A plethora of sites

Ghia Euskirchen & Michael Snyder

Understanding how gene expression is regulated on a global scale is important for determining how basic processes such as cell proliferation, cell differentiation and responses to environmental signals are controlled. Three papers now show that it is possible to identify binding sites for key transcription factors in human cells on a chromosome level.

Cellular and developmental processes are controlled in large part by transcription factors whose binding sites are small (average size 6–9 bp) and highly degenerate. How, then, does one find the regulatory elements in a sea of base pairs numbering in the billions?

The chromatin immunoprecipitation and microarray (ChIP-chip) method involves immunoprecipitating chromatin associated with a transcription factor of interest and then probing a genomic DNA array to identify sites bound by the factor (**Fig. 1**). This strategy was originally established in yeast using genomic DNA arrays containing all 6,000 intergenic regions spotted on a single microscope slide[1,2]. Several groups recently adapted this technology to identify transcription factor binding sites in human cells for selected regions of the human genome[3–5]. Two teams, one involving a collaboration between Affymetrix and Harvard[6] and the other at Yale[7,8], have now extended this technology to map binding sites along entire human chromosomes.

Martone *et al.*[7] and Euskirchen *et al.*[8] mapped the binding sites of NF-κB (p65) and CREB, respectively, along human chromosome 22, and Cawley *et al.*[6] mapped the binding sites of Sp1, c-Myc and p53 on chromosomes 21 and 22. In each case, genomic tiling arrays were used. The Affymetrix-Harvard group used oligonucleotide arrays containing 25-bp oligonucleotide pairs, with an average spacing of 35 bp. The Yale group used a PCR-based tiling array containing 21,000 products with an average size of 820 bp.

In the cases of NF-κB, CREB, c-Myc and Sp1, the researchers found an extraordinary number of binding sites along each chromosome studied. The numbers extrapolate to a total of 12,000 and 25,000 binding sites across the entire nonrepetitive regions of the genome. A considerable number of such sites (1,600) are probably also present for p53. Binding was noted near well-annotated genes as well as new transcribed regions whose function is unknown. For the case of annotated genes, the results provide new potential insights into the functions of these transcription factors. For example, many putative CREB targets are involved in neuronal function or signal transduction, with the potential to up- or downregulate the CREB signaling pathway. In addition, several potential targets of CREB and NF-κB are themselves transcription factors, suggestive of the existence of possible regulatory cascades.

Wide distribution

One notable observation is that binding sites lie not only in the immediate vicinity of the transcription start site, but also at distal sites and within genes. These studies show that 9–27% of binding sites lie within 1 kb of the transcription start sites of genes. The results for NF-κB and CREB also show that an additional 25% lie further upstream of the gene (within 10 kb), and 40% are located within introns. This information is important, as many laboratories are currently building and working with proximal promoter regions. Though powerful, these arrays probably uncover only a portion of the potential regulatory information present in these genes.

Many binding sites (36%) lie within or near the 3′ ends of genes. Cawley *et al.*[6] speculate that such sites may regulate antisense transcription, and in several cases, they found that same factor binds near both the 5′ and 3′ ends of genes. For example, the Ewing sarcoma gene *EWSR1*, the tumor suppressor gene *EP300* and the mitogen-activated protein kinase gene *MAPK1* each use Sp1 or c-Myc to regulate both recognized and new transcripts. Cawley *et al.*[6] examined regions in these genes with internal or 3′ binding sites and experimentally verified noncoding or new transcripts that overlap with coding transcripts.

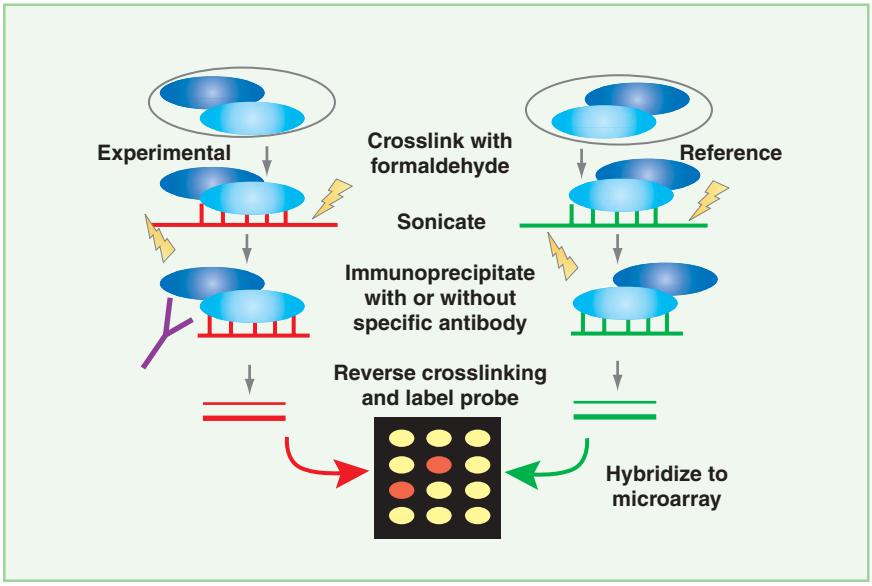

Figure 1 The ChIP-chip method for mapping transcription factor binding sites. To determine transcription factor binding sites along a human chromosome, cells expressing the factor of interest are treated with formaldehyde to promote crosslinking, and protein-DNA complexes are immunoselected from nuclear extracts using antibodies against the factor of interest. As a reference, parallel immunoprecipitations are performed without antibodies or with control antibodies (*e.g.*, preimmune sera). Purified and labeled DNA is then hybridized to a whole-chromosome tiling DNA microarray.

Ghia Euskirchen and Michael Snyder are in the Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut.
e-mail: michael.snyder@yale.edu

Not exact matches

The wealth of information generated from these studies also allows for a comparison of the observed transcription factor binding sites with consensus binding sequences. These data are important, as many groups are using computational approaches to identify potential binding sites based on consensus sequences. The results for each of the factors show that only a fraction (2–35%) of binding sites exactly match consensus sequences. Presumably, then, a considerable proportion of binding occurs at sequences similar but not identical to the consensus sites or is mediated through other factors.

Martone *et al.*[7] and Euskirchen *et al.*[8] further correlate the location of NF-κB and CREB binding sites with expression changes induced by conditions that activate the factors. They found that binding occurs near a proportion of the genes whose expression is induced, as expected. But the correlation is not absolute, as many induced genes lack binding sites, suggestive of indirect modes of regulation. Binding sites were also found near genes whose expression was downregulated after induction. This result was unexpected for NF-κB (p65), which was not previously thought to have a role in repression. Therefore, many, if not most, factors can probably function as both activators and repressors, depending on the genomic context.

Each of these studies identified a plethora of binding sites. A key challenge ahead will be to determine the contribution of each of these sites to the regulation of target genes. Many sites will probably contribute quantitatively to the control of nearby genes, although the presence of redundant binding by regulatory factors seems likely as well. It is also possible that some binding sites regulate expression of nonfunctional messages, providing a unknown level of regulatory 'noise' to the system.

In summary, these studies show for the first time that it is possible to map globally the binding sites of key transcription factors across large chromosomal regions *in vivo*. Scaling of these methods is expected to provide a comprehensive analysis of binding sites across the entire human genome. Further analysis of the ~1,500 human transcription factors in each of the ~250 mammalian cell types is expected to provide a comprehensive picture of the transcriptional control in human cells.

1. Iyer, V.R. *et al. Nature* **409**, 533–538 (2001).
2. Horak, C.E. & Snyder, M. *Meth. Enzymol.* **350**, 469–483 (2002).
3. Horak, C.E. *et al. Proc. Natl. Acad. Sci. USA* **99**, 2924–2929 (2002).
4. Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. & Farnham, P.J. *Genes Dev.* **16**, 235–244 (2002).
5. Li, Z. *et al. Proc. Natl. Acad. Sci. USA* **100**, 8164–8169 (2003).
6. Cawley, S. *et al. Cell* **116**, 499–509 (2004).
7. Martone, R. *et al. Proc. Natl. Acad. Sci. USA* **100**, 12247–12252 (2003).
8. Euskirchen, G. *et al. Mol. Cell. Biol.* in the press.