions in the insert sequences, spanning reads also are good supporting modalities. Supporting reads are beneficial in the sense that they help in eliminating false candidates; however, the risk of true-positive results which are simply expressed at low levels being removed increases at the same rate. This problem is tackled by the availability of scoring functions in the tools. These functions are dependent on factors like, read depth, mapping quality, and number of supporting reads. Final scores are derived via empirical analysis (FusionSeq) or machine learning modalities (deFuse).

Here we describe the basic applications for commonly used tools.

6.1.1.1 FusionSeq

FusionSeq is a computational suite designed to detect candidate chimeric RNA/gene fusions through analysis of paired-end RNA seq data and offers high ease-of-use given that it is able to function irrespective of the mapping approach. The output of FusionSeq is a list of high confidence fusion candidates which are scored to provide for ease of follow-up validation studies. The results are accessible through a web browser. Drawbacks to the use of FusionSeq arise when considering the high CPU time and memory usage, particularly when analyzing large numbers of samples in parallel. This is because FusionSeq selects for all possible exons involved in the junction sequence and produces a junction library from all possible pairs of "tiles" which cover the exons and are each offset by one nucleotide. RNA-seq reads are mapped to these junctions but particularly with higher exon counts, this approach can be time consuming due to somewhat inefficient screening of false-positives (Sboner et al. 2010).

6.1.1.2 TopHat

TopHat is an algorithm to identify chimeric RNA transcripts representing fusion gene products. TopHat-Fusion is the most recent and updated version of this tool and offers the ability to align reads across fusion junctions. The software accepts and aligns RNA-seq reads but critically, does not rely on gene annotation. This is relevant as it allows for the tool to identify novel fusions which are derived from parental genes which are known, unknown, or unannotated variants of known genes (Kim and Salzberg 2011).

6.1.1.3 JAFFA

Frequently, methods used for the identification of chimeric RNA are designed for use with short read lengths. JAFFA is a software tool which compares cancer transcriptomes to references, as opposed to the genome and is optimized for read

lengths that are 100 bp or greater. The cancer transcriptome is inferred through long reads or de novo assembly of short reads.

JAFFA operates through a pipeline in which RNA-seq reads serve as the input and candidate fusion genes with breakpoint sequences are the output. Features include the presence of three modes which vary in appropriateness based on input read length: Assembly (wherein short reads are assembled de novo into contigs prior to detecting fusions), Direct (RNA-seq reads which do not map to known transcripts are employed), and Hybrid (combination of direct and assembly approaches) (Davidson et al. 2015).

6.1.1.4 EricScript

EricScript (chimERIC tranSCRIPT detection algorithm) is a tool for the detection of chimeric transcripts in paired-end RNA seq data (Benelli et al. 2012). This software differs from other prediction tools in that it is highly efficient due to its use of an exon junction reference which allows for reduced run times. Importantly, the package presents scores that allow for highly efficient detection of true from false-positive transcripts, which is a common challenge when distinguishing between potential fusions. For researchers, this scoring mechanism allows for efficient screening of potential output transcripts and allows for a reduced number of targets for data analyses. A study performed by Kumar et al. identified that EricScript was distinguished in its balance between time and memory requirements relative to sensitivity (Kumar et al. 2016).

6.1.1.5 SOAPfuse

SOAPfuse is an open-source tool that may be applied for the detection of fusion transcripts from paired-end RNA-seq data inputs. It can identify features of RNA-seq datasets such as insert size and read length, so full homogeneity of the dataset is less critical. This software was developed in perl and is limited in that it is only executable in Linux OS. SOAPfuse functions through alignment of RNA-seq paired-end reads against human reference sequences to detect candidate fusions. It employs both discordant mapping paired end reads as well as junction reads to confirm the sites. A junction library is constructed and is used to filter out false-positive fusions. The output of the program is a list of high likelihood fusions as well as their locations, junction sequences with single-nucleotide resolution, and diagrams displaying the varying location of reads relative to junction sequences and exon expression levels. This output data allows for effective follow-up analysis (Jia et al. 2013).

6.1.1.6 STARChip

A rapidly expanding field of research states that circular isoforms of RNA are expressed across the genome and may be correlated with disease. The value in detecting such nonlinear RNA alignments lies in the fact that it allows for more rapid detection of chromosomal rearrangements which are commonly associated with cancer. STAR Chimeric Post (STARChip) is a software package which applies the STAR aligner to chimeric alignments in order to produce annotated circRNA and

fusions. This tool is effective for high-dimensional datasets and offers high performance at relatively low computing time (Akers et al. 2018).

6.1.1.7 FuSeq

FuSeq is a fusion detection method which applies a recent quasi-mapping method for alignment which allows it to operate with far lower computational time than many other tools. The tool functions through a pipeline for mapped read-pairs and another junction split-reads. Following the process, false-positive results are minimized through application of a range of filters (Vu et al. 2018). Additional tools are summarized in Table 6.1.

6.2 Fusion Transcripts Databases

There are several fusion transcripts databases available for scientific community. Almost all these resources are freely available, harboring the information of fusion coordinates, tissue, condition, sample information, cancer type, etc. Our research group also developed a database of fusion transcripts for model plant *Arabidopsis thaliana*. Most popular fusion transcripts databases are mentioned in the Table 6.2.

6.3 Validation of Transcripts

Following the generation of potential fusion transcript lists, there is a wide range of possible approaches to validating the chimeric RNA and ensuring that they are not false-positive results. Some of the most readily applied approaches are described below:

- In-Silico Validation: By utilizing the predicted junction sequence at the breakpoint between parental gene sequences, it is possible to identify commonly expressed chimeric RNA. This is performed by searching for the junction sequence in the raw RNA sequencing reads using string-matching software.
- Validation Through Query of Online Databases: Databases containing chimeric expressed sequence tags and junction sequences and may be queried for certain sequences to determine if they have been previously validated. Table 6.2 describes some of these databases.
- Application of Wet Lab Approaches: Reverse-transcription polymerase chain reaction (RT-PCR) may be used to detect and measure the expression of chimeric RNA transcripts. After isolating RNA from a sample and creating cDNA, PCR may be applied specifically to the junction sequence to determine expression levels. Primers may be designed such that they flank this unique junction sequence and allow the researcher to amplify the sequence if present.

Table 6.1 Summarizes the tools available to identify fusion transcripts along with their methodologies

Method	Brief overview of methodology
Arriba (Uhrig 2019)	Arriba extracts gene fusions from the chimeric alignments reported by STAR (Dobin et al. 2013) by applying a collection of filters which recognize frequent types of artifacts found in RNA-Seq data
ChimeraScan (Iyer et al. 2011)	Identifies candidate fusions from discordant Bowtie (Langmead et al. 2009) genome alignments. Unmapped reads are trimmed and realigned. Junction breakpoint reads are resolved by aligning to candidate fused exons. Fusions are filtered based on an abundance of fusion-supporting reads
ChimPipe (Rodriguez-Martin et al. 2017)	The GEMtools RNA-seq pipeline (GEMTools 2019) and GEM alignment utility (Marco-Sola et al. 2012) are used to capture discordant and chimeric read alignments, and fusion candidates are filtered according to fusion evidence and additional gene-based filters
deFuse (McPherson et al. 2011)	Aligns reads to spliced and unspliced gene sequences using Bowtie (Langmead et al. 2009), resolves split read junctions using a novel dynamic programming algorithm, and uses an AdaBoost classifier to discriminate between likely true versus false fusions
EricScript (Benelli et al. 2012)	BWA (Li and Durbin 2009) is used to align reads to the genome. Discordant reads are used to identify candidate gene fusions. BLAT (Kent 2002) is then used in an iterative local alignment step to define precise fusion breakpoints by aligning to customized targets of fused exons. An AdaBoost classifier trained with synthetic data is used to score and rank fusion predictions
FusionCatcher (Nicorici et al. 2014)	Leverages a collection of alignment utilities including Bowtie (Langmead et al. 2009), Bowtie2 (Langmead and Salzberg 2012), BLAT (Kent 2002), and STAR (Dobin et al. 2013) with a collection of customized target databases to identify and characterize fusion candidates. Rigorous filtering of fusion predictions according to gene and fusion annotations is employed
FusionHunter (Li et al. 2011)	First uses Bowtie to align reads to the genome and identify candidate fusions based on discordant read pairs. Then creates a "pseudoreference" by positioning candidate fusion genes with canonical ordering, realigns reads using a custom algorithm, and identifies both split and spanning reads providing evidence for gene fusions
InFusion (Okonechnikov et al. 2016)	Reads are first aligned to the reference transcriptome using Bowtie2. Unaligned and discordantly aligned reads are further examined in the context of the genome and transcriptome to cluster evidence and define candidate fusions
JAFFA-Assembly (Davidson et al. 2015)	After removing intronic and intergenic region aligning reads defined by Bowtie genome alignments, the remaining reads are assembled using Oases (Schulz et al. 2012) and the assembled contigs are mapped directly to the transcriptome using BLAT. Chimeric BLAT alignments are further assessed as fusion candidates

(continued)

Table 6.1 (continued)

Method	Brief overview of methodology
JAFFA-Direct (Davidson et al. 2015)	After removing intronic and intergenic region aligning reads defined by Bowtie genome alignments, the remaining reads are mapped directly to the transcriptome using BLAT. Chimeric BLAT alignments are further assessed as fusion candidates
JAFFA-Hybrid (Davidson et al. 2015)	After removing intronic and intergenic region aligning reads defined by Bowtie genome alignments, the remaining reads are assembled using Oases. Both the assembled transcripts and the original reads that failed to map to the genome are then mapped directly to the transcriptome using BLAT. Chimeric BLAT alignments are further assessed as fusion candidates
MapSplice (Wang et al. 2010)	An RNA-seq aligner based on Bowtie similar to TopHat (Trapnell et al. 2009) and includes fusion-finding capabilities, although specific algorithmic details are lacking
nFuse (McPherson et al. 2012)	Designed for use with WGS-seq and RNA-seq but can be executed with RNA-seq only, leveraging its included deFuse with Bowtie2
Pizzly (Melsted et al. 2017)	Uses a k-mer-based strategy to examine reads that do not map to isoforms consistently via kallisto (Bray et al. 2016) pseudoalignment
PRADA (Torres-Garcia et al. 2014)	Reads are aligned to a combined genome and transcriptome reference using BWA. Discordant reads identify fusion candidates, and junction reads are identified by mapping to a database of all possible 5′-3′ chimeric exon junction database
SOAP-fuse (Jia et al. 2013)	The SOAP2 aligner (Hurgobin 2016) is used to map reads to genomes and spliced transcripts to identify fusion candidates
STARChip (Akers et al. 2018)	Uses chimeric reads reported by STAR aimed primarily at identifying circular RNAs but also reports fusion candidates
STAR-Fusion (Haas 2019a)	Uses chimeric read alignments reported by STAR in its Chimeric.out.junction file to identify candidate fusions followed by extensive filtering of likely artifacts
STAR-SEQR (STAR-SEQR 2019)	Uses chimeric reads reported by STAR to find fusions
TopHat-Fusion (Kim and Salzberg 2011)	A modified execution of the TopHat aligner (Trapnell et al. 2009; Kim et al. 2013) to examine initially unmapped reads as supporting fusion events
TrinityFusion-C (Haas 2019b)	De novo assembles only the chimeric reads defined by STAR using the Trinity assembler (Tomczak et al. 2015), and subsequently leverages GMAP (Jang et al. 2020; Kim and Zhou 2019) for chimera candidate detection
TrinityFusion-D (Haas 2019b)	De novo assembles all input reads using Trinity, and subsequently leverages GMAP for chimera candidate detection
TrinityFusion-UC (Haas 2019b)	De novo assembles both chimeric and unmapped reads defined by STAR using the Trinity assembler, and subsequently leverages GMAP for chimera candidate

Table 6.2 Different databases available to identify fusion transcripts along with their methodologies

Database	Brief overview of database
The Cancer Genome Atlas (TCGA) (Tomczak et al. 2015)	TCGA seeks to create a comprehensive profile of genomic alterations associated with cancers through profiling human tumor cohorts
ChimerDB (Jang et al. 2020)	ChimerDB is one of the most comprehensive databases available for the study of gene fusions. It includes deep sequencing data as well as information from publications
Fusion Gene Annotation Database (FusionGDB) (Kim and Zhou 2019)	FusionGDP provides functional annotations as well as information on protein structure, fusion transcript amino acid sequences, breakpoint mapping, and the like for a range of known fusion genes
FusionCancer (Wang et al. 2015)	FusionCancer is a database based on gene fusion identification from RNA-seq datasets in human cancers. This is a query engine with annotated information of cancer fusion genes and which offers high ease of use for researchers
FusionHub (Panigrahi et al. 2018)	This is a web platform which allows for querying of multiple gene fusion databases. It allows for multiple visualization approaches and allows for ease of annotation
AtFusionDB (Singh et al. 2019)	AtFusionDB is a comprehensive database which contains fusion transcript information specific to Arabidopsis thaliana. There are a variety of annotation tools, search modules, and visualization approaches which facilitate the study of plant genomes
ChiTaRS (Balamurali et al. 2020)	ChiTaRS is an incredibly comprehensive chimeric transcript database with annotated information from eight species' genomes. A number of features exist within the database including information on druggable fusion targets and transcripts with clinical correlates

6.4 Conclusion

In the years following the discovery of the Philadelphia chromosome, there has been an explosion of evidence supporting gene rearrangements as correlates and/or causative agents of oncogenesis. This has led to databases, software tools, and biochemical techniques which have allowed for increasingly efficient and effective analysis of the novel field of chimeric RNA production. While the majority chimeric RNA is of unclear functional significance, advances in genomic editing approaches may expand the potential for novel explorations of the functional genome.

References

Akers NK, Schadt EE, Losic B (2018) STAR chimeric post for rapid detection of circular RNA and fusion transcripts. Bioinformatics 34(14):2364–2370. https://doi.org/10.1093/bioinformatics/bty091

Andrews S (n.d.) FASTQC: a quality control tool for high-throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Balamurali D, Gorohovski A, Detroja R, Palande V, Raviv-Shay D (2020) Milana Frenkel-Morgenstern. ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. Nucleic Acids Res 48(1):825–834. https://doi.org/10.1093/nar/gkz1025

Benelli M et al (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics 28(24):3232–3239

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527

Davidson NM, Majewski IJ, Oshlack A (2015) JAFFA: high sensitivity transcriptome-focused fusion gene detection. Genome Med 7(1):43. https://doi.org/10.1186/s13073-015-0167-x

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21

Elfman J, Pham L-P, Li H (2020) The relationship between chimeric RNAS and gene fusions: Potential implications of reciprocity in cancer. J Genet Genomics 47(7):341–348

GEMTools (2019) GEMTools 2019. Available from http://gemtools.github.io/

Haas BJ (2019a) STAR-fusion code and documentation on GitHub. Available from https://github.com/STAR-Fusion/STAR-Fusion/wiki

Haas BJ (2019b) TrinityFusion - fusion and foreign transcript detection via RNA-seq de novo assembly. Available from https://github.com/trinityrnaseq/TrinityFusion/wiki

Hurgobin B (2016) Short read alignment using SOAP2. Methods Mol Biol 1374:241–252

Iyer MK, Chinnaiyan AM, Maher CA (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics 27(20):2903–2904

Jang YE, Jang I, Kim S, Cho S, Kim D, Kim K, Kim J, Hwang J, Kim S, Kim J, Kang J, Lee B, Lee S, Chimer DB (2020) 4.0: an updated and expanded database of fusion genes. Nucleic Acids Res 48(1):817–824

Jia W, Qiu K, He M et al (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol 14:12. https://doi.org/10.1186/gb-2013-14-2-r12

Jia Y, Xie Z, Li H (2016) Intergenically spliced chimeric RNAs in cancer. Trends Cancer 2(9):475–484. https://doi.org/10.1016/j.trecan.2016.07.006

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12(4):656–664

Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12(8):72. https://doi.org/10.1186/gb-2011-12-8-r72

Kim P, Zhou X (2019) FusionGDB: fusion gene annotation DataBase. Nucleic Acids Res 47(1):994–1004. https://doi.org/10.1093/nar/gky1067

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36

Kumar S, Vo A, Qin F et al (2016) Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep 6:21597. https://doi.org/10.1038/srep21597

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):25

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760

Li Y, Chien J, Smith DI, Ma J (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 27(12):1708–1710

Marco-Sola S, Sammeth M, Guigo R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9(12):1185–1188

McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG et al (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 7(5):e1001138

McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC (2012) nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res 22(11):2250–2261

Melsted P, Hateley S, Joseph IC, Pimentel H, Bray N, Pachter L (2017) Fusion detection and quantification by pseudoalignment. bioRxiv. 2017:166322. https://doi.org/10.1101/166322

Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, Virtanen S, Kilkku O et al (2014) FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 011650. https://doi.org/10.1101/011650

Okonechnikov K, Imai-Matsushima A, Paul L, Seitz A, Meyer TF, Garcia-Alcalde F (2016) InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data. PLoS One 11(12):e0167417

Pandey RV, Pabinger S, Kriegner A et al (2016) ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. BMC Bioinf 17:56. https://doi.org/10.1186/s12859-016-0915-y

Panigrahi P, Jere A, Anamika K (2018) FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. PLoS ONE 13(5):e0196588. https://doi.org/10.1371/journal.pone.0196588

Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE 7(2):e30619. https://doi.org/10.1371/journal.pone.0030619

Ren R (2005) Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. Nat Rev Cancer 5(3):172–183. https://doi.org/10.1038/nrc1567

Rodriguez-Martin B, Palumbo E, Marco-Sola S, Griebel T, Ribeca P, Alonso G et al (2017) ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data. BMC Genomics 18(1):7

Sboner A, Habegger L, Pflueger D et al (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11:R104. https://doi.org/10.1186/gb-2010-11-10-r104

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092

Singh A, Zahra S, Das D, Kumar S (2019) AtFusionDB: a database of fusion transcripts in Arabidopsis thaliana. Database 2019:135. https://doi.org/10.1093/database/bay135

Singh S, Qin F, Kumar S, Elfman J, Lin E, Pham LP, Yang A, Li H (2020) The landscape of chimeric RNAs in non-diseased tissues and cells. Nucleic Acids Res 48(4):1764–1778

STAR-SEQR (2019) STAR-SEQR code and documentation on GitHub 2019. Available from https://github.com/ExpressionAnalysis/STAR-SEQR

Tomczak K, Czerwińska P, Wiznerowicz M (2015) The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 19(1):68–77. https://doi.org/10.5114/wo.2014.47136

Torres-Garcia W, Zheng S, Sivachenko A, Vegesna R, Wang Q, Yao R et al (2014) PRADA: pipeline for RNA sequencing data analysis. Bioinformatics 30(15):2224–2226

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111

Uhrig S (2019) Arriba - fast and accurate gene fusion detection from RNA-Seq data. Available from https://github.com/suhrig/arriba

Vu T, Deng W, Trac Q et al (2018) A fast detection of fusion genes from paired-end RNA-seq data. BMC Genomics 19:786. https://doi.org/10.1186/s12864-018-5156-1

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL et al (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38(18):178

Wang Y, Wu N, Liu J et al (2015) FusionCancer: a database of cancer fusion genes derived from RNA-seq data. Diagn Pathol 10:131. https://doi.org/10.1186/s13000-015-0310-4

Yan Y, Park SS, Janz S, Eckhardt LA (2007) In a model of immunoglobulin heavy-chain (IGH)/MYC translocation, the Igh 3′ regulatory region induces MYC expression at the immature stage of B cell development. Genes Chromosomes Cancer 46(10):950–959. https://doi.org/10.1002/gcc.20480

Wu K, Liao X, Gong Y et al (2019) Circular RNA F-circSR derived from SLC34A2-ROS1 fusion gene promotes cell migration in non-small cell lung cancer. Mol Cancer 18(1):98. https://doi.org/10.1186/s12943-019-1028-9