Chapter 26

Methods to Study Splicing from High-Throughput RNA Sequencing Data

Gael P. Alamancos, Eneritz Agirre, and Eduardo Eyras

Abstract

The development of novel high-throughput sequencing (HTS) methods for RNA (RNA-Seq) has provided a very powerful mean to study splicing under multiple conditions at unprecedented depth. However, the complexity of the information to be analyzed has turned this into a challenging task. In the last few years, a plethora of tools have been developed, allowing researchers to process RNA-Seq data to study the expression of isoforms and splicing events, and their relative changes under different conditions. We provide an overview of the methods available to study splicing from short RNA-Seq data, which could serve as an entry point for users who need to decide on a suitable tool for a specific analysis. We also attempt to propose a classification of the tools according to the operations they do, to facilitate the comparison and choice of methods.

Key words RNA-Seq, Splicing, Alternative splicing, Isoform, Quantification, Reconstruction

1 Introduction

The development of novel high-throughput sequencing (HTS) methods for RNA (RNA-Seq) has facilitated the discovery of many novel transcribed regions and splicing isoforms [1] and has provided evidence that a large fraction of the transcribed RNA in human cells undergo alternative splicing [2, 3]. RNA-Seq thus represents a very powerful tool to study alternative splicing under multiple conditions at unprecedented depth. However, the large datasets produced and the complexity of the information to be analyzed has turned this into a challenging task. In the last few years, a plethora of tools have been developed (Fig. 1), allowing researchers to process RNA-Seq data to study the expression of isoforms and splicing events, and their relative changes under different conditions. In this chapter, we provide an overview of the methods available to study alternative splicing from short RNA-Seq data.

Klemens J. Hertel (ed.), *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*, Methods in Molecular Biology, vol. 1126, DOI 10.1007/978-1-62703-980-2_26,

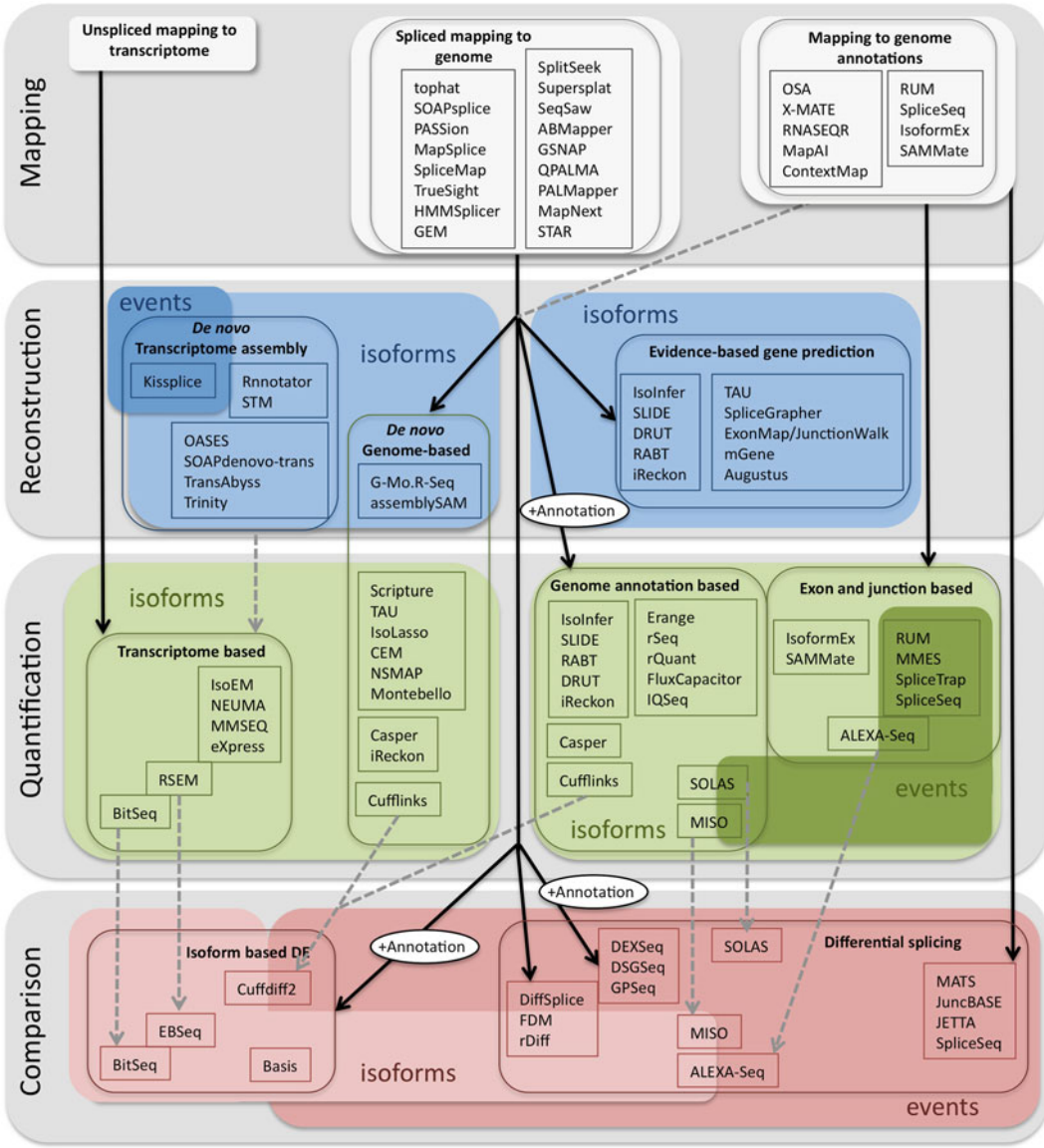

Fig. 1 Graphical representation of methods to study splicing from RNA-Seq. Methods are divided according to whether they perform mapping, reconstruction of events/isoforms, quantification of events/isoforms and whether they can perform a comparison between two or more conditions of event/isoform relative abundances, or of isoform expression. We only list the mapping methods that are spliced-mappers or the ones that use some heuristics to map to known exons and junctions. Mapping methods that also perform quantification are repeated in both levels. Methods for reconstruction (*blue*), quantification (*green*), and comparison (*red*) are divided according to whether they work with isoforms (*lighter color*) or with events (*darker color*); when they work at both levels, events and isoforms, they are overlapped by the two color tones, darker and lighter, respectively. Methods are also grouped by rounded rectangles according to the tables in Subheading 2. Some methods perform reconstruction and quantification and are grouped with those that only perform reconstruction. Methods that require an annotation are indicated. Quantification methods that work with or without annotation are in different groups. *Solid arrows* connect Mapping methods to the tools in the other three levels; since, in principle, any mapping method producing BAM as output could be fed to methods reading BAM as input. Some methods perform mapping and quantification or mapping and differential splicing and are connected with a *solid arrow* too. We indicate with *dashed gray arrows* those cases when a comparison method can use the output from a quantification method

We will group the methods according to the different questions they address:

1. Assignment of the sequencing reads to their likely gene of origin. This is addressed by methods that map reads to the genome and/or to the available gene annotations (Subheading 2.1).
2. Quantification of events and isoforms. Either using an annotation (Subheadings 2.2 and 2.3) or after reconstructing transcripts (Subheading 2.4), many methods estimate the expression level or the relative usage of isoforms and/or events.
3. Recovering the sequence of splicing events and isoforms. This is addressed by transcript reconstruction and de novo assembly methods (Subheadings 2.4, 2.5, and 2.6).
4. Providing an isoform or event view of differential splicing or expression. These include methods that compare relative event/isoform abundance or isoform expression across two or more conditions (Subheadings 2.7 and 2.8).
5. Visualizing splicing regulation. Various tools facilitate the visualization of the RNA-Seq data in the context of alternative splicing (Subheading 2.9).

In this review, we use transcript or isoform to refer to a distinct RNA molecule transcribed from a gene locus. We use gene to refer to the set of isoforms transcribed from the same genomic region and the same strand, sharing some exonic sequence; and a gene locus refers to this genomic region. A splicing event refers to the exonic region of a gene that shows variability across two or more of its isoforms. Splicing events generally include exon skipping (or cassette exon), alternative 5′ and 3′ splice-sites, mutually exclusive exons, retained introns, alternative first exons and alternative last exons (*see* for example [4]), although other events may occur as a combination of two or more of these ones. In this review, we do not enter into the details of the specific mathematical models behind each method; for a comparative analysis of the mathematical models behind many of these methods *see* ref. 5. Our aim is rather to provide an overview that could serve as an entry point for users who need to decide on a suitable tool for a specific analysis. We also attempt to propose a classification of the tools according to the operations they do, to facilitate the comparison and choice of methods.

2 Materials

This section includes the list of methods described in subsequent sections.

2.1 Spliced-Mappers

In Table 1, we provide a list of mapping tools that are able to locate exon–intron boundaries. Some of the methods use annotation information for mapping (OSA, X-MATE, SAMMate, IsoformEx,

Table 1
A list of mapping tools that are able to locate exon–intron boundaries

Method	Type	Uses annotation	Paired-end reads	Splice site model	Reference	Web site
TopHat	Exon-first	Optional	Yes	Exact match to GT/C-AG	[6]	http://tophat.cbcb.umd.edu/
SOAPsplice	Exon-first	No	Yes	Exact match to GT-AG, GC-AG, AT-AC	[7]	http://soap.genomics.org.cn/soapsplice.html
PASSion	Exon-first	No	Only paired-end	Exact match to GT-AG, GC-AG, AT-AC	[8]	https://trac.nbic.nl/passion
MapSplice	Exon-first. Seed-and-extend for spliced reads	No	Yes	Unbiased	[9]	http://www.netlab.uky.edu/p/bioinfo/MapSplice
SpliceMap	Exon-first. Seed-and-extend for spliced reads	No	Yes	Exact match to GT-AG, GC-AG, AT-AC	[10]	http://www.stanford.edu/group/wonglab/SpliceMap/
HMMSplicer	Exon-first. Seed-and-extend for spliced reads	No	Yes	Hidden Markov Model	[11]	http://derisilab.ucsf.edu/index.php?software=105
TrueSight	Exon-first. Seed-and-extend for spliced reads	No	Yes	Exact match to GT-AG, GC-AG, AT-AC	[12]	http://bioen-compbio.bioen.illinois.edu/TrueSight/
GEM	Exon-first. Seed-and-extend for splice reads	Optional	Yes	User defined regular expression and known junctions (optional)	[13]	http://algorithms.cnag.cat/wiki/The_GEM_library
SplitSeek	Seed-and-extend	No	Yes	Unbiased	[14]	http://solidsoftwaretools.com/gf/project/splitseek
Supersplat	Seed-and-extend	No	No	Unbiased	[15]	https://github.com/mocklerlab/supersplat

SeqSaw	Seed-and-extend	No	Yes	Unbiased	[16]	http://bioinfo.au.tsinghua.edu.cn/software/seqsaw
ABMapper	Seed-and-extend	No	Yes	Exact match to GT-AG, GC-AG, AT-AC	[17]	http://abmapper.sourceforge.net/
MapNext	Seed-and-extend	Optional	No	Known-junctions and GT-AG for novel ones	[18]	http://evolution.sysu.edu.cn/english/software/mapnext.htm
STAR	Seed-and-extend	Optional	Yes	Exact match to GT-AG, GC-AG, AT-AC and unbiased	[19]	http://gingeraslab.cshl.edu/STAR/
GSNAP	Seed-and-extend	No	Yes	Exact match to GT-AG, GC-AG, AT-AC	[20]	http://research-pub.gene.com/gmap/
QPALMA	Seed-and-extend	No	No	SVM model for splice-sites	[21]	http://www.raetschlab.org/suppl/qpalma
PALMapper	GenomeMapper + QPalma	No	Yes	Qpalma model	[22]	http://galaxy.raetschlab.org/
CRAC	Seed-and-extend	No	No	Unbiased	[23]	http://crac.gforge.inria.fr/
OLEgo	Multi-seed	No	Yes	Combined model of splice-site sequence and intron length	[24]	http://zhanglab.c2b2.columbia.edu/index.php/OLego
Subread	Multi-seed	No	Yes	Exact match to GT-AG	[25]	http://bioconductor.org/packages/release/bioc/html/Rsubread.html
OSA	Seed-and-extend	Yes	Yes	Known and splice-sites and exact match to GT-AG, GC-AG, AT-AC	[26]	http://omicsoft.com/osa/

(continued)

Table 1
(continued)

Method	Type	Uses annotation	Paired-end reads	Splice site model	Reference	Web site
X-MATE	Recursive mapping to genome and junctions	Yes	No	Known splice-sites	[27]	http://grimmond.imb.uq.edu.au/X-MATE/
RNASEQR	Bowtie and BLAT on transcripts and genome	Yes	Yes	Known splice-sites and BLAT model	[28]	https://github.com/rnaseqr/RNASEQR
MapAI	Bowtie alignments to transcripts	Yes	No	Known splice-sites	[29]	http://www.bioinf.boku.ac.at/pub/MapAl/
SAMMate	Bowtie to exons and junctions	Yes	Yes	Known splice-sites	[30]	http://sammate.sourceforge.net/
IsoformEx	Bowtie to exons and junctions	Yes	No	Known splice-sites	[31]	http://bioinformatics.wistar.upenn.edu/isoformex
RUM	Bowtie and BLAT on transcripts and genome	Yes	Yes	Known splice-sites and BLAT model	[32]	http://www.cbil.upenn.edu/RUM/userguide.php
SpliceSeq	Bowtie alignments to Splicing graphs	Yes	Yes	Known splice-sites	[33]	http://bioinformatics.mdanderson.org/main/SpliceSeq:Overview
PASTA	Bowtie alignment of read fragments	No	Yes	Logistic-regression model for splice-sites	[34]	http://genome.ufl.edu/rivalab/PASTA
ContextMap	Genome alignments from other methods	No	No	Unbiased	[35]	http://www.bio.ifi.lmu.de/softwareservices/contextmap

RNASEQR, RUM, SpliceSeq, MapAI), some can use annotation as an option (GEM, MapNext, STAR, TopHat), and others (the rest) work directly with the genome reference. Additionally, some methods perform quantification (Subheading 2.2) (SAMMate, IsoformEx, RUM, SpliceSeq) and are included here since they provide an independent method for mapping. We also indicate whether the method can map paired-end reads, the type of splice-site model used, the reference where the method is described and the URL where the software is available.

2.2 Genome-Based Quantification of Known Events and Isoforms

In Table 2, we give a list of methods that can be used to quantify known splicing events (RUM, SpliceSeq, MMES, SpliceTrap), known isoforms (SAMMate, IsoformEx, Erange, rSeq, rQuant, FluxCapacitor, IQSeq, Cufflinks, Casper, CEM, IsoInfer, SLIDE, RABT, DRUT, iReckon), or both (MISO, ALEXA-Seq, SOLAS) when a genome-based annotation is available. Some include the mapping step (RUM, SpliceSeq, SAMMate, IsoformEx). Some isoform-based methods can quantify known and novel isoforms simultaneously (IsoInfer, SLIDE, RABT, DRUT, iReckon) or choose between quantifying known or novel isoforms (Cufflinks, Casper, CEM, IsoLasso). We indicate the type of input used by each method, whether they exploit paired-end read information in the calculation and what type of quantification is given. We also provide the reference where the method is described, and the URL (or email) where the software is available.

2.3 Isoform Quantification Guided by a Transcriptome

Table 3 includes methods that quantify isoforms using a transcriptome annotation and reads mapped with a non-spliced mapper. All the methods listed used bowtie to map reads to transcripts in the original publication. Although they generally work with reads mapped to a transcriptome, some methods (RSEM, MMSEQ) can work with reads mapped to a genome. We indicate the type of input used by the method, whether they exploit paired-end read information in the calculation and what type of isoform quantification is given. We also provide the reference where the method is described, and the URL where the software is available.

2.4 Genome-Based Reconstruction and Quantification Without Annotation

Table 4 includes methods to reconstruct (all methods) and to quantify (all methods except for G-Mo.R-Se and assemblySAM) multiple isoforms from genome-mapped reads without using any gene annotation. Some methods can also be run with annotations for quantification (Cufflinks, IsoLasso, Casper, CEM). Some perform simultaneously the reconstruction and quantification of novel isoforms (NSMAP, Montebello, IsoLasso). We indicate the type of input used by each method, whether they exploit paired-end read information in the calculation and what type of isoform quantification is given. We also provide the reference where the method is described and the URL or email where the software is available.

Table 2
A list of methods that can be used to quantify known splicing events (RUM, SpliceSeq, MMES, SpliceTrap), known isoforms (SAMMate, IsoformEx, Erange, rSeq, rQuant, FluxCapacitor, IQSeq, Cufflinks, Casper, CEM, IsoInfer, SLIDE, RABT, DRUT, iReckon), or both (MISO, ALEXA-Seq, SOLAS) when a genome-based annotation is available

Method	Type	Input used in publication	Uses paired-end reads	Quantification	Reference	Web site
RUM	Exon and junction quantification	Bowtie and BLAT on transcripts and genome	Yes	Read counts and RPKM of exons and junctions	[32]	http://www.cbil.upenn.edu/RUM/userguide.php
SpliceSeq	Exon and junction quantification	Bowtie alignments to Splicing graphs	Yes	Inclusion level of exons and junctions	[33]	http://bioinformatics.mdanderson.org/main/SpliceSeq:Overview
MMES	Junction quantification	SOAP alignments to junctions	No	Junction scores	[36]	Email to Wang.Liguo@mayo.edu
SpliceTrap	Exon and junction quantification	Bowtie on inclusion/skipping events	Yes (models insert sizes)	Exon inclusion level	[37]	http://rulai.cshl.edu/splicetrap/
SAMMate	Isoform quantification	Bowtie on genome and junctions	Yes	RPKM/FPKM	[30]	http://sammate.sourceforge.net/
IsoformEx	Isoform quantification	Bowtie on genome and junctions	No	Isoform expression (~RPKM)	[31]	http://bioinformatics.wistar.upenn.edu/isoformex
MISO	Event and isoform quantification	Bowtie on genome and junctions	Yes	Isoform PSI value	[38]	http://genes.mit.edu/burgelab/miso/

ALEXA-Seq	Event and isoform quantification	Reads mapped to genome and junctions	Yes	Event and isoform expression level	[39]	http://www.alexaplatform.org/alexa_seq/
SOLAS	Event and isoform quantification	Reads mapped to genome	No	Isoform expression (~RPKM)	[40]	http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/
Erange	Isoform quantification	Bowtie on genome and junctions	No	Isoform RPKM	[41]	http://woldlab.caltech.edu/rnaseq
rSeq	Isoform quantification	SeqMap alignments to exons and exon–exon junctions	Yes (in latest version)	Isoform RPKM	[42]	http://www.personal.umich.edu/~jianghui/rseq/
rQuant	Isoform quantification	Reads mapped to genome	No	Isoform average read coverage and RPKM	[43]	http://galaxy.raetschlab.org/
FluxCapacitor	Isoform quantification	Reads mapped to genome	Yes	Isoform relative abundance (~PSI)	[44]	http://flux.sammeth.net/capacitor.html
IQSeq	Isoform quantification	GFF/MRF/BED	Yes	Isoform RPKM	[45]	http://archive.gersteinlab.org/proj/rnaseq/IQSeq/
Cufflinks	Known or novel Isoform quantification	TopHat alignments	Yes	FPKM	[46]	http://cufflinks.cbcb.umd.edu/
Casper	Known or novel Isoform quantification	TopHat alignments	Yes	Isoform PSI value	[47]	https://sites.google.com/site/rosselldavid/software
CEM	Known or novel Isoform quantification	TopHat alignments	Yes	Isoform expression	[48]	http://alumni.cs.ucr.edu/~liw/cem.html

(continued)

Table 2
(continued)

Method	Type	Input used in publication	Uses paired-end reads	Quantification	Reference	Web site
IsoLasso	Known or novel Isoform quantification	TopHat alignments	Yes	RPKM	[49]	http://alumni.cs.ucr.edu/~liw/isolasso.html
IsoInfer	Known and novel isoform quantification	TopHat alignments	Yes	Isoform RPKM	[50]	http://www.cs.ucr.edu/~jianxing/IsoInfer.html
SLIDE	Known and novel isoform quantification	modEncode spliced mappings	Yes	Isoform RPKM	[51]	https://sites.google.com/site/jingyijli/SLIDE.zip
RABT	Known and novel isoform quantification	TopHat alignments	Yes	Isoform FPKM	[52]	http://cufflinks.cbcb.umd.edu/
DRUT	Known and novel isoform quantification	Bowtie/TopHat alignments to transcriptome/genome	No	FPKM (computed by IsoEM)	[53]	http://www.cs.gsu.edu/~serghei/?q=drut
iReckon	Known and novel isoform quantification	TopHat alignments	Yes	Isoform RPKM	[54]	http://compbio.cs.toronto.edu/ireckon/

Table 3
Methods that quantify isoforms using a transcriptome annotation and reads mapped with a non-spliced mapper

Method	Input reads format	Uses paired-end reads	Isoform quantification	Reference	Web site
RSEM	BAM/SAM	Yes (models insert size)	"Expected number of fragments per isoform" and "fraction of transcripts represented by the isoform"	[55]	https://github.com/bli25wisc/RSEM/
IsoEM	SAM	Yes (models insert size)	Isoform expression	[56]	http://dna.engr.uconn.edu/?page_id=105
NEUMA	Fastq/Fasta mapped with Bowtie	Yes	FVKM (fragments per virtual kilobase per million sequenced reads)	[57]	http://neuma.kobic.re.kr
BitSeq	SAM	Yes (models insert size)	Isoform expression	[58]	http://www.bioconductor.org/packages/2.11/bioc/html/BitSeq.html
MMSEQ	Sorted BAM	Yes	Haplotype and isoform-specific expression	[59]	http://bgx.org.uk/software/mmseq.html
eXpress	BAM	Yes	FPKM, estimated counts	[60]	http://bio.math.berkeley.edu/eXpress/

Table 4
Methods to reconstruct (all methods) and to quantify (all methods except for G-Mo.R-Se and assemblySAM) multiple isoforms from genome-mapped reads without using any gene annotation

Method	Type	Input used in publication	Uses paired-end reads	Isoform quantification	Reference	Web site
G-Mo.R-Se	De novo isoform reconstruction	SOAP alignments	No	No	[61]	http://www.genoscope.cns.fr/externe/gmorse/
assemblySAM	De novo isoform reconstruction	Own heuristics for read-mapping using Bowtie	Yes	No	[62]	http://sammate.sourceforge.net/assemblysam.html
TAU	De novo isoform reconstruction and quantification	Supersplat alignments	No	Average per-base sequencing depth	[63]	Email to HPriest@danforthcenter.org
Scripture	De novo isoform reconstruction and quantification	TopHat alignments	Yes (models insert size)	RPKM	[64]	http://www.broadinstitute.org/software/Scripture/
Cufflinks	Known or novel isoform quantification	TopHat alignments	Yes (models insert size)	FPKM	[46]	http://cufflinks.cbcb.umd.edu/
Casper	Known or novel Isoform quantification	TopHat alignments	Yes	Isoform PSI value	[47]	https://sites.google.com/site/rosselldavid/software

CEM	Known or novel isoform quantification	TopHat alignments	Yes	Isoform expression	[48]	http://alumni.cs.ucr.edu/~liw/cem.html
IsoLasso	Known or novel isoform quantification	TopHat alignments	Yes	RPKM	[49]	http://alumni.cs.ucr.edu/~liw/isolasso.html
Montebello	Novel isoform reconstruction and quantification	SpliceMap alignments	Yes	Isoform expression	[65]	http://www.stanford.edu/group/wonglab/Montebello/Montebello_0.8.tar.gz
NSMAP	Novel isoform reconstruction and quantification	TopHat alignments	Yes (models insert size)	RPKM	[66]	https://sites.google.com/site/nsmapforrnaseq

2.5 Evidence-Based Alternatively Spliced Gene Prediction

Table 5 includes methods that could be used to perform alternatively spliced gene prediction from RNA-Seq data. Besides the de novo reconstruction and quantification methods from Subheading 2.4 and those from Subheading 2.2 that can predict novel and known isoforms simultaneously (IsoInfer, SLIDE, RABT, DRUT, iReckon), we also include methods that can use various sources of evidence to predict alternatively spliced genes (TAU, SpliceGrapher, ExonMap/JunctionWalk) and methods that predict alternatively spliced protein-coding genes from multiple evidences (Augustus, mGene). We also include classical protein-coding gene prediction methods that could potentially use RNA-Seq as evidence (Gaze, JigSaw, EVM, Evigan). For each method, we indicate the type of input used, whether they exploit paired-end read information in the calculation or provide any isoform quantification. We also give the reference where the method is described and the URL or email where the software is available.

2.6 De Novo Transcriptome Assembly

Table 6 includes methods for de novo transcriptome assembly. Some of these methods produce multiple isoforms per assembled gene (OASES, SOAPdenovo-trans, TransAbyss, Trinity) and only two quantify the alternative isoforms (TransAbyss, Trinity). Nonetheless, these methods could be coupled with transcriptome-based quantification methods (Subheading 2.3). KisSplice assembles alternatively spliced events rather than isoforms and quantifies the read coverage of these events. We indicate whether they exploit paired-end read information in the calculation, the *k*-mer approach (single/multiple), whether they detect multiple isoforms per gene and whether they perform isoform quantification. We also provide the reference where the method is described and the URL (or email) where the software is available.

2.7 Differential Splicing

Table 7 includes methods that measure relative changes in inclusion/usage between two or more conditions at the exon level (DEXSeq, DSGSeq, GPSeq, SOLAS), event level (MATS, JuncBASE, JETTA, SpliceSeq), and isoform region level (DiffSplice, SplicingCompass, FDM, rDiff) or at both, isoform and event levels (MISO, ALEXA-Seq). We indicate whether the methods perform any quantification per sample, the measure of differential splicing provided, whether they exploit paired-end read information in the calculation, the reference where the method is described and the URL where the software is available.

2.8 Isoform-Based Differential Expression

Table 8 includes methods that measure differential expression at the transcript level between two or more conditions, allowing multiple transcripts per gene. Cuffdiff2, additionally, can calculate significant changes in the relative abundance of isoforms. For each method, we indicate the quantification performed per sample, whether it exploits paired-end read information in the calculation,

Table 5
Methods to perform alternatively spliced gene prediction from RNA-Seq data

Method	Type	Input used in publication	Uses paired-end reads	Isoform quantification	Reference	Web site
IsoInfer	Known and novel isoform quantification	TopHat alignments	Yes	Isoform RPKM	[50]	http://www.cs.ucr.edu/~jianxing/IsoInfer.html
SLIDE	Known and novel isoform quantification	modEncode spliced mappings	Yes	Isoform RPKM	[51]	https://sites.google.com/site/jingyijli/SLIDE.zip
RABT	Known and novel isoform quantification	TopHat alignments	Yes	Isoform FPKM	[52]	http://cufflinks.cbcb.umd.edu/
DRUT	Known and novel isoform quantification	Bowtie (TopHat) alignments to transcriptome/genome	No	FPKM (computed by IsoEM)	[53]	http://www.cs.gsu.edu/~serghei/?q=drut
iReckon	Known and novel isoform quantification	TopHat alignments	Yes	Isoform RPKM	[54]	http://compbio.cs.toronto.edu/ireckon/
TAU	Evidence-based isoform reconstruction and quantification	Supersplat alignments	No	Average per-base sequencing depth	[63]	Email to hpriest@danforthcenter.org
SpliceGrapher	Evidence-based isoform reconstruction	TopHat alignments	Yes	No	[67]	http://SpliceGrapher.sf.net

(continued)

Table 5
(continued)

Method	Type	Input used in publication	Uses paired-end reads	Isoform quantification	Reference	Web site
ExonMap/ JunctionWalk	Evidence-based isoform reconstruction	Reads mapped to exons and junctions	Handled by SpliceMap	No	[68]	http://gluegrant1.stanford.edu/~DIC/RNASeqArray/TranscriptConstruction.html
mGene	Evidence-based alternatively spliced gene prediction	Reads mapped to genome	Yes	No	[69]	http://mgene.org/
Augustus	Evidence-based alternatively spliced gene prediction	Spliced evidences	No	No	[70]	http://bioinf.uni-greifswald.de/augustus/
Gaze	Evidence-based gene prediction	Evidence in GFF format	No	No	[71]	http://www.sanger.ac.uk/resources/software/gaze/
JigSaw	Evidence-based gene prediction	Spliced evidences	No	No	[72]	http://www.cbcb.umd.edu/software/jigsaw/
EVM	Evidence-based gene prediction	PASA alignments	No	No	[73]	http://evidencemodeler.sourceforge.net/
Evigan	Evidence-based gene prediction	Spliced evidences	No	No	[74]	http://www.seas.upenn.edu/~strctlrn/evigan/evigan.html

Table 6
Methods for de novo transcriptome assembly

Method	Uses paired-end reads	Graph approach	Detects alternative isoforms	Isoform quantification	Reference	Web site
Rnnotator	Yes	Variable *k*-mer	No	No	[75]	Email to vtdelapuente@lbl.gov
STM	Yes	Variable *k*-mer	No	No	[76]	http://www.surget-groba.ch/downloads/stm.tar.gz
OASES	Yes	Variable *k*-mer	Yes	No	[77]	http://www.ebi.ac.uk/~zerbino/oases/
SOAPdenovo-trans	Yes	Variable *k*-mer	Yes	No	[78]	http://sourceforge.net/projects/soapdenovotrans/
TransAbyss	Yes	Variable *k*-mer	Yes	Isoform read coverage	[79]	http://www.bcgsc.ca/platform/bioinfo/software/
Trinity	Yes	Single *k*-mer	Yes	Yes (uses RSEM)	[80]	http://TrinityRNASeq.sourceforge.net
KisSplice	No	Single *k*-mer	Events only	Event read coverage	[81]	http://alcovna.genouest.org/kissplice/

Table 7
Methods that measure relative changes in inclusion/usage between two or more conditions at the exon level (DEXSeq, DSGSeq, GPSeq, SOLAS), event level (MATS, JuncBASE, JETTA, SpliceSeq), and isoform region level (DiffSplice, SplicingCompass, FDM, rDiff) or at both isoform and event levels (MISO, ALEXA-Seq)

Method	Type	Quantification	Uses paired-end reads	Models biological variability	Differential quantification	Reference	Web site
DEXSeq	Exon level	No	No	Yes	Differential exon inclusion	[82]	http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
DSGSeq	Exon level	Isoform relative abundances	No	Yes	Differential exon inclusion	[83]	http://bioinfo.au.tsinghua.edu.cn/software/DSGseq
GPSeq	Exon level	Exon splicing ratio	No	Yes	Differential exon splicing index	[84]	http://cran.r-project.org/web/packages/GPseq/index.html
SOLAS	Exon level	Event and isoform inclusion	No	No	Differential exon inclusion	[40]	http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/
MATS	Event level	Event inclusion	Handled by mapping method	Yes	Differential event inclusion	[85]	http://rnaseq-mats.sourceforge.net/
JuncBASE	Event level	Event inclusion	Yes	No	Differential event inclusion	[86]	http://compbio.berkeley.edu/proj/juncbase/Home.html
JETTA	Event level	SeqMap alignments	Handled by mapping method	No	Differential event inclusion	[87]	http://igenomed.stanford.edu/~junhee/JETTA/rnaseq.html

SpliceSeq	Event level	Inclusion level of exons and junctions	Yes	No	Differential event inclusion	[33]	http://bioinformatics.mdanderson.org/main/SpliceSeq:Overview
MISO	Event and isoform levels	PSI	Yes	No	Differential event/isoform PSI	[38]	http://genes.mit.edu/burgelab/miso/
Alexa-Seq	Event and isoform levels	Gene, transcript, and event expression levels	Yes	No	Differential relative event/isoform expression	[39]	http://www.alexaplatform.org/alexa_seq/
SplicingCompass	Isoform region level	Normalized exon density	Handled by mapping method	No	Differential relative isoform abundance	[88]	http://www.ichip.de/software/SplicingCompass.html
DiffSplice	Isoform region level	Expression of "Alternative Splicing Modules"	Yes	Yes	Differential Expression of "Alternative Splicing Modules"	[89]	http://www.netlab.uky.edu/p/bioinfo/DiffSplice
FDM	Isoform region level	Isoform region relative expression	No	Yes	Differential relative isoform abundance	[90]	http://csbio-linux001.cs.unc.edu/nextgen/software/FDM
rDiff	Isoform region level	Isoform region relative expression	Yes	Yes	Differential relative isoform abundance	[91]	http://bioweb.me/rdiff

Table 8
Methods that measure differential expression at the transcript level between two or more conditions, allowing multiple transcripts per gene

Method	Quantification	Uses paired-end reads	Models biological variability	Differential quantification	Reference	Web site
BitSeq	Isoform expression	Yes	Yes	Differential isoform expression	[58]	http://www.bioconductor.org/packages/2.11/bioc/html/BitSeq.html
BASIS	Isoform relative expression	No	No	Differential isoform expression	[92]	http://www.rcf.usc.edu/~liangche/software.html
Cuffdiff2	Isoform expression	Yes	Yes	Differential isoform expression	[93]	http://cufflinks.cbcb.umd.edu/
EBSeq	Isoform expression quantified by input method	Handled by quantification method	Yes	Differential isoform expression	[94]	http://www.biostat.wisc.edu/~kendzior/EBSEQ/

the measure of differential expression provided, the reference where the method is described and the URL where the software is available.

2.9 Visualization of Alternative Splicing

Table 9 includes some of the available tools for the visualization of alternative splicing using RNA-Seq data. Some of them can be used as command line tools that are included in the distribution of the analysis tools (RSEM, SpliceGrapher, DiffSplice, DEXSeq, SplicingCompass) or provided separately (Sashimi Plots), whereas others are Graphical User Interfaces (Savant, ALEXA-Seq, SpliceSeq).

3 Methods

3.1 Spliced-Mapping Short Reads

Event and Isoform quantification are very much dependent on the correct assignment of RNA-Seq reads to the molecule of origin. Accordingly, we will start by reviewing some of the read mappers that are splice-site aware, and therefore, can be used to detect exon–intron boundaries and connections between exons. This alignment problem has been addressed in the past by tools that combine fast heuristics for sequence matching with a model for splice-sites, for example, Exonerate [97], BLAT [98], or GMAP [99]. These methods, however, are not competitive enough to map all reads from a sequencing run in a reasonable time. In the last few years, a myriad of methods have been developed for mapping short reads to a reference genome [100]. Those that are splice-site aware and incorporate intron-like gaps are generally called spliced-mappers, split-mappers, or spliced aligners. Their main challenge is that reads must be split into shorter pieces, which may be harder to map unambiguously; and although introns are marked by splice-site signals, these occur frequently by chance in the genome.

Spliced-mappers have been classified previously into two main classes [101], *exon-first* and *seed-and-extend* (Subheading 2.1). *Exon-first* methods map reads first to the genome using an unspliced approach to find read-clusters; unmapped reads are then used to find connections between these read-clusters. These methods include TopHat [6], SOAPsplice [7], PASSion [8], MapSplice [9], SpliceMap [10], HMMsplicer [11], TrueSight [12], and GEM [13]. *Seed-and-extend* methods generally start by mapping part of the reads as *k*-mers or substrings; candidate matches are then extended using different algorithms and potential splice-sites are located. These methods include SplitSeek [14], Supersplat [15], SeqSaw [16], ABMapper [17], MapNext [18], STAR [19], GSNAP [20], and PALMapper [22]. A generalization of seed-and-extend methods is represented by the multi-seed methods, like CRAC [23],

Table 9
Some of the available tools for the visualization of alternative splicing using RNA-Seq data

Method	Type	Used with	Input data	Visualization	Reference	Web site
Sashimi Plots	Command line tool	MISO	GFF3	Splicing events and read coverage	[38]	http://genes.mit.edu/burgelab/miso/docs/sashimi.html
RSEM	Command line tool	RSEM	Transcript BAM file	Read profiles (WIG)	[55]	https://github.com/bli25wisc/RSEM/
SpliceGrapher	Command line tool	SpliceGrapher	GFF files	Isoforms	[67]	http://SpliceGrapher.sf.net
DEXSeq	Command line tool	DEXSeq	DEXSeq results	Differential exon usage	[82]	http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
SplicingCompass	Command line tool	SplicingCompass	SplicingCompass results	Differential exon usage	[88]	http://www.ichip.de/software/SplicingCompass.html
DiffSplice	Command line tool	DiffSplice	GTF (graphs)	Isoforms	[89]	http://www.netlab.uky.edu/p/bioinfo/DiffSplice
SpliceSeq	GUI	SpliceSeq	SpliceSeq processed data	Isoforms and alternatively spliced events	[33]	http://bioinformatics.mdanderson.org/main/SpliceSeq:Overview
ALEXA-Seq viewer	GUI	ALEXA-Seq	Alexa-seq database	Splicing events and expression levels	[39]	http://www.alexaplatform.org/alexa_seq/
Savant Browser	GUI	iReckon	GFF	Isoforms	[95]	http://genomesavant.com/savant/
SplicingViewer	GUI	Splicing Viewer	BAM	Splicing events coverage	[96]	http://bioinformatics.zj.cn/splicingviewer/

OLego [24], and Subread [25], which consider multiple subreads within each read. Similarly, ABMapper consider multiple read-splits for mapping. Some methods actually use a hybrid strategy, following an exon-first approach for unspliced reads, and then using seed-and-extend approach for spliced reads, like MapSplice, SpliceMap, HMMSplicer, TrueSight, GEM, and PALMapper; the latter being a combination of GenomeMapper [102] and QPalma [21] for spliced reads. *Exon-first* methods depend strongly on sufficient coverage on potential exons to incorporate spliced reads, but are generally faster than *seed-and-extend* methods. On the other hand, *seed-and-extend* methods tend to be less dependent on recovering exon-like read-clusters and may recover more novel splice-sites. However, the storage of k-mers for long reads requires sufficient computer memory for large k, and the mapping has limited accuracy for small k [7].

There is also a different class of tools, which use the annotation and/or some heuristics to map reads. These include OSA [26], X-Mate [27], RNASEQR [28], MapAI [29], SAMMate [30], IsoformEx [31], RUM [32], SpliceSeq [33], and PASTA [34]. RNASEQR and RUM use Bowtie [103] to map reads to the transcriptome and genome; and then identify novel junctions from the unmapped reads using BLAT [98]. Similarly, SAMMate and IsoformEx use Bowtie to locate reads in exons and junctions, whereas SpliceSeq uses Bowtie to map reads to a graph representation of the annotation; X-Mate uses its own heuristics to trim and map reads recursively to locate reads on exons and junctions. On the other hand, PASTA does not use any gene annotation; it uses Bowtie and a splice-site model to locate read fragments on exon junctions. Among these methods, SAMMate, IsoformEx, RUM, and SpliceSeq also provide some level of quantification for exons, events, or isoforms (Subheading 2.2) (Fig. 1), which makes them convenient as a pipeline tool. OSA is actually a seed-and-extend mapping method but relies on an annotation. OSA avoids splitting reads into subreads which helps improving speed; and like other annotation-guided methods, also split-maps reads that are not located in the provided annotation using the seed-and-extend approach. Finally, unlike the other methods, MapAI and ContextMap use reads already mapped to a reference genome. MapAI uses reads mapped to a transcriptome to assign them to their genomic positions, whereas ContextMap refines the genome mappings using the read context, extending to all reads the context approach used by methods like MapSplice or GEM for spliced reads. In the newest version, ContextMap can also be used as a standalone read-mapping tool. Annotation-guided mapping methods are possibly the best option to accurately assign reads to gene annotations, whereas de novo mapping tools are convenient for finding new splicing junctions.

Besides the differences in the mapping procedure, de novo mapping tools detect splice-sites using a variety of approaches, which may determine the reliability of the splice-sites detected and the possibility of obtaining novel ones. Most tools search for an exact match of the flanking intronic dinucleotides to the canonical splice-sites GT-AG, GC-AG, AT-AC (*see* Subheading 2.1). Tools like MapNext and Tophat use a two-step approach, first mapping to the known junctions and then locating novel ones with GT-AG dinucleotides, whereas tools like MapSplice, Supersplat, SpliceMap, and HMMSplicer use a gapped-alignment approach that allows the detection of junctions regardless of the exon coverage. HMMSplice, QPalma, PASTA, and OLego use a more complex representation for splice-sites. HMMSplice is based on a hidden Markov model, QPalma on a Support Vector Machine, PASTA on a logistic regression, and OLego in the combined logistic modeling of sequence bias and intron-size; all of which are trained on known splice-sites. In contrast, MapSplice, SeqSaw, STAR, SplitSeek, and CRAC can do an unbiased search of splice-junctions, not necessarily looking for the splice-site motif and generally using support from multiple reads; hence, they can potentially recover noncanonical splice-sites. Annotation-guided methods will accurately assign reads to known splice-sites, but will miss novel ones, unless they use some heuristics for novel junctions like RUM and RNASEQR. Mapping methods like STAR, GEM, MapNext, and TopHat accept annotations as optional input, which will guide the initial mapping of reads. Other parameters may be important too, like the search range of intron lengths. Most models impose restrictions in the minimum and maximum intron lengths, but methods like MapSplice does not impose any restriction and OSA has a specific search for novel exons using distal fragments. The decision of which tool to use depends very much on whether the aim is to assign reads to known annotations or to find novel splice-sites.

3.2 Definition and Quantification of Events and Isoforms

First reports using RNA-Seq to quantify splicing followed an approach analogous to splicing junction arrays [104]. They were based on the analysis of junctions built from known gene annotations [2, 3, 105–108]. In these and later methods, reads aligning to candidate alternative exons and its junctions are considered as inclusion reads, whereas reads mapping to flanking exons and to junctions skipping the candidate alternative exon are considered as skipping or exclusion reads. These reads are then used to provide an estimate of the relative inclusion of the regulated exon [109], generally called inclusion level. This approach has shown a reasonable agreement with microarrays and can be modified to include exon-body reads and variable exon lengths [2, 109]

An alternative measure, "percent spliced in" (PSI), has been defined as the number of isoforms that include the exon over the total isoforms [110], or equivalently, as the fraction of mRNAs

that represent the inclusion isoform [38]. If the PSI value is calculated for a particular splicing event, it can be considered equivalent to the inclusion level. Isoform quantification can be expressed in terms of either a global measure of expression [58], which may provide a global ranking comparable across genes in one sample, or in terms of a relative measure of expression, which is normalized per gene locus and comparable across conditions. The global measure is generally given in terms of RPKM or FPKM (Reads or Fragments Per Kilobase of transcript sequence per Millions mapped reads); and the relative measure is given in terms of a PSI value or a similar value.

Besides the original approaches [2, 3, 105–108], various tools have been developed recently to quantify events and isoforms. These range from simply quantifying the inclusion of events to the reconstruction and quantification of novel isoforms. Some of the tools that reconstruct isoforms also estimate their quantification, and some tools may quantify either known isoforms or novel ones, or both simultaneously. Accordingly, we classify the methods depending on whether they use annotation or not and on the type of input and output:

1. Event/isoform quantification using known (genome-based) gene annotations (Subheading 2.2).
2. Isoform quantification using a transcriptome annotation (Subheading 2.3).
3. De novo isoform reconstruction with a genome reference, either purely focused on reconstruction or also providing isoform quantification (Subheading 2.4).
4. Isoform reconstruction and quantification guided by annotation. These methods use a gene annotation as a guide and can complete the annotation with new exons, new isoforms, or even with some new gene loci (Subheading 2.5).
5. Finally, some of the de novo transcript assembly methods also quantify isoforms (Subheading 2.6).

3.2.1 Event and Isoform Quantification Guided by Gene Annotation

Various tools have been developed for event quantification from a single condition (Subheading 2.2) (Fig. 1): RUM [32], SpliceSeq [33]. MMES [36], SpliceTrap [37], MISO [38], ALEXA-Seq [39], and SOLAS [40]. RUM provides quantification of genes, exons, and junctions in terms of read-counts and RPKM (reads per kilobase per million mapped reads), whereas SpliceTrap and MMES use the reads mapped to junctions and employ a statistical model to calculate exon inclusion levels and junction scores, RUM and MMES also provide the mapping step. RUM has its own heuristics (Subheading 2.1), whereas MMES maps reads to exon–exon junctions using SOAP [111]. Similarly, SpliceSeq maps reads to a splicing-graph to obtain exon and junction inclusion levels. MISO and

ALEXA-Seq use reads on exons and junctions, whereas SOLAS uses only reads on exons. MISO provides PSI values, while ALEXA-Seq and SOLAS event and isoform expression levels. MISO, ALEXA-Seq, and SOLAS can also estimate isoform relative abundances and can be further used for differential splicing (Subheading 2.7).

Quantification of isoforms is more complicated than that of events, as it requires the correct assignment of reads to isoforms sharing part of their sequence. One of the first attempts to do this was Erange [41], where reads mapped to the genome and known junctions were distributed in isoforms according to the coverage of the genomic context, and isoform expression was defined in terms of RPKM. However, the uncertainty in the assignment of reads shared by two or more isoforms must be appropriately modeled. Accordingly, a number of methodologies have been proposed to address this issue (Subheading 2.2): SAMMate [30], IsoformEx [31], MISO [38], ALEXA-Seq [39], SOLAS [40], rSeq [42], rQuant [43], FluxCapacitor [44], IQSeq [45], Cufflinks [46], Casper [47], CEM [48], IsoLasso [49], IsoInfer [50], SLIDE [51], RABT [52], DRUT [53], and iReckon [54]. Isoform quantification is generally given in terms of RPKM, FPKM, some equivalent *isoform expression level* value, PSI, or an equivalent *relative expression* value.

SAMMate and IsoformEx use the reads mapped to exons and junctions by their own methods to quantify gene and isoform expression in terms of RPKM values. SAMMate incorporates two quantification methods, one that is not sensitive to coverage, so it can be used on early sequencing platforms [112] and a recent one that is aimed for deeper coverage and uses a filtering of non-expressed transcripts [113]. SOLAS and rSeq use reads on exons to estimate isoform expression levels, whereas rQuant uses the position-wise density of mapped reads to calculate two abundance estimates: the RPKM and the estimated average read coverage for each transcript. IQSeq provides a statistical model that facilitates the incorporation of data from multiple technologies; and FluxCapacitor, unlike other methods, does not account for the mapping variability across isoforms and directly solves the constraints derived from distributing the reads in isoforms according to the splicing graph built from the read evidence.

IsoInfer, SLIDE, RABT, DRUT, and iReckon can quantify the known annotation and at the same time predict and quantify novel isoforms in known gene loci. RABT quantifies known and novel isoforms, taking into account existing gene annotations and using the same graph assembly algorithm of Cufflinks, combining the sequencing reads with reads obtained by fragmenting known transcripts. RABT is part of the Cufflinks distribution, but here we distinguish it from the original Cufflinks, which quantifies abundances of either only annotated or only novel isoforms [46, 52]. Similar to RABT, SLIDE uses RNA-Seq data and existing gene annotation to discover novel isoforms and to estimate the abundance of known

and new isoforms. Additionally, it can use other sources of evidence, like RACE, CAGE, and EST, or even the output from other isoform reconstruction algorithms. IsoInfer uses the transcript start and end sites, plus exon–intron boundaries to enumerate all possible isoforms, estimate their expression levels and then choose the subset of isoforms that best explain the observed reads, predicting novel isoforms from the existing exon data. On the other hand, iReckon can work with just transcript start and end sites or with full annotations; it models multimapped reads, intron-retention and unspliced pre-mRNAs and performs reconstruction and quantification simultaneously. DRUT uses a modified version of the IsoEM algorithm [56] in combination with a de novo reconstruction method similar to Cufflinks to complete partial existing annotations as well as to estimate isoform frequencies. Casper, similar to Cufflinks, estimates abundances of known or novel isoforms separately, but unlike other methods, uses information of the connectivity of more than two exons. Generally, known isoform quantification methods show a high level of agreement with experimental validation [54] and can be improved using annotation-guided methods for read mapping [29].

3.2.2 Isoform Quantification Guided by a Transcriptome

A number of methods consider reads mapped to a transcriptome for isoform quantification (Subheading 2.3); these include RSEM [55], IsoEM [56], NEUMA [57], BitSeq [58], MMSEQ [59], and eXpress [60]. Although these methods depend on a transcriptome annotation, they can use a standard (non-spliced) mapper to obtain the input data. Additionally, they can work also with predicted isoforms from transcript assembly methods (Fig. 1). All of them provide a measure of global isoform expression, similar to RPKM. Moreover, RSEM also calculates the fraction of transcripts represented by the isoform, equivalent to PSI. RSEM and IsoEM use both an Expectation–Maximization algorithm and model paired-end fragment size. RSEM models the mapping uncertainty to transcripts and provides confidence intervals of the abundance estimates. IsoEM uses the fragment-size information to disambiguate the assignment of reads to isoforms. BitSeq is based on a Bayesian approach, incorporates the mapping step to the transcriptome, models the nonuniformity of reads, and provides an expression value per isoform. BitSeq can also be used for differential isoform expression (see below). MMSEQ also takes into account the nonuniform read distribution and deconvolutes the mapping to isoforms to estimate isoform-expression and haplotype-specific isoform-expression. The method eXpress is in fact a general tool for quantifying abundances of a set of sequences in a generic experiment and can be used with a reference genome or transcriptome. For RNA-Seq reads mapped to a transcriptome, eXpress provides isoform quantification in terms of FPKM. Finally, NEUMA is different from the other methods, as it does not use any probabilistic

description and assumes uniformity of the reads along transcripts. NEUMA labels reads according to whether they are isoform or gene specific and calculates a measure of isoform quantification defined as the number of fragments per virtual kilobase per million reads (FVKM). Transcript-based methods can be generally applied to the transcripts obtained from genome annotations, so that the correspondence of transcripts to gene loci is maintained. Additionally, they can be used in combination with de novo transcript assembly methods (see below) to estimate isoform abundance in genomes without a reference.

3.2.3 Genome-Based Transcript Reconstruction and Quantification Without Annotation

These methods use the reads mapped to the genome to reconstruct isoforms de novo. They are generally based on previous approaches to transcript reconstruction from ESTs [114–117]. As for ESTs [118], accuracy is limited by the lengths of the input reads; hence, the use of paired-end sequencing becomes crucial. Additionally, as RNA abundance spans a wide range of values, the correct recovery of lowly expressed isoforms requires sufficient sequencing coverage. Although these methods work independently of the mapping procedure, they strongly rely on the accuracy of the spliced-mapper.

Purely reconstruction methods, without isoform quantification, include G-Mo.R-Se [61] and assemblySAM [62]. Methods that reconstruct isoforms as well as estimate their abundances include Cufflinks [46], Casper [47], CEM [48], IsoLasso [49], TAU [63], Scripture [64], Montebello [65], and NSMAP [66]. G-Mor.R-Se, Scripture, and TAU proceed in a similar way by first obtaining candidate exons from read-clusters and then connecting them using reads spanning exon–exon junctions. Subsequently, all possible isoforms from the graph of connected exons are computed. As they explore all possible connections between potential exons, they ensure a high sensitivity but at the cost of a high false-positive rate. In contrast, Cufflinks first connects predicted exons trying to identify the minimum number of possible isoforms using a graph generated from the reads; expression levels are then calculated using a statistical model [42]. IsoLasso also tries to obtain the minimal set of isoforms from predicted exons, but maximizing the number of reads included in each isoform. CEM model takes into account positional, sequencing and mappability biases of the RNA-Seq. Casper follows a heuristics similar to Cufflinks but exploiting the reads that connect more than 2 exon. Some of these methods use paired-end reads and/or model the insert-size distribution, which improve the reconstruction accuracy [119]. NSMAP, IsoLasso, and Montebello perform identification of the exonic structures and estimation of the isoform expression levels simultaneously in a single probabilistic model; iReckon does so too, but was not included in this section as it needs at least the transcript start and end positions. The rest of methods perform reconstruction and quantification independently.

Although reasonable overlap among methods has been reported [120], there are still many predictions unique to each method. Interestingly, given a fixed number of sequenced bases, sequencing longer reads does not seem to lead to more accurate quantifications [55, 56], although exonic structures may be better predicted [48]. These de novo reconstruction and quantification methods seem a good option for finding novel isoforms [64], alternatively spliced genes in a genome with partial annotation [61] and for quantifying isoforms under various conditions [46]. However, they depend much on coverage. Accordingly, if the aim is to obtain the expression of known isoforms, gene-based methods may be a better choice. Alternatively, for protein-coding gene finding there are other options available, as discussed next.

3.2.4 Evidence-Based Alternatively Spliced Gene Prediction

The methods described above are mainly focused on isoform quantification, using available annotation, or on the de novo reconstruction and quantification of isoforms, using reads mapped to the genome. Quantification methods based solely on gene annotations could miss many novel genes and isoforms, whereas de novo approaches not using annotations may produce many false positives. Combined approaches that discover novel isoforms in known and new loci and, at the same time, quantify them, could help improving the gene annotation. Some of the annotation-based quantification methods can also reconstruct and quantify new isoforms in known gene loci (Subheading 2.5): IsoInfer [50], SLIDE [51], RABT [52], DRUT [53], and iReckon [54]. Some of these methods can work with partial evidence, like iReckon. However, they do not predict new isoforms in new gene loci. To this end, a number of methods can use RNA-Seq and other sources of evidence to predict the exon–intron structures of isoforms, or even to predict full protein-coding gene structures. These methods include (Subheading 2.5) TAU [63], SpliceGrapher [67], mGene [69], and the method described in ref. 68. The method mGene is an SVM-based gene predictor (*see*, e.g., [121]) that first reconstructs a high-quality gene set, which then uses to train a gene model that is applied using RNA-Seq data in addition to the previously determined genomic signal predictors. In contrast, SpliceGrapher and TAU incorporate into the same graph model information from ESTs and RNA-Seq reads to complete known gene annotations and produce novel variants. ExonMap/JunctionWalk proposed in ref. [68] combine SpliceMap [10] alignments with known annotations to complete known isoforms and obtain novel ones without quantification (Fig. 1).

Some of these methodologies are reminiscent of the evidence-based gene prediction methods. These are generally based on probabilistic models of protein-coding genes, which can incorporate external spliced evidence like ESTs and cDNAs into the model to guide the prediction of the exon–intron structure, and some of which can predict multiple isoforms in a gene

locus. Accordingly, evidence-based gene prediction methods could still be useful for splicing analysis from RNA-Seq. In particular, Augustus [70] is an evidence-based protein-coding gene prediction method, capable of finding multiple isoforms per gene, which has been shown to be highly accurate using a blind test set [122, 123]. Other evidence-based prediction methods include (Subheading 2.5) GAZE [71], JigSaw [72], EVM [73], and Evigan [74]. Although these four methods do not explicitly model alternative isoforms, they can still produce multiple transcripts in a locus.

Evidence-based gene prediction methods can take as input transcripts reconstructed by other methods and generate protein-coding isoforms. They do not provide a quantification of isoforms, but in combination with quantification methods (Subheadings 2.2 and 2.3) they could be a powerful approach to annotate and quantify alternatively spliced protein-coding genes from newly sequenced genomes using RNA-Seq data.

3.2.5 De Novo Transcript Assembly

De novo transcript assemblers put together reads into transcriptional units without mapping the reads to a genome reference, similar to building Unigene clusters from ESTs prior to having a genome reference [124]. A transcriptional unit can be defined as the set of RNA sequences that are transcribed from the same genome locus and share some sequence, i.e., the set of RNA isoforms from the same gene. This is generally represented as a sequence-based graph, where paths along the graph potentially resolve the different isoforms. Methods for transcript assembly include (Subheading 2.6) Rnnotator [75], STM [76], OASES [77], SOAPdenovo-trans [78], TransAbyss [79], Trinity [80], and Kissplice [81]. Although KisSplice focuses on recovering alternative splicing events, we include it here as it follows a similar approach to the other methods. *See* ref. 125 for a recent comparison between some of these methods.

The main challenge of these methods is not only to distinguish sequence errors from polymorphisms but also to distinguish close paralogues from alternative isoforms, which requires correctly capturing the exonic variability. All these methods are based on a graph built from *k*-mer overlaps between read sequences. The choice of *k*-mer length affects the assembly, being more sensitive at low values of *k* and more specific at high values. Accordingly, some use a variable *k*-mer approach. Isoforms are recovered as paths through the graph with sufficient read coverage. Not all methods can provide multiple isoforms from the same gene (Subheading 2.6).

Genome-independent methods are useful when there is no genome reference sequence available, and could also be valuable when the RNA is expected to contain much variation, like in a cancer cell with many copy number alterations, mutations and genome rearrangements compared to the reference genome. De novo assembly methods tend to be more sensitive to sequencing errors

and low coverage, and generally require more computational resources, although full parallelization of the graph algorithms can alleviate this issue [126]. Some of the methods also consider the comparison to reference sets of DNA or protein sequences [76]. In fact, mapping assembled transcripts to a reference genome, even from a related species, seems to improve accuracy in transcript quantification [127]. KisSplice is explicitly designed to obtain and quantify de novo alternative splicing events, which may potentially be coupled with other methods to study differential splicing. On the other hand, OASES, TransAbyss, Trinity, and SOAPdenovo-trans can produce multiple isoforms, but only TransAbyss and Trinity perform quantification. Nonetheless, multiple assembled isoforms can be quantified with transcript-based methods (Subheading 2.3) or further processed with isoform-based differential expression methods (Subheading 2.8).

3.3 Comparing Splicing Across Samples

The comparison of events and isoforms across two or more conditions provide valuable information to understand the regulation of alternative splicing. However, it is important to distinguish differential isoform relative abundance, from differential isoform expression. Changes in relative abundance of isoforms, regardless of the expression change, indicate a splicing-related mechanism. On the other hand, there can be measurable changes in the expression of isoforms across samples, without necessarily changing the relative abundance, which possibly indicates a transcription-related mechanism. With this in mind, we can consider two types of methods, those that measure relative event or isoform usage (Subheading 2.7) and those that measure isoform-based changes in expression (Subheading 2.8).

3.3.1 Differential Splicing

Most of these methods are focused on splicing events, thereby summarizing the isoform relative abundance into two possible splicing outcomes in a local region of the gene (Fig. 1). They use a predetermined set of splicing events, generally calculated from gene annotations and additional EST and cDNA data; hence, they are suitable for studying splicing variation in well-annotated genomes. They all consider exon-skipping events (cassette exons), and some also include alternative 5′ and 3′ splice-sites, mutually exclusive exons and retained introns; and in very few cases, multiple-cassette exons, alternative first exons and alternative last exons [38]. Potential novel events are sometimes built by considering hypothetical exon–exon junctions from the annotation [85].

Methods that calculate differential relative abundance of events or exons under at least two conditions include (Subheading 2.7) SpliceSeq [33], MISO [38], ALEXA-Seq [39], SOLAS [40], DEXSeq [82], DSGSeq [83], GPSeq [84], MATS [85], JuncBase [86], JETTA [87], SplicingCompass [88], DiffSplice [89], FDM [90] rDiff [91, 128], and the methods from ref. 129. ALEXA-Seq

estimates inclusion levels on a set of pre-calculated events using only unambiguous reads, i.e., reads that map to one unique event, and calculates various measures of differential expression, including the splicing index, i.e., a measure of change in expression of an event between two conditions relative to the change in expression of the entire gene locus between the same two conditions. On the other hand, SOLAS uses single-reads and only takes into account those mapping within exons, disregarding reads spanning exon–exon junctions, to detect differentially spliced events between two conditions. DEXSeq, DSGSeq, and GPSeq use read counts on exons to calculate those genes with differential splicing between two conditions. They do not provide any event or isoform information and report the exons with significant change (Fig. 1). MATS and MISO use both a Bayesian approach to calculate the differential inclusion of splicing events between two samples, using reads that map to exons and to the inclusion and skipping exon junctions. JuncBASE also uses reads mapped to exon junctions and uses a Fisher exact test to compare the read count in the inclusion and exclusion forms in two conditions. JETTA estimates the differential inclusion between two conditions from pre-calculated expression values for genes, exons, and junctions, which the authors obtain using SeqMap [130] and rSeq [49]. SpliceSeq calculates read coverage along genes, exons, and junctions for each sample, which are then compared to identify significant changes in splicing across samples. SpliceSeq also includes the evaluation of the impact of alternative splicing on protein products and a visualization of the events (see below). Besides all these methods, various methods were proposed in ref. 129 based on reads over exon junctions to find robust estimates of PSI, taking into account the positional bias of reads relative to the junction.

Some of these methods can also measure the change in the relative abundance of isoforms (Fig. 1): MISO can measure changes in isoform relative abundances from previously calculated isoform PSI values; ALEXA-Seq uses the events that are differentially expressed to infer isoform abundance differences between two conditions. Finally, rDiff, FDM, and DiffSplice are methods that work with a more general definition of event and that can operate without an annotation. FDM and DiffSplice are graph-based methods and both identify regions of differential abundance of transcripts between two samples using the variability of reads that define a splicing graph. Similarly, rDiff uses a Maximum Mean Discrepancy test [131] to estimate regions that have a significant distance between the read distributions in the two conditions. Alternatively, rDiff can work with an annotation; it considers reads in exonic regions that are not in all isoforms and groups those regions according to whether they occur in the same set of isoforms. Finally, SplicingCompass uses a geometric approach to detect differentially spliced genes and quantifies relative exon usage.

In summary, these methods test whether events, isoforms, or genic regions, change their relative abundances between two or more conditions, and so directly address the question of differential splicing.

When comparing two or more conditions, biological variability becomes an important issue, which has been shown to be relevant for studying expression [132] and splicing [82] from RNA-Seq data. However, not all methods take this into account. From the methods described here, DEXSeq, DSGSeq, GPSeq, DiffSplice, FDM, rDiff, and a newer version of MATS accept multiple replicates and model biological variability in different ways. In contrast, the initial methods for calculating splicing changes from RNA-Seq data [2, 3, 105], as well as MISO, ALEXA-Seq, JETTA, SpliceSeq, SOLAS, and SplicingCompass, do not work with multiple replicates. On the other hand, JuncBASE can work with replicated data but does not seem to model variability. As the cost of sequencing continue to decrease, it will be more common to include replicates in the differential splicing analysis, which will prove relevant to discern actual regulatory changes from biological variability.

3.3.2 Isoform-Based Differential Expression

Current methods to study differential splicing at the event level show a high validation rate [2, 85]. However, their agreement with microarray-based methods is not as high as one may expect [2]. This limitation could be due to the simplification of considering only events, rather than full RNA isoforms. An improvement in this direction would be to quantify changes in isoform expression. A possible approach is to combine methods that quantify isoforms with methods for differential gene expression. However, as previously pointed out [5, 90, 93], this may be problematic, since tools for differential gene expression analysis do not generally take into account the uncertainty of mapping reads to isoforms. We will not discuss here the many methods that have been proposed to study differential gene expression analysis from RNA-Seq data; for a recent review *see* refs. 5, 133.

A number of methods have been proposed to detect expression changes at the isoform level (Subheading 2.8): BitSeq [58], BASIS [92], Cuffdiff2 [93], and EBSeq [94]. Cuffdiff2, BitSeq, and EBSeq take into account the read-mapping uncertainty, accept multiple replicates and model biological variability. BASIS does not accept replicates, but it models variability along genes. Cuffdiff2 and BitSeq provide quantification and differential expression of isoforms from genome-mapped and transcriptome-mapped reads, respectively. Cuffdiff2 can use reads directly mapped to the genome or can use the results from Cufflinks on two conditions after using cuffcompare [46] (Fig. 1), which gives equivalent transcripts in both conditions. On the other hand, EBSeq relies on the isoform quantification from other methods, like RSEM or Cufflinks, and is actually included in the current release of RSEM;

whereas BASIS uses coverage over exon regions that are isoform-specific to calculate differential expression of isoforms. These methods rely on an annotation, either genome-based (Cuffdiff2, BASIS, and EBSeq) or transcriptome-based (BitSeq and EBSeq). Except for Cuffdiff2, these methods do not explicitly address the question of whether the relative abundance of these isoforms change across samples (Fig. 1). Accordingly, if there is an increase of transcription but the relative abundance of isoforms remain constant, they can detect changes in isoform expression, even though there might not be an actual change in splicing. On the other hand, if there are changes in the relative abundance of isoforms, they may possibly detect expression changes, but they will not provide information about the change of the relative abundances, and therefore do not directly address the question of differential splicing.

3.4 Visualizing Alternative Splicing

Being able to visualize the complexity of alternative splicing is an important aspect of the analysis. In the past, there have been multiple efforts to store and visualize alternative isoforms from ESTs and cDNAs [134, 135]. Visualization for RNA-Seq requires specialized tools that can efficiently process large amount of data from multiple samples. This has triggered the development of specialized tools to visualize alternative isoforms and events from RNA-Seq data (Subheading 2.9). Perhaps the simplest way to visualize isoforms and events is to generate track files for a genome browser. For instance, RSEM produces WIG files that can be viewed as tracks in the UCSC browser [136]. Similarly, SpliceGrapher and DiffSplice produce files in GFF-like formats (http://gmod.org/wiki/GFF), which can be uploaded into visualization tools like GBrowse [137] or Apollo [138]. On the other hand, SpliceGrapher and Alexa-Seq have their own visualization utilities. Other tools have been developed independently from the analysis method. For instance, the Sashimi plot toolkit to visualize isoforms and events and their relative coverage was used with MISO but can be used with the results from other tools (Subheading 2.8). Similarly, the browser Savant [95] has been used in conjunction with iReckon, but can be used independently for multiple HTS data formats. Finally, SpliceSeq [33] and SplicingViewer [96] are stand-alone tools that, besides mapping reads and quantifying events, also provide a visualization of results.

4 Conclusions and Outlook

The rapid development of short-read RNA sequencing technologies has triggered the development of new methods for data analysis. In this review, we have tried to provide an overview of methods applicable to the study of alternative splicing. These provide a way to detect and quantify exon–exon junctions, transcript isoforms,

and differential splicing. Despite the many tools available not all are necessarily applicable to every purpose. For instance, for genomes with good annotation coverage, like human, the expression of known isoforms and possibly their changes under several conditions might be more accurately assessed using annotation-guided methods. Similarly, if sufficient annotation is available, there are also hybrid methods that can quantify known isoforms and predict novel ones simultaneously. For newly sequenced genomes, there are effective methods to perform de novo reconstruction and quantification of isoforms. However, if one is specifically interested in protein-coding genes, there are also evidence-based gene prediction methods available, which can be quite effective for isoform prediction.

One can identify some open questions and areas of improvement. For instance, not all of the de novo transcript assembly methods describe multiple isoforms per gene and only few actually quantify them. These are still two hard problems to solve, as incompleteness or absence of transcriptomes can lead to many reconstruction and quantification errors [139]. There are different approaches to improve these questions, either by a combination of methods and homology searches [140] or by using error correction of sequencing reads before assembly [141]. These tools are of great relevance for non-model organisms and we will probably see substantial improvements in the near future. Accurate reconstruction and quantification of isoforms is crucial for downstream analysis and in particular, for differential analysis of isoform abundances. Methods to estimate differential splicing at the event level seem to provide accurate measures as shown by experimental validation. However, differential expression at the isoform level is still an active area of development.

Extending de novo transcriptome assembly methods to calculate differential expression of isoforms between two or more conditions could facilitate the analysis of isoform expression for non-model organisms. Although this may be done currently with a combination of methods, a tool that integrates all these could provide a powerful approach to study expression and splicing in tumor samples, where multiple genome rearrangements and copy number alterations are expected to have occurred. On a different direction, considering that a reference genome sequence does not represent all DNA that can be possibly transcribed in a cell, unmapped RNA reads may come from functional RNAs not represented in the genome annotation. Tools that map reads to a genome reference and simultaneously attempt to perform transcript assembly will be also quite useful to perform systematic analyzes of RNA in cancer samples as well as in genomes that are partly assembled.

Besides the technical improvements, there is probably also a need to improve the comparison and evaluation of current methods. Transcript reconstruction methods should be evaluated using

manual gene annotation sets, as proposed previously for gene prediction methods [123] and currently by RGASP for RNA-Seq based methods (http://www.gencodegenes.org/rgasp). Additionally, these comparisons should use measures that take into account alternative splicing [123, 142]. Similarly, there is the need to develop an experimental gold standard dataset for isoform quantification and differential isoform expression [143].

As a final question, we may ask for how long some of these methods will be needed. There are new technologies for single-molecule sequencing that soon will be used to probe the transcriptome. This may preclude the need to perform reconstruction of isoforms. Nonetheless, short-read RNA-Seq may still be necessary for efficient quantification. On the other hand, single-molecule sequencing technologies will open up a whole new set of problems, like that of reconciling new cell-specific RNA sequences with the information available for the genome sequence and its annotation. In fact, we will be in the position to quantify multiple transcriptomes and to revisit previous studies of differential splicing and expression in cancer, as the DNA and transcription complexity of the tumor cell is fully revealed.

With this review, we have aimed to provide an overview of the different tools to study different aspects of alternative splicing from RNA-Seq data, organized such that it is useful for the end user to navigate through the list of methods. All of them have their advantages and disadvantages, but are certainly useful to answer specific questions. We also hope that this review makes it easier to identify the tools that are still missing in order to improve the study of splicing with RNA-Seq.

Acknowledgements

We thank Y. Xing, K. Hertel, J.R. González, M. Kreitzman, and P. Drewe for comments and suggestions. This work was supported by the Spanish Ministry of Science with grants BIO2011-23920 and CSD2009-00080 and by Sandra Ibarra Foundation for Cancer with grant FSI 2011-035.

References

1. Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489(7414):101–108
2. Wang ET, Sandberg R, Luo S et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456(7221):470–476
3. Pan Q, Shai O, Lee LJ et al (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40(12): 1413–1415
4. Chen L (2011) Statistical and computational studies on alternative splicing. In: Horng-Shing Lu H et al (eds) Handbook of statistical bioinformatics. Springer, New York. doi:10.1007/978-3-642-16345-6_2

5. Pachter L (2011) Models for transcript quantification from RNA-Seq. arXiv:1104.3889v2 (http://arxiv.org/abs/1104.3889)
6. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111
7. Huang S, Zhang J, Li R et al (2011) SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. Front Genet 2(July):46
8. Zhang Y, Lameijer EW, 't Hoen PA et al (2012) PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. Bioinformatics 28(4):479–486
9. Wang K, Singh D, Zeng Z et al (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38(18):e178
10. Au KF, Jiang H, Lin L et al (2010) Detection of splice junctions from paired-end RNA seq data by SpliceMap. Nucleic Acids Res 38(14):4570–4578
11. Dimon MT, Sorber K, DeRisi JL (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. PloS one 5(11):e13875
12. Li Y, Li-Byarlay H, Burns P et al (2013) TrueSight: a new algorithm for splice junction detection using RNA-seq. Nucleic Acids Res 41(4):e51
13. Marco-Sola S, Sammeth M, Guigó R et al (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9(12):1185–1188
14. Ameur A, Wetterbom A, Feuk L et al (2010) Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol 11(3):R34
15. Bryant DW, Shen R, Priest HD et al (2010) Supersplat– spliced RNA-seq alignment. Bioinformatics 26(12):1500–1505
16. Wang L, Wang X, Wang X et al (2011) Observations on novel splice junctions from RNA sequencing data. Biochem Biophys Res Commun 409(2):299–303
17. Lou SK, Ni B, Lo LY et al (2011) ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. Bioinformatics 27(3):421–422
18. Bao H, Xiong Y, Guo H et al (2009) MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. BMC Genomics 10(Suppl 3):S13
19. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21
20. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881
21. De Bona F, Ossowski S, Schneeberger K et al (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24(16): i174–i180
22. Jean G, Kahles A, Sreedharan VT et al. (2010) RNA-Seq read alignments with PALMapper. Curr Protoc Bioinformat Chapter 11:Unit 11.6
23. Philippe N, Salson M, Commes T et al (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. Genome Biol 14(3):R30
24. Wu J, Anczuków O, Krainer AR et al (2013) OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucl Acids Res 41(10):5149–5163
25. Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41(10):e108
26. Hu J, Ge H, Newman M, Liu K (2012) OSA: a fast and accurate alignment tool for RNA-Seq. Bioinformatics 28(14):1933–1934
27. Wood DL, Xu Q, Pearson JV et al (2011) X-MATE: a flexible system for mapping short read data. Bioinformatics 27(4):580–581
28. Chen LY, Wei KC, Huang AC et al (2012) RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Res 40(6):e42
29. Labaj PP, Linggi BE, Wiley HS et al (2012) Improving RNA-Seq Precision with MapAl. Front Genet 3:28
30. Xu G, Deng N, Zhao Z et al (2011) SAMMate: a GUI tool for processing short read alignments in SAM/BAM format. Source Code Biol Med 6(1):2
31. Kim H, Bi Y, Pal S et al (2011) IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-seq data. BMC Bioinforma 12:305
32. Grant GR, Farkas MH, Pizarro AD et al (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27(18): 2518–2528
33. Ryan MC, Cleland J, Kim R et al (2012) SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics 28(18):2385–2387
34. Tang S, Riva A (2013) PASTA: splice junction identification from RNA-Sequencing data. BMC Bioinforma 14(1):116
35. Bonfert T, Csaba G, Zimmer R et al (2012) A context-based approach to identify the most likely mapping for RNA-seq experiments. BMC Bioinforma 13(Suppl 6):S9
36. Wang L, Xi Y, Yu J et al (2010) A statistical method for the detection of alternative splicing using RNA-seq. PLoS one 5(1):e8529

37. Wu J, Akerman M, Sun S et al (2011) SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics 27:3010–3016
38. Katz Y, Wang ET, Airoldi EM et al (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7(12):1009–1015
39. Griffith M, Griffith OL, Mwenifumbo J et al (2010) Alternative expression analysis by RNA sequencing. Nat Methods 7(10):843–847
40. Richard H, Schulz MH, Sultan M et al (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucl Acids Res 38(10):e112
41. Mortazavi A, Williams BA, Mccue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):1–8
42. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25(8):1026–1032
43. Bohnert R, Behr J, Rätsch G (2009) Transcript quantification with RNA-Seq data. BMC Bioinforma 10(Suppl 13):P5
44. Montgomery SB, Sammeth M, Gutierrez-Arcelus M et al (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464(7289): 773–777
45. Du J, Leng J, Habegger L et al (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One 7(1):e29175
46. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515
47. Rossell D, Attolini CSO, Kroiss M et al. (2012) Quantifying alternative splicing from paired-end RNA-sequencing data. COBRA Preprint Series. Working Paper 97 http://biostats.bepress.com/cobra/art97
48. Li W, Jiang T (2012) Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics 28(22):2914–2921
49. Li W, Feng J, Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18(11):1693–1707
50. Feng J, Li W, Jiang T (2010) Inference of isoforms from short sequence reads. In: Berger B (ed) Research in computational molecular biology, lecture notes in computer science, vol 6044. Springer, Heidelberg, pp 138–157
51. Li JJ, Jiang CR, Brown JB et al (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. PNAS 108(50):19867–19872
52. Roberts A, Pimentel H, Trapnell C et al (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27(17):2325–2329
53. Mangul S, Caciula A, Glebova O et al (2012) Improved transcriptome quantification and reconstruction from RNA-Seq reads using partial annotations. Silico Biol 11(5): 251–261
54. Mezlini AM, Smith EJ, Fiume M et al (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res 23(3):519–529
55. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma 12:323
56. Nicolae N, Mangul S, Mandoiu I et al (2011) Estimation of alternative splicing isoform frequencies from RNA-seq data. Algorithms Mol Biol 6:9
57. Lee S, Seo CH, Lim B et al (2011) Accurate quantification of transcriptome from RNA-seq data by effective length normalization. Nucleic Acids Res 39(2):e9
58. Glaus P, Honkela A, Rattray M (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28(13):1721–1728
59. Turro E, Su SY, Gonçalves Â et al (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol 12(2):R13
60. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10(1): 71–73
61. Denoeud F, Aury JM, Da Silva C et al (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol 9(12):R175
62. Zhao Z, Nguyen T, Deng N et al. (2011) SPATA: a seeding and patching algorithm for de novo transcriptome assembly. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshop (IEEE BIBMW'11) pp. 26–33
63. Filichkin S, Priest H, Givan S et al (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20(1): 45–58
64. Guttman M, Garber M, Levin JZ et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28(5):503–510
65. Hiller D, Wong WH (2012) Simultaneous isoform discovery and quantification from RNA-Seq. Stat Biosci 5(1):100–118

66. Xia Z, Wen J, Chang CC et al (2011) NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. BMC Bioinforma 12:162
67. Rogers MF, Thomas J, Reddy AS et al (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biol 13(1):R4
68. Seok J, Xu W, Jiang H et al (2012) Knowledge-based reconstruction of mRNA transcripts with short sequencing reads for transcriptome research. PLoS ONE 7(2):e31440
69. Behr J, Bohnert R, Zeller G et al (2010) Next generation genome annotation with mGene. ngs. BMC Bioinforma 11(Suppl 10):O8
70. Stanke M, Schöffmann O, Morgenstern B et al (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinforma 7:62
71. Howe KL, Chothia T, Durbin R (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res 12(9):1418–1427
72. Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21(18): 3596–3603
73. Haas BJ, Salzberg SL, Zhu W et al (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9(1):R7
74. Liu Q, Mackey AJ, Roos DS et al (2008) Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics 24(5):597–605
75. Martin J, Bruno VM, Fang Z et al (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11:663
76. Surget-Groba Y, Montoya-Burgos J (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20(10):1432–1440
77. Schulz MH, Zerbino DR, Vingron M et al (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092
78. Xie Y, Wu G, Tang J et al. (2013) SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. arXiv:1305.6760 [q-bio.GN] (http://arxiv.org/abs/1305.6760)
79. Robertson G, Schein J, Chiu R et al (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7(11):909–912
80. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652
81. Sacomoto GA, Kielbassa J, Chikhi R et al (2012) KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinforma 13(Suppl 6):S5
82. Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res 22(10): 2008–2017
83. Wang W, Qin Z, Feng Z et al (2013) Identifying differentially spliced genes from two groups of RNA-seq samples. Gene 518(1):164–170
84. Srivastava S, Chen L (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res 38(17):e170
85. Shen S, Park JW, Huang J et al (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res 40(8):e61
86. Brooks AN, Yang L, Duff MO et al (2011) Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res 21(2):193–202
87. Seok J, Xu W, Gao H et al (2012) JETTA: junction and exon toolkits for transcriptome analysis. Bioinformatics 28(9):1274–1275
88. Aschoff M, Hotz-Wagenblatt A, Glatting KH et al (2013) SplicingCompass: differential splicing detection using RNA-Seq data. Bioinformatics 29(9):1141–1148
89. Hu Y, Huang Y, Du Y et al (2013) DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res 41(2):e39
90. Singh D, Orellana CF, Hu Y et al (2011) FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics 27(19):2633–2640
91. Drewe P, Stegle O, Hartmann L et al (2013) Accurate detection of differential RNA processing. Nucl Acids Res 41(10):5189–5198
92. Zheng S, Chen L (2009) A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res 37(10):e75
93. Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31(1):46–53
94. Leng N, Dawson JA, Thomson JA et al (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29(8):1035–1043
95. Fiume M, Williams V, Brook A et al (2010) Savant: genome browser for high-throughput sequencing data. Bioinformatics 26(16): 1938–1944

96. Liu Q, Chen C, Shen E et al (2012) Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer. Genomics 99(3):178–182
97. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinforma 6:31
98. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12(4):656–664
99. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875
100. Fonseca NA, Rung J, Brazma A et al (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28(24):3169–3177
101. Garber M, Grabherr MG, Guttman M et al (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8(6):469–477
102. Schneeberger K, Hagmann J, Ossowski S et al (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biol 10(9):R98
103. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
104. Clark TA, Sugnet CW, Ares M Jr (2002) Genome wide analysis of mRNA processing in yeast using splicing-specific microarrays. Science 296(5569):907–910
105. Sultan M, Schulz MH, Richard H et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321(5891):956–960
106. Cloonan N, Forrest ARR, Kolle G et al (2008) Stem cell transcriptome profiling via massive scale mRNA sequencing. Nat Methods 5(7):613–619
107. Cloonan N, Xu Q, Faulkner GJ et al (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics 25(19):2615–2616
108. Tang F, Barbacioru C, Wang Y et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6(5):377–382
109. Chen L (2012) Statistical and computational methods for high-throughput sequencing data analysis of alternative splicing. Stat Biosci 5(1):138–155
110. Venables JP, Klinck R, Bramard A et al (2008) Identification of alternative splicing markers for breast cancer. Cancer Res 68(22): 9525–9531
111. Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25(15):1966–1967
112. Deng N, Puetter A, Zhang K et al (2011) Isoform-level microRNA-155 target prediction using RNA-seq. Nucleic Acids Res 39(9):e61
113. Nguyen TC, Deng N, Zhu D (2013) SASeq: a selective and adaptive shrinkage approach to detect and quantify active transcripts using RNA-Seq. arXiv:1208.3619v2 [q-bio.QM] (http://arxiv.org/abs/1208.3619v2)
114. Heber S, Alekseyev M, Sze SH et al (2002) Splicing graphs and EST assembly problem. Bioinformatics 18(Suppl 1):S181–S188
115. Haas BJ, Delcher AL, Mount SM et al (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666
116. Xing Y, Resch A, Lee C (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. Genome Res 14(3):426–441
117. Xing Y, Yu T, Wu YN et al (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Res 34(10):3150–3160
118. Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8(1):6–21
119. Salzman J, Jiang H, Wong WH (2011) Statistical modeling of RNA-Seq data. Stat Sci 26(1):62–83
120. Li B, Ruotti V, Stewart R et al (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493–500
121. Sonnenburg S, Schweikert G, Philips P et al (2007) Accurate splice site prediction using support vector machines. BMC Bioinforma 8(Suppl 10):S7
122. Stanke M, Keller O, Gunduz I et al (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34(Web Server issue):W435–W439
123. Guigó R, Flicek P, Abril JF et al (2006) EGASP: the human ENCODE genome annotation assessment project. Genome Biol 7(Suppl 1):S2.1–31
124. Pontius JU, Wagner L, Schuler GD (2003) UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/books/NBK21083/
125. Zhao QY, Wang Y, Kong YM et al (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinforma 12(Suppl 14):S2
126. Jackson B, Schnable P, Aluru S (2009) Parallel short sequence assembly of transcriptomes. BMC Bioinforma 10(Suppl 1):S14
127. Vijay N, Poelstra JW, Künstner A et al (2013) Challenges and strategies in transcriptome

assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol 22(3):620–634
128. Stegle O, Drewe P, Bohnert R et al (2010) Statistical tests for detecting differential rna-transcript expression from read counts. Nat Preced. doi:10.1038/npre.2010.4437.1
129. Kakaradov B, Xiong HY, Lee LJ et al (2012) Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. BMC Bioinforma 13(Suppl 6):S11
130. Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24(20):2395–2396
131. Borgwardt KM, Gretton A, Rasch MJ et al (2006) Integrating structured biological data by Kernel Maximum Mean Discrepancy. Bioinformatics 22(14):e49–e57
132. Hansen KD, Wu Z, Irizarry RA et al (2011) Sequencing technology does not eliminate biological variability. Nat Biotechnol 29: 572–573
133. Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol 11(12):220. doi:10.1186/gb-2010-11-12-220
134. Bhasi A, Philip P, Sreedharan VT et al (2009) AspAlt: A tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. Genomics 94(1):48–54
135. Martelli PL, D'Antonio M, Bonizzoni P et al (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. Nucleic Acids Res 39(Database issue):D80–D85
136. Karolchik D, Hinrichs AS, Kent WJ (2012) The UCSC Genome Browser. Curr Protoc Bioinformatics Chapter 1:Unit1.4
137. Donlin MJ. (2009) Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics, Chapter 9:Unit 9.9
138. Lee E, Harris N, Gibson M et al (2009) Apollo: a community resource for genome annotation editing. Bioinformatics 25: 1836–1837
139. Pyrkosz AB, Cheng H, Brown CT. (2013) RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates. arXiv:1303.2411 [q-bio.GN] (http://arxiv.org/abs/1303.2411)
140. Birzele F, Schaub J, Rust W et al (2010) Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic Acids Res 38(12):3999–4010
141. MacManes MD, Eisen MB (2013) Improving transcriptome assembly through error correction of high-throughput sequence reads. arXiv:1304.0817 [q-bio.GN] (http://arxiv.org/abs/1304.0817) (3/April/2013)
142. Eyras E, Caccamo M, Curwen V et al (2004) ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res 14(5):976–987
143. Lovén J, Orlando DA, Sigova AA et al (2012) Revisiting global gene expression analysis. Cell 151(3):476–482