Prospects for Epigenetic Epidemiology

Debra L. Foley, Jeffrey M. Craig, Ruth Morley, Craig J. Olsson, Terence Dwyer, Katherine Smith, and Richard Saffery

Initially submitted May 16, 2008; accepted for publication October 9, 2008.

Epigenetic modification can mediate environmental influences on gene expression and can modulate the disease risk associated with genetic variation. Epigenetic analysis therefore holds substantial promise for identifying mechanisms through which genetic and environmental factors jointly contribute to disease risk. The spatial and temporal variance in epigenetic profile is of particular relevance for developmental epidemiology and the study of aging, including the variable age at onset for many common diseases. This review serves as a general introduction to the topic by describing epigenetic mechanisms, with a focus on DNA methylation; genetic and environmental factors that influence DNA methylation; epigenetic influences on development, aging, and disease; and current methodology for measuring epigenetic profile. Methodological considerations for epidemiologic studies that seek to include epigenetic analysis are also discussed.

DNA methylation; environment; epigenesis, genetic; folic acid

Abbreviation: DNMT, DNA methyltransferases.

The association between environmental factors and disease risk has been investigated for a wide variety of disorders. There have been notable successes, such as the discovery of strong links between smoking, radon, and asbestos exposure and lung cancer (1). For many other common diseases that contribute substantially to the global health burden, however, environmental factors account for only a proportion of total disease risk and may interact with genes to exert their effect. Even the association between smoking and lung cancer may be moderated by genotype (2–4). Conversely, environmental modulation of genetic effects may act at many different levels and can influence the phenotype of even the most penetrant single-gene disorders.

Genetic epidemiology is now an integral component of the epidemiology paradigm (5–7). Some diseases, such as cystic fibrosis, are almost wholly caused by single genetic mutations (8). Most common diseases, however, cannot be explained by single genetic risk factors in isolation. Advances in the application of genetic methodologies will no doubt continue to identify new genetic variants with small independent effects, but a knowledge of both polygenic effects and complex interactions may be required to fully understand disease causation (2). In some cases, specific environmental exposures are associated with increasing mutation rate, with a dose-dependent increase in genetic damage and disease risk (9–12).

Environmental agents can also modify gene expression independently of the primary DNA sequence through a process known as epigenetics. Epigenetic modifications are mitotically heritable chemical/structural changes that regulate gene activity in the absence of underlying changes to DNA sequence. These modifications are the likely mediators of gene-environment interaction because genetic factors can modify the epigenetic response to the environment (13), and faulty epigenetic silencing can have downstream genetic consequences (14). The primary DNA sequence is generally fixed at conception, but epigenetic marks are dynamic and modifiable, probably throughout the life course. Recent work with human cell lines has shown evidence of dynamic reprogramming of epigenetic markings during the cell cycle (15–17). The expression of genetic risk is therefore likely to show varying penetrance over
time in response to epigenetic profile. Such dynamism raises the possibility of novel preventive and/or therapeutic opportunities (18).

This review covers the following 8 topics:

1. Epigenetic mechanisms
2. Genetic factors that influence DNA methylation
3. Epigenetics and development
4. Aging and the epigenome
5. Epigenetics and disease
6. Environmental factors and the epigenome
7. Measuring epigenetic profile
8. Considerations for epidemiologic studies involving epigenetic analysis

## EPIGENETIC MECHANISMS

Several epigenetic mechanisms regulate genes. The most robust and readily measured modification is DNA methylation. Its assessment requires only small amounts of genomic DNA with little specialized sample processing, and a wide range of fresh and archived tissues can be used. Other epigenetic marks include changes to histone proteins, around which DNA is packed, or involve functional noncoding RNAs. The interested reader is referred elsewhere for a detailed account of these mechanisms (19, 20). At present, analysis of histone proteins and noncoding RNAs is more technically challenging than analysis of DNA methylation and requires cells to be cryopreserved, snap-frozen in liquid nitrogen, or stored in RNA preservative.

The DNA double helix is a ladder-like molecule in which each side, the backbone, is made up of phosphate groups and sugar molecules, and the rungs are composed of 4 bases: adenine (A), guanine (G), cytosine (C), and thymine (T). These bases are linked in pairs (known as base pairs) on opposite strands: C-G and A-T. In a CpG site, C and G are linked by a phosphate molecule in the “backbone” (Figure 1). CpG islands are regions of DNA that contain a high density of CpG sites (21). In many cases, these are located in the control region of genes or in association with repetitive DNA elements (22). In general, low levels of DNA methylation (hypomethylation) are associated with higher gene activity and high levels of methylation with gene silencing (23). Repeat associated CpG islands and non-regulatory CpG sites generally exist in a methylated state (22, 24).

## GENETIC FACTORS THAT INFLUENCE DNA METHYLATION

One-carbon metabolism is a process by which methyl groups are passed from one donor molecule to the next (25, 26) (Figure 2). This process produces S-adenosylmethionine, which donates its methyl group to cytosine in a reaction catalyzed by DNA methyltransferases (DNMTs; refer to the information below). The rate of passage through this cycle can be influenced by genetic polymorphisms that encode the enzymes involved (27). The C-to-T substitution at nucleotide 677 of the methylenetetrahydrofolate reductase gene, MTHFR (677C>T), for example, results in a more thermolabile enzyme, and TT homozygous individuals, compared with CC homozygous individuals, have lower levels of DNA methylation (28, 29).

The DNMT family regulates DNA methylation. DNMT1 maintains methylation levels following DNA replication, whereas DNMT3A and DNMT3B act de novo to add DNA methylation. These enzymes regulate the dynamic methylation of genes during the establishment of imprinting (parent of origin gene expression) or cell differentiation (30). Variants in DNMT1 have been identified as risk factors for disease, including, in a case-control study, systemic lupus erythematosus (31). A genetic deficiency of DNMT3B causes a recessive human disorder characterized by immunodeficiency, centromere instability, and facial anomalies (32). In case-control studies, variants in other DNMTs (i.e., DNMT3L, DNMT1) have been associated with cancers (33–36).

CpG sites themselves are subject to genetic variation that can alter the sequence of gene regulatory regions and potential methylation levels. The C allele of the 102 T>C variant of the serotonin receptor gene (5HT2A), for example, contains 2 additional CpG dinucleotides thought to facilitate greater methylation levels and lower gene expression (37) and has been associated with psychiatric phenotypes (reviewed by Serritti et al. (38)). Removal of CpG sites can potentially abolish binding sites for proteins involved in transcriptional regulation. In 2 prospective cohorts of colon cancer cases, a specific variant (C>T) in the O-6-methylguanine-DNA methyltransferase tumor suppressor gene (MGMT), for example, has been strongly associated with O-6-methylguanine-DNA methyltransferase promoter methylation and gene silencing (39).

DNA methylation levels may also be modified through genetic variation at non-CpG sites within, or in close

---

**Figure 1.** Diagrammatic representation of CpG methylation (where cytosine (C) and guanine (G) are linked by a phosphate molecule). Approximately 60% of genes have CpG islands in a region called the promoter. This segment of DNA acts to control the activity, or expression, of that gene. An important property of CpG sites is that the cytosine can be methylated, by adding a methyl (CH₃) molecule, to form methyl-CpG. Note that, because DNA is double stranded, with strands running in opposite directions, cytosine bases on one strand pair with guanine on the other, CpG sites are methylated on both strands.
proximity to gene regulatory regions. Increasing methylation of a CpG island in the serotonin transporter gene (5HTT), for example, is associated with decreasing levels of gene expression, but this effect is evident only when the 5HTTLPR genotype (length of an upstream DNA repeat) is included (40).

EPIGENETICS AND DEVELOPMENT

DNA methylation levels and profile are very dynamic, especially during the epigenetic remodeling that takes place early in embryogenesis (Figure 3). The partially methylated egg and sperm genomes are globally demethylated soon after fertilization. Methylation is then reestablished progressively, starting in the early postconception period. Imprinted genes, however, retain the methylation profile of the parent of origin (41).

Epigenetic marks can be stably passed from one cell to its descendants (42) and, in some cases, when such marks survive the epigenetic remodeling of gametogenesis and early embryogenesis, from parent to offspring (43–45). This process has been demonstrated in animal studies, but limited evidence for germline transmission in humans has also been reported and, in one case, has been linked to an increased risk of cancer (45).

Epimutation is estimated to be 100 times more frequent than genetic mutation (46, 47) and may occur randomly or in response to the environment. Periods of rapid cell division and epigenetic remodeling are likely to be most sensitive to stochastic or environmentally mediated epimutation.

AGING AND THE EPIGENOME

Both genome-wide and specific methylation profile/patterns change with age, and this may be genetically controlled (48). A generalized decrease in DNA methylation with age has been reported in mice and in cell lines (49–51), although this decrease may be tissue and/or gene specific (52–54). Decreased methylation may be accompanied by reactivation of previously silenced genes (46). Age-related methylation changes have been described in cancer (55) and in gene promoters associated with cancer risk (56, 57; reviewed by Issa (58)). Age-related epigenetic changes have also been demonstrated in sperm cells, but the direction of change over time appears gene specific (59). Within-pair differences in DNA methylation are greater in older than younger monozygotic twins (60). The contribution of genetic, environmental, and random factors to this cumulative discordance is unknown.

EPIGENETICS AND DISEASE

Disruption of epigenetic profile is a feature of most cancers (61–63) and is speculated to play a role in the etiology of other complex diseases (13, 64, 65), including asthma (66), allergy (67), obesity, type 2 diabetes, coronary heart disease, autism spectrum disorders (68), and bipolar disorder and schizophrenia (69–73). The potential to identify distinct epigenetic biomarkers associated with eating disorders has also been explored (74, 75). Disruption of epigenetic profile is also implicated in some adverse health outcomes for subjects conceived by means of assisted reproductive technologies (76).

The most striking disease-associated epigenetic change is seen in cancer. The increase in tumor suppressor gene CpG island methylation between cancerous and noncancerous tissue is often close to 100% (77) and is usually reversed at CpG sites associated with tandemly repetitive and interspersed repeat DNA in the same tumor cells (78–81). Specific profiles of methylation have also been associated with factors that predict prognosis (82).

To our knowledge, the dramatic difference in methylation levels observed in cancerous versus noncancerous tissue has not been found in other complex diseases, where methylation at any given CpG island or specific CpG sites in affected versus unaffected individuals may vary by less than 10% (83–85). Interpretation of functional consequences is therefore problematic. Results from small studies must also be interpreted with caution even when supported by in vitro functional demonstration of changes in gene activity with methylation change (86). However, for some genes, evidence exists that a small change in the level of DNA methylation, especially in the lower range, can dramatically alter gene expression (87, 88).
ENVIRONMENTAL FACTORS AND THE EPIGENOME

Diet

Diet is an important modifier of epigenetic profile (Figure 2; reviewed by Davis and Uthus (25), Cobiac (89), Lahiri et al. (90), and Muskiet (91)). Specific micronutrients involved in one-carbon metabolism include folate, an important primary methyl donor, whose availability is directly correlated with DNA methylation levels (91, 92). Low folate levels lead to hyperhomocysteinemia, which inhibits key metabolites of the one-carbon pathway (91, 93) and is associated with coronary heart disease and cancer (91, 94–96). To our knowledge, there is no evidence to date that these associations are mediated epigenetically. Diet can also change the epigenome via factors that alter the profile of histone modifications in cells, thereby altering gene expression levels (reviewed by Davis and Ross (97) and Herceg (98)). Given the demonstrated role of diet in regulating the epigenome, recommended dietary allowances of micronutrients have been proposed for maintenance of genome/epigenome stability (99).

The timing of nutritional insufficiency or other environmental exposures may also be critical (100). Epidemiologic evidence suggests transgenerational effects on health outcomes and mortality that are sensitive to the timing of environmental exposures. These effects may also be sex specific (101–105). Such associations have received a great deal of attention because of speculation about the underlying mechanisms. Pembrey et al. (105) identified several theoretical possibilities to explain the findings, including prions, viruses, RNA, responsive DNA sequences, or epigenetic changes. There is, however, no direct evidence that they are mediated via modification to epigenetic profile.

Animal studies have been more convincing. Maternal protein restriction in rats during pregnancy leads to a loss of methylation in the offspring at gene promoters associated with glucose metabolism in tissues such as liver, lung, and kidney (106, 107).
Other lifestyle factors

Human case-control studies have demonstrated that alcohol consumption increases methylation at gene promoters (108–111) and is associated with methylation-induced silencing of tumor suppressor genes in colorectal cancer (112) and hyperhomocysteinemia (113). Alcohol consumption is also associated with altered levels of methyltransferase activity and DNA methylation (71, 109). In both animal and in vitro studies, alcohol has been shown to impede the bioavailability of dietary folate and inhibit folate-dependent biochemical reactions (114, 115).

Cigarette smoking is associated with increased methylation at tumor suppressor genes in human case-control studies (116) and in mice (117) and with loss of methylation at oncogenes in human cancer cell lines (118). Disruption of DNMTs is implicated in the mechanism of induced change (118).

Endocrine disruptors

Endocrine disruptors are chemicals that interfere with the function of hormones by mimicking, blocking, or disrupting their synthesis, transport, or elimination. Hormones are chemical messengers that travel through the blood to target cells where they interact with receptors that, in turn, directly influence gene activity, usually via epigenetic mechanisms. Any endocrine disruption may therefore have epigenetic consequences. Diethylstilbestrol is a synthetic estrogen. Prenatal exposure to diethylstilbestrol has been shown to increase the risk of cervical and vaginal cancer and pregnancy-related problems in women and testicular abnormalities in men (119). Diethylstilbestrol exposure in animals decreases promoter DNA methylation in reproductive tissues and increases methylation (and hence decreases activity) of DNMT (120). Such exposure also decreases methylation of the cancer-causing oncogene c-fos (121) and the estrogen-responsive gene lactoferrin in mice (122). In animals, in utero or neonatal exposure to bisphenol A is associated with higher body mass, altered reproductive function, increasing cancer risk, and specific DNA methylation changes (123–125).

Genistein is an estrogen-like polyphenol found in soybeans that alters DNMT function and changes the DNA methylation status of several genes, including some tumor suppressors (126–128). Genistein may be protective against certain prostate and mammary cancers (127). Animal studies have also shown that transient exposure to vinclozolin (a fungicide) or methoxychlor (a pesticide) is associated with changes in DNA methylation of several genes (129) and decreased fertility in male offspring over several generations (130).

Other environmental agents

Heavy metals such as nickel, cadmium, and arsenic and ionizing and ultraviolet radiation (reviewed by Herceg (98)) have all been associated with an altered epigenetic profile. Chromium exposure in male mice, for example, causes hypomethylation of sperm genomic DNA and is associated with increasing risk of tumors and other abnormalities in progeny (131, 132).

Maternal behavior

One of the most interesting studies of environment-induced epigenetic change involved newborn rats exposed to differing degrees of maternal care (133, 134). Low-level maternal care was associated with decreased glucocorticoid receptor promoter methylation (increased gene activity) in the hippocampus and altered stress response in the young. This effect was reversible by either pharmacologic or dietary intervention in adulthood (135). Similar findings have been reported for the estrogen receptor alpha gene (136). No human version of these studies has yet been reported, but, given how widely stress is implicated in disease onset and relapse (137–141), it is difficult to overstate the potential importance of such findings.

Infective agents and the epigenome

Bacterial infection may directly modify the epigenetic profile of the host animal (142). Aberrant methylation of gastric mucosa genes is a common finding in humans infected with Helicobacter pylori and is an early event in gastric carcinogenesis (143–145). Insulin-like growth factor 2 methylation imprinting profile in the placenta is altered in mice infected with Campylobacter rectus during pregnancy (142). Similar alterations in DNA methylation have been reported in human cells and cell lines following viral or helminth infection (146–152). Methylation may be increased or decreased depending on the infectious agent (142, 153). Change in epigenetic profile in response to infection may play a role in the development of immune-related disorders and many cancers previously associated with infectious agents (e.g., gastric cancer).

MEASURING EPIGENETIC PROFILE

DNA methylation can be assessed by using very small amounts of genomic DNA obtained from fresh or archived tissue samples (154), including blood spots stored for many years (155). The “gold standard” for DNA methylation analysis involves sodium bisulfite treatment of genomic DNA for selective conversion of unmethylated cytosine residues to uracil, leaving methylated cytosine unchanged (156). This specific chemical change can be assessed by DNA sequencing, polymerase chain reaction amplification, or mass spectroscopy methods.

DNA methylation can also be studied on a genome-wide scale using a variety of enzyme-based or antibody affinity techniques that enrich for either methylated or unmethylated fractions of genomic DNA (157–162). These fractions can then be hybridized to DNA microarrays or sequenced en masse. Low-resolution array methodology has been used to identify genes consistently differentially methylated in a case-control study of psychosis (73) and has the potential to quickly (although not inexpensively) identify genes differentially methylated in a range of complex disorders in case-control studies.

As an alternative to mapping levels of DNA methylation at specific sites, the overall level of methylation in a tissue can be determined by directly measuring the amount of cytosine and methyl-cytosine (81, 163–165) or by a proxy using a combination of methylation-dependent DNA digestion and fluorescent tagging (166). Another alternative is to use other methods that assess levels of repeat-based methylation (167).

The development of ultra-high-throughput DNA sequencing technologies for the direct sequencing of enriched, methylated DNA fragments or of bisulfite-converted genomic DNA (168, 169) will ultimately permit measurement of comprehensive methylation profiles. The dynamic nature of epigenetic markings obviously warrants longitudinal biospecimen sampling where possible. In 2008, the National Institutes of Health made a significant commitment to research involving epigenomics (http://www.nih.gov/news/health/jan2008/od-22.htm) as part of an NIH Roadmap established in recognition of the importance of developing high-throughput and cost-effective methods for epigenetic analysis.

CONSIDERATIONS FOR EPIDEMIOLOGIC STUDIES INVOLVING EPIGENETIC ANALYSIS

Genetic-epidemiologic studies provide a framework for understanding the joint impact of genotype and environmental exposure on disease risk. Addition of epigenetic data will help clarify the functional basis underlying such joint effects by providing unbiased biologic measures arising from the combined effects of such interactions (and stochastic effects) that may lead to altered gene activity. The common disease genetic and epigenetic model provides a starting framework for including epigenetic data in genetic studies (13, 170).

Genetic and epigenetic analysis

Understanding the similarities and differences in genetic versus epigenetic data will help with planning the next wave of epidemiologic studies that prospectively incorporate both. Any peripheral tissue samples can be used for genotyping an individual. The level and pattern of epigenetic marks vary across different tissue and cell types, however, posing a formidable challenge for epigenetic analysis. Epigenetic marks in readily accessible cells (e.g., blood, buccal, skin), for example, may not reflect those in generally inaccessible tissues (e.g., brain). Analysis of postmortem tissue will provide valuable insight into the epigenetic profile of inaccessible tissues, but the identification and validation of peripheral epigenetic marker will be required for epidemiologic studies to benefit from incorporation of epigenetic data.

Genetic epidemiology involves the analysis of single nucleotide polymorphism alleles (usually binary) or variable numbers of simple repeats (usually tandem nucleotide elements), which can alter gene function via changes in coding sequence, RNA processing, or changing gene promoter sequence. A potentially methylatable CpG-rich region, in contrast, can comprise many dozens of individual CpG sites, each of which may have the potential to influence gene activity. CpG sites may be individually or coordinately methylated, but this is likely to vary from gene to gene and tissue to tissue. The relative biologic significance of methylation at a particular site versus average methylation across the entire region remains to be determined in most cases. A recent study of interindividual variation in methylation suggests that measurement of average methylation levels may suffice to characterize the methylation state of CpG-rich islands (171).

The genotype of an individual is generally fixed at conception, barring subsequent genetic mutations. The epigenotype, in contrast, is tissue and cell-type specific and may vary over time as a function of environmental exposure, aging, and random processes (48, 60, 172, 173). Analysis of epigenetic data, therefore, cannot rely upon the assumptions of Mendelian randomization (174, 175), which is predicated on the random assortment of genes transmitted from parents to offspring during gamete formation before disease onset. Associations between genotype and disease are therefore not usually biased because of reverse causation or confounding, unless linkage disequilibrium, pleiotropy, genetic heterogeneity, or population stratification are involved (174). Cross-sectional assessment of epigenetic profile does not permit any inference regarding direction of causation. Longitudinal data will be required to make causal inferences about observed associations between epigenetic marks and disease in cross-sectional data, and the timing of epigenetic sample collection will therefore be of paramount importance.

Modeling epigenetic data

Epigenetics is a relatively new field, and there is currently no consensus as to the most appropriate way to model methylation data. Very few epidemiologic studies currently integrate genetic (dichotomous variable), epigenetic (continuous variable), and environmental exposure data. Separate associations may exist between genotype and epigenotype, genotype and phenotype, and epigenotype and phenotype. Environmental exposures (e.g., cigarette smoking or alcohol consumption) may influence epigenotype and have independent effects upon the phenotype. The appropriate statistical model will thus depend on the question of interest and substantive knowledge concerning the relevant biologic pathways. If methylation is not allele specific, however, and that needs to be tested on a gene-by-gene basis, then there is no need to specify a model of inheritance (e.g., dominant, recessive, additive) for diseases caused by methylation changes.

Methods for modeling correlated outcomes include generalized estimating equations, mixed models, or the simpler approach of analyzing summary measures of the data such as means. Because potentially methylatable CpG sites often exist in clusters and may show coordinated methylation changes, analytic methods for summarizing correlated data may be required. Data reduction techniques such as principal components analysis have been used to model correlated CpG sites in studies in which DNA methylation is the exposure of interest (176).

Most analysis of epigenetic data summarizes methylation at individual CpG sites as proportions. The relation between methylation and gene expression may not be linear and may differ from site to site. Modern statistical software permits the modeling of proportional outcomes within the generalized...
linear framework (e.g., STaTa’s GLM command (Stata Corporation, College Station, Texas), but the resulting estimates may be difficult for researchers to interpret. When epigenetic status or change in status over time is the outcome, then models for either threshold-based dichotomies or proportional data will be required. Threshold models, defined by a given level or pattern of methylation or a degree of change in methylation over time, will benefit from relevant functional data to identify meaningful thresholds.

**Multiple testing**

Similar to genome-wide association studies, a genome-wide search for epigenetic risk factors may investigate associations between a phenotype and thousands of biologic variables. As the number of tests performed increases, so does the probability of obtaining statistically significant results by chance. Tools such as the false discovery rate (177) permit identification of as many true associations as possible while minimizing the overall proportion of false-positive tests and can be applied to epigenetic analysis. Disentangling genetic and epigenetic effects in human studies is likely to prove challenging. Simple approaches might utilize standard epidemiologic analytic tools, such as examining associations between outcome and epigenotype within strata defined by different genotypes or examining interactions between different genetic and epigenetic factors, but these strategies are likely to require large sample sizes to achieve adequate power.

Studies of monozygotic twins provide an opportunity to examine the impact of unshared environmental exposures (e.g., discordance for cigarette smoking) on epigenetic profile while controlling for genotype and common familial environmental factors. An estimate of the association between changes in methylation and the outcome of interest not confounded by genetic background and shared environment can be obtained from the within-pairs coefficient of a suitable regression model. Carlin et al. (178) discuss models for estimating this coefficient and its interpretation in their review of regression modeling of twin data. Allelic effects on methylation against different genetic backgrounds can also be tested in dizygotic twins who share an allele in common versus dizygotic twins who differ at the allele of interest.

The dynamic nature of the epigenome may help explain the variable age at onset, progression, and outcomes associated with many common diseases and may provide new insights into the role of the environment in helping shape risk profiles. In this review, we have introduced epigenetics and discussed its role as a mediator of environmental effects on health and disease. Now that we are equipped with standard epidemiologic tools and statistical models, together with ever-advancing technology, the time is right to begin to answer important questions in this field.

**ACKNOWLEDGMENTS**

Author affiliations: Orygen Youth Health Research Centre & Department of Psychiatry, University of Melbourne, Australia (Debra L. Foley); Developmental Epigenetics, Australia and Royal Children's Hospital & Department of Paediatrics, University of Melbourne, Australia (Jeffrey M. Craig, Ruth Morley, Richard Saffery); Centre for Adolescent Health, Australia and Royal Children's Hospital & Department of Paediatrics, University of Melbourne, Australia (Craig J. Olsson); Clinical Epidemiology and Biostatistics Unit, Australia and Royal Children's Hospital & Department of Paediatrics, University of Melbourne, Australia (Katherine Smith); and Murdoch Childrens Research Institute, Australia and Royal Children's Hospital & Department of Paediatrics, University of Melbourne, Australia (Jeffrey M. Craig, Ruth Morley, Craig J. Olsson, Terence Dwyer, Katherine Smith, Richard Saffery).

Conflict of interest: none declared.

**REFERENCES**


