An encyclopedia of mouse DNA elements (Mouse ENCODE)

Mouse ENCODE Consortium, John A Stamatoyannopoulos1, Michael Snyder2, Ross Hardison3, Bing Ren4, Thomas Gingeras5, David M. Gilbert6, Mark Groudine7, Michael Bender8, Rajinder Kaul9, Theresa Canfield10, Erica Giste11, Audra Johnson12, Mia Zhang13, Gayathri Balasundaram14, Rachel Byron15, Vaughan Roach16, Peter J Sabo17, Richard Sandstrom18, A Sandra Stehling19, Robert E Thurman20, Sherman M Weissman21, Philip Cayting22, Manoj Hariharan23, Jin Lian24, Yong Cheng25, Stephen G Landt26, Zhihai Ma27, Barbara J Wold28, Job Dekker29, Gregory E Crawford30, Cheryl A Keller31, Weisheng Wu32, Christopher Morrissey33, Swathi A Kumar34, Tejaswini Mishra35, Deepthi Jain36, Marta Byrsk-Bishop37, Daniel Blankenberg38, Bryan R Lajoie39, Gaurav Jain40, Amartya Sanyal41, Kaun-Bei Chen42, Olger Denas43, James Taylor44, Gerd A Blobel45, Mitchell J Weiss46, Max Pimkin47, Wulan Deng48, Georgi K Marinov49, Brian A Williams50, Katherine I Fisher-Aylor51, Gilberto Desalvo52, Anthony Kiralusha53, Diane Trout54, Henry Armhein55, Ali Mortazavi56, Lee Edsall57, David McCleary58, Samantha Kuan59, Yin Shen60, Feng Yue61, Zhen Ye62, Carrie A Davis63, Chris Zaleski64, Sonali Jha65, Chenghai Xue66, Alex Dobin67, Wei Lin68, Meagan Fastuca69, Huaien Wang70, Roderic Guigo71, Sarah Djebali72, Julien Lagarde73, Tyrone Ryba74, Takayo Sasaki75, Venkat S Malladi76, Melissa S Cline77, Vanessa M Kirkup78, Katrina Learned79, Kate R Rosenbloom80 and W James Kent81, Elise A Feingold82, Peter J Good83, Michael Pazin84, Rebecca F Lowdon85 and Leslie B Adams86

Abstract
To complement the human Encyclopedia of DNA Elements (ENCODEx) project and to enable a broad range of mouse genomics efforts, the Mouse ENCODE Consortium is applying the same experimental pipelines developed for human ENCODE to annotate the mouse genome.

Keywords ENCODE Project, mouse genome, DNaseI hypersensitive sites, histone modifications, transcriptome, transcription factor binding sites, comparative genomics, ChIP-seq, RNA-seq

Background
The laboratory mouse is the premier mammalian model organism for the study of human disease, and it has played a vital role in both the annotation of the human genome and the study of gene function and regulation. Similar to humans, mice naturally develop diverse diseases that affect the hematologic, nervous, cardiovascular, endocrine, musculoskeletal, renal and other systems, providing excellent experimental paradigms for studying the pathogenesis of cancer, autoimmune disease, diabetes, obesity, atherosclerosis, hypertension, gastrointestinal disorders and diverse neurodegenerative states. Mouse models are currently available for hundreds of human disorders [1-4], spanning diverse quantitative and behavioral phenotypes and physiological systems. These comprise both inbred strains and genetically engineered mutants, many of which have been extensively characterized. For these reasons, the mouse has emerged as a premier system for translating basic human genetic, genomic and physiologic research into paradigms for therapeutic development.

The mouse genome has been uniquely useful in annotating the human genome and advancing the understanding of human gene functions. At 2.7 Gb, the mouse genome is of comparable size and structure with the human genome, and 99% of mouse genes have human orthologs. Because of the availability of inbred strains and the facile and rapid features of mouse breeding, the mouse has played a vital role in decoding fundamental features of gene function and regulation during developmental and differentiation intervals that are either difficult or impossible to study systematically in humans. An ideal evolutionary distance for human comparative genomics (circa 200 million years) has made the mouse...
The stability of the sequences that regulate most genes [7–12]. The apparent witnessed marked divergence in the non-coding DNA largely stable during an evolutionary interval that regulation with striking precision, indicating that the mouse germline recapitulate features of human gene of the mouse system. Many human genes integrated into are routinely taken for granted emphasize the great utility fundamental aspects of transgenic gene regulation that demonstrate the potential of the mouse genome to inform studies of human gene regulation; indeed, transgenic mice have become a routine part of the repertoire of modern molecular and developmental biology. Many transgenic mice have become a routine part of the repertoire of modern molecular and developmental biology. Many fundamental aspects of transgenic gene regulation that are routinely taken for granted emphasize the great utility of the mouse system. Many human genes integrated into the mouse germline recapitulate features of human gene regulation with striking precision, indicating that the trans-acting regulatory environment has remained largely stable during an evolutionary interval that witnessed marked divergence in the non-coding DNA sequences that regulate most genes [7–12]. The apparent stability of the trans-acting regulatory environment renders the mouse uniquely useful for studies of transcriptional regulation by mutagenesis of human DNA that is then transferred into mouse. Engineered mutations in transgenic mice frequently show phenotypes analogous to those of naturally occurring mutations in humans.

The Mouse ENCODE Project Consortium

By undertaking a parallel Mouse ENCODE Project that utilizes the same technologies and pipelines developed for the human ENCODE Project [13–15], the Mouse ENCODE Consortium aims to (i) enhance the value of the human ENCODE Project through relevant comparisons; (ii) access cell types, tissues, and developmental time points that are not addressable by the human project; and (iii) provide a general resource to inform and accelerate ongoing efforts in mouse genomics and disease modeling with human translational potential.

The organization of the Mouse ENCODE Consortium includes data production centers and a data coordination center (DCC). Production centers generally focus on different data types, including transcription factor and polymerase occupancy, DNase hypersensitivity, histone modification and RNA transcription. The DCC is co-located with the human ENCODE Project DCC [15] at the University of California Santa Cruz (UCSC), USA.

A web-based portal site (MOUSE ENCODE [16]) has been established to consolidate and distribute information on Mouse ENCODE consortium goals, data, protocols and publications.

Mouse ENCODE data types

The Mouse ENCODE Project is analyzing primary mouse cells and tissues spanning a range of tissue types, developmental time points, as well as model cell lines. To ensure consistency, the project is focusing on C57BL/6-derived cells and tissues, except for the case of certain widely used model cell lines. Primary tissues are harvested from age-matched mice using standardized protocols on mice either bred locally or obtained from standard sources (The Jackson Laboratory, Bar Harbor, Maine, USA; Charles River Laboratories, Wilmington, Massachusetts, USA). Following the practice of the human ENCODE Project [14], model cell lines are cultured using standard operating procedures that are reviewed for consistency and clarity. Among the cell lines in use are those selected as analogs to several human ENCODE common cell lines [14], including K562 (mouse erythroleukemia cell line MEL (ATCC)), GM12878 (mouse lymphoid cell line CH12 (ATCC)), and H1 embryonic stem cells (E14 mouse embryonic stem cells).

Accessing Mouse ENCODE data

The Mouse ENCODE Project has already generated and released hundreds of data sets through the UCSC browser [17,18] (Table S1 in Additional file 1). All data sets are also deposited with the Gene Expression Omnibus (GEO) repository after public release through the UCSC browser. The data sets shown in Table S1 in Additional file 1 span many high-utility data types generated using state-of-the-art approaches, including DNase hypersensitive sites by DNase-seq [19], DNase footprints by digital genomic footprinting [20], RNA-seq [21], histone modifications by ChIP-seq [22], transcription factor and polymerase occupancy sites by ChIP-seq [23], and DNA replication timing by Repli-chip [24]. In addition, selected chromosomal regions will be interrogated for chromatin interactions by 5C [25], including the entirety of mouse chromosome 12. All data are collected from at least two biological replicates, and all replicate data are also available through the Mouse ENCODE repository at UCSC. An up-to-date log of Mouse ENCODE data releases can be found [16], and it is also linked through the home page of the ENCODE project [26]. Submissions are ongoing, and an updated summary timeline for major data types is available [18].
To ensure the quality and consistency of experimental procedures used at each data production center, the Consortium has selected a single reference cell type (MEL) on which all experimental approaches are being applied. For other cell and tissue types, the data types vary, with DNasel sensitivity, histone modifications and RNA-seq focused mainly on primary tissues, and transcription factor binding generally focused on model cell lines (Table S1 in Additional file 1). A comprehensive collection of cell culture and tissue sample preparation protocols utilized by the Consortium is available online [27].

Data production standards and assessment of data quality
The Mouse ENCODE Consortium is applying the same data generation, quality control, analysis pipelines and data standard developed for the human ENCODE Project. Working copies of data standard documents are available as an appendix to the recently published User’s Guide to ENCODE Data [14] and at the home page of the ENCODE project [26]. Consortium data undergo quality review at the level of the production centers to ensure experimental success and generation of high-quality data, and subsequently at the DCC (see below) to ensure accurate visualization, and links to primary data files and metadata.

Data availability
Mouse ENCODE data are available online through the UCSC browser mm9 mouse genome sequence build [17] and through a dedicated Mouse ENCODE mirror browser linked to the portal site [18]. Data in the UCSC browser can be viewed readily in the context of other genome annotations available for the mouse genome. An online tutorial developed for facilitating the viewing of human ENCODE data is also directly applicable to the Mouse ENCODE data [28]. Detailed instructions are also provided for the data download and analysis functions available in the browser. DNA sequence reads from Mouse ENCODE ChIP-seq, DNase-seq and RNA-seq are available for direct retrieval from the UCSC browser archive [29] and the GEO repository [30].

Data release and use policy
The Mouse ENCODE data are rapidly released soon after they are verified (that is, shown to be reproducible) to facilitate their immediate utility to the broader community. A log of data releases is available at the Mouse ENCODE portal site [18] and through the main UCSC browser [17]. The terms of data use are described under the ENCODE Data Release and Use Policy [31]. As with human ENCODE, data are made available following quality review and standardization of formatting. While Mouse ENCODE data are made freely available for viewing and pre-publication analysis upon release, data use for genome-wide analysis in papers, abstracts or public presentations is restricted during the first 9 months following public release. The expiration of this ‘embargo’ period for genome-wide analyses is clearly marked in the track titles of Mouse ENCODE data in the UCSC browser. Mouse ENCODE data are immediately available for analysis of individual gene loci.

Data analysis plans
Production groups are engaged in analysis of the individual data types generated by each group. In addition, the Mouse ENCODE Consortium is currently in the planning stages of an integrated analysis. Integration of multiple mouse ENCODE data types will be performed to assess the extent of annotation of the mouse genome, and to illuminate general features of mouse gene and chromosomal regulation. Mouse ENCODE data will also be extensively integrated with human ENCODE data in order to study the evolution of gene regulatory mechanisms, and to cross-validate findings within both the human and mouse projects. Integration with data from invertebrates (Drosophila melanogaster and Caenorhabditis elegans) generated under the ModENCODE project may also yield insights into common gene regulatory mechanisms and conserved pathways. While it is expected that broad features of regulatory mechanisms will be conserved across animal phyla, the integrative and comparative analyses enabled by the Mouse ENCODE project will provide a unique opportunity for systematic study of both conservation of function and biochemical activity relative to conservation of sequence per se. The Consortium expects to conduct global analyses with an emphasis on integration with the human ENCODE Project, and not to focus on specific genes, genomic regions, tissues/cell states or pathways.

Joining the Mouse ENCODE Consortium
Following on the model of the human ENCODE Consortium, which currently counts hundreds of members worldwide, the Mouse ENCODE Consortium is an open scientific venture that welcomes scientists at all levels and with all types of relevant expertise. More information on joining the human or mouse ENCODE Consortia is available [26].

Perspective
In summary, the laboratory mouse is a powerful tool for the investigation of human gene function and for dissecting the genetic and transcriptional regulatory circuits controlling development and homeostasis of mammals. The Mouse ENCODE Project aims to potentiate both the utility of the mouse as a model for regulatory genomics and the human ENCODE project effort to advance annotation of the human genome.
Additional file 1

Table S1. Overview of Mouse ENCODE data. Snapshot of data generated by the Mouse ENCODE Consortium and released through University of California Santa Cruz (UCSC) browser. Vertical axis: cell lines and ex vivo cells and tissues. The originating cell type is shown in parentheses next to each line. For mouse embryonic or fetal tissues, the developmental day of harvest is shown in parentheses. Unless otherwise noted, all other tissues are from adult animals. Horizontal axis: experimental assays, including DNasel hypersensitivity sites (DNasel), DNasel footprints by digital genomic footprinting (DGF), miRNA-seq; DNA replication type by Repli-chIP; modifications to histone H3, including bulk acetylation (H3ac) or modification to specific lysine positions on the H3 tail; ChIP-seq for polymerase (Pol2, Pol2-4H8), histone acetyltransferase p300, and diverse transcription factors; and ChIP input control (far right lane). Filled cells indicate that an assay has been performed by the indicated production center(s) and data released to UCSC.

Abbreviations
ATCC, American Type Culture Collection; DCC, data coordination center; GEO, Gene Expression Omnibus; UCSC, University of California Santa Cruz.

Competing interests
The authors declare that they have no competing interests.

Mouse ENCODE Consortium Authors
Writing Group:
John A Stamatoyannopoulos1, Michael Snyder3, Ross Hardison7, Bing Ren12

University of Washington-Fred Hutchinson Cancer Research Center Group:
John A Stamatoyannopoulos1, Mark Groudine2, Michael Bender2, Rajinder Kaul, Theresa Canfield, Erica Gstei, Audra Johnson, Mia Zhang, Gayathri Balasundaram1, Rachel Byron1, Vaughan Roach1, Peter Sabo1, Richard Sandstrom1, A Sandra Stehling1, Bob Thurman1

Stanford-Yale Group:
Michael Snyder1, Sherman M Weissman1, Philip Cayting2,5,6, Manoj Hariharan3, John A Stamatoyannopoulos1, Mark Groudine2, Michael Bender2, Bing Ren12, Feng Yue12, Zhen Ye12

Penn State/University of Massachusetts/Duke University/Emory University/California Institute of Technology/University of California, Irvine/Children's Hospital of Philadelphia Group:
Ross Hardison1, Barbara J Wold4, Joe Dekker1, Gregory Crawford1, Cheryl A Kelli1, Weisheng Wu1, Christopher Morrissey1, Swathi A Kamar1, Tejaswini Mishra3, Deepthi Jain4, Marta Byrska-Bishop1, Daniel Blankenberg1, Bing Ren12, Weisheng Wu7, Christopher Morrissey7, Swathi A Kumar7

University of California Santa Cruz Group:
Bing Ren1, Lee Edsall1, David McCleary1, Samantha Kuan1, Yen Shen1, Feng Yue1, Zhen Ye1

Cold Spring Harbor Laboratory/CRG Group:
Thomas R Gingeras1, Carne A Davis1, Chris Zaleski1, Sonali Jha1, Chenghai Xue1, Alex Dobin1, Wei Lin1, Meagan Fastuca1, Huaien Wang1, Roderic Guigo3, Sarah Djebali1, Juliet Lagarde1, Daniel Blankenberg1, Cheryl A Keller7, Weisheng Wu7, Christopher Morrissey7, Swathi A Kumar7

Florida State University Group:
David M Gilbert5, Tyrone Ryba1, Takayo Sasaki1, Maria Cherry1

Data Coordination Center at University of California Santa Cruz:
Venkat S Malladi1, Melissa S Cline1, Vanessa M Kirkup1, Katrina Learned1, Kate R Rosenbloom1 and W James Kent1,2

NHGRI Project Management Group:
Elise A Feingold1, Peter J Good1, Michael Pazin1, Rebecca F Lowdon1, Leslie B Adams1

Author details
1 Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America. 2 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America. 3 Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America. 4 Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California San Diego, La Jolla, California, United States of America. 5 Dept. of Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America. 6 Department of Biological Science, Florida State University, Tallahassee, Florida, United States of America. 7 Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America. 8 National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America. 9 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America. 10 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America. 11 Department of Pediatrics, Duke University, Durham, North Carolina, United States of America. 12 Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America. 13 Div. of Biology, California Institute of Technology, Pasadena, California, United States of America. 14 Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America. 15 Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America. 16 Department of Development and Cell Biology, University of California Irvine, Irvine California, United States of America. 17 Division of Bioinformatics and Genomics, Center for Genomic Regulation, Barcelona, Catalunya, Spain. 18 Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz (UCSC), Santa Cruz, California, United States of America. 19 National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America.

Published: X Month 2012

References
1. Hardouin SN, Nagy A. Mouse models for human disease. Clin Genet 2000, 57:237–244.
2. Bedell MA, Langaesaeda DA, Jenkins NA, Copeland NG. Mouse models of human disease. Part II: Recent progress and future directions. Genes Dev 1997, 11:41–43.
3. Rees DA, Alcolado JC. Animal models of diabetes mellitus. Diabet Med 2005, 22:559–70.
4. Holt, BD, Nadeau JR. Phenotype-driven genetic approaches in mice: high-throughput phenotyping for discovering new models of cardiovascular disease. Trends Cardiovasc Med 2001, 11:52–89.
5. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal A, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Caminca P et al.: Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420:530–562.
6. Miller W, Makova KD, Neufeldt RN, Hardison RC. Comparative genomics. Annu Rev Genomics Hum Genet 2004, 5:15–56.
7. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Maucci E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungeisen K, Kent WJ, Kossida D, Lara M, et al.: A high-resolution map of human evolutionary constraint using 29 mammals. Nature 2011, 478:476–482.
8. Moreno C, Lazar J, Jacob HJ, Kvittek AE. Comparative genomics for detecting human disease genes. Adv Genet 2008, 60:655–697.
9. Visel A, Akyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA. Functional...
autonomy of distant-acting human enhancers. Genomics 2009, 93:509-513.

10. Pennacchio LA, Visel A: Limits of sequence and functional conservation. Nat Genet 2010, 42:557-558.

11. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM: Deletion of ultraconserved elements yields viable mice. PLoS Biol 2007, 5:e234.

12. Cheng JF, Priett JR, Pennacchio LA: Comparative genomics: a tool to functionally annotate human DNA. Methods Mol Biol 2007, 366:229-251.

13. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004, 306:636-640.

14. ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Zweig AS, Haussler D, Fejes Tóth K, Kellis M, Ren B: Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009, 459:108-112.

15. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh B, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Haussler D, Kent WJ: ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 2011, 39:D871-875.

16. UCSC Genome Bioinformatics [http://genomewiki.ucsc.edu]

17. MOUSE ENCODE Data [http://www.mouseencode.org]

18. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet 2011, 43:264-268.

19. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Nebh S, Kuehn MS, Noble WS, FIELDS S, Stamatoyannopoulos JA: Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 2009, 6:283-289.

20. Moritz H, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5:621-628.