The beginning of the end for microarrays?

Jay Shendure

Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

For over a decade, DNA microarrays have provided a powerful approach to achieve parallel interrogation of biological systems at a genomic scale. But two new reports in this issue of *Nature Methods*[1,2] demonstrate that massively parallel DNA sequencing may be on its way to supplanting microarrays as the technology of choice for quantifying and annotating transcriptomes.

Key applications of microarrays have included transcriptome analysis, profiling of protein-DNA interactions, and characterization of both small-scale (for example, single-nucleotide polymorphism) and large-scale (for example, copy-number variation) genetic variation. From the molecular profiling of tumors to genome-wide association studies, microarrays have made their impact scientifically but also culturally, as a much increased fraction of the community is now adept at collecting and analyzing large-scale datasets.

From a technical perspective, the fundamental reliance of microarrays on nucleic-acid hybridization results in several inherent limitations: knowledge of the sequences being interrogated is a prerequisite for array design; analysis of highly related sequences is problematic because of cross-hybridization; and the analog nature of the signal makes it difficult to confidently detect and quantify low-abundance species. Additionally, microarray users have encountered major challenges with respect to the reproducibility of results between laboratories and across platforms.

Since 2004, massively parallel DNA sequencing technologies have exploded onto the scene, offering dramatically lower per-base costs than had previously been possible with electrophoretic sequencing[3]. The two papers in this issue of *Nature Methods*[1,2] describe the application of next-generation sequencing to characterize several mouse poly(A)$^+$ transcriptomes with unprecedented depth and resolution: Sean Grimmond and colleagues apply the Applied Biosystems' SOLiD platform to embryonic stem cells before and after differentiation, and Barbara Wold and colleagues apply the Illumina GA (Solexa) platform to the transcriptomes of mouse brain, liver and muscle.

The basic approach of shotgun transcriptome sequencing with short-read technology has been widely dubbed 'RNA-Seq', and other groups recently reported analogous results in plants[4] and yeast[5,6]. Sequencing-based characterization of a transcriptome is appealing because it effectively surmounts the limitations of microarrays listed above. Prior sequence knowledge is not required (though it still helps); paralogous sequences can be distinguished; and quantitation is 'digital' rather than 'analog', with ensuing benefits for dynamic range and sample comparison. Although the experiment has not yet been done, it is likely that the digital nature of the quantitation will also lead to superior inter-platform reproducibility.

Of course, transcriptome sequencing by itself is nothing new. Sequencing of expressed sequence tags (ESTs)[7] provided an early means of discovering coding sequences in the absence of a reference genome and subsequently for annotation of transcriptional units. The high cost of deep EST sequencing motivated the development of serial analysis of gene expression (SAGE)[8], which lowered costs by minimizing the amount of information collected per transcript. Even with SAGE, however, the cost of transcriptome analysis with conventional sequencing remains high relative to that of microarray analysis. The introduction of next-generation sequencing technology into this area represents a major leap toward a leveling of the playing field. For example, tens of millions of independently derived sequencing tags can now be obtained at a cost similar to what tens of thousands used to cost.

The RNA-Seq approach also brings a qualitative and quantitative improvement to transcriptome analysis. For example, by taking a shotgun approach (rather than restriction digestion of transcript-identifying tags, as with SAGE), the groups of Grimmond and Wold can discover new alternative splice junctions and transcriptional units simultaneously with gene expression measurements. The Grimmond group additionally demonstrates that transcribed single-nucleotide polymorphisms can be recovered from RNA-Seq data[1]. With regard to quantitation, in a key experiment in which RNA standards are spiked in at known concentrations, the Wold group demonstrates quantitative linearity over a broad dynamic range (five orders of magnitude)[2].

How do we compare the various reports on shotgun transcriptome sequencing to one another? The sequencing platforms of Applied Biosystems and Illumina are conceptually related, and currently offer approximately similar read lengths, error rates and per-tag costs. The more pertinent distinctions are in the shotgun library construction protocols and in the data analysis. For example, protocols used by Grimmond and colleagues[1] and in the previous report in plants[4] retain information about transcript directionality, useful for annotation and

Jay Shendure is in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.
e-mail: shendure@u.washington.edu

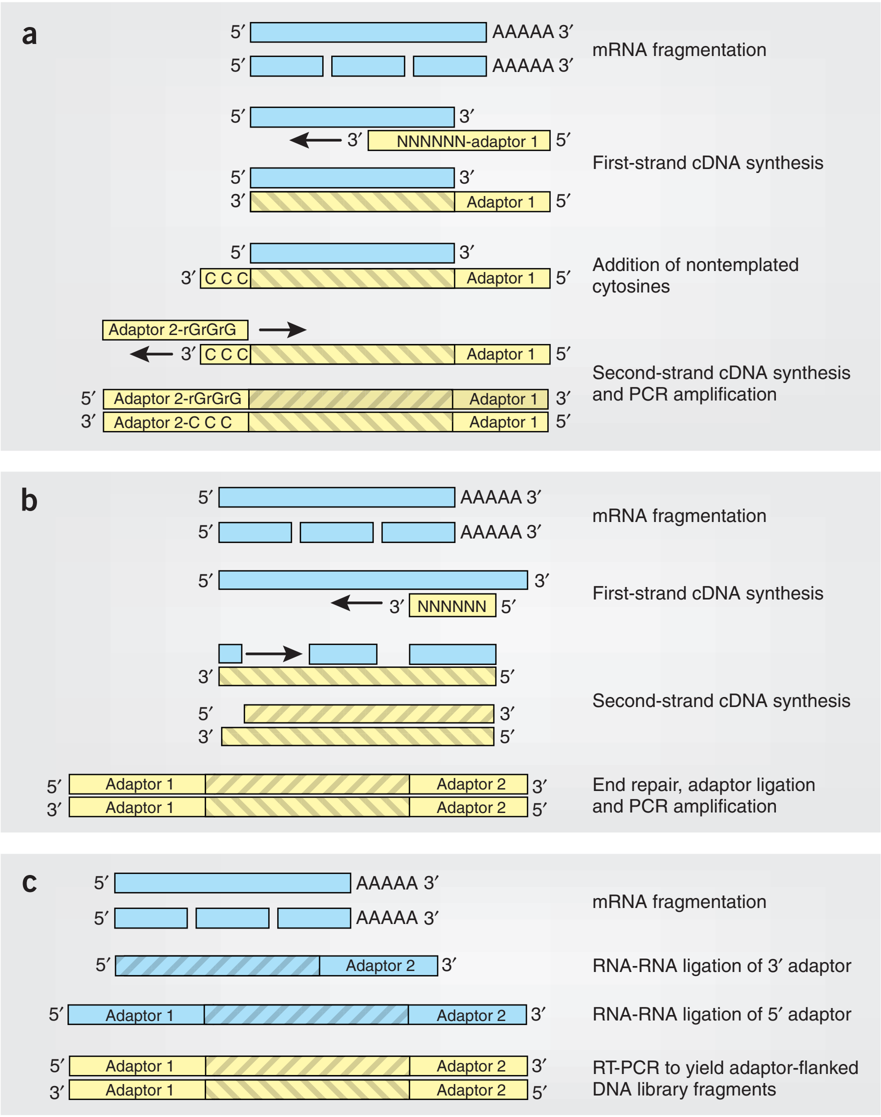
a
5′ AAAAA 3′
5′ AAAAA 3′
mRNA fragmentation
5′ 3′
3′ NNNNNN-adaptor 1 5′
5′ 3′
3′ Adaptor 1 5′
First-strand cDNA synthesis
5′ 3′
3′ C C C Adaptor 1 5′
Addition of nontemplated cytosines
Adaptor 2-rGrGrG
3′ C C C Adaptor 1 5′
5′ Adaptor 2-rGrGrG Adaptor 1 3′
3′ Adaptor 2-C C C Adaptor 1 5′
Second-strand cDNA synthesis and PCR amplification
b
5′ AAAAA 3′
5′ AAAAA 3′
mRNA fragmentation
5′ 3′
3′ NNNNNN 5′
First-strand cDNA synthesis
3′ 5′
5′ 3′
3′ 5′
Second-strand cDNA synthesis
5′ Adaptor 1 Adaptor 2 3′
3′ Adaptor 1 Adaptor 2 5′
End repair, adaptor ligation and PCR amplification
c
5′ AAAAA 3′
5′ AAAAA 3′
mRNA fragmentation
5′ Adaptor 2 3′
RNA-RNA ligation of 3′ adaptor
5′ Adaptor 1 Adaptor 2 3′
RNA-RNA ligation of 5′ adaptor
5′ Adaptor 1 Adaptor 2 3′
3′ Adaptor 1 Adaptor 2 5′
RT-PCR to yield adaptor-flanked DNA library fragments

Figure 1 | One of the differences between RNA-Seq approaches is the protocol used to construct adaptor-flanked sequencing libraries from poly(A)$^+$-enriched RNA. (**a**) Grimmond and colleagues[1] perform first-strand synthesis on fragmented mRNA by priming with randomized bases linked to one adaptor, such that directional information is captured. The second adaptor is added by template-switching. (**b**) Wold and colleagues[2] perform first-strand synthesis on fragmented mRNA by priming with randomized hexamers. Conventional second-strand synthesis is followed by end repair and adaptor ligation. The protocol is straightforward but loses directional information. (**c**) Ecker and colleagues[4] serially perform RNA-RNA ligations to add the 3′ and then 5′ adaptor such that directional information is captured. Reverse transcription–PCR then yields an adaptor-flanked DNA library.

necessary for identifying overlapping antisense transcription, whereas this is not the case in the approach of Wold and colleagues (**Fig. 1**). Also, all groups[1,2,4–6] observe substantial unevenness in coverage across the length of any given transcript, potentially resulting from RNA secondary structure or priming bias. The Wold group[2] does a nice job of recognizing and quantifying this problem, and then demonstrates that interventions such as RNA fragmentation can substantially mitigate it. Additional optimization and direct comparison of the various protocols will be helpful in bringing us closer to the goal of truly unbiased transcriptome profiles.

As the two papers in this issue demonstrate, the differences between sequencing and microarray-based transcriptome analysis may be especially poignant in the context of mammalian genomes, where the genome size, the number of genes, the frequency of alternative splicing, and the relative fraction of repetitive and paralogous sequences is much greater than in model organisms such as yeast. Unsurprisingly, the complexity of the mammalian transcriptional landscape renders its analysis correspondingly more difficult. Common challenges faced by Wold group and Grimmond group included: (i) mapping of short sequence reads to the genome; (ii) appropriate assignment of ‘multi-mapping’ reads; (iii) identification of new alternative splice junctions; (iv) classification of reads mapping outside annotated boundaries (for instance, distinguishing genomic DNA contamination versus heterogeneous nuclear RNA versus new transcriptional units versus belonging to adjacent transcriptional units); and (v) comparison of samples to identify differentially expressed genes. They applied varying strategies to these tasks, and additional work will be necessary to identify the optimal approaches. Aspects of their work that may catch on broadly include the ‘reads per kilobase of exon model per million mapped reads’ (RPKM) metric defined by the publicly available software from the Wold group as well as their probabilistic handling of ambiguously mapping reads. Finally, it is important to note that modest improvements to the underlying sequencing methods (for example, longer, more accurate mate-paired reads) will directly mitigate the algorithmic challenges.

Over the past year, commercial implementations of massively parallel short-read sequencing platforms have been applied to profile protein-DNA interactions, cytosine methylation, genetic variation, genomic rearrangements and now transcriptomes. As access disseminates and costs continue to drop, it seems probable that a steadily increasing fraction of the community will begin to use sequencing, rather than microarrays, to interrogate biological phenomena at the genomic scale. However, the point at which sequencing and microarrays achieve cost equivalence varies by application. Some areas, such as expression profiling, may be getting close to this stage, but for others (for example, custom genotyping and copy-number variation), microarrays may prove more resilient. Also, there have been recent reports of ‘preparative’, rather than ‘analytical’ applications of microarrays, including selective genomic capture[9] and precursor production for DNA synthesis[10]. It is possible that in the long run, such unanticipated roles for microarrays may prove to be more important than the original intent around which the technology was developed.

Several technical challenges still separate us from truly comprehensive, quantitative transcriptome profiles. It will be important to achieve long-range continuity such that we know not only what initiation, termination and splice sites are used but how often, in each possible combination. Also, improving efficiency and reducing bias in library construction may eventually allow the transcriptomes of single cells to be comprehensively interrogated. And lastly, methods and algorithms that enable full exploration of the non-poly(A)$^+$ and antisense transcriptomes still await development.

Although these new technologies may improve the quality of transcriptome profiling, we will continue to face what has probably been the larger challenge with microarrays—how best to generate biologically meaningful interpretations of complex datasets that are sufficiently interesting to drive follow-up experimentation.

COMPETING INTERESTS STATEMENT

The author declares competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

1. Cloonan, N. *Nat. Methods* **5**, 613–619 (2008).
2. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
3. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. *Nat. Rev. Genet.* **5**, 335–344 (2004).
4. Lister, R. *et al. Cell* **133**, 523–536 (2008).
5. Nagalakshmi, U. *et al. Science* **320**, 1344–1349 (2008).
6. Wilhelm, B.T. *et al. Nature*, published online 18 May 2008 (doi10.1038/nature07002).
7. Adams, M.D. *et al. Science* **252**, 1651–1656 (1991).
8. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. *Science* **270**, 484–487 (1995).
9. Olson, M. *Nat. Methods* **4**, 891–892 (2007).
10. Tian, J. *et al. Nature* **432**, 1050–1054 (2004).

Hunting hidden transcripts

Piero Carninci

Strategies for the comprehensive identification of transcript isoforms produced from specific genomic loci make use of and expand existing tools and resources.

Eukaryotic genes produce a multitude of RNAs. These include protein-coding mRNAs, non–protein coding RNAs and many variants per genomic locus. In this issue of *Nature Methods*, two complementary studies outline new strategies that will be instrumental to comprehensively decode the primary structure of transcribed RNA variants and also to provide physical cDNA clones for functional experiments[1,2].

Understanding the coding potential of the genome cannot be achieved by analyzing its sequence alone but relies extensively on the experimental identification of its transcribed RNA fraction (the transcriptome). In the early days of genomics, the sequence of known genes, proteins and expressed sequence tags—the latter prepared by sequencing randomly picked cDNAs—were widely used for gene identification. Despite the usefulness of these data to identify expressed regions, accurate identification of mRNA structure has been hampered by the presence of a large amount of truncated cDNAs in conventional cDNA libraries. A first technological evolution was the development and large-scale sequencing of full-length cDNA libraries[3]. After the publication of the first version of the human genome, full-length cDNAs have been used to unambiguously annotate human[4] and subsequently mouse genes[5].

As a surprise to many of us, this annotation process identified only a small number of protein-coding genes (less than 25,000), similar to the number in other organisms that we consider less complex (for instance, the *Caenorhabditis elegans* genome with ~19,000 genes). How could so few genes possibly control very complex biological phenomena, such as vertebrate embryonic development and the human brain?

Some answers to this question came from high-throughput studies[6] involving short sequence tags derived from 5′ cDNA ends (CAGE tags) and paired-end tags deriving from both 5′ and 3′ cDNA ends. This study found that >63% of the mouse genome is transcribed, that there are at least 78,000 protein isoforms generated by alternative splicing and promoter usage, and that more than half of the genome's output is constituted by non–protein coding RNAs[6]. In parallel, tiling arrays, in which all nonrepeated genome regions are represented with a dense array of oligonucleotide probes, have also identified transcribed fragments (named 'transfrags'), which independently demonstrates that a large part of the genome is expressed[7]. The combined use of such experimental approaches in the Encyclopedia of the DNA Elements (ENCODE) project has shown that up to 93% of the genome is expressed[8].

Taken together, although there is a growing consensus that the genome does indeed produce many more transcripts than the number of protein-coding genes, there is little consensus on the upper bound of such complexity. How many different types of RNAs are produced by the use of alternative promoter sites, termination sites and alternative splicing? How many different RNAs and proteins are produced in each cell type of our body?

Notably, precise combinations of multiple exons cannot be unambiguously detected by whole-transcriptome analysis using either tiling arrays or RNA shotgun sequencing. This problem is particularly severe in the case of complex splicing patterns, where combinations of multiple non-neighboring alternative exons cannot be unambiguously assigned without isolating and sequencing individual cDNAs. This approach, however, is laborious and expensive, particularly because RNA expression typically varies widely (at least four orders of magnitude).

Solutions to this problem have been proposed in two articles in this issue of *Nature Methods*[1,2]. The first solution, from a group coordinated by Roderic Guigo and Tom Gingeras, targets the identification of all RNA variants in distinct genomic loci[1]. The design is suitable for identification not only of protein-coding RNAs but also of alternative isoforms that are controlled by novel upstream promoters or terminated at thus-far unknown downstream sites. The other approach, from the groups of Marc Vidal, Fritz Roth and Kourosh Salehi-Ashtiani, focuses on protein-coding variants. Their approach will additionally yield enlarged collections of open reading frame (ORF) clones (ORFeomes), representing a much

Piero Carninci is at the Omics Science Center, RIKEN, Yokohama Institute, Yokohama City, Kanagawa 230-0045, Japan.
e-mail: carninci@riken.jp