Decoding the human genome

Kelly A. Frazer[1]

Moores UCSD Cancer Center, Department of Pediatrics and Rady Children's Hospital, University of California at San Diego, La Jolla, California 92093, USA

Interpreting the human genome sequence is one of the major scientific endeavors of our time. In February 2001, when the human genome reference sequence was initially released (Lander et al. 2001), our understanding of the encoded contents was surprisingly limited. It was perplexing to many in the scientific community when we realized that the human genome contains only ~21,000 distinct protein-coding genes (Claverie 2001; Hollon 2001; Pennisi 2003; Clamp et al. 2007), as other less complex species like the nematode *Caenorhabditis elegans* were known to have a similar number of protein-coding genes (Hillier et al. 2005). It quickly became apparent that the developmental and physiological complexity of humans would not be explained solely by the number of protein-coding genes, and the quest to understand the contents of the human genome began full force.

The Encyclopedia of DNA Elements (ENCODE) Project was launched in September of 2003 with the daunting task of identifying all the functional elements encoded in the human genome sequence. To accomplish this task, the National Human Genome Research Institute (NHGRI) organized The ENCODE Project Consortium, which consists of an international group of scientists with diverse expertise in experimental and computational methods for generating and analyzing high-throughput genomic data (The ENCODE Project Consortium 2004). During the initial four years, the consortium conducted a pilot project which focused on annotating functional elements in a defined 1% of the human genome consisting of ~30 Mb divided among 44 genomic regions. On June 14, 2007, a report summarizing the findings of the pilot project revealed pervasive transcription of the human genome, with the majority of nucleotides represented in transcripts in at least a limited number of cell types at some time (The ENCODE Project Consortium 2007). Many of these transcripts comprised novel noncoding RNA genes. Importantly, The ENCODE Pilot Project assigned function to 60% of the evolutionarily constrained bases in the 44 genomic regions and identified many additional functional elements seemingly unconstrained across mammalian evolution. Integration of the various experimental data generated by The ENCODE Pilot Project provided further insights into connections between chromatin structure (modifications and accessibility) and gene expression (The ENCODE Project Consortium 2007; Koch et al. 2007; Thurman et al. 2007; Zhang et al. 2007) and the timing of replication (Karnani et al. 2007).

Armed with increased knowledge about the types of functional elements contained within the human genome sequence and with the advent of massively parallel sequencing, in 2007 the ENCODE project was expanded to study the entire human genome. This month, *Nature* published a paper entitled "An integrated encyclopedia of DNA elements in the human genome," which reports the production and initial analysis of 1640 data sets focused on two major classes of annotations: genes (both coding and noncoding) along with their corresponding RNA transcripts, and transcriptional regulatory regions. This paper (The ENCODE Project Consortium 2012), along with companion papers in *Nature*, *Genome Research*, and *Genome Biology*, provides much more than a mere inventory of sequence elements but rather presents an integrated analysis providing important insights into the functional organization of the human genome. In alignment with the tradition of large consortia sponsored by the NHGRI, the ENCODE project has made all data and derived results available through a freely accessible database (Rosenbloom et al. 2010).

The following sections describe some of the highlights of the ENCODE project, including technical accomplishments, high quality data sets, and integrated analyses with other resources, such as disease-associated variants identified through genome-wide association studies.

Methodology

ENCODE has placed emphasis on developing technologies and generating data using standardized guidelines for each type of assay (The ENCODE Project Consortium 2011). As an example, ENCODE has performed thousands of individual chromatin immunoprecipitation (ChIP) reactions followed by high-throughput sequencing (ChIP-seq), and through these experiences, the investigators have developed a set of guidelines addressing antibody validation, experimental replication, sequence depth, data reporting, and data quality assessment (Landt et al. 2012). For all ENCODE data types, standardized guidelines are established and protocol descriptions are available at http://www.encodeproject.org/. Importantly, these guidelines enable investigators to readily generate additional data sets to compare with ENCODE data of the same type as well as perform integrative analyses across multiple data types.

Genes and transcripts

The ENCODE project has produced a reference gene set referred to as GENCODE (Harrow et al. 2012), which has been constructed by a merge of manual and automated annotations of all human protein-coding genes, pseudogenes, and noncoding RNAs, including all splice isoforms. The quality of the gene and transcript models in GENCODE are excellent, with the majority of exon-exon junctions experimentally examined validated (Howald et al. 2012). To develop a comprehensive RNA expression catalog, ENCODE sequenced RNA (RNA-seq) from different cell lines (Djebali et al. 2012) and multiple subcellular fractions. Analysis of the RNA-seq data for RNA editing determined that most sequence variants are A-to-G(I) edits located in introns and UTRs, with only a fraction of sites reproducibly edited across multiple cell lines; however, there is an association between editing and specific genes, suggesting that the editing of a transcript is more important than the editing of any individual site (Park et al. 2012). Interestingly, there are ~9500 long noncoding RNAs (lncRNAs) in GENCODE that, by and large, are not translated (Bánfai et al. 2012), biased toward two-exon transcripts, and predominately localized in the chromatin and nucleus; one-third of these RNAs

[1]Corresponding author
E-mail kafrazer@ucsd.edu
Article is at http://www.genome.org/cgi/doi/10.1101/gr.146175.112.
Freely available online through the *Genome Research* Open Access option.

appear to have arisen in the primate lineage (Derrien et al. 2012). Deep sequencing of the subcellular RNA fractions shows that splicing occurs predominantly during transcription, but for alternate exons and lncRNAs, splicing tends to occur later, and in some cases, lncRNAs remain unspliced (Tilgner et al. 2012). Although ENCODE has made considerable headway in annotating and understanding noncoding RNAs, clearly, additional novel functions will be discovered by further analysis of this diverse class of molecules.

Regulatory regions

A diverse group of regulatory elements (promoters, enhancers, and insulators) work collectively to modulate gene expression in a temporal and spatial-specific manner. The ENCODE project has used multiple approaches to identify and characterize regulatory regions. As regulatory regions have an open chromatin structure, their sequences are hypersensitive to DNase I. ENCODE generated DNase-seq data from 19 diverse cell types and analyzed it in combination with expression data to identify proximal and distal regulatory elements at a genome-wide scale (Natarajan et al. 2012).To directly identify regulatory regions, the positions of 119 different DNA binding proteins and a number of RNA polymerase components were mapped in 72 cell types using ChIP-seq (Gerstein et al. 2012). Additionally, the chromosomal locations of 12 histone modifications correlated with distinct classes of regulatory elements and/or transcription states were determined. Several integrative analyses of these data were performed by ENCODE, demonstrating that the locations of DNA binding proteins overlap with GC-rich, DNase I sensitive regions (Arvey et al. 2012; Cheng et al. 2012; J Wang et al. 2012) and resulting in a catalog of cell-type-specific regulatory regions. A clustering procedure, termed the Clustered Aggregation Tool (CAGT), revealed extensive heterogeneity in how histone modifications are deposited and how nucleosomes are positioned around protein binding sites, suggesting that most regulatory elements have directionality (Kundaje et al. 2012). ENCODE also demonstrated that there can be cell-selective regulation of CTCF occupancy at binding sites (H Wang et al. 2012); previously, it was widely believed that binding patterns of this ubiquitously expressed regulatory factor were largely invariant. These analyses, as well as additional data and derived results provided by ENCODE (The ENCODE Project Consortium 2012), offer insights into how distinctive functional domains in the genome of a cell are associated with specific chromatin states. Ultimately, each of the several hundred human cell types will be examined using these approaches and those data layered atop the human genome sequence; from this comprehensive regulatory map, we will have a detailed description of the genomic physiological differences and similarities across all cell types.

Integration of ENCODE data with human genetic variation

The ENCODE project also explored the potential functional effects of individual variants in regulatory and protein-coding regions. An analysis of 69 fully sequenced genomes using RegulomeDB, a database which integrates the large collection of ENCODE regulatory information as well as other data, assigned thousands of variants to potential regulatory functions and indicated that there is at least as much variation affecting regulatory function as affecting gene function (Boyle et al. 2012). Likewise, a combined analysis of regulatory regions marked by DNase I hypersensitivity sites and the whole-genome sequences of 53 individuals suggests that individuals likely contain more functionally important regulatory variants than protein-coding variants (Vernot et al. 2012). It is important to note that, although there appear to be more regulatory variants in the human genome than coding variants, they are likely to have, on average, smaller effect sizes. Additionally, a combined analysis of multiple types of ENCODE data and the ~4800 human variants that have been associated through GWAS with human diseases and/or phenotypes enabled putative functional annotations for up to 80% of all previously reported associations (Schaub et al. 2012). These results provide a glimpse into the future, as the combined analysis of ENCODE and other large genomic data sets will likely, over the course of the next decade, provide insights into the genetic and epigenetic factors underlying the development and progression of human diseases such as heart disease, diabetes, cancer, and mental illnesses.

Beyond this current release of the ENCODE project, coordinated efforts will need to continue in order to understand all of the functional elements contained within the human genome. These efforts should continue to examine the ubiquitous transcription of the human genome to uncover the multitude of different roles that noncoding RNAs are likely to play. They should expand the number of DNA binding proteins and histone modifications studied and, importantly, also examine a greater number of cell types under a variety of experimental conditions. As these immense data sets continue to be generated, we will eventually be able to identify all of the functional elements in the human genome and be better armed to predict how natural and disease-related genetic variation influences their function. Although there is much that we do not understand at this point in time, one thing is perfectly clear—there will be many, many more unanticipated and exciting findings that will emerge as we continue to decode the human genome.

References

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res* (this issue). doi: 10.1101/gr.127712.111.

Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling W, Kundaje A, Gunawardena HP, Yu Y, Xie L, et al. 2012. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* (this issue). doi: 10.1101/gr.134767.111.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* (this issue). doi: 10.1101/gr.137323.112.

Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan K-K, Dong X, Djebali S, Ruan Y, et al. 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* (this issue). doi: 10.1101/gr.136838.111.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104:** 19428–19433.

Claverie JM. 2001. Gene number. What if there are only 30,000 human genes? *Science* **291:** 1255–1257.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (this issue). doi: 10.1101/gr.132159.111.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* (in press).

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* (in press).
Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* (in press).
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* (this issue). doi: 10.1101/gr.135350.111.
Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in *C. elegans*: So many genes, such a little worm. *Genome Res* **15:** 1651–1660.
Hollon T. 2001. Human genes: How many? *Scientist* **15:** 1.
Howald C, Tanzer A, Chrast J, Kokocinski F, Derrien T, Walters N, Gonzalez JM, Frankish A, Aken BL, Hourlier T, et al. 2012. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res* (this issue). doi: 10.1101/gr.134478.111.
Karnani N, Taylor C, Malhotra A, Dutta A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res* **17:** 865–876.
Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, et al. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17:** 691–707.
Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* (this issue). doi: 10.1101/gr.136366.111.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* (this issue). doi: 10.1101/gr.136184.111.
Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type–specific gene expression from regions of open chromatin. *Genome Res* (this issue). doi: 10.1101/gr.135129.111.
Park E, Williams B, Wold B, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res* (this issue). doi: 10.1101/gr.134957.111.
Pennisi E. 2003. Human genome. A low number wins the GeneSweep Pool. *Science* **300:** 1484.
Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* **38:** D620–D625.
Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* (this issue). doi: 10.1101/gr.136127.111.
Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* **17:** 917–927.
Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* (this issue). doi: 10.1101/gr.134445.111.
Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM. 2012. Personal and population genomics of human regulatory variation. *Genome Res* (this issue). doi: 10.1101/gr.134890.111.
Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. 2012. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* (this issue). doi: 10.1101/gr.136101.111.
Wang J, Zhuang J, Iyer S, Lin XY, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* (this issue). doi: 10.1101/gr.139105.112.
Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein MB. 2007. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res* **17:** 787–797.