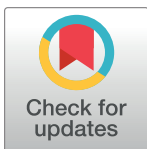# PLOS GENETICS

# Using epigenomics to understand cellular responses to environmental influences in diseases

**Julia J. Wattacheril[1], Srilakshmi Raj[2], David A. Knowles[3,4,5], John M. Greally[2]***

**1** Department of Medicine, Center for Liver Disease and Transplantation, Columbia University Irving Medical Center, New York Presbyterian Hospital, New York, New York, United States of America, **2** Division of Genomics, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, United States of America, **3** New York Genome Center, New York, New York, United States of America, **4** Department of Computer Science, Columbia University, New York, New York, United States of America, **5** Department of Systems Biology, Columbia University, New York, New York, United States of America

* john.greally@einsteinmed.edu

## Abstract

It is a generally accepted model that environmental influences can exert their effects, at least in part, by changing the molecular regulators of transcription that are described as epigenetic. As there is biochemical evidence that some epigenetic regulators of transcription can maintain their states long term and through cell division, an epigenetic model encompasses the idea of maintenance of the effect of an exposure long after it is no longer present. The evidence supporting this model is mostly from the observation of alterations of molecular regulators of transcription following exposures. With the understanding that the interpretation of these associations is more complex than originally recognised, this model may be oversimplistic; therefore, adopting novel perspectives and experimental approaches when examining how environmental exposures are linked to phenotypes may prove worthwhile. In this review, we have chosen to use the example of nonalcoholic fatty liver disease (NAFLD), a common, complex human disease with strong environmental and genetic influences. We describe how epigenomic approaches combined with emerging functional genetic and single-cell genomic techniques are poised to generate new insights into the pathogenesis of environmentally influenced human disease phenotypes exemplified by NAFLD.

## Introduction

Many human diseases have a clear environmental contribution. For decades, it has been assumed that the environment can influence the regulation of gene expression through "epigenetic" mechanisms, which can be interpreted as meaning the molecular regulators of transcription, such as DNA methylation and chromatin states. The extracellular environmental cue acting to modify epigenetic states has even been given its own name, the "epigenator" [1]. Evidence supporting this model has mostly come from findings of altered epigenetic patterns following environmental exposures. It is a widely accepted model that epigenetic alterations are the primary mediators of environmental influences, potentially propagating these effects

long after the cessation of exposure because of the assumed property that epigenetic regulation can self-propagate through cell division.

We have, however, become more critical in our interpretation of genome-wide studies of epigenetic mediators, described as epigenome-wide association studies (EWAS). We have assumed that a change in a transcriptional regulator like DNA methylation, found when comparing pools of cells, reflects individual cells undergoing reprogramming. This interpretation is now recognised to be too simplistic and ignores other ways the same outcome can occur. Evidence linking environmental exposures with epigenetic modifications needs to be reassessed as part of this updated perspective.

In this review, we take a fresh look at the evidence for epigenetic mediation of environmental influences. We use nonalcoholic fatty liver disease (NAFLD) including nonalcoholic steatohepatitis (NASH), as our disease focus. NAFLD is an excellent paradigm for a common disease about which much is known in terms of environmental and genetic risk factors. We describe how NAFLD has been studied to date with exploratory EWAS as an example of how the over-simplified interpretation of results may be misleading.

The constructive way of thinking about the nonepigenetic effects influencing our interpretation of EWAS is that they are not merely impediments to interpretation, but instead offer alternative insights into disease pathophysiology. We will review how a systematic change in either a cell subtype proportion or a genetic variant can lead to a DNA methylation, transcriptional or other changes that we collectively refer to as "molecular genomic" processes, in preference to the ambiguous term "epigenetic." Such systematic changes reflect cell subtype or DNA variant associations with the exposure or disease and would be valuable to understand as potential contributors to the phenotype. The study of the relationship between genetic and molecular genomic variation is typically described as functional genomics and encompasses the identification of sequence variants that influence molecular genomic processes, including gene expression levels (expression quantitative trait loci (eQTLs)), DNA methylation (meQTLs), and chromatin accessibility (caQTLs). Some functional variants are only revealed following an environmental exposure, uncovering a potential way that individuals can differ in their responses to the same environmental challenges [2]. Furthermore, now that we are in the era of increasingly sophisticated single-cell genomic assays, we are finding unexpected heterogeneity within canonical cell types, and effects of functional sequence variants that are restricted to subtypes of cells [3], representing unprecedented insights into how environmental influences act within a tissue.

It is now timely to rethink how we use molecular genomic assays to understand how the environment influences cells, and how this influence can vary because of differences in the DNA sequence between individuals. NAFLD is a useful focus, representing a disease of major public health importance with strong environmental influences, multiple extrahepatic manifestations, and variable interindividual susceptibility to disease development and progression. For example, the fibrosis that is part of more advanced NAFLD is influenced by glucose as an environmental agent and represents a major target for intervention to prevent the morbidity and mortality of end-stage liver disease. The goal of this review is to prompt innovative ways of thinking about how to gain therapeutic insights into environmentally driven diseases such as NAFLD, using novel molecular genomic and cellular approaches.

## The epigenome and the environment

### Foundational studies

In this review, we define the environment as the influences extrinsic to and influential for the cell, tissue, or organism. Epigenomic studies can be defined as genome-wide mapping of

molecular mediators of transcriptional regulation, given that such mediators have been traditionally described to have "epigenetic" properties [4].

An inherent property of epigenetic regulation of transcription is its reversibility—if a gene is activated, it can be subsequently switched off, or vice versa. Molecular regulators of transcription such as chromatin states, DNA modifications, and transcription factor (TF) activity alter as part of cellular differentiation, demonstrating this molecular malleability.

The ability of the same DNA to behave in different ways in different cell types appears to indicate a layer of information residing on top of the DNA sequence, for which the back-translation of epi- (above, upon) -genetics (DNA sequence) seemed a useful descriptive term. It has also become commonplace to use the word "epigenetic" to describe events that appear to occur variably in cells and organisms in ways that are not mediated by DNA sequence variability, making "epigenetic" synonymous with "nongenetic" in a further use of the term.

The idea of a regulatory layer of information, acting to influence cellular properties without causing DNA changes, puts molecular epigenetic processes in the spotlight as potential mediators of differences between cells or organisms that share the same DNA sequence. The puzzle of discordance of monozygotic twins for conditions that clearly had a genetic contribution prompted speculation that the environment could be different (nonshared) for each twin and that this difference in exposures caused only one to develop the disease [5]. Early studies of DNA methylation showed what appeared to be a progressive difference with age between monozygotic twins [6], an association that suggested a mechanism for this discordance.

Earlier work also supported DNA methylation responding to environmental influences. In 1984, rats were fed a methyl-deficient diet with the idea that it would promote neoplasia. The diet was associated with the development of decreased DNA methylation in the livers of the animals [7]. An influential model was the "viable yellow" mouse, animals with a mutation with variable effects on phenotype, ranging from no apparent phenotype at all (pseudoagouti) to a phenotype of yellow fur, obesity, hyperinsulinaemia, and an increased rate of malignancies. In 1998, Wolff showed that feeding pregnant dams a diet designed to increase DNA methylation increased the proportion of pseudoagouti offspring [8], subsequently shown to be associated with increasing DNA methylation at the mutation site, the insertion of an IAP retroelement upstream of the *Nonagouti* gene [9,10]. This was a fascinating paradigm, indicating that maternal diet during pregnancy could influence the eventual adult phenotype of her offspring, and represents part of the foundation for the field of the Developmental Origins of Health and Disease (DOHaD) [11]. The idea that the organism retains a memory of a past exposure involved a different use of the term "epigenetic," to describe molecular regulatory processes that were heritable from parent to daughter cells [12]. Intrauterine nutritional deprivation during the Dutch Famine of 1944–1945 was described to be associated with obesity in male offspring in young adulthood, limited to those exposed during the first half of pregnancy, whereas exposure in the last trimester and the first months of life was found to have significantly reduced rates of obesity [13]. While the epidemiologists performing these studies were concerned that there could be some confounding effects on these findings, such as the greater fertility and fecundity of women from higher socioeconomic classes [14,15], the model emerged of a foetus adapting to a stressful environment in utero and retaining this memory of exposure in a way that is maladaptive postnatally. This finding became a model for epigenetic mediation of environmental exposures long after the environmental exposure was no longer present.

It would be misleading to give the impression that the only discoveries in this field were being made in mammals or in more economically developed countries. In 1966, Madeleine Charnier, working at the University of Dakar, Senegal, found that sex determination in a reptile species (the rainbow Agama lizard, *Agama agama*) depended on the temperature at which the embryo develops [16]. Temperature-dependent sex determination occurs in many amphibians and fish, another example of a phenotype dependent on environment and not

determined by DNA differences. Plant biology was revealing numerous striking examples of exposures apparently being "remembered" by the organism, such as vernalisation, the response of plants to the prolonged cold exposure of winter, which was found to induce the silencing of the *FLC* gene [17], thus maintaining a memory of the past exposure.

The result of these and other observations was a model of environmental influences acting through "epigenetic" regulatory mechanisms, self-propagating their new patterns of organisation to maintain a memory of the past exposure.

## How molecular regulators of transcription may respond to the environment

### DNA and histone modifications

"Epigenetic" molecular regulators of transcription are numerous. The pattern throughout most of the mammalian genome is one of DNA wrapped around octamers of histones to form a nucleic acid:protein complex called a nucleosome. Most of the cytosines located at CG dinucleotides genome-wide are modified by the addition of a methyl group to form 5-methylcytosine. Where the genome departs from these default patterns typically represents the locations of regulatory elements, where sequence-specific proteins like TFs bind to the DNA. The histones in the flanking nucleosomes acquire patterns of posttranslational modifications (PTMs) that flag these sites as promoters, enhancers, or other regulatory elements. Further distinctive patterns of DNA and histone modifications are found at loci undergoing transcription, while repressive histone modifications or histone variants represent a more macro level of organisation, defining heterochromatic regions of the genome.

DNA methylation remains the paradigm for a transcriptional regulator that can maintain a biochemical memory through cell division, as 5-methylcytosine can be propagated from a parent chromatid to both daughter chromatids. There is also evidence for propagation of a repressive histone modification (histone H3 lysine 27 trimethylation, H3K27me3) through a number of generations of the nematode *C. elegans* [18]. TFs appear to remain bound to their DNA target sites through DNA replication, a phenomenon described as "mitotic bookmarking" [19], representing a further way that memory of the molecular organisation of the chromatid in a parent cell can be passed to replicated chromatids in daughter cells.

We assume that a cell is faithfully able to propagate its transcriptional regulatory organisation to daughter cells as a way of transmitting the memory of a prior environmental exposure. However, there is another factor worth considering—it appears that some histone modifications and histone variants can directly mediate responses to the environment. In **Table 1,** we summarise some of these possible environmental influences and molecular regulatory responses.

### Two molecular genomic problems to resolve: Cellular memory and sequence specificity

These observations indicate that the environment may be able to act through regulators of chromatin and modifiers of DNA to influence gene expression. While these are intriguing findings and provide clear candidates for the mediation of environmental effects on transcriptional regulation, there are two problems that remain to be overcome before invoking any as primary mediators of long-term, stable cellular reprogramming.

The first is cellular memory. With the exception of DNA methylation, there are no clear biochemical mechanisms for self-propagation of the specific molecular events in **Table 1** to daughter chromatids following cell division. What needs to be invoked to support the idea that these chromatin constituents have long-term consequences is a two-step model, one involving an

**Table 1. Examples of environmental exposures that directly affect molecular regulators of transcription.**

| Environmental exposure | Molecular regulatory changes | Effect | Genomic location of effects | Reference |
|---|---|---|---|---|
| Glucose (through hexosamine biosynthetic pathway) | GlcNAcylation of histones | Unclear | Unclear | (See footnote*) |
| | GlcNAcylation of EZH2 | Enhances protein stability and enzymatic activity | Heterochromatin (H3K27me3) | [20] |
| | GlcNAcylation of TET enzymes | Enhances TET1 activity | Loci with 5-hydroxymethylcytosine | [21] |
| Lactic acid (anaerobic metabolism) | Lactylation of histones | Possibly gene activation | Active gene promoters | [22] |
| Serotonin | Serotonylation of histones | Possibly enhanced gene expression | Gene promoters | [23] |
| Ethanol | Acetylation of histones | Induction of expression of genes involved in signal transduction and learning and memory | New loci of H3K9ac and H3K27ac formed (promoters/enhancers) | [24] |
| NAD+ metabolism | Enhanced SIRT1 activity | Protein (histone) deacetylation | Unclear | (See footnote**) |
| | MacroH2A | Reverse effect: macroH2A1 binds to PARP-1 and limits its availability in the cell | None expected | [25] |
| Hypoxia | G9A | Increased protein stability | Gene promoters | [26] |
| | KDM5A | Inhibition of activity | Loci with H3K4me3 and H3K36me3 | [27] |
| Short chain fatty acids (e.g., produced by gut fermentation of dietary fibre) | Inhibition of histone deacetylases (HDACs) | Increased histone crotonylation | Histone crotonylation is enriched at gene promoters | [28] |
| Dietary folic acid, vitamins B$_6$ and B$_{12}$ | Increased production of S-adenosyl methionine (SAM) | Increased methylation of DNA and histones | Genome-wide | [29] |
| Vitamin C | TET cofactor | Increased 5-hydroxymethylcytosine, decreased 5-methylcytosine | Genome-wide | [30] |

*While GlcNAcylation of histones is often described as an example of a metabolic sensor in chromatin, the concern has been raised that GlcNAcylation does not occur on mammalian histones [31]. More information appears to be needed to resolve this issue.

**The relationship between NAD+ and SIRT1 activity has been debated for a long time, as described in this review [32].

https://doi.org/10.1371/journal.pgen.1010567.t001

initial environmental perturbation, followed by the maintenance of a new equilibrium of transcriptional regulators that have the capacity to maintain their patterns long term and through cell division. For example, is it possible that a hyperglycaemic event that increases O-GlcNAcylation acutely then induces EZH2 activity nearby in the genome, causing de novo patterns of formation of H3K27me3 that are then maintained, even after the hyperglycaemia subsides?

The second problem is that these global regulators of transcription lack sequence specificity and therefore do not have the ability to choose specific loci for selective activity. Once again, a separate mediators has to be active initially, for example, when short chain fatty acid exposure causes specific subgroups of promoters to be selected for increased histone crotonylation [28]. Later, we make the case that this is likely to be mediated by sequence-specific TFs establishing regulatory patterns in response to environmental challenges, patterns that are passed on to the global regulators of transcription, allowing TFs to have a primary role in the model of epigenomic responsiveness to environmental stimuli.

## Epigenetic association studies

### Epigenetic studies and NAFLD

When attempting to understand whether epigenetic changes occur associated with traits or disease phenotypes, an EWAS has been the typical approach. As introduced in **Box 1**, an

### Box 1. The epigenome-wide association study (EWAS).

The epigenome-wide association study (EWAS) refers to the testing of samples from individuals with a phenotype or exposure of interest using an assay testing the molecular regulators of gene transcription. This broad picture narrows significantly in practice, as most EWAS involve testing human blood leukocytes and the use of DNA methylation microarrays to survey across the entire genome. "Epigenome-wide" should not be taken to indicate comprehensive coverage of the genome, as microarrays typically represent no more than a few percent of the CG dinucleotides at which DNA methylation occurs. The groups are compared for consistent differences in DNA methylation levels between them, a positive outcome reflected by loci showing changes that are statistically significant. The underlying hypothesis driving these studies is the assumption that the phenotype being studied involves a reprogramming of the transcriptional regulation of the cell, not DNA sequence changes. By identifying DNA methylation differences between the groups, we not only gain evidence for this model, we also find the genes involved in changing the properties of the cell as part of the development of the phenotype. As will be discussed in the main text, it is not straightforward to interpret the results of these studies, for a number of reasons. The positive view is that the factors that make EWAS difficult to interpret, if understood, allow insights into the development of the phenotype, but at the expense of the hypothesis of cellular reprogramming.

EWAS is generally performed on multiple individuals in a test group (with the phenotype) and a control group that is matched for potentially influential variables like age, sex, and ancestry. The same tissue type is sampled in all individuals, and an epigenome-wide assay is performed, usually studying DNA methylation at numerous loci in the genome, ranging from tens of thousands to millions of sites. For environmental studies, testing the same individuals in conditions when exposed and not exposed is an alternative study design that eliminates the influence of genetic variability, as discussed later, and it may be possible to quantify the exposure so that you do not have to compare two groups, but can instead correlate exposure as a continuous variable with the DNA methylation changes. When there is a locus or set of loci that has a difference of DNA methylation that appears statistically nonrandom, this represents a positive outcome for the EWAS. Generally, the result of a positive EWAS is one of numerous loci showing significant differences in DNA methylation, which generally leads to an attempt to interpret why this group of genomic regions underwent regulatory changes, involving linking each site with a gene, and then looking for a coherence of properties of these genes, using gene ontology or pathway information.

We are using NAFLD as the focal point for the discussions to follow. NAFLD represents a major complication of the obesity epidemic, with NAFLD now the most common form of liver disease worldwide [33,34], estimated to affect 6 to 30 million people in the United States, including 600,000 who have advanced to developing cirrhosis [35]. A spectrum of histological stages characterizes the disease, ranging from simple fat accumulation (or steatosis) to an inflammatory phenotype, NASH, which can progress to involve fibrosis and cirrhosis [36]. Clinical outcomes with cirrhosis include decompensation, portal hypertension, liver transplantation, hepatocellular carcinoma, and death. Long-term follow-up of NAFLD patients confirms that NASH patients have a higher risk of liver-related mortality than non-NASH patients [37]. The economic burden of NASH is significant—the lifetime costs in the USA for NASH patients alone in 2017 was estimated to exceed US$200 billion [38].

The pathogenesis of NAFLD involves both environmental exposures and genetic predisposition [39,40]. The causative environmental exposures are generally involved in causing obesity, such as the Western diet, but distinct gut microbiome complements in NAFLD patients are also invoked as possible contributory exposures [41]. It comes as no surprise that DOHaD models that are linked to obesity are also linked to NAFLD, with exposures during foetal and early neonatal life to maternal under- and overnutrition, excess glucocorticoids, and environmental pollutants linked to offspring NAFLD [42]. Whether these exposures act primarily to cause offspring obesity, with NAFLD as a secondary consequence, or whether early developmental programming predisposes to liver damage independently remains unclear.

Genetic susceptibility is a factor of importance in the development of NAFLD, which has a strong heritable component and involves several known loci [43]. NASH can develop in lean individuals, especially when carrying genetic risk alleles [44]. DNA sequence changes identified through genome-wide association studies (GWAS) have been causally associated with NAFLD [45–48], discussed in more detail below.

NAFLD helps to illustrate an important point—the environmental exposure extrinsic to the organism may not be the same exposure experienced by the cells in the affected tissue (**Fig 1**). In the case of NAFLD, the exposure extrinsic to the cell are the lipotoxic lipid species that influence hepatocytes in a poorly understood stress model. More relevant is the extrinsic exposure that causes NAFLD to progress to significant fibrosis, exemplified by high-fructose corn syrup [49,50], while the equivalent cellular exposure inducing fibrosis is the inflammatory cytokine TGFß [51].

We describe below several studies demonstrating tissue- and genomic region-dependent variation in DNA methylation in NAFLD, the foundation for proposals that DNA methylation has a role in the pathogenesis of NAFLD [52]. There have been several human EWAS performed on cohorts of patients with NAFLD that are worth examining as examples of a broader field of use of epigenomic assays to study environmental effects. Studies involving DNA methylation assays were chosen as this is at present by far the most commonly used epigenomic assay in EWAS. In **Table 2,** we list eight studies involving DNA methylation studies of hepatic
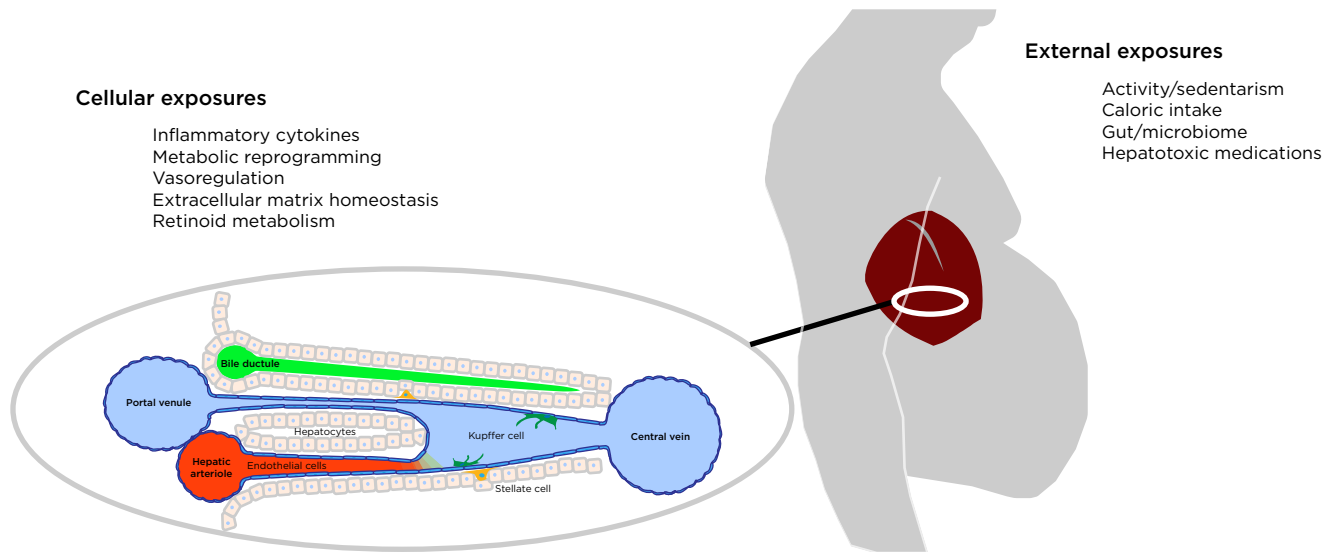


**Fig 1. Environmental exposures associated with disease are not necessarily the same when considering those extrinsic to the organism, and those directly acting on the cellular microenvironment, depicted with examples relevant to NAFLD.**

**Table 2. Examples of prior EWAS performed in NAFLD and NASH.**

| Phenotype(s) | Number of individuals studied | Tissue type(s) | Epigenomic assay(s) | Citation |
|---|---|---|---|---|
| NAFLD: mild | 33 | Liver | DNA methylation | [53] |
| NAFLD: severe | 23 | | Gene expression | |
| Normal controls | 18 | Liver | DNA methylation | [54] |
| Healthy obese | 18 | | Gene expression | |
| NAFLD | 12 | | | |
| NASH | 15 | | | |
| Serum GGT, AST, and ALT levels | 731 (discovery) | Peripheral | DNA methylation | [55] |
| Hepatic steatosis (ultrasound of replication cohort) | 719 (replication) | blood | | |
| Healthy obese | 35 | Liver | DNA methylation | [56] |
| NAFLD | 34 | | Gene expression | |
| NASH | 26 | | | |
| Normal controls | 30 | Peripheral | DNA methylation | [57] |
| NAFLD | 35 | blood | | |
| NAFLD: mild | 35 | Liver | DNA methylation | [58] |
| NAFLD: advanced | 25 | | | |
| Normal controls | 15 | Liver | DNA methylation | [59] |
| NAFLD and fibrosis | 15 | | Gene expression | |
| Healthy controls | 30 | Peripheral | DNA methylation | [60] |
| NAFLD | 18 | blood | | |
| NASH | 17 | | | |
| NAFLD: severe fibrosis | 119 | Liver | DNA methylation | [61] |
| NAFLD: no fibrosis | 206 | | | |

EWAS, epigenome-wide association studies; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis.

steatosis. None of these include extrahepatic phenotyping and are instead focused solely on liver disease.

## EWAS design issues

These studies can be used to help illustrate some of the strengths and weaknesses of epigenetic association studies, as have been extensively reviewed in prior publications [62–65]. We break down some of the major issues in the following five categories.

**Surrogate tissue sampling.** Some human tissues are very accessible (e.g., peripheral blood leukocytes, buccal epithelium, hair follicles), but many common diseases are mediated by organs that would require sampling using invasive and risky procedures (e.g., liver, lungs, brain). A frequent question is whether an accessible tissue like peripheral blood leukocytes can report the epigenetic events occurring in an inaccessible tissue in the same individual [66,67]. What tends to be missing in EWAS projects using surrogate tissues is a clear rationale at the outset for the choice of the tissue used. If all that is required from the study is the development of a robust biomarker, in which DNA methylation changes reflect or predict a disease but without the need to understand why the DNA methylation change occurred, the use of surrogate, easily accessible tissues is highly desirable. If the question is whether the mechanism of the disease can be revealed through studies of a tissue like peripheral blood, that is when the rationale should be explicit—is the assumption that all cells in the body are changing their transcriptional regulatory patterns in the same way, both in the accessible and inaccessible tissues? Three of the NAFLD studies in **Table 2** use peripheral blood leukocyte DNA for DNA

methylation assays [55,57,60], making these studies examples of the use of surrogate reporter tissues. For their results in peripheral blood leukocytes to be informative about the physiology of cells in the liver, the genes and pathways would have to be active in both of these very distinctive tissue types and dysregulated in the same manner in response to an environmental exposure that may or may not be comparable in these distinct tissues. At present, such an explicit description of the rationale for the study or any discussion about whether this model is realistic is typically absent from reports of epigenetic studies of surrogate tissues.

**Cohort sizes.** The median number of individuals in each cohort in **Table 2** is 65, falling to 61.5 in the studies using liver samples. Obtaining well-characterised, high-quality liver samples for molecular studies is very challenging, reflected by these limited numbers. The question about how many samples are needed to power an epigenetic association study is frequently raised [68,69]. There are three major interacting variables that can be considered—the degree of change of DNA methylation, the number of sites tested, and the number of samples compared. When you use fewer samples, you can only confidently attribute changes of DNA methylation that are of greater magnitude and at fewer sites. It is a reasonable conclusion that these studies in **Table 2** detected only a limited subset of the total likely number of DNA methylation changes occurring in these individuals.

**Reverse causation.** A further issue is whether the DNA methylation changes can be assumed to cause the phenotype, or whether they are caused by the phenotype, the latter being an example of "reverse causation." This is a limitation inherent to cross-sectional study designs in which one group compared already has the phenotype [64]. Examples of reverse causation include obesity-related phenotypes that have been found to change DNA methylation of peripheral blood leukocytes (studying body mass index (BMI) [70,71] or blood lipid profiles [72]). Reverse causation represents another way that we can overinterpret epigenetic association studies, by making the incorrect assumption that the DNA methylation changes are causing the phenotype.

It has been proposed that a good way of addressing this issue of confounded interpretation of reverse causation is through the use of a more difficult study design: longitudinal sampling of the same patients over time [73]. The Ahrens study in **Table 2** accomplished this in a clever manner, sampling liver before and after bariatric surgery [53]. They found DNA methylation changes to be partially reversible at a subset of loci when liver biopsies were compared after bariatric surgery and associated dramatic weight loss (averaging 40 kg per person). By including a "healthy obese" group in their study, the possibility that the environmental exposure of obesity represents an independent and confounding influence on DNA methylation in the liver is addressed, allowing a focus on a subset of loci that is more likely to be involved in disease pathogenesis [54].

**Cell subtype proportional composition.** It is possible for cell subtype proportions to vary within the tissue studied and between the groups tested and lead to a result showing changes of DNA methylation without any cells in the samples having undergone molecular reprogramming. For a DNA methylation change to be attributable to this influence, the cell subtype proportional change has to be consistently differing between the groups—in other words, the different proportion of a specific cell subtype needs to be nonrandomly present across the individuals in the test group compared with the controls.

Of the studies in **Table 2**, the Nano group [55] applied the *minfi* software package [74] to account for six leukocyte subtypes in analysing their results of peripheral blood studies, a mainstream approach in EWAS at present. Had cell subtypes been studies in liver itself, multiple changes would be expected, confirmed by the Johnson study that used the *EpiDISH* cell subtype deconvolution approach [75], revealing a decrease in epithelial cells and increases in immune cells, in particular natural killer (NK) T lymphocytes, with the progression of fibrosis

in their liver samples. In the earlier stages of NAFLD, reprogramming of cellular properties should be reflected by the accumulation of large amounts of lipids in the cytoplasm of hepatocytes, manifested histologically by the displacement of the nucleus to a peripheral intracellular location. Cell subtype compositional changes would be likely to include the influx of inflammatory cells into the liver parenchyma with disease progression, and the transdifferentiation of hepatic stellate cells to the myofibroblasts that produce the extracellular matrix proteins that cause fibrotic scarring of the organ. The cells composing the liver affected by NAFLD or NASH are therefore substantially different to those in normal, healthy livers, both in terms of their innate properties and the representations of different subtypes. The development of single-cell transcription profiles from these kinds of tissues is likely to generate new ways of testing for both reprogramming and cell subtype proportional changes in disease like NAFLD [76].

**Genetic effects on DNA methylation.** Finally, the influence of genetic sequence variation between the individuals studied needs to be considered. It is now widely recognised that DNA sequence variability can be associated with DNA methylation differences between individuals, defining functional variants referred to as meQTLs [77]. These are usually revealed when both genotyping and DNA methylation studies are performed on cohorts of individuals, testing whether differences in DNA methylation (the quantitative trait) are associated with the presence of different alleles for a variant at a locus within the flanking 1 Mb. In **Fig 2,** we illustrate the example of a locus with a C/C genotype on both alleles in 81% of people in the population, a heterozygous C/A in 18% and a homozygous A/A in 1%. DNA methylation at a nearby site averages 40% in the C/C individuals, 30% in the C/A, and 20% in the A/A people, revealing a trend in DNA methylation associated with genotype. It may not be the variant itself with the C:A polymorphism that mediates the effect on the DNA methylation, as that locus is transmitted with a substantial amount of flanking DNA, representing a haplotype in which a separate functional variant could also be carried.

When the proportion of DNA methylation variation attributable to DNA sequence variation has been estimated, using different assays, tissue types, and ways of making these estimates, very substantial effects have been found, estimated to vary between 14% and 80% [78–85]. Of the studies in **Table 2**, that by Nano and colleagues was distinctive for identifying which DNA methylation changes were attributable to DNA sequence variation. They found four sites of distinctive DNA methylation that survived their other rigorous filtering criteria, of which three were attributable to DNA sequence variation, leaving them with a single locus near the *SLC7A11* gene where DNA methylation appeared to be associated with hepatic steatosis on its own and not genetic variability between the groups tested [55].

In performing these kinds of EWAS, the ancestries of the patients studied should be described. Two of the **Table 2** studies were on Han Chinese [57,60]; one was on "Caucasian women" [59], and one on Japanese patients [58]. The effect of DNA sequence variability on DNA methylation variability is so strong that ancestry has been shown to be predictable from DNA methylation assays [86]. Furthermore, with clear differences in susceptibility to NAFLD and NASH as complications of obesity between Hispanic and Black American patients [87,88], not accounting for ancestry in study design could be a strong influence on the results obtained if the affected individuals are disproportionately of one ancestral category. The meQTLs in a specific tissue type can differ between individuals of different genetic ancestries [89,90], which could cause the appearance of DNA methylation differences associated with a disease if the individuals studied in the disease group are disproportionately from distinct ancestral origins to the controls. The variability in DNA methylation and gene expression can even occur at fine-scale geographic levels, as demonstrated in a study identifying within-island DNA methylation and gene expression differences in individuals of diverse ancestry in Indonesia [91].
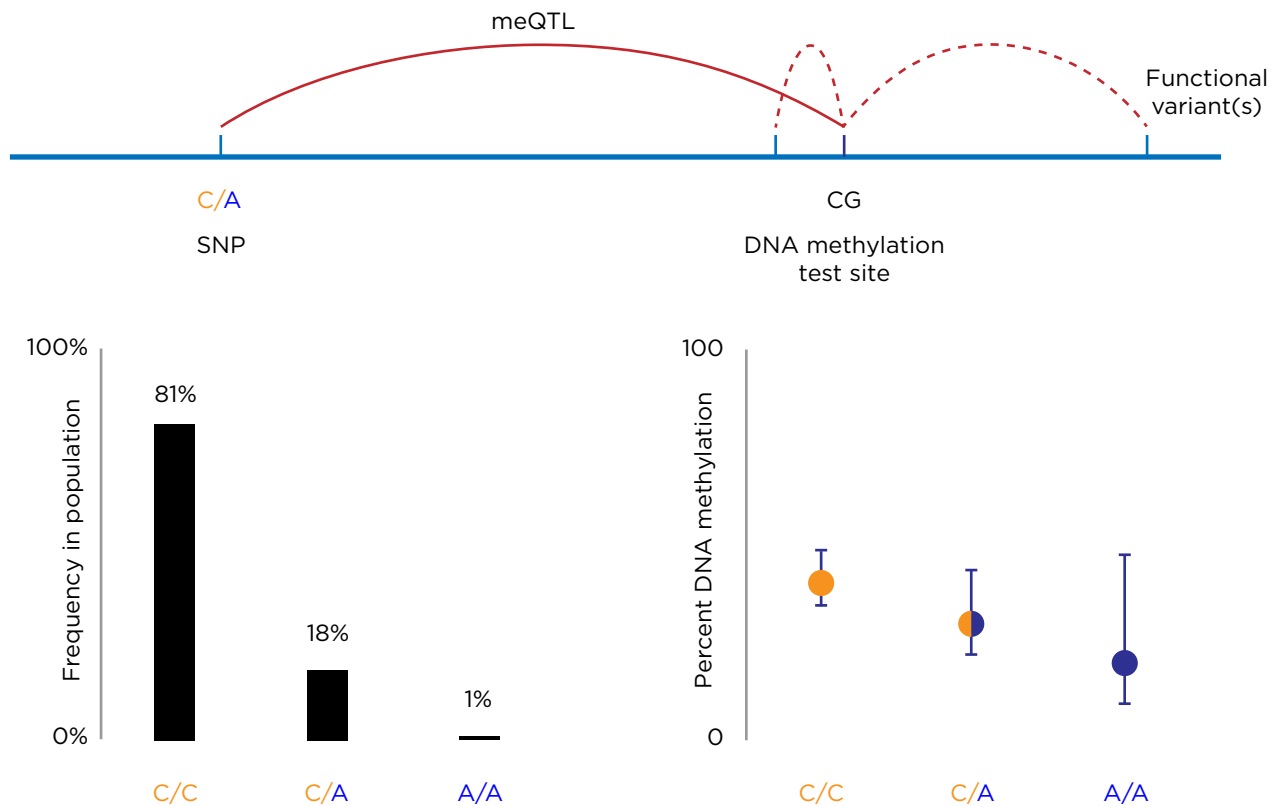
**Fig 2. An illustration of a methylation quantitative trait locus (meQTL).** These are identified by finding a difference in DNA methylation between individuals that correlates with having one or other allele for a DNA sequence variant, in this case showing a variant on the same chromosome (a *cis*-meQTL). A significant change in DNA methylation associated with these differing allelic states defines the meQTL. Whether the meQTL causes the DNA methylation change is less certain—within the haplotype containing the meQTL and the locus where DNA methylation was tested will include other DNA sequence variants, one or more of which could be directly influencing the DNA methylation as a functional variant.

https://doi.org/10.1371/journal.pgen.1010567.g002

Finally, as mentioned above, if we know something about the genetic susceptibility affecting individuals within the cohort affected by the disease, that information is probably worth including. NAFLD has been found to be associated with protein-coding sequence variation, including the I148M missense variant in *PNPLA3* that compromises the individual's ability to hydrolyse triglycerides [92]. It would be unwise to assume that this subgroup of patients with NAFLD or NASH is experiencing the same environmental exposures at the tissue and cellular level as individuals with different or no recognisable pathogenic variants. Including protein-coding pathogenic variant information as a likely source of variability in EWAS is prudent as a major influence on phenotypic outcomes. It is likely that the pathogenic burden in the genome (genetic variation contributing to disease) is concentrated in loci of open chromatin in disease-relevant tissues [93]. In this "omnigenic model" of complex traits, genes with modest effects throughout the genome are influencing disease etiology through cell- and tissue-specific effects. This model supports the findings with NAFLD that genetic and epigenetic mechanisms interact with environment in a variable, tissue- and cell-dependent manner that introduces variability to EWAS outcomes.

The nine studies of **Table 2** are used to represent the broader field of EWAS. Like most EWAS, each study has its own strengths, as defined above, but equally no individual study can be said to have addressed all of the problems inherent to how EWAS are currently performed. The potential outcome that results from failing to account for these problems is that changes in DNA methylation could be overinterpreted, assumed to represent cellular reprogramming responding to an environmental provocation, and causative of the phenotype, whereas the DNA methylation change could instead be due to changes in cell subtype composition within the tissue, genetic differences between individuals, or the consequence of the hepatic phenotype.

## The long road from EWAS to individual disease prediction

Both GWAS and EWAS have been touted as having great promise for precision medicine, particularly for polygenic disorders such as NAFLD and NASH [94]. For polygenic traits, risk prediction through a risk score offers one such avenue. Typically, genetic risk scores (GRS; also described as polygenic risk scores (PRS)) are constructed to identify high-risk individuals based on GWAS results. The GRS is calculated as the weighted sum of the risk alleles for a trait in an individual, using weights determined by the best statistically powered GWA study for the trait. GRS have shown mixed utility in the ability to identify high-risk groups of individuals [95], and limited translation across ethnicities [96]. Recent NAFLD GWAS from the UK Biobank have identified >90 associated genomic loci, allowing the development of a GRS that successfully identified high-risk cases as having an odds ratio of 2.1 compared to individuals with the lowest NAFLD GRS score [94]. This difference is more modest than GRS for other diseases, such as the 3-fold higher risk for coronary artery disease [97].

DNA methylation risk scores (MRS) are proposed to capture the effects of the environment in generating a risk score for a phenotype [98]. Like a GRS, the MRS summarises information across multiple informative loci in the genome to generate a single score, but based on risks associated with DNA methylation values, not with genetic variants. There are numerous factors that limit the utility and widespread calculation of MRS. The first is that there are too few large-scale EWAS with concurrent genotyping to provide adequate external datasets to calculate MRS [98]. The DNA methylation arrays may also introduce bias that skews interpretability due to the microarray design. For traits such as NAFLD that involve strong environmental effects, the hope would be that MRS would capture these influences and would offer predictive power beyond sequence variation alone.

It is encouraging that the few genome-wide DNA methylation studies published in recent years suggest that blood DNA methylation measurements provide a stable and accurate assessment of risk when calculating MRS, for a range of phenotypes [99–102]. One recent study including the largest reported MRS cohort (*n* = 831) indicated that DNA methylation scores outperform baseline risk and GRS, improving imputation of 139 outcomes, compared with just 22 improved through GRS [103]. Many of these MRS replicated in external cohorts of different ethnicities and showed variable but robust replication across kidney-related traits in diverse populations. This remains an emerging area of research but looks to have room to improve when the accuracy of MRS is enhanced with increased sample sizes.

## Embracing the sources of variation

### Spurious associations reveal systematic changes

If it sounds daunting to consider the possibility that we need to go beyond the DNA methylation change and understand its cause in terms of cell subtype or DNA sequence variants, there is an important point to consider. For one of these influences to influence the results, sending

the DNA methylation in a specific direction so that it differs between groups, the spurious influence needs to be nonrandomly distributed between the groups. For example, if you are testing liver samples and there are more Kupffer cells in samples from one group compared with the other, this will cause the appearance of a DNA methylation change related to the loci that are distinctively methylated in that cell type. Likewise, in the situation of the meQTL above, where the C allele is associated with increased and the A allele with decreased DNA methylation, you will only see a resulting difference in DNA methylation between groups if the C allele is overrepresented in one group and not the other. These typically unrecognized influences, if studied, reveal cellular changes and genetic associations with the disease being studied, which represent potentially valuable insights into its pathogenesis.

## Examples of cell subtype changes following environmental exposures

It should be no surprise that environmental cues prompt responses that involve the composition of a tissue changing in terms of cell subtypes. Endocrine disrupting chemicals are characterised by altering sexual differentiation during development, a stark example where entire organs form differently in response to an environmental exposure.

We have previously highlighted [65] an interesting model of a cell subtype response to vitamin A deficiency during intrauterine development. In this mouse model, the associated tissue phenotype was quite subtle, involving hyperplasia of smooth muscle surrounding distal airways, established during lung development and persisting into adulthood, causing bronchoconstriction in response to a methylcholine challenge [104]. A very similar phenotype that recapitulates a component of the human asthma phenotype was found following prenatal deficiency of vitamin D [105]. While relatively subtle from a histological point of view, the change in cell subtype composition and organisation in the lung tissue was enough to cause a measurable phenotype.

EWAS of environmental exposures have also revealed cell subtype changes in peripheral blood. In a study of low-level intrauterine exposure to arsenic, DNA methylation of umbilical cord blood leukocytes was tested and analysed using an approach comparable to that described to be used by Nano group [55] earlier. These researchers found that arsenic exposure was associated with a higher proportion of CD8+ T lymphocytes [106]. The authors noted that the deconvolution approach based on DNA methylation profiles of adult reference cell types does not work as expected in cord blood, as was later systematically reviewed [107], and could therefore call their results into question. However, a separate study of arsenic exposure to adult mice showed an increase in CD8+ T cell proportions in bronchoalveolar lavage samples [108], suggesting that the exposure has consistent effects to increase this proportion of lymphocytes in different tissues.

A consistently robust association from blood leukocyte DNA methylation studies is with a history of cigarette smoking [109]. One of the loci at which DNA methylation is distinctive in smokers is at the gene for the GPR15 surface marker. When this specific marker was studied further, a complex picture emerged. Initially, it appeared that GPR15 defined a specific T lymphocyte subtype in blood that increased in proportion following smoking [110]. However, the same researchers followed up with a study that showed GPR15 to be present on many T lymphocyte subtypes following smoking, indicating that the protein is induced by cigarette smoke exposure in many cell subtypes and does not necessarily reflect a change in cell subtype proportions, as they had originally proposed [111]. It should be noted that DNA methylation studies of peripheral blood generate a robust biomarker of smoking, a very useful indicator of a relevant environmental exposure when studying a phenotype like lung cancer [112].

When performing studies on human liver tissue, given the increasing availability of single-cell RNA-seq data from both healthy [116] and diseased [117] livers, it should be possible to go back to the results of those studies from **Table 2** that included gene expression analyses of liver tissue [53,54,56,59] and get an indication of what proportion of the observed DNA methylation differences are due to cell subtype changes (**Box 2**). Rather than regarding this as undermining the results of an EWAS, such an additional layer of information potentially enhances insights into disease pathogenesis, by revealing cell subtype proportion changes occurring as part of the development or progression of the disease. Furthermore, if this kind of approach permits removal of DNA methylation changes that are due to cell subtype changes, the remaining signal is much more likely to represent the kind of cellular reprogramming events typically sought in an EWAS.

## Box 2. How can cell subtype proportions be measured?

The complete blood count and differential white cell measurement provides quantification of lymphocytes, granulocytes (neutrophils, eosinophils, and basophils), and monocytes. This represents a restricted group of major subtypes of white blood cells, recognising that there are many subtypes of lymphocytes in particular, and monocytes are also inherently heterogeneous [113]. If DNA methylation microarray data are available from a tissue and if subtypes of cells have been tested to identify loci within the microarray with distinctive DNA methylation patterns in the different cell subtypes, the data can be analysed in a way that allows estimation of the proportions of each cell type present. This technique has been developed primarily for white blood cells, testing a different set of cell subtypes than is reported by the clinical differential white cell count: granulocytes, monocytes, and four subtypes of lymphocytes, B cells, natural killer cells, and CD4 and CD8 T cells [114]. If gene expression data are available, a comparable approach to allow estimation of cell subtypes can be performed, a salient example being CIBERSORT, which allows 22 different white blood cell subtype proportions to be estimated [115]. When reference gene expression data have not been generated from isolated cell subtypes in a tissue, the results of single-cell transcription studies can instead be used [76], allowing cell subtype estimations to be broadened beyond blood to other organs and tissues.

If the vitamin A or vitamin D deficiency studies of the mouse lung described above were performed by a currently typical EWAS approach, sampling the bulk lung tissue from animals in each exposure group, testing DNA methylation patterns, and removing those changes attributable to cell subtype changes, we would be eliminating from further consideration the smooth muscle changes that mediate the bronchoconstriction phenotype in these animals. Likewise, the finding from DNA methylation studies of an increased proportion of CD8+ T lymphocytes following arsenic exposure is consistent with what has been found through immunological studies of the effects of toxicity of this heavy metal [118]. These cell subtype proportion changes occurring nonrandomly in patients with a disease are not, therefore, an artefact to discard, but instead represent an insight into disease pathogenesis, and should be harvested as useful information.

### Genetic variation modifying the response to an environmental exposure

Some of the earliest documentation that different people respond to the same environmental exposure in different ways may go back over 2,500 years to the apocryphal story of Pythagoras (570–495 BCE) apparently recognising that only some people developed a fatal reaction to fava beans. The acute haemolytic anaemia causing these deaths is now recognised to be due to the vicine and convicine in the bean inducing reactive oxygen species within cells. While this induction occurs in anyone eating fava beans, in the erythrocytes of individuals with the X-linked glucose-6-phosphate dehydrogenase (G6PD) deficiency, this exposure causes profound haemolysis, anemia, and death.

With the introduction of the 8-aminoquinoline antimalarial drugs in the 20th century, it was noted as early as 1926 that some individuals had fatal responses following their administration [119]. A 1952 study described pamaquine to be associated with haemolytic anaemia in a subset of patients tested: "It may be noted that all six acute hemolytic anemias occurred among 76 pigmented individuals, while only one subacute anemia was observed among 81 white subjects," with the six "pigmented" individuals separately described: "Five acute hemolytic anemias occurred in negroes and one in a Chinese" [120]. While their terminology is racist, a couple of important lessons emerged from this study, that a subset of people can have severe adverse drug reactions and that this risk can differ in frequency depending on your ancestry (**Box 3**). The susceptibility to haemolytic anaemia following exposure to these antimalarial drugs was subsequently found to be due to G6PD deficiency [121], which is more common in populations originating from regions of the world where malaria is endemic and heterozygosity for the deficiency is protective [122]. In case the repeated mention of malaria causes confusion, the evolutionary selection for G6PD deficiency has nothing to do with the availability of 8-aminoquinoline drugs in the last half century, but instead should have to do with conferring resistance to *Plasmodium falciparum* infection over thousands of years, possibly through the increased phagocytosis of infected erythrocytes that are G6PD deficient [123].

### Box 3. The value of including diverse populations in genetic studies of disease.

It was emphasised that the example of G6PD deficiency reveals population-specific risks, due to the higher prevalence in populations with a long duration of exposure to malaria. Interestingly, using the Geography of Genetic Variants (GGV) browser [131], we see that the *PNPLA3* I148M pathogenic variant is also very heterogeneous in its frequency in different world populations (**Fig 5**). Because the original Dallas Heart Study cohort was diverse, they recognised the increased frequency of the *PNPLA3* I148M pathogenic variant in the Hispanic subset of their patients [129] who are at higher risk for NAFLD than other ethnic groups. This finding highlights the increasingly appreciated value in studying diverse populations to reveal loci mediating susceptibility to genetic diseases [132]. Increasing the genetic variability of the individuals studied can influence not only how genome-wide association studies (GWAS) perform but also how transcriptomic and epigenomic assays are interpreted. Fortunately, this challenge has been turned into a positive, allowing genetic diversity to inform us about specific loci mediating phenotypes and environmental responses.
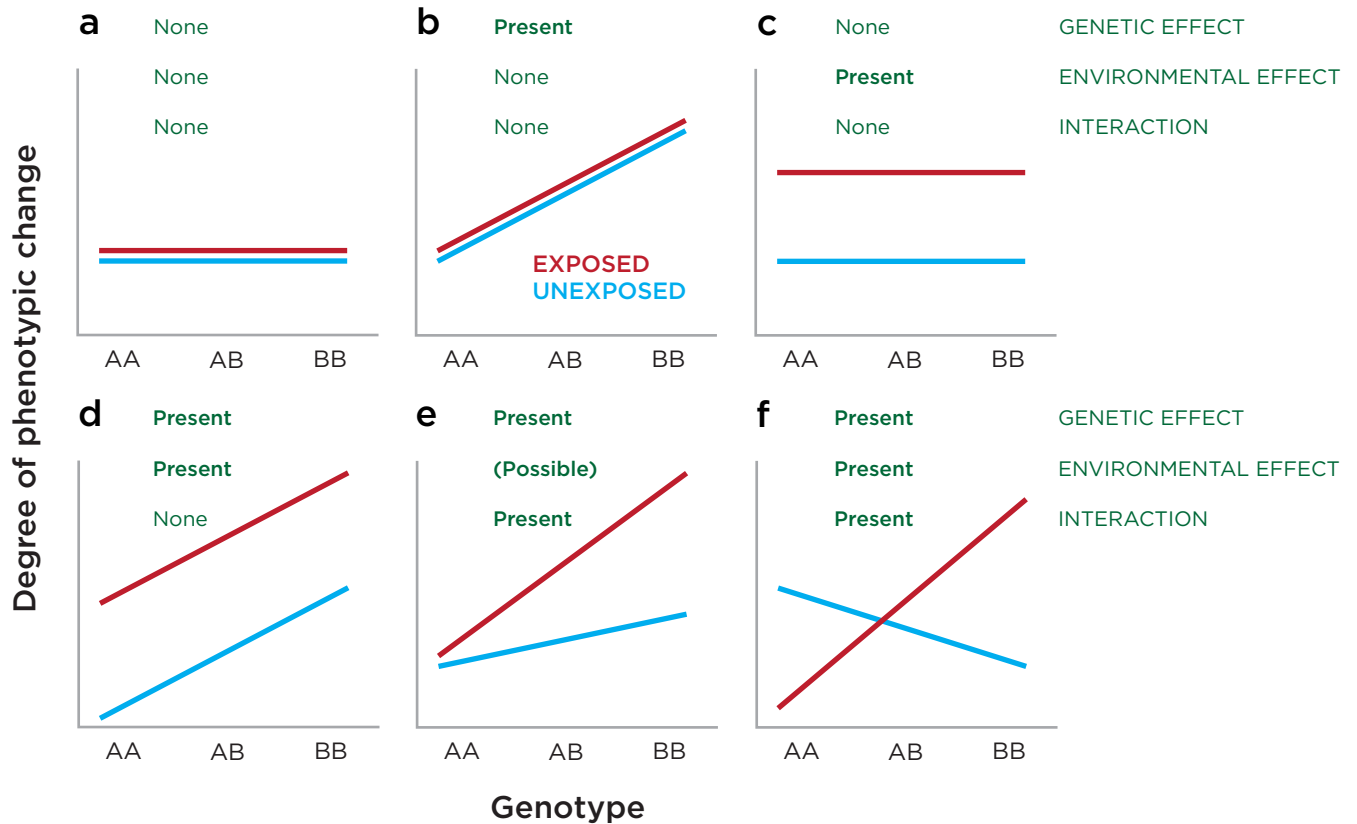
**Fig 3.** Relating phenotypic change ($y$ axis) to genotype ($x$ axis, major allele A and minor allele B) in three situations (green), genetic effects, environmental effects, and interactions.

The study of how genetic sequence variation between individuals influences the response to drug exposure was founded upon observations like this antimalarial drug association with haemolytic anaemia in G6PD deficiency. The field became known as pharmacogenetics, in which the functions of individual genes could be linked with drug responses, the most straightforward means currently for the delivery of personalised medicine, the tailoring of treatment that takes into account the individuality of the patient.

As our ability to explore the DNA sequence polymorphism throughout the genome became facilitated through microarray technologies, studies could be performed that did not need to be anchored by a specific candidate gene but could instead test the entire genome for loci where the response to an environmental exposure was significantly associated with variability at a specific locus. In **Fig 3** (adapted from Dempfle and colleagues [124]), we show examples of the kinds of results that could occur at such a locus. The more common sequence in the population is shown as an A, the less common (minor, alternative) allele as a B. The degree of phenotypic change is plotted on the $y$ axis, comparing people with the A allele on both chromosomes (AA), or the minor allele on one (AB) or both (BB) chromosomes. A change in phenotype associated with genotype at this locus should cause the lines to deviate from the horizontal, while an effect of the environment on the phenotype should cause the exposed (red) and unexposed (blue) lines to separate. When the locus itself is helping to mediate the susceptibility to the environmental exposure, you would expect to see both occurring, as well as evidence for interaction (**Fig 3F**). If the phenotypic measure was haemolysis, and we studied

a loss of function variant within the *G6PD* gene, and considered only females (who have two copies of this X chromosome gene), we should find the blue line to slope gently upwards, as there is a chronic, low-level hemolysis in individuals with G6PD deficiency. Following exposure to fava beans or an 8-aminoquinoline antimalarial drug, the red line should separate strongly upwards in the AB and BB individuals, demonstrating the interaction at this locus. Due to the interaction of genotype and phenotype via environmental interactions, studying them together could be a more powerful method to study the contribution of DNA methylation to disease outcomes. Indeed, DNA methylation at variably methylated regions in neonatal cord blood could be explained by both genetic and environmental effects studied in an integrated model [125].

Gene-environment-wide interaction studies (GEWIS) typically emerged from GWAS, which by themselves attempt to link phenotype only with genotype (as would be exemplified in **Fig 3B**). When the environmental exposure information is available for the people studied, the extra dimension of GEWIS can be added and allows exposures beyond medications to be studied. The exposure does not need to be pharmacological. The field of gene–environment (GxE) interaction research involves major methodological and statistical challenges [126], including the ability to detect these interactions with confidence [127] and suffers from imprecision of the use of terms describing the field [124], but has been successful in many associations, in particular when involving genes with metabolic functions [128,129]. While subsequent GWAS have identified further risk loci [47], a further analysis of the Dallas Heart Study participants revealed the striking results shown in **Fig 4A** (reproduced from [130]). This figure was used to demonstrate vividly how the *PNPLA3* genotype interacts with BMI. **Fig 4B** illustrates how BMI, when taken as a proxy for its causative environmental exposures, generates a plot comparable with the models in **Fig 3**, as a way of illustrating the effects of both genetic and environmental factors
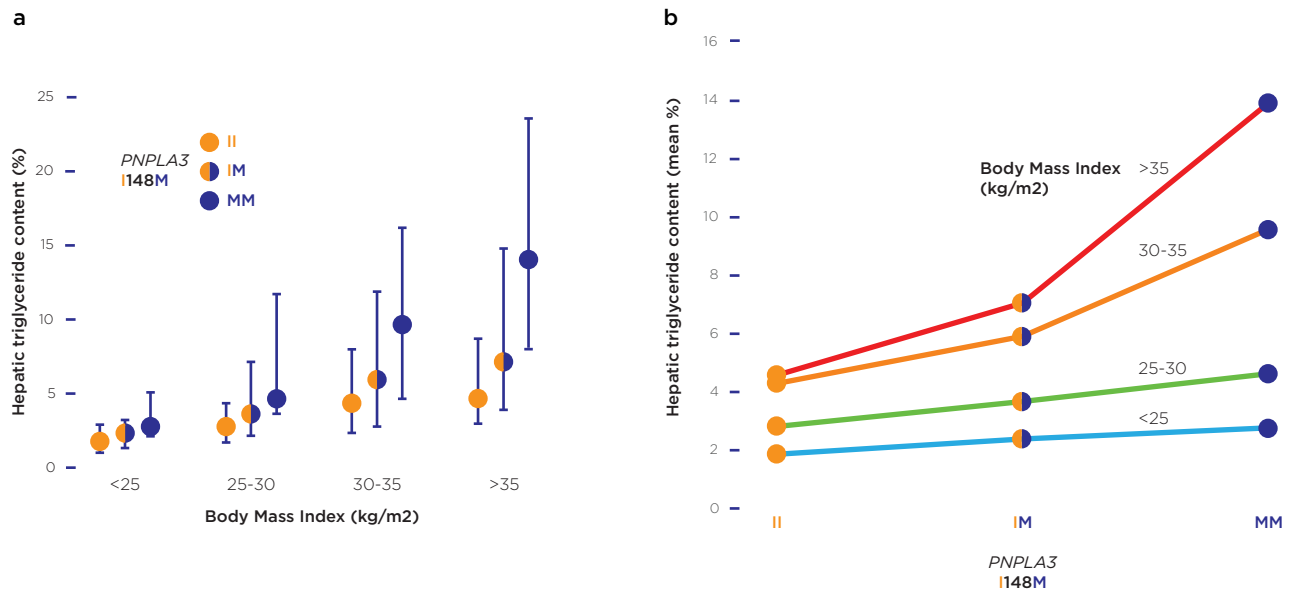


**Fig 4.** (**a**) The interaction of genotype and body mass index to cause MAFLD (quantified by hepatic triglyceride content, *y* axis) is vividly revealed using data from the Dallas Heart Study [116]. (**b**) Replotting the same data using the format of **Fig 2** to show the combined genetic and environmental influences in MAFLD. The WebPlotDigitizer tool (https://automeris.io/WebPlotDigitizer/) was used to extract the raw data from the source image in the published work [115], in the disease.

https://doi.org/10.1371/journal.pgen.1010567.g004

## chr22:44,324,727 G/C



**Frequency Scale = Proportion out of 1**
The pie below represents a minor allele frequency of 0.25

Sample sizes below 30 become increasingly transparent to represent uncertain frequencies, i.e.

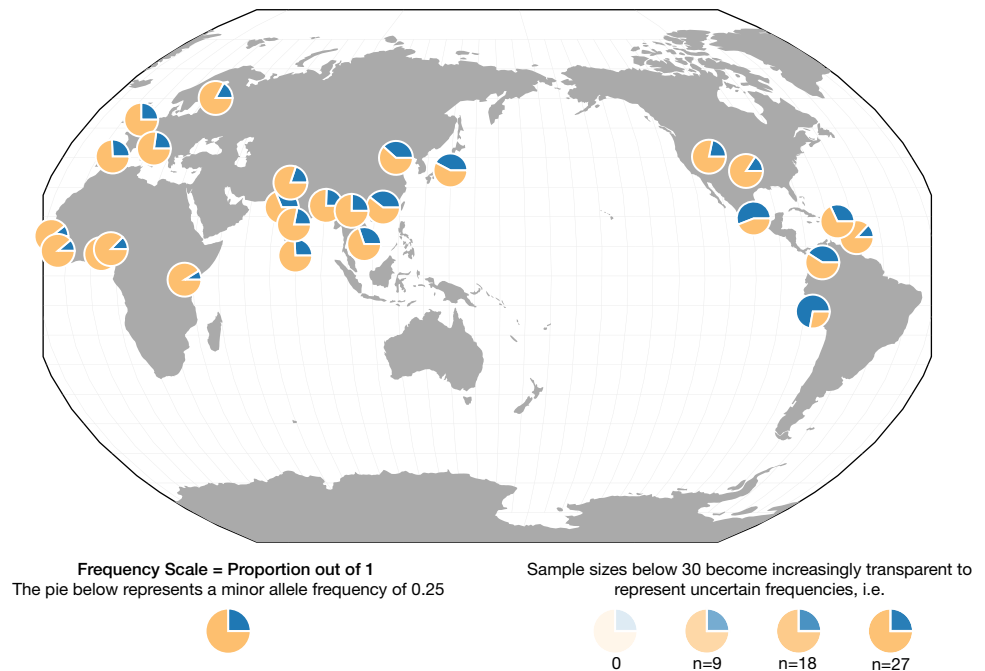0        n=9        n=18        n=27

**Fig 5. The distribution of the *PNPLA3* I148M variant (blue) worldwide shows the greatest enrichment in populations from Central and South America, while being a common variant in all populations.**

### How to test for influences by DNA sequence variants

An issue with any GWAS, or its extension to study environmental effects through GEWIS and GxE approaches, is that it does not point us to a specific locus at nucleotide resolution when it finds a phenotypic association. Instead, it implicates a haplotype, usually at least tens of kilobases in size and containing potentially thousands of variants [133]. To refine the search for the causative locus within the haplotype, one strategy has been to exploit the differences in phenotypic susceptibility between individuals of different ancestries to fine-map the region [134]. Another strategy uses the effects described earlier for DNA sequence variability influencing transcription and its regulatory mechanisms, identifying the subset of functional variants in these regions, allowing them to be prioritised as potentially mediating the phenotype.

The most studied effect of functional variants is on gene expression. By treating the level of expression of a gene as a quantitative phenotype and correlating it with genetic variation, a locus where variation of the DNA sequence is associated with a change in gene expression can be defined. The sequences tested can be on the same chromosome and relatively close (typically <1 Mb) to the gene whose expression level is measured, a *cis* relationship, or further away or on another chromosome, a *trans* relationship. The outcome sought is a change similar to that in **Fig 2**, with the variant described as an eQTL and the target gene an eGene. By identifying the eQTLs in a haplotype implicated by an EWAS, the number of variants can be reduced from hundreds to a very limited number [133]. An even more powerful way of refining the search for causal variants is through studies that look for effects of a variant present on one but not the other allele, leading to an imbalance of expression of a linked gene, referred to as "allele-specific expression" [135]. This helps to safeguard against attributing function to a

variant in a region of the genome in high linkage disequilibrium with neighbouring variants, one of which may instead be mediating the functional effect.

Something interesting happens when you look for eQTLs after challenging the cell with an exposure. The groundbreaking study that revealed "response" eQTLs used an exposure of dendritic cells to infection by *Mycobacterium tuberculosis*. The authors found that while most eQTLs remained the same before and after exposure, a subset was present only in the uninfected or in the infected cells, and were described as response eQTLs. Within the panel of eQTLs genome wide, the response eQTLs were enriched at loci identified by GWAS as being involved in susceptibility to tuberculosis [136].

Response eQTLs have now been identified for multiple different exposures and cell types. Human monocytes have been tested following Toll-like receptor 4 (TLR4) stimulation [137], human dendritic cells were exposed to *E. coli* lipopolysaccharide, influenza, or interferon-β (IFN-β) [138], human monocytes were treated with interferon-γ (IFN-γ) or lipopolysaccharide [139], monocyte-derived macrophages were infected with *Listeria monocytogenes* or *Salmonella typhimurium* [140], while primary monocytes were exposed to ligands activating Toll-like receptor pathways (TLR1/2, TLR4, and TLR7/8) and to influenza virus, in samples from Africans and Europeans [141]. A study that exposed human macrophages to IFNγ, *Salmonella enterica* serovar Typhimurium, or a combination of the two exposures implicated certain TFs in mediating the response to infection, with a primary effect on PU.1 and secondary effects on stimulus-specific TFs, such as NF-κB and STAT2 [142]. A study of whole blood from 1,000 individuals, using exposures to three bacteria, a fungus, a live virus, and a superantigen, demonstrated that variability in responses between individuals was less influenced by age and sex and more by genetic factors, identifying response eQTLs enriched at loci implicated in autoimmune and inflammatory disorders [143]. Allele-specific expression has also proven valuable for discovering response eQTLs to factors such as BMI and exercise in large observational cohorts [144].

As well as these studies of infection and the immune system, we performed a study identifying response eQTLs following exposure of cardiomyocytes to anthracycline [2]. These response eQTLs were more enriched than preexposure eQTLs for loci implicated by GWAS for anthracycline-induced cardiotoxicity. The approach to study allele-specific expression has also been successfully scaled to a high number (50) of environmental exposures in five different cell types, revealing a large number of genes with GxE effects [145]. What all of these studies have in common is that they represent in vitro exposures by agents known to induce responses by the cells used. An obvious question that arises is how to apply these approaches to other diseases. Maintaining our focus on NAFLD, earlier, we described the environmental exposures to the individual may end up translating into exposures to the cells of the liver, including the effects of lipotoxicity on mixed cells types within the liver. We need to ask whether we can study one cell type in isolation, or whether effects of an exposure require the physical relationship between the cells that compose the normal liver. Rather than sampling primary liver cells from human subjects, which has substantially more risk than a blood draw to sample immune cells, we can use the approach of the anthracycline/cardiomyocytes study, which generated cardiomyocytes from induced pluripotent stem cells (iPSCs) from multiple individuals [2]. iPSCs can also be used to generate organoids that contain many of the cell subtypes of an organ, including liver organoids [146], which may be another avenue worth pursuing in studies of responses to exposures, although we highlight the caution that the field of organoid research is still in its early stages [147].

Once we have our exposure and cell system in place, we can move to molecular studies. The genotyping of samples is a foundation for understanding how genetic variability influences environmental responses and the transcriptional and epigenomic data generated.

Furthermore, as mentioned above, the genotypic contribution to disease outcome is likely expressed in a tissue- and cell-specific manner [93]. Genotype information can be inferred from DNA methylation microarrays [148], while it has also been found that Gap Hunter can reveal ancestry information [86]. However, these approaches generates much less genotyping information than more mainstream approaches such as the use of microarrays that represent common variants in the genome, ideally designed to be as informative as possible across diverse populations [149]. An alternative is low-coverage whole genome sequencing, at a depth that is insufficient to identify every variant in the genome but allows the identification of many common variants through imputation methods that leverage large reference panels, and the additional revelation of some lower frequency variants that would not be detected by microarrays [150].

The next molecular assay is typically gene expression analysis, allowing eQTL identification, as described already. However, there are other molecular assays that reveal the effects of functional variants. While we have described meQTLs earlier as loci responding to functional variants by changing their DNA methylation, in practice, these have not been used to test for environmentally responsive loci in a way comparable with eQTL studies. Chromatin accessibility QTLs (caQTLs) represent an intriguing alternative that is understudied at present. An ingenious pooling approach was used to reveal caQTLs in lymphoblastoid cell lines from 1,000 individuals from 10 populations, revealing population-specific caQTLs [151], although this study did not involve any in vitro exposure. A recent, groundbreaking study combined the comparison of cells before and after exposure, testing multiple immune cell types, and identifying both response eQTLs and response caQTLs. The authors found evidence for function of candidate causal variants that would have been undetectable using more mainstream approaches studying resting cells [152]. A problem with eQTLs is that the expression of a gene is generally influenced by multiple *cis* regulators [153], making it difficult to link local sequence variation with the quantitative trait of gene expression. Chromatin accessibility, on the other hand, is likely to vary at the locus containing the functional variant, lending itself to allele-specific studies [151,154]. A shift in focus from eQTLs to caQTLs may be fruitful for functional variant analyses.

Finally, it should be borne in mind that the relatively common variants that we can identify through microarray or low-coverage whole genome sequencing with imputation may only contribute a small proportion of variability of gene expression. Even with detailed haplotype information (in mostly European populations), the imputation can only predict variants down to a frequency of $1/1,000$ ($10^{-3}$) [155]. A study of 360 LCLs derived from European individuals combined deep whole genome sequencing and RNA sequencing to identify which variants influenced the heritability of levels of gene expression. Approximately 90% of the heritability was found to be associated with sequence variants that occurred only once in the cohort (singletons) and at a minor allele frequency of <0.01% in the gnomAD database [156]. If we assume that the variants identified in this study represent those with effects on DNA methylation (meQTLs) and chromatin accessibility (caQTLs), we can expect that only a small proportion of variants causing changes in expression, DNA methylation and chromatin accessibility will be revealed by current approaches, that deep whole genome sequencing will reveal many more variants affecting these molecular phenotypes, and that most functional variants will be present in the genome in a heterozygous state because of their population rarity.

## Insights from the epigenome into cell signalling

To close the circle, we return to the use of epigenomic assays but now dissociated from their ability to reveal information about DNA sequence variants in mediating differences between

individuals in their responses to environmental influences. A question that is often left unaddressed in epigenomic studies is why specific loci undergo changes, whether of DNA methylation or chromatin accessibility. We have made the point previously that such targeting implies a primary role for TFs [65]. The reason that functional variants have their properties appears often to be due to the effect of the DNA sequence to change local TF binding [154]. In a recent study of oestradiol's effects on the ventral hippocampus of female mice, we showed that the hormone acts on a cell surface receptor to initiate a cytoplasmic cell signalling cascade that culminates in the activity of the Egr1 TF to change chromatin structure and gene expression [157]. The Alasoo and colleagues' study described earlier implicated specific TFs in mediating the response eQTLs they identified [142]. Of the environmental influences with potential effects on epigenetic regulators listed in **Table 1**, hypoxia [158], hyperglycaemia [159], ethanol [160], and lactic acid [161] are good examples of environmental cellular stresses known to influence cell signalling pathways, with potential consequences on TF activity and nuclear localisation. With this TF-centred perspective, we can return to some of the foundational studies in the field of environmental epigenomics and ask whether they could be viewed alternatively with a TF-centric perspective.

One very understudied area of potential importance is the PTM of TFs. It is known that some TFs can be acetylated or deacetylated by the same enzymes that act on histones [162], with methylation and demethylation likewise mediated by histone-modifying enzymes [163]. The effects of environmental exposures that disrupt histone acetylation, methylation, or other PTMs could also therefore be acting on TFs. The paradigm of dietary folic acid having effects on transcriptional regulation may not be solely due to its effects to augment DNA methylation but could be mediated through the property of the folic acid receptor to act as a TF [164]. The phenomenon of temperature-dependent sex determination took over 50 years to find that the transcription factor *Dmrt1* (*doublesex and mab3-related transcription factor 1*) is likely to be a primary mediator of the temperature response in another amphibian, the red-eared slider turtle *Trachemys scripta* [165]. Considering the effects of the environment as being mediated through TFs helps to explain the sequence specificity of environmental responses by the epigenome.

If TFs are central in directing the response by the cell to an environmental exposure, the next question is what controls the TFs? Some are directly bound in the nucleus by ligands, the nuclear receptors [166], but many TFs act in response to cell signalling pathways [167]. This suggests that transcriptomic and epigenomic assays are defining two sets of information following an environmental exposure. Typically, we look for a coherence in the loci where changes are occurring in the genome, whether genes changing expression, or the genes linked to loci where DNA methylation or chromatin states are changing. The coherence is expressed in terms of the gene properties (through overrepresentation of specific gene ontology terms) or through known interactions between the protein products of genes.

If, however, this represents downstream of TF activity selecting these loci in the genome, it could be said that this represents a secondary response to the environmental stimulus. By identifying the TF(s) mediating the response and working upstream, we can define the primary response to the environmental influence. This logic was used to develop SPAGI (Signalling Pathway Analysis for putative Gene regulatory network Identification), a tool that infers from TFs implicated in a transcriptional response the pathways that activated those TFs in the first place [168]. We illustrate these ideas of secondary and primary pathway responses in **Fig 6**, representing a further way of exploiting epigenomic and transcriptomic information when studying environmental responses by the cell. With a goal of defining targets for therapeutic intervention, defining cell signalling pathways involved in environmental responses offers clear opportunities, but it should be noted that the TFs themselves are no longer considered
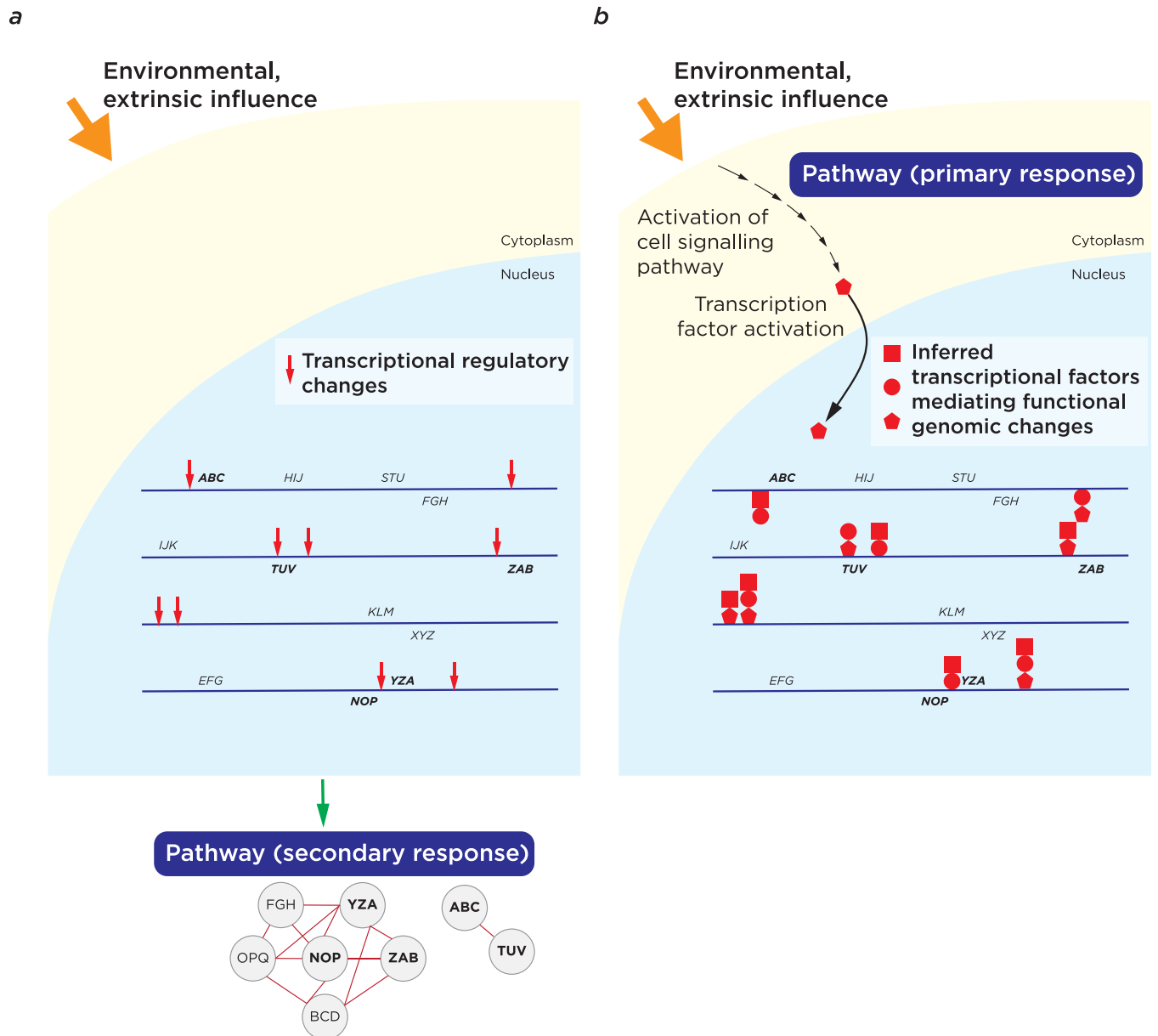
**Fig 6.** (**a**) When testing the response to an extrinsic influence using genomic approaches, we typically identify the loci at which changes are most pronounced and attempt to understand how they have biological coherence by studying the genes implicated for their biological properties, including protein–protein interactions. While this is part of the cellular response to the extrinsic perturbation, an alternative perspective is shown in (**b**), considering the role of TFs to select loci for altered function and, in turn, the cell signalling pathways that regulate the TFs. The activation of TFs in this model would be the primary response to the extrinsic influence, with the transcriptional regulatory changes a secondary event.

"undruggable" [169–171] and that, in the case of NAFLD and NASH, there are some nuclear receptor TFs that have been targeted for therapy (Peroxisome proliferator-activated receptor (PPAR) proteins and the Farnesoid X receptor (FXR) [172]). Epigenomic assays, used with the idea that they reveal cell signalling and TFs mediating environmental responses, could be valuable in defining targets for therapeutic intervention.

With a TF-centric perspective that leads us to consider the role of cell signalling, we can return to a fundamental question about epigenetics and heritability—does the cell signalling state of the parent cell influence that of daughter cells? This would represent a further

mechanism for inheriting cellular properties through cell division. There is evidence for such a mechanism. Alterations of parental cell stress or mitogenic activity in parent cells has been found to influence cell cycle commitment in the daughter cells, mediated by transmission of the p53 protein and the cyclin D1 (*CCND1*) mRNA through mitosis [173]. It is probably reasonable to consider this as an example of a broader phenomenon of nongenetic heritability through cell division, although probably an uncomfortable fit for those who would describe epigenetic heritability purely in terms of nuclear information.

## Conclusions and future directions

The study of the epigenome and its response to environmental exposures can be much more encompassing than current models would indicate, making these studies more complex but also more promising and exciting in terms of the potential insights to be gained. As a field, our initial hope was that focusing studies on the regulators of the genome would be enough to give direct clues to the mechanisms and outcomes of environmental exposures. While this remains possible, in many cases, these studies are more likely to be revealing the effects of variation of DNA sequence, or of cell subtype proportions, and the effects of TFs. The mainstream interpretation of EWAS that defines these influences as sources of error is excessively restrictive. By embracing these "spurious" influences instead as sources of insights into the physiological or pathological effects of environmental exposures, we expand the opportunity to discover more mechanisms underlying the associated phenotype. Epigenomics assays can become a greater part of the repertoire of approaches being used to follow up GWAS to identify the genetic loci involved. To gain new insights into human cell and tissue types that are normally inaccessible, we can leverage advances in iPSC differentiation to create many different cell types or even organoids, permitting the detailed dissection of molecular events responding to in vitro exposures in these highly controlled systems. Additionally, we have the opportunity to use epigenomic approaches to not only reveal the TFs central to mediating environmental responses, but also to infer their upstream cell signalling regulators, revealing possible targets for therapeutic interventions.

This more expansive approach to the use of epigenomics to understand environmental influences in disease comes with clear challenges. The issue of very rare genetic variants influencing environmental responses, and the resulting phenotypes of the individual, their cells and the molecular regulators is a significant problem. The identification of the ultrarare variants that may mediate a significant proportion of these interindividual differences requires deep whole genome sequencing. However, this creates another opportunity for epigenomic approaches—while the individual DNA sequence variants will be rarely observed more than once even in large cohorts, many different rare variants at a locus can have the convergent outcome of a change in a functional property of the locus, whether chromatin accessibility, DNA methylation, or an effect on nearby gene expression. Multiple different rare variants at a locus are likely to converge functionally as the same kind of change in a functional genomic property. The most productive strategy is therefore to associate the polymorphism of molecular genomic phenotypes with exposures or cellular/organismal phenotypes as the primary association, which should be relatively more easily detected. The more typical approach that links DNA sequence variability with the molecular genomic phenotype thus becomes the secondary association. This convergence of multiple rare variant effects on a molecular genomic outcome in effect "collapses" the multiple DNA sequence variants into a common outcome to increase association power.

There are deficiencies in our insights into the repertoire of DNA motifs bound by TFs, the upstream regulatory influences upon TFs by cell signalling pathways, and the related issue of

the types, mediators, and effects of PTMs of TFs. Any in vitro studies will need to be designed in a way that represents the best guess about the exposures at the cellular level in vivo and the one or more cell types responding to mediate the disease or other phenotype. Animal models will continue to have major value as a parallel to these direct studies of human cells. Broadening the repertoire of pluripotent stem cell resources to represent more diverse racial and ethnic groups with distinctive risks of environmentally responsive phenotypes will be another pressing need.

Finally, it should be borne in mind that just because GWAS or the extended GEWIS and GxE studies do not explain all susceptibility to a disease, this is not by itself a justification for performing epigenomic studies of disease. The gap between what GWAS findings can explain and the heritability estimated from twin and other studies has been described as "missing heritability." To fill this gap, environmental interactions are invoked, and epigenetic dysregulation as a consequence of the environmental influences or as a separate source of variability is also considered. Missing heritability may be due to many factors, including the inflation of the estimate of heritability from twin studies, limited sample size, the polygenicity of phenotypes [174], and the rarity of the genetic variants causing the conditions [175]. What we tend to overlook is that chance is likely to be an additional factor in phenotypes [176], but that should be easily embraced by those interested in epigenetics, as the ball rolling down Waddington's epigenetic landscape was not predestined to end up in a specific creode following a series of bifurcations; the future lineage commitment of the cell being represented was probabilistic and, therefore, subject to random variability. Nondeterminism is therefore at the core of the original idea of an epigenetic model for phenotypic variability.

## References

1. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes Dev. 2009; 23:781–783. doi: https://doi.org/10.1101/gad.1787609 PMID: 19339683

2. Knowles DA, Burrows CK, Blischak JD, Patterson KM, Serie DJ, Norton N, et al. Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes. elife. 2018: 7. doi: https://doi.org/10.7554/eLife.33480 PMID: 29737278

3. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nat Commun. 2020; 11:810. doi: https://doi.org/10.1038/s41467-020-14457-z PMID: 32041960

4. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003; 33(Suppl):245–254. doi: https://doi.org/10.1038/ng1089 PMID: 12610534

5. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. Genome Med. 2014; 6:60. doi: https://doi.org/10.1186/s13073-014-0060-z PMID: 25484923

6. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci USA. 2005; 102:10604–10609. doi: https://doi.org/10.1073/pnas.0500398102 PMID: 16009939

7. Wilson MJ, Shivapurkar N, Poirier LA. Hypomethylation of hepatic nuclear DNA in rats fed with a carcinogenic methyl-deficient diet. Biochem J. 1984; 218:987–990. doi: https://doi.org/10.1042/bj2180987 PMID: 6721844

8. Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. FASEB J. 1998; 12:949–957. PMID: 9707167

9. Cooney CA, Dave AA, Wolff GL. Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. J Nutr. 2002; 132:2393S–2400S. doi: https://doi.org/10.1093/jn/132.8.2393S PMID: 12163699

10. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999; 23:314–318. doi: https://doi.org/10.1038/15490 PMID: 10545949

11. Wadhwa PD, Buss C, Entringer S, Swanson JM. Developmental origins of health and disease: brief history of the approach and current focus on epigenetic mechanisms. Semin Reprod Med. 2009; 27:358–368. doi: https://doi.org/10.1055/s-0029-1237424 PMID: 19711246

12. Russo VEA, Martienssen RA, Riggs AD, editors. Epigenetic Mechanisms of Gene Regulation. Illustrated ed. Cold Spring Harbor Laboratory Press; 1996.

13. Ravelli GP, Stein ZA, Susser MW. Obesity in young men after famine exposure in utero and early infancy. N Engl J Med. 1976; 295:349–353. doi: https://doi.org/10.1056/NEJM197608122950701 PMID: 934222

14. Stein Z, Susser M. Fertility, fecundity, famine: food rations in the dutch famine 1944/5 have a causal relation to fertility, and probably to fecundity. Hum Biol. 1975; 47:131–154. PMID: 1126699

15. Lumey LH, Ravelli AC, Wiessing LG, Koppe JG, Treffers PE, Stein ZA. The Dutch famine birth cohort study: design, validation of exposure, and selected characteristics of subjects after 43 years follow-up. Paediatr Perinat Epidemiol. 1993; 7:354–367. doi: https://doi.org/10.1111/j.1365-3016.1993.tb00415.x PMID: 8290375

16. Charnier M. Action de la température sur la sex-ratio chez l'embryon d'Agama agama (Agamidae, Lacertilien). Comptes Rendus des Séances de la Société de Biologie de l'Ouest Africain, Paris. 1966; 160:620–622.

17. Song J, Angel A, Howard M, Dean C. Vernalization—a cold-induced epigenetic switch. J Cell Sci. 2012; 125:3723–3731. doi: https://doi.org/10.1242/jcs.084764 PMID: 22935652

18. Gaydos LJ, Wang W, Strome S, Gene repression. H3K27me and PRC2 transmit a memory of repression across generations and during development. Science. 2014; 345:1515–1518. doi: https://doi.org/10.1126/science.1255023 PMID: 25237104

19. Michelotti EF, Sanford S, Levens D. Marking of active genes on mitotic chromosomes. Nature. 1997; 388:895–899. doi: https://doi.org/10.1038/42282 PMID: 9278053

20. Lo P- W, Shie J- J, Chen C- H, Wu C- Y, Hsu T- L, Wong C- H. O-GlcNAcylation regulates the stability and enzymatic activity of the histone methyltransferase EZH2. Proc Natl Acad Sci USA. 2018; 115:7302–7307. doi: https://doi.org/10.1073/pnas.1801850115 PMID: 29941599

21. Hrit J, Goodrich L, Li C, Wang B- A, Nie J, Cui X, et al. OGT binds a conserved C-terminal domain of TET1 to regulate TET1 activity and function in development. elife. 2018: 7. doi: https://doi.org/10.7554/eLife.34870 PMID: 30325306

22. Zhang D, Tang Z, Huang H, Zhou G, Cui C, Weng Y, et al. Metabolic regulation of gene expression by histone lactylation. Nature. 2019; 574:575–580. doi: https://doi.org/10.1038/s41586-019-1678-1 PMID: 31645732

23. Farrelly LA, Thompson RE, Zhao S, Lepack AE, Lyu Y, Bhanu NV, et al. Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3. Nature. 2019; 567:535–539. doi: https://doi.org/10.1038/s41586-019-1024-7 PMID: 30867594

24. Mews P, Egervari G, Nativio R, Sidoli S, Donahue G, Lombroso SI, et al. Alcohol metabolism contributes to brain histone acetylation. Nature. 2019; 574:717–721. doi: https://doi.org/10.1038/s41586-019-1700-7 PMID: 31645761

25. Posavec Marjanović M, Hurtado-Bagès S, Lassi M, Valero V, Malinverni R, Delage H, et al. MacroH2A1.1 regulates mitochondrial respiration by limiting nuclear NAD+ consumption. Nat Struct Mol Biol. 2017; 24:902–910. doi: https://doi.org/10.1038/nsmb.3481 PMID: 28991266

26. Casciello F, Al-Ejeh F, Kelly G, Brennan DJ, Ngiow SF, Young A, et al. G9a drives hypoxia-mediated gene repression for breast cancer cell survival and tumorigenesis. Proc Natl Acad Sci USA. 2017; 114:7077–7082. doi: https://doi.org/10.1073/pnas.1618706114 PMID: 28630300

27. Batie M, Frost J, Frost M, Wilson JW, Schofield P, Rocha S. Hypoxia induces rapid changes to histone methylation and reprograms chromatin. Science. 2019; 363:1222–1226. doi: https://doi.org/10.1126/science.aau5870 PMID: 30872526

28. Fellows R, Denizot J, Stellato C, Cuomo A, Jain P, Stoyanova E, et al. Microbiota derived short chain fatty acids promote histone crotonylation in the colon through histone deacetylases. Nat Commun. 2018; 9:105. doi: https://doi.org/10.1038/s41467-017-02651-5 PMID: 29317660

29. Mentch SJ, Locasale JW. One-carbon metabolism and epigenetics: understanding the specificity. Ann N Y Acad Sci. 2016; 1363:91–98. doi: https://doi.org/10.1111/nyas.12956 PMID: 26647078

30. Blaschke K, Ebata KT, Karimi MM, Zepeda-Martínez JA, Goyal P, Mahapatra S, et al. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. Nature. 2013; 500:222–226. doi: https://doi.org/10.1038/nature12362 PMID: 23812591

31. Gagnon J, Daou S, Zamorano N, Iannantuono NVG, Hammond-Martel I, Mashtalir N, et al. Undetectable histone O-GlcNAcylation in mammalian cells. Epigenetics. 2015; 10:677–691. doi: https://doi.org/10.1080/15592294.2015.1060387 PMID: 26075789

32. Imai S- I, Guarente L. It takes two to tango: NAD+ and sirtuins in aging/longevity control. npj Aging Mech Dis. 2016; 2:16017. doi: https://doi.org/10.1038/npjamd.2016.17 PMID: 28721271

**33.** Ray K. NAFLD-the next global epidemic. Nat Rev Gastroenterol Hepatol. 2013; 10:621. doi: https://doi.org/10.1038/nrgastro.2013.197 PMID: 24185985

**34.** Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology. 2016; 64:73–84. doi: https://doi.org/10.1002/hep.28431 PMID: 26707365

**35.** Lazo M, Hernaez R, Eberhardt MS, Bonekamp S, Kamel I, Guallar E, et al. Prevalence of nonalcoholic fatty liver disease in the United States: the Third National Health and Nutrition Examination Survey, 1988–1994. Am J Epidemiol. 2013; 178:38–45. doi: https://doi.org/10.1093/aje/kws448 PMID: 23703888

**36.** McPherson S, Hardy T, Henderson E, Burt AD, Day CP, Anstee QM. Evidence of NAFLD progression from steatosis to fibrosing-steatohepatitis using paired biopsies: implications for prognosis and clinical management. J Hepatol. 2015; 62:1148–1155. doi: https://doi.org/10.1016/j.jhep.2014.11.034 PMID: 25477264

**37.** Stepanova M, Rafiq N, Makhlouf H, Agrawal R, Kaur I, Younoszai Z, et al. Predictors of all-cause mortality and liver-related mortality in patients with non-alcoholic fatty liver disease (NAFLD). Dig Dis Sci. 2013; 58:3017–3023. doi: https://doi.org/10.1007/s10620-013-2743-5 PMID: 23775317

**38.** Younossi ZM, Tampi R, Priyadarshini M, Nader F, Younossi IM, Racila A. Burden of illness and economic model for patients with nonalcoholic steatohepatitis in the united states. Hepatology. 2019; 69:564–572. doi: https://doi.org/10.1002/hep.30254 PMID: 30180285

**39.** Liebe R, Esposito I, Bock HH, Vom Dahl S, Stindt J, Baumann U, et al. Diagnosis and management of secondary causes of steatohepatitis. J Hepatol. 2021; 74:1455–1471. doi: https://doi.org/10.1016/j.jhep.2021.01.045 PMID: 33577920

**40.** Arab JP, Arrese M, Trauner M. Recent Insights into the Pathogenesis of Nonalcoholic Fatty Liver Disease. Annu Rev Pathol. 2018; 13:321–350. doi: https://doi.org/10.1146/annurev-pathol-020117-043617 PMID: 29414249

**41.** Kolodziejczyk AA, Zheng D, Shibolet O, Elinav E. The role of the microbiome in NAFLD and NASH. EMBO Mol Med. 2019: 11. doi: https://doi.org/10.15252/emmm.201809302 PMID: 30591521

**42.** Lynch C, Chan CS, Drake AJ. Early life programming and the risk of non-alcoholic fatty liver disease. J Dev Orig Health Dis. 2017; 8:263–272. doi: https://doi.org/10.1017/S2040174416000805 PMID: 28112071

**43.** Sookoian S, Pirola CJ. Genetic predisposition in nonalcoholic fatty liver disease. Clin Mol Hepatol. 2017; 23:1–12. doi: https://doi.org/10.3350/cmh.2016.0109 PMID: 28268262

**44.** Younes R, Bugianesi E. NASH in lean individuals. Semin Liver Dis. 2019; 39:86–95. doi: https://doi.org/10.1055/s-0038-1677517 PMID: 30654392

**45.** Namjou B, Lingren T, Huang Y, Parameswaran S, Cobb BL, Stanaway IB, et al. GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. BMC Med. 2019; 17:135. doi: https://doi.org/10.1186/s12916-019-1364-z PMID: 31311600

**46.** Anstee QM, Darlay R, Cockell S, Meroni M, Govaere O, Tiniakos D, et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort. J Hepatol. 2020; 73:505–515. doi: https://doi.org/10.1016/j.jhep.2020.04.003 PMID: 32298765

**47.** Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet. 2011; 7:e1001324. doi: https://doi.org/10.1371/journal.pgen.1001324 PMID: 21423719

**48.** Gawrieh S, Guo X, Tan J, Lauzon M, Taylor KD, Loomba R, et al. A Pilot Genome-Wide Analysis Study Identifies Loci Associated With Response to Obeticholic Acid in Patients With NASH. Hepatol Commun. 2019; 3:1571–1584. doi: https://doi.org/10.1002/hep4.1439 PMID: 31832568

**49.** Abdelmalek MF, Suzuki A, Guy C, Unalp-Arida A, Colvin R, Johnson RJ, et al. Increased fructose consumption is associated with fibrosis severity in patients with nonalcoholic fatty liver disease. Hepatology. 2010; 51:1961–1971. doi: https://doi.org/10.1002/hep.23535 PMID: 20301112

**50.** Jensen T, Abdelmalek MF, Sullivan S, Nadeau KJ, Green M, Roncal C, et al. Fructose and sugar: A major mediator of non-alcoholic fatty liver disease. J Hepatol. 2018; 68:1063–1075. doi: https://doi.org/10.1016/j.jhep.2018.01.019 PMID: 29408694

**51.** Dooley S, ten Dijke P. TGF-β in progression of liver disease. Cell Tissue Res. 2012; 347:245–256. doi: https://doi.org/10.1007/s00441-011-1246-y PMID: 22006249

**52.** Hyun J, Jung Y. DNA methylation in nonalcoholic fatty liver disease. Int J Mol Sci. 2020:21. doi: https://doi.org/10.3390/ijms21218138 PMID: 33143364

53. Murphy SK, Yang H, Moylan CA, Pang H, Dellinger A, Abdelmalek MF, et al. Relationship between methylome and transcriptome in patients with nonalcoholic fatty liver disease. Gastroenterology. 2013; 145:1076–1087. doi: https://doi.org/10.1053/j.gastro.2013.07.047 PMID: 23916847

54. Ahrens M, Ammerpohl O, von Schönfels W, Kolarova J, Bens S, Itzel T, et al. DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. Cell Metab. 2013; 18:296–302. doi: https://doi.org/10.1016/j.cmet.2013.07.004 PMID: 23931760

55. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. Gastroenterology. 2017; 153:1096–1106.e2. doi: https://doi.org/10.1053/j.gastro.2017.06.003 PMID: 28624579

56. de Mello VD, Matte A, Perfilyev A, Männistö V, Rönn T, Nilsson E, et al. Human liver epigenetic alterations in non-alcoholic steatohepatitis are related to insulin action. Epigenetics. 2017; 12:287–295. doi: https://doi.org/10.1080/15592294.2017.1294305 PMID: 28277977

57. Zhang R- N, Pan Q, Zheng R- D, Mi Y- Q, Shen F, Zhou D, et al. Genome-wide analysis of DNA methylation in human peripheral leukocytes identifies potential biomarkers of nonalcoholic fatty liver disease. Int J Mol Med. 2018; 42:443–452. doi: https://doi.org/10.3892/ijmm.2018.3583 PMID: 29568887

58. Hotta K, Kitamoto T, Kitamoto A, Ogawa Y, Honda Y, Kessoku T, et al. Identification of the genomic region under epigenetic regulation during non-alcoholic fatty liver disease progression. Hepatol Res. 2018; 48:E320–E334. doi: https://doi.org/10.1111/hepr.12992 PMID: 29059699

59. Gerhard GS, Malenica I, Llaci L, Chu X, Petrick AT, Still CD, et al. Differentially methylated loci in NAFLD cirrhosis are associated with key signaling pathways. Clin Epigenetics. 2018; 10:93. doi: https://doi.org/10.1186/s13148-018-0525-9 PMID: 30005700

60. Wu J, Zhang R, Shen F, Yang R, Zhou D, Cao H, et al. Altered DNA Methylation Sites in Peripheral Blood Leukocytes from Patients with Simple Steatosis and Nonalcoholic Steatohepatitis (NASH). Med Sci Monit. 2018; 24:6946–6967. doi: https://doi.org/10.12659/MSM.909747 PMID: 30270343

61. Johnson ND, Wu X, Still CD, Chu X, Petrick AT, Gerhard GS, et al. Differential DNA methylation and changing cell-type proportions as fibrotic stage progresses in NAFLD. Clin Epigenetics. 2021; 13:152. doi: https://doi.org/10.1186/s13148-021-01129-y PMID: 34353365

62. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011; 12:529–541. doi: https://doi.org/10.1038/nrg3000 PMID: 21747404

63. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013; 10:949–955. doi: https://doi.org/10.1038/nmeth.2632 PMID: 24076989

64. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. PLoS Genet. 2016; 12:e1006105. doi: https://doi.org/10.1371/journal.pgen.1006105 PMID: 27336614

65. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017; 18:441–451. doi: https://doi.org/10.1038/nrg.2017.32 PMID: 28555657

66. Lowe R, Gemma C, Beyan H, Hawa MI, Bazeos A, Leslie RD, et al. Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. Epigenetics. 2013; 8:445–454. doi: https://doi.org/10.4161/epi.24362 PMID: 23538714

67. Lin X, Teh AL, Chen L, Lim IY, Tan PF, MacIsaac JL, et al. Choice of surrogate tissue influences neonatal EWAS findings. BMC Med. 2017; 15:211. doi: https://doi.org/10.1186/s12916-017-0970-x PMID: 29202839

68. Tsai P- C, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. Int J Epidemiol. 2015; 44:1429–1441. doi: https://doi.org/10.1093/ije/dyv041 PMID: 25972603

69. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. BMC Genomics. 2019; 20:366. doi: https://doi.org/10.1186/s12864-019-5761-7 PMID: 31088362

70. Richmond RC, Sharp GC, Ward ME, Fraser A, Lyttleton O, McArdle WL, et al. DNA methylation and BMI: investigating identified methylation sites at HIF3A in a causal framework. Diabetes. 2016; 65:1231–1244. doi: https://doi.org/10.2337/db15-0996 PMID: 26861784

71. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017; 541:81–86. doi: https://doi.org/10.1038/nature20784 PMID: 28002404

72. Dekkers KF, van Iterson M, Slieker RC, Moed MH, Bonder MJ, van Galen M, et al. Blood lipids influence DNA methylation in circulating cells. Genome Biol. 2016; 17:138. doi: https://doi.org/10.1186/s13059-016-1000-6 PMID: 27350042

**73.** Ng JWY, Barrett LM, Wong A, Kuh D, Smith GD, Relton CL. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. Genome Biol. 2012; 13:246. doi: https://doi.org/10.1186/gb-2012-13-6-246 PMID: 22747597

**74.** Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30:1363–1369. doi: https://doi.org/10.1093/bioinformatics/btu049 PMID: 24478339

**75.** Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017; 18:105. doi: https://doi.org/10.1186/s12859-017-1511-5 PMID: 28193155

**76.** Kong Y, Rastogi D, Seoighe C, Greally JM, Suzuki M. Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data. PLoS ONE. 2019; 14:e0215987. doi: https://doi.org/10.1371/journal.pone.0215987 PMID: 31022271

**77.** Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, et al. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet. 2010; 86:411–419. doi: https://doi.org/10.1016/j.ajhg.2010.02.005 PMID: 20215007

**78.** Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016; 167:1398–1414.e24. doi: https://doi.org/10.1016/j.cell.2016.10.026 PMID: 27863251

**79.** Bell JT, Tsai P- C, Yang T- P, Pidsley R, Nisbet J, Glass D, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012; 8:e1002629. doi: https://doi.org/10.1371/journal.pgen.1002629 PMID: 22532803

**80.** van Dongen J, Nivard MG, Willemsen G, Hottenga J- J, Helmer Q, Dolan CV, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016; 7:11115. doi: https://doi.org/10.1038/ncomms11115 PMID: 27051996

**81.** Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, 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. 2013; 93:876–890. doi: https://doi.org/10.1016/j.ajhg.2013.10.004 PMID: 24183450

**82.** Gunasekara CJ, Scott CA, Laritsky E, Baker MS, MacKay H, Duryea JD, et al. A genomic atlas of systemic interindividual epigenetic variation in humans. Genome Biol. 2019; 20:105. doi: https://doi.org/10.1186/s13059-019-1708-1 PMID: 31155008

**83.** Cheung WA, Shao X, Morin A, Siroux V, Kwan T, Ge B, et al. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. Genome Biol. 2017; 18:50. doi: https://doi.org/10.1186/s13059-017-1173-7 PMID: 28283040

**84.** Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12:R10. doi: https://doi.org/10.1186/gb-2011-12-1-r10 PMID: 21251332

**85.** Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, et al. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS Genet. 2011; 7:e1002228. doi: https://doi.org/10.1371/journal.pgen.1002228 PMID: 21852959

**86.** Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. "Gap hunting" to characterize clustered probe signals in Illumina methylation array data. Epigenetics Chromatin. 2016; 9:56. doi: https://doi.org/10.1186/s13072-016-0107-z PMID: 27980682

**87.** Pan J- J, Fallon MB. Gender and racial differences in nonalcoholic fatty liver disease. World J Hepatol. 2014; 6:274–283. doi: https://doi.org/10.4254/wjh.v6.i5.274 PMID: 24868321

**88.** Rich NE, Oji S, Mufti AR, Browning JD, Parikh ND, Odewole M, et al. Racial and Ethnic Disparities in Nonalcoholic Fatty Liver Disease Prevalence, Severity, and Outcomes in the United States: A Systematic Review and Meta-analysis. Clin Gastroenterol Hepatol. 2018; 16:198–210.e2. doi: https://doi.org/10.1016/j.cgh.2017.09.041 PMID: 28970148

**89.** Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014; 15:145. doi: https://doi.org/10.1186/1471-2164-15-145 PMID: 24555763

**90.** Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat Genet. 2022; 54:18–29. doi: https://doi.org/10.1038/s41588-021-00969-x PMID: 34980917

**91.** Natri HM, Bobowik KS, Kusuma P, Crenna Darusallam C, Jacobs GS, Hudjashov G, et al. Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental

differences across the Indonesian archipelago. PLoS Genet. 2020; 16:e1008749. doi: https://doi.org/10.1371/journal.pgen.1008749 PMID: 32453742

92. He S, McPhaul C, Li JZ, Garuti R, Kinch L, Grishin NV, et al. A sequence variation (I148M) in PNPLA3 associated with nonalcoholic fatty liver disease disrupts triglyceride hydrolysis. J Biol Chem. 2010; 285:6706–6715. doi: https://doi.org/10.1074/jbc.M109.064501 PMID: 20034933

93. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017; 169:1177–1186. doi: https://doi.org/10.1016/j.cell.2017.05.038 PMID: 28622505

94. Miao Z, Garske KM, Pan DZ, Koka A, Kaminska D, Männistö V, et al. Identification of 90 NAFLD GWAS loci and establishment of NAFLD PRS and causal role of NAFLD in coronary artery disease. HGG Adv. 2022; 3:100056. doi: https://doi.org/10.1016/j.xhgg.2021.100056 PMID: 35047847

95. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018; 19:581–590. doi: https://doi.org/10.1038/s41576-018-0018-x PMID: 29789686

96. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019; 51:584–591. doi: https://doi.org/10.1038/s41588-019-0379-x PMID: 30926966

97. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018; 50:1219–1224. doi: https://doi.org/10.1038/s41588-018-0183-z PMID: 30104762

98. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics. 2020; 15:1–11. doi: https://doi.org/10.1080/15592294.2019.1644879 PMID: 31318318

99. Xu C- J, Söderhäll C, Bustamante M, Baïz N, Gruzieva O, Gehring U, et al. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. Lancet Respir Med. 2018; 6:379–388. doi: https://doi.org/10.1016/S2213-2600(18)30052-3 PMID: 29496485

100. Story Jovanova O, Nedeljkovic I, Spieler D, Walker RM, Liu C, Luciano M, et al. DNA Methylation Signatures of Depressive Symptoms in Middle-aged and Elderly Persons: Meta-analysis of Multiethnic Epigenome-wide Studies. JAMA Psychiatry. 2018; 75:949–959. doi: https://doi.org/10.1001/jamapsychiatry.2018.1725 PMID: 29998287

101. Kresovich JK, Xu Z, O'Brien KM, Shi M, Weinberg CR, Sandler DP, et al. Blood DNA methylation profiles improve breast cancer prediction. Mol Oncol. 2022; 16:42–53. doi: https://doi.org/10.1002/1878-0261.13087 PMID: 34411412

102. Odintsova VV, Rebattu V, Hagenbeek FA, Pool R, Beck JJ, Ehli EA, et al. Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. Front Psychiatry. 2021; 12:688464. doi: https://doi.org/10.3389/fpsyt.2021.688464 PMID: 34393852

103. Thompson M, Hill BL, Rakocz N, Chiang JN, et al. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. medRxiv. 2022. doi: https://doi.org/10.1038/s41525-022-00320-1 PMID: 36008412

104. Chen F, Marquez H, Kim Y- K, Qian J, Shao F, Fine A, et al. Prenatal retinoid deficiency leads to airway hyperresponsiveness in adult mice. J Clin Invest. 2014; 124:801–811. doi: https://doi.org/10.1172/JCI70291 PMID: 24401276

105. Foong RE, Bosco A, Jones AC, Gout A, Gorman S, Hart PH, et al. The effects of in utero vitamin D deficiency on airway smooth muscle mass and lung function. Am J Respir Cell Mol Biol. 2015; 53:664–675. doi: https://doi.org/10.1165/rcmb.2014-0356OC PMID: 25867172

106. Koestler DC, Avissar-Whiting M, Houseman EA, Karagas MR, Marsit CJ. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. Environ Health Perspect. 2013; 121:971–977. doi: https://doi.org/10.1289/ehp.1205925 PMID: 23757598

107. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clin Epigenetics. 2019; 11:125. doi: https://doi.org/10.1186/s13148-019-0717-y PMID: 31455416

108. Kozul CD, Ely KH, Enelow RI, Hamilton JW. Low-dose arsenic compromises the immune response to influenza A infection in vivo. Environ Health Perspect. 2009; 117:1441–1447. doi: https://doi.org/10.1289/ehp.0900911 PMID: 19750111

109. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011; 88:450–457. doi: https://doi.org/10.1016/j.ajhg.2011.03.003 PMID: 21457905

110. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015; 7:81. doi: https://doi.org/10.1186/s13148-015-0113-1 PMID: 26246861

**111.** Bauer M, Hackermüller J, Schor J, Schreiber S, Fink B, Pierzchalski A, et al. Specific induction of the unique GPR15 expression in heterogeneous blood lymphocytes by tobacco smoking. Biomarkers. 2019; 24:217–224. doi: https://doi.org/10.1080/1354750X.2018.1539769 PMID: 30387691

**112.** Bakulski KM, Dou J, Lin N, London SJ, Colacino JA. DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood. Sci Rep. 2019; 9:4576. doi: https://doi.org/10.1038/s41598-019-40963-2 PMID: 30872662

**113.** Villani A- C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017:356. doi: https://doi.org/10.1126/science.aah4573 PMID: 28428369

**114.** Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012; 13:86. doi: https://doi.org/10.1186/1471-2105-13-86 PMID: 22568884

**115.** Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:453–457. doi: https://doi.org/10.1038/nmeth.3337 PMID: 25822800

**116.** MacParland SA, Liu JC, Ma X- Z, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018; 9:4383. doi: https://doi.org/10.1038/s41467-018-06318-7 PMID: 30348985

**117.** Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu NT, et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. Nature. 2019; 575:512–518. doi: https://doi.org/10.1038/s41586-019-1631-3 PMID: 31597160

**118.** Dangleben NL, Skibola CF, Smith MT. Arsenic immunotoxicity: a review. Environ Health. 2013; 12:73. doi: https://doi.org/10.1186/1476-069X-12-73 PMID: 24004508

**119.** Cordes W. Experiences with plasmochin in malaria (preliminary reports). 15th Annual Report. Boston, MA: Boston United Fruit Company; 1926. p. 66–71.

**120.** Earle DP, Bigelow FS, Zubrod CG, Kane CA. Studies on the chemotherapy of the human malarias. ix. effect of pamaquine on the blood cells of man. J Clin Invest. 1948; 27:121–129. doi: https://doi.org/10.1172/JCI101950 PMID: 16695624

**121.** Alving AS, Carson PE, Flanagan CL, Ickes CE. Enzymatic deficiency in primaquine-sensitive erythrocytes. Science. 1956; 124:484–485. doi: https://doi.org/10.1126/science.124.3220.484-a PMID: 13360274

**122.** Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001; 293:455–462. doi: https://doi.org/10.1126/science.1061573 PMID: 11423617

**123.** Cappadoro M, Giribaldi G, O'Brien E, Turrini F, Mannu F, Ulliers D, et al. Early phagocytosis of glucose-6-phosphate dehydrogenase (G6PD)-deficient erythrocytes parasitized by Plasmodium falciparum may explain malaria protection in G6PD deficiency. Blood. 1998; 92:2527–2534. doi: https://doi.org/10.1182/blood.V92.7.2527 PMID: 9746794

**124.** Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schäfer H. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet. 2008; 16:1164–1172. doi: https://doi.org/10.1038/ejhg.2008.106 PMID: 18523454

**125.** Czamara D, Eraslan G, Page CM, Lahti J, Lahti-Pulkkinen M, Hämäläinen E, et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nat Commun. 2019; 10:2548. doi: https://doi.org/10.1038/s41467-019-10461-0 PMID: 31186427

**126.** Gauderman WJ, Mukherjee B, Aschard H, Hsu L, Lewinger JP, Patel CJ, et al. Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. Am J Epidemiol. 2017; 186:762–770. doi: https://doi.org/10.1093/aje/kwx228 PMID: 28978192

**127.** Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005; 6:287–298. doi: https://doi.org/10.1038/nrg1578 PMID: 15803198

**128.** Ritz BR, Chatterjee N, Garcia-Closas M, Gauderman WJ, Pierce BL, Kraft P, et al. Lessons Learned From Past Gene-Environment Interaction Successes. Am J Epidemiol. 2017; 186:778–786. doi: https://doi.org/10.1093/aje/kwx230 PMID: 28978190

**129.** Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet. 2008; 40:1461–1465. doi: https://doi.org/10.1038/ng.257 PMID: 18820647

**130.** Stender S, Kozlitina J, Nordestgaard BG, Tybjærg-Hansen A, Hobbs HH, Cohen JC. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. Nat Genet. 2017; 49:842–847. doi: https://doi.org/10.1038/ng.3855 PMID: 28436986

131. Marcus JH, Novembre J. Visualizing the geography of genetic variants. Bioinformatics. 2017; 33:594–595. doi: https://doi.org/10.1093/bioinformatics/btw643 PMID: 27742697

132. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019; 570:514–518. doi: https://doi.org/10.1038/s41586-019-1310-4 PMID: 31217584

133. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013; 93:779–797. doi: https://doi.org/10.1016/j.ajhg.2013.10.012 PMID: 24210251

134. Mägi R, Horikoshi M, Sofer T, Mahajan A, Kitajima H, Franceschini N, et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum Mol Genet. 2017; 26:3639–3650. doi: https://doi.org/10.1093/hmg/ddx280 PMID: 28911207

135. Zou J, Hormozdiari F, Jew B, Castel SE, Lappalainen T, Ernst J, et al. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. PLoS Genet. 2019; 15:e1008481. doi: https://doi.org/10.1371/journal.pgen.1008481 PMID: 31834882

136. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci USA. 2012; 109:1204–1209. doi: https://doi.org/10.1073/pnas.1115761109 PMID: 22233810

137. Kim S, Becker J, Bechheim M, Kaiser V, Noursadeghi M, Fricker N, et al. Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. Nat Commun. 2014; 5:5236. doi: https://doi.org/10.1038/ncomms6236 PMID: 25327457

138. Lee MN, Ye C, Villani A- C, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014; 343:1246980. doi: https://doi.org/10.1126/science.1246980 PMID: 24604203

139. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014; 343:1246949. doi: https://doi.org/10.1126/science.1246949 PMID: 24604202

140. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell. 2016; 167:657–669.e21. doi: https://doi.org/10.1016/j.cell.2016.09.025 PMID: 27768889

141. Quach H, Rotival M, Pothlichet J, Loh Y- HE, Dannemann M, Zidane N, et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. Cell. 2016; 167:643–656.e17. doi: https://doi.org/10.1016/j.cell.2016.09.024 PMID: 27768888

142. Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, Kundu K, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018; 50:424–431. doi: https://doi.org/10.1038/s41588-018-0046-7 PMID: 29379200

143. Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc Natl Acad Sci USA. 2018; 115:E488–E497. doi: https://doi.org/10.1073/pnas.1714765115 PMID: 29282317

144. Knowles DA, Davis JR, Edgington H, Raj A, Favé M- J, Zhu X, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods. 2017; 14:699–702. doi: https://doi.org/10.1038/nmeth.4298 PMID: 28530654

145. Moyerbrailean GA, Richards AL, Kurtz D, Kalita CA, Davis GO, Harvey CT, et al. High-throughput allele-specific expression across 250 environmental conditions. Genome Res. 2016; 26:1627–1638. doi: https://doi.org/10.1101/gr.209759.116 PMID: 27934696

146. Fiorotto R, Amenduni M, Mariotti V, Fabris L, Spirli C, Strazzabosco M. Liver diseases in the dish: iPSC and organoids as a new approach to modeling liver diseases. Biochim Biophys Acta Mol basis Dis. 2019; 1865:920–928. doi: https://doi.org/10.1016/j.bbadis.2018.08.038 PMID: 30264693

147. Huch M, Knoblich JA, Lutolf MP, Martinez-Arias A. The hope and the hype of organoid research. Development. 2017; 144:938–941. doi: https://doi.org/10.1242/dev.150201 PMID: 28292837

148. Philibert RA, Terry N, Erwin C, Philibert WJ, Beach SR, Brody GH. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. Clin Epigenetics. 2014; 6:28. doi: https://doi.org/10.1186/1868-7083-6-28 PMID: 25859287

149. Bien SA, Wojcik GL, Zubair N, Gignoux CR, Martin AR, Kocarnik JM, et al. Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. PLoS ONE. 2016; 11:e0167758. doi: https://doi.org/10.1371/journal.pone.0167758 PMID: 27973554

150. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low-depth whole-genome sequencing in complex trait association studies. Bioinformatics. 2019; 35:2555–2561. doi: https://doi.org/10.1093/bioinformatics/bty1032 PMID: 30576415

151. Tehranchi A, Hie B, Dacre M, Kaplow I, Pettie K, Combs P, et al. Fine-mapping cis-regulatory variants in diverse human populations. elife. 2019: 8. doi: https://doi.org/10.7554/eLife.39595 PMID: 30650056

152. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019; 51:1494–1505. doi: https://doi.org/10.1038/s41588-019-0505-9 PMID: 31570894

153. Vijayabaskar MS, Goode DK, Obier N, Lichtinger M, Emmett AML, Abidin FNZ, et al. Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. PLoS Comput Biol. 2019; 15:e1007337. doi: https://doi.org/10.1371/journal.pcbi.1007337 PMID: 31682597

154. Johnston AD, Simões-Pires CA, Thompson TV, Suzuki M, Greally JM. Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding. Nat Commun. 2019; 10:3472. doi: https://doi.org/10.1038/s41467-019-11412-5 PMID: 31375681

155. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48:1279–1283. doi: https://doi.org/10.1038/ng.3643 PMID: 27548312

156. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. Nat Genet. 2019; 51:1349–1355. doi: https://doi.org/10.1038/s41588-019-0487-7 PMID: 31477931

157. Jaric I, Rocks D, Greally JM, Suzuki M, Kundakovic M. Chromatin organization in the female mouse brain fluctuates across the oestrous cycle. Nat Commun. 2019; 10:2851. doi: https://doi.org/10.1038/s41467-019-10704-0 PMID: 31253786

158. Lee JW, Ko J, Ju C, Eltzschig HK. Hypoxia signaling in human diseases and therapeutic targets. Exp Mol Med. 2019; 51:1–13. doi: https://doi.org/10.1038/s12276-019-0235-1 PMID: 31221962

159. Stefano GB, Challenger S, Kream RM. Hyperglycemia-associated alterations in cellular signaling and dysregulated mitochondrial bioenergetics in human metabolic disorders. Eur J Nutr. 2016; 55:2339–2345. doi: https://doi.org/10.1007/s00394-016-1212-2 PMID: 27084094

160. Ron D, Messing RO. Signaling pathways mediating alcohol effects. Curr Top Behav Neurosci. 2013; 13:87–126. doi: https://doi.org/10.1007/7854_2011_161 PMID: 21877259

161. Sun S, Li H, Chen J, Qian Q. Lactic Acid: No Longer an Inert and End-Product of Glycolysis. Physiology (Bethesda). 2017; 32:453–463. doi: https://doi.org/10.1152/physiol.00016.2017 PMID: 29021365

162. Park J-M, Jo S-H, Kim M-Y, Kim T-H, Ahn Y-H. Role of transcription factor acetylation in the regulation of metabolic homeostasis. Protein Cell. 2015; 6:804–813. doi: https://doi.org/10.1007/s13238-015-0204-y PMID: 26334401

163. Carr SM, Poppy Roworth A, Chan C, La Thangue NB. Post-translational control of transcription factors: methylation ranks highly. FEBS J. 2015; 282:4450–4465. doi: https://doi.org/10.1111/febs.13524 PMID: 26402372

164. Boshnjaku V, Shim K-W, Tsurubuchi T, Ichi S, Szany EV, Xi G, et al. Nuclear localization of folate receptor alpha: a new role as a transcription factor. Sci Rep. 2012; 2:980. doi: https://doi.org/10.1038/srep00980 PMID: 23243496

165. Ge C, Ye J, Zhang H, Zhang Y, Sun W, Sang Y, et al. Dmrt1 induces the male pathway in a turtle species with temperature-dependent sex determination. Development. 2017; 144:2222–2233. doi: https://doi.org/10.1242/dev.152033 PMID: 28506988

166. Sever R, Glass CK. Signaling by nuclear receptors. Cold Spring Harb Perspect Biol. 2013; 5:a016709. doi: https://doi.org/10.1101/cshperspect.a016709 PMID: 23457262

167. Edwards DR. Cell signalling and the control of gene transcription. Trends Pharmacol Sci. 1994; 15:239–244. doi: https://doi.org/10.1016/0165-6147(94)90318-2 PMID: 7940986

168. Kabir MH, Patrick R, Ho JWK, O'Connor MD. Identification of active signaling pathways by integrating gene expression and protein interaction data. BMC Syst Biol. 2018; 12:120. doi: https://doi.org/10.1186/s12918-018-0655-x PMID: 30598083

169. Hagenbuchner J, Ausserlechner MJ. Targeting transcription factors by small compounds—Current strategies and future implications. Biochem Pharmacol 2016; 107:1–13. doi: https://doi.org/10.1016/j.bcp.2015.12.006 PMID: 26686579

170. Lambert M, Jambon S, Depauw S, David-Cordonnier M-H. Targeting transcription factors for cancer treatment. Molecules. 2018: 23. doi: https://doi.org/10.3390/molecules23061479 PMID: 29921764

171. Bushweller JH. Targeting transcription factors in cancer—from undruggable to reality. Nat Rev Cancer. 2019; 19:611–624. doi: https://doi.org/10.1038/s41568-019-0196-7 PMID: 31511663

172. Tanaka N, Aoyama T, Kimura S, Gonzalez FJ. Targeting nuclear receptors for the treatment of fatty liver disease. Pharmacol Ther. 2017; 179:142–157. doi: https://doi.org/10.1016/j.pharmthera.2017.05.011 PMID: 28546081

173. Yang HW, Chung M, Kudo T, Meyer T. Competing memories of mitogen and p53 signalling control cell-cycle entry. Nature. 2017; 549:404–408. doi: https://doi.org/10.1038/nature23880 PMID: 28869970

174. Plomin R. Commentary: missing heritability, polygenic scores, and gene-environment correlation. J Child Psychol Psychiatry. 2013; 54:1147–1149. doi: https://doi.org/10.1111/jcpp.12128 PMID: 24007418

175. Young AI. Solving the missing heritability problem. PLoS Genet. 2019; 15:e1008222. doi: https://doi.org/10.1371/journal.pgen.1008222 PMID: 31233496

176. Smith GD. Epidemiology, epigenetics and the "Gloomy Prospect": embracing randomness in population health research and practice. Int J Epidemiol. 2011; 40:537–562. doi: https://doi.org/10.1093/ije/dyr117 PMID: 21807641