Received: 16 May 2022 | Revised: 27 June 2022 | Accepted: 19 July 2022

DOI: 10.1002/gepi.22496

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Investigating the prediction of CpG methylation levels from SNP genotype data to help elucidate relationships between methylation, gene expression and complex traits

James J. Fryett[1] | **Andrew P. Morris**[2] | **Heather J. Cordell**[1]

[1]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK

[2]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, University of Manchester, Manchester, UK

Correspondence
Heather J. Cordell, Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.
Email: heather.cordell@newcastle.ac.uk

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/M011186/1; Versus Arthritis, Grant/Award Number: 21754; Wellcome Trust, Grant/Award Number: 102858/Z/13/Z

Abstract

As popularised by PrediXcan (and related methods), transcriptome-wide association studies (TWAS), in which gene expression is imputed from single-nucleotide polymorphism (SNP) genotypes and tested for association with a phenotype, are a popular approach for investigating the role of gene expression in complex traits. Like gene expression, DNA methylation is an important biological process and, being under genetic regulation, may be imputable from SNP genotypes. Here, we investigate prediction of CpG methylation levels from SNP genotype data to help elucidate relationships between methylation, gene expression and complex traits. We start by examining how well CpG methylation can be predicted from SNP genotypes, comparing three penalised regression approaches and examining whether changing the window size improves prediction accuracy. Although methylation at most CpG sites cannot be accurately predicted from SNP genotypes, for a subset it can be predicted well. We next apply our methylation prediction models (trained using the optimal method and window size) to carry out a methylome-wide association study (MWAS) of primary biliary cholangitis. We intersect the regions identified via MWAS with those identified via TWAS, providing insight into the interplay between CpG methylation, gene expression and disease status. We conclude that MWAS has the potential to improve understanding of biological mechanisms in complex traits.

KEYWORDS
MetaXcan, MWAS, PrediXcan, TWAS

1 | INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified regions of the genome associated with a range of phenotypes (MacArthur et al., 2017). However, for many of these findings, the mechanism by which variants affect their associated phenotype remains unknown (Gallagher & Chen-Plotkin, 2018). Most trait-associated variants identified by GWAS fall in regulatory regions of the genome (Maurano et al., 2012), and are

hypothesised to act by altering gene expression rather than the protein code. Indeed, enrichment of expression quantitative trait loci (eQTLs) at known GWAS risk loci (Nicolae et al., 2010), and overlaps between GWAS risk variants and genomic loci affecting markers of genome regulation (such as histone modifications) have been identified (Chen et al., 2016; Tehranchi et al., 2016; X. Zhang, Joehanes, et al., 2015), reinforcing this hypothesis. For this reason, an approach to improve understanding of mechanisms underlying GWAS findings is to integrate GWAS and gene expression data. One such approach is the transcriptome-wide association study (TWAS), implemented in the software packages PrediXcan (Gamazon et al., 2015), S-PrediXcan (Barbeira et al., 2018) and FUSION (Gusev et al., 2016). This approach uses known relationships between single-nucleotide polymorphisms (SNPs) and gene expression (estimated from a reference panel with matched genotype and gene expression data) to impute expression into GWAS samples. Imputed expression is then tested for association with the phenotype to identify phenotype-relevant genes. This method has been widely used to investigate the role of gene expression in complex traits (Ioannidis et al., 2018; Khawaja et al., 2018; Mancuso et al., 2018; Roselli et al., 2018), and represents a powerful approach for interpretation of GWAS findings.

CpG methylation, which refers to the addition of a methyl (–CH3) group to cytosine residues in cytosine-guanine dinucleotides, is known to regulate the expression of nearby genes. For example, increased methylation at CpG sites in promoter regions is often associated with decreased expression at a nearby gene (although the relationship is often more complex than this) (Luo et al., 2018; Schubeler, 2015). Additionally, aberrant methylation at CpG sites has been implicated as a potential mechanism in complex diseases (Dhana et al., 2018; Story Jovanova et al., 2018; Xu et al., 2018). Like gene expression, DNA methylation is under genetic regulation. Twin and family-based studies have identified a significant heritable component of CpG methylation, with estimates of heritability ranging from 16% to 20% (Bell et al., 2012; Grundberg et al., 2013; Hannon, Knox, et al., 2018; van Dongen et al., 2016). Studies estimating CpG methylation heritability using SNPs also find a significant heritable component, although estimates vary depending on which SNPs are used. For example, a large study using SNPs across the whole genome found an estimate of 19% (van Dongen et al., 2016), similar to estimates from twin studies, whereas studies focussing on heritability attributable to SNPs proximal to CpG sites generated smaller estimates (Quon et al., 2013; Rowlatt et al., 2016). In addition, studies have consistently identified relationships between CpG methylation and genotypes at individual SNPs, termed methylation quantitative trait loci (mQTLs) (Gaunt et al., 2016; Grundberg et al., 2013; Richardson et al., 2016; Volkov et al., 2016). The presence of these mQTLs and the non-zero heritability estimates of CpG methylation indicate that it may be possible to predict methylation from SNP genotypes.

Here, we apply the PrediXcan approach, originally designed for testing for association between predicted gene expression levels and a phenotype, to the problem of testing for association between predicted methylation levels and a phenotype, assuming that genome-wide SNP data is available (in both training and test data sets) to inform the prediction. We start by investigating how well CpG methylation can be predicted from SNP genotypes local to CpG sites, using data from the Accessible Resource for Integrated Epigenomics Studies (ARIES) (Relton et al., 2015), a study within the Avon Longitudinal Study of Parents and Children (ALSPAC) (Boyd et al., 2013; Fraser et al., 2013) and data from the Understanding Society study (Hannon, Gorrie-Stone, et al., 2018). We compare the performance of three penalised regression methods and investigate prediction accuracy at five window sizes to identify an optimal method and window size for prediction model training. For CpG sites where methylation can be predicted well, we then generate prediction models and illustrate their use in a methylome-wide association study (MWAS) of the autoimmune liver disease primary biliary cholangitis (PBC). Finally, we investigate the relationships between regions identified via MWAS and those identified via TWAS, providing insight into the interplay between CpG methylation, gene expression and disease status.

2 | MATERIALS AND METHODS

2.1 | ARIES data

We obtained approval to access genotype data and CpG methylation data measured at an antenatal clinic for 855 mothers as part of the ARIES study, a study within ALSPAC (Boyd et al., 2013; Fraser et al., 2013). The ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Consent for

biological samples has been collected in accordance with the Human Tissue Act (2004).

Details of the collection and processing of the ARIES data can be found in the Supporting Information Text. Following processing and quality control, we were left with matched genotype and CpG methylation data at the "antenatal" time point for 841 ARIES samples to be taken forward for statistical modelling.

2.2 | Understanding Society data

We also obtained approval to access genotype data (University of Essex et al., 2015) and methylation data (University of Essex et al., 2017) previously generated by Understanding Society: the UK Household Longitudinal Study. Details of the collection and processing of the Understanding Society data can be found in the Supporting Information Text. Following processing and quality control, we had matched genotype and methylation data for 1120 samples to be taken forward for downstream analysis.

2.3 | Training and testing CpG methylation prediction models

CpG methylation prediction models were generated separately in the ARIES and Understanding Society data sets by regressing methylation levels on genotype dosages of all SNPs within a specified distance (window size) of the CpG site, using three different penalised regression methods (see Supporting Information Text). The methods considered were ridge regression (Hoerl & Kennard, 2000), LASSO (Tibshirani, 1996) and elastic net with mixing parameter α set to 0.5 (Zou & Hastie, 2005). Although, in principle, one could consider the mixing parameter α as a parameter to be estimated (e.g., by performing a grid search), fixing its value at 0.5 has the advantage of reducing the computational complexity and matches what was done in the original PrediXcan publication (Gamazon et al., 2015), effectively providing a balance between the level of penalisation employed by ridge regression and LASSO. The prediction models were trained in R using the *glmnet* package (Friedman et al., 2010). For all methods, a value for the regularisation parameter λ was selected using 10-fold cross-validation. Any values of λ that produced a prediction model that did not contain any SNPs were excluded. Of the remaining values of λ, the value at which the minimum mean squared error between predicted and observed methylation was achieved in the cross-validation was then selected.

The ARIES data were used to carry out a comparison of penalised regression approaches, while both ARIES and Understanding Society data were used to carry out a comparison of SNP window sizes. To compare the three penalised regression approaches, 50% of ARIES samples were designated as the model training set, and 20% as the test set. (The remaining 30% of samples were saved for use later as a prediction model testing set). For each CpG site, prediction models were trained by regressing CpG methylation on SNP genotypes for all SNPs within 1 Mb of the CpG site using each method, the resulting models were applied to the test set, and the correlation (R) between predicted and observed methylation levels was calculated.

For the comparison of five window sizes, CpG methylation prediction models were trained by regressing CpG methylation on genotypes of SNPs within 250 kb, 500 kb, 1 Mb, 2 Mb or 3 Mb of the CpG site, using elastic net with $\alpha = 0.5$ and using the same training and testing sets as used previously to compare penalised regression methods. By limiting the maximum window size to 3MB, we effectively focus on SNPs that act as *cis* mQTLs; although a role has been demonstrated for *trans* mQTLs (including both interchromosomal effects and intra-chromosomal effects, operating at distances >5 Mb of the CpG site) (Min et al., 2020), they represent a much smaller percentage of total mQTLs (8.5% compared to the 92.5% represented by *cis* mQTLs) and, moreover, the effect sizes for *trans* mQTLs are lower than for *cis* mQTLs, meaning that much larger sample sizes are required to reliably identify them (Min et al., 2020).

Having identified an optimal method from the three methods considered, and a CpG-specific window size for prediction model training, new CpG methylation prediction models were trained using the ARIES and Understanding Society data sets, to establish how accurately methylation could be predicted. The 50% of ARIES samples that had previously been used as a prediction model training set and the 20% that had previously been used for prediction model testing were combined and used as the prediction model training set here. The remaining 30% of samples that had not been used before this point were used as a prediction model testing set to evaluate overall predictive accuracy. The same procedure was used to generate training and testing sets for assessing predictive accuracy with the Understanding Society data.

2.4 | Enrichment testing

Having estimated predictive accuracy, enrichment tests were performed to determine the extent to which five

pre-specified functional annotations were more highly represented among the set of well-predicted CpG sites than among the background set of all CpG sites that passed quality control. Separately, we also tested whether the same five annotations were more highly represented among the set of trait-associated CpG sites than among all CpG sites tested in the MWAS. Enrichment tests were applied to the results obtained using ARIES data and Understanding Society data using annotations taken from the Illumina manifest files. As the annotations listed in the 450k chip and EPIC chip manifest files were slightly different, separate enrichment tests were performed for the results obtained using ARIES data and Understanding Society data. For full details of the procedure, see the Supporting Information Text.

2.5 | Heritability estimation

The heritability of methylation at each CpG site was estimated using restricted maximum likelihood (REML) analysis in GCTA (Yang et al., 2010, 2011). Heritability estimates were generated separately for the ARIES data and the Understanding Society data. For each CpG, the SNPs within the optimal window size of the CpG site's genomic location were used to construct a genetic relationship matrix (GRM). The proportion of the variance of CpG methylation explained by these SNPs (the narrow-sense heritability) was then estimated using REML analysis in GCTA. Heritability estimates were restricted to fall within the [0, 1] range.

2.6 | Training and validating a final set of CpG methylation prediction models

Having obtained an estimate of prediction accuracy using the optimal method and window size for each CpG site, our final step was to train a set of CpG methylation prediction models that could be used in MWAS to investigate the relationship between predicted CpG methylation and complex traits.

To maximise the prediction accuracy of these final CpG methylation prediction models (and thus improve the power of the subsequent tests of association between predicted methylation and phenotype), the sample size of the prediction model training set was increased by combining the 70% training set and 30% testing set. This resulted in two training sets (one comprised of ARIES data and one of Understanding Society data), each consisting of 100% of their samples. Using these training sets, prediction models were then trained for the CpG sites where a prediction accuracy estimate ≥ 0.1 had been achieved at the optimal method and window size. This resulted in 78,250 CpG methylation prediction models trained using 100% of the ARIES data and 207,525 CpG methylation prediction models trained using 100% of the Understanding Society data.

The prediction models were then validated through application to the other data set on which the models had not been trained (see Supporting Information Text). The prediction models trained on 100% of ARIES data were applied to 100% of the Understanding Society samples, and the correlation between predicted and observed methylation was calculated. Similarly, prediction models trained on 100% of Understanding Society data were applied to 100% of the ARIES samples, and the correlation between predicted and observed methylation was calculated. Models which failed to meet a prediction accuracy R estimate ≥ 0.1 in their respective validation data set were discarded from further consideration.

2.7 | MWAS of PBC

As an illustration of our approach, the final CpG methylation prediction models were used within the S-PrediXcan (Barbeira et al., 2018) software package, together with summary statistics from our recent genome-wide meta-analysis of the autoimmune liver disease PBC (Cordell et al., 2021), to perform an MWAS, testing for association between predicted methylation at up to 78,250 ARIES and 207,525 Understanding Society CpGs and disease status. This analysis represents an updated version of the analysis previously described (Cordell et al., 2021) which had used earlier, unoptimised, versions of the methylation prediction models. We also used the most recent (GTEx v8) gene expression and splicing (eQTL and sQTL) prediction models from the PredictDB Data Repository (https://predictdb.org/) in S-PrediXcan, together with the same set of PBC summary statistics, to test for association between disease status and predicted gene expression/splicing, allowing us to intersect the regions identified via MWAS with those identified via TWAS and a splicing site wide association study (SWAS), respectively.

3 | RESULTS

3.1 | Sparse methods outperform polygenic methods at prediction of CpG methylation from SNP genotypes

We first compared the performance of three penalised regression approaches for the prediction of CpG

methylation from local SNP genotypes. Overall, sparse approaches (elastic net and LASSO) outperformed the more polygenic ridge regression approach (Figure 1a). Across the CpG sites successfully modelled with all three methods, higher average prediction accuracy estimates were achieved with LASSO (mean $R = 0.097$, SD = 0.191) and elastic net (mean $R = 0.098$, SD = 0.191) than with ridge regression (mean $R = 0.080$, SD = 0.164). Estimates from all three methods were highly correlated (Figure 1b). While the difference between methods was hard to see when considering all CpGs, it could be seen more clearly when looking at the 10,004 CpGs for which an R estimate ≥ 0.5 was achieved by any of the three methods (Supporting Information: Figure S1).

3.2 | The optimal window size for fitting CpG methylation prediction models is CpG-specific

We next sought to investigate whether changing the window size used to select SNPs used for model fitting could improve the models' accuracy. Prediction accuracy estimates obtained at the five window sizes were highly correlated with one another (Figure 2a,b). On average, a slight decrease in the mean accuracy achieved across the CpG sites successfully modelled at all five window sizes was observed with an increase in the window size (Table 1). However, when looking at the results on a CpG-by-CpG basis, no clear pattern was observed, with some CpG sites showing greater prediction accuracy at the larger window sizes and other CpG sites showing greater prediction accuracy at the smaller window sizes. The same result was observed when restricting the comparison to those CpG sites for which an R estimate ≥ 0.5 was achieved at any of the five window sizes (Supporting Information: Figure S2). This suggests that the optimal window size for training CpG methylation prediction models is a CpG-specific quantity. For each CpG site, the optimal window size was therefore determined as the window size at which the maximum prediction accuracy was achieved.

The same comparison of five window sizes was performed on the other data set considered, the Understanding Society study, to identify optimal window sizes for those CpG methylation measurements. 50% of Understanding Society samples were designated as the training set, with 20% of samples assigned to the testing set. Overall, prediction accuracy estimates at the five window sizes were highly correlated with one another (Supporting Information: Figure S3). Again, some CpG sites showed greater prediction at larger window sizes, while others showed greater accuracy at the smaller window sizes, reinforcing the conclusion that the optimal window size for CpG methylation prediction model training is CpG-specific. The same conclusion was once again reached when the comparison was restricted to just the CpG sites where a prediction accuracy estimate ≥ 0.5 was achieved with any of the five window sizes (Supporting Information: Figure S4). For each CpG site in the Understanding Society data set, the optimal window size was therefore determined as the window size at which the maximum prediction accuracy was achieved.

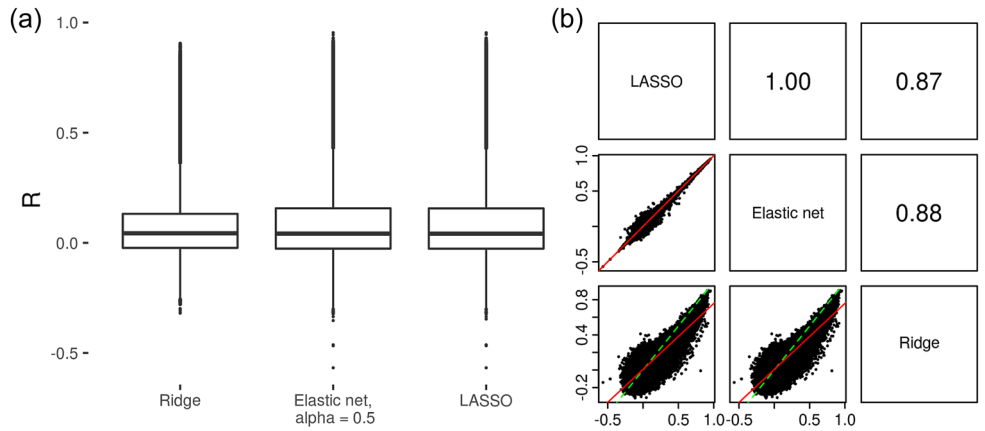

FIGURE 1 Comparison of penalised regression approaches for predicting CpG methylation. (a) Box plots of prediction accuracy estimates (R) from training and testing prediction models using 3 forms of penalised regression (ridge regression, elastic net, LASSO) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (b) Correlation plots between prediction accuracy estimates achieved using the three penalised regression approaches. In the lower panels, each point represents a CpG site, with the R achieved by two methods displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between x and y (red solid line). Upper panels show the pairwise correlations between the R values achieved using the three methods.

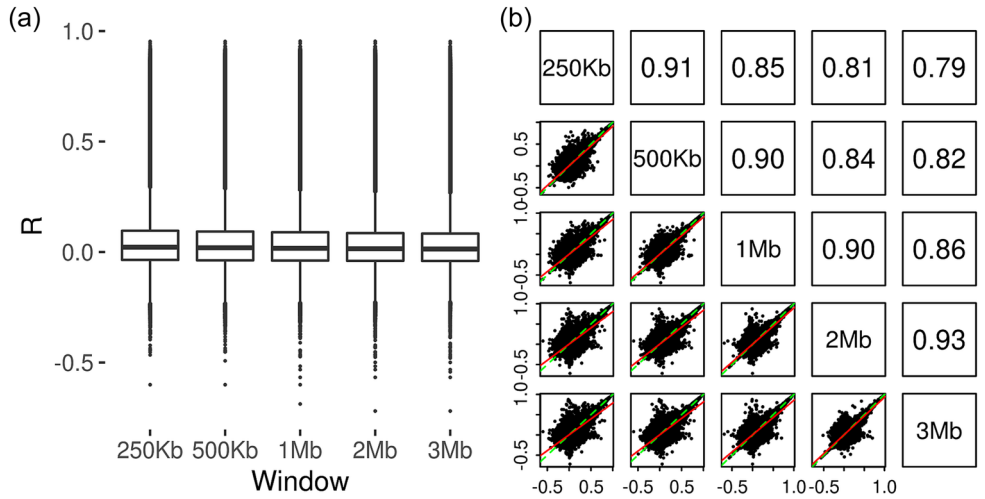

FIGURE 2 Comparison of window sizes for predicting CpG methylation. (a) Box plots of prediction accuracy estimates (*R*) from training and testing prediction models using elastic net with single-nucleotide polymorphisms selected using five window sizes (250 kb, 500 kb, 1 Mb, 2 Mb and 3 Mb) on ARIES data. The line within the box represents the median, with the edges of the box the upper and lower quartiles. (b) Correlation plots between prediction accuracy estimates achieved using the five window sizes. In the lower panels, each point represents a CpG site, with the *R* achieved at the two window sizes displayed on the axes. Also shown are the line of equality (green dashed line) and a best fit line between *x* and *y* (red solid line). Upper panels show the pairwise correlations between the *R* values achieved at the five window sizes.

TABLE 1 Average prediction accuracy estimates achieved when training and testing CpG methylation prediction models using five different window sizes using ARIES data.

Window size	Average prediction accuracy
250 kb	0.0548
500 kb	0.0525
1 Mb	0.0502
2 Mb	0.0481
3 Mb	0.0467

3.3 | Methylation at most CpG sites cannot be accurately predicted from SNP genotypes

Having identified an optimal method (elastic net with α set to 0.5) from the three methods considered, and a CpG-specific window size for prediction model training, new CpG methylation prediction models were trained in the 70% of the ARIES and Understanding Society data sets using these optimal values. The prediction models were then applied to their respective 30% testing sets (i.e., ARIES-trained models applied to the ARIES testing set and Understanding Society-trained models applied to the Understanding Society testing set), and the correlation between predicted and measured methylation was calculated.

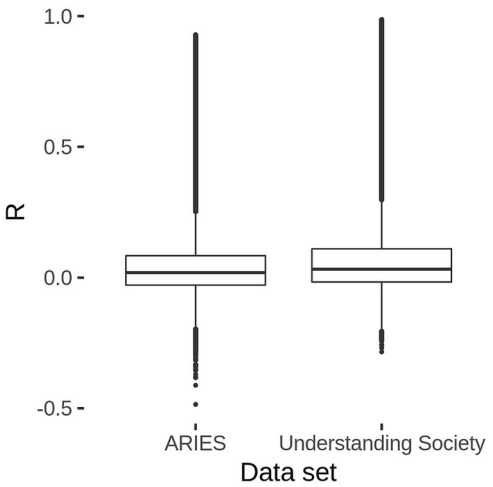

FIGURE 3 Prediction accuracy of CpG methylation prediction models trained using elastic net with $\alpha = 0.5$ and with a CpG-specific window size. Box plots of prediction accuracy estimates (*R*) from training and testing prediction models using elastic net (with $\alpha = 0.5$) with a CpG-specific window size using data from ARIES and Understanding Society. The line within the box represents the median, with the edges of the box the upper and lower quartiles.

We found that methylation at most CpG sites could not be accurately predicted from SNP genotypes (Figure 3), although there existed a subset of CpG sites for which methylation could be predicted with some accuracy. Reassuringly, prediction accuracy estimates from ARIES data were highly correlated ($r = 0.757$) with those obtained

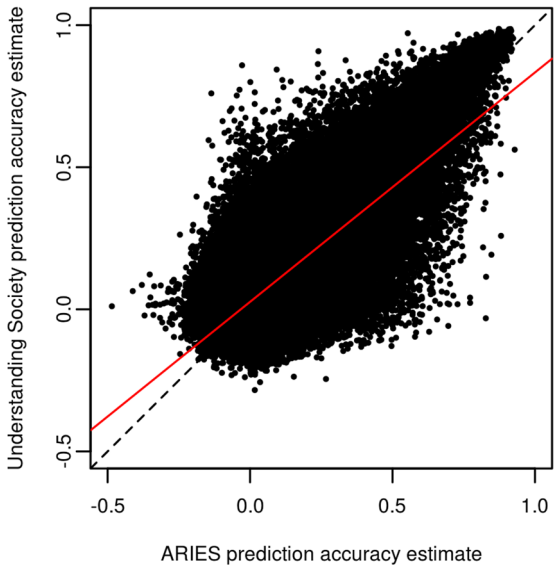

FIGURE 4 Comparison of prediction accuracy estimates from ARIES and Understanding Society data sets. Each point represents a CpG site, with its prediction accuracy estimate obtained from training and testing a prediction model using the ARIES data shown on the *x* axis, and its prediction accuracy estimate obtained from training and testing a prediction model using the Understanding Society data shown on the *y* axis. The red line represents a best fit line, and the dashed line represents the line of equality ($y = x$).

from Understanding Society data (Figure 4). This was especially the case for the well-predicted CpG sites such as cg16906346, which showed a prediction accuracy of 0.924 when examined using ARIES data and a prediction accuracy estimate of 0.953 when examined using Understanding Society data. In total, prediction models for 78,250 CpG sites from ARIES (assayed using the HumanMethylation450 BeadChip) and 207,525 CpG sites from Understanding Society (assayed using the denser MethylationEPIC array) showed a prediction accuracy ≥ 0.1 when predicted using the optimal method and window size. These CpG sites were taken forward for further analysis.

Of particular interest were the CpG sites where methylation could be predicted with a high degree of accuracy. 10,220 ARIES-trained models and 30,865 Understanding Society-trained models showed prediction accuracy ≥ 0.5 in their respective test sets, representing a set of well-predicted CpG sites. To learn more about these well-predicted CpG sites, enrichment testing was conducted. When considering the CpGs that were predicted well when using ARIES data, CpG sites tagged to genes (odds ratio [OR] = 0.682, $p = 4.08 \times 10^{-68}$), CpG sites located at CpG islands (OR = 0.883, $p = 1.72 \times 10^{-9}$) and CpG sites tagged to promoters (OR = 0.636, $p = 1.13 \times 10^{-61}$) were all depleted among the set of well-predicted CpG sites (when compared to the background set of all CpG sites), while CpGs at enhancer regions (OR = 1.42, $p = 9.62 \times 10^{-53}$) and CpGs at DNAse1 hypersensitivity sites (OR = 1.40, $p = 8.57 \times 10^{-33}$) were enriched among the well-predicted CpG sites. Reassuringly, these enrichments and depletions were replicated when looking at those CpG sites that were well-predicted when using the Understanding Society data (Supporting Information: Figure S5).

The theoretical upper limit on how accurately CpG methylation can be predicted from SNP genotypes is equivalent to its narrow-sense heritability. We estimated the heritability of methylation at each CpG site using SNPs within the CpG-specific optimal window size and compared heritability estimates with estimates of prediction accuracy obtained with the optimal method and window size. Overall, heritability estimates were highly correlated and concordant with prediction accuracy estimates, with the exception of a small number of CpG sites where heritability estimates were much greater than the prediction accuracy estimates). This was observed for both the ARIES and the Understanding Society data sets (Supporting Information: Figure S6), suggesting that the upper bound on prediction accuracy had been reached for most CpG sites.

3.4 | Training and validating a final set of CpG methylation prediction models

Having obtained an estimate of prediction accuracy using the optimal method and window size for each CpG site, the final step was to use the full set (100%) of samples in the ARIES and Understanding Society data sets to train and validate a set of CpG methylation prediction models that could be used in MWAS to investigate the relationship between predicted CpG methylation and complex traits. Following this procedure (see Supplementary Text for full details), we ended up with a total of 232,356 prediction models. These 232,356 prediction models covered 193,315 unique CpG sites, with 39,041 CpG sites represented by both an ARIES-trained and an Understanding Society-trained prediction model.

3.5 | MWAS in PBC identifies regions that overlap with TWAS and SWAS signals

MWAS was carried out to test for association between predicted methylation and disease status at the 48,658 (out of a possible 78,250) ARIES CpGs and 172,008 (out of a possible 207,525) Understanding Society CpGs for which sufficient SNPs were available in the PBC summary statistics to inform the tests. Results are shown

in Figure 5, along with TWAS and SWAS results from similar tests of association between PBC and predicted gene expression and splicing, respectively. The association results generated at the Understanding Society CpGs (outermost circle) were considerably stronger than those generated at the sparser set of ARIES CpGs (second circle), resulting in 782 significant Understanding Society CPGs ($p < 2.91 \times 10^{-7}$, corresponding to $p = 0.05$ Bonferroni-corrected for the 172,008 tests performed). As expected, these significant Understanding Society CpGs corresponded to GWAS association signals of association between SNPs and PBC (innermost circle), but also, in many cases, to TWAS and SWAS signals (i.e., regions of association between PBC and predicted gene expression [third circle] and/or splicing [fourth circle]). Figure 6 shows the implicated genes ($p < 7.79 \times 10^{-6}$,

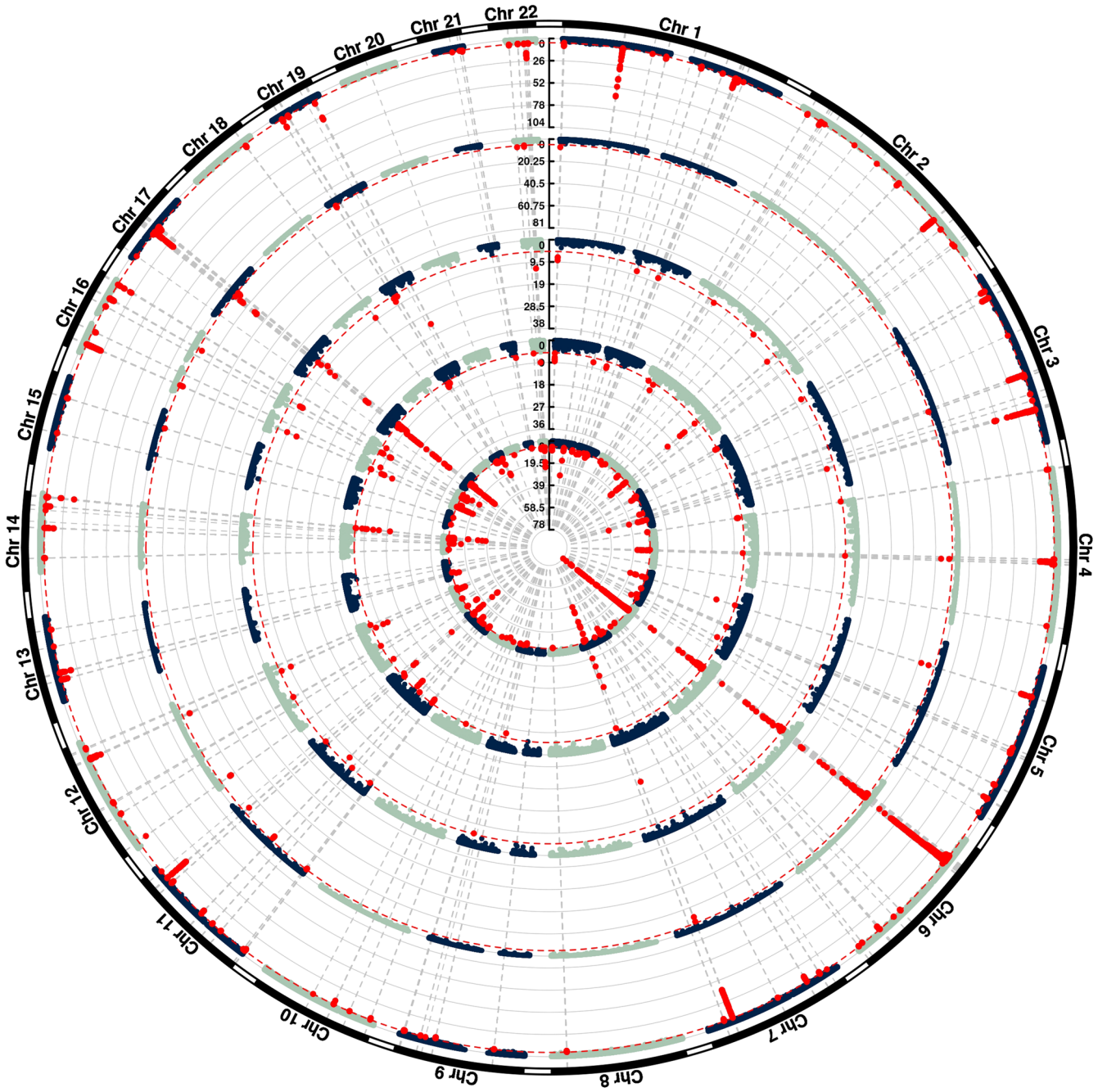

FIGURE 5 Circular Manhattan plot. Shown are the −log10 p values from tests of association with PBC and predicted methylation at Understanding Society CpGs (outermost circle), predicted methylation at ARIES CpGs (second circle), predicted gene expression (third circle), predicted splicing (fourth circle) and measured (genotyped or imputed) SNPs (innermost circle). PBC, primary biliary cholangitis; SNP, single-nucleotide polymorphism.

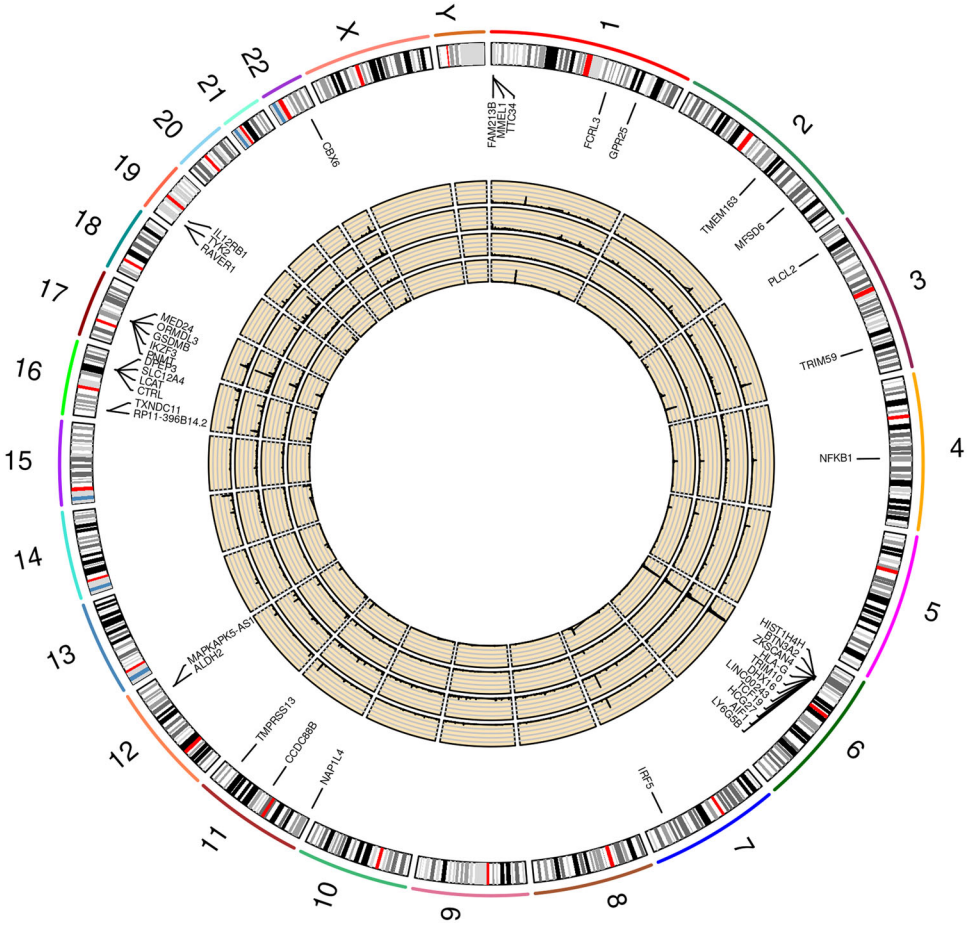

FIGURE 6 Circos plot. Shown via links to the ideogram are the significant genes along with inner circles showing the −log10 p values from tests of association with PBC and predicted methylation at Understanding Society CpGs (outermost circle), predicted gene expression (second circle), predicted splicing (third circle) and measured (genotyped or imputed) SNPs (innermost circle). PBC, primary biliary cholangitis; SNP, single-nucleotide polymorphism.

corresponding to $p = 0.05$ Bonferroni-corrected for the 6419 genes tested)—except for some regions on chromosomes 6 and 17 where there were too many significant genes to be plotted—along with MWAS results at Understanding Society CpGs (outermost circle), TWAS (second circle) results, SWAS results (third circle) and GWAS results (innermost circle). A full list of the 63 implicated genes is given in Supporting Information: Table 1, while Supporting Information: Table 2 additionally includes the significant CPGs and significant splicing sites ($p < 6.39 \times 10^{-6}$, corresponding to $p = 0.05$ Bonferroni-corrected for the 7825 tests performed). Further detailed analysis of each of the 63 implicated gene regions, using fine-mapping and complementary approaches such as co-localisation (Giambartolomei et al., 2018), Mendelian Randomisation (Zuber et al., 2022), and Bayesian Network analysis (Howey et al., 2020, 2021) is beyond the scope of the current study, but will be carried out in the near future to help elucidate the underlying causal relationships between the different biological phenomena investigated here and the extent to which the same genetic factor(s) may underpin the associations observed.

4 | DISCUSSION

Here we have extended the approach originally popularised by PrediXcan (Gamazon et al., 2015) for testing association between predicted gene expression levels and a phenotype, to instead test for association between predicted *methylation* levels and a phenotype, allowing one to carry out MWAS to identify associations between the trait and imputed methylation at CpGs across the genome. In the original PrediXcan publication (Gamazon et al., 2015), the authors considered prediction models based on LASSO, elastic net with $\alpha = 0.5$ and polygenic scores. They found that LASSO performed similarly to elastic net and both methods outperformed polygenic scores; they subsequently focused on prediction models

using elastic net because it performed well and was more robust to slight changes in input SNPs.

Comparing the performance of penalised regression approaches for predicting methylation, similarly to what we had seen previously for gene expression (Fryett et al., 2020), we found that sparse models (trained with LASSO or elastic net with $\alpha = 0.5$) tended to outperform more polygenic models (trained with ridge regression), suggesting that the underlying local genetic architecture of CpG methylation is sparse. While there has been no formal investigation of the genetic architecture of CpG methylation, mQTL studies have found that most CpG sites have few mQTLs (if any), each with a large effect size (Gaunt et al., 2016), which is indicative of a sparse local architecture. Gene expression has also been shown to have a sparse local architecture (Wheeler et al., 2016), indicating this local sparsity may a feature shared by multiple cellular traits.

A comparison of prediction accuracy estimates when training CpG methylation prediction models using a range of different window sizes showed that increasing window size led to a marginal decrease on average prediction accuracy across many CpG sites, although the effect on the prediction accuracy of individual CpG sites varied. Interestingly, there was no consistent direction of effect to this, with some CpG sites benefitting from a smaller window size, and others benefitting from a larger window size, suggesting that the optimal window size for the prediction of CpG methylation is a CpG-specific quantity. Before our analysis, there had been no investigation into the effect of window size on the accuracy with which CpG methylation can be predicted from SNP genotypes. However, given that mQTL studies have shown that methylation at a small number of CpGs is regulated by SNPs distal to the CpG sites (Gaunt et al., 2016), it is perhaps unsurprising that increasing the window size to the point where some of these more distal regulatory SNPs can be included in the prediction models could improve prediction accuracy for some CpG sites. In contrast, increasing the window size for the CpG sites where methylation is not known to be regulated by distal SNPs could lead to increased noise in the CpG methylation prediction model fitting procedure, leading to poorer estimation of the prediction model coefficients, and subsequently poorer prediction accuracy. Given that these distal regulatory SNPs are only known to exist for some, not all, CpG sites, this may explain why the average prediction accuracy across all CpG sites examined here fell slightly as the window size increased.

A crucial finding from our study is that methylation at most CpG sites cannot be accurately predicted from SNP genotypes. Through comparison with heritability estimates obtained using GCTA, we found that prediction accuracy estimates for most CpG sites examined here approached their upper bound, and so are unlikely to be increased much further at the current window size and sample size. Despite most CpG sites showing poor prediction accuracy and heritability estimates, there exist a set of CpG sites where methylation can be predicted with accuracy, with some CpG sites showing high degrees of prediction accuracy. We found that these well-predicted CpG sites are enriched at enhancers and depleted at promoters, matching previously identified enrichments for CpG sites with mQTLs (Banovich et al., 2014; Gutierrez-Arcelus et al., 2013). Importantly, prediction accuracy estimates for well-predicted CpG sites replicated when tested on an independent data set, indicating that we have identified real, reliable relationships between genotype and measured methylation. For these robustly predicted CpG sites, our approach represents a powerful method to investigate their role in complex traits.

As an illustration of our approach, we applied our final CpG methylation prediction models to GWAS summary data from PBC, identifying 782 significant CpGs, many of which localised with significant regions of association between predicted gene expression and/or splicing. Further interrogation of these regions using advanced co-localisation and causal modelling analysis techniques will help elucidate the interplay between CpG methylation, gene expression and disease status, thus improving our understanding of underlying biological mechanisms. This PBC data set was chosen largely for convenience (as we had easy access to the GWAS summary statistics, generated previously by ourselves) but also partly because our previous work with PBC led us to expect the existence of strong GWAS signals, making this a good data set to illustrate the approach; in principle any disease exhibiting similarly strong GWAS signals should work equally well.

We conclude by briefly discussing some limitations of our study. We trained CpG methylation prediction models using methylation data measured in blood, since the only large-scale data sets available with both genome-wide SNP data and genome-wide methylation are blood-based. However, blood is unlikely to be the true causal tissue of interest for many complex traits. Studies have found strong concordance between the effects of SNPs on methylation in blood and the effects of those same SNPs on methylation in other tissues (Hannon et al., 2017; Lin et al., 2018; Shi et al., 2014), however, the number of tissues studied and the sample sizes used in these studies has been limited. Thus, it is difficult to say how well our prediction models or association results translate to tissues of interest. Additionally, both data sets used to generate prediction models are from populations of

British ancestry. To date, there has been little study of how genetic effects on methylation differ across populations. In TWAS, gene expression prediction accuracy is reduced when using prediction models trained using samples of a different ancestry to the samples in the GWAS data (Mikhaylova & Thornton, 2019; Mogil et al., 2018). Should the CpG methylation prediction models generated here be used in an MWAS of a non-European population, a similar reduction in prediction accuracy would likely be observed.

In our study, we focussed on comparing a limited set of (penalised regression based) prediction methods and a limited set of five possible window sizes for choosing the SNP predictors. This was in part motivated by the success of such approaches for predicting gene expression (Gamazon et al., 2015) for use in subsequent association testing (via TWAS) with a phenotype of interest. Although alternative methods (e.g., based on support vector machines or deep learning) have been developed to predict DNA methylation (Bhasin et al., 2005; Levy et al., 2020; Tang et al., 2020; Tian et al., 2019; W. Zhang, Spector, et al., 2015; Zhou et al., 2012), these methods are not generally designed to predict methylation from SNP genotype data alone (as would be needed to take the models forward for MWAS, in conjunction with individual-level SNP genotypes or GWAS summary statistics for a phenotype of interest), but they rather make use of additional features (such as full DNA sequence data, histone modification marks or transcription factor binding sites) to inform the prediction. This is a much richer set of features than would generally be available in publicly (or privately) available GWAS data sets, limiting the applicability for subsequent MWAS of any models that encompass these features.

Limiting the search for SNP predictors to five possible (CpG-specific) window sizes was largely a pragmatic choice. It is possible that improved prediction for any given CpG could be achieved through use of a window size not considered here, or through a more complicated scheme such as adapting the window size to the local LD pattern. However, we note that optimal prediction of methylation per se is not our ultimate goal; we are more interested in the power to detect associations through MWAS, which relies not only on the accuracy of the methylation imputation, but also on the sample size of the GWAS data (or summary statistics) to be used. Thus, an association can still be detected for CpGs with low prediction R, provided a GWAS with a sufficiently large sample size is used.

In conclusion, our study suggests that MWAS based on imputed methylation levels represents a potentially powerful approach for aiding the interpretation of GWAS data and interrogating the relationship between CpG methylation and gene expression. Further development and application of this method may help improve understanding of the role of CpG methylation and gene expression in complex trait biology and to identify potential targets for disease therapy.

AUTHOR CONTRIBUTIONS

Heather J. Cordell and Andrew P. Morris conceived and designed the project and played an important role in interpreting the results. James J. Fryett and Heather J. Cordell carried out data analysis and drafted the manuscript. All authors contributed to revising the manuscript and approved the final paper.

ACKNOWLEDGEMENTS

This study includes data from ALSPAC. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. This study was funded in whole, or in part, by the Wellcome Trust (Grant numbers: 102858/Z/13/Z and 219424/Z/19/Z). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. J. J. F. was funded by a BBSRC DTP studentship (BB/M011186/1). H. J. C. was funded by a Wellcome Senior Research Fellowship in Basic Biomedical Science (102858/Z/13/Z) and a Wellcome Investigator Award in Science (219424/Z/19/Z). A. P. M. acknowledges support from Versus Arthritis (grant reference 21754). This study includes data from ALSPAC. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Collection of genotype data from the ARIES mothers was funded by the Wellcome Trust (WT088806). Collection of the ARIES methylation data was funded by the BBSRC (BBI025751/1 and BB/I025263/1), the MRC Integrative Epidemiology Unit at the University of Bristol (MC_UU_12013/1 & MC_UU_12013/2 & MC_UU_12013/8), the LLHW via MRC (G1001357) and the Wellcome Trust (WT092830/Z/10/Z). This study includes data from Understanding Society: The UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council (Grant Number: ES/M008592/1). The data were collected by NatCen and the genome wide scan data were analysed by the Wellcome Trust

Sanger Institute. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/. Data governance was provided by the METADAC data access committee, funded by ESRC, Wellcome, and MRC. (2015-2018: Grant Number MR/N01104X/1 2018-2020: Grant Number ES/S008349/1). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

ARIES (ALSPAC) and Understanding Society data are available by application to ALSPAC (http://www.bristol.ac.uk/alspac/researchers/access/) and Understanding Society (https://www.understandingsociety.ac.uk/documentation/health-assessment/accessing-data), respectively. Results generated during this study can be found within the published article and its supplementary files, with the final ARIES and Understanding Society methylation prediction models that we developed (.db files, ready for use with S-PrediXcan) freely available for download from https://www.staff.ncl.ac.uk/heather.cordell/MethPaper.html.

ORCID

Heather J. Cordell http://orcid.org/0000-0002-1879-5572

REFERENCES

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., & Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics*, *10*(9), e1004663. https://doi.org/10.1371/journal.pgen.1004663

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., Stahl, E. A., Huckins, L. M., GTEx, C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1), 1825. https://doi.org/10.1038/s41467-018-03621-1

Bell, J. T., Tsai, P. C., Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S. Y., Dempster, E. L., Murray, R. M., Grundberg, E., Hedman, A. K., Nica, A., Small, K. S., MuTHER, C., Dermitzakis, E. T., ... Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genetics*, *8*(4), e1002629. https://doi.org/10.1371/journal.pgen.1002629

Bhasin, M., Zhang, H., Reinherz, E. L., & Reche, P. A. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Letters*, *579*(20), 4302–4308. https://doi.org/10.1016/j.febslet.2005.07.002

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—The index offspring of the Avon Longitudinal Study of parents and children. *International Journal of Epidemiology*, *42*(1), 111–127. https://doi.org/10.1093/ije/dys064

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A. L., Fernandez, J. M., Rowlston, S., Wilder, S. P., Farrow, S., ... Soranzo, N. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, *167*(5), 1398–1414. https://doi.org/10.1016/j.cell.2016.10.026

Cordell, H. J., Fryett, J. J., Ueno, K., Darlay, R., Aiba, Y., Hitomi, Y., Kawashima, M., Nishida, N., Khor, S. S., Gervais, O., Kawai, Y., Nagasaki, M., Tokunaga, K., Tang, R., Shi, Y., Li, Z., Juran, B. D., Atkinson, E. J., Gerussi, A., ... UK-PBC, C. (2021). An international genome-wide meta-analysis of primary biliary cholangitis: Novel risk loci and candidate drugs. *Journal of Hepatology*, *75*(3), 572–581. https://doi.org/10.1016/j.jhep.2021.04.055

Dhana, K., Braun, K., Nano, J., Voortman, T., Demerath, E. W., Guan, W., Fornage, M., van Meurs, J., Uitterlinden, A. G., Hofman, A., Franco, O. H., & Dehghan, A. (2018). An Epigenome-Wide Association Study (EWAS) of obesity-related traits. *American Journal of Epidemiology*, *187*, 1662–1669. https://doi.org/10.1093/aje/kwy025

van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J. J., Helmer, Q., Dolan, C. V., Ehli, E. A., Davies, G. E., van Iterson, M., Breeze, C. E., Beck, S., BIOS, C., Suchiman, H. E., Jansen, R., van Meurs, J. B., Heijmans, B. T., Slagboom, P. E., & Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, *7*, 11115. https://doi.org/10.1038/ncomms11115

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., & Lawlor, D. A. (2013). Cohort profile: The Avon Longitudinal Study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*, *42*(1), 97–110. https://doi.org/10.1093/ije/dys066

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Fryett, J. J., Morris, A. P., & Cordell, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. *Genetic Epidemiology*, *44*(5), 425–441. https://doi.org/10.1002/gepi.22290

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS era: From association to function. *American Journal of Human*

Genetics, *102*(5), 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx, C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/ng.3367

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W. L., Ho, K., Ring, S. M., Evans, D. M., Davey Smith, G., & Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*, 61. https://doi.org/10.1186/s13059-016-0926-z

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., CommonMind, C., Pasaniuc, B., & Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, *34*(15), 2538–2545. https://doi.org/10.1093/bioinformatics/bty147

Grundberg, E., Meduri, E., Sandling, J. K., Hedman, A. K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K. S., Ge, B., Caron, M., Shin, S. Y., Multiple Tissue Human Expression Resource, C., Lathrop, M., Dermitzakis, E. T., ... Deloukas, P. (2013). Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *American Journal of Human Genetics*, *93*(5), 876–890. https://doi.org/10.1016/j.ajhg.2013.10.004

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., ... Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. https://doi.org/10.1038/ng.3506

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., ... Dermitzakis, E. T. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*, *2*, e00523. https://doi.org/10.7554/eLife.00523

Hannon, E., Gorrie-Stone, T. J., Smart, M. C., Burrage, J., Hughes, A., Bao, Y., Kumari, M., Schalkwyk, L. C., & Mill, J. (2018). Leveraging DNA-Methylation Quantitative-Trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *American Journal of Human Genetics*, *103*(5), 654–665. https://doi.org/10.1016/j.ajhg.2018.09.007

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C., Belsky, D. W., Corcoran, D. L., Arseneault, L., Moffitt, T. E., Caspi, A., & Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genetics*, *14*(8), e1007544. https://doi.org/10.1371/journal.pgen.1007544

Hannon, E., Weedon, M., Bray, N., O'Donovan, M., & Mill, J. (2017). Pleiotropic effects of trait-associated genetic variation on DNA methylation: Utility for refining GWAS loci. *American Journal of Human Genetics*, *100*(6), 954–959. https://doi.org/10.1016/j.ajhg.2017.04.013

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Howey, R., Clark, A. D., Naamane, N., Reynard, L. N., Pratt, A. G., & Cordell, H. J. (2021). A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships. *PLoS Genetics*, *17*(9), e1009811. https://doi.org/10.1371/journal.pgen.1009811

Howey, R., Shin, S. Y., Relton, C., Davey Smith, G., & Cordell, H. J. (2020). Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. *PLoS Genetics*, *16*(3), e1008198. https://doi.org/10.1371/journal.pgen.1008198

Ioannidis, N. M., Wang, W., Furlotte, N. A., Hinds, D. A., Research, T., Bustamante, C. D., Jorgenson, E., Asgari, M. M., & Whittemore, A. S. (2018). Gene expression imputation identifies candidate genes and susceptibility loci associated with cutaneous squamous cell carcinoma. *Nature Communications*, *9*(1), 4264. https://doi.org/10.1038/s41467-018-06149-6

Khawaja, A. P., Cooke Bailey, J. N., Wareham, N. J., Scott, R. A., Simcoe, M., Igo RP, J. r, Jr., Song, Y. E., Wojciechowski, R., Cheng, C. Y., Khaw, P. T., Pasquale, L. R., Haines, J. L., Foster, P. J., Wiggs, J. L., Hammond, C. J., Hysi, P. G., UK Biobank Eye and Vision, C., & NEIGHBORHOOD, C. (2018). Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nature Genetics*, *50*(6), 778–782. https://doi.org/10.1038/s41588-018-0126-8

Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., & Christensen, B. C. (2020). MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics*, *21*(1), 108. https://doi.org/10.1186/s12859-020-3443-8

Lin, D., Chen, J., Perrone-Bizzozero, N., Bustillo, J. R., Du, Y., Calhoun, V. D., & Liu, J. (2018). Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia. *Genome Medicine*, *10*(1), 13. https://doi.org/10.1186/s13073-018-0519-4

Luo, C., Hajkova, P., & Ecker, J. R. (2018). Dynamic DNA methylation: In the right place at the right time. *Science*, *361*(6409), 1336–1340. https://doi.org/10.1126/science.aat6806

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog. *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K. L., Kote-Jarai, Z., Eeles, R., Freedman, M., Haiman, C., & Pasaniuc, B. (2018). Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature Communications*, *9*(1), 4079. https://doi.org/10.1038/s41467-018-06302-1

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

Mikhaylova, A. V., & Thornton, T. A. (2019). Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations. *Frontiers in Genetics*, *10*, 261. https://doi.org/10.3389/fgene.2019.00261

Min, J. L., Hemani, G., Hannon, E., Dekkers, K. F., Castillo-Fernandez, J., Luijk, R., & Relton, C. L. (2020). Genomic and phenomic insights from an Atlas of genetic effects on DNA methylation. *medRxiv*, 2020. https://doi.org/10.1101/2020.09.01.20180406

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., Johnson, W. C., Im, H. K., Liu, Y., & Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genetics*, *14*(8), e1007586. https://doi.org/10.1371/journal.pgen.1007586

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. https://doi.org/10.1371/journal.pgen.1000888

Quon, G., Lippert, C., Heckerman, D., & Listgarten, J. (2013). Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Research*, *41*(4), 2095–2104. https://doi.org/10.1093/nar/gks1449

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., Woodward, G., Lyttleton, O., Evans, D. M., Reik, W., Paul, Y. L., Ficz, G., Ozanne, S. E., Wipat, A., Flanagan, K., Lister, A., Heijmans, B. T., Ring, S. M., & Davey Smith, G. (2015). Data resource profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International Journal of Epidemiology*, *44*(4), 1181–1190. https://doi.org/10.1093/ije/dyv072

Richardson, T. G., Shihab, H. A., Hemani, G., Zheng, J., Hannon, E., Mill, J., Carnero-Montoro, E., Bell, J. T., Lyttleton, O., McArdle, W. L., Ring, S. M., Rodriguez, S., Campbell, C., Smith, G. D., Relton, C. L., Timpson, N. J., & Gaunt, T. R. (2016). Collapsed methylation quantitative trait loci analysis for low frequency and rare variants. *Human Molecular Genetics*, *25*(19), 4339–4349. https://doi.org/10.1093/hmg/ddw283

Roselli, C., Chaffin, M. D., Weng, L. C., Aeschbacher, S., Ahlberg, G., Albert, C. M., Almgren, P., Alonso, A., Anderson, C. D., Aragam, K. G., Arking, D. E., Barnard, J., Bartz, T. M., Benjamin, E. J., Bihlmeyer, N. A., Bis, J. C., Bloom, H. L., Boerwinkle, E., Bottinger, E. B., ... Rienstra, M. (2018). Multi-ethnic genome-wide association study for atrial fibrillation. *Nature Genetics*, *50*(9), 1225–1233. https://doi.org/10.1038/s41588-018-0133-9

Rowlatt, A., Hernández-Suárez, G., Sanabria-Salas, M. C., Serrano-López, M., Rawlik, K., Hernandez-Illan, E., Alenda, C., Castillejo, A., Soto, J. L., Haley, C. S., & Tenesa, A. (2016). The heritability and patterns of DNA methylation in normal human colorectum. *Human Molecular Genetics*, *25*(12), 2600–2611. https://doi.org/10.1093/hmg/ddw072

Schubeler, D. (2015). Function and information content of DNA methylation. *Nature*, *517*(7534), 321–326. https://doi.org/10.1038/nature14192

Shi, J., Marconett, C. N., Duan, J., Hyland, P. L., Li, P., Wang, Z., Wheeler, W., Zhou, B., Campan, M., Lee, D. S., Huang, J., Zhou, W., Triche, T., Amundadottir, L., Warner, A., Hutchinson, A., Chen, P. H., Chung, B. S., Pesatori, A. C., ... Landi, M. T. (2014). Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nature Communications*, *5*, 3365. https://doi.org/10.1038/ncomms4365

Story Jovanova, O., Nedeljkovic, I., Spieler, D., Walker, R. M., Liu, C., Luciano, M., Bressler, J., Brody, J., Drake, A. J., Evans, K. L., Gondalia, R., Kunze, S., Kuhnel, B., Lahti, J., Lemaitre, R. N., Marioni, R. E., Swenson, B., Himali, J. J., Wu, H., ... Amin, N. (2018). DNA methylation signatures of depressive symptoms in middle-aged and elderly persons: Meta-analysis of multiethnic epigenome-wide studies. *JAMA Psychiatry*, *75*(9), 949–959. https://doi.org/10.1001/jamapsychiatry.2018.1725

Tang, J., Zou, J., Zhang, X., Fan, M., Tian, Q., Fu, S., Gao, S., & Fan, S. (2020). PretiMeth: Precise prediction models for DNA methylation based on single methylation mark. *BMC Genomics*, *21*(1), 364. https://doi.org/10.1186/s12864-020-6768-9

Tehranchi, A. K., Myrthil, M., Martin, T., Hie, B. L., Golan, D., & Fraser, H. B. (2016). Pooled ChIP-Seq links variation in transcription factor binding to complex disease risk. *Cell*, *165*(3), 730–741. https://doi.org/10.1016/j.cell.2016.03.041

Tian, Q., Zou, J., Tang, J., Fang, Y., Yu, Z., & Fan, S. (2019). MRCNN: A deep learning model for regression of genome-wide DNA methylation. *BMC Genomics*, *20*(Suppl. 2), 192. https://doi.org/10.1186/s12864-019-5488-5

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

University of Essex, Institute for Social and Economic Research, N. S. R., & University of Exeter Medical School. (2017). *Understanding Society: DNA methylation data based on Illumina methylation EPIC array*.

University of Essex, Institute for Social and Economic Research, N. S. R., & Wellcome Trust Sanger Institute. (2015). *Understanding Society: Genome Wide SNP data based on the Illumina human core exome array*.

Volkov, P., Olsson, A. H., Gillberg, L., Jørgensen, S. W., Brøns, C., Eriksson, K. F., Groop, L., Jansson, P. A., Nilsson, E., Rönn, T., Vaag, A., & Ling, C. (2016). A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. *PLoS One*, *11*(6), e0157776. https://doi.org/10.1371/journal.pone.0157776

Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., Gtex, C., Cox, N. J., Nicolae, D. L., & Im, H. K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS Genetics*, *12*(11), e1006423. https://doi.org/10.1371/journal.pgen.1006423

Xu, C. J., Soderhall, C., Bustamante, M., Baiz, N., Gruzieva, O., Gehring, U., & Koppelman, G. H. (2018). DNA methylation in

childhood asthma: An epigenome-wide meta-analysis. *Lancet Respiratory Medicine*, *6*(5), 379–388. https://doi.org/10.1016/s2213-2600(18)30052-3

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., & Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology*, *16*(1), 14. https://doi.org/10.1186/s13059-015-0581-9

Zhang, X., Joehanes, R., Chen, B. H., Huan, T., Ying, S., Munson, P. J., Johnson, A. D., Levy, D., & O'Donnell, C. J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nature Genetics*, *47*(4), 345–352. https://doi.org/10.1038/ng.3220

Zhou, X., Li, Z., Dai, Z., & Zou, X. (2012). Prediction of methylation CpGs and their methylation degrees in human DNA sequences. *Computers in Biology and Medicine*, *42*(4), 408–413. https://doi.org/10.1016/j.compbiomed.2011.12.008

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320.

Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E., Patel, A., Wallace, C., & Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *American Journal of Human Genetics*, *109*(5), 767–782. https://doi.org/10.1016/j.ajhg.2022.04.001

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fryett, J. J., Morris, A. P., & Cordell, H. J. (2022). Investigating the prediction of CpG methylation levels from SNP genotype data to help elucidate relationships between methylation, gene expression and complex traits. *Genetic Epidemiology*, 46, 629–643. https://doi.org/10.1002/gepi.22496