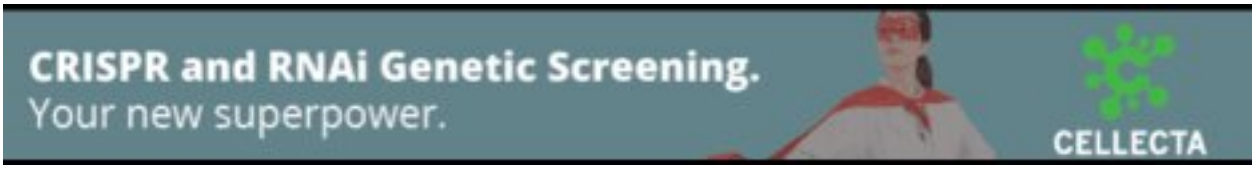
ENCODE: More genomic empowerment

George M. Weinstock

Genome Res. 2007 17: 667-668
Access the most recent version at doi:10.1101/gr.6534207



Commentary

ENCODE: More genomic empowerment

George M. Weinstock[1]

Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

The completion of The ENCODE (Encyclopedia of DNA Elements) Pilot Project marks an important step forward in enriching the human genome sequence with functional information. The Human Genome Project (HGP) produced a sequence that was annotated for potential genes and other genomic features (Lander et al. 2001; Venter et al. 2001; The International Human Genome Sequencing Consortium 2004). This watershed advance was nevertheless only a first step in "whole-genome" approaches to progressing the understanding of human biology. Numerous other genomes that were subsequently sequenced sharpened the definition of key genomic components under evolutionary constraint: presumed functional elements such as protein-coding genes and regulatory regions. The International HapMap Project also probed the nature and structure of the variation of the human sequence, providing additional dimensionality (The International HapMap Consortium 2003, 2005). Now, the ambitious ENCODE project (The ENCODE Project Consortium 2004, 2007) adds to these by addressing many more types of sequence-based function than ever before, further connecting the genetic code to the mechanistic biochemical understanding of human biology.

The task of the ENCODE Project is formidable. The goal is to map a variety of sequence elements including genes, promoters, enhancers, repressor or silencer sequences, exons, replication origin and termination sites, transcription factor binding sites, methylation sites, DNase I hypersensitive sites, chromatin modifications, conserved sequences, and RNA transcripts, to name only those considered in the pilot project. Moreover, given the differences in the occupancy of binding sites or types of modifications in each tissue, during different developmental stages, and in response to environmental stimuli and other states that control the genome, elucidating the status of each element in each state is a mammoth undertaking. In the ENCODE pilot project (http://www.genome.gov/10005107), this diversity of sequence elements was mapped by a consortium of 35 groups engaged in producing high-throughput mapping data as well as developing technology and computational approaches in anticipation of extending the project to the whole genome. The pilot project limited attention to 1% (~30 Mb) of the genome, represented by 44 regions selected by diverse criteria such as containing well-studied genes and other elements and the presence of comparative sequence data, and also included a sampling of different gene densities and levels of non-exonic conservation. The aim was mainly to enumerate sequence elements, and not determine or validate their function. In this sense ENCODE follows the now familiar strategy of producing a community resource, a large-scale data set consisting of a list of features and their status in different tissues or cellular states, that the research community can use to answer functional and mechanistic questions.

Notwithstanding this philosophy, the results from the ENCODE pilot project have been anything but a mere inventory. Demonstration of the data quality and utility of the ENCODE approach required not only production of large data sets but also their integration to find correlations between the differing data types. The consortium produced >200 data sets, representing >400 million data points and 200 Mb of comparative sequence. Rapid data release guidelines were followed, and the principal data repository (http://genome.ucsc.edu/ENCODE) addressed the challenges of integrating disparate experimental results. The conclusions from the project underscore the success in this effort.

Numerous themes have emerged or been sharpened by this limited foray into the human functional genome. The project found that nearly the entire genome may be represented in primary transcripts that extensively overlap and include many non-protein-coding regions. The idea of a network of transcripts has been suggested before (e.g., Cheng et al. 2005; Carninci et al. 2006), but data from the ENCODE project provide firmer footing for further investigation of this challenge to the concept of lone transcription units (Denoeud et al. 2007; Emanuelsson et al. 2007; Rozowsky et al. 2007; Ruan et al. 2007; Trinklein et al. 2007). New transcription start sites (TSSs) were identified, and the arrangement of regulatory sequences (and binding of transcription factors) around TSSs was more broadly described to show the range of locations (Denoeud et al. 2007; Trinklein et al. 2007; Xi et al. 2007; Zhang et al. 2007). Some enhancers that previously mapped distal to the known TSSs were found by the ENCODE project to be near the newly described TSSs and transcripts, suggesting a role in regulating proximal expression. A richer view of the connections between chromatin structure, regulation of transcription, and replication has emerged from integrating these data sets. For example, earlier knowledge of the types of histone modifications that correlate with gene expression has been amplified to create a predictor of expression based on these modifications and chromatin accessibility (The ENCODE Project Consortium 2007; Koch et al. 2007; Thurman et al. 2007; Zhang et al. 2007), and correlations between the timing of replication and chromatin structure were also described (Karnani et al. 2007). Another result from integrating data sets shows that ~60% of the bases found to be under evolutionary constraint in genome comparisons are correlated with functional sites identified by the project's experimental approaches (Margulies et al. 2007). One anticipates this proportion will increase as binding sites for additional transcription factors and other elements are mapped.

The ENCODE project has enriched the annotation of the human DNA sequence by describing the functional elements encoded therein. But has the pilot project demonstrated that it is possible to build an encyclopedia of sequence elements and correlate biochemical functions? The 60% of constrained sequences that map to functional elements suggests that this will be possible, but more (and different) elements need to be mapped experimentally to account for the remaining 40%. Will there be predictive value in knowing for each region of the genome the encyclopedia entries and state of the chromatin? The initial results from this pilot project are encouraging, with demonstra-

[1]Corresponding author.
E-mail gwstock@bcm.tmc.edu; fax (713) 798-4373.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.6534207.
Freely available online through the *Genome Research* Open Access option.

tions of such correlations for gene expression or replication timing. How might these ENCODE data and their correlations be applied? Certainly one class of genes, perhaps the major class at present, that could benefit is those of unknown function. Future studies of a gene of interest with unknown function can be informed by the ENCODE data about that genomic region, complete with locations of transcripts, TSSs, regulatory factor binding sites, and chromatin structural information that should be considered. These may include factors that have been previously shown to be mainly regulating transcripts in a specific tissue. Hence, one may begin to regard the gene of interest as having a role in that tissue—guilt by association—and a hint about function. So just as the human genome sequence provided a gene list with only computational prediction of function, and the HapMap project provided alleles and haplotypes without connection to phenotype, so the ENCODE project provides a catalog of genome features, which provides tantalizing clues without directly linking to function. In all of these projects, the data sets empower the research community to engage in deeper analyses of sequence elements that will ultimately lead to biologically significant insights.

Beyond the pilot project is the scale-up of the ENCODE project. This extension of the pilot effort will increase the data set in multiple dimensions: Not only will the full genome landscape be studied, but also the range of the study will expand by increasing the number of factors and modifications, as well as the number of tissues and states to be considered. With these extensive lists of transcription factor binding sites, chromatin modifications, origins of replication, and other elements, function-oriented researchers can now begin to enjoy "drinking from the fire hose," as sequence-oriented informaticians have since the HGP ended.

Acknowledgments

I thank the stimulating intellectual environment of the Baylor College of Medicine Human Genome Sequencing Center and the generous support from the NIH-NHGRI.

References

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626–635.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Chrast, J., et al. 2007. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* (this issue) doi: 10.1101/gr.5660607.

Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, J.S., Urban, A.E., Du, J., Lian, Z., Stole, V., Weissman, S., Snyder, M., et al. 2007. Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res.* (this issue) doi: 10.1101/gr.5014606.

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* (in press).

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

The International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Karnani, N., Taylor, C., Malhotra, A., and Dutta, A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res.* (this issue) doi: 10.1101/gr.5427007.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S., Karaöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., et al. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* (this issue) doi: 10.1101/gr.5704207.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., 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.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* (this issue) doi: 10.1101/gr.6034307.

Rozowsky, J., Newburger, D., Sayward, F., Wu, J., Jordan, G., Korbel, J.O., Nagalakshmi, U., Yang, J., Zheng, D., Guigo, R., et al. 2007. The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res.* (this issue) doi: 10.1101/gr.5696007.

Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* (this issue) doi: 10.1101/gr.6018607.

Thurman, R.E., Day, N., Noble, W.S., and Stamatoyannopoulos, J.A. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* (this issue) doi: 10.1101/gr.6081407.

Trinklein, N.D., Karaöz, U., Wu, J., Halees, A., Force Aldred, S., Collins, P.J., Zheng, D., Zhang, Z.D., Gerstein, M., Snyder, M., et al. 2007. Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res.* (this issue) doi: 10.1101/gr.5716607.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., and Weng, Z. 2007. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.* (this issue) doi: 10.1101/gr.5754707.

Zhang, Z.D., Paccanaro, A., Fu, Y., Weissman, S., Weng, Z., Chang, J., Snyder, M., and Gerstein, M. 2007. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.* (this issue) doi: 10.1101/gr.5573107.