**REVIEW**

**Computational Analysis of Epigenetic Data**

Lydia Auch | Northeastern University, Boston, Massachusetts

**ABSTRACT**

The field of epigenetics has received massive attention in recent years in an attempt to understand the technicalities of gene expression that are not linked to DNA sequences alone. Two main methods of epigenetic regulation have been determined: DNA methylation and histone modification. DNA modification involves adding or removing methyl groups from DNA to allow transcription of the gene. Histone modifications change the accessibility of DNA strands by changing the tightness of chromatin through the addition or removal of extra molecules on the histone proteins. Novel methods of gathering  epigenetic data has caused an unprecedented influx of information that biologists are unprepared to sift through. These datasets must be well analyzed and stored properly so that they are accessible, generalizable, and presentable to the scientific community as a whole. Through the collaborative efforts of bioinformatics, genetics, and modeling techniques, epigenetic data can be correctly maintained and will provide the most benefit to research in the future.

**INTRODUCTION**

There is a multitude of genetic information that is not explainable by DNA base-pairs alone. Epigenetics helps account for gaps genetics cannot explain such as how DNA is packaged and how cells specialize from a single fertilized egg. Epigenetics is defined as the study of mitoically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence (Bock & Lengauer, 2007). Cellular regulation of self-renewal and differentiation depend heavily on transcriptional networks, but also on the on the epigenetic landscape, chromatin properties, within the cell (Tollervey & Lunyak, 2012). This synergistic epigenetic landscape allows gene expression to be altered without changes to the DNA sequence. Using multiple layers of epigenetics events, adaptable but precise control over regulatory genes is possible to facilitate coordinated changes at defined stages of development (Feil & Fraga, 2012). The field of epigenetics has received massive attention in the past decade and is one of the fastest moving areas in biology (Bock & Lengauer, 2007). This surge of interest gives rise to the challenge of analyzing the massive amount of data gathered. Bioinformatics and computational methods must evolve to make data accessible and relevant to other researchers. This review aims to discuss the rapid progress of the field of epigenetics and the issues that computational biologists are now facing in regards to how to store and maintain the influx of epigenetic data.

**MECHANISMS OF EPIGENETIC REGULATION**

The methods of epigenetic regulation contain many gaps but there are two mechanisms that are fairly well understood: DNA methylation and histone modification. Both of these are important to understanding phenotypic regulation since differentiated cells maintain their phenotypes through cell division. This implies that specialized genes, which determine phenotype, are permanently turned on or off in different types of differentiated cells (Holliday, 2006). Establishment of epigenetic signatures requires modifications to be heritable and needs the coordinated efforts of 'writers', 'erasers', and 'readers' of epigenetic features (Tollervey & Lunyak, 2012). These mechanisms are controlled within the cell but it is now clear that the outside environment can induce DNA methylation patterns and alter histone modifications (Feil & Fraga, 2012).

**DNA Methylation.** DNA methylation patterns are directed and preserved by a family of enzymesknown as DNA methyltransferases (DNMT's). These DNMT's add methyl groups to DNA base pairs to change the expression levels of genes. CpG dinucleotides account for the sites of almost all DNA methylation in mammals. CpG islands, clusters of CpG's, are commonly found near promoters and first exons yet are generally not methylated in the germline and most somatic tissue. Despite the unmethylated state of most CpG islands, investigations of DNA methylation in mammals have focused on these regions (Jones & Takai, 2001). Hypomethylation usually correlates with increased gene expression while increased methylation tends to result in transcriptional repression (Madrigano et al., 2012). DNA methylation was originally thought to be irreversible and stable. But recently, evidence suggests that DNA demethylation can occur passively when DNMT's which control maintenance of methylation are blocked from accessing methylation sites, or actively through selective recruitment of DNA demethylases (Tollervey & Lunyak, 2012).

**Histone Modification.** The packaging of chromosomal DNA is comprised of a basic unit of the DNA/protein complex, called chromatin. This unit, the nucleosome, contains a bead-like structure made up of 8 histone proteins and DNA wrapped around the surface of the histone complex. Depending on how tightly or loosely the nucleosome is packaged, three main states of chromatin can occur: euchromatin, hererochromatin, and bivalent chromatin. Post-translational modifications of histones (histone PTMs) are a key epigenetic mechanism because they facilitate different chromatin states (Tollervey & Lunyak, 2012).  Different PTMs include acetylation, methylation, and phosphorylation among others. Each modification to a histone can induce or inhibit additional modifications. This type of interaction is called cross talk and can operate on one single nucleosome or cause interactions among multiple nucleosomes. Most histone modifications are reversible and are constantly maintained, established, and removed to uphold the needed chromatin state (Feil & Fraga, 2012).

Different histone PTM profiles are associated with different chromosome states. PTMs that are transcriptionally permissive are associated with “open” euchromatin, while repressive PTMs are associated with “silent” heterochromatin. Bivalent chromatin are characterized by the presence of both repressive and activating PTMs which creates a silent but transcriptionally available state (Tollervey & Lunyak, 2012).

**Non-histone proteins.** Another mechanism of epigenetic regulation is through RNA molecules. The mechanisms of this type of regulation are still largely unknown however it is becoming increasingly evident that RNAi pathways and non-coding RNA's have a role in epigenetic states such as transmitting signals between two cells (Holliday, 2006; Tollervey & Lunyak, 2012).  Yet, all of these epigenetic mechanisms are interconnected and work together to create the full epigenetic landscape.

**PROCESSES FOR GATHERING DATA**

Research on these methods of epigenetic regulation increased rapidly in part due to the expanding use of next-generation DNA sequencing technologies. Precise mapping of epigenetic data is necessary to understanding the complex layers of regulation. DNA samples are gathered from a collection of cells. Therefore, the methylation pattern on each cells' DNA will vary.  Additionally, methylated cytosine is not readily distinguished from unmethylated cytosine (Laird, 2010). Some methods of genetic sequencing have been adapted or combined for use in gathering epigenetic data while other methods have been created in response to insufficiencies by using enzyme-based and affinity-based methods.

**Methylation-Specific PCR.** Methylation-Specific PCR (MSP) is a way of mapping DNA methylation patterns in CpG islands. MSP assesses the methylation status of CpG sites without needed to know the levels of methylation specific enzymes.  Previous PCR based approaches to measuring methylation patterns often gave false positives due to their reliance on restriction enzyme cleavage to distinguishmethylated and unmethylated DNA (Herman et al., 1996). However, the use of MSP is decreasing due to its reliance on gel-electrophoresis and inherent labor-intensive method (Laird, 2010).

**Bisulphite conversion.** Treating denatured genomic DNA with sodium bisulphite chemically deaminates unmethylated cytosine more rapidly than unmethylated cytosine. This treatment converts unmethylated Cs to Ts, essentially turning an epigenetic difference into a genetic difference (Smallwood et al., 2014). This then allows for analysis of DNA using many more techniques such as Sanger sequencing of PCR products (Laird, 2010). Biosulphite sequencing produces helpful base-pair resolution DNA methylation data and is well suited for further sequencing approaches. However, it is less easily adapted to array hybridization because biosulphite treatment reduces sequence complexity, increasing sequence redundancy which conflicts with standard array hybridization techniques (Laird, 2010).

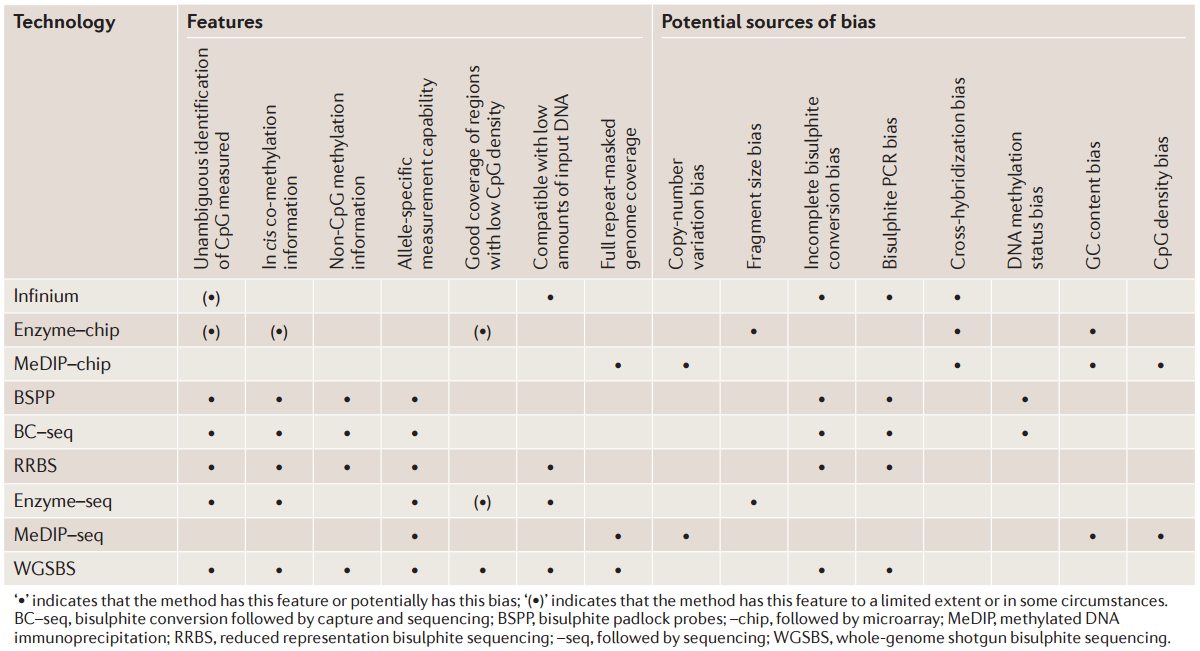
**Affinity Enrichment.** Affinity enrichment of methylated regions by using methyl-binding proteins that have a preference for methylated DNA is proving to be a very powerful tool for profiling DNA. Chromatin immunoprecipitation of denatured DNA with an antibody specific for methylated cytosine is followed by either microarray hybridization (ChIP-chip) or next-generation sequencing (ChIP-seq). Approaches using higher affinity methyl-binding proteins and protein complexes through the methylated CpG island recovery assay (MIRA) have been gaining popularity (Laird, 2010; Colyer, Dellett, & Mills, 2012). Affinity enrichment methods label the input DNA and the enriched DNA with different fluorescent dyes to determine the different strands. This process of gathering epigenetic data is useful for rapid genome-wide assessment but it does not offer information on individual CpG dinucleotides (Laird, 2010).

Table 1 | Features and sources of bias for various techniques

**Endonuclease Digestion.** Endonuclease digestion relies on the fact that each sequence-specific restriction enzyme is accompanied by a DNMT. The DNA methyltransferases protect endogenous DNA by methylating bases in the recognition site. The cutting patterns of restriction enzymes can shine light on DNA methylation because some restriction enzymes are inhibited by methylated cytosine (Mao & Chou, 2010). Methylation sensitive restriction often causes incomplete digestion for reasons other than DNA methylation yet the technique is extremely prone to false-positive results (Kaput & Sneider, 1997;  Laird, 2010).

The afore mentioned processes for gathering epigenetic data are not necessarily the most important or most popular. Yet they each offer novel methods of mapping epigenetic patterns. A comparison of additional technologies features and potential areas of bias is shown in Table 1 (Laird, 2010). The techniques by which data is gathered is less important than how that data is controlled and stored since one universal procedure is impossible and even undesired.

**DATA ANALYSIS**

The techniques introduced above are rapidly producing large volumes of data yet there is insufficient forethought about what that data will be used for and how it will be analyzed (Cushman, 2014). Often, researchers have the intentions to compare data that must be highly manipulated to provide insights.

For example, in many techniques for determining DNA methylation profiles, measurements represent absolute measurements for a given sample, unlike gene expression measurements which find a differential comparison between samples. A measurement scale for methylation of molecules is then represented in the form of the ratio M/(M+U), where M represents the signal for methylated molecules and U represents the signal for unmethylated molecules (Laird, 2010). This is different than the infinite scale used in gene expression array analysis which is more influenced by the variance of measurements with a mean near to the middle of the range. Because of the resulting bias towards features with mean methylation in the middle of the range, selecting probes with high standard deviation and thus reducing the number of features is a common practice (Laird, 2010). However, the distance metrics used to compare measurements across samples is not consistent and this inconsistency needs to be given careful consideration when selecting metrics.

As referenced above, the total amount of CpG methylation can differ significantly among samples. When normalization methods are applied to the gathered data, data loss can occur because the techniques assume similar total signals across samples, resulting in ignoring important biological information (Laird, 2010). Normalization algorithms must therefore be improved or altered before applying them to epigenetic data. One such algorithm is called Functional Epigenetic Modules (FEM) and is used to retrieve epigenetically deregulated genes. This algorithm integrates the analysis of Illumina generated DNA methylation data and matched or unmatched gene expression data (Jiao, Widschwendter, & Teschendorff, 2014). FEM is one example of an analysis package that can identify new key genes or unfamiliar signaling pathways through underlying epigenetic mechanisms.

Understanding the issues that are faced when doing statistical analysis on epigenetic data or when attempting to integrate different forms of data is the first step in creating tools that are useful to a growing number of research areas. The power of supervised algorithms in biological investigations has been previously demonstrated (Chuang et al., 2007; Zhang et al., 2007; ENCODE Project Consortium, 2007) and continued connections between bioinformatics and epigenetics is necessary to represent data in useful and meaningful ways.

**CONTROL OF EPIGENETIC DATA**

Depending on what question one is trying to answer, the data can be represented in numerous ways.

In the same way that there are countless techniques for gathering epigenetic data, it is also true that there is no standard or widely used method of storing or representing the epigenetic measurements taken. The sheer vastness of data is overwhelming and focusing on data storage and curation can cause researchers to lose sight of the original intentions for the data – more data does not necessarily lead to better understanding (Cushman, 2014).

Development of computational resources for epigenetic data analysis has accelerated rapidly. However creating a system for bioinformatic analysis is a complex task due to the considerations needed for each data gathering technique. Concerns for sequence-based analyses include alignment to a reference genome, read counts, and collapsing of clonal reads (Laird, 2010). Yet, some data is well maintained thanks to massive efforts of the biological community. Studies developing new epigenetic data often use well-documented, normalized data from sources such as The Cancer Genome Atlas data portal (Chiappenelli et al., 2015). The issue with using large databases is that it can be difficult to know what data to ignore or set aside when it is not relevant to the analyses at hand.

The intersection of genomic and epigenomic datasets, controlled experimentation, and simulation modeling provides many challenges but also the most viable place to find a solution to data control. Data by itself is not helpful or informative; experiments alone are not generalizable to the rest of the scientific community; models without data are not compelling. Epigenomic datasets will be the most useful when they allow research to form a priori hypotheses, link to controlled and replicated experiments, and integrate powerful modeling techniques to make the information accessible (Cushman, 2014). Results and their analyses need to be stored in a way that they can be generalized to explore their implications across scales of biological organization.

**CONCLUSIONS – DIRECTIONS FOR THE FUTURE**

The broad nature of epigenetics offers the unique opportunity to influence many subsections of biology and other life sciences.  Because the implications for the future of this field are unknown, research will go wherever biologists find passion. One area that will most likely continue to see an increase of attention is in the ways that environmental factors can affect gene expression. To conclude that the epigenome is altered because of the environment itself is difficult. Utilizing genome-wide approaches to computation, will aid in defining loci that are susceptible to environmental impacts, therefore aiding in research connected to diet such as obesity and diabetes. Validating epigenetic changes in large-scale human studies over wide periods of time could help monitor the effects of extrinsic factors (Feil & Fraga, 2012). The ability to monitor epigenetic changes over large populations would also aid in determining epigenetic mechanisms that play a role in cancer, mental disorders, autoimmune diseases, and protein pathways (Bock & Lengauer, 2007). However if the data is not properly stored and categorized, collaboration on efforts such as these would be for naught.

Whatever direction biological research takes, it is clear that computational studies will broaden and deepen. It will be increasingly possible to use epigenetic data analysis to reverse engineer genetic regulation leading to advances in controlling cell fate. The decreasing cost of epigenome mapping will allow further studies into quantitative analysis of epigenetic variation in human populations, potentially due to environmental factors. Additionally, controlling quality of epigenetic data can add evolutionary perspective through mapping the epigenomes of multiple species. It is expected that comparative epigenomics will improve researchers ability to identify functionally important sites in human genomes (Bock & Lengauer, 2007).

The deluge of information facing the biology and epigenetic communities needs to be addressed with a coordinated effort between national and international bodies. Data production has seen massive innovation and the bottle neck is shifting away from generating data and towards analyzing data. The development of powerful and easily usable programs for modeling and analysis of epigenomes is extremely important to providing platforms for scientists, without strong mathematics or programming backgrounds, the ability to present data in useful ways.

Genomics, epigenomics, bioinformatics, modeling, and experimentation are all drawing closer together and when research occurs at the intersection of these fields, solutions to biological mysteries will be uncovered.

**ACKNOWLEDGEMENTS**

This review would not be possible without the diligent comments provided by Sourabh and Darnell along with the impeccable teaching prowess of Tom.

**REFERENCES**

Bock, C., & Lengauer, T. (2007). Computational epigenetics. *Bioinformatics*,24(1), 1-10. doi:10.1093/bioinformatics/btm546

Chiappinelli, K., Strissel, P., Desrichard, A., Li, H., Henke, C., Akman, B., . . . Strick, R. (2015). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell*, 162(5), 974-986. doi:10.1016/j.cell.2015.07.011

Chuang, H., Lee, E., Liu, Y., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol Molecular Systems Biology*, 3. doi:10.1038/msb4100180

Colyer, H. A., Dellett, M., & Mills, K. I. (2012). Detecting DNA Methylation Using the Methylated CpG Island Amplification and Microarray Technique. *Methods in Molecular Biology Cancer Epigenetics*, 329-339. doi:10.1007/978-1-61779-612-8\_21

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 799–816.

Feil, R., & Fraga, M. F. (2012). Epigenetics and the environment: Emerging patterns and implications. *Nat Rev Genet Nature Reviews Genetics*. doi:10.1038/nrg3142

Herman, J. G., Graff, J. R., Myohanen, S., Nelkin, B. D., & Baylin, S. B. (1996). Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences,* 93(18), 9821-9826. doi:10.1073/pnas.93.18.9821

Holliday, R. (2006). Epigenetics: A Historical Overview. *Epigenetics*, 1(2), 76-80. doi:10.4161/epi.1.2.2762

Jiao, Y., Widschwendter, M., & Teschendorff, A. E. (2014). A systems-level integrative framework for genome- wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, 30(16), 2360- 2366. doi:10.1093/bioinformatics/btu316

Jones, P., & Takai, D. (2001). The Role of DNA Methylation in Mammalian Epigenetics. *Science*, 293(5532), 1068-1070. doi:10.1126/science.1063852

Kaput, J., & Sneider, T. W. (1979). Methylation of somatic vs germ cell DNAs analyzed by restriction endonuclease digestions*. Nucl Acids Res Nucleic Acids Research,* 7(8), 2303-2322. doi:10.1093/nar/7.8.2303

Madrigano, J., Baccarelli, A. A., Mittleman, M. A., Sparrow, D., Vokonas, P. S., Tarantini, L., & Schwartz, J. (2012). Aging and epigenetics: Longitudinal changes in gene-specific DNA methylation. *Epigenetics*, 7(1), 63-70. doi:10.4161/epi.7.1.18749

Mao, R., & Chou, L. (2010). Methylation Analysis by Restriction Endonuclease Digestion and Real-Time PCR. *Clinical Chemistr*y, 56(7), 1050-1052. doi:10.1373/clinchem.2010.146654

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., . . . Kelsey, G. (2014). Single- cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods Nat Meth*, 11(8), 817-820. doi:10.1038/nmeth.3035

Tollervey, J. R., & Lunyak, V. V. (2012). Epigenetics. *Epigenetics*, 7(8), 823-840. doi:10.4161/epi.21141

Zhang, Z. D., Paccanaro, A., Fu, Y., Weissman, S., Weng, Z., Chang, J., . . . Gerstein, M. B. (2007). Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Research,* 17(6), 787-797. doi:10.1101/gr.5573107