



HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2014 August 26.

Published in final edited form as:

Nat Biotechnol. 2010 October ; 28(10): 1053–1056. doi:10.1038/nbt1010-1053.

Prospects for discovery by epigenome comparison

Aleksandar Milosavljevic

The NIH Epigenomics Roadmap Data Analysis and Coordination Center Molecular and Human Genetics Department Baylor College of Medicine, Houston, Texas 77030 amilosav@bcm.edu

Abstract

Epigenomic analysis efforts have so far focused on the multiple layers of epigenomic information within individual cell types. With the rapidly increasing diversity of epigenomically mapped cell types, unprecedented opportunities for comparative analysis of epigenomes are opening up. One such opportunity is to map the bifurcating tree of cellular differentiation. Another is to understand the epigenomically mediated effects of mutations, environmental influences, and disease processes. Comparative analysis of epigenomes therefore has the potential to provide wide-ranging fresh insights into basic biology and human disease. The realization of this potential will critically depend on availability of a cyberinfrastructure that will scale with the volume of data and diversity of applications and a number of other computational challenges.

Introduction

Multiple layers of epigenomic information are being mapped at rapidly decreasing cost using ChIP-seq, MeDIP-seq, RRBS-seq, bisulfite-seq, RNA-seq, small-RNA-seq, DNase-hypersensitivity-seq, 3C-seq, 5C and other “*-seq” assays and their “*-chip” cousins. Integrative analysis of epigenomic information has so far focused on multiple layers of epigenomic information within individual sample types^{1 2 3 4}. The most recent achievement in this regard is computational inference of a code-book of chromatin states⁵ defined by combinations of histone marks.

Until recently, epigenomic information was only available from a few samples, but the situation is about to change. The systematic sampling of a diversity of cell types by the NIH Epigenomics Roadmap Initiative⁶ and by the increasing applications of “*-seq” assays by the wider research community is likely to produce on the order of hundred epigenomes within a year. The DNA sequencing cost, which is still limiting for the adoption of “*-seq” assays is halving every six months. This strong and steady trend is enabling mapping of epigenomes in the context of smaller, frequently disease-focused projects. The ever denser sampling of the space of epigenomic variation by large and small projects alike is opening unprecedented opportunities for discovery through comparative analysis of epigenomes.

Comparative analysis of epigenomes has two clear precedents – analysis of genomic variation and evolutionary conservation and analysis of perturbations of gene expression patterns. Methodological achievements such as Gene Set Enrichment analysis⁷, and the Connectivity Map⁸ point to analogous opportunities in the analysis of epigenomic perturbations. Yet the unique character and diversity of epigenomic variation calls for fresh approaches. Epigenomes come mostly in the form of quantitative measurements at various

levels of resolution, from the basepair-level resolution of methylation levels detected by whole-genome bisulfite sequencing, to the >100bp level resolution maps of MeDIP-seq methylation maps or histone marks⁹. Unlike the gene-centric measurements expression levels, epigenomic signals are spread throughout the genome and not uniquely attached to any specific genomic element.

Epigenomic variation spans unusually wide physical and time scales and cuts across a variety of biological processes (Fig. 1). Many molecular processes, including transcriptional regulation, splicing, DNA recombination, replication, and repair¹⁰ leave footprints within multiple layers of epigenomic information. Epigenomic diversity spans multiple time scales, ranging from short-term physiological processes such as memory formation¹¹ and cell differentiation¹², to long term processes such as aging¹³ and evolutionary variation¹⁴. Additional dimensions of epigenomic variation are influenced by genetic, environmental, disease-associated and experimental perturbations.

In the following we first review prospects for two specific applications of comparative analysis of epigenomes and then consider relevant computational and cyberinfrastructure challenges.

Mapping the bifurcating tree of cell differentiation

Waddington's iconic epigenetic landscape (Fig 1) provides a visual analogy for understanding cell differentiation in systems biology terms. The ball on the top is a cell at a particular point in a multi-dimensional phase space ("state space" in modern systems biology terms). Each dimension in the state space represents a quantitative measurement of a molecular state such as the expression level of a gene, degree of methylation of a particular enhancer, promoter, or other chemical modification such as acetylation of a histone tail in a particular nucleosome. Waddington's trick was to reduce the state space that has billions of dimensions into a three-dimensional representation that still captures the two key aspects of the dynamics of the system – the bifurcating branching pattern and canalisation (the degree of stabilizing constraint acting along particular trajectories, referred to as "creodes" by Waddington). Deep valleys indicate a high degree of "canalisation" while shallow valleys indicate a low degree of canalisation, and therefore an elevated sensitivity to perturbations due to environmental influences or mutations.

An interesting question is whether the epigenomes from a diversity of related cell types will provide sufficient information to infer the bifurcating branching patterns of the epigenetic landscape, as illustrated in Fig 1. While the question is far from being settled, recent studies of differentiation mediated by the Polycomb-Trithorax system provide hints that this will be possible. It is now recognized that many CpG island genes developmentally regulated by the Polycomb-Trithorax system in embryonic stem cells reside in the "bivalent" or "poised" state defined by the presence of both activating H3K4me3 and inactivating H3K27me3 marks¹⁵. The marks are resolved into an active or inactive state upon differentiation. The understanding of the epigenomic footprint of the Polycomb-Trithorax system was recently applied to identify an extensive H3K27me3 program shared by pancreatic beta cells and acinar cells, reflecting their common developmental history¹⁶. In sharp contrast to the

extensive H3K27me3 program indicating endodermal pancreatic origin of beta cells, their gene expression signature largely resembles that of ectoderm-derived neural tissues. The study provided a definitive example of an epigenomic signature of cell lineage control by showing that the neural expression program was co-opted during late pancreatic cell differentiation through selective activation of a small number of transcriptional regulators involving removal of inactivating H3K27me3 marks. Interestingly, in this case not only was the reconstruction of the lineage informative but so was the deviation from the bifurcating pattern.

Assuming that epigenomes contain sufficient information to reconstruct the presumably bifurcating patterns, the question of methodology arises. One candidate is the cladistic method¹⁷. This method was fruitful in recovering evolutionary branching patterns of speciation based on genomic sequence, and with some adaptations to epigenomic data, may prove useful for inferring epigenetic branching patterns of cellular differentiation.

Detecting and comparing epigenomic perturbations

Epigenome comparisons will deepen our understanding of consequences of genetic, environmental, and disease-related perturbations (Fig 1). Comparisons of epigenomes following such perturbations are likely to provide insights into the mechanistic basis of their phenotypic expressions. For example, a sequence variant mimicking the epigenomic effect of a drug may point to the drug target or drug's mechanism of action.

The pervasiveness and nature of effects of human genomic variation on the epigenome are currently just beginning to be understood¹⁸, with the exception of relatively well understood genomic variants causing human diseases. Two types of such variants are known to exist – those acting *in trans* via changes in epigenome maintenance such as the Rett and ICF syndromes, and those acting *in cis* such as the Fragile X and Facioscapulohumeral Muscular Dystrophy (FSHMD). A key question is how frequently do sequence variants cause changes in the epigenome? Even subtle mutations *in trans* may affect the whole epigenome and therefore be detectable. The detection of epigenomic phenotypes of mutations *in cis* will be aided by the proximity of the mutations and their phenotypes, as is the case in the analogous context of eQTL mapping.

The information about DNA sequence variation provided as a “side-benefit” of *-seq assays provides an exceptional opportunity to understand the effects of genomic variation on the epigenome. In heterozygous SNP loci, the “*-seq” assays provide allele-specific information about epigenomic states. The epigenomic differences between two alleles (homologous chromosomes) may be “sequence dependent”, i.e. may associate with a sequence variant on the same chromosome. The sequence variant may in this case even be the cause of the allele-specific epigenomic change. An opposite situation, due to imprinting, may be recognized by an “allelic flip” – a situation where two alleles at a specific locus display reverse biases in two samples due to reversed parents of origin.

The first high-resolution genome-wide survey of sequence-dependent and allele-specific methylation in humans was reported earlier this year¹⁹. The study found that sequence variants have pervasive effects on the epigenome. A large degree of allelic variation

between epigenomes was observed. A large fraction of this variation was not attributable to imprinting or X inactivation, and is therefore likely due to sequence variants acting *in cis*. This indicates that epigenome comparisons may shed light of the functional consequences of an increasing catalogue of sequence variants within the >98% non-protein-coding fraction of the human genome. In this regard, a specific plan has been proposed²⁰ to utilize the epigenomic allelic imbalance data to identify SNPs of functional significance within critical regions detected by GWAS studies.

Epigenome comparisons may also help identify functional consequences of structural variants. A recent study²¹ provides indirect evidence that the effects of copy number variants on the epigenome may be widespread. The study reports that the effects of copy number variants on gene expression are not limited to the genes within copy-altered loci, as commonly assumed. In fact, most of the genes affected reside far away from the structural alteration, leading the authors to hypothesize that the effects of structural variants may be mediated by local changes in chromatin structure. Epigenome comparisons will likely be employed to test this hypothesis.

Computational and engineering challenges ahead

An inherent challenge for epigenome comparison comes from the fact that epigenomic signals are not as structured as is the sequence of letters of genomic DNA or the set of quantitative measurements of gene expression obtained by a gene chip experiment. The level of resolution of epigenomic signals varies either due to the nature of the assay – from the basepair level of resolution of methylation levels of specific cytosines determined by bisulfite sequencing to the at most hundred basepair level of resolution of MeDIP assays – or is inherent as is the case for modifications of histone tails which do not correspond to any specific basepair of genomic DNA. There are a number of different solutions to this problem. One is as averaging signals over specific window sizes or over genomic features such as exons, introns or enhancer elements. Another alternative is the natural parsing of epigenomic signals into discrete peaks – a solution suitable for punctate peaks such as trimethylation of H3K4 but not suitable for the broad peaks associated with many other signals such as trimethylation of H3K36. There will likely be numerous ways in which the genome-wide signals will be transformed into a set of numbers for epigenome comparison, each transformation being suitable for specific set of purposes.

One can envision two aspects of comparison – similarity search and detection of specific differences. Regarding the similarity search, using whole-genome comparison as an analogy, we may hypothesize that a combination of global and local methods may emerge. Unlike genomic sequence, which provides a convenient concept of “locality” in the one-dimensional basepair coordinate system, epigenomes may benefit from being compared through a prism of non-contiguous loci such as those containing binding sites of specific master regulators, sets of genes related to a particular differentiation pathway or gene elements such as promoters.

Interpretation of specific differences detected between two epigenomes will depend on our understanding of the natural variation in the signal. In analogy to DNA sequence

comparisons, we need to understand which loci are “conserved” and which are under looser constraint. The differences of the same magnitude in constrained loci would then be interpreted to be of potentially higher functional significance than those at inherently variable loci. Of course, the immediate problem is that we do not have such knowledge. Gradual accumulation of data will solve this problem but probably not in a definitive way since variation may be highly context-dependent – variation during development in a particular cell lineage may for example have very different meaning than physiological variation in a different lineage or variation due to aging. One consequence is that the differences detected at a certain level of analysis will be open to context-dependent reinterpretation with accumulating data.

The issue of standards for representing, visualizing and sharing the results of comparisons will inevitably arise (denoted as Data Level 4 in Box 1). Like genomic sequence, epigenomic signal differences may be projected in the one-dimensional basepair coordinate system. On the other hand, like the results of a gene expression array experiment, they may also be meaningfully summarized as a set of numbers obtained by one or the other transformation of the genome-wide epigenomic signal.

The comparative interpretation of epigenomic signals will also pose a number of technical and engineering challenges that are often grouped under the term “cyberinfrastructure”. This includes the standards, resources, and tools for computer-aided discovery, data sharing and collaboration over the web. The problems of high-volume data capture, visualization, interpretation and reuse are currently recognized as key limiting factors across scientific disciplines²². A comprehensive review of specific bioinformatic cyberinfrastructures²³ summarizes the state of the art as an “archipelago” of isolated systems.

One practical cyberinfrastructure challenge is to enable effective data exchange and reuse. The first step in this direction is to develop a unifying framework for the multiple layers of heterogeneous information generated by the “*-seq” and *-chip” assays. Data standards – see Data Levels in Box 1 -- are emerging in coordination between the pilot TCGA, 1000 Genomes, ENCODE, and the NIH Epigenomics Roadmap projects. The abstract Data Levels codify commonalities across the diversity of assays and technologies.

Another important element of cyberinfrastructure are metadata standards. One example is the SRA XML schema developed by the NCBI and adapted by the NIH Epigenomics Roadmap Initiative for epigenomic data. The metadata is a key requirement for reuse of epigenomic data in the public domain for comparative analyses because it provides the biological and experimental context in which the data was generated (Box 1).

Another practical challenge is to ensure reproducibility of reported analysis results²⁴. This problem may be tackled by encapsulating all aspects of in-silico analyses in the form of workflow descriptions (Box 1) and distributing them in conjunction with analysis results as metadata (Box 1).

Epigenome comparisons and higher-level interpretations will be intensive in terms of data storage and computing power. The use of multiple data and computing resources that are geographically distributed over the web and of cloud computing and programming

frameworks such as GATK²⁵ will be required. As the diversity of derived data and knowledge increases, advanced knowledge representation and exchange methods such as the RDF derived in the context of the Semantic Web will be increasingly applied²⁶. For a more complete list of infrastructure requirements and concepts relevant (but not necessarily specific) to epigenome research we refer the interested reader to Box 1.

How will a cyberinfrastructure be used to specifically facilitate epigenome comparison? Fig. 2 illustrates a hypothetical scenario where a number of projects, possibly involving clinical researchers utilize a hosted (Web 2.0) service to process epigenomic data to perform comparative analysis. Software-as-service is appealing because a minimum of local resources may be required. This is particularly important for adoption of epigenomics in the context of translational and disease-focused studies where local bioinformatic resources and expertise may be limited. In this scenario, many projects utilize shared remote compute hardware (cloud computing) and well-tested pipelines with built-in quality characterization steps that take in sequencing data (Data Level 0) as it is delivered from the sequencers and generate epigenomic signals at the level of individual samples (Data Levels 1-3). Tools for comparison against the Human Epigenome Atlas and possibly other types of visualization and analysis are provided. Upon publication, the data is deposited to NCBI GEO or SRA archives but also included in more specialized data- and knowledge-bases such as the Human Epigenome Atlas. The data is re-usable because of attached metadata which provides the record of the original experiment, including valuable details about the sample and assay performed. The results of higher level analyses are also attached to the data. The higher-level analyses, typically reported in journal publications, are reproducible using workflows attached to analysis results.

One open issue is to identify the best way of involving the research community in the continued development and maintenance of data- and knowledge-bases such as the Human Epigenome Atlas. Such resources will provide essential context for interpretation of newly obtained epigenomes. To stimulate the contribution of smaller projects to this data- and knowledge commons, in close collaboration with the NCBI, the NIH Epigenomics Roadmap Consortium is developing standards for epigenomic metadata and defining reference pipelines for uniform processing and characterization of quality of a variety of epigenomic assays.

Conclusion

The mapping of epigenomes is likely to provide many novel insights through comparative analysis. The mapping of the bifurcating tree of cellular differentiation will provide a new key reference for understanding organismal development. Precise and comprehensive mapping of epigenomic perturbations will reveal consequences of genomic mutations and environmental influences on human development and disease.

A number of challenges lie ahead. One more specific set of challenges is to develop conceptual and computational approaches for comparative analysis of epigenomes. Another, more general set relates to the engineering of a cyberinfrastructure, including shared data-

and knowledge-bases that will scale with the unprecedented volume of data and diversity of applications.

Box 1
Cyberinfrastructure for epigenome research: key concepts

The components of the emerging cyberinfrastructure are organized around six general requirements listed in the first column.

	Requirement	Concept	Description and examples of relevance for epigenomics	
1.	Data reuse and integration	Data Level	This abstraction captures commonalities and facilitates development of data formats and tools for a diversity of genomic and epigenomic assays. Examples below focus on “dash-seq” assays.	
			Data Level 0	Refers to the DNA sequence reads, typically in SRF or fastq formats.
			Data Level 1	Refers to reads mapped to a reference assembly, typically in SAM/BAM or BED formats. Level 1 data can be used to identify both genomic and epigenomic variation. It also includes unmapped (repetitive) fraction of reads.
			Data Level 2	“Raw epigenomic signal” such as read density plots, CpG methylation counts ²⁸ or other statistics, frequently in the bigWig UCSC Genome Browser format ²⁹ .
			Data Level 3	Typically discrete data such as ChIP-seq peak calls or HMM segmentations of the genome into chromatin states. Obtained by analyzing individual or multiple marks from a single sample. Depending on data volume, stored either in high-density or in simple tab delimited (GFF, LFF) formats.
			Data Level 4	Results of epigenome comparisons. Syntax and semantics for this data level is still under development.
		Syntax	Data formats to meet the often conflicting requirements of storage efficiency for high-volume data (bigWig), simplicity (tab delimited,) and machine readability (JSON, XML).	
		Semantics	Theory of meaning. This term is commonly used in connection with controlled vocabularies and ontologies such as the widely used Gene Ontologies and other ontologies produced by the OBO Foundry and other projects.	
		Semantic Web (Web 3.0)	Set of technologies including RDF for knowledge representation developed by the World Wide Web Consortium (W3C) allowing programmatic communication and automated reasoning about the information shared across the web.	
		Metadata	Data about data. Key requirement for data reuse. Various minimal standards have been recommended by groups such as the MIBBI	

	Requirement	Concept	Description and examples of relevance for epigenomics
			Project. In coordination with EBI and DDBJ, and based on the feedback from the NIH Epigenomics Roadmap Initiative and other users, NCBI has now developed Version 1.2 of an SRA-XML metadata format for assays with sequencing readouts. Shared metadata formats will be essential for a successful coordination of international epigenome projects.
2.	Tool integration	Pipeline	Set of analysis tools that are invoked sequentially to perform a data analysis task. Galaxy30 is a software suite with an interactive interface and an online service for pipeline design. One example is integration of the EpiGRAPH software for epigenome analysis using Galaxy31 to identify epigenetic modifications that are characteristics of highly polymorphic (SNP-rich) promoters.
Workflow		Formal, portable, programmatically executable description of a data analysis process. May be used as metadata to document and ensure reproducibility of data analysis. Projects developing workflow systems include Galaxy, GenePattern, and Taverna.	
Workbench		An environment for integration of data analysis and visualization tools and data sets (e.g., CLC Genomics Workbench and Genboree Workbench).	
3.	Web services and programmatic interoperability	URI and URL	Address system of the Web. Used to uniquely identify objects such as web pages and epigenome maps for access by web browsers and other computer programs via the HTTP and other protocols.
REST API		Representational State Transfer Application Programming Interface. A programming interface, typically implemented using the HTTP protocol that is developed using a set of design principles to ensure efficient communication of computer programs over the web. Provides access to data and computing resources over the web using scripts written in a programming language such as Pearl, Python, Ruby, or JavaScript.	
4.	Access to computing resources and services	Cloud Computing	Access to “elastic”, on-demand computing and storage services over the web (e.g., Amazon and Rackspace cloud computing services)
Software-as-a-Service (SaaS)		Access to software applications over the web such as those for epigenomic data processing and comparison (Fig. 2). This is a key aspect of Web 2.0 (see below).	
5.	Collaboration and publication	Authentication protocol	Protocol (e.g., OpenID) allowing users or computer programs acting as their agents to be recognized by multiple web servers.
Web 2.0		Web hosting of collaborative processes such as grant review at the NIH or epigenomic data processing and comparison (Fig. 2).	
6.	Databases knowledge-bases and archival sites.		Examples include NCBI GEO and SRA archives, Ensembl, UCSC Genome Browser, and the more specialized Human Epigenome Atlas.

References

1. Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007; 447(7146):799–816. [PubMed: 17571346]
2. Barski A, Cuddapah S, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007; 129(4):823–837. [PubMed: 17512414]
3. Heintzman ND, Stuart RK, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007; 39(3):311–318. [PubMed: 17277777]
4. Mikkelsen TS, Ku M, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*. 2007; 448(7153):553–560. [PubMed: 17603471]
5. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010; 28(8):817–825. [PubMed: 20657582]
6. Article by Bernstein et al. on the NIH Epigenomics Roadmap Initiative in the same issue of *Nature Biotechnology*.
7. Subramanian A, Tamayo P, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102(43):15545–15550. [PubMed: 16199517]
8. Lamb J, Crawford ED, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313(5795):1929–1935. [PubMed: 17008526]
9. Reference to the article by R. A. Harris et al. in the same issue of *nature Biotechnology*.
10. Kouzarides T. Chromatin modifications and their function. *Cell*. 2007; 128(4):693–705. [PubMed: 17320507]
11. Levenson JM, Roth TL, et al. Evidence that DNA (cytosine-5) methyltransferase regulates synaptic plasticity in the hippocampus. *J Biol Chem*. 2006; 281(23):15763–15773. [PubMed: 16606618]
12. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*. 2007; 447(7143):425–432. [PubMed: 17522676]
13. Rakyan VK, Down TA, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res*. 2010; 20(4):434–439. [PubMed: 20219945]
14. Bernstein BE, Kamal M, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*. 2005; 120(2):169–181. [PubMed: 15680324]
15. Bernstein BE, Meissner A, et al. The mammalian epigenome. *Cell*. 2007; 128(4):669–681. [PubMed: 17320505]
16. van Arensbergen J, Garcia-Hurtado J, et al. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res*. 2010; 20(6):722–732. [PubMed: 20395405]
17. Ridley, M. *Evolution and classification : the reformation of cladism*. Essex; England: 1989.
18. Meaburn EL, Schalkwyk LC, et al. Allele-specific methylation in the human genome Implications for genetic studies of complex disease. *Epigenetics*. 2010; 5(7)
19. Schalkwyk LC, Meaburn EL, et al. Allelic skewing of DNA methylation is widespread across the genome. *Am J Hum Genet*. 2010; 86(2):196–212. [PubMed: 20159110]
20. Tycko B. Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. *Am J Hum Genet*. 2010; 86(2):109–112. [PubMed: 20159108]
21. Cahan P, Li Y, et al. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet*. 2009; 41(4):430–437. [PubMed: 19270704]
22. Tony Hey, ST.; Tolle, Kristin, editors. *The fourth Paradigm: data-Intensive Scientific Discovery*. Microsoft Research; 2009.
23. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet*. 2008; 9(9):678–688. [PubMed: 18714290]
24. Mesirov JP. Computer science. Accessible reproducible research. *Science*. 2010; 327(5964):415–416. [PubMed: 20093459]

25. McKenna A, Hanna M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20(9):1297–1303. [PubMed: 20644199]
26. Wang X, Gorkitsky R, et al. From XML to RDF: how semantic web technologies will change the design of ‘omic’ standards. *Nat Biotechnol.* 2005; 23(9):1099–1103. [PubMed: 16151403]
27. Waddington, CH. *The strategy of the genes; a discussion of some aspects of theoretical biology.* Allen & Unwin; London: 1957.
28. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics.* 2009; 10:232. [PubMed: 19635165]
29. Rosenbloom KR, Dreszer TR, et al. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.* 2010; 38(Database issue):D620–625. [PubMed: 19920125]
30. Blankenberg D, Von Kuster G, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol.* 2010; 10:11–21. Chapter 19: Unit 19.
31. Bock C, Von Kuster G, et al. Web-based analysis of (Epi-) genome data using EpiGRAPH and Galaxy. *Methods Mol Biol.* 2010; 628:275–296. [PubMed: 20238087]

	Perturbations				
	Natural Variation	Genetic	Environmental	Disease-related	Experimental
Minutes	Physiology	✓	✓	✓	✓
Days	Development	✓	✓	✓	✓
Years	Population	✓	✓	✓	
10 ³ years	Evolution	✓			
10 ⁶ years					

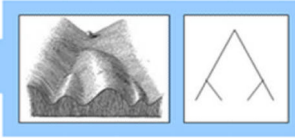


Figure 1. The Scope of Epigenomic Variation

The spectrum of epigenomic variation is wide, spanning biological processes on all time scales, from rapid physiological homeostatic processes that may occur at the scale of minutes to the diversity across species separated by tens of millions of years of evolution. Waddington's epigenetic landscape²⁷ and the bifurcating tree of cellular differentiation corresponding to the landscape are on the right, highlighted by the light blue background.

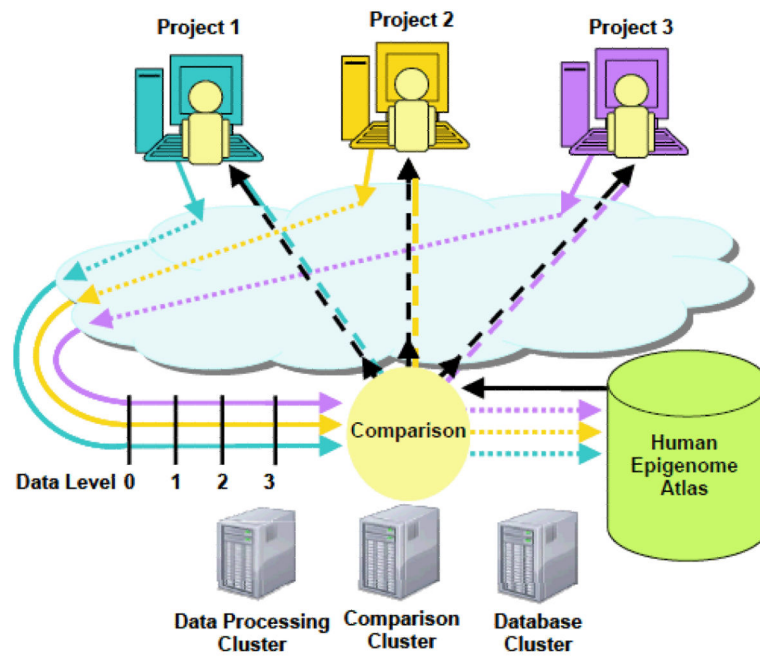


Figure 2. A Cyberinfrastructure for Epigenome Analysis and Comparison

The cyberinfrastructure seamlessly connects users and resources that are geographically distributed over the network. A clinical researcher conducting a study of disease-related epigenomic perturbations may rely almost completely on remote resources distributed over the web for primary processing of the data (Data Levels 0-3) and comparative analysis using the Human Epigenome Atlas. Upon publication of results, individual projects contribute data to the Human Epigenome Atlas, thus enhancing the utility of this shared resource for future users.