Published online 30 October 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D871–D875***
doi:10.1093/nar/gkq1017

ENCODE whole-genome data in the UCSC genome browser (2011 update)

Brian J. Raney[1,*], Melissa S. Cline[1], Kate R. Rosenbloom[1], Timothy R. Dreszer[1], Katrina Learned[1], Galt P. Barber[1], Laurence R. Meyer[1], Cricket A. Sloan[1], Venkat S. Malladi[1], Krishna M. Roskin[1], Bernard B. Suh[1], Angie S. Hinrichs[1], Hiram Clawson[1], Ann S. Zweig[1], Vanessa Kirkup[1], Pauline A. Fujita[1], Brooke Rhead[1], Kayla E. Smith[1], Andy Pohl[1], Robert M. Kuhn[1], Donna Karolchik[1], David Haussler[1,2] and W. James Kent[1]

[1]Center for Biomolecular Science and Engineering, School of Engineering and [2]Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

Received September 15, 2010; Accepted October 9, 2010

ABSTRACT

The ENCODE project is an international consortium with a goal of cataloguing all the functional elements in the human genome. The ENCODE Data Coordination Center (DCC) at the University of California, Santa Cruz serves as the central repository for ENCODE data. In this role, the DCC offers a collection of high-throughput, genome-wide data generated with technologies such as ChIP-Seq, RNA-Seq, DNA digestion and others. This data helps illuminate transcription factor-binding sites, histone marks, chromatin accessibility, DNA methylation, RNA expression, RNA binding and other cell-state indicators. It includes sequences with quality scores, alignments, signals calculated from the alignments, and in most cases, element or peak calls calculated from the signal data. Each data set is available for visualization and download via the UCSC Genome Browser (http://genome.ucsc.edu/). ENCODE data can also be retrieved using a metadata system that captures the experimental parameters of each assay. The ENCODE web portal at UCSC (http://encodeproject.org/) provides information about the ENCODE data and links for access.

BACKGROUND

Since the Hershey–Chase experiments in the early 1950s (1), it has been known that DNA encodes heritable traits. These traits result from genes in DNA being transcribed into RNA, which might be spliced, transported to an appropriate cellular compartment, and translated into proteins. This process is regulated at many levels, including DNA methylation, chromatin modification, binding of transcription factors to the DNA, binding of splicing factors to the RNA and RNA transport. Arguably our heritable traits are determined as much through changes in regulation as differences in gene content (2).

The goal of the ENCODE project is to catalog the functional elements in the human genome that control these processes, through direct measurement using a variety of genomic technologies and detailed integrative analyses. ENCODE began with a pilot phase that focused on 1% of the genome (3), and scaled up to a genome-wide analysis in 2007.

The role of the ENCODE Data Coordination Center (DCC) is to organize and display the data generated by the labs in the consortium, and to ensure that the data meets specific quality standards when it is released to the public. Before a lab submits any data, the DCC and the lab draft a data agreement that defines the experimental parameters and associated metadata. The DCC validates incoming data to ensure consistency with the agreement. It then loads the data onto a test server for preliminary inspection, and coordinates with the labs to organize the data into a consistent set of tracks. When the tracks are ready, the DCC Quality Assurance team performs a series of integrity checks, verifies that the data is presented in a manner consistent with other browser data, and perhaps most importantly, verifies that the metadata and accompanying descriptive text are presented in a way that is useful to our users. The data is released on the public

*To whom correspondence should be addressed. Tel: +1 831 459 1477; Fax: +1 831 459 1809; Email: braney@soe.ucsc.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

UCSC Genome Browser website only after all of these checks have been satisfied.

In parallel, data is analyzed by the ENCODE Data Analysis Center, a consortium of analysis teams from the various production labs plus other researchers. These teams develop standardized protocols to analyze data from novel assays, determine best practices, and produce a consistent set of analytical methods such as standardized peak callers and signal generation from alignment pile-ups.

EXPERIMENTAL DATA AVAILABILITY

The ENCODE labs are submitting a wide range of different experiment types to the DCC (Table 1).

In addition to primary experimental data sets, the ENCODE data also includes the Gencode V4 gene annotation set (4). This set contains both manually curated and annotated gene models, as well as automatically generated annotations for regions where manual annotation has not yet been performed, along with computationally identified pseudogenes.

Other data sets estimate the mapability of regions of the human genome for given sequencing read lengths. This mapability is determined by how many times a certain DNA sequence appears in the human genome, where higher numbers limit the ability to accurately map a sequence in that region.

For experimental data to be meaningful, the user must have some information on the context of the experiment. To communicate this contextual information, the DCC has established a controlled vocabulary of experimental metadata that includes such information as cell type and experiment type, as well as experiment-specific data such as the antibody used for transcription factor-binding experiments, insert length for paired-tag experiments and cellular compartment for organelle-specific measures of transcript presence. This metadata is presented in the Genome Browser track display, and is available in the download area of each ENCODE track.

Table 1. ENCODE data types

Data type	Description
BIP	Bi-directional promoters, identified informatically
CAGE	5′ cap analysis of gene expression
ChIP-seq	DNA fragments from ChIP purifications, measured by sequencing
CNV	Copy number variation across common cell types
DGF	Digital DNase genomic footprinting
DNA PET	DNA fragments measured by paired-end di-tag sequencing
DNase-seq	Sequencing of DNase-digested DNA
Exon-array	RNA expression measured by Affymetrix exon microarrays
FAIRE-seq	Formaldehyde assisted isolation of regulatory elements
Genes	Gene annotation by Gencode
Mapability	Level of sequence uniqueness within the genome
Methyl-seq	DNA methylation, measured by sequencing
Methyl-RRBS	DNA methylation, measured by reduced representation bisulfite sequencing
Methyl27	DNA methylation, measured by Illumina bead arrays
NRE	Negative regulatory elements
ORChID	Predicted hydroxyl radical cleavage intensity on naked DNA
RIP-chip	RNA–protein interactions, measured by microarrays
RNA-chip	Tiling arrays measuring expression in various cell compartments
RNA PET	RNA expression measured by paired-end di-tag sequencing
RNA-seq	RNA expression measured by sequencing

Currently, most ENCODE tracks represent primary experimental data, as recorded by an experimental assay. As the ENCODE project matures, the DCC has begun publishing analysis tracks derived from processing the primary data. The first set of this data is displayed in Figure 1 as a *rainbow multi-wiggle* track. This composite track integrates six to eight different primary tracks representing RNA expression and histone modifications into a single display in which the signal from each track is rendered as a semi-transparent signal of a given color defined by the cell type assayed. This display facilitates the visual identification of constitutive and tissue-specific signals of cellular state.

Within each multi-cellular organism, the DNA in each cell is virtually identical, yet the regulatory signals from the configuration of this DNA yield a variety of cellular phenotypes depending on tissue type and developmental stage. To capture the breadth of these signals, the ENCODE project is surveying a variety of cellular contexts: over 100 different cell types have been designated for use. These are organized into three priority tiers: Tiers 1, 2 and 3. Table 2 lists the Tier 1 and 2 cell lines. All experiments are performed on the Tier 1 cell lines if possible. Most experiments are also performed on the Tier 2 cell lines; and selected (often lab-specific) experiments are performed on Tier 3 cell lines. This experimental design facilitates integrative analysis while capturing some breadth of cellular state.

Since late 2009 (5), the DCC has more than doubled the number of publicly available ENCODE tracks, to 863 as of August 2010. This includes 305 additional ChIP-Seq tracks that collectively now report on 98 sets of transcription factor-binding sites and 10 histone modifications. The DCC is hosting several new types of data this year, including Digital DNase Genomic Footprinting (6), ORChID predicted hydroxyl radical cleavage intensity on naked DNA (7), RIP-chip identification of RNA–protein-binding sites (8) and integrated regulatory tracks (see Figure 1). The RNA-Seq data sets have been augmented with longer reads, more paired reads and with data sets where the direction of transcription can be determined.

The human reference genome has transitioned from the hg18/NCBI36 assembly to hg19/GRCh37, and most tracks originally submitted on hg18 have been re-mapped to hg19. ENCODE data visualization has been enhanced through the deployment of new Genome Browser features, such as the ability to reorder tracks by dragging and dropping and support for BAM format (see the UCSC Genome Browser paper in this issue). The DCC ENCODE portal (http://genome.ucsc.edu/encode/) now

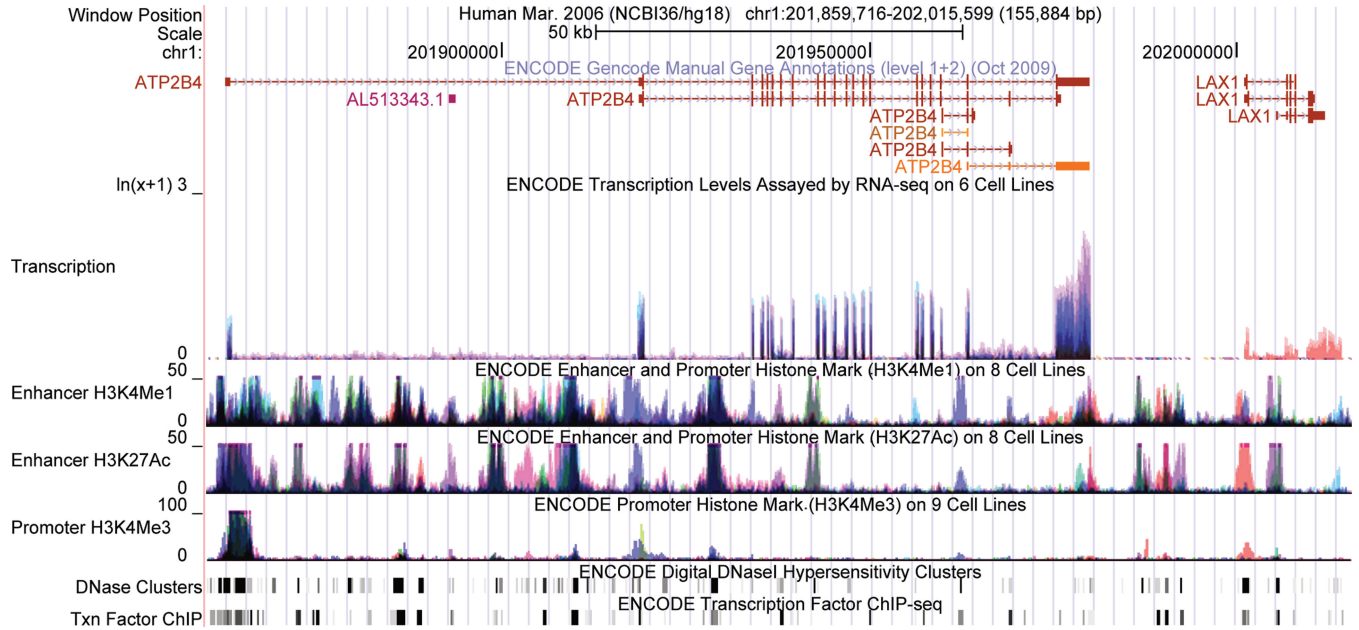

Figure 1. Constitutive and tissue-specific gene regulation under the ENCODE Integrated Regulation tracks. These tracks compare transcription, promoter marks and enhancer marks across the ENCODE Tier 1 and Tier 2 cell lines. Each cell line is rendered in one color, and darker regions indicate signals present in many cell types. ATP2B4 (left) is expressed constitutively, with the strongest expression in NHEK (lilac). Consistent with this, its leftmost promoter appears poised for activity in many tissues (indicated by the black signal in the Promoter H3K4Me3 track). Txn Factor ChIP shows abundant transcription factor binding at the promoter, and Enhancer H3K27Ac and Enhancer H3K4Me1 both show marks suggesting enhancer activity in many cell types. For contrast, LAX1 (right) shows expression only in GM12878 (orange). At the first promoter (shown in the upper isoforms), Promoter H3K4Me3 and Enhancer H3K27Ac show marks only in GM12878. At the second promoter (shown in the bottom isoform), Enhancer H3K27Ac appears active in NHEK (lilac), but Promoter H3K4Me3 does not appear to be active, suggesting that a transcriptional enhancer may be poised for activity but the promoter is not.

Table 2. The ENCODE Tier 1 and Tier 2 cell types

Tier	Cell type	Description
Tier 1	GM12878	Lymphoblastoid cells, 1000 genomes sample
	K562	Chronic myeloid leukemia
	H1-hESC	Embryonic stem cell
Tier 2	HepG2	Hepatocellular carcinoma
	HeLa-S3	Cervical carcinoma
	HUVEC	Umbilical vein endothelial cells

provides additional features for interpreting ENCODE data, such as a listing of all registered metadata variables

We have added a set of sample Genome Browser sessions to the UCSC wiki (http://genomewiki.ucsc.edu/index.php/Encode_scenarios). These sessions illustrate how ENCODE data can be used in conjunction with other Genome Browser data to support biological inferences. Each image is a screenshot that links to a Genome Browser session (9) with a preconfigured set of tracks and display parameters. Within the session, a user can access all the standard Genome Browser functionality, such as moving to a nearby region or adding a custom track to the display.

ACCESSING THE DATA

Data produced by the labs affiliated with the ENCODE project is submitted to the DCC for display and data mining. Upon release, new ENCODE tracks are announced in the News section of the ENCODE DCC portal page and on the encode-announce mailing list (https://lists.soe.ucsc.edu/mailman/listinfo/encode-announce). ENCODE data is freely available to the public under the terms of the data release policy (http://genome.ucsc.edu/ENCODE/terms.html). Data publication is restricted for nine months following the initial submission to the DCC. During this time, interested researchers are encouraged to contact the data producers to explore potential collaborations. Following this period, the data may be used without restrictions. The Data Submission Status spreadsheet (Supplementary Data S1) contains the complete list of ENCODE experiments submitted to the DCC as of August 2010.

The ENCODE whole-genome data is interspersed among non-ENCODE data tracks on the Genome Browser. As of August 2010 this data was available primarily on the NCBI36/hg18 human assembly, but new data is being mapped to the GRCh37/hg19 assembly. Each type of data from each lab is organized into one composite track, which can contain multiple sub-tracks representing experimental conditions such as cell type and other assay specific attributes. If a track has many sub-tracks, then only selected sub-tracks will be displayed by default. The sub-track display can be configured via the track configuration pages. More information about browser use and ENCODE track configuration can be found in the Genome Browser user's guide: http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html.

All data submitted to the DCC is maintained in its original state on a file server accessible through FTP and HTTP. Downloadable files are arranged by submitting lab, with each lab/data-type combination in its own directory. Each submitted file can be associated with its metadata either visually by viewing the index file with a web browser or programmatically using the *files.txt* file present in each download directory. The download area is accessible via the Download links in the track details pages, or directly at http://genome.ucsc.edu/ENCODE/downloads.html. The ENCODE data is also being accessioned at NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html).

Most ENCODE data is represented in one of three different file types, all of which can be loaded as custom tracks: BAM (10) for alignment storage, bigWig (11) for signal graphs, and extended BED files containing peak data. The bigWig and BAM data formats are loss-less, randomly accessible, and indexed so that only the data needed for the current view is read from the file, which can be local or accessed over the internet (http://genome.ucsc.edu/FAQ/FAQformat).

For users investigating the worm (*Caenorhabditis elegans*) and fruitfly (*Drosophila melanogaster*) model organisms, we highly recommend the data from the modENCODE project (http://www.modencode.org/) (12). To explore further data on human genomic variation, we recommend the 1000 Genomes Project (http://www.1000genomes.org/) (13).

To provide additional information and training on using the Genome Browser, including access and use of ENCODE data, the DCC has contracted with OpenHelix (http://www.openhelix.com/). They offer training sessions on-site and at several highly attended worldwide meetings throughout the year, as well as online tutorials and reference materials.

FUTURE WORK

In 2011, the ENCODE DCC plans to enhance both the accessibility and availability of new data. A new interface for the ENCODE portal will allow users to select tracks based on metadata, simplifying the task of determining which tracks may be relevant to their inquiries. To facilitate data browsing, UCSC is developing a track search tool that will streamline the process of locating specific data sets through the use of both free text search as well as a guided search through specific terms provided by the contributing labs to describe experiments. Microarray and sequence data generated by ENCODE will be submitted directly to the GEO and the Short Read Archive.

Over the next year, the DCC expects to provide several new forms of data, in addition to expanding the existing data sets to include new transcription factors and cell types. The new forms of data fall into three categories: analysis tracks, experimental validation and new ENCODE projects funded in 2009 through the American Recovery and Reinvestment Act (ARRA).

The ARRA-funded ENCODE projects will generate data on proteogenomics, ChIP-Seq experiments with epitope-tagged antibodies, and data derived from the mouse genome. The proteogenomics data contains peptides identified through mass spectrometry and mapped onto the genome with an HMM-based analysis (14). This extends RNA expression data by providing information on not only what is transcribed but also what is translated. Epitope tagging offers an alternative approach for generating antibodies, which can be a major challenge in ChIP-Seq experimentation. This approach offers the possibility of ChIP analysis of proteins that are recalcitrant to ChIP-grade antibody production, thus allowing the probing of a wider range of DNA-binding factors.

In 2009, the Mouse ENCODE project was created as an adjunct to ENCODE, with the goal of further understanding the human genome through a comparative analysis of mouse using analogous cell types and assays. The rule of thumb in comparative genomics is that genomic conservation suggests functional significance. Prior work suggests this to be true in RNA expression (15): RNA isoforms that occur consistently in related species largely seem more functional than species-specific RNA isoforms. However, the situation is more complex in regulatory regions, where overall conservation tends to be weaker. General associations between transcription factors and target genes can be strongly conserved, even when specific binding-site utilization is not (16). Therefore, identifying the molecular events that are conserved in human and mouse demands a careful analysis that includes measuring analogous cell types in both species. As much as possible, mouse experiments have been planned as analogs to human ENCODE experiments, to maximize the value of cross-species analysis. This experimental design focuses on three analogs to the human Tier 1 cell lines: MEL (analog of K562), CH12 (analog of GM12878) and CJ7 (analog of H1hESC). Experiments on many other mouse cell types are also planned. Many of these experiments take advantage of experimental systems that are not feasible in human, such as genetic knockouts and experiments on embryonic tissue.

CONTACTING US

Questions and feedback about the ENCODE data at UCSC should be directed to our ENCODE mailing list: encode@soe.ucsc.edu. General questions about the Genome Browser should be sent to the mailing lists described in the Genome Browser companion paper in this issue. We announce releases of new ENCODE data via the ENCODE announcement list, encode-announce@soe.ucsc.edu; to subscribe, visit https://lists.soe.ucsc.edu/mailman/listinfo/encode-announce.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the staff of the Center for Biomolecular Science and Engineering, our system administrators, Jorge Garcia, Erich Weiler, Victoria Lin and Alex Wolfe, as well as all the members of the labs in the ENCODE consortium.

FUNDING

The National Human Genome Research Institute (grant 5P41HG002371-09 to the UCSC Center for Genomic Science); (grant 5U41HG004568-02 to the UCSC ENCODE Data Coordination Center); Howard Hughes Medical Institute [(HHMI) to D.H.]. Funding for open access charge: HHMI (to D.H.).

Conflict of interest statement. All the authors except M.S.C., C.A.S., V.S.M., B.B.S. and V.K., receive royalties from the sale of UCSC Genome Browser source code licenses to commercial entities.

REFERENCES

1. Hershey,A. and Chase,M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, **36**, 39–56.
2. King,M. and Wilson,A. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
3. Birney,E., Stamatoyannopoulos,J., Dutta,A., Guigó,R., Gingeras,T., Margulies,E., Weng,Z., Snyder,M., Dermitzakis,E., Thurman,R. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
4. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C., Chrast,J., Lagarde,J., Gilbert,J., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7(Suppl. 1)**, S4.1–S4.9.
5. Rosenbloom,K., Dreszer,T., Pheasant,M., Barber,G., Meyer,L., Pohl,A., Raney,B., Wang,T., Hinrichs,A., Zweig,A. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.
6. Hesselberth,J., Chen,X., Zhang,Z., Sabo,P., Sandstrom,R., Reynolds,A., Thurman,R., Neph,S., Kuehn,M., Noble,W. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
7. Greenbaum,J., Pang,B. and Tullius,T. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.*, **17**, 947–953.
8. Baroni,T., Chittur,S., George,A. and Tenenbaum,S. (2008) Advances in RIP-chip analysis: RNA-binding protein immunoprecipitation-microarray profiling. *Methods Mol. Biol.*, **419**, 93–108.
9. Kuhn,R., Karolchik,D., Zweig,A., Wang,T., Smith,K., Rosenbloom,K., Rhead,B., Raney,B., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
10. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Subgroup,G.P.D.P. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
11. Kent,W., Zweig,A., Barber,G., Hinrichs,A. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
12. Celniker,S., Dillon,L., Gerstein,M., Gunsalus,K., Henikoff,S., Karpen,G., Kellis,M., Lai,E., Lieb,J., MacAlpine,D. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
13. Via,M., Gignoux,C. and Burchard,E. (2010) The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med.*, **2**, 3.
14. Khatun,J., Hamlett,E. and Giddings,M. (2008) Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification. *Bioinformatics*, **24**, 674–681.
15. Kan,Z., Garrett-Engele,P., Johnson,J. and Castle,J. (2005) Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Res.*, **33**, 5659–5666.
16. Weirauch,M. and Hughes,T. (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.*, **26**, 66–74.