Journal of
Next Generation Sequencing & Applications

Zhang, et al., Next Generat Sequenc & Applic 2016, 3:3
DOI: 10.4172/2469-9853.1000140

Review Article Open Access

Bioinformatics Tools for RNA-seq Gene and Isoform Quantification

Chi Zhang[1], **Baohong Zhang**[1], **Michael S Vincent**[2] **and Shanrong Zhao**[1*]

[1]*Early Clinical Development, Pfizer Worldwide R&D, Cambridge, MA, USA*

[2]*Inflammation and Immunology RU, Pfizer Worldwide R&D, Cambridge, MA, USA*

[*]**Corresponding author:** Shanrong Zhao, Early Clinical Development, Pfizer Worldwide R&D, Cambridge, MA, 02139, USA, Tel: + 1-212-733-2323; E-mail: Shanrong.Zhao@pfizer.com

Rec date: Oct 27, 2016; **Acc date: Dec 15**, 2016; **Pub date:** Dec 17, 2016

Abstract

In recent years, RNA-seq has emerged as a powerful technology in estimation of gene or transcript expression. 'Union-exon' and transcript based approaches are widely used in gene quantification. The 'Union-exon' based approach is simple, but it does not distinguish between isoforms when multiple alternatively spliced transcripts are expressed from the same gene. Because a gene is expressed in one or more transcript isoforms, the transcript based approach is more biologically meaningful than the 'union exon'-based approach. However, estimating the expression of individual isoform is intrinsically challenging because different isoforms of a gene usually have a high proportion of genomic overlap. Recently, a number of tools have been developed for RNA-seq isoform quantification. We review those methods and their main features to serve as guidance for users to choose suitable tools for their specific projects.

Keywords: Bioinformatics tools; RNA-seq gene; Spliced transcripts; Isoforms; Union-exon; Human transcriptome; Annotation databases

Introduction

Production of distinct mRNA isoforms from the same gene locus is common phenomena in metazoans. According to gene annotation in GENCODE Release 25, there are around 60,000 annotated genes, including protein coding, lncRNA and processed pseudogene, which produce about 200,000 transcripts in the human transcriptome (http://www.gencodegenes.org). Of those annotated genes, 20,000 are protein coding genes, which in turn produce about 144,000 transcripts. The current annotation gives an estimate of 7 transcript isoforms per protein coding gene; however, the annotation is far from complete. A recent study suggests that more than 1/3 of tissue dependent transcripts have complex local splicing variations (LSVs), where an exon can be involved in more than two alternative junctions [1]. These LSVs can produce a large number of alternatively spliced isoforms per gene locus, generating a much more diverse human transcriptome than previously estimated.

Alternative splicing is a process by which exons or portions of exons or noncoding regions within a pre-mRNA transcript are differentially joined or skipped, resulting in multiple isoforms being encoded by a single gene. The process of splicing occurs in a large ribonucleoprotein (RNP) machine called the spliceosome, which functions in a dynamic assembly-disassembly cycle involving five small nuclear ribonucleoprotein (snRNP) complexes (U1, U2, U4/U6, and U5).

New insights suggest that constitutive splicing primarily occurs co-transcriptionally in the nucleus, whereas alternative splicing mainly occurs post-transcriptionally [2-4]. Alternative splicing generates a tremendous amount of proteomic diversity in humans and significantly affects various functions in cellular processes, tissue specificity, developmental states, and disease conditions. Errors in alternative splicing can lead to various diseases, including muscle disorders [5,6] and cancers [7-9], emphasizing the need to accurately quantify isoform expression.

RNA sequencing (RNA-seq) is emerging as a new technology in transcriptome profiling. Beyond quantifying gene expression, the data generated by RNA-seq facilitates discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele specific expression [10-12]. Compared with microarray technology, RNA-seq not only overcomes some of the technical limitations including varying probe performance and nonspecific hybridization, but can also detect alternative splicing isoforms and subtle changes of splicing under different physiological conditions [13].

Furthermore, RNA-seq allows for the detection of novel transcript species in well studied organisms, such as unique transcripts in certain tissues or in rare cell types, and has been instrumental to catalog the diversity of novel transcript species including long non-coding RNA, miRNA, siRNA, and other small RNA classes [14].

Gene Quantification

Recently, quite a number of methods have been developed for the inference of gene and isoform abundance. In general, these methods can be divided into two categories: 'union exon'-based approach and transcript-based approach, as illustrated in Figure 1. The 'union exon'-based methods, such as FeatureCounts and HTSeq is widely used in RNA-seq gene quantification because of its simplicity [15,16].

First, all overlapping exons from a gene are merged into 'union exons'. Then, per-gene read counts are aggregated by intersecting all mapped reads with 'union exons' of the gene. Compared with transcript level quantification, reads can generally be assigned to genes with higher confidence. However, gene-level counts do not distinguish between isoforms when multiple alternatively spliced transcripts are expressed from the same gene. Consequently, some important events such as isoform switches and minor isoform expression changes are masked at the gene level, as shown in Figure 2. For UHRR (Universal

Citation: Zhang C, Zhang B, Vincent MS, Zhao S (2016) Bioinformatics Tools for RNA-seq Gene and Isoform Quantification. Next Generat Sequenc & Applic 3: 140. doi:10.4172/2469-9853.1000140

Human Reference RNA), both transcripts are expressed, and the expression of the short isoform ENST00000375512 is supported by 48 exon-exon spanning reads between exons #7 and #8 (colored in indigo). However, in human brain sample HBRR_C4, only the long transcript ENST00000356884.6 is expressed.

At the gene level, overall, there is not much difference between sample UHRR_C1 and HBRR_C4. As a result, the expression change in minor isoform ENST00000375512 is masked at the gene level. Therefore, ignoring transcript isoforms and counting reads at the gene level does not always give correct answers [17,18].

The side-by-side comparison of 'union exon'-based approach and transcript-based method revealed that gene expression levels are significantly underestimated by 'union exon'-based approach in some cases, and the average of RPKM from 'union exons'-based method is less than 50% of the mean expression obtained from transcript-based approach [19]. Consequently, the 'union exons'-based approach in gene quantification is discouraged despite of its popularity. Other studies also suggest that tools that quantify expressions at the transcript level give better results than at the gene level [20].

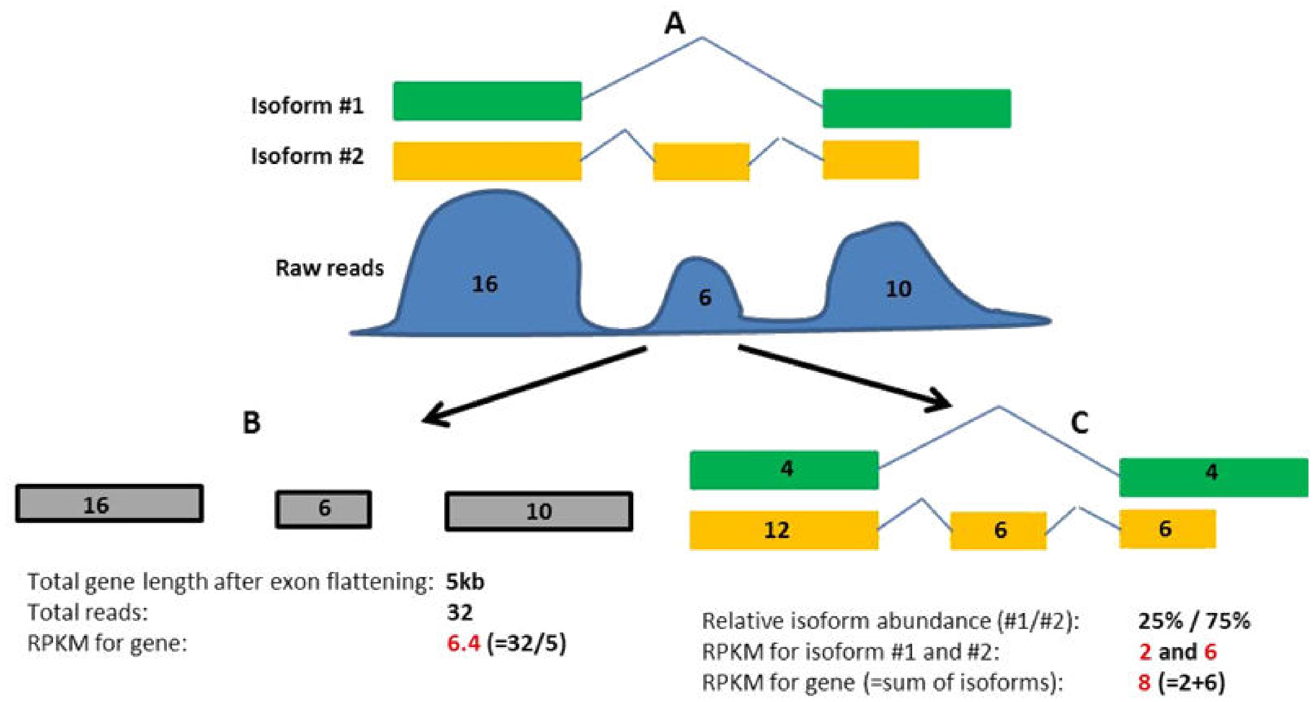

Figure 1: Illustration of different gene quantification methods. A) A hypothetical gene, its two isoforms and read coverage profile. Assuming that the sum of mapped reads from all genes is 1 million, and each small and large exon is 1 kb and 2 kb long, respectively. B) 'Union-exon' based approach. After exon flattening, the 'union exons' are 2 kb, 1 kb and 2 kb long, respectively. The calculated RPKM is 6.4. C) transcript-based approach. Mapped reads are first assigned to individual isoforms, and the corresponding expressions for the two isoforms are 2 RPKM and 6 RPKM, respectively. In this calculation, the entire gene expression is 8 RPKM.

Multiple mRNA transcripts can be generated from a gene locus by the usage of alternative transcription start site, alternative splicing and alternative polyadenylation. Different isoforms of the gene typically have a high proportion of overlapping exons (Figure 3). The transcriptome landscape is further complicated by the prevalence of gene overlap on the same or opposite strands in DNA [21].

Therefore, accurate estimation of expression levels of individual isoforms is intrinsically very difficult. In the following sections, we will review methods available for isoform quantification with a focus on recent advancement in this field. We classify the methods into two main categories: tools that consider only known transcripts and those that incorporate novel transcript discovery.

Isoform Quantification of Known Transcripts

If the transcriptome of a species is annotated, its annotation databases can be leveraged to map and quantify the expression of most common transcripts. For the human genome, RefGene [22], Ensembl [23] and the UCSC [24] annotation database are most frequently used. The choice of gene annotation can have a significant impact on quantification even at the gene level [25]. In addition to annotation databases, it is possible to incorporate novel transcripts discovered from *de novo* assembly into existing annotations. One of the most popular assemblers is the TRINITY package [26].

A large number of RNA-seq specific mapping algorithms have been developed to align large numbers of sequence reads to a reference genome and/or transcriptome, including Bowtie [27,28], TopHat2 [29], STAR [30], GSNAP [31], and Map Splice [32]. Most of the tools require SAM/BAM files as inputs, which are generated by the aligners described above. Some new isoform quantification tools have their own built-in mapping algorithms and can take sequencing reads as inputs directly, instead of aligned SAM/BAM files [33-36].

Nearly all isoform quantification algorithms use either maximum likelihood (ML) estimate or Bayesian inference and expectation-maximization (EM) methods to assign ambiguously mapped reads to transcript isoforms, and they differ mainly in EM convergence speed and some other important features.

Table 1 summarizes these tools and their features, including run time, bias correction, multi-threading, stranded protocol support and indel read alignments. The output of these tools usually reports the uncertainty of isoform quantification as well. By appropriately accounting for uncertainty in quantification, more accurate downstream differential analyses were obtained at both the gene and isoform levels. Sleuth [18] is a downstream analyses tool specifically developed for differential analyses at the transcript level.

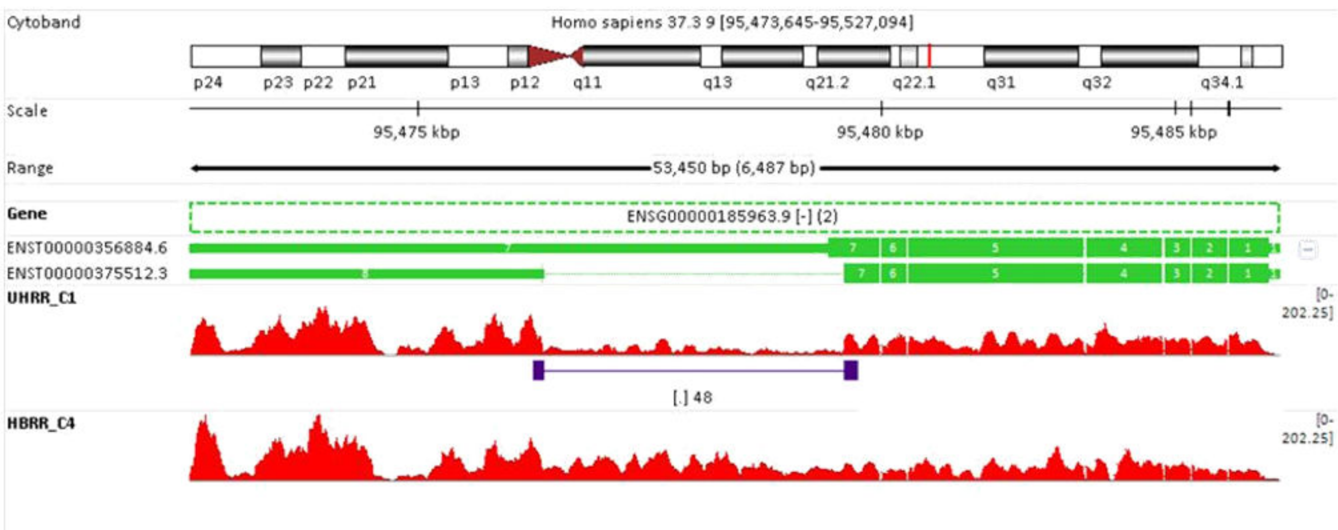

Figure 2: Minor isoform changes are masked at the gene level. Gene BICD2 (Ensembl ID: ENSG00000185963.9) consists of two very similar isoforms. At the gene level, overall, there is not much difference between sample UHRR_C1 and HBRR_C4, as shown in read coverage profiles. However, the difference is dramatic at the transcript level. For human brain sample HBRR_C4, only the long transcript ENST00000356884.6 is expressed; while in sample UHRR_C1, both isoforms are present. Note in sample UHRR_C1, there are 48 reads that span across the junction site between exons #7 and #8 (colored in indigo), and such reads can only originate from the short transcript ENST00000375512. The stranded RNA-seq dataset were downloaded from Illumina's Base Space, consisting of 2 UHRR (Universal Human Reference RNA) and 2 HBRR (Human Brain Reference RNA) samples. Only UHRR_C1 and HBRR_C4 are shown in the figure.

RSEM is an accurate and user-friendly software tool for quantifying transcript abundances from RNA-seq data [37]. It estimates the ML of relative abundances of the transcript isoforms and then fractionally assigns reads to the isoforms based on these abundances. The assignments of reads to isoforms come from iterations of EM method. According to the comparative assessment [20], RSEM is relatively slow. Recent tools, such as eXpress [38], aim to reduce the computational burden of isoform quantification by substantially altering the EM algorithm.

It improves the convergence speed using an online-EM algorithm that models indels, fragment length, sequencing errors and corrects sequence-specific fragment biases. Tigar2 utilizes Bayesian inference as an alternative to ML estimation and provides better accuracy for longer reads and supports for variable-length reads produced by Ion Torrent PGM sequencer [39,40]. However, Tigar2 is by far the slowest algorithm, taking more than 9 hours to process 100 million reads even with multi-thread processing [20]. BitSeq also uses Bayesian inference and a user can choose either markov chain monte carlo (MCMC) or variational Bayesian (VB) methods for quantification [41,42]. MCMC is generally slower than VB method, but provides better accuracy.

For the isoform quantification, not all information in the alignment is necessary. Simply the knowledge of the transcripts and positions to which a given read maps is sufficient for quantification purpose. In support of such 'analysis-efficient' computation, the concept of quasi-mapping, lightweight-alignment or pseudo alignment has been introduced recently [33-36]. With appropriate optimization, such tools can map and quantify 100 million reads in 10 minutes even on a laptop. Sailfish was initially implemented using a k-mer approach [35].

It entirely skips read mapping, a time-consuming step, and provides quantification estimates much faster than existing approaches. However, shredding reads into k-mers discards valuable information present in whole reads, and accordingly, results in loss of accuracy, since each k-mer can potentially align to more transcripts than the read itself. To circumvent this issue, Kallisto uses fast hashing of k-mers together with the transcriptome de Bruijn graph to produce a list of transcripts that are compatible with each read while avoiding alignment of individual bases [36].

K-mer based Sailfish has now been deprecated, and the latest Sailfish uses the quasi-mapping procedure provided by RapMap [34], as opposed to individual k-mer counting, for transcript-level quantification to increase accuracy without sacrificing speed.

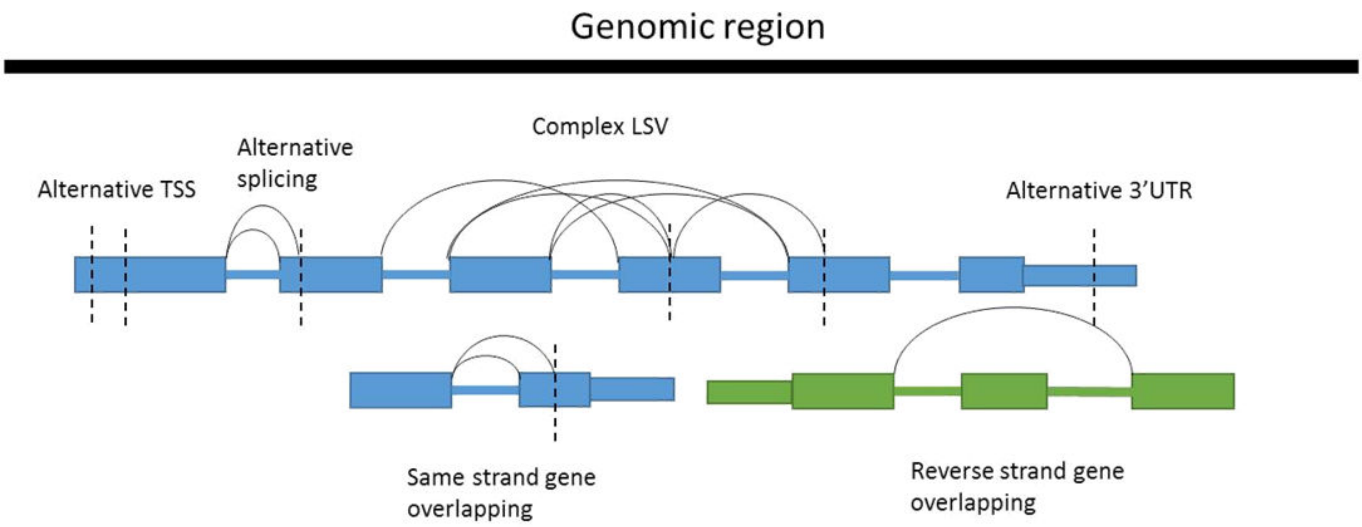

Figure 3: Diagram of the complexity of gene structures. Multiple isoforms generated by alternative splicing, together with overlapping genes, make accurate quantification at the transcript level difficult. Assigning ambiguous reads in overlapping regions to its true origin is hard. Genes on the plus strand are colored in blue while genes on the minus strand in green.

	Run time	Algorithm	Bias correction	Stranded protocol	Multi-thread support	Indel alignments
Cufflinks	+++	ML	No	Yes	Yes	Yes
RSEM	+++	ML	No	Yes	Yes	No
eXpress	++	ML	Yes	Yes	No	Yes
Tigar2	+++++	VB	No	No	Yes	Yes
BitSeq	VB:+++ MCMC:++++	MCMC/ VB	Yes	No	Yes	Yes
Kallisto	+	ML	Yes	Yes	Yes	Yes
Salmon	+	VB/ML	Yes	Yes	Yes	Yes
Sailfish	+	VB/ML	Yes	Yes	Yes	Yes

Note: The run time, algorithm used, support for bias correction, stranded library preparation protocol, multi-thread computing and indel alignments of popular tools are compared. For a sample with 50 million reads running on a typical HPC with multi-thread computing turned on, run time over 8 hours is considered "+++++", less than 3 hours but more than 1 hour is considered "+++", less than 30 minutes but more than 10 minutes is considered "++", and less than 10 minutes is considered "+"

Table 1: Features of isoform quantification packages.

Interestingly, the developer of Sailfish introduces Salmon [33], another novel method and software tool for isoform quantification that exhibits state-of-the-art accuracy while being significantly faster than most other tools. Salmon achieves this through a two-phase inference procedure, a reduced data representation, and the quasi-mapping algorithm of RapMap. During its online phase, in addition to performing streaming inference of transcript abundances, Salmon also constructs a highly-reduced representation of the sequencing experiment.

Specially, Salmon constructs "rich" equivalence classes over all of the sequenced fragments to greatly reduce the time required to perform iterative optimization. Both Salmon and Sailfish allow users to choose between ML estimate and VB inference for isoform quantification. However, there are a few main differences between Sailfish and Salmon. One is that Salmon implements a dual-phase inference algorithm, consisting of both an online and offline phase, while Sailfish uses only an offline algorithm. Salmon also accepts alignment files in SAM/BAM format, making it a flexible tool for isoform quantification, but Sailfish does not. Another difference is that Salmon contains richer models for bias correction, whereas Sailfish does not.

Isoform Quantification in Conjunction with Novel Transcript Discovery

Transcriptome analysis from RNA-seq data typically involves two sub-problems, i.e. identification of the set of isoforms and estimation of the abundance of these isoforms. If no annotation for a species of interest is available or novel transcript discovery is desired, isoform structures have to be constructed from RNA-seq data first [43]. One popular approach is to include novel transcript discovery and quantification in the same package. These tools include Cufflinks [44], Scripture [45], IsoLasso [46], NSMAP [47], SLIDE [48], iReckon [49], Traph [50], MiTie [51], and FlipFlop [52]. These tools either assemble the transcriptome ab initio, or use existing annotations to infer new splicing junctions. The discovered novel transcripts can then be used or added to existing annotations for quantification. However, recent reviews indicate that the quantification algorithms provided by these tools are not on par with tools from the previous category [20,53]. This is because identification of isoforms from RNA-seq data is far from being solved and is still challenging, due in particular to the incomplete nature of RNA-seq reads and the fact that the number of potential candidate isoforms is very large, growing almost exponentially with the number of exons. As a result, the performance reported by the state-of-the-art algorithms is often unsatisfactory.

Cufflinks is perhaps one of the most popular tools for novel transcripts discovery and quantification. It assembles transcripts ab initio and merges them with existing annotations and then quantify the transcripts [44]. In a sense, the quantification strategy in Cufflinks is similar to one iteration of the EM algorithm used in RSEM [37]. During assembly, Cufflinks attempts to explain the observed reads with minimum number of isoforms. iReckon [49] is introduced for simultaneous determination of the isoforms and estimation of their abundances. Their probabilistic approach incorporates multiple biological and technical phenomena, including novel isoforms, intron retention, unspliced pre-mRNA, PCR amplification biases, and multimapped reads. iReckon utilizes regularized EM to accurately estimate the abundances of known and novel isoforms [49].

The strategy of simultaneously discovering and quantifying transcripts is also adopted by many other state-of-the-art methods (e.g. SLIDE [48], StringTie [54], IsoLasso [46] and CIDANE [55]. Similar to iReckon, SLIDE requires existing annotations and cannot perform *de novo* assembly. A unique advantage of SLIDE is that it has the flexibility of incorporating other transcriptomic data, such as RACE, CAGE, and EST, to increase isoform discovery accuracy. IsoLasso [46] is based on the well-known LASSO algorithm, a multivariate regression method designated to seek a balance between the maximization of prediction accuracy and the minimization of interpretation. By including some additional constraints in the quadratic program involved in LASSO, IsoLasso is able to make the set of assembled transcripts as complete as possible StringTie [54]. It is another popular assembler developed by Salzberg group. It uses a network flow algorithm for the simultaneous discovery and quantification of transcripts. Another advantage of StringTie is that it is part of the HISAT2-StringTie-Ballgown workflow and requires less effort to setup the entire RNA-seq data analysis pipeline [56]. CIDANE [55] is a novel framework for genome-based transcript reconstruction and quantification from RNA-seq reads. Its algorithmic core not only reconstructs transcripts ab initio, but also allows the use of the growing annotation of known splice sites, transcription start and end sites, or full-length transcripts, which are available for most model organisms. Accurately estimating isoforms in multiple samples is an important preliminary step to differential expression analysis at the level of isoforms. One promising direction to improve isoform identification and quantification is to exploit several samples at the same time, such as biological replicates or time course experiments. If some isoforms are shared by several samples, potentially with different abundances, the identifiability issue may vanish and the statistical power of the deconvolution methods may increase due to the availability of more data for estimation. The joint RNA isoform detection and quantification from multiple RNA-seq samples is effective in reducing false positive transcript discoveries [43,57].

Conclusions

In this review, we summarize the tools for gene and transcript isoform quantification and provide guidance for end users to choose the tools with desired features. Each of the tools offers unique features that are suitable for answering specific research questions. RNA-seq is emerging as a powerful approach for identification and quantification of transcript isoforms. However, short read fragments that cover only part of the transcript make it difficult to reconstruct full-length transcripts, especially for those expressed at low levels.

References

1. Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, et al. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife 5: e11752.
2. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, et al. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22: 1616-1625.
3. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, et al. (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol 18: 1435-1440.
4. Zaghlool A, Ameur A, Nyberg L, Halvardson J, Grabherr M, et al. (2013) Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. BMC biotechnology 13: 99.
5. Batra R, Charizanis K, Manchanda M, Mohan A, Li M, et al. (2014) Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. Mol Cell 56: 311-322.
6. de Klerk E, Venema A, Anvar SY, Goeman JJ, Hu O, et al. (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res 40: 9089-9101.
7. Fu Y, Sun Y, Li Y, Li J, Rao X, et al. (2011) Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 21: 741-747.
8. Mayr C, Bartel DP (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673-684.
9. Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, et al. (2012) An in-depth map of polyadenylation sites in cancer. Nucleic Acids Res 40: 8460-8471.
10. Costa V, Angelini C, De Feis I, Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol 2010: 853916.
11. Lee JH, Gao C, Peng G, Greer C, Ren S, et al. (2011) Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. Circ Res 109: 1332-1341.
12. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

13. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS one 9: e78644.
14. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics 9: e1003569.
15. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166-169.
16. Liao Y, Smyth GK, Shi W (2014) Feature counts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923-930.
17. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31: 46-53.
18. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L (2016) Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv.
19. Zhao S, Xi L, Zhang B (2015) Union exon based approach for RNA-Seq gene quantification: To Be or Not to Be? PloS one 10: e0141910.
20. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, et al. (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol 16: 150.
21. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, et al. (2015) Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics 16: 675.
22. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.
23. Yates A, Akanni W, Amode MR, Barrell D, Billis K, et al. (2016) Ensembl 2016. Nucleic Acids Res 44: D710-16.
24. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, et al. (2006) The UCSC known genes. Bioinformatics 22: 1036-1046.
25. Zhao S, Zhang B (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16: 97.
26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652.
27. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
28. 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: R25.
29. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, et al. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: R36.
30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.
31. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873-881.
32. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, et al. (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38: e178.
33. Patro R, Duggal G, Kingsford C (2015) Salmon: Accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-Alignment. bioRxiv.
34. Srivastava A, Sarkar H, Gupta N, Patro R (2016) RapMap: A rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics 32: i192-i200.
35. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32: 462-464.
36. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34: 525-527.
37. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.
38. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10: 71-73.
39. Nariai N, Kojima K, Mimori T, Sato Y, Kawai Y, et al. (2014) TIGAR2: Sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. BMC Genomics 15: S5.
40. Nariai N, Hirose O, Kojima K, Nagasaki M (2013) TIGAR: Transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. Bioinformatics 29: 2292-2299.
41. Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M (2015) Fast and accurate approximate inference of transcript expression from RNA-seq data. Bioinformatics 31: 3881-3889.
42. Glaus P, Honkela A, Rattray M (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28: 1721-1728.
43. Bernard E, Jacob L, Mairal J, Viara E, Vert JP (2015) A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples. BMC Bioinformatics 16: 262.
44. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.
45. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503-510.
46. Li W, Feng J, Jiang T (2011) IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18: 1693-1707.
47. Xia Z, Wen J, Chang CC, Zhou X (2011) NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq. BMC Bioinformatics 12: 162.
48. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci USA 108: 19867-19872.
49. Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, et al. (2013) iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res 23: 519-529.
50. Tomescu AI, Kuosmanen A, Rizzi R, Makinen V (2013) A novel min-cost flow method for estimating transcript expression with RNA-Seq. BMC Bioinformatics 14: S15.
51. Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P, et al. (2013) MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. Bioinformatics 29: 2529-2538.
52. Bernard E, Jacob L, Mairal J, Vert JP (2014) Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. Bioinformatics 30: 2447-2455.
53. Teng M, Love MI, Davis CA, Djebali S, Dobin A, et al. (2016) A benchmark for RNA-seq quantification pipelines. Genome Biol 17: 74.
54. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33: 290-295.
55. Canzar S, Andreotti S, Weese D, Reinert K, Klau GW (2016) CIDANE: Comprehensive isoform discovery and abundance estimation. Genome Biol 17: 16.
56. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11: 1650-1667.
57. Suo C, Calza S, Salim A, Pawitan Y (2014) Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data. Bioinformatics 30: 506-513.