Fonds National de la Recherche Luxembourg

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Proteome and Genome Research Unit

Non-negative Matrix Factorization for Methylation Data Deconvolution

Update on the Data Challenge and the related paper: Lutsik et al

Petr Nazarov

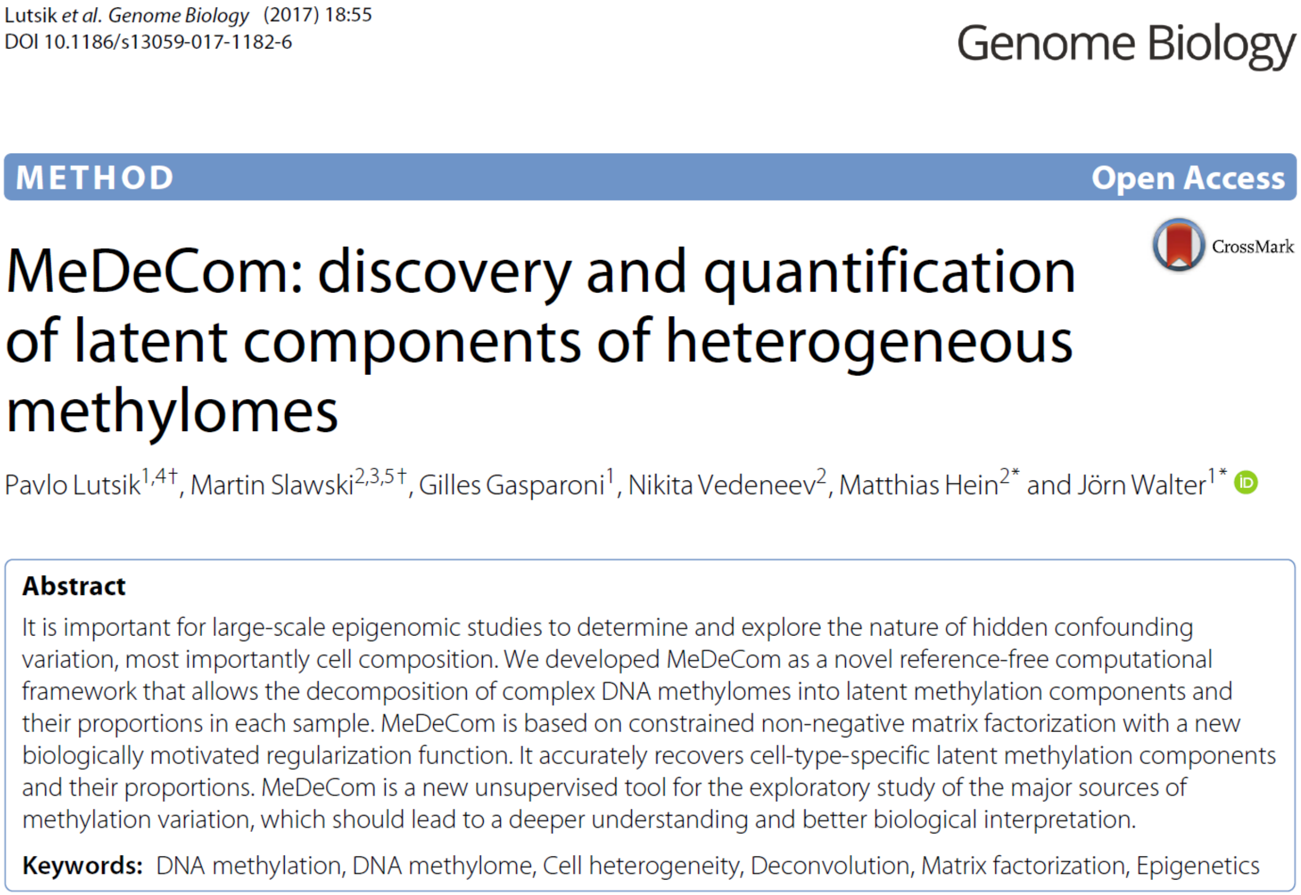
petr.nazarov@lih.lu

LIH, Strassen, Luxembourg

2019-01-22

Outline

- ➢ Concept of NMF
- ➢ Update on the Data Challenge
- ➢ Paper



Lutsik *et al. Genome Biology* (2017) 18:55
DOI 10.1186/s13059-017-1182-6

Genome Biology

METHOD Open Access

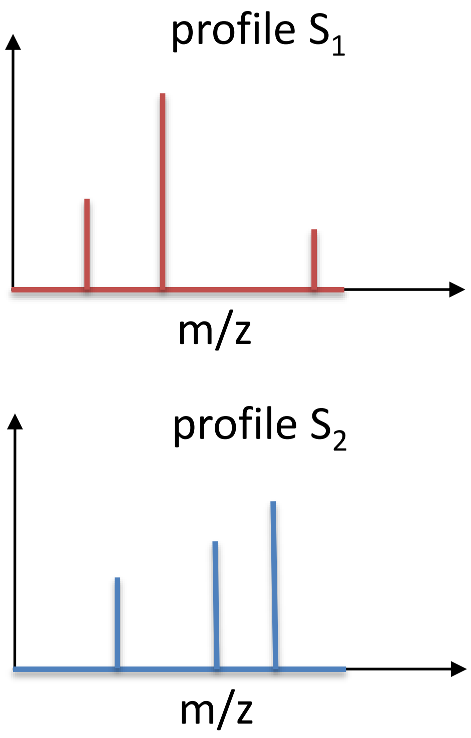
MeDeCom: discovery and quantification of latent components of heterogeneous methylomes

Pavlo Lutsik[1,4†], Martin Slawski[2,3,5†], Gilles Gasparoni[1], Nikita Vedeneev[2], Matthias Hein[2*] and Jörn Walter[1*]

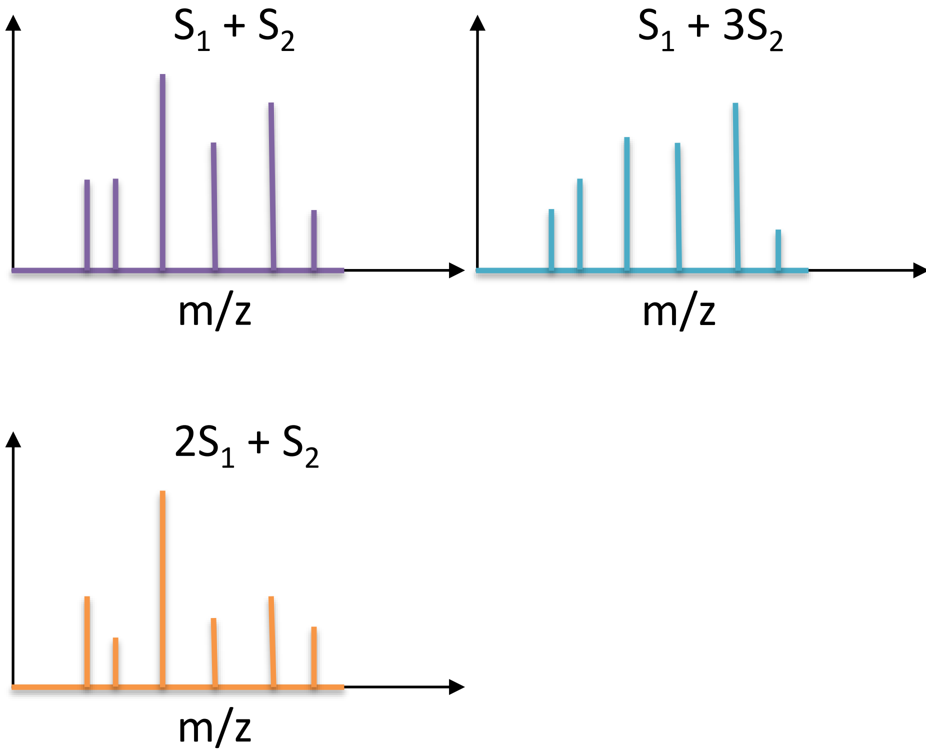
Abstract
It is important for large-scale epigenomic studies to determine and explore the nature of hidden confounding variation, most importantly cell composition. We developed MeDeCom as a novel reference-free computational framework that allows the decomposition of complex DNA methylomes into latent methylation components and their proportions in each sample. MeDeCom is based on constrained non-negative matrix factorization with a new biologically motivated regularization function. It accurately recovers cell-type-specific latent methylation components and their proportions. MeDeCom is a new unsupervised tool for the exploratory study of the major sources of methylation variation, which should lead to a deeper understanding and better biological interpretation.

Keywords: DNA methylation, DNA methylome, Cell heterogeneity, Deconvolution, Matrix factorization, Epigenetics

Concept: NMF



Mixing



$$\mathbf{X} = \mathbf{S} \times \mathbf{M}$$

$$\mathbf{M} = \begin{array}{|c|c|c|}\hline 1 & 2 & 1 \\ \hline 1 & 1 & 3 \\ \hline \end{array}$$

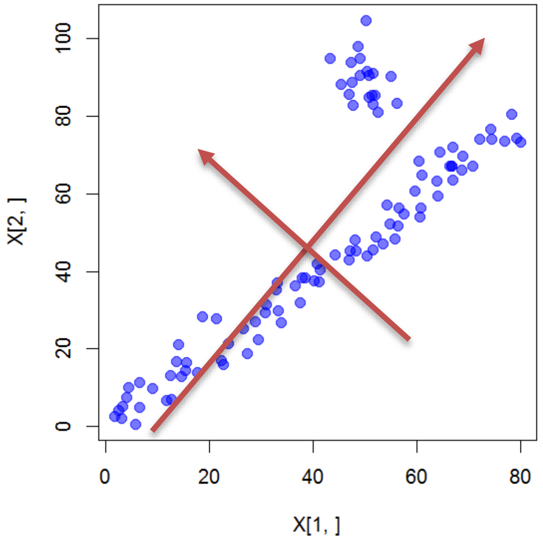
Non-negative **Non-negative** **Non-negative**

$$\mathbf{X} \approx \mathbf{T} \times \mathbf{A}$$

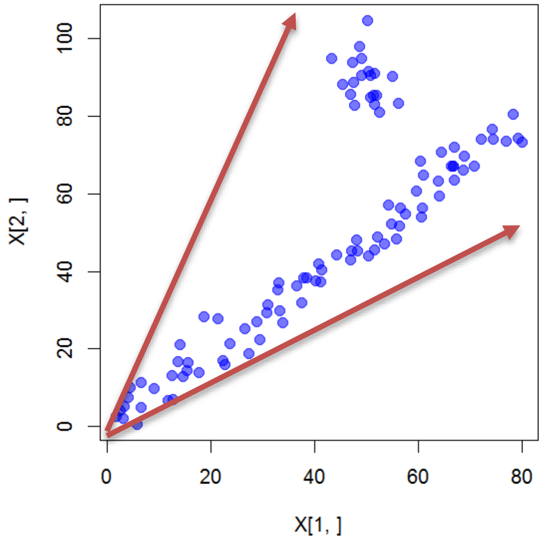
~S ~M

Concept: NMF

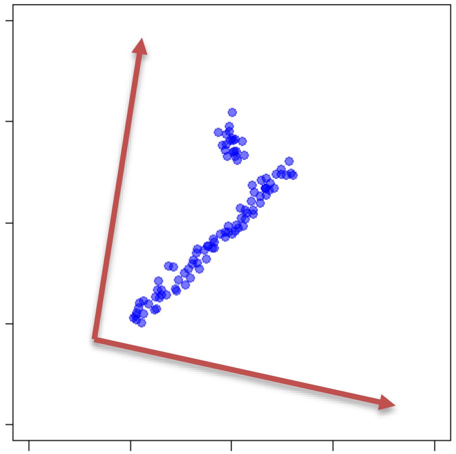
PCA



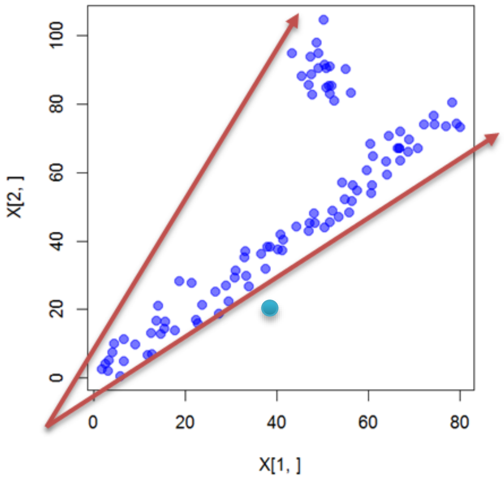
NMF



NMF: issue 1



NMF: issue 2



- Multiple solutions
- Is the minimal description stable?

⇒ we need:

- ➢ additional restrictions
- ➢ regularizations during fitting

Data Challenge

Place & Participants

Aussois, Dec 2018

Invited speakers:

- E. Andres **Houseman**, independent data scientist, USA, **RefFreeEWAS**
- Pavlo **Lutsik**, from DKFZ, Heildeberg, Germany, **MeDeCom**
- Eugene **Lurie**, from BCM, Houston, USA, **Edec**

Participants:

- 9 (10) commands 3-4 members: FR, DE, US, RU, LU, NL, … ?

Data Challenge

Structures

2 sub-challenges: **X = S x M + noise**

- **Training**: 3 cell types, 100 synthetic samples, no confounding variables
- **Main**: *k* cell types (*k*=5), 100 synthetic samples, *y* confounding variables

Our team:

- Fabian Bergmann (MSc student) – IT, submits, fine tuning, RefFreeEWAS
- Tony Kaoma – wide search for alternative algorithms, MeDeCom
- Petr Nazarov – ICA, moderating FB ☺

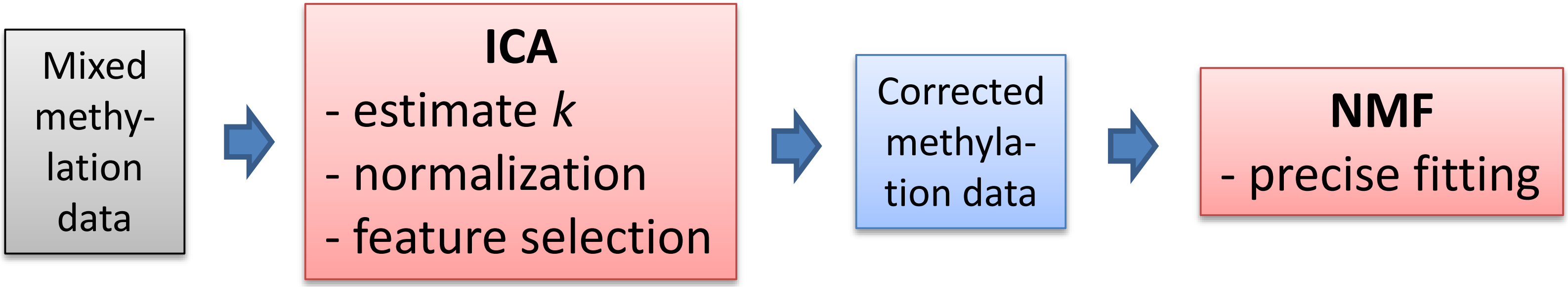
Reasons why we won sub-challenge 1:

- our tuning of the parameters was more efficient – RefFreeEWAS overfits!
- search for methods for initial estimation by TK helped

…and ICA was not needed at al ☺ !

Data Challenge

Winning strategy for sub-challenge 2



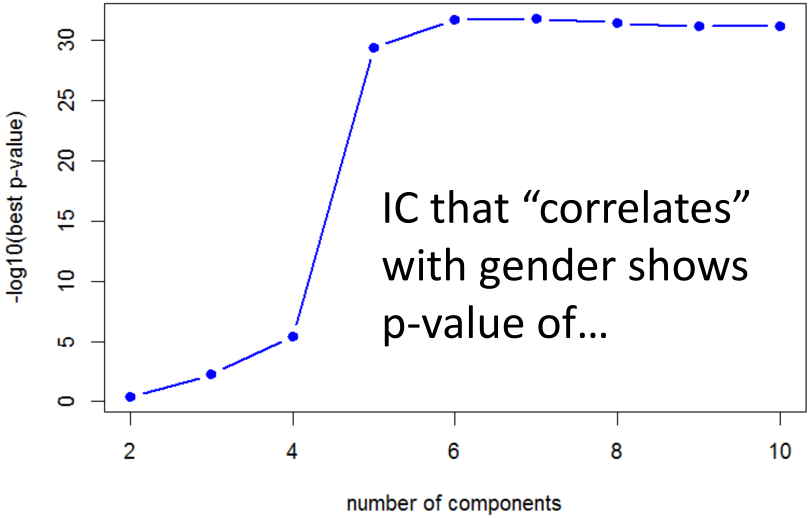
Mixed methylation data

ICA
- estimate *k*
- normalization
- feature selection

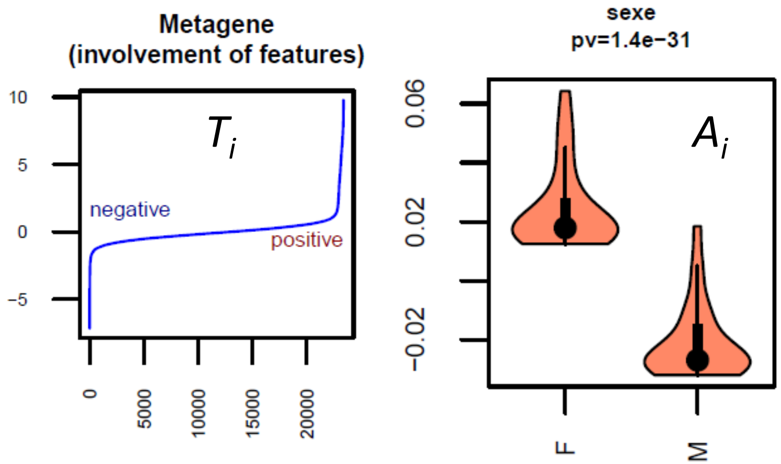
Corrected methylation data

NMF
- precise fitting

Gender was one of the confounders



IC that "correlates" with gender shows p-value of…



X = T x A

Data Challenge

Teams

Pavlo Lutsik: MeDeCom

Eugene Lurie: Edec

Almost the only biologist ☺

Data Challenge

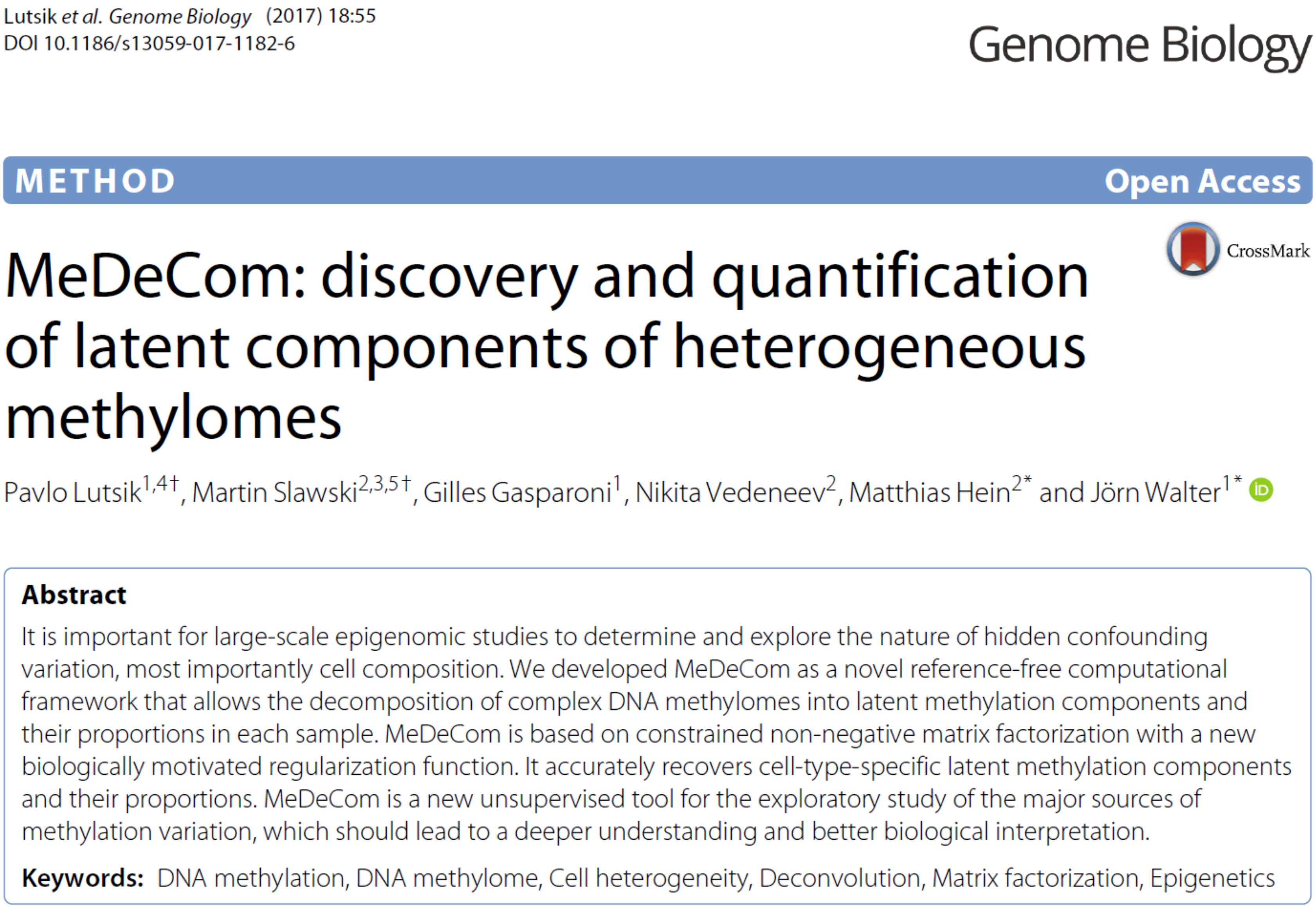
Conclusions & Results

- Even the top methods on NMF must be fine-tuned in order to give good results
- ICA is of no help for simple tasks. But can be useful in more complex situations (e.g. confounders)
- Pavlo Lutsik, the developer of MeDeCom and co-author of RnBeads, was “as**toni**shed” by Tony’s results on his own tool and proposed to work together on the protocols paper
- The general paper based on the challenge was planned, but i.m.h.o., the chances are vague
- It brings new knowledge and simply… a lot of fun ☺

E.Andreas Houseman: RefFreeEWAS

supported by Luxembourg National Research Fund (C17/BM/11664971/DEMICS)

MeDeCom Paper



Lutsik *et al. Genome Biology* (2017) 18:55
DOI 10.1186/s13059-017-1182-6

Genome Biology

METHOD Open Access

MeDeCom: discovery and quantification of latent components of heterogeneous methylomes

Pavlo Lutsik[1,4†], Martin Slawski[2,3,5†], Gilles Gasparoni[1], Nikita Vedeneev[2], Matthias Hein[2*] and Jörn Walter[1*]

Abstract

It is important for large-scale epigenomic studies to determine and explore the nature of hidden confounding variation, most importantly cell composition. We developed MeDeCom as a novel reference-free computational framework that allows the decomposition of complex DNA methylomes into latent methylation components and their proportions in each sample. MeDeCom is based on constrained non-negative matrix factorization with a new biologically motivated regularization function. It accurately recovers cell-type-specific latent methylation components and their proportions. MeDeCom is a new unsupervised tool for the exploratory study of the major sources of methylation variation, which should lead to a deeper understanding and better biological interpretation.

Keywords: DNA methylation, DNA methylome, Cell heterogeneity, Deconvolution, Matrix factorization, Epigenetics

MeDeCom Paper

D = T×A + E

Main idea (sorry, it is simple but… ☺)

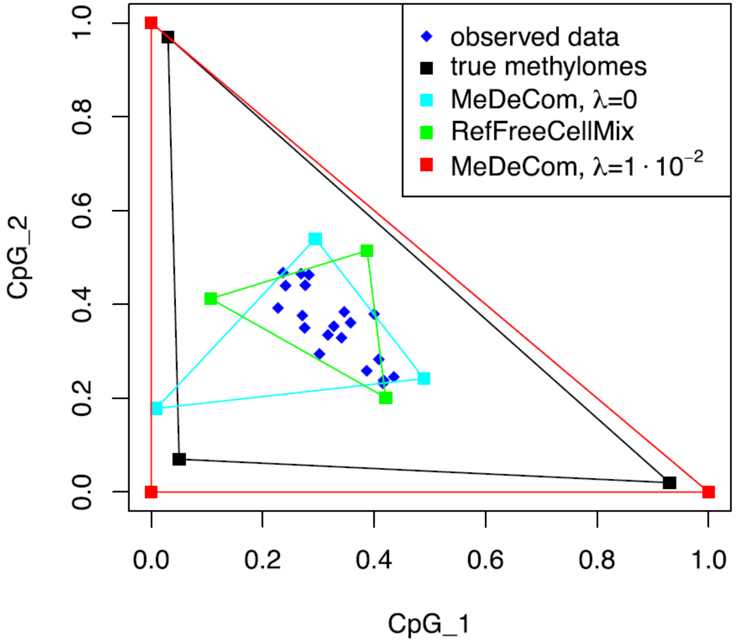
Standard NMF:

$$\begin{aligned} \min_{T,A} \|D - TA\|_F^2 &= \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n} (D_{ij} - (TA)_{ij})^2 \\ \text{subject to} \quad & 0 \le T_{is} \le 1 \ \ \forall i, s \\ & A_{sj} \ge 0 \ \ \forall s, j \\ & \textstyle\sum_{s=1}^{k} A_{sj} = 1 \ \ \forall j. \end{aligned}$$

MeDeCom's regularization:

$$\min_{T,A} \|D - TA\|_F^2 + \boxed{\lambda \sum_{i=1}^{m} \sum_{s=1}^{k} \omega(T_{is})}, \text{with}\, \omega(x) = x(1-x)$$

$$\begin{aligned} \text{subject to } 0 \le T_{is} \le 1 \ \ & \forall i, s \\ A_{sj} \ge 0 \ \ & \forall s, j \\ \sum_{s=1}^{k} A_{sj} = 1 \ \ & \forall j, \end{aligned}$$



MeDeCom Paper

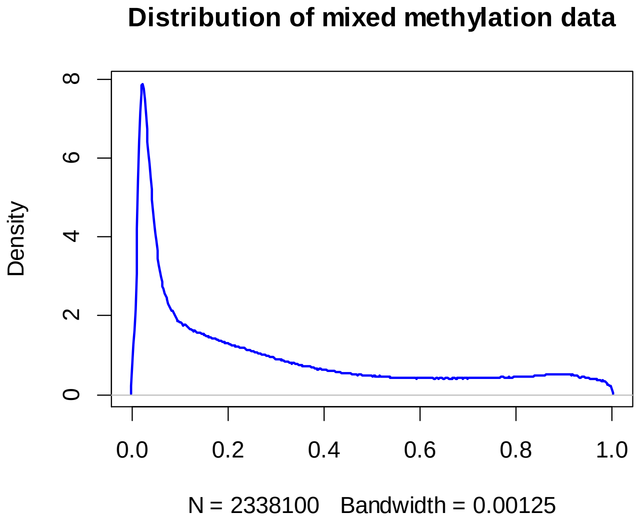
Assumptions / Requirments

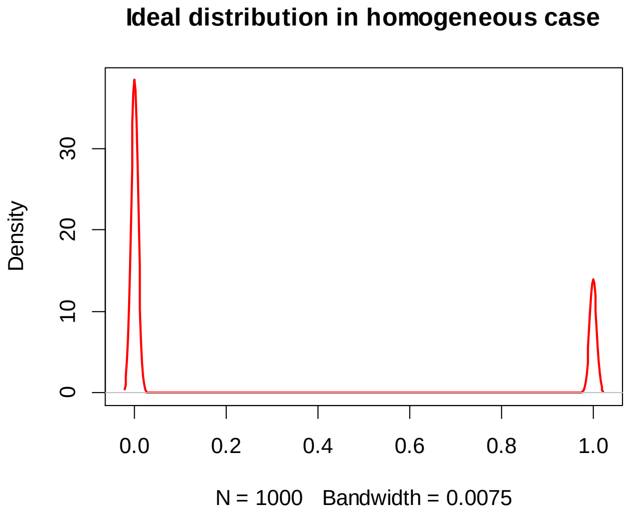
(1) Cell population consists of finite (and small) number of sub-populations.

(2) Each cell subpopulation have homogenous methylome profile => ∀CpG can be either 0 or 1.

(3) Population mixtures are variable b/w samples.

(4) Low level of technical noise and high level of biological variability.

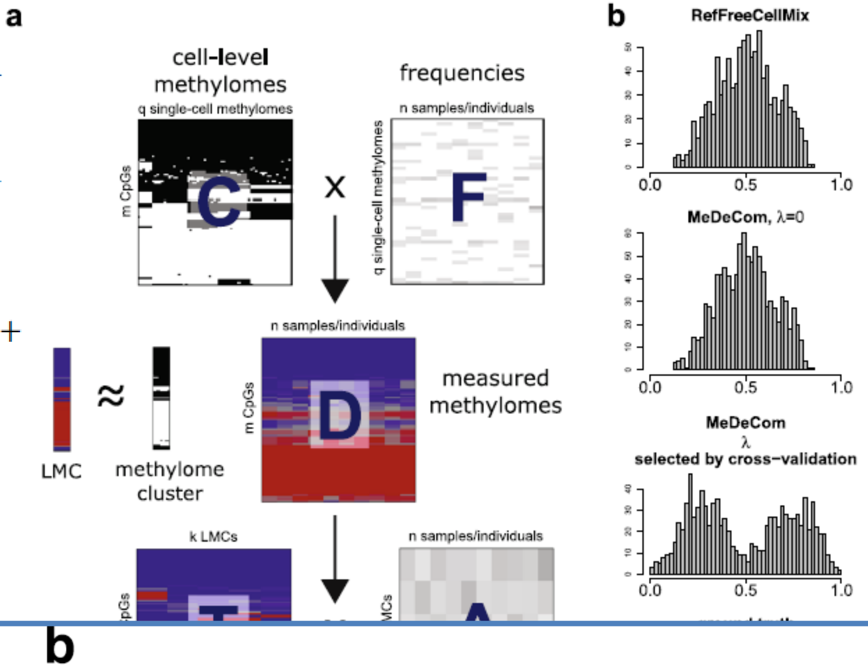


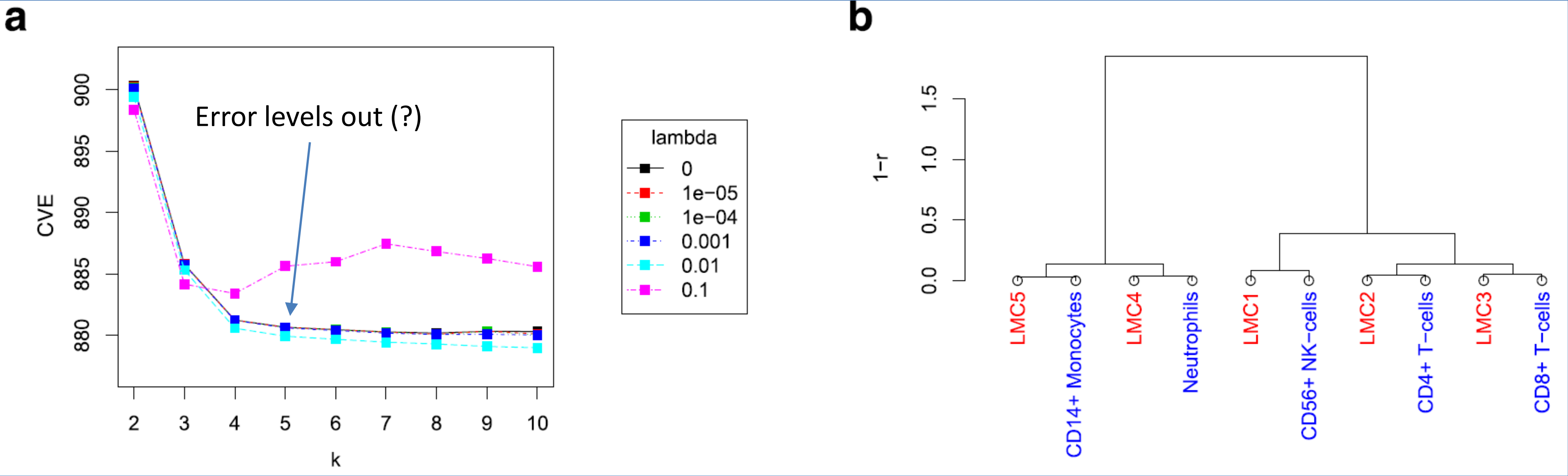


MeDeCom Paper

Synthetic data

- $k_{\text{sim}} = 2$ with two distant cell types (neutrophils and CD4+ T cells).
- $k_{\text{sim}} = 2$ with two similar cell types (neutrophils and monocytes).
- $k_{\text{sim}} = 3$ with two similar cell types and one distant from the first two (neutrophils, monocytes and CD4+ T cells).
- $k_{\text{sim}} = 5$ with all major blood cell types, excluding eosinophils and B cells.





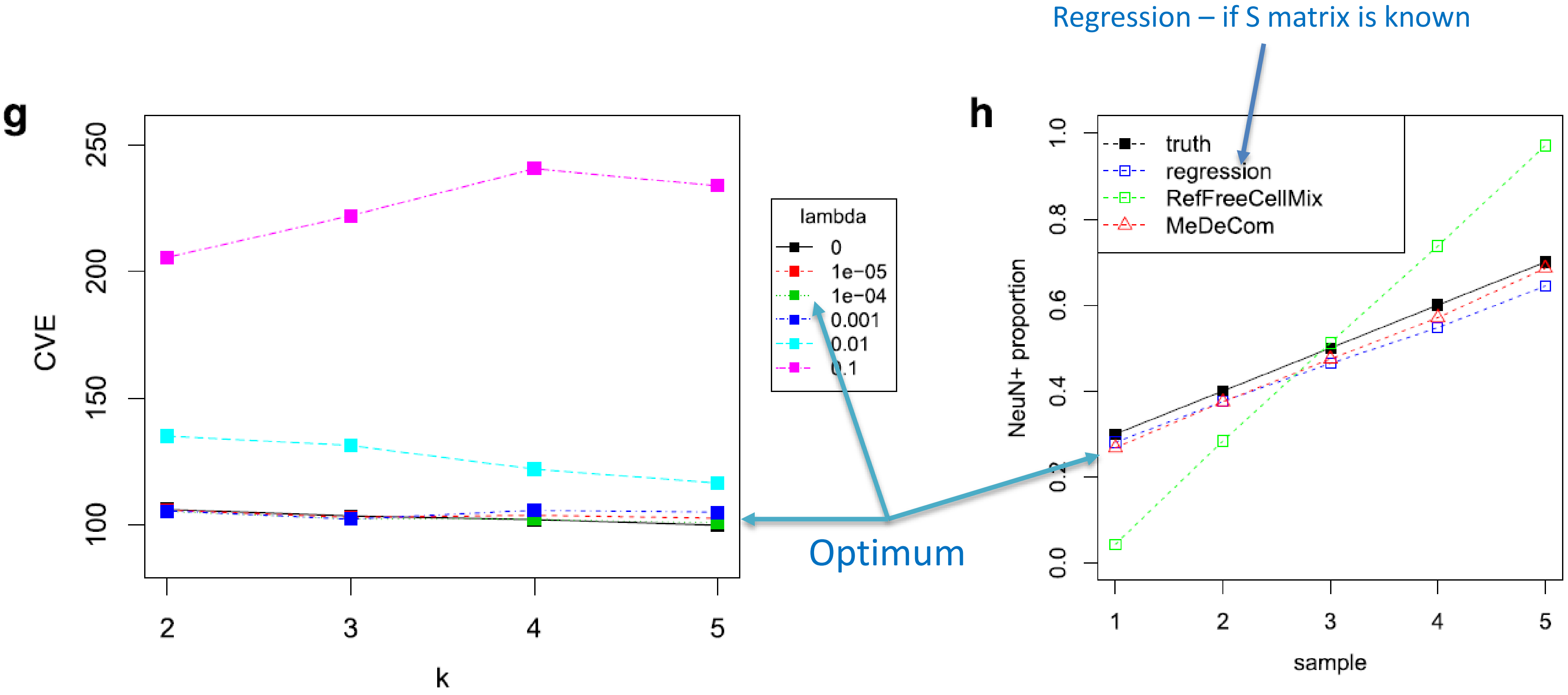
Error levels out (?)

MeDeCom Paper

Cell mixture

Dataset ArtMixN: cell sorting into NeuN+ and NeuN- cells.

NeuN = RBFOX3 protein



Regression – if S matrix is known

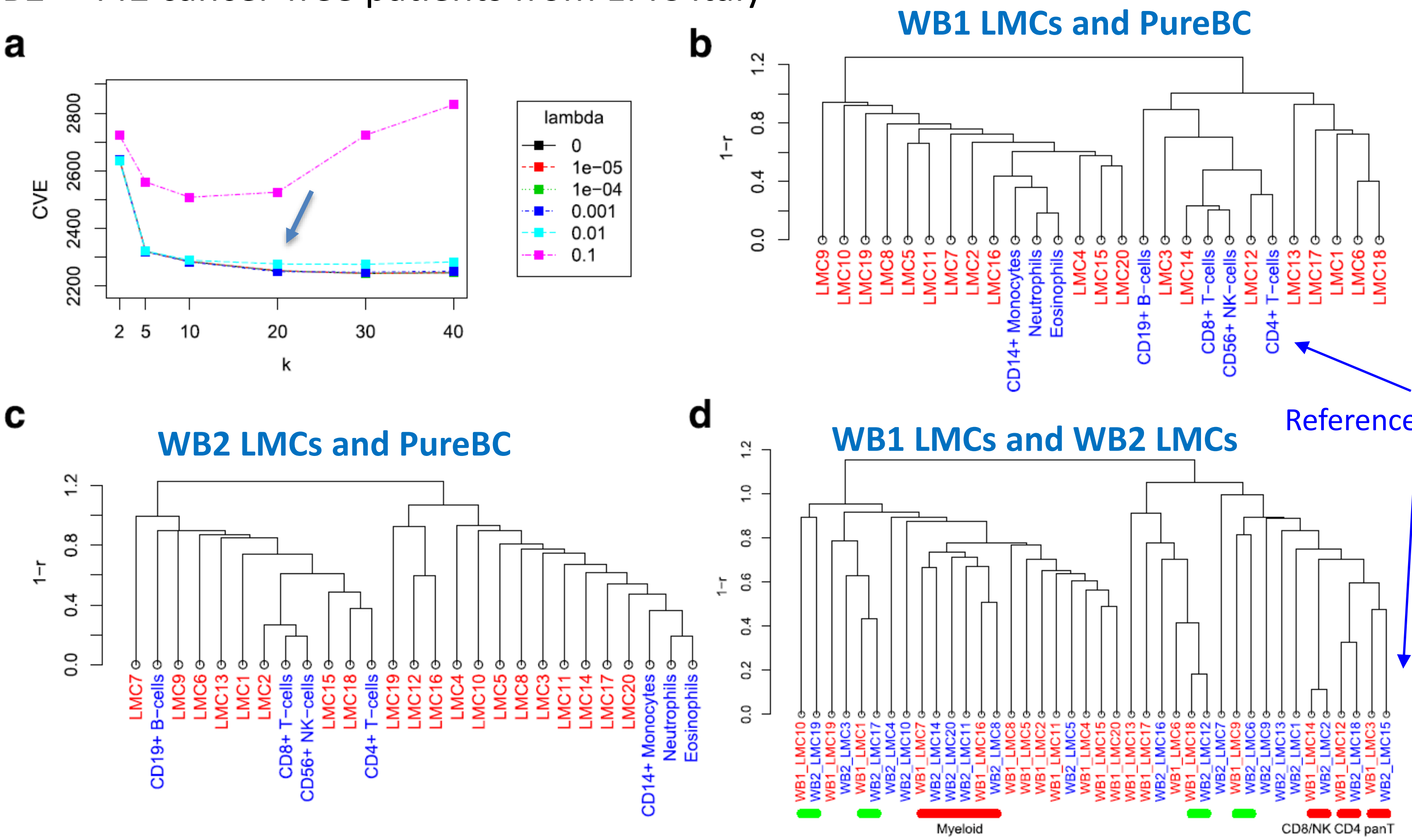
Optimum

MeDeCom Paper

Blood samples

Whole blood samples were used.

- **PureBC** – 7 MACS-purified cell types: neutro, mono, B cells ,CD4+, CD8+, NK, eosinophils
- WB1 – 87 rheumatoid arthritis patients
- WB2 – 442 cancer-free patients from EPIC Italy



WB1 LMCs and PureBC

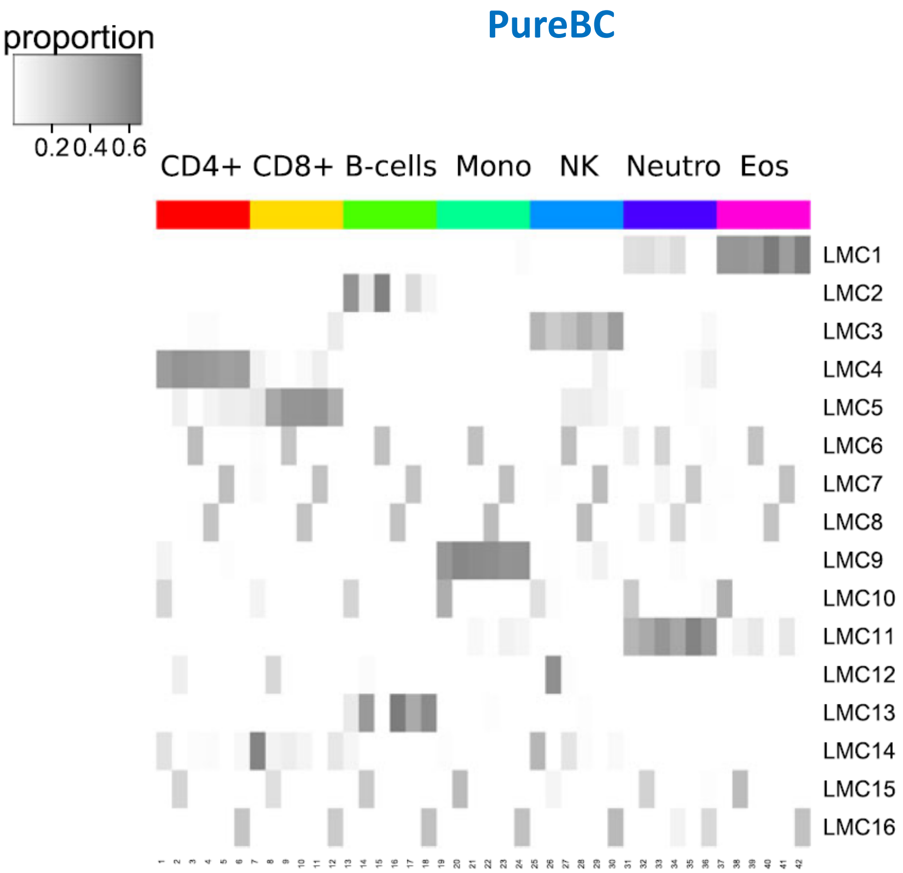
WB2 LMCs and PureBC

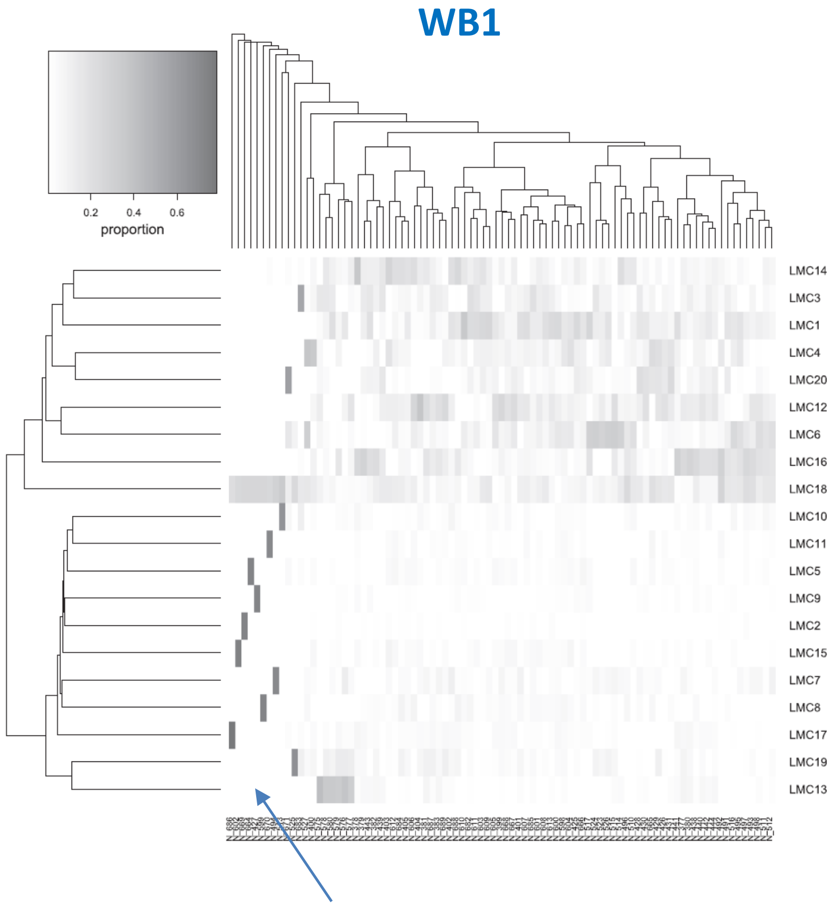
WB1 LMCs and WB2 LMCs

Reference: PureBC

MeDeCom Paper

PureBC samples





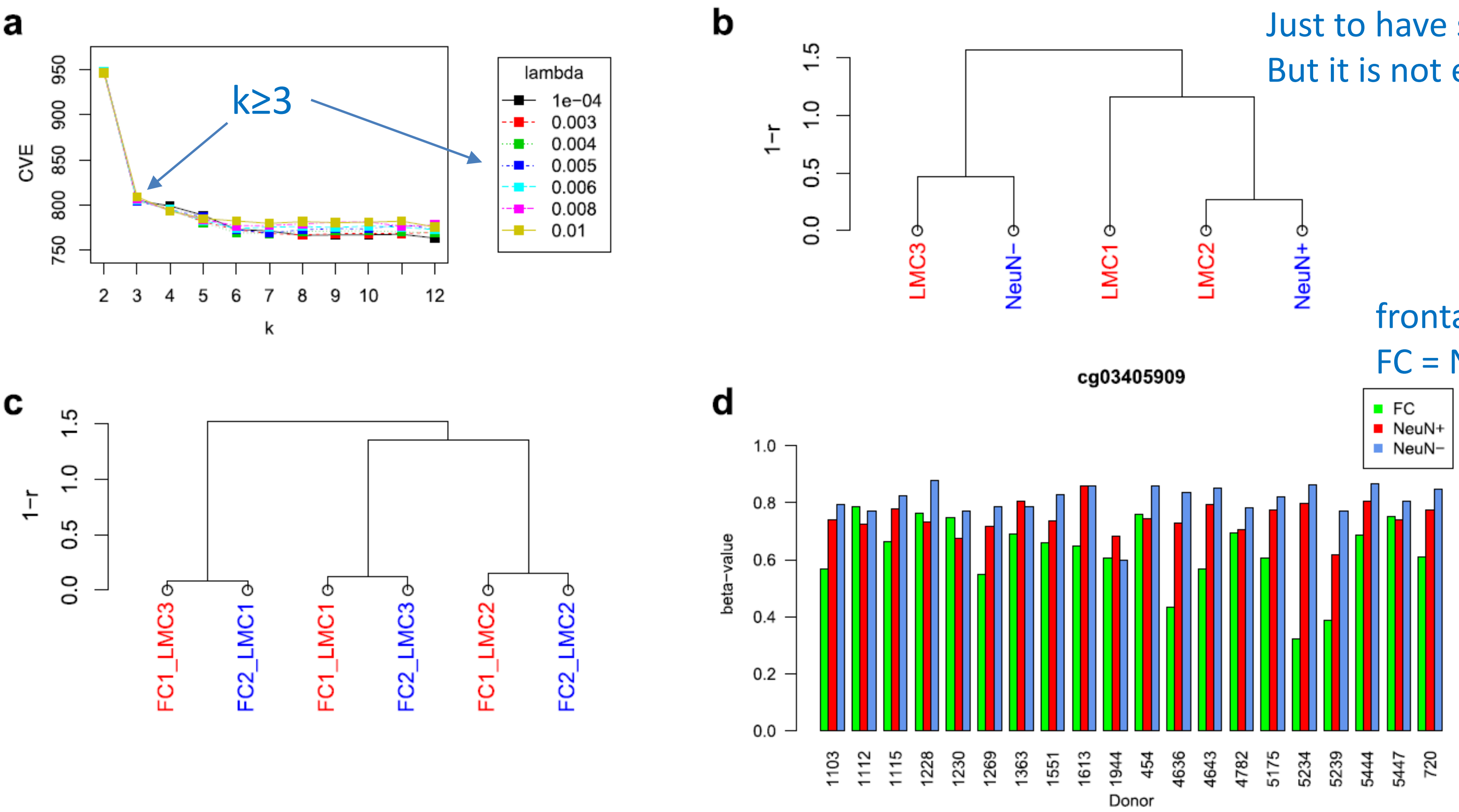
Specific to a single donor? Mutations?

MeDeCom Paper

Brain samples

Whole blood samples were used.

- **PureN** – 2x29 NeuN+/- fractions of 29 healthy controls
- FC1 – 2x10 frontal cortex of MDD (major depression disorder) patients
- FC1 – 114 frontal cortex of AD (Alzheimer’s disease) patients



k≥3

Why k=3 ?

Just to have stable result…

But it is not enough.

frontal cortex

FC = NeuN$^+$ + NeuN$^-$

MeDeCom Paper

Conclusions

MeDeCom

(1) provides significant advances compared to other methods;

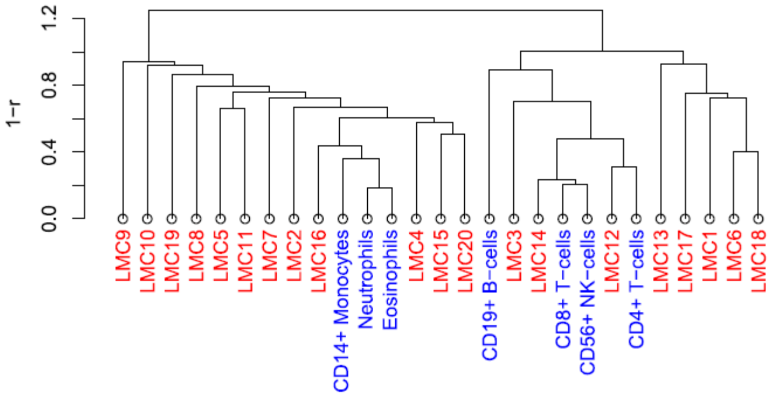
(2) uses with biologically relevant constrains and its LMCs are more interpretable;

(3) acts robustly on artificial and real data;

(4) identifies key methylation signatures;

(5) (IMHO) Low k is more dangerous than high k.

But: Separation of specific blood cell subtypes (similar methylomes) becomes challenging


1−r
LMC9
LMC10
LMC19
LMC8
LMC5
LMC11
LMC7
LMC2
LMC16
CD14+ Monocytes
Neutrophils
Eosinophils
LMC4
LMC15
LMC20
CD19+ B−cells
LMC3
LMC14
CD8+ T−cells
CD56+ NK−cells
LMC12
CD4+ T−cells
LMC13
LMC17
LMC1
LMC6
LMC18