D912–D917 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 9 November 2011*
doi:10.1093/nar/gkr1012

ENCODE whole-genome data in the UCSC Genome Browser: update 2012

Kate R. Rosenbloom[1,*], Timothy R. Dreszer[1], Jeffrey C. Long[1], Venkat S. Malladi[1], Cricket A. Sloan[1], Brian J. Raney[1], Melissa S. Cline[1], Donna Karolchik[1], Galt P. Barber[1], Hiram Clawson[1], Mark Diekhans[1], Pauline A. Fujita[1], Mary Goldman[1], Robert C. Gravell[1], Rachel A. Harte[1], Angie S. Hinrichs[1], Vanessa M. Kirkup[1], Robert M. Kuhn[1], Katrina Learned[1], Morgan Maddren[1], Laurence R. Meyer[1], Andy Pohl[1,2], Brooke Rhead[1], Matthew C. Wong[1], Ann S. Zweig[1], David Haussler[1,3] and W. James Kent[1]

[1]Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA, [2]Centre for Genomic Regulation (CRG), Barcelona, Spain and [3]Howard Hughes Medical Institute, UCSC, Santa Cruz, CA 95064, USA

Received September 15, 2011; Revised October 18, 2011; Accepted October 20, 2011

ABSTRACT

The Encyclopedia of DNA Elements (ENCODE) Consortium is entering its 5th year of production-level effort generating high-quality whole-genome functional annotations of the human genome. The past year has brought the ENCODE compendium of functional elements to critical mass, with a diverse set of 27 biochemical assays now covering 200 distinct human cell types. Within the mouse genome, which has been under study by ENCODE groups for the past 2 years, 37 cell types have been assayed. Over 2000 individual experiments have been completed and submitted to the Data Coordination Center for public use. UCSC makes this data available on the quality-reviewed public Genome Browser (http://genome.ucsc.edu) and on an early-access Preview Browser (http://genome-preview.ucsc.edu). Visual browsing, data mining and download of raw and processed data files are all supported. An ENCODE portal (http://encodeproject.org) provides specialized tools and information about the ENCODE data sets.

INTRODUCTION

Following a 4-year pilot phase aimed at identifying functional elements in selected regions comprising 1% of the human genome (1–2), the Encyclopedia of DNA Elements (ENCODE) project expanded to a whole-genome scope in September 2007 (3). Now beginning the 5th year of its mission to explore the 'dark matter' of the human genome, ENCODE contains an unprecedented range of diverse genomic data. With additional NHGRI support from the federal American Recovery and Reinvestment Act of 2009, complementary study of the mouse genome by ENCODE groups is underway. Previous manuscripts in this publication (4–5) have described the overall project and how the ENCODE Data Coordination Center at the University of California, Santa Cruz works with ENCODE labs worldwide to import their data sets, supporting documentation and metadata, and to make the data accessible to the broader biomedical community. A companion paper in this issue, 'The UCSC Genome Browser database: Extensions and updates 2012', provides background information about the UCSC Genome Browser database and infrastructure (6–7) that underlies ENCODE support at UCSC. This article focuses on ENCODE data and access tools introduced in 2011.

NEW DATA AVAILABILITY

With the increasing flood of ENCODE data production and the inevitable delays during quality review of submitted data, there arose a demand for an early access site for pre-reviewed data. In February 2011 UCSC deployed a Preview Browser (http://genome-preview.ucsc.edu) to serve this function. The Preview Browser is a weekly mirror of the UCSC internal development server. Data is made available on this site with the caveat that it is subject to change and has undergone only cursory review.

*To whom correspondence should be addressed. Tel: +1 831 459 1472; Fax: +1 831 459 1809; Email: kate@soe.ucsc.edu

The year 2011 marked the first release of Mouse ENCODE data to the public. The Mouse ENCODE project serves to complement the Human ENCODE project, furthering the understanding of human functional elements through comparative analysis. Mouse experiments aim to be analogous to those in the Human ENCODE project, as well as address experimental conditions not feasible in human, such as genetic knockouts and embryonic tissues. On the public UCSC server this year, we released mouse ENCODE results identifying transcription factor binding sites and histone marks by ChIP-seq, regions of transcription by RNA-seq, and open chromatin by DNase-seq. Data sets representing these functional elements in additional cell and tissue types, developmental stages and treatment conditions are hosted on the Preview Browser in preparation for quality review.

During the previous year the ENCODE Consortium undertook a coordinated effort to remap and re-analyze all data sets from the initial phase of data production (referenced to the March 2006 NCBI36/hg18 human genome assembly) to the current standard human reference genome (February 2009 GRCh37/hg19). At the same time, data file formats were transitioned to newer standards [BAM (8) and bigWig/bigBed (9)]. The hg19 versions of all ENCODE data are now available at UCSC.

The ENCODE human data repertoire expanded with the addition of 90 additional cell types (for a total of 235) and 57 additional transcription factor and histone modifications assayed (for a total of 177). Table 1 shows how data sets are distributed across the most intensively studied cell types.

New types of data available provided by UCSC this year include chromatin interaction maps by 5C (10) and ChIA-PET (11), nucleosome positioning by Mnase-seq , deep-sequenced DNAseI hypersensitive sites, SNP data for cell lines assayed for copy number variation, and three additional assays of RNA-binding proteins.

The Gencode Gene set (12) has been updated to version 7 (May 2011). This version features 25% more manual annotation, along with improved organization and display of the annotation to make it more intuitive to biologists. Details pages for the annotated elements show evidence used to build the annotation such as UniProt (13), CCDS (14), RefSeq (15) and GenBank (16) sequences, and PubMed IDs for published experimental evidence.

A notable addition this year was the first proteomics data within ENCODE. The new proteogenomics track features mappings of tandem mass spectrometry peptide profiles to the genome (17), complementing transcriptional evidence from RNA-based assays. The scope of DNA-binding site identification has been expanded by the introduction of epitope tagging of proteins (18) where antibodies suitable for chromatin immunoprecipitation are not available.

This year also featured two new integrative tracks provided by ENCODE analysts: a segmentation of the genome into 15 states based on the chromatin state in 9 cell lines (19) and a synthesis of multiple sources of the open chromatin state in 7 cell lines. As integrative analysis is now a major focus of Consortium efforts, more analysis tracks integrating function across primary data sets are expected in the coming year.

Table 2 lists the number of data sets currently available for each ENCODE data type.

Validation data sets to accompany primary data sets are now available for open chromatin and transcription factor binding site experiments.

NEW ACCESS INFORMATION AND TOOLS

The ENCODE portal (http://encodeproject.org), which is the centralized resource for accessing the information and tools described in this section, was extensively upgraded this year. An entire section for Mouse ENCODE resources has been added. The experimental guidelines and data standards developed by the ENCODE Consortium this year for a broad range of whole-genome assays

Table 1. ENCODE experiments in the human genome are focused on a set of cell lines selected by the Consortium for intensive study

Cell lines	Karyo	Tissue	Description	Datasets
Tier 1				
GM12878	Normal	Blood	Lymphoblastoid	166
H1-hESC	Normal	Embryonic stem	Embryonic stem	89
K562	Cancer	Blood	Leukemia	253
Tier 2 existing				
HeLa-S3	Cancer	Uterine cervix	Cervical carcinoma	118
HepG2	Cancer	Liver	Liver carcinoma	135
HUVEC	Normal	Umbilical endothelium	Umbilical vein endothelial	54
Tier 2 added in 2011				
A549	Cancer	Lung	Lung carcinoma	35
CD14+	Normal	Blood	Monocyte	2
IMR90	Normal	Lung	Lung fibroblast	3
MCF-7	Cancer	Breast	Breast carcinoma	33
SK-N-SH	Cancer	Brain	Neuroblastoma	25
Tier 3				
219 additional				928 total

All assays are performed in Tier 1; Tier 2 cell types are designated as the next level of priority.

Table 2. ENCODE encompasses a diverse set of assays

Data type	No. of experiments
Chromatin Interactions	
5C	4
ChIA-PET	6
DNA methylation	
Methyl array	63
Methyl RRBS	93
Methyl-seq	20
Histone modifications	
ChIP-seq	221
ChIP-seq (MOUSE)	28
Open chromatin	
DNase-DGF	19
DNase-seq	135
Dnase-seq (MOUSE)	27
FAIRE-seq	27
RNA profiling	
CAGE	45
Exon array	120
RNA-chip	26
RNA-PET	22
RNA-seq	151
RNA-seq (MOUSE)	27
Transcription factor binding sites	
Epitope-tag ChIP-seq	12
ChIP-seq	745
ChIP-seq (MOUSE)	92
Other	
Bi-directional promoters	1
DNA cleavage	1
DNA-PET	6
Gencode genes	5
Genotype	64
Negative regulatory elements	2
Nucleosome positioning	2
Proteogenomics	5
RNA binding proteins	49
Short read mapability	13

Descriptive overviews along with methods and references are included in the description page that accompanies all datasets.

(RNA-seq, ChIP-seq, DNase-seq, DNA methylation assays) are hosted on a dedicated portal *Data Standards* page, along with platform characterization summaries and references.

A key resource for learning about ENCODE data is the OpenHelix ENCODE tutorial (openhelix.com/ENCODE), a free Online resource released in November 2010. This tutorial provides an overview of the ENCODE project, summarizes the types of data available through ENCODE, and details methods for accessing ENCODE data via the UCSC Genome Browser. The tutorial, and accompanying instructional material, is free to the public and is sponsored by the DCC. Other resources for learning about ENCODE data usage can be found on the new ENCODE portal *Education and Outreach* page.

The DCC devoted considerable engineering effort this year to developing tools to enable users to easily locate data of interest within the overwhelming set of ENCODE data tracks and subtracks. For an overview of ENCODE data, the DCC now provides a *Data Summary* page on the ENCODE portal. This page includes a spreadsheet in multiple formats itemizing ENCODE experiments by lab, data type, cell type and other experimental variables.

The premier methods for locating ENCODE data are the new *Track Search* and *File Search* tools, available from the ENCODE portal and Genome Browser web pages. Both of these tools allow free-text searching by keyword, coupled with an advanced search feature that provides selectable lists of terms from the ENCODE controlled vocabulary (described below) to guide the search. Multiple terms can be applied in both 'and' and 'or' combinations. For example, in a single advanced search, a user can locate tracks showing evidence of the enhancer-associated histone modifications 'H3K4me1' and 'H3K27Ac' in either 'NHLF' or 'IMR90' lung cell lines. The Track Search tool is described more fully in the companion Genome Browser paper in this issue. The File Search tool locates downloadable files for analysis across the full range of ENCODE data sets, and the related track File Downloads tool (available from the track configuration page) selects files within a single track. The *Downloads* page of many ENCODE tracks include hundreds and even thousands of files. Using controlled vocabulary terms relevant for each experiment set, the files are now listed in a sortable and filterable table.

In a related effort, the DCC this year implemented an accessioning scheme to group related files and tracks within logical experiments. These accessions make it easier to relate associated files and provide a short, stable identifier for citations. Each experiment groups a set of data from a single providing laboratory for a single assay in a single cell type and set of experimental conditions. All replicates and levels of data (raw sequence files and mappings to multiple genome assemblies, processed data such as peak calls or putative transcription isoforms) associated with a single logical experiment are assigned the same accession. The DCC accession is visible everywhere metadata for a track or file appears. As of this writing, ENCODE comprises 1861 experiments in human and 174 experiments in mouse.

The ENCODE DCC controlled vocabulary (CV) is a mechanism for associating metadata with ENCODE experiments. Metadata terms are added as needed, and the metadata controlled vocabularies have been expanded this year for both human and mouse. There are currently 23 metadata controlled vocabularies. The largest vocabularies are 'Antibody' (199 terms) and 'Cell Line' (235 human and 34 mouse cell types). The CV has received extensive curation and quality review this year to ensure completeness and eliminate duplicate and confusing terms. This effort has led to a more informative set of metadata associated with each track, including links to term descriptions and supporting documents. Two specific areas where the CV was improved are the cell type karyotype and lineage terms. The karyotype term has been simplified to describe cell lines that are derived from normal or cancerous tissues. At present 72 cell lines have been annotated as normal and 47 cell lines as cancerous. The lineage term has been used to describe the progenitor tissue type from which the source tissue type has differentiated. The values ectoderm, endoderm,

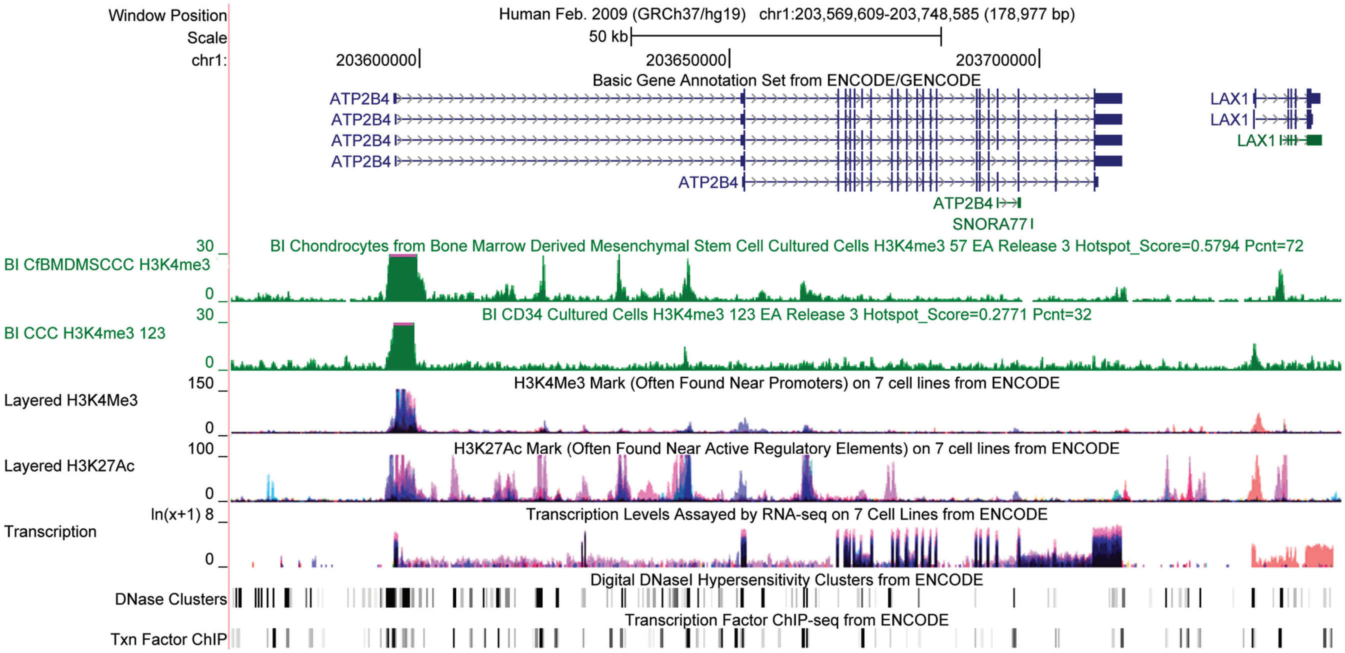

Window Position
Scale
chr1:
Human Feb. 2009 (GRCh37/hg19) chr1:203,569,609-203,748,585 (178,977 bp)
50 kb
203600000
203650000
203700000
Basic Gene Annotation Set from ENCODE/GENCODE
ATP2B4
ATP2B4
ATP2B4
ATP2B4
ATP2B4
ATP2B4
SNORA77
LAX1
LAX1
LAX1
30
BI CfBMDMSCCC H3K4me3
0
BI Chondrocytes from Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells H3K4me3 57 EA Release 3 Hotspot_Score=0.5794 Pcnt=72
30
BI CCC H3K4me3 123
0
BI CD34 Cultured Cells H3K4me3 123 EA Release 3 Hotspot_Score=0.2771 Pcnt=32
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
ln(x+1) 8
Transcription
0
Transcription Levels Assayed by RNA-seq on 7 Cell Lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE

Figure 1. ENCODE data displayed in the UCSC Genome Browser together with two annotations from the Roadmap Epigenomics Release III data hub. The genomic region contains two protein coding genes, plasma membrane calcium ATPase 4a (ATP2B4) and lymphocyte transmembrane adaptor 1 isoform a (LAX1). The GENCODE Genes track shows multiple variant transcripts for both genes as well as a snoRNA in the region. The Epigenomics Roadmap tracks just below the GENCODE track show H3K4me3, a histone mark associated with promoters, in two cell lines not assayed by the ENCODE project. These tracks show support for the short, non-coding form of LAX1 in mesenchymal stem cells, and support for the longer isoform in CD34 cells, based on peaks at likely promoter regions. The next three tracks are transparent overlays from seven cell lines assayed by the ENCODE project showing the H3K4me3 mark again, the H3K27Ac mark associated with active regulatory regions, and a log plot of transcription levels in the same cell lines. The histone marks and pattern of transcription show coordinated, cell-type-specific activity; the ATP2B4 gene is most active in NHEK (purple) and K562 (blue) cells, while LAX1 is most active in GM12878 (orange) cells. The DNAse and Transcription Factor ChIP-seq clusters shown in the last two tracks summarize data from a much wider range of cell lines and indicate a large number of regulatory regions. Additional details for these annotations are available on click-through.

mesoderm and inner cell mass are associated with 36, 45, 90 and 12 cell lines, respectively.

A new Genome Browser feature, *Data Hubs*, supports display of off-site annotations alongside ENCODE data. The first publicly provided hub presents the Roadmap Epigenomics (20) catalog of data sets, enabling close comparison of the voluminous and complementary results from these two consortia. Figure 1 shows a Genome Browser screen showcasing ENCODE and Roadmap Epigenomics data together. For more information about the *Data Hubs* feature, see the Genome Browser update in this issue.

The DCC effort to pass quality-reviewed ENCODE data to the NCBI Gene Expression Omnibus (GEO) (21) and Short Read Archive (SRA) as an auxiliary data repository has made considerable progress in the past year. Since September 2010 we have accessioned 916 GEO Samples, in 15 GEO Series in human and mouse over 3 assemblies (NCBI36/hg18, GRCh37/hg19 and NCBI37/mm9). To further organize the data and facilitate access, NCBI BioProjects have been created for ENCODE.

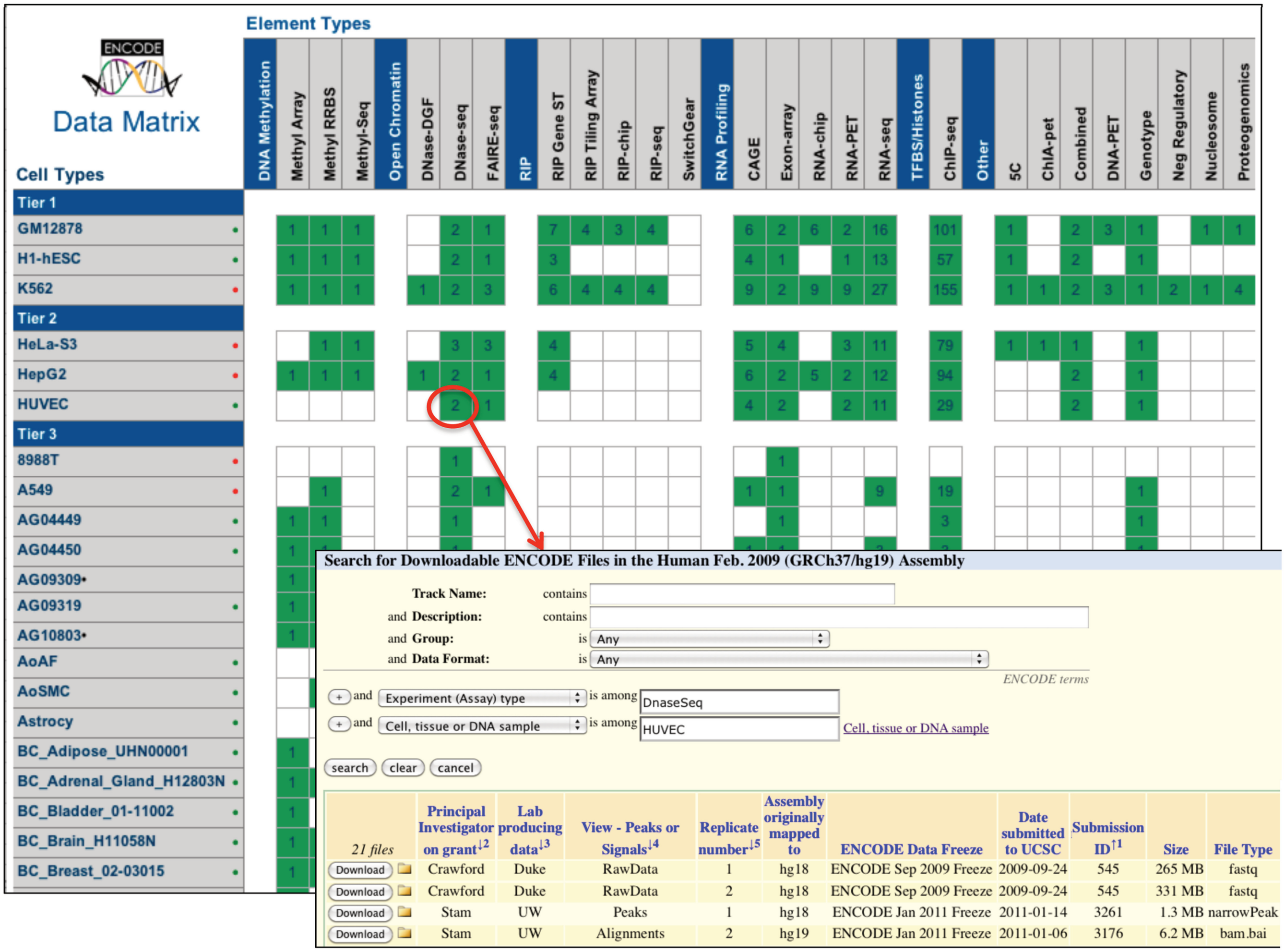
Table 3. ENCODE vital statistics, as of September 2011

Category	Human	Mouse
Experiments	1861	174
Assay types	29	3
Cell and tissue types	235	34
ChIP antibodies	179	30

ACCESSING ENCODE DATA

ENCODE data availability is summarized in Tables 1–3 in this article, and a comprehensive spreadsheet of experiments available from the ENCODE portal *Data Summary* page. Data sets marked as having 'released' status are available from the UCSC public server, http://genome.ucsc.edu. Data sets marked 'displayed' or 'reviewing' can be viewed at the preview site, http://genome-preview.ucsc.edu. Human ENCODE data is available on two human genome assemblies: NCBI36/hg18 and GRCh37/hg19. Mouse ENCODE data is provided on the mouse NCBI37/mm9 assembly.

All ENCODE data is subject to the Consortium data policy, which places some restrictions on use for the 9 months after the data becomes publicly available. Restriction timestamps for all experiments are prominently displayed on the track and file information pages, as well as being listed on the Data Summary spreadsheet. The data policy is described in detail on the *Data Policy* page of the ENCODE portal.



Cell Types	DNA Methylation			Open Chromatin			RIP					RNA Profiling					TFBS/Histones	Other							
	Methyl Array	Methyl RRBS	Methyl-Seq	DNase-DGF	DNase-seq	FAIRE-seq	RIP Gene ST	RIP Tiling Array	RIP-chip	RIP-seq	SwitchGear	CAGE	Exon-array	RNA-chip	RNA-PET	RNA-seq	ChIP-seq	5C	ChIA-pet	Combined	DNA-PET	Genotype	Neg Regulatory	Nucleosome	Proteogenomics
Tier 1																									
GM12878 ●	1	1	1		2	1	7	4	3	4		6	2	6	2	16	101	1		2	3	1		1	1
H1-hESC ●	1	1	1		2	1	3					4	1		1	13	57	1		2		1			
K562 ●	1	1	1	1	2	3	6	4	4	4		9	2	9	9	27	155	1	1	2	3	1	2	1	4
Tier 2																									
HeLa-S3 ●		1	1		3	3	4					5	4		3	11	79	1	1	1		1			
HepG2 ●	1	1	1	1	2	1	4					6	2	5	2	12	94			2		1			
HUVEC ●					2	1						4	2		2	11	29			2		1			
Tier 3																									
8988T ●					1								1												
A549 ●		1			2	1						1	1			9	19					1			
AG04449 ●	1	1			1								1				3					1			
AG04450 ●	1																								
AG09309•	1																								
AG09319 ●	1																								
AG10803•	1																								
AoAF ●																									
AoSMC ●																									
Astrocy ●																									
BC_Adipose_UHN00001 ●	1																								
BC_Adrenal_Gland_H12803N ●	1																								
BC_Bladder_01-11002 ●	1																								
BC_Brain_H11058N ●	1																								
BC_Breast_02-03015 ●	1																								

Search for Downloadable ENCODE Files in the Human Feb. 2009 (GRCh37/hg19) Assembly

Track Name: contains

and **Description:** contains

and **Group:** is Any

and **Data Format:** is Any

ENCODE terms

+ and Experiment (Assay) type is among DnaseSeq

+ and Cell, tissue or DNA sample is among HUVEC Cell, tissue or DNA sample

21 files		Principal Investigator on grant↓2	Lab producing data↓3	View - Peaks or Signals↓4	Replicate number↓5	Assembly originally mapped to	ENCODE Data Freeze	Date submitted to UCSC	Submission ID↑1	Size	File Type
Download		Crawford	Duke	RawData	1	hg18	ENCODE Sep 2009 Freeze	2009-09-24	545	265 MB	fastq
Download		Crawford	Duke	RawData	2	hg18	ENCODE Sep 2009 Freeze	2009-09-24	545	331 MB	fastq
Download		Stam	UW	Peaks	1	hg18	ENCODE Jan 2011 Freeze	2011-01-14	3261	1.3 MB	narrowPeak
Download		Stam	UW	Alignments	2	hg19	ENCODE Jan 2011 Freeze	2011-01-06	3176	6.2 MB	bam.bai

Figure 2. Data matrix display and selection of files for download. This feature will be linked to the ENCODE portal, and will navigate to the Advanced Search features of File and Track Search.

ENCODE GEO submissions are listed on the GEO ENCODE summary page, http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html. ENCODE has been assigned NCBI BioProject identifiers to further organize the data: PRJNA30707 for Human ENCODE (with the subproject PRJNA63443 for Production phase data) and PRJNA50617 for Mouse ENCODE. Data in each project is further categorized as *epigenomic*, *functional genomics* or *transcriptome*.

FUTURE WORK

Highlights of the fifth and final year of this phase of the ENCODE project will be the fruition of ongoing integrative analysis efforts and dissemination of the results to the DCC, promotion of an additional collection of cell types for Consortium-wide use (see Table 1), expansion of the transcription factor space based on community input, selected new experiment types in high-value areas such as single-cell assays, and additional validation data sets. The Mouse ENCODE project makes its future experiment planning publicly available on the ENCODE portal Mouse Data Summary page.

DCC efforts during the 5th year will continue to emphasize data accessibility and usability. We have scheduled an update to the OpenHelix ENCODE tutorial, and are contracting for the design and production of ENCODE Quick Reference Cards. A new *Data Matrix* web application on the portal will provide table and matrix-based display of the breadth of ENCODE data, with click-through access to search results for selected experiments. Figure 2 shows a snapshot as of September 2011. We expect to release this feature on the ENCODE portal by late fall 2011.

In upcoming months we expect the new data hub feature will be adopted more widely, and we anticipate that the larger ENCODE production groups will migrate to hub-based hosting of much of their data. The DCC will be implementing search across data hubs to further enhance the synergy between UCSC-hosted and remote data sources.

CONTACT INFORMATION

General questions and feedback about ENCODE data at UCSC should be directed to the ENCODE mailing list:

encode@soe.ucsc.edu. General questions about the Genome Browser should be sent to the UCSC browser mailing list: genome@soe.ucsc.edu. Specific questions about details of laboratory methods or data interpretation should be directed to the ENCODE laboratory contact listed on the description page for that data set. We announce releases of new ENCODE data via the ENCODE announcement list. To subscribe, visit https://lists.soe.ucsc.edu/mailman/listinfo/encode-announce.

ACKNOWLEDGEMENTS

We would like to thank the systems administration staff at the Center for Biomolecular Science and Engineering: Jorge Garcia, Erich Weiler, Victoria Lin and Gary Moro, for their dedication and support, keeping high-volume ENCODE data flowing to our public site while assuring our servers are reliable and available. Thanks also to members of the ENCODE Consortium for providing these valuable data sets.

FUNDING

National Human Genome Research Institute (grants 5P41HG002371-10 and 3P41HG002371-10S1 to the UCSC Center for Genomic Science, and grant 5U41HG004568-04 and 3U41HG004568-03S1 to the UCSC ENCODE Data Coordination Center); Howard Hughes Medical Institute (to D.H.). Funding for the open access charge: The Howard Hughes Medical Institute.

Conflict of interest statement. The authors receive royalties from the sale of UCSC Genome Browser source code licenses to commercial entities.

REFERENCES

1. ENCODE Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
2. The ENCODE Project Consortium, Birney,E., Stamatoyannopoulos,J., Dutta,A., Guigó,R., Gingeras,T., Margulies,E., Weng,Z., Snyder,M., Dermitzakis,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
3. Myers,R.M., Stamatoyannopoulos,J., Snyder,M., Dunham,I., Hardison,R.C., Bernstein,B.E., Gingeras,T.R., Kent,W.J., Birney,E., Wold,B. *et al.* (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
4. Rosenbloom,K.R., Dreszer,T.R., Pheasant,M., Barber,G.P., Meyer,L.R., Pohl,A., Raney,B.J., Wang,T., Hinrichs,A.S., Zweig,A.S. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.
5. Raney,B.J., Cline,M.S., Rosenbloom,K.R., Dreszer,T.R., Learned,K., Barber,G.P., Meyer,L.R., Sloan,C.A., Malladi,V.S., Roskin,K.M. *et al.* (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.
6. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
7. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
8. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009). 1000 Genome Project Data Processing Subgroup. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
9. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
10. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
11. Li,G., Fullwood,M.J., Xu,H., Mulawadi,F.H., Velkov,S., Vega,V., Ariyaratne,P.N., Mohamed,Y.B., Ooi,H.S., Tennakoon,C. *et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, **11**, R22.
12. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7(Suppl. 1)**, S41–S49.
13. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
14. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
17. Krug,K., Nahnsen,S. and Macek,B. (2011) Mass spectrometry at the interface of proteomics and genomics. *Mol Biosyst.*, **7**, 284–291.
18. Poser,I., Sarov,M., Hutchins,J.R., Heriche,J.K., Toyoda,Y., Pozniakovsky,A., Weigl,D., Nitzsche,A., Hegemann,B., Bird,A.W. *et al.* (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat Methods.*, **5**, 409–415.
19. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
20. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
21. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.