CelFEER

Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads

Master Thesis

Pia Keukeleire

Delft University of Technology

TU Delft

CelFEER

Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads

by

Pia Keukeleire

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday July 6 at 10 AM

Student number:	4550676	
Project duration:	November 2022 - July 2022	
Degree:	Master Computer Science at the EEMCS faculty	
Thesis committee:	Dr. ir. Stavros Makrodimitris,	Daily supervisor
	Prof. dr. ir. Marcel Reinders,	Supervisor
	Prof. dr. ir. Megha Khosla	

TU Delft

Cell type deconvolution of methylated cell-free DNA at the resolution of individual reads

Pia Keukeleire, Stavros Makrodimitris, and Marcel Reinders

Delft University of Technology

ABSTRACT

Cell-free DNA (cfDNA) are DNA fragments originating from dying cells that enter the plasma. Uncontrolled cell death, for example caused by cancer, induces an elevated concentration of cfDNA. As a result, determining the cell type origins of cfDNA can provide information about an individual's health. This research looks into how to increase the sensitivity of a methylation-based cell type deconvolution method. We do this by adapting an existing method, CelFiE, which uses the methylation values of individual CpG sites to estimate cell type proportions. Our new method, named CelFEER, instead differentiates cell types by the average methylation values of individual reads. We additionally improved the originally reported performance of CelFiE by using a new approach for finding marker regions in the genome that are differentially methylated between cell types. This approach compares the methylation values over 500 bp regions instead of at single CpG sites and solely takes hypomethylated regions into account. We show that CelFEER estimates cell type proportions with a higher correlation ($r^2 = 0.94 \pm 0.04$) than CelFiE ($r^2 = 0.86 \pm 0.09$) on simulated mixtures of cfDNA. Moreover, we found that it can find a significant difference between skeletal muscle cfDNA in ALS patients ($n = 4$) and a control group ($n = 4$); $t(6) = 3.54, p = 0.01$.

1 INTRODUCTION

As cells die, small fragments of their DNA can enter the bloodstream. Consequently, our blood contains traces of multiple different cell types. These fragments of DNA are called cell-free DNA (cfDNA). The cfDNA in our plasma is mostly composed of DNA originating from blood cells [20]. Some diseases, however, cause cells to die in an uncontrolled manner, leaving the DNA incompletely degraded and more prone to enter the bloodstream. The discovery of such disease-derived cell types in cfDNA provides a minimally invasive alternative for tissue biopsies, and is thus frequently referred to as a liquid biopsy [15]. Commonly researched applications of liquid biopsies are prenatal testing, organ transplant monitoring and tumor discovery and monitoring [21]. In all of these applications, however, we know the cell type of interest in advance. Cell type deconvolution, on the other hand, aims to give the full composition of the cell types of circulating cfDNA. An example of a use case in which this type of analysis is especially desirable is finding tumor locations in patients with a cancer of unknown primary [1]. Additionally, characterizing changes in cell type proportions is helpful in understanding disease development and progression [12].

One method for characterizing the cell type origins of cfDNA is the analysis of methylation signatures. Methylation occurs when a methyl-group is added to the fifth carbon of cytosines (5mC), often with the purpose of silencing gene transcription [7]. This process happens mostly in the context of CpG sites, and usually over regions spanning multiple CpG sites [15]. Adjacent CpG sites have been found to correlate highly in methylation status [19]. Because the silencing of gene transcription often happens in a cell type specific manner, these methylation signatures have been found to reveal the cell type origins of cfDNA [21].

Traditionally, cell type deconvolution methods calculate the average methylation of all sequencing reads per CpG site, and use these averages as model input [4, 9, 18]. These averages are often referred to as β values. Although these methods usually do take the correlation between CpG sites into account by averaging over the β values in a region, the value at each individual CpG site is assumed to be independent. In a similar problem setting, namely tumor fraction estimation, Li et al. devised an approach to better incorporate the correlation between sites [14]. Their method, named CancerDetector, calculates the average methylation per individual sequencing read instead of the average methylation per CpG site. They showed how this method outperforms a similar previous method that uses β values [9]. Figure 1b illustrates how rare cell types can be more sensitively detected using read averages than using β values [14]. In this figure, the tumor-derived read makes up 25% of the cfDNA, whereas such rare cell types are far less prevalent in biological data. Since it is essential that our method can deconvolute lowly abundant cell types, β values might not be appropriate.

A read-based approach has been adopted in multiple other tumor fraction estimation methods, such as DISMIR [13] and EpiClass [17]. Even though the effectiveness of this approach has been shown for tumor fraction estimation, it has not yet been used in the related task of cell type deconvolution.

We hypothesize that read averages can increase the sensitivity of methylation-based cell type deconvolution methods. In order to evaluate the effects of using read averages without being affected by other model decisions, we decided to adapt the method CelFiE (CELl Free DNA Estimation via expectation-maximization) by Caggiano et al. [4]. CelFiE has the advantages that it is able to estimate missing cell types and that it can estimate cell type proportions of cfDNA with a low read coverage. Caggiano et al. demonstrated possible clinical applications of CelFiE by showing its ability to differentiate between pregnant and non-pregnant women by their

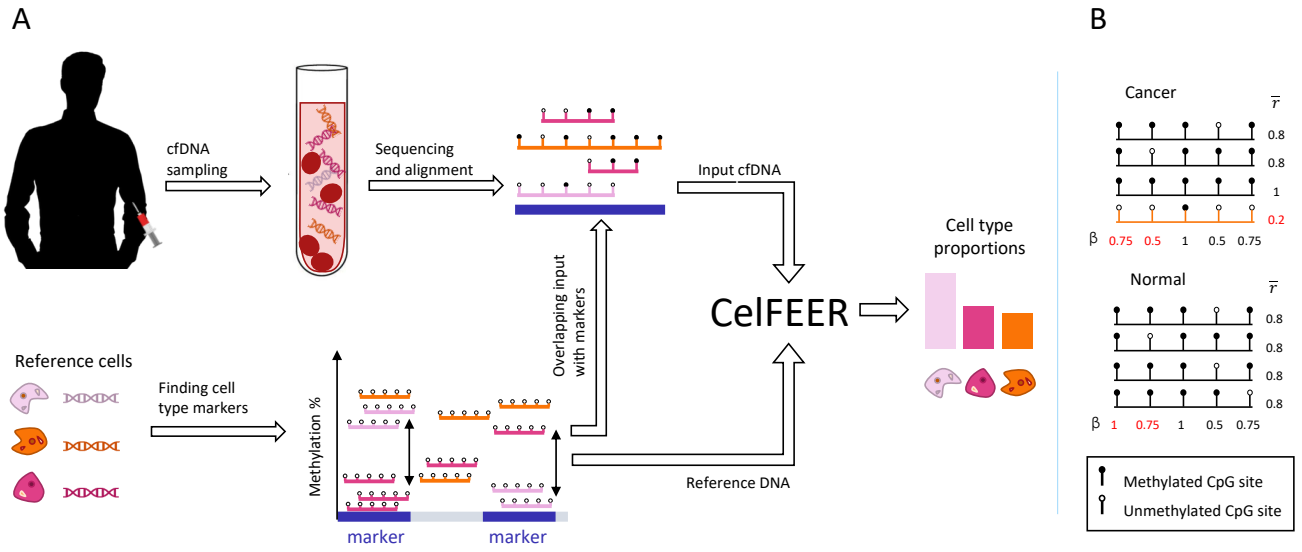


Figure 1. **(A)** Workflow of cell type deconvolution with CelFEER. Sequenced and aligned cfDNA fragments are intersected with cell type marker regions in the genome that are found using reference cell type data. The reference cell data and the cfDNA input data are used as model input for CelFEER, which subsequently outputs the estimated cell type proportions in the cfDNA. **(B)** Toy example illustrating how a tumor-derived read (in orange) can be distinguished from other reads more easily by comparing read averages ($\bar{r}$) instead of CpG site averages (β). Values in red are differential between the cancer and normal sample.

proportion of placenta derived cfDNA, as well as between ALS patients and healthy individuals by their proportion of skeletal muscle cfDNA. In their work, Caggiano et al. used whole genome bisulfite sequencing (WGBS) of reference cell type DNA and input cfDNA. Since WGBS data covers the entire genome, it has the benefit that it can be used for cell type specific biomarker discovery by comparing the methylation in all CpG sites [16].

We find that the selection of appropriate cell type markers is of crucial importance for the model performance. Using the entire genome as model input is not only computationally infeasible, but it will also likely have a negative impact on performance when CpG sites that are not informative of the cell type origin are included. By redefining the cell type informative markers, we improved CelFiE and were able to achieve better results than those reported in the original publication. The new set of markers is found using regions of 500 bp instead of single CpG sites, and only includes hypomethylated markers.

In this research, we adapted CelFiE to work at the resolution of single reads by changing the input to the average methylation value of single reads and by changing the underlying distributional assumptions accordingly. The complete workflow of the resulting method, named CelFEER (CELl Free DNA Estimation via Expectation-maximization on a Read resolution), is depicted in Figure 1a. We compared CelFEER to CelFiE on generated data and on simulated cell type mixtures composed of real WGBS data. We further applied CelFEER on the cfDNA of four ALS patients and four controls, and found that CelFEER detects a significant difference in the proportion of skeletal muscle cfDNA. Our experiments indicate that read averages can indeed more sensitively detect rare cell types.

2 METHODS

2.1 CelFiE overview

As CelFEER is an adaptation of CelFiE, understanding this original method is essential for understanding CelFEER. CelFiE uses an expectation-maximization (EM) algorithm to solve a Bayesian model of the cell type proportions of cfDNA mixtures. It does this by learning these proportions simultaneously with the average methylation percentage of each cell type at each CpG site. The methylation percentages correspond to the fraction of reads that are methylated at a specific CpG site, and are initialized by transforming the reference data into fractions. The methylation percentages are estimated because the reference cell type data is assumed to be imperfect and incomplete; CelFiE aims to learn the true methylation percentages from both the cfDNA input and the reference cell type data. The reference data consists of the methylation counts of T cell types indexed by t at M CpG sites indexed by m. More precisely, it takes the form of two $T \times M$ matrices: Y and D^Y, where Y_{tm} and D^Y_{tm} are the number of methylated and total reads at CpG site m, respectively, in reference cell type t. The reference data is assumed to be drawn from a binomial distribution where the number of trials equals the reference read depth and the probabilities the true

methylation percentage in the cell type of origin.

CelFiE learns the cell type proportions of multiple individuals simultaneously, allowing the method to infer information from other individuals' methylation values. The cfDNA data from N individuals indexed by n is given in two $N \times M$ matrices, X and D^X. X_{nm} and D^X_{nm} are the number of methylated and total reads at CpG site m, respectively, for individual n. An example of how these matrices are formatted is given in Figure 2. Each x_{nmc} refers to the methylation value of a specific read c and can thus be either 0 or 1. These methylation values are assumed to be drawn from a Bernoulli distribution governed by the methylation percentage in the cell type of origin. D^X_{nm} consists of the sum of all x_{nmc} while X_{nm} is the sum of all x_{nmc} that are equal to 1.

CelFiE estimates two parameters: α and β, where α is the final output of the model. α_{nt} is the fraction of cfDNA in person n that originated from cell type t, and β_{tm} is the true unknown methylation percentage of cell type t at position m.

CelFiE models the input cfDNA as a mixture of different cell types. Whether this input originates from a given (or unknown) cell type is modeled using a latent variable z, where $z_{tmc} = 1$ when x_{mc} originates from cell type t and 0 otherwise. The objective is thus to describe the joint distribution $P(X, z, Y|\alpha, \beta)$. For the complete mathematical description of the model and its underlying assumptions, refer to Appendix A and [4].

The model iteratively relates the input to probable cell types in the expectation step, and calculates the parameters that make the input and reference data most likely in the maximization step. More mathematically put, in the expectation step the posterior distribution $\tilde{p}$ of z given the input data x and parameters α and β is calculated. These parameters are then updated by the α and β values that maximize the expectation of the joint likelihood under the calculated posterior.

2.2 Read-based approach

CelFEER uses essentially the same model as CelFiE but with read averages as input. This changes the underlying distributions of the model, while the overall structure of the algorithm remains the same. The algorithm is visualized in Figure 3. In CelFEER, the single counts per CpG site are replaced by five counts per 500 bp region. Each count $\hat{x}_{nmi} \in \hat{X}_{nm}$ for individual n mapping to region m equals the number of reads with a discretized read average i, where $i \in \{0, 0.25, 0.5, 0.75, 1\}$. A read average is calculated by dividing the number of methylated CpG sites by the total number of CpG sites on a read, where only reads with three or more CpG sites are used. This heuristic is adopted from previous methods [13, 14]. The read average is then rounded to the closest value i. E.g. a read c from individual n mapping to region m with one out of three CpG sites methylated (and therefore a read average of 1/3) would be represented as $\hat{x}_{nmc} = \{0, 1, 0, 0, 0\}$. Hence each read is effectively one-hot encoded. Summing all one-hot encoded reads that fall into the same 500 bp region results in the five counts which are used as input to the model. This process is depicted in Figure 2. Binning reads with a similar read average substantially speeds up the method, because this means we only need to estimate the distribution over five possible read averages instead of all possible read averages. In the worst case scenario, the number of possible read averages equals the read depth. Moreover, binning ensures we have more evidence for each of the five distributions to be estimated.

The reference data has the same composition as the input data, but instead of a set of counts per individual per site, the reference data contains a set of counts per cell type per site. Since the reference data has a different format in CelFEER compared to CelFiE, the $\hat{\beta}$ values take on a different form as well. $\hat{\beta}_{tmi}$ is now the proportion of reads originating from cell type t and mapping to region m that have a read average i.

As in CelFiE, the model aims to describe the joint distribution of the input $\hat{X}$, the reference $\hat{Y}$ and the latent variable z, which are all assumed to be independent. In order to describe the full data likelihood, we first split it into three parts: $P(\hat{X}, z, \hat{Y}|\alpha, \hat{\beta}) = P(\hat{X}|z, \hat{\beta})P(z|\alpha)P(\hat{Y}|\hat{\beta})$.

The first part, $P(\hat{X}|z, \hat{\beta})$, describes the likelihood of observing the read averages given that we know what cell type each read originates from and how the read averages of each cell type are distributed across the 500 bp windows. In this likelihood we look at each read c individually, and not yet at the total counts of all reads in a region. The probability for a read c at region m to have the value $\hat{x}_{mc}$ can be described as a categorical distribution where each category corresponds to a possible read average and $\hat{\beta}_{tmi}$ is the probability of originating from cell type t and belonging to category i. This holds for every individual n:

$$\hat{x}_{nmc}|\hat{\beta}_{tm}, z_{ntmc} \overset{\text{iid}}{\sim} \prod_i \hat{\beta}_{tmi}^{z_{ntmc} \cdot \hat{x}_{ci}} \tag{1}$$

The different cell types, individuals, reads and regions are all assumed to be independent. The log-likelihood of the first part can hence be calculated as follows:

$$\begin{aligned} \log P(\hat{X}|z, \hat{\beta}) &= \sum_{n,t,m,c} \log P(\hat{x}_{nmc}|z_{ntmc}, \hat{\beta}_{tm}) \\ &= \sum_{n,t,m,c} z_{ntmc} \left(\sum_i \hat{x}_{nmci} \log \hat{\beta}_{mti} \right) \end{aligned} \tag{2}$$

The second part of the full likelihood describes how likely it is that a read c originates from each cell type t. The probability of observing a specific cell type in the cfDNA is governed by the cell type proportions. This probability can be described using a Bernoulli distribution:

$$z_{ntmc}|\alpha_{nt} \overset{\text{iid}}{\sim} \alpha_{nt}^{z_{ntmc}} \tag{3}$$

Which makes the second part of the log-likelihood:

$$\log P(z|\alpha) = \sum_{n,t,m,c} \log P(z_{ntmc}|\alpha) = \sum_{n,t,m,c} z_{ntmc} \log \alpha_{nt} \tag{4}$$

The final term is the only term that does not depend on the latent variables z. The reference data is assumed to be multinomially sampled with probabilities $\hat{\beta}_{tmi}$ and a number of

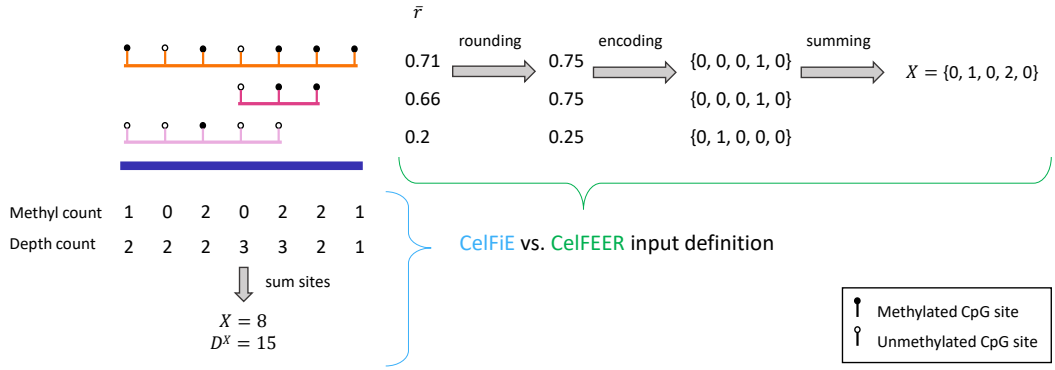


Figure 2. Formatting of the input for CelFiE (bottom left) and CelFEER (top right). On the top left, three partially methylated reads aligning to a 500 bp marker are depicted. For CelFiE, the input is given in two numbers, one equalling the sum of methylated reads at each CpG site and the other equalling the sum of the total amount of reads at each CpG site. For CelFEER, the read averages ($\bar{r}$) are first rounded to the closest value in $\{0, 0.25, 0.5, 0.75, 1\}$, then one-hot encoded and summed to obtain the input. The reference data is formatted in the same way.

trials equal to the reference read depth, which can be obtained by summing over all read average counts:

$$\hat{Y}_{tm}|\hat{\beta}_{tm} \overset{\text{iid}}{\sim} \frac{(\sum_i \hat{Y}_{tmi})!}{\prod_i \hat{Y}_{tmi}!} \prod_i \hat{\beta}_{tmi}^{\hat{Y}_{tmi}} \tag{5}$$

Which makes the third part of the full data likelihood equal to:

$$\log P(\hat{Y}|\hat{\beta}) = n\left(\log(\sum_i \hat{Y}_{tmi})! - \sum_i \log(\hat{Y}_{tmi}!) + \sum_i \hat{Y}_{tmi} \log \hat{\beta}_{tmi}\right) \tag{6}$$

Because of the presence of the latent variables z, there is no closed form solution for maximizing the log-likelihood [3] . Instead, we maximize the expected value of the log-likelihood under the posterior distribution of these latent variables. The posterior distribution of the latent variable z_{ntmc} is calculated by applying the Bayes rule as follows:

$$\begin{aligned} P(z_{ntmc} = 1|\hat{x}_{nmc}, \hat{\beta}, \alpha) \\ &= \frac{P(\hat{x}_{nmc}|z_{ntmc} = 1, \hat{\beta}) P(z_{ntmc} = 1|\alpha)}{P(\hat{x}_{nmc}|\hat{\beta})} \\ &= \frac{\alpha_{nt} \prod_i \hat{\beta}_{tmi}^{\hat{x}_{nmci}}}{\sum_t \alpha_{nt} \prod_i \hat{\beta}_{tmi}^{\hat{x}_{nmci}}} =: \tilde{p}_{ntmc}(\alpha, \hat{\beta}) \end{aligned} \tag{7}$$

Where the distribution of $P(\hat{x}_{nmc}|\hat{\beta})$ follows from the fact that each read originates from only one cell type t, thus summing over all cell types gives the full data distribution of the reads.

Since the read averages are one-hot encoded, there will be five possible values for the posterior $\tilde{p}_{ntmc}$. Following from this fact, we can remove the read index c and can start looking at the total sum of reads that have either of the five possible read averages. For each read c where $\hat{x}_{nmci} = 1$, $\tilde{p}_{ntmc}$ will be equal to:

$$\frac{\alpha_{nt} \hat{\beta}_{tmi}}{\sum_t \alpha_{nt} \hat{\beta}_{tmi}} := p_{ntmi}(\alpha, \hat{\beta}) \tag{8}$$

For the expectation step in the EM algorithm, we need to define the expectation of the latent variable z over the full data likelihood at iteration j.

Let $\alpha^{(j)}$ and $\beta^{(j)}$ equal the parameters estimated at iteration j, and $p^{(j)} := p(\alpha^{(j)}, \beta^{(j)})$. The expectation, also called the Q function, is derived in Appendix A and is defined as follows:

$$\begin{aligned} Q_j(\alpha, \hat{\beta}) &:= \mathbb{E}_{z|\hat{x},\alpha^{(j)},\hat{\beta}^{(j)}} \log P(\hat{x}, z, \hat{Y}|\alpha^{(j)}, \hat{\beta}^{(j)}) \\ &= \sum_{n,t,m,i} ((p^{(j)}_{ntmi} \hat{x}_{nmi} + \hat{Y}_{tmi}) \log \hat{\beta}^{(j)}_{tmi}) + \sum_{n,t,m,i} p^{(j)}_{ntmi} \hat{x}_{nmi} \log \alpha^{(j)}_t \\ &+ n \sum_{t,m} \left[\log(\sum_i \hat{Y}_{tmi})! - \sum_i \log(\hat{Y}_{tmi}!)\right] \end{aligned} \tag{9}$$

Finally α and $\hat{\beta}$ are updated by maximizing $Q_j(\alpha, \hat{\beta})$, resulting in the following update formulas. For the full derivation, see Appendix A.

$$\alpha_{nt} = \frac{\sum_{m,i} p_{ntmi} \hat{x}_{nmi}}{\sum_{m,k,i} p_{nkmi} \hat{x}_{nmi}} \tag{10}$$

$$\hat{\beta}_{tmi} = \frac{\sum_n (p_{ntmi} \hat{x}_{nmi} + \hat{Y}_{tmi})}{\sum_{n,i} (p_{ntmi} \hat{x}_{nmi} + \hat{Y}_{tmi})} \tag{11}$$

Each run of CelFEER performs the optimization 10 times independently, because EM is not guaranteed to converge to a global optimum. The log-likelihood is compared for each restart and CelFEER returns the output from the restart with the highest log-likelihood. In all simulations, we run CelFEER 50 times to capture the variance of the model output.

When including unknown cell types in simulations, we create the true cell type proportions and true cell type methylation in the same fashion as usual. In the reference data that is passed to the model, the methylation values for the unknown cell type are set to $\{0, 0, 0, 0, 0\}$. This way, the estimated methylation percentages for an unknown cell type are initialized to $\{0.2, 0.2, 0.2, 0.2, 0.2\}$.

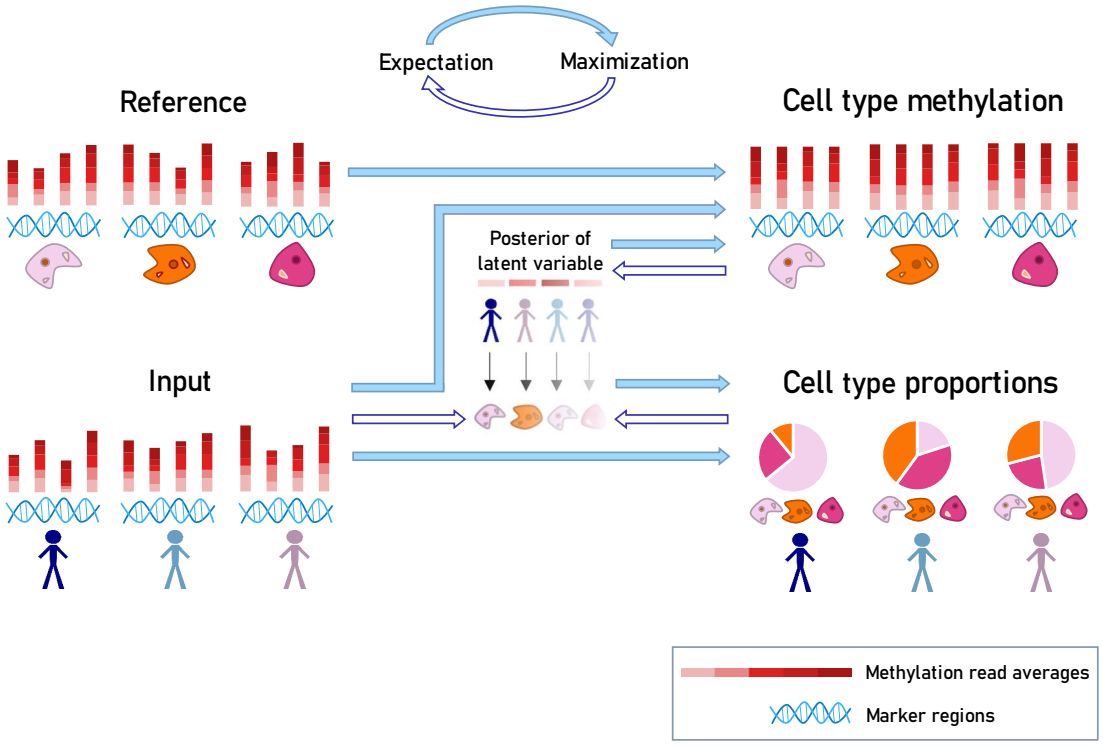


Figure 3. An illustration of the workings of CelFEER for three individuals and three cell types. On the left site of the figure, the reference and input data are depicted. On the right side, the estimated methylation percentages (top) and estimated cell type proportions (bottom) are depicted. In the middle the the latent variable z, which indicates what cell type each individual read (i.e. the methylation average of each read and of each individual) is derived from.

2.3 Marker selection

The markers define which CpG sites will be used as input to the model. The methylation values of CpG sites at marker locations should be consistently different for different cell types, such that the methylation values at these sites can be used to distinguish between cell types. The markers are found using an adaptation of the method used by Caggiano et al.. The complete process of adapting the markers is described in Appendix D. The original method (before adaptation) works as follows. All CpG sites are compared by measuring the distance between the methylation percentage of one cell type to the median methylation percentage of all cell types. The 100 markers with the largest distance are then selected as markers. The total amount of markers found consequently equals 100 times the number of cell types. The markers have to satisfy three requirements in the original method; the first is that a marker is only allowed to be a marker of one cell type. If the same CpG site is in the top 100 of two or more cell types, that site is not used as a marker. The second requirement is that each cell type should have at least one read at a marker location. The last requirement enforces that the median read depth of all cell types at a marker position equals at least 15.

This last requirement, however, still allows the cell type for which the CpG site is a marker to have a read depth less than 15, as long as the median read depth of all cell types is sufficient. A CpG site could be a marker for a cell type as long as it is covered by at least one read of that cell type. To remove the possibility of getting this type of marker, we introduced an extra check to ensure this cell type has a read depth at least as large as the median read depth threshold. Besides, we included one more requirement to ensure marker uniqueness. Instead of comparing only the top 100 markers of each cell type, we compared the top 150 markers of each cell type. After this comparison, again only the top 100 markers are used. This extra step prevents the situation where a marker is in the top 101 of one cell type and in the top 99 of another, which could lead to the inclusion of less differential markers.

The original method should, in theory, be able to find both hypo- and hypermethylated markers. In practice, it finds almost solely hypomethylated markers. Comparing each cell type's methylation percentage to the median methylation percentage can make markers less distinct, as is shown in Figure 4c. Therefore, we adapted the method to compare each cell type's methylation percentage to the minimum methylation percentage of all other cell types, as is shown in Figure 4d. We found that hypomethylated markers are best at differentiating between cell types (Appendix D).

Originally, CelFiE uses as input and as reference data the methylation values at the marker CpG sites summed with the methylation values of CpG sites in the ± 250 bp surrounding

the marker sites. We improved CelFiE by first summing the CpG sites into 500 bp windows which are subsequently used to find marker regions. Otherwise, markers on regions are found using the exact same approach as markers found on single CpG sites. The difference between finding markers on single CpG sites and on regions is shown in Figure 4a and Figure 4b. As there is no requirement for the amount of CpG sites in a region and only for the minimum read coverage of a region, the amount of CpG sites per marker can differ. Because summing the CpG sites into 500 bp windows substantially increased the read coverage at potential marker regions, we increased the read depth threshold to 150. To find the value for this threshold, we tried a range of increasing values and compared the resulting markers by their distance between cell types.

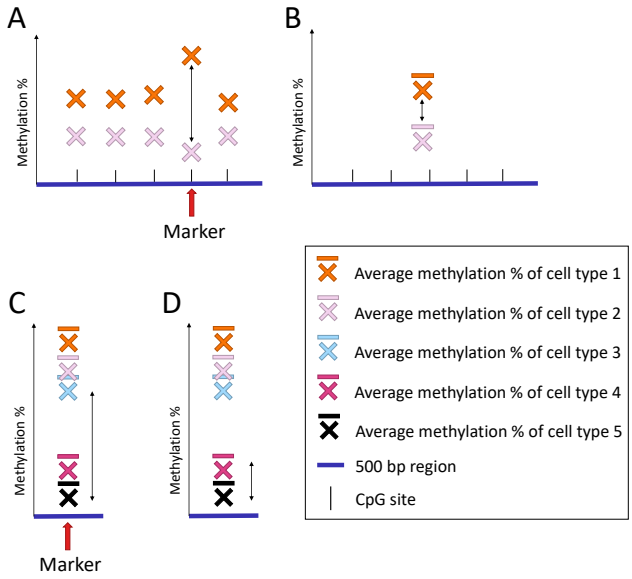


Figure 4. Illustration of the two principal changes to the approach for findings markers in the genome. The arrows indicate the difference between cell types. Figures **(A)** and **(C)** illustrate how the markers are found originally, and Figures **(B)** and **(D)** how they are found after improvements. Figures **(A)** and **(B)** show how measuring the distance between single CpG sites **(A)** results in different markers than measuring the distance between 500 bp regions **(B)**. Cell types 1 and 2 do not have a large distance when regarding their average over the entire region, making this region an unsuitable marker. Figures **(C)** and **(D)** show that the distance from the median cell type **(C)** is different from the distance from the min cell type **(D)**. Using the median would result in a marker that does not differentiate well between cell types 4 and 5.

Finding the markers using the read average data largely follows the same approach. First, the chromosome is split into 500 bp windows into which the reads are mapped. For each

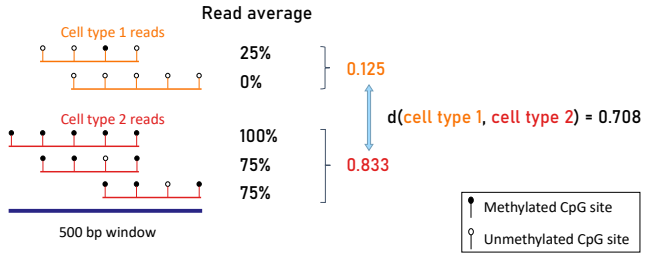


Figure 5. Illustration of method for determining the distance between two cell types. First, the average of all read averages is determined for each cell type. These are then compared to find the distance between cell types.

cell type, the read averages are averaged over all reads that map to the same window. The CelFEER markers are found by comparing these averages. This process is illustrated in Figure 5. For the read averages, we again optimized the read depth threshold and observed that the best markers were found using a read depth threshold of 20. The large difference with the read depth threshold for the CelFiE input (after summing in 500 bp windows) can be explained by the large difference in the scale of the input of CelFiE and CelFEER. This difference in scale is due to the fact that all CpG sites on a read contribute to a single value in CelFEER, and to multiple values in CelFiE.

Since the approach to summarize read averages into bins is slightly different from the approach used to bin the CpG count data, we bin the CpG count data in the same manner as the read averages when comparing CelFiE and CelFEER in subsection 3.2.

2.4 Generated data simulations

In order to validate if CelFEER works under the model assumptions, simulations with artificial data were set up as follows.

The input and reference data are generated according to the distributions assumed by the model. The simulations use the same parameters as originally used by Caggiano et al. in their artificial simulations. In each random restart, α is randomly initialized by drawing from a uniform distribution and normalizing to ensure the values sum to one. β_i is initialized by taking $\frac{Y_i}{\sum_j Y_j}$. This was done for both CelFiE and CelFEER.

2.5 Simulations on WGBS data

To further evaluate the method, we simulated cfDNA data by mixing WGBS data of different cell types. The cell type data was obtained from ENCODE [5] and Blueprint [6], and is composed of T-cell CD4, monocyte, macrophage, memory B cell, neutrophil, adipose, pancreas, small intestine, stomach and tibial nerve data. The sample identifiers of the used data can be found in Table B.1. The data is a mixture of paired-end and single-end reads, and consists of the same datasets used by Caggiano et al.. For each cell type, one sample was used

to compose the reference matrix and one to simulate a cfDNA mixture. Both sex chromosomes were removed, to make the reference matrix applicable to both sexes and to ensure that random methylation due to X chromosome inactivation is not seen as relevant. Furthermore, all SNPs in dbSNP [10] were removed.

To ensure that each dataset contained an equal amount of reads before creating a mixture, the total read coverage of each cell type was normalized by dividing by the total amount of reads of all cell types and multiplying with the average amount of reads. Next, the methylation values of each cell type were multiplied with the desired proportion for that cell type. These proportions were always ensured to add up to one by dividing each cell type's proportion by the sum of all cell types' proportions.

In the original publication [4], WGBS mixtures were created in a similar manner. However, there are two differences in their method for creating the mixtures compared to our approach. First off, Caggiano et al. do not normalize the read coverage. This has as an effect that the mixtures are not actual mixtures, as multiplying the read coverage of cell type X with e.g. 10% does not ensure that 10% of the mixture will be composed of cell type X. Therefore we decided to first normalize the input data. Secondly, Caggiano et al. did not directly multiply the input data with the desired proportions, but multiplied their input matrix X (containing the methylated read counts) with the desired proportions twice, and their D^X matrix (containing the read depths) with the desired proportions once. The reason for doing this is unclear, as this completely changes the methylation percentages of the input.

The mixtures of read averages were made in a similar fashion. First, all read counts were normalized such that each cell type occurred in equal quantities before multiplying the input with the desired proportions.

For both methods the reference data was not normalized. During parameter convergence, the only equation where the reference data is used is Equation 11, where it is transformed to a proportion. The absolute counts of the reference data only matter in their proportion to the input data in Equation 11. It does, however, make sense to not normalize the reference data here since it is logical that reference data with a higher coverage is more reliable and should therefore weigh more in the calculation of β.

3 RESULTS

3.1 Simulations using generated data

To test whether CelFEER works as expected, we followed Caggiano et al. as closely as possible in generating data to simulate cfDNA input and cell type DNA reference data. Using generated data, they showed that CelFiE (i) estimates proportions correlated to the true cell type proportions, (ii) is able to detect small differences between two groups of individuals and (iii) is able to estimate the proportions of unknown cell types (i.e. cell types that are present in the input data, but not in the reference).

The results of these simulations are not an accurate reflection of the model performance, as the simulations for neither CelFiE nor CelFEER model any correlation between sites. As a result, the input of adjacent sites is not summed together as is done for WGBS data, even though Caggiano et al. have shown that the original method does not return sensible results on WGBS data without summing adjacent sites. The simulations do serve as a way of investigating whether CelFEER has the same three properties (which are described above) as CelFiE.

CelFEER estimates of generated data correlate to true proportions

As a first evaluation of the read based method, the performance of CelFEER is compared to the performance of CelFiE on generated data. The simulations use the same input as in [4], meaning that 50 replicates were run, each with 25 cell types, 6000 CpG sites and 1 individual. The read depth at each CpG site was drawn from a Poisson distribution centred around 10.

CelFEER performed slightly worse, with a mean Pearson's correlation $r^2 = 0.84 \pm 0.05$ compared to $r^2 = 0.87 \pm 0.07$ for CelFiE. The result of CelFiE found by us is, however, not as good as the result reported in [4], where the supposedly same simulations result in $r^2 = 0.96 \pm 0.01$.

CelFEER and CelFiE do not detect a significant difference between two groups

Even in individuals with cfDNA originating from aberrant cell types, most of the cfDNA is derived from hematopoietic origins [18]. In other words, the actual amount of cfDNA from an aberrant cell type can be very small. Therefore, it is important to be able to differentiate between a group that does not have this cell type and a group that has only a very small amount of it. To this end, we simulated a cell type that made up a proportion of 0.01 of the cfDNA of five individuals (group A) and 0 of the cfDNA of five other individuals (group B). Ten cell types were used in total on an input of 1000 CpG sites. The remaining nine cell types had a true proportion drawn from a uniform distribution between 0.5 and 1, which were then normalized such that all proportions summed to one.

Figure 7 shows the estimated proportion of the rare cell type for both groups, using both CelFiE and CelFEER. Averaged over 50 replicates, CelFiE estimated a proportion of 0.03 ± 0.01 in group A and 0.025 ± 0.007 in group B, while CelFEER estimated proportions of 0.031 ± 0.01 and 0.026 ± 0.008 for the two groups respectively. A two-samples t-test done for each individual showed no significant difference between the average proportions estimated by both methods in neither groups ($p > 0.1$ for all individuals). Moreover, the proportions of the rare cell type are highly overestimated in both groups.

CelFEER estimates proportions of unknown cell types

One of the advantages of CelFiE over previous deconvolution methods is its ability to infer cell type information from the methylation states of other individuals. This way it can

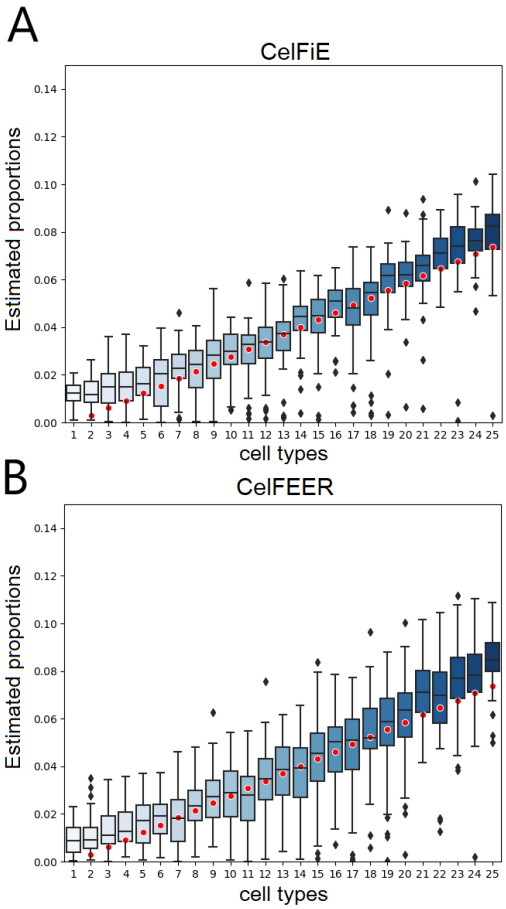


Figure 6. Simulations on generated data for one individual. Each boxplot displays the estimated proportion of a cell type for replicate model runs. The red dots indicate the true cell type proportions for 25 cell types.

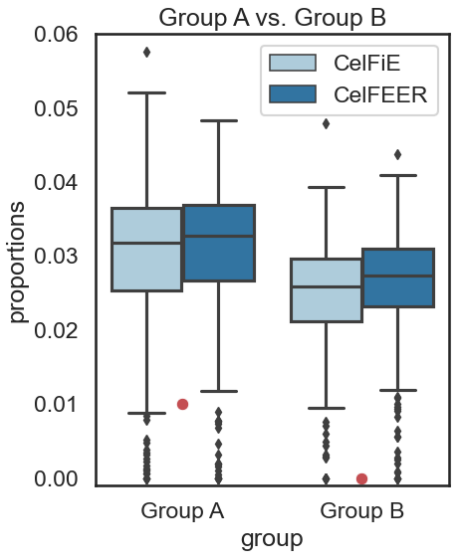


Figure 7. Estimates of the proportion of a rare cell type (1%) that is present in group A but not in group B, estimated over 50 replicate runs using CelFiE and CelFEER. Only the estimated and true proportions of this rare cell type are plotted. The true proportions are represented by the red dots.

estimate the cell type proportions of cell types that are not present in the reference data. As in the original paper, we generated cfDNA for 1000 CpG sites, 10 cell types and 10 individuals at a read depth of 10. In the reference data, we set the methylation states of the last cell type to 0 at each CpG site. The true proportion of this unknown cell type was drawn from a normal distribution centred around 0.2 with a standard deviation of 0.1, and clipped if smaller than 0 or larger than 1. All other cell type proportions were drawn from a uniform distribution between 0 and 1, and together with the unknown cell type the proportions were made to sum to 1. This was done for each individual separately.

We measured the root mean squared error (RMSE) of the estimated proportion of the missing cell type. Averaged over all individuals, CelFEER resulted in an RMSE of 0.0009, and CelFiE in an RMSE of 0.0010. This shows that CelFEER is also capable of estimating proportions of unknown cell types in generated data.

3.2 Results of simulations using WGBS data

Since there are no ground truth cell type proportions available for real cfDNA data, it is impossible to know if the estimated proportions of cfDNA correlate with the true cell type proportions. Therefore, we simulated mixtures of cfDNA by mixing WGBS data of different cell types.

First, we aimed to replicate the results of the original paper [4] before comparing these to the results of our method. For an impartial comparison we used the same data used by Caggiano et al., given in Table B.1. When comparing CelFiE to CelFEER, we were limited to using seven different cell types because of the availability of read data at the time of testing.

Discrepancy with results of original paper

In the original paper [4], the true cell type proportions are made by drawing a proportion of T cells from a normal distribution centered around 20%, a proportion of small intestine cells from a normal distribution centered around 10% and proportions of the eight remaining cells from a random uniform distribution. These eight remaining cell types were normalized such that, together with the proportion of T cells and small intestine cells, they would sum to one. This was repeated 50 times for 100 individuals. They reported to have obtained a Pearson's correlation of $r^2 = 0.83 \pm 0.16$ between the estimated and true cell type proportions, and a correlation of $r^2 = 0.96 \pm 0.01$ between the estimated and true methylation values.

Surprisingly, there is a large discrepancy between the reported results and our replicated results. Following their

approach as accurately as possible, the obtained correlation between the true and estimated cell types was $r^2 = 0.32 \pm 0.19$ and between the true and estimated methylation percentages $r^2 = 0.98 \pm 0.01$. We did find several mistakes in the code and equations published in [4]. After adapting the source code, the correlation was slightly lower for the estimated cell types ($r^2 = 0.30 \pm 0.18$) and slightly higher for the estimated methylation percentages ($r^2 = 0.99 \pm 0.06\text{e}{-2}$). The correlation between the true and estimated methylation percentages is much higher than the correlation between the true and estimated cell type proportions, meaning that an accurate reconstruction of the methylation profiles of the reference data has little effect on the ability to discriminate between cell types. This is an indication that the markers are not representative of their cell type, and that markers which can better discriminate between cell types will result in a more accurate cell type proportion estimation.

When we used our improved set of markers and removed the mistakes from the code, we obtained a higher correlation than reported for both the estimated cell types and the estimated methylation values; $r^2 = 0.87 \pm 0.05$ and $r^2 = 0.99 \pm 0.06\text{e}{-2}$ respectively. These results are not only an improvement on the results we previously obtained, but are also better than the results reported in the original paper [4].

Comparison between CelFiE and CelFEER

To compare the performance of CelFEER to the performance of CelFiE, we again simulated cfDNA mixtures by artificially mixing WGBS data of different cell types. Although we use three cell types less due to data availability, we followed the same approach to create the true cell type proportions for 100 individuals. The marker regions of both models were found using their reference data and were therefore different for the two models, since one set of regions was found by comparing CpG site averages and the other by comparing read averages of different cell types.

In Figure 8, the results of 50 replicate runs for a randomly selected individual are shown. The corrected version of CelFiE with optimized marker regions was used, i.e. not the version that was published by Caggiano et al.. Without unknown cell types in the reference data, CelFEER results in a correlation of $r^2 = 0.94 \pm 0.04$ while CelFiE results in a correlation of $r^2 \pm 0.86 \pm 0.09$. We find that the difference in correlation between CelFEER and CelFiE is significant; $t(9998) = 58.11, p < 0.001$. To examine whether this would go at the expense of runtime, we measured the time it takes each method to run one replicate. On our system, CelFEER requires ~ 1.1 times the time needed by CelFiE.

Since one of the assets of CelFiE is its ability to infer the proportions of unknown cell types, we expected CelFEER to outperform CelFiE on this aspect as well. Similar to the original experiments in [4], we masked T cells in the reference data by setting all T cell reference methylation values to 0. CelFEER highly overestimates the missing cell type proportion and therefore estimates proportions that are less correlated to the true cell type proportions than CelFiE does, although CelFiE also overestimates considerably (see Table 1). When small intestine cells are masked as well, the correlation between the estimated and true cell type proportions decreases even more.

Table 1. Pearson's correlation (r^2) between true and estimated cell type proportions (α estimates) of a simulated mixture of seven different cell types.

Unknowns	CelFiE r^2	CelFEER r^2
0	0.86 ± 0.09	0.94 ± 0.04
1	0.60 ± 0.19	0.48 ± 0.25
2	0.30 ± 0.34	0.19 ± 0.29

In addition to comparing the estimated cell type proportions and their correlation to the true proportions, we investigated the estimated cell type methylation values. We measured the correlation between the estimated cell type methylation percentages and the methylation percentages obtained by normalizing the methylation values of the reference data to sum to one. It is remarkable how this correlation is consistently higher for CelFiE (Table 2). This implies that the methylation percentages estimated by CelFiE diverge only very little from the reference methylation. This probably means that CelFEER takes the input of other individuals more into account when estimating the methylation values, and therefore indirectly when estimating the cell type proportions.

Another advantage of CelFiE over previous methods is that it works with low coverage input data. A higher read coverage means higher sequencing costs, and it is therefore desirable that CelFEER performs sufficiently on low coverage data as well. To test this, before mixing the cell types we normalized the read coverage of each cell type to equal the total amount of input regions multiplied with a constant, n. This way, each cell type covered each region with n reads on average. For each $n \in \{2, 5, 10, 50\}$ the average correlation over 50 replicates and 100 individuals was measured. The cell type proportions were generated in the same manner as before, and no unknowns were estimated. The relation between the correlation and the coverage is shown in Figure 9. We can conclude that for a stable performance, the coverage should be 10 or higher. Interestingly, the correlation between the estimated and true cell type proportions increases a little for CelFiE when $n = 5$. It is possible that lowering the coverage acts as a noise reduction on the CelFiE input. Even on the lowest coverage, CelFEER outperforms CelFiE, showing that CelFEER is a suitable method for low coverage data.

Markers found on read averages are different from markers found on count input

Finally we were interested in comparing the markers found using read averages to the markers found using CpG site averages. We hypothesised that CelFEER works better with markers found on the read averages of the reference data, on the grounds that CelFEER differentiates cell types by their

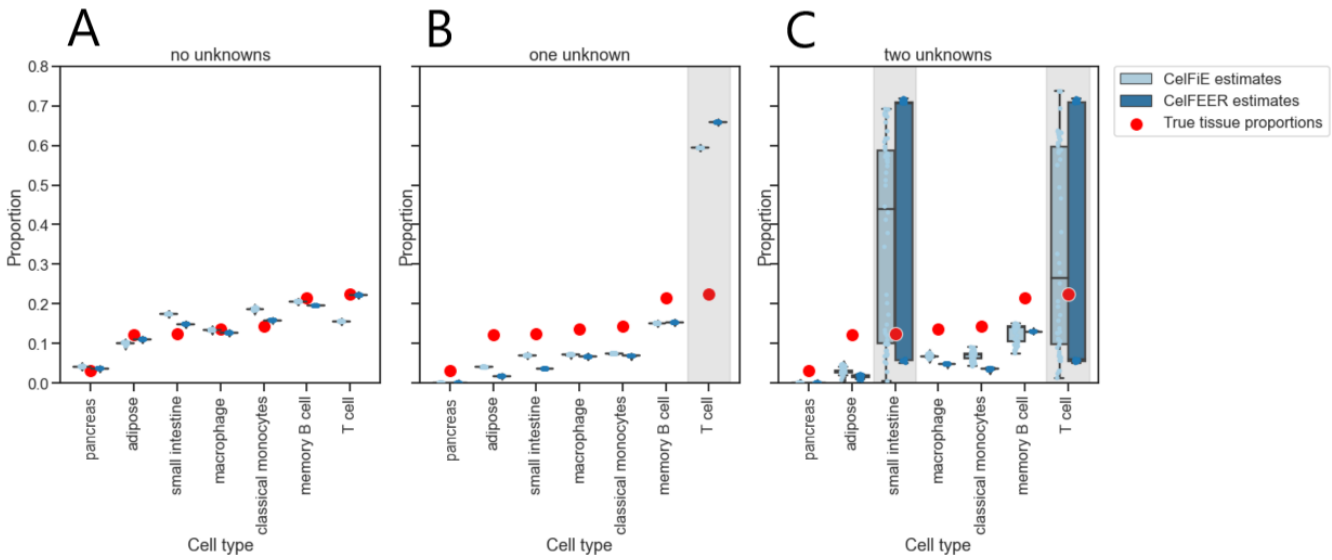


Figure 8. Cell type proportions estimated by CelFiE and CelFEER for zero, one and two unknowns respectively. The boxplots visualize the estimated proportions of 50 replicates for a randomly chosen individual. On top of the boxplots, the individual datapoints are plotted.

Table 2. Pearson's correlation (r^2) between reference methylation and estimated methylation values (β estimates) of a simulated mixture of seven different cell types.

Unknowns	CelFiE r^2	CelFEER r^2
0	$9.98e{-1} \pm 0.03e{-1}$	0.93 ± 0.03
1	0.92 ± 0	0.89 ± 0.11
2	0.85 ± 0.25	0.77 ± 0.26

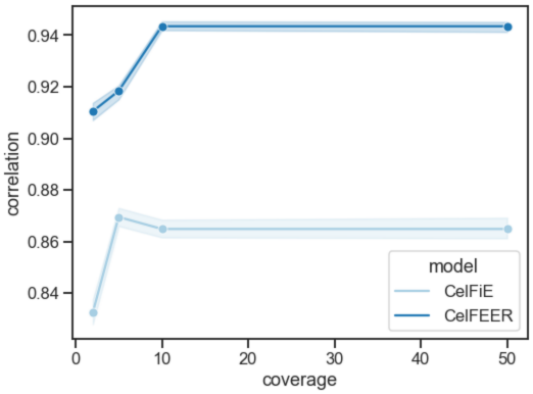


Figure 9. Relation between the input coverage and the correlation between the estimated and true cell type proportions. The full range of the correlations of 100 individuals and 50 replicates is highlighted.

read averages. Additionally, as reasoned in the introduction, read averages should be more sensitive to differences in methylation status between cell types. We again performed the same experiments, using a simulated mixture of seven different cell types.

We firstly checked the overlap in markers found using both methods. Of the 700 markers, 130 markers were found by both. Each of the seven cell types has markers that are found by both methods. There are no markers that are a marker for one cell type in one method and a marker for another cell type in the other method.

Using the markers found by CelFiE, CelFEER performed similarly with a correlation of $r^2 = 0.94 \pm 0.04$ (Figure C.1). The correlation between the cell type proportions estimated by CelFiE using CelFEER's markers is $r^2 = 0.69 \pm 0.21$, indicating that the markers found by CelFEER are not suitable for the input of CelFiE. Averaged over all cell types, the difference in methylation percentage between cell types at CelFiE's marker locations is 0.65 for both the reference and input data, where the reference data showed slightly less variation with a standard deviation of 0.19 compared to 0.20 for the input data. For CelFEER, this difference is 0.66 ± 0.20 for the input and 0.64 ± 0.22 for the reference. Figure C.2 does show that for some cell types the variation in the distance from the median is substantially larger for the CelFEER markers.

3.3 Application in ALS

Caggiano et al. showed that CelFiE is able to differentiate between Amyotrophic Lateral Sclerosis (ALS) patients and a control group by the estimated proportion of skeletal muscle derived cfDNA. Although it is interesting to see if CelFEER is also able to distinguish between the ALS and the control group, it is hard to evaluate the method based on its cell type proportion estimates since there are no ground truth cell type proportions available. Moreover, while Caggiano et al. used

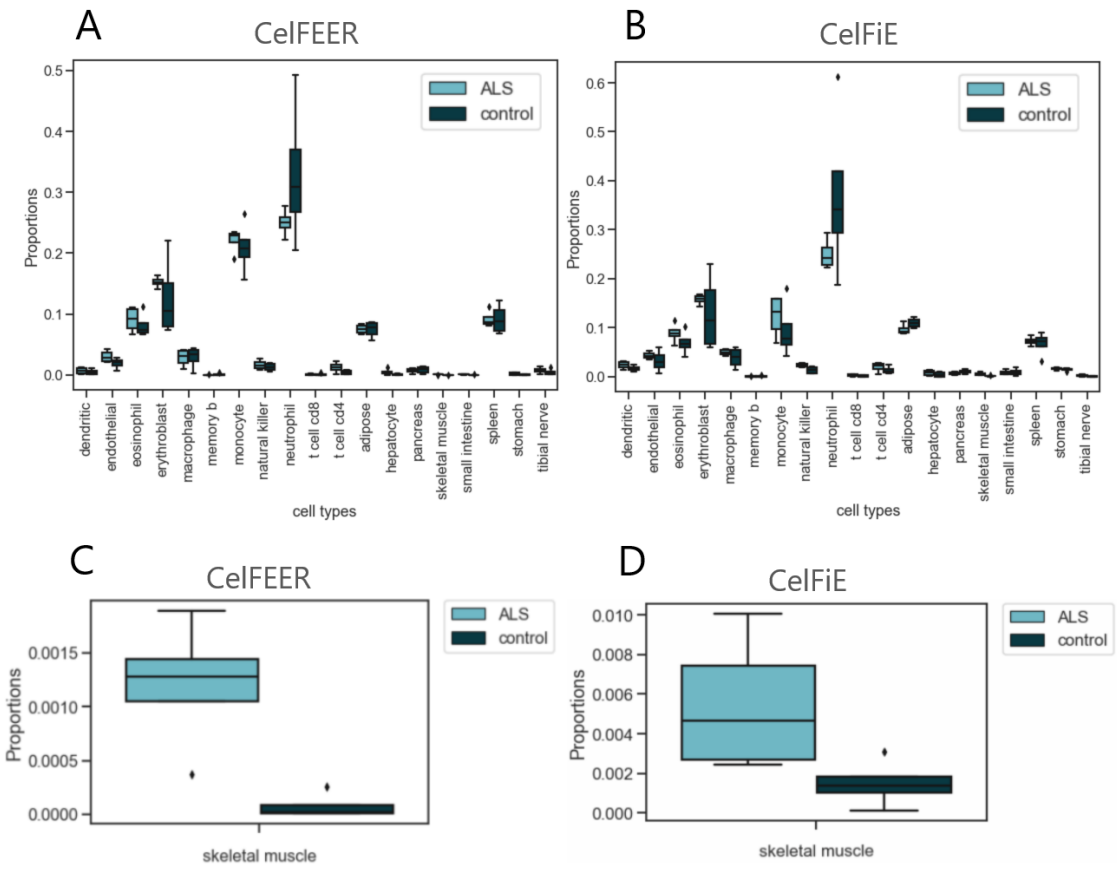


Figure 10. Estimated proportions of cfDNA in ALS patients ($n = 4$) and a control group ($n = 4$). **Figures A, B** The complete cell type decomposition of CelFEER and CelFiE respectively. **Figures C, D** The estimated proportion of skeletal muscle cfDNA by CelFEER and CelFiE.

28 case and 25 control samples, we only used four case and four control samples. The reference data consists of all 19 cell types given in Table B.1.

We firstly decomposed the cfDNA without unknown cell types in the reference data, thus estimating the proportions of each of the 19 cell types present in the reference. The five cell types with the highest proportions estimated by CelFEER were, in both groups, the following: neutrophil, monocyte, erythroblast, spleen and eosinophil. CelFiE estimated similar proportions, but instead of spleen it estimated adipose to be the fourth highest in proportion. In their own work [4], however, neither spleen nor adipose, but macrophage cells are in this top five. Still, these results mostly correspond to the findings of Moss et al. [18]. The full decomposition can be seen in Figure 10a and Figure 10b.

Next, we specifically examined the skeletal muscle cell proportions in both groups. CelFiE estimated an average proportion of $5.5\text{e}{-3} \pm 3.1\text{e}{-3}$ in the ALS case group, and $1.5\text{e}{-3} \pm 1.1\text{e}{-3}$ in the control group (Figure 10d). A two-sample t-test did not indicate a significant difference between the two groups; $t(6) = 2.09, p = 0.08$. CelFEER, on the contrary, did find a significant difference, with an average proportion of $1.2\text{e}{-3} \pm 5.4\text{e}{-4}$ for the ALS case group and $7.7\text{e}{-5} \pm 1\text{e}{-4}$ for the control group (Figure 10c); $t(6) = 3.54, p = 0.01$. Clearly, CelFEER is able to detect small fractions of rare cell types in cfDNA.

As Caggiano et al. estimated the cell type proportions in [4] with one unknown cell type, we repeated the experiments with one unknown. However, as shown in subsection 3.2, CelFEER is not adept for estimating unknown cell types, and estimated an extremely high proportion of unknown cell types in both the ALS and control groups (0.93 ± 0.01 and 0.86 ± 0.12 respectively). CelFiE estimates more likely proportions (0.18 ± 0.02 and 0.17 ± 0.05 for case and control respectively). Consequently, the estimated skeletal muscle proportion was extremely small for CelFEER; $1.4\text{e}{-12} \pm 1.5\text{e}{-12}$ for case and $4.9\text{e}{-17} \pm 6.4\text{e}{-17}$ for control (Figure C.3b). These estimates were higher for CelFiE; $1.3\text{e}{-4} \pm 2\text{e}{-4}$ for case and $1.2\text{e}{-5} \pm 9.5\text{e}{-6}$ for control (Figure C.3a). Both methods were still able to differentiate between the case and control groups by the proportion of skeletal muscle, although the difference was not significant for either method

($t(6) = 1.04, p = 0.34$ for CelFiE and $t(6) = 1.63, p = 0.15$ for CelFEER). The extreme overestimation of the unknown cell type proportion in CelFEER means that a large part of the ALS input consists of a more or less equal mixture of different read averages. Since the unknown cell type is initialized with methylation percentages set to make each read average equally likely, the estimated proportions will converge to a large proportion of unknown cells. This is another indication that the unknown cell type may need to be initialized differently.

4 DISCUSSION AND OUTLOOK

The analysis of cfDNA has some attractive properties, such as the possibility to detect and monitor disease without undertaking aggressive surgery [15]. By retrieving the cell types of origin of cfDNA, it is possible to obtain a complete overview of all cells that shed cfDNA, and even of the amount of cfDNA each cell type yields. An inquiry in the cell type proportions can indicate the presence of aberrant cell types, such as tumor cells, in the cfDNA. Yet, detection of aberrant cell types can be difficult, especially in early stages of disease. Recent methods use the methylation states at CpG sites that cause a differential gene expression in different cell types. In this research, we adapted one such method, CelFiE [4], to instead use differential methylation averages of individual reads. The intuition behind this approach is that the methylation averages of individual reads differentiate more than CpG site averages, since aberrant reads are almost undetectable when averaged with healthy reads. This new method, named CelFEER, uses an expectation-maximization algorithm and a reference cell type dataset to estimate the true cell type proportions of cfDNA mixtures.

We first compared the performances of CelFiE and CelFEER on a generated dataset, where the cfDNA mixtures and reference cell types are drawn from the assumed underlying distributions of the models. Although these simulations can provide us insight in whether CelFEER returns sensible results, the simulations do not model the input as a sum over multiple correlated CpG sites. For this reason, it makes sense that CelFiE performs better, as the model does not fully exploit the correlation between neighbouring CpG sites. CelFEER performs similarly, although slightly worse than CelFiE.

To evaluate the model in a more realistic scenario, we created mixtures of different cell type WGBS datasets. We then evaluated the correlation between the model estimates and the artificially created mixture proportions. Surprisingly, we could not replicate the results of Caggiano et al. when running CelFiE on these mixtures. It is not clear why this discrepancy between the results exists. Although it is possible that the authors forgot to mention important steps and decisions, it seems unlikely given that their code has been made public. It is possible that some of the smaller mistakes the authors made, for instance failing to normalize the read coverage before creating a simulated cell type mixture, coincidentally improved performance. After improving CelFiE, we found that CelFEER nonetheless estimates proportions that better correlate with the true cell type proportions than CelFiE's estimates. We additionally showed that CelFEER suffers less from a low read coverage, and performs well even with an artificially induced coverage of two reads per 500 bp window. Strangely, CelFEER performs badly when estimating unknown cell types.

We showed that CelFEER functions on actual cfDNA mixtures as well by running the model on the cfDNA of four ALS patients and four controls. As expected from previous literature, the main cell types found are from hematopoietic origin. CelFEER finds a significant difference between the two groups by differentiating between the estimated proportions of skeletal muscle cfDNA.

The model's performance is highly reliant on the quality of the input regions, where the quality is defined by the difference in methylation between cell types at an input region. In pursuit of improving CelFiE's model performance, we improved the original method for finding markers by applying the following changes: (i) we differentiated between 500 bp regions instead of single CpG sites, (ii) we focused on hypomethylated regions and (iii) we applied stricter rules to marker regions. To find marker regions for CelFEER, we devised a method that largely follows the same approach as CelFiE but instead uses the read averages of the reference data.

The read averages are formulated in a way that one read average, so one single value, summarizes multiple CpG sites. For this reason, the range of the input is much lower for CelFEER than for CelFiE. In addition, CelFEER filters out reads covering less than 3 CpG sites, which decreases the range even more. It may be interesting to investigate whether allowing for reads with a lower CpG site coverage gives improvements to the model. Low read quality is one of the disadvantages of working with WGBS data, as the bisulfite conversion is known to be detrimental to the DNA [11]. Another way for compensating for the smaller range would be to increase the amount of samples used in the reference dataset. Currently, each reference cell type consists of the DNA of a single individual.

If the reference data does not include all of the cell types found in the cfDNA, the proportions of the cell types that are included will be overestimated. Since actual cfDNA is likely to contain a component of cell types that are absent from the reference data [4], it is useful to estimate proportions of unknown cell types. However, CelFEER currently greatly overestimates the proportions of unknown cell types. It may be possible to improve this by changing the input for unknown cell types, as we presently employ CelFiE's method of setting unknown cell types to 0, which may not work for CelFEER. In relation to that, we may need to change the initial values for the estimated methylation percentages for unknown cell types. Currently, the methylation percentages for unknown cell types are initialized with $\hat{\beta}_{tmi} = 0.2$ for every i.

Despite the improvements made to the selected marker regions, there is potential for more distinct markers, perhaps

by adopting a completely new approach. After all, the method for finding markers was optimized for CpG count data and then translated almost exactly to read average data. Read averages may, however, require a completely different approach for finding markers, such as the switching reads defined by Li et al. [13]. An adequate set of differential regions not only improves model performance but also allows for targeted sequencing of these regions only, for example using RRBS, and can thus reduce the sequencing cost [2].

Although the input size of CelFEER is larger than the input size of CelFiE (read averages are described by five counts instead of the two counts used by CelFiE), it suffers only from a minor increase in runtime. Like CelFiE, CelFEER is an efficient method that scales linearly in the size of the input and reference. Even so, it could be beneficial to consider CelFEER's performance when using more or less counts. Using less counts, i.e. rounding the read averages more before summing similar averages, would likely decrease model performance but speed up computations. Using more counts, on the other hand, may give an increase in performance that is worth the added computation time.

Finally, the use of CelFEER in practical applications should be investigated further by testing the model on more cfDNA data. A first step would be to use more samples in the ALS experiment. Eventually the model could be tested on, for instance, pregnancy and cancer samples.

With CelFEER, we showed that a cell type deconvolution method can more sensitively estimate cell type proportions when using read averages instead of CpG site averages, even at a low input read coverage.

A EQUATIONS

Original CelFiE equations

Posterior distribution:

$$\begin{aligned} p_{ntm1}(\alpha,\beta) &:= p_{ntmc}(\alpha,\beta) \qquad \text{if } x_{nmc}=1 \\ &= \frac{\beta_{tm}\alpha_{nt}}{\sum_k \beta_{kt}\alpha_{nk}} \\ p_{ntm0}(\alpha,\beta) &:= p_{ntmc}(\alpha,\beta) \qquad \text{if } x_{nmc}=0 \\ &= \frac{(1-\beta_{tm})\alpha_{nt}}{\sum_k (1-\beta_{kt})\alpha_{nk}} \end{aligned} \tag{12}$$

α and β update formula:

$$\alpha_{nt} = \frac{\sum_m \left(x_{nm}p_{ntm1} + (D^X_{nm} - x_{nm})p_{ntm0}\right)}{\sum_{km}\left(x_{nm}p_{nkm1} + (D^X_{nm} - x_{nm})p_{nkm0}\right)} \tag{13}$$

$$\beta_{tm} = \frac{\sum_n p_{ntm1}X_{nm} + nY_{tm}}{\sum_n p_{ntm0}(D^X_{nm} - X_{nm}) + nD^Y_{tm} + \sum_n p_{ntm1}X_{nm}} \tag{14}$$

Log-likelihood formulation:

$$\begin{aligned} Q(\alpha,\beta) = &\sum_{n,t,m}\left[(Y_{tm} + p_{ntm1}X_{nm})\log(\beta_{tm}) + \left(D^Y_{tm} - Y_{tm} + p_{ntm0}(D^X_{nm} - X_{nm})\right)\log(1-\beta_{tm})\right] \\ &+ \sum_{n,t,m}\left(X_{nm}p_{ntm1} + (D^X_{nm} - X_{nm})p_{ntm0}\right)\log\alpha_{nt} \end{aligned} \tag{15}$$

Derivation of full data log-likelihood

$$\begin{aligned} Q(\alpha,\hat\beta) &:= \mathbb{E}_{z|\hat X,\alpha,\hat\beta}\log P(\hat X, z, Y|\alpha,\beta) \\ &= \mathbb{E}_{z|\hat X,\alpha,\hat\beta}\left(\log P(\hat X|z,\hat\beta) + \log P(z|\alpha) + \log P(\hat Y|\hat\beta)\right) \\ &= \sum_{n,t,m,c}\mathbb{E}_{z|\hat X,\alpha,\hat\beta}\left[z_{ntmc}\sum_i \hat x_{nmci}\log\hat\beta_{tmi} + z_{ntmc}\log\alpha_{nt}\right] \\ &+ \sum_{n,t,m}\left(\log(\sum_i \hat Y_{tmi}!) - \sum_i \log(\hat Y_{tmi}!) + \sum_i \hat Y_{tmi}\log\hat\beta_{tmi}\right) \\ &= \sum_{n,t,m,c}\tilde p_{ntmc}\left[\sum_i \hat x_{nmci}\log\hat\beta_{tmi} + \log\alpha_{nt}\right] \\ &+ \sum_{n,t,m}\left(\log(\sum_i \hat Y_{tmi}!) - \sum_i \log(\hat Y_{tmi}!) + \sum_i \hat Y_{tmi}\log\hat\beta_{tmi}\right) \\ &= \sum_{n,t,m}\left[\sum_i p_{ntmi}\hat x_{nmi}\log\hat\beta_{tmi} + \sum_i p_{ntmi}\hat x_{nmi}\log\alpha_{nt}\right] \\ &+ \sum_{n,t,m}\left[\log(\sum_i \hat Y_{tmi}!) - \sum_i \log(\hat Y_{tmi}!) + \sum_i \hat Y_{tmi}\log\hat\beta_{tmi}\right] \\ &= \sum_{n,t,m,i}((p_{ntmi}\hat x_{nmi} + \hat Y_{tmi})\log\hat\beta_{tmi}) + \sum_{n,t,m,i} p_{ntmi}\hat x_{nmi}\log\alpha_{nt} + n\sum_{t,m}\left[\log(\sum_i \hat Y_{tmi})! - \sum_i \log(\hat Y_{tmi}!)\right] \end{aligned} \tag{16}$$

Derivation of α and $\hat\beta$ update formulas

Maximization of the log-likelihood w.r.t α and $\hat\beta$ can be done using the following fact that for a probability simplex $S_K \subset \mathbb{R}^K$ and any $a \in \mathbb{R}^K_{++}$:

$$\arg\max_{p\in S_K}\sum_k a_k\log p_k = (a_1,...,a_K)/\sum_{k=1}^K a_k$$

To derive α_t, we let $a_t = \sum_i p_{tmi}\hat x_i$ s.t.

$$\alpha_{nt} = \frac{\sum_{m,i} p_{ntmi}\hat x_{nmi}}{\sum_{m,t,i} p_{ntmi}\hat x_{nmi}} \tag{17}$$

For $\hat{\beta}_{tmi}$ we let $a_i = p_{tmi}\hat{x}_i + \hat{Y}_{tmi}$ s.t.

$$\hat{\beta}_{tmi} = \frac{\sum_n (p_{ntmi}\hat{x}_{nmi} + \hat{Y}_{tmi})}{\sum_{n,i} (p_{ntmi}\hat{x}_{nmi} + \hat{Y}_{tmi})} \tag{18}$$

B DATA

Table B.1. WGBS cell type data and sources

Cell type	Database	Sample 1	Sample 2
CD4-positive, alpha-beta T cell	Blueprint	S007G7	S007DD
CD8-positive, alpha-beta T cell	Blueprint	C003VO	C00256
endothelial cell of umbilical vein (resting)	Blueprint	S00DCS	S00BJM
monocyte	Blueprint	S01MAPA1	S01E03A1
erythroblast	Blueprint	S002S3	S002R5
macrophage	Blueprint	S0022I	S00390
mature eosinophil	Blueprint	S00V65	S006XE
memory B cell	Blueprint	C003N3	S017RE51
cytotoxic CD56-dim natural killer cell	Blueprint	C006G5	C002CT
mature neutrophil	Blueprint	C0010K	C000S5
conventional dendritic cell	Blueprint	S00CP651	S00D71
adipose	ENCODE	ENCFF318AMC	ENCFF477GKI
HepG2	ENCODE	ENCFF847OWL	ENCFF064GJQ
pancreas	ENCODE	ENCFF753ZMQ	ENCFF500DKA
small intestine	ENCODE	ENCFF266NGW	ENCFF122LEF
spleen	ENCODE	ENCFF550FZT	ENCFF333OHK
stomach	ENCODE	ENCFF435SPL	ENCFF497YOO
tibial nerve	ENCODE	ENCFF843SYR	ENCFF699KTW
skeletal muscle myoblast primary cell	ENCODE	ENCFF774GXJ	-

C SUPPLEMENTARY FIGURES

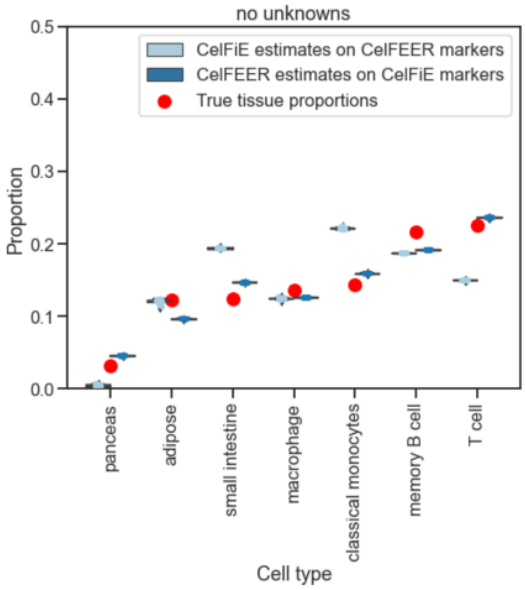
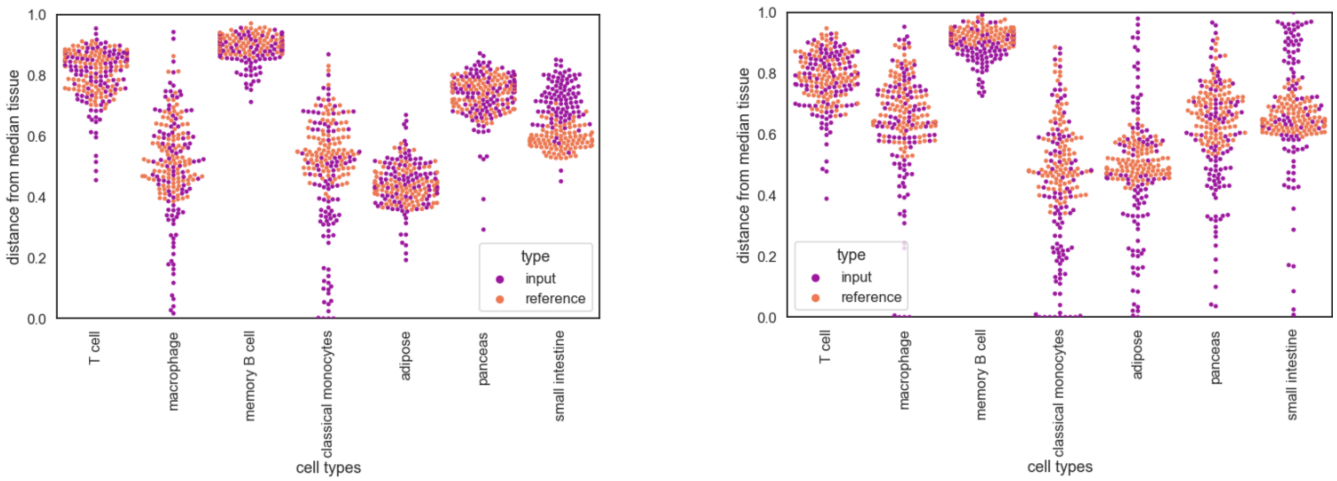


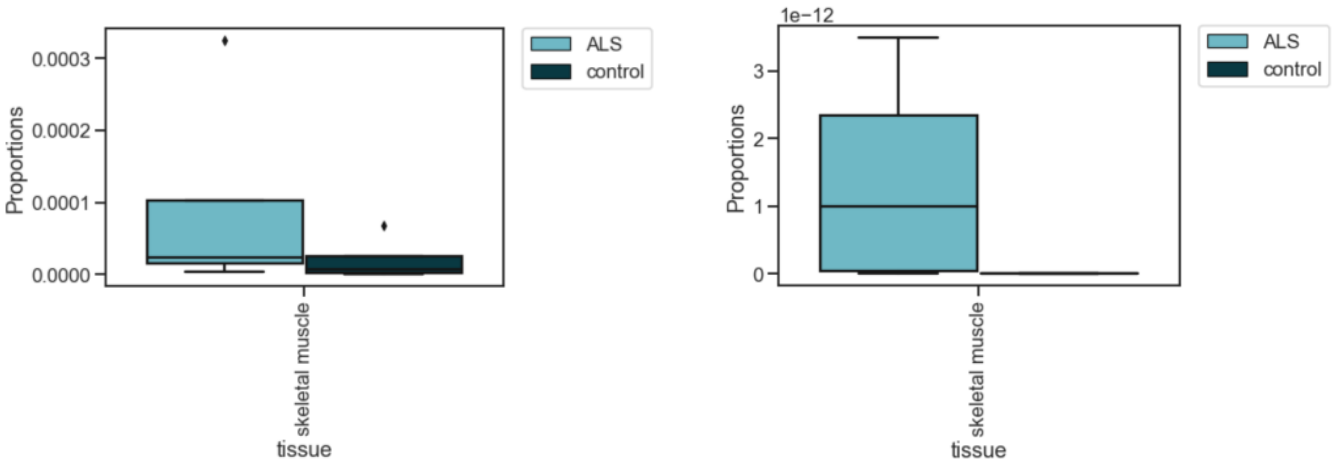
Figure C.1. CelFiE and CelFEER run on different markers.



(a) CelFiE markers

(b) CelFEER markers

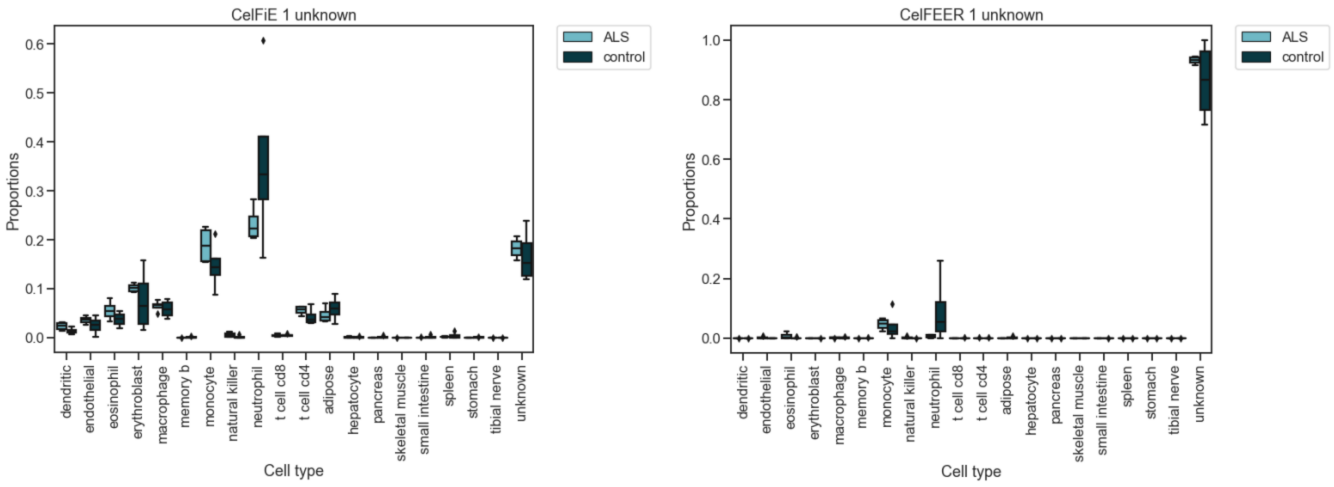
Figure C.2. Markers found by **(a)** CelFiE and **(b)** CelFEER for seven different cell types.



(a) CelFiE

(b) CelFEER

Figure C.3. Estimated proportions of skeletal muscle cfDNA when one unknown cell type is estimated in addition to the total 19 cell types in the reference data.



(a) CelFiE ALS and control cell type decomposition

(b) CelFEER ALS and control cell type decomposition

Figure C.4. Complete cell type decomposition when one unknown cell type is estimated.

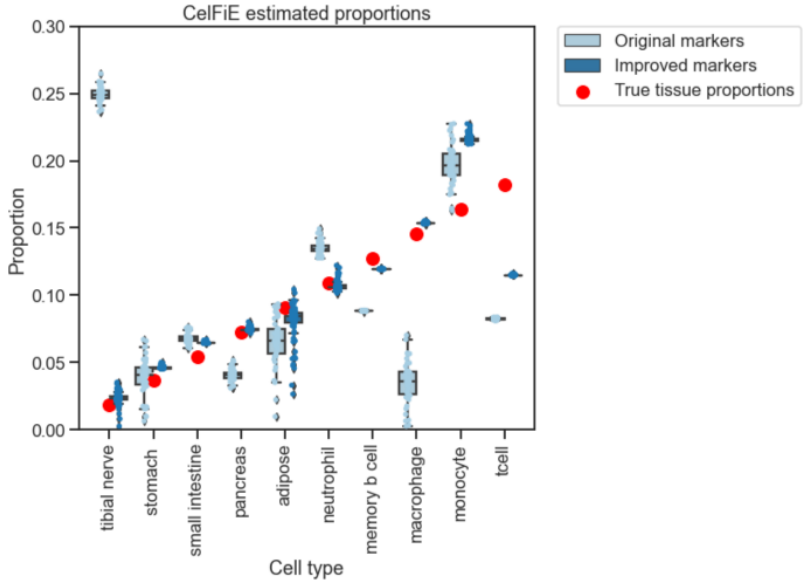


Figure C.5. CelFiE estimated cell type proportions on a simulated cfDNA mixture using WGBS cell type data, using both the markers found as described in [4] and the markers found using our improved method. For visualisation purposes the true cell type proportions are a simple incremental array summing to one. The results of 50 replicate runs on 10 individuals are displayed.

D SELECTION OF CELL TYPE INFORMATIVE MARKERS

A crucial step in predicting the cell type of origin is selecting markers in the genome that represent the cell types. Not only does a set of distinct markers improve prediction, it can make sequencing of cfDNA less expensive since only the DNA overlapping the markers needs to be sequenced. Methylation markers that span multiple CpG sites are in literature often referred to as differentially methylated regions. To find cell type informative markers, we started by analyzing the markers found using the method created by Caggiano et al., which is described in subsection 2.3. This method was then improved to find more informative markers. In this section and the following we refer to the absolute counts of methylated CpG sites as methylation values, and to the fraction of methylated to unmethylated CpG sites as methylation percentages.

Regions are more robust markers than single sites

Caggiano et al. use the traditional approach of using single CpG sites as markers. This method, however, decreases the ability to differentiate between different cell types as it is sensitive to both biological and technical noise. In order to reduce noise, the CpG sites 250 bp upstream and 250 bp downstream of the markers are added to the markers' methylation counts. The authors showed that their method only returns sensible results when the methylation values are thus summed into regions. It nonetheless happens that the 500 bp surrounding the markers contain little CpG sites. This method does not exploit earlier findings that the methylation status is highly coupled between adjacent CpG sites [8]. Moreover, regions where CpG sites are clustered in high numbers, called CpG islands (CGIs), are known to be epigenetic regulatory regions that can be cell type specific [22].

According to these findings, it makes more sense to compare regions containing multiple CpG sites instead of single CpG sites to find differential markers. To test this hypothesis, CpG sites were grouped in a simple fashion: CpG sites were summed if they were in a 500 bp vicinity of each other. The starting location of each 500 bp window was set to be the first CpG site which contained measurements and did not fall in a previous bin. This strategy has the downside that it may split clusters in two, but if this is the case and if this cluster is differential, it is not harmful for the method to use both parts of the cluster as markers.

In addition to summing over 500 bp windows, we also summed over 10 bp windows with the idea of removing noise while still looking at mostly local methylation. After finding markers on the 10 bp windows, the surrounding CpG sites were summed to nevertheless obtain a total window of 500 bp. In order to compare the markers' ability to differentiate between cell types, we looked at the absolute difference between the methylation percentage of each marker's cell type and the median methylation percentage of all cell types. To test the generalizability of the markers, we did this for both the reference data (which was used to find the markers) and for the input data. As can be seen in Figure D.1, the markers are most differential when they are first summed in 500 bp windows, and the variance in distance has substantially decreased. This strategy also seems to result in markers that generalize relatively well to unseen data, as the input and reference data have a similar distance to the median of other cell types. Although summing in 500 bp windows seems to return better markers than summing in 10 bp windows, it is remarkable how much improvement can be seen compared to the original method, especially for the tibial nerve cells. This is probably the effect of the decrease in noise which appears even if we sum over such small intervals. The results confirm the belief that markers are more differentiable when CpG sites are first summed compared to when they are summed after selecting individual sites. For this reason, all future experiments on markers are done on sites summed in 500 bp regions. In this section, we used only hypomethylated markers as they promised to be most distinguishing between cell types.

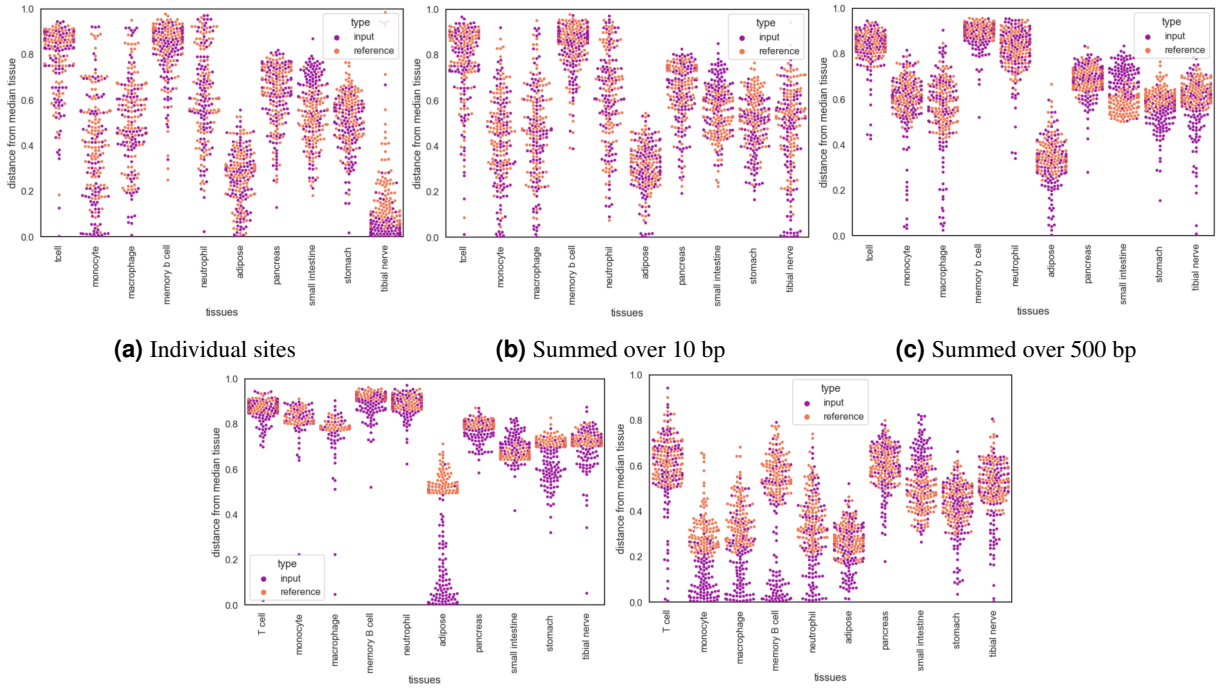
Hypomethylated sites are easier to differentiate than hypermethylated sites or than a mixture of both

Caggiano et al. originally determined the best markers for each cell type by comparing the distances between the methylation percentages of each individual cell type to the median methylation percentage of all cell types. This should, in theory, result in a mixture of hypo- and hypermethylated markers. A sufficiently large distance to the median is, however, not a very strict requirement as it does not remove the probability of having two or more cell types with a very similar methylation percentage (especially as the number of cell types in the reference grows). Moreover, in practice almost all of the markers found using this method are hypomethylated, so there is little benefit in also allowing for hypermethylated markers.

To make the markers more differential, we measured the distance between the methylation percentage of each cell type and the minimum methylation percentage of all other cell types. This approach was compared to the original approach (where the distance from the median is measured instead) as well as to a similar approach where we looked only for hypermethylated markers (and thus compared to the maximum of all other cell types). When comparing the markers' distances from the median, the original method seems to result in the best markers for all cell types except adipose (Figure D.1d). Hypomethylated markers, on the other hand, have a slightly smaller distance from the median for all cell types except for adipose, for which the distance is larger (Figure D.1e). Hypermethylated markers have overall the smallest distance from the median (Figure D.1c).

However, as reasoned above, the distance from the median may not be the best metric for defining the ability to differentiate between cell types. Therefore, we can not assume that the distance from the median also translates to the best cell type deconvolution results. For this reason, we looked at the results on a simulated mixture of the WGBS data of 10 cell types and measured the Pearson's correlation between the true and estimated cell type proportions of 50 replicate runs for 10 individuals. We set the true cell type proportions to a linearly incrementing array that sums to one. While the hypomethylated markers resulted in a correlation of $r^2 = 0.86 \pm 0.01$, the hypermethylated and original method resulted in a correlation of $r^2 = 0.68 \pm 0.04$ and $r^2 = 0.58 \pm 0.03$ respectively. This confirms the idea that the distance from the median is not the best metric for obtaining differentiable markers.

This can additionally be observed from the amount of markers found by each metric. The method for finding markers works in such way that it first finds the 100 best markers for each cell type and then removes the markers that are overlapping multiple cell types. As can be seen in Figure D.2, the original method finds less markers which means that the markers it finds have a high amount of overlap between cell types. Especially monocytes and macrophage cells seem to have much overlap, which makes sense given the fact that macrophage cells are differentiated monocyte cells [23]. Hypo- and hypermethylated markers are nevertheless able to differentiate these two cell types. To test whether the markers found using the original method would



(a) Individual sites **(b)** Summed over 10 bp **(c)** Summed over 500 bp

(d) Distance measured from median **(e)** Hypermethylated

Figure D.1. Distance from median methylation percentage for three different strategies; Purple dots represent the input at different marker locations and orange dots represent the reference at the same marker locations. The reference data was used to find the marker locations.
Row 1: Comparison between single CpG site markers which are summed with their 500 bp neighbouring sites (D.1a), 10 bp markers which are summed with their 490 bp neighbouring sites (D.1b) and 500 bp markers (D.1c).
Row 2: Comparison between markers defined by their distance from the median methylation percentage (D.1d), distance from the maximum (D.1e) and distance from the minimum (D.1c).
All figures in row one use hypomethylated markers, and all figures in row two are first summed over 500 bp.

result in better performance if more markers were included, we first tested for uniqueness of the 200 best markers of each cell type and then included the 100 best markers. This way each cell type had 100 markers. This resulted in a negligible increase in performance.

As the hypomethylated markers seem to give the best results, all experiments in this section, including the previous section, use hypomethylated markers.

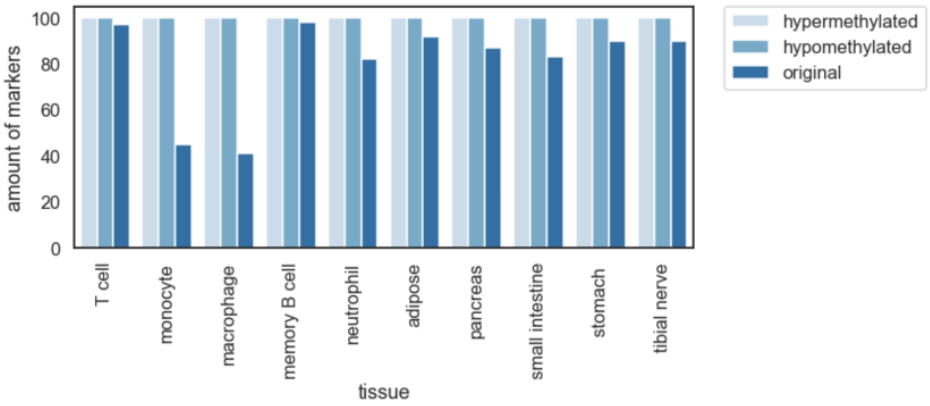


Figure D.2. The bar chart shows the amount of markers found for each cell type using each of the three different ways to measure the distance between cell types.

Additional improvements for increased differentiation between cell types

In addition to the improvements discussed in the previous two sections, there were two possible unwanted outcomes in the original method for finding markers. The first of which is that the authors introduced only a requirement for the median read depth of all cell types at a candidate marker site. This means that if one cell type is covered by one single read only at a candidate CpG site, this CpG site can still become a marker for that cell type as long as all other cell types have sufficient coverage. A simple adjustment was made to the method by setting a minimum depth threshold for cell types at their potential marker sites. This threshold was set equal to the median depth threshold.

The second possible undesirable behaviour is caused by the manner of checking for the uniqueness of the markers. As only the top 100 markers of all cell types is checked for overlapping markers, it is possible that the same site is the 100th best marker for cell type x and the 101st best marker for cell type y. This situation was prevented by keeping a list of the 150 best markers for each cell type which are all checked for uniqueness, such that the 100th best marker for cell type x could not even be the 150th best marker for cell type y.

The effects of both changes were measured by calculating the Pearson's correlation between the true and estimated cell type proportions for 10 individuals and 10 cell types of 50 replicate runs. The true cell type proportions were drawn from a uniform distribution and made to sum to one. Using no improvements, the correlation between the true and estimated cell types was $r^2 = 0.87 \pm 0.09$. Using only the additional uniqueness criterion did not change the results, and resulted in the same amount of correlation. The stricter depth criterion, however, improved the correlation to $r^2 = 0.91 \pm 0.06$. Combining both improvements resulted in the same correlation. This means that the situation described above does not occur, and the markers are already sufficiently unique. This is perhaps a consequence of using hypomethylated markers only.

REFERENCES

[1] Barefoot, M. E., Loyfer, N., Kiliti, A. J., McDeed IV, A. P., Kaplan, T., and Wellstein, A. (2021). Detection of cell types contributing to cancer from circulating, cell-free methylated dna. *Frontiers in genetics*, 12.

[2] Beck, D., Ben Maamar, M., and Skinner, M. K. (2022). Genome-wide cpg density and dna methylation analysis method (medip, rrbs, and wgbs) comparisons. *Epigenetics*, 17(5):518–530.

[3] Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

[4] Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B. L., Henderson, R., Lomen-Hoerth, C., Dahl, A., and Zaitlen, N. (2021). Comprehensive cell type decomposition of circulating cell-free dna with celfie. *Nature communications*, 12(1):1–13.

[5] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.

[6] Fernández, J. M., de la Torre, V., Richardson, D., Royo, R., Puiggròs, M., Moncunill, V., Fragkogianni, S., Clarke, L., Flicek, P., Rico, D., et al. (2016). The blueprint data analysis portal. *Cell systems*, 3(5):491–495.

[7] Greenberg, M. V. and Bourc'his, D. (2019). The diverse roles of dna methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, 20(10):590–607.

[8] Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., and Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna. *Nature genetics*, 49(4):635–642.

[9] Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., Grimes, B., Krysan, K., Yu, M., Wang, W., et al. (2017). Cancerlocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free dna. *Genome biology*, 18(1):1–12.

[10] Kitts, A. and Sherry, S. (2002). The single nucleotide polymorphism database (dbsnp) of nucleotide sequence variation. *The NCBI handbook. McEntyre J, Ostell J, eds. Bethesda, MD: US national center for biotechnology information*.

[11] Kurdyukov, S. and Bullock, M. (2016). Dna methylation analysis: choosing the right method. *Biology*, 5(1):3.

[12] Li, B., Pei, G., Yao, J., Ding, Q., Jia, P., and Zhao, Z. (2021a). Cell-type deconvolution analysis identifies cancer-associated myofibroblast component as a poor prognostic factor in multiple cancer types. *Oncogene*, 40(28):4686–4694.

[13] Li, J., Wei, L., Zhang, X., Zhang, W., Wang, H., Zhong, B., Xie, Z., Lv, H., and Wang, X. (2021b). Dismir: Deep learning-based noninvasive cancer detection by integrating dna sequence and methylation information of individual cell-free dna reads. *Briefings in bioinformatics*, 22(6):bbab250.

[14] Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong, W. H., et al. (2018). Cancerdetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free dna methylation sequencing data. *Nucleic acids research*, 46(15):e89–e89.

[15] Lo, Y. M. D., Han, D. S. C., Jiang, P., and Chiu, R. W. K. (2021). Epigenetics, fragmentomics, and topology of cell-free dna in liquid biopsies. *Science*, 372(6538).

[16] Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. (2022). A human dna methylation atlas reveals principles of cell type-specific methylation and identifies thousands of cell type-specific regulatory elements. *Biorxiv*.

[17] Miller, B. F., Pisanic II, T. R., Margolin, G., Petrykowska, H. M., Athamanolap, P., Goncearenco, A., Osei-Tutu, A., Annunziata, C. M., Wang, T.-H., and Elnitski, L. (2020). Leveraging locus-specific epigenetic heterogeneity to improve the performance of blood-based dna methylation biomarkers. *Clinical epigenetics*, 12(1):1–19.

[18] Moss, J., Magenheim, J., and Neiman, D. e. a. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease.

[19] Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome. *Genome research*, 20(7):883–889.

[20] Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., and Shendure, J. (2016). Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1-2):57–68.

[21] Sun, K., Jiang, P., Chan, K. A., Wong, J., Cheng, Y. K., Liang, R. H., Chan, W.-k., Ma, E. S., Chan, S. L., Cheng, S. H., et al. (2015). Plasma dna tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences*, 112(40):E5503–E5512.

[22] Tahir, R. A., Zheng, D., Nazir, A., and Qing, H. (2019). A review of computational algorithms for cpg islands detection. *Journal of biosciences*, 44(6):1–11.

[23] Yang, J., Zhang, L., Yu, C., Yang, X.-F., and Wang, H. (2014). Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. *Biomarker research*, 2(1):1–9.