Sequence census methods for functional genomics

Barbara Wold & Richard M Myers

Next-generation sequencing technologies are beginning to facilitate genome sequencing. But in addition, new applications and new assay concepts have emerged that are vastly increasing our ability to understand genome function.

When Thomas Edison invented the phonograph, playing music was well down his list of possible uses. Something similar may now be happening in the genome world. A new generation of massively parallel DNA sequencing platforms is here (see Primer). They aim to replace the workhorse capillary sequencing systems that brought us—very beautifully—the human genome sequence. These machines, led by 454 and Illumina (formerly Solexa), and lately joined by ABI[1], have emerged in DNA sequencing centers over the past two years, promising vastly more sequence (>1 gigabase of sequence per run) than standard capillary-based technology can produce. Still other new machines are on the way. Their development is driven by the US National Institutes of Health and National Human Genome Research Institute challenges for DNA sequencing at costs of $100,000 and then of less than $1,000 per human or human-size genome. 'If you build them, we will buy them' was implied, and other near-infinite sequencing appetites such as those of microbial metagenomics researchers have added fuel. As hoped, the new instruments are being explored in the world's genome centers for rapid and cheap genome sequencing. Read length limits, error rates and assembly algorithm issues, among other problems, mean that these new kids on the sequencing block have not fully reduced whole-genome sequencing to practice. Not yet, at least.

Barbara Wold is in the Division of Biology, California Institute of Technology, Pasadena, California 91125, USA. Richard M. Myers is in the Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.
e-mail: bwold@caltech.edu or myers@shgc.stanford.edu

PUBLISHED ONLINE 19 DECEMBER 2007; DOI:10.1038/NMETH1157

But something different and remarkable happened on the way to inexpensive whole-genome sequencing: as music was to Edison's phonograph, a new family of 'sequence census' counting assays is to this new generation of DNA sequencers. If you need to take the measure of an RNA or DNA 'ome', microarrays are no longer the only way to do it. A new and rapidly growing family of assays for measuring the global, genome-wide profiles of mRNAs, small RNAs, transcription-factor binding, chromatin structure, DNase hypersensitivity and DNA methylation status are now being implemented by applying one of the massively parallel, ultrahigh-throughput DNA sequencing systems.

The principle behind these global 'sequence census' methods is disarmingly simple: to learn the content of a complex nucleic acid sample, just sequence it. Sequence it directly—without bacterial cloning as a prerequisite—and do it with the aim of getting just enough sequence to assign the site of origin in the genome for each read rather than trying to determine its entire sequence. A single short sequence read (or sometimes a pair of reads, one from each end) is determined for millions of nucleic acid molecules from a biological sample. You need not sequence the entirety of each molecule in the starting mix, because a small snippet of 25–35 base pairs allows you to use informatics to identify the location of each fragment in the reference genome. Once mapped, you 'count' the hits and analyze their distribution throughout the genome. The key for these uses is the very high number of individual reads, each corresponding to a different molecule in the starting sample. Conceptually, this builds on the ideas behind earlier methods such as serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS)[2,3], with the new assays being substantially less expensive, more general and capable of delivering vastly more information.

These next-generation, 'Seq-based' methods are a natural fit for functional genomics applications because they generate huge numbers of short sequencing reads quickly and cheaply, and, critically for their success, they all focus on a 'reduced genome' input. That is, they do not require sequencing an entire large genome but rather a small fraction of the total that appears as mRNA, as methylated or unmethylated fragments, as DNA or RNA bound by specific proteins, or DNA regions that are hypersensitive to nucleases. The platforms acquire sequence data from amplified single DNA fragments rather than from fragments cloned in plasmids (see Primer). Although cost is a perpetual moving target—and there are likely to be improvements and new competing sequencers—it is now possible to do a thorough measurement of a nucleic acid profile for the same cost or less than that of using hybridization to microarrays. The Seq-based methods bypass some longstanding technical problems of microarrays, including a requirement to synthesize microarrays with millions of DNA probes that does not scale well with large genome size, considerable cross-hybridization, and difficulties with quantitation owing to the continuous nature the hybridization signals. However, these new methods do not solve everything. For example, even though a larger fraction of the genome is accessible by Seq-based methods than by contemporary microarray hybridization, still 15–20% of the reads in the human genome cannot be unambiguously mapped to a single location because they occur more than once in the genome.

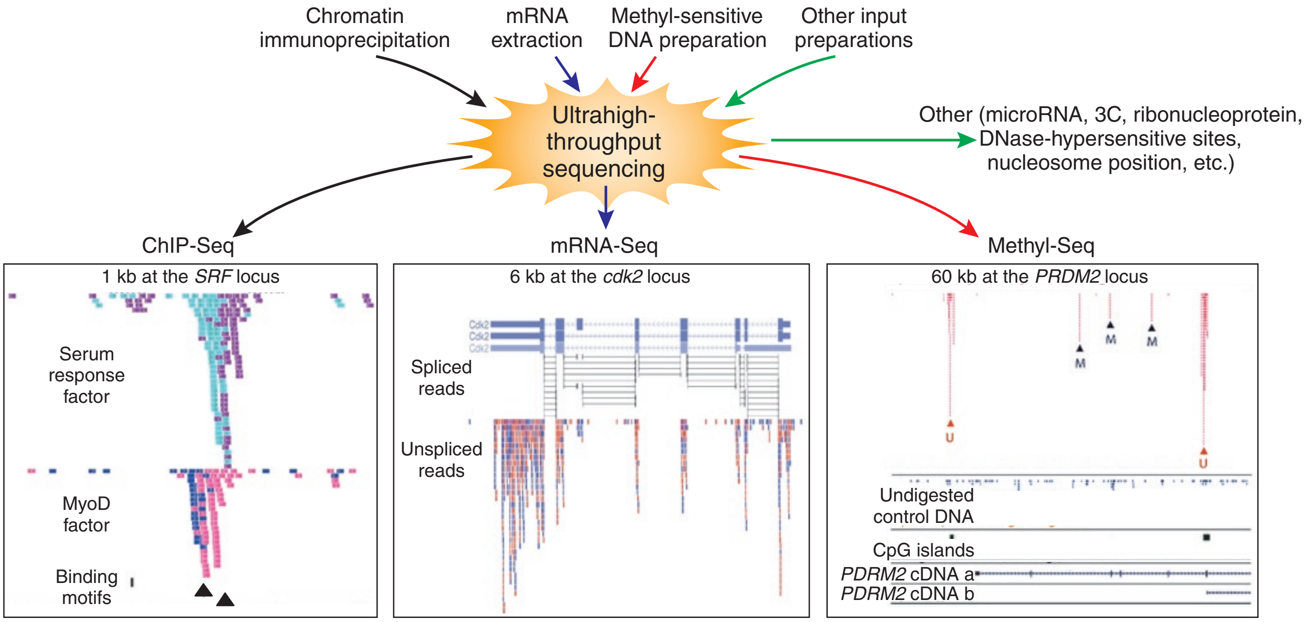

Figure 1 | Sequence census assays. Each colored rectangle corresponds to a single sequence read obtained by Illumina (Solexa) sequencing and mapped to its site of origin at the locus indicated in each panel as displayed on UCSC browser tracks. Chip-Seq (left) was performed for serum response factor (SRF) and myogenic transcription factor (MyoD) in C2C12 cells (B. Williams, G. Kwan, A. Mortazavi, S. Sharp and B.W.; unpublished observation). Read directionality is indicated by blue and pink (MyoD), and green and purple (SRF). Triangles indicate the matches to the transcription factors' respective binding-site motifs. Note that the ChIP-Seq distributions are centered directly over their motif match positions. For a single ChIP-Seq site, read direction is expected to segregate to the left and right of the binding point, as indicated by the colors. The RNA-Seq transcriptome map (middle) was performed on poly(A)$^+$ RNA from myogenic cells (A. Mortazavi, B. Williams and B.W.; unpublished observation). Reads that span RNA splices are shown in black; red and blue indicate sequence read directionality, which is expected to be random for this RNA-Seq protocol. Two exons correspond to alternate splice isoforms. All splices from RefSeq and other gene models were detected. The Methyl-Seq map (right) shows reads at sites that are restriction digested by *Hpa*II or *Msp*I, if they are unmethylated (U) or methylated (M), respectively (D. Johnson and R.M.M.; unpublished observation).

The sequence census application that is farthest along is chromatin immunoprecipitation, or ChIP, and several publications have appeared describing ChIP-Seq in the past few months[5–8] (**Fig. 1**). The goal of these experiments is to map all *in vivo* DNA sites occupied by a DNA-binding protein of interest. To do this, an antibody specifically recognizing a DNA binding protein is used to immunoprecipitate the protein that has been cross-linked to its DNA-binding sites in living cells, bringing the bound DNA fragments along for the ride. The protein of interest can be part of the general transcription machinery, including RNA polymerase or any of its accessory factors, particular versions of modified histones, sequence-specific transcription factors or other DNA-interacting proteins such as those involved in replication or repair.

Until recently, global scoring of ChIPed DNA fragments was almost always done by hybridizing the mixture to microarrays that tile part or most of the genome being studied (so-called ChIP-chip). Although much valuable data have been generated this way, ChIP-chip has the microarray-related limitations mentioned above. ChIP-Seq overcomes some of the biggest problems. As microarray design and fabrication are bypassed, any organism for which a genome sequence is available is fully accessible for all Seq-based assays. Because they do not rely on DNA probes chosen by the experimenter, ChIP-Seq data are 'agnostic', although each platform and application has to be evaluated for possible sequence bias. ChIP-Seq does not suffer from false or uncertain signals resulting from cross-hybridization, and quantification is potentially more accurate because counting sequence reads is 'digital' rather than continuous. Finally, ChIP-Seq can home in on a binding site at higher resolution than is typical for ChIP-chip. A size-selection step and computational features (**Fig. 1**) typically allow binding sites to be localized to regions of 40 base pairs or smaller.

RNA-Seq is another application of ultrahigh-throughput sequencing that is being developed and tested in multiple laboratories, and it seems likely to see even wider use than ChIP-Seq. This began by accelerating the discovery of small RNAs with the 454 platform[9,10], and the other sequencing machines are now in use for the same purpose. For profiling mRNA populations, microarrays have dominated for more than a decade, bringing us most of what we know about entire transcriptomes from yeast and bacteria to mouse, man and mustardweed. RNA-Seq can offer more. For example, RNA splices have long eluded even the densest tiling arrays, yet they can be nicely mapped by Seq-based methods (**Fig. 1**). Also, the Seq-based methods can, by sheer brute force of high sampling, detect RNAs from very low abundance classes (or a rare subpopulation of cells contributing to the sample) and do it unambiguously. But these are very early days for mRNA-Seq, and there are limits and problems—some already appreciated, others still lurking. Long-range transcript mapping to sort out which splices go with each other will, for example, require robust paired-end reads and/or use of longer reads than those obtainable today. All RNA-Seq datasets present a new slew of informatics challenges, including the conundrum of how best to interpret and use reads that map to multiple sites because of gene paralogy. And the quality of a given transcript's map will depend on its RNA prevalence class because rare transcripts do not provide enough reads to map their splices or even to verify the presence of specific exons, unless the input RNA sample is put through prior normalization.

Although there are no published results yet, several groups have begun to use ultra-high-throughput sequencing to help measure the methylation status of DNA at CpG clusters in the human genome (**Fig. 1**). These methods produce a reduced genome sample of CpG island fragments, which in some methods are treated with bisulfite, and the fragments are then sequenced. Although none so far give complete coverage of all CpGs in the human genome, they all provide a massive increase in the number of regions that can be assayed in a single, simple experiment compared to even the most efficient presently available assays. Similar approaches are being used to assay DNase-hypersensitive sites, which are strongly associated with regulatory regions, on a genome-wide scale in living cells.

Lest anyone get too carried away, it is important to recognize current limitations of Seq-based methods as well as the fact that there are applications for which ultra-high-throughput sequencing is not yet (or perhaps ever) the way to go. Even for the most promising uses, the methods are very new, and robust protocols to support production efforts such as that of the National Institutes of Health ENCODE Project are still being developed and refined. Although algorithms for read placement and calling peaks of positive signals have been developed, there is room for improvement. The machines themselves are far from bulletproof, and the yield of usable reads is lower than one would like it to be. Paired-end read sequencing, which should be especially powerful for mapping RNA splice isoforms, is in relatively early stages on the Illumina (Solexa) and ABI (SOLiD) systems.

Even with the uncertainties, an exciting frontier is just beginning to emerge, in which regulatory biology based on a complete whole-genome knowledge of a given cell type and state. The aim, after all, is not to merely annotate the genome, although that is a most useful step. The bigger goal is to understand how the genome specifies all the different cell types and their states of behavior. Being able to assay the regulatory inputs and outputs of the genome routinely and comprehensively, under normal and experimentally or genetically varied conditions, promises to change how we understand the driving combinatorics. At a purely technical level, having the many different kinds of functional genomic measurements made possible or made better by these new Seq-based methods will facilitate long-wanted integration and synthesis of gene circuits and networks.

1. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).
2. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. *Science* **270**, 484–487 (1995).
3. Brenner, S. *et al. Nat. Biotechnol.* **18**, 630–634 (2000).
4. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
5. Barski, A. *et al. Cell* **129**, 823–837 (2007).
6. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. *Science* **316**, 1497–1502 (2007).
7. Mikkelsen, T.S. *et al. Nature* **448**, 553–560 (2007).
8. Robertson, G. *et al. Nat. Methods* **4**, 651–657 (2007).
9. Lu, C. *et al. Genome Res.* **16**, 1276–1288 (2006).
10. Ruby, J.G. *et al. Cell* **127**, 1193–1207 (2006).