

Unlocking the secrets of the genome

Despite the successes of genomics, little is known about how genetic information produces complex organisms. A look at the crucial functional elements of fly and worm genomes could change that.

Susan E. Celniker, Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, Manolis Kellis, Eric C. Lai, Jason D. Lieb, David M. MacAlpine, Gos Micklem, Fabio Piano, Michael Snyder, Lincoln Stein, Kevin P. White and Robert H. Waterston, for the modENCODE Consortium

The primary objective of the Human Genome Project was to produce high-quality sequences not just for the human genome but also for those of the chief model organisms: *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*) and mouse (*Mus musculus*). Free access to the resultant data has prompted much biological research, including development of a map of common human genetic variants (the International HapMap Project)¹, expression profiling of healthy and diseased cells² and in-depth studies of many individual genes. These genome sequences have enabled researchers to carry out genetic and functional genomic studies not previously possible, revealing new biological insights with broad relevance across the animal kingdom^{3,4}.

Nevertheless, our understanding of how the information encoded in a genome can produce a complex multicellular organism remains far from complete. To interpret the genome accurately requires a complete list of functionally important elements and a description of their dynamic activities over time and across different cell types. As well as genes for proteins and non-coding RNAs, functionally important elements include regulatory sequences that direct essential functions such as gene expression, DNA replication and chromosome inheritance.

Although geneticists have been quick to decode the functional elements in the yeast *S. cerevisiae*, with its small compact genome and powerful experimental tools⁵⁻⁶, our understanding of the more complex genomes of human, mouse, fly and worm is still rudimentary. Intrinsic signals that define the boundaries of protein-coding genes can only be partly recognized by current algorithms, and signals for other functional elements are even harder to find and interpret. Experimental approaches, notably the sequencing of complementary

DNA and expressed sequence tags, have been invaluable, but unfortunately these data sets remain incomplete⁷. Non-coding RNA genes present an even greater challenge⁸⁻¹⁰, and many remain to be discovered, particularly those that have not been strongly conserved during evolution. Flies and worms have roughly the same number of known transcription factors as humans¹¹, but comprehensive molecular studies of gene regulatory networks have yet to be tackled in any of these species.

In an attempt to remedy this situation, the National Human Genome Research Institute (NHGRI) launched the ENCODE (Encyclopedia of DNA Elements) project in 2003, with the goal of defining the functional elements in the human genome. The pilot phase of the project focused on 1% of the human genome and a parallel effort to foster technology development¹². The initial ENCODE analysis revealed new findings but also made clear just how complex the biology is and how our grasp of it is far from complete¹³. On the basis of this experience, the NHGRI launched two complementary programmes in 2007: an expansion of the human ENCODE project to the whole genome (www.genome.gov/ENCODE) and the model organism ENCODE (modENCODE) project to generate a comprehensive annotation of the functional elements in the *C. elegans* and *D. melanogaster* genomes (www.modencode.org; www.genome.gov/modENCODE).

These two model organisms, with their ease of husbandry and genetic manipulation, are pillars of modern biological research, and a systematic catalogue of their functional genomic elements promises to pave the way to a more complete understanding of the human genome. Studies of these animals have provided key insights into many basic metazoan processes, including developmental patterning, cellular signalling, DNA replication and inheritance, programmed cell death and RNA interference (RNAi). The genomes are small enough to be investigated comprehensively with current technologies and findings can be validated *in vivo*. The research communities that study these two organisms will rapidly make use of the modENCODE results, deploying powerful experimental approaches that are often not possible or practical in mammals, including genetic, genomic, transgenic, biochemical and RNAi assays. modENCODE, with its potential for biological validation, will add value to the human ENCODE effort by illuminating the relationship between molecular and biological events.

The modENCODE project (Table 1) complements other systematic investigations into these highly studied organisms. In both organisms, RNAi collections have been developed and used to uncover novel gene functions¹⁴⁻¹⁸. Mutants are being recovered through insertional mutagenesis¹⁹ and targeted deletions (<http://celeganskoconsortium.omrf.org>;

TABLE 1 MODENCODE CONSORTIUM

Elements	Worm	Fly	Primary experimental data
Transcripts (mRNAs, non-coding RNAs, transcription start sites, untranslated regions, miRNAs)	Robert Waterston (University of Washington), Fabio Piano (New York University)	Susan Celniker (LBNL), Eric Lai (Sloan-Kettering Institute)	Tiling arrays, RNASeq, RT-PCR/RACE, mass spectrometry, 3' untranslated region clone library, UAS-miRNA flies, knockdowns of RNA-binding proteins
Transcription-factor-binding sites	Michael Snyder (Yale University)	Kevin White (University of Chicago)	ChIP-chip, ChIP-seq, transcription-factor-tagged strains, anti-transcription factor antibodies
Chromatin marks	Jason Lieb (University of North Carolina), Steven Henikoff (University of Washington)	Gary Karpen (LBNL), Steven Henikoff	ChIP-chip and ChIP-seq of chromosome-associated proteins and nucleosomes
DNA replication		David MacAlpine (Duke University Medical Center)	ChIP-chip and ChIP-seq of essential initiator proteins, origin mapping and DNA copy number in differentiated tissues

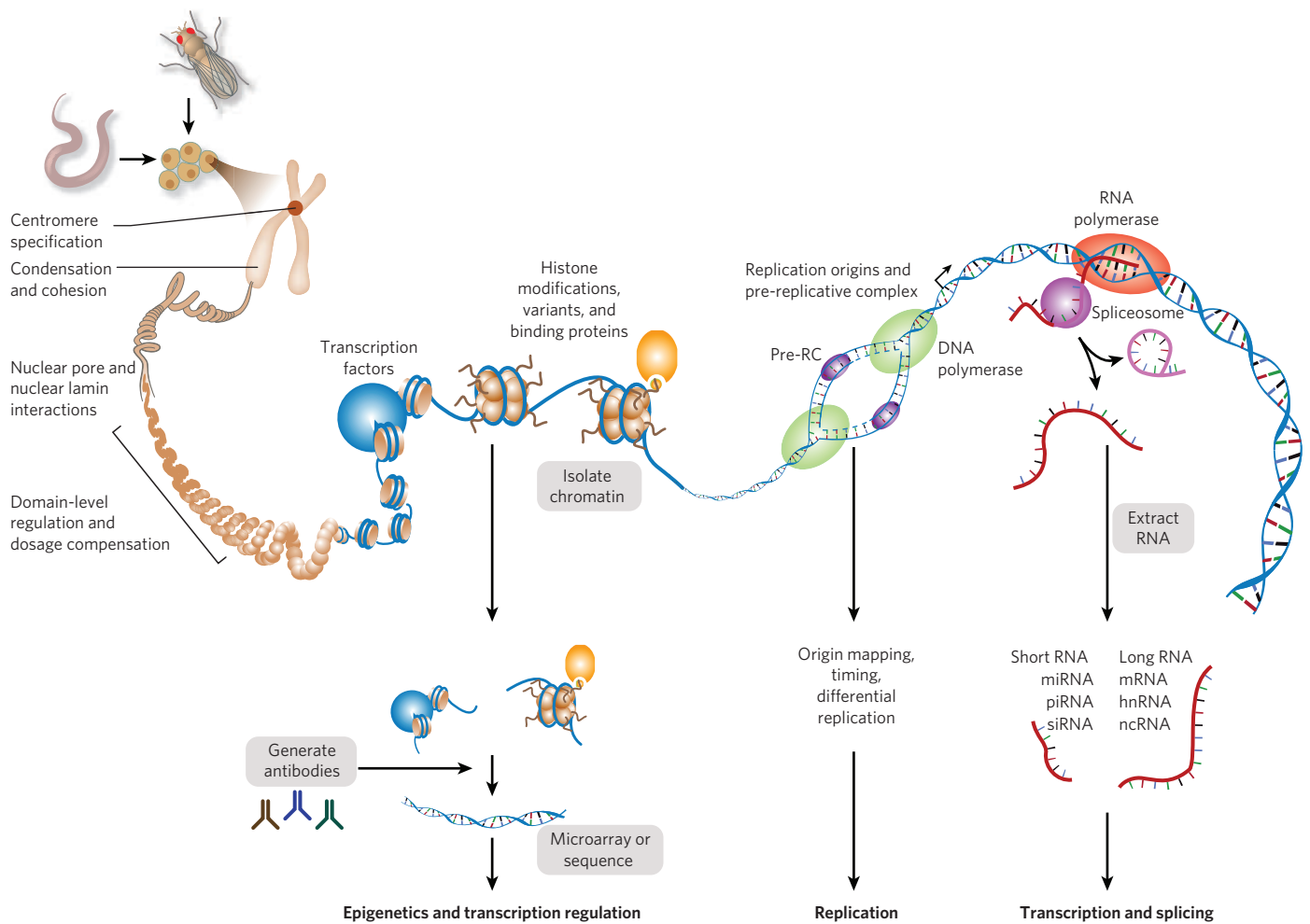


Figure 1 | DNA element functions and identification process.

www.shigen.nig.ac.jp/c.elegans), with the eventual goal of one for every known gene. Genome sequences of related species are now also available for both fly^{20,21} and worm²², and multiple independent wild isolates are being characterized (T. MacKay, personal communication, www.dpgp.org²³; R.H.W.). First-generation catalogues have been assembled of gene expression patterns during development and in different tissues^{24–34}.

Research and analysis

The modENCODE project will operate as an open consortium and participants can join on the understanding that they will abide by the set criteria (www.genome.gov/26524644). An important aim of the project is to respond to the needs of the broader *Drosophila* and *C. elegans* scientific communities, and several avenues will be open for suggestions on which experiments to prioritize. For example, researchers can visit www.modencode.org/Vote.shtml now to help prioritize transcription factors for studies using chromatin immunoprecipitation followed by DNA microarray or DNA sequencing (ChIP-chip and ChIP-seq), and can also indicate whether they have useful antibodies. We will seek community input on

other issues as the opportunities arise.

The core of the modENCODE project consists of ten groups who use high-throughput methods to identify functional elements (see Table 1). A Data Coordinating Center (DCC) will collect, integrate and display the data. Together, the groups expect to identify the principal classes of functional element for *D. melanogaster* and *C. elegans*. They will work closely together to complete the precise annotation of protein-coding genes, identify small RNAs and non-coding RNA transcripts, map transcription start sites, identify promoter motif elements, elucidate functional elements within 3' untranslated regions, and identify alternatively spliced transcripts as well as the signals required for splicing. Genomic sites bound by sequence-specific transcription factors will also be comprehensively identified. Charting the chromatin 'landscapes' will include the characterization of key histone modifications and variants, nucleosome phasing, RNA polymerase II isoforms and proteins involved in dosage compensation, centromere function, replication, homologue pairing, recombination and associations of chromosomes with the nuclear envelope.

Integrative analysis of these data across

the different types of functional element will be used to reveal fundamental principles of fly and worm genome biology and to begin to uncover the emergent properties of these complex genomes. Some topics the modENCODE groups, along with interested members of the wider community, intend to explore are outlined below, but these are only a beginning. Our intention is to create a resource that will provide the foundation for ongoing analysis by scientists for years to come.

Our two model organisms share many similarities with other metazoans, including humans. They also differ from other organisms in some striking ways, particularly in details of the establishment and maintenance of cellular identity, centromere biology and heterochromatin function. To help understand how the similarities and differences in worm and fly biology are reflected in their genome sequences and how they are specified by genome function at the molecular level, we will carry out comparative analyses of transcription, splicing, *cis*-regulatory and post-transcriptional elements and chromatin function. We will subsequently investigate how our findings apply to the control of gene expression in the human genome.

We also plan to use genome-wide data on pre- and post-transcriptional functional elements to expand our understanding of gene-regulatory networks. We will study how these two layers of control complement or reinforce each other during development. For example, the availability of full-length transcripts and promoter structures for microRNA (miRNA) genes will enable us to develop models of regulatory circuits that integrate the upstream regulation of miRNA genes with that of other regulatory factors (such as transcription factors) and the effects of miRNAs on their downstream targets. We will search global patterns identified in the regulatory programs for emerging principles of gene regulation within and across species; as part of this endeavour, we will evaluate evidence for the modular structure of regulatory networks.

Because several developmental stages and diverse tissues will be sampled in both animals, we will be able to investigate the global and dynamic activities of functional elements across the entire genome in multiple cell types and stages of differentiation. We aim to define the characteristics and rules that distinguish regulatory programs in different cell types and developmental stages at the DNA, chromatin, and post-transcriptional levels. This will enable us to identify the types of element that function together in various spatio-temporal environments and find new types of functional element, perhaps including those used in restricted developmental contexts.

An important objective is to generate specific biological hypotheses that can be refined and tested experimentally by the broader scientific community. For example, these analyses might identify transcribed regions with novel regulatory roles, structural regions that function in the establishment of chromatin structure or three-dimensional conformation, enhancers far away from the gene they control, and alternative promoter regions. In addition, we will use comparative analyses of the sequenced genomes from different species to clarify the extent of conservation and the functional constraints associated with potential new classes of element and to characterize their evolutionary signatures²¹.

Another objective of the modENCODE project is the creation of reference data sets of maximum utility. We have agreed that, whenever possible, a common set of reagents will be used to facilitate comparison of data sets generated by different groups. For example, the fly and worm groups using ChIP-chip and related methods to map the genome-wide distributions of histone modifications will use a common set of validated antibodies. In addition, we will use common fly and worm strains, and in the case of *Drosophila*, the common cell lines Kc167, S2-DRSC, CME W1 CL.8+ and ML-DmBG3-c2.

The fly and worm genomes are about a thirtieth of the size of their mammalian counterparts, making current methods for high-

throughput genomic analysis cost-effective. We will use high-density tiling DNA microarrays to interrogate the genome on a single microarray (*C. elegans*, 26 base pair (bp) median spacing; *D. melanogaster*, 38 bp median spacing) at a resolution sufficient for ChIP-chip experiments. Denser arrays (*D. melanogaster*, 7 bp median spacing), which promise higher resolution, will be used in a move to high-throughput sequencing platforms such as the Illumina Genome Analyzer to generate sufficient sequence coverage for transcript mapping and miRNA and ChIP experiments.

The biological significance of the genomic features identified will be tested in experiments designed to evaluate the accuracy and functionality of subsets of the structural and regulatory annotations. For example, we will carry out ChIP experiments on extracts from whole animals or cells that lack selected regulators (using mutants or RNAi). The tissue-specific DNA-binding patterns of selected regulators will be validated in transgenic animals. Figure 1 summarizes the DNA elements to be interrogated and the methods to be used.

Data management and accessibility

Data generated by the modENCODE Consortium, including those from validation experiments, will be collected, quality checked, integrated and distributed through the modENCODE DCC (www.modencode.org). The DCC will collate detailed metadata for each submitted data set to ensure broad and long-term usability. Where appropriate, the data will also be submitted to public databases, for example, GenBank (www.ncbi.nlm.nih.gov) and the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) or Array Express (www.ebi.ac.uk/microarray-as/aer/entry) and the University of California, Santa Cruz Genome Bioinformatics Site (<http://genome.ucsc.edu>). The DCC will also work

closely with WormBase (www.wormbase.org) and FlyBase (www.flybase.org) to facilitate integration of the modENCODE data with selected data from these databases and with other information about these organisms.

All data will be available for bulk download through an FTP site and through a number of Generic Model Organism Database tools (www.gmod.org): BioMart (www.biomart.org) will provide powerful data-mining capabilities, and InterMine (www.intermine.org) will provide a flexible interface for complex querying of the data, a library of canned queries, and powerful list-based tools and operations (<http://intermine.modencode.org>). As for the ENCODE pilot project data (www.genome.gov/10005107), new data can be examined alongside existing data using interactive genome browsers³⁵ for both the fly (www.modencode.org/cgi-bin/gbrowse/fly) and the worm (www.modencode.org/cgi-bin/gbrowse/worm).

The *Drosophila* and *C. elegans* communities have thrived because of their open culture. In keeping with this tradition and with those of the genome sequencing projects, HapMap and the ENCODE pilot project, modENCODE is a 'community resource project' subject to the NHGRI's data-sharing policy. The success of this policy is based on mutual and independent responsibilities for the production and use of the resource. We will release data rapidly (Table 2), before publication, once they have been established to be reproducible (verification; see www.modencode.org/PublicationPolicy link' for the criteria), even if the data have not been sampled to determine if there is biological meaning (validation). In turn, users are asked to recognize the source of the data and to respect the legitimate interest of the resource producers to publish an initial report of their work (see www.genome.gov/modencode for more details). Finally, the funding agencies

TABLE 2 GLOBAL ANALYSIS GOALS

Elements and processes	Specific examples
Transcribed regions	Define cell- and tissue-specific transcriptional landscapes. Annotate transcription start sites, exons, untranslated region structures, small regulatory RNAs and short single-exon open reading frames
Gene regulation, transcriptional regulation	Identify transcription-factor binding sites in various cell and tissue types. Correlate chromatin structure marks and transcriptional activities for protein-coding and non-protein-coding genes
Post-transcriptional regulation	Identify tissue-specific binding sites for miRNAs and other small RNAs, RNA secondary structures and alternative splicing regulatory motifs
Chromatin structure and function	Identify sites of association between DNA and chromosomal proteins involved in centromere specification, meiotic recombination, dosage compensation, nuclear envelope and matrix interactions and chromosome condensation. Identify sites of incorporation of histone variants and specifically modified histones. Correlate transcription maps for meta-analysis of developmental chromatin dynamics.
DNA replication	Identify cell- and tissue-specific origins of replication. Correlate with cell- and tissue-specific transcription and chromatin marks

recognize the need to support the analysis and dissemination of the data.

In addition, a variety of physical resources (for example, DNA constructs and transgenic strains) will be produced that are likely to be of use to the broader community and to which that community will have unrestricted access. We expect to cooperate with data users in the worm and fly communities to set the gold standard for data release and openness.

Conclusion

The Human Genome Project benefited enormously from the technology developed and the experience acquired in sequencing the significantly smaller genomes of model organisms, particularly *C. elegans* and *D. melanogaster*. The modENCODE project is dedicated to the next phase of decoding the information stored in these genomes: the comprehensive identification of sequence-based functional elements. Having laid the foundation for the discovery of many of the genetic programs underlying metazoan development and behaviour, *Drosophila* and *Caenorhabditis* will serve as ideal model systems to identify DNA-based functional ele-

ments on a genome-wide basis. In the future, these data will provide a powerful platform for characterizing the functional networks that direct multicellular biology, thereby linking genomic data with the biological programs of higher organisms, including humans. ■

1. Sabeti, P. C. *et al. Nature* **449**, 913–918 (2007).
2. Neve, R. M. *et al. Cancer Cell* **10**, 515–527 (2006).
3. Chintapalli, V. R., Wang, J. & Dow, J. A. *Nature Genet.* **39**, 715–720 (2007).
4. Nichols, C. D. *Pharmacol. Ther.* **112**, 677–700 (2006).
5. Ross-Macdonald, P. *et al. Nature* **402**, 413–418 (1999).
6. Boone, C., Bussey, H. & Andrews, B. J. *Nature Rev. Genet.* **8**, 437–449 (2007).
7. Celniker, S. E. & Rubin, G. M. *Annu. Rev. Genomics Hum. Genet.* **4**, 89–117 (2003).
8. Tupy, J. L. *et al. Proc. Natl Acad. Sci. USA* **102**, 5495–5500 (2005).
9. Ruby, J. G. *et al. Cell* **127**, 1193–1207 (2006).
10. Ruby, J. G. *et al. Genome Res.* **17**, 1850–1864 (2007).
11. Reece-Hoyes, J. S. *et al. Genome Biol.* **6**, R110 (2005).
12. The ENCODE Project Consortium *Science* **306**, 636–640 (2004).
13. Birney, E. *et al. Nature* **447**, 799–816 (2007).
14. Boutros, M. *et al. Science* **303**, 832–835 (2004).
15. Kamath, R. S. *et al. Nature* **421**, 231–237 (2003).
16. Rual, J. F. *et al. Genome Res.* **14**, 2162–2168 (2004).
17. Sonnichsen, B. *et al. Nature* **434**, 462–469 (2005).
18. Dietzl, G. *et al. Nature* **448**, 151–156 (2007).
19. Bellen, H. J. *et al. Genetics* **167**, 761–781 (2004).
20. Clark, A. G. *et al. Nature* **450**, 203–218 (2007).

21. Stark, A. *et al. Nature* **450**, 219–232 (2007).
22. Stein, L. D. *et al. PLoS Biol.* **1**, E45 (2003).
23. Hillier, L. W. *et al. Nature Methods* **5**, 183–188 (2008).
24. Tomancak, P. *et al. Genome Biol.* **3**, research0088.1–0088.14 (2002).
25. Arbeitman, M. N. *et al. Science* **297**, 2270–2275 (2002).
26. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. *Science* **302**, 249–255 (2003).
27. Li, T. R. & White, K. P. *Dev. Cell* **5**, 59–72 (2003).
28. Stolic, V. *et al. Science* **306**, 655–660 (2004).
29. Manak, J. R. *et al. Nature Genet.* **38**, 1151–1158 (2006).
30. Tomancak, P. *et al. Genome Biol.* **8**, R145 (2007).
31. Jiang, M. *et al. Proc. Natl Acad. Sci. USA* **98**, 218–223 (2001).
32. Reinke, V., Gil, I. S., Ward, S. & Kazmer, K. *Development* **131**, 311–323 (2004).
33. Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L. & Hunter, C. P. *Development* **130**, 889–900 (2003).
34. Kim, S. K. *et al. Science* **293**, 2087–2092 (2001).
35. Stein, L. D. *et al. Genome Res.* **12**, 1599–1610 (2002).

Supplementary Information A full list of names and addresses of current consortium participants is linked to the online version of this feature at <http://tinyurl.com/modENCODE>

Acknowledgements We thank Brenda Andrews and Tim Hughes for discussions on the status of yeast functional genomics.

Author Information Correspondence should be addressed to S.E.C. (celniker@fruitfly.org).

Authors

Susan E. Celniker¹, Laura A. L. Dillon², Mark B. Gerstein^{3,4}, Kristin C. Gunsalus⁵, Steven Henikoff⁶, Gary H. Karpen⁷, Manolis Kellis^{8,9}, Eric C. Lai¹⁰, Jason D. Lieb¹¹, David M. MacAlpine¹², Gos Micklem¹³, Fabio Piano⁵, Michael Snyder¹⁴, Lincoln Stein¹⁵, Kevin P. White^{16,17}, Robert H. Waterston¹⁸

¹Department of Genome Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ²Division of Extramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. ³Program in Computational Biology and Bioinformatics, ⁴Department of Computer Science and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. ⁵Center for Genomics and Systems Biology, New York University, New York, New York 10003, USA. ⁶Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. ⁷Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. ⁸Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA. ⁹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ¹⁰Sloan-Kettering Institute, New York, New York 10065, USA. ¹¹Department of Biology and Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ¹²Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA. ¹³Department of Genetics, University of Cambridge, CB2 3EH, UK, and Cambridge Systems Biology Centre, Tennis Court Road, Cambridge CB2 1QR, UK. ¹⁴Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06824, USA. ¹⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11542 USA. ¹⁶Institute for Genomics & Systems Biology, University of Chicago, Chicago, Illinois 60637, USA. ¹⁷Institute for Genomics & Systems Biology, Argonne National Laboratory, Argonne, Illinois 60439, USA. ¹⁸Department of Genome Sciences and University of Washington School of Medicine, Seattle, Washington 98195, USA.

Unlocking the secrets of the genome

Supplementary information to article doi:459927a

Full list of modENCODE consortium participants and affiliations.

Caenorhabditis elegans

Transcripts (New York University) Marco Mangone⁵, Ting Han¹⁹, Marlon Stoeckius²⁰, Philip MacMenamin⁵, Arun Prasad Manoharan¹⁹, Jonas Maaskola²⁰, Emily Mis⁵, Dongping Wei¹⁹, Vishal Khivansara¹⁹, Kevin Chen⁵, Oliver Attie⁵, Wei Chen²¹, Nikolaus Rajewsky²⁰, Kristin C. Gunsalus⁵, John Kim¹⁹, Fabio Piano⁵, (University of Washington, Seattle) LaDeana W. Hillier¹⁸, Robert H. Waterston (Principal Investigator)¹⁸, Phil Green¹⁸, Brent Ewing¹⁸, David Gordon¹⁸, Colleen Davis¹⁸, Michael MacCoss¹⁸, Daniela Tomazela¹⁸, Gennifer Merrihew¹⁸, Barbara Frewen¹⁸, Jesse Canterbury¹⁸, Mark B. Gerstein^{3,4}, Rajkumar Sasidharan⁴, Ashish Agarwal⁴, Mike Wilson⁴, Guoneng Zhong⁴, Valerie Reinke²², Jeanyoung Jo²², David M. Miller III²³, Joseph D. Watson²³, William C. Spencer²³, Rebecca D. McWhirter²³, Stephen E. Von Stetina²³, Sarah C. Anthony²³, Frank J. Slack²⁴, Masaomi Kato²⁴; **Transcription-factor-binding sites** (Yale University) Michael Snyder (Principal Investigator)¹⁴, Debasish Raha¹⁴, Mei Zhong¹⁴, Wei Nui¹⁴, Valerie Reinke²², Judith Janette²², Jeanyoung Jo²², Mark B. Gerstein^{3,4}, Rajkumar Sasidharan⁴, Ashish Agarwal⁴, Mike Wilson⁴, Guoneng Zhong⁴, Stuart K. Kim²⁵, Cindie Slightam²⁵, Min Jiang²⁵, Xiao Xu²⁵, Mihail Sarov²⁶, A. Francis Stewart²⁷, Tony Hyman²⁶, Robert H. Waterston¹⁸, John I. Murray¹⁸, Elicia A. Preston¹⁸, Dionne Vafeados¹⁸; **Chromatin marks** (Fred Hutchinson Cancer Research Center) Steven Henikoff (Principal Investigator)⁶, Kami Ahmad²⁸, Siew-Loon Ooi⁶, Akiko Sakai²⁸, Jorja G. Henikoff⁶, (The University of North Carolina at Chapel Hill) Jason D. Lieb (Principal Investigator)¹¹, Christina M. Whittle¹¹, Sevinc Ercan¹¹, Kohta Ikegami¹¹, Morten B. Jensen¹¹, Saianand Balu¹¹, Everett Zhou¹¹, Jennifer Brennan¹¹, Lindsay Dick¹¹, Susan Strome²⁹, Taryn Phippen²⁹, Teruaki Takasaki²⁹, Thea Egelhofer²⁹, A. Leo Iniguez³⁰, Heather Holster³⁰, Heidi Rosenbaum³⁰, X. Shirley Liu³¹, Tao Liu³¹, Hyunjin Shin³¹, Yong Zhang³¹, Arjun K. Manrai³¹, Zhenhua Wu³¹, Eran Segal³², Yaniv Lubling³², Noam Kaplan³², Yair Field³², Abby Dernburg³³, Hoang Pham³³, Arshad Desai³⁴, Reto Gassman³⁴, Karen Yuen³⁴, Julie Ahinger³⁵, Anne Canonge³⁵, Paulina Kolasinska-Zwierz³⁵, Isabel Latorre³⁵

Drosophila melanogaster

Transcripts (Lawrence Berkeley National Laboratory) Angela N. Brooks³⁶, Kasper D. Hansen³⁷, Sandrine Dudoit^{37,38}, Steven E. Brenner^{36,39}, Roger Hoskins¹, Ann S. Hammonds¹, Joseph W. Carlson¹, Jane Landolin¹, Kenneth H. Wan¹, Charles Yu¹, Benjamin Booth¹, Susan E. Celniker (Principal Investigator)¹, Peter Cherbas^{40,41}, Justen Andrews⁴⁰, Lucy Cherbas^{40,41}, Dayu Zhang⁴⁰, David Miller⁴⁰, Andreas Rechsteiner⁴¹, Thomas C. Kaufman⁴⁰, Justin P. Kumar⁴⁰, Laura Langton⁴², Marijke J. van Baren⁴², Aaron E. Tenney⁴², Charles L. G. Comstock⁴², Michael Brent⁴², Aaron T. Willingham⁴³, Philipp Kapranov⁴³, Srinika Ghosh⁴³, Alex Dobin⁴⁴, Christopher P. Zalski⁴⁴, Wei Lin⁴⁴, Carrie A. Davis⁴⁴, Thomas R. Gingeras^{43,44}, Michael O. Duff⁴⁵, Li Yang⁴⁵, Brenton R. Graveley⁴⁵, Norbert Perrimon^{46,47}, Stephanie Mohr⁴⁶, Bernard Mathey-Prevot⁴⁶, (Memorial Sloan-Kettering Cancer Center) Eric C. Lai¹⁰, Wei-Jen Chung¹⁰, Raquel Martin¹⁰, Gregory J. Hannon⁴⁸, Michelle Rooks⁴⁸, Eugene Berezikov⁴⁹, Martin Hirst⁵⁰, Marco Marra⁵⁰, Yongjun Zhao⁵⁰, Thomas Zeng⁵⁰; **Transcription-factor-binding sites** (University of Chicago) Richard P. Auburn⁵¹, Hugo J. Bellen⁵², Chris Bristow⁹, Christopher D. Brown¹⁶, Jia Chen⁵³, Keith Ching⁵⁴, Marc H. Domanus¹⁷, Lee Edsall⁵⁴, Robert Grossman⁵³, Jason Ernst^{8,9}, David Hanley⁵³, Elizabeth Heinz¹⁷, Roger Hoskins¹, Haruhiko Ishii⁵⁵, Manolis Kellis^{8,9}, Pouya Kheradpour⁹, Zirong Li⁵⁵, Michael F. Lin^{8,9}, Xiangjun Liu^{53,56}, Folker Meyer^{57,58}, Steven W. Miller⁵⁹, Carolyn A. Morrison¹⁶, Nicolas Negre¹⁶, James W. Posakony⁵⁹, Bing Ren⁶⁰, Steven Russell⁵¹, Douglas A. Scheftner¹⁶, Rachel Sealon⁹, Lionel Senderowicz¹⁶, Parantu K. Shah¹⁶, Rebecca F. Spokony¹⁶, Alexander Stark^{8,9}, Feng Tian^{53,56}, Koen J.T. Venken⁶¹, Ulrich Wagner⁵⁴, Kevin P. White (Principal Investigator)^{16,17}, Robert White⁶², Jared Wilkening⁵⁸, Wenjun Wu⁵⁸, Zhen Ye⁵⁴, Sui Zhang⁵⁹, Yizhong Zhang¹⁷; **Chromatin marks** (Fred Hutchinson Cancer Research Center) Steven Henikoff (Principal Investigator)⁶, Kami Ahmad²⁸, Siew-Loon Ooi⁶, Akiko Sakai²⁸, Jorja G. Henikoff⁶ (Lawrence Berkeley National Laboratory) Gary H. Karpen⁷, Sarah C. R. Elgin⁶³, Mitzi I. Kuroda⁶⁴, Peter J. Park⁶⁵⁻⁶⁷, Vincenzo Pirrotta⁶⁸, Artyom A. Alekseyenko⁶⁴, Andrey A. Gorchakov⁶⁴, Cameron Kennedy⁷, Peter V. Kharchenko^{65,66}, Ok-Kyung Lee⁶⁴, Sarah E. Marchetti⁶³, Aki Minoda⁷, Shouyong Peng⁶⁶, Nicole C Riddle⁶³, Yuri B. Schwartz⁶⁸, Gregory Shanower⁶⁸, Michael Y. Tolstorukov⁶⁶; **DNA replication** (Duke University) David M. MacAlpine (Principal Investigator)¹², Leyna C. DeNapoli¹², Matthew L. Eaton¹², Heather K. MacAlpine¹², Queying Ding¹², Thomas Eng⁶⁹, Helena Kashevsky⁶⁹, Noa Sher⁶⁹, Terry L. Orr-Weaver⁶⁹

Data Coordination Center (Cold Spring Harbor Laboratory) Galt Barber⁷⁰, Sergio Contrino¹³, Francois Guillier¹³, Angie Hinrichs⁷⁰, W. James Kent⁷⁰, Suzanna Lewis⁷¹, Sheldon McKay¹⁵, Gos Micklem¹³, Christopher J. Mungall⁷¹, Kate Rosenbloom⁷⁰, Kim Rutherford¹³, Richard Smith¹³, Lincoln Stein (Principal Investigator)¹⁵, E. O. Stinson⁷¹, Nicole L. Washington⁷¹, Zheng Zha¹⁵

modENCODE project scientific management Francis S. Collins⁷², Laura A. L. Dillon², Elise A. Feingold², Peter J. Good², Mark S. Guyer²

Affiliations

¹Department of Genome Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ²Division of Extramural Research, National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Suite 4076, Bethesda, Maryland 20892, USA. ³Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA; Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. ⁴Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. ⁵Center for Genomics and Systems Biology, Department of Biology, New York University, 1009 Silver Center, 100 Washington Square East, New York, New York 10003, USA. ⁶Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. ⁷Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Department of Molecular and Cell Biology, University of California, Berkeley, One Cyclotron Road, Berkeley,

California 94720, USA. ⁸Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA. ⁹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ¹⁰Sloan-Kettering Institute, 1275 York Avenue, Box 252, New York, New York 10065, USA. ¹¹Department of Biology and Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ¹²Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA. ¹³Department of Genetics, University of Cambridge, CB2 3EH, UK; Cambridge Systems Biology Centre, Tennis Court Road, Cambridge CB2 1QR, UK. ¹⁴Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06824, USA. ¹⁵Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ¹⁶Institute for Genomics & Systems Biology, Department of Human Genetics, University of Chicago, 920 E. 58th Street, Chicago, Illinois 60637, USA. ¹⁷Institute for Genomics & Systems Biology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, USA. ¹⁸Department of Genome Sciences and University of Washington School of Medicine, William H. Foege Bldg., 1705 N.E. Pacific Street, Seattle, Washington 98195-5065, USA. ¹⁹Life Sciences Institute, Department of Human Genetics, University of Michigan, 210 Washtenaw Avenue, Ann Arbor, Michigan 48109, USA. ²⁰Max-Delbrück-Centrum für Molekulare Medizin (MDC), Systems Biology, Berlin-Buch, Robert-Rössle-Str. 10, 13092 Berlin, Germany. ²¹Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany. ²²Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA. ²³Department of Cell and Developmental Biology, Vanderbilt University, 465 21st Avenue South, Nashville, Tennessee 37232, USA. ²⁴Department of Molecular, Cellular and Developmental Biology, PO Box 208103, Yale University, New Haven, Connecticut 06520, USA. ²⁵Stanford University Medical Center, Department of Developmental Biology, 279 Campus Drive, Stanford, California 94305, USA. ²⁶Max Planck Institute of Molecular Cell Biology and Genetics, Proteinhauerstrasse 108, 01307 Dresden, Germany. ²⁷Technical University Dresden, BiolInnovations Centre, Genomics Department, Tatzberg 47, 01307 Dresden, Germany. ²⁸Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. ²⁹Molecular, Cell, and Developmental Biology, Sinheimer Labs, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. ³⁰Roche NimbleGen, Inc., 500 South Rosa Road, Madison, Wisconsin 53719, USA. ³¹Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, Massachusetts 02115, USA. ³²Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel. ³³Life Sciences Division, Lawrence Berkeley National Laboratory and Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA. ³⁴Ludwig Institute for Cancer Research & Department of Cellular and Molecular Medicine, University of California, San Diego, 9500 Gilman Drive, MC 0653, La Jolla, California 92093, USA. ³⁵Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. ³⁶Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. ³⁷Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA. ³⁸Department of Statistics, University of California, Berkeley, California 94720, USA. ³⁹Department of Plant & Microbial Biology, University of California, Berkeley, California 94720, USA. ⁴⁰Department of Biology, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA. ⁴¹Center for Genomics and Bioinformatics, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA. ⁴²Center for Genome Sciences, Washington University, 4444 Forest Park Boulevard, Saint Louis, Missouri 63108, USA. ⁴³Affymetrix Inc, Santa Clara, California 95051, USA. ⁴⁴Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ⁴⁵Department of Genetics and Developmental Biology, University of Connecticut Health Center, 263 Farmington, Connecticut 06030-3301, USA. ⁴⁶Department of Genetics and Drosophila RNAi Screening Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ⁴⁷Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ⁴⁸Cold Spring Harbor Laboratory, Watson School of Biological Sciences and Howard Hughes Medical Institute, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ⁴⁹Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. ⁵⁰BC Cancer Agency Genome Sciences Centre, 570 West 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada. ⁵¹Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK. ⁵²Department of Molecular and Human Genetics, Howard Hughes Medical Institute, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁵³National Center for Data Mining, University of Illinois at Chicago, 700 Science and Engineering Offices, MC 249, 851 South Morgan Street, Chicago, Illinois 60607, USA. ⁵⁴Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁵⁵Ludwig Institute for Cancer Research, University of California San Diego Biology Division, 59500 Gilman Drive, La Jolla, California 92093-0653, USA. ⁵⁶Institute of Biomedical Informatics, School of Medical Sciences, Tsinghua University, Beijing, China 100084. ⁵⁷Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, MCS/221, Argonne, Illinois 60439, USA. ⁵⁸University of Chicago, Computation Institute, Research Institute Suite 405, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA. ⁵⁹Division of Biological Sciences, Section of Cell & Developmental Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁶⁰Ludwig Institute for Cancer Research, University of California San Diego School of Medicine, Department of Cellular and Molecular Medicine, University of California San Diego Moores Cancer Center, 59500 Gilman Drive, La Jolla, California 92093, USA. ⁶¹Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁶²Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK. ⁶³Department of Biology CB-1137, Washington University, Saint Louis, Missouri 63130, USA. ⁶⁴Harvard-Partners Center for Genetics & Genomics, Brigham & Women's Hospital, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ⁶⁵Children's Hospital Informatics Program, Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, 300 Longwood Ave, Boston, Massachusetts 02115, USA. ⁶⁶Harvard-Partners Center for Genetics and Genomics, New Research Building Rm 250, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ⁶⁷Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St Boston, Massachusetts 02115, USA. ⁶⁸Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854, USA. ⁶⁹Department of Biology, Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ⁷⁰Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA. ⁷¹Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mailstop 64-121, Berkeley, California 94720, USA. ⁷²Office of the Director, National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Suite 4B09, Bethesda, Maryland 20892, USA.