GENOMICS

Search party

New assays for the rapid analysis of transcription factor binding site preferences, in combination with a simple yet powerful computational genomic analysis tool, could yield a bounty of valuable data.

Only a small number of target genes have been identified for the thousands of known and putative DNA-binding proteins in the human genome. The identification of regulatory modules is limited both by sparse information on binding preferences, and by the time- and tissue-specific nature of regulation. "To completely understand the complexity of gene regulation by even a single transcription factor, one would have to study every single cell type in the body," explains Jussi Taipale, a biochemist at the University of Helsinki. "But since we have these huge genome [sequences] now, one could computationally find these binding sites."

Taipale, aware of other groups' success in the computational analysis of the fly genome, sought the help of a colleague in the computer science department, Esko Ukkonen, to develop a similar system for sifting through mammalian genomes to identify gene enhancers. The method is traditional: comparing genomic data from different species to identify conserved regions. The twist is in the analysis—rather than just comparing sequences, they performed combinatorial analysis of the relative placement of transcription factor binding sites that are conserved across species. This eliminates noise resulting from sequence variations; the authors compare it to performing alignments with peptide sequences rather than gene sequences. They used their algorithm, enhancer element locator (EEL), with transcription factor binding data both from the JASPAR2 database and from an assay the group developed for characterizing the relative affinity of transcription factors for variants on a canonical binding sequence (Hallikas *et al.*, 2005).

After setting parameters for relative positioning and conservation of binding sites, the Finnish group used data from 107 transcrip-

Figure 1 | Tissue-specific expression of the N-*Myc* gene, driven by two different enhancers. EEL analysis identified two regions near the gene with pairs of GLI transcription factor binding sites; coupling these putative enhancers to *LacZ* revealed markedly different expression patterns. Images reprinted with permission from Elsevier.

tion factors to screen the human genome against the mouse genome, and identified a number of confirmed and putative enhancers. In many cases, they identified enhancers that used a common transcription factor but had different expression profiles (**Fig. 1**), highlighting the importance of cooperation between binding factors. The authors say their approach offers analytical efficiency for mammalian genomes that was previously restricted to simpler organisms. "Our basic scoring scheme is very efficient," explains Ukkonen. "People [have been] trying to use much more computationally heavy probabilistic models that are not easy to use on a mammalian genomic scale." The authors now hope to enhance their accuracy of prediction by expanding EEL to perform alignments from larger numbers of species.

Taipale also hopes to obtain more biological data to bolster the effectiveness of EEL: "I would be very excited if we could get together... some sort of genome-wide human genetic code project where we would figure out binding sites for every transcription factor in the human genome." He may not have long to wait, thanks to new work from Aseem Ansari's team at the University of Wisconsin. While working with synthetic transcription factors, Ansari became frustrated by the inability to obtain comprehensive binding

affinity data. Systematic evolution of ligands by exponential enrichment (SELEX) enriches high-affinity target sequences, but often misses lower-affinity, but biologically relevant, variants. Other techniques, like chromatin immunoprecipitation, limit investigators to a particular tissue or developmental stage. "[We wanted] a sense of the contribution of each nucleotide to binding or recognition of DNA ligands," he says, "and there wasn't really anything available that would do that in an unbiased manner."

Ansari developed a method using microarrays of double-stranded DNAs containing every possible combination of eight or ten base pairs. By treating these chips with fluorophore-conjugated DNA-binding factors it becomes possible to analyze the full spectrum of sequence determinants for recognition and the importance of each nucleotide for binding (Warren *et al.*, 2005). Ansari's group tested their chips with synthetic and natural transcription factors, and obtained relative binding profiles for both molecules that surpass the levels of detail that were previously possible. Notably, the system can also be used to analyze cooperative binding and the impact that different cofactors can have on binding specificity for a given sequence. Ansari also believes this system could be used for studying the affinities of factors that bind specifically to DNA sequences that include bulges, mismatches or chemical modifications.

Like Taipale, Ansari's biggest interest now is in how outside groups—in his case, bioinformaticians—use data from his technique. "It's such a bridge for computational problems," he explains. And indeed, it seems that the intersection of such powerful biological and computational strategies may herald a promising new phase in genomics.

Michael Eisenstein

RESEARCH PAPERS
Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
Warren, C.L. *et al.* Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. USA* **103**, 867–872 (2006).