Human Molecular Genetics, 2014, Vol. 23, Review Issue 1 **R83–R88**
doi:10.1093/hmg/ddu284
Advance Access published on June 13, 2014

Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies

Liming Liang[1,2] and William O.C. Cookson[3,*]

[1]Department of Epidemiology and [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA and [3]National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK

Received May 29, 2014; Revised and Accepted June 5, 2014

Platform technologies for measurement of CpG methylation at multiple loci across the genome have made ambitious epigenome-wide association studies affordable and practicable. In contrast to genetic studies, which estimate the effects of structural changes in DNA, and transcriptomic studies, which measure genomic outputs, epigenetic studies can access states of regulation of genome function in particular cells and in response to specific stimuli. Although many factors complicate the interpretation of epigenetic variation in human disease, cell-specific methylation patterns and the cellular heterogeneity present in peripheral blood and tissue biopsies are anticipated to cause the most problems. In this review, we suggest that the difficulties may be exaggerated and we explore how cellular heterogeneity may be embraced with appropriate study designs and analytical tools. We further suggest that systematic mapping of the loci influenced by age, sex and genetic polymorphisms will bring important biological insights as well as improved control of epigenome-wide association studies.

INTRODUCTION

Epigenetics broadly defines the study of changes that affect DNA function without altering the nucleotide sequence. Epigenetics is now a hot area in medical genomics. Large-scale projects such as the International Human Epigenome Consortium (http://ihec-epigenomes.org/) are generating comprehensive measurements of the variation in DNA methylation, histone modification, nucleosome occupancy and coding and non-coding RNA expression from normal and diseased cells. These data intersect with the ENCyclopedia Of DNA Elements (ENCODE) project that is building a comprehensive list of the functional elements of the human genome by studying >4000 different datasets covering a wide range of cells, conditions and technologies (http://genome.ucsc.edu/ENCODE/index.html).

These resources offer wonderful opportunities for understanding and mapping genome function in health and disease. Although there are many epigenetic markers that may be of use in systematic studies, the degree of CpG methylation at multiple sites (loci) across the genome is at the moment most easily applicable to understanding disease. Hypermethylation in regulatory regions such as CpG islands, which are often associated with transcription start sites, are generally associated with genes that are not expressed, although other CpG sites in untranscribed genes may exhibit hypomethylation (1).

It is not yet clear to what extent methylation in regulatory regions of genes is directly modifying gene expression or function and how much is the consequence of the binding of classical transcription and enhancer factors. Nevertheless, it is generally true that genome-wide methylation assays capture robust biological information about the functional state of cells and tissues. It is reasonable to expect that this information can be used to detect novel genes and pathways underlying disease and to develop systematic understanding of essential processes such as inflammation and repair.

The effective use of genome-wide data of this extent and complexity is however demanding. Known but relatively uncharted effects on epigenetic marks at multiple sites include age, sex, the environment and DNA variants such as Single Nucleotide Polymorphisms (SNPs). Investigators will also have to deal with batch effects that may arise from the platforms used to quantify epigenetic variation.

A current perception is that the major problem with epigenome-wide studies arises because epigenetic changes determine (or reflect) cellular differentiation into specific lineages (2–4), requiring purified cells to be studied (5) or that the cell-specific

*To whom correspondence should be addressed at: Imperial College London, Guy Scadding Building, Royal Brompton Campus, Dovehouse Street, London SW3 6LY, UK. Tel: +44 2075942943; Email: w.cookson@imperial.ac.uk

patterns be evaluated before attempting association studies (6). If true, given that most tissues are complex mixtures of different cells in varying stages of activity, then progress will be slow. Happily, there are ways to deal with cellular heterogeneity that do not involve exhaustive isolation of every cell type.

The main purpose of this review is to explore practical and analytical approaches to CpG methylation that may be used to control for cellular mixtures, whilst taking into account other major potential confounders of epigenome-wide surveys.

EXPERIMENTAL DESIGN: EXERCISING THE GENOME

A full discussion of experimental design is beyond the remit of this review, but the concept of expression space of an individual genome is worth considering in strategies to extract the maximum information from any study (7). The expression space (or the epigenomic space) of a genome represents the spectrum of various functions that the genome may carry out. Exercising the genome captures states that are relevant to the question under study. Experiments that exercise the epigenome may examine the genome at different points in time after a stimulus, or at different concentrations of a stimulus (and considering whether the stimulus is physiological or pharmacological), or testing in different subjects with different genetic and epigenetic backgrounds and examining genome function in different cells and tissues (7). To this list can be added environmental and disease-related stimuli that may strongly influence the epigenome.

Most epigenome-wide association studies will be based on a cross-sectional observational design that is similar to that applied to genetic association studies. However, even this simple structure may draw on the epigenomic space in unexpected ways.

Whilst DNA sequence changes are stable over the lifetime of an individual, epigenetic changes may represent dynamic responses to known and unknown confounding factors, such as age, sex, the environment and disease that may need to be considered in study design.

If these factors confound the outcome or exposure of interest in a cross-sectional study, they can inflate false-positive rates as well as reduce statistical power to capture real associations. In common with gene expression studies and expression quantitative loci (eQTL) mapping, it may be anticipated that statistical power will be increased by modelling unknown as well as known factors in association studies (8). We consider these below.

BATCH AND PLATFORM EFFECTS

The technology for genome-wide measures of CpG methylation is steadily evolving. Illumina provides a popular platform for the robust measurement of methylation status at 450 000 CpG residues across the genome, with matching probes comparing sequences for native DNA compared with sequences after bisulphite conversion. This technology is likely to be replaced by whole-genome bisulphite sequencing, with or without a means for enriching for loci showing significant variation between cells, tissues and individuals (9) [it is worth noting that other cytosine modifications such as 5-hydroxymethylcytosine and 5-formylcytosine are found in genomic DNA of a wide range of mammalian cells and may carry additional information (10)]. The outcome of these measurements is a quantitative parameter that captures the degree of methylation at a particular site [in the case of Illumina platform, the parameter is β, measured on a scale of 0 (completely unmethylated) to 1 (completely methylated)].

As with any microarray platform that measures quantitative traits, batch effects, such as those associated with chips, plates, runs, time and other experimental and biological conditions, are expected to cause a large variation in parameter measurement. Adjusting effectively for these can increase power and reduce false positives. Adjustment can be accomplished by treating batch ID as a factor variable included as a fixed effect in the regression model to test association. When the sample size is small, an empirical Bayes method can be used to estimate batch effects jointly using all probes on the array (11). An alternative is to model the batch ID as a random intercept in the regression model, saving degrees of freedom and potentially increasing power.

Confounding factors such as environmental exposures and technical variations may commonly be represented in the array data without being known or observable. High dimension methods such as principal component analyses (PCA) and multidimensional scaling (MDS) can be used effectively to estimate and adjust for such factors when they are sufficiently strong (12,13). Similar adjustments have provided marked improvements in power to detect eQTL associations in genome-wide surveys (8,14).

AGE AND SEX

The age of subjects is known to influence methylation at loci across the genome. Although monozygotic twins may be indistinguishable by methylation status at birth, consistent age-related changes (both negative and positive) may be detected as children grow older (15). Gene ontology analyses indicate that these loci affect genes for developmental processes and immune functions and that there is overlap with loci that change with age in adult studies (15). In healthy adults (postmenopausal women), most age-related changes involve locus hypermethylation and are not associated with the disabilities that variably accompany ageing (16).

In contrast, other studies of adults suggest enrichment for structural motifs such as bivalent chromatin domains in age-associated DNA hypermethylation (17) and that these changes are not immune or haemopoietic cell-type specific (17). It has also been suggested that polycomb group proteins, which are supressed in stem cells, may show increased methylation in post-menopausal women, predisposing to malignancy (18). Other authors have opined that genome-wide methylation profiles may be usefully employed to quantify human rates of aging (19).

These papers reflect different schools of interpretation of genome-wide methylation signals, the best of which aims for reproducibility, meta-analyses, and mapped and confirmed loci that eventually can be functionally interrogated. Importantly, at least some of the age-related changes in methylation at specific loci may be attributed with varying composition of different

white cells in the peripheral blood leukocytes (PBLs) that are the source of DNA for many genetic and epigenetic association studies (20). These difficulties arising from cellular heterogeneity are discussed in detail below.

The effects of sex on methylation status are well documented (21,22), beginning with the recognition that X-chromosome inactivation in females is accompanied by widespread CpG hypermethylation (23). Interestingly, the pattern of genes showing methylation on the X chromosome is tightly regulated and variable between individuals (24) may even be tissue specific (25). Sex-related changes in methylation are also recognized at autosomal loci (21,22), but these effects have not yet been systematically mapped or studied. A further complexity arises from the recognition that environmental effects may produce different methylation consequences in males and females, for example genes influencing glucose metabolism, obesity and diabetes amongst neonates in rural Gambia (26).

GENETIC POLYMORPHISM

Genetic polymorphisms such as SNPs are recognized to affect methylation at individual loci (27) and genome wide (28,29). These effects are strongest in *cis* but are also present in *trans*. The mechanisms for SNP effects are not well understood but are likely to relate to allelic differences in the binding or expression of regulatory factors, including non-coding RNAs.

True SNP effects need to be differentiated from genetic variants that underlie the methylation probe sequence or the CpG interrogation site. These polymorphisms can cause confounding between methylation and the outcome of interest, as the perceived association with methylation is in fact the association with genotypes of the genetic variant. Such methylation probes are easily removed during quality control steps of analysis.

Genetic variants may confound associations even when the probe sequence or the CpG site does not overlap with any polymorphism. A genetic variant can be a regulator of the disease outcome at the same time as influencing methylation at a CpG site. In this case, association between outcome and methylation reflects sharing of genetic controls rather than causality. One can use mediation analysis (30–33) to estimate how much SNP–outcome association is through methylation as an intermediate phenotype and Mendelian randomization (34) to assess the causal direction between SNP, methylation and outcome.

CELL-SPECIFIC EFFECTS

Lineage commitment to particular cell types can be identified by methylation status at specific loci (2–4). There has been considerable concern that studying heterogeneous mixtures of cells from tissues and peripheral blood will cause serious confounding of epigenome-wide association studies (5,6) and that as a consequence purified cells should be studied (5) or that the cell-specific pattern for the gene region of interest should be evaluated before attempting association studies (6).

If this assumption is correct, then larger-scale epigenome studies will prove very technically difficult to perform. It is worth observing that cell specificity applies equally to studies of global gene expression, and, despite the old canard that gene expression in peripheral blood measured by microarrays was really an expensive cell count, there have been enormous advances in mapping eQTLs from lymphoblastoid cell lines and tissues with heterogeneous composition (8,35,36). Such studies have been particularly helpful in understanding systematically the function of the numerous disease-associated loci discovered by genome-wide association studies (GWAS) (35).

Before exploring means of resolving or controlling for cellular heterogeneity, there are two further complicating factors that are not usually considered. First, not all PBL constituents will be treated equally by the process of DNA extraction: buffy coats are retrieved after centrifugation that separates some granulocytes (such as eosinophils and basophils) from lymphocytes, monocytes and neutrophils, and red cell lysis methods may have differential effects on monocytes compared with other PBLs. It is therefore essential that cases and controls in a study have DNA extracted by matching protocols and that the results from different subjects and panels are amalgamated by meta-analyses rather than simple pooling.

Secondly, and equally of relevance, most PBLs exist in activated and unactivated states, including neutrophils (37), eosinophils (38,39), monocyte–macrophages (40) and T cells and B cells (41). Based on knowledge of gene expression, activation will in each case be accompanied by important changes in methylation profiles (Fig. 1). The manipulation of cells during their purification may also perturb their methylation profile and gene expression. In these circumstances, simplistic isolation of pure cell types and their use as references may generate additional ungovernable influences on study outcomes.

Will it then be possible to detect meaningful associations from unfractionated cells? Assuming experimental variables (ancestry, DNA extraction methods, batch and platform effects) are controlled, then the power to detect cell-specific associations from PBLs and other tissues will depend on the proportion of each cell type, the effect size in specific cells and the sample size (Fig. 1). Assuming the variance at a locus arises from a particular cell type in a sample, then the overall variance in DNA from PBL or a tissue will produce an attenuation of the effect size in specific cells that would mask associations rather than magnify them. Similarly, the study of the wrong cell types (such as in some circumstances, the Epstein–Barr virus-transformed lymphoblastoid cell lines that are commonly used as a renewable source of DNA for genetic studies) will produce null results rather than systematic false positives.

Pushing on and ignoring heterogeneity for the moment, a positive association result may then be entirely due to the presence of particular cells or to the activation of particular cells, or as is perhaps most likely, a combination of these effects (Fig. 1) (it is also quite conceivable that the effects may be general and not cell specific at all). These primary results may themselves be of interest, such as the highly reproducible effects of cigarette smoking on genes potentially affecting coagulation pathways (42,43), but a general next step becomes to engage cellular heterogeneity and to differentiate between possibilities.

Although whole blood is often the only available tissue in large-scale epidemiology studies, it is fortunate that phenotyping of subjects often includes full blood counts (FBCs, also complete blood counts), which are routinely measured by clinically standardized automated procedures. The FBC gives absolute counts and proportions of all major constituents of PBL, with high reproducibility and well-understood normal values.

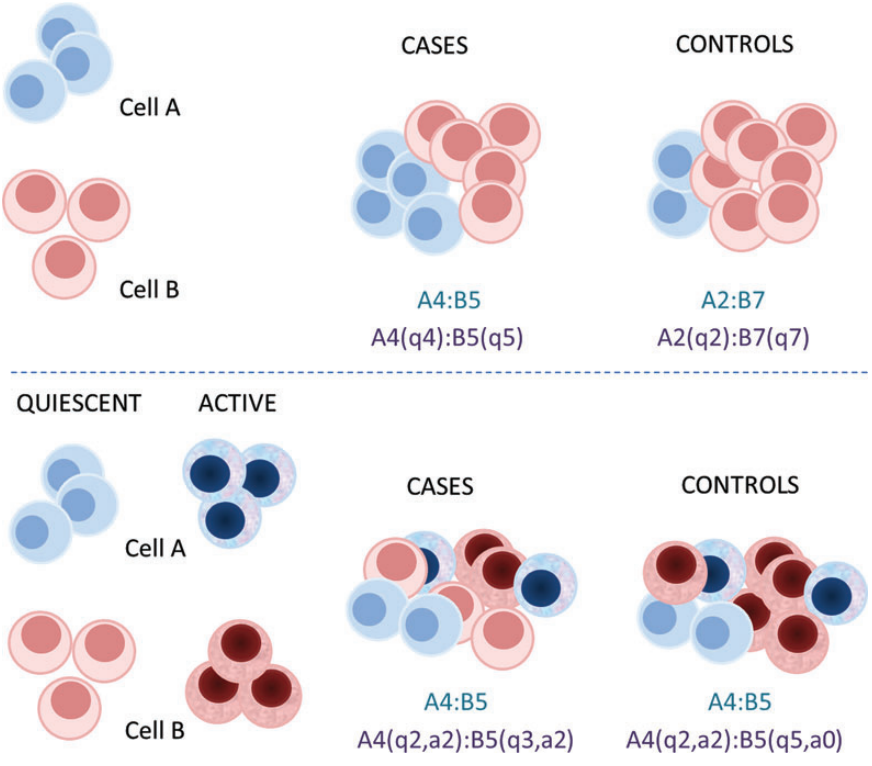

Figure 1. Cellular heterogeneity and epigenetic association. The top panel of the figure shows how two types of cells (**A** and **B**) may be in different proportions in tissue samples from cases (for example, A4 : B5) and controls (A2 : B7). If epigenetic variation between cases and controls is confined to these cells, then their proportions (shown in blue) will be reflected in their epigenetic ratios (shown in purple: **q** indicates the number of cells that are quiescent). This does not, however, render the results of association testing unimportant. The lower panel illustrates the case when cell type A and cell type B are capable of quiescent and active states (as, for example, is usual for immunological cells). Here, simple cell counts will miss significant alterations in the epigenetic profile of cell type B (shown in purple: **q** indicates the number of cells that are quiescent and **a** the number that are active). In both examples, the measure of epigenetic changes provides an accurate estimate of differences between case and control samples.

To account for cellular heterogeneity in epigenetic association studies, one commonly used approach is to include these white cell counts (or their proportions) as fixed effects in regression models (44). These factors are often included as linear terms in the regression model and can adjust for confounding in most situations, although non-linear effects from cell counts should be considered when examining interactions (45).

In the absence of an FBC, or for other factors that are known but not observed, association models can in theory be built on factors that are estimated from known methylation patterns in particular cell types such as CD8+, CD4+ and natural killer T cells, based on a cell-sorted methylation data (5,20). These achieve reliable estimates of fractions of major cell types in tightly controlled experiments, but their practical utility may be confounded by disease-relevant cellular activation in patients and population samples (Fig. 1). Additionally, reference panels based on sorted cells are not available from important sources such as adipose or tumour tissue.

In the absence of reference data, major components in genome-wide DNA methylation patterns from PBL can be estimated and used as surrogates of cell proportions (46,47), even to the suggested extent that it may be helpful in non-hematopoietic cancers (48).

Strong associations between methylation and exposure or outcome should be taken into account when estimating surrogate components, using analyses such as those implemented in RefFreeEWAS (49), surrogate variable analysis (50) in SVA (12) or SVA-PLS (51) and the Bayesian factor analysis package PEER (13,52). Surrogate components, estimated explicitly with PCA or MDS or implicitly using linear mixed models such as EWASher (53), may lose power to detect true association with large effect sizes but are of most use when effect size is smaller than confounding factors (8,14). Genomic control remains a final approach to correct for systematic inflation in false positives in an association study (54).

A further approach to detecting for cell-specific effects may lie in network analyses, which can identify co-regulated gene modules that represent functional biological units of a system (55). Correlation-based networks discovered through the WCGNA package (56) have previously been shown to strongly correlate with particular cell types using global gene expression (57,58) as well as global CpG methylation patterns (59).

Whilst certainly helpful, it is likely that the statistical inferences described earlier will never be better than direct cell counts, and any models incorporating measured or imputed cell counts may never be completely convincing, at least to referees.

For epigenetic associations to be substantiated, the onus still falls on validation of primary associations in secondary panels of subjects, ideally followed by the demonstration of effects in isolated cells from cases and controls. In understanding the functional consequences of epigenetic associations,

it may become important to appreciate that methylation shows a complex and currently unpredictable relationship to gene expression (60).

CONCLUSIONS

The use of epigenome-wide information in association with other studies is therefore far from facile. In addition to effects of age, sex, genetic polymorphism, cellular heterogeneity and the environment are added DNA extraction methods, batch effects and the recognition that activated immune, and other cells exhibit a very different epigenotype to their resting namesakes. Nevertheless, these obstacles are surmountable with appropriate study designs and analytical tools. We believe that systematic mapping of loci influenced by age, sex and SNPs will allow direct control of their influence in future studies, as well as examining important biological processes.

Finally, it should be realized that dynamic epigenetic changes may follow rather than initiate disease processes, so that association with epigenetic variation may not indicate causality in the same way as an SNP association. Nevertheless, epigenomics provides the opportunity for remarkable insights into genome function. Global gene expression is of proven value in interpreting the functional consequence of disease associations (36), but its measurement is expensive and requires stringent sample collection and storage. In contrast, DNA from PBL is often available from historic population and case–control studies, so that genome-wide mapping of CpG methylation may directly inform on mechanisms of diverse diseases, even from DNA that is 30 000 years old (61).

Conflict of Interest statement. None declared.

FUNDING

The Freemasons' Grand Charity, the Wellcome Trust under WT 077959 and WT096964 and the NIH R01 HL101251–01. Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

REFERENCES

1. Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.
2. Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.
3. Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R. and Bird, A. (2011) Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.*, **21**, 1074–1086.
4. Calvanese, V., Fernandez, A.F., Urdinguio, R.G., Suarez-Alvarez, B., Mangas, C., Perez-Garcia, V., Bueno, C., Montes, R., Ramos-Mejia, V., Martinez-Camblor, P. *et al.* (2012) A promoter DNA demethylation landscape of human hematopoietic differentiation. *Nucleic Acids Res.*, **40**, 116–131.
5. Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A. *et al.* (2013) Recommendations for the design and analysis of epigenome-wide association studies. *Nat. Methods*, **10**, 949–955.
6. Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S.E., Greco, D., Soderhall, C., Scheynius, A. and Kere, J. (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE*, **7**, e41361.
7. Kohane, I.S., Kho, A.T. and Butte, A.J. (2002) *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA.
8. Liang, L., Morar, N., Dixon, A.L., Lathrop, G.M., Abecasis, G.R., Moffatt, M.F. and Cookson, W.O. (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.*, **23**, 716–726.
9. Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
10. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. and Balasubramanian, S. (2014) Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.*, **6**, 435–440.
11. Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
12. Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.
13. Stegle, O., Parts, L., Durbin, R. and Winn, J. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, **6**, e1000770.
14. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.
15. Alisch, R.S., Barwick, B.G., Chopra, P., Myrick, L.K., Satten, G.A., Conneely, K.N. and Warren, S.T. (2012) Age-associated DNA methylation in pediatric populations. *Genome Res.*, **22**, 623–632.
16. Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A. *et al.* (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.*, **8**, e1002629.
17. Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M. *et al.* (2010) Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.*, **20**, 434–439.
18. Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M. *et al.* (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.*, **20**, 440–446.
19. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y. *et al.* (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell.*, **49**, 359–367.
20. Jaffe, A.E. and Irizarry, R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R31.
21. Liu, J., Morgan, M., Hutchison, K. and Calhoun, V.D. (2010) A study of the influence of sex on genome wide methylation. *PLoS ONE*, **5**, e10028.
22. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
23. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
24. Carrel, L. and Willard, H.F. (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, 400–404.
25. Cotton, A.M., Lam, L., Affleck, J.G., Wilson, I.M., Penaherrera, M.S., McFadden, D.E., Kobor, M.S., Lam, W.L., Robinson, W.P. and Brown, C.J. (2011) Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum. Genet.*, **130**, 187–201.
26. Khulan, B., Cooper, W.N., Skinner, B.M., Bauer, J., Owens, S., Prentice, A.M., Belteki, G., Constancia, M., Dunger, D. and Affara, N.A. (2012) Periconceptional maternal micronutrient supplementation is associated with widespread gender related changes in the epigenome: a study of a unique resource in the Gambia. *Hum. Mol. Genet.*, **21**, 2086–2101.
27. Heijmans, B.T., Kremer, D., Tobi, E.W., Boomsma, D.I. and Slagboom, P.E. (2007) Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum. Mol. Genet.*, **16**, 547–554.
28. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E. *et al.* (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.*, **40**, 904–908.

29. Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S. and Ophoff, R.A. (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS ONE*, **4**, e6767.
30. Vanderweele, T.J. and Vansteelandt, S. (2010) Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.*, **172**, 1339–1348.
31. Lin, D.Y., Fleming, T.R. and De Gruttola, V. (1997) Estimating the proportion of treatment effect explained by a surrogate marker. *Stat. Med.*, **16**, 1515–1527.
32. Lange, T. and Hansen, J.V. (2011) Direct and indirect effects in a survival context. *Epidemiology*, **22**, 575–581.
33. Baron, R.M. and Kenny, D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, **51**, 1173–1182.
34. Relton, C.L. and Davey Smith, G. (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.*, **41**, 161–176.
35. Consortium, G.T. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
36. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
37. Mantovani, A., Cassatella, M.A., Costantini, C. and Jaillon, S. (2011) Neutrophils in the activation and regulation of innate and adaptive immunity. *Nat. Rev. Immunol.*, **11**, 519–531.
38. Rothenberg, M.E., Owen, W.F. Jr., Silberstein, D.S., Woods, J., Soberman, R.J., Austen, K.F. and Stevens, R.L. (1988) Human eosinophils have prolonged survival, enhanced functional properties, and become hypodense when exposed to human interleukin 3. *J. Clin. Invest.*, **81**, 1986–1992.
39. Kita, H. (2011) Eosinophils: multifaceted biological properties and roles in health and disease. *Immunol. Rev.*, **242**, 161–177.
40. Farina, C., Theil, D., Semlinger, B., Hohlfeld, R. and Meinl, E. (2004) Distinct responses of monocytes to Toll-like receptor ligands and inflammatory cytokines. *Int. Immunol.*, **16**, 799–809.
41. Feske, S. (2007) Calcium signalling in lymphocyte activation and disease. *Nat. Rev. Immunol.*, **7**, 690–702.
42. Breitling, L.P., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.*, **88**, 450–457.
43. Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A. and Demeo, D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.*, **21**, 3073–3082.
44. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M. *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, **31**, 142–147.
45. Cornelis, M.C., Tchetgen, E.J., Liang, L., Qi, L., Chatterjee, N., Hu, F.B. and Kraft, P. (2012) Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am. J. Epidemiol.*, **175**, 191–202.
46. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. and Kelsey, K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.
47. Koestler, D.C., Christensen, B., Karagas, M.R., Marsit, C.J., Langevin, S.M., Kelsey, K.T., Wiencke, J.K. and Houseman, E.A. (2013) Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics.*, **8**, 816–826.
48. Koestler, D.C., Christensen, B., Karagas, M.R., Marsit, C.J., Langevin, S.M., Kelsey, K.T., Wiencke, J.K. and Houseman, E.A. (2012) Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. *Cancer Epidemiol. Biomarkers Prev.*, **21**, 1293–1302.
49. Houseman, E.A., Molitor, J. and Marsit, C.J. (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, **30**, 1431–1439.
50. Teschendorff, A.E., Zhuang, J. and Widschwendter, M. (2011) Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, **27**, 1496–1505.
51. Chakraborty, S., Datta, S. and Datta, S. (2012) Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics*, **28**, 799–806.
52. Stegle, O., Kannan, A., Durbin, R. and Winn, J. (2008) *Accounting for non-Genetic Factors Improves the Power of eQTL Studies*. Vingron, M. and Wong, L. (eds): RECOMB 2008, LNBI 4955. Springer-Verlag, Berlin, Heidelberg, pp. 411–422.
53. Zou, J., Lippert, C., Heckerman, D., Aryee, M. and Listgarten, J. (2014) Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods*, **11**, 309–311.
54. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
55. Schadt, E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.
56. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
57. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.
58. Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H. *et al.* (2010) The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, **463**, 318–325.
59. van Eijk, K.R., de Jong, S., Boks, M.P., Langeveld, T., Colas, F., Veldink, J.H., de Kovel, C.G., Janson, E., Strengman, E., Langfelder, P. *et al.* (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, **13**, 636.
60. Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T. and Blanchette, M. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.*, **15**, R37.
61. Gokhman, D., Lavi, E., Prufer, K., Fraga, M.F., Riancho, J.A., Kelso, J., Paabo, S., Meshorer, E. and Carmel, L. (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, **344**, 523–527.