Multi-omic statistical inference of cellular heterogeneity - COBICA 2024

Hugo Barbot, David Causeur, Yuna Blum, Magali Richard

Multi-omic statistical inference of cellular heterogeneity

Hugo Barbot[1], David Causeur[1], Yuna Blum[2], Magali Richard[3]

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

Cell Deconvolution

Cellular heterogeneity in a bulk:

→ refers to the variety of cell types within the bulk,

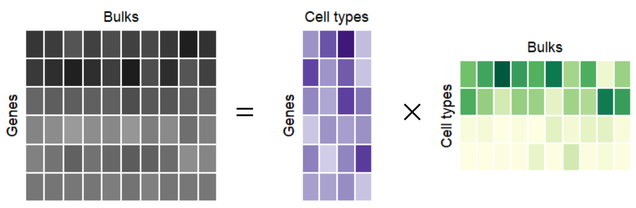
→ reflects progression of **disease state**,

→ is a **complex mixture** signal,

→ is **difficult to assess** from bulk molecular profiles.

⇒ Cell deconvolution **infers** relative abundance of cell types using **one or more -omic data** [1].

Y matrix $M \times N$ = X matrix $M \times K$ × β matrix $K \times N$



Statistical Framework

Cell deconvolution often results from Ordinary Least Squares (OLS) optimization:

$$\begin{cases} \forall i \in [\![1; N]\!] \quad Y_i = X\beta_i + \varepsilon_i, \\ \mathcal{L}(\varepsilon_i) = \mathcal{N}(0, \sigma^2 I_m). \end{cases}$$

Normality; Independence; Homoscedasticity

under constraints for each β_i $\begin{cases} \sum_{k=1}^{K} \beta_{ik} = 1, \\ 0 \leq \beta_{ik} \leq 1. \end{cases}$ Leads to a variety of algorithmic solutions.

Independence, homoscedasticity and normality are **highly questionable assumptions** in the present situation.

Problems:

1. **Intrinsic nature** of the data (counts, percentages, ...)
2. Network-based **dependence** across genes
3. Some key genes are **more influent** on deconvolution accuracy

Our approach:

1. **Respect** the inherent characteristics of biological data
2. Define a **multi-omic likelihood-based** objective function
3. Introduce gene **weights** in optimization

Benchmark Data

Models are fitted to cell deconvolution **data generated in vitro**, using cell types commonly found in pancreatic ductal adenocarcinoma (**PDAC**). A profile of 21104 gene expressions and more than 800000 CpG sites are available on $N = 30$ **independent** bulks, using $K = 9$ cell types.
The true proportions β_i of each cell type in each bulk are controlled and therefore can be assumed to be known.

⇒ The aim of those dataset is to serve as a **benchmark reference for comparison** of cell deconvolution algorithms.

Fig. 1: Heatmap of the true proportions (between 0% and 60%) of each cell types (columns) within the 30 bulks (rows).

To favor the **simultaneous** use of DNA methylation and gene expression data in the cell deconvolution task, those methylation rates are **aggregated into gene-level** measurements by averaging over all values at CpG sites in the **promoter region** of each gene.

Data integration methodology

⇒ Gene expression values are **overdispersed counts** data, so a **negative binomial** regression model is assumed.

⇒ DNA methylation values are **rates**, so a **Beta** regression model is assumed.

- In a uni-omic case, the maximisation of the log-likelihood is the objective function ($\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y)$ for gene expression, $\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ for methylation rates).
- In a multi-omic case, the maximisation of the sum of the log-likelihoods is the objective function ($\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) + \mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$).
- A Cyclic Coordinate Descent (CCD) is performed to optimize the objective function, with respect to all parameters.
- Weights ω_y and ω_z are introduce for each gene or CpG sites and aims to give **more importance** to some influential genes.

Early results

Five cell deconvolution algorithms are applied on PDAC benchmark data:

- Two OLS based methods on gene expression data, `nnls` and `rlm`, chosen in [2] and are computationally fast.
- Two negative binomial regression on gene expression data (`NBR`). The weighted variant `w-NBR` have weights set to zero for genes with low methylation, resulting from clustering of aggregated gene-level methylation rates (5833 genes).
- A multi-omic methodology with the former weighting strategy on both omic data types and maximizing the multi-omic log-likelihood (`w-NBR-Beta`).

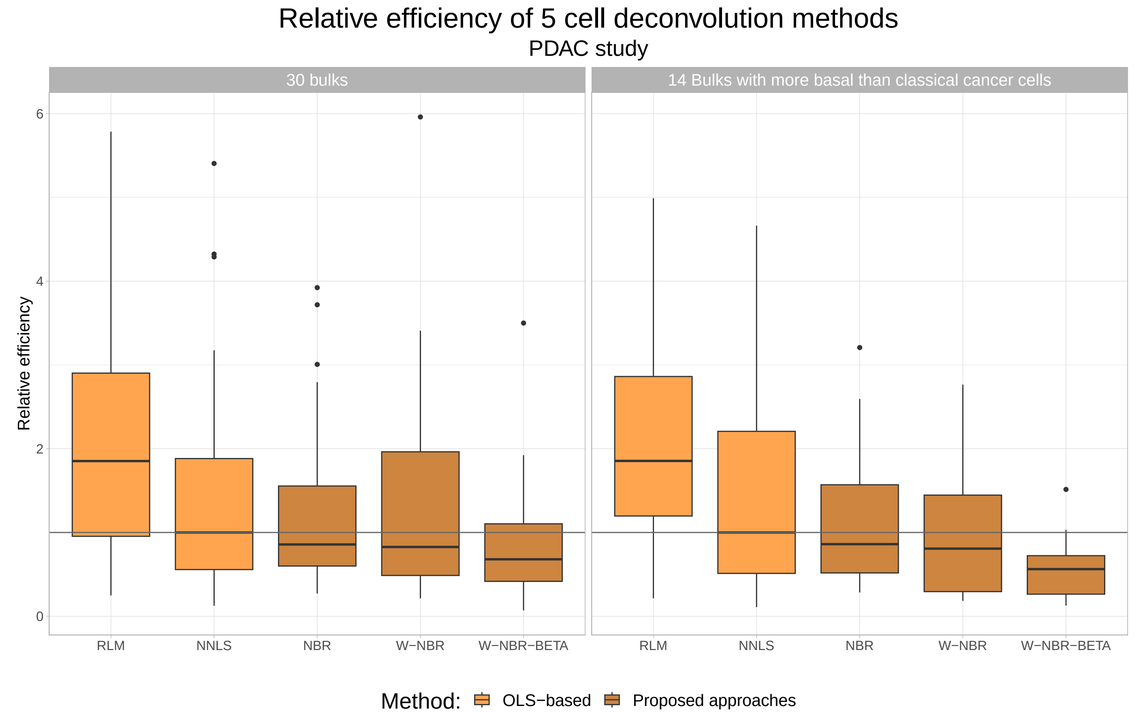


Fig. 2: Relative efficiencies of the OLS-based cell deconvolution methods and three proposed methods. the estimation accuracy metric obtained by dividing the Mean Squared Error (MSE) for each bulk by the median MSE of the `nnls` method over all bulks. Left plot: all bulks. Right plot: bulks with more basal than classical cancer cells.

Key messages

- **Integration of multi-omics data and the use of *ad-hoc* probability distribution improve the estimation of cell types proportions.**
- **This methodology can be applied to other data types, with a proper probability distribution, and to other cancer types, with a proper signature matrix and with the posibility for variable selection.**

Perspectives

Ongoing works:

A larger comparative study conducted on simulations and benchmark datasets is on progress. The goal is to identify in which condition the added value of a multi-omic approach can be expected.

Next step

- Improve the exploration of the solution space to find approximate global optimum.
- Introduction of a dependence model between gene expressions and methylation rates.

[1] 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.

[2] 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.

L'INSTITUT agro Rennes Angers

65 rue de Saint Brieuc,
35042 Rennes,
hugo.barbot@institut-agro.fr