An Encyclopedia of Mouse DNA Elements (Mouse ENCODE)

Mouse ENCODE Project Consortium*

** Author list at end of manuscript*

Abstract

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. While the mouse genome sequence has been available for nearly a decade, little progress has been made in systematically annotating the non-coding regions that comprise 98% of the genome. To complement the human Encyclopedia of DNA Elements (ENCODE) project and to enable a broad range of mouse genomics efforts, the Mouse ENCODE Consortium is now applying many of the same experimental pipelines developed for human ENCODE in order to annotate the mouse genome. The Mouse ENCODE Consortium has already generated and released hundreds of genome-wide data sets profiling chromatin structure, transcription factor binding, and RNA transcription landscapes across multiple tissue types, developmental time points, and model cell lines. The Consortium is also developing tools to facilitate comparative analysis of mouse and human functional genomic annotations. The resulting publicly available resources should be of broad utility for the study of both human and mouse gene regulation and the molecular basis of human disease.

Background

The laboratory mouse *Mus musculus* is now well-established as the leading mammalian model system for the investigation of human diseases. 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 2.7Gb, 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 (c. 200M years) has made the mouse genome a standard for comparative genomic analyses seeking to illuminate human functional DNA [5-7].

Less than 2% of the mouse genome is currently believed to comprise protein-coding regions. Among the vast non-coding sequences lie numerous yet-to-be-identified functional DNA elements that regulate diverse genomic processes, including transcriptional regulation, meiotic recombination, and DNA replication and repair. A major focus of the MouseENCODE Project is to identify comprehensively transcriptional regulatory elements in the mouse genome, providing a valuable resource for understanding the genetic circuitry that controls animal development and lineage specification. It is expected that millions of *cis*-regulatory elements lie within mouse non-coding regions, many of which are conserved in human DNA. As such, comprehensive illumination of mouse elements should greatly facilitate the functional annotation of human genome.

Hundreds of human-to-mouse transgenic studies 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 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, 8-12]. 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 comparative studies; (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, DNaseI hypersensitivity, histone modification, and RNA transcription. The DCC is co-localized with the human ENCODE Project DCC[15] at the University of California Santa Cruz.

A web-based portal site (http://www.mouseencode.org) 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 using standardized protocols on mice either bred locally or obtained from standard sources (Jackson Labs, Charles River Laboratories). 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 (ES) cells (E14 Mouse ES cells).

Accessing Mouse ENCODE Data

The Mouse ENCODE Project has already generated and released hundreds of data sets through the UCSC browser (http://genome.ucsc.edu, http://www.mouseencode.org/data) (**Figure 1**). All data sets are also deposited with the GEO repository after public release through the UCSC browser. The data sets shown in **Figure 1** span many high-utility data types generated using state-of-the-art approaches including DNaseI hypersensitive sites by DNase-seq[16]; DNaseI footprints by Digital Genomic Footprinting[17]; RNA-seq[18]; histone modifications by ChIP-seq[19]; and transcription factor and polymerase occupancy sites by ChIP-seq[20]. In addition, selected chromosomal regions will be interrogated for chromatin interactions by 5C [21], 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 at http:.//www.mouseencode.org, and is also linked through the home page at http://www.encodeproject.org. Submissions are ongoing, and an updated summary timeline for major data types is available at http://www.mouseencode.org/data/summary.

To ensure the quality and consistency of experimental procedures used at each Data Production Center, the Consortium is 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 DNaseI sensitivity, histone modifications, and RNA-seq focused mainly on primary tissues, and transcription factor binding generally focused on model cell lines (**Figure 1**). A comprehensive collection of cell culture and tissue sample preparation protocols utilized by the Consortium is available online (http://www.mouseencode.org/protocols).

Data production standards and assessment of data quality

The MouseENCODE 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 http://www.encodeproject.org. 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 Data Coordination Center (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 (http://genome.ucsc.edu) and through a dedicated Mouse ENCODE mirror browser linked to the portal site (http://www.mouseencode.org/data). 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 (http://www.openhelix.com/ENCODE). 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 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/) and the GEO repository (http://www.ncbi.nlm.nih.gov/geo/).

Data release and use policy

The Mouse ENCODE data are rapidly released soon after they are verified (i.e., 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 (www.mouseencode.org/data/releases) and through the main UCSC browser. The terms of data use are described under the ENCODE Data Release and Use Policy (http://www.encodeproject.org/ENCODE/terms.html). 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 (*D. melanogaster* and *C. 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 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 are available at http://www.encodeproject.org.

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 MouseENCODE 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.

FIGURES

FIGURE 1. Overview of Mouse ENCODE Data. Snapshot of data generated by the Mouse ENCODE Consortium and released through UCSC browser. V*ertical axis*: Cell lines and *ex vivo* cells and tissues. The originating cell type is shown in parenthesis 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 DNaseI hypersensitivie sites (DNaseI), DNaseI footprints by Digital Genomic Footprinting (DGF); standard mRNA-seq (RNA-seq); long mRNA-seq (long RNA-seq); 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 transcrition factors; and ChIP input control (far right lane). Filled cells indicate that an assay has been performed by the indicated Production Center(s) (color legend) and data released to UCSC.

REFERENCES

1. Hardouin SN, Nagy A: **Mouse models for human disease**. *Clin Genet*. 2000, 57:237-44. .
2. Bedell MA, Largaespada DA, Jenkins NA, Copeland NG: **Mouse models of human disease. Part II: Recent progress and future directions**. *Genes Dev*. 1997, 11:11-43.
3. Rees DA, Alcolado JC: **Animal models of diabetes mellitus.** *Diabet Med*. 2005, 22:359-70.
4. Holt, BD, Nadeau JH: **Phenotype-driven genetic approaches in mice: high-throughput phenotyping for discovering new models of cardiovascular disease**. *Trends Cardiovasc Med*. 2001, 11:82-89.
5. Chinwalla AT, et. al: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, 420:520-562.
6. Miller W, Makova KD, Nekrutenko A, Hardison RC: **Comparative genomics**. *Annu Rev Genomics Hum Genet*. 2004, 5:15-56.
7. PMID: 21993624 Lindblad-Toh K, et. al: **A high-resolution map of human evolutionary constraint using 29 mammals**. *Nature* 2011, 478:476-82.
8. Moreno C, Lazar J, Jacob HJ, Kwitek AE: **Comparative genomics for detecting human disease genes**. *Adv Genet* 2008, 60:655-97.
9. Visel A, Akiyama 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(9):e234.
12. Cheng JF, Priest JR, Pennacchio LA: **Comparative genomics: a tool to functionally annotate human DNA**. *Methods Mol Bio*l. 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, Crawford GE: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol.* 2011, 9(4):e1001046.
15. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent WJ. **ENCODE whole-**

genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res*. 2011, 39:D871-5.

16. 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, 43:264-8.
17. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nat Methods.* 2009, 6:283-9.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, 5:621-8.
19. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, 459:108-12.
20. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S**Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, 4:651-7.
21. van Berkum NL, Dekker J: **Determining spatial chromatin organization of large genomic regions using 5C technology.** *Methods Mol Biol.* 2009, 567:189-213.

Mouse ENCODE Consortium Authors

Writing Group:
John A. Stamatoyannopoulos[i], Michael Snyder[3], Ross Hardison[7], Bing Ren[12]

University of Washington-Fred Hutchinson Cancer Research Center Group:
John A. Stamatoyannopoulos[1], Mark Groudine[2], Michael Bender[2], Rajinder Kaul[1], Theresa Canfield[1], Erica Giste[1], Audra Johnson[1], Mia Zhang[2], Gayathri Balasundaram[2], Rachel Byron[2], Vaughan Roach[1], Peter Sabo[1], Richard Sandstrom[1], A. Sandra Stehling[1], Bob Thurman[1]

Stanford-Yale Group:
Michael Snyder[3], Sherman M. Weissman[4], Philip Cayting[4,5,6] , Manoj Hariharan[3], Jin Lian[5], Yong Cheng[3], Stephen G. Landt[3], Zhihai Ma[3]

Penn State / University of Massachusetts / Duke University / Emory University / California Institute of Technology / University of California, Irvine/Children's Hospital of Philadelphia Group:
Ross Hardison[7], Barbara J. Wold[16], Job Dekker[8], Gregory Crawford[9,10], Cheryl A. Keller[7], Weisheng Wu[7], Christopher Morrissey[7], Swathi A. Kumar[7], Tejaswini Mishra[7], Deepti Jain[7], Marta Byrska-Bishop[7], Daniel Blankenberg[7], Bryan R. Lajoie[8], Gaurav Jain[8], Amartya Sanyal[8], Kaun-Bei Chen[9], Olgert Denas[9], James Taylor[11] Gerd A. Blobel[15], Mitchell J. Weiss[15], Max Pimkin[15], Wulan Deng[15], Georgi K. Marinov[16], Brian A. Williams[16], Katherine I. Fisher-Aylor[16], Gilberto Desalvo[16], Anthony Kiralusha[16], Diane Trout[16], Henry Amrhein[16], Ali Mortazavi[17]

University of California San Diego Group: Bing Ren[12], Lee Edsall[12], David McCleary[12], Samantha Kuan[12], Yin Shen[12], Feng Yue[12], Zhen Ye[12]

Cold Spring Harbor Laboratories / CRG Group:
Thomas R Gingeras[18], Carrie A. Davis[18], Chris Zaleski[18], Sonali Jha[18], Chenghai Xue[18], Alex Dobin[18], Wei Lin[18], Meagan Fastuca[18], Huaien Wang[18], Roderic Guigo[19], Sarah Djebali[19], Julien Lagarde[19]

Data Coordination Center at University of California Santa Cruz: Venkat S. Malladi[13], Melissa S. Cline[13], Vanessa M. Kirkup[13], Katrina Learned[13], Kate R. Rosenbloom[13] and W. James Kent[13]

NHGRI Project Management Group: Elise A. Feingold[14], Peter J. Good[14], Michael Pazin[14], Rebecca F. Lowdon[14], Leslie B. Adams[14]

Author Affiliations

[1] Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America

[2] Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America,

[3] Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America,

[4] Department of Genetics, Yale University, New Haven, Connecticut, United States of America,

[5] Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America,

[6] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America,

[7] Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America,

[8] Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachussetts, United States of America,

[9] Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America,

[10] Department of Pediatrics, Duke University, Durham, North Carolina, United States of America,

[11] Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America

[12] Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California San Diego, La Jolla, California, United States of America

[13] Center for Biomolecular Science and Engineering, School of Engineering , University of California Santa Cruz (UCSC), Santa Cruz, California, United States of America

[14] National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America

[15] Div. of Hematology, Children's Hospital of Philadelphia, Abramson Research Center, Philadelphia, Pennsylvania, United States of America

[16] Div. of Biology, California Institute of Technology, Pasadena, California, United States of America

[17] Dept. of Developmental and Cell Biology, University of California Irvine, Irvine California, United States of America

[18] Dept. of Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America

[19] Division of Bioinformatics and Genomics, Center for Genomic Regulation, Barcelona, Catalunya, Spain

Figure 1 - Overview of MouseENCODE Data