



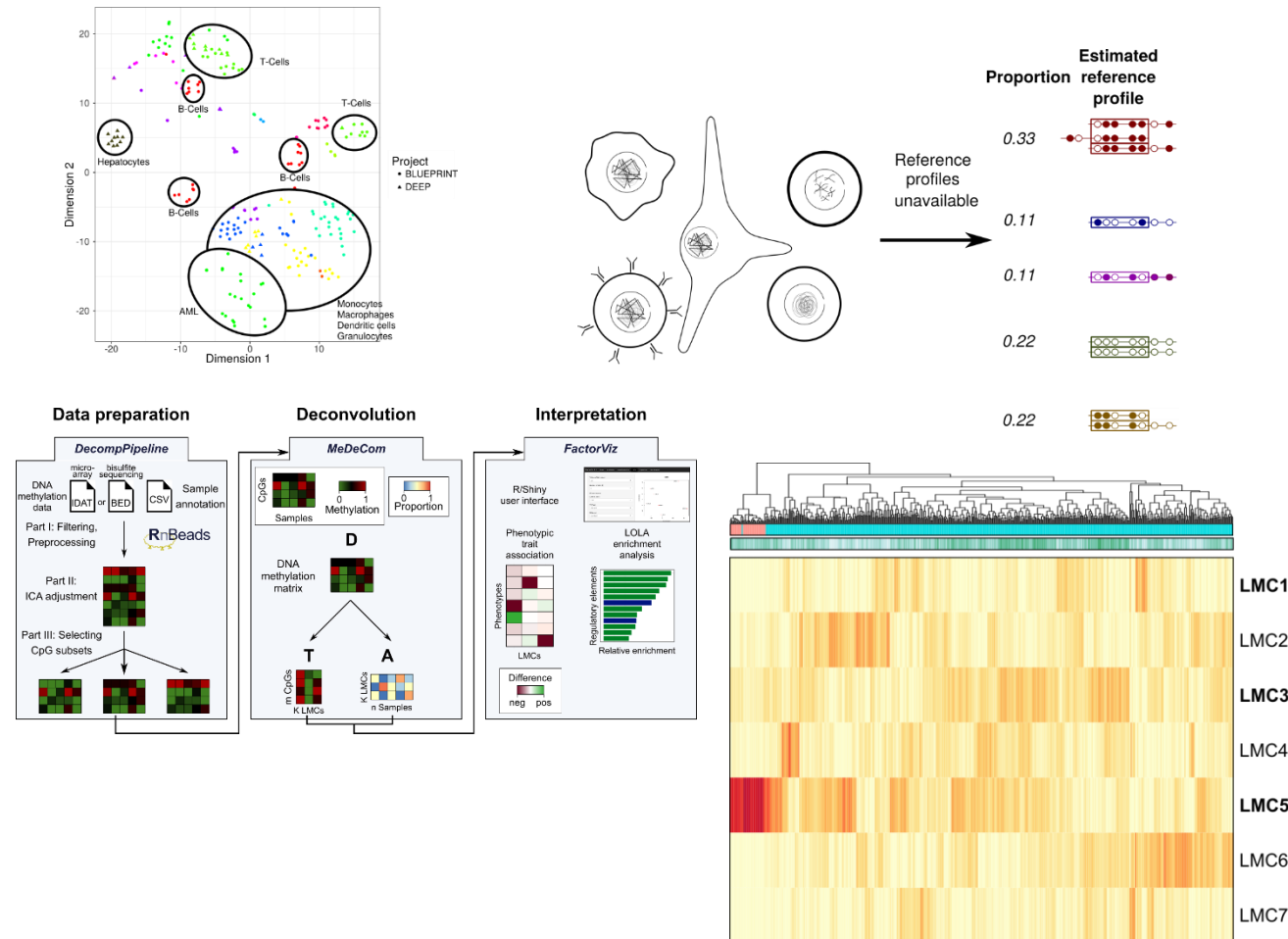
Saarland University
Department of
Genetics/Epigenetics

Reference-free deconvolution of complex DNA methylation data - a systematic protocol

Michael Scherer
HADACA, Aussois
11/26/2019

Overview

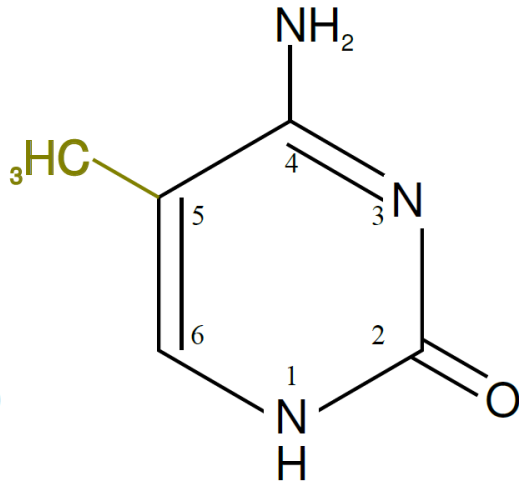
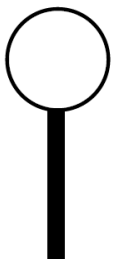
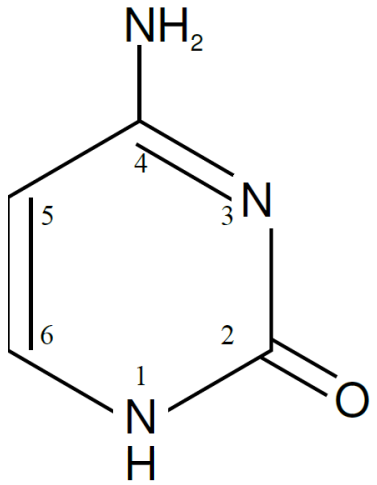
- Introduction into DNA methylation
- DNA methylation-based deconvolution
- Systematic protocol for DNA-methylation based deconvolution using *MeDeCom*
- Application of the proposed protocol on TCGA data
- Conclusions



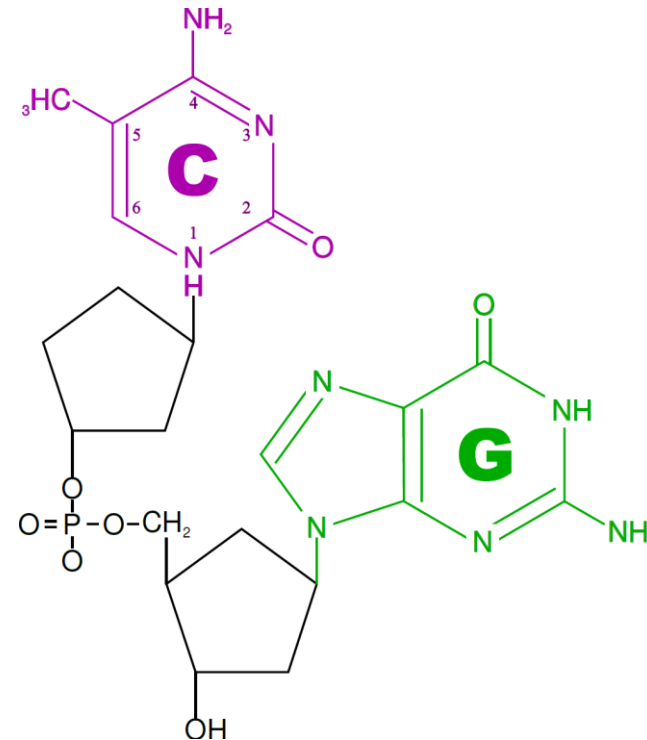
DNA methylation

Cytosine

5-Methyl-
Cytosine



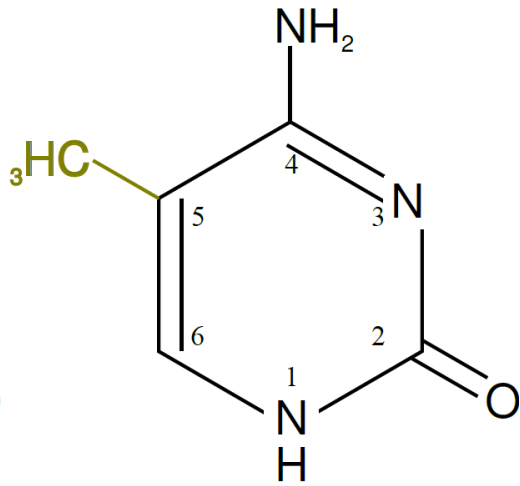
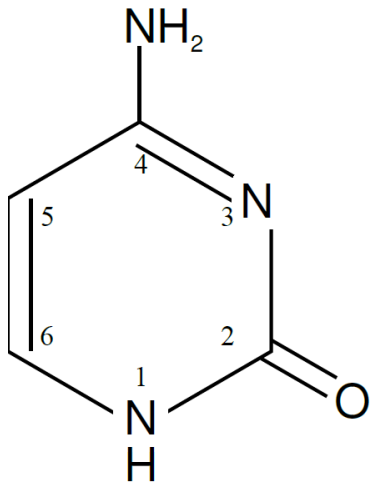
- Reversible epigenetic modification
- Almost exclusively in CpG context



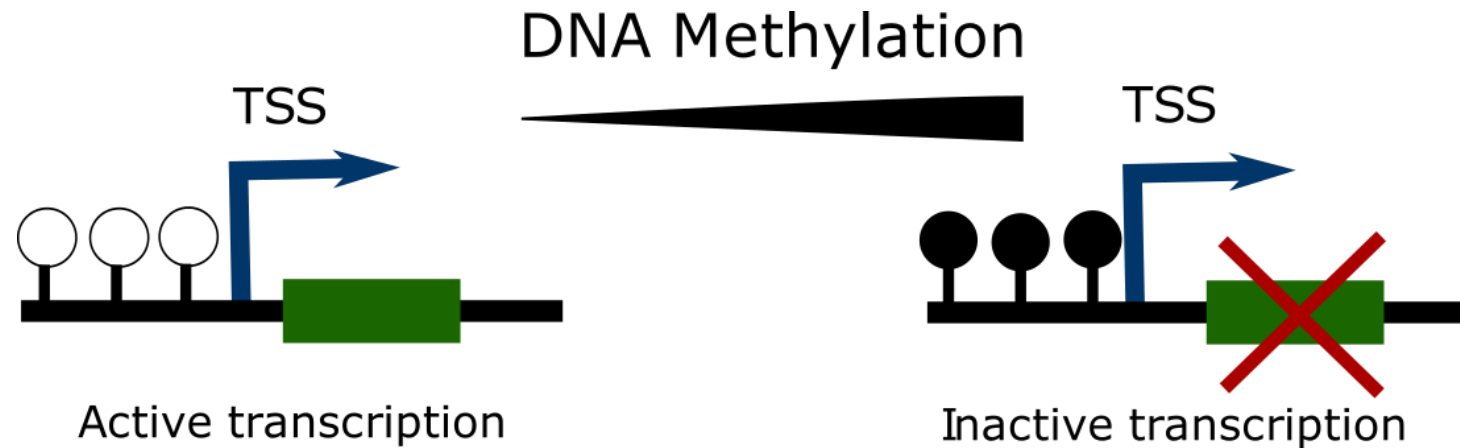
DNA methylation

Cytosine

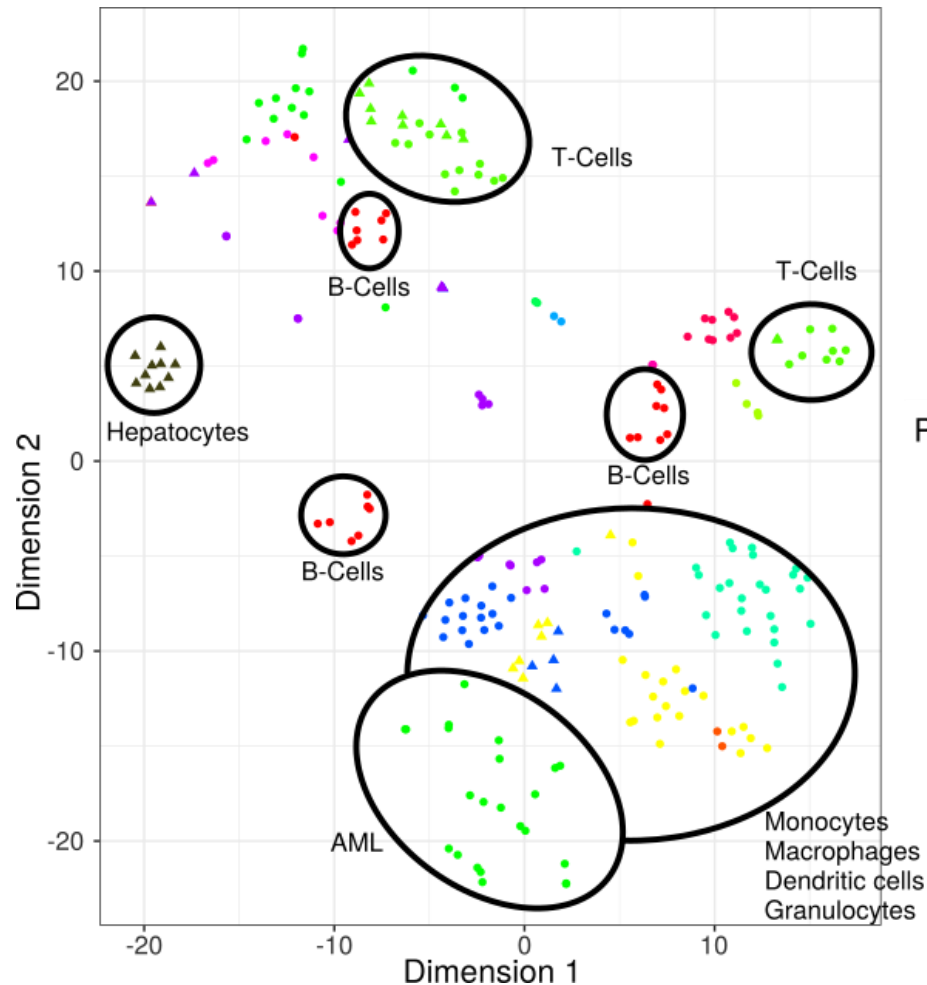
5-Methyl-
Cytosine



- Reversible epigenetic modification
- Almost exclusively in CpG context
- Transcriptional repression in promoter regions



DNA methylation



- Reversible epigenetic modification
- Almost exclusively in CpG context
- Transcriptional repression in promoter regions
- Highly cell type specific

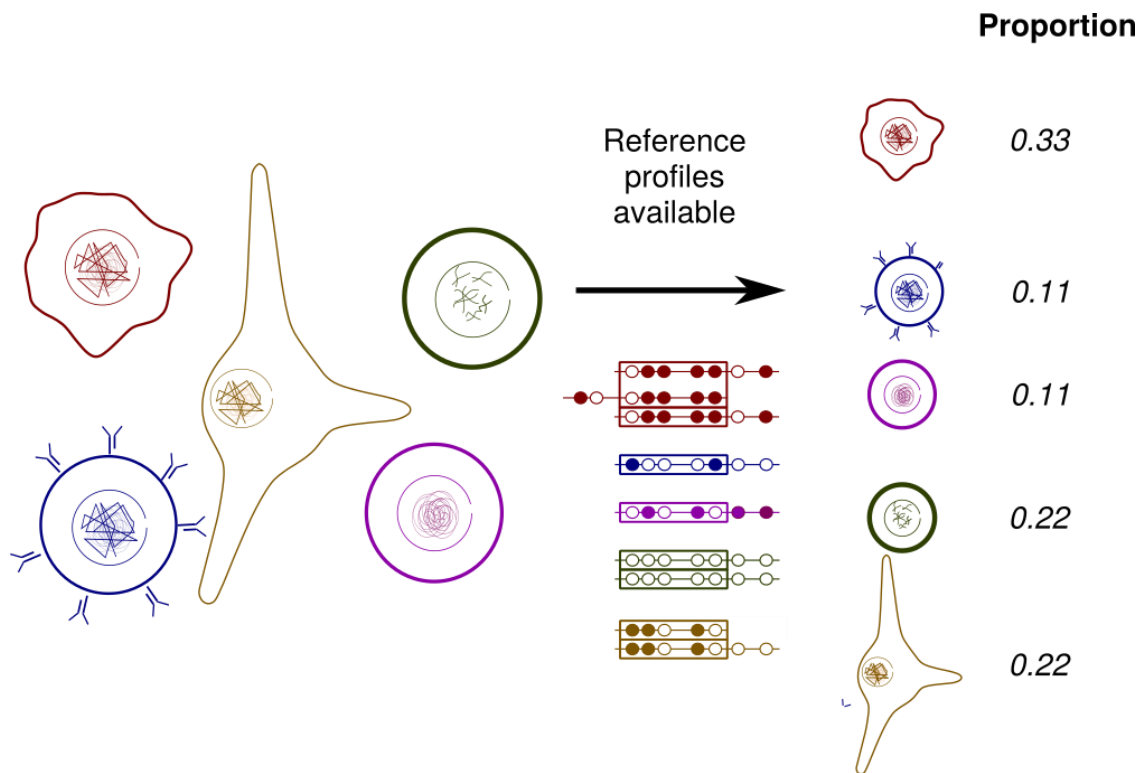
Figure: tSNE plot of WGBS data from different cell types assayed in the DEEP¹ and BLUEPRINT² consortia

¹ <http://www.deutsches-epigenom-programm.de/>

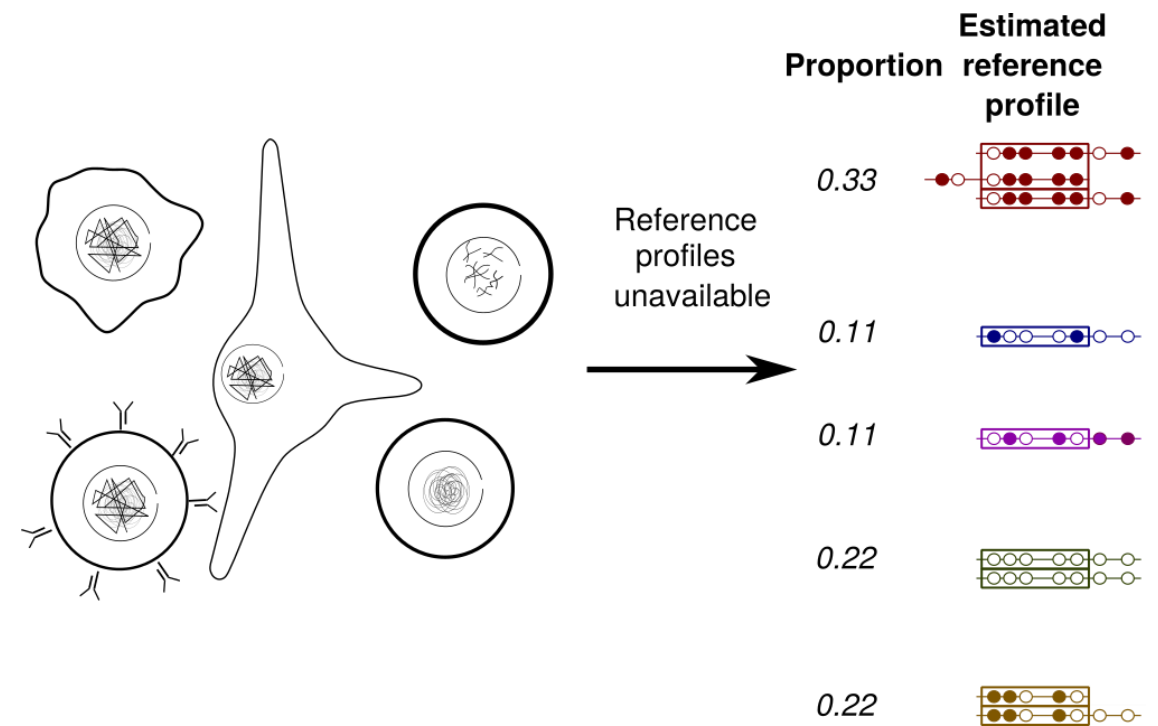
² <http://www.blueprint-epigenome.eu/>

DNA methylation based deconvolution

Reference-based deconvolution



Reference-free deconvolution



DNA methylation based deconvolution

Reference-based deconvolution

- Houseman approach¹
- MethylCIBERSORT²
- EpiDISH³

¹ Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, (2012).

² Chakravarthy, A. *et al.* Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.* **9**, (2018).

³ Teschendorff, A. E *et al.* A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).

Reference-free deconvolution

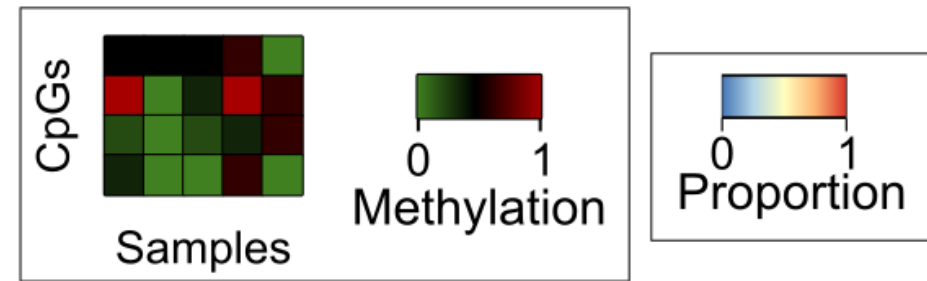
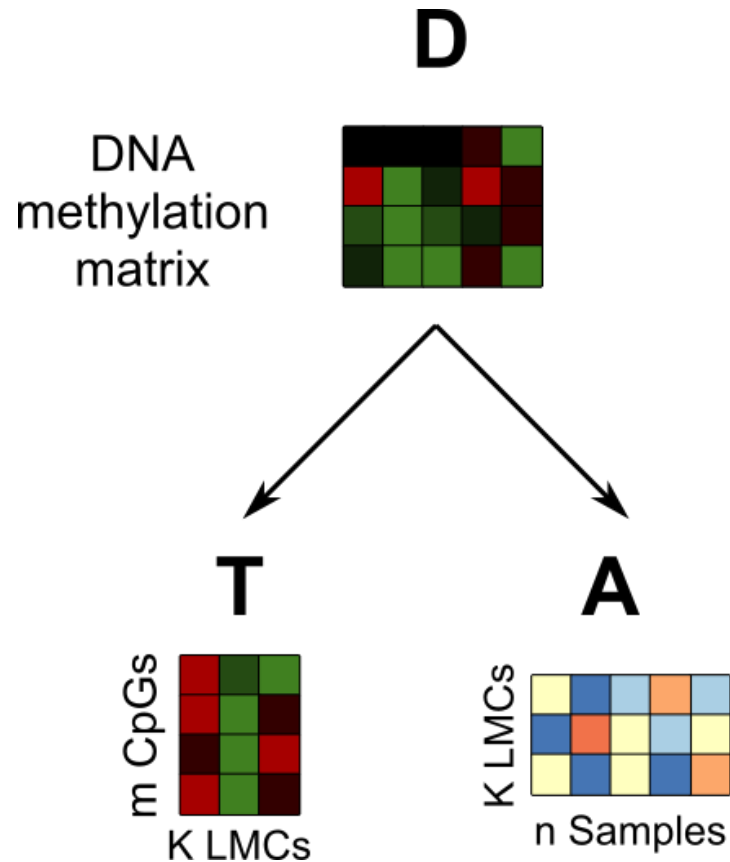
- RefFreeCellMix⁴
- EDec⁵
- MeDeCom⁶

¹ Houseman, E. A. *et al.* Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431-1439 (2014).

² Onuchic, V. *et al.* Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Rep.* **17**, 2075-2086 (2016).

³ Lutsik, P. *et al.* MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 55 (2017).

Non-negative matrix factorization



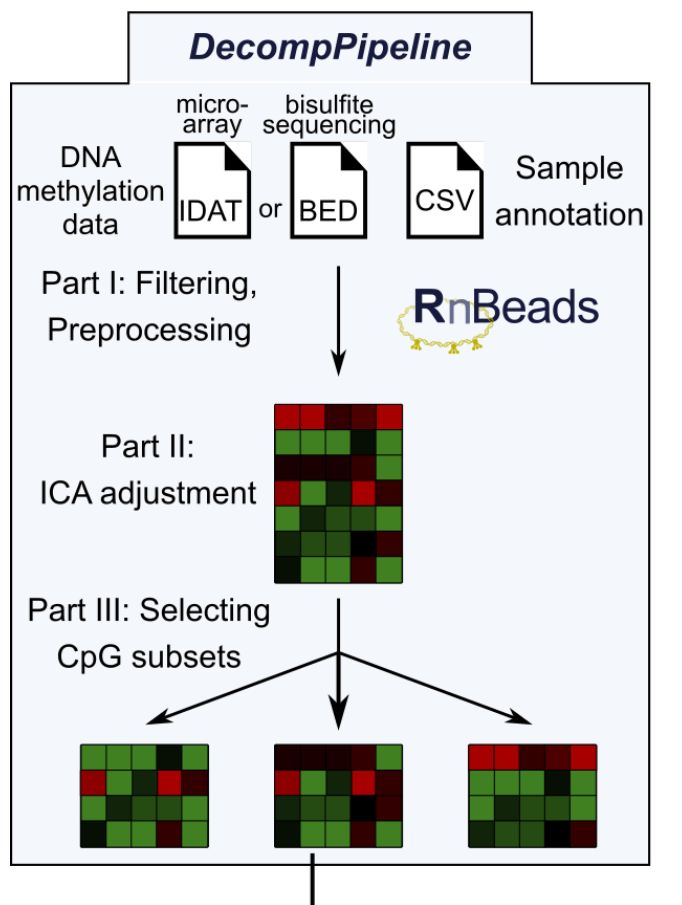
Key messages from HADACA 2018

- Only small performance differences between the three available reference-free deconvolution tools (*RefFreeCellMix*, *EDec*, *MeDeCom*) on *in-silico* mixed data
- Thorough data processing more important than choice of the deconvolution tool
- Accounting for confounding factors critical for obtaining biologically plausible results¹

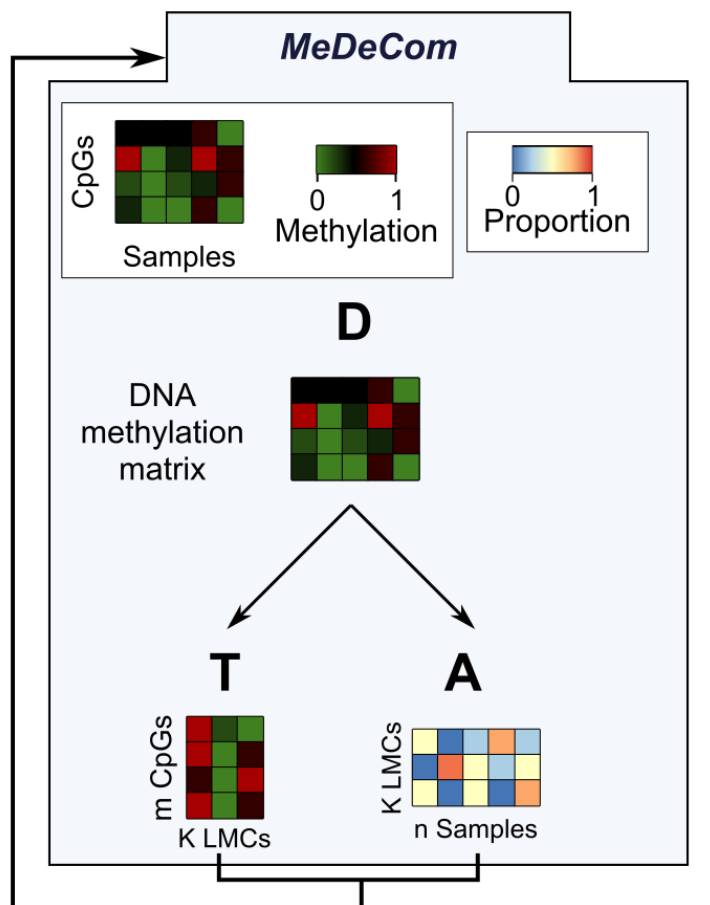
¹ Decamps, C. *et al.* Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. Preprint at <https://www.biorxiv.org/content/10.1101/698050v1.abstract> (2019).

Systematic protocol for DNA methylation based deconvolution

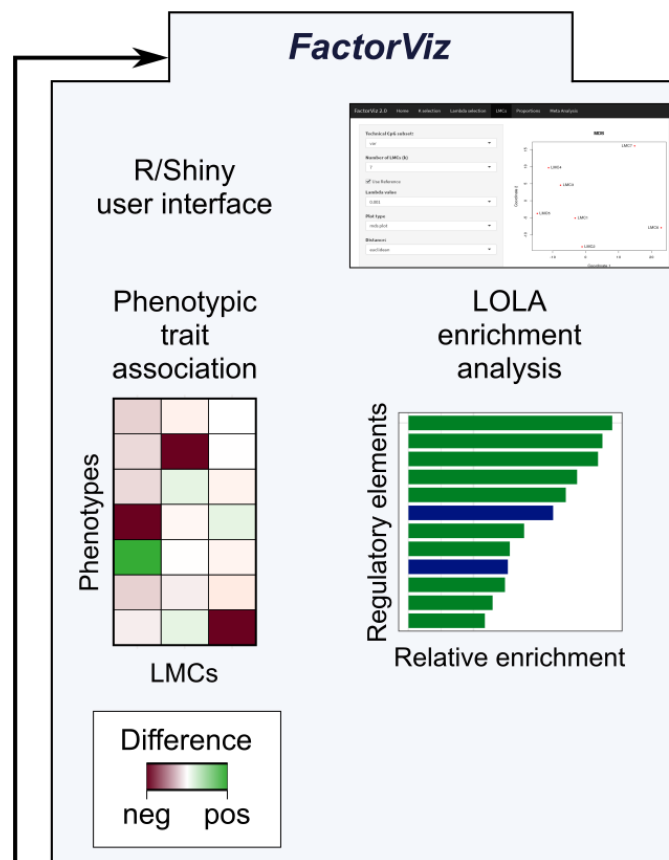
Data preparation



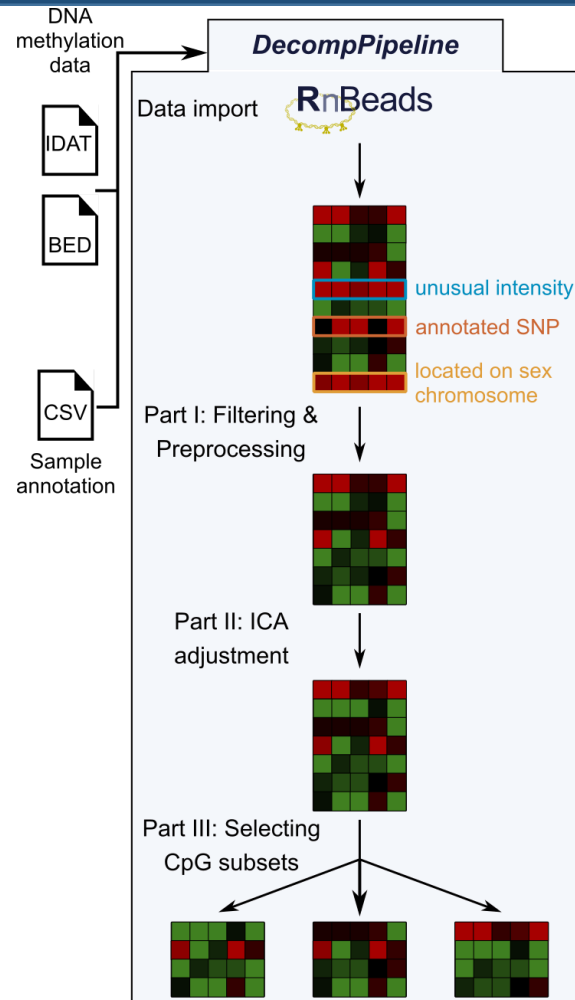
Deconvolution



Interpretation



DecompPipeline¹



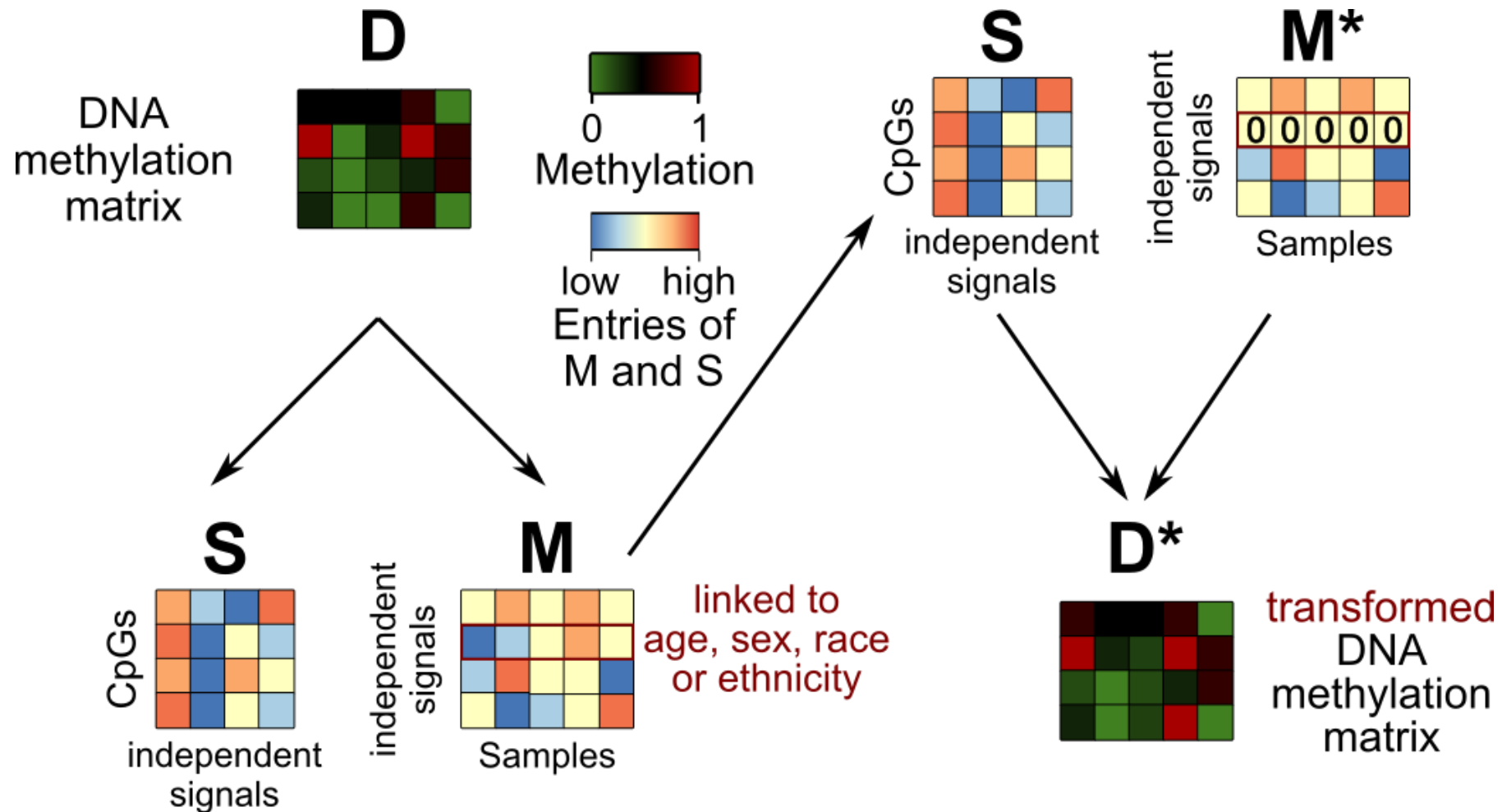
- Data import using the widely-used *RnBeads*² software package
- Three-step procedure
 - Quality-aware filtering
 - Accounting for confounding factors using independent component analysis (ICA³)
 - Selecting potentially informative CpGs

¹ <https://github.com/lutsik/DecompPipeline>

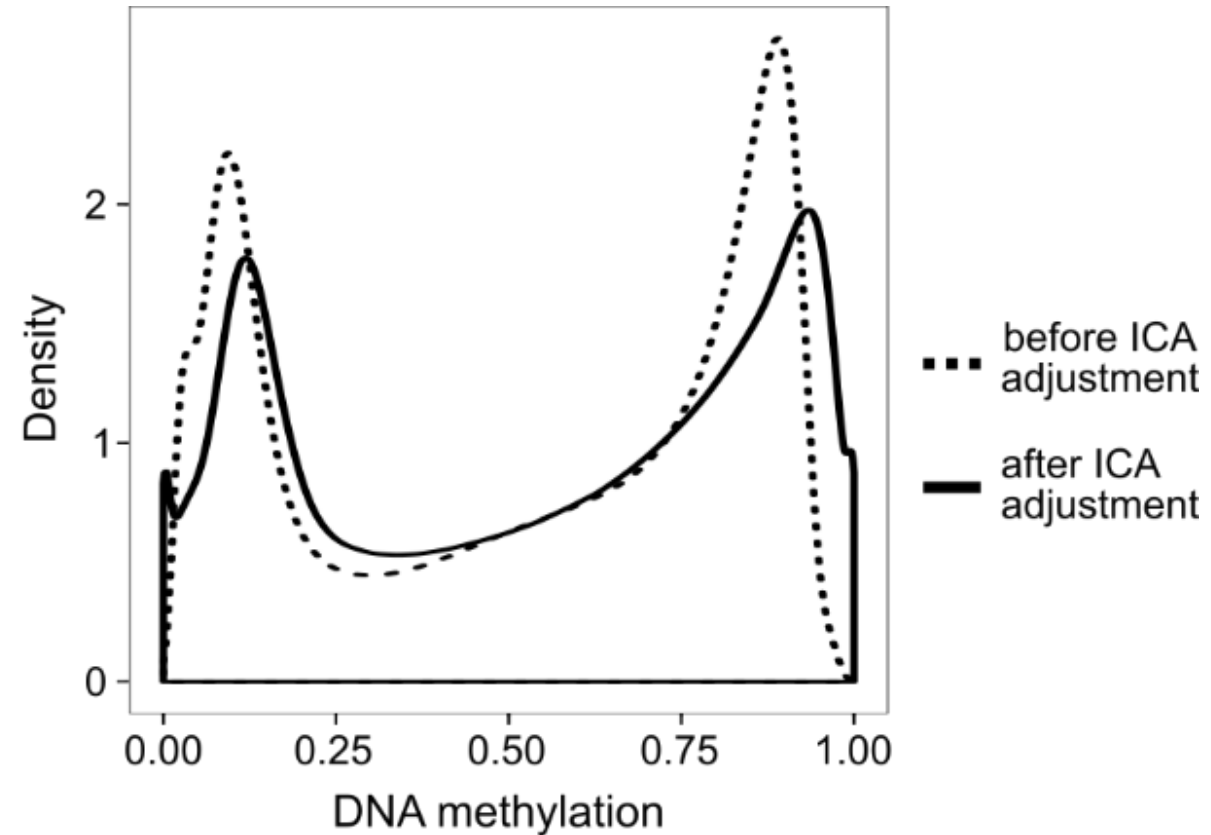
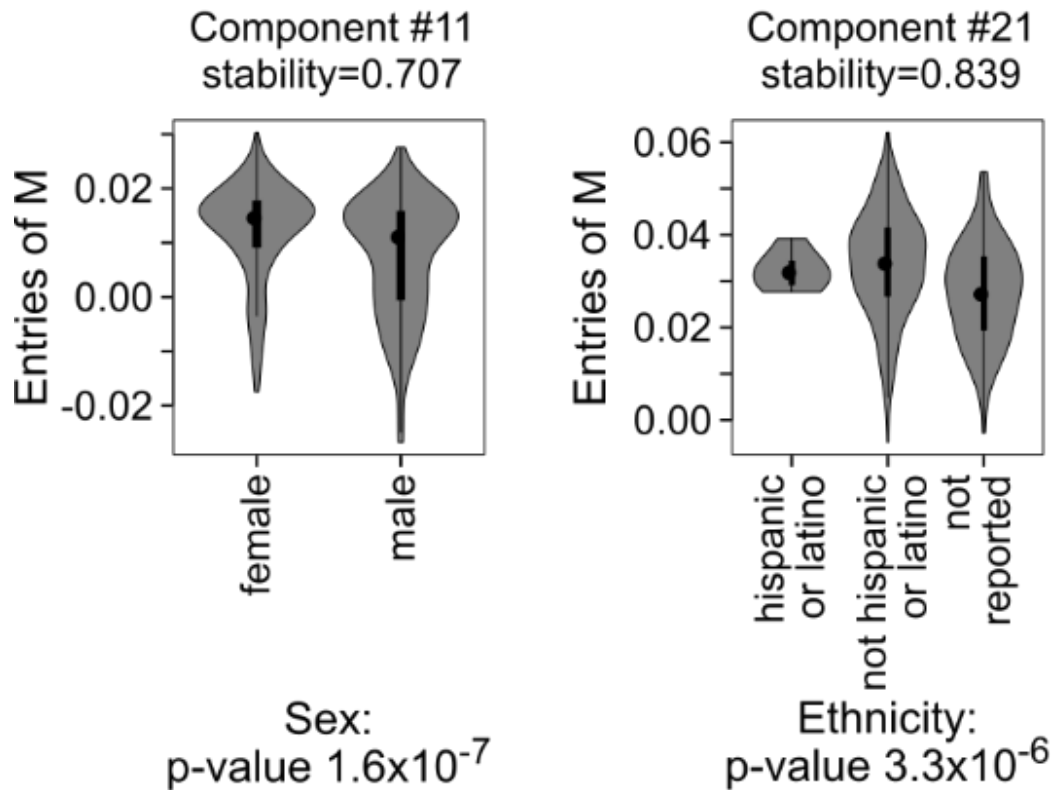
² Müller, F. *et al.* RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 20, 55 (2019).

³ Nazarov, P. V *et al.* Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. *BMC Med. Genomics* 12, 132 (2019).

Confounding factor adjustment using ICA

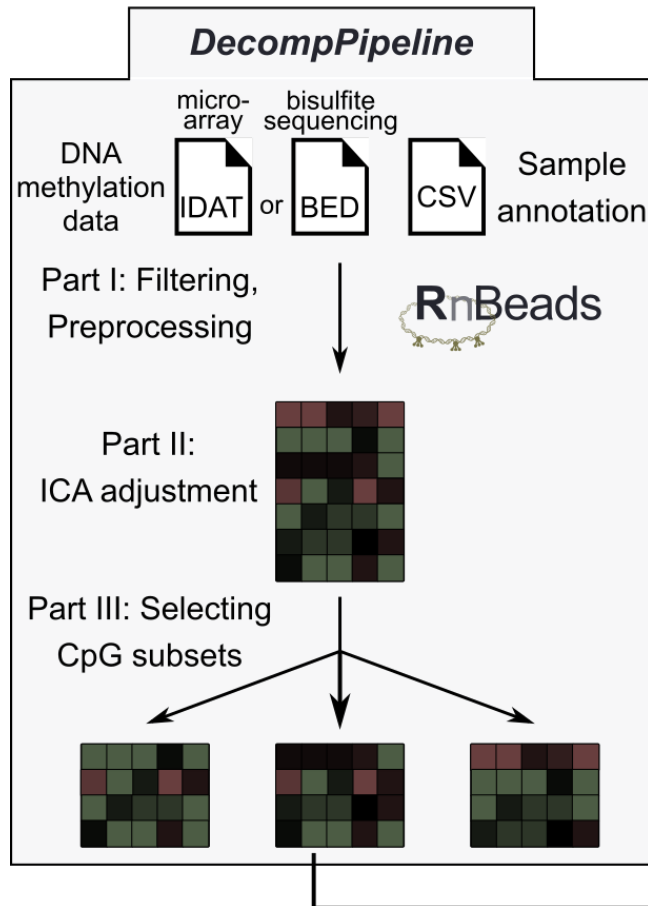


Confounding factor adjustment using ICA

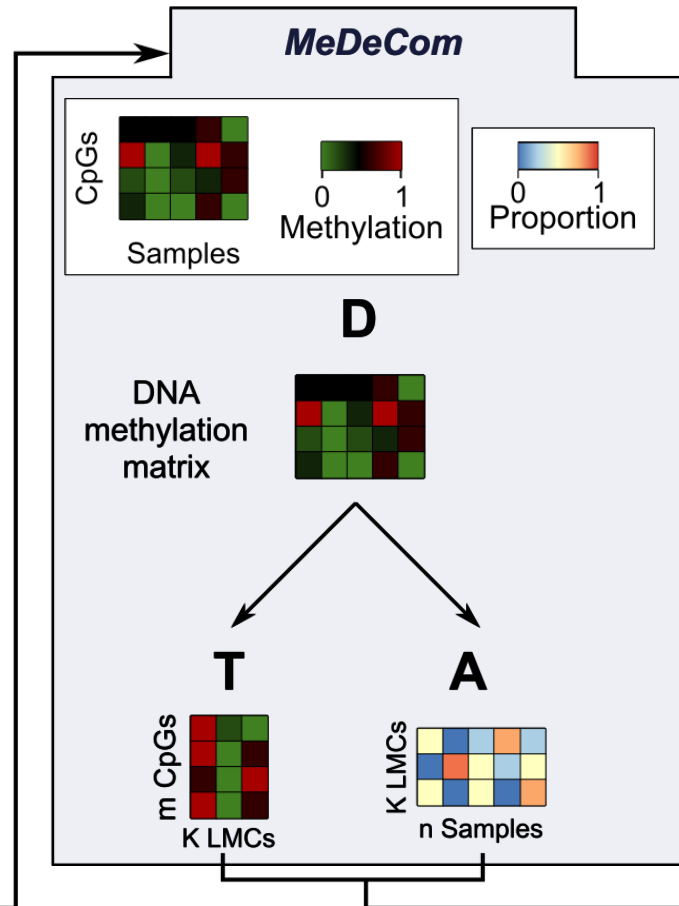


Protocol overview

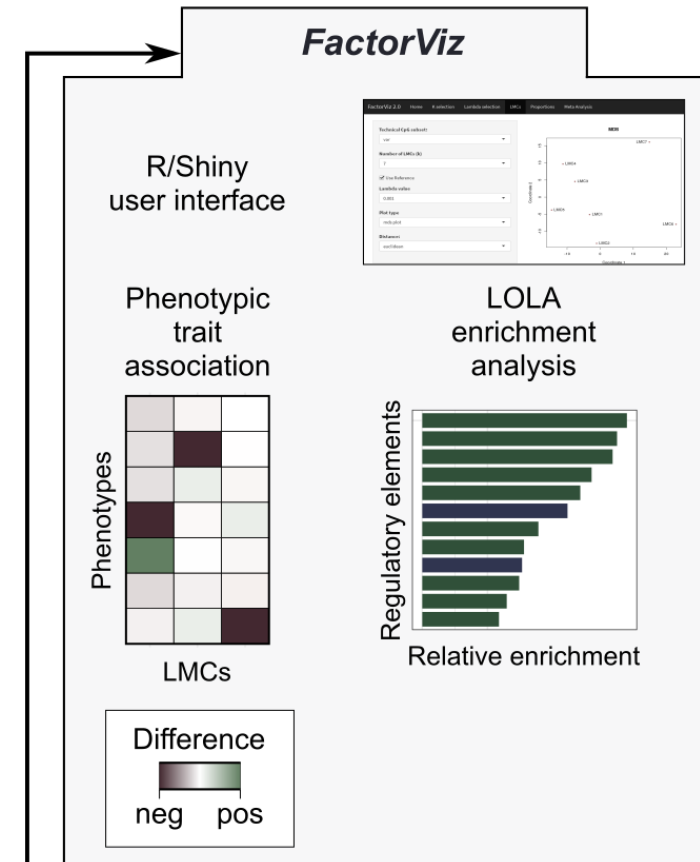
Data preparation



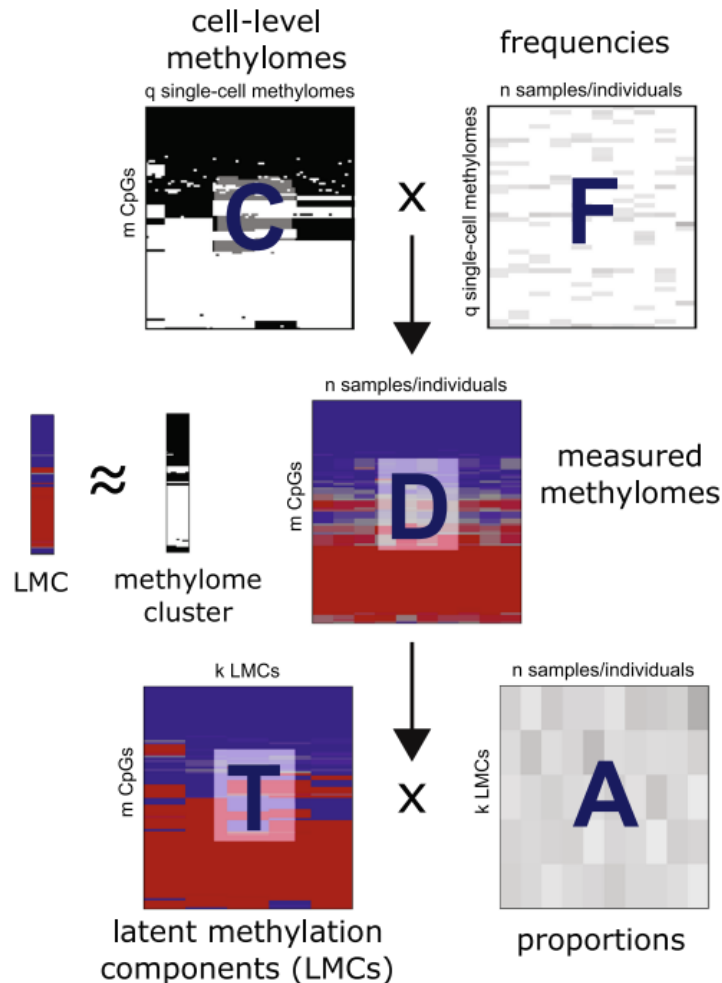
Deconvolution



Interpretation



MeDeCom¹



- Regularized non-negative matrix factorization
- Critical parameter choices:
 - Number of latent methylation components (LMCs, K)
 - Regularization parameter (λ)
- Optimized using an alternate optimization scheme
- Cross validation error computed

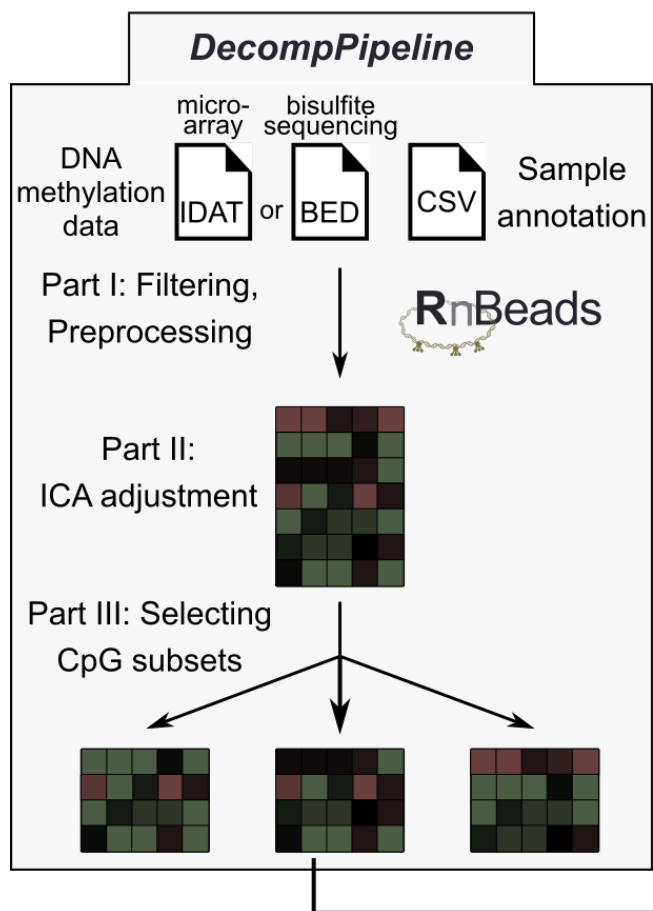
¹ Lutsik, P. *et al.* MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 55 (2017).

RefFreeCellMix and *EDec*

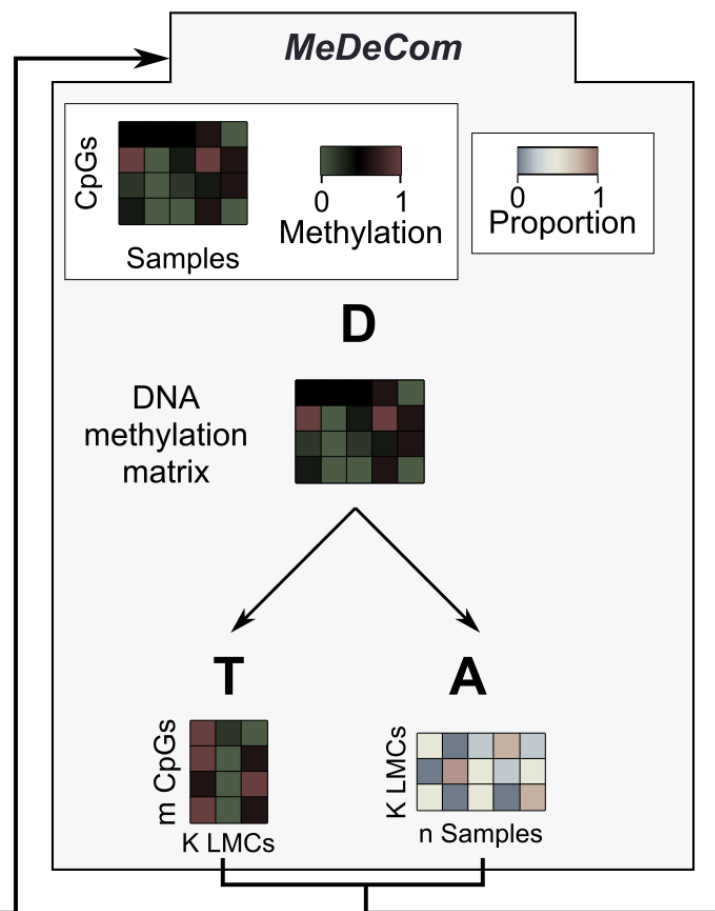
- Similar approaches as *MeDeCom*
- Seamless integration into the protocol

Protocol overview

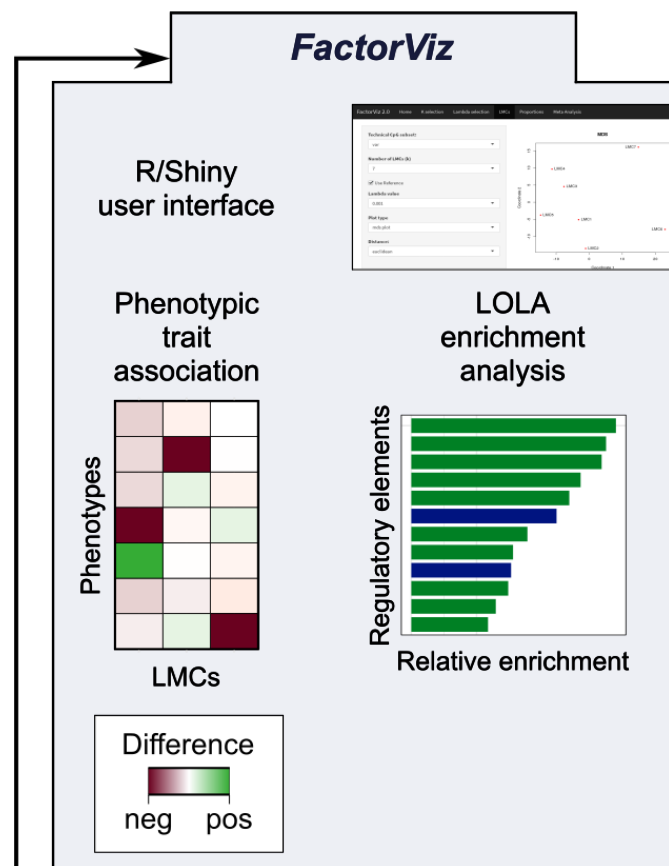
Data preparation



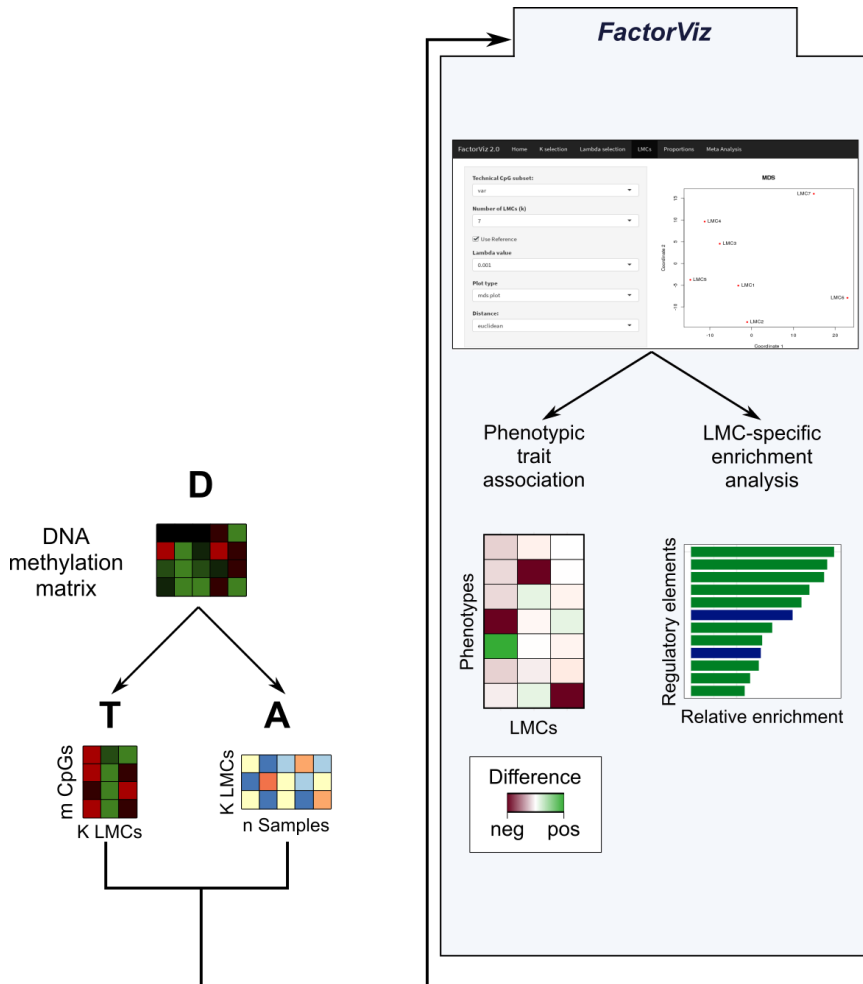
Deconvolution



Interpretation



FactorViz¹ overview



- R/Shiny application to visualize deconvolution results
- Evaluation and interpretation functions
- Proportions and LMC matrix biologically interpreted

¹ <https://github.com/lutsik/FactorViz>

FactorViz: Interface

FactorViz 2.0

Home

Choose Directory

OR

Path

Note:

If both path (as text input) and directory
(chosen via the file manager) is provided only
the path will be considered

☐ Non DeComp-Pipeline Input

Load Datasets

Files in the directory

```
[1] "ann_C.RData"      "ann_S.RData"      "medecom_set.RData"  
[4] "meth_data.RData"
```

FactorViz: Functions

FactorViz 2.0

Home

K selection

Lambda selection

LMCs

Proportions

Meta Analysis

Technical CpG subset:

var

Number of LMCs (k)

7

☒ Use Reference

Lambda value

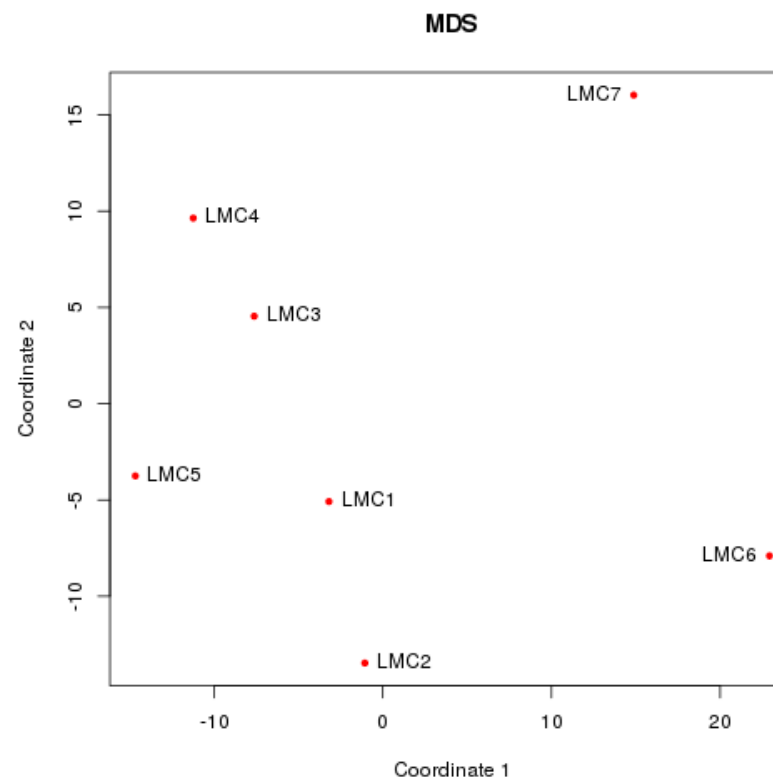
0.001

Plot type

mds plot

Distance:

euclidean



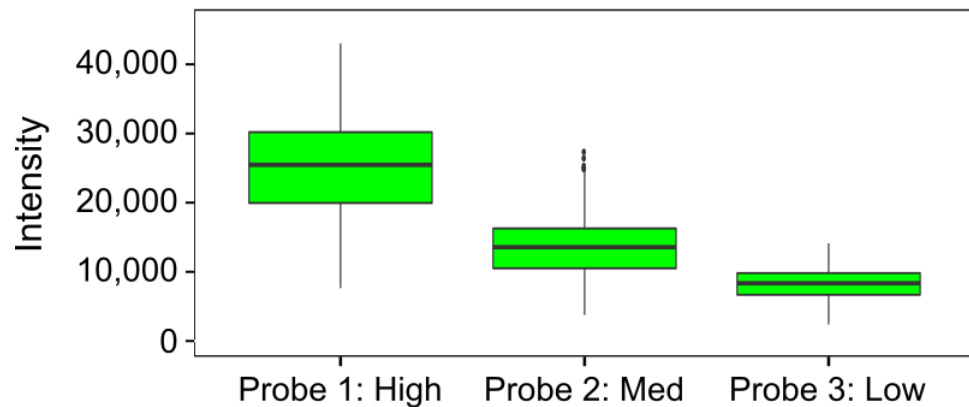
Application to TCGA LUAD dataset

- 461 samples from the lung adenocarcinoma dataset from TCGA¹
- Assayed using the Illumina Infinium 450k BeadChip

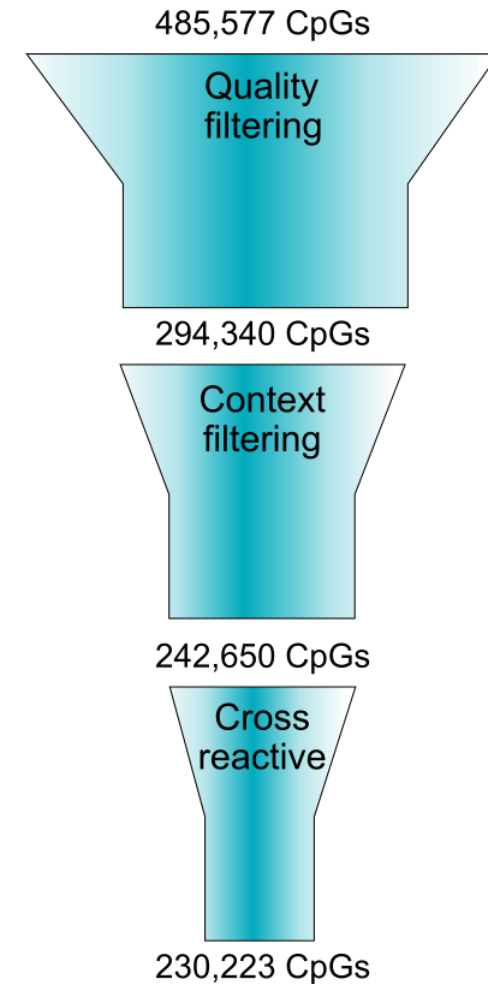
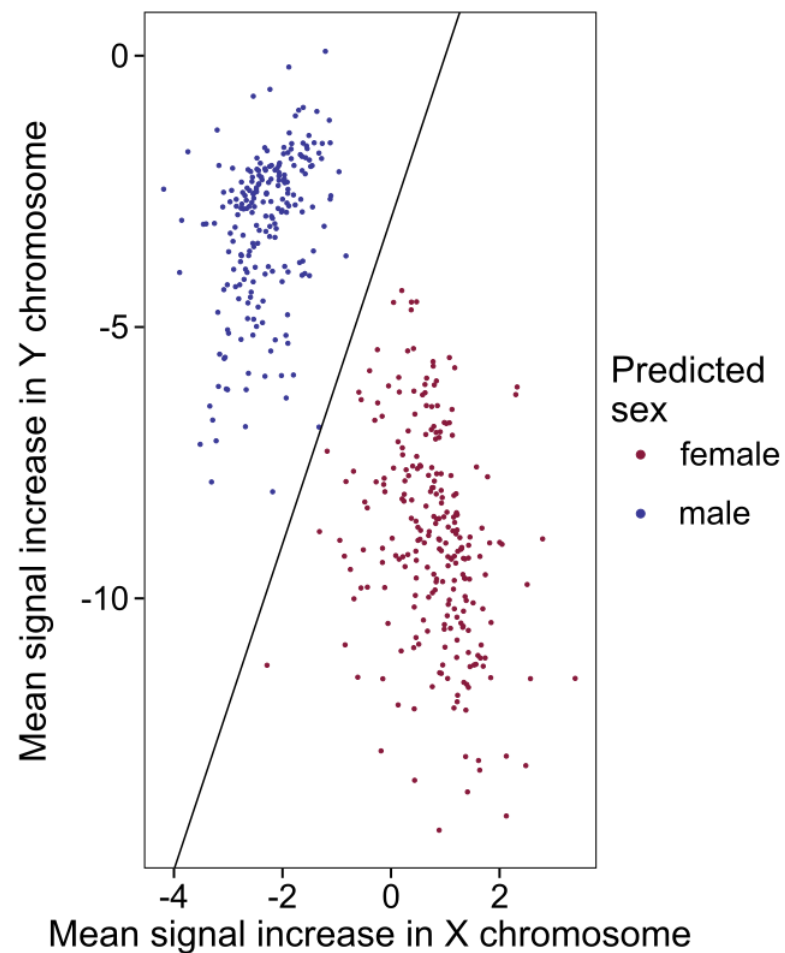
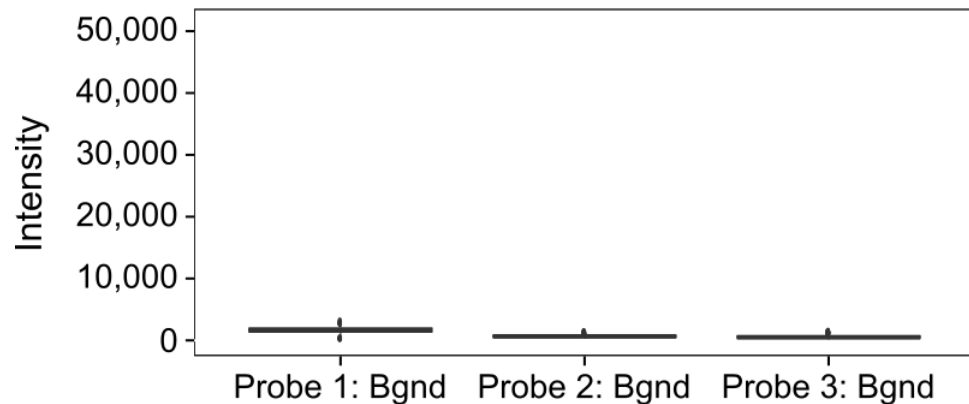
¹ <https://cancergenome.nih.gov/>

QC on TCGA data

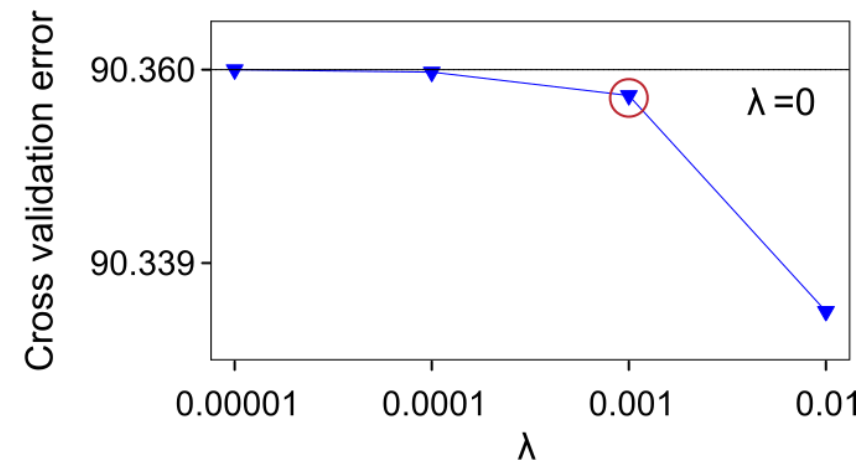
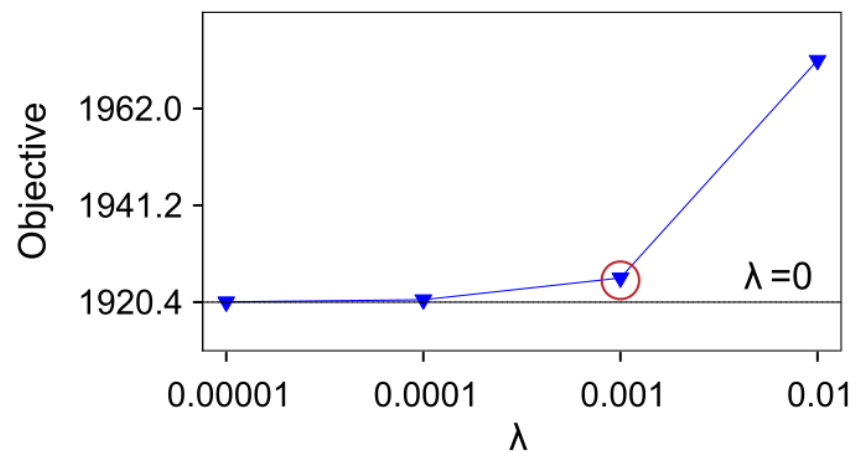
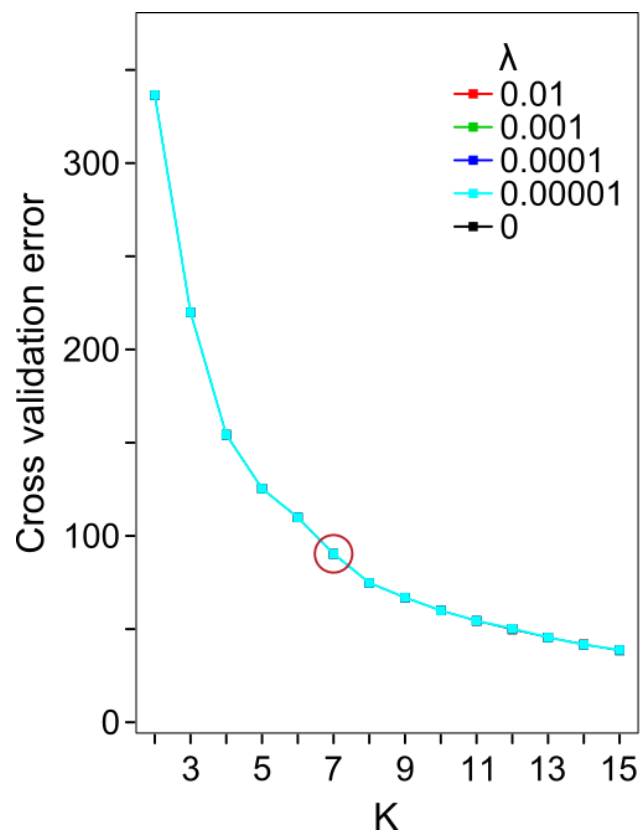
HYBRIDIZATION: green channel



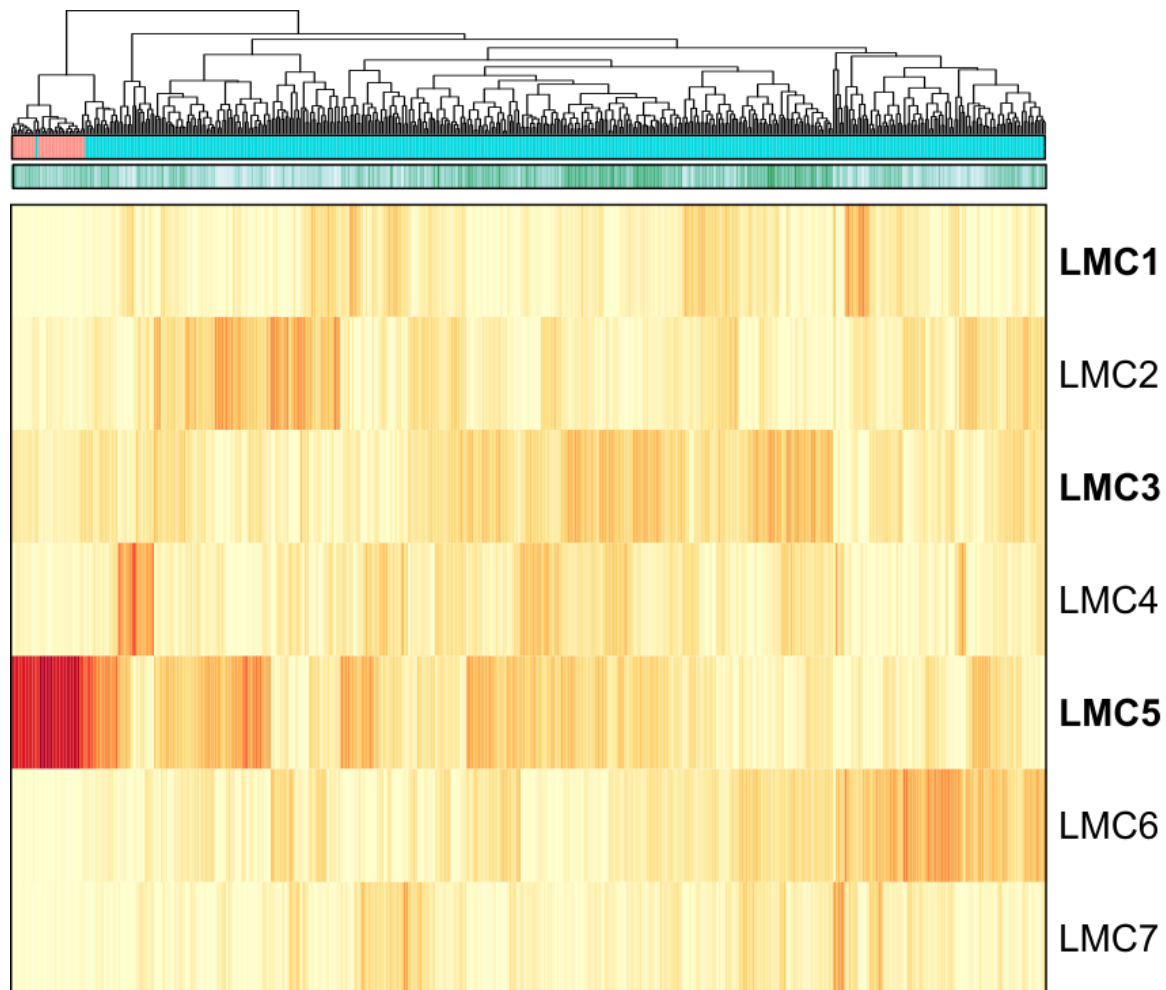
HYBRIDIZATION: red channel



Parameter selection

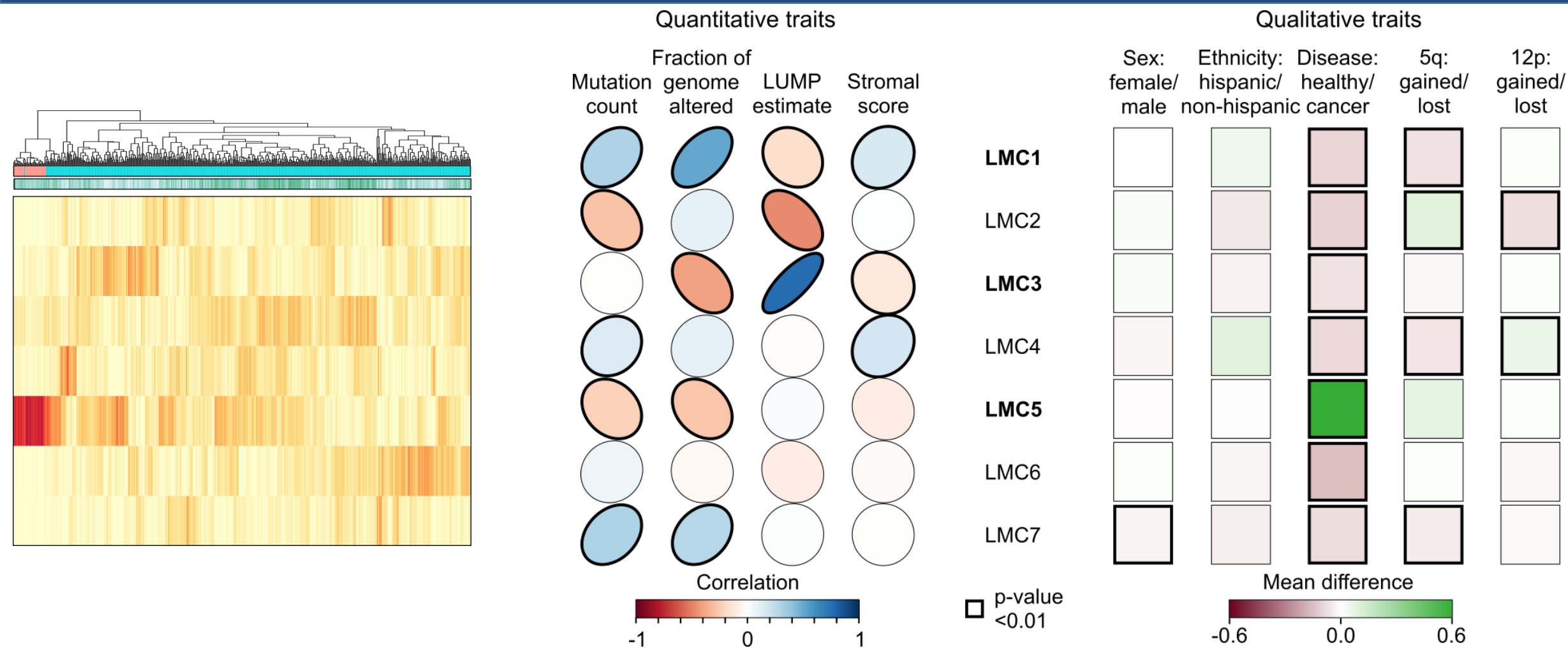


Proportions heatmap

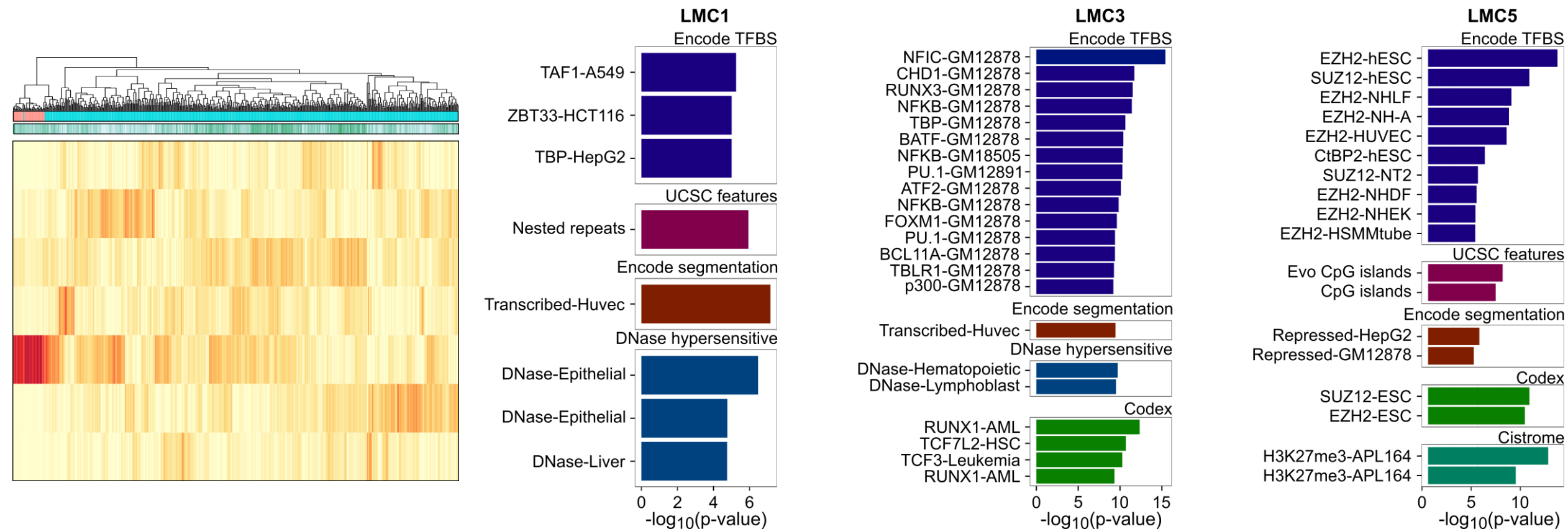


¹ Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 1-11 (2015).

Phenotypic trait associations

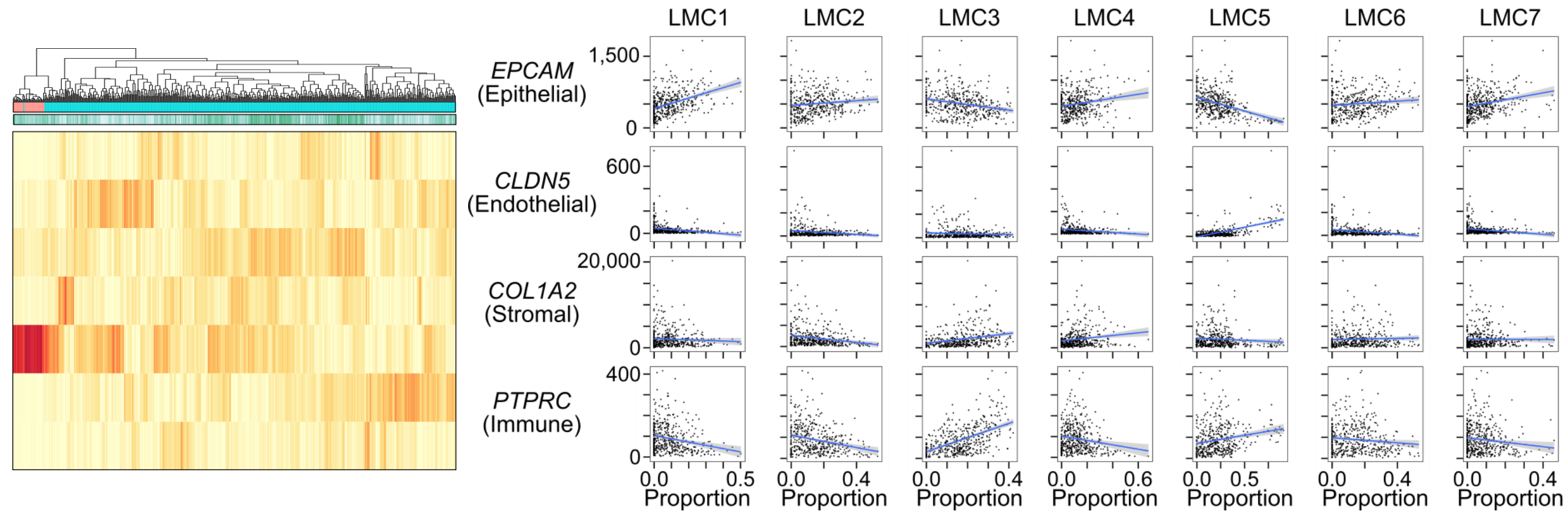


LMC LOLA¹ enrichment analysis



¹Sheffield, N. & Bock, C. LOLA:Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 32, 587-589 (2016).

Sample-specific marker gene expression



Conclusions

- Thorough data processing and biologically guided interpretation more critical than the deconvolution tool itself
- Three-stage protocol
 - Quality-adapted CpG filtering and confounding factor adjustment with ICA using *DecompPipeline*
 - Methylome deconvolution using *MeDeCom*, *RefFreeCellMix* or *EDec*
 - Validation and interpretation of deconvolution results with *FactorViz*
- Deconvolution of TCGA LUAD dataset shows indications of immune cell infiltration, stromal, and epithelial components

Acknowledgements



Pavlo Lutsik
Reka Toth
Valentin Maurer
Christoph Plass



Petr V. Nazarov
Tony Kaoma



Jörn Walter
Shashwat Sahay



Thomas Lengauer

