HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 August 19.

Published in final edited form as:

Nature. 2015 February 19; 518(7539): 317–330. doi:10.1038/nature14248.

Integrative analysis of 111 reference human epigenomes

A full list of authors and affiliations appears at the end of the article.

Correspondence should be addressed to: manoli@mit.edu.
†Equal contributors, integrative analysis.
‡Equal contributors, data production and processing.
¥Equal contributors, joint senior authors.

Author contributions and full list of contributors:
Integrative analysis coordination: Anshul Kundaje[1,2,3], Wouter Meuleman[1,2], Jason Ernst[1,2,4]; **Integrative analysis leads:** Misha Bilenky[5,], Angela Yen[1,2], Pouya Kheradpour[1,2], Zhizhuo Zhang[1,2], Alireza Heravi-Moussavi[5], Yaping Liu[1,2], Viren Amin[6], Michael J Ziller[2,7], John W Whitaker[8], Matthew D Schultz[9], Richard S Sandstrom[10], Matthew L Eaton[1,2], Yi-Chieh Wu[1,2], Jianrong Wang[1,2], Lucas D Ward[1,2], Abhishek Sarkar[1,2], Gerald Quon[1,2], Andreas Pfenning[1,2], Xinchen Wang[11,1,2], Melina Claussnitzer[1,2]; **Data production and processing leads:** Cristian Coarfa[6], R Alan Harris[6], Noam Shoresh[2], Charles B Epstein[2], Elizabeta Gjoneska[12,2], Danny Leung[8], Wei Xie[8], R David Hawkins[8], Ryan Lister[9], Chibo Hong[13], Philippe Gascard[14], Andrew J Mungall[5], Richard Moore[5], Eric Chuah[5], Angela Tam[5], Theresa K Canfield[10], R Scott Hansen[10], Rajinder Kaul[15], Peter J Sabo[10]; **Integrative analysis co-leads:** Mukul S Bansal[1,2,16], Annaick Carles[17], Jesse R Dixon[8], Kai-How Farh[2], Soheil Feizi[1,2], Rosa Karlic[18], Ah-Ram Kim[1,2], Ashwinikumar Kulkarni[19], Daofeng Li[20], Rebecca Lowdon[20], Tim R Mercer[21], Shane J Neph[10], Vitor Onuchic[6], Paz Polak[2,22], Nisha Rajagopal[8], Pradipta Ray[19], Richard C Sallari[1,2], Kyle T Siebenthall[10], Nicholas Sinnott-Armstrong[1,2], Michael Stevens[20], Robert E Thurman[10], Jie Wu[23,24], Bo Zhang[20], Xin Zhou[20]; **Analysis and production contributors:** Nezar Abdennur[1,2], Mazhar Adli[25,26], Martin Akerman[24], Luis Barrera[1,2], Jessica Antosiewicz-Bourget[27], Tracy Ballinger[28], Michael J Barnes[14], Daniel Bates[10], Robert JA Bell[13], David A Bennett[30], Katherine Bianco[31], Christoph Bock[2], Patrick Boyle[2], Jan Brinchmann[32], Pedro Caballero-Campo[33], Raymond Camahort[34], Marlene J Carrasco-Alfonso[34], Timothy Charnecki[6], Huaming Chen[9], Zhao Chen[8], Jeffrey B Cheng[29], Stephanie Cho[5], Andy Chu[5], Wen-Yu Chung[19], Chad Cowan[34], Athena Deng[5], Vikram Deshpande[25], Morgan Diegel[10], Bo Ding[8], Timothy Durham[2], Lorigail Echipare[35], Lee Edsall[36], David Flowers[37], Olga Genbacev-Krtolica[31], Casey Gifford[2], Shawn Gillespie[25], Erika Giste[10], Ian A Glass[39], Andi Gnirke[2], Matthew Gormley[31], Hongcang Gu[2], Junchen Gu[20], David A Hafler[40], Matthew J Hangauer[41], Manoj Hariharan[9], Meital Hatan[2], Eric Haugen[10], Yupeng He[9], Shelly Heimfeld[37], Sarah Herlofsen[32], Zhonggang Hou[27], Richard Humbert[10], Robbyn Issner[2], Andrew R Jackson[6], Haiyang Jia[8], Peng Jiang[27], Audra K Johnson[10], Theresa Kadlecek[42,43], Baljit Kamoh[5], Mirhan Kapidzic[31], Jim Kent[28], Audrey Kim[8], Markus Kleinewietfeld[40], Sarit Klugman[8], Jayanth Krishnan[1,2], Samantha Kuan[36], Tanya Kutyavin[10], Ah-Young Lee[36], Kristen Lee[10], Jian Li[6], Nan Li[8], Yan Li[8], Keith Ligon[44], Shin Lin[9], Yiing Lin[9], Jie Liu[8], Yuxuan Liu[19], C John Luckey[34], Yussanne Ma[5], Cecile Maire[44], Alexander Marson[29], John S Mattick[45], Michael Mayo[5], Michael McMaster[31], Hayden Metsky[1,2], Tarjei Mikkelsen[2], Diane Miller[5], Mohammad Miri[25], Eran Mukamel[9], Raman P Nagarajan[13], Fidencio Neri[10], Joseph Nery[9], Tung Nguyen[8], Henriette O'Geen[35], Sameer Paithankar[6], Thalia Papayannopoulou[15], Mattia Pelizzola[9], Patrick Plettner[5], Nicholas E Propson[27], Sriram Raghuraman[6], Brian Raney[28], Anthony Raubitschek[46], Alex P Reynolds[10], Hunter Richards[41], Kevin Riehle[6], Paolo Rinaudo[33], Joshua F Robinson[31], Evan Rosen[34], Eric Rynes[10], Jacquie Schein[5], Renee Sears[20], Terrence Sejnowski[9], Anthony Shafer[10], Li Shen[8], Robert Shoemaker[8], Mahvash Sigaroudinia[14], Igor Slukvin[27], Sandra Stehling-Sun[10], Ron Stewart[27], SaiLakshmi Subramanian[6], Kran Suknuntha[27], Scott Swanson[27], Shulan Tian[27], Hannah Tilden[31], Linus Tsai[34], Mark Urich[9], Ian Vaughn[41], Jeff Vierstra[10], Shinny Vong[10], Ulrich Wagner[36], Hao Wang[10], Tao Wang[8], Yunfei Wang[19], Arthur Weiss[42], Holly Whitton[2], Andre Wildberg[8], Heather Witt[35], Kyoung-Jae Won[8], Mingchao Xie[20], Xiaoyun Xing[20], Iris Xu[1,2], Zhenyu Xuan[19], Zhen Ye[36], Chia-an Yen[36], Pengzhi Yu[27], Xian Zhang[8], Xiaolan Zhang[2], Jianxin Zhao[14], Yan Zhou[31], Jiang Zhu[25], Yun Zhu[8], Steven Ziegler[46]**; Co-Principal Investigators:** Arthur E Beaudet[47], Laurie A Boyer[11], Philip De Jager[34], Peggy J Farnham[35], Susan J Fisher[31], David Haussler[28], Steven Jones[5,48], Wei Li[49], Marco Marra[5,17], Michael T McManus[41], Shamil Sunyaev[2,22,34], James A Thomson[27], Thea D Tlsty[14], Li-Huei Tsai[12,2], Wei Wang[8], Robert A Waterland[50], Michael Zhang[19]**; Scientific Management:** Lisa H Chadwick[51,ǂ] **; Principal Investigators:** Bradley E Bernstein[2,43,25], Joseph F Costello[13], Joseph R Ecker[9], Martin Hirst[5,17], Alexander Meissner[2], Aleksandar Milosavljevic[6], Bing Ren[8], John A Stamatoyannopoulos[10], Ting Wang[20], Manolis Kellis[1,2].

Supplementary Materials

All datasets and analysis results are available from http://compbio.mit.edu/roadmap/. Browsable views of all datasets (as shown in Fig. 3) are available from the WashU Epigenome Browser[100] at http://epigenomegateway.wustl.edu/ and the UCSC Genome Browser[101] at http://genome.ucsc.edu/cgibin/hgTracks?db=hg19&hubUrl=http://vizhub.wustl.edu/VizHub/RoadmapReleaseAll.txt. All primary datasets and protocols are available at REMC portal[102] at http://www.roadmapepigenomics.org, GEO datasets at http://ncbi.nlm.nih.gov/geo/roadmap/epigenomics, and the Human Epigenome Atlas at http://epigenomeatlas.org. Epigenomic annotations and motif predictions are incorporated into HaploReg for mining GWAS at http://compbio.mit.edu/haploreg. Extended data (1-12), supplementary figures (S1-S13), supplementary tables (S1-S6) are available in the online version of the paper.

Abstract

The reference human genome sequence set the stage for studies of genetic variation and its association with human disease, but a similar reference has lacked for epigenomic studies. To address this need, the NIH Roadmap Epigenomics Consortium generated the largest collection to-date of human epigenomes for primary cells and tissues. Here, we describe the integrative analysis of 111 reference human epigenomes generated as part of the program, profiled for histone modification patterns, DNA accessibility, DNA methylation, and RNA expression. We establish global maps of regulatory elements, define regulatory modules of coordinated activity, and their likely activators and repressors. We show that disease and trait-associated genetic variants are enriched in tissue-specific epigenomic marks, revealing biologically-relevant cell types for diverse human traits, and providing a resource for interpreting the molecular basis of human disease. Our results demonstrate the central role of epigenomic information for understanding gene regulation, cellular differentiation, and human disease.

Introduction

While the primary sequence of the human genome is largely preserved in all human cell types, the epigenomic landscape of each cell can vary considerably, contributing to distinct gene expression programs and biological functions[1-4]. Epigenomic information, such as covalent histone modifications, DNA accessibility, and DNA methylation can be interrogated in each cell and tissue type using high-throughput molecular assays[2,5-8]. The resulting maps have been instrumental for annotating cis-regulatory elements and other non-coding genomic features with characteristic epigenomic signatures[9-10], and for dissecting gene regulatory programs in development and disease[7,9,11-14]. Despite these technological advances, we still lack a systematic understanding of how the epigenomic landscape contributes to cellular circuitry, lineage specification, and the onset and progression of human disease.

To facilitate and spearhead these efforts, the NIH Roadmap Epigenomics Program was established, with the goal of elucidating how epigenetic processes contribute to human biology and disease. One of the major components of this program consists of the Reference Epigenome Mapping Centers[15], which systematically characterized the epigenomic landscapes of representative primary human tissues and cells. We used a diversity of assays, including chromatin immunoprecipitation (ChIP)[9-10,16-17], DNA digestion by deoxyribonuclease I (DNase)[7,18], bisulfite treatment [1-2,19-20], methylated DNA immunoprecipitation (MeDIP)[21], methylation-sensitive restriction enzyme digestion (MRE)[22], and RNA profiling[8], each followed by massively-parallel short-read sequencing (-seq). The resulting datasets were assembled into publicly-accessible websites and databases, which serve as a broadly useful resource for the scientific and biomedical community. Here, we report the integrative analysis of 111 reference epigenomes (**Fig. 1**, Extended Data 1a-d), which we analyze jointly with an additional 16 epigenomes previously reported by the ENCyclopedia Of DNA Elements (ENCODE) project[9,23].

We integrate information about histone marks, DNA methylation, DNA accessibility, and RNA expression to infer high-resolution maps of regulatory elements annotated jointly

across a total of 127 cell and tissue types. We use these annotations to recognize epigenome differences that arise during lineage specification and cellular differentiation, to recognize modules of regulatory regions with coordinated activity across cell types, and to identify key regulators of these modules based on motif enrichments and regulator expression. In addition, we study the role of regulatory regions in human disease by relating our epigenomic annotations to genetic variants associated with common traits and disorders. These analyses demonstrate the importance and wide applicability of our data resource, and lead to important insights into epigenomics, differentiation, and disease. Specifically:

- Histone mark combinations show distinct levels of DNA methylation and accessibility, and predict differences in RNA expression levels that are not reflected in either accessibility or methylation.
- Megabase-scale regions with distinct epigenomic signatures show strong differences in activity, gene density, and nuclear lamina associations, suggesting distinct chromosomal domains.
- Approximately 5% of each reference epigenome shows enhancer and promoter signatures, which are 2-fold enriched for evolutionarily-conserved non-coding elements on average.
- Dynamics of epigenomic marks in their relevant chromatin states allow a data-driven approach to learn biologically-meaningful relationships between cell types, tissues, and lineages.
- Enhancers with coordinated activity patterns across tissues are enriched for common gene functions and human phenotypes, suggesting they represent coordinately-regulated modules.
- Regulatory motifs are enriched in tissue-specific enhancers, enhancer modules, and DNA accessibility footprints, providing an important resource for gene-regulatory studies.
- Genetic variants associated with diverse traits show epigenomic enrichments in trait-relevant tissues, providing an important resource for understanding the molecular basis of human disease.

1. Reference epigenome mapping across tissues and cell types

The Reference Epigenome Mapping Centers generated a total of 2,805 genome-wide datasets, including 1,821 histone modification datasets, 360 DNA accessibility datasets, 277 DNA methylation datasets, and 166 RNA-seq datasets, encompassing a total of 150.21 billion mapped sequencing reads corresponding to 3,174-fold coverage of the human genome.

In this manuscript, we focus on a subset of 1,936 datasets (**Fig. 2**) comprising 111 reference epigenomes (**Fig. 2a-d**), which we define as having a core set of five histone modification marks (**Fig. 2e**). The five marks consist of: H3K4me3, associated with promoter regions[10,24]; H3K4me1, associated with enhancer regions[10]; H3K36me3, associated with transcribed regions; H3K27me3, associated with Polycomb repression[25]; and H3K9me3,

associated with heterochromatin regions[26]. Selected epigenomes also contain a subset of additional epigenomic marks, including: acetylation marks H3K27ac and H3K9ac, associated with increased activation of enhancer and promoter regions[27-29] (**Fig. 2f**); DNase hypersensitivity[7,18], denoting regions of accessible chromatin commonly associated with regulator binding (**Fig. 2g**); DNA methylation, typically associated with repressed regulatory regions or active gene transcripts[4,30] and profiled using whole-genome bisulfite sequencing (WGBS)[19], reduced-representation bisulfite sequencing (RRBS)[20], and mCRF-combined[31] methylation-sensitive restriction enzyme(MRE)[22] and immuno-precipitation based[21] assays (**Fig. 2h**); and RNA expression levels[8], measured using RNA-seq and gene expression microarrays (**Fig. 2i**). Our definition of 111 reference epigenomes is very similar to that used by the International Human Epigenome Consortium (IHEC), which required RNA-seq, WGBS, and H3K27ac that are only available in a subset of epigenomes here. Lastly, an additional 16 histone modification marks on average were profiled across 7 deeply-covered cell types (**Fig. 2j**).

We jointly processed and analyzed our 111 reference epigenomes with 16 additional epigenomes from ENCODE[9,23]. We generated genome-wide normalized coverage tracks, peaks and broad enriched domains for ChIP-seq and DNase-seq[7,32], normalized gene expression values for RNA-seq[33], and fractional methylation levels for each CpG site[31,34-35]. We computed several quality control measures (**Fig. 2**, Table S1) including: number of distinct uniquely mapped reads; the fraction of mapped reads overlapping areas of enrichment[18,36]; genome-wide strand cross-correlation[37] (**Fig. 2e-g**); inter-replicate correlation; multidimensional scaling of datasets from different production centers (Fig. S1); correlation across pairs of datasets (Extended Data 1e); consistency between assays carried out in multiple mapping centers (Table S2); and read mapping quality for bisulfite-treated reads[38-39]. Outlier datasets were flagged, removed or replaced, and lower-coverage datasets were combined when possible (See Methods).

The resulting datasets provide global views of the epigenomic landscape in a wide range of human cell and tissue types (**Fig. 3**), including: the largest and most diverse collection to date of chromatin state annotations (**Fig. 3a**); some of the deepest surveys of individual cell types using diverse epigenomic assays (with 21-31 distinct epigenomic marks for seven deeply-profiled epigenomes, **Fig. 3b**); and some of the broadest surveys of individual epigenomic marks across multiple cell types (**Fig. 3c**). These datasets enable genome-wide epigenomic analyses across multiple dimensions (**Fig. 3d**). All datasets, standards and protocols are publicly available from web portals, linked from the main consortium homepage (http://www.roadmapepigenomics.org), including the supplementary website for this paper (http://compbio.mit.edu/roadmap).

2. Chromatin states display highly specific DNA methylation and accessibility

As a foundation for integrative analysis, we learned a common set of combinatorial chromatin states[40] across all 111 epigenomes, plus 16 additional epigenomes generated by the ENCODE project (127 epigenomes in total), using the core set of five histone modification marks that were common to all. We learned a 15-state model (**Fig. 4a,b**, Table S3a) consisting of 8 active states and 7 repressed states (**Fig. 4c**) that were recurrently

recovered (Extended Data 2a), and showed distinct levels of DNA methylation (**Fig. 4d**), DNA accessibility (**Fig. 4e**), regulator binding (Extended Data 2b, Fig. S2), and evolutionary conservation (**Fig. 4f**, Fig. S3). The active states (associated with expressed genes) consist of active TSS-proximal promoter states (TssA, TssAFlnk), a transcribed state at the 5' and 3' end of genes showing both promoter and enhancer signatures (TxFlnk), actively-transcribed states (Tx, TxWk), enhancer states (Enh, EnhG), and a state associated with zinc finger protein genes (ZNF/Rpts). The inactive states consist of constitutive heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, EnhBiv), repressed Polycomb states (ReprPC, ReprPCWk), and a quiescent state (Quies) which covers on average 68% of each reference epigenome. Enhancer and promoter states cover approximately 5% of each reference epigenome on average, and show enrichment for evolutionarily-conserved non-coding regions[41].

To capture the greater complexity afforded by additional marks, we learned additional chromatin state models in subsets of cell types. In the subset of 98 reference epigenomes that also included H3K27ac data, we also learned an 18-state model (Extended Data 2c, Table S3b), enabling us to distinguish enhancer states containing strong H3K27ac signal (EnhA1, EnhA2), which showed higher DNA accessibility (Extended Data 3a), lower methylation (Extended Data 3b), and higher TF binding (Extended Data 2c) than enhancers lacking H3K27ac. In a subset of 7 epigenomes with an average of 24 epigenomic marks, we learned separate 50-state chromatin state models based on all the available histone marks and DNA accessibility in each epigenome (Fig. S4), which additionally distinguished: a DNase-state with distinct TF binding enrichments (Fig. S4f), including for mediator/cohesin components[42] (even though CTCF was not included as an input track to learn the model) and repressor NRSF; transcribed states showing H3K79me1 and H3K79me2 and associated with the 5' ends of genes and introns; and a large number of putative regulatory and neighboring regions showing diverse acetylation marks even in absence of the H3K4 methylation signatures characteristic of enhancer and promoter regions.

We used chromatin states to study the relationship between histone modification patterns, RNA expression levels, DNA methylation, and DNA accessibility. Consistent with previous studies[19,23,43-44], we found low DNA methylation and high accessibility in promoter states, high DNA methylation and low accessibility in transcribed states, and intermediate DNA methylation and accessibility in enhancer states (**Fig. 4d-e**, Extended Data 3a,b). These differences in methylation level were stronger for higher-expression genes than for lower-expression genes, leading to a more pronounced DNA methylation profile (Extended Data 3c, Fig. S5, Table S4f). Genes proximal to H3K27ac-marked enhancers show significantly higher expression levels (Extended Data 3d), and conversely, higher-expression genes were significantly more likely to neighbor H3K27ac-containing enhancers (Extended Data 3e).

Chromatin states sometimes captured differences in RNA expression that are missed by DNA methylation or accessibility. For example, TxFlnk, Enh, TssBiv, and BivFlnk states show similar distributions of DNA accessibility but widely differing enrichments for expressed genes (**Fig. 4c,d**). Enh and ReprPC states show intermediate DNA methylation, but very different distributions of DNA accessibility and different enrichments for expressed genes (**Fig. 4c-e**). Lack of DNA methylation, typically associated with de-repression, is

associated with both the active TssA promoter state and the bivalent TssBiv and BivFlnk states. Bivalent states TssBiv and BivFlnk also showed overall lower DNA methylation and higher DNA accessibility than enhancer states Enh and EnhG and binding by both activating and repressive regulatory factors (Extended Data 2b). These results also held for alternate methylation measurement platforms (Extended Data 4a-c), and for the 18-state chromatin state model (Extended Data 4d-e). Overall, these results highlight the complex relationship between DNA methylation, DNA accessibility, and RNA transcription and the value of interpreting DNA methylation and DNA accessibility in the context of integrated chromatin states that better distinguish active and repressed regions.

Given the intermediate methylation levels of tissue-specific enhancer regions, we directly annotated intermediate methylation (IM) regions, based on the complementary DNA methylation assays of MeDIP[31,45] and MRE-Seq[22,39] in 19 reference epigenomes and 6 additional samples which had both assays available[46]. This resulted in more than 18,000 IM regions, showing 57% CpG methylation on average, that are strongly enriched in genes, enhancer chromatin states (EnhBiv, EnhG, Enh), and evolutionarily-conserved regions. IM was associated with intermediate levels of active histone modification and DNaseI hypersensitivity. Near TSSs, IM correlated with intermediate gene expression, and in exons it was associated with an intermediate level of exon inclusion[46]. IM signatures were equally strong within tissue samples, peripheral blood, and purified cell types, suggesting that IM is not simply reflecting differential methylation between cell types, but likely reflects a stable state of cell-to-cell variability within a population of cells of the same type.

3. Methylation and mark differences across differentiation and cell types

We next studied the relationship between DNA methylation dynamics and histone modifications across 95 epigenomes with methylation data, extending previous studies that focused on individual lineages[19,47-49]. We found that the distribution of methylation levels for CpGs in some chromatin states varied significantly across tissue and cell types (**Fig. 4g**, Extended Data 4f, Table S4a). For example: TssAFlnk states are largely unmethylated in terminally-differentiated cells and tissues, but frequently methylated for several pluripotent and ESC-derived cells (Bonferroni-corrected F-test $p<.01$); Enh and EnhG states are highly methylated in pluripotent cells, but show a broader distribution of intermediate methylation in differentiated cells and tissues ($p<.01$); EnhBiv states are unmethylated in most primary cells and tissues, but show a broader distribution of methylation levels in pluripotent cells, possibly reflecting cell-to-cell heterogeneity ($p<.01$); the repressed state ReprPC shows varying methylation levels among epigenomes; the Het state showed high levels of methylation in almost all epigenomes.

We also studied DNA methylation changes in three different systems. First, we studied DNA methylation changes during Embryonic Stem Cell (ESC) differentiation[49-50]. We identified regions that lost methylation (Differentially Methylated Regions, DMRs, Table S4c) upon differentiation of ESCs (E003) to mesodermal (E013), endodermal (E011), and ectodermal (E012) lineages (**Fig. 4h**). Each lineage showed a largely distinct set of ~2200-4400 DMRs, that are enriched for distinct transcription factor binding events (**Fig. 4h**, right column)[51], consistent with their distinct developmental regulation. Upon further

differentiation, ectodermal DMRs remained hypomethylated in three neural progenitor populations [52], despite the usage of distinct hESC lines, and mesodermal and endodermal DMRs remained highly methylated (**Fig. 4h**), highlighting the lineage-specific nature of changes in DNA methylation during early differentiation[49,53].

Second, we studied DNA methylation changes associated with breast epithelia differentiation[44]. Ectoderm to breast epithelia differentiation was dominated by DNA methylation loss (1.3M CpGs lost methylation vs. 0.2M gained), consistent with other primary somatic cell types[50]. Distinguishing luminal vs. myoepithelial cells by flow sorting, and comparing a set of DMRs (Table S4d) defined specifically in epithelial lineages[44], we found differences in nearest-gene enrichments[54] (mammary gland epithelium development vs. actin filament bundle, respectively), and differences in motif density (luminal DMRs show greater motif density for 51 TFs and lower for 0 TFs). Proximal DMRs were highly associated with increased transcription, consistent with regulatory element de-repression associated with DNA methylation loss.

Third, we asked whether tissue environment or developmental origin is the primary driving factor in DNA methylation differences observed in more differentiated cell types[55], using epigenomes from skin cell types (keratinocytes E057/058, melanocytes E059/E061, and fibroblasts E055/056) that share a common tissue environment but possess distinct embryonic origins (surface ectoderm, neural crest, and mesoderm, respectively). We found that despite the shared tissue environment, these three cell types displayed lower overlap in their DNA methylation and histone modification signatures, and instead were more similar to other cell types with a shared developmental origin. Using a set of DMRs (Table S4e) defined specifically in the skin cell types[55], keratinocytes shared 1392 (18%) of DMRs with surface ectoderm-derived breast cell types (Hypergeometric P-value$<10^{-6}$), and 97% of these were hypomethylated. These shared DMRs were enriched for regulatory elements and cell-type relevant genes, suggesting a common gene regulatory network and shared signaling pathways and structural components[55]. These results suggest that common developmental origin can be a primary determinant of global DNA methylation patterns, and sometimes supersedes the immediate tissue environment in which they are found.

In addition, we examined coordinated changes in chromatin marks associated with cellular differentiation[56]. We found that enhancers showing coordinated differences in multiple marks are enriched near genes showing common tissue-specific expression, and common knockout phenotypes based on their mouse orthologs. For example, enhancers that showed higher H3K27ac and H3K4me3 (**Fig. 4i**, Cluster C2) in left ventricle (E095) relative to their ESCs (E003) and mesendodermal (E004) precursor lineages were enriched for heart ventricle expression and cardiac and muscle phenotypes in their mouse orthologs.

4. Epigenomic dynamics reveal most variable states and distinct chromosomal domains

We next sought to characterize the overall variability of each chromatin state across the full range of cell and tissue types. We first evaluated the observed consistency of each chromatin state at any given genomic position across all 127 epigenomes (**Fig. 5a**). We found that H3K4me1-associated states (including TxFlnk, EnhG, EnhBiv, Enh) are the most tissue-specific, with 90% of instances present in at most 5-10 epigenomes, followed by bivalent

promoters (TssBiv), and repressed states (ReprPC, Het). In contrast, active promoters (TssA) and transcribed states (Tx, TxWk) were highly constitutive, with 90% of regions marked in as many as 60-75 epigenomes, and quiescent regions (Quies) were the most constitutive, with 90% of Quies regions consistently marked as Quies in most of the 127 epigenomes. These results held in the 18-state chromatin state model (Extended Data 5a), and in the subset of highest-quality epigenomes (Fig. S6a,b).

Adjusting for the overall coverage and variability of each state, we then studied differences in the relative fraction of the genome annotated to each chromatin state between cell types (**Fig. 5b**, Extended Data 5b, S6c-e). Immune cells show a consistent and previously unrecognized depletion of active and bivalent promoters (TssA, TssBiv) and weakly transcribed states (TxWk), which may be related to their capacity to generate sub-lineages, replicate, and enter quiescence (reversible G0 phase). ESCs and iPSCs show enrichment of TssBiv, consistent with previous studies[57], and a depletion of ReprPCWk (defined by weak H3K27me3), possibly due to restriction of H3K27me3-establishing Polycomb proteins to promoter regions. Surprisingly, IMR90 fetal lung fibroblasts, which were previously used as a somatic reference cell type[58] are in fact a strong outlier in multiple ways, showing higher levels of Het, ReprPC and EnhG, and a depletion of Quies chromatin states.

We next studied the relative frequency with which different chromatin states switch to other states across different tissues and cell types (**Fig. 5c**), relative to switching across samples of the same tissue or cell type (Fig. S7a,b). This revealed a relative switching enrichment between active states and repressed states, consistent with activation and repression of regulatory regions. The only exception was significant switching between transcribed states and active promoter and enhancer states, possibly due to alternative usage of promoters[22] and enhancers[59] embedded within transcribed elements. These chromatin state switching properties were also found in the 18-state model incorporating H3K27ac marks (Extended Data 5c) and in the subset of 16 ENCODE reference epigenomes using both models (Fig. S7c,d). We found that enhancers and promoters maintained their identity, except for a small subset of regions switching between enhancer signatures and promoter signatures[60]. Luciferase assays showed that these regions indeed possess both enhancer and promoter activity[60], consistent with their epigenomic marks.

While our chromatin state analysis focused at the nucleosome resolution (200-bp), we also studied the overall co-occurrence of chromatin states across tissues at a larger 2Mb resolution to recognize higher-order properties (**Fig. 5d**). This analysis revealed that 2Mb segments rich in active enhancers are constrained to approximately 40% of the genome (clusters c1-c6), with the remainder marked predominantly by inactive regions (c7-c11), consistent with the identification of two large chromatin conformation compartments[12,61]. However, both compartments can be further subdivided by their chromatin state composition: inactive regions separate into predominantly quiescent (40%; c9, c11), heterochromatic (10%, c10), or bivalent (10%, c7-c8) marked regions; and active regions separate into regions rich in multiple marks (c3 and c6, showing a large diversity of active, ReprPC, and bivalent states), weakly-transcribed regions (c5, showing primarily Enh and TxWk states), and regions of intermediate activity (c1, c2, c4). As these subdivisions are based on average state density across a large diversity of cell types, we expected them to be

stable chromosomal features, and indeed, they showed strong differences in gene density, CpG island occupancy, lamina association[62-63] and cytogenetic bands (**Fig. 5d**, Extended Data 5d).

5. Relationships between marks and lineages reveal expanded epigenomic space

We next used epigenome similarity to study the relationship between tissues and cell types, based on the similarity of diverse histone modification marks evaluated in their relevant chromatin states. Hierarchical clustering of our 111 reference epigenomes using H3K4me1 signal in Enh (**Fig. 6a**) showed consistent grouping of biologically-similar cell and tissue types, including ESCs, iPSCs, T-cells, B-cells, adult brain, fetal brain, digestive, smooth muscle, and heart. We also found several initially surprising but biologically-meaningful groupings: fetal brain and germinal matrix samples clustered with neural stem cells rather than adult brain, consistent with fetal neural stem cell proliferation; many ES-derived cells clustered with ESCs and iPS cells rather than the corresponding tissues, suggesting that those are still closer to pluripotent states than corresponding somatic states; adult and fetal thymus samples clustered with T-cells rather than other tissues, consistent with roles in T-cell maturation and immunity. Several marks successfully recovered these biological groupings when evaluated in their relevant chromatin states (Fig. S8), including H3K4me1 in TssA, H3K27me3 in ReprPC state, and H3K27ac in Enh states, suggesting that the signal of each mark in relevant chromatin states is highly indicative of cell type and tissue identity. These alternative clusterings also showed some differences; for example, H3K4me3 in TssA states grouped several fetal samples together with each other, in a cluster sister to ESCs and iPSCs, rather than in separate tissue groups.

We applied this approach to compare the Roadmap Epigenomics reference epigenomes with the 16 ENCODE 2012 samples with broad mark coverage (Extended Data 6). We found that H3K4me1 signal in enhancer chromatin states correctly groups primary cells from similar tissues across the two projects, emphasizing the robustness of our annotations and signal tracks across projects (Extended Data 6a). For example, epidermal keratinocytes NHEK group with other keratinocytes, mammary epithelial cells HMEC with other skin cells, and skeletal muscle myoblasts HSMM and osteoblasts with bone marrow. Some cancer cell lines also grouped with corresponding primary tissues, including hepatocellular carcinoma HepG2 with liver tissue, primary lung fibroblasts NHLF with the IMR90 lung fibroblast cell line, and T cell leukemia Dnd41 with Thymus, while in other cases cancerous cell lines grouped together, e.g. HeLa-S3 cervical carcinoma with A549 lung carcinoma. Similarly, H3K27me3 signal in Polycomb-repressed states grouped five immortalized cell lines together (Extended Data 6c), despite their T-cell, Lung, Cervical, Leukemia, and Hepatocellular origins[12,64]. This larger trees spanning ENCODE 2012 and Roadmap Epigenomics also highlighted the large number of lineages not previously covered by reference epigenomes, including brain, muscle, smooth muscle, heart, mucosa, digestive tract, and fetal tissues.

To understand the relationship among different tissue/cell samples beyond the constraints of a tree representation, we also studied the full similarity matrix of each mark in relevant chromatin states (Fig. S9) and also visualized the principal dimensions of epigenomic

variation using multidimensional scaling (MDS) analysis (Fig. S10). The pairwise similarity matrices of different marks were most effective in distinguishing different subsets of the samples, with H3K4me1 in Enh primarily capturing immune cell similarities, and H3K27me3 in ReprPC capturing pluripotent cell similarities (Fig. S9). In the MDS analysis, the first four dimensions of variation for most marks separated several major sample groups (Extended Data 7a-i), with some subtle differences between marks. For example, pluripotent cells and immune cells were two strong outliers in the first two dimensions of H3K4me1 variation in Enh (**Fig. 6b**), but H3K27me3 in ReprPC showed more uniform spreading of reference epigenomes (**Fig. 6c**), consistent with the coverage distributions of immune and pluripotent cells for the corresponding chromatin states (Fig. 5b). For most marks, the first five dimensions captured most of the variance, with additional dimensions capturing only 4-6% for each mark (Extended Data 7).

6. Epigenome dynamics reveal enhancer modules and their putative regulators

We next exploited the dynamics of epigenomic modifications at *cis*-regulatory elements to gain insights into gene regulation. We focused on 2.3M regions (12.6% of the genome) showing DNA accessibility in any reference epigenome and regulatory (promoter or enhancer) chromatin states, considering enhancer-only, promoter-only, or enhancer-promoter alternating states separately (Fig. S11). We clustered enhancer-only elements (Enh, EnhBiv, EnhG) into 226 enhancer modules of coordinated activity (**Fig. 7a**), promoter-only elements into 82 promoter modules (Fig. S11a) and promoter/enhancer 'dyadic' elements into 129 modules (Fig. S11b), enabling us to distinguish ubiquitously-active, lineage-restricted, and tissue-specific modules for each group. Focusing on the enhancer-only clusters, we found that the neighboring genes of enhancers in the same module showed significant enrichment for common functions[65] (**Fig. 7b**, Fig. S11c,d), common genotype-phenotype associations[66] (**Fig. 7c**), and common expression in their mouse orthologs (Fig. S12), each annotation type showing strong consistency with the known biology of the corresponding tissues. For example, stem-cell enhancers are enriched near developmental patterning genes, immune cell enhancers near immune response genes, and brain enhancers near learning and memory genes (**Fig. 7b**). Sub-clustering of individual modules continued to reveal distinct enrichment patterns of individual sub-modules (Fig. S11e), suggesting increased diversity of regulatory processes beyond the 226 modules used here.

The genome sequence of enhancers in the same module showed substantial enrichment for sequence motifs[67] associated with diverse transcription factors (Fig. S13a). We found 84 significantly enriched motifs in 101 modules (Extended Data 8), indicating that enhancer modules likely represent co-regulated sets, and proposing candidate upstream regulators for nearly half of all modules. Direct application of the same approach and thresholds to the putative regulatory regions annotated in each of the 111 reference epigenomes led to significant enrichment for only 10 enriched motifs in 15 reference epigenomes (Fig. S13b,c) of which 8 are blood samples, and focusing on the regions unique to each of the 17 tissue groups (Fig. 2b) only led to 19 enriched motifs in 10 tissue groups (Fig. S13d,e), emphasizing the importance of studying regulatory motif enrichments at the level of enhancer modules.

We next sought to distinguish likely activator and repressor motifs, by identifying regulators whose expression pattern across cell/tissue types shows a strong (positive or negative) correlation with the activity of enhancers in the enriched module[9]. We focused on the 40 most strongly expression-correlated regulators (Extended Data 9a), and used the module-level motif enrichments to link each regulator to the cell/tissue types that define each module (**Fig. 8**). We found that many of the inferred links correspond to known regulatory relationships, including: OCT4 (also known as POU5F1) in pluripotent cells, HNF1B and HNF4A1 in liver and other digestive tissues, RFX4 in neurosphere and neuronal cells, and MEF2D in muscle. The most enriched regulators showed primarily positive correlations, suggesting they function as transcriptional activators, while a subset of factors showed a negative correlation, with the factor expressed in the lineages where its motif showed enhancer depletion, suggesting a repressive role. For example, REST (also known as NRSF), a known repressor of neuronal lineages was least expressed in neuronal tissues, where its motif was most enriched in enhancers, and a similar signature was found for ZBTB1B, a known repressor of myogenesis and brain development.

Regulatory motifs predicted to be drivers of enhancer activity patterns showed significant enrichment in tissue-specific high-resolution (6bp-40bp) DNase digital genomic footprints (DGF)[68] in matching cell types (Extended Data 9b, Table S5b), providing DNA accessibility evidence that the motifs are indeed bound in these cell types. In addition, they showed positional bias relative to both the center of DGF locations, and relative to their boundaries (Extended Data 10), a property not found for shuffled motifs[69]. These positional biases were highly tissue- and cell type-specific for most activating factors (Extended Data 9c), including POU5F in iPSCs, MEF2D in heart, HNF1B in GI tissues, BHLH in brain, SPI1 in immune cells, and MEF2 in heart and muscle, in each case matching the tissues that showed the highest enrichment. In contrast, for repressive factors and CTCF, positional biases were found in large numbers of tissues, even when the motifs were not enriched in active enhancers. For example, REST (NRSF) was positionally biased in DGF sites in nearly all tissues except brain (Extended Data 9c), even though it was only enriched in active enhancers in brain (Extended Data 9a), consistent with widespread repressive binding in non-brain tissues.

Overall, these enhancer modules, motif enrichments, and regulatory predictions provide an unbiased map that can help guide studies of candidate master regulators for fetal and adult lineage establishment and cell type identity.

7. Impact of DNA sequence and genetic variation on epigenomic state

We next studied the impact of primary DNA sequence on the epigenomic landscape, across genomic regions and between the two alleles of a given individual. First, we evaluated whether histone modifications and DNA methylation can be predicted by the underlying DNA sequence using DNA motifs for TFs expressed in ESCs and four ES-derived cell types. Using the area under the receiver operating curve (AUROC), we found between 71% predictive power for H3K4me1 peaks and 98% for H3K4me3 peaks (average of 85% across six marks and methylation-depleted regions)[70]. The most predictive motifs were those of factors associated with specific histone modifications or specific cell-types, and were found

within peak regions enriched for chromatin marks and at their boundaries. As an example of a boundary enrichment, H3K4me3 peaks were flanked by motifs consisting of a continuous stretch of A and T followed by a G and C cap, which may play a role in nucleosome positioning or recruiting promoter-associated TFs, such as nuclear receptors. Enhancer and promoter-predictive motifs were enriched in high-resolution DNase hypersensitive sites (Table S5a), suggesting they correspond to TF-bound sequences.

Second, we studied how sequence variants between the two alleles of the same individual can lead to allelic biases in histone modifications, DNA methylation, and transcript levels. We reconstructed chromosome-spanning haplotypes for ESCs, four ESC-derived cell lines[71] and 20 tissue samples[60], and we resolved allele-specific activity and structure for each. We found widespread allelic bias in both transcript levels and epigenomic marks for each epigenome. For example, 24% of all testable genes that contain exonic variants demonstrate allelic transcription in one or more ESC or ESC-derived cell lineages, and the majority of these genes also exhibit allelic epigenomic modifications in promoters (71%) and Hi-C-linked enhancers (69%)[71]. Similarly, as many as 11% of the testable enhancers display allelic bias in histone modification H3K27ac in the 20 tissue samples with allele-resolved transcription and chromatin states[60]. Allelic histone acetylation at enhancers is highly specific to individual genotypes, and often occurs near sequence variants that alter transcription factor binding, suggesting *cis*-acting sequence drivers[60,71].

8. Complex trait variants are enriched in diverse epigenomic marks

We next used our tissue-specific epigenomic datasets to study the regulatory annotation enrichments of phenotype-associated variants from genome-wide association studies (GWAS) of diverse traits and disorders. Previous studies showed that disease-associated variants are enriched in specific regulatory chromatin states[9], evolutionarily-conserved elements[72], histone marks[73], and accessible regions[14]. We expanded these analyses using the diversity of primary tissues surveyed by our epigenomic maps, applied to a compendium of disease-associated variants from the NHGRI GWAS catalog[74]. We intersected the set of variants identified in each curated study with peaks of H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3 across each of the 127 epigenomes, and H3K27ac, H3K9ac, and DNase when available (Extended Data 11-12, Table S6), and we searched for significant enrichment in their overlap relative to what would be expected given the NHGRI GWAS catalog as background (see Methods).

For enhancer-associated H3K4me1 peaks, we found 58 studies (**Fig. 9a**, Extended Data 11a) with significant enrichments in at least one tissue at 2% FDR (Hypergeometric $P<10^{-3.9}$). Upon manual curation, the enriched cell types were consistent with our current understanding of disease-relevant tissues for the vast majority of cases. For example, diverse immune traits were enriched in immune cell enhancers, including rheumatoid arthritis, celiac disease, type 1 diabetes, systemic lupus erythematosus, chronic lymphocytic leukemia, allergy, multiple sclerosis, and Graves' disease[75-81]. A large number of metabolic trait variants are enriched in liver enhancer marks, including LDL, HDL, total cholesterol, lipid metabolism phenotypes, and metabolite levels[82-83]. Fasting glucose was most enriched for pancreatic islet enhancer marks, and insulin-like growth factors in placenta, consistent with

their endocrine regulatory roles[84-85]. Several cardiac traits were enriched in heart tissue enhancers, including the PR heart repolarization interval, blood pressure, and aortic root size. Interestingly, inflammatory bowel disease and ulcerative colitis variants show enrichment in both immune and gastrointestinal enhancer marks, suggesting dysregulation of both organs may underlie disease predisposition. Both attention deficit hyperactivity disorder and adiponectin levels were enriched in brain regions, consistent with causal roles in brain dysregulation[86-87]. In contrast, late-onset Alzheimer's disease variants were enriched in immune cell enhancers, rather than brain, consistent with recent evidence of a possible immune and inflammatory basis[88-90].

For active enhancer-associated H3K27ac peaks (available in 98 cell types), we found a similar number of enriched studies (47 at 2% FDR, Extended Data 12b), but for promoter-associated H3K4me3 and H3K9ac peaks, we found only 25 and 18 enriched studies, respectively (Extended Data 12a,b), suggesting that enhancer-associated marks are more informative for tissue-specific disease enrichments than promoter-associated marks. For DNase peaks, we only found 9 enriched studies (Extended Data 12c), partly because they were only available in 53 reference epigenomes (restricting H3K4me1 to the same 53 resulted in 25 enriched studies, Table S6), and possibly due to lack of distinction between enhancer and promoter regions. For transcription-associated H3K36me3, we found 15 enriched studies (Extended Data 12d), indicating that these help capture additional biologically-meaningful variants outside annotated promoter and enhancer regions. In contrast, we found no enriched study for either Polycomb-associated H3K27me3 peaks or heterochromatin-associated H3K9me3 peaks (Extended Data 12e,f). These results indicate that enhancer-associated marks have the greatest ability to distinguish tissue-specific enrichments for regulatory regions, but promoter-, open-chromatin-, and transcription-association marks are also significantly enriched, indicating that disease variants affect a wide range of processes.

These results illustrate that the epigenomic annotations provided here across a broad range of primary tissues and cells, will be of great utility for interpreting genetic changes associated with complex traits. We make all these epigenomic annotations of GWAS regions publicly searchable and browsable through the Roadmap Epigenome Browser[91] and an updated version of the HaploReg database[92].

Discussion

The Reference Epigenome Mapping Consortium has been working to improve epigenomic assays, generate reference epigenomic maps, and use them to understand gene regulation, differentiation, reprogramming, and human disease (see http://www.roadmapepigenomics.org/publications). This paper constitutes the first integrative analysis of all the reference epigenomes generated by the consortium, and represents an early component of the International Human Epigenome Consortium (http://ihec-epigenomes.org/) which seeks to extend such epigenomic maps to more than a thousand reference human epigenomes[93].

In this paper, we use this resource to gain insights into the epigenomic landscape, its dynamics across cell types, tissues, and development, and its regulatory circuitry. We find that combinations of histone modification marks are highly informative of the methylation and accessibility levels of different genomic regions, while the converse is not always true. Genomic regions vary greatly in their association with active marks, with approximately 5% of each epigenome marked by enhancer or promoter signatures on average, which show increased association with expressed genes, and increased evolutionary conservation, while two thirds of each reference epigenome on average are quiescent, and enriched in gene-poor and nuclear lamina-associated stably-repressed regions. Even though promoter and transcription associated marks are less dynamic than enhancer mark, each mark recovers biologically-meaningful cell type groupings when evaluated in relevant chromatin states, allowing a data-driven approach to learn relationships between cell types, tissues, and lineages. The coordinated activity patterns of enhancer regions enable us to cluster them into co-regulated modules, which are proximal to genes with common functions and phenotypes and enriched in regulatory motifs, enabling us to predict candidate upstream regulators.

We also demonstrate the usefulness of the resulting regulatory annotations for interpreting human genetic variation and disease. In an unbiased sampling across the GWAS catalog, we find that genetic variants associated with complex traits are highly enriched in epigenomic annotations of trait-relevant tissues, providing mechanistic insights on the likely relevant cell types underlying genome-wide significant loci. The GWAS enrichments in our analysis were strongest for enhancer-associated marks, consistent with their highly tissue-specific nature. However, promoter-associated and transcription-associated marks were also enriched, implicating several gene-regulatory levels as underlying genetic variants associated with complex traits. These results suggest that our datasets will be valuable in the study of human disease, as several companion papers explore in the context of autoimmune disorders[94-95], Alzheimer's Disease[90,96-97] and cancer[98-99].

Overall, our epigenomic datasets, regulatory annotations, and integrative analyses have resulted in the most comprehensive map of the human epigenomic landscape to date across the largest collection of primary cells and tissues. We expect this map will be of broad use to the scientific and biomedical communities, for studies of genome interpretation, gene regulation, cellular differentiation, genome evolution, genetic variation, and human disease.

Online Methods

1. Data matrix, primary analysis and processing, quality control

All genome-wide maps of histone modifications, DNA accessibility, DNA methylation and RNA expression are freely available online. Raw sequencing data deposited at the Short Read Archive or dbGAP is linked from http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/. All primary processed data (including mapped reads) for profiling experiments are contained within Release 9 of the Human Epigenome Atlas (http://genboree.org/EdaccData/Release-9/). Complete metadata associated with each dataset in this collection is archived at GEO and describes samples, assays, data processing details and quality metrics collected for each profiling experiment.

Release 9 of the compendium contains uniformly pre-processed and mapped data from multiple profiling experiments (technical and biological replicates from multiple individuals and/or datasets from multiple centers). In order to reduce redundancy, improve data quality and achieve uniformity required for our integrative analyses, experiments were subjected to additional processing to obtain comprehensive data for **111 consolidated epigenomes** (See methods sections below for additional details). Numeric epigenome identifiers EIDs (e.g. E001) and mnemonics for epigenome names were assigned for each of the consolidated epigenomes. **Table S1** (QCSummary sheet) summarizes the mapping of the individual Release 9 samples to the consolidated epigenome IDs. Key metadata such as age, sex, anatomy, epigenome class (see below), ethnicity and solid/liquid status were summarized for the consolidated epigenomes. Datasets corresponding to **16 cell-lines from the ENCODE project** (with epigenome IDs ranging from E114-E129) were also used in the integrative analyses[23]. All datasets from the **127 consolidated epigenomes** were subjected to processing filters to ensure uniformity in terms of read length based mappability and sequencing depth as described below.

Each of the 127 epigenomes included consolidated ChIP-seq datasets for a **core set of histone modifications** - H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K9me3 as well as a corresponding whole-cell extract sequenced control. 98 epigenomes and 62 epigenomes had consolidated H3K27ac and H3K9ac histone ChIP-seq datasets respectively. A smaller subset of epigenomes had ChIP-seq datasets for additional histone marks, giving a total of 1319 consolidated datasets (**Table S1**, QCSummary sheet). 53 epigenomes had DNA accessibility (DNase-seq) datasets. 56 epigenomes had mRNA-seq gene expression data. For the 127 consolidated epigenomes, a total of 104 DNA methylation datasets across 95 epigenomes involved either bisulfite treatment (WGBS or RRBS assays) or a combination of MeDIP-seq and MRE-seq assays. In addition to the 1936 datasets analyzed here across 111 reference epigenomes, the NIH Roadmap Epigenomics Project has generated an additional 869 genome-wide datasets, linked from GEO, the Human Epigenome Atlas, and NCBI, and also publicly and freely available.

1.1 RNA-seq uniform processing and quantification for consolidated epigenomes—

We uniformly reprocessed mRNA-seq datasets from 56 reference epigenomes that had RNA-seq data. For RNA-seq analysis, after library construction[44], we aligned 75bp or 100bp long reads using the BWA aligner, and generated read coverage profiles separately for positive and negative strand strand-specific libraries. We used several QC metrics for the RNA-seq library, including intron-exon ratio, intergenic reads fraction, strand specificity (for stranded RNA-seq protocols), 3'-5' bias, GC bias, and RPKM discovery rate (**Table S1**, RNAseqQCSummary sheet). We quantified exon and gene expression using a modified RPKM measure[8], whereby we used the total number of reads aligned into coding exons for the normalization factor in RPKM calculations, and excluded reads from the mitochondrial genome, reads falling into genes coding for ribosomal proteins, and reads falling into top 0.5% expressed exons. RPKM for a gene was calculated using the total number of reads aligned into all merged exons for a gene normalized by total exonic length. The resulting files contain RPKM values for all annotated exons and coding and non-coding genes (excluding ribosomal genes), as well as introns (Gencode V10

annotations were used). We also report the coordinates of all significant intergenic RNA-seq contigs not overlapping the annotated genes.

1.2 ChIP-seq and DNase-seq uniform reprocessing for consolidated epigenomes

a. Read mapping: Sequenced datasets from the Release 9 of the Epigenome Atlas involved mapping a total of 150.21 billion sequencing reads onto hg19 assembly of the human genome using the PASH read mapper[34]. These read mappings were used (except for RNA-seq data sets which were mapped as described above) for constructing the 111 consolidated epigenomes. Only uniquely mapping reads were retained and multiply-mapping reads were filtered out. BED files containing the mapped reads were obtained from http://genboree.org/EdaccData/Release-9/. Alignment parameters for each assay type and experiment are specified in the associated publicly accessible Release 9 metadata archived at GEO. For the ENCODE datasets, BAM files containing mapped reads were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/. Only uniquely-mapping reads were retained and multiply-mapping reads were discarded. Replicates were pooled.

b. Mappability filtering, pooling and subsampling: The raw Release 9 read alignment files contain reads that are pre-extended to 200 bp. However, there were significant differences in the original read lengths across the Release 9 raw datasets reflecting differences between centers and changes of sequencing technology during the course of the project (36 bp, 50 bp, 76 bp and 100 bp). To avoid artificial differences due to mappability, for each consolidated dataset, the raw mapped reads were uniformly truncated to 36 bp and then refiltered using a 36 bp custom mappability track to only retain reads that map to positions (taking strand into account) at which the corresponding 36-mers starting at those positions are unique in the genome. Filtered datasets were then merged across technical/biological replicates, and where necessary to obtain a single consolidated sample for every histone mark or DNase-seq in each standardized epigenome. **Table S1**. summarizes the mapping of the individual Release 9 primary data sample files to the consolidated data files corresponding to the 127 consolidated reference epigenomes.

To avoid artificial differences in signal strength due to differences in sequencing depth, all consolidated histone mark datasets (except the additional histone marks the 7 deeply profiled epigenomes, Fig. 2j) were **uniformly subsampled to a maximum depth** of 30 million reads (the median read depth over all consolidated samples). For the 7 deeply-profiled reference epigenomes (**Fig. 2j**), histone mark datasets were subsampled to a maximum of 45 million reads (median depth). The consolidated DNase-seq datasets were subsampled to a maximum depth of 50 million reads (median depth). These uniformly subsampled datasets were then used for all further processing steps (peak calling, signal coverage tracks, chromatin states).

c. Peak Calling: For the histone ChIP-seq data, the **MACSv2.0.10** peak caller was used to compare ChIP-seq signal to a corresponding whole cell extract (WCE) sequenced control to identify **narrow regions of enrichment** (peaks) that pass a Poisson *p*-value threshold 0.01, **broad domains** that pass a broad-peak Poisson *p*-value of 0.1 and **gapped peaks** which are

broad domains ($p < 0.1$) that include at least one narrow peak ($p < 0.01$) (https://github.com/taoliu/MACS/)[32]. Fragment lengths for each dataset were pre-estimated using strand cross-correlation analysis and the SPP peak caller package (https://code.google.com/p/phantompeakqualtools/)[37] and these fragment length estimates were explicitly used as parameters in the MACS2 program (--shift-size=fragment_length/2).

For DNase-seq data, we used two methods to identify DNaseI-accessible sites. First, the **Hotspot algorithm** was used to identify fixed-size (150bp) DNase hypersensitive sites, and more general-sized regions of DNA accessibility (hotspots) using an FDR of 0.01 (http://www.uwencode.org/proj/hotspot)[103]. MACSv2.0.10 was also used to call narrow peaks using the same settings specified above for the histone mark narrow peak calling.

Narrow peaks and broad domains were also generated for the unconsolidated, 36 bp mappability filtered histone mark ChIP-seq and DNase-seq Release 9 datasets using MACSv2.0.10 with the same settings as specified above.

d. Genome-wide signal coverage tracks: We used the signal processing engine of the **MACSv2.0.10** peak caller to generate genome-wide signal coverage tracks. Whole cell extract was used as a control for signal normalization for the histone ChIP-seq coverage. Each DNase-seq dataset was normalized using simulated background datasets generated by uniformly distributing equivalent number of reads across the mappable genome. We generated 2 types of tracks that use different statistics based on a Poisson background model to represent per-base signal scores. Briefly, reads are extended in the 5' to 3' direction by the estimated fragment length. At each base, the observed counts of ChIP-seq/DNaseI-seq extended reads overlapping the base are compared to corresponding dynamic expected background counts (λ_{local}) estimated from the control dataset. λ_{local} is defined as $max(\lambda_{BG}, \lambda_{1K}, \lambda_{5K}, \lambda_{10K})$ where λ_{BG} is the expected counts per base assuming a uniform distribution of control reads across all mappable bases in the genome and λ_{1K}, λ_{5K}, λ_{10K} are expected counts estimated from the 1 kb, 5 kb and 10 kb window centered at the base. λ_{local} is adjusted for the ratio of the sequencing depth of ChIP-seq/DNase-seq dataset relative to the control dataset. The two types of signal score statistics computed per base are as follows.

(1) **Fold-enrichment** ratio of ChIP-seq or DNase counts relative to expected background counts λ_{local}. These scores provide a direct measure of the **effect size** of enrichment at any base in the genome. (2) negative log10 of the **Poisson *p*-value** of ChIP-seq or DNase counts relative to expected background counts λ_{local}. These **signal confidence scores** provides a measure of **statistical significance** of the observed enrichment.

The -log10(*p*-value) scores provide a convenient way to threshold signal (e.g. 2 corresponds to a *p*-value threshold of 1e-2), similar to what is used in identifying enriched regions (peak calling). We recommend using the signal confidence score tracks for visualization. A universal threshold of 2 provides good separation between signal and noise. Both types of signal tracks were also generated for the unconsolidated datasets using the same parameter settings described above.

e. Quality Control: For the primary Release 9 datasets, data quality enrichment scores were computed as the fraction of the uniquely mapped reads overlapping with areas of enrichment. Several methods were employed to select signal enrichment regions. The **SPOT quality score** was computed based on regions identified with the HotSpot peak caller[103]; the **FindPeaks quality score** was inferred based on peak calls made using the FindPeaks[36] software; finally, a **Poisson metric** was derived by modeling the read distribution in genome-tiling 1000 basepair windows with a Poisson process and selecting as enriched regions windows with $p < 0.05$. All the quality scores in Release 9 are in agreement, with strong pairwise correlation (Pearson correlation > 0.9). **Concordance between centers** was confirmed and data analysis pipeline was validated at the outset of the project using datasets for the H1 cell line. The same pipeline was subsequently used to produce Release 9 data. ChIP-seq data for 6 histone modifications (H3K4me3, H3K27me3, H3K9ac, H3K9me3, H3K36me3, and H3K4me1) were independently generated for the H1 cell line by three REMCs (Broad, UCSD, UCSF-UBC). To quantify concordance, the reads from each experiment were mapped (Level 1 data), read density tracks (Level 2 data) were generated using the EDACC's primary data processing pipeline, and finally Pearson correlation coefficients were computed between each pair of experiments, as well as between experiments and H1 input acting as a control for background correlation between signals (**Table S2**). The methylome processing pipeline was characterized experimentally on four independent samples[38-39]. The same pipeline was used to process bisulfite-treated reads in Release 9 and the same read mappings were used for consolidated epigenomes.

For the uniformly reprocessed and consolidated ChIP-seq and DNase-seq datasets, **strand cross-correlation measures** were used to estimate signal-to-noise ratios (https://code.google.com/p/phantompeakqualtools/)[37]. Datasets for each mark were rank ordered based on the normalized strand cross-correlation coefficient (*NSC*) and flagged if the scores were significantly below the median value or in the range of *NSC* values for WCE extract controls. Consolidated datasets with extremely low **sequencing depth** (< 10M reads) were also flagged. Each standardized epigenome was then manually assigned a subjective **quality flag** of 1 (high), 0 (medium) or −1 (low), based on the number of flagged datasets it contained. The **SPOT, FindPeaks and Poisson** quality scores were also recomputed for the consolidated datasets. We observed high correlations of the NSC scores with the SPOT (Pearson correlation of 0.7) and FindPeaks scores (Pearson correlation of 0.65). All QC measures are provided in **Table S1** (Sheets QCSummary and AdditionalQCScores).

To identify potential antibody cross-reactivity or mislabeling issues, a **pairwise correlation heatmap** (**Extended Data 1e**) was computed across all consolidated datasets for H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac, and DNase. We computed the Pearson correlation between all pairs of the signal tracks based on signal in chr 1-22 and chrX. We used the signal confidence score tracks (-log10(Poisson *p*-value)) where we first computed the average signal scores within each consecutive 25-bp interval. To order the experiments in the heatmap we defined the distance between two pairs of experiments as 1-correlation value and used a traveling salesman problem formulation[104].

1.3 Methylation data cross-assay equalization and uniform processing for consolidated epigenomes—We used PASH[38] alignments for the WGBS and RRBS read alignments. From the number of converted and unconverted reads at each individual CpGs the total coverage and fractional methylation were reported. The data were uniformly post-processed and formatted into two matrices for each chromosome. One matrix contained read coverage information for each base (C and G) in every CpG (row) and for each reference epigenome (column). Another matrix similarly contained fractional methylation ranging from 0 to 1. For the locations where coverage was <=3 we considered data as missing. For MeDIP/MRE methylation data we used the output of the mCRF tool[31] that reports fractional methylation in the range from 0 to 1 and uses an internal BWA mapping. The mCRF results were combined in a single matrix per chromosome for all reference epigenomes where available.

2. Chromatin state learning

In order to capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states) across 127 epigenomes, we used ChromHMM[105], which is based on a multivariate Hidden Markov Model.

2.1 'Core' 15-state model—A ChromHMM model applicable to all **127 epigenomes** was learned by virtually concatenating consolidated data corresponding to the core set of **5 chromatin marks** assayed in all epigenomes (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3). The model was trained on 60 epigenomes with highest-quality data (Fig. 2k), which provided sufficient coverage of the different lineages and tissue types (**Table S1** - Sheet QCSummary). The ChromHMM parameters used were as follows: Reads were shifted in the 5' to 3' direction by 100 bp. For each consolidated ChIP-seq dataset, read counts were computed in non-overlapping 200 bp bins across the entire genome. Each bin was discretized into two levels, 1 indicating enrichment and 0 indicating no enrichment. The binarization was performed by comparing ChIP-seq read counts to corresponding whole-cell extract control read counts within each bin and using a Poisson *p*-value threshold of 1e-4 (the default discretization threshold in ChromHMM). We trained several models with the number of states ranging from 10 states to 25 states. We decided to use a 15-state model (**Fig. 4a-f, Extended Data 2b**) for all further analyses since it captured all the key interactions between the chromatin marks, and because larger numbers of states did not capture sufficiently distinct interactions. The trained model was then used to compute the posterior probability of each state for each genomic bin in each reference epigenome. The regions were labeled using the state with the maximum posterior probability.

2.2 'Expanded' 18-state model—A second "expanded" model applicable to **98 epigenomes** that also have an H3K27ac ChIP-seq dataset, was learned by virtually concatenating consolidated data corresponding to the core set of **5 chromatin marks and H3K27ac**. The model was trained on 40 high quality epigenomes using the same parameters as those used for the primary model (**Table S1** - Sheet QCSummary). We trained several models with the number of states ranging from 15 states to 25 states. An 18 state model was used for further analyses (**Extended Data 2c**) based on similar considerations.

2.3 State labels, interpretation and mnemonics—In order to assign biologically meaningful mnemonics to the states, we used the ChromHMM package to compute the overlap and neighborhood enrichments of each state relative to various types of functional annotations (**Fig. 4b-c,f, Extended Data 2b,c, Fig. S2**).

For any set of genomic coordinates representing a genomic feature and a given state, the fold enrichment of overlap is calculated as the ratio of "the joint probability of a region belonging to the state and the feature" vs. "the product of independent marginal probability of observing the state in the genome" times "the probability of of observing the feature", namely the ratio between the (#bases in state AND overlap feature)/(#bases in genome) and the [(#bases overlap feature)/(#bases in genome) X (#bases in state)/(#bases in genome)]. The neighborhood enrichment is computed for genomic bins around a set of single base pair anchor locations in the genome e.g. transcription start sites.

For the **overlap enrichment** plots in the figures, the enrichments for each genomic feature (column) across all states is normalized by subtracting the minimum value from the column and then dividing by the max of the column. So the values always range from 0 (white) to 1 (dark blue) i.e. its a column wise relative scale. For the **neighborhood positional enrichment** plots, the normalization is done across all columns i.e. the minimum value over the entire matrix is subtracted from each value and divided by the maximum over the entire matrix.

The **functional annotations** used were as follows (All coordinates were relative to the hg19 version of the human genome): (1) CpG islands obtained from the UCSC table browser. (2) Exons, genes, introns, transcription-start-sites (TSSs) and transcription end sites (TESs), 2Kb windows around TSSs and 2Kb windows around TESs based on the GENCODEv10 annotation (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV10/) restricted to long transcripts. (3) Expressed and non-expressed genes, their TSSs and TESs. Genes were classified into the expressed or non-expressed class based on their RNA-seq expression levels in the H1-ESC (Fig. 4c) and GM12878 (Extended Data 2b) cell-lines. A gaussian mixture model with 2 components was fit on expression levels of all genes to obtain thresholds for the two classes. (4) Zinc finger genes (obtained by searching the ENSEMBL annotation for genes with gene names starting with ZNF). (5) Transcription factor binding sites (TFBS) based on ENCODE ChIP-seq data in the H1-ESC cell-line. The uniformly processed TF ChIP-seq peak locations were downloaded from the ENCODE repository: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/. We also computed % TF binding site coverage for states calls in the GM12878 and K562 cell-lines using corresponding TF ChIP-seq data data from ENCODE which matched and supported the mnemonics and state interpretations obtained from the H1 cell-line (**Fig. S2**). (6) Conserved GERP elements based on 34 way placental mammalian alignments http://mendel.stanford.edu/SidowLab/downloads/gerp/ (**Fig. S3**). (7) Conserved non-coding GERP elements obtained by subtracting parts of the above mentioned GERP elements that overlap exons.

2.4 Comparison to chromatin states learned on individual epigenomes—We also learned independent 15-state models individually on each of the 127 epigenomes using

the core set of 5 marks and the same parameter settings as for the primary model. In order to compare the individual models to the joint 15 state primary model, we stacked the emission vectors for all states from all the models and hierarchically clustered them using Euclidean distance and Ward linkage (**Extended Data 2a**). The individual epigenome models consistently and repeatedly identified states that were also recovered by the joint model (Extended Data 2a). Two additional clusters which included states recovered by the independent models learned in individual cell types, but not recovered in the joint model, were HetWk, characterized by weak presence of H3K9me3, and Rpts, characterized by presence of H3K9me3 along with a diversity of other marks, which was enriched in a large number of repeat elements.

2.5 Expanded chromatin states using large numbers of histone marks—For each of the seven deeply-profiled reference epigenomes (Fig. 2j) we independently learned chromatin states on observed data for all available histone marks or variants, and DNase in the reference epigenome. The same binarization and model learning procedure was followed as for the core set of 5 marks. We chose to consistently focus on a larger set of 50-states to capture the additional state distinctions afforded by using additional marks (**Fig. S4**). Enrichments for annotations, including some of those described above for the 15-state model, were computed using ChromHMM. The HiC domains were obtained from [106], the lamina associated domains are described below, conserved element sets were the hg19 lift-over from [72], repetitive elements are from RepeatMasker.

3. Relationship between histone marks, methylation, and DNase

The distribution of DNA methylation (percent CpG methylation from WGBS data) and DNA accessibility (DNase-seq -log10(p-value) signal confidence scores) was computed using regions belonging to each of the 15 chromatin states based on the core set of 5 marks across all reference epigenomes for which these datasets were available. (**Fig. 4d,e**).

CpGs with a minimum read coverage of 5 were used to calculate the average methylation percentages within genomic regions labeled with each chromatin state from the 15 state primary model. Only regions containing more than 3 CpGs with at most 200bp between consecutive CpGs were used. Plots were generated using ggplot2 package for R (v.3.02). The average methylation levels for the chromatin states across DNA methylation platforms (WGBS, RRBS and mCRF) were analyzed using Standard Least Square models in JMP (v. 11.0; SAS Ins.). The model included the platforms (3 levels), chromatin states (15 levels) and the interactions (Extended Data 4).

4. Calling of Lamina Associated Domains

Genome-wide DamID binding data for human Lamin B1 in SHEF-2 ESCs were obtained from GEO series GSE22428[62]. Lamina Associated Domains were determined using a similar method to the one described in [63]. First, hg18-based data coordinates were converted to hg19-based coordinates using UCSC's liftOver tool. Data was smoothed using a running median filter with a window size of 5 probes, after which domains were detected by estimating border and domain positions and comparing these to domains defined on 100

randomized instances of the same data set. Parameters are chosen such that the False Discovery Rate (FDR) for detected domains is 1%.

5. Cell-type specificity and switching of chromatin states

5.1 Chromatin state variability—For each state ***s*** for the core 15 state joint model, we computed the number of genomic bins that were labeled with that state in at least one epigenome (G_s). From amongst these bins we counted the number of bins ($g_{s,i}$) that were labeled as being in state ***s*** in exactly ***i*** epigenomes (i=1..127). We converted these counts to fractions ($g_{s,i} / G_s$) and computed the cumulative fraction that is consistently labeled with the same chromatin state in at most ***N*** epigenomes (N=1..127). States whose cumulative fractions rise faster than others represent those that are less constitutive (more variable). We repeated the same procedure restricted to 43 high quality and non-redundant Roadmap epigenomes (using only 1 representative epigenome from those corresponding to the ESC lines, iPS lines or epigenomes for the same tissue type from different individuals and excluding ENCODE cell-lines) (**Table S1** - Sheet VariationAnalysis) (**Fig. S6a**). Analogous analysis was performed on states from the 18 state expanded model (**Extended Data 5a, Fig. S6b**).

The observed cumulative fractions of cell-type specificity are a function of the composition of cell-types in the compendium and do depend to some extent on the variability of data quality for the different marks. For example, the enhancer mark (H3K4me1) does have a much better signal-to-noise ratio than the transcribed mark (H3K36me3). One might expect this to result in more spurious variation of states associated with the transcribed mark. However, contrary to this expectation, the cumulative fractions for states involving only the transcription mark (Tx and TxWk) and not the enhancer mark indicate that these states are in fact less variable and more constitutive across cell types. On the other hand, all states composed of the enhancer mark (H3K4me1), irrespective of whether they do (TxFlnk, EnhG) or do not (EnhBiv, Enh, BivFlnk, TssAFlnk) include the transcription mark (H3K36me3), are far more cell-type specific. These observation indicate that the increased variability of states are largely due to the enhancer mark (H3K4me1) than the transcribed mark (H3K36me3). As replicates are not available in all epigenomes, we did not correct for inter-replicate variation in this analysis, but in the state switching analysis below, we utilize samples from the same tissue as quasi-replicates.

5.2 Chromatin state switching—To avoid spurious switching due to differences in data quality, we restricted this analysis to chromatin states from the 43 high quality and non-redundant Roadmap epigenomes (see above). Using the 15 state primary model, we computed the empirical switching frequency of any pair of states across all pairs of 43 epigenomes. For a given pair of states A and B, we counted the number of genomic bins that were labeled as (A,B) or (B,A) in all pairs of epigenomes. The switching frequency matrix (which is symmetric) was then row-normalized in order to convert the switching frequencies to switching probabilities. This is done to avoid a dependence on the total number of epigenomes. Also, the switching probabilities unlike switching frequencies are not dominated by states that are highly prevalent (e.g. quiescent state). **Fig. S7b**. shows the empirical switching probabilities for all pairs of states across the 43 epigenomes. In order to

differentiate between chromatin state dynamics across tissues (inter-tissue) relative to variation of states across individuals or replicates from the same tissue (intra-tissue), we also computed analogous switching frequencies by restricting to subgroups of epigenomes from the same tissue type (**Table S1** - Sheet VariationAnalysis). The frequencies were added across all sub-groups and then row normalized to switching probabilities. **Fig. S7a**. shows the intra-tissue switching probabilities. We then computed the relative enrichment of state switches as the log10 ratio of inter-tissue switching probability across the 43 epigenomes relative to the intra-tissue switching probabilities (**Fig. 5c**). We repeated this analysis on the 16 ENCODE cell-lines and obtained similar conclusions regarding relative enrichment of state switches (**Fig. S7c**). Analogous analyses were performed using the 18-state expanded model in Roadmap Epigenomics samples (**Extended Data 5c**) and ENCODE samples (**Fig. S7d**).

5.3 Large scale chromatin structure—To study large-scale chromatin structure we first calculated ChromHMM (15 states model) state-frequencies identified in 200bp genome-wide bins across 127 epigenomes. Then we averaged state frequencies over the 2Mb genomic regions, thus defining 1458 long vector for each state. The unsupervised clustering of a 15×1458 matrix (using Pearson correlation as a similarity measure and complete linkage) revealed 11 distinct genomic clusters enriched in different subsets of chromatin states (**Fig. 5d**, top heat map). Clusters had different sizes, with the smallest one (c1) containing only 27 bins, while the largest cluster (c9), occupied predominantly by a 'Quiescent' state for all epigenomes, had 377 bins. For each 2Mb bin in each cluster we calculated average gene density, Lamin B1 signal (see section 4 above), overlap with different cytogenetic bands (**Fig. 5d**, bottom, that displays also average levels across each cluster). We also show chromosomal locations of the clusters as well as distributions of CpG island frequency across the 2Mb bins in each cluster (**Extended Data 5d**).

6. Differentially Methylated Regions (DMRs) and DNA methylation variation

6.1 DMR calls across reference epigenomes—For the global epigenomic comparisons, we defined DMRs using the Lister *et al* method[107], combining all differentially methylated sites within 250bp of one another into a single DMR and excluded any DMR with less than 3 DMSs. For each DMR in each sample, we computed its average methylation level, weighted by the number of reads overlapping it[108]. This resulted in a methylation level matrix with rows of DMRs and columns of samples.

6.2 DMRs in hESC differentiation (Fig. 4h)—For analyzing differentiation of hESCs in **Fig. 4h**, we used a second set of DMRs. We used a pairwise comparison strategy between ESCs and three *in vitro* derived cell types representative of the three germ layers (mesoderm, endoderm, ectoderm) and performed DMR calling as previously described [52]. Only DMRs losing more than 30% methylation compared to the ESC state at a significance level of $p \leq 0.01$ were retained. Subsequently, we computed weighted methylation levels for all three DMR sets across HUES64, mesoderm, endoderm and ectoderm as well as three consecutive stages of in vitro derived neural progenitors (please see companion[52] paper for details on the cell types). Finally, we plotted the corresponding distribution using the R function vioplot in the vioplot package. In order to identify potential regulators associated

with the loss of DNA methylation at these regions, we determined binding sites of a compendium of transcription factors profiled in distinct cell lines and types (see Ziller, 2013 #45 for details) that overlapped with each set of hypomethylated DMRs. Next, we determined a potential enrichment over a random genomic background by randomly sampling 100 equally sized sets of genomic regions, respecting the chromosomal and size distribution of the different DMR sets and determined their overlap with the same transcription factor binding site compendium to estimate a null distribution. Only transcription factors that showed fewer binding sites across the control regions in 99 of the cases were considered for further analysis. Next, we computed the average enrichment over background for each TF with respect to the 100 sets of random control regions for each germ layer DMR and report this enrichment level in **Fig. 4h** right, where we capped the relative enrichment at 12.

6.3 Additional DMR calls—For studying breast epithelia differentiation, DMRs were called from WGBS, requiring at least 5 aligned reads to call differentially-methylated CpG, and at least 3 differentiallymethylated CpGs within a distance of 200 bp of each other[44]. For studying tissue environment vs. developmental origin, DMRs were called from MeDIP and MRE data using the M&M algorithm[55].

6.4 DNA methylation variation—For variation in methylation of each chromatin state across epigenomes (**Fig. 4g, Extended Data 4f**), we first excluded any contiguous chromatin state region containing less than 3 CpG sites. Then, the mean of the methylation level for all contained CpG sites was calculated for each region, and for each epigenome, density values were calculated for these mean methylation values between 0% and 100%, with density values estimated over n=1000 points with a gaussian kernel, with a default bandwidth of ‘nrd0’. Finally, for each chromatin state, we plotted the ln(density+1) for each epigenome as rows, with the color scale set with white as the minimum ln(density+1) value and red, green, or blue, for WGBS, mCRF, and RRBS, respectively, set as the maximum ln(density+1) value in the matrix. Rows were ordered by the epigenomic lineage and grouping ordering shown in **Fig. 2a**; In **Extended Data 4f**, epigenomes were first grouped by methylation platform, and then ordered by **Fig. 2a** within each platform. The chromatin state methylation profiles in the cell lines vs primary cells/tissue cells were analyzed using a mixed model with repeated measures. Overall effect of the group (cell lines vs primary cells/ tissue cells) was tested using epigenomes within group as the error term. Bonferroni correction was used for adjusting the p-values.

7. Identifying coordinated changes in chromatin marks during development

To identify patterns of coordinated changes of histone marks over enhancers during heart muscle development, we compared ESCs, Mesendoderm cells, and Left Ventricle tissue[56]. We identified relevant enhancers as those that show changes in at least one histone mark between a specific cell type cluster (heart muscle in our case) and other cell types using LIMMA (Linear Model for Microarray Analysis). We applied FDR corrected P-value significance threshold of 0.05 to obtain cluster-specific enhancers. For each tissue type (heart muscle in our case) we then clustered the enhancers into five clusters (C1-C5) based on their multi-mark epigenomic profiles using the k-means algorithm implemented in the

Spark tool (**Fig. 4i**). The tools used to generate **Fig. 4i** are integrated into the Epigenomic Toolset within the Genboree Workbench and are accessible for online use at www.genboree.org.

8. Clustering of epigenomes reveals common lineages, common properties

For each analyzed mark, we calculated Pearson correlation values between all pairwise combinations of reference epigenomes using the mark's signal confidence scores (-log10(Poisson *p*-value)) within 200bp of the genomic regions deemed relevant for that mark. Relevance of regions is determined by whether a region was called in a particular (mark-matched) chromatin state with posterior probability of > 0.95 in any of the reference epigenomes. For H3K4me1, H3K27ac and H3K9ac we used state Enh, for H3K4me3 state TssA, for H3K27me3 state ReprPC, for H3K36me3 state Tx and for H3K9me3 state Het, unless otherwise noted (all based on the 15-state core model).

The resulting correlation matrices were used as the basis for a distance matrix for complete-linkage hierarchical clustering, followed by optimal leaf ordering[109]. Bootstrap support values are derived from 1,000 random samplings with replacement from all regions considered for a particular mark and a bootstrap tree was estimated for each resampling. The bootstrap support for a branch corresponds to the fraction of bootstrapped trees that support the bipartition induced by the branch.

In parallel to this, all correlation matrices mentioned above were used to perform Multi-Dimensional Scaling analyses using R.

9. Delineation of DNaseI-accessible regulatory regions

For each of the 39 Roadmap reference epigenomes with DNase data, peak positions are combined across reference epigenomes by defining peak island areas, defined by stacking all DNase peak positions across epigenomes, and considering the Full Width at Half Maximum (FWHM). Note that for this we are only considering peak locations, not intensities. The goal of this is to obtain an estimate of the area of open chromatin, not to quantify the level of 'openness', as these data are not available for all reference epigenomes. In cases when peak islands overlap, they are merged because it means that the original DNase peak area populations overlap at least for half of the epigenomes with DNase peaks in that area (given the FWHM approach). Peak island summits are defined as the median peak summit of all peak island member DNase peaks. This results in a total of 3,516,964 DNase enriched regions across epigenomes.

We then annotate each of the ~3.5M DNase peaks with the chromatin states they overlap with in each of the 111 Roadmap reference epigenomes, using the core 15-state chromatin state model, and focusing on states TssA, TssAFlnk, and TssBiv for promoters, and EnhG, Enh, and EnhBiv for enhancers, and state BivFlnk (flanking bivalent Enh/Tss) for ambiguous regions. Out of these, ~2.5M regions are called as either enhancer or promoter across any of the 111 Roadmap reference epigenomes. Note that because DNase data is not available for all Roadmap epigenomes, the set of regulatory regions defined may exclude DNase regions active in cell types for which DNase was not profiled (Fig. 2g). Although

most regions are undisputedly called exclusively promoter or enhancer, there are ~530k regions that needed further study to decide whether they should be called promoters, enhancers, or both ('dyadic'). We arbitrate on these regions by first clustering them (using the methods in the following section) with an expected cluster size of 10,000 regions, and then for each cluster calculating (a) the mean posterior probabilities for promoter and enhancer calls separately, and (b) the mean number of reference epigenomes in which regions were called promoter or enhancer. Clusters of regions for which the differences in mean posterior probabilities (a) is smaller than 0.05, *or* for which the absolute log2-ratio of the number of epigenomes called as promoter or enhancer (b) is smaller than 0.05 are called true 'dyadic' regions, along with a small number of 'ambiguous' regions in state BivFlnk. Note that this particular clustering is only to arbitrate on these regions using group statistics instead of one-by-one; the final clusterings are described next. Overall, we define ~2.3M putative enhancer regions (12.63% of genome), ~80k promoter regions (1.44% of genome) and ~130k dyadic regions (0.99% of genome), showing either promoter or enhancer signatures across epigenomes.

10. Clustering of DNaseI-accessible regulatory regions to identify modules of coordinated activity

In order to cluster regulatory (i.e., enhancer, promoter or dyadic) regions based on their activity patterns across all reference epigenomes, we expressed each region in terms of a binary vector of length n*x*s, where *n* is the number of reference epigenomes (111) and *s* is the number of chromatin states considered. For enhancers and promoters, *s*=3, as both of these types of regions are made up of 3 chromatin states in the 15-state ChromHMM model (enhancers: EnhG, Enh & EnhBiv, promoters: TssA, TssAFlnk & TssBiv).

The thus obtained binary matrices are subsequently clustered using a variation of a k-centroid clustering algorithm[110]. Instead of Euclidean distance we use a Jaccard-index based distance. This is done to be able to correctly cluster highly cell type restricted regions. From a computational point of view, we optimized the method to both deal with the size of the used data matrices and leverage their sparsity, in order to efficiently compute and update distances for matrices with sizes on the order of $10^6 \times 10^3$. The algorithm has been further modified to converge when less than 0.01% of cluster assignments change between iterations.

We selected the number of clusters *k* by tuning the expected number of regions within each cluster to be approximately 1000 for promoter and dyadic regions, and approximately 10,000 for enhancer regions, given their much larger count (81k, 129k, and 2.3M for promoter, dyadic, and enhancer respectively). This results in a value of *k*=233 for enhancer clusters (for ~10k elements per cluster), and the algorithm converged on *k*=226 non-empty clusters, which are used for subsequent analyses.

Clusters are visualized (**Fig. 7a**) by 'diagonalizing' when possible. First, 'ubiquitous' clusters (defined as having at least 50% of epigenomes with an enhancer/promoter density of > 25%) are shown. Then, the remaining clusters are ordered according to which epigenome has the maximum enhancer density.

Enrichment analyses of proximity to gene members of a catalogue of gene sets (Gene Ontology (GO), Human Phenotype Ontology (HPO)) have been performed using the GREAT tool[54]. In particular, the GREAT web API was used to automatically submit region descriptions and retrieve results for subsequent parsing. We restricted ourselves to interpretation of results with an enrichment ratio of at least 2, and multiple hypothesis testing corrected, p-values < 0.01 for both the binomial and the hypergeometric distribution based tests.

For visualization of a representative subset of enriched terms in **Fig. 7b** and **Fig. 7c**, we select representative terms for display (after diagonalizing the enrichment matrix by re-ordering the rows). We do this using a weighted bag-of-words approach to select highly-enriched terms that contain many words that are overrepresented in gene-set labels showing similar enrichment patterns. Briefly, sliding along the row names (gene-set terms) of the, diagonalized, enrichment matrices, we collect word counts and multiply these by integer-rounded -log10(q-values) obtained from GREAT. We do this in sliding windows of size 33 for **Fig. 7b** (resulting in 35 terms) and size 16 for **Fig. 7c** (resulting in 15 terms). For each word in a window, these values are expressed relative to the same words across all row names, registering to what extend they are over-represented. Each gene-set term in the window is then assigned a score based on the mean over-representation of all words it consists of. Lastly, gene-sets are co-ranked based on this mean over-representation and their GREAT significance. The best-ranked gene set label is selected as the representative label for that window. All terms are shown in Fig. S11d and available for download on the supplementary website.

11. Predicting regulators active in each tissue / cell type / lineage

We collect 1,772 known TF recognition motifs (position weight matrices) from primarily large-scale databases[67,111-116] and measure their enrichment in the enhancers for each enhancer module compared to the union of the 226 enhancer modules (as described in [67] and [9]) using a 0.3 conservation-based confidence cutoff[69,72]. We cluster motifs using a 0.75 correlation cutoff resulting in 300 motif clusters[67] and select for each motif cluster the motif with the highest enrichment in any enhancer module for further analysis.

We compute an expression score for each enhancer module and transcription factor as the Pearson correlation between the TF expression across cell types with expression data (quantile-normalized log(RPKM) with zeros replaced by log(0.0005)) and the ‘center’ of a module. For each enhancer module, its center is defined as a vector of length 111, containing the fraction of regions in that module called as (any type of) enhancer in each of the 111 epigenomes analyzed. This expression score is meant to act as the “expression” of a transcription factor within a module of cell types. We then compute an expression-enrichment value for each transcription factor as the correlation of this expression score and the enrichment of the corresponding motif across enhancer modules. The top 40 motifs in terms of their absolute expression-enrichment correlation and the clusters with log2 enrichment or depletion of at least log2=1.5 for at least one motif are shown in **Fig. 8** and **Extended Data 8a** (only one motif is shown in **Fig. 8** for each factor).

We show all 84 motifs that were significantly enriched (log2>=1.5) in any enhancer modules, across the full set of 226 enhancer modules (Fig. S13a) and in the 101 modules in which they were significantly enriched (Extended Data 8a). Similarly, we show all 10 enriched motifs across the full set of 111 individual reference epigenomes (Fig. S13b) and specifically in the 15 enriched epigenomes (Fig. S13c). Lastly, we show all 19 enriched motifs across the full set of 17 tissue groups (Fig. S13d), and specifically within the 10 groups that showed significant enrichments (Fig. S13e).

For visualization of regulator-cell type links (Fig. 8), we compute edge weights between each cell type and motif using these motif-module enrichments. For each motif and cell type, we compute the sum across all modules of the product of the log2 motif enrichment and the value of the cell type within the module center (only consider the highly associated cell types by replacing values <0.7 with 0). We show all resulting edge weights of at least 1.5 and visualize the network using Cytoscape[117].

Based on the same motif enrichment method mentioned above, we computed the motif enrichment in the tissue-specific DGF regions in each library. The tissue-specific DGF regions were identified by selecting the DGF region occurring no more than 20 DGF libraries among 42 DGF libraries. To generate **Extended Data 9b**, we standardized the motif enrichment in each library into z-scores for each motif (row) and color each DGF library (column) based on their tissue type.

12. DNA Motif Positional Bias in Digital Genomic Footprinting Sites

We compute the positional enrichment of each driver motif (**Extended Data 9c, Extended Data 10**) related to the Digital Genomic Footprinting(DGF) sites in each cell type (Table S5b). For each driver TF motif, we generated two views corresponding to the motif position (the center of the motif instance) relative to the center of closest DGF site (center view) and the motif position relative to the boundary of closest DGF site (boundary view). We only consider the motif instances with closest DGF site within 100bp. For center view, we plotted the motif occurrence density respect to the distance to DGF center for different cell type. For the boundary view, we considered the shortest distance of the center of a motif instance to the both side of DGF boundary, and gave a negative distance value if the motif instance is inside the DGF, otherwise the distance value is positive. Similar to center view, we plot the motif density respect to the derived distance value in the boundary view for each cell type.

To access the significance of the motif concentration within DGF in each cell type, we compute the DGF enrichment ratio as the ratio between the number of motif instance with distance less than 20bp to the DGF center and that number in the immediate flanking window, that is, the number of motif instance with distance to the DGF center larger than 20bp and smaller than 40bp. As control, we randomly sampled the same number of motif instances from the shuffled versions of the given motif, and obtained the DGF enrichment ratio for the shuffled motif instances. The DGF enrichment ratio of the true motif is further converted to z-score by mean and standard deviation from the DGF enrichment ratios of shuffled motif from 1000 times random sampling. Then the adjusted p-value is further computed from z-score and bonferroni correction for number of cell types.

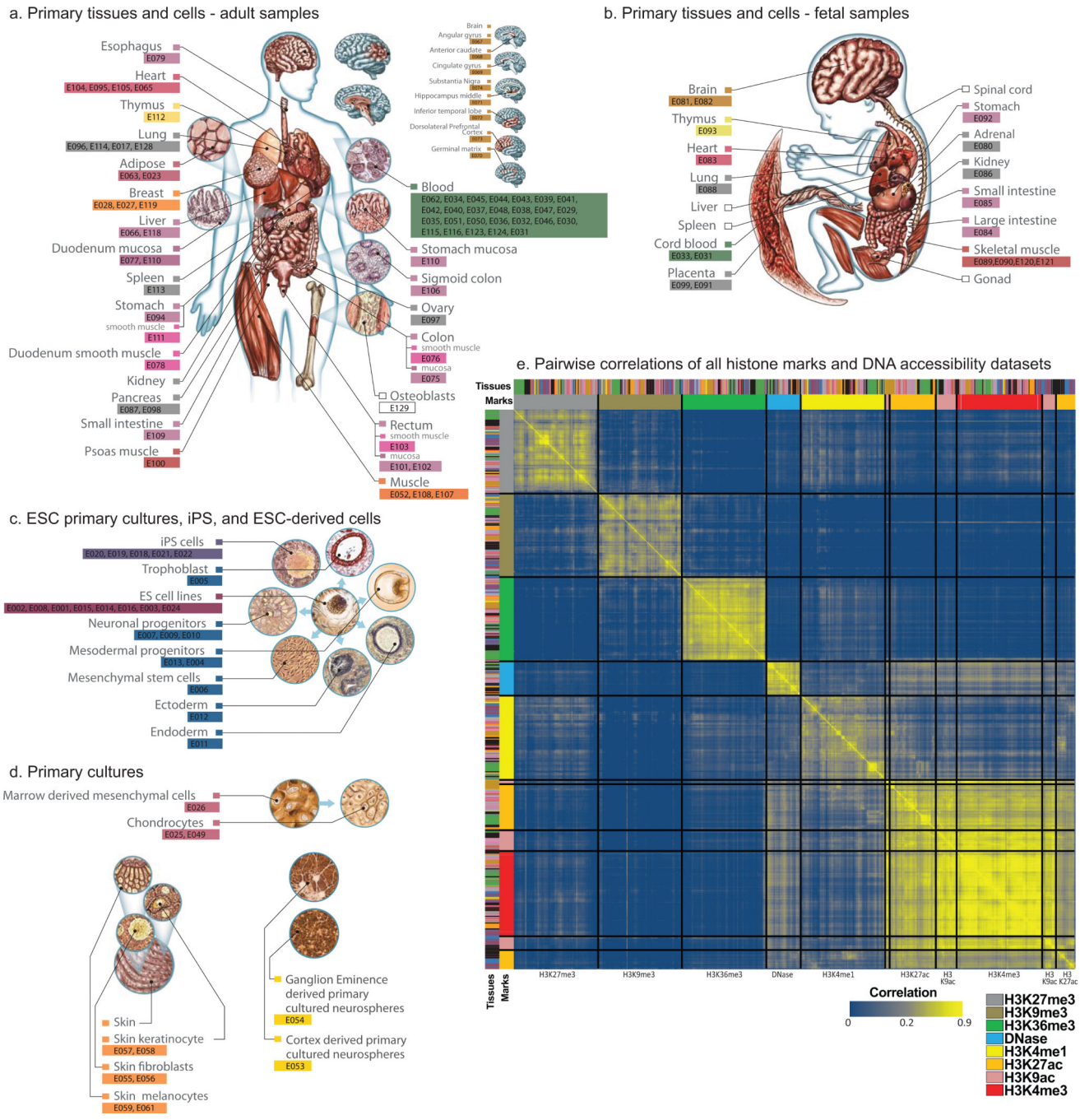
13. Comparing Digital Genomic Footprinting with DNA motifs that are predictive of epigenomic modification

The motifs that were predictive of epigenomic modifications[70] were compared to Digital Genomic Footprinting data (DGF) in Table S5. This was done in three cell types where both DGF and predictive motifs were available: 'H1 BMP4 Derived Mesendoderm Cultured Cells' (E004), 'H1 BMP4 Derived Trophoblast Cultured Cells' (E005), and 'H1 Derived Mesenchymal Stem Cells' (E006). The motifs that were predictive of the following seven modifications were considered: H3K27me3, H3K27ac, H3K9me3, H3K36me3, H3K4me1, H3K4me3 and DNA methylation valleys (DMV)[11]. To identify overlaps the predictive motifs were scanned against the modification peaks of the corresponding modification and the location of the best match between motif and sequence was recorded. Then we counted the number of times the locations of the best motif matches overlapped a DGF by at least one bp. These counts were compared to the number of overlaps identified randomly, which was calculated by comparing DGF to random locations within the modifications peaks. The reported random frequency was the average of 100 repeats. To calculate the fold enrichment we divided the observed frequency by the random frequency.

14. Tissue-specific activity of disease-associated regions

We tested the enrichment of SNPs from individual Genome-wide Association Studies (GWAS) for the gapped peak call sets based on histone marks H3K4me1, H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K9ac, and H3K27ac as well as the DNase peak call set based on MACS2 in each reference epigenome where available. The SNPs used were curated into the NHGRI GWAS catalog[74] and obtained through the UCSC Table Browser[118] on September 12, 2014. We restricted the enrichment analysis to chr1-22 and chrX. We defined a study to be a unique combination of annotated trait and pubMedID. To reduce dependencies between pairs of SNPs assigned to the same study, we pruned SNPs such that no two SNPs were within 1MB of each other on the same chromosome. The pruning procedure considered each SNP in ranked order of p-value with the the most significant coming first, and we retained a SNP if there was no already retained SNP on the same chromosome within 1MB. We computed hypergeometric p-values for the enrichment of each pruned set of SNPs overlapping peak calls against the pruned GWAS catalog as the background. We estimated separately for each mark a mapping from a p-value to a false discovery rate across tests for all study and reference epigenome combinations by generating 100 randomized versions of the pruned GWAS catalogs shuffling which SNPs were assigned to which study and computing the average fraction of reference epigenome–study combinations that reached that level of significance (in a continuous mapping of p-values to FDR) using randomized catalogs divided by the number based on the actual GWAS catalog.

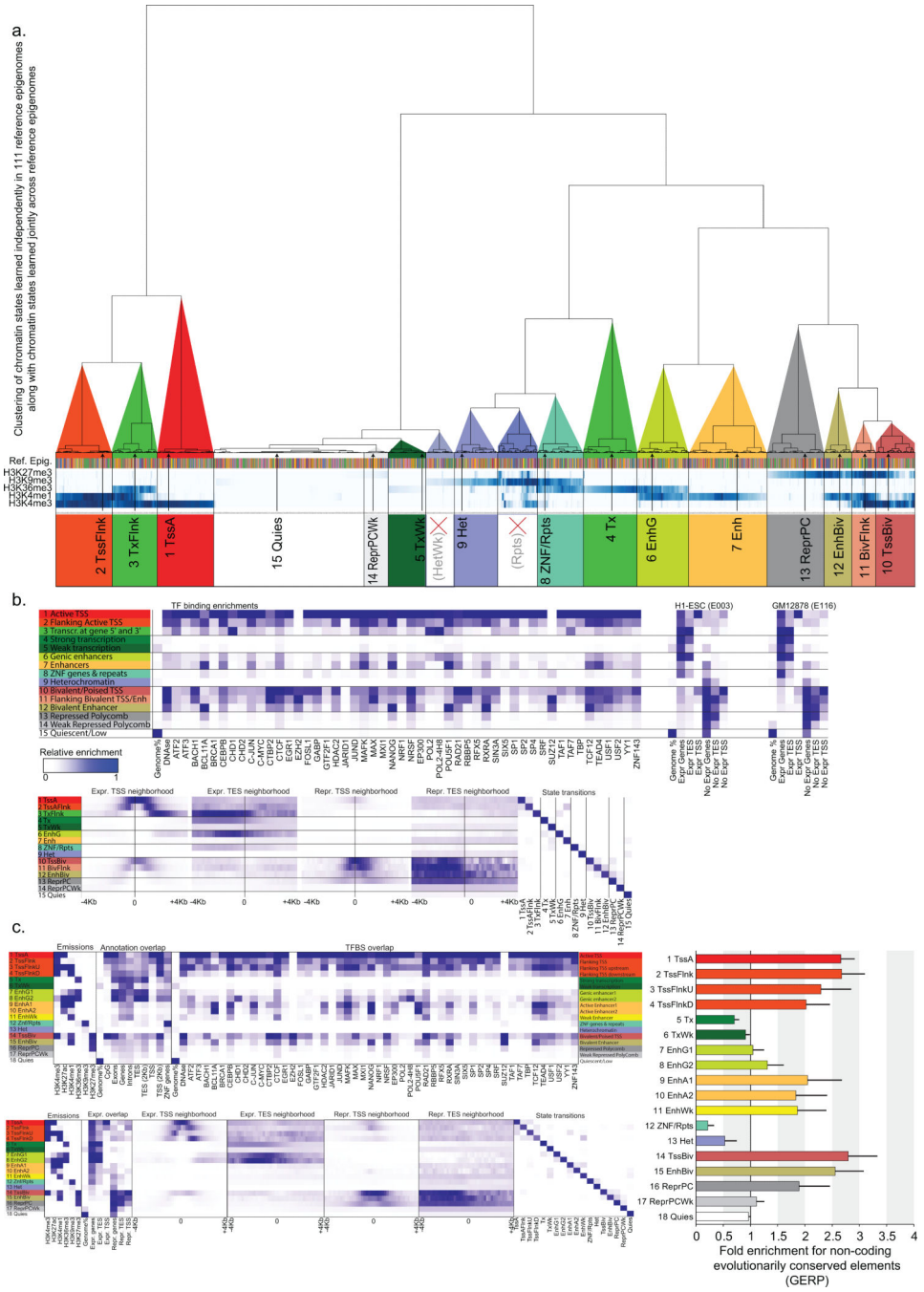
Extended Data



Extended Data 1.

a-d. Tissues and Cell Types of Reference Epigenomes. Comprehensive listing of all 111 reference epigenomes generated by the consortium, along with epigenome identifiers (EIDs), including: (a) adult samples; (b) fetal samples; (c) ESC, iPSC, and ESC-derived cells; and (d) primary cultures. Colors indicate the groupings of tissues and cell types (as in Fig. 2b, and throughout the manuscript). For five samples (adult osteoblasts, and fetal liver,

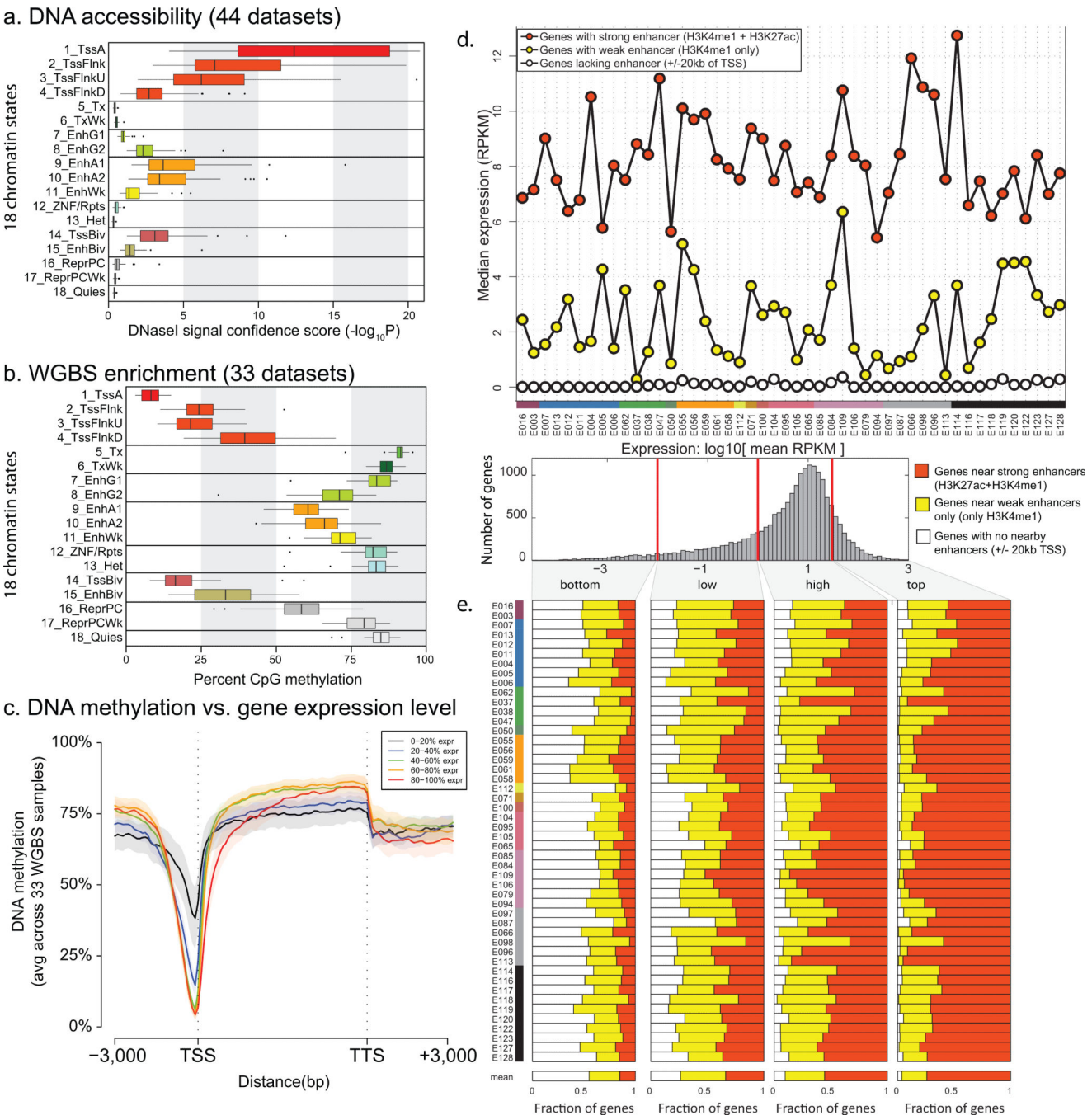
spleen, gonad, and spinal cord), no color is present, indicating that these are not part of the 111 reference epigenomes (ENCODE 2012 samples, or not all five marks in the core set were present), but datasets from these samples are high quality and were sometimes used in companion paper analyses, and are available to the public. **e. Assay correlations.** Heatmap of the pairwise experiment correlations for the core set of five histone modification marks (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3) across all 127 reference epigenomes, the two common acetylation marks (H3K27ac and H3K9ac), and DNA accessibility (DNase) across the reference epigenomes where they are available. Yellow indicates relatively higher correlation and blue lower correlation. Rows and columns were ordered computationally to maximize similarity of neighboring rows and columns (see Methods). All experiments for H3K9me3, H3K27me3, H3K36me3, DNase, and H3K4me1 are consistently ordered into distinct and contiguous groups. For H3K4me3, H3K9ac, and H3K27ac, experiments group primarily based on the mark, but in some cases, the correlations and ordering appear more cell type driven.



Extended Data 2. Chromatin state model robustness and enrichments

a. Chromatin state model robustness. Clustering of 15-state ‘core’ chromatin state model learned jointly across reference epigenomes (Fig. 4a) with chromatin state models learned independently in 111 reference epigenomes. We applied ChromHMM to learn a 15-state ChromHMM model using the five core marks in each of the 111 reference epigenomes generated by the Roadmap Epigenomics program, and clustered the resulting 1680 state emission probability vectors (leaves of the tree) with the 15 states from the joint model (indicated by arrows). We found that the vast majority of states learned across cell types

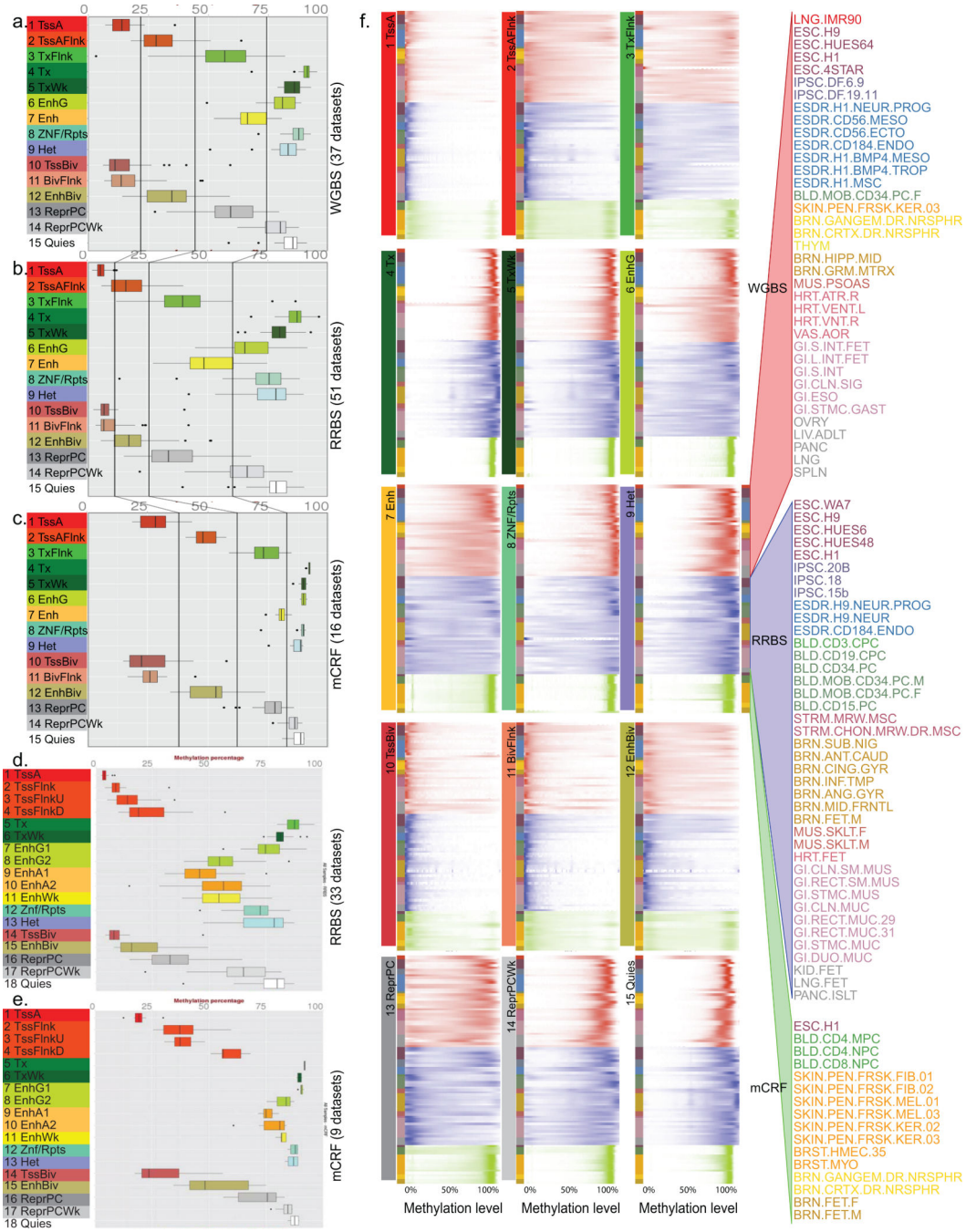
clustered into 15 clusters, corresponding to the joint model states, validating the robustness of chromatin states across cell types. This analysis revealed two new clusters (red crosses) which are not represented in the 15 states of the jointly-learned model: 'HetWk', a cluster showing weak enrichment for H3K9me3; and 'Rpts', a cluster showing H3K9me3 along with a diversity of other marks, and enriched in specific types of repetitive elements (satellite repeats) in each cell type, which may be due to mapping artifacts. This joint clustering also revealed subtle variations in the relative intensity of H3K4me1 in states TxFlnk, Enh, and TssBiv, and H3K27me3 in state TssBiv. Overall, this analysis confirms that the 15-state chromatin state model based on the core set of five marks provides a robust framework for interpreting epigenomic complexity across tissues and cell types. b. Enrichments for 15-state model based on five histone modification marks. Top Left: TF binding site overlap enrichments of 15 states in H1-ESC from the 'core' model for transcription factor binding sites (TFBS) based on ChIP-seq data in H1-ESC. TF binding coverage for other cell-types based on matched TF ChIP-seq data is shown in Fig. S2. Top Right: Enrichments for expressed and non-expressed genes in H1-ESC and GM12878. Bottom: Positional enrichments at the transcription start site (TSS) and transcription end site (TES) of expressed (expr.) and repressed (repr.) genes in H1-ESC. Transition probabilities show frequency of co-occurrence of each pair of chromatin states in neighboring 200-bp bins. d. Definition and enrichments for 18-state 'expanded' model that also includes H3K27ac associated with active enhancer and active promoter regions, but which was only available for 98 of the 127 reference epigenomes. Inclusion of H3K27ac distinguishes active enhancers and active promoters. Top: TFBS enrichments in H1-ESC (E003) chromatin states using ENCODE TF ChIP-seq data in H1-ESC . Bottom: Positional enrichments in H1-ESC for genomic annotations, expressed and repressed genes, TSS and TES, and state transitions as in Extended Data 2b and Fig. 4a-c. Right: Average fold-enrichment (colors bars) and standard deviation (black line) across 98 reference epigenomes (Fig. S3d) for the fold enrichment for non-coding of genomic segments (GERP) in each chromatin state (rows) in the 18-state model. Even after excluding protein-coding exons (see Fig. S3b vs. Fig. S3d), the TSS-proximal states show the highest levels of conservation, followed by EnhBiv and the three non-transcribed enhancer states. In contrast, Tx and TxWk elements are weakly depleted for conserved regions, and Znf/Rpts, and Het are strongly depleted for conserved elements.



Extended Data 3. Relationship between histone marks, DNA methylation, DNA accessibility, and gene expression

a. H3K27ac-marked ‘active’ enhancers show higher levels of DNA accessibility, based on enrichment of DNase-seq signal confidence scores (-log10(Poisson *p*-value))for elements in each chromatin state in our extended 18-state model that includes the core five histone modification marks and H3K27ac, similar to Fig. 4e. **b.** Level of whole-genome bisulfite methylation for all chromatin states in the 18-state model shows that H3K27ac-marked ‘active’ enhancers associated with H3K27ac in addition to H3K4me1 show lower methylation levels, consistent with higher regulatory activity. The whiskers in a. and b. show

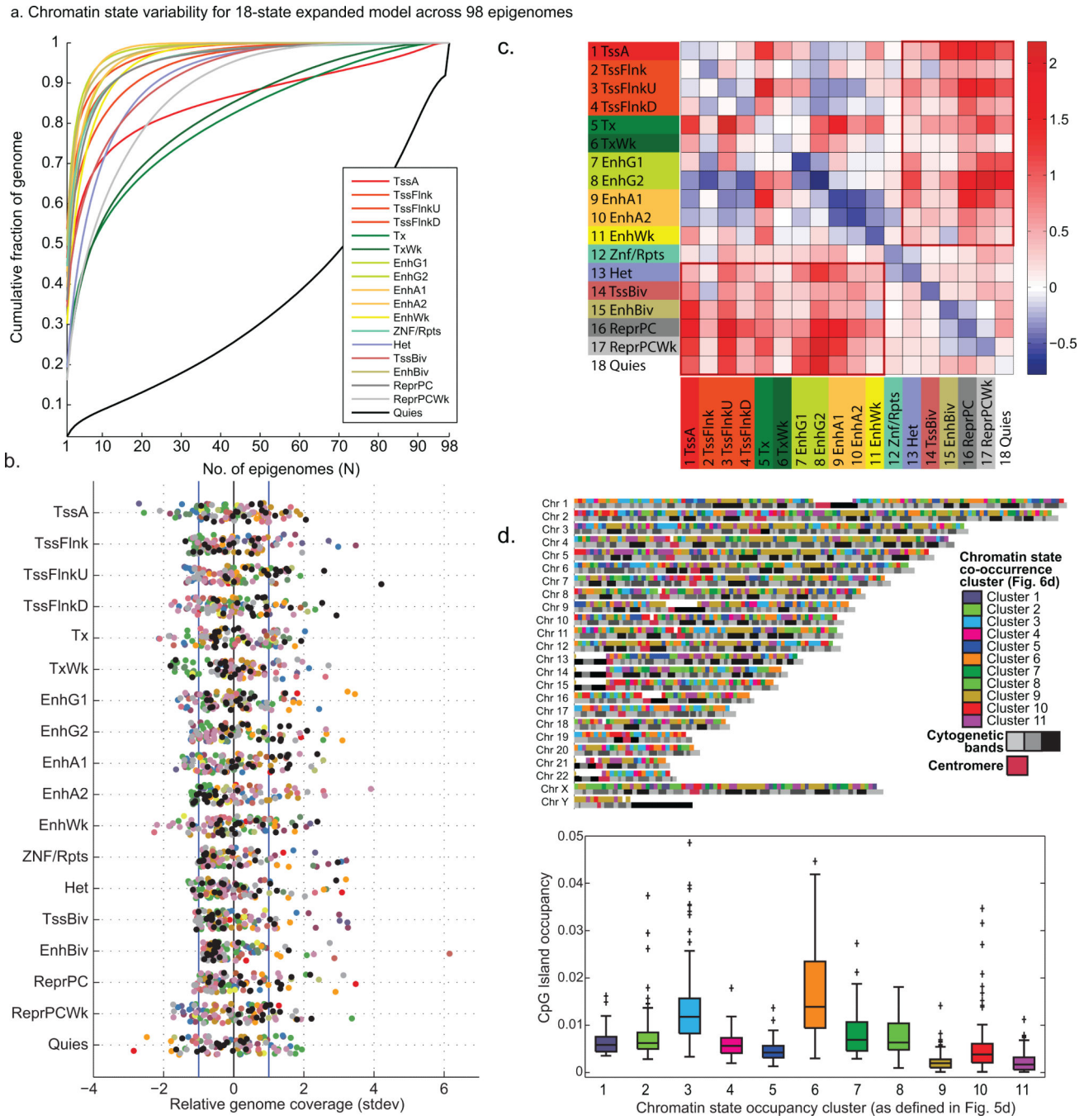
1.5 x IQR (interquartile range) and the filled circles are individual outliers c. DNA methylation levels for genes showing different expression levels. The depletion of DNA methylation in promoter regions, and the enrichment of DNA methylation in transcribed regions, are both more pronounced for highly expressed genes. The enrichment for high DNA methylation is more pronounced in the 3' ends of the most highly expressed genes. d. Genes associated with active enhancer states have consistently significantly higher expression. 'Active enhancer' associated genes have at least one EnhA1 and/or EnhA2 +/ −20Kb from TSS (18-state model). 'Weak-enhancer'genes are associated with EnhG1, EnhG2, EnhWk, EnhBiv. Lowest expression have genes that are not associated with any enhancer. Plots with red markers show median expression of genes associated with 'active' enhancers, yellow markers 'weak' enhancers, and white markers no association with any enhancer state. e. Higher-expression genes show greater association with H3K27ac-marked 'active' enhancers. Highly expressed genes are consistently more frequently associated with H3K27ac-marked active enhancers (EnhA1 and EnhA2) across all cell types. Fraction of genes associated with H3K27ac-marked 'active' enhancers (red), H3K27ac-lacking 'weak' enhancers only (yellow), or no enhancers (white) for genes of varying expression levels in each cell type with RNA-seq data.



Extended Data 4. Methylation relationship with chromatin state

a-c. DNA methylation levels in 15-state model across technologies. We observed significant differences in the average methylation levels observed that were correlated with the different DNA methylation platforms used, but their relative relationships in average chromatin state methylation were conserved. Relative to WGBS (panel a, repeated from Fig. 4d for comparison purposes), RRBS (panel b) showed the lowest overall methylation levels (as expected given its CpG island enrichment), while mCRF showed the highest (panel c). This highlights the importance of recognizing and potentially correcting for DNA methylation

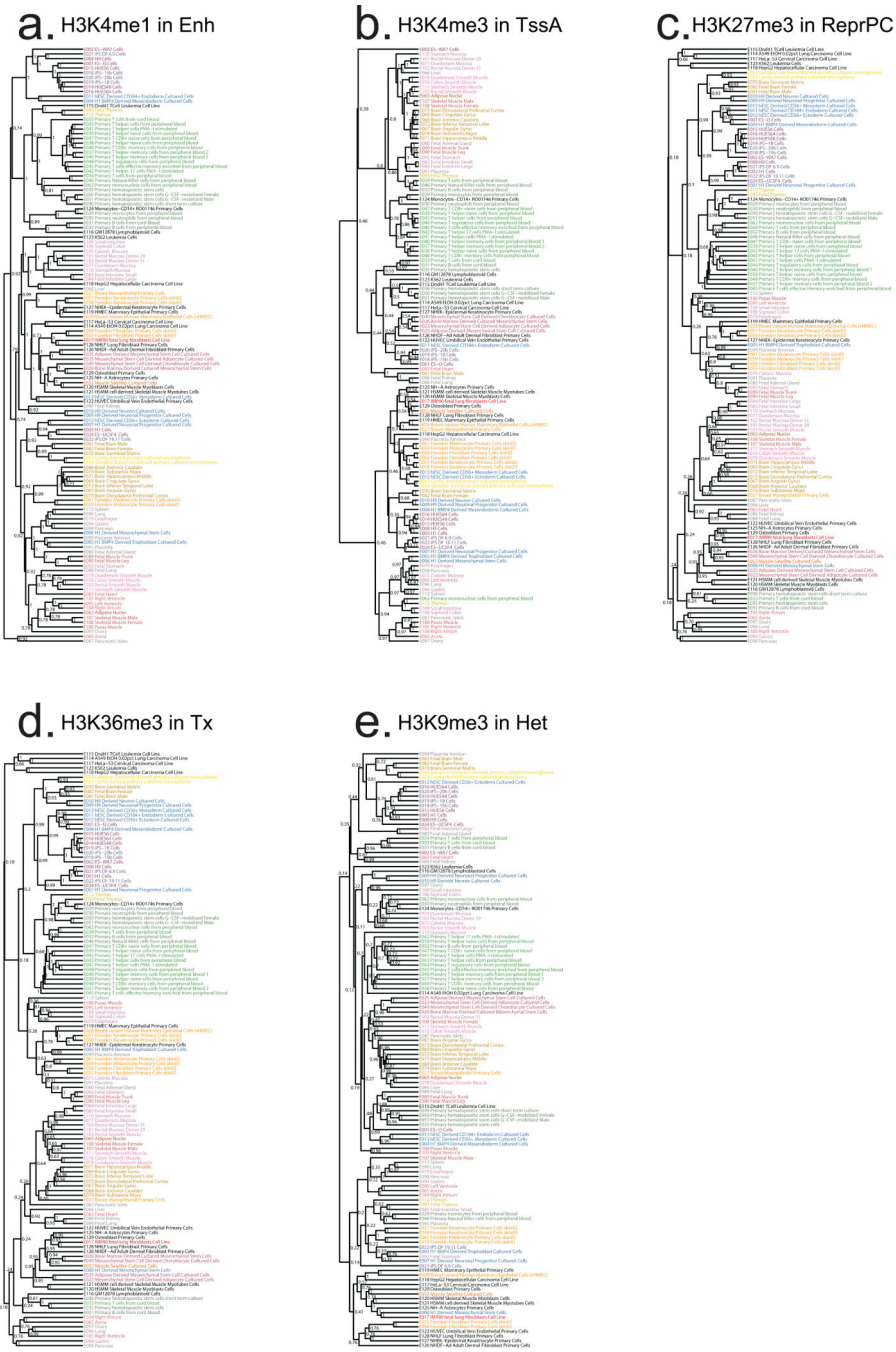
platform specific biases prior performing integrative analyse. **d,e.** Distribution of DNA methylation levels measured using RRBS and mCRF in 18-state model (defined in Extended Data 2c). WGBS is shown in Extended Data 3b. The whiskers in a., b., c., d., and e. show 1.5 x IQR (interquartile range) and the filled circles are individual outliers **f. DNA methylation variation across cell types.** Density plots denote distribution of DNA methylation levels from 0% to 100% for each chromatin state across the 95 reference epigenomes profiled for whole-genome bisulfite (WGBS, red), reduced representation bisulfite (RRBS, blue), or MeDIP/MRE (mCRF, green). The respective color (red, blue, or green) was set to the maximum ln(density+1) value for each chromatin state and respective platform, with intermediate values colored on a natural log scale. For each panel, epigenomes are listed in the same order, shown on the right, with abbreviations of samples in the order of Fig. 2 for each technology.



Extended Data 5. Chromatin state variability, switching, and genomic coverage

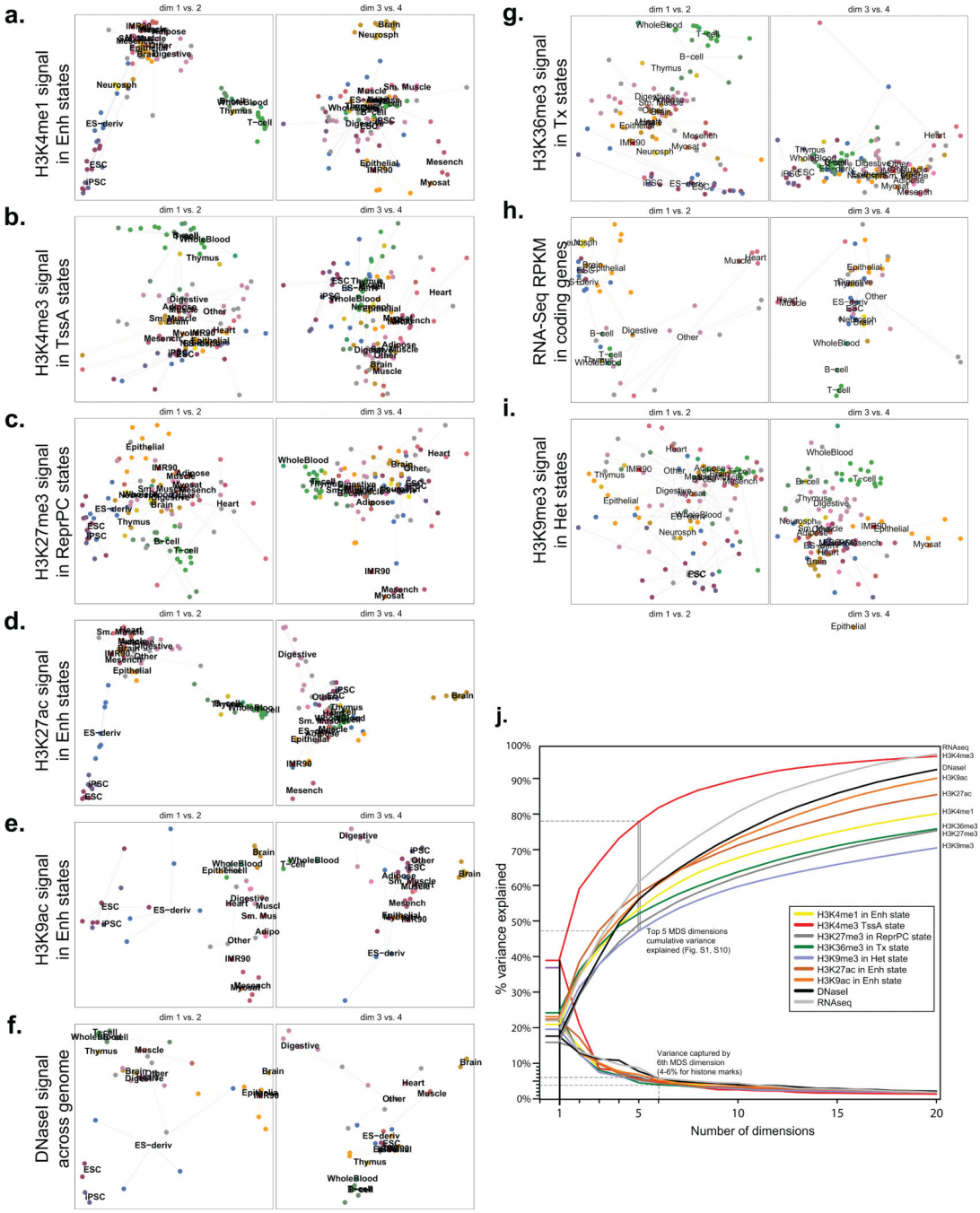
a. Variability level for 18-state model. Chromatin state variability (similar to Fig. 5a), quantified based on the fraction of the genomic coverage (y-axis) of each state (color) that is consistently labeled with that state in at most N (ranging from 1 to 98) reference epigenomes, using the 18-state model learned based on 6 chromatin marks, including H3K27ac. **b. Chromatin state over- and under-representation** for 18-state expanded model. **c.** Log-ratio ($\log_{10}$) of **chromatin state switching probabilities** for the 18-state expanded model across 34 high-quality, non-redundant epigenomes that have H3K27ac

data, relative to intra-tissue switching probabilities across replicates or samples from multiple individuals. d. Chromatin state coverage grouped by epigenomic domains. Top: Chromosome ‘painting’ of 11 clusters shown in Fig. 5d and discovered based on chromatin state co-occurrence at the 2Mb scale across reference epigenomes. Bottom: Enrichment of CpG islands in each cluster clearly showing higher CpG density ‘active’ clusters 3 and 6 comparing to passive clusters 9-11. Each box plot shows a distribution of CpG total occupancy in 2Mb bins in each cluster (with box boundaries indicate 25th and 75th percentiles the whiskers extend to the most extreme datapoints the algorithm considers to not be outliers. Points are drawn as outliers if they are larger than Q3+W*(Q3-Q1) or smaller than Q1-W*(Q3-Q1), where Q1 and Q3 are the 25th and 75th percentiles, respectively.).



Extended Data 6. Hierarchical clustering of epigenomes using diverse marks

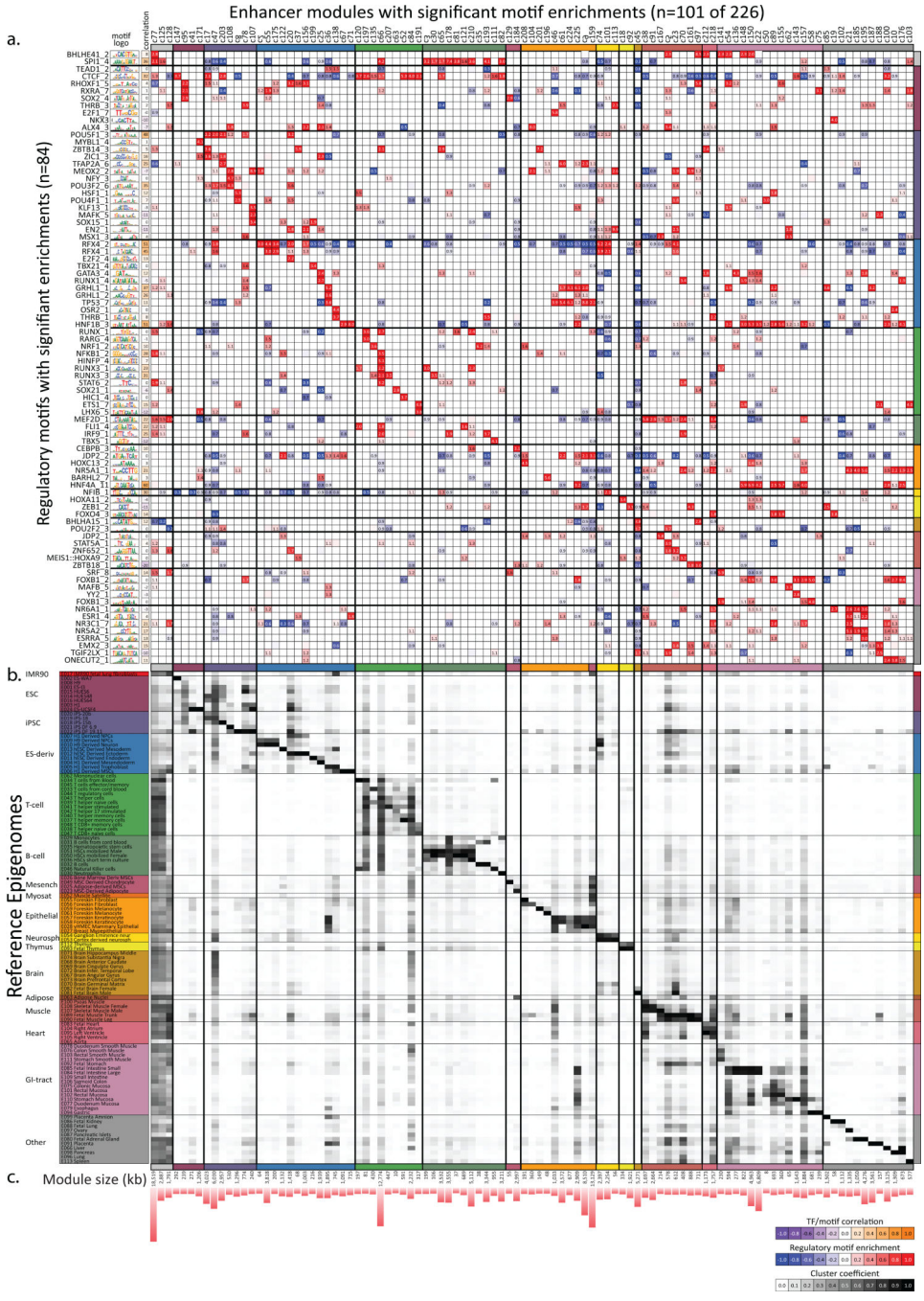
a-e. Clustering of all 127 reference epigenomes, including ENCODE samples, using H3K4me1, H3K4me3, H3K27me3, H3K36me3 and H3K9me3 signal in Enh, TssA, ReprPC, Tx and Het chromatin states, respectively. All panels show hierarchical clustering with optimal leaf ordering. Colors indicate sample groups, as defined in Fig. 2. Numbers on internal nodes represent bootstrap support scores over 1,000 bootstrap samples.



Extended Data 7.

a-i. Multidimensional scaling (MDS) plots showing tissue/cell type similarity using different epigenomic marks. Multi-Dimensional Scaling (MDS) analysis results, showing reference epigenomes using their group coloring defined in Fig. 2. Thin lines connect same-group reference epigenomes. The first 4 axes of variation are shown in pairs. Marks are assessed in regions with relevant chromatin states (see Methods). **j. Variance explained by each MDS dimension.** The first 5 dimensions shown in Fig. S10 (Fig. 6b,c) explain between 45% and 80% of the total epigenometo-epigenome variance for all histone modification mark

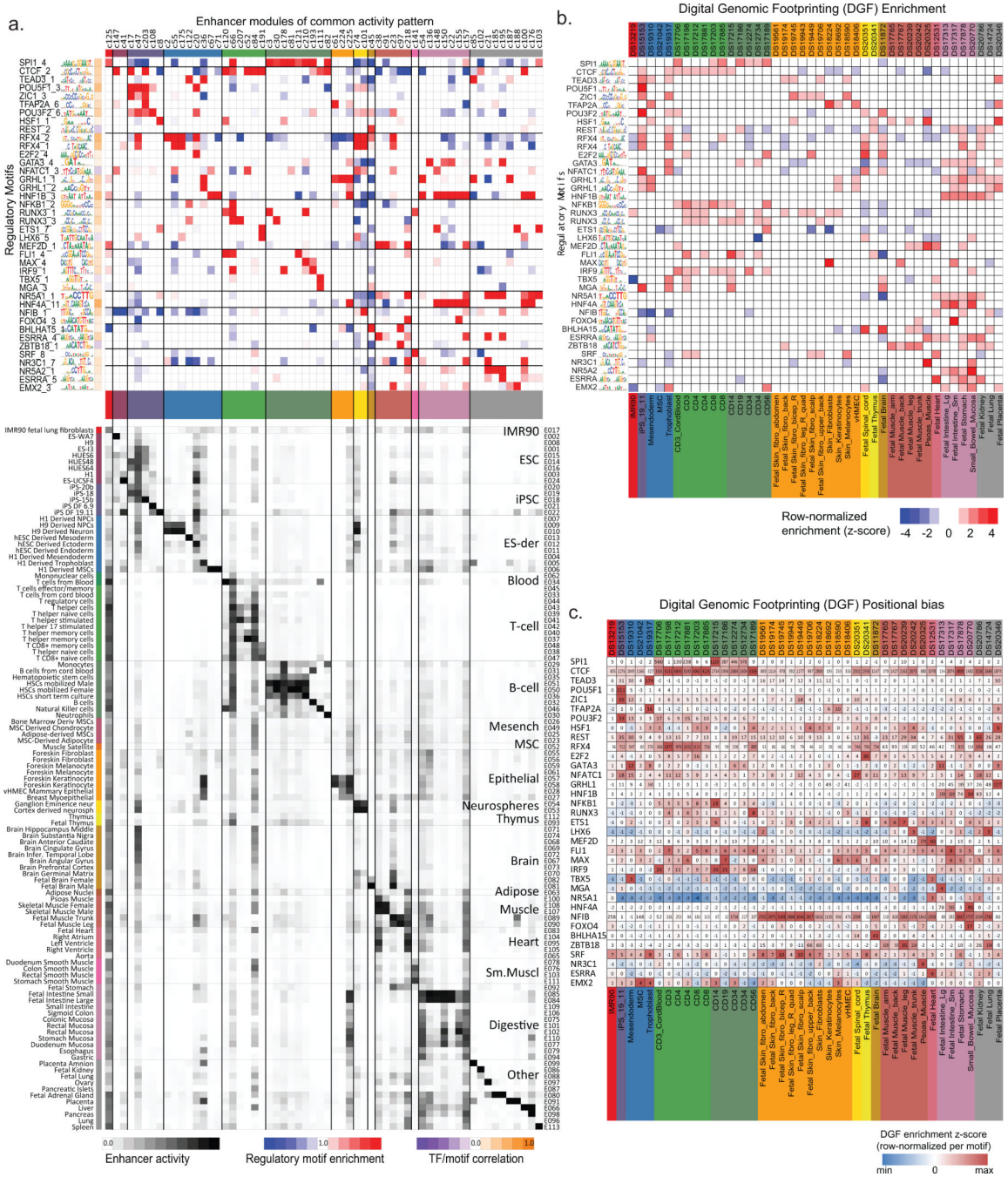
correlations, and additional dimensions explain less than 10%. Only a few components of H3K4me3 in TssA chromatin states explains a much larger fraction of the variance than other marks, possibly due to its stability across cell types.



Extended Data 8.

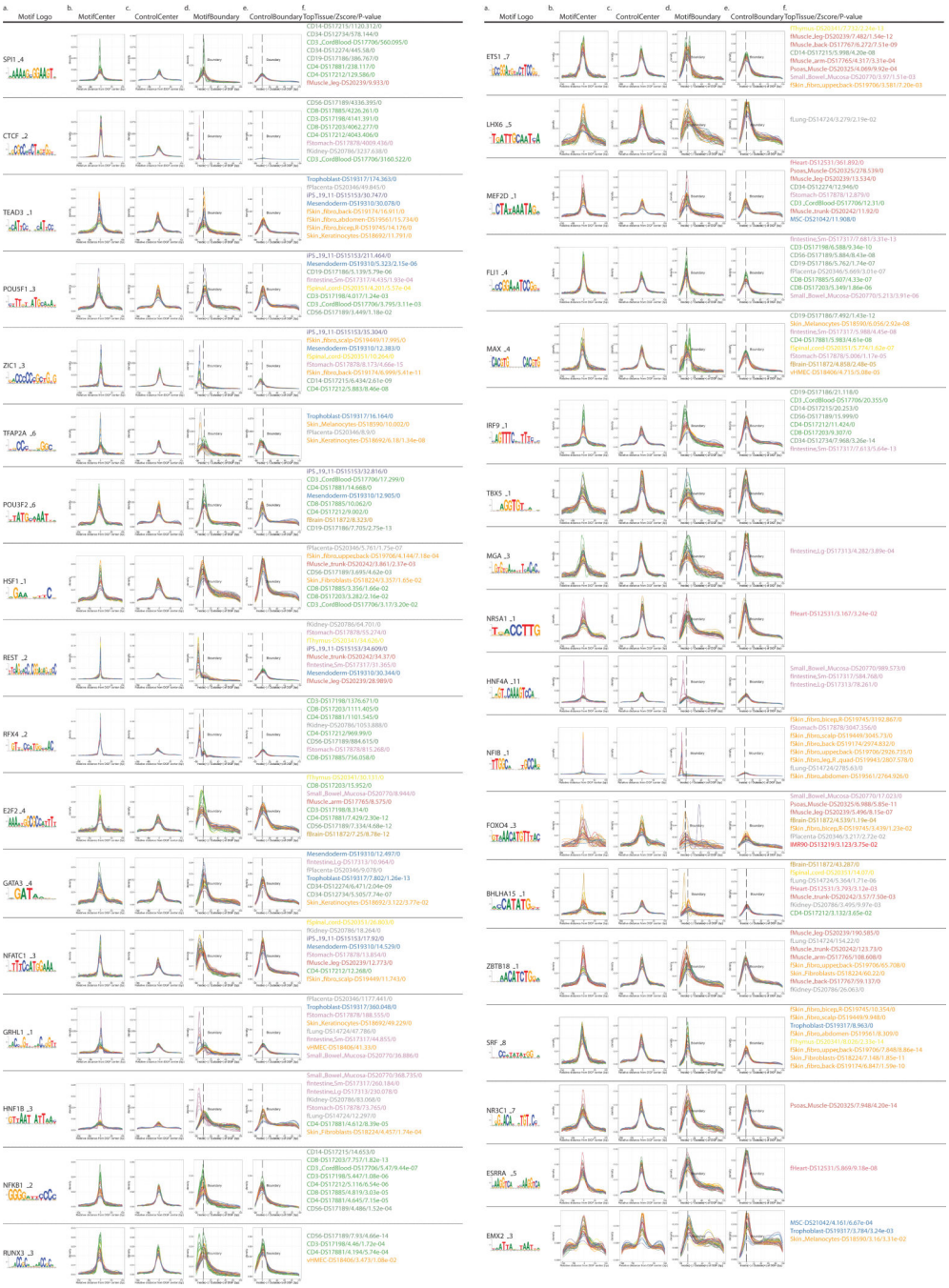
a. Regulatory motifs enriched in clusters. Enrichment (red) or depletion (blue) of regulatory motifs (rows) in the enhancer modules (columns) relative to shuffled control motifs. For each motif is shown the motif name, consensus logo, and correlation between

regulator expression and module activity: positive correlation (orange) is indicative of activators, and negative correlation (purple) indicates a repressive role for the factor. Only clusters with enrichment or depletion of at least 2^1.5-fold for one motif are shown. **b.** Average activity level of enhancers of each module in each reference epigenome (black=high, white=low). Bottom: Total size of each enhancer module showing enrichment (in kb).



Extended Data 9.

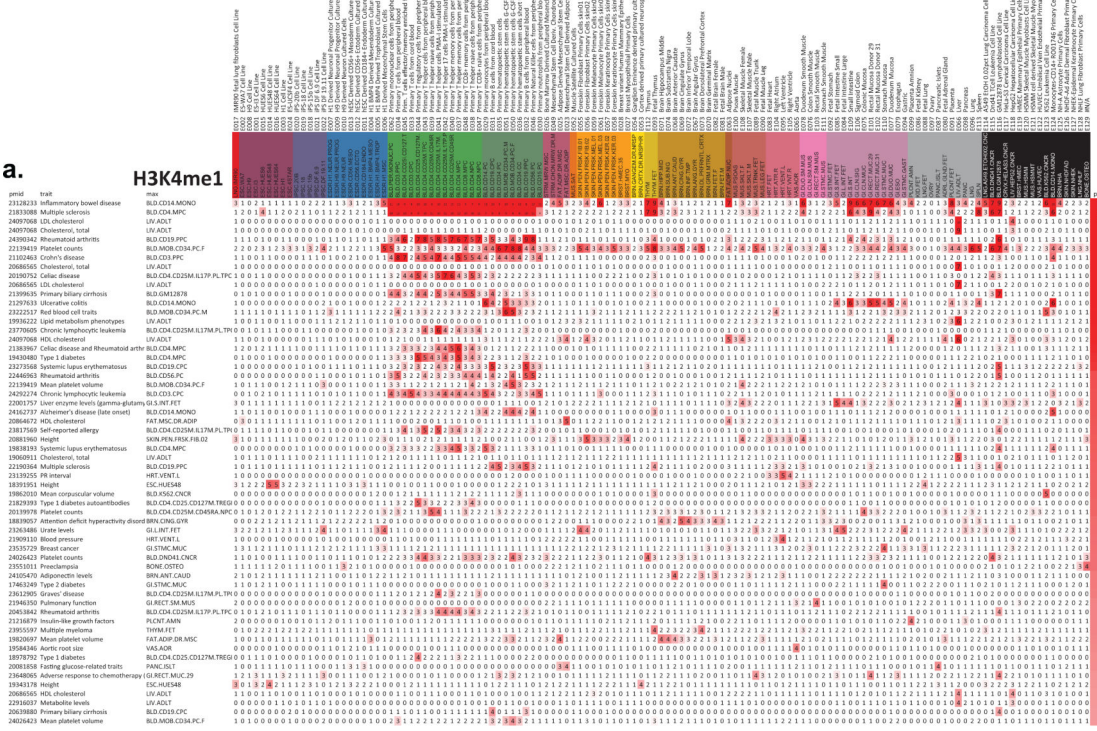
a. Regulatory motif enrichment, DGF enrichment, and positional bias for predicted driver motifs, based on strong (positive or negative) correlations between TF expression and enhancer module activity. **a. Regulatory motif enrichments** for the 40 regulators showing the strongest absolute correlation between TF expression and module activity. Of these, 36 were also recovered solely based on their motif enrichment scores (Extended Data 8), but six were only discovered based on their correlations (Esrra_4, Max_4, Mga_3, Nfatc1_3, Rest_2, and Tead3_1), illustrating the importance of studying motif enrichments in the context of TF expression and enhancer activity patterns. b. Predicted driver regulatory motifs are enriched in high-resolution DNase footprints. Enrichment of predicted driver motif instances (Fig. 8 and Extended Data 9a) in 42 high-resolution (6bp-40bp) Digital Genomic Footprinting (DGF) libraries from deeply sequenced DNase datasets[68] shows consistent tissue preferences in matching cell types. For example, POU5F1 in iPS cells, HNF1B and HNF4A1 in digestive tissues, RFX4 in mesendoderm and neural lineages, MFE2B in muscle. c. Matrix of significant positional bias across factors and cell types. For each Digital Genomics Footprinting (DGF) dataset (columns), positional bias score (heatmap) of predicted driver regulatory motifs (rows) found to be significantly enriched (Fig. 8, Extended Data 9a) in enhancer modules (Fig. 7a).



Extended Data 10. Positional biases of predicted driver motifs relative to high-resolution DNase footprint centers and boundaries

a. Driver TF motif instance logo, as in Fig. 8 and Extended Data 9a. **b.** Distribution of motif instances relative to the center of the high-resolution DNase sites (digital genome footprints, DGF, lengths range from 6bp to 40bp), each curve colored according to the cell/tissue type (from Fig. 2, Table S5b). **c.** Distribution of shuffled motifs that match composition and number of conserved occurrences in the genome[69,72]. **d.** Positional bias relative to boundary of DGF region for true motifs, similar to b. **e.** Positional bias relative to boundary of DGF

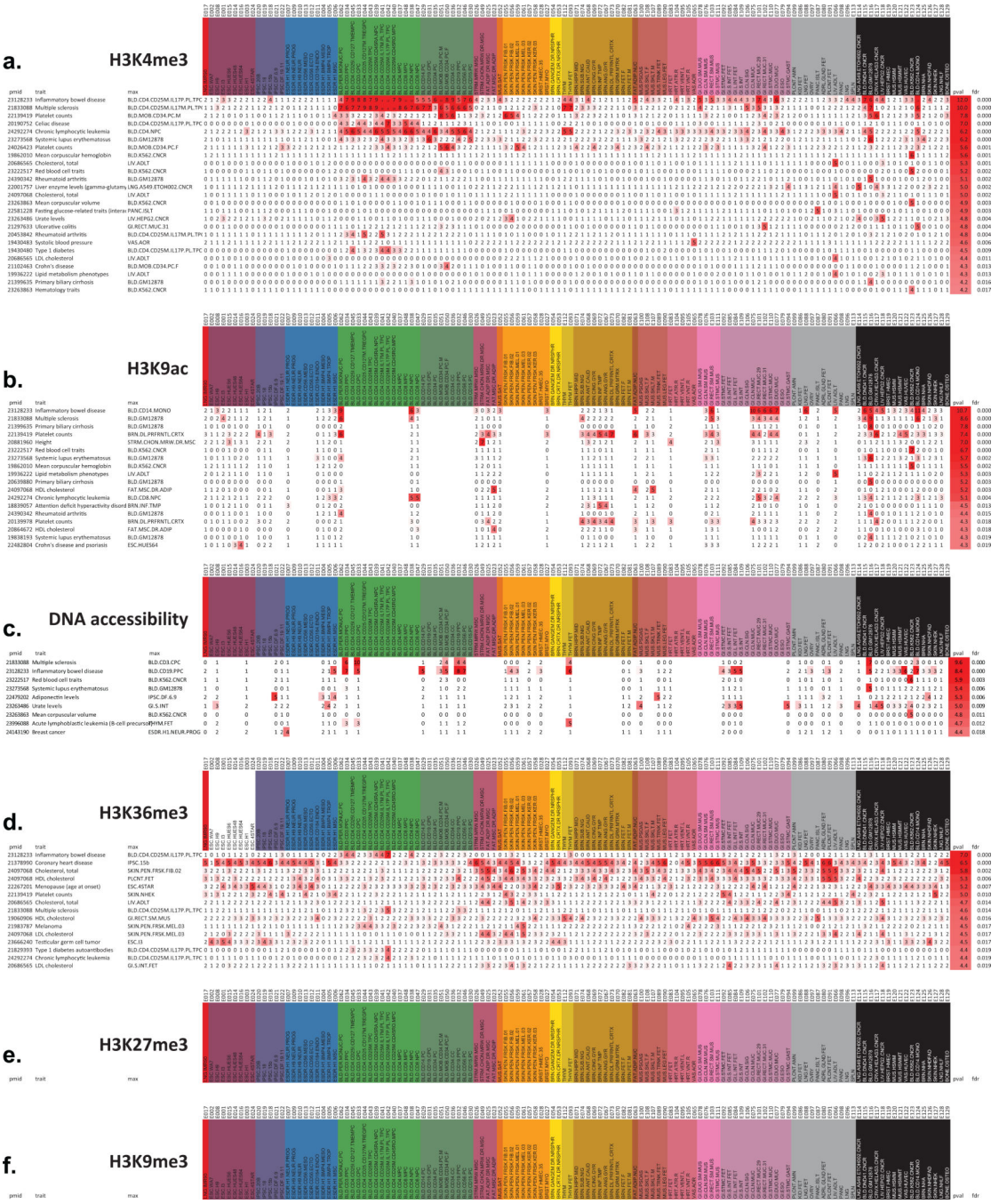
region for shuffled motifs, similar to c. **f.** Cell types showing significant positional bias after multiple testing correction, colored according to Fig. 2 and Table S5b.



Extended Data 11. Epigenomic enrichments of genetic variants associated with diverse traits

Tissue-specific enrichments for peaks of diverse epigenomic marks for genetic variants associated with complex disease, expanding Fig. 9. Enrichments are shown for: **a.** H3K4me1 peaks (enhancers). This panel includes all the data shown in Fig. 9, but expands the enrichments shown to all reference epigenomes (columns), and additional traits (rows)

that did not meet the FDR=0.02 threshold. **b.** H3K27ac peaks (active enhancers). **a-b.** Studies were defined by a set of SNPs annotated in the GWAS catalog with the same combination of a trait (far left column) and publication shown by the Pubmed ID (far right column), uncorrected p-value (in -$\log_{10}$), and estimated FDR.



Extended Data 12. Epigenomic enrichments of genetic variants associated with diverse traits

Tissue-specific enrichments for peaks of diverse epigenomic marks for genetic variants associated with complex disease, expanding Fig. 9. Enrichments are shown for: **a.**

H3K4me3 peaks (promoters). **b.** H3K9ac peaks (active promoters and active enhancers). **c.** DNase peaks (accessible regions). **d.** H3K36me3 peaks (transcribed regions). **e.** H3K27me3 peaks (Polycomb-repressed regions). **f.** H3K9me3 peaks (heterochromatin regions). **a-f.** Studies were defined by a set of SNPs annotated in the GWAS catalog with the same combination of a trait (far left column) and publication shown by the Pubmed ID (far right column), uncorrected p-value (in -$\log_{10}$), and estimated FDR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Roadmap Epigenomics Consortium[†], Anshul Kundaje[1,2,3,†], Wouter Meuleman[1,2,†], Jason Ernst[1,2,4,†], Misha Bilenky[5,†], Angela Yen[1,2,†], Pouya Kheradpour[1,2,†], Zhizhuo Zhang[1,2,†], Alireza Heravi-Moussavi[5,†], Yaping Liu[1,2,†], Viren Amin[6,†], Michael J Ziller[2,7,†], John W Whitaker[8,†], Matthew D Schultz[9,†], Richard S Sandstrom[10,†], Matthew L Eaton[1,2,†], Yi-Chieh Wu[1,2,†], Jianrong Wang[1,2,†], Lucas D Ward[1,2,†], Abhishek Sarkar[1,2,†], Gerald Quon[1,2,†], Andreas Pfenning[1,2,†], Xinchen Wang[11,1,2,†], Melina Claussnitzer[1,2,†], Cristian Coarfa[6,‡], R Alan Harris[6,‡], Noam Shoresh[2,‡], Charles B Epstein[2,‡], Elizabeta Gjoneska[12,2,‡], Danny Leung[8,‡], Wei Xie[8,‡], R David Hawkins[8,‡], Ryan Lister[9,‡], Chibo Hong[13,‡], Philippe Gascard[14,‡], Andrew J Mungall[5,‡], Richard Moore[5,‡], Eric Chuah[5,‡], Angela Tam[5,‡], Theresa K Canfield[10,‡], R Scott Hansen[10,‡], Rajinder Kaul[15,‡], Peter J Sabo[10,‡], Mukul S Bansal[1,2,16], Annaick Carles[17], Jesse R Dixon[8], Kai-How Farh[2], Soheil Feizi[1,2], Rosa Karlic[18], Ah-Ram Kim[1,2], Ashwinikumar Kulkarni[19], Daofeng Li[20], Rebecca Lowdon[20], Tim R Mercer[21], Shane J Neph[10], Vitor Onuchic[6], Paz Polak[2,22], Nisha Rajagopal[8], Pradipta Ray[19], Richard C Sallari[1,2], Kyle T Siebenthall[10], Nicholas Sinnott-Armstrong[1,2], Michael Stevens[20], Robert E Thurman[10], Jie Wu[23,24], Bo Zhang[20], Xin Zhou[20], Arthur E Beaudet[47], Laurie A Boyer[11], Philip De Jager[34], Peggy J Farnham[35], Susan J Fisher[31], David Haussler[28], Steven Jones[5,48], Wei Li[49], Marco Marra[5,17], Michael T McManus[41], Shamil Sunyaev[2,22,34], James A Thomson[27], Thea D Tlsty[14], Li-Huei Tsai[12,2], Wei Wang[8], Robert A Waterland[50], Michael Zhang[19], Lisa H Chadwick[51,¥], Bradley E Bernstein[2,43,25,¥], Joseph F Costello[13,¥], Joseph R Ecker[9,¥], Martin Hirst[5,17,¥], Alexander Meissner[2,¥], Aleksandar Milosavljevic[6,¥], Bing Ren[8,¥], John A Stamatoyannopoulos[10,¥], Ting Wang[20,¥], and Manolis Kellis[1,2,¥]

Affiliations

[1] Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge MA 02139, USA

[2] The Broad Institute of Harvard and MIT, 415 Main Street, Cambridge MA 02142, USA.

[3] Department of Genetics, Department of Computer Science, 300 Pasteur Dr., Lane Building, L301, Stanford, CA 94305-5120.

[4] Department of Biological Chemistry, University of California, Los Angeles, 615 Charles E Young Dr South, Los Angeles, CA 90095, USA.

[5] BC Cancer Agency, Canada's Micheal Smith Genome Sciences Centre, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada.

[6] Epigenome Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[7] Department of Stem Cell and Regenerative Biology, 7 Divinity Ave, Cambridge, MA 02138, USA.

[8] Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, Moores Cancer Center, Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

[9] Genomic Analysis Laboratory, Howard Hughes Medical Institute & The Salk Institute for Biological Studies, 10010 N.Torrey Pines Road, La Jolla, CA 92037, USA.

[10] Department of Genome Sciences, University of Washington, 1705 NE Pacific Street Seattle WA 98195 USA 206-267-1091, USA.

[11] Biology Department, Massachusetts Institute of Technology, 31 Ames St, Cambridge, MA 02142, USA.

[12] Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 43 Vassar St, Cambridge, MA 02139, USA.

[13] Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, 1450 3rd Street, San Francisco, CA, 94158, USA.

[14] Department of Pathology, University of California San Francisco, 513 Parnassus Avenue, San Francisco CA 94143-0511, USA.

[15] Department of Medicine, Division of Medical Genetics, University of Washington, 2211 Elliot Avenue Seattle WA 98121 USA 206-267-1091, USA.

[16] Department of Computer Science & Engineering, University of Connecticut , 371 Fairfield Way, Storrs CT 06269, USA.

[17] Department of Microbiology and Immunology and Centre for High-Throughput Biology, University of British Columbia, 2125 East Mall, Vancouver, BC, V6T 1Z4, Canada.

[18] Bioinformatics Group, Division of Biology, Faculty of Science, Zagreb University, Horvatovac 102a, 10000 Zagreb, Croatia.

[19] Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas, Dallas, NSERL, RL10, 800 W Campbell Road, Richardson, TX 75080, USA.

[20] Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, 4444 Forest Park Ave. Saint Louis, MO 63108, USA.

[21] Institute for Molecular Bioscience, University of Queensland, St Lucia Queensland 4072, Australia.

[22] Brigham & Women's Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.

[23] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA.

[24] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

[25] Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, USA.

[26] University of Virginia, School of Medicine, 1340 Jefferson Park Ave. Charlottesville, VA, 22908, USA.

[27] Morgridge Institute for Research, 330 N. Orchard St., Madison, WI 53707, USA, USA.

[28] Center for Biomolecular Sciences and Engineering, University of Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA.

[29] UCSF School of Medicine, 513 Parnassus Avenue, San Francisco CA 94143, USA.

[30] Rush University Medical Center, 1653 W Congress Pkwy, Chicago, IL 60612, USA.

[31] OB/GYN & Reproductive Sciences, University of California San Francisco, 35 Medical Center Way, San Fancisco, CA 94143, USA.

[32] Rikshospitalet University Hospital, Sognsvannsveien 20, 0372 Oslo, Norway.

[33] Reproductive Endocrinology and Infertility, University of California San Francisco, 2356 Sutter St, San Francisco, CA, 94115, USA.

[34] Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA.

[35] Department of Biochemistry, Keck School of Medicine, University of Southern California, 1450 Biggy Street, Los Angeles, CA 90089-9601, USA.

[36] Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA 92093, USA, USA.

[37] Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, USA.

[38] Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. North Seattle WA 98109 U 206-667-4004, USA.

[39] Department of Pediatrics, Seattle Children's Hospital/University of Washington, 4800 Sand Point Way NE Seattle WA 98105 USA, USA.

[40] Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA.

[41] Department of Microbiology and Immunology, Diabetes Center, University of California, San Francisco, 513 Parnassus Ave., San Francisco, CA 94143-0534, USA.

[42] School of Medicine, University of California San Francisco, 513 Parnassus Avenue, San Francisco CA 94143, USA.

[43] Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815-6789, USA.

[44] Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute/Brigham and Women's Hospital, 450 Brookline Avenue, Boston, MA 02215, USA.

[45] Vincent's Clinical School, University of New South Wales, Level 2, ASGM Building/Botany St, Sydney NSW 2052, Australia.

[46] Immunology Research Program, Benaroya Research Institute, 1201 Ninth Avenue Seattle WA 98101 USA, USA.

[47] Molecular and Human Genetics Department, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[48] Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada.

[49] Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

[50] USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030, USA.

[51] National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, N.C. 27709, USA.

Acknowledgements

This work was supported by the NIH Common Fund as part of the NIH Roadmap Epigenomics Program through U01ES017155 (BB/AM), U01ES017154 (JC/MM), U01ES017166 (BR), U01ES017156 (JS), U01DA025956 (AM/AB), and by NHGRI through RC1HG005334, R01HG004037 (MK), RO1NS078839 (L-HT). This work was also supported by NIH fellowship grants F32HL110473 and K99HL119617 (S.L.), and NSF CAREER award 1254200 (J.E.). We acknowledge program leadership by members of the NIH Epigenomics Workgroup, especially John S. Satterlee, Frederick L. Tyson, Joni Rutter, Kimberly A. McAllister, Astrid Haugen, Christine Colvis (NCATS), James Battey (NIDCD), Linda Birnbaum (NIEHS), and Nora Volkow (NIDA). We acknowledge feedback from our External Scientific Panel members Marisa Bartolomei, Stephen Baylin, Stephan Beck, Aravinda Chakravarti, Laurie Jackson-Grusby, Jason Lieb, Steve Peckman, John Quackenbush, and Steve Stice. Sample procurement was supported by grants 5R24HD000836 (IAG) for staged fetal tissues; P30AG10161, R01AG15819, R01AG17917 (DAB) and U01AG46152 (PLD, DAB) for adult brain samples.

References

1. Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013; 155:39–55. [PubMed: 24074860]
2. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011; 12:7–18. [PubMed: 21116306]
3. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012; 13:484–492. [PubMed: 22641018]

4. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013; 14:204–220. [PubMed: 23400093]
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
6. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009; 10:669–680. [PubMed: 19736561]
7. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]
8. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
9. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. [PubMed: 21441907]
10. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]
11. Xie W, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013; 153:1134–1148. [PubMed: 23664764]
12. Zhu J, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell. 2013; 152:642–654. [PubMed: 23333102]
13. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489:83–90. [PubMed: 22955618]
14. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]
15. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]
16. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage- committed cells. Nature. 2007; 448:553–560. [PubMed: 17603471]
17. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
18. John S, et al. Genome-scale mapping of DNase I hypersensitivity. Curr Protoc Mol Biol. 2013 Chapter 27, Unit 21 27.
19. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]
20. Meissner A, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005; 33:5868–5877. [PubMed: 16224102]
21. Weber M, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005; 37:853–862. [PubMed: 16007088]
22. Maunakea AK, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010; 466:253–257. [PubMed: 20613842]
23. ENCODE_Project_Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
24. Bernstein BE, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 2005; 120:169–181. [PubMed: 15680324]
25. Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. Science. 2010; 330:612–616. [PubMed: 21030644]
26. Peters AH, et al. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell. 2003; 12:1577–1589. [PubMed: 14690609]
27. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–112. [PubMed: 19295514]
28. Rada-Iglesias A, et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011; 470:279–283. [PubMed: 21160473]
29. Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010; 107:21931–21936. [PubMed: 21106759]

30. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009; 10:295–304. [PubMed: 19308066]
31. Stevens M, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Genome Res. 2013; 23:1541–1553. [PubMed: 23804401]
32. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
33. Butterfield YS, et al. JAGuaR: Junction Alignments to Genome for RNA-Seq Reads. PLoS One. 2014; 9:e102398. [PubMed: 25062255]
34. Coarfa C, et al. Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics. 2010; 11:572. [PubMed: 21092284]
35. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
36. Fejes AP, et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008; 24:1729–1730. [PubMed: 18599518]
37. Landt SG, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–1831. [PubMed: 22955991]
38. Kunde-Ramamoorthy G, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res. 2014; 42:e43. [PubMed: 24391148]
39. Harris RA, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol. 2010; 28:1097–1105. [PubMed: 20852635]
40. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
41. Davydov EV, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010; 6:e1001025. [PubMed: 21152010]
42. Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2010; 467:430–435. [PubMed: 20720539]
43. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480:490–495. [PubMed: 22170606]
44. Gascard P, et al. Epigenetic and transcriptional determinants of mammary gland development. Companion Manuscript. 2015
45. Mohn F, Weber M, Schubeler D, Roloff TC. Methylated DNA immunoprecipitation (MeDIP). Methods Mol Biol. 2009; 507:55–64. [PubMed: 18987806]
46. Elliott G, et al. Intermediate DNA Methylation is a Conserved Signature of Genome Regulation. Companion Manuscript. 2015
47. Ji H, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010; 467:338–342. [PubMed: 20720541]
48. Meissner A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454:766–770. [PubMed: 18600261]
49. Gifford CA, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell. 2013; 153:1149–1163. [PubMed: 23664763]
50. Ziller MJ, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013; 500:477–481. [PubMed: 23925113]
51. Tsankov AM, et al. Modular and context dependent rewiring of transcription factor networks during human ESC differentiation. Companion Manuscript. 2015
52. Ziller MJ, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature. 2014
53. Xie M, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013; 45:836–841. [PubMed: 23708189]
54. McLean CY, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28:495–501. [PubMed: 20436461]

55. Lowdon RF, et al. Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. Nat Commun. 2014; 5:5442. [PubMed: 25421844]
56. Amin V, et al. Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. Nature Communications. 2015
57. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–326. [PubMed: 16630819]
58. Hawkins RD, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell. 2010; 6:479–491. [PubMed: 20452322]
59. Varley KE, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013; 23:555–567. [PubMed: 23325432]
60. Leung D, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Companion Manuscript. 2015
61. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
62. Meuleman W, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013; 23:270–280. [PubMed: 23124521]
63. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–951. [PubMed: 18463634]
64. Antequera F, Boyes J, Bird A. High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. Cell. 1990; 62:503–514. [PubMed: 1974172]
65. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
66. Kohler S, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014; 42:D966–974. [PubMed: 24217912]
67. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014; 42:2976–2987. [PubMed: 24335146]
68. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009; 6:283–289. [PubMed: 19305407]
69. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 2007; 17:1919–1931. [PubMed: 17989251]
70. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2014
71. Dixon JR, et al. Global Reorganization of Chromatin Architecture during Embryonic Stem Cell Differentiation. Companion Manuscript. 2015
72. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
73. Trynka G, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013; 45:124–130. [PubMed: 23263488]
74. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–1006. [PubMed: 24316577]
75. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–1125. [PubMed: 21102463]
76. Cooper JD, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat Genet. 2008; 40:1399–1401. [PubMed: 18978792]
77. Berndt SI, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet. 2013; 45:868–876. [PubMed: 23770605]
78. Stahl EA, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42:508–514. [PubMed: 20453842]
79. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009; 41:703–707. [PubMed: 19430480]
80. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]

81. Yang W, et al. Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. Am J Hum Genet. 2013; 92:41–51. [PubMed: 23273568]
82. Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466:714–719. [PubMed: 20686566]
83. Willy PJ, et al. LXR, a nuclear receptor that defines a distinct retinoid response pathway. Genes Dev. 1995; 9:1033–1045. [PubMed: 7744246]
84. Pasquali L, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet. 2014; 46:136–143. [PubMed: 24413736]
85. Dalcik H, et al. Expression of insulin-like growth factor in the placenta of intrauterine growth-retarded human fetuses. Acta Histochem. 2001; 103:195–207. [PubMed: 11368100]
86. Lesch KP, et al. Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. J Neural Transm. 2008; 115:1573–1585. [PubMed: 18839057]
87. Repunte-Canonigo V, et al. A potential role for adiponectin receptor 2 (AdipoR2) in the regulation of alcohol intake. Brain Res. 2010; 1339:11–17. [PubMed: 20380822]
88. Sawcer S, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476:214–219. [PubMed: 21833088]
89. Heneka MT, Kummer MP, Latz E. Innate immune activation in neurodegenerative disease. Nat Rev Immunol. 2014; 14:463–477. [PubMed: 24962261]
90. Gjoneska E, Pfenning AR, Kundaje A, Tsai L-H, Kellis M. Conserved epigenomic signatures between mouse and human elucidate immune basis of Alzheimer's disease. Nature, Companion Manuscript. 2015
91. Zhou X, et al. Epigenomic annotation of genetic variants using the Roadmap EpiGenome Browser. Nat Biotechnol. 2015
92. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–934. [PubMed: 22064851]
93. Satterlee JS, Schubeler D, Ng HH. Tackling the epigenome: challenges and opportunities for collaboration. Nat Biotechnol. 2010; 28:1039–1044. [PubMed: 20944594]
94. Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014
95. Seumois G, et al. Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol. 2014; 15:777–788. [PubMed: 24997565]
96. De Jager PL, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014; 17:1156–1163. [PubMed: 25129075]
97. Lunnon K, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat Neurosci. 2014; 17:1164–1170. [PubMed: 25129077]
98. Polak P, et al. Cell type of origin chromatin organization shapes the mutational landscape of cancer. Companion Manuscript. 2015
99. Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. Nat Commun. 2014; 5:5114. [PubMed: 25268989]
100. Zhou X, et al. The Human Epigenome Browser at Washington University. Nat Methods. 2011; 8:989–990. [PubMed: 22127213]
101. Karolchik D, et al. The UCSC Genome Browser Database. Nucleic Acids Res. 2003; 31:51–54. [PubMed: 12519945]
102. Chadwick LH. The NIH Roadmap Epigenomics Program data resource. Epigenomics. 2012; 4:317–324. [PubMed: 22690667]
103. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011; 43:264–268. [PubMed: 21258342]
104. Ernst J, Kellis M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. 2013; 23:1142–1154. [PubMed: 23595227]

105. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012; 9:215–216. [PubMed: 22373907]
106. Dixon JR, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–380. [PubMed: 22495300]
107. Lister R, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013; 341:1237905. [PubMed: 23828890]
108. Schultz MD, Schmitz RJ, Ecker JR. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends Genet. 2012; 28:583–585. [PubMed: 23131467]
109. Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics. 2001; 17(Suppl 1):S22–29. [PubMed: 11472989]
110. Leisch, F. A toolbox for KK-centroids cluster analysis. Computational Statistics and Data Analysis. 2006. http://dx.doi.org/10.1016/j.csda.2005.10.006
111. Matys V, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003; 31:374–378. [PubMed: 12520026]
112. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004; 32:D91–94. [PubMed: 14681366]
113. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]
114. Berger MF, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008; 133:1266–1276. [PubMed: 18585359]
115. Jolma A, et al. DNA-binding specificities of human transcription factors. Cell. 2013; 152:327–339. [PubMed: 23332764]
116. Badis G, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324:1720–1723. [PubMed: 19443739]
117. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]
118. Karolchik D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004; 32:D493–496. [PubMed: 14681465]
119. Garber M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–62. [PubMed: 19478016]
120. Osborne JD, et al. Annotating the human genome with Disease Ontology. BMC Genomics. 2009; 10(Suppl 1):S6. [PubMed: 19594883]
121. Hill DP, et al. The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Res. 2004; 32:D568–571. [PubMed: 14681482]

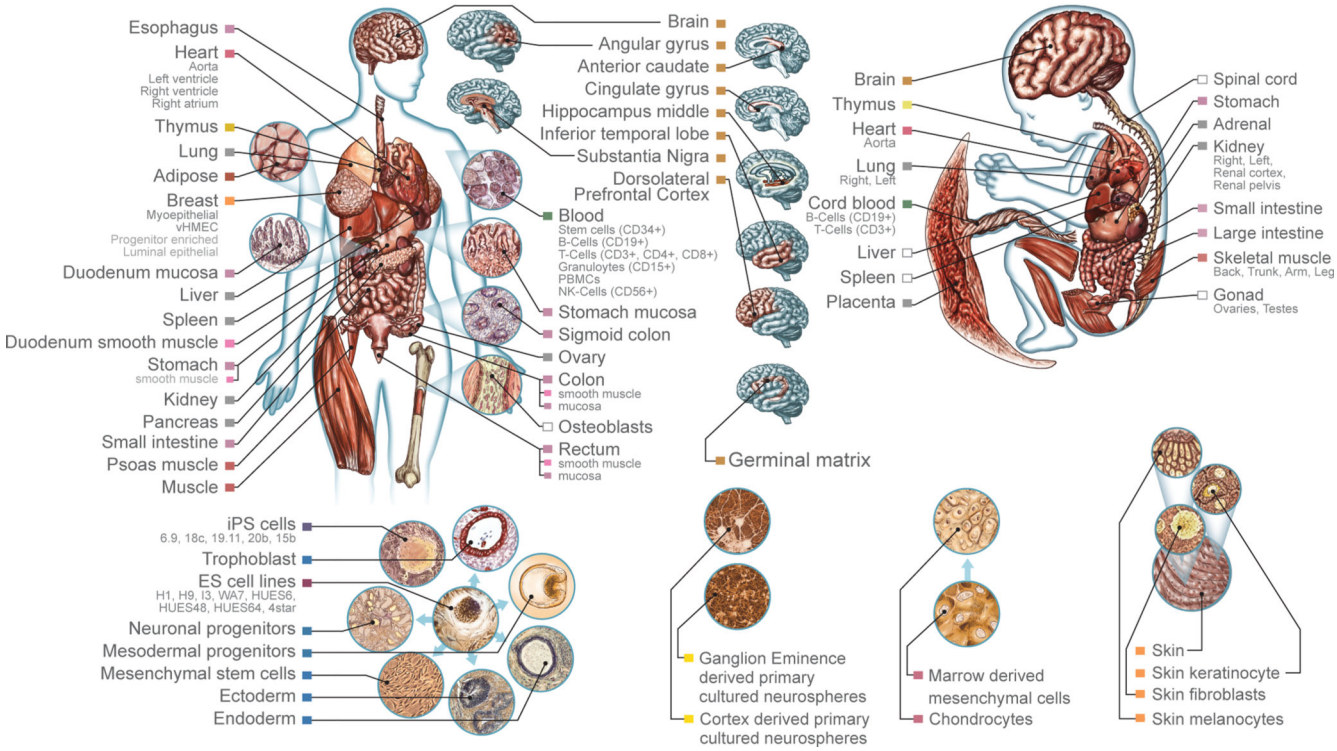


Figure 1. Tissues and cell types profiled in the Roadmap Epigenomics Consortium

Primary tissues and cell types representative of all major lineages in the human body were profiled, including multiple brain, heart, muscle, GI-tract, adipose, skin, and reproductive samples, as well as immune lineages, ESCs and induced Pluripotent Stem (iPS) cells, and differentiated lineages derived from ESCs. Box colors match groups shown in Fig. 2b. Epigenome identifiers (EIDs, Fig. 2c) for each sample shown in Extended Data 1.

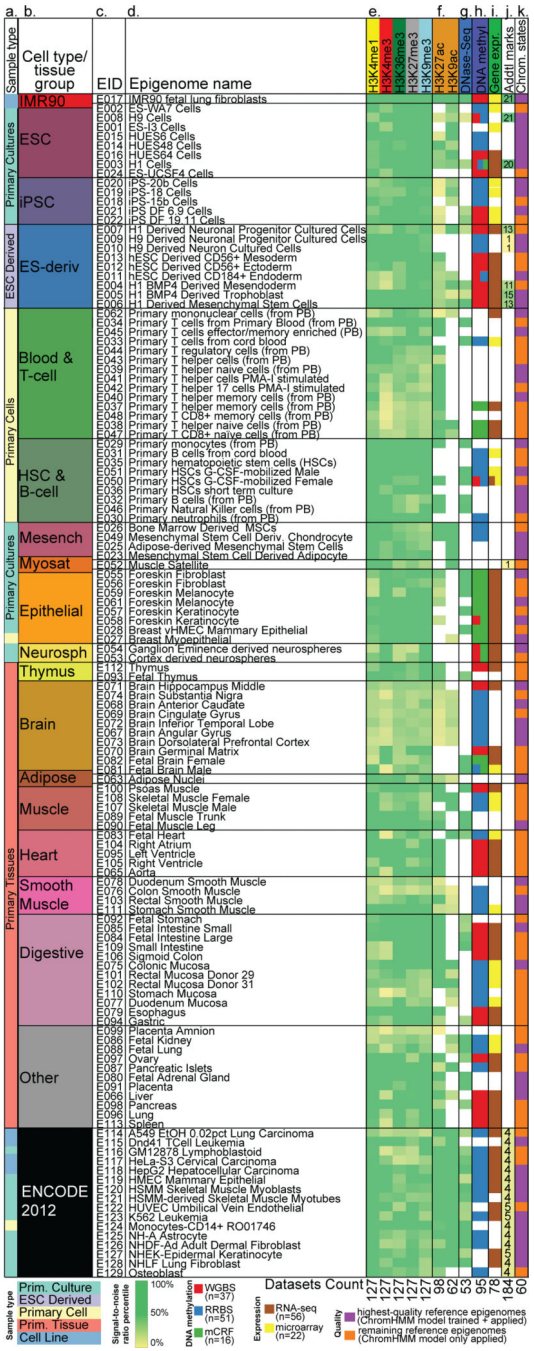


Figure 2. Datasets available for each reference epigenome

List of 127 epigenomes including 111 by the Roadmap Epigenomics program (E001-E113) and 16 by ENCODE (E114-E129). Full list of names and quality scores in Table S1. **a-d:** Tissue and cell types grouped by type of biological material (a), anatomical location (b), showing reference epigenome identifier (EID, c), and abbreviated name (d). PB=Peripheral Blood. ENCODE 2012 reference epigenomes shown separately. **e-g.** Normalized strand cross-correlation quality scores (NSC)[37] for the core set of five histone marks (e), additional acetylation marks (f) and DNase-seq (g). **h.** Methylation data by WGBS (red), RRBS (blue),

and mCRF (green). 104 methylation datasets available in 95 distinct reference epigenomes. **i.** Gene expression data using RNA-seq (Brown) and microarray expression (Yellow). **j.** 26 epigenomes contain a total of 184 additional histone modification marks. **k.** 60 highest-quality epigenomes (purple) were used for training the core chromatin state model, which was then applied to the full set of epigenomes (purple and orange).

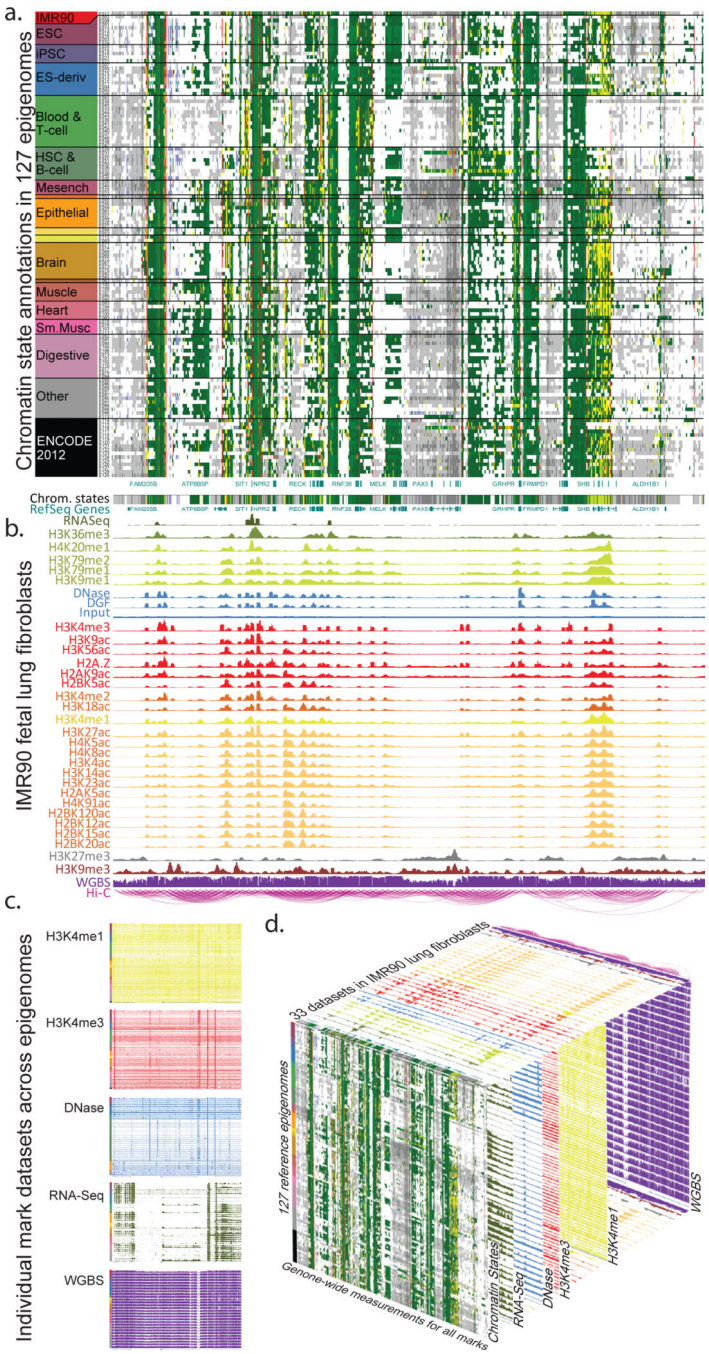


Figure 3. Epigenomic information across tissues and marks

a. Chromatin state annotations across 127 reference epigenomes (rows, Fig. 2) in a ~3.5Mb region on chromosome 9. Promoters are primarily constitutive (red vertical lines), while enhancers are highly dynamic (dispersed yellow regions). **b.** Signal tracks for IMR90 showing RNA-seq, a total of 28 histone modification marks, whole-genome bisulfite DNA methylation, DNA accessibility, Digital Genomic Footprints (DGF), input DNA, and chromatin conformation information[71]. **c.** Individual epigenomic marks across all

epigenomes in which they are available. **d.** Relationship of figure panels highlights dataset dimensions.

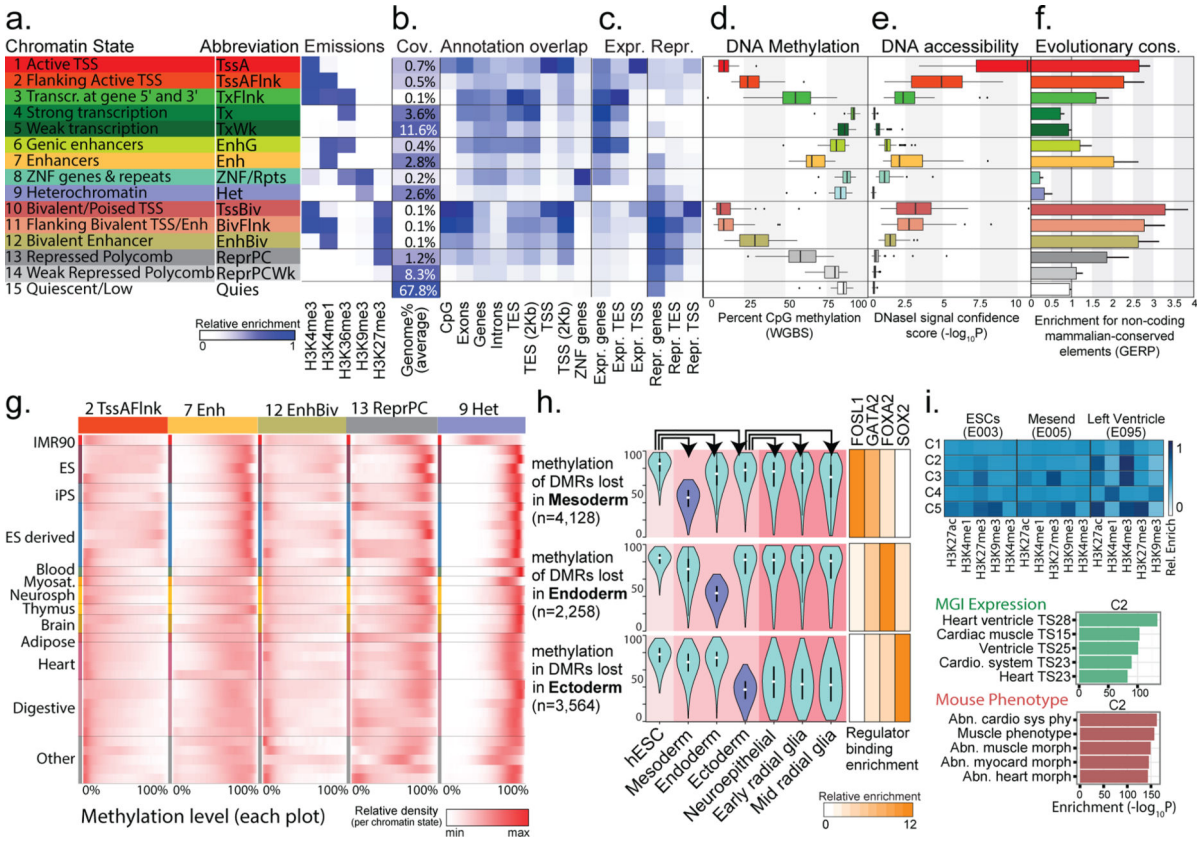


Figure 4. Chromatin states and DNA methylation dynamics

a. Chromatin state definitions, abbreviations, and histone mark probabilities. **b.** Average genome coverage. Genomic annotation enrichments in H1-ESC. **c.** Active and inactive gene enrichments in H1-ESC (see Extended Data 2b for GM12878). **d.** DNA methylation. **e.** DNA accessibility. **d-e.** Whiskers show 1.5 * interquartile range. Circles are individual outliers. **f.** Average overlap fold enrichment for GERP evolutionarily conserved non-coding regions. Bars denote standard deviation. **g.** DNA methylation (WGBS) density (color, ln scale) across cell types. red=max ln(density+1). Left column indicates tissue groupings, full list shown in Extended Data 4f. **h.** DNA methylation levels (left) and TF enrichment (right) during ESC differentiation. **i.** Chromatin mark changes during cardiac muscle differentiation. Heatmap=average normalized mark signal in Enh. C5 cluster enrichment[54].

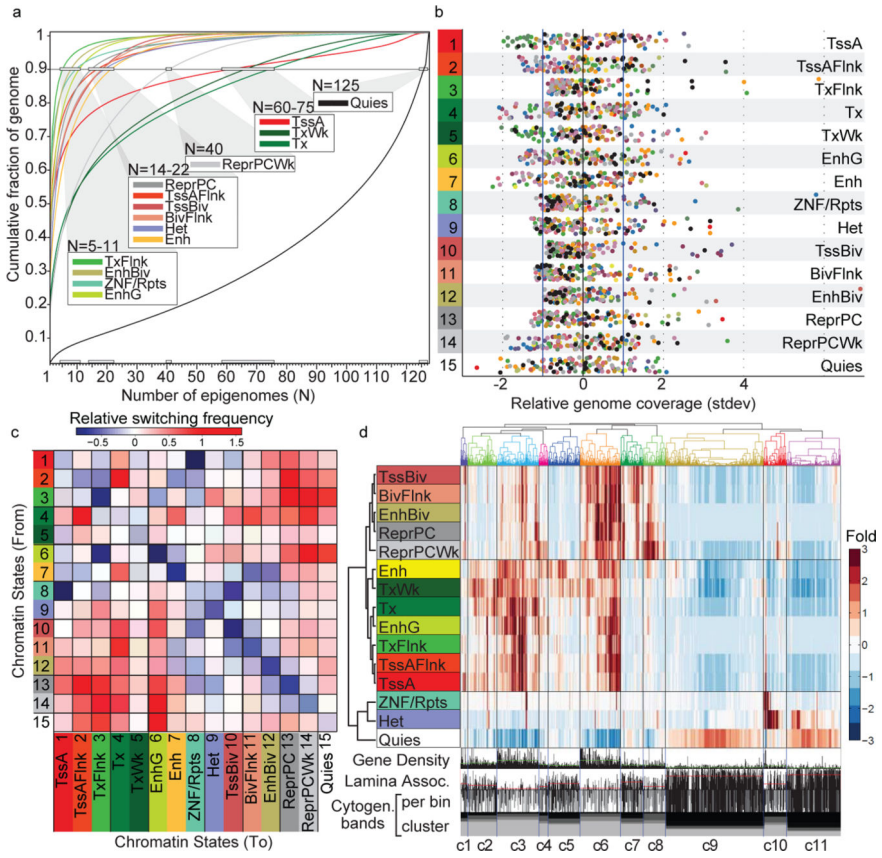


Figure 5. Cell type differences in chromatin states

a. Chromatin state variability, based on genome coverage fraction consistently labeled with each state. **b.** Relative chromatin state frequency for each reference epigenome. **c.** Chromatin state switching $\log_{10}$ relative frequency (inter-cell-type vs. inter-replicate). **d.** Clustering of 2Mb intervals (columns) based on relative chromatin state frequency (fold enrichment), averaged across reference epigenomes. LaminB1 occupancy profiled in ESCs. Red lines show cluster average.

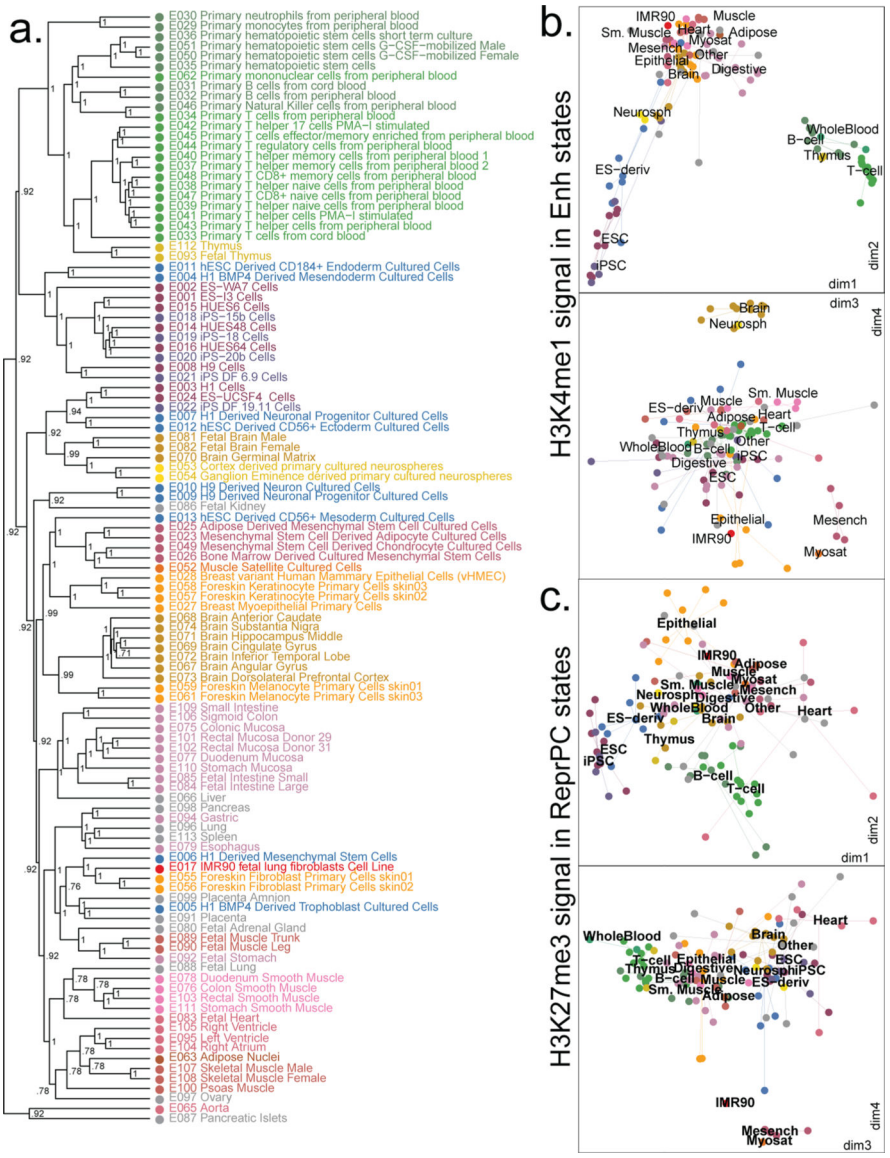


Figure 6. Epigenome relationships

a. Hierarchical epigenome clustering using H3K4me1 signal in Enh states. Numbers indicate bootstrap support scores over 1,000 samplings. **b-c.** Multidimensional scaling (MDS) plot of cell type relationships based on similarity in H3K4me1 signal in Enh states (b) and H3K27me3 signal in ReprPC states (c). First four dimensions shown as dim1 vs. dim2 and dim3 vs. dim4.

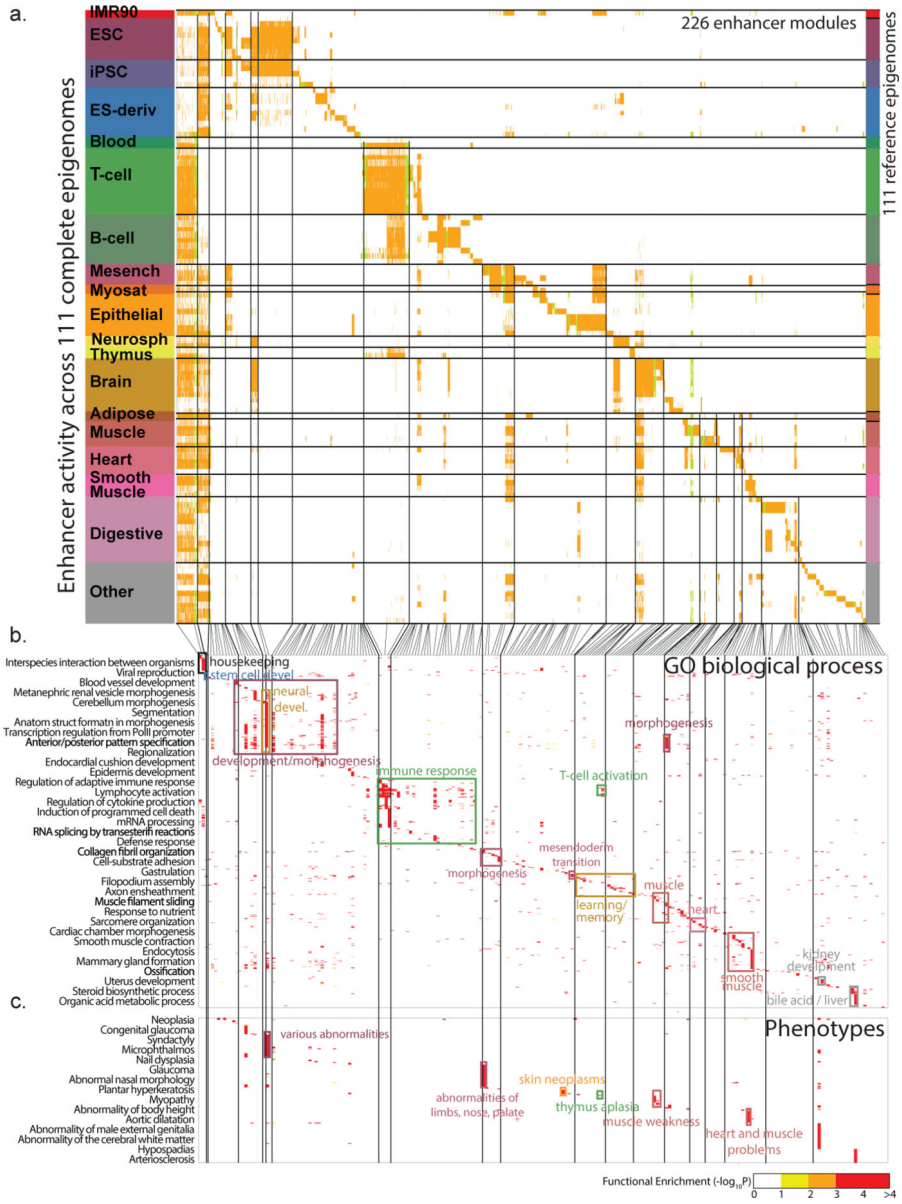


Figure 7. Regulatory modules from epigenome dynamics

a. Enhancer modules by activity-based clustering of 2.3 million DNase-accessible regions classified as Enh, EnhG or EnhBiv (color) across 111 reference epigenomes. Vertical lines separate 226 modules. Broadly-active enhancers shown first. Module IDs shown in Fig. S11c. **b-c.** Proximal gene enrichments[54] (b) for each module using gene ontology (GO) biological process (panel b) and human phenotypes (panel c). Rectangles pinpoint enrichments for selected modules. Representative gene set names (left) selected using bag-of-words enrichment.

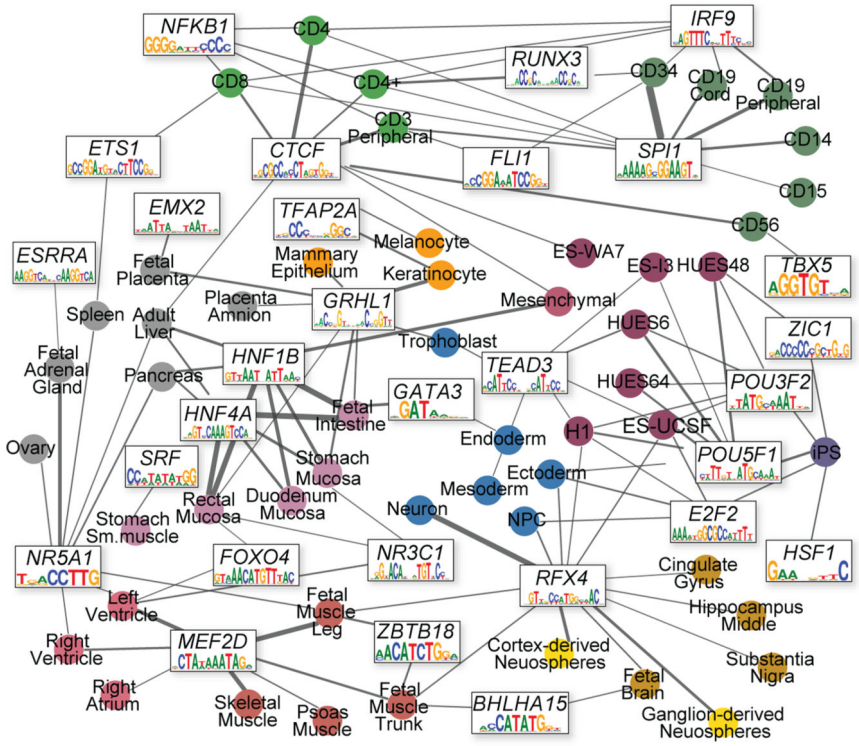


Figure 8. Linking regulators to their target enhancers

Module-level regulatory motif enrichment (Fig. S11) and correlation between regulator expression and module activity patterns (Extended Data 8a) are used to link regulators (boxes) to their likely target tissue and cell types (circles). Edge weight represents motif enrichment in the reference epigenomes of highest module activity.

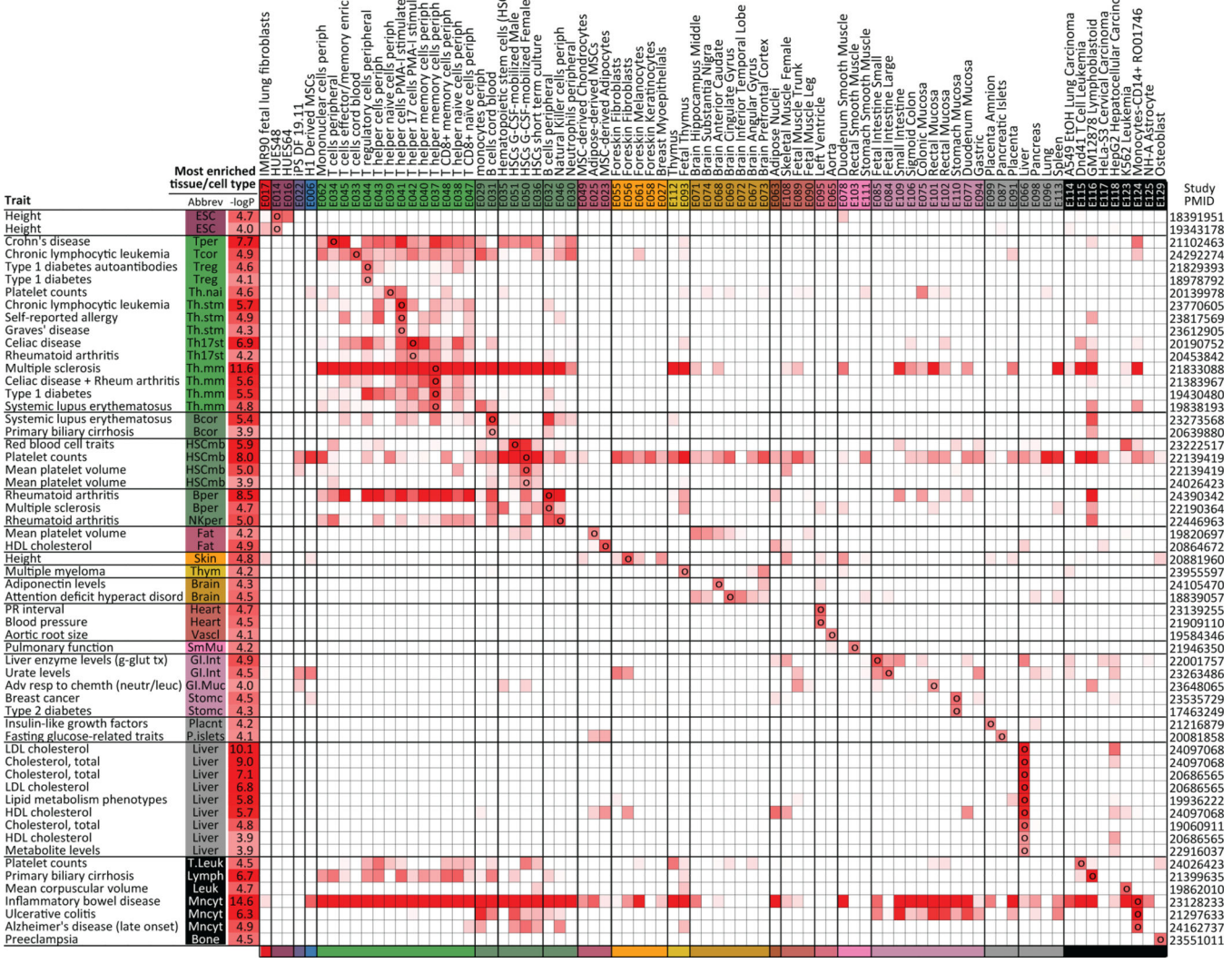


Figure 9. Epigenomic enrichments of genetic variants associated with diverse traits

Tissue-specific H3K4me1 peak enrichment for genetic variants associated with diverse traits. Circles denote reference epigenome (column) of highest enrichment for SNPs reported by a given study (row), defined by trait and publication (PubMed identifier, PMID). Tissue (Abbrev) and p-value ($-\log_{10}$) of highest enrichment are shown. Only rows and columns containing a value meeting a FDR of 2% are shown (Full matrix for all studies showing at least 2% FDR in Extended Data 11-12).