# Institutional Profile

# The Einstein Center for Epigenomics: studying the role of epigenomic dysregulation in human disease

There is increasing interest in the role of epigenetic and transcriptional dysregulation in the pathogenesis of a range of human diseases, not just in the best-studied example of cancer. It is, however, quite difficult for an individual investigator to perform these studies, as they involve genome-wide molecular assays combined with sophisticated computational analytical approaches of very large datasets that may be generated from various resources and technologies. In 2008, the Albert Einstein College of Medicine in New York, USA established a Center for Epigenomics to facilitate the research programs of its investigators, providing shared resources for genome-wide assays and for data analysis. As a result, several avenues of research are now expanding, with cancer epigenomics being complemented by studies of the epigenomics of infectious disease and a neuroepigenomics program.

**KEYWORDS: bioinformatics ■ chromatin ■ cytosine methylation ■ epigenetics**

**Andrew S McLellan[1], Robert A Dubin[1], Qiang Jing[1], Shahina B Maqbool[1], Raul Olea[1], Gael Westby[1], Pilib Ó Broin[1,2], Melissa J Fazzari[1], Deyou Zheng[1], Masako Suzuki[1] & John M Greally[1†]**

*[†]Author for correspondence:*
*[1]Albert Einstein College of Medicine, Price Center for Genetics and Translational Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA*
*Tel.: +1 718 678 1234*
*Fax: +1 718 678 1016*
*john.greally@einstein.yu.edu*
*[2]National Center for Biomedical Engineering Science, NUI, Galway, University Road, Galway, Ireland*

While mutation of a gene is an obvious means of causing cellular pathology, the abnormal silencing or activation of a gene is another means of inducing similar problems. Transcriptional regulation is influenced not only by DNA sequence, but also by sequence-specific mechanisms (e.g., binding of transcription factors or noncoding RNAs) and relatively sequence-nonspecific mechanisms, such as cytosine methylation, post-translational modifications of histones, nucleosomal positioning, and probably other influences such as histone variants [1]. Since at least some of the latter sequence-nonspecific processes mediate epigenetic regulatory processes such as genomic imprinting and X chromosome inactivation [2,3], they have been referred to as epigenetic regulators, extending the definition to encompass genome-wide studies as 'epigenomic'.

As we have pointed out in a previous review [4], this definition is inaccurate but convenient. There is at present a substantial surge in interest in testing how these epigenetic regulatory mechanisms may contribute to human diseases, for a number of reasons. These include the recognition from gene-expression microarray experiments that transcriptional dysregulation of specific genes appears to characterize certain diseases, prompting investigators to take the next step to understand the regulatory processes influencing these patterns. Furthermore, we now appreciate that epigenetic regulatory processes can be influenced by a number of environmental factors, age and sex, and that diseases other than cancer appear to involve a contribution of epigenetic dysregulation [4].

If an investigator wants to study these epigenomic regulators in their system of interest, they have the choice of focusing on a locus of interest for gene-specific studies, or extending the studies genome-wide. Increasingly, the latter is the preferred primary option for a number of reasons, not only because there may be many loci of interest from the outset, but also because at a specific gene the regulatory sites are probably not restricted to the canonical promoter sequence. As initiatives such as the ENCODE project [5] are beginning to demonstrate the presence of potentially functional elements in nonpromoter sequences and the feasibility of genome-wide assays to

future medicine part of fsg

study epigenetic and transcriptional regulators, the usual strategy is now shifting to an initial 'discovery' approach to define candidate loci at which subsequent targeted, quantitative studies are focused.

However, it is difficult for individual investigators to establish these assays in their own laboratories, in part because the molecular assays are reasonably specialized (e.g., microarray or massively-parallel sequencing [MPS]), but even more due to the difficulty in managing, organizing and analyzing the extremely large datasets resulting from these experiments.

## Einstein Center for Epigenomics

To overcome both of these hurdles, the Albert Einstein College of Medicine (NY, USA) established a Center for Epigenomics during 2008. The Center is located in the institution's new Center for Genetics and Translational Medicine (Figure 1) on the Einstein campus. This building houses not only the 'Epigenomics Shared Facility' (ESF) where microarray and MPS-based assays are performed, but also the Computational and Statistical Epigenomics Group and their high-performance computing resources. The components of the Center for Epigenomics are depicted in Figure 2.

**Figure 1. The Einstein Center for Epigenomics is located in the new Michael F Price Center for Genetics and Translational Medicine located in the Harold and Muriel Block Research Pavilion.** This building brings together investigators in a number of disciplines with the goal of fostering their interactions, with epigenomics being a major emphasis.

The organization of the Center is designed to facilitate both the ability to perform experiments and the interactions between the scientists involved. By performing assays through a shared facility, from which the data flow to servers where analyses are performed, we move the experimental burden from the investigator to the shared facilities, thus allowing the investigator to focus on the biological aspects of their projects.

## Epigenomics shared facility

This shared facility is staffed by one PhD-level junior faculty member and two MS-level research assistants. The technologies used at present include Roche-NimbleGen (WI, USA) long oligonucleotide microarrays, and both Illumina GA II (Solexa, CA, USA) and Roche FLX (454) MPS technologies. The epigenomic assays offered by the ESF have been rolled out in a sequential manner, with the early priority placed on chromatin immunoprecipitation (ChIP) studies, both microarray-based (ChIP-chip) and MPS-based (ChIP-seq), cytosine methylation assays (microarray or MPS-based HELP [6]) and MPS-based microRNA assays (miRNA-seq). Samples are submitted following information capture through a web-based laboratory information management system (LIMS), with capture of experimental information in the ESF to the same LIMS. Following completion of the assay, the data quality is assessed by ESF staff and the datasets transferred to the Epigenomics high-performance computing (HPC) resource. Data captured from each stage of the process, through sample submission, processing and analysis, are stored in a MySQL database.

## Virtualization of the center environment

As the Center for Epigenomics is not a physical entity, but includes investigators from the several Einstein campuses and external collaborators, we rely heavily on web-based resources to allow the interactions necessary for a productive community of researchers. As interactions between investigators, ESF and computational epigenomics personnel are encouraged, we maintain an email LISTSERV to announce upcoming meetings, seminars, tutorials or other opportunities for

interaction. Protocols, sample submission guidelines, contact information and analytical resources are available as web-based resources prior to initiating an experiment.

The sample submission system designed for the ESF consists of an integrated semantic wiki and web application environment developed using a combination of leading technologies such as Semantic MediaWiki, Ruby on Rails and AJAX. WikiLIMS, developed by the BioTeam [101] is a LIMS system developed specifically for MPS experiments, which we have further modified with integrated dynamic scripting to allow intelligent experiment-specific capture of sample information and other metadata into the relational database. The user interface allows the investigator to submit sample information easily, to monitor experimental progress in the ESF, and to access results of the experiment, not only data quality metrics, but also links to genome browser tracks or other visualization tools. The wiki environment allows users to create their own pages in a free-form manner, and facilitates the implementation of a sophisticated electronic laboratory notebook. Within their personal wikispace, users can create text, upload images and even add links to data files. We are developing this resource to anticipate greater overall use of electronic laboratory notebooks in coming years, with the epigenomics implementation as

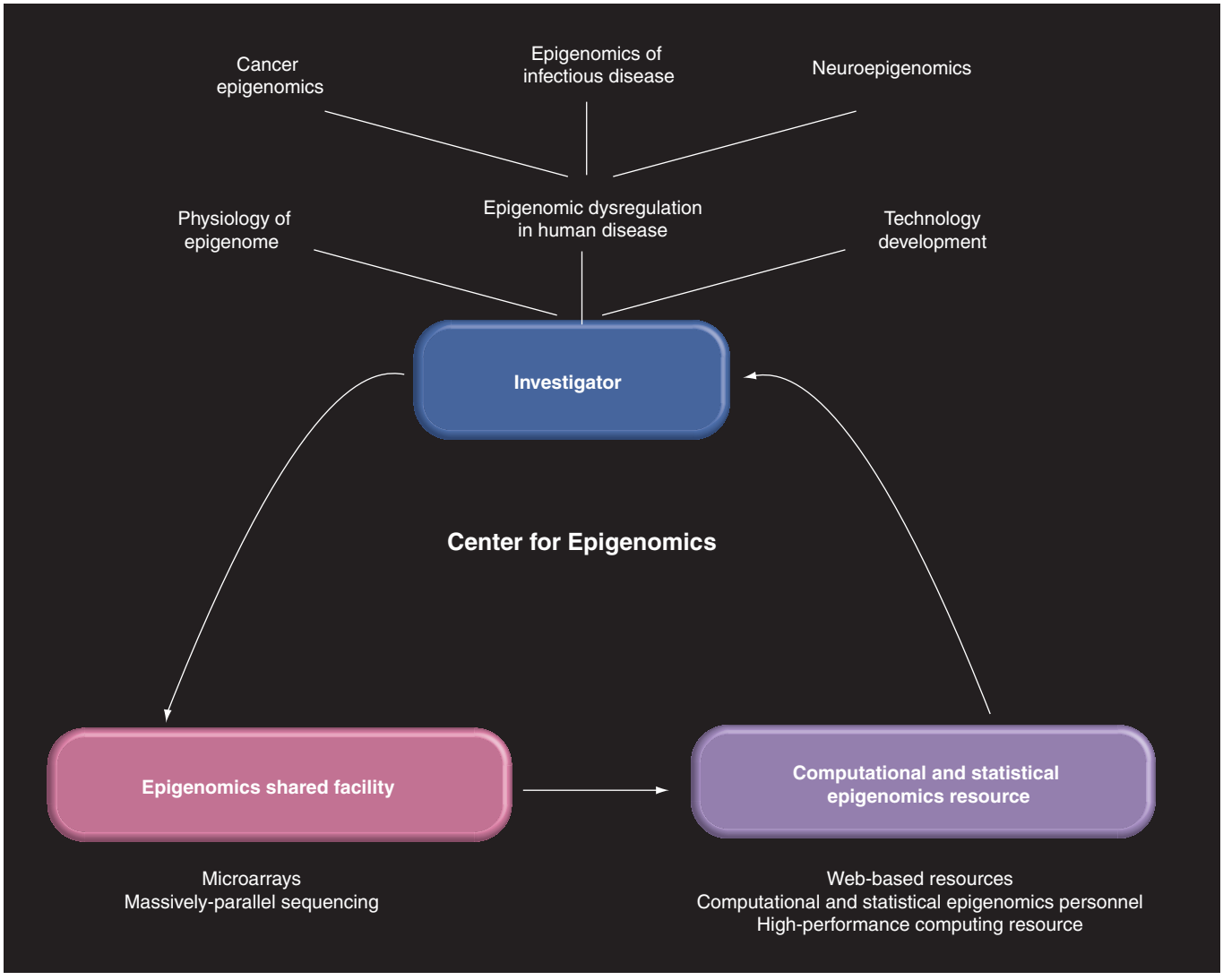


**Figure 2. The organization of Einstein's Center for Epigenomics is designed to foster several avenues of research, of which studies of human diseases is a major emphasis**. By tightly integrating the activities of the data generating and data analytical components of the Center, a streamlined system of sample handling and analysis can be created that allows investigators to perform epigenomics projects that would be otherwise difficult to initiate.

a test case for wider institutional use. We also expect that our wikiLIMS will significantly facilitate data sharing and public data deposition.

The management of epigenomic datasets generated using microarrays or MPS is an increasingly complicated issue. These datasets are based on imaging data at the outset, requiring substantial data storage resources unless strategic decisions are made to discard the primary images, which in turn requires that sufficient information has been extracted from the images prior to their disposal. We maintain a buffer of several months' worth of image data, and store these and derived data within the central HPC resource. This HPC resource includes an Infiniband® connected cluster of 24 dual quad-core AMD Opteron processors with over 650 GB RAM and 10 TB of local storage, coupled with a Sun AMD 8-core server with 100 TB RAID for medium-term storage of experimental data. The cluster is configured with all of the necessary bioinformatic tools to perform varied data analysis, including homology and similarity search, read-mapping, genome assembly, protein and sequence analysis and visualization. This resource provides a single location for both the storage and analysis of data obtained from high-throughput experiments, allowing the computational epigenomicists to perform analysis remotely, avoiding the large volume of network traffic and potential for error inherent in downloading data to their own machines. The ESF also maintains a full mirror of the UCSC genome browser and all associated databases, further reducing the need for transfer of large files across the network. Automated pipeline analyses to process data are placed in a secured production environment, and there is a constant drive to evaluate additional and better methods of analysis. In addition, there is ongoing development in parallel of new analytical approaches within the same computing resource. While we have developed and adopted some computational tools to meet our short-term need for helping investigators to inspect and analyze data from our current experiments (e.g., HELP, ChIP-seq and ChIP-chip), we also commit computational personnel for mid- to long-term software development and support. With time, new versions of the pipeline analytical approaches will be moved to the production environment, and more sophisticated data analysis and integration software is expected to be developed. By encouraging the use of our experimental and computational resources by researchers attempting to develop new data exploration and analytical algorithms, we hope to provide a fertile environment for the ongoing development and implementation of epigenomic analytical tools.

We recognize that the single greatest difficulty with epigenomics research today is the computational and statistical analysis of these datasets, prompting our focus on investing in this specific area as our major priority. We have put in place several dedicated computational researchers and plan to recruit more, so that we can balance our computational support with new research in the computation area. As byproducts of this emphasis, we are creating a substantial virtual environment for the Center, in which we hope researchers and trainees can interact productively, moving their research from a primary molecular focus towards a more bioinformatic emphasis. In addition to software support, we also plan for outreach by training the Einstein research community.

## Research areas

Epigenomics research tends to break down into three major categories: the study of the normal physiology of the epigenome; its dysregulation in disease; and the development of novel technologies with which to study the epigenome. The tools used for each category are molecular assays and bioinformatic and/or statistical analysis. At Einstein, all of these areas are active, but we are attempting to drive development towards two areas of perceived strength – computational epigenomics and studies of human diseases other than cancer. There have been several publications of computational analytical approaches, epigenomic annotations or tools from Einstein researchers [7–11], and with the establishment of Einstein's new Department of Systems and Computational Biology, we hope to continue to build on this foundation by facilitating the work of the expanded group of researchers hired to this department.

While cancer epigenomics has been the major human disease application to date, including for Einstein researchers [12–14], our goal is to promote the study of other diseases that are currently less recognized as involving epigenomic dysregulation. In particular, we would like to develop ongoing work in the epigenomics of infectious disease [15,16] and in neuroepigenomics [17]. It is this desire to facilitate epigenomic studies by researchers for whom there is minimal or no precedent for such studies in their field that has prompted our efforts to create the ESF and analytical resources described above.

## Clinical epigenomics

The purpose of the Center for Epigenomics is, ultimately, to drive the understanding of how epigenetic dysregulation contributes to human disease. At some point, therefore, the work performed will move from animal models or cell lines to clinical samples. This will cause problems, not least of which will be the limited amount of sample from human subjects, and the greater degree of contamination of the desired cell type by admixture with other cells. Miniaturization of assays to allow limited cell numbers to be tested is therefore a major priority, and has been an area of progress for the Center investigators recently [6].

We also face a practical issue of the choice of an epigenomic assay to perform on the clinical samples. While cytosine methylation is only one facet of the epigenome, it is substantially easier to study than chromatin structure or constituents from biopsies, not only because the amount of sample needed often exceeds that possible from minimally invasive biopsy techniques, but mostly because the preparation of the samples for ChIP requires immediate processing through a lengthy series of steps. This is why cytosine methylation remains the best-studied epigenomic regulator in human disease studies.

The other problem that is emerging is more troublesome. It appears that some disease-associated changes in cytosine methylation are statistically significant, but marginal in degree and potentially heterogeneous across cell populations [18], and may occur at nonpromoter locations [19]. This requires that our epigenomic assays be as comprehensive as possible in terms of genomic coverage, while retaining the ability to discriminate effects that may be occurring in subpopulations of cells. While the quantitative aspects of ChIP-chip and ChIP-seq have not been established and remain in an exploratory phase, there is some recent evidence for MPS-based cytosine methylation assays to be reasonably quantitative (to the point that they may be able to resolve as little as 20% differences in methylation [20]). There is clearly room for improvement as we move epigenomic studies into the clinical arena.

## Future perspective

In the next several years we expect to see a pronounced increase in interest in the study of epigenomic dysregulation in human disease, working back from disease associations to studies of the potential dysregulatory influences, whether toxic, dietary, age-related, infectious or other. The challenge will decreasingly be to perform molecular assays, and increasingly to develop robust means of interpreting the results. While in part this will unavoidably require personnel from molecular laboratories to gain expertise in bioinformatics and statistics, an institution benefits from shared facilities that shoulder this burden. Einstein's Center for Epigenomics is likely to become predominantly a computational and statistical epigenomics resource over time, with the goal of allowing a productive environment for testing and exploring hypotheses about the epigenome in a range of organisms and over a wide spectrum of human diseases.

## Financial & competing interests disclosure

*Pilib Ó Broin acknowledges the support of the Science Foundation Ireland (# RFP/05/CMS001). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.*

*No writing assistance was utilized in the production of this manuscript.*

## Bibliography


1 Bernstein BE, Meissner A, Lander ES: The mammalian epigenome. *Cell* 128(4), 669–681 (2007).

2 Reik W, Walter J: Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* 2(1), 21–32. (2001).

3 Latham KE: X chromosome imprinting and inactivation in preimplantation mammalian embryos. *Trends Genet.* 21(2), 120–127 (2005).

4 Hatchwell E, Greally JM: The potential role of epigenomic dysregulation in complex human disease. *Trends Genet.* 23(11), 588–595 (2007).

5 The ENCODE project consortium: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), 799–816 (2007).

6 Oda M, Glass JL, Thompson RF *et al.*: High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.* (2009).

7 Figueroa ME, Reimers M, Thompson RF *et al.*: An integrative genomic and epigenomic approach for the study of transcriptional regulation. *PLoS ONE* 3(3), e1882 (2008).

8 Glass JL, Thompson RF, Khulan B *et al.*: CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* 35(20), 6798–6807 (2007).

9 Thompson RF, Reimers M, Khulan B *et al.*: An analytical pipeline for genomic representations used for cytosine methylation studies. *Bioinformatics* 24(9), 1161–1167 (2008).

10 Sohal D, Yeatts A, Ye K *et al.*: Meta-analysis of microarray studies reveals a novel hematopoietic progenitor cell signature and demonstrates feasibility of inter-platform data integration. *PLoS ONE* 3(8), e2965 (2008).

11 Thompson RF, Suzuki M, Lau KW, Greally JM: A pipeline for the quantitative analysis of CG dinucleotide methylation using mass spectrometry. *Bioinformatics* 25(17), 2164–2170 (2009).

12 Figueroa ME, Wouters BJ, Skrabanek L *et al.*: Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/t-lymphoid features. *Blood* 113(12), 2795–2804 (2009).

13 Polo JM, Juszczynski P, Monti S *et al.*: Transcriptional signature with differential expression of bcl6 target genes accurately identifies bcl6-dependent diffuse large B cell lymphomas. *Proc. Natl Acad. Sci. USA* 104(9), 3207–3212 (2007).

14 Lopes EC, Valls E, Figueroa ME *et al.*: Kaiso contributes to DNA methylation-dependent silencing of tumor suppressor genes in colon cancer cell lines. *Cancer Res.* 68(18), 7258–7263 (2008).

15 Gissot M, Choi SW, Thompson RF, Greally JM, Kim K: *Toxoplasma gondii* and *Cryptosporidium parvum* lack detectable DNA cytosine methylation. *Eukaryot. Cell* (2008).

16 Gissot M, Kelly KA, Ajioka JW, Greally JM, Kim K: Epigenomic modifications predict active promoters and gene structure in toxoplasma gondii. *PLoS Pathog.* 3(6), E77 (2007).

17 Formisano L, Noh KM, Miyawaki T, Mashiko T, Bennett MV, Zukin RS: Ischemic insults promote epigenetic reprogramming of μ-opioid receptor expression in hippocampal neurons. *Proc. Natl Acad. Sci. USA* 104(10), 4170–4175 (2007).

18 Heijmans BT, Tobi EW, Stein AD *et al.*: Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl Acad. Sci. USA* 105(44), 17046–17049 (2008).

19 Abe M, Ohira M, Kaneda A *et al.*: Cpg island methylator phenotype is a strong determinant of poor prognosis in neuroblastomas. *Cancer Res.* 65(3), 828–834 (2005).

20 Ball MP, Li JB, Gao Y *et al.*: Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* 27(4), 361–368 (2009).


## ■ Website


101 BioTeam website www.bioteam.net/