D908–D912 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 12 November 2010*
doi:10.1093/nar/gkq1146

NCBI Epigenomics: a new public resource for exploring epigenomic data sets

Ian M. Fingerman, Lee McDaniel, Xuan Zhang, Walter Ratzat, Tarek Hassan, Zhifang Jiang, Robert F. Cohen and Gregory D. Schuler*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA

Received August 27, 2010; Revised October 14, 2010; Accepted October 25, 2010

ABSTRACT

The Epigenomics database at the National Center for Biotechnology Information (NCBI) is a new resource that has been created to serve as a comprehensive public resource for whole-genome epigenetic data sets (www.ncbi.nlm.nih.gov/epigenomics). Epigenetics is the study of stable and heritable changes in gene expression that occur independently of the primary DNA sequence. Epigenetic mechanisms include post-translational modifications of histones, DNA methylation, chromatin conformation and non-coding RNAs. It has been observed that misregulation of epigenetic processes has been associated with human disease. We have constructed the new resource by selecting the subset of epigenetics-specific data from general-purpose archives, such as the Gene Expression Omnibus, and Sequence Read Archives, and then subjecting them to further review, annotation and reorganization. Raw data is processed and mapped to genomic coordinates to generate 'tracks' that are a visual representation of the data. These data tracks can be viewed using popular genome browsers or downloaded for local analysis. The Epigenomics resource also provides the user with a unique interface that allows for intuitive browsing and searching of data sets based on biological attributes. Currently, there are 69 studies, 337 samples and over 1100 data tracks from five well-studied species that are viewable and downloadable in Epigenomics.

INTRODUCTION

Interest in the field of epigenetics has exploded over the previous decade. Epigenetics, strictly defined, refers to the study of stable and heritable changes in gene expression that are not mediated by the primary DNA sequence (1,2). Epigenetic mechanisms participate in processes such as regulating gene expression, homologous recombination and DNA repair. Individual epigenetic features are often called 'marks' because they are stably tied to specific genomic locations and may be propagated through many rounds of cell division, yet have the ability to be 'erased' later as cells undergo differentiation or are exposed to extra-cellular stimuli and environmental cues (3–6). Several types of epigenetic marks have been identified and intensely studied. These include post-translational modifications of histone proteins, DNA methylation, chromatin organization (DNase hypersensitivity) and non-coding regulatory RNA (3). Just as ordinary mutations are known to contribute to disease, so too can corruption of normal epigenetic states ('epimutations'). Indeed, many cancers, degenerative diseases and metabolic disorders have been shown to be the result of mis-targeting of DNA methylation or deficiencies in pathways for histone modification (7–9). We may consider the collection of epigenetic marks across the genome to constitute the cellular 'epigenome'. The study of the epigenome, or epigenomics, refers to identifying what is literally 'on' the genome (the prefix *epi-* indicating above), and how these phenomena impact global gene expression, DNA mediated processes, and subsequently, development. Studying epigenomes presents more of a challenge because each cell type may have a different configuration of features and, undoubtedly, additional epigenetic marks have yet to be discovered.

The development of chromatin immunoprecipitation (ChIP) as an experimental technique was a major breakthrough for the field of epigenetics (10,11). ChIP allows for the genomic localization of chromatin associated proteins. Typically, growing cells are treated with formaldehyde to induce protein–DNA crosslinking, followed by lysis and physical/enzymatic disruption of the chromatin.

*To whom correspondence should be addressed. Tel: +1 301 435 7226; Fax: +1 301 480 5779; Email: schuler@ncbi.nlm.nih.gov

Antibodies that specifically recognize epigenetic features are used to immunoprecipitate the protein–DNA complexes. These antibodies can be specific to modified histones, histone modifying enzymes, transcription factors, or even modified nucleotides. Following the immunoprecipitation, the DNA is isolated from the protein–DNA complexes for analysis. If a particular epigenetic feature is localized to a specific genomic region, DNA representing that region will be enriched in the immunoprecipitate. In conjunction with microarray analysis and more recently, high-throughput (or next-gen) sequencing, these genomic regions can be identified. It is common to represent epigenetic data as genome 'tracks'. The output of a next-gen sequencing experiment is millions of short DNA sequences, which are then aligned to a genome sequence. Sequences that are enriched in the experimental material (an immunoprecipitate, say) will occur multiple times and form visually discernable peaks when represented graphically in a genome viewer. A commonly used data structure for this type of track is the 'wiggle' format developed for use with the UCSC Genome Browser (12). Because files in this format are usually given a .wig file extension, we will hereafter refer to them as 'WIG files'. Further advancement of these technologies has enabled genome-wide epigenetic analysis and as a result, massive amounts of data have been generated characterizing genomic localization of histone modifications, DNA methylation, smRNA and miRNA expression, and chromatin accessibility in numerous organisms and cell types.

With the renewed interest in epigenetics and the methodological advances in whole-genome analysis, in 2007, the NIH launched the Roadmap Epigenomics Project (http://nihroadmap.nih.gov/epigenomics/). Among its aims are the development of reference epigenome maps from a variety of cell types and unraveling the relationships between the epigenomic landscape and human disease. Complementary to this effort are the ENCODE (ENCyclopedia of DNA Elements) project and the corresponding modENCODE project for model organisms (13,14). Although they are focused on identifying functional DNA elements in the genome, many of these sites may be epigenetically regulated or participate in epigenetic regulation. In addition to these large projects, data from individual laboratory projects are incorporated on an ongoing basis. The Epigenomics database is being created as public resource to provide access to these data. It aims to provide both users familiar with the epigenetics field and novice users with a simple and intuitive interface to view, explore, analyze and manipulate these data.

THE EPIGENOMICS DATABASE

Content in the Epigenomics resource is derived primarily from data originally submitted to archival databases at the NCBI, specifically, the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA) (15,16). Although originally established to warehouse gene expression data, GEO has become a general-purpose database for molecular abundance data from wide variety of experiment types, including those aimed at epigenomics. In addition to the raw abundance data, extensive meta data may be provided with GEO submissions in order to fully describe the biological and experimental context. In recent years, measurements of molecular abundance are increasingly being generated by the use of next-gen sequencing-based approaches. SRA serves as a repository for raw next-gen sequence data, together with detailed information on the sequencing instrument and other experimental variables.

In constructing the new Epigenomics database, we have identified the subset of GEO and SRA data that pertain to epigenomics, subjected them to additional review, and reorganized them in a fashion that is more useful for epigenomics researchers. In many cases, data producers provide WIG files as part of their GEO submissions, which allows them to be directly leveraged in the Epigenomics resource. However, because some submissions either lack WIG files or have files based on older genome assemblies, we have developed a pipeline for generating epigenetic tracks that primarily uses the processed output from the Bowtie aligner (17). The Epigenomics database currently has data tracks for epigenetic features, including histone modifications, DNA methylation, chromatin accessibility and expression of small non-coding RNAs. Data is also available for several chromatin associated factors such as histone modifying enzymes, transcription factors, and components of the core transcriptional machinery. More of these data type will be included as they become available. Furthermore, gene expression data for relevant biological samples will also be included in Epigenomics.

The two fundamental types of database records are 'studies' and 'samples', both of which are assigned unique accession numbers (with prefixes ESS and ESM, respectively). A study refers to one or more experiments with a common set of scientific aims. Most often an epigenomic study will correspond to a publication or to a publicly available data set. For each study, there is a brief summary of the scientific design, together with a listing of the biological samples that were studied and the epigenetic features examined. Full data source information is provided, including the submitter's institution, links to the original data submissions in GEO and SRA, links to literature citations in PubMed and (where available) links to the full-text articles in PubMed Central.

Each study is associated with a collection of samples. A sample corresponds to the biological material that was examined and includes detailed biological source attributes with values drawn from controlled vocabularies. In order to unify and consistently assign biological attributes, extensive manual curation is performed using submitted meta data as a foundation. This process may include examining primary literature and researching on-line repositories of cell lines, mouse strains, and tissues. There are over 20 biological attribute fields available, including strain, cultivar, ecotype, individual, gender, age, developmental stage, cell line, cell type, tissue type, health status and many others.

The Epigenomics database was first released in June of 2010. As of this writing, it contains 69 studies, 337 samples and over 1100 data tracks from five well-studied species (Table 1). Currently, data tracks for global expression of micro and small RNAs (194 tracks), histone modifications (626 tracks), DNA methylation (128 tracks), chromatin accessibility (60 tracks) and various chromatin associated factors (including RNA polymerase, transcription factors, and various histone modifying enzymes) (140 tracks), are available among others.

Table 1. Epigenomics database current record holdings[a]

Species	Studies	Samples	Tracks
Arabidopsis thaliana	9	21	27
Caenorhabditis elegans	5	22	27
Drosophila melanogaster	5	18	141
Homo sapiens	29	159	656
Mus musculus	25	117	295
Totals	69[b]	337	1146

[a]Holdings as of 14 October 2010.
[b]Several studies contain data from multiple species.

THE EPIGENOMICS WEBSITE

The entry point for exploring these data is the Epigenomics home page (www.ncbi.nlm.nih.gov/epigenomics/). In addition to basic database searching functionality, it includes links to a series of tutorial documents that explain how to use the database as well as scientific background documents that cover fundamental topics in epigenetics, such as an introduction to histone modifications. The Epigenomics database is part of the NCBI's umbrella Entrez search system, which supports both free-text and fielded queries, together with a uniform system for representing links between related records in different databases (18).

To simplify browsing of the database contents, we have developed a unique Sample Browser tool, which lists samples in a tabular (spreadsheet-like) display (Figure 1). The user-configurable columns correspond to various biological and experimental attributes while the rows are the epigenomic samples. The table may be sorted on any column and the entire table may be exported in a spreadsheet-compatible format (comma-separated values). Pre-set filters provide easy browsing

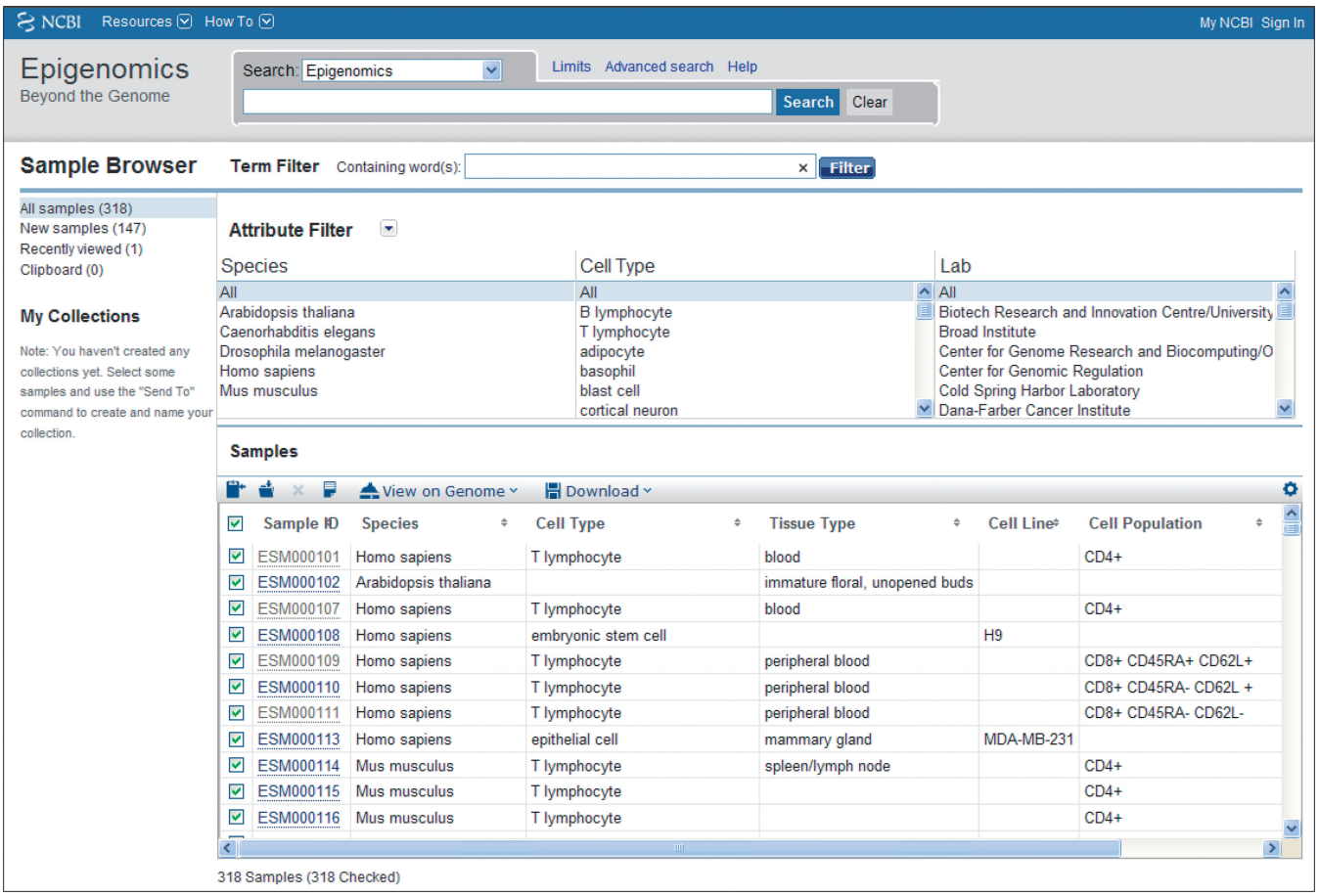

Figure 1. The Sample Browser interface for navigating samples in the Epigenomics database. Samples can be filtered by free text by using the Term Filter. Additionally, three pre-set attribute filters based on species, cell type, or lab (submitting institution) are available. The samples panel displays samples with detailed biological attributes. This view is customizable and samples can be selected and sorted in this interface. The Sample Browser toolbar displays a series of icons that provide additional functionality for manipulating samples (from left to right): add record to clipboard, add record to a collection, remove record (applies to clipboard and collections), view sample information, view on genome, download track data/export browser data as table, and configure (allows column selection within the Sample Browser window).

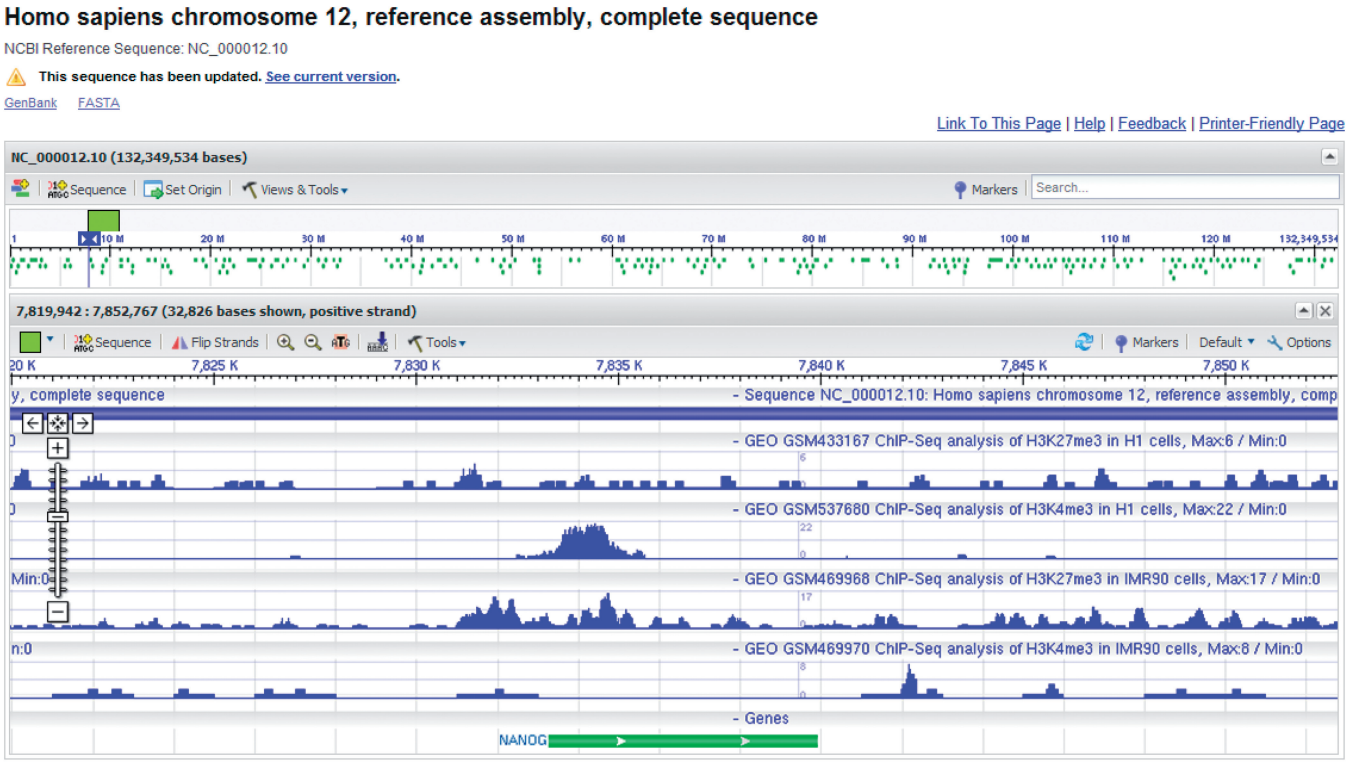

Figure 2. Visualization of epigenetic features in the vicinity of the NANOG gene locus using the NCBI sequence viewer. Epigenomic data tracks can be displayed allowing peaks at specific features to be visualized and compared across different samples. Epigenomic track data are displayed in the middle (blue) tracks. In this example, the track labels indicate we are comparing the epigenetic marks H3K4me3 and H3K27me3 at the NANOG gene locus in both H1 embryonic stem cells (top two tracks) and IMR90 fibroblasts (bottom two tracks). Peaks in the tracks indicate areas of the genome that are enriched for a particular epigenetic feature. The NANOG gene product participates in maintaining stem cell pluripotency. In H1 embryonic stem cells, where the NANOG gene is expressed, the genome tracks show an enrichment of H3K4me3 (a mark associated with actively transcribed genes), and a lack of H3K27me3 (a mark of repressed chromatin). Conversely, in IMR90 fibroblasts the NANOG gene is no longer expressed. In this instance, levels of H3K4me3 are reduced, and there is an enrichment of H3K27me3 at the NANOG locus.

by species, cell type and submitting institution, while a free-text filtering feature allows for fine control of the displayed samples. Sets of samples may be stored temporarily using a clipboard feature or saved permanently (with a free NCBI login) in named collections. The Sample Browser also serves as a hub for connecting to other tools that act on sets of samples, specifically, graphical rendering and bulk downloading.

One of the most common tasks performed with track data involves simple inspection of graphical views and several popular genome browsers have been developed for this purpose. These tools allow peaks at specific features (e.g. promoters and enhancers) to be visualized and compared across different samples. For example, Figure 2 shows the differences in the epigenetic marks, H3K4me3 and H3K27me3, at the developmentally regulated NANOG gene locus in both embryonic stem cells and terminally differentiated fibroblasts. The Epigenomics website provides an easy way for users to visualize a chosen set of tracks using either the NCBI Sequence Viewer or the UCSC Genome Browser.

Advanced users may prefer to download track data to their own systems for local analysis. A bulk downloading tool may be used to retrieve any chosen set of tracks and have them delivered in the form of a compressed archive (ZIP) file containing the corresponding WIG files, together with a 'read me' file with further details about the samples. Track data for Epigenomics is also available for download via an anonymous FTP site (ftp://ftp.ncbi.nih.gov/epigenomics/).

CONCLUSION

The Epigenomics database at NCBI has been established to serve as a comprehensive public resource for epigenetic and epigenomic data sets. Data are being collected from several large scale projects, including the NIH Roadmap Epigenomics project, ENCODE and modENCODE projects as well as from smaller single laboratory studies. With interest in the field of epigenomics expanding and the amount of data increasing dramatically, it is important that this information be readily available and easily accessed by all members of the scientific community. Epigenomics introduces the Sample Browser, a new and unique tool at NCBI, which provides an intuitive interface to the database. We hope to provide users with all levels of knowledge and expertise in the field of epigenetics the ability to examine and analyze these data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge and thank Tanya Barrett and Alexandra Soboleva for their input into establishing the Epigenomics resource. The authors would also like to thank members of the NCBI GEO, SRA and Seq-viewer teams for contributing resources to the Epigenomics database.

FUNDING

Funding for open access charge: The Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Waddington,C.H. (1942) The epigenotype. *Endeavour*, **1**, 18–20.
2. Berger,S.L., Kouzarides,T., Shiekhattar,R. and Shilatifard,A. (2009) An operational definition of epigenetics. *Genes Dev.*, **23**, 781–783.
3. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
4. Ng,R.K. and Gurdon,J.B. (2008) Epigenetic inheritance of cell differentiation status. *Cell Cycle*, **7**, 1173–1177.
5. Probst,A.V., Dunleavy,E. and Almouzni,G. (2009) Epigenetic inheritance during the cell cycle. *Nature Rev. Mol. Cell Biol.*, **10**, 192–206.
6. Roloff,T.C. and Nuber,U.A. (2005) Chromatin, epigenetics and stem cells. *Eur. J. Cell Biol.*, **84**, 123–135.
7. Schneider,R., Bannister,A.J. and Kouzarides,T. (2002) Unsafe SETs: histone lysine methyltransferases and cancer. *Trends Biochem. Sci.*, **27**, 396–402.
8. Esteller,M. (2008) Epigenetics in cancer. *N. Engl. J. Med.*, **358**, 1148–1159.
9. Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat. Rev. Cancer*, **4**, 143–153.
10. O'Neill,L.P. and Turner,B.M. (1996) Immunoprecipitation of chromatin. *Methods Enzymol.*, **274**, 189–197.
11. Braunstein,M., Rose,A.B., Holmes,S.G., Allis,C.D. and Broach,J.R. (1993) Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Gene Dev.*, **7**, 592–604.
12. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–619.
13. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
14. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
15. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for functional genomics data sets - 10 years on. *Nucleic Acids Res.*, **37**, D885–890.
16. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
17. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
18. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.