The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

The ENCyclopedia Of DNA Elements (ENCODE) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

With the complete human genome sequence now in hand (*1*–*3*), we face the enormous challenge of interpreting it and learning how to use that information to understand the biology of human health and disease. The ENCyclopedia Of DNA Elements (ENCODE) Project is predicated on the belief that a comprehensive catalog of the structural and functional components encoded in the human genome sequence will be critical for understanding human biology well enough to address those fundamental aims of biomedical research. Such a complete catalog, or "parts list," would include protein-coding genes, non–protein-coding genes, transcriptional regulatory elements, and sequences that mediate chromosome structure and dynamics; undoubtedly, additional, yet-to-be-defined types of functional sequences will also need to be included.

To illustrate the magnitude of the challenge involved, it only needs to be pointed out that an inventory of the best-defined functional components in the human genome—the protein-coding sequences—is still incomplete for a number of reasons, including the fragmented nature of human genes. Even with essentially all of the human genome sequence in hand, the number of protein-coding genes can still only be estimated (currently 20,000 to 25,000) (*3*). Non–protein-coding genes are much less well defined. Some, such as the ribosomal RNA and tRNA genes, were identified several decades ago, but more recent approaches, such as cDNA-cloning efforts (*4*, *5*) and chip-based transcriptome analyses (*6*, *7*), have revealed the existence of many transcribed sequences of unknown function. As a reflection of this complexity, about 5% of the human genome is evolutionarily conserved with respect to rodent genomic sequences, and therefore is inferred to be functionally important (*8*, *9*). Yet only about one-third of the sequence under such selection is predicted to encode proteins (*1*, *2*). Our collective knowledge about putative functional, noncoding elements, which represent the majority of the remaining functional sequences in the human genome, is remarkably underdeveloped at the present time.

An added level of complexity is that many functional genomic elements are only active or expressed in a restricted fashion—for example, in certain cell types or at particular developmental stages. Thus, one could envision that a truly comprehensive inventory of functional elements might require high-throughput analyses of every human cell type at all developmental stages. The path toward executing such a comprehensive study is not clear and, thus, a major effort to determine how to conduct such studies is warranted.

To complement ongoing large-scale projects that are contributing to the identification of conserved elements in the human genome (*10*) (www.intlgenome.org) and to the isolation and characterization of human full-length cDNAs (*11*), the ENCODE Project (www.genome.gov/ENCODE) was launched to develop and implement high-throughput methods for identifying functional elements in the human genome. ENCODE is being implemented in three phases—a pilot phase, a technology development phase, and a production phase. In the pilot phase, the ENCODE Consortium (see below) is evaluating strategies for identifying various types of genomic elements. The goal of the pilot phase is to identify a set of procedures that, in combination, can be applied cost-effectively and at high-throughput to accurately and comprehensively characterize large regions of the human genome. The pilot phase will undoubtedly reveal gaps in the current set of tools for detecting functional sequences, and may reveal that some methods being used are inefficient or unsuitable for large-scale utilization. Some of these problems should be addressed in

*Affiliations for all members of the ENCODE Consortium can be found on *Science* Online at www.sciencemag.org/cgi/content/full/306/5696/636/DC1.
†To whom correspondence should be addressed. E-mail: elise_feingold@nih.gov
ENCODE Project Scientific Management:
National Human Genome Research Institute (E. A. Feingold, P. J. Good, M. S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F. S. Collins).
Initial ENCODE Pilot Phase Participants:
Affymetrix, Inc. (T. R. Gingeras, D. Kampa, E. A. Sekinger, J. Cheng, H. Hirsch, S. Ghosh, Z. Zhu, S. Patel, A. Piccolboni, A. Yang, H. Tammana, S. Bekiranov, P. Kapranov, R. Harrison, G. Church, K. Struhl); Ludwig Institute for Cancer Research (B. Ren, T. H. Kim, L. O. Barrera, C. Qu, S. Van Calcar, R. Luna, C. K. Glass, M. G. Rosenfeld); Municipal Institute of Medical Research (R. Guigo, S. E. Antonarakis, E. Birney, M. Brent, L. Pachter, A. Reymond, E. T. Dermitzakis, C. Dewey, D. Keefe, F. Denoeud, J. Lagarde, J. Ashurst, T. Hubbard, J. J. Wesselink, R. Castelo, E. Eyras); Stanford University (R. M. Myers, A. Sidow, S. Batzoglou, N. D. Trinklein, S. J. Hartman, S. F. Aldred, E. Anton, D. I. Schroeder, S. S. Marticke, L. Nguyen, J. Schmutz, J. Grimwood, M. Dickson, G. M. Cooper, E. A. Stone, G. Asimenos, M. Brudno); University of Virginia (A. Dutta, N. Karnani, C. M. Taylor, H. K. Kim, G. Robins); University of Washington (G. Stamatoyannopoulos, J. A. Stamatoyannopoulos, M. Dorschner, P. Sabo, M. Hawrylycz, R. Humbert, J. Wallace, M. Yu, P. A. Navas, M. McArthur, W. S. Noble); Wellcome Trust Sanger Institute (I. Dunham, C. M. Koch, R. M. Andrews, G. K. Clelland, S. Wilcox, J. C. Fowler, K. D. James, P. Groth, O. M. Dovey, P. D. Ellis, V. L. Wraight, A. J. Mungall, P. Dhami, H. Fiegler, C. F. Langford, N. P. Carter, D. Vetrie); Yale University (M. Snyder, G. Euskirchen, A. E. Urban, U. Nagalakshmi, J. Rinn, G. Popescu, P. Bertone, S. Hartman, J. Rozowsky, O. Emanuelsson, T. Royce, S. Chung, M. Gerstein, Z. Lian, J. Lian, Y. Nakayama, S. Weissman, V. Stolc, W. Tongprasit, H. Sethi).
Additional ENCODE Pilot Phase Participants:
British Columbia Cancer Agency Genome Sciences Centre (S. Jones, M. Marra, H. Shin, J. Schein); Broad Institute (M. Clamp, K. Lindblad-Toh, J. Chang, D. B. Jaffe, M. Kamal, E. S. Lander, T. S. Mikkelsen, J. Vinson, M. C. Zody); Children's Hospital Oakland Research Institute (P. J. de Jong, K. Osoegawa, M. Nefedov, B. Zhu); National Human Genome Research Institute/Computational Genomics Unit (A. D. Baxevanis, T. G. Wolfsberg); National Human Genome Research Institute/Molecular Genetics Section (F. S. Collins, G. E. Crawford, J. Whittle, I. E. Holt, T. J. Vasicek, D. Zhou, S. Luo); NIH Intramural Sequencing Center/National Human Genome Research Institute (E. D. Green, G. G. Bouffard, E. H. Margulies, M. E. Portnoy, N. F. Hansen, P. J. Thomas, J. C. McDowell, B. Maskeri, A. C. Young, J. R. Idol, R. W. Blakesley); National Library of Medicine (G. Schuler); Pennsylvania State University (W. Miller, R. Hardison, L. Elnitski, P. Shah); The Institute for Genomic Research (S. L. Salzberg, M. Pertea, W. H. Majoros); University of California, Santa Cruz (D. Haussler, D. Thomas, K. R. Rosenbloom, H. Clawson, A. Siepel, W. J. Kent).
ENCODE Technology Development Phase Participants:
Boston University (Z. Weng, S. Jin, A. Halees, H. Burden, U. Karaoz, Y. Fu, Y. Yu, C. Ding, C. R. Cantor); Massachusetts General Hospital (R. E. Kingston, J. Dennis); NimbleGen Systems, Inc. (R. D. Green, M. A. Singer, T. A. Richmond, J. E. Norton, P. J Farnham, M. J. Oberley, D. R. Inman); NimbleGen Systems, Inc. (M. R. McCormick, H. Kim, C. L. Middle, M. C. Pirrung); University of California, San Diego (X. D. Fu, Y. S. Kwon, Z. Ye); University of Massachusetts Medical School (J. Dekker, T. M. Tabuchi, N. Gheldof, J. Dostie, S. C. Harvey).

the ENCODE technology development phase (being executed concurrently with the pilot phase), which aims to devise new laboratory and computational methods that improve our ability to identify known functional sequences or to discover new functional genomic elements. The results of the first two phases will be used to determine the best path forward for analyzing the remaining 99% of the human genome in a cost-effective and comprehensive production phase.

ENCODE Targets

The defining feature of the ENCODE pilot phase is the uniform focus on a selected 30 Mb of the human genome. Each pilot-phase participant will study the entire set of ENCODE targets—44 discrete regions that together encompass ~1% of the human genome. All approaches will thus be tested on a relatively large amount of genomic sequence, allowing an assessment of the ability of each to be applied at large scale. The use of a single target set will allow the results of different approaches to be directly compared with one another.

The set of ENCODE targets was chosen to represent a range of genomic features (www.genome.gov/10005115). It was first decided that a number of smaller regions (0.5 to 2 Mb) distributed across many different chromosomes should be chosen, as opposed to (for example) a single 30-Mb region. To ensure that existing data sets and knowledge would be effectively utilized, roughly half of the 30 Mb was selected manually. The two main criteria used for the manual selection were as follows: (i) the presence of extensively characterized genes and/or other functional elements; and (ii) the availability of a substantial amount of comparative sequence data. For example, the genomic segments containing the α- and β-globin gene clusters were chosen because of the wealth of data available for these loci (*12*). On the other hand, the region encompassing the *CFTR* (cystic fibrosis transmembrane conductance regulator) gene was selected because of the extensive amount of multispecies sequence data available (*13*). Once the manual selections had been made, the remaining targets were chosen at random by means of an algorithm that ensured that the complete set of targets represented the range of gene content and level of nonexonic conservation (relative to mouse) found in the human genome. The locations and characteristics of the 44 ENCODE target regions (along with additional details about their selection) are available at the UCSC ENCODE Genome Browser (www.genome.ucsc.edu/ENCODE/regions_build34.html).

The ENCODE Consortium

The pilot phase is being undertaken by a group of investigators, the ENCODE Consortium, who are working in a highly collaborative way to implement and evaluate a set of computational and experimental approaches for identifying functional elements. The pilot phase began in September 2003 with the funding of eight projects (table S1) that involve the application of existing technologies to the large-scale identification of a variety of functional elements in the ENCODE targets, specifically genes, promoters, enhancers, repressors/silencers, exons, origins of replication, sites of replication termination, transcription factor binding sites, methylation sites, deoxyribonuclease I (DNase I) hypersensitive sites, chromatin modifications, and multispecies conserved sequences of yet unknown function (Fig. 1). Genetic variation within the conserved sequences is also being determined. The methodological approaches being employed in the pilot phase include transcript and chromatin immunoprecipitation–microarray hybridization (ChIP-chip; see below) analyses with different microarray platforms, computational methods for finding genes and for identifying highly conserved sequences, and expression reporter assays.

In addition to the initial eight, other groups have since joined the ENCODE Consortium. These include groups doing comparative sequencing specifically for ENCODE, groups coordinating databases for sequence-related and other types of ENCODE data, and groups conducting studies on specific sequence elements (table S1). Beyond these current

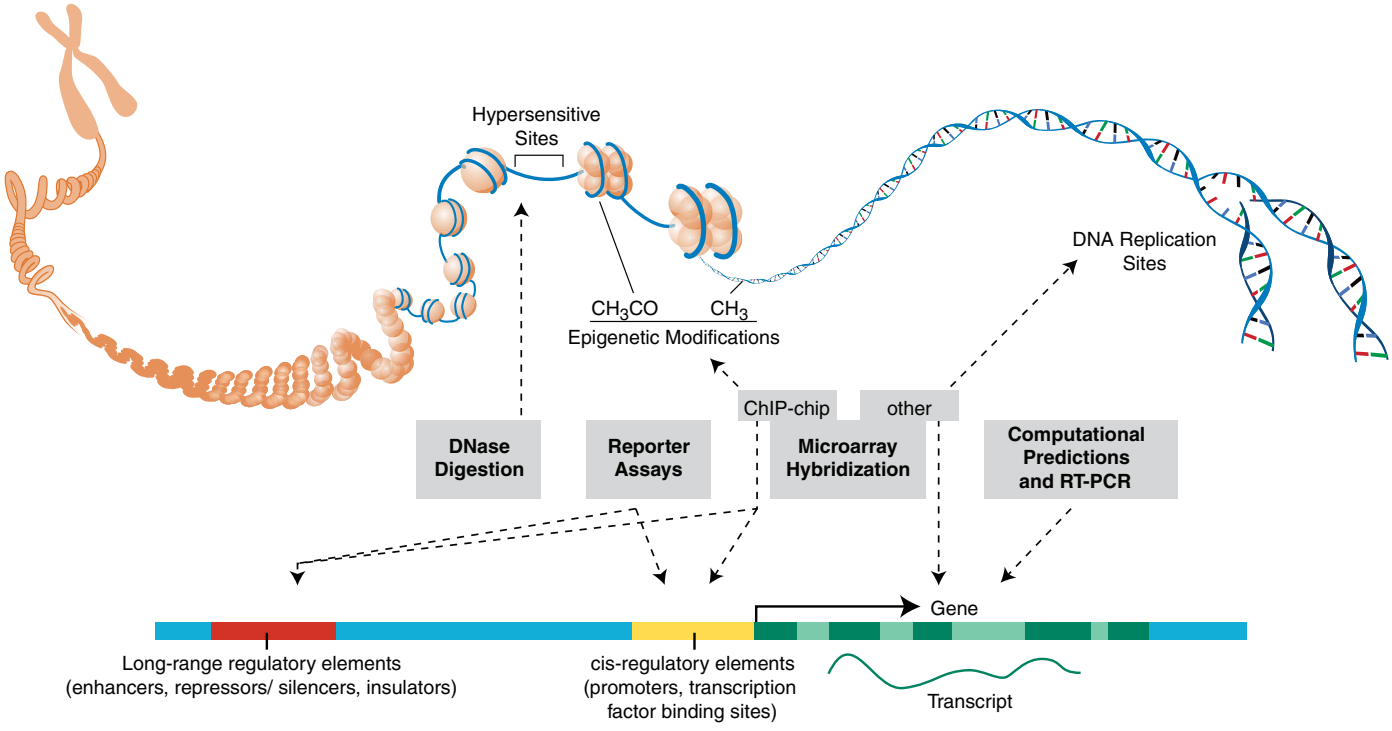

Fig. 1. Functional genomic elements being identified by the ENCODE pilot phase. The indicated methods are being used to identify different types of functional elements in the human genome.

participants, the ENCODE Consortium is open to all interested academic, government, and private-sector investigators, as long as they abide by established Consortium guidelines (www.genome.gov/10006162), which require the commitment to work on the entire set of ENCODE targets, to participate in all Consortium activities, to make a significant intellectual contribution, and to release data in accordance with the policies specifically established for the project (see below).

The parallel technology development phase (table S2) is intended to expand the "tool box" available for high-throughput identification of functional genomic elements, and includes projects to develop new methods both for more efficient identification of known elements and for identification of heretofore unknown elements.

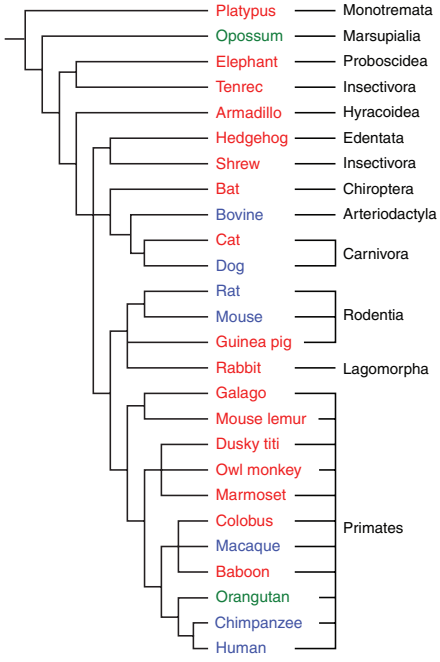

Fig. 2. Mammals for which genomic sequence is being generated for regions orthologous to the ENCODE targets. Genomic sequences of the ENCODE targets are being generated for the indicated mammalian species. The current plans are to produce high-quality finished (blue), comparative-grade finished (red), or assembled whole-genome shotgun (green) sequence, as indicated. High-quality finished reflects highly accurate and contiguous sequence, with a best-faith effort used to resolve all difficult regions (*26*). Comparative-grade finished reflects sequence with greater than eightfold coverage that has been subjected to additional manual refinement to ensure accurate order and orientation of all sequence contigs (*16*). In the case of whole-genome shotgun sequence, the actual coverage and quality may vary. Other vertebrate species for which sequences orthologous to the ENCODE targets are being generated include chicken, frog, and zebrafish (not shown). A complete list of the ENCODE comparative sequencing efforts is provided in table S3.

Research Plans

Each group is using one or more high-throughput approaches to detect a specific genomic element(s). In some cases, multiple platforms are being evaluated in comparable experiments. For example, several types of microarrays [e.g., oligonucleotide arrays made by different technologies, polymerase chain reaction (PCR)–amplicon arrays] are being used to identify transcribed regions. The pilot project participants are primarily using a restricted number of cell lines to identify functional elements. This approach is being taken for practical purposes, but it has the limitation that not all cell types will be surveyed, and therefore some elements with tissue-restricted function may not be identified in the pilot phase. To facilitate comparison of data generated on different platforms and by different approaches, a common set of reagents is being included whenever appropriate. So far, the common reagents chosen include two cell lines (HeLa S3, a cervical adenocarcinoma, and GM06990, an Epstein-Barr virus–transformed B-lymphocyte) and two antibodies [one for the general transcription factor $TAF_{II}250$ (*14*) and another for the inducible transcription factor STAT-1 (*15*)].

The ENCODE pilot phase also includes a component that is generating sequences of the genomic regions that are orthologous to the ENCODE target regions from a large set of nonhuman vertebrates (www.nisc.nih.gov/open_page.html?/projects/encode/index.cgi) (*16*). This will allow ENCODE to identify the quality and amount of comparative sequence data necessary to accurately identify evolutionarily conserved elements, and to develop more powerful computational tools for using comparative sequence data to infer biological function. An initial set of 10 vertebrates have been selected for targeted sequencing on the basis of multiple factors, including phylogenetic position (Fig. 2 and table S3) and the availability of a bacterial artificial chromosome (BAC) library. In addition to this ENCODE-specific effort, comparative sequence data are also being captured from ongoing whole-genome sequencing projects, including those for mouse, rat, dog, chicken, cow, chimpanzee, macaque, frog, and zebrafish. A unique RefSeq accession number (*17*) is being assigned for the sequence of each ENCODE-orthologous target region in each species, with periodic data freezes.

A feature of the evolutionarily conserved elements to be assayed is sequence variation. This will be accomplished by resequencing PCR-amplified fragments from genomic DNA of 48 individuals, the same samples being used by the HapMap Consortium to determine common patterns of genetic variation (*18*). This will result in a quantitative view of the evolutionary constraints on conserved regions.

Data Management and Analysis

Capturing, storing, integrating, and displaying the diverse data generated will be challenging. Data that can be directly linked to genomic sequence will be managed at the UCSC Genome Browser (www.genome.ucsc.edu/ENCODE) (*19*). Other data types will be stored either at available public databases [e.g., the GEO (Gene Expression Omnibus) (www.ncbi.nlm.nih.gov/geo) and ArrayExpress (www.ebi.ac.uk/arrayexpress) sites for microarray data] or on publicly accessible Web sites specifically developed by ENCODE Consortium participants. An ENCODE portal will also be established to index these data, allowing users to query different data types regardless of location. Access to metadata associated with each experiment will be provided. The ENCODE pilot phase will make use of the MAGE standard for representing microarray data (*20*), and data standards for other data types will be developed as needed.

Figure 3 uses the early pilot-phase data for one of the ENCODE target regions (ENr231) to illustrate how the ENCODE data will be presented in a way that will capture the innovation of the ENCODE Project's goal of developing a full representation of functional genomic elements. The results of the different analyses are presented as parallel tracks aligned with the genomic sequence. For example, the gene structures for known genes are shown; future ENCODE data will include the precise locations of 5′ transcription start sites, intron/exon boundaries, and 3′ polyadenylation sites for each gene in the ENCODE targets. In addition, efforts will be made to confirm all gene predictions in these regions. Positions of the evolutionarily conserved regions, as detected by analyzing sequences from several organisms, are shown and can be correlated with results of other studies. This presentation will allow a comprehensive depiction of the results from all of the ENCODE components, enabling both comparisons of different methods and of the reproducibility of the data from different laboratories. The former is illustrated by comparison of two different methods for localizing promoter regions, one based on reporter constructs containing sequences around putative transcription initiation sites (*21*) and the other involving chromatin immunoprecipitation (ChIP) with an antibody to RNA polymerase (RNAP) and hybridization to DNA microarrays (chip) (*22*) (so-called ChIP-chip) to identify sequences bound by components of the transcriptional machinery. Reproducibility is illustrated by the comparison of

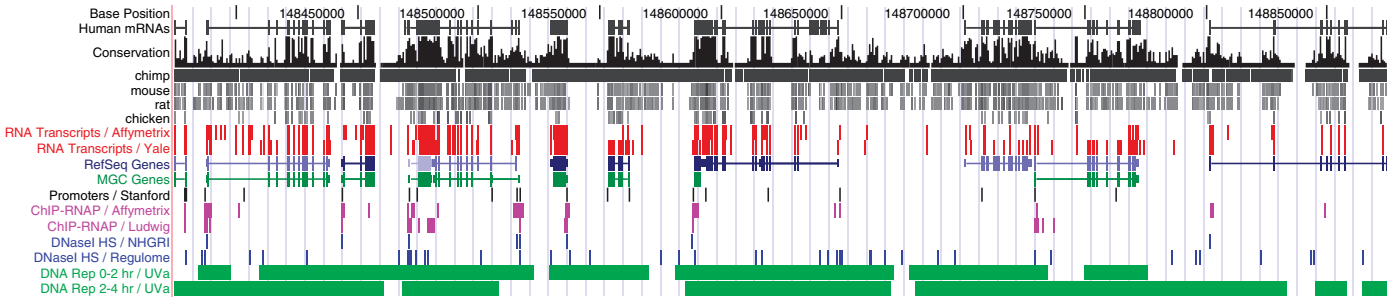

Fig. 3. UCSC Genome Browser display of representative ENCODE data. The genomic coordinates for ENCODE target ENr231 on chromosome 1 are indicated along the top. The different tracks are labeled at the left with the source of the data. The Conservation track shows a measure of evolutionary conservation based on a phylogenetic hidden Markov model (phylo-HMM) (*27*). Multiz (*28*) alignments of the human, chimpanzee, mouse, rat, and chicken assemblies were used to generate the species tracks. RefSeq Genes, MGC Genes indicate the mapping of mRNA transcripts from RefSeq (*17*) and MGC (*24*) projects, respectively, whereas the track labeled Human mRNAs represents all mRNAs in GenBank. Other tracks represent verified data from the ENCODE Consortium (*23*). The track with the location of sequences tested for promoter activity in a reporter assay is labeled as Promoters/Stanford. The positions of transcripts identified by oligonucleotide microarray hybridization (RNA Transcripts/Affymetrix and RNA Transcripts/Yale) and sequences detected by ChIP/chip analysis (ChIP-RNAP/Ludwig and ChIP-RNAP/Affymetrix) are indicated. The DNA replication tracks show segments that are detected to replicate during specified intervals of S phase in synchronized HeLa cells. The 0-2 hours and 2-4 hours tracks show segments that replicate during the first and second 2 hours periods of S phase. The DNA fragments released by DNase I cleavage were identified either from $CD4^+$ cells (DNaseI HS/NHGRI) or from K562 cells (DNaseI HS/Regulome).

studies conducted by two laboratories within the ENCODE Consortium, which analyzed different biological starting materials using the ChIP-chip approach and found an 83% concordance in the identified RNAP-binding sites in the region (*23*). Representation of the ENCODE data in this manner will afford a synthetic and integrative view of the functional structure of each of the target regions and, ultimately, the entire human genome.

Each research group will analyze and publish its own data to evaluate the experimental methods it is using and to elucidate new information about the biological function of the identified sequence elements. In addition, the Consortium will organize, analyze, and publish on all ENCODE data available on specific subjects, such as multiple sequence alignments, gene models, and comparison of different technological platforms to identify specific functional elements. At the conclusion of the pilot phase, the ENCODE Consortium expects to compare different methods used by the Consortium members and to recommend a set of methods to use for expanding this project into a full production phase on the entire genome.

Data Release and Accessibility

The National Human Genome Research Institute (NHGRI) has identified ENCODE as a "community resource project" as defined at an international meeting held in Fort Lauderdale in January 2003 (www.wellcome.ac.uk/doc_WTD003208.html). Accordingly, the ENCODE data-release policy (www.genome.gov/ENCODE) stipulates that data, once verified, will be deposited into public databases and made available for all to use without restriction.

Two concepts associated with this data-release policy deserve additional discussion. First, "data verification" refers to the assessment of the reproducibility of an experiment; ENCODE data will be released once they have been experimentally shown to be reliable. Because different types of experimental data will require different means of demonstrating such reliability, the Consortium will identify a minimal verification standard necessary for public release of each data type. These standards will be posted on the ENCODE Web site. Subsequently, ENCODE pilot-phase participants will use other experimental approaches to "validate" the initial experimental conclusion. This enriched information will also be deposited in the public databases.

Second, the report of the Fort Lauderdale meeting recognized that deposition in a public database is not equivalent to publication in a peer-reviewed journal. Thus, the NHGRI and ENCODE participants respectfully request that, until ENCODE data are published, users adhere to normal scientific etiquette for the use of unpublished data. Specifically, data users are requested to cite the source of the data (referencing this paper) and to acknowledge the ENCODE Consortium as the data producers. Data users are also asked to voluntarily recognize the interests of the ENCODE Consortium and its members to publish initial reports on the generation and analyses of their data as previously described. Along with these publications, the complete annotations of the functional elements in the initial ENCODE targets will be made available at both the UCSC ENCODE Genome Browser and the ENSEMBL Browser (www.ensembl.org).

Conclusion

ENCODE will play a critical role in the next phase of genomic research by defining the best path forward for the identification of functional elements in the human genome. By the conclusion of the pilot phase, the 44 ENCODE targets will inevitably be the most well-characterized regions in the human genome and will likely be the basis of many future genome studies. For example, other large genomics efforts, such as the Mammalian Gene Collection (MGC) program (*24*) and the International HapMap Project (*25*), are already coordinating their efforts to ensure effective synergy with ENCODE activities. Although the identification of all functional genomic elements remains the goal of the ENCODE project, attaining such comprehensiveness will be challenging, especially in the short term. For example, not all types of elements, such as centromeres, telomeres, and other yet-to-be defined elements, will be surveyed in the pilot project. Nonetheless, this project will bring together scientists with diverse interests and expertise, who are now focused on assembling the functional catalog of the human genome.

References and Notes

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. International Human Genome Sequencing Consortium, *Nature*, in press.
4. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
5. T. Ota *et al.*, *Nat. Genet.* **36**, 40 (2004).
6. J. L. Rinn *et al.*, *Genes Dev.* **17**, 529 (2003).
7. P. Kapranov, V. I. Sementchenko, T. R. Gingeras, *Brief Funct. Genomic Proteomic* **2**, 47 (2003).
8. International Rat Sequencing Consortium, *Nature* **428**, 493 (2004).
9. International Mouse Sequencing Consortium, *Nature* **420**, 520 (2002).
10. D. Boffelli, M. A. Nobrega, E. M. Rubin, *Nat. Rev. Genet.* **5**, 456 (2004).
11. T. Imanishi *et al.*, *PLoS Biol.* **2**, 856 (2004).
12. T. Evans, G. Felsenfeld, M. Reitman, *Annu. Rev. Cell Biol.* **6**, 95 (1990).
13. J. W. Thomas *et al.*, *Nature* **424**, 788 (2003).
14. S. Ruppert, E. H. Wang, R. Tjian, *Nature* **362**, 175 (1993).

15. K. Shuai, C. Schindler, V. R. Prezioso, J. E. Darnell Jr., *Science* **258**, 1808 (1992).
16. R. W. Blakesley *et al.*, *Genome Res.* **14**, 2235 (2004).
17. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **31**, 34 (2003).
18. International HapMap Consortium, *Nat. Rev. Genet.* **5**, 467 (2004).
19. W. J. Kent *et al.*, *Genome Res.* **12**, 996 (2002).
20. P. T. Spellman *et al.*, *Genome Biol* **3**, RESEARCH0046 (2002).
21. N. D. Trinklein *et al.*, *Genome Res.* **14**, 62 (2004).
22. B. Ren, B. D. Dynlacht, *Methods Enzymol.* **376**, 304 (2004).
23. ENCODE Consortium, unpublished data.
24. Mammalian Gene Collection (MGC) Project Team, *Genome Res.***14**, 2121 (2004).
25. International HapMap Consortium, *Nature* **426**, 789 (2003).
26. A. Felsenfeld, J. Peterson, J. Schloss, M. Guyer, *Genome Res.* **9**, 1 (1999).
27. A. Siepel, D. Haussler, in *Statistical Methods in Molecular Evolution*, R. Nielsen, Ed. (Springer, New York, in press).
28. M. Blanchette *et al.*, *Genome Res.* **14**, 708 (2004).
29. The Consortium thanks the ENCODE Scientific Advisory Panel for their helpful advice on the project: G. Weinstock, G. Churchill, M. Eisen, S. Elgin, S. Elledge, J. Rine, and M. Vidal. We thank D. Leja, and M. Cichanowski for their work in creating figures for this paper. Supported by the National Human Genome Research Institute, the National Library of Medicine, the Wellcome Trust, and the Howard Hughes Medical Institute.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5696/636/DC2
Tables S1 to S3

VIEWPOINT

Systems Biology and New Technologies Enable Predictive and Preventative Medicine

Leroy Hood,[1]* James R. Heath,[2,3] Michael E. Phelps,[3] Biaoyang Lin[1]

Systems approaches to disease are grounded in the idea that disease-perturbed protein and gene regulatory networks differ from their normal counterparts; we have been pursuing the possibility that these differences may be reflected by multiparameter measurements of the blood. Such concepts are transforming current diagnostic and therapeutic approaches to medicine and, together with new technologies, will enable a predictive and preventive medicine that will lead to personalized medicine.

Biological information is divided into the digital information of the genome and the environmental cues that arise outside the genome. Integration of these types of information leads to the dynamic execution of instructions associated with the development of organisms and their physiological responses to their environments. The digital information of the genome is ultimately completely knowable, implying that biology is unique among the sciences, in that biologists start their quest for understanding systems with a knowable core of information. Systems biology is a scientific discipline that endeavors to quantify all of the molecular elements of a biological system to assess their interactions and to integrate that information into graphical network models (*1*–*4*) that serve as predictive hypotheses to explain emergent behaviors.

The genome encodes two major types of information: (i) genes whose proteins execute the functions of life and (ii) cis control elements. Proteins may function alone, in complexes, or in networks that arise from protein interactions or from proteins that are interconnected functionally through small molecules (such as signal transduction or metabolic networks). The cis control elements, together with transcription factors, regulate the levels of expression of individual genes. They also form the linkages and architectures of the gene regulatory networks that integrate dynamically changing inputs from signal transduction pathways and provide dynamically changing outputs to the batteries of genes mediating physiological and developmental responses (*5*, *6*). The hypothesis that is beginning to revolutionize medicine is that disease may perturb the normal network structures of a system through genetic perturbations and/or by pathological environmental cues, such as infectious agents or chemical carcinogens.

Systems Approaches to Model Systems and Implications for Disease

A model of a metabolic process (galactose utilization) in yeast was developed from existing literature data to formulate a network hypothesis that was tested and refined through a series of genetic knockouts and environmental perturbations (*7*). Messenger RNA (mRNA) concentrations were monitored for all 6000 genes in the genome, and these data were integrated with protein/protein and protein/DNA interaction data from the literature by a graphical network program (Fig. 1).

The model provided new insights into the control of a metabolic process and its interactions with other cellular processes. It also suggested several concepts for systems approaches to human disease. Each genetic knockout strain had a distinct pattern of perturbed gene expression, with hundreds of mRNAs changing per knockout. About 15% of the perturbed mRNAs potentially encoded secreted proteins (*8*). If gene expression in diseased tissues also reveals patterns characteristic of pathologic, genetic, or environmental changes that are, in turn, reflected in the pattern of secreted proteins in the blood, then perhaps blood could serve as a diagnostic window for disease analysis. Furthermore, protein and gene regulatory networks dynamically changed upon exposure of yeast to an environmental perturbation (*9*). The dynamic progression of disease should similarly be reflected in temporal change(s) from the normal state to the various stages of disease-perturbed networks.

Systems Approaches to Prostate Cancer

Cancer arises from multiple spontaneous and/or inherited mutations functioning in networks that control central cellular events (*10*–*12*). It is becoming clear from our research that the evolving states of prostate cancer are reflected in dynamically changing expression patterns of the genes and proteins within the diseased cells.

A first step toward constructing a systems biology network model is to build a comprehensive expressed-mRNA database on the cell type of interest. We have used a technology called multiple parallel signature sequencing (MPSS) (*13*) to sequence a complementary DNA (cDNA) library at a rate of a million sequences in a single run and to detect mRNA transcripts down to one or a few copies per cell. A database containing more than 20 million mRNA signatures was constructed for normal prostate tissues and an androgen-sensitive prostate cancer cell line, LNCaP, in four states: androgen-starved,

[1]Institute for Systems Biology, Seattle, WA, USA. [2]Department of Chemistry, California Institute of Technology, Pasadena, CA, USA. [3]Department of Molecular and Medical Pharmacology, The David Geffen School of Medicine at the University of California Los Angeles, Los Angeles, CA, USA.

*To whom correspondence should be addressed. E-mail: lhood@systemsbiology.org