The ENCODE Project was launched in 2003, as the first nearly complete human genome sequence was reported. 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, our knowledge of specific elements, particularly with regards to non-protein coding genes and regulatory regions, was restricted to a few well-studied loci.

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. 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. 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.

**ENCODE phase III**
ENCODE 3 (2012–2017) expanded production and added new types of assays (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.
integrated and synthesized into the first version of an encyclopedia, part of ENCODE 3 is that the regulatory mapping efforts have now been increased greatly since the project began (Fig. 2a–c). An important component is uniform data processing. Data from the 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 implemented. 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.

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 (SC, 13). Chromatin accessibility includes DNase-seq (524), assay for transposable-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). 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). 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, 33), 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 (Replica-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.

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.
ENCODE as a resource

The purpose of ENCODE is to provide valuable, accessible resources to the community. ENCODE data and 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 elements13,42, ENCODE-derived features, 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.
During ENCODE 2, mapping of mouse epigenomic and transcriptomic features was conducted in adult mouse tissues and cell lines through the Mouse ENCODE Project\(^4\), which identified 21,978 protein-coding genes, 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\(^6\,7\)) 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\(^8\). 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\(^9\,10\) (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\(^1\). 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\(^11\,12\).

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\(^13\).

Finally, comparisons of epigenome and transcriptome maps across species have led to insights into the evolution of transcriptional regulatory elements and regulatory information\(^14\,15\). 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\(^16\).

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.
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. Single-cell transcriptome capture agents and open chromatin assays are also being applied to increase our understanding of the cellular heterogeneity of different tissues and samples. These efforts will complement the many related activities that are also being pursued by HCA, HuBMAP and others. 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. Full-length transcript isoforms are being elucidated in different cell types using long-read sequencing technologies. 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 to validate about 400 cCREs are summarized in a bar chart, with the bars indicating the proportion of candidate cCREs in each rank tier that showed reproducible reporter staining in the expected tissue (grey) or any tissue (pink).

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 to validate about 400 cCREs are summarized in a bar chart, with the bars indicating the proportion of candidate cCREs in each rank tier that showed reproducible reporter staining in the expected tissue (grey) or any tissue (pink).
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. Kells, M. et al. Defining functional DNA elements in the human genome. Proc. Natl Acad. Sci. USA 111, 6191–6196 (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). Nucleic Acids Res. 44, D301–D315 (2016).
6. Birney, E. et al. Identification and functional analysis of elementary functional units in the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007).
7. 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.
8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
9. 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 type and tissue types.
10. The ENCODE Project Consortium et al. Expanded encyclopedia of DNA elements in the human and mouse genomes. Nature https://doi.org/10.1038/nature15968 (2020).
11. Partridge, E. C. et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature https://doi.org/10.1038/nature15968 (2020).
12. Breschi, A. et al. A limited set of transcriptional programs define major cell types. Preprint at https://doi.org/10.1101/857690 (2020).
13. Grubert, F. et al. Landscape of cohesin-mediated chromatin loops in the human genome. Nature https://doi.org/10.1038/nature15968 (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/nature15968 (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. Science 316, 1596–1600 (2007).
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. Roberts, R. 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).
22. A consortium-wide effort to standardize performance, quality control and outputs of CHIP-seq experiments, including validation of antibodies, to facilitate experimental reproducibility and data utility.
23. Sundararaman, B. et al. Resources for the comprehensive discovery of functional RNA elements. Mol. Cell 61, 903–913 (2016).
24. Mauro, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1910–1915 (2012).
25. 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).
26. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).
27. Results of a large-scale effort of the mouse ENCODE consortium, presenting regulatory and transcript maps of the mouse.
28. 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. Calhii, M. Kells 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, U01HG006992, U10HG006993, U10HG006994, U10HG006985, U10HG006989, U10HG006987, U10HG006968, U10HG007000, U10HG007001, U10HG007002, U10HG007003, U10HG007024, US4HG006981, US4HG006987, US4HG006998, US4HG007004, US4HG007005, US4HG007010 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, HibiBio, Arsenal Biosciences, Cell Signaling Technologies, BioMillenia, and Nahla 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, Minvr, 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 SensOmics, Mirvie, Qbio, January, Filtricine, and Genome Heart. He serves on the scientific advisory board of these companies and Genapsys and Jupiter.

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.
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.