6-2008

Molecular Biology: Power Sequencing

Brenton R. Graveley
University of Connecticut School of Medicine and Dentistry

Published in final edited form as:

Nature. 2008 June 26; 453(7199): 1197–1198. doi:10.1038/4531197b.

Molecular Biology:

Power Sequencing

Brenton R. Graveley

Department of Genetics and Developmental Biology, University of Connecticut Stem Cell Institute, University of Connecticut Health Center, Farmington, Connecticut 06030, USA

Brenton R. Graveley: graveley@neuron.uchc.edu

Abstract

Advances in DNA-sequencing technology provide unprecedented insight into the entire collection of four genomes' transcribed sequences; they herald a new era in the study of gene regulation and genome function.

Genomes are the blueprints of life: they contain all the information necessary to build and operate their hosts. But we still have much to learn about the language of DNA to interpret the billions of the Gs, As, Ts and Cs, the DNA bases that spell out life. The information-containing portions of genomes are transcribed into two RNA classes: messenger RNAs, which are translated into proteins; and non-coding RNAs, which have regulatory, structural or mechanical roles. So studying the transcribed portion of the genome — the transcriptome — significantly aids gene identification, as well as providing insight into the inner workings of the genome and the biology of an organism. Five papers[1–5], including one on page XXX of this issue by Wilhelm *et al.*[1], describe how advances in DNA-sequencing technology can be harnessed to explore transcriptomes in remarkable detail.

The concept of sequencing large numbers of randomly selected mRNAs is not new. It forms the basis of the controversial, yet revolutionary EST method[6], which was originally used to identify genes in the reference copy of the human genome. In this technique, genes are quickly identified through sequencing small fragments of large numbers of mRNAs. Although EST sequencing remains useful, it is relatively slow, requires considerable resources and generally cannot identify mRNAs that are expressed at low levels.

DNA microarrays are also powerful tools for transcriptome analysis. Particularly informative are tiling arrays, which are dotted with DNA sequences derived from defined intervals (for example, every 35 base pairs) throughout the genome. Fluorescently labelled RNA is then allowed to bind to the arrays, and the transcribed portions of the genome are identified by determining which DNA sequences pair with the RNA. But tiling arrays also have several shortcomings. First, they can only be used for organisms with known genome sequences. Second, their limited sensitivity, specificity and dynamic range – the ratio between the smallest and largest fluorescent signal – make it difficult to identify low-abundance mRNAs and to distinguish between highly similar mRNA sequences. Finally, the number of DNA probes that fit on a microarray is limited, putting constrains on the

minimum feasible genomic distance between the probes, and so on the resolution at which a genome can be analysed.

Enter the trio of next-generation sequencing technologies — 454, Illumina (Solexa) and ABI (SOLiD) — which can generate gigabases of sequence in a single experiment[7]. They differ from traditional sequencing methods in two ways. First, rather than sequencing individual DNA clones, hundreds of thousands (454) to tens of millions (Illumina and ABI) of DNA molecules are sequenced in parallel. Second, the sequences obtained are much shorter (25–50 nucleotides for Illumina and ABI, and 200–400 nucleotides for the 454 system) than those generated by traditional sequencing (typically more than 800 nucleotides). The shorter sequences generated by these new sequencing platforms make it more difficult to unambiguously match them to the reference genome, but this is a relatively minor trade-off that is more than made up for by the massive amount of total sequence generated. These technologies have already revolutionized the study of chromatin structure, DNA-binding proteins, DNA methylation, genome organization and small RNAs[7]. But how useful would they be for studying transcriptomes was not known.

Five teams have now used a method called mRNA-Seq (Fig. 1) to sequence at varying levels of detail the transcriptomes of four organisms — the fission yeast *Saccharomyces pombe*[1], the budding yeast *Saccharomyces cerevisiae*[2], the plant *Arabidopsis thaliana*[3] and the mouse[4,5]; for the sequencing step, four of the groups used the Illumina Genome Analyzer system[1-4] and one used the ABI SOLiD system[5]. Between 30 and 125 million sequences 25 to 39-base-pairs in length were obtained in each study. The most comprehensive of these was performed byWilhelm *et al.* who generated 122 million 39-base-pair sequences for *S. pombe*[3], corresponding to nearly five gigabases of sequence or 250 equivalents of this organism's genome.

But how comprehensively do these analyses cover the known genes? In the one-billion bases of sequence obtained for *S. cerevisiae*, only some 91% of the known genes were detected. By contrast, sequencing five billion bases of the *S. pombe* transcriptome allowed Wilhelm *et al.* to identify 99.3% of known genes. So although ‘moderate’ sequencing of the transcriptome can quickly detect most genes, identification of all genes requires extraordinarily deep sequencing.

The mRNA-Seq method can also detect previously unidentified genes. In *S. pombe*, 453 new transcripts are identified, of which 427 seem to be non-coding. Similarly, in the *S. cerevisiae* transcriptome 204 previously undetected transcripts are identified. Although these numbers sound relatively small, they are noteworthy because the organisms studied were already studied extensively in the past. Undoubtedly, mRNA-Seq will also identify unknown genes in organisms that are not typically studied in the laboratory.

Genes consist of exons, which are separated by shorter sequences known as introns. Following transcription, introns are spliced out of mRNA to form mature mRNA containing only exons. One limitation that the spacing of DNA probes in tiling arrays pose is that the short introns cannot be confidently identified. mRNA-Seq, by contrast, provides unparalleled resolution as many sequence ‘reads’ contain one portion mapping to the end of

one exon and the remainder of the read maps to the beginning of the exon on the other end of the intron. These reads not only identify introns, but also precisely delineate the ends of exons and introns. Analysing such intron-spanning reads confirmed 78% and 93% of known introns in *S. cerevisiae* and *S. pombe*, respectively. What's more, Wilhelm and colleagues discovered[1] 20 new introns in *S. pombe*.

The dynamic range, sensitivity and specificity of mRNA-Seq make it also ideal for quantitatively analysing different aspects of gene regulation including differences in transcript abundance. For example, a comparison of the transcriptome of normal *A. thaliana* with those of three strains of this plant defective in various aspects of DNA methylation — a modification that regulates gene expression — reveals[3] scores of genes, some of them new, that are differentially expressed when DNA methylation pathways are perturbed. For example, consistent with the known role of DNA methylation in gene silencing, the number of reads that mapped to transposons and pseudogenes was significantly higher in the strain defective in DNA methylation than in the wild-type strain.

The efficiency of intron removal is another aspect of gene regulation that can be monitored by comparing the number of reads that span an intron with those that span the corresponding exon–intron junctions (Fig. 2). Wilhelm and colleagues compared[1] *S. pombe* transcriptomes from proliferating cells with those of cells undergoing different stages of meiotic cell division. They identify 314 introns from 254 genes that are spliced more efficiently during meiosis than in rapidly proliferating cells; only 12 such meiotically spliced genes were previously known. Further analysis of this dataset also reveals a striking correlation between transcription levels and splicing efficiency — the higher the level of transcripts, the more efficiently they are processed to mature mRNAs. Moreover, in higher eukaryotes such as the mouse, exons can be spliced together in different patterns to generate multiple mRNAs from a single gene – a process called alternative splicing. The intron-spanning reads obtained by mRNA-Seq can also be used to identify cases of alternative splicing, and to quantitate changes in alternative splicing that occur in different samples[4,5]. For example, by comparing the mRNA-Seq transcriptomes obtained from mouse brain and muscle, Mortazavi et al.[4] was able to clearly identify an exon in the *Mef2d* gene that is spliced specifically in muscle.

For transcriptome mining by mRNA-Seq, this is just the beginning of things to come. Technological improvements such as longer reads, paired-end reads (the ability to obtain sequence from both ends each molecule and the distance between those sequences)[8], enrichment for sequences of interest[9], DNA-strand-specific sequencing of the mRNA transcripts[1], and methods to sequence all RNAs[1] and not just mRNA, will all make mRNA-Seq even more powerful. Algorithms that can accurately assemble short-sequence reads into longer stretches[10] will further allow sequencing of the transcriptome of organisms for which a reference genome is not available. Together, these advances will provide even greater insight into the transcriptional landscapes, regulation of gene expression and alternative splicing. Most importantly, next-generation sequencing has the potential to turn individual laboratories into small genome centres and to allow an individual scientist to determine the entire transcriptome of any source (any organism, tumour samples, tissues from patients with neurodegenerative disorders and so on) in a matter of days, and for only a few thousand

US dollars. Indeed, this technology will have a long-lasting impact on the way and speed with which we do science.

References

1. Wilhelm BT, et al. Nature. 2008; 453:1239–1243. [PubMed: 18488015]
2. Nagalakshmi U, et al. Science. 200810.1126/science.1158441
3. Lister R, et al. Cell. 2008; 133:523–536. [PubMed: 18423832]
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Nature Methods. 200810.1038/nmeth.1226
5. Cloonan N, et al. Nature Methods. 200810.1038/nmeth.1223
6. Adams MD, et al. Science. 1991; 252:1651–1656. [PubMed: 2047873]
7. Wold B, Myers RM. Nature Methods. 2008; 5:19–21. [PubMed: 18165803]
8. Campbell PJ, et al. Nature Genet. 2008; 40:722–729. [PubMed: 18438408]
9. Hodges E, et al. Nature Genet. 2007; 39:1522–1527. [PubMed: 17982454]
10. Zerbino DR, Birney E. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

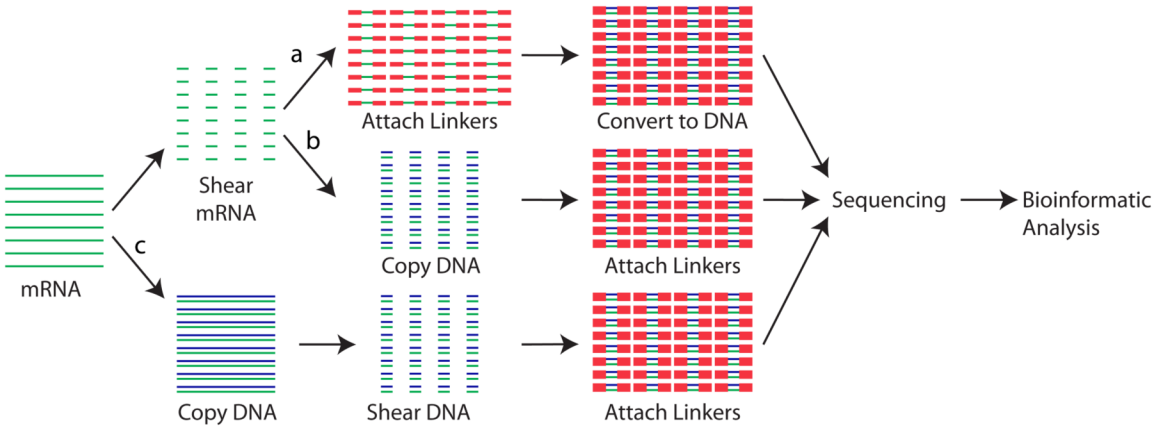

Figure 1. mRNA-Seq

In this technique, which was used for analysis of transcriptomes of five organisms[1–5], the isolated mRNA is analysed by one of three procedures. **a**, In the first procedure, mRNAs are randomly sheared, linker molecules are attached to their ends, and they are then converted to DNA. **b**, Alternatively, after shearing, the mRNA fragments are converted to DNA, followed by the attachment of linker molecules. **c**, In a third procedure, mRNAs are first copied into DNA sequences, which are then sheared and attached to linkers. In all three cases, the resulting DNA are analysed by next-generation sequencing technology and the data are compared with the reference genome for that particular organism using bioinformatics, to determine the genomic regions from which the sequences were derived.

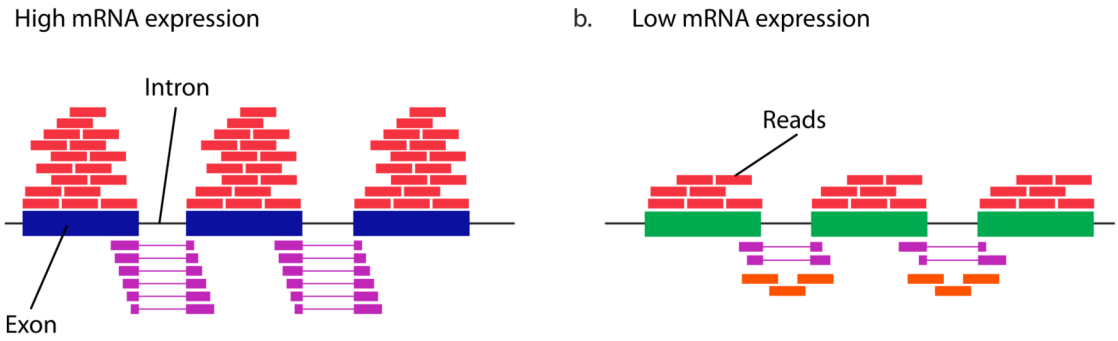

Figure 2. Determining mRNA expression levels using mRNA-Seq

a, For a gene expressed highly, several sequence reads (orange) map to each of its exons and some reads (pink) to the two exons spanning an intron. No or few reads map to introns, suggesting that intron removal in this case is very efficient. **b**, But when a gene is expressed at low levels, fewer sequence reads map to the exons or to two exons spanning an intron, whereas some reads map to the introns (brown). That almost equal number of reads map to regions within the intron and to those spanning it suggests that splicing of these gene transcripts is inefficient.