Published in final edited form as:

Transl Res. 2015 January ; 165(1): 7–11. doi:10.1016/j.trsl.2014.09.011.

Translational Research Epigenomics

Joseph M. Replogle and Philip L. De Jager

This issue of *Translational Research* features articles reviewing the progress and promise of epigenomics in the context of human health and disease. These articles provide examples of epigenomics, the study of genomic modifications causing and maintaining heritable changes in gene expression that cannot be attributed to changes in the primary DNA sequence, in a wide range of disorders from cancer (Nickel et al., Figueroa et al., Costa et al. Stadler et al., Langevin et al., Kishi et al.) to neurodegenerative (Bennett et al.) and metabolic (Evans-Molina et al.) diseases. This diversity in diseases highlights the breadth of the potential for clinical applications of epigenomics. At their most basic level, epigenomic studies help to elucidate disease etiology and pathogenesis. Building on this foundation, epigenomic insights can guide the development of diagnostic and prognostic tools. As epigenetic marks can be responsive to the environment, there is a lot of interest in their potential role mediators of the effect of non-genetic risk factors for disease; these mechanistic insights into the consequences of environmental and other risk factors may provide targets for drug development. Further, (Arnett et al.) the cell-type specificity of epigenomic marks suggests that drugs that specifically target diseased epigenomic states, such as histone deacetylase (HDAC) and DNA methyltransferase (DNMT) inhibitors, may be useful in the context of cancers and inflammatory diseases (Lopez et al.). Finally, tools arising from engineered epigenomic states, such as induced pluripotent stem cells, hold potential to fundamentally alter drug testing, disease modeling, tissue repair, and transplantation (Kobayashi et al.).

Translational epigenomics ultimately seeks to leverage associations between epigenomic marks and clinical outcomes. This field is still in its infancy and will require parallel efforts to (1) improve and reduce the cost of epigenotyping technologies, (2) develop new analytic methods, and (3) establish the fundamental lexicon that relates epigenomic marks to one another and establishes functional units for each mark. All three efforts have recently accelerated thanks to large projects such as the Encyclopedia of DNA Elements (ENCODE) (<https://www.encodeproject.org>) and the National Institutes of Health's Roadmap Epigenomics Project (<http://www.roadmapepigenomics.org>); however, much remains to be done before a large-scale epigenome-wide association study (EWAS) become an approach that is not limited to a small number of specialized laboratories. Also, while these large public projects have generated tremendous resources, they have sampled only a relatively modest number of individuals, cell types and particularly cell states: the extent of

Correspondence Philip L. De Jager, M.D., Associate Professor of Neurology, Department of Neurology, Brigham and Women's Hospital, pdejager@rics.bwh.harvard.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

interindividual variation in the landscape of healthy profiles (particularly at the extremes of age) remains poorly understood and diseased epigenomic states are only beginning to be sampled. In this commentary, we discuss the methodological insights gained from previous epigenetic and genetic studies, particularly EWAS, genome-wide association studies (GWAS), and expression quantitative trait locus (eQTL) studies, in the hope that future studies will translate into novel disease insights and therapeutics.

Mechanisms and Dimensions of Epigenetic Regulation

Epigenetic regulation provides an essential and complex step between genetic information and the diverse spectrum of cellular phenotypes observed within an individual. Therefore, human cells employ multiple mechanisms of epigenetic control in order to regulate differentiation and maintain phenotypic stability (Dressler et al.). At the level of DNA nucleotides, cells directly methylate or hydroxymethylate cytosine residues, predominantly at cytosine-guanine dinucleotides (CpGs) (Barreiro et al.). Additionally, cells covalently modify histones, the alkaline proteins that interact with DNA to assemble nucleosomes. Combinations of post-translational amino acid modifications of histones, including methylation, acetylation, phosphorylation, ubiquitination, and citrullination, code for specific changes in transcription, DNA repair, and other cellular processes. These basic epigenetic modifications interact with ATP-dependent nucleosome remodeling enzymes, transcription factor binding, and scaffold proteins to influence higher-level nucleosome positioning and chromatin architecture. Finally, small and large non-coding RNAs play roles in epigenomic control of transcription, and post-transcriptional chemical modifications alter messenger and non-coding RNA functions (Liu et al.). All of these epigenetic states may vary over many dimensions, including age, cell type, and environmental stimulation (Nilsson et al.), and modulate transcription. They are thus relevant to the study of disease susceptibility and pathogenesis. However, the feasibility of high-throughput, genome-wide profiling is limited for many marks because of current technologies which make scaling to study hundreds or thousands of samples difficult. More suitable for EWAS currently, CpG methylation can be profiled genome-wide using bisulphite treatment, which converts unmethylated cytosine to uracil without affecting methylcytosine residues, followed by sequencing or a high-throughput automated epigenotyping platforms.

Such technology is not widely available for the histone marks that have been a primary focus of many genome-wide reference maps of epigenetic information. Generally, chromatin immunoprecipitation, which uses antibodies to precipitate modified histones or chromatin proteins covalently bound to DNA, followed by DNA sequencing (ChIP-seq) can be used to provide a genome-wide profile of a chromatin mark. However, for a single disease, it is often unclear which chromatin mark might influence susceptibility and progression. Additionally, many marks must be evaluated using a combinatorial framework in order to understand their function at a genomic locus because marks act cooperatively in order to regulate transcription. Therefore, in order to characterize the effect of chromatin marks on disease, many ChIP-seq experiments must be performed on each sample, and EWAS of chromatin marks with a the necessary sample size are difficult and costly today. Similarly, studies of chromatin conformation, DNAase I hypersensitivity, and nucleosome positioning may inform transcriptional regulation and ultimately provide insights into disease

susceptibility, but current technologies have limited their application on larger scale. Finally, expression of noncoding RNA can be assayed genome-wide using RNA sequencing technologies, and these studies generally employ statistical techniques and experimental designs originally implemented in studies of mRNA variation. For the remainder of this commentary, we focus primarily on the application of methodological insights from published epigenomic and genetic studies as they relate to implementing future EWAS.

Lessons from GWAS

GWAS correlate variation in DNA sequence with common, polygenic traits such as susceptibility to Alzheimer's disease and diabetes. In the last decade, GWAS have unveiled thousands of genetic loci associated with human phenotypes.¹ Nonetheless, a majority of the genetically driven variance of disease susceptibility probably remains to be discovered, and characterizing epigenetic elements that modulate disease susceptibility and progression promises to provide new mechanistic insights and therapeutic targets. While genetic variation may drive epigenomic variation related to disease in certain cases, recent EWAS suggest that the effect of both types of variation may be largely independent.² Studies of methylation patterns have begun to unveil new loci and mechanisms associated with common diseases, but future epigenomic studies will benefit from the issues addressed by the earlier generation of GWAS.³⁻⁵

GWAS provide an initial framework with which to guide the statistical considerations and study design for EWAS. In determining the sample size necessary for a GWAS, generally two parameters must be estimated: the frequency of the variant in the study sample and the effect size of the variant on the phenotype of interest. In the case of EWASs, sources of biological and measurement variability must also be considered in power calculations in addition to the effect size: in particular, power will be very dependent on the proportion of cells within the profiled sample that are in an altered state relative to disease. If only a small proportion of cells are in the altered state, very large sample sizes will be necessary to find robust associations; this echoes the large sample sizes required to find lower frequency variants of moderate or modest effect that have minor allele frequencies < 0.05 . Luckily, while mean differences in methylation level between case and control subjects at a given CpG in a recent EWAS for Alzheimer's disease (AD) were small at $\sim 1\%$, the associated CpGs' effect size was substantially higher than those of typical common genetic variants: the average CpG explained an average 5% of the variance in AD susceptibility, which compares to $< 1\%$ for genetic variants.² That study suggests that sample sizes of 500-1000 subjects may be a reasonable target study design for certain EWAS; however, this estimate is likely to be dependent on the trait of interest and the tissue or cell type being sampled. Overall, the variability in methylated regions may be more important than the absolute methylation levels⁶, and power calculations that incorporate different effect sizes, and variance in methylation level will play a crucial role in designing future EWASs that are well powered to detect true positives and eschew false positives.

A related issue in determining power is understanding the number of independent hypotheses being tested in an EWAS. In GWAS, we understand the correlation structure of the data: linkage disequilibrium (LD) exists among SNPs and makes correction for every

single common genetic variant excessive in a GWAS. An estimate of ~1 million independent common genetic effects led to the determination of a genome-wide threshold of significance for GWAS at $p=5 \times 10^{-8}$ given an $\alpha=0.05$. At this point, we have not yet clearly delineated relationships between neighboring CpGs or other chromatin marks across individuals, leading to the implementation of relatively safe but overly conservative strategies for accounting for the testing of multiple hypotheses, such as Bonferroni corrections.² However, with the availability of large datasets of epigenotype data, we can begin to define empirically driven units of methylation or “methylation blocks” or mBlocks following the terminology of the linkage disequilibrium literature (De Jager unpublished). Recent attempts to evaluate this structure are beginning to be reported⁷ but are limited by the available technologies and sample only a small fraction of potentially methylated CpGs. Ultimately, comprehensive genome-wide DNA methylation profiles need to be generated in large numbers of individuals, similar to the Haplotye Map effort⁸, so that comprehensive correlation maps can be developed and guide the development of analysis methods and technological platforms. The corollary is that such maps should also enable the implementation of imputation strategies for EWAS data.

In order to further limit the error rate, association studies must also follow standard sound principles of study design to avoid and correct for confounding variables that lead to spurious associations. Factors to consider including population stratification, the systematic ancestry differences between cases and controls that could drive spurious genetically-driven differences in methylation level.⁹ Also, samples must be processed and analyzed in a manner that limits batch effects so that technical biases can be separated from biological differences. Use of surrogate variables that empirically capture structure in the data to adjust for known and unknown confounders is a reasonable strategy, but one should remain cautious about using unannotated surrogate variables that could well capture aspects of the disease or trait being studied.

Lessons from eQTL Studies

Although EWAS resemble GWAS in their statistical considerations, genetic studies cannot inform the dimensionality of EWAS. Except in the cases of cancer cells, germ cells, somatic recombination, and sporadic mutation, genetic variants are considered to be largely constant within an individual’s cells. On the other hand, transcriptional states vary widely to produce the diversity of cellular phenotypes and behaviors observed within the human body. As epigenomic states are integral to the regulation of transcription, transcriptomic studies offer insights into considerations for robust epigenomic studies. In particular, recent eQTL studies, which correlate genotypes with mRNA levels, have addressed multidimensional transcription regulation in the context of human disease, and these concepts can be extended to inform epigenomic studies. Primarily, eQTL studies have demonstrated the importance of context-specific transcription regulation in disease.

Early eQTL discovery focused primarily on *in vitro* lymphoblastoid cell lines (LCLs) and hematologic cell types. Although GWAS signals were enriched for regulatory variants in these studies, this enrichment was driven primarily by immunity related phenotypes suggesting that studying the cell type implicated in the disease rather than a surrogate cell

type would allow for better characterization of non-immune diseases.^{10,11} Building upon this conclusion, Raj and colleagues identified eQTLs in highly purified monocytes and T cells to highlight that disease and trait-associated cis-eQTLs are more cell-type specific than average cis-eQTLs, and Fairfax and colleagues and Lee and colleagues identified eQTLs in monocytes exposed to inflammatory stimuli to discover examples of disease associated eQTLs that were stimulus-specific.

The trajectory of epigenomic studies resembles the trajectory of eQTL studies: the ENCODE project initially provided a foundation of epigenomic marks in cell lines while the subsequent NIH Roadmap project used tissues and primary cells to identify more context specific alterations in epigenomic marks. Context specific analysis of methylation is likely to be crucial for understanding human disease. Nonetheless, if a crucial cell type is difficult to obtain, such as lung tissue in the case of idiopathic pulmonary fibrosis (Yang et al.), profiling surrogate cell types may still be useful. For diseases where the focal cell type is unclear, variants identified by GWAS can be combined with data from the NIH Roadmap Consortium to nominate relevant cell types and epigenetic marks.^{12,13} Additionally, in cases where multiple cell types may be important for disease, such as Alzheimer's disease (Bennett et al.), profiling tissues, which contain a mixture of cell types, can be useful for eQTL and epigenomic studies. Importantly, the magnitude of cell type heterogeneity varies across tissues as well as disease states and is a critical consideration in study design. However, EWAS are worth conducting even in the absence of clear estimates of the proportion of different cell types in the tissue sample of a given individual: results may indicate which of the constituent cell types plays a role in disease and disease-related change that is independent of cell proportion is discoverable if an adequate sample size is provided. Secondary analyses that include terms for the proportion of constituent cell types derived from the epigenomic profiles offer one strategy to deconvolute the two types of association; however, the results of such analyses must be interpreted cautiously given the rudimentary nature of current cell-specific models and our lack of understanding to which a cell type-specific signature correlates signatures of different states of cell activation that may be relevant to disease. We are beginning to see examples in which studies in peripheral blood cells have elucidated associations with hypertension and cancer despite the cell specificity of epigenetic marks (Friso et al.).

Challenges for EWAS

Epigenomic variation may contribute to the onset of a disease or may be the consequence of disease processes or drugs. Thus, as with transcriptomic studies, a central limitation of EWAS design relates to the interpretation of the results in terms of causality. This limitation of association studies is usually ignored for GWAS given the assumption that, aside from specific sites undergoing somatic recombination and mutation, an individual's complement of genetic variation is largely established at the time of the zygote's formation. Thus, while case-control and cross-sectional studies are highly informative for GWAS, such designs for an EWAS cannot address causality, and longitudinal studies that carefully consider temporality are essential in order to address this point in EWAS. Nonetheless, even a cross-sectional EWAS in tissue samples such as brain that cannot be accessed longitudinally are meaningful as the association of a locus with a trait of interest provides a critical lead for

investigations into disease pathophysiology that will require studies in *in vitro* or *in vivo* model systems to explore the issue of causality and permit the definition of the role of an associated epigenomic variation as a risk factor or a biomarker for disease.

Beyond discovery studies, epigenomic studies offer an important advantage over GWAS: the potential that the epigenomic change determined to be a risk factor for disease can be changed. Many studies have now demonstrated that the epigenome is much more plastic than we originally appreciated and that several known drugs alter the epigenome. However, the challenge here lies in that current drugs have global, pleiotropic effects, and that in some diseases such as in the hemoglobinopathies, it is desirable to only alter the epigenomic and transcriptional landscape of a single gene (Ginder) or at specific sites in the genome. Thus, much remains to be learned in the strategies with which to manipulate the epigenome.

Future outlook

Current studies are beginning to define the considerations necessary for the successful execution of an EWAS that reports robust, reproducible results. A recent example is a pair of independent AD studies that cross-replicated their results despite differences in the definitions of the primary phenotype.^{2,14} Lessons learned in such studies regarding realistic effect sizes as well as refined understanding of the correlation structure of the epigenome are beginning to set the stage for making EWAS a more generic study design, but it will never become as simple as GWAS given the additional parameters that influence epigenomic states. As large-scale high throughput profiling technologies for the epigenome become cheaper, more comprehensive, and miniaturized in terms of cellular material, studies of appropriate size and complexity will be conducted more easily, preferentially targeting the culprit cell type or tissue instead of a tissue of convenience such as whole blood. In addition, these advances will facilitate the implementation of longitudinal studies that can have the opportunity to address the causal role of epigenomic variation.

As in the early days of GWAS, we are now on the cusp of the rapid development of EWAS studies that will offer important new dimensions of information to current genome-wide disease studies that have relied primarily on genetic variation and the measurement of transcriptional products. The epigenome offers an important link that, while plastic, may provide a richer perspective on disease than the snapshot of a transcriptional profile as the epigenome captures not just information on genes that are actively transcribed but also on those that are poised to become transcribed given a specific stimulus and those that are in a conformation that is not accessible to the transcriptional machinery. Nonetheless, it is already clear that the best studies will integrate all three pieces of information to identify the different sources of genomic variation that influence disease and interact to produce changes in transcription that are the proximal functional consequences linking genetic and non-genetic risk factors to disease biology.

This richer perspective will doubtlessly identify more promising targets for disease diagnosis, prognosis, and therapeutic intervention. It will also spur efforts to design molecules or strategies that will be capable of targeted epigenomic modification to reverse the effect of a risk factor or perhaps to block the effect of a genetic risk factor in certain

cases where rendering a genetic variant or a key chromosomal feature inaccessible by promoting the formation of heterochromatin in a targeted manner is a possible strategy (Lopez article).

Acknowledgement

There were no sources of editorial support in the preparation of this manuscript. Both authors have read the journal's authorship agreement and have reviewed and approved the manuscript.

Bibliography

1. Hindorff LA, et al. Potential etiologic and functional implications of genomewide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009; 106:9362–9367. [PubMed: 19474294]
2. De Jager PL, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience*. 2014; 17:1156–1163. [PubMed: 25129075]
3. Rakyan VK, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet*. 2011; 7:e1002300. [PubMed: 21980303]
4. Bock C. Analysing and interpreting DNA methylation data. *Nature reviews. Genetics*. 2012; 13:705–719.
5. Michels KB, et al. Recommendations for the design and analysis of epigenome-wide association studies. *Nature methods*. 2013; 10:949–955. [PubMed: 24076989]
6. Feinberg AP, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. *Science translational medicine*. 2010; 2:49ra67.
7. Liu Y. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet*. 2014; 94:485–495. [PubMed: 24656863]
8. International HapMap C. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–861. [PubMed: 17943122]
9. Barfield RT, et al. Accounting for population stratification in DNA methylation studies. *Genetic epidemiology*. 2014; 38:231–241. [PubMed: 24478250]
10. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010; 6:e1000888. [PubMed: 20369019]
11. Nica AC, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics*. 2010; 6:e1000895. [PubMed: 20369022]
12. Trynka G, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*. 2013; 45:124–130. [PubMed: 23263488]
13. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–1195. [PubMed: 22955828]
14. Lunnon K, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nature neuroscience*. 2014; 17:1164–1170. [PubMed: 25129077]