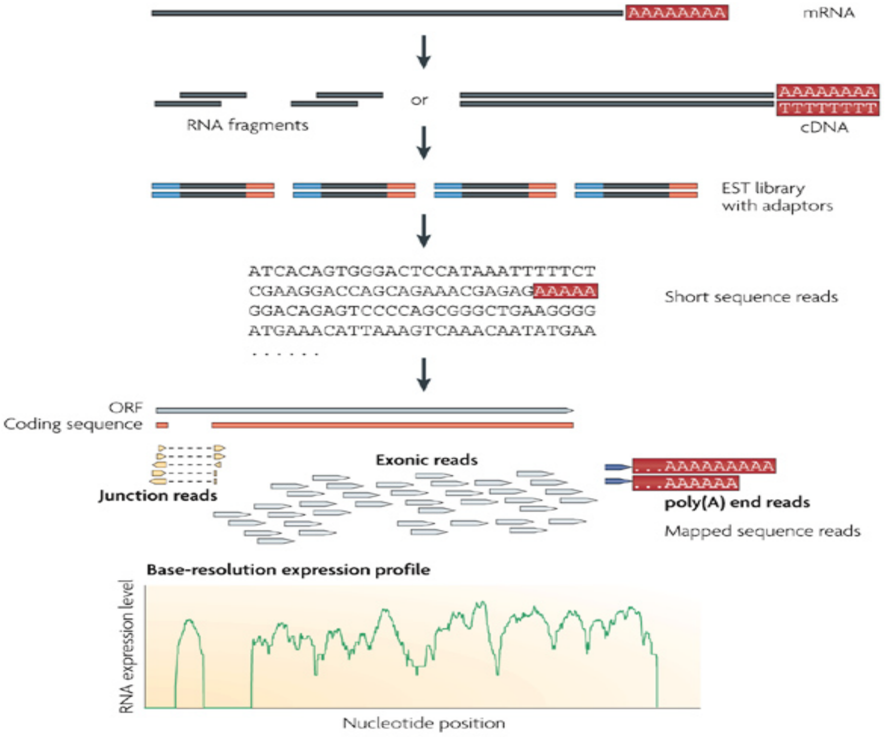
Bioc218 final paper
Bing Wu 3/13/2014

Review of Computation for RNA-Seq Studies: Development and Improvement

A comprehensive picture of the transcriptome – the identity and amount of each individual RNA molecule, for each cell type and state has been a holy grail for biologists working on a wide range of topics. For example, analysis of the transcriptome of embryos provide new information about molecular mechanisms underlying cell differentiation and organism development [1]; transcriptome of cancer cells extend our understanding of carcinogenesis [2] and hopefully assist in selecting drug targets.

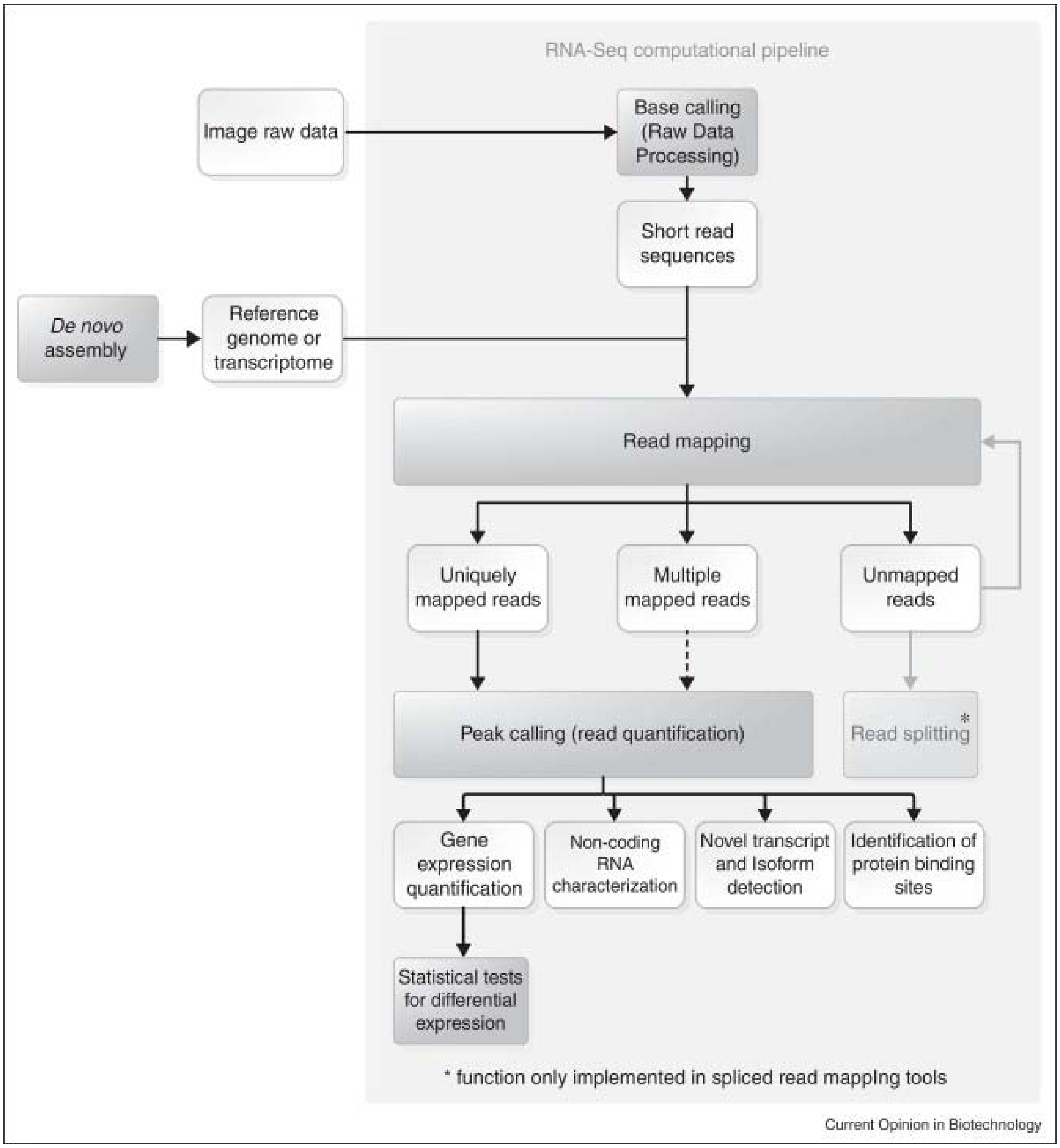
RNA-seq (RNA Sequencing) is the technique that reveals the sequences and quantities of the RNA present in the sample, allowing subsequent computational analysis to reconstruct the transcriptome. A typical RNA-seq experiment is illustrated in the diagram below. In this schematic, mRNA (Poly(A)+) is used as an example; however, in general a total ensemble of RNA (including mRNA, rRNA, tRNA and other non-coding RNA) can be fragmented into short sequences (200-500bp) and converted into a library of complementary DNA (cDNA). The DNA fragments are then attached to adaptors and rendered to high-throughput sequencing [4]. Variable methods in library construction creates different biases, complicating the analysis of the sequencing results [5]; an ideal approach should be capable of directly reading each individual RNA sequence, long or short. However, this review will only focus on the computational techniques and challenges after acquiring the short sequence reads using the current approaches.



Nature Reviews | Genetics

Computational analysis in RNA-seq

The analysis pipeline of RNA-seq is summarized in the diagram below. Providing that a reference genome (or transcriptome) is pre-existent, short sequencing reads usually need to be mapped to the reference in the first place. However, short reads can undergo *de novo* assembly too, which will be described later. The aligned reads are then assembled to reconstruct the transcriptome. Note that besides the short reads, RNA-seq also generates read count data, from which transcript expression levels can be estimated. Normalizing and quantifying expression values will be followed by statistical analyses that assign significance to differential expression levels.


RNA-Seq computational pipeline
Image raw data
Base calling (Raw Data Processing)
Short read sequences
De novo assembly
Reference genome or transcriptome
Read mapping
Uniquely mapped reads
Multiple mapped reads
Unmapped reads
Peak calling (read quantification)
Read splitting *
Gene expression quantification
Non-coding RNA characterization
Novel transcript and Isoform detection
Identification of protein binding sites
Statistical tests for differential expression

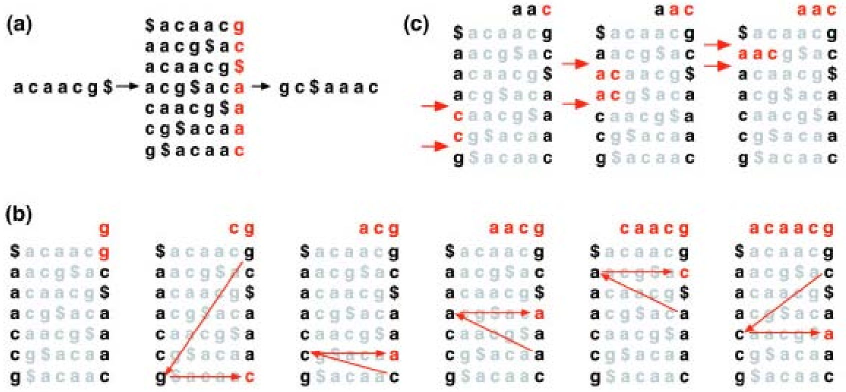
* function only implemented in spliced read mapping tools

Current Opinion in Biotechnology

1. Read mapping

Aligning reads to a reference genome or transcriptome resembles classic alignment problems. However, the short RNA-seq reads make the sequencing error rate, the genuine differences between reference and query organisms, and the RNA spanning exon-exon junctions more considerable. Currently RNA-seq read mapping approaches can be classified into two categories: “unspliced read aligners”, which align reads to reference without allowing large gaps, and “spliced aligners”, which aligning reads to the entire genome permitting large gaps for intron-spanning reads.

Bowtie alignment program [6] uses “Burrows-Wheeler transform methods” - one of the two main algorithms for “unspliced read aligners”, to index the reference genome, as shown below. Such Burrows-Wheeler transform allows large texts to be searched with economic memory footprint.

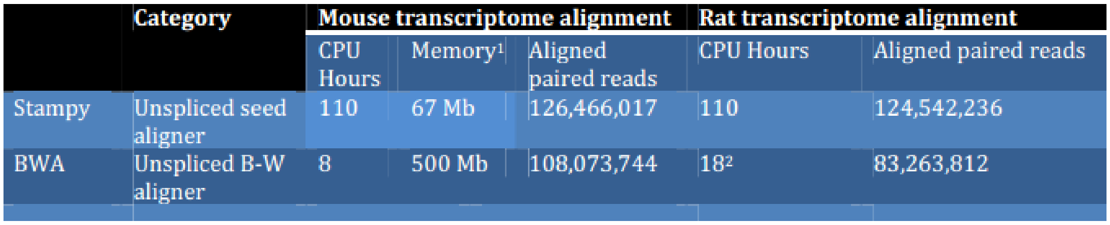


Burrows-Wheeler transform. **(a)** The Burrows-Wheeler matrix and transformation for 'acaacg'. **(b)** Steps taken by EXACTMATCH to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'. **(c)** UNPERMUTE repeatedly applies the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column).

Adapted from Langmead *et al. Genome Biology* 2009 **10**:R25

Bowtie further introduces some extensions such as a backtracking algorithm to allow sequencing errors or genetic variations, and a “double indexing” strategy to prevent over-backtracking. Alternatively, “seed methods” are used by another set of match finding programs, such as MAQ [7] and Stampy [8]. In these methods, reads are broken down into shorter subsequences – “seeds”, assuming that at least some seeds are matched perfectly to the reference, to reveal candidate mapping regions. Afterwards, other alignment methods, such as Smith-Waterman are used to extend the alignments.

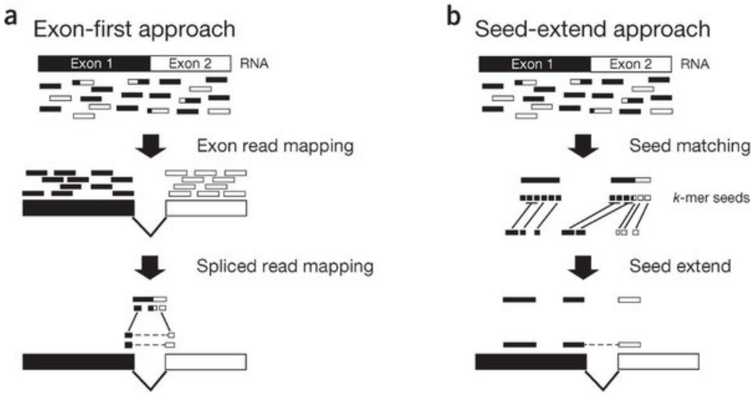
Compared to “seed methods”, “Burrows-Wheeler transform methods” are generally faster when the exact reference is available. However, when only an imprecise reference (for instance, a genome of a distant species) is available, “seed methods” turn out more sensitive, yielding more aligned reads. The table below shows comparison of unspliced Seed (Stampy) and Burrows-wheeler (BWA) aligners for mapping reads to both the mouse and rat transcriptome consisting of 8,557 genes expressed in mES that have a rat ortholog.



	Category	Mouse transcriptome alignment			Rat transcriptome alignment	
		CPU Hours	Memory[1]	Aligned paired reads	CPU Hours	Aligned paired reads
Stampy	Unspliced seed aligner	110	67 Mb	126,466,017	110	124,542,236
BWA	Unspliced B-W aligner	8	500 Mb	108,073,744	18[2]	83,263,812

The unspliced read aligners are specialized at identifying known exons and isoforms; however, they are not able to identify novel splicing events, which require reads to be spliced and aligned separately. “Seed methods” can actually be adapted into a “spliced aligner”. Seeds from the same read can be placed onto different regions onto the genome. The subsequent extension and merging of extended seeds will determine the full spliced alignment for the read [9]. Another common algorithm of “spliced aligners” is the “exon first” method, including two steps. Firstly, the complete reads are mapped to the reference with “unspliced read aligners”. Secondly, the remaining unmapped reads are broken down into shorter segments and mapped independently, revealing the spliced points [10].

Since the “exon first” strategy requires to steps, in which the reads succeeding in unspliced alignment will not go through the second step of spliced aligning, spliced alignments of reads that can also be mapped to the genome contiguously will be missed. In contrast, “seed methods” evaluate spliced and unspliced alignments at the same time, and thereby the bias toward unspliced alignments is prevented.



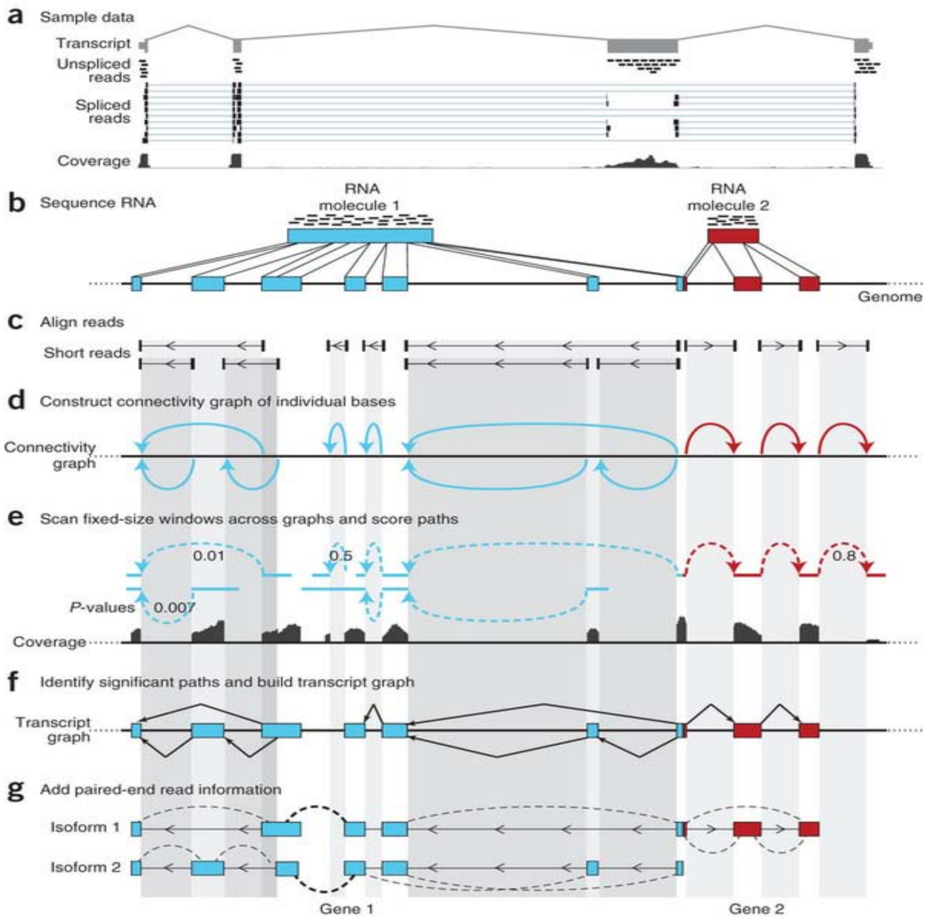
Exon-first methods (**a**) map full, unspliced reads (exonic reads), and remaining reads are divided into smaller pieces and mapped to the genome. An extension process extends mapped pieces to find candidate splice sites to support a spliced alignment. Seed-and-extend methods (**b**) store a map of all 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.

Adapted from Nature Methods 8,469–477(2011) [3]

2. Transcriptome reconstruction

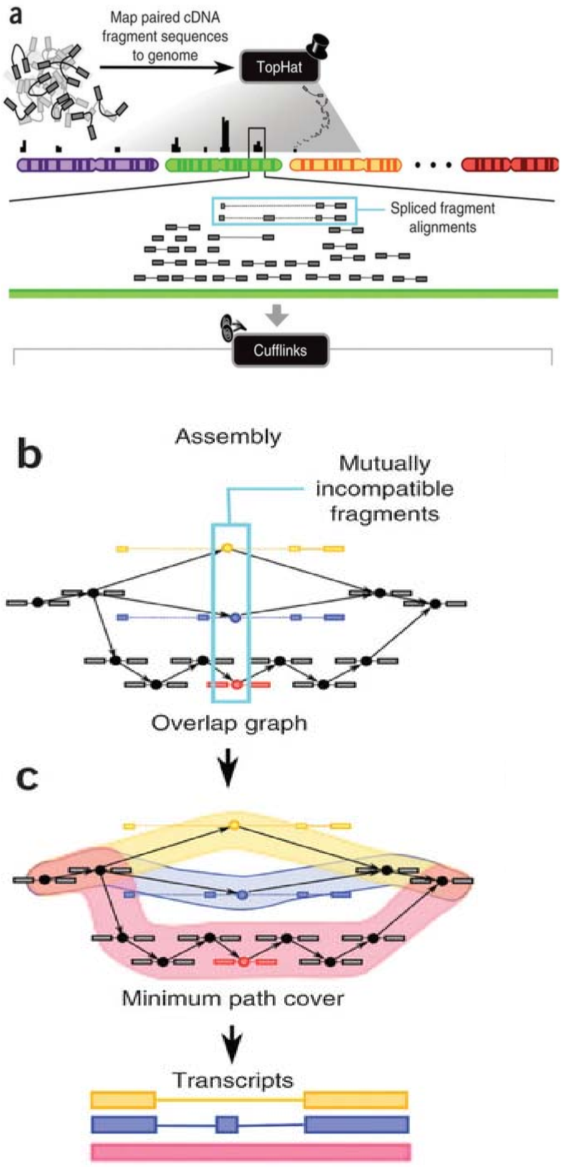
After the short reads from RNA-seq are mapped onto the reference genome, a graph will be built to represent closely adjacent/overlapping reads for each genomic locus. Traversing the graph will finally combine the connected exons into transcriptional units, determining the isoforms, revealing novel transcripts, and reconstruct the complete transcriptome.

Scripture is a method developed upon “spliced aligners”. Two bases will be connected if they are immediate neighbors either in the genomic sequence or within a short sequencing read. Reads covering overlapping locus will be clustered to generate a connectivity graph. Then both spliced and unspliced reads will be used to identify paths in the connectivity graph [11].



(**a**) Spliced and unspliced reads. Unspliced reads (black bars) fall within a single exon, whereas spliced reads (bars broken into 'dumbbells') span exon–exon junctions (thin horizontal lines connect the alignment of a read to the exons it spans). (**b**) Shown are transcripts from two different genes (blue and red boxes). The grayscale vertical shading in subsequent panels is shown for visual tracking. (**c**) Spliced reads. (**d**) Connectivity graph construction. Scripture builds a connectivity graph by drawing an edge (curved arrow) between any two bases that are connected by a spliced read gap. (**e**) Path scoring. Scripture scans the graph with fixed-sized windows and uses coverage from all reads (spliced and unspliced; bottom track) to score each path for significance (*P*-values shown as edge labels). (**f**) Transcript graph construction. Scripture merges all significant windows and uses the connectivity graph to give significant segments a graph structure (three graphs, in this example). (**g**) Refinement with paired-end data. Scripture uses paired-end (dashed curved lines) to join previously disconnected graphs (gene 1, bold dashed line), find breakpoint regions within contiguous segments (detectable in this example by the lack of dashed lines between genes 1 and 2) and eliminate isoforms that result in paired-end reads mapping at a distance with low likelihood.

Cufflinks [12] is another program for transcript discovery and transcriptome reconstruction. Reads aligned across splice junctions are taken in and clustered based on the genomic loci onto which they are mapped. Compatible fragments will be connected in the overlap graph and merged into complete isoforms, whereas those incompatible reads must be originated from distinct spliced isoforms. Cufflinks implements Dilworth's Theorem (the number of mutually incompatible reads is the same as the minimum number of transcripts needed to 'explain' all the fragments) and produces a minimal set of paths covering all fragments in the overlap graph by finding the largest set of reads in which no two could have originated from the same isoform.


a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts

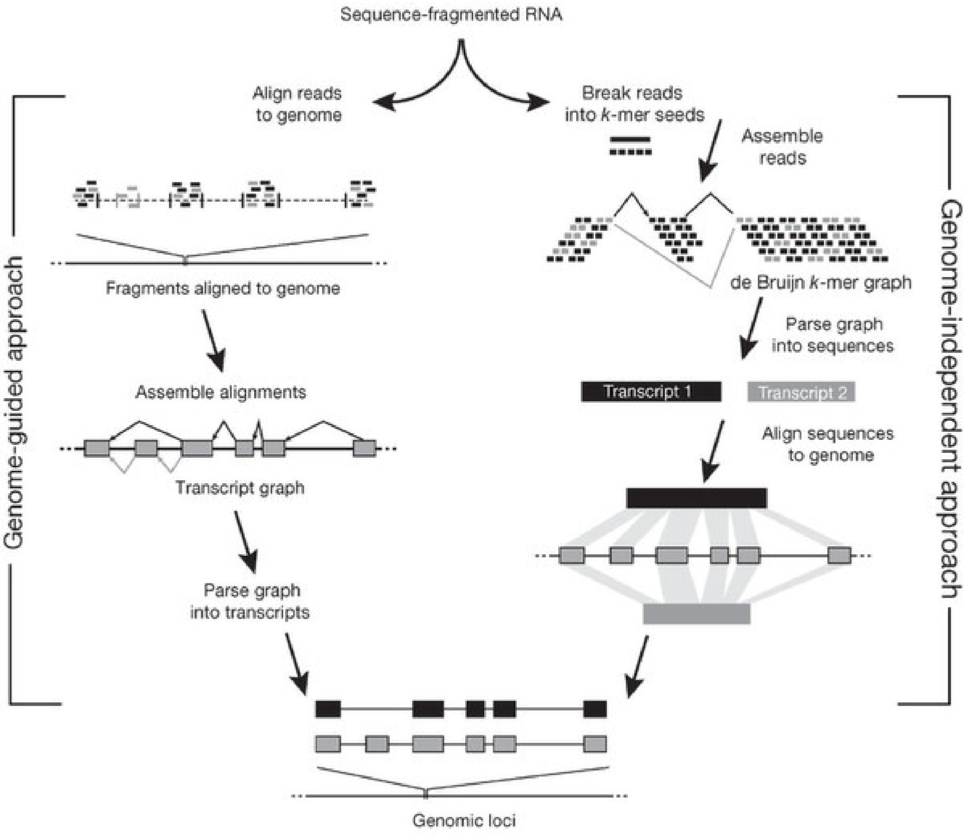
(**a**) The algorithm takes as input cDNA fragment sequences that have been aligned to the genome by software capable of producing spliced alignments, such as TopHat. The first step in fragment assembly is to identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms (**b**). Fragments are connected in an 'overlap graph' when they are compatible and their alignments overlap in the genome. In this example, the yellow, blue and red fragments must have originated from separate isoforms, but any other fragment could have come from the same transcript as one of these three. Isoforms are then assembled from the overlap graph (**c**).

Scripture and Cufflinks differ majorly in graph construction and traversal methods. Cufflinks is more conservative in its choice of which transcripts to re-construct, since it sticks to “the minimum number of transcripts”; whereas Scripture may produce a larger set of transcripts for the same genomic region.

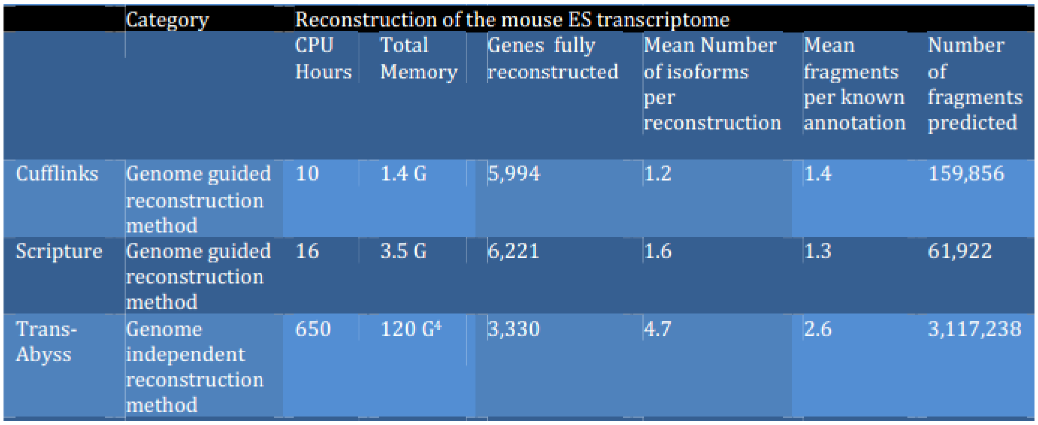
Both Scripture and cufflinks are reference-based transcriptome assembly strategies, which are advantageous in detecting and assembling transcripts of low abundance with their high sensitivity. Another bonus is that since the underlying genome sequence is already known, small gaps within the transcript caused by a lack of read coverage can be filled in using the reference.

However, in some cases a reference genome is not available, or is of low quality (with considerable genome misassemblies and genomic deletions). To address such problems, *de novo* transcriptome assembly strategy has been developed. Instead of mapping reads onto the reference genome, *de novo* take advantage of the redundancy of the short reads and merge these partially overlapping reads into complete transcripts. The common strategy *de novo* assemblers use is de Bruijn graph [13], in which reads are broken into shorter overlapping subsequences of length k base pairs, termed 'k-mers' (k consecutive nucleotides). This reduces the complexity associated with handling millions of reads to a fixed number of possible k-mers. Next, paths are traversed in the graph, eliminating false branch points introduced by k-mers that are not supported by reads. The remaining paths through the graph are then reported as individual transcripts.

More recently, the Trans-ABySS method has been developed which specialized at assembling non-normalized transcriptome data (with intrinsic variability in transcript abundance), by assembling k-mers at different k values, to achieve higher sensitivity [14].



Ideally, a hybrid approach incorporating both the genome-independent and genome-guided strategies will cope with both capturing known information with high sensitivity, as well as detecting novel variation. Practically, genome-independent methods require significantly greater computational resources compared to genome-guided methods [3].

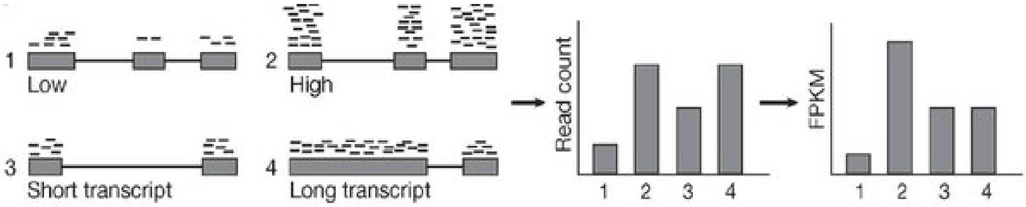


	Category	Reconstruction of the mouse ES transcriptome					
		CPU Hours	Total Memory	Genes fully reconstructed	Mean Number of isoforms per reconstruction	Mean fragments per known annotation	Number of fragments predicted
Cufflinks	Genome guided reconstruction method	10	1.4 G	5,994	1.2	1.4	159,856
Scripture	Genome guided reconstruction method	16	3.5 G	6,221	1.6	1.3	61,922
Trans-Abyss	Genome independent reconstruction method	650	120 G[4]	3,330	4.7	2.6	3,117,238

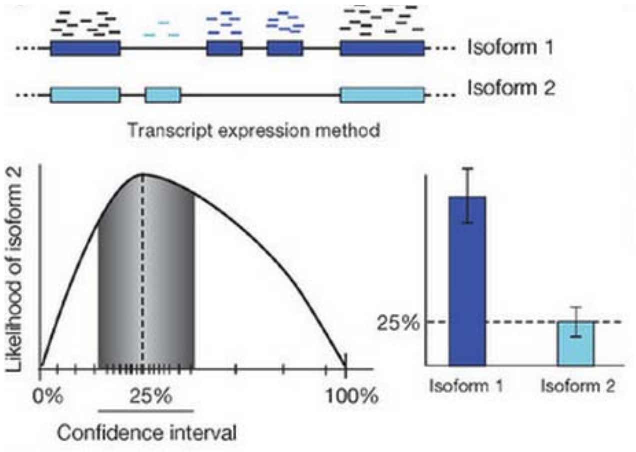
3. Expression level quantification

As another set of information generated by RNA-seq, the read counts are informative in estimating the differential expression level of the transcripts. A highly expressed transcript is represented by a larger RNA copy number, which will be detected by more reads in sequencing.

However, a number of systematic variability complicates the correlation between expression levels and read counts. For instance, the length of the transcript also contributes to the number of reads[15]. To address this issue, the reads per kilobase of transcript per million mapped reads (RPKM) metric is usually used to normalize the read counts by both the length and the number of mapped reads of each transcript [7, 12].

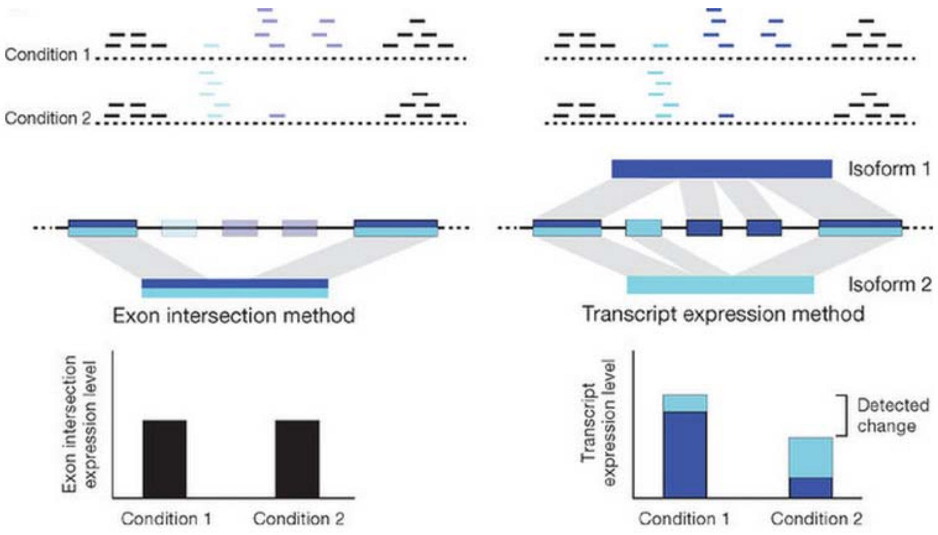


A second challenge is that many reads cannot be assigned to a transcript unequivocally when several isoforms are produced from the corresponding genes[16]. One strategy is to quantify expression level by counting only the reads that are aligned to a unique isoform [17]. However, this method will not work for the transcripts that do not contain any unique isoform. To handle this situation, “Isoform expression methods” are developed to achieve the maximum likelihood estimate (MLE) in a "likelihood function" modeling the sequencing process and assign isoform abundance that can best explain the reads obtained; in addition, expression quantity will also be modified by "sampling" alternative abundance estimates around the MLE to improve the robustness of this method for genes expressed at low levels [12, 18].



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.

Practically, estimating the differential expression for each isoform might not be a necessary goal. "Exon intersection method" [19], which counts reads mapped to its constitutive exons, and the "exon union method" [7], which counts all reads mapped to any exon in any of the gene's isoforms are used to simplify the quantification of gene expression level. However, this simplification may miss the expression difference between two samples, as illustrated by the hypothetical gene shown below. In other words, if the gene-level read counts are similar in two samples, but distributed differently among the isoforms, differential expression results will differ depending on the counting method used.



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.

Conclusions and suggested improvements

Computation in RNA-seq has developed rapidly with the improvement of sequencing technologies and biological questions. However, no consensus has been achieved on the best pipeline of identifying spliced isoforms and determining the expression levels accordingly. Extending the read length could alleviate this problem, since a longer read would more likely span multiple junctions and provide evidence for more spliced events. However, as read length continues to increase, new mapping methods will need to align hundreds of millions of long reads spanning a growing but uncertain number of junctions, which will be a daunting task.

Here I suggest that differential read counts between exons from the same gene could indicate splicing events, and therefore help in interpreting isoforms. Current analysis approaches usually assemble the transcriptome, identifying isoforms according to spliced alignments, and then assigning abundance to each isoform based on read counts. Conversely, the base-resolution expression profile is potentially capable of revealing differential expression between two adjacent exons. If the normalized abundance of one exon is statistically smaller than that of the one preceding it, it is a strong implication that the preceding exon can be spliced together with the subsequent sequences, skipping the second exon. This source of splicing information can serve as complement/feedback in transcriptome reconstruction, especially when the mapping procedure is less sensitive than catching every possible spliced alignment.

Reference

1. Xue, Z., et al., *Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing.* Nature, 2013. **500**(7464): p. 593-7.
2. Sinicropi, D., et al., *Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue.* PLoS One, 2012. **7**(7): p. e40092.
3. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nat Methods, 2011. **8**(6): p. 469-77.
4. Mutz, K.O., et al., *Transcriptome analysis using next-generation sequencing.* Curr Opin Biotechnol, 2013. **24**(1): p. 22-30.
5. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
6. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
7. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
8. Lunter, G. and M. Goodson, *Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.* Genome Res, 2011. **21**(6): p. 936-9.
9. Wu, T.D. and S. Nacu, *Fast and SNP-tolerant detection of complex variants and splicing in short reads.* Bioinformatics, 2010. **26**(7): p. 873-81.

10. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
11. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.
12. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
13. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
14. Robertson, G., et al., *De novo assembly and analysis of RNA-seq data.* Nat Methods, 2010. **7**(11): p. 909-12.
15. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**: p. 14.
16. Li, B., et al., *RNA-Seq gene expression estimation with read mapping uncertainty.* Bioinformatics, 2010. **26**(4): p. 493-500.
17. Griffith, M., et al., *Alternative expression analysis by RNA sequencing.* Nat Methods, 2010. **7**(10): p. 843-7.
18. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.
19. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.