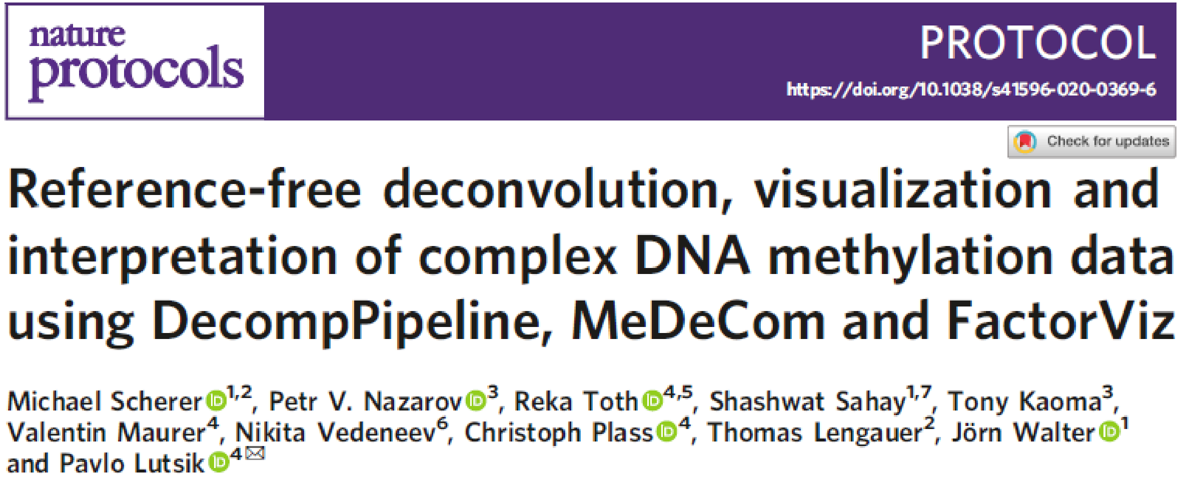
dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

mpi max planck institut informatik

LUXEMBOURG INSTITUTE OF HEALTH
Multiomics Data Science, Quantitative Biology Unit

Reference-free deconvolution of complex DNA methylation data

Michael Scherer, Petr V. Nazarov, ...,
Tony Kaoma, ..., Pavlo Lutsik



nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-020-0369-6

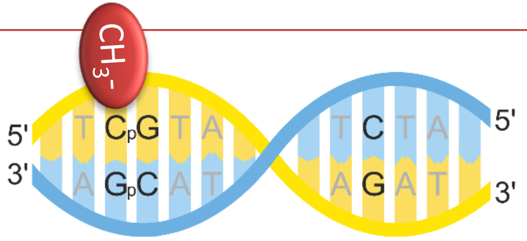
Check for updates

Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz

Michael Scherer[1,2], Petr V. Nazarov[3], Reka Toth[4,5], Shashwat Sahay[1,7], Tony Kaoma[3], Valentin Maurer[4], Nikita Vedeneev[6], Christoph Plass[4], Thomas Lengauer[2], Jörn Walter[1] and Pavlo Lutsik[4]✉

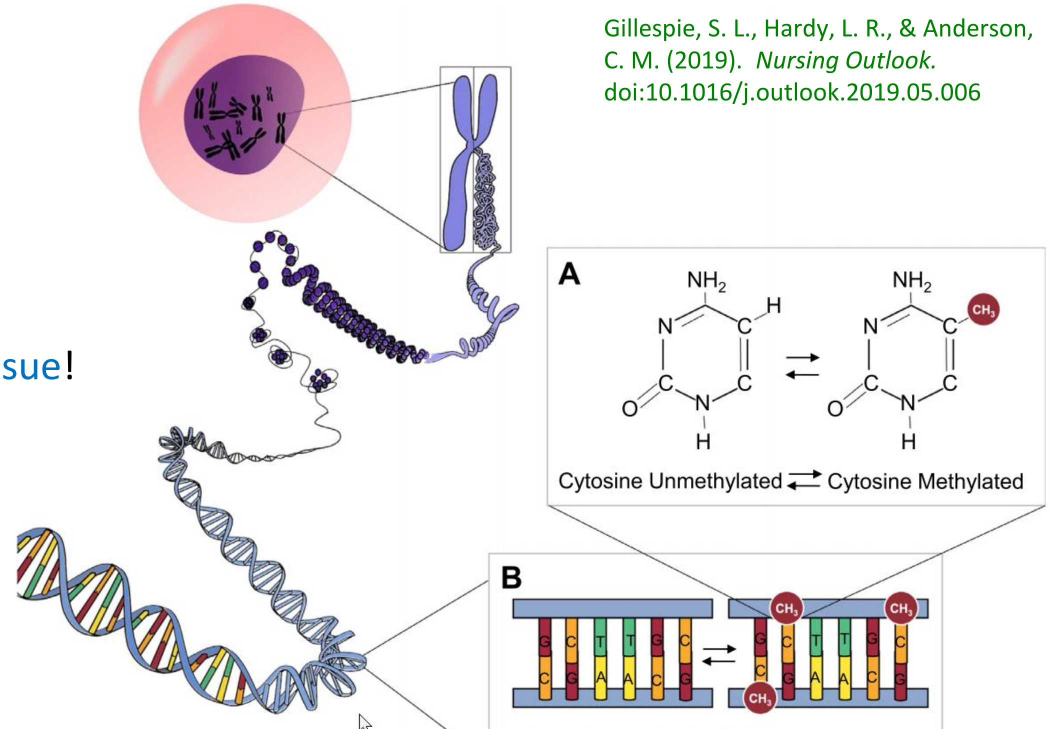
Fonds National de la Recherche Luxembourg
C17/BM/11664971/DEMICS

Background: DNA Methylation



CpG is shorthand for *5'—C—phosphate—G—3'*

Gillespie, S. L., Hardy, L. R., & Anderson, C. M. (2019). *Nursing Outlook.* doi:10.1016/j.outlook.2019.05.006

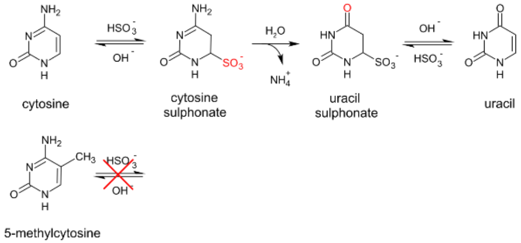

A
NH2
N
C
H
O
CH3
Cytosine Unmethylated ⇄ Cytosine Methylated
B
G C T T G C
C G A A C G
CH3
Gene Transcribed ⇄ Gene Repressed

Main features

- Responsible for tissue differentiation and is specific to tissue!
- Can be changed by external factors and life style
- Typically repress transcription (if in promoter)
- Is strongly involved in carcinogenesis
- DNAm signature is much more stable than RNA – works even for paraffin-embedded samples

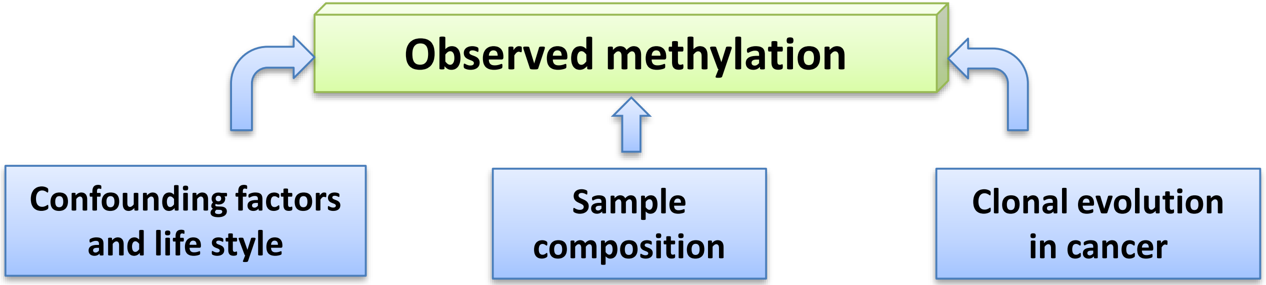
Methods

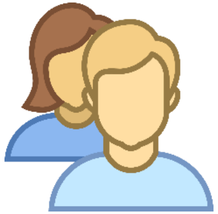
- Standard: "bisulfite" (HSO_3^-) treatment: unmethylated CpG→UpG
- Illumina arrays: 450k and EPIC (850k)
- Sequencing: RRBS, WGBS

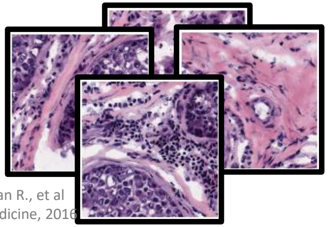


Background: Heterogeneity

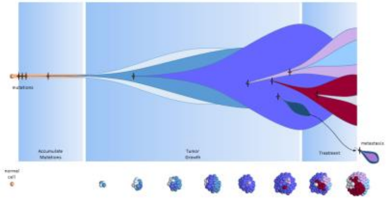
Heterogeneity in methylation data


Observed methylation
Confounding factors and life style
Sample composition
Clonal evolution in cancer





Natrajan R., et al
PLoS Medicine, 2016

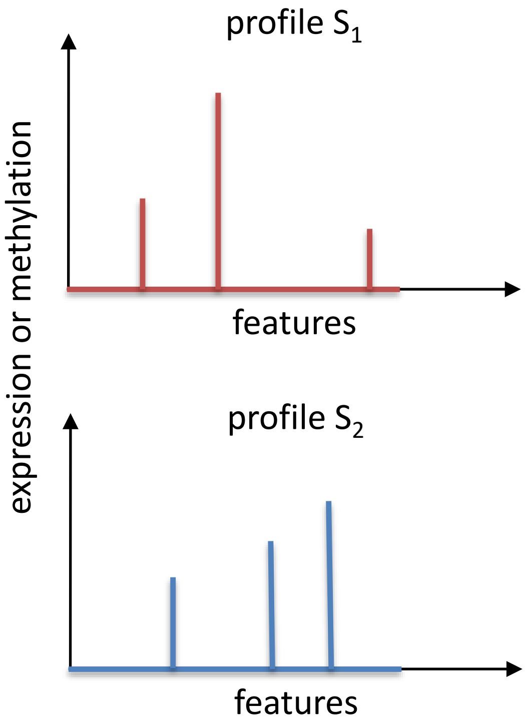


https://www.pancanology.com/clonal-evolution-model/

- Gender, ethnicity, age, lifestyle
- Natural tissue heterogeneity
- Inter/intra tumor heterogeneity due to clonal evolution

It is important to disentangle these effects! Ideally in a **reference-free** manner

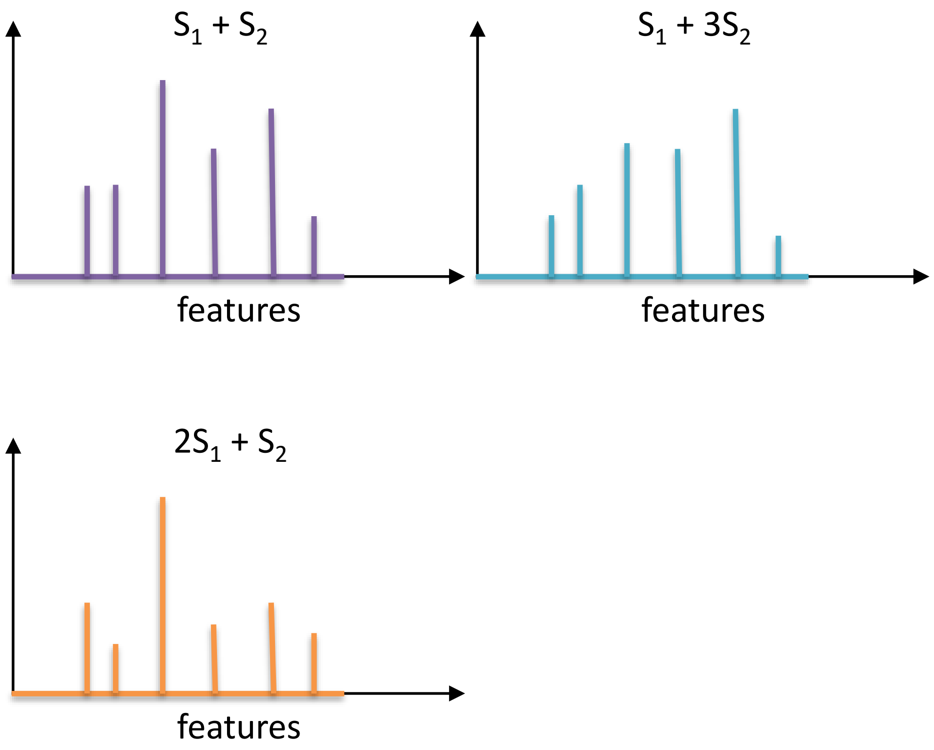
Mixing and Non-negative Matrix Factorization (NMF)



Mixing

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

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



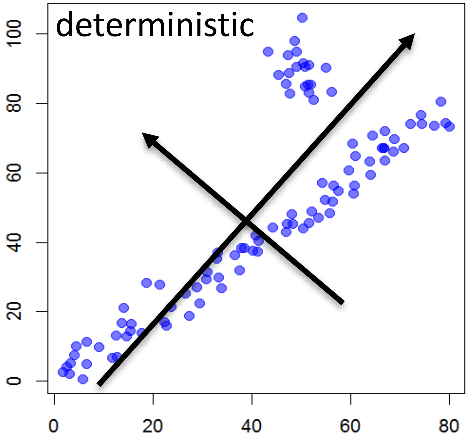
NMF: $\mathbf{D} \approx \mathbf{T} \times \mathbf{A}$ all elements ≥ 0

T estimates profiles **S**

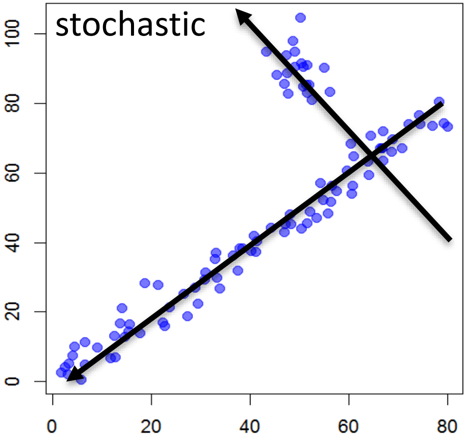
A estimates mixing **M**

Advantages and Issues of NMF

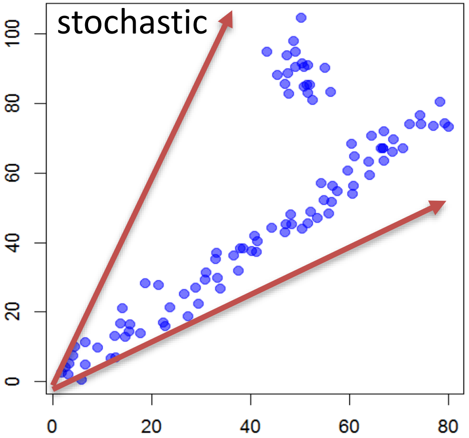
PCA



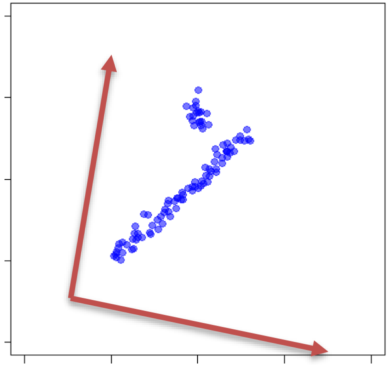
ICA



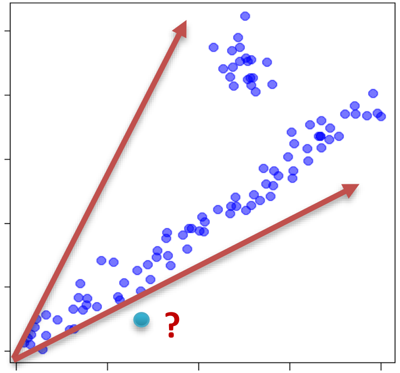
NMF



NMF: issue 1



NMF: issue 2



Advantages of NMF:

- Fits physical principles
- Easy to interpret

Issues of NMF:

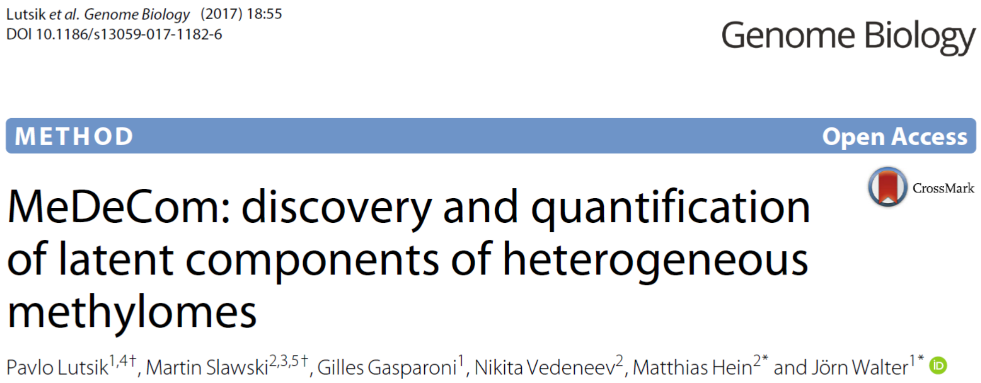
- Multiple solutions
- Is the minimal description stable?

⇒ **we need:**

- ➢ additional restrictions
- ➢ regularizations during fitting

Sompairac el al, Int J Mol Sci, 2019 (link)
Cantini el al, Bioinformatics, 2019 (link)

MeDeCom: Core Algorithm



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

Genome Biology

METHOD Open Access

CrossMark

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*]

Hypothesis: in a pure cell population, methylation should be either 0 or 1

D = T×A + e

Standard NMF:

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

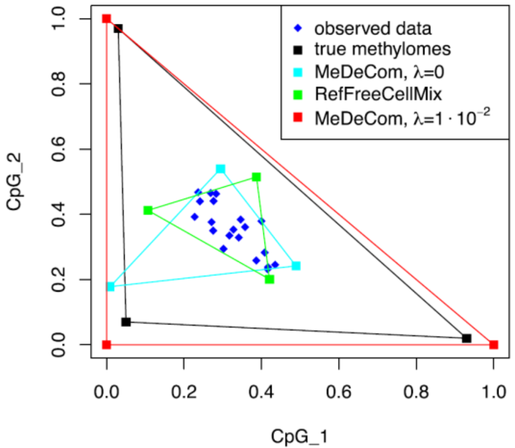
MeDeCom's regularization:

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

$$\text{subject to } 0 \le T_{is} \le 1 \ \ \forall i,s$$

$$A_{sj} \ge 0 \ \ \forall s,j$$

$$\sum_{s=1}^{k} A_{sj} = 1 \ \ \forall j,$$



Other reference-free tools:

RefFreeCellMix – Houseman, BMC Bioinformatics, 2016 (link)
EDec – Onuchic, Cell Rep., 2016 (link)

MeDeCom: Issues

Assumptions & Requirements

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

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

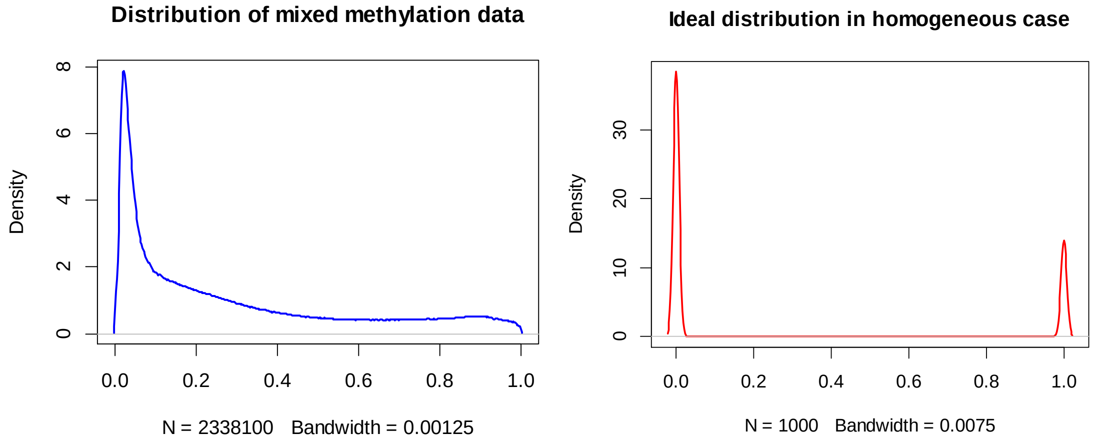
(3) Population mixtures are variable b/w samples.

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

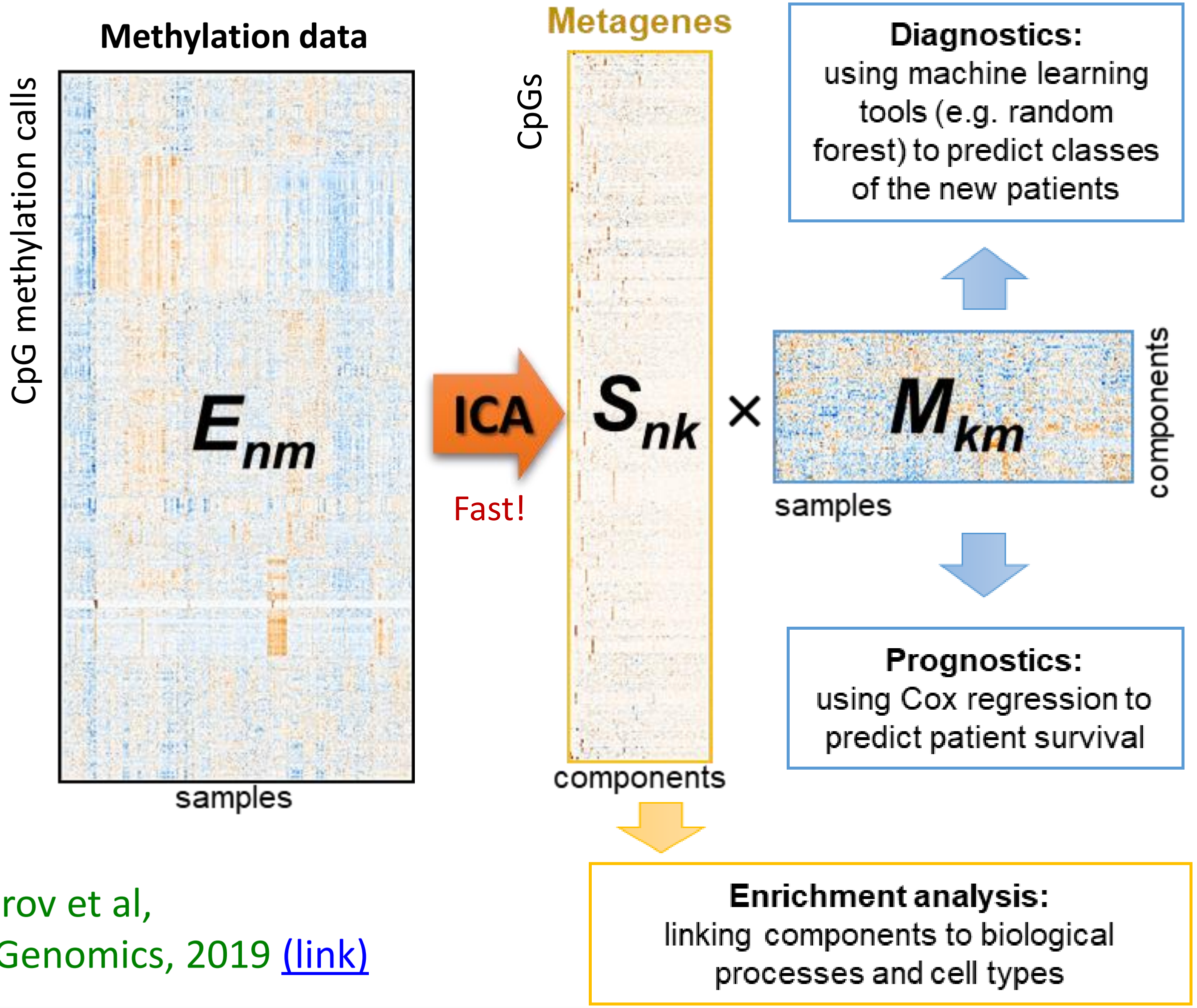
Issues

(1) Extremely time / memory consuming, runs on HPC (easily can reach 10^4 runs to cover hyperparameter space)

(2) Sensitive to technical noise and confounding factors (gender, age,..)



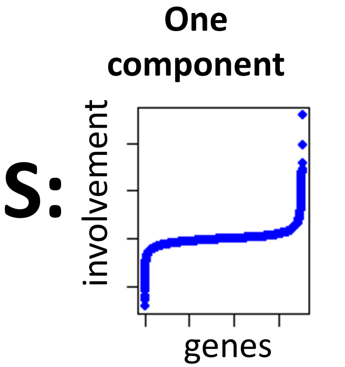
Consensus Independent Component Analysis (consICA)

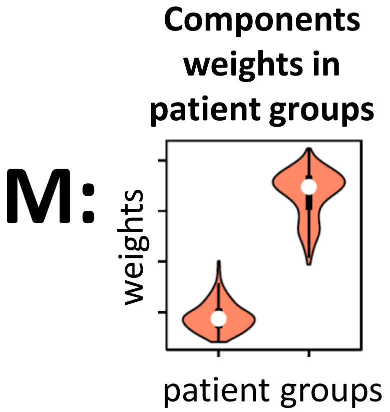


Diagnostics: using machine learning tools (e.g. random forest) to predict classes of the new patients

Prognostics: using Cox regression to predict patient survival

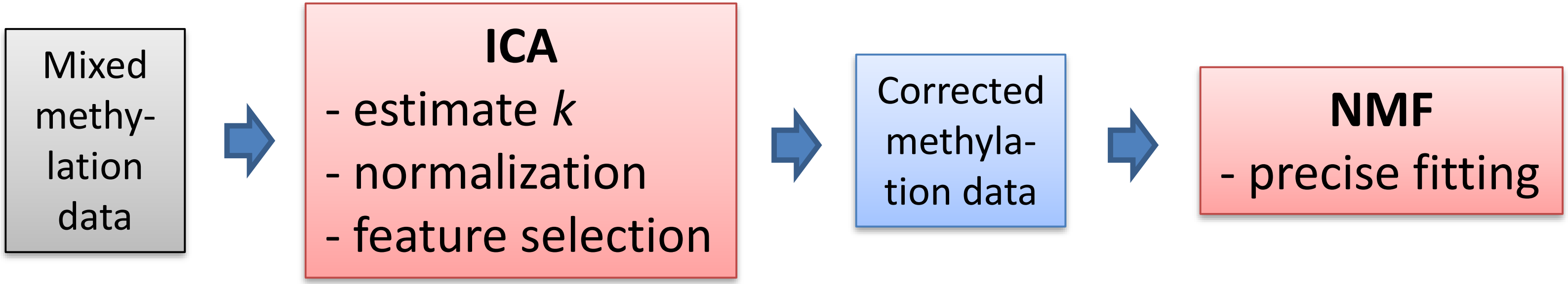
Enrichment analysis: linking components to biological processes and cell types





consICA: Nazarov et al, BMC Medical Genomics, 2019 (link)

Deconvolution Data Challenge, 2018



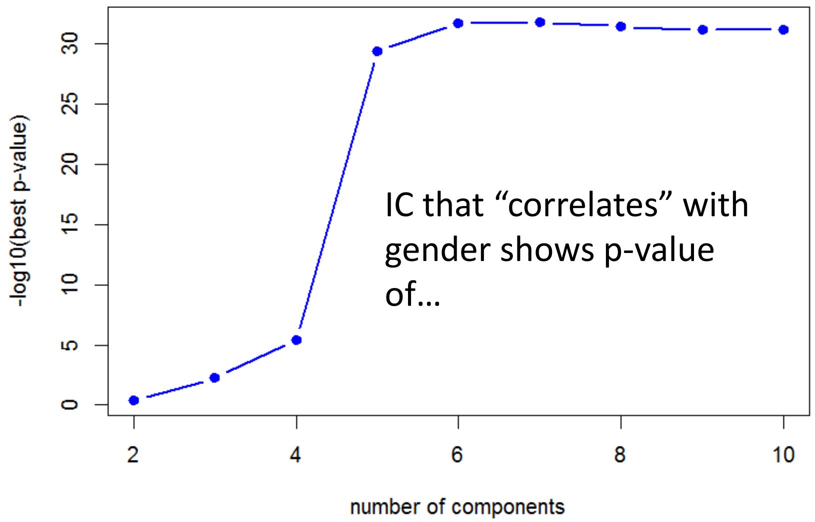
Mixed methylation data

ICA
- estimate *k*
- normalization
- feature selection

Corrected methylation data

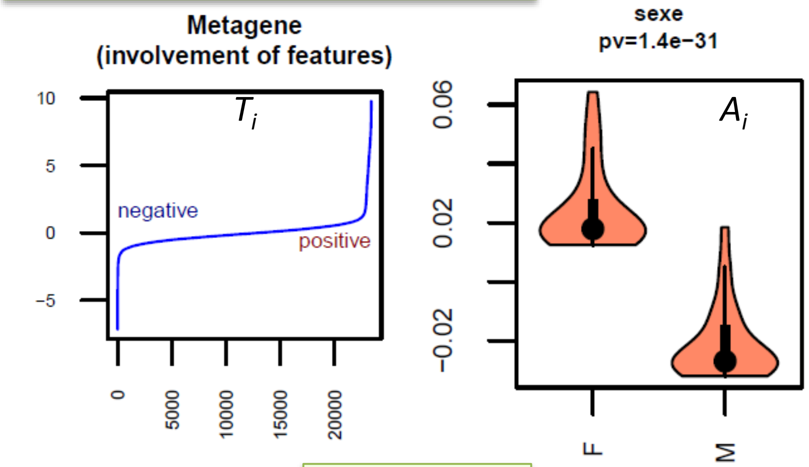
NMF
- precise fitting

Captures gender - one of the confounders



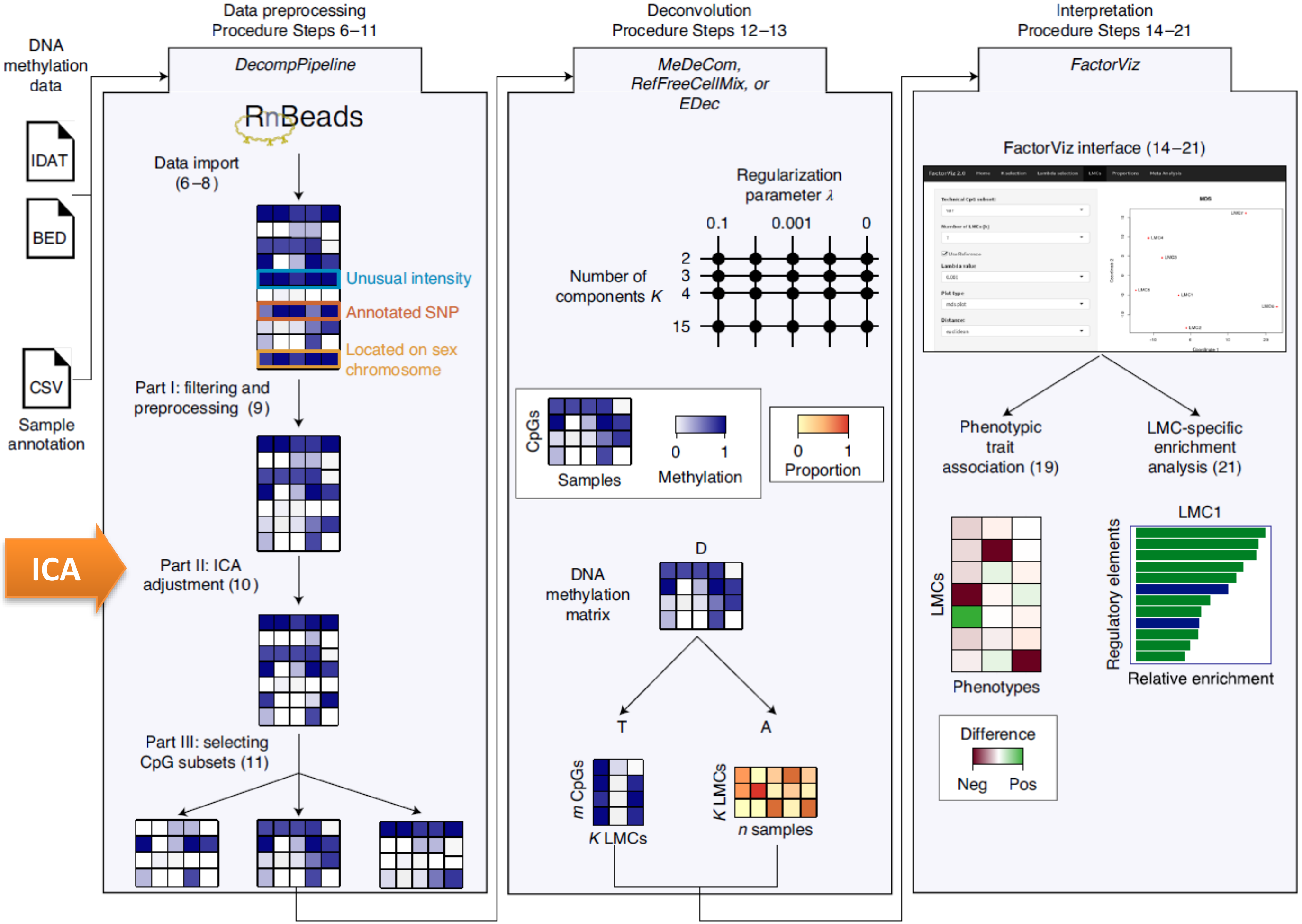
IC that “correlates” with gender shows p-value of…

Can reduce number of features



X = T x A

Pipeline Overview

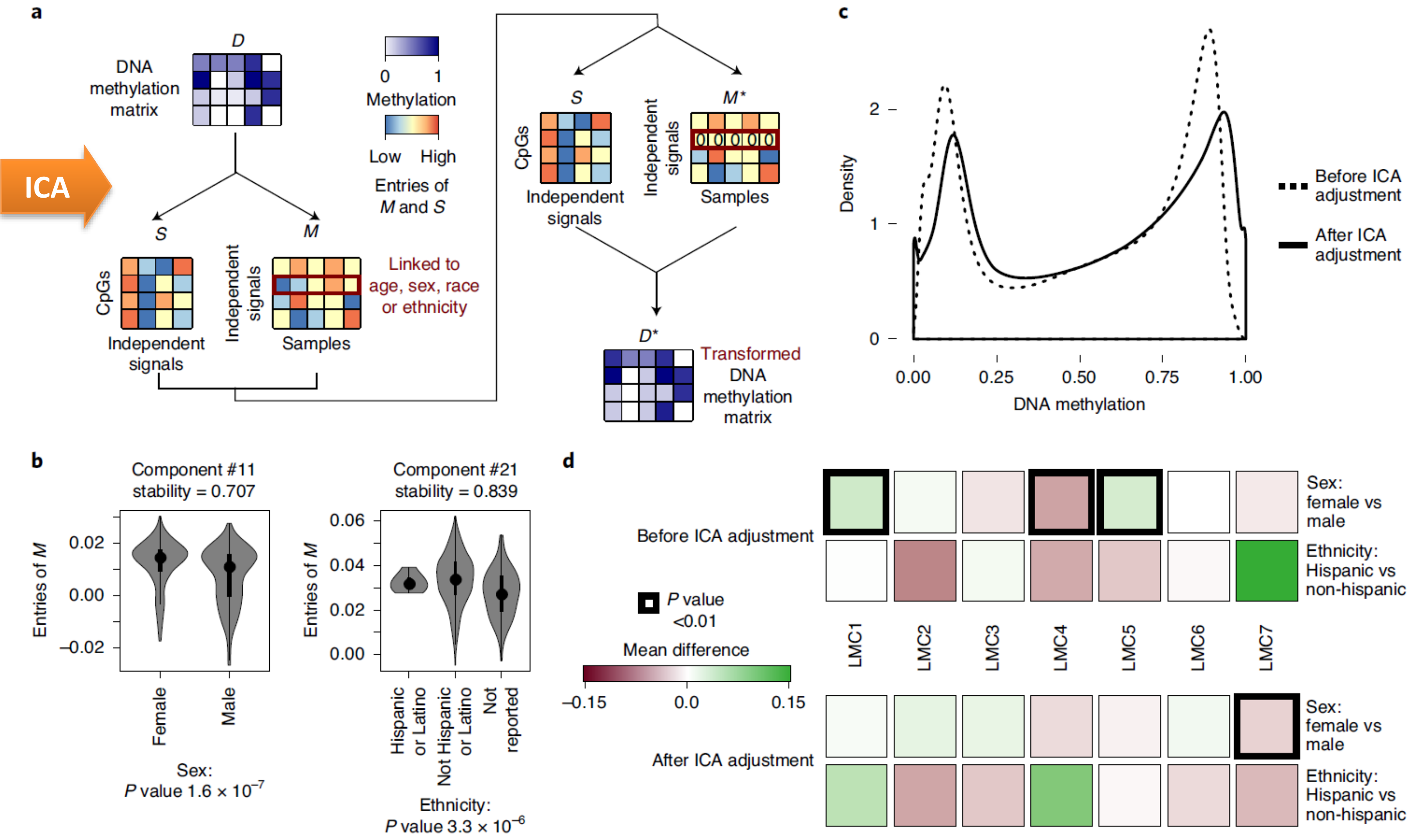


(1) Any methylation technology. *DecompPipeline:* data import, preprocessing, accounting for confounders and feature selection by ICA.

(2) *MeDeCom (RefFreeCellMix* or *Edec)* performs deconvolution of data into the **latent methylation components** (LMCs) and the proportions matrix. λ and K should be identified.

(3) The results are interpreted using the R/Shiny visualization tool *FactorViz*

ICA Results: Preprocessing



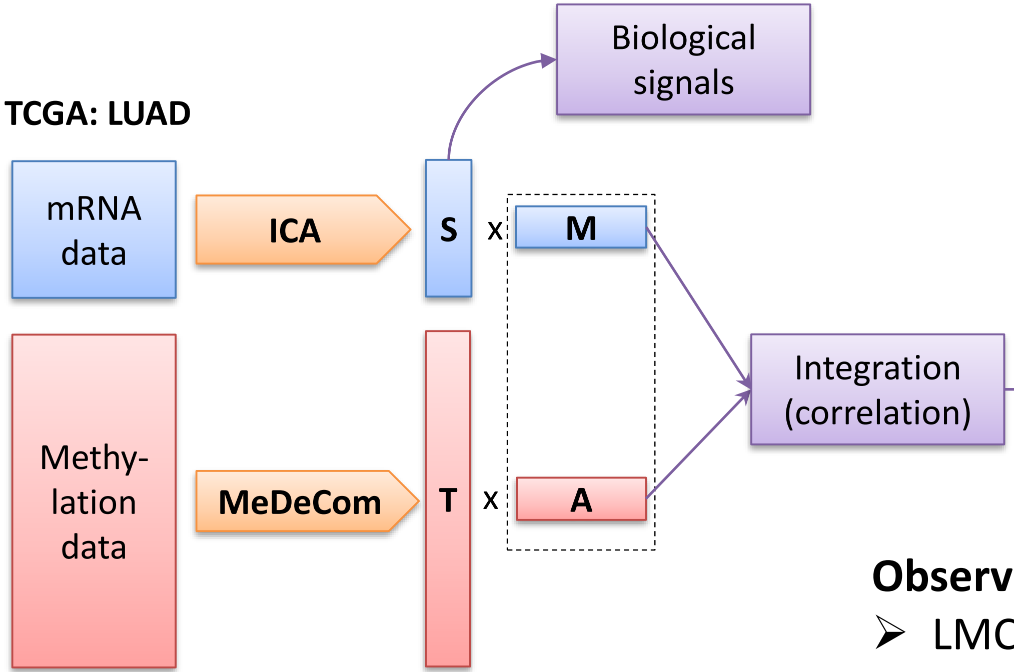
Evaluation of ICA on TCGA LUAD dataset.
(a,b) ICA deconvolution: components linked to confounding factors are detected and removed.

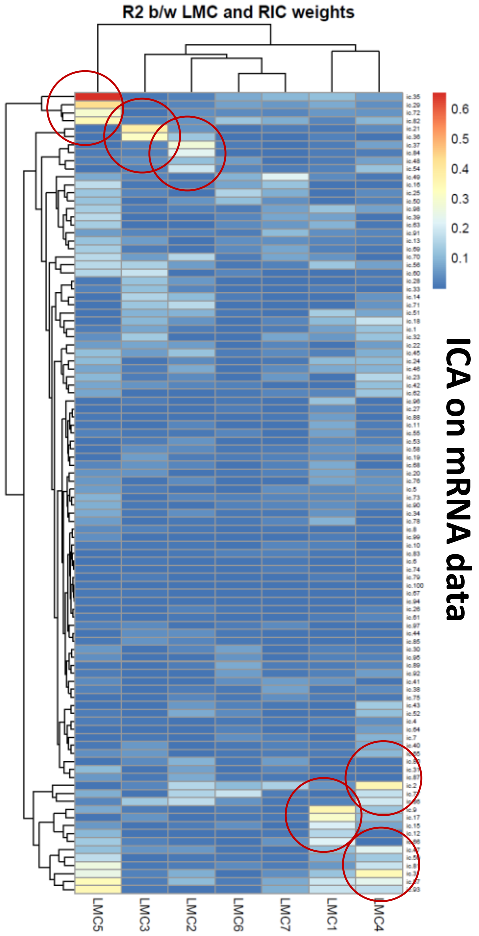
(c) Distributions of the transformed (D*) and original (D) methylation matrices.

(d) Associations between LMC proportions and qualitative phenotypic traits. (□ - significant)

ICA Results: Integration with RNAseq

TCGA: LUAD





MeDeCom on methylation data

Observations:

- ➢ LMC5 ↔ IC35
- ➢ LMC3 ↔ IC21
- ➢ LMC1 ↔ IC9
- ➢ LMC4 ↔ IC2, IC3

Direct functional annotation of methylation is challenging – we end to map CpGs onto promoter regions. In paper: LOLA (region-based) and GO on hypomethylated sites – IMHO can be improved

Recommendations?

Interpretation

LMC5 ↔ IC35

LMC5 was correlated with marker gene CLDN5 (Endothelial), pv = 1e-42

Functional annotation of **IC35** is:

GO:BP pos : 59 terms(FDR<0.01)

Term

regulation of vasoconstriction
extracellular structure organization
regulation of receptor activity
regulation of ERK1 and ERK2 cascade
angiogenesis
positive regulation of cell proliferatio...

LMC3 ↔ IC21

LMC3 was correlated with marker gene PTPRC (Immune), pv = 1e-32

Functional annotation of **IC21** is:

GO:BP pos : 78 terms(FDR<0.01)

Term

immune response
B cell activation
inflammatory response
positive regulation of lymphocyte prolif...
B cell receptor signaling pathway
chemokine−mediated signaling pathway
lymphocyte migration

LMC1 ↔ IC9

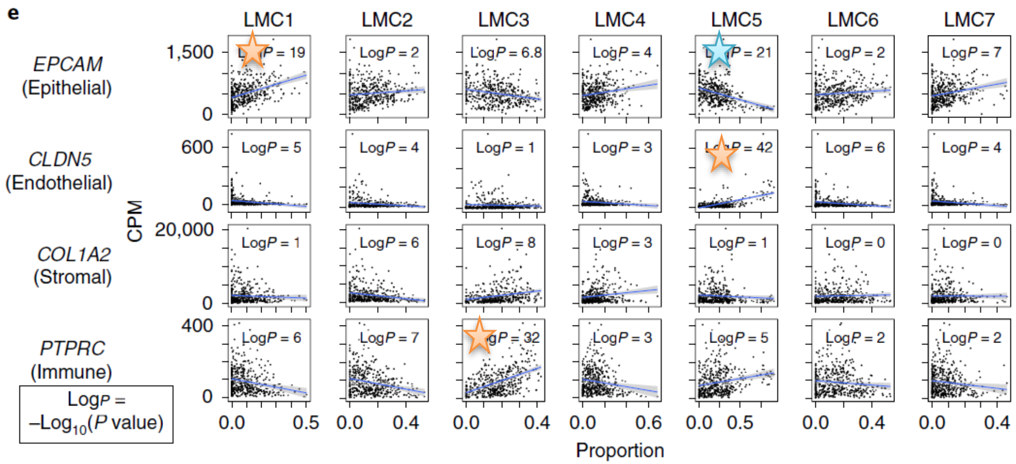
LMC3 was correlated with marker gene EPCAM (Epithelial), pv = 1e-19

Functional annotation of **IC9** is: (???)

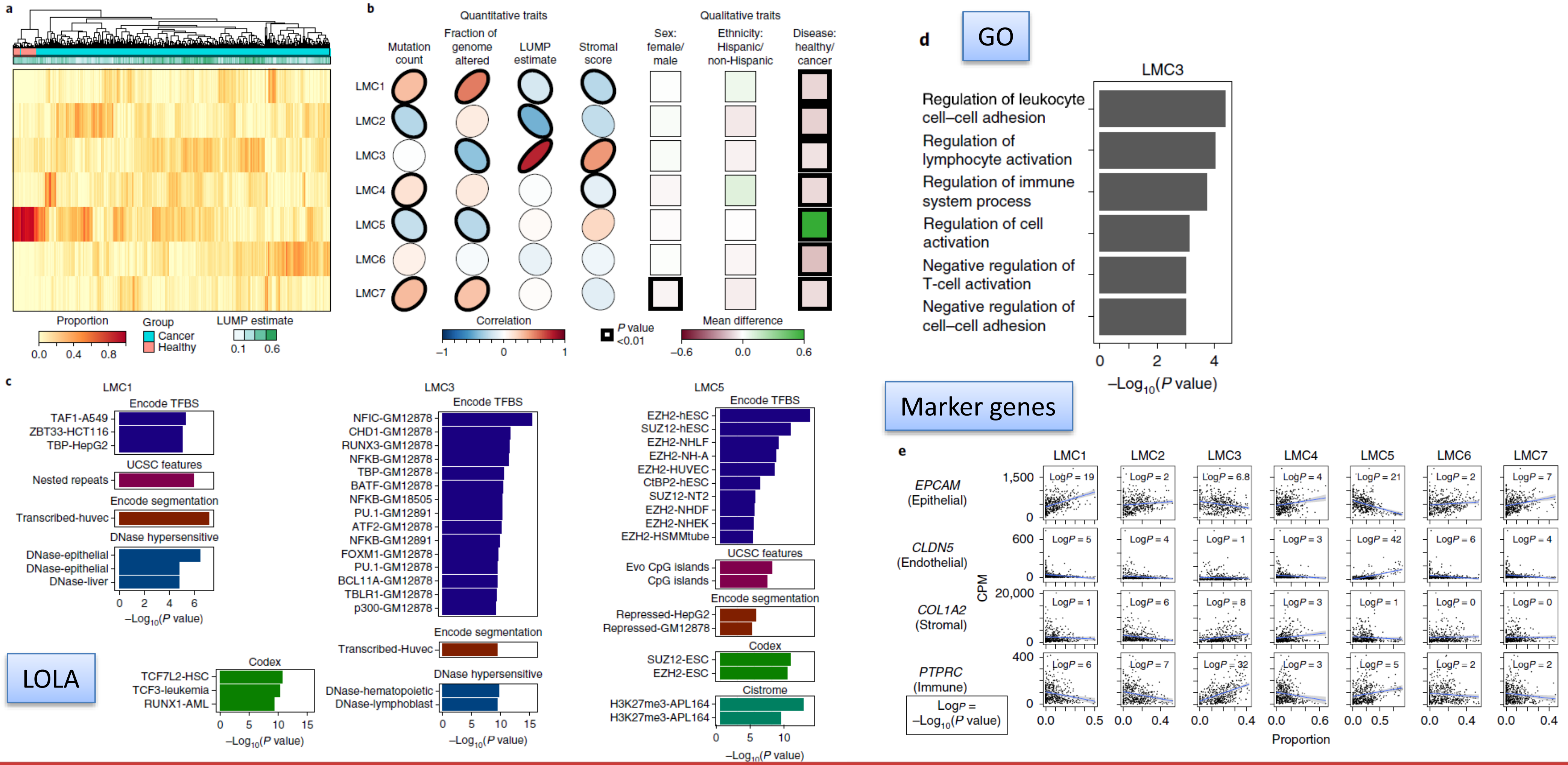
GO:BP pos : 53 terms(FDR<0.01)

Term

regionalization
embryonic organ morphogenesis
embryonic skeletal system development
positive regulation of transcription fro...
limb development
neuron fate specification
nervous system development



Pipeline Output: LUAD, Illumina

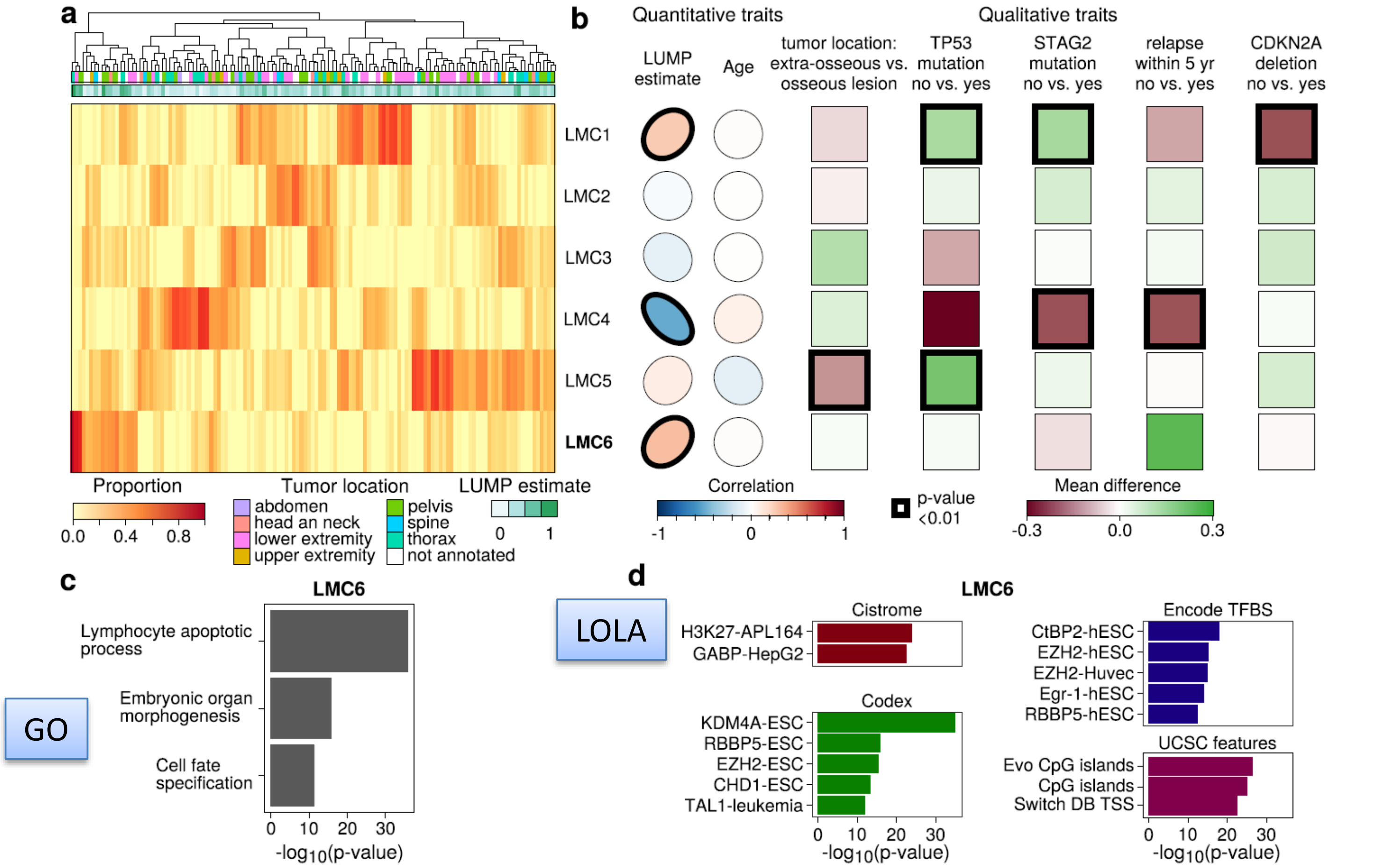


GO

Marker genes

LOLA

Pipeline Output: Ewing sarcoma, RRBS data

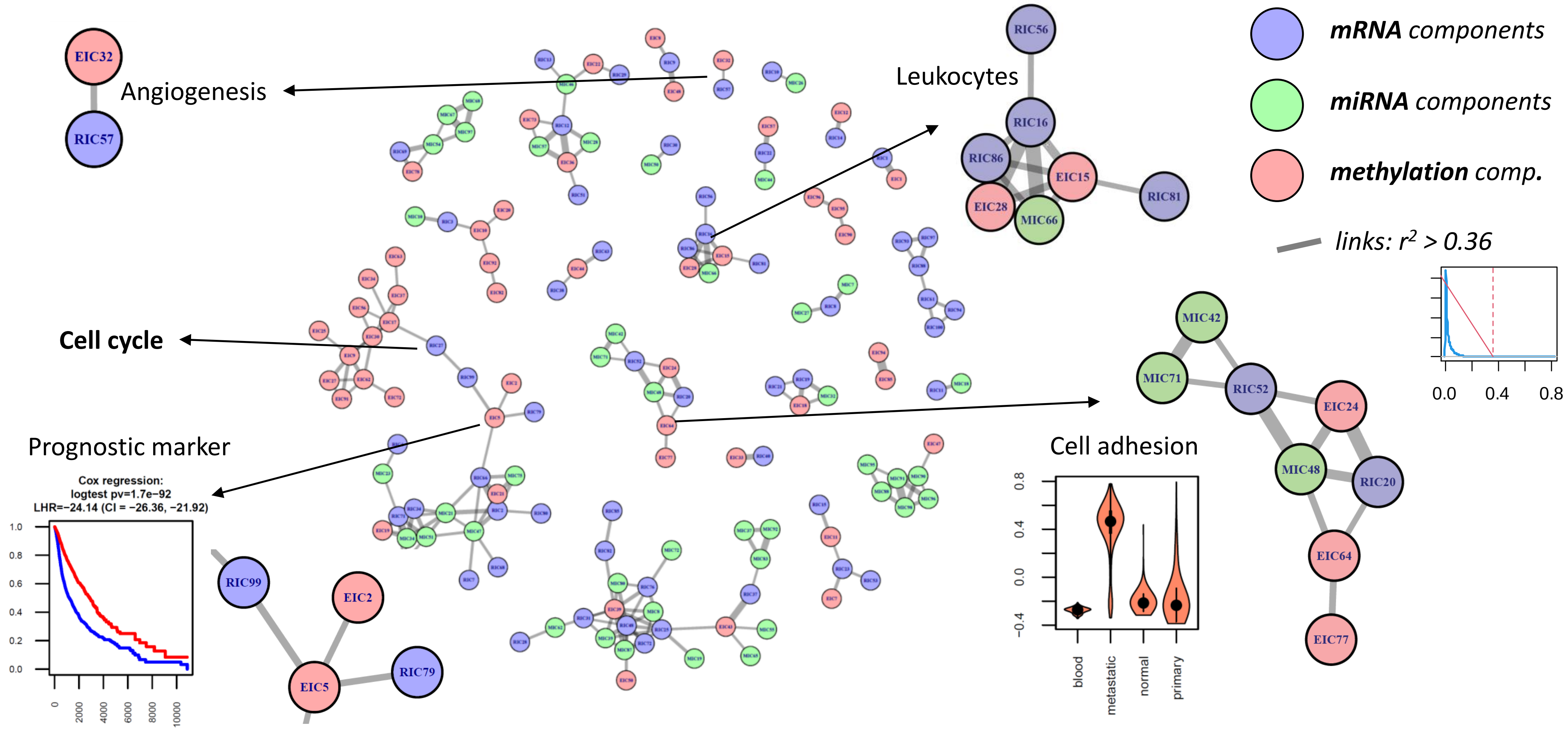


GO

LOLA

RRBS Reduced-representation bisulfite sequencing. A next-generation sequencing strategy yielding CpG methylation calls in CpG-dense regions of the genome.

How about ICA alone ?



Angiogenesis

Leukocytes

Cell cycle

Prognostic marker

Cell adhesion

mRNA components

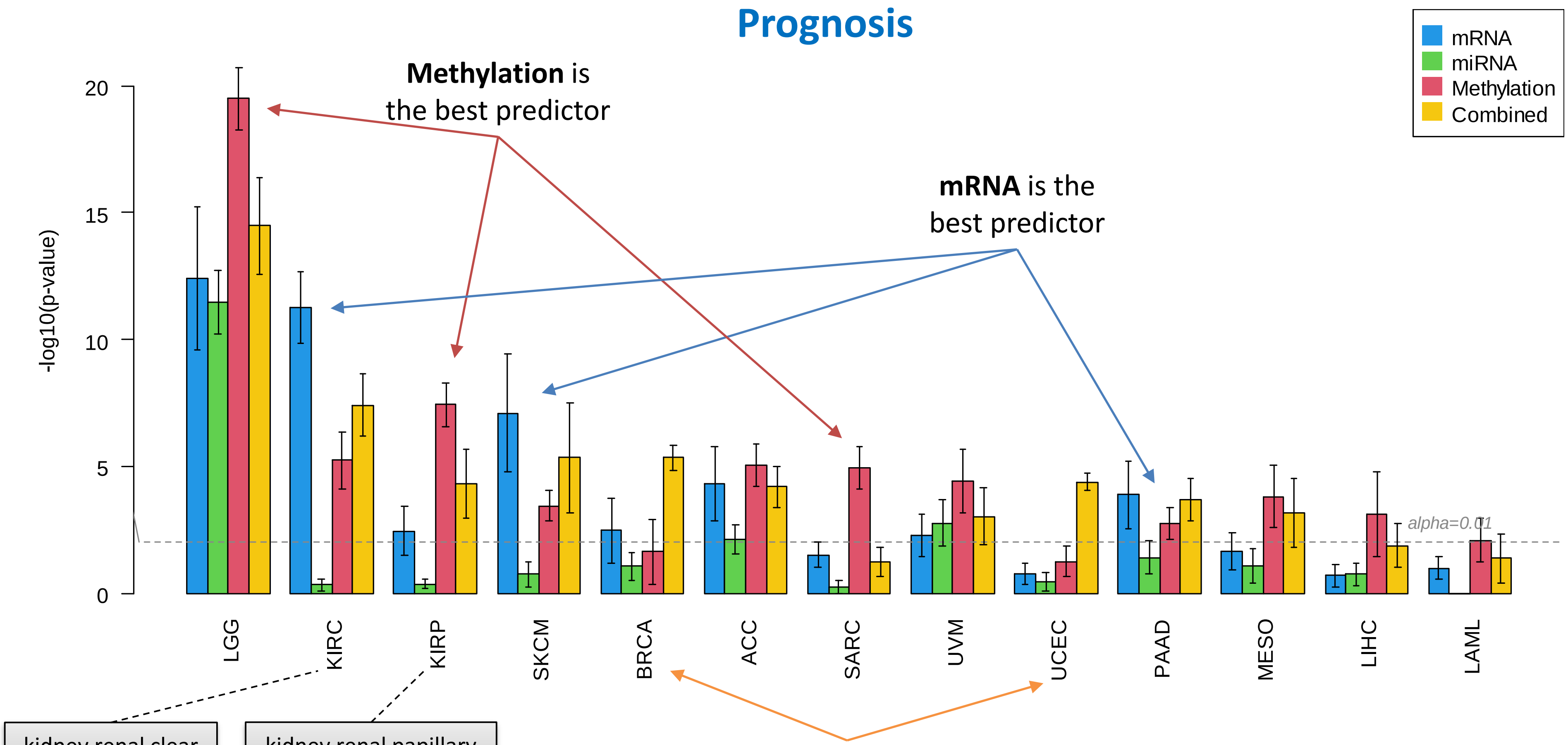
miRNA components

methylation comp.

links: $r^2 > 0.36$

Cox regression:
logtest pv=1.7e−92
LHR=−24.14 (CI = −26.36, −21.92)

How about ICA alone ?



Methylation is the best predictor

mRNA is the best predictor

kidney renal clear cell carcinoma

kidney renal papillary cell carcinoma

Combined features give good predictions for breast and uterine carcinomas

Conclusions

Presented pipeline: DecompPipeline + MeDeCom + FactorViz:

(1) provides a complete pipeline of combining top available tools

(2) is applicable for bisulphate sequencing data

(3) (early) MeDeCom was tested on synthetic and experimental data

(4) When in the pipeline, similar results with RefFreeCellMix

Limitations of the approach:

- low number of components (usually <10)
- may be tricky to interpret without RNA-seq data
- missing some important subpopulations: proliferating tumor cells (though, cell division may be not affecting methylation?..)

Our **consICA approach** can be applicable to methylation data as well. Despite it does not estimate concentrations as precise as MeDeCom, but it can extract a lot more meaningful biological signals!