Estimating the rate of cell type degeneration from epigenetic sequencing of cell-free DNA

Christa Caggiano[1], Barbara Celona[2], Fleur Garton[3], Joel Mefford[1], Brian Black[2], Naomi Wray[3], Catherine Lomen-Hoerth[2], Andrew Dahl[1], and Noah Zaitlen[1]

[1] University California Los Angeles, CA, USA
[2] University of California San Francisco, CA, USA
[3] University of Queensland, Brisbane, Australia

Abstract. Circulating cell-free DNA (cfDNA) in the bloodstream originates from dying tissues and is a promising non-invasive biomarker for cell death. The purpose of this work is to develop a method that can accurately estimate the relative abundances of cell types that contribute to cfDNA in the blood. To do this, we leverage the distinct DNA methylation profile of each tissue type throughout the body, and use this information to estimate the contribution of each of these cell types to the cfDNA mixture. Decomposing these mixtures, however, is difficult, as cfDNA of a relevant cell type may only be present in the blood in small amounts. We developed an EM algorithm, CelFiE, that estimates cell type proportion from both whole genome cfDNA input and reference data. Notably, CelFiE can handle missing and low count data, and does not rely on CpG site curation. CelFiE can also estimate an arbitrary number of 'unknown' cell type categories. We show in simulations that CelFiE can accurately estimate known and unknown cell type of origin of cfDNA mixtures in low coverage and noisy data. Simulations also demonstrate that we can effectively estimate cfDNA originating from 'rare' cell types composing less than 1% of the total cfDNA. To validate CelFiE's ability to correctly decompose real cfDNA mixtures, we use cfDNA extracted from pregnant and non-pregnant women. In the non-pregnant controls, CelFiE correctly estimates the absence of a placental component (p=1.01e-07). This indicates that CelFie can decompose real cfDNA mixtures consistent with true biology. CelFiE's ability to provide reliable estimates of cfDNA composition can be a tool for biomarker discovery, where the learned cell type degeneration rates can be used to quantitatively monitor disease.

Keywords: cfDNA · decomposition · epigenetics.

1 Introduction

Cells die at different rates as a function of disease state, age, environmental exposure, and behavior [1][2]. Knowing the rate at which cells die is a fundamental scientific question, with direct translational applicability. A quantifiable indication of cell death could streamline admission into clinical trials, facilitate disease diagnosis and prognosis, and offer a method for evaluating treatment effectiveness [3][4][5][6]. Circulating cell-free DNA (cfDNA) is a promising candidate for understanding cell type specific death. When a cell dies, DNA is released into the bloodstream in short fragments (approximately 160bp) [7]. All people are thought to have a low level of cfDNA in their blood at a given moment [8] [9]. In healthy individuals, cfDNA in the blood likely arises from normal cell turnover. In individuals with a disease, cfDNA can come from illness specific apotosis and necrosis [10]. As a result, cfDNA levels have been shown to be elevated in individuals with cancer, autoimmune diseases, transplantation responses, and trauma [11] [12][13][14]. Indeed, cfDNA has become the clinical standard for non-invasive prenatal testing [15], and many companies and research groups are sequencing cfDNA to identify the presence of somatic mutations related to tumors [16][17][18].

To understand what is driving changes in the cfDNA of people with disease, we can decompose the cfDNA mixture into the cell types from which the cfDNA is originating. This can give a noninvasive picture of cell death characterizing an individual's disease at a particular moment. While each cell type has the same DNA sequence, and does not give us information on where a cfDNA fragment is arising from, DNA methylation is cell type specific [19]. Subsequently, there is a rich literature of decomposition approaches using DNA methylation, often focusing on estimating the contribution of blood cell types to whole blood [20] [21][22][23]. More recent work has attempted to use methylated cfDNA to decompose cfDNA [24][25][26].

These approaches, however, fail to address some of the unique challenges of cfDNA. Previous work was designed for reference and input data from a methylation chip. Since cfDNA is only present in the blood in small amounts, methylation chips are not ideal for clinical use, because to get the required amount of input DNA, an onerous amount of blood must be extracted from a patient [27]. In this work, we turn to using whole genome bisulfite sequencing (WGBS) to assess the methylation of cfDNA. In contrast with methylation chips that target specific genomic locations, WGBS covers the entire genome, resulting in lower coverage and increased noise relative to chip data. Current methods are ill-equipped to handle stochasticity in either the reference or input.

Previous methods are also limited by which methylation sites (CpGs) are chosen. Methylation chips survey a limited number of CpGs, which may not be maximally informative of cell types. Some approaches also rely on selecting a set of CpGs designed for a particular dataset [24][26]. While curated site selection is useful for specific biological queries, it can lead to bias when generalizing across diseases. Choosing which sites to include in a decomposition can substantially influence which tissues are predicted (i.e., choosing sites only informative for pancreas will lead to an overestimated pancreas component). Another important limitation of previous cfDNA decomposition methods is that the results are restricted to the cell types included in the reference panel. The decomposition results using these methods will estimate the cfDNA mixture as being exactly composed of the tissues given in the input. There are many hundreds of cell types throughout the body and it would be impossible to incorporate all into a reference panel. Thus, the choice of reference tissues can lead to obvious bias in decomposition results of these methods.

In this work, we develop an efficient EM algorithm, CelFiE (Cell Free dna Estimation via expectation-maximization) for cfDNA decomposition that allows for low coverage and noisy data. Our method can also estimate unknown cell types not included in a reference panel and is not dependent on cherry-picking input methylation sites. We show in realistic, data-driven simulations that we can accurately estimate known and unknown cell types at low coverage and relatively few number of sites. We also can estimate cell types that contribute to only a small fraction of the total cfDNA. Decomposition of complex in-silico mixtures demonstrates that CelFiE is robust to many violations of our model assumptions. We design an algorithm for unbiased site selection that increases performance and decreases computation. Finally, we apply CelFiE to cfDNA extracted from pregnant women. Since placenta is only expected in pregnant women and not in non-pregnant controls, this data provides an inherent validation for our method. Overall, we demonstrate that CelFiE has the potential to decompose cfDNA in realistic conditions. This has broad translational utility for understanding the biology of cell death, and in applications such as quantitative biomarker discovery or in the non-invasive monitoring and diagnosis of disease.

2 Methods

The objective of this work is to decompose a cfDNA sample drawn from a patient into its cell types of origin. We assume that we are provided with a bisulfite sequenced reference data set of T cell types indexed by t, at M CpG sites indexed by m. Bisulfite sequencing produces read counts from specific cell types in two $T \times M$ matrices Y, D^Y with Y_{tm}, D^Y_{tm} the number of methylated and total reads at CpG m in reference cell types t. Together, these two matrices represent our reference data.

We are also provided with cfDNA extracted from N individual indexed by n. The bisulfite sequencing read counts of the cfDNA are given in two $N \times M$ matrices X, D^X with X_{nm}, D^X_{nm} the number of methylated and total reads at CpG m in the cfDNA from individual n. These two matrices represent our cfDNA data.

The goal of this work is to develop an algorithm that will take as input Y, D^Y, X_{nm}, D^X_{nm}, and output a matrix α, such that $\alpha[n,t]$ is the fraction of the cfDNA that originated in cell type t. We will also permit $t > T$, so that we can estimate the proportion of cell type in the cfDNA not provided in the reference panel.

3 Model

We model cfDNA of the individuals as a mixture of DNA from cell types in the reference panel and unknown cell type absent from the reference panel. We assume that the individuals are exchangeable.

We assume that reference data are drawn from a binomial distribution:

$$Y_{tm} | D^Y_{tm}, \beta_{tm} \sim Binomial(\beta_{tm}, D^Y_{tm})$$

where $\beta_{tm} \in [0,1]$ is the true, unknown proportion of DNA in a cell type that is methylated at position m. For now we assume that β_{tm} is shared across all individuals.

Now we model the samples in the cfDNA data. We assume each cfDNA molecule is drawn from some cell type, and in turn that its methylation value is drawn from a Bernoulli distribution governed by the cell type of origin:

$$x_{nmc}|\beta, Z_{nmc} \sim \text{Bernoulli}(\beta_{Z_{nmc}m})$$

where x_{nmc} is the c-th read from person n at position m, Z_{nmc} is the cell type of origin for this read, $\beta_{Z_{nmc}m}$ is the methylation proportion for the cell type of origin, Z_{nmc}.

We define the total number of reads, $X_{nm} := \sum_c x_{nmc}$, as the sum over all these reads from person n at site m. Note that c varies over 1 up to D^X_{nm}, and varies per person and methylation site–and, as a special case, if $D^X_{nm} = 0$ the summation is formally set to 0.

We assume that the cell type of origin of each cfDNA molecule is drawn independently from some individual-specific Multinomial distribution, governed by α_n, a length-T vector, where each entry is between 0 and 1 and all entries for each individual sum to 1:

$$Z_{nmc}|\alpha_n \overset{\text{ind}}{\sim} \text{Multinomial}(\alpha_{n1}, \ldots, \alpha_{nT})$$

where α_{nt} is the probability a random read from person n comes from cell type t. That is, the cell type of origin for each read is drawn from one of T bins, each with probability given by α_n.

3.1 EM algorithm

For simplicity, we first describe the algorithm for decomposing the cfDNA of a single individual using a complete reference panel with all relevant cell types. We then extend this approach to allow for multiple individuals as well as unknown cell types. Full details of both algorithms are given in the Appendix.

One Sample: Assume there is only one sample in the cfDNA data (i.e. $N = 1$). Let z_{tmc} indicate that read c for CpG m derives from cell type t. In relation to Z above, $z_{tmc} = 1$ if $Z_{1mc} = t$, and otherwise 0. That is, Z_{1mc} is a factor, and z_{tmc} is a binary indicator for which values Z_{1mc} takes.

The full data likelihood is:

$$P(x, z, Y|\alpha, \beta) = P(x|z, \beta)P(z|\alpha)P(Y|\beta)$$

with log probability we can show is equal to:

$$\sum_{t,m} \left(Y_{tm} \log \beta_{tm} + (D^Y_{tm} - Y_{tm}) \log(1 - \beta_{tm}) \right)$$

Calculating the Q function requires the conditional for z:

$$P(z_{tmc} = 1|x_{mc}, \beta, \alpha) \propto P(x_{mc}|z_{tmc} = 1, \beta)P(z_{tmc} = 1|\alpha) \propto \left(\beta_{tm}^{x_{mc}} (1 - \beta_{tm})^{1-x_{mc}} \right) \alpha_t \implies$$
$$P(z_{tmc} = 1|x_{mc}, \beta, \alpha) = \frac{\left(\beta_{tm}^{x_{mc}} (1 - \beta_{tm})^{1-x_{mc}} \right) \alpha_t}{\sum_k \left(\beta_{kt}^{x_{mc}} (1 - \beta_{kt})^{1-x_{mc}} \right) \alpha_k} =: p_{tmc}(\alpha, \beta)$$

This term depends on c only through x_{mc}, and in particular takes only one of two possible values:

$$\frac{\beta_{tm}\alpha_t}{\sum_k \beta_{kt}\alpha_k} =: p_{tm1}(\alpha, \beta) = p_{tmc}(\alpha, \beta) \quad \text{if } x_{mc} = 1$$
$$\frac{(1 - \beta_{tm})\alpha_t}{\sum_i (1 - \beta_{kt})\alpha_k} =: p_{tm0}(\alpha, \beta) = p_{tmc}(\alpha, \beta) \quad \text{if } x_{mc} = 0$$

E step: The Q function is defined at iteration i by:

$$Q_i(\beta, \alpha) := \mathbb{E}_{z|x,\alpha^{(i)},\beta^{(i)}} \left(\log P(x, z, y|\alpha, \beta)\right)$$

Let $p_{tm}^{(i)} := p_{tm1}(\alpha^{(i)}, \beta^{(i)})$, i.e. evaluated at current parameters for iteration i. Then we can show:

$$\begin{aligned} Q_i(\beta, \alpha) = \sum_{t,m} \left[\left(Y_{tm} + p_{tm1}^{(i)} x_m \right) \log\left(\beta_{tm}\right) + \left(D_{tm}^Y - Y_{tm} + p_{tm0}^{(i)} (D_m^X - x_m) \right) \log\left(1 - \beta_{tm}\right) \right] \\ + \sum_{t,m} \left(x_m p_{tm1}^{(i)} + (D_m^X - x_m) p_{tm0}^{(i)} \right) \log \alpha_t \end{aligned}$$

The first line in this equation captures expected total number of methylated reads (first term in the sum) and total number expected unmethylated reads (second term). The second line captures the deviation of the empirical cell type proportions from the prior expectation.

M step: For α, we must optimze under the constraint that α is a probability vector (i.e. sums to one with nonnegative entries) is:

$$\alpha_t = \frac{\sum_m \left(x_m p_{tm1}^{(i)} + (D_m^X - x_m) p_{tm0}^{(i)} \right)}{\sum_{k,m} \left(x_m p_{km1}^{(i)} + (D_m^X - x_m) p_{km0}^{(i)} \right)}$$

For β,

$$\begin{aligned} \nabla_{\beta_{tm}} Q_i(\beta, \alpha) &= \nabla_{\beta_{tm}} \left[\sum_{t,m} \left[\left(Y_{tm} + p_{tm1}^{(i)} x_m \right) \log\left(\beta_{tm}\right) + \left(D_{tm}^Y - Y_{tm} + p_{tm0}^{(i)} (D_m^X - x_m) \right) \log\left(1 - \beta_{tm}\right) \right] \right] \\ &= \frac{Y_{tm} + p_{tm1}^{(i)} x_m}{\beta_{tm}} - \frac{D_{tm}^Y - Y_{tm} + p_{tm0}^{(i)} (D_m^X - x_m)}{1 - \beta_{tm}} \end{aligned}$$

Setting this to zero to solve the first order condition gives:

$$\beta_{tm} = \frac{p_{tm1}^{(i)} x_m + Y_{tm}}{p_{tm0}^{(i)} (D_m^X - x_m) + D_{tm}^Y - Y_{tm} + p_{tm1}^{(i)} x_m + Y_{tm}}$$

EM algorithm with multiple cfDNA individuals We now go back to X rather than x. Each person has their own responsibility function, because each person has their own α:

$$P(z_{ntmc} = 1 | X_{nmc}, \beta, \alpha) = \frac{\left(\beta_{tm}^{X_{nmc}} (1 - \beta_{tm})^{1 - X_{nmc}} \right) \alpha_t}{\sum_k \left(\beta_{kt}^{X_{nmc}} (1 - \beta_{kt})^{1 - X_{nmc}} \right) \alpha_k} =: p_{ntmc}(\alpha_n, \beta)$$

As before, $p_{ntmc}(\alpha_n, \beta) := p_{ntm0}(\alpha_n, \beta)$ if $X_{nmc} = 0$ (and anologous for $X_{nmc} = 1$). In turn, as before, we set $p_{ntmc}^{(i)} = p_{ntmc}(\alpha_n^{(i)}, \beta^{(i)})$

The E-step becomes:

$$\begin{aligned} Q_i(\alpha, \beta) = \sum_{n,t,m} \left[\left(Y_{tm} + p_{ntm1}^{(i)} X_{nm} \right) \log\left(\beta_{tm}\right) + \left(D_{tm}^Y - Y_{tm} + p_{ntm0}^{(i)} (D_{nm}^X - X_{nm}) \right) \log\left(1 - \beta_{tm}\right) \right] \\ + \sum_{n,t,m} \left(X_{nm} p_{ntm1}^{(i)} + (D_{nm}^X - X_{nm}) p_{ntm0}^{(i)} \right) \log \alpha_{nt} \end{aligned}$$

The M-step for α is functionally identical, just applied to each individual in turn to estimate an α_n for each person n.

For β_{tm}, the M-step just amounts to aggregating reads over all people at position m that are expected to derive from cell type t:

$$\beta_{tm} = \frac{\sum_n p^{(i)}_{ntm1} X_{nm} + Y_{tm}}{\sum_n p^{(i)}_{ntm0}(D^X_{nm} - X_{nm}) + D^Y_{tm} - Y_{tm} + \sum_n p^{(i)}_{ntm1} X_{nm} + Y_{tm}}$$

Unknown sources: For each unknown cell type desired, add a new zero column to D^Y and Y. This produces an EM that is mathematically similar to the STRUCTURE model of mixtures of human populations (CITE). Note that if the number of unknown cell types is greater than the number of individuals, then the problem is not identified.

Regularization: In practice, we add a methylated and unmethylated pseudocount to every entry of X and Y/D^X and D^Y to account prevent issues driven by perfectly methylated or unmethylated sites.

4 Results

4.1 Evaluation using simulated cfDNA mixtures

We began by simulating cfDNA mixtures informed by realistic sequencing conditions: namely, low read count and noisy data. To analyze the performance of CelFiE under these conditions, we compared the results of CelFiE with a previous cfDNA deconvolution method, MethAtlas [25]. We simulated data that closely matched the input and reference data provided in MethAtlas. The true methylation proportion of 8,000 CpGs was drawn from a random uniform distribution, so that the methylation of one CpG was between 0% and 100%. Methylation proportions were drawn for 25 cell types. To characterize the performance of MethAtlas and CelFiE across both rare and abundant cell types, we simulated the true cell type proportion vector as $(1, \ldots, i)/\binom{i}{2}$, where i is the number of cell types truly in the mixture.

For CelFiE, the input data is number of methylated reads and read depth. The reference read depths were drawn from a random Poisson, centered at 10x, a relatively low sequencing depth. The number of methylated reads for a given CpG in each of the 25 cell types was drawn from a random binomial, where the probability of success was the true methylation value in that cell type, and the number of trials was the read depth at that locus. CfDNA input data was simulated one read at a time. Read depths were simulated from a random Poisson distribution centered at 10x, and then the reads for a CpG were assigned to originate from a cell type based on the cell type proportion vector for the cfDNA mixture. A read was determined to be either methylated or unmethylated given the true methylation proportion in that read's cell type of origin for that CpG. Since MethAtlas was designed for methylChip data, we calculated methylation proportion for a CpG by dividing the methylated reads by the depth at that locus.

CelFiE performs substantially better at low read depths (Fig 1). The correlation between the true and estimated proportions of MethAtlas is low, with a mean r^2 of 0.22 across cell types, in contrast with CelFiE's mean r^2 of 0.96. Furthermore, MethAtlas, is constrained to only estimating the proportion of cell type in the reference. This will create bias in the decomposition of MethAtlas that we seek to address with CelFiE.

After demonstrating CelFiE's performance relative to MethAtlas, we further characterized the properties of our approach. We simulate data as in figure 1, but with 100, 1000, and 10000 informative CpG sites. We fix one cell type at a percentage between 0% and 100% of the total mixture, allowing the other nine cell types to be random. As the number of sites increases, the capability of CelFiE to accurately decompose the cfDNA mixtures improves (Figure 2A).

Previous work suggests that a large portion of cfDNA originates from white blood cells [24]. This means that a potential cell type of clinical significance may only be present at a low proportion in the mixture. To assess the ability of CelFiE to estimate these rare cell types, we fixed one cell type and 0.01%, and simulated read depths at 10x, 100x, and 1000x coverage. Even at low read depths, CelFiE can decompose extremely low incidence cell types (Figure 2).

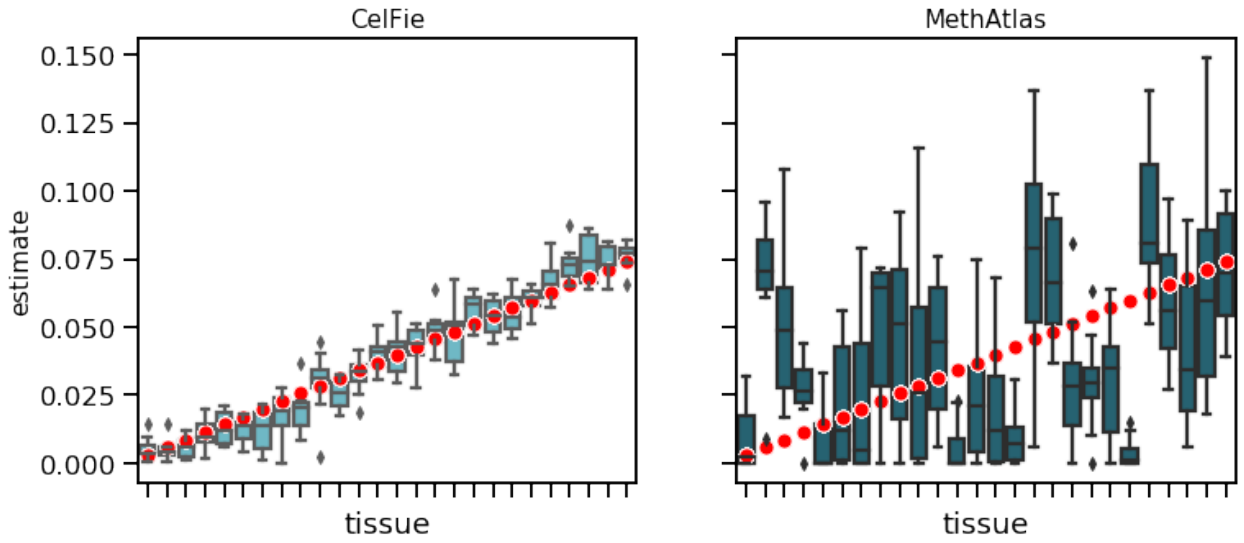

Fig. 1: Decomposition of simulated cfDNA mixtures by CelFiE (left) and MethAtlas (right). 50 simulations were performed for each algorithm and the estimates were plotted. The true cell type proportion is depicted in red.

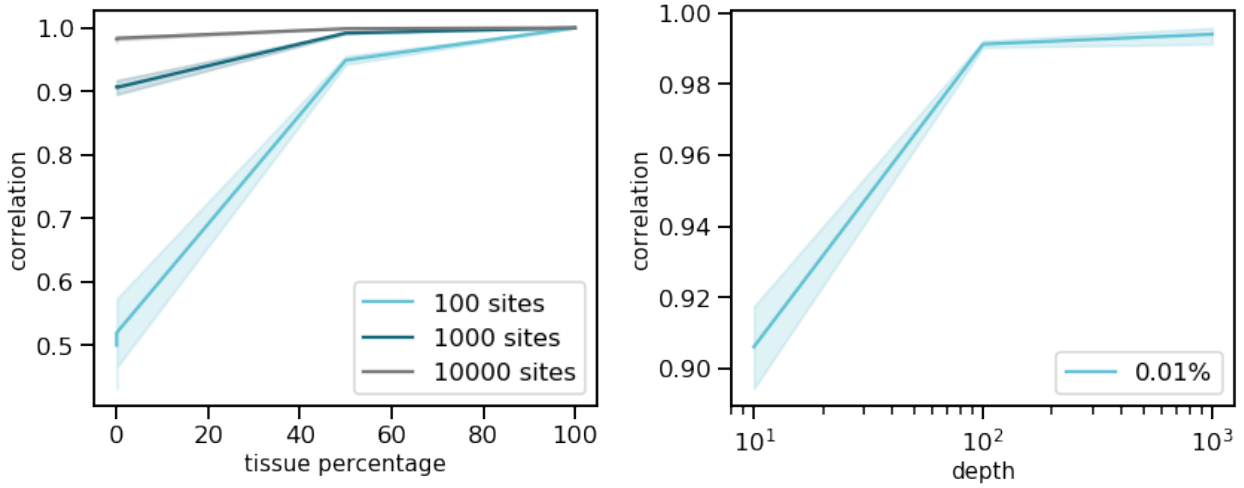

Fig. 2: Correlation between real and estimated cell type proportion. We fix a cell type at a proportion between 0% and 100%, and estimate the cell type proportion for 100, 1000, and 10000 sites (A). For a rare cell type of 0.01%, we plot the correlation between the true and estimated proportion of that cell type for mixtures with 10x, 100x, and 1000x read depths (B).

Next, we turned to understanding the behavior of CelFiE when estimating unknown cell types. We create reference and cfDNA reads for 1000 CpGs, at 10x depth. For the cfDNA mixtures, ten cell types were truly in the mixture. One cell type was excluded from the reference. We show that as the number of people included in the decomposition increase, the performance of CelFiE improves (3A). When two cell types are excluded from the reference, more people are needed to accurately estimate the unknown component (REDOING THIS PLOT USED WRONG VECTOR (3B).

4.2 Performance on in-silco cfDNA mixtures

After demonstrating the performance of CelFiE in data-driven simulations, we consider mixtures made from real WGBS data. We use biological replicates for ten WGBS datasets, downloaded from the ENCODE project [28][29]. These data encompass both tissue and cell types. Since tissues are heterogeneous mixtures of numerous cell types, decomposing tissues can contribute to error in our decomposition results.

Since nearly 80% of CpG sites in the human genome do not vary between cell types, we developed a method for unbiasedly choosing a set of informative CpGs [30]. We selected 7,237 tissue informative markers (TIMs) (section 7.3). Selecting TIMs improves performance a random subset of 7,000 CpGs. Furthermore, since CpG sites are locally coordinated in their methylation states [31], we summed all methylated and unmethylated reads for all CpGs +/-250bp of a TIM. Summing sites increased average read depth from 30x to 300x (INSERT REAL DEPTHS). Overall, this served to increases signal and accuracy (SUPP FIG?).

We then generated cfDNA mixtures for 100 individuals. For each individual, the true proportions vector was simulated so that cell types had increasing proportions. Lung and esophagus were excluded from the reference. CelFiE's ability to discriminate unknowns increases as the number of people in the experiment increases (Figure 4).

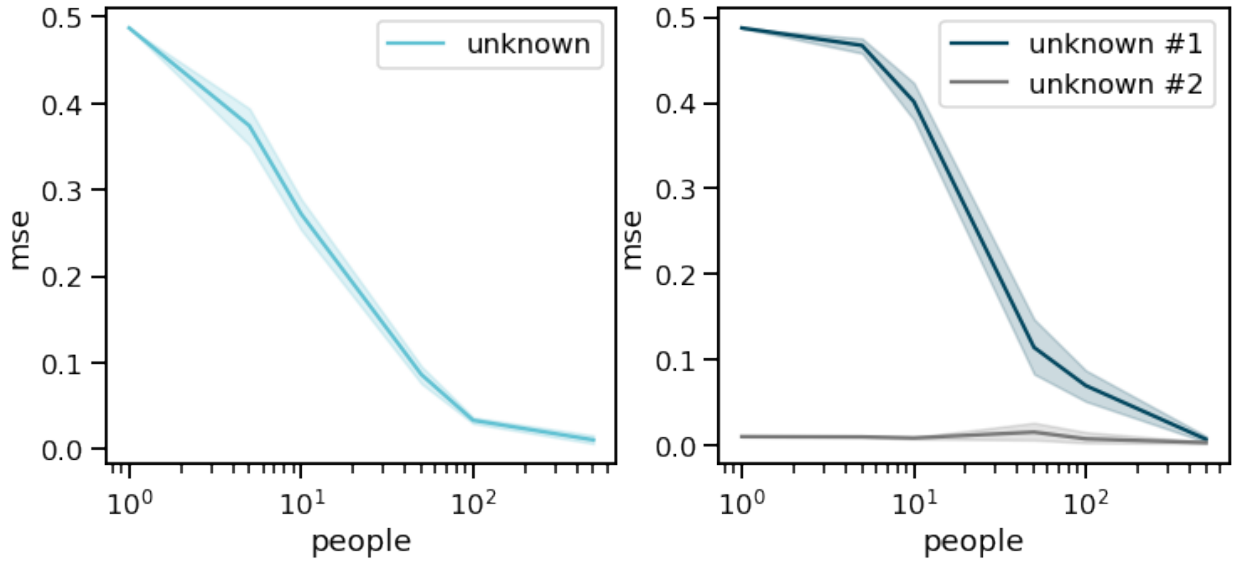

Fig. 3: Decomposition results for cfDNA mixtures with missing cell types in the reference. We simulate cfDNA for 1 to 500 people, and exclude one (A) or two (B) cell types truly in the cfDNA from the reference. We calculate the correlation between the estimated unknown component and the true unknown component for 50 simulation experiments.

4.3 Application to real data

Finally, we applied CelFiE to real cfDNA data. We choose pregnancy data as our real data application because it is one of the few examples of patient populations with a reliable true positive [32]. Unlike decomposition of

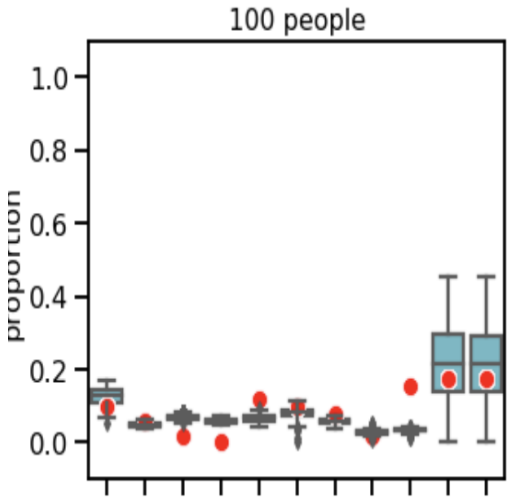

Fig. 4: PLACEHOLDER FIGURE Tissue proportion estimates for real WGBS cfDNA mixtures. For true random mixtures of cfDNA (red dots), we estimate the cell type composition (blue boxes) for reference and two unknowns.

cell types in blood, there is no FACS or similar existing gold standard. However, we know that non-pregnant women and men will not have placenta cfDNA in their bloodstream.

To do this, we download publicly available WGBS cfDNA of 7 pregnant and 8 non-pregnant women [33]. We subset the WGBS sites to the same TIMs we use in Section 4.2 and sum +/- 250 bp. Fourteen WGBS tissue samples from the ENCODE project were chosen for the reference panel, representing tissues throughout the body and blood. We estimate no placenta in non-pregnant women (Fig 5) (p=1.01e-07).

5 Discussion

During times of disease or increased cell turnover, increased levels of cfDNA can be detected in the blood. For example, increases in the amount of cfDNA has been detected in patients with multiple types of cancer, autoimmune diseases, as well as acute episodes of myocardial infarction, trauma, transplantation response and exercise (Swarup and Rajeswari, 2007, Velders et al., 2014, Schwarzenbach et al., 2011). Correspondingly, the utility of cfDNA as a biomarker diagnosis has been demonstrated in increasing number of settings, including prenatal testing (Jiang and Lo, 2016) and the detection tumor specific mutations (Li et al., 2016, Pentsova et al., 2016). Of greatest interest however, is that qualitative assessments of the cfDNA can now also provide information about cfDNA cellular origin (Snyder et al., 2016, Lehmann-Werman et al., 2016). This type of quantitative and qualitative assessment demonstrates an individualized, unbiased approach to understanding cellular turnover over time. However, these technologies are nascent, noisy, and expensive.

In this work we presented an algorithm, CelFiE, to decompose complex cfDNA mixtures into their cell types of origin. CelFiE can accurately decompose mixtures with low read count and missing data in both the reference cell types and the patient cfDNA samples. Furthermore, when large cohorts are available, it can estimate an arbitrary number of unknown cell types, which minimizes bias and increases confidence in the decomposition. Finally, because the algorithm is based on an EM, it is computationally efficient, and can scale.

The accuracy of CelFiE depends on a number of factors including the read depth, the number of sites considered in the reference panel, the informativeness of those sites, the concentration of the cell type of interest, and the number or reference sets and patient samples. Recent technologies for digesting or capturing specific regions of CelFiE [cite], may allow deeper sequencing at informative CpGs. Our algorithm for selecting such TIM CpGs demonstrated marked improvement in accuracy and could be use to select sites for capture.

There are number of areas for improvement going forward. Many of the reference cell types are complex mixtures of cell types and could be modeled as such. Our current ENCODE based results showed a high

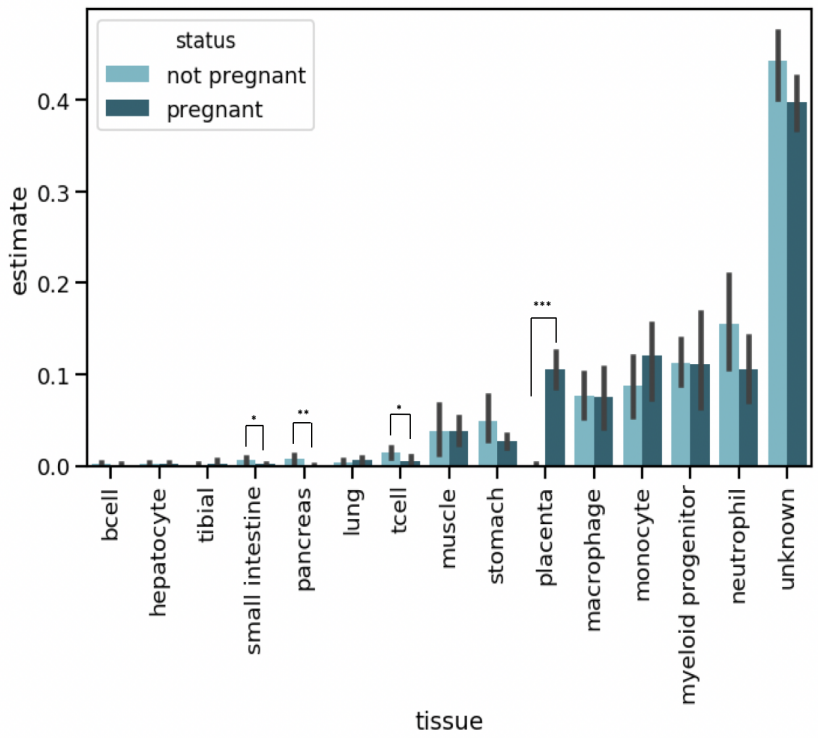

Fig. 5: CelFiE decomposition estimates for 8 non-pregnant (light blue) and 7 pregnant women (teal).

degree of correlation between replicates, but we believe modeling this heterogeneity will likely improve the results further. Even within cell types there may be heterogeneity of CpG methylation between individuals and disease states, and this could also be added to our model. We currently account for local correlation of CpG methylation by summing proximal CpG methylation states, but this may not always be optimal. Cell types are correlated in their methylation profiles and it could be interesting to consider a hierarchical output in which the composition can be considered at different levels of the cell type phylogeny. The addition of non CpG methylation and cfDNA fragment length may provide additional sources of information about cell types of origin.

While our initial application to pregnancy data demonstrated its accuracy, the intended future use is as a tool for biomarker discovery in disease contexts. Thus, both the the known and unknown cell type estimates can be tested for association with disease status, progression, and other quantitate and dichotomous phenotypes. The CelFiE algorithm is freely available on Github.

6 Code Availability

Method available for academic use at https://github.com/christacaggiano/celfie

References

1. Shigekazu Nagata. Apoptosis by Death Factor. *Cell*, 88(3):355–365, February 1997.
2. Pascal Meier, Andrew Finch, and Gerard Evan. Apoptosis in development. *Nature*, 407(6805):796–801, October 2000.
3. Diana Joka, Kristin Wahl, Sarah Moeller, Jerome Schlue, Bernhard Vaske, Matthias J. Bahr, Michael P. Manns, Klaus Schulze-Osthoff, and Heike Bantel. Prospective biopsy-controlled evaluation of cell death biomarkers for prediction of liver fibrosis and nonalcoholic steatohepatitis. *Hepatology*, 55(2):455–464, 2012.
4. Miquel Vila and Serge Przedborski. Targeting programmed cell death in neurodegenerative diseases. *Nature Reviews Neuroscience*, 4(5):365–375, May 2003.

5. Martin R. Turner, Robert Bowser, Lucie Bruijn, Luc Dupuis, Albert Ludolph, Michael Mcgrath, Giovanni Manfredi, Nicholas Maragakis, Robert G. Miller, Seth L. Pullman, Seward B. Rutkove, Pamela J. Shaw, Jeremy Shefner, and Kenneth H. Fischbeck. Mechanisms, models and biomarkers in amyotrophic lateral sclerosis. *Amyotrophic lateral sclerosis & frontotemporal degeneration*, 14(0 1):19–32, May 2013.
6. Robert Bowser, Martin R. Turner, and Jeremy Shefner. Biomarkers in amyotrophic lateral sclerosis: opportunities and limitations. *Nature Reviews Neurology*, 7(11):631–638, November 2011.
7. Maurice Stroun, Pierre Maurice, Valeri Vasioukhin, Jacqueline Lyautey, Christine Lederrey, François Lefort, Alain Rossier, Xu Qi Chen, and Philippe Anker. The Origin and Mechanism of Circulating DNA. *Annals of the New York Academy of Sciences*, 906:161–8, May 2000.
8. Danny Laurent, Fiona Semple, Philip J. Starkey Lewis, Elaine Rose, Holly A. Black, Stuart J. Forbes, Mark J. Arends, James W. Dear, and Timothy J. Aitman. Absolute measurement of the tissue origins of cell-free DNA in the healthy state and following paracetamol overdose. *bioRxiv*, page 715888, July 2019.
9. Anatoli Kustanovich, Ruth Schwartz, Tamar Peretz, and Albert Grinshpun. Life and death of circulating cell-free DNA. *Cancer Biology & Therapy*, 20(8):1057–1067, August 2019.
10. Sabine Jahr, Hannes Hentze, Sabine Englisch, Dieter Hardt, Frank O. Fackelmayer, Rolf-Dieter Hesch, and Rolf Knippers. DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells. *Cancer Research*, 61(4):1659–1665, February 2001.
11. Xiao Han, Junyun Wang, and Yingli Sun. Circulating Tumor DNA as Biomarkers for Cancer Detection. *Genomics, Proteomics & Bioinformatics*, 15(2):59–72, April 2017.
12. Suzan Tug, Susanne Helmig, Julia Menke, Daniela Zahn, Thomas Kubiak, Andreas Schwarting, and Perikles Simon. Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cellular Immunology*, 292(1):32–39, November 2014.
13. Iwijn De Vlaminck, Lance Martin, Michael Kertesz, Kapil Patel, Mark Kowarsky, Calvin Strehl, Garrett Cohen, Helen Luikart, Norma F. Neff, Jennifer Okamoto, Mark R. Nicolls, David Cornfield, David Weill, Hannah Valantine, Kiran K. Khush, and Stephen R. Quake. Noninvasive monitoring of infection and rejection after lung transplantation. *Proceedings of the National Academy of Sciences*, 112(43):13336–13341, October 2015.
14. Mikail Gögenur, Jakob Burcharth, and Ismail Gögenur. The role of total cell-free DNA in predicting outcomes among trauma patients in the intensive care unit: a systematic review. *Critical Care*, 21, January 2017.
15. Thomas J. Musci, Genevieve Fairbrother, Annette Batey, Jennifer Bruursema, Craig Struble, and Ken Song. Noninvasive prenatal testing with cell-free DNA: US physician attitudes toward implementation in clinical practice. *Prenatal Diagnosis*, 33(5):424–428, 2013.
16. Stanislav Volik, Miguel Alcaide, Ryan D. Morin, and Colin Collins. Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. *Molecular Cancer Research*, 14(10):898–908, October 2016.
17. Sinja Taavitsainen, Matti Annala, Elisa Ledet, Kevin Beja, Patrick J. Miller, Marcus Moses, Matti Nykter, Kim N. Chi, Oliver Sartor, and Alexander W. Wyatt. Evaluation of Commercial Circulating Tumor DNA Test in Metastatic Prostate Cancer. *JCO Precision Oncology*, (3):1–9, June 2019.
18. Cormac Sheridan. Investors keep the faith in cancer liquid biopsies. *Nature Biotechnology*, 37:972–974, August 2019.
19. Kaie Lokk, Vijayachitra Modhukur, Balaji Rajashekar, Kaspar Märtens, Reedik Mägi, Raivo Kolde, Marina Koltšina, Torbjörn K Nilsson, Jaak Vilo, Andres Salumets, and Neeme Tõnisson. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology*, 15(4):r54, 2014.
20. Eugene Andres Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, May 2012.
21. Eugene Andres Houseman, John Molitor, and Carmen J. Marsit. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, May 2014.
22. Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G. Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Sparse PCA Corrects for Cell-Type Heterogeneity in Epigenome-Wide Association Studies. *Nature methods*, 13(5):443–445, May 2016.
23. Elior Rahmani, Regev Schweiger, Liat Shenhav, Eleazar Eskin, and Eran Halperin. A Bayesian Framework for Estimating Cell Type Composition from DNA Methylation Without the Need for Methylation Reference. In S. Cenk Sahinalp, editor, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 207–223, Cham, 2017. Springer International Publishing.
24. Roni Lehmann-Werman, Daniel Neiman, Hai Zemmour, Joshua Moss, Judith Magenheim, Adi Vaknin-Dembinsky, Sten Rubertsson, Bengt Nellgård, Kaj Blennow, Henrik Zetterberg, Kirsty Spalding, Michael J. Haller, Clive H. Wasserfall, Desmond A. Schatz, Carla J. Greenbaum, Craig Dorrell, Markus Grompe, Aviad Zick, Ayala Hubert, Myriam Maoz, Volker Fendrich, Detlef K. Bartsch, Talia Golan, Shmuel A. Ben Sasson, Gideon Zamir, Aharon Razin, Howard Cedar, A. M. James Shapiro, Benjamin Glaser, Ruth Shemer, and Yuval

Dor. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proceedings of the National Academy of Sciences*, 113(13):E1826–E1834, March 2016.
25. Joshua Moss, Judith Magenheim, Daniel Neiman, Hai Zemmour, Netanel Loyfer, Amit Korach, Yaacov Samet, Myriam Maoz, Henrik Druid, Peter Arner, Keng-Yeh Fu, Endre Kiss, Kirsty L. Spalding, Giora Landesberg, Aviad Zick, Albert Grinshpun, A. M. James Shapiro, Markus Grompe, Avigail Dreazan Wittenberg, Benjamin Glaser, Ruth Shemer, Tommy Kaplan, and Yuval Dor. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications*, 9(1):1–12, November 2018.
26. Xiaomeng Liu, Jie Ren, Nan Luo, Huahu Guo, Yuxuan Zheng, Jingyi Li, Fuchou Tang, Lu Wen, and Jirun Peng. Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by methylated CpG tandem amplification and sequencing (MCTA-Seq). *Clinical Epigenetics*, 11(1):93, June 2019.
27. Ruth Pidsley, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208, October 2016.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.
29. Carrie A. Davis, Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, Kriti Jain, Ulugbek K. Baymuradov, Aditi K. Narayanan, Kathrina C. Onate, Keenan Graham, Stuart R. Miyasato, Timothy R. Dreszer, J. Seth Strattan, Otto Jolanki, Forrest Y. Tanaka, and J. Michael Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018.
30. Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Phil L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, August 2013.
31. Cecilia Lövkvist, Ian B. Dodd, Kim Sneppen, and Jan O. Haerter. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Research*, 44(11):5123–5132, June 2016.
32. E.S. Taglauer, L. Wilkins-Haug, and D.W. Bianchi. Review: Cell-free fetal DNA in the maternal circulation as an indication of placental health and disease. *Placenta*, 35(Suppl):S64–S68, February 2014.
33. Taylor J. Jensen, Sung K. Kim, Zhanyang Zhu, Christine Chin, Claudia Gebhard, Tim Lu, Cosmin Deciu, Dirk van den Boom, and Mathias Ehrich. Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biology*, 16:78, April 2015.
34. Roman Kosoy, Rami Nassir, Chao Tian, Phoebe A White, Lesley M. Butler, Gabriel Silva, Rick Kittles, Marta E. Alarcon-Riquelme, Peter K. Gregersen, John W. Belmont, Francisco M. De La Vega, and Michael F. Seldin. Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America. *Human mutation*, 30(1):69–78, January 2009.
35. Michael F. Seldin and Alkes L. Price. Application of Ancestry Informative Markers to Association Studies in European Americans. *PLOS Genetics*, 4(1):e5, January 2008.

7 Appendix

7.1 EM algorithm

For simplicity, we first describe the algorithm for decomposing the cfDNA of a single individual using a complete reference panel with all relevant cell types. We then extend this approach to allow for multiple individuals as well as cell types.

One Sample: Let t index T cell types in our reference panel, n index the N individuals in the cfDNA data set, and m index the M CpG sites considered.

Assume there is only one sample in the cfDNA dataset (i.e. $N = 1$). Let z_{tmc} indicate that read c for CpG m derives from cell type t. In relation to Z above, $z_{tmc} = 1$ if $Z_{1mc} = t$, and otherwise 0. I.e. Z_{1mc} is a factor, and z_{tmc} is a binary indicator for which values Z_{1mc} takes.

The full data likelihood is:

$$P(x, z, Y|\alpha, \beta) = P(x|z, \beta)P(z|\alpha)P(Y|\beta)$$

The complicated term is the first one:

$$\begin{aligned}\log P(x|z,\beta) &= \sum_{t,m,c} \log P(x_{mc}|z_{tmc},\beta_{tm}) \\ &\equiv \sum_{t,m,c} \log\left[(\beta_{tm})^{z_{tmc}\cdot x_{mc}}\,(1-\beta_{tm})^{z_{tmc}\cdot(1-x_{mc})}\right] \\ &\equiv \sum_{t,m,c} z_{tmc}\left[x_{mc}\log(\beta_{tm}) + (1-x_{mc})\log(1-\beta_{tm})\right]\end{aligned}$$

The second ingredient in the likelihood is:

$$\log P(z|\alpha) = \sum_{t,m,c} \log P(z_{tmc}|\alpha) = \sum_{t,m,c} \log(\alpha_t^{z_{tmc}}) = \sum_{t,m,c} z_{tmc}\log\alpha_t$$

Finally,

$$\log P(Y|\beta) = \sum_{t,m}\left(Y_{tm}\log\beta_{tm} + (D^Y_{tm} - Y_{tm})\log(1-\beta_{tm})\right)$$

Calculating the Q function requires the conditional for z:

$$\begin{aligned}&P(z_{tmc}=1|x_{mc},\beta,\alpha) \propto P(x_{mc}|z_{tmc}=1,\beta)P(z_{tmc}=1|\alpha) \propto \left(\beta_{tm}^{x_{mc}}(1-\beta_{tm})^{1-x_{mc}}\right)\alpha_t \implies \\ &P(z_{tmc}=1|x_{mc},\beta,\alpha) = \frac{\left(\beta_{tm}^{x_{mc}}(1-\beta_{tm})^{1-x_{mc}}\right)\alpha_t}{\sum_k\left(\beta_{kt}^{x_{mc}}(1-\beta_{kt})^{1-x_{mc}}\right)\alpha_k} =: p_{tmc}(\alpha,\beta)\end{aligned}$$

This term depends on c only through x_{mc}, and in particular takes only one of two possible values:

$$\begin{aligned}\frac{\beta_{tm}\alpha_t}{\sum_k \beta_{kt}\alpha_k} &=: p_{tm1}(\alpha,\beta) = p_{tmc}(\alpha,\beta) \quad \text{if } x_{mc}=1 \\ \frac{(1-\beta_{tm})\alpha_t}{\sum_i (1-\beta_{kt})\alpha_k} &=: p_{tm0}(\alpha,\beta) = p_{tmc}(\alpha,\beta) \quad \text{if } x_{mc}=0\end{aligned}$$

E step: The Q function is defined at iteration i by:

$$Q_i(\beta,\alpha) := \mathbb{E}_{z|x,\alpha^{(i)},\beta^{(i)}}\left(\log P(x,z,y|\alpha,\beta)\right)$$

To evaluate this, we break it into three parts. Let $p^{(i)}_{tm} := p_{tm1}(\alpha^{(i)},\beta^{(i)})$, i.e. evaluated at current parameters for iteration i. Then:

$$\begin{aligned}\mathbb{E}_{z|x,\alpha^{(i)},\beta^{(i)}}\left(\log P(x|z,\alpha,\beta)\right) &\equiv \sum_{t,m,c} \mathbb{E}_{z|x,\alpha^{(i)},\beta^{(i)}}\left(z_{tmc}\right)\left[x_{mc}\log(\beta_{tm}) + (1-x_{mc})\log(1-\beta_{tm})\right] \\ &\equiv \sum_{t,m,c} p^{(i)}_{tmc}\left[x_{mc}\log(\beta_{tm}) + (1-x_{mc})\log(1-\beta_{tm})\right] \\ &\equiv \sum_{t,m,c}\left[p^{(i)}_{tm1}x_{mc}\log(\beta_{tm}) + p^{(i)}_{tm0}(1-x_{mc})\log(1-\beta_{tm})\right] \\ &\equiv \sum_{t,m}\left[p^{(i)}_{tm1}x_{m}\log(\beta_{tm}) + p^{(i)}_{tm0}(D^X_m - x_m)\log(1-\beta_{tm})\right]\end{aligned}$$

Then, the second part is:

$$\begin{aligned}\mathbb{E}_{z|x,\alpha^{(i)},\beta^{(i)}}\left(\log P(z|\alpha)\right) &\equiv \sum_{t,m,c} p^{(i)}_{tmc}\log\alpha_t \\ &\equiv \sum_{t,m}\left(x_m p^{(i)}_{tm1} + (D^X_m - x_m)p^{(i)}_{tm0}\right)\log\alpha_t\end{aligned}$$

The third part is $P(Y|\beta)$.

Finally, adding the three parts together:

$$\begin{aligned}
Q_i(\beta,\alpha) &= \sum_{t,m}\left[p^{(i)}_{tm1}x_m\log(\beta_{tm}) + p^{(i)}_{tm0}(D^X_m - x_m)\log(1-\beta_{tm})\right] \\
&\quad + \sum_{t,m}\left(x_m p^{(i)}_{tm1} + (D^X_m - x_m)p^{(i)}_{tm0}\right)\log\alpha_t \\
&\quad + \sum_{t,m}\left(Y_{tm}\log\beta_{tm} + (D^Y_{tm} - Y_{tm})\log(1-\beta_{tm})\right) \\
&= \sum_{t,m}\left[\left(Y_{tm} + p^{(i)}_{tm1}x_m\right)\log(\beta_{tm}) + \left(D^Y_{tm} - Y_{tm} + p^{(i)}_{tm0}(D^X_m - x_m)\right)\log(1-\beta_{tm})\right] \\
&\quad + \sum_{t,m}\left(x_m p^{(i)}_{tm1} + (D^X_m - x_m)p^{(i)}_{tm0}\right)\log\alpha_t
\end{aligned}$$

The first line in the final equation captures expected total number of methylated reads (first term in the sum) and total number expected unmethylated reads (second term). The second line captures the deviation of the empirical cell type proportions from the prior expectation.

M step: For α, the optimizer under the constraint that α is a probability vector (i.e. sums to one with nonnegative entries) is:

$$\alpha_t = \frac{\sum_m \left(x_m p^{(i)}_{tm1} + (D^X_m - x_m)p^{(i)}_{tm0}\right)}{\sum_{k,m}\left(x_m p^{(i)}_{km1} + (D^X_m - x_m)p^{(i)}_{km0}\right)}$$

For β,

$$\begin{aligned}
\nabla_{\beta_{tm}} Q_i(\beta,\alpha) &= \nabla_{\beta_{tm}}\left[\sum_{t,m}\left[\left(Y_{tm} + p^{(i)}_{tm1}x_m\right)\log(\beta_{tm}) + \left(D^Y_{tm} - Y_{tm} + p^{(i)}_{tm0}(D^X_m - x_m)\right)\log(1-\beta_{tm})\right]\right] \\
&= \frac{Y_{tm} + p^{(i)}_{tm1}x_m}{\beta_{tm}} - \frac{D^Y_{tm} - Y_{tm} + p^{(i)}_{tm0}(D^X_m - x_m)}{1-\beta_{tm}}
\end{aligned}$$

Setting this to zero to solve the first order condition gives:

$$\beta_{tm} = \frac{p^{(i)}_{tm1}x_m + Y_{tm}}{p^{(i)}_{tm0}(D^X_m - x_m) + D^Y_{tm} - Y_{tm} + p^{(i)}_{tm1}x_m + Y_{tm}}$$

EM algorithm with multiple cfDNA individuals We now go back to X rather than x. Each person has their own responsibility function, because each person has their own α:

$$P(z_{ntmc} = 1|X_{nmc},\beta,\alpha) = \frac{\left(\beta_{tm}^{X_{nmc}}(1-\beta_{tm})^{1-X_{nmc}}\right)\alpha_t}{\sum_k\left(\beta_{kt}^{X_{nmc}}(1-\beta_{kt})^{1-X_{nmc}}\right)\alpha_k} =: p_{ntmc}(\alpha_n,\beta)$$

As before, $p_{ntmc}(\alpha_n,\beta) := p_{ntm0}(\alpha_n,\beta)$ if $X_{nmc} = 0$ (and anologous for $X_{nmc} = 1$). In turn, as before, we set $p^{(i)}_{ntmc} = p_{ntmc}(\alpha^{(i)}_n,\beta^{(i)})$

The E-step becomes:

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

The M-step for α is functionally identical, just applied to each individual in turn to estimate an α_n for each person n.

For β_{tm}, the M-step just amounts to aggregating reads over all people at position m that are expected to derive from cell type t:

$$\beta_{tm} = \frac{\sum_n p^{(i)}_{ntm1} X_{nm} + Y_{tm}}{\sum_n p^{(i)}_{ntm0}(D^X_{nm} - X_{nm}) + D^Y_{tm} - Y_{tm} + \sum_n p^{(i)}_{ntm1} X_{nm} + Y_{tm}}$$

7.2 WGBS Data Processing

Adult tissue and cell type WGBS/RRBS bedMethyl files were obtained from the ENCODE Project (). All bed file coordinates were harmonized to hg38 using hgLiftOver (). For each tissue or cell type, the file was restructured to report the number of methylated reads and read depth for each CpG locus. To remove sites that were non-informative of tissue of origin (i.e. the same methylation state across all tissues), CpG sites with a variance in percent methylation across tissues of less than 0.005 were removed.

7.3 Site Selection and Summing

Tissue informative markers: Only 21.8% of autosomal CpGs vary by cell type [30]. Selecting sites that do vary can maximize information on tissue of origin and reduce computation burden. We propose unbiasedly selecting tissue informative markers (TIMs), an approach inspired by ancestry informative markers in population genetics [34] [35]. For each reference tissue, for each CpG, the distance between the percent methylation of that cell type and the median percent methylation for that CpG was calculated. Only CpGs where the median depth was greater than 15 and had no missing data were considered. The top N CpGs with the greatest distance per cell type were selected. TIMs provide increased accuracy in decomposition, and vastly improve computation time.

Site combination: DNA methylation of nearby CpGs is coordinated [31]. WThus, we can combine related CpG sites to increase read depth and overcome noise. After selecting TIMs, we combine reads for all CpGs 250bp upstream and downstream of the TIM.