The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium

Running Title: The ENCODE Project seeks to identify all functional sequence elements in the human genome

Corresponding Author: Elise A. Feingold, Ph.D.
5635 Fishers Lane
Suite 4076 MSC 9305
Bethesda, MD 20892-9305
(301) 496-7531
(301) 480-2770 (FAX)
Elise_Feingold@nih.gov

ENCODE Project Consortium

ENCODE Project Scientific Management E.A. Feingold[1], P.J. Good[1], M. Guyer[1], S. Kamholz[1], L. Liefer[1], K. Wetterstrand[1], F.S. Collins[2]

Initial ENCODE Pilot Phase Participants:

Affymetrix, Inc. T.R. Gingeras[3], D. Kampa[3], E.A. Sekinger[4], J. Cheng[3], H. Hirsch[4], S. Ghosh[3], Z. Zhu[5], S. Patel[3], A. Piccolboni[3], A. Yang[4], H. Tammana[3], S. Bekiranov[3], P. Kapranov[3], G. Church[5], K. Struhl[4];

Ludwig Institute for Cancer Research B. Ren[6], T.H. Kim[6], L.O. Barrera[6], C. Qu[6], S. Van Calcar[6], R. Luna[7], C.K. Glass[7], M.G. Rosenfeld[8];

Municipal Institute of Medical Research R. Guigo[9], S. Antonarakis[10], E. Birney[11], M. Brent[12], L. Patcher[13], A. Reymond[10, 14], M. Dermitzakis[15], C. Dewey[13], D. Keefe[11], F. Denoeud[9], J. Lagarde[9], J. Ashurst[15], T. Hubbard[15], J.J. Wesselink[9], R. Castelo[9], E. Eyras[9];

Stanford University R.M. Myers[16], A. Sidow[16, 17], S. Batzoglou[18], N.D. Trinklein[16], S.J. Hartman[16], S.F. Aldred[16], E. Anton[16], D.I. Schroeder[19], S.S. Marticke[16], L. Nguyen[16], J. Schmutz[20], J. Grimwood[20], M. Dickson[20], G.M. Cooper[16], E.A. Stone[16], G. Asimenos[18], M. Brudno[18];

University of Virginia A. Dutta[21], N. Karnani[21], C.M. Taylor[21, 22], H.K. Kim[21], G. Robins[22];

University of Washington G. Stamatoyannopoulos[23, 24], J.A. Stamatoyannopoulos[25], M. Dorschner[25], P. Sabo[25], M. Hawrylycz[25], R. Humbert[25], J. Wallace[25], M. Yu[23], P. Navas[23], M. Olson[23, 24], M. McArthur[25], W.S. Noble[24];

Wellcome Trust Sanger Institute I. Dunham[15], C.M. Koch[15], R.M. Andrews[15], G.K. Clelland[15], S. Wilcox[15], J.M. Fowler[15], K.D. James[15], P. Groth[15], O.M. Dovey[15], P.D. Ellis[15], V.L. Wraight[15], A.J. Mungall[15], P. Dhami[15], H. Fiegler[15], C.F. Langford[15], N.P. Carter[15], D. Vetrie[15];

Yale University M. Snyder[26], G. Euskirchen[26], A.Urban[26], U. Nagalakshmi[26], J. Rinn[26], G. Popescu[26], P. Bertone[26], J. Rozowsky[27], O. Emanuelsson[27], T. Royce[27], S. Chung[27], M. Gerstein[27], Z. Lian[28], J. Lian[28], Y. Nakayama[28], S. Weissman[28], V. Stolc[26, 29], W. Tongprasit[30], H. Sethi[30]

Additional ENCODE Pilot Phase Participants:

British Columbia Cancer Agency Genome Sciences Centre S. Jones[31], M. Marra[31], H. Shin[31], J. Schein[31];

Broad Institute M. Clamp[32], K. Lindblad-Toh[32], J. Chang[32], D.B. Jaffe[32], M. Kamal[32], E.S. Lander[32], T. Mikkelson[32], J. Vinson[32], M.C. Zody[32];

Children's Hospital Oakland Research Institute P.J. de Jong[33], K. Osoegawa[33], M. Nefedov[33], B. Zhu[33];

National Human Genome Research Institute / Bioinformatics and Scientific Programming Core A.D. Baxevanis[34], T.G. Wolfsberg[34];

National Human Genome Research Institute / Molecular Genetics Section F.S. Collins[34], G.E. Crawford[34];

NIH Intramural Sequencing Center/National Human Genome Research Institute E.D. Green[35], G.G. Bouffard[35], E.H. Margulies[35], M.E. Portnoy[35], N.F. Hansen[35], P.J. Thomas[35], J.C. McDowell[35], B. Maskeri[35], A.C. Young[35], J.R. Idol[35], R.W. Blakesley[35];

National Library of Medicine G. Schuler[36];

Pennsylvania State University W. Miller[37-39], R. Hardison[37, 40], L. Elnitski[37, 39], P. Shah[37, 39];

The Institute for Genomic Research S.L. Salzberg[41], M. Pertea[41], W.H. Majoros[41];

University of California, Santa Cruz D. Haussler[42, 43], D. Thomas[42], K. Rosenbloom[42], H. Clawson[42], A. Siepel[42], W.J. Kent[42]

ENCODE Technology Development Phase Participants:

Boston University Z. Weng[44, 45], S. Jin[44, 45], A. Halees[44], H. Burden[44, 45], U. Karaoz[44], Y. Fu[44], Y. Yu[44, 45], C. Ding[44], C.R. Cantor[44, 45];

Massachusetts General Hospital R.E. Kingston[46, 47], J. Dennis[46, 47];

NimbleGen Systems, Inc. R.D. Green[48], M.A. Singer[48], T.A. Richmond[48], J.E. Norton[48], P.J Farnham[49], M.J. Oberley[50], D.R. Inman[51];

NimbleGen Systems, Inc. M.R. McCormick[48], H. Kim[52], C.L. Middle[48], M.C. Pirrung[52];

University of California, San Diego X.D. Fu[7], Y.S. Kwon[7], Z. Ye[7];

University of Massachusetts Medical School J. Dekker[53], T.M. Tabuchi[53], N. Gheldof[53], J. Dostie[53], S.C. Harvey[54]

Affiliations for Participants

1 Division of Extramural Research, National Human Genome Research Institute, Bethesda, MD 20892-9305

2 Office of the Director, National Human Genome Research Institute, Bethesda, MD 20892

3 Affymetrix, Inc., Santa Clara, CA 92024

4 Dept of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

5 Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

6 Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA 92093-0653

7 Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093-0651

8 Howard Hughes Medical Institute, University of California, San Diego, La Jolla, CA 92093

9 Grup de Recerca en Informatica Biomedica, Institut Municipal d'Investigacio Medica and Centre de Regulacio Genomica, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

10 Department of Genetic Medicine and Development, University of Geneva Medical School and University Hospitals of Geneva, 1211 Geneva, Switzerland

11 European Bioinformatics Institute, Hinxton Cambridge, United Kingdom

12 Laboratory for Computational Genomics, Washington University, St. Louis, MO 63130

13 Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720

14 Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

15 The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

16 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

17 Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305

18 Department of Computer Science, Stanford University, Stanford, CA 94305

19 Biomedical Informatics Program, Stanford University School of Medicine, Stanford, CA 94305

20 Stanford Human Genome Center, Stanford University School of Medicine, Stanford, CA 94305

21 Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908

22 Department of Computer Science, University of Virginia, Charlottesville, VA 22908

23 Division of Medical Genetics, University of Washington, Seattle, WA 98195

24 Department of Genomic Sciences, University of Washington, Seattle, WA 98195

25 Regulome, Seattle, WA 98121

26 Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520

27 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520

28 Department of Genetics, Yale University, New Haven, CT 06520

29 Center for Nanotechnology, NASA Ames Research Center, Moffett Field, CA 94036

30 Eloret Corporation, NASA Ames Research Center, Sunnyvale, CA 94087

31 Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

32 Broad Institite, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02141

33 Children's Hospital Oakland Research Institute, Oakland, CA 94609-1673

34 Genome Technology Branch, Division of Intramural Research, National Human Genome Research Institute, Bethesda, MD 20892-8002

35 NISC Comparative Sequencing Program, Genome Technology Branch and NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, Bethesda, MD 20892

36 National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20892

37 Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802

38 Department of Biology, Pennsylvania State University, University Park, PA 16802

39 Department of Computer Science, Pennsylvania State University, University Park, PA 16802

40 Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802

41 The Institute for Genomic Research, Rockville, MD 20850

42 Genome Bioinformatics Group, Center for Biomolecular Science & Engineering, Baskin School of Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064

43 Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064

44 Bioinformatics Program, Boston University, Boston, MA 02215

45 Biomedical Engineering Department, Boston University, Boston, MA 02215

46 Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114

47 Department of Genetics, Harvard Medical School, Boston, MA 02115

48 NimbleGen Systems, Inc., Madison, WI 53711

49 Department of Medical Pharmacology, University of California, Davis Genome Center, Davis, CA 95616-8816

50 Health Sciences Learning Center, University of Wisconsin Medical School, Madison, Wisconsin 53705-2221

51 University of Wisconsin Medical School, Madison, WI 53706

52 Department of Chemistry, Duke University, Durham, NC 27708

53 Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605-2324

54 School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230

ABSTRACT

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 Mb (approximately 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.

INTRODUCTION

Both genetic and environmental factors contribute to almost every human disease, thereby offering the potential for development of interventions based on understanding the genetic factors, the non-genetic factors, and their interactions. The rationale for initiating the Human Genome Project and for determining the sequence of the human genome was predicated on the belief that knowledge of the sequence would lead to a better understanding of the genetic factors underlying human disease and, ultimately, to improvement of human health. Having now determined the sequence of human DNA, we are faced with the enormous challenge of interpreting it and understanding how to use that information to understand the biology of human health and disease.

The ENCyclopedia Of DNA Elements (ENCODE) Project, which is described in this paper, is predicated on the belief that a comprehensive catalogue of the structural and functional components encoded in the human genome sequence will be critical for understanding human biology well enough to unravel the genetic basis of disease and for using that information to address human health problems. Such a complete catalogue, 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 be included.

To illustrate the magnitude of the challenge involved in developing such an encyclopedia of DNA elements, 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 – 25,000) (*1*). Non-protein-coding genes are much less well defined. Some, such as the rRNA and tRNA genes, were identified several decades ago, but more recent approaches, such as cDNA-cloning efforts (*2, 3*) and chip-based transcriptome analyses (*4, 5*), have revealed the existence of many transcribed sequences of unknown function. Neither the number of these sequences nor their function has been systematically determined. 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 (*6, 7*).

Yet only about one-third of the sequence under such selection is predicted to encode proteins (*8, 9*); the remaining two-thirds is inferred to represent elements with other functions, such as non-protein-coding RNA transcripts or transcriptional regulatory elements. Our collective knowledge about these putative functional, non-coding elements, which represent the majority of the 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 towards executing such a comprehensive study is not clear and, thus, a major effort to determine how to conduct such studies is warranted.

The National Institutes of Health (NIH) already supports several large-scale projects that are contributing to the identification of functional elements in the human genome. The National Human Genome Research Institute (NHGRI) directs a large-scale sequencing program (www.nhgri.nih.gov/10001691) that is determining the sequence of many additional vertebrate (particularly mammalian) genomes. The resulting data are being used for comparative sequence analyses to identify evolutionarily conserved genomic regions, which represent strong candidates for functional elements (*10*). The Mammalian Gene Collection (MGC), a trans-NIH effort, aims to identify and sequence a single full-length cDNA clone for every human (and mouse) protein-coding gene (*11*); as of

September 2004, clones and sequences for over 12,000 human genes were available (*12*). The ENCODE Project (www.genome.gov/ENCODE) is a new effort intended to develop and implement an effective set of high-throughput biological, biochemical, and computational 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 rigorously evaluating and comparing a number of strategies for comprehensively identifying various types of genomic elements. Each of the strategies will be assessed for its accuracy, comprehensiveness, and ability to be applied cost-effectively and at high-throughput to large regions of the human genome. The pilot phase should also 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 deficiencies are expected to be addressed in ENCODE's 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. Specifically, all pilot-phase participants have agreed to study the entire set of ENCODE targets — 44 discrete regions that together encompass approximately 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 common test set will allow the results of different approaches to be directly compared with one another, providing an opportunity to identify the most effective set of techniques for the analysis of the entire genome.

The set of ENCODE targets was chosen to represent a range of genomic features (www.genome.gov/10005115). To begin with, it was 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 main criteria used for the manual selection were: 1) the presence of extensively characterized genes and/or other functional elements; and 2) the availability of a substantial amount of comparative sequence data. For example, the genomic segments containing the alpha and beta globin gene clusters were chosen because of the wealth of data available for these loci (*13*). On the other hand, the region encompassing the *CFTR* gene was selected because of the extensive amount of multi-species sequence data available for this genomic segment (*14*). Once the manual selections had been made, the remaining targets were chosen at random using an algorithm that ensured that the complete set of targets represented the range of gene content and level of non-exonic

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 (see www.genome.gov/ENCODE), who are working together in a highly interactive way to implement and evaluate a set of computational and experimental approaches for identifying functional elements in the targeted genomic regions. The results obtained using these different approaches will be compared and, where appropriate, followed up with additional experimental and computational analyses in an effort to cross-validate different results. Such cross-validation will be abetted by the use of common reagents (see below).

The ENCODE pilot phase began in September 2003 with the funding of eight projects (see 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, DNaseI hypersensitive sites, chromatin modifications, and multi-species 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 using different microarray platforms, computational methods for finding genes and for identifying highly conserved sequences, and expression reporter assays.

In addition to these eight, other groups have 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 initial participants, the ENCODE Consortium is open to all interested academic, government, and private-sector investigators, as long as they are willing to participate according to established Consortium guidelines (www.genome.gov/10006162). Participation requires 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 novel methods both for more efficient identification of known elements and for identification of heretofore unknown elements. Interactions

between investigators participating in the first two phases of ENCODE are encouraged to promote rapid implementation of promising new techniques.

RESEARCH PLANS

Each group participating in the ENCODE pilot phase 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 and PCR-amplicon arrays) are being used to identify transcribed regions. 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 and GM06990) and two antibodies (one for the general transcription factor $TAF_{II}250$ (*15*) and another for the inducible transcription factor NF-κB (*16*)). All verified data generated by the Consortium will be publicly available (see below). In addition, the sources of common reagents will be identified at www.genome.gov/ENCODE and the Consortium plans to make other reagents available as feasible.

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 non-human vertebrates. This will allow ENCODE to identify the quality and amount of comparative sequence data necessary to accurately identify evolutionarily conserved

elements. It will also provide a data set for developing more powerful computational tools for using comparative sequence data to infer biological function. This sequencing effort involves the isolation and sequencing of bacterial artificial chromosome (BAC) clones spanning ENCODE targets in multiple species (www.nisc.nih.gov/open_page.html?/projects/encode/index.cgi) to produce "comparative-grade" sequence data (*17*). To date, ten vertebrates have been selected on the basis of multiple factors, including phylogenetic position (Fig. 2 and table S3) and the availability of a BAC library. In addition to this ENCODE-specific effort, comparative sequence data are also being captured from a number of ongoing whole-genome sequencing projects, including those for mouse, rat, dog, chicken, cow, chimpanzee, macaque, frog, and zebrafish. A unique RefSeq accession number (*18*) is being assigned for the sequence of each ENCODE-orthologous target region in each species, with periodic data freezes instituted that will allow analyses to be performed on identical data sets. In the future, selection of additional species for ENCODE-specific, BAC-based sequencing will be coordinated with the broader NHGRI process for choosing organisms for whole-genome sequencing (www.genome.gov/10002189).

One feature of the evolutionarily conserved elements to be assayed is sequence variation (table S1). 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 (*19*). This aspect of the ENCODE project will result in a quantitative view of the evolutionary constraints on conserved non-protein-coding and protein-coding regions.

The ENCODE pilot phase will produce an inventory of functional elements in the targeted 30 Mb of the human genome. As a starting point, gene structures, including the precise locations of 5' transcription start sites, intron/exon boundaries and 3' polyadenylation sites, will be determined for each gene in the ENCODE targets, both known and predicted (table S1). An early example of other types of data being obtained for each ENCODE target is presented in Fig. 3. The positions of the evolutionarily conserved regions, as detected by analyzing sequences from several organisms, are shown, and can be correlated with the results of other studies. The experimental methods highlighted in Fig. 3 include microarray analyses to detect transcribed sequences (*20*) and to determine the replication profile of the segment (*21*), high-throughput assays for DNase I hypersensitive sites, and two methods to localize promoters. For the promoter studies, the first method employs reporter constructs containing sequences around putative transcription initiation sites, measuring expression of a reporter gene (*22*). The second involves chromatin immunoprecipitation (ChIP) with an antibody to RNA polymerase (RNAP) and hybridization to DNA microarrays (chip) (*23*) (so-called ChIP-chip) to identify sequences bound by components of the transcriptional machinery. Using this second approach, two laboratories within the ENCODE Consortium have analyzed different biological starting materials, yet there was a striking 83% concordance in the identified RNAP-binding sites (*24*). For the microarray analysis of RNA transcripts, two groups using different RNA samples but the same Affymetrix genome tiling arrays detect overlapping patterns of RNA transcripts (*25*). DNA replicated during indicated two-hour intervals of S phase was hybridized to Affymetrix genome tiling

arrays to determine the replication profile of these genomic segments (*26*). The results indicate that the gene-dense Enr231 segment is replicated in the early part of S phase and that at least one early firing origin of replication resides in a 20 kb inter-genic region located at co-ordinate 148,550,000.

Another sequence feature being assayed in the ENCODE pilot phase is DNaseI hypersensitivity, which is known to be associated with *cis*-regulatory sequences such as enhancers, promoters, insulators, and locus control regions (*27*). Two groups are using similar library-based approaches for identification of DNaseI hypersensitive sites (*28, 29*). These methods involve isolation of DNA fragments that result from individual DNaseI cutting events in nuclear chromatin. These fragments are then localized by high-throughput DNA sequencing. Many of these sites overlap many of the predicted promoters in the region (Fig. 3). A separate approach, quantitative chromatin profiling (*28*) is being applied across the ENCODE target regions. Tiled amplicons (~250 bp each) are assayed by quantitative PCR (qPCR) to measure DNaseI cleavage across each ENCODE target. Preliminary analysis on the beta-globin region demonstrates the ability of this assay to detect known hypersensitive regions (fig. S1).

DATA MANAGEMENT AND ANALYSIS

The ENCODE pilot phase will generate large sets of several different data types. Capturing, storing, integrating, and displaying such diverse data will be challenging. Data that can be directly linked to genomic sequence will be managed at the UCSC Genome

Browser (*30*), where an ENCODE-specific component has been established (www.genome.ucsc.edu/ENCODE). Other data types will be stored either at available public databases [e.g., the GEO (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 since these will be key to analyzing and comparing data from different platforms and laboratories. The ENCODE pilot phase will make use of the MAGE standard for representing microarray data (*31*), while data standards for other data types will be developed as needed.

Each research group will analyze its own data to evaluate the experimental methods being used and to elucidate new information about the biological function of the identified sequence elements. In addition, the Consortium will organize and analyze all ENCODE data available on specific subjects, such as multiple species 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 ACCESSIBILTY

The NHGRI has identified ENCODE as a "community resource project." This important concept was defined at an international meeting held in Ft. Lauderdale in January 2003 (http://www.wellcome.ac.uk/en/1/awtpubrepdat.html) as a research project specifically devised and implemented to create a set of data, reagents, or other material whose primary utility will be as a resource for the broad scientific community. Accordingly, the ENCODE data release policy (www.genome.gov/ENCODE_data_release) stipulates that data, once verified, will be deposited into public databases and made available for all to use without restriction.

There are two concepts associated with this data release policy that deserve additional discussion. First, "data verification" refers to assessing 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 website. 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 statement ”made available for all to use without restriction” is used in the sense discussed in the report of the Ft. Lauderdale meeting, which 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 ENCODE data be regarded as unpublished data, with users adhering 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. Such studies will likely consist of comparisons of different experimental and computational methods for finding functional genomic elements. In addition, at the conclusion of the pilot phase, the ENCODE Consortium expects to publish its overall comparative analysis of the different methods used by Consortium members, and a recommendation for expanding the program to the entire human genome. It is expected that the results generated by specific methods will be published by the individual ENCODE participants, including descriptions of the biological insights gained from their analyses. 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 all 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 MGC program and the International HapMap Project (*32*), are already coordinating their efforts to ensure effective synergy with ENCODE activities. ENCODE has successfully brought together scientists with diverse interests and expertise, all intensely focused on tackling the immense challenge of assembling a functional catalog of the human genome. Its interactive and highly collaborative nature, along with its commitment to rapid data release, is intended to encourage innovation, thereby helping to advance genome science; the lessons learned from the ENCODE Project will undoubtedly be useful in the analysis of other genomes. Knowledge of all of the functional elements in the human genome, along with information about how they interact in pathways and networks, is crucial to understanding how both genetic and environment factors influence disease. ENCODE's ultimate aim is to provide that knowledge.

References

1. International Human Genome Sequencing Consortium, *submitted* (2004).
2. Y. Okazaki *et al.*, *Nature* **420**, 563-73 (2002).
3. T. Ota *et al.*, *Nat Genet* **36**, 40-5 (2004).
4. J. L. Rinn *et al.*, *Genes Dev* **17**, 529-40 (2003).
5. P. Kapranov, V. I. Sementchenko, T. R. Gingeras, *Brief Funct Genomic Proteomic* **2**, 47-56 (2003).
6. International Rat Sequencing Consortium, *Nature* **428**, 493-521 (2004).
7. International Mouse Sequencing Consortium, *Nature* **420**, 520-62 (2002).
8. J. C. Venter *et al.*, *Science* **291**, 1304-51 (2001).
9. International Human Genome Sequencing Consortium, *Nature* **409**, 860-921 (2001).
10. D. Boffelli, M. A. Nobrega, E. M. Rubin, *Nat Rev Genet* **5**, 456-65 (2004).
11. Mammalian Gene Collection (MGC) Program Team, *Proc Natl Acad Sci U S A* **99**, 16899-903 (2002).
12. Mammalian Gene Collection (MGC) Project Team, *Genome Res* (2004 in press).
13. T. Evans, G. Felsenfeld, M. Reitman, *Annu Rev Cell Biol* **6**, 95-124 (1990).
14. J. W. Thomas *et al.*, *Nature* **424**, 788-93 (2003).
15. S. Ruppert, E. H. Wang, R. Tjian, *Nature* **362**, 175-9 (1993).
16. R. Martone *et al.*, *Proc Natl Acad Sci U S A* **100**, 12247-52 (2003).
17. R. W. Blakesley *et al.*, *Genome Res* (2004 in press).
18. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **31**, 34-7 (2003).

19. International HapMap Consortium, *Nat Rev Genet* **5**, 467-75 (2004).

20. P. Kapranov *et al.*, *Science* **296**, 916-9 (2002).

21. Y. Jeon *et al.*, (submitted).

22. N. D. Trinklein *et al.*, *Genome Res* **14**, 62-6 (2004).

23. B. Ren, B. D. Dynlacht, *Methods Enzymol* **376**, 304-15 (2004).

24. T. R. Gingeras, K. Struhl, B. Ren, unpublished data

25. T. R. Gingeras, M. Snyder, unpublished data

26. A. Dutta, unpublished data

27. D. S. Gross, W. T. Garrard, *Annu Rev Biochem* **57**, 159-97 (1988).

28. G. Stamatoyannopoulos, unpublished data

29. F. S. Collins, unpublished data

30. W. J. Kent *et al.*, *Genome Res* **12**, 996-1006 (2002).

31. P. T. Spellman *et al.*, *Genome Biol* **3**, RESEARCH0046 (2002).

32. International HapMap Consortium, *Nature* **426**, 789-96 (2003).

33. A. Felsenfeld, J. Peterson, J. Schloss, M. Guyer, *Genome Res* **9**, 1-4 (1999).

34. A. Siepel, D. Haussler, in *Statistical Methods in Molecular Evolution* N. R., Ed. (Springer, 2004 in press).

35. M. Blanchette *et al.*, *Genome Res* **14**, 708-15 (2004).

36. R. M. Myers, unpublished data

37. T. R. Gingeras, unpublished data

38. M. Snyder, unpublished data

Acknowledgements:

The Consortium thanks the ENCODE Scientific Advisory Panel for their helpful advice on the project: George Weinstock, Gary Churchill, Mike Eisen, Sarah Elgin, Steve Elledge, Jasper Rine, and Marc Vidal. We thank Darryl Leja and Mike Cichanowski for their work in creating figures for this paper. Marco Marra's group acknowledges the assistance of Ian Bosdet, Carrie Mathewson, Darlene Lee, and Readman Chiu. Mark McCormick's group acknowledges the critical assistance of Todd Richmond with bioinformatics and array design. Francis Collins' group would like to thank Tom Vasicek, Daixing Zhou, Shujun Luo, and Lynx Therapeutics for sequencing DNase hypersensitive sites using MPSS and Ingeborg Holt and James Whittle for bioinformatics support. Eric Green's group would like to thank all participants of the NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program for generating and analyzing comparative sequence data. This work was supported by the National Human Genome Research Institute, the National Library of Medicine, the Wellcome Trust and the Howard Hughes Medical Institute.

Figure 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.

Figure 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 (*33*). Comparative-grade finished reflects sequence with greater than 8-fold coverage that has been subjected to additional manual refinement to ensure accurate order and orientation of all sequence contigs (*17*). 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.

Figure 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 source of the data. The Conservation track shows a measure of evolutionary conservation based on a phylogenetic hidden Markov model (phylo-HMM) (*34*). Multiz (*35*) alignments of the

human, chimpanzee, mouse, rat, and chicken assemblies were used to generate the species tracks. RefSeq, MGC indicate the mapping of mRNA transcripts from RefSeq (*18*) and MGC (*12*) projects, respectively, while the track labeled Human mRNAs represent all mRNAs in GenBank. The track with the location of sequences tested for promoter activity in a reporter assay is labeled as Promoters/Stanford (*36*). The positions of transcripts identified by oligonucleotide microarray hybridization (RNA Transcripts/Affymetrix; (*37*)) and RNA Transcripts/Yale; (*38*)) and sequences detected by ChIP/chip analysis by the Ren and Gingeras/Struhl laboratories (ChIP-RNAP/Ludwig and ChIP-RNAP/Affymetrix, respectively (*24*)) are indicated. The DNA replication tracks show segments that are detected to replicate during specified intervals of S phase in synchronized HeLa cells (*26*). The 0-2 hr and 2-4 hr tracks show segments that replicate during the first and second 2 hr periods of S phase. The DNA fragments released by DNaseI cleavage were identified from either CD4+ cells (DNaseI HS/NHGRI; (*29*)) or from K562 cells (DNaseI HS/Regulome; (*28*)).

Supporting Online Material

Figure S1

Tables S1, S2, S3

Figure S1. Quantitative chromatin profile of the human beta-globin locus.

Computed signal-to-noise (SNR) ratios of DNaseI hypersensitivity were measured across 90 kb of ENCODE target ENm009 in K562 erythroid cells (28). The peaks correspond to DNaseI hypersensitivity sites, which reflect candidates for promoters, enhancers, insulators, and LCR elements.

Table S1: Groups participating in the ENCODE Pilot phase.

Initial ENCODE Pilot Phase Participants

Research Group	Institution	Research Goals
Ian Dunham	Wellcome Trust Sanger Institute	Map origins of replication, DNA methylation, chromatin modifications, transcription factor binding sites, primarily with ChIP/chip assays using spotted DNA microarrays.
Anindya Dutta	University of Virginia	Identify early and late origins of replication, sites of replication termination and pause sites for replication forks. Replication products mapped by hybridization to Affymetrix microarrays.
Thomas Gingeras	Affymetrix, Inc.	Map RNA transcripts, binding sites for transcription factors and chromosomal proteins using Affymetrix microarrays and ChIP/chip assays.
Roderic Guigo	Municipal Institute of Medical Research	Identify all protein-coding genes. Combine computational prediction with experimental RT/PCR confirmation of gene models.
Richard Myers	Stanford University	Identify promoters and enhancers with transfection of reporter constructs into cell lines. Identify transcription factor binding sites and chromatin modifications with ChIP/Chip assays using Affymetrix arrays. Identify conserved domains with comparative genomic data. Test the function of conserved domains by mapping polymorphisms to these domains and assaying for the effect of these polymorphisms in reporter assays for enhancers in transfected cells.
Bing Ren	Ludwig Institute for Cancer Research	Identify promoters, enhancers, repressors/silencers using ChIP/chip assays and mapping on spotted DNA microarrays.
Michael Snyder	Yale University	Map RNA transcripts and binding sites for transcription factors and chromosomal proteins using DNA microarrays and ChIP/chip assays. Comparison of Affymetrix, NimbleGen and spotted DNA arrays platforms.
George Stamatoyannopoulos	University of Washington	Map DNase I hypersensitive sites using quantitative, real time PCR.

Additional ENCODE Pilot Phase Participants

Research Group	Institution	Research Goals

Andy Baxevanis	National Human Genome Research Institute	Develop an ENCODE data portal for non-sequence based data including coordinated data deposition and dissemination.
Kerstin Lindblad-Toh/ Michelle Clamp	Broad Institute	Develop methodologies, algorithms and software to generate regional alignments of multiple genomes in the ENCODE regions.
Greg Crawford/ Francis Collins	National Human Genome Research Institute	Identify DNase hypersensitive sites; develop high-throughput Massively Parallel Signature Sequencing (MPSS) assay for DNase hypersensitive sites.
Pieter De Jong	Children's Hospital Oakland Research Institute	Create clone resources to support comparative sequencing.
Eric Green	NIH Intramural Sequencing Center/ National Human Genome Research Institute	Isolate BAC clones for ENCODE regions in multiple organisms; generate multispecies comparative genome sequence data for these ENCODE regions; develop computational tools for analysis of comparative genome sequences.
Ross Hardison	Pennsylvania State University	Develop tools to analyze comparative genomic sequences and integrate functional data with the genome sequence.
David Haussler	University of California, Santa Cruz	Develop ENCODE-specific views of the human genome using the Santa Cruz UCSC Browser; develop tools to analyze comparative genomic sequences and integrate functional data with the genome sequence.
Steven Jones	British Columbia Cancer Agency Genome Sciences Centre	Generate whole genome data on gene expression; develop tools to identify regulatory elements from co-expressed genes.
Marco Marra	British Columbia Cancer Agency Genome Sciences Centre	Generate fingerprint maps and tiling paths for BACs isolated from the ENCODE regions in different species; identify alternatively spliced transcripts for genes in the ENCODE regions.
Webb Miller	Pennsylvania State University	Develop tools to analyze comparative genomic sequences and integrate functional data with the genome sequence.
Steve Salzberg	The Institute for Genomic Research	Develop computational tools to analyze comparative genomic Sequences, to find genes and to assemble genomes.
Greg Schuler	National Center for Biotechnology Information (NCBI), National Library of Medicine	Coordinate ENCODE comparative genomic sequence data with NCBI.

Table S2: Groups participating in the ENCODE Technology Development phase.

ENCODE Technology Development Phase Participants

Research Group	Institution	Research Goals
Job Dekker	University of Massachusetts Medical School	Develop PCR strategy to identify regions in chromosomes that interact through protein complex binding using the Chromosome Conformation Capture (3C) technology.
Xiang-Dong Fu	University of California, San Diego	Improve sensitivity and specificity of the ChIP/chip technology using single stranded oligonucleotide microarrays and DASL (DNA Annealing Selection and Ligation) technology.
Roland Green	NimbleGen Systems, Inc.	Test ability of NimbleGen's Maskless Array Synthesis technology to map transcription factor binding sites and first exon/promoter identification in ChIP and microarray assays.
Robert Kingston	Massachusetts General Hospital	Develop high-throughput methods for mapping chemical and enzymatic DNA cleavage sites in chromatin at nucleotide resolution.
Mark McCormick	NimbleGen Systems, Inc.	Develop "exon-linkage assay" to study alternative splicing using NimbleGen's oligonucleotide array platform.
Zhiping Weng	Boston University	Develop computational methods to identify cis-regulatory elements in alternative promoters and confirm these elements by competitive PCR and reporter-construct assays in transfected cells.
To be added in September 2004		
To be added in September 2004		
To be added in September 2004		
To be added in September 2004		
To be added in		

September 2004		
To be added in September 2004		

Table S3: Anticipated Comparative Sequencing Datasets for the ENCODE Target Regions

Name	Latin Name	Sequence Quality	Source
Armadillo (nine-banded)	*Dasypus novemcinctus*	Comparative grade finished*	NIH Intramural Sequencing Center
Baboon (olive)	*Papio cynocephalus anubis*	Comparative grade finished	NIH Intramural Sequencing Center
Bat (greater horseshoe)	*Rhinolophus ferrumequinum*	Comparative grade finished	NIH Intramural Sequencing Center
Cat	*Felis catus*	Comparative grade finished	NIH Intramural Sequencing Center
Dusky Titi	*Callicebus moloch*	Comparative grade finished	NIH Intramural Sequencing Center
Elephant (African)	*Loxodonta africana*	Comparative grade finished	NIH Intramural Sequencing Center
Galago (small-eared)	*Otolemur garnetti*	Comparative grade finished	NIH Intramural Sequencing Center
Guinea Pig	*Cavia porcellus*	Comparative grade finished	NIH Intramural Sequencing Center
Hedgehog (middle-African)	*Atelerix albiventris*	Comparative grade finished	NIH Intramural Sequencing Center
Lemur (gray mouse)	*Microcebus murinus*	Comparative grade finished	NIH Intramural Sequencing Center
Marmoset (white-tufted ear)	*Callithrix jacchus*	Comparative grade finished	NIH Intramural Sequencing Center
Monkey (colobus)	*Colobus guereza*	Comparative grade finished	NIH Intramural Sequencing Center
Monkey (owl)	*Aotus nancymaae*	Comparative grade finished	NIH Intramural Sequencing Center
Platypus (duck-billed)	*Ornithorhynchus anatinus*	Comparative grade finished	NIH Intramural Sequencing Center
Rabbit	*Oryctolagus cuniculus*	Comparative grade finished	NIH Intramural Sequencing Center
Shrew (European common)	*Sorex araneus*	Comparative grade finished	NIH Intramural Sequencing Center
Tenrec (lesser hedgehog)	*Echinops telfairi*	Comparative grade finished	NIH Intramural Sequencing Center
Bovine	*Bos taurus*	High-quality finished**	Baylor College of Medicine Human Genome Sequencing Center
Chimpanzee	*Pan troglodytes*	High-quality finished	Washington University Genome Sequencing Center, Broad Institute/MIT Center for Genomic Research
Dog	*Canis familiaris*	High-quality	Broad Institute/MIT Center for

		finished	Genomic Research
Frog	*Xenopus tropicalis*	High-quality finished	US Department of Energy Joint Genome Institute
Macaque	*Macaca mulatta*	High-quality finished	Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, J. Craig Venter Joint Technology Center
Mouse	*Mus musculus*	High-quality finished	Washington University Genome Sequencing Center, Broad Institute/MIT Center for Genomic Research, Wellcome Trust Sanger Institute
Rat	*Rattus norvegicus*	High-quality finished	Baylor College of Medicine Human Genome Sequencing Center
Zebrafish	*Danio rerio*	High-quality finished	Wellcome Trust Sanger Institute
Chicken	*Gallus gallus*	High-coverage whole genome shotgun***	Washington University Genome Sequencing Center
Opossum	*Monodelphis domestica*	High-coverage whole genome shotgun	Broad Institute/MIT Center for Genomic Research
Orangutan	*Pongo pygmaeus*	High-coverage whole genome shotgun	TBD

* Comparative grade finished sequencing involves shotgun sequencing to 8X-10X coverage with additional manual refinement to order and orient contigs. The product is at an intermediate level between purely shotgun and high-quality finished sequence.

** High-quality finished reflects highly accurate and contiguous sequence, with a best-faith effort used to resolve all difficult regions.

*** Sequence orthologous to the human ENCODE targets generated from whole genome efforts in other organisms will be incorporated into the ENCODE dataset where available. The ultimate product of these efforts may vary in terms of depth of shotgun coverage.