Paulina Jedynak
Team of Environmental Epidemiology
Institute for Advanced Biosciences
Centre de Recherche UGA, Grenoble, France
paulina.jedynak@univ-grenoble-alpes.fr

Tutorial on how to apply the reference free method for estimation of the cell mix proportions in a sample using RefFreeEWAS package for R (Houseman et al. 2016)

10/05/20

The concept of methylome matrix deconvolution (Houseman et al. 2016):

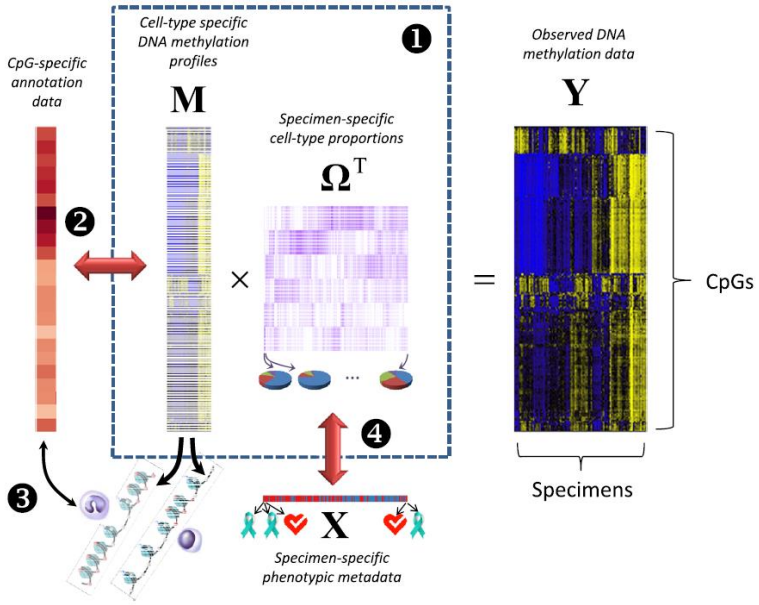

Fig. 1 Overview of proposed Methods. If associations between DNA methylation data **Y** and phenotypic metadata **X** factor through the decomposition $\mathbf{Y} = \mathbf{M}\boldsymbol{\Omega}^T$, and the data in **M** serve to distinguish cell types by their associations with relevant annotation data, then associations between **X** and **Y** are explained in whole or in part by differences in the distribution of constituent cell types. Numbers indicate steps in analysis: (1) deconvolution; (2) determining discriminating loci; (3) gene-set analysis; (4) analysis of associations with phenotype

There are 4 important terms: Y – observed methylation matrix; K – number of cell types (a scalar); Omega (Ω) – cell mix proportion matrix; Mu (M) - matrix representing CpG-specific methylation states for each of K cell types. Also to remember: m – number of CpGs, n – number of subjects.

To deconvolute the methylation matrix Y we need to solve the equation: Y = M x Ω^T, where $M = (\mu_j^T)_{j\in\{1,\ldots,m\}}^T = (\mu_{jk})_{j\in\{1,\ldots,m\},k\in\{1,\ldots,K\}}$ is an *unknown* m x K matrix with each column representing methylation profile for each cell type and $\Omega = (\omega_i^T)_{i\in\{1,\ldots,n\}}^T = (\omega_{ik})_{i\in\{1,\ldots,m\},k\in\{1,\ldots,K\}}$ is an *unknown* n x K matrix representing subject-specific cell-type distributions (each row representing the cell-type proportions for a given subject, i.e. the entries of Ω lie within [0, 1] and the rows of Ω sum to values less than one).

Paulina Jedynak
Team of Environmental Epidemiology
Institute for Advanced Biosciences
Centre de Recherche UGA, Grenoble, France
paulina.jedynak@univ-grenoble-alpes.fr

PART I

To be able to deconvolute Y, we first need to estimate K (number of cell types). K is the number of columns in Ω matrix. Choosing the right K is crucial for the correct estimation of the cell-type proportions.

Estimation of K can be done in several ways, here I present two methods: based on (Houseman et al. 2016) and (Decamps et al. 2020).

Houseman et al. 2016

Estimation of K is done by testing several values of K (from 1 to K, where 1 means one cell type and K means K cell types) and choosing the optimal one. In practice it goes like this:

For each K (from 1 to K) of assumed cell types, we estimate M and Ω as follows:

1. Start with an initial estimate of M.
2. Fixing M, construct a new $\Omega = (\omega_i^T)_{i\in\{1,\dots,n\}}^T$: for each $i \in \{1, \dots, n\}$, minimize $|y_i^{(c)} - M\omega_i|^2$ subject to the constraints $0 \leq \omega_{ik} \leq 1$ and $\sum_{k=1}^{k} \omega_{ik} \leq 1$.
3. Fixing Ω, construct a new $M = (\mu_j^T)_{j\in\{1,\dots,m\}}^T$: for each j$\in \{1, \dots, m\}$, minimize $|y_j^{(r)} - \Omega\mu_j|^2$ subject to the constraints $0 \leq \mu_{jk} \leq 1$.
4. Repeat steps 1.-2. a specific number of times.

This is an extremely heavy computation (each step is done in several iterations to minimize the error). That is why, it is established to use, instead of the entire methylome, only the most informative part of it, i.e. 10k most variable CpGs (Houseman et al. 2016).

Now, from all tested Ks we need to choose the right one, Houseman et al. propose to use deviance statistic estimated with bootstrap using the RefFreeCellMixArrayDevianceBoots function from the RefFreeEWAS package.

Decamps et al. 2020

Decamps et al. propose an alternative method of estimation of K. They first recommend to select most informative probes based not on their general variability, but filtering out those probes whose methylation is affected by factors that are not influencing cell proportions. In this approach one needs to decide which variables may influence methylation but are not affecting the cell mix. An example for placental methylation can be: delivery mode – may affect methylation in placenta but happens after the constitution of the cell mix so it cannot influence it; batch effects - may affect methylation in placenta but cannot affect the cell proportions in placenta, etc. As the factors are defined, a univariate regression is run between methylation matrix and each of the factors and all the probes associated with these factors with the significance level of 15% are discarded.

Paulina Jedynak
Team of Environmental Epidemiology
Institute for Advanced Biosciences
Centre de Recherche UGA, Grenoble, France
paulina.jedynak@univ-grenoble-alpes.fr

K is chosen after visual inspection of the scree plot drawn on the PCA result for the selected probes. Cattell's rule is applied to choose the right K (K = PCs + 1, (Cattell 1966)).

PART II

As we have chosen the number of cell-types K we can run the cell proportion estimation using the RefFreeCellMix function from the RefFreeEWAS package. Depending on which method we have chosen in the previous step, a relevant version of the RefFreeCellMix needs to be applied.

Houseman et al. 2016

We estimate the cell proportion mix for the K defined in the previous step using either 10k most variable CpGs or all available probes.

Decamps et al. 2020

We estimate the cell proportion mix for the K defined in the previous step using selected probes.

In this part we will obtain 2 resulting matrices: cell-type proportion estimates (Ω), which is the outcome of interest, and a complementing matrix (M) representing CpG-specific methylation states for the given K. Multiplication of these 2 matrices will recreate the input matrix Y.

Implementation in R:

A practical step-by-step tutorial in R can be found here:
https://rpubs.com/paujedynak/reffree_cell_mix_tutorial

References:

Cattell RB. 1966. The Scree Test For The Number Of Factors. Multivariate Behav Res 1:245–276; doi:10.1207/s15327906mbr0102_10.

Decamps C, Privé F, Bacher R, Jost D, Waguet A, Achard S, et al. 2020. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. BMC Bioinformatics 21:16; doi:10.1186/s12859-019-3307-2.

Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. 2016. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics 17:259; doi:10.1186/s12859-016-1140-4.

Supplementary material from (Houseman et al. 2016) where detailed explanation of the K and cell-type proportion matrix estimation is provided.

Reference-free deconvolution of DNA methylation data and mediation by cell composition effects – Supplementary Information

Paulina Jedynak
Team of Environmental Epidemiology
Institute for Advanced Biosciences
Centre de Recherche UGA, Grenoble, France
paulina.jedynak@univ-grenoble-alpes.fr

E. Andres Houseman[1], Molly Kile[1], David C. Christiani[2], Tan A. Ince[3], Karl T. Kelsey[4], Carmen J. Marsit[5]

Section S1 – Convex Deconvolution of DNA Methylation Data

We assume an $m \times n$ matrix $\mathbf{Y}$ representing methylation data collected for n subjects or specimens, each measured on an array of m CpG loci, and that the measured values are constrained to the unit interval $[0,1]$. We explicitly write $\mathbf{Y}$ in terms of its row vectors $\mathbf{Y} = (\mathbf{y}_j^{(r)\mathrm{T}})_{j\in\{1,\ldots,m\}}^{\mathrm{T}}$ and its column vectors $\mathbf{Y} = (\mathbf{y}_i^{(c)})_{i\in\{1,\ldots,n\}}$. We also assume the following decomposition: $\mathbf{Y} = \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$, where $\mathbf{M} = (\boldsymbol{\mu}_j^{\mathrm{T}})_{j\in\{1,\ldots,m\}}^{\mathrm{T}} = (\mu_{jk})_{j\in\{1,\ldots,m\},k\in\{1,\ldots,K\}}$ is a *unknown* $m \times K$ matrix representing m CpG-specific methylation states for each of K cell types (with row vectors representing profiles each individual CpG) and $\mathbf{\Omega} = (\boldsymbol{\omega}_i^{\mathrm{T}})_{i\in\{1,\ldots,n\}}^{\mathrm{T}} = (\omega_{ik})_{i\in\{1,\ldots,n\},k\in\{1,\ldots,K\}}$ is an *unknown* $n \times K$ matrix representing subject-specific cell-type distributions (each row representing the cell-type proportions for a given subject, i.e. the entries of $\mathbf{\Omega}$ lie within $[0,1]$ and the rows of $\mathbf{\Omega}$ sum to values less than one). For a fixed number K of assumed cell types, we estimate $\mathbf{M}$ and $\mathbf{\Omega}$ as follows:

1. Start with an initial estimate of $\mathbf{M}$.
2. Fixing $\mathbf{M}$, construct a new $\mathbf{\Omega} = (\boldsymbol{\omega}_i^{\mathrm{T}})_{i\in\{1,\ldots,n\}}^{\mathrm{T}}$: for each $i \in \{1,\ldots,n\}$, minimize $\left\|\mathbf{y}_i^{(c)} - \mathbf{M}\boldsymbol{\omega}_i\right\|^2$ subject to the constraints $0 \le \omega_{ik} \le 1$ and $\sum_{k=1}^{K} \omega_{ik} \le 1$.
3. Fixing $\mathbf{\Omega}$, construct a new $\mathbf{M} = (\boldsymbol{\mu}_j^{\mathrm{T}})_{j\in\{1,\ldots,m\}}^{\mathrm{T}}$: for each $j \in \{1,\ldots,m\}$, minimize $\left\|\mathbf{y}_j^{(r)} - \mathbf{\Omega}\boldsymbol{\mu}_j\right\|^2$ subject to the constraints $0 \le \mu_{jk} \le 1$.
4. Repeat steps (1)-(2) a specific number of times.

The constrained optimizations in steps (1) and (2) can easily be achieved using a quadratic programming algorithm[1] implemented in the *R* library *quadprog*. We note that if $\mathbf{M}$ is chosen reasonably well, a relatively few number of iterations will be necessary to achieve near-convergence. For the present analysis, 25 iterations were used; Figure S2.1 displays box-and-whisker plots for the distribution of absolute differences (absolute values of the entries of $\mathbf{Y} - \mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$) between the last two iterations of the $K = 2$ fit, while Figure S2.2 displays the corresponding plot for $K = K^* = \max(2, \hat{K})$, where $\hat{K}$ was the estimated number of classes as described below in Section S3. As suggested by the figures, the error was typically less than 0.01, and often about 0.001 or less.

For the present analysis, we have initialized $\mathbf{M}$ step (0) as follows: we used hierarchical clustering to cluster the columns of $\mathbf{Y}$ (i.e. using a Manhattan metric and Ward's method of clustering), formed K classes from the resulting dendrogram, and initialized $\mathbf{M}$ as the K mean methylation vectors corresponding to each class. In this way, $\mathbf{M}$ was initialized in a manner consistent with the RPMM algorithm[2], widely used in DNA methylation analysis.

Paulina Jedynak
Team of Environmental Epidemiology
Institute for Advanced Biosciences
Centre de Recherche UGA, Grenoble, France
paulina.jedynak@univ-grenoble-alpes.fr

A substantial portion of the variation between cell-type specific methylomes will be driven only by the most evidently variable CpG loci, with the remaining loci contributing only noise; consequently, for the present analysis we have selected the $m = 5000$ most variable CpGs (within each data set) for the 27K data sets, and the $m = 10{,}000$ most variable CpGs (within each data set) for the 450K data sets. However, subsequent to step (3) in the algorithm, with the value of $\boldsymbol{\Omega}$ estimated, we constructed a new $\mathbf{M}$ for the full array, as in step (2).