An atlas of genetically-regulated DNA methylation in normal human cell types

Evan Wu[1], Temidayo Adeluwa[2], Saideep Gona[2], Yichao Zhou[2], Katherine A. Aracena[1], Luis B. Barreiro[3], Mengjie Chen[1,3], Emma Thompson[4], Carole Ober[4], Hae Kyung Im[3,*]

1 Center for Research Informatics, The University of Chicago, Chicago, IL 60637, USA
2 Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, IL 60637, USA
3 Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
4 Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA

*** Correspondence to: haky@uchicago.edu**

Abstract

DNA methylation (DNAm) is an important epigenetic process involved in the etiology of many complex diseases, but its heterogeneity across specific cell-type and disease contexts limits the transferability of available DNAm datasets. To address this, we developed a novel approach, methENPACT, that predicts epigenetic features from the reference genome using the transformer-class model Enformer then uses this output to train cell-type specific elastic net models of DNAm. Leveraging a recently available whole-genome bisulfite sequencing atlas, we trained across 80 distinct cell-types and found that methENPACT models can predict variation in DNAm across the genome with performance ranging from AUC 0.74-0.89 and correlation 0.44-0.61. These predictions explain variant-level effects in DNAm when correlated with independent observed data, despite never explicitly observing across-individual variation. We confirmed that this predictive power was independent of ancestry effects and concordant with the predictions of the most recent linear model of DNAm. Applying methENPACT-predicted DNAm to downstream phenotypes like gene expression and lymphoblastoid growth rate recapitulated expected biological signals and suggested independent pathways of investigation compared to the transcriptome. The convenience and portability of methENPACT to predict DNAm using independent sources of data may allow for the investigation of the functional relevance of DNAm in many more tissue-specific and disease contexts.

Introduction

The advent of accessible whole-genome sequencing technology has led to a revolution in studying the genetic etiology of complex diseases, with more powerful genome-wide association studies (GWAS) continuing to discover novel susceptibility loci. However, biological interpretation of these statistical signals remains elusive as most loci are discovered in noncoding regions [1, 2], which suggests they may be altering the regulation of gene expression rather than the translated sequences. In humans, DNA methylation (DNAm) is the direct modification of DNA by methylation of cytosine residues at approximately 30 million 5'-cytosine-phosphate-guanine-3' (CpG) dinucleotide sites genome-wide, and is thought to primarily repress gene expression by preventing transcription factor (TF) binding and/or recruiting chromatin remodeling complexes [2–4]. Notably, DNAm is also modulated through genetic and environmental variation [4–6], and epigenome-wide studies have linked genomic variation to differential DNAm levels by identifying methylation quantitative trait loci (meQTLs) that may play a role in various diseases. A recent study by Oliva et al. showed that tissue-specific meQTLs are enriched in complex traits and colocalized with GWAS loci, in some cases independent of gene expression QTLs [7]. However, DNAm profiles differ across cell-types and biological conditions and often require invasive procedures to access causal tissues, which is prohibitive to obtaining the large number of samples needed to discover genome- and methylome- wide signals. The specificity of methylomes to tissue and disease conditions restricts the application of existing DNAm data to new studies or meta-analyses compared to GWAS data. To address these limitations, we propose methENPACT (methylation + ENformer + imPACT), a computational method that predicts the cell-type- and individual- specific unobserved methylomes given only DNA sequences and bypassing the need to retrain expensive neural networks for a particular biological context. Using this method we provide an atlas of methENPACT models across 80 cell types, which can be further used to elucidate mechanisms of DNAm underlying a variety of complex biological phenotypes.

Results

Cell-type-specific models of DNAm can predict across the reference genome

MethENPACT follows the ENPACT approach proposed by Adeluwa et al. (preprint under preparation), which sought to predict transcription factor (TF) occupancy using coregulated epigenetic annotations as was demonstrated with the IMPACT method [8]. The key difference is that instead of using observed epigenetic data, ENPACT-based models are trained on the genetically-predicted epigenome predicted by Enformer, a transformer class deep learning model that outputs 5,313 total tracks for human sequences including: 2,313

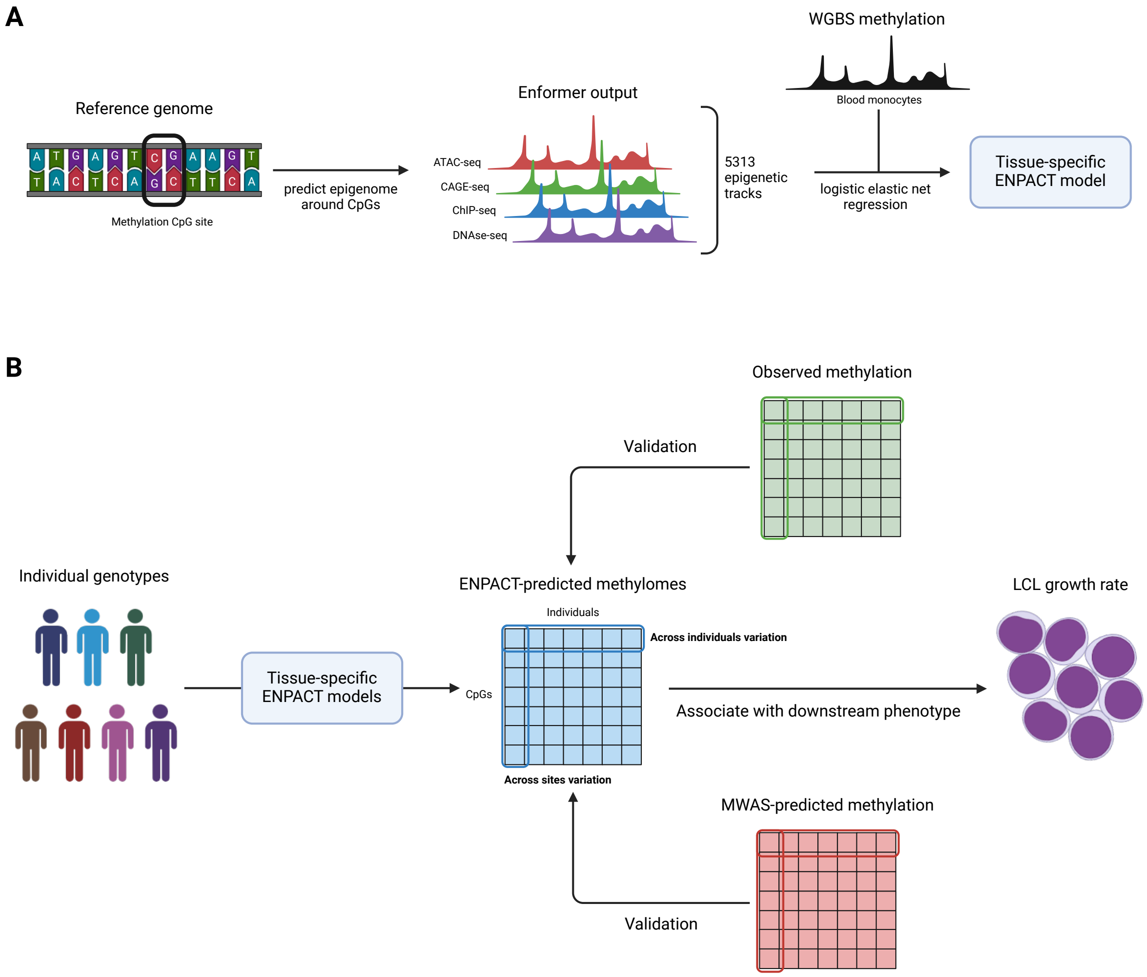
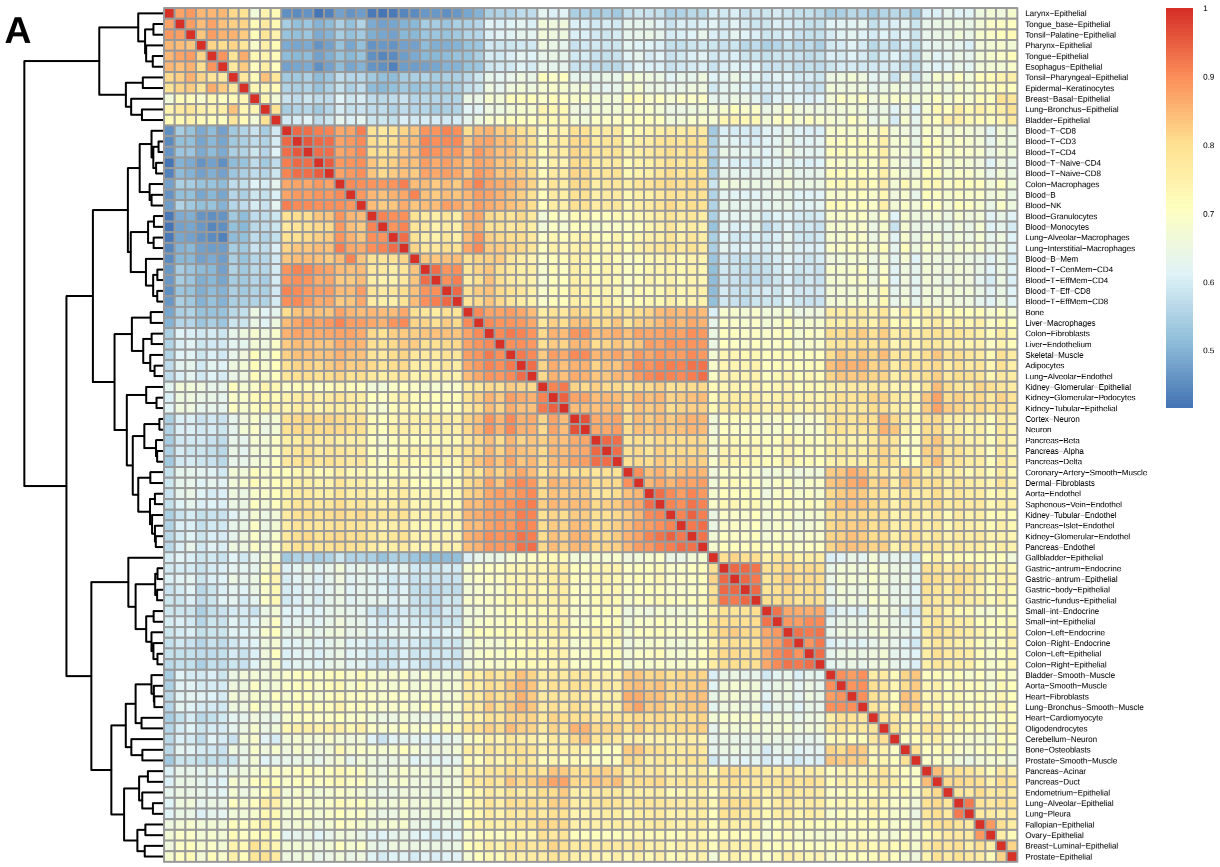


Figure 1. Study design. (A) Training schematic of methENPACT DNAm models. **(B)** MethENPACT validation methods. Figures created in BioRender.com.

TF binding CHIP-seq, 1,860 histone modification CHIP-seq, 684 chromatin accessibility DNAse-seq or ATAC-seq, and 638 gene expression CAGE-seq tracks across various cell types [9]. Training a methENPACT model entails generating the reference epigenome by inputting the human reference genome into Enformer, subsetting the local 512 bp context around CpG training sites, and finally regressing observed DNAm levels against the reference epigenome (Figure 1A). Enformer was trained on sequence variation in the reference genome (across-sites variation) and did not explicitly account for variation seen in different genomes (across-individuals variation). Despite this, we hypothesize that using Enformer as an intermediary may still account for genetic effects across individuals if the effects of motif variation are similar to SNP effects. Using the reference epigenome means that methENPACT models require only DNAm data for training, tolerate any number of biological replicates by aggregating DNAm measurements, and circumvent the problem of aggregating model parameters across individually-trained models. After generating reference Enformer outputs, we trained methENPACT elastic net regression models on whole-genome bisulfite sequencing data from a recent DNAm atlas study by Loyfer et al., encompassing 205 samples across 80 distinct cell types [10]. DNAm levels were quantified as the percentage of reads methylated at each CpG, leading to ambiguity as to what model type would perform best on this dataset. We trained with a 80:20 test split across $\approx 46,000$ CpGs and considered three elastic net model types: untransformed linear regression, linear regression with logit-transformed DNAm values, and logistic regression with binarized DNAm values. After comparing prediction performance, we decided to use a binarized outcome with 15% as the threshold to call methylation as done by Loyfer et al. [10]. Although binarizing the data at an arbitrary cutoff reduces information available and introduces biases, binarized methENPACT models outperformed untransformed and logit-transformed outcomes and showed more favorable qualities, balancing the bimodally distributed DNAm data and avoiding extreme prediction values, making it an acceptable tradeoff given the structure of the data.

Here we show the performance of all cell-type specific methENPACT models. The models showed slight performance differences across the various cell-types, with model AUCs and correlations between predicted DNAm probabilities and observed DNAm in the test set ranging from 0.7445-0.8880 and 0.4366-0.6105,



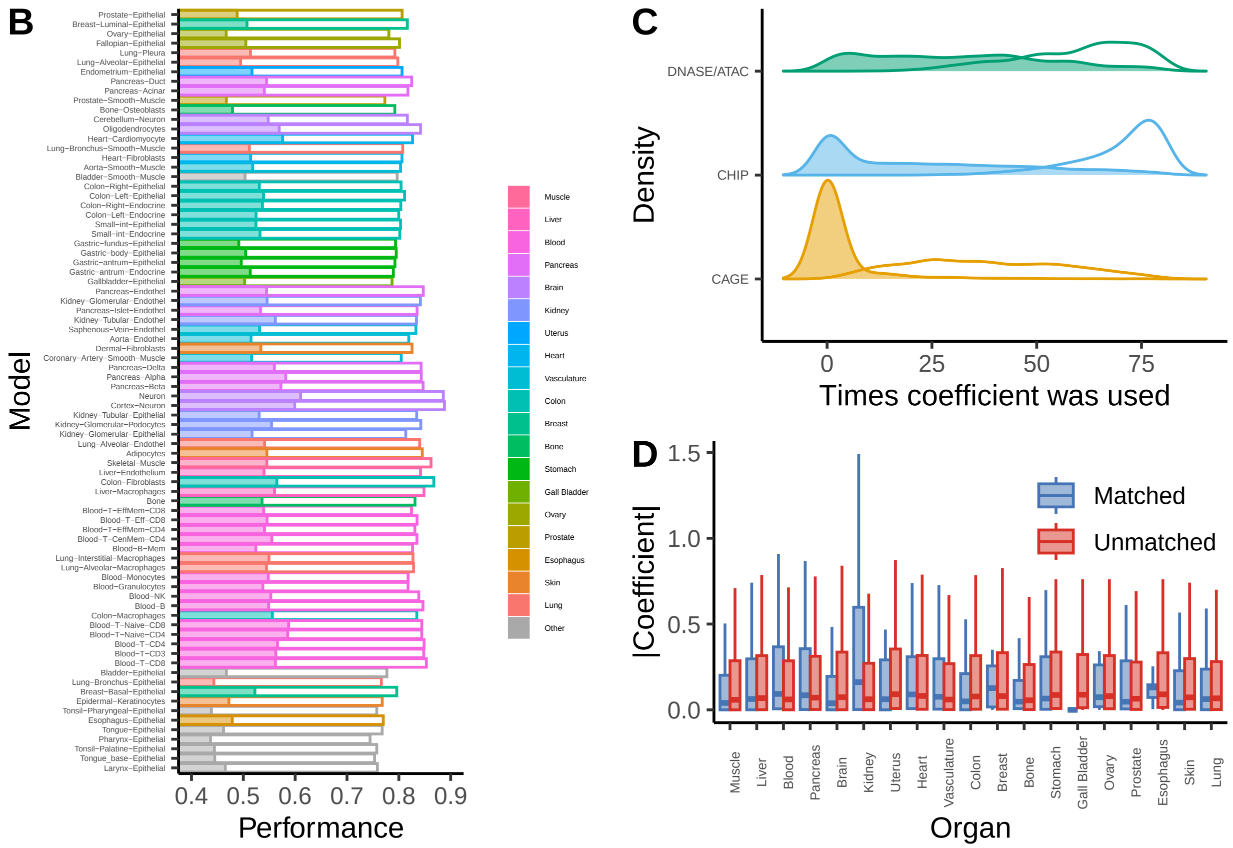


Figure 2. MethENPACT model performance. (A) MethENPACT performance between predicted and observed data. Correlations are shown as filled bars and AUCs shown as outlined bars. **(B)** Correlations between predicted DNAm across 80 cell type-specific models. **(C)** Distribution of the number of times an Enformer track was used as a coefficient at least once across all models. The outlined distribution counts coefficients with a cutoff of $|\beta| > 0$ while the filled distribution counts coefficients with a cutoff of $|\beta| > |\bar{\beta}$ with the mean calculated for each model. **(D)** Enrichment of tracks in matched organs versus unmatched organs.

respectively (Figure 2B). Performance was generally worse in epithelial tissues compared to immune or organ-specific cells, which could be due to variation in tissue preparation methods [13] or intrinsically higher noise in DNAm for these cell-types. Interestingly, the neuronal models had the highest predictive performance, which could be a consequence of either their nonreplicating nature or unique mechanisms of DNAm in the brain that contribute to higher predictability from the local epigenome.

MethENPACT-predicted DNAm also preserves the expected relationships between cell-type and tissue. For all 80 models, we predicted at 1,000 random CpG sites and calculated pairwise correlations (Figure 2A). Models trained on data from both the same tissue and cell type clustered together with higher correlations (≈ 0.7) than unrelated tissues (≈ 0.4). This suggests that DNAm is somewhat shared across tissues but still exhibits heterogeneous profiles across cell types, in line with the results from Loyfer et al. [13]. Analysis of which tracks inform each of the models reveals similar results. There some tracks shared across nearly all cell-types but with low magnitude, but when subsetting for weights that contribute at least the mean absolute magnitude across all tracks for a given model, the distribution is shifted towards lower sharing, suggesting that cell-type specific patterns may be more informative to the models (Figure 2C). *A priori* we would expect that Enformer tracks that share similar biology to the modeled cell-types would be more informative to methENPACT models. To test this, we assigned organ labels for Enformer tracks and models and calculated the distribution of absolute model weights for groups of coefficients where the labels either matched or mismatched separately. There were differences in number and variety of cell-types represented by a single organ label from the methENPACT and Enformer training data which bias the distributions. Seven out of nineteen shared organ labels displayed higher median coefficient magnitude when matching, particularly in kidney models (Figure 2D). Particular cell-type models showed variance within the same organ group but overall organ-level patterns tended to be preserved (not shown). However, many Enformer tracks are highly correlated due to sourcing from the same markers or cell-types, so we should not over-interpret the weights in the elastic net prediction models.

MethENPACT models can predict across sites and individuals from the same cell-type context

Next, we investigated whether methENPACT models can predict across individuals. For this purpose we obtained genotype and DNAm levels in monocyte-derived macrophages across 35 individuals—14 of African and 21 of European ancestry [11]—and assessed the performance of the blood monocyte methENPACT model at 1,000 random CpGs using individualized Enformer predictions (Figure 1B). We first compared predictions across sites and found a small but significant correlation of 0.1575 on average between predicted and observed DNAm (Figure 3A), reassuring us that predictions based on training data can be applied to predict DNAm in an independent dataset.

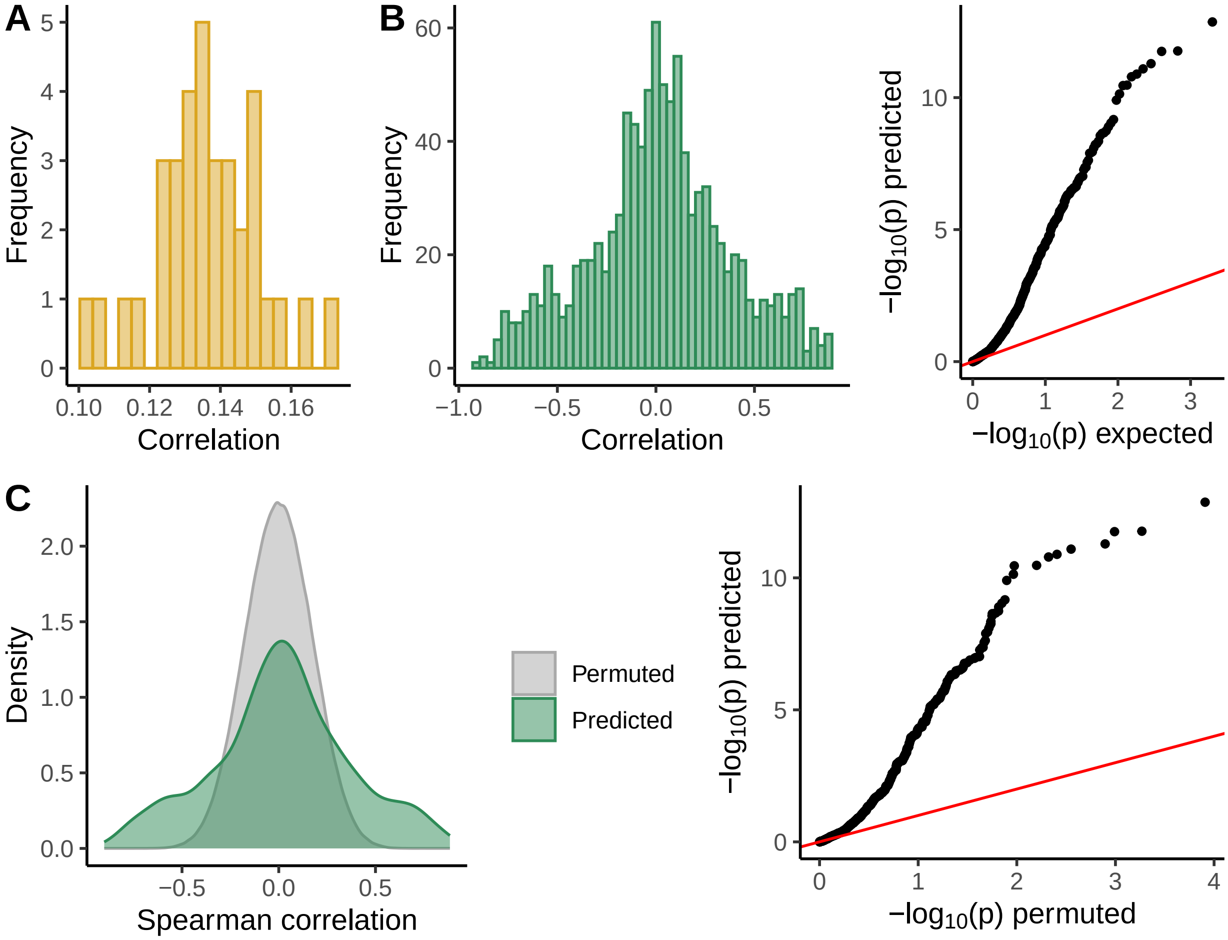


Figure 3. Validation with observed DNAm across individuals. (A) Spearman correlations between predicted DNAm probability and observed DNAm across sites. **(B)** Spearman correlations between predicted DNAm probability and observed DNAm across individuals and Q-Q plot of correlation p values against the uniform null. **(C)** Overlay of correlation densities for predicted DNAm and 1,000 permutations and Q-Q plot of prediction p values against permuted correlation p values.

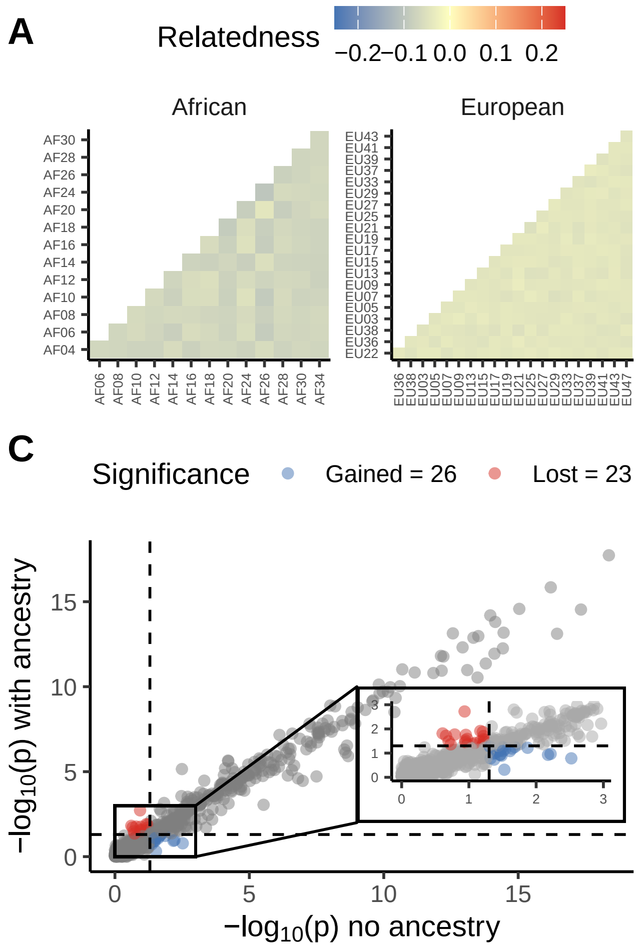


Figure 4. Ancestry effects do not contribute to methENPACT predictions. **(A)** Genetic relatedness calculated within each ancestry group, with only unique pairwise relatedness values shown. **(B)** Across-site and across-individual correlations stratified by individual ancestry. **(C)** P values of predicted DNAm coefficients for models with and without ancestry. Blue and and dots indicate coefficients that gained or lost significance in ancestry-adjusted models compared to ancestry-agnostic models. **(D)** R^2 of ancestry-adjusted against ancestry-agnostic models.

We then assessed methENPACT predictions across individualized variation by computing correlations between predicted and observed DNAm for each of the 1,000 CpG sites (Figure 3B). Around half of the CpGs exhibit a significant negative correlation across individuals, which suggests that these sites are genetically predictable by methENPACT but we obtain the wrong sign of the effect. This sign confounding can be attributed to Enformer ascribing some epigenetic differences to the wrong allele, and has been similarly found in Enformer gene expression predictions [12, 13]. As such, we consider the magnitude of the across-individual correlations as indicative of predictive power available to our model, but the direction of effect is variably informative depending on whether the correct allele was represented during training. We further confirmed the ability of our model to predict across individuals by simulating the null hypothesis that there is no predictive power. This was achieved by permuting individual labels for the prediction data at each site and recalculating the correlation with observed data and its p value (Figure 3C). The longer-tailed distribution of our predictions and more significant correlation p values suggest that our model is able to predict across individuals beyond the null expectation.

Ancestry does not contribute to putative predictive ability

Population and relatedness structures in our validation dataset could have contributed an ancestry effect that masks the true genetic effects on DNAm. We first confirmed that there was unlikely to be relatedness effects between individuals by calculating their genetic relatedness (Figure 4A). We calculated relatedness separately within African and European samples because the Plink estimation method relies on population-level allele frequencies, which inflates estimates of relatedness within each population when distinct populations are analyzed simultaneously. We found that all individuals were unrelated among each other with negative relatedness (< -0.03) across all pairwise estimates. We also confirmed that there was no significant difference in across-site prediction performance between the two populations, suggesting that genetically-predicted DNAm markers are consistent across ancestry groups (Figure 4B).

To assess whether across-individual predictions were affected by ancestry, we linearly regressed observed and predicted DNAm levels for each CpG test site both with and without an ancestry coefficient. Few sites exhibited a significant ancestry coefficient (9.8% of CpGs), suggesting that generally predicted DNAm is not affected by ancestry though particular sites that base predictions on ancestry-specific alleles may be unavoidably affected. The significance of the DNAm prediction coefficient changed for only 5.00% of sites with the inclusion of ancestry information (Figure 4C). This is corroborated by model R-squared values being highly correlated for sites with and without ancestry (Pearson $r = 0.973$) and only increased slightly for sites we could not predict well in the first place (Figure 4D), suggesting that ancestry effects generally do not explain the significant correlations observed in across individual DNAm levels.

MethENPACT predictions are concordant with MWAS predictions across individuals

A recent method that predicts DNAm from genotype data is the methylation-wide association study (MWAS), which uses a reference database of individuals with known genotypes and DNAm levels and applies it to impute the unobserved methylome in independent genomic datasets. A recent implementation of this method [14] trained linear predictors of DNAm in whole blood across $\approx 200,000$ genetically-predictable CpGs, which we used to benchmark the performance of our blood monocyte methENPACT model. We selected 1,000 test CpG sites that were available to the MWAS model and generated methENPACT and MWAS predictions using CEU genotype data from the 1000 Genomes phase 3 release [15]. We did not find significant across-site correlations within individuals between the two models, perhaps relating to model architecture, cell-type, and output score differences. However, we found that the correlations across individuals were highly significant, even more so than our evaluation against observed DNAm (Figure 5A). This suggests that our model picks up many of the same genetic signals as MWAS at genetically-predictable sites; we can readily see concordant effects that appear to be additive and SNP-based between both of the models when looking at the most significantly correlated CpGs (Figure 5B).

MethENPACT-predicted DNam can generate hypotheses by association with downstream phenotypes

DNAm occurs across the genome and its regulatory functions and effects of gene expression vary greatly depending on whether they are present on promoters, enhancers, or gene bodies. CpG sites and CpG islands are numerous at gene promoters upstream of the gene transcriptional start site (TSS) and repress gene expression when DNAm is present [2, 3, 5], which is a readily interpretable functional mechanism by which DNAm may affect downstream biological conditions. Therefore, we investigated whether methENPACT DNAm predictions explain changes in gene expression by correlating the predicted methylome at 18,054 gene TSS sites across 435 individuals from the GEUVADIS study [16], which has provided publicly available matched genotype and gene expression data as well as downstream phenotype data. Across all gene TSS sites of assessed individuals, we see that the correlation between predicted DNAm and gene expression is small but significantly negative (Figure 6A), in line with our expectations that more methylated TSS regions are associated with lower gene expression. Again, the directionality of the DNAm-gene expression relationship cannot be confirmed across individuals due to the variant-agnostic training approach of Enformer, but we still see correlations more significant than the null expectation (Figure 6B) suggesting that at least for some genes, changes in DNAm is associated with changes in gene expression.

Having confirmed that methENPACT-predicted DNAm exhibits expected associations with gene expression, we tested whether differential DNAm regulation can be associated with more downstream phenotypes. One such phenotype with data available to HapMap samples is lymphoblastoid cell line (LCL) growth rate, which was generated based on modeling observed cellular growth rates of immortalized LCL cultures using sample metadata [17]. Different normalized LCL growth rates could be due to individual genetic variation that affects cellular growth processes such as cell cycle and metabolic gene pathways. To test whether DNAm might underlie these genetic changes, we linearly regressed LCL growth rate against gene-wise methENPACT predictions across the 157 individuals shared between the GEUVADIS and HapMap datasets. This approach does not account for the sign confounding noted above, which may be better addressed by integrating association approaches like S-PrediXcan [18]. No genes were found to be significant after FDR correction, but we detected 662 genes significant at the nominal level of $p < 0.05$. Since each individual signal is unlikely to be a true association, we did not focus on individual genes and instead assessed whether the putatively significant genes as a whole displayed patterns that might inform differences in the LCL growth phenotype. We performed an over-representation analysis with Gene Ontology and detected 39 significant gene sets that can be assigned broadly to 5 categories (Figure 6C): nitrogen metabolism, RNA metabolism, development, and transcription factor binding. Nitrogenous compounds like nucleotides and amino acids are essential to facilitate proper cell growth and proliferation [19], and RNA is

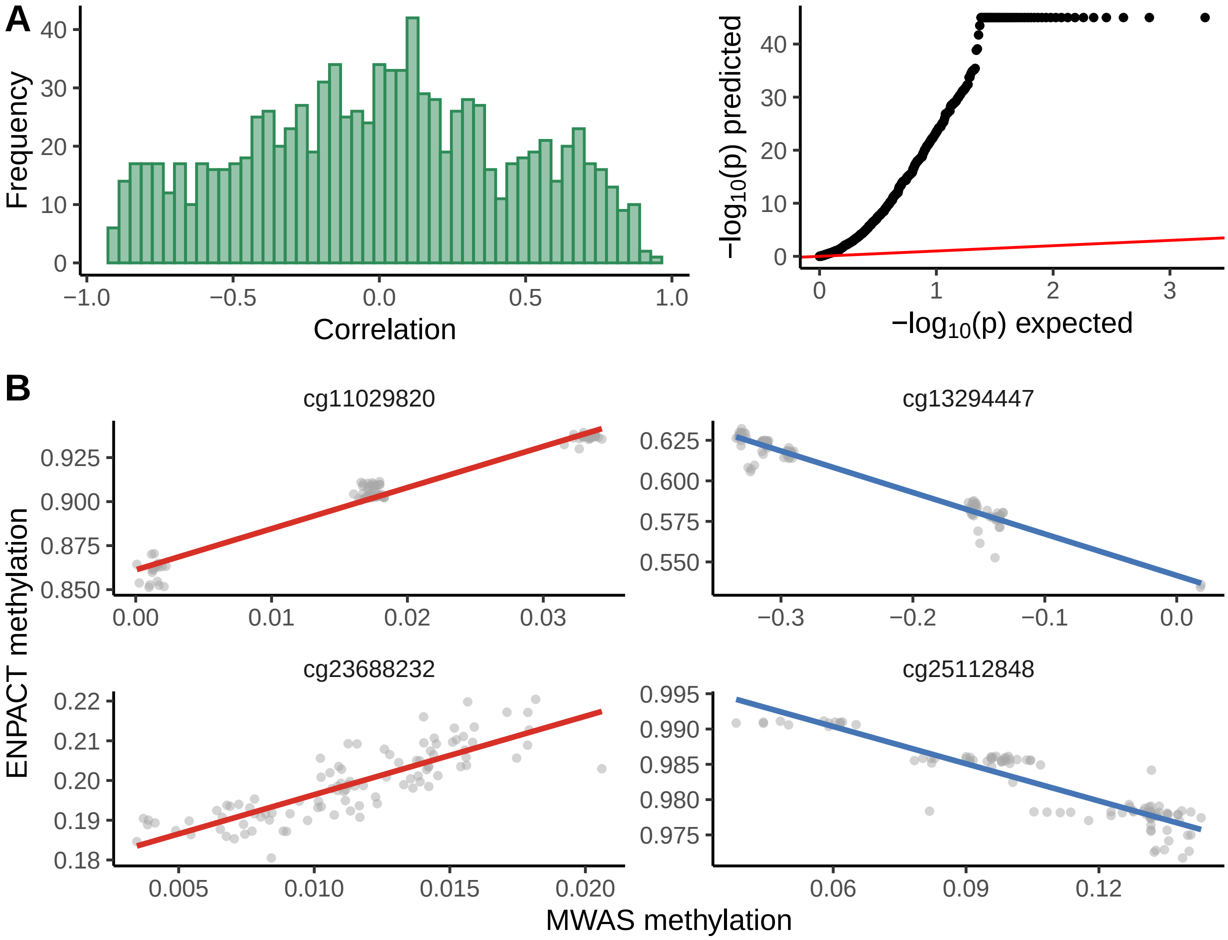


Figure 5. Validation with MWAS model predictions. **(A)** Spearman correlations between methENPACT predicted DNAm probability and MWAS predicted DNAm and Q-Q plot of their p values against the uniform null, values capped at 45. **(B)** Highly correlated predictions across methENPACT and MWAS models.

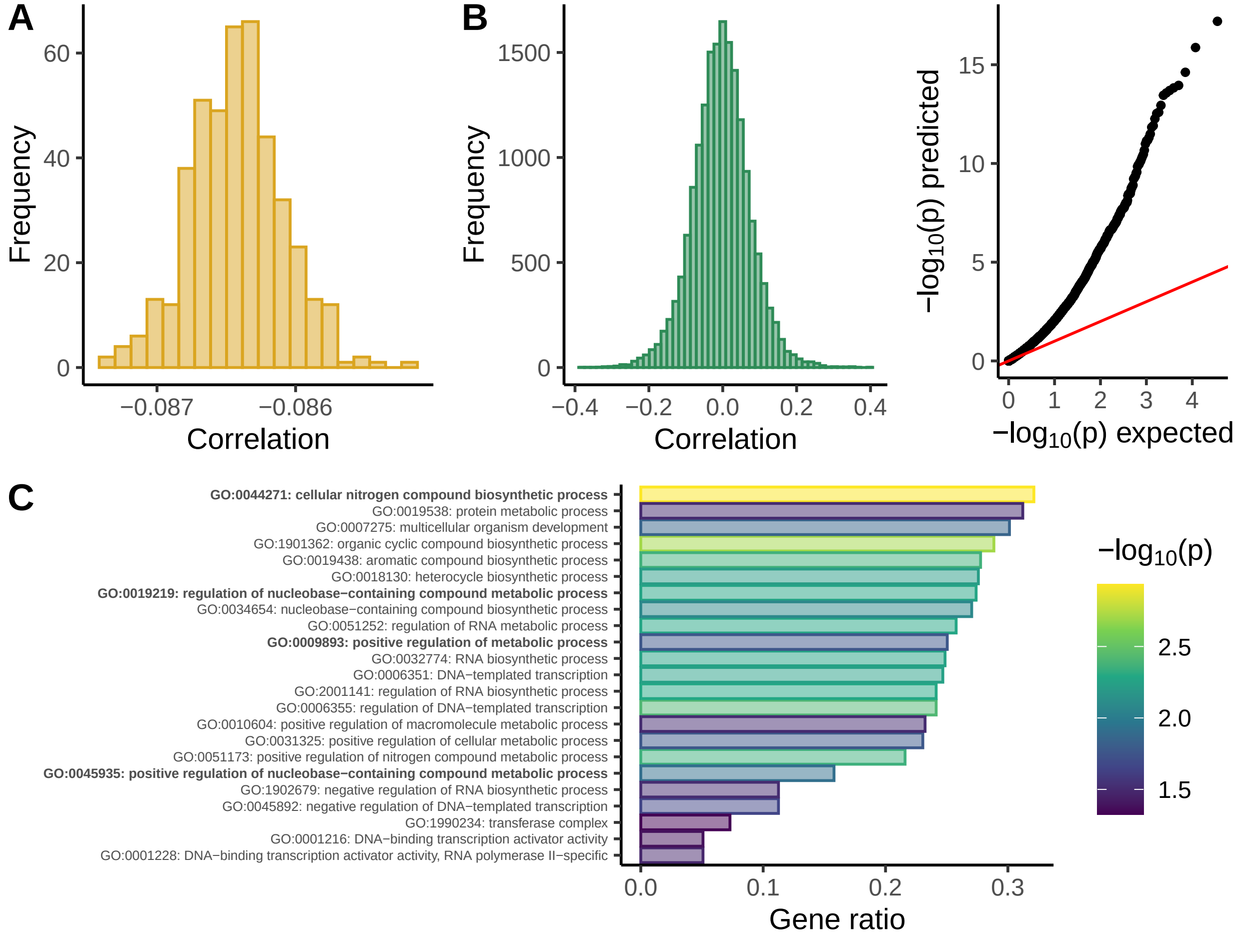


Figure 6. Genetic associations using methENPACT predictions. **(A)** Spearman correlations between methENPACT-predicted DNAm near the TSS and gene expression across sites. **(B)** Spearman correlations between methENPACT-predicted DNAm near the TSS and gene expression across individuals, and Q-Q plot of p values of the correlations against the uniform null. **(C)** Gene ontology enrichment of genes whose predicted DNAm is associated with LCL growth rates (FDR < 0.05).

one such molecule in high demand as it is required for protein synthesis and must be synthesized within the cell [20]. As such, differences in DNAm that affect the regulation of metabolic machinery may cause imbalances in nitrogenous compounds or RNA that slow or speed up LCL proliferation across individuals. These processes were not detected when simply looking at gene expression associated with the LCL growth phenotype, which primarily implicated cell-cycle related gene sets [17], suggesting that probing the genetically-predicted DNAm axis may suggest independent signals and mechanisms of biological significance.

Discussion

In this study, we present methENPACT as a method that is able to computationally predict the genetically-regulated DNAm across sites and individuals. Using methENPACT, we were able to produce 80 cell-type specific models that will allow for the interrogation of the unobserved methylomes in many new contexts, necessitating only genotype data. This framework can further test for associations with traits using GWAS data to generate hypotheses about mechanisms of action for those biological states, as demonstrated with the LCL growth phenotype. New methENPACT models may also be trained with high throughput using just the reference epigenome and a source of DNAm data, which can expand the scope of such analyses into any cell types and diseases not covered here. However, because predicted differential signals are computationally inferred they necessitate follow-up with biological experiments to validate their associations.

We have noted some limitations of methENPACT, in particular the binarization of DNAm training data at 15% and uninformative direction of effect across individuals. However, the good performance of the models during training and validation as well as the ability to propose candidate CpGs associated with traits through unsigned association tests suggests that methENPACT is still applicable to real-world data. An additional limitation is that DNAm at many CpGs likely cannot be genetically predicted due to the limited heritability of DNAm, with average h^2 across the genome at 10-20% [6, 14] and wide variation found across CpGs [14], which might have affected the performance of our predictions. Hidden gene-environment interaction effects that were unaccounted for may also bias our models when trying to apply models across different experiments or datasets. Nevertheless, the applicability and convenience of methENPACT may allow for broader assessments of the role of DNAm in complex diseases.

Methods

Training cell-type-specific methENPACT models

We downloaded bigwig files from a recent DNAm atlas study [10] which included 205 samples covering 80 different cell-types and average DNAm levels across 28 million CpGs whole-genome. Initially we were interested in investigating sites pertaining to allergy and asthma, and chose to focus on 45,887 CpGs from a custom allergy panel selected for functional relevance in allergy and asthma [21]. Bigwig files were converted to BED format and overlapped with the custom panel sites, leaving us with $\approx 44,000$ sites per sample to proceed with model training. We downloaded publicly available Enformer weights [9] and generated outputs in the immediate context of our selected CpGs using the GrCh37 reference genome sequence as input, averaging the two 128 base pair bins around each side of the CpGs (4 bins total) and saving outputs for all 5,313 epigenomic tracks available for prediction. The DNAm data was aggregated within each cell-type of interest, with most models averaging DNAm across three individuals. We then trained a logistic elastic net regression model for each cell-type data available:

$$\log \frac{p_c}{1-p_c} = \beta_0 + \sum_{i=1}^{5313} \beta_i X_i$$

Where p_c is the probability of DNAm specific to a given cell-type model, interpreted as the average bulk DNAm level of a given CpG; β_i are the coefficients of the model for each of the 5,313 Enformer output features; and X_i are the 5,313 reference Enformer output features. We minimize the objective function of this model using coordinate descent:

$$\min_{\beta}(-\frac{1}{N}[Y(\beta_0 + \beta X) - \log{(1 + e^{\beta_0 + \beta X}}] + \lambda[\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2])$$

Where N is the number of training data points, Y is the vector of observed DNAm training data, α is the elastic net penalty controlling the strength of the $L1$ and $L2$ regularization terms, and λ is the hyperparameter controlling the overall strength of regularization. Training was performed using the `glmnet` package in R [22, 23] with an 80:20 training/test split between all allergy CpGs, $\alpha = 0.5$ to reduce the dimensionality of the Enformer predictors, and 5-fold cross-validation and model performance evaluated using AUC, with final model weights saved at $\lambda_{1\text{se}}$. Model validation across all 80 cell types was performed at 1,000 randomly chosen CpGs, where we generated reference Enformer outputs and predicted using each of the models.

0.1 Validation of methENPACT across individuals controlling for ancestry

We obtained whole-genome genotypes and blood monocyte whole-genome bisulfite sequencing data in 35 individuals: 14 of African and 21 of European ancestry [11]. We confirmed that individuals were unrelated using `Plink` [24] to generate the genetic relatedness matrix within African and European ancestry groups. To assess model performance, we selected the top 1,000 CpGs that were a known allergy meQTL [21] by p value and generated Enformer outputs using individual phased genotypes around these sites. We saved the two bins flanking each side of the site and averaged values across both predicted haplotypes, then used these track values to predict DNAm in all 35 individuals. To validate observed across-individual correlations, we permuted sample labels of the predicted values randomly, as well as preserving ancestry groups, 1,000 times and correlated with observed data to generate an empirical null distribution, which we then compared with the observed correlations. We tested for ancestry effects across individuals by linearly modeling observed and predicted DNAm both with and without ancestry values:

$$\text{With ancestry covariate: } M_{\text{obs},i} = \beta_{0,i} + \beta_{1,i} M_{\text{pred},i} + \beta_{2,i} A_i$$
$$\text{Without ancestry covariate: } M_{\text{obs},i} = \beta_{0,i} + \beta_{1,i} M_{\text{pred},i}$$

Where M_{obs} and M_{pred} are vectors of observed and predicted DNAm levels across the 35 individuals, respectively; β denotes the coefficient of the covariates; and A is the ancestry encoding $[0 : \text{European}, 1 : \text{African}]$ for the individuals; for all $i \in [1, 1000]$ denoting each test site. We then assessed whether the ancestry models had significant β_1 and β_2 coefficients compared to the without ancestry models with only β_1.

0.2 Validation of methENPACT with MWAS

To validate our model against the existing model we predicted in a dataset independent of either model's training data: the publicly available phased genotype data from 99 CEU individuals available from the 1000 Genomes phase 3 release in VCF format [15]. We downloaded the Understanding Society MWAS model [14] and used the MetaXcan pipeline [18] to predict DNAm at 203,000 heritable CpGs. We then selected 1,000 random test sites that were both predictable by MWAS and known meQTLs to generate Enformer outputs and methENPACT predictions using the blood monocyte model at those sites.

0.3 Testing for genetically-regulated DNAm associations with downstream phenotypes

We downloaded LCL gene expression levels from GEUVADIS across 462 individuals and 18,959 genes [16]. For a subset of 435 individuals we had genotype information for, we generated individualized Enformer predictions at the gene TSS sites of 19,689 protein coding genes retrieved using the `biomaRt` package in R [25] and averaged the four central bins of each prediction. We subsetted the genes such that there was at least one CpG within the upstream 256 bp 2-bin window, and predicted DNAm using the methENPACT blood B cell model leaving us with predictions for 19,162 genes. After intersection with the GEUVADIS dataset we were left with 18,054 genes between the gene expression and predicted DNAm data with which to calculate correlations. A subset of 157 of these individuals had LCL growth phenotype available through HapMap [17], which we used to test for linear association with DNAm at each gene TSS using the following model:

$$P_i = \beta_{0,j} + \beta_{1,j} M_{\text{pred},j}$$

Where P is the vector of LCL growth, M_{pred} is the vector of predicted DNAm at the gene TSS, β_0 is the intercept, and β_1 is the test coefficient for association; for all genes $j \in [1, 18054]$. P values of β_1 were adjusted for multiple testing using FDR. We performed an overrepresentation analysis for the Gene Ontology genesets with the genes that had β_1 coefficient significance $p < 0.05$ using the `gprofiler2` package in R [26], with p values adjusted using FDR.

Data availability

The methENPACT training pipeline is available on Github: https://github.com/hakyimlab/methylation-prediction. Trained models from across the 80 atlas cell-types are made available through a Google Colab module, which allows for inputting individual genotypes at a given CpG site and prediction with a chosen model. The list of genes tested for enrichment are provided.

References

1. Abdellaoui, A., Yengo, L., Verweij, K. J. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110,** 179–194. ISSN: 0002-9297. http://dx.doi.org/10.1016/j.ajhg.2022.12.011 (Feb. 2023).
2. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics and Chromatin* **8.** ISSN: 1756-8935. http://dx.doi.org/10.1186/s13072-015-0050-4 (Dec. 2015).
3. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38,** 23–38. ISSN: 1740-634X. http://dx.doi.org/10.1038/npp.2012.112 (July 2012).
4. Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6,** 597–610. ISSN: 1471-0064. http://dx.doi.org/10.1038/nrg1655 (Aug. 2005).
5. Villicaña, S. & Bell, J. T. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biology* **22.** ISSN: 1474-760X. http://dx.doi.org/10.1186/s13059-021-02347-6 (Apr. 2021).
6. Kader, F. & Ghai, M. DNA methylation-based variation between human populations. *Molecular Genetics and Genomics* **292,** 5–35. ISSN: 1617-4623. http://dx.doi.org/10.1007/s00438-016-1264-2 (Nov. 2016).
7. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics* **55,** 112–122. ISSN: 1546-1718. http://dx.doi.org/10.1038/s41588-022-01248-z (Dec. 2022).
8. Amariuta, T. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *The American Journal of Human Genetics* **104,** 879–895. ISSN: 0002-9297. http://dx.doi.org/10.1016/j.ajhg.2019.03.012 (May 2019).
9. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18,** 1196–1203. ISSN: 1548-7105. http://dx.doi.org/10.1038/s41592-021-01252-x (Oct. 2021).
10. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613,** 355–364. ISSN: 1476-4687. http://dx.doi.org/10.1038/s41586-022-05580-6 (Jan. 2023).
11. Aracena, K. A. *et al.* Epigenetic variation impacts individual differences in the transcriptional response to influenza infection. *Nature Genetics* **56,** 408–419. ISSN: 1546-1718. http://dx.doi.org/10.1038/s41588-024-01668-z (Feb. 2024).
12. Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics* **55,** 2056–2059. ISSN: 1546-1718. http://dx.doi.org/10.1038/s41588-023-01574-w (Nov. 2023).
13. Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics* **55,** 2060–2064. ISSN: 1546-1718. http://dx.doi.org/10.1038/s41588-023-01524-6 (Nov. 2023).
14. Fryett, J. J., Morris, A. P. & Cordell, H. J. 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. ISSN: 1098-2272. http://dx.doi.org/10.1002/gepi.22496 (Aug. 2022).
15. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature15393 (Sept. 2015).
16. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature12531 (Sept. 2013).
17. Im, H. K. *et al.* Mixed Effects Modeling of Proliferation Rates in Cell-Based Models: Consequence for Pharmacogenomics and Cancer. *PLoS Genetics* **8** (ed Akey, J. M.) e1002525. ISSN: 1553-7404. http://dx.doi.org/10.1371/journal.pgen.1002525 (Feb. 2012).

18. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9.** ISSN: 2041-1723. `http://dx.doi.org/10.1038/s41467-018-03621-1` (May 2018).

19. Kurmi, K. & Haigis, M. C. Nitrogen Metabolism in Cancer and Immunity. *Trends in Cell Biology* **30,** 408–424. ISSN: 0962-8924. `http://dx.doi.org/10.1016/j.tcb.2020.02.005` (May 2020).

20. Diehl, F. F. *et al.* Nucleotide imbalance decouples cell growth from cell proliferation. *Nature Cell Biology* **24,** 1252–1264. ISSN: 1476-4679. `http://dx.doi.org/10.1038/s41556-022-00965-1` (Aug. 2022).

21. Morin, A. *et al.* A functional genomics pipeline to identify high-value asthma and allergy CpGs in the human methylome. *Journal of Allergy and Clinical Immunology* **151,** 1609–1621. ISSN: 0091-6749. `http://dx.doi.org/10.1016/j.jaci.2022.12.828` (June 2023).

22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33.** ISSN: 1548-7660. `http://dx.doi.org/10.18637/jss.v033.i01` (2010).

23. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software* **106.** ISSN: 1548-7660. `http://dx.doi.org/10.18637/jss.v106.i01` (2023).

24. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81,** 559–575. ISSN: 0002-9297. `http://dx.doi.org/10.1086/519795` (Sept. 2007).

25. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4,** 1184–1191. ISSN: 1750-2799. `http://dx.doi.org/10.1038/nprot.2009.97` (July 2009).

26. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9,** 709. ISSN: 2046-1402. `http://dx.doi.org/10.12688/f1000research.24956.2` (Nov. 2020).