Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.

Sequencing of RNA has long been recognized as an efficient method for gene discovery[1] and remains the gold standard for annotation of both coding and noncoding genes[2]. Compared with earlier methods, massively parallel sequencing of RNA (RNA-Seq)[3] has vastly increased the throughput of RNA sequencing and allowed global measurement of transcript abundance. Two reports in this issue introduce approaches for RNA-Seq analysis that capture genome-wide transcription and splicing in unprecedented detail. Trapnell *et al.*[4] describe a software package, Cufflinks, for simultaneous discovery of transcripts and quantification of expression levels and apply it to study gene expression and splicing during the differentiation of mouse myoblast cells. Taking a similar approach, Guttman *et al.*[5] use software called Scripture to reannotate the transcriptomes of three mouse cell lines, defining complete gene models for hundreds of new large intergenic noncoding RNAs (lincRNAs)[6].

Although transcript sequencing has been possible for nearly 20 years, until recently it required the construction of clone libraries. Projects to determine full-length gene structures for human, mouse and other important models have taken years to complete[7]. With new sequencing technologies, no cloning is needed, allowing direct sequencing of cDNA fragments. In a matter of days and at a small fraction of the cost of earlier projects, one can achieve reasonably complete coverage of a transcriptome[8]. But this approach has been hindered by a substantial challenge: without cloning, one cannot know a priori which reads came from which transcripts. Recent studies analyzed gene expression and alternative splicing by mapping short RNA-Seq reads to previously known or predicted transcripts[9,10]. Although highly informative, such studies are inherently limited to known genes and to alternative splicing across previously identified splice junctions. To fully leverage RNA-Seq data for biological discovery, one should be able to reconstruct transcripts and accurately measure their relative abundance without reference to an annotated genome.

Previous efforts to reconstruct transcripts from short RNA-Seq reads have followed two general strategies (**Fig. 1**). The first, a *de novo* assembly approach implemented in the ABySS software[11], reduces the annotation problem to that of aligning full-length cDNAs, which is well handled by several algorithms. This method is also applicable to the discovery of transcripts that are missing or incomplete in the reference genome and to RNA-Seq data from organisms lacking a genome reference.

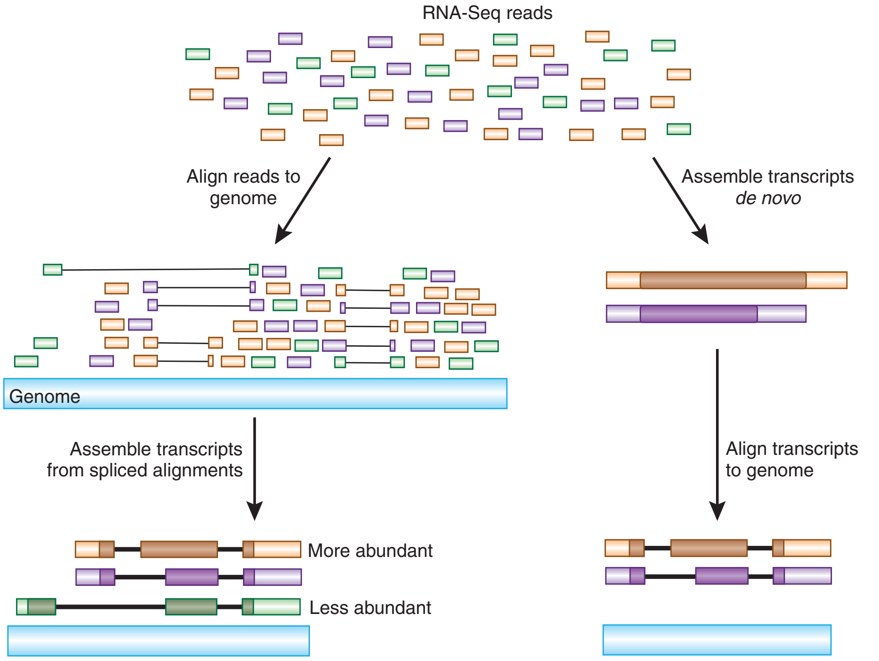
RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Figure 1 Strategies for reconstructing transcripts from RNA-Seq reads. The ‘align-then-assemble’ approach (left) taken by Trapnell *et al.*[4] and Guttman *et al.*[5] first aligns short RNA-Seq reads to the genome, accounting for possible splicing events, and then reconstructs transcripts from the spliced alignments. The ‘assemble-then-align’ approach (right) first assembles transcript sequences *de novo*—that is, directly from the RNA-Seq reads. These transcripts are then splice-aligned to the genome to delineate intron and exon structures and variations between alternatively spliced transcripts. As *de novo* assembly is likely to work only for the most abundant transcripts, the align-then-assemble method should be more sensitive, although this warrants further investigation. RNA-Seq reads are colored according to the transcript isoform from which they were derived. Protein-coding regions of reconstructed transcript isoforms are depicted in dark colors.

Brian J. Haas and Michael C. Zody are at the Broad Institute, Cambridge, Massachusetts, USA. e-mail: bhaas@broadinstitute.org or mczody@broadinstitute.org

However, assembly of short reads is itself difficult, and only the most abundant transcripts are likely to be fully assembled.

The second strategy involves splice-aware alignment of individual short RNA-Seq reads to the genome followed by transcript reconstruction[12,13]. This is the approach taken by Trapnell *et al.*[4] in Cufflinks and by Guttman *et al.*[5] in Scripture. Both programs use the TopHat aligner[14] to generate spliced alignments to the genome. Whereas earlier RNA-Seq experiments produced 25–32-base reads, the 75-base or longer reads now available can be aligned in segments, allowing reads whose ends are anchored in different exons to define splice sites without relying on prior annotations. Both programs then build directed graphs and traverse the graphs to identify distinct transcripts, using paired end information to link sparsely covered transcripts and filter out unlikely isoforms.

There are also notable differences in the details of the algorithms. For example, Cufflinks uses a rigorous mathematical model to identify the complete set of alternatively regulated transcripts at each locus and to assign coverage to each transcript; Scripture employs a statistical segmentation model to distinguish expressed loci and filter out experimental noise. More extensive testing of Cufflinks, Scripture and *de novo* assembly methods such as ABySS will be required to determine whether some methods perform better in certain applications.

Strikingly, despite the extensive prior annotation of the mouse genome (which was based on millions of expressed sequence tags (ESTs) and thousands of full-length cDNAs), both studies identify thousands of novel transcripts, including novel isoforms of known genes and completely novel coding and noncoding genes.

Trapnell *et al.*[4] discover 3,724 high-confidence isoforms of known genes that are absent from existing automated and manually curated gene sets. They also demonstrate that independently determining the expression of each isoform with high accuracy is an important prerequisite for subsequent analysis. It has been shown that RNA-Seq can accurately detect gene expression levels over a wide dynamic range[3,9], but previous experiments have relied on known or predicted isoforms. By reconstructing all isoforms directly from the RNA-Seq read alignments and accurately classifying individual paired read fragments according to their isoform of origin, Trapnell *et al.*[4] are able to measure the expression levels of individual isoforms within a single gene with high accuracy. They further show that correct assignment of RNA-Seq fragments to novel isoforms can substantially affect the computed expression levels of known isoforms of the same gene.

Measuring the expression of individual isoforms makes it possible to study regulatory changes in greater detail than was previously feasible. Regulatory changes may be transcriptional, indicated by isoforms with different transcription start sites, or post-transcriptional, indicated by isoforms with the same start site that show alternative internal splicing. Trapnell *et al.*[4] identify large numbers of genes that undergo significant changes of both types over the time course of their experiment. The ability to examine regulation of expression at such a fine scale over an entire genome allows important new insights into genome function. For example, data at this level of detail could vastly improve our ability to model regulatory networks or to infer regulatory motifs based on correlation of the expression and splicing of individual isoforms rather than genes.

Guttman *et al.*[5] also identify a number of novel splice isoforms but focus their analysis on novel transcripts, particularly lincRNAs. Previous work using ChIP-Seq and whole-genome tiling arrays[6] identified loci that encode lincRNAs but lacked the resolution to produce accurate models. With the Scripture predictions, Guttman *et al.*[5] were able to construct gene models for 609 known loci and identify and generate structures for over 1,000 novel lincRNAs. They also identified 469 antisense transcripts of protein-coding genes.

Determining gene models for these noncoding RNAs opens the door to functional analysis. For example, Guttman *et al.*[5] examined the conservation levels of transcripts. Consistent with previous observations, the lincRNAs were typically more conserved than intronic sequences but less conserved than protein-coding genes. Conversely, the antisense transcripts showed no conservation outside of that resulting from overlap with coding exons, suggesting that these two classes of transcripts have very different functions and constraints. The RNA-Seq data also revealed expression patterns of noncoding transcripts and showed that the lincRNAs are not only less abundant than protein-coding genes but also less broadly expressed, with a greater fraction showing tissue specificity compared with protein-coding genes in the same cell lines. More generally, the determination of precise gene models and expression patterns for noncoding RNAs will facilitate their inclusion in regulatory network and gene interaction models, an important step toward understanding their functions.

The number of novel transcripts discovered by Trapnell *et al.*[4] and Guttman *et al.*[5] may leave us wondering: why do existing annotations fall so short? Known isoforms account for almost 80% of the RNA-Seq fragments in the Trapnell *et al.*[4] data, indicating that these are highly expressed genes that were easily identified from clone-based cDNA sequencing (Guttman *et al.*[5] do not provide an identical breakdown, but the high level of coverage shown for the most abundant transcripts suggests similar numbers). Another 11% of fragments map to novel isoforms of known genes, 62% of which are supported by previous EST or mRNA sequence but are not annotated as distinct transcripts. These less abundant isoforms may have been sampled sparsely in previous studies, or may not have been fully sequenced or annotated because of similarity to known transcripts at the same locus. Similarly, 43% of the novel lincRNAs found by Guttman *et al.*[5] were found in a previous mouse cDNA project[15]. Given the apparent tissue-specificity of lincRNAs, the remainder may not have been seen previously due to relatively limited tissue sampling. The emphasis of earlier large-scale transcript sequencing projects on protein-coding genes also explains the absence of annotation for most of these features, even where evidence has existed. Clear definition of these novel coding and noncoding transcripts is made possible by the unbiased nature of RNA-Seq combined with the unbiased discovery methods of Trapnell *et al.*[4] and Guttman *et al.*[5]

Cufflinks, Scripture and similar tools provide a great opportunity to improve the annotation of both well-studied genomes and poorly annotated genomes that have not received extensive traditional EST and full-length mRNA sequencing. However, there are still substantial challenges in using RNA-Seq for annotation. A large number of transcripts identified by Cufflinks and Scripture were consistent with known isoforms but incomplete due to lack of coverage. Just as RNA-Seq allows reconstruction of transcripts that are only weakly supported by EST data, many less highly or less broadly expressed transcripts are only weakly or incompletely supported by current RNA-Seq.

As technology allows increasingly deeper sequencing of the transcriptome, it will be possible to identify more transcripts with higher confidence. However, more sophisticated methods for separating functional low-abundance transcripts from transcriptional noise and process artifacts will be needed. Also, although Cufflinks and Scripture will be useful tools for annotating new genomes, different genomes may pose different algorithmic challenges owing to variation in characteristics such as gene density, intron content and length, and prevalence of alternative splicing. It remains to be seen how well Cufflinks and Scripture will perform on genomes that are very different from mouse.

Massively parallel sequencing technology has already revolutionized the way we study genomes, and the capacity and quality of sequencing data continue to improve at a rapid pace. Trapnell *et al.*[4] and Guttman *et al.*[5] have demonstrated the power of RNA-Seq combined with novel transcript discovery to greatly improve the annotation of an already well-studied genome and to add substantially to our understanding of transcriptional and post-transcriptional regulation. By making their software available, they provide powerful tools that will facilitate future RNA-Seq studies.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Adams, M.D. *et al. Science* **252**, 1651–1656 (1991).
2. Haas, B. J. *et al. Genome Biol* **3**, RESEARCH0029 (2002).
3. Nagalakshmi, U. *et al. Science* **320**, 1344–1349 (2008).
4. Trapnell, C. *et al. Nat. Biotechnol.* **28**, 503–519 (2010).
5. Guttman, M. *et al. Nat. Biotechnol.* **28**, 511–515 (2010).
6. Guttman, M. *et al. Nature* **458**, 223–227 (2009).
7. Temple, G. *et al. Genome Res.* **19**, 2324–2333 (2009).
8. Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).
9. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
10. Wang, E.T. *et al. Nature* **456**, 470–476 (2008).
11. Birol, I. *et al. Bioinformatics* **25**, 2872–2877 (2009).
12. Denoeud, F. *et al. Genome Biol.* **9**, R175 (2008).
13. Yassour, M. *et al. Proc. Natl. Acad. Sci. USA* **106**, 3264–3269 (2009).
14. Trapnell, C., Pachter, L. & Salzberg, S.L. *Bioinformatics* **25**, 1105–1111 (2009).
15. Carninci, P. *et al. Science* **309**, 1559–1563 (2005).

Haploidy with histones

Gregory P Copenhaver & Daphne Preuss

An engineered centromere-specific histone could enable homozygous diploid lines to be generated at high frequency, simplifying crop breeding.

Sexually reproducing plants carrying a set of chromosomes from each parent are the rule in nature, but, for crop breeders, haploid plants represent a more useful resource. Arising either spontaneously at very low frequencies or generated by protracted cross-breeding or tissue-culture methods, haploid plants allow fully homozygous lines to be screened for desirable traits in one generation. A recent study in *Nature* reports that haploid plants can now be rapidly produced through the introduction of a single genetic alteration. Ravi and Chan[1] show that perturbing a centromeric histone in the model plant *Arabidopsis thaliana* makes it possible to reliably create haploid plants and 'doubled haploid' progeny from those plants. If this approach can be translated to crop species, it would find immediate application in agricultural biotechnology, shortening crop breeding programs by years.

In most eukaryotic organisms, the movement of a chromosome during cell division is regulated by its centromere, which is bound by the centromeric histone H3 (CENH3), a variant of the more ubiquitous histone H3. After DNA replication, CENH3 is loaded onto the newly formed daughter strands, targeting epigenetic marks in the centromere region[2,3]. In a zygote, the centromeres of the maternal and paternal chromosomes are bound by CENH3 proteins from the maternal and paternal germ cells, respectively. Normally, these two sets of CENH3 help to move the maternal and paternal chromosomes with equal efficiency in the first few mitotic divisions that form the developing embryo. Ravi and Chan[1] show that altering CENH3 from one parent can induce targeted elimination of the chromosomes inherited from that parent (**Fig. 1**).

The authors modify CENH3 in two ways. In the first, green fluorescent protein (GFP) is fused to the N terminus of CENH3. In the second, the N-terminal tail of CENH3 is replaced with the corresponding domain from histone H3, and GFP is fused to the new tail (**Fig. 1a**). Both the H3 and CENH3 N-terminal tails are targets for multiple post-translational modifications and are thought to regulate chromatin structure. The modified CENH3s do retain some function, but their recognition of the chromosome segregation machinery is diminished. As a result, the only chromosomes in the zygote that are moved properly are those that harbor CENH3 from the wild-type parent (**Fig. 1b**).

As new histone synthesis takes place within a developing embryo, one would expect that DNA strands are loaded with a mixture of CENH3 proteins encoded by the maternal and paternal alleles. Consistent with this view, Ravi and Chan[1] find that the distinction between chromosomes having maternal or paternal CENH3 is lost after the first few divisions and the remaining divisions are able to proceed normally throughout development, resulting in a haploid plant. These haploids can produce diploid (doubled-haploid) progeny, presumably either through somatic chromosome doubling or rare non-reductional divisions during meiosis.

For nearly a century, crop breeders have recognized that haploid plants can be used to accelerate the development of new inbred lines[4]. In a typical program, genetically diverse parents are crossed to create hybrids (F$_1$), and populations of their offspring (F$_2$, F$_3$, F$_4$ and so on) are surveyed to identify desirable traits and to select individual plants for further propagation. After several generations, the traits under selection become fixed, and the inbred line is typically homozygous for chromosomal regions of interest. Incorporating doubled haploids into a breeding program has the advantage of saving considerable time by achieving homozygosity more quickly; however, this strategy requires that more lines be planted and screened in a single generation, allowing a sufficiently complete survey of genetic combinations.

Although haploids occur spontaneously in many crop species, they are extremely rare, often forming prezygotically from gametophyte cells that develop into a mature plant. Haploids can be formed at a higher (albeit still extremely low) frequency from 'inducer' lines, from gametophytes cultured *in vitro*, or from intra- or interspecies hybrids that undergo post-zygotic chromosome elimination. What is most exciting about the breeding approach described by Ravi and Chan[1] is the high frequency at which they recover haploid plants from a diploid parent (~1–10% of a normal seed set in *A. thaliana*). In addition, they show that the same scheme can be used to create diploid plants from tetraploids, which may be useful for breeding crops with complex ploidy, such as hexaploid wheat.

These results raise several questions about chromosome dynamics during cell division. What is the nature of the competition between centromeres bound to different CENH3s? The authors suggest that the modified CENH3s

Gregory P. Copenhaver is in the Department of Biology and the Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. Daphne Preuss is at Chromatin, Inc., Chicago, Illinois, USA.
e-mail: dpreuss@chromatininc.com