The Vast, Interconnected Regulatory Landscape of the Human Genome

October 22, 2012

ME Arnegard

Big questions have emerged from the beginnings of a genomic revolution that is still picking up steam. Completed in 2003, the Human Genome Project identified and mapped roughly 20,000 protein coding genes composing the human 'exome', which is encoded by little more than 1% of our entire genome. This raised the obvious question: How much of the remaining 99% is useless 'junk' DNA *versus* functional elements of a 'regulome' that orchestrates gene regulation? Genome-wide association studies have found that the human exome contains only one in ten single nucleotide polymorphisms (SNPs) associated with the risk of various disease states. How do the remaining nine in ten SNPs underlie disease? A burst of new answers to these fundamental questions have recently been provided by the Encyclopedia of DNA Elements (ENCODE) project, to which Human Biology Division scientists Dr. Muneesh Tewari and Dr. Kavita Garg contributed their talents and expertise in predicting promoters of microRNAs, important regulators of gene expression whose own regulation is poorly characterized.

After completion of a pilot phase that began in 2003, the National Human Genome Research Institute of NIH awarded $80 million to ENCODE consortium scientists in 2007 to better understand how DNA is differentially packaged and regulated in more than 150 human cell lines. As illustrated in the accompanying figure, the 24 standard experimental approaches of ENCODE include: assays for DNaseI hypersensitivity, ENCODE's workhorse technique; next generation sequencing of transcribed RNA (RNAseq); assays of DNA methylation patterns; and sequencing of tiny genomic fragments bound to modified histones or specific transcription factors, as found by chromatin immunoprecipitation (ChIPseq). On September 6th, 2012, the findings of this massive effort were presented in a set of 30 papers simultaneously appearing in *Nature*, *Genome Biology* and *Genome Research*.

The publication to which Drs. Tewari and Garg contributed was directed by ENCODE group leader Dr. John Stamatoyannopoulos (Department of Genome Sciences, University of Washington) and first-authored by Dr. Robert Thurman (Senior Scientist in the Stamatoyannopoulos Lab). This particular piece of the ENCODE project aimed to construct genome-wide maps of active regulatory regions in 125 different cell types by identifying their telltale signatures of DNaseI hypersensitive sites (DHSs), which have guided the discovery of all types of *cis*-regulatory elements including

promoters, enhancers, silencers and locus control regions. The resulting cell-specific maps contained nearly 3 million noncoding sites exposed to regulatory factors, including almost all previously identified*cis*-regulatory sequences and a vast landscape of new elements, most of which are highly specific to different cell types.

Thurman *et al.* were able to annotate DHSs into categories such as repeat DHSs or DHSs corresponding to promoters of protein-coding genes, but they also detected numerous DHSs far from any protein-coding genes. The authors suspected that some of these remote DHSs might correspond to promoters for microRNAs. Given the prior success of the Tewari Lab in predicting promoters of microRNAs which are overexpressed in ovarian carcinoma (Knouf *et al.*, 2011), the ENCODE group turned to Drs. Tewari and Garg for help in identifying transcription start sites (TSS) for all known human microRNAs. Of the 329 such TSS that were identified, 300 were found to be less than 500 base pairs (bp) from a DHS. Thurman *et al.* also found that some microRNA promoters resided in the accessible chromatin compartments of only a few cell types whereas others were localized in the accessible compartments of nearly all cell types and, thus, appeared to be constitutively active (see Fig. 1c of the original *Nature*Article). These important contributions from the Tewari Lab helped Thurman *et al.* annotate DHSs into a category for microRNA promoters, as well as a category for distal DHSs not associated with microRNAs or protein-coding genes (*i.e.*, intergenic DHSs located at least 10,000 bp away from a TSS).

In addition to providing spatially precise maps of all categories of accessible chromatin in numerous cell types, Thurman *et al.*demonstrate that the binding of key transcription factors in place of canonical nucleosomes largely drives chromatin accessibility. Their work similarly advances our understanding of the connection between DNA methylation and DNA accessibility. Moreover, the authors link over 500,000 distal DHSs (*i.e.*, candidate enhancers) to the promoters with which they become synchronously hypersensitive to DNaseI, and they demonstrate elevated mutation rates in the accessible chromatin compartment of cells with high proliferative potential, such as malignant cancer cells.

The findings of Thurman and his many co-authors, together with reports from other ENCODE consortium groups, bring us closer to solving a major conundrum facing genetics: How can the enormous phenotypic differences seen between humans and chimpanzees be encoded by nucleotide differences comprising a mere 1.2% of our genome? Although divergence in protein-coding genes is clearly a part of the solution to this mystery, new results from the ENCODE consortium suggest that the answer will also largely involve the combinatorial power of the regulome for economically orchestrating profound differences in cell fate and organismal phenotype.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature*489:75-82.

Also see: Knouf EC, Garg K, Arroyo JD, Correa Y, Sarkar D, Parkin RK, Wurz K, O'Briant KC, Godwin AK, Urban ND, Ruzzo WL, Gentleman R, Drescher CW, Swisher EM, Tewari M. 2011. An integrative genomic approach identifies p73 and p63 as activators of miR-200 microRNA family transcription. *Nucleic Acids Res.* 40:499-510.

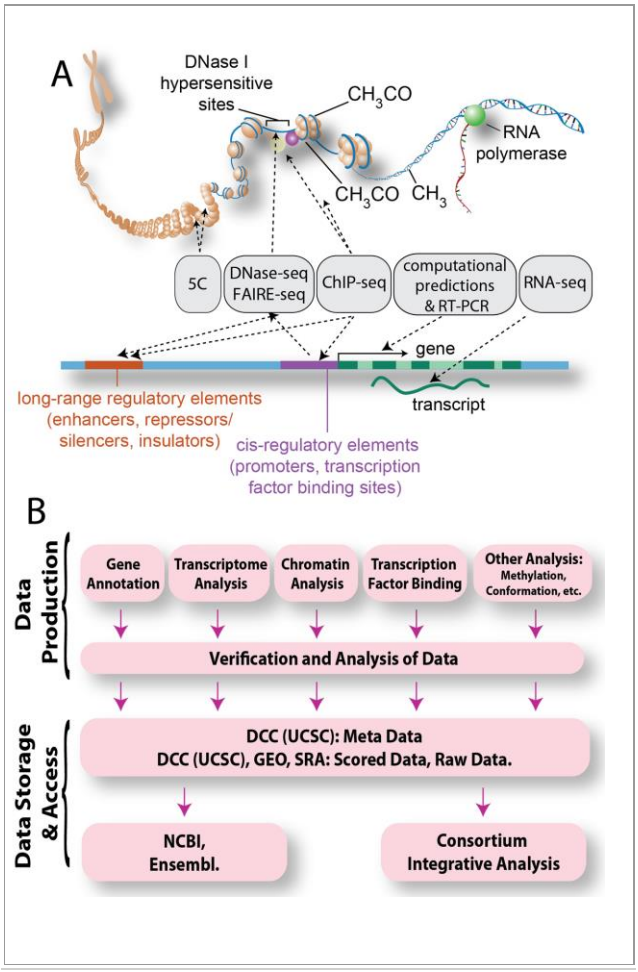

Image courtesy of Darryl Leja (NHGRI) and Ian Dunham (EBI)

Overview of the Encyclopedia of DNA Elements (ENCODE) project. (A) Genomic elements that are the targets of the ENCODE project (dashed arrows) and some of the methods (gray rounded boxes) used to quantify them in more than 150 human cell lines. (B) Summary of ENCODE's workflow. To gain access to human ENCODE data, navigate to the UCSC Genome Browser, select the February 2009 assembly, and jump to your genomic region of interest. ENCODE data can be found in the 'Expression and Regulation' and the 'Mapping, Genes, and Variation' track groups.