Epigenome-Wide Association Studies (EWAS): Past, Present, and Future

James M. Flanagan

Abstract

Just as genome-wide association studies (GWAS) grew from the field of genetic epidemiology, so too do epigenome-wide association studies (EWAS) derive from the burgeoning field of epigenetic epidemiology, with both aiming to understand the molecular basis for disease risk. While genetic risk of disease is currently unmodifiable, there is hope that epigenetic risk may be reversible and or modifiable. This review will take a look back at the origins of this field and revisit the past early efforts to conduct EWAS using the 27k Illumina methylation beadarrays, to the present where most investigators are using the 450k Illumina beadarrays and finally to the future where next generation sequencing based methods beckon. There have been numerous diseases, exposures and lifestyle factors investigated with EWAS, with several significant associations now identified. However, much like the GWAS studies, EWAS are likely to require large international consortium-based approaches to reach the numbers of subjects, and statistical and scientific rigor, required for robust findings.

Key words EWAS, Methylation, Risk, Biomarker, Cancer, Peripheral blood, Exposures

1 Origins and Definitions

Genetic epidemiology aims to use the natural variation in the genome, namely single-nucleotide polymorphisms (SNPs) and also copy number variants (CNVs) to look for associations with cancer risk, prognosis and in pharmacogenetic studies to explain drug response or resistance. It has been used to investigate why some patients have tumours that metastasize or have a particular pathology and has been used to look at environmental interactions. Over the last decade, starting with the Human Genome Project (HGP) and HAPMAP catalogue of human variation the development of experimental study design and statistical analysis to perform genome-wide association studies (GWAS) within large international consortia has begun to bear significant fruit with numerous diseases and phenotypes [1]. Using the recent iCOGs consortium as an example approximately 70 novel genetic associations for
breast [2], ovarian [3], and prostate cancers [4] were identified in a large international effort. The current challenges in genetic epidemiology are now in translation of this new information into risk prediction and or stratified medicine-based approaches that can be used in the clinic.

New research over the last few years has aimed to look further into the genome at the natural variation in the epigenome, in particular at DNA methylation, which is the addition of a methyl group to cytosines as a mechanism of gene regulation. For all of the natural epigenetic variation in the human population you can ask all of the same questions as in genetic epidemiology studies. The term epigenetic epidemiology was first mentioned in the literature as a future research topic by Issa in a review of colon cancer epigenetics in 2000, in which was described the need to understand the interaction between dietary folate and colon cancer risk [5]. The concept was further developed and defined in a review by Jablonka and Lamb in 2002 in which the authors argue that inheritance of epigenetic information may provide a mechanism of inherited traits including risk of disease [6]. More recently numerous reviews have defined the area and various aspects of epigenetic epidemiology [7–13].

The abbreviation EWAS is now widely used to mean epigenome-wide association study, initiated by the review by Rakyan and colleagues [14] and typically used to define the analysis of DNA methylation. However, we must acknowledge that the term “epigenome” should also encompass a wide array of histone modifications that also influence gene function as part of the epigenome and have been largely ignored to date in EWAS due to technological constraints and appropriate sample availability. Furthermore, the abbreviation EWAS has also been used as an acronym for “environment-wide association studies” [15, 16], “exposome-wide association studies” [17], and even “early-walking aids” [18] and potentially other definitions. However, in the majority of studies published to date, and in this review, the term EWAS is used to define studies of DNA methylation to identify the common normal variation in the DNA methylome using genome-wide technologies most frequently using peripheral blood DNA.

The motivation for pursuing genome-wide approaches came from studies of DNA methylation of candidate genes measured in peripheral blood DNA. One of the earliest of these was the association between loss of imprinting (LOI) of the insulin growth factor II (IGF2) gene and colorectal cancer risk in cross-sectional studies [19–22]. However, IGF2 hypomethylation in peripheral blood was not related to colorectal or breast cancer risk in a prospective cohort study using pre-diagnostic samples, indicating that this may rather be a biomarker of disease, but not of risk [23]. Furthermore, epimutations have been identified in the MLH1 and MSH2 genes associated with familial colon cancer, Lynch syndrome patients [24, 25].
However, in both of these cases the DNA methylation phenotype (epimutation) was driven by rare genetic polymorphisms in cis [26, 27]. This argues against the hypothesis that epimutations themselves may be heritable across generations as a mechanism to explain a proportion of the missing heritability of these diseases. Subsequently, candidate gene studies conducted by us [28] and others [29–32] have reported associations between candidate gene methylation and cancer risk. Specifically, using the ATM gene example we have investigated common epigenetic variation focussing on regions of the genome that are more variable, such as gene-body or intragenic sequences [28]. Much like GWAS, it is in the common variation in the population, rather than in rare epimutations, that EWAS will be most powerful for identifying robust biomarkers.

2 EWAS Reviews of Methods

The goal of this chapter is not to reiterate many of the numerous reviews of methods for EWAS, study design, statistical and bioinformatic approaches nor the various plagues that blight them [8, 14, 33–36]. There have been many excellent reviews on novel pre-processing methods and study design issues [9, 37–45]. This chapter rather focuses on the initial studies that have been performed, predominantly on white blood cell (WBC) DNA and what has been discovered thus far that have been robustly replicated and or validated. It is not intended to be an exhaustive review of all published studies. Studies have been identified in the literature using the search terms “EWAS,” “epigenome-wide association,” and “epigenetic epidemiology,” supplemented with some known studies that did not contain these keywords. However, several important methodological issues that have been consistently raised as concerns, caveats, and/or complications in EWAS as compared to GWAS need to be discussed.

The first concern is often on the use of WBC DNA where methylation variation may reflect differences in the proportions of different blood cell types that could obscure or confound associations. While it may be preferable to identify a more appropriate cell type [46], the fact remains that most large cohort studies have only collected blood samples that are thus the only DNA source available for identifying incident disease cases prior to disease onset. However, at this stage current evidence suggests that cell composition explains only a minor proportion of variation in blood DNA methylation [43]. Data generated on the 27k Illumina array [47, 48] and 450k array [49] using fractionated blood samples can be used to infer the cell population in WBC DNA [50] and to determine if identified biomarkers vary across cell type, most successfully used in the recent investigation of the autoimmune disease rheumatoid arthritis [51]. Apart from this disease, which
plausibly could be explained by the hematological nature of the disease, cell type composition has not explained a large proportion of significant associations in most EWAS analyses of non-hematological diseases published thus far. However, it remains important to consider cell type as a potential confounding factor.

There is a strong case to be made for longitudinal studies not only for risk studies (including DNA samples prior to disease onset), but also in intervention studies using pre- and posttreatment with specific exposures to investigate environmental impacts on the epigenome [40]. However, only few studies have investigated temporal variation of WBC DNA methylation in serial samples from the same individuals reporting associations with BMI or age [52–54]. Only methylation traits that are stable over time can be well characterised using a single blood sample. It is, therefore, important to establish the temporal variability of methylation markers in WBC DNA in order to select informative markers, i.e., those that are variable in the population, yet stable over time when only one blood sample is available. Failing this, it would be vital to use multiple serial blood samples from the same individuals to report robust associations with a particular phenotype.

The extent to which WBC DNA methylation is genetically determined and/or represents acquired changes through life is not known. A genetic component is supported by evidence in families [52] and twins [55, 56], in numerous studies showing allele-specific methylation [57, 58], and lastly with the examples of genetically driven epimutations in MSH2 and MLH1 [26, 59]. However, strong conclusions about a meiotically heritable mechanism for an epigenetic trait remain elusive. Age is associated with methylation and is a key risk factor for cancer and other disease; thus it is important to evaluate its potential confounder effects [60]. However, without controlled longitudinal studies it will be almost impossible to rule out any accumulated exposures throughout life from the effect of aging alone [54].

Lastly, unlike GWAS studies in which independent replication of genotyping has become less necessary (although still required for poor-performing SNPs), largely because the genotype calling is categorical, DNA methylation is a continuous quantitative variable and the various technical and statistical processing in generating this data strongly argues for the need for validation of the reported methylation values using independent quantitative methods such as Epityper or Pyrosequencing. It is not surprising that the many associations that have been reported and repeatedly validated in subsequent studies were often independently validated by alternative methods, such as the methylation at F2RL3 associated with tobacco smoking which was validated by Epityper in the original study and in subsequent reports [61–63].
The Illumina 27k Illumina array covers on average 2 CpG sites in the promoter regions of approximately 14,000 genes and represents less than 0.1% of the 28 million CpG sites in the human genome. Unlike GWAS studies that can use tag SNPs that can conceivably cover the majority of the haplotype blocks in the genome, the 27k array falls far short from being representative of “epigenome-wide.” Nevertheless some studies have used these arrays to identify blood-based DNA methylation signatures of risk in ovarian cancer [64], bladder cancer [65], head and neck squamous cell carcinoma (HNSCC) [66], and breast cancer [67]. Although these epigenome-wide cancer studies report intriguing associations, independent replication is required before any associations can be considered robust, since they were small studies (~200 cases/200 controls), and in some cases used blood samples collected after diagnosis or treatment. Furthermore, none of these early EWAS using the 27k used independent validation to verify the associated probes, and only time will tell if they can be further replicated.

In the first of these studies Teschendorff and colleagues used a mixture of at-diagnosis (n=113) and posttreatment blood samples from ovarian cancer patients (n=122) compared to unmatched healthy controls (n=148). This study design was not ideal given that the healthy controls were not matched to the cases and as such were significantly biased on several subject characteristics including age, OC use, and pregnancy as well as the cases being confounded either by the presence of disease or prior treatment with chemotherapeutic cytotoxic drugs. Nevertheless, using a supervised classification system the investigators identify a signature using the top 100 probes that can accurately predict case–control status using the at-diagnosis set that also predicted ovarian cancer in the posttreatment samples. However, from this first report one interesting observation was a bias in the differences between cases and controls towards non-CpG island probes (which were significantly under-represented in this array design), arguing strongly for the use of the latterly designed 450k array which does cover non-CpG islands with a higher density of probes.

Several studies from Kelsey and colleagues have reported signatures for both bladder and HNSCC from case–control studies using the 27k array [65, 66]. In the first of these they report on n=223 cases blood DNA collected approximately 1 year after diagnosis and treatment, compared to n=237 healthy unmatched (and untreated) controls that were also biased by smoking status and gender. This study reports a signature of the top nine probes that differentiate cases from controls with a predictive value of AUC=0.7 to predict the presence of bladder cancer [65]. In a smaller study of 92 cases of HNSCC compared to 92 healthy
unmatched controls (biased on smoking, alcohol, and HPV sero-positivity), using a similar supervised analysis they report a 6-CpG signature that predicts the presence of cancer in the patients (AUC = 0.73) [66]. A reanalysis of these data using a 50-probe signature associated with different blood cell populations more accurately predicted the case–control status of the subjects, suggesting that much of this association is driven by blood cell changes that may have occurred as a result of the presence of disease or prior treatment in the patients [48]. Like the previous ovarian cancer study, these studies have not used independent methods to validate these small probe-set signatures, and have thus far not been replicated by independent studies.

In a more recent study Xu and colleagues used the 27k Illumina DNA methylation array to perform an EWAS on prospective incident breast cancer cases (n = 298) and unmatched controls (n = 612) from the Sister Study [67]. Results suggest 250 probe associations of methylation detected in WBC DNA with breast cancer risk (FDR p < 0.05). However, without independent replication or methodological validation, conclusions from this study need to be tempered with caution. As described in the accompanying editorial [68], the strongly zero-inflated p-value distribution might either reflect yet-undiscovered potential confounding factors, or failure to capture and correct for the correlation between probes. An interesting observation in this study was that the majority of associated probes showed a loss of methylation reminiscent of the global hypomethylation observed in total methylation in cancer patients compared to controls [69]. If validated, this association and those reported for other tumour types may yet prove useful as risk biomarkers if independent of other known risk factors.

4 Present: 450k Studies

The Illumina 450k array is the most widely used platform in the last two years for studies reporting EWAS. The array still only covers less than 2 % of the CpG sites in the genome, but does attempt to cover all known genes with a high density of probes in the promoters (including CpG islands and surrounding sequences), but also covers with a lower density across the gene bodies, 3′UTRs, and other intergenic sequences.

There have been two EWAS using the 450k array to investigate diabetes, in specific subpopulations of WBC including CD14+ monocytes for type 1 diabetes in which several markers were further analyzed by pyrosequencing but were shown not to validate the reported association individually [70]. An important advance in this study, however, was the use of multiple study designs including discordant monozygotic twin pairs and singleton case–control group in which many of the associated markers were replicated [70].
CD4+ T cells were analyzed in an EWAS for type 2 diabetes, in which one gene *ABCG1* was associated with the phenotypic correlate insulin levels and a model for insulin resistance [71]. Further validation in independent cohorts for both of these findings is therefore warranted.

The first 450k-based EWAS to gain from adjusting for the blood cell proportions was performed by Liu and colleagues investigating *n*= 354 *rheumatoid arthritis* cases compared to *n*= 337 healthy matched controls reporting 4 CpG sites in the MHC cluster that epigenetically mediates a genetic association with the disease [51]. In addition to peripheral blood, validation specifically in monocytes suggested the possibility of a myeloid mechanism driving the peripheral blood association. While no studies have replicated this methylation phenotype yet, several GWAS reports have identified SNPs in the MHC locus associated with rheumatoid arthritis validating this genetic association with the disease (www.genome.gov/gwastudies). This supports the view that combining both genetic (GWAS) and epigenetic (EWAS) approaches in parallel may prove a fruitful approach for understanding mechanisms of disease risk [72].

The most robustly validated findings thus far with EWAS have been the associations between DNA methylation in blood and smoking status [61, 62, 73–83]. There may be many reasons why smoking associated methylation changes have been so strongly validated; firstly, that unlike other environmental exposures smoking is much easier to categorize subjects into presence or absence of the exposure and most cohort studies will have collected this variable; and secondly the methylation difference between smokers and nonsmokers for most probes is >5% which is a large enough change to be detected on multiple platforms. The genes that have shown the most robust associations across multiple studies (at least 3) are in *AHR*, 2q37.1, 6p21.33, *F2RL3*, *GPR15*, *GFI1*, *CYP1A1*, *MYO1G*, and *CNTNAP2*. These have been replicated in peripheral blood DNA, but also in lung tissue [62], cord blood of smoking mothers [79], and different genetic populations [82]. A number of interesting observations have been made regarding this association with smoking. Firstly, the methylation alterations are detectable in blood DNA even in former smokers that have quit up to 10–20 years prior to giving blood [62, 73, 83]. We have used this knowledge to create a predictive model of former smoking status that might be more quantitative and useful as an epidemiological variable (AUC = 0.80) to study risk of disease associated with former smoking [73]. Mechanistically, given that the majority of white blood cell types have life-spans of ~30 days, this suggests that the exposure must also be affecting the hematopoietic stem and progenitor cells which perpetuate the epigenetic alterations in the daughter differentiated cells. If exposures such as smoking can increase an individual’s risk of cancer, even in former smokers, then
it is conceivable that those exposures throughout life must also affect the adult stem and progenitor cells and potentially the cell of origin for the initial carcinogenic events. This provides an interesting model for other exposures that also induce long term increased risk of disease. The second interesting observation is that the most consistently altered gene across all studies is the aryl hydrocarbon receptor repressor (AHRR), and the majority of changes occur in an intragenic CpG island showing a loss of methylation >20%. The AHRR gene is involved in the regulation of the aryl hydrocarbon receptor (AHR) pathway which metabolizes the carcinogenic dioxins and dioxin-like compounds and therefore is a highly plausible candidate. However, while we have shown that hypomethylation of this intragenic locus is associated with over-expression of the gene in lung tissue, whether this results in a further repression of the pathway is not yet clear [62].

There have now been numerous studies that have reported an association between WBC DNA methylation and increasing age. Initially using the 27k array for EWAS two studies reported methylation signatures of aging in developmental genes defined by those that are bivalently marked or targeted by the PRC2 polycomb group complex [60, 84]. This association with age was also reported in a twin study using the 27k array, which showed further evidence for a genetic component of epigenetic variation [57] and was replicated in an interactome analysis of several 27k array datasets [85]. More recently using the 450k array Florath and colleagues showed that the majority of probes (78/94) that associate with age in a cross-sectional study design were also associated with age in a longitudinal study design (+8 years) [54]. This is important in ruling out confounding factors not accounted for in the cross-sectional study design in unrelated individuals. These investigators also developed a 17-probe set signature to estimate an individual’s age with average difference between predicted and observed ages of ±2.8 years [54]. Using a similar approach, Hannum and colleagues have developed a 71-probe set signature derived from the 450k array to predict age in multiple tissues with an accuracy of ± 3.9 years [86]. Importantly, this study showed that in tumour tissue DNA, the apparent age of the cells and ageing rate are far higher than their matched normal tissue counterparts providing direct evidence for an epigenetic mechanism for increased cancer risk with increased age. Intriguingly, males appeared to age faster than females, with no apparent hypothesis for a mechanism at this stage [86]. The overlap between the 71-probe signature and another dataset comparing newborns and centenarians [87] had 70/71 overlapping, while at least 39/78 probes defined by Florath and colleagues were also overlapping with these datasets. Across all of these studies several genes appear most consistently associated with age, including ELOVL2, CCDC102B, OTUD7A, and FHL2. Most importantly, some of
these associations have also been replicated using alternative technologies for ELOVL2 and FHL2 [88]. Therefore, apart from clearly displaying the need to adjust for age in any other EWAS analyses, these studies have shown some important findings regarding the consequences of increasing age in human biology and provided a novel molecular epigenetic measure of cellular age [89].

5 Future: Bis-Seq and Consortium-Based Approaches

Thus far, there have been no sequencing based EWAS reported; however, it is an obvious evolution for future EWAS. While the 450k array is an advance on the previous 27k version, it is clear that targeting only 2% of the CpG sites is only scratching the surface of the epigenome. Several reports have shown that the most variable sites of the genome are in intragenic CpG sites (gene-body) which are only sparsely covered by the 450k array. We have previously hypothesised that the most variable region of the genome is in the first few kb of a gene, which can only be targeted using sequencing based methods [90]. Given that NGS based methods will soon be used for clinical genetics, if any of the epigenetic markers we identify could become clinically useful, then using this same technology would be feasible. Therefore, the future challenges will be finding an appropriate method at a reasonable cost to allow such studies with large enough study sizes to have sufficient power for EWAS [91]. This new approach will undoubtedly lead to further debate about appropriate study design, statistical methods and bioinformatic processing of this data with new biases and confounding factors to take into account. Already, it is clear that the international consortium based approaches will be needed to combine datasets, whether for 450k arrays or Bis-Seq data, to approach the level of replication and robustness that have been achieved with GWAS consortia, and several EWAS consortia have recently been initiated to do just that.

Acknowledgements

JMF is funded by Breast Cancer Campaign and Cancer Research UK.

References


42. Ramagopalan SV, Rakyan VK (2013) The promise and challenges of blood spot methyloomics. Epigenetics 8:775–777


DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect 120:1425–1431


81. Philibert RA, Beach SR, Lei MK, Brody GH (2013) Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. Clin Epigenetics 5:19


