Multi-omic statistical inference of cellular heterogeneity - JdS 2024

Hugo Barbot, David Causeur, Yuna Blum, Magali Richard

Multi-omic statistical inference of cellular heterogeneity

Hugo Barbot [1] & David Causeur [2] & Yuna Blum [3] & Magali Richard [4]

[1] *IRMAR - UMR CNRS 6625, France, hugo.barbot@institut-agro.fr*
[2] *IRMAR - UMR CNRS 6625, France, david.causeur@institut-agro.fr*
[3] *IGDR - UMR CNRS 6290, France, yuna.blum@univ-rennes.fr*
[4] *TIMC - UMR CNRS 5525, France, magali.richard@univ-grenoble-alpes.fr*

Résumé. L'hétérogénéité de la composition en types cellulaires d'échantillons biologiques est un marqueur important de la progression d'une maladie, utile pour son diagnostic. Cette composition cellulaire est cependant difficile à évaluer à partir de profils moléculaires d'un échantillon composite, la contribution de chaque type cellulaire aux signaux observés étant inconnue. La déconvolution cellulaire vise à estimer les proportions des différents types cellulaires à partir de ces profils moléculaires, plusieurs types de données omiques pouvant être utilisés dans cet objectif, tels que l'expression des gènes ou leur taux de méthylation de l'ADN. La déconvolution cellulaire s'appuie sur l'hypothèse que le profil moléculaire de l'échantillon composite peut être approché par une somme pondérée de profils moléculaires spécifiques des mêmes gênes pour chaque type cellulaire considéré, les poids étant les proportions inconnues de ces types cellulaires. La plupart des méthodes statistiques utilisées pour la déconvolution cellulaire sont basées sur des extensions de l'algorithme des moindres carrés ordinaires, sous les contraintes de positivité et de somme à un sur les coefficients du mélange. L'utilisation de cet algorithme suppose implicitement l'indépendance, l'homoscédasticité et la normalité des erreurs résiduelles, conditions sous lesquelles il offre des garanties d'optimalité. Dans le cas présent, chacune de ces trois hypothèses est discutable. D'une part, la nature intrinsèque des données omiques requiert des modèles mieux adaptés à leur sur-dispersion : l'expression des gènes par séquençage de l'ARN est par exemple une donnée de comptage et le taux de méthylation de l'ADN un pourcentage. D'autre part, la structure de dépendance induite par le réseau de régulation des gènes est très forte. Le but de ce travail est de proposer un cadre statistique respectant les caractéristiques inhérentes des données biologiques, et permettant d'intégrer plusieurs types de données omiques.

L'intégration des données multi-omiques pour la déconvolution cellulaire vise à tirer parti de points de vue complémentaires sur l'hétérogénéité cellulaire. Le cadre statistique général que nous proposons est spécialement conçu pour l'intégration de deux types de données omiques fréquemment utilisés pour la déconvolution cellulaire et cités précédemment, l'expression des gènes par séquençage de l'ARN, pour lesquels un modèle contraint de régression binomiale négative est considéré, et le taux de méthylation de l'ADN, utilisant un modèle contraint de régression pour distribution beta. Plusieurs stratégies d'optimisation simultanée sont considérées, basées sur la maximisation sous contrainte de la vraisemblance pondérée, les poids associés aux gènes étant introduits pour renforcer l'influence de certaines combinaisons spécifiques d'expressions et de taux de méthylation de l'ADN, ou sur une sélection de gènes. Une étude comparative de méthodes de déconvolution cellulaire, utilisant conjointement plusieurs types de données omiques ou non, est menée sur des données dites *benchmark*,

utilisant neuf types cellulaires communs dans les études sur le cancer du pancréas (PDAC). Les résultats confirment à la fois le gain de l'utilisation d'une approche multi-omiques et de distributions de probabilités *ad hoc* pour chaque type de données omiques. Finalement, des perspectives d'améliorations fondées sur des modèles de dépendance entre les erreurs d'approximation par les deux types de données omiques pour chaque gène sont présentées.

Mots-clés. Déconvolution cellulaire, inférence en grande dimension, Intégration de données multi-omiques, Régression

Abstract. Cellular heterogeneity in biological tissues reflects progression of disease state and is therefore useful for improved diagnostic and prognosis. Cellular composition of tissues is however difficult to assess from bulk molecular profiles, with all cells present in the tissue contributing to the recorded signals. Cell deconvolution is a common approach to unravel the heterogeneous molecular profiles observed in bulk tissues, by inferring the underlying relative abundance of individual cell types using one or more omics data, such as RNA-seq gene expressions or DNA methylation rates. So far, cellular deconvolution assumes that bulk omic profiles result from weighted sums of so-called signature cell-specific omic profiles, weights being the unknown proportions of those cell types. Consistently, most statistical methods used for cellular deconvolution are based on extensions of the Ordinary Least Squares (OLS) optimization algorithm, under nonnegativity and sum-to-one constraints on those unknown mixing coefficients. Using OLS implicitly assumes independence, homoscedasticity and normality of the residual errors, conditions under which OLS optimization guarantees optimal estimation. In cellular deconvolution applied to bulk molecular profile, all three assumptions are highly questionable. Indeed, strong violations of those assumptions may be due to the instrinsic nature of omics data, RNA-seq data being overdispersed read counts and DNA methylation rates being percentages for example, or to the dependence structure induced by the gene regulatory network, some key genes being more influent on deconvolution accuracy than others. The goal of this work is to provide a well defined statistical framework that respects the inherent characteristics of biological data for deconvolution, using multi-omic data.

Multi-omic data integration for cellular deconvolution aims at leveraging complementary viewpoints on cellular heterogeneity. The general statistical framework we propose is especially designed for integration of two frequently used omic data types for cell deconvolution mentioned previously, RNA-seq gene expression data, for which a constrained negative binomial regression model is assumed, and DNA methylation rates, using a constrained beta regression model. Many simultaneous optimization strategies are considered, either based on constrained and weighted maximum likelihood, weights being introduced to strengthen the influence of some genes based on their specific combination of signature expressions and DNA-methyation rates, or on gene selection. An extensive comparative study of cell deconvolution performance with leading single or multi-omic methods is conducted on *benchmark* data and using nine cell types commonly found in PDAC (pancreatic cancer). Results confirm both the gain in a multi-omic integration approach and in the use of *ad-hoc* probability distributions for each -omic data type. Additional improvements based on dependence models between approximation errors by the two -omic data types for each gene are finally discussed.

Keywords. Cell deconvolution, High-dimensional inference, Multi-omic data integration,

1 Introduction to single-omic cell deconvolution

The basic principles of cell deconvolution are introduced hereafter, based on a single genomic profile of gene expressions. 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. Let $y_i = (y_{i1}, \ldots, y_{im})'$ denote the complete expression profile for bulk i. The signature expressions $x_j = (x_{j1}, \ldots, x_{jK})',\ j = 1, \ldots, m$ of all genes for K cell types of interest is also available. In standard so-called supervised cell deconvolution models, the gene expression profiles are assumed to result from a linear combination of the signature expressions, the mixing coefficients of this linear decomposition being the unknown proportions $\beta_i = (\beta_{1i}, \ldots, \beta_{Ki})'$ of each cell type within each bulk, up to an additive error term. Consistently, most cell deconvolution approaches are variants of the following constrained least-squares minimization issue:

$$(\hat{\beta}_{0i}, \hat{\beta}_i') \quad = \quad \text{argmin}_{(\beta_{0i}, \beta_i')} \sum_{j=1}^{m} (y_{ij} - \beta_{0i} - x_j' \beta_i)^2,$$

where β_{0i} is an intercept and the coefficients β_i are constrained to lie within the $K-$simplex $\mathcal{S}_K = \left\{ \beta = (\beta_1, \ldots, \beta_K),\ 0 \leq \beta_k \leq 1, \sum_{k=1}^{K} \beta_k = 1 \right\}$.

The nonnegativity and sum-to-one constraints on β makes the above constrained optimization issue more challenging than its unconstrained version, which explains the variety of algorithmic solutions available for this task [1]. Moreover, variants of the least-squares objective function have been proposed, aiming for example at more robustness regarding outliers or inspired by popular machine learning methods such as penalized or support vector regression.

Choosing a least-squares type objective function is convenient since well-studied and proven unconstrained minimization algorithms can be used to inspire cell deconvolution methods incorporating nonnegativity and sum-to-one constraints. Moreover, in the maximum-likelihood estimation theory, ordinary least-squares guarantees desirable properties, such as unbiasedness and minimum variance, under the standard assumptions of the linear regression model:

$$y_i \quad = \quad \beta_{0i} \mathbf{1}_m + x \beta_i + \varepsilon_i, \tag{1}$$

where $\mathbf{1}_m$ is the $m-$vector whose entries are all equal to 1, x is the $m \times K$ signature expression matrix whose jth row is x_j and ε_i is an error term assumed to be normally distributed with mean $\mathbf{0}_m$, the $m-$vector whose all entries are zero, and positive variance-covariance matrix $\Sigma = \sigma^2 I_m$, with $\sigma > 0$. In other words, independence, homoscedasticity and normality of the residual error terms are required to guarantee optimality of unconstrained ordinary least squares estimation of the regression coefficients β_i. In the present situation, all three assumptions are highly questionable. Indeed, whereas in the standard approach of gene

expression data analysis, genes are usually considered as features measured on independent statistical units being different biological samples, in cell deconvolution methods, statistical units are genes and features are bulks. Yet, gene expressions notoriously show different levels of variability and are driven by a gene regulatory network that induces a graph-structured stochastic dependence pattern across genes. Moreover, their distribution is highly skewed, especially when expression data are read counts deduced from RNA-sequencing methods.

For an illustrative purpose, model (1) is fitted to cell deconvolution data in which a profile of $m = 21104$ gene expressions is available on $n = 30$ independent bulks and the signature matrix contains the gene expressions of those m genes in $K = 9$ cell types. The former dataset is generated with the aim of serving as a benchmark reference for comparison of cell deconvolution algorithms. Therefore, it is obtained under a strict control of the true proportions of each of the 9 cell types in the composition of each bulk. Figure 1 displays a heatmap of those true proportions, after a reordering of both cell types and bulks so that similar bulks in terms of cellular composition are grouped in clusters. The plot shows that bulks have different cellular compositions: fibroblasts are obviously dominant in all bulks and the bulks can be divided into two clusters, one with a much larger proportion of classical than basal cancer cells and the other one with more basal than classical cancer cells.

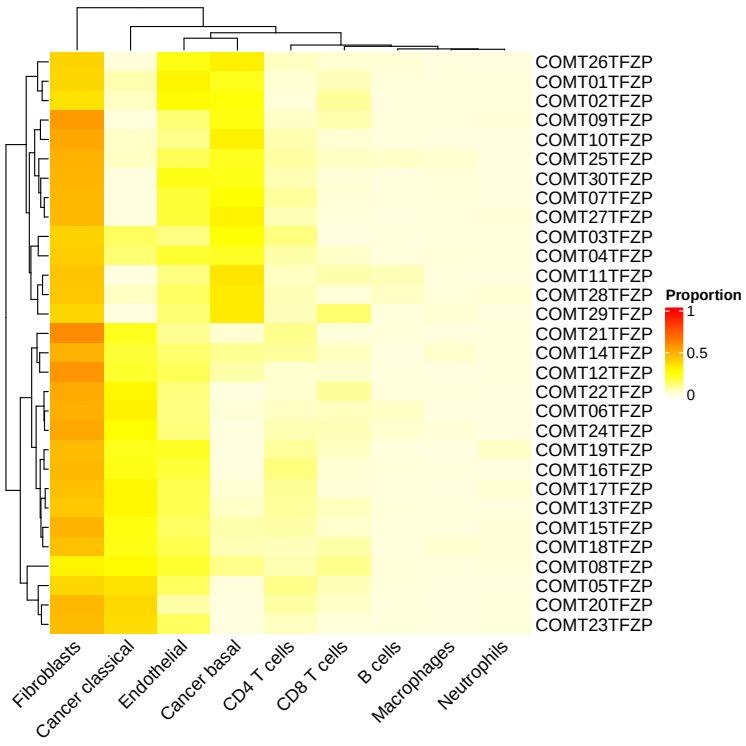

Figure 1: Heatmap of the true proportions of each cell types (columns) within the 30 bulks (rows).

Constrained least-squares approximations of the proportions are now calculated, using

the `R` package `nnls` [2], implementing a nonnegative least-squares estimation algorithm for linear regression models, widely used for cell deconvolution. For each gene, $n = 30$ values of residual errors are calculated as the differences between observed gene expressions in each bulk and linear scores of the signature expressions resulting from the `nnls` algorithm. Figure 2 displays histograms of residual standard deviations (log-transformed for a clearer visualization) and correlations. It shows both a strong heteroscedasticity and dependence across genes, with a strong imbalance between positive and negative correlations and a marked peak of correlations close to 1.

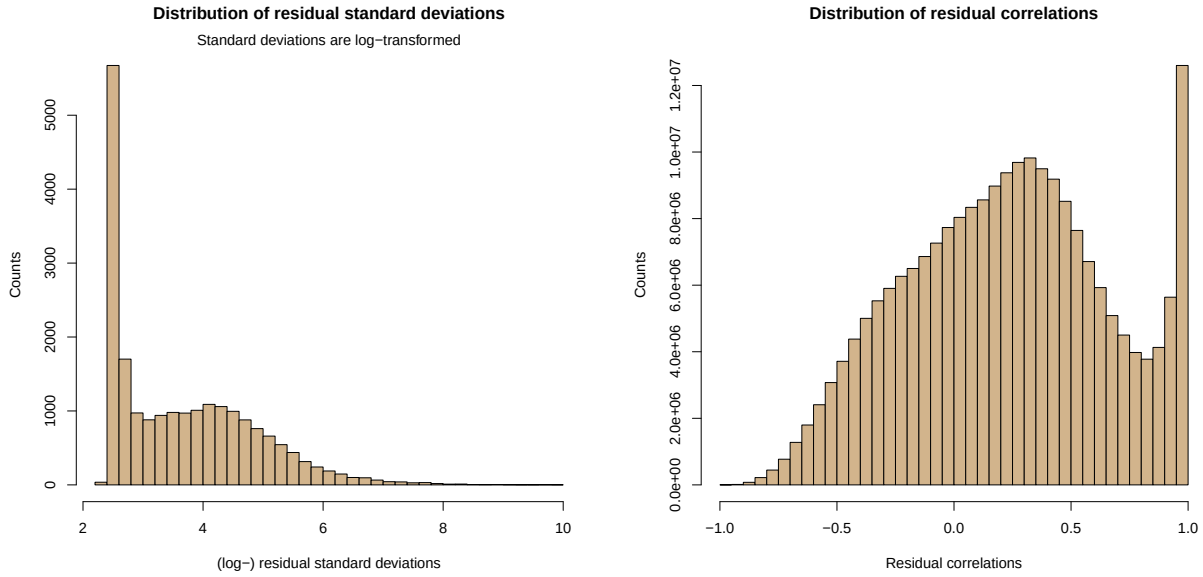

Figure 2: Histogram of log-transformed residual standard deviations (left plot) and residual correlations (right plot).

2 A general statistical framework for multi-omic cell deconvolution

Gene expression values obtained using RNA-sequencing technologies are overdispersed read counts between the start and stop codons of each gene. Most of the models used in statistical genomics for analysing such data are based on assumptions of a nonnormal distribution, either Poisson or negative binomial for the most popular.

Similarly, the following constrained negative binomial regression model [3, 4] is now assumed for Y_{ij}:

$$\mathbb{P}(Y_{ij} = y_{ij} \mid x_j) \;=\; \frac{\Gamma(y_{ij} + \frac{1}{\alpha_i})}{\Gamma(y_{ij}+1)\Gamma(\frac{1}{\alpha_i})}\Big(\frac{1}{1+\alpha_i\mu_i(x_j)}\Big)^{\frac{1}{\alpha_i}}\Big(\frac{\alpha_i\mu_i(x_j)}{1+\alpha_i\mu_i(x_j)}\Big)^{y_{ij}}, \qquad (2)$$

where $\alpha_i > 0$, $\mu_i(x_j) = \mathbb{E}(Y_{ij} \mid x_j) = \beta_{0i} + \beta_{1i}x_{j1} + \ldots + \beta_{Ki}x_{jK} > 0$, $\beta_{0i} > 0$ and $\beta_i = (\beta_{1i}, \ldots, \beta_{Ki})'$ is the vector of proportions of each cell type, with, for all $k = 1, \ldots, K$, $0 \leq \beta_{ki} \leq 1$ and $\sum_{k=1}^{K} \beta_{ki} = 1$.

Overdispersion with respect to a Poisson regression model that might result from unobserved heterogeneity within the gene expression data is accounted for by parameter α_i:

$$\mathrm{Var}(Y_{ij} \mid x_j) \;=\; \mu_i(x_j)(1 + \alpha_i\mu_i(x_j)) > \mu_i(x_j).$$

Under assumption that the gene expression values are independent given the signature expressions, the weighted log-likelihood $\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y)$ of model (2) is given below (index i for the bulk has been omitted):

$$\begin{aligned}\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) \;=\; & \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\Gamma(y_j + \frac{1}{\alpha}) - \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\Gamma(y_j + 1) - m\mathrm{log}\Gamma(\frac{1}{\alpha}) - \\ & \frac{1}{\alpha} \sum_{j=1}^{m} \omega_{yj} \mathrm{log}\big(1 + \alpha\mu(x_j)\big) + \sum_{j=1}^{m} y_j\omega_{yj} \mathrm{log}\big(\alpha\mu(x_j)\big) - \\ & \sum_{j=1}^{m} y_j \mathrm{log}\big(1 + \alpha\mu(x_j)\big),\end{aligned}$$

where the weights $\omega_y = (\omega_{y1}, \ldots, \omega_{ym})$ are positive, with $\sum_{j=1}^{m} \omega_{yj} = m$. Introducing weights for each gene in the expression of the log-likelihood aims at giving more importance to some influential genes, or even selecting subsets of active genes.

In the PDAC benchmark study introduced above, M-values of methylation levels at more than 800,000 CpG sites over the genome are also available for each of the 30 bulks and correspondingly for the 9 cell types. In order 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.

Given the $K-$profile $\tilde{x}_j = (\tilde{x}_{j1}, \ldots, \tilde{x}_{jK})'$ of signature methylation rates for gene j in the cell types of interest, it is now assumed that the M-values Z_{ij} of methylation rates in bulk i are distributed according to a Beta distribution with density [5, 6] :

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

where $\phi_i > 0$, $\mu_i(\tilde{x}_j) = \mathbb{E}(Z_{ij} \mid \tilde{x}_j) = \tilde{\beta}_{0i} + \beta_{1i}\tilde{x}_{j1} + \ldots + \beta_{pi}\tilde{x}_{jp} > 0$, $\tilde{\beta}_{0i}$ is an intercept parameter and $\beta_i = (\beta_{i1}, \ldots, \beta_{iK})'$ is the vector of proportions of each cell type, with, for all $k = 1, \ldots, K$, $0 \leq \beta_{ik} \leq 1$ and $\sum_{k=1}^{K} \beta_{ik} = 1$.

Under assumption that the M-values are independent given the signature methylation

rates, the weighted log-likelihood $\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ of the above model is given below:

$$\begin{aligned}\mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z) &= m\log\Gamma(\phi) - \sum_{j=1}^{m} \omega_{zj} \log\Gamma(\mu(\tilde{x}_j)\phi) - \sum_{j=1}^{m} \omega_{zj} \log\Gamma((1-\mu(\tilde{x}_j))\phi) + \\ & \sum_{j=1}^{m} \omega_{zj} \big(\mu(\tilde{x}_j)\phi - 1\big)\log(z_j) + \sum_{j=1}^{m} \omega_{zj} \big((1-\mu(\tilde{x}_j))\phi - 1\big)\big)\log(1-z_j).\end{aligned}$$

As above for the weighted log-likelihood of the gene expression values, gene weights $\omega_z = (\omega_{z1}, \ldots, \omega_{zm})'$ are also introduced here in order to adjust individual gene contributions to the estimation of the cell deconvolution model.

In a multi-omic data integration perspective, we propose a Cyclic Coordinate Descent (CCD) algorithm to optimize $\mathcal{L}(\alpha, \beta_0, \beta; y, x, \omega_y) + \mathcal{L}(\phi, \tilde{\beta}_0, \beta; z, \tilde{x}, \omega_z)$ with respect to all parameters, for given weights ω_y and ω_z. Initial values of the proportion parameters and intercept are obtained by any standard cell deconvolution algorithm, such as `nnls` [2] or `rlm` [7]. Those two algorithms are computationally fast, `rlm` being the most efficient for the time of execution among the three methods handling outlier with bulk data in the benchmark of Avila Cobos *et al* [8]. In order to ensure nonnegativity of estimation, at each update of an estimated proportion parameter, if the marginal maximization of the log-likelihood provides a negative update, then the current value of the proportion parameter is set to zero. Also at each update, the updated vector of proportion parameters is scaled so that it sums to one.

Weights ω_y and ω_z can be used to select genes based on their signature profiles of DNA methylation and/or expression values. Indeed, for an illustrative purpose of the former point, a standard hierarchical clustering algorithm applied on the signature profiles of $M-$values provides four clusters showing a gradient of methylation rates for all cell types: in the first cluster, genes have low methylation rates whereas, at the opposite, in the fourth cluster, they have large methylation rates. The weighting strategy considered in the comparative study reported below consists in selecting genes with low methylation rates to fit the cell deconvolution model by setting to zero all weights for genes out of the first cluster, containing 5833 genes.

3 A taste of a comparative study

In the present situation where the true proportions β_i of each cell type in each bulk are controlled and therefore can be assumed to be known, cell deconvolution methods can be compared using their Mean-Squared-Errors (MSEs) of Estimation, for each bulk over the nine cell types. Part of a large comparative study is reported below, focusing on five cell deconvolution algorithms: first, constrained least-squares approximations of the proportions are calculated for each of the 30 bulks using only the gene expression dataset of the PDAC study, with the `R` package `nnls`, implementing a nonnegative least-squares estimation algorithm for linear regression models, and function `rlm` in the `R` package `MASS`, implementing a robust estimation of a linear regression model using an M-estimator. In the latter case,

negative estimates of the proportions are set to zero and, in both cases, the resulting vector of nonnegative estimated proportions is scaled so that it sums to one. The unweighted log-likelihood of the negative binomial cell deconvolution model (2) is also maximized to provide alternative estimations (`NBR`) of the proportions of cell types using only the gene expression data. M-values in the DNA methylation rates of the PDAC study are introduced in two ways: first by a weighted variant `w-NBR` of the `NBR` algorithm, where, as mentioned previously, the weights are set to zero for genes outside of the first cluster of genes with low methylation rates, and then by combining the former weighting strategy for both omic data types and maximizing the multi-omic log-likelihood (`w-NBR-Beta`).

Figure 3 displays boxplots of an estimation accuracy metric obtained by dividing the MSE for each bulk by the median MSE of the `nnls` method over all bulks. The former relative efficiency measure is introduced in order to figure out the gain with respect to the best OLS-based method in the present study. Additionally, boxplots of the former relative efficiency measures is also provided only for the 14 bulks with more basal than classical cell types. First, it turns out that `nnls` shows better estimation accuracy than `rlm`, and is outperformed by the cell deconvolution approaches we propose. This is especially true when a selection of genes with low methylation rates is introduced in the estimation algorithm and even more when the two -omics data types are simultaneously accounted for in the estimation of the proportions of cell types. The gain in using a multi-omic approach is more obvious when the comparison is restricted to bulks for which the proportion of basal cancer cells exceeds those of basal cancer cells.

4 Perspectives

The partial results shown above of a larger comparative study we have conducted based on simulations and on the benchmark datasets of the PDAC study confirms that multi-omic approaches can improve cell deconvolution. The presentation will discuss in which conditions the added value of a multi-omic approach can be expected. Moreover, it will compare a large panel of gene weighting or selection strategies. Finally, the introduction of a dependence model between gene expressions and methylation rates within the statistical framework introduced above will be presented as a possible extension.

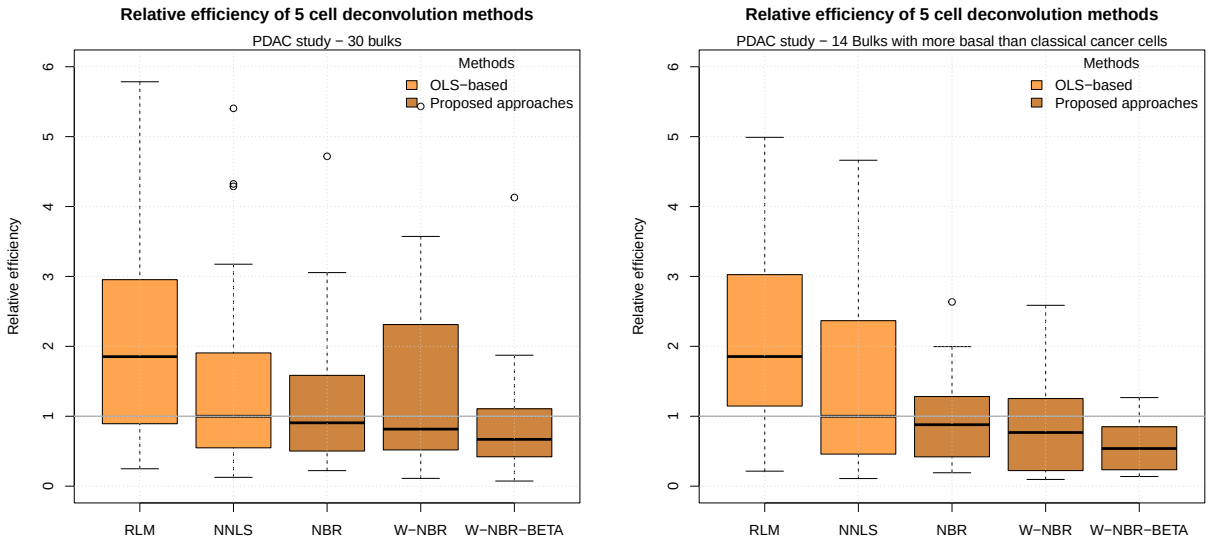

Figure 3: Relative efficiencies of the OLS-based cell deconvolution methods (`rlm` and `nnls`) and three proposed methods: unweighted negative binomial cell deconvolution (`nbr`) based on gene expressions, a weighted negative binomial cell deconvolution (`w-nbr`) also based on gene expressions with weights set to 0 for genes with large methylation rates and a weighted multi-omic (negative binomial + beta regression) cell deconvolution algorithm (`w-nbr-beta`). Left plot: all bulks. Right plot: bulks with more basal than classical cancer cells.

References

[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] Katharine M. Mullen and Ivo H. M. van Stokkum. *nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS)*, 2023. R package version 1.5.

[3] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[4] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[5] Timothy J Triche Jr, Peter W Laird, and Kimberly D Siegmund. Beta regression improves the detection of differential dna methylation for epigenetic epidemiology. *BioRxiv*, page 054643, 2016.

[6] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

[7] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[8] 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.