Transcription Factor Footprints Cover Much of the Human Genome

October 22, 2012

GMR Deyter

The sequenced human genome contains a wealth of information that must be deciphered to provide a comprehensive understanding of genes and their regulatory networks. The regulation of transcription is one of the fundamental questions regarding gene expression, and decades of research have achieved many feats, such as discovering the basal transcriptional machinery and defining promoter elements that bind sequence-specific factors. One technique used to define key DNA binding transcriptional regulators is DNAse I footprinting, where the enzyme DNase I is able to cleave DNA hypersensitive sites (DHSs) at regions that are devoid of chromatin-compacting nucleosomes but not at DNA sites bound by proteins. However, a more global analysis of protein binding to cognate DNA regulatory elements is needed to understand the genome-wide breadth of transcription factor binding, cell type-specific transcription factor function, and even the identification of novel binding motifs. The cataloging of such research from many labs encompasses the ENCODE project.

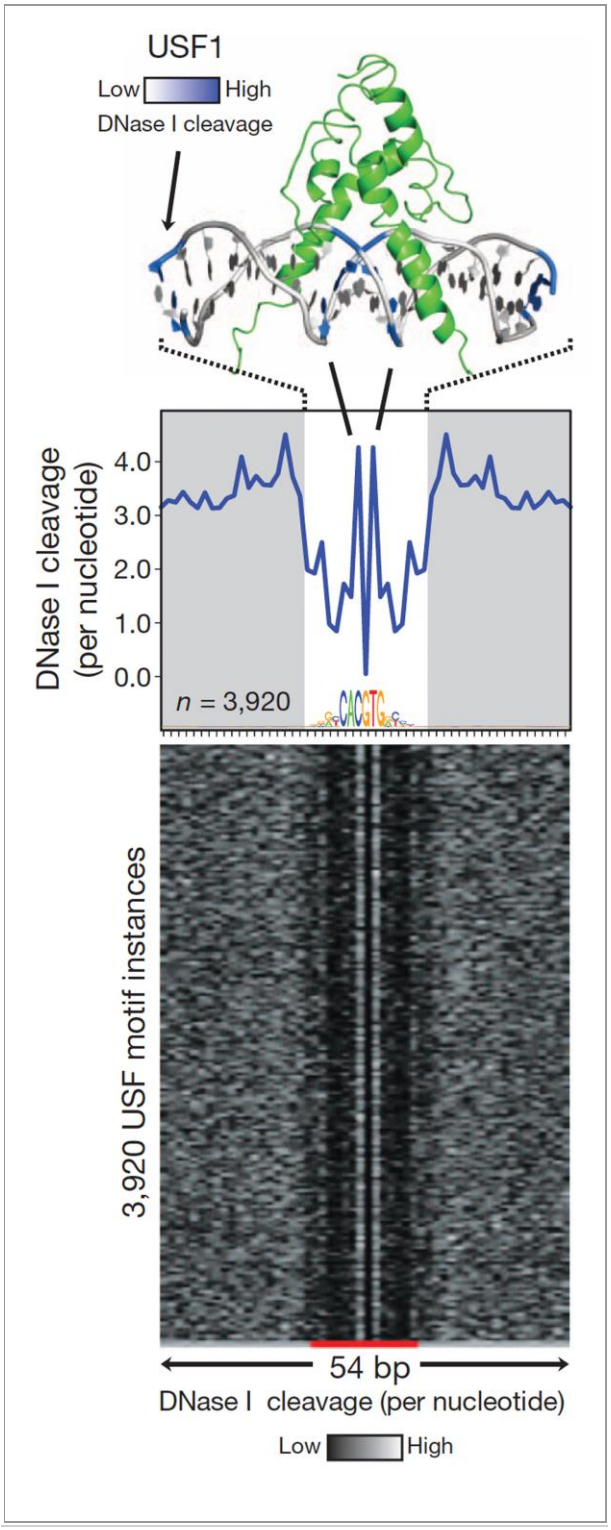
A substantial collaboration between multiple departments at the University of Washington and the Fred Hutchinson Cancer Center (Mark Groudine lab, Basic Sciences Division) was executed to comprehensively define human regulatory DNA elements and associated transcription factor binding. To achieve this, nuclei from 41 different cell and tissue types were isolated and treated with DNase I to generate DNase I cleavage libraries that were analyzed by deep sequencing of DNA. 14.9 billion Illumina sequence reads were obtained, ~ 75% of which mapped to unique genomic locations. A detection bioinformatic algorithm was implemented to define the DNase I footprints, revealing 8.4 million distinct footprinted elements existing throughout the genome. Interestingly, nearly all of the DHSs had at least one footprint, suggesting that DHSs are continually populated with proteins and are rarely nucleosome-free regions.

The analysis of the DNase I footprint sequences revealed striking features of the genomic protein landscape. A comparison of the sequencing results with public databases of transcription factor binding motifs revealed that the DNase I footprints were enriched for known transcription factor sites. A footprint occupancy score (FOS) was computed to determine the magnitude of transcription factor residence within each footprint. The authors computed the FOS for known transcription factor

motifs in the DHSs and established that FOS correlated with transcription factor occupancy identified by ChIP-seq, a method that involves immunoprecipitation of specific transcription factors bound to DNA fragments, followed by sequencing of the bound DNA. FOS also paralleled vertebrate conservation of the core motif region, indicating evolutionary selection on factor occupancy.

To determine if footprint structures parallel transcription factor structure, DNase I cleavage patterns at thousands of specific transcription factor binding motifs were superimposed on the transcription factor's co-crystal structure with DNA. The data clearly show low DNase I cleavages at protein-DNA contacts, and evolutionary conservation of the motif nucleotides that mediate the protein interaction. An alignment of the upstream regions of the transcription start sites of many genes revealed yet another unprecedented high-resolution phenomenon: a conserved ~50 base-pair footprint made by the transcription initiation machinery. The footprint sequencing data also uncovered 289 novel motifs that were found to have a conserved footprint signature in mouse liver tissue. Astoundingly, many of the novel motifs appeared to be occupied in a cell type-specific manner indicating the discovery of recognition sequences for developmentally important but uncharacterized transcriptional regulators. The identification of the novel transcription factors will be an important future endeavor to provide insight on the regulation of the cellular genetic program.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489(7414):83-90.



Adapted from the manuscript.

An example of the exquisite resolution obtained from DNase I mapping of USF1 transcription factor binding motifs. A ribbon diagram of the USF1 (green)/DNA co-crystal structure (top) is superimposed on the DNase I cleavage pattern from nearly four-thousand USF1 genome-wide motifs. Notice that the DNA regions of the USF1 motif that interact with the alpha-helices of USF1 have decreased DNase I cleavage while two medial regions are hypersensitive to DNase I (middle graph). Bottom: A heat map shows the nearly invariant nature of DNase I cleavages (per nucleotide) at USF1 motifs throughout the genome.