Analysis and annotation of the transcriptional regulatory sequences of higher eukaryotes: the point of view from the wet lab

Jean Imbert – Centre de Recherche en Cancérologie de Marseille – INSERM U599 – 27 boulevard Leï Roure – 13009 Marseille – Tel.: 04 91 75 84 04 – Fax: 04 91 26 03 64 – Email: imbert@marseille.inserm.fr

For genes that have been successfully delineated within the human genome sequence, most regulatory sequences that control their transcription remain to be elucidated. Hence, comprehensive identification of the *cis*-acting regulatory elements is one of the major challenges of genome biology. Pennachio and Rubin noted in 2001 that "Regulatory sequences constitute a small fraction of the roughly 95% of the human genome that does not encode proteins, but they determine the level, location and chronology of gene expression. Despite the importance of these non-coding sequences in gene regulation, our ability to identify and predict functions for this category of DNA is extremely limited" (1). Until recently indeed, efficient searches for *cis*-regulatory elements and identification of their respective *trans*-acting DNA-binding factors have been based on laborious trial-and-error strategies. These time-consuming experimental approaches, usually targeted at a single gene or locus, include complementary low-throughput *in vivo* and *in vitro* studies (*for detailed protocols and strategies see* (2):

1. Generation of deletion constructs to determine the minimal sequences necessary for transcription of a reporter gene in cell-transfection assays. Then, site-directed and saturation mutagenesis are required to finely define the regulatory elements contained in the minimal fragment that sustains transcription.
2. Mapping of DNAse I hypersensitive sites to identify sequences potentially accessible for transcription factor (TF) binding.
3. *In vivo* and *in vitro* genomic footprinting assays to identify the sequences bound by various regulatory proteins.
4. Electromobility shift assays (EMSAs) to identify the specific protein complex(e)s bound to a given *cis*-regulatory element.
5. Enhancer trapping with various selection vectors such as Cre-lox site-specific recombination system.
6. *In vivo* screen in transgenic mice or using episomic vectors to isolate and characterize *cis*-regulatory sequences.
7. *In vivo* protein-DNA crosslinking combined with immunoprecipitation (Chromatin immunoprecipitation assay, ChIP) to identify and clone the genomic targets of any specific DNA-binding regulatory proteins.

While ChIP assays have proven particularly powerful to analyze the recruitment of specific TFs as well as chromatin modifications, their resolution is limited to a small number of target genes. More recently, coupling chromatin immunoprecipitation to micro-arrays that contain genomic regions ("ChIP-on-chip") has provided investigators with the ability to identify, in a high-throughput manner, regulatory regions and promoters directly bound by specific TFs (3). This opens new prospects for a global analysis of the regulatory pathways of gene expression as well as of the functional chromatin organization of the genome. Although these approaches still raises sensitivity and accessibility issues, in particular with higher organisms, very encouraging results were recently obtained in several laboratories and new pangenomic DNA microarrays are currently available from several companies. In parallel, the design of these specific microarrays, and the management and interpretation of the data generated stimulate the development of specific computational suites for the processing, the integration, the analysis and the modeling of Chip-on-chip data sets.

Theoretically, ChIP-on-chip assay allows the researcher to take a snapshot of all genomic occupancies of a given transcription factor or chromatin component in living cells. So far, different types of genomic microarrays have been used for high throughput screening of ChIP samples. They can be classified in three categories following their composition: selected promoters, a random selection of CpG islands, or portions of continuous genomic sequences. Selected promoters and portions of genomic sequences, usually in the range of 1 Kb, are generally obtained by PCR amplification using specific pairs of primers. The same genomic regions can be also covered by tiling 50-60-mers oligonucleotides that provide a cost-effective and less error-prone alternative. CpG islands microarrays are all derived from the original CpG islands libraries built by Bird and coworkers (4). These micoarrays are based on the observations that CpG dinucleotides are under represented in the mammalian genome (20% of the expected frequency) whereas CpG islands of 200-2000 base-pairs in length are located close to the 5' end of ~60% of known genes. Furthermore, studies of the role of

tumor suppressor genes in cancer development and progression has generated increasing evidence for a crucial role of CpG island hypermethylation (5).

An alternative approach, known under the acronym DamID (6) uses the unique properties of *E. coli* DNA adenine methyl transferase (dam). A fusion protein associating dam and a chromatin protein of interest will methylate adenines in DNA sequence in the vicinity of the chromatin protein binding sites. Purification by methyl-specific PCR amplification provides highly specific genomic DNA fragments to screen promoter or pangenomic microarrays. While this approach requires ectopic expression of an artificial fusion protein that might affect its outcome, it presents the obvious advantage to be available for any DNA binding protein without requirement of a highly specific antibody. Other approaches known as GMAT (7) or SACO (8) combine ChIP and SAGE-like library cloning for high-resolution genome-wide mapping of chromatin protein occupancy.

Large scale identification of DNAse I hypersensitive sites using active chromatin sequence libraries might also provide an efficient tool to identify and clone important regulatory regions at genome scale (9).

Several computational approaches have also been proposed to guide our search for *cis*-regulatory regions at the level of individual gene or whole genome:

1. Inter-species sequence comparisons: identification of non-coding sequences with reasonable chances having regulatory properties. Sequences that regulate gene expression tend to be conserved among species as illustrated by many transgenic experiments where genes from various mammals are nearly always expressed similarly to their expression in their natural host when transferred as large genomic fragments.
2. Sequence analysis of co-regulated genes within a species: most TFs bind to conserved sites in several genes to coordinate their expression. The assumption is that gene co-expression depends on similar regulatory pathways triggering binding of similar sets of shared transcription factors to conserved *cis*-regulatory elements.
3. Screening of putative regulatory regions with databases of known transcription binding sites.

Severe limitations however impair a general and easy use of *in silico* approaches such as phylogenetic footprinting. Among them, we can mention either a too high degree of conservation between two related species with no clear “islands” of highly conserved non-coding sequences, or absence of significant similarities. Furthermore, functional conservation of gene expression is not sufficient to assure the evolutionary preservation of corresponding *cis*-regulatory elements (1,10). For example, even transcription start site prediction softwares such as Eponine and MatInspector can only detect approximately 50% of well characterized promoters (11). Finally, several experimentally characterized regulatory elements are not conserved between species (12).

Another important limitation is linked to the fact that binding sites for transcription factors are often degenerate and better characterized as a probability (position weight matrix) than as a consensus sequence (13). Consequently, quality of the databases collecting the transcription factor binding sites relies largely on the number of functionally well-defined DNA binding sites available for a given transcription factor. On one hand, continuous accumulation of biochemical and molecular biology approaches are mandatory to improve size and quality of these databases. On another hand, and notwithstanding real limitations mentioned above, ever improving bioinformatic tools and databases offer a solid support to the wet lab approach.

During my lecture, I will provide some examples how combining adequate *in vitro* and *in vivo* functional assays and some easily accessible bioinformatic tools helped me in deciphering the architecture of the complex regulatory regions characteristic of the human genes (14,15).

References

1. Pennacchio, L. A., and E. M. Rubin. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2: 100-109.
2. Carey, M., and S. T. Smale. 2000. *Transcriptional regulation in eukaryotes - Concepts, strategies and techniques*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
3. Kirmizis, A., and P. J. Farnham. 2004. Genomic approaches that aid in the identification of transcription factor target genes. *Exp. Biol. Med.* 229: 705-721.

4. Cross, S. H., J. A. Charlton, X. Nan, and A. P. Bird. 1994. Purification of CpG islands using a methylated DNA binding column. *Nat Genet.* 6: 236-244.
5. Esteller, M. 2005. DNA methylation and cancer therapy: new developments and expectations. *Curr. Opin. Oncol.* 17: 55-60.
6. van Steensel, B., and S. Henikoff. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* 18: 424-428.
7. Roh, T. Y., W. C. Ngau, K. Cui, D. Landsman, and K. Zhao. 2004. High-resolution genome-wide mapping of histone modifications. *Nat Biotechnol.* 22: 1013-1016.
8. Impey, S., S. R. McCorkle, H. Cha-Molstad, J. M. Dwyer, G. S. Yochum, J. M. Boss, S. McWeeney, J. J. Dunn, G. Mandel, and R. H. Goodman. 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 119: 1041-1054.
9. Sabo, P. J., R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. U. S. A.* %19;..
10. Ludwig, M. Z. 2002. Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.* 12: 634-639.
11. Down, T. A., and T. J. Hubbard. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* 12: 458-461.
12. Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross- species sequence comparisons. *Science* 288: 136-140.
13. Wasserman, W. W., and W. Krivan. 2003. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften.* 90: 156-166.
14. Kim, H. P., J. Imbert, and W. J. Leonard. 2006. Both integrated and differential regulation of components of the IL-2/IL-2 receptor system. *Cytokine Growth Factors Rev*, 17:349-366, 2006.
15. Grange, T., J. Imbert, and D. Thieffry. 2005. Epigenomics: large scale analysis of chromatin modifications and transcription factors/genome interactions. *Bioessays* 27: 1203-1205.