

## Epigenome-wide association studies for common human diseases

Vardhman K. Rakyan\*, Thomas A. Down†, David J. Balding§ and Stephan Beck||

**Abstract** | Despite the success of genome-wide association studies (GWASs) in identifying loci associated with common diseases, a substantial proportion of the causality remains unexplained. Recent advances in genomic technologies have placed us in a position to initiate large-scale studies of human disease-associated epigenetic variation, specifically variation in DNA methylation. Such epigenome-wide association studies (EWASs) present novel opportunities but also create new challenges that are not encountered in GWASs. We discuss EWAS design, cohort and sample selections, statistical significance and power, confounding factors and follow-up studies. We also discuss how integration of EWASs with GWASs can help to dissect complex GWAS haplotypes for functional analysis.

Elucidating the genetic and non-genetic determinants of human complex diseases represents one of the principal challenges of biomedical research. In recent years, genome-wide association studies (GWASs) have uncovered >800 SNP associations for more than 150 diseases and other traits<sup>1</sup>. Although the complete genetic basis is not yet known for any human complex disease, resequencing of exomes — and ultimately whole genomes — holds promise for identifying most of the causal genetic variations. However, there is now increasing interest in exploring how non-genetic variation, including epigenetic factors, could influence complex disease aetiology<sup>2–4</sup>.

The epigenome of a cell is highly dynamic, being governed by a complex interplay of genetic and environmental factors<sup>5</sup>. Normal cellular function relies on the maintenance of epigenomic homeostasis, which is further highlighted by numerous reported associations between epigenomic perturbations and human diseases, notably cancer<sup>4</sup>. However, most studies of such associations to date have been performed either with inadequate genome coverage (for example, tens to hundreds of loci) but adequate sample size, or with coverage that is closer to being genome-wide (thousands of loci) but inadequate sample size. Consequently, for any human complex disease, we remain unaware of the proportion of phenotypic variation that is attributable to inter-individual epigenomic variation. This problem can only be elucidated by large-scale, systematic epigenomic equivalents of GWASs — epigenome-wide

association studies (EWASs), as first proposed in 2008 (REF. 6). At least for DNA methylation (DNAm), technology is now available that is directly comparable in resolution and throughput to the highly successful GWAS chips that allow genotyping of around 500,000 (500K) SNPs.

But how does one conduct an EWAS? In addition to considerations that are common to both GWASs and EWASs (for example, adequate technology and sample size), the design of EWASs has specific considerations with respect to sample selection. DNAm patterns are specific to tissues and developmental stages, and they also change over time. Furthermore, EWAS associations can be causal as well as consequential for the phenotype in question — a difference from GWASs that presents considerable challenges. Here, we discuss these considerations in the context of designing and analysing an effective EWAS, keeping in mind that EWASs are likely to evolve, much like GWASs did, as information and experience accumulate.

### Epigenetic variation and complex disease

**Types of epigenetic information.** Epigenetic information in mammals can be transmitted in multiple forms<sup>5</sup>, including mitotically stable DNAm, post-translational modifications of histone proteins and non-coding RNAs (ncRNAs). For DNAm, the predominant form is methylation of cytosines in the context of cytosine-guanine dinucleotides (CpGs). However, recent results suggest that CpH methylation (where H = C/A/T)

\*Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary, University of London, 4 Newark Street, London E1 2AT, UK.

†Wellcome Trust Cancer Research UK Gurdon Institute and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK.

§Genetics Institute, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK.

||UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK.

Correspondence to V.K.R., D.J.B. and S.B.

e-mails: [v.rakyan@qmul.ac.uk](mailto:v.rakyan@qmul.ac.uk);

[d.balding@ucl.ac.uk](mailto:d.balding@ucl.ac.uk);

[s.beck@ucl.ac.uk](mailto:s.beck@ucl.ac.uk).

doi:10.1038/nrg3000

Published online 12 July 2011

**Genome-wide association studies (GWAS).** These are genome-wide studies that are designed to identify genetic associations with an observable trait, disease or condition, such as diabetes.

**Exome**  
The part of a genome that encodes exons for translation into proteins.

may be more common than previously appreciated<sup>7,8</sup>. Catalysed by the ten-eleven translocation (TET) methylcytosine dioxygenases, 5-hydroxymethylation<sup>9,10</sup> of cytosines (hmC) is yet another form of DNAm. Although details are still unclear, increasing evidence suggests a role of hmC in gene regulation and differentiation<sup>11</sup>. Histone modifications include, to name but a few, mono-, di- or trimethylation, acetylation and citrullination of one or more amino acids in the amino-terminal tails of core histones<sup>5</sup>. More recently, it has been discovered that ncRNAs can self-propagate and be transmitted independently of the underlying DNA; in other words, they can ‘epigenetically’ transmit regulatory information<sup>12,13</sup>. Such ncRNAs include short microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs) and large intergenic non-coding RNAs (lincRNAs), among others<sup>12</sup>.

**Epigenetic variation in health and disease.** The full range of epigenetic marks is currently unknown but is potentially enormous, considering that the diploid human epigenome contains >10<sup>8</sup> Cs (of which >10<sup>7</sup> are CpGs) and >10<sup>8</sup> histone tails that can all potentially vary. The most studied epigenetic mark is DNAm, and BOX 1 discusses the most common features and contexts in which DNAm varies. DNAm variation at a single CpG site is known as a methylation variable position (MVP), which can be considered as the epigenetic equivalent of a SNP<sup>14</sup>. Very rarely, CpGs on only one of the two strands of DNA per allele are methylated. This is known as hemimethylation, and it probably reflects post-replication lag in DNAm maintenance in proliferating cells. If DNAm is altered at multiple adjacent CpG sites, this is referred to as a differentially methylated region (DMR). DMRs vary considerably in length:

### Box 1 | Definition of features known to vary in DNA methylation

This rapidly increasing list of features is not meant to be complete but intends to show the key loci and contexts in which DNA methylation (DNAm) is known to vary.

**Methylation variable position (MVP).** A CpG site that shows differential methylation — for example, between different disease states, as illustrated in the figure. Given recent findings on non-CpG methylation, potentially all Cs could be MVPs.

**Differentially methylated region (DMR).** A region of the genome at which multiple adjacent CpG sites show differential methylation. DMRs can occur in many different contexts, such as:

- iDMR — imprinting-specific differentially methylated region
- tDMR — tissue-specific differentially methylated region
- rDMR — reprogramming-specific differentially methylated region
- cDMR — cancer-specific differentially methylated region
- aDMR — ageing-specific differentially methylated region.

**Variably methylated region (VMR).** These regions are defined by increased variability rather than gain or loss of DNAm.

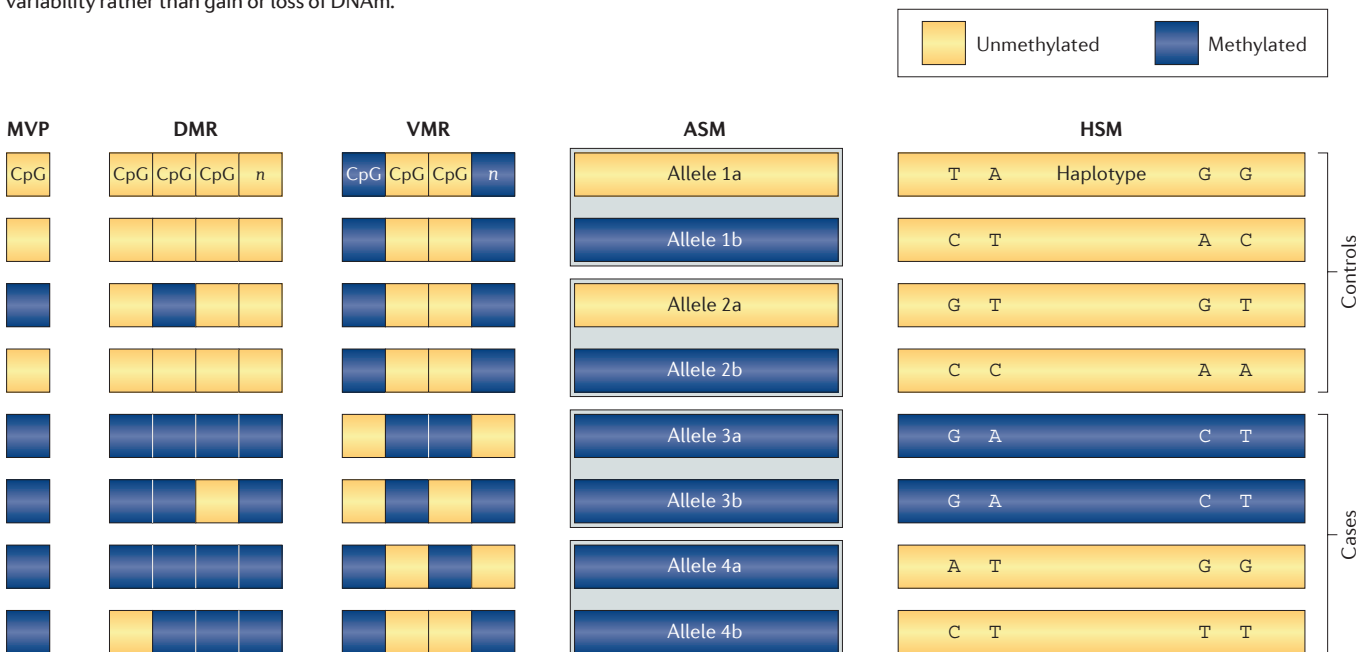
**Allele-specific methylation (ASM).** These are positions or regions that vary in DNAm depending on the parent-of-origin, the presence of a polymorphism or as a result of a stochastic event.

**Haplotype-specific methylation (HSM).** This is a differentially methylated region that is defined by a set of co-inherited SNPs (a haplotype).

**CpG islands (CGIs).** These are regions enriched for CpG sites. Most CGIs are unmethylated in all cell types.

**CGI shores.** These are regions immediately adjacent to CGIs and display higher variation in DNAm than CGIs despite their lower density of CpG sites.

The figure shows different types of DNAm variation that can be identified with epigenome-wide association studies. The notation *n* is used to indicate the variable size of the regions shown. For the purpose of this simplified illustration, the cases and controls are assumed to have methylated or unmethylated CpG states only. Real samples will contain populations of different cells and hence display much more heterogeneous methylation levels across the full dynamic range between 0% and 100%.



**Epigenome**

The complete collection of epigenetic marks, such as DNA methylation and histone modifications, and other molecules that can transmit epigenetic information, such as non-coding RNAs, that exist in a cell at any given point in time.

**Core histones**

The proteins that form the nucleosome, which is composed of two copies each of the histones 2A, 2B, 3 and 4. Together, they form a histone octamer around which 147 bases of genomic DNA are wrapped.

**Core promoters**

Regions upstream and downstream of the transcriptional start site (TSS), typically defined as the interval -60 to +40 bases from the TSS.

**CpG islands**

(CGIs). Regions of the genome (typically 500 bp–2 kb) that contain a higher than expected frequency of CpG sites. CGIs are frequently unmethylated and found near promoter regions.

**Imprinted**

This term refers to genes that are expressed in a parent-of-origin-specific manner.

**Loss-of-imprinting**

(LOI). Parental imprinting results in the epigenetic silencing of one allele of a gene owing to its parental origin. Aberrant disruption of imprinting leads to both alleles being expressed; that is, loss-of-imprinting.

**Satellite DNA**

A type of non-coding, repetitive DNA that is a component of functional centromeres and the main structural constituent of heterochromatin.

**Methylation quantitative trait loci**

(methQTLs). DNA variants that influence the DNA methylation state either in *cis* or in *trans*.

**Allele-specific methylation**

(ASM). The presence of DNA methylation on only one of the two alleles present in a cell. This could be due to parental imprinting, random methylation of one allele or genetic effects.

they are typically <1 kb, but they can exceed 1 Mb<sup>15</sup>. Until recently, MVPs and DMRs were mostly studied in the context of core promoters, CpG islands (CGIs) and imprinted differentially methylated regions (iDMRs); however, it is becoming increasingly clear that DNAm is highly dynamic, even outside such regions. For example, a recent study found that tissue- and cancer-specific DMRs preferentially occur in regions adjacent to CGIs — so-called CGI shores<sup>16</sup>. DNAm also has a key role in silencing repeat elements, which may also have an impact on disease aetiology<sup>17,18</sup>.

The role of DNAm variation in complex disease has mainly been explored in the context of cancer, in what may be considered as early EWASs. Findings from these studies have been extensively discussed<sup>4,19</sup>, the key general conclusions being that tumour development is associated with gain of DNAm at CGIs, loss-of-imprinting and epigenetic remodelling of repeat elements, particularly loss of DNAm at satellite DNA<sup>20,21</sup>. For non-malignant, common complex diseases, such as diabetes or autoimmunity, the epigenetic component is only just beginning to be investigated. Observations that support the involvement of an epigenetic component in these diseases include the following. First, monozygotic twin concordance for any complex disease is almost never 100%. Recent small-scale EWASs of monozygotic twins who are discordant for systemic lupus erythematosus<sup>22</sup> and autism-spectrum disorders<sup>23</sup> have found disease-associated epigenetic differences within monozygotic pairs. Second, the incidence of several complex diseases — such as type 1 diabetes<sup>24</sup> — is rising in the general population and is frequently altered in migrant populations, suggesting a role for non-genetic factors. Third, epidemiological evidence suggests that a suboptimal *in utero* or early childhood environment can have an impact on disease outcomes (such as type 2 diabetes) in adulthood, a phenomenon termed 'developmental reprogramming' (REF. 25). Currently, the prime candidate for the molecular memory of the *in utero* environment is epigenetic modifications, including DNAm<sup>26–28</sup>.

**Epigenetic variation as a consequence or cause of disease.** As mentioned above, epigenetic variation can be causal for disease or can arise as a consequence of disease. Epigenetic variation could arise, either directly or indirectly, as a consequence of disease — examples of this could include long-term alterations in immune-related cells in autoimmune disorders, altered metabolic regulation in type 2 diabetes or somatic-mutation-induced epigenetic alterations in cancer. However, distinguishing this from epigenetic variation that is causative of the disease process is not straightforward (as we discuss in greater detail below), but is nevertheless crucial; this is because it will help to elucidate the functional role of the disease-associated variation and its potential utility in terms of diagnostics or therapeutics. A key step towards achieving this goal is to determine whether the variation is present prior to any overt signs of disease. In this regard, it is useful to consider how such epigenetic variation could arise

prior to disease. First, it could be inherited and hence be present in all tissues including the germ line (that is, transgenerational epigenetic inheritance), although the extent of this phenomenon is not fully known. Second, it could arise stochastically and be present soma-wide if it happens in early (for example, *in utero*) development<sup>29,30</sup>, or it could be limited to one or a few tissues<sup>31,32</sup> if it were to happen postnatally or during adult life. Third, it could be environmentally induced, either by adult lifestyle-related factors, such as diet or smoking<sup>33</sup>, or even *in utero*; that is, developmental reprogramming (described above).

It is also possible that the underlying genotype influences epigenetic variation, as recently demonstrated by several studies<sup>34–39</sup>. Loci harbouring genetic variants that influence methylation state have been termed methylation quantitative trait loci (methQTLs)<sup>34</sup>. In most methQTLs, the correlations with *cis*-genotype are the most pronounced ones. There is some evidence that genetic variation can also influence epigenetic states in *trans*, but this does not seem to be as prevalent as *cis*-effects<sup>38</sup>. Also, it is important to note that, in most of these previous studies, the true causative genetic variant was not unequivocally identified, and most methQTLs did not demonstrate a strict one-to-one relationship between the *cis*-genotype and the epigenotype; rather, a specified genotype generates an increased probability of methylation. Feinberg and Irizarry<sup>2</sup> have recently argued for the existence of genetic variants in mouse and human genomes that do not change the mean phenotype but rather the variability of phenotype; this could be mediated epigenetically through variably methylated regions (VMRs, see BOX 1). The existence of methQTLs provides a strong argument for integrated GWASs and EWASs to uncover genotypes that exert their function through epigenetic variation (discussed later).

These methQTLs can also affect allele-specific methylation (ASM, see BOX 1). In this context, the steady-state methylation levels differ across the two alleles within the same cell. However, ASM can also occur in the absence of any specific genotype–epigenotype correlations. For example, parental imprinting, X-chromosome inactivation and random monoallelic methylation of one allele are all instances of ASM that are not caused by differences in the underlying genotype between methylated and unmethylated alleles.

Finally, it is also worth considering the possibility that, in some cases, disease-associated epigenetic variation could arise prior to disease onset but may not be causative for the disease per se. This type of epiphenomenon could be due to confounding, in which an environmental factor (such as smoking) or a genetic variant induces both aberrant epigenetic states and disease.

These potential relationships between epigenetic variation and complex disease have important implications for the design and analysis of EWASs. First, they will determine the most relevant tissue and cell types to be sampled. Second, reverse causation and confounding are particular issues for EWAS design. Despite the considerable evidence of epigenetic perturbations in cancer<sup>4</sup> and emerging evidence in other non-malignant

## Reverse causation

Refers to an association between A and B that is due to B causing A rather than the presumed A causing B.

## Methylation-sensitive restriction enzyme digestion

Procedure that cleaves dsDNA depending on the methylation status of the enzyme's recognition site. Some enzymes only cleave when the recognition site is methylated and others only when the site is unmethylated.

## Affinity enrichment

In this context, this term refers to a procedure to enrich methylated DNA fragments from a pool of methylated and unmethylated fragments using affinity reagents such as antibodies against 5-methylcytosine or other methyl-binding proteins.

## RRBS

(Reduced representation bisulphite sequencing). A procedure for single base resolution methylation analysis using bisulphite DNA sequencing of a representative part of a genome, typically 5–10%.

## Box 2 | Profiling technologies for epigenome-wide association studies

Lack of suitable technology has been a major bottleneck for epigenome-wide association studies (EWASs) in the past. Fortunately, this is no longer the case and a variety of both array- and sequencing-based methods are now readily available. As these have already been extensively reviewed<sup>44,47,80</sup> and benchmarked<sup>45,46,81,82</sup>, they are only briefly described here along with some additional technologies that may also be suitable for EWASs as guidance for the variety of choices available.

### Array-based technologies

- **CHARM (comprehensive high-throughput relative methylation)**<sup>83</sup>. This technique uses methylation-sensitive restriction enzymes.
- **Infinium**<sup>84</sup>. This assay uses two different bead types (for methylated and unmethylated DNA) to detect CpG methylation of bisulphite treated DNA. It also uses chemical conversion of DNA.

### Technologies that can be used in conjunction with arrays (chip) or sequencing (seq)

- **HELP-chip/seq (HpaII tiny fragment enrichment by ligation-mediated PCR combined with arrays or sequencing)**<sup>85</sup>. This technique uses methylation-sensitive restriction enzymes.
- **MethylCap-chip/seq (methyl capture using the methyl binding domain of protein MeCP2 combined with arrays or sequencing)**<sup>86</sup>. This technique uses affinity enrichment.
- **MBD-chip/seq (methyl capture using complex of methyl binding proteins MBD2 and MBD3L1 combined with arrays or sequencing)**<sup>87,88</sup>. This technique uses affinity enrichment.
- **MeDIP-chip/seq (methylated DNA immunoprecipitation with antibody against 5-methylcytosine combined with arrays or sequencing)**<sup>89,90</sup>. This technique uses affinity enrichment.

### Sequencing-based technologies

- **Whole-genome BS-seq (bisulphite sequencing)**<sup>8</sup>. This technique uses chemical conversion of DNA.
- **RRBS (reduced representation bisulphite sequencing)**<sup>91</sup>. This technique uses chemical conversion of DNA.

### Choice of profiling technologies

Of these, the BS-seq approach — bisulphite conversion of randomly fragmented DNA followed by sequencing — provides the highest level of coverage and resolution, negligible bias towards CpG dense regions and a direct readout of non-CpG methylation<sup>92,93</sup>. Like all methods based on bisulphite conversion, BS-seq is not capable of distinguishing between methylated and hydroxymethylated cytosine bases<sup>94</sup>. Except for the reduced representation method (RRBS), which provides 5–10% genome coverage, whole-genome BS-seq is currently too expensive for EWAS profiling, although costs keep falling rapidly. Affinity-based enrichment methods such as MeDIP-, MethylCap- and MBD-chip/seq are more economical and highly automatable<sup>95</sup> but are less quantitative and do not provide single-base resolution. In our view, the recently released Infinium 450K BeadArrays seem well suited for EWAS profiling with respect to throughput, cost, resolution and accuracy. However, like other non-sequencing-based methods, the readout of this assay is susceptible to certain polymorphisms that were not known or considered at the time the array was designed.

Of course, the trade-off with all of these methods is that many CpG sites are not profiled. As there is no epigenomic equivalent of the HapMap project, which helped to elucidate some of the genetic variation in the human genome<sup>77,78</sup>, we are not aware of the level of normal epigenetic variation that exists in human populations or even which sites are the most relevant for disease aetiology. A true understanding of complex-disease epigenomics will therefore only be realized when whole-genome methods become more affordable, possibly using techniques such as nanopore<sup>96</sup> and single-molecule real-time<sup>97</sup> sequencing: methods that are currently being developed. These will allow direct (that is, no bisulphite, restriction or enrichment modifications required) and simultaneous determination of DNA methylation, DNA hydroxymethylation and DNA sequence in a single reaction.

diseases<sup>22,23,40–42</sup>, none of these studies has been able to conclusively distinguish causal from consequential epigenetic variants: a problem that has long been recognized<sup>43</sup>. Although any EWAS association with disease is potentially an advance, being able to identify the direction of causality will greatly aid in determining the usefulness of epigenetic variation as, for example, a marker of disease progression, a target for reversal by treatment with an epi-drug (that is, a drug that has an effect on the epigenome), or a measure of drug response by monitoring the kinetics of drug-induced epigenetic changes.

### Profiling epigenetic variation

One of the major developments that enabled large-scale GWASs was the introduction of powerful but affordable genetic profiling technologies, in particular SNP arrays. Only recently have epigenomic profiling

technologies reached the stage at which large-scale EWASs are becoming feasible. For such studies to be possible, the mark or molecule must be stable, amenable to high-throughput analysis and easily accessible in routine clinical samples. Automatable whole-genome profiling methods must also be available. Currently, DNAm (and specifically CpG methylation) is the most suitable mark for EWASs. Other epigenetic marks are by just as important as DNAm (or more so) but are, as of yet, neither as easily accessible in clinical specimens nor as amenable to high-throughput processing. In addition, there are numerous well-established correlations between different epigenetic marks and hence profiling DNAm can, albeit indirectly, provide information about histone modification states and RNA dynamics<sup>5</sup>.

In principle, sequencing- and array-based profiling technologies can be used for EWASs. The most common






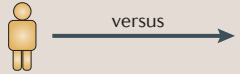
	Key advantage	Key disadvantage
<b>Case versus control (singletons)</b> 	Many cohorts exist	Cannot easily control for environmental and genetic confounders
<b>Families</b> 	Can study potential inheritance	Few large cohorts of this type exist
<b>Disease-discordant monozygotic twins</b> 	Can control for genetics	Few large cohorts of this type exist
<b>Prospectively sampled, longitudinal</b> 	Can establish causality	Slow and difficult to establish

Figure 1 | The different types of sample cohorts that could be used in an epigenome-wide association study. Refer to the main text for a full discussion.

platforms of these two technologies have been extensively reviewed<sup>44</sup> and independently benchmarked<sup>45,46</sup> and are listed in BOX 2. As is typical for this type of study, the choice comes down to balancing coverage, resolution, accuracy, specificity, throughput and cost<sup>47</sup>. Ultimately, sequencing-based technologies are likely to prevail, but array-based methods (such as those used for GWASs) are, in our view, currently the most suitable methods for EWASs. As described in BOX 2, there are options for custom and off-the-shelf platforms covering the choices described above. Of these, the recently released Illumina 450K Infinium Methylation BeadChip seems to be, in our opinion, the most promising for the first wave of EWASs, offering a good balance of genome-wide coverage (>450K CpG sites), resolution (single base pair) and throughput (12 samples per chip and up to 96 samples per run).

### Study designs for EWASs

In this section, we discuss the most informative study designs for EWASs with respect to types of study subjects and addressing the issue of reverse causation. FIGURE 1 illustrates some of the advantages and disadvantages of the four examples discussed.

**Retrospective (case-control).** The most commonly used GWAS design involves unrelated individuals recruited on the basis of their phenotype (for example, cases and controls). Many case-control samples are already available, in some cases with genotype and expression data that can be integrated with epigenomic data. However, a retrospective study cannot determine whether the identified epigenetic variants are due to disease-associated genetic differences, post-disease processes or disease-associated drug interventions. Early examples of using

case-control studies to identify associations between epigenetic variation and clinically relevant phenotypes have included studies on metabolic dysfunction<sup>48</sup> and treatment with tamoxifen<sup>49</sup>.

**Parent-offspring pairs.** These could be useful in EWASs that aim to identify transgenerational transmission of epigenetic marks (BOX 3). It has recently been demonstrated that feeding  $F_0$  generation male mice either a high-fat or low-protein diet from weaning to the time of mating results in  $F_1$  offspring with altered metabolic phenotypes<sup>28,50</sup>. Given that sperm pass on very little, if any, cytoplasmic material to the offspring, these examples suggest the transgenerational transmission of epigenetic variants induced by the suboptimal diet of the  $F_0$  males. A similar strategy using epigenomic profiling of parent-offspring trios could be used in humans. For example, if there is evidence to suggest that paternal environment influences phenotypic outcomes in the offspring, one could perform integrated epigenomic and genomic profiling in the offspring to identify altered epigenetic variants. The genetic information could then be used to eliminate the possibility that genetic modifiers are causing the epigenetic variation. Such study designs will need to use profiling methods that are able to detect allele-specific differences, will need to be adequately powered and will need to have reliable measures of parental environmental exposures.

**Monozygotic twins.** Monozygotic twins who are discordant for a disease of interest represent a useful resource for EWASs, as any identified disease-associated epigenetic variant cannot be caused by germline genetic variation<sup>32,51</sup>. However, unless the twins are recruited longitudinally, which is rarely possible, these studies cannot be used to distinguish between cause and consequence for the reasons discussed earlier. Recruiting large numbers of discordant monozygotic twins for a well-powered study is a potential problem, but some large twin resources are available (see Further information).

**Longitudinal cohorts.** Longitudinal cohort designs follow initially disease-free people (ideally from birth) over the course of many years, recording disease events and other phenotypic changes and taking biological samples. They are expensive to establish, but many such studies are already underway, some of which involve appropriate tissues for EWASs (see Further information). For example, the British 1946 birth cohort<sup>52</sup> offers samples and data spanning 65 years (so far) for over 5,000 individuals. Two major advantages of such studies, compared with many case-control designs, are the avoidance of confounding due to differences in the recruitment of cases and controls and of bias due to case-control differences in the measurement of risk factors. Longitudinal studies can also be invaluable for establishing the temporal origins and stability of disease-associated epigenetic variation, thereby helping to distinguish causal epigenetic variants from consequential ones. If environmental influences are also recorded, it may be possible to relate these to epigenetic changes.

**Box 3 | Transgenerational epigenetic inheritance**

In mammals, epigenetic states are extensively reprogrammed between generations, and this is associated with the reinstatement of the pluripotent state that exists in early development. However, a few studies have shown that epigenetic states are occasionally not completely reprogrammed, resulting in the transgenerational transmission of epigenetic states. The strongest evidence for this phenomenon in mammals comes from various mouse models such as  $A^y$ , and  $Axin^{Fl}$  (REFS 29,30). In these models, the characteristic phenotype is associated with DNA methylation variation at the relevant locus. Interestingly, these states are not always completely reprogrammed between generations, thereby resulting in the range of phenotypes in the offspring being influenced by the phenotype of the parent, even in the absence of genetic heterogeneity. Establishing transgenerational epigenetic inheritance in humans is a far more challenging task, as the outbred nature of human populations means that it is difficult to distinguish true epigenetic inheritance from the inheritance of genetic variants that determine variable epigenetic states. Nevertheless, several reports suggest that transgenerational epigenetic inheritance in humans may occur. If this is true, then we may need to reconsider whether some estimates of heritability are confounded by transgenerational epigenetic inheritance. For example, a specific epigenetic state may be induced in the germ line by environmental factors, such as diet, and these states are passed on to the next generation, ultimately influencing phenotypic outcomes<sup>98</sup>. Indeed, it has recently been demonstrated in rats that a high-fat diet in fathers alters  $\beta$ -islet function in the daughters<sup>28</sup>. The true extent of this phenomenon is expected to become clearer in the coming years.

Longitudinal cohorts of disease-discordant monozygotic twins would convey the additional advantage of ruling out genetic influences on disease-associated epigenetic variation, but such cohorts are rarely available for EWASs of common diseases. A compromise two-phase study design is discussed below, which involves a disease-discordant monozygotic-twin cohort for the discovery phase and a different longitudinal cohort for the replication phase.

**Choice of tissue for EWASs**

In GWASs, most tissue types are suitable for identifying germline genetic variation and DNA extracted from patient blood- or blood-cell-derived cell lines is usually used. However, disease-associated epigenetic variation can be tissue-specific. As most EWASs use live individuals, DNA samples can only be easily accessed from certain sources, such as blood, buccals, saliva, hair follicles, urine and faeces. Blood and blood subtypes, for instance, are relevant for autoimmune diseases or blood-based cancers, and any tissue will suffice if the epigenetic variant is present soma-wide (as will be the case if induced during developmental reprogramming in early embryogenesis).

However, for many diseases, alternative tissue sources need to be explored. These could include assaying cell-free serum DNA — comprising DNA from proliferating cells that is shed into the blood (as happens for most cancers) — or post-mortem DNA, although the latter is a less suitable choice if the aim is to establish causality. In fact, until epigenomic profiling can be routinely performed in a non-invasive manner (for example, through imaging techniques<sup>53</sup>) and/or using small tissue biopsies<sup>54</sup>, it will remain challenging to perform effective EWASs for brain-based and certain other diseases.

Another important issue is tissue heterogeneity. All tissues are composed of multiple cell types (for example, blood contains >50 distinct cell types). If the

disease-associated variation is restricted to a certain cell type that represents only a small proportion of the tissue sampled, then the variation may not be detected. The disease state itself can also alter the composition of cell types in a tissue (for example, inflamed tissue will have a slightly different composition of cell types from non-inflamed tissue). Hence measured epigenetic differences between cases and controls may only reflect differences in cell-type composition and not true epigenetic differences.

Finally, blood-spot (or Guthrie) cards are another valuable source of DNA. These are routinely created in many developed countries immediately after birth using either cord- or heel-prick blood. Biobanks that include DNA and possibly other tissue, as well as phenotypic information, have been set up in several countries (see Further information).

**Examples of EWAS design**

There is not a single EWAS design that will suit all purposes; rather, the most suitable design depends on the required outcome. This is best demonstrated in the form of two hypothetical examples from the many possible EWAS designs that could be conducted.

**An EWAS for disease-risk epigenetic markers.** Let us assume that we are interested in identifying DNAm variants that arise prior to the onset of an autoimmune disease. We could start by performing genome-wide DNAm analysis of monozygotic twins who are discordant for the disease to identify disease-associated MVPs in immune-effector cells (that is, a disease-relevant blood-cell subset) that cannot be due to genetic variation. We could then take these MVPs and assay them in the same type of immune-effector cells from a prospective cohort to look at DNAm at these sites in unrelated individuals who were sampled both before and after disease onset. Any MVPs that can be validated prior to disease onset are then candidate causal variations and cannot be attributed to post-disease effects, such as long-term medication or immune-related effects. Key follow-up studies could include correlation with gene expression and other epigenetic marks to investigate the affected pathways. Overall, this EWAS design combines analysis of a disease-relevant tissue from two independent cohorts that allow for discovery and validation of MVPs and elimination of various confounding factors.

**An EWAS for drug-response epigenetic markers.** Several cancer studies have identified epigenetic variants that can potentially be used to monitor disease progression and even response to treatment<sup>4</sup>. Some of these variants were detected by assaying DNA shed by the primary tumour into the patient's serum, hence providing a relatively straightforward means of assessing progression<sup>55</sup>. An EWAS could also measure the DNAm state in serum from singleton patients who suffer from a specified form of cancer prior to, during and following drug treatment. This could potentially identify epigenetic markers that predict the best response to treatment in real time. The root cause of the cancer-associated epigenetic variants (that is, genetic or environmental) need not be known,

nor would the primary tumour need to be directly analysed, for the variant to be an effective measure of progression or response.

### Statistical considerations for EWASs

**Sample size and power.** In 2005, just as the GWAS wave was about to break, Wang *et al.*<sup>56</sup> published an influential Review arguing for large sample sizes to detect small effects, and they highlighted the role of both minor allele frequency (MAF) and effect size in determining the power of a test of SNP association. They also discussed predictions from population genetics theory of the MAF spectrum over SNPs within a population and the (limited) theory and data to predict effect size distributions. The corresponding arguments are no less compelling for EWASs, but the relevant parameters are even more difficult to predict because of the paucity of data and relevant theory. DNA alleles do not typically vary across cells and can now be typed with low error rates. By contrast, methylation states may be tissue-specific and can vary over cells within a tissue, over alleles within a cell (ASM) and, in rare cases, over DNA strands within an allele (hemimethylation). Thus, for a tissue sample from one individual, the methylation state measured at a CpG site lies between zero and one, as it is an average over cells, alleles and strands and is further blurred by measurement error. Here, we use the limited available information about frequency spectra of DNAm variants, and their effect sizes for common disease, to tentatively propose power calculations under three scenarios. It remains unclear how realistic the proposed scenarios are, but we hope at least to stimulate further discussion and investigation into this important aspect of EWAS design.

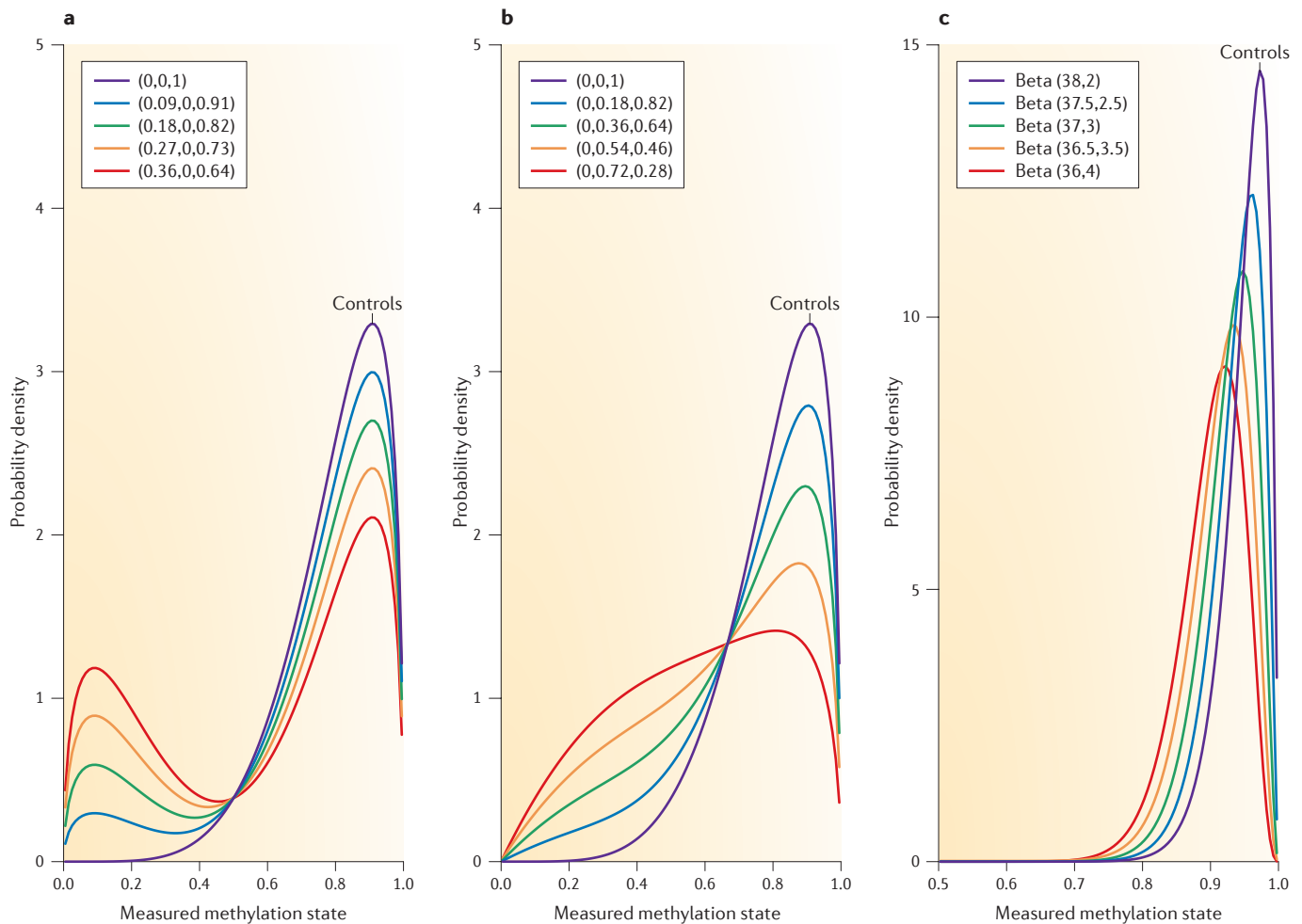
A recent methylome analysis reported that, on average, 68% of CpG sites were methylated in human peripheral blood mononuclear cells<sup>57</sup>. There was great variation across genomic contexts: CpG sites in regions of high CpG density were almost always unmethylated, as were CGIs and 5'-UTRs. By contrast, 3'-UTRs, introns and repetitive elements were predominantly methylated. The rate of ASM was estimated to be between 0.3% and 0.6% (more than that attributable to imprinting alone). Hemimethylation was found to be rare (<0.2%, which included non-CpG methylation and incomplete bisulphite conversion). The methylation spectrum was not symmetric: there were few sites close to being 100% methylated but almost entirely unmethylated sites were not uncommon.

In FIG. 2a,b, hypothetical methylation spectra for three different classes of individuals ('methylated', 'intermediate' and 'unmethylated') have been combined to generate overall frequency spectra in cases and controls. These form the basis of the power simulations reported in TABLE 1. The difference in mean methylation rate between cases and controls provides a popular summary of effect size, but it does not reflect differences in variances or other features of the methylation spectrum. It also does not reflect the relative magnitude of methylation rates, whereas, if a rare epigenotype in controls is almost absent in cases, this is likely to be more important than the same difference of mean rates for a more common epigenotype.

Odds ratios are well-established measures of genetic effect sizes for binary phenotypes. If we regard the mean methylation rate at a site in cases (or controls) to represent the methylation probability for a randomly chosen DNA strand in the case (or control) tissue samples, then we can compute a methylation odds ratio (methOR). This methOR is the same as the ordinary odds ratio, except that the sampling unit is a DNA strand rather than an individual. Thus, the methOR is the odds for a random DNA strand in the tissue sample from a random case to be methylated, divided by the same odds for controls. This provides a measure of effect size that incorporates relative magnitudes but, like the mean difference in rates, it also does not allow for difference between cases and controls of features of the methylation spectrum, such as its variance. As for other odds ratios, methOR is comparable across prospective and retrospective studies, and its value only measures association and does not imply causation.

TABLE 1 gives simulation-based power estimates for the three sets of methylation spectra from FIG. 2. They have similar methORs, although the case-control differences in mean methylation rates are the same for a and b but not for c. The fact that the power values differ between a and b emphasizes that there is no single-number measure of effect size, as power depends on the entire methylation spectrum in cases and controls. However, for the logistic regression analysis conducted in our simulations, methOR gives a better guide to power than the difference in rates. When methOR is around 1.25, a sample size of 800 cases + 800 controls is adequate to achieve 80% power at a significance level of  $\alpha = 10^{-6}$  for scenario c but not a or b (see the next section for a discussion of genome-wide significance for EWASs). When methOR is around 1.5, a sample size of 400 + 400 gives 80% power at  $\alpha = 10^{-6}$  for b and c but not a.

Little is currently known about actual differences in methylation spectra at epigenetic variants implicated in disease, and recommendations about sample size will need to evolve with emerging data. A recent report<sup>58</sup> on the effects of smoking on methylation identified one strong association at a CpG site located in coagulation factor II (thrombin) receptor-like 3 (*F2RL3*), for which the median methylation rates were 95% for those who had never smoked and 83% for heavy smokers, giving a difference of 12% and methOR = 2.7. Methylation status was much less variable in those who had never smoked than in heavy smokers (interquartile ranges 0.94–0.96 and 0.78–0.88, respectively). For such a strong effect, the sample size of 65 heavy smokers and 56 non-smokers was adequate to detect the association. However, smoking is known to be among the most important environmental factors for health, so other effect sizes of interest are likely to be much smaller. If we regard 1.5 to be a target methOR value, then it would not seem to be cost-effective to pursue an EWAS with fewer than 400 cases and 400 controls, and 800 of each would be preferable to achieve good power. This is much less than the 2,000 cases and controls that became the *de facto* standard minimum sample size for GWASs following the



**Figure 2 | Hypothetical DNA methylation frequency spectra in cases and controls.** Methylation states in controls (purple curve) and cases for four effect sizes (other curves) are shown under three scenarios. In panels **a** and **b**, numbers in the parentheses in the key show the proportions of individuals who are unmethylated, intermediate or methylated, respectively. The purple line shows controls and the other four lines show cases. The distributions of measured methylation states are assumed to follow the following beta distributions: unmethylated individuals have a beta(1.5,6) distribution, which has mean = 0.2 and sd = 0.14; intermediate individuals have a beta(2,2) distribution, which has mean = 0.50 and sd = 0.22; and methylated individuals have a beta(6,1.5) distribution, which has mean = 0.80 and sd = 0.14. In **c**, the methylation spectrum is assumed to follow a single beta distribution for controls and each set of cases, and its parameters are shown in the key.

**Bayesian**

The two main statistical schools are the classical (or frequentist) school, which dominated twentieth century science and measures the strength of evidence against a hypothesis using *P* values, and the Bayesian school, which was developed in the nineteenth century but is currently undergoing a resurgence and attempts to compute the posterior probability that the hypothesis is true.

Wellcome Trust Case Control Consortium (WTCCC) study<sup>59</sup>, reflecting the fact that effect sizes for EWASs and GWASs are not directly comparable. It seems likely that effect sizes and hence power will vary substantially according to genomic context, in which case genome-wide ranking by *P* values is unsatisfactory<sup>60</sup> and Bayesian measures of support that take power into account are more appropriate. Currently, however, there remains little information to inform Bayesian prior distributions of effect sizes.

**Genome-wide significance.** In GWASs, the establishment of genome-wide thresholds for significance is complicated by correlations between the genotyped SNPs<sup>61</sup>. In EWASs, there are analogous correlations among DNAm sites in DMRs, but these correlations typically extend to, at most, a few kilobases — although

to date they have only been reported in non-disease contexts. Based on what has been discussed above on co-methylation, ASM and hemimethylation, most CpG methylation can be expected to be symmetric across strands and across alleles in somatic cells. Thus, the ~28 million CpG sites in the haploid human genome correspond to substantially fewer independent methylation states, owing to correlation within DMRs and methylation symmetry. If a set of 500K CpG sites was evenly spaced, the average spacing between sites may be large enough to allow an assumption of independence. In such a case, a significance level  $\alpha = 10^{-6}$  per site gives probability 0.36 of no false positives (type 1 error) under the null hypothesis, and this might be regarded as a liberal threshold for a possible EWAS association. If 5 million CpG sites were assayed, we would expect five false positives under the null at this  $\alpha$ -value. Correlation



Table 1 | Power simulations for epigenome-wide association studies

	(Number of cases, number of controls)			
	(100,100)	(200,200)	(400,400)	(800,800)
<b>Scenario a</b>				
methOR = 1.24 md = 3.6%	0 0	0 0	1 0	20 3
methOR = 1.49 md = 7.2%	0 0	4 0	66 21	100 99
methOR = 1.78 md = 10.8%	2 0	55 11	100 98	100 100
methOR = 2.10 md = 14.4%	18 1	99 78	100 100	100 100
<b>Scenario b</b>				
methOR = 1.24 md = 3.6%	0 0	0 0	2 0	33 8
methOR = 1.49 md = 7.2%	1 0	16 2	85 51	100 100
methOR = 1.78 md = 10.8%	13 1	84 46	100 100	100 100
methOR = 2.10 md = 14.4%	60 19	100 97	100 100	100 100
<b>Scenario c</b>				
methOR = 1.27 md = 1.25%	1 0	7 1	50 19	98 88
methOR = 1.54 md = 2.5%	37 10	95 78	100 100	100 100
methOR = 1.82 md = 3.75%	95 77	100 100	100 100	100 100
methOR = 2.11 md = 5.0%	100 99	100 100	100 100	100 100

The data in this table show the power (%) to detect a methylation variable position (MVP) at significance level  $\alpha = 10^{-6}$  (left entry in each cell) and  $\alpha = 10^{-8}$  (right entry) for the stated sample sizes under scenarios a, b and c of FIG. 2. Analysis is via a Wald test in logistic regression implemented in the R software. md, difference in mean methylation rate between cases and controls; methOR, methylation odds ratio (the odds for a random DNA strand in the tissue sample from a random case to be methylated, divided by the same odds for controls).

among neighbouring sites means that a specific calculation is required to identify a stringent standard for epigenome-wide significance (global type 1 error  $<0.05$ ), which will typically lie between  $10^{-8}$  and  $10^{-7}$ .

**Confounding in EWASs.** GWASs can be affected by two sources of confounding. First, with retrospective ascertainment there is a risk of systematic differences between cases and controls in the handling or processing of samples (known as technical confounding, which includes batch effects)<sup>62,63</sup>. Similar problems are possible for EWASs. Second, confounding can arise because the ancestry of cases differs systematically from that of controls (known as population structure and cryptic relatedness)<sup>64</sup>. This causes confounding in GWASs because any polygenic contribution to disease causation is correlated with ancestry, and environmental exposures may also be correlated with ancestry (for example, owing to different geographic locations of ancestors). Whether ‘polyepigenetic’ effects exist seems unclear, but environmental exposures correlated with ancestry seem likely to affect epigenetic studies.

Unlike GWASs, environmental factors can also directly confound an EWAS by affecting both epigenotype and phenotype, which can inflate type 1 error and exaggerate effect size estimates. Potential confounders, such as age<sup>65</sup> and smoking behaviour, should be adjusted for in a regression analysis if this is possible. Even if a measured covariate is not a confounder but, for example, has an independent effect on phenotype, then adjusting for it can allow better delineation of the direct epigenetic effect.

Fortunately, the large numbers of SNPs in a GWAS allow many possibilities to detect and correct confounding<sup>63</sup>, including genome-wide adjustment of association statistics, regression adjustment using principal coordinates and mixed regression models<sup>64</sup>. Similar methods are likely to be effective to detect and adjust for confounding in EWASs. For example, leading principle coordinates of genome-wide methylation states may encapsulate unmeasured confounders, so if these are also correlated with phenotype, then it may be appropriate to include them as covariates in a regression analysis, as is common for GWAS analyses. Indeed, if GWAS data are also available on the EWAS individuals, it may be appropriate to adjust for leading principle coordinates of both genetic and epigenetic states.

**Analysis of multi-stage studies.** The values in TABLE 1 assume a single-stage study but, as discussed above, the possibilities of confounding, of correlation with genotype and of reverse causation often argue for a two-stage study design; for example, by including a discordant monozygotic-twin stage followed by a longitudinal cohort stage. In simple settings, it is optimal if the sample size in each stage is inversely proportional to the square root of the cost per individual in that stage<sup>66</sup>. The question arises of whether the second stage should assay all the sites from the first stage or whether costs can be reduced by only assaying a limited set of ‘hits’ in stage two. Because of the relatively low cost of assaying all hits in stage two, and the additional information that is provided, this strategy seems generally preferable. The exception would be scenarios in which stage one is large

**Principal coordinates**

Analysis of principal coordinates is a multivariate statistical technique that is related to principal components analysis but investigates individuals rather than variables. It is often used to investigate population structure in a sample of individuals whose relatedness has been estimated from genome-wide genotype data.

enough to eliminate all but a handful of potential hits. In either case, broadly speaking, it is optimal to conduct a single joint analysis of results from both stages. If stage one involves monozygotic-twin pairs, a paired analysis may be appropriate (such as a paired *t*-test) if there is substantially more variation among twin pairs than within them. A combined two-sample case-control analysis is then not appropriate, but it is straightforward to combine test statistics from the two stages using standard meta-analysis techniques.

**Replication for EWASs.** Particularly in the early days of GWASs, replication of hits in an independent study was important in weeding out false positives that arose through technical or design flaws in the initial study. Arguably, GWAS design has improved to the extent that replication is less crucial now as there are many checks available on the quality of the primary study, but replication is still seen as highly desirable and is typically relatively easy to achieve. Ideally, replication should be carried out by an independent group of researchers studying the same polymorphism in the same population and with the same phenotype definition, but preferably using a different study design and different laboratory techniques. In practice, it is impossible to demand all of this, and what constitutes a satisfactory compromise is a matter of debate, although there are some broad points of consensus<sup>67</sup>. For EWASs, the same issues arise, and the issues of correlation with genotype and reverse causation should both be addressed in replicate analyses. Thus, a replication is potentially more demanding for EWASs than for GWASs, but limited availability of tissue samples and study subjects mean that replication will be harder to achieve. As the EWAS field begins to develop, it would be inappropriate for reviewers and editors to impose overly strict replication requirements that are analogous to those used in the current mature phase of GWASs. In particular, we should avoid any encouragement for researchers to hold back samples or resources from the primary study in order to use them later to claim ‘replication’. Lessons should be learned from the GWAS experience: the primary study needs to be well-powered, and rigorous quality checks need to be imposed on the EWAS data. If replication is not immediately feasible, this should not preclude publication, but the need for further confirmation of results should be acknowledged. The appropriate level of tolerance of false positives from the primary study depends on several factors, including the costs of follow-up analyses. If these costs are not too excessive, it may be optimal to initially tolerate some false positives in order to minimize false negatives. The field of EWASs needs to develop in a similar fashion to GWASs, such that standards tighten over time as lessons are learnt from the accumulated experience of the research community.

#### Post-EWAS follow-up studies

The ultimate aim of EWASs, like GWASs, is to provide a better understanding of disease aetiology and to lead to the development of novel therapeutics and diagnostics. Typical follow-up experiments to determine the

aetiological role of disease-associated epigenetic variation could include correlation with other epigenetic modifications and collectively how they have an impact on gene expression. This could be achieved using ChIP-seq experiments, either for the many histone modifications known to correlate with DNAm<sup>68</sup> or for transcription factors whose binding may be modulated — either positively or negatively — by methylation at their target sites<sup>69</sup>. If a large effect size can be determined for a single site, then one could validate the link to the disease-associated phenotype by modulating the expression of the gene in question either in *in vitro* systems or model organism studies. However, a more likely scenario is of many disease-associated epigenetic variants each conferring only a small disease risk, as is suggested by the few small-scale EWASs conducted to date<sup>22,23,40–42</sup>. In this case, it may be more fruitful to use approaches that integrate both computational and experimental methodologies to look at perturbations of entire transcriptional networks. The issue of reverse causation is also important in post-EWAS experiments, both in terms of which variants to follow-up and in terms of the experimental approaches.

Even if the aetiological role of any identified epigenetic variant proves elusive, it may still be possible to use them as predictive biomarkers. In this regard, the combination of chemical stability and ontogenetic plasticity makes DNAm ideally suited as a biomarker. Translating any molecular marker, including DNAm differences, into clinically informative biomarkers has turned out to be more challenging<sup>70</sup> than had been expected, but progress has been made. Following earlier setbacks, a multi-centre study identified, validated and replicated hypermethylation at septin 9 (*SEPT9*) as a blood-based DNAm biomarker for colorectal cancer in 2008 (REF. 71), leading to a commercial test in early 2010 (REF. 72). However, enthusiasm is tempered with caution, as highlighted by the problems encountered by the cancer research community in identifying biomarkers that predict which patients would benefit from a particular therapy<sup>70</sup>. The main problem has been the inability to select patients with a molecularly well-defined disease phenotype owing to, in large part, the heterogeneity of cancer tissues. Molecular heterogeneity is also an issue for the common diseases that are being targeted by the first wave of EWASs, although it is expected to be a smaller problem than for cancer.

Based on this experience, a systematic approach, such as the recently launched [OncoTrack](#) project (see Further information), is needed to advance the field. Two bodies in particular — the Biomarkers Consortium and the AACR-FDA-NCI Cancer Biomarkers Collaborative (a partnership between the American Association of Cancer Researchers (AACR), the US Food and Drug Administration (FDA) and the US National Cancer Institute (NCI)) — have recently issued a comprehensive report on the current state of affairs and future directions<sup>73</sup>. The response of the community has been positive, prompting calls such as “Bring on the biomarkers” (REF. 74) and pledges to replace the patchy framework of fragmented research with a coordinated ‘big science’ approach (such as OncoTrack), which has proved

#### ChIP-seq

(Chromatin immunoprecipitation followed by sequencing). A method for mapping the distribution of histone modifications and chromatin-associated proteins genome wide that relies on immunoprecipitation with antibodies to modified histones or other chromatin proteins. The enriched DNA is sequenced to create genome-wide profiles.

**Bivalency**

A property of chromatin that contains both activating and repressing epigenetic modifications at the same locus.

**Multivariate hidden Markov analysis**

A statistical method for modelling multidimensional data by one of a small number of hidden Markov states, each of which is associated with a multivariate probability distribution.

successful for efforts such as the human and cancer genome projects. Based on this and on other efforts, we can be cautiously optimistic that similar progress will also be made for epigenetic biomarkers.

**Integration of EWASs and GWASs**

The correlations that have been observed between genotype and epigenotype (methQTLs) are encouraging for the prospects of further integrated analysis. A recent study<sup>39</sup> analysing SNPs, gene expression and DNAm in 77 HapMap cell lines identified SNPs that affect both gene expression and DNAm, thus providing evidence for shared genetic and epigenetic mechanisms affecting multiple QTLs. In this way, EWASs can be used to investigate genetic predispositions that exert their function through epigenetic mechanisms. One possible strategy involves the design of a custom array tiled across haplotypes identified by disease-associated GWAS SNPs, profiling it for differential DNAm and analysing the data stratified for risk SNPs rather than cases and controls. Using this strategy, a recent study<sup>75</sup> successfully integrated GWAS and EWAS data to identify haplotype-specific DNAm (HSM) in a type 2 diabetes and obesity susceptibility locus. In the future, it may well be possible to do similar analyses for additional and combinatorial epigenetic marks to capture certain chromatin disease states — for example, based on altered bivalency status — that are currently not easily captured by DNAm. Using multivariate hidden Markov analysis of recurrent and spatially coherent combinations of epigenetic marks, a recent study<sup>76</sup> reported 51 distinct chromatin states for human T cells that look highly promising for possible integration with GWAS data of blood-based diseases.

**Conclusions and future directions**

The success of GWASs in identifying disease-associated genetic variations clearly warrants the development of complementary approaches for identifying additional variations that cannot be captured with GWASs. As outlined in this article, EWASs have the potential to do just that by capturing disease-associated epigenetic variations, such as differential DNAm.

The single most useful resource empowering GWASs was the availability of a detailed SNP map of the human genome<sup>77,78</sup>, which allowed the selection of so-called tag SNPs for comprehensive variation coverage and cost-efficient profiling. DNAm is correlated over tissue-specific blocks of CpG sites spanning up to 1 kb<sup>79</sup>. Knowledge of this block structure for different tissues and cell types has and will continue to improve the selection of CpG sites for EWASs as new methylome maps become available. Currently, such high-resolution maps are available for human embryonic stem cells, fetal fibroblasts and peripheral blood monocytes<sup>8,57</sup>, informing potential EWASs on early developmental disorders and blood-based diseases. As part of the recently launched International Human Epigenome Consortium (IHEC), 1,000 reference epigenomes (including methylome maps) will be generated for human tissues and cell types over the coming years. In this context, these maps can be considered as the epigenetic equivalent to the human haplotype map and can be expected to significantly accelerate and improve our ability to conduct EWASs for many common diseases.

In addition to improving study design — for which we have discussed the key issues in this Review — the main challenge for EWASs will be access to appropriate samples. A useful starting point would be to establish the proposed [Biobank Central](#) resource, which will allow researchers to electronically search for specific combinations of samples and associated data as required for EWASs. Initiation of new birth and other longitudinal cohorts should also be encouraged and existing collections should ensure that samples are suitable for EWASs and related studies that are likely to require chromatin (not just DNA) in the future. Finally, appropriately powered and designed EWASs need to be conducted to enable the development of tools for the analysis, interpretation and integration of EWAS data. To achieve this will require close cooperation between scientists, clinicians, resource providers and funding agencies, as pioneered for GWASs. At the time of writing, the first wave of EWASs was still underway and an [international conference](#) has been arranged for later this year to discuss first results.

- Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
- 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). **This paper proposes a mechanism whereby genetic variants that do not change the mean phenotype could change the variability of the phenotype, which could be mediated epigenetically.**
- Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).
- Kulis, M. & Esteller, M. DNA methylation and cancer. *Adv. Genet.* **70**, 27–56 (2010).
- Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
- MacArthur, D. Why do genome-wide scans fail? *Genetic Future* [online], <http://www.genetic-future.com/2008/03/why-do-genome-wide-scans-fail.html> (2008).
- Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA* **97**, 5237–5242 (2000).
- Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009). **This paper describes the first human methylome to be mapped at single-base resolution, demonstrating extensive DNAm at non-CpG sites in stem cells.**
- Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
- Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
- Veron, N. & Peters, A. H. Epigenetics: Tet proteins in the limelight. *Nature* **473**, 293–294 (2011).
- Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128**, 763–776 (2007).
- Rassoulzadegan, M. *et al.* RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441**, 469–474 (2006).
- Rakyan, V. K. *et al.* DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* **2**, e405 (2004). **This is the first systematic study of DNAm profiles in the human genome.**
- Frigola, J. *et al.* Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nature Genet.* **38**, 540–549 (2006).
- Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genet.* **41**, 178–186 (2009).
- Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.* **20**, 972–980 (2010).
- Fabris, S. *et al.* Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. *Epigenetics* **6**, 188–194 (2011).



19. Lechner, M., Boshoff, C. & Beck, S. Cancer epigenome. *Adv. Genet.* **70**, 247–276 (2010).
20. Ting, D. T. *et al.* Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593–596 (2011).
21. Feber, A. *et al.* Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. *Genome Res.* **21**, 515–524 (2011).
22. Javierre, B. M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.* **20**, 170–179 (2010).
23. Nguyen, A., Rauch, T. A., Pfeifer, G. P. & Hu, V. W. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. *FASEB J.* **24**, 3036–3051 (2010).
24. Bach, J. F. The effect of infections on susceptibility to autoimmune and allergic diseases. *N. Engl. J. Med.* **347**, 911–920 (2002).
25. Barker, D. J. Maternal nutrition, fetal nutrition, and disease in later life. *Nutrition* **13**, 807–813 (1997).
26. Thompson, R. F. *et al.* Experimental intrauterine growth restriction induces alterations in DNA methylation and gene expression in pancreatic islets of rats. *J. Biol. Chem.* **285**, 15111–15118 (2010).
27. Heijmans, B. T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl Acad. Sci. USA* **105**, 17046–17049 (2008).
28. Ng, S. F. *et al.* Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. *Nature* **467**, 963–966 (2010).
29. Rakyan, V. K. *et al.* Transgenerational inheritance of epigenetic states at the murine *Axin<sup>fl</sup>* allele occurs after maternal and paternal transmission. *Proc. Natl Acad. Sci. USA* **100**, 2538–2543 (2003).
30. Morgan, H. D., Sutherland, H. G., Martin, D. I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genet.* **23**, 314–318 (1999).
31. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl Acad. Sci. USA* **102**, 10604–10609 (2005).
32. Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genet.* **41**, 240–245 (2009).  
**These two papers represent key analyses of DNAm differences between monozygotic twin pairs. They provided first evidence for epigenetic metastability in humans that is unlikely to be explained by genetic heterogeneity.**
33. Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.* **5**, e1000602 (2009).
34. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).
35. Kerkel, K. *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature Genet.* **40**, 904–908 (2008).  
**This was the first genome-wide survey to establish sequence-dependent ASM to be a recurrent phenomenon outside imprinted regions. This finding has implications for mapping and interpreting associations of non-coding SNPs and haplotypes with human phenotypes.**
36. Hellman, A. & Chess, A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics Chromatin* **3**, 11 (2010).
37. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
38. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20**, 883–889 (2010).
39. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
40. Feinberg, A. P. *et al.* Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci. Transl. Med.* **2**, 49ra67 (2010).
41. Bell, C. G. *et al.* Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Med. Genomics* **3**, 33 (2010).
42. Mill, J. *et al.* Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am. J. Hum. Genet.* **82**, 696–711 (2008).
43. Baylín, S. & Bestor, T. H. Altered methylation patterns in cancer cell genomes: cause or consequence? *Cancer Cell* **1**, 299–305 (2002).
44. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Rev. Genet.* **11**, 191–203 (2010).
45. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotech.* **28**, 1097–1105 (2010).
46. Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotech.* **28**, 1106–1114 (2010).  
**These two papers benchmarked and compared six of the most commonly used methods for DNAm analysis.**
47. Beck, S. Taking the measure of the methylome. *Nature Biotech.* **28**, 1026–1028 (2010).
48. Ulrey, C. L., Liu, L., Andrews, L. G. & Tollefsbol, T. O. The impact of metabolism on DNA methylation. *Hum. Mol. Genet.* **14** (Suppl. 1), R139–R147 (2005).
49. Widschwendter, M. *et al.* Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. *Cancer Res.* **64**, 3807–3813 (2004).
50. Carone, B. R. *et al.* Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell* **143**, 1084–1096 (2010).
51. Bell, J. T. & Spector, T. D. A twin approach to unraveling epigenetics. *Trends Genet.* **27**, 116–125 (2011).
52. Pearson, H. Epidemiology: study of a lifetime. *Nature* **471**, 20–24 (2011).
53. Yamagata, K. DNA methylation profiling using live-cell imaging. *Methods* **52**, 259–266 (2010).
54. Paliwal, A., Vaissiere, T. & Herceg, Z. Quantitative detection of DNA methylation states in minute amounts of DNA from body fluids. *Methods* **52**, 242–247 (2010).
55. Levenson, V. V. DNA methylation as a universal biomarker. *Expert Rev. Mol. Diagn.* **10**, 481–488 (2010).
56. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
57. Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**, e1000533 (2010).
58. Breiting, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. Tobacco-smoking-related differential DNA methylation: 27k discovery and replication. *Am. J. Hum. Genet.* **88**, 450–457 (2011).  
**This is the first example of a well-designed EWAS. The authors used a combination of a discovery cohort and technical validation using a different platform, followed by replication, to identify a single CpG site that displays an extremely significant correlation with smoking status.**
59. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
60. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–690 (2009).
61. Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179–185 (2008).
62. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
63. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genet.* **37**, 1243–1246 (2005).
64. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 11 (2009).
65. 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).
66. van Belle, G. *Statistical Rules of Thumb* 2nd edn (Wiley, Hoboken, New Jersey, 2008).
67. Chanock, S. J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
68. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10**, 295–304 (2009).
69. Palacios, D., Summerbell, D., Rigby, P. W. & Boyes, J. Interplay between DNA methylation and transcription factor availability: implications for developmental activation of the mouse Myogenin gene. *Mol. Cell. Biol.* **30**, 3805–3815 (2010).
70. Sawyers, C. L. The cancer biomarker problem. *Nature* **452**, 548–552 (2008).
71. Grutzmann, R. *et al.* Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS ONE* **3**, e3759 (2008).
72. Payne, S. R. From discovery to the clinic: the novel DNA methylation biomarker <sup>5</sup>SETP9 for the detection of colorectal cancer in blood. *Epigenomics* **2**, 575–585 (2010).
73. Khleif, S. N., Doroshow, J. H. & Hait, W. N. AACR-FDA-NCI Cancer Biomarkers Collaborative consensus report: advancing the use of biomarkers in cancer drug development. *Clin. Cancer Res.* **16**, 3299–3318 (2010).
74. Poste, G. Bring on the biomarkers. *Nature* **469**, 2 (2011).
75. Bell, C. G. *et al.* Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS ONE* **5**, e14040 (2010).
76. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotech.* **28**, 817–825 (2010).
77. Altschuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
78. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
79. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38**, 1378–1385 (2006).  
**The first study to show that DNA methylation is correlated in blocks of up to 1 kb. This finding enables the design of cost-effective EWASs with comprehensive genome coverage.**
80. Beck, S. & Rakan, V. K. The methylome: approaches for global DNA methylation profiling. *Trends Genet.* **24**, 231–237 (2008).
81. Li, N. *et al.* Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **52**, 203–212 (2010).
82. Robinson, M. D., Statham, A. L., Speed, T. P. & Clark, S. J. Protocol matters: which methylome are you actually studying? *Epigenomics* **2**, 587–598 (2010).
83. Irizarry, R. A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18**, 780–790 (2008).
84. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics* **1**, 177–200 (2009).
85. Suzuki, M. & Grealia, J. M. DNA methylation profiling using *HpaII* tiny fragment enrichment by ligation-mediated PCR (HELP). *Methods* **52**, 218–222 (2010).
86. Brinkman, A. B. *et al.* Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **52**, 232–236 (2010).
87. Rauch, T. A. & Pfeifer, G. P. DNA methylation profiling using the methylated-CpG island recovery assay (MIRA). *Methods* **52**, 213–217 (2010).
88. Serre, D., Lee, B. H. & Ting, A. H. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* **38**, 391–399 (2010).
89. Mohn, F., Weber, M., Schubeler, D. & Roloff, T. C. Methylated DNA immunoprecipitation (MeDIP). *Methods Mol. Biol.* **507**, 55–64 (2009).
90. Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotech.* **26**, 779–785 (2008).
91. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protoc.* **6**, 468–481 (2011).
92. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
93. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
94. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).



95. Butcher, L. M. & Beck, S. AutoMeDIP-seq: a high-throughput, whole genome, DNA methylation assay. *Methods* **52**, 223–231 (2010).
96. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol.* **4**, 265–270 (2009).
97. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7**, 461–465 (2010).
98. Pembrey, M. E. *et al.* Sex-specific, male-line transgenerational responses in humans. *Eur. J. Hum. Genet.* **14**, 159–166 (2006).

#### Acknowledgements

S.B. was supported by the Wellcome Trust (084071) and a Royal Society Wolfson Research Merit Award.

#### Competing interests statement

The authors declare no competing financial interests.

#### FURTHER INFORMATION

Vardhman K. Rakan's homepage:

<http://www.icms.qmul.ac.uk/Profiles/Diabetes/Rakan%20Vardhman.htm>

Thomas A. Down's homepage:

<http://www.gurdon.cam.ac.uk/down.html>

David J. Balding's homepage:

<http://www.zebfontaine.eclipse.co.uk/djb.htm>

Stephan Beck's homepage: <http://www.ucl.ac.uk/cancer/research-groups/medical-genomics>

Avon Longitudinal Study of Parents and Children:

<http://www.bristol.ac.uk/alspac>

BioBank Central: [http://en.wikipedia.org/wiki/BioBank\\_Central](http://en.wikipedia.org/wiki/BioBank_Central)

Biomarker Consortium: <http://www.thebiomarkersconsortium.org>

Canadian Biosample Repository: <http://biosample.ca>

Catalog of Published Genome-Wide Association Studies: <http://www.genome.gov/26525384>

EuroBioBank: <http://www.eurobiobank.org>

Epigenomics of Common Diseases Conference:

<http://www.wellcome.ac.uk/conferences/epigenomics>

GenomEUtwin project: <http://www.genomeutwin.org>

International Cancer Genome Consortium:

<http://www.icgc.org>

International Human Epigenome Consortium:

<http://www.ihec-epigenomes.org>

Longitudinal cohorts listed on Wikipedia:

[http://en.wikipedia.org/wiki/Longitudinal\\_study](http://en.wikipedia.org/wiki/Longitudinal_study)

OncoTrack Consortium: <http://www.oncotrack.org>

Public Population Project Observatory:

<http://www.p3gobservatory.org>

The Twins UK cohort: <http://www.twinsuk.ac.uk/cohort.html>

UK Biobank: <http://www.ukbiobank.ac.uk>

UK HALCyon cohorts: <http://www.halcyon.ac.uk/?q=cohorts>

US National Twin Registry: <http://www.niehs.nih.gov/news/events/pastmtg/2005/twin/index.cfm>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF