CrossMark

Journal of Developmental Origins of Health and Disease (2015), **6**(1), 10–16.
© Cambridge University Press and the International Society for Developmental Origins of Health and Disease 2014
doi:10.1017/S2040174414000506

ORIGINAL ARTICLE

Measuring epigenetics as the mediator of gene/environment interactions in DOHaD

M.-L. Ong, X. Lin and J. D. Holbrook*

*Singapore Institute for Clinical Sciences (SICS), A*STAR, Singapore, Singapore*

Analysis of DNA methylation data in epigenome-wide association studies provides many bioinformatics and statistical challenges. Not least of these, are the non-independence of individual DNA methylation marks from each other, from genotype and from technical sources of variation. In this review we discuss DNA methylation data from the Infinium450K array and processing methodologies to reduce technical variation. We describe recent approaches to harness the concordance of neighbouring DNA methylation values to improve power in association studies. We also describe how the non-independence of genotype and DNA methylation has been used to infer causality (in the case of Mendelian randomization approaches); suggest the mediating effect of DNA methylation in linking intergenic single nucleotide polymorphisms, identified in genome-wide association studies, to phenotype; and to uncover the widespread influence of gene and environment interactions on methylation levels.

Received 10 June 2014; Revised 12 September 2014; Accepted 20 September 2014; First published online 15 October 2014

Key words: DNA methylation, epigenomics, gene–environment interactions

Introduction

The causes of most diseases can be thought of as belonging in two broad categories: inherited (genetic) factors, and environmental exposures which can occur just before the disease or much earlier in development.

It has long been known that disease risks can be passed down through variation in DNA sequence within families and across generations. However, the extent of this heritability remains largely unresolved. The genetic basis of disease has been widely explored. Genome-wide association studies (GWAS) have discovered polymorphisms associated with certain diseases or risk factors, however, these account for only a small proportion of variance in the risk for common diseases such as major depression, Type II diabetes and obesity.

Unlike DNA marks, epigenetic marks encode information from both the inherited genotype[1–3] and environmental exposures,[4,5] and thus present a promising approach to explain multifactorial diseases. Epigenetic marks may be biomarkers for risk stratification and disease diagnosis. DNA methylation is one of the epigenetic changes, which has drawn much attention. In humans, it occurs mainly in the context of CpG dinucleotides. Advances in microarray technology and next-generation sequencing have made it possible to measure and quantify DNA methylation at a high resolution on a genome-wide scale and across multiple samples. These technologies open up exciting opportunities to perform epigenome-wide association studies (EWAS), however, they also pose huge bioinformatics challenges particularly in the areas of data processing,[6] statistical analyses/power[7,8] and integration with other genome-wide molecular datasets (i.e. genotype and transcriptome data). Issues of heterogeneous methylation across cell and tissue types and the unique statistical properties of DNA methylation measurements pose further challenges to the analysis.

This review focuses on computational and statistical methods in DNA methylation analyses and interpretation, and the challenges of gene–environment interaction analyses, and multiple genome-wide molecular data integration.

DNA methylation measurements

Several methods have been developed to profile DNA methylation on a genome-wide scale. Widely used methods are MeDIP-seq, MBD-seq, reduced representation bisulphite sequencing (RRBS) and the Illumina Infinium HumanMethylation27 and HumanMethylation450 arrays.[9,10] Protocols and commercial kits for all four methods are available. MeDIP-seq and MBD-seq employ an antibody or methyl-binding protein, respectively, to create a genomic library enriched for methylated genomic regions that are then sequenced. In both methods, cytosine coverage is high but the methods have poor resolution since they measure the relative enrichment of methylated DNA across regions rather than at individual residues, and hybridization to the antibody or methyl-binding protein is not necessarily linear to per cent methylation. Both RRBS and the Infinium arrays use bisulfite treatment, which converts cytosine residues to uracils, while 5-methylcytosines remain unaffected. RRBS uses methylation-sensitive restriction enzymes to create a genomic fragment library of methylated regions that are then bisulphite converted and sequenced. RRBS has single cytosine

*Address for correspondence: J. Holbrook, Singapore Institute for Clinical Sciences (SICS), Brenner Centre for Molecular Medicine, 30 Medical Drive, Singapore 117609, Singapore.
(Email Joanna_Holbrook@sics.a-star.edu.sg)

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

base resolution and offers much higher coverage than Infinium arrays, but tends to bias towards regions that are moderately or highly methylated and does not cover all the hypomethylated regions that are thought to be important in inter-individual variation. Interestingly, a recent analysis on genome-wide, inter-individual variation in DNA methylation in humans suggest that despite the greater coverage of RRBS, RRBS and the Infinium HumanMethylation450 array capture a comparable percentage of variably methylated CpGs.[11] The different technologies have been applied on cells and tissues and benchmarked against one another.[9,12]

In this review, we focus mainly on the Infinium HumanMethylation450 array platform (Infinium450K), which is capable of measuring the methylation status of more than 450,000 cytosines in humans. Previous reports have shown the accuracy of Infinium450K data when compared with data generated using the HumanMethylation27 array, GoldenGate and whole genome bisulphite sequencing.[13] We have also applied RRBS and Infinium450K to clinical samples and showed a high concordance between the two.[14] Although Infinium450K arrays offer far lesser coverage as compared with sequencing-based methods, their cost effectiveness, throughput and resolution have made them an increasingly popular choice for EWAS. To date, there have been more than 6000 Infinium450K data sets deposited in public repositories (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534), and the number is growing.

A crucial aspect of an EWAS is the study design which is driven by the scientific question of interest and we refer the reader to reviews by Rakyan *et al.*,[15] Michels *et al.*[16] and Mill *et al.*[17] on various aspects of study design such as cohort and sample size considerations and tissue type selection. We note that given the large number of statistical tests to be conducted, it is important for an EWAS to be well powered. Similar to a well-designed GWAS, significant EWAS findings should be validated in an independent cohort. Another way to add confidence to associative observations is functional validation.[18] For example, functional evaluation of candidate epigenetic marks can be conducted in animal models, where their epigenetic states can be perturbed and the changes relevant to the disease can be observed. These functional studies can also provide evidence for the causative link between the phenotype and the phenotype-associated epigenetic change. Measurement of the epigenetic mark in animal species could be conducted using *de novo* sequencing techniques such as methylation-sensitive pyrosequencing or MeDIP-seq (easier where a reference genome exists). However even the Infinium450K array designed for human has been validated for use with modification and filtering on some great ape species,[19] *Cynomolgus macaques*[20] and mice.[21]

Infinium450K data processing and quality control

The Infinium450K array employs two different assay designs that is Type I and Type II. The Type I assay uses two probes per CpG locus, corresponding to the methylated and unmethylated alleles, and both signals are measured in the same colour channel. On the other hand, the Type II assay uses only one probe, which detects both alleles and the methylated and unmethylated signals are generated in the green and red channels, respectively. In both cases, the methylation value (β value) for a CpG is computed as the ratio between the methylated signal and the sum of the methylated and unmethylated signals.[13] The logit transformed β values are referred to as M values.[22] Several groups have reported technical differences between Type I and Type II probes that is the β values from Type II probes exhibit a narrower range as compared with the Type I probes.[10]

Besides sample and probe quality control procedures,[23] the processing of methylation data involves additional steps to correct for (1) intensity differences between the red and green channels using measurements from control probes on the array, (2) inter-sample variability and technical differences between Type I and Type II probes and, (3) batch effects.

Various methods have been proposed to correct for (2) inter-sample variability and technical differences between Type I and Type II probes.[14,23–25] Wu *et al.*[26] compared the relative performance of using raw un-normalized methylation values and normalized methylation values from four different normalization methods, (i) β mixture quintile normalization (BMIQ),[24] (ii) subset-quantile within array normalization (SWAN),[25] (iii) complete pipeline[23] and (iv) Illumina's method as implemented in GenomeStudio software. In terms of reproducibility of methylation values across technical replicates, they found that BMIQ and SWAN had better overall performance, but the improvement was only modest compared with using raw un-normalized data. Benchmarking the performance of Infinium450K data against RRBS on the same samples, Pan *et al.*[14] showed that the greatest improvement in agreement is achieved through Type I and II correction, followed by quantile-normalization and colour adjustment. They also showed that their improved version of GenomeStudio's normalization algorithm generally performed better than the original and to SWAN[24] on these samples.

The normalization procedures for correcting for inter-sample variability and technical differences between Type I and Type II probes can also correct for minor batch effects.[14,27] For major batch effects, batch effect correction can be applied after normalization.[28,29] We note also that the careful study design can minimize batch effects.[26] Careful processing on Infinium450K data is necessary because the methylation effect sizes discovered in the DOHaD field so far have been rather small (1–2% difference) so sensitivity in the measure of methylation and removal of technical bias is essential.

Confounding effects of cell heterogeneity

When an EWAS is conducted using heterogeneous tissue types such as blood or umbilical cord, accounting for cellular heterogeneity in the analysis is important.[30] Different cell lineages

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

have different methylation profiles,[31,32] and without accounting for cellular heterogeneity, an observed association between a phenotype and methylation could be due to differences in the cell type distributions across samples. Statistical algorithms for inferring cell mixture proportions belong to two main categories that is a reference-based[32] and reference-free[33,34] approach. The critical step of the reference-based approach is identifying the set of methylation signatures of the major cell types in the tissue. These cell-type-specific CpGs can then serve as a high dimensional multivariate surrogate for the distribution of cell type proportions in the heterogeneous tissue. This method has been used to estimate blood cell type proportions in a study investigating the differential methylation of rheumatoid arthritis *v.* controls in whole blood.[35] The authors included this inferred blood cell type proportions as a covariate in a linear regression model to correct for the effects of cell type heterogeneity. More recently, two different reference-free methods have been proposed for complex tissues where the reference methylation signatures of constituent cell types are not easily obtainable, or the relevant cell types are yet unknown. The FaST-LMM-EWASher approach[34] uses a combination of linear mixed model and principle components to correct for cell-type heterogeneity, where the pairwise similarity between individuals is used as a proxy for cell-type composition. The Houseman method[33] uses a latent surrogate variable approach to adjust for cell type effects. These statistical methods facilitate the discovery of genuine associations of interest by removing potential spurious associations due to cell-type heterogeneity. This is especially important in investigating the origins of diseases which may be associated with cell type differences for instance inflammatory changes in obesity.

Association analysis

Differentially methylated CpGs (DMCs)

To identify individual CpGs that are associated with a phenotype, a simple approach is to test each CpG individually for association with the phenotype. Published studies have conducted the analysis either using the methylation levels as the predictor variable and the phenotype as the outcome variable[36,37] or using the methylation levels as the outcome variable and the phenotype as the predictor variable.[38] For the former, when the phenotype is continuous or binary, linear regression or logistic regression can be used for the analysis, respectively. For longitudinal phenotypes, for example, when a cohort is followed over time or correlated phenotypes and when related individuals are recruited, mixed models or generalized estimating equations can be used for estimating the respective linear or logistic regression models. We note that when the longitudinal outcome is binary, the regression coefficients estimated from mixed models and generalized estimating equations have different interpretation and the choice of analysis method depends on the scientific question of interest. In the latter when the methylation levels are used as the outcome, since methylation is continuous, linear regression can be employed. As before, if methylation levels are measured longitudinally or among related individuals, mixed models or generalized estimating equations can be used.

After all CpGs on the Infinium450K array have been tested individually for association with the phenotype, the epigenome-wide *P*-values can be displayed graphically using either a quantile-quantile plot or manhatten plot. To assess statistical significance of the epigenome-wide *P*-values, the individual *P*-values are typically corrected for multiple testing, usually using a Bonferroni or Benjamini-Hochberg[39] correction. However, this approach is very conservative as it disregards the correlation between the test variables. Examples of EWAS associations which have passed multiple testing corrections include the association of *HIF3A* methylation in blood with obesity[37] and *AHRR* and *F2RL3* methylation in blood as a consequence of smoking or exposure to smoke *in utero*.[40–44]

Differentially methylated regions (DMRs)

DNA methylation at individual CpGs have been shown to be highly correlated over short chromosomal distances using high-density measures of the methylome. This characteristic of co-methylation allows neighbouring CpGs to be grouped into regions, thereby increasing statistical power due to the smaller number of tests and reducing false positives due to singular noisy signals. A method leveraging on this property has been developed for high coverage array CHARM and sequencing-based platforms, and termed 'bump hunting'.[45] Briefly, the method involves first regressing the methylation measures arranged by chromosomal position over the outcome of interest, then smoothing the regression slopes to reduce noise, and finally selecting the candidate DMRs whose spatially contiguous CpG signals lie consistently above the pre-set signal threshold. A 'bump hunting' method for Infinium450K, which takes into account the sparsity and spatial irregularity of the probes on the array, has also been developed.[46] Using Infinium450K data from blood of individuals of different ages, Ong and Holbrook showed that their region analysis achieved greater specificity compared with a DMC approach, as their method increased the extent of common findings between independent aging studies. And they showed the power increase from 39% from a DMC approach to 61% with region detection, as the method reduces the number of tests from 450 to 55K. Alternative DMR approaches such as sliding window analysis coerce the data into arbitrary fixed sizes, and carries out differential analyses on these pre-fixed regions. The 'bump hunting' method of Jaffe *et al.*[45] and the 'region discovery' approach of Ong and Holbrook,[46] eliminates having to impose an arbitrary constant size for each region.

VMRs

The 'bump hunting' and 'region discovery' approaches have also been adapted to search for variably methylated regions (VMRs) across individuals.[46,47] In region discovery,[47] the signal is determined solely by the median absolute deviation in the

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

methylation values across samples, regardless of the outcome of interest. This statistic gives us a measure of the degree of inter-individual variation in methylation. Prioritizing the analysis to VMRs allows one to significantly reduce the number of tests, which reduces the multiple testing problems inherent in genome-wide studies. Interestingly, the number of VMRs detected in blood increase dramatically as a function of human age.[46] The VMR genes of the older age population tend to cluster into neuorosignalling pathways. Some of the neurosignalling genes containing VMRs (e.g. *POMC* and *OXTR*) have previously been shown to be methylated in response to environment.

Integration of multiple genome-wide molecular data sets

The overarching goal of most developmental 'omics' studies is to uncover the biological mechanisms that underlie developmental outcomes. To that end, when different types of molecular datasets, for example genomic, epigenomic, gene expression, and metabolomics data are available, an integrated analysis can be conducted. For example, gene expression quantitative trait loci (eQTL) studies seek to identify single nucleotide polymorphisms (SNPs) that are associated with gene expression,[48] while methylation quantitative trait loci (methQTL) studies seek to identify SNPs that are associated with DNA methylation.[49] eQTL/methQTL SNPs, gene expression and methylation can then be modelled jointly for their effects on phenotype.[50,51] A similar approach can be undertaken with metabolites.[52,53]

With multiple data sets, one can also explore the causal relationships between the different layers of biological control. For example, Gutierrez-Arcelus *et al.*[54] explored the causal relationship between genotype, DNA methylation and gene expression by combining data from RNA-seq, SNP genotyping and the Infinium450K array performed on umbilical cords of newborn infants. Using a Bayesian network and relative likelihood method, they found that a SNP is most likely to independently affect expression and DNA methylation, with SNP driving expression, which in turn affects methylation being the least likely model. The Bayesian network models have also been applied to reconstruct causal pathways.[55] Other causal inference methods such as the likelihood-based causality model selection test, which computes conditional correlation measures have also been used successfully to infer causal associations between gene expression and disease.[56]

Gene-DNA methylation–environment interplay

The effects of genotype on DNA methylation have been studied extensively and a large number of methQTLs have been found in human tissues.[1–3] Recently Liu *et al.*[57] found that methQTLs can incorporate contiguous and non-contiguous CpG clusters (which they term GeMes).

Some genotype associations with phenotype may be mediated by the influence of a genotype on the epigenome and its subsequent impact on phenotype. This is a promising paradigm for investigating intergenic SNPs associated with phenotype in a GWAS, but with no obvious direct route to perturb the transcriptome.[57] Liu *et al.*[35] identified a mediating role of HLA DNA methylation in the aetiology of rheumatoid arthritis. Using a causal inference test (CIT),[58] the group found CpG loci, which mediate the effect of previously reported associative genotype on rheumatoid arthritis risk. Briefly, the CIT requires four conditions to be satisfied (1) SNP is associated with disease (2) SNP is associated with methylation after adjusting for disease (3) Methylation is associated with disease after adjusting for SNP (4) SNP is independent of disease after adjusting for methylation.

The association of genotype with methylation allows for causal inference approach, Mendelian randomisation to be applied to EWAS hits.[59] Dick *et al.*,[37] reported a significant association between blood DNA methylation within the *HIF3A* gene and body mass index (BMI), they also showed *HIF3A* methylation was in a methQTL with SNPs in the *HIF3A* gene. As the *HIF3A* SNPs were not associated with adult BMI, they proposed that DNA methylation changes are driven by BMI under the assumption of Mendelian randomization. However, an alternative model is that both BMI and DNA methylation share a yet undiscovered causal factor (excluding genetic variant). Possible modifiers of both BMI and DNA methylation could include environmental factors such as diet and exercise. These type of confounding factors are often present and violate the assumptions of Mendelian randomization. Other factors that represent violations are the presence of linkage disequilibrium, genetic heterogeneity, pleiotrophy, population stratification and canalization.[60] Again, these factors are nearly always present in biological data sets. Additional strong assumptions are linearity of all relationships and no interactions. However, interactions of gene and environment are very common indeed in biology and in DoHAD.

Gene–environment interplay in complex diseases that is gene environment interaction ($G \times E$ model) can be conceptualized as the genotypic predisposition to one's degree of sensitivity to environmental influences. A striking example is the interaction of *FKBP5* genotype and early childhood trauma to affect methylation of *FKBP5* intron 7, *FKBP5* expression and subsequent deregulation of glucocorticoid receptor signalling.[61] Yousefi *et al.*[62] found that interactions between maternal smoking and leptin receptor (*LEPR*) SNPs affected *LEPR* methylation levels, and these *LEPR* SNPs, in interaction with *LEPR* methylation, associated with leptin levels at 18 years. Teh *et al.*,[63] used a genome-wide survey of DNA methylation with DNA obtained from umbilical cords in relation to a wide range of measures of antenatal maternal health and well-being, including maternal mood. They identified 1423 VMRs[46] and used statistical modelling to examine whether the variability in DNA methylation at individual VMRs was best explained by sequence-based genetic variation, antenatal maternal environmental influences or the interaction between the two factors. The results revealed that variation in DNA methylation was best explained by genetic factors in ~25% of the VMRs. Commonly, these effects involved genetic variants in close proximity to VMR. In contrast, ~75% of the VMRs were best

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

explained by a G × E interaction model. Interestingly, in no cases were VMRs best explained by environmental conditions alone, acting independent of the genome.

Other than interaction analyses, a less stringent segregated analysis may also allow one to determine CpG loci where the association between DNA methylation and environment is dependent on genotype. For instance, genotype at a polymorphism in the brain-derived neurotrophic factor (*BDNF*) gene strongly affects whether multiple CpGs in the neonate methylome co-vary with pre-natal maternal anxiety. The methylomes of neonates with differential *BDNF* genotypes reflect inter-individual variation in the neonatal volumes of different brain regions.[64] These results underscore the importance of integrating genotype data in EWAS. In addition to G × E models, one can also test for E × E interactions effects. In the study of maternal tobacco use on the infant's DNA methylation, Suter *et al.*[65] found DMCs which showed significant association with smoking status in interaction with infant birth weight. It is also biologically plausible that DNA methylation interacts with genotype to influence disease outcomes. For example, Soto-Ramírez *et al.*[66] found that the interaction of *IL4R* genetic variants and *IL4R* DNA methylation increases the risk of asthma at age 18 years.

Future directions

As sequencing-based methods become more cost-effective, the shift from array-based platforms to next generation sequencing (NGS) methods will likely accelerate. NGS methods offer a much more comprehensive picture of the human methylome (the Infinium450K array covers only 2% of all human CpGs), but it will increase the computational load of analysis. Another advantage of NGS over array methods is that batch effect problems are much reduced assuming proper experimental design, and this is especially pertinent for large-scale EWAS studies. However, the tremendous increase in the amount of methylation data generated per sample would greatly exacerbate the multiple testing problem. It will therefore be crucial to employ data reduction methods such as region analysis[46] to boost statistical power. Analogous to linkage disequilibrium (LD) in genetics, high-density methylation measurements can allow correlated CpGs to be reliably clustered into methylation blocks.[45,46,57] These methylation structures can then be comprehensively interrogated for their associations with LD blocks and co-expression modules, without compromising power significantly.

There is a massive increase in the number of methylomes being generated by individual laboratories and public consortia such as the International Human Epigenome Consortium and the Encyclopedia of DNA Elements initiative. With the huge forthcoming methylome data publicly available, reference maps of human variably methylated CpG blocks (VMRs) across different cell types and conditions can be established. This would be a valuable resource for prioritizing the discovery efforts to these specific regions. Identification of methQTLs would add important information about the genetic influence on these methylation maps. However, the growing evidence of the prevalence of the combined effects of genetics and the environment on methylation[61–63] suggests that the binary classification of genetically driven CpGs (methQTLs) and non-genetically driven CpGs (non-methQTLs) is too simplistic. It is thus likely that consideration of methQTLs will evolve from a qualitative discrete analysis to a quantitative continuum analysis. Development of new statistical methodologies to test combined and interacting effects of genotype and environment on DNA methylation, and environment, methylation and genotype on phenotype, will be necessary.

As genome-wide molecular datasets consisting of genome sequence, DNA methylome, metabolome, proteome, microbiome, transcriptome, etc. become more readily available, the next challenge would be to develop methods that efficiently integrate these data as a whole and perform a joint analysis, which maximizes the use of data and allows one to explore the interplay of these layers of regulation. Furthermore, the cross talk between layers of regulation can change under different conditions and over time, and can pose major challenges to understanding the causes and consequences of these molecular changes. The development of powerful approaches for integrative data analysis, taking into account the specific noise, biases and statistical properties of individual data sets, across cell types, time and conditions is likely to be a primary focus of the field of computational biology going forward.

Acknowledgements

MLO, XL and JDH are supported by the Singapore Institute for Clinical Sciences, Agency for Science Technology and Research (A*STAR), Singapore.

Financial Support

This work was supported the Singapore Institute for Clinical Sciences (SICS) – Agency for Science, Technology and Research (A*STAR), Singapore.

Conflicts of Interest

None.

References

1. Gibbs JR, van der Brug MP, Hernandez DG, *et al*. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*. 2010; 6, e1000952.
2. Zhang D, Cheng L, Badner JA, *et al*. Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet*. 2010; 86, 411–419.
3. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. *Genome Biol*. 2012; 13, R8.
4. Lam LL, Emberly E, Fraser HB, *et al*. Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci USA*. 2012; 109 (Suppl 2), 17253–17260.

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

5. McGowan PO, Sasaki A, D'Alessio AC, *et al*. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat Neurosci*. 2009; 12, 342–348.
6. Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genetics*. 2012; 13, 705–719.
7. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011; 12, 529–541.
8. Heijmans BT, Mill J. Commentary: the seven plagues of epigenetic epidemiology. *Int J Epidemiol*. 2012; 41, 74–78.
9. Harris RA, Wang T, Coarfa C, *et al*. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*. 2010; 28, 1097–1105.
10. Dedeurwaerder S, Defrance M, Calonne E, *et al*. Evaluation of the infinium methylation 450K technology. *Epigenomics*. 2011; 3, 771–784.
11. Ziller MJ, Gu H, Muller F, *et al*. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013; 500, 477–481.
12. Bock C, Tomazou EM, Brinkman AB, *et al*. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol*. 2010; 28, 1106–1114.
13. Bibikova M, Barnes B, Tsan C, *et al*. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011; 98, 288–295.
14. Pan H, Chen L, Dogra S, *et al*. Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics*. 2012; 7, 1173–1187.
15. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011; 12, 529–541.
16. Michels KB, Binder AM, Dedeurwaerder S, *et al*. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods*. 2013; 10, 949–955.
17. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet*. 2013; 14, 585–594.
18. Meaney MJ, Ferguson-Smith AC. Epigenetic regulation of the neural transcriptome: the meaning of the marks. *Nat Neurosci*. 2010; 13, 1313–1318.
19. Hernando-Herraez I, Prado-Martinez J, Garg P, *et al*. Dynamics of DNA methylation in recent human and great ape evolution. *PLoS Genet*. 2013; 9, e1003763.
20. Ong ML, Tan PY, MacIsaac JL, *et al*. Infinium monkeys: Infinium 450K array for the *Cynomolgus macaque* (*Macaca fascicularis*). *G3*. 2014; 4, 1227–1234.
21 Wong NC, Ng J, Hall NE, *et al*. Exploring the utility of human DNA methylation arrays for profiling mouse genomic DNA. *Genomics*. 2013; 102, 38–46.
22. 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*. 2010; 11, 587.
23. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*. 2012; 4, 325–341.
24. Teschendorff AE, Marabita F, Lechner M, *et al*. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013; 29, 189–196.
25. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol*. 2012; 13, R44.
26. Wu MC, Joubert BR, Kuan PF, *et al*. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics*. 2013; 9, 318–329.
27. Sun Z, Chai HS, Wu Y, *et al*. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med Genom*. 2011; 4, 84.
28. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010; 11, 733–739.
29. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007; 8, 118–127.
30. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014; 15, R31.
31. 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*. 2012; 7, e41361.
32. Houseman EA, Accomando WP, Koestler DC, *et al*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012; 13, 86.
33. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *BMC Bioinformatics*. 2014; 30, 1431.
34. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods*. 2014.
35. 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*. 2013; 31, 142–147.
36. Joubert BR, Håberg SE, Nilsen RM, *et al*. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect*. 2012; 120, 1425.
37. Dick KJ, Nelson CP, Tsaprouni L, *et al*. DNA methylation and body-mass index: a genome-wide analysis. *Lancet*. 2014; 383, 1990–1998.
38. Engel SM, Joubert BR, Wu MC, *et al*. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian mother and child cohort. *Am J Epidemiol*. 2014; 179, 834–842.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*. 1995; 57, 289–300.
40. Wan ES, Qiu W, Baccarelli A, *et al*. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Genet*. 2012; 21, 3073–3082.
41. Zhang Y, Yang R, Burwinkel B, *et al*. F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int J Epidemiol*. 2014; 43, 1215–1225.
42. Sun YV, Smith AK, Conneely KN, *et al*. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. *Hum Genet*. 2013; 132, 1027–1037.

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press

43. Monick MM, Beach SR, Plume J, *et al*. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. *Am J Med Genet B Neuropsychiatr Genet*. 2012; 159B, 141–151.
44. Shenker NS, Ueland PM, Polidoro S, *et al*. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*. 2013; 24, 712–716.
45. Jaffe AE, Murakami P, Lee H, *et al*. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol*. 2012; 41, 200–209.
46. 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*. 2014; 13, 142–155.
47. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*. 2012; 13, 166–178.
48. Morley M, Molony CM, Weber TM, *et al*. Genetic analysis of genome-wide variation in human gene expression. *Nature*. 2004; 430, 743–747.
49. Gamazon E, Badner J, Cheng L, *et al*. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry*. 2013; 18, 340–346.
50. Huang Y-T, VanderWeele TJ, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann Appl Statist*. 2014; 8, 352.
51. Huang Y-T. Integrative modeling of multiple genomic data from different types of genetic association studies. *Biostatistics*. 2014; 15, 587–602.
52. Petersen A-K, Zeilinger S, Kastenmüller G, *et al*. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum Mol Genet*. 2014; 23, 534–545.
53. Gieger C, Geistlinger L, Altmaier E, *et al*. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*. 2008; 4, e1000282.
54. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, *et al*. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*. 2013; 2, e00523.
55. Sachs K, Perez O, Pe'er D, *et al*. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. 2005; 308, 523.
56. Schadt EE, Lamb J, Yang X, *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005; 37, 710.
57. Liu Y, Li X, Aryee MJ, *et al*. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet*. 2014; 94, 485–495.
58. Millstein J, Zhang B, Zhu J, *et al*. Disentangling molecular relationships with a causal inference test. *BMC Genet*. 2009; 10.
59. Relton CL, Smith GD. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol*. 2012; 41, 161–176.
60. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med*. 2008; 5, e177.
61. Klengel T, Mehta D, Anacker C, *et al*. Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nat Neurosci*. 2013; 16, 33–41.
62. Yousefi M, Karmaus W, Zhang H, *et al*. The methylation of the LEPR/LEPROT genotype at the promoter and body regions influence concentrations of leptin in girls and BMI at age 18 years if their mother smoked during pregnancy. *Int J Mol Epidemiol Genet*. 2013; 4, 86.
63. Teh AL, Pan H, Chen L, *et al*. The effect of genotype and in utero environment on inter-individual variation in neonate DNA methylomes. *Genome Res*. 2014; 24, 1064–1074.
64. Chen L, Pan H, Tuan TA, *et al*. Infant BDNF Val66Met influences the association of the DNA methylome with maternal anxiety and neonatal brain volumes. *Development and Psycopathology*. 2014 (in press).
65. Suter M, Ma J, Harris A, *et al*. Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics*. 2011; 6, 1284.
66. Soto-Ramírez N, Arshad SH, Holloway JW, *et al*. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics*. 2013; 5, 1.

https://doi.org/10.1017/S2040174414000506 Published online by Cambridge University Press