

REVIEW

Open Access



An overview of DNA methylation-derived trait score methods and applications

Marta F. Nabais^{1,2}, Danni A. Gadd³, Ellis Hannon², Jonathan Mill², Allan F. McRae¹ and Naomi R. Wray^{1,4*} 

*Correspondence:
Naomi.wray@uq.edu.au

¹ Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

² University of Exeter Medical School, RILD Building, RD&E Hospital Wonford, Barrack Road, Exeter EX2 5DW, UK

³ Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK

⁴ Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia

Abstract

Microarray technology has been used to measure genome-wide DNA methylation in thousands of individuals. These studies typically test the associations between individual DNA methylation sites (“probes”) and complex traits or diseases. The results can be used to generate methylation profile scores (MPS) to predict outcomes in independent data sets. Although there are many parallels between MPS and polygenic (risk) scores (PGS), there are key differences. Here, we review motivations, methods, and applications of DNA methylation-based trait prediction, with a focus on common diseases. We contrast MPS with PGS, highlighting where assumptions made in genetic modeling may not hold in epigenetic data.

Introduction

The most characterized epigenetic marker is DNA methylation (DNAm). DNAm is a reversible modification to DNA involving the covalent addition of methyl groups (CH₃) to the fifth carbon position of cytosine by DNA methyltransferases. In mammals, DNAm predominantly occurs at cytosine-guanine dinucleotides (CpGs). In some instances, DNAm can block the binding of transcription factors to DNA and is therefore associated with reduced gene expression [1]. DNAm is primarily detected through the conversion of unmethylated cytosines with sodium bisulphite to uracil, allowing methylated and unmethylated cytosines to be distinguished using array-based or sequencing-based technologies. On each individual DNA molecule, cytosines are either methylated or not, and so measurements of DNAm in bulk tissue (such as DNA extracted from whole blood) are averages across DNA molecules from many cells. Hence, reported measures of DNAm are proportion-related values. DNAm arrays capture a small proportion (~ 2%) of possible DNAm sites (including some non-CpG sites) in the genome, with each site targeted by a “probe.” However, in current commercial arrays, these probes have been selected to be informative, being annotated to 96% of coding genes, as well as targeting known enhancer and promoter elements. The measurement of DNAm by array technology is cheaper and more high-throughput than by sequencing [2] and so relatively large human



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

data sets have been generated from array technology [3] to identify probes which show differential DNAm at CpG sites associated with traits. These methylation-wide association study (MWAS) data sets are many-fold smaller than genome-wide association study (GWAS) data sets that rely predominantly on SNP array data. Consortia are being established to bring together data sets for meta-analysis (e.g., [4, 5]).

While the DNA sequence is stable across cell types throughout lifetime (other than somatic mutations), DNAm varies between cell types (within an individual) and between people (within a cell type) for several reasons (Fig. 1). First, there are cell type-specific DNAm patterns which provide “fingerprints” of cell-type lineages [6, 7]; hence, DNAm at relevant sites can be used to determine the cell type of origin in mixed cell-type samples [8]. Second, at some genomic locations, DNA polymorphisms are associated with DNAm [9–12]. These polymorphisms are termed methylation quantitative trait loci (mQTLs). While most SNPs only confer a small effect on DNAm variation [4], some associations are strong (up to 2 standard deviation units/allele [4]) and likely imply a

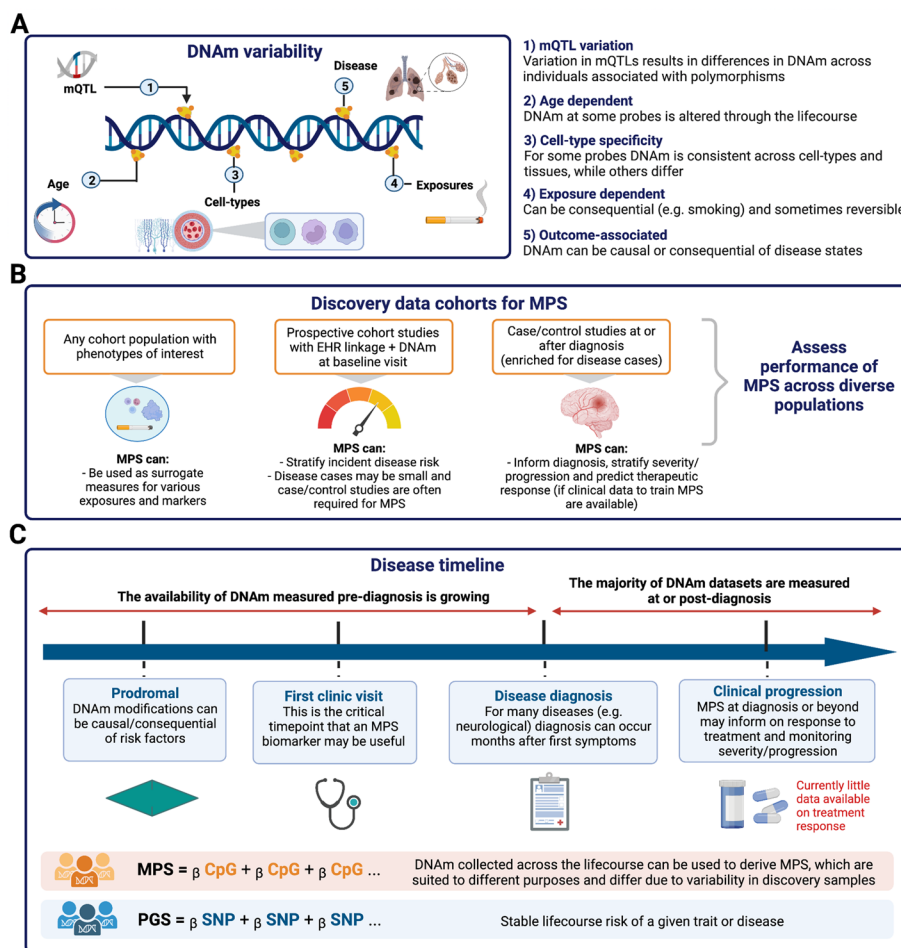


Fig. 1 Methylation profile scores (MPS). **a** Sources of variability in DNAm measurements that inform the signal captured by scoring approaches. **b** Types of discovery cohort samples and their uses for the development of MPS. **c** Disease timeline: utility of DNAm scores depends on when blood is sampled. Created with BioRender.com. MPS, methylation profile score; PGS, polygenic score; mQTLs, methylation quantitative trait loci

direct causal relationship between DNA variation and DNAm in *cis* (where *cis* implies a close proximity on the chromosome between the DNA polymorphism and the site of methylation). Measures of DNAm at mQTLs are correlated between relatives, i.e., they are heritable quantitative traits [13]. Third, at some CpG sites, the DNA methylation levels are strongly associated with age [14], whereas other sites undergo changes in DNAm in response to environmental exposures. The effect of smoking on DNAm is most well-characterized [15–18]. While DNAm is, in general, considered to be a reversible modification, analyses of longitudinal DNAm have shown that DNAm at some sites can be variable between people but have high within-person consistency over time, including at sites not known to be influenced by a mQTL [13]. If DNAm inter-individual variation in the blood is associated with disease, or with non-measured risk factors for disease, it could be a biomarker for disease or disease progression.

It is notable that the use of DNAm technology in the context of a cancer diagnosis from blood samples is already well-advanced [19]. The cell-free DNA isolated from the blood carries tissue-specific signatures that track the cancer-associated cell death. The Epi proColon[®] (for the detection of colorectal cancer) already has FDA approval, and the Galleri[®] multi-cancer test is in clinical trials. Detection of DNAm signatures in cell-free DNA from other cell death-associated diseases, such as neurodegenerative disease, is an active area of research [8].

In this review, we focus on DNAm measured by array technology. We consider how MWAS data sets can be used to calculate trait-associated methylation profiles scores (MPS) in independent data sets that have DNAm measured. MPS are also referred to as episcores [20] or simply epigenetic or DNAm predictors [15]. While there many parallels between MPS and polygenic (risk) scores (PGS) (calculated for trait prediction from the results of GWAS), there are key differences. Our target audiences are those familiar with the construction and interpretation of PGS. Here, we review the motivations, methods, and applications of MPS. We contrast MPS with PGS, highlighting where assumptions made in genetic modeling may not hold in epigenetic data [21–24]. For another recent review on the use of MPS in health applications, see Yousefi et al. [25]. We start by introducing the concept of MPS, then step through technical considerations that contribute to differences in MPS and PGS.

Methylation profile scores (MPS)

Evaluation of trait prediction using DNAm data requires at least two independent data sets with measures of both genome-wide DNAm and the trait of interest (Fig. 2). The MWAS “discovery” sample is used to identify DNAm probes associated with the trait, resulting in a list of probes and weights. These can be used to construct a MPS for each individual in the independent “target” sample. The utility of the trait prediction is evaluated by the association of the MPS with the directly measured trait in this target sample. For MPS, as for PGS, there is no requirement that the scores represent functional or causal mechanisms, but simply that an association is found in independent data. If a trait association is demonstrated, then MPS can be calculated in individuals who have DNAm data but for whom the trait value is unknown and hence used as a trait biomarker. For both PGS and MPS, the accuracy of prediction may be limited, but the signals carried by the scores could still have utility. Accuracy of

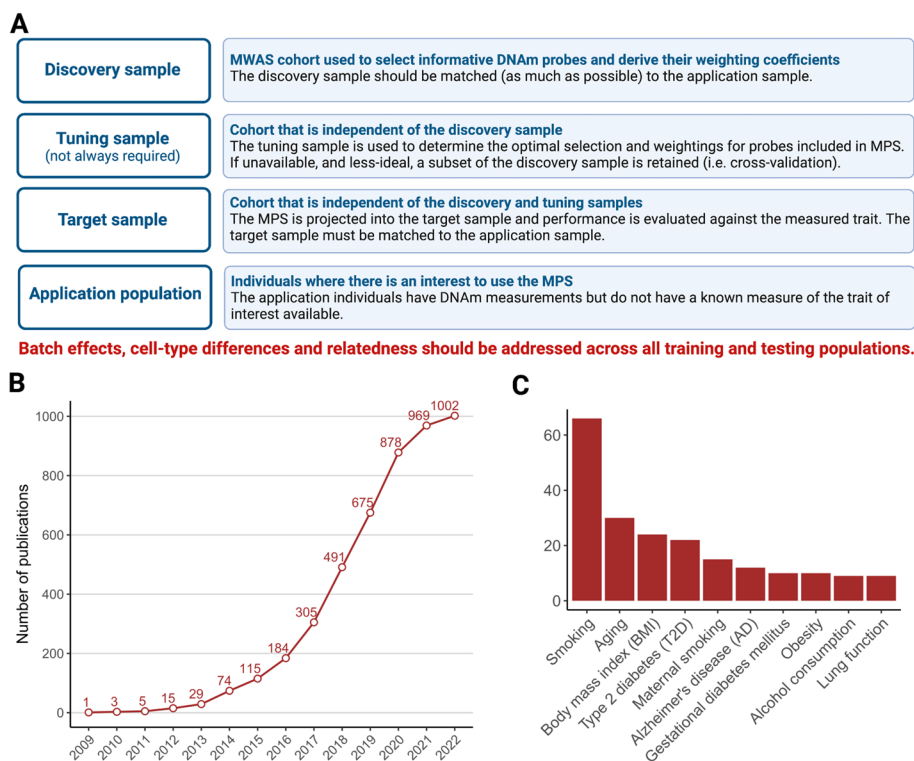


Fig. 2 MPS data sets. **a** Data set definitions. **b** Curated statistics on number of methylome-wide association studies (MWAS) publications (y-axis) per-year and **c** per-trait (excluding cancer) taken from EWAS Atlas database, on 10 January 2023 (<https://ngdc.cncb.ac.cn/ewas/downloads>). **a** Created with BioRender.com

prediction can be maximized by combining PGS, MPS, and other known risk factors. Even such a combined predictor is likely to have high prediction error for a specific individual and so the utility is likely to be at the level of stratification, in which a high-risk group will be enriched for those who go on to have disease. Sometimes, a “tuning” sample is needed in the derivation of MPS; an MWAS data set independent of the discovery and target samples is used to optimize probe selection. Since such data sets are not usually available, a subset of the discovery sample can be removed for use as a tuning sample (also known as cross-validation). We also use the term “application” sample to refer to the population in which an MPS could be used as a biomarker.

We use the word trait “prediction” for MPS with hesitation since it is important to emphasize a key difference between MPS and PGS. In principle, PGS can be calculated at birth and will not change over lifetime (unless the SNPs and their weights used to construct the PGS are updated). PGS can therefore be considered as predictors of future, not-yet observed events. Since measured DNAm levels can fluctuate, any trait-specific MPS for an individual can change throughout an individual’s lifetime. Hence, much more thought is needed, compared to PGS applications, about the time point at which biological samples (e.g., blood) are taken in the discovery, target, and final application samples as to whether the MPS developed can be considered as a predictor of a future event. While some MPS scenarios will fit the criteria of prediction, we emphasize that this is not always the case. Consider the goal

of using DNAm as a blood biomarker to aid in early risk stratification of subsequent (incident) disease onset. To have clinical utility as a diagnostic biomarker, the MPS needs to be valid in blood samples taken at or before the time diagnosis is achieved under current practice (i.e., prospective samples) (Fig. 1), whereas most MWAS disease cohorts have been collected after diagnosis and MPS derived from these may be confounded with consequences of later disease processes (including treatment). Such prospective DNAm data sets are not yet common. One of the largest cohorts to date is Generation Scotland (an adult community cohort) which has 18,413 individuals and electronic health linkage spanning 15 years after blood sampling and has been used to investigate MPS associations with the incidence of 11 major morbidities, including type 2 diabetes [3, 20]. The CHARGE consortium has reported MPS in relation to incident coronary heart disease in a follow-up of 11.2 years, on average, following sample collection using data from 9 cohorts comprising 11,461 individuals [26]. Another example of prospectively measured DNAm is from the use of Guthrie card blood spots, stored in some countries for all babies born. These provide an unbiased resource allowing study designs that contrast DNAm at birth in those that go on to get diagnoses in later childhood compared to matched controls. For example, the MINERVA study measured DNAm at birth in 1,293 individuals, 50% of whom were later diagnosed with autism spectrum disorders [27] (although in this example, no differences in DNAm at birth associated with the autism diagnosis were found). Currently, most MWAS of disease use biological samples collected post-diagnosis, and so, the MWAS-identified disease-associated DNAm probes may reflect an advanced stage of the disease trajectory or even a consequence of diagnosis (including treatment). Hence, MPS derived and evaluated in post-diagnosis samples may not be useful for disease prediction in currently healthy individuals or people presenting with the first symptoms of the disease. Careful evaluation is required using blood samples taken at a time relevant to diagnosis in real-life settings. DNAm taken at the time of diagnosis could be useful in predicting disease severity, disease progression, or therapeutic response post-diagnosis if longitudinal clinical data are available to develop predictors.

Tissue sample considerations

For GWAS and PGS, the tissue (or cell type) from which DNA is derived is rarely considered since germline-inherited DNA polymorphisms are the same in all tissues (somatic mutations require non-standard analysis to detect from array data). While tissue source (e.g., blood vs saliva) can be identified in the principal component analysis of SNP array data pre-quality control (QC), differences are handled through routine QC steps and do not impact upon genotype calls. In contrast, cell type-specific DNAm is a major contributor to DNAm variation, and tissues from which DNA is usually isolated for genomic analysis comprise a heterogeneous mix of different cell types. In a later section, we consider issues associated with cell-type proportions from the whole blood in the context of MPS.

The majority of large MWAS studies, to date, have quantified DNAm in the whole blood, which is an easily accessible tissue relevant to biomarker development. Notably, DNAm can be measured accurately from dried blood spots with good concordance with

measures from matched blood samples [28]. To achieve large cohorts for MWAS and MPS in the future, less-invasive sampling alternatives may need to be adopted. Nasal, buccal, and saliva samples from adults can be collected in a decentralized way through postal kits. These tissue types are also well-suited to DNAm studies of preterm infants, babies, and children, and distinct DNAm signatures have been identified in buccal and saliva samples for gestational age and preterm status [29, 30]. Leukocytes and squamous epithelial cells are typically found in the oral cavity [31], and several methods exist to adjust for this cellular heterogeneity [32, 33]. However, the transferability of MPS derived from the blood to DNAm derived from less-invasive tissues requires specific investigation. A comparison of DNAm from whole blood, buccal epithelial, and nasal epithelial (sampled at the same time from the same individuals) showed good concordance at mQTL but not at other DNAm variable sites used in MPS [34]. Notably, the commonly used Horvath DNAmAge Epigenetic Clock (which was derived from DNAm measurements in 51 tissue types) exhibited variability across these 3 tissues and epigenetic age MPS from the blood were the closest to actual age [34]. In summary, MPS developed in one tissue type is not necessarily transferable to another tissue type without careful evaluation.

Parallels and differences in array-based measures of DNAm and SNPs

To date, Illumina is the only commercial provider of DNAm arrays, and their most recent “EPIC” array is designed to measure DNAm at over 850K genomic locations [35, 36]. Array-based measures of DNAm are considered to have some advantages compared to current whole-genome bisulphite sequencing technologies. First, if the study goal is MWAS, then cost is a big factor; the Illumina list price per sample for DNAm array is USD330 compared to the list price of USD5000 for whole-genome bisulphite sequencing (of course dependent on read depth). Other advantages include their fixed content (designed to have probes for CpG islands in the gene-regulatory regions of the majority of genes), considerably smaller data files which are cheaper to store and analyze, and standardized quantification measured by *M*-values and beta values. *M*-values are the \log_2 ratio of the intensities of the methylated versus unmethylated probes at a given CpG site, across all cells sampled, with positive values indicating a site is more methylated than unmethylated [37]. Beta values represent the proportion of methylated probes across all cells in the sample and therefore range from 0 to 100%. Conversions between these two DNAm measurement types can be performed with ease [37]. Beta values are considered more interpretable owing to their scale [37].

The Illumina iScan System is designed to read both SNP and DNAm arrays, although the probe design is more complex for DNAm analysis [38]. While between-batch technical differences now have minimal impact on SNP-genotype calling, DNAm batch effects are much more apparent. Careful pre-processing of DNAm data is required prior to downstream analyses, and we refer readers to key papers [38–41], but where possible, it is best to employ consistent laboratory protocols. Briefly, DNAm is assayed on plates (e.g., comprising 4 Illumina chips, 8 samples per chip), with multiple sets of plates that are run at the same time forming a batch. In turn, when many samples are available from a study, multiple batch runs are required which combine to form DNAm sets. If all samples in a given cohort are not run at the same time, the effects of processing batch

may be a major confounder in the analyses [42]. Hence, the DNAm at each probe are pre-processed and normalized prior to downstream analyses, but since substantial batch effects can still remain [43] set and batch are often also fitted as covariates in MWAS. Successful evaluation of MPS from discovery to target to application samples is more likely if similar technical protocols are used. Systematic evaluation of 41 MPS across 101 different DNAm pre-processing and normalization strategies has highlighted the impact of technical variation on MPS [44].

It is notable that DNAm arrays have traditionally been priced several-fold higher than SNP arrays, which has been a contributing factor to the smaller sample sizes for MWAS compared to GWAS. As a likely consequence of the smaller market, there has been less activity in the development of DNAm array content compared to SNP arrays. We understand that Thermo Fisher has a product in development which may provide competition in array content and price, which may drive DNAm measurement in larger cohorts. Although the EPIC array includes 850K probes, 120K are reported as non-variable between individuals in the blood [40], and further QC steps that retain only variable sites useful for use in MPS have reduced the number of probes used in practice to ~370K [45]. Currently, many studies use a combination of Illumina EPIC and 450K arrays and so use the intersection of probes (~450K) and of these ~200K probes are retained as being sufficiently variable for use in blood-based MWAS and MPS [45, 46].

Parallels and differences of MWAS and GWAS

The probes and their weights used in MPS are generated from the discovery MWAS data; hence, guidelines for the optimal design of MWAS [23] are critical when the goal is trait prediction. We do not consider the design of MWAS but refer readers to a set of key reviews [19, 20, 47–49]. These are essential reading, since quoting Mill and Heijmans [19] “It would be naive to assume that we can simply undertake MWAS analyses on samples that have been previously used in GWAS.” GWAS designs have relatively few constraints since DNA polymorphisms (mostly) do not change over the lifetime, and indeed control samples can be deliberately selected to include people who are older and hence past the age of onset for the disease studied. In contrast, optimal MWAS designs need to follow the sort of practices implemented for observational studies including appropriate ascertainment matching of cases and controls. For example, the selection of older controls is not suitable for an MWAS study given many sites change methylation levels with age. For many diseases and disorders, case status may be associated with body mass index or smoking (e.g., [50]). As discussed above (and Figs. 1 and 2), the discovery sample MWAS used to develop MPS ideally has the properties relevant to the target and final application samples where the MPS are validated and properties relevant to the situations where the MPS are applied.

In GWAS, the phenotype is always the dependent variable ($y \sim \text{SNP}$), but in MWAS, sometimes DNAm is analyzed as the dependent variable ($\text{DNAm} \sim y$), reflecting the two possible directions of dependency. For the purposes of trait prediction, we assume the $y \sim \text{DNAm}$ model for analysis. For example, while DNAm changes associated with smoking most likely reflect smoking being causal for the changes, supporting the logic of the $\text{DNAm} \sim y$ model, when the goal is to develop an MPS in independent data for those with

unknown smoking status the analysis model must be $y \sim \text{DNAm}$. While DNAm measures are continuous they may not be normally distributed [40].

In GWAS, the genomic inflation factor (λ , ratio of the observed median test statistic to the expected median test) is used to demonstrate that the quality control pipeline has retained no residual population stratification associated with the trait that could generate the identification of false positives [51]. Under no residual stratification, λ is expected to be 1, since it is reasonable to assume that fewer than half of the sites are truly associated. MWAS data must be afforded more consideration in this regard. The simulations conducted by Zhang et al. [52] illustrate the issues. They used real MWAS data from a healthy cohort of volunteers from the Lothian (Scotland) birth cohorts, so ancestrally homogeneous. They simulated causal associations for a simulated trait onto measures of DNAm made at probes located only on even chromosomes. They then conducted MWAS analysis examining evidence for association only on odd chromosomes, finding strong evidence for the association for their simulation scenario ($\lambda = 7.67$). This observation reflects, in part, cell type proportion differences between individuals with cell type-specific methylation patterns correlated across chromosomes. However, including cell-type proportion values (directly measured, not MPS predicted) as covariates only reduced the λ to 4.95, implying other factors (technical or biological) generate a correlation of DNAm across chromosomes. Many MWAS methods have been introduced with the goal to control for unmeasured cell-type proportions and other unmeasured potential confounders, e.g., [53, 54]. In the Zhang et al. simulation, if the first 5 principal components from a correlation matrix of DNAm across individuals were included as covariates, still the λ only reduced to 1.67. They introduced linear mixed model methods to estimate the effect of each probe while controlling for the background genome-wide DNAm, which reduced λ to the expected value of 1. These simulations were conducted for a quantitative trait. In real data, while technical confounding with a quantitative trait is not expected, for binary disease traits, this can be problematic. For example, in case-control cohorts, the DNA collection and processing of cases and controls are frequently achieved by different protocols which can lead to technical confounding in DNAm levels (more so than for DNA polymorphisms that use the same array technology [55]). Statistical methods can be employed to account for confounding and to avoid the detection of false-positive associations, but for binary traits, the confounding can be too complete, so careful experimental design is a more effective approach to avoid the potential consequences of confounding [42].

Parallels and differences of MPS and PGS

PGS are calculated as a weighted sum of trait-associated alleles [56], so for the i th individual, the PGS is $PGS_i = \sum_j^{m_{PGS}} \hat{\beta}_j \times SNP_{ij}$, where $\hat{\beta}_j$ is the estimated effect size for SNP j which has values of $SNP_{ij} = 0, 1$ or 2 alleles in individual i . Similarly, MPS are calculated as $MPS_i = \sum_j^{m_{MPS}} \hat{b}_j \times CpG_{ij}$ where \hat{b}_j is the estimated effect size for probe j and CpG_{ij} is the methylation value of probe j in the i th individual. CpG_{ij} is a continuous measure (proportion of cells that are methylated). From the same GWAS data, different statistical (or machine learning) methods can be used to generate PGS with the methods making different choices about how many SNPs to include, which SNPs to include, and what weights to allocate to the risk alleles (e.g., see [57] for a comparison of PGS

methods). Similarly, from the same MWAS data, different statistical methods can be used to generate MPS with the methods differing on how many probes, which probes, and what weights to allocate to DNAm values (Table 1, Additional file 1). Neither PGS nor MPS are restricted to SNPs/probes that are associated at the level of genome-wide significance, and SNP/probes selected may not be biologically meaningful. For both PGS and MPS, many combinations of SNP/probes can give very similar out-of-sample

Table 1 Methods used in publication derivation of DNA methylation profile scores (MPS)

Method: software	Brief summary (see Additional file 1 for a more detailed summary)	Literature examples of the application of MPS
CP+T after linear regression	Marginal effects of DNAm sites derived from linear regression. Selected probes have MWAS association p -value less than a defined threshold. In a greedy algorithm, the most associated probe is selected first. Other probes are selected if correlation (r) with any genomically local probe already selected is less than a defined threshold. The results are often reported from the p -value threshold that generates the highest out-of-sample prediction, but to avoid a winner's curse effect, a single p -value threshold should be applied in the target cohort identified from MPS results applied in an independent tuning cohort.	BMI, height [58] schizophrenia [59] C-reactive protein levels [60] Interleukin-6 [61]
Penalized linear regression: glmnet [62, 63]	In ridge regression/lasso/elastic, net probe effect sizes are estimated jointly. In ridge regression, linear regression estimates are shrunk (dependent on penalty parameter λ_1). In lasso, a proportion of probes have an effect size set to 0 (dependent on penalty parameter λ_2). Elastic net regression requires the estimation of two penalty parameters (λ_1, λ_2), such that ridge regression and lasso are special cases of elastic net regression (when $\lambda_2 = 0$ or $\lambda_1 = 0$, respectively).	Major depressive disorder [64] Smoking [65] Alzheimer's disease ^a [66] Incident diabetes [46] Alcohol consumption, body fat percent, body mass index, lipoprotein cholesterol, waist-to-hip ratio [15, 67] 109 proteins [20] Electronic health records [68]
Linear mixed model BLUP ^b : OSCA [52]	All probes have a predicted effect size with effect sizes assumed to be drawn from a normal distribution with the total variance attributed to DNAm estimated from a restricted maximum likelihood (REML) analysis of the data.	ALS [69]
Linear mixed model ^b : OSCA [52] lme4 [70]	Effect size of each probe estimated while fitting the joint effect of probes genome-wide in one (or several) random effects to control for unidentified background confounders.	ALS [69] Parkinson's disease [71] Alzheimer's and Parkinson's disease, ALS, schizophrenia, rheumatoid arthritis [45]
Bayesian inference model ^b : BayesRR [72]	A linear mixed model, but with Bayesian framework to model any epigenetic genetic architecture. Probe effects are assumed drawn from one of multiple normal distributions (including null). Genetic effects can be modeled simultaneously.	BMI, smoking [72] Cognitive ability [73]

C+PT- clumping + P -value thresholding of MWAS summary statistics, BLUP-Best linear unbiased prediction

^a MPS derived from postmortem brain cortex

^b These methods account for correlations in DNAm between people resulting from family relationships

prediction results, and methods that minimize the number of features selected are likely to be the most useful in biomarker tests.

For PGS, individual-level GWAS data are often not available to researchers due to logistical and privacy concerns. Thus, the choice of SNPs and their weights are usually derived from meta-analyses of GWAS summary statistics (i.e., SNP identification number, risk allele, risk allele frequency, risk allele effect size and its standard error, p -value of association). The correlation structure between SNPs is integrated through knowledge of linkage disequilibrium (LD) among genetic variants, derived from a reference panel with individual-level genotypic data. The simplest PGS method selects independent SNPs associated with a p -value less than a specified threshold to select the m_{PGS} SNPs and uses the GWAS association effect size estimates as the $\hat{\beta}_j$ (the so-called clumping and p -value thresholding method, denoted here C+PT). Other PGS methods, in essence, use the genome-wide set of GWAS summary statistics to learn the trait-specific genetic architecture that can lead to choices of m_{PGS} and updated values of the $\hat{\beta}_j$ that give higher out-of-sample prediction than C+PT (e.g., [57, 74]). Some PGS methods also use functional (or other) SNP annotations as prior information to increase the chances that SNPs selected are causal variants. This is particularly important if the goal is to increase the transferability of PGS across ancestry groups, as sets of SNPs that are highly correlated in one ancestry (hence which of the SNPs is selected has little impact on the efficacy of the predictor) may not be so highly correlated in other ancestries. Hence increasing the probability of selecting the causal SNP, which is likely to be causal in all ancestries [75] is important.

There are many-fold fewer MWAS data sets compared to GWAS data sets. However, MWAS data have traditionally been less hampered by data privacy issues [76], and so, a high proportion is shared in databases which allow direct download (from repositories such as Gene Expression Omnibus (GEO) [77] or ArrayExpress [78]) of individual-level data from the discovery MWAS. This means that cross-validation (splitting a tuning sample (Fig. 2) out of the discovery MWAS) can be applied to determine the optimum selection of probes into the MPS. Basic MPS approaches have adopted the C+PT method used in PGS. MPS methods are summarized in Table 1 and Additional file 1. Briefly, linear mixed model approaches such as OSCA MOA and MOMENT [52] estimate the effect of each probe in turn while fitting the joint effects of genome-wide DNAm and the correlation structure in DNAm between people, which is effective at accounting for unknown batch/confounder effects, and parallels mixed linear models used in GWAS (such as GCTA -mlma [79], fastGWA [80], BOLT-LMM [81]). Penalized regression methods utilize cross-validation to directly select probes and derive probe weights for MPS. These methods are little used for PGS generation, owing to the large number of genetic features that would need to be accommodated by the models. Penalized and Bayesian regression methods may mitigate against the winner's curse effect, a problem in one probe at a time analyses, by considering all methylation sites jointly and applying shrinkage factors. MethylDetectR provides scripts and/or an online tool for the calculation of MPS for many traits, housing weights generated by several MPS studies [67]. MethylPipeR is a tool that facilitates the automated application of penalized regression models for MPS generation, in addition to tree-based statistical learning methods [46]. The primary issue facing those using these tree-based neural net methods is that the number of features is too large to build networks and current research is investigating if

the list of probes considered can be reduced. BayesRR is a Bayesian approach that jointly models all probe and SNP effects that has also been shown to implicitly adjust for the presence of unknown confounders such as batch and cell-type effects [72] (see below).

With increasing numbers of MWAS publications [82] (Fig. 2) and data sets—accompanied by increasing privacy concerns [83]—studies using MPS derived from meta-analyzed MWAS summary statistics are starting to be published. The development of new MPS methods based on MWAS summary statistics is likely to be an area of active research. Derivation of GWAS summary statistics-based approximations of methods that use individual-level data is possible because the correlation (LD) structure between SNPs reflects population history (such as genetic drift, migration, mutation, population bottlenecks), and this can be assumed to be trait independent. Hence, the correlation structure between SNPs can be inferred from ancestry-matched LD reference samples. However, the repeatability of the correlation structure between DNAm probes across samples drawn from the same population is likely more complex [84] and likely to vary between cell types. Moreover, there is an additional correlation structure between probes at a considerable genomic distance (more than 300 kb in some cases) with intermediary blocks uncorrelated [84]. While some of the correlation was found to be genetically controlled, the authors hypothesized that the dispersed correlation could reflect a structural basis in nuclear organization. Moreover, the odd/even chromosome simulations of Zhang et al. [52] (described above) exposed extensive cross-chromosome correlation. The lack of reference data sets to phase DNAm by haplotype and the influence of environmental exposures on the correlation structure [84] mean that more data are needed to establish if a trait-independent correlation structure can be assumed for application with MWAS summary statistics. The LD correlation of SNPs means that some PGS methods (e.g., SBayesR [85], PRS-CS-auto [86]) have been optimized without the need for tuning samples (samples independent of both discovery and target samples used to obtain parameter estimates needed in the choice of the SNPs and their weights). However, future optimization of MPS methods will likely need to use tuning samples which must be ascertained to have properties similar to the target and application samples.

Applications of MPS for trait prediction

We identify four key applications for use of MPS for the prediction of an unmeasured/unknown phenotype (Fig. 3). First, in the context of research where some key phenotypes have not been, or were inaccurately, recorded. For example, smoking is a key risk factor relevant to epidemiological analyses. When smoking status is not recorded, smoking can be predicted accurately from DNAm data, AUC statistic = 0.98 (where AUC can be interpreted as the probability that a smoker ranks higher than a non-smoker on the MPS) [15, 17, 72, 87, 88]. Moreover, the MPS smoking measure may be a more accurate quantification of smoking exposure (both direct and passive) than self-report data. Prediction of unrecorded phenotypes is important in association studies, where fitting confounder variables can help reduce the false-positive rate [52].

A second application of MPS is the potential use of DNAm as a direct outcome measure in the clinic [83], for example, to measure the effectiveness of smoking

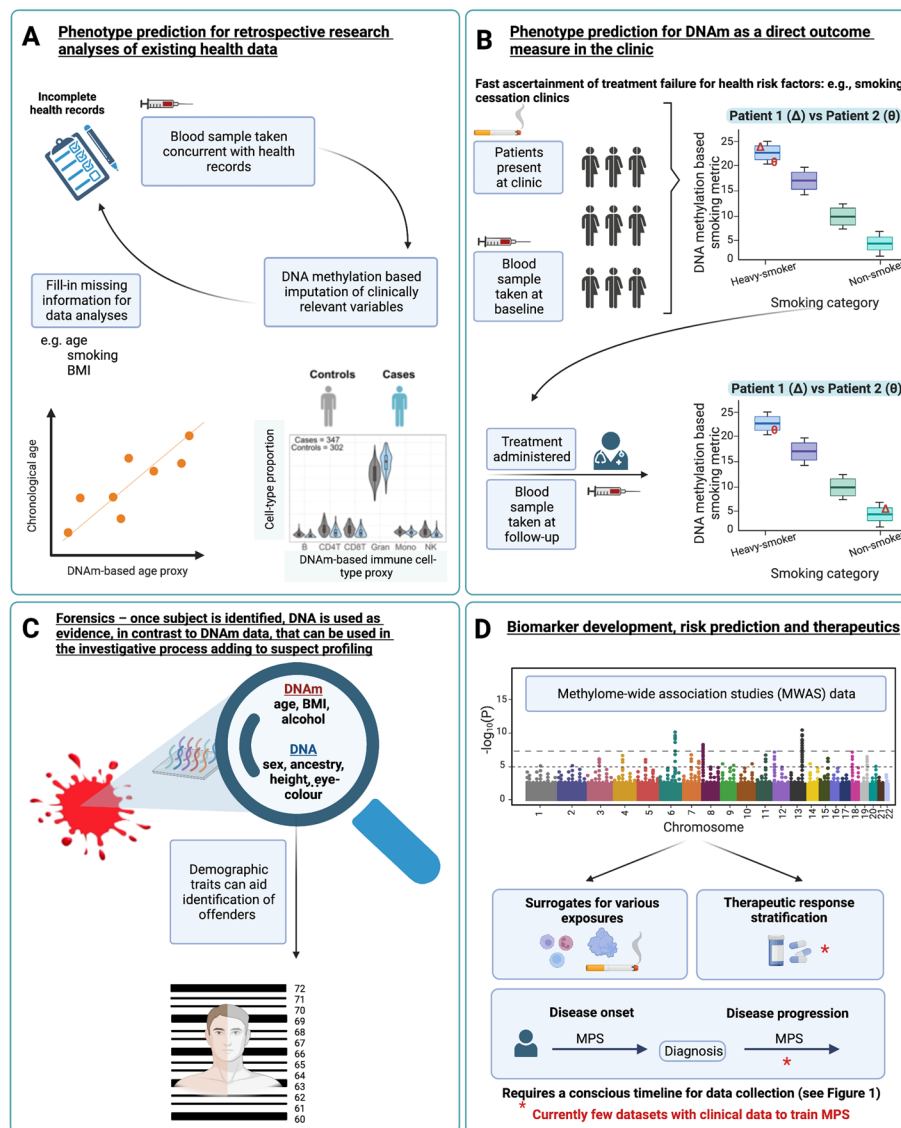


Fig. 3 Applications of DNA methylation-based trait prediction. Prediction of non-recorded phenotypes **A** in research data; **B** in objective quantification of participant compliance through longitudinal prediction of traits; **C** in forensics, where trait prediction could contribute to investigative rather than formal evidence-based procedures; and **D** as biomarkers to aid disease diagnosis and in future (if suitable discovery samples become available) for choice of therapeutics. Created with BioRender.com

cessation therapy both qualitatively (e.g., current vs never-smokers) [17, 65] and quantitatively (e.g., assessing the reversibility of smoking-induced methylation changes) [89]. Body weight loss has also been associated with differences in DNAm [90]. DNAm data could then be used to show patients objective quantified results, which may improve the chances of successful behavior change due to positive reinforcement. MPS may also facilitate clinical ascertainment of treatment failure, if DNAm data are measured pre- and post-therapeutic intervention. Though these multi-time point, longitudinal data sets are rarely available, recent studies have demonstrated differential DNAm signatures associated with therapeutic intervention 4–12 weeks after initiation [91, 92].

A third application for phenotype prediction could be in forensics: when a biological sample is available, but the person associated with the sample is unidentified. In contrast to DNA profiling, which is used as evidence, MPS can be used in the investigative process adding to suspect profiling (since once a person is identified, DNA profiling is sufficient). In criminal investigations, demographic traits can be crucial to help identify offenders. Whereas highly heritable traits such as height can be predicted from genetic data, less heritable traits such as weight, or body mass index (BMI) could be better predicted by MPS or a combination of MPS and PGS [58, 93]. A promising example is a prediction of age from DNAm data, where prediction is highly accurate even with current DNAm array platforms [93, 94]. Indeed, variance in age was found to be fully explained by MPS in the Generation Scotland cohort (i.e., $\rho^2 = 1$, $SE = 0.0036$) [93].

Finally, DNAm data could prove useful in clinical settings as biomarkers of disease risk (Fig. 1). DNAm differences associated with incident diseases such as type 2 diabetes are detectable many years prior to formal diagnoses [46]. These signatures may represent the consequences of risk exposures; DNAm at smoking-associated genes have been linked to lung cancer development [95]. In such instances, DNAm may lie on causal pathways to disease. Equally, DNAm signatures may represent a record of exposures to factors such as smoking, without being directly causal for a disease. MPS generally do not need to delineate between the reasons why DNAm differences are associated with the outcome, as their purpose is risk stratification. However, future research questions are likely to investigate if MPS predicting relevant exposure traits could be part of an overall risk algorithm. MPS derived from blood samples taken at first diagnosis could be useful in predicting disease progression, but this is only achievable with the generation of discovery data sets with longitudinal clinical information. More translational research is needed to evaluate the utility of MPS in health settings.

Cell type proportions in MPS

When DNAm is measured in bulk tissue such as whole blood, trait associations could reflect a mixture of intrinsic and extrinsic signals (Fig. 4). The “intrinsic” signal of DNAm represents a change in DNAm that is directly associated with the trait (which could affect one or more cell types) [96, 97]. The “extrinsic” [98] signal can represent the differences in cell-type proportions, given that some DNAm is cell type-specific. In MWAS analyses, cell-type proportion differences are generally considered to be confounders whose contribution to trait variation should be adjusted. A close interplay between circulating immune cells and DNAm exists [99]. As such, adjustments for DNAm-derived immune cell proportions are critical in blood-based MWAS analyses [6]. These adjustments are imperfect, as they do not include every immune cell and rare subpopulations therefore likely exist that are unaccounted for. For example, two MWAS of blood protein levels reported associations between pappasylin (PAPPA) and various DNAm sites; however, after adjustment for eosinophil proportions, these signals were attenuated [100, 101]. An expanded cell-type deconvolution panel for 56 immune cell profiles has recently become available and may aid in the separation of extrinsic and intrinsic signals [102].

Careful consideration of study aims is needed when deciding the relevance of trait-associated cell-type proportions. If the goal is a biological interpretation of DNAm differences, then correction for cell-type proportions is appropriate. However, when

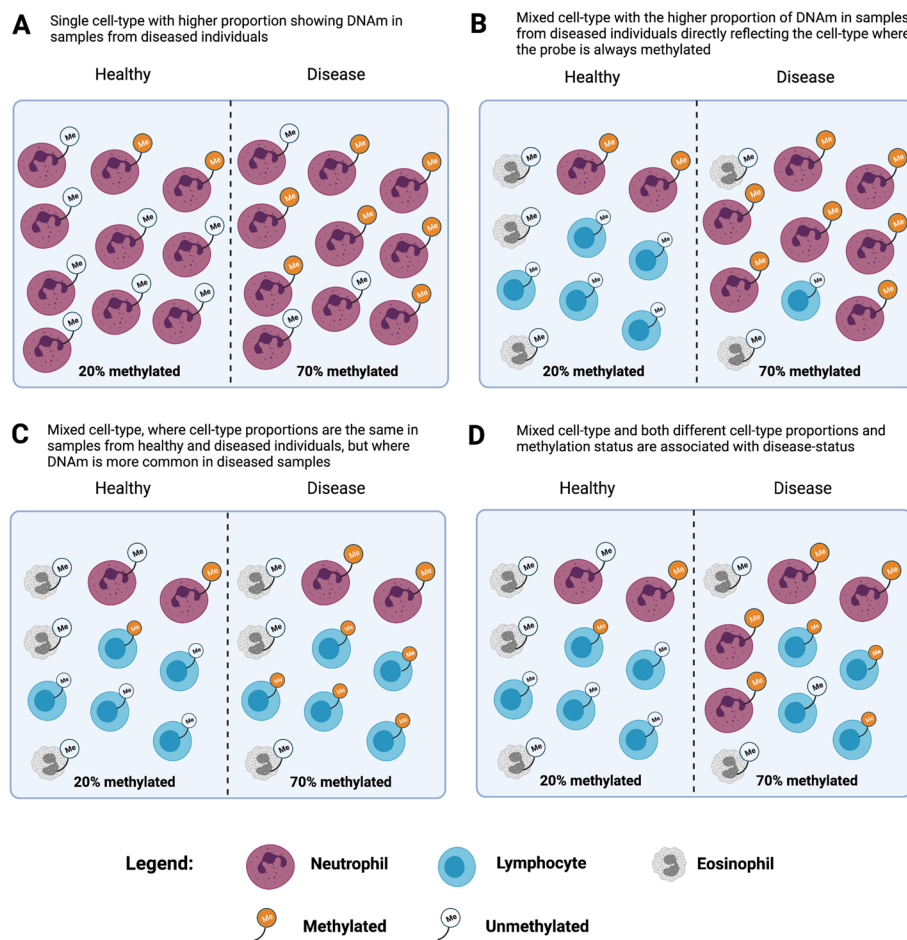


Fig. 4 Cellular heterogeneity and its effects on DNAm measurements in bulk tissue. All scenarios show the same difference in DNAm between healthy and diseased samples. The DNAm differences between healthy and disease samples can reflect increased DNAm associated with disease (**A**, **C**, and lymphocytes in **D**) or cell-type proportion differences associated with disease status (**B**, **D**). Recreated with BioRender.com and inspired by Holbrook et al. [96]

DNAm in the blood is simply considered as a biomarker, and when the goal is trait prediction, any signal that can be derived from the DNAm that maximizes out-of-sample prediction accuracy should be included. For example, in our MWAS of amyotrophic lateral sclerosis [69], we found that predicted cell-type proportion differences between cases and controls replicated between cohorts (higher proportion of neutrophil granulocytes, N case/control of 612/782 and 1159/637 in discovery and target cohorts, respectively). Moreover, we found that MPS based on weights applied to DNAm-derived cell-type proportions generated an AUC of 0.67, higher than our MPS approach and close to the maximum AUC achieved from combining the predicted cell-type proportions with the MPS (AUC = 0.69). However, a higher proportion of neutrophil granulocytes estimated from DNAm in cases compared to controls was found for many diseases (Alzheimer’s disease, Parkinson’s disease, schizophrenia, and rheumatoid arthritis), demonstrating non-specificity [45]. We tested a published MPS for major depressive disorder (derived from non-smokers in both cases and controls) [64] on these same diseases

and also found non-specificity, with higher AUC for Parkinson's disease (AUC = 0.58) and schizophrenia (SCZ1) in (AUC = 0.57) cohorts than had been reported for depression (AUC = 0.53) [64]. Although we hypothesize this non-specificity may be driven by different cell-type proportions shared between major depressive disorder and the other traits, there may be other factors contributing to the observed result. Statistical MWAS methods (such as TCA [103] and CellDMC [104]) have been developed to identify trait associations specific to cell types. These methods could be used to improve trait prediction, but further development and critical evaluation are needed.

Evaluation metrics for MPS

First, we consider the evaluation metrics for PGS when calculated in target samples (those where the MPS trait has been measured) and then draw parallels for MPS. For quantitative traits (which are typically normally distributed or can be transformed to be so), the accuracy of prediction of a PGS is evaluated as R^2 —the proportion of phenotypic variance (σ_p^2) in the trait explained by the PGS. The R^2 is on the same scale as and can be contrasted to heritability (h^2) which is the proportion of σ_p^2 that is explained by genetic variation (σ_G^2), i.e., σ_G^2/σ_p^2 (and where the phenotypic variance is the variance of the trait y after removing variation attributed to fixed effects, such as sex and age). Whereas h^2 reflects all factors contributing to genetic variation, the PGS can only capture the genetic variation associated with the SNPs measured in the GWAS, so the upper bound on the R^2 from PGS (under assumptions of the same trait drawn from the same population) is the SNP-based heritability (h_{M}^2 , where the M refers to the number of independent SNPs represented by the genome-wide SNP array). Given that effect sizes of individual SNPs are estimated with error the R^2 that is expected from a PGS can be approximated as $R_M^2 \approx \frac{h_M^2}{1+M/(Nh_M^2)}$ [105], where N is the sample size (of the discovery sample). In theory, as sample sizes increase, the variance explained by PGS will also increase and tend to the SNP-based heritability [105, 106].

For binary traits, a logistic regression pseudo- R^2 statistic, Nagelkerke's R^2 (R_{NK}^2), is often reported to evaluate PGS. While this can be useful for comparing the efficacy of different PGS methods applied to the same target data set, the values cannot be fairly compared across target data sets as the statistic depends on the proportion of cases in the sample. Instead, an R_{CC}^2 estimate can be made under a linear regression model of the binary (CC: case-control) data, which is then converted to the liability scale (R_l^2) accounting for the proportion of cases in the sample (P) and the lifetime risk of disease (K), $R_l^2 = R_{CC}^2 \frac{[K(1-K)]^2}{z^2 P(1-P)}$, where z is the height of the normal curve when thresholded by proportion K [107]. Another evaluation metric for binary disease traits is the AUC which has the advantage of not being dependent on the proportion of cases in the sample, but its scale is less intuitive to understand (it is related to the square root of R^2 [108] so increases in AUC associated with increased discovery sample size are smaller for AUC than for R^2 statistics [57]).

In the context of MPS, R^2 for quantitative traits, and R_{NK}^2 , R_{CC}^2 and AUC for binary traits are typically used to assess the accuracy of trait prediction [58, 59, 69, 71]. The proportion of phenotypic variance explained by all DNAm markers analyzed in an MWAS (ρ^2) is now sometimes reported [45, 52, 69] and is an upper bound on the variance explained by MPS in an independent sample (with the same proportion of cases,

and hence the same phenotypic variance). Although whether such a maximum can be achieved even with increasing sample sizes depends on what extent discovery sample confounding factors contribute to the ρ^2 estimate. It is important to recognize that estimates of R_{CC}^2 cannot be converted to R_I^2 , because the underlying genetic theory [107, 109] that generates the conversion equation does not apply. Therefore, for disease traits the AUC and AUC-related statistics may be the best metric for the comparison of MPS applied across different target samples.

Combining MPS with PGS

An early study [58] evaluated both MPS and PGS for BMI and height in Lothian Birth Cohort participants ($n = 1,366$). The PGS and MPS explained 8% and 7%, respectively, of the variance in BMI and 14% when fitted jointly demonstrating that PGS and MPS contributions were mostly independent and additive. Most notably, the discovery sample was $\sim 350K$ for the PGS but only 750 for MPS. The same study reported that height MPS explained no variation in height. The difference in the success of MPS for BMI and height likely reflects that the current diet impacts blood DNAm which is associated with concurrent BMI. In contrast, variation in height between older people is likely not captured in their blood (but may be in childhood [110]). It is likely that much more variation in BMI could be explained with larger discovery samples. Results from studies with larger sample sizes have reported the combination of PGS+MPS to be less than additive [15, 73], likely reflecting that some SNPs associated with traits could be mQTL (i.e., both are capturing the same underlying genetic risk).

The BayesRR [72] method uses a linear mixed model in a Bayesian framework to model any epigenetic genetic architecture and genetic architecture simultaneously (by assuming probe effects and SNP effects are drawn from one of the multiple normal distributions with different variances and different mixing proportions estimated from the data) in discovery samples that have both GWAS and MWAS data. BayesRR was applied to BMI and smoking behavior (pack-years) measured in 9,488 individuals in Generation Scotland to give estimates of genetic and epigenetic architecture from the same traits [72]. Their statistics were all provided with 95% confidence intervals, here, for simplicity, we report the rounded point estimates (N.B. these statistics are ρ^2 (see above) not out-of-sample R^2). For smoking behavior defined as the number of pack-years, they reported that 46% of the phenotypic variance was captured by methylation probes and 6% by genome-wide SNPs (i.e., SNP-based heritability, h_{SNP}^2). For BMI, the estimates were $\rho^2 = 76\%$ and $h_{SNP}^2 = 16\%$. For BMI, the number of contributing probes was higher, and the effect size attributed to each was smaller. For example, the 17 probes of the largest effect explained about 10% of the variance of BMI, whereas 15 probes were estimated to explain 27% of the phenotypic variance of smoking behavior. Using the BayesRR derived DNAm score, out-of-sample trait prediction was associated with 18% of the variance in BMI and 38% of the variance in smoking (ARIES cohort adult males). The variance explained in BMI in ARIES cohort children at birth, 7 years, and 15 years were 3%, 2%, and 10%, respectively (important given the caveats of Fig. 1). The out-of-sample results are impressive given the discovery sample. The BayesRR authors calculated that with a discovery sample of 100,000, out-of-sample R^2 from DNAm alone could be 60%, which could increase to 80% when MPS are added to PGS. Although BayesRR presented an

integrated model for DNAm and SNP array data, in reality, larger discovery samples will be available (and needed) for GWAS compared to MWAS and so PGS and MPS will be generated independently. Tuning samples will be needed to determine how to weigh PGS and MPS when combined into a single predictor.

Ancestry

It is well-recognized that most GWAS discovery samples are from participants of European ancestry [111]. The cost-effective paradigm of GWAS utilizes the LD correlation structure enabling a very high proportion of genomic variation associated with the 3 billion base pairs ($\times 2$ chromosomes) of a genome to be captured by $\sim 500\text{K}$ SNPs. Inevitably, associations point to genomic regions rather than individual causal SNPs (at least until GWAS samples are very large and incorporate sufficient recombination events to pinpoint causal associations). PGS out-of-sample prediction is robust to whether the SNPs included are the causal variants or variants very highly correlated with them. However, the different population histories of people of different ancestries mean that the correlation structure between SNPs differs between ancestries, and so causal SNPs (or those held in tight LD with them across ancestries) need to be prioritized in the PGS calculations. One driving component of LD differences between ancestries is allele frequency differences. Even when trait-associated alleles have large frequency differences between ancestries, the effect size estimates can be similar (see Figs. 1 and 2 in Liu et al. [112] for nice visualizations in application to inflammatory bowel disease). If the effect size estimates are different across ancestries, interaction with genetic or non-genetic risk factors is implied. Consistency of trait definition is an important consideration, both between and within ancestry samples, and so applications of cross-ancestry PGS should always be benchmarked against within-ancestry results. Increasing GWAS data sets from different ancestry groups is currently a key priority for many funding agencies.

There are few data sets that compare DNAm across ancestries. DNAm age predictors were originally derived from multi-ancestry samples [113, 114], and this may be why these MPS are accurate across ancestries [93, 98]. However, ancestry-specific differences in epigenetic aging (difference between chronological and DNAm predicated age) have been reported [98]. High replication of effect sizes has been reported between Chinese and European for blood mQTL [115] and between Europeans and African-Americans for probes associated with C-reactive protein [116]. More DNAm data sets from diverse ancestries are needed to be able to draw informed conclusions about the transferability of MPS across ancestries and because environmental and cultural differences may directly impact DNAm. Moreover, the lack of diversity in GWAS is raising concerns about exacerbating health inequality now that the clinical translation of PGS is being evaluated [111]. Prospective studies are now starting to show that MPS biomarkers could have clinical utility [46], so more efforts are needed to diversify MWAS data sets [117].

Conclusions

MPS are increasingly being developed to stratify risk associated with diseases and traits associated with diseases. The methodology used to calculate these scores parallels that used to define PGS for common complex diseases, which are derived from common single nucleotide variants. PGS can be considered trait or disease predictors,

since in principle they can be calculated at any time over the lifespan. In contrast, DNAm can be dynamic across the lifespan and therefore must be developed using data relevant to a specific application context, or at least evaluated in the application context before the real-life utility can be confirmed. MPS likely capture signatures that are a consequence of the trait, in addition to potential trait-associated causal pathways and MPS development must recognize this nuance. For MPS where DNAm is largely a direct outcome of the trait, e.g., smoking and BMI, MPS capture a large proportion of variation in out-of-sample evaluation, much more than PGS, despite much smaller sample sizes. In the context of disease prediction, less variance is often explained by MPS, but there is sufficient evidence to date to support further evaluation of DNAm as a biomarker of disease onset or disease progression. This requires careful ascertainment of cases and controls and efforts in all stages of experimental design to minimize batch effects and to understand contributions arising from cell-type effects. MPS can also be trained to predict disease-associated traits, which (at least in preliminary evaluation) could be more cost-effective or simply more feasible than direct measurement of the trait. For example, MPS for 109 plasma proteins trained in independent cohorts ($N \geq 725$) and projected into the Generation Scotland cohort ($N = 9,537$) were associated with the incidence of 11 major age-related morbidities, with 137 MPS disease associations reported for 11 common diseases over 14 years of electronic health linkage [20]. Investment in the generation of DNAm data in large prospective cohorts with linkage to health records such as the UK Biobank (or, added in proof, UCLA Health Biobank [68]) would allow further characterization of MPS as early biomarkers of incident diseases. This may be cost-effective relative to grant funding that is spent on biomarker research by international funding agencies in other settings. However, as data sets available for measurement of DNAm increase in size, samples will likely be processed over extended periods of time generating technical variability. This can impact the detection of small DNAm effect sizes which is a growing concern calling for the development of improved technology for the array-based measurement of DNAm. Nonetheless, the generation of DNAm data in cohorts such as the UK Biobank that are already genetically informative would drive the development of new technologies as well as the development of new methods focussed on joint MPS and PGS modeling.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-02855-7>.

Additional file 1.

Additional file 2. Review history.

Peer review information

Anahita Bishop and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

Planning: MFN, EH, JM, AFMcR, and NRW. First draft: MFN and NRW. Submission draft: MFN, EH, JM, AFMcR, and NRW. Revisions: DAG and NRW. Figures: MFN and DAG. Final version: The authors read and approved the final manuscript.

Authors' Twitter handles

Twitter handles: @RandomWalkin (Marta Nabais); @dannigadd (Danni Gadd); @PsyEpigenetics (Jon Mill); @WrayNaomi (Naomi Wray).

Funding

This work is supported by a University of Queensland/University of Exeter (QUEX) joint initiative that provided a PhD scholarship to MFN. It is also supported by National Health and Medical Research Council grants (1113400,1173790, 1151854 (EU-JPND BRAIN-MND)) to NRW. DAG is supported by the Wellcome Trust 4-year PhD in Translational Neuroscience (108890/Z/15/Z).

Declarations**Competing interests**

DAG has received consultancy fees from Optima partners.

Received: 13 April 2022 Accepted: 17 January 2023

Published online: 16 February 2023

References

- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012;13:484–92.
- Seiler Vellame D, Castanho I, Dahir A, Mill J, Hannon E. Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation. *BMC Genomics.* 2021;22:446.
- Bernabeu E, McCartney DL, Gadd DA, Hillary RF, Lu AT, Murphy L, et al. Campbell A, Harris SE, Liewald D, et al. Refining epigenetic prediction of chronological and biological age. *bioRxiv* 2022. <https://doi.org/10.1101/2022.09.08.507115>.
- Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, Carnero-Montoro E, Lawson DJ, Burrows K, Suderman M, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet.* 2021;53:1311–21.
- Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, et al. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet.* 2016;9:436–47.
- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13:86.
- Sehoulil J, Loddenkemper C, Cornu T, Schwachula T, Hoffmuller U, Grutzkau A, Lohneis P, Dickhaus T, Grone J, Kruschewski M, et al. Epigenetic quantification of tumor-infiltrating T-lymphocytes. *Epigenetics.* 2011;6:236–46.
- Caggiano C, Celona B, Garton F, Mefford J, Black BL, Henderson R, Lomen-Hoerth C, Dahl A, Zaitlen N. Comprehensive cell type decomposition of circulating cell-free DNA with CellFIE. *Nat Commun.* 2021;12:2717.
- McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, Gibson J, Henders AK, Bowdler L, Painter JN, et al. Identification of 55,000 replicated DNA methylation QTL. *Sci Rep.* 2018;8:17605.
- Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* 2010;6:e1000952.
- Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, Kumari M, Schalkwyk LC, Mill J. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am J Hum Genet.* 2018;103:654–65.
- Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, Troakes C, Turecki G, O'Donovan MC, Schalkwyk LC, et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci.* 2016;19:48–54.
- Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, Pattie A, Corley J, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* 2014;24:1725–33.
- Luo C, Hajkova P, Ecker JR. Dynamic DNA methylation: in the right place at the right time. *Science.* 2018;361:1336–40.
- McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, Morris SW, Birmingham ML, Campbell A, Murray AD, et al. Epigenetic prediction of complex traits and death. *Genome Biol.* 2018;19:136.
- Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, Weidinger S, Lattka E, Adamski J, Peters A, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One.* 2013;8:e63812.
- Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, Davey Smith G, Hughes AD, Chaturvedi N, Relton CL. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics.* 2014;6:4.
- Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, Davies G, Ritchie CW, Gale CR, Starr JM, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry.* 2018;8:99.
- Locke WJ, Guanzone D, Ma C, Liew YJ, Duesing KR, Fung KYC, Ross JP. DNA methylation cancer biomarkers: translation to the clinic. *Front Genet.* 2019;10:1150.
- Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Cheng Y, Fawns-Ritchie C, Nangle C, Campbell A, Flaig R, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *Elife.* 2022;11:e71802.
- Greally JM. A user's guide to the ambiguous word 'epigenetics'. *Nat Rev Mol Cell Biol.* 2018;19:207–8.

22. Lappalainen T, Grealis JM. Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet.* 2017;18:441–51.
23. Birney E, Smith GD, Grealis JM. Epigenome-wide association studies and the interpretation of disease -omics. *PLoS Genet.* 2016;12:e1006105.
24. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet.* 2013;14:585–94.
25. Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL. DNA methylation-based predictors of health: applications and statistical considerations. *Nat Rev Genet.* 2022;23(6):369–83.
26. Agha G, Mendelson MM, Ward-Caviness CK, Joehanes R, Huan T, Gondalia R, Salfati E, Brody JA, Fiorito G, Bressler J, et al. Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease. *Circulation.* 2019;140:645–57.
27. Hannon E, Schendel D, Ladd-Acosta C, Grove J, iP-BASDG, Hansen CS, Andrews SV, Hougaard DM, Bresnahan M, Mors O, et al. Elevated polygenic burden for autism is associated with differential DNA methylation at birth. *Genome Med.* 2018;10:19.
28. Walker RM, MacGillivray L, McCafferty S, Wrobel N, Murphy L, Kerr SM, Morris SW, Campbell A, McIntosh AM, Porteous DJ, Evans KL. Assessment of dried blood spots for DNA methylation profiling. *Wellcome Open Res.* 2019;4:44.
29. Merid SK, Novoloaca A, Sharp GC, Kupers LK, Kho AT, Roy R, Gao L, Annesi-Maesano I, Jain P, Plusquin M, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. *Genome Med.* 2020;12:25.
30. Wheeler ENW, Galdi P, McCartney DL, Blesa M, Sullivan G, Stoye DQ, Lamb G, Sparrow S, Murphy L, Wrobel N, et al. DNA methylation in relation to gestational age and brain dysmaturation in preterm infants. *Brain Commun.* 2022;4:fcac056.
31. Theda C, Hwang SH, Czajko A, Loke YJ, Leong P, Craig JM. Quantitation of the cellular content of saliva and buccal swab samples. *Sci Rep.* 2018;8:6944.
32. van Dongen J, Ehli EA, Jansen R, van Beijsterveldt CEM, Willemsen G, Hottenga JJ, Kallsen NA, Peyton SA, Breeze CE, Kluff C, et al. Genome-wide analysis of DNA methylation in buccal cells: a study of monozygotic twins and mQTLs. *Epigenetics Chromatin.* 2018;11:54.
33. Eipel M, Mayer F, Arent T, Ferreira MR, Birkhofer C, Gerstenmaier U, Costa IG, Ritz-Timme S, Wagner W. Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. *Aging (Albany NY).* 2016;8:1034–48.
34. Hannon E, Mansell G, Walker E, Nabais MF, Burrage J, Kepa A, Best-Lane J, Rose A, Heck S, Moffitt TE, et al. Assessing the co-variability of DNA methylation across peripheral cells and tissues: implications for the interpretation of findings in epigenetic epidemiology. *PLoS Genet.* 2021;17:e1009443.
35. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, et al. High density DNA methylation array with single CpG site resolution. *Genomics.* 2011;98:288–95.
36. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Dijk S, Muhlhäuser B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* 2016;17:208.
37. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics.* 2010;11:587.
38. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* 2014;30:1363–9.
39. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, Teschendorff AE. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics.* 2017;33:3982–4.
40. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, Hannon E. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics.* 2019;20:366.
41. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics.* 2018;34(23):3983–9.
42. Buhule OD, Minster RL, Hawley NL, Medvedovic M, Sun G, Viali S, Deka R, McGarvey ST, Weeks DE. Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Front Genet.* 2014;5:354.
43. Sun Z, Chai HS, Wu Y, White WM, Donkena KV, Klein CJ, Garovic VD, Therneau TM, Kocher JP. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med Genomics.* 2011;4:84.
44. Ori APS, Lu AT, Horvath S, Ophoff RA. Significant variation in the performance of DNA methylation predictors across data preprocessing and normalization strategies. *Genome Biol.* 2022;23:225.
45. Nabais MF, Laws SM, Lin T, Vallerga CL, Armstrong NJ, Blair IP, Kwok JB, Mather KA, Mellick GD, Sachdev PS, et al. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. *Genome Biol.* 2021;22:90.
46. Cheng Y, Gadd DA, Gieger C, Monterrubio-Gómez K, Zhang Y, Berta I, Stam MJ, Szlachetka N, Lobzaev E, Wrobel N, et al. Development and validation of DNA Methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. *medRxiv* 2022. <https://doi.org/10.1101/2021.11.19.21266469>.
47. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, Jokubaitis VG, Lea RA. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin Epigenetics.* 2021;13:214.
48. Rakyán VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet.* 2011;12:529–41.
49. Paul DS, Beck S. Advances in epigenome-wide association studies for common diseases. *Trends Mol Med.* 2014;20:541–3.
50. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, St Clair D, Mustard C, Breen G, Therman S, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* 2016;17:176.

51. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet.* 2004;36:388–93.
52. Zhang F, Chen W, Zhu Z, Zhang Q, Nabais MF, Qi T, Deary IJ, Wray NR, Visscher PM, McRae AF, Yang J. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biology.* 2019;20:107.
53. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, Oh S, Burchard EG, Eskin E, Zou J, Halperin E. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods.* 2016;13:443–5.
54. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods.* 2014;11:309–11.
55. Hop PJ, Zwamborn RAJ, Hannon EJ, Dekker AM, van Eijk KR, Walker EM, Iacoangeli A, Jones AR, Shatunov A, Khleifat AA, et al. Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. *NAR Genom Bioinform.* 2020;2:lqaa105.
56. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009;460:748–52.
57. Ni G, Zeng J, Revez JA, Wang Y, Zheng Z, Ge T, Restuadi R, Kiewa J, Nyholt DR, Coleman JRI, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol Psychiatry.* 2021;90:611–20.
58. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, Harris SE, Liewald D, Henders AK, Mendelson MM, et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *Am J Hum Genet.* 2015;97:75–85.
59. Watkeys OJ, Cohen-Woods S, Quide Y, Cairns MJ, Overs B, Fullerton JM, Green MJ. Derivation of poly-methylomic profile scores for schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry.* 2020;101:109925.
60. Barker ED, Cecil CAM, Walton E, Houtepen LC, O'Connor TG, Danese A, Jaffee SR, Jensen SKG, Pariante C, McArdle W, et al. Inflammation-related epigenetic risk and child and adolescent mental health: a prospective study from pregnancy to middle adolescence. *Dev Psychopathol.* 2018;30:1145–56.
61. Stevenson AJ, Gadd DA, Hillary RF, McCartney DL, Campbell A, Walker RM, Evans KL, Harris SE, Spires-Jones TL, McRae AF, et al. Creating and validating a DNA methylation-based proxy for interleukin-6. *J Gerontol A Biol Sci Med Sci.* 2021;76:2284–92.
62. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33:1–22.
63. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw.* 2011;39:1–13.
64. Barbu MC, Shen X, Walker RM, Howard DM, Evans KL, Whalley HC, Porteous DJ, Morris SW, Deary IJ, Zeng Y, et al. Epigenetic prediction of major depressive disorder. *Mol Psychiatry.* 2021;26:5112–23.
65. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics.* 2019;11:1469–86.
66. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, Wheildon G, Mastroeni D, Gasparoni G, Riemschneider M, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. *Nat Commun.* 2021;12:3517.
67. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Res.* 2020;5:283.
68. Thompson M, Hill BL, Rakocz N, Chiang JN, Geschwind D, Sankararaman S, Hofer I, Cannesson M, Zaitlen N, Halperin E. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *NPJ Genom Med.* 2022;7:50.
69. Nabais MF, Lin T, Benyamin B, Williams KL, Garton FC, Vinkhuyzen AAE, Zhang F, Vallerga CL, Restuadi R, Freydenzon A, et al. Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis. *npj Genom Med.* 2020;5:10.
70. Bates D, Machler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67:1–48.
71. Vallerga CL, Zhang F, Fowdar J, McRae AF, Qi T, Nabais MF, Zhang Q, Kassam I, Henders AK, Wallace L, et al. Analysis of DNA methylation associates the cystine–glutamate antiporter SLC7A11 with risk of Parkinson's disease. *Nat Commun.* 2020;11:1238.
72. Trejo Banos D, McCartney DL, Patxot M, Ancheri L, Battram T, Christiansen C, Costeira R, Walker RM, Morris SW, Campbell A, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat Commun.* 2020;11:2865.
73. McCartney DL, Hillary RF, Conole ELS, Banos DT, Gadd DA, Walker RM, Nangle C, Flaig R, Campbell A, Murray AD, et al. Blood-based epigenome-wide analyses of cognitive abilities. *Genome Biol.* 2022;23:26.
74. Zhang Q, Privé F, Vilhjálmsón B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Comm.* 2021;12:4192.
75. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun.* 2020;11:3865.
76. Joly Y, Dyke SO, Cheung WA, Rothstein MA, Pastinen T. Risk of re-identification of epigenetic methylation data: a more nuanced response is needed. *Clin Epigenetics.* 2015;7:45.
77. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013;41:D991–995.
78. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kummeren P, Lara GG, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2003;31:68–71.
79. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88:76–82.

80. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, Yang J. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet.* 2019;51:1749–55.
81. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet.* 2018;50:906–8.
82. Xiong Z, Li M, Yang F, Ma Y, Sang J, Li R, Li Z, Zhang Z, Bao Y. EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res.* 2020;48:D890–5.
83. Phillibert RA, Beach SRH, Brody GH. The DNA methylation signature of smoking: an archetype for the identification of biomarkers for behavioral illness. In: Stoltenberg SF, editor. *Genes and the Motivation to Use Substances*. New York: Springer, New York; 2014. p. 109–27.
84. Liu Y, Li X, Aryee MJ, Ekstrom TJ, Padyukov L, Klareskog L, Vandiver A, Moore AZ, Tanaka T, Ferrucci L, et al. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet.* 2014;94:485–95.
85. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, Wang H, Zheng Z, Magi R, Esko T, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun.* 2019;10:5086.
86. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun.* 2019;10:1776.
87. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, Flanagan JM, Vineis P. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology.* 2013;24:712–6.
88. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ Res.* 2016;146:395–403.
89. Dugue PA, Jung CH, Joo JE, Wang X, Wong EM, Makalic E, Schmidt DF, Baglietto L, Severi G, Southey MC, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. *Epigenetics.* 2020;15:358–68.
90. Keller M, Yaskolka Meir A, Bernhart SH, Gepner Y, Shelef I, Schwarzfuchs D, Tsaban G, Zelicha H, Hopp L, Muller L, et al. DNA methylation signature in blood mirrors successful weight-loss during lifestyle interventions: the CENTRAL trial. *Genome Med.* 2020;12:97.
91. Nair N, Plant D, Verstappen SM, Isaacs JD, Morgan AW, Hyrich KL, Barton A, Wilson AG, investigators M: Differential DNA methylation correlates with response to methotrexate in rheumatoid arthritis. *Rheumatology (Oxford).* 2020;59:1364–71.
92. Julia A, Gomez A, Lopez-Lasanta M, Blanco F, Erra A, Fernandez-Nebro A, Mas AJ, Perez-Garcia C, Vivar MLG, Sanchez-Fernandez S, et al. Longitudinal analysis of blood DNA methylation identifies mechanisms of response to tumor necrosis factor inhibitor therapy in rheumatoid arthritis. *EBioMedicine.* 2022;80:104053.
93. Zhang Q, Vallergera CL, Walker RM, Lin T, Henders AK, Montgomery GW, He J, Fan D, Fowdar J, Kennedy M, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* 2019;11:54.
94. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, Christensen BC, Gladyshev VN, Heijmans BT, Horvath S, et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* 2019;20:249.
95. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun.* 2015;6:10192.
96. Holbrook JD, Huang RC, Barton SJ, Saffery R, Lillycrop KA. Is cellular heterogeneity merely a confounder to be removed from epigenome-wide association studies? *Epigenomics.* 2017;9:1143–50.
97. Shireby G, Dempster E, Policicchio S, Smith RG, Pishva E, Chioza B, Davies JP, Burrage J, Lunnon K, Seiler-Vellame D, et al. DNA methylation signatures of Alzheimer's disease neuropathology in the cortex are primarily driven by variation in non-neuronal cell-types. *Nat Comm.* 2022;13:5620.
98. Horvath S, Gurvey M, Levine ME, Trumble BC, Kaplan H, Allayee H, Ritz BR, Chen B, Lu AT, Rickabaugh TM, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* 2016;17:171.
99. Bergstedt J, Azzou SAK, Tsuo K, Jaquaniello A, Urrutia A, Rotival M, Lin DTS, MacIsaac JL, Kobor MS, Albert ML, et al. The immune factors driving DNA methylation variation in human blood. *Nat Commun.* 2022;13:5895.
100. Zaghlool SB, Kuhnelt B, Elhadad MA, Kader S, Halama A, Thareja G, Engelke R, Sarwath H, Al-Dous EK, Mohamoud YA, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat Commun.* 2020;11:15.
101. Gadd DA, Hillary RF, McCartney DL, Shi L, Stolcyn A, Robertson NA, Walker RM, McGeachan RI, Campbell A, Xueyi S, et al. Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health. *Nat Commun.* 2022;13:4670.
102. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun.* 2022;13:761.
103. Rahmani E, Schweiger R, Rhead B, Criswell LA, Barcellos LF, Eskin E, Rosset S, Sankaraman S, Halperin E. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat Commun.* 2019;10:3417.
104. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods.* 2018;15:1059–66.
105. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One.* 2008;3:e3395.
106. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, Visscher PM. From basic science to clinical application of polygenic risk scores: A Primer. *JAMA Psychiatry.* 2021;78:101–9.
107. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.* 2011;88:294–305.
108. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet.* 2014;15:765–76.

109. Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics*. 1950;35:212–36.
110. Relton CL, Groom A, St Pourcain B, Sayers AE, Swan DC, Embleton ND, Pearce MS, Ring SM, Northstone K, Tobias JH, et al. DNA methylation patterns in cord blood DNA and body size in childhood. *PLoS One*. 2012;7:e31821.
111. 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;51:584–91.
112. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47:979–86.
113. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, Klotzle B, Bibikova M, Fan JB, Gao Y, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49:359–67.
114. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14:R115.
115. Kassam I, Tan S, Gan FF, Saw WY, Tan LWL, Moong DKN, Soong R, Teo YY, Loh M. Genome-wide identification of cis DNA methylation quantitative trait loci in three Southeast Asian Populations Human. *Mol Genet*. 2021;30:603–18. <https://doi.org/10.1093/hmg/ddab038>.
116. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, Colicino E, Waite LL, Joehanes R, Guan W, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biology*. 2016;17:255.
117. Breeze CE, Wong JYY, Beck S, Berndt SJ, Franceschini N. Diversity in EWAS: current state, challenges, and solutions. *Genome Med*. 2022;14:71.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

