Review

For reprint orders, please contact: reprints@futuremedicine.com

Epigenomics

How to make DNA methylome wide association studies more powerful

Genome-wide association studies had a troublesome adolescence, while researchers increased statistical power, in part by increasing subject numbers. Interrogating the interaction of genetic and environmental influences raised new challenges of statistical power, which were not easily bested by the addition of subjects. Screening the DNA methylome offers an attractive alternative as methylation can be thought of as a proxy for the combined influences of genetics and environment. There are statistical challenges unique to DNA methylome data and also multiple features, which can be exploited to increase power. We anticipate the development of DNA methylome association study designs and new analytical methods, together with integration of data from other molecular species and other studies, which will boost statistical power and tackle causality. In this way, the molecular trajectories that underlie disease development will be uncovered.

Xinyi Lin[1], Sheila Barton[2] & Joanna D Holbrook*,[1]

[1]Singapore Institute for Clinical Sciences (SICS), Agency for Science & Technology Research (A*STAR), Brenner Centre for Molecular Medicine, 30 Medical Drive, 117609, Singapore

[2]MRC Lifecourse Epidemiology Unit, Faculty of Medicine, University of Southampton, Southampton, SO16 6YD, UK

*Author for correspondence: joanna_holbrook@sics.a-star.edu.sg

First draft submitted: 23 February 2016; Accepted for publication: 23 March 2016; Published online: 7 April 2016

Keywords: epigenome-wide association study (EWAS) • developmental trajectories • DNA methylation • DNA methylation association studies (methWAS) • epigenetic epidemiology • gene × environment (GxE)

The influences of nature and nurture on human health have been studied for hundreds of years. In the last few decades, derivations of genome-wide association studies (GWAS) have interrogated the interplay of the two, either by incorporating genetic and environmental measures in gene × environment screens or by screening the DNA methylome (as the latter can be thought of as the product of genetic and environmental influences). Both approaches are afflicted by problems of low statistical power to detect significant associations, and in both cases increasing subject numbers to boost power has practical impediments. As we enter the era of the methylome-wide association-study (methWAS), we discuss ways to increase statistical power appropriate to these types of studies.

Health is a result of the interaction of genes with environment

Increasingly, we understand that genetics and prior environmental exposures determine an individual's sensitivity and resistance to extrinsic influences. This can be expressed in terms of negative symptomology, for example, a genotypic group could be more sensitive to the consequences of a high-fat diet while another genotype group is relatively resistant. For example, Asians experience higher risk of hypertension, cardiovascular disease and diabetes at lower BMIs compared with other racial groups [1–3]. It can also be expressed in terms of plasticity wherein a genotypic group does worse in a 'bad' environment but better in a 'good environment'. For instance, the long allele (5-HTTLPR L) genotype of the variable tandem repeat (VNTR) within

SLC6A4 is associated with an increased risk for affective disorders under adverse conditions but also with a decreased risk under more favorable settings [4]. In another example, polymorphisms in *ESR1* moderate the effects of family cohesion on age of menarche. Specifically, girls with a GG genotype at *ESR1* polymorphism rs9304799 experience puberty later in a high quality family environment and much earlier puberty in a low quality family environment. In contrast, the impact of family environment on age of menarche in AG girls is less and has no effect on age of menarche in girls with a AA genotype [5,6].

Diseases such as obesity, diabetes, hypertension, depression, schizophrenia and coronary heart disease are major public health issues with high economic costs and significant consequences on quality of life. Their pathogenesis starts long before the symptoms are apparent and their etiology comprises both genetic and environmental components. They are the consequence of the interplay of genes and environment and cannot be sufficiently explained by their separate (or marginal) effects. Detecting which environmental exposures and which genetic variants are causative and in which combinations, is an important task to enable intervention and prevention [7].

Screens for genetic, environmental & GxE influences are afflicted by low statistical power

Screens for environmental influences on health have tended to be hypothesis driven, but hypothesis-free, large-scale screens for genetic variants have had their own productive era in terms of GWAS [8]. The number of tests inherent in a GWAS has highlighted problems of statistical power to detect significant associations. Typically millions of genetic variants are assayed exceeding the number of subjects contained within the study, leading to low statistical power. Once a statistical modeling approach is chosen, factors influencing statistical power include:

- Required significance level (chosen to protect overall Type 1 error) – less stringent level implies higher statistical power.
- Magnitude of true effect size – higher effect size implies higher statistical power.
- Sample size – higher sample size implies higher statistical power.
- Noise (unwanted variability) – more noise implies less statistical power.

A type 1 error rate of 0.05 is typically used and multiple testing corrections are applied to ensure this error rate is maintained across all tests conducted. For a GWAS, which assumes all common genetic variation has been covered, the uncorrected p-value required to claim significance for an individual test is $p < 5 \times 10^{-8}$ (this corresponds to a genome-wide 0.05 type 1 error rate maintained across one million independent tests) [9–11]. GWAS researchers have tackled the problem of a stringent required significance level, by increasing the number of subjects, either in individual studies or by combining studies through meta-analysis, thereby boosting statistical power.

However when interaction with environmental components is included in a genome-wide screen, increasing subject numbers is not always feasible. Adding more subjects for genetic characterization has economies of scale but environmental observations on more subjects do not [12]. Genetic information can be obtained from a one-time collection of DNA sample and running genotyping assays for the additional subjects incurs only incremental cost and little additional time. In contrast, collecting information on environmental exposures from additional subjects, such as measuring their fasting blood glucose levels, costs the same, has the same subject burden and takes the same amount of clinic time for subject 2000 as it did for subject 1. Also and importantly, testing for an interaction itself requires larger sample sizes to achieve the same statistical power due to additional variability associated with the estimated interaction term [13]. An oft-quoted ‘rule of thumb’ is that detection of an interaction requires a sample size at least four-times larger than that required for the detection of a main effect of comparable magnitude [14]. This may be one of the reasons why the wealth of literature on GxE effects (especially in the psychological sciences) has not translated to a raft of searches for GxE associations in genome-wide data. These studies, variously called genome–environment-wide interaction (GEWI) and gene × environment wide association (GxEWA) studies have been suggested [15,16]. However in actuality, candidate gene approaches, for example [17]; or dimension reduction using polygenic risk scores, for example [18]; or GxE test of candidates identified in a GWAS, are usually applied.

Utilizing DNA methylation marks as a proxy for GxE

Happily, there are intermediate markers, which integrate gene and environment effects and can be tested against phenotype. Genome-wide DNA methylation marks can be assayed by microarray [19,20] (although these covers only a small fraction of the methylome) or emerging NGS approaches [21,22] and therefore methWAS can be conducted. DNA methylation marks are a product of the interaction between genes and environments. One unusually complete example is the methylation state of *FKBP5*, which is decreased in response to childhood trauma only in carriers of a risk allele. Methylation state goes on to predict stress reactivity and risk of psychopathology in adulthood [23,24]. Teh *et al.* [25] showed that the majority of variation in the neonate DNA methylome is best explained by an interaction of an SNP and a prenatal environment (compared with the main effect of a SNP or an environment).

A phenotype may be the result of many polymorphisms and environmental factors. It is difficult to model all these possibilities in a screen. However DNA methylation marks are downstream of all these factors and therefore combine multiple inputs. MethWAS can be conducted as an alternative to GEWIS, without the need to collect complex environmental data or model the different ways interactions can occur [26]. This is notwithstanding the fact that DNA methylation marks work in concert with each other and may interact to cause phenotypes. Similarly, epigenetic marks may act in concert with genetic or environmental factors to affect phenotype as elegantly described by Ladd-Acsta and Fallin [12]. There is certainly an argument for conducting methylation × genetics or methylation × environments or methylation × methylation studies. However testing methylation marks against phenotype is already a more comprehensive study of genetic and environmental influences than can be hoped for using a finite number of environmental measures and simplistic statistical models to integrate them with genotype, as would be done in GEWIS.

As has been ably covered elsewhere [27–29], researchers conducting methWAS must tackle decisions such as: the platform to use to measure the DNA methylome, the appropriate tissue to sample and the timepoint to measure both the methylome and environmental exposure or phenotype. Besides these crucial decisions to ensure a well-designed study, they must find ways of boosting statistical power and appropriately analyzing DNA methylation which is a continuous data-type (a marked contrast from categorical DNA polymorphism data). There is reason to hope that effect sizes will be larger for epigenetic marks than for genotypic ones, as some published studies seem to suggest [30–32]. Already significant and replicated associations have been reported from large studies, for example, *HIF3A* methylation associated with BMI [33] and *AHRR* methylation for smoking [34], and extended by independent groups, for example [35–39] and [40], respectively. When possible, the study should be sufficiently powered with adequate sample size. However, due to the increased difficulty and cost in sampling the relevant tissue (at sufficient depth and resolution) for methWAS compared with GWAS, it is unlikely that methWAS studies will achieve the same sample numbers as GWAS. Therefore, statistical power must be boosted by other means, by exploiting the intrinsic characteristics of DNA methylation data to reduce unwanted variability and by efficient statistical modeling.

Increasing power of methWAS by reducing (unwanted) variability

One of the factors that negatively affects statistical power is unwanted variability. For methylation data, one of the key sources of unwanted variability is cellular heterogeneity [41–43]. The most efficient way to reduce unwanted variation caused by cellular heterogeneity is to investigate methylation in more homogeneous samples. For example, blood can be fractionated and DNA methylation investigated in specific cell types, for example [44], and more precisely from specific subpopulations, for example [45–47]. Another possibility for blood-based study is to directly measure the cell count in the DNA sample and adjust methylation data accordingly, for example [42,48]. However, such measures are not always feasible, for instance when studying previously frozen bloods or tissues that are by nature heterogeneous and not easily fractionated (e.g., placenta). In these scenarios, cellular heterogeneity has to be accounted for using statistical approaches, either by deriving an estimate of cellular proportions using an independent dataset of methylation profiles of the individual cell types and adjusting for these derived cell type proportions (reference-based adjustment) [49], or directly adjusting without inferring the cellular proportions (reference-free adjustment) [50]. These approaches, although absolutely necessary in methylome data from mixed cell type samples [42,43], have limitations. Reference-based adjustments require the use of an appropriate cell-specific methylation reference panel. When it is unclear if the available reference-panel is appropriate for the specific study (e.g., an adult blood reference panel might not be appropriate for use in investigating infant cord blood methylation [51]) it might be useful to construct reference-panels that are appropriate to the study. For example, a study investigating infant cord blood methylation might fractionate and assess methylation in a few fresh (non-study) cord blood samples to construct a reference panel, which is less tedious than fractionating or obtaining cell counts in all the study samples. It is also important to optimize the performance of this procedure by selecting the most informative cell type markers from the methylome [52]. The reference-free approaches calculated from the dataset under study, are limited in their precision and may remove (wanted) biological variability from the data and hence reduce statistical power.

Other major sources of unwanted variability in methylome data are batch effects. In methWAS known sources of batch effects are bisulfite conversion batch, experimental batch, chip and position on chip for array studies [53] and reagent set and order for sequencing runs [54,55]. Minimizing batch effects from these sources can be done by designing the study to ensure the phenotype of interest is not confounded by predictable batch effects, optimizing laboratory procedures and by statistical approaches to correct observed batch effects in the data [20]. There is scope to further improve the processing of DNA methylation data. For instance current processing methodologies for Illumina450K data assume that the methylated and unmethylated signals form independent gamma distributions, which is obviously not true [63]. Last, there are known sources of variability that are not necessarily the variable of interest, such as sex [56], ethnicity [42] and age [57,58], which can be adjusted for in downstream analyses.

Increasing power of methWAS by appropriate statistical modeling

At an individual site, in an individual cell, methylation is either 0 or 100%. However in tissue samples, a mixture of cells is assayed to give an average methylation percentage at any given site. Thus percent methylation values are continuous and range from 0 to 100. Methylation levels at particular sites are often not normally (Gaussian) distributed across samples; they can be bimodal, profoundly skewed, or multimodal. Extreme values of highly methylated and highly unmethylated sites show reduced variance compared with intermediate values [59,60]. Both of these facts violate assumptions made by classical statistical techniques such as ordinary least squares regression which assume normality and constant variance of model residuals (errors) [61]. Violation of statistical assumptions can lead to large numbers of false-negative results and therefore loss of power [62]. Various approaches can be used to surmount problems arising from methylation value distributions. One solution is to logit transform the percent methylation values (often termed beta values) to M values [59] but this does not always result in an appropriate error distribution with constant variance [63]. Other solutions include modeling using robust regression to minimize the effect of outliers or robust standard errors for regression coefficients to deal with heteroskedasticity [62], modeling using non-normal errors [63,64] or using nonparametric techniques which rely on ranks rather than the methylation values [65]. Bayesian approaches have also been used to shrink estimated sample variances toward a pooled estimate [66].

Many modern platforms for methylome assay (e.g., methyl-capture sequencing, Infinium arrays) measure methylation sites at single base resolution. Most published methWAS analyzed each site individually for association with exposure/phenotype of interest, for example [22,33]. The individual site analysis is then corrected for multiple testing using Bonferroni [67] or false discovery rate (FDR) [68,69]. Recently, a new approach to FDR that is less stringent has been suggested [70]. Its appropriateness in methWAS hinges on the assumption of unimodality, in other words, whether the effect size has a common mode. The assumption might be violated when comparing old age to young age samples because there are global methylome changes with aging, where both hypomethylation and hypermethylation can occur. However, it may be reasonable for phenotypes, which have subtle locus-specific effects on the methylome.

As an alternative to individual site analysis, grouped-site analysis, where one assesses the collective association between a group of sites and phenotype, can be employed. The advantages a grouped analysis confers over individual marker analysis have been well investigated in the context of GWAS and many of the same reasons apply here. Briefly, an individual marker analysis could be an erroneous signal or the analysis suffers from power loss because the testing procedure ignores the correlation of the tested markers and does not allow joint effects of marks to be modeled. The inefficiency due to multiple testing becomes particularly relevant as the number of sites (in methWAS usually CpGs) being tested increase with the newer arrays and sequencing platforms. For example, the widely used Illumina Infinium HumanMethylation450 beadchip array includes 485,512 sites [19] while the new Illumina Infinium MethylationEPIC beadchip includes 856,187 sites [71]. Sequencing technologies can in principle assay all approximately 28 million CpGs in the human genome, but in practice are likely to assay an order of magnitude less (a recent study using methyl-capture sequencing of clinical samples assayed approximately 2–3 million methylation sites [72]). As methylation at CpGs show patterns of correlation across the epigenome, in principle, we can estimate the effective number of independent tests/CpGs across the epigenome and adjust for this number, as opposed to all the 2–28 million CpGs, similar to what was done in GWAS [9–11]. However, this is not a trivial task as, unlike genotype, DNA methylation is variable across cell-type, tissue and age so this estimate would have to be performed for each combination of cell combination, tissue and age tested.

Grouped-CpG analysis decreases the number of tests by aggregating co-varying CpGs. This is especially attractive, for higher coverage methods to reduce the number of tests while retaining the broader survey of interindividual variation. Another advantage is that multiple-testing corrections often assume independence between tests and grouping-CpGs improves the fit of the data to this assumption. Existing grouped-CpG analysis can be classified into two classes based on how the groups are defined: groups are defined in testing procedure and groups are formed *a priori*. Methods that group the CpGs during the testing procedure include region discovery [73], bump hunting [74] and comb-p [75]. These tests typically first test each CpG for association between exposure/phenotype, and then define the region of differential methylation using the individual marker analyses. They have the advantage of prioritizing association with the phenotype to form the groups. However, when the groups are defined as part of the testing procedure, computationally intensive permutation procedures are generally required to obtain a p-value that maintains the correct Type 1 error control [76]. On the other hand, when the groups are defined and formed *a priori* (i.e., without using exposure/phenotype to define the groups), multiple testing correction is straightforward via standard procedure over the total number of groups tested, providing a computationally efficient alternative. Examples of tests include forming groups using genomic annotations [77], such as genes, pathways or CpG islands. For example, if groups were formed using CpG island context (~29K CpG islands in genome, each island consisting of all CpGs within that island would represent one test), the bonferroni threshold to maintain a Type 1 error rate at 0.05 would be $0.05/(29 \times 10^3)$. However, this approach risks reducing power, unless the collapsing approach is carefully chosen (see next paragraph). Alternatively, groups can be formed using the methylation data to find spatially correlated CpGs data such as in adjacent sites clustering (A-clustering) [78], or using methylation data to group CpGs based on co-varying networks as was done in weighted correlation network analysis (WGCNA) [79]. The choice of how groups are formed can affect both the statistical power and interpretation of results. The ideal choice remains an open research question, and is likely to depend on the choice of platform for methylation assay which determines the scarcity of the CpG measures as well as underlying CpG co-varying architecture and phenotype under study. For example, combining a CpG with an effect on phenotype with other non co-varying CpGs that do not affect the phenotype could introduce substantial noise and reduce the power.

Another factor that affects statistical power, is how information is aggregated across the CpGs within the group. In the simplest case, information is collapsed at the CpG level into a single summary variable (e.g., mean methylation) [80,81]; and one tests for association between this summary variable and the phenotype. This test highly resembles the collapsing/burden tests used in the analysis of rare variants sequencing association studies (an example of a collapsing test is to use the mean or total number of rare variants in the group as a summary variable), and implicitly assumes that all CpGs within the group share the same effect size (and hence can be represented using a single regression coefficient) [82,83]. However, if the effects of the CpGs on phenotype have opposite signs or if only few CpGs within the group show association with the phenotype, this simple collapsing approach would be highly inefficient. Another approach to aggregating information at the CpG-level is to conduct a principal component analysis (PCA) on the CpGs within the group and test the resulting principal components (PC) for association with the phenotype. WGCNA [79] utilizes such an approach where the first PC for CpGs within the module (denoted the eigengene/eigenCpG) is compared with the phenotype. This analysis offers the chance to determine 'emergent properties' of the data. However, the power of this approach is highly dependent on the appropriate number of PCs used, for example, if only the first PC is used and important information is captured in the second PC, this approach can suffer from power loss [84]. An alternative method is to aggregate information at the test statistics level, such as in kernel machine regression [77]. This method offers an advantage by allowing effects to have different signs and boosts power by varying the degrees of freedom of the test depending on the correlation of tested terms. Another closely related method aggregates information at the p-value level and uses weighted inverse χ^2 method to account for correlation in the p-values [85].

Increasing power of methWAS by leveraging data from multiple molecular species

A complementary approach to increase power in methWAS is to integrate omics data from other molecular species such as genetics, sequence metrics and chromatin states and transcriptomics since DNA methylation is influenced by genotype [25], the broader characteristic of the sequence context [86], and chromatin state [87] and both influences and is influenced by transcription [88–90]. When integrating information from the transcriptome, it is essential that the transcriptome is measured in the same tissue as interrogated for the DNA methylome and appropriate for the phenotype in question. In the simplest case, data from other molecular species is used as a filtering criterion to reduce the number of CpGs tested. For example, if genotype and methylation data are both available, analysis could be restricted to only methylation marks showing association with genotype. In this way power is increased as the number of tests is reduced. For example, in a study investigating methylation as an intermediary for genetic risk in rheumatoid arthritis, Liu *et al.* [91] required candidate CpG ‘mediators’ to have methylation levels which co-vary with in *cis* genotype. If transcription and methylation data are available, one could restrict the CpGs tested to those affecting transcription (i.e., CpG showing association between transcription and methylation) [92]. However, this approach is very conservative as a methylation mark could affect transcription in different tissues or conditions to the one tested, and could do so in a nonlinear way, which may not be detected by the analysis. So false positives are reduced at the expense of increasing false negatives. Both of these examples do not examine the association between phenotype and other molecular species. Alternatively, the associative screen is first conducted for the methylome and other molecular species separately and then the nominally significant marks are mapped to a common identifier, often genes or pathways. Identifiers that show statistical significance for multiple species are meta-analyzed and prioritized. These methods increase power by decreasing the required significance level for any one type of molecular species and instead requiring association of phenotype to be present across molecular species, for example [93,94] but again they risk missing true positive because of the stringent criteria that the molecular levels be linearly correlated in the cell type(s) and conditions tested.

The different molecular species can also be jointly analyzed in the association analysis. When jointly analyzing different molecular species with measurements of the same nature (e.g., continuous measurements for both methylation and transcription), aggregating information at the data level is possible. Methodologies suggested for this type of analysis include factor analysis performed on data from all molecular types, and then constrained (for example by linear discriminate analysis or SVA) to the phenotype [95]. Another possibility is to apply WGCNA [79] with the co-varying network modules constructed using all molecular species together (each molecular species is appropriate scaled). When jointly analyzing molecular species with different types of measurements (e.g., categorical data from genotype vs. continuous data from methylation), analysis methods that aggregate information at the test statistics-level are generally more appropriate. For example, Zhao *et al.* [77] employed kernel machine regression and constructed a test of the joint effects of a group of SNPs and a group of CpGs on a phenotype. New approaches may also integrate stand-specific information from DNA methylome and transcriptome data and detect the association of allele-specific methylation (ASM) and expression (ASE) with disease. Regions of ASM and ASE are variable across individuals (as they are often a consequence of in *cis* polymorphisms affecting DNA methylation states) and across tissues within the same individual [96] suggesting tissue-specific functional roles. Again there is an assumption that all molecular species in a network will show coordinated behavior in a manner that is amenable to the statistical tools and in the tissue and condition tested.

Increasing power & addressing causality in methWAS by leveraging information from multiple timepoints

Simplistically there are three possibilities that can explain a true observed association between two variables (besides chance and selection bias). (i) The first variable causes the second variable; (ii) the second variable causes the first variable; (iii) a third factor (confounder) is a common cause of both variables. In interpreting findings from observational studies, one has to address the question of whether the associations are likely to be causal (scenario i or ii) or due to confounding (scenario iii). In this respect, GWAS findings are more readily interpreted because genetic variation occurs upstream of other influences and can not be influenced by them. Therefore, while causal mediation methods have been applied in GWAS [97] their application in methWAS is less straightforward. As DNA methylation is dynamic across the human lifecourse, methWAS findings have to contend with issues of confounding by genetic and/or environmental factors and, in the absence of confounding, directionality of causation. Furthermore, the causal possibilities can co-occur, for instance a methylation mark can cause disease but development of the disease can then affect methylation at the same CpG. Therefore, attempting to determine directionality of effects from a observational methWAS study with data collected from a single time point is extremely challenging. For example, Figure 1A–C presents a simple scenario where methylation at a disease-associated CpG is stable with age and measured at/after disease onset (as happens in cross-sectional or case–control studies, for example [91]). We can observe a positive correlation between methylation and disease at T_1 but it is not possible to discern between causality in either direction (Figure 1A vs 1B) or the confounding case (Figure 1C). If the DNA sample was collected prior to disease onset, for example in prospective cohort studies, the temporal relationship is easier to establish (Figure 1D–F) and it is possible to discern between the casual scenarios (Figure 1D & 1E). Nevertheless, it remains impossible to rule out confounding (Figure 1F). Some, studies have

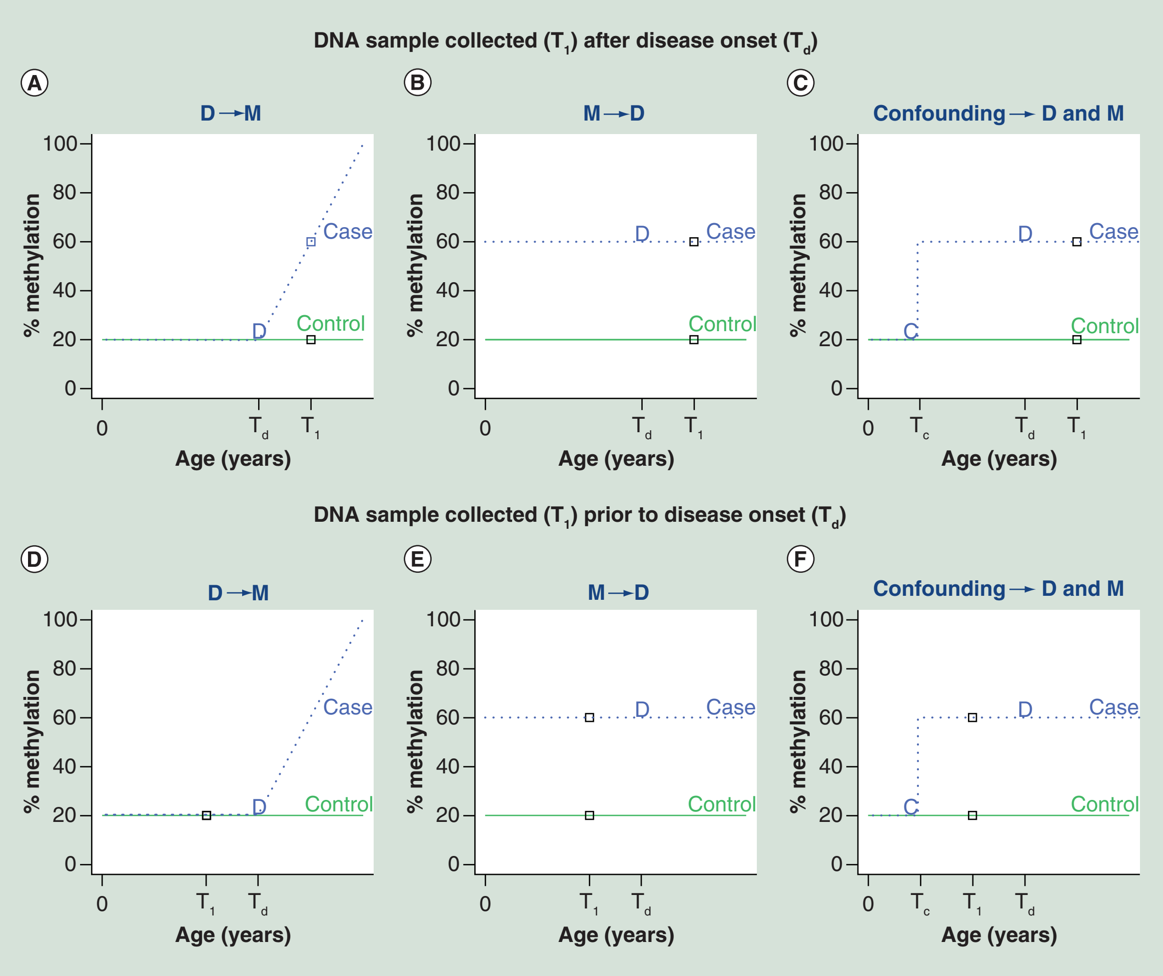

Figure 1. Simplified DNA methylation trajectories for a subject without (green solid line) and with disease (dotted blue line) where (A & D) methylation changes as a consequence of disease, or (B & E) disease occurs as a consequence of methylation, or (C & F) there is no causal relationship between disease and methylation (a confounding factor independently affects both disease status and methylation), respectively. T1 and Td represent times when DNA sample is collected and disease occurred, respectively (these events are also represented with a square and 'D'). Tc is the time when a confounding factor (e.g., environmental exposures) affects methylation (denoted with a 'C'). In the top panel **(A–C)**, DNA sample is collected after disease onset, we observe a positive association between methylation and disease, but cannot distinguish between each of the three scenarios D->M, M-> D and confounding. In the bottom panel **(D–F)**, DNA sample was obtained prior to disease onset, allowing us to rule out D -> M, but both M -> D and confounding are still possibilities.

used Mendelian randomization [39,98] and mediation analysis [91] to address causality. These methods require strong assumptions including the assumption that the genetic instrument (or any polymorphisms they are in linkage disequilibrium with) can affect the phenotype only through methylation and not through any other pathways. They also require very large sample sizes or very high effect sizes to achieve adequate statistical power [99].

Jointly interpreting findings from multiple observational studies with different study designs could provide better clues as to the directionality of causation. In a recent study, DNA methylation was implicated as a mediator of the effects of genetic variants on the development of hypertension. SNPs significantly associated with hypertension were identified in a replicated GWAS study in adults. In a subpopulation of those adults, methylation marks in *cis* with the associated SNPs were also strongly associated with the phenotype. Additionally, in a (separate) population of neonates who are not affected by hypertension, the same methylation marks still associated with the identified SNPs. This suggests that the variation in methylation is upstream of the phenotype, not a consequence of it [100]. Data from different age groups could thus allow us to better understand the directionality of the associations. In another example, a replicated methWAS, identified the association of methylation of *HIF3A* and adult BMI [33]. The authors speculated that *HIF3A* hypermethylation was a consequence of increasing adiposity. However, the association of *HIF3A* methylation and weight was also detected in neonates suggesting it was not solely a consequence of adult acquired adiposity.

Alternatives to interpreting findings from multiple observational studies are longitudinal studies that prospectively sample the methylome and collect environmental influences at multiple time points. Longitudinal studies better allow us to map methylation trajectories alongside disease progression, and environmental influences. Therefore, they are powerful in examining causality [27,101]. This aim is central to efforts of longitudinal cohorts with epigenetic sampling [39,102–104]. Additionally, multigenerational studies will allow investigation of epigenetic mediation of transgenerational transmission, for example, of metabolic profile [105–107], lifespan [108] or psychophysiological trauma [109].

Future perspective

Boosting statistical power in individual studies is desirable. Statistical power can be boosted by increasing sample size to logistical limits, although it is difficult to estimate the required sample size *a priori* (as is often required by grant awarding bodies) as variance of continuous methylation levels and effect size on phenotypes are usually unknown. Statistical power

Executive summary

Health is a result of the interaction of genes with environment

- Complex diseases are a product of the interaction of genetic predisposition and environmental influences.

Screens for GxE influences are afflicted by low statistical power

- Hypothesis-free large-scale screens for genetic variants influencing disease (genome-wide association studies [GWAS]) require thousands of subjects to achieve statistical power to detect true associations.
- Investigating interactions with environmental influences in these screens, reduced power further.
- Moreover, increasing number of subjects is nontrivial as collection of environmental measures (especially longitudinal measurements) is costly and burdensome.

Utilizing DNA methylation marks as a proxy for GxE

- An alternative is to study the DNA methylome.
- DNA methylation marks are putatively downstream of multiple causative genetic and environmental factors and upstream of disease.
- However, as the DNA methylome must be measured in an appropriate tissue and at appropriate time(s), methylome-wide association studies (methWAS) are challenging to perform in massive numbers of subjects. Therefore, statistical power must be boosted by other means.

Increasing power of methWAS by reducing (unwanted) variability

- A key source of unwanted variability in methWAS studies is introduced by cellular heterogeneity. This can be tackled by study design, collection of cell count data and statistical adjustment.
- Other sources of unwanted variability can also be partially bested by statistical modeling.

Increasing power of methWAS by appropriate statistical modeling

- Statistical models employed within methWAS must account for the non-normal distribution and heteroskedascity of the data.
- Appropriate multiple testing procedures must also be applied, without causing excessive loss of power.
- Grouping CpGs and testing grouped-CpGs as a single unit reduces the number of variables to test and so increases power.

Increasing power of methWAS by leveraging data from multiple molecular species

- Combining methWAS with data from other molecular species (e.g. genotype and transcriptome) can decrease required significance levels and so increase statistical power.
- It can also detect the ‘emergent properties’ of the data.

Increasing power & addressing causality in methWAS by leveraging information from multiple timepoints

- Causality and confounding are important issues in methWAS.
- It is challenging to address them in single timepoint data.
- However, there are promising clues emerging from studies that combine observations made at different stages of the lifecourse.
- We look forward to longitudinal and multi-generational studies that incorporate DNA methylation measures.

Future perspective

- Replication across independent studies and sample sets is critical for the future of methWAS, especially to address confounding.
- We look forward to taking full advantage of existing datasets to boost the power of methWAS studies and uncover molecular trajectories that underlie disease.

can be boosted through careful study design and data analysis as described here. Lessons can be learnt from the evolution of GWAS studies but DNA methylation is not analogous to DNA polymorphisms and different approaches are necessary. Statistically significant results remain vulnerable to confounding by unknown factors in the data, which could drive a spurious association between phenotype and DNA methylation marks. Replication in independent datasets, as has become standard for GWAS is perhaps even more critical for methWAS. Additionally, meta-analysis can be conducted across studies to boost the total sample size. Large-scale epigenome mapping initiatives have provided hugely important information about DNA methylation across tissues and individuals and their relationship with other molecular marks [110–112]. These provide methWAS studies with important context, which could be utilized to enhance statistical power, for instance in more appropriate grouping of CpGs and modeling of the interaction with other molecular species. As the field goes on to produce more data from samples on the continuum of health and disease, development of analytical methods for the interpretation and integration of methWAS data while conserving power, will become an even more important research challenge. MethWAS offers hope of the confident identification of DNA methylation marks, which are downstream of genetic and environmental influences but upstream of disease. Their discovery, in part relies on increasing the statistical power of methWAS.

Financial & competing interests disclosure

This work is supported by the Translational Clinical Research (TCR) Flagship Program on Developmental Pathways to Metabolic Disease funded by the National Research Foundation (NRF) and administered by the National Medical Research Council (NMRC), (Singapore- NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014). SJ Barton is supported by the Epigen Consortium. This work was supported by grants from the Medical Research Council (MC_U147585827, MC_ST_U12055), British Heart Foundation (RG/07/009), Arthritis Research UK, National Osteoporosis Society, International Osteoporosis Foundation, Cohen Trust, and NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

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

Open access

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Stommel M, Schoenborn CA. Variations in bmi and prevalence of health risks in diverse racial and ethnic populations. *Obesity (Silver Spring)* 18(9), 1821–1826 (2010).

2 Consultation WHOE. Appropriate body-mass index for asian populations and its implications for policy and intervention strategies. *Lancet* 363(9403), 157–163 (2004).

3 Chambers JC, Loh M, Lehne B *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case–control study. *Lancet Diabetes Endocrinol.* 3(7), 526–534 (2015).

4 Caspi A, Sugden K, Moffitt TE *et al.* Influence of life stress on depression: moderation by a polymorphism in the 5-htt gene. *Science* 301(5631), 386–389 (2003).

5 Hartman S, Widaman KF, Belsky J. Genetic moderation of effects of maternal sensitivity on girl's age of menarche: replication of the Manuck *et al.* study. *Dev. Psychopathol.* 27(3), 747–756 (2015).

6 Manuck SB, Craig AE, Flory JD, Halder I, Ferrell RE. Reported early family environment covaries with menarcheal age as a function of polymorphic variation in estrogen receptor-alpha. *Dev. Psychopathol.* 23(1), 69–83 (2011).

7 Holbrook JD. An epigenetic escape route. *Trends Genet.* 31(1), 2–4 (2015).

8 Arking D, Rommens J. Editorial overview: molecular and egnetic bases of disease: enter the post-GWAS era. *Curr. Opin. Genet. Dev.* 33, 77–79 (2015).

9 Barsh GS, Copenhaver GP, Gibson G, Williams SM. Guidelines for genome-wide association studies. *PLoS Genet.* 8(7), e1002812 (2012).

10 Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* 32(4), 381–385 (2008).

11 Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* 32(2), 179–185 (2008).

12 Ladd-Acosta C, Fallin MD. The role of epigenetics in genetic and environmental epidemiology. *Epigenomics* 8(2), 271–283 (2015).

13 Leon AC, Heo M. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Comput. Stat. Data Anal.* 53(3), 603–608 (2009).

14 Smith PG, Day NE. The design of case–control studies: the influence of confounding and interaction effects. *Int. J. Epidemiol.* 13(3), 356–365 (1984).

15 Thomas D. Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* 11(4), 259–272 (2010).

16 Khoury MJ, Wacholder S. Invited commentary: From genome-wide association studies to gene–environment-wide interaction studies–challenges and opportunities. *Am. J. Epidemiol.* 169(2), 227–230; discussion 234–225 (2009).

17 Melkonian SC, Daniel CR, Ye Y *et al.* Gene–environment interaction of genome-wide association study-identified susceptibility loci and meat-cooking mutagens in the etiology of renal cell carcinoma. *Cancer* 122(1), 108–115 (2016).

18 Mullins N, Power RA, Fisher HL *et al.* Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychol. Med.* 46(4), 759–770 (2015).

19 Bibikova M, Barnes B, Tsan C *et al.* High density DNA methylation array with single cpg site resolution. *Genomics* 98(4), 288–295 (2011).

20 Pan H, Chen L, Dogra S *et al.* Measuring the methylome in clinical samples: improved processing of the infinium human methylation450 beadchip array. *Epigenetics* 7(10), 1173–1187 (2012).

21 Teh AL, Pan H, Lin X *et al.* Comparison of methyl-capture sequencing vs. Infinium humanmethylation450 array for methylome analysis in clinical samples. *Epigenetics* 11(1), 36–48 (2016).

22 Allum F, Shao X, Guenard F *et al.* Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat. Commun.* 6, 7211 (2015).

23 Klengel T, Binder EB. Fkbp5 allele-specific epigenetic modification in gene by environment interaction. *Neuropsychopharmacology* 40(1), 244–246 (2015).

24 Klengel T, Binder EB. Allele-specific epigenetic modification: a molecular mechanism for gene–environment interactions in stress-related psychiatric disorders? *Epigenomics* 5(2), 109–112 (2013).

25 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).

•• **Demonstration that GxE is the best explanation for the majority of interindividual variation in neonate DNA methylomes.**

26 Hernandez L, Blazer D. Study design and analysis for assessment of interactions - definitions of interactions. In: *Genes, Behaviour, and the Social Envrionement: Moving Beyond the Nature/Nurture Debate.* Hernandez L, Blazer D (Eds). National Academires Press (US), Washington, DC, USA (2006).

27 Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 12(8), 529–541 (2011).

28 Heijmans BT, Mill J. Commentary: the seven plagues of epigenetic epidemiology. *Int. J. Epidemiol.* 41(1), 74–78 (2012).

29 Michels KB, Binder AM, Dedeurwaerder S *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nat. Methods* 10(10), 949–955 (2013).

•• **Authoritative recommendations for design and analyses of EWAS.**

30 Godfrey KM, Sheppard A, Gluckman PD *et al.* Epigenetic gene promoter methylation at birth is associated with child's later adiposity. *Diabetes* 60(5), 1528–1534 (2011).

31 Lillycrop KA, Costello PM, Teh AL *et al.* Association between perinatal methylation of the neuronal differentiation regulator hes1 and later childhood neurocognitive function and behaviour. *Int. J. Epidemiol.* 44(4), 1263–1276 (2015).

32 Desai M, Jellyman JK, Ross MG. Epigenomics, gestational programming and risk of metabolic syndrome. *Int. J. Obes.* 39(4), 633–641 (2015).

33 Dick KJ, Nelson CP, Tsaprouni L *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* 383(9933), 1990–1998 (2014).

34 Philibert RA, Beach SR, Lei MK, Brody GH. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clin. Epigenetics* 5(1), 19 (2013).

35 Pan H, Lin X, Wu Y *et al.* Hif3a association with adiposity: the story begins before birth. *Epigenomics* 7(6), 937–950 (2015).

36 Huang T, Zheng Y, Qi Q *et al.* DNA methylation variants at hif3a locus, b vitamins intake, and long-term weight change: gene-diet interactions in two US cohorts. *Diabetes* 64(9), 3146–3154 (2015).

37 Monastero R, Garcia-Serrano S, Lago-Sampedro A *et al.* Methylation patterns of VEGFB promoter are associated with gene and protein expression levels: the effects of dietary fatty acids. *Eur. J. Nutr.* doi:10.1007/s00394-015-1115-7 (2015) (Epub ahead of print).

38 Demerath EW, Guan W, Grove ML *et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum. Mol. Genet.* 24(15), 4464–4479 (2015).

39 Richmond RC, Sharp GC, Ward ME *et al.* DNA methylation and body mass index: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* pii: db150996 (2016) (Epub ahead of print).

40 Novakovic B, Ryan J, Pereira N, Boughton B, Craig JM, Saffery R. Postnatal stability, tissue, and time specific effects of ahrr methylation change in response to maternal smoking in pregnancy. *Epigenetics* 9(3), 377–386 (2014).

41 Liang L, Cookson WO. Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. *Hum. Mol. Genet.* 23(R1), R83–R88 (2014).

42 Lam LL, Emberly E, Fraser HB *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proc. Natl Acad. Sci. USA* 109(Suppl. 2), 17253-17260 (2012).

43 Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 15(2), R31 (2014).

44 Reinius LE, Acevedo N, Joerink M *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* 7(7), e41361 (2012).

45 Kulis M, Merkel A, Heath S *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* 47(7), 746–756 (2015).

46 Martino D, Joo JE, Sexton-Oates A *et al.* Epigenome-wide association study reveals longitudinally stable DNA methylation differences in CD4$^+$ T cells from children with IGE-mediated food allergy. *Epigenetics* 9(7), 998–1006 (2014).

47 Heyn H, Li N, Ferreira HJ *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc. Natl Acad. Sci. USA* 109(26), 10522–10527 (2012).

48 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).

49 Houseman EA, Accomando WP, Koestler DC *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).

- **Highlightes the issue of cellular heterogeneity and suggests a solution.**

50 Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30(10), 1431–1439 (2014).

51 Yousefi P, Huen K, Quach H *et al.* Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environ. Mol. Mutagen.* 56(9), 751–758 (2015).

52 Koestler DC, Jones MJ, Usset J *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics* 17(1), 120 (2016).

53 Sun Y, Vestergaard M, Pedersen CB, Christensen J, Basso O, Olsen J. Gestational age, birth weight, intrauterine growth, and the risk of epilepsy. *Am. J. Epidemiol.* 167(3), 262–270 (2008).

54 Leek JT, Scharpf RB, Bravo HC *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11(10), 733–739 (2010).

55 Leek JT. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* 42(21), e161 (2014).

56 Singmann P, Shem-Tov D, Wahl S *et al.* Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin* 8, 43 (2015).

57 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).

58 Zampieri M, Ciccarone F, Calabrese R, Franceschi C, Burkle A, Caiafa P. Reconfiguration of DNA methylation in aging. *Mech. Ageing Dev.* 151, 60–70 (2015).

59 Du P, Zhang X, Huang CC *et al.* Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11, 587 (2010).

60 Relton CL, Groom A, St Pourcain B *et al.* DNA methylation patterns in cord blood DNA and body size in childhood. *PLoS ONE* 7(3), e31821 (2012).

61 Bland M. *An Introduction to Medical Statistics (3rd Edition)*. Oxford University Press, Oxford, UK (1986).

62 Barton SJ, Crozier SR, Lillycrop KA, Godfrey KM, Inskip HM. Correction of unexpected distributions of p values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics* 14, 161 (2013).

- **Demonstration that it is possible to account for noncompliance with parametric model assumptions by using alternative modeling strategies.**

63 Saadati M, Benner A. Statistical challenges of high-dimensional methylation data. *Stat. Med.* 33(30), 5347–5357 (2014).

64 Pleasants AB, Wake GC, Shorten PR *et al.* A new, improved and generalizable approach for the analysis of biological data generated by -omic platforms. *J. Dev. Orig. Health Dis.* 6(1), 17–26 (2015).

65 Siegel S, Castellan NJ. *Nonparametric Statistics For The Behavioural Sciences.* McGraw-Hill Humanities, NY, USA (1988).

66 Ritchie ME, Phipson B, Wu D *et al.* Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* 43(7), e47 (2015).

67 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 800–802 (1988).

68 Storey JD. A direct approach to false discovery rates. *J. R. Stat. Soc. B* 64, 479–498 (2002).

69 Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* 100(16), 9440–9445 (2003).

70 Stephens M. False discovery rates: a new deal (2016). http://biorxiv.org/content/early/2016/01/29/038216

71 Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics* 8(3), 389–399 (2015).

72 Teh AL, Pan H, Lin X *et al.* Comparison of methyl-capture sequencing vs. Infinium humanmethylation450 array for methylome analysis in clinical samples. *Epigenetics* 11(1), 36–48 (2016).

73 Ong ML, Holbrook JD. Novel region discovery method for Infinium 450k DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. *Aging Cell* 13(1), 142–155 (2014).

- **Case study showing how grouping CpGs into regions can boost power.**

74 Jaffe AE, Murakami P, Lee H *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41(1), 200–209 (2012).

75 Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated p-values. *Bioinformatics* 28(22), 2986–2988 (2012).

76 Robinson MD, Kahraman A, Law CW *et al.* Statistical methods for detecting differentially methylated loci and regions. *Front. Genet.* 5, 324 (2014).

77 Zhao N. Kernel machine methods for analysis of genomic data from different sources. Department of Biostatistics PhD, (2014).
https://cdr.lib.unc.edu

78 Sofer T, Schifano ED, Hoppin JA, Hou L, Baccarelli AA. A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* 29(22), 2884–2891 (2013).

79 Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

80 Wang D, Yan L, Hu Q *et al.* IMA: An R package for high-throughput analysis of Illumina's 450k infinium methylation data. *Bioinformatics* 28(5), 729–730 (2012).

81 Warden CD, Lee H, Tompkins JD *et al.* Cohcap: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res.* 41(11), e117 (2013).

82 Pongpanich M, Neely ML, Tzeng JY. On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. *Front. Genet.* 2, 110 (2011).

83 Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* 35(7), 606–619 (2011).

84 Zhao Y, Chen F, Zhai R, Lin X, Diao N, Christiani DC. Association test based on snp set: logistic kernel machine based test vs. principal component analysis. *PLoS ONE* 7(9), e44978 (2012).

85 Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with rnbeads. *Nat. Methods* 11(11), 1138–1140 (2014).

86 Schubeler D. Function and information content of DNA methylation. *Nature* 517(7534), 321–326 (2015).

87 Sun D, Yi SV. Impacts of chromatin states and long-range genomic segments on aging and DNA methylation. *PLoS ONE* 10(6), e0128517 (2015).

88 Lewsey MG, Hardcastle TJ, Melnyk CW *et al.* Mobile small rnas regulate genome-wide DNA methylation. *Proc. Natl Acad. Sci. USA* (113(6), e801–e810 (2016).

89 Plank JL, Dean A. Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell* 55(1), 5–14 (2014).

90 Huang B, Jiang C, Zhang R. Epigenetics: the language of the cell? *Epigenomics* 6(1), 73–88 (2014).

91 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).

92 Gevaert O. Methylmix: an R package for identifying DNA methylation-driven genes. *Bioinformatics* 31(11), 1839–1841 (2015).

93 Rauscher GH, Kresovich JK, Poulin M *et al.* Exploring DNA methylation changes in promoter, intragenic, and intergenic regions as early and late events in breast cancer formation. *BMC Cancer* 15, 816 (2015).

94 Li Y, Camarillo C, Xu J *et al.* Genome-wide methylome analyses reveal novel epigenetic regulation patterns in schizophrenia and bipolar disorder. *BioMed Res. Int.* 2015, 201587 (2015).

95 Liu Y, Devescovi V, Chen S, Nardini C. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst. Biol.* 7, 14 (2013).

96 Marzi SJ, Meaburn EL, Dempster EL *et al.* Tissue-specific patterns of allelically-skewed DNA methylation. *Epigenetics* 11(1), 24–23 (2016).

97 Vanderweele TJ, Asomaning K, Tchetgen Tchetgen EJ *et al.* Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *Am. J. Epidemiol.* 175(10), 1013–1020 (2012).

98 Relton CL, Davey Smith G. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics* 7(8), 1239–1243 (2015).

- **Discussion of the use of Mendelian randomization in EWAS and applications to causality.**

99 Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23(R1), R89–R98 (2014).

100 Kato N, Loh M, Takeuchi F *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* 47(11), 1282–1293 (2015).

101 Chadwick LH, Sawa A, Yang IV *et al.* New insights and updated guidelines for epigenome-wide association studies. *Neuroepigenetics* 1, 14–19 (2015).

102 Soh SE, Chong YS, Kwek K *et al.* Insights from the growing up in singapore towards healthy outcomes (gusto) cohort study. *Ann. Nutr. Metab.* 64(3–4), 218–225 (2014).

103 Soh SE, Tint MT, Gluckman P *et al.* Cohort profile: the gusto birth cohort study. *Int. J. Epidemiol.* 43(5), 1401–1409 (2012).

104 Ng JW, Barrett LM, Wong A, Kuh D, Smith GD, Relton CL. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol.* 13(6), 246 (2012).

105 Painter RC, Osmond C, Gluckman P, Hanson M, Phillips DI, Roseboom TJ. Transgenerational effects of prenatal exposure to the dutch famine on neonatal adiposity and health in later life. *BJOG* 115(10), 1243–1249 (2008).

106 Huang C, Li Z, Narayan KM, Williamson DF, Martorell R. Bigger babies born to women survivors of the 1959–1961 chinese famine: a puzzle due to survival selection? *J. Dev. Orig. Health Dis.* 1(6), 412–418 (2010).

107 Pembrey ME, Bygren LO, Kaati G *et al.* Sex-specific, male-line transgenerational responses in humans. *Eur. J. Hum. Genet.* 14(2), 159–166 (2006).

108 Marttila S, Kananen L, Jylhava J *et al.* Length of paternal lifespan is manifested in the DNA methylome of their nonagenarian progeny. *Oncotarget* 6(31), 30557–30567 (2015).

109 Yehuda R, Daskalakis NP, Bierer LM *et al.* Holocaust exposure induced intergenerational effects on fkbp5 methylation. *Biol. Psychiatry* doi:10.1016/j.biopsych.2015.08.005 (2015) (Epub ahead of print).

110 Kundaje A, Meuleman W *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539), 317–330 (2015).

•• **Description of rich and complex datasets available to EWAS researchers.**

111 Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414), 57–74 (2012).

112 Cancer Genome Atlas Research N, Weinstein JN, Collisson EA *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45(10), 1113–1120 (2013).