GENOMICS

Hacking the genome

Like computer hackers who cooperate in developing and using tools to understand and manipulate the inner workings of computer software, researchers are developing sophisticated biological methods that will allow them to crack the function of the genome.

Genome sequencing has been a tremendous achievement, and scientists are now faced with the daunting prospect of determining the function of all the sequences. Algorithms designed to make functional assignments have been helpful but inadequate. This is particularly true for the noncoding sequences, which make up the majority of the genome.

'Wet' methods used to determine the function of DNA are not amenable to scaling the way sequencing is. As a result, researchers have found it necessary to develop new techniques to probe the function of the genome. To date these methods have relied on two forms of technology. The first involves the sequencing of cDNA libraries derived from mRNA, and the second relies on microarray-mediated hybridization.

Yoshihide Hayashizaki and colleagues at the RIKEN Genomic Sciences Center have been one of the leaders in the development and application of mRNA sequencing technologies for functional genomics. This is epitomized by their cap analysis of gene expression (CAGE) method (**Fig. 1**). CAGE permits the exact determination of the transcriptional start site (TSS) of potentially every mRNA in a complex sample by creating a library of 20-base-pair tags, each representing the 5′ end of a different transcript. Sequencing of these tags and alignment to the genome reveals all the potential TSSs.

Although the basic CAGE technique was first published in 2003, it has since been refined, and these improvements have allowed Hayashizaki and colleagues to perform the first comprehensive TSS analysis of the human and mouse genomes (Carninci *et al.*, 2006). Using this technology, they collected and sequenced 11 million and 8 million tags from the mouse and human genomes, respectively. The number of human tags has since increased to 13 million. "We obtained a total of more than 180,000 distinct TSSs," says Hayashizaki. "We also analyzed the number of genes, and the total number was more than 45,000." Thus, nearly every gene has several start sites, each potentially under the control of a different promoter.

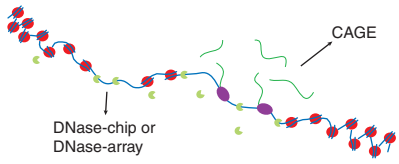

Figure 1 | Techniques for determining genome function. The CAGE technique analyzes populations of mRNA molecules (green) to determine the exact positions of the TSSs on the genome. DNase-chip and DNase-array both use treatment of chromatin with DNase I (light green) to digest regions of DNA lacking histones (red circles). The regions are localized on the genome using DNA-tiling arrays.

Hayashizaki adds, "Living cells discriminate different promoters even if those promoters are in the same coding gene, [therefore] we have to separate the transcriptional activity promoter by promoter." In contrast, if one measures expression at the level of the entire gene, for example by using microarrays, the measurement is just a summation of the total and will miss vital information.

Although microarrays may not be ideal for analyzing the expression of transcripts derived from multiple start sites, they are very powerful tools for other applications. One of these applications may be the elucidation of active regulatory regions of the genome involved in modulating promoter function.

Identification of active transcriptional regulatory elements is traditionally done by analyzing the sensitivity of chromatin to nucleases at specific genomic locations. Regulatory elements are typically characterized by an increased accessibility to regulatory proteins, and thus to nucleases, compared to sites of inactive chromatin. The gold-standard assay consists of treating nuclei with DNase I and measuring the extent of cutting by Southern blot hybridization. For several years, the field has been struggling with the challenge of scaling up such methods that are impractical to use genome-wide. Sequencing-based approaches have been evaluated but remain limited by high background and cost.

In this issue of *Nature Methods*, two groups led by Francis Collins and John Stamatoyannopoulos describe different methods, DNase-chip and DNase-array, that combine the conventional DNAse I assay with a readout using tiling microarrays covering 1% of the human genome (**Fig. 1**; Crawford *et al.*, 2006 and Sabo *et al.*, 2006). Scaling up these methods to the whole genome should require only the generation of more comprehensive tiling microarrays, which seems feasible. These methods thus promise to pave the way for genome-wide identification of DNase I hypersensitive sites, the hallmark of transcriptional regulatory regions.

As an aid to the development of technologies to map the functional genome and distribute the resulting information, several initiatives have been organized. Projects originating in the US, Encyclopedia of DNA Elements (ENCODE), and Japan, Functional Annotation of the Mouse (FANTOM) and Genome Network (GN), have now become global cooperative initiatives.

By exploiting the sequence information that has been collected and by the cooperative development of new tools and sharing of information, genome hackers are on the way to finally cracking the function of the genome.

Daniel Evanko

RESEARCH PAPERS

Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genetics* **38**, 626–635 (2006).

Crawford, G.E. *et al.* DNase-chip: a high-resolution method to identify DNAse I hypersensitive sites using tiled microarrays *Nat. Methods* **3**, 503–509 (2006).

Sabo, P.J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).