EPIGENETICS

Statistical and integrative system-level analysis of DNA methylation data

Andrew E. Teschendorff[1–3] and Caroline L. Relton[4]

Abstract | Epigenetics plays a key role in cellular development and function. Alterations to the epigenome are thought to capture and mediate the effects of genetic and environmental risk factors on complex disease. Currently, DNA methylation is the only epigenetic mark that can be measured reliably and genome-wide in large numbers of samples. This Review discusses some of the key statistical challenges and algorithms associated with drawing inferences from DNA methylation data, including cell-type heterogeneity, feature selection, reverse causation and system-level analyses that require integration with other data types such as gene expression, genotype, transcription factor binding and other epigenetic information.

Bisulfite conversion
A technique in which DNA is treated with bisulfite, resulting in modification (upon amplification) of unmethylated cytosines into thymines, whereas methylated cytosines are protected from modification.

[1]*Department of Women's Cancer, University College London, 74 Huntley Street, London WC1E 6AU, UK.*
[2]*UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK.*
[3]*Chinese Academy of Sciences (CAS) Key Laboratory of Computational Biology, CAS–Max Planck Gesellschaft (MPG) Partner Institute for Computational Biology, 320 Yue Yang Road, Shanghai 200031, China.*
[4]*Medical Research Council Integrative Epidemiology Unit (MRC IEU), School of Social & Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK.*

Correspondence to A.E.T.
a.teschendorff@ucl.ac.uk

doi:10.1038/nrg.2017.86
Published online 13 Nov 2017

DNA methylation (DNAm) refers to the covalent attachment of a methyl (CH_3) group to DNA bases, which for eukaryotes is usually 5-methylcytosine (5mC) in the context of cytosine–guanine dinucleotides (CpGs). Like other epigenetic modifications, DNAm is mitotically heritable and plays a key role in embryonic development and regulation of gene expression[1]. As such, DNAm is highly cell-type-specific. DNAm is also influenced by genotype and can be altered by exposure to external factors, such as smoking and diet[2–6]. Like somatic mutations, DNAm changes accrue with age[4,7,8] and are thought to mediate the effects of environmental risk factors on disease incidence and to contribute to disease progression and treatment resistance[9,10]. Irrespective of their potential causal role, DNAm-based biomarkers offer great promise for risk prediction, early detection and prognosis[9]. Their discovery is facilitated by technologies that allow genome-wide measurement of DNAm in a high-throughput manner[11]. Importantly, the metastability of DNAm and the DNA-based nature of the assays provide important technical advantages over measuring histone modifications or mRNA expression. In particular, DNAm assays based on bisulfite conversion are highly quantitative and reproducible, offering high sensitivity to detect small (~1%) changes in DNAm from samples with limited amounts of available DNA. Among these, the Illumina BeadChip microarray technology[12,13] offers a good compromise between cost and coverage and is so far the most popular choice for epigenome-wide association studies (EWAS), which require DNAm measurements in hundreds if not thousands of samples[13]. By contrast, the higher coverage and cost of whole-genome bisulfite sequencing (WGBS) and reduced-representation bisulfite sequencing (RRBS) make these the optimal technologies for mapping reference DNA methylomes, as generated by international consortia such as the US National Institutes of Health (NIH) Roadmap Epigenomics Project, the International Human Epigenome Consortium (IHEC) and BLUEPRINT[14,15], or for measuring genome-wide DNAm patterns from low-yield DNA samples such as cell-free DNA (cfDNA) in plasma[16].

Rigorous and reliable inference from DNAm data is key to a wide range of downstream tasks in EWAS, including the identification of disease biomarkers and causal relationships. These tasks require careful statistical analyses, starting with quality control steps that assess the reliability of the data, followed by intra-sample normalization to adjust for sample-specific technical biases (for example, incomplete bisulfite conversion and background correction). Beyond the obvious importance and need for such normalization, downstream statistical analyses need to deal with other challenges, notably including batch effects and other confounding factors, feature selection and integration with other types of omic data. Given that DNAm is highly cell-type-specific, cell-type heterogeneity of complex tissues (for example, blood or breast) constitutes a major confounder, requiring the application of cell-type deconvolution algorithms. These algorithms offer a form of *in silico* or virtual microdissection, allowing inference of DNAm changes that are not driven by alterations in tissue composition. Other DNAm alterations have been found to be reproducibly associated with different environmental factors (for example, smoking and obesity)[17–19], which can also cause confounding in EWAS. Reverse causation also poses challenges, as observed in the case of the relationship between obesity and DNA methylation, where the prevailing evidence points to the phenotype of interest altering DNAm rather than vice versa[18,20]. The interpretability of an EWAS is also limited

by DNAm being an imperfect measure of gene activity, thus requiring integration with other types of data (for example, mRNA expression or chromatin immunoprecipitation followed by sequencing (ChIP–seq)) in order to help improve causal inference and interpretation. Although statistical methods for such integrative analyses are underdeveloped, the technical reliability of DNAm measurements makes DNAm the ideal epigenetic focal point for such system-level analyses.

Here, we discuss the aforementioned statistical challenges and review the corresponding computational algorithms and software, focusing throughout on downstream analyses, that is, after intra-sample normalization. We first consider confounding factors, owing to the need to determine the major sources of inter-sample variation, with an emphasis on cellular heterogeneity and cell-type deconvolution algorithms. Next, we turn to the main task of an EWAS, which is feature selection. To help with the interpretation of EWAS data, we subsequently describe methods for integrating DNAm with other types of omic data, such as genotype, mRNA expression and transcription factor (TF) binding data, including approaches to strengthen causal inference. We end with an outlook on outstanding statistical challenges and a prediction of how the field will develop. Details of technologies for generating DNAm data and associated intra-sample normalization methods are not covered here, as they were recently reviewed elsewhere[21–24].

Cell-type heterogeneity and deconvolution

EWAS seek to identify differentially methylated cytosines (DMCs) between cases and controls. This task is hampered by variations in the proportions of cell types that make up the tissue where DNAm is measured. These proportions may vary substantially between cases and controls, and while this variation may be biologically and clinically important[25,26], they often reflect changes that are consequential of the disease state, hampering the identification of alterations that may drive disease risk or progression[27–29]. For example, rheumatoid arthritis (RA) was shown to be associated with a shift in the granulocyte-to-lymphocyte ratio, leading to thousands of DMCs, most of which disappeared upon correction for cell-type composition[30].

In general, cell-type deconvolution methods are needed to address any of the following four aims: estimation of absolute or relative cell-type fractions within the samples of interest; identification of DMCs that are not the result of changes in cell-type composition; identification of DNAm profiles representing cell types in the tissue of interest; and identification of the cell type (or types) carrying the DMCs. Broadly speaking, statistical paradigms for cell-type deconvolution fall into two main categories, called 'reference-based' (REF. 31) (if it uses *a priori* defined DNAm reference profiles of representative cell types in the tissue of interest) and 'reference-free' (REF. 32) (BOX 1). Other work has developed a third paradigm ('semi-reference-free')[33,34], which circumvents some of the disadvantages of both reference-free and reference-based methods (BOX 1).

Reference-based cell-type deconvolution. The main requirements underlying reference-based inference are that the main constituent cell types of the tissue are known and that reference molecular profiles representing these cell types are available. Importantly, the reference profiles need to be defined only over features that are informative of differences between cell types; for example, in the DNAm context, they should ideally represent cell-type-specific DNAm markers or be highly discriminative of the different cell subtypes in the tissue of interest. The construction of such reference profiles usually needs to be completed in advance of the study, and it typically requires the generation of genome-wide DNAm data of cell populations purified by fluorescence-activated cell sorting (FACS) or magnetic-activated cell sorting (MACS), followed by statistical analysis to select DMCs between cell subtypes. The importance of constructing a high-quality reference profile database has recently been highlighted[35]. For instance, similar cell types are likely to have highly collinear profiles, which may result in unstable parameter estimation[36]. This is of particular concern if quality control causes a relatively large number of CpGs present in the reference database to drop out, which may further aggravate the collinearity. Hence, it has been proposed that a reference database should maximize the condition number of the matrix it defines[37], which in effect ensures maximal stability of the inference to random loss of features in the reference database.

Assuming a reference database exists, there are then two approaches to infer cell-type fractions within a sample of interest. Both methods effectively run a multivariate regression of the DNAm profile of the sample against the reference DNAm profiles as covariates, with the estimated regression coefficients corresponding to cell-type fractions (if appropriately normalized) (FIG. 1Aa). A widely known technique named constrained projection (CP) (also called quadratic programming (QP)) performs least-squares multivariate regression while imposing normalization constraints on the regression coefficients, which allows the estimated coefficients to be directly interpreted as cell-type proportions within the sample[31,38]. An alternative 'non-constrained' approach is to impose the non-negativity and normalization constraints after estimation of the regression coefficients. This is the approach taken by CIBERSORT, which implements a penalized multivariate regression, originally presented in the context of gene expression data[37]. A similar non-constrained approach can be taken with robust partial correlation (RPC) (a robust form of multivariate regression)[37,39]. A recent comparative DNAm study of CP, CIBERSORT and RPC concluded that for realistic noise levels, RPC and CIBERSORT might be preferable over CP[39], consistent with findings obtained for gene expression data[37].

Methods such as CP or CIBERSORT use reference DNAm profiles defined as the average DNAm over biological replicates, using DMCs that maximize the differences in mean methylation between cell types. Ideally, these DMCs would also exhibit very stable (that is, ultra-low variance) DNAm profiles within

Epigenome-wide-association studies
(EWAS). A study design that seeks associations between DNA methylation at many sites across the genome and an exposure, trait or disease of interest.

Intra-sample normalization
The procedure of adjusting the raw data profile of a biological sample for technical biases and artefacts. This is often followed by inter-sample normalization, in which adjustments are made to the data for technical and biological factors that otherwise cause unwanted (and often confounding) data variation across samples.

Confounding
When the relationship between an exposure and an outcome is not causal but is due to the effects of a third variable (the confounder) on the exposure and the outcome. White blood cell heterogeneity can act as a confounder in many epigenetic studies.

Feature selection
The statistical procedure of identifying features which, in some broad sense, correlate with an exposure or phenotype of interest (POI).

Differentially methylated cytosines
(DMCs). Cytosines (usually in a CpG context) that exhibit a statistically significant difference in DNA methylation between two groups of samples, according to some statistical test.

Condition number
In the context of reference-based cell-type deconvolution, the condition number of a reference matrix represents an index of the numerical stability of the inference. Formally, it measures the sensitivity of the regression parameters (also known as cell weights) to small perturbations or errors in the reference matrix.

Box 1 | **Statistical inference paradigms for cell-type deconvolution**

Reference-based cell-type deconvolution tools

These methods correct for cell-type heterogeneity by using an existing reference DNA methylation (DNAm) database of cell types that are thought to be present in the tissue of interest. If the main underlying cell types of the tissue are known, then estimates of the absolute cell-type fractions are possible; otherwise, estimated fractions are relative. The estimated absolute or relative cell-type fractions can then be used as covariates in supervised multivariate regression models to infer differentially methylated cytosines (DMCs) that are independent of changes in cell-type composition.

Advantages

- Absolute or relative cell-type fractions can be estimated in each individual sample.
- If required, they can be easily combined with batch-correction methods such as COMBAT.
- The model itself is relatively assumption free.

Disadvantages

- The tools require knowledge of the main cell types that are present in the tissue. Reliable reference DNAm profiles must be available for these cell types.
- On their own, they cannot deal with unknown confounding factors.
- They assume that cell–cell interactions in the sample do not affect the DNAm profiles of the individual cell types.
- Reference profiles could be confounded by factors such as age or genotype.

Reference-free cell-type deconvolution tools

These methods correct for cell-type heterogeneity by inferring from the full data matrix 'surrogate variables', which include sources of data variation that are driven by cell-type composition. These surrogate variables are inferred from the data without the need for a reference DNAm database and are used as covariates in the final supervised multivariate regression model to infer DMCs that are independent of changes in cell-type composition and other cofounders.

Advantages

- There is no requirement to know the main cell types in a tissue or to have reference DNAm profiles; hence, in principle, they are applicable to any tissue type.
- *De novo* (unsupervised) discovery of novel cell subtypes.
- They allow for the possibility that cell–cell interactions alter the profiles of individual cell types.
- They can adjust simultaneously for other confounding factors, known or unknown.

Disadvantages

- Without further biological input, they cannot provide estimates of cell-type fractions in individual samples.
- Performance is strongly dependent on model assumptions, which are often not satisfied.

Semi-reference-free cell-type deconvolution tools

This is a third paradigm that corrects for cell-type heterogeneity by inferring surrogate variables representing variation due to cell-type composition but that, unlike a purely 'reference-free' approach, does so by using partial prior biological knowledge of which cytosine–guanine dinucleotides (CpGs) differ between cell types. Typically, these tools infer the surrogate variables from the reduced data matrix, projected on this set of selected features.

Advantages

- They allow for the possibility that cell–cell interactions alter the DNAm profiles of individual cell types.
- If required, they can be combined with batch-correction methods such as COMBAT.
- They are more robust to incomplete knowledge of underlying cell types in the tissue of interest.
- They can provide approximate relative estimates of cell-type fractions in individual samples.

Disadvantages

- Performance is still strongly dependent on model assumptions, which may not be satisfied.
- Inference of absolute cell-type fractions in individual samples remains challenging.
- The ability to resolve highly similar cell types is limited.

cell types, appearing as strongly bi-modal profiles. However, depending on the tissue and cell types, such bi-modal DMCs may not be present, so it may also be necessary to include the variance in DNAm when performing reference-based deconvolution. For instance, an algorithm called CancerLocator models reference DNAm profiles using beta distributions, generating beta-distribution references for healthy plasma DNA and solid tumours, subsequently using a two-state beta-mixture model to infer tumour burden and tissue of origin of circulating tumour DNA (ctDNA) in plasma[40] (FIG. 1Ab). Similarly, algorithms for inferring tumour purity of primary cancers also use explicit beta distributions and have been shown to provide accurate estimates, in line with gold-standard estimates derived from copy-number data[41–43].

Reference-free cell-type deconvolution. To date, there are two main types of reference-free methods (BOX 1), which differ greatly in terms of their model assumptions. One class is widely known as surrogate variable analysis (SVA)[44–46], an approach developed originally to address general unknown confounding factors and that has also gained considerable favour for cell-type deconvolution[47–49]. SVA uses the phenotype of interest (POI) from the outset and attempts to construct 'surrogate variables' that capture confounding variation of any sort (that is, not just cell-type compositional changes but, for example, also batch effects) in the space of variation that is 'orthogonal' to that associated with the POI[44,45,50]. A variant of SVA, called RefFreeEWAS[32], assumes an explicit mixture-modelling structure (as required for modelling cell-type composition) and has been demonstrated to work well[32,51]. Another variant of SVA, called independent surrogate variable analysis (ISVA)[50], is similar to SVA but uses a blind source separation (BSS) algorithm (independent component analysis (ICA)[52]) instead of principal component analysis (PCA) in the residual variation space, which may help to identify a more relevant subspace of confounding variation (that is, a subset of surrogate variables). The need for this subspace selection step may arise if the model describing the effect of the POI on the data is a poor one, as this may result in variation associated with the POI being found in the surrogate variable subspace[50]. Unlike PCA, BSS is designed to disentangle independent sources of variation[52] and is therefore better suited for deconvolving the residual biological variation associated with the POI from potential confounding variation.

Another set of reference-free approaches, exemplified by methods such as EWASher[53] or ReFACTor[54], do not use the phenotype of interest when inferring latent components associated with cell-type composition. This is only possible if certain assumptions are made. Specifically, EWASher and ReFACTor assume that the top principal component of variation in the data is associated with changes in cell-type composition, an assumption that will not hold if the POI accounts for a larger proportion of data variance. Thus, the applicability of these two methods is critically dependent on the POI and the underlying tissue type (FIG. 1B).

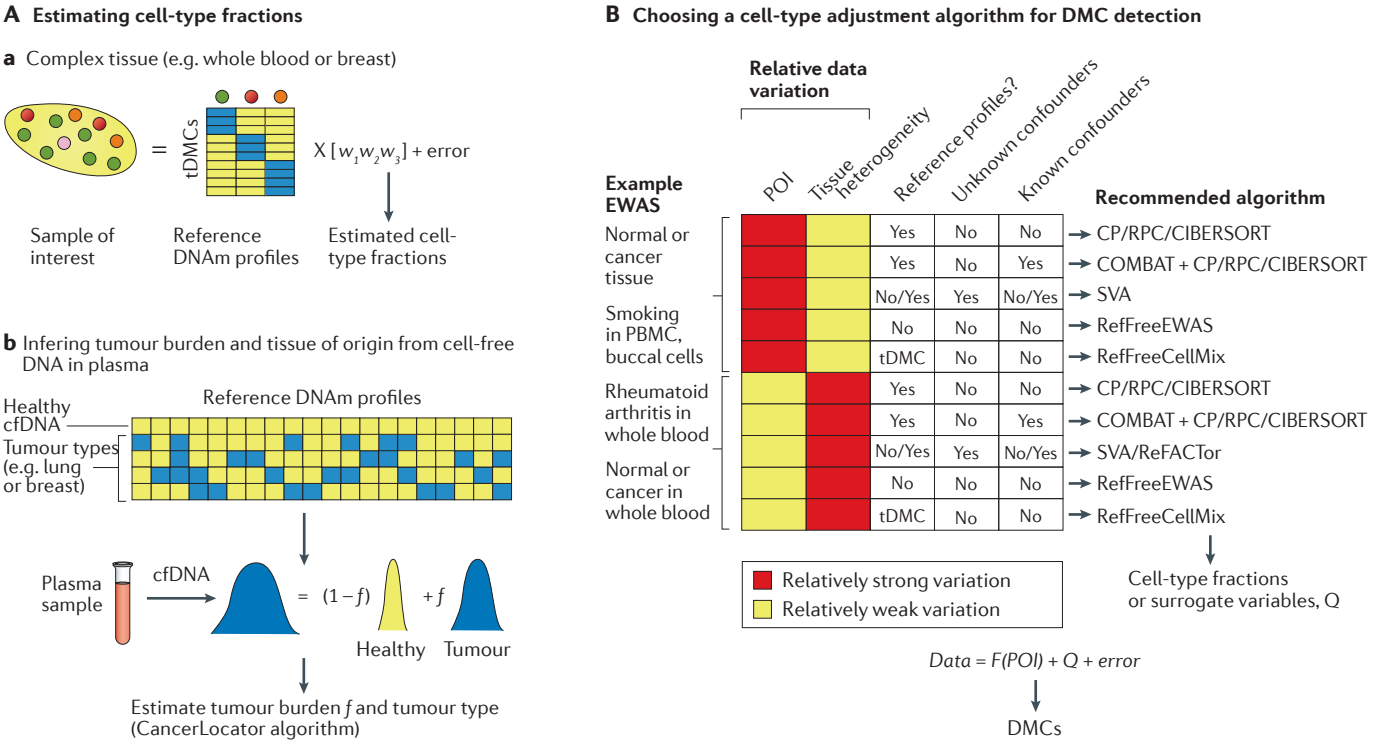


Figure 1 | **DNA methylation analysis of cell-type heterogeneity.** **Aa** | Estimating cell-type fractions in a sample for which a genome-wide DNA methylation (DNAm) profile is available is an important task, as changes in these proportions can have biological and clinical importance or can confound analyses. Constrained projection (CP) infers these proportions by running a constrained multivariate regression model of the sample's DNAm profile against reference DNAm profiles for the cell types of interest, with the estimated regression coefficients (w_1, w_2 and w_3) representing cell proportions. **Ab** | From a plasma sample, estimating the relative fractions of cell-free DNA (cfDNA) from healthy cells versus circulating tumour DNA (ctDNA) presents a novel promising clinical application for non-invasive early detection and disease monitoring. The CancerLocator algorithm (TABLE 1) allows estimation of the tumour burden (denoted *f*) and the type of tumour. **B** | Cell-type heterogeneity may cause confounding and compromise the identification of differentially methylated cytosines (DMCs) in epigenome-wide association studies (EWAS). The diagram presents recommendations as to which statistical algorithms might be better suited for different EWAS scenarios. This depends on whether reference DNAm profiles are available, the presence of unknown confounders and technical batch effects (known confounders). When reference profiles are available, reference-based methods are recommended unless there is evidence of other confounding variation, in which case surrogate variable analysis (SVA)-like methods are preferable. If partial prior information is available, such as if cell-type-specific DMCs (tDMCs) are known but no reference profiles are available, a semi-reference-free approach like RefFreeCellMix is recommended. Relative data variation between the phenotype of interest (POI) and that due to cell-type heterogeneity is important when deciding between reference-free methods. Finally, DMCs are inferred using a multivariate regression of the data against the POI (*F* denotes the link function) and cell-type fractions or surrogate variables as covariates (denoted *Q*). Note that regression coefficients have been omitted for the sake of clarity. **C** | A third important task is the quantification of epigenetic heterogeneity within a given cell type, for instance, quantifying clonal heterogeneity within tumour cells. Given that DNAm normally exhibits strong spatial correlations on scales up to approximately 500 bp and that tumours are characterized by widespread deviations from the DNAm ground state, one way to approximate clonal epigenetic heterogeneity is to measure the proportion of discordant reads (PDR). Tumours characterized by high epigenetic clonal heterogeneity have been found to exhibit worse clinical outcome (see the main text). For specific algorithms mentioned in this figure, see TABLE 1. CpG, cytosine–guanine dinucleotide; CR, concordant reads; DR, discordant reads; PBMC, peripheral blood mononuclear cells; RPC, robust partial correlations; RRBS, reduced-representation bisulfite sequencing; RUV, removing unwanted variation; WGBS, whole-genome bisulfite sequencing.

Constrained projection
(CP). Also known as quadratic programming (QP). A widely used technique for performing multivariate linear regression with constraints (such as non-negativity and normalization) imposed on the regression coefficients. In the context of cell-type deconvolution, the coefficients correspond to cell-type proportions in a sample. By definition, these proportions are non-negative, and their sum must be ≤1.

Beta distributions
The distributions of beta values. The beta value is a statistical term used to describe the quantification of DNA methylation at a given cytosine, as the ratio of methylated alleles to the total number of alleles (methylated + unmethylated), a number that by definition must lie between 0 (fully unmethylated) and 1 (fully methylated).

Surrogate variable analysis
(SVA). A widely used technique for selecting features associated with a factor of interest, which is not confounded by other factors. SVA uses a model to identify the data variation that is orthogonal to the factor of interest and subsequently uses principal component analysis (PCA) on this orthogonal variation matrix to construct 'surrogate variables', which in theory should capture confounding sources of variation.

Phenotype of interest
(POI). The factor or variable of interest in an epigenome-wide association study (EWAS). This factor is often binary, representing case–control status, but could also represent an ordinal variable (for example, genotype) or be continuous (for example, age).

Blind source separation
(BSS). The problem of inferring the sources of variation gives rise to a data matrix without using any prior information ('blind'). Algorithms that can achieve this are called BSS algorithms, of which independent component analysis (ICA) is one example.

For instance, the assumption underlying EWASher and ReFACTor may hold in whole blood for a wide range of phenotypes because the granulocyte fraction varies substantially, even among healthy individuals (see, for example, REF. 39), yet in a less complex tissue such as peripheral blood, which is devoid of granulocytes, cell-type compositional changes could account for a much smaller proportion of total data variance. Similarly, in diseases such as cancer, which are characterized by large-scale changes in DNAm, involving most of the genome, only a smaller fraction of these changes are due to changes in cell-type composition[48,55]. Thus, methods such as ReFACTor or EWASher may not offer the level of sensitivity required for many types of EWAS[48].

Semi-reference-free cell-type deconvolution. A promising third paradigm, which remains underexplored, can be viewed as semi-reference-free (BOX 1). Conceptually, it adapts the removing unwanted variation (RUV) framework[56], in that it attempts to infer 'empirical control features', that is, features affected by confounding variation but not associated with the POI, which can subsequently be used to adjust the data. In the context of cell-type deconvolution, a pre-specified set of cell-type-specific DMCs (for example, DMCs that differ between blood cell subtypes) could serve as empirical control features[34,57]. A recent algorithm, called RefFreeCellMix, which uses a constrained form of non-negative matrix factorization (NMF), can be easily adapted in this semi-reference-free manner to infer cell-type proportions[33]. By performing NMF on the reduced data matrix obtained by selecting cell-type-specific DMCs, RefFreeCellMix can obtain estimates of cell-type fractions, from which DMCs associated with a POI can subsequently be inferred using supervised regression. This approach was recently applied to the deconvolution of breast cancer samples (EDec algorithm)[34]. More recently, a regularized version of RefFreeCellMix, called MeDeCom[58], which favours latent factors (representing cell-type-specific DNAm profiles) that exhibit bi-modal (that is, fully unmethylated or methylated) methylation states, has been shown to lead to improved modelling of cell-type composition. All these algorithms also offer a means of identifying the specific cell types carrying the DNAm alterations, although this remains largely unexplored.

Comparison of cell-type deconvolution algorithms. For a given EWAS, the choice of cell-type deconvolution algorithm depends mainly on the availability of a suitable reference DNAm database. The database could be confounded by external factors such as age or genotype, rendering the references less useful for application to data sets where these factors might be very different (for example, using adult blood cell subtype reference profiles to estimate cell subtype fractions in umbilical cord blood[59]); in other cases, reference profiles generated on purified cell populations may not capture important *in vivo* cell–cell interactions, which are known to alter molecular profiles[60] (BOX 1). Beyond these limitations, there are three additional factors to consider when choosing a cell-type deconvolution method: first, the specific information desired (for example, DMCs, cell-type fractions or unsupervised discovery of novel cell types); second, the presence of additional confounding factors and whether these are known or unknown; and third, the POI and tissue type, which determines the relative data variance associated with the POI and cell-type composition. Recommendations and guidelines for different scenarios are provided (see FIG. 1B) and are largely in agreement with those of recent comparative studies[47–49,61]. Briefly, for DMC detection in tissues for which the main underlying cell types are known, reference-based methods, which are relatively assumption free and which can be combined with batch-correction methods such as COMBAT[62], are recommended, unless confounders are unknown, in which case a method like SVA is preferable. Reference-free or semi-reference-free methods are necessary for tissues for which no reference DNAm profiles are available. Because reference-free methods are more dependent on model assumptions, special care must be taken in selecting the most appropriate method, which will depend by and large on the relative data variance carried by the POI and cell-type composition, as well as on the presence of unknown confounders (FIG. 1B). For estimating cell-type fractions, a reference-based algorithm is most appropriate, although semi-reference-based algorithms such as RefFreeCellMix or MeDeCom could also be used if the inferred latent components are uniquely mappable to underlying cell types[33]. Finally, one may also wish to perform cell-type deconvolution in order to discover novel cell types in a tissue of interest. This unsupervised application would require application of methods such as RefFreeCellMix or MeDeCom on the full set of available CpGs rather than on an informed subset of cell-type-specific DMCs.

Epigenetic heterogeneity within cell types. Epigenetic heterogeneity also manifests itself within specific cell types[63], notably pluripotent cells[64] and cells of the immune system[65], but also within haematological cancers[66,67] and the epithelial compartments of solid tumours[55,68]. In the context of precursor cancer lesions, such epigenetic heterogeneity is believed to be an important driver of cancer risk, whereas in cancer, clonal heterogeneity determines disease progression and response to drug treatment[66]. Thus, there is substantial interest in developing statistical measures that can quantify epigenetic clonal heterogeneity. Such quantification is best done using WGBS or RRBS data, because associated reads (representing strings of binary methylated or unmethylated calls at single-nucleotide resolution) have the required spatial resolution to allow epiallelic diversity to be estimated (FIG. 1C). Also of particular importance is the detection of shifts in the proportions of specific epialleles, for which algorithms (for example, methclone[69]) have been developed. In the context of Illumina methylation bead arrays, identifying epigenetic loci marking shifts in epigenetic subclones is possible using statistical tests for detecting methylation outliers[55].

Table 1 | **Algorithms and software for downstream statistical analyses of DNA methylation data**

Name	Description	Programming language	Web links	Refs
Cell-type deconvolution algorithms				
CP/QP	Reference-based method using constrained projection	R	https://github.com/sjczheng/EpiDISH	31
RPC	Reference-based robust partial correlations	R	https://github.com/sjczheng/EpiDISH	39
CIBERSORT	Reference-based support vector regressions	R	https://github.com/sjczheng/EpiDISH	37
SVA	Surrogate variable analysis (reference-free)	R	www.bioconductor.org SVA package	44
ISVA	Independent surrogate variable analysis (reference-free)	R	https://cran.r-project.org/package=isva	50
RefFreeEWAS	Reference-free deconvolution	R	https://cran-r-project.org/package=RefFreeEWAS	32
RefFreeCellMix	Reference-free or semi-reference-free NMF using recursive QP	R	https://cran-r-project.org/package=RefFreeEWAS	33
MeDeCom	Reference-free or semi-reference-free constrained and regularized NMF	R	http://github.com/lutsik/MeDeCom	58
EDec	Like RefFreeCellMix but applied to breast cancer or tissue	R	https://github.com/BRL-BCM/EDec	34
RUV/RUVm	Removing unwanted variation	R	http://www.bioconductor.org missMethyl package	56,208
CancerLocator	Inference of tumour burden and tissue of origin from plasma cfDNA	Java	https://github.com/jasminezhoulab	40
MethylPurify	Tumour purity estimation from WGBS or RRBS data	Python	https://pypi.python.org/pypi/MethylPurify	41
InfiniumPurify	Tumour purity estimation from Illumina Infinium data	Python	https://bitbucket.org/zhengxiaoqi/	42
Algorithms for feature selection				
BSSeq and BSmooth	DMR finder	R	http://www.bioconductor.org bsseq package	209
Bumphunter (minfi)	DMR finder	R	http://www.bioconductor.org minfi package	86,87
DMRcate	DMR finder	R	http://www.bioconductor.org	95
COMETgazer/COMETvintage	Regions of co-methylation and DMC or DMRs	C++ and R	https://github.com/rifathamoudi/COMETgazer https://github.com/rifathamoudi/COMETvintage	83
EVORA/iEVORA	Differentially variable CpGs	R	https://cran.r-project.org/package=evora	55,68, 98,103
DiffVar	Differentially variable CpGs	R	www.bioconductor.org missMethyl package	100
GALMSS	Generalized additive linear model for location, scale and shape	R	https://cran.r-project.org/package=galmss	101
GSEA, pathway, integrative and system-level analysis				
Gometh/gseameth (missMethyl)	Gene ontology and gene set enrichment analysis	R	http://www.bioconductor.org missMethyl package	110
extractAB (minfi)	Estimation of open and closed chromatin regions	R	http://www.bioconductor.org minfi package	178
FEM/EpiMods	Functional epigenetic modules (DNAm and mRNA)	R	http://www.bioconductor.org FEM package	134
SMITE	Significance-based modules integrating transcriptome and epigenome	R	http://www.bioconductor.org SMITE package	160
ME-Class	Methylation-based expression classification and prediction	Python	https://github.com/cschlosberg/me-class	85
ELMER	Enhancer linking by methylation/expression relationships	R	http://www.bioconductor.org ELMER package	147

Table 1 (cont.) | **Algorithms and software for downstream statistical analyses of DNA methylation data**

Name	Description	Programming language	Web links	Refs
GSEA, pathway, integrative and system-level analysis (cont.)				
TENET	Tracing enhancer networks using epigenetic traits	R	http://farnhamlab.com/software http://www.bioconductor.org TENET package	150
TEPIC	Integration of open-chromatin data (for example, NOMe-Seq or DHS) to predict gene expression	Python or C++	https://github.com/schulzlab/TEPIC	210
iCluster/iCluster+	Integrative clustering	R	http://www.bioconductor.org iClusterPlus package	137
PARAFAC (multiway)	Parallel factor analysis and non-Bayesian tensor decomposition	R	https://cran.r-project.org/package=multiway	168
SDA	Sparse decomposition analysis and Bayesian tensor decomposition	Linux executable	https://jmarchini.org/sda	169
JIVE	Joint and individual variation explained	R	https://cran.r-project.org/package=r.jive	166
Methods for causal inference				
MR-Base	An analytical platform that uses curated GWAS data to perform Mendelian randomization tests and sensitivity analyses	R	http://www.mrbase.org	211
JLIM	Joint likelihood mapping	R	http://github.com/cotsapaslab/jlim/	212
Bayesian coloc	Bayesian test for colocalization	R	https://cran.r-project.org/package=coloc	213
gwas-pw	Joint analysis of GWAS signals	R	https://github.com/joepickrell/gwas-pw	214
HEIDI	Heterogeneity in dependent instruments	C++	http://cnsgenomics.com/software/smr/	215

cfDNA, cell-free DNA; CP, constrained projection; CpGs, cytosine–guanine dinucleotides; DHS, DNase-hypersensitive site; DMC, differentially methylated CpG; DMRs, differentially methylated regions; GSEA, gene set enrichment analysis; GWAS, genome-wide association study; NMF, non-negative matrix factorization; NOMe-seq, nucleosome occupancy and methylome sequencing; QP, quadratic programming; RRBS, reduced-representation bisulfite sequencing; WGBS, whole-genome bisulfite sequencing.

Independent component analysis
(ICA). An unsupervised dimensionality reduction algorithm that decomposes the data matrix into a sum of linear components of variation, which are as statistically independent from each other as possible. Statistical independence is a stronger condition than the linear uncorrelatedness of principal component analysis (PCA) components, allowing improved modelling of sources of variation in complex data.

Principal component analysis
(PCA). An unsupervised dimensionality reduction algorithm that decomposes the data matrix into a sum of linear principal components (PCs) of variation, ranked by decreased variance and uncorrelated to each other.

Latent components
Components or sources of data variation that are 'hidden' (or latent) and that are inferred from the data using an unsupervised algorithm.

Feature selection and interpretation

The most common task in analysing omic data is feature selection. For any given EWAS, it is useful to think of CpG DNAm profiles as belonging to specific 'families', each characterized by a particular pattern or shape and each linked to an underlying putative biological (or technical) factor. For instance, DNAm variation of CpGs marking specific cell types will typically exhibit patterns of DNAm variation that correlate linearly with the underlying cell-type fractions, whereas those driven by genetic variants will not. Given that current technologies allow measurement of DNAm in effectively one million to several million CpG sites, small differences in feature selection methods can have a dramatic impact on the specific ranking and selection of CpGs. An appreciation of the intricacies of feature selection is therefore critically important.

Variably methylated cytosines. A popular unsupervised feature selection strategy is to rank and filter features by variance or by a robust version such as the median absolute deviation; the aim is to select the most variably methylated cytosines (VMCs), while also removing those that exhibit little or no variance (which are assumed to represent noise)[70]. However, applying this strategy to DNAm data could bias the selection of features, given that DNAm data are usually quantified in terms of a beta value, which by construction is heteroscedastic. In fact, for beta values, variance is maximal at a value of 0.5 (REF. 71); hence, filtering by variance could favour genomic regions with intermediate mean levels of DNAm. Filtering tools that avoid this bias have been developed[72]. Alternatively, DNAm may be quantified in terms of M-values[71], which can be obtained directly from the log-ratio of intensities of methylated to unmethylated alleles or indirectly from beta values by applying the logit transformation. In principle, M-values are more homoscedastic, although care must be taken with features that have methylation beta values close to 0 or 1, as the logit transformation can turn these into significant outliers[71,73].

In general, VMCs will exhibit a large range of DNAm values and will include those driven by single-nucleotide polymorphisms (SNPs). For a substantial number of these VMCs, the variation will be driven by a SNP affecting the interrogated cytosine (or another cytosine located within the probe body in the case of Illumina bead arrays), and such VMCs are normally removed during quality control[74,75]. For other VMCs, the SNP driving the variation will not be located at the interrogated cytosine (nor in the underlying probe), thus defining methylation quantitative trait loci (mQTLs)[76] (FIG. 2a). Although mQTLs are highly variable, they are not always prominent features driving top components in a PCA unless the study cohort consists of populations stratified by ancestry[18,76,77]. This is because principal components represent components of maximal covariation, so that mQTLs (especially

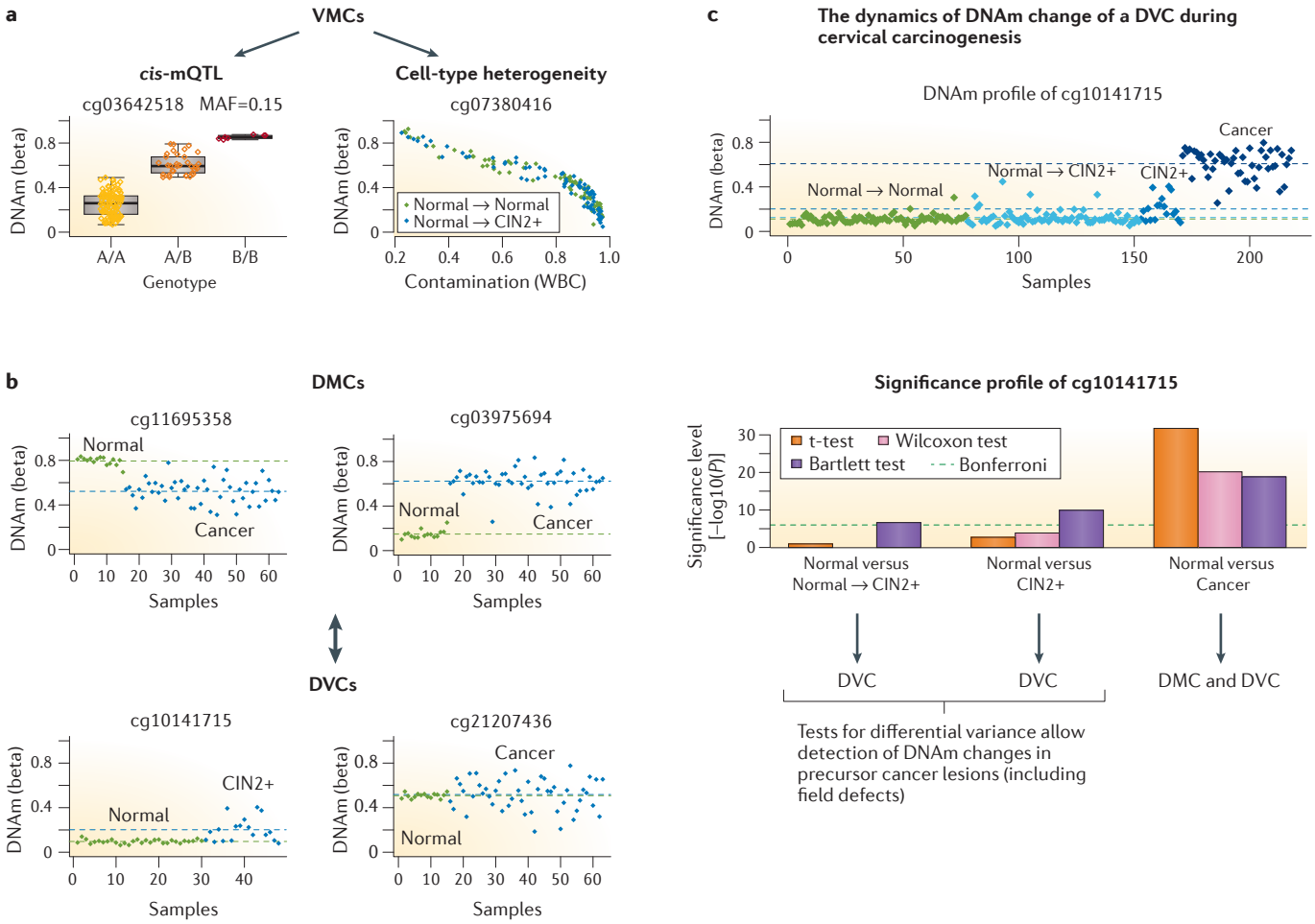


Figure 2 | **Variability, differential means and differential variability in DNA methylation data. a** | Two examples of variably methylated cytosines (VMCs), one driven by single-nucleotide polymorphisms (SNPs) located in *cis* with the indicated cytosine–guanine dinucleotide (CpG) (defining a well-known *cis* methylation quantitative trait locus (*cis*-mQTL)) (left panel) and another driven by variation in immune-cell contamination (right panel). Both profiles of CpG DNA methylation (DNAm) derive from an Illumina Infinium DNAm data set encompassing 152 normal cervical smear samples[68]. For the mQTL, samples are grouped according to the predicted genotype. For the other VMC, blue denotes normal cervical smears from women who 3 years after sample collection developed a cervical intraepithelial neoplasia of grade two or higher (CIN2+), whereas green denotes normal cervical smears from women who remained healthy. This particular VMC is unmethylated in all white blood cells (WBC) but not in cervical epithelial cells, and so the variation in the cervical smear is due to variation in WBC contamination. Panels illustrate how SNPs and cell-type composition can drive large variation in DNAm, but variation that may not correlate with case versus control status. **b** | Contrast between differentially methylated cytosines (DMCs) and differentially variable cytosines (DVCs). Two examples of each are given, drawn from Illumina Infinium DNAm data from normal cervical smears (green) and either cervical intraepithelial neoplasia (CIN2+) or cervical cancer (both blue). The average levels are shown as horizontal dashed lines. Observe how a DMC is typically characterized by most samples in one phenotype exhibiting a deviation in DNAm value. By contrast, a DVC is characterized by a very stable DNAm profile in one phenotype but by DNAm outliers driving large variation in the other. **c** | Example of a CpG that exhibits progression in DNAm between successive stages in cervical carcinogenesis. When comparing normal cervical smears that progress to CIN2+ (Normal→CIN2+) to those that do not (Normal→Normal), this CpG can be identified (that is, with a highly significant *P*-value) only via a test for differential variance (or for deviation from normality) such as Bartlett's test. When comparing CIN2+ to normal cervical smears, differential variance is still the main distinguishing feature. Only when comparing (invasive) cervical cancer to normal cervix does this CpG exhibit a stronger difference in average DNAm, therefore enabling its identification using, for example, *t*-tests or Wilcoxon tests. Thus, this panel illustrates how the DNAm profile of the same CpG changes during cervical carcinogenesis and emphasizes the importance of selecting the appropriate statistical test, as the choice of test will have a dramatic impact on feature selection. All data shown represent real DNAm data derived from REF. 68, with the corresponding CpG identifier given above each panel.

Supervised
Of statistical inferences, using the phenotype of interest from the outset, for instance, when identifying features correlating with a phenotype.

those with low minor allele frequencies) account for only relatively smaller fractions of data covariance. Other VMCs that will appear more prominently in top principal components may be associated with other biological factors such as cell-type composition (FIG. 2a) or may exhibit strongly bi-modal profiles such as those seen in cancer.

Differentially methylated cytosines and regions. The most common supervised feature selection procedure is to select CpGs for which there is a significant difference in the average between phenotypes, defining DMCs (FIG. 2b). The simplest method for selecting DMCs is that based on the absolute difference in mean beta values, which is analogous to the log-fold-change

used in the gene expression context. However, because of the heteroscedasticity of beta values, such filtering may again bias selection against CpGs with very low or very high mean levels of methylation[71]. A much safer option is to apply such thresholding on differences in mean beta value only after having ranked or selected features based on some formal statistic, as the statistic incorporates information about the spread of the data within phenotypes. One option is to use non-parametric Wilcoxon rank sum tests, as these consider only the relative ranking of beta values, although a caveat is that these tests are less powered. Another option is to use *t*-tests. Although *t*-tests require the data within the phenotypes being compared to be Gaussian distributed (an assumption not satisfied with beta-valued data), nevertheless, in practice, this does not impose any more of a limitation than the non-Gaussian nature of, for example, gene expression data from microarrays or RNA sequencing (RNA-seq), for which empirical Bayesian frameworks built on regularized *t*-statistics have proved extremely popular[78–80]. For feature selection, what matters is the distribution of values across samples, and for both DNAm and mRNA expression data, this distribution is approximately Gaussian. Confirming this, *t*-statistics and moderated *t*-statistics have been successfully applied to beta-valued data and shown to lead to very similar rankings compared to the application of the same statistics to M-values[73]. An important exception is when using Bayesian models, which are naturally more sensitive to underlying model assumptions (often Gaussian distributions). For instance, in studies with small sample sizes, empirical Bayes models are necessary for obtaining improved estimates of variance, thus favouring M-values[71,73]. DMCs derived from *t*-tests or regularized *t*-tests may or may not exhibit large differences in average DNAm, since a CpG exhibiting a small (for example, 5%) difference in mean methylation but with low variance within phenotypes may still have a large *t*-statistic. Many smoking-associated DMCs identified in whole blood are of this type[17]. Cancer DMCs, on the other hand, generally exhibit much larger differences in mean DNAm (>30%, FIG. 2b).

Differential methylation can also be called at the regional level. There are a number of reasons why identifying differentially methylated regions (DMRs) is desirable. First, due to the processivity of DNA methyltransferases and other enzymes modifying the epigenome, DNAm is generally highly correlated on scales up to approximately 500 bp and beyond[16,81]. DNAm alterations associated with disease phenotypes and age typically also exhibit such spatially correlated patterns, albeit much weaker[16]. Thus, calling DMRs removes some of the spatial redundancy, helping to reduce the dimensionality of the data. Second, calling differential methylation at the regional level may offer increased robustness, especially in the context of limited-coverage WGBS data[82,83]. Third, although still controversial, DNAm alterations that extend to the regional level are thought to be more functionally important than alterations that affect only isolated sites[84,85]. Statistical algorithms for calling DMRs include bumphunter[86,87], an algorithm originally designed for high-resolution DNAm data (for example, WGBS or CHARM[88]) but that has also been successfully adapted for Illumina Infinium BeadChips and that can allow detection of small (~1–5 kb) DMRs, as well as larger (~100 kb–2 Mb) DMRs, termed differentially methylated blocks (DMBs)[89–94]. A more recent algorithm tailored for WGBS data, and which exploits the spatial correlation structure of DNAm, identifies regions of covariation in methylation (COMETs)[82,83], which can then be used as regional features for differential methylation analysis. Using COMETs to call differential methylation can result in improvements in sensitivity of greater than 40–50% compared with DMC calling, even in WGBS data with 30× coverage[82,83]. Spatial correlation of methylation across different tissues and cell types has also been recently used to define 'methylation haplotype blocks', which facilitates the identification of the tissue of origin of ctDNA in serum[16]. More recently, adopted methods for identifying DMRs are DMRcate[95] and Comb-p[96]. It is noteworthy that each DMR method differs in the assumptions made and statistical approach taken and that different methods therefore very rarely identify precisely the same DMRs.

Differentially variable cytosines and regions. An entirely different feature selection paradigm is based on features that exhibit differential variance in methylation between two phenotypes, so-called differentially variable cytosines (DVCs). This approach computes the variance across samples belonging to the same phenotype and then compares this variance between two or more phenotypes using a statistical test for differential variance[97] (BOX 2). It is important to appreciate that DVCs may not be DMCs (and vice versa) and that there are also different types of DVCs (FIG. 2b).

The importance of differential variance has been most clearly demonstrated in the context of early carcinogenesis[68,98], where differential variance between normal cells from healthy individuals and normal cells at risk of neoplastic transformation is critical to the identification of DNAm alterations that define field defects in breast[55] and cervical cancer[68] (FIG. 2c). These DNAm alterations are characterized by relatively large changes in DNAm (typically 20–30% or higher), defining outliers, that occur predominantly, or exclusively, in the samples at risk of neoplastic transformation (FIG. 2c). As might be expected from DNAm alterations in cells that have not yet undergone neoplastic transformation, these outlier events are relatively infrequent and exhibit a stochastic pattern[55]. However, in cells that have undergone neoplastic transformation or turned invasive, the pattern of DNAm variation becomes more homogeneous and deterministic, in the sense that effectively all (or most) cancer samples exhibit a difference in DNAm (FIG. 2c). By combining differential-variance-based feature selection with an adaptive index classification algorithm[99] in an approach called epigenetic variable outliers for risk-prediction analysis (EVORA)[68], such DVCs have been demonstrated to allow prediction of the prospective risk of cervical cancer (BOX 2). A modification of EVORA, called iEVORA, which offers improved control of the

Variably methylated cytosines
(VMCs). Cytosines (usually in a CpG context) that exhibit a significant amount of variance in DNA methylation, as assessed across independent samples and relative to other CpG sites.

Heteroscedastic
Of a statistical distribution or of a random sample thereof, the expected variance, or spread, being dependent on the mean.

Logit transformation
A mathematical transformation that takes values defined on the unit interval (0,1) (for example, beta values (β)) into values defined on the open interval $(-\infty, +\infty)$, termed M-values. Mathematically, $M = \log_2[\beta/(1-\beta)]$.

Methylation quantitative trait loci
(mQTLs). CpG sites whose DNA methylation level is correlated with a single-nucleotide polymorphism (SNP). If the SNP occurs close to the CpG (for instance, within a 10 kb window), it is called *cis*-mQTL, otherwise *trans*-mQTL.

Differentially variable cytosines
(DVCs). Cytosines (usually in a CpG context) that exhibit a statistically significant difference in the variance of DNA methylation between two groups of samples, according to some statistical test.

Field defects
Genetic or epigenetic alterations that are thought to predate the development of cancer and that are usually seen in the normal tissue found adjacent to cancer.

Type 1 error rate
The probability of erroneously calling the result of a test significant (positive) when the underlying true hypothesis is the null. It corresponds to the fraction of true negatives that are called positive, also known as the false-positive rate.

Variably methylated regions (VMRs). Contiguous genomic regions where DNA methylation is highly variable relative to a normal 'ground state'. A VMR can be defined for one given sample.

type 1 error rate, was recently used to demonstrate the existence of DNAm field defects in the normal tissue adjacent to breast cancer[55]. Given the growing importance of differential variance, a number of other algorithms[100–102] have been proposed that offer an improved control of the type 1 error rate over the test implemented in EVORA. However, with a stricter control of the type 1 error rate, these other differential variance algorithms may also lack the sensitivity to detect DNAm alterations in precursor cancer lesions[103]. Thus, their application appears limited to other phenotypes (for example, neoplasia or invasive cancer).

An altogether different phenotype for which differential variance has recently been demonstrated to lead to novel insight is age[77]. Specifically, the Breusch–Pagan test for heteroscedasticity was used to identify CpGs whose DNAm variability increases with age, identifying sites that are very different to those making up age-predictive epigenetic clocks[8,104] and that appear to be more relevant for understanding ageing mechanisms[77].

Box 2 | **Differential variability: a novel feature-selection paradigm**

Differential variance

Differential variance (DV) is a novel statistical paradigm for feature selection that has been shown to be valuable in studies seeking DNA methylation (DNAm) field defects, that is, DNAm alterations that appear in the normal cell of origin of epithelial cancers and that become enriched in cancer. A test for DV identifies cytosine–guanine dinucleotides (CpGs) for which the variance in DNAm differs significantly between phenotypes, defining differentially variable cytosines (DVCs). Hypervariable DVCs exhibit increased variance (conversely, hypovariable DVCs exhibit decreased variance) in the disease phenotype compared to normal controls. Depending on the specific test for DV, DVCs typically contain varying numbers of outliers, which occur exclusively or predominantly in one phenotype. DVCs may also exhibit ultra-stable (that is, very low variance) DNAm in one phenotype but not in the other.

Statistical tests for DV

Bartlett's test. This test assumes normality for each of two underlying distributions being compared and is therefore sensitive to outliers. Although it suffers from a high type 1 error rate, its sensitivity to outliers (that is, deviations from normality) makes it an attractive choice because in precursor cancer lesions, DNAm outliers have been shown to be biologically relevant. This test is used in epigenetic variable outliers for risk-prediction analysis (EVORA) and iEVORA and was instrumental to identifying DNAm field defects in cervical and breast cancer (TABLE 1).

The Levene and Brown–Forsythe tests. Levene's test compares the absolute spread of values from the mean in each group, using a one-way ANOVA *F*-test, whereas the Brown–Forsythe test uses the median instead of the mean, rendering it more robust. Both tests are less sensitive to departures from normality than Bartlett's test. Levene's test is implemented in the DiffVar package (TABLE 1).

Breusch–Pagan test. This is a test for heteroscedasticity or differential variability in a response variable (here, DNAm) as a function of an independent variable with continuous values (for example, age). It works by correlating the independent variable with the residuals of a linear regression of the response variable against the independent variable. This test has been used to identify CpGs exhibiting age-associated increases in DNAm variance (see the main text).

EVORA

EVORA is a statistical framework that uses differential variability in DNAm to identify CpGs that exhibit outlier DNAm values in normal cells that are at risk of neoplastic transformation compared to normal cells that are not at risk. For a given risk-marker CpG, this method assumes that DNAm outliers may exhibit stochasticity — that is, they define infrequent events across independent samples. Feature selection using DV is combined with an adaptive index classification algorithm (effectively, a counting scheme for the number of outliers in a sample) to construct a risk score.

As with differential methylation, differential variance may also be defined at the regional level. First, it has been possible to demonstrate that there are genomic regions of increased DNAm variability, so-called variably methylated regions (VMRs)[105], also termed regions of high methylation disorder or entropy[106]. Regions that constitute VMRs in one phenotype (for example, cancer) but not in another (for example, normal tissue) are differentially variable regions (DVRs)[105]. DVR detection is possible using dedicated functions in software packages such as minfi[87] or DMRcate[95], although the implemented differential variance tests are aimed only at controlling the type 1 error rate and may thus be underpowered for detecting epigenetic field defects in cancer studies[55].

Interpreting DNA methylation changes. Beyond cell-type composition[107], observed DNAm alterations could be associated with deregulation of specific genes or signalling pathways in individual cell types[34,108]. Thus, there is a strong rationale for testing the enrichment of identified features for specific gene ontology (GO) terms and signalling pathways. As multiple DMCs or DVCs may map to the same gene, it is critical to adjust for differential representation[109] to avoid spurious over-representation in certain pathways by virtue of a higher probe or CpG density in those genes involved. This adjustment can be done with the gometh/gseameth algorithm[110]. An alternative approach is to assign a DNAm value to a given gene, such as by focusing on the average DNAm within a certain distance of the transcription start site (TSS)[111], and to then identify differentially methylated genes, which can be subsequently fed into popular gene set enrichment analysis (GSEA) methods[112,113]. With a DNAm value assigned to each gene, one may also perform differential methylation analysis at the level of signalling pathways or search for differentially methylated gene modules (called 'EpiMods') within protein–protein interaction (PPI) networks[111]. For instance, such an approach demonstrated that the WNT signalling pathway, a key developmental pathway, is a hot spot of age-associated DNAm deregulation[111].

Integration of DNAm with other types of omic data

There are many factors that limit the interpretability of the DNAm data generated in a typical EWAS[114,115]. Besides cell-type heterogeneity, genetic variation and reverse causation (that is, alterations to measured DNAm levels caused by the phenotype itself) can also cause confounding[18,116]. As a predictor of gene expression, DNAm is also limited and outperformed by chromatin state information encoded by histone modification marks[117,118]. Thus, enhancing interpretability requires integration with other types of omic data, including genotype or gene expression matched to the same samples for which DNAm is available.

Integration of DNAm with genotype. Total heritability of DNAm has been estimated at 20%[76,119], with common SNPs accounting for approximately 37% of this heritability[76]. In line with this, many studies have demonstrated that mQTLs are widespread[76,120,121], accounting

for almost 40% of assayed CpG sites and explaining approximately 20% of the inter-individual variation in DNAm, with environmental effects accounting for the remaining 80%[76]. Thus, adjusting for DNAm variation induced by genetic variation is a common procedure in EWAS, which can be achieved using PCA on the matched genotype data[76,77,122] or directly from DNAm data if no matched genotype information is available[123]. Beyond being a source of confounding, genetically driven DNAm variation provides a useful resource for interrogating the functional role of DNAm variation in disease-associated loci. For example, functional inferences can be made by ascertaining whether disease-associated genetic variants from genome-wide association studies (GWAS) are also mQTLs (and may thus be influencing disease risk partly via epigenetic pathways) or by using genotype as a causal anchor to strengthen causal inference regarding the role of DNAm in mediating pathways to disease[124–126] (BOX 3; FIG. 3A). As a concrete example, genetic variants associated with blood lipid levels were used to demonstrate a causal effect of lipid levels on DNAm in blood, whereas mQTLs associated with lipid-level DMCs in blood excluded an effect in the reverse direction[116]. Such inference can thus help to establish causal directionality in an EWAS of a disease risk factor, determining whether DNAm may mediate that risk.

Integration of DNAm with gene expression. The relationship between DNAm and gene expression is complex. From a modelling perspective, the first challenge is that it is not only the DNAm profile of the gene itself but also the DNAm levels at distal regulatory elements, notably enhancers, that dictate the expression level of a gene. In the context of cancer, distal regulation by DNAm patterns at enhancers appears to account for more of the intertumour expression variation than corresponding DNAm changes at promoters[127]. However, expression variation should be assessed primarily against the normal tissue reference (which is often not done), and adjustment for cell-type heterogeneity is imperative, as enhancers are among the most cell-type-specific regions[108,128]. Also problematic is that most enhancers loop over their nearest genes to target genes much further away, causing uncertainty as to which genes an enhancer may regulate. Although improved statistical methods for linking enhancers to their putative gene targets are emerging[129], these still need further improvement. Focusing on the gene itself, a third challenge is to ascertain which part of a gene's DNAm profile is most predictive of its transcript level, as this may also depend on biological context and is still a matter of debate, with some studies suggesting gene-body methylation levels as being more predictive than the more classical TSS region[130–132]. However, a meta-analysis of human genome-wide methylation, expression and chromatin data has demonstrated that the relationship between gene-body methylation and gene expression is non-monotonic, with the genes expressed at the lowest and highest levels exhibiting the highest levels of gene-body methylation[133]. This meta-analysis is consistent with other studies demonstrating that it is the TSS, first exon and 3′ end that exhibit the strongest monotonic associations[85,134,135]. At the TSS and first exon, the correlation is usually negative, characterized by a highly nonlinear 'L'-shape function: that is, methylated promoters are generally associated with gene silencing, whereas unmethylated promoters associate with both transcribed and untranscribed states[136]. Focusing on a specific predictive region such as the first exon or TSS allows assignment of a DNAm value to each gene, such as by averaging DNAm values for CpGs in this region. The monotonic relation (be it linear or nonlinear) between DNAm and transcription in these regions further facilitates subsequent integration with gene expression or with other gene-level omic data (for example, copy-number variants). Importantly, the procedure of assigning a DNAm value to a gene is a necessary preliminary step for integrative clustering analyses using tools such as iCluster+, which perform joint clustering of samples over a common set of features (usually genes) defined for different data types[137–139].

Other attempts at integration of DNAm and gene expression do not assign a unique DNAm value to a gene; instead, they use information about the spatial shape of the DNAm profile over a gene (and beyond) as a predictor of gene expression[84,85]. Such an approach requires DNAm data at high resolution (for example, WGBS) to then perform unsupervised clustering of gene-based spatial DNAm profiles, typically centred on a 10–30 kb window around the TSS of genes, and subsequently using special distance metrics to quantify the similarity of spatial DNAm profiles[84]. This novel approach identified 4–5 spatially distinct DNAm shapes, each correlating with underexpression or overexpression in *cis*[84], further confirming that DNAm patterns that extend well beyond the 5′ and 3′ ends of a gene are equally informative of gene expression[15,108]. More recently, a supervised version of this spatial clustering method, which uses a random-forest classifier called ME-Class, has been shown to improve the prediction of gene expression, highlighting the importance of the TSS and 3′ end as the most predictive gene regions[85].

System-level integration of DNAm. A powerful system-level integrative approach is to exploit the well-known association of DNAm at regulatory elements with TF binding[140–145] to infer patterns of regulatory activity in development and disease. Although DNAm at regulatory sites has traditionally been viewed as dictating TF binding affinity, the converse (that is, DNAm levels at regulatory sites being a reflection of binding activity) is also frequently observed[115,142]. Furthermore, whereas for most classes of TFs, in which DNAm inhibits or is inversely correlated with binding, there are other classes of TFs (for example, those belonging to the homeodomain, POU and NFAT families) that prefer binding to methylated sequences[143]. Thus, although the relationship between DNAm and TF binding is undoubtedly complex, two recent key observations have helped to spur a number of novel system epigenomics methods for inferring TF binding activity. One key observation is that tissue-specific TFs can be identified as those with enrichment

Differentially variable regions
(DVRs). Contiguous genomic regions containing a statistically significant number of differentially variable cytosines (DVCs). This is different from a variably methylated region (VMR) in that a DVR is derived by comparing a fairly large number of cases and controls.

Gene set enrichment analysis
(GSEA). A widely used statistical procedure to assess whether a derived gene list of interest is enriched for specific biological terms, usually including gene ontologies, signalling pathways, specific transcriptomic signatures or targets of gene regulators.

System epigenomics
An emerging field whereby cellular phenotypes in normal development and disease are modelled as complex systems, using tools from complexity science (for example, dynamical system theory or statistical physics) to understand them.

Box 3 | **Statistical approaches for establishing mediation by DNA methylation**

DNA methylation (DNAm) is a molecular phenotype that is influenced by endogenous and exogenous factors as well as disease processes themselves, and this presents challenges in understanding the correlations between measures of interest. A variety of statistical methods have been applied to dissect causal relationships and to construct causal pathways involving molecular intermediates including DNAm. These methods have been applied to differentially methylated cytosines (DMCs) only and have yet to be extended to consider the mediating role of differentially methylated regions (DMRs).

Exposure–outcome mediation

The most commonly applied approach in epidemiology is a regression-based method originally proposed by Baron and Kenny[199] that aims to distinguish the degree of mediation of an exposure (E) on an outcome (Y) by an intermediate (M). The Sobel test is applied to ascertain whether the effect of E on Y is statistically significant once adjusted for M.

Advantages

- It is simple to administer.
- The proportion of mediation can be quantified.

Disadvantages

- It requires strong assumptions that are often violated when applying it to molecular mediators. These assumptions include (i) that Y and M are continuous and (ii) that there is no measurement error in the mediator.
- This method should be applied only in the context of complete (not partial) mediation, which is usually not the case when considering DNAm.
- Other, more flexible methods have been applied to DNAm data, including linear equations, structural equation models, marginal structural models and G-computation; however, these approaches all require assumptions of no measurement error and no unmeasured confounding, which are violated in analyses involving DNAm.

Causal inference test (CIT)

This popular approach for exploring causal links in DNAm analyses uses genetic variation as a causal anchor. It is analogous to the Baron and Kenny approach in its use of a series of regression analyses to establish mediated effects but uses genotype (G) in place of the exposure (E). This approach has been used to infer the causal effect of methylation quantitative trait loci (mQTLs) on a particular outcome[30].

Advantages

- It avoids confounding and reverse causation in the mediator–outcome relationship by using genotype as a causal anchor.
- It is simple to apply.

Disadvantages

- It relies on a *P*-value to determine the causal effect and does not estimate the magnitude of the mediated effect.
- It is vulnerable to measurement error in the mediator or outcome.
- It cannot differentiate between a mediated effect and a situation in which the genetic variant directly influences the outcome via an alternative biological pathway (pleiotropy).

Mendelian randomization

This form of instrumental variable (IV) analysis makes use of genetic variants that are robustly associated with the exposure (E) or mediator (M) of interest. It can also be applied in the reciprocal direction to evaluate the direction of cause from a postulated outcome (Y) on the apparent exposure or mediator. The assumptions of Mendelian randomization (MR) are detailed at length elsewhere[200]. Its application in the context of DNAm is becoming more widespread[116,201–203], and an automated platform for MR analysis is freely available (http://www.mrbase.org/) to facilitate this (see TABLE 1).

Advantages

- It provides an estimate of the magnitude of the mediated effect.
- It overcomes the issue of measurement error in the mediator because genotype is usually measured accurately.
- It is readily applicable through online tools.

Disadvantages

- It is reliant on the identification of *cis*-mQTLs to tag the differentially methylated site of interest.
- It has low power, which necessitates the use of large sample sizes.
- The potential pleiotropy of genetic variants, although strategies can be adopted to counter this limitation[204,205].

Pleiotropy
A phenomenon that occurs when a genetic variant is associated with multiple traits. Vertical pleiotropy occurs where the traits are all on the same pathway (and is generally less of a problem), whereas horizontal pleiotropy exists where a genetic variant is associated with multiple traits via separate pathways.

Expression quantitative trait loci
(eQTLs). Genes whose expression levels are correlated with single-nucleotide polymorphisms (SNPs). If the SNP occurs near (definitions vary, but it could range from 10 kb to a 1 Mb window centred on the transcription start site) the gene, it is called a *cis*-eQTL; otherwise, it is a *trans*-eQTL.

in unmethylated or relatively hypomethylated binding sites[108]. Although this was demonstrated by integrating WGBS and Encyclopedia of DNA Elements (ENCODE) ChIP–seq data across multiple different cell types[108], other studies have shown that similar inferences are possible with lower resolution Infinium methylation bead arrays[91].

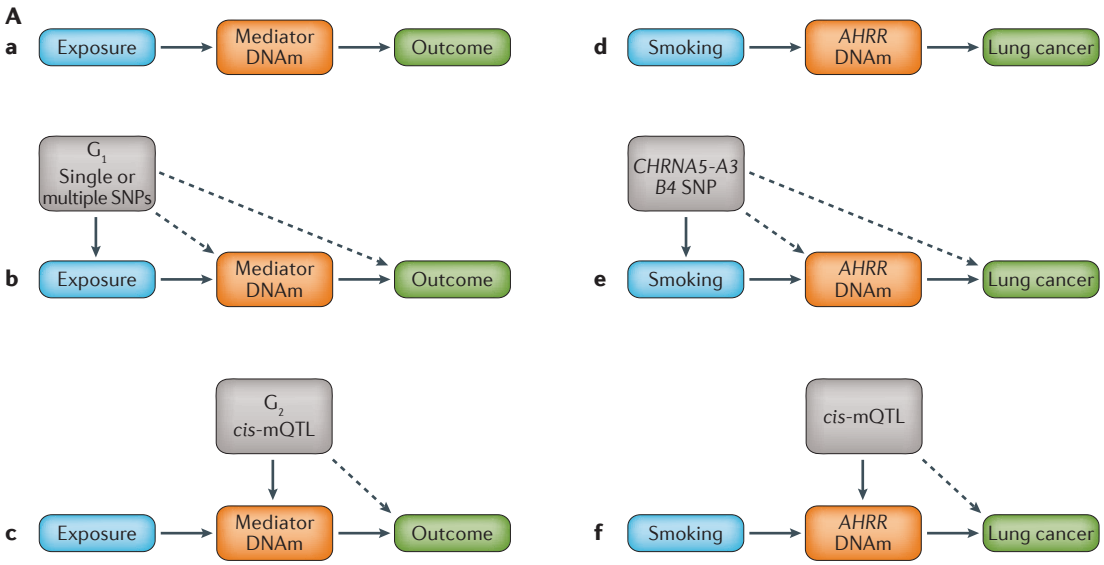
A second key observation is that integration of *trans*-mQTLs with *cis* expression quantitative trait loci (*cis*-eQTLs) can reveal coordinated DNAm alterations at binding sites of a TF whose expression is altered by the SNP, thus providing an important novel paradigm for elucidating the downstream effects of non-coding GWAS SNPs[122] (FIG. 3B).

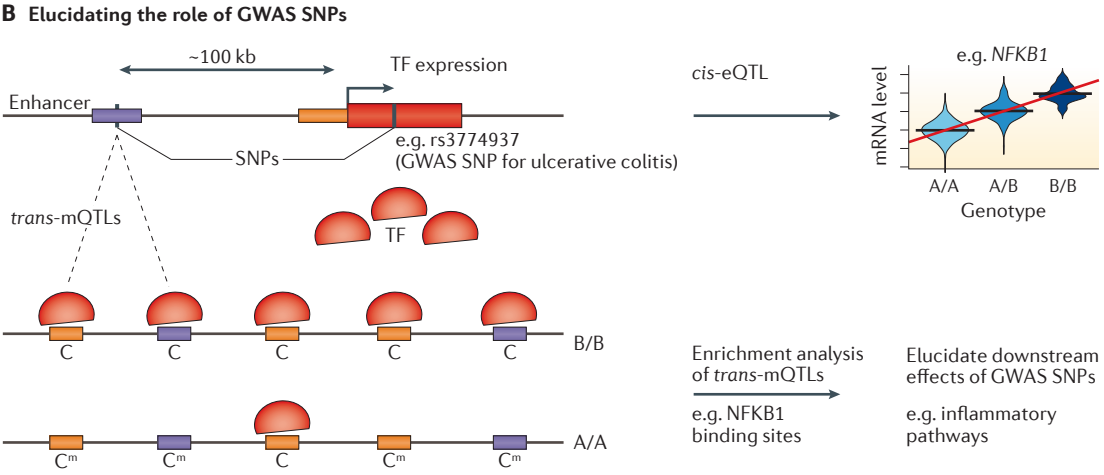
This inverse correlation between DNAm and regulatory-element activity can be exploited by computational tools to infer disrupted regulatory networks associated with disease risk factors[51,91,122,146] and disease itself[127,147,148]. For instance, the enhancer linking by methylation/expression relationships (ELMER) algorithm[147] (TABLE 1) begins by identifying enhancers (annotated by ENCODE and the Roadmap Epigenomics Mapping Consortium[15,149]) whose DNAm levels are altered in cancer. It then uses the matched mRNA expression of putative gene targets to construct cancer-specific enhancer–gene networks. ELMER subsequently uses TF-binding motif enrichment analysis for correlated enhancers and mRNA expression of enriched TFs to identify cancer-specific activated TFs. Other similar approaches, such as tracing enhancer networks using epigenetic traits (TENET)[150] and RegNetDriver[151], have recently been proposed (TABLE 1). RegNetDriver constructs tissue-specific regulatory networks by integrating cell-type-specific open-chromatin data with regulatory elements from ENCODE and RMEC, allowing active regulatory elements in a tissue to be identified. Mapping disease-associated molecular alterations in that tissue onto the corresponding tissue-specific network can reveal which TFs are deregulated in disease[151]. All these tools can lead to important novel hypotheses (for example, ELMER identified RUNX1 as a key TF determining clinical outcome in kidney cancer), as well as novel insights (for example, RegNetDriver revealed that most of the functional alterations of TFs in prostate cancer were associated with DNAm changes but that TF hubs were preferentially altered at the copy-number level). However, obvious limitations remain: the sets of enhancer regions used are usually not cell-type-specific or were generated in unrepresentative cell-line models, while linking genes to enhancers and vice versa is challenging as most enhancers skip their nearest promoter to link to genes that are much further away (contact distances can range from 40 kb to 3 Mb with a median distance of ~180 kb[152,153]). Although tools like ELMER and TENET use correlations between enhancer DNAm and mRNA target expression to hone in on the more likely targets, these correlations are themselves subject to potential confounders such as cell-type heterogeneity.

Another valuable system-level integrative strategy, exemplified by the functional epigenetic modules (FEM) algorithm (TABLE 1), has been to integrate DNAm and gene expression data in the context of a gene function network, for instance a PPI network, to identify hot spots (gene modules) where there is significant epigenetic deregulation in relation to some phenotype of interest[134,154] (FIG. 3C). There are two main reasons why integration of DNAm with a PPI network is meaningful. First, PPI networks encode information about which proteins interact together and which are therefore more likely to be co-expressed as part of a common biological process or signalling pathway. This co-expression is likely to be under epigenetic control and therefore potentially measurable from DNAm patterns at the corresponding genes[111]. Indeed, like gene expression, DNAm also exhibits modularity in the context of a PPI network, whereby promoter DNAm levels of genes whose proteins interact are on average more highly correlated than those of non-interacting proteins[111] (FIG. 3C). Second, using a functional network from the outset and searching for subnetworks where there is simultaneous differential methylation and differential expression can help to identify biological pathways or processes that are epigenetically deregulated, which in turn may lead to novel insight. This is not dissimilar to performing a direct form of GSEA but using a network instead of an external database of biological terms. Similar supervised functional network algorithms have been extensively applied in the gene expression context, leading to important novel insights[155–157]. As an example of the insights gained using FEM, it successfully identified two separate gene modules with the main targets of epigenetic silencing mapping to a target *(HAND2)* and co-activator (*TGFB1I1*) of the progesterone receptor, a key tumour suppressor pathway for which inactivation is thought to contribute causally to the development of endometrial cancer [134,154]. More recently, other algorithms that extend or modify FEM have been proposed[158,159] (TABLE 1). The algorithm 'significance-based modules integrating the transcriptome and epigenome' (SMITE)[160] can identify DNAm-mediated altered cellular states (for example, gene modules) without the need for direct integration with a PPI network, thus allowing a larger gene-space to be explored. In summary, although these methods can substantially improve the interpretation of DNAm changes in EWAS, they are nevertheless limited by the quality of the modelling between methylation and gene expression.

Another set of integrative algorithms are tailored for integrating DNAm data that are generated in conjunction with other data types for the same samples: for instance, this may include mutations, copy-number variants (CNVs), mRNA, microRNAs (miRNAs) and protein expression[161]. Analysing individual data types separately and subsequently correlating resulting clusters has been a popular strategy[162]; however, performing simultaneous inference using all data types together offers, in principle, a much more powerful and unbiased framework in which to identify system-level associations and extract novel biological insight. For instance, simultaneous inference may help to identify genes that are deregulated epigenetically or through CNVs in a mutually exclusive fashion[151,163]. Although many statistical algorithms for multi-omic integrative analyses exist, their application to multi-omic data remains challenging, owing to the high-dimensional nature of the data but also because the effect of confounders on the inference is poorly understood. So far, a joint NMF algorithm was applied to the matched DNAm, mRNA and miRNA expression data sets for ovarian cancer from The Cancer Genome Atlas (TCGA), revealing novel perturbed pathways[164]. An integrative DNAm and mRNA analysis of oestrogen receptor $(ER)^+$ breast cancer used a joint latent variable algorithm, called iCluster/iCluster+[137,139,165], demonstrating that ER^+ breast cancer transcriptomic subtypes differ epigenetically mainly only in terms of the level of DNAm deregulation[138]. Other algorithms and techniques for joint multi-omic matrix factorization analyses are available,

TF hubs
In the context of a regulatory network where edges represent regulatory interactions between transcription factors (TFs) and target genes, those TFs with the largest number of interactions.





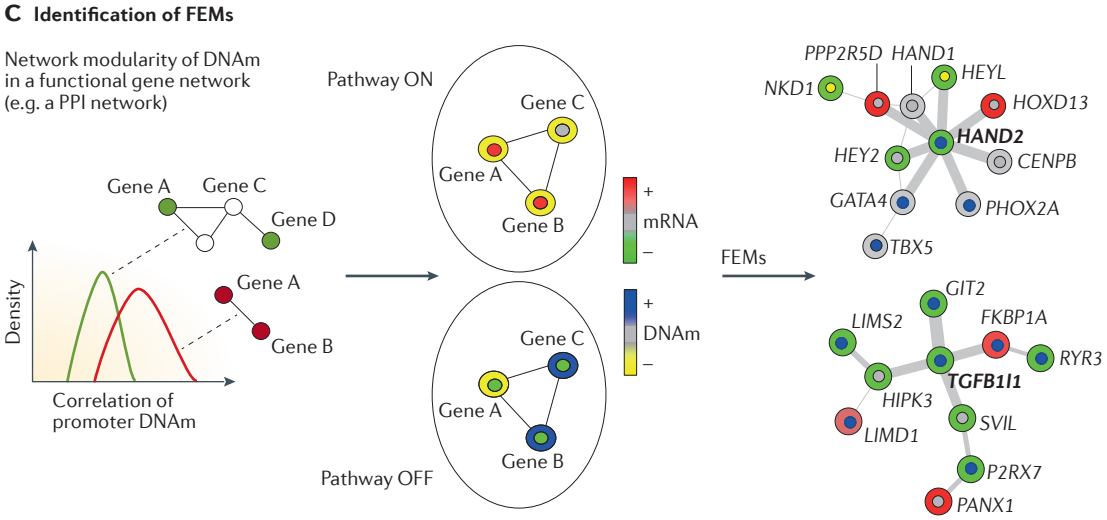


Figure 3 | **Examples of system-level integrative analysis of DNA methylation data. Aa** | To establish causal pathways for observed associations between an exposure, mediator and outcome, genotype can be used as a causal anchor. **Ab** | To strengthen causal inference from exposure to outcome and from exposure to mediator, a genetic variant (G1) or combination of multiple variants that robustly correlate with the exposure can be used. Solid lines represent the established association of the instrumental variable (single-nucleotide polymorphism (SNP)) with the factor for which it is acting as a proxy, and dashed lines represent the relationships being tested in the Mendelian randomization (MR) framework. The association of G1 with the outcome (and mediator) provides evidence of a causal impact of the exposure on these factors. **Ac** | When considering the causal pathway from the mediator (DNA methylation (DNAm)) to the outcome, a second genetic variant (G2) or combination of multiple variants can be used. G2 is a *cis* methylation quantitative trait locus (*cis*-mQTL) that robustly correlates with the DNAm site of interest. Details of the statistical methods to implement this MR approach are further described in BOX 3. G1 and G2 analyses can, if desired, be conducted in entirely different sample sets with causal inference remaining valid. **Ad–f** | An application of this conceptual framework is shown in which the exposure–outcome setting is smoking and lung cancer and the proposed mediator is DNAm at the *AHRR* gene locus[173]. SNPs at the *CHRNA* locus are an established proxy for smoking heaviness and have been used in an MR framework[206]. Their application here can corroborate established evidence for the causal role of smoking in lung cancer as well as interrogate the causal role for methylation as a mediating mechanism. **B** | Integration of DNAm data with matched SNPs and mRNA expression can be used to elucidate the role of genome-wide association study (GWAS) SNPs. For instance, a genetic variant defining a *cis* expression QTL (*cis*-eQTL) for a transcription factor (TF) can be found to be associated with a large number of *trans*-mQTLs. For *cis*-eQTLs associated with increased TF activity, these *trans*-mQTLs exhibit a skew towards hypomethylation (loss of methylation is indicated by the transition C^m (methylated cytosine) to C (cytosine)) and are enriched for binding sites of this TF and for *cis* expression quantitative trait methylation loci (*cis*-eQTMs) defined by the corresponding TF gene targets. An example of a SNP associated with ulcerative colitis illustrates how relevant disrupted pathways can be identified[122]. **C** | Like mRNA expression, promoter DNAm exhibits modularity, that is, stronger correlations between genes that interact in a gene-functional network (for example, a protein–protein interaction (PPI) network). This modularity and the association between promoter DNAm and mRNA expression can be exploited to identify gene modules that are significantly deregulated at both transcriptomic and epigenetic levels. The Functional Epigenetic Modules (FEM) algorithm (TABLE 1) can be used to identify such hotspots of deregulation. A successful application of FEM to endometrial cancer uncovered the gene *HAND2*, a target of the progesterone receptor, which is hypermethylated and silenced in pre-neoplastic lesions and in cancer and which has been shown to drive endometrial carcinogenesis[134,154]. Another gene module is centred around *TGFB1I1* (also known as *HIC5*), a known co-activator of the progesterone receptor. Part **C** is adapted with permission from REF. 207, Springer.

Expression quantitative trait methylation loci (eQTMs). Genes whose expression levels are correlated with the DNA methylation level of a CpG. If the CpG occurs close to the gene (within a 250 kb window), it is called a *cis*-eQTM.

Tensor
A multi-dimensional array with the number of dimensions often called the 'order' or 'rank' of the tensor and for which linear decomposition algorithms are available, analogous to linear matrix factorization algorithms for data matrices. Scalars, vectors and matrices are tensors of order 0, 1 and 2, respectively.

yet they remain largely unexplored in a DNAm context. For instance, joint and individual variation explained (JIVE) (TABLE 1) is a powerful multi-dimensional matrix factorization algorithm that can identify sources of data variation that are common to multiple data types, as well as those that are unique to each data type[166,167]. If multi-omic data are matched across all dimensions (for example, the same genes and samples measured for two different tissues or data types), they can be packed into a multi-dimensional array known as a tensor, for which non-Bayesian (parallel factor analysis (PARAFAC)[168]) and Bayesian (sparse decomposition analysis (SDA)[169]) tensor decomposition algorithms are available (TABLE 1). By approximating the data tensor as a sum of products of simple latent component vectors, one for each data type, these models are readily interpretable, with the Bayesian version less prone to overfitting. A recent study applied SDA to a third-order tensor of expression values defined over 20,000 genes, 845 individuals and 3 tissue types (skin, adipose and lymphoblasts), subsequently correlating the latent components to SNPs and revealing *trans*-eQTL gene networks that were either common or unique to different tissue types, thus helping to delineate tissue-specific functional effects of GWAS SNPs[169]. Thus, tensorial methods should also be particularly suitable for elucidating tissue-specific and tissue-independent mQTLs in EWAS profiling multiple tissue types.

Conclusions and future directions

Recent studies underline the importance of DNAm as a focal point for elucidating and understanding diverse phenomena, including ageing phenotypes[8,77,170–172], functional effects of GWAS variants[30,122], the causal pathways between environmental factors and disease risk[18,51,154,173,174], cell-type heterogeneity and stochasticity[63,68,174,175], cancer evolution and metastasis[66,67,69,176,177] and 3D chromatin architecture[106,178]. Furthermore, they highlight potential downstream applications, including cancer risk prediction[68,179,180], prediction of frailty and all-cause mortality[181–183] and non-invasive detection of cancer and tissue of origin from ctDNA in blood plasma[16,40,184]. For many of these efforts, cell-type heterogeneity and deconvolution will continue to pose challenges. Indeed, most of the algorithms for system-level integration that compute correlations between features do not adjust for cell-type heterogeneity, yet this adjustment is paramount for correct interpretability. Another outstanding challenge is that current algorithms do not allow for the identification of the specific cell type (or types) carrying the DMCs, thus requiring laborious follow-up experimental validation in purified samples. The potential limitation of cell–cell interactions for the accuracy of reference profiles used in reference-based inference also needs to be assessed. Hybrid approaches that generate reference DNAm (or RNA-seq) profiles for different types of single cells in a small number of individuals could be a fruitful strategy for constructing improved reference profiles that are tailored to the tissue of interest[185]. Ultimately, the level of resolution required by cell-type deconvolution strategies also needs to be determined, as epigenetic and DNAm heterogeneity exists right down to the single-cell level[186]. Thus, quantification of functional epigenetic heterogeneity will be a key problem for the future. Related to this, it is also unclear whether DNAm or mRNA expression is better suited for cell-type deconvolution[26,34,37] and whether joint analysis of data types could further improve inference. The generation of gold-standard data sets, artificial and real, is challenging yet absolutely necessary to ensure objective comparisons of existing and upcoming statistical algorithms[48]. In particular, a large comparative and comprehensive analysis of cell-type deconvolution algorithms, including novel semi-reference-free methods, which are particularly amenable for Bayesian treatment, is urgently needed.

Feature selection and inference of causality in EWAS also remain a considerable challenge, even when

Mendelian randomization
A technique to estimate the effect of an exposure on an outcome using genetic variants and instrumental variables for the exposure. This approach can also be applied to assessing mediation.

adjustment for cell-type heterogeneity is possible, as features may still be susceptible to reverse causation or confounding by other unknown factors. Longitudinal prospective studies can avoid some reverse causation effects, and using genotype as a causal anchor via Mendelian randomization can further help to exclude the effects of confounders, but all this does not currently provide a panacea to the problem. Causal inference methods often rely on model assumptions (for example, linearity) that may not hold and that may lead to residual confounding and to wrong or conflicting conclusions. Measurement errors, such as in epidemiological variables, further exacerbate this problem. Thus, as recently proposed[115], causal inference methods may need to incorporate prior biological information from the outset in order to strengthen inference: for instance, guided by recent studies demonstrating that *trans*-mQTLs at TF binding sites could help to delineate the effects of non-coding GWAS SNPs[122], it will be of great interest to extend causal-inference methodology to such multi-locus scenarios. Alternatively, breakthrough experimental techniques that allow single-locus and multi-locus epigenome editing[187] will shed new light on epigenetic function and causality, yet these will also require the development of novel statistical procedures to fully interpret the effects of epigenetic perturbations. Another emerging challenge for feature selection is the presence of stochastic epigenetic perturbations, exemplified by DNAm outliers in normal tissue that predate disease onset and that may be indicators of disease risk (for example, normal tissue at risk of cancer development)[55]. A particular challenge is distinguishing DNAm outliers that mark shifts in the epiallele composition of a tissue (contributing to epigenetic mosaicism) from DNAm outliers driven by technical or other confounders.

More generally, analysing DNAm in conjunction with other epigenetic and functional data promises to improve our understanding of 'system epigenomics'. However, this will require sophisticated statistical modelling, which could benefit from harnessing innovative approaches used in other fields, such as engineering, artificial intelligence and physics. Although the value of advanced machine-learning methods (for example, deep neural networks) is undeniable[15,129,188,189], extracting novel biological insight from them is often limited. Thus, we envisage that phenomenological models inspired or built on physical models[190–193] could capture the right level of complexity to extract and harness useful biological insight. Along these lines, integrative analysis of multi-omic data, potentially at the single-cell level and within the framework of statistical mechanics models[186,191,194–198], may allow construction of epigenetic landscapes as envisaged by Waddington[106,193], which in turn may help to elucidate systems-biological principles underlying diverse phenomena such as tissue homeostasis and cancer.

The rapid growth and availability of statistical tools to integrate, analyse and make inferences about DNAm data are encouraging. Such developments continue to address the challenges faced by the field, and fundamental to these developments is an understanding of both the statistical characteristics of the data being used as well as the biological phenomena they represent.

1. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
2. Ahuja, N., Li, Q., Mohan, A. L., Baylin, S. B. & Issa, J. P. Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res.* **58**, 5489–5494 (1998).
3. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl Acad. Sci. USA* **102**, 10604–10609 (2005).
4. Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–446 (2010).
5. Rakyan, V. K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–439 (2010).
6. Maegawa, S. *et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res.* **20**, 332–340 (2010).
7. Ahuja, N. & Issa, J. P. Aging, methylation and cancer. *Histol. Histopathol.* **15**, 835–842 (2000).
8. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
9. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).
10. Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**, 21–33 (2006).
11. Beck, S. Taking the measure of the methylome. *Nat. Biotechnol.* **28**, 1026–1028 (2010).
12. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).
13. 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**, 389–399 (2016).
14. Stunnenberg, H. G., The International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
15. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
16. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017).
This paper demonstrates how DNAm patterns detected from cell-free DNA in blood plasma can be used to detect cancer and its tissue of origin.
17. Gao, X., Jia, M., Zhang, Y., Breitling, L. P. & Brenner, H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin. Epigenet.* **7**, 113 (2015).
18. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
19. Joehanes, R. *et al.* Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
20. Zwamborn, R. A. *et al.* Prolonged high-fat diet induces gradual and fat depot-specific DNA methylation changes in adult mice. *Sci. Rep.* **7**, 43261 (2017).
21. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Genet.* **13**, 705–719 (2012).
22. Morris, T. J. & Beck, S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* **72**, 3–8 (2015).
23. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
24. Albrecht, F., List, M., Bock, C. & Lengauer, T. DeepBlueR: large-scale epigenomic analysis in R. *Bioinformatics* **33**, 2063–2064 (2017).
25. Liang, L. *et al.* An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature* **520**, 670–674 (2015).
26. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
27. Teschendorff, A. E. *et al.* An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE* **4**, e8274 (2009).
28. Langevin, S. M. *et al.* Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. *Epigenetics* **9**, 884–895 (2014).
29. Koestler, D. C. *et al.* Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1293–1302 (2012).
30. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142–147 (2013).
This paper presents an EWAS demonstrating the dramatic impact adjusting for cell-type heterogeneity can have on the number of discoveries.
31. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformat.* **13**, 86 (2012).
This paper presents a reference-based cell-type deconvolution algorithm for EWAS.
32. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–1439 (2014).
33. Houseman, E. A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformat.* **17**, 259 (2016).
34. Onuchic, V. *et al.* Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell Rep.* **17**, 2075–2086 (2016).

35. Koestler, D. C. *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformat.* **17**, 120 (2016).
36. Chatfield, C. Model uncertainty, data mining and statistical inference. *J. R. Statist. Soc. A* **158**, 419–466 (1995).
37. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
38. Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H. & Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol.* **15**, R50 (2014).
39. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformat.* **18**, 105 (2017).
40. Kang, S. *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol.* **18**, 53 (2017).
41. Zheng, X. *et al.* MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.* **15**, 419 (2014).
42. Zhang, N. *et al.* Predicting tumor purity from methylation microarray data. *Bioinformatics* **31**, 3401–3405 (2015).
43. Zheng, X., Zhang, N., Wu, H. J. & Wu, H. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* **18**, 17 (2017).
44. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
This paper presents SVA, a powerful framework for feature selection in the presence of confounders, including cell-type composition and unknown factors.
45. Leek, J. T. & Storey, J. D. A general framework for multiple testing dependence. *Proc. Natl Acad. Sci. USA* **105**, 18718–18723 (2008).
46. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
47. McGregor, K. *et al.* An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.* **17**, 84 (2016).
48. Zheng, S. C. *et al.* Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat. Methods* **14**, 216–217 (2017).
49. Kaushal, A. *et al.* Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformat.* **18**, 216 (2017).
50. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496–1505 (2011).
51. Teschendorff, A. E. *et al.* Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncol.* **1**, 476–485 (2015).
52. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000).
53. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods* **11**, 309–311 (2014).
54. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 443–445 (2016).
55. Teschendorff, A. E. *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat. Commun.* **7**, 10478 (2016).
56. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).
57. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
58. Lutsik, P. *et al.* MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 55 (2017).
59. Bakulski, K. M. *et al.* DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics* **11**, 354–362 (2016).
60. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
61. Hattab, M. W. *et al.* Correcting for cell-type effects in DNA methylation studies: reference-based method outperforms latent variable approaches in empirical studies. *Genome Biol.* **18**, 24 (2017).
62. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
63. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
64. Singer, Z. S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).
65. Busslinger, M. & Tarakhovsky, A. Epigenetic control of immunity. *Cold Spring Harb. Perspect. Biol.* **6**, a019307 (2014).
66. Landau, D. A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
This paper uses WGBS data to estimate epigenetic clonal heterogeneity in cancer and to show that increased epigenetic heterogeneity is associated with a poor clinical outcome.
67. Li, S. *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–799 (2016).
68. Teschendorff, A. E. *et al.* Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* **4**, 24 (2012).
This paper demonstrates that the risk of an epithelial cancer can be predicted from the DNAm patterns measured in normal cells, years before neoplastic transformation. The detection of DNAm risk markers was only possible using differential variability as a novel feature-selection paradigm in a risk prediction algorithm called EVORA.
69. Li, S. *et al.* Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol.* **15**, 472 (2014).
70. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
71. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformat.* **11**, 587 (2010).
72. Wang, X., Laird, P. W., Hinoue, T., Groshen, S. & Siegmund, K. D. Non-specific filtering of beta-distributed data. *BMC Bioinformat.* **15**, 199 (2014).
73. Zhuang, J., Widschwendter, M. & Teschendorff, A. E. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformat.* **13**, 59 (2012).
74. Dedeurwaerder, S. *et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings Bioinformat.* **15**, 929–941 (2014).
75. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
76. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
77. Slieker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17**, 191 (2016).
This paper demonstrates the importance of differentially variable DNAm patterns in the context of ageing, linking age-associated DVCs to age-associated transcriptional changes. It provides a novel paradigm for understanding the role of age-associated DNAm changes in disease aetiology.
78. Wettenhall, J. M. & Smyth, G. K. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**, 3705–3706 (2004).
79. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
80. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
81. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).
82. Libertini, E. *et al.* Saturation analysis for whole-genome bisulfite sequencing data. *Nat. Biotechnol.* **34**, 691–693 (2016).
83. Libertini, E. *et al.* Information recovery from low coverage whole-genome bisulfite sequencing. *Nat. Commun.* **7**, 11306 (2016).
84. VanderKraats, N. D., Hiken, J. F., Decker, K. F. & Edwards, J. R. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res.* **41**, 6816–6827 (2013).
85. Schlosberg, C. E., VanderKraats, N. D. & Edwards, J. R. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Res.* **45**, 5100–5111 (2017).
86. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).
87. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
88. Irizarry, R. A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18**, 780–790 (2008).
89. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44**, 40–46 (2012).
90. Timp, W. *et al.* Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* **6**, 61 (2014).
91. Yuan, T. *et al.* An integrative multi-scale analysis of the dynamic DNA methylation landscape in aging. *PLoS Genet.* **11**, e1004996 (2015).
92. Vandiver, A. R. *et al.* Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biol.* **16**, 80 (2015).
93. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genet.* **43**, 768–777 (2011).
94. Hansen, K. D. *et al.* Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res.* **24**, 177–184 (2014).
95. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8**, 6 (2015).
96. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**, 2986–2988 (2012).
97. Snedecor, G. W. & Cochran, W. G. *Statistical Methods* (Wiley-Blackwell, 1989).
98. Teschendorff, A. E. & Widschwendter, M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487–1494 (2012).
99. Tian, L. & Tibshirani, R. Adaptive index models for marker-based risk stratification. *Biostatistics* **12**, 68–86 (2011).
100. Phipson, B. & Oshlack, A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol.* **15**, 465 (2014).
101. Wahl, S. *et al.* On the potential of models for location and scale for genome-wide DNA methylation data. *BMC Bioinformat.* **15**, 232 (2014).
102. Ahn, S. & Wang, T. A powerful statistical method for identifying differentially methylated markers in complex diseases. *Pac. Symp. Biocomput.* **2013**, 69–79 (2012).
103. Teschendorff, A. E., Jones, A. & Widschwendter, M. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformat.* **17**, 178 (2016).
104. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
105. Jaffe, A. E., Feinberg, A. P., Irizarry, R. A. & Leek, J. T. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* **13**, 166–178 (2012).
106. Jenkinson, G., Pujadas, E., Goutsias, J. & Feinberg, A. P. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat. Genet.* **49**, 719–729 (2017).

107. Breeze, C. E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep.* **17**, 2137–2150 (2016).
108. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
109. Geeleher, P. *et al.* Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* **29**, 1851–1857 (2013).
110. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
111. West, J., Beck, S., Wang, X. & Teschendorff, A. E. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci. Rep.* **3**, 1630 (2013).
112. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
113. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
114. Birney, E., Smith, G. D. & Greally, J. M. Epigenome-wide association studies and the interpretation of disease -omics. *PLoS Genet.* **12**, e1006105 (2016).
115. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* **18**, 441–451 (2017).
116. Dekkers, K. F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17**, 138 (2016).
117. Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA* **107**, 2926–2931 (2010).
118. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
119. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
120. Banovich, N. E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
121. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
122. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
This paper demonstrates how genetic variants that affect the activity of a transcription factor in *cis* are associated in *trans* with coherent DNAm alteration at its binding sites. This principle provides a new strategy for elucidating the role of non-coding GWAS SNPs.
123. Rahmani, E. *et al.* Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin* **10**, 1 (2017).
124. Relton, C. L. & Davey Smith, G. 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 (2012).
This is paper proposes the use of genotype as a causal anchor to strengthen causal inference in epigenetic studies. It sets out the principle of two-step Mendelian randomization for molecular mediation.
125. Richardson, T. G. *et al.* Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. *Am. J. Hum. Genet.* **101**, 590–602 (2017).
126. Caramaschi, D. *et al.* Exploring a causal role of DNA methylation in the relationship between maternal vitamin B_{12} during pregnancy and child's IQ at age 8, cognitive performance and educational attainment: a two-step Mendelian randomization study. *Hum. Mol. Genet.* **26**, 3001–3013 (2017).
127. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
128. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
129. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
130. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).
131. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2** (Suppl. 1), S4–S11 (2005).
132. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
133. Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462–474 (2012).
134. Jiao, Y., Widschwendter, M. & Teschendorff, A. E. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* **30**, 2360–2366 (2014).
135. Brenet, F. *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS ONE* **6**, e14524 (2011).
136. Walsh, C. P. & Bestor, T. H. Cytosine methylation and mammalian development. *Genes Dev.* **13**, 26–34 (1999).
137. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
138. Gao, Y. *et al.* The integrative epigenomic-transcriptomic landscape of ER positive breast cancer. *Clin. Epigenet.* **7**, 126 (2015).
139. Mo, Q. *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl Acad. Sci. USA* **110**, 4245–4250 (2013).
140. Maurano, M. T. *et al.* Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.* **12**, 1184–1195 (2015).
141. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
142. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
143. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
144. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
145. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
146. Guilhamon, P. *et al.* Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. *Nat. Commun.* **4**, 2166 (2013).
147. Yao, L., Shen, H., Laird, P. W., Farnham, P. J. & Berman, B. P. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 105 (2015).
148. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).
149. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
150. Rhie, S. K. *et al.* Identification of activated enhancers and linked transcription factors in breast, prostate, and kidney tumors by tracing enhancer networks using epigenetic traits. *Epigenetics Chromatin* **9**, 50 (2016).
151. Dhingra, P. *et al.* Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. *Genome Biol.* **18**, 141 (2017).
152. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
153. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
154. Jones, A. *et al.* Role of DNA methylation and epigenetic silencing of *HAND2* in endometrial cancer development. *PLoS Med.* **10**, e1001551 (2013).
This is paper uses a system-level integrative analysis of DNAm data, identifying *HAND2* promoter methylation as a driver event in endometrial carcinogenesis. It presents an example of an epigenetically deregulated gene linking ageing and obesity, the two main risk factors for endometrial cancer.
155. Dutkowski, J. & Ideker, T. Protein networks as logic functions in development and cancer. *PLoS Computat. Biol.* **7**, e1002180 (2011).
156. Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Systems Biol.* **3**, 140 (2007).
157. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385–1389 (2010).
158. Ruan, P., Shen, J., Santella, R. M., Zhou, S. & Wang, S. NEpiC: a network-assisted algorithm for epigenetic studies using mean and variance combined signals. *Nucleic Acids Res.* **44**, e134 (2016).
159. Ma, X., Liu, Z., Zhang, Z., Huang, X. & Tang, W. Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformat.* **18**, 72 (2017).
160. Wijetunga, N. A. *et al.* SMITE: an R/Bioconductor package that identifies network modules by integrating genomic and epigenomic information. *BMC Bioinformat.* **18**, 41 (2017).
161. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
162. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
163. Teschendorff, A. E. *et al.* The multi-omic landscape of transcription factor inactivation in cancer. *Genome Med.* **8**, 89 (2016).
164. Zhang, S. *et al.* Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* **40**, 9379–9391 (2012).
165. Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, e35236 (2012).
166. O'Connell, M. J. & Lock, E. F. R.JIVE for exploration of multi-source molecular data. *Bioinformatics* **32**, 2877–2879 (2016).
167. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Statist.* **7**, 523–542 (2013).
168. Harshman, R. A. & Lundy, M. E. PARAFAC: Parallel factor analysis. *Comput. Stat. Data Anal.* **18**, 39–72 (1994).
169. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**, 1094–1100 (2016).
170. Wang, T. *et al.* Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol.* **18**, 57 (2017).
171. Cole, J. J. *et al.* Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. *Genome Biol.* **18**, 58 (2017).
172. Hahn, O. *et al.* Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism. *Genome Biol.* **18**, 56 (2017).
173. Fasanelli, F. *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).
174. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1757–1764 (2010).
175. Issa, J. P. Epigenetic variation and cellular Darwinism. *Nat. Genet.* **43**, 724–726 (2011).
176. McDonald, O. G. *et al.* Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nat. Genet.* **49**, 367–376 (2017).
177. Zhuang, J. *et al.* The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS Genet.* **8**, e1002517 (2012).
178. Fortin, J. P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
179. Levine, M. E. *et al.* DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging* **7**, 690–700 (2015).

180. Yang, Z. *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* **17**, 205 (2016).
181. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).
182. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 14617 (2017).
183. Breitling, L. P. *et al.* Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. *Clin. Epigenet.* **8**, 21 (2016).
184. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl Acad. Sci. USA* **113**, E1826–E1834 (2016).
185. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355,** eaai8478 (2017).
186. Cheow, L. F. *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* **13**, 833–836 (2016).
187. Stricker, S. H., Koferle, A. & Beck, S. From profiles to function in epigenomics. *Nat. Rev. Genet.* **18**, 51–66 (2017).
188. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
189. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).
190. MacArthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154**, 484–489 (2013).
191. Teschendorff, A. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **8**, 15599 (2017).
192. Teschendorff, A. E. *et al.* The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Computat. Biol.* **10**, e1003709 (2014).
193. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Computat. Biol.* **10**, e1003734 (2014).
194. Mojtahedi, M. *et al.* Cell fate decision as high-dimensional critical state transition. *PLoS Biol.* **14**, e2000640 (2016).
195. Mar, J. C. & Quackenbush, J. Decomposition of gene expression state space trajectories. *PLoS Computat. Biol.* **5**, e1000626 (2009).
196. Teschendorff, A. E., Sollich, P. & Kuehn, R. Signalling entropy: a novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* **67**, 282–293 (2014).
197. Garcia-Ojalvo, J. & Martinez Arias, A. Towards a statistical mechanics of cell fate decisions. *Curr. Opin. Genet. Dev.* **22**, 619–626 (2012).
198. Stumpf, P. S., Ewing, R. & MacArthur, B. D. Single cell pluripotency regulatory networks. *Proteomics* **16**, 2303–2312 (2016).
199. Baron, R. M. & Kenny, D. A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
200. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
201. Mendelson, M. M. *et al.* Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian randomization approach. *PLoS Med.* **14**, e1002215 (2017).
202. Morales, E. *et al.* Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *Int. J. Epidemiol.* **45**, 1644–1655 (2016).
203. Allard, C. *et al.* Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. *Epigenetics* **10**, 342–351 (2015).
204. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
205. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
206. Taylor, A. E. *et al.* Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: the CARTA consortium. *BMJ Open* **4**, e006141 (2014).
207. Teschendorff, A. E. in *Computational and Statistical Epigenomics* (ed. Teschendorff, A. E.) 161–185 (Springer, 2015).
208. Maksimovic, J., Gagnon-Bartsch, J. A., Speed, T. P. & Oshlack, A. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Res.* **43**, e106 (2015).
209. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).
210. Schmidt, F. *et al.* Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* **45**, 54–66 (2017).
211. Hemani, G. *et al.* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. Preprint at *bioRxiv* http://dx.doi.org/10.1101/078972 (2016).
212. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
213. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
214. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
215. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

Author contributions
Both authors contributed to all aspects of manuscript researching, discussion, writing and editing.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
European BLUEPRINT Epigenome Mapping Consortium: http://www.blueprint-epigenome.eu
European Genome–Phenome Archive (EGA): https://www.ebi.ac.uk/ega
Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo
Human Epigenome Atlas: http://www.epigenomeatlas.org
ICGC Data Portal: http://dcc.icgc.org
International Human Epigenome Consortium: http://www.ihec-epigenomes.org
US NIH Roadmap Epigenomics Mapping Consortium: http://www.roadmapepigenomics.org
Genetics of DNA Methylation Consortium: http://www.godmc.org.uk/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF