Genome Biology

RESEARCH

Open Access

Systematic evaluation of methylation-based cell type deconvolution methods for plasma cell-free DNA

Tongyue Sun[1†], Jinqi Yuan[1†], Yacheng Zhu[1], Jingqi Li[1], Shen Yang[1], Junpeng Zhou[1], Xinzhou Ge[2], Susu Qu[3], Wei Li[4*], Jingyi Jessica Li[5,6,7,8,9*] and Yumei Li[1*]

†Tongyue Sun and Jinqi Yuan contributed equally to this work.

*Correspondence:
wei.li@uci.edu; lijy03@g.ucla.edu; ymli12@suda.edu.cn

[1] School of Basic Medical Sciences, Suzhou Medical College, Soochow University, Suzhou 215123, China
[2] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA
[3] Chinese Institute for Brain Research, Beijing 102206, China
[4] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[5] Department of Statistics and Data Science, University of California, Los Angeles, CA 90095, USA
[6] Interdepartmental Program in Bioinformatics, University of California, Los Angeles, CA 90095, USA
[7] Department of Human Genetics, University of California, Los Angeles, CA 90095, USA
[8] Department of Computational Medicine, University of California, Los Angeles, CA 90095, USA
[9] Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

Abstract

Background: Plasma cell-free DNA (cfDNA) is derived from cellular death in various tissues. Investigating the tissue origin of cfDNA through cell type deconvolution, we can detect changes in tissue homeostasis that occur during disease progression or in response to treatment. Consequently, cfDNA has emerged as a valuable noninvasive biomarker for disease detection and treatment monitoring. Although there are many methylation-based methods for cfDNA cell type deconvolution, a comprehensive and systematic evaluation of these methods has yet to be conducted.

Results: In this study, we benchmark five methods: MethAtlas, cfNOMe toolkit, CelFiE, CelFEER, and UXM. Utilizing deep whole-genome bisulfite sequencing data from 35 human cell types, we generate in silico cfDNA samples with ground truth cell type proportions to assess the deconvolution performance of the five methods under multiple scenarios. Our findings indicate that multiple factors, including reference marker selection, sequencing depth, and reference atlas completeness, jointly influence the deconvolution performance. Notably, an incomplete reference with missing markers or cell types leads to suboptimal results. We observe performance differences among methods under varying conditions, underscoring the importance of tailoring cfDNA deconvolution analyses. To increase the clinical relevance of our findings, we further evaluate each method's performance in potential clinical applications using real-world datasets.

Conclusions: Based on the benchmark results, we propose general guidelines to choose the suitable methods based on sequencing depth of the cfDNA data and completeness of the reference atlas to maximize the performance of methylation-based cfDNA cell type deconvolution.

Keywords: Benchmark, Cell-free DNA, DNA methylation, Deconvolution

Sun *et al. Genome Biology*   (2024) 25:318

Page 2 of 26

Background

Plasma cell-free DNA (cfDNA) predominantly consists of double-stranded DNA fragments released by dying cells from various tissues [1–3]. In healthy individuals, plasma cfDNA is primarily derived from apoptosis of normal hematopoietic cells, with minimal contributions from other cell types [4–6]. However, the progression and treatment of numerous diseases, including cancers [7–9], can impact the cell-type-derived cfDNA proportions. Consequently, an abnormal proportion distribution can indicate altered tissue homeostasis resulted from disease progression or treatment. As a result, cfDNA has emerged as a valuable noninvasive biomarker for disease detection and treatment monitoring [10–12]. Elucidating the cell-type-derived cfDNA proportions in patients will contribute to the discovery of noninvasive biomarkers and enhance our understanding of disease development. Cell type deconvolution is the analysis that estimates the proportions of cell types presenting in cfDNA.

While each cell possesses nearly identical DNA sequence, DNA methylation signatures exhibit cell type specificity, providing a means to trace back to the cell type origin of cfDNA [6, 13, 14]. The most prevalent form of DNA methylation is 5-methylcytosine (5mC), which predominantly occurs at CpG sites in humans [15–17]. Due to its high base resolution, whole-genome bisulfite sequencing (WGBS) has become the standard method for capturing genome-wide DNA methylation profiles. Several studies have introduced methylation-based deconvolution methods to estimate the proportions of cell types in cfDNA [5, 6, 18–20]. MethAtlas [5] utilizes methylation ratio of a CpG site as the methylation metric for deconvolution, initially developed for array methylomes but adapted for methylation sequencing data analysis [21–23]. It models the plasma cfDNA methylation ratio profile as a linear combination of the methylation ratio profiles of cell types in a reference atlas. Subsequently, the relative contribution of each cell type to cfDNA is determined using the non-negative least squares (NNLS) linear regression. cfNOMe toolkit [19] follows a similar approach to MethAtlas but utilizes linear least squares linear regression to estimate each cell type's relative contribution. CelFiE [18] utilizes the number of methylated and all reads at CpG sites as input for deconvolution. Then, it uses an expectation–maximization (EM) algorithm to estimate the parameters of a Bayesian mixture model including the cell type proportions. The method CelFiE is notably for being able to estimate the contributions from multiple unknown cell types not available in the reference atlas. CelFEER [20], a method adapted from CelFiE, uses essentially the same model as CelFiE but with read averages (the ratio of methylated CpG sites to the total number of CpG sites on a read) as input. UXM [6] is a fragment-level deconvolution method using the percentage of unmethylated fragments in a genomic region as the methylation metric. Unmethylated fragments are defined as fragments in which no more than 25% of CpGs as methylated. Then, NNLS is applied to infer the relative contribution of each cell type to cfDNA.

Nevertheless, a comprehensive assessment of the performance of these methods has not yet been conducted. Although a preprint manuscript assessed the performance of two reference-based deconvolution methods, MethAtlas and cfDNAme, it utilized artificial DNA mixtures consisting of only two cell types [23], and its method coverage was incomplete: methods, such as CelFiE and CelFEER et al., were not included in the assessment, and cfDNAme method was not a standalone program for cfDNA deconvolution

[24]. Furthermore, it focused on optimizing the number of reference markers and sequencing depths for deconvolution. Additionally, each method necessitates specific preprocessing procedures and relies on distinct reference methylation atlases. The potential impact of these procedures on deconvolution results remains unknown. To bridge this knowledge gap, we conducted a thorough comparison and evaluation of the above five peer-reviewed methylation-based cfDNA cell type deconvolution methods with publicly available and executable standalone programs, namely MethAtlas, cfNOMe (referred to as an abbreviation for cfNOMe toolkit in this study), CelFiE, Cel-FEER, and UXM (Fig. 1 and Additional file 1: Table S1). Our evaluation includes their performance in terms of reference marker selection and deconvolution accuracy under various scenarios, utilizing in silico mixtures of cell types' WGBS data with ground truth cell type proportions. To enhance the relevance of our findings for clinical applications, we also included two real-world datasets from diseased patients [25, 26], providing direct evidence of the methods' effectiveness in actual clinical settings. Based on the evaluation results, we summarize the strengths and weaknesses of the five methods, providing general guidelines for users to choose the suitable method based on the sequencing depth of their cfDNA data and the completeness of the reference atlas.

Results

Benchmark study design

Our benchmark study utilized a recently published DNA methylation atlas of normal human cell types generated by deep WGBS [6]. As shown in Fig. 1, from this atlas, we selected 35 cell types, totaling 182 samples, each with at least two biological replicates. To ensure independent reference and testing datasets, we randomly divided all samples into two halves: one for generating the reference methylation atlas and the other for creating in silico cfDNA samples. The reference methylation atlas was generated following the procedures specified by each deconvolution method. To assess the impact of cell type proportion distribution, we created three in silico cfDNA datasets ($n = 100$ in silico cfDNA samples in each dataset) using the WGBS data of 35 cell types: (1) a uniform distribution dataset, where the cell type proportions of each sample were first

(See figure on next page.)

Fig. 1 Schematic overview of the benchmark study design. In this benchmark study, WGBS data from 182 samples representing 35 cell types is initially randomly divided into two halves. One half is designated for the creation of the reference methylation atlas, while the other is utilized for generating in silico cfDNA samples. Ground truth cell type proportions are then generated using either a uniform distribution, Dirichlet distribution, or a constrained random distribution with blood cells as the primary cell types. Subsequently, the deconvolution performance is rigorously assessed under various influencing factors, including reference marker selection, sequencing depth, and reference completeness. Five evaluation metrics, root-mean-square error (RMSE), Pearson's correlation coefficient, Spearman's rank correlation, Lin's concordance correlation coefficient (CCC), and Jensen–Shannon divergence (JSD), are employed to scrutinize the accuracy of predicted proportions (P_p) against ground truth proportions (P_g). Additionally, two real-world datasets including cfDNA samples from both patients and controls were included to evaluate performance on real clinical applications. Reference methylation atlas generated from all the 182 samples were used for the deconvolution of real-world datasets. Two metrics were applied to evaluate the deconvolution performance in these datasets: (1) the statistical difference in the cfDNA fraction of affected tissues between diseased and healthy individuals and (2) the ROC-AUC of machine learning models for disease detection based on the estimated fractions of all cell types

Fig. 1 (See legend on previous page.)

independently sampled from the Uniform(0,1) distribution and then normalized to sum up to 1 (Additional file 1: Table S2 and the "Methods" section). This resulted in homogeneous proportion distributions allowing us to conduct unbiased performance comparisons among cell types. (2) A constrained random distribution dataset, where all samples

had blood cells as the major cell types to mimic the cfDNA in blood samples of healthy individuals (Additional file 1: Table S3). For each sample in this dataset, the cell type proportions were generated in the following steps (the "Methods" section): (1) in addition to five immune cell types, we randomly selected another 1–10 cell types as the contributing cell types; (2) we generated the cell type proportions by independently sampling floating numbers from a random distribution and then normalized the proportions to sum up to 1; (3) we randomly assigned the five largest proportion values to the five immune cell types and the remaining proportion values to the other cell types. (3) A Dirichlet distribution dataset, where the cell type proportions of each sample were independently sampled from the Dirichlet distribution to provide moderately diverse mixtures with varying contributions from different cell types (Additional file 1: Table S4 and Additional file 2: Fig. S1). With the cell type proportions determined for each sample in the three datasets, we generated the in silico samples by sampling sequencing fragments from each cell type based on the proportion. We used the cell type proportions in these in silico cfDNA samples as the ground truths for method evaluation.

We assessed the performance of the deconvolution methods using five metrics that compare each method's predicted cell type proportions with ground truth proportions used to generate the in silico cfDNA samples: (1) root-mean-square error (RMSE), which captures the absolute error between true and predicted values; (2) Pearson's correlation coefficient, which assesses overall agreement, with high sensitivity to the magnitude of variation; (3) Spearman's rank correlation, a metric less sensitive to outliers and better suited for capturing rank-based relationships; (4) Lin's concordance correlation coefficient (CCC), which balances both precision and accuracy in quantifying agreement between estimated and actual values; and (5) Jensen–Shannon divergence (JSD), which measures the similarity between predicted and actual cell type fraction distributions, making it especially useful for datasets with widely varying cell type proportions. These metrics were compiled into a composite measure, the normalized performance score, which integrates the min–max normalized values of all five metrics (the "Methods" section). For researchers interested in the individual metrics, we provide detailed results in the supplementary data. Additionally, we comprehensively evaluated the influence of multiple factors on the deconvolution result, including reference marker selection, sequencing depth, and reference atlas completeness (Fig. 1).

Furthermore, to evaluate performance on real clinical applications, we introduced two real-world datasets including cfDNA samples from both patients and controls. Given the absence of ground truth for these real-world datasets, we leveraged established knowledge from the literature to evaluate the performance of various methods (Fig. 1). One metric we used is the statistical difference in the cfDNA fraction derived from affected tissues between diseased and healthy individuals. Another metric is the area under the receiver operating characteristic curve (ROC-AUC), which evaluates the accuracy of disease detection based on the estimated fractions of all cell types from each deconvolution method.

Evaluation of deconvolution accuracy and computing resource requirements
To begin with, we assessed the overall performance of each deconvolution method using their default settings. For the uniform distribution and Dirichlet distribution datasets,

Sun *et al. Genome Biology*   (2024) 25:318

Page 6 of 26

CelFEER outperformed others, achieving exceptionally low median RMSE values of 0.0099 and 0.014 respectively, with UXM as the next best performer (Fig. 2A and Additional file 2: Fig. S2). Notably, CelFEER demonstrated lower standard deviations across the 100 cfDNA samples, indicating its superior stability compared to other methods (Additional file 1: Table S5). For the constrained random distribution dataset, UXM led in performance, outperforming the other four methods, with CelFEER and CelFiE following (Fig. 2A and Additional file 2: Fig. S2). Nevertheless, MethAtlas and cfNOMe consistently ranked lowest for at least one of the three datasets. These findings underscore the influence of cfDNA sample composition on deconvolution performance.

Given that sequencing depth significantly impact the analysis cost, we next investigated this by varying the sequencing depth of the samples being analyzed. We generated additional in silico cfDNA datasets with sample sequencing depths of 1, 5, 10, and 15, alongside our existing datasets of 22 (the lowest sequencing depth among data used to generate the in silico samples). Results indicated diminished performance across all methods for low-depth samples (Fig. 2B). Notably, cfNOMe struggled at a sequencing depth of 1, showing the poorest performance across all datasets. While other methods performed relatively better than cfNOMe at this depth, they still exhibited marked performance declines (Additional file 2: Fig. S3). These findings underscore the challenges all methods face in low-depth scenarios, particularly for samples with sequencing depths below 5 (Fig. 2B and Additional file 2: Fig. S3).

Turning to computing resource requirements, UXM and MethAtlas significantly outpaced the other three methods in terms of running time for analyzing a set of 100 cfDNA samples (Fig. 2C). Unsurprisingly, MethAtlas, with the most straightforward optimization problem, was the fastest, completing the analysis of 100 samples in under 30 s. Conversely, cfNOMe was the slowest, requiring an average of 1.4 min to analyze one sample. In terms of memory usage, MethAtlas had the highest memory requirements, followed by cfNOMe (Fig. 2D). CelFiE, on the other hand, exhibited minimal memory usage as low as 585 MB, making it suitable for execution on a personal computer (Fig. 2D).

Evaluation of reference marker selection across methods

Selecting reference markers for each cell type plays a pivotal role in shaping the final deconvolution results [27]. While all methods successfully identified markers specific to each cell type (Additional file 2: Fig. S4), substantial differences were observed among the selected markers by different methods (Fig. 3A). This discrepancy indicates that each method may capture distinct cell-type-specific methylation features.

To evaluate the quality of reference markers selected by each method, we first compared the marker specificity (measured by both absolute and statistical differences) across methods (the "Methods" section). Results showed that robust marker selection (e.g., CelFEER and UXM), those with high specificity, directly contributes to more accurate cell type deconvolution (Fig. 3B, C). Additionally, we hypothesized that ideal reference markers should align with the reported cell-type-specific CpGs. To test this, we assessed reference marker selection results by comparing them with the cell-type-specific methylation marks in the MethyMark database [28]. MethyMark integrated 50 methylomes across 42 human tissues/cell types, 15 of which were also included in the reference methylation atlas of our study (Additional file 1: Table S6). For each
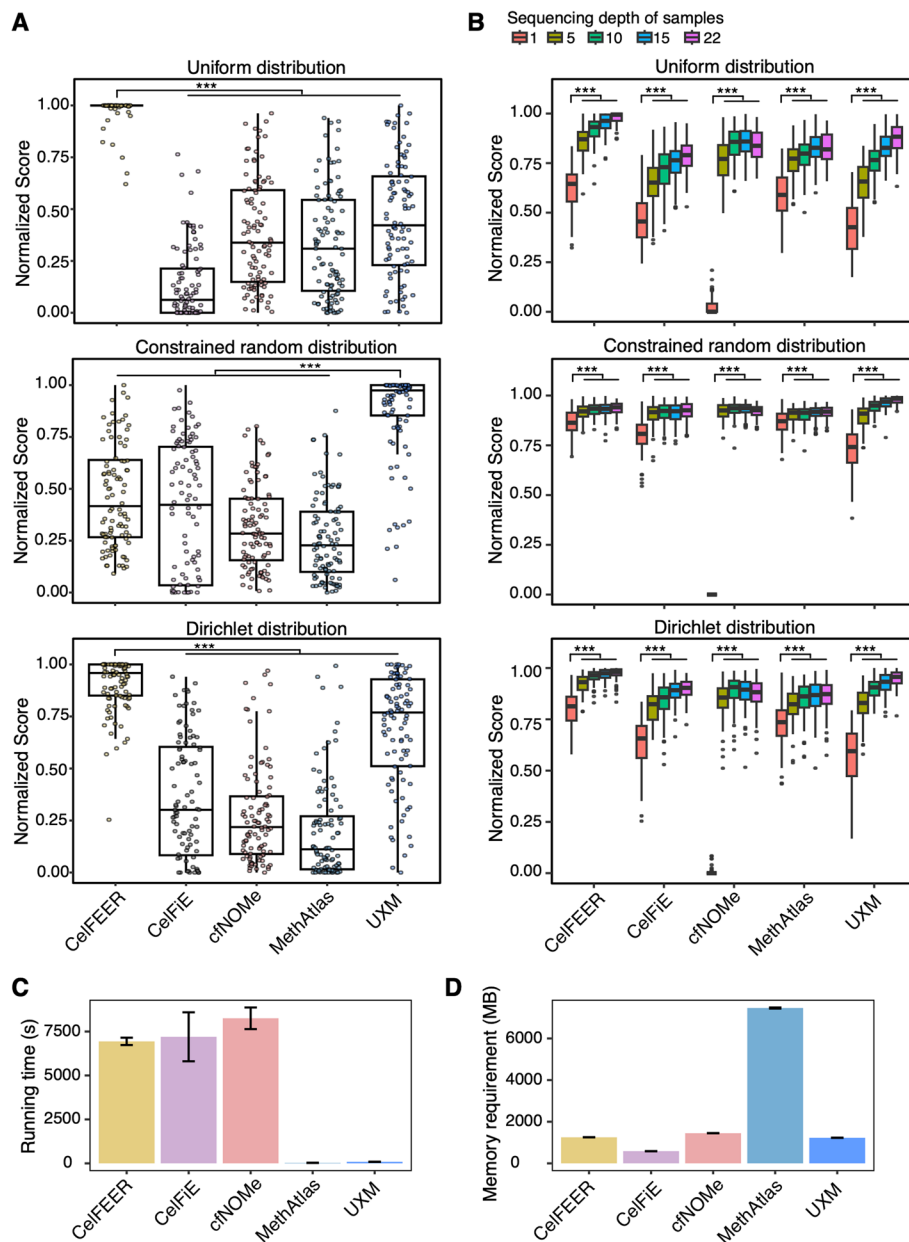
Fig. 2 Deconvolution results under the default setting of evaluated methods. **A** Boxplots showing normalized performance score of deconvolution results from various methods in the uniform distribution (upper), constrained random distribution (middle), and Dirichlet distribution (lower) datasets. **B** Boxplots showing normalized performance score of deconvolution results from various methods for samples in different sequencing depth in the uniform distribution (upper), constrained random distribution (middle), and Dirichlet distribution (lower) datasets. *P*-values were determined using Wilcoxon rank-sum test. ***$P < 0.001$. **C** Running time for each deconvolution method. **D** Memory requirements for each deconvolution method

of the 15 overlapping cell types, we examined the concurrence between reference markers identified by each deconvolution method and cell-type-specific markers in MethyMark. Notably, UXM identified a higher proportion of reference markers that corresponded to cell-type-specific markers in MethyMark (Additional file 2: Fig.
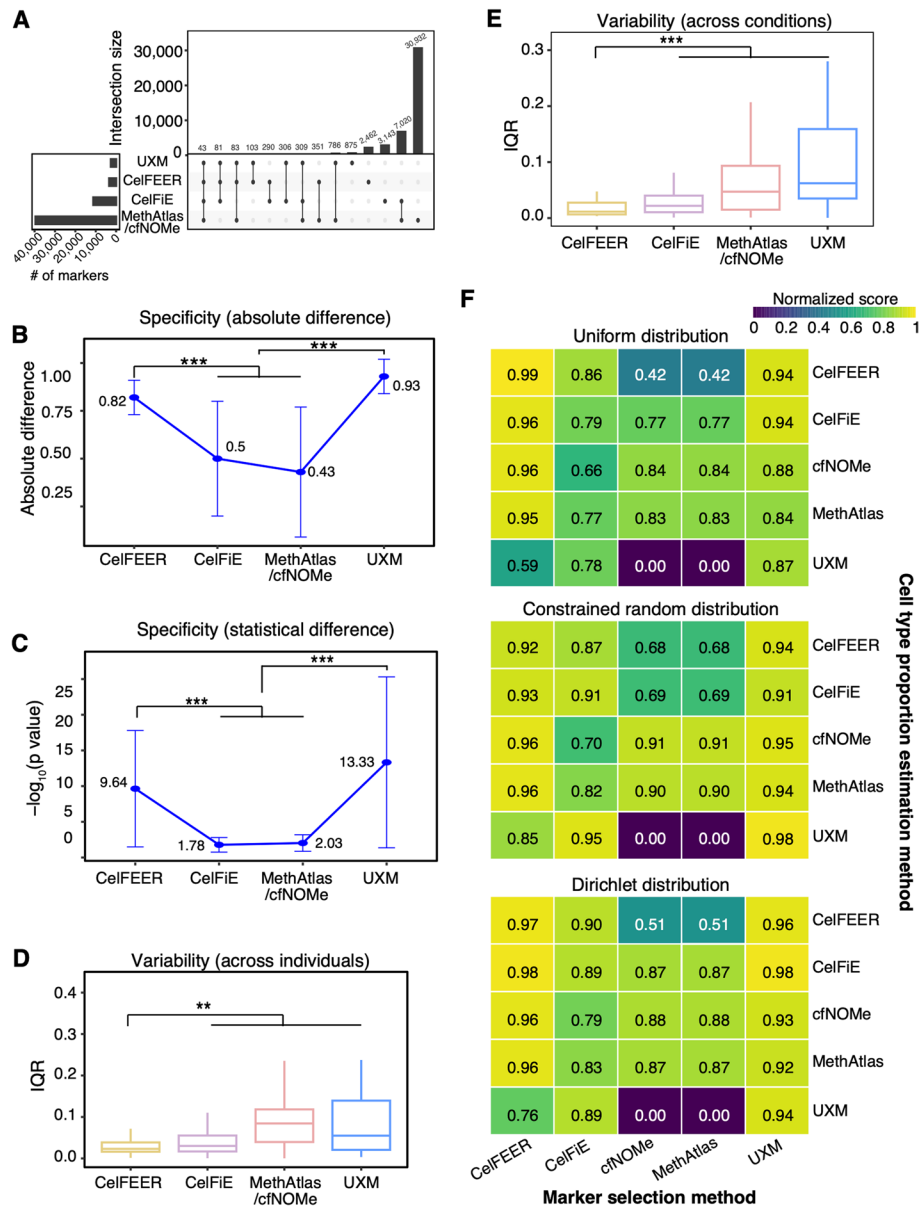
Sun *et al. Genome Biology*   (2024) 25:318

Page 8 of 26

Fig. 3 Evaluation of reference marker selection in each deconvolution method. **A** Upset plots illustrate the intersections of selected markers from various methods. cfNOMe used the same set of reference markers as MethAtlas. **B** Markers' specificity measured by absolute methylation level difference between the target cell type and all the others. Data are represented as mean ± standard deviations (SD). **C** Markers' specificity measured by statistical difference between the target cell type and all the others. Statistical difference was calculated as -log$_{10}$(*P*-value) using Student *t*-test. Data are represented as mean ± SD. **D** The variability of the selected markers across healthy individuals was measured by interquartile range (IQR). IQR data were sourced from the ImmuMethy database, focusing on four immune cell types. **E** The variability of the selected markers across diverse healthy states was measured by IQR. Data for IQR were calculated based on data extracted from the ImmuMethy database. **F** Normalized performance score across 25 combinations of marker selection and cell type proportion estimation modules. The median normalized performance score, calculated from 100 in silico cfDNA samples, was represented by color and labeled accordingly. Rows represent cell type proportion estimation methods, and columns represent marker selection methods. *P*-values between different methods in **B**–**E** were determined using Wilcoxon rank-sum test. **P < 0.01, ***P < 0.001

Sun *et al. Genome Biology*   (2024) 25:318

Page 9 of 26

S5A), suggesting its potential preference for capturing cell type specificity in methylation levels.

Moreover, an optimal reference methylation atlas should include markers that remain stable across individuals and diverse conditions. Consequently, we assessed reference marker selection results by examining marker variability based on data in ImmuMethy [29], which collected DNA methylation data for immune cells in different conditions. Variability was quantitatively measured using the interquartile range (IQR) and standard deviations (SD). Among the four immune cell types (B cell, granulocyte, monocyte, and T cell) included in the reference methylation atlas of our study and also collected in ImmuMethy with more than three conditions, CelFEER exhibited overall lower variability across individuals in each condition, as evidenced by both IQR and SD metrics (Fig. 3D and Additional file 2: Figs. S5-S9). CelFEER also demonstrated overall lower variability across different conditions (Fig. 3E and Additional file 2: Fig. S5C). These lower variabilities may contribute, at least in part, to CelFEER's superior performance in both uniform distribution and Dirichlet distribution datasets. However, UXM, which efficiently captured a significant proportion of cell-type-specific markers in MethyMark (Additional file 2: Fig. S5A) and demonstrated relatively superior deconvolution performance (Fig. 2A), displayed the greatest variability across diverse conditions (Fig. 3D, E). In conclusion, while the variability of the reference methylation markers could influence the performance of deconvolution methods, it is not the sole determining factor.

Since all evaluated methods can be viewed as consisting of two modules—a marker selection module and a cell type proportion estimation module, we conducted a detailed analysis to assess the relative importance of each. For this, we fed the markers selected by each of the five methods into the cell type proportion estimation modules of all methods, resulting in 25 unique combinations (the "Methods" section). Interestingly, markers selected by CelFEER maintained relatively high performance even when cell type proportion estimation was performed by other methods, such as CelFiE, MethAtlas, and cfNOMe (Fig. 3F and Additional file 2: Fig. S10). An expectation was UXM, which performed best when its own selected markers were used for cell type proportion estimation (Fig. 3F and Additional file 2: Fig. S10), highlighting the importance of the compatibility between these two components for optimal performance with UXM. These results revealed the critical role of marker selection module: markers selected by CelFEER were able to compensate the lower performance of CelFiE, MethAtlas, and cfNOMe.

Impact of sequencing depth of reference data on deconvolution results

The sequencing depth filter threshold utilized during the identification of reference methylation markers can significantly affect the quality and total number of the final selected markers. Therefore, we sought to assess whether varying sequencing depth filter thresholds impacted deconvolution performance. Starting from the default sequencing depth filter threshold (15X) for most deconvolution methods, we explored different thresholds up to 300X. Subsequently, we ran the deconvolution methods under each sequencing depth threshold, using all the three datasets. Notably, UXM showed stable performance as the filter became more stringent, likely due to the fixed low number of selected markers (Fig. 4). Conversely, for the other methods, the number of markers dropped sharply when the sequencing depth thresholds exceeded 100, resulting in

Sun *et al. Genome Biology*   (2024) 25:318

Page 10 of 26

significantly worse performance (Fig. 4). Moreover, the dramatic reduction in the feature selection search space resulted in only a small fraction—or none—of the original markers (at a sequencing depth threshold of 15) being retained, causing a notable decline in marker specificity (Additional file 2: Fig. S11). These results highlighted the critical role that both the quality and quantity of selected markers play in determining the accuracy of cell type proportion estimation.

Additionally, we assessed the impact of sequencing depth in the raw reference data used to generate the reference atlas on marker selection quality and, consequently, on deconvolution performance. We downsampled the reference data to average CpG site coverages of 85, 71, 57, 43, 28, 14, and 3, corresponding to 5/6, 2/3, 1/2, 1/3, 1/6, and 1/30 of the original coverage, respectively. Consistently, deconvolution performance declined sharply when reference data coverage was as low as 3 (Fig. 5A–C and Additional file 2: Fig. S12). Most methods (e.g., CelFEER, cfNOMe, MethAtlas, and
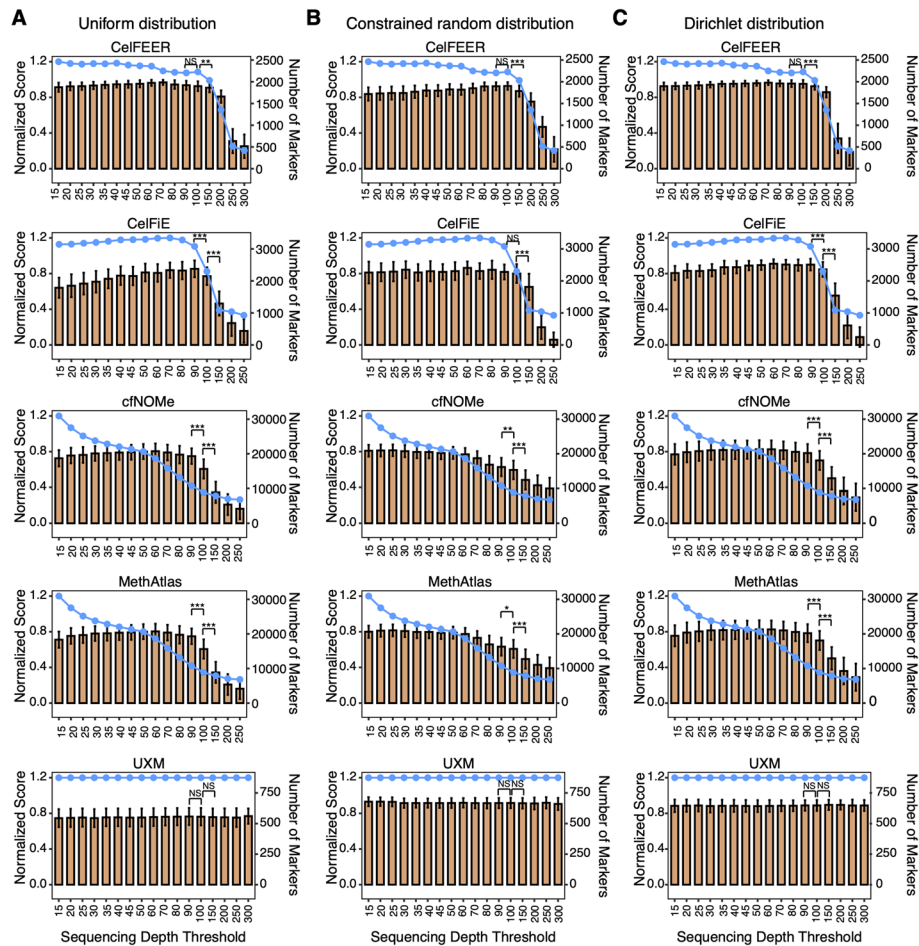

Fig. 4 Impact of sequencing depth filter on deconvolution results. **A**–**C** Normalized performance score of deconvolution results from various methods across different sequencing depth thresholds for marker selection in the uniform distribution (**A**), constrained random distribution (**B**), and Dirichlet distribution (**C**) datasets. The line plot depicts the number of markers selected. MethAtlas, CfNOMe, and CelFiE were unable to find reference markers when the sequencing depth threshold was set to 300. *P*-values were determined using Wilcoxon rank-sum test. NS, $P \geq 0.5$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$

UXM) showed improved deconvolution performance with increased coverage. However, CelFiE achieved its best performance at a reference data coverage of 14. These differing impacts were related to variations in marker specificity (Fig. 5D).

Impact of reference atlas completeness

Ideally, a comprehensive reference atlas should encompass all informative methylation markers and include as many cell types as possible. In practical terms, however, the presence of missing reference markers or cell types is inevitable due to the constraints imposed by the availability of high-coverage sequencing data. Thus, we conducted an assessment to determine whether the completeness of the reference methylation atlas impacted deconvolution results.

Firstly, we ran the five deconvolution methods using reference atlases with varying proportions of markers (10%, 20%, 30%, 40%, and 50%) removed from the full atlas. This approach simulated scenarios where input information for deconvolution methods might be incomplete. Under all three datasets, UXM exhibited dramatic worsen performance with increasing proportions of missing markers, while CelFEER and CelFiE displayed moderate decreasing performance (Fig. 6A and Additional file 2: Fig.
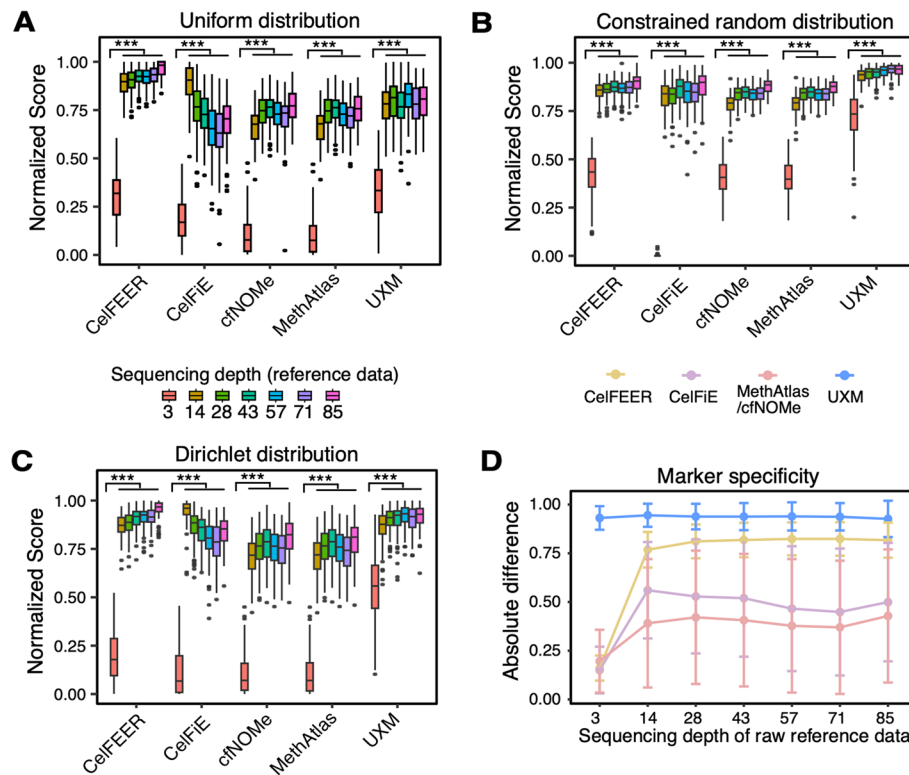

Fig. 5 Impact of sequencing depth in the raw reference data. **A**–**C** Normalized performance score of deconvolution results from various methods across different sequencing depth in raw reference data used to generate the reference atlas for the uniform distribution (**A**), constrained random distribution (**B**), and Dirichlet distribution (**C**) datasets. **D** Line and scatter plot showing the markers' specificity measured by absolute methylation level difference between the target cell type and all the others. Data are represented as mean ± Standard Deviations (SD). *P*-values were determined using Wilcoxon rank-sum test. ****P* < 0.001
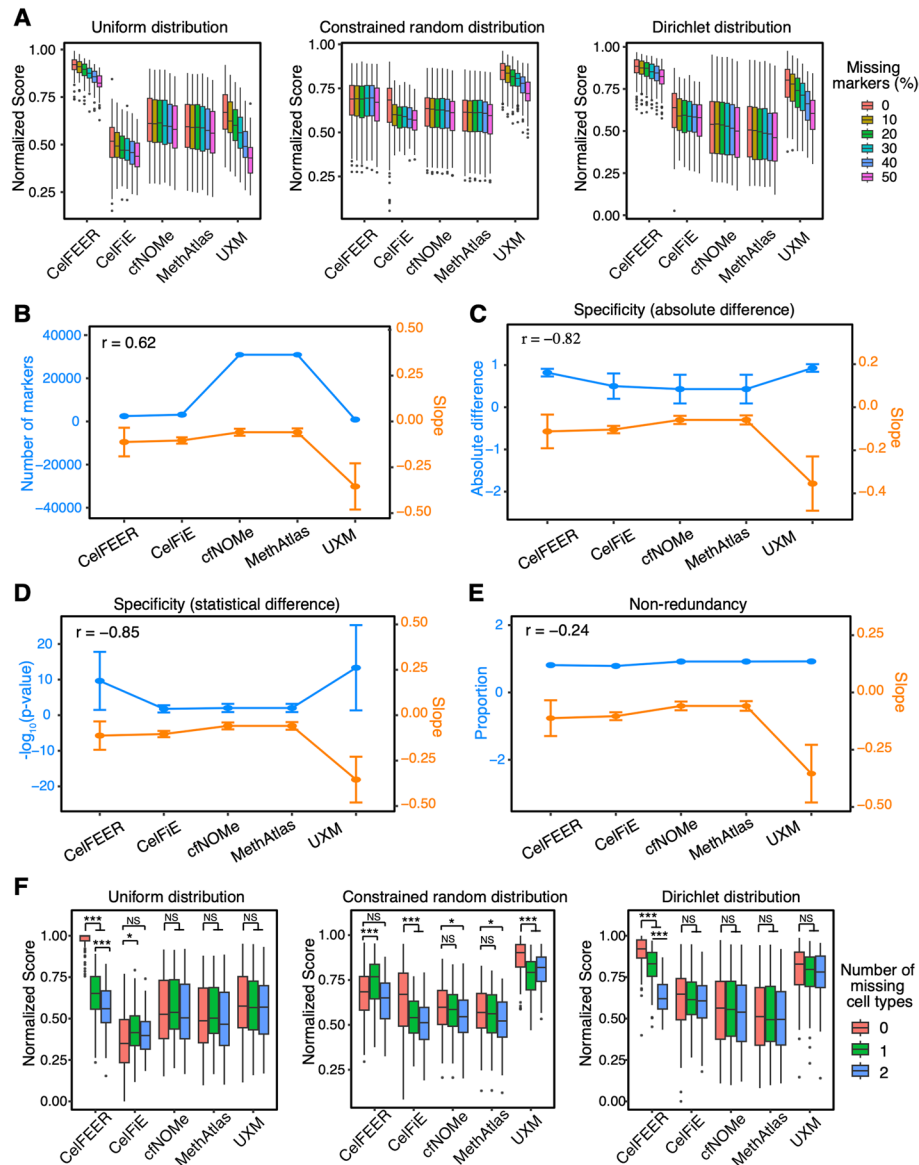
Fig. 6 Impact of reference marker completeness on deconvolution results. **A** Boxplots showing the normalized performance score from various deconvolution methods for the uniform distribution (left), constrained random distribution (middle), and Dirichlet distribution (right) datasets. Different proportions of markers are intentionally omitted in the reference. **B**–**E** Line and scatter plots showing the relationship between the impact of missing markers (orange) and the number (**B**), specificity (**C**, **D**), and non-redundancy (**E**) of markers (blue). Specificity was measured by absolute methylation level difference and -log10(*P*-value) from a Student *t*-test comparing the target cell type to all others. Non-redundancy was assessed as the proportion of paired markers with a Pearson correlation coefficient below 0.2. The impact of missing markers was quantified as the slope of normalized performance score over marker missing proportions using linear regression. Data are represented as mean ± SD with the Pearson correlation coefficient between the means of the two lines shown at the top left. **F** Boxplots showing the normalized performance score from various deconvolution methods with no, one, or two cell types intentionally missing in the reference. *P*-values were determined using Wilcoxon rank-sum test. NS, $P \geq 0.5$, *$P < 0.05$, ***$P < 0.001$

Sun *et al. Genome Biology*   (2024) 25:318

Page 13 of 26

S13). In contrast, MethAtlas and cfNOMe showed consistent performance under different proportions of missing markers.

To further investigate what factors are associated with the impact of missing markers, we assessed three factors: the number of markers, marker specificity, and marker non-redundancy. We did observe a relationship between the number of markers and the impact of missing markers (Fig. 6B). When fewer markers are used (e.g., in UXM), missing markers tend to have a more significant effect on deconvolution performance, as each marker contributes more to the overall signal (Fig. 6B). An additional analysis, which evaluated how varying the number of selected markers (25 per cell type as recommended in the original paper, as well as 50, 100, 200, and 300) impacts the performance of UXM, displayed a clear relationship between the number of selected markers and the impact of missing markers (Additional file 2: Fig. S14). On the other hand, with a larger number of markers (e.g., cfNOMe and MethAtlas), the impact of individual missing markers is diluted (Fig. 6B). These results demonstrate that method with smaller marker set, while potentially more efficient, may come at the cost of robustness to missing data. Moreover, both specificity and non-redundancy of markers could indicate the impact of missing markers (Fig. 6C–E). UXM, which selected highly specific and non-redundant markers, is more sensitive to missing markers, as these markers are likely critical to distinguishing cell types. In contrast, tools like cfNOMe, MethAtlas, and CelFiE, which use a more redundant marker set, are relatively more robust to missing markers, though this could come at the cost of reduced specificity.

Subsequently, we examined the impact of missing cell types by removing one or two cell types from the reference atlas while keeping the in silico cfDNA samples unchanged. Among the five methods, only CelFiE and CelFEER can estimate the proportions of unknown cell types not available in the reference atlas. However, CelFEER showed reduced performance when one cell type was missing and even worse when two cell types were missing in the uniform distribution and Dirichlet distribution datasets, while showed similar or slightly better performance in constrained random distribution dataset (Fig. 6F). CelFiE only maintained the performance in uniform distribution and Dirichlet distribution datasets, while achieving significantly decreased performance in constrained random distribution dataset. Conversely, the performance of MethAtlas and cfNOMe, although worse, remained relatively stable (Fig. 6F). For the constrained random distribution dataset, we noted that when the missing cell types included blood cells, which are the major cell types, the performance became worse (Additional file 2: Fig. S15). To evaluate the impact of missing major cell types in the reference atlas, we subsequently removed each of the five blood cell types or one blood cell type coupled with another randomly selected cell type for the constrained random distribution dataset. All methods, except CelFEER, exhibited decreased performances compared to the results when there were no missing cell types (Additional file 2: Fig. S15). CelFEER showed comparable performance in scenarios where one cell type was missing, potentially benefiting from its ability to estimate unknown cell type proportions. In conclusion, the completeness of the reference atlas could impact deconvolution results to varying degrees for different methods and different proportions of the missing cell types.

Evaluation of deconvolution results on real-world clinical settings

Considering all previous analyses were based on in silico datasets, they may not fully reflect the complexities present in real-world cfDNA samples. Diseases such as cancers can influence cell death in affected tissues and cell types [1]. Deconvoluting the cell type composition in cfDNA from these patients can reveal altered homeostasis in affected cell types [5, 30, 31]. Additionally, changes in cell type fractions may serve as indicators for disease detection [32]. To build on this knowledge and strengthen the validity of our findings, we expanded our assessment by incorporating two real-world datasets: (1) a hepatocellular carcinoma (HCC) dataset [25], comprising WGBS data of cfDNA samples from 24 HCC patients and 32 controls, and (2) a multi-cancer dataset, featuring cfMethyl-Seq (a revised reduced representative bisulfite sequencing technology) data of cfDNA samples from 225 cancer patients across five cancer types and 193 controls [26].
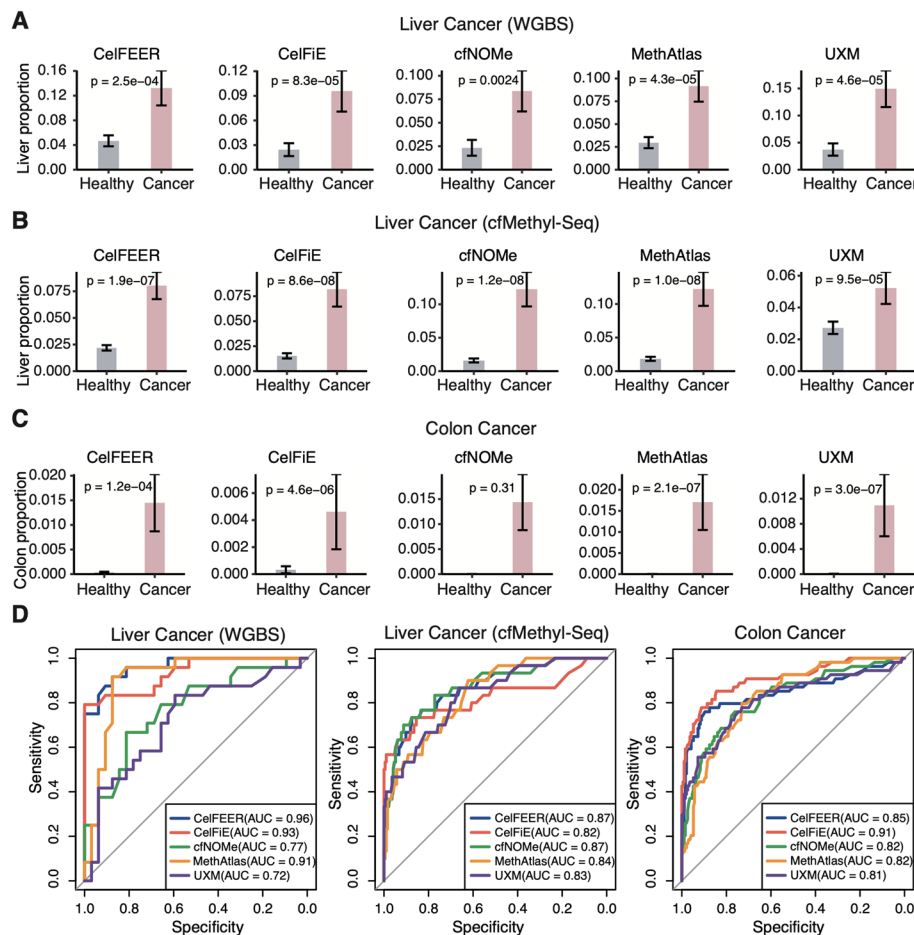

Fig. 7 The deconvolution results of cfDNA from the diseased patients and healthy individuals. **A** The liver-derived cfDNA fractions from liver cancer patients (WGBS) and healthy individuals. **B** The liver-derived cfDNA fractions from liver cancer patients (cfMethyl-Seq) and healthy individuals. **C** The colon-derived cfDNA fractions from colon cancer patients and healthy individuals. **D** The receiver operating characteristic (ROC) curves showing disease detection performance using the estimated fractions of all cell types. *P*-values were determined using a one-sided Wilcoxon rank-sum test. The area under ROC curve (AUC) was labeled for each deconvolution method

We then applied all evaluated methods to assess whether tissue fractions in cfDNA could indicate disease presence. Across all methods in most datasets, we observed a significantly higher fraction of affected tissue in diseased patients compared to healthy individuals (Fig. 7 and Additional file 2: Fig. S16). Using statistical significance as a metric, we found that MethAtlas performed best for identifying elevated tissue fractions in diseased patients (Additional file 1: Table S7). Additionally, we evaluated disease detection performance using the estimated cell type fractions as predictors. Based on ROC-AUC, CelFEER and CelFiE demonstrated the overall highest accuracy for disease detection (Fig. 7, Additional file 2: Fig. S16, and Additional file 1: Table S7). The inclusion of these real-world datasets provides direct evidence of the methods' efficacy in clinical settings.

Summary of the cfDNA deconvolution methods' performance

To summarize our benchmark results, we developed a scoring-ranking system based on the evaluation metrics on the three in silico datasets and two real-world datasets (the "Methods" section). CelFEER demonstrated the best overall performance, while cfNOMe exhibited the worst overall performance (Fig. 8). However, when considered for various evaluation aspects, each method displayed distinct characteristics. CelFEER showcased relatively high accuracy and robust performance, yet it had a slower running speed. UXM emerged as a superior method in terms of accuracy and sequencing depth but exhibited sensitivity to missing markers and cell types in the reference atlas. CelFiE demonstrated the median overall performance, but with the best performance for disease detection effectiveness. Although cfNOMe and MethAtlas did not excel in overall performance, they proved to be the most stable methods for missing reference markers. Despite MethAtlas's lowest accuracy, it had the fastest processing speed and the best performance on affected tissue detection of real-world datasets.

In brief, for cfDNA deconvolution analysis, we recommend users to (1) choose a method stable to sequencing depth, especially with low-coverage data; (2) employ a stringent marker selection strategy to retain the most informative and specific markers; and (3) utilize a comprehensive reference atlas that encompasses as many cell types as possible that present in the cfDNA.
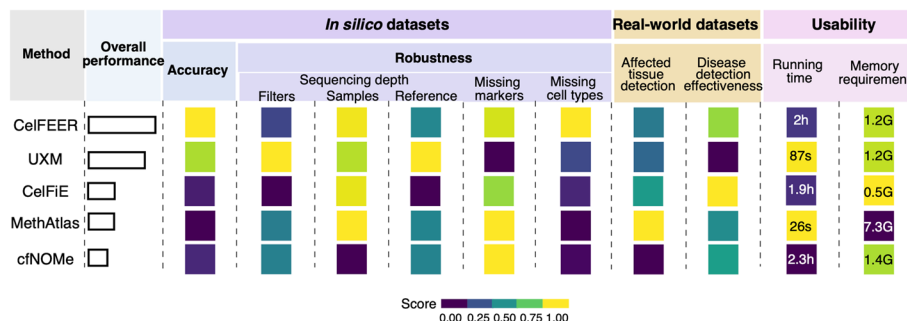

Fig. 8 Overall summary of the methods' performance. The color-coded table displays a concise overview of the benchmark results for the five methylation-based cfDNA deconvolution methods. Refer to the "Methods" section for detailed information on the ranking criteria and methodology

Sun *et al. Genome Biology*   (2024) 25:318

Page 16 of 26

Discussion

We conducted a comprehensive evaluation of the performance of five methylation-based cfDNA deconvolution methods using in silico cfDNA samples and assessed their performance using five metrics. Given that cfDNA primarily originates from apoptosis of diverse cell types, we opted to utilize methylation data of cell types for generating in silico cfDNA samples. However, previous methylation datasets often cover only a fraction of genomic regions and are typically based on tissues containing a mixture of cell types [32–34]. A recent contribution by Loyfer et al. provided a human methylome atlas encompassing 39 cell types derived from 205 healthy tissue samples through deep WGBS [6]. This atlas proves to be a valuable resource for benchmarking methylation-based cfDNA deconvolution methods. Notably, the inclusion of biological replicates for most cell types in this atlas allows us to split replicates into two halves for constructing the reference atlas and generating in silico cfDNA samples, thereby avoiding the potential "double dipping" problem [35]. Leveraging the deep sequencing depth of this methylome atlas, we can also assess the impact of sequencing depth and reference atlas completeness. Based on the comprehensive benchmarking results, we provide general guidelines for researchers to choose suitable methods.

Our analyses were conducted on three in silico cfDNA datasets to encompass diverse biological scenarios and assess the robustness of these methods. The constrained random distribution dataset aimed to mimic the actual cfDNA compositions in healthy individuals [5, 6], while the uniform distribution dataset was employed to evaluate the methods' robustness to different cell type compositions. Additionally, the Dirichlet distribution dataset, which ensures simulated fractions sum to one and allows for adjustable variability in cell type proportions, provided moderately diverse mixtures with varying contributions from different cell types. Each method exhibited distinct performances among those datasets, underscoring the importance of employing evaluation on different in silico cfDNA datasets. In practical scenarios, researchers often cannot provide a complete reference atlas containing all the markers and cell types. Therefore, we also evaluated the influence of missing markers and cell types in the reference atlas. While some methods, such as MethAtlas, cfNOMe, and CelFEER, demonstrated stable performances under these factors, others proved to be sensitive. Consequently, researchers should carefully consider the completeness of their data when selecting tools for cfDNA deconvolution analysis. In particular, our benchmark results revealed significant performance disparities between CelFEER and CelFiE, despite both employing the same model and optimization algorithm. The two methods' primary distinctions lie in the methylation metric used for deconvolution and the selection procedure for reference markers. Hence, the two methods' different performance underscores the critical roles of the input methylation metric and reference markers in cfDNA deconvolution.

While the benchmarked methods in our study yielded reasonable deconvolution results in most analyses, certain limitations require attention. Firstly, no single method consistently outperformed others across all evaluated factors, suggesting that the optimization of these methods might prioritize specific considerations. Secondly, the cell type specific markers are still far from fully captured, as evidenced by the limited overlaps between reference markers selected by these methods and cell type specific markers in the external datasets. Third, none of these methods accounts for the uncertainty

of deconvolution results, potentially impacting analytical outcomes [36]. Fourth, when deploying cfDNA deconvolution tools, the importance of software maintenance and user support cannot be overstated. It is worth noting that some of our benchmarked methods lacked a comprehensive user guide and were less responsive to user inquiries. Finally, although benchmarking using in silico data offers a controlled environment and ground truth, they may not fully capture the complexity and variability inherent in real-world cfDNA samples, such as heterogeneity in methylation patterns caused by biological factors like tissue-specific cell death, disease progression, or inflammation. These factors can substantially affect methylation profiles, making in silico data an imperfect stand-in for clinical samples. To address these limitations, adding real-world datasets is essential for a more accurate assessment of deconvolution methods' clinical relevance and robustness. However, real-world datasets also come with challenges, including the absence of a ground truth for validation and potential biases from sample collection methods or patient demographics.

Taken together, our benchmark results provide a general guideline on how to conduct cell type deconvolution for cfDNA methylation sequencing data, paving the way for methodological enhancements and refinements. Future endeavors should prioritize the extraction of comprehensive and accurate cell-type-specific markers, while accounting for factors like sequencing depth, biological conditions [37], cellular heterogeneity [38], and inter-individual variability [39, 40]. Additionally, integrating methylation-based cfDNA deconvolution with other omics data modalities, such as nucleosome footprints [9, 41] and chromatin accessibility [42], can yield holistic insights into cellular specificity and bolster deconvolution performance by harnessing complementary information from diverse omics datasets. Furthermore, with the increasing availability of single-cell DNA methylation sequencing data in the foreseeable future [43], aggregating such data will enrich the completeness, intricacy, and reliability of the reference atlas utilized in cfDNA deconvolution, thereby facilitating the development of more advanced methods.

Conclusions

We benchmarked five methylation-based methods for cell type deconvolution from cfDNA, highlighting CelFEER and UXM as the top-performing approaches. Importantly, the selection of the most appropriate method should be guided by critical factors, such as sequencing depth and the comprehensiveness of the reference atlas. Careful consideration of these factors allows researchers to align their methodology with the specific requirements of their study, thereby enhancing the accuracy and effectiveness of cell type deconvolution in cfDNA analyses.

Methods

Data collection and processing

We curated a comprehensive human DNA methylation dataset, including 39 normal human cell types from Gene Expression Omnibus (GEO) with accession number GSE186458 [6]. Our selection criteria focused on including cell types having a minimum of two biologically independent replicates of high quality. This rigorous selection process yielded a refined dataset comprising 35 distinct cell types, encompassing a total of 182 samples (Additional file 1: Table S7). Subsequently, we partitioned all samples into two

halves. Replicates from the same cell type in each half were merged into unified samples utilizing the wgbstools [6]. This process generated two independent datasets, each serving a specific purpose. One dataset was employed for the creation of in silico cfDNA samples, while the other functioned as the reference methylation atlas.

We collected WGBS data of the plasma cfDNA samples from 32 healthy individuals and 24 liver cancer patients under the accession code EGAD00001000856 in the European Genome-Phenome Archive (EGA) [25]. We also collected the cfMethyl-Seq data of the plasma samples from 193 healthy individuals and 225 cancer patients (67 lung adenocarcinoma, 41 lung squamous cell carcinoma, 30 liver cancer, 54 colon cancer, and 33 stomach cancer patients) under the accession code EGAD00001009003 in the EGA [26]. These real-world data were used for the assessment in clinical settings. For the cfMethly-Seq dataset, TrimGalore (v0.6.10) was used to trim the default Illumina adapters and low-quality bases from both ends of the reads. For the WGBS dataset, we trimmed the first 8 bases from the 5′ end of each read and filtered out reads with a quality score below 20. The trimmed reads from both datasets were then aligned to the hg38 genome using Bismark (v0.24.2). All aligned BAM files were deduplicated using the *deduplicate_bismark* function from Bismark (v0.24.2) and filtered with Samtools (v1.1). Finally, the bam files were converted to pat files using the *bam2pat* function from wgbstools, with the reference genome being hg38, for subsequent processes.

The generation of in silico cfDNA samples and reference methylation atlas

To investigate the impact of cell type proportion distribution, we generated three sets of in silico cfDNA datasets, each consisting of 100 samples. These sets were based on three distinct distributions: uniform distribution, Dirichlet distribution, and constrained random distribution. For the uniform distribution dataset, we employed the *random.uniform* function from the Python package *numpy* (v 1.24.3) [44] to create a 35×100 matrix, with each column sum normalized to 1. To generate the Dirichlet distribution dataset, we utilized the IMIFA (v 2.1.10) in R, setting the distribution parameter $\alpha = 0.5$ to create a 35×100 matrix. The constrained random distribution dataset was generated through the following steps: (1) selection of the five immune cell types and random inclusion of *n* additional cell types, where *n* is an integer ranging from 1 to 10; (2) generation of a list of $(n + 5)$ random floating numbers using *random* function from Python package *random*; (3) division of all numbers by the sum of the random numbers to ensure that cell type proportions summed to 1. Then, the proportion (P_i) for a cell type *i* was calculated as $P_i = \frac{r_i}{\sum_1^{n+5} r_i}$, where r_i was the random number; (4) assignment of the five largest proportion values to the five immune cell types, with the remaining values allocated to the *n* additional cell types; (5) assignment of a value 0 to all unselected cell types. This process was repeated 100 times, resulting in 100 in silico cfDNA samples where immune cell types were predominant. Subsequently, the *mix_pat* command in wgbstools was used to generate each cfDNA sample according to the corresponding proportions.

The reference methylation atlas was generated for the five deconvolution methods by following the specific instructions provided by the respective method. In summary:

Sun *et al. Genome Biology*   (2024) 25:318

Page 19 of 26

(1) MethAtlas selected the 100 most hypermethylated and hypomethylated CpGs based on methylation ratio of the CpG site (the fraction of reads that are methylated), along with neighboring CpGs up to 50 bp for each cell type. Pairwise-specific CpGs were then identified through 500 iterations, focusing on pairs of cell types with the smallest Euclidean distance in each iteration

(2) cfNOMe used the same set of reference markers as MethAtlas

(3) CelFiE calculated the distance between the methylation ratio of each cell type and the median methylation ratio of all cell types for each CpG. Then, the top 100 CpGs with the largest distance for each cell type, along with the proximal CpGs (± 250 bp) around these CpGs, were selected to form the final reference methylation atlas

(4) CelFEER adapted CelFiE's method with the following major improvements: (a) use 500-bp windows instead of CpG sites to find marker regions, (b) employ read averages (the ratio of methylated CpG sites to the total number of CpG sites on a read) instead of methylation ratio as the methylation quantification metric, (c) focus on hypomethylated regions, (d) select the top 150 markers and remove markers that are in the top 100 of two or more cell types before narrowing down to the top 100 markers for each cell type

(5) UXM firstly found genomic regions based on the segmentation algorithm that optimizes the CpG sites homogeneity score in each genomic region using the *segment* function (–min_cpg 5 –genome hg38 –betas) from wgbstools. Then, UXM selected the top 25 specifically unmethylated genomic regions with at least 5 CpGs with size ranging from 10 to 1500 bp using the *find_markers* function (–pval 0.05 –min_cpg 5 –min_bp 10 –max_bp 1500 –sort_by delta_means –tg_quant 0.25 –bg_quant 0.025 –top 25 –only_hypo) from wgbstools

Implementation of each cfDNA deconvolution method

All the five methods took in silico cfDNA samples and reference methylation atlas as inputs for the deconvolution process. To execute MethAtlas (https://github.com/nloyfer/meth_atlas), the *deconvolve.py* script was utilized. For cfNOMe (https://github.com/FlorianErger/cfNOMe), the *methylation_deconvolution.py* script was employed for deconvolution. For CelFiE (https://github.com/christacaggiano/celfie), *celfie.py* was used with –unknowns set to 0, unless explicitly specified. Similarly, for CelFEER (https://github.com/pi-zz-a/CelFEER), the *celfeer.py* was employed with –unknowns set to 0, unless otherwise specified. Lastly, for UXM (https://github.com/nloyfer/UXM_deconv), the *uxm deconv* command was executed for deconvolution.

Evaluation of reference markers selection

We evaluated the cell type specificity, non-redundancy, and variability of the reference markers identified by each deconvolution method. Specificity was measured in three ways: (1) the absolute difference in methylation levels between the target cell type and all others; (2) the statistical difference expressed as $-\log_{10}(P\text{-value})$, calculated using the Student *t*-test, comparing the target cell type with all others; (3) the overlapped ratio with cell-type-specific methylation marks (CpG sites) derived from MethyMark database (http://fame.edbc.org/methymark/). We first utilized liftOver from the UCSC Genome

Browser [45] to convert the genome coordinates of CpGs from hg19 to hg38. Subsequently, BEDTools [46] (v2.31.0, bedtools intersect -wa -u) was employed to intersect the reference markers identified by each deconvolution method with cell-type-specific markers in MethyMark for each of the 15 overlapping cell types. The overlapped ratio was employed to assess the markers' cell type specificity, with higher ratios indicating better specificity. The non-redundancy of the reference markers was assessed as the proportion of paired markers with a Pearson correlation coefficient below 0.2.

To evaluate the cell type variability of the reference markers, we downloaded DNA methylation variability data of immune cells from ImmuMethy (http://immudb.bjmu.edu.cn/immumethy/index.jsp). The MethyMark database integrates 50 methylomes across 42 human tissues/cell types, with 15 cell types overlapping with those in our reference methylation atlas (Additional file 1: Table S6). To quantify variability in ImmuMethy, we measured it in two ways: interquartile range (IQR) and standard deviations (SD). Four immune cell types (B cell, granulocyte, monocyte, and T cell) from ImmuMethy, having more than three conditions overlapping with the reference cell types in our study, were selected for subsequent analysis. SD and IQR across different individuals of the same cell type and conditions were extracted directly from ImmuMethy. SD and IQR across different conditions of the same cell types were calculated based on the mean methylation value of each condition. BEDTools was used to obtain SD and IQR values for the reference markers selected by each deconvolution method. Higher SD or IQR values corresponded to increased variability.

Evaluation of effects of sequencing depth

Firstly, to evaluate the influence of sequencing depth filters, we implemented 17 depth filter thresholds (15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 150, 200, 250, and 300) during the generation of the reference methylation atlas. We excluded the threshold of 300 for MethAtlas, cfNOMe, and CelFiE, as it resulted in no remaining markers for analysis. Specifically:

- For MethAtlas and cfNOMe, the reference methylation atlases were generated by applying each of the 17 thresholds to the median depth per CpG site across all samples
- In case of CelFiE and CelFEER, reference methylation atlases were generated by adjusting the *depth_filter* parameter accordingly
- For UXM, we simply modified the –min_cov parameter for the *find_markers* in wgbstools, keeping the other parameters unchanged

Secondly, to assess the impact of sequencing depth in the raw reference data used to generate the reference atlas on marker selection quality and deconvolution performance, we downsampled the reference data to average CpG site coverages of 85, 71, 57, 43, 28, 14, and 3, corresponding to 5/6, 2/3, 1/2, 1/3, 1/6, and 1/30 of the original coverage, using the wgbstools cview with *–sub_sample* parameter. After generating the new reference data, we generated the reference atlas for all five methods as described before and repeated the deconvolution analyses.

Finally, we investigated impact of the sequencing depth of the samples being analyzed. Based on the previously-mentioned three distributions—uniform distribution, Dirichlet distribution, and constrained random distribution—we generated additional in silico datasets with sample sequencing depths of 1, 5, 10, and 15, alongside our existing datasets of 22 (the lowest sequencing depth among data used to generate the in silico samples) using the coverage parameter (-c) in the *mix pat* function of wgbstools. Then, all five deconvolution methods were run on these in silico datasets with reference atlas generated using default settings.

Evaluation of effects of missing markers or cell types in the reference methylation atlas

To assess the impact of missing markers, we introduced five gradient missing proportions (10%, 20%, 30%, 40%, and 50%) for reference markers. The corresponding numbers of missing markers were randomly removed to represent a diverse range of genomic locations and ensure unbiased results. Each missing proportion was repeated 20 times to ensure the robustness of our findings. For the evaluation of missing cell types, we randomly deleted one cell type five times to generate five reference atlases with one missing cell type. Similarly, we randomly deleted two cell types five times to generate five atlases with two missing cell types. To evaluate the impact of missing major cell types in the reference atlas, we also deleted each of the five blood cell types or one blood cell type coupled with another randomly selected cell type for the constrained random distribution dataset to generate ten atlases. Subsequently, each deconvolution method was executed using reference methylation atlases with missing markers/cell types. For CelFiE and CelFEER, –unknowns was set to 1 or 2 for atlas with one or two missing cell types respectively. All these analyses were done under a sequencing depth filter of 15. For UXM, an additional analysis, which evaluated how varying the number of selected markers (25 per cell type as recommended in the original paper, as well as 50, 100, 200, and 300) impacts the performance of UXM, was done by setting –top parameter of the *find_markers* function of wgbstools.

Evaluation of the marker selection and cell type proportion estimation modules

To assess the relative importance of the marker selection module and cell type proportion estimation module within each method, we fed the markers selected by each of the five methods into the cell type proportion estimation modules of all methods, resulting in 25 unique combinations. When estimating cell type proportions with MethAtlas and cfNOMe, which utilize CpG sites as markers, we extracted all CpG sites falling within the marker regions defined by CelFiE, CelFEER, and UXM. These sites were then employed to construct the reference atlases. In the case of CelFiE, which employs genomic regions for marker selection, we compiled the marker regions (with the CpG sites of MethAtlas/cfNOMe treated as 1-bp regions) identified by the other four methods into an input reference atlas, as required by CelFiE. CelFEER and UXM adopted a similar approach to that of CelFiE for their procedures.

Assessment of deconvolution performance on real-world datasets

For the real-world datasets, we consolidated all data of the 35 cell types to generate full reference methylation atlases for the evaluation of deconvolution performance. Then, the

Sun *et al. Genome Biology*   (2024) 25:318

Page 22 of 26

reference atlases were used to perform deconvolution on the real-world cfDNA samples using each of the five methods. Given the absence of ground truth for these real-world datasets, we leveraged established knowledge from the literature to evaluate the deconvolution performance. One metric we used is the statistical difference in the cfDNA fraction derived from affected tissues between diseased and healthy individuals, calculated as $-\log_{10}(P\text{-value})$ using a one-sided Wilcoxon rank-sum test. Another metric is the ROC-AUC of Random Forest models for disease detection constructed using the estimated cell type fractions from each deconvolution method as predictors.

Assessment of running time and memory requirements

The running time and memory requirements were evaluated using memory-profiler (v0.61.0), a third-party Python package [47]. For analysis, we selected a set of 100 cfDNA samples under uniform distribution with sequencing depth filter of 15 for the subsequent analysis. The total runtime and peak memory usage for all 100 samples were systematically recorded for each deconvolution method. To enhance the reliability of our results, we executed each deconvolution method ten times under consistent computational conditions: utilizing an Intel(R) Xeon(R) Gold 5218R processor with 40 threads and 192 GB of memory, operating on Ubuntu 22.04 LTS with Python version 3.10.11.

Evaluation metrics

To assess the performance of each method, we computed root mean squared errors (RMSE), Pearson's correlation coefficient, Spearman's rank correlation, Lin's concordance correlation coefficient (CCC), and Jensen–Shannon divergence (JSD) between cell type proportions predicted by each cfDNA deconvolution method with the ground truth cell type proportions. Higher Pearson correlation, Spearman's rank correlation, and CCC, along with lower RMSE and JSD values, were indicative of superior performance. These metrics were compiled into a composite measure in each dataset, the normalized performance score (NPS), which integrates the min–max normalized values of all five metrics using the following equations:

$$S_{rmse}^{ijk} = 1 - \frac{X_{rmse}^{ijk} - \min\_j(X_{rmse}^{ijk})}{\max\_j(X_{rmse}^{ijk}) - \min\_j(X_{rmse}^{ijk})} \tag{1}$$

$$S_{P}^{ijk} = \frac{X_{P}^{ijk} - \min\_j(X_{P}^{ijk})}{\max\_j(X_{P}^{ijk}) - \min\_j(X_{P}^{ijk})} \tag{2}$$

$$S_{JSD}^{ijk} = 1 - \frac{X_{JSD}^{ijk} - \min\_j(X_{JSD}^{ijk})}{\max\_j(X_{JSD}^{ijk}) - \min\_j(X_{JSD}^{ijk})} \tag{3}$$

$$S_{CCC}^{ijk} = \frac{X_{CCC}^{ijk} - \min\_j(X_{CCC}^{ijk})}{\max\_j(X_{CCC}^{ijk}) - \min\_j(X_{CCC}^{ijk})} \tag{4}$$

$$S_{Spearman}^{ijk} = \frac{X_{Spearman}^{ijk} - \min\_j(X_{Spearman}^{ijk})}{\max\_j(X_{Spearman}^{ijk}) - \min\_j(X_{Spearman}^{ijk})} \tag{5}$$

Here, X_{rmse}^{ijk}, X_P^{ijk} X_{JSD}^{ijk}, X_{CCC}^{ijk} and $X_{Spearman}^{ijk}$ represent the original values of RMSE, Pearson correlation, JSD, CCC, and Spearman correlation of method j on the k-th sample of i-th dataset, respectively. Lower RMSE and JSD corresponds to higher values of S_{rmse}^{ijk} and S_{JSD}^{ijk}. And higher Pearson correlation, CCC, and Spearman correlation corresponds to higher values of S_P^{ijk}, S_{CCC}^{ijk}, and $S_{Spearman}^{ijk}$. The scores were then aggregated across different evaluation metrics to the NPS by calculating the arithmetic mean.

$$S_{NPS}^{ijk} = \frac{S_{rmse}^{ijk} + S_P^{ijk} + S_{JSD}^{ijk} + S_{CCC}^{ijk} + S_{Spearman}^{ijk}}{5} \tag{6}$$

Then, to summarize the overall performance of each method across different datasets and evaluation factors, we developed a scoring-ranking system based on the NPS.

- *Accuracy*: For each dataset, we initially calculated the median values of NPS (under a sequencing depth filter of 15) and then aggregated these median values across different datasets by calculating the arithmetic mean:

$$S_{accuracy}^{j} = \frac{1}{3}\sum_{i=1}^{3}(median\_k(S_{NPS}^{ijk})) \tag{7}$$

- *Robustness*: We first calculated the median values of NPS and then aggregated these median values across different datasets by calculating the arithmetic mean:

$$S_{overall}^{jd} = \frac{1}{3}\sum_{i=1}^{3}(median\_k(S_{NPS}^{ijdk})) \tag{8}$$

Here, S_{NPS}^{ijdk} represents NPS values for method j on sample k of dataset i under evaluated factor d (where d represents 17 filter thresholds for *Sequencing depth (filter)*, five depth thresholds for *Sequencing depth (samples)*, seven depth thresholds for *Sequencing depth (reference)*, six missing proportions for *Missing markers*, and 11 categories for *Missing cell types*, respectively). Subsequently, to assess the impact of *Sequencing depth* and *Missing markers* on the deconvolution results of each method, we calculated the variance of the above-defined $S_{overall}^{jd}$ under different depth thresholds (V_{filter}^{j}, $V_{samples}^{j}$, $V_{reference}^{j}$) and missing markers proportions ($V_{markers}^{j}$), separately. For *Missing cell types*, the relative difference to the score with no cell type missing was calculated using:

$$D_{cell}^{j} = \frac{\overline{S_{overall}^{jd}} - S_{overall}^{j0}}{S_{overall}^{j0}} \tag{9}$$

where $S_{overall}^{j0}$ is the score when there is no cell type missing, and $\overline{S_{overall}^{jd}}$ is the average score for all others except no-cell-type-missing category. A higher D_{cell}^{j} represents more robust performance.

Sun *et al. Genome Biology*   (2024) 25:318

Page 24 of 26

- *Real-world datasets:* The aggregate score for evaluation metrics of affected tissue detection (S_p^j) and disease detection effectiveness (S_{auc}^j) were calculated using the same approach as for *Accuracy*
- *Usability*: The aggregated score for running time (S_{time}^j) and memory requirements (S_{mem}^j) were calculated using the same approach as for *Accuracy*

Finally, we normalized the aggregated scores z^j of each evaluation factor for comparison and aggregation across different evaluation factors:

$$z^{j\prime} = \frac{z^j - \min\_j(z^j)}{\max\_j(z^j) - \min\_j(z^j)} \tag{10}$$

Here, z^j represents $S_{accuracy}^j, V_{filter}^j, V_{samples}^j, V_{reference}^j, V_{markers}^j, D_{cell}^j, S_p^j, S_{auc}^j, S_{time}^j$, and S_{mem}^j respectively. A higher score corresponds to a better deconvolution performance. The overall performance for each method was calculated based on the normalized scores using

$$S_{overall}^j = 0.5 \times S_{accuracy}^{j\prime} + 0.2 \times \left(\frac{V_{filter}^{j\prime} + V_{samples}^{j\prime} + V_{reference}^{\prime j} + V_{markers}^{j\prime} + D_{cell}^{j\prime}}{5} \right) + 0.2 \times$$ We

$$\left(\frac{S_{auc}^{j\prime} + S_p^{j\prime}}{2} \right) + 0.1 \times (\frac{S_{time}^{j\prime} + S_{mem}^{j\prime}}{2})$$

assigned different weights because we think accuracy are much more important than other evaluating factors in practice.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03456-8.

Additional file 1. This file includes multiple supplementary tables ranging from Tables S1 to S8.

Additional file 2. This file includes multiple supplementary figures ranging from Figs S1 to S16.

Data availability
The WGBS data of human cell types are publicly available on Gene Expression Omnibus (GEO) under the accession number GSE186458 [48]. Further details regarding the samples employed in this study can be found in Additional file 1: Table S8. The cfDNA methylation sequencing data are available on European Genome-phenome Archive (EGA) under the accession numbers EGAD00001000856 [49] and EGAD00001009003 [50]. All the codes used to generate results can be found at Zenodo [51] with the URL https://doi.org/10.5281/zenodo.14241756, licensed under the MIT license. The codes are also openly available at GitHub [52] with the URL https://github.com/LiymLab/cfDNA_deconv_benchmark, licensed under the MIT license.

Sun *et al. Genome Biology*   (2024) 25:318

Page 25 of 26

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Jingyi Jessica Li is an Editorial Board Member for Genome Biology but was not involved in the editorial process of this manuscript.

References

1. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-free DNA. Cancer Biol Ther. 2019;20:1057–67.
2. Oberhofer A, Bronkhorst AJ, Uhlig C, Ungerer V, Holdenrieder S. Tracing the origin of cell-free DNA molecules through tissue-specific epigenetic signatures. Diagnostics. 2022;12:1834.
3. Stroun M, Maurice P, Vasioukhin V, Lyautey J, Lederrey C, Lefort F, Rossier A, Chen XQ, Anker P. The origin and mechanism of circulating DNA. Ann N Y Acad Sci. 2000;906:161–8.
4. Lui YY, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YM. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clin Chem. 2002;48:421–7.
5. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, Samet Y, Maoz M, Druid H, Arner P, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9:5068.
6. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613:355–64.
7. Volik S, Alcaide M, Morin RD, Collins C. Cell-free DNA (cfDNA): clinical significance and utility in cancer shaped by emerging technologies. Mol Cancer Res. 2016;14:898–908.
8. Corcoran RB, Chabner BA. Application of cell-free DNA analysis to cancer treatment. N Engl J Med. 2018;379:1754–65.
9. Stanley KE, Jatsenko T, Tuveri S, Sudhakaran D, Lannoo L, Van Calsteren K, de Borre M, Van Parijs I, Van Coillie L, Van Den Bogaert K, et al. Cell type signatures in cell-free DNA fragmentation profiles reveal disease biology. Nat Commun. 2024;15:2220.
10. Gao Q, Zeng Q, Wang Z, Li C, Xu Y, Cui P, Zhu X, Lu H, Wang G, Cai S, et al. Circulating cell-free DNA for cancer early detection. Innovation (Camb). 2022;3:100259.
11. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372:eaaw3616.
12. Ma N, Jeffrey SS. Deciphering cancer clues from blood. Science. 2020;367:1424–5.
13. Zhou J, Sears RL, Xing XY, Zhang B, Li DF, Rockweiler NB, Jang HS, Choudhary MNK, Lee HJ, Lowdon RF, et al. Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. BMC Genomics. 2017;18:724.
14. Lokk K, Modhukur V, Rajashekar B, Martens K, Magi R, Kolde R, Koltsina M, Nilsson TK, Vilo J, Salumets A, Tonisson N. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. 2014;15:r54.
15. Schmitz RJ, Lewis ZA, Goll MG. DNA methylation: shared and divergent features across eukaryotes. Trends Genet. 2019;35:818–27.
16. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
17. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
18. Caggiano C, Celona B, Garton F, Mefford J, Black BL, Henderson R, Lomen-Hoerth C, Dahl A, Zaitlen N. Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. Nat Commun. 2021;12:2717.
19. Erger F, Norling D, Borchert D, Leenen E, Habbig S, Wiesener MS, Bartram MP, Wenzel A, Becker C, Toliat MR, et al. cfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Med. 2020;12:54.
20. Keukeleire P, Makrodimitris S, Reinders M. Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads. NAR Genom Bioinform. 2023;5:lqad048.
21. Fox-Fisher I, Piyanzin S, Ochana BL, Klochendler A, Magenheim J, Peretz A, Loyfer N, Moss J, Cohen D, Drori Y, et al. Remote immune processes revealed by immune-derived circulating cell-free DNA. Elife. 2021;10:e70520.
22. Katsman E, Orlanski S, Martignano F, Fox-Fisher I, Shemer R, Dor Y, Zick A, Eden A, Petrini I, Conticello SG, Berman BP. Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. Genome Biol. 2022;23:158.
23. Hill T, Redekar N, Andargie TE, Jang MK, Agbor-Enoh S. Benchmarking and optimization of cell-free DNA deconvolution. bioRxiv. 2023;2023.07.17.549353. https://doi.org/10.1101/2023.07.17.549353.
24. Cheng AP, Cheng MP, Gu W, Sesing Lenz J, Hsu E, Schurr E, Bourque G, Bourgey M, Ritz J, Marty FM, et al. Cell-free DNA tissues of origin by methylation profiling reveals significant cell, tissue, and organ-specific injury related to COVID-19 severity. Med. 2021;2(411–422):e415.

25. Chan KC, Jiang P, Chan CW, Sun K, Wong J, Hui EP, Chan SL, Chan WC, Hui DS, Ng SS, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci U S A. 2013;110:18761–8.
26. Stackpole ML, Zeng W, Li S, Liu CC, Zhou Y, He S, Yeh A, Wang Z, Sun F, Li Q, et al. Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer. Nat Commun. 2022;13:5566.
27. Wu MC, Joubert BR, Kuan PF, Haberg SE, Nystad W, Peddada SD, London SJ. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics. 2014;9:318–29.
28. Liu H, Liu X, Zhang S, Lv J, Li S, Shang S, Jia S, Wei Y, Wang F, Su J, et al. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. Nucleic Acids Res. 2016;44:75–94.
29. Qi H, Song S, Wang P. ImmuMethy, a database of DNA methylation plasticity at a single cytosine resolution in human blood and immune cells. Database (Oxford). 2022;2022:baac020.
30. Barefoot ME, Loyfer N, Kiliti AJ. McDeed APt, Kaplan T, Wellstein A: Detection of cell types contributing to cancer from circulating, cell-free methylated DNA. Front Genet. 2021;12:671057.
31. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, Chan WK, Ma ES, Chan SL, Cheng SH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112:E5503-5512.
32. Li S, Zeng W, Ni X, Liu Q, Li W, Stackpole ML, Zhou Y, Gower A, Krysan K, Ahuja P, et al. Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proc Natl Acad Sci U S A. 2023;120:e2305236120.
33. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.
34. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
35. Song D, Li K, Ge X, Li JJ. ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping. bioRxiv. 2023;2023.07.21.550107. https://doi.org/10.1101/2023.07.21.550107.
36. Hamada M. Fighting against uncertainty: an essential issue in bioinformatics. Brief Bioinform. 2014;15:748–67.
37. Gopalakrishnan S, Van Emburgh BO, Robertson KD. DNA methylation in development and human disease. Mutat Res. 2008;647:30–8.
38. Huan Q, Zhang Y, Wu S, Qian W. HeteroMeth: a database of cell-to-cell heterogeneity in DNA methylation. Genomics Proteomics Bioinformatics. 2018;16:234–43.
39. Bock C, Walter J, Paulsen M, Lengauer T. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. Nucleic Acids Res. 2008;36:e55.
40. Jacoby M, Gohrbandt S, Clausse V, Brons NH, Muller CP. Interindividual variability and co-regulation of DNA methylation differ among blood cell populations. Epigenetics. 2012;7:1421–34.
41. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016;164:57–68.
42. Sun K, Jiang P, Cheng SH, Cheng THT, Wong J, Wong VWS, Ng SSM, Ma BBY, Leung TY, Chan SL, et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res. 2019;29:418–27.
43. Zong W, Kang H, Xiong Z, Ma Y, Jin T, Gong Z, Yi L, Zhang M, Wu S, Wang G, et al. scMethBank: a database for single-cell whole genome DNA methylation maps. Nucleic Acids Res. 2022;50:D380–6.
44. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. Array programming with NumPy. Nature. 2020;585:357–62.
45. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34:D590-598.
46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
47. Memory Profiler. https://github.com/pythonprofilers/memory_profiler.
48. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T, et al. A human DNA methylation atlas reveals principles of cell type-specific methylation and identifies thousands of cell type-specific regulatory elements. Datasets. Gene Expression Omnibus. 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186458.
49. Chan KC, Jiang P, Chan CW, Sun K, Wong J, Hui EP, Chan SL, Chan WC, Hui DS, Ng SS, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Datasets. European Genome-Phenome Archive. 2016. https://ega-archive.org/datasets/EGAD00001000856.
50. Stackpole ML, Zeng W, Li S, Liu CC, Zhou Y, He S, Yeh A, Wang Z, Sun F, Li Q, et al. Methylome sequencing of cell-free DNA and RRBS of solid tissue. Datasets. European Genome-Phenome Archive. 2022. https://ega-archive.org/datasets/EGAD00001009003.
51. Tongyue S, Jinqi Y, Yacheng Z, Jingqi L. Systematic evaluation of methylation-based cell type deconvolution methods for plasma cell-free DNA. 2024. Zenodo. https://doi.org/10.5281/zenodo.14241756.
52. Tongyue S, Jinqi Y, Yacheng Z, Jingqi L. Systematic evaluation of methylation-based cell type deconvolution methods for plasma cell-free DNA. Github. 2024. https://github.com/LiymLab/cfDNA_deconv_benchmark.

Publisher's Note