Hicks and Irizarry *Genome Biology* (2019) 20:261
https://doi.org/10.1186/s13059-019-1827-8

Genome Biology

METHOD Open Access

methylCC: technology-independent estimation of cell type composition using differentially methylated regions

Stephanie C. Hicks[1] and Rafael A. Irizarry[2,3]*

Abstract

A major challenge in the analysis of DNA methylation (DNAm) data is variability introduced from intra-sample cellular heterogeneity, such as whole blood which is a convolution of DNAm profiles across a unique cell type. When this source of variability is confounded with an outcome of interest, if unaccounted for, false positives ensue. Current methods to estimate the cell type proportions in whole blood DNAm samples are only appropriate for one technology and lead to technology-specific biases if applied to data generated from other technologies. Here, we propose the technology-independent alternative: *methylCC*, which is available at https://github.com/stephaniehicks/methylCC.

Keywords: DNA methylation, Whole blood, Cell composition, Microarray, HumanMethylation27 BeadChip, HumanMethylation450 BeadChip, Whole genome bisulfite-sequencing, Reduced representation bisulfite-sequencing

Background

DNA methylation (DNAm) is a type of chemical modification occurring at CpG dinucleotide sites that is involved in controlling gene expression and has been shown to play an important role in distinguishing cell lineages [1]. High-throughput DNAm assays have been widely applied among researchers as well as large consortia to further our understanding of basic biology and health implications [2]. However, a major challenge in extracting information from these DNAm datasets is variability introduced from intra-sample cellular heterogeneity observed in samples of heterogeneous cell composition. Specifically, individual cell types encode unique cell type-specific DNAm signatures to distinguish between the cell lineages. Therefore, when we measure DNAm on samples with a heterogeneous cell composition, we actually observe a convolution of the DNAm profiles of each cell type [3]. It is common for variability in cell type proportions to explain most of the observed sample-to-sample variability.

Cell composition induced variability is particularly problematic in epigenome-wide association studies (EWAS) [4] because, due to convenience, these are most frequently performed on whole blood, a highly heterogeneous tissue. In a seminal paper, Houseman et al. [3] describe a statistical method that accurately estimates the relative proportions of cell type components in whole blood. Jaffe et al. [5] used this approach to demonstrate that reported age-related changes of blood DNAm profiles [6–12] could be explained with high levels of confounding between age-related variability and cell composition, demonstrating the importance of accounting for this source of variability. As the consequential effect of this source of variability started to be recognized, interest in statistical methods for estimating and accounting for intra-sample cellular heterogeneity grew accordingly. There are currently two major types of approaches. The first, originally developed by Houseman et al. [3], assumes that the observed heterogeneous blood profiles are a linear combination of the cell type-specific DNAm profiles, assumes these DNAm profiles are known, and then estimates the unknown proportions using a standard estimation procedures. To be able to assume cell type-specific DNAm profiles are known, a rather complex experiment, in which cells of the same cell

*Correspondence: rafa@ds.dfci.harvard.edu
[2]Department Data Sciences, Dana-Farber Cancer Institute, 450 Brookline Ave,, Boston, USA
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, USA
Full list of author information is available at the end of the article

type are sorted and then used to obtain high-throughput measurements of the reference samples, is conducted. Methods that make use of the sorted cell type-specific DNAm profiles are referred to as *reference-based*. Alternatively, other methods that do not use external reference profiles, referred to as *reference-free* methods, have been developed for DNAm data [13, 14] and for more general types of data such as Surrogate Variable Analysis (SVA) [15], or Remove Unwanted Variability (RUV) [16] to account for batch effects [17].

Reference-based approaches have been shown to greatly outperform reference-free procedures [18]. Here, we consider them to be the state of the art. However, in this paper, we demonstrate that a limitation of reference-based approaches is the presence of a technology-specific bias, which can influence the estimates of cell composition; namely, when using cell type-specific DNAm profiles measured using one technology, for example a microarray platform, to estimate the cell type proportions in samples measured from another technology, for example a sequencing platform. Here, we introduce a statistical method, referred to as *methylCC*, that removes this technical bias using a latent variable model and accurately estimates the cell composition in a platform-agnostic manner. To achieve this, we identify regions of the genome in which each cell type are either clearly methylated or unmethylated, and we model these as latent states. These latent states are biologically driven and therefore technology-independent, which allows us to estimate binary, platform-independent profiles that can be successfully be applied across technologies. To study the improvements in estimates of cell composition using methylCC, we evaluated the difference between the true and estimated proportions of cell types with a Monte Carlo simulation. Specifically, we studied how using cell type-specific DNAm profiles measured on a microarray platform to estimate the cell type proportions in samples measured on a sequencing platform can lead to inaccurate estimates of cell composition. Furthermore, we demonstrate how our platform-agnostic approach provides an overall improvement in estimates of cell composition. Although due to the availability of data all our examples are from whole blood, the approach can be generalized to other tissues.

Results

Consider a set of high-throughput data Y_{ij} representing a heterogeneous tissue sample, such as whole blood, from $i \in (1, \ldots, N)$ individuals containing DNAm measurements at CpG sites $j \in (1, \ldots, J)$. Suppose the heterogeneous tissue is a combination of K cell types, which we index with $k \in (1 \ldots K)$. Houseman et al. [3] proposed the following statistical model to estimate the proportions of K cell types in whole blood DNAm samples, for each individual i:

$$\boldsymbol{Y}_i = \sum_{k=1}^{K} \pi_{ik} \boldsymbol{X}_k + \boldsymbol{\varepsilon_i}. \tag{1}$$

Here π_{ik} represents the proportion of cell type k in individual i, which is the parameter of interest. The $\boldsymbol{X}_k$ represents the kth cell type-specific DNAm profile with measurements at the same J CpG sites as $\boldsymbol{Y}_i$. The measurement error and other unexplained biological variability is represented by $\boldsymbol{\varepsilon}_i$. The cell type proportions for individual i are assumed to be nonnegative, $\pi_{ik} \geq 0$, and sum to 1, $\sum_{k=1}^{K} \pi_{ik} = 1$. To develop a practical tool, Houseman et al. [3] sorted whole blood samples into $K = 6$ cell types that make up the majority of this tissue and obtained a DNAm profile for each cell type. They used Illumina's HumanMethylation27 BeadChip (Illumina 27K), which measures DNAm at approximately 27,000 CpG sites [19]. This experimental data provided plug-in estimates for the cell type-specific DNAm profile, $\boldsymbol{X}_k$, and with these in place then they estimated the π_{ik} using a constrained least square algorithm. Soon after the development of this method, Illumina released a new platform that measured approximately 450,000 CpG sites: the HumanMethylation450 BeadChip (Illumina 450K) [20]. Jaffe et al. [5] leveraged publicly available data of sorted cell types measured with this new Illumina 450K platform [1] to implement the Houseman et al. method [3].

Although the Illumina 450K microarray platform has been the most widely used platform, two new sequencing technologies are being increasingly adapted by the research community: Whole-genome Bisulfite Sequencing (WGBS) and Reduced Representation Bisulfite Sequencing (RRBS) [21]. Furthermore, Illumina has recently released a new version of their BeadChip, which measures approximately 850,000 CpG sites. However, similar experiments with sorted cells processed at the same time are not yet available from these new platforms, which implies we do not have plug-in estimates for $\boldsymbol{X}_k$ on these platform technologies. Currently, the only way the Houseman et al. approach [3] can be applied to DNAm data measured on these new platforms is by assuming that the cell type-specific DNAm profiles $\boldsymbol{X}_k$ derived for the 450K platform applies to others. Here, we show this assumption does not hold.

Across platforms estimates are inaccurate

To determine if the Houseman et al. method [3], as implemented by Jaffe et al. [5], which was specifically developed for the Illumina 450K array platform, is applicable across platforms, we obtained a dataset for which whole blood samples from N=10 adult males were run on both the Illumina 450K and RRBS platforms (referred to below as the *two-platform dataset*). We applied the Houseman method to the whole blood samples in the *two-platform dataset* and expected similar cell composition estimates for each individual across platforms as these were the same whole blood samples just measured on two platforms. Because the Houseman method has been shown to provide reliable cell composition estimates for DNAm data measured on the Illumina 450K platform, in this specific case we considered the cell composition estimates from the Houseman method to be the gold-standard or ground truth, as done by Rahmani et al. [22]. However, we found that the resulting cell composition estimates between the $N = 10$ whole blood samples measured on the Illumina 450K and RRBS platforms did not agree (Fig. 1).

To determine the cause of this disagreement, we examined this dataset more closely and found two limitations with the Houseman approach when applied to technologies other than the Illumina microarrays: (1) DNAm measurements vary across platforms and (2) different platforms measure different CpGs. These two limitations are discussed in the following two sections, respectively. We then describe a statistical solution to overcome these two limitations using a general latent class model to estimate the cell composition of heterogeneous samples agnostic to platform technology. We also provide a software implementation of our method available at https://bioconductor.org/packages/release/methylCC.

DNAm measurements vary across platforms

The first limitation is that there is a platform-dependent bias. We can observe this bias by simply plotting and comparing the raw DNAm measurements using the whole blood samples in the *two-platform dataset*. We commonly observe genomic regions in which both platforms seem to indicate a change from unmethylated to methylated states, but the observed DNAm levels differ substantially across platforms (for example, Fig. 2a). A more systematic demonstration is obtained by first using the reference cell sorted dataset [1] to identify regions that are clearly unmethylated in all purified cell types and regions that are clearly methylated in all purified cell types, and then plotting the empirical DNAm distribution across all whole blood samples within these regions for both platforms (Fig. 2b) and noting the different distributions. We note in particular that observed DNAm levels measured on RRBS tends to have values closer to 0 and 1, compared to the Illumina 450K array attenuating these values away from the edges.

Different platforms measure different CpGs

The second limitation is that different platforms measure different CpGs. The human genome contains over

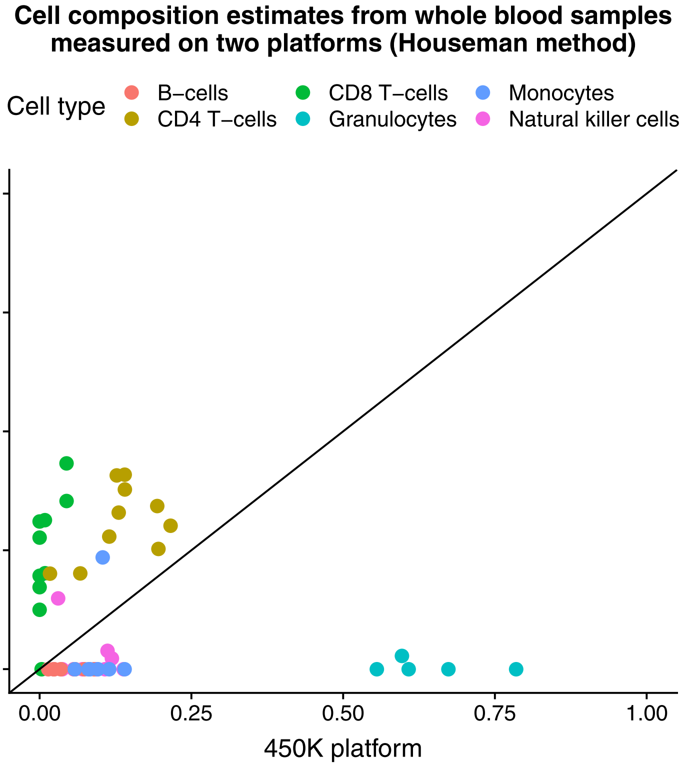

Fig. 1 Across platforms estimates are inaccurate. Cell composition estimates ($K = 6$ cell types) from $N = 10$ whole blood samples (*two-platform dataset* – GEO Accession GSE95163) measured on the Illumina 450k microarray platform (*x*-axis) and the RRBS platform (*y*-axis). The statistical method proposed by Houseman et al. [3], as implemented by Jaffe et al. [5], and was used to estimate the cell composition

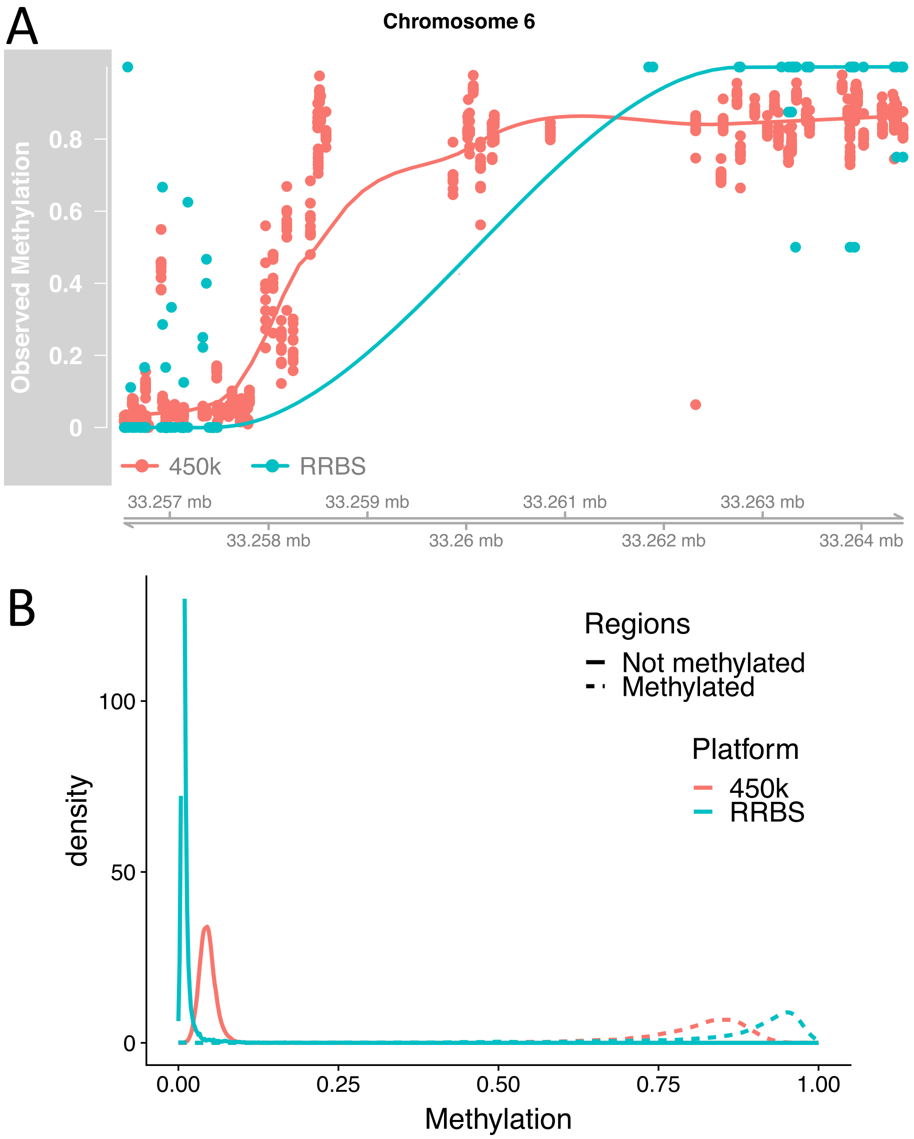

Fig. 2 Evidence for platform-dependent bias. **a** Observed DNAm levels from one region in same *N*=10 individuals (*two-platform dataset* – GEO Accession GSE95163) measured on two platform technologies: Illumina 450k (red) and RRBS (blue). **b** Density of DNAm levels measured on the 450k platform (red) and RRBS platform (blue) in regions that are either methylated (dashed line) or not methylated (solid line). Regions were identified by searching for regions in purified whole blood cell types from Reinius et al. [1] that appeared either methylated or not methylated in all six purified cell types

20,000,000 CpG sites, and each platform includes a subset of these which, for logistical reasons, differs across platforms. For example, RRBS [21] uses restriction enzymes to enrich for the areas of the genome that have a high CpG content, while the Illumina 450k platform selects CpG sites that are more uniformly distributed across the genome. Therefore, to apply the Houseman model to samples measured on platforms other than the Illumina 450K array, we have to restrict ourselves to the intersection of the CpGs measured Illumina 450K array and the alternative platform, because the j^{th} CpG in the whole blood sample $\boldsymbol{Y}_i$ must match the j^{th} CpG in the cell type-specific DNAm profile $\boldsymbol{X}_k$. As a result, in our 10 RRBS samples we only have measurements from 102 of the 600 CpGs in the cell type-specific DNAm profiles used by the 450K implementation of the Houseman method. This results in a loss of power since informative cell type-specific CpGs may be left out (for example Fig. 3).

methylCC estimates cell composition in DNAm samples agnostic to platform technology

To adjust for the platform-specific biases, we introduce a model that accounts for these biases directly and models methylation states using latent variables. To account for

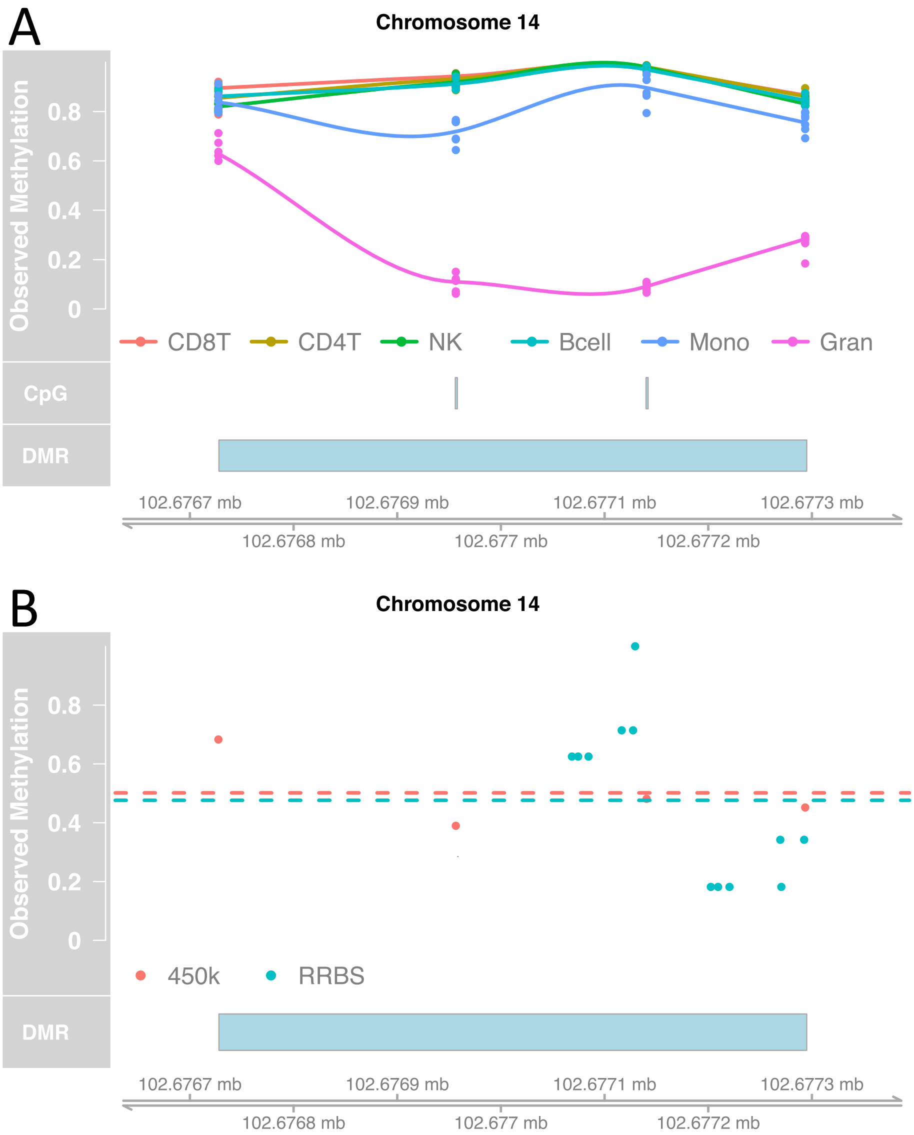

Fig. 3 Example of Houseman cell type-specific CpGs measured on the 450k microarray platform, but not measured on the RRBS platform. **a** The cell type-specific CpG and the cell type-specific region of CpGs (or differentially methylated regions, DMR) were both identified using the purified CD8T cell, CD4T cell, Natural Killer (NK), B cell, Monocytes (Mono) and Granulocytes (Gran) cell types from [1]. **b** In the whole blood DNAm data from one individual (GEO Accession GSE95163) measured on two platforms (450k and RRBS), the cell type-specific CpG from Houseman et al. [3] is not measured on the RRBS platform. However, using a cell type-specific region (DMR), we are able to measure the methylation level by averaging across the region in both the 450k and RRBS platform for this sample (dotted lines)

the fact that different platforms measure different CpG sites, we model the latent classes at the region level rather than the CpG level. Specifically, we propose the following statistical model:

$$\boldsymbol{Y}_i = \sum_{k=1}^{K} \pi_{ik} \left\{ (1 - \boldsymbol{Z}_k)\boldsymbol{\delta}_0 + \boldsymbol{Z}_k \boldsymbol{\delta}_1 \right\} + \boldsymbol{\varepsilon}_i \tag{2}$$

where $\boldsymbol{Y}_i$ is the observed DNAm level in the heterogeneous tissue, in this case whole blood, for the i^{th} individual $i \in (1, \ldots, N)$, but now measured DNAm levels in R genomic regions $r \in (1, \ldots, R)$, as opposed to J individual CpGs in the Houseman model. Similar to the Houseman model, π_{ik} represents the proportion of cell type k in individual i, which is the parameter of interest. In addition, we assume that the cell type proportions for individual i are nonnegative, $\pi_{ik} \geq 0$, and sum to 1, $\sum_{k=1}^{K} \pi_{ik} = 1$. Here, $\boldsymbol{Z}_k = (Z_{1k}, \ldots, Z_{Rk})$ is a vector of latent variables for the k^{th} cell type where each latent variable, Z_{rk}, is an indicator that is equal to 1 if the region r is methylated in cell type k and 0 otherwise. The platform-specific biases are represented with random effects $\boldsymbol{\delta}_0 = (\delta_{0,1}, \ldots, \delta_{0,R})$

and $\boldsymbol{\delta}_1 = (\delta_{1,1}, \ldots, \delta_{1,R})$, which are assumed to follow multivariate normal distributions $N\left(\alpha_0 \mathbf{1}, \sigma_0^2 I_{(R\times R)}\right)$ and $N\left(\alpha_1 \mathbf{1}, \sigma_1^2 I_{(R\times R)}\right)$, respectively. Measurement error and other unexplained biological variability is represented with $\boldsymbol{\varepsilon}_i$, which we assume follows a multivariate normal distribution $N\left(0, \tau^2 I_{(R\times R)}\right)$. Note that in our model the random effects $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$ are assumed to be platform-dependent: they represent the technology-dependent bias with different mean and variances in different platforms (Fig. 2b). However, the $\boldsymbol{Z}_k$s are not platform-dependent: they are latent classes determined by biology.

The statistical model in Eq. 2 can be thought of as a generalization of Eq. 1 if we restrict the Houseman approach to only include CpGs that are either methylated or unmethylated in each cell type. In this case, region r would simply be a single CpG site and the k^{th} cell type-specific DNAm profile, $\boldsymbol{X}_k$, would be defined by $X_{rk} = \delta_{0,r}$ if region r is unmethylated and $X_{rk} = \delta_{1,r}$ if region r is methylated.

A significant advantage of our model is that instead of directly measuring the cell type-specific DNAm profiles, $\boldsymbol{X}_k$, for each platform, we account for region-to-region variability using a latent random variable and therefore do not need to measure it directly with each new platform. Instead, all we need is to identify R regions for which each k^{th} cell type is either clearly methylated ($Z_{rk} = 1$), or not methylated ($Z_{rk} = 0$) for $r \in (1, \ldots, R)$. We define $\boldsymbol{Z}$ to be the matrix with entries Z_{rk} in the rth row and kth column, which needs to be full rank for the parameters of interest, π_{ik} to be identifiable. Because $\boldsymbol{Z}$ is entirely determined by biology, not by the platform technology, we only have to identify these regions once for each type of heterogeneous (biological) sample. This requires experimental data from cell sorted samples measured on only one platform. To demonstrate the utility this approach for estimating cell composition in whole blood samples, we searched for these genomic regions in the purified cell type data described in [1], which were measured on the Illumina 450K array platform. This dataset includes B cells, monocytes, granulocytes, CD8T cells, CD4T cells and natural killer (NK) cells. We identified $R = 210$ regions satisfying our criteria (Additional file 1: Figure S1). Finally, with the R regions in place, the estimation of the proportion of cell types, π_{ik}, reduces to a missing data problem. We use an EM algorithm with a constrained linear model to estimate the parameters $\boldsymbol{\theta} = \left(\alpha_0, \alpha_1, \sigma_0^2, \sigma_1^2, \tau^2\right)$ and $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ for individuals $i \in (1, \ldots, N)$ (see the "Methods" section for complete details on estimation procedure).

methylCC improves estimates of cell composition of DNAm samples measured on other platform technologies

To demonstrate the improvements in the estimates of cell composition provided by our platform-agnostic approach, we applied our method to the *two-platform dataset*. Specifically, we fit our model to the 10 whole blood samples measured on both the Illumina 450K array and RRBS platforms. Similar to Fig. 1, we considered the estimates of cell composition from the Houseman model in the Illumina 450K samples to be the gold-standard reference. In Fig. 1, we demonstrated that directly applying the Houseman approach [3], as implemented by Jaffe et al. [5], to the RRBS data led to biased cell composition estimates. However, our new approach substantially improves estimates of cell composition (Fig. 4).

Furthermore, we evaluate the performance of our platform-agnostic approach with the goal of estimating the proportion of cell types in heterogeneous tissue samples. Here, we performed a Monte Carlo simulation study to illustrate the improvements in estimates of cell composition by our platform-agnostic approach compared to the Houseman approach for heterogeneous samples measured on a sequencing platform (described in detail in the Methods Section). For the simulations study, we created cell type-specific DNAm profiles for a microarray platform, $\boldsymbol{X}_k^{450K}$, and a sequencing platform, $\boldsymbol{X}_k^{RRBS}$, by simulating platform-dependent random effects with different means and variances (Additional file 11: Figure S2A). Then, we simulate whole blood samples with a relative proportion of cell types $\boldsymbol{\pi}_i$ and measurement error $\boldsymbol{\varepsilon}_i$ to create the observed DNAm level in whole blood samples measured on in the 450k array platform $\boldsymbol{Y}_i^{450k}$ and the RRBS platform $\boldsymbol{Y}_i^{RRBS}$. We estimate the cell composition in the whole blood samples measured on both platform using the reference-based Houseman method and our platform-agnostic method. Then, we evaluate the difference between the true and estimated proportion of cell types using either our approach or the Houseman approach.

For whole blood samples measured on the 450K array platform, we found the Houseman approach, which was specifically developed for the array platform, and our approach perform similarly (Additional file 1: Figure S2B). However, for whole blood samples measured on a sequencing platform, our platform-agnostic model results in significantly improved estimates of cell composition (Additional file 1: Figure S2C). This is because our model accounts for the platform-specific biases directly and models methylation states using latent variables.

methylCC accurately estimates cell composition of DNAm samples measured on WGBS platforms

We evaluated our platform-agnostic approach using WGBS reference methylome data from the BLUEPRINT Epigenome Database [23]. We downloaded $N = 44$ samples from seven purified whole blood cell types, specifically B cells, CD4T cells, CD8T cells, neutrophils,

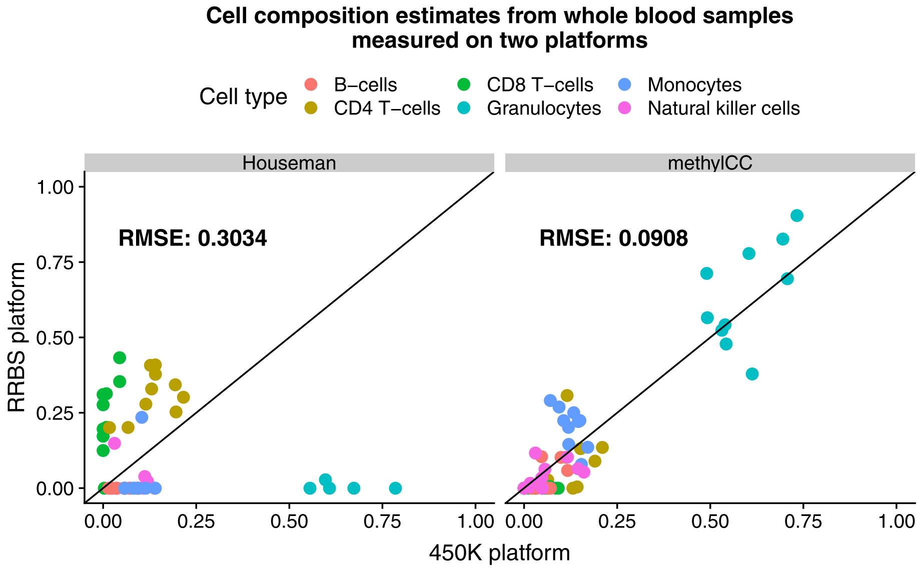

Fig. 4 methylCC: A latent variable model with region-specific and platform-dependent random effects improves estimates of cell composition. Cell composition estimates from $N = 10$ whole blood samples ($K = 6$ cell types) measured on the Illumina 450k microarray platform (*x*-axis) and the RRBS platform (*y*-axis). Two methods were used to estimate the cell composition: (1) the model proposed by Houseman et al. [3], as implemented by Jaffe et al. [5], that was developed for samples measured on the Illumina 450k microarray platform (left), and (2) our proposed method that is independent of platform technology (right). Root mean squared error (RMSE) is the difference between the cell composition estimates for whole blood DNAm samples measured on the Illumina 450k array platform and on the RRBS platform (averaged across cell types). Within each plot, the RMSE is shown for the reference-based method from Houseman et al. [3] (left) and our proposed model (methylCC) (right)

eosinophils, monocytes, and natural killer cells. For a given WGBS sample (e.g., CD8T cells), we assumed the "gold standard" cell composition to be 100% CD8T cells and 0% for the other cell types. We fit our model to $N = 44$ purified cell types measured on the WGBS platform. We found our platform-agnostic approach closely matches the expected cell composition estimates from the purified whole blood WGBS samples (Fig. 5).

Next, we used the BLUEPRINT reference methylomes to construct a set of cell type-specific DMRs to investigate whether DMRs identified with the purified cell types measured on the WGBS platform can lead to improved estimates of cell composition with methylCC as opposed to DMRs identified with the purified cell types measured on the Illumina 450K array platform. Using the WGBS "gold standard" data, we found the DMRs identified with the purified cell types measured on the WGBS platform resulted performed better than DMRs identified with the purified cell types measured on the Illumina 450K array platform (Additional file 1: Figure S3). Using the *two-platform dataset*, we found methylCC results in a substantial improvement over the Houseman approach with either set of DMRs, but using the DMRs identified with the 450K reference methylomes performs slightly better (Additional file 1: Figure S4). Data exploration of the BLUEPRINT data reveals that this is likely due to a batch effect in the BLUEPRINT data (Additional file 1: Figure S5).

methylCC accurately estimates cell composition of DNAm samples measured on Illumina 450K array platforms

To validate the results of our simulation study using whole blood samples measured on the Illumina 450K array platform, we compared the cell composition estimates from our model and the Houseman model using two publicly available data sets with DNAm whole blood samples measured on the Illumina 450K array platform. In the first data set [24], the N=78 whole blood samples had their cell composition independently estimated using flow cytometry, which can be considered as a "gold standard" [22]. We found our platform-agnostic approach closely matches the independent cell composition measurements (Additional file 1: Figure S6).

Next, we used a second data set [25] with N=689 whole blood samples, which did not have independent measurements of cell composition. Here, we considered the cell composition estimates from the Houseman model to be the "gold standard" for the purposes of this assessment because the Houseman model was specifically designed for the Illumina 450K array platform and it has been previously considered as a "gold standard" [22]. Using this data, we found our platform-agnostic approach closely

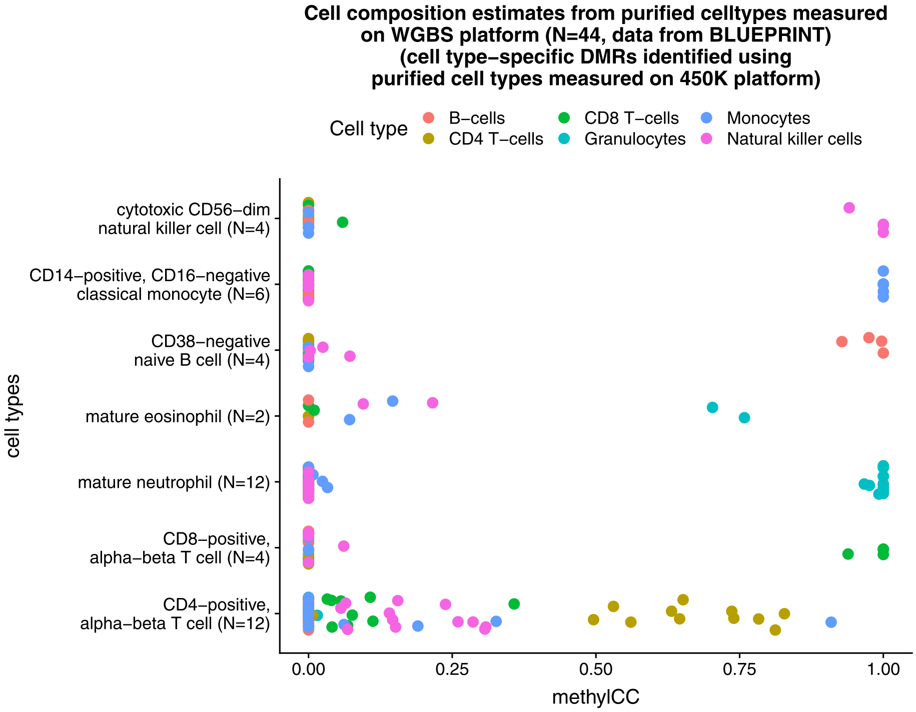

Fig. 5 methylCC accurately estimates cell composition of DNAm samples measured on WGBS platforms. Cell composition estimates using *N*=44 WGBS samples from the BLUEPRINT Epigenome Database [23] from seven purified whole blood cell types, specifically B cells, CD4T cells, CD8T cells, neutrophils, eosinophils, monocytes, and natural killer cells. For a given WGBS sample (e.g., CD8T cells), we assumed the "gold standard" cell composition to be 100% CD8T cells and 0% for the other cell types. We fit our model to $N = 44$ purified cell types measured on the WGBS platform

matches the referenced-based approach (Additional file 1: Figure S7).

Discussion

A major challenge in measuring DNAm is variability introduced from intra-sample cellular heterogeneity, which is a convolution of DNAm profiles across cell types. This is particularly problematic in epigenome-wide association studies for human disease performed on whole blood, a heterogeneous tissue. Accounting for this source of variability is a first step to determine the actual cell proportions of each sample. Currently, the most effective approach is based on fitting a linear model in which one assumes the DNAm profiles of the representative cell types are known for a specific platform technology, the Illumina microarray platform. Although this method works well in practice, we have demonstrated that if the DNAm data was generated on a new platform technology, such as RRBS or WGBS, this can lead to technology-specific biases in the cell composition estimates.

To address this, we have developed a latent variable model with region-specific and platform-dependent random effects to accurately estimate the cell composition in DNAm whole blood samples measured from any platform technology. By using informative genomic regions that are either methylated or unmethylated for each purified cell type, our model can account for the platform-specific biases directly and model methylation states using latent variables. We have illustrated how we can estimate the cell composition across platform technologies as cell types preserve their methylation state in regions independent of platform, despite observed measurements being platform-dependent. Note that, our current model assumes that the random effects and measurement error are normally distributed. Although these assumptions were a practical approximation that led to an improvement for RRBS data and accurately identified purified cell types in WGBS data, the model may need to be generalized to other distributions, such as count data for which negative binomial models may be more appropriate. Given that sequencing platform technologies are poised to become more widely used for studies measuring DNAm in whole blood, this suggests that our method is an needed contribution.

Conclusions

We demonstrated that our method accurately estimates the cell composition from whole blood samples and is applicable across multiple platforms, including microarray and sequencing platforms. Specifically, we illustrated how our method significantly improves the estimates of the cell composition compared to the reference-based method in whole blood samples measured on a sequencing platform using real and simulated whole blood samples, in addition to purified whole blood cell types measured on a sequencing platform. Our approach is agnostic to platform because it first uses experimental data to identify regions in which each cell type is clearly methylated or unmethylated, and then models these as latent states. While the continuous measurements used in the linear model approaches are affected by platform-specific biases, the latent states are biologically driven and therefore technology independent, implying that experimental data only needs to be collected once. We have implemented our method into the *methylCC* R-package providing researchers a tool to estimate the cell composition in the analysis of their own whole blood DNAm data.

Methods

Using cell sorted experimental data to identify informative genomic regions in $\boldsymbol{Z}$

Cell sorted experimental data is needed to identify R informative genomic regions for which the kth cell type is either clearly methylated ($Z_{rk} = 1$) or not methylated ($Z_{rk} = 0$) for regions $r \in (1, \ldots, R)$. This is step is done only once for each type of heterogeneous (biological) sample, such as whole blood, and does not depend on the platform technology. In addition, this matrix $\boldsymbol{Z}$ needs to be full rank for the parameters of interest, π_{ik} to be identifiable.

In application for estimating the cell composition in whole blood samples, we used cell sorted data described in [1], which were measured on the Illumina 450K array platform. This dataset includes six biological replicates for each of the six purified cell type (B cells, monocytes, granulocytes, CD8T cells, CD4T cells, and natural killer (NK) cells). We used the `bumphunter` [26] Bioconductor [27] package to identify differentially methylated regions (DMRs) across cell types. For example, to search for DMRs such that the six granulocytes samples are unmethylated and the other cell types are methylated (Fig. 3), we fit a linear model $Y_{ij} = \beta_0(l_j) + \beta_1(l_j)X_j + \varepsilon_{ij}$ at each jth genomic position (or CpG site) where Y_{ij} represents observed DNAm level in the i^{th} biological replicate for a purified cell type at position j with a covariate of interest, X_j, (for example $X_j = 0$ for granulocytes and $X_j = 1$ for other cell types). Then, we searched for regions of CpGs such that $\beta_1(l) \neq 0$. For more details on identifying DMRs, we refer the reader to [26, 28].

We searched for regions that were not overlapping so they would be considered independent observations. In certain pairwise cell type comparisons, the only regions found contained just one CpG; however, we prioritized regions with more than one CpG whenever possible. In addition to these cell type-specific DMRs, our method has the option for a user to search for and include additional cell type-specific CpGs along with the DMRs, if too few DMRs are found. Following these steps, we identified $R = 210$ regions satisfying our criteria (Additional file 1: Figure S1).

We also used the reference methylomes from the BLUEPRINT Epigenome Database (http://www.blueprint-epigenome.eu) [23], which contained $N = 44$ samples from seven purified whole blood cell types, specifically B cells, CD4T cells, CD8T cells, neutrophils, eosinophils, monocytes, and natural killer cells. We combined neutrophils, eosinophils as one group called granulocytes. We used the `bsseq` [29] Bioconductor [27] package to store the WGBS data and we used the `dmrseq` [30] package to identify DMRs across the six cell types.

In the next section, we describe our estimation procedure to obtain the cell composition estimates, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, and we note that we assume these regions $\boldsymbol{Z}$ are known here. This is because if we fit the model only to these regions, then the estimation procedure reduces to a missing data problem with random effects $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$.

Estimation procedure

Using the $R = 210$ informative genomic regions identified above, we estimate the parameters of interest, namely the proportion of cell types $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$ for the $i \in (1, \ldots, N)$ individuals, and the parameters $\boldsymbol{\theta} = \left(\alpha_0, \alpha_1, \sigma_0^2, \sigma_1^2, \tau^2\right)$ in the proposed latent variable model (Eq. 2) using an EM algorithm with constraints $\sum_{k=1}^{K} \pi_{ik} = 1$ and $\pi_{ik} \geq 0$ for all k.

Obtain initial parameter estimates $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\pi}_i^{(0)}$ at step $t = 0$

To obtain initial parameter estimates for the $\alpha_0^{(0)}$ and $\left(\sigma_0^2\right)^{(0)}$ at step $t = 0$, we use the reference cell sorted dataset [1], which has six biological replicates for each cell type, to identify a set of R^0 genomic regions that are clearly unmethylated ($Z_{rk} = 0$) in all K purified whole blood cell types. In these unmethylated regions, the expected DNAm level is

$$E(Y_{ir}) = \sum_{k=1}^{K} \pi_{ik} E(\delta_{0,r}) + E(\varepsilon_{ir}) = \sum_{k=1}^{K} \pi_{ik} \alpha_0 = \alpha_0$$

and we use Jensen's inequality to estimate an upper bound on the variance of Y_{ir}:

$$\begin{aligned} Var(Y_{ir}) &= Var\left(\sum_{k=1}^{K} \pi_{ik} \delta_{0,r}\right) + Var(\varepsilon_{ir}) \\ &\leq \sum_{k=1}^{K} \pi_{ik} Var(\delta_{0,r}) + Var(\varepsilon_{ir}) = \sigma_0^2 + \tau^2 \end{aligned}$$

Therefore, we obtain initial parameter estimates

$$\hat{\alpha}_0^{(0)} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{R^0}\sum_{r=1}^{R^0} Y_{ir}\right]$$

$$(\hat{\sigma}_0^2)^{(0)} \geq \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{R^0-1}\sum_{r=1}^{R^0}(Y_{ir}-\bar{Y}_i)^2\right]$$

where the measurement error τ^2 is assumed to be small. The argument is similar for the initial parameter estimates of $\alpha_1^{(0)}$ and $\left(\sigma_1^2\right)^{(0)}$ by identifying genomic regions (R^1) where the CpGs are all methylated $(Z_{rk} = 1)$ for all K purified cell types.

To obtain initial parameter estimates for the proportion of cell types, $\boldsymbol{\pi}_i^{(0)}$ with constraints $\sum_{k=1}^{K} \pi_{ik}^{(0)} = 1$ and $\pi_{ik}^{(0)} \geq 0$, we use the fact that $\hat{\boldsymbol{\pi}}_i = \text{argmin}_{\pi_i} \log L = \text{argmax}_{\pi_i}(-\log L)$ and $-\log L \propto (Y - X\pi)^T(Y - X\pi)$. This non-negative least squares (NNLS) problem with constraints is equivalent to the quadratic programming problem $\text{argmin}_{\pi_i}\left(\frac{1}{2}\pi_i^T Q\pi_i + a^T\pi_i\right)$ where $Q = (X^TX)$ and $a = \left(-X^TY\right)$ [31, 32]. Therefore, we calculate $\hat{\boldsymbol{X}}^{(0)} = (1-\boldsymbol{Z})\hat{\alpha}_0^{(0)} + \boldsymbol{Z}\hat{\alpha}_1^{(0)}$ and apply quadratic programming [31, 32] to solve for $\hat{\boldsymbol{\pi}}_i^{(0)} = \left(\pi_{i1}^{(0)}, \ldots, \pi_{iK}^{(0)}\right)$. We use the `solve.QP()` function from the R package `quadprog` [33] to implement the quadratic programming. Finally, to obtain an initial parameter estimate for $(\tau^2)^{(0)}$, we calculate

$$(\hat{\tau}^2)^{(0)} = \frac{1}{RN}\sum_{i=1}^{N}\sum_{r=1}^{R}\left(Y_{ir} - \sum_{k=1}^{K}\hat{X}_{rk}^{(0)}\hat{\pi}_{ik}^{(0)}\right)^2$$

EM algorithm to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$

To construct an EM algorithm to obtain maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, we define the complete-data vector $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1)$ where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ represents the observed DNAm levels for individuals $i \in (1, \ldots, N)$ each of length R regions. The complete-data likelihood is given by

$$f(\boldsymbol{Y}^*|\boldsymbol{\theta},\boldsymbol{\pi}) = \prod_{i=1}^{N} f_1(\boldsymbol{Y}_i|\boldsymbol{\delta}_0,\boldsymbol{\delta}_1,\boldsymbol{\theta},\boldsymbol{\pi}_i) f_2(\boldsymbol{\delta}_0|\boldsymbol{\theta}) f_3(\boldsymbol{\delta}_1|\boldsymbol{\theta})$$

where $f_1 \sim N\left(\sum_{k=1}^{K}\pi_{ik}\{(1-\boldsymbol{Z}_k)\boldsymbol{\delta}_0 + \boldsymbol{Z}_k\boldsymbol{\delta}_1\}, \tau^2 I_{(R\times R)}\right)$, $f_2 \sim N\left(\alpha_0, \sigma_0^2 I_{(R\times R)}\right)$, and $f_3 \sim N\left(\alpha_1, \sigma_1^2 I_{(R\times R)}\right)$. It is easy to show the log of the complete-data likelihood is linear in the following complete-data sufficient statistics: $T_1 = \sum_{r=1}^{R}\delta_{0,r}$, $T_2 = \sum_{r=1}^{R}\delta_{1,r}$, $T_3 = \sum_{r=1}^{R}(\delta_{0,r})^2$, $T_4 = \sum_{r=1}^{R}(\delta_{1,r})^2$, and $T_5 = \sum_{r=1}^{R}(u_{ir})^2$ where $u_{ir} = Y_{ir} - \sum_{k=1}^{K}\pi_{ik}\{(1-Z_{rk})\delta_{0,r} + Z_{rk}\delta_{1,r}\}$.

The EM algorithm alternates between the following two steps:

1 **E-Step**

We can consider the two joint distributions $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{\delta}_0)$ and $\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{\delta}_1)$ separately since $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$ are independent. The joint distributions are also normally distributed

$$\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{\delta}_0) \sim N\left(\begin{bmatrix}\boldsymbol{X\pi}\\ \alpha_0\mathbf{1}\end{bmatrix}_{((RN+R)\times 1)}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}_{((RN+R)\times(RN+R))}\right)$$

$$\boldsymbol{Y}^* = (\boldsymbol{Y}, \boldsymbol{\delta}_1) \sim N\left(\begin{bmatrix}\boldsymbol{X\pi}\\ \alpha_1\mathbf{1}\end{bmatrix}_{((RN+R)\times 1)}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}_{((RN+R)\times(RN+R))}\right)$$

where $\boldsymbol{Y}$ is a matrix of dimension $R \times N$, but we convert this into a vector of length RN, $\boldsymbol{X} = (1-\boldsymbol{Z})\alpha_0 + \boldsymbol{Z}\alpha_1$ is an $R \times K$ matrix and $\boldsymbol{\pi}$ is a $K \times N$ matrix. We convert the $\boldsymbol{X\pi}$ matrix into a vector of length RN. To derive the conditional distributions of $\boldsymbol{\delta}_0|\boldsymbol{Y}$ and $\boldsymbol{\delta}_1|\boldsymbol{Y}$, we use Theorem 3.2.3 and 3.2.4 in [34]:

$$\boldsymbol{\delta}_0|\boldsymbol{Y} \sim N\left(\alpha_0\mathbf{1} + \Sigma_{21}\Sigma_{11}^{-1}[\,\boldsymbol{Y} - \boldsymbol{X\pi}\,]\,, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

where

- $\boldsymbol{X} = (1-\boldsymbol{Z})\alpha_0 + \boldsymbol{Z}\alpha_1$ is an $R \times K$ matrix. $\boldsymbol{\pi}$ is a $K \times N$ matrix.
- $\Sigma_{11} = Cov(\boldsymbol{Y})$ is an $RN \times RN$ covariance matrix with entries

$$\begin{aligned} Cov(Y_{ir}, Y_{i'r'}) &= W_{0ri}^2\sigma_0^2 + W_{1ri}^2\sigma_1^2 + \tau^2 && \texttt{if } r = r', i = i' \\ &= W_{0ri}W_{0ri'}\sigma_0^2 + W_{1ri}W_{1ri'}\sigma_1^2 && \texttt{if } r = r', i \neq i' \\ &= 0 && \texttt{if } r \neq r', i \neq i' \end{aligned}$$

  where $W_{0ri} = \sum_{k=1}^{K}\pi_{ik}(1-Z_{rk})$, and $W_{1ri} = \sum_{k=1}^{K}\pi_{ik}Z_{rk}$
- $\Sigma_{12} = Cov(\boldsymbol{Y}, \boldsymbol{\delta}_0)$ is an $RN \times R$ covariance matrix with entries

$$\begin{aligned} Cov(Y_{ir}, \delta_{0,r'}) &= W_{0ri}\sigma_0^2 && \texttt{if } r = r' \\ &= 0 && \texttt{if } r \neq r' \end{aligned}$$

  Note: $\Sigma_{12}^T = \Sigma_{21}$.
- $\Sigma_{22} = Cov(\boldsymbol{\delta}_0)$ is an $R \times R$ matrix with $Var(\delta_{0,r}) = \sigma_0^2$ and $Cov(\delta_{0,r}, \delta_{0,r'}) = 0$

We use the conditional distribution $\boldsymbol{\delta}_0|\boldsymbol{Y}$ to calculate the tth iteration in the E-Step when computing $E_\theta(T_1|\boldsymbol{Y})$ and $E_\theta(T_3|\boldsymbol{Y})$.

$$T_1^{(t)} = \sum_{r=1}^{R}\left[\hat{\alpha}_0^{(t)}\mathbf{1} + \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}\left[\boldsymbol{Y}_i - \left\{(1-\boldsymbol{Z})\hat{\alpha}_0^{(t)} + \boldsymbol{Z}\hat{\alpha}_1^{(t)}\right\}\hat{\boldsymbol{\pi}}\right]\right]$$

$$T_3^{(t)} = \sum_{r=1}^{R}\left[\hat{\alpha}_0^{(t)}\mathbf{1} + \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}[\,\boldsymbol{Y}_i - \left\{(1-\boldsymbol{Z})\hat{\alpha}_0^{(t)} + \boldsymbol{Z}\hat{\alpha}_1^{(t)}\right\}\hat{\boldsymbol{\pi}}\,]\right]^2 + diag\left(\hat{\Sigma}_{22}^{(t)} - \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}\hat{\Sigma}_{12}^{(t)}\right)$$

Similarly, we can show

$$\boldsymbol{\delta}_1|\boldsymbol{Y} \sim N\left(\alpha_1\mathbf{1} + \Sigma_{21}\Sigma_{11}^{-1}[\,\boldsymbol{Y} - \boldsymbol{X\pi}\,]\,, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

where

- $\boldsymbol{X} = (1 - \boldsymbol{Z})\alpha_0 + \boldsymbol{Z}\alpha_1$ is an $R \times K$ matrix. $\boldsymbol{\pi}$ is a $K \times N$ matrix.
- $\Sigma_{11} = Cov(\boldsymbol{Y})$ is same as defined above.
- $\Sigma_{12} = Cov(\boldsymbol{Y}, \boldsymbol{\delta}_1)$ is an $RN \times R$ covariance matrix with entries

$$\begin{aligned} Cov(Y_{ir}, \delta_{1,r'}) &= W_{1ri}\sigma_1^2 & \texttt{if } r = r' \\ &= 0 & \texttt{if } r \neq r' \end{aligned}$$

  Note: $\Sigma_{12}^T = \Sigma_{21}$.
- $\Sigma_{22} = Cov(\boldsymbol{\delta}_1)$ is an $R \times R$ matrix with $Var(\delta_{1,r}) = \sigma_1^2$ and $Cov(\delta_{1,r}, \delta_{1,r'}) = 0$

We use the conditional distribution $\boldsymbol{\delta}_1|\boldsymbol{Y}$ to calculate the t^{th} iteration in the E-Step when computing $E_\theta(T_2|\boldsymbol{Y})$ and $E_\theta(T_4|\boldsymbol{Y})$.

$$T_2^{(t)} = \sum_{r=1}^{R}\left[\hat{\alpha}_1^{(t)}\mathbf{1} + \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}\left[\boldsymbol{Y} - \left\{(1-\boldsymbol{Z})\hat{\alpha}_0^{(t)} + \boldsymbol{Z}\hat{\alpha}_1^{(t)}\right\}\hat{\boldsymbol{\pi}}\right]\right]$$

$$T_4^{(t)} = \sum_{r=1}^{R}\left[\hat{\alpha}_1^{(t)}\mathbf{1} + \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}\left[\boldsymbol{Y} - \left\{(1-\boldsymbol{Z})\hat{\alpha}_0^{(t)} + \boldsymbol{Z}\hat{\alpha}_1^{(t)}\right\}\hat{\boldsymbol{\pi}}\right]\right]^2 + diag\left(\hat{\Sigma}_{22}^{(t)} - \hat{\Sigma}_{21}^{(t)}\left(\hat{\Sigma}_{11}^{(t)}\right)^{-1}\hat{\Sigma}_{12}^{(t)}\right)$$

2. **M-Step**

   The complete-data maximum likelihood estimates (MLEs) were calculated by using the log of the complete-data likelihood, taking the derivative with respect to the individual parameters, setting the likelihood equal to zero and solving for the MLEs.

$$\begin{aligned} \hat{\alpha}_0 &= \frac{T_1}{R} \\ \hat{\alpha}_1 &= \frac{T_2}{R} \\ \hat{\sigma}_0^2 &= \frac{T_3}{R} - (\hat{\alpha}_0)^2 \\ \hat{\sigma}_1^2 &= \frac{T_4}{R} - (\hat{\alpha}_1)^2 \end{aligned}$$

   Using these MLEs, we can substitute the sufficient statistics calculated in the E-Step:

$$\begin{aligned} \hat{\alpha}_0^{(t+1)} &= \frac{T_1^{(t)}}{R} \\ \hat{\alpha}_1^{(t+1)} &= \frac{T_2^{(t)}}{R} \\ (\hat{\sigma}_0^2)^{(t+1)} &= \frac{T_3^{(t)}}{R} - \left(\hat{\alpha}_0^{(t+1)}\right)^2 \\ (\hat{\sigma}_1^2)^{(t+1)} &= \frac{T_4^{(t)}}{R} - \left(\hat{\alpha}_1^{(t+1)}\right)^2 \end{aligned}$$

   To estimate $\boldsymbol{\pi}_i$, we apply quadratic programming [31, 32] (see section on "Obtain initial parameter estimates $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\pi}_i^{(0)}$ at step $t=0$" for details) with the constraints $\sum_{k=1}^K \pi_{ik} = 1$ and $\pi_{ik} \geq 0$ for all k. We calculate $\boldsymbol{X}^{(t)}$ using the t^{th} iteration of the conditional expectations $E_\theta(\boldsymbol{\delta}_0|\boldsymbol{Y})$ and $E_\theta(\boldsymbol{\delta}_1|\boldsymbol{Y})$ then apply quadratic programming [31, 32] to solve for $\hat{\boldsymbol{\pi}}_i^{(t+1)} = \left(\pi_{i1}^{(t+1)}, \ldots, \pi_{iK}^{(t+1)}\right)$. We use the `solve.QP()` function from the R package `quadprog` [33] to implement the quadratic programming.

   Finally, the MLE for τ^2, was calculated by using the log of the complete-data likelihood, taking derivative with respect to τ^2, setting likelihood equal to zero and solving.

$$(\hat{\tau}^2)^{(t+1)} = \frac{1}{R * N}\sum_{i=1}^{N}\sum_{r=1}^{R}\left(Y_{ir} - \sum_{k=1}^{K} X_{rk}^{(t)}\pi_{ik}^{(t+1)}\right)^2$$

Details for simulation studies

We created platform-dependent cell type-specific DNAm profiles for the k^{th} cell type ($\boldsymbol{X}_k^{450k}$ and $\boldsymbol{X}_k^{RRBS}$) where $\boldsymbol{X}_k^* = (1 - \boldsymbol{Z}_k)\boldsymbol{\delta}_0^* + \boldsymbol{Z}_k\boldsymbol{\delta}_1^*$ by simulating platform-dependent random effects $\left(\boldsymbol{\delta}_l^* \sim N(\alpha_l^*, (\sigma_l^2)^* I_{(R\times R)}\right)$ for both $l = 0, 1$ (Fig. 2b). For each whole blood DNAm sample ($N = 200$), we simulate a relative proportion of cell types ($\boldsymbol{\pi}_i$) and measurement error ($\boldsymbol{\varepsilon}_i$) to create the observed DNAm level in the 450k array platform $\left(\boldsymbol{Y}_i^{450k} = \sum_{k=1}^K \boldsymbol{\pi}_{ik}\boldsymbol{X}_k^{450k} + \boldsymbol{\varepsilon}_i\right)$ and the RRBS platform $\left(\boldsymbol{Y}_i^{RRBS} = \sum_{k=1}^K \boldsymbol{\pi}_{ik}\boldsymbol{X}_k^{RRBS} + \boldsymbol{\varepsilon}_i\right)$.

Assessment of performance

Next, we estimate the cell composition of, for example, the 450k array and RRBS samples using both the reference-based Houseman method and our platform-agnostic method. We do not scale the cell compositions estimates to 1 to allow for potential unaccounted cell types.

We calculate the cell type-specific $RMSE_k$ as

$$RMSE_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\pi}_{ik} - \pi_{ik})^2}$$

where π_{ik} is the true cell composition and $\hat{\pi}_{ik}$ is the estimated cell composition (using either Houseman model or our proposed model) in the i^{th} sample and kth cell type. The cell type-specific $RMSE_k$ is averaged across cell types and recorded as the mean RMSE. We repeat the above $n_{sims} = 100$ times to calculate the distribution of mean RMSE.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1827-8.

Additional file 1: Supplementary Figures S1-S7.

Open access

Authors' contributions

SCH and RAI developed the method methylCC. SCH wrote the methylCC R package, analyzed DNAm data and performed the simulation studies. SCH and RAI wrote the manuscript. Both authors read and approved the final manuscript.

Funding

SCH and RAI were supported by NIH R01 grants GM083084 and RR021967/GM103552 and HG005220.

Availability of data and materials

Software implementing the presented method to estimate the cell composition of whole blood samples measured from DNAm is available as an R package on GitHub (https://github.com/stephaniehicks/methylCC) and Zenodo (https://doi.org/10.5281/zenodo.3408423) [35], but also has been submitted to Bioconductor. The source code to reproduce the results presented are available on GitHub (https://github.com/stephaniehicks/methylCCPaper). The source code and methylCC package are all made available under the Creative Commons Attribution-NonCommercial 3.0 United States (CC BY-NC 3.0 US).
All the data used is publicly available including:

- Data from [1] measured cell sorted DNAm data from whole blood samples on the Illumina 450K array platform (GSE35069). This dataset includes six biological replicates for each of the six purified cell type ($N =$ 36 samples): B cells, monocytes, granulocytes, CD8T cells, CD4T cells, and natural killer (NK) cells. This data was used to build the Houseman et al. model [3]. This data is also available as an Bioconductor [27] data package [36].
- Data from Liu et al. [25] measures the DNAm levels in whole blood samples ($N = 689$) on the Illumina 450K array platform. This dataset from Liu et al. studied methylation differences between rheumatoid arthritis patients and normal controls (GSE42861). Here, we only considered the normal controls and used the proportion of cell types estimated using the Houseman approach as a gold standard or the true cell composition in each whole blood sample.
- Data from Rahmani et al. [24] from the Gala II population [37] measures DNAm levels in whole blood samples ($N = 78$) on the Illumina 450K array platform (GSE77716). These samples had their cell composition independently measured using flow cytometry.
- We validated our model by comparing the estimates cell composition of the same whole blood samples measured on two platform technologies from Carmona et al. [38] (GSE95163): (1) Illumina 450k microarray platform and (2) reduced representation bisulfite sequencing (RRBS) platform. The whole blood samples were derived from ten male individuals resulting in N =10 microarray samples and N=10 RRBS samples.
- We also validated our model using WGBS data from the BLUEPRINT Epigenome Database (http://www.blueprint-epigenome.eu) [23]. We downloaded $N = 44$ samples from seven purified whole blood cell types, specifically B cells, CD4T cells, CD8T cells, neutrophils, eosinophils, monocytes, and natural killer cells. We combined neutrophils, eosinophils as one group called granulocytes. For a given WGBS sample (e.g., CD8T cells), we assumed the "gold standard" cell composition to be 100% CD8T cells and 0% for the other cell types.

Ethics approval and consent to participate

All results in this paper are based only on publicly available data and do not require ethics approval.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St,, Baltimore, USA. [2]Department Data Sciences, Dana-Farber Cancer Institute, 450 Brookline Ave,, Boston, USA. [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, USA.

Received: 17 June 2019 Accepted: 19 September 2019
Published online: 29 November 2019

References

1. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE. 2012;7(7):41361. https://doi.org/10.1371/journal.pone.0041361.
2. Schübeler D. Epigenomics: Methylation matters. Nature. 2009;462(7271): 296–7. https://doi.org/10.1038/462296a.
3. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86. https://doi.org/10.1186/1471-2105-13-86.
4. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41. https://doi.org/10.1038/nrg3000.
5. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):31. https://doi.org/10.1186/gb-2014-15-2-r31.
6. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T.-P., Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20(4): 434–9. https://doi.org/10.1101/gr.103101.109.
7. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440–6. https://doi.org/10.1101/gr.103606.109.
8. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST. Age-associated dna methylation in pediatric populations. Genome Res. 2012;22(4):623–32. https://doi.org/10.1101/gr.125187.111.
9. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, Shin S-Y, Dempster EL, Murray RM, Grundberg E, Hedman AK, Nica A, Small KS, MuTHER Consortium, Dermitzakis ET, McCarthy MI, Mill J, Spector TD, Deloukas P. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012;8(4): 1002629. https://doi.org/10.1371/journal.pgen.1002629.
10. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J.-B., Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–367. https://doi.org/10.1016/j.molcel.2012.10.016.
11. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K, van den Berg LH, Ophoff RA. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol. 2012;13(10):97. https://doi.org/10.1186/gb-2012-13-10-r97.

12. Johansson A, Enroth S, Gyllensten U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. PLoS ONE. 2013;8(6): 67378. https://doi.org/10.1371/journal.pone.0067378.
13. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30(10):1431–9. https://doi.org/10.1093/bioinformatics/btu029.
14. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11(3):309–11. https://doi.org/10.1038/nmeth.2815.
15. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35. https://doi.org/10.1371/journal.pgen.0030161.
16. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52. https://doi.org/10.1093/biostatistics/kxr034.
17. McGregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, Greenwood CMT. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. Genome Biol. 2016;17:84. https://doi.org/10.1186/s13059-016-0935-y.
18. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, Irizarry RA, Teschendorff AE. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. Nat Methods. 2017;14(3):216–7. https://doi.org/10.1038/nmeth.4187.
19. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1(1):177–200. https://doi.org/10.2217/epi.09.14.
20. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan J-B, Shen R. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4): 288–95. https://doi.org/10.1016/j.ygeno.2011.07.007.
21. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18): 5868–77. https://doi.org/10.1093/nar/gki901.
22. Rahmani E, Schweiger R, Shenhav L, Wingert T, Hofer I, Gabel E, Eskin E, Halperin E. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol. 2018;19(1):141. https://doi.org/10.1186/s13059-018-1513-2.
23. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016;34(7):726–37. https://doi.org/10.1038/nbt.3605.
24. Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, Oh S, Burchard EG, Eskin E, Zou J, Halperin E. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat Methods. 2016;13(5):443–5. https://doi.org/10.1038/nmeth.3809.
25. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekström TJ, Feinberg AP. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31(2):142–7. https://doi.org/10.1038/nbt.2487.
26. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200–9. https://doi.org/10.1093/ije/dyr238.
27. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115–21. https://doi.org/10.1038/nmeth.3252.
28. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9. https://doi.org/10.1093/bioinformatics/btu049.
29. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):83. https://doi.org/10.1186/gb-2012-13-10-r83.
30. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics. 2019;20(3): 367–383. https://doi.org/10.1093/biostatistics/kxy007.
31. Goldfarb D, Idnani A. Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs. In: Hennart JP, editor. Numerical Analysis. Lecture Notes in Mathematics, Vol 909. Springer; 1982. https://doi.org/10.1007/bfb0092976.
32. Goldfarb D, Idnani A. A numerically stable dual method for solving strictly convex quadratic programs. Math Program. 1983;27(1):1–33. https://doi.org/10.1007/BF02591962.
33. Turlach BA, Weingessel A. Quadprog: Functions to Solve Quadratic Programming Problems. 2013. R package version 1.5-5. https://CRAN.R-project.org/package=quadprog.
34. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. San Diego: Academic Press; 1979.
35. Hicks SC, Irizarry RA. stephaniehicks/methylCC: Estimate the cell composition of whole blood in DNA methylation samples. GitHub. 2019. https://doi.org/10.5281/zenodo.3408423.
36. Jaffe AE. FlowSorted.Blood.450k: Illumina HumanMethylation Data on Sorted Blood Cell Populations. 2018. R package version 1.20.0. https://doi.org/10.18129/B9.bioc.FlowSorted.Blood.450k. https://www.bioconductor.org/packages/FlowSorted.Blood.450k.
37. Pino-Yanes M, Thakur N, Gignoux CR, Galanter JM, Roth LA, Eng C, Nishimura KK, Oh SS, Vora H, Huntsman S, Nguyen EA, Hu D, Drake KA, Conti DV, Moreno-Estrada A, Sandoval K, Winkler CA, Borrell LN, Lurmann F, Islam TS, Davis A, Farber HJ, Meade K, Avila PC, Serebrisky D, Bibbins-Domingo K, Lenoir MA, Ford JG, Brigino-Buenaventura E, Rodriguez-Cintron W, Thyne SM, Sen S, Rodriguez-Santana JR, Bustamante CD, Williams LK, Gilliland FD, Gauderman WJ, Kumar R, Torgerson DG, Burchard EG. Genetic ancestry influences asthma susceptibility and lung function among Latinos. J Allergy Clin Immunol. 2015;135(1):228–35. https://doi.org/10.1016/j.jaci.2014.07.053.
38. Carmona JJ, Accomando Jr. WP, Binder AM, Hutchinson JN, Pantano L, Izzi B, Just AC, Lin X, Schwartz J, Vokonas PS, Amr SS, Baccarelli AA, Michels KB. Empirical comparison of reduced representation bisulfite sequencing and Infinium BeadChip reproducibility and coverage of DNA methylation in humans. NPJ Genom Med. 2017;2:13. https://doi.org/10.1038/s41525-017-0012-9.

Publisher's Note

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