Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

High-throughput RNA sequencing (RNA-seq) promises a comprehensive picture of the transcriptome, allowing for the complete annotation and quantification of all genes and their isoforms across samples. Realizing this promise requires increasingly complex computational methods. These computational challenges fall into three main categories: (i) read mapping, (ii) transcriptome reconstruction and (iii) expression quantification. Here we explain the major conceptual and practical challenges, and the general classes of solutions for each category. Finally, we highlight the interdependence between these categories and discuss the benefits for different biological applications.

Defining a precise map of all genes along with their alternative isoforms and expression across diverse cell types is critical for understanding biology. Until recently, production of such data was prohibitively expensive and experimentally laborious. The major method for annotating a transcriptome required the slow and costly process of cloning cDNAs or expressed sequence tag (EST) libraries, followed by capillary sequencing[1–3]. Owing to the high cost and limited data yield intrinsic to this approach, it only provided a glimpse of the true complexity of cell type–specific splicing and transcription[4,5]. Analysis of these data required sophisticated computational tools, many of which[6–9] provide the basis for the programs used today for high-throughput RNA sequencing (RNA-seq) data. Alternative strategies, such as genome-wide tiling arrays, allowed for the identification of transcribed regions at a larger and more cost-efficient scale but with limited resolution[3,10]. Splicing arrays with probes across exon-exon junctions enabled researchers to analyze predefined splicing events[11,12] but could not be used to identify previously uncharacterized events. Expression quantification required hybridization of RNA to gene-expression microarrays, a process that is limited to studying the expression of known genes for defined isoforms[13,14].

Recent advances in DNA sequencing technology have made it possible to sequence cDNA derived from cellular RNA by massively parallel sequencing technologies, a process termed RNA-seq[5,15–23]. We use the term RNA-seq to refer to experimental procedures that generate DNA sequence reads derived from the entire RNA molecule. Specific applications such as small RNA sequence analysis require special approaches, which we do not address here. In theory, RNA-seq can be used to build a complete map of the transcriptome across all cell types, perturbations and states. To fully realize this goal, however, RNA-seq requires powerful computational tools. Many recent studies have applied RNA-seq to specific biological problems, including the quantification of alternative splicing in tissues[5], populations[5,24] and disease[25], discovery of new fusion genes in cancer[18,26], improvement of genome assembly[27], and transcript identification[16,23,28,29].

Here we focus on the computational methods needed to address RNA-seq analysis core challenges. First, we describe methods to align reads directly to a reference transcriptome or genome ('read mapping'). Second, we discuss methods to identify expressed genes and isoforms ('transcriptome reconstruction'). Third, we present methods for estimation of gene and isoform abundance, as well as methods for the analysis of differential expression across samples ('expression quantification').

Because of ongoing improvements in RNA-seq data generation, there is great variability in the maturity of available computational tools. In some areas, such as read mapping, a wealth of algorithms exists but in others, such as differential expression analysis, solutions are only beginning to emerge. Rather than comprehensively describing each method, we highlight the key common principles as well as the critical differences underlying

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to M.G. (mgarber@broadinstitute.org).

PUBLISHED ONLINE 27 MAY 2011; DOI:10.1038/NMETH.1613

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

[a]This list is not meant to be exhaustive as many different programs are available for short-read alignment. Here we chose a representative set capturing the frequently used tools for RNA-seq or tools representing fundamentally different approaches.

each approach and their application to RNA-seq analysis. We also discuss how these different methodologies can impact the results and interpretation of the data. Although we discuss each of the three categories as separate units, RNA-seq data analysis often requires using methods from all three categories. The methods described here are largely independent of the choice of library construction protocols, with the notable exception of 'paired-end' sequencing (reading from both ends of a fragment), which provides valuable information at all stages of RNA-seq analysis[28–30].

As a reference for the reader, we provide a list of currently available methods in each category (**Table 1**). To provide a general indication of the compute resources and tradeoffs of different methods, we selected a representative method from each category and applied it to a published RNA-seq dataset consisting of 58 million paired-end 76-base reads from mouse embryonic stem cell RNA[28] (**Supplementary Table 1**).

Mapping short RNA-seq reads

One of the most basic tasks in RNA-seq analysis is the alignment of reads to either a reference transcriptome or genome. Alignment of reads is a classic problem in bioinformatics with several solutions specifically for EST mapping[8,9]. RNA-seq reads, however, pose particular challenges because they are short (~36–125 bases), error rates are considerable and many reads span exon-exon junctions. Additionally, the number of reads per experiment is increasingly large, currently as many as hundreds of millions. There are two major algorithmic approaches to map RNA-seq reads to a reference transcriptome. The first, to which we collectively refer as 'unspliced read aligners', align reads to a reference without allowing any large gaps. The unspliced read aligners fall into two main categories, 'seed methods' and 'Burrows-Wheeler transform methods'. Seed methods[31–38] such as mapping and assembly with quality (MAQ)[33] and Stampy[35] find matches for short subsequences, termed 'seeds', assuming that at least

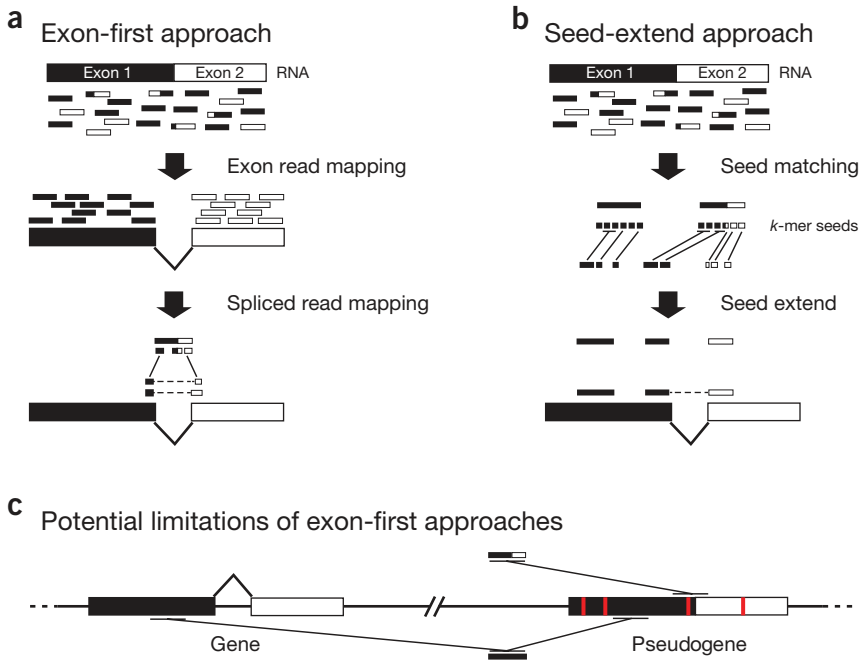

Figure 1 | Strategies for gapped alignments of RNA-seq reads to the genome. (**a**,**b**) An illustration of reads obtained from a two-exon transcript; black and gray indicate exonic origin of reads. Exon-first methods (**a**) map full, unspliced reads (exonic reads), and remaining reads are divided into smaller pieces and mapped to the genome. An extension process extends mapped pieces to find candidate splice sites to support a spliced alignment. Seed-and-extend methods (**b**) store a map of all small words (*k*-mers) of similar size in the genome in an efficient lookup data structure; each read is divided into *k*-mers, which are mapped to the genome via the lookup structure. Mapped *k*-mers are extended into larger alignments, which may include gaps flanked by splice sites. (**c**) A potential disadvantage of exon-first approaches illustrated for a gene and its associated retrotransposed pseudogene. Mismatches compared to the gene sequence are indicated in red. Exonic reads will map to both the gene and its pseudogene, preferring gene placement owing to lack of mutations, but a spliced read could be incorrectly assigned to the pseudogene as it appears to be exonic, preventing higher-scoring spliced alignments from being pursued.

one seed in a read will perfectly match the reference. Each seed is used to narrow candidate regions where more sensitive methods (such as Smith-Waterman) can be applied to extend seeds to full alignments. In contrast, the second approach includes Burrows-Wheeler transform methods[39–41] such as Burrows-Wheeler alignment (BWA)[40] and Bowtie[39], which compact the genome into a data structure that is very efficient when searching for perfect matches[42,43]. When allowing mismatches, the performance of Burrows-Wheeler transform methods decreases exponentially with the number of mismatches as they iteratively perform perfect searches[39–41].

Unspliced read aligners are ideal for mapping reads against a reference cDNA databases for quantification purposes[5,20,26,44,45]. If the exact reference transcriptome is available, Burrows-Wheeler methods are faster than seed-based methods (in our example, ~15× faster requiring ~110 central processing unit (CPU) hours) and have small differences in alignment specificity (~10% lower) **Supplementary Table 1**). In contrast, when only the reference transcriptome of a distant species is available, 'seed methods' can result in a large increase in sensitivity. For example, using the rat transcriptome as a reference for mouse reads resulted in 40% more reads aligned at a cost of ~7× more compute time, yielding a comparable alignment success rate as when aligning to the actual reference mouse transcriptome (**Supplementary Table 1** and **Supplementary Figs. 1** and **2**). Similarly, an increase in sensitivity using seed methods has been observed when aligning reads to polymorphic regions in a species for quantification of allele-specific gene expression[46].

Unspliced read aligners are limited to identifying known exons and junctions, and do not allow for the identification of splicing events involving new exons. Alternatively, reads can be aligned to the entire genome, including intron-spanning reads that require large gaps for proper placement. Several methods exist, collectively referred to as 'spliced aligners', that fall into two main categories: 'exon first' and 'seed and extend'. Exon-first[47–49] methods such as MapSplice[49], SpliceMap[47] and TopHat[48] use a two-step process. First, they map reads continuously to the genome using the unspliced read aligners (**Fig. 1a**). Second, unmapped reads are split into shorter segments and aligned independently. The genomic regions surrounding the mapped read segments are then searched for possible spliced connections. Exon-first aligners are very efficient when only a small portion of the reads require the more computationally intensive second step. Alternatively, seed-extend methods[8,50,51] such as 'genomic short-read nucleotide alignment program' (GSNAP)[50] and 'computing accurate spliced alignments' (QPALMA)[51] break reads into short seeds, which are placed onto the genome to localize the alignment (**Fig. 1b**). Candidate regions are then examined with more sensitive methods, such as the Smith-Waterman algorithm[51] or iterative extension and merging of initial seeds[8,50] to determine the exact spliced alignment for the read (**Fig. 1b**). Many of these alignment methods[47–51] also support paired-end read mapping, which increases alignment specificity.

Exon-first approaches are faster and require fewer computational resources compared to seed-extend methods. For example, a seed-extend method (GSNAP) takes ~8× longer (~340 CPU hours) than an exon-first method (TopHat) resulting in ~1.5× more spliced reads (**Supplementary Table 1**). However, the biological meaning of these additional splice junctions has not been demonstrated.

Exon-first approaches can miss spliced alignments for reads that also map to the genome contiguously, as can occur for genes that have retrotransposed pseudogenes (**Fig. 1c**). In contrast, seed-extend methods evaluate spliced and unspliced alignments in the same step, which reduces this bias toward unspliced alignments, yielding the best placement of each read. Seed-extend methods perform better than exon-first approaches when mapping reads from polymorphic species[52].

Transcriptome reconstruction

Defining a precise map of all transcripts and isoforms that are expressed in a particular sample requires the assembly of these reads or read alignments into transcription units. Collectively, we refer to this process as transcriptome reconstruction. Transcriptome reconstruction is a difficult computational task for three main reasons.

First, gene expression spans several orders of magnitude, with some genes represented by only a few reads. Second, reads originate from the mature mRNA (exons only) as well as from the incompletely spliced precursor RNA (containing intronic sequences), making it difficult to identify the mature transcripts. Third, reads are short, and genes can have many isoforms, making it challenging to determine which isoform produced each read.

Several methods exist to reconstruct the transcriptome, and they fall into two main classes: 'genome-guided' and 'genome-independent' (**Fig. 2**). Genome-guided methods rely on a reference genome to first map all the reads to the genome and then assemble overlapping reads into transcripts. By contrast, genome-independent methods assemble the reads directly into transcripts without using a reference genome.

Genome-guided reconstruction. Existing genome-guided methods can be classified in two main categories: 'exon identification' and 'genome-guided assembly'[28,29] approaches.

Exon identification[16,23] methods such as G.mor.se[16] were developed early when reads were short (~36 bases) and few aligned to exon-exon junctions. They first define putative exons as coverage islands, and then use spliced reads that span across these coverage islands to define exon boundaries and to establish connections between exons. Exon identification methods provided a first approach to solve the transcript reconstruction problem best suitable for short reads, but they are underpowered to identify full-length structures of lowly expressed, long and alternatively spliced genes.

To take advantage of longer read lengths, genome-guided assembly methods such as Cufflinks[29] and Scripture[28] have been developed. These methods use spliced reads directly to reconstruct the transcriptome[28,29]. Scripture initially transforms the genome into a graph topology, which represents all possible connections of bases in the transcriptome either when they occur consecutively or when they are connected by a spliced read. Scripture uses this graph topology to reduce the transcript reconstruction problem to a statistical segmentation problem of identifying significant transcript paths across the graph[28]. Scripture provides increased sensitivity to identify transcripts expressed at low levels by working with significant paths, rather than significant exons[28]. Cufflinks uses an approach originally developed for EST assembly[7], to connect aligned reads into a graph based on the location of their spliced alignments[29]. Scripture and Cufflinks build conceptually similar assembly graphs but differ in how they parse the graph into transcripts. Scripture reports all isoforms that are compatible with the read data (maximum sensitivity)[28], whereas Cufflinks reports the minimal number of compatible isoforms (maximum precision)[29]. Specifically, Scripture enumerates all possible paths through the assembly graph that are consistent with the spliced reads and the fragment size distribution of the paired end reads. In contrast, Cufflinks chooses a minimal set of paths through the graph such that all reads are included in at least one path. Each path defines an isoform, so this minimal set of paths is a minimal assembly of reads. As there can be many minimal sets of

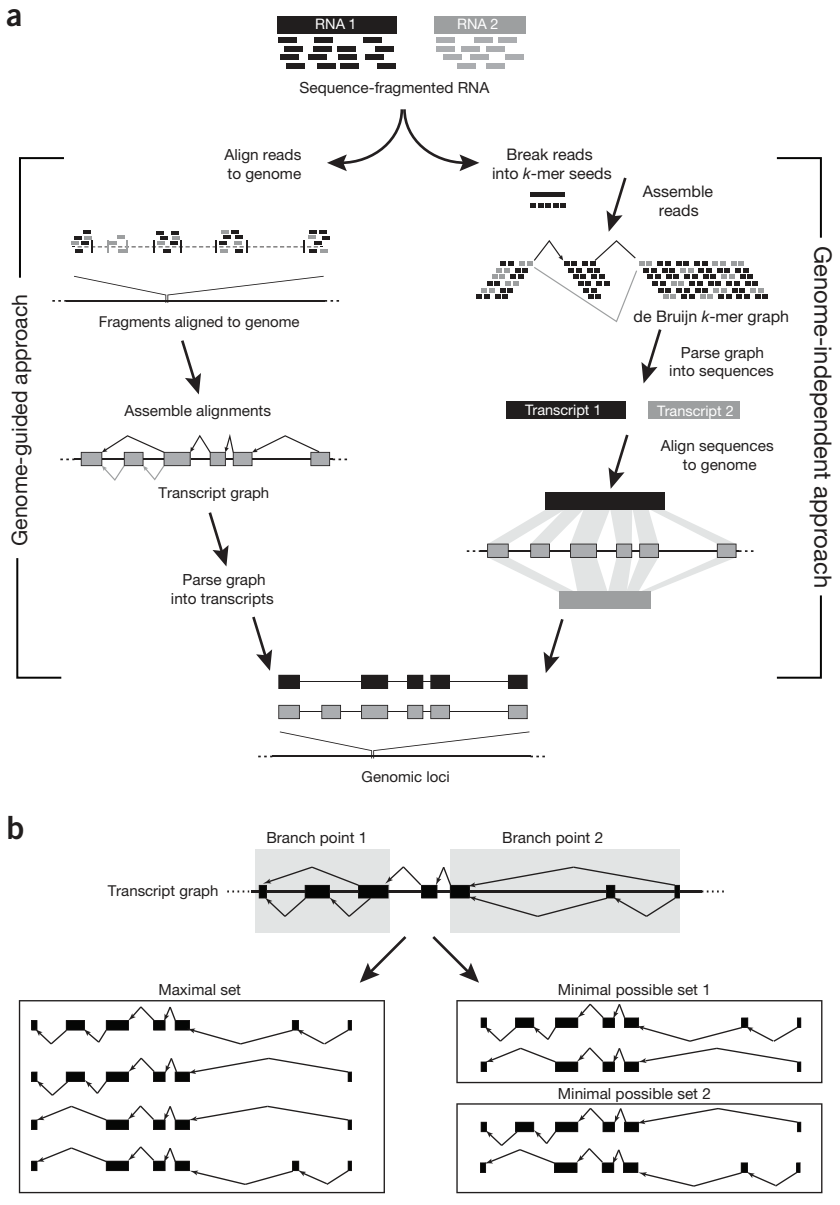

Figure 2 | Transcriptome reconstruction methods. (**a**) Reads originating from two different isoforms of the same genes are colored black and gray. In genome-guided assembly, reads are first mapped to a reference genome, and spliced reads are used to build a transcript graph, which is then parsed into gene annotations. In the genome-independent approach, reads are broken into *k*-mer seeds and arranged into a de Bruijn graph structure. The graph is parsed to identify transcript sequences, which are aligned to the genome to produce gene annotations. (**b**) Spliced reads give rise to four possible transcripts, but only two transcripts are needed to explain all reads; the two possible sets of minimal isoforms are depicted.

isoforms, Cufflinks uses read coverage across each path to decide which combination of paths is most likely to originate from the same RNA[29] (**Fig. 2b**).

Scripture and Cufflinks have similar computational requirements, and both can be run on a personal computer. Both assemble similar transcripts at the high expression levels but differ substantially for lower expressed transcripts where Cufflinks reports 3× more loci (70,000 versus 25,000) most of which do not pass the statistical significance threshold used by Scripture (**Supplementary Table 1** and **Supplementary Fig. 3**). In contrast, Scripture reports more isoforms per locus (average of 1.6 versus 1.2) with difference arising only for a handful of transcripts (**Supplementary Table 1**). In the most extreme case, Scripture reports over 300 isoforms for a single locus whereas Cufflinks reports 11 isoforms for the same gene.

Genome-independent reconstruction. Rather than mapping reads to a reference sequence first, genome-independent transcriptome reconstruction algorithms such as transAbyss[53] use the reads to directly build consensus transcripts[53–55]. Consensus transcripts can then be mapped to a genome or aligned to a gene or protein database for annotation purposes. The central challenge for genome-independent approaches is to partition reads into disjoint components, which represent all isoforms of a gene. A commonly used strategy is to first build a de Bruijn graph, which models overlapping subsequences, termed '*k*-mers' (*k* consecutive nucleotides), rather than reads[55–58]. This reduces the complexity associated with handling millions of reads to a fixed number of possible *k*-mers[57,58]. The overlaps of $k - 1$ bases between these *k*-mers constitute the graph of all possible sequences that can be constructed. Next, paths are traversed in the graph, guided by read and paired-end coverage levels, eliminating false branch points introduced by *k*-mers that are shared by different transcripts but not supported by reads and paired ends. Each remaining path through the graph is then reported as a separate transcript (**Fig. 2**).

Although genome-independent reconstruction is conceptually simple, there are two major complications: distinguishing sequencing errors from variation, and finding the optimal balance between sensitivity and graph complexity. Unlike the mapping-first strategy, sequencing errors introduce branch points in the graph that increase its complexity. To eliminate these artifacts, genome-independent methods look at the coverage of different paths in the graph and apply a coverage cutoff to decide when to follow a path or when to remove it[53,59]. In practice, the choice of the *k*-mer length for this analysis can greatly affect the assembly[53]. Smaller values of *k* result in a larger number of overlapping nodes and a more complex graph, whereas larger values of *k* reduce the number of overlaps and results in a simpler graph structure. An optimal choice of *k* depends on coverage: when coverage is low, small values of *k* are preferable because they increase the number of overlapping reads contributing *k*-mers to the graph. But when coverage is large, small values of *k* are overly sensitive to sequencing errors and other artifacts, yielding very complex graph structures[59].

To cope with the variability in transcript abundance intrinsic to expression data, several methods, such as transABySS, use a variable *k*-mer strategy to gain power across expression levels to assemble transcripts[53,55], albeit at the expense of CPU power and requiring parallel execution.

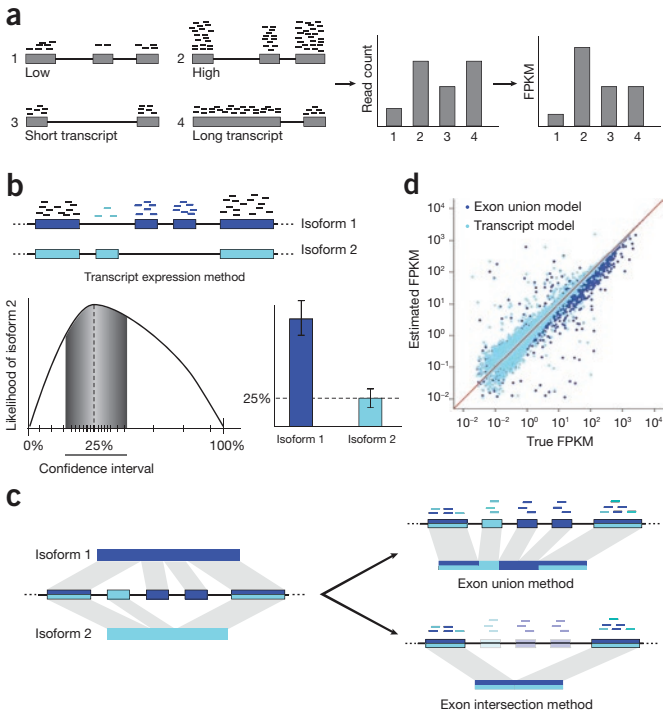

Figure 3 | An overview of gene expression quantification with RNA-seq. (**a**) Illustration of transcripts of different lengths with different read coverage levels (left) as well as total read counts observed for each transcript (middle) and FPKM-normalized read counts (right). (**b**) Reads from alternatively spliced genes may be attributable to a single isoform or more than one isoform. Reads are color-coded when their isoform of origin is clear. Black reads indicate reads with uncertain origin. 'Isoform expression methods' estimate isoform abundances that best explain the observed read counts under a generative model. Samples near the original maximum likelihood estimate (dashed line) improve the robustness of the estimate and provide a confidence interval around each isoform's abundance. (**c**) For a gene with two expressed isoforms, exons are colored according to the isoform of origin. Two simplified gene models used for quantification purposes, spliced transcripts from each model and their associated lengths, are shown to the right. The 'exon union model' (top) uses exons from all isoforms. The 'exon intersection model' (bottom) uses only exons common to all gene isoforms. (**d**) Comparison of true versus estimated FPKM values in simulated RNA-seq data. The $x = y$ line in red is included as a reference.

Reconstruction strategies compared. Both genome-guided and genome-independent algorithms have been reported to accurately reconstruct thousands of transcripts and many alternative splice forms[28,29,53,55]. The question as to which strategy is most suitable for the task at hand is strongly governed by the particular biological question to be answered. Genome-independent methods are the obvious choice for organisms without a reference sequence, whereas the increased sensitivity of genome-guided approaches makes them the obvious choice for annotating organisms with a reference genome. In the case of genomes or transcriptomes that have undergone major rearrangements, such as in cancer cells[26], the answer to the above question becomes less clear and depends on the analytical goal. In many cases, a hybrid approach incorporating both the genome-independent and genome-guided strategies might work best for capturing known information as well as capturing novel variation. In practice, genome-independent methods require considerable computational resources (~650 CPU hours and >16 gigabytes of random-access

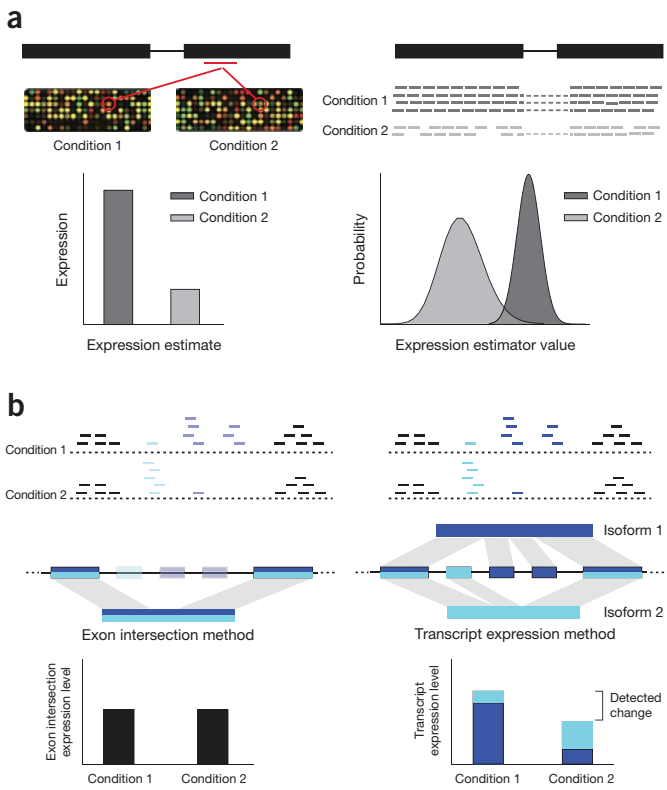

Figure 4 | Overview of RNA-seq differential expression analysis. (**a**) Expression microarrays rely on fluorescence intensity via a hybridization of a small number of probes to the gene RNA. RNA-seq gene expression is measured as the fraction of aligned reads that can be assigned to the gene. (**b**) A hypothetical gene with two isoforms undergoing an isoform switch between two conditions is shown. The total number of reads aligning to the gene in the two conditions is similar, but its distribution across isoforms changes. Differential expression using the simplified exon union or exon intersection methods reports no changes between conditions while estimating read counts and expression for the individual isoforms detects both differential expression at the gene and isoform level.

memory (RAM)) compared to genome-guided methods (~4 CPU hours and <4 gigabytes RAM; **Supplementary Table 1**).

Estimating transcript expression levels

Expression quantification has long been an important application. Over the past decade, DNA microarrays have been the technology of choice for high-throughput transcriptome profiling. When using RNA-seq to estimate gene expression, read counts need to be properly normalized to extract meaningful expression estimates[5,15,20–22,60–63]. There are two main sources of systematic variability that require normalization. First, RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample[19,64,65] (**Fig. 3a**). Second, the variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples[19,20] (**Fig. 3a**).

To account for these issues, the reads per kilobase of transcript per million mapped reads (RPKM) metric normalizes a transcript's read count by both its length and the total number of mapped reads in the sample[20] (**Fig. 3a**). When data originate from paired-end sequencing, the analogous fragments per kilobase of transcript per million mapped reads (FPKM) metric accounts for the dependency between paired-end reads in the RPKM estimate and as such is the metric of choice for both gene and isoform quantification[29].

As many genes have multiple isoforms, many of which share exons, and many genes families have close paralogs, some reads cannot be assigned unequivocally to a transcript (**Fig. 3b**). This 'read assignment uncertainty' affects expression quantification accuracy[29,66,67]. One strategy, used in the alternative expression analysis by RNA sequencing (Alexa-seq) method[44], is to estimate isoform-level expression values by counting only the reads that map uniquely to a single isoform. Although this works for some alternatively spliced genes, it fails for genes that do not contain unique exons from which to estimate isoform expression. Alternative methods termed 'isoform-expression methods' such as Cufflinks[29] and mixture of isoforms (MISO)[30], handle uncertainty by constructing a 'likelihood function' that models the sequencing process and identifies isoform abundance estimates that best explain the reads obtained in the experiment[29,30,53,66] (**Fig. 3b**). This estimate, defined as the isoform abundance that maximizes the likelihood function, is termed the maximum likelihood estimate (MLE). For genes expressed at low levels, the MLE is not an accurate expression estimate; Bayesian inference improves the robustness of expression quantification by 'sampling' alternative abundance estimates around the MLE while also providing a confidence measure on the estimate (**Fig. 3b**).

We note that the number of potential isoforms greatly impacts the results, with incorrect or misassembled isoforms introducing uncertainty. As such, when working with methods that produce the maximal isoform sets, it is necessary to prefilter transcripts before expression estimation for some genes. This applies to both genome-guided as well as genome-independent algorithms.

Often, the objective is to estimate expression per gene rather than for each isoform or transcript[19,20,63]. A gene's expression is defined as the sum of the expression of all of its isoforms. However, calculating isoform abundance can be computationally challenging especially for complex loci. Rather than computing isoform abundances, it is possible to define simplified schemes for quantifying gene expression. The two most commonly used counting schemes are (**Fig. 3c**): the 'exon intersection method'[68], which counts reads mapped to its constitutive exons, and the 'exon union method'[20,44], which counts all reads mapped to any exon in any of the gene's isoforms. The exon intersection method is analogous to expression microarrays, which typically probe expression signal in constitutive regions of each gene. Although convenient, these simplified models come at a cost; the exon union model underestimates expression for alternatively spliced genes[29,69] (**Fig. 3d**), and the intersection can reduce power for differential expression analysis, as discussed below.

Differential expression analysis with RNA-seq

Having quantified and normalized expression values, an important question is to understand how these expression levels differ across conditions. The last decade saw the development of extensive methodology for the statistical analysis of differential expression using microarrays[70–72] (**Fig. 4a**). Although in principle these approaches are directly applicable to RNA-seq data as well, using read coverage to quantify transcript abundance provides additional information such as a distribution for expression estimates in a single sample (**Fig. 4a**). Moreover, the power to detect differential

expression depends on the sequencing depth of the sample, the expression of the gene, and even the length of the gene[64,68].

To accommodate the count-based nature of RNA-seq data, initial methods modeled the observed reads using count-based distributions such as the Poisson distribution[19,29,66]. However, several studies have reported that these distributions do not account for biological variability across samples[73,74]. Ideally, if one had enough replicates the variability across replicates could be estimated empirically using a permutation-derived approach[70–72], similar to that used by the Myrna method[73]. However, to date few RNA-seq expression studies have generated a sufficient number of replicates to achieve this goal. To overcome this, many methods attempt to model biological variability and provide a measure of significance in the absence of a large number of biological replicates. These methods, such as EdgeR[75], differential expression analysis of count data (DESeq)[76] and recent versions of Cuffdiff[29], model the count variance across replicates as a nonlinear function of the mean counts using various different parametric approaches (such as the normal and negative binomial distributions)[19,29,65–67,76,77].

It is important to note although these approaches can assign significance to differential expression, the biological conclusions must be interpreted with care. For example, although the variability of the sequencing process is low compared to microarray hybridization[19], measurements can vary substantially because of differences in library construction protocols[78] and most importantly because of intrinsic variability in biological samples. As with any biological measurement, biological replicates provide the only measure of intrinsic, nontechnical transcript expression variability and thus are as critical as ever for differential expression analysis.

Implications of quantification strategies on differential gene expression analysis. When performing differential expression analysis, most methods take as input the normalized read count for each gene in each condition. Using simplified gene quantification models, such as the exon intersection method or the exon union method, can lead to unexpected conclusions. When a gene has multiple isoforms, a change in the gene's expression may not result in a corresponding change in raw gene-level counts. Consider the case of a gene with two isoforms (**Fig. 4b**), one substantially longer than the other. If the gene-level read counts are similar between conditions but distributed differently among the isoforms, differential expression results will differ depending on the counting method used. Differential expression analysis based on the isoform expression method will identify differential expression at both the isoform and gene levels (**Fig. 4b**). In contrast, no expression changes for this gene would be detected when using the exon union method and exon intersection method (**Fig. 4b**). Conversely, a differentially spliced gene could maintain a constant overall expression but generate a substantially different number of reads in two conditions. Indeed, in simulations we observed that, in the absence of differential splicing, the isofom expression and union methods performed similarly and were more sensitive that the exon intersection method. But for genes with multiple isoforms that were differentially spliced, the isoform expression method performed considerably better than the both exon union and exon intersection methods (94% versus 15% and 30% of genes detected, respectively) (**Supplementary Figs. 4** and **5**).

Conclusions and anticipated future developments

As sequencing technologies mature, existing computational tools will need to evolve to meet new requirements, and new tools will emerge to enable new applications. For example, as read length continues to increase, new mapping methods will need to efficiently align hundreds of millions of long reads—a daunting task. As longer reads often span multiple exon-exon junctions, transcript reconstruction and quantification methods would benefit by incorporating the more complete isoform information encoded in longer reads. Standard RNA-seq methods are not suited to annotate the 5′ start site and 3′ ends of transcripts by using specialized RNA-seq libraries[79–81] that identify the ends; transcriptome reconstruction methods will improve transcript annotation. Methods for estimating expression from RNA-seq data need to be improved to better handle the increasing availability of biological replicate experiments and would ideally model (and automatically subtract) systematic sources of bias that are introduced by laboratory methods (such as 3′-end biases). By providing the sequence of expressed transcripts, RNA-seq encodes information about allelic variation and RNA processing, so reconstruction methods should be adapted to account for this variability and report it. The ongoing cycle of improvements in technology, both in the laboratory as well as computationally, will continue to expand the possibilities of RNA-seq, making this technology applicable to an increasing variety of biological problems.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank L. Gaffney for help with figures; B. Haas for making available scripts to run transAbyss and for many discussions; Y. Katz, C. Nusbaum, A. Pauli and M. Zody for helpful discussions and comments on the manuscript; and J. Alfoldi, C. Burge, M. Cabili, K. Lindblad-Toh, J. Rinn, L. Pachter, S. Salzberg and O. Zuk for helpful comments on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Marra, M. *et al.* An encyclopedia of mouse genes. *Nat. Genet.* **21**, 191–194 (1999).
2. Carninci, P. *et al.* Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273–1289 (2003).
3. de Souza, S.J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **97**, 12690–12693 (2000).
4. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
5. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
6. Adams, M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
7. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
8. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
9. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
10. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
11. Pan, Q. *et al.* Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16**, 929–941 (2004).

12. Castle, J.C. *et al.* Expression of 24,426 human alternative splicing events and predicted *cis* regulation in 48 tissues and cell lines. *Nat. Genet.* **40**, 1416–1425 (2008).
13. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
14. Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
15. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
16. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
17. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
18. Maher, C.A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
19. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
First systematic comparison of expression arrays and RNA-seq revealed that technical variability between RNA-seq runs is extremely low; the authors developed the first methods for principled differential analysis of expression with read counts.
20. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
One of the first papers to describe the RNA-seq experimental protocol and provided the foundations for the computational analysis of quantitative transcriptome sequencing by introducing the RPKM expression metric.
21. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
22. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
23. Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 3264–3269 (2009).
24. Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).
25. Wilhelm, B.T. *et al.* RNA-seq analysis of two closely related leukemia clones that differ in their self-renewal capacity. *Blood* **117**, e27–e38 (2010).
26. Berger, M.F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).
27. Mortazavi, A. *et al.* Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res.* **20**, 1740–1747 (2010).
28. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
This paper describes a spliced alignment–based genome-guided transcript reconstruction methods that allow discovery of novel genes and isoforms from RNA-seq data.
29. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
This paper describes a spliced alignment–based genome-guided transcript reconstruction methods that allow discovery of novel genes and isoforms from RNA-seq data and provided a method for estimating the expression of each reconstructed isoform.
30. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
This paper describes a computational method that estimates isoform expression making use of both single and paired-end reads, and provides a Bayesian approach for detecting differential isoform expression.
31. Homer, N., Merriman, B. & Nelson, S.F. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4**, e7767 (2009).
32. Jiang, H. & Wong, W.H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).
A statistical algorithm to calculate isoform abundances for alternatively spliced genes is described.
33. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
34. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
35. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* advance online publication 27 October 2010 (doi:10.1101/gr.111120.110).
36. Rizk, G. & Lavenier, D. GASSST: global alignment short sequence search tool. *Bioinformatics* **26**, 2534–2540 (2010).
37. Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, e1000386 (2009).
38. Smith, A.D., Xuan, Z. & Zhang, M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**, 128 (2008).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
Introduced short read alignment with the Burrows-Wheeler transform, allowing the construction of the first fast alignment pipelines for RNA-seq.
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
42. Burrows, M. & Wheeler, D.J.A. Block-sorting lossless data compression algorithm. *Digital SRC Reports* 124, [AU: provide an article ID number or page numbers, or some other identifying information for this paper, such as a doi number or Pubmed or CrossRef ID] (1994).
43. Ferragina, P. & Manzini, G. An experimental study of a compressed index. *Inf. Sci.* **135**, 13–28 (2001).
44. Griffith, M. *et al.* Alternative expression analysis by RNA sequencing. *Nat. Methods* **7**, 843–847 (2010).
45. Cloonan, N. *et al.* RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics* **25**, 2615–2616 (2009).
46. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
47. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38**, 4570–4578 (2010).
48. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
This method combined fast read alignment using Burrows-Wheeler transform alignment with novel junction discovery, was one of the first scalable RNA-seq alignment programs, and paved the way for gene discovery and transcript reconstruction with RNA-seq.
49. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
50. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
51. De Bona, F., Ossowski, S., Schneeberger, K. & Ratsch, G. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, i174–i180 (2008).
52. Mikkelsen, T.S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
53. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
Described a variable *k*-mer approach for genome-independent reconstruction that allows for transcript discovery without a reference genome.
54. Birol, I. *et al.* De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872–2877 (2009).
55. Surget-Groba, Y. & Montoya-Burgos, J.I. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.* **20**, 1432–1440 (2010).
56. De Bruijn, N.G. A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* **46**, 6 (1946).
57. Pevzner, P.A. 1-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.* **7**, 63–73 (1989).
58. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
59. Zerbino, D.R. Using the Velvet *de novo* assembler for short-read sequencing technologies. *Curr. Protoc. Bioinformatics* **31**, 11.5.1–11.5.12 (2010).
60. Blencowe, B.J., Ahmad, S. & Lee, L.J. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.* **23**, 1379–1386 (2009).

61. Lister, R., Gregory, B.D. & Ecker, J.R. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.* **12**, 107–118 (2009).
62. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**, S22–S32 (2009).
63. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
64. Oshlack, A. & Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4**, 14 (2009).
65. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
66. Jiang, H. & Wong, W.H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**, 1026–1032 (2009).
67. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
68. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
69. Wang, X., Wu, Z. & Zhang, X. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J. Bioinform. Comput. Biol.* **8** (Suppl. 1), 177–192 (2010).
70. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
71. Grant, G.R., Manduchi, E. & Stoeckert, C.J. Jr. Analysis and management of microarray gene expression data. *Curr. Protoc. Mol. Biol.* **19** 6 (2007).
72. Grant, G.R., Liu, J. & Stoeckert, C.J. Jr. A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics* **21**, 2684–2690 (2005).
73. Langmead, B., Hansen, K.D. & Leek, J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* **11**, R83 (2010).
74. Robinson, M.D. & Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007). **Provided a statistical framework that is well suited to differential expression testing when a small number of RNA-seq replicates are available, and which also works well for larger experiments.**
75. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
76. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
77. Wang, L., Feng, Z., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2010).
78. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
79. Jan, C.H., Friedman, R.C., Ruby, J.G. & Bartel, D.P. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469**, 97–101 (2011).
80. Mangone, M. *et al.* The landscape of *C. elegans* 3′UTRs. *Science* **329**, 432–435 (2010).
81. Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* **7**, 528–534 (2010).
82. Lee, S. *et al.* Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.* **39**, e9 (2010).