Chapter 3 – Transcriptional regulation

Ch 3 - L 1.1

Transcriptional regulation

ENCODE transcriptomics

Leading Edge
Review

Textbook

Cell

Looping Back to Leap Forward: Transcription Enters a New Era

Michael Levine,[1,*] Claudia Cattoglio,[1,2] and Robert Tjian[1,2,*]
[1]Department of Molecular and Cell Biology
[2]Howard Hughes Medical Institute, CIRM Center of Excellence, Li Ka Shing Center for Biomedical and Health Sciences
University of California, Berkeley, Berkeley, CA 94707, USA
*Correspondence: mlevine@berkeley.edu (M.L.), jmlim@uclink4.berkeley.edu (R.T.)
http://dx.doi.org/10.1016/j.cell.2014.02.009

Comparative genome analyses reveal that organismal complexity scales not with gene number but with gene regulation. Recent efforts indicate that the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers. Proliferation of *cis*-regulatory DNAs is accompanied by increased complexity and functional diversification of transcriptional machineries recognizing distal enhancers and core promoters and by the high-order spatial organization of genetic elements. We review progress in unraveling one of the outstanding mysteries of modern biology: the dynamic communication of remote enhancers with target promoters in the specification of cellular identity.

Levine, 2014

Introduction

Transcription regulation is the premier mechanism underlying differential gene activity in animal development and disease.

differential gene activity

Comparative genome analyses reveal that organismal complexity scales not with gene number but with gene regulation. Recent efforts indicate that the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers. Proliferation of *cis*-regulatory DNAs is accompanied by increased complexity and functional diversification

... the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers.

This is one of the most impacting results of **ENCODE**

ENCODE is one of the largest international projects of ***functional genomics***

Started at the completion of Human Genome sequencing

ENCODE 2007 with traditional sequencing + microarrays
1% of the Human Genome
finding promoters, enhancers, transcripts etc

Scaled up rapidly after the introduction of NGS

Vol 447|14 June 2007|doi:10.1038/nature05874

nature

thirteen years ago...

ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

We report the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further integrated and augmented by a number of evolutionary and computational analyses. Together, our results advance the collective knowledge about human genome function in several major areas. First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view of chromatin structure has emerged, including its inter-relationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded new mechanistic and evolutionary insights concerning the functional landscape of the human genome. Together, these studies are defining a path for pursuit of a more comprehensive characterization of human genome function.

ARTICLE 6 SEPTEMBER 2012 | VOL 489 | NATURE | 57

doi:10.1038/nature11247

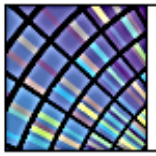
eight years ago...

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with



ENCODE
Encyclopedia of DNA Elements
nature.com/encode

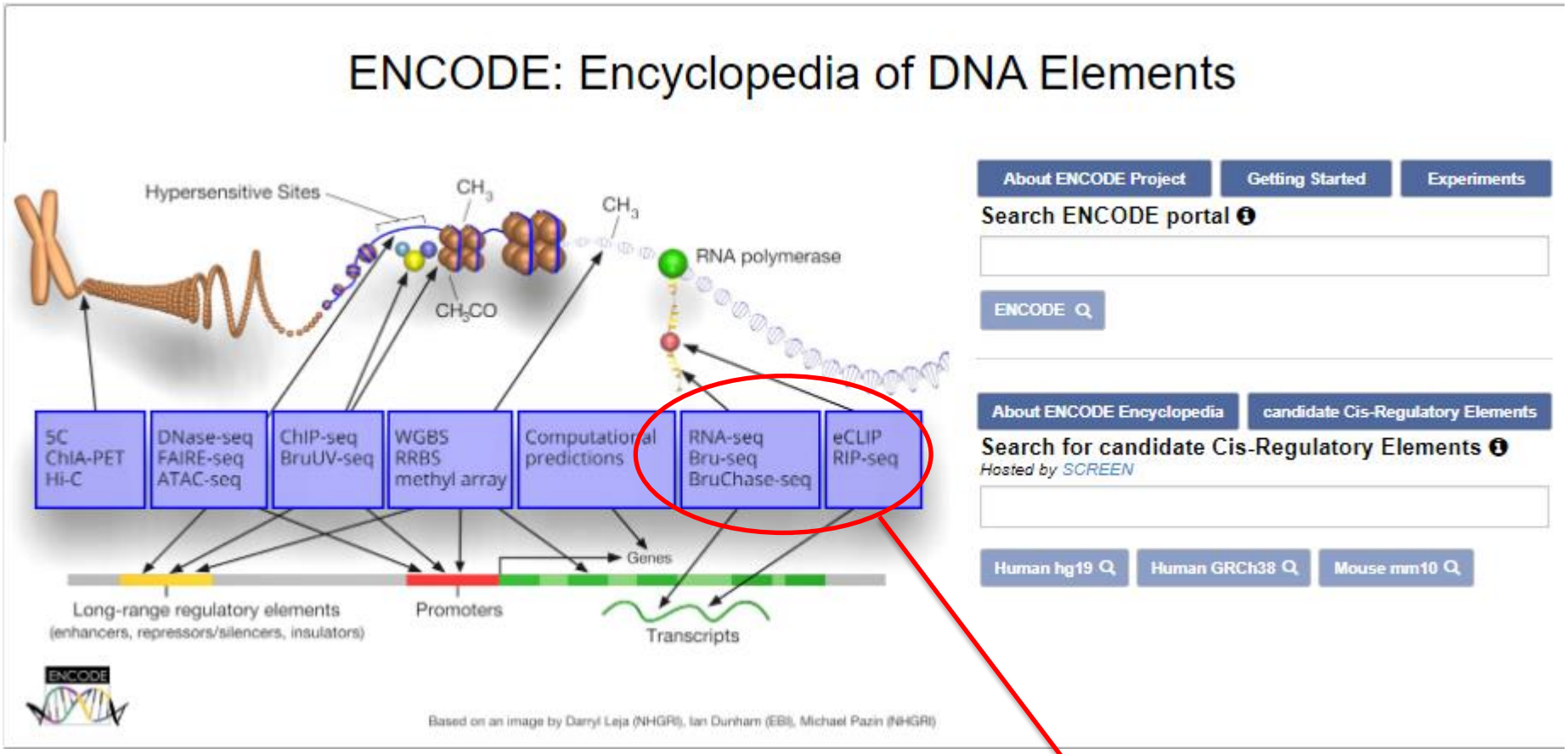
95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

*This Article was accompanied by **30** specific articles published in the same month !*

in brief...

- ✓ 80.4% of the human genome participate in at least one biochemical event
- ✓ two third of genomic sequences are represented in RNA
- ✓ 400,000 sites have chromatin features of enhancers
- ✓ 70,300 regions have promoter-like features

ENCODE project website:



Transcriptomics
RNA metabolism

connect to Encode experiment matrix

https://www.encodeproject.org/matrix/?type=Experiment&status=released

ARTICLE

doi:10.1038/nature11212

An expansive human regulatory lexicon encoded in transcription factor footprints

Shane Neph[1]*, Jeff Vierstra[1]*, Andrew B. Stergachis[1]*, Alex P. Reynolds[1]*, Eric Haugen[1], Benjamin Vernot[1], Robert E. Thurman[1], Sam John[1], Richard Sandstrom[1], Audra K. Johnson[1], Matthew T. Maurano[1], Richard Humbert[1], Eric Rynes[1], Hao Wang[1], Shinny Vong[1], Kristen Lee[1], Daniel Bates[1], Morgan Diegel[1], Vaughn Roach[1], Douglas Dunn[1], Jun Neri[1], Anthony Schafer[1], R. Scott Hansen[1,2], Tanya Kutyavin[1], Erika Giste[1], Molly Weaver[1], Theresa Canfield[1], Peter Sabo[1], Miaohua Zhang[3], Gayathri Balasundaram[3], Rachel Byron[3], Michael J. MacCoss[1], Joshua M. Akey[1], M. A. Bender[3,4], Mark Groudine[3,5], Rajinder Kaul[1,2] & John A. Stamatoyannopoulos[1,6]

Regulatory factor binding to genomic DNA protects the underlying sequence from cleavage by DNase I, leaving nucleotide-resolution 'footprints'. Using genomic DNase I footprinting across 41 diverse cell and tissue types, we detected 45 million transcription factor occupancy events within regulatory regions, representing differential binding to 8.4 million distinct short sequence elements. Here we show that this small genomic sequence compartment, roughly twice the size of the exome, encodes an expansive repertoire of conserved recognition sequences for DNA-binding proteins that nearly doubles the size of the human *cis*-regulatory lexicon. We find that genetic variants affecting allelic chromatin states are concentrated in footprints, and that these elements are preferentially sheltered from DNA methylation. High-resolution DNase I cleavage patterns mirror nucleotide-level evolutionary conservation and track the crystallographic topography of protein–DNA interfaces, indicating that transcription factor structure has been evolutionarily imprinted on the human genome sequence. We identify a stereotyped 50-base-pair footprint that precisely defines the site of transcript origination within thousands of human promoters. Finally, we describe a large collection of novel regulatory factor recognition motifs that are highly conserved in both sequence and function, and exhibit cell-selective occupancy patterns that closely parallel major regulators of development, differentiation and pluripotency.

ARTICLE

doi:10.1038/nature11212

An expansive human regulatory lexicon encoded in transcription factor footprints

what is this ?

nucleotide-resolution genomic DNase I footprinting
in 41 diverse cell and tissue types

45 million transcription factor occupancy events within regulatory regions,
binding to 8.4 million distinct short sequence elements

genetic variants affecting allelic chromatin states are concentrated in footprints, and sheltered from DNAmethylation.

a stereotyped 50-base-pair footprint that precisely defines the site of transcript origination within thousands of human promoters

a large collection of novel regulatory factor recognition motifs

ARTICLE

doi:10.1038/nature11245

Architecture of the human regulatory network derived from ENCODE data

Mark B. Gerstein[1,2,3]*, Anshul Kundaje[4]*, Manoj Hariharan[5]*, Stephen G. Landt[5]*, Koon-Kiu Yan[1,2]*, Chao Cheng[1,2]*, Xinmeng Jasmine Mu[1]*, Ekta Khurana[1,2]*, Joel Rozowsky[2]*, Roger Alexander[1,2]*, Renqiang Min[1,2,6]*, Pedro Alves[1]*, Alexej Abyzov[1,2], Nick Addleman[5], Nitin Bhardwaj[1,2], Alan P. Boyle[5], Philip Cayting[5], Alexandra Charos[7], David Z. Chen[3], Yong Cheng[5], Declan Clarke[8], Catharine Eastman[5], Ghia Euskirchen[5], Seth Frietze[9], Yao Fu[1], Jason Gertz[10], Fabian Grubert[5], Arif Harmanci[1,2], Preti Jain[10], Maya Kasowski[5], Phil Lacroute[5], Jing Leng[1], Jin Lian[11], Hannah Monahan[7], Henriette O'Geen[12], Zhengqing Ouyang[5], E. Christopher Partridge[10], Dorrelyn Patacsil[5], Florencia Pauli[10], Debasish Raha[7], Lucia Ramirez[5], Timothy E. Reddy[10]†, Brian Reed[7], Minyi Shi[5], Teri Slifer[5], Jing Wang[1], Linfeng Wu[5], Xinqiong Yang[5], Kevin Y. Yip[1,2,13], Gili Zilberman-Schapira[1], Serafim Batzoglou[4], Arend Sidow[14], Peggy J. Farnham[9], Richard M. Myers[10], Sherman M. Weissman[11] & Michael Snyder[5]

Transcription factors bind in a combinatorial fashion to specify the on-and-off states of genes; the ensemble of these binding events forms a regulatory network, constituting the wiring diagram for a cell. To examine the principles of the human transcriptional regulatory network, we determined the genomic binding information of 119 transcription-related factors in over 450 distinct experiments. We found the combinatorial, co-association of transcription factors to be highly context specific: distinct combinations of factors bind at specific genomic locations. In particular, there are significant differences in the binding proximal and distal to genes. We organized all the transcription factor binding into a hierarchy and integrated it with other genomic information (for example, microRNA regulation), forming a dense meta-network. Factors at different levels have different properties; for instance, top-level transcription factors more strongly influence expression and middle-level ones co-regulate targets to mitigate information-flow bottlenecks. Moreover, these co-regulations give rise to many enriched network motifs (for example, noise-buffering feed-forward loops). Finally, more connected network components are under stronger selection and exhibit a greater degree of allele-specific activity (that is, differential binding to the two parental alleles). The regulatory information obtained in this study will be crucial for interpreting personal genome sequences and understanding basic principles of human biology and disease.

ARTICLE

doi:10.1038/nature11245

Architecture of the human regulatory network derived from ENCODE data

Transcription factors bind in a combinatorial fashion to specify the on-and-off states of genes

the genomic binding information of 119 transcription-related factors in over 450 distinct experiments.

combinatorial, co-association of transcription factors: distinct combinations of factors bind at specific genomic locations.

there are significant differences in the binding proximal and distal to genes.

Regulatory regions

Promoters *versus* Enhancers

the main feature is that promoters are always in the region immediately preceding and overlapping the Transcriptional Start Site (TSS)

while

Enhancers are placed in virtually indifferent regions around the gene, i.e. up to 100,000 bp upstream or downstream, in introns, with aparent no deal of distance with function.

Definitions:

Promoter = the minimal sequence sustaining transcription and correct initiation, usually 50-150 bp 5'-upstream TSS

Upstream regulatory sequence: sequences 5'-adjacent to promoter that regulate promoter utilization (500-2,000 bp, usually a downstream part to +100 is also included). Also sometimes indicated as «UAS», «proximal regulatory element» or «proximal enhancer».

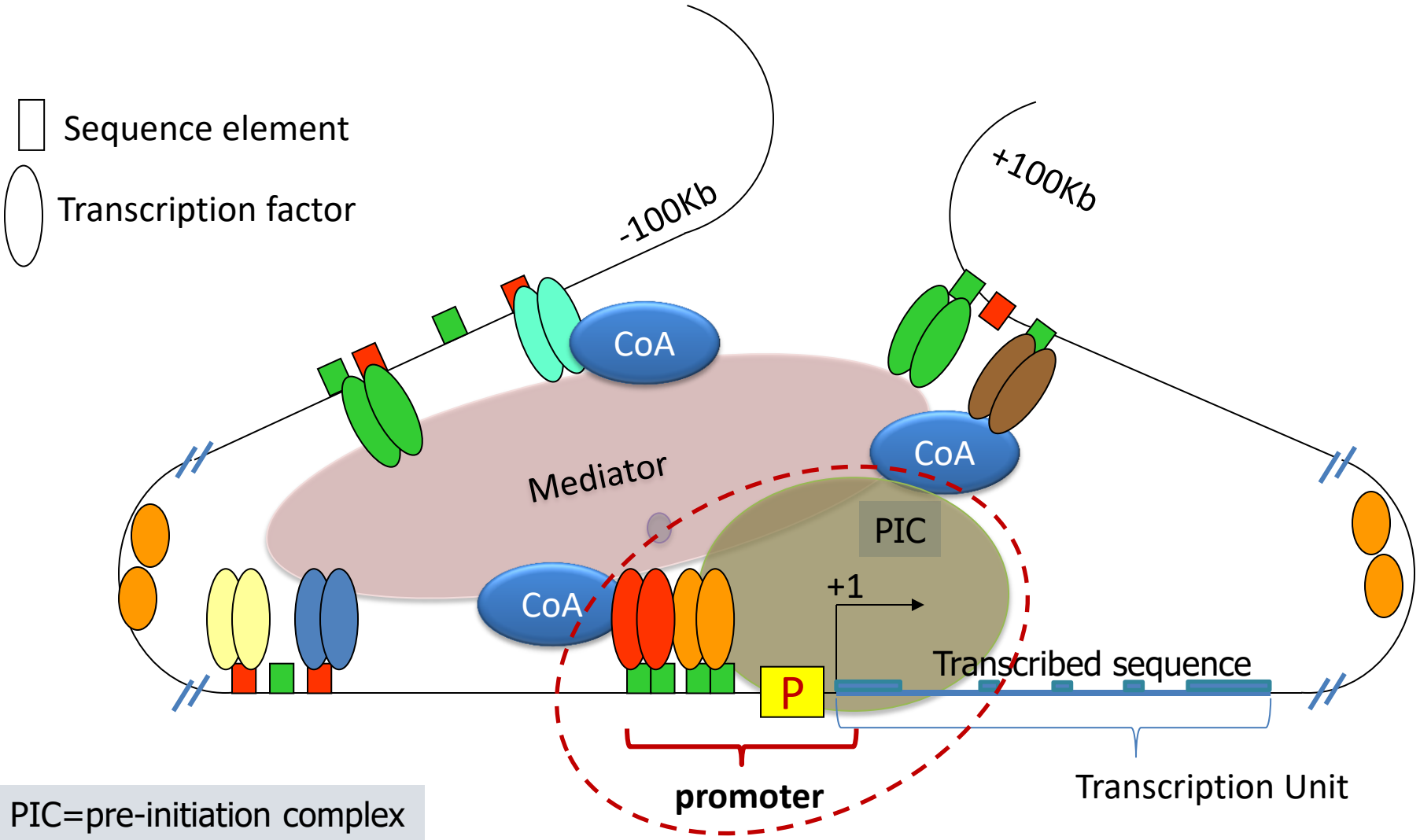
Enhancers: regulatory sequences or «modules» laying virtually at any distance and position from the regulated («cognate») TSS or promoter. Note: even though «enhancer» means «something that increases», enhancers may display repressing activity.

Minimal or «core» promoters are defined as the region bound by **General Transcription Factors** and RNA Polymerase, that is roughly -40 to +40 bp in respect to TSS.

Essentially, it is the region footprinted by RNA Pol II and GTFs.

Normally however the promoter is accompanied by a proximal regulatory region, that Aa place somewhere at -1,000 to +100 bp respect the TSS.

Schematics of eukaryotic gene regulatory sequences and proteins



PIC=pre-initiation complex

Regulatory modules (enhancers, proximal regulatory elements, etc.)

DNA segments where short sequence motifs, 4 to 15 base long, called Response Elements and recognized by **Transcription Factors** are juxtapposed.

Response Elements = **TFBS** (Transcription Factor Binding Sites)