EPIGENETICS

OPINION

Associating cellular epigenetic models with human phenotypes

Tuuli Lappalainen and John M. Greally

Abstract | Epigenetic association studies have been carried out to test the hypothesis that environmental perturbations trigger cellular reprogramming, with downstream effects on cellular function and phenotypes. There have now been numerous studies of the potential molecular mediators of epigenetic changes by epigenome-wide association studies (EWAS). However, a challenge for the field is the interpretation of the results obtained. We describe a second-generation EWAS approach, which focuses on the possible cellular models of epigenetic perturbations, studied by rigorous analysis and interpretation of genomic data. Thus refocused, epigenetics research aligns with the field of functional genomics to provide insights into environmental and genetic influences on phenotypic variation in humans.

The first association of epigenetic processes with human disease occurred in the early 1980s when it was found that the DNA of cancer cells had less 5-methylcytosine than normal counterpart cells[1,2] and that within this globally hypomethylated cancer genome the promoters of tumour suppressor genes had paradoxically increased DNA methylation[3]. The increased DNA methylation of promoters of tumour suppressor genes provided a reasonable mechanism for silencing of those genes and for their contribution to increased cellular growth. These observations raised the intriguing possibility that events other than DNA sequence mutations were occurring in cancer and caused cells to become chronically altered in terms of transcriptional regulation.

The question then arose whether similar cellular changes were taking place in disease phenotypes other than cancer. Of particular interest were those in which environmental exposures of various types were linked to altered disease susceptibility, in particular those occurring a long time before the phenotype emerged. The observation that DNA methylation could be transmitted from parent to daughter cells[4] had prompted this transcriptional regulatory mechanism to be described as epigenetic[5,6] (BOX 1) and could thus be a molecular mediator propagating the memory of past cellular perturbations.

Taking advantage of new assays to interrogate increasing proportions of the genome, DNA methylation studies were carried out on samples of cells from individuals affected by the phenotype of interest, comparing these profiles with comparable samples from individuals lacking the phenotype. The genome-wide association study (GWAS) was used as a model in these surveys, but instead of looking at the relative allele frequencies of DNA variants, these epigenome-wide association studies (EWAS) tested whether the DNA methylation at individual or groups of adjacent cytosines in the genome differed systematically in those with the phenotype of interest.

EWAS has become a fast-growing area of research[7]. The typical results published for EWAS are those of significant changes in DNA methylation of modest degree, substantially less than the potential maximum change of 100%. It is generally possible to infer associations between loci changing DNA methylation and the likely functions of proteins encoded by nearby genes, indicating that not only are specific loci changing DNA methylation consistently in multiple individuals but also that these changes select for specific genes within the genome that may have importance for disease risk or aetiology.

Despite what sound like positive results of these EWAS approaches, we[7,8] and others[9,10] have noted that this typical EWAS approach suffers from multiple problems. These problems can be grouped into two categories: the robust identification of molecular (usually DNA methylation) differences associated with the phenotype studied, and the interpretability of these results. The robustness issue is the typical focus of reviews addressing the problems of designing and carrying out EWAS[7,9,10] and usually includes discussion of cohort selection, statistical approaches used when dealing with high-dimensional data and caution about the technical artefacts that can occur in all genome-wide studies[11–13]. In this Opinion article, we focus instead on the issue of interpretability of even the highest confidence DNA methylation changes. Although our focus is on DNA methylation, as the most commonly studied regulatory mechanism in human EWAS, the lessons apply to the broadening group of transcriptional regulators studied in human diseases.

First, at the core of the interpretability concern are the often vague definitions and terminologies used when discussing epigenetics, the biological hypothesis being studied and the potential cellular and molecular processes mediating the phenotypic changes, as we discuss in detail below. Second, DNA methylation can change in response to a diverse range of influences and does not solely reflect a chronic alteration of transcriptional regulation in response to a perturbation. For example, a well-recognized source of variability of DNA methylation is the presence of systematic differences in cell-subtype proportions between the groups tested[14,15]. DNA sequence polymorphism is also increasingly recognized to have a very powerful influence on transcriptional regulation, estimated to account for a substantial proportion of differences of DNA methylation between individuals. Twin studies show the heritability of DNA

Box 1 | **A history of the word 'epigenetics'**

The adjective 'epigenetic' is derived from 'epigenesis', a theory of embryology that describes specialized tissue structures developing from non-specialized precursors. Although the need for such a theory is perhaps puzzling today, it was the counter to the preceding theory of 'preformationism', which instead proposed that development started with a miniature version of the organism in a gamete. Aristotle (384–322 BC) was an early voice dissenting from the theory of preformationism, having made observations of development in chickens that were inconsistent with a preformationism process. With the development of microscopy in the 17th century, it was confirmed that mature organs were formed from structurally different precursors, referred to as an emerging property of the developing organ, or epigenesis.

In the 20th century, Waddington combined training in *Drosophila melanogaster* genetics (in the laboratory of T. H. Morgan) and embryology, which involved microsurgical manipulations and grafting of embryos in tissue culture. His focus was on the regulators of cellular differentiation during development, attempting to understand how cells changed their differentiation fates, responding to inductive effects evoked by humoral factors acting on cells with different competence to respond. He described these effects in terms of what he called the epigenetic landscape[67,98], depicted in terms of a hillside within which a valley branches out into a continuously bifurcating delta of valleys, down which a ball could be imagined to roll, representing cell differentiation and distinctive fates (FIG. 1a). Waddington's goal was to bridge the gap between the two fields of embryology and genetics[99], with the downhill slope the process of epigenesis and the delta of valleys of cell fates formed because of the influence of genes. His goal to reconcile the embryologists' model of epigenesis and the potential role of genetic influences in development was reflected by his describing the cell fate landscape as epigenetic, fusing the words epigenesis and genetic.

Waddington's idea was re-interpreted in 1958 by Nanney, who made the assumption that epigenetic systems were those responsible for cellular memory, whether differentiation in a vertebrate cell or the memory of a prior exposure to galactoside by *Escherichia coli*[33]. In the 1970s, this idea of cellular memory as an epigenetic property of cells was taken a step further by both Riggs[5] and Holliday[6], who noted that DNA methylation offered a potential transcriptional regulatory mechanism that could be maintained through cell division. The further re-interpretation by Riggs and Holliday described epigenetic properties more in terms of Nanney's definition focused on cellular memory and less attributable to Waddington's model reconciling the model of epigenesis with genetics, what could be described today as the field of developmental genetics. By describing DNA methylation as epigenetic, the definition shifted from a cellular to a molecular focus, involving processes that are heritable in daughter chromatids following cell division.

The two models of genomic imprinting[100] and X chromosome inactivation[101] share the characteristics of, first, silencing of one of a pair of alleles and, second, occurring early during development (during gametogenesis or early embryogenesis, respectively) and persisting through the lifetime. The findings that DNA methylation and other transcriptional regulatory mechanisms were distinctively organized on the active and silent alleles, subsequently allied with the explosion of new assays for studying such regulatory processes genome-wide, prompted the latest definition of epigenetics. In this definition, epigenetics is back-translated to its imagined original Greek roots, "the inheritance of variation (-genetics) above and beyond (epi-) changes in the DNA sequence" (REF. 102). We describe this as the 'epi+genetics' definition, encompassing nearly all regulatory differences between cells, interchangeable with the generic description 'transcriptional regulatory', which is probably a more appropriate term to use in most cases of current use of the adjective epigenetic. At present, the cellular inheritance aspect of the definition of Riggs and Holliday is often implied but not explicitly tested, whereas the original Nanney and Waddington emphases on cellular memory and cellular properties involved in development have been lost.

methylation to be between 18%[16] and 37%[17], whereas studies in human HapMap lymphoblastoid cell lines (LCLs) estimated 22–63% of the DNA methylation variability to be due to DNA sequence polymorphism[18], a study of a three-generation family increasing this estimate to 80%[19], and more recent studies estimating DNA sequence polymorphism effects to account for 13.9% of DNA methylation variability in monocytes[20] and 65.7% in a study of various cell types[21]. The importance of genetic effects is also emphasized by studies of DNA methylation quantitative trait loci (meQTLs; also known as mQTLs) that have identified tens of thousands of CpG dinucleotides at which DNA methylation levels are affected by common genetic variation in human populations[18,20,22–24]. However, many EWAS do not measure or account for genetic effects on DNA methylation. A third issue is reverse causation. If the starting hypothesis is that DNA methylation changes are part of a chronic alteration in transcriptional regulation leading to the phenotype studied, it is often not be possible to find conclusive evidence for this hypothesis in the usual cross-sectional study design that includes people who have already developed a phenotype. It has now been shown that DNA methylation in peripheral blood leukocytes can be altered in response to increased body mass index[25,26] or altered blood lipid profiles[27]. Reverse causation occurring at the molecular level occurs when DNA methylation is altered by transcription through the locus[28,29], reflecting rather than causing the transcriptional differences observed.

As a result of these issues, we have proposed that no EWAS to date, which includes our own studies, can be said to be fully interpretable[8]. The goal of this article is to focus on how this situation can be improved and how the concerns of vaguely defined terminology and biological hypotheses, confounding factors and ambiguous causation can be addressed. Although we focus on humans and on diseases other than cancer, and on the analysis of DNA methylation rather than differences in gene expression or chromatin states between cases and controls, many of these insights should apply more broadly. We propose that, although we currently face substantial challenges in generating results that are interpretable in terms of their underlying biological models, these challenges all appear to be surmountable, allowing unprecedented characterization of the sources and mechanisms underlying phenotypic variation in humans in this post-GWAS era.

The complex definition of epigenetics

Some of the current debate about the applications, interpretation and study design of EWAS is because of the extremely vague definition of the word 'epigenetics'. In BOX 1 we describe the evolution of the use of this term, which is also the focus of some excellent reviews[30,31]. What is apparent from this historical summary is that the definition has changed to reflect what we could study at the time: the early definitions were proposed in an era when we were limited to testing cellular developmental events during morphogenesis, whereas more recent definitions are derived from the major advances in understanding the biochemistry of transcriptional regulation and by computational biologists using genome-wide data from assays studying these regulators, in each case treating transcriptional regulatory processes as a type of information separate from and controlling the function of DNA sequence. This transition over time from developmental biology to molecular biology and informatics is substantial, making it

unsurprising that there is currently room for major differences of opinion about the most appropriate use of this term.

Cellular reprogramming. Rather than discussing the relative merits of different definitions of epigenetics, acknowledging their wide range of historical roots (BOX 1), we prefer to focus on the most useful definition for epigenetics when carrying out association studies with phenotypes and the implications of such definitions in terms of the underlying biological models tested by these studies. Although such epigenetic association studies are frequently proposed or published without first defining a clear hypothesis about what is happening to cells being tested, it is probably reasonable to generalize that most such studies aim to test how different environmental conditions or perturbations lead to changes in the properties of one or more cell types that are associated with the development of the phenotype. A specific example of this is the Developmental Origins of Health and Disease (DOHaD) hypothesis[32] of early developmental events, such as perturbations of the nutritional state *in utero*, having a considerable effect on disease risk much later in life. This implicitly refers to the definition of epigenetics by Nanney[33], which focuses on cellular memory of past events affecting their later functional state and to the proposed plasticity of cell states during development[34–36]. Overall, the cellular model underlying epigenetic association studies can generally be described as one of cellular reprogramming, akin to, but much more limited in degree, than the cellular reprogramming that occurs in the production of induced pluripotent stem cells[37]. We make the case below that transcription factors (TFs) are likely to have a major mechanistic role in mediating the chronic alterations of transcriptional regulation within a canonical cell type sought in a typical EWAS. A more overt focus on the idea of cellular reprogramming should be of value in designing and interpreting EWAS, whereas currently this cellular model is generally not stated explicitly as the hypothesis being tested, either in descriptions of EWAS designs or in reports of the results of the studies.

Equating epigenetics with molecular processes. In an EWAS, evidence supporting changes in cellular reprogramming is sought by testing molecular characteristics of that cell type. The first problem that emerges from the use of the current, broad definition that we describe as epi+genetics (BOX 1) is that it equates 'epigenetic' with 'transcriptional regulatory' molecular processes, encompassing any non-DNA candidate genomic regulator that distinguishes cells, including a diverse range of chromatin properties and small RNAs. However, it is not reasonable to infer that all molecular genomic regulators have the ability to mediate cellular memory and thus be capable of mediating cellular reprogramming. For example, unlike plants, animals do not appear to have RNA-dependent RNA polymerases[38], making it difficult to understand how small RNAs could by themselves propagate effects over multiple rounds of animal cell division. However, small RNAs should be reasonable reporters of the effects of cellular reprogramming, and it is plausible that by inducing changes of chromatin states they may exert effects beyond the period of their own presence in a cell lineage[39]. The second challenge is that the molecular mechanisms that have been shown to mediate cellular memory are not limited to this function. For example, DNA methylation has been convincingly shown to associate with classical paradigms of epigenetic regulation such as imprinting and X chromosome inactivation[40–42], but this does not imply that DNA methylation occurring elsewhere in the genome is always propagating an epigenetic cellular memory.

Redefining epigenetic properties as those of cells. We currently describe many molecular genomic processes as epigenetic, whether heritable or not. This laxity of definition allows us to make the logical leap that any changes we see of these transcriptional regulators in a population of cells are mediating a cellular memory of prior perturbation. In turn, this allows an EWAS to be founded upon the hypothesis that by identifying differences in any of this broad group of molecular epigenetic regulators, we have identified a mediator of cellular reprogramming. This leads to a problem of interpretation of the completed EWAS. It should be emphasized that it is perfectly reasonable to use any of the broad range of candidate transcriptional regulators as reflective of cellular reprogramming, but without any requirement that they mediate this process.

As an alternative approach, we propose the following: the EWAS should be based initially on clearly defining a model for the cellular events that are believed to be occurring to mediate the phenotype of interest and then molecular studies should be designed that can test this cellular hypothesis, as discussed in detail below. This shifts the emphasis productively, defining epigenetic properties being those of cells and not equated loosely with a broad range of molecular transcriptional regulators, in effect returning the definition of epigenetics to its original roots. An advantage of an emphasis on cellular epigenetic properties is that it is likely to be simpler to develop testable hypotheses focused on cell states and fates, which can then be explored in terms of the diverse molecular mediators involved in transcriptional regulation.

Refocusing on TFs

The current epi+genetics definition, as we describe it, encompassing all kinds of putative transcriptional regulators has one surprisingly universal omission. A successful EWAS typically defines loci where DNA methylation differs in multiple individuals with the phenotype compared with individuals without the phenotype.

Glossary

Canalization
The maintenance of a cell and its progeny within a differentiation lineage. The term refers to the canal-like structures depicted by Waddington in his epigenetic landscape, which he described not as canals but as creodes, a neologism from biblical Greek words meaning 'necessary' and 'path'.

Epigenetic
We define an epigenetic property as that of a cell, mediated by genomic regulators, conferring on the cell the ability to remember a past event.

Epigenome-wide association studies
(EWAS). Studies of the epigenome for nonrandom association of a difference in organization of a genomic regulator, comparing individuals with a phenotype with individuals lacking the phenotype. The epigenome is itself defined as the genome-wide distribution of transcriptional regulators believed to mediate the memory of past cellular events.

Genome-wide association study
(GWAS). A study that looks for association between genetic variation and a high-level trait such as disease or biomarker across individuals, typically scanning millions of genetic variants genome-wide for association signals.

Polycreodism
A systematic variability of cell fate decisions to create a distinctive repertoire of cells in a tissue.

Quantitative trait loci
(QTLs). Loci in the genome at which genetic variation is associated with molecular variation across individuals. For example, individuals with a particular single nucleotide variant have altered expression levels of a gene (eQTL), altered DNA methylation (meQTL; also known as mQTL) or altered chromatin state (chromQTL).

These DNA methylation changes therefore occur in a sequence-specific manner. However, the enzymes responsible for DNA methylation and other changes often associated with gene regulation (such as histone modifications) do not have DNA-binding domains with the ability to target restricted sets of sequences in this way. Targeting transcriptional regulatory events to specific DNA sequences in mammalian genomes typically involves the activities of TFs. TFs have all of the properties required to direct transcriptional regulation[43], to mediate environmental influences[44] and to maintain cellular memory[45], doing so in a sequence-specific way[46]. Possibly because of their large numbers and the technical challenges in mapping binding of a particular TF throughout the genome[47], they are understudied when looking for mediators of cell perturbations and memory. Another factor is probably the assumed relationship between TF binding and DNA methylation, the likely preconception being that DNA methylation directs TF binding and not vice versa. However, recent studies of the dynamic changes in both TF binding and DNA methylation in differentiating embryonic stem cells support a model in which TF binding induces the local loss of DNA methylation in many situations[48,49]. DNA methylation assays can therefore be regarded as a way of footprinting TF binding patterns in the mammalian genome. However, there are situations in which the presence of DNA methylation exerts a primary role to influence TF binding[50], making it difficult to generalize about the complex interrelationship of these molecular events.

The role of TFs in paradigms of epigenetic regulation. Returning to the major paradigms of epigenetic events in the genome, both genomic imprinting[51] and X chromosome inactivation[52] involve specific TFs functioning at an early stage as part of a multi-component process. However, once established, the two alleles with distinct transcriptional regulatory states are identically exposed to a repertoire of TFs but function in a manner that is dependent on their pre-existing states. Therefore, although TFs have a primary role in establishing transcriptional regulatory patterns, there are other mechanisms to mediate a memory of prior cellular events that can overcome the ongoing presence of TFs. Major candidates for mediating this resistance to TF activity are the Polycomb repressive complexes (PRCs), with PRC2 recruited at an early stage of mammalian X inactivation[53]. Polycomb group proteins were first recognized in *Drosophila melanogaster* to have the properties of maintaining long-term silencing of homeotic genes during development[54,55]. In mammalian cells, the histone H3 lysine 27 trimethylation (H3K27me3) modification deposited by the PRC2 complex is enriched at genes encoding the TFs that could potentially transdifferentiate the cell[56]. The processes that target Polycomb to specific loci to confer the memories of prior cell states appear to involve TFs in *D. melanogaster*[57–61], but in mammals the sequence-specific targeting of Polycomb is less well understood. In addition to a role for TFs, long non-coding RNAs (lncRNAs) and pre-existing histone modifications also appear to contribute, with CG dinucleotide-richness of target sequences being another feature of mammalian Polycomb targets (as previously reviewed in REFS 62,63). Altogether many of the specific mechanisms of the Polycomb complex and its inheritance through DNA replication are unknown[64], Polycomb appears to have a general role in the canalization of cell fate decisions in Waddington's epigenetic landscape model.

We can therefore regard the mediators of memory of past cellular states as involving TFs regulating targeting to specific sequences, with a role for lncRNAs in some cases and with Polycomb involved to mediate long-term silencing events. The examples of X chromosome inactivation and genomic imprinting also involve TFs at early stages, with Yy1 helping to establish X chromosome inactivation[52], whereas Zfp57 has a comparable role in selecting loci for genomic imprinting[51]. As each of these TFs bind or function elsewhere in the genome at loci not undergoing these paradigmatically epigenetic regulatory processes[65,66], the point is reinforced that the mediators of memory of past cellular events do not act exclusively to confer what could be called the epigenetic properties of the cell. Understanding the precise molecular mechanisms conferring cellular memory in different contexts is a key requirement in epigenetics research and should be open to the possibility that TFs are involved.

Cellular epigenetic models

Cellular reprogramming. As described above, the typical, current cellular model of epigenetic perturbation is the reprogramming of one or more cell types in individuals with a phenotype, compared with the same cells in individuals without the phenotype. As depicted by Waddington's original epigenetic landscape[67] (FIG. 1a), this model can be shown as the emergence of a minor, mosaic subpopulation of distinctively altered cells occurring to slightly expand an existing canal (or creode, as Waddington described them[67]) that depicts a canonical cell type (FIG. 1b). This model includes cellular reprogramming events that lead to regulatory changes of a variable number of genes, without identifiably changing the overall cell type, as discussed in further detail below. This cellular reprogramming model can also be applied to stable genomic regulatory changes in long-lived postmitotic cells, which removes the need for its molecular mediators to be heritable through cell division.

Polycreodism. However, there is a second cellular model to consider. When the cellular reprogramming event occurs in differentiating cells as a response to perturbation, during embryonic or later development, it may influence cell fate decisions, leading to differences in the repertoire of cell subtypes formed within a tissue, but without necessarily involving changes in the molecular characteristics of any of the mature cells composing the differentiated tissue. We describe this as polycreodism — the variation in proportions of cell types formed during development as a response to a perturbation. The depiction of this model is different from that of the cellular reprogramming affecting more differentiated cells, with no change in the location or structure of each creode, but a change in their relative depths (FIG. 1c). This model of variability of cell fate decisions was proposed by Waddington as a means by which genetic variation or environmental changes could influence development[67]. This polycreodism model is of particular interest for DOHaD studies. Although such studies have generally focused on testing the cellular reprogramming model, long-term memory of a perturbation during development can also be created by an altered repertoire of cells in an organ or system that can become fixed and persist throughout the lifespan of the individual. In practice, it should be stressed that the distinction between the cellular reprogramming and polycreodism models will be limited by our ability to define and distinguish cell types, and that testing for polycreodism in an EWAS-like study design can be very challenging. Despite these caveats — which are discussed further below — we view polycreodism as a fundamentally important cellular model of human disease that should be more widely considered.

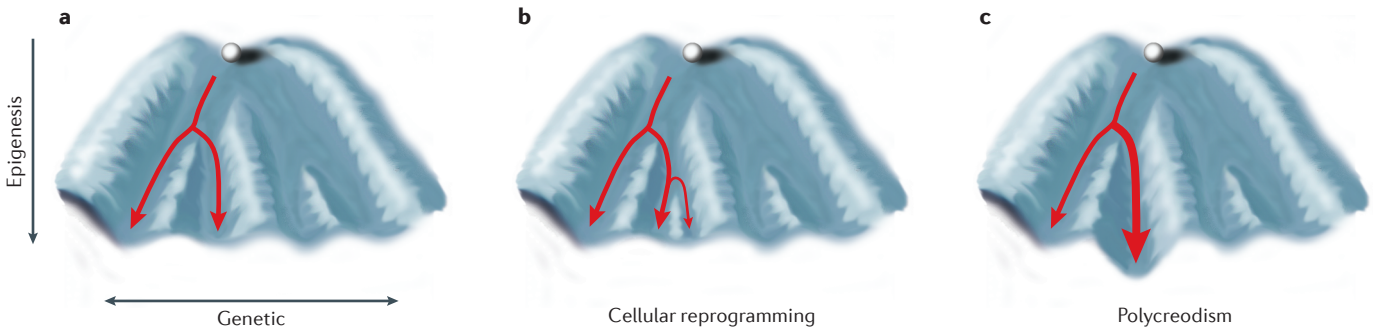

Figure 1 | **Epigenetic landscapes of cell reprogramming and cell-fate changes.** The cell fate decisions made during development are represented by Waddington's epigenetic landscape, which consists of a ball rolling down a hill within channels ('creodes'), with bifurcations representing differentiation choices. **a** | In this lineage decision, the cell has an equal likelihood of ending up in one of two different creodes. **b** | Cellular reprogramming changes the epigenetic landscape, in which the emergence of a mosaic subset of a canonical cell type is represented by a new channel within an existing creode of differentiated cells into which a small minority of cells move but retain the apparent identity of cells of the original creode. **c** | The deepening of one creode depicts the increasing likelihood of cells ending up in that differentiated state. Such variability in creode use between individuals is described in this Opinion article as polycreodism and represents an alternative model for perturbations affecting genomic regulation during development, with potential phenotypic consequences.

Empirical support for the polycreodism model comes from some fascinating recent examples of developmental perturbations. One such observation has been made about micronutrient deficiency during mouse development as a mechanism for later susceptibility to an adult asthma-like phenotype. A retinoid-deficient environment during embryonic days 9.5–14.5 in a mouse model was shown to result in increased amounts of smooth muscle around the airways in mice that appeared otherwise healthy[68]. Pulmonary function tests carried out on these mice in adulthood revealed increased airway resistance[68], representing the non-inflammatory part of a phenotype resembling human reactive airways disease or asthma. A very similar phenotype of increased smooth muscle around the airway has also been found in mice exposed to a vitamin D-deficient environment *in utero*[69]. Both vitamin A, the dietary precursor of retinoic acid, and vitamin D are essential micronutrients that, in combination with retinoid X receptors, function as nuclear receptors. These nuclear receptors represent well-known examples of TFs that regulate gene expression and when deficient have the unexpected effect of causing changes in cell-type proportions within a tissue during development through altered cell fate decisions. The phenotypic effects of vitamin A deficiency during development may not be limited to pulmonary manifestations: mild retinoid deficiency during pregnancy is also associated with decreased numbers of nephrons in the kidneys[70], supporting the possibility that this micronutrient deficiency during development also predisposes to renal disease later in life.

Perturbations of cell fate decisions acting through TFs are not limited to the effects of micronutrients. Endocrine-disrupting chemicals (EDCs) modify normal endocrine system function and represent a major area of interest in epigenetics research[71]. One well-studied example of EDCs is the organotin family, members of which are used as pesticides and in the manufacture of polyvinylchloride (PVC) pipes. *In utero* exposure of mice or *Xenopus laevis* to the organotin tributyltin chloride leads to accumulation of fat from birth to adulthood, leading these compounds to be called obesogens[72]. Mechanistically, tributyltin appears to function at least partially through the peroxisome proliferator-activated receptor-γ (PPARγ) and the retinoid X receptor (RXR)[72], causing mesenchymal stem cells to differentiate preferentially into the adipocyte lineage[73]. Again, the phenotypic effect appears to be mediated by a change in choice of lineage commitment rather than requiring the reprogramming of a specific cell type. Such a change in the proportions of cell subtypes forming an organ or a system and contributing to disease risk is a familiar model for immunologists, who have made many observations linking immune repertoire formation and risk of inflammatory or autoimmune diseases[74].

Polycreodism and cell-subtype adjustment strategies. Variability of cell fate potential is never the cellular model tested in current EWAS. Within the heterogeneous tissues unavoidably sampled in EWAS, differences in the proportions of cell subtypes between cases and controls would be reflected by changes in DNA methylation across a large number of loci. However, this has legitimately been thought of as an influence merely confounding the interpretation of the results of a study testing the cellular reprogramming hypothesis[14,15,75]. Differences in cell-type proportions between cases and controls are often thought of as consequences or correlates of the phenotype rather than causes and are eliminated analytically to the greatest extent possible in EWAS to avoid false-positive findings[14,15,75,76]. However, such well-intentioned rigour applied to the retinoid-deficient mouse example described above would eliminate the disease-mediating outcome of the developmental perturbation: the accumulation of smooth muscle surrounding the airways of the lung. The polycreodism cell model is certainly tenable in any phenotype in which the alteration of the repertoire of cells forming the organ may be mechanistically important in causing the phenotype.

The new challenge of defining a cell type

Successfully carrying out a study associating epigenetic properties, however defined, with a phenotype can therefore be appreciated to be intimately related to definition and quantification of the cell subtypes present in samples tested. Not only is this needed to overcome the influence of systematic changes in cell-subtype proportions as a confounding influence on DNA methylation studies[14,15,77] but also to generate evidence supporting a cellular reprogramming versus a polycreodism

cellular model. However, there is a key issue to consider: how do we define a cell type in the first place? Is what we call cellular reprogramming flexibly modifying the function of a canonical cell type, or are cell identities discrete and discontinuous[78], so that cellular reprogramming generates outcomes comparable to those occurring during cell fate determination that result in polycreodism?

Three levels of cell-type definition. Cell types can be thought of as being distinguishable at three levels. The traditional approach with lowest resolution is based on histology, which involves looking for morphological differences. By using different types of markers, histologically identical cells can be discriminated further. For example, although lymphocytes look very homogeneous by histology, they can be subdivided using cell surface markers into numerous different B cell and T cell subpopulations. However, the new era of single-cell transcriptomics has revealed a much finer level of resolution at the molecular level, discriminating beyond even cell surface marker classifications and uncovering an increasing number of cell types and subtypes[79]. The resolution to distinguish cell types depends not only on the cell-type diversity of studied tissues but also on technical factors, with greater capacity to discriminate cell subtypes when more cells are tested and deeper sequencing is carried out. For these technical reasons, and sometimes also because of biological continuity between cell types, the patterns of single cell profiles can sometimes defy easy classifications into discrete categories[79,80].

Current EWAS cell-subtype adjustment strategies. The potential influence of unrecognized cell-subtype diversity is apparent from the results of EWAS studies themselves. In every EWAS to date, without known exception, statistically significant associations have been based on modest changes of DNA methylation levels between cases and controls. DNA methylation cannot change to such a limited extent within an individual cell, as the cytosine on a pair of chromosomes can be methylated on neither (representing 0% methylation), both (representing 100% methylation) or (relatively rarely) one of the alleles (representing 50% methylation). Thus, a modest change in DNA methylation of, for example, 20% can only occur if there is a phenotype-associated change in the proportion of a mosaic subset of cells with distinctive DNA methylation, at sites where DNA methylation is different between the cell subtypes present.

Many EWAS to date have made efforts to adjust for the confounding effects of cell-subtype composition, which is essential for identifying cellular reprogramming events. Some studies use cell sorting or selection based on surface markers to limit cell-type diversity, although we have shown that this is not by itself sufficient to generate populations that are homogeneous in terms of molecular regulators[81]. The computational approaches that aim to infer cell-subtype proportions are divided into those that use reference DNA methylation data from the cell subtypes that are believed to be present[14], and those that assume covariation of changes in DNA methylation at multiple loci in the genome to reflect the effects of a presumed number of latent cell subtypes present[75,76]. However, even after carrying out these corrective measures, the residual DNA methylation changes have continued to be of only modest degree. This indicates that either cellular reprogramming is affecting a mosaic subpopulation of cells or that these analyses have remained vulnerable to the influences of cell subtype, which may not have been recognized in the lack of molecular genomic level of resolution.

Using single-cell sequencing to measure cell subtypes. Single-cell sequencing has quickly emerged as a powerful approach to characterize cell-subtype composition at higher resolution and accuracy than by using previous strategies, providing direct insights into the functional state of each cell and allowing characterization of cell populations that are not well captured by cell sorting. Although single-cell bisulfite sequencing technologies are being developed to allow DNA methylation analysis at the single-cell level[82,83], by far the most mature and scalable single-cell technology is single cell RNA sequencing (scRNA-seq). We show its use for cell-type composition quantification in FIG. 2 with a fairly simple analysis. We used reference scRNA-seq data from 33,000 peripheral blood mononuclear cells (PBMCs) to identify differentially expressed genes between four major CD4$^+$ T cell subtypes[84], and we then used the expression signatures of these 156 genes for estimation of the proportions of the four CD4$^+$ T cell types in a public RNA-seq data set of bulk CD4$^+$ T cells in individuals of different ages[85,86] (for additional methodological details and data sources, see FIG. 2). The analysis shows a robust signal of the known shift from naive towards memory CD4$^+$ T cell types with age[87], although we note that the proportions of the memory CD4$^+$ T cell subpopulations were sensitive to the choice and processing of the reference data sets.

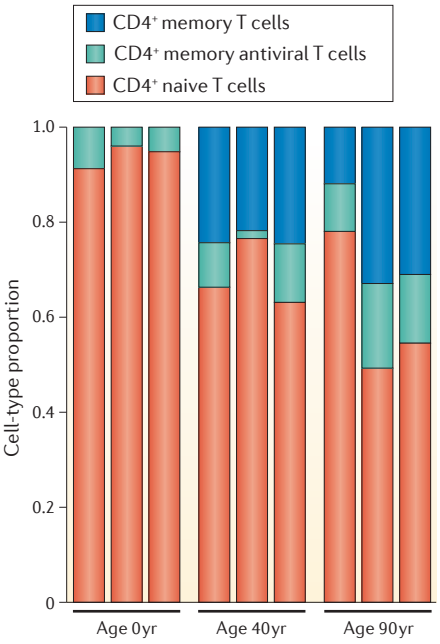

Figure 2 | **Preliminary data from scRNA-seq analysis.** We show the results of deconvolution of bulk cell population RNA sequencing (RNA-seq) data, using CD4$^+$ T cell subpopulations detected in a reference single-cell RNA-seq (scRNA-seq) resource. Specifically, we used reference scRNA-seq data from 33,000 peripheral blood mononuclear cells (PBMCs) from the 10x Genomics Chromium v1 system (see 10x Genomics reference single-cell RNA-seq data), from which we identified 31 cell subtypes using the Seurat software[84] and selected 156 genes that are differentially expressed between three major CD4$^+$ T cell subtypes. We then used the Cibersort software with default settings[85] to apply the expression signatures of these 156 genes for estimation of the proportions of the three CD4$^+$ T cell types in a public RNA-seq data set of bulk CD4$^+$ T cells in individuals of different ages (see Gene Expression Omnibus accession number GSE65515)[86]. We note the age-associated shift from naive to memory T cells and the ability of scRNA-seq to distinguish multiple CD4$^+$ T memory cell types. DNA methylation studies of these cells would probably have shown moderate changes between samples and age groups, reflecting these unrecognized, mosaic cell-subtype changes.

This analysis shows that although decomposition analyses based on scRNA-seq are not yet fully mature, selection using the CD4 surface marker alone is not enough

to enrich a uniform cell population for interpretable epigenetic association studies and that cell-subtype proportions can vary substantially, even between individuals of the same age. It could be argued that scRNA-seq is not essential to quantify $CD4^+$ T cell subtypes, as they can be quantified by flow cytometry, but as scRNA-seq has even revealed cell types for which surface markers are not known[79] and can be applied to tissues for which flow cytometric markers are less developed than haematopoietic cells it looks to be a promising approach to apply in functional genomics studies in general. We conclude that scRNA-seq now allows fine-grained analysis and correction for cell-type composition in epigenetic association studies and that emerging single-cell epigenomic assays will provide an additional layer of information not only for cell-subtype analysis but also for studying mechanisms of gene regulation at the cellular level.

The second-generation EWAS

Defining the hypothesis. We propose several approaches and practices for improving the interpretability of EWAS, which we refer to collectively as the second-generation EWAS. The first step is basic: to define the hypothesis being tested. This article emphasizes the need to clarify the proposed cellular epigenetic hypothesis being tested as the priority, as discussed above, indicating how cellular changes are mechanistically causal to a disease. In biomarker studies in which the question of causality is not of interest, many components of second-generation EWAS will apply, but much more straightforward study designs may be sufficient as interpretability is not the priority. Of the cellular epigenetic models, the hypothesis of cellular reprogramming is broadly amenable to EWAS-like study designs that use observational molecular data of people with appropriate phenotype and environmental exposure data, with the caveats and opportunities discussed below. By contrast, the hypothesis underlying the polycreodism model is not as straightforward to test because it is likely to be very difficult to distinguish whether the change in cell proportions is the cause or consequence of the phenotype. Approaches to study causal polycreodism are likely to include animal models, such as those discussed above, but these methods are not yet established. Therefore, in this section, we focus on the usual cellular reprogramming model (summarized in BOX 2), with studies aiming to pinpoint loci in the genome that are responsible for modified regulatory properties and function of cells, without changing the overall cell type.

Study subjects and biospecimens. Cross-sectional study designs commonly used in EWAS — typically case–control studies — are affected by reverse causation, because any association between a cellular phenotype such as DNA methylation and a disease phenotype can be a result of causality in either direction. One way to define causality is by revealing the temporal development of cellular changes using longitudinal sampling, with other analytical options discussed further below. The necessary sample size depends on cell-type diversity, biological models and the desired analysis. Large numbers of samples increase the sample collection challenges, but are essential to enable the discovery of smaller effects that may be biologically interesting and, importantly, allow more advanced statistical modelling and correction of a large number of covariates, as discussed below. The choice of tissue sampled needs a clear biological justification. Is the hypothesis that the cells in this tissue mediate the phenotype, or do they report an exposure that causes the phenotype in a different tissue? If the cells are likely to merely reflect the consequences of an existing phenotype, they may be suitable for biomarker use but not for testing for cellular reprogramming. Furthermore, it is likely to be a more informative approach if the cells being tested have been found to have a distinctive phenotype. Although samples are always

Box 2 | **Testing for cellular reprogramming**

We divide these guidelines into two parts: those focused on revealing high-confidence differences between the groups studied and those specifically needed to allow interpretation of the results as involving cellular reprogramming.

Revealing high-confidence differences

Design of EWAS. Questions at the outset of an epigenome-wide association study (EWAS) should include whether the specific hypothesis is best tested with cross-sectional, longitudinal or twin cohorts. Can the cell type tested be proposed to have a biologically realistic link to the exposure and/or phenotype? Is the cohort size adequate, not only for the primary association but also for the correction of other confounding factors (such as methylation quantitative trait loci (meQTLs))? What are the analytical approaches and resources in place to handle large, multidimensional data sets? What is the plan for validating the findings?

Replication. Significant findings should be replicated in an independent cohort to protect against spurious findings. It should be noted that biases might replicate across data sets.

Interpretation of the results in light of cellular reprogramming

Accounting for sources of variability. Cell-type composition, genetic variation and transcription should be measured in all subjects. In data creation and analysis, technical artefacts and batch effects need to be considered and controlled for. As much quantitative information as possible should be captured about other factors that are associated with the phenotype and possibly with the molecular trait (for example, metadata such as age, sex and diet) to avoid spurious correlations.

Analysis. Testing for association between phenotype and molecular traits is fairly straightforward per se, but normalization and correction of confounders adds substantial complexity. It is essential for epigenetic association studies to establish statistical quality control tools and best practices. A quantile–quantile plot of *P*-values (the most classical genome-wide association study (GWAS) quality control plot) is extremely informative about the distribution of effects genome-wide: few significant loci with the vast majority of the data following the null hypothesis would pinpoint specific potentially reprogrammed loci, whereas inflated *P*-values genome-wide are a sign of extremely widespread association with the phenotype, or (more likely) confounders that have been insufficiently accounted for.

Interpretation. High-confidence changes in a molecular characteristic such as DNA methylation can indicate several interesting underlying processes. The confounding variables are by themselves of major potential value in understanding processes leading to phenotypes, so that when a DNA methylation change is discarded because it is secondary to, for example, transcription through that locus, what has been revealed is a gene or non-coding RNA expression difference that is nonrandomly associated with the individuals with the phenotype and is thus potentially involved in mediating the phenotype. It is worth starting with the view that all associations are due to events other than cellular reprogramming, and by systematically excluding everything attributable to confounding factors any remaining changes can be defined as indicative of cellular reprogramming. It remains important to be cautious about inferring causality even at this stage: extensive downstream studies will be required to test causality for an association, which should not be a requirement of an EWAS.

going to be heterogeneous in terms of cell subtypes, even after purification[81], it is essential that the cell type of interest is as free from contaminating cell types as possible to obtain maximum sensitivity to detect DNA methylation changes of modest degree and to facilitate correction for cell-subtype diversity.

Cellular subtype quantification. The point has been stressed already, but a central component in the testing of cellular reprogramming models is determining the cell-subtype composition of samples as a major confounding variable. It appears that scRNA-seq represents a generally useful approach, being both widely accessible and fairly affordable, and could potentially be applied to tissues in which cell subtypes have not been well studied. scRNA-seq allows identification, discrimination and quantification of cell subtypes present in a sample. scRNA-seq may not need to be carried out on every sample, but may instead be performed on a limited subset of samples to identify cell subtypes and their distinctive gene expression profiles. These scRNA-seq data can then be used to estimate through analytical approaches the cell-type composition of samples on which bulk RNA sequencing has also been carried out. Failing the availability of single-cell molecular studies, characterization of the cell population with flow cytometry or at least quantitative histology will allow some degree of insight into the cell-subtype composition present in the samples tested. These estimates of cell-type proportions can then be included in association models to remove their effect on EWAS findings, in order to hone in to the more limited number of loci at which regulatory changes indicative of cellular reprogramming events are taking place. Finally, we note that even though it is possible to test for association between cell-type proportions and the phenotype of interest, the necessary null assumption is that the differences are a consequence rather than a cause of the phenotype. However, such information may be useful for forming additional hypotheses of biological processes that contribute to symptoms of a disease, or of possible polycreodism, which both require separate studies to unravel.

Multiple complementary genome-wide assays. Testing for the cellular reprogramming model requires not only estimates of cell-type composition but also the results of molecular assays testing genomic regulators that are markers of loci likely to drive cellular reprogramming. Genome-wide assays testing regulatory features of the genome are extremely diverse and numerous, raising the question of whether some are better than others for identifying cellular reprogramming. To detect low-level mosaic changes within a cell population, DNA methylation studies are inherently more sensitive than the broad group of chromatin immunoprecipitation (ChIP)-based assays, which are not nearly as quantitative. As cellular reprogramming could be associated with the relocation of *cis*-regulatory elements within the genome, there is no a priori way of predicting which loci to test, requiring that molecular profiling be as comprehensive as possible throughout the genome, rather than limited to predefined loci. Whole-genome bisulfite sequencing (WGBS) is the assay that allows most potential *cis*-regulatory loci to be tested for DNA methylation (5-methylcytosine and a smaller contribution by 5-hydroxy-methylcytosine)[88]. Parallel use of the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)[89] generates a comprehensive map of loci of open chromatin, which are the candidate *cis*-regulatory loci for that cell type and are valuable for pinpointing the locations where DNA methylation changes are more likely to have regulatory effects. Furthermore, measuring gene expression by RNA-seq can be valuable for linking changes in the chromatin to active gene expression and for taking into account the reverse causation effects on DNA methylation of transcription through a locus[28,29]. The RNA-seq data can also be used with reference single-cell transcriptomic information to estimate cell-subtype proportions present, as described earlier. Genotyping is also needed to account for the effects of DNA sequence polymorphism on methylation and other molecular phenotypes. The first step of this analysis is genetic association analysis to identify *cis*-meQTLs where a common genetic variant in the studied population sample is associated with the DNA methylation level of a given CpG dinucleotide in *cis*. For sites with significant meQTLs, the genotype can then be included in the EWAS models as a covariate to account for the genetic source of variance.

Generating these multiple data sets presents challenges, as accounting for several covariates typically requires several hundreds of individuals to be analysed, the additional assays increase biospecimen type and volume requirements and the volume of data generated presents computational challenges. However, we believe that the costs involved are ultimately justified by the markedly improved interpretability of results.

This combination of assays should have greatly improved sensitivity and specificity to identify *cis*-regulatory loci that differ between cases and controls in all or a subset of cells being studied, compared with traditional DNA methylation analyses carried out in isolation. As changes in DNA methylation and loci of open chromatin may be merely reporting alterations of TF binding that represent the primary causal molecular change, as discussed above, these assays can reveal the loci where TF binding is altered in association with a phenotype. The subsequent inference of the potential mechanisms of reprogramming involving these loci should include those potentially mediated by TFs. One approach to identify TFs that mediate cellular reprogramming is to look for motifs that are significantly enriched at the loci with high-confidence changes in genomic regulation, which can help to identify known binding sites for the one or more TFs that may be regulating cellular reprogramming. This can allow preliminary insights into the upstream signalling involved in the cellular reprogramming process.

Interpreting the results. The loci that emerge from this stringent analysis looking for cellular reprogramming should be expected to include those that cause the phenotype, changes associated with the phenotype but due to reverse causation (in which the phenotype induces molecular changes in the cells[8,25,27]) and loci showing correlations with confounding factors that have not been recognized or fully accounted for. We show some examples of potential interpretations of certain types of results in FIG. 3. Although the latter class of loci can be reduced — but probably not entirely eliminated — by the approaches described above, a cross-sectional design comparing individuals with and without a phenotype cannot on its own distinguish reverse causation effects from cellular reprogramming[8]. Reverse causation effects can sometimes be biologically and medically interesting and may be informative about symptoms, progression and complications of a disease, but rarely represent the main reason for doing an EWAS. To detect reverse causation, genetic analysis can make use of the fact that genetic associations with phenotypes are always causal. Mendelian randomization is probably the best-known

analytical approach for using genetic associations to infer causality relationships between phenotypes[90]. In addition, as mentioned above, molecular analysis of longitudinal samples collected before and after the development of the phenotype can be informative about the temporal order of events[8]. In the absence of evidence for reverse causation, following up the association found in an EWAS will probably require experiments with animal or cellular models to confirm causality. This is also true for polycreodism hypotheses. Although laborious, working out the causality of associations is valuable for a deeper biological understanding of how the phenotype came about and for developing interventions for disease. A well-carried out EWAS of the cellular reprogramming model is a valuable step towards this goal, allowing the proposition of solid, testable hypotheses of causality.

Harvesting the confounders

We envision that WGBS, ATAC-seq, RNA-seq, scRNA-seq and genotyping of the samples in a second-generation EWAS represent a necessary panel of assays to allow interpretation of the results obtained, although the pioneering empirical studies to test the value of this comprehensive approach are still lacking. We estimate that this panel of assays costs at least tenfold more per sample than a simple DNA methylation survey assay, a substantial investment when carried out on a sizeable cohort. This cost increase is analogous to many early GWAS and genome-sequencing studies, in which initial small studies were found to be inadequate, requiring their own second-generation approaches involving substantially higher investments to achieve scientific success with robust findings.

We anticipate that a substantial proportion of the loci with DNA methylation changes will be attributable to confounding variables. However, in a second-generation EWAS, confounding variables can be not only corrected for but also harvested for additional biological insights. For example, reverse causation because of transcription through a locus makes changes of DNA methylation at that locus less likely to be driving cellular reprogramming, but indicates a transcriptional difference associated with the phenotype, which is potentially an insight into why the phenotype developed. The data will also reveal gene expression QTLs (eQTLs)[91–93], meQTLs[24] and chromatin state QTLs (chrQTLs)[94] for

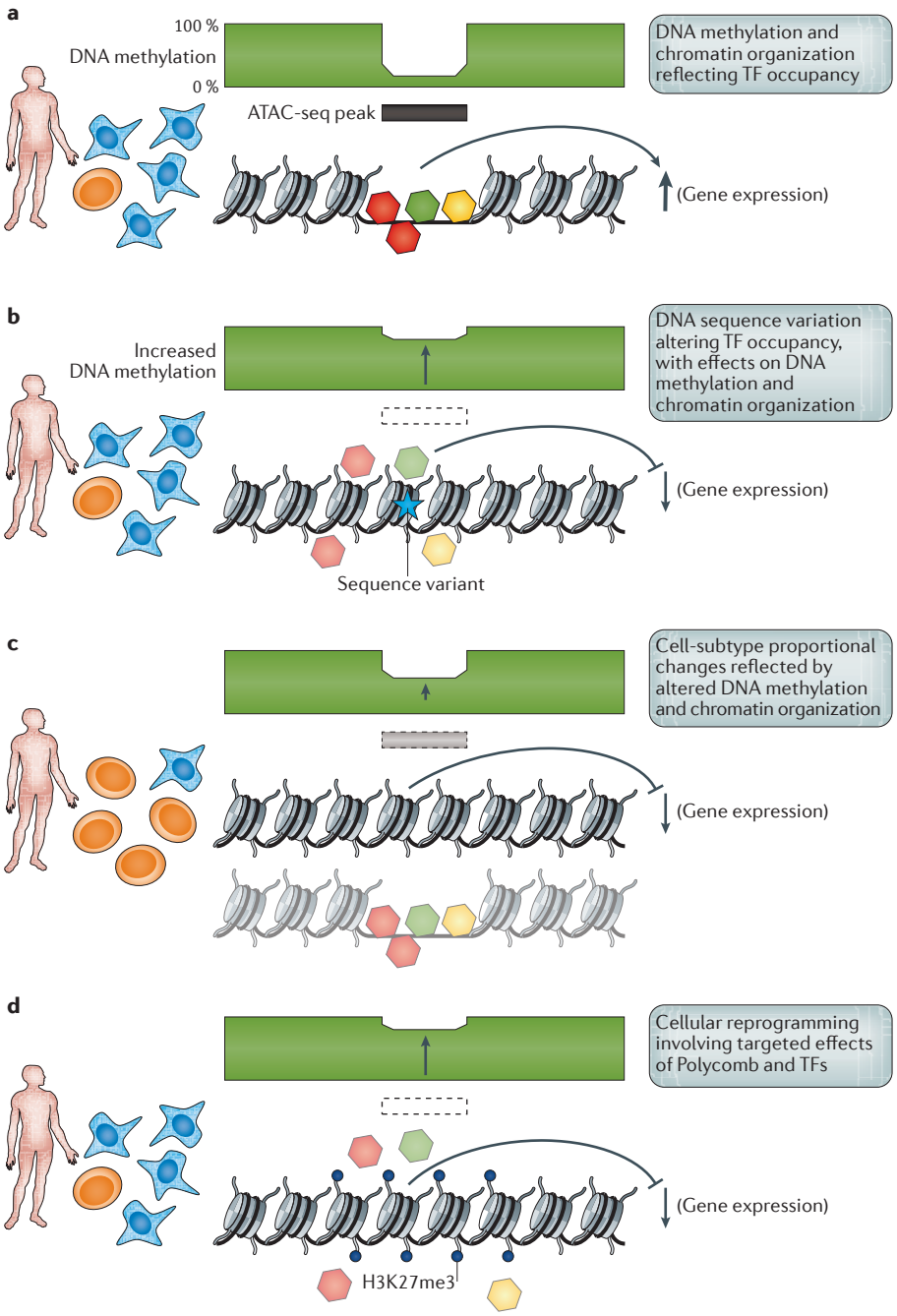

Figure 3 | **Interpreting the results of a second-generation EWAS.** We show a few examples of how to interpret the results of genome-wide assays of transcriptional regulators when these are used to characterize a locus of interest from an epigenome-wide association study (EWAS). **a** | A transcription factor (TF)-centric perspective, in which low methylated regions (LMRs)[49] and loci of open chromatin represent the footprints of TFs (hexagons) binding at *cis*-regulatory elements, shown in the figure to have a positive effect on gene expression. **b** | The presence of a DNA sequence variant within the locus that destabilizes the binding of a TF and leads to loss of the group of TFs from the locus, with acquisition of DNA methylation, loss of open chromatin and loss of the partial stimulatory effect on gene expression. **c** | An alteration in cell-subtype composition could be mistaken for true changes in DNA methylation and chromatin profiles within a cell type, for a locus that is distinctively organized in the contaminating cell subtype and to a degree that is dependent on the proportion of contaminating cells. **d** | An example of how targeted Polycomb-mediated silencing to that locus could prevent TF binding, again leading to changes in DNA methylation, chromatin structure and gene expression, but now because of the type of cellular reprogramming event sought in many EWAS. Only when the competing models that result from confounding effects are excluded can this last model be predicted with confidence to be occurring. ATAC-seq, assay for transposase-accessible chromatin with high-throughput sequencing; H3K27me3, histone H3 lysine 27 trimethylation.

the cell type studied. In a common disease setting, genetic differences between the cases and controls are typically very small (this is why GWAS cohorts of thousands of individuals are needed to find such genetic associations with disease). Thus, ignoring genetic effects in EWAS is unlikely to lead to many DNA methylation–phenotype associations that are actually driven by genetic effects rather than environmentally driven cellular reprogramming. However, at many loci, genetic effects introduce a major additional source of variation and accounting for it markedly improves statistical power to detect true epigenetic associations. Furthermore, molecular QTLs that colocalize with GWAS associations with diseases and other traits form particularly strong hypotheses of molecular changes that cause increased disease risk, with the genetic association functioning as the causality anchor[95–97]. Second-generation EWAS adds value to other ongoing efforts to decipher mechanisms of loci implicated in GWAS, especially by rigorous control of cell-type diversity, and we envision it will become an important part of the post-GWAS toolkit.

Therefore, although adjusting for the confounding variables in a second-generation EWAS reduces the number of loci at which DNA methylation is associated with the phenotype, the loci that survive the corrections are much more likely to indicate genuine cellular reprogramming. Even the loci associated with different confounders are of inherent value in understanding the processes underlying the development of a phenotype. In addition to contributing to post-GWAS analysis, second-generation EWAS with careful models and analysis of cellular reprogramming will be essential to understand environmental perturbations as a major cause of phenotypic variation in humans.

Conclusions

The future of testing cellular epigenetic models will require all of the rigour described in this article to generate interpretable results, but will yield extraordinary insights into the joint effects of DNA sequence variants, environmental effects and consequent genomic regulatory changes and cellular outcomes that underlie human phenotypes. The EWAS field can be regarded as having entered a period comparable to that encountered by the genetic association field in the early 2000s, when it was realized that candidate gene association studies and early GWAS did not produce replicable, robust and interpretable results. This was due to inadequate sample sizes, poor statistical practices and failure to correct for population stratification and technical batch effects. The GWAS field self-corrected and overcame what appeared to be daunting obstacles to generate what are now valuable and high-confidence insights. Although the specific challenges in EWAS are different from GWAS, as discussed above, the need for comparable self-correction is now apparent in the EWAS field, but once this has been achieved there is substantial potential to reveal new insights into genomic mechanisms in human phenotypes.

Tuuli Lappalainen is at the New York Genome Center, 101 Avenue of the Americas, New York, New York 10013, USA; and at the Department of Systems Biology, Columbia University, 1130 Street Nicholas Avenue, New York, New York 10032, USA.

John M. Greally is at the Departments of Genetics, Medicine and Pediatrics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, New York 10461, USA.

tlappalainen@nygenome.org; john.greally@einstein.yu.edu

doi:10.1038/nrg.2017.32
Published online 30 May 2017

1. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89–92 (1983).
2. Gama-Sosa, M. A. *et al.* The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **11**, 6883–6894 (1983).
3. Greger, V., Passarge, E., Höpping, W., Messmer, E. & Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum. Genet.* **83**, 155–158 (1989).
4. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
5. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25 (1975).
6. Holliday, R. A new theory of carcinogenesis. *Br. J. Cancer* **40**, 513–522 (1979).
7. Michels, K. B. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nat. Methods* **10**, 949–955 (2013).
8. Birney, E., Smith, G. D. & Greally, J. M. Epigenome-wide association studies and the interpretation of disease-omics. *PLOS Genet.* **12**, e1006105 (2016).
9. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
10. Heijmans, B. T. & Mill, J. Commentary: the seven plagues of epigenetic epidemiology. *Int. J. Epidemiol.* **41**, 74–78 (2012).
11. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
12. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
13. Kraft, P., Zeggini, E. & Ioannidis, J. P. A. Replication in genome-wide association studies. *Stat. Sci.* **24**, 561–573 (2009).
14. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
15. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
16. Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLOS Genet.* **8**, e1002629 (2012).
17. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).
18. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
19. Gertz, J. *et al.* Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLOS Genet.* **7**, e1002228 (2011).
20. Chen, L. *et al.* Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
21. Cheung, W. A. *et al.* Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol.* **18**, 50 (2017).
22. 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).
23. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).
24. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).
25. Richmond, R. C. *et al.* DNA methylation and BMI: investigating identified methylation sites at HIF3A in a causal framework. *Diabetes* **65**, 1231–1244 (2016).
26. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
27. Dekkers, K. F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17**, 138 (2016).
28. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69 (2007).
29. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009).
30. Pisco, A. O., d'Herouel, A. F. & Huang, S. Conceptual confusion: the case of epigenetics. Preprint at *bioRxiv* http://biorxiv.org/content/early/2016/05/12/053009 (2016).
31. Haig, D. Commentary: the epidemiology of epigenetics. *Int. J. Epidemiol.* **41**, 13–16 (2012).
32. Gillman, M. W. *et al.* Meeting report on the 3rd International Congress on Developmental Origins of Health and Disease (DOHaD). *Pediatr. Res.* **61**, 625–629 (2007).
33. Nanney, D. L. Epigenetic control systems. *Proc. Natl Acad. Sci. USA* **44**, 712–717 (1958).
34. Wu, H., Hauser, R., Krawetz, S. A. & Pilsner, J. R. Environmental susceptibility of the sperm epigenome during windows of male germ cell development. *Curr. Environ. Health Rep.* **2**, 356–366 (2015).
35. Marsit, C. J. Influence of environmental exposure on human epigenetic regulation. *J. Exp. Biol.* **218**, 71–79 (2015).
36. Perera, F. & Herbstman, J. Prenatal environmental exposures, epigenetics, and disease. *Reprod. Toxicol.* **31**, 363–373 (2011).
37. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
38. Zong, J., Yao, X., Yin, J., Zhang, D. & Ma, H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447**, 29–39 (2009).
39. Holoch, D. & Moazed, D. RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.* **16**, 71–84 (2015).
40. Kaslow, D. C. & Migeon, B. R. DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. *Proc. Natl Acad. Sci. USA* **84**, 6210–6214 (1987).
41. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* **21**, 1592–1600 (2011).
42. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).

43. Smith, N. C. & Matthews, J. M. Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors. *Curr. Opin. Struct. Biol.* **38**, 68–74 (2016).
44. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* **46**, 43–68 (2012).
45. Alon, U. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461 (2007).
46. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).
47. Savic, D. *et al.* CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res.* **25**, 1581–1589 (2015).
48. Feldmann, A. *et al.* Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLOS Genet.* **9**, e1003994 (2013).
49. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
50. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
51. Li, X. *et al.* A maternal-zygotic effect gene, *Zfp57*, maintains both maternal and paternal imprints. *Dev. Cell* **15**, 547–557 (2008).
52. Donohoe, M. E., Zhang, L.-F., Xu, N., Shi, Y. & Lee, J. T. Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol. Cell* **25**, 43–56 (2007).
53. Wang, J. *et al.* Imprinted X inactivation maintained by a mouse Polycomb group gene. *Nat. Genet.* **28**, 371–375 (2001).
54. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by Polycomb and trithorax proteins. *Cell* **128**, 735–745 (2007).
55. Grimaud, C., Nègre, N. & Cavalli, G. From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. *Chromosome Res.* **14**, 363–375 (2006).
56. Davis, F. P. & Eddy, S. R. Transcription factors that convert adult cell identity are differentially Polycomb repressed. *PLOS ONE* **8**, e63407 (2013).
57. Orsi, G. A. *et al.* High-resolution mapping defines the cooperative architecture of Polycomb response elements. *Genome Res.* **24**, 809–820 (2014).
58. Frey, F. *et al.* Molecular basis of PRC1 targeting to Polycomb response elements by PhoRC. *Genes Dev.* **30**, 1116–1127 (2016).
59. Sipos, L., Kozma, G., Molnár, E. & Bender, W. *In situ* dissection of a Polycomb response element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **104**, 12416–12421 (2007).
60. Kozma, G., Bender, W. & Sipos, L. Replacement of a *Drosophila* Polycomb response element core, and *in situ* analysis of its DNA motifs. *Mol. Genet. Genomics* **279**, 595–603 (2008).
61. Busturia, A. *et al.* The MCP silencer of the *Drosophila* Abd-B gene requires both Pleiohomeotic and GAGA factor for the maintenance of repression. *Development* **128**, 2163–2173 (2001).
62. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20**, 1147–1155 (2013).
63. Blackledge, N. P., Rose, N. R. & Klose, R. J. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat. Rev. Mol. Cell Biol.* **16**, 643–649 (2015).
64. Francis, N. J., Follmer, N. E., Simon, M. D., Aghia, G. & Butler, J. D. Polycomb proteins remain bound to chromatin and DNA during DNA replication *in vitro*. *Cell* **137**, 110–122 (2009).
65. Anvar, Z. *et al.* ZFP57 recognizes multiple and closely spaced sequence motif variants to maintain repressive epigenetic marks in mouse embryonic stem cells. *Nucleic Acids Res.* **44**, 1118–1132 (2016).
66. Guo, A. M., Sun, K., Su, X., Wang, H. & Sun, H. YY1TargetDB: an integral information resource for Yin Yang 1 target loci. *Database (Oxford)* **2013**, bat007 (2013).
67. Waddington, C. H. *The Strategy of the Genes: a Discussion of Some Aspects of Theoretical Biology* (Allen & Unwin, 1957).
68. Chen, F. *et al.* Prenatal retinoid deficiency leads to airway hyperresponsiveness in adult mice. *J. Clin. Invest.* **124**, 801–811 (2014).
69. Foong, R. E. *et al.* The effects of *in utero* vitamin D deficiency on airway smooth muscle mass and lung function. *Am. J. Respir. Cell. Mol. Biol.* **53**, 664–675 (2015).
70. Lelièvre-Pégorier, M. *et al.* Mild vitamin A deficiency leads to inborn nephron deficit in the rat. *Kidney Int.* **54**, 1455–1462 (1998).
71. Greally, J. M. & Jacobs, M. N. *In vitro* and *in vivo* testing methods of epigenomic endpoints for evaluating endocrine disruptors. *ALTEX* **30**, 445–471 (2013).
72. Grün, F. *et al.* Endocrine-disrupting organotin compounds are potent inducers of adipogenesis in vertebrates. *Mol. Endocrinol.* **20**, 2141–2155 (2006).
73. Kirchner, S., Kieu, T., Chow, C., Casey, S. & Blumberg, B. Prenatal exposure to the environmental obesogen tributyltin predisposes multipotent stem cells to become adipocytes. *Mol. Endocrinol.* **24**, 526–539 (2010).
74. Woodsworth, D. J., Castellarin, M. & Holt, R. A. Sequence analysis of T cell repertoires in health and disease. *Genome Med.* **5**, 98 (2013).
75. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–1439 (2014).
76. 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).
77. Adalsteinsson, B. T. *et al.* Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PLOS ONE* **7**, e46705 (2012).
78. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).
79. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
80. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
81. Wijetunga, N. A. *et al.* The meta-epigenomic structure of purified human stem cell populations is defined at *cis*-regulatory sequences. *Nat. Commun.* **5**, 5195 (2014).
82. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
83. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
84. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
85. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
86. Zhao, M. *et al.* Distinct epigenomes in CD4$^+$ T cells of newborns, middle-ages and centenarians. *Sci. Rep.* **6**, 38411 (2016).
87. Qi, Q. *et al.* Diversity and clonal selection in the human T cell repertoire. *Proc. Natl Acad. Sci. USA* **111**, 13139–13144 (2014).
88. Ulahannan, N. & Greally, J. M. Genome-wide assays that identify and quantify modified cytosines in human disease studies. *Epigenetics Chromatin* **8**, 5 (2015).
89. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
90. 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).
91. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
92. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLOS Genet.* **1**, e78 (2005).
93. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
94. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
95. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLOS Genet.* **11**, e1004857 (2015).
96. Lappalainen, T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* **25**, 1427–1431 (2015).
97. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
98. Waddington, C. H. The epigenotype. 1942. *Int. J. Epidemiol.* **41**, 10–13 (2012).
99. Waddington, C. H. *Organisers and Genes* (Cambridge Univ. Press, 1940).
100. Li, Y. & Sasaki, H. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. *Cell Res.* **21**, 466–473 (2011).
101. Gendrel, A.-V. & Heard, E. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annu. Rev. Cell Dev. Biol.* **30**, 561–580 (2014).
102. Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* **330**, 612–616 (2010).

Acknowledgements
The authors thank S. Henikoff for sharing his insights and acknowledge his creation of the useful term epi + genetics, as well as R. Satija for advice in single-cell RNA sequencing data analysis.

Competing interests statement
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FURTHER INFORMATION
10x Genomics reference single-cell RNA-seq data: https://support.10xgenomics.com/single-cell/datasets
Cibersort software: https://cibersort.stanford.edu/index.php
Gene Expression Omnibus data from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65515
Seurat software: http://satijalab.org/seurat/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF