1. Introduction

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. In a review published in 2001, Pennachio and Rubin noted 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” (28).

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 (6):

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(es) 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.

Few other medium- or high-throughput approaches are also proposed to isolate *cis*-regulatory sequences but still require more experimental confirmations to be fully validated. For example, a limited number of large scale promoter and enhancer screening have been carried out (4,9,10). More recently, the so-called ChIP-on-chip assay combining ChIP and DNA microarrays has been successfully used to identify the binding sites for specific transcription factors throughout the yeast (29,31) and human genomes (40,41). Alternatively, large scale identification of DNAse I hypersensitive sites using active chromatin sequence libraries might provide an efficient tool to identify and clone important regulatory regions at genome scale (33).

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 (24,28). For example, even transcription start site prediction softwares
such as Eponine and MatInspector can only detect approximately 50% of well characterized promoters (8). Finally, several experimentally characterized regulatory elements are not conserved between species (23).

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 (39). 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 this course, I will provide some examples how combining adequate *in vitro* and *in vivo* functional assays and some simple bioinformatic tools can help in deciphering the architecture of two complex regulatory region characteristic of the human genes.

2. Basic concepts in transcriptional regulation in eukaryotes

<table>
<thead>
<tr>
<th>Table I: <em>cis</em>-acting regulatory elements (from (28))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Promoter</strong>: Sequence of DNA near the 5’ end of a gene that acts as a binding site for RNA polymerase and from which transcription is initiated.</td>
</tr>
<tr>
<td><strong>Enhancer</strong>: Control element that elevates the levels of transcription from a promoter, independent of orientation or distance.</td>
</tr>
<tr>
<td><strong>Locus control region (LCR)</strong>: Confers tissue-specific temporally regulated expression of linked genes. LCRs function independently of position, but they are copy number dependent and open the nucleosome structure so that other factors can bind. LCRs affect replication timing and origin usage.</td>
</tr>
<tr>
<td><strong>Boundary element/insulator</strong>: DNA sequence that prevents the activation or inactivation of transcription by blocking the effects of surrounding chromatin.</td>
</tr>
<tr>
<td><strong>Silencer</strong>: Control element that suppresses gene expression independent of orientation or distance.</td>
</tr>
<tr>
<td><strong>Matrix attachment region (MAR)/scaffold attachment region (SAR)</strong>: DNA sequence that binds the nuclear scaffold and can affect transcription. These elements probably form higher-order looped structures within chromosomes and influence gene expression by separating chromosomes into regulatory domains.</td>
</tr>
</tbody>
</table>
Higher eukaryote genes contain highly structured regulatory DNA sequences that direct complex patterns of expression in many different cell types during development and in response to multiple stimuli (22). Two main classes of regulatory regions can be defined: the core promoter and the regulatory regions. The core promoter is a short stretch of DNA near the transcription start sequences (TSS) at the 5’ end of a gene that acts as a binding site for RNA polymerase and from which transcription is initiated. The regulatory regions modulate the level of transcription from a core promoter independently from orientation and distance. They include various functionally distinct regulatory elements that can be structurally related (enhancer, LCR, silencer, etc., see Table I). A wealth of studies has shown that modular multi-binding site architecture is required for the function of most eukaryotic regulatory regions (3,20,27).

In metazoan, gene activation involves several events including chromatin opening, activator binding to regulatory regions, recruitment of basal transcription factors and RNA polymerase to the core promoter, and transcription elongation. A general model has emerged in which activators function to stabilize or modulate transcription through interactions with histone modifying and chromatin remodeling enzymes (Fig. 1).
In this model, the initial step in transcription activation involves the recognition of the promoter/enhancer region by sequence-specific DNA-binding factors. These factors recruit ATP-using chromatin remodeling factors to the template. In some cases, sequence-specific DNA-binding factors and remodeling factors may bind to chromatin in a concerted fashion. Chromatin-remodeling factors catalyze the mobilization of nucleosomes, as is needed for the binding of additional transcription factors and coregulators to the DNA template. In addition, the promoter/enhancer-binding factors mediate, directly or indirectly, the association of acetyltransferases (such as CBP/p300, PCAF/Gcn5, SRC/p160) that modify core histones and other proteins essential for transcription initiation. Protein acetylation probably acts through multiple mechanisms to promote unfolding of the chromatin, to modulate the affinity of DNA-binding factors, and to regulate activities of transcription factors and cofactors. Promoter/enhancer-binding factors also interact with a large multisubunit complex, which has related forms known as TRAP, ARC, DRIP, SMCC, NAT, SRB, and Mediator. In turn, this complex interacts with RNA polymerase II to facilitate communication with the basal transcriptional machinery. There are also direct interactions between the TAF subunits (TBP-associated factors) of the TFIID complex and promoter/enhancer-binding factors. Many, and possibly all, of these interactions are required to achieve productive transcription initiation. Transcriptional elongation through chromatin requires the action of the P-TEFb (CycT-Cdk9) and FACT (Spt16-SSRP1) complexes, and is marked by the progressive phosphorylation of the carboxy-terminal domain (CTD) of RNA polymerase II. Core promoter elements, such as the TATA box and DPE, can be important for enhancer-to-core promoter communication. Boundary/insulator elements demarcate domains of gene activity. Many of these activation steps can be blocked by specific repressors to prevent inappropriate gene expression or dampen the response to inducers (not shown). The fortuitous placement of heterochromatin near a euchromatic gene can have a repressive effect on transcription (position-effect variegation).

3. Examples of identification of cis-acting regulatory regions in two human genes

3.1. Transcriptional control of CD25/IL-2Rα gene expression

The IL-2/IL-2 receptor system is critical to the proper regulation of T-cell expansion and contraction following antigen encounter. Resting lymphocytes do not produce IL-2; however, production of IL-2 by CD4+ T cells is potently induced following exposure to antigenic or mitogenic stimulation. Concomitantly, high-affinity IL-2 receptors are expressed on activated T cells. The binding of IL-2 to high-affinity IL-2 receptors serves as a major T-cell growth factor controlling antigen-mediated cell growth. IL-2 also plays an essential role in mediating antigen-induced cell death (AICD). In addition to its actions on T lymphocytes, IL-2 can boost the cytolytic activity of natural killer (NK) cells, can augment immunoglobulin production by B cells, and can exert actions, at least in vitro, on neutrophils and macrophages. Thus, the IL-2/IL-2 receptor system has pleiotropic actions.

Expression of the IL-2Rα gene is controlled at the levels of transcription initiation and mRNA stabilization(21,37). Analysis of its cis-acting regulatory regions has defined several negative and positive regulatory elements, located upstream and downstream of the major transcription initiation site (Fig. 2).
These elements are respectively responsible for its repression in resting T lymphocytes and its expression after activation. Five positive regulatory regions (PRR) and a poorly characterized negative regulatory region have been identified in the *human* gene. PRRI and PRRII are involved respectively in the activation and the control of its tissue specific expression. PRRI contains consensus binding sites for SRF and the NF-κB transcription factor family (1,7). PRRII is recognized by Elf-1, a member of the Ets transcription factor family, preferentially expressed in the lymphoid lineage (13). PRRIII and PRRIV are responsible for IL-2-amplification of its own high affinity receptor (13,16,19,36). The distal *cis*-regulatory region CD28rE is a potent transcriptional enhancer activated by the CD28 and CD3/TcR pathways (43). In the mouse IL-2Rα gene, a Base Unpairing Region (BUR) has been identified in the first intron that is bound specifically by the factor SATB-1. SATB1 targets chromatin remodelling to the IL-2Rα gene, which is ectopically transcribed in SATB1 null thymocytes (2). SATB1 recruits the histone deacetylase contained in the NURD chromatin remodelling complex to a SATB1-bound site in the IL-2Rα locus, and mediates the specific deacetylation of histones in a large domain within the locus. SATB1 also targets ACF1 and ISWI, subunits of CHRAC and ACF nucleosome mobilizing complexes, to this specific site and regulates nucleosome positioning over seven kilobases (42). I will illustrate during my lecture how using simple bioinformatic tools has greatly helped in the delineation of this complex genomic architecture (Fig. 3 and 4).
3.2. Selection of CD19 B-cell specific regulatory sequences and design of CD19-GFP lentiviral vectors (26).

The CD19 cell surface molecule regulates signal transduction events critical for B lymphocyte development and humoral immunity. The CD19 gene is expressed early and throughout B-lymphoid differentiation, to be lost only at the terminal stage of maturation in plasma cells; it takes part to the CD21, CD81 and Leu-13 complex which regulates thresholds of intracellular signals (38). Its expression is primarily regulated at the transcriptional level (17,25,44).
To determine the minimal genomic sequences that are functionally important and may thus confer B-cell specificity when inserted in a lentiviral vector, we used a computational sequence analysis approach (28) to study the CD19 gene regulatory regions. Contigs containing the human and mouse CD19 genes were first obtained from Genbank and Ensembl genome databases (www.ensembl.org, (12)). Using the RepeatMasker software (35), repetitive sequences were masked in contig AC109460 containing the human CD19 gene. Second, comparison with the mouse contig AC125169 was carried out, using the Pipmaker software (34). Two well-conserved genomic fragment were identified in human and mouse CD19 enhancer/promoter regions (18), including a previously non identified 5’ upstream fragment. The two conserved regions, thereafter named CR1 and CR2, were contained in a contiguous 1274 bp fragment of the human CD19 gene (Fig. 5). CR2 (198 bp) was located –1 Kb upstream of human CD19 gene initiation site, and at a not yet localized upstream position in the mouse gene, since the sequence of contig AC125169 was not yet entirely known. To
allow for a more precise comparison using Pipmaker, we created a 1529 bp fragment for the mouse CD19 gene where CR2 was inserted 5' upstream and artificially separated with 100 Ns from the sequenced mouse CD19 promoter. The CD19 enhancer/promoter region contains several well characterized regulatory elements required for B-cell specific expression such as the binding sites for the B-cell specific transcription factor BSAP/Pax5 (11,15,18,32). Both human and mouse CD19 upstream regions were further screened for putative regulatory elements using the TESS (5) and MatInspector bioinformatic tools (30) (Fig. 5).

Based upon this computational approach, we created three lentiviral vectors, in which the GFP gene was placed under the control of human CD19 gene enhancer/promoter fragments. All CD19 sequences were produced by PCR methods from human primary cell genomic DNA. The S.CD19 sequence contains the proximal region of homology (CR1) for a total length of 230 bp. The composite M.CD19 sequence of 464 bp length results from the artificial juxtaposition of the two fragments containing the proximal and distal highlighted homologous regions (CR1 and CR2). The last sequence named L.CD19 corresponds to the entire native CD19 sequence from the initiation site to the CR2 region, and is 1274 bp long. In order to produce B-specific lentiviral vectors, we cloned the three CD19 sequences upstream GFP marker gene between MluI and BamHI restriction sites of the pTRIPΔU3-EF1-GFP plasmid in place of the ubiquitous EF1α promoter (235 bp): the subsequent recombinant lentiviral vectors were respectively called S.CD19-GFP, M.CD19-GFP and L.CD19-GFP (Fig. 6).

![Figure 6. Recombinant CD19-GFP lentiviral vectors.](image)

Ubiquitous EF1α promoter (235 bp) of pTRIPΔU3-EF1α-GFP vector plasmid was substituted by the S.CD19 (230 bp), M.CD19 (464 bp) or L.CD19 (1274 bp) sequence by cloning between MluI(▲) and BamHI(●) sites. New plasmids were respectively called pTRIPΔU3-S.CD19-GFP, pTRIPΔU3-M.CD19-GFP, pTRIPΔU3-L.CD19-GFP and gave rise to the S.CD19-GFP, M.CD19-GFP and L.CD19-GFP lentiviral vectors. LTR: Long Terminal Repeat; SD/SA: Splice Donor / Splice Acceptor; Ψ: encapsidation signal (on truncated GAG sequence); RRE: Rev Responsive Element; Triplex: Central Termination Sequence (CTS) and central PolyPurine Tract (cPPT) of HIV-1 virus; ΔU3: deletion of 400 bp in the 3’LTR U3 region (Self Inactivated vector).

These CD19-GFP lentiviral vectors efficiently transduced cord blood CD34+ progenitor cells. The M.CD19 and especially L.CD19 sequences, preferentially targeted GFP expression to in
vitro and in vivo differentiated CD19+ progeny; moreover, transgene expression was detected from the CD34+ pro/pre-B cells to the mature peripheral IgM+ B-cell stage (Fig. 7). In contrast, GFP expression was weak or absent in primary T-lymphoid and uncommitted progenitor cells, or in erythroid, Natural Killer and myeloid differentiated cells.

**Figure 7. GFP expression in the progeny of transduced CD34+ progenitor cells differentiated in vitro.**
Transduced and un-transduced CD34+ placental progenitor cells were submitted to in vitro differentiation by co-culture with the MS-5 stromal cell line (B-lymphoid, myeloid and NK differentiation) or in liquid culture (erythroid differentiation). Dot plots come from representative experiments showing typical GFP expression according to cell phenotype and transduction lentiviral vector. CD19+ (B-lymphoid), CD33+ (myeloid) and CD56+ (NK) differentiated cells came from the same progenitors transduction protocol; CD19+ and CD33+ cells appeared in parallel in the same culture well. GFP expression was detected by flow cytometry after 10 days in culture for erythroid (GP-A+) differentiation, 3 weeks for NK differentiation and 5 weeks for B-lymphoid/myeloid differentiation. Analysis were done inside the ToPRO-3−/hCD45+ gate; boxes on dot plots delimit lineage/GFP+ cells. On the right, specificity index (SI) of each promoter sequence according to cell lineage was illustrated by histograms relatively to CD19+ cells: subsequently, a lower SI indicates a better specificity for B-cells. Mean values ±SE of 7 (EF1-GFP, dashed bar), 4 (S, closed bar and M.CD19-GFP, grey bar) and 3 (L.CD19-GFP open bar) experiments are indicated. SI could not be calculated for GP-A+ cells because no CD19+ differentiated cells coming from common progenitor transduction protocol were obtained in parallel.

4. References


