MeDeCom discovers and quantifies latent components of heterogeneous methylomes

Pavlo Lutsik[1,*], Martin Slawski[2,*], Gilles Gasparoni[1], Matthias Hein[3] and Jörn Walter[1,#]

[1] *Department of EpiGenetics, Saarland University, Germany*
[2] *Department of Statistics & Biostatistics, Department of Computer Science, Rutgers University, USA*
[3] *Machine Learning Group, Saarland University, Germany*

*contributed equally
[#]correspondence: j.walter@mx.uni-saarland.de

Abstract

Large-scale DNA methylation studies on blood or tissue samples are confronted with sample-specific confounding factors, such as heterogeneous cell composition and individual genetic variation. Aiming to address these issues, the large-scale profiling efforts, one of which is IHEC, are generating a range of high-quality reference DNA methylomes. However, the limitations of the mainstream experimental methods and complex experimental setups call for an unbiased discovery and exploration of the confounding sources.

We developed MeDeCom, a reference-free computational framework that uses non-negative matrix factorization to decompose complex DNA methylation profiles of cell mixtures. MeDeCom generates interpretable latent methylation components (LMCs) and provides estimates of LMC proportions in each sample. LMCs enable biological exploration by correlating them to reference epigenomes or other annotated cell type-specific molecular signatures. The estimated proportions can help to interpret sample-specific variation as well as disease or age-related phenotypes. MeDeCom can be applied to any WGBS or Infinium 450k/EPIC array data set that has a sufficient technical quality.

We demonstrate the performance of MeDeCom on artificial cell mixtures and complex biological data sets of whole blood, partially purified blood cell populations and the cortical brain tissue. We show that our framework facilitates a deeper understanding of genome-scale DNA methylation data and can be used to enhance the reference methylomes.