HHS Public Access

Author manuscript

Brief Bioinform. Author manuscript; available in PMC 2023 November 19.

Published in final edited form as:

Brief Bioinform. 2022 November 19; 23(6): . doi:10.1093/bib/bbac449.

A systematic assessment of cell type deconvolution algorithms for DNA methylation data

Junyan Song, PhD,

Department of Applied Mathematics and Statistics, Stony Brook University

Pei-Fen Kuan [Associate Professor]

Department of Applied Mathematics and Statistics, Stony Brook University, Her research activities have been directed at developing statistical methodologies to facilitate the analysis, integration and interpretation of high-throughput omics data.

Abstract

We performed systematic assessment of computational deconvolution methods that play an important role in the estimation of cell type proportions from bulk methylation data. The proposed framework methylDeConv (available as an R package) integrates several deconvolution methods for methylation profiles (Illumina HumanMethylation450 and MethylationEPIC arrays) and offers different cell-type-specific CpG selection to construct the extended reference library which incorporates the main immune cell subsets, epithelial cells and cell-free DNAs. We compared the performance of different deconvolution algorithms via simulations and benchmark datasets and further investigated the associations of the estimated cell type proportions to cancer therapy in breast cancer and subtypes in melanoma methylation case studies. Our results indicated that the deconvolution based on the extended reference library is critical to obtain accurate estimates of cell proportions in non-blood tissues.

Keywords

DNA methylation; cell type heterogeneity; deconvolution; epigenetics; EWAS

Introduction

Cellular components play an important role in therapeutic response and disease diagnosis since many physiological processes involve cell motility and differentiation [1]. The extent of immune cell infiltration into tumors, an important component in tumor microenvironment, may give rise to different immune therapeutic responses [2]. For instance, the increase of lymphocyte infiltration in melanoma is associated with higher drug response rates and serves as predictive biomarkers for disease development [3]. Conventional methods for

Corresponding author: Pei-Fen Kuan, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Nicolls Road, Math Tower, Room 1-106, Stony Brook, NY 11794, USA. Tel.: +1-631-632-1419; Fax: +1-631-632-8490; peifen.kuan@stonybrook.edu.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

determining the cell compositions have several limitations. This includes the traditional flow cytometry method which is infeasible for large sample studies due to the high cost and difficulty in obtaining fresh tissues [4]. Additionally, physical cell type separation approaches, such as laser capture microdissection (LCM), fluorescence-activated cell sorting (FACS) and translating ribosome affinity purification (TRAP), have technical difficulties including the lack of good surface markers and cell-type-specific promoters [5]. These have rendered computational-based deconvolution methods for cell type proportion estimation of bulk epigenomics and transcriptomics data an attractive approach and active areas of research. Both DNA methylation and gene expression have emerged as important hallmark of numerous diseases including cancer [6-12]. DNA methylation profiling has several advantages over gene expression, including the stability and smaller amount of DNA required from formalin-fixed, paraffin-embedded samples compared with RNA [13]. The current state-of-the-art approach for DNA methylation profiling includes the Illumina Infinium HumanMethylation450 (450 k) array, and more recently the Infinium MethylationEPIC (EPIC) array which extends the CpG coverage to transcription factor binding sites, chromatin and enhancer regions.

In epigenome-wide association study (EWAS), cellular composition has been found to play a critical role in understanding the association between phenotypic of interest and DNA methylation. For instance, cellular composition was confounded with the strength of association between age, the primary phenotype and methylation levels [14]. Similarly, EWAS investigating the breastfeeding and epigenetic variation in buccal cells from 1006 twins was shown to be affected by the proportions of epithelial cells, leukocytes and natural killer (NK) cells [15]. The standard approach is to include the estimated cell proportions as adjustment factors to uncover the true association with the phenotypic outcome of interests, as evident from the smoking [16, 17] and neonatal EWAS [18]. Besides playing a role as confounder in EWAS, the estimated cell proportions from DNA methylation data can also be utilized to understand the biological mechanisms underlying the phenotypes. For example, Ankur *et al.* [4] showed that the ratio of CD8T to regulatory T cells ratio was elevated in hot tumors pan-cancer, whereas Hannon *et al.* [19] showed the several immune cell proportions were associated with psychosis and treatment-resistant schizophrenia. Recently, methods for inferring cell-specific differentially methylated CpGs (DMCs) from bulk tissues are emerging by incorporating the estimated cell proportions [20, 21]. For example, CellDMC [21] models the interactions between the phenotype and the estimated cell proportions, whereas tensor composition analysis (TCA) [20] uses tensor composition analysis for detection of cell-specific DMCs.

Cell type deconvolution algorithms for DNA methylation data

Cell type deconvolution algorithms can be divided into two main categories: the reference-based methods and reference-free methods [22]. Reference-based methods utilize pre-defined cell-type-specific differentially methylated regions (DMRs), whereas reference-free methods rely on unsupervised methods to infer putative cell proportion confounding factors. An advantage of the reference-based methods is that since the identities of inferred cell type proportions are known, one can further correlate the inferred cell type proportions to clinical attributes to ascertain if specific cell type has diagnostic or prognostic value.

Additionally, the reference-based methods tend to be less computationally intensive and provide more accurate estimates. Specifically, reference-based methods assume that the DNA methylation profile of a sample is a weighted sum of cell-type-specific reference profiles. These methods utilize a least-squares minimization to uncover the weights for the observed sample methylation profile. One commonly used algorithm is the quadratic programming (QP) or constrained projection (CP) [23], in which the weights are constrained to be non-negative and sum to one. In contrast, the CIBERSORT (CBS) algorithm [24], which one of the most widely used cell type deconvolution software for gene expression is based on a non-constrained reference-based method through a support vector regression with linear kernel, and is generalizable to various genomics features besides gene expression, as well as allowing users to create a custom signature matrix [2, 25]. On the other hand, EpiDISH [26] provides a unified computational pipeline for cell type deconvolution on DNA methylation data by combining three reference-based methods, namely Houseman's QP/CP [23], Robust Partial Correlation (RPC) and CBS [24]. MethylResolver [27] is another popular deconvolution method for DNA methylation that is based on least trimmed squares (LTS) regression to select an optimal set of CpGs for each cell type. ARIC [28] is a recently proposed deconvolution method based on a weighted support vector regression (SVR) that aims to provide robust estimation of rare cell types and uses a two-step feature selection strategy, which includes CpG collinearity and outlier elimination. On the other hand, examples of reference-free methods include RefFreeEWAS [29], BayesCCE [30] and TOAST [31]. RefFreeEWAS is an extension of Houseman's QP/CP [23] by approximating the matrix of cell-specific methylation using a two-stage regression analysis based on non-negative matrix factorization; BayesCCE follows a semi-supervised Bayesian framework with a prior estimated from external data, whereas TOAST is based on a iterative procedure that improves CpG selection, followed by deconvolution using existing algorithms such as RefFreeEWAS.

This paper aims to improve the deconvolution of DNA methylation array by integrating the deconvolution algorithms with alternative cell-type-specific CpG selection methods and extending the reference library to include both immune and epithelial cells. We develop a computational deconvolution tool methylDeConv available as an R package (https://github.com/jysonganan/methylDeConv) to provide a comprehensive and integrated pipeline for methylation data. Specifically, our proposed pipeline allows for a number of options for analyzing both the Illumina 450 k and EPIC arrays, including tissue types and reference libraries. Besides the choice of deconvolution algorithms, a critical first step in cell type deconvolution is the selection of informative CpGs, i.e. CpGs that contain information about the different cell types. Our pipeline methylDeConv also offers multiple cell-type-specific CpG selection methods based on differential testing or machine learning classification, and an alternative method of representing cell type proportions based on the predicted class probabilities of multi-class elastic net.

Methods

Deconvolution algorithms

We incorporated several deconvolution algorithms including the reference-based methods in EpiDISH [26], Houseman's QP/CP [23], RPC [26], CBS [24], MethylResolver [27] as well as the recently proposed method ARIC [28]. Two reference-free methods, namely RefFreeEWAS [29] and TOAST [31], were also included in our study. For RefFreeEWAS and TOAST, we used myRefFreeCellMix and csDeconv functions implemented in the TOAST R package, respectively These methods were implemented in R, except for ARIC which was written in Python. These methods are summarized in Table 1.

In estimating cell type proportions for samples profiled on the 450 k or EPIC methylation arrays, the reference-based deconvolution algorithms can be performed using reference profiles from Bioconductor packages FlowSorted.Blood.450 k [32] and FlowSorted.Blood.EPIC [33], respectively, which consisted of antibody-bead sorted and purified neutrophils, B cells, monocytes, NK cells, CD4T cells and CD8T cells. There were on average six replicates for each cell type in each platform. We argued that the selection of CpGs in the reference profiles was an important step in the deconvolution algorithms [34], which motivated us to consider several CpG selection strategies in the following subsection.

Cell-type-specific CpG selection

We considered several strategies to identify cell-type-specific CpGs from the reference profiles, namely (1) differential methylation analyses based on t-test [35, 36] and moderated *t*-test [4], (2) machine learning classification methods and (3) highly variable CpGs. In strategy (1), we first filtered the CpGs with *P*-values smaller than a user-defined threshold, followed by ranking them according to either the mean differences or the absolute mean differences to obtain a pre-specified number of top CpGs. In strategy (2), we considered two machine learning models, namely elastic net [37] and random forest [38]. For elastic net model, we retained the CpGs with non-zero coefficients from five-fold cross-validations. For random forest, the variables were selected based on either the variable importance score or the recursive feature elimination (RFE) algorithm in Caret package [39] which performed random forest iteratively with different subsets of CpGs. On the other hand, strategy (3) (i.e. choosing highly variable CpGs) was based on the observation that cell-type-specific CpGs were in general most variable since they corresponded to CpGs with distinct profiles across different cell types [31].

In strategy (1), we further considered (i) one-versus-all (comparing reference profiles of one cell type to the rest) and (ii) pairwise comparisons (comparing reference profiles for each pair of cell type). Similarly, in strategy (2), we also considered (i) two-class classification for each pair of cell type and (ii) multi-class classification for all cell types. These were summarized as methods 1–5 in Table 2.

Additionally, we also combined strategies (1) and (2) by proposing a two-step CpG selection framework. In the first step, we utilized the one-versus-all t-test to preselect a larger set of candidate CpGs (e.g. 300 top ranking CpGs per cell type, yielding at most 1800 candidate CpGs in six immune cell types). In the second step, we fitted a multi-class elastic net model

or a multi-class random forest model to further eliminate redundant CpGs. These were summarized as methods 6–8 in Table 2. A schematic diagram illustrating the main steps involved in DNA methylation based cell deconvolution is provided in Figure 1.

In Supplementary Appendix S1 (see Supplementary Data available online at https://academic.oup.com/bib), we also evaluated the performance of other strategies including using the predicted class probabilities and application of the enrichment-score based method as alternative estimates of cell type proportions.

Extension of reference library to non-immune cells

In non-blood tissues (e.g. saliva), epithelial cells made up a large fraction of cells. Additionally, in blood of cancer samples, there could be an increased level of cell-free DNA (cfDNA) in circulation. Therefore, we extended the reference library in FlowSorted.Blood.EPIC by adding 10 purified epithelial and 10 randomly selected cfDNA from Moss *et al.* [40] (available at Gene Expression Omnibus (GEO) under accession number GSE122126). In FlowSorted.Blood.450 k, we added 11 purified epithelial and 7 fibroblast cells from the ENCODE project [41] (GEO accession number GSE40699). Besides providing estimates of epithelial cell proportion for non-blood tissues and cfDNA for blood of cancer samples, we hypothesized that extending the reference library can also significantly improve the estimation of immune cell proportions.

Benchmark datasets

We included several benchmark datasets to evaluate the performance of the different deconvolution algorithms and CpG selection strategies. The first benchmark dataset (BenchmarkData1) consisted of 12 *in silico* artificially reconstructed mixtures of 6 immune cell types from the FlowSorted.Blood.EPIC package. The second benchmark dataset (BenchmarkData2) consisted of the remaining 48 purified cfDNA samples from Moss *et al.* [40] that were not used in building the extended reference library.

Using Monte-Carlo simulation, we generated additional benchmark datasets by varying the proportions of non-immune cells. We considered two types of mixture models, namely (i) beta mixture and (ii) Gaussian mixture. We also considered five different ranges of non-immune proportions, i.e. 0, 0.1–0.2, 0.2–0.5, 0.5–0.8 and 0.8–0.9. For certain applications, cell types that were present in very small fractions may be of interest. For example, the tumor infiltrating lymphocytes which had prognostic value [42] were present in low fractions in some cancer types [43]. Following Zhang *et al.* [28], we extended our simulation study by considering rare non-immune proportions, i.e. rare cell type setting 0.01, 0.03, 0.05, 0.07, 0.1; and very rare cell type setting 0.001, 0.003, 0.005, 0.007 and 0.01. Details of the simulation setup were provided in Supplementary Appendix S2 (see Supplementary Data available online at https://academic.oup.com/bib). We compared the performance of the different deconvolution strategies via (i) the Spearman correlation coefficients between the estimated cell type proportions $\hat{p}_k$ versus the true proportions p_k, (ii) the root mean square error (RMSE), where $\text{RMSE} = \sqrt{\sum_{k=1}^{K}(\hat{p}_k - p_k)^2}$ and (iii) the symmetric mean absolute percentage error (sMAPE), where $\text{sMAPE} = \frac{1}{K}\sum_{k=1}^{K}\frac{|\hat{p}_k - p_k|}{\hat{p}_k + p_k}$.

We reported the performance metrics for within cell type (in this case, *k* denoted sample k) and for within sample comparisons (in this case, *k* denoted cell type k). Pathway analysis was also performed to identify significantly over-represented Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene ontology (GO) gene sets at false discovery rate (FDR) < 0.05 among the selected cell-specific CpGs via the function methylgometh in package methylGSA [44].

Due to the large overlap of CpGs (i.e. 452 567 common CpGs) between 450 k and EPIC arrays, we also evaluated the accuracy of the cell proportions estimation by borrowing the information from the existing 450 k reference library for scenarios in which we did not have an EPIC reference library. That is, CpG selection and the average reference profiles were performed on the 450 k reference library and incorporated into cell type deconvolution of EPIC arrays. The six matched 450 k and EPIC blood, as well as saliva samples of Braun *et al.* [45] (GEO accession number GSE111165), enabled us to use the estimated cell proportions from 450 k array as gold standard to validate the deconvolution of EPIC methylation array.

Data preprocessing

All the methylation data in this study, accessible from GEO or Bioconductor packages, were available in the IDAT format. We preprocessed the data using preprocessNoob from the minfi package [46], followed by the function getBeta to obtain the Beta value matrix. All subsequent analyses and deconvolution were performed on the methylation Beta value matrix.

Tool

The different deconvolution algorithms, CpG selection, reference library extension as well as the simulation procedures were included in our computational deconvolution tool R package methylDeConv, along with the code to perform the analyses in this study available at (https://github.com/jysonganan/methylDeConv).

Results

Comparison of deconvolution algorithms and CpG selection methods

The following CpG selection methods exhibited high correlation coefficients (>0.9), low RMSE and low sMAPE between estimated and true proportions: oneVsAlltest, onevsAllLimma, pairwiseGlmnet, multiGlmnet, glmnetpreselect and Rfepreselect across the five reference-based deconvolution algorithms (i.e. Houseman, RPC, CBS, MethylResolver and ARIC) in BenchmarkData1 (Figure 2). On the other hand, the two reference-free algorithms (i.e. RefFreeEWAS and TOAST-csDeconv) had poorer performance than the reference-based methods. Rfepreselect and RFpreselect were two-step CpG selection methods which utilized multi-class random forest modeling in the second step (Table 2). Rfepreselect showed a better performance compared with RFpreselect, which was consistent with the fact that Rfepreselect searched for the optimal subset of CpGs, whereas RFpreselect selected the top CpGs without taking into account multicollinearity among the selected CpGs. However, pairwiseGlmnet, multiGlmnet and Rfepreselect incurred

longer computational time; thus, we recommended oneVsAlltest, oneVsAllLimma or glmnetpreselect as CpG selection method. The CpGs selected by these three methods showed that a large degree of overlap (Figure 3). Similarly, our simulation studies also showed that MethylResolver required significantly longer computational time and TOAST-csDeconv was the slowest among the seven methods for deconvolution of large number of mixture samples. We summarized the key features and findings of the different deconvolution algorithms in Table 1.

A sensitivity analysis was conducted to evaluate the effect of the number of top-ranking CpGs included in the deconvolution algorithms. Results showed that retaining top 100 CpGs per cell type yielded higher correlation coefficients between estimated and true proportions compared with settings which retained top 50,150 or 200 CpGs (Supplementary Figure S1, see Supplementary Data available online at https://academic.oup.com/bib).

The CpGs selected by oneVsAlltest, glmnetpreselect as well as the overlapping CpGs between these two methods were enriched in gene sets and pathways associated with immune response and cell differentiation, consistent with existing literatures [18, 47] (Supplementary Table S4, see Supplementary Data available online at https://academic.oup.com/bib).

Cell proportions estimation using 450 k reference library

Within each cell type, the top 100 hypermethylated (hypomethylated) CpGs from 450 k reference library were ranked highly (lowly) in EPIC reference library with median rank of 229 (865836), indicating that cell-type-specific CpGs were consistent across the two arrays. Hypergeometric tests also showed a significant overlap between the top 100 CpGs from 450 k reference library and top 100 CpGs from EPIC reference library (P-values $< 10^{-10}$). Furthermore, using either 450 k or EPIC reference library yielded similar performance on BenchmarkData1 with average Spearman correlation coefficients >0.9, low RMSE and low sMAPE for the five reference-based methods (Figure 4), suggesting that one can rely on 450 k reference library for scenarios where EPIC reference library was unavailable to train the deconvolution algorithms for estimating cell proportions on EPIC arrays.

Validation of extended reference library

The low-dimensional t-SNE and heatmap visualization (Supplementary Figures S2-S4, see Supplementary Data available online at https://academic.oup.com/bib) of the extended EPIC reference library (six immune cell types plus epithelial and cfDNA) showed that the purified reference profiles were segregated into cell-specific clusters. As expected, epithelial cells were more distinct compared with the other immune cell types, whereas CD4T and CD8T were more similar to each other Application of the different deconvolution algorithms and CpG selection strategies on the 48 cfDNA samples in BenchmarkData2 using the extended EPIC reference library indicated that most of the algorithms and CpG selection strategies achieved high accuracy. The details were provided in Supplementary Appendix S3 (see Supplementary Data available online at https://academic.oup.com/bib).

Next, in scenarios where the DNA methylation profiles were obtained from non-cancer samples, we extended the reference library by only adding the epithelial cells. The

t-SNE and heatmap visualization showed that the different cell types remained well separated (Supplementary Figures S5-S7, see Supplementary Data available online at https://academic.oup.com/bib).

Notably, in BenchmarkData1 which consisted of only immune cell type mixtures, the extended reference library (Figure 5) remained robust and showed comparable performance to the EPIC reference library (Figure 1). Next, we evaluated whether extending the reference library significantly improved the estimation of immune cell proportions when the mixture profiles contained non-immune components. Results showed that as the proportions of non-immune cells increased, the deconvolution based on only immune cells reference library yielded inaccurate estimated proportions of immune cells. On the other hand, the reference-based deconvolution algorithms using extended reference library yielded accurate estimated proportions for the six immune cell types, epithelial and cfDNAs across all ranges of non-immune proportions (Supplementary Figures S8-S13, see Supplementary Data available online at https://academic.oup.com/bib). Among the reference-based deconvolution algorithms, MethylResolver showed the best performance when the proportion of non-immune cells were high, which is consistent with prior results that MethylResolver was robust to unknown content of cell mixtures [27]. On the other hand, RPC was computationally more efficient and showed comparable deconvolution performance to MethylResolver, whereas Houseman's algorithm had lower average correlation coefficient in Gaussian mixture simulations compared with other reference-based methods (Supplementary Figure S11, see Supplementary Data available online at https://academic.oup.com/bib). The two reference-free algorithms (i.e. RefFreeEWAS and TOAST-csDeconv) had poorer performance than the reference-based methods. Similar conclusions were obtained when we evaluated the effect of extended reference library on the six matched 450 k and EPIC blood, as well as saliva samples of Braun *et al.* [45], i.e. expanding the reference library by adding the epithelial cells was important for accurate deconvolution of cell types in non-blood tissues or tissues with high non-immune cell proportions. In terms of estimating rare cell type proportion, our simulation results showed that the reference-based methods were able to estimate the proportion of epithelial accurately in beta mixture simulation. However, in very rare epithelial Gaussian mixture simulation setting, Houseman and ARIC had lower correlation (Supplementary Figure S2l, see Supplementary Data available online at https://academic.oup.com/bib). Details were provided in Supplementary Appendix S4 (see Supplementary Data available online at https://academic.oup.com/bib).

Additional results including sensitivity analysis to evaluate the consistency of selected CpGs and performance of alternative cell proportion estimation strategies were provided in Supplementary Appendices S5 and S1 (see Supplementary Data available online at https://academic.oup.com/bib), respectively. We also studied the utility of a dual-net architecture of operator and selector [48], a deep learning algorithm for CpG selection in Supplementary Appendix S6 (see Supplementary Data available online at https://academic.oup.com/bib). Our results indicated that the reference-based deconvolution methods applied on the CpGs selected by the deep learning algorithm achieved comparable performance to the CpGs selected by oneVsAlltest with average Spearman correlation coefficients >0.9 (Supplementary Figure S24, see Supplementary Data available online at https://

academic.oup.com/bib). However, the dual-net deep network algorithm incurred a much longer computational time.

Case studies

Based on the deconvolution results using extended reference library, within sample and within cell type comparisons were both robust and reliable to the varying degree of non-immune proportions. RPC algorithm and CpG selection based on oneVsAlltest were advantageous over other combinations of deconvolution algorithms and CpG selection methods because of the faster computational time while maintaining a high accuracy. For the case studies, we performed the deconvolution using the RPC algorithm and CpGs selected from oneVsAlltest.

Early-stage breast cancer DNA methylation study

Our first case study involved an early-stage breast cancer DNA methylation study from Sehl *et al.* [49] (GEO accession number GSE140038). This study consisted of paired peripheral blood samples of 72 breast cancer patients collected at pre-treatment and post-treatment. Among these patients, 37 patients received radiation therapy-only, whereas 35 patients received radiation plus chemotherapy. We applied the deconvolution to estimate the cell type proportions. Among the patient samples who received radiation therapy-only treatment, we observed a significant decrease in the proportions of B-cells, CD4T and CD8T, and increase in neutrophils in post-treatment compared with the pre-treatment (Bonferroni adjusted $P < 0.05$ from paired sample *t*-tests) (Figure 6A). On the other hand, among the patients who received radiation plus chemotherapy, the proportion of monocyte was significantly higher in post-treatment samples (Figure 6B). For each patient and cell type, we computed the change in cell type proportion as the difference between post-treatment and pre-treatment. A linear regression was fitted to the change in cell type proportion as outcome and the therapy group as independent variable, adjusting for age at pre-treatment. Significant differences in the change of CD8T and monocytes cell proportions were observed between the radiation therapy versus radiation plus chemotherapy (Figure 6C). The observed changes in proportion of monocytes suggested that the combination of radiation and chemotherapy could indicate a compensatory mechanism to immunosuppression in post-treated breast cancer.

To illustrate the confounding effect of cellular composition, we performed an EWAS to identify DMCs between pre- and post-treatment among patients who received radiation plus chemotherapy. The beta values were transformed to M-values (M) using logit function. Due to the pairing structure, i.e. pre- and post-treatment methylation profiles per patient, we performed the EWAS on the difference, $M_{\text{post}} - M_{\text{pre}}$, adjusting for $\hat{p}_{k,\text{post}} - \hat{p}_{k,\text{pre}}$, where $\hat{p}_k$ was the estimated proportion of cell k. To avoid multicollinearity, i.e. the proportions summed to one, we excluded neutrophils from the adjustment. Without adjustment for the proportion of immune cells, 25 203 DMCs were identified at FDR < 0.05, whereas after adjustment the number of DMCs reduced to 7081, of which 3327 were in common. This suggested that an EWAS analysis without accounting for cellular composition could result in identification of a large number of false-positive DMCs.

To understand the biological mechanisms underlying the changes in DNA methylation associated with combination of radiation plus chemotherapy, we performed gene set analysis on the ranked list of CpGs from the model which adjusted for cellular composition using the methylglm function from methylGSA [44]. Both the GO [50] and KEGG canonical pathway [51] gene sets were tested. The minimum and maximum gene set sizes were 15 and 500. Three KEGG gene sets and 78 GO terms (71 in biological process [BP], 4 in cellular component [CC] and 3 in molecular function [MF]) were significantly enriched at FDR < 0.05 (Supplementary Table S5, see Supplementary Data available online at https://academic.oup.com/bib). The three KEGG gene sets were cytokine-cytokine receptor interaction, JAK-STAT signaling pathway and chemokine signaling pathway, respectively. The top GO terms included several immune response terms such as leukocyte proliferation and activation. We further clustered the significant BP-GO terms using REVIGO [52] to reduce functional redundancies by identifying representative terms. The interaction between the representative BP terms is shown in Figure 6D. Overall, the results showed that radiation and chemotherapy had significant impact on the immune system and circulating lymphocytes.

Melanoma DNA methylation study

Our second case study involved a study profiling 15 samples of desmoplastic melanoma and 15 samples of superficial malignant peripheral nerve tumor (MPNST) from Jour *et al.* [53] (GEO accession number GSE112308). All samples had high proportions of epithelial cells. MPNST had higher proportions of epithelial cells but lower proportions of CD4T, CD8T, monocytes and NK cells compared with desmoplastic melanoma, although they were not statistically significant after Bonferroni adjustment (Figure 7A).

We further performed EWAS to identify DMCs between desmoplastic melanoma and MPNST, and compared the results from models with and without cell proportion adjustment. A total of 46 579 and 131 550 DMCs were identified at FDR < 0.05 from the models with and without cell proportion adjustment, respectively. Similar to the first case study, the results showed that the confounding effect of cellular proportions needed to be accounted for appropriately to reduce detection of false-positive DMCs. Among the 46 579 DMCs from the model with cellular composition adjustment, 83.5% of them were hypermethylated in desmoplastic melanoma (Figure 7B).

To further illustrate the utility of the estimated cell proportions, we performed the cell-specific differential methylation analysis via CellDMC [21] to identify cell-specific DMCs. The number of hypermethylated and hypomethylated DMCs in MPNST at FDR < 0.05 within each cell type is given in Table 3. Specifically among the cell-specific DMCs, a higher proportion of CpGs were hypermethylated in MPNST relative to desmoplastic melanoma within B cell and NK, whereas the opposite pattern was observed in epithelial and monocytes. Gene set analysis on the ranked list of CpGs within each cell type identified 94, 12, 1, 75 and 1 GO terms at FDR < 0.05 in Bcell, CD8T, epithelial, neutrophils and NK cells, respectively (Supplementary Table S6, see Supplementary Data available online at https://academic.oup.com/bib). On the other hand, 2, 1, 1 and 3 KEGG gene sets were identified at FDR < 0.05 within Bcell, CD8T, epithelial, neutrophils and NK

cells, respectively (Supplementary Table S6, see Supplementary Data available online at https://academic.oup.com/bib). Of particular interest, the three KEGG gene sets significant within neutrophils were regulation of actin cytoskeleton, pathways in cancer and WNT signaling pathway. Previous study showed that infiltrating neutrophils were prognostic in melanoma [54], whereas WNT signaling played an important role in melanoma progression [55], suggesting that the differences between MPNST and desmoplastic melanoma could be attributed to the regulation of the WNT signaling pathways in neutrophils.

Since the DNA methylation was profiled in tumor which was a heterogeneous mixture of different cell types such as cancer, stromal and infiltrating immune cells, we also computed the tumor purity score using InfiniumPurify [56]. The estimated tumor purity score was positively correlated to the proportions of CD4T, CD8T, monocytes and NK cells (Figure 7C). EWAS with adjustment for tumor purity identified 49 275 DMCs at FDR < 0.05, of which 87% were hypermethylated in desmoplastic melanoma and 26 974 were in common with the earlier model that adjusted for immune and epithelial cell proportions. All 26 974 DMCs have consistent estimated effect size direction regardless of the types of adjustment. Among these DMCs, 88% were hypermethylated in desmoplastic melanoma. We performed gene set analysis on these 26 974 DMCs using the over-representation analysis approach via methylgometh [57] function implemented in methylGSA. None of the KEGG gene sets was identified, whereas 101 GO terms (79 in BP, 12 in CC and 10 in MF) were significantly enriched at FDR < 0.05 (Supplementary Table S7, see Supplementary Data available online at https://academic.oup.com/bib). The representative terms among these 79 BP were extracellular matrix organization, modulation of chemical synaptic transmission and homophilic cell adhesion via plasma membrane adhesion molecules (Figure 7D).

Discussion

In this paper, we compared several reference-based and reference-free deconvolution algorithms and CpG selection methods through extensive simulations and benchmark datasets with varying proportions of non-immune components. Among the different deconvolution algorithms and cell-type-specific CpG selection methods, we found that the RPC algorithm applied on CpGs selected by oneVsAlltest method yielded robust results and fast computational time. Our results also showed that the reference-based deconvolution algorithms using the extended reference library (i.e. adding epithelial and/or cfDNAs) is important in to obtain accurate estimates of cell proportions including rare cell types in non-blood samples. We anticipate that over time, additional methylation data on purified cell types (e.g. obtained via LCM or FACS) will become available to extend the reference library. In scenarios where one expects a high proportion of unknown cell mixtures (e.g. >0.8), we recommend using MethylResolver to obtain the estimated proportion of other known cell types.

Alternative CpG selection strategies that can be considered for future work include statistical methods for inferring DMRs [58] and deep learning algorithms [48]. Statistical methods for DMRs capitalize on the fact that nearby CpGs are correlated. Once a DMR is identified, one can either use the average methylation or the area of the bump or peak as a candidate feature. On the other hand, deep learning, a subfield of machine learning based on artificial

neural networks, is gaining popularity in scientific computing for accurate prediction and classification of large data sets in recent years [59]. In our current study which included fewer than 10 cell types, conventional differential testing method such as oneVsAlltest was able to select cell-specific CpGs that yielded accurate estimation of cell proportions. Thus, the deep learning algorithm did not offer significant improvement despite incurring a much longer computational time. However, we envisioned that as more purified cell types become available to extend the reference library, as well as the availability for higher resolution methylation platform, deep learning will potentially become a powerful class of algorithm for CpG selection.

A closely related issue in bulk DNA methylation data analysis of solid tumor samples is the heterogeneity within each tumor That is, a solid tumor usually consists of a mixture of cancer cells, stromal, adjacent normal cells and infiltrating lymphocytes. Several classes of methods for estimating tumor purity, i.e. the proportion of cancer cells in a tumor have been proposed. The most straightforward approach is based on matching somatic copy number alterations (CNAs) or single nucleotide variants (SNVs) [60-62]; however, this requires the availability of matching CNAs or SNVs. Another class of methods estimates the tumor purity as a function of stromal and immune cells infiltration scores [63, 64]; however, this approach is suboptimal because it ignores other cell types in tumor. In theory, if the purified cancer cell types are available, e.g. obtained via LCM, methylDeConv can be used to address the tumor heterogeneity issue by extending the reference library. However, due to the vast differences across multiple cancers, it is challenging to find a common purified cancer cell reference library for estimating the cancer cell proportions. The third class of approaches for estimating the tumor purity is by using normal samples [56, 62, 65]. As illustrated by Zheng *et al.* [56], one can utilize the normal samples from the large consortium as universal normal to estimate the tumor purity of different cancer types, rendering this approach versatile and cost-effective. As the estimated tumor purity usually has already accounted for the immune cell proportions implicitly, tumor purity and immune cell proportions tend to be correlated. Thus, we recommend to either adjust for tumor purity or immune cell proportions, but not both in the EWAS analysis. A sensitivity analysis should be carried out to compare the consistency of the identified DMCs.

Our computational framework methylDeConv is a unified analysis pipeline for the integrated deconvolution on methylation data which can be updated easily to incorporate additional purified cell types to further extend the reference library as they become available.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

National Institute for Occupational Safety and Health (NIOSH) awards U01OH011478 and U01OH012257 (PI: P.F.K.). The findings and conclusions presented in this article are those of the authors and do not represent the official position of NIOSH, the CDC or the U.S. Public Health Service.

Data availability

The data underlying this article are available in the Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/, and can be accessed with accession numbers GSE122126, GSE40699, GSE111165, GSE140038 and GSE112308.

References

1. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol 2013;25:571–8. [PubMed: 24148234]
2. Chen B, Khodadoust MS, Liu CL, et al. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol 2018;1711:243–59. [PubMed: 29344893]
3. Uryvaev A, Passhak M, Hershkovits D, et al. The role of tumor-infiltrating lymphocytes (TILs) as a predictive biomarker of response to anti-PD1 therapy in patients with metastatic non-small cell lung cancer or metastatic melanoma. Med Oncol 2018;35:1–9.
4. Chakravarthy A, Furness A, Joshi K, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun 2018;9:1–13. [PubMed: 29317637]
5. Zhong Y, Wan YW, Pang K, et al. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics 2013;14:89. [PubMed: 23497278]
6. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015;19:A68.
7. Hao X, Luo H, Krawczyk M, et al. DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci USA 2017;114:7414–9. [PubMed: 28652331]
8. Zhao Q, Shi X, Xie Y, et al. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform 2015;16:291–303. [PubMed: 24632304]
9. Guo Y, Sheng Q, Li J, et al. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS One 2013;8:e71462. [PubMed: 23977046]
10. Danaher P, Warren S, Lu R, et al. Pan-cancer adaptive immune resistance as defined by the Tumor Inflammation Signature (TIS): results from The Cancer Genome Atlas (TCGA). J Immunother Cancer 2018;6:1–17. [PubMed: 29298730]
11. Wang L, Saci A, Szabo PM, et al. EMT-and stroma-related gene expression and resistance to PD-1 blockade in urothelial cancer. Nat Commun 2018;9:1–12. [PubMed: 29317637]
12. Li Y, Kang K, Krahn JM, et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017;18:1–13. [PubMed: 28049423]
13. Moran S, Vizoso M, Martinez-Cardús A, et al. Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the Infinium HumanMethylation450 Microarray. Epigenetics 2014;9:829–33. [PubMed: 24732293]
14. Slieker RC, van Iterson M, Luijk R, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol 2016;17:1–13. [PubMed: 26753840]
15. Odintsova VV, Hagenbeek FA, Suderman M, et al. DNA methylation signatures of breastfeeding in buccal cells collected in mid childhood. Nutrients 2019;11:2804. [PubMed: 31744183]
16. Gao X, Jia M, Zhang Y, et al. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics 2015;7:113. [PubMed: 26478754]
17. Bauer M, Linsel G, Fink B, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics 2015;7:81. [PubMed: 26246861]
18. Lin X, Tan JYL, Teh AL, et al. Cell type-specific DNA methylation in neonatal cord tissue and cord blood: a 850K-reference panel and comparison of cell types. Epigenetics 2018;13:941–58. [PubMed: 30232931]

19. Hannon E, Dempster EL, Mansell G, et al. DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. Elife 2021;10:e58430. [PubMed: 33646943]
20. Rahmani E, Schweiger R, Rhead B, et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nat Commun 2019;10:1–11. [PubMed: 30602773]
21. Zheng SC, Breeze CE, Beck S, et al. Identification of differentially methylated cell types in epigenome-wide association studies. Nat Methods 2018;15:1059–66. [PubMed: 30504870]
22. Titus AJ, Gallimore RM, Salas LA, et al. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum Mol Genet 2017;26:R216–24. [PubMed: 28977446]
23. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012;13:86. [PubMed: 22568884]
24. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7. [PubMed: 25822800]
25. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 2016;48:1193–203. [PubMed: 27526324]
26. Teschendorff AE, Breeze CE, Zheng SC, et al. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics 2017;18:105. [PubMed: 28193155]
27. Arneson D, Yang X, Wang K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Commun Biol 2020;3:1–13. [PubMed: 31925316]
28. Zhang W, Xu H, Qiao R, et al. ARIC: accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data. Brief Bioinform 2022;23. https://academic.oup.com/bib/article-abstract/23/1/bbab362/6361035?redirectedFrom=fulltext.
29. Houseman EA, Kile ML, Christiani DC, et al. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics 2016;17:259. [PubMed: 27358049]
30. Rahmani E, Schweiger R, Shenhav L, et al. BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol 2018;19:141. [PubMed: 30241486]
31. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. Genome Biol 2019;20:190. [PubMed: 31484546]
32. Jaffe AE, Jaffe MAE. Package FlowSorted. Blood 2014;450k. https://bioconductor.org/packages/release/data/experiment/html/FlowSorted.Blood.450k.html.
33. Salas LA, Koestler D, Butler R, et al. FlowSorted. Blood. EPIC, dim 2018;575719:289. https://bioconductor.org/packages/release/data/experiment/html/FlowSorted.Blood.EPIC.html.
34. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics 2017;9:757–68. [PubMed: 28517979]
35. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 2014;15:1–9.
36. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 2012;7:e41361. [PubMed: 22848472]
37. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67:301–20.
38. Breiman L. Random forests. Mach Learn 2001;45:5–32.
39. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26. [PubMed: 27774042]
40. Moss J, Magenheim J, Neiman D, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun 2018;9:1–12. [PubMed: 29317637]

41. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 2018;46:D794–801. [PubMed: 29126249]
42. Gao G, Wang Z, Qu X, et al. Prognostic value of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: a systematic review and meta-analysis. BMC Cancer 2020;20:179. [PubMed: 32131780]
43. Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep 2018;23(181–193):e187.
44. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. Bioinformatics 2019;35:1958–9. [PubMed: 30346483]
45. Braun PR, Han S, Hing B, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. Transl Psychiatry 2019;9:1–10. [PubMed: 30664621]
46. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30:1363–9. [PubMed: 24478339]
47. Salas LA, Koestler DC, Butler RA, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol 2018;19:1–14. [PubMed: 29301551]
48. Wojtas M, Chen K. Feature importance rankingfor deep learning. Adv Neural Inform Process Syst 2020;33:5105–14.
49. Sehl ME, Carroll JE, Horvath S, et al. The acute effects of adjuvant radiation and chemotherapy on peripheral blood epigenetic age in early stage breast cancer patients. NPJ Breast Cancer 2020;6:1–5. [PubMed: 31934613]
50. Gene OC. The Gene Ontology (GO) project in 2006. Nucleic Acids Res 2006;34:D322–6. [PubMed: 16381878]
51. Ogata H, Goto S, Sato K, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34. [PubMed: 9847135]
52. Supek F, Bosnjak M, Skunca N, et al. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6:e21800. [PubMed: 21789182]
53. Jour G, Vasudevaraja V, Prieto VG, et al. BCAT1 and miR-2504: novel methylome signature distinguishes spindle/desmoplastic melanoma from superficial malignant peripheral nerve sheath tumor. Mod Pathol 2019;32:338–45. [PubMed: 30310175]
54. Jensen TO, Schmidt H, Moller HJ, et al. Intratumoral neutrophils and plasmacytoid dendritic cells indicate poor prognosis and are associated with pSTAT3 expression in AJCC stage I/II melanoma. Cancer 2012;118:2476–85. [PubMed: 21953023]
55. Gajos-Michniewicz A, Czyz M. WNT signaling in melanoma. Int J Mol Sci 2020;21:4852. [PubMed: 32659938]
56. Zheng X, Zhang N, Wu HJ, et al. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome Biol 2017;18:17. [PubMed: 28122605]
57. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics 2016;32:286–8. [PubMed: 26424855]
58. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 2012;41:200–9. [PubMed: 22422453]
59. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. [PubMed: 26017442]
60. Su X, Zhang L, Zhang J, et al. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics 2012;28:2265–6. [PubMed: 22743227]
61. Gusnanto A, Wood HM, Pawitan Y, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics 2012;28:40–7. [PubMed: 22039209]
62. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012;30:413–21. [PubMed: 22544022]
63. Yoshihara K, Shahmoradgoli M, Martinez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013;4:2612. [PubMed: 24113773]

64. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971. [PubMed: 26634437]

65. Zhang N, Wu HJ, Zhang W, et al. Predicting tumor purity from methylation microarray data. Bioinformatics 2015;31:3401–5. [PubMed: 26112293]

66. Barrell D, Dimmer E, Huntley RP, et al. The GOA database in 2009–an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009;37:D396–403. [PubMed: 18957448]

Key Points

- Accurate estimation of cellular composition is an important step in EWAS analysis using the Illumina DNA Methylation BeadArrays.
- In this work, we conducted a systematic assessment of various computational techniques for cellular deconvolution.
- Results from simulations and real benchmark datasets indicated that using an appropriate extended reference library is critical for accurate estimation of cellular composition.
- We provided a software methylDeConv which offered a unified framework by integrating several deconvolution algorithms and different cell-type-specific CpG selection methods to construct the extended reference library.

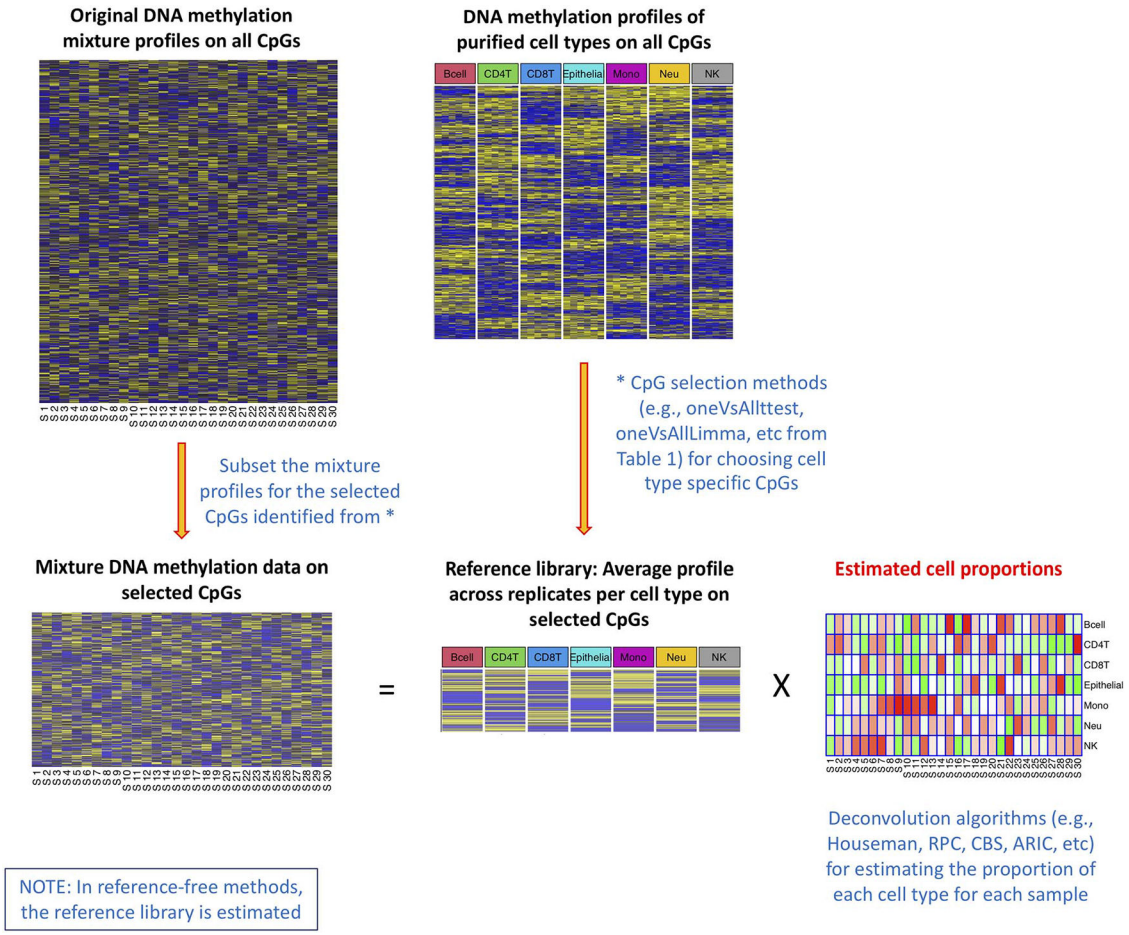

Figure 1.

Schematic plot illustrating the workflow of cell type deconvolution for estimating the cell proportions from the mixture profiles.

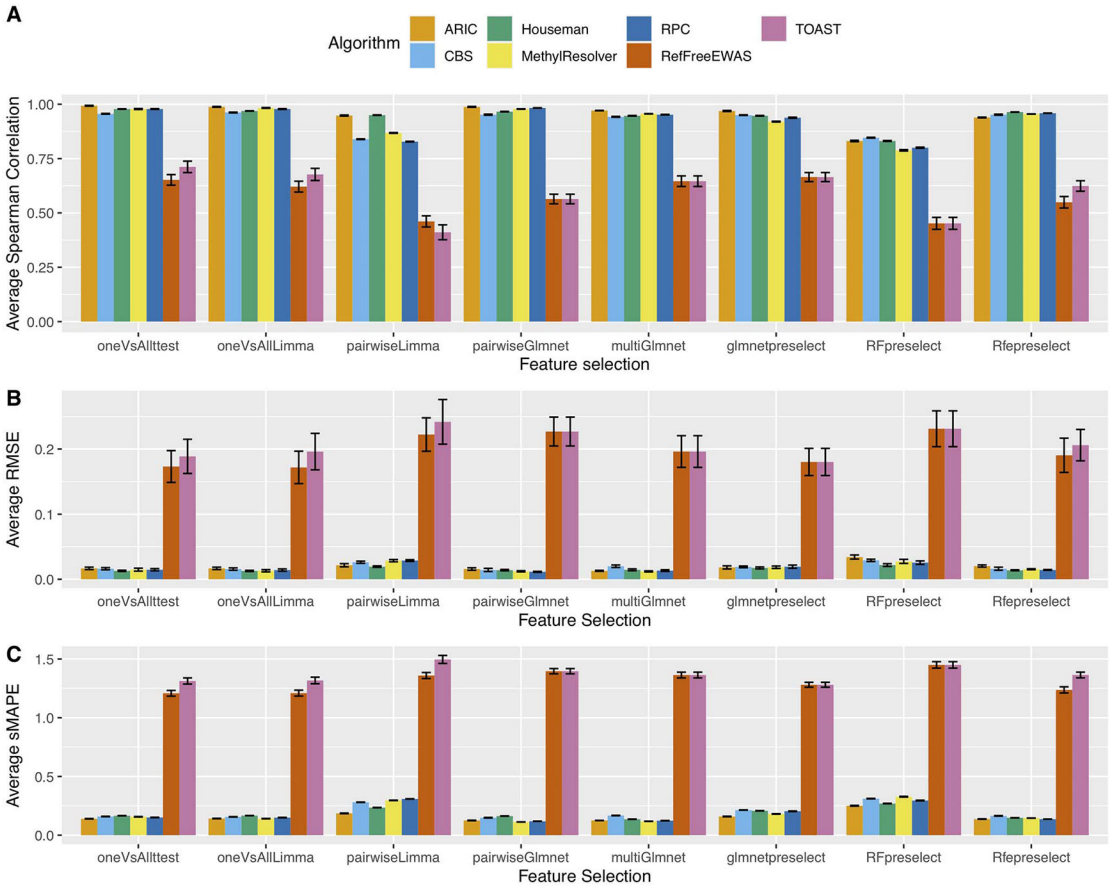

Figure 2.

(**A**) Average spearman correlation coefficients. (**B**) Average RMSE. (**C**) Average sMAPE, along with the standard deviation error bars between estimated and true proportions within each of the 12 samples in BenchmarkData1 for different CpG selection and deconvolutional algorithms using the FlowSorted.Blood.EPIC reference library (six immune cell types).

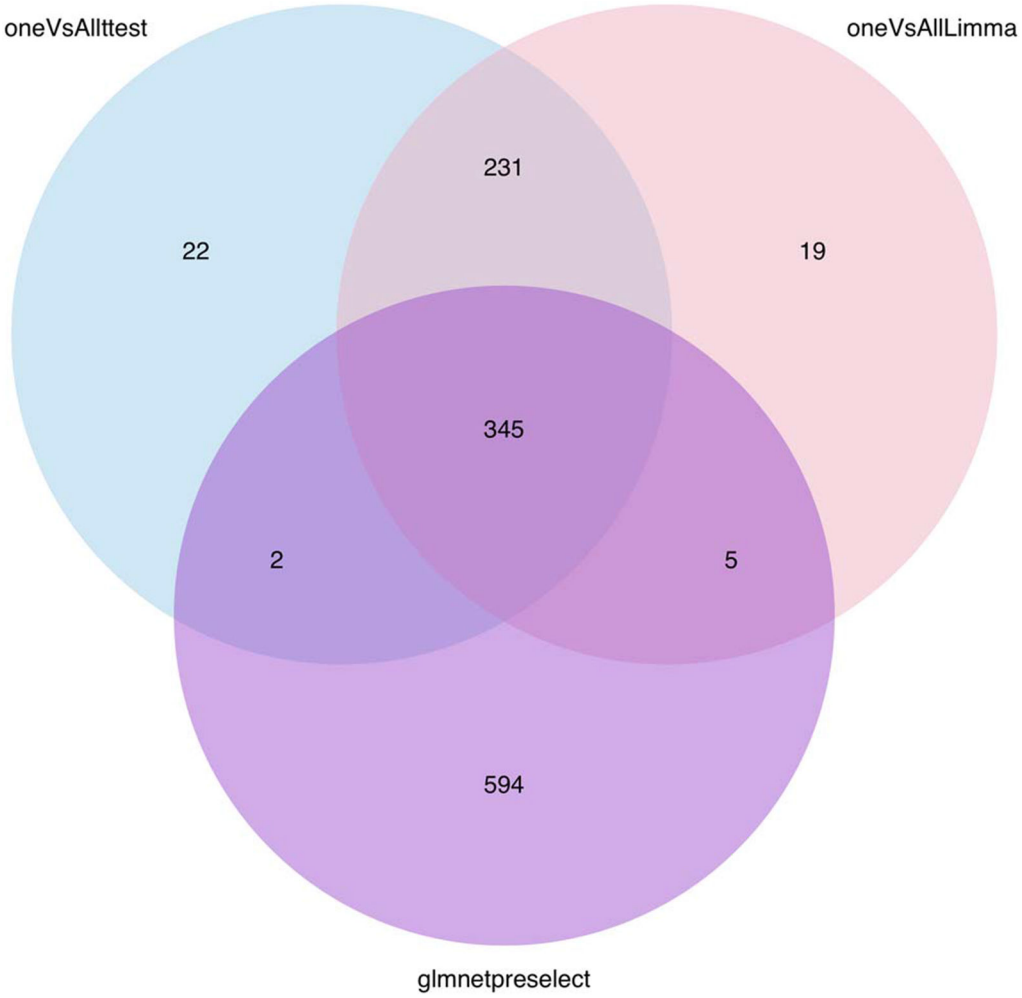

Figure 3.
Venn diagram comparing the CpGs selected by oneVsAlltest, oneVsAllLimma and glmnetpreselect onFlowSorted.Blood.EPIC reference library.

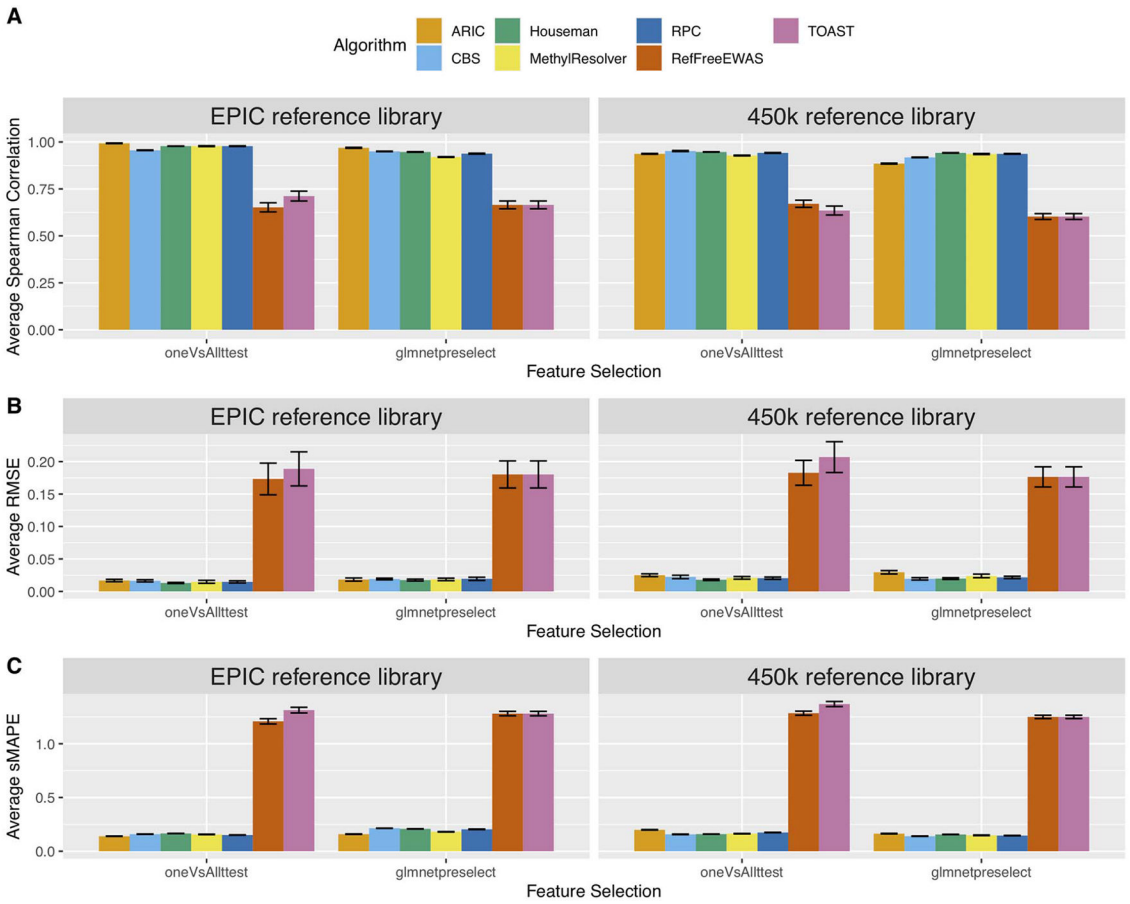

Figure 4.

(**A**) Average Spearman correlation coefficients. (**B**) Average RMSE. (**C**) Average sMAPE, along with the standard deviation error bars between estimated and true proportions within each of the 12 samples in BenchmarkData1 for oneVsAlltest and glmnetpreselect across different deconvolutional algorithms using the FlowSorted.Blood.EPIC reference library (six immune cell types, left panel) and FlowSorted.Blood.450 k reference library (right panel).

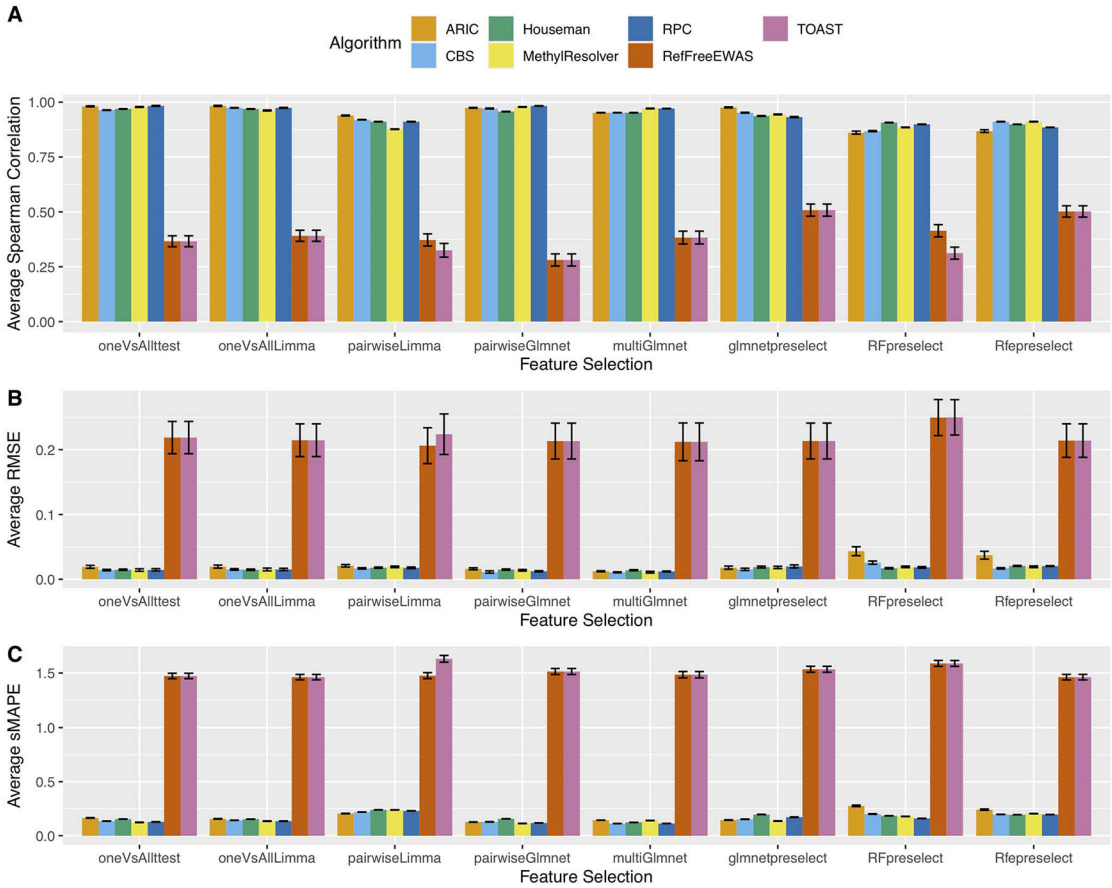

Figure 5.

(**A**) Average Spearman correlation coefficients. (**B**) Average RMSE. (**C**) Average sMAPE, along with the standard deviation error bars between estimated and true proportions within each of the 12 samples in BenchmarkData1 for different CpG selection and reference-based deconvolutional algorithms using the extended reference library (six immune cell types plus epithelial).

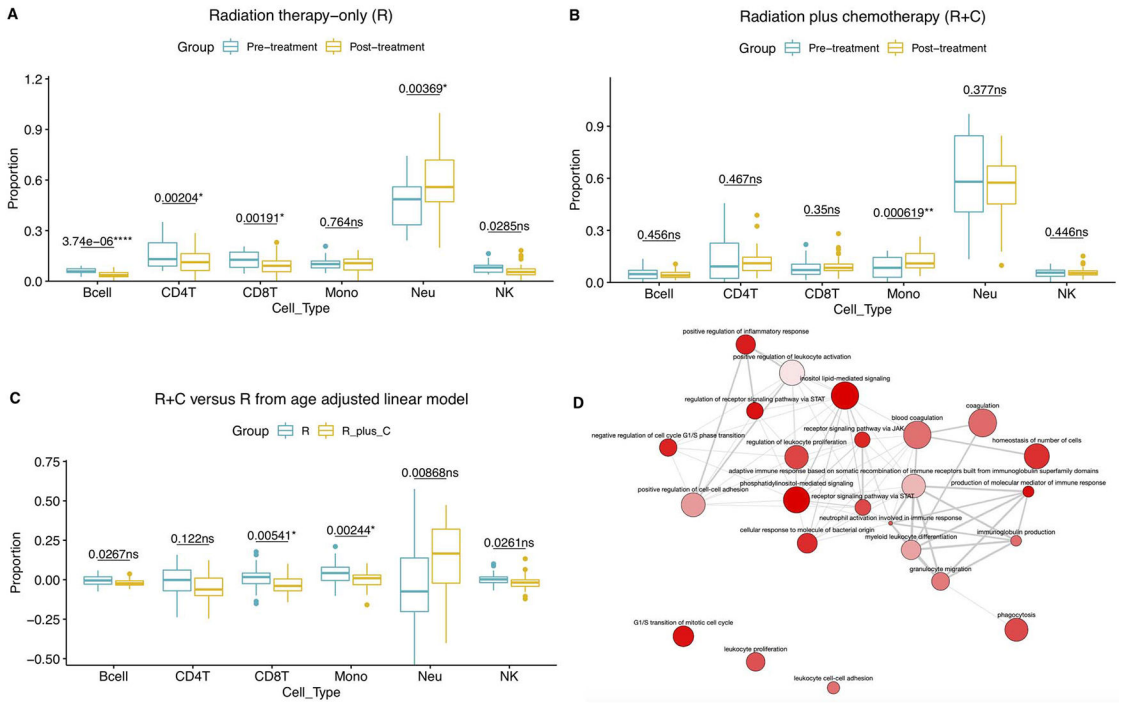

Figure 6.

(**A**–**C**) Comparison of the estimated proportions of cell types from the breast cancer DNA methylation case study, raw *P*-values from paired sample *t*-tests (**A**,**B**) and age-adjusted linear model (**C**) were printed above the boxplots for each cell type, * denoted Bonferroni adjusted *P*-value <0.05. (**A**) Pre- versus post-treatment in patients who received radiation therapy-only treatment (R). (**B**) Pre- versus post-treatment in patients who received radiation plus chemotherapy treatment (R + C). (**C**) R + C versus R from age adjusted linear model. (**D**) Interactive graph view of the representative BP terms from EWAS of pre- versus post-treatment in patients who received R + C. Bubble size indicates the frequency of the GO term in the underlying GOA [66] database, whereas the line width indicates the degree of similarity.

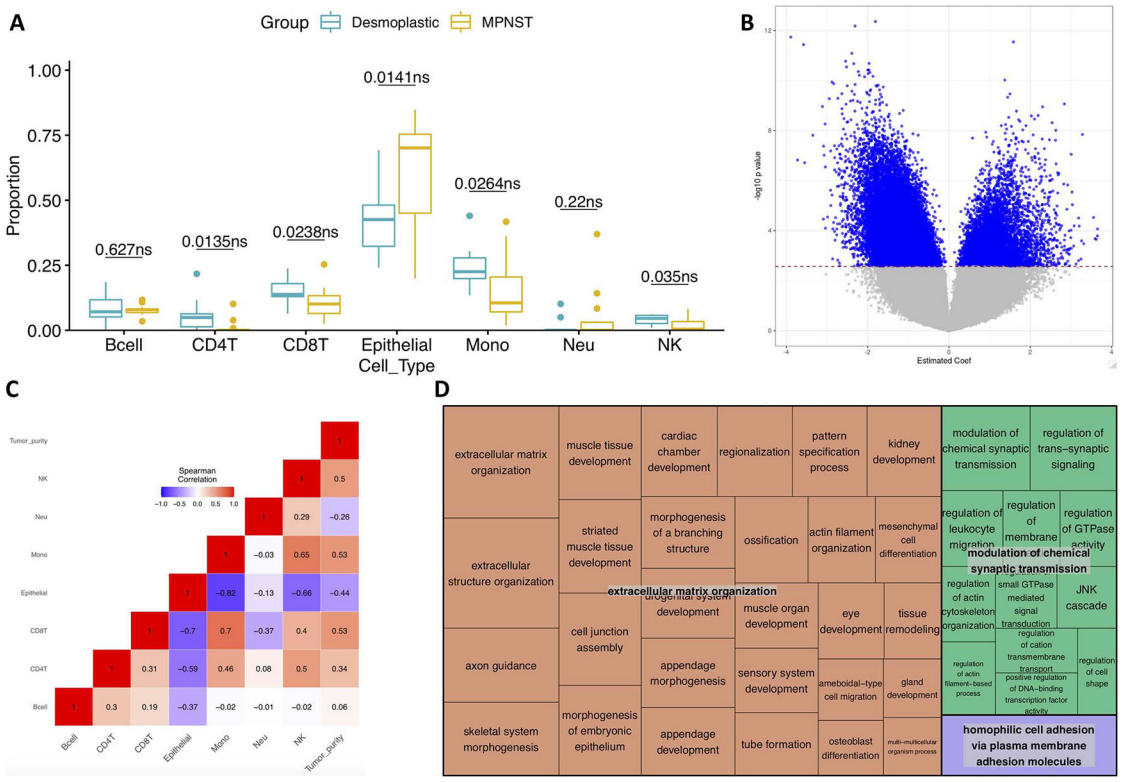

Figure 7.

(**A**) Comparison of the estimated proportions of cell types from the melanoma DNA methylation case study, raw *P*-values from linear model from two sample *t*-tests were printed above the boxplots for each cell type. (**B**) Volcano plot from EWAS comparing desmoplastic melanoma to MPNST. Blue dots denote DMCs at FDR < 0.05. (**C**) Spearman correlation matrix plot of the estimated cell proportions and tumor purity score. (**D**) TreeMap depicting the clustering of BP terms from GO analysis on the common DMCs. Representative terms are joined into clusters of related terms, denoted with different colors.

Table 1.

Description of the different deconvolution algorithms

Deconvolution algorithms	Description	Remark
1. Houseman's QP/CP [23]	Reference-based method using linear CP. Implemented in minfi [46] and EpiDISH [26] R packages.	One of the earliest method for Illumina methylation arrays deconvolution.
2. EpiDISH [26]	Reference-based method using RPC. Also implements several reference-based deconvolution algorithms including CBS [24] and Houseman's QP/CP in the R package.	RPC algorithm is computationally fast.
3. MethylCIBERSORT [4]	Reference-based method using CBS algorithm. Implemented as an R package.	Only version 0.2.0 is publicly available on Zenodo.
4. MethylResolver [27]	Reference-based method using LTS regression. Implemented as an R package.	Robust to unknown content, however, incurs longer computational time for large datasets.
4. ARIC [28]	Reference-based method using weighted SVR. Implemented in Python.	Computationally fast.
5. RefFreeEWAS [29]	Reference-free method, extension of Houseman's QP/CP [23]. Implemented as an R package. Also implemented in TOAST [31] R package.	Computationally fast.
6. TOAST-csDeconv [31]	Reference-free method by improving CpG selection, followed by deconvolution using other reference-free algorithm such as RefFreeEWAS [29]. Implemented as an R package.	High memory usage and incurs long computational time for large datasets.

Table 2.

Description of the different cell-specific CpG selection methods

CpG/feature selection methods	Description	Remark
1. oneVsAlltest	Select top 100 CpGs using one-versus-all t-test per cell type.	Computationally fastest among the eight methods. Combined with RPC, they yield robust and accurate estimation from our simulation studies.
2. oneVsAllLimma	Select top 100 CpGs using one-versus-all moderated t-test per cell type.	Computationally fast.
3. pairwiseLimma	Select top 100 CpGs using pairwise moderated t-test per cell type.	
4. pairwiseGlmnet	Select CpGs with non-zero coefficients from pairwise two-class elastic net model.	Incurs longer computational time.
5. multiGlmnet	Select CpGs with non-zero coefficients from multi-class elastic net model.	Incurs longer computational time.
6. glmnetpreselect	Step 1. Preselect top 300 CpGs from one-versus-all *t*-test per cell type. Step 2. Select CpGs with non-zero coefficients from multi-class elastic net model fitted with preselected CpGs.	
7. RFpreselect	Step 1. Preselect top 300 CpGs from one-versus-all *t*-test per cell type. Step 2. Select the most important variables (top 100 CpGs per cell type) from multi-class random forest model fitted with preselected CpGs.	
8. Rfepreselect	Step 1. Preselect top 300 CpGs from one-versus-all *t*-test per cell type. Step 2. Select CpGs based on RFE algorithm applied to the multi-class random forest model fitted with preselected CpGs.	Incurs longer computational time.

Brief Bioinform. Author manuscript; available in PMC 2023 November 19.

Table 3.

Number of hypermethylated and hypomethylated DMCs in MPNST within each cell type

Cell type	Number of hypomethylated CpGs	Number of hypermethylated CpGs
Bcell	1166	4141
CD4T	28	2
CD8T	0	0
Epithelial	979	324
Monocytes	507	29
Neutrophils	7	2
NK	391	4693