

Multi-omic statistical inference of cellular heterogeneity

PEPR Santé Numérique - IRP 3

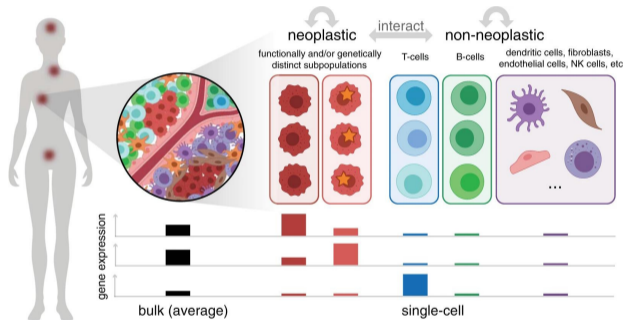
————— H. Barbot¹ D. Causeur¹ Y. Blum² M. Richard³ —————

¹IRMAR - UMR CNRS 6625, ²IGDR - UMR CNRS 6290, ³TIMC - UMR CNRS 5525



JOBIM, June 2024

Cellular heterogeneity



[1]

Cellular heterogeneity in bulk:

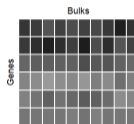
- refers to the variety of cell types within the bulk,
- reflects progression of **disease state**,
- is a **complex mixture** signal,
- is **difficult to assess** from bulk molecular profiles.

50+ algorithms exist and many benchmarks to compare them [2], [3], [4]

Supervised cell deconvolution

Cellular deconvolution **assumes** that bulk omic profiles result from **weighted sums** of so-called signature cell-specific omic profiles.

Y matrix
 $M \times N$

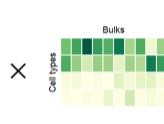


X matrix
 $M \times K$



~

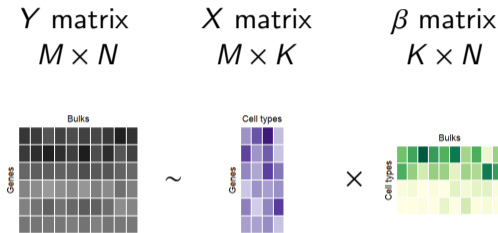
β matrix
 $K \times N$



×

Supervised cell deconvolution

Cellular deconvolution **assumes** that bulk omic profiles result from **weighted sums** of so-called signature cell-specific omic profiles.



Most of supervised methods results from Ordinary Least Squares (OLS) optimization.

$$\begin{cases} \forall i \in \llbracket 1; N \rrbracket & Y_i = X\beta_i + \varepsilon_i, \\ \mathcal{L}(\varepsilon_i) = \mathcal{N}(0, \sigma^2 I_M). \end{cases}$$

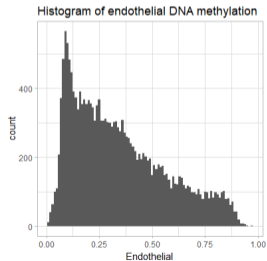
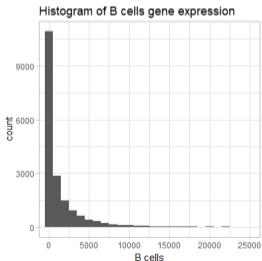
under constraints for each β_i $\begin{cases} \sum_{k=1}^K \beta_{ik} = 1, \\ 0 \leq \beta_{ik} \leq 1. \end{cases}$

Two particular omics

Cell deconvolution framework based on extensions of OLS are not designed for a specific omic data or for estimating a vector lying within the K -simplex.

Cell deconvolution is frequently used on two omic data types :

- RNA-seq gene expression read counts.
- DNA methylation rates (*beta values*).



Benchmark data

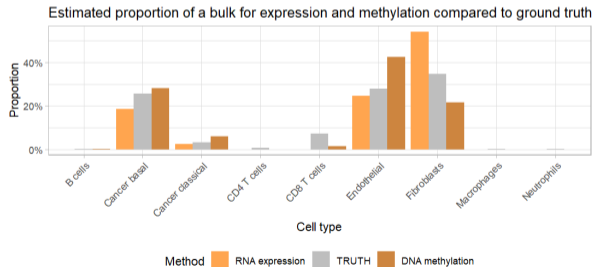
Data generated in vitro, using cell types commonly found in pancreatic ductal adenocarcinoma (**PDAC**), from TIMC MAGE team. \Rightarrow

- 21104 gene expressions
- 800000 CpG sites
- $N = 30$ **independent** bulks
- $K = 9$ cell types

Benchmark data

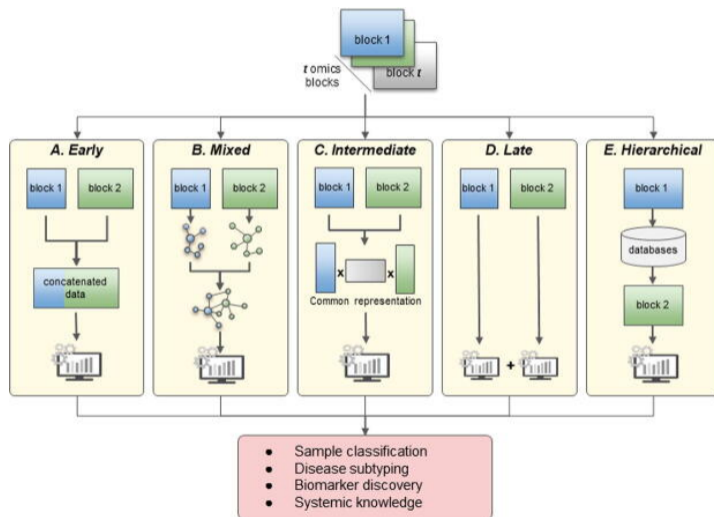
Data generated in vitro, using cell types commonly found in pancreatic ductal adenocarcinoma (**PDAC**), from TIMC MAGe team. \Rightarrow

- 21104 gene expressions
- 800000 CpG sites
- $N = 30$ **independent** bulks
- $K = 9$ cell types



For a bulk example, same global accuracy but two different insights.

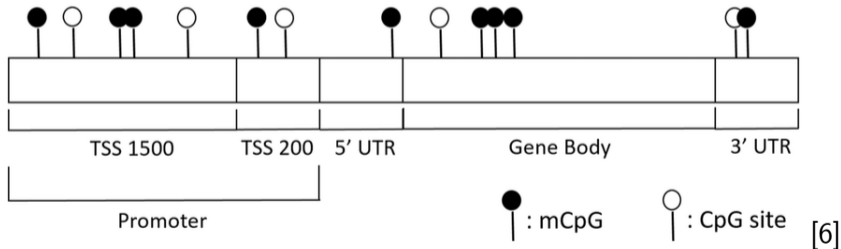
How can we do multi-omic cell deconvolution



[5] M. Picard, MP. Scott-Boyer, A. Bodein, O. Périn, A. Droit.

Common representation

→ Methylation rates are **aggregated into gene-level** measurements by averaging over all values at CpG sites in the **promoter region** of each gene.



A negative binomial framework for cell deconvolution on expression

For $1 \leq i \leq N$ and $1 \leq j \leq M$, let y_{ij} denote the expression level of gene j for bulk i and x_j the vector of expression level of all K cell types.

→ A **negative binomial** regression model is assumed for **overdispersed** gene expression counts.

$$\mathbb{P}(Y_{ij} = y_{ij} \mid x_j) = \frac{\Gamma(y_{ij} + \frac{1}{\alpha_i})}{\Gamma(y_{ij} + 1)\Gamma(\frac{1}{\alpha_i})} \left(\frac{1}{1 + \alpha_i \mu_i(x_j)} \right)^{\frac{1}{\alpha_i}} \left(\frac{\alpha_i \mu_i(x_j)}{1 + \alpha_i \mu_i(x_j)} \right)^{y_{ij}},$$

- $\alpha_i > 0$ is a scale parameter: $\text{Var}(Y_{ij} \mid x_j) = \mu_i(x_j)(1 + \alpha_i \mu_i(x_j))$,
- $\mu_i(x_j) = \mathbb{E}(Y_{ij} \mid x_j) = \beta_{0i} + \beta_{1i}x_{j1} + \dots + \beta_{Ki}x_{jK} > 0$,
- $\beta_i = (\beta_{1i}, \dots, \beta_{Ki})'$ is the vector of proportions of each cell type, constrained to lie within the K -simplex.

The maximisation of the **weighted log-likelihood** $\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y)$ is the objective function.

A beta framework for cell deconvolution on methylation

For $1 \leq i \leq N$ and $1 \leq j \leq M$, let z_{ij} denote the methylation rates of aggregated cites j for bulk i and \tilde{x}_j the vector of methylation rates of all K cell types.

→ A **Beta** regression model is assumed for DNA methylation **rates**.

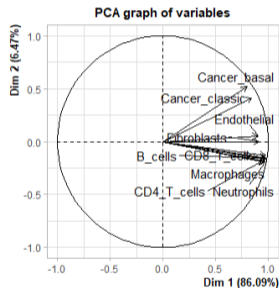
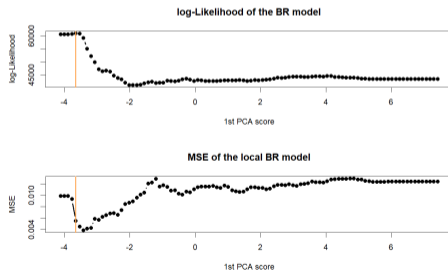
$$\varphi(z_{ij} \mid \tilde{x}_j) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i(\tilde{x}_j)\phi_i)\Gamma((1 - \mu_i(\tilde{x}_j))\phi_i)} z^{\mu_i(\tilde{x}_j)\phi_i - 1} (1 - z)^{(1 - \mu_i(\tilde{x}_j))\phi_i - 1},$$

- $\phi_i > 0$ is a precision parameter: $\text{Var}(Z_{ij} \mid x_j) = \frac{\mu_i(x_j)(1 - \mu_i(x_j))}{1 + \phi_i}$,
- $\mu_i(\tilde{x}_j) = \mathbb{E}(Z_{ij} \mid \tilde{x}_j) = \tilde{\beta}_{0i} + \beta_{1i}\tilde{x}_{j1} + \dots + \beta_{Ki}\tilde{x}_{jK} > 0$,
- $\beta_i = (\beta_{i1}, \dots, \beta_{iK})'$ is the vector of proportions of each cell type, constrained to lie within the K -simplex.

The maximisation of the **weighted log-likelihood** $\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ is the objective function.

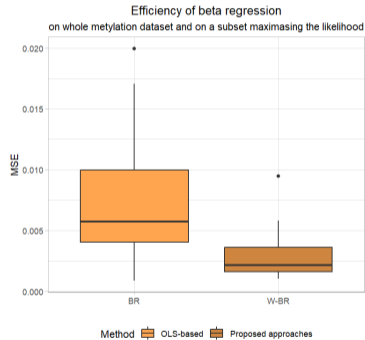
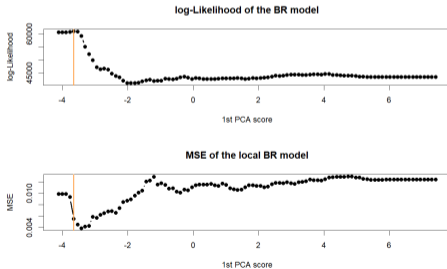
A different weighting strategy

- PCA on the reference matrix of methylation data and first,
- Evaluate the log likelihood and the precision of estimation on subset of aggregated methylation data.



A different weighting strategy

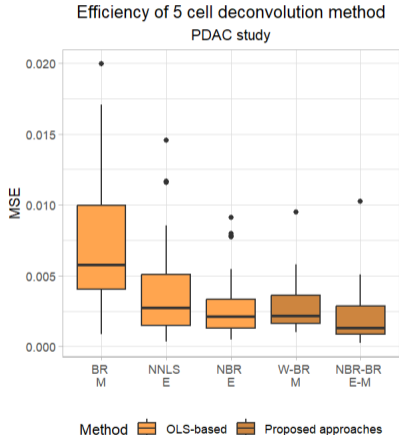
- PCA on the reference matrix of methylation data and first,
- Evaluate the log likelihood and the precision of estimation on subset of aggregated methylation data.



Intermediate multi-omic framework

Cyclic Coordinate Descent (CCD) to maximise $(\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) + \mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z))$.

Multi-omic integration and use of *ad-hoc* probability distribution improve the estimation of cell types proportions.



Next step

Ongoing works

A larger comparative study:

- with new and mixed integration strategies,
- on simulations and other benchmark dataset.

- Accounting for the gene regulation network in the estimation,
- Introduction of a dependence model between expression and methylation data.

HADACA 3rd edition

→ Registration open from 01/09/2024 to 06/10/2024
(<https://hadaca3.sciencesconf.org/>)

HEALTH DATA CHALLENGE 2024

M4D
HADACA 3rd edition

Multimodal data integration to quantify tumor heterogeneity in cancer

02-06 december 2024







Centre Paul Langevin - Aussois
French Alps

3 invited speakers
data integration / deconvolution / tumor heterogeneity
40 participants expected

Registration & more details

anr® Aix-Marseille Université
FRANCE 2030
oviesan
TIMC IAB
CRICM
UGA Université Grenoble Alpes
PERSYVAL-2
RIS
CNRS

References

- 
- Jean Fan, Kamil Slowikowski, and Fan Zhang.
Single-cell transcriptomics in cancer: computational challenges and opportunities.
Experimental & Molecular Medicine, 52(9):1452–1465, 2020.
- 
- Clémentine Decamps, Alexis Arnaud, Florent Petitprez, et al.
DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification.
BMC Bioinformatics, 22(1):473, October 2021.
- 
- Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, et al.
Benchmarking of cell type deconvolution pipelines for transcriptomics data.
Nature communications, 11(1):5650, 2020.
- 
- Lana X Garmire, Yijun Li, Qianhui Huang, et al.
Challenges and perspectives in computational deconvolution of genomics data.
Nature Methods, pages 1–10, 2024.
- 
- Milan Picard, Marie-Pier Scott-Boyer, Antoine Bodein, et al.
Integration strategies of multi-omics data for machine learning analysis.
Computational and Structural Biotechnology Journal, 19:3735–3746, 2021.
- 
- Alexander J Titus, Gregory P Way, Kevin C Johnson, and Brock C Christensen.
Deconvolution of dna methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes.
Scientific reports, 7(1):11594, 2017.