The role of epigenetics in genetic and environmental epidemiology

Epidemiology is the branch of science that investigates the causes and distribution of disease in populations in order to provide preventative measures and promote human health. The fields of genetic and environmental epidemiology primarily seek to identify genetic and environmental risk factors for disease, respectively. Epigenetics is emerging as an important piece of molecular data to include in these studies because it can provide mechanistic insights into genetic and environmental risk factors for disease, identify potential intervention targets, provide biomarkers of exposure, illuminate gene–environment interactions and help localize disease-relevant genomic regions. Here, we describe the importance of including epigenetics in genetic and environmental epidemiology studies, provide a conceptual framework when considering epigenetic data in population-based studies and touch upon the many challenges that lie ahead.

First draft submitted: 29 July 2015; Accepted for publication: 6 October 2015; Published online: 27 October 2015

Keywords: environmental epidemiology • epidemiology • epigenomics • genetic epidemiology • genomics

Epigenetics, the study of mitotically heritable and reversible molecular information outside of the DNA sequence, plays a critical role in transcription, nuclear organization, genome stability and imprinting. First described by Waddington as the “study of the interactions between genes and their products that bring the phenotype into being,” epigenetics provides the link between genes and phenotype [1]. Genes or environmental exposures alone are not usually sufficient to completely explain a disease process but instead they are one component of a larger complex molecular system that, as a whole, results in the phenotype. Since epigenetics is an integral component of this molecular system, it may play a mechanistic role in the disease process. Indeed, altered epigenetic states have been shown to be associated with a wide range of diseases including cancer [2], cardiovascular [3], aging and neurodegeneration [4], psychiatric [5–7] and diseases of the immune

system [8,9]. Incorporation of epigenetics into the field of epidemiology is likely to help elucidate biological mechanisms for genetic and environmental risk factors for disease. As described in detail below, inclusion of epigenetics may also result in identification of disease-relevant loci and/or exposures that would otherwise be missed via traditional analyses that did not incorporate epigenetic context.

In addition to the mechanistic causal role epigenetics may play in disease etiology, it may have an equal impact in the field of epidemiology by serving as a biomarker of exposure or disease.

Often in environmental epidemiology, studies cannot rigorously assess the influence of exposures on health outcomes due to a lack of reliable, prospectively collected environmental exposure data collected at the appropriate risk timing window. For later onset diseases, establishing prospective cohorts with long

Christine Ladd-Acosta¹
& M Daniele Fallin^{*,2}

¹Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD 21205, USA

²Department of Mental Health, Johns Hopkins School of Public Health, Baltimore, MD 21205, USA

*Author for correspondence:

Tel.: +1 410 955 3908

Fax: +1 410 614 7469

dfallin@jhu.edu

Future
Medicine

part of
fsg

follow-up durations are very costly. Even studies that collect prospective biospecimens and exposure data have concerns about when biosamples are collected, whether they accurately capture exposures with short half-lives or whether they capture biologically effective doses (for exposure measurements based on external data such as geographic information system based air pollution exposure metrics). There is emerging evidence that epigenetic changes are related to specific exposures and that some of these changes may persist over time, suggesting epigenetics may serve as a long-term biomarker of exposure that can surmount some of these limitations. For years, epigenetics has served as a clinical biomarker of disease, with epigenetic profiles being used to aid in cancer diagnosis and prognosis as well as for predicting how a patient will respond to a particular treatment; it is likely that this will extend to other diseases and clinical settings. It is also possible that epigenetic profiles could be used as a metric in clinical trials to assess treatment response.

A framework for incorporating epigenetics into epidemiology

The relationships between genes, environmental exposures, epigenetics and health outcomes can be com-

Figure 1. A framework describing how epigenetics could be involved, at a mechanistic level, in the etiology of disease. (A) Epigenetics may be causally involved in a disease pathway by mediating genetic or environmental risk. (B) Epigenetics could act as a modifier of genetic or environmental risk. (C) Epigenetics provides a biological mechanism to explain how genetic and environmental factors may, in combination, be involved in the disease process.

plex. Therefore, we provide a conceptual framework (Figures 1 & 2) to describe the relationships that can be used to guide epidemiology study designs, define specific testable hypotheses and aid in result interpretation. We view epigenetic data as informative to the field of epidemiology in two major, equally impactful, ways: by implicating a biological mechanism for disease (Figure 1) and by serving as a biomarker of disease or environmental exposure, even if not directly involved in the causal pathway (Figure 2). When designing studies, it is important to consider the nature of the question, that is, it mechanistic or biomarker-focused, since this has implications when considering the most suitable type of tissue and the appropriate timeframe to obtain DNA methylation (DNAm) measurements. We discuss these considerations in more detail below in the section entitled, 'Challenges and unmet needs.' We view the scenarios shown in Figure 1 as mechanistic in nature where epigenetics plays a causal role in the etiology of disease. In contrast, Figure 2 highlights various pathways in which epigenetics can serve as a biomarker of disease or environmental exposure, that is, epigenetic changes are not directly involved in the causal disease pathway. Below we describe each of these pathways, in detail, and provide specific examples to highlight the importance of incorporating this type of information in genetic and environmental epidemiology. Although our examples focus on DNAm, one type of epigenetic mark, the concepts presented also apply to other types of epigenetic data.

Epigenetics as mechanism for genetic etiology of disease

Genome-wide association studies (GWAS) have identified numerous common and rare variants associated with complex human diseases [10]. While some disease-associated genetic variants have obvious functional consequences such as nonsynonymous amino acid changes and truncated proteins, a large portion of the GWAS identified variants are not predicted to result in downstream functional changes at the protein level or are located within introns or intergenic regions. Understanding how these variants may be involved in disease is an important area of active investigation and there is growing evidence that epigenetics may provide a mechanistic link.

First, it is possible that epigenetics may mediate genetic risk, that is, the methylation state of a specific locus is driven by a nearby genetic variant and the differences in DNAm subsequently lead to disease (Figure 1A & Box 1). There is empirical evidence to support this concept. Numerous studies have identified methylation quantitative trait loci (meQTLs), that is, SNPs that control methylation levels at particular

CpG sites [11–17]. Several studies have examined both local and distal SNP–methylation correlations at CpG sites across a range of tissue types [11,17–21]. In general, they have shown that most meQTLs control multiple methylation sites, typically within 3000 base pairs of the genetic variant. However, some meQTLs have also been observed to be strongly associated with DNAm levels at distal CpG sites. To our knowledge no study has reported genetic variants controlling DNAm at non-CpG sites. Using causal inference testing methods, DNAm has specifically been shown to mediate the association between genetic risk variants and inflammatory diseases including rheumatoid arthritis [22] and peanut allergy [23]. It has also been shown to mediate SNP associations with gene expression and insulin secretion in islet cells [24].

Second, as shown in **Figure 1B & Box 2**, genetic variants can be modified by epigenetic marks, that is, the expression of a particular risk variant depends upon the methylation state of that particular allele. GWAS, that examine genotype alone, are likely to miss these loci because affected and unaffected individuals have the exact same genotype. The involvement of epigenetics provides one possible explanation for disease heterogeneity and incomplete penetrance in genetic models. While there are numerous lines of evidence demonstrating epigenetic regulation of allele-specific expression [25–27], there is little empirical evidence highlighting the importance of this mechanistic pathway; this is likely due to technical limitations with detecting allele-specific expression. However, with the recent emergence of affordable RNA-sequencing platforms as well as appropriate analytic methods, we expect more studies will be able to investigate epigenetic modification of genetic variants to determine the role of this pathway in disease etiology [28].

Epigenetics as a mechanism for environmental etiology of disease

The field of environmental epidemiology examines relationships between environmental exposures and health outcomes. Here, we use the term environmental exposure broadly to encompass a wide range of exposures spanning from exogenous environmental toxicants to *in utero* environment conditions – to socioeconomic status – to endogenous exposures, such as hormone levels. Because epigenetic marks regulate cellular plasticity, that is, the ability of a cell to rapidly respond to its environment, and are key regulators of critical cell processes including genome stability and gene transcription, they offer an attractive biological mechanism to explain how environmental exposures can cause disease. For example, as depicted in **Figure 1A & Box 3**, an environmental exposure may

Figure 2. Epigenetics may not play a direct causal role in a disease pathway but instead may serve as a useful biomarker. (A) Epigenetics, as either a cause or consequence of disease, has utility as a clinical biomarker. **(B)** Epigenetics has the potential to serve as a biomarker of environmental exposure, even if it is not part of the disease pathway.

lead to epigenetic changes that subsequently cause disease; we term this type of mechanism epigenetic mediation of an environmental exposure. It is also possible that epigenetic marks may modify, rather than mediate, environmental risk, that is, the response of a particular gene or cell to environmental stimulus is dependent upon the cells underlying epigenetic state (**Figure 1B & Box 4**).

Pivotal experiments in the Agouti mouse model have shown that DNAm in offspring is altered at the A(vy) locus in response to prenatal exposure to genistein and results in differential susceptibility of the pups to obesity in adulthood [29]. Thus, epigenetic changes have been shown to mediate environmental exposures and adult health outcomes. In humans, there is growing evidence of epigenetic alterations due to a wide range of environmental exposures [30]. For example, a recent study that performed a crossover trial of controlled-human exposure among healthy adults showed that exposure to coarse ambient particulate matter decreased DNAm levels, globally, at *Alu* repetitive elements including at the *TLR4* gene and was associated with increased systolic blood pressure [31]. Further mediation analyses in this study revealed that the changes in acute blood pressure were mediated, at least in part, by exposure-related changes in DNAm [31]. This is consistent with global loss of DNAm leading to genomic instability through activation of repetitive elements, such as retrotransposons, microsatellites and *Alu* repeat elements. Loss of DNAm at repetitive elements can lead to their activation and potentially play a causal role in disease. A second study showed loss of DNAm at the *F1* gene mediates the effects of black carbon on fibrinogen protein levels, a biomarker of cardiovascular function. Additionally, they showed DNAm levels at the *ICAM1* locus mediate the relationship between sul-

Box 1.

Epigenetic mediation of disease-related genetic risk factor. At this particular locus, the unaffected individual with the AA genotype (shown in green) is methylated (shown as magenta superscript 'm') at a nearby CpG site and the associated disease gene silenced. In contrast, the affected individual has a TT genotype (shown in red) and is unmethylated at the neighboring CpG site, leading to transcription of the disease gene. The inter-individual difference in epigenetic state at this particular locus, related to differential expression of a nearby disease gene, is driven by genotype differences among the two individuals at a nearby SNP. Thus, in this example, the epigenetic state is mechanistically involved in mediating genetic risk for disease.

fate and ozone exposure and ICAM-1 protein levels, another cardiovascular biomarker [32].

Should these types of biological mechanisms play a role in disease etiology, and as briefly discussed above there is evidence suggesting they do, failure to include epigenetic data in environmental epidemiology studies could result in the inability to detect environmental risk factors, underestimation of the environmental contribution to disease and a missed opportunity to understand exposure biology.

Epigenetics as a biological mechanism for gene–environment interactions

There is emerging evidence that both genes and environmental exposures are likely to contribute, including in combination, to disease risk in the human population. Therefore, numerous epidemiology studies have begun to examine gene–environment interactions (GxE) at both the candidate gene and genome-wide levels. Epigenetics sits at the interface between genes and environment and, thus, provides a potential mechanism to explain how GxE may occur at the molecular level (Figure 1C & Boxes 5 & 6).

One scenario is that the expression of a disease risk variant is dependent upon the DNAm state and the DNAm state is altered in response to a particular environmental exposure (Box 5). An alternative scenario is that epigenetic differences, driven by underlying genetic variation, as opposed to environmental exposures shown in Box 5, affect the ability of a cell

to respond to a particular exposure (Box 6). In both of these examples, risk for disease depends on both genes and exposure and involves an epigenetic mechanism. In fact, recent work by Teh *et al.* showed that 75% of variably methylated regions, that is, regions of the genome where methylation levels were highly variable among 237 neonates examined, could be explained by GxE [33].

Epigenetics as a biomarker of disease

As we describe in detail above, changes in DNAm patterns associated with disease could play a causal role in disease etiology. However, it is also possible that differences in DNAm are a consequence of the disease process itself, a situation in etiologic epidemiology termed 'reverse causality.' Yet, even in this situation, epigenetic marks still serve as important biomarkers of disease and have clinical utility (Figure 2A). Epigenetic patterns have been shown to be useful in cancer detection, prognosis and predicting response to treatment [34]. For example, Cologuard is the first US FDA approved stool DNA screen that uses DNAm levels to inform colorectal cancer diagnoses via inclusion in a composite risk score [35,36]. DNAm measurements at the *GSTPI* promoter region have also been shown to provide a molecular diagnostic tool for prostate cancer; individuals with prostate cancer show increased DNAm levels at this locus in numerous bodily fluids including plasma, serum, urine and postprostatic massage urethral washes [37–40]. Additionally, DNAm patterns at three genes (*GDF15*, *TMEFF2* and *VIM*), measured in urine samples, have been shown to be highly specific and sensitive for detection of bladder cancer [41]. *GSTPI* promoter methylation status has also been shown to predict cancer recurrence among men with prostate cancer [42–44]. Among individuals diagnosed with glioma, increased methylation at the *MGMT* gene promoter has been shown to be associated with longer survival time, thus, highlighting the prognostic utility of epigenetic biomarkers [45]. Finally, epigenetic alterations have also been shown to predict response to cancer treatments including radiation, chemo- and hormone therapy. For example, *MGMT* promoter methylation has been shown to predict response to temozolomide [45,46] and carmustine [47], both chemotherapeutic agents as well as radiotherapy [48] among glioblastoma patients.

Epigenetics as a biomarker of environmental exposure

In addition to the role epigenetics may play in disease processes, it could also serve as a useful and practical biomarker of environmental exposure (Figure 2B) [49]. Should epigenetic patterns provide a reliable measure

of exposure, in lieu of having actual exposure data at the appropriate time window, they could provide a useful measurement tool for epidemiology studies that have difficulty obtaining reliable measures of exposure, enable use of extensive existing biospecimens and inform new study designs, in a cost-effective manner. For example, the field of genetic epidemiology is becoming more interested in assessing genetic risk for disease in the context of environmental heterogeneity, that is, investigating GxE. Tens to hundreds of thousands of biospecimens have been banked and used for GWAS; however, most samples do not have corresponding environmental exposure data readily available. Collection of exposure data typically requires recontacting study participants to ascertain measures of exposure via questionnaire or obtaining new biosamples suitable for direct measurement of a particular exposure. This approach is costly, subject to recall bias for questionnaire-based exposure ascertainment methods and exposure measurements from newly collected biospecimens in an ongoing cohort often do not reflect the disease-relevant exposure window. An epigenetic signature of exposure that reflects prior exposure with good sensitivity and specificity, using existing banked specimens would help to overcome several of these limitations, enabling incorporation of exposure data into epidemiology studies that did not collect the data during initial enrollment that would otherwise be missed opportunities.

There is extensive literature documenting epigenetic changes associated with a wide range of environmental exposures [30]. More recent work has identified DNAm patterns that reflect past exposure including prenatal smoking exposure [50–54], cumulative lifetime tobacco exposure [55] and lifetime measures of lead exposure [56]. Thus, there is empirical evidence to support the concept that epigenetics can serve as a biomarker of environmental exposure.

Challenges

The integration of epigenetics into epidemiology holds considerable promise for furthering our understanding of disease mechanisms and, in addition, for the potential to serve as a useful biomarker. However, as we discuss in detail below, there are many challenges and considerations that need to be taken into account when undertaking these types of studies.

Study design considerations

Tissue type

Studies seeking to investigate the mechanistic role of epigenetics in a particular disease may need to examine epigenetic patterns in the affected tissue itself (Figure 1). While this is often feasible in studies where surgical

excision of the affected tissue is standard of care, for example, cancer, or where the affected tissue is readily accessible, such as blood cell disorders, for many other diseases, for example, psychiatric illnesses, obtaining a biopsy of the affected tissue is not possible. Therefore, these types of diseases often must rely on measuring epigenetic changes in post-mortem tissue samples, which introduces another set of concerns, or in surrogate tissues that can be ethically and conveniently sampled. There is some evidence that surrogate tissues may provide valuable insights into mechanistic-based disease questions when the affected tissue cannot be obtained. For example, Davies *et al.* showed that while absolute magnitude of DNAm at specific loci often differs across brain and blood tissues, interindividual changes in DNAm, which may be related to disease,

Box 2.

Epigenetic modification of a disease-related genetic risk factor. At this particular locus, both individuals have the same genotype (AT) but differ in their allele-specific methylation states, i.e., the first individual is methylated on the T allele and second individual is unmethylated on the T allele. In this particular example, the differential methylation states result in no expression of the T risk variant in individual 1 (with the methylated T allele) and expression of the mutated protein in individual 2 (with the unmethylated T allele).

Genotype	Outcome		
	Affected	Unaffected	Result
Traditional GWAS			
AT	✓	✓	No association
GWAS that accounts for allele methylation state			
^m AT		✓	
A ^m T	✓		Association

Detection of genetic risk factors. Traditional GWAS studies only examine genotype and do not include allele-specific methylation information. Thus, disease-relevant genetic variants that are modified by epigenetics could be missed via traditional GWAS analyses because both the affected and unaffected individuals have the same genotype (AT).

Box 3.

Epigenetic mediation of an environmental risk factor for disease. The model above shows a particular region of the genome, near a disease-relevant gene, that is susceptible to epigenetic changes upon exposure to a particular environmental factor. Individuals that are exposed become hypermethylated relative to unexposed individuals. This change in the methylation state causes differential expression of the disease-related gene.

are maintained across tissues within an individual [57]. A recent genome-scale screen using whole blood-derived DNA samples from individuals with bipolar or schizophrenia revealed psychosis-associated DNAm changes in the promoter region of *ST6GALNAC1*; these DNAm differences were replicated in an independent set of case–control brain tissue samples [58]. Candidate gene-based studies have also detected similar disease-associated effects between tissues. For example, Kaminsky *et al.*, examined DNAm levels in the *HCG9* gene and observed bipolar-associated differences in DNAm across sperm, post-mortem brain and peripheral blood tissues [59]. Additional evidence stems from a recent study showing age-related DNAm modules can be detected across multiple tissues including brain and blood [60]. For multiple types of cancer, disease-associated DNAm changes, initially detected using primary tumor tissue samples, have also been shown to be present in peripheral blood samples [61–63].

More research to comprehensively evaluate the utility of surrogate tissues in epidemiology-based research studies is needed. It is likely that a combination of approaches, including both post-mortem and surrogate tissues, will be needed to address mechanistic questions when affected tissue is not readily accessible. Further, surrogate tissues such as blood or buccal, may be particularly well suited for studies interested in using epigenetics as a biomarker of exposure or

disease since they represent a readily accessible tissue source and have been shown to harbor changes related to disease and environmental exposure [49,58,63]. We have incorporated this important consideration into our conceptual framework (see notation at bottom of Figures 1 & 2).

Timing

Unlike the DNA sequence, epigenetics are potentially dynamic over the life span, at least at some genomic sites. As we have already described, they may be influenced by environmental exposures over time, or by aging itself [60]. To ensure appropriate causal inference, repeated longitudinal epigenetic measurements that examine and compare epigenetic measurements before and after the onset of disease are needed to clarify the role of epigenetics in the disease process. Longitudinal epidemiology studies with repeated biosampling that start prior to the onset of disease can be extremely costly since they require following and sampling a large population repeatedly for decades and in the end a relatively small proportion of all the individuals included in the study will go on to develop the disease of interest. Thus, it will be important to also consider other approaches, such as partnering with bench scientists to utilizing *in vitro* assays and animal models, to help tease apart the molecular mechanisms and circuits with causal involvement in the disease pathway.

Cell composition

Epigenetics is critical to cellular differentiation and maintenance of cellular identity after mitosis; thus, each cell contains cell-type specific epigenetic patterns. There has been considerable concern in the field that disease or exposure-related changes in epigenetic marks may simply reflect interindividual shifts in cell composition. Analytic methods have recently been developed [64–69] and are being implemented across a wide range of epigenetic studies to address this concern. In addition to correcting for potential confounding by cell type, it is also possible that adjustment may simply help remove noise from true disease or exposure signals. Alternatively, several studies have performed cell sorting to isolate a specific cell type and subsequently perform methylation measurements using that specific cell type. This is attractive if the biologically relevant epigenetic changes related to disease are confined to a particular cell type. One major disadvantage to this approach is that it is often unclear which cell type is most likely to be relevant to a particular disease and/or exposure and examining all cell types is not feasible for cost and practical collection reasons.

It is important to understand if disease and/or exposure related differences in DNAm reflect underlying cell composition changes and these results should be interpreted carefully. However, even if DNAm differences are related to shifts in cell proportions, this may be informative for understanding disease, for example, may reveal a cell type that differs between affected and unaffected individuals, or environmental exposure, for example, identify cell type that is particularly sensitive to exposure, processes. Additionally, DNAm can serve as a biomarker for overall cell composition state and may be useful when trying to monitor changes in cell proportions or disease progression.

Measurement technologies

Although there are a wide range of highly sensitive and specific measurement tools available for epigenetic measurements, sample throughput, cost, coverage and collection methods can be barriers to implementation in large-scale epidemiology studies. Comprehensive, site-specific, genome-wide measurement technologies to measure DNAm, such as whole-genome bisulfite sequencing, cost thousands of dollars per sample; thus, it

Box 4.

Epigenetic modification of an environmental risk factor for disease. As depicted above, the ability of a cell to respond to a particular exposure, e.g., stress, may be dependent upon the underlying epigenetic state. Here, the cell shown on the left is methylated in the promoter region of a gene involved in responding to stress; thus, the response gene is silenced in this particular cell and the cell is unable to appropriately respond to stress. In contrast, the cell on the right is unmethylated and the exposure-relevant gene is being actively transcribed. When this cell is exposed to stress the cell is able to respond appropriately.

Box 5.

Epigenetics as a mechanism for gene–environmental interactions (GxE): exposure mediated GxE. The production of a disease-relevant mutant protein may be dependent on both the underlying genetic variant and a particular environmental exposure. Exposure changes the methylation state from methylated to unmethylated and, as a result, the disease-relevant gene is actively transcribed among exposed individuals; the bottom two scenarios shown above. The individual with the C variant makes a protein that is inappropriate, leading to disease. However, the individual that was exposed and harbors an underlying T variant does not go on to produce the protein or have disease because that particular genetic variant results in formation of a stop codon, inducing gene silencing.

is cost prohibitive for population-based studies interested in examining thousands or tens of thousands of samples. To overcome this issue, most epigenetic epidemiology studies to date have used a more cost-effective technology, the Illumina 450K Human Methylation BeadChip, focused on a subset of high-value genomic regions [70,71]. Other array-based or reduced representation sequencing-based technologies are also available [72,73]. Chromatin assays require crosslinking DNA and proteins from fresh tissue samples and a relatively large number of cells. It is often difficult in large-scale epidemiology studies, where samples are primarily collected in a clinical and/or home setting, to perform these types of laboratory techniques immediately after sample collection.

Conclusion & future perspective

Array-based DNAm measurement has now become mainstream and quality control is improving. For large-

scale epigenetic epidemiological studies, this is likely to be a mainstay tool for many years, allowing query of specific CpGs, set of CpGs in pathways or global DNAm. However, as epigenetic measurement technologies increase and the cost of sequencing continues to decrease, it will likely be possible to comprehensively measure and assess epigenetic data, as well as its relationship with genetic and environmental exposure data, at a population level via full genome sequencing. We expect this to provide a wealth of information, but will require continued development of new statistical tools for sequencing-based data. Additionally, as more studies seek to integrate different types of genomic data inclusive of epigenomic, DNA sequencing, metabolomic and transcriptomic data, we will need new methods and frameworks for handling the magnitude and dealing with the multiplicity of these large datasets. As the field moves in this direction, careful consideration of study design, specific research ques-

tion and statistical approaches will be critical for proper interpretation of results.

In addition to developing new statistical tools and methods, considerable advances will likely be made in the next several years to further our understanding, at a foundational level, of the structures and relationships that exist within the epigenome and across different types of genomic data. This type of work has already begun with the ENCODE [74] and the Roadmap Epigenomics [75,76] projects that seek to provide reference maps of epigenetic patterns in primary tissues. In addition to mapping tissue-specific patterns it will also be important to understand interindividual differences within and across tissues as well as over time. Research has also begun to define and understand structural relationships between epigenetic data, such as correlation of DNAm levels at neighboring loci [11,77], relationships

between various types of epigenetic data including DNAm, chromatin modifications and RNA expression [16,18,78] and relationships with DNA sequence. For specific locations in the genome, emerging evidence suggests DNAm at the single CpG level [14,16,17,22,23,79–81] as well as at DNAm clusters [11] is controlled by underlying genetic variation. Environmentally labile regions of the genome have also been identified [49,79], although further studies to characterize specific regions for specific exposures, are ongoing.

Similar to the International HapMap Project [82] and 1000 Genomes Project [83] reference datasets that inform population-specific linkage disequilibrium, provide expectations on variation and enable imputation in genetic epidemiology studies, epigenetic epidemiology needs comprehensive maps of epigenetic data, an understanding of epigenetic relationships and

Box 6.

Epigenetics as a mechanism for gene–environment interactions (GxE): genotype mediated GxE. As shown in the example above, gene expression of an exposure response gene may be driven by underlying genetic variation, i.e., individuals with an 'A' allele express the gene and those with a 'C' allele do not. Upon exposure to a particular environmental factor, individuals that carry the 'A' allele and express the gene are able to appropriately respond to the exposure. However, individuals with the 'C' allele are not able to respond to the exposure because they do not have expression of the exposure response gene.

reliable, predictable epigenetic patterns. As our understanding of these foundational relationships grows and standardization of genomic and epigenomic data becomes the norm, we will likely begin to understand the downstream biological targets of disease-associated SNPs identified in GWAS and biological pathways involved in environmental risk factors for disease.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial inter-

est in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

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

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Executive summary

- Incorporation of epigenetic data in epidemiology may further our understanding of disease mechanisms including epigenetic mediation of genetic or environmental disease risk factors.
- Investigating epigenetics as a modifier of genetic or environmental risk for disease can expand our understanding of disease mechanisms.
- Epigenetics changes provide a potential biological mechanism to explain gene–environment interactions associated with disease.
- Epigenetic data may be used clinically as a biomarker for diagnosis, prognosis or response to treatment and, in addition, it has potential utility as a performance metric in clinical trials.
- Epigenetics has the potential to serve as a robust biomarker of exposure for epidemiology studies that seek to examine environmental risk factors for disease.

References

Papers of special note have been highlighted as:

• of interest; •• of considerable interest

- 1 Waddington CH. Canalization of development and the inheritance of acquired characters. *Nature* 150, 3 (1942).
- 2 Esteller M. Epigenetics in cancer. *N. Engl. J. Med.* 358(11), 1148–1159 (2008).
- 3 Abi Khalil C. The emerging role of epigenetics in cardiovascular disease. *Ther. Adv. Chronic Dis.* 5(4), 178–187 (2014).
- 4 Lardenoije R, Iatrou A, Kenis G *et al.* The epigenetics of aging and neurodegeneration. *Prog. Neurobiol.* 131, 21–64 (2015).
- 5 Akbarian S. Epigenetic mechanisms in schizophrenia. *Dialogues Clin. Neurosci.* 16(3), 405–417 (2014).
- 6 Darby MM, Sabunciyar S. Repetitive elements and epigenetic marks in behavior and psychiatric disease. *Adv. Genet.* 86, 185–252 (2014).
- 7 Krishnan HR, Sakharkar AJ, Teppen TL, Berkel TD, Pandey SC. The epigenetic landscape of alcoholism. *Int. Rev. Neurobiol.* 115, 75–116 (2014).
- 8 Martino D, Kesper DA, Amarasekera M, Harb H, Renz H, Prescott S. Epigenetics in immune development and in allergic and autoimmune diseases. *J. Reprod. Immunol.* 104–105, 43–48 (2014).
- 9 Zhang Z, Zhang R. Epigenetics in autoimmune diseases: pathogenesis and prospects for therapy. *Autoimmun. Rev.* 14(10), 854–863 (2015).
- 10 Manolio TA. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* 14(8), 549–558 (2013).
- 11 Liu Y, Li X, Aryee MJ *et al.* GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am. J. Hum. Genet.* 94(4), 485–495 (2014).
- Foundational paper that maps correlations between DNAm at various loci across the genome and, in addition, examines their relationships to genetic variants and environmental exposure to smoking.
- 12 Drong AW, Nicholson G, Hedman AK *et al.* The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS ONE* 8(2), e55923 (2013).
- 13 Grundberg E, Meduri E, Sandling JK *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* 93(5), 876–890 (2013).
- 14 Lemire M, Zaidi SH, Ban M *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* 6, 6326 (2015).
- 15 Heyn H, Sayols S, Moutinho C *et al.* Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep.* 7(2), 331–338 (2014).
- 16 Shi J, Marconett CN, Duan J *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat. Commun.* 5, 3365 (2014).
- 17 Smith AK, Kilaru V, Kocak M *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* 15, 145 (2014).

- 18 Banovich NE, Lan X, Mcvicker G *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* 10(9), e1004663 (2014).
- 19 Bell JT, Pai AA, Pickrell JK *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 12(1), R10 (2011).
- 20 Gibbs JR, Van Der Brug MP, Hernandez DG *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* 6(5), e1000952 (2010).
- 21 Shin J, Bourdon C, Bernard M *et al.* Layered genetic control of DNA methylation and gene expression: a locus of multiple sclerosis in healthy individuals. *Hum. Mol. Genet.* 24(20), 5733–5745 (2015).
- 22 Liu Y, Aryee MJ, Padyukov L *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* 31(2), 142–147 (2013).
- **One of the first lines of evidence in humans showing epigenetic mediation of a genetic risk factor for disease.**
- 23 Hong X, Hao K, Ladd-Acosta C *et al.* Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat. Commun.* 6, 6304 (2015).
- 24 Olsson AH, Volkov P, Bacos K *et al.* Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets. *PLoS Genet.* 10(11), e1004735 (2014).
- 25 Chess A. Random and non-random monoallelic expression. *Neuropsychopharmacology* 38(1), 55–61 (2013).
- **Excellent review discussing allele-specific expression, epigenetics and its potential impact on genotype–phenotype relationships.**
- 26 Tycko B. Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.* 19(R2), R210–R220 (2010).
- 27 Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* 11(8), 533–538 (2010).
- 28 Costa V, Aprile M, Esposito R, Ciccociolla A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* 21(2), 134–142 (2013).
- 29 Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. *Environ. Health Perspect.* 114(4), 567–572 (2006).
- **Seminal findings showing epigenetic mediation of a prenatal environmental exposure lead to differential risk for obesity in adulthood.**
- 30 Bakulski KM, Fallin MD. Epigenetic epidemiology: promises for public health research. *Environ. Mol. Mutagen.* 55(3), 171–183 (2014).
- 31 Bellavia A, Urch B, Speck M *et al.* DNA hypomethylation, ambient particulate matter, and increased blood pressure: findings from controlled human exposure experiments. *J. Am. Heart Assoc.* 2(3), e000212 (2013).
- 32 Bind MA, Lepeule J, Zanobetti A *et al.* Air pollution and gene-specific methylation in the Normative Aging Study: association, effect modification, and mediation analysis. *Epigenetics* 9(3), 448–458 (2014).
- 33 Teh AL, Pan H, Chen L *et al.* The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Res.* 24(7), 1064–1074 (2014).
- 34 Rodriguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. *Nat. Med.* 17(3), 330–339 (2011).
- **Excellent review that summarizes how epigenetics is currently being used in clinical medicine as a biomarker of disease.**
- 35 A stool DNA test (Cologuard) for colorectal cancer screening. *JAMA* 312(23), 2566 (2014).
- 36 A stool DNA test (Cologuard) for colorectal cancer screening. *Med. Lett. Drugs Ther.* 56(1453), 100–101 (2014).
- 37 Goessl C, Krause H, Muller M *et al.* Fluorescent methylation-specific polymerase chain reaction for DNA-based detection of prostate cancer in bodily fluids. *Cancer Res.* 60(21), 5941–5945 (2000).
- 38 Crocetto LE, Korn D, Kretzner L *et al.* Prostate cancer molecular markers GSTP1 and hTERT in expressed prostatic secretions as predictors of biopsy results. *Urology* 64(4), 821–825 (2004).
- 39 Woodson K, O'reilly KJ, Hanson JC, Nelson D, Walk EL, Tangrea JA. The usefulness of the detection of *GSTP1* methylation in urine as a biomarker in the diagnosis of prostate cancer. *J. Urol.* 179(2), 508–511; discussion 511–502 (2008).
- 40 Stewart GD, Van Neste L, Delvenne P *et al.* Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: results of the MATLOC study. *J. Urol.* 189(3), 1110–1116 (2013).
- 41 Costa VL, Henrique R, Danielsen SA *et al.* Three epigenetic biomarkers, GDF15, TMEFF2, and VIM, accurately predict bladder cancer from DNA-based analyses of urine samples. *Clin. Cancer Res.* 16(23), 5842–5851 (2010).
- 42 Partin AW, Van Neste L, Klein EA *et al.* Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. *J. Urol.* 192(4), 1081–1087 (2014).
- 43 Maldonado L, Brait M, Loyo M *et al.* *GSTP1* promoter methylation is associated with recurrence in early stage prostate cancer. *J. Urol.* 192(5), 1542–1548 (2014).
- 44 Mahon KL, Qu W, Devaney J *et al.* Methylated glutathione S-transferase 1 (mGSTP1) is a potential plasma free DNA epigenetic marker of prognosis and response to chemotherapy in castrate-resistant prostate cancer. *Br. J. Cancer* 111(9), 1802–1809 (2014).
- 45 Hegi ME, Diserens AC, Godard S *et al.* Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clin. Cancer Res.* 10(6), 1871–1874 (2004).
- 46 Paz MF, Yaya-Tur R, Rojas-Marcos I *et al.* CpG island hypermethylation of the DNA repair enzyme

- methyltransferase predicts response to temozolomide in primary gliomas. *Clin. Cancer Res.* 10(15), 4933–4938 (2004).
- 47 Esteller M, Garcia-Foncillas J, Andion E *et al.* Inactivation of the DNA-repair gene *MGMT* and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* 343(19), 1350–1354 (2000).
 - 48 Rivera AL, Pelloski CE, Gilbert MR *et al.* *MGMT* promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neurooncology* 12(2), 116–121 (2010).
 - 49 Ladd-Acosta C. Epigenetic signatures as biomarkers of exposure. *Curr. Envir. Health Rpt* 2(2), 117–125 (2015).
 - 50 Joubert BR, Haberg SE, Nilsen RM *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* 120(10), 1425–1431 (2012).
 - 51 Breton CV, Siegmund KD, Joubert BR *et al.* Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS ONE* 9(6), e99716 (2014).
 - 52 Richmond RC, Simpkin AJ, Woodward G *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum. Mol. Genet.* 24(8), 2201–2217 (2015).
 - Provides some of the first evidence in humans of persistent epigenetic changes associated with previous exposure, thus, highlighting the potential utility for epigenetics to serve as a long-term biomarker of exposure.
 - 53 Novakovic B, Ryan J, Pereira N, Boughton B, Craig JM, Saffery R. Postnatal stability, tissue, and time specific effects of AHRH methylation change in response to maternal smoking in pregnancy. *Epigenetics* 9(3), 377–386 (2014).
 - 54 Markunas CA, Xu Z, Harlid S *et al.* Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* 122(10), 1147–1153 (2014).
 - 55 Shenker NS, Ueland PM, Polidoro S *et al.* DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology* 24(5), 712–716 (2013).
 - Provides some of the first evidence in humans of persistent epigenetic changes associated with previous exposure, thus highlighting the potential utility for epigenetics to serve as a long-term biomarker of exposure.
 - 56 Wright RO, Schwartz J, Wright RJ *et al.* Biomarkers of lead exposure and DNA methylation within retrotransposons. *Environ. Health Perspect.* 118(6), 790–795 (2010).
 - 57 Davies MN, Volta M, Pidsley R *et al.* Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 13(6), R43 (2012).
 - 58 Dempster EL, Pidsley R, Schalkwyk LC *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* 20(24), 4786–4796 (2011).
 - 59 Kaminsky Z, Tochigi M, Jia P *et al.* A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. *Mol. Psychiatry* 17(7), 728–740 (2012).
 - 60 Horvath S, Zhang Y, Langfelder P *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 13(10), R97 (2012).
 - 61 Wong IH, Lo YM, Zhang J *et al.* Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Res.* 59(1), 71–73 (1999).
 - 62 Cui H, Cruz-Correa M, Giardiello FM *et al.* Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science* 299(5613), 1753–1755 (2003).
 - 63 Coppede F, Lopomo A, Spisni R, Migliore L. Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World J. Gastroenterol.* 20(4), 943–956 (2014).
 - 64 Houseman EA, Accomando WP, Koestler DC *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).
 - Describes epigenetic signatures of six specific cell types in blood and, in addition, provides a critical advance and commonly used method to adjust for differences in blood cell composition in epigenetic epidemiology studies.
 - 65 Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol.* 15(3), R50 (2014).
 - 66 Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30(10), 1431–1439 (2014).
 - 67 Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10), 1363–1369 (2014).
 - 68 Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 8(3), 290–302 (2013).
 - 69 Montano CM, Irizarry RA, Kaufmann WE *et al.* Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol.* 14(8), R94 (2013).
 - 70 Sandoval J, Heyn H, Moran S *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6(6), 692–702 (2011).
 - 71 Bibikova M, Barnes B, Tsan C *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 98(4), 288–295 (2011).
 - 72 Fouse SD, Nagarajan RO, Costello JF. Genome-scale DNA methylation analysis. *Epigenomics* 2(1), 105–117 (2010).
 - 73 Callinan PA, Feinberg AP. The emerging science of epigenomics. *Hum. Mol. Genet.* 15(Spec No 1), R95–R101 (2006).
 - 74 Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696), 636–640 (2004).
 - 75 Roadmap Epigenomics C, Kundaje A, Meuleman W *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539), 317–330 (2015).

- 76 Bernstein BE, Stamatoyannopoulos JA, Costello JF *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28(10), 1045–1048 (2010).
- 77 Eckhardt F, Lewin J, Cortese R *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* 38(12), 1378–1385 (2006).
- 78 Meissner A, Mikkelsen TS, Gu H *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454(7205), 766–770 (2008).
- 79 Bell JT, Tsai PC, Yang TP *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* 8(4), e1002629 (2012).
- 80 Putku M, Kals M, Inno R *et al.* *CDH13* promoter SNPs with pleiotropic effect on cardiometabolic parameters represent methylation QTLs. *Hum. Genet.* 134(3), 291–303 (2015).
- 81 Zhi D, Aslibekyan S, Irvin MR *et al.* SNPs located at CpG sites modulate genome–epigenome interaction. *Epigenetics* 8(8), 802–806 (2013).
- 82 International HapMap C. The International HapMap Project. *Nature* 426(6968), 789–796 (2003).
- 83 Genomes Project C, Abecasis GR, Auton A *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422), 56–65 (2012).