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]

(1) Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142

(2) Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139

(3) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge MA 02138

[†] To whom correspondence should be addressed. mgarber@broadinstitute.org

Abstract

Recent studies have shown that high-throughput RNA sequencing (RNA-Seq) can be used to address biological problems that were challenging using previous techniques. RNA-Seq promises a comprehensive picture of the transcriptome allowing for the complete annotation and quantification of all genes and their isoforms in any given sample. However, realizing this promise requires increasingly complex computational methods. In this review we present the computational challenges and describe current methods to address them. We break these methods into three categories: (i) read alignment, (ii) transcriptome reconstruction, and (iii) expression quantification. For each category, we explain the major conceptual and practical challenges, as well as the general classes of solutions to these problems. We highlight the interdependence between the three different categories and discuss the benefits of different approaches for different biological applications. Since RNA-Seq technologies continue to improve, we point to further computational developments required to accommodate technological developments.

Introduction

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, followed by capillary sequencing[1-3]. Due to the high cost and limited data yield intrinsic to this approach, it only provided a glimpse into the true complexity of alternative splicing and cell-type specific transcription[4,5]. The higher-throughput Expressed Sequence Tag (EST) libraries required sophisticated computational tools, some of which[6-9], are the basis of many of the programs we review here. Alternative strategies, such as genome-wide tiling arrays, allowed for the identification of transcribed regions at a larger and more cost-efficient scale, but were unable to identify connected gene units or alternative splicing events that might exist[3,10]. Splicing arrays with probes across exon-exon junctions enabled researchers to analyze a predefined set of splicing events[11,12] but were of limited utility 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 revolutionized many genomic analyses. DNA sequencing can also be applied to generate short read data from total RNA, a process termed RNA-Seq[5,15-23]. We use the term RNA-Seq to refer to all experimental procedures coupled to sequencing methods that generate reads, of variable lengths, from the entire RNA molecule. In theory, RNA-Seq can be used to build a complete map of the transcriptome across all cell types and states. However, to fully realize this goal, RNA-Seq requires powerful experimental and 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 novel fusion genes in cancer[18,26], genome assembly improvements[27], and transcript identification[16,23,28,29]. Each of these applications required slightly different computational methodologies, but they all relied on a set of core approaches including alignment to a reference genome or transcriptome, assembly into transcriptional units, and quantification of expression.

In this review, we focus on the computational methods to address these three core computational challenges. First, we describe methods to align RNA-Seq reads directly to a reference transcriptome or genome (read mapping). Second, we discuss methods to identify and assemble expressed genes and isoforms either by using the genome-mapped reads, or directly from unmapped reads (transcriptome reconstruction). Third, we describe methods for abundance estimation of genes and their isoforms, as well as methods for the analysis of differential expression of these transcripts (expression quantification).

Several methods have been developed to address these computational challenges. Due to ongoing improvements in RNA-Seq data generation, there is great variability in the level of maturity of available computational tools. In some areas, such as read mapping, a wealth of algorithms has been developed, while in others, such as differential expression analysis, solutions are only beginning to emerge. Rather than providing a comprehensive description of each method, we highlight the key common principles, as well as the critical differences underlying the different methods and their application to RNA-Seq analysis. Furthermore, we discuss how these different methodologies can impact the results and the 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. Within each category we chose a

representative method, and applied it to a published RNA-Seq data set[28]. We report these results to provide a general indication of the compute resources required by the methods. To help guide the reader, we provide a list and brief description of currently available methods for each of the three categories in Table 1.

Even though experimental procedures involved in library construction for RNA-Seq are critical, they have already been described in previous works[30-32] and are not the focus of this review. The methods described here are largely independent of the choice of library construction protocols, with the notable exception of "paired-end" sequencing (reading RNA from both ends of a fragment), that provide valuable information in all stages of RNA-Seq analysis[28,29,33].

Mapping short RNA-Seq reads

One of the most basic tasks in RNA-Seq analysis is the alignment of reads to either a reference genome or known transcriptome. Finding regions of similarity between sequences is a classic problem, and several solutions exists for EST mapping[8,9], however RNA-Seq reads pose particular challenges because they are short (~36-125 bases), error rates can be considerable, and reads can span exon-exon junctions. In addition, the number of reads per experiment is increasingly larger, currently on the order of hundreds of millions. A number of alignment programs exist to address the challenge of mapping short reads to a known transcriptome which we collectively refer to as **unspliced read aligners** because they align reads to reference without allowing any large gaps between exons. The **unspliced read aligners** fall into two main categories, **Seed methods** and **Burrows-Wheeler transform**[34] methods. **Seed methods**[35-42] such us MAQ[37] and Stampy[39] find matches for short sub-sequences, termed 'seeds', assuming that at least one seed within a read will *perfectly* match the reference. Candidate alignments for each read are then computed and ranked and the best scoring alignment is then reported. In contrast, **Burrows-Wheeler transform methods**[43-45] like BWA[44] and Bowtie[43], compact the genome into a search-efficient data structure[46], which allows for faster alignment, as long as the number of mismatches is small [43-45] (less than 3).

Seed and BW mappers are not optimal for direct alignment of RNA-Seq reads to the genome since many reads that originate from exon-exon junctions map discontinuously to the genome. However, short read aligners are ideal for mapping reads against a known cDNA databases for quantification purposes[5,20,26,47,48]. If the exact reference transcriptome is available, BW methods are more commonly used as they are significantly faster than seed based methods (~15X faster when aligning 58 million paired-end reads) and result in minimal difference in

alignment specificity (~10% lower). In contrast, if the exact reference is not available requiring alignment to a more divergent reference, such as mouse reads on a rat reference, seed methods, while ~7X slower, result in a 40% gain in sensitivity and yield a comparable alignment success rate as when aligning to the actual reference transcriptome. Similarly, an increase in sensitivity using seed methods has been observed when aligning reads to polymorphic regions within a species for quantification of allele-specific gene expression [49].

Unspliced read aligners are limited to identifying known exons and junctions, and do not allow for the identification of splicing events involving novel exons. Alternatively, reads can be aligned to the entire genome, including intron-spanning reads that require large gaps for proper placement. Several methods, collectively referred to as **spliced aligners**, have been proposed and can be grouped into two main categories, **exon-first** and **seed-and-extend**. **Exon-first**[50-52] methods like MapSplice[52], SpliceMap[50] and TopHat[51] employ a two-step process. First, they map reads to the genome, disallowing large gaps (figure 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**[8,53,54] like GSNAP [53] and QPALMA[54] break reads into short seeds, which are placed onto the genome to localize the alignment (Figure 1b). Candidate regions are then examined with more sensitive methods, such as the Smith-Waterman algorithm[54], or iterative extension and merging of initial seeds[8,53] to determine the exact spliced alignment for the read (figure 1b).

Exon-first approaches are faster and require fewer computational resources compared to **seed-extend** methods. For example, a **seed-extend** method (GSNAP) takes ~8X longer (~340

CPU hours) than an **exon-first** method (TopHat) resulting in ~1.5X more spliced reads. However, as the ground truth is not known it is unclear whether these additional splice junctions are biologically meaningful.

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 (Figure 1c). In contrast, seed-extend methods will align equally well to gapped and ungapped locations, yielding the best placement of each read. **Seed-extend** methods perform better than exon-first approaches when mapping reads from polymorphic species[55]. Many of these alignment methods[50-54] also support paired-end read mapping which can increase the specificity of the alignment and these paired-end alignments are used by many programs that process RNA-Seq data for both reconstruction and quantification.

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, genes can exist at many different expression levels, spanning several orders of magnitude, and some genes may be represented by only a few reads. Second, reads may 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 have been proposed that reconstruct the transcriptome, and they fall into two main classes: genome-guided and genome-independent (Figure 2). **Genome-guided** methods rely on a reference genome to first map all the reads to the genome, before proceeding to assemble overlapping reads into transcripts. By contrast, **genome-independent** methods assemble the reads directly into potential 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 Gmorse[16] were developed early when reads were short (~36 bases) and few aligned to exon-exon junctions. The strategy is to first define putative exons as coverage islands, 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 that are hard to map when spanning exon-exon junctions, however 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] were 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 utilizes an approach originally developed for EST assembly[7], by connecting fragments into a graph if the overlapping fragments agree on their spliced alignment locations[29]. Both Scripture and Cufflinks build conceptually similar assembly graphs but differ in how they parse this graph into transcripts. Scripture reports *all* isoforms that are compatible with the read data (maximum sensitivity)[28], while 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 fragment size distribution of the paired end reads. In contrast, Cufflinks finds the set of all *incompatible assemblies of reads*, splice sites do not overlap. These sets of incompatible assemblies represent the minimum possible number of reconstructed isoforms. Since this minimal transcript set is not guaranteed to be unique, coverage level compatibility across the graph is used to decide between minimal sets of transcripts[29].

Scripture and Cufflinks have similar computational requirements and both can be run on a personal computer. Both assemble nearly identical transcripts at the highest expression levels but differ significantly for lower expressed transcripts where Cufflinks reports 3X more loci (70,000 vs 25,000) most of which do not pass the statistical significance threshold used by Scripture. In contrast, Scripture reports more isoforms per loci (with an average of 1.6 versus 1.2) that differ mostly for a handful of transcripts. In the most extreme case, Scripture reports over 300 isoforms for a single locus while 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 Trans-Abyss[56] use the reads to directly build consensus transcripts[56-58]. 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[58-61]. This reduces the complexity associated with handling millions of reads to a fixed number of possible *k*-mers[60,61]. The overlaps of *k*-1 bases between these *k*-mers constitute a graph that encodes all possible sequences that can be constructed. Next, paths are traversed within 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 (figure 2).

While conceptually simple, there are two major complications in genome-independent reconstruction: distinguishing sequencing errors from variation, and finding the optimal balance between sensitivity and graph complexity. Unlike the mapping-first strategy, sequencing errors introduce a number of "bubbles" –short deviating paths 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[56,62]. In practice, the choice of the *k*-mer length for this analysis can greatly affect the assembly[56]. 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 levels: when coverage is low, small values of *k* are preferable because they increase the number of overlapping reads contributing *k*-

mers to the graph. However, when coverage is large, small values of *k* are overly sensitive to sequencing errors and other artifacts yielding very complex graph structures[62].

To cope with the variability in transcript abundance intrinsic to expression data, several methods, such as Trans-ABySS, uses a variable *k*-mer strategy to gain power across expression levels to assemble transcripts[56,58], albeit at the expense of CPU power and requiring parallel execution.

Reconstruction strategies compared. Both genome-guided and genome-independent algorithms have been reported to accurately reconstruct thousands of transcripts and many alternative splice forms[28,29,56,58]. 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 sensitivity of genome-guided approaches makes these algorithms preferable for annotating high-quality genomes and expanding the catalog of expressed transcripts. In the case of genomes that are of lesser quality or transcriptomes that underwent 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 significant computational resources (~650 CPU hours and >16Gb RAM) compared to genome-guided methods (~4 CPU hours and <4Gb RAM).

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. RNA-Seq is increasingly used to quantify gene expression by measuring the number of reads originating from a gene[5,15,20-22,32,63-65]. The challenge, however, is that read-counts need to be properly normalized to extract meaningful expression estimates. 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[19,66,67] (Figure 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] (figure 3a).

To account for these issues, the RPKM (Reads Per Kilobase of transcript per Million mapped reads) metric normalizes a transcript's read count by both its length and the total number of mapped reads in the sample[20] (Figure 3a). When data originates from paired-end sequencing the analogous FPKM metric (Fragments Per Kilobase of transcript per Million mapped reads) accounts for the dependency between paired-ends reads in the RPKM estimate and as such is a more natural metric for both gene and isoform quantification[29].

Since many genes have multiple isoforms, many of which share exons, and many genes families have close paralogs, reads cannot be assigned unequivocally to a transcript (Figure 3c). This *read assignment uncertainty* affects expression quantification accuracy[29,68,69]. One strategy, used by Alexa-Seq[47], to estimate isoform level expression values works by counting only the reads that map uniquely to a single isoform. While 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, such as Cufflinks[29] and MISO[33], 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[29,33,56,68] (Figure 3c). This estimate is defined as the abundance that maximizes the likelihood function and is termed the maximum likelihood estimate (MLE). Genes expressed at low levels may not generate enough sequencing data for robust quantification using only the MLE so that a Bayesian inference model can improve the robustness of expression quantification by "sampling" alternative abundance estimates around the MLE (figure 3c). This estimate is generally more accurate and provides a measure of confidence that can be used during differential expression analysis. Collectively, we term this quantification approach the *transcript expression method*.

We note that the number of potential isoforms greatly impacts robustness, and incorrect or misassembled isoforms can introduce further uncertainty. In the transcriptome reconstruction section we discussed methods that produce the minimal versus maximal possible isoform sets: for some genes, the maximal set can be very large, resulting in low-confidence isoform abundance estimates. It is therefore critical to pre-filter transcripts prior to any expression estimation when working with reconstruction methods that report maximal isoform sets. This applies to both genome-guided as well as genome-independent algorithms.

Often, the objective is to estimate expression on a *per-gene* level rather than for each isoform or transcript[19,20,32]. A gene's expression is defined as the sum of the expression of all of its isoforms. However, calculating isoform abundances 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 (see figure 3b): The *exon intersection method*[70] which counts reads mapped to its constitutive exons, and the *exon union method*[20,47] 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. While convenient, these simplified models come at a cost; the union model underestimates expression for alternatively spliced genes[29,71] (Figure 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[72-74] (Figure 4). While 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 within a single sample (Figure 4a). Moreover, the power to detect differential expression depends on the sequencing depth of the sample, the expression level of the gene, and even the length of the gene[66,70].

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[19,29,68]. However, several studies have reported that these distributions, exemplified by Poisson, do not account for biological variability across samples[75,76]. Ideally, if one had enough replicates the variability across replicates could be estimated empirically using a permutation-derived approach, similar to that used by the Myrna method[75]. 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[77], DESeq[78], and recent versions of Cuffdiff[29], model the count variance across replicates as a nonlinear function of the mean counts and use various different parametric approaches (such as the Normal and Negative Binomial distributions)[19,29,67-69,78,79].

It is important to note while these approaches can assign significance to differential expression, the biological conclusions must be interpreted with care. For example, while the variability of the sequencing process is low compared to microarray hybridization[19], measurements can vary substantially due to differences in library construction protocols[30] and most importantly due to intrinsic variability within biological samples. As with any biological measurement, biological replicates provide the only measure of intrinsic, non-technical 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. Therefore, the choice of quantification method is critical for differential expression analysis. 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 (Figure 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. Using the *transcript expression method* differential

expression analysis will identify differential isoform expression as well as differential gene expression (Figure 4b). In contrast, the *exon union method* and *exon intersection method* will not detect any expression changes for this gene (Figure 4b). Conversely, a differentially spliced gene could maintain a constant overall expression but generate a significantly different number of reads in two conditions. Indeed in simulations we observed that for genes with multiple isoforms, that were differentially spliced, the *transcript expression method* performed slightly better than the *exon union method* (86% versus 79% of genes detected) both of which performed significantly better than the *exon intersection method* (53% of genes detected).

Conclusions and anticipated future developments

In this review, we provided a description of the computational challenges involved in various aspects of RNA-Seq analysis, along with a survey of current approaches. 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, further improvements or novel methods will be needed to map and take advantage of reads that span multiple exon-exon junctions, so that longer reads will improve isoform reconstruction by providing more complete information about which isoform is expressed and which reads originate from each isoform. Current transcript reconstruction methods are not suited to annotate the 5'-start site and 3'-ends of transcripts, thus specialized RNA-Seq libraries to identify transcript ends[80-82] are increasingly being developed and utilized. Transcriptome reconstruction methods will improve reconstructions gene by handling and modeling these data appropriately. Methods for estimating expression levels from RNA-Seq need to further improve in order to better handle the increasing availability of biologically replicated 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, 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 a variety of additional biological problems.

Acknowledgments

We would like to thank Leslie Gaffney for help with figures, Brian Haas, Yarden Katz, Andrea Pauli, and Mike Zody for helpful discussions and comments on the manuscript and to Jessica Alfoldi, Chris Burge, Moran Cabili, Kerstin Lindblad-Toh, Chad Nusbaum, John Rinn, Lior Pachter, Steven Salzberg, and Or Zuk for helpful comments on the manuscript.

Figure Legends

Figure 1. Strategies for gapped alignments of RNA-Seq reads to the genome.

An illustration of reads obtained from a two-exon transcript; black and gray indicate exonic origin of reads. b) exon-first methods first map full, unspliced reads (exonic-reads). Remaining reads are divided into smaller pieces and mapped against the genome. An extension process extends mapped pieces to find candidate splice sites to support a spliced alignment. c) Seed-and-extend methods 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. d) Potential disadvantage of exon-first approaches: An illustration of a gene and its associated retrotransposed pseudogene. Red indicates mismatches compared to the gene sequence. Exonic reads will map to both the gene and its pseudogene preferring gene placement due to lack of mutations, however a spliced read could be incorrectly assigned to the pseudogene as it appears to be "exonic", preventing higher scoring spliced alignments from being pursued.

Figure 2. Transcriptome reconstruction methods.

Reads originating from two different isoforms of the same genes are colored in black and gray. Left arrow presents the 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. The right arrow points to the genome independent approach. Reads are broken into *k*-mers seeds and arranged into a de Brujin graph structure. The graph is parsed to identify transcript sequences, which are aligned to the genome to produce gene annotations.

Figure 3. An overview of gene expression quantification with RNA-Seq.

(a) Reads are mapped to four transcripts with different lengths and expression levels. The center panel shows the total read counts observed for each transcript. The right panel shows the read counts normalized by the transcript length using the FPKM metric. Although transcripts 2 and 4 have comparable read-counts, transcript 2 has a significantly higher normalized expression level, which is reflected in the normalized score. (b) Depicts a gene with two expressed isoforms. Exons are colored according to the isoform of origin, black for isoform 1 only, gray for isoform 2 only, and mixed black and gray for exons common to both isoforms. Depicted below are two simplified gene models used for quantification purposes. The *exon union model* uses exons from all isoforms. Spliced transcripts from each model, and their associated lengths, are shown below each model. The *exon intersection model* uses only exons common to all gene isoforms. (c) Reads from alternatively spliced genes may be attributable to a single isoform or more than one isoform, and are color-coded to indicate their isoform of origin. Dotted reads indicate read mapping uncertainty. To estimate the counts for each isoform, current methods estimate isoform abundances that best explain the observed read counts under a generative model. Abundance levels are sampled near the original maximum likelihood estimate (dashed line) to improve the robustness of the estimate and provide a confidence interval around each isoform's abundance. (d) Comparison of true versus estimated FPKM values in simulated RNA-Seq data. Gray dots represent FPKM estimated using the deconvolution approach. Black dots represent FPKM values estimated using the gene union model.

Figure 4. An overview of RNA-Seq differential expression analysis.

a) Comparison of microarray and RNA-Seq data. Expression microarrays measure expression by fluorescence intensity via a small number of probes per gene. RNA-Seq measures gene expression as the fraction of reads from the library that originated from the gene. b) The method used to count and normalize reads impacts power to detect differential gene expression. The figure depicts a hypothetical gene with two isoforms undergoing an isoform switch between two conditions. The total number of reads between 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 will report no changes between conditions, because the gene-level count remains nearly the same. Estimating read counts and expression for the individual isoforms will detect both differential expression at the gene and isoform level.

Table 1. Selected Software Packages

Package	Notes	Input	Category	Uses
Read Mapping				
BFAST[35]	A sensitive, seed method that can align long reads (e.g. 454) as well as short reads allowing for gaps. It does not incorporate read quality but handles paired end reads.	Reference Genome or reference transcriptome Sequence reads	Unspliced seed aligners	Expression quantification of a known set of transcripts with reads from low quality data or of species with high mutation rates that may contain indels at high frequency SNP discovery in expressed transcripts Long reads (e.g. from Roche 454)
GASSST[40]	Handles short indels. It does not incorporate read base quality (supports Fasta input only).			
RMAP[42]	Incorporates read quality scores in the alignment process.			
SeqMap[36]	Handles short indels. It does not utilize sequence base quality.			
SHRiMP[41]	Combines user configurable spaced seeding with a sensitive Smith-Waterman extension step. Can be memory intensive. It does not utilize sequence base quality.			
Stampy[39]	Supports a wide range of input			

	formats. Handles base and mapping quality information.			
Bowtie[43]	Burrows-Wheeler transform based aligners. BWA has very good handling of base quality information while the others do not utilize base quality information.	Burrows-Wheeler transformed reference genome or transcriptome Sequence reads	BWT short read aligners	Alignment to an existing transcriptome for quantification. Some spliced-read aligners build on top of these fast aligners to map unspliced reads or read segments.
BWA[44]				
SOAP2[45]				
MapSplice[52]	Works with multiple short read aligners and can be configured to run unspliced and spliced alignments simultaneously at the expense of execution time.	Burrows-Wheeler transformed reference genome Sequence reads	Exon first spliced aligners	The sensitivity, speed and relatively low resources required by these aligners makes them ideal for genome guided transcript reconstruction and expression quantification methods.
SpliceMap[50]	Works with multiple short read aligners to map short (25bp) read segments, which are extended across large gaps.			
TopHat[51]	Tightly integrated with Bowtie to first map exonic reads and then mapping reads across gaps.			
GSNAP[53]	Aligns reads with high substitution and indel rates and can use SNP databases to increase alignment	Reference Genome	Seed-extend	Transcript reconstruction or isoform quantification with Long reads (e.g. from Roche 454) or

	sensitivity.	Sequece Reads	spliced aligners	low quality reads.
QPALMA[54]	Follows an unspliced read mapping step with a very sensitive extension of Smith-Watterman to handle large gaps (introns). Handles base and mapping quality information.			
X-MATE[48]	Optimized for SOLiD color-space data it includes upfront quality filtering of reads, and alignments.	Reads and quality scores in SOLiD color space	Alignment to Splice junction databases.	RNA-Seq analysis of SOLiD data.
Transcript reconstruction				
Scripture[28]	Defines all possible transcripts supported by the aligned reads. It also provides a score and confidence for each reconstruction.	Aligned reads to a reference genome	Genome-guided transcript reconstruction methods	Transcript reconstruction, expression quantification. Ideal to find novel, previously unannotated transcripts.
Cufflinks[29]	Finds the minimal set of transcripts that are supported by previously aligned spliced reads.			Transcript reconstruction, gene/transcript expression quantification and differential analysis of expression alternative splicing.
ALLPATHS[83]	Preserves ambiguities from allelic variation and unresolved repeats.	Sequence reads	Whole-genome assemblers	Reconstructing genomes from short-read data at high coverage. These methods can be used in principle for transcriptomes but they are not optimized to do so.
Velvet[61] (OASES)	Can be parameterized to accommodate transcript reconstruction and has been			

	previously used for this purpose.			
Trans-ABySS[56]	Variable *k*-mer approach. Recovers alternatively spliced isoforms and allelic variants. Execution is highly parallel.	Sequence reads	Transcriptome assembler	Reconstructing transcripts including alternative isoforms. This method is optimized to reconstruct a transcriptome without using a known genome reference.
Transcript quantification				
Alexa-Seq[47]	Quantifies alternative splicing events.	Reference or reconstructed transcriptome Sequence reads	Gene quantification	Find differentially included exons.
ERANGE[20]	A set of RNASeq analysis scripts for expression quantification of known set transcripts.			Quantify gene expression.
NEUMA[84]	Normalizes expression using uniquely mappable reads.			Quantify transcript and gene expression.
Cufflinks[29]	Uses maximum likelihood and posterior distribution sampling to estimate relative isoform expression.	Reference or reconstructed transcriptome Aligned reads	Transcript quantification	Quantify transcript expression, differential isoform expression and alternative splicing analysis.
MISO[33]	Uses maximum likelihood and posterior distribution sampling to estimate relative isoform expression.			
RSEM[69]	Maximum likelihood based method to isoform expression that can			

	incorporate non-uniformity of coverage in the model.			
Cuffdiff[29]	Uses the Cufflinks quantification model to identify differentially expressed or spliced genes and transcripts.	Reference or reconstructed transcriptome Aligned reads	Differential Expression	Find differentially expressed or spliced genes
DegSeq[79]	Assumes a normal distribution of read counts and estimates variance and mean from replicates.	Reference or reconstructed transcriptome Aligned reads		Find differentially expressed genes
EdgeR[77]	General package for differential expression with digital data.	Pre-computed fragment counts for each gene and sample		
DESeq[78]	General package for differential expression with digital data.			
Myrna[75]	Cloud-based permutation based analysis, scales to many replicates	Sequence reads		

References

Reference descriptions:

1) **Mortazavi et al.** – This paper was one of the first 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.

2) **Marioni et al** – Did the first systematic comparison between expression arrays and RNA-Seq. They reported that the technical variability between RNA-Seq runs is extremely low and developed the first methods for principled differential analysis of expression with read counts.

3) **Langmead et al/Li et al** – Introduced short read alignment with the Burrows-Wheeler transform, allowing the construction of the first fast alignment pipelines for RNA-Seq.

4) **Trapnell et al. 2009** – Combined fast read alignment using Burrows-Wheeler transform alignment with novel junction discovery. This method was one of the first scalable RNA-Seq alignment programs and paved the way for gene discovery and transcript reconstruction with RNA-Seq.

5) **Jiang and Wong** – Described a statistical algorithm to calculate isoform abundances for alternatively spliced genes.

6) **Guttman et al. 2010/Trapnell et al. 2010** – Described two genome-guided transcript reconstruction methods that allow discovery of novel genes and isoforms. Trapnell et al. also provided a method for estimating the expression of each reconstructed isoform.

7) **Robertson et al.** – Described a variable *k*-mer approach for genome-independent reconstruction that allows for transcript discovery without a reference genome.

8) **Katz et al. 2010** – Provided 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

9) **Robinson and Smyth 2010** – 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.

1. M. Marra, L. Hillier, T. Kucaba et al., *Nat Genet* **21** (2), 191 (1999).
2. P. Carninci, K. Waki, T. Shiraki et al., *Genome Res* **13** (6B), 1273 (2003).
3. S. J. de Souza, A. A. Camargo, M. R. Briones et al., *Proc Natl Acad Sci U S A* **97** (23), 12690 (2000).
4. M. Guttman, I. Amit, M. Garber et al., *Nature* **458** (7235), 223 (2009).
5. E. T. Wang, R. Sandberg, S. Luo et al., *Nature* **456** (7221), 470 (2008).
6. M. D. Adams, J. M. Kelley, J. D. Gocayne et al., *Science* **252** (5013), 1651 (1991).
7. B. J. Haas, A. L. Delcher, S. M. Mount et al., *Nucleic Acids Res* **31** (19), 5654 (2003).
8. W. J. Kent, *Genome Res* **12** (4), 656 (2002).
9. T. D. Wu and C. K. Watanabe, *Bioinformatics* **21** (9), 1859 (2005).
10. P. Kapranov, S. E. Cawley, J. Drenkow et al., *Science* **296** (5569), 916 (2002).
11. Q. Pan, O. Shai, C. Misquitta et al., *Mol Cell* **16** (6), 929 (2004).
12. J. C. Castle, C. Zhang, J. K. Shah et al., *Nat Genet* **40** (12), 1416 (2008).
13. M. Schena, D. Shalon, R. W. Davis et al., *Science* **270** (5235), 467 (1995).
14. T. R. Golub, D. K. Slonim, P. Tamayo et al., *Science* **286** (5439), 531 (1999).
15. N. Cloonan, A. R. Forrest, G. Kolle et al., *Nat Methods* **5** (7), 613 (2008).
16. F. Denoeud, J. M. Aury, C. Da Silva et al., *Genome Biol* **9** (12), R175 (2008).
17. R. Lister, R. C. O'Malley, J. Tonti-Filippini et al., *Cell* **133** (3), 523 (2008).
18. C. A. Maher, C. Kumar-Sinha, X. Cao et al., *Nature* **458** (7234), 97 (2009).
19. J. C. Marioni, C. E. Mason, S. M. Mane et al., *Genome Res* **18** (9), 1509 (2008).
20. A. Mortazavi, B. A. Williams, K. McCue et al., *Nat Methods* **5** (7), 621 (2008).
21. U. Nagalakshmi, Z. Wang, K. Waern et al., *Science* **320** (5881), 1344 (2008).
22. M. Sultan, M. H. Schulz, H. Richard et al., *Science* **321** (5891), 956 (2008).
23. M. Yassour, T. Kaplan, H. B. Fraser et al., *Proc Natl Acad Sci U S A* **106** (9), 3264 (2009).

24. R. Blekhman, J. C. Marioni, P. Zumbo et al., *Genome Res* **20** (2), 180 (2009).
25. B. T. Wilhelm, M. Briau, P. Austin et al., *Blood* (2010).
26. M. F. Berger, J. Z. Levin, K. Vijayendran et al., *Genome Res* **20** (4), 413 (2010).
27. A. Mortazavi, E. M. Schwarz, B. Williams et al., *Genome Res* **20** (12), 1740 (2010).
28. M. Guttman, M. Garber, J. Z. Levin et al., *Nat Biotechnol* **28** (5), 503 (2010).
29. C. Trapnell, B. A. Williams, G. Pertea et al., *Nat Biotechnol* **28** (5), 511 (2010).
30. J. Z. Levin, M. Yassour, X. Adiconis et al., *Nat Methods* **7** (9), 709 (2010).
31. U. Nagalakshmi, K. Waern, and M. Snyder, *Curr Protoc Mol Biol* **Chapter 4**, Unit 4 11 1 (2010).
32. Z. Wang, M. Gerstein, and M. Snyder, *Nat Rev Genet* **10** (1), 57 (2009).
33. Y. Katz, E. T. Wang, E. M. Airoldi et al., *Nat Methods* **7** (12), 1009 (2010).
34. M. burrows and D. J. Wheeler, 1994.
35. N. Homer, B. Merriman, and S. F. Nelson, *PLoS One* **4** (11), e7767 (2009).
36. H. Jiang and W. H. Wong, *Bioinformatics* **24** (20), 2395 (2008).
37. H. Li, J. Ruan, and R. Durbin, *Genome Res* **18** (11), 1851 (2008).
38. R. Li, Y. Li, K. Kristiansen et al., *Bioinformatics* **24** (5), 713 (2008).
39. G. Lunter and M. Goodson, *Genome Res* (2010).
40. G. Rizk and D. Lavenier, *Bioinformatics* **26** (20), 2534 (2010).
41. S. M. Rumble, P. Lacroute, A. V. Dalca et al., *PLoS Comput Biol* **5** (5), e1000386 (2009).
42. A. D. Smith, Z. Xuan, and M. Q. Zhang, *BMC Bioinformatics* **9**, 128 (2008).
43. B. Langmead, C. Trapnell, M. Pop et al., *Genome Biol* **10** (3), R25 (2009).
44. H. Li and R. Durbin, *Bioinformatics* **25** (14), 1754 (2009).
45. R. Li, C. Yu, Y. Li et al., *Bioinformatics* **25** (15), 1966 (2009).
46. Paolo Ferragina and Giovanni Manzini, *Information Sciences* **135** (1-2), 13 (2001).
47. M. Griffith, O. L. Griffith, J. Mwenifumbo et al., *Nat Methods* **7** (10), 843 (2010).
48. N. Cloonan, Q. Xu, G. J. Faulkner et al., *Bioinformatics* **25** (19), 2615 (2009).
49. J. F. Degner, J. C. Marioni, A. A. Pai et al., *Bioinformatics* **25** (24), 3207 (2009).
50. K. F. Au, H. Jiang, L. Lin et al., *Nucleic Acids Res* **38** (14), 4570 (2010).
51. C. Trapnell, L. Pachter, and S. L. Salzberg, *Bioinformatics* **25** (9), 1105 (2009).
52. K. Wang, D. Singh, Z. Zeng et al., *Nucleic Acids Res* **38** (18), e178 (2010).
53. T. D. Wu and S. Nacu, *Bioinformatics* **26** (7), 873 (2010).
54. F. De Bona, S. Ossowski, K. Schneeberger et al., *Bioinformatics* **24** (16), i174 (2008).
55. T. S. Mikkelsen, M. J. Wakefield, B. Aken et al., *Nature* **447** (7141), 167 (2007).
56. G. Robertson, J. Schein, R. Chiu et al., *Nat Methods* **7** (11), 909 (2010).
57. I. Birol, S. D. Jackman, C. B. Nielsen et al., *Bioinformatics* **25** (21), 2872 (2009).
58. Y. Surget-Groba and J. I. Montoya-Burgos, *Genome Res* **20** (10), 1432 (2010).
59. N. G. De Bruijn, *Koninklijke Nederlandse Akademie v. Wetenschappen* **46**, 6 (1946).
60. P. A. Pevzner, *J Biomol Struct Dyn* **7** (1), 63 (1989).
61. D. R. Zerbino and E. Birney, *Genome Res* **18** (5), 821 (2008).
62. D. R. Zerbino, *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 5.
63. B. J. Blencowe, S. Ahmad, and L. J. Lee, *Genes Dev* **23** (12), 1379 (2009).
64. R. Lister, B. D. Gregory, and J. R. Ecker, *Curr Opin Plant Biol* **12** (2), 107 (2009).
65. S. Pepke, B. Wold, and A. Mortazavi, *Nat Methods* **6** (11 Suppl), S22 (2009).
66. A. Oshlack and M. J. Wakefield, *Biol Direct* **4**, 14 (2009).
67. M. D. Robinson and A. Oshlack, *Genome Biol* **11** (3), R25 (2010).
68. H. Jiang and W. H. Wong, *Bioinformatics* **25** (8), 1026 (2009).
69. B. Li, V. Ruotti, R. M. Stewart et al., *Bioinformatics* **26** (4), 493 (2010).
70. J. H. Bullard, E. Purdom, K. D. Hansen et al., *BMC Bioinformatics* **11**, 94 (2010).
71. X. Wang, Z. Wu, and X. Zhang, *J Bioinform Comput Biol* **8 Suppl 1**, 177 (2010).

72. V. G. Tusher, R. Tibshirani, and G. Chu, *Proc Natl Acad Sci U S A* **98** (9), 5116 (2001).
73. G. R. Grant, E. Manduchi, and C. J. Stoeckert, Jr., *Curr Protoc Mol Biol* **Chapter 19**, Unit 19 6 (2007).
74. G. R. Grant, J. Liu, and C. J. Stoeckert, Jr., *Bioinformatics* **21** (11), 2684 (2005).
75. B. Langmead, K. D. Hansen, and J. T. Leek, *Genome Biol* **11** (8), R83 (2010).
76. M. D. Robinson and G. K. Smyth, *Bioinformatics* **23** (21), 2881 (2007).
77. M. D. Robinson, D. J. McCarthy, and G. K. Smyth, *Bioinformatics* **26** (1), 139 (2009).
78. S. Anders and W. Huber, *Genome Biol* **11** (10), R106 (2010).
79. L. Wang, Z. Feng, X. Wang et al., *Bioinformatics* **26** (1), 136 (2009).
80. C. H. Jan, R. C. Friedman, J. G. Ruby et al., *Nature* **469** (7328), 97 (2010).
81. M. Mangone, A. P. Manoharan, D. Thierry-Mieg et al., *Science* **329** (5990), 432 (2010).
82. C. Plessy, N. Bertin, H. Takahashi et al., *Nat Methods* **7** (7), 528 (2010).
83. I. Maccallum, D. Przybylski, S. Gnerre et al., *Genome Biol* **10** (10), R103 (2009).
84. S. Lee, C. H. Seo, B. Lim et al., *Nucleic Acids Res* (2010).

FIGURE 1

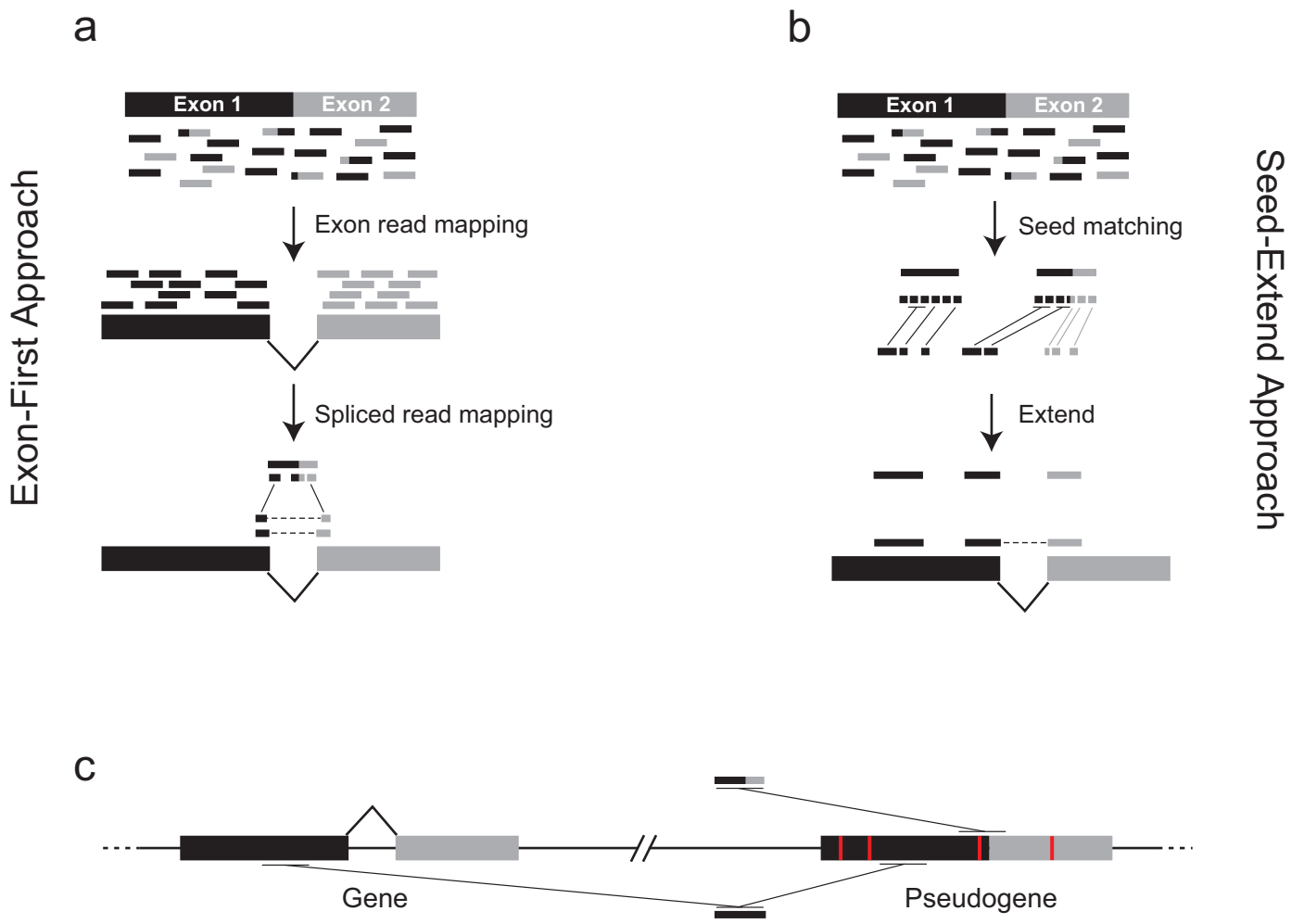
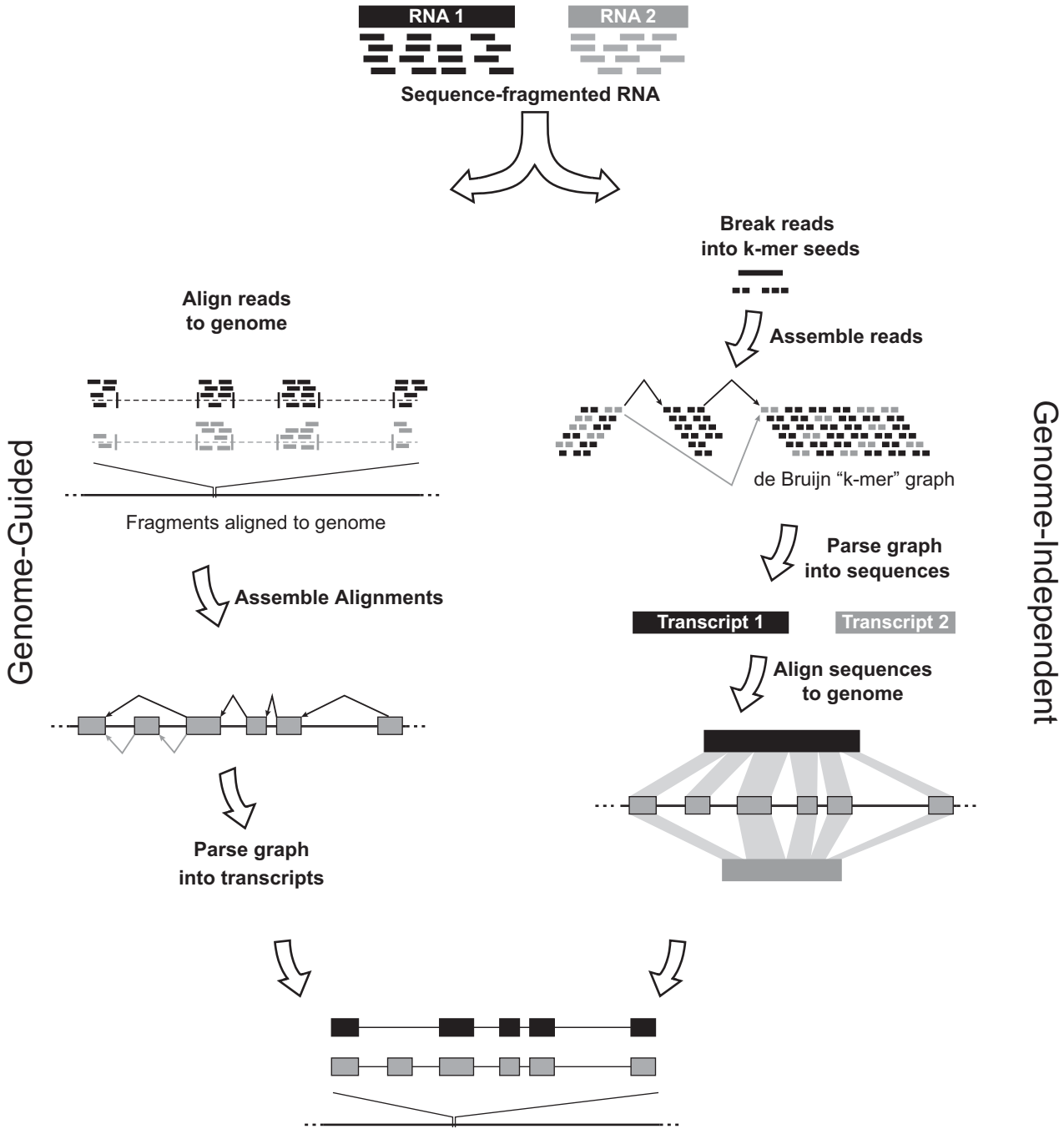
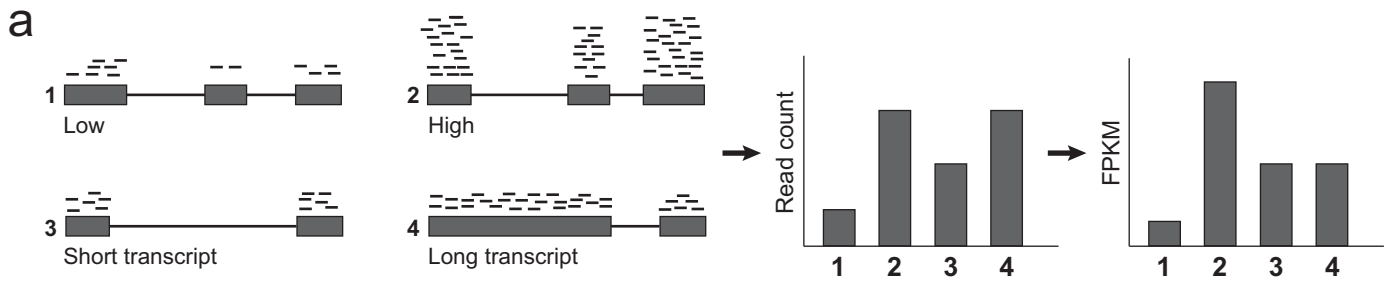
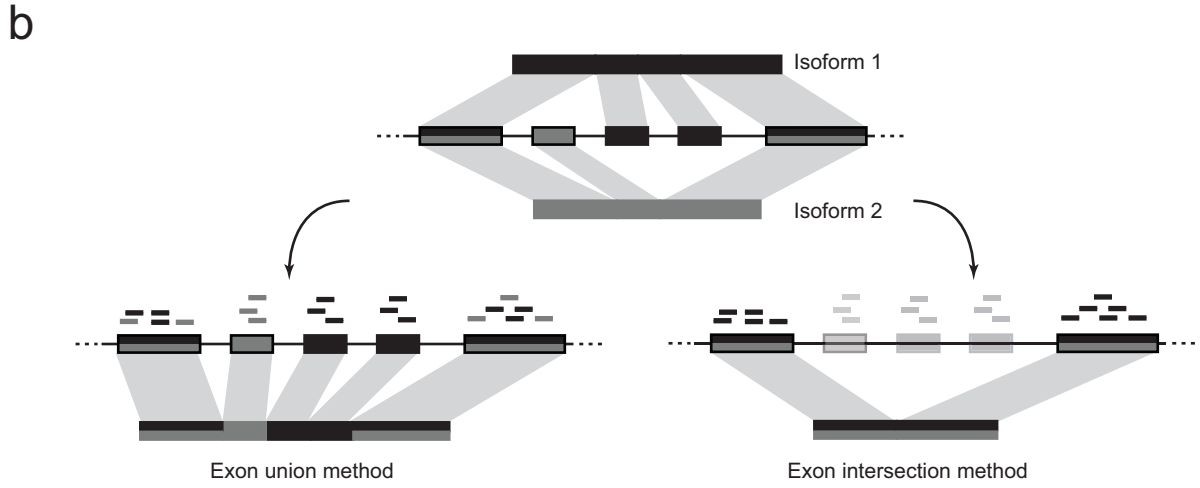
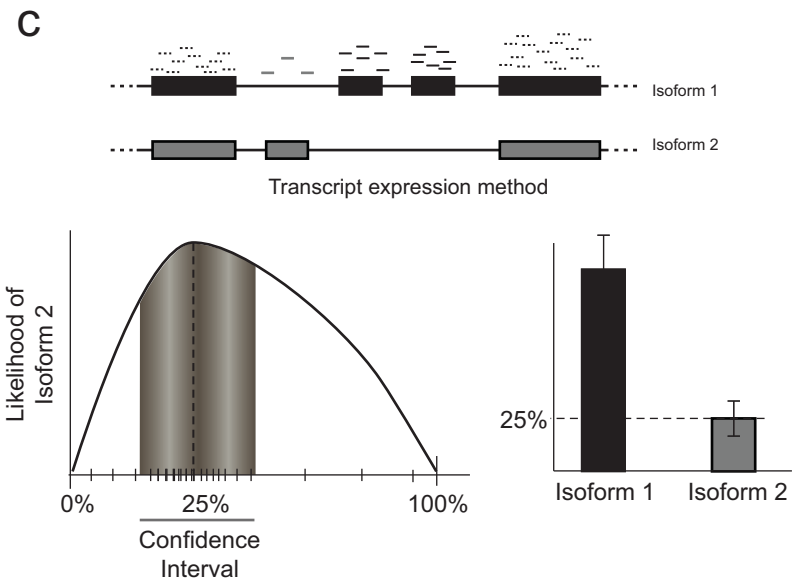

a
b
Exon-First Approach
Seed-Extend Approach
Exon 1
Exon 2
Exon read mapping
Spliced read mapping
Seed matching
Extend
c
Gene
Pseudogene

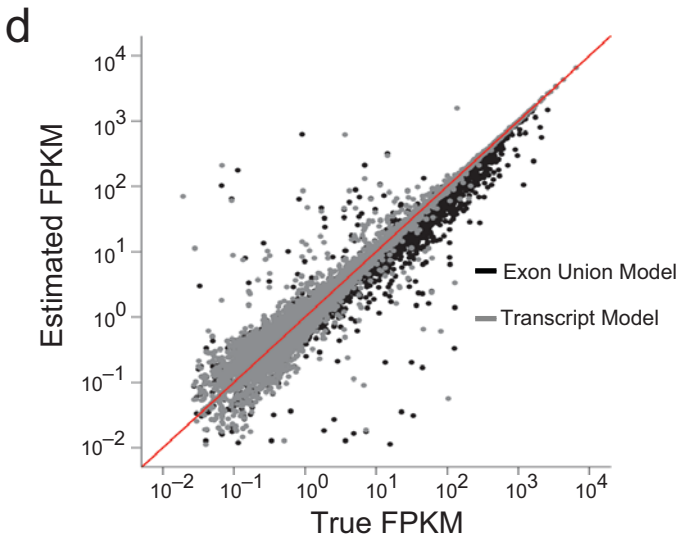
FIGURE 2

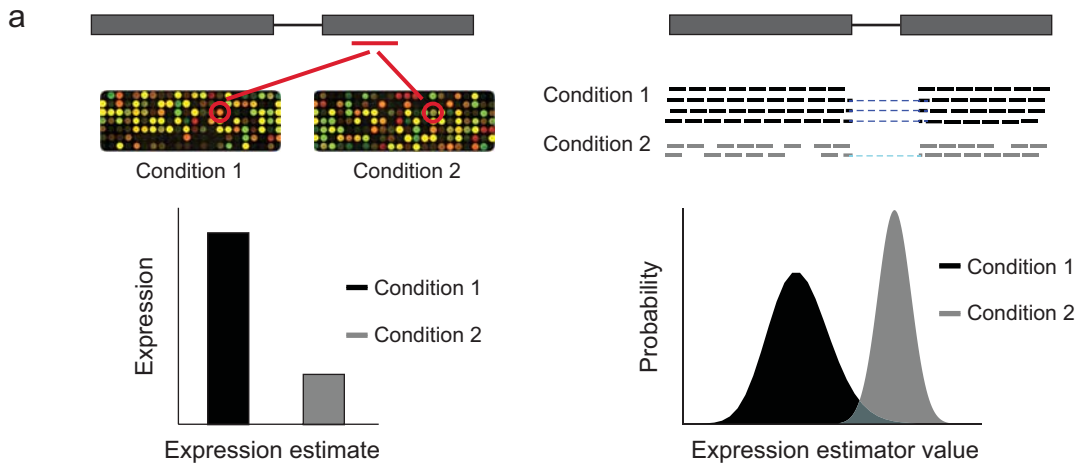

RNA 1
RNA 2
Sequence-fragmented RNA
Break reads into k-mer seeds
Assemble reads
de Bruijn "k-mer" graph
Parse graph into sequences
Transcript 1
Transcript 2
Align sequences to genome
Genome-Independent
Align reads to genome
Fragments aligned to genome
Assemble Alignments
Parse graph into transcripts
Genome-Guided

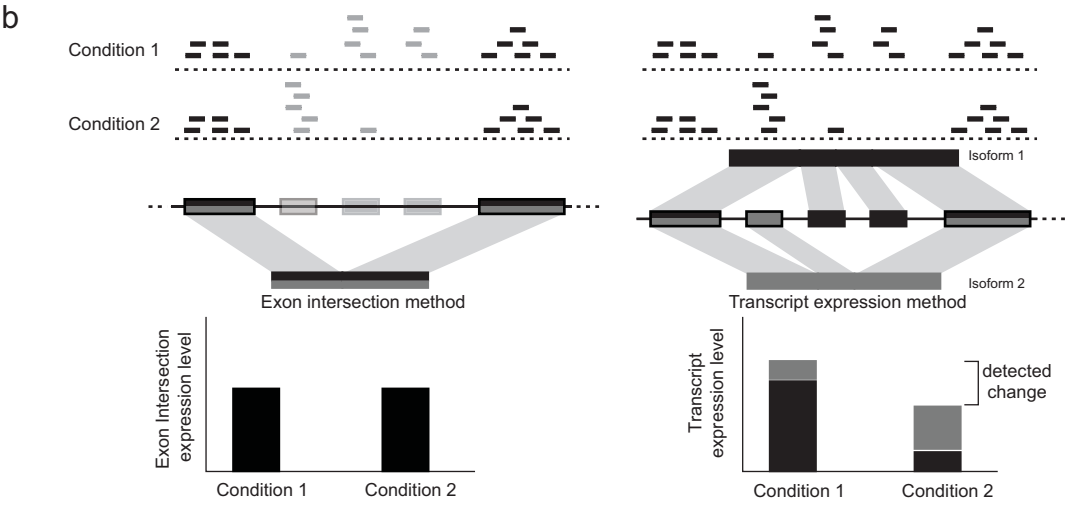

a
1
Low
2
High
3
Short transcript
4
Long transcript
Read count
1 2 3 4
FPKM
1 2 3 4


b
Isoform 1
Isoform 2
Exon union method
Exon intersection method


c
Isoform 1
Isoform 2
Transcript expression method
Likelihood of Isoform 2
0%
25%
100%
Confidence Interval
25%
Isoform 1
Isoform 2


d
Estimated FPKM
True FPKM
Exon Union Model
Transcript Model


a
Condition 1
Condition 2
Expression
Condition 1
Condition 2
Expression estimate
Condition 1
Condition 2
Probability
Condition 1
Condition 2
Expression estimator value


b
Condition 1
Condition 2
Exon intersection method
Exon Intersection expression level
Condition 1
Condition 2
Isoform 1
Isoform 2
Transcript expression method
Transcript expression level
detected change
Condition 1
Condition 2