Perspectives on ENCODE

https://doi.org/10.1038/s41586-020-2449-8

Received: 7 December 2019

Accepted: 5 May 2020

Published online: 29 July 2020

Open access

Check for updates

The ENCODE Project Consortium*, Michael P. Snyder[1,2]✉, Thomas R. Gingeras[3], Jill E. Moore[4], Zhiping Weng[4,5,6], Mark B. Gerstein[7], Bing Ren[8,9], Ross C. Hardison[10], John A. Stamatoyannopoulos[11,12,13], Brenton R. Graveley[14], Elise A. Feingold[15], Michael J. Pazin[15], Michael Pagan[15], Daniel A. Gilchrist[15], Benjamin C. Hitz[1], J. Michael Cherry[1], Bradley E. Bernstein[16], Eric M. Mendenhall[17,18], Daniel R. Zerbino[19], Adam Frankish[19], Paul Flicek[19] & Richard M. Myers[18]

The Encylopedia of DNA Elements (ENCODE) Project launched in 2003 with the long-term goal of developing a comprehensive map of functional elements in the human genome. These included genes, biochemical regions associated with gene regulation (for example, transcription factor binding sites, open chromatin, and histone marks) and transcript isoforms. The marks serve as sites for candidate *cis*-regulatory elements (cCREs) that may serve functional roles in regulating gene expression[1]. The project has been extended to model organisms, particularly the mouse. In the third phase of ENCODE, nearly a million and more than 300,000 cCRE annotations have been generated for human and mouse, respectively, and these have provided a valuable resource for the scientific community.

The ENCODE Project was launched in 2003, as the first nearly complete human genome sequence was reported[2]. At that time, our understanding of the human genome was limited. For example, although 5% of the genome was known to be under purifying selection in placental mammals[3,4], our knowledge of specific elements, particularly with regards to non-protein coding genes and regulatory regions, was restricted to a few well-studied loci[2,5].

ENCODE commenced as an ambitious effort to comprehensively annotate the elements in the human genome, such as genes, control elements, and transcript isoforms, and was later expanded to annotate the genomes of several model organisms. Mapping assays identified biochemical activities and thus candidate regulatory elements.

Analyses of the human genome in ENCODE proceeded in successive phases (Extended Data Fig. 1). Phase I (2003–2007) interrogated a specified 1% of the human genome in order to evaluate emerging technologies[6]. Half of this 1% was in regions of high interest, and the other half was chosen to sample the range of genomic features (such as G+C content and genes). Microarray-based assays were used to map transcribed regions, open chromatin, and regions associated with transcription factors and histone modification in a wide variety of cell lines, and these assays began to reveal the basic organizational features of the human genome and transcriptome. Phase II (2007–2012) introduced sequencing-based technologies (for example, chromatin immunoprecipitation with sequencing (ChIP–seq) and RNA sequencing (RNA-seq)) that interrogated the whole human genome and transcriptome[7]. General assays such as transcript, open-chromatin and histone modification mapping were used on a wide variety of cell lines, while more specific assays, such as mapping transcription factor binding regions, were performed extensively on a smaller number of cell lines to provide detailed annotations on, and to investigate the relationships of, many regulatory proteins across the genome. Transcriptome analysis of subcellular compartments (the nucleus, cytosol and subnuclear compartments) of these cells enabled the locations of transcripts to be analysed[7].

ENCODE phase III

ENCODE 3 (2012–2017) expanded production and added new types of assays[8] (Fig. 1, Extended Data Fig. 1), which revealed landscapes of RNA binding and the 3D organization of chromatin via methods such as chromatin interaction analysis by paired-end tagging (ChIA-PET) and Hi-C chromosome conformation capture. Phases 2 and 3 delivered 9,239 experiments (7,495 in human and 1,744 in mouse) in more than 500 cell types and tissues, including mapping of transcribed regions and transcript isoforms, regions of transcripts recognized by RNA-binding proteins, transcription factor binding regions, and regions that harbour specific histone modifications, open chromatin, and 3D chromatin interactions. The results of all of these experiments are available at the ENCODE portal (http://www.encodeproject.org). These efforts, combined with those of related projects and many other laboratories, have produced a greatly enhanced view of the human genome (Fig. 2), identifying 20,225 protein-coding and 37,595 noncoding genes

[1]Department of Genetics, School of Medicine, Stanford University, Palo Alto, CA, USA. [2]Cardiovascular Institute, Stanford School of Medicine, Stanford, CA, USA. [3]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [4]University of Massachusetts Medical School, Program in Bioinformatics and Integrative Biology, Worcester, MA, USA. [5]Department of Thoracic Surgery, Clinical Translational Research Center, Shanghai Pulmonary Hospital, The School of Life Sciences and Technology, Tongji University, Shanghai, China. [6]Bioinformatics Program, Boston University, Boston, MA, USA. [7]Yale University, New Haven, CT, USA. [8]Ludwig Institute for Cancer Research, University of California, San Diego, La Jolla, CA, USA. [9]Center for Epigenomics, University of California, San Diego, La Jolla, CA, USA. [10]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA. [11]Altius Institute for Biomedical Sciences, Seattle, WA, USA. [12]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [13]Department of Medicine, University of Washington, Seattle, WA, USA. [14]Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Health, Farmington, CT, USA. [15]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [16]Broad Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [17]Biological Sciences, University of Alabama in Huntsville, Huntsville, AL, USA. [18]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [19]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. *An alphabetical list of authors and their affiliations appears online. A formatted list of authors appears in the Supplementary Information. ✉e-mail: mpsnyder@stanford.edu

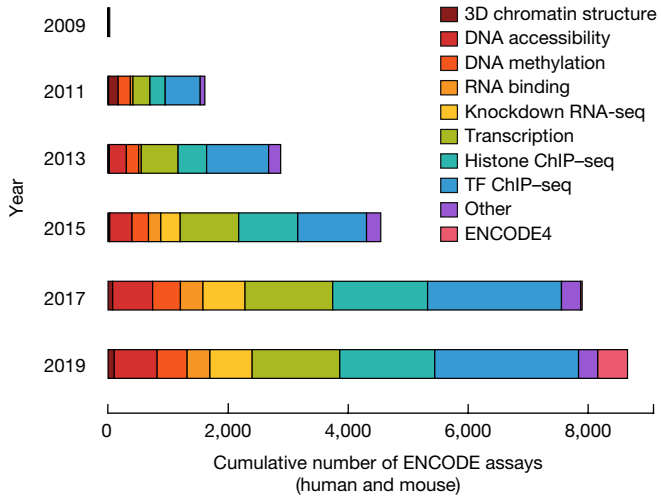


Fig. 1 | ENCODE assays by year. Accumulations of assays over the three phases of ENCODE. 3D chromatin structure includes ChIA-PET (62 experiments), Hi-C (31), and chromatin conformation capture carbon copy (5C, 13). Chromatin accessibility includes DNAase-seq (524), assay for transposase-accessible chromatin using sequencing (ATAC-seq, 129), transcription activator-like effector nuclease (TALEN)-modified DNAase-seq (40), formaldehyde-assisted isolation of regulator elements with sequencing (FAIRE-seq, 37) and micrococcal nuclease digestion with deep sequencing (MNase-seq, 2). DNA methylation includes DNAme arrays (259), WGBS (124), reduced-representation bisulfite sequencing (RRBS, 103), methylation-sensitive restriction enzyme sequencing (MRE-seq, 24) and methylated DNA immunoprecipitation coupled with next-generation sequencing (MeDIP-seq, 4). Histone modification includes ChIP–seq (1,605) on histone and modified histone targets. Knockdown transcription includes RNA-seq preceded by small interfering RNA (siRNA, 54), short hairpin RNA (shRNA, 531), clustered regularly interspaced short palindromic repeats (CRISPR, 50) or CRISPR interference (CRISPRi, 77). RNA binding includes enhanced cross-linking immunoprecipitation (eCLIP, 349), RNA bind-n-seq (158), RNA immunoprecipitation sequencing (RIP-seq, 158), RNA-binding protein immunoprecipitation-microarray profiling (RIP-chip, 32), individual nucleotide-resolution CLIP (iCLIP, 6) and Switchgear (2). Transcription includes RNA annotation and mapping of promoters for the analysis of gene expression (RAMPAGE, 155), cap analysis gene expression (CAGE, 78), RNA paired-end tag (RNA-PET, 31), microRNA-seq (114), microRNA counts (114), more classical RNA-seq (900) and RNA-microarray (170), including 112 experiments at single-cell resolution. Transcription factor (TF) binding is ChIP–seq on non-histone targets (2,443). Other assays include genotyping array (123), nascent DNA replication strand sequencing (Repli-seq, 104), replication strand arrays (Repli-chip, 63), tandem mass spectrometry (MS/MS, 14), genotyping by high-throughput sequencing (genotyping HTS, 12) and DNA-PET (6) can be looked at in detail at https://www.encodeproject.org.

(Fig. 2a), 2,157,387 open chromatin regions, 750,392 regions with modified histones (mono-, di- or tri-methylation of histone H3 at lysine 4 (H3K4me1, H3K4me2 or H3K4me3), or acetylation of histone 3 at lysine 27 (H3K27ac)), 1,224,154 regions bound by transcription factors and chromatin-associated proteins (Fig. 2c), 845,000 RNA subregions occupied by RNA-binding proteins, and more than 130,000 long-range interactions between chromatin loci. These annotations have greatly enhanced our view of the human genome from its original annotation in 2003 to a much richer and higher-resolution view (for example, Fig. 2d, e). Indeed, although the number of human protein-coding genes known has changed only modestly, the number of transcript isoforms, long noncoding RNAs (lncRNAs), and potential regulatory regions identified has increased greatly since the project began (Fig. 2a–c). An important part of ENCODE 3 is that the regulatory mapping efforts have now been integrated and synthesized into the first version of an encyclopedia, highlighting a registry of 0.9 million cCREs in human and 0.3 million cCREs in mouse. Details can be found in the accompanying ENCODE paper[8] and companion papers in this issue and other journals[9–14].

Technology, quality control and standards

Reaching the present annotation required a substantial expansion of technology development, from ENCODE groups and others, as well as the establishment of standards to ensure that the data are reproducible and of high quality. Most ENCODE 2 assays used sequence-based readouts (for example, RNA-seq[15,16] and ChIP–seq[17,18]) rather than the array-based methods[19,20] used in the pilot phase, and in ENCODE 3, methods such as global mapping of 3D interactions[13] and RNA-binding regions[14] were added. Throughout the project, computational and visualization approaches were developed for mapping reads and integrating different data types (Supplementary Note 1).

A key feature of ENCODE is the application of data standards, including the use of independent replicates (separate experiments on two or more biological samples[5,21]), except when precluded by the limited availability of materials (for example, postmortem human tissues). Of the 8,699 ENCODE 2 and ENCODE 3 experiments, 6,101 have independent replicates. Of equal importance was the use of well-characterized reagents, such as antibodies for mapping sites of transcription factor binding, chromatin modifications and protein–RNA interactions[22]. ENCODE developed protocols to test each antibody 'lot' to demonstrate their experimental suitability, captured extensive metadata, and implemented controlled vocabularies and ontologies. Standards for reagents, experimental data, and metadata are on the ENCODE website: https://www.encodeproject.org/data-standards/.

Many metrics, including sequencing depth, mapping characteristics, replicate concordance, library complexity, and signal-to-noise ratio, were used to monitor the quality of each data set, and quality thresholds were applied[21]. A minority of experiments that fell short of the standards (for example, insufficiently validated antibodies) are still reported, but are marked with a badge to indicate that an issue was found. This is a compromise for having some data versus none when an experiment did not meet ENCODE-defined thresholds.

An important component is uniform data processing. Data from the major ENCODE assays (ChIP–seq, DNase I hypersensitive sites sequencing (DNase-seq), RNA-seq, and whole-genome bisulfite sequencing (WGBS)) are uniformly processed and the processing pipelines are available for users to apply to their own data, by downloading the code from the GitHub (http://github.com/ENCODE-DCC) or by accessing the pipelines at the DNAnexus cloud provider. The standards and pipelines will continue to evolve as new technologies arise and are implemented.

The ENCODE Consortium is a good example of how large-scale group efforts can have a large impact on the scientific community, and many other national and international projects—including the NIH Roadmap Epigenomics Program, The Cancer Genome Atlas (TCGA), the International Human Epigenome Consortium (IHEC), BLUEPRINT, the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC), the Genotype and Tissue Expression Project (GTEx), PsychENCODE, Functional Annotation of Animal Genomes (FAANG), the Global Alliance for Genomics and Health (GA4GH), the 4D Nucleome Program (4DN), the Human Cell Atlas and the FANTOM consortium—have now formed (Supplementary Note 1). ENCODE has engaged with most of these consortia to share standards for data quality control, submission, and uniform processing and has helped to facilitate the use of common ontologies with some of these consortia. Data from the now-completed NIH Roadmap Epigenomics Program have been reprocessed and are available in the ENCODE database and are part of the Encyclopedia annotation. ENCODE continues to work with other consortia, individually and as part of the IHEC and GA4GH (for example, http://epishare-project.org) to increase data interoperability and the value of its resources.

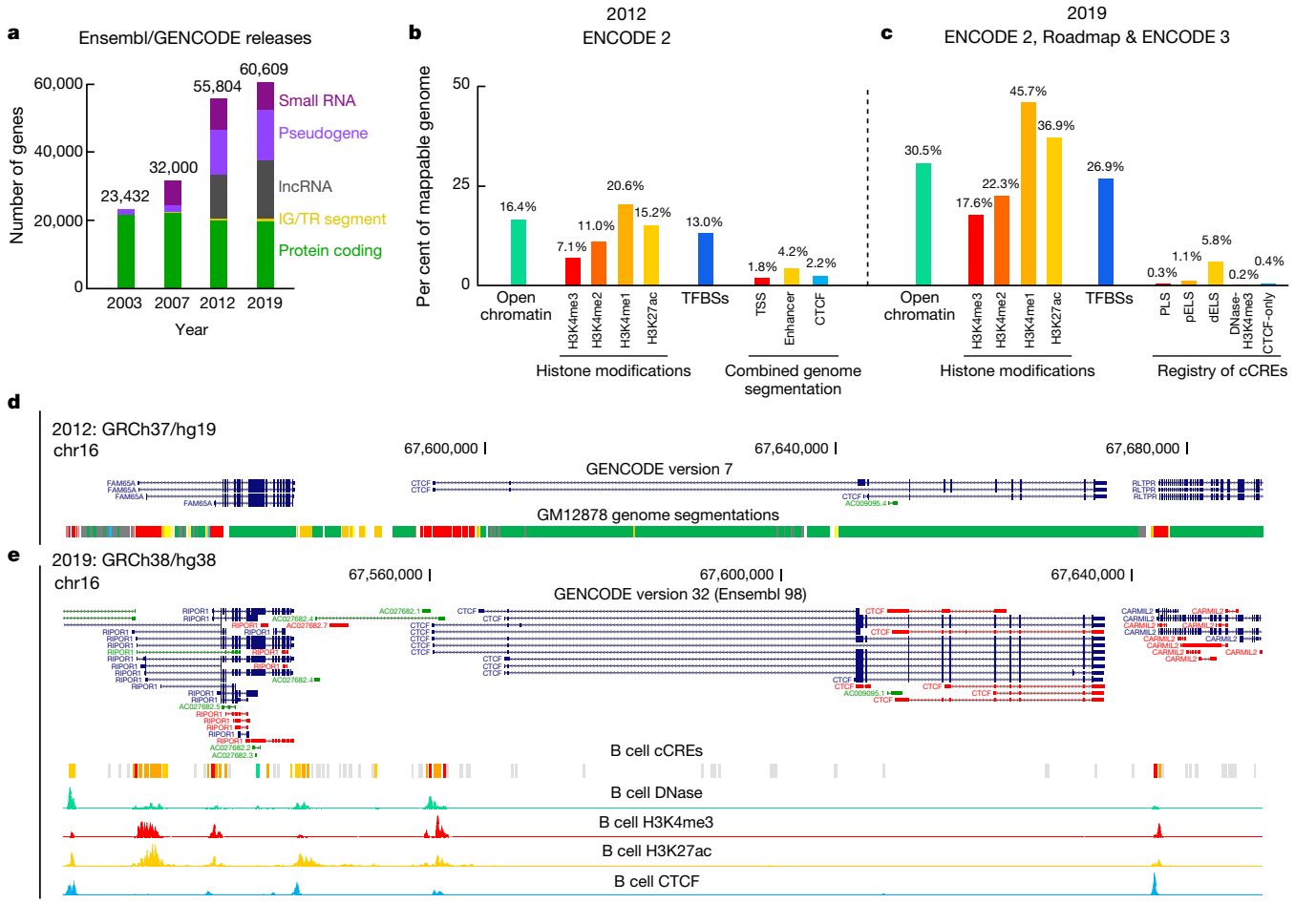


Fig. 2 | Progress in annotating the human genome. Link to high-resolution PDF file: https://www.dropbox.com/s/rjdrcqygz15p034/perspective.pdf?dl=0. **a**, Improvement of gene annotations in the past 15 years by GENCODE, an international gene annotation group that uses ENCODE data[42]. **b**, ENCODE annotations in 2012 with phase II data. Bars show the percentages of the mappable human genome (3.1 billion nucleotides; hg19) that were annotated as open chromatin by DNase-seq data, enriched in four types of active histone mark according to ChIP–seq data, and annotated as transcription factor binding sites (TFBSs) according to ChIP–seq data. Also shown are percentages of the genome assigned as transcription start sites (TSSs), enhancers and the insulator-binding protein (CTCF) by combining ChromHMM and Segway genome segmentations[7]. **c**, ENCODE annotations in 2019 with ENCODE 2, Roadmap, and ENCODE 3 data. The registry of cCREs developed during phase III defines 0.3%, 1.1%, 5.8%, 0.2% and 0.4% of the human genome as cCREs with promoter-like signatures (PLS), proximal enhancer-like signatures (pELS), distal enhancer-like signatures (dELS), with high DNase, high H3K4me3 and low H3K27ac signals (DNase-H3K4me3), and bound by CTCF, respectively. **d**, A UCSC genome browser view of GENCODE genes (V7) coloured by transcript annotation (blue for coding, green for noncoding, and red for problematic) and combined genome segmentation (TSSs in red, enhancers in orange, weak enhancers in yellow, transcription in green, repressed in grey) at the *CTCF* locus on the hg19 human genome. **e**, The UCSC genome browser view of GENCODE genes (V28, coloured as in **d**) and cCREs at the *CTCF* locus on the hg38 human genome[8]. Promoter-like, enhancer-like, and CTCF-only cCREs annotated in B cells are in red, yellow, and blue, respectively. The last four tracks show the DNase, H3K4me3, H3K27ac, and CTCF signals in B cells.

ENCODE as a resource

The purpose of ENCODE is to provide valuable, accessible resources to the community. ENCODE data and derived features are available from a publicly accessible data portal (https://www.encodeproject.org), and consent was obtained from donors to make data freely available to the public. Raw and processed data are available directly from the cloud as an Amazon Public Data Set (https://registry.opendata.aws/encode-project/). The data are widely used by the scientific community—more than 2,000 publications from researchers outside of ENCODE have used ENCODE data to study diverse topics (Fig. 3). Because most disease-associated common variants are noncoding and show substantial enrichment in candidate cell-type-specific *cis* regulatory elements[23,24], ENCODE-derived resources, both in isolation and in conjunction with data from other resources (for example, GTEx), can help to identify and interpret disease-associated noncoding variants (Fig. 3a). Users engage with the data in many ways, ranging from downloads of multiple data sets to detailed investigations of specific loci. Anyone navigating a major genome browser has access to thousands of biochemical, functional, and computational annotations to display at any genomic scale or to overlay on any sequence variant. Maps of epigenomic features relevant to gene regulation have been integrated to form a registry of discrete elements that are candidates for enhancers, promoters, or other regulatory elements. A specialized browser, SCREEN (http://screen.encodeproject.org), is an interface that can be used to identify and study these cCREs and associated ENCODE data and other annotations. This dynamic registry will be regularly updated as additional information is acquired.

Mouse ENCODE and modENCODE

Model organism studies have produced essential insights into almost every aspect of biology, including genome organization and function.

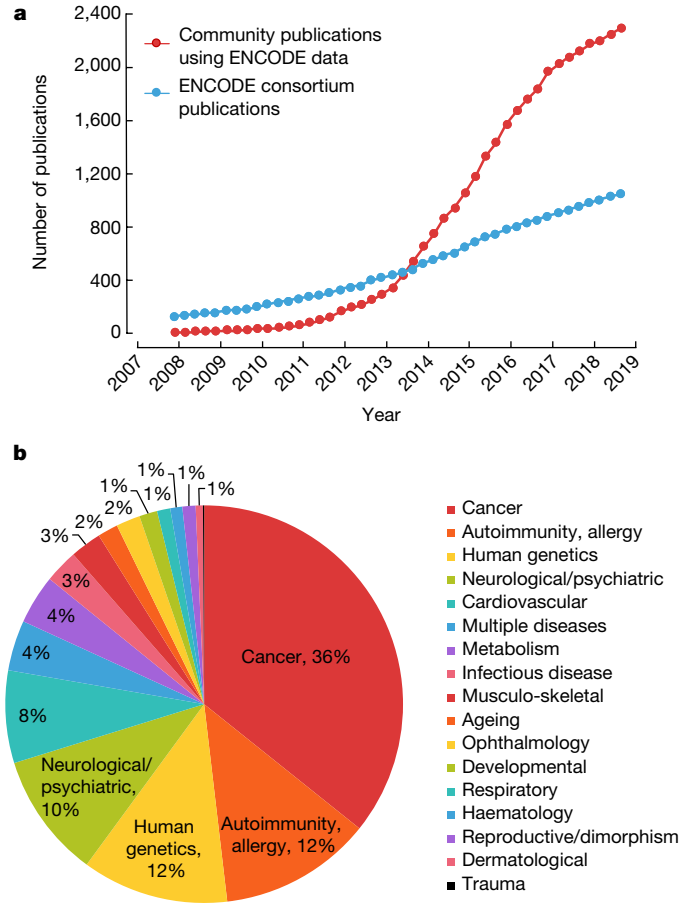


Fig. 3 | Publications using ENCODE data. The National Human Genome Research Institute (NHGRI) has identified a list of publications that used ENCODE data. This list is publicly shared to provide examples illustrating how the resource has been used (https://www.encodeproject.org/publications/). **a**, Publications over time. Community publications appear to use ENCODE data and do not report ENCODE grant support in PubMed; consortium publications report ENCODE grant support in PubMed. In brief, community publications are identified using two steps; first, candidates are identified through automated searches for citation of ENCODE accession numbers, ENCODE flagship papers, or resources such as HaploReg and RegulomeDB; second, candidates are manually evaluated to determine whether ENCODE data were actually used. Consortium papers are identified through automated searches of PubMed for publications that were supported at least in part by ENCODE awards, and are not further evaluated or annotated. **b**, Human disease example publications. The subset of community publications that were annotated as 'human disease' (other categories are basic biology, software tool, fly/worm data) were further manually categorized by disease aetiology.

During ENCODE 2, mapping of mouse epigenomic and transcriptomic features was conducted in adult mouse tissues and cell lines through the Mouse ENCODE Project[25], which identified 21,978 protein-coding regions, 32,168 noncoding genes, 1,192,301 open chromatin regions, 722,334 regions with modified histones H3K4me1, H3K4me2, H3K4me3, or H3K27ac, and 686,294 regions bound by transcription factors.

During ENCODE 2, a model organism ENCODE project (modENCODE[26,27]) was conducted to characterize the transcriptome, epigenome, and transcription factor binding sites in *Drosophila melanogaster* and *Caenorhabditis elegans* tissues, developmental stages and cell lines (Extended Data Fig. 1). These organisms provided the opportunity to develop detailed records of epigenomic features and transcriptome maps throughout development, which is difficult to accomplish in humans. Deep mapping of the spatial and temporal transcriptomes of these species has substantially enhanced the annotation of both genomes. Similarly, detailed mapping of the regulatory circuits that govern gene regulation in *Drosophila* and *C. elegans* has provided insights into general principles of genome organization and function. Mapping of transcription factor binding sites in *Drosophila* and *C. elegans* has continued after modENCODE ended in a project called model organism Encyclopedia of Regulatory Networks (modERN) and to date has characterized more than 262 transcription factors in *Drosophila* and 217 transcription factors in *C. elegans*[28]. Collectively, the modENCODE Project has provided new insights about how the genomes of multicellular organisms direct development and maintain homeostasis.

In ENCODE phase III, experiments were carried out to characterize dynamic histone marks and accessibility, DNA methylomes, and transcriptomes in samples taken during eight mouse fetal developmental stages with up to twelve tissues per stage[28–30] (Fig. 4). The resulting more than 1,500 datasets comprise, to our knowledge, the most comprehensive study of epigenomes and transcriptomes during the prenatal development of a mammal. Integrative analysis of these datasets has expanded our knowledge of the transcriptional regulatory networks that regulate mammalian development and underscored the role of gene regulatory mechanisms in human disease. At least 214,264 of the candidate enhancers identified in fetal mouse tissues are conserved in the human genome[8]. The human orthologues of these potential regulatory elements are significantly enriched for genetic variants that are associated with common illnesses in a tissue-restricted manner, providing information for investigations of the molecular basis of human disease[29,30].

The mouse data from ENCODE 3 also include the results of more than 400 experiments using transgenic reporter mice designed to assess the function of cCREs in three embryonic tissues at two developmental stages. The results of this systematic study have helped to predict the in vivo activities of cCREs. For example, stronger enrichment for epigenetic signatures of enhancer activity correlated with higher rates of validation in the corresponding tissue[29,31].

Finally, comparisons of epigenome and transcriptome maps across species have led to insights into the evolution of transcribed regions and regulatory information[25,32]. Combinatorial histone modification patterns at *cis*-regulatory elements and other genomic features are broadly conserved in metazoans. These chromatin states and transcript levels are highly correlated across tissues and developmental stages in all species examined. However, a notable fraction of specific *cis*-regulatory elements undergoes sequence and functional turnover during evolution, indicating that some regulatory components show substantial plasticity in their evolution while operating in a conserved regulatory network[33].

Current limitations: phase IV and beyond

It is now apparent that elements that govern transcription, chromatin organization, splicing, and other key aspects of genome control and function are densely encoded in the human genome; however, despite the discovery of many new elements, the annotation of elements that are highly selective for particular cell types or states is lagging behind. For example, very few examples of condition-specific activation or repression of transcriptional control elements are currently annotated in ENCODE. Similarly, information from human fetal tissue, reproductive organs and primary cell types is limited. In addition, although many open chromatin regions have been mapped, the transcription factors that bind to these sequences are largely unknown, and little attention has been devoted to the analysis of repetitive sequences. Finally, although transcript heterogeneity and isoforms have been described in many cell types, full-length transcripts that represent the isoform structure of spliced exons and edits have been described for only a small number of cell types.

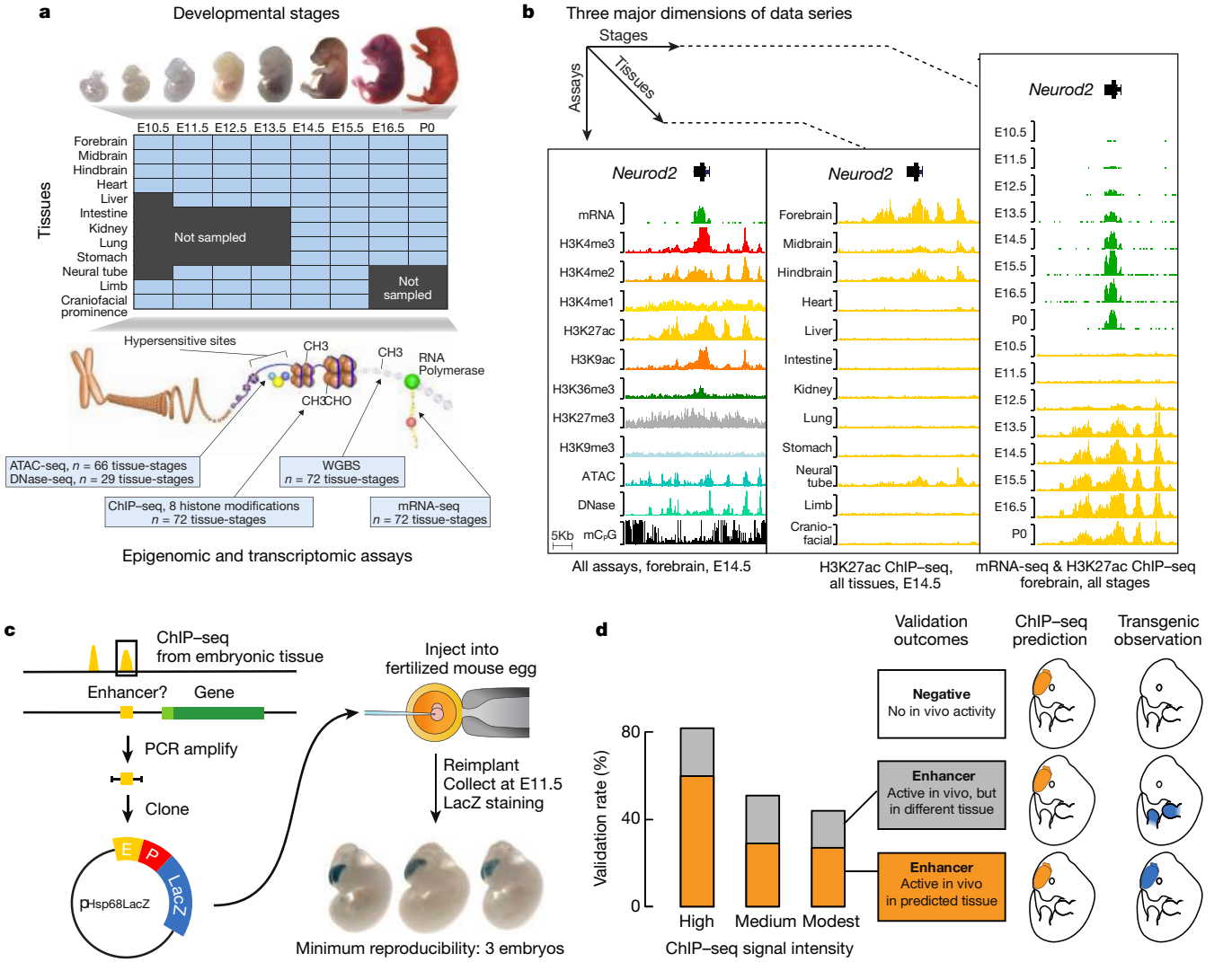


Fig. 4 | An overview of the mouse ENCODE Project in the current phase. **a**, Schematic representation of ENCODE 3 mouse developmental data series. The chromatin graphic is adapted from an image by Darryl Leja (NHGRI), Ian Dunham (EBI), and M.J.P. (NHGRI). The embryo image second from the right in was adapted from ref. [43], an Open Access article distributed under the terms of the Creative Commons Attribution License 2.0. **b**, Three major axes of the data series: assays, tissues, and developmental stages. The region shown is chr11:98,307,637–98,344,383, mm10. **c**, A schematic diagram of the transgenic assays used to validate and characterize the function of cCREs in E11.5 and E12.5 mouse embryos. The cCREs were selected on the basis of ChIP–seq data and cloned into a reporter vector that was then introduced into fertilized mouse eggs. The activities of the CRE were validated by tissue-specific expression patterns of the reporter gene. **d**, Results from recent transgenic assays[8,29] to validate about 400 cCREs are summarized in a barchart, with the bars indicating the proportion of candidate CREs in each rank tier that showed reproducible reporter staining in the expected tissue (grey) or any tissue (pink).

Thus, as part of ENCODE 4, considerable effort is being devoted to expanding the cell types and tissues analysed (see URLs in Supplementary Note 1) as well as mapping the binding regions for many more transcription factors and RNA-binding proteins. These efforts are largely focused in a few reference cell lines, with the hope that improved knowledge will help with imputation or predictions in other cell states[34]. Single-cell transcriptome capture agents[35] and open chromatin assays[36] are also being applied to increase our understanding of the cellular heterogeneity of different tissues and samples. These efforts will supplement the many related activities that are also being pursued by HCA, HuBMAP and others[37,38]. Extensive mapping efforts of all types will continue in both the human and mouse, and parallel efforts to map transcription factor binding sites are being pursued in the *Drosophila* and *C. elegans* by the modERN Project[28]. Full-length transcript isoforms are being elucidated in different cell types using long-read sequencing technologies[39]. ENCODE will continue to work with other consortia, and the data from different groups and individual laboratories will need to be consolidated into a common repository.

Importantly, although very large numbers of noncoding elements have been defined, the functional annotation of ENCODE-identified elements is still in its infancy. High-throughput reporter-based assays[40], CRISPR-based genome and epigenome editing methods[41], and other high-throughput approaches are being used in the current phase of ENCODE to assess the functions of many thousands of elements and to relate those functional results to their biochemical signatures. These targeted functional assays, combined with the large-scale annotation of biochemical features, should further enhance the value of ENCODE data.

Through these and other efforts, it is expected that many more elements in the human genome will be identified across a variety of cell types and conditions, their activities will be revealed (often at the single-cell level), and their biological functions will be inferred more accurately. The development of a systems-wide understanding

of function and integration with genetic information associated with human traits will greatly enhance our understanding of human biology and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2449-8.

1. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
2. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
3. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
4. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
5. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
6. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
   The results of the pilot phase of ENCODE included extensive functional assays across a selected one per cent of the human genome with experiments conducted on a variety of cell lines and largely with array-based technology.
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
   The results of the second phase of ENCODE were based mostly on a large number of genome-wide assays that leveraged high-throughput sequencing technologies and were done across two 'tier one' cell lines with large-scale assays across several hundred cell and tissue types.
8. The ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* https://doi.org/10.1038/s41586-020-2493-4 (2020).
9. Partridge, E. C. et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* https://doi.org/10.1038/s41586-020-2023-4 (2020).
10. Meuleman, W. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* https://doi.org/10.1038/s41586-020-2559-3 (2020).
11. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* https://doi.org/10.1038/s41586-020-2528-x (2020).
12. Breschi, A. et al. A limited set of transcriptional programs define major cell types. Preprint at https://doi.org/10.1101/857169 (2020).
13. Grubert, F. et al. Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* https://doi.org/10.1038/s41586-020-2151-x (2020).
14. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA binding proteins. *Nature* https://doi.org/10.1038/s41586-020-2077-3 (2020).
15. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
16. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
17. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
18. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
19. Iyer, V. R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
20. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
21. Landt, S. G. et al. ChIP–seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
   A consortium-wide effort to standardize performance, quality control and outputs of ChIP–seq experiments, including validation of antibodies, to facilitate experimental reproducibllity and data utility.
22. Sundararaman, B. et al. Resources for the comprehensive discovery of functional RNA elements. *Mol. Cell* **61**, 903–913 (2016).
23. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
24. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
25. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
   Results of a large-scale effort of the mouse ENCODE consortium, presenting regulatory and transcript maps of the mouse.
26. Gerstein, M. B. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
27. The modENCODE Consortium et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).
28. Kudron, M. M. et al. The ModERN Resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics* **208**, 937–949 (2018).
29. Gorkin, D. U. et al. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* https://doi.org/10.1038/s41586-020-2093-3 (2020).
30. He, P. A. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* https://doi.org/10.1038/s41586-020-2536-x (2020).
31. He, Y. et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* https://doi.org/10.1038/s41586-020-2119-x (2020).
32. Cheng, Y. et al. Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).
33. Stefflova, K. et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**, 530–540 (2013).
34. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.* **20**, 9 (2019).
35. Tang, F., Lao, K. & Surani, M. A. Development and applications of single-cell transcriptome analysis. *Nat. Methods* **8** (Suppl), S6–S11 (2011).
36. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
37. Hu, B. C.; HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
38. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
39. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
40. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
41. Klein, J. C., Chen, W., Gasperini, M. & Shendure, J. Identifying novel enhancer elements with CRISPR-based screens. *ACS Chem. Biol.* **13**, 326–332 (2018).
42. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
43. Paudyal, A. et al. The novel mouse mutant, chuzhoi, has disruption of Ptk7 protein and exhibits defects in neural tube, heart and lung development and abnormal planar cell polarity in the ear. *BMC Dev. Biol.* **10**, 87 (2010).

Acknowledgements We thank S. Moore, E. Cahill, M. Kellis and J. Li for their assistance, and B. Wold for helpful comments. This work was supported by grants from the NIH: U01HG007019, U01HG007033, U01HG007036, U01HG007037, U41HG006992, U41HG006993, U41HG006994, U41HG006995, U41HG006996, U41HG006997, U41HG006998, U41HG006999, U41HG007000, U41HG007001, U41HG007002, U41HG007003, U41HG007234, U54HG006991, U54HG006997, U54HG006998, U54HG007004, U54HG007005, U54HG007010 and UM1HG009442.

Author contributions The role of the NHGRI Project Management Group in the preparation of this paper was limited to coordination and scientific management of the ENCODE consortium. All other authors contributed to the concepts, writing and/or revisions of this manuscript.

Competing interests B.E.B. declares outside interests in Fulcrum Therapeutics, 1CellBio, HiFiBio, Arsenal Biosciences, Cell Signaling Technologies, BioMillenia, and Nohla Therapeutics. P.F. is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd. M.P.S. is cofounder and scientific advisory board member of Personalis, SensOmics, Mirvie, Qbio, January, Filtricine, and Genome Heart. He serves on the scientific advisory board of these companies and Genapsys and Jupiter. Z.W. is a cofounder of Rgenta Therapeutics and she serves on its scientific advisory board. R.M.M. is an advisor to DNAnexus and Decheng Capital, and has outside interests in IMIDomics, Accuragen and ReadCoor, Inc. The authors declare no other competing financial interests.

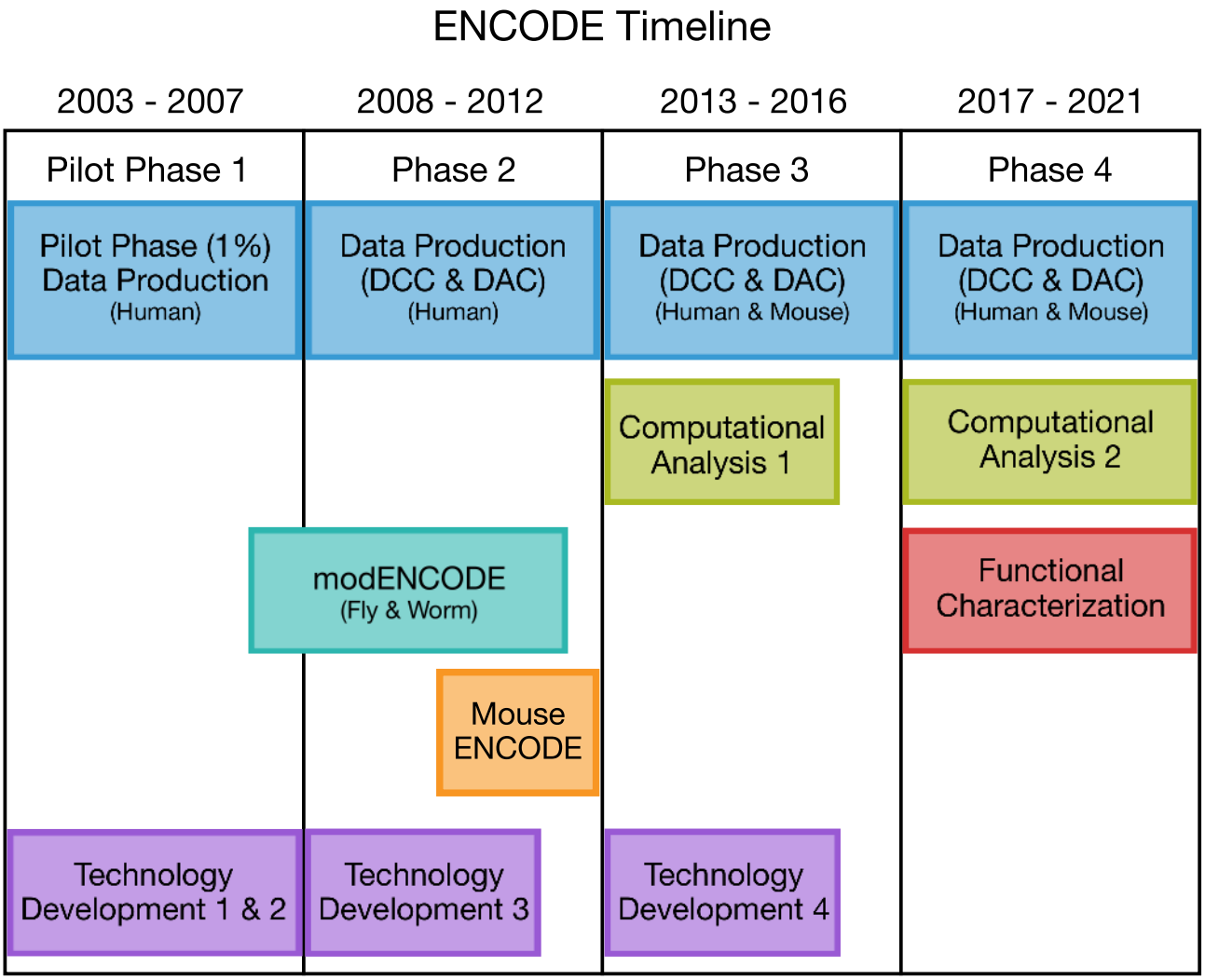
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2449-8.
Correspondence and requests for materials should be addressed to M.P.S.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The ENCODE Project Consortium

Federico Abascal[95], Reyes Acosta[11], Nicholas J. Addleman[1], Jessika Adrian[1], Veena Afzal[49], Bronwen Aken[19], Jennifer A. Akiyama[49], Omar Al Jammal[15], Henry Amrhein[43], Stacie M. Anderson[44], Gregory R. Andrews[4], Igor Antoshechkin[43], Kristin G. Ardlie[20], Joel Armstrong[100], Matthew Astley[95], Budhaditya Banerjee[87], Amira A. Barkal[86], If H. A. Barnes[19], Iros Barozzi[49], Daniel Barrell[19], Gemma Barson[95], Daniel Bates[11], Ulugbek K. Baymuradov[1], Cassandra Bazile[35], Michael A. Beer[98,99], Samantha Beik[20], M. A. Bender[76], Ruth Bennett[19], Louis Philip Benoit Bouvrette[37,38,39], Bradley E. Bernstein[16], Andrew Berry[19], Anand Bhaskar[1], Alexandra Bignell[19], Steven M. Blue[36], David M. Bodine[44], Carles Boix[26], Nathan Boley[1], Tyler Borrman[4], Beatrice Borsari[29], Alan P. Boyle[58,59], Laurel A. Brandsmeier[18], Alessandra Breschi[29], Emery H. Bresnick[93], Jason A. Brooks[4], Michael Buckley[11], Christopher B. Burge[35], Rachel Byron[76], Eileen Cahill[15], Lingling Cai[5], Lulu Cao[1], Mark Carty[96], Rosa G. Castanon[48], Andres Castillo[11], Hassan Chaib[1], Esther T. Chan[1], Daniel R. Chee[11], Sora Chee[8], Hao Chen[4,6], Huaming Chen[48], Jia-Yu Chen[41], Songjie Chen[1], J. Michael Cherry[1], Surya B. Chhetri[17,18], Jyoti S. Choudhary[107], Jacqueline Chrast[101], Dongjun Chung[92], Declan Clarke[7], Neal A. L. Cody[37,38,39], Candice J. Coppola[17,18], Julie Coursen[15], Anthony M. D'Ippolito[46], Stephen Dalton[111], Cassidy Danyko[3], Claire Davidson[19], Jose Davila-Velderrain[26], Carrie A. Davis[3], Job Dekker[77], Alden Deran[100], Gilberto DeSalvo[43], Gloria Despacio-Reyes[95], Colin N. Dewey[89], Diane E. Dickel[49], Morgan Diegel[11], Mark Diekhans[100], Vishnu Dileep[108], Bo Ding[50], Sarah Djebali[29,30], Alexander Dobin[3], Daniel Dominguez[35], Sarah Donaldson[19], Jorg Drenkow[32], Timothy R. Dreszer[1], Yotam Drier[21], Michael O. Duff[14], Douglass Dunn[11], Catharine Eastman[60], Joseph R. Ecker[48,57], Matthew D. Edwards[26], Nicole El-Ali[47], Shaimae I. Elhajjajy[4], Keri Elkins[36], Andrew Emili[61], Charles B. Epstein[20], Rachel C. Evans[18], Iakes Ezkurdia[102], Kaili Fan[4], Peggy J. Farnham[63], Nina Farrell[20], Elise A. Feingold[15], Anne-Maud Ferreira[101], Katherine Fisher-Aylor[43], Stephen Fitzgerald[95], Paul Flicek[19], Chuan Sheng Foo[79], Kevin Fortier[4], Adam Frankish[19], Peter Freese[40], Shaliu Fu[5], Xiang-Dong Fu[41], Yu Fu[4,78], Yoko Fukuda-Yuzawa[49], Mariateresa Fulciniti[22], Alister P. W. Funnell[11], Idan Gabdank[1], Timur Galeev[7], Mingshi Gao[4], Carlos Garcia Giron[19], Tyler H. Garvin[49], Chelsea Anne Gelboin-Burkhart[36], Grigorios Georgolopoulos[11], Mark B. Gerstein[7], Belinda M. Giardine[10], David K. Gifford[26], David M. Gilbert[108], Daniel A. Gilchrist[15], Shawn Gillespie[21], Thomas R. Gingeras[3], Peng Gong[62], Alvaro Gonzalez[96], Jose M. Gonzalez[19], Peter Good[117], Alon Goren[20], David U. Gorkin[8,9], Brenton R. Graveley[14], Michael Gray[95], Jack F. Greenblatt[61,71], Ed Griffiths[95], Mark T. Groudine[76], Fabian Grubert[1], Mengting Gu[7], Roderic Guigó[29], Hongbo Guo[61], Yu Guo[63], Yuchun Guo[26], Gamze Gursoy[7], Maria Gutierrez-Arcelus[87], Jessica Halow[11], Ross C. Hardison[10], Matthew Hardy[19], Manoj Hariharan[48], Arif Harmanci[7], Anne Harrington[49], Jennifer L. Harrow[106], Tatsunori B. Hashimoto[26], Richard D. Hasz[112], Meital Hatan[20], Eric Haugen[11], James E. Hayes[94], Peng He[43], Yupeng He[48], Nastaran Heidari[1,64], David Hendrickson[20], Elisabeth F. Heuston[44], Jason A. Hilton[1], Benjamin C. Hitz[1], Abigail Hochman[35], Cory Holgren[65], Lei Hou[26], Shuyu Hou[5], Yun-Hua E. Hsiao[97], Shanna Hsu[20], Hui Huang[8], Tim J. Hubbard[105], Jack Huey[4], Timothy R. Hughes[61,73], Toby Hunt[19], Sean Ibarrientos[11], Robbyn Issner[20], Mineo Iwata[11], Osagie Izuogu[19], Tommi Jaakkola[26], Nader Jameel[65], Camden Jansen[47], Lixia Jiang[1], Peng Jiang[81,82], Audra Johnson[11], Rory Johnson[29,33], Irwin Jungreis[20,26], Madhura Kadaba[65], Maya Kasowski[1], Mary Kasparian[65], Momoe Kato[49], Rajinder Kaul[11,13], Trupti Kawli[1], Michael Kay[19], Judith C. Keen[113], Sunduz Keles[88,89], Cheryl A. Keller[10], David Kelley[25], Manolis Kellis[20,26], Pouya Kheradpour[26], Daniel Sunwook Kim[1], Anthony Kirilusha[43], Robert J. Klein[94], Birgit Knoechel[22,24], Samantha Kuan[8], Michael J. Kulik[110], Sushant Kumar[7], Anshul Kundaje[1], Tanya Kutyavin[11], Julien Lagarde[29], Bryan R. Lajoie[77], Nicole J. Lambert[35], John Lazar[11], Ah Young Lee[8], Donghoon Lee[7], Elizabeth Lee[49], Jin Wook Lee[1], Kristen Lee[11], Christina S. Leslie[96], Shawn Levy[18], Bin Li[8], Hairi Li[41], Nan Li[50], Xiangrui Li[5], Yang I. Li[1], Ying Li[5], Yining Li[1], Yue Li[26], Jin Lian[62], Maxwell W. Libbrecht[80], Shin Lin[1], Yiing Lin[66], Dianbo Liu[26], Jason Liu[7], Peng Liu[89], Tingting Liu[53], X. Shirley Liu[81,82], Yan Liu[5], Yaping Liu[26], Maria Long[10], Shaoke Lou[7], Jane Loveland[19], Aiping Lu[5], Yuheng Lu[96], Eric Lécuyer[37,38,39], Lijia Ma[65], Mark Mackiewicz[18], Brandon J. Mannion[49], Michael Mannstadt[21], Deepa Manthravadi[95], Georgi K. Marinov[43], Fergal J. Martin[19], Eugenio Mattei[4], Kenneth McCue[43], Megan McEown[18], Graham McVicker[48], Sarah K. Meadows[18], Alex Meissner[27], Eric M. Mendenhall[17,18], Christopher L. Messer[18], Wouter Meuleman[11], Clifford Meyer[81,82], Steve Miller[95], Matthew G. Milton[65], Tejaswini Mishra[1], Dianna E. Moore[18], Helen M. Moore[114], Jill E. Moore[4], Samuel H. Moore[15], Jennifer Moran[65], Ali Mortazavi[47], Jonathan M. Mudge[19], Nikhil Munshi[22], Rabi Murad[47], Richard M. Myers[18], Vivek Nandakumar[11], Preetha Nandi[15], Anil M. Narasimha[1], Aditi K. Narayanan[1], Hannah Naughton[15], Fabio C. P. Navarro[7], Patrick Navas[11], Jurijs Nazarovs[88], Jemma Nelson[11], Shane Neph[11], Fidencio Jun Neri[11], Joseph R. Nery[48], Amy R. Nesmith[18], J. Scott Newberry[18], Kimberly M. Newberry[18], Vu Ngo[50], Rosy Nguyen[18], Thai B. Nguyen[36], Tung Nguyen[50], Andrew Nishida[11], William S. Noble[12], Catherine S. Novak[49], Eva Maria Novoa[26], Briana Nuñez[15], Charles W. O'Donnell[26], Sara Olson[14], Kathrina C. Onate[1], Ericka Otterman[11], Hakan Ozadam[77], Michael Pagan[15], Tsultrim Palden[35], Xinghua Pan[62,67,68], Yongjin Park[26], E. Christopher Partridge[18], Benedict Paten[100], Florencia Pauli-Behn[18], Michael J. Pazin[15], Baikang Pei[7], Len A. Pennacchio[49,54,56], Alexander R. Perez[96], Emily H. Perry[19], Dmitri D. Pervouchine[29,31], Nishigandha N. Phalke[4], Quan Pham[49], Doug H. Phanstiel[69,70], Ingrid Plajzer-Frick[49], Gabriel A. Pratt[36], Henry E. Pratt[4], Sebastian Preissl[8], Jonathan K. Pritchard[1], Yuri Pritykin[96], Michael J. Purcaro[4], Qian Qin[23,84], Giovanni Quinones-Valdez[97], Ines Rabano[36], Ernest Radovani[61], Anil Raj[1], Nisha Rajagopal[87], Oren Ram[20], Lucia Ramirez[1], Ricardo N. Ramirez[47], Dylan Rausch[21], Soumya Raychaudhuri[87], Joseph Raymond[20], Rozita Razavi[71], Timothy E. Reddy[45,46], Thomas M. Reimonn[4], Bing Ren[8,9], Alexandre Reymond[101], Alex Reynolds[11], Suhn K. Rhie[63], John Rinn[28], Miguel Rivera[21], Juan Carlos Rivera-Mulia[108,109], Brian Roberts[18], Jose Manuel Rodriguez[103], Joel Rozowsky[7], Russell Ryan[21], Eric Rynes[11], Denis N. Salins[1], Richard Sandstrom[11], Takayo Sasaki[108], Shashank Sathe[36], Daniel Savic[42], Alexandra Scavelli[3], Jonathan Scheiman[23], Christoph Schlaffner[95], Jeffery A. Schloss[15], Frank W. Schmitges[71], Lei Hoon See[3], Anurag Sethi[7], Manu Setty[96], Anthony Shafer[11], Shuo Shan[4], Eilon Sharon[1], Quan Shen[62,72], Yin Shen[8,51], Richard I. Sherwood[87], Minyi Shi[1], Sunyoung Shin[90], Noam Shoresh[20], Kyle Siebenthall[11], Cristina Sisu[7,104], Teri Slifer[1], Cricket A. Sloan[1], Anna Smith[115], Valentina Snetkova[49], Michael P. Snyder[1,2], Damek V. Spacek[1], Sharanya Srinivasan[87], Rohith Srivas[1], George Stamatoyannopoulos[75], John A. Stamatoyannopoulos[11,12,13], Rebecca Stanton[36], Dave Steffan[65], Sandra Stehling-Sun[11], J. Seth Strattan[1], Amanda Su[35], Balaji Sundararaman[36], Marie-Marthe Suner[19], Tahin Syed[26], Matt Szynkarek[65], Forrest Y. Tanaka[1], Danielle Tenen[20], Mingxiang Teng[85], Jeffrey A. Thomas[116], Dave Toffey[65], Michael L. Tress[103], Diane E. Trout[43], Gosia Trynka[95], Junko Tsuji[4], Sean A. Upchurch[43], Oana Ursu[1], Barbara Uszczynska-Ratajczak[29,34], Mia C. Uziel[20], Alfonso Valencia[103], Benjamin Van Biber[11], Arjan G. van der Velde[4,6], Eric L. Van Nostrand[36], Yekaterina Vaydylevich[15], Jesus Vazquez[102], Alec Victorsen[65], Jost Vielmetter[43], Jeff Vierstra[11], Axel Visel[49,54,55], Anna Vlasova[29], Christopher M. Vockley[20], Simona Volpi[15], Shinny Vong[11], Hao Wang[11], Mengchi Wang[50], Qin Wang[5], Ruth Wang[36], Tao Wang[50], Wei Wang[50], Xiaofeng Wang[37,38,39], Yanli Wang[53], Nathaniel K. Watson[1], Xintao Wei[14], Zhijie Wei[5], Hendrik Weisser[95], Sherman M. Weissman[62], Rene Welch[89], Robert E. Welikson[11], Zhiping Weng[4,5,6], Harm-Jan Westra[87], John W. Whitaker[50], Collin White[18], Kevin P. White[74], Andre Wildberg[50], Brian A. Williams[43], David Wine[20], Heather N. Witt[63], Barbara Wold[43], Maxim Wolf[26], James Wright[95], Rui Xiao[41], Xinshu Xiao[97], Jie Xu[53], Jinrui Xu[7], Koon-Kiu Yan[7], Yongqi Yan[11], Hongbo Yang[53], Xinqiong Yang[1], Yi-Wen Yang[97], Galip Gürkan Yardımcı[12], Brian A. Yee[36], Gene W. Yeo[36], Taylor Young[4], Tianxiong Yu[5], Feng Yue[52,53], Chris Zaleski[3], Chongzhi Zang[81,82,83], Haoyang Zeng[26], Weihua Zeng[47], Daniel R. Zerbino[19], Jie Zhai[1], Lijun Zhan[14], Ye Zhan[77], Bo Zhang[53], Jialing Zhang[62], Jing Zhang[7], Kai Zhang[50], Lijun Zhang[53], Peng Zhang[5], Qi Zhang[91], Xiao-Ou Zhang[4], Yanxiao Zhang[8], Zhizhuo Zhang[26], Yuan Zhao[8], Ye Zheng[88], Guoqing Zhong[61], Xiao-Qiao Zhou[15], Yun Zhu[50] & Jared Zimmerman[108]

[20]The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [21]MGH, Boston, MA, USA. [22]Dana-Farber Cancer Institute, Boston, MA, USA. [23]Harvard Medical School, Boston, MA, USA. [24]Boston Children's Hospital, Boston, MA, USA. [25]Harvard University, Cambridge, MA, USA. [26]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [27]Max Planck Institute for Molecular Genetics, Department of Genome Regulation, Berlin, Germany. [28]University of Colorado Boulder, Boulder, CO, USA. [29]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology and Universitat Pompeu Fabra, Barcelona, Spain. [30]IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, U1220, CHU Purpan, CS60039, Toulouse, France. [31]Skolkovo Institute for Science and Technology, Moscow, Russia. [32]Functional Genomics, Cold Spring Harbor Laboratory, Woodbury, NY, USA. [33]Department of Clinical Research, University of Bern, Bern, Switzerland. [34]International Institute of Molecular and Cell Biology, Warsaw, Poland. [35]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [36]Department of Cellular and Molecular Medicine, Institute for Genomic Medicine, Stem Cell Program, Sanford Consortium for Regenerative Medicine, University of California, San Diego, La Jolla, CA, USA. [37]Département de Biochimie et Médecine Moléculaire, Université de Montréal, Montréal, Quebec, Canada. [38]Division of Experimental Medicine, McGill University, Quebec, Canada. [39]Institut de Recherches Cliniques de Montréal (IRCM), Montréal, Quebec, Canada. [40]Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [41]Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California, San Diego, San Diego, CA, USA. [42]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN, USA. [43]Division of Biology, California Institute of Technology, Pasadena, CA, USA. [44]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [45]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. [46]Center for Genomic and Computational Biology, Duke University, Durham, NC, USA. [47]Biological Sciences, University of California, Irvine, Irvine, CA, USA. [48]Salk Institute for Biological Studies, La Jolla, CA, USA. [49]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [50]Department of Chemistry and Biochemistry, Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, USA. [51]Institute for Human Genetics, Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [52]Department of Biochemistry and Molecular Genetics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [53]Penn State Health Milton S. Hershey Medical Center, Hershey, PA, USA. [54]US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [55]School of Natural Sciences, University of California, Merced, Merced, CA, USA. [56]Comparative Biochemistry Program, University of California, Berkeley, CA, USA. [57]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA, USA. [58]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [59]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [60]Department of Molecular and Cellular Physiology, School of Medicine, Stanford University, Palo Alto, CA, USA. [61]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [62]Department of Genetics, School of Medicine, Yale University, New Haven, CT, USA. [63]Department of Biochemistry and Molecular Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [64]Department of Radiation Oncology, School of Medicine, Stanford University, Palo Alto, CA, USA. [65]Department of Human Genetics, Institute for Genomics and Systems Biology, The University of Chicago, Chicago, IL, USA. [66]Division of General Surgery, Section of Transplant Surgery, School of Medicine, Washington University, St. Louis, MO, USA. [67]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Southern Medical

University, Guangzhou, China. [68]Guangdong Provincial Key Laboratory of Single Cell Technology and Application, Guangzhou, China. [69]Department of Cell Biology & Physiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [70]Thurston Arthritis Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [71]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [72]School of Medicine, Jiangsu University, Zhenjiang, China. [73]Department of Molecular Genetics, Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [74]Tempus Labs, Chicago, IL, USA. [75]Department of Medicine, University of Washington, Seattle, WA, USA. [76]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [77]HHMI and Program in Systems Biology, University of Massachusetts Medical School, Albert Sherman Center, Worcester, MA, USA. [78]University of Massachusetts Amherst, Amherst, MA, USA. [79]Institute for Infocomm Research, Singapore, Singapore. [80]Simon Fraser University, Burnaby, British Columbia, Canada. [81]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. [82]Department of Data Sciences, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA, USA. [83]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [84]Molecular Pathology Unit & Cancer Center, Boston, MA, USA. [85]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA. [86]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA. [87]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [88]Department of Statistics, Medical Sciences Center, Madison, WI, USA. [89]Department of Biostatistics and Medical Informatics, Madison, WI, USA. [90]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA. [91]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA. [92]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA. [93]Department of Cell and Regenerative Biology, UW-Madison Blood Research Program, Carbone Cancer Center, University of Wisconsin School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA. [94]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [95]Wellcome Sanger Institute, Cambridge, UK. [96]Program in Computational Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [97]610 Charles E. Young Drive S, Terasaki Life Sciences Building, Room 2000E, Los Angeles, CA, USA. [98]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. [99]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [100]University of California, Santa Cruz, Santa Cruz, CA, USA. [101]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [102]Centro Nacional de Investigaciones Cardiovasculares (CNIC) and CIBER de Enfermedades Cardiovasculares (CIBERCV), Madrid, Spain. [103]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [104]Brunel University London, London, UK. [105]King's College London, Guy's Hospital, London, UK. [106]ELIXIR Hub, Wellcome Genome Campus, Cambridge, UK. [107]Institute of Cancer Research, London, UK. [108]Department of Biological Science, Florida State University, Tallahassee, FL, USA. [109]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Medical School, Minneapolis, MN, USA. [110]Center for Vaccines and Immunology University of Georgia, Athens, GA, USA. [111]Center for Molecular Medicine and Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA. [112]Gift of Life Donor Program, Philadelphia, PA, USA. [113]American Society for Radiation Oncology, Arlington, VA, USA. [114]National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [115]Leidos Biomedical, Inc, Frederick, MD, USA. [116]National Disease Research Interchange (NDRI), Philadelphia, PA, USA. [117]4407 Puller Drive, Kensington, MD, USA.



Extended Data Fig. 1 | ENCODE timeline. Pilot phase: September 2003–September 2007; ENCODE 2: September 2007–September 2012; ENCODE 3: September 2012–January 2017; ENCODE 4: February 2017–present; modENCODE: April 2007–April 2012; mouse ENCODE: 2009–2012.