Genomics, Proteomics & Bioinformatics, 2024, **22(3)**, qzae044
https://doi.org/10.1093/gpbjnl/qzae044
Advance access publication: 10 June 2024
Method

Genomics, Proteomics & Bioinformatics
OXFORD

MethylGenotyper: Accurate Estimation of SNP Genotypes and Genetic Relatedness from DNA Methylation Data

Yi Jiang [1,2,#], **Minghan Qu** [1,2,#], **Minghui Jiang** [1,2], **Xuan Jiang** [1,2], **Shane Fernandez** [3,4], **Tenielle Porter** [3,4,5], **Simon M. Laws** [3,4,5], **Colin L. Masters** [6], **Huan Guo** [1,7], **Shanshan Cheng** [1,2,*], **Chaolong Wang** [1,2,*]

[1] Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[2] Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[3] Centre for Precision Health, Edith Cowan University, Perth, WA 6027, Australia
[4] Collaborative Genomics and Translation Group, School of Medical and Health Sciences, Edith Cowan University, Perth, WA 6027, Australia
[5] Curtin Medical School, Bentley, WA 6102, Australia
[6] The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Melbourne, VIC 3052, Australia
[7] Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

*Corresponding authors: chaolong@hust.edu.cn (Wang C), sscheng@hust.edu.cn (Cheng S).

#Equal contribution.

Handling Editor: Yajie Zhao

Abstract

Epigenome-wide association studies (EWAS) are susceptible to widespread confounding caused by population structure and genetic relatedness. Nevertheless, kinship estimation is challenging in EWAS without genotyping data. Here, we proposed MethylGenotyper, a method that for the first time enables accurate genotyping at thousands of single nucleotide polymorphisms (SNPs) directly from commercial DNA methylation microarrays. We modeled the intensities of methylation probes near SNPs with a mixture of three beta distributions corresponding to different genotypes and estimated parameters with an expectation-maximization algorithm. We conducted extensive simulations to demonstrate the performance of the method. When applying MethylGenotyper to the Infinium EPIC array data of 4662 Chinese samples, we obtained genotypes at 4319 SNPs with a concordance rate of 98.26%, enabling the identification of 255 pairs of close relatedness. Furthermore, we showed that MethylGenotyper allows for the estimation of both population structure and cryptic relatedness among 702 Australians of diverse ancestry. We also implemented MethylGenotyper in a publicly available R package (https://github.com/Yi-Jiang/MethylGenotyper) to facilitate future large-scale EWAS.

Key words: DNA methylation; Genotype calling; Genetic relatedness; Population structure; Epigenome-wide association study.

Introduction

DNA methylation (DNAm), which involves transferring a methyl group on to the C5 position of the cytosine, is an important mechanism of gene regulation and is dynamically variable in response to environmental changes. It has become the most widely studied type of epigenetic modifications owing to the development of high-throughput DNAm microarrays. Specifically, the Infinium HumanMethylation450 (450K) and HumanMethylationEPIC (EPIC) arrays can simultaneously assay DNAm levels at hundreds of thousands of Cytosine-phosphate-Guanine (CpG) sites across the genome. Using these arrays, epigenome-wide association studies (EWAS) have identified numerous CpG sites associated with complex diseases or environmental exposures, leading to a better understanding of disease etiology [1–3]. Similar to those observed in genome-wide association studies (GWAS), population structure and cryptic relatedness can lead to spurious association signals in EWAS. DNAm levels may differ between populations due to the impacts of distinct genetic and environmental factors [4]. In contrast, related samples are likely to share similar environmental exposures and thus similar DNAm levels [1,5]. Despite the potential confounding effect, cryptic relatedness is often overlooked in EWAS due to the lack of methods to infer genetic relatedness based on DNAm data. Even in cohorts where both DNAm and GWAS data have been generated, the sample overlap between the two datasets is unlikely perfect due to their separate quality control (QC) procedures or other logistic factors. Therefore, methods to infer genetic relatedness directly from DNAm data will be very useful to facilitate large-scale EWAS.

With sufficient single nucleotide polymorphism (SNP) genotypes, we can easily estimate genetic relatedness in samples with or without population structure using existing tools [6–10]. While the Infinium methylation arrays are designed to incorporate SNP probes to facilitate the identification of sample swapping [11,12], the number of SNPs is too small to obtain accurate kinship estimates (*e.g.*, EPIC array consists of 59 SNP probes, including 6 on the X chromosome). On the other hand, tens of thousands of CpGs adjacent to common SNPs [*i.e.*, minor allele frequency (MAF) > 0.01] are often discarded by standard QC, because nearby SNPs can introduce mismatches to the probe sequence and thus interfere with the measured methylation intensity [13–15]. It has been demonstrated

Received: 8 March 2024; **Revised:** 26 May 2024; **Accepted:** 6 June 2024.

that the measured methylation intensities at these probes frequently show multi-modal distributions depending on the SNP genotypes, especially when common SNPs are present at the extension base [16–18]. Thus, we speculate that it is possible to infer SNP genotypes, as well as subsequently population structure and genetic relatedness, based on methylation intensity data.

By design, there are two types of Infinium methylation probes. Type I probes use two beads at each locus to measure the methylated and unmethylated signals separately. Fluorescent colors are determined by the nucleotide at the extension base (*i.e.*, red for A and T alleles, green for G and C alleles). Thus, SNPs (except for A/T and G/C SNPs) at the extension base will cause color channel switching (CCS). Genotypes for these SNPs can be accurately determined by comparing signal intensity from different color channels [13]. Type II probes use a single bead at each locus, with the extension base occurring at the target CpG. After bisulfite conversion, the red-labeled A and green-labeled G nucleotides will bind to the unmethylated and methylated alleles, respectively. In the presence of a SNP at the target CpG, a red color will be detected if the target C allele is mutated to A or T despite no effect of bisulfite conversion. Conversely, a green color will remain detected if C is mutated to G. It is much more challenging to infer genotypes for Type II probes than for Type I probes, because both methylation and mutation can affect the fluorescent color of the single bead at Type II probes. To the best of our knowledge, there are no existing methods to estimate genotypes for SNPs adjacent to Type II probes, although Infinium methylation arrays consist of a significantly greater number of Type II probes than Type I probes [14].

In this study, we developed a novel method, MethylGenotyper, to perform genotype calling based on DNAm data for SNP probes, Type I probes, and Type II probes (**Figure 1**). For each type of probes, we first converted the methylation intensity signals to the ratio of alternative allele intensity (RAI), initially proposed by Zhou et al. [13] for Type I probes targeting CCS SNPs. RAI is expected to follow a three-modal distribution with peaks near 0, 0.5, and 1, corresponding to three genotypes, respectively. We modeled RAI for each type of probes with a mixture of three beta distributions and one uniform distribution, and employed an expectation-maximization (EM) algorithm to obtain the maximum likelihood estimates (MLEs) of model parameters and genotype probabilities. The performance of the method in parameter estimation and genotype calling was evaluated by extensive simulations with different sample sizes and numbers of SNPs. Subsequently, we applied MethylGenotyper to two empirical datasets with both DNAm data (EPIC array) and GWAS data: the Dongfeng–Tongji (DFTJ) cohort [19], consisting of 4662 Chinese samples, and the Australian Imaging, Biomarker and Lifestyle (AIBL) study [20,21], consisting of 702 samples with diverse ancestry. In both datasets, we demonstrated that MethylGenotyper can infer high-quality genotypes at over 4000 SNPs, enabling accurate estimation of individual ancestry and pairwise relatedness. We also implemented MethylGenotyper into a publicly available R package (https://github.com/Yi-Jiang/MethylGenotyper) to facilitate future large-scale EWAS.

Method

DNAm and GWAS data from the DFTJ cohort

The DFTJ cohort is a prospective cohort of retired workers from the Dongfeng Motor Corporation in Shiyan, Hubei Province, China [19]. A total of 38,295 participants were enrolled in 2013, and leukocytes from 5200 samples were selected for DNAm profiling using the Infinium HumanMethylationEPIC v1.0 BeadChips (Illumina, San Diego, CA) in two batches (Figure S1). The samples with low detection quality (detection $P > 0.01$ in over 1% probes), mismatched methylation-inferred and self-reported sexes, discordant genotypes at > 10 SNP probes compared with GWAS data, and those lacking GWAS data were excluded, resulting in 4662 samples. These 4662 samples correspond to 4542 unique individuals, including 114 with duplicate measurements and three with triplicate measurements. The raw intensity data (IDAT) files were processed by background correction with the noob method [22,23] and dye-bias correction with the regression on logarithm of internal control probes (RELIC) method [24]. Probes on the sex chromosomes were excluded.

The GWAS data of 33,114 samples in the DFTJ cohort were assayed using the Infinium OmniZhonghua-8 v1.4 arrays (Illumina). After excluding low-quality samples (> 0.05 discordance rate at duplicate sites, call rate < 0.9, inbreeding coefficient < −0.1 or > 0.3 based on autosomal SNPs or < −0.2 based on the X chromosome), duplicated samples, and sex-mismatched samples, a total of 31,155 individuals passed QC. The SNPs with call rate < 0.95, minor allele count (MAC) < 3, or Hardy–Weinberg equilibrium (HWE) $P < 1 \times 10^{-6}$ were excluded, leaving 775,059 autosomal SNPs and 24,134 X chromosomal SNPs. Then, the GWAS data were phased and imputed by Eagle (v2.4.1) [25] and Minimac4 [26], with a whole-genome sequenced reference panel consisting of 3931 East Asian samples from the 1000 Genomes Project (1KGP) [27] and the SG10K project [28]. After excluding 20,640 variants with > 0.2 difference in allele frequency (AF) compared to East Asians in 1KGP, 14,495,888 variants with imputation $R^2 \geq 0.3$ and MAF $\geq$ 0.001 were kept for downstream analyses.

DNAm and GWAS data from the AIBL study

The AIBL study (https://aibl.org.au) is a consortium between Austin Health, Commonwealth Scientific and Industrial Research Organisation, Edith Cowan University, the Florey Institute (The University of Melbourne), and the National Ageing Research Institute in Australia, aiming to improve the understanding of Alzheimer's disease [20,21]. We downloaded DNAm array data (EPIC array) of 726 samples from the Gene Expression Omnibus repository (GEO: GSE153712) [29]. We downloaded IDAT files and processed the data with the noob method [22,23] and the RELIC method [24] for background correction and dye-bias correction, respectively. A total of 716 samples in the AIBL cohort were genotyped using the Infinium OmniExpressHumanExome+ v1.0 arrays (Illumina) [30,31], including 702 samples that overlapped with the samples having DNAm data.

Association between DNAm intensity and nearby SNPs

The association between DNAm β-values and the genotypes of nearby SNPs was assessed based on the DFTJ data. We focused on biallelic SNPs with MAF > 0.01 in the East Asian samples of 1KGP [27] and excluded probes with multiple SNPs within 5 bp from the 3′ end of the probe. For each probe, the β-values were pre-adjusted by regressing out sex, age, body mass index (BMI), smoking status, sample plates, and six immune cell type proportions estimated by Bigmelon [32]. The squared Pearson correlation (R^2) between the residualized β-values and the

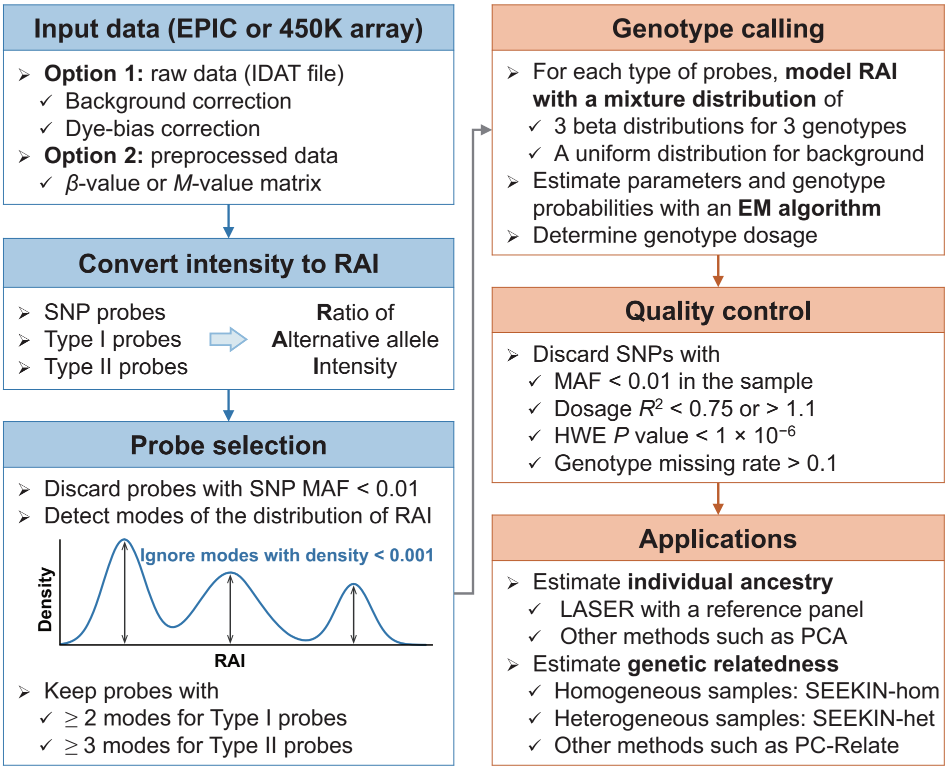
Input data (EPIC or 450K array)
➢ Option 1: raw data (IDAT file)
✓ Background correction
✓ Dye-bias correction
➢ Option 2: preprocessed data
✓ β-value or M-value matrix
Convert intensity to RAI
➢ SNP probes
➢ Type I probes
➢ Type II probes
Ratio of Alternative allele Intensity
Probe selection
➢ Discard probes with SNP MAF < 0.01
➢ Detect modes of the distribution of RAI
Ignore modes with density < 0.001
Density
RAI
➢ Keep probes with
✓ ≥ 2 modes for Type I probes
✓ ≥ 3 modes for Type II probes
Genotype calling
➢ For each type of probes, model RAI with a mixture distribution of
✓ 3 beta distributions for 3 genotypes
✓ A uniform distribution for background
➢ Estimate parameters and genotype probabilities with an EM algorithm
➢ Determine genotype dosage
Quality control
➢ Discard SNPs with
✓ MAF < 0.01 in the sample
✓ Dosage R² < 0.75 or > 1.1
✓ HWE P value < 1 × 10⁻⁶
✓ Genotype missing rate > 0.1
Applications
➢ Estimate individual ancestry
✓ LASER with a reference panel
✓ Other methods such as PCA
➢ Estimate genetic relatedness
✓ Homogeneous samples: SEEKIN-hom
✓ Heterogeneous samples: SEEKIN-het
✓ Other methods such as PC-Relate

Figure 1 The workflow of MethylGenotyper

MethylGenotyper takes either raw intensity data (recommended) or preprocessed data (only works for SNP probes and Type II probes) to calculate the RAI of each probe. Probes with MAF $\geq$ 0.01 in the corresponding population in 1KGP are kept, and the RAI of each probe possessing two or three modes is required. An EM algorithm is developed to fit the RAIs of each type of probes with a mixture of three beta distributions and one uniform distribution. The three beta distributions correspond to three genotypes and their weights are probe-specific based on the assumption of HWE. The uniform distribution represents background noise with a constant weight across the same type of probes. RAI, ratio of alternative allele intensity; MAF, minor allele frequency; 1KGP, the 1000 Genomes Project; EM, expectation-maximization; HWE, Hardy–Weinberg equilibrium; PCA, principal component analysis; SNP, single nucleotide polymorphism; IDAT, intensity data; LASER, locating ancestry from sequencing reads.

genotype dosages of the nearby SNP was computed as a function of the SNP position relative to the probe.

Calculation of RAI

We considered three types of probes for SNP genotyping: (1) SNP probes by design, (2) Type I probes targeting CCS SNPs introduced by Zhou et al. [13], and (3) Type II probes with a SNP at the extension base. RAI is defined for each type of probes separately.

For SNP probes, the reference and alternative alleles are targeted by different probes and the RAI can be calculated as:

$$\mathrm{RAI} = \frac{S\left(p^{ALT}\right)}{S\left(p^{REF}\right) + S\left(p^{ALT}\right)} \tag{1}$$

where $S\left(p^{REF}\right)$ and $S\left(p^{ALT}\right)$ denote probe signals supporting the reference allele and the alternative allele, respectively.

For Type I probes with CCS SNPs, we followed Zhou et al. [13] to calculate RAI as:

$$\mathrm{RAI} = \frac{S^{oob}\left(p^{M}\right) + S^{oob}\left(p^{U}\right)}{S^{oob}\left(p^{M}\right) + S^{oob}\left(p^{U}\right) + S^{ib}\left(p^{M}\right) + S^{ib}\left(p^{U}\right)} \tag{2}$$

where p^M and p^U denote the proportions of methylated and unmethylated probes, respectively, S^{oob} represents the out-of-band signal supporting the alternative allele, and S^{ib} represents the in-band signal supporting the reference allele.

For Type II probes, the extension base targets C of a CpG. Without mutation at the target site, a red color signal will be detected after bisulfite treatment when there is no methylation, while a green color signal will be detected when C is methylated. When C is mutated to A or T, a red color signal will always be detected; when C is mutated to G, a green color signal will always be detected. Thus, we have

$$\beta = \frac{S^{Grn}}{S^{Grn} + S^{Red}} = \begin{cases} (1 - \mathrm{RAI}) \times p^{M} & \text{for C/A or C/T SNPs} \\ (1 - \mathrm{RAI}) \times p^{M} + \mathrm{RAI} & \text{for C/G SNPs} \end{cases} \tag{3}$$

where S^{Red} and S^{Grn} represent the red and green color intensities, respectively, p^M represents the proportion of methylated C alleles, and β is the standard beta value calculated as the proportion of green color intensity. RAI roughly corresponds to 0, 0.5, and 1 for reference homozygotes, heterozygotes, and alternative homozygotes, respectively. Therefore, we expect β to follow a three-modal distribution with modes near (0, $0.5p^M$, and p^M) for C/A or C/T SNPs and (p^M, $0.5p^M + 0.5$, and 1) for C/G SNPs, under the assumption that p^M is stable across samples for each CpG. In practice, we used the "multimode" method to check the distribution of β-values for each CpG, including the number of modes, mode locations, and the density of each mode [33]. Probes that exhibited at least two modes with density height > 0.001 (bandwidth = 0.04) were retained. For probes with more than three modes detected, we removed the lowest modes until only three modes remained. We defined l_{het} as the location of the central mode of β (or the mode closest to 0.5 if only two modes are detected), which corresponds to heterozygotes. We then proposed to estimate p^M for each CpG as:

$$\widehat{p^M} = \begin{cases} \min(2l_{\text{het}}, 1) & \text{for C/A or C/T SNPs} \\ \max(2l_{\text{het}} - 1,\ 0) & \text{for C/G SNPs} \end{cases} \tag{4}$$

Finally, RAI was calculated as:

$$\text{RAI} = \begin{cases} 1 - \beta/\widehat{p^M} & \text{for C/A or C/T SNPs} \\ (\beta - \widehat{p^M})/(1 - \widehat{p^M}) & \text{for C/G SNPs} \end{cases} \tag{5}$$

We truncated RAI values at 0.01 or 0.99 for those outside the range of 0.01–0.99.

For both Type I and Type II probes, an additional filtering step was applied based on the distribution of RAI values, and only Type I probes with at least two modes and Type II probes with at least three modes were retained. A more stringent threshold was applied to Type II probes, because the SNPs for which we call genotypes correspond directly to the methylation target sites, which are more susceptible to the influence of methylation β-values and potential confounding.

Modeling the distribution of RAI

For each type of probes, we coded the RAI values by an $m \times n$ matrix $\boldsymbol{X}$, where m and n indicate the numbers of probes and samples, respectively. We assumed that $\boldsymbol{X}$ follows a mixture of three beta distributions corresponding to three genotypes, and a uniform distribution represents the background noise [12]:

$$X_{ij} \sim (1-\lambda)\sum_{k=0}^{2} w_{ik}\text{Beta}(\alpha_k, \beta_k) + \lambda\text{U}(0,1) \tag{6}$$

$$w_{ik} = \binom{2}{k}\phi_i^k(1-\phi_i)^{2-k} \tag{7}$$

where X_{ij} represents the RAI value at probe i of sample j; k represents the genotype dosage, coded as 0, 1, and 2 for reference homozygotes, heterozygotes, and alternative homozygotes, respectively; $\text{Beta}(\alpha_k, \beta_k)$ represents the beta distribution with parameters α_k and β_k; $\text{U}(0,1)$ represents the standard uniform distribution; λ represents the probability that RAI comes from background noise; $(1-\lambda)\ w_{ik}$ represents the probability that RAI comes from genotype k with weight w_{ik} specified by the Hardy–Weinberg proportions at probe i; and ϕ_i represents the AF at probe i.

The EM algorithm

Let $B_{ijk} = \text{Beta}\left(X_{ij};\ \alpha_k, \beta_k\right)$ denote the probability density of $\text{Beta}(\alpha_k, \beta_k)$ at X_{ij}. Assuming X_{ij} is independent, the log-likelihood function can be written as:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \lambda) = \sum_{i=1}^{m}\sum_{j=1}^{n}\log\left(\lambda + (1-\lambda)\sum_{k=0}^{2} w_{ik}B_{ijk}\right) \tag{8}$$

We developed an EM algorithm to estimate the parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, λ), of which the initial values were set as $\alpha_0 = 5,\ \beta_0 = 60,\ \alpha_1 = \beta_1 = 30,\ \alpha_2 = 60,\ \beta_2 = 5,\ \phi_i = 0.2$ for any i, and $\lambda = 0.01$.

In the E-step, we calculated the probability of X_{ij} from $\text{U}(0,1)$ as:

$$\widehat{\Lambda}_{ij} = \frac{\lambda}{\lambda + (1-\lambda)\sum_{k=0}^{2} w_{ik}B_{ijk}} \tag{9}$$

and the probability of X_{ij} from $\text{Beta}(\alpha_k, \beta_k)$ as:

$$\widehat{P}_{ijk} = \frac{(1-\lambda)w_{ik}B_{ijk}}{\lambda + (1-\lambda)\sum_{k=0}^{2} w_{ik}B_{ijk}} = \frac{w_{ik}B_{ijk}}{\sum_{k=0}^{2} w_{ik}B_{ijk}}\left(1 - \widehat{\Lambda}_{ij}\right) \tag{10}$$

In the M-step, we updated the parameters with their moment estimators:

$$\widehat{\lambda} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\Lambda}_{ij} \tag{11}$$

$$\widehat{\phi}_i = \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=0}^{2}\frac{k\widehat{P}_{ijk}}{1 - \widehat{\Lambda}_{ij}} \tag{12}$$

$$\widehat{\alpha}_k = \widehat{M}_k\left(\frac{\widehat{M}_k\left(1 - \widehat{M}_k\right)}{\widehat{V}_k} - 1\right) \tag{13}$$

$$\widehat{\beta}_k = (1 - \widehat{M}_k)\left(\frac{\widehat{M}_k\left(1 - \widehat{M}_k\right)}{\widehat{V}_k} - 1\right) \tag{14}$$

where

$$\widehat{M}_k = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{P}_{ijk}X_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{P}_{ijk}} \tag{15}$$

$$\widehat{V}_k = \frac{mn}{mn-1}\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{P}_{ijk}X_{ij}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{P}_{ijk}} - \widehat{M}_k^2\right) \tag{16}$$

We iterated the E-step and M-step until the log-likelihood converged to its maximum, and thus yielded the maximum likelihood estimates of ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, λ).

Genotype calling and QC

With the estimated genotype probabilities $\widehat{P}_{ijk}$ and background probabilities $\widehat{\Lambda}_{ij}$, we set the genotypes with $\widehat{\Lambda}_{ij} > \max\limits_{0 \le k \le 2}\widehat{P}_{ijk}$ as missing and then updated each $\widehat{P}_{ijk}$ by dividing it with $(1 - \widehat{\Lambda}_{ij})$ to ensure $\sum_{k=0}^{2}\widehat{P}_{ijk} = 1$ for any probe i and sample j. For the other non-missing genotypes, we defined $\widehat{G}_{ij}$ as the most probable genotype and $\widehat{D}_{ij} = \sum_{k=0}^{2} k\widehat{P}_{ijk}$ as the genotype dosage. Following the methodology of Li et al. [34], we computed the AF $\widehat{q}_i$ and the dosage $\widehat{R}_i^2$ as follows:

$$\widehat{q}_i = \frac{\sum_{j=1}^{n} \widehat{D}_{ij}}{2n} \tag{17}$$

$$\widehat{R}_i^2 = \frac{\mathrm{Var}(\widehat{D}_{i\cdot})}{2\widehat{q}_i(1-\widehat{q}_i)} = \frac{\sum_{j=1}^{n} \widehat{D}_{ij}^2 - 4\widehat{q}_i^2 n}{2\widehat{q}_i(1-\widehat{q}_i)n} \tag{18}$$

where n is the sample size and $\mathrm{Var}(\widehat{D}_{i\cdot})$ is the variance of $\widehat{D}_{i\cdot}$ across samples. We calculated SNP-level missing rate as the proportion of genotypes being missing or with $\max_{0 \le k \le 2} \widehat{P}_{ijk} < 0.9$. To ensure the data quality, we excluded SNPs with MAF < 0.01, HWE P value $< 1 \times 10^{-6}$, $\widehat{R}_i^2 < 0.75$ or $\widehat{R}_i^2 > 1.1$, or missing rate > 0.1.

Simulation data

To examine the performance of our genotype calling algorithm, we conducted simulations with two parameter settings, mimicking the RAI distributions of Type I probes ($\alpha_0 = 3,\ \beta_0 = 35,\ \alpha_1 = \beta_1 = 20,\ \alpha_2 = 65,\ \beta_2 = 4,\ \lambda = 0.025$) and Type II probes ($\alpha_0 = 2,\ \beta_0 = 20,\ \alpha_1 = 35,\ \beta_1 = 40,\ \alpha_2 = 40,\ \beta_2 = 3,\ \lambda = 0.015$), respectively. Let m and n be the numbers of probes and samples, respectively. We chose four values of m (50, 100, 200, and 400 for Type I probes; 500, 1000, 2000, and 4000 for Type II probes) and six values of n (100, 200, 400, 800, 1600, and 3200) in different simulations. In each simulation, we first randomly drew AFs of m common SNPs (MAF > 0.05) from the 1KGP data, denoted as q_i for SNP i. We then simulated an $m \times n$ genotype matrix ($\boldsymbol{T}$) by drawing genotypes of SNP i from a binomial distribution with probability q_i. Next, we randomly set a small fraction (λ) of genotypes in $\boldsymbol{T}$ to missing. Finally, we simulated an $m \times n$ RAI matrix ($\boldsymbol{X}$) by drawing X_{ij} from $\mathrm{U}(0,1)$ if T_{ij} is missing and from $\mathrm{Beta}(\alpha_k, \beta_k)$ if $T_{ij} = k$, where $k = 0,\ 1,$ or 2. In total, we generated 48 (= 2 × 4 × 6) sets of simulations based on combinations of parameter settings, sample sizes, and numbers of probes. For each simulation, we repeated 20 times to evaluate the mean and standard error (SE) of the accuracy of our methods in estimating the parameters and genotypes.

Inference of population structure

We inferred population structure for the AIBL samples using the locating ancestry from sequencing reads (LASER) method [35,36]. Briefly, we first defined an ancestral space using the top 4 principal components (PCs) of principal component analysis (PCA) of the 1KGP samples, because the top 4 PCs of 1KGP could separate major continental groups [28]. We then used the trace program in LASER to project each study sample into the ancestral space based on genotypes from MethylGenotyper.

Estimation of kinship coefficient

For DNAm data, we estimated kinship coefficients by SEEKIN [6], which accounts for the uncertainty in the inferred genotypes through the dosage $\widehat{R}_i^2$ for each SNP. We used the SEEKIN-hom and SEEKIN-het estimators for samples from the DFTJ and AIBL cohorts, respectively. The SEEKIN-het estimator accounts for the diverse ancestry by introducing individual-specific AFs [6]. Briefly, given an ancestral space defined by the top 4 PCs of the 1KGP samples, we modeled genotypes with a linear function of the top 4 PCs:

$$G_{ij}^R \sim \beta_{i0} + \beta_{i1}v_{j1}^R + \beta_{i2}v_{j2}^R + \beta_{i3}v_{j3}^R + \beta_{i4}v_{j4}^R \tag{19}$$

where G_{ij}^R indicates the genotype at SNP i of individual j from the reference samples, $v_{j\cdot}^R$ indicates the PC coordinates of individual j, and $\beta_{i\cdot}$ are the regression coefficients for SNP i. Denoting $\widehat{v}_{j\cdot}$ as the projected PC coordinates of the j-th study sample, the individual-specific AF $\widehat{q}_{ij}$ can be estimated as:

$$\widehat{q}_{ij} = \frac{1}{2}\left(\widehat{\beta}_{i0} + \widehat{\beta}_{i1}\widehat{v}_{j1} + \widehat{\beta}_{i2}\widehat{v}_{j2} + \widehat{\beta}_{i3}\widehat{v}_{j3} + \widehat{\beta}_{i4}\widehat{v}_{j4}\right) \tag{20}$$

We truncated $\widehat{q}_{ij}$ at 0.001 and 0.999 for values outside the boundary.

For comparison, we estimated kinship coefficients using GWAS data of the DFTJ and AIBL cohorts. For the DFTJ cohort, we applied SEEKIN-hom [6] to GWAS data of 286,727 SNPs with MAF > 0.01 and linkage disequilibrium (LD) $r^2 < 0.5$. For the AIBL cohort, we applied PC-Relate [10] to GWAS data of 113,690 SNPs with MAF > 0.05 and LD $r^2 < 0.2$. Based on kinship coefficients estimated from GWAS data, we classified relatedness into duplicated, 1st-degree, 2nd-degree, 3rd-degree, and unrelated pairs according to the cutoffs described by Manichaikul and his colleagues [8]. In addition, we grouped 2nd-degree or closer relatedness (kinship coefficient $> 2^{-3.5}$) as the positive set and the rest as the negative set to compute the following statistics for DNAm-based kinship classification:

$$\mathrm{Precision} = \frac{\mathrm{True\ positive}}{\mathrm{True\ positive} + \mathrm{False\ positive}} \tag{21}$$

$$\mathrm{Recall} = \frac{\mathrm{True\ positive}}{\mathrm{True\ positive} + \mathrm{False\ negative}} \tag{22}$$

$$F_1 = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{23}$$

Results

Correlation between DNAm intensity and SNP genotypes

We first examined squared correlation between the measured DNAm β-values and genotypes (coded as 0, 1, and 2) of nearby SNPs in the DFTJ dataset (Figure S1) [19]. We focused on biallelic autosomal SNPs with MAF > 0.01 in East Asian samples of 1KGP, and excluded probes with multiple SNPs within 5 bp from the 3′ end of the probe. After regressing out age, sex, BMI, smoking status, sample plates, and six immune cell type proportions, the highest R^2 was observed for Type II probes with a SNP at the extension base (median $R^2 = 0.90$ at position −1) (Figure S2A), which was expected because Type II probes measured DNAm at the extension base. For Type I probes, the highest R^2 was observed at the 3′ end (median $R^2 = 0.48$ at position 0), where DNAm was measured. Importantly, the number of Type II probes with a SNP at the extension base ($n = 7619$) is the largest among all probe categories that we examined (Figure S2B). Considering both the R^2 and the number of probes, we chose to focus on Type II probes with a SNP at the extension base, in addition to the SNP probes and Type I probes targeting CCS SNPs [13].

Performance of MethylGenotyper in simulation data

We conducted two sets of simulations to examine the performance of MethylGenotyper based on the parameters of Type I probes and Type II probes obtained from real data (see next section). Our method fit the simulated RAI distributions perfectly for both probe types (Figure S3). The estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (parameters of three beta distributions), λ (relative intensity of the background noise), and $\boldsymbol{\phi}$ (AF), approached their simulated true values as the sample size and the number of SNPs increased (**Figure 2**, Figure S4). For Type II probes, the error rates of the parameter estimates began to stabilize as the sample size reached 800 (Figure 2A–G). For a dataset consisting of 3200 samples and 4000 SNPs, the estimation errors relative to the true values were 0.0046 [95% confidence interval (CI): 0.0036–0.0055) for $\boldsymbol{\alpha}$, 0.0041 (95% CI: 0.0033–0.0049) for $\boldsymbol{\beta}$, 0.086 (95% CI: 0.084–0.088) for λ, and 0.031 (95% CI: 0.031–0.031) for $\boldsymbol{\phi}$. The genotype concordance rates were ~ 98.4% for different numbers of samples and SNPs (Figure 2I). For Type I probes, the genotype concordance rates were slightly lower (~ 97.7%), possibly due to a higher level of background noise ($\lambda = 0.025$ for Type I probes *versus* $\lambda = 0.015$ for Type II probes) and a smaller number of simulated SNPs (Figure S4I).

Inferring genotypes using EPIC data from the DFTJ cohort

Following the probe selection procedure (Figure 1; see Method), we retained 53 SNP probes, 168 Type I probes, and 5050 Type II probes for genotype calling using the EPIC data of 4662 Chinese samples from the DFTJ cohort. The RAI distributions for each type of probes were fit well with parameter values shown in **Figure 3**. The relative intensities of the background noise λ were estimated to be 0.014, 0.024, and 0.017 for SNP probes, Type I probes, and Type II probes, respectively. After QC, we called genotypes at 4319 SNPs, including 53 from SNP probes, 111 from Type I probes, and

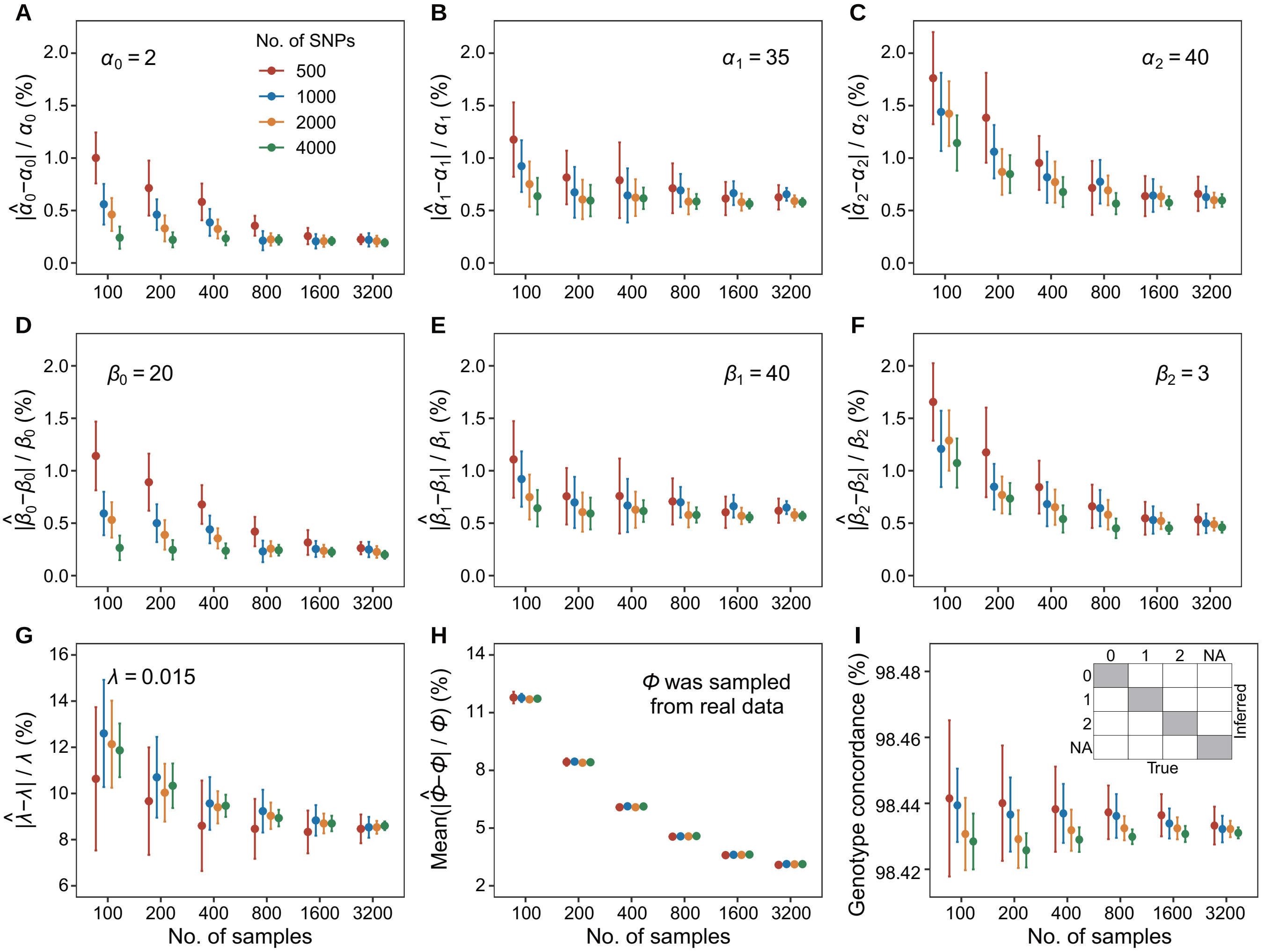

Figure 2 Performance of MethylGenotyper on simulation data mimicking the RAI distribution of Type II probes
The genotype matrices were simulated under the assumption of HWE, with AFs ($\boldsymbol{\phi}$) randomly sampled from the 1KGP data. Conditional on these genotype matrices, RAI matrices were generated from a mixture distribution with parameters matching characteristics of Type II probes in the DFTJ dataset. **A**.–**G**. Error rates of the estimated parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, λ). True parameter values were labeled in each panel. **H**. Mean error rates of the estimated AFs. **I**. Concordance of the inferred genotypes. As shown in the inserted panel, concordance was computed by dividing the number of genotypes in the shade areas by the total number of genotypes. In each panel, dots and vertical bars represent the means and 95% confidence intervals (±1.96 SE), calculated from 20 repeats of simulation. AF, allele frequency; DFTJ, Dongfeng–Tongji; SE, standard error.

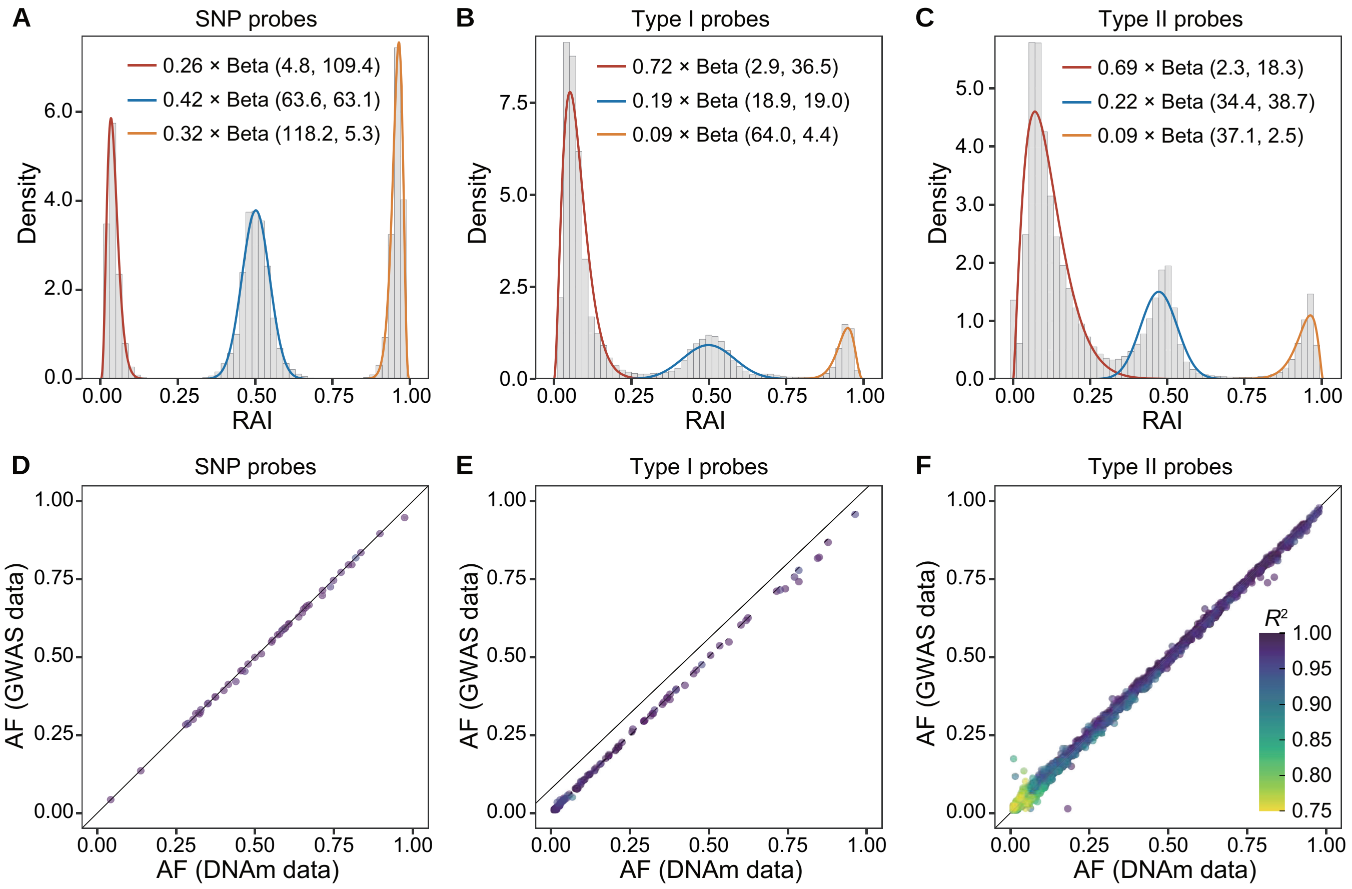

Figure 3 Performance of MethylGenotyper in the DFTJ dataset
A.–**C**. Fitted distributions of RAI for SNP probes (A), Type I probes (B), and Type II probes (C), respectively. Histograms show the distributions of RAI for all selected probes, and smooth lines indicate the fitted beta distributions, with weights averaged across probes. **D**.–**F**. Comparison of AFs derived from DNAm data with those from GWAS data for SNP probes (D), Type I probes (E), and Type II probes (F), respectively. Each point represents a SNP, colored by the estimated dosage R^2. Only SNPs passing QC were shown in the bottom panels. DNAm, DNA methylation; GWAS, genome-wide association study; QC, quality control.

Table 1 Accuracy of genotypes called by MethylGenotyper in the DFTJ dataset

Probe type	Number of SNPs	Genotype concordance	Heterozygote concordance
SNP probe	53 (53)	98.89%	98.53%
Type I probe	111 (109)	98.26%	96.93%
Type II probe	4155 (4092)	98.25%	96.59%
C-to-T	3875 (3818)	98.27%	96.62%
C-to-A	193 (191)	98.24%	96.37%
C-to-G	87 (83)	97.75%	96.05%
Overall	4319 (4254)	98.26%	96.64%

Note: SNPs with MAF < 0.01, HWE $P < 1 \times 10^{-6}$, $R^2 < 0.75$ or > 1.1, or missing rate > 0.1 were excluded. Concordance was evaluated by comparison to the array genotyping data of the same samples. Number of SNPs with array genotyping data were shown in the parentheses. SNP, single nucleotide polymorphism; DFTJ, Dongfeng–Tongji; MAF, minor allele frequency; HWE, Hardy–Weinberg equilibrium.

4155 from Type II probes, with high accuracy (Table 1). Compared to the imputed GWAS data, the overall genotype concordance was 98.26%, and the heterozygote concordance was 96.64%. It is noteworthy that the genotyping accuracy reported here may be underestimated, because erroneous genotypes could be present in the imputed GWAS data. For Type II probes, the C allele at the extension base might be mutated to T, A, or G alleles, of which C-to-T mutations accounted for 93.3% of all SNPs. We found that SNPs with C-to-T or C-to-A mutations exhibited similar genotype concordance rates at ∼ 98.27%, while C-to-G SNPs showed lower genotype concordance at 97.75%. Given the high genotyping accuracy of MethylGenotyper, it was not surprised that the estimated AFs were highly consistent with those based on the GWAS data for different probe types (Figure 3D–F).

Estimation of genetic relatedness in the DFTJ dataset

Based on GWAS data, we identified 123 duplicated pairs, 110 pairs of 1st-degree, 22 pairs of 2nd-degree, and 53 pairs of 3rd-degree relatedness among 4662 DNAm samples in the DFTJ cohort (**Figure 4A**). In contrast, based on genotypes inferred from 53 SNP probes and 111 Type I probes, it was difficult to distinguish related pairs from the huge number of unrelated pairs, with almost zero precision to identify 2nd-degree or closer relatedness (Figure 4B and C, Figure S5A and B). However, by incorporating genotypes inferred from 4155 Type II probes, the variance of kinship estimates was dramatically reduced, leading to a clear separation of different degrees of relatedness (Figure 4D, Figure S5C).

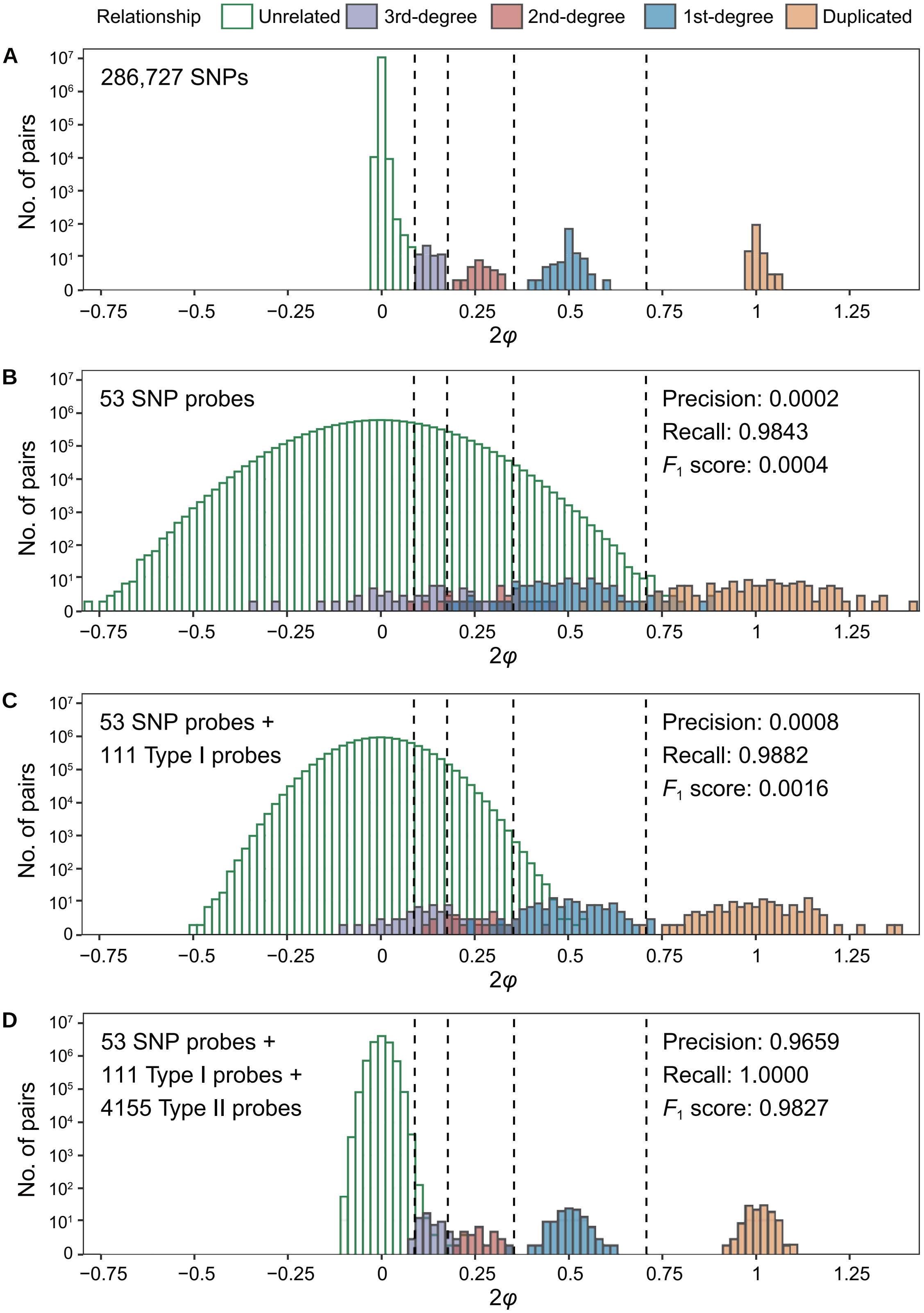

Figure 4 Estimation of genetic relatedness among samples in the DFTJ dataset

A. Kinship estimates based on 286,727 genome-wide SNPs (gold standard). **B**. Kinship estimates based on 53 SNP probes. **C**. Kinship estimates based on 53 SNP probes and 111 Type I probes. **D**. Kinship estimates based on 53 SNP probes, 111 Type I probes, and 4155 Type II probes. Relationship types were determined by the gold standard kinship estimates in (A), with the inference criteria indicated by vertical dashed lines. In (B–D), precision, recall, and F_1 score were calculated by comparison to the gold standard, treating 2nd-degree or closer relatedness as positive. "φ" represents kinship coefficient.

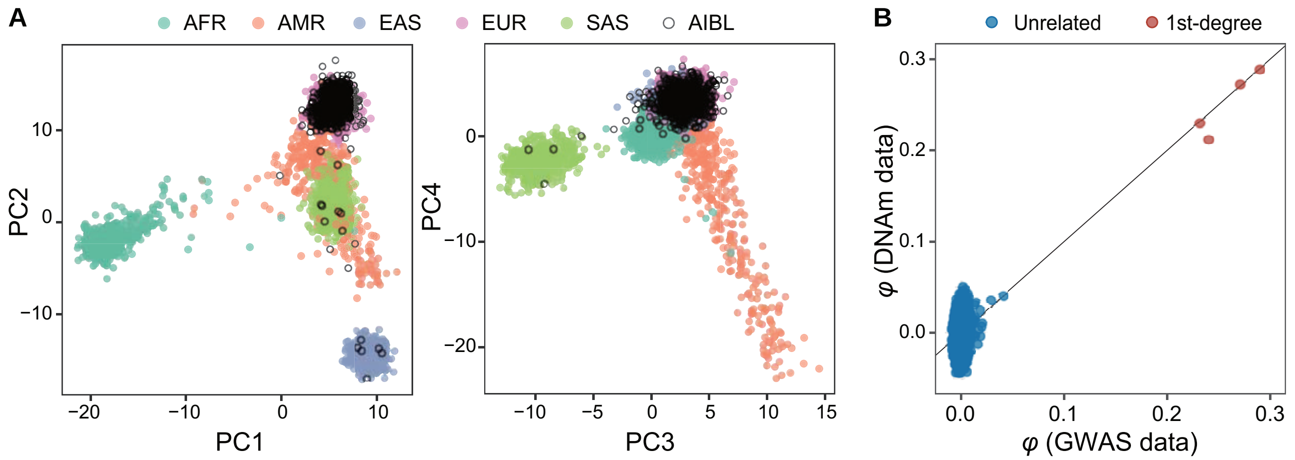

Figure 5 Performance of MethylGenotyper in the AIBL dataset

A. Inferred ancestry of 702 AIBL samples in the ancestral space generated by the top 4 PCs of the 1KGP samples. Analysis was based on 4217 SNPs called by MethylGenotyper. **B**. Comparison of kinship estimates based on genotypes called from DNAm data and those called from GWAS data. Relationship types were determined based on array genotyping data. "φ" represents kinship coefficient. AIBL, Australian Imaging, Biomarkers and Lifestyle; PC, principal component; AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian.

Compared to the benchmark established using GWAS data, our method achieved a precision of 0.9659 and a perfect recall rate of 1.0000 ($F_1 = 0.9827$) in identifying 2nd-degree or closer relatedness.

We also assessed whether our method could allow for accurate kinship estimation using the Infinium 450K array. By design, over 90% of the probes on the 450K array were included in the EPIC array [14]. Thus, we extracted DNAm data of these probes from the EPIC data of the DFTJ cohort and applied MethylGenotyper to call genotypes. A total of 2212 SNPs were identified with high-quality genotypes, including 53 from SNP probes, 104 from Type I probes, and 2055 from Type II probes. The performance in kinship estimation based on these SNPs was similar to that based on the full EPIC data, albeit with a slightly larger variance (Figure S6).

Inference of population structure and cryptic relatedness in the AIBL dataset

We further validated MethylGenotyper using data from the AIBL study [29], in which both DNAm data (EPIC array) and GWAS data were available for 702 Australian samples. Based on the DNAm data, we obtained high-quality SNP genotypes at 4217 probes, including 52 SNP probes, 135 Type I probes, and 4030 Type II probes. The AFs estimated from DNAm data were highly consistent with those from 1KGP European data, except for a small number of SNPs (Figure S7).

We first investigated the ancestral background of the AIBL samples using the LASER method with 1KGP data as the reference panel [35]. Surprisingly, while most samples were clustered with Europeans, a handful of the samples were clustered with East Asians, South Asians, or in between (**Figure 5A**). To account for the diverse ancestry background, we estimated kinship coefficients among AIBL samples using estimators for heterogeneous samples [6,10]. We found that the kinship estimates were highly consistent between those derived from DNAm data and from GWAS data (Figure 5B). Additionally, we identified four pairs of 1st-degree relatedness that were previously unknown. These results demonstrate the robustness of MethylGenotyper and its potential applications in the inference of population structure and cryptic relatedness among samples from diverse populations.

Discussion

Population structure and cryptic relatedness are major confounding factors in both GWAS and EWAS, where hundreds of thousands of sites are tested for phenotype association [1]. However, genetic data are often not available or incomplete in EWAS samples. Several studies have explored the potential to infer population structure directly from DNAm data, mostly based on PCA of CpG sites near known SNPs [37–39]. The methods developed in these studies, without properly modeling the relationship between SNP genotypes and DNAm signals, have limited resolution to infer population structure, because DNAm intensity can be affected by many other factors, including batch effects. Furthermore, to date, no study has attempted to infer genetic relatedness directly from DNAm data, even though close genetic relatedness often implies shared environmental exposures that can affect both DNAm and phenotypes. In this study, we developed a novel method, MethylGenotyper, to accurately infer genotypes at thousands of SNPs based on DNAm data that are frequently discarded by standard QC. Our results demonstrate that SNP genotypes inferred by our method allow for accurate inferences of both population structure and genetic relatedness, thus addressing a major confounding issue in EWAS.

While it has been noted that SNPs near CpG target sites can interfere with methylation intensity [16–18], few studies have extensively explored genotype calling at these SNPs, except for two studies [12,13]. Heiss and Just [12] developed ewastools to call genotypes specifically for tens of SNP probes incorporated into the Illumina 450K and EPIC arrays, and showed that these SNPs can be used to identify mislabeled or contaminated samples. Zhou et al. [13] proposed a method to infer genotypes at hundreds of SNPs that can cause CCS at Type I probes. Nevertheless, based on EPIC array data, we showed that SNP probes and Type I CCS probes together were not sufficient for accurate kinship estimation to separate closely related and unrelated pairs. In contrast, we expanded the number of genotyped SNPs by 25 times by incorporating thousands of Type II probes.

We processed SNP probes, Type I probes, and Type II probes separately but under a unified statistical framework. We first generalized the RAI statistic proposed by Zhou et al. [13] for Type I probes to all three types of probes. We then modeled RAI for each type of probes with a mixture of three beta distributions and one uniform distribution, similar to the model in ewastools [12], except that we introduced probe-specific weights based on AFs from external source to improve genotyping accuracy. With a sophisticated model and an EM algorithm, our method can infer genotypes with over 98% concordance rate for over 4000 SNPs from EPIC array, allowing for almost perfect identification of $\leq$ 2nd-degree relatedness in the DFTJ dataset. Notably, similar performance in kinship estimation can be achieved even when we used DNAm probes available on the Illumina 450K array, supporting wide applicability of MethylGenotyper to different methylation arrays.

Based on EPIC data from the AIBL cohort, we further illustrated that SNP genotypes inferred by MethylGenotyper can be used to infer population structure and close relatedness among samples with diverse ancestry. We used the LASER method [35] to estimate individual ancestry in a reference ancestral space of worldwide populations, and the SEEKIN-het estimator [6] for kinship estimation, accounting for individual-specific ancestry background. The analysis workflow incorporating LASER and SEEKIN methods has been implemented in the MethylGenotyper package to facilitate the research community. Furthermore, while unexplored in the present study, we expect that high-quality genotypes from over 4000 SNPs will be sufficient to identify fine-scale population structure within continental groups based on downsampling experiments in a previous study [40].

In addition to estimating population structure and genetic relatedness, the accurate genotypes called by MethylGenotyper can be used in many downstream analyses, including the estimation of inbreeding coefficient and the detection of sample contamination or sample swapping. Nevertheless, the utility of these genotypes in methylation quantitative trait locus (meQTL) mapping is limited due to the relatively small number of SNPs and the potential impacts of these SNPs on the methylation measurements of nearby CpGs.

The computational cost of our method increases linearly with the numbers of samples (n) and candidate probes (m), resulting in a complexity of $O(mn)$. Taking EPIC data of 1000 samples as an example, the raw data preprocessing (background and dye-bias correction) takes $\sim$ 17 min with 10 central processing units (CPUs), while genotype calling takes $\sim$ 13 min with 1 CPU. The test was run on a high-performance computing cluster with Intel Xeon CPUs (2.30 GHz).

In conclusion, we have developed MethylGenotyper to accurately infer genotypes at thousands of SNPs directly from DNAm microarray data. Our findings demonstrate that these SNP genotypes can be used to accurately estimate population structure and genetic relatedness, beyond simple tasks such as identifying mislabeled or contaminated samples. One limitation of MethylGenotyper is that we only focus on SNPs at the extension base of both Type I and Type II probes. Future studies might consider incorporating SNPs present at other nearby positions, which are also known to interfere with the measured DNAm intensity. Given the widespread confounding effects caused by population structure and genetic relatedness, we recommend the research community to incorporate MethylGenotyper into the standard analysis pipeline of EWAS to maximize statistical power and avoid spurious association signals.

Code availability

The R package of MethylGenotyper is publicly available at GitHub (https://github.com/Yi-Jiang/MethylGenotyper) and BioCode (https://ngdc.cncb.ac.cn/biocode/tool/BT007466).

Data availability

The DFTJ methylation data of candidate probes for MethylGenotyper have been deposited in the Open Archive for Miscellaneous Data [41] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (OMIX: OMIX006294), and are publicly accessible at https://ngdc.cncb.ac.cn/omix.

CRediT author statement

Yi Jiang: Formal analysis, Funding acquisition, Visualization, Software, Writing – original draft. **Minghan Qu:** Formal analysis, Visualization, Writing – original draft. **Minghui Jiang:** Formal analysis. **Xuan Jiang:** Formal analysis. **Shane Fernandez:** Validation. **Tenielle Porter:** Validation. **Simon M. Laws:** Funding acquisition, Validation. **Colin L. Masters:** Validation. **Huan Guo:** Data curation. **Shanshan Cheng:** Conceptualization, Data curation, Project administration, Supervision, Writing – review & editing. **Chaolong Wang:** Funding acquisition, Conceptualization, Data curation, Methodology, Project administration, Supervision, Writing – review & editing. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzae044).

Competing interests

The authors have declared no competing interests.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (Grant Nos. 82325044 and 82021005), the China Postdoctoral Science Foundation (Grant No. 2021M701318), the Natural Science Fund for Distinguished Young Scholars of Hubei Province, China (Grant No. 2022CFA046), and the Fundamental Research Funds for the Central Universities, China (Grant Nos. 2019kfyXJJS036 and 2023BR030 of HUST). Genetic and DNAm data in the AIBL study were funded by the National

Health and Medical Research Council in Australia (Grant Nos. GNT1161706 and GNT1151854). We thank all the investigators of the DFTJ cohort and the AIBL study for their contributions.

ORCID

0000-0002-1196-0280 (Yi Jiang)
0009-0003-3155-679X (Minghan Qu)
0000-0002-7888-2408 (Minghui Jiang)
0009-0001-1580-3186 (Xuan Jiang)
0000-0002-4881-245X (Shane Fernandez)
0000-0002-7887-6622 (Tenielle Porter)
0000-0002-4355-7082 (Simon M. Laws)
0000-0003-3072-7940 (Colin L. Masters)
0000-0002-7838-2585 (Huan Guo)
0000-0003-2676-7341 (Shanshan Cheng)
0000-0003-3945-1012 (Chaolong Wang)

References

[1] Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet 2011;12:529–41.

[2] Wei S, Tao J, Xu J, Chen X, Wang Z, Zhang N, et al. Ten years of EWAS. Adv Sci (Weinh) 2021;8:e2100727.

[3] Fraszczyk E, Spijkerman AMW, Zhang Y, Brandmaier S, Day FR, Zhou L, et al. Epigenome-wide association study of incident type 2 diabetes: a meta-analysis of five prospective European cohorts. Diabetologia 2022;65:763–76.

[4] Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol 2012;13:R8.

[5] Gross A, Tonjes A, Scholz M. On the impact of relatedness on SNP association analysis. BMC Genet 2017;18:104.

[6] Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet 2017; 13:e1007021.

[7] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559–75.

[8] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics 2010;26:2867–73.

[9] Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet 2012;91:122–38.

[10] Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. Am J Hum Genet 2016; 98:127–48.

[11] Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods 2014;11:1138–40.

[12] Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. Clin Epigenetics 2018;10:73.

[13] Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res 2017;45:e22.

[14] Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol 2016;17:208.

[15] McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genom Data 2016;9:22–4.

[16] Daca-Roszak P, Pfeifer A, Zebracka-Gala J, Rusinek D, Szybinska A, Jarzab B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies. BMC Genomics 2015;16:1003.

[17] LaBarre BA, Goncearenco A, Petrykowska HM, Jaratlerdsiri W, Bornman MSR, Hayes VM, et al. MethylToSNP: identifying SNPs in Illumina DNA methylation array data. Epigenetics Chromatin 2019;12:79.

[18] Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. "Gap hunting" to characterize clustered probe signals in Illumina methylation array data. Epigenetics Chromatin 2016;9:56.

[19] Wang F, Zhu J, Yao P, Li X, He M, Liu Y, et al. Cohort profile: the Dongfeng–Tongji cohort study of retired workers. Int J Epidemiol 2013;42:731–40.

[20] Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009;21:672–87.

[21] Fowler C, Rainey-Smith SR, Bird S, Bomke J, Bourgeat P, Brown BM, et al. Fifteen years of the Australian Imaging, Biomarkers and Lifestyle (AIBL) study: progress and observations from 2359 older adults spanning the spectrum from cognitive normality to Alzheimer's disease. J Alzheimers Dis Rep 2021;5:443–68.

[22] Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 2013;41:e90.

[23] Zhou W, Triche TJ Jr, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res 2018;46:e123.

[24] Xu Z, Langie SA, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics 2017;18:4.

[25] Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016;48:1443–8.

[26] Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284–7.

[27] 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74.

[28] Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. Cell 2019;179:736–49.e15.

[29] Nabais MF, Laws SM, Lin T, Vallerga CL, Armstrong NJ, Blair IP, et al. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. Genome Biol 2021;22:90.

[30] Porter T, Burnham SC, Savage G, Lim YY, Maruff P, Milicic L, et al. A polygenic risk score derived from episodic memory weighted genetic variants is associated with cognitive decline in preclinical alzheimer's disease. Front Aging Neurosci 2018;10:423.

[31] Porter T, Burnham SC, Milicic L, Savage G, Maruff P, Lim YY, et al. Utility of an Alzheimer's disease risk-weighted polygenic risk score for predicting rates of cognitive decline in preclinical Alzheimer's disease: a prospective longitudinal study. J Alzheimers Dis 2018;66:1193–211.

[32] Gorrie-Stone TJ, Smart MC, Saffari A, Malki K, Hannon E, Burrage J, et al. Bigmelon: tools for analysing large DNA methylation datasets. Bioinformatics 2019;35:981–6.

[33] Ameijeiras-Alonso J, Crujeiras RM, Rodriguez-Casal A. multimode: an R package for mode assessment. J Stat Softw 2021;97:1–32.

[34] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–34.

[35] Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am J Hum Genet 2015;96:926–37.

[36] Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, et al. Ancestry estimation and control of population stratification for sequence-based association studies. Nat Genet 2014;46:409–15.

[37] Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, et al. Accounting for population stratification in DNA methylation studies. Genet Epidemiol 2014; 38:231–41.

[38] Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. Genome-wide methylation data mirror ancestry information. Epigenetics Chromatin 2017;10:1.

[39] Yuan V, Price EM, Del Gobbo G, Mostafavi S, Cox B, Binder AM, et al. Accurate ethnicity prediction from placental DNA methylation data. Epigenetics Chromatin 2019;12:51.

[40] Wang C, Zollner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet 2012;8:e1002886.

[41] Chen T, Chen Xu, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021; 19:578–83.

Genomics, Proteomics & Bioinformatics, 2024, 22, 1–12
https://doi.org/10.1093/gpbjnl/qzae044
Method