Department of Computer Science

BSc Computer Science and Statistics with Emphasis on Data Science - Final Project

Academic Year 2020 - 2021

A Bayesian approach to reveal the origins of cell-free DNA by human cell-type methylation atlas

Guy Kornblit, Shir Molina

Advisors: Prof. Tommy Kaplan, Netanel Loyfer

A report submitted in partial fulfilment of
the requirements for the degree of Bachelor of Science

The Hebrew University of Jerusalem
Department of Computer Science

Abstract

Methylation patterns of circulating cell-free DNA (cfDNA) contain rich information about recent cell death events in the body. Here, we present an approach for determination of the tissue origins of cfDNA, using a reference methylation atlas of human tissues and cell types. This method is based on features and statistics that have been researched in Prof. Tommy Kaplan's lab in the Hebrew University of Jerusalem, and evaluated using similar methods to previous studies of models of that kind. We show that a mixture model can identify cellular contribution to the blood stream, but presents strong bias and insufficient resolution. Hence, we show that our model is not better than a baseline model suggested by the lab.

Contents

Chapter 1

Introduction

Small fragments of DNA circulate freely in the peripheral blood of healthy and diseased individuals. **These cell-free DNA (cfDNA) molecules are thought to originate from dying cells and thus reflect ongoing cell death taking place in the body**. This understanding has led to the emergence of diagnostic tools, which are impacting multiple areas of medicine. A key limitation of theses technologies is that sequencing does not reveal the tissue origins of cfDNA, precluding the identification of tissue-specific cell death, which is critical in many types of diseases. For example, in oncology, it is often important to determine the tissue origin of the tumor. Identification of the tissue origins of cfDNA may also provide insights into collateral tissue damage (e.g., toxicity of drugs in genetically normal tissues), a key element in drug development and monitoring of treatment response.

Prof. Tommy Kaplan's lab focuses on tracing the tissue sources of plasma cfDNA, based on tissue-specific epigenetic signatures. Namely, DNA methylation patterns. **Methylation** of cytosine adjacent to guanine (CpG sites) is an essential component of cell type-specific gene regulation, and hence **is a fundamental mark of cell identity** (Fig. 1.1). The lab has demonstrated [4] the potential of a robust and accurate deconvolution of plasma methylation from as little as 20 ml of blood, as a diagnostic modality for early detection and monitoring of disease.

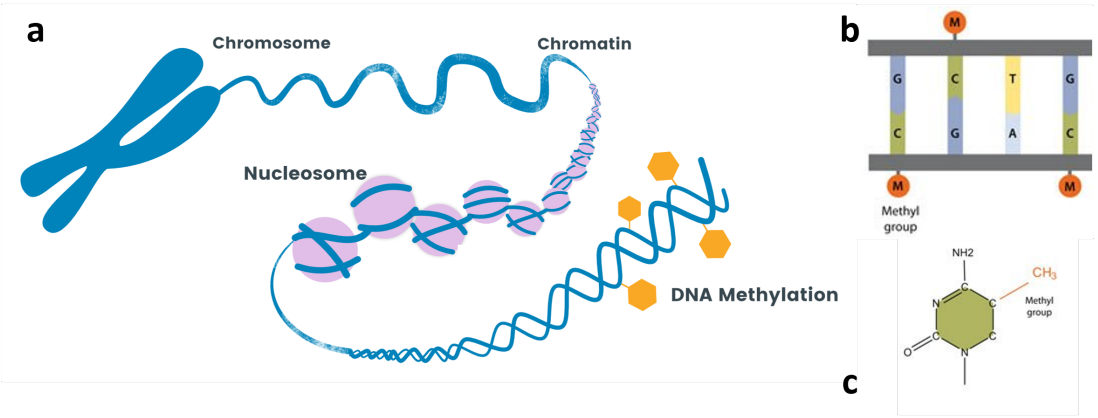

Figure 1.1: DNA methylation. **a** DNA molecule location in a cell nucleus. **b**. A methyl group addition to the cytosine-phosphate-guanine (CpG). **c** methyl group molecule.

1.1 Aims and Objectives

We aim to study potential contribution of a novel deconvolution method to decipher the origin of DNA in blood circulation. Namely:

1. Develop a generative model that describe the creation of cfDNA samples, that will be able to identify abnormalities from healthy cfDNA.
2. Design evaluation methods to compare with baseline model and to test different characteristics of our model.
3. Design and implement a framework to experiment the model under various configurations.

1.2 Project Approach

We developed a mixture model that uses a statistic for methylation data, with different data considered for each observation than before. We evaluated our model using in-silico simulations, to model ability estimating for each cell-type and for all cell types and tissues simultaneously. In addition, we tested aspects of the model performance on real and unsupervised data.

1.3 Dissertation Outline

Chapter 2 discusses previous work that our project directly relies on, and set the settings for our approach for the problem. Chapter 3 presents the data used in this work. Chapter 4 explains how the project will be undertaken. Chapter 5 presents details of our suggested solution and evaluation methods, and Chapters 6 and 7 discuss the results and summarize.

Chapter 2

Related work

Deconvolution of cell-free DNA based on methylation patterns is an ongoing field of research. As such, different approaches are being examined to leverage knowledge about methylation patterns and effect. Specifically, we rely on some of the lab's key contributions.

Development of a DNA methylation atlas. The lab published [4] a comprehensive DNA methylation database of key human cell types. The atlas contains the **methylomes of healthy human tissues , which are expected to be universally conserved** (that is, be nearly identical among cells of the same type, among individuals, throughout life, and be largely retained even in pathologies). The methylomes for every cell type and tissue in the atlas were collected from isolated surgical samples. For better granularity of methylation signatures, data from specific cell types was preferred rather than whole tissue methylome since those are a composite of multiple heterogeneous cell types, that are not uniquely linked to the target tissue (for example, different types of epithelial cells or white blood cells). In addition, highly similar methylomes from the same tissue were merged.

Healthy baseline for plasma cfDNA. Previous deconvolution approach was demonstrated to accurately identify cell type-specific cfDNA in healthy and pathological conditions. Based on different distribution across CpG sites, the algorithm used was Non Negative Least Squares regression (NNLS). The model estimation, based on plasma cfDNA from healthy donors ($n = 23$), is that most of the cfDNA originates from white blood cells (50%), vascular endothelial cells (7%) and platelets (30%) (Fig.2.1). The cfDNA signal from these tissues reflects the sum of multiple parameters: total cell number in these organs, the degree of baseline turnover, and the fact that cfDNA from these tissues is apparently cleared via blood. The absence of a cfDNA signal from other tissues in the body, known to have a high turnover rate, likely reflects alternative routes to clear cfDNA rather than to blood. For example, in lung, kidney and skin.

Discriminative Markers. For each cell type or tissue in atlas, previous work discovered windows of adjacent CpG sites that showed unique methylation pattern when compared to other cell types and tissues. Choosing only a subset of the methylome for deconvolution was made to increase sensitivity and affordability of the routine. In fact, **deconvolution based on a defined subset of informative sites performed better than an approach taking into account all sites**, including those that are not differentially methylated between tissues and hence contributed mostly noise.

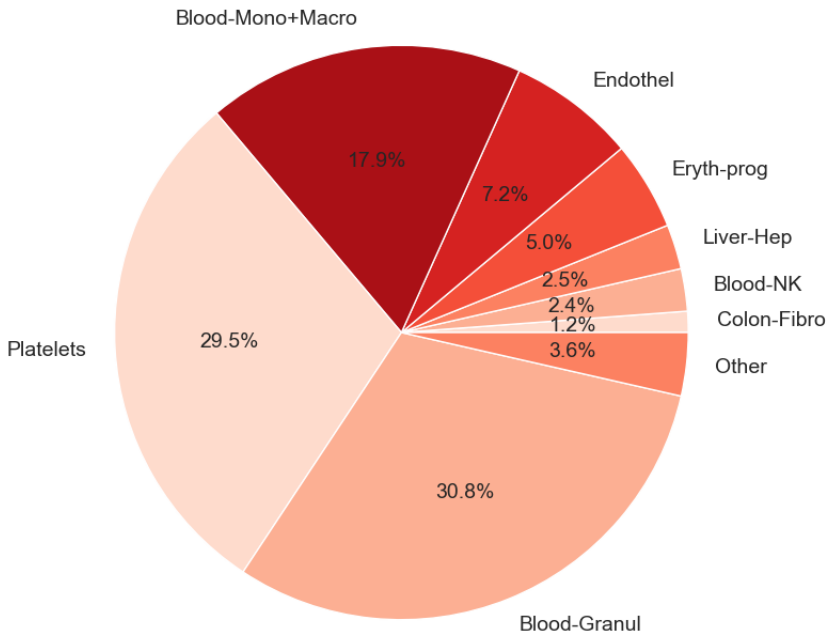

Figure 2.1: Baseline estimate for healthy patients from the NNLS algorithm described in [4]. Based on plasma cfDNA sample collected from 23 healthy individuals.

Chapter 3

Data

3.1 Data description

A single Observation, also referred as a Read, is a fragment of DNA that spans some genomic loci and contains various methylation (CpG) sites. Each site contains marking if it is methylated or not, where other nucleotide sequences that are not CpG sites are omitted. In addition, every read contains information about its genomic index (3B), and its methylome index (30M). **Methylation profile** is basically a collection of reads.

In this work, methylation profiles include reads obtained by the following procedures:

1. **Labeled Tissues Data**: sampled from isolated cell-types from tissues of interest. Therefore, multiple observations per cell-type can represent the latter methylation patterns. Since DNA is extracted directly from undamaged cells, raw pool of labeled reads covers the entire genome loci.

2. **Plasma cfDNA**: originated from various cells of unknown origin, hence covers random and partial loci. Thus, different reads that covers the same CpG sites can represent contradicting methylation patterns.

3. **Whole-blood DNA**: multiple types of white blood cell (leukocytes) are collected and sequenced to obtain methylation profile of DNA. These profiles are used to reflect plasma cfDNA healthy profile, since they include heterogeneous mixture of cell types that assumed to contribute the majority of cfDNA to plasma.

3.2 Data set

- **Baseline for tissue contribution to plasma cfDNA**, based on preceding NNLS regression model on updated atlas data. This baseline estimation contains 39 different cell types (see chapter 2).

- **Reference Atlas**: methylation profiles for every labeled tissue and cell-type. For every tissue, we sample 70% of its observations to be considered as a **train atlas**, and the other 30% of observations to be used as the **test atlas**.

- **Markers (or window)**: A set of about 25 markers per cell-type of methylome indices that the latter showed different methylation pattern against all other cell-types in the atlas (see chapter 2).
- **Whole-blood samples**: white blood cells methylation profiles extracted from 23 healthy individuals.
- **plasma cfDNA samples**: real plasma samples from 23 healthy individuals, 10 samples from liver cancer patients, and 12 samples from Covid-19 patients.

Preprocessing

Constructing methylation profile (Fig. 3.1). We define a statistic to express methylation patterns as a distribution over the markers. Given a collection of reads $\mathcal{R}$, and a set of markers:

1. link each read to a marker, if the read contains at least 2 CpG sites that overlap with the marker indices. Reads without overlapping sites are omitted.
2. for each read, count the number of methylated and unmethylated CpG sites, and label the entire read by a threshold β:

$$\begin{cases} M, & \frac{\text{methylated sites}}{\text{all sites}} \geq \beta \\ U, & \frac{\text{methylated sites}}{\text{all sites}} \leq 1-\beta \\ X, & \text{otherwise} \end{cases}$$

3. count the number of reads per marker and label (U,M,X), denoted n_i^U, n_i^M, n_i^X for every window i, s.t $n(i) = \sum_l n_i^l$.
4. Estimate marker multinomial distribution, by adding pseudo counts α and normalizing. Namely, for marker i and label l, calculate $\hat{p}_i(l) = \frac{n_i^l + \alpha}{n(i)+3\alpha}$.

Remarks:

1. we set $\beta = 0.75$, and $\alpha = 1$.
2. stages 1-3 are being preformed in the lab, where the last stages are preformed within our code.

Besides adding pseudo counts to the methylation profiles, we also add pseudo counts to the baseline estimate for tissue contribution, denoted as $\gamma = 0.01$. This stage has been added to avoid zero probabilities later in the model.

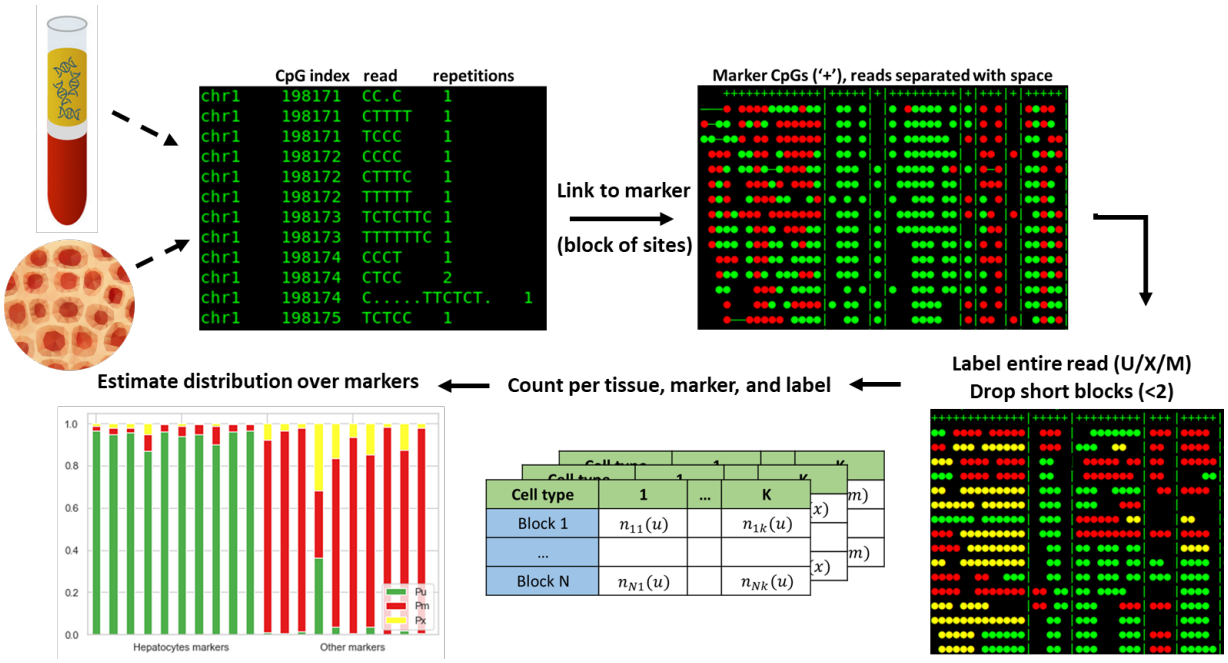

Figure 3.1: Reads acquired from plasma cfDNA or an isolated group of a certain cell type, are attached to discriminative blocks of sites, labeled by methylation average within the read sites, and aggregated to form multinomial distribution over the 3 labels (U/X/M).

3.3 Preliminary analysis

About 25 markers were assigned per cell-type, that covers roughly 5-10 CpG sites (within a larger set of base pairs). For all cell types, most of the markers showed low methylation rate against other cell types and tissues. Namely, most of the CpG sites in these windows were not methylated for samples of some cell type, but were hyper methylated for all other cell types (Fig. 3.2). Due to the methylation mechanism role in gene regulation, its expected that low methylation windows actually allows the expression of a specific gene for a given cell type. Some of the markers found can separate well a single cell-type of tissue from all others, but in some cases the markers found can differentiate between 2-3 cell types and all others. These marker shows similar methylation pattern among their group. Consequently, these cell type are linked to more markers, and therefore are exposed to more noise.

Observations in the atlas are collected in different manners, hence the amount of data we hold per cell-type and tissue varies (Fig. 3.3). This data later use to estimate distribution for every cell-type, thus can affect accuracy. The bias seems towards cell types who are the main contributors of cfDNA to the plasma. Comparing amounts of observations for cell type-specific markers against all others, there is no significant difference in distribution (Fig. 3.4).

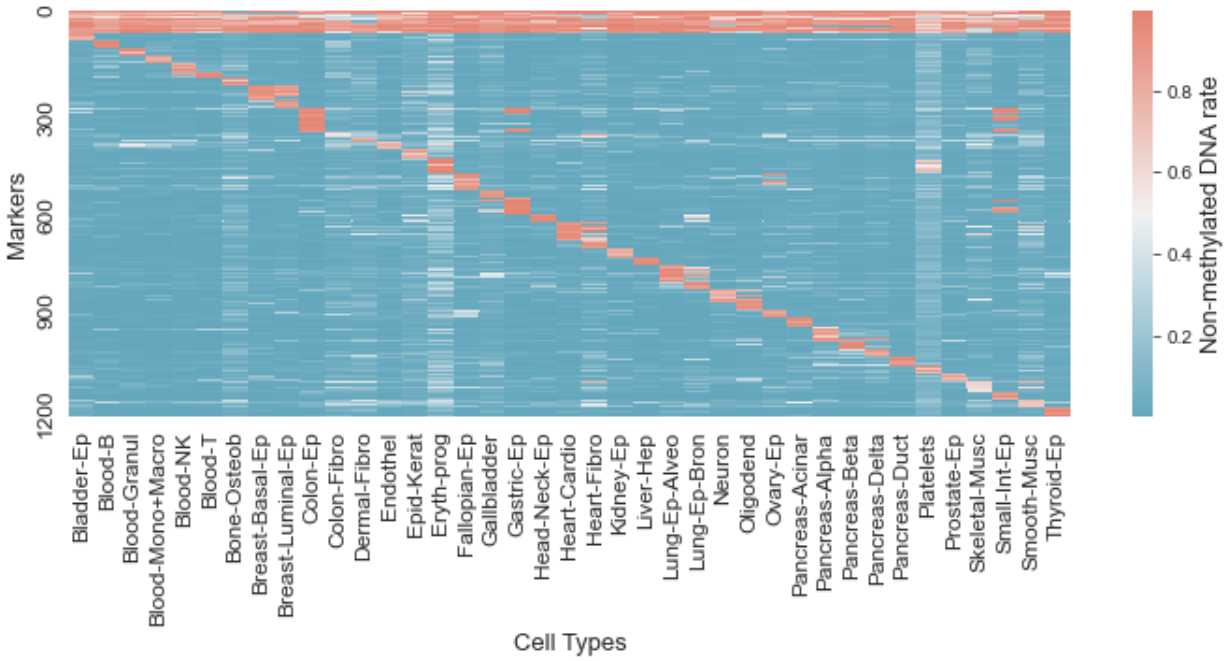

Figure 3.2: Unmethylated DNA rate by markers. Every DNA observation labeled as unmethylated if 75% of its CpG sites were unmethylated, averaging unmethylated observations per window gives the Unmethylated rate that shown on the map. High values (in red) indicate low methylation rate. The upper region of high values correspond to markers that differentiate between cell types by being highly methylated, and show low variance of unmethylated rate.

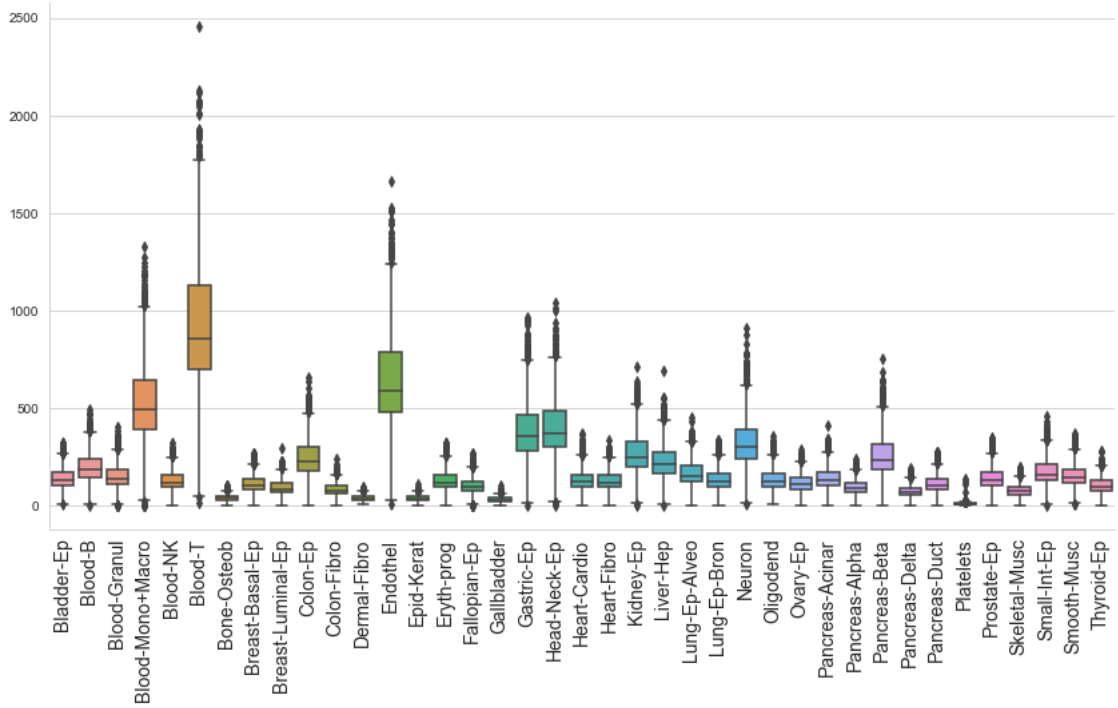

Figure 3.3: Depth (amount of reads) for every tissue and cell-type in atlas, distribution over entire set of markers linked to the shown cell-types.

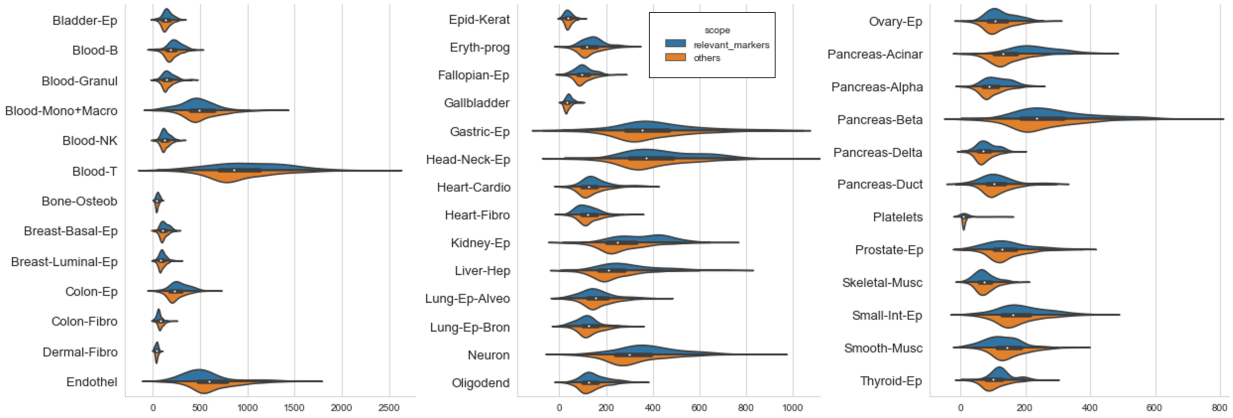

Figure 3.4: Depth distribution for every tissue and cell-type in atlas, comparing cell type-specific markers against all others.

Chapter 4

Approach

4.1 Deconvolution method

To analyze novel DNA methylation samples, composed of admixed methylomes from various cell types, we devised a computational deconvolution algorithm. We developed a generative mixture model [3] for plasma cfDNA methylation profile, based on the statistics defined in chapter 3 to describe methylation profiles in the reference atlas. In a mixture model we assume that a measurement x can be drawn from one of K data generating processes, each with their own set of parameters. For our problem, we assume two categorical distributions. First, a distribution for a read to be drawn from one of the K cell types, and second, a distribution of a read's label given it's cell type and window. Both distributions are parametric, where those parameters reflect a prior for the categories (cell type or label). The model is given the baseline estimate of relative contributions in healthy plasma cfDNA, and the estimate of the underlying distribution from the atlas. Then, given a mixed cfDNA sample, the model learns the posteriors of both distributions, using maximum likelihood estimation of the parameters (Fig. 4.1).

4.2 Feature selection

Our model parameters describe distributions of different cell-types over markers. Therefore, we used data only from cell type-significant markers to estimate parameters for each cell type and tissue. Thus, we aim to design a method that operates only on a fixed subset of the genome. Usually, mixture model suffer from identifiability[1]problem, where the underlying distributions for each category are not separable enough. In that case, one can virtually permute the order of categories, and derive the same conclusions. Marker selection in our model tackles this problem by prioritizing distribution estimates from cell type-specific markers. In practice, we are given the set of markers for cell types and tissues included in this work, hence this set is considered as the whole feature set (we simulate methylation profile for all of these markers for example), and use cell type-selection within the estimation procedure.

[1]Existence of a unique characterization for any one of the models in the class being considered. Estimation procedures may not be well-defined and asymptotic theory may not hold if a model is not identifiable.

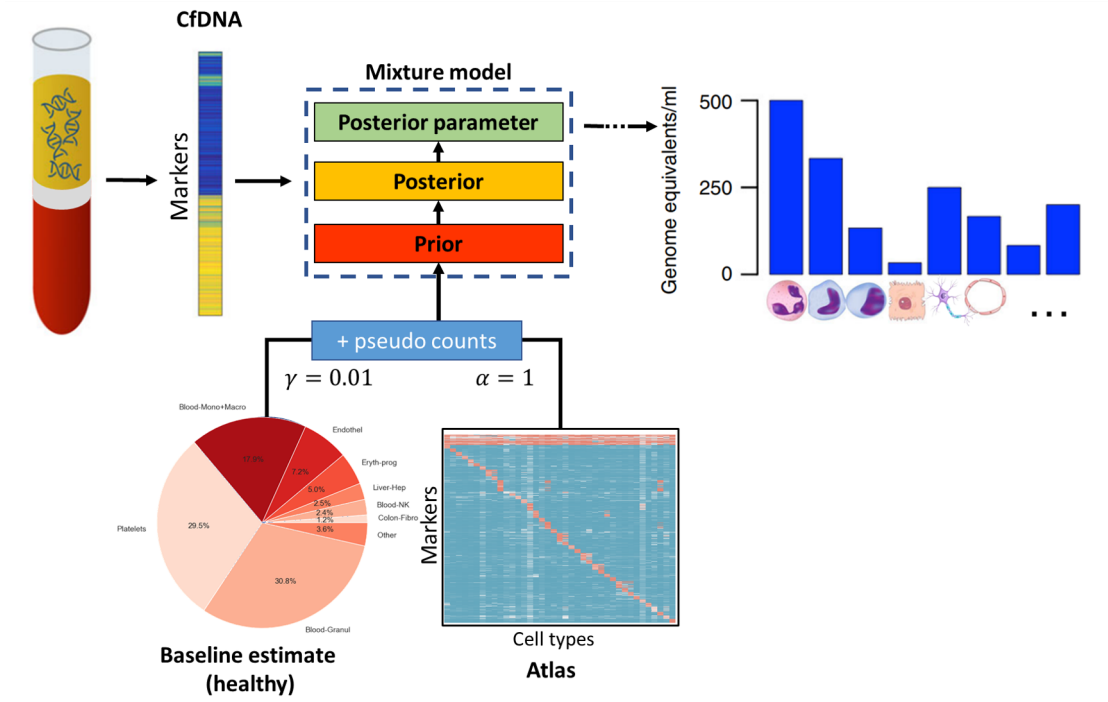

Figure 4.1: Relative contribution can be multiplied by the total concentration of cfDNA in plasma to obtain the absolute concentrations of cfDNA originating from each cell type (genome equivalents/ml).

4.3 Evaluation

Since the problem is unsupervised, we aim our model will be able to identify cfDNA contributions in high resolution for all cell-types, specifically from tissues appear rarely and linked to different pathologies. We test this ability in two levels of difficulty. First, we mix low levels of a single cell type methylome to an isolated whole-blood sample. Second, we generate a complete plasma cfDNA sample based on our model by different priors and test our model's ability to identify complete distribution under various conditions. Prior knowledge supports that most of plasma cfDNA originates from white blood cells. Therefore, we use white blood cells methylomes as an isolated sample to mix-in rare tissues. Generative simulation can test different random settings by sampling the contribution vector. To asses performance on realistic settings, we use real plasma cfDNA data and define statistical tests for our model to identify liver cancer and covid-19 effect on related tissues.

4.4 Robustness

Previous work targeted analysing DNA methylation data in an unbiased manner, where it is unclear which cell types contribute to cfDNA and which underlying diseases a patient may have. Our model operates under the null hypothesis that a plasma cfDNA comes from a healthy individual, and tries to identify a significant change to that prior. To test our inference method, we create plots and logs that keep track of important metrics for our algorithm correctness (for example, making sure likelihood is monotonically increasing). We test our model predictions on simulated

and real data to see if the estimations are consistent and statistically significant.

Chapter 5

Design and Implementation

5.1 Model

We use the following notation:

- K - number of tissues in the experiment.
- $\theta \in \mathbb{R}^K \cap$Simplex- distribution vector of tissues relative contributions (target).
- N - number of Markers (or windows).
- $n(i)$ - number of reads for the i'th window; a.k.a window's depth.
- n_i^U, n_i^M, n_i^X - number of labeled reads in the i'th window such that $n(i) = n_i^U + n_i^M + n_i^X$.
- $p_i^k(l) = \mathbb{P}(r = l|\text{r came from tissue k and marker i})$ for some read r.
- $\mathcal{P} \in \mathbb{R}^{N \times K \times 3}$ is the collection of all multinomial parameters $p_i^k(l)$ $(\forall l \in \{u, m, x\}, k \in [K], i \in [N])$.

5.1.1 Assumptions

1. We assume that a read's loci is also a stochastic process that depends on its originating cell-type or tissue. Namely, given a cell-type, there is a non-uniform distribution that determines which marker a read is from, hence θ reflects relative contribution over entire profile and not per window.

2. The number of CpG sites within a read is also random (between 2 sites to the size of the marker). Therefore, reads does not contribute the same amount of information, and does not introduce the same noise.

3. Read labels (the random variable we consider in the model) are completely independent random variables (between and within windows).

4. For all $k \in [K]$ it exists a group of markers $N_k \in [N]$ that encodes a unique methylation pattern for tissue k, that differs from all or most of the other tissues. Therefore, all other markers does not contribute relevant data to identify k, they are considered as noise and therefore ignored (feature selection).

5.1.2 Read Mixture Model

Let $i \in [N]$ and $j \in [n(i)]$ be the indices of a read $R_j^i \in \{u, m, x\}$ that was mapped to the i'th window, as the j'th observation in that window. Let $Z_j^i \sim$ Categorical$(\theta_1, \theta_2, ..., \theta_K)$ be a latent variable, corresponding to the mixture component of R_j^i . It holds that $R_j^i | Z_j^i = k \sim Categorical\left(p_i^k(u), p_i^k(m), p_i^k(x)\right)$. Therefore, the **likelihood of a single read conditioned by theta** is

$$\mathbb{P}_\theta\left(R_j^i = l\right) = \sum_{k=1}^{K} \mathbb{P}_\theta\left(Z_j^i = k\right) \mathbb{P}\left(R_j^i = l | Z_j^i = k\right) = \sum_{k=1}^{K} \theta_k \cdot \mathbb{P}\left(R_j^i = l | Z_j^i = k\right)$$

where $\mathbb{P}\left(R_j^i = l | Z_j^i = k\right) = p_i^k(u)^{\mathbf{1}_{\{l=u\}}} p_i^k(m)^{\mathbf{1}_{\{l=m\}}} p_i^k(x)^{\mathbf{1}_{\{l=x\}}}$. Under independence settings, the likelihood function of a collection of reads $\mathcal{R} = \{\{R_1^i, R_2^i, ..., R_{n(i)}^i\}\}_{i=1}^N$, is

$$L(\mathcal{R}) = \prod_{i=1}^{N} \prod_{j=1}^{n(i)} \sum_{k=1}^{K} \theta_k \cdot \mathbb{P}\left(R_j^i = l | Z_j^i = k\right)$$

$$\Rightarrow \ell(\mathcal{R}) = \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \log \left[\sum_{k=1}^{K} \theta_k \cdot \mathbb{P}\left(R_j^i = l | Z_j^i = k\right)\right]$$

Maximum likelihood estimator

Analytically deriving the estimate for θ by deriving the log-likelihood gradient under constraints does not result in a closed-form solution. We estimate the MLE of θ using the **Expectation Maximization algorithm** [3, 1]: The log-likelihood gradient with respect to $\Theta = (\theta, \mathcal{P})$ is

$$\begin{aligned}\frac{d}{d\Theta}\ell(r; \Theta) &= \mathbb{E}_{p(z|r)}\left[\log\left(\mathbb{P}(z)\,\mathbb{P}(r|z)\right)\right] \\ &= \sum_{k=1}^{K} p(z = k|r) \cdot \left[\log p(z = k) + \log \mathbb{P}(r|z = k)\right]\end{aligned}$$

where $p(z|r)$ is the posterior distribution of a category given the data. We define a surrogate $Q(\Theta \mid \Theta^{(t)}) = \mathbb{E}_{Z|R,\Theta^{(t)}}\left[\log L(\Theta; R, Z)\right]$, that our algorithm works to improve, rather than directly improving $\ell(\mathcal{R}; \Theta)$. Improvements to the former, imply improvements to the latter [2].

Given $\theta^{(0)}, \mathcal{P}^{(0)}$ as initial parameters, repeat t iterations until convergence:

1. **E-step**: compute the posterior distribution of the latent variables, also called responsibilities.

$$r_k^{ij} = \mathbb{P}\left(Z_j^i = k | R_j^i\right) = \frac{\theta_k \mathbb{P}\left(R_j^i | Z_j^i = k\right)}{\sum_{k'} \theta_{k'} \mathbb{P}\left(R_j^i | Z_j^i = k'\right)}$$
$$(\forall k \in [K], i \in [N], j \in [n(i)])$$

   note that r_k^{ij} actually depends only on i, k, and the reads label $l \in \{u, m, x\}$. In addition, $\theta_{k'}, \mathbb{P}\left(R_j^i | Z_j^i = k'\right)$ are estimates from the $(t-1)$ iteration.

2. **M-step**: given the responsibilities, we maximize $Q(\Theta \mid \Theta^{(t-1)})$, s.t

$$\Theta^{(t)} = \text{argmax}_{\Theta} \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \sum_{k=1}^{K} r_k^{ij} \left[\log \mathbb{P}\left(Z_j^i = k\right) + \log \mathbb{P}\left(R_j^i | Z_j^i = k\right) \right]$$

   (a) **Estimate** $\theta^{(t)}$: by solving the constrained problem we get $\hat{\theta}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n(i)} r_k^{ij}}{\sum_{i=1}^{N} n(i)}$, and by selecting subset of relevant markers for each tissue, $N_k \subseteq [N]$, we clean noise from the estimator by

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} \mathbf{1}_{\{i \in [N_k]\}} \sum_{j=1}^{n(i)} r_k^{ij}}{\sum_{i=1}^{N} \mathbf{1}_{\{i \in [N_k]\}} n(i)}$$

   (b) **Estimate** $\mathcal{P}^{(t+1)}$: The result of solving the constrained problem is

$$\hat{p}_i^k = \frac{1}{\sum_{l \in \{u,m,x\}} n_i^r r_k^{ir}} \begin{bmatrix} n_i^U r_k^{iU} \\ n_i^M r_k^{iM} \\ n_i^X r_k^{iX} \end{bmatrix}$$

Notes:

1. θ_0 used in our model is the given estimate for relative contributions in healthy patients (chapters 2,3). Using it as a prior in our model is equivalent to stating that the null hypothesis is that the patient is healthy. $\mathcal{P}_0$ is our estimates from the reference atlas (chapter 3).

2. To estimate θ, we allow updates of $\mathcal{P}$ to enable more degrees of freedom for the model to learn, comparing to keeping $\mathcal{P}$ fixed.

3. It's possible to show that each iteration of the algorithm increases the log-likelihood. We tested our program by tracking this metric and the convergence process of all parameters.

4. **Profile mixture model**. As a first attempt, we implemented a mixture model based on the complete formulation of mixture density, without considering latent variables. This method was able to find estimation using local search algorithms, but was outperformed by EM.

5. **Complete derivation and implementation** details available in `github.cs.huji.ac.il/guy-korn/Methylation_project`

5.2 Evaluation

We consider evaluation as testing the performance of the model in identifying cell types and tissues which composes small percentage of the entire cfDNA. Therefore, we group the methylome data of all white blood cells under a single category, and sum their relative contribution in θ.

5.2.1 In-silico mix-in simulations

In this simulation, we take methylation profile from white blood cells and mix unseen observations from a single tissue profile such that the tissue will comprise $\alpha \in \{0\%, ..., 10\%\}$ of the final profile. This kind of simulation was used to evaluate the baseline NNLS model, but since our statistics over the data were different, we designed a method for deriving the amount of samples required from the cell type and whole blood sample methylation profiles.
Notes:

1. To relax the conditions of this simulation, for every cell type evaluated, we used subsetting such that the model will have to estimate the relative contributions of the specific cell type and white blood cells.
2. We compared different sampling schemes of reads from methylation profiles, we chose random choice of reads per window since its more realistic.
3. complete details in `https://github.cs.huji.ac.il/guy-korn/Methylation_project`

5.2.2 Multi-class Evaluation

We generated plasma cfDNA samples according to our model, based on different priors for the cell-types relative contribution. We tested our model predictions for all tissues simultaneously.

Generative simulation

1. Given a prior $F(\theta)$, we sample distribution vector θ.
2. For every window $i \in [N]$:
   (a) Sample depth $n(i) \sim N(\mu, \sigma^2)$.
   (b) Sample $R_j^i \sim \sum_{k=1}^{K} \theta_k \cdot \text{Categorical}\left(p_i^k(u), p_i^k(m), p_i^k(x)\right)$ for every $j \in [n(i)]$.

Evaluation: given the generated methylation profile for cfDNA sample, we use the model to estimate $\hat{\theta}$ and record error metrics:

$$L_2(\theta, \hat{\theta}) = \sum_{k=1}^{K} (\theta_k - \hat{\theta}_k)^2$$

$$\text{single class error} = \frac{\theta_k - \hat{\theta_k}}{\theta_k} \quad \forall k \in [K]$$

Notes:

1. Equivalently, we could first sample $n_{i,1}, ..., n_{i,K} \sim$ Multinomial $(n(i), \theta)$, and then sample labels for observations from each tissue, i.e., $n^U_{i,k}, n^M_{i,k}, n^X_{i,k} \sim$ Multinomial $\left(n_{i,k}, p^\theta_i(u), p^\theta_i(m), p^\theta_i(x)\right)$ where $p^\theta_i(l) = \sum_{k=1}^K \theta_k p^k_i(l)$.

2. Sampling in each window depends both on θ and the distribution over that window per category, according to our model.

5.2.3 Hypothesis Testing

Using the plasma cfDNA from different conditions (healthy, liver cancer, and covid-19), we test the following:

Model is better than random

We denote our model as $M(\mathcal{R}; \theta_0, \mathcal{P}_0)$ where $\mathcal{R}$ is a plasma cfDNA sample (by our definition of methylation profile, chapter 3), θ_0 is the baseline prior for healthy plasma cfDNA and $\mathcal{P}_0$ is the set of estimated categorical distribution from the atlas (chapters 2,3). We mark our model's output as $\hat{\theta}$. Given

$$M(\mathcal{R}; \theta_0, \mathcal{P}_0)|\mathcal{R} \text{ is real data} \sim F$$
$$M(\mathcal{R}; \theta_0, \mathcal{P}_0)|\mathcal{R} \text{ is random data} \sim G$$

we define our null hypothesis $\mathcal{H}_0 : F = G$. Given a data set of estimation based on permuted real data, we used Kolmogorov–Smirnov test for goodness of fit for every cell type and tissue.

Identification of disease related tissue

Student's-T test. Compare means of a disease related tissue. where $\mathcal{H}_0 : \theta_k^{\text{healthy}} \geq \theta_k^{\text{disease}}$. For Liver cancer we test Liver-Hepatocytes, which already shown increased contribution for other models. For covid-19, we testes 2 types of Lung cells in the atlas.
Permutation test. We consider two random variables. $X \in \{\text{healthy}, \text{disease}\}$ and $Y = \hat{\theta}_k$, where $\mathcal{H}_0 : P(XY) = P(X)P(Y)$. Given a set of estimates from both healthy and diseased samples:

1. Permute labels across estimates for cell type k.

2. For each permutation, compute average estimate for k in each group and save the difference of means $D = \{d_1, ..., d_M\}$.

3. p_value $= \frac{1}{M} \sum_{j=1}^M \mathbf{1}_{\{d_j \geq d_{\text{real}}\}}$

Chapter 6

Experiments and Results

Single class identification. To asses the performance of the model in determining the relative contribution of various cell types to a methylation profile of DNA. For realistic and exhaustive assessment, we used whole-blood samples from 23 individuals. We then computationally mixed-in methylation profiles of individual samples of cell types and tissues at varying admixtures, reapplied the feature selection and deconvolution algorithm using an atlas from which the individual mixed-in sample was removed. We then compared the actual percentage with the predicted one. We simulated such data for every cell type in the reference methylation atlas, except for white blood cells, at mixing levels varying from 0% to 10% (in 1% intervals) across 115 replicates (23 independent whole-blood samples, times 5 replicates for each cell type). As shown in Fig.6.1, the model manged to identify a trend for all cell types, but with low accuracy of estimating the actual mix rate. Importantly, most cell type were detected even when composing $< 1\%$ of the mixture. Namely, when analysing pure leukocytes sample (Fig.6.1, leftmost site of each plot). Yet, it seems that the model did identify specific trends for every cell types, which indicates that it was able to learn some cell type-specific patterns. Namely, some cell types were over estimated for low mixture rate ($< 5\%$), but their estimation accuracy improved when more examples of the cell type methylome were added to the simulated sample (e.g., in Bladder-Ep, Pancreas-Beta).

Relative to the NNLS baseline model (chapter 2), our model performed worse for all tissues. To our understanding, one reason may be strong bias of our model towards its initial estimate of relative contribution, which is considered as a prior. For the results in Fig. 6.1, the model's prior was added with pseudo counts of $\gamma = 0.01$. As a comparison, we repeated the evaluation process using $\gamma = 0.001$. As shown in the example in Fig. 6.2. It seems that the effect of the prior adjustment was close to affine transformation of the trends identified by our model. The model sensitivity might occur due to the fact that the samples the model was given did not contained a lot of observations per marker. We constrained the simulations to generate samples with average of 30 observation across markers, for the purpose of developing a method that does not require deep and expensive sequencing of blood samples. Under these settings, the model's learning process appeared almost degenerate (Fig. 6.3).

Multi class evaluation. To asses the performance of the model when dealing with deviation from its prior while considering multiple cell types. We compared the performance of the model in deconvolution of samples that mixes all of the cell

types in the atlas. We compared samples generated by the prior given to the model against samples that follow uniform distribution over the cells. As shown in Fig.6.4 , the prior bias has strong significance, especially for tissues with high prior belief (endothel, white blood cells). **Model abilities on real data**. To asses performance on real scenarios, we compared estimates of plasma cfDNA samples taken from 23 healthy patients, 10 samples taken from liver cancer patients, and 12 samples from Covid-19 patients. Namely, we tested if our model can identify abnormalities that are related to a disease. For liver cancer, we compared Hepatocytes, which comprise approximately 70% of the liver tissue, and for Covid we compared two types of cells from the lungs. We found that the difference in means between each two groups (healthy, disease) was significant (p value < 0.05) for both cancer and Covid. On the contrary, permutation test (n=3000) did not rejected the null hypothesis that estimates are dependent between healthy and diseased. Further evidence supplied in Fig. 6.5, the main effect that can indicate a change in distribution appears in higher contribution the model estimate for liver hepatocytes. It appears that the signal of Lung cells for Covid-19 samples is weak, which corresponds with other models tested for this data. We can conclude that our model may estimated higher relative contribution of liver or lungs cell in samples from relevant diseases, but it seems that the model did not learned significant different distribution based on cfDNA in non healthy setting. Even though, we tested if our model estimate distribution is different comparing between real and permuted data, and as expected the null hypothesis was rejected (p value < 0.05) in a singe hypothesis test for every cell type. To conclude, these results reflect again that our model is able to identify trends in the data, but with low resolution and not good enough for realistic settings.

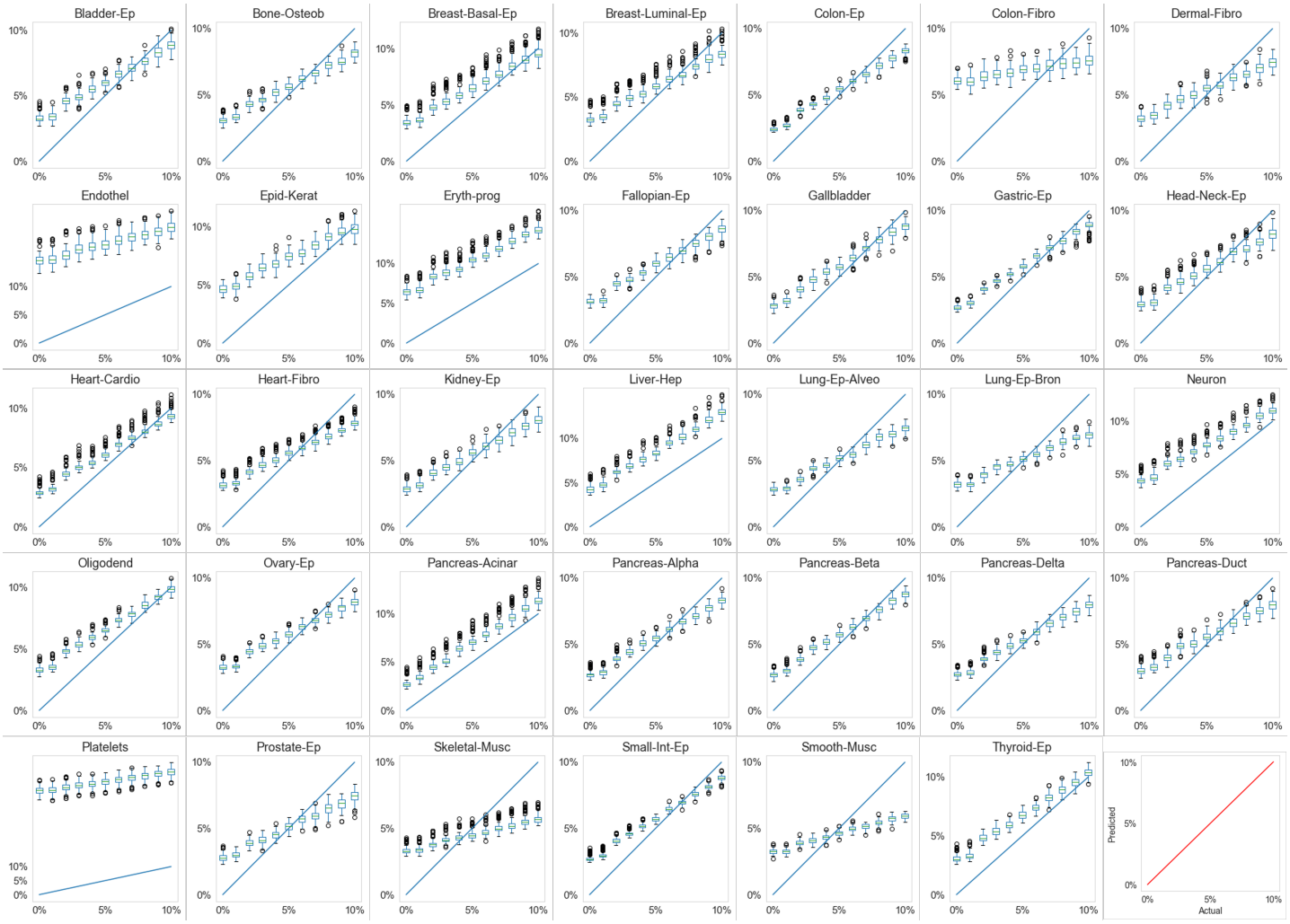

Figure 6.1: Deconvolution of simulated admixed samples. The methylome of each cell type was mixed in silico with the methylome of leukocytes such that it contributed between 0% and 10% of DNA, in 1% intervals (x-axis of each plot) and compared to the prediction of deconvolution using the reference methylation atlas (y-axis). Box plots represent median, IQR and outliers of predicted contribution for each mixed-in level, across 115 replicates for each cell type (5 replicates of measured cell type methylomes, each mixed within any of 23 leukocyte replicates).

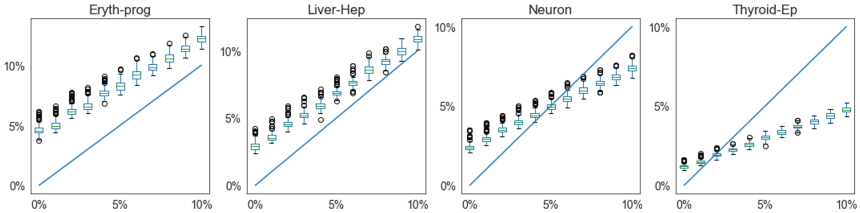

Figure 6.2: Mix in simulation results example with pseudo counts $\gamma = 0.001$ added to the relative contribution prior. Near affine transformation in compare to results in Fig. 6.1

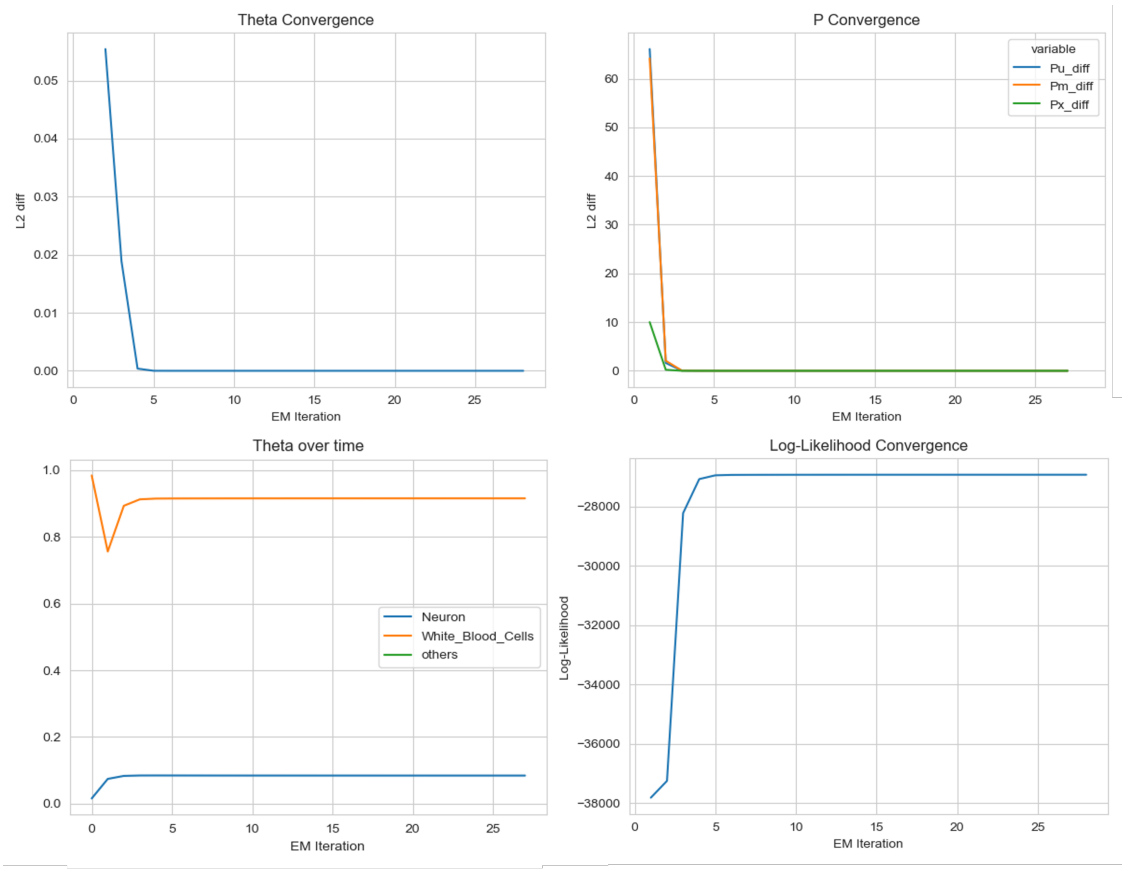

Figure 6.3: EM algorithm during model deconvolution of simulated sample when Neuron methylome composing 1% of the reads in the sample. First row: squared differences over iterations between both parameters $(\theta, \mathcal{P})$ that the model is trying to estimate. Second row: relative contribution of white blood cells against Neuron over iterations. And log-likelihood value over iterations.

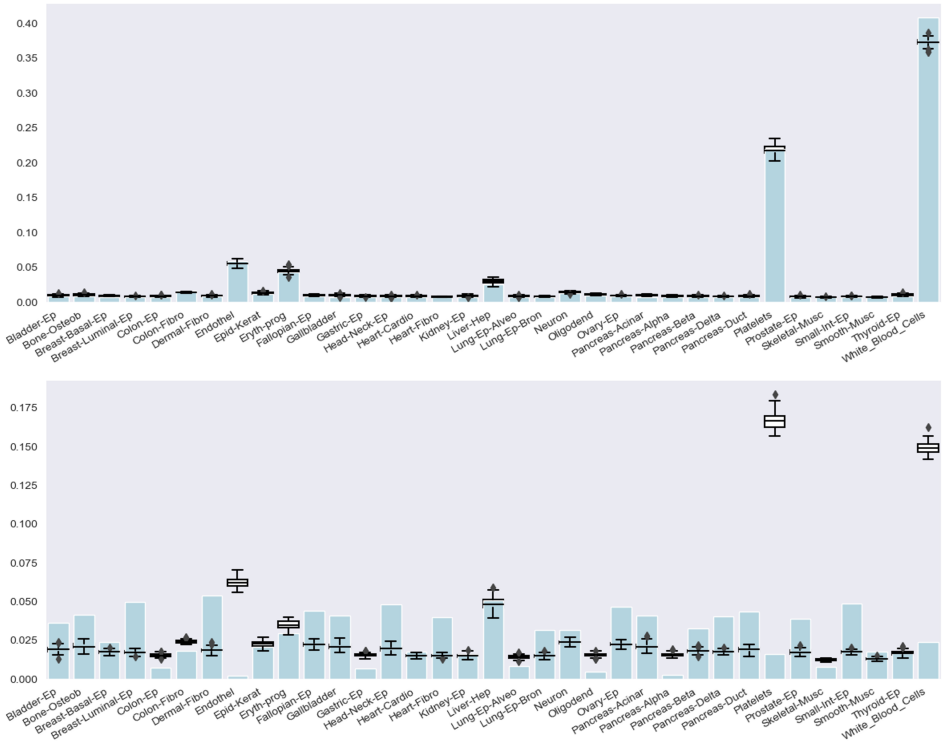

Figure 6.4: Model estimates over samples generated by different relative contribution vectors. For a given distribution vector (in blue), we simulated 100 samples following the generative procedure defined by our model, and by using the estimated distribution for every cell type from the reference atlas. The relative contributions estimated by our model are shown in the black box plots.

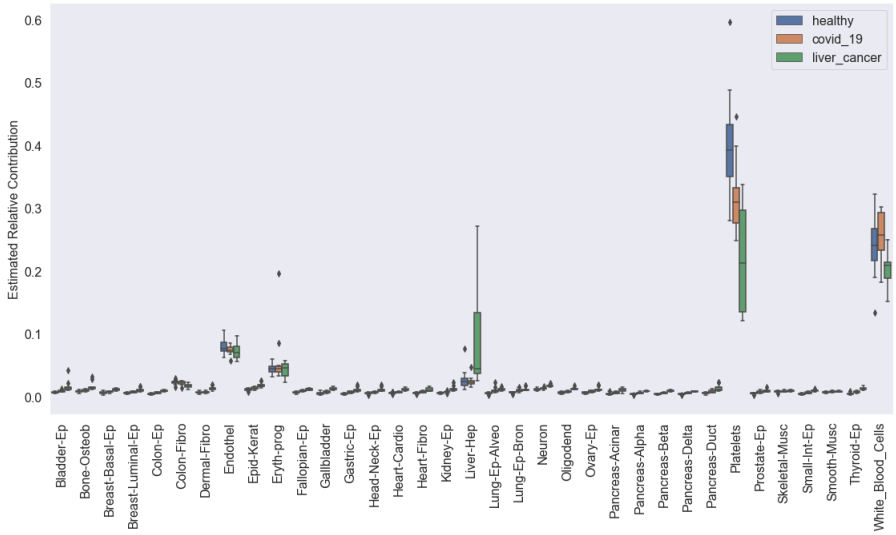

Figure 6.5: Model estimation over real plasma cfDNA methylation profiles, colors reflects different populations (healthy, liver cancer, Covid-19). Each box plot describes the estimation distribution over all samples in a pathology group, for every cell type.

Chapter 7

Discussion

In this work, we present a method to decipher the cellular origin of cfDNA by deconvoluting genome-wide methylation profiles, and use it to determine which cells release DNA into blood. Comparing to previous suggested models [4], our model showed **insufficient results for the required objective**. In particular, high bias effect that was probably affected by relatively low sample size (low and similar number of observation per window), those caused almost degenerate learning process for the model. However, our model did identify some trends in simulated and real data, in different settings.

To our understanding, some major flaws in model design could have caused this effect. First, comparing to previous work, **our model considered less data from every observation**. Every read is a sequence of multiple CpG sites, that each is methylated or not. Our statistics consider only the average of the methylated sites in a read. Therefore, we **ignore potential dependencies** between sites and information that might be encoded even in the amount of sites within an observation. Second, using feature selection in the model estimation procedure caused **integration of fewer examples** for our model to learn from. Although feature selection appeared to improve the model performance, it may be that performing weighted feature selection would improve our model performance. Lastly, our model design encompasses **bias by design**, so given the constraints on the number of observations it might have been that our **model choice was not suitable for the aim.**

7.1 Future Work

In attempt to reduce the impact of bias on our model, we suggest to expand the model configuration to encompass a prior distribution for the categorical distributions. Specifically, this model will consider distributions for the parameters of the categorical distributions in the model ($\theta \sim F, \mathcal{P} \sim G$). Preferably, a Dirichlet prior is used to describe distribution of probabilities, and the addition of pseudo counts for both parameters are already equivalent of estimating a posterior, given the atlas and baseline estimate of θ. Using this approach, one can infer posterior distributions for the parameters after clustering observations in a DNA methylation profile, and by that, can model uncertainty in estimations. A different approach can be to isolate inference for every cell type against all others (one vs. all), thus lowering the dimension of the problem, potentially increasing model sensitivity to minor changes

in contribution from different tissues. Additional features of the data, such as markers significance ranking per cell type, can be included in the model to supply the model with more information to base its estimations.

Appendix A

Personal Reflection

A.1 Reflection on Project

During the development of the model, we encountered problems regarding the proper form to formulate the model distributions. In particular, we spent a lot of time on deriving a solution to the complete formulation of the mixture density ([3]), until we learned about the missing data approach, which introduces latent variables and enabled the use of Expectation Maximization algorithm. Reflecting on our progress, we could have implemented a complete Bayesian inference over the distribution of the target parameter, as intended. We should have planned a wider background learning period to make better decisions.

Another major delay in our planning was caused due to an attempt to build a dynamic framework for experimentation that will allow the model to run on subsets of cell types in the atlas. Although this approach was proven to be helpful, we would have gone further if we decided on a specified subset from the beginning and focus on improving our model, or pivoting to a different one.

Appendix B

Appendices

B.1 More relevant material

- Data used in the project is under medical confidentiality, thus not attached to this work. Can be supplied by request.
- The entire code and further derivations and proofs regarding the design of the model and evaluation can be found in `https://github.cs.huji.ac.il/guy-korn/Methylation_project`.

Appendix C

Implementation

C.1 EM Review

Given $\theta^{(0)}, \mathcal{P}^{(0)}$ as initial parameters, repeat for iteration t until convergence:

1. **E-step**: compute the posterior distribution of the latent variables, also called responsibilities.

$$r_k^{ij} = \mathbb{P}\left(Z_j^i = k | R_j^i\right) = \frac{\theta_k \mathbb{P}\left(R_j^i | Z_j^i = k\right)}{\sum_{k'} \theta_{k'} \mathbb{P}\left(R_j^i | Z_j^i = k'\right)}$$
$$(\forall k \in [K], i \in [N], j \in [n(i)])$$

   note that r_k^{ij} actually depends only in i, k, and the reads label $l \in \{u, m, x\}$. In addition, $\theta_{k'}, \mathbb{P}\left(R_j^i | Z_j^i = k'\right)$ are estimates from the $(t-1)$ iteration.

2. **M-step**: given the responsibilities, we maximize $Q(\Theta \mid \Theta^{(t-1)})$, s.t

$$\Theta^{(t)} = \text{argmax}_\Theta \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \sum_{k=1}^{K} r_k^{ij} \left[\log \mathbb{P}\left(Z_j^i = k\right) + \log \mathbb{P}\left(R_j^i | Z_j^i = k\right)\right]$$

   (a) **Estimate** $\theta^{(t)}$: since the term $\log \mathbb{P}\left(R_j^i | Z_j^i = k\right)$ does not depend on θ, it holds that

$$\hat{\theta} = \text{argmax}_{\theta \in \text{Simplex}} \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \sum_{k=1}^{K} r_k^{ij} \log \theta_k$$

   by solving the constrained problem we get $\hat{\theta}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n(i)} r_k^{ij}}{\sum_{i=1}^{N} n(i)}$, and by selecting subset of relevant markers for each tissue, $N_k \subseteq [N]$, we clean noise from the estimator by

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} \mathbf{1}_{\{i \in [N_k]\}} \sum_{j=1}^{n(i)} r_k^{ij}}{\sum_{i=1}^{N} \mathbf{1}_{\{i \in [N_k]\}} n(i)}$$

   (b) **Estimate** $\mathcal{P}^{(t+1)}$: as before, the term $\log \mathbb{P}\left(Z_j^i = k\right)$ does not depends on $\mathcal{P}$, so it holds that

$$\hat{\mathcal{P}} = \text{argmax}_p \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \sum_{k=1}^{K} r_k^{ij} \log \mathbb{P}\left(R_j^i | Z_j^i = k\right)$$

and due to the independence in distribution for every window i and tissue k, the problem is equivalent to solving

$$\hat{p}_i^k = \text{argmax}_{p_i^k \in \text{Simplex}_3} \sum_{j=1}^{n(i)} r_k^{ij} \log \mathbb{P}\left(R_j^i | Z_j^i = k\right)$$

and by solving the constrained problem it holds that

$$\hat{p}_i^k = \frac{1}{\sum_{l \in \{u,m,x\}} n_i^r r_k^{ir}} \begin{bmatrix} n_i^U r_k^{iU} \\ n_i^M r_k^{iM} \\ n_i^X r_k^{iX} \end{bmatrix}$$

Notes:

1. θ_0 used in our model is the given estimate for relative contributions for healthy patients (chapters 2,3). Using it as a prior for our model is equivalent of stating the null hypothesis for this model is that the patient is healthy. $\mathcal{P}_0$ is our estimates from the reference atlas (chapter 3).

2. To estimate θ, we allow updates of $\mathcal{P}$ to allow more degrees of freedom for the model to learn, comparing to keeping $\mathcal{P}$ fixed.

3. It's possible to show that each iteration of the algorithm increases the log-likelihood. We tested our program by tracking this metric and the convergence process of all parameters.

4. Let V be a table of indicators that reflects the markers selection for every tissue. Follow the algorithm design, feature selection effects directly on estimating θ by filtering noise data. Even though, we tested approaches to allow soft selection by weighting V. For example, we tried to score every marker by its difference methylation distribution, and to reduce the effect of unbalanced amount of reads between tissues and markers (which might reflect another latent quality).

5. **Profile mixture model**. As a first attempt, we implemented a mixture model based on the complete formulation of mixture density, without considering latent variables. This method was able to find estimation using local search algorithms that were out preformed by EM.

6. **Complete derivation and implementation** details available in `github.cs.huji.ac.il/guy-korn/Methylation_project`

C.2 EM implementation

We denote the following:

1. $N_l = \left(n_1^l, .., n_N^l\right)^T$ be a column vector of l-labeled observations per window.

2. $P(u), P(m), P(x) \in \mathbb{R}^{N \times K}$ probability matrices, s.t $[P(l)]_{i,k} = p_i^k(l)$.

3. $V \in \{0,1\}^{N\times K}$ indicators matrix s.t every column vector $v_k \in \{0,1\}^N$ marks relevant markers for the k'th tissue. For tissue k' without relevant markers, we consider all markers, i.e., $v_{k'} = \vec{1}$.

4. $d_k = (n(1), ..., n(N)) \cdot v_k \in \mathbb{R}$, is the sum of all observations in markers relevant to tissue k, and $d^{-1} = \left(\frac{1}{d_1}, ..., \frac{1}{d_K}\right)^T$.

5. $v * u$ as element wise multiplication between vectors. $\overset{row}{*}, \overset{col}{*}$ will be used for multiplying element of a vector in every element in a row (or column) of a matrix.

C.2.1 E-step

we compute 3 matrices of size $N \times K$, such that for every $i \in [N], k \in [K], l \in \{u, m, x\}$ it holds

$$r_k^{il} = \sum_{j\in n(i), R_j^i = l} r_k^{ij} = \sum_{j\in n(i), R_j^i = l} \frac{\theta_k \mathbb{P}\left(R_j^i = l | Z_j^i = k\right)}{\sum_{k'} \theta_{k'} \mathbb{P}\left(R_j^i | Z_j^i = k'\right)}$$

for some l, it follows from the conditional likelihood that

$$r_k^{il} \propto \theta_k n_i^l p_i^k(l)$$

we define $\tilde{W}_l \in \mathbb{R}^{N\times K}$ such that $(\tilde{W}_l)_{i,k} = n_i^l \; p_i^k(l) \; \theta_k$ by

$$\tilde{W}_l = \begin{pmatrix} n_1^l \\ \vdots \\ n_N^l \end{pmatrix} \overset{row}{*} \begin{pmatrix} p_1^1(l) & \cdots & p_1^K(l) \\ \vdots & \ddots & \vdots \\ p_N^1(l) & \cdots & p_N^K(l) \end{pmatrix} \overset{col}{*} \begin{pmatrix} \theta_1 & \cdots & \theta_k \end{pmatrix} = N_l \overset{row}{*} P(l) \overset{col}{*} \theta$$

Now, the final result r_k^{il} can be obtained after normalizing every row in $\tilde{W}_l$ (sum of window l distribution over all tissues). we denote the final result as $W_l \in \mathbb{R}^{N\times K}$ where $r_k^{il} = (W_l)_{i,k}$.

C.2.2 M-step

θ computation

it holds that

$$\begin{aligned} hat\theta_k &= \frac{\sum_{i=1}^{N} \mathbf{1}_{\{i\in N_k\}} \sum_{j=1}^{n(i)} r_k^{ij}}{\sum_{i=1}^{N} n(i)\mathbf{1}_{\{i\in N_k\}}} \\ &= \frac{1}{d_k} \cdot \sum_{i=1}^{N} \mathbf{1}_{\{i\in N_k\}} \sum_{l\in\{u,m,x\}} n_i^l \cdot r_k^{il} \\ &= \frac{1}{d_k} \cdot \sum_{l} \sum_{i=1}^{N} \mathbf{1}_{\{i\in N_k\}} n_i^l \cdot r_k^{il} \\ &= \frac{1}{d_k} \sum_{l} v_k^T * \begin{pmatrix} n_1^l & \cdots & n_N^l \end{pmatrix} \begin{pmatrix} r_k^{1l} \\ \vdots \\ r_k^{Nl} \end{pmatrix} \\ &= \frac{1}{d_k} \sum_{l\in\{U,M,X\}} [W_l]_{\cdot k}^T (v_k * N_l) \end{aligned}$$

therefore

$$\begin{aligned} \hat{\theta} &= d^{-1} * \sum_{l\{u,m,x\}} \begin{pmatrix} [W_l]_{\cdot 1}^T (v_1 * N_l) \\ \vdots \\ [W_l]_{\cdot K}^T (v_K * N_l) \end{pmatrix} \\ &= d^{-1} * \sum_{l\in\{u,m,x\}} \text{sum\_rows} \left\{ W_l^T \overset{row}{*} \left(V \overset{col}{*} N_l \right)^T \right\} \\ &= d^{-1} * \sum_{l\in\{u,m,x\}} \text{sum\_rows} \left\{ \left(W_l \overset{col}{*} \left(V \overset{col}{*} N_l \right) \right)^T \right\} \end{aligned}$$

where all column of $S_l = V \overset{col}{*} N_l \in \mathbb{N}^{N\times K}$ represent feature selection of l-labeled observations for every tissue, and the operator $\text{sum\_rows}\left(W_l^T \overset{row}{*} S_l^T\right) \in \mathbb{R}^K$ represent dot product between the k-columns of W_l and S_l.

$\hat{\mathcal{P}}$ computation:

for every $i \in [N], k \in [K]$ we need to compute

$$\hat{p}_i^k = \frac{1}{\sum_{l\in\{u,m,x\}} n_i^r r_k^{ir}} \begin{bmatrix} n_i^U r_k^{iU} \\ n_i^M r_k^{iM} \\ n_i^X r_k^{iX} \end{bmatrix}$$

we could compute 3 matrices $\tilde{P}(l) \in \mathbb{R}^{N\times K}$ by

$$\tilde{P}(l) = \begin{bmatrix} N_l * r_1^{\cdot l} & \cdots & N_l * r_K^{\cdot l} \end{bmatrix} = N_l \overset{col}{*} W_l$$

and then normalize every i, k coordinate by the sum of that coordinate in all 3 tables, hence

$$\hat{P}(l)_{i,k} = \frac{\tilde{P}(l)_{i,k}}{\tilde{P}(u)_{i,k} + \tilde{P}(m)_{i,k} + \tilde{P}(x)_{i,k}}$$

C.2.3 Validation

to validate the process work as expected, we compute the log-likelihood function, which supposed to be monotonically increasing during EM iterations.

$$\ell(R) = \sum_{i=1}^{N}\sum_{j=1}^{n(i)} \log\left[\sum_{k=1}^{K}\theta_k \cdot \mathbb{P}\left(R_j^i = l | Z_j^i = k\right)\right] = \sum_{i=1}^{N}\sum_{l\in\{u,m,x\}} n_i^l \cdot \log\left[\sum_{k=1}^{K}\theta_k \cdot p_i^k(l)\right]$$

for some l, every element in the middle sum can be represented by

$$\sum_{i=1}^{N} N_l^T \cdot \log\begin{bmatrix}\theta^T p_1(l)\\ \vdots \\ \theta^T p_N^k(l)\end{bmatrix} = \sum_{i=1}^{N} N_l^T \cdot \log\left[\text{sum\_rows}\left(\theta^T \overset{col}{*} P(l)\right)\right]$$

hence $\ell(R) = \sum_l \sum_{i=1}^{N} N_l^T \cdot \log\left[\text{sum\_rows}\left(\theta^T \overset{col}{*} P(l)\right)\right]$.

C.2.4 Summary - EM algorithm

: Input:

1. initial parameters: $\theta \in \mathbb{R}^K$ and $P(u), P(m), P(x) \in \mathbb{R}^{N\times K}$.
2. data: methylation profile $S \in \mathbb{N}^{N\times 3}$ such that $S = [N_U, N_M, N_X]$, and a table $V \in \{0,1\}^{N\times K}$.

Preprocessing:

1. compute $S_l = V \overset{col}{*} N_l \in \mathbb{N}^{N\times K}$, s.t $[S_l]_{\cdot k}$ is a l-labeled observations vector, contains zeros where markers aren't relevant for a tissue.
2. compute $d^{-1} = \left(V^T N\right)^{-1} \in \mathbb{R}^K$, where $N = \sum_l N_l \in \mathbb{N}^N$.

Repeat until convergence:

1. E-step: compute $\tilde{W}_l = N_l \overset{row}{*} P(l) \overset{col}{*} \theta \in \mathbb{R}^{N\times K}$ and normalize by row, return W_l.
2. M-step:
   (a) compute $\hat{\theta} = d^{-1} * \sum_{l\in\{u,m,x\}} \text{sum\_rows}\left\{\left(W_l \overset{col}{*} S_l\right)^T\right\}$.
   (b) compute $\tilde{P}(l) = N_l \overset{col}{*} W_l \in \mathbb{R}^{N\times K}$, normalize for every i,k and get 3 tables for $l \in \{u,m,x\}$.

C.3 Evaluation

We consider evaluation as testing the performance of the model in identifying cell types and tissues which composes small percentage of the entire cfDNA. Therefore, we group the methylome data of all white blood cells under a single category, and sum their relative contribution in θ.

C.3.1 In-silico mix-in simulations

In this simulation, we take methylation profile from white blood cells, and mix unseen observations from a single tissue profile such that the tissue will comprise $\alpha \in \{0\%, ..., 10\%\}$ of the final profile.
Definition: coverage of a profile is the average depth over its windows, i.e., $\mu = \frac{1}{N}\sum_{i=1}^{N} n(i)$. we set coverage of the simulated profile $\mu_S = 30$ as hyper parameter.
Simulation process: let there be tissue profile S_k, mixing level α and a blood sample profile S_w.

1. We compute the coverage of S_k and S_w, μ_k and μ_w.

2. compute the percentage of observation needed to obtain α of tissue reads and $1-\alpha$ of blood sample reads in the mixed profile:

$$S_k\text{-Mix rate} = \alpha\frac{\mu_S}{\mu_k},\ S_w\text{-Mix rate} = (1-\alpha)\frac{\mu_S}{\mu_h}$$

3. Now, calculate the number of reads for every window $i \in [N]$, in the simulated profile$\tilde{n}(i) = \tilde{n}_k(i) + \tilde{n}_h(i)$ such that

$$\tilde{n}_k(i) = n_k(i) \cdot S_k\text{-Mix rate}\tilde{n}_w(i) = n_w(i) \cdot S_w\text{-Mix rate}$$

4. we randomly draw $\tilde{n}_k(i)$, $\tilde{n}_w(i)$ reads from the profiles of tissue k and blood sample w.

Notes:

1. To relax the conditions of this simulation, for every cell types evaluated, we used subsetting such that the model will have to estimate relative contributions only for the cell type and for white blood cells.

2. we want that the mix rates will be in $[0, 1]$, otherwise we could want require more than 100% than the available reads. In this case, the sampling becomes deterministic and repetition loses its meaning. The problem occurs only when $\mu_S \geq \mu_k, \mu_w$, which can happen due to randomness of tissue and blood profiles.

3. We compared different sampling schemes of reads from methylation profiles, we chose random choice of reads per window since its more realistic.

4. proof of correctness in `https://github.cs.huji.ac.il/guy-korn/Methylation_project`.

C.3.2 Mix in simulation validity proof

let S be the simulated profile by the described algorithm, it holds that

$$\begin{aligned}\sum_{i=1}^{N}\tilde{n}(i) &= \sum_{i=1}^{N}\alpha\cdot n_k(i)\frac{\mu_S}{\mu_k}+(1-\alpha)\cdot n_h(i)\frac{\mu_S}{\mu_w}\\ &= \alpha\cdot\frac{\mu_S}{\mu_k}\sum_{i=1}^{N}n_k(i)+(1-\alpha)\cdot\frac{\mu_S}{\mu_w}\sum_{i=1}^{N}n_w(i)\\ &= \alpha\cdot\mu_S\cdot N+(1-\alpha)\cdot\mu_S\cdot N=\mu_S\cdot N\\ &\Rightarrow \mu_S=\frac{\sum_{i=1}^{N}\tilde{n}(i)}{N}\end{aligned}$$

and

$$\frac{\sum_{i=1}^{N}\tilde{n}_k(i)}{\sum_{i=1}^{N}\tilde{n}(i)}=\frac{S_k\text{-Mix rate}\cdot\sum_{i=1}^{N}n_k(i)}{\sum_{i=1}^{N}\tilde{n}(i)}=\alpha\frac{\mu_S\cdot\sum_{i=1}^{N}n_k(i)}{\mu_k\sum_{i=1}^{N}\tilde{n}(i)}=\alpha\frac{\mu_S\cdot\mu_k\cdot N}{\mu_k\cdot\mu_S\cdot N}=\alpha$$

as needed.

C.3.3 Multi-class Evaluation

We generated plasma cfDNA samples following our model, based on different priors for the cell-types relative contribution. We tested our model predictions for all tissues simultaneously.

Generative simulation

1. given a prior $F(\theta)$, we sample distribution vector θ.
2. for every window $i\in[N]$:
   (a) sample depth $n(i)\sim N(\mu,\sigma^2)$.
   (b) sample $R_j^i\sim\sum_{k=1}^{K}\theta_k\cdot\text{Categorical}\left(p_i^k(u),p_i^k(m),p_i^k(x)\right)$ for every $j\in[n(i)]$.

Evaluation: given the generated methylation profile for cfDNA sample, we use the model to estimate $\hat{\theta}$ and record error metrics:

$$L_2(\theta,\hat{\theta})=\sum_{k=1}^{K}(\theta_k-\hat{\theta}_k)^2$$

$$\text{single class error}=\frac{\theta_k-\hat{\theta_k}}{\theta_k}\quad\forall k\in[K]$$

Notes:

1. Equivalently, we could first sample $n_{i,1},...,n_{i,K}\sim\text{Multinomial}\,(n(i),\theta)$, and then sample labels for observations from each tissue, i.e., $n_{i,k}^U,n_{i,k}^M,n_{i,k}^X\sim\text{Multinomial}\left(n_{i,k},p_i^\theta(u),p_i^\theta(m),p_i^\theta(x)\right)$ where $p_i^\theta(l)=\sum_{k=1}^{K}\theta_k p_i^k(l)$.

2. Note that sampling in each window depends both on θ and the distribution over that window for all category, following our model.

3. In every generated profile, we want the frequency of reads from every tissue will reflect its frequency in θ. The suggested simulation introduce noise, since $n(i)$ are generally low number of variables sampled by θ. Therefore, we tried an approach where given $n(i)$ we get $n_{i,k} = \lfloor n(i) \cdot \theta_k \rfloor$ for all $k \in [K]$.

C.3.4 Hypothesis Testing

Using the plasma cfDNA from different conditions (healthy, liver cancer, and covid-19), we test the following:

Model is better than random

We denote our model as $M(\mathcal{R}; \theta_0, \mathcal{P}_0)$ where $\mathcal{R}$ is a plasma cfDNA sample (by our definition of methylation profile, chapter 3), θ_0 is the baseline prior for healthy plasma cfDNA and $\mathcal{P}_0$ is the set of estimated categorical distribution from the atlas (chapters 2,3). We mark our model's output as $\hat{\theta}$. Given

$$M(\mathcal{R}; \theta_0, \mathcal{P}_0) | \mathcal{R} \text{ is real data} \sim F$$
$$M(\mathcal{R}; \theta_0, \mathcal{P}_0) | \mathcal{R} \text{ is random data} \sim G$$

we define our null hypothesis $\mathcal{H}_0 : F = G$, and use permutations test:

1. estimate $\hat{\theta}$ for every sample.

2. permute each sample in every marker (between u,m,x).

3. estimate $\hat{\theta}$ using the model.

4. for every $k = 1, .., K$, use Kolmogorov–Smirnov test for goodness of fit.

Identification of disease related tissue

Student's-T test. compare means of a disease related tissue. where $\mathcal{H}_0 : \theta_k^{\text{healthy}} \geq \theta_k^{\text{disease}}$. For Liver cancer we test Liver-Hepatocytes, which already shown increased contribution for other models. For covid-19, we testes 2 types of Lung cells in the atlas. **Permutation test**. We consider two random variables. $X \in \{\text{healthy}, \text{disease}\}$ and $Y = \hat{\theta}_k$, where $\mathcal{H}_0 : P(XY) = P(X)P(Y)$. Given a set of estimates from both healthy and diseased samples:

1. permute labels across estimates for cell type k.

2. for each permutation, compute average estimate for k in each group, and save the difference of means $D = \{d_1, ..., d_M\}$.

3. p_value $= \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{\{d_j \geq d_{\text{real}}\}}$

Consistency of estimation

We plot estimation for all tissues over the same sample, and look at the standard deviation.

Bibliography

[1] Roger Grosse and Nitish Srivastava. *CS231 course: Mixture models*. 2015. URL: https://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf.

[2] Donald B. Little Roderick J.A.; Rubin. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons. pp, 1987, pp. 134–136.

[3] Jean-Michel Marin, Kerrie Mengersen, and Christian P. Robert. "Bayesian Modelling and Inference on Mixtures of Distributions". In: *Bayesian Thinking*. Ed. by D.K. Dey and C.R. Rao. Vol. 25. Handbook of Statistics. Elsevier, 2005, pp. 459–507. DOI: https://doi.org/10.1016/S0169-7161(05)25016-2.

[4] J. Moss et al. "Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease." In: *Nature Communications* 9.5068 (2018), pp. 2041–1723. DOI: https://doi.org/10.1038/s41467-018-07466-6.