A Bayesian Framework for Estimating Cell Type Composition from DNA Methylation Without the Need for Methylation Reference

Elior Rahmani[1], Regev Schweiger[1], Liat Shenhav[1], Eleazar Eskin[2], and Eran Halperin[2]

[1] Tel Aviv University, Tel Aviv, Israel,
elior.rahmani@gmail.com
[2] University of California Los Angeles, Los Angeles CA, USA,
ehalperin@cs.ucla.edu

Abstract. Genome-wide DNA methylation levels measured from a target tissue across a population have become ubiquitous over the last few years, as methylation status is suggested to hold great potential for better understanding the role of epigenetics. Different cell types are known to have different methylation profiles. Therefore, in the common scenario where methylation levels are collected from heterogeneous sources such as blood, convoluted signals are formed according to the cell type composition of the samples. Knowledge of the cell type proportions is important for statistical analysis, and it may provide novel biological insights and contribute to our understanding of disease biology. Since high resolution cell counting is costly and often logistically impractical to obtain in large studies, targeted methods that are inexpensive and practical for estimating cell proportions are needed. Although a supervised approach has been shown to provide reasonable estimates of cell proportions, this approach leverages scarce reference methylation data from sorted cells which are not available for most tissues and are not appropriate for any target population. Here, we introduce BayesCCE, a Bayesian semi-supervised method that leverages prior knowledge on the cell type composition distribution in the studied tissue. As we demonstrate, such prior information is substantially easier to obtain compared to appropriate reference methylation levels from sorted cells. Using real and simulated data, we show that our proposed method is able to construct a set of components, each corresponding to a single cell type, and together providing up to 50% improvement in correlation when compared with existing reference-free methods. We further make a design suggestion for future data collection efforts by showing that results can be further improved using cell count measurements for a small subset of individuals in the study sample or by incorporating external data of individuals with measured cell counts. Our approach provides a new opportunity to investigate cell compositions in genomic studies of tissues for which it was not possible before.

Keywords: DNA methylation, epigenetics, Bayesian model, cell type composition, cell type proportions, tissue heterogeneity

1 Introduction

Epigenome-Wide Association Studies (EWAS), where genome-wide methylation levels are measured across a population and compared to a phenotype of interest, have become ubiquitous over the last few years. Many associations between methylation sites and disease status have been reported (e.g., multiple sclerosis [1], schizophrenia [2], and type 2 diabetes [3]), suggesting an important role for DNA methylation in complex diseases. Thus, DNA methylation status holds great potential for better understanding the role of epigenetics, potentially leading to better clinical tools for diagnosing and treating patients.

In a typical EWAS, we obtain a large matrix in which each entry corresponds to the methylation level (a number between 0 and 1) at a specific genomic position for a specific individual. This level represents the fraction of the probed DNA molecules that were found to have an additional methyl group at the specific position for the specific individual. In such studies, we typically search for rows of the methylation matrix (each corresponding to one genomic position) that are significantly correlated with a phenotype of interest. The analysis of EWAS is complicated by the fact that the studied tissue is typically a mixture of different cell types. Since each cell type may have a distinct methylation pattern, the resulting DNA methylation matrix is a convolution of the signals arising from the different cell types. As a result, a large number of false discoveries may be found in the common case where the cell type composition is correlated with the phenotype [4].

In principle, one can avoid false discoveries by adding high-resolution cell counts to a regression model commonly used in an EWAS. Unfortunately, such cell counting for a large cohort may be costly and often logistically impractical (e.g., in some tissues, such as blood, reliable cell counting can be obtained from fresh samples only). In order to overcome this problem and to allow correcting methylation data for cell type composition, several statistical and computational methods have been proposed [5–9]. These methods take either a supervised approach, in which reference data of methylation patterns from sorted cells are obtained and used for predicting cell compositions [5], or an unsupervised approach (reference-free) [6–9].

The main advantage of the reference-based method is that it provides direct (absolute) estimates of the cell counts, while current unsupervised methods are only capable of inferring components that capture linear combinations of the cell counts. However, the reference-based method can only be applied when relevant reference data exist. Currently, reference data only exist for blood [10], breast [11] and brain [12], for a small number of individuals (e.g., six samples in the blood reference [10]). In addition, the individuals in most data sets do not match the reference individuals in their methylation-altering factors such as age [13] and sex [14, 15]. This problem was recently highlighted in a study showing that available blood reference collected from adults fails to estimate cell proportions of newborns [16]. It is therefore often the case that unsupervised methods are either the only option or are a better option for the analysis of EWAS.

As opposed to the supervised approach, although can be applied for any tissue in principle, the reference-free methods do not provide direct estimates of the cell type proportions. A few reference-free methods allow us to infer a set of components, or general axes, which were shown to be linearly correlated with cell type composition [8, 9]. Unlike cell proportions, while linearly correlated components are useful in linear analyses such as linear regression, they cannot be used in any nonlinear downstream analysis (e.g., when studying specific cell types). Cell proportions may provide novel biological insights and contribute to our understanding of disease biology, and we therefore need targeted methods that are practical and low in cost.

Here, we propose an alternative strategy that utilizes prior knowledge about cell counts to improve upon the performance of reference-free methods, while addressing some of their limitations. We present a Bayesian semi-supervised method, BayesCCE (Bayesian Cell Count Estimation), which encodes experimentally obtained cell count information as a prior on the distribution of the cell type composition in the data. As we demonstrate here, the required prior is substantially easier to obtain compared with standard reference data from sorted cells. We can estimate this prior from general cell counts collected in previous studies, without the need for corresponding methylation data or any other genomic data.

We evaluate our method on two large data sets and on simulated data, and show that our method produces a set of components that can be used as cell composition estimates. We observe that each component of BayesCCE can be regarded as corresponding to a linear transformation of exactly one cell type (i.e. high absolute correlation with one cell type, but not necessarily good estimates in absolute terms). Considering existing reference-free methods as a baseline for estimating cell proportions, we find that BayesCCE provides improvement of up to 50% in correlation. We also consider the case where both methylation and cell count information are available for a small subset of the individuals in the sample, or for a group of individuals from external data. We show that our proposed Bayesian model can leverage such additional information, and we demonstrate that it allows us to impute missing cell counts in absolute terms. Testing this case on both real and simulated data, we find that measuring cell counts for a small group of samples (a couple of dozens) can lead to a further increase in the correlation of BayesCCE's components with the cell types composition. We therefore propose that future studies will consider measuring cell counts for at least a small number of the samples in the study, if possible, or incorporate into their analysis external data of samples with both methylation and measured cell counts from the same tissue.

2 Methods

2.1 Model

Let $O \in \mathbb{R}^{m \times n}$ be an m sites by n samples matrix of DNA methylation levels coming from heterogeneous source consisted of k cell types. For methylation

levels, we consider what is commonly referred to as beta-normalized methylation levels, which are defined for each samples in site as the proportion of methylated probes out of the total number of probes. Put differently, $O_{ji} \in [0,1]$ for each site j and sample i. We denote $M \in \mathbb{R}^{m \times k}$ as the cell type specific mean methylation levels for each site, and denote a row of this matrix, corresponding to the jth site, using $M_{j,\cdot}$. We denote $R \in \mathbb{R}^{n \times k}$ as the cell type proportions of the samples in the data, and denote $X \in \mathbb{R}^{n \times p}$ as a matrix of p covariates for each individual and a p-length row vector β_j as their corresponding effects in the jth site. If the measurements of O_{ji} were the true values of the methylation levels, then $O_{ji} = M_{j,\cdot}R_i^T + \beta_j X_i^T$. Due to measurement noise and other unmodeled factors, we incorporate an error term ϵ_{ji}. Thus, the full model for the observed methylation levels is

$$O_{ji} = M_{j,\cdot}R_i^T + \beta_j X_i^T + \epsilon_{ji} \tag{1}$$

$$\epsilon_{ji} \sim N(0, \sigma^2) \tag{2}$$

$$\forall i \forall h : R_{ih} \geq 0 \tag{3}$$

$$\forall i : \sum_{h=1}^{k} R_{ih} = 1 \tag{4}$$

$$\forall j \forall h : 0 \leq M_{jh} \leq 1 \tag{5}$$

The constraints in (3) and in (4) require the cell proportions to be positive and to sum up to one in each sample, and the constraints in (5) require the cell type specific mean levels to be in the range $[0,1]$. We note that the above formulation of the problem is similar to the one previously suggested in the context of reference-based estimation of cell proportions from DNA methylation by Houseman et al. [5]. The reference-based method first obtains estimates of M from reference methylation data collected from sorted cells of the cell types composing the studied tissue. Once M is known, R can be estimated by solving a standard quadratic program.

If the matrix M is not known, which is a reference-free version of the problem, the above formulation of the problem can be regarded as a version of non-negative matrix factorization (NNMF) problem. NNMF has been suggested in several applications in biology; notably, the problem of inference of cell type composition from methylation data has been recently formulated as a NNMF problem [9]. In order to optimize the model, the authors use an alternative optimization procedure in which M or R are optimized while the other is kept fixed. However, as demonstrated by the authors [9], their version of NNMF results in the inference of a linear combination of the cell proportions R. Put differently, more than one component of the NNMF is required for explaining each cell type in the data. Another recent reference-free method for estimating cell composition in methylation data, ReFACTor [8], performs a feature selection step followed by a principal components analysis (PCA). Similarly as in the NNMF solution, ReFACTor is an unsupervised method and it only finds principal components (PCs) that form linear combinations of the cell proportions rather than directly estimates the cell proportion values [8].

Here, we suggest a more detailed model by adding a prior on R and taking into account potential covariates. Specifically, we assume that

$$R_i^T \sim Dirichlet(\alpha_1, ..., \alpha_k) \tag{6}$$

where $\alpha_1, ..., \alpha_k$ are assumed to be known. In practice, the parameters are estimated from external data in which cell type proportions of the studied tissue are known. Such experimentally obtained cell type proportions were used to test the appropriateness of the Dirichlet prior in describing cell composition distribution (data not shown). We are interested in estimating R. Deriving a maximum likelihood-based solution for this model and repeating the constrains for completeness results in the following optimization problem:

$$\begin{aligned}
\hat{R}, \hat{M}, \{\hat{\beta}_j\}_{j=1}^m = \underset{R,M,\{\beta_j\}_{j=1}^m}{\operatorname{argmin}} \quad & \frac{1}{2\sigma^2} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(O_{ji} - M_{j,.} R_i^T - \beta_j X_i^T\right)^2 && (7) \\
& - \sum_{i=1}^{n} \sum_{h=1}^{k} (\alpha_h - 1) \log(R_{ih}) \\
\text{s.t} \quad & \forall i \forall h : R_{ih} \geq 0 && (8) \\
& \forall i : \sum_{h=1}^{k} R_{ih} = 1 && (9) \\
& \forall j \forall h : 0 \leq M_{jh} \leq 1 && (10)
\end{aligned}$$

Our intuition in this model is that since the priors on R are estimated from real data, incorporating them will push the solution of the optimization to retrun estimates of R which are closer to the true values as opposed to a linear combination of them.

2.2 Algorithm

Our algorithm uses ReFACTor as a starting point. Specifically, we use ReFACTor's PCs (ReFACTor components) in order to estimate R by finding an appropriate linear transformation of the ReFACTor components. In principle, both ReFACTor and NNMF could be used as the starting point for our method. However we found that ReFACTor captures a larger portion of the cell composition variance compared with the NNMF solution (see Results).

Applying ReFACTor on our input matrix O we get a list of t sites that are most informative with respect to the cell composition in O. Let $\tilde{O} \in \mathbb{R}^{t \times n}$ be a truncated version of O containing only the t sites selected by ReFACTor. We apply PCA on $\tilde{O}$ to get $L \in \mathbb{R}^{t \times d}, P \in \mathbb{R}^{n \times d}$, the loadings and scores of the first d ReFACTor components. Then, we reformulate the original optimization

problem in terms of linear transformations of L and P as follows:

$$\hat{A}, \hat{V}, \hat{B} = \underset{A,V,B}{\operatorname{argmin}} \ \frac{1}{2\sigma^2}||\tilde{O} - LAV^T P^T - LBX^T||_F^2 \tag{11}$$
$$-\sum_{i=1}^{n}\sum_{h=1}^{k}(\alpha_h - 1)\log\left(\sum_{l=1}^{d} P_{il}V_{lh}\right)$$
$$\text{s.t} \quad \forall i \forall h : \sum_{l=1}^{d} P_{il}V_{lh} \geq 0 \tag{12}$$
$$\forall i : \sum_{h=1}^{k}\sum_{l=1}^{d} P_{il}V_{lh} = 1 \tag{13}$$
$$\forall j \forall k : 0 \leq \sum_{l=1}^{d} L_{jl}A_{lh} \leq 1 \tag{14}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm, $A \in \mathbb{R}^{d\times k}$ is a transformation matrix such that $\tilde{M} = LA$ ($\tilde{M}$ being a truncated version of M with the t sites selected by ReFACTor), $V \in \mathbb{R}^{d\times k}$ is a transformation matrix such that $R = PV$ and $B \in \mathbb{R}^{d\times p}$ is a transformation matrix such that LB corresponds to the effects of each covariate on the methylation levels in each site. The constraints in (12) and in (13) correspond to the constraints in (8) and in (9), and the constraints in (14) correspond to the constraints in (10).

Given $\hat{V}$, we simply return $\hat{R} = P\hat{V}$ as the estimated cell proportions. Note that in the new formulation we are now required to learn only $d(2k+p)$ parameters - d, k and p being small constants - a dramatically decreased number of parameters compared with the original problem which requires $nk + m(k+p)$ parameters. By taking that approach, we make an assumption that $\tilde{O}$ consists of a low rank structure that captures the cell composition using d orthogonal vectors. While a natural value for d would be $d = k$, d is not bounded to be k. Particularly, in cases where substantial additional cell composition signal is expected to be captured by latter ReFACTor components (i.e. components beyond the first k), we would expect to benefit from increasing d. Clearly, overly increasing d is expected to result in overfitting and thus a decrease in performance. Finally, taking into account covariates with potentially dominant effects in the data should alleviate the risk of introducing noise into $\hat{R}$ in case of mixed low rank structure of cell composition signal and other unwanted variation in the data. We note, however, that similarly to the case of correlated explaining variables in regression, considering covariates that are expected to be correlated with the cell type composition may result in underestimation of A, V and therefore to a decrease in the quality of $\hat{R}$.

2.3 Imputing cell counts using a subset of samples with measured cell counts

In practice, we observe that each of BayesCCE's components corresponds to a linear transformation of one cell type rather than to an estimate of that cell type in absolute terms. That is, it still lacks the right scaling (multiplication by a constant and addition of a constant) for transforming it into cell proportions. Furthermore, we would like the ith BayesCCE component to correspond to the ith cell type described by the prior using the α_i parameter. Empirically, that is not necessarily the case, especially in scenarios where some of the α_i values are similar. In order to address these two caveats, we suggest incorporating measured cell counts for a subset of the samples in the data.

Assume we have n_0 reference samples in the data with known cell counts $R^{(0)}$ and n_1 samples with unknown cell counts $R^{(1)}$ ($n = n_0 + n_1$). This problem can be regarded as an imputation problem, in which we aim at imputing cell counts for samples with unknown cell counts. We can find $\hat{M}$ by solving the problem in (7) under the constraints in (10) for the n_0 reference samples while replacing R with $R^{(0)}$ and keeping it fixed. Then, given $\hat{M}$, we can now solve the problem in (11), after replacing LA with $\hat{M}$ (i.e. we find only V, B now), under the following constraints

$$\forall(1 \leq i \leq n_0)\forall h : \sum_{l=1}^{d} P_{il}^{(0)} V_{lh} = R_{ih}^{(0)} \tag{15}$$

$$\forall(1 \leq i \leq n_1)\forall h : \sum_{l=1}^{d} P_{il}^{(1)} V_{lh} \geq 0 \tag{16}$$

$$\forall(1 \leq i \leq n_1) : \sum_{h=1}^{k} \sum_{l=1}^{d} P_{il}^{(1)} V_{lh} = 1 \tag{17}$$

where $P^{(0)}$ contains n_0 rows corresponding to the reference samples in P, and $P^{(1)}$ contains n_1 rows corresponding to the remaining samples in P. In this case, both problems of estimating M and solving (11) while keeping $\hat{M}$ fixed are convex - the first problem takes the form of a standard quadratic problem and the latter results in an optimization problem of the sum of two convex terms under linear constraints. Using $\hat{M}$, estimated from cell counts and corresponding methylation levels of a group of samples, as well as adding the constraints in (15), are expected to direct the inference of R towards a set of components such that each one corresponds to one known cell type with a proper scale.

2.4 Implementation and practical issues

We estimate σ^2 in (11) as the mean squared error of predicting $\tilde{O}$ with P and X. The $\alpha_1, ..., \alpha_k$ Dirichlet parameters of the prior can be estimated from cell proportions using maximum likelihood estimators. In practice, we add a column of ones to both L and P in (11) in order to assure feasibility of the problem -

these constant columns are used to compose the mean methylation level per site across all cell types and the mean cell proportion fraction in each cell type across all samples. In addition, we slightly relax some of the constraints in the problem to avoid problems due to numeric instability and inconsistent noise issues. First, we do not impose the equality constraints in (13) and in (17) but rather allow a small deviation from equality (5%). In addition, the inequality constraints in (12) and in (16) are changed to require the cell proportions to be greater than $\epsilon > 0$, as a result of the logarithm term in the objective ($\epsilon = 0.0001$). Finally, given cell counts for a subset of the samples, we allow a small deviation from the equality constraints in (15) due to expected inaccuracies of cell counts measurements (1%).

We performed all the experiments in this paper using a Matlab implementation of BayesCCE. Specifically, we solved the optimization problems in BayesCCE using the *fmincon* function with the default interior-point algorithm, and we used the *fastfit* [17] Matlab package for calculating maximum likelihood estimates of the Dirichlet priors. All executions of BayesCCE required several minutes on a 64-bit Mac OS X computer with 3.1GHz and 16GB of RAM. Corresponding code is available at: https://github.com/cozygene/bayescce.

2.5 Evaluation of performance

The fraction of cell composition variation (R^2) captured by each of the reference-free methods, ReFACTor and NNMF, was computed for each cell type using a linear predictor fitted with the first k components provided by the method. In order to evaluate the performance of BayesCCE, for each component i we calculated its correlation with the ith cell type, and reported the mean absolute correlation (MAC) across the k estimated cell types. Empirically, we observed that in the case of $k = 6$ with no known cell counts for a subset of samples, the ith BayesCCE component did not necessarily correspond to the ith cell type. Put differently, the labels of the k cell types had to be permuted before calculating the MAC. In this case we considered the permutation of the labels which resulted with the highest MAC as the correct permutation. In the rest of the cases, we did not apply such permutation (all the experiments using $k = 3$ and all the experiments using $k = 6$ with known cell counts for a subset of the samples). For evaluating ReFACTor and NNMF, reference-free methods which do not attribute their components to specific cell types in any scenario, we considered the permutation leading to the highest MAC in all experiments when compared with BayesCCE. In addition, we considered the mean absolute error (MAE) as an additional quality measurement. When calculating absolute errors for the ReFACTor components, we scaled each component to be in the range $[0, 1]$. Finally, in experiments where cell counts were assumed to be known for a subset of the samples, MAC and MAE were calculated using only the samples for which cell counts were assumed to be unknown.

2.6 Implementation of ReFACTor and NNMF

The ReFACTor components were calculated for each data set using the parameters $k = 6$ and $t = 500$ and according to the implementation of ReFACTor described at http://glint-epigenetics.readthedocs.io, while accounting for known covariates in each data set. More specifically, in the Liu et al. data [18] we accounted for age, sex, smoking status and batch information, and in the Hannum et al. data [19] we accounted for age, sex, ethnicity and batch information. We used the first six ReFACTor components ($d = 6$) for simulated data in order to accommodate with the number of simulated cell types, and the first ten components ($d = 10$) for real data, as real data are typically more complex and are therefore more likely to contain substantial signal in latter components. The NNMF components were computed for each data set using the RefFreeEWAS R package from the subset of 10,000 most variable sites in the data set, as performed in the NNMF paper by the authors [9].

2.7 Implementation of the reference-based algorithm

We implemented the reference-based algorithm according to Houseman et al. [5], using 300 highly informative methylation sites defined in a recent study [20] and using reference data collected from sorted blood cells [10].

2.8 Data sets

We evaluated the performance of BayesCCE using three data sets collected with the Illumina 450K DNA methylation array. All three data sets are publicly available and preprocessed versions of the data were downloaded from the Gene Expression Omnibus (GEO) database. The first data set (accession GSE42861) was studied in a recent association study of DNA methylation with rheumatoid arthritis (RA) by Liu et al. ($n = 686$) [18]. The second data set (accession GSE40279) was originally used in a study of aging rates by Hannum et al. ($n = 656$) [19]. In addition, we used a reference data set of sorted cell types collected in six individuals from whole blood tissue (accession GSE35069) [10]. The latter was used for generating simulated data sets and for estimating the cell type specific mean levels in the implementation of the reference-based algorithm. We excluded from each data set sites coming from the sex chromosomes, as well as polymorphic and cross-reactive sites, as was previously reported [21]. Two samples in the Hannum et al. data were detected as outliers by PCA and were therefore excluded. When running BayesCCE on the data sets by Liu et al. and Hannum et al. we considered known batch information in the analysis.

2.9 Data simulation

We simulated data following a model that was previously described in details elsewhere [8]. Briefly, we used methylation levels from sorted blood cells [10] and, assuming normality, estimated maximum likelihood parameters for each site in

each cell type. Cell type specific DNA methylation data were then generated for each simulated individual from normal distributions with the estimated parameters, conditional on the range [0,1], for six cell types and for each site. Cell proportions for each individual were generated using a Dirichlet distribution. The parameters for the Dirichlet were fitted using the cell proportions estimated for the individuals in the Liu etl al. [18] and Hannum et al. [19] data sets using the reference-based method [5]. Finally, observed DNA methylation levels were composed from the cell type specific methylation levels and cell proportions for each individual, and a random normal noise was added to every data entry to simulate technical noise ($\sigma = 0.01$).

3 Results

3.1 Benchmarking existing reference-free methods

We first demonstrate that existing reference-free methods can estimate components that are correlated with the tissue composition in methylation data collected from heterogeneous sources. For the experiments in this paper, we used the whole-blood data set by Liu et al. [18] ($n = 686$) and the whole-blood data set by Hannum et al.[19] ($n = 654$; see Methods). In addition, we simulated data based on reference data set of methylation levels from sorted leukocytes cells [10] (see Methods). While cell proportions were known for each sample in the simulated data, cell counts were not available for the two real data sets. We therefore estimated the cell type composition of six major blood cell types (granulocytes, monocytes and four subtypes of lymphocytes) using a reference-based method [5], which was shown to reasonably estimate leukocyte cell proportions from whole blood methylation data collected from adult individuals [22, 16, 20]. Due to the absence of large publicly available data with measured cell counts, these estimates were considered as the ground truth for evaluating the performance of the different methods.

We considered two reference-free methods, ReFACTor [8] and NNMF [9], both allowing to generate components that were shown to capture cell type composition information in methylation. We evaluated the first six components of ReFACTor and the six components provided by NNMF - six being the number of estimated cell types composing the ground truth. We found both methods to capture a large portion of the cell composition in all data sets ; particularly, we observed that ReFACTor performed considerably better than NNMF in all data sets (Figure 1 a-c). Yet, in spite of the fact that both ReFACTor and NNMF capture a large portion of the cell composition variance, each component provided by these methods is a linear combination of the cell types in the data rather than an estimate of the proportions of a single cell type. As a result, as we show in the following experiments, both methods, in general, perform poorly when their components are considered as estimates of cell proportions.

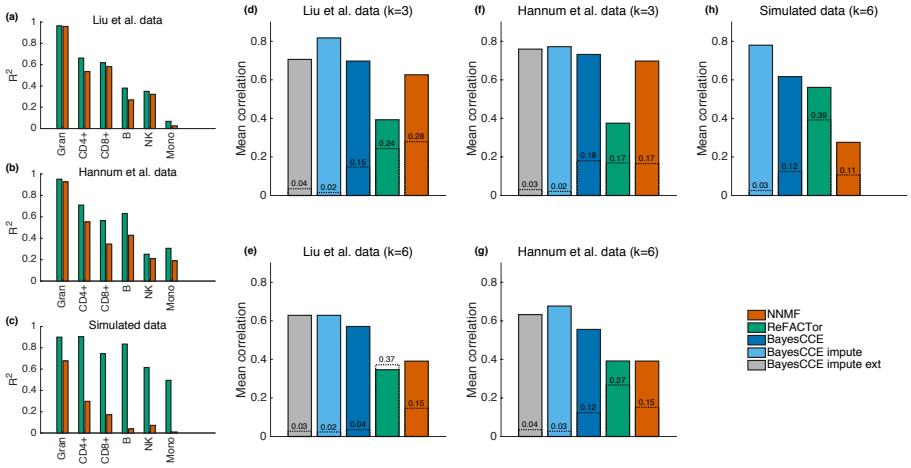

Fig. 1. Capturing cell type composition in real and simulated data. **(a)-(c)** The fraction of variance explained (R^2) by the first six components of NNMF and ReFACTor in each one of the six cell types in the Liu et al. data ($n = 686$), the Hannum et al. data ($n = 654$) and in simulated data ($n = 650$; average over 10 simulations). **(d)-(h)** Considering a single component for estimating each cell type in real data and in simulated data (average over 10 simulations), using existing reference-free methods (NNMF and ReFACTor) and BayesCCE, as well as using BayesCCE with known cell counts for 5% of the samples in the data (BayesCCE impute) and BayesCCE with cell counts and methylation for a group of samples from external data (BayesCCE impute ext). The bar plots present the mean absolute correlation of the components with the cell types in case of three assumed cell types ($k = 3$) and in case of six assumed cell types ($k = 6$). Dashed line on each bar plot indicates the mean absolute error of the estimates across all cell types. For more details about the evaluation of performance see Methods.

3.2 Evaluation of BayesCCE on real and simulated data

We evaluated BayesCCE under various scenarios. The results of the experiments described hereafter are summarize in Figure 1 d-h. In the first and most common scenario, we assume that no appropriate reference methylation data of sorted cells exist for the studied tissue, but we do have information about the distribution of the cell composition in the studied tissue. Such information can be inferred from cell counts collected in previous studies of the same tissue (without the need for any additional genomic data). This information can be then used by BayesCCE for tuning the prior required for the model (see Methods). In order to demonstrate this, we used cell counts collected from 35 healthy adults in a recent study [23]. These cell counts measured levels of lymphocytes, monocytes and three subtypes of granulocytes. Since our ground truth, compiled using the reference-based method, contained only the total granulocyte levels, we collapsed the three subtypes of granulocytes into a total measurement of granulocytes.

We applied BayesCCE on the real data sets under the assumption that three cell types compose the data ($k = 3$). Since each component of BayesCCE is expected to correspond to a linear transformation of one cell type, we report absolute linear correlations (see Methods). BayesCCE provided excellent estimates of the levels of granulocytes and lymphocytes in both data sets ($r = 0.96$ and $r = 0.98$ in the Liu et al. data, and $r = 0.94$ and $r = 0.98$ in the Hannum et al. data; see Figure 2). In contrast, we observed poor estimates of the monocyte levels ($r = 0.14$ in the Liu et al. data and $r = 0.26$ in the Hannum et al. data). We note that poor performance in capturing some cell type may be partially derived by inaccuracies introduced by the reference-based estimates which are used as the ground truth in our experiments. For example, several recent studies consisted of a relatively large number of samples for which both methylation levels and cell count measurements were available, demonstrated that while the reference-based estimates of the overall lymphocyte and granulocyte levels were found to be highly correlated with the true levels, the accuracy of the estimates of monocytes was found to be substantially lower [16, 8, 24]. Such inaccuracies in estimating a specific cell type by the reference-based approach may be the result of utilizing inappropriate reference. More specifically, cell types with highly variable methylation patterns across different populations may not be well represented for the target population by existing reference (coming from a specific population). Another possible driver for low quality estimates is having cell types with methylation profiles that do not distinct them well enough from other cell types in the tissue, or failing to select a set of informative features that mark some of the cell types.

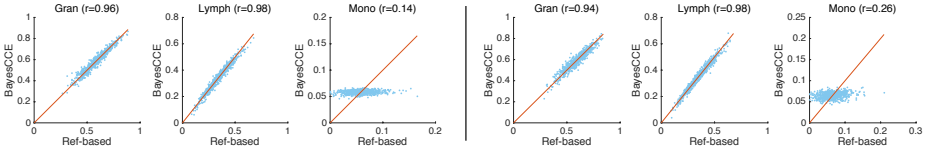

Fig. 2. BayesCCE captures cell type composition under the assumption of three cell types in the data ($k = 3$): granulocytes, lymphocytes and monocytes. Each BayesCCE component was linearly transformed to match its corresponding cell type in scale. Left: the results for the Liu et al. data. Right: the results for the Hannum et al. data.

As a second validation of our method, we used the reference-based estimates of the six cell types for learning the prior. For each one of the two real data sets, we used the cell proportion estimates of the other data set for learning the prior. We then applied BayesCCE on each data set under the assumption of six cell types ($k = 6$) and measured the correlation with the reference-based estimates. The mean absolute correlation across the six cell types was found to be 0.57 in the Liu et al. data and 0.56 in the Hannum et al. data (Figure 3). In addition to the real data analysis, we further conducted a similar experiment on simulated data ($n = 650$). In this case, we estimated the prior from a group of 50 samples

that were generated from the true distribution. We applied BayesCCE on ten different simulated data sets, and found the mean absolute correlation across all cell types and across all the simulated data sets to be 0.62. As expected, applying BayesCCE on increased sample size resulted in an improved performance (Supplementary Figure S1).

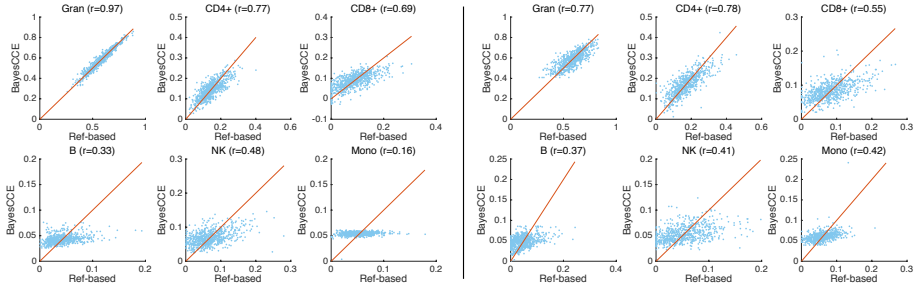

Fig. 3. BayesCCE captures cell type composition under the assumption of six cell types in the data ($k = 6$): granulocytes, four subtypes of lymphocytes (CD4+, CD8+, B cells and NK cells) and monocytes. Each BayesCCE component was linearly transformed to match its corresponding cell type in scale. Left: the results for the Liu et al. data. Right: the results for the Hannum et al. data.

3.3 Evaluation of cell count imputation

Next, we considered the scenario in which cell counts are known for a small subset of the samples in the data. This problem can be viewed as an imputation problem of the missing cell count values (see Methods). We repeated the previous experiments ($k = 3$ and $k = 6$), only this time we used the values of the estimated cell counts for randomly selected 5% of the samples in each data set. As opposed with the previous experiments, in which each one of BayesCCE's components formed a linear transformation of one of the cell types, here we get that the BayesCCE components form absolute estimates of the cell proportions (i.e. low absolute error). In addition, we observed up to 22% improvement in the mean correlation values compared with our previous experiments (Supplementary Figure S2 and Figure S3). We further tested this approach on simulated data ($n = 650$), while assuming known cell counts for 5% of the samples in the data, and found the mean correlation across different cell types and across ten different simulated data set to be 0.78. Applying this approach with an increased number of samples for which cell counts are known, reveals that the cell count estimates can be improved using a relatively small subset of a couple of dozens of samples with known cell counts (Supplementary Figure S4).

In the absence of cell counts for a subset of the individuals in the data, external data with samples for which both methylation levels and cell counts

are available can be added to the analysis. Again, we repeated the previous experiments ($k = 3$ and $k = 6$), only this time for each data set we added randomly selected 5% of the samples from the other data set, and used both their methylation levels and estimated cell counts in the analysis. Unlike in the previous experiments, here we potentially introduce new batch effects into the analysis, as in each experiment the original sample is combined with external data. We therefore accounted for the new batch information by adding it as a new covariate into BayesCCE. We observed up to 14% improvement in the mean correlation values compared with our previous experiments not taking any cell counts into account (Supplementary Figure S5 and Figure S6), showing that incorporating external samples with both methylation and cell counts can be a practical and useful way for estimating cell counts.

4 Discussion

We introduce BayesCCE, a Bayesian method that estimates cell type composition from heterogeneous methylation data using a prior on the cell composition distribution. In contrast to previous methods, using BayesCCE we can generate components such that each component corresponds to a linear transformation of a single cell type. These components can allow researchers to perform downstream analysis that is not possible using existing reference-free methods.

Our approach is based on finding a suitable linear transformation of the components found by ReFACTor [8]. Thus, it is limited by the quality of the ReFACTor components, and particularly BayesCCE will provide the exact same result as ReFACTor if used for correcting for potential cell type composition confounder in methylation data. We therefore suggest to use ReFACTor for correction and BayesCCE for cases in which a study of individual cell types is performed. We note that several supervised and unsupervised deconvolution methods have been suggested for estimating cell composition from gene expression [25–29]. However, these were refined for gene expression data and, to the best of our knowledge, none of these methods takes into account prior knowledge about the cell composition distribution as in BayesCCE. It remains of interest to investigate whether BayesCCE can be adapted for estimating cell composition from gene expression without the need for purified expression profiles.

The parameters of the prior required for BayesCCE can be estimated by utilizing previous studies that collected cell counts from the tissue of interest. Since no other genomic information is required, obtaining such data is relatively easy for many tissues, such as brain [30], heart [31] and adipose tissue [32]. Particularly, such data should be substantially easier to obtain compared to reference data from sorted cells for the corresponding tissues. Ideally, one would want to use cell counts coming from the same population as the target population in the study, especially when the cell composition distribution of the studied tissue may vary substantially across different populations. While this may be a potential limitation of BayesCCE in cases where cell counts from the target population are not available, our results using priors estimated from three differ-

ent data sets empirically show that priors estimated from a different population than the target population can still provide good estimates.

Since no large data with measured cell counts are currently publicly available, we used a supervised method [5] for obtaining cell type proportion estimates, which were used as the ground truth in our experiments. Even though the method used for obtaining these estimates was shown to reasonably estimate leukocyte cell proportions from whole blood methylation data in several independent studies [22, 16, 20], these estimates may have introduced biases into the analysis. Particularly, in the presence of systematic biases, the estimates could have affected the estimated priors, which in turn could have affected the results. However, we believe that our results on several independent data sets, including simulated data, and the use of priors estimated from several sources, including real cell counts, provide a compelling evidence for the utility of BayesCCE.

Finally, we demonstrate that imputation of the cell counts can be highly accurate even when cell counts are available for only a relatively small number of individuals. Moreover, in the general setting of BayesCCE, each component is correlated to one cell type, and the identity of that cell type may not be known, while in the case of imputation BayesCCE is able to reconstruct the cell counts up to a small absolute error (i.e. each component corresponds to a known cell type and is scaled to form cell proportion estimates of that cell type). We therefore recommend that in future studies either the cell counts be measured for at least a couple of dozens of the samples or external data of samples with measured cell counts be utilized in the analysis.

Acknowledgments. We would like to thank Lana Martin for feedback on the manuscript. This research was partially supported by the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. E.H., E.R., L.S. and R.S. were supported in part by the Israel Science Foundation (Grant 1425/13), E.H., L.S. and R.S. by the United States Israel Binational Science Foundation grant 2012304. E.R. and L.S. were supported by Len Blavatnik and the Blavatnik Research Foundation. R.S. was supported by the Colton Family Foundation. E.E. was supported by National Science Foundation grants 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants R01-GM083198, R01-ES021801, R01-MH101782, R01-ES022282 and U54EB020403.

Appendix

The supplementary materials can be found at:
https://github.com/cozygene/BayesCCE/blob/master/BayesCCE-SI.pdf.

References

1. Koch, M.W., Metz, L.M., Kovalchuk, O.: Epigenetic changes in patients with multiple sclerosis. Nature Reviews Neurology 9(1), 35–43 (2013)

2. Ikegame, T., Bundo, M., Sunaga, F., Asai, T., Nishimura, F., Yoshikawa, A., Kawamura, Y., Hibino, H., Tochigi, M., Kakiuchi, C., et al.: Dna methylation analysis of bdnf gene promoters in peripheral blood cells of schizophrenia patients. Neuroscience research 77(4), 208–214 (2013)
3. Toperoff, G., Aran, D., Kark, J.D., Rosenberg, M., Dubnikov, T., Nissan, B., Wainstein, J., Friedlander, Y., Levy-Lahad, E., Glaser, B., et al.: Genome-wide survey reveals predisposing diabetes type 2-related dna methylation variations in human peripheral blood. Human molecular genetics 21(2), 371–383 (2012)
4. Jaffe, A.E., Irizarry, R.A.: Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15(2), R31 (2014)
5. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T.: DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics (2012)
6. Houseman, E.A., Molitor, J., Marsit, C.J.: Reference-free cell mixture adjustments in analysis of dna methylation data. Bioinformatics 30(10), 1431–1439 (2014)
7. Zou, J., Lippert, C., Heckerman, D., Aryee, M., Listgarten, J.: Epigenome-wide association studies without the need for cell-type composition. Nature methods 11(3), 309–11 (Mar 2014)
8. Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J., et al.: Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. Nature methods 13(5), 443–445 (2016)
9. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., Marsit, C.J.: Reference-free deconvolution of dna methylation data and mediation by cell composition effects. BMC Bioinformatics 17(1), 259 (2016)
10. Reinius, L.E., 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 7(7), e41361 (2012)
11. Teschendorff, A.E., Gao, Y., Jones, A., Ruebner, M., Beckmann, M.W., Wachter, D.L., Fasching, P.A., Widschwendter, M.: Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nature communications 7 (2016)
12. Guintivano, J., Aryee, M.J., Kaminsky, Z.A.: A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics 8(3), 290–302 (2013)
13. Horvath, S.: Dna methylation age of human tissues and cell types. Genome biology 14(10), R115 (2013)
14. Singmann, P., Shem-Tov, D., Wahl, S., Grallert, H., Fiorito, G., Shin, S.Y., Schramm, K., Wolf, P., Kunze, S., Baran, Y., et al.: Characterization of whole-genome autosomal differences of dna methylation between men and women. Epigenetics & chromatin 8(1), 1–13 (2015)
15. Yousefi, P., Huen, K., Davé, V., Barcellos, L., Eskenazi, B., Holland, N.: Sex differences in dna methylation assessed by 450 k beadchip in newborns. BMC genomics 16(1), 1 (2015)
16. Yousefi, P., Huen, K., Quach, H., Motwani, G., Hubbard, A., Eskenazi, B., Holland, N.: Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. Environmental and molecular mutagenesis 56(9), 751–758 (2015)
17. Minka, T.: Estimating a dirichlet distribution (2000)

18. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al.: Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature biotechnology 31(2), 142–147 (2013)
19. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al.: Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell 49(2), 359–367 (2013)
20. Koestler, D.C., Jones, M.J., Usset, J., Christensen, B.C., Butler, R.A., Kobor, M.S., Wiencke, J.K., Kelsey, K.T.: Improving cell mixture deconvolution by id entifying o ptimal dna methylation l ibraries (idol). BMC bioinformatics 17(1), 1 (2016)
21. Chen, Y.a., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R.: Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. Epigenetics 8(2), 203–209 (2013)
22. Koestler, D.C., Christensen, B.C., Karagas, M.R., Marsit, C.J., Langevin, S.M., Kelsey, K.T., Wiencke, J.K., Houseman, E.A.: Blood-based profiles of dna methylation predict the underlying distribution of cell types: a validation analysis. Epigenetics 8(8), 816–826 (2013)
23. Chomczynski, P., Wilfinger, W.W., Eghbalnia, H.R., Kennedy, A., Rymaszewski, M., Mackey, K.: Inter-individual differences in rna levels in human peripheral blood. PloS one 11(2), e0148260 (2016)
24. Cardenas, A., Allard, C., Doyon, M., Houseman, E.A., Bakulski, K.M., Perron, P., Bouchard, L., Hivert, M.F.: Validation of a dna methylation reference panel for the estimation of nucleated cells types in cord blood. Epigenetics (just-accepted), 00–00 (2016)
25. Lu, P., Nakorchevskiy, A., Marcotte, E.M.: Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. Proceedings of the National Academy of Sciences 100(18), 10370–10375 (2003)
26. Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., Clark, H.F.: Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS one 4(7), e6098 (2009)
27. Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L., Luthi-Carter, R.: Population-specific expression analysis (psea) reveals molecular changes in diseased brain. Nature methods 8(11), 945–947 (2011)
28. Zuckerman, N.S., Noam, Y., Goldsmith, A.J., Lee, P.P.: A self-directed method for cell-type identification and separation of gene expression microarrays. PLoS Comput Biol 9(8), e1003189 (2013)
29. Steuerman, Y., Gat-Viks, I.: Exploiting gene-expression deconvolution to probe the genetics of the immune system. PLoS Comput Biol 12(4), e1004856 (2016)
30. Azevedo, F.A., Andrade-Moraes, C.H., Curado, M.R., Oliveira-Pinto, A.V., Guimarães, D.M., Szczupak, D., Gomes, B.V., Alho, A.T., Polichiso, L., Tampellini, E., et al.: Automatic isotropic fractionation for large-scale quantitative cell analysis of nervous tissue. Journal of neuroscience methods 212(1), 72–78 (2013)
31. Pinto, A.R., Ilinykh, A., Ivey, M.J., Kuwabara, J.T., DAntoni, M.L., Debuque, R., Chandran, A., Wang, L., Arora, K., Rosenthal, N.A., et al.: Revisiting cardiac cellular composition. Circulation research 118(3), 400–409 (2016)
32. Divoux, A., Tordjman, J., Lacasa, D., Veyrie, N., Hugol, D., Aissat, A., Basdevant, A., Guerre-Millo, M., Poitou, C., Zucker, J.D., et al.: Fibrosis in human adipose tissue: composition, distribution, and link with lipid metabolism and fat mass loss. Diabetes 59(11), 2817–2825 (2010)