Next-generation transcriptome assembly

Jeffrey A. Martin[a], Zhong Wang[a]

[a]US Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598

[a]*To whom correspondence may be addressed. E-mail:* jamartin@lbl.gov or zhongwang@lbl.gov

September 14, 2011

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

Review for *Nature Reviews Genetics* — 'Study Designs' series

Next-generation transcriptome assembly

Jeffrey Martin, Zhong Wang

Jeffrey Martin
Bioinformatics Systems Analyst
DOE Joint Genome Institute
2800 Mitchell Dr., MS100
Walnut Creek, CA 94598, USA
Tel: (925)-927-2908
Email: JAMartin@lbl.gov

Zhong Wang, Ph.D.
Staff Scientist, Group Lead for Genome Analysis
DOE Joint Genome Institute
2800 Mitchell Dr., MS100
Walnut Creek, CA 94598, USA
Tel: (925)-296-5795
Email: ZhongWang@lbl.gov

Transcriptomics studies require a high quality, comprehensive reference transcriptome that includes all transcripts, coding and noncoding, large and small. Recent advances have enabled the *de novo* reconstruction of the entire transcriptome by deep RNA-Seq, even without a reference genome. However, transcriptome assembly from billions of RNA-Seq reads, often very short, poses a significant informatics challenge. This review summarizes recent developments in transcriptome assembly strategies, along with some perspectives on transcriptome assembly in the near future.

Studying the dynamics and regulation of the transcriptome holds the key to understanding the function of a genome and the underlying biological processes. For a long time the accuracy and comprehensiveness of transcriptomics studies have been limited because our knowledge of the transcriptome has been partial and biased, since it is largely derived from gene prediction and limited EST evidence. Whole transcriptome sequencing by next-generation sequencing (NGS) technologies or RNA-Seq has started to reveal the complex landscape and dynamics of the transcriptome from yeast to human at an unprecedented level of sensitivity and accuracy[1-4].

Compared to traditional low-throughput EST sequencing by Sanger technology, the enormous [**it would be better to be more precise about the sequencing depth rather than say 'enormous**', **especially in light of the queries of the referees about the deep sequencing of rare transcripts**]sequencing depth of a typical RNA-Seq experiment offers a near complete snapshot of a transcriptome, including the rare transcripts that play regulatory roles. In contrast to alternative high-throughput technologies such as microarrays, RNA-Seq achieves base-pair-level resolution, much higher dynamic range, and is capable of *de novo* annotation[1,2]. Despite these advantages, sequence reads obtained from the common NGS platforms, including Illumina, SOLiD, and 454, are often very short, ranging from 35bp to 500bp[5]. As a result, it is necessary to reconstruct the full-length transcripts by transcriptome assembly. Small classes of RNA (such as microRNAs, piRNAs, snoRNAs, siRNAs **[these small RNAs could be rolled into one glossary definitio**n]) are shorter than the sequencing length and do not require assembly.

Reconstructing a comprehensive transcriptome from short reads has many informatics challenges. Similar to short-read genome assembly, transcriptome assembly involves piecing together short, relatively low quality reads. Typical NGS datasets are very large (several gigabases to terabases), which poses a stringent requirement for computing systems to have large memory and/or many cores to run parallel algorithms. Several short-read assemblers have been developed to tackle these challenges, including Velvet[6], ABYSS[7], ALLPATHS[8] and several

others[9]. These tools have achieved reasonable success in the assembly of genomes[9,10]. However, they may not be directly applied to transcriptome assembly mainly because of three considerations. First, whereas DNA sequencing depth is expected to be the same across a genome, the sequencing depth of transcripts can vary by several orders of magnitude. Many short-read genome assemblers use sequencing depth information for discerning repetitive regions of the genome, a feature that is problematic for transcriptome assembly. Sequencing depth is also used by assemblers to calculate an optimal set of parameters for genome assembly, which likely results in only a small set of transcripts being favoured in transcriptome assembly. Second, unlike genomic sequencing, where both strands are sequenced, RNA-Seq experiments are usually strand-specific. To be effective, transcriptome assemblers will need to take advantage of strand information to resolve overlapping sense and anti-sense transcripts[11-14]. Finally, it is generally difficult for short read assemblers to resolve repeat structures in a genome assembly; this problem is exacerbated during transcriptome assembly because transcript variants from the same gene can share many exons.. Given the complexity of most transcriptomes and the above challenges, reconstructing all the transcripts and their variants exclusively from short reads has been viewed as being very difficult.

In the past three years, several breakthroughs have been made to address the above challenges, thanks to improvements in data quality and the rapid evolution of assembly algorithms. In this review, we summarize these exciting breakthroughs that have resulted in a wealth of assembled transcriptomes from short reads[16-27], while providing practical guidelines for implementing a transcriptome assembly experiment. We discuss the experimental and informatics considerations that need to be made before assembly, such as RNA-Seq library construction, data pre-processing and how to assess the assembly quality. Three assembly strategies will be discussed: assembly based upon a reference genome, *de novo* assembly, and a hybrid approach that combines both approaches. We focus on the strengths and weaknesses of the three strategies, in the context of small, gene-dense transcriptomes and large transcriptomes

with pervasive alternative splicing. Finally, we give some perspectives on the future of transcriptome assembly, in light of the rapid evolution in sequencing technology and high performance computing.

Considerations prior to assembly

To ensure a high quality transcriptome assembly, special considerations should be made in designing the RNA-Seq experiment prior to assembly. The steps of a typical transcriptome assembly experiment are shown in **Figure 1**. In the data generation phase, total RNAs or mRNAs are fragmented and converted into a library of cDNAs with sequencing adapters. The cDNA library is then sequenced by NGS sequencers to produce millions to billions of short reads from one end or both ends of the cDNA fragments. In the data analysis phase, these short reads are pre-processed to remove sequencing errors and other artifacts, and subsequently assembled to reconstruct the original RNAs and assess their abundance ('expression counting'). The library construction methods, sequencing technologies, and data pre-treatment techniques are known to influence the accuracy and precision of gene expression counting[28]. Likewise, these factors can also impact the quality of assembled transcriptomes, as discussed below.

Library construction. To increase the number of assembled transcripts, especially the less abundant ones, ribosomal RNA (rRNA) and abundant transcripts are removed during the first steps of library construction. Poly(A) selection is very effective at enriching mRNAs in eukaryotes, but this selection approach will miss noncoding RNAs (ncRNA) and mRNAs that lack a poly(A) tail. In order to include RNAs without a poly(A) tail in the assembled transcriptome, rRNA contamination can be removed by hybridization-based depletion methods[29,30]. These normalization techniques can reduce the representation of highly abundant transcripts by many fold[31], thereby increasing the opportunity for assembling rare transcripts. Another consideration during library construction is to eliminate the PCR amplification step from the standard protocols. Recently it has shown that amplification-free protocols can reduce

the bias [which type of bias?] originated from PCR amplification[32,33]. Sequencing coverage of the transcriptome from these protocols is more even and contiguous across transcripts, making it easier to construct full-length transcripts. Lastly, strand-specific protocols[34] allow overlapping transcripts derived from opposite strands to be separated. This consideration is especially important for gene-dense genomes, such as prokaryotes and lower eukaryotes, where overlapping genes are very common.

Sequencing. Each of the current NGS technologies has been used to successfully assemble transcriptomes[35-37], and they differ mostly in throughput and cost. In general, the assembly of large and complex transcriptomes (plants and mammals) requires more sequencing depth and is frequently done on Illumina or SOLiD platforms. However, the 454 technology offers longer reads and it can be used in combination with the other two platforms for "hybrid assembly", where short reads with greater sequencing depth assemble into contigs and long reads help to scaffold the contigs and resolve variants[38,39]. It is worth noting that the short read problem can also be alleviated by using a paired-end protocol, where DNA fragments (100-250bp) are sequenced 75-150bp from both ends, and the overlapping reads are joined together to form a much longer read[40]. Paired reads from long inserts (500-1000bp) also offer long range connectivity, similar to 454 reads. Some assemblers, such as ALLPATHS, require at least two libraries with different insert sizes, for this reason[8].

Data preprocessing. Removing artifacts from RNA-Seq datasets prior to assembly improves the read quality, while also improving assembly accuracy and computational efficiency. This step is relatively straightforward and can be executed using several tools[41-44]. In general, three types of artifacts should be removed from raw RNA-Seq data: i) sequencing adapters[43,44], which originate from failed or short DNA insertions during library preparation, ii) low-complexity reads[43], and iii) near identical reads derived from PCR amplification[16]. Adapter and low complexity sequences can lead to misassemblies. PCR duplicates are more common in long insert libraries, and their presence can skew mate-pair statistics that are used by many

assemblers for scaffolding. rRNA and contaminant DNA should also be removed to improve assembly speed, although contaminant DNA may not always be detected if the contaminants are unknown. Sequencing errors can also be inferred in the dataset, based upon k-mer frequencies or quality scores. Rare k-mers are generated by sequencing errors or low-abundant transcripts. Reads containing errors can either be removed or trimmed to improve assembly quality and decrease the computational memory required[10,16,42]. However, k-mer based error removal carries a side-effect, in that reads derived from rare transcripts are also removed.

Transcriptome assembly strategies

Depending upon whether or not a reference genome assembly is available, current transcriptome assembly strategies generally fall into one of three categories: reference-based, *de novo*, or a hybrid assembly strategy that combines the two (**Figure 2**). Please note that the hybrid strategy we refer to here is different from the "hybrid assembly" often seen in the literature, which refers to the use of both long and short sequencing reads for assembly. In the following sections we discuss each of these three strategies in detail, including how they work and their pros and cons in the context of both simple and complex transcriptome assembly.

Reference-based strategy

When a reference genome for the target transcriptome is available, the transcriptome assembly can be built upon the reference genome. In general, this strategy involves three steps: aligning the RNA-Seq reads to a reference genome using a splice-aware aligner such as TopHat[45], SpliceMap[46], MapSplice[47], or GSNAP[48] (**Box 1**); clustering overlapping reads from each locus to build a graph representing all possible isoforms, and traversing the graph to resolve individual isoforms (**Figure 2a**). Examples of methods employing this strategy include Cufflinks[22], Scripture[17], and others[18,49] (**Table 1**).

Advantages. The reference-based transcriptome assembly strategy has several advantages. It transforms a large assembly problem (millions of reads) into many smaller

problems (local assembly of each locus, having thousands of reads or less). In this way, assembly can be solved efficiently using parallel computing. Contamination or sequencing artifacts are not a major concern because they are not expected to align to the reference genome. More importantly, the reference-based strategy is very sensitive and can detect genes with low expression levels. Full-length variants can be assembled from only a few folds of sequencing depth[22], and small gaps in read coverage can be filled using the reference sequence[18]. Similarly, this strategy tends to generate longer UTRs, since it recovers the ends of the transcripts, which usually have lower sequencing coverage[17].

Applications. Reference-based transcriptome assembly is easier to perform for the simple transcriptomes of prokaryotic and lower eukaryotic organisms since these organisms have few introns and little alternative splicing. Transcription boundaries can be inferred from regions of contiguous read coverage[37,50,51]. Alternative transcription start and stop sites can also be inferred based upon the 5' cap or poly(A) signals within the mapped reads[50,52]. However, complications arise due to the gene-dense nature of these genomes. Many genes often overlap, resulting in adjacent genes being assembled into one transcript, even though they are not from a polycistronic RNA. Strand-specific RNA-Seq has been used to successfully separate adjacent overlapping genes from opposite strands in the genome[50,51]. However, overlapping genes transcribed from the same strand with comparable expression levels cannot be easily separated without knowledge about their starts and ends.

Plant and mammalian transcriptomes have complex alternative splicing patterns and are challenging to accurately assemble from short reads. Several assemblers, including Cufflinks[22] and Scripture[17], have been developed for efficiently re-constructing transcripts from mammalian-sized datasets. Both algorithms use Tophat[45] to align reads to the genome, but use different graph construction and traversal methods to assemble splicing isoforms[17,22]. A recent study suggested that Cufflinks had higher sensitivity and specificity than Scripture, when detecting previously annotated introns[19], but a comprehensive comparison of the performance of these programs is

needed, as discussed in a later section. Also, it is not known how well these programs perform on polyploid plant transcriptomes, in which different alleles from each subgenome need to be resolved.

Disadvantages. There are a few drawbacks to the reference-based strategy. The success of reference-based assemblers depends on the quality of the reference genome being used. Many genome assemblies contain hundreds to thousands of mis-assemblies and large genomic deletions[53], which may lead to misassembled or partially assembled transcriptomes. Errors introduced by short-read aligners are also carried over into the assembled transcripts. Spliced reads that span large introns can be missed because aligners often only search for introns smaller than a fixed length, to reduce the computation. Reference-based transcriptome assembly is of course not possible without a reference genome. In rare cases, it is possible to use the reference from a closely related species. The strawberry reference genome, for example, was used to assemble the raspberry transcriptome[54]; however in these applications, transcripts from divergent genomic regions would be missed. Lastly, reference-based approaches cannot assemble trans-spliced genes, in which two pre-mRNAs are spliced together into a single mature mRNA[55]. Detection of trans-spliced genes has been shown to be critical for understanding the genetic pathways involved in some cancers[56], such as prostate cancer[57].

In summary, reference-based assembly is generally preferable for cases in which a high quality reference genome already exists. From our experience, these methods are very accurate and sensitive, as they can assemble full-length transcripts at a sequencing depth as low as 10x. When combined with gene predictions, reference-based assembly represents a powerful tool for comprehensive transcriptome annotation.

De novo strategy

When a reference genome is not available or is incomplete, RNA-Seq reads can be *de novo* assembled. A handful of *de novo* transcriptome assemblers have been developed (**Table 1**).

The Rnnotator[16], Multiple-k[21], and Trans-ABySS[19] assemblers follow the same strategy; they assemble the dataset multiple times using a De Bruijn graph-based approach[6-8,58] to reconstruct transcripts from a broad range of expression levels, and then post-process the assembly to merge contigs and remove redundancy (**Figure 2b**). By contrast, other assemblers (Trinity[59], and Oases[20]) traverse the De Bruijn graph directly to assemble each isoform.

Advantages. Compared to the reference-based strategy, *de novo* transcriptome assembly is advantageous in several ways. The obvious advantage is that *de novo* assembly does not depend on a reference genome. Except for a few model organisms most organisms do not have a high quality, finished genome. For such cases, *de novo* assembly can provide an initial set of transcripts, from which RNA-Seq expression studies can be carried out. Sometimes *de novo* assembly should be performed even if a reference genome is available, since it can recover transcripts that are transcribed from highly repetitive genomic regions that are not in the genome assembly, or detect transcripts from contaminants or an unknown source. Another advantage of *de novo* assembly is that it does not depend upon known canonical splice sites[60] or the prediction of novel splicing sites, as required by reference-based assemblers. Similarly, long introns are not a concern for *de novo* assemblers.

Applications. The *de novo* assembly of prokaryotic and lower eukaryotic transcriptomes is relatively easy. Yeast transcriptomes that are sequenced to sufficient depth can be very accurately reconstructed from short 35bp reads. with the majority of the transcripts being assembled to full length[16]. Overlapping genes transcribed from opposite strands in these compact genomes can also be effectively resolved by the alignment of strand-specific reads to the assembled contigs[16], or by not constructing the reverse compliment k-mers in the De Bruijn graph in the first place [**more explanation is needed of this process here and in the figure**]. For overlapping transcripts from the same strand, the *de novo* strategy faces the same challenge as the reference-based approach. In theory, differences in sequencing depth (expression),

signatures of transcription start and end sites, and coding potentials can all be used to separate such cases.

De novo assembly of higher eukaryotic transcriptomes is much more challenging, not only because of the larger size of the datasets, but also due to the difficulty of identifying alternatively spliced variants. Millions to billions of RNA-Seq reads are needed to comprehensively annotate the transcriptome of plants and other large eukaryotes. For large datasets, De Bruijn graph assemblers can easily consume hundreds of gigabytes of RAM, and can run for days to weeks. This problem is alleviated by parallel De Bruijn graph implementations[7,8] that distribute memory over a cluster of computational nodes. Various strategies have been adopted to infer transcript splicing isoforms by interrogating the De Bruijn graph. Oases[20], for example, traverses the De Bruijn graph by applying paired-end read information to assemble isoforms at each locus[25,61]. Trinity[59], on the other hand, implements a step-wise strategy by first greedily assembling the most abundant variants, and then assembling each locus independently with its own De Bruijn graph. Trinity and Trans-ABySS also have a speed advantage because they parallelize the De Bruijn graph construction and traversal for each locus. [**more detail is also needed here to explain the process**]

Disadvantges. Besides the fact that the computing resources needed to *de novo* assemble large transcriptomes can be overwhelming, there are several aspects of the *de novo* assembly strategy that need to be further improved. In general, *de novo* transcriptome assembly requires much higher sequencing depth for full-length gene assembly. A reference-based assembler can reconstruct full-length transcripts with < 10x sequencing coverage[19]. In contrast, a *de novo* assembler usually requires more than 30x coverage for the same task[16]. Furthermore, *de novo* transcriptome assemblers are very sensitive to sequencing errors and to the presence of chimeric molecules in the dataset[62]. Although algorithms have been developed to filter out or correct error-containing reads from abundant transcripts, it is difficult to distinguish these reads from

those derived from low abundance transcripts. So far there is no effective way to discriminate chimeric reads that are artifacts of library preparation from true trans-spliced reads.

Hybrid strategy

Reference-based and *de novo* strategies can be used together, in a hybrid approach, to give a more comprehensive annotation of the transcriptome. By combining these two complementary strategies, one can take advantage of the high sensitivity of reference-based assemblers while leveraging the ability of *de novo* assemblers to detect novel and trans-spliced transcripts. Generally, the hybrid assembly strategy can be carried out by aligning the reads to the reference genome first or *de novo* assembling the reads first[63] (**Figure 2c**). It has not been systematically evaluated to determine which strategy is better, and the choice is likely dependent upon several factors discussed below.

Align-then-assemble. Intuitively, when a high quality reference genome assembly is available, the hybrid approach should start by assembling the dataset using the reference, followed by *de novo* assembly of the reads that failed to align to the genome (**Figure 2c**). As mentioned earlier, *de novo* assembly requires more computing resources, particularly memory, compared to the alignment-based reference strategy. With a nearly complete reference, most of the reads will be assembled, leaving only a small fraction of the reads for *de novo* assembly. This approach is also the preferred method to quickly filter out unwanted sequences, for example in pathogen detection[64], where reads of human origin that form the bulk of the data are filtered out first. When computing resources are limited, the align-then-assemble approach can be used to overcome this limitation.

Assemble-then-align. If the quality of the reference genome is called into question or the reference genome is from a different, but closely related species, *de novo* assembly should be performed first, followed by alignment of the contigs to the reference to extend and scaffold

contigs (**Figure 2c**). The major advantage of this approach is that errors in the genome assembly do not get propagated into the assembled transcripts. As mentioned earlier, *de novo* assembly generates more fragmented transcripts than reference-based assembly. By aligning the assembled transcripts and the unassembled reads to the reference genome, or a closely related one, incomplete transcripts can be extended to form longer, possibly full-length, transcripts. Gaps between fragments of the same transcript can also be joined and filled in using the genomic sequence. Note that one can carry out the alignment step to protein sequences, in cases where the sequence similarity at the RNA level is not high enough for alignment. In a recent study, catfish transcripts were aligned to the stickleback proteome to achieve significantly longer transcripts (the N50 size increased by 27%)[21]. The mosquito transcriptome was scaffolded using the same technique[24].

To date, there are no automated software pipelines that can carry out the hybrid assembly strategy. A systematic study is needed to explore which errors are introduced by hybrid assembly approaches. In the align-then-assemble approach, methods need to be developed to detect the errors in the reference assemblies, in order to prevent them from being propagated into the final assembly. In the assemble-then-align approach, measures must be taken to avoid incorrectly joining segments of different genes (i.e., chimeras).

Assessing assembly quality

While criteria to assess genome assemblies is still under development[53,65], standards for assessing the quality of transcriptome assemblies have not been established, except in a recent study where standardized metrics for assessing the quality of transcriptome assemblies was proposed for a simple transcriptome in which alternative splicing is rare[16]. Here we propose to extend these metrics for both simple and complex transcriptomes. These metrics include accuracy, completeness, contiguity, chimeric, and variant resolution metrics, and they allow for the direct comparison between different assemblies and optimization of assembly parameters

(**Box 2**). All of these metrics can be estimated by using a set of known transcripts as a reference. Among them, the variant resolution metric, for the evaluation of transcriptomes with extensive alternative splicing, is particular challenging because a set of genes with all known isoforms is often not available, as this is one of the problems transcriptome assembly is trying to address. A reference set of transcripts can also be derived from complementary experimental methods. For example, the degree to which full-length protein coding genes are assembled can be evaluated by checking whether or not the alternative isoforms encode full-length ORFs, and by validating the isoforms using proteomics assays[26]. Untranslated regions (UTRs) can be evaluated through other experimental approaches, such as RACE[66].

Conclusions and future perspectives

In summary, many important milestones have been reached which bring us closer to comprehensively annotating and accurately quantifying any transcriptome. Advances in both reference-based and *de novo* transcriptome assembly have expanded RNA-Seq applications to practically any genome. This is particularly important because currently only a small number of species have a high quality reference genome available. The majority of species, especially polyploid plants, lack a reference genome, owing to their genome size and complexity. Another area that is expected to be significantly improved by the advances in *de novo* transcriptome assembly is metatranscriptomics, where thousands of transcriptomes from an entire microbial community are investigated simultaneously.

Advances in high performance computing (HPC) will greatly reduce the time required to assemble large transcriptome or metatranscriptome datasets. Most of the currently available transcriptome assemblers have some level of built-in parallelism that takes advantage of high-performance computing clusters with thousands of computing cores. Alternatively, cloud computing[67] is an attractive framework for parallel computing, since computing resources can be rented as a service on an as-needed basis. A cloud-based genome assembler has already been

developed[68], and hopefully cloud-based transcriptome assemblers will emerge as scalable solutions to the large transcriptome assembly problem.

Meanwhile, experimental RNA-Seq and sequencing protocols are constantly improving and can greatly reduce the informatics challenges. For example, RNA-Seq reads from third generation sequencers, like PacBio[69], are much longer. PacBio sequencers are capable of sequencing a single transcript to full-length in a single read. If this technology reaches comparable throughput to the second generation technologies, the need for transcriptome assembly will likely be eliminated. Hopefully, the future of transcriptome assembly will be “no assembly required”.

Table 1 1 | **A list of splice-aware short-read aligners**

Many splice-aware aligners have been developed for aligning transcripts to a genomic reference. The advantages of "seed and extend" algorithms are that they can align sequences with more errors, that can be missed by BWT aligners. BWT aligners, on the other hand, are able to align sequences quickly, and using less memory.

Aligner	Paired end?	Algorithm	Finds non-canonical splice sites?	Output format	Availability
Blat[70]	No	seed and extend	yes	PSL	http://users.soe.ucsc.edu/~kent/src/
TopHat[45]	Yes	BWT	yes	BAM	http://tophat.cbcb.umd.edu/
GSNAP[48]	Yes	seed and extend	no	SAM	http://research-pub.gene.com/gmap/
SpliceMap[46]	Yes	BWT	no	SAM	http://www.stanford.edu/group/wonglab/SpliceMap/
MapSplice[47]	Yes	BWT	yes	SAM	http://www.netlab.uky.edu/p/bioinfo/MapSplice

Box 2 | **Proposed quality metrics for assessing transcriptome assemblies**

We suggest five metrics for evaluating the quality of an assembled transcriptome, given a set of reference transcripts derived from the same transcriptome, or a reference genome:

1. The accuracy metric is defined as the percentage of the correctly assembled bases estimated using the reference transcripts (*N*). If reference transcripts are not available, then the reference genome can be used as an alternative. Accuracy can be formally written as:

$$Accuracy = 100 \times \frac{\sum_{i=1}^{M} A_i}{\sum_{i=1}^{M} L_i}$$

where L_i is the length of alignment between a reference transcript and an assembled transcript T_i, A_i is the correct bases in transcript T_i, and M represents the number of best alignments between assembled transcripts and reference.

2. The completeness metric is defined as the percentage of reference transcripts covered by all the assembled transcripts, and is written as:

$$Completeness = 100 \times \frac{\sum_{i=1}^{N} I(C_i \geq \delta)}{N}$$

where the indicator function, I, represents whether(1) or not(0) Ci (the percentage of a reference transcript, i, that is covered by assembled transcripts) is greater than some arbitrary threshold δ, for example 80%.

3. The contiguity metric is defined as the percentage of reference transcripts covered by a single, longest assembled transcripts, and is similarly written as:

$$Contiguity = 100 \times \frac{\sum_{i=1}^{N} I(C_i \geq \delta)}{N}$$

where the indicator function, I, represents whether(1) or not(0) Ci (the percentage of a reference transcript, i, that is covered by a single, longest assembled transcript) is greater than some arbitrary threshold δ, for example 80%.

4. The percentage of chimeras among all the assembled transcripts. A chimeric transcript is one that contains non-repetitive parts from two or more different reference genes. They can arise from biological (gene fusions, transplicing), experimental (inter-molecular ligation) or informatics (misassemblies) sources. The first two should be constant for a given sample, so this metric is a direct measure of the misassembled transcripts, when comparing assemblies.

5. The percentage of transcript variants assembled. This can be calculated by the average of the percentage of assembled variants within the reference set as:

$$Variants = 100 \times \frac{\sum_{i=1}^{N} \frac{C_i - E_i}{V_i}}{N}$$

where Ci and Ei are the number of correctly or incorrectly assembled variants for

reference gene i, respectively; and Vi is the total number of variants for i.

Figure 1 | **The data generation and analysis steps of an RNA-Seq experiment. a** | Data generation. To generate an RNA-Seq dataset, RNA (light blue) is first extracted and fragmented into short fragments. The RNA fragments are then reverse transcribed into cDNA (yellow), and sequencing adaptors (blue) are ligated, followed by fragment size selection Finally, the ends of the cDNAs are sequenced using NGS technologies to produce many short reads (dark red). **b** | Data analysis. After sequencing, reads are pre-processed by removing sequencing errors (red X's) and low-quality reads. Artifacts, such as adapter sequence (blue), contaminant DNA (green), and PCR duplicates should also be removed to improve the assembly and reduce the amount of computing resources needed. The pre-processed reads are then assembled into transcripts (orange) and polished by post-assembly processes. The expression level of each transcript is then estimated for further downstream analysis.

Figure 2 | **Overview of the next-generation transcriptome assembly strategies. a** | The reference-based strategy using a reference genome (blue). Reads (red) are first splice-aligned to a reference genome. Then, a connectivity or splice graph is constructed to represent all possible isoforms at a locus. Finally, the graph is traversed to assemble the most likely isoforms (orange). **b** | The *de novo* assembly strategy without a reference genome. A De Bruijn graph is constructed from all overlapping k-mers within a read. Here, a simple example using 4-mers is shown to illustrate two possible paths through a De Bruijn graph. The De Bruijn graph is then trimmed for errors and isoforms (orange) are assembled by traversing the graph. **c** | Alternative approaches for hybrid transcriptome assembly. The left choice depicts the align-then-assemble strategy in which reference-based assembly is followed by de novo assembly of reads which failed to align to the genome. The right choice depicts the assemble-then-align strategy in which the reads are first *de novo* assembled and then scaffolded and extended using a reference genome. RNA-Seq reads are shown in red, while assembled transcripts are shown in orange.

Table 1 | **A comparison of the features of existing software for transcriptome assembly**. MP: Multiple Processor support (assembler takes advantage of many cores from a single computer). MPI, Message Passing Interface support (assembler runs in parallel on multiple computers within a cluster).

Please remember to include the following items with your revision. Examples are given in the accompanying letter:

- An autobiography: Please provide a brief (approx 100 words) potted history of the research career of each author, including the interests of your lab. This will be linked to the authors' affiliation in the online version.
- Online summary: In contrast to the preface, which is intended to entice the passing reader, this summary will provide a bullet-pointed reminder of what the review covers, in about 10 points. We hope that our readers will come back to these to jog their memories some time after they have read the reviews.

- Reference comments: Please provide one sentence to describe the importance of important papers cited (around 10 will do).
- Please provide any copyright information that is associated with the diagrams we have reproduced. We will take care of obtaining copyright clearance, but in order to do so we need the full citation of the work in which the diagrams were originally published.

Glossary terms

BWT The Burrows-Wheeler transform algorithm. Introduced in 1994 by Michael Burrows and David Wheeler for data compression, it is widely used by many short read aligners.

Cloud computing The abstraction of the underlying hardware architectures (for example, servers, storage and networking) that enable convenient, on-demand network access to a shared pool of computing resources that can be readily provisioned and released.

De Bruijn graph A graph with vertices represented as a sequence of symbols (e.g., A,C,T,G) of length *k*. A directed edge connects two vertices if removing the first symbol from one vertex and then appending another symbol creates the sequence from the second vertex.

Greedily assembling An assembly algorithm in which choices are made based upon a series of locally optimal solutions. This approach may eventually lead to a sub-optimal global solution.

K-mer frequencies The number of times each k-mer (substring of length k) appears in a set of DNA sequences.

Low-complexity reads Short DNA sequences composed of stretches of homopolymer nucleotides or simple sequence repeats. Some are artifacts generated from NGS platforms. Low-complexity reads often cause misassemblies.

Normalization techniques Methods that can increase the representation of rare transcripts by reducing the highly represented ones, in an effort to equalize the representions of all RNA species.

Paired-end protocol A library construction and sequencing strategy that allows the sequencing of both ends of a DNA fragment, to produce “paired-end” reads. Overlapping paired-ends can be joined to produce a longer sequence read. Pairs of longer DNA fragments (several kbs) are usually termed “mate-pairs” and are very useful in assembly in that they provide physical connectivity between contigs.

RNA-Seq A technology that uses NGS technologies to sequence the transcriptome, to determine the identity of each transcript and its relative abundance.

Traversing A method for visiting all nodes in a graph.

References

1. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
3. Marguerat, S. & Bahler, J. RNA-seq: from technology to biology. *Cell Mol Life Sci* **67**, 569-79 (2010).
4. Wilhelm, B.T. & Landry, J.R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249-57 (2009).
5. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).
6. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
7. Simpson, J.T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-23 (2009).
8. Butler, J. et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**, 810-20 (2008).
9. Paszkiewicz, K. & Studholme, D.J. De novo assembly of short sequence reads. *Brief Bioinform* **11**, 457-72 (2010).
10. Miller, J.R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-27 (2010).
11. Makalowska, I., Lin, C.F. & Makalowski, W. Overlapping genes in vertebrate genomes. *Comput Biol Chem* **29**, 1-12 (2005).
12. Normark, S. et al. Overlapping genes. *Annu Rev Genet* **17**, 499-525 (1983).
13. Johnson, Z.I. & Chisholm, S.W. Properties of overlapping genes are conserved across microbial genomes. *Genome Res* **14**, 2268-72 (2004).
14. Fukuda, Y., Washio, T. & Tomita, M. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. *Nucleic Acids Res* **27**, 1847-53 (1999).
15. 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**, 1009-15 (2010).
16. Martin, J. et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).
17. 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**, 503-10 (2010).
18. Denoeud, F. et al. Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**, R175 (2008).
19. Robertson, G. et al. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-12 (2010).
20. Zerbino, D. & Schulz, M. Oases: a transcriptome assembler for very short reads. in *http://www.ebi.ac.uk/~zerbino/oases/* (2010).
21. Surget-Groba, Y. & Montoya-Burgos, J.I. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res* **20**, 1432-40 (2010).
22. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
23. Birol, I. et al. De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872-7 (2009).
24. Crawford, J.E. et al. De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. *PLoS One* **5**, e14202 (2010).
25. Garg, R., Patel, R.K., Tyagi, A.K. & Jain, M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res* **18**, 53-63 (2011).
26. Adamidi, C. et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res.* 2 May 2011 (10.1101/gr.113779.110).
27. Yassour, M. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A* **106**, 3264-9 (2009).
28. Sam, L.T. et al. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS One* **6**, e17305 (2011).
29. He, S. et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* **7**, 807-12 (2010).
30. Chen, Z. & Duan, X. Ribosomal RNA Depletion for Massively Parallel Bacterial RNA-Sequencing Applications. *Methods Mol Biol* **733**, 93-103 (2011).
31. Christodoulou, D.C., Gorham, J.M., Herman, D.S. & Seidman, J.G. Construction of Normalized RNA-seq Libraries for Next-Generation Sequencing Using the Crab Duplex-Specific Nuclease. *Curr Protoc Mol Biol* **Chapter 4**, Unit4 12 (2011).

32. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18 (2011).
33. Mamanova, L. et al. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* **7**, 130-2 (2010).
34. Levin, J.Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-15 (2010).
35. Chen, S. et al. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS One* **5**, e15633 (2010).
36. Schwartz, T.S. et al. A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics* **11**, 694 (2010).
37. Passalacqua, K.D. et al. Structure and complexity of a bacterial transcriptome. *J Bacteriol* **191**, 3203-11 (2009).
38. Dalloul, R.A. et al. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol* **8,** e1000475 (2010).
39. Jackman, S.D. & Birol, I. Assembling genomes using short-read sequencing technology. *Genome Biol* **11**, 202 (2010).
40. Rodrigue, S. et al. Unlocking short read sequencing for metagenomics. *PLoS One* **5**, e11840 (2010).
41. Shi, H., Schmidt, B., Liu, W. & Muller-Wittig, W. A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. *J Comput Biol* **17**, 603-15 (2010).
42. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* **11**, R116 (2010).
43. Falgueras, J. et al. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* **11**, 38 (2010).
44. Lassmann, T., Hayashizaki, Y. & Daub, C.O. TagDust--a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839-40 (2009).
45. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
46. 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**, 4570-8 (2010).
47. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178 (2010).
48. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-81 (2010).
49. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8 (2008).
50. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-9 (2008).
51. Perkins, T.T. et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet* **5**, e1000569 (2009).
52. Ozsolak, F. et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018-29 (2010).
53. Salzberg, S.L. & Yorke, J.A. Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320-1 (2005).
54. Ward, J. & Weber, C. Comparative RNA-Seq For The Investigation Of Gene Expression In Phytophthora-Challenged Red Raspberry. in *Plant & Animal Genomes XIX Conference* (Town & Country Convention Center, 2011). Is this a conference abstract? If so then it should not be included in the reference list. It can be cited in the main text as a personal communication.
55. Kinsella, M., Harismendy, O., Nakano, M., Frazer, K.A. & Bafna, V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* **27**, 1068-75 (2011).
56. McPherson, A. et al. deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLoS Comput Biol* **7**, e1001138 (2011).
57. Tomlins, S.A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-9 (2007).
58. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**, 9748-53 (2001).
59. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 15 May 2011 (doi:10.1038/nbt.1883).
60. Burset, M., Seledtsov, I.A. & Solovyev, V.V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* **28**, 4364-75 (2000).
61. Jager, M. et al. Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing. *BMC Genomics* **12**, 158 (2011).
62. Cocquet, J., Chong, A., Zhang, G. & Veitia, R.A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127-31 (2006).

63. Haas, B.J. & Zody, M.C. Advancing RNA-Seq analysis. *Nat Biotechnol* **28**, 421-3 (2010).
64. Greninger, A.L. et al. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One* **5**, e13381 (2010).
65. Meader, S., Hillier, L.W., Locke, D., Ponting, C.P. & Lunter, G. Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Res* **20**, 675-84 (2010).
66. Schaefer, B.C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal Biochem* **227**, 255-73 (1995).
67. Taylor, R.C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics* **11 Suppl 12**, S1 (2010).
68. Schatz, M.C., Sommer, D.D., Kelley, D.R. & Pop, M. De Novo Assembly of Large Genomes using Cloud Computing. in *CSHL Biology of Genomes conference* (2010).
69. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-8 (2009).
70. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).