The NIH Roadmap Epigenomics Mapping Consortium

Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen & James A Thomson

The NIH Roadmap Epigenomics Mapping Consortium aims to produce a public resource of epigenomic maps for stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.

Recent years have seen remarkable progress in understanding of human genetics, enabled by the availability of the human genome sequence and increasingly high-throughput technologies for DNA analysis[1]. Yet despite their breadth and comprehensiveness, purely DNA sequence–level investigations do not shed light on a crucial component of human biology: how the same genome sequence can give rise to over 200 different cell types through remarkably consistent differentiation programs. This process of developmental specification, classically termed 'epigenesis', is now known to involve differential regulation of genes and their products[2]. Aberrant regulation of such phenomena has been extensively linked to human diseases and, additionally, can be influenced by environmental inputs[3–5].

Gene regulation and genome function are intimately related to the physical organization of genomic DNA and in particular to the way it is packaged into chromatin, a complex nucleoprotein structure comprising histones, DNA binding factors, accessory protein complexes and noncoding RNAs[6–9] (**Fig. 1**). Chromatin is a dynamic entity that is subject to modification of both its DNA and protein components, with direct structural and functional consequences. The term 'epigenome' is used to describe the way in which these modifications and structural features are distributed across the genome in a given cell population. The epigenomic landscapes and the associated gene expression programs are maintained within a given cell lineage through complex processes that involve transcription factors, chromatin regulators, histone modifications and variants, and RNAs[10–12], but that remain poorly understood in mammals.

Although the mechanisms remain obscure, a now overwhelming body of evidence supports central roles for epigenomic changes in disease susceptibility and pathogenesis. Multiple disease processes, including cancer, are now well known to be associated with characteristic alterations in the patterns of chromatin, DNA methylation and gene expression[3,5]. In addition, epidemiological studies have linked early environmental exposures, such as *in utero* starvation, to long-term health consequences ranging from metabolic disorders to psychiatric diseases[13]. A causal role for epigenomic aberrations is supported by several lines of evidence, including mutations of genes encoding chromatin regulators in developmental disorders and cancer[4,14–16], and by the therapeutic efficacy of small-molecule inhibitors of DNA methyltransferases and histone-modifying enzymes[17].

Major epigenomic features can now be interrogated comprehensively by combining cellular, biochemical and molecular techniques with high-throughput sequencing. Production of genome-wide maps of cytosine methylation, histone modifications, chromatin accessibility and RNA transcripts represents a powerful and

Bradley E. Bernstein, Alexander Meissner, Manolis Kellis, Eric S. Lander and Tarjei S. Mikkelsen are at the Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA; Bradley E. Bernstein is also at the Howard Hughes Medical Institute, Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; and Alexander Meissner is in the Department of Stem Cell and Regenerative Biology at Harvard University, Cambridge, Massachusetts, USA. John A. Stamatoyannopoulos is in the Departments of Genome Sciences and Medicine, University of Washington School of Medicine, Seattle, Washington, USA. Joseph F. Costello is in the Department of Neurosurgery, University of California at San Francisco, San Francisco, California, USA. Bing Ren is at the Ludwig Institute for Cancer Research, University of California San Diego School of Medicine, La Jolla, California, USA. Aleksandar Milosavljevic and Arthur L. Beaudet are in the Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. Marco A. Marra and Martin Hirst are at the Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada. Joseph R. Ecker is in the Genomic Analysis Laboratory, Salk Institute for Biological Studies, La Jolla, California, USA. Peggy J. Farnham is at the Genome Center, University of California at Davis, Davis, California, USA. James A. Thomson is at the University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA.
e-mail: Bernstein.Bradley@mgh.harvard.edu

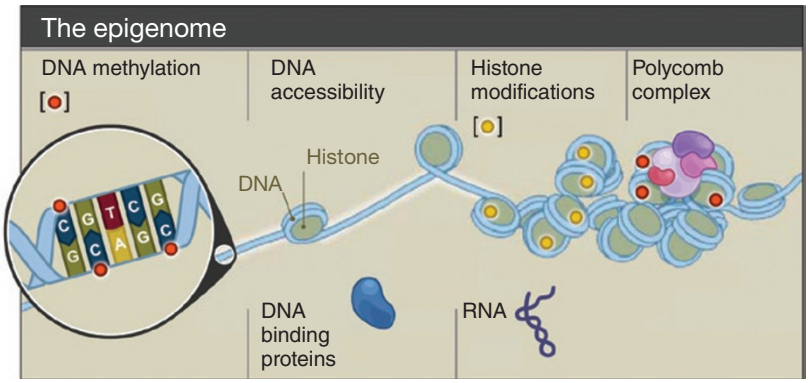

Figure 1 Layers of genome organization. Genome function and cellular phenotypes are influenced by DNA methylation and the protein-DNA complex known as chromatin. In mammals, DNA methylation occurs on cytosine bases, primarily in the context of CpG dinucleotides. Accessible chromatin that is hypersensitive to DNase I digestion marks promoters and functional elements bound by transcription factors or other regulatory proteins. Histone modifications, associated proteins such as Polycomb repressors and noncoding RNAs constitute an additional layer of chromatin structure that affects genome function in a context-dependent manner.

general approach for surveying the regulatory state of the genome in a cell type of interest. The resulting data define the locations and activation states of diverse functional elements, including genes and their transcriptional control elements (e.g., promoters, enhancers and insulators), noncoding transcripts and epigenetic effectors, such as imprinting control regions[18–25]. More globally, such maps can provide insight into developmental state and potential, for example of a stem cell population, and shed light on aberrant regulatory programs in diseased tissues.

Here we describe the aims and scope of the US National Institutes of Health (NIH) Roadmap Epigenomics Mapping Consortium, which has set out to provide a publicly accessible resource of epigenomic maps in stem cells and primary *ex vivo* tissues. These maps will detail the genome-wide landscapes of DNA methylation, histone modifications and related chromatin features, and are intended to provide a reference for studies of the genetic and epigenetic events that underlie human development, diversity and disease. Below, we describe the organizational structure, goals and anticipated deliverables of the consortium.

A coordinated study of human epigenomes

In 2008, the NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. The consortium leverages experimental pipelines built around next-generation sequencing technologies to map DNA methylation, histone modifications, chromatin accessibility and RNA transcripts in stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease. The mapping of such normal epigenomes is being undertaken by four Epigenomics Mapping Centers and supported by a Data Analysis and Coordinating Center, which collectively coordinate experimental and analytical efforts to maximize consistency, data quality and overall coverage of the epigenomic landscape.

Because the epigenomic landscape varies markedly across tissue types (and between individuals), there is no single 'reference' epigenome. Rather, the consortium expects to deliver a collection of normal epigenomes for different tissues and individuals, intended to provide a framework or reference for comparison and integration within a broad array of future studies. A core goal of the consortium is to close the gap between data generation and its public dissemination by rapid release of raw sequence data, profiles of epigenomic features and higher-level integrated maps, in coordination with the US National Center for Biotechnology Information (NCBI). The consortium is also committed to the development, standardization and dissemination of protocols, reagents and analytical tools to enable the research community to utilize, integrate and expand upon this body of data (**Fig. 2**).

Reference maps for major epigenomic features

The Epigenomics Mapping Centers have collaboratively established data collection pipelines to produce high-quality, comprehensive epigenomic maps. Specific data types have been prioritized that offer broad insight into genome regulation, are generally applicable to diverse cell populations and can be evaluated comprehensively and accurately by high-throughput sequencing. These include genomic maps for DNA methylation, histone modifications, chromatin accessibility and RNA expression. The Mapping Centers work with the Data Analysis and Coordination Center to evaluate, compare and integrate the different data types and formats to ensure data quality and standards that enable the larger community to build upon these data.

The first of these data types, DNA methylation, is assayed by sequencing DNA that has been treated with sodium bisulfite (BS-seq), or enriched by methylcytosine pull-down (methylated DNA immunoprecipitation (MeDIP)-seq) or methylation-sensitive restriction enzymes (MRE-seq). BS-seq, applied either to whole genomes or to reduced-representation samples, has been designated as a primary assay because it provides accurate and consistent nucleotide-resolution data. The consortium is implementing MeDIP-seq and MRE-seq on a more limited basis to benchmark and compare these widely applied approaches.

A second type of data, histone modifications, are assayed by sequencing DNA enriched by chromatin immunoprecipitation with modification-specific histone antibodies (ChIP-seq). The consortium has implemented rigorous specificity tests that use arrays of differentially modified histone tail peptides to ensure antibody specificity. In addition, common cell sources are collectively profiled and compared, ensuring consistency between the different data-collection centers.

Chromatin accessibility is assayed by sequencing DNase I cleavage sites in nuclear chromatin. These assays are performed at high sequencing depth to provide a global survey of accessible regions as well as high-resolution information regarding the protein occupancy of specific sequences[24].

Finally, RNA expression is assayed by sequencing mRNAs or size-selected small RNA fractions to high depths. These expression data are intended to augment and illuminate the functional output of the epigenomic profiles.

Given its mandate to deliver epigenomic maps for hundreds of different cell populations, the consortium must balance breadth of cell coverage with the depth to which different epigenomic features are investigated. High-value cell types, such as human embryonic stem cells (hESCs), will be subjected to deep exploration of a very broad range of histone modifications and comprehensive, single nucleotide–resolution analysis of

DNA methylation. Although it is not yet possible to specify a definitive set of features that represent a minimal epigenome, the consortium has initially identified DNA methylation, six major histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3 and H3K36me3). chromatin accessibility and RNA as essential features that will be assayed in most or all designated cell populations. This combination of deep and broad analysis is expected to maximize coverage of cellular diversity and disease-relevant human tissues, while ensuring that a broad range of epigenomic features is explored.

Prioritized cells and tissues

The consortium will investigate a diverse collection of cell and tissue models, including hESCs and adult stem cells and their differentiated progeny; induced pluripotent stem cells; and primary *ex vivo* human fetal and adult tissues. These cells and tissues were prioritized on the basis of broad scientific and biomedical interest, tractability, phenotypic diversity and under-representation in other collaborative projects.

Because of their biomedical importance, hESCs and major lineage derivatives have been selected for intensive investigation. The resulting data will offer insight into the distributions, dynamics and inter-relationships among epigenomic features, and catalyze study of their functions in development, epigenetic control and genome regulation. The consortium will also target additional stem cell models, including mesenchymal and neural stem cells, and reprogrammed cells, as *in vitro* models of development with particular relevance to regenerative medicine.

Broader coverage of human cellular diversity will be achieved through study of primary cells and tissues relevant to metabolic and cardiovascular disease, cancer, neuropsychiatric disease, aging and other leading health issues. These will be acquired from primary sources and sorted or otherwise manipulated to obtain suitably homogeneous cell populations that will be directly channeled to data-collection pipelines. Prioritized cell types include sorted hematopoietic lineages, liver, muscle and adipose, as well as selected cell types from breast and neural tissues. In addition, fetal tissues will be analyzed for insight into epigenomic landscapes of early development. Maps for such primary *ex vivo* tissues are urgently needed because most of our current knowledge has come from either transformed cell lines or cultured cells, both of which experience marked nonphysiologic changes to their chromatin environment, including aberrant DNA hypermethylation and loss of heterochromatin integrity. Collectively, profiles for these diverse cell models should offer unprecedented insight into the breadth and dynamics of human epigenomes and provide a durable framework for future explorations of epigenomic changes associated with human disease.

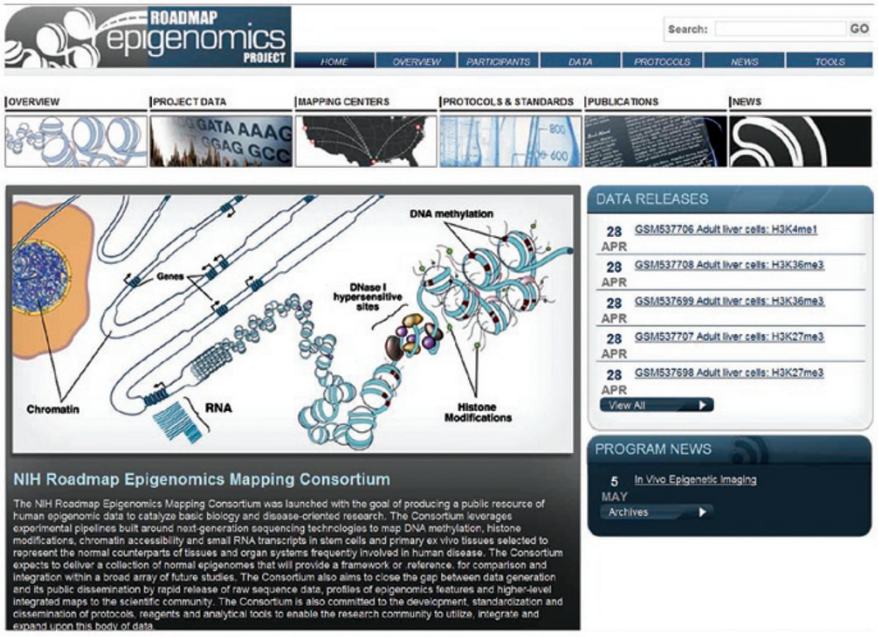

Figure 2 Portal for the NIH Roadmap Epigenomics Mapping Consortium. A public portal (http://www.roadmapepigenomics.org/) provides general information about the consortium and its participants, along with links to experimental protocols, consortium data and interfaces for visualizing epigenomic maps.

Integration and dissemination of human epigenomes

The consortium aims to provide the scientific community ready access to a critical mass of high-quality epigenomic data for cells and tissues representative of normal human biology. These data will comprise multiple levels of information, from raw sequencing data and epigenomic profiles for an individual epigenomic feature in a single cell or tissue type, to integrated epigenomic maps that represent a composite of multiple epigenomic profiles for an individual cell type or, alternatively, that capture biological variation of such features across different cell types. The consortium will also develop and disseminate software tools and algorithms to facilitate use of these results by the community—for example, through the ability to search for epigenomic signatures common across genes or loci, or to identify distinguishing features of cell lineages, developmental stages, cellular environments or derivation history. The latter may also be used to classify disease states or to identify aberrant epigenomic features or regulatory programs that underlie human pathology.

The primary web portal for the consortium (http://www.roadmapepigenomics.org/) offers detailed descriptions of the overall project, target cell and tissue types, and epigenomic assays used by the consortium. The portal links to companion sites managed by NCBI (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/) and the Data Analysis and Coordination Center (http://www.epigenomeatlas.org/) that provide access to raw and processed consortium data along with tools for visualization, analysis and integration of epigenomic data.

Progress and challenges

The use of established technologies and approaches has enabled the consortium to rapidly initiate data production. Notable progress during the initial phase included production of comprehensive DNA methylomes for an hESC line (H1) and primary fibroblasts[26], and generation of hundreds of data sets—for major histone modifications, targeted DNA methylation analysis, RNA expression and chromatin accessibility—representing dozens of cell types, including multiple stem cell lines and *ex vivo* adult and developing tissues. These data sets are now available for download and viewing at the web portals referenced above. Guided by other NIH genomics projects[27], the consortium has adopted a data release policy under which users will have immediate access to the data

but are expected to abide by a moratorium on submission or presentation of works that incorporate these data for the 9 months following their release.

Any effort of this scope inevitably faces challenges and obstacles. The chief issues have revolved around cell-type selection and acquisition, assay standardization and developing the infrastructures for integration and dissemination of epigenome-scale data sets.

Cell type selection and acquisition. A key ongoing challenge relates to the identification and prioritization of cells and tissues by the consortium. Ideally, models are selected on the basis of pervasive biological and medical importance. However, the decisions are confounded by issues of tractability. Many high-value primary tissues are available in limited quantities that push the detection boundaries of current technologies. In addition, isolating relatively homogeneous populations from certain complex tissues can involve extensive preparative steps that may themselves effect changes to the epigenome. Finally, our relatively crude understanding of inter-individual epigenomic variation leaves open the question of how many samples of a given tissue type must be analyzed to yield a representative map. These challenges highlight the importance of technology development, including effective procedures for isolating homogeneous cell populations, interrogating small samples and increasing the throughput of the assays.

Standardization of assays. The consortium is implementing the latest epigenomic technologies based on next-generation sequencing technology. Because these technologies continue to evolve and are inherently dependent on preparative steps, there is an ongoing need to benchmark and validate assays. In the case of histone modification assays, substantial resources must be committed to procurement and validation of high-quality antibody reagents, including confirmation of biochemical specificity and ChIP-seq efficacy. In the case of DNA methylation, there is a need to benchmark and standardize different assay types, including BS-seq applied either to reduced representations of the genome or to the whole genome, as well as various enrichment methods in widespread use by the scientific community[28].

Data integration and dissemination. Several challenges have emerged at the level of data handling and analysis. First, a clearer understanding of the underlying data sets in terms of sensitivity, specificity and precision is needed and is being pursued as a joint effort among the centers. Second, the sheer volume and complexity of consortium-generated data has pushed the limits of existing analytical and visualization tools. Thus, the development of a new generation of tools for integration, dissemination and interpretation of epigenomic data is vital to the overall success of the program.

Future and context

The long-term goal of epigenomics research is a fuller understanding of how global changes in diverse functional features superimposed on the human genome sequence contribute to cellular phenotypes in health and disease. This is a complex and ambitious undertaking, the realization of which will ultimately require systematic dissection and analysis of tissues, characterization of disease models and detailed exposition of regulatory mechanisms through model-organism studies. The efforts of the Roadmap Epigenomics Mapping Consortium to establish an expansive resource of epigenomic maps of normal cell and tissue phenotypes represents an important step in this direction. By catalyzing subsequent mechanistic studies of chromatin, DNA methylation and transcription, these efforts should provide a springboard for disease-focused studies, such as those currently being pursued under the parallel Roadmap program Epigenomics of Human Health and Disease. These Roadmap efforts will also be complemented by other major initiatives, such as the International Human Epigenome Consortium, which was established to accelerate and coordinate epigenomics research worldwide (see accompanying paper[29]).

More broadly, the consortium aims to foster synergistic interactions with related collaborative projects, including the Encyclopedia of DNA Elements (ENCODE) Consortium[18], the International HapMap Project and the 1000 Genomes Project[30]. The Epigenomics Mapping Consortium is distinguished from these efforts by the broad set of normal primary tissues and stem cell–derived developmental models that it will survey. As such, it will provide a highly complementary resource through which the *in vivo* state and behavior of DNA elements catalogued under ENCODE or implicated in studies of genome variation may be understood. Such information will be essential for appreciating the relevance of detected genomic elements and variants to normal development and human disease.

In the coming years, the Roadmap Epigenomics Program and other complementary efforts should vastly improve understanding of the organization of the human epigenome and how it varies across tissues, individuals and disease states—information that may translate directly into the identification of aberrant epigenetic events that underlie susceptibility to specific diseases and environmental exposures.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

ACKNOWLEDGMENTS
We thank R. Waterland, C. Epstein, N. Shoresh and all consortium members, as well as the NIH Epigenomics Implementation Group, for discussions and feedback in the drafting of this document.

1. Altshuler, D., Daly, M.J. & Lander, E.S. *Science* **322**, 881–888 (2008).
2. Bird, A. *Nature* **447**, 396–398 (2007).
3. Feinberg, A.P. *Nature* **447**, 433–440 (2007).
4. Jaenisch, R. & Bird, A. *Nat. Genet.* **33** Suppl, 245–254 (2003).
5. Jones, P.A. & Baylin, S.B. *Cell* **128**, 683–692 (2007).
6. Kouzarides, T. *Cell* **128**, 693–705 (2007).
7. Bernstein, B.E., Meissner, A. & Lander, E.S. *Cell* **128**, 669–681 (2007).
8. Fraser, P. & Bickmore, W. *Nature* **447**, 413–417 (2007).
9. Zaratiegui, M., Irvine, D.V. & Martienssen, R.A. *Cell* **128**, 763–776 (2007).
10. Schwartz, Y.B. & Pirrotta, V. *Nat. Rev. Genet.* **8**, 9–22 (2007).
11. Grewal, S.I. & Moazed, D. *Science* **301**, 798–802 (2003).
12. Henikoff, S. *Nat. Rev. Genet.* **9**, 15–26 (2008).
13. Jirtle, R.L. & Skinner, M.K. *Nat. Rev. Genet.* **8**, 253–262 (2007).
14. Hess, J.L. *Crit. Rev. Eukaryot. Gene Expr.* **14**, 235–254 (2004).
15. Hansen, R.S. *et al. Proc. Natl. Acad. Sci. USA* **96**, 14412–14417 (1999).
16. Dalgliesh, G.L. *et al. Nature* **463**, 360–363 (2010).
17 Batty, N., Malouf, G. G. & Issa, J. P. *Cancer Lett.* **280**, 192–200 (2009).
18. Birney, E. *et al. Nature* **447**, 799–816 (2007).
19. Heintzman, N.D. *et al. Nature* **459**, 108–112 (2009).
20. Eckhardt, F. *et al. Nat. Genet.* **38**, 1378–1385 (2006).
21. Meissner, A. *et al. Nature* **454**, 766–770 (2008).
22. Cokus, S.J. *et al. Nature* **452**, 215–219 (2008).
23. Barski, A. *et al. Cell* **129**, 823–837 (2007).
24. Hesselberth, J.R. *et al. Nat. Methods* **6**, 283–289 (2009).
25. Mikkelsen, T.S. *et al. Nature* **448**, 553–560 (2007).
26. Lister, R. *et al. Nature* **462**, 315–322 (2009).
27. Toronto International Data Release Workshop Authors *Nature* **461**, 168–170 (2009).
28. Suzuki, M.M. & Bird, A. *Nat. Rev. Genet.* **9**, 465–476 (2008).
29. Satterlee, J. *Nat. Biotechnol.* **28**, 1039–1044 (2010).
30. Frazer, K.A. *et al. Nature* **449**, 851–861 (2007).