An updated State-of-the-Art Overview of transcriptomic Deconvolution Methods

Bastien Chassagnol, Grégory Nuel, Etienne Becht

An updated State-of-the-Art Overview of transcriptomic Deconvolution Methods

Bastien Chassagnol[1,2,*], Grégory Nuel[2], Etienne Becht[1]

1 Institut De Recherches Internationales Servier (IRIS), FRANCE
2 LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), Sorbonne Université, 4, place Jussieu, 75252 PARIS, FRANCE

*** bastien_chassagnol@laposte.net**

Abstract

Although bulk transcriptomic analyses have significantly contributed to an enhanced comprehension of multifaceted diseases, their exploration capacity is impeded by the heterogeneous compositions of biological samples. Indeed, by averaging expression of multiple cell types, RNA-Seq analysis is oblivious to variations in cellular changes, hindering the identification of the internal constituents of tissues, involved in disease progression. On the other hand, single-cell techniques are still time, manpower and resource-consuming analyses.

To address the intrinsic limitations of both bulk and single-cell methodologies, computational deconvolution techniques have been developed to estimate the frequencies of cell subtypes within complex tissues. These methods are especially valuable for dissecting intricate tissue niches, with a particular focus on tumour microenvironments (TME).

In this paper, we offer a comprehensive overview of deconvolution techniques, classifying them based on their methodological characteristics, the type of prior knowledge required for the algorithm, and the statistical constraints they address. Within each category identified, we delve into the theoretical aspects for implementing the underlying method, while providing an in-depth discussion of their main advantages and disadvantages in supplementary materials.

Notably, we emphasise the advantages of cutting-edge deconvolution tools based on probabilistic models, as they offer robust statistical frameworks that closely align with biological realities. We anticipate that this review will provide valuable guidelines for computational bioinformaticians in order to select the appropriate method in alignment with their statistical and biological objectives.

We ultimately end this review by discussing open challenges that must be addressed to accurately quantify closely related cell types from RNA sequencing data, and the complementary role of single-cell RNA-Seq to that purpose.

1 Introduction

1.1 Main sources of transcriptomic variability

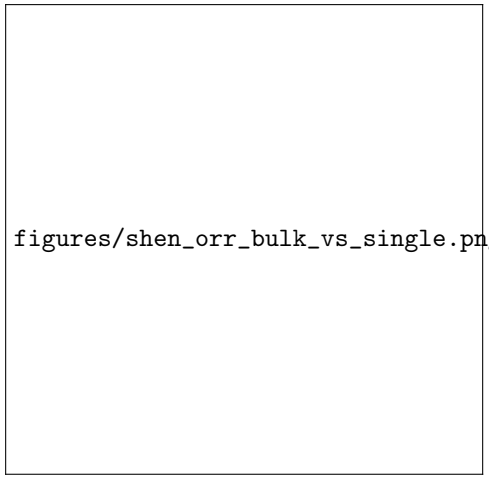
The transcriptome refers to the complete set of RNA transcripts, expressed within a biological sample. By providing a snapshot of gene expression patterns, studying its variations across phenotypical conditions provide valuable insights into the regulatory mechanisms of gene expression that underlie disease progression and individual responses to treatments.

The main biological sources of transcriptomic expression, between individuals and within tissues, proceed from three main biological factors, summarised in Figure 1: the global environmental and topic condition of the sample, encompassing disease state and tissue location; the genotype condition, involving single-nuclear polymorphisms, haplotypes, and comparable genetic aspects; and the cellular composition. Changes of cell composition are notably driven by intertwined physiological processes activating *cell motility* and *cell differentiation* mechanisms ([SG13]). In addition, the pertinent biological signal is often entangled with extraneous technical noise, requiring specific corrections in subsequent downstream analyses.

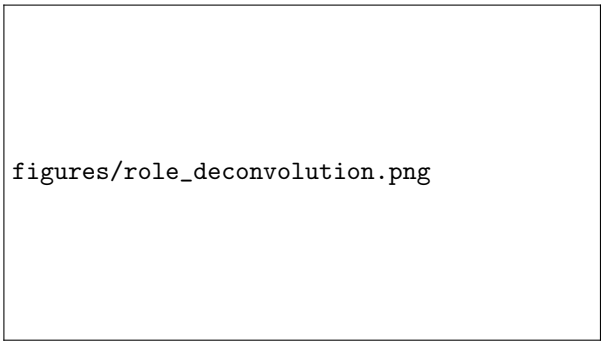
In addition, intrinsic heterogeneity is also present at the cell population level itself, arising from the presence of unspecified and infrequent population subtypes, coexistence of different developmental *cell states* (see Figure 1, bottom subfigure), or asynchronous biological processes (such as the cell cycle or circadian rhythm). Lastly, the kinetics of transcriptome regulation is inherently stochastic [Bue+15] (see Figure 1, top subfigure).

Figure 1. Main sources of transcriptomic variability, illustrated by the the intricacy of tumoral environments. The diversity of molecular profiles proceeds from a combination of intrinsic and extrinsic factors. *Intrinsic factors* encompass stochastic genetic, transcriptional, and proteomic mechanisms, while *extrinsic factors* include interactions between the resident cell populations and the surrounding microenvironment. The interconnection between these factors requires a systematic and multi-layered approach to comprehensively understand the intricacy of such biological environments. Figure reproduced from [Kas+22, Fig. 1]

While the analysis of the transcriptome through bulk RNA-Seq reveals meaningful co-expression patterns, by averaging measurements over several cell populations, it tends to ignore the intrinsic heterogeneity and complexity inherent to biological samples. Accordingly, bulk RNA-based methods are usually not able to determine whether significant changes in gene expression stem from a change of cell composition, from phenotype-induced variations or a combination of these factors ([Kuh+12]).



(a) Deconvolution methods. Physical methods for dealing with sample heterogeneity require a preliminary isolation step at the single cell level, perturbing their physical integrity. On the other hand, profiling the global heterogeneous sample directly, as performed in standard bulk RNA-Seq analyses, provides a systematic and comprehensive overview, yet, without the individual cell characteristics. To that end, computational deconvolution algorithms seem to find the sweet spot and capture simultaneously local and global information. Reproduced from [SG13, Fig .2].



(b) Changes in cell composition impact the transcriptomic expression: here, at least two distinct biological mechanisms can likely explain the increased expression of transcriptomic activity observed for a given marker gene. In the scenario (A), the cell composition is unchanged, but previously inactivated cells are stimulated and released the TF in the biological medium. In scenario (B), there is a change of cell composition, with the infiltration of a second cell type in the sample. Reproduced from [Sho+12, Fig. 1].

Hence, failure to account for changes of the cell composition is likely to result in a loss of *specificity* (genes mistakenly identified as differentially expressed, while they only reflect an increase in the cell population naturally producing them) and *sensibility* (genes expressed by minor cell populations are amenable being masked by highly variable expression from dominant cell populations), as simply illustrated in Section 1.1. Overall, the intrinsic heterogeneity of complex tissues, above all tumoral ones, reduces the robustness and reproducibility of downstream analyses, notably differential gene expression analysis or clustering of co-expression networks [1].

Various computational methodologies have emerged in recent years to estimate automatically cell type proportions in biological samples from bulk transcriptomic profiles, alleviating the high costs of single-cell RNA-Seq technologies or enabling the exploitation of archived patient datasets whose original material is not anymore available [Avi+18]. Furthermore, by requiring prior isolation of cell populations single-cell technologies hinder the analysis of interactions occurring between them. In contrast to bulk RNA-Seq and single cell methodologies, computational techniques can simultaneously capture systemic and cell-specific information, respectively. Accordingly, by dissecting the intricacy of tissues, they reveal a strong potential to identify causal drivers and provide insights on regulation

[1] [Whi+03] notably exhibits that most of the variability of gene expression in whole blood samples proceeds from relative changes of the composition in neutrophils, the most abundant immune cell type.

mechanisms.

Overview of numerical deconvolution methods

Deconvolution generally speaking names the process that consists in retrieving from a mixture its individual sub-components, popularised as the "cocktail party problem" [Che53]. In a biological sample (whole blood, tissue, ...), this consists generally in retrieving the distinct cell populations (immune, stromal...) composing it, but it can be directly extended to identify the different sources of the RNA production (for instance, many studies investigate on estimating a tumour purity score returning the proportion of malignant cells in [Yos+13]) or, at higher resolution, identify the cycle stages within a cell population (see Figure 3).

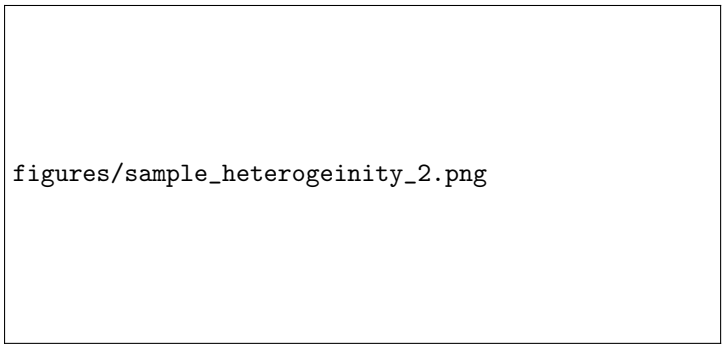


Figure 3. We detail some common applications of deconvolution methods, ordered by tier of resolution, from the least detailed resolution: *tissue* level ([QM09, Fig .1]), to the most detailed one, *cell cycles* ([LNM03, Fig .1]), through the *cell population* strata ([Fin+19a, Fig .1]).

Traditionally, deconvolution models assume that the total bulk expression is linearly related to the individual cell profiles. Precisely, they posit that the global expression can reconstructed by summing the distinct contributions of every cellular population weighed by their respective abundance within the sample (Equation (1)):

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X} \times \boldsymbol{p}_i \quad \text{matricial form} \\ y_{gi} &= \sum_{j=1}^{J} x_{gj} \times p_j \quad \text{algebraic form} \end{aligned} \tag{1}$$

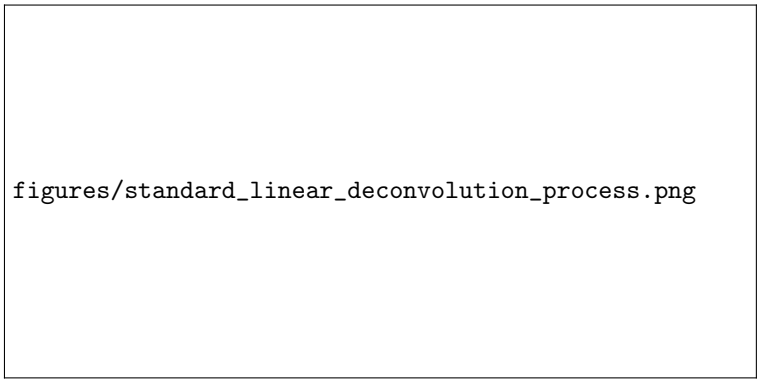
, with the following notations:

- $(\boldsymbol{y} = (y_{gi}) \in \mathbb{R}_+^{G \times N}$ is the global bulk transcriptomic expression, measured in N individuals.
- $\boldsymbol{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G \times J}}$ the signature matrix of the mean expression of G genes in J purified cell populations.
- $\boldsymbol{p} = (p_{ji}) \in]0, 1[^{J \times N}$ the unknown relative proportions of cell populations in N samples

Overall, the system includes G linear equations with J unknowns (the cellular proportions). In addition, most deconvolution problems explicitly integrate the *compositional* nature of cell ratios, enforcing in the estimation process the *unit-simplex constraint* (Equation (2)):

$$\begin{cases} \sum_{j=1}^{J} p_{ji} = 1 \\ \forall j \in \widetilde{J} \quad p_{ji} \geq 0 \end{cases} \tag{2}$$

Implicitly, Equation (2) implies that no other, unknown cell population could contribute to the measured bulk mixture. The main classes of deconvolution methods, defined on the basis of their



(a) Deconvolution principle.Infographic showing how to use the unit-simplex constraint Equation (2) and the proportional relation correlating cell populations with their respective purified transcriptomic profiles and the measured global bulk mixture Equation (1), here illustrated in the context of inferring cellular ratios, in a standard linear regression framework.

(b) Deconvolution ecosystem to disentangle complex and heterogeneous biological samples.The deconvolution methods are classified according to their input data requirements as well as the output type and resolution they provide. Supervised, alternatively named partial methods, methods utilise markers, signatures, or cytometry proportions, to achieve cell detection (A), estimating cell proportions (B), correcting heterogeneity (C), or estimating cell type-specific expression profiles (D), ranked from the simplest to the most challenging task. On the other hand, complete deconvolution methods sequentially estimate proportions from cell type-specific expression and reciprocally. They require nonetheless comprehensive prior knowledge on proportions or expression profiles (signatures, markers) and make a bench of assumptions to ensure the identifiability and consistency of the output. Reproduced from [SG13, Fig. 3].

biological objectives, are summarised in Figure 4(b), ranging from the approaches requiring the most information to the most unsupervised approaches:

In the following Section 2, we focus on *partial deconvolution* methods, that require individual cellular expression profiles to infer cell composition [Stu+04]. Besides, in the remainder of this paper, we posit, as most deconvolution algorithms, that the samples are uncorrelated with each other (independence assumption), allowing simultaneous and parallel cell ratio estimations. While this assumption reduces computational complexity, [Efr09] demonstrates cross-correlation across samples in real-world transcriptomic profiles.

2 Reference-based Approaches: Deciphering Cell Mixture through Expression Signatures

2.1 Regression-based approaches

The system of linear equations, given in Equation (1) rarely holds in practice, due to technical noise or unaccounted environmental variations. Most deconvolution algorithms model explicitly the error with a residual unobserved term, added to each individual transcriptomic measure, ϵ_g.

Subsequently, the usual approach is to retrieve the ordinary least squares (OLS) estimate which minimise the sum of squares (SSE) between predicted values fitted by the linear model: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{p}$ and the actually observed and measured values: $\boldsymbol{y}$:

$$\hat{\boldsymbol{p}}^{\text{OLS}} \equiv \arg\min_{\boldsymbol{p}} ||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2 = \arg\min_{\boldsymbol{p}} ||\boldsymbol{X}\boldsymbol{p} - \boldsymbol{y}||^2 = \sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} x_{gj} p_j \right)^2 \tag{3}$$

with $\hat{\boldsymbol{p}}$ the unknown *coefficients* to estimate, $\boldsymbol{y}$ known as the *predicted, response variable* in a linear regression context and $\boldsymbol{X}$ the *design matrix*, storing the J purified profiles. Note that the "Rouché-Capelli" theorem states that the uniqueness of a solution to Equation (3) requires that the number of genes is at least equal to the number of cell ratios to estimate (see `appendiX`). The OLS estimator,$\hat{p}_{\text{OLS}}$ is explicitly given by the *Normal equations* (see `appendiX`):

Interestingly, if we consider a generative approach, in which the error term is described by a white-*Gaussian* process (homoscedastic, null-centred), the *Gaussian-Markov* theorem (see `appendiX`) states that the OLS estimate is unique and equal to the Maximum Likelihood Estimate (see `appendiX`).

Linear modelling, whose cellular ratios are the ones returned in Equation (11), has first been used as such in [Abb+09] paper, using the lsfit function. The same method is used in [Li+16],to identify subgroups of melanomas characterised by varying levels of TCD8 subsets and correlate them with prognostic factors. To avoid accounting for tumoral cells when asserting ratios of infiltrated cells, only genes both highly correlated to the cell types of the sample and negatively correlated to the *tumour purity*, defined as the ratio of *aneuploid cells* exhibiting a non canonical number of chromosomes.

However, assumption of homoscedasticity of the residuals makes standard linear approaches sensitive to outliers, while they do not endorse explicitly the unit-simplex constraint (Equation (2)), requiring posterior normalisation of the coefficients.

2.1.1 Weighted linear approaches

The presence of an unknown cell population might be relaxed by including a constant intersection term p_0, adding in practice a column of ones in the design matrix. To account for potential heteroscedascity (variance of the errors depends on the gene value), weighted linear approaches allow users to add prior weights to modify the *leverage* (contribution) of each gene to the computation of the OLS estimate. Considering $\boldsymbol{W}$ the diagonal matrix of weights, the Weighted version of the Least Square estimate Equation (11) is given by Equation (4):

$$\hat{p}_{\text{wOLS}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{y} \tag{4}$$

EPIC [Rac+17] combines this weighted approach with the addition of a column characterising the tumour profile in the signature matrix. [Rac+17] notably provides two signatures of circulating and tumour-infiltrating immune cells, CAFs (cancer-associated fibroblasts) and epithelial cells, respectively designed for whole-blood and solid tumoral tissues, aggregating bulk and scRNA-Seq data.

Instead, the quanTIseq [Fin+19a] algorithm integrates an additional constant intersection term to quantify the contribution of the unknown tumoral content. In addition, to address the issue of cell "drop-outs" (cell populations, generally infrequent and/or exhibiting a strong correlation with other cell types, that are wrongly estimated as absent), a heuristic approach is employed whereby the final Tregs estimate is computed as the average of two Tregs measures, in the presence and absence of the TCD4+ subset in the design matrix. Tregs are indeed highly correlated with TCD4+ cell populations.

In weighted linear approaches, individual gene contributions are usually provided by the user. Without prior knowledge, the usual approach is then to give less importance to genes exhibiting strong variability within a cell population. However, assigning appropriate weights to each gene typically necessitates either prior knowledge or strong assumptions about the dataset's distribution. We subsequently review in next Section 2.1.2 robust linear regression methods that compute the weights or trim outlying gene expression in a automated manner.

2.1.2 Review of Robust Regression and SVR Methodologies for Data-Driven Transcript Feature Engineering

In the previously described approaches, the inclusion of all genes in the regression framework may yield biased estimates when the expression of some genes significantly differ, due to significant changes of sequencing protocol or phenotype condition between the bulk mixture and purified expression profiles. Unfortunately, outlying genes in least-square approaches have the strongest influence on the parameters estimation, in reason of the Euclidean metric used to evaluate the prediction error.

Several robust methods, making a compromise between *efficiency* and *robustness* of the estimate (see Appendix), have been proposed. They are usually classified into *M-estimates* (see Appendix), whereby an adaptive function is enforced on the residuals, giving less weights to those with strong leverage, and *LTS estimates*, where a user-provided ratio of aberrant genes is automatically identified and trimmed (see Appendix).

With both methods, the weights assigned to each observation depend on the estimator which in turn depend on the weights. As a result, the robust estimator must be computed sequentially, these methods are accordingly referred to as Iteratively Reweighted Least Squares (IRLS) approaches. Uniform weights are usually assigned to each observation, subsequently, a standard least regression estimate is computed. Once the OLS obtained, each observation is reweighted, using the transformation induced by the *influence function*, and which usually depends on its leverage on the regression framework. The subsequent IRLS estimates are then computed with those new weights, and the process continues until convergence [Yoh87]).

Of note, a variant of the LTS (least trimmed squares) approach has been implemented by the FARDEEP algorithm [Hao+19]. It has notably been modified to ensure convergence towards a final set of trimmed observations, in a linearly growing number of iterations. However, the algorithm is highly sensitive to the tuning parameter that controls the final number of observations trimmed during the regression. And while convergence and consistency of the algorithm is guaranteed, there's no theoretical guarantee that the final estimate returned is indeed optimal.

Overall, all the variants proposed in this section are proned to overfitting. Indeed, since these weights are derived from the model's performance, they are highly sensible to dataset-specific patterns, leading to potential inconsistent and poor results on newly observed datasets. In addition, they are less efficient than the standard OLS estimate in case the Gauss-Markov assumptions hold. For instance, the LAD estimate (see Appendix) as a relative efficiency of 0.64 compared to the OLS estimate.

Support-vector-regression are supervised machine learning algorithm featuring an alternative

strategy to select genes. It turned out that in real-world experiences, they tend to exhibit increased robustness to noisy observations. The first historical mention to SVR approach, termed ϵ-SVR [CV95], uses a insensitive loss function, whose parameter ϵ is provided by the user to control the error rate tolerated on the outputs (see Appendix).

CIBERSORT (Cell Type Identification By Estimating Relative Samples Of RNA Transcripts), developed by [New+15], utilises the the $\nu-$SVR ([CC02]) variant. Instead of optimising the precision (error rate tolerance), the ν parameter controls the proportion of Support Vectors integrated in the regression framework ([Sch+00]) [2]. Compared to standard robust linear regression approaches, [New+15] exhibits the better performance of SVR methods with "spillover effects" (see Section 5), enabling them to integrate more closely related cell types in their analysis while providing a more robust and explainable model.

In practice, CIBERSORT implements the $nu-$ SVR approach with the svm function from R package `e1071` ([Mey+21]). CIBERSORT additionally provides a standalone web application, and relevant purified signatures. The most popular is the LM22 profile, a meta transcriptomic collection of 6 studies of 22 distinct immune cell types (see Section 4.2). The ImmuCC algorithm ([Che+17]) harnesses the implementation from CIBERSORT algorithm, with a new reference signature aggregating 25 cell types and tailored for murine deconvolution.

2.1.3 Correcting the Uncoupling Between RNA and Cytometry Fractions

It appears that most of the existing deconvolution algorithms estimate the fraction of mRNA coming attributable to each cell type, rather than the underlying cell proportion itself. In other words, they assume *homogeneous* cell populations, e.g. they consider that each cell subtype exhibits the same RNA library depth ([Sos+21]). However, in real-world settings, this premise usually does not hold, for both technical and biological reasons. For instance, the RNA extraction efficiency may depend on the cell type, and its survival capacity to the lysis and extraction phase. Once the average production of total transcriptomic expression has been estimated (or phsyically measured), it becomes feasible to subsequently re-normalise the inferred cellular transcriptomic ratios, such that they align with the anticipated, biologically interpretable cellular ratios (see Equation (5)):

$$\hat{p_j^*} = K\frac{\hat{p_j}}{r_j}, \quad K = \frac{1}{\sum_{j=1}^{J} \frac{\hat{p_j}}{r_j}} \tag{5}$$

with r_j the average number of transcripts extracted per cell type, and K the normalisation constant.

Post-correction of this uncoupling is accounted in [Rac+17] and [Fin+19a] studies, with direct measures of the total expression of cell subtypes, as quantified with RNAeasy mini kit (Qiagen) and he Proteasome Subunit Beta 2, respectively [3].

When direct measures are not available, the MMAD (microarray microdissection with analysis of differences, [LHP14]) proposes an iterated approach for estimating the coefficient extraction efficiency, r_j. Yet, the regression framework is not anymore linear, and the new cellular estimate is computed using a non-linear conjugate gradient search algorithm.

2.1.4 Linear Regression Approaches with Explicit Unit-Simplex Constraint

All the previously described algorithms do not explicitly integrate the unit-simplex constraint Equation (2) during the estimation process, and re-normalise instead, posterior to the estimation, the inferred ratios.

The NNLS (Non Negative Least Squares) estimate relies on the Lawson Hanson algorithm [HH81], and its output is often provided as a reference in most review papers benchmarking deconvolution

[2] [CC02] demonstrates the equivalence between the two approaches: increasing the ν hyper-parameter results in a smaller ϵ-tube and a higher precision on the results. Asymptotically, determining the ν-proportion of support vectors reaching a given precision $\hat{\epsilon}$, is even equal to the output of the ϵ-SVR with that degree of precision.

[3] In the back-end, they utilise the expression of the *housekeeping genes* as a surrogate variable of the absolute number of transcripts produced by the cell population

algorithms ([Stu+19b], [JL21]). The nnls function from the R `limSolve` package can be used to solve this optimisation problem.

The Least Squares with Equality and Inequality Constraints (LSEI) generalises this approach by enforcing both non-negativity and sum-to-one constraints. The *lsei* function in R, from limSolve package, can be used to solve the corresponding optimisation problem. The Matlab *lsqlin* function, returning the same output as *lsei*, is used by the Bioconductor package `DeconRNASeq` ([Gon+11], [GS13]) .

Both algorithms belong to the class of *QP* (quadratic programming), which aims at optimising a system of linear, convex functions, with a guaranteed unique solution.

2.1.5 Regularised linear regression

When the number of cell types J exceeds the number of transcripts G, the deconvolution problem stated in Equation (1) is *undetermined*, with potential infinite set of solutions verifying the set of G equations. Several regularised linear approaches have been implemented to deal specifically with problems where the number of unknowns exceeds the number of variables (see appendix).

The DCQ algorithm [Alt+14] uses in particular the **Elastic Net** regularisation, a compromise between the L1 and L2 penalties proposed by the Lasso and Ridge methods. In R, the *glmnet* [Fri+11] offers a straightforward and versatile implementation of the method. The benchmark study led by [JL21] exhibits the reduced performance of deconvolution methods applying these regularised approaches. However, a comprehensive analysis of the settings used to conduct the benchmark study show that they somehow miss the point: penalised linear regression approaches are not intended to retrieve the cell ratios of a given biological sample, but rather retrieve the optimal *support* of cell populations that induce transcriptomic variations from a biological state to another. Implicitly, these methods assume that the proportions of most cell populations do not vary over time.

To illustrate the point, DCQ has been used to identify the dynamical evolution of immune cell ratios during influenza infection. Indeed, dozens of immune cell types coordinate their efforts to maintain tissue homeostasis. Precisely, DCQ studied the evolution dynamics of up to to 213 immune cell subpopulations in mice lungs for ten time points and retrieve significant changes in 70 immune cell type ratios.

Two years after, the ImmQuant package [Fri+16] offers a user-friendly tool for inferring immune cells in both human and mice organisms. The pipeline includes automatic data import and cleansing, selection of the marker genes, deconvolution of the biological samples provided and visualisation of the output.

2.2 Probabilistic-based approaches

The second family of methods for inferring cellular ratios from purified reference profiles utilises probabilistic models to capture the generative process underlying the bulk expression production. Interestingly, these approaches naturally address the unit-simplex constraint (Equation (2)), provide a more accurate representation of the discrete nature of transcript counts and can even account for an unknown cell population or individual variations of the gene expression. In particular, these approaches accurately reproduce the commonly observed correlation between the mean and the variance of the gene expression ([Lob+08]).

Since a large number of parameters might be introduced in these models, it is common practice to represent the conditional independence relating them using a directed acyclic graph and the homogenised notation illustrated in Section 2.2.2.

2.2.1 Discrete probabilistic approaches

Latent Dirichlet Allocation (LDA) is a straightforward approach to model abundances (see also [BNJ03] and appendix). The NNML (Non-negative maximum likelihood model) algorithm, by [Qia+12], extends the frequentist LDA model adopting a Bayesian approach. Precisely, the prior

distribution of the cell ratios is modelled by a symmetric Dirichlet distribution. This kind of distributions exhibits several advantages: it naturally endorses the unit-simplex constraint Equation (2) and streamlines the integration of prior knowledge, such as equibalanced hypothesis or inclusion of cytometry measures [4]

Extensions of the NNML algorithm introduce generative models that relax controversial assumption, such as the completeness (no unknown cell population) or the validity (no sample-specific variations of the purified signatures) of the reference profile. However, these probabilistic frameworks often require **regularisation** strategies, classified as "hard" and "soft" constraints, to ensure problem *identifiability*. Practical regularisation strategies often rely on strong constraints and assumptions about the distribution of purified expression profiles. They must balance the trade-off between introducing too much bias and risk overfitting, or insufficiently define the problem and suffer from *ill-conditioned* modelling.

To that end, the `ISOLATE` algorithm ([QM09]) assumes that the expression profile of any gene of the unknown cell type can be rewritten as the expression of one of the cell types already described, up to an additional multiplicative perturbation described by an uninformative Gamma prior. In a tumoral context, this constraint can be interpreted as a change of gene expression induced by heterotypic tumoral conditions, on an unique cell population subset, termed CSO in the paper (cancer site of origin). The basic framework described above has been extended in the `ISOpure` algorithm ([Quo+13]). Unlike the naive approach, `ISOpure` not only computes a shared cancer profile common across all samples but also refines it to incorporate sample-specific variations in tumoral expression. However, the CSO assumption only holds if the mutations concern only one cell line, an assumption that usually does not hold in intricate TMEs, wehre both tumoral and normal cell lines expression are impacted by the clonal growth.

Accordingly, the NNML_{np} algorithm ([Qia+12] and Section 2.2.2) assumes instead that the transcriptomic profile of the unknown cell type can be rewritten as a potential convex combination of all (possibly a subset) the included cell populations. Biologically, this approach hypothesises that the tumoral part of the sample is not a new cell line, but rather a mixture itself of the original cell populations, whose expression has been altered upon tumoral mutations, or changes induced by the new conditions of the medium. Their approach is nonetheless hindered by the stringent regularisation assumption that the perturbation factor for a given gene is the same across cell populations.

The PERT algorithm ([Qia+12] and Section 2.2.2) relaxes the strong assumption that the purified cell expression profiles are representative of the expression profiles of the mixture. Specifically, the vector representing the expression profile of a cell population is altered through a multiplicative perturbation factor $\boldsymbol{\rho}_G$, which is gene-specific and sampled from a non-informative Gamma distribution with an average value of 1.

TEMT (Transcript Estimation from Mixed Tissue samples, Section 2.2.2), by [LX13], harnesses directly the reads (sequence of nucleotides) themselves, instead of raw RNA-Seq counts. This approach enables to account for multiple transcripts resulting from *alternative splicing* (refer to Biological introduction, in the PhD manuscript) and technical biases issued from read sequencing itself [5]. The methodology is thus particularly relevant for decomposing, and correcting technical artefacts from relevant biological signal, and can be used as an alternative normalisation method (see also Appendix 1, in the PhD manuscript).

This approach uniquely incorporates technical artefacts into the deconvolution process, addressing the assumption made by other methods that input data has been corrected for such noise. Additionally, it estimates an unknown cell profile, in a process similar to the NNML_{np} approach.

The complexity of the likelihood or the posterior function requires specific optimisation methods to retrieve the relevant parameters: PERT and NNML uses a conjugate gradient descent algorithm, while

[4] To note, the Beta distribution is a variant of Dirichlet distribution with two-component mixtures, used as prior for binomial distributions.

[5] Technical artefacts in RNA-Seq encompass length, positional and amino bias. For instance, longer transcripts may yield more counts ("effective length"), while sequence-related biases include over-transcription around transcript ends.

TEMT and the ISOLATE algorithm utilise a variational online EM [DLR77]. Since diverse regularisation strategies do not address the same biological constraints, and often require different optimisation strategies, [QM09] suggests to systematically benchmark the method against manually annotated tumours, as evaluated by pathologists.

2.2.2 Continuous probabilistic approaches

The `Demix` generative model, by [Ahn+13], and its direct `DemixT` extension,by [Wan+18], infer the proportion and expression profile of the tumoral content, in a two and three-component mixture, respectively. Briefly, Demix(T) models the distribution of the bulk expression for each gene as a convolution (sum of independent variables) univariate log Normal distributions (see Section 2.2.2), each purified profile parametrised by its own parameters, inferred prior to the study. For the sake of comparison, a generative model based on a convolution of Normal distributions is also compared to the log Normal approach. This model streamlines the estimation process as a closed-form can be derived for the log-likelihood. However, the $\log_2$-transformation required to endorse the assumptions of the model is likely to disrupt the fundamental linearity deconvolution assumption (Equation (1)).

Modelling the mixture problem as a convolution offers several advantages, including the elimination of a residual error term to account for the stochasticity of the resulting bulk profile, and the utilisation of distributions that accurately depict the inherent compositional characteristics of RNA-Seq datasets.

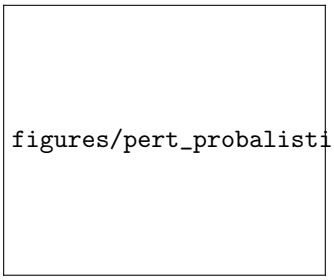
However, no explicit form for the convolution of $\log_2$-normalised variables is known, and an iterated conditional modes-like ([Bes86]) approach [6] is used to maximise the log-likelihood of the resulting generative model:

- The unknown general parameters of interest (cellular proportions and mean and variance of the tumoral profile), are determined by maximising the log-likelihood of the generative model depicting the convolution, conditioned on the previously known mean and variance for healthy cell populations. Since the closed form of the log-likelihood is not known for a convolution of log-Normal, it is approximated through numerical integration (not needed with a convolution of Normal distributions), and the MLE is obtained using a *Nelder-Mead* procedure.
- In a second time, tumoral profiles are estimated by plugging-in the parameters estimated in the previous step. With a two-component model, the unit-simplex constraint (Equation (2)) and the fundamental linear deconvolution assumption (Equation (1)), only one degree of freedom, or unknown, namely the tumoral content, must be inferred (see [Ahn+13, Eq.1]).

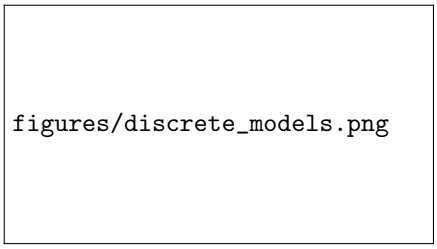
[Erk+10] implements instead a Bayesian framework, `Dsection` (see Section 2.2.2, in which the bulk expression of each gene in each sample, y_{gi}, follows a Normal distribution whose parameters are stochastic variables rather than point values. For instance, the distribution of the inverse of the variance, referred to *precision* in the paper, is modelled by a Gamma distribution.

The posterior distribution of individual cell-specific expressions and bulk gene variances is identifiable to known density distributions (*conjugate* priors). However, the posterior distribution of cellular ratios lacks a known density distribution due to the intractable integration of the normalising constant. The Metropolis-Hasting algorithm is employed to sample this posterior distribution, which is only known up to a normalising constant, while Gibbs sampling is used to retrieve simultaneously the joined posterior distributions of the whole set of parameters composing the generative model. Note that in opposition to the Demix(T) approach ([Ahn+13]), the variance of the bulk expression is uncoupled to the individual variance of the purified cellular profiles.

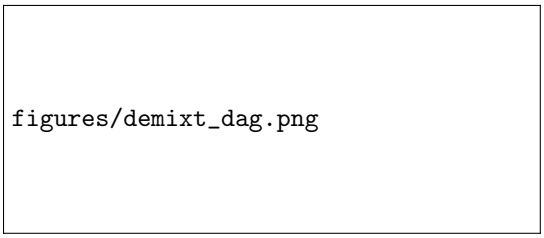
[6] The parameters are iteratively maximised, conditioned on the current updated value of the remaining subset of parameters, rather than simultaneously



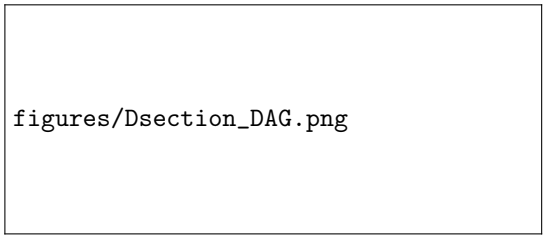
(a) Schematic of deconvolution probabilistic methods.The non-negative least squares model (NNLS) and the non-negative maximum likelihood model (NNML) can only predict proportions of prespecified reference populations. In scenario ii), the non-negative maximum likelihood new population model (NNMLnp) can additionally account for a new unobserved reference population, while in scenario iii) the perturbation model (PERT) can integrate batch or environmental tissue-specific factors using a genome-wide perturbation vector $\boldsymbol{rho}$. Reproduced from [Qia+12, Fig. 1].



(b) DAGs displaying the generative model of closely related generative deconvolution models..All the shapes and parameters with black outline are shared by any of the described probabilistic models, all derived from the ancestral method, namely LDA, better known as the "bag-of-words" method, the other colours enhancing the differences between the various approaches. NNML, NNML_{NP} and PERT were all introduced in paper [Qia+12]. The TEMT model [LX13] clearly sets apart from the others, as it accounts for, at least to normalise the purified samples, the biases introduced by commonly known technical artefacts, such as the tendency of longer reads to overcrowd the RNA library, as they provide by purely physical causes more initiation sites for the RNA polymerase.



(c) Graphical representation of the Demix(T) ([Ahn+13] and [Wan+18]) probabilistic model.



(d) Graphical representation of the Dsection [Erk+10] probabilistic model.

(e) Directed Acyclic Graph (DAG) legend.

Figure 5. Partial probabilistic models to infer cellular ratios.We follow the RevBayes convention to homogenise indexes and parameters across a set of generative models. Notably, the *likelihood* density functions describing the distribution of the observations, are in green colour while the prior distributions of the parameters to estimate are in red colour.

3 Pathway Enrichment Analysis and other Marker-Based Scoring Methods

Some deconvolution algorithms simplify the estimation process by adopting a marker-based paradigm. The definition of "markers" genes has gradually broadened, from designing genes uniquely expressed in a cell a population to include genes comprehensively expressed in one cell type relatively to other cell groups. Marker-based relied historically on strong definitions of *marker* genes ([GPT07], [CSC10]), however, nowadays, *weak* markers approaches are favoured (markers are only required to be consistently over-expressed in a given cell population), since they also enable to delineate closely related cell types.

These markers can be derived through either knowledge-driven approaches ([Ang+15], [Roo+15]) or data-driven methods [CZS15], [Bec+16], [Zha+17]. The initial data-driven strategy for identifying marker genes involved identifying genes whose mean expression value in a give cell population consistently exceeded the expression value measured across other cell types ([Sho+12], [CZS15]). More robust statistical approaches, evaluating the relevance of selected markers through the computing of empirically estimated$p$-values, have been developed since then, ranging from SNR (signal-to-noise) ratios [Bec+16], to the F-statistic ([Wan+10]) through the Gini index ([Zha+17]).

Integrating the definition of a gene marker into the fundamental presumption of linear deconvolution simplifies framework Equation (1)) into Equation (6):

$$
\begin{aligned}
y_{\forall g \in \widetilde{G_j}} &= \sum_{j^{'}=1}^{J} x_{gj'} \times p_{j'} = x_{gj} p_j, \\
&\text{since by definition } x_{gj'} = 0, \forall j^{'} \neq j \\
\begin{pmatrix} \boldsymbol{y}_{\widetilde{G_1}} \\ \boldsymbol{y}_{\widetilde{G_2}} \\ \vdots \\ \boldsymbol{y}_{\widetilde{G_J}} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{x}_{\widetilde{G_1},1} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_{\widetilde{G_2},2} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{x}_{\widetilde{G_J},J} \end{pmatrix} \times \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_J \end{pmatrix}
\end{aligned}
\tag{6}
$$

with the following notations:

- $\widetilde{G} = \{1, \dots, G\}$ is the set indexing the total number of genes selected in the signature matrix (we introduce the tilde as a shorthand indicator for a set).
- $\widetilde{G_j} \subset \widetilde{G}$ is the subset of genes expressed uniquely in cell population $j \in \widetilde{J}$
- We additionally assume the unique existence of a *partition* $\widetilde{G}$, shared across samples, such that $\widetilde{G_j} \cap \widetilde{G_l} = \varnothing$, $\forall (l, j) \in \widetilde{J}$, $l \neq j$ and $\bigcup_{j=1}^{J} \widetilde{G_j} = \widetilde{G}$.
- We introduce the shorthand $\boldsymbol{y}_{\widetilde{G_j}}$ and $\boldsymbol{X}_{\widetilde{G_j},j}$ to respectively denote the measured expression of the market set $\widetilde{G_j}$ in the bulk mixture, and its respective expression in the purified cell population j.

If eq. (6) holds, the bulk expression associated to a gene marker set is proportional to the expression of the cell population associated to this marker, the multiplicative constant being the ratio associated to this cell type, p_j.

However, as already specified in Section 2, the presence of technical noise or intrinsic biological stochasticiy usually renders the system of equations inconsistent. Assuming the same framework detailed in Section 2.1, the Normal equations, outlined in appendix, give the following OLS solution (Equation (7)):

$$
\hat{p_j} = \frac{1}{|G_j|} \sum_{g \in \widetilde{G_j}} \frac{y_g}{x_{gj}} \tag{7}
$$

with $|G_j|$ the module, namely the number of genes composing the marker set of a cell population.

Once specific markers for each population have been identified, the estimation of cellular ratios relies either on *abundance score* (see Section 3.1) or *enrichment score* (see Section 3.2 and Section 3.3).

3.1 Abundance scores

Historical endeavours, by [GPT07] and [CSC10], assume the strong definition of a marker (section 3) holds, and the cellular ratios that were returned correspond to the estimates given in eq. (7). [CSC10] only differed by the addition of a *link function*, precisely a *log2* transformation to reduce the noise bias associated to small ratio values, applied to the bulk and purified profiles.

Later, the MCP (Micro-environment Cell Populations)-counter, by [Bec+16], adopts a weak marker paradigm, and replaces the abundance score given in Equation (7), by the geometric mean of the genes characterising a given cell population (eq. (8)):

$$\mathrm{ES}(\widetilde{G_j} \in \widetilde{G}) = \left(\prod_{g \in \widetilde{G_j}} y_j \right)^{1/|\widetilde{G_j}|} \propto p_j \tag{8}$$

3.2 Enrichment scores, based on KS metric

Most of the methods computing an enrichment score rely on a variant of the weighted enrichment-based method named ssGSEA, for single-sample gene set enrichment analysis ([Sub+05] and [Bar+09]). The computation of enrichment scores, based on the Kolmogorov–Smirnov metric, is reported in appendix, while its main limitations.

[Yos+13] implements the ESTIMATE metric to compute immune and stromal enrichment scores in tumoral samples. Te best link function coupling the purity score (proportion of tumoral cells) with the ESTIMATE measure was computed with the https://en.wikipedia.org/wiki/Eureqa software. [ASB15] implements an extension of this method integrating orthogonal modalities. Precisely, the tumour purity score is computed from four distinct sources: the ESTIMATE score itself , ABSOLUTE (quantify the proportion of cancer cells based on the number and location of somatic copy-number mutations), LUMP (correlation between the degree of methylation and the tumour proportion) and immunehistochemistry image analysis.

[Roo+15] and [Ang+15] uses GSEA-based metrics to compute the tumoral activity and relate it to mechanisms involved in immune tumour resistance. [Ang+15] notably demonstrates the co-existence of two kinds of tumoural environments, distinguishing hypermutated tumours showing upregulation of immunoinhibitory molecules from non-hypermutated and stagnant tumours, enriched with immunosuppressive cells.

[Şen+16] infers gene markers for 24 distinct cell populations in 19 cancer types. With these enrichment scores, they demonstrate that the over-expression of Th17, CD8+ and Tregs increases chances of survival, while strong activity of Th2 cells is correlated with a negative prognostic.

Ultimately, the `xCell` algorithm, by [AHB17], claims to identify up to 64 distinct cell types, including immune and stromal ones, derived from a compendium of 1822 purified transcriptomic cell lines. *Calibration*, using a power link function to couple abundance scores with true cell ratios, and reduction of the multi-collinearity of the signature matrix to avoid "spillover" effects, underlie the originality, and robustness of the method.

Finally, TIminer, by [Tap+17], is a free Docker pipeline, aggregating the marker sets of [AHB17], [Ang+15] and [Cha+17]. It was initially designed for estimating the proportion of infiltrated immune cell types, along with neoantigen prediction and tumour immunogenicity.

3.3 Enrichment scores, based on alternative metrics

We present alternative strategies for calculating enrichment scores, emphasising that any method capable of comparing two distributions could be utilised for this purpose (for a theoretical definition of

these methods, report to appendix).

[BUK11] implements SPEC (Subset Prediction from Enrichment Correlation) to predict which cell population is more likely to contribute to an observed change in the gene expression, based on Pearson correlation. SPEC notably demonstrates that the main resistance mechanism of the gold-standard treatment against Hepatis C was the cross-interaction between the myeloid cells and the anti-interferon therapy.

[Sho+12] uses the z-score (negative log10 of p-value), resulting from a Fisher's exact test (see appendix).

The Bioconductor package BioQC, by [Zha+17], computes abundance scores by evaluating the relevance of median differential expressions with a non-parametric *Wilcoxon-Mann-Whitney* test.

In conclusion, marker-based methodologies provide abundance scores that are only proxy of relative cellular ratios. [AHB17] and [Yos+13] attempt to mitigate this issue, by learning a link function coupling these two features. Overall, these restrictions render marker-based methods impractical for intra-sample comparisons, in contrast to the signature-based methods discussed in previous Section 2.

4 Reference-Free Approaches: Simultaneous Deconvolution of Cell Fractions and Purified Expression Profiles

Complete deconvolution algorithms attempt to simultaneously estimate both the proportions and the pure expression profile of cell types [SG13] from the bulk profile alone, namely minimising the following quantity (Equation (9)):

$$\left(\hat{\boldsymbol{P}}, \hat{\boldsymbol{X}}\right) = \arg\min_{\boldsymbol{P},\boldsymbol{X}} \{|\boldsymbol{Y} - \boldsymbol{X} \times \boldsymbol{P}|\} \quad \boldsymbol{Y} \in \mathbb{R}_+^{G\times N},\ \boldsymbol{X} \in \mathbb{R}_+^{G\times J},\ \boldsymbol{P} \in \mathbb{R}_+^{J\times N} \tag{9}$$

Without further information, the system of equations described in Equation (9) is *undetermined*, having either an infinite set of solutions or no one at all. Hence, the identifiability of the unsupervised deconvolution problem require strong assumptions on the distribution.

4.1 Unsupervised approaches

[Ven+01] proposes the first version of a reference-free approach, inspired from Gaussian mixtures, to deconvolve colon cancer samples, from which two clusters, on a total of four identified, could be labelled with strong evidence as hematopoietic and fibroblast cells. [Ven+01] also demonstrates that the marker-based assumption (see Section 3) is a necessary condition for the existence and uniqueness of the system of equations (Equation (9)).

Repsilber and colleagues then extended the method proposed by [Ven+01], by solving Equation (9) using a Non-Negative Matrix Factorisation algorithm. NMF notably guarantees that both $\boldsymbol{X}$ and $\boldsymbol{P}$ are strictly non-negative (see details in appendix and [Rep+10]), as reported in Equation (10):

$$\begin{aligned} &\min_{\boldsymbol{P},\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{X}\|_F^2 \\ &\text{subject to the non-negativity constraints:} \\ &\boldsymbol{P} \geq 0,\ \boldsymbol{X} \geq 0 \end{aligned} \tag{10}$$

Variants of the NMF approach were used in UNDO, by [Wan+15] and CAM, by [Wan+16], methodologies. The Convex Analysis of Mixtures (CAM) enforces both the non-negativity of the outputs returned, and the unit-simplex constraint Equation (2) for the ratios. Precisely, these convex geometry-based methods project the resulting bulk expression matrix $\boldsymbol{Y}$ into a J-dimensional *polytope*, whereby each cell population profile forms a convex hull whose vertices are the marker genes of the so-called cell population. The final set of convex solutions are the ones covering the most precisely the facets of the convex hulls derived from the bulk profile. CAMTHC, by [Che19], for Convex Analysis of

Mixtures for Tissue Heterogeneity Characterisation, and *CAMfree*, by [JL21], are both R package implementing the CAM methodology.

4.2 Semi-supervised approaches integrating prior information

Since then, semi-supervised approaches, coupling partial prior knowledge of markers associated with a cell type with numerically inferred `de-novo` molecular markers, enable to increase the identifiability of the problem by reducing the set of possible solutions. Semi-approaches directly extending [Wan+16] have been implemented in R, as packages *CAMmarker* and *CellMix* [7] The usual approach to integrate prior information is to constrain all input values of the purified expression profile to zero, except whether the gene has formally been associated with a cell population.

Closely related is the semi-CAM approach, by [Don+20]. In details, the semi-CAM approach is a two-step estimation procedure; first, it identifies the final gene partition for the deconvolution process, assigning each unlabelled gene to its most probable cell type, given the already identified marker genes. To achieve this, it enhances the k-means clustering employed by the CAMfree approach, whereby the initial centroids are the vertices covering the most the convex hulls, by incorporating known marker information into the cluster centre construction. Whenever known marker genes for partially described cell types are available, [Don+20] demonstrates that the semi-CAM method outperforms the unsupervised historical CAMfree method.

The Digital Sorting Algorithm (DSA, [Zho+13]), is another semi-supervised approach, adopting a EM-like approach. Precisely, the cellular ratios and the purified expression profiles are iteratively estimated, conditioned on the current update of the remaining parameters, until convergence. Prior information can easily be integrated as initial values for either cellular ratios or purified expression profiles. However, the identifiability of the problem still requires the marker assumption.

Overall, all the methods described in this section are much more sensitive to the quality of data provided, especially when no prior information is provided.

Outline of the Cellular Deconvolution Procedure

The estimation of the composition of a biological sample is only one of the steps composing the deconvolution framework. In the remainder of the text, we define as *pipeline* this whole process, ranging from the pre-processing and collection of purified profiles to the downstream analyses, while the term "algorithm" only refers to the estimation stage itself.

A standard cellular deconvolution pipeline typically involves the following main steps:

1. **Data Preprocessing and Marker Gene Selection:** This step (see stage 1, in Section 5.2.2) involves the formatting of gene expression profiles obtained through RNA-seq or microarray, ranging from quality control to data transformation transformation and normalisation, and the removal of unwanted batch effects induced by technical artefacts.

2. **Construction of purified signature matrices** Partial methods inferring cell ratios requires an additional step consisting of identifying and characterising a subset of genes, able to delineate all the cell populations ought to compose the mixture. This step is illustrated in Section 5.2.2, part 2.

3. **Parameter Estimation:** This step refers to the deconvolution algorithm itself (stage 3, Section 5.2.2). The type of tissue or/and organism to deconvolve along with the objective biological goal guide the final choice of the algorithm used.

[7] *CellMix*, by [Gau13], benchmarks a whole set of deconvolution methods, in particular, *ssKL* and *ssFrobenius* that solve optimisation problem Equation (10) by minimising the KullBack Leibler divergence and the Frobenius norm, respectively.

4. **Evaluate the output:** This step involves the formulation of statistical tests to assess the presence of a cell population within the sample (intra-sample comparison) or to compare two cell fractions across different biological conditions (inter-sample comparison). Surprisingly, there is a notable absence of robust and widely accepted methods proving theoretically the consistency and precision of the outputs returned by most deconvolution methods. Alternatively, it is possible to benchmark the performance of a new deconvolution algorithm against gold-standard deconvolution methods and against cytometry data.
5. **Visualisation and biological interpretation:** Ultimately, various visualisations and expert validations play a pivotal role in verifying the precision and biological relevance of the algorithm in deciphering disease mechanisms, or providing new biomarkers (see stage 4, in Section 5.2.2). All these aspects are listed comprehensively in review paper, by [Che+18], and we provide a practical example in appendix.

In this section, we notably focus on the methods used for selecting the minimal subset of genes, that best discriminate the cell populations included in the deconvolution study. Overall, they fall under the general *feature-engineering* machine-learning concept, which refers to the preprocessing stage that filters irrelevant variables before applying the model [GE03].

Precisely, partial deconvolution methods based on signature profiles (Section 2) typically employ the "one-vs-all strategy" to identify the minimal set of transcripts consistently expressed in a given cell population, compared to all others. This strategy notably aims to reduce gene expression variance within a given cell type while simultaneously maximising the variance between different cell populations. However, once concatenated, the number of identified markers is still usually intractable to perform deconvolution tasks, and the resulting signature matrix often exhibits strong multicollinearity. Thus, most partial deconvolution approaches integrate an additional step to refine the purified references, which usually enables faster computation, increases the Signal-to-Noise Ratio (SNR) and increases the robustness and reproducibility of the model.

To select the genes in a global approach, the most common approach, for models based on regression optimisation, relies on optimising the *condition number* of the final reference matrix. In short, the idea is to identify the subset of quantified genes whose combined expression in the transcriptomic expression profile has the smallest condition number (see appendix for the definition and theoretical proof of the relevance of Condition Number with a OLS approach).

5 Main Challenges in the automated quantification of cell populations from RNA sequencing data

Several benchmarks have recently been developed to compare the performances of numerical deconvolution methods in relation with the biological objective ([Stu+19b]), the preprocessing protocol chosen to normalise datasets ([Fa+20]) or the noise structure and magnitude ([JL21]).

5.1 Impact of normalisation techniques

[Fa+20] defines *data normalisation* as the set of techniques to make samples' distribution comparable, including universal scaling methods (min-max, z-score, row or column-wise). It also encompasses more specific methods, such as *TPM* or *FPKM*, to account for variations of the library size and depth. On the other hand, *data transformation* refers to the *link function* applied on raw datasets, such that the assumptions underlying the generative model hold.

[Fa+20] exhibits that *scaling* methods, such as *row scaling*, or *z-score*, which are used to smooth extreme values, decrease overall the performance of the deconvolution algorithms. In addition, [Fa+20] demonstrates that applying log-normalisation leads to suboptimal performances while the best results are reached without transforming the data, conclusions consistent to the findings from [Zho+13].

Indeed, [Hof+06] shows that the *log2* transformation, while better guaranteeing the normality requirements on the distribution of the residuals, breaks the fundamental linear assumption (Equation (1)).

[JL21] suggests to apply the same transformations on both the purified signature matrix and the bulk matrix expression, with the best performances obtained with the Transcripts Per Kilobase Million (TPM) transformation (see Appendix A, in the manuscript). [Rac+17] indeed suggests that the TPM normalisation, as a *linear mapping*, naturally enforces the unit-simplex constraint Equation (2).

Regarding the construction of a signature matrix, [Avi+18] emphasises that pre-filtering genes exhibiting the strongest differences between cell types improves the robustness and reproducibility of the algorithm. With LLS-based methods (see Section 2.1), [New+15] notably demonstrates the relevance of minimising the *condition number* of the signature matrix, by reducing its multicollinearity (see appendix).

To counterbalance technical biases induced by the transcriptomic quantification technology, either RNA-Seq or microarray, some deconvolution methodologies, such as `CibersortX` ([New+19]) propose automated batch correction effect with the ComBat function, prior to the deconvolution process. Interestingly, [JL21] demonstrates that Cibersort [New+15], CibersortX [New+19] and MuSiC [Wan+19] were less sensitive to the choice of normalisation and sequencing platform, compared to other methods benchmarked.

5.2 General guidelines for constructing the reference matrix

5.2.1 Guidelines for the Selection of Cell Populations for Profiling

Many deconvolution methods are highly sensitive to the absence of cell subtypes in the reference signature, yielding the best estimates when the reference profile faithfully represents the actual composition of the biological sample [Stu+19b].

These discrepancies, most pronounced in the absence of closely correlated or orthogonal cellular profiles, lead to the "spillover" phenomena ([SG13], [Fa+20]). For instance, [Hao+19] demonstrates substantial reduction in estimating the cellular ratios of moncotyes, when myeloid dendritic cells are not included in the reference profile, despite being truly present in the mixture.

On the other hand, *background prediction* refers to erroneous identification of a cell population as being present in a mixture. This issue is even more pronounced with marker-based methods (section 3), assuming transcriptomic markers are associated with an unique cell population.

Overall, Cibersort [New+15], CibersortX [New+19] and MuSiC [Wan+19] are the least sensitive to the presence of undescribed highly-correlated or rare cell types in the mixture ([JL21]).

5.2.2 Guidelines for Phenotype and Tissue Selection in Data Collection

To mitigate the recommendations of constructing the most representative cell signature, we should highlight that comprehensive and simultaneous estimation of the whole array of cell populations composing the mixture is usually infeasible.

Firstly, some rare cell types may remain unprofiled, in particular, tumoral profiles are complex to dissect. Tumoral microenvironments display significant variability and plasticity, characterised by distinct mutation patterns, and intra-tumour heterogeneity resulting from the joint presence of diverse tumoral subclones ([Bok+22]). In addition, somatic mutations in native cell lines may lead to the loss of certain markers, posing challenges in defining pro-metastatic immune cell subsets ([Boe+22]), especially for marker-based approaches.

TIMERtumour, by [Li+16] and *EPICabsolute*, by [Rac+17], are computational methodologies specifically tailored to quantify the level of infiltration and contamination of tumoral tissues by immune cells. Yet, none of the existing deconvolution methodologies address the intra tumoral heterogeneity, stemming from the potential presence of distinct tumoral subclones ([Yu22]).

[Rac+17] additionally pinpoints that the actual deconvolution solutions for unravelling tumoral heterogeneity are targeted towards decomposition of *solid tumours*, rather than "liquid" tumours, such as haematological malignancies (leukaemia).

Secondly, there is no unique and consistent nomenclature for identifying immune cell subsets, as translating functional insights into reliable phenotypic definitions based on protein markers is challenging ([ALH21]). We describe computational solutions in Appendix A of the PhD manuscript to integrate updated cell atlases ([Lew20]), dictionary of immunological terms ([Uni23]) and ontologies in tree-like, highly scalable structures in an automated framework.

Thirdly, it is strongly deterred to incorporate cell populations from different hierarchical levels in the analysis, as this may lead to increased multicollinearity or even violate the independence assumption between purified expression profiles. The best results are typically achieved by constructing signature matrices at the finest level of granularity, as they mitigate "dropouts" effects by better delineating closely related cell types.

In order to compute back the contributions of the parental and higher-ranked cell lines, [Stu+19b] provides the R function map_result_to_celltypes in the `immunedeconv` package, which automatically aggregates estimated descendant ratios to compute the parental fraction (or even cell lines separated by further layers of lineage).

Ultimately, bad characterisation of cell populations may stem from existing intra-variability within a cell population, which results from asynchronous dynamics, such as the coexistence of different phases of the cell cycle.

While in controlled conditions, such as cell cultures, chemical arrest or nutrient starvation can achieve synchronisation of the cell cycles [Bar+08], it becomes a challenging task when profiling living tissue [8].

Sample-specific events, such as *heterotypic* contamination (for instance, infiltrates of blood circulating immune cells, [Cha+19]), disease-induced ([Gau13]) or microenvironment dysregulations ([TPZ20]) may additionally alter the transcriptomic profiles of purified cell lines.

Accordingly, to mitigate the significant loss of performance commonly observed between artificial benchmarks and real-world conditions, it is recommended to collect purified profiles in a variety of tissues, or at least representative of the phenotype condition of the bulk profiles to deconvolve [9]. The performance of deconvolution algorithms in real conditions depends more on the representativeness of cell types profiled in the signature and environmental conditions than the choice of the regression or probabilistic framework, as discrepancies between the phenotype and tropic conditions of purified samples, compared to bulk profiles, can introduce significant bias and reduce model accuracy ([SFL20], [Cai+22]).

As a final note, we quote [Stu+19b], who believed that the "improvements made to signature matrices largely outweigh potential algorithmic improvement". We refer the reader to Section 5.2.2 providing general guidelines on the best signature to harness, with respect to the cell populations profiled.

On the contrary, [Avi+18] and [Fa+20] single-cell-based deconvolution methods, capitalising on virtually reconstructed signature profiles from scRNA-Seq data, do not show significant improvement over more classical methods based on bulk-deconvolution methods.

On average, [JL21] shows that penalized regression approaches, including Lasso, Ridge and Elastic Net approaches, the latter formally implemented in the DCQ algorithm [Alt+14], underperformed,

[8] For instance, the CD3 marker, commonly used to define T cell subsets, may exhibit variable expression levels or even be entirely absent, depending on the cell cycle phase.

[9] Unfortunately, this recommendation is rarely observed, for instance, the expression profile of eosinophils, in the LM22 signature of Cibersort ([New+15]) was solely estimated from three distinct samples, from the same cohort.

while on the contrary, standard OLS, see Section 2.1, and robust regression approaches (RLR, FARDEEP, SVR, see Section 2.1.2) partial deconvolution methods, exhibit overall the best performances.

Interesting review papers encompass the works by [Fin+19b], [Pet+18], [Avi+18] and [Bla+21].

Perspectives: the Fate of Deconvolution Algorithms with the Development of Spatial Transcriptomics and single cell RNA-Seq

5.2.3 Overview of Spatial Transcriptomics and Single-Cell RNA Sequencing

Spatial transcriptomics enables the simultaneous profiling of gene expression at a high spatial resolution *in-situ*, while preserving the global cellular layout. ST reveals notably useful to determine the general layout of cell populations within a tissue and to identify hotspots, also known as "niches" (localised microenvironments in which stem cells prevail over fully differentiated cell subtypes) [10].

However, the design of the lattice of spots in ST technologies, such as HDST [Vic+19] or Slide-Seq [Rod+19]), is constrained by physical limitations that directly alleviate the final *resolution* (namely the distance between capture spots). Hence, it is not uncommon that the mRNA collected at a given sport constitutes a mixture of cell types, rather than representing a single cell.

Thus, SRT techniques have to meet a middle ground between cellular resolution and the depth and coverage of the RNA library. For instance, approaches like SeqFISH+ ([Eng+19]) and MERFISH ([Che+15]) provide subcellular resolution but are limited in throughput. Conversely, Spatial Transcriptomics ([Stå+16]) and FISSEQ ([Lee+14]) exhibit larger coverage of the genome, yet they cannot achieve single-cell resolution sequencing and are further constrained by high detection thresholds [11].

Single-cell RNA sequencing (scRNA-Seq) provides a high-resolution view of the transcriptome, by quantifying RNA content at the single-cell level. scRNA-Seq enabled to uncover cellular heterogeneity, identify rare cell populations, and capture complex dynamic changes in gene expression, that were typically obscured in bulk RNA-Seq analysis.

However, scRNA-Seq is costly and time-consuming, making it challenging to scale up for large sample sizes. In addition, the sparse nature of scRNA-Seq outputs, resulting from "drop-outs" and the complexity of the technology, renders the analysis challenging and prone to higher technical biases and variability. Hence, going down to the single cell level, scRNA-Seq typically exhibits lower coverage and depth compared to bulk RNASeq (but still higher compared to SRT).

Coupling scRNA-Seq with spatial transcriptomic data streamlines the understanding of the mechanisms relating gene expression patterns with changes of cell populations within tissues, by bridging the advantages of both methodologies while mitigating their major limitations. However, *mismatch*, designing the discordance between the cell types inferred from expression profiles derived from single-cell RNA sequencing and SRT, is commonly observed. Mismatch usually results from pre-sequencing and post-sequencing artefacts. Pre-sequencing mismatch can stem from *sampling bias* of the tissue section (lower depth with spatial barcoding or lower access to intertwined tissue structures with HPRI) or from an artificial and ectopic stimuli perturbing the cellular expression profile (stress response, or less likely, alteration of cell phenotype due to the disruption of *in situ* spatial dynamics resulting from tissue dissociation).

[10] It is common to use the abbreviation "SRT", for Spatially Resolved Transcriptomics, when referring to the general spatial sequencing framework, in order to mitigate nomenclature confusion with the specific and corporate technology "Spatial Transcriptomics" ([Stå+16])/

[11] A minimal number of 200 mRNA molecules per cell is required to detect the expression of a transcript, excluding practically a large amount of genes involved only in specific phases of the cell cycle

5.2.4 Integrating Spatial Transcriptomics with Single-Cell RNA-Seq Data Through Deconvolution Approaches

Recent alternative to mitigate the low detection threshold of scRNA in SRT and better handle mismatch issues, involve two primary approaches: *deconvolution* algorithms and *mapping* (report to appendix).

Spatial deconvolution tools, a close synonym to *stochastic profiling* techniques, estimate the cell composition for each capture spot. While sequencing the transcriptome at the single cell level is usually infeasible in a spatial context, aggregating the expression of a random pool of cells (usually rather small, aggregating no more than a dozen of them) automatically increases the depth and coverage of the RNA library, which in turn counterbalances the intrinsic noisiness and low resolution of scRNA-Seq methods.

Spatial deconvolution algorithms usually capitalise on reference signatures obtained from single-cell RNA sequencing profiles (see section 5.2.5), instead of bulk expression. The final signature is finally computed by summing the individual cellular contributions in order to reconstitute a "pseudo-bulk" mixture.

Nonetheless, spatial deconvolution algorithms necessitate specific adjustments compared to traditional approaches, as conventional deconvolution algorithms, designed for bulk transcriptome, often yield suboptimal results when dealing with sparse expression matrices, inherent to the SRT framework ([Kle+20]). In addition, spatial deconvolution methods face similar challenges to traditional deconvolution algorithms, as they too, cannot obtain absolute estimation of cell ratios, thus limiting their applicability for meaningful intra-sample comparisons.

The most population spatial deconvolution methods encompass, ranked by analytical complexity:

- The most basic methods calculate "enrichment scores" that indicate the degree of association between an individual spatial location and a specific cell type. These scores are computed using the same techniques outlined in Section 3. For example, in Seurat, by [Kis+17], each spatial location is assigned to the cell type whose expression profile, composed of the markers within its gene set, exhibits the highest similarity.

  Taking a more advanced approach, the Multimodal Intersection Analysis (MIA, [Mon+20]) combines gene pathway information inferred from single cell RNA-Sequencing (scRNA-Seq) data with gene modules that are identified as enriched through spatial barcoding techniques.

- SPOTlight [Elo+21] and SpatialDWLS [DY21] are both regression-based models that used linear solvers to estimate cellular ratios while enforcing the unit-simplex constraint, through the non-negative least squares (NNLS) algorithm.

- *Probabilistic models*, represent the mixture as a convolution of parametric distributions whose estimated cell ratios are the MLE (alternatively the MAP whereby a prior distribution is assigned to the cell ratios) of the distribution. Stereoscope ([Kho+21], also illustrated in section 5.2.4) and Cell2location ([Kle+20]) fit the distribution with a mixture of negative binomial(NB) distributions, while Robust cell-type decomposition (RCTD, [Cab+22]) utilises Poisson distributions.

- NMF regression (NMFref) is an unsupervised algorithm used both by SlideSeq [XHB16] and SPOTLight [Gul+13] to infer simultaneously cellular ratios and individual expression profiles.

- More exotic and recent methods explore alternative ways, such as DSTG [He+20] algorithm using *mutual nearest neighbour clustering* or deep-learning methods, with Tangram [Ber+20].

To conclude, we should mention promising studies extending the investigation ability of spatial transcriptomics, by coupling high-resolution *tissue images* with *histological annotations* (cell sizes and shapes, for instance) and *SRT* data ([Lar+22]).

5.2.5 Construction of reference signatures, based on single Cell RNA-Seq profiles

On the other hand, single-cell RNA sequencing technologies empower cellular deconvolution algorithms, by enabling the derivation of signature matrices more representative of the phenotype condition.

Indeed, by capturing gene expression profiles at the single-cell level, scRNA-Seq allows better discrimination of closely related cell types, and identification of rare cell type variants, which are likely to be confused with noise using bulk RNA-Seq.

Even better, the stronger granularity of scRNA-Seq outputs enables to capture the heterogeneity within cell populations, including unravelling asynchronous states of a cell population.

5.2.6 Integrate other omic modalities

To close this discussion, we should point out that the common observation of lack of reproducibility of any deconvolution method might be mitigated by coupling scores obtained from distinct biological sources.

For instance, epigenomics (DNA methylation and CpG distribution patterns) has recently been used by *EpiDISH* ([Tes+17]) and *methyICC* ([HI19]) to deconvolve cell populations, using purified methylated profiles of cell populations. Similarly, BayesCCE [Rah+18], Edec [Onu+16], RefFreeEwas [Hou+16], and MeDeCom [Lut+17] determine both cell ratios and methylome reference profiles, but adopting a reference-free approach, leveraging on variants of the non-negative matrix factorisation (NNMF, section 4) optimisation.

`SpaDecon`, by [Col+23], is one of the most promising spatial integrated approach, coupling histological annotations with metabolic and transcriptomic activity. `34P`, by [Occ+23], even claims to be able to dissect intra-tumour heterogeneity in luminal breast cancer by integrating morphological annotations, SRT data and whole slide images to a neural network architecture.

It is hence believed that the integration of multiple biological inputs in a spatially resolved context is poised to elucidate the as-yet-unsolved biological processes conducting the spatial organisation of tissue niches, and notably the key drivers controlling the level of immune cell infiltration ([Roz+17]).

However, [Tes+17] highlighted the absence of a comprehensive benchmark comparing the deconvolution performance of transcriptomic-based versus methylation-based approaches.

To conlude this section, we mention the review papers from [Rao+21], [Lon+21], [Kre21] and [Wil+22], that describe comprehensively a whole array of methods integrating spatial transcriptomic, scRNA-Seq technologies and imagery annotations.

References

[Che53] E. Colin Cherry. “Some Experiments on the Recognition of Speech, with One and with Two Ears”. In: *The Journal of the Acoustical Society of America* (Sept. 1, 1953). ISSN: 0001-4966. DOI: 10.1121/1.1907229. URL: https://asa.scitation.org/doi/10.1121/1.1907229.

[HK70] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* (Feb. 1, 1970). ISSN: 0040-1706. DOI: 10.1080/00401706.1970.10488634. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the *EM* Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* (Sept. 1977). ISSN: 00359246. DOI: 10.1111/j.2517-6161.1977.tb01600.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1977.tb01600.x.

[HH81] Karen H. Haskell and Richard J. Hanson. "An Algorithm for Linear Least Squares Problems with Equality and Nonnegativity Constraints". In: *Mathematical Programming* (Dec. 1, 1981). ISSN: 1436-4646. DOI: 10.1007/BF01584232. URL: https://doi.org/10.1007/BF01584232.

[Rou85] Peter Rousseeuw. "Multivariate Estimation With High Breakdown Point". In: *Mathematical Statistics and Applications Vol. B* (Jan. 1, 1985). ISSN: 978-94-010-8901-2. DOI: 10.1007/978-94-009-5438-0_20.

[Bes86] Julian Besag. "On the Statistical Analysis of Dirty Pictures". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1986). ISSN: 0035-9246. URL: https://www.jstor.org/stable/2345426.

[Yoh87] Victor J. Yohai. "High Breakdown-Point and High Efficiency Robust Estimates for Regression". In: *The Annals of Statistics* (June 1987). ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176350366. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-15/issue-2/High-Breakdown-Point-and-High-Efficiency-Robust-Estimates-for-Regression/10.1214/aos/1176350366.full.

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Machine Learning* (Sept. 1, 1995). ISSN: 1573-0565. DOI: 10.1007/BF00994018. URL: https://doi.org/10.1007/BF00994018.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996). ISSN: 0035-9246. URL: https://www.jstor.org/stable/2346178.

[Fue00] Angel de la Fuente. *Mathematical Methods and Models for Economists*. In collab. with Library Genesis. Cambridge, UK ; New York, NY : Cambridge University Press, 2000. 835 pp. ISBN: 978-0-521-58529-3. URL: http://archive.org/details/mathematicalmeth00fuen.

[Sch+00] Bernhard Schölkopf et al. "New Support Vector Algorithms". In: *Neural Computation* (May 1, 2000). ISSN: 0899-7667. DOI: 10.1162/089976600300015565. URL: https://doi.org/10.1162/089976600300015565.

[Ven+01] D. Venet et al. "Separation of Samples into Their Constituents Using Gene Expression Data". In: *Bioinformatics* (June 1, 2001). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.suppl_1.S279. URL: https://doi.org/10.1093/bioinformatics/17.suppl_1.S279.

[CC02] Chang Cc and Lin Cj. "Training Nu-Support Vector Regression: Theory and Algorithms". In: *Neural computation* (Aug. 2002). ISSN: 0899-7667. DOI: 10.1162/089976602760128081. URL: https://pubmed.ncbi.nlm.nih.gov/12180409/.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* (Mar. 1, 2003). ISSN: 1532-4435.

[GE03] Isabelle Guyon and André Elisseeff. "An Introduction of Variable and Feature Selection". In: *J. Machine Learning Research Special Issue on Variable and Feature Selection* (Jan. 1, 2003). DOI: 10.1162/153244303322753616.

[LNM03] Peng Lu, Aleksey Nakorchevskiy, and Edward M. Marcotte. "Expression Deconvolution: A Reinterpretation of DNA Microarray Data Reveals Dynamic Changes in Cell Populations". In: *Proceedings of the National Academy of Sciences* (Sept. 2, 2003). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1832361100. URL: https://www.pnas.org/content/100/18/10370.

[Whi+03] Adeline R. Whitney et al. "Individuality and Variation in Gene Expression Patterns in Human Blood". In: *Proceedings of the National Academy of Sciences of the United States of America* (Feb. 18, 2003). ISSN: 0027-8424. DOI: 10.1073/pnas.252784499.

[CM04] Vladimir Cherkassky and Yunqian Ma. "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression". In: *Neural Networks: The Official Journal of the International Neural Network Society* (Jan. 2004). ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00169-2.

[Stu+04] Robert O. Stuart et al. "In Silico Dissection of Cell-Type-Associated Patterns of Gene Expression in Prostate Cancer". In: *Proceedings of the National Academy of Sciences of the United States of America* (Jan. 13, 2004). ISSN: 0027-8424. DOI: 10.1073/pnas.2536479100.

[Sub+05] Aravind Subramanian et al. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles". In: *Proceedings of the National Academy of Sciences of the United States of America* (Oct. 25, 2005). ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

[ZH05] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (2005). ISSN: 1369-7412. URL: https://www.jstor.org/stable/3647580.

[Hof+06] Martin Hoffmann et al. "Robust Computational Reconstitution – a New Method for the Comparative Analysis of Gene Expression in Tissues and Isolated Cell Fractions". In: *BMC Bioinformatics* (Aug. 4, 2006). ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-369. URL: https://doi.org/10.1186/1471-2105-7-369.

[Lam+06] Justin Lamb et al. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease". In: *Science* (Sept. 29, 2006). DOI: 10.1126/science.1132939. URL: https://www.science.org/doi/full/10.1126/science.1132939.

[RV06] PETER J. ROUSSEEUW and KATRIEN VAN DRIESSEN. "Computing LTS Regression for Large Data Sets". In: *Data Mining and Knowledge Discovery* (Jan. 1, 2006). ISSN: 1573-756X. DOI: 10.1007/s10618-005-0024-4. URL: https://doi.org/10.1007/s10618-005-0024-4.

[GPT07] Mark Gosink, Howard Petrie, and Nicholas Tsinoremas. "Electronically Subtracting Expression Patterns from a Mixed Cell Population". In: *Bioinformatics. 23* (2007).

[Lam07] Justin Lamb. "The Connectivity Map: A New Tool for Biomedical Research". In: *Nature Reviews. Cancer* (Jan. 2007). ISSN: 1474-175X. DOI: 10.1038/nrc2044.

[Bar+08] Ziv Bar-Joseph et al. "Genome-Wide Transcriptional Analysis of the Human Cell Cycle Identifies Genes Differentially Regulated in Normal and Cancer Cells". In: *Proceedings of the National Academy of Sciences* (Jan. 22, 2008). ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0704723105. URL: https://www.pnas.org/content/105/3/955.

[Lob+08] Edward K. Lobenhofer et al. "Gene Expression Response in Target Organ and Whole Blood Varies as a Function of Target Organ Injury Phenotype". In: *Genome Biology* (June 20, 2008). ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-6-r100. URL: https://doi.org/10.1186/gb-2008-9-6-r100.

[Aba+09] Luca Abatangelo et al. "Comparative Study of Gene Set Enrichment Methods". In: *BMC bioinformatics* (Sept. 2, 2009). ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-275.

[Abb+09] Alexander R. Abbas et al. "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus". In: *PloS One* (July 1, 2009). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006098.

[Bar+09] David A. Barbie et al. "Systematic RNA Interference Reveals That Oncogenic KRAS-driven Cancers Require TBK1". In: *Nature* (Nov. 5, 2009). ISSN: 1476-4687. DOI: 10.1038/nature08460.

[Efr09] Bradley Efron. "Are a Set of Microarrays Independent of Each Other?" In: *The Annals of Applied Statistics* (Sept. 2009). ISSN: 1932-6157, 1941-7330. DOI: 10.1214/09-AOAS236. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-3/issue-3/Are-a-set-of-microarrays-independent-of-each-other/10.1214/09-AOAS236.full.

[QM09] Gerald Quon and Quaid Morris. "ISOLATE: A Computational Strategy for Identifying the Primary Origin of Cancers Using High-Throughput Sequencing". In: *Bioinformatics (Oxford, England)* (Nov. 1, 2009). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp378.

[Wri09] Daniel Wright. "Ten Statisticians and Their Impacts for Psychologists". In: *Perspectives on Psychological Science* (Nov. 1, 2009). DOI: 10.1111/j.1745-6924.2009.01167.x.

[CSC10] Jennifer Clarke, Pearl Seo, and Bertrand Clarke. "Statistical Expression Deconvolution from Mixed Tissue Samples". In: *Bioinformatics (Oxford, England)* (Apr. 15, 2010). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq097.

[Erk+10] Timo Erkkilä et al. "Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues". In: *Bioinformatics* (Oct. 15, 2010). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq406. URL: https://doi.org/10.1093/bioinformatics/btq406.

[Rep+10] Dirk Repsilber et al. "Biomarker Discovery in Heterogeneous Tissue Samples -Taking the in-Silico Deconfounding Approach". In: *BMC Bioinformatics* (Jan. 14, 2010). ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-27. URL: https://doi.org/10.1186/1471-2105-11-27.

[Wan+10] Yipeng Wang et al. "In Silico Estimates of Tissue Components in Surgical Samples Based on Expression Profiling Data". In: *Cancer research* (Aug. 15, 2010). ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-10-0021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411177/.

[BUK11] Christopher R. Bolen, Mohamed Uduman, and Steven H. Kleinstein. "Cell Subset Prediction for Blood Genomic Studies". In: *BMC Bioinformatics* (June 24, 2011). ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-258. URL: https://doi.org/10.1186/1471-2105-12-258.

[Fri+11] Jerome Friedman et al. *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Version 4.1-3. 2011. URL: https://CRAN.R-project.org/package=glmnet.

[Gon+11] Ting Gong et al. "Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples". In: *PloS One* (2011). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027156.

[SL11] Richard Simard and Pierre L'Ecuyer. "Computing the Two-Sided Kolmogorov-Smirnov Distribution". In: *Journal of Statistical Software* (Mar. 9, 2011). ISSN: 1548-7660. DOI: 10.18637/jss.v039.i11. URL: https://doi.org/10.18637/jss.v039.i11.

[Kuh+12] Alexandre Kuhn et al. "Cell Population-Specific Expression Analysis of Human Cerebellum". In: *BMC Genomics* (Nov. 12, 2012). ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-610. URL: https://doi.org/10.1186/1471-2164-13-610.

[Qia+12] Wenlian Qiao et al. "PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions". In: *PLOS Computational Biology* (Dec. 20, 2012). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002838. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002838.

[Sho+12] Jason E. Shoemaker et al. "CTen: A Web-Based Platform for Identifying Enriched Cell Types from Heterogeneous Microarray Data". In: *BMC Genomics* (Sept. 6, 2012). ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-460. URL: https://doi.org/10.1186/1471-2164-13-460.

[Ahn+13] Jaeil Ahn et al. "DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data". In: *Bioinformatics (Oxford, England)* (Aug. 1, 2013). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt301.

[Gau13] R. Gaujoux. "An Introduction to Gene Expression Deconvolution and the CellMix Package A Comprehensive Framework for Gene Expression Deconvolution". In: *undefined* (2013). URL: https://www.semanticscholar.org/paper/An-introduction-to-gene-expression-deconvolution-A-Gaujoux/980b8ac01435d2faa76eb1e0bc94e0a83b27b7a3.

[GS13] Ting Gong and Joseph D. Szustakowski. "DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data". In: *Bioinformatics (Oxford, England)* (Apr. 15, 2013). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt090.

[Gul+13] Ankur Gulati et al. "Association of Fibrosis with Mortality and Sudden Cardiac Death in Patients with Nonischemic Dilated Cardiomyopathy". In: *JAMA* (Mar. 6, 2013). ISSN: 1538-3598. DOI: 10.1001/jama.2013.1363.

[LX13] Yi Li and Xiaohui Xie. "A Mixture Model for Expression Deconvolution from RNA-seq in Heterogeneous Tissues". In: *BMC bioinformatics* (2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-S5-S11.

[Quo+13] Gerald Quon et al. "Computational Purification of Individual Tumor Gene Expression Profiles Leads to Significant Improvements in Prognostic Prediction". In: *Genome Medicine* (Mar. 28, 2013). ISSN: 1756-994X. DOI: 10.1186/gm433. URL: https://doi.org/10.1186/gm433.

[SR13] Igor R. Shafarevich and Alexey O. Remizov. "Linear Equations". In: *Linear Algebra and Geometry*. Ed. by Igor R. Shafarevich and Alexey O. Remizov. Berlin, Heidelberg: Springer, 2013. ISBN: 978-3-642-30994-6. DOI: 10.1007/978-3-642-30994-6_1. URL: https://doi.org/10.1007/978-3-642-30994-6_1.

[SG13] Shai S. Shen-Orr and Renaud Gaujoux. "Computational Deconvolution: Extracting Cell Type-Specific Information from Heterogeneous Samples". In: *Current Opinion in Immunology* (Oct. 2013). ISSN: 1879-0372. DOI: 10.1016/j.coi.2013.09.015.

[Yos+13] Kosuke Yoshihara et al. "Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data". In: *Nature Communications* (Oct. 11, 2013). ISSN: 2041-1723. DOI: 10.1038/ncomms3612. URL: https://www.nature.com/articles/ncomms3612.

[Zho+13] Yi Zhong et al. "Digital Sorting of Complex Tissues for Cell Type-Specific Gene Expression Profiles". In: *BMC Bioinformatics* (Mar. 7, 2013). ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-89. URL: https://doi.org/10.1186/1471-2105-14-89.

[Alt+14] Zeev Altboum et al. "Digital Cell Quantification Identifies Global Immune Cell Dynamics during Influenza Infection". In: *Molecular Systems Biology* (Feb. 28, 2014). ISSN: 1744-4292. DOI: 10.1002/msb.134947. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4023392/.

[Lee+14] Je Hyuk Lee et al. "Highly Multiplexed Subcellular RNA Sequencing in Situ". In: *Science* (Mar. 21, 2014). DOI: 10.1126/science.1250212. URL: https://www.science.org/doi/full/10.1126/science.1250212.

[LHP14] David A. Liebner, Kun Huang, and Jeffrey D. Parvin. "MMAD: Microarray Microdissection with Analysis of Differences Is a Computational Tool for Deconvoluting Cell Type-Specific Contributions from Tissue Samples". In: *Bioinformatics (Oxford, England)* (Mar. 1, 2014). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt566.

[YYB14] Chun Yu, Weixin Yao, and Xue Bai. "Robust Linear Regression: A Review and Comparison". Apr. 24, 2014. URL: http://arxiv.org/abs/1404.6274.

[Zho+14] Quan Zhou et al. "A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing". Sept. 5, 2014. URL: http://arxiv.org/abs/1409.1976.

[Ang+15] Mihaela Angelova et al. "Characterization of the Immunophenotypes and Antigenomes of Colorectal Cancers Reveals Distinct Tumor Escape Mechanisms and Novel Targets for Immunotherapy". In: *Genome Biology* (Mar. 31, 2015). ISSN: 1474-760X. DOI: 10.1186/s13059-015-0620-6.

[ASB15] Dvir Aran, Marina Sirota, and Atul J. Butte. "Systematic Pan-Cancer Analysis of Tumour Purity". In: *Nature Communications* (Dec. 4, 2015). ISSN: 2041-1723. DOI: 10.1038/ncomms9971.

[Bue+15] Florian Buettner et al. "Computational Analysis of Cell-to-Cell Heterogeneity in Single-Cell RNA-sequencing Data Reveals Hidden Subpopulations of Cells". In: *Nature Biotechnology* (Feb. 2015). ISSN: 1546-1696. DOI: 10.1038/nbt.3102. URL: https://www.nature.com/articles/nbt.3102.

[Che+15] Kok Hao Chen et al. "RNA Imaging. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells". In: *Science (New York, N.Y.)* (Apr. 24, 2015). ISSN: 1095-9203. DOI: 10.1126/science.aaa6090.

[CZS15] Maria Chikina, Elena Zaslavsky, and Stuart C. Sealfon. "CellCODE: A Robust Latent Variable Approach to Differential Expression Analysis for Heterogeneous Cell Populations". In: *Bioinformatics (Oxford, England)* (May 15, 2015). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv015.

[Gen+15] Andrew J. Gentles et al. "The Prognostic Landscape of Genes and Infiltrating Immune Cells across Human Cancers". In: *Nature Medicine* (Aug. 2015). ISSN: 1546-170X. DOI: 10.1038/nm.3909. URL: https://www.nature.com/articles/nm.3909.

[New+15] Aaron Newman et al. "Robust Enumeration of Cell Subsets from Tissue Expression Profiles". In: *Nature methods* (Mar. 30, 2015). DOI: 10.1038/nmeth.3337.

[Roo+15] Michael S. Rooney et al. "Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity". In: *Cell* (Jan. 15, 2015). ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.12.033.

[Wan+15] Niya Wang et al. "UNDO: A Bioconductor R Package for Unsupervised Deconvolution of Mixed Gene Expressions in Tumor Samples". In: *Bioinformatics (Oxford, England)* (Jan. 1, 2015). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu607.

[Bec+16] Etienne Becht et al. "Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression". In: *Genome Biology* (Oct. 20, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1070-5. URL: https://doi.org/10.1186/s13059-016-1070-5.

[Fri+16] Amit Frishberg et al. "ImmQuant: A User-Friendly Tool for Inferring Immune Cell-Type Composition from Gene-Expression Data". In: *Bioinformatics* (Dec. 15, 2016). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw535. URL: https://doi.org/10.1093/bioinformatics/btw535.

[Hou+16] E. Andres Houseman et al. "Reference-Free Deconvolution of DNA Methylation Data and Mediation by Cell Composition Effects". In: *BMC Bioinformatics* (June 29, 2016). ISSN: 1471-2105. DOI: 10.1186/s12859-016-1140-4. URL: https://doi.org/10.1186/s12859-016-1140-4.

[Li+16] Bo Li et al. "Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy". In: *Genome Biology* (Aug. 22, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1028-7.

[Onu+16] Vitor Onuchic et al. "Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types". In: *Cell reports* (Nov. 15, 2016). ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.10.057. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5115176/.

[Şen+16] Yasin Şenbabaoğlu et al. "Tumor Immune Microenvironment Characterization in Clear Cell Renal Cell Carcinoma Identifies Prognostic and Immunotherapeutically Relevant Messenger RNA Signatures". In: *Genome Biology* (Nov. 17, 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-1092-z. URL: https://doi.org/10.1186/s13059-016-1092-z.

[Stå+16] Patrik L. Ståhl et al. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics". In: *Science (New York, N.Y.)* (July 1, 2016). ISSN: 1095-9203. DOI: 10.1126/science.aaf2403.

[Wan+16] Niya Wang et al. "Mathematical Modelling of Transcriptional Heterogeneity Identifies Novel Markers and Subpopulations in Complex Tissues". In: *Scientific Reports* (Jan. 7, 2016). ISSN: 2045-2322. DOI: 10.1038/srep18909. URL: https://www.nature.com/articles/srep18909.

[XHB16] Ling Xu, Dan He, and Ying Bai. "Microglia-Mediated Inflammation and Neurodegenerative Disease". In: *Molecular Neurobiology* (Dec. 2016). ISSN: 1559-1182. DOI: 10.1007/s12035-015-9593-4.

[AHB17] Dvir Aran, Zicheng Hu, and Atul Butte. "xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape". In: *Genome Biology* (Dec. 1, 2017). DOI: 10.1186/s13059-017-1349-1.

[Cha+17] Pornpimol Charoentong et al. "Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade". In: *Cell Reports* (Jan. 3, 2017). ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.12.019.

[Che+17] Ziyi Chen et al. "Inference of Immune Cell Composition on the Expression Profiles of Mouse Tissue". In: *Scientific Reports* (Jan. 13, 2017). ISSN: 2045-2322. DOI: 10.1038/srep40508. URL: https://www.nature.com/articles/srep40508.

[Kis+17] Vladimir Yu Kiselev et al. "SC3: Consensus Clustering of Single-Cell RNA-seq Data". In: *Nature Methods* (May 2017). ISSN: 1548-7105. DOI: 10.1038/nmeth.4236.

[Lut+17] Pavlo Lutsik et al. "MeDeCom: Discovery and Quantification of Latent Components of Heterogeneous Methylomes". In: *Genome Biology* (Mar. 24, 2017). ISSN: 1474-760X. DOI: 10.1186/s13059-017-1182-6. URL: https://doi.org/10.1186/s13059-017-1182-6.

[Rac+17] Julien Racle et al. "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data". In: *eLife* (Nov. 13, 2017). Ed. by Alfonso Valencia. ISSN: 2050-084X. DOI: 10.7554/eLife.26476. URL: https://doi.org/10.7554/eLife.26476.

[Roz+17] Orit Rozenblatt-Rosen et al. "The Human Cell Atlas: From Vision to Reality". In: *Nature* (Oct. 2017). ISSN: 1476-4687. DOI: 10.1038/550451a. URL: https://www.nature.com/articles/550451a.

[Tap+17] Elias Tappeiner et al. "TIminer: NGS Data Mining Pipeline for Cancer Immunology and Immunotherapy". In: *Bioinformatics (Oxford, England)* (June 15, 2017). DOI: 10.1093/bioinformatics/btx377.

[Tes+17] Andrew E. Teschendorff et al. "A Comparison of Reference-Based Algorithms for Correcting Cell-Type Heterogeneity in Epigenome-Wide Association Studies". In: *BMC Bioinformatics* (Feb. 13, 2017). ISSN: 1471-2105. DOI: 10.1186/s12859-017-1511-5. URL: https://doi.org/10.1186/s12859-017-1511-5.

[Zha+17] Jitao David Zhang et al. "Detect Tissue Heterogeneity in Gene Expression Data with BioQC". In: *BMC Genomics* (Apr. 4, 2017). ISSN: 1471-2164. DOI: 10.1186/s12864-017-3661-2. URL: https://doi.org/10.1186/s12864-017-3661-2.

[Avi+18] Francisco Avila Cobos et al. "Computational Deconvolution of Transcriptomics Data from Mixed Cell Populations". In: *Bioinformatics (Oxford, England)* (June 1, 2018). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty019.

[CCI18] CCIB. *Gene Enrichment Profiler*. 2018. URL: http://xavierlab2.mgh.harvard.edu/EnrichmentProfiler/help.html.

[Che+18] Binbin Chen et al. "Profiling Tumor Infiltrating Immune Cells with CIBERSORT". In: *Methods in molecular biology (Clifton, N.J.)* (2018). ISSN: 1064-3745. DOI: 10.1007/978-1-4939-7493-1_12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5895181/.

[Lee+18] Junseok Lee et al. "Ensemble Modeling for Sustainable Technology Transfer". In: *Sustainability* (July 2, 2018). DOI: 10.3390/su10072278.

[Pet+18] Florent Petitprez et al. "Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine". In: *Frontiers in Oncology* (2018). ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00390.

[Rah+18] Elior Rahmani et al. "BayesCCE: A Bayesian Framework for Estimating Cell-Type Composition from DNA Methylation without the Need for Methylation Reference". In: *Genome Biology* (Sept. 21, 2018). ISSN: 1474-7596. DOI: 10.1186/s13059-018-1513-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6151042/.

[Wan+18] Zeya Wang et al. "Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration". In: *iScience* (Nov. 2, 2018). ISSN: 2589-0042. DOI: 10.1016/j.isci.2018.10.028. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249353/.

[Cha+19] Wennan Chang et al. *ICTD: A Semi-Supervised Cell Type Identification and Deconvolution Method for Multi-Omics Data*. Dec. 5, 2019. DOI: 10.1101/426593. URL: https://www.biorxiv.org/content/10.1101/426593v3. preprint.

[Che19] L. Chen. *CAMTHC: Convex Analysis of Mixtures for Tissue Heterogeneity Characterization*. 2019. URL: https://rdrr.io/bioc/CAMTHC/.

[Eng+19] Chee-Huat Linus Eng et al. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+". In: *Nature* (Apr. 2019). ISSN: 1476-4687. DOI: 10.1038/s41586-019-1049-y. URL: https://www.nature.com/articles/s41586-019-1049-y.

[Fin+19a] Francesca Finotello et al. "Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-seq Data". In: *Genome Medicine* (May 24, 2019). ISSN: 1756-994X. DOI: 10.1186/s13073-019-0638-6. URL: https://doi.org/10.1186/s13073-019-0638-6.

[Fin+19b] Francesca Finotello et al. "Next-Generation Computational Tools for Interrogating Cancer Immunity". In: *Nature Reviews Genetics* (Dec. 2019). ISSN: 1471-0064. DOI: 10.1038/s41576-019-0166-7. URL: https://www.nature.com/articles/s41576-019-0166-7.

[Hao+19] Yuning Hao et al. "Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles Using Least Trimmed Squares". In: *PLoS computational biology* (May 2019). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006976.

[HI19] Stephanie C. Hicks and Rafael A. Irizarry. "methylCC: Technology-Independent Estimation of Cell Type Composition Using Differentially Methylated Regions". In: *Genome Biology* (Nov. 29, 2019). ISSN: 1474-7596. DOI: 10.1186/s13059-019-1827-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6883691/.

[Kor+19] Ilya Korsunsky et al. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony". In: *Nature Methods* (Dec. 2019). ISSN: 1548-7105. DOI: 10.1038/s41592-019-0619-0. URL: https://www.nature.com/articles/s41592-019-0619-0.

[New+19] Aaron M. Newman et al. "Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry". In: *Nature Biotechnology* (July 2019). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0114-2. URL: https://www.nature.com/articles/s41587-019-0114-2.

[Rod+19] Samuel G. Rodriques et al. "Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution". In: *Science* (Mar. 29, 2019). DOI: 10.1126/science.aaw1219. URL: https://www.science.org/doi/10.1126/science.aaw1219.

[Stu+19a] Tim Stuart et al. "Comprehensive Integration of Single-Cell Data". In: *Cell* (June 13, 2019). ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.05.031.

[Stu+19b] Gregor Sturm et al. "Comprehensive Evaluation of Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology". In: *Bioinformatics (Oxford, England)* (July 15, 2019). ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz363.

[Vic+19] Sanja Vickovic et al. "High-Definition Spatial Transcriptomics for in Situ Tissue Profiling". In: *Nature Methods* (Oct. 2019). ISSN: 1548-7105. DOI: 10.1038/s41592-019-0548-y. URL: https://www.nature.com/articles/s41592-019-0548-y.

[Wan+19] Xuran Wang et al. "Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference". In: *Nature Communications* (Jan. 22, 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-018-08023-x. URL: https://www.nature.com/articles/s41467-018-08023-x.

[Wel+19] Joshua D. Welch et al. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". In: *Cell* (June 13, 2019). ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.05.006.

[Yan19] Xin-She Yang. "7 - Support Vector Machine and Regression". In: *Introduction to Algorithms for Data Mining and Machine Learning*. Ed. by Xin-She Yang. Academic Press, Jan. 1, 2019. ISBN: 978-0-12-817216-2. DOI: 10.1016/B978-0-12-817216-2.00014-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128172162000144.

[Ber+20] Ludvig Bergenstråhle et al. *Super-Resolved Spatial Transcriptomics by Deep Data Fusion*. Mar. 13, 2020. DOI: 10.1101/2020.02.28.963413. URL: https://www.biorxiv.org/content/10.1101/2020.02.28.963413v2. preprint.

[CN20] Zixuan Cang and Qing Nie. "Inferring Spatial and Signaling Relationships between Cells from Single Cell Transcriptomic Data". In: *Nature Communications* (Apr. 29, 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-15968-5. URL: https://www.nature.com/articles/s41467-020-15968-5.

[Don+20] Li Dong et al. "Semi-CAM: A Semi-Supervised Deconvolution Method for Bulk Transcriptomic Data with Partial Marker Gene Information". In: *Scientific Reports* (Mar. 25, 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-62330-2. URL: https://www.nature.com/articles/s41598-020-62330-2.

[Fa+20] Cobos Fa et al. "Comprehensive Benchmarking of Computational Deconvolution of Transcriptomics Data". In: (Jan. 10, 2020). DOI: 10.1101/2020.01.10.897116. URL: https://europepmc.org/article/ppr/ppr108248.

[He+20] Bryan He et al. "Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning". In: *Nature Biomedical Engineering* (Aug. 2020). ISSN: 2157-846X. DOI: 10.1038/s41551-020-0578-x.

[Kle+20] Vitalii Kleshchevnikov et al. *Comprehensive Mapping of Tissue Cell Architecture via Integrated Single Cell and Spatial Transcriptomics*. Nov. 17, 2020. DOI: 10.1101/2020.11.15.378125. URL: https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1. preprint.

[Lew20] Julius M. Cruse Lewis Robert E. *Illustrated Dictionary of Immunology*. 3rd ed. Boca Raton: CRC Press, June 30, 2020. 816 pp. ISBN: 978-0-429-12407-5. DOI: 10.1201/9780849379888.

[Mon+20] Reuben Moncada et al. "Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas". In: *Nature Biotechnology* (Mar. 2020). ISSN: 1546-1696. DOI: 10.1038/s41587-019-0392-8. URL: https://www.nature.com/articles/s41587-019-0392-8.

[Sac+20] Pallavi Sachdev et al. "Abstract 1924: Genetic Analysis of Responses to Eribulin versus Vinorelbine and Paclitaxel in 100 Cancer Cell Lines from the Cancer Cell Line Encyclopedia (CCLE)". In: *Cancer Research* (Aug. 15, 2020). ISSN: 0008-5472. DOI: 10.1158/1538-7445.AM2020-1924. URL: https://doi.org/10.1158/1538-7445.AM2020-1924.

[SFL20] Gregor Sturm, Francesca Finotello, and Markus List. "In Silico Cell-Type Deconvolution Methods in Cancer Immunotherapy". In: *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*. Ed. by Sebastian Boegel. Methods in Molecular Biology. New York, NY: Springer US, 2020. ISBN: 978-1-07-160327-7. DOI: 10.1007/978-1-0716-0327-7_15. URL: https://doi.org/10.1007/978-1-0716-0327-7_15.

[TPZ20] Daiwei Tang, Seyoung Park, and Hongyu Zhao. "NITUMID: Nonnegative Matrix Factorization-Based Immune-TUmor MIcroenvironment Deconvolution". In: *Bioinformatics* (Mar. 1, 2020). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz748. URL: https://doi.org/10.1093/bioinformatics/btz748.

[Tra+20] Hoa Thi Nhu Tran et al. "A Benchmark of Batch-Effect Correction Methods for Single-Cell RNA Sequencing Data". In: *Genome Biology* (Jan. 16, 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1850-9.

[ALH21] Michiel C. van Aalderen, Rene A. W. van Lier, and Pleun Hombrink. "How to Reliably Define Human CD8+ T-Cell Subsets: Markers Playing Tricks". In: *Cold Spring Harbor Perspectives in Biology* (Jan. 11, 2021). ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a037747. URL: http://cshperspectives.cshlp.org/content/13/11/a037747.

[Arm+21] Erick Armingol et al. "Deciphering Cell–Cell Interactions and Communication from Gene Expression". In: *Nature Reviews Genetics* (Feb. 2021). ISSN: 1471-0064. DOI: 10.1038/s41576-020-00292-x. URL: https://www.nature.com/articles/s41576-020-00292-x.

[Bla+21] Andrea Blasco et al. "Improving Deconvolution Methods in Biology through Open Innovation Competitions: An Application to the Connectivity Map". In: *Bioinformatics* (Mar. 22, 2021). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab192. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8479655/.

[DY21] Rui Dong and Guo-Cheng Yuan. "SpatialDWLS: Accurate Deconvolution of Spatial Transcriptomic Data". In: *Genome Biology* (May 10, 2021). ISSN: 1474-760X. DOI: 10.1186/s13059-021-02362-7. URL: https://doi.org/10.1186/s13059-021-02362-7.

[Elo+21] Marc Elosua-Bayes et al. "SPOTlight: Seeded NMF Regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes". In: *Nucleic Acids Research* (May 21, 2021). ISSN: 1362-4962. DOI: 10.1093/nar/gkab043.

[JL21] Haijing Jin and Zhandong Liu. "A Benchmark for RNA-seq Deconvolution Analysis under Dynamic Testing Environments". In: *Genome Biology* (Apr. 12, 2021). ISSN: 1474-760X. DOI: 10.1186/s13059-021-02290-6. URL: https://doi.org/10.1186/s13059-021-02290-6.

[Kho+21] Combiz Khozoie et al. *scFlow: A Scalable and Reproducible Analysis Pipeline for Single-Cell RNA Sequencing Data*. preprint. Preprints, Aug. 16, 2021. DOI: 10.22541/au.162912533.38489960/v1. URL: https://www.authorea.com/users/226952/articles/480342-scflow-a-scalable-and-reproducible-analysis-pipeline-for-single-cell-rna-sequencing-data?commit=921426e3a377f7897ba262b5fc2bf0ef3680570a.

[Kre21] Ivan Krešimir Lukić. "Bioinformatics Approach to Spatially Resolved Transcriptomics". In: *Emerging Topics in Life Sciences* (Aug. 9, 2021). ISSN: 2397-8554. DOI: 10.1042/ETLS20210131. URL: https://doi.org/10.1042/ETLS20210131.

[Lon+21] Sophia K. Longo et al. "Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics". In: *Nature Reviews Genetics* (Oct. 2021). ISSN: 1471-0064. DOI: 10.1038/s41576-021-00370-8. URL: https://www.nature.com/articles/s41576-021-00370-8.

[MAL21] KAVITA MALI. *Everything You Need to Know about Linear Regression!* Analytics Vidhya. Oct. 4, 2021. URL: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/.

[Mey+21] David Meyer et al. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. Version 1.7-9. Sept. 16, 2021. URL: https://CRAN.R-project.org/package=e1071.

[Rao+21] Anjali Rao et al. "Exploring Tissue Architecture Using Spatial Transcriptomics". In: *Nature* (Aug. 2021). ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9. URL: https://www.nature.com/articles/s41586-021-03634-9.

[Sos+21] Olukayode A. Sosina et al. "Strategies for Cellular Deconvolution in Human Brain RNA Sequencing Data". In: (Aug. 4, 2021). DOI: 10.12688/f1000research.50858.1. URL: https://f1000research.com/articles/10-750.

[Boe+22] Maximilian Boesch et al. "OMIP 077: Definition of All Principal Human Leukocyte Populations Using a Broadly Applicable 14-Color Panel". In: *Cytometry Part A* (2022). ISSN: 1552-4930. DOI: 10.1002/cyto.a.24481. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.24481.

[Bok+22] A. A. Bokil et al. "32P Discovering Markers That Identify Pro-Metastatic Immune Cell Subsets". In: *Immuno-Oncology and Technology*. Abstract Book of the ESMO Immuno-Oncology Congress 2022 7-9 December 2022 (Dec. 1, 2022). ISSN: 2590-0188. DOI: 10.1016/j.iotech.2022.100137. URL: https://www.sciencedirect.com/science/article/pii/S2590018822000685.

[Cab+22] Dylan M. Cable et al. "Robust Decomposition of Cell Type Mixtures in Spatial Transcriptomics". In: *Nature Biotechnology* (Apr. 2022). ISSN: 1546-1696. DOI: 10.1038/s41587-021-00830-w. URL: https://www.nature.com/articles/s41587-021-00830-w.

[Cai+22] Manqi Cai et al. "Robust and Accurate Estimation of Cellular Fraction from Tissue Omics Data via Ensemble Deconvolution". In: *Bioinformatics* (May 26, 2022). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac279. URL: https://doi.org/10.1093/bioinformatics/btac279.

[Kas+22] Aditya Kashyap et al. "Quantification of Tumor Heterogeneity: From Data Acquisition to Metric Generation". In: *Trends in Biotechnology* (June 1, 2022). ISSN: 0167-7799, 1879-3096. DOI: 10.1016/j.tibtech.2021.11.006. URL: https://www.cell.com/trends/biotechnology/abstract/S0167-7799(21)00267-5.

[Lar+22] Ludvig Larsson et al. "SnapShot: Spatial Transcriptomics". In: *Cell* (July 21, 2022). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2022.06.002. URL: https://www.cell.com/cell/abstract/S0092-8674(22)00707-3.

[Wil+22] Cameron G. Williams et al. "An Introduction to Spatial Transcriptomics for Biomedical Research". In: *Genome Medicine* (June 27, 2022). ISSN: 1756-994X. DOI: 10.1186/s13073-022-01075-1. URL: https://doi.org/10.1186/s13073-022-01075-1.

[Yu22] Xiaoqing Yu. "Estimation of Tumor Immune Signatures from Transcriptomics Data". In: *Handbook of Statistical Bioinformatics*. Ed. by Henry Horng-Shing Lu et al. Springer Handbooks of Computational Statistics. Berlin, Heidelberg: Springer, 2022. ISBN: 978-3-662-65902-1. DOI: 10.1007/978-3-662-65902-1_16. URL: https://doi.org/10.1007/978-3-662-65902-1_16.

[Ber+23] Matthew N. Bernstein et al. *Monkeybread: A Python Toolkit for the Analysis of Cellular Niches in Single-Cell Resolution Spatial Transcriptomics Data*. Sept. 15, 2023. DOI: 10.1101/2023.09.14.557736. URL: https://www.biorxiv.org/content/10.1101/2023.09.14.557736v1. preprint.

[Col+23] Kyle Coleman et al. "SpaDecon: Cell-Type Deconvolution in Spatial Transcriptomics with Semi-Supervised Learning". In: *Communications Biology* (Apr. 7, 2023). ISSN: 2399-3642. DOI: 10.1038/s42003-023-04761-x. URL: https://www.nature.com/articles/s42003-023-04761-x.

[Occ+23] N. Occelli et al. "34P Investigating Morphological Heterogeneity in Luminal Breast Cancer Integrating Artificial Intelligence and Spatial Transcriptomics". In: *ESMO Open* (May 1, 2023). ISSN: 2059-7029. DOI: 10.1016/j.esmoop.2023.101258. URL: https://www.esmoopen.com/article/S2059-7029(23)00484-2/fulltext.

[Uni23] Japan University. *Encyclopedia of Immunology*. 2023. URL: https://rnavi.ndl.go.jp/mokuji_html/000003269982.html.

[Xu+23] Xinlan Xu et al. "Short Text Classification Based on Hierarchical Heterogeneous Graph and LDA Fusion". In: *Electronics* (Jan. 2023). ISSN: 2079-9292. DOI: 10.3390/electronics12122560. URL: https://www.mdpi.com/2079-9292/12/12/2560.

[ZR23] Kangmei Zhao and Seung Yon Rhee. "Interpreting Omics Data with Pathway Enrichment Analysis". In: *Trends in Genetics* (Apr. 1, 2023). ISSN: 0168-9525. DOI: 10.1016/j.tig.2023.01.003. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(23)00018-5.

Appendix, describing relevant statistical concepts

Linear regression and Gauss-Markov theorem

If relation Equation (1) holds perfectly (no additional technical noise nor stochastic transcriptomic expression, no additional cell content, ...), the number of solutions is given by the "Rouché-Capelli" theorem [SR13], detailed in theorem .1.

Theorem .1: Rouché-Capelli theorem

The number of solutions for a system of linear equations depend on both the rank of its augmented matrix with respect to its coefficient matrix, and the number of unknowns with respect to the number of equations:

- *Uniqueness of the solution generally implies that the number of genes G is equal to the number of cell types studied J, and that the expression of any given individual gene can not be rewritten as the linear combination of other genes expressed within the sample (matricially, this implies that the reference signature, $\boldsymbol{X}$, also termed as the* coefficient matrix*, is invertible).*
- *If the number of equations is less than the number of unknowns, in other words the number of rows of the coefficient matrix is inferior to the number of columns, then in most cases an infinite number of equally probable solutions hold, rendering the system* undetermined *and consequently irrelevant. In practice, with the development of efficient sequencing techniques able to quantify simultaneously the expression of thousands of genes, this situation is rarely encountered.*
- *When the number of genes exceeds those of cell populations, the system is said* overdetermined*, and the existence of a solution to the corresponding system of equations, requires collinear redundancy in the coefficient matrix (in other words, the information from at least one of the G equations can be rewritten as a linear combination of the others (summing lines over or/and multiplying them by real constants). Otherwise, the system is* inconsistent *(alternatively* degenerate*) if there is no set of solutions that satisfies simultaneously all the G equations.*

All these statements can be encompassed into the following more general statement: a system of linear equations with J unknowns has a solution if and only if the rank *(dimension of the space spanned by its columns) of its coefficient matrix, here $\boldsymbol{X}$, is equal to the rank of its augmented matrix, here $[\boldsymbol{X}|\boldsymbol{y}]$. If any set of solutions exist, their projection is a subspace of $\mathbb{R}^J$, of dimension $J - \text{Rank}(\boldsymbol{X})$ (actually, the unit simplex constraint, see Equation (2), decreases by one dimension the projected subspace of solutions). This set of solutions is unique, provided $J = \text{Rank}(\boldsymbol{X})$, otherwise an infinite number of solutions hold. An example of all three possible scenarios with an overdetermined system is illustrated in Figure 10, in dimension 2.*

The general principle underlying lls regression and the hyperplane obtained from maximising the parameters is illustrated in Section 3.3.

Theorem .2: Normal equations

The ***Normal equations****, regardless of the distribution of the error term, provide the following Ordinary Least Squares (OLS) estimate Equation* (11)*:*

$$\hat{p}^{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{11}$$

whose existence implies that the design matrix $\boldsymbol{X}$ is invertible, alternatively that its rank is equal to the number of its columns J for an overdetermined system.
In a cellular deconvolution context, this statement particularly enforces that a cell transcriptomic profile cannot be rewritten as a linear combination of the other expression vectors (no multicollinearity*), alternatively that you should not simultaneously infer the cellular ratios of overlapping cell subsets, notably child cell lines mixed with parent cell lines.*

Theorem .3: Gauss-Markov theorems

The Gauss-Markov assumptions encompass:

1. ***Strong exogeneity**: The cell type-specific expression profiles are not random variables but rather fixed and constant observations, underlying implicitly that cell populations do no interact:* $\forall i \in \widetilde{J}, \forall j \in \widetilde{J}, i \neq j, \quad \mathbb{C}ov\left[\boldsymbol{x}_{.\boldsymbol{i}}, \boldsymbol{x}_{.\boldsymbol{j}}\right] = 0.$

2. ***Gaussian-Markov noise:** This hypothesis postulates that the residual error term is described by a white Gaussian noise process, characterised by null mean and variance that is independent on the gene, thus homoscedastic, which yields, in mathematical terms:*

$$y_g = \sum_{j=1}^{J} x_{gj} p_j \ + \epsilon_g, \ \epsilon_g \sim \mathcal{N}\left(0, \sigma_g^2\right)$$

.

By integrating the exogeneity and homoscedasticity assumptions, it is possible to derive the distribution of each transcript, which reveals Gaussian as articulated in Equation (12)*:*

$$\begin{aligned} y_g|\boldsymbol{x}_{g.} &\sim \mathcal{N}(\sum_{j=1}^{J} x_{gj} p_j, \sigma^2) = \varphi(y_g|\boldsymbol{x}_{g.}, \boldsymbol{p}, \sigma) \quad \textit{(univariate formula)} \\ \boldsymbol{y}_{1:G}|\boldsymbol{X} &\sim \mathcal{N}_G(\boldsymbol{X}\boldsymbol{p}, \sigma^2 \mathbf{I}_G) \quad \textit{2: multivariate formula)} \end{aligned} \tag{12}$$

The second line is the matricial representation of the equation. It highlights that the conditional distribution is identifiable to a spherical multivariate Gaussian distribution, parametrised by a diagonal covariance matrix $\boldsymbol{\Sigma}$ *with only one constant diagonal term.*

3. ***Independence:** From the aforementioned Gaussian-Markov and exogeneity assumptions, we readily deduce that the gene expressions of the bulk measures are independent:* $\forall j \in \widetilde{G}, \forall k \in \widetilde{G}, j \neq k, \quad \mathbb{C}ov\left[y_j, y_k\right] = 0.$

4. ***Completeness:** We assume no additional latent variable, such as a non-observed cell population.*

If they hold, the MLE estimate is then equal to the OLS estimate given by the ***Normal equations*** *(Equation* (11)*). Additionally, the MLE is the unique BLUE (best linear unbiased estimator), i.e. the unbiased estimator with the lowest variance.*

Proof .4: Gauss-Markov proof

Under the Gaussian-Markov assumptions (see theorem .3 and notably Equation (12)) and assumption of independence between samples, then, the global log-likelihood distribution of the response variable $\boldsymbol{y}$ conditioned on $\boldsymbol{X}$ is given by Equation (13):

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MLE}} &= \ell_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{X}) \\ &= \arg\max_{\boldsymbol{\theta}} \left[\sum_{g=1}^{G} \log \left(\mathbb{P}_{\boldsymbol{\theta}}(y_g | \boldsymbol{x}_{g.}) \right] \right. \\ &= \arg\max_{\boldsymbol{\theta}} \left[\sum_{g=1}^{G} \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_g - \sum_{j=1}^{J} x_{gj} p_j\right)^2}{2\sigma^2}} \right) \right] \\ &= K - G\log(\sigma) - \sum_{g=1}^{G} \left(\frac{(y_g - \sum_{j=1}^{J} x_{gj} \times p_j)^2}{2\sigma^2} \right)\end{aligned} \tag{13}$$

with $K = -\frac{G}{2}\log(2\pi)$, the *normalising constant*. Finding the values for which the derivative of the function Equation (13) cancels yield the same estimate returned by the OLS method Equation (11).
The MLE estimate provides additionally an estimate of the standard deviations:

$$\hat{\sigma}^2 = \frac{1}{G} \sum_{g=1}^{G} y_g - \sum_{j=1}^{J} x_{gj} \times \hat{p_j}$$

Ultimately, to prove that the estimate $\boldsymbol{p}$ is indeed the unique global maxima of the log-likelihood function Equation (13), we just have to differentiate the equation once more, and show that the resulting Hessian matrix is indeed *positive definite*.

.1 Robust regression approaches

To evaluate robust least-squares regression methods, two metrics are generally used: the relative *efficiency* of the robust estimate, compared to the OLS estimate when the assumptions of the Gaussian-Markov theorem apply (the OLS estimate is indeed asymptotically efficient estimate, in the sense of attaining the Cramér-Rao bound), and the breakdown point (BP), which is the minimal proportion of outliers in the dataset required so that the estimate does not converge anymore. The OLS estimate has a small BP of $\frac{1}{G}$, implying that only one single unusual observation can contribute to the mean of the estimated ratios [Rou85].

Definition .5: M-estimates Regression

M-estimates, short for "maximum likelihood estimates" design the class of estimators that maximise a likelihood function. In practice, M-estimates *replaces the equally weighted observations from lls regression with an adaptive function of the residuals:*

$$\hat{\boldsymbol{p}}_M = \arg\min_{\boldsymbol{p}} \sum_{g=1}^{G} \rho\left(y_g | \boldsymbol{x}_{g.}; \boldsymbol{p}\right) \tag{14}$$

where ρ is the robust loss function and $\psi = \rho'$ is its derivative called the influence function. *Setting $\rho(x) = \frac{1}{2}t^2$, we return on the original OLS problem. Different loss functions lead to different properties of M-estimators, and the choice of the loss function depends on the distribution of the dataset and the desired properties of the estimator:*

- Huber's loss *was the first used, in 1981 Equation* (15)*:*

$$\rho(\boldsymbol{x}; \boldsymbol{p}) = \begin{cases} \frac{1}{2}(\boldsymbol{x} - \boldsymbol{p})^2, & \text{if } |\boldsymbol{x} - \boldsymbol{p}| \leq c \\ c \cdot |\boldsymbol{x} - \boldsymbol{p}| - \frac{1}{2}c^2, & \text{if } |\boldsymbol{x} - \boldsymbol{p}| > c \end{cases} \tag{15}$$

  The Huber loss is a compromise between the squared loss (least squares) and the absolute loss (L1 loss), the latter being less sensitive to outliers.

- Tukey's bisquare function *is a softer smoothing function Equation* (16)*:*

$$\rho(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{\boldsymbol{x}-\boldsymbol{p}}{c}\right)^2\right)^3\right], & \text{if } |\boldsymbol{x} - \boldsymbol{p}| \leq c \\ \frac{c^2}{6}, & \text{if } |\boldsymbol{x} - \boldsymbol{p}| > c \end{cases} \tag{16}$$

  With $c = 4.6885$, its efficiency *is equal to the Huber's estimate (95% of an OLS estimate). Although not implemented independently in any deconvolution paper, the standard rlm (for robust linear modelling) function in the R MASS package, which performs the Tukey's biweight iterative regression, is often used as a gold-standard robust linear regression method in most of the deconvolution benchmark papers ([Stu+19b], [Gau13]).*

- *The bisquare loss is similar to Tukey's loss but has a more compact support. It also downweights outliers with a stronger penalty Equation* (17)*:*

$$\rho(\boldsymbol{x}; \boldsymbol{p}) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{\boldsymbol{x}-\boldsymbol{p}}{c}\right)^2\right)^3\right], & \text{if } |\boldsymbol{x} - \boldsymbol{p}| \leq c \\ 0, & \text{if } |\boldsymbol{x} - \boldsymbol{p}| > c \end{cases} \tag{17}$$

- *The Least Absolute Deviation (LAD) minimises the absolute differences of the residuals (L1 distance) rather than their squared differences (L2 distance):*

$$\hat{\boldsymbol{p}}_{MAE} = \arg\min_{\hat{p}} \left(|\boldsymbol{x}\boldsymbol{p} - \boldsymbol{y}|\right) \tag{18}$$

  where MAE stands for Mean Absolute Deviation. A distribution of these functions is reported in Section 3.3.

  The RCR (Robust Computational Reconstitution) algorithm, presented in [Hof+06], employs LAD regression with additional trimming and the enforcement of non-negativity and sum-to-one constraints.

Definition .6: Least Trimmed Squared Regression

The LTS method was first proposed in [Rou85], with the idea to select the gene subset that exhibit the smallest residuals altogether. Practically, the estimate is given by Equation (19)*:*

$$\hat{\boldsymbol{p}}_{LTS} = \arg\min_{\boldsymbol{p}} \sum_{g=1}^{G} {}^{*}r_g(\boldsymbol{p})^2 \tag{19}$$

with $\left|\widetilde{G^*}\right| = G(1-\alpha)+1$ *with* α *the trimming proportion, and* $r_g(\boldsymbol{p})$ *the residuals ordered by increasing order. Taking* $\alpha = \frac{G}{2}$*, LTS asymptotically displays a strong BP of 0.5, implying it is robust to outliers, but a very low efficiency of 0.08. In addition, LTS is an NP-hard problem [Rou85], as any combination of* $\binom{G}{\left|\widetilde{G^*}\right|}$ *observations should be tested, to find the* $\left|\widetilde{G^*}\right|$ *genes with the minimal residual error. [RV06] hence extends the method in high dimension, or with a large number of observations, by proposing a stochastic and faster version of this algorithm. However, its performance is highly dependent on the initial random* $\left|\widetilde{G^*}\right|$*-subset chosen. Last but not least, the trimming ratio is an additional hyper-parameter that plays a key role on the accuracy of the estimate.*

A comprehensive review of robust linear estimates is supplied in [YYB14], with 10 influence functions benchmarked. It notably demonstrates that MM-estimates and RWLSE estimates have overall the best performance in terms of robustness and asymptotic efficiency.

Definition .7: Support Vector Regression

As in classical linear regression, linear SVR identifies the hyperplane that fits as many data points as possible, but contrary to classical linear regression approaches, only a subset of data points, termed as "support vectors" (SVs) impact the prediction.
Precisely, like most penalised approaches, the optimisation function Equation (20) *reflects the need of finding the sweet spot between minimising the error (herein, the difference between the estimated and observed transcriptomic values) and maintaining a controlled level of complexity to prevent overfitting:*

$$\tau(p,\zeta,\epsilon) = \underbrace{\frac{1}{2}\sum_{j=1}^{J} p_j^2}_{L2\ metric} + \underbrace{C\left(\nu\epsilon + \frac{1}{G}\sum_{g=1}^{G}(\zeta_g + \zeta_g^*)\right)}_{\nu-insensitive\ function} \tag{20}$$

where C is a regularisation parameter controlling the trade-off between complexity and error control, $\boldsymbol{p}$ the estimates, referred to as the weights of the model, and ζ_g and ζ_g^* slack variables *to control the number of points outside the ϵ-tube. The* penalty function of the $L2$-norm *in Equation* (20)*, which is identical to that employed in ridge regression, penalises the model complexity by putting less weight on the estimated ratios of highly correlated cell types [CM04].*
Finally, each pair of observation and covariates, $y_g, \boldsymbol{x}_{g.}$, are subjected to the following constraints Equation (21)*:*

$$\begin{aligned} y_g - \boldsymbol{p}^T\boldsymbol{x}_{g.} - b &\leq \epsilon + \xi_g \\ \boldsymbol{p}^T\boldsymbol{x}_g + b - y_g &\leq \epsilon + \xi_g^* \end{aligned} \tag{21}$$

with b is the bias term *(corresponding to the intercept in linear regression models), ϵ the margin of tolerance, and slack variables ξ_i and ξ_i^* the allowed deviations from the margin (see Section 3.3 for an univariate visualisation of the constraints induced by the SVR dual optimisation problem described in Equation* (21)*). The bias term corresponds to the null intercept in standard linear regression framework, and is usually negative in SVM models ([Yan19]).*
Furthermore, SVR models can further transform the input data using a kernel function*, allowing to find non-linear decision boundaries and intricate relations, even though assumption Equation* (1) *suggests to use the default identity kernel. For a comprehensive tutorial on SVR based methods, we refer the reader to Introduction to SVR modelling).*

.2 Regularised linear approaches

Definition .8: Regularised linear regression

Historically, the Ridge regression *[HK70] employs a L2-penalty Equation* (22)*:*

$$\begin{cases} \hat{\boldsymbol{p}}_{Ridge} = \arg\min_{\boldsymbol{p}} \left[\underbrace{\sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} p_j x_{gj} \right)^2}_{\text{linear regression}} + \underbrace{\lambda \sum_{j=1}^{J} p_j^2}_{\text{penalty function}} \right] \\ \text{subject to } \sum_{j=1}^{J} p_j^2 \leq c \end{cases} \tag{22}$$

where λ is a constant to apply the Lagrange multiplier optimisation theorem [Fue00]. Ridge regression shrinks the coefficients but not necessarily to zero, implying that there is no hard feature selection: a particularly problematic concern in high-dimensional datasets. Otherwise, Ridge is particularly useful when multicollinearity is a concern.
Subsequently, the Lasso regression [Tib96] uses a L1-penalty, which allows a hard variable selection:

$$\begin{cases} \hat{\boldsymbol{p}}_{Lasso} = \arg\min_{\boldsymbol{p}} \left[\sum_{g=1}^{G} \left(y - \sum_{j=1}^{J} p_j x_{gj} \right)^2 + \lambda \sum_{j=1}^{J} |p_j| \right] \\ \text{subjected to } \sum_{j=1}^{J} |p_j| \leq c \end{cases} \tag{23}$$

Efficiency of this optimisation approach relies strongly on the sparsity of the dataset, inducing that most of the coefficients are truly null. The set of coefficients with non-null values is called the true support*, with an increase of the Lagrangian multiplier being associated to a more stringent feature selection.*
However, Lasso regression underperforms and shows inconsistency when estimating closely related cell types with highly correlated transcriptomic profiles, since it tends to arbitrarily choose one out of a group of correlated features. Especially, the true support can not be found when the irrepresentable condition *assumption is violated, namely when the correlation between the explanatory (those belonging to the true support) and non-explanatory variables is smaller that the intra correlation between the variables associated to the true support. Finally, if the number of features J is much larger than the number of observations G, Lasso might overfit the model.*
Elastic net [ZH05] has been developed to keep the middle ground of both worlds Equation (24)*:*

$$\hat{\boldsymbol{p}}_{ElasticNet} = \arg\min_{\boldsymbol{p}} \left[\underbrace{\sum_{g=1}^{G} \left(y_g - \sum_{j=1}^{J} p_j x_{gj} \right)^2}_{\text{regression function}} + \underbrace{\lambda \sum_{j=1}^{J} (1-\alpha) p_j^2 + \alpha |p_j|}_{\text{penalise complexity}} \right] \tag{24}$$

in which α is a trade-off parameter between the L2-penalty ($\alpha = 0$) and the L1-penalty ($\alpha = 1$). This formulation enables continuous shrinkage, including hard feature selection, and can even by deployed with highly correlated cell expression profiles. However, enhanced versatility of the ElasticNet regression comes at the expense of cumbersome hyper-parameter tuning to find the right balance between L1 and L2 penalties, notably, if the L1 penalty is not strong enough, ElasticNet might not perform effective feature selection.

Interesting, [Zho+14] demonstrates that the Elastic net problem is identifiable to a linear SVR under specific reparametrisation, allowing to utilise highly-scalable and parallel SVM solvers.

.3 Probabilistic approaches

Definition .9: Latent Dirichlet Allocation: introduction

LDA, as a generative probabilistic model, has first been used in natural language processing and topic modelling, with the goal of inferring the distribution of topics across documents. Precisely, LDA assumes that documents are mixtures of topics, and topics are mixtures of words. Applied to our cellular deconvolution context, the documents, for which only the respective number of words is available, represent each patient or sample bulk transcriptomic profile and the distribution of words represent read counts. Finally, the latent topics describe cell populations that make up each document.
Formally, let's introduce the following couple of independent random variables (T, Z) in the probabilistic framework, along with $L = \sum_{g=1}^{G} y_g$, the total number of counts in the sample (aka the library depth):

- $Z = Z_{1:L} \in \{1 : J\}^L$*: a discrete latent variable identifying from which reference population each count originates. With that modelling, the cell ratios (document-topic proportions) can be recovered with:*
$$p_j = \frac{\sum_{l=1}^{L} z_l \mathbb{1}_{z_l=j}}{L}$$
, note that this framework naturally enforces the unit-simplex constraint Equation (2).
- $T = T_{1:L} \in \{1 : G\}^L$*: it its the vectorised transcriptomic expression profile $\boldsymbol{y}$. The total expression of a given gene g is retrieved by summing all transcripts from T associated to this gene: $y_g = \sum_{l=1}^{L} t_l \, \mathbb{1}_{t_l=g}$.*
- *Finally, let's introduce the individual purified expression profile for a gene g produced by a given cell population j: $x_{gj} = \sum_{l=1}^{L} t_l \, \mathbb{1}_{(t_l=g) \cap (z_l=j)}$, then the ratio of this specific gene over the total transcriptomic expression for population j (topic-word proportions) is given by: $\beta_{gj} = \frac{x_{gj}}{L_j}$, with L_j the total number of counts in population j and $\boldsymbol{\beta}_j$ its multidimensional generalisation.*

Definition .10: Latent Dirichlet Allocation: estimation

With this modelling approach, the joint distribution of (T, Z) in the LDA model for a given sample is given by Equation (25):

$$\mathbb{P}_{\boldsymbol{\theta}}(Z_{1:L}, T_{1:L}) = \mathbb{P}(\boldsymbol{p}) \times \prod_{l=1}^{L} \sum_{j=1}^{J} \overbrace{\mathbb{P}(Z = j)}^{p_j} \quad \overbrace{\mathbb{P}(T = g|Z = j)}^{\beta_{g,j}} \tag{25}$$

which corresponds to a parametric mixture model of multinomial (the generalisation of binomial distributions, with more than two outputs for each generation) distributions, and $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{\beta})$ the minimal set of parameters to estimate (all other quantities of interest can be deduced from them).
Simultaneously optimising both sets of parameters is analytically intractable. Instead, adopting a process similar to the EM algorithm (see Section `Introduction to Mixture modelling` in the PhD manuscript), the optimisation is computed sequentially until convergence:

1. ***Initialisation:*** *LDA requires an initialisation step, initial parameters $\boldsymbol{\theta}_0 = (\boldsymbol{p}_0, \boldsymbol{\beta}_0)$ being drawn from prior distributions that should generate candidates in the support of the solution space.*

2. ***E-step:*** *At step (q), MAP (maximum a posteriori) is used to assign the production of each gene g for the count $l \in \{1, \dots L\}$ to a given population j Equation (26):*

$$\hat{Z}_l^q,\ l \in \{1, \dots, L\} = \arg\max_{j \in \tilde{J}} \left[\mathbb{P}(Z_l = j|T_l = g)\right] \tag{26}$$

   , using the prior inferred parameters of the mixture of multinomial distributions $\boldsymbol{\theta}^{q-1}$

3. ***M-step:*** *Injecting the latent variables inferred in the previous estimation step, the parametric vector $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{X})$ which maximised the conditional distribution $\mathbb{P}_{\boldsymbol{\theta}}(T_{1:L}|Z_{1:L}) = \prod_{l=1}^{L} \mathbb{P}_{\boldsymbol{\theta}}(T_l|Z_l)$ is returned.*

The main advantage of LDA relies on its versatility, since this approach can be applied to various types of data (provided it has been discretised), can be easily interpreted and is close to the biological process. We refer the reader to [Lee+18] and [Xu+23] for a comprehensive report of the main features and limitations provided by this "bag-of-words" approach.
However, determining the number of cell types J can be challenging without proper biological annotation, the method is highly sensitive to preprocessing choices, and struggles with sparse data or short documents (in a biological context, this implies that this method should not be used to characterise rare cell populations, contributing poorly to the final pool of transcripts).

A Enrichment-based methods

Definition A.1: Principles of GSEA

Gene Set Enrichment Analysis (GSEA) is a bioinformatics method used initially to determine whether a predefined set of genes shows significant differences between two biological states. They hence differ from Differential Gene Expression Analysis (DGEA) analyses, since GSEA operates on groups of genes associated with a biological function or process rather than considering independently one gene after the other. In a second time, GSEA assigns a statistical significance score to each gene set which evaluates the null hypothesis of randomly distributed throughout the ranked gene list against the alternative hypothesis of a clustering pattern at the top or bottom of the ranked list.
The enrichment score (ES) for each gene set returned by GSEA analyses, reflecting the degree to which genes are unequally distributed in the ties of the ranked list, is given by the following running sum statistic, assuming beforehand that $\widetilde{G}$ and $\widetilde{G_j}$ are ranked by decreasing order of fold change (or any relevant metric) Equation (27):

$$\begin{aligned} ES\left(\widetilde{G_j} \in \widetilde{G}\right) &= \sup_{g=1}^{|\widetilde{G_j}|} \left| F^*_{g\in\widetilde{G_j}}(g) - F_{g\in\widetilde{G}}(g) \right| \\ \text{with } F^*_{g\in\widetilde{G_j}}(g) = \mathbb{P}^*(\widetilde{G_j} \leq g) &= \frac{R^*(g)}{\left|\widetilde{G_j}\right|}, \quad F_{g\in\widetilde{G}}(g) = \frac{R(g)}{\left|\widetilde{G}\right|} \end{aligned} \tag{27}$$

with $\left|\widetilde{G}\right| = G$ the number of genes (I commonly use the second notation for consistency and conciseness reasons, since there is no real risk of confusion), $|\widetilde{G_j}|$ the module, namely the number of genes composing the gene set associated to cell population j, $F^*_{g\in\widetilde{G_j}}(g) = \frac{\text{index of gene } g\text{, alternatively number of genes higher ranked}}{|\widetilde{G_j}|}$ and $F_{g\in\widetilde{G}}(g)$ are the cumulative distribution functions (CDF) of the gene rankings/positions (ordered by decreasing order of fold change) of gene set G_j ($R^*(g)$ being the index of gene g in gene module $\widetilde{G_j}$), respectively within the module itself and with respect to the total set of genes quantified in the study $\widetilde{G}$ (note the asterisk to set apart both distributions).
Note that this score, without weights, is the standard Kolmogorov-Smirnov running sum statistic, used traditionally to compare empirical distributions and for which the existence for an asymptotic one-sided statistical test of the null hypothesis distribution is known [SL11], and that ES scores can be easily computed in R with the gsva function.

Definition A.2: Main limitations of GSEA

GSEA approaches are yet hindered by the requirement of carefully identifying all genes involved in the biological process of interest, GSEA identifies associations and correlations between gene sets and phenotypes rather than causal mechanistic information and is sensitive to the choice of the metric. In addition, GSEA does not specify the sense of variation induced by the phenotypical condition on the gene set expression, namely whether it is up- or down-regulated. Ultimately, to evaluate the significance of the enrichment score returned, it is required that the size of the gene set is not too large, neither too small. Indeed, chances of considering smaller pathways as significantly enriched is biased upwards by chance, leading to increased chances of returning false positives. In addition, neither asymptotic tests nor bootstrap-computed p-values are tailored to small gene subsets, in the latter case due to the impossibility of computing the required number of permutations to compute the null distribution.
However, these limitations can be partly alleviated by coupling several complementary metrics, for example by combining the index returned with the fold-change ranks of gene expression with weights obtained from the $p-$values retrieved from DGEA. Regarding the lack of information on the sense of variation, the Connectivity Map [Lam07] proposes a direct extension of the ES scores in Equation (27), by computing two distinct metrics, one for the genes down regulated within the pathway, and one for the up regulated. If the sign of both outputs is congruent, [Lam+06] suggests to conclude of the absence of any significant enrichment (no matter the negative or positive feedback role of a given gene within the pathway, its expression is impacted similarly).

Definition A.3: Using Hypergeometric Laws for Gene Pathway Enrichment Analysis:

Hypergeometric distribution is commonly used in gene pathway enrichment analysis, such as in the Gene Ontology (GO) database. The main purpose is to assess whether a particular set of genes, often the set of differentially expressed genes, is statistically over-represented in a predefined gene pathway compared to what would be expected by chance. If the observed overlap is larger than expected, it suggests that the pathway is enriched.
Mathematically, the hypergeometric distribution returns the probability of observing $X = k \equiv \left|\widetilde{G_j}^{\text{diff}}\right|$ genes (likely the subset of genes differentially expressed) from the set of interest in a pathway of size $\left|\widetilde{G_j}\right|$, drawn randomly without replacement from the total set of genes marked as differentially expressed in Differential Gene Expression Analysis (DGEA) $\widetilde{G}^{\text{diff}} \in \widetilde{G}$, and is computed by the following probability mass function Equation (28):

$$\mathbb{P}(X = k) = \frac{\binom{|\widetilde{G_j}|}{k} \cdot \binom{|\widetilde{G}|-|\widetilde{G}_j|}{|\widetilde{G}^{\text{diff}}|-k}}{\binom{|\widetilde{G}|}{|\widetilde{G}^{\text{diff}}|}} \tag{28}$$

Before the advent of efficient computational tools, it was common to approximate the hypergeometric distribution with a standard binomial distribution, $X \sim \text{Binom}(p, n)$, parametrised by $p = \frac{|\widetilde{G}_j|}{|\widetilde{G}|}$ the probability of a success, defined here as drawing by chance a gene associated to pathway $\widetilde{G_j}$ from the universe of genes $\widetilde{G}$, and $n = \left|\widetilde{G}^{\text{diff}}\right|$ the number of trials (this strategy is comparable to perform an experience with replacements).
*Ultimately, any test used to compare equality of proportions, here the number of differentially expressed genes respectively within the pathway and in the whole gene population, can be used, ranging from contingency tables evaluated by asymptotic χ^2 statistical test to **Fisher's exact test**. With quantitative gene expression available, any test comparing two continuous statistical distributions including the Pearson correlation score, could alternatively be employed as well.*

Definition A.4: Main limitations of hyper-geometric approaches

The hypergeometric test is roughly subjected to the same limitations as GSEA analyses, making similar assumptions. Namely, application of the method assumes that genes are selected independently for inclusion in the pathway (while in the real world, genes cooperate each other to perform intricate biological processes), it is sensitive to the gene size [Aba+09] and to the Background Set, namely the choice of the gene universe $\widetilde{G}$ sequenced, and like any type of downstream analyses, multiple testing correction is necessary to control the family-wise error rate when performing statistical evaluation independently for multiple items simultaneously.
In addition, the method is even less informative than GSEA, since the hypergeometric test treats all genes in the pathway equally, regardless of their relevance or the magnitude of gene expression changes (according to our latest state-of-the-art review, there is no implementation of a weighted hypergeometric test). However, this set of limitations, likely to decrease the statistical power of the tool, is counterbalanced by its heightened versatility, as it remains agnostic to the inherent characteristics of the evaluated datasets. A brief overview of methods aiming at quantifying the level of activation of gene pathways, in a given cell population or phenotype condition, is detailed in Section 3.3.

B Unsupervised and reference-free approaches appendix

Definition B.1: Principles of LS-NMF

Least-Square Non-Negative Matrix Factorization (LS-NMF) is originally a dimensionality reduction technique, based on factorising a given non-negative data matrix into a product of two non-negative matrices. LS-NMF enables to reduce the dimensionality of the original data in a meaningful manner, by representing the data as a product of two lower-dimensional matrices while keeping the fundamental linear assumption of linearity in cell deconvolution methods (see Equation (1)) and enforcing non-negativity in both the factor matrices $\boldsymbol{P}$ and $\boldsymbol{X}$ (indeed, in both cases, negative values can not be interpreted).
More generally, it is a powerful method to extract relevant features or components of the data (here we assume that the subdimensional features match the individual cellular profiles, $\boldsymbol{X}$), while the coefficients in $\boldsymbol{P}$ represent the weights of these features for each data point (in a deconvolution framework, they are assimilated to cellular ratios). The number of hidden components/spanning dimensions J, which is also the rank of $\boldsymbol{X}$ are interpreted as the number of cell populations at the same lineage level.
Given a non-negative data matrix $\boldsymbol{Y} \in \mathbb{R}_+^{G \times N}$, LS-NMF seeks to determine the best two-terms matrix factorisation that approximate $\boldsymbol{Y} \sim \boldsymbol{X}\boldsymbol{P}$, both non-negative matrices, by minimising the Frobenius norm of the difference [a] Equation (29):

$$\begin{aligned} &\min_{\boldsymbol{P},\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{X}\|_F^2 \\ &\textit{subject to the non-negativity constraints:} \\ &\boldsymbol{P} \geq 0,\ \boldsymbol{X} \geq 0 \end{aligned} \tag{29}$$

This optimisation problem is often intractable, and thus typically solved iteratively using algorithms like multiplicative updates or gradient descent.
However, LS-NMF suffers from two main limitations: it is highly sensitive to the initial set of values provided for $\boldsymbol{P}$ and $\boldsymbol{X}$, and different initialisation can lead to different factorisation and convergence to local optima. The choice of the number of components, which can be interpreted as the rank of $\boldsymbol{X}$ is often arbitrary and critical.

[a] Instead of the Frobenius norm, it is also possible to employ the Kullback-Leibler divergence, as in [Don+20]

Acronyms

DGEA Differential Gene Expression Analysis 43, 44

LLS Linear Least Squares 48

LS-NMF Least-Square Non-Negative Matrix Factorization 46

MLE Maximum Likelihood Estimator 49

scRNA-Seq single cell RNA-Sequencing 19, 21, 22

ST Spatial transcriptomics 51

TPM Transcripts Per Kilobase Million 18

Acronyms

DGEA Differential Gene Expression Analysis 43, 44

LLS Linear Least Squares 48

LS-NMF Least-Square Non-Negative Matrix Factorization 46

MLE Maximum Likelihood Estimator 49

scRNA-Seq single cell RNA-Sequencing 19, 21, 22

ST Spatial transcriptomics 51

TPM Transcripts Per Kilobase Million 18

C Feature-engineering and condition numbers

Definition C.1: Condition number: general definition

The condition number is especially employed in the field of linear regression to quantify the sensitivity of the output to perturbations of the input, which could interpret as the expected error made on their measure.
The condition number is defined more precisely to be the maximum ratio of the relative error in the measured value to the relative error made on the input. Consider for an explicit mathematical formula the following variables: $\boldsymbol{p}$ *is the input of our problem,* $\boldsymbol{y}$ *(alternatively* $f(\boldsymbol{p})$ *the measured value, and* $\tilde{f}(\boldsymbol{p})$*) (alternatively* $\hat{\boldsymbol{y}}$*) the predicted value by any algorithm or predictive function. Then, the relative condition number is formally defined by Equation* (30)*:*

$$\kappa(f, \boldsymbol{p}) = \lim_{\epsilon \to 0^+} \sup_{\|\delta \boldsymbol{p} \leq \epsilon\|} \frac{\|\delta f(\boldsymbol{p})\| \,/\, \|f(\boldsymbol{p})\|}{\|\delta \boldsymbol{p}\| \,/\, \|\boldsymbol{p}\|} \tag{30}$$

with $||\quad||$*, namely the double vertical bars, the usual typology used to mark any matrix norm* [a] *and* $\|\delta f(\boldsymbol{p})\| = \left\| f(\boldsymbol{p}) - \tilde{f}(\boldsymbol{p}) \right\|$ *the relative error.*

[a] See definitions, properties and popular matrix norm definitions on this Wikipedia page: Matrix Norm.

Theorem C.2: Application of the Condition Number as a predictive quality metric for linear-based regression problems

Now, let's focus on an overdetermined linear regression problem, as defined in Equation (3) *and whose Linear Least Squares (LLS) solution,* $\boldsymbol{p}_{\hat{OLS}}$*, is given by Equation* (11)*. Then, we can show that the condition number associated to this is given by Equation* (31) [a] *:*

$$\kappa(\boldsymbol{X}) = ||\boldsymbol{X}|| \times ||\boldsymbol{X}^\top|| \tag{31}$$

It is then possible to show the following inequality, derived directly from the definition of a matrix norm, holds Equation (32)*:*

$$||\boldsymbol{X}|| \times ||\boldsymbol{X}^\top|| \geq ||\boldsymbol{X}\boldsymbol{X}^\top|| \geq \left\|\boldsymbol{X}\boldsymbol{X}^{-1}\right\| = 1 \tag{32}$$

, which provides an upper bound on the precision we can achieve with linear regression in the best case scenario. This bound is only reached if, and only if, the condition number of $\boldsymbol{X}$ *is equal to 1.*
Defining $\|\|$ *as the L2 or Euclidean norm, and building the design matrix such that it is normal yields an explicit general formula relating the condition number of the matrix to its eigen values Equation* (33)*:*

$$\kappa(\boldsymbol{X}) \equiv cond(\boldsymbol{X}) = \frac{\lambda_{max}}{\lambda_{min}} \tag{33}$$

with λ_{max} *and* λ_{min} *respectively the largest and smallest eigenvalues resulting from the singular value decomposition of* $\boldsymbol{X}$*. In R, this condition number can be easily computed with the kappa function.*
As such, this metric assesses how small perturbations in the input data can affect the stability and robustness of the regression model, successfully identifying ill-posed or multicollinear regression problems. Indeed, a matrix associated with a high condition number indicates that matrix $\boldsymbol{X}^\top\boldsymbol{X}$ *is close to being singular and often exhibits strong Multicollinearity, rendering the task of correlating the variations of the response variable with the dependent challenging.*
When coping with high condition number matrix, it is thus often recommended to capitalise on Regularised linear methods, such as the Lasso or Elastic regression methods described in Section 2.1.2, to reduce the multicollinearity by removing irrelevant features, with the purpose of stabilising the model and prevent overfitting.

[a] Note that this definition of the condition number recovers the usual formula of the condition number, defined with respect to matrix inversion of a linear system Equation (1): $||\boldsymbol{X}|| \times ||\boldsymbol{X}^{-1}||$, provided the design matrix $\boldsymbol{X}$ is orthogonal (in that case, the solution returned by the Normal equations simplifies to $\boldsymbol{X}^{-1}\boldsymbol{y}$, since $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is then equal to the identity matrix), implying that all covariates are independent to each other.

Theorem C.3: Correlating Condition Number with a MLE approach

A probabilistic approach provides meaningful insights to reconsider the LLS regression problem. Remember that we model the error by explicitly adding an error term following a null-centred and homoscedastic Gaussian distribution: $\epsilon \sim \mathcal{N}(0, \sigma^2)$ [a]*. Considering that the Maximum Likelihood Estimator (MLE) estimate* $\hat{\boldsymbol{p}}_{mle}$ *is unbiased, we can show that the associated variability is given by Equation* (34)*:*

$$\mathbb{V}ar\left[\hat{\boldsymbol{p}}\right] = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \sigma^2 \tag{34}$$

, then, we have the following equality Equation (35)*:*

$$\begin{aligned} \mathbb{V}ar\left[\hat{\boldsymbol{p}}\right] = \sigma^2 \\ \Leftrightarrow \\ \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = 1 \\ \Leftrightarrow \\ \left\|\boldsymbol{X}^\top \boldsymbol{X}\right\|^{-1} = \left\|\boldsymbol{X}^\top \boldsymbol{X}\right\| = 1 \\ \Leftrightarrow \\ \kappa(\boldsymbol{X}) = 1 \end{aligned} \tag{35}$$

In other words, the variability $\mathbb{V}\text{ar}\left[\hat{\boldsymbol{p}}\right]$*, which we can interpret here as the error made by the algorithm oracle, of the estimated ratios, is equal to the measure error made on the response variable* σ^2*, if, and only if, the condition number of the design matrix is equal to 1. In addition, the precision we can achieve on the estimates is bounded by the precision on the response variable. However, the condition number as a predictive metric for the robustness of a model suffers from specific limitations. First, hampered by its global encompassing approach of a problem, it can not be used to determine which variables are most influential. In addition, it is often diverted from its original purpose and misused to quantify and predict the impact of numerical stability. However, it should not be used to take into account round-off numerical errors nor floating-point accuracy of the computer since the condition number is a property of the matrix that strictly depends on the error performed on the measure of the response variable.*
Its value is highly dependant on the nature and the scale of the datasets analysed, making comparisons across different conditions or sequencing technologies null and void. Paired with that issue, there is no universal threshold that would discriminate ill-conditioned matrices with a high CN from well-designed experiments with low CN. Researchers often rely on domain knowledge and heuristics to determine whether the condition number is problematic for their specific analysis. Ultimately, the condition number does not return the expected inaccuracy to expect when solving a problem, but rather a maximal upper bound, making this metric rather conservative.

[a] remember from the Gaussian-Markov theorem Proof .4 proves that both approaches are equivalent

D Practical use case: construction of the LM22 signature

Let's take a practical example by deriving the process used to generate the most popular signature matrix, the LM22 signature which gathers the transcriptomic fingerprints of 22 functionally defined human hematopoietic cell lines profiled for $G = 547$ genes using Affymetrix HGU133A microarray data, isolated from peripheral blood [New+15]. This signature notably includes seven T cell types, naïve and memory B cells, plasma cells, NK cells, and myeloid subsets.

1. First of all, Robust Multi-array Average (RMA) normalisation was performed to aggregate probe sets information at the gene level, uniformed using the HUGO annotation.

Briefly, RMA consists of three steps: *background correction* aims at correcting for the inherent non-specific binding of probes, *quantile normalisation* ensures the intensity distributions of all samples are comparable (a key feature when conducting comparative analyses across different conditions), relying on the assumption that most genes are not differentially expressed across samples, and finally an optional log2-transformation to stabilise the variance, notably towards highly or poorly expressed transcripts, and render the data approximately bell-shaped and close to a Normal distribution.

2. Batch correction was performed to remove differences related to the source of the transcriptomic GEP (gene expression profiles), since a collection of datasets, all from the public domain, were required to get a comprehensive number of replicates for each cell population.
3. Differentially expressed genes were identified using a two-sided unequal variance t-test (better referenced as the Welch test), all transcripts exhibiting an adjusted q-value inferior to 0.3 being considered as markers of the related leukocyte subset and ordered by decreasing fold change.
4. Subsequently, for each leukocyte subset, the intersection of the top g^* marker genes from each cell subset were combined into a signature matrix, iterating g from 50 to 200 across all subsets.
5. The signature matrix that exhibits the lowest condition number, denoted as the LM22 signature, identifies a total of 22 whole blood cell populations, by a staggering reduced subset of 547 distinct genes.
6. Additional filtering steps, tissue and study-specific, may be carried out to further discard genes displaying non specific expression, or likely to be expressed in non identified/characterised cell populations, such as tumoral cell lines. For instance, [Che+18] suggests adding this two-step filtering protocol, when applying CIBERSORT to deconvolve tumoral samples:
    - Exclude genes presenting a strongly enriched expression in non-hematopoietic cell lines, based on the scores computed by the Gene enrichment profiler [CCI18]. The underlying idea is to exclude genes expressed in non-hematopoietic cells, namely that do no originate from blood cell lines, and thus should not appear as differentially expressed in whole blood samples. A strong score is thus generally the marker of an ectopic transcriptomic expression.
    - Second step is removing all genes consistently expressed in cell lines profiled in the Cancer Cell Line Encyclopaedia (CCLE [12]). The idea underlying the removal of genes significantly enriched in CCLE is to exclude genes that are expressed in both healthy and tumoral cell lines, inn order to prevent potential biases in the estimation of endogenous cell populations. To enhance the performance of CIBERSORT, it's advisable to incorporate purified tumour cell profiles for even stronger filtering of confounding genes.

Interestingly, the SVR approach chosen for CIBERSORT incorporates an additional round of feature selection, controlled by the ν hyper-parameter, within the deconvolution process itself. [Gen+15] demonstrates that this additional feature selection increases the robustness and versatility of the methodology, by discarding genes displaying an ectopic transcriptomic expression in a given phenotype.

E Mapping

Mapping, generally employed for highplex RNA imaging assays, consists first to assign each spatially detected cell to its corresponding (scRNA-seq) profile and secondarily, infer a pattern predicting the location of each scRNA-seq cell based on its transcriptome.

[12] CCLE is a database of 100 human cancer cell lines storing their genetic and molecular fingerprints. Its main purpose is to study antiproliferative activities of various drugs and anticancer agents against these cell lines and identify pathways associated with drug resistance [Sac+20].

Mapping workflow can be subdivided into four main stages, often referred to as the four A's ([Lon+21, Fig. 4]):

- **Adopt** From literature, a subset of the tissues or the populations of interest, with intricate spatial patterns, is selected for further analysis.
- **Assay** Survey the same tissue (to keep the same phenotypical conditions and limit technical variability) with scRNA-sequencing (its higher coverage and unbiased nature makes it a promising candidate for the selection of candidate genes) and spatial barcoding to locate their prevailing location within the tissue. Then, track the spatial and temporal dynamics of this subset of genes with HPRI imaging (recall that this method requires to know in advance the sequence of the genes).
- **Assemble** Using deconvolution and mapping algorithms, generate maps that assigns each coordinate to one cell type. Matching histology images may reveal informative landmarks and help denoising complex areas, such as the tumour leading edge, transition region between cancer and normal tissue.
- **Analyse** The high-dimensionality of ST datasets was use to corroborate ligand-receptor interactions involved in cellular signalling, or to survey evolving dynamics occurring in a disease progressing condition.

Mapping methods, by coupling scRNA-Seq with SRT, reveals intricate intercellular communication networks, by capturing the emphco-localisation of cell subtypes. Co-localisation patterns are further used to evaluate the strength of the interactions between a receptor and its ligand. Receptor-ligand bounds are usually predicted using computational tools, such as co-expression network analysis or molecular screening, and prior information from literature review.

In practice, if two cell subtypes are spatially distant, the likelihood of intercellular communication, and accordingly ligand-receptor binding, is small, even though physically possible. Indeed, [Arm+21] demonstrates that cellular signalling primarily occurs within the proximity of the secreting cell, predominantly at the juxtacrine and paracrine levels.

[Tra+20] benchmarks 14 mapping algorithms, and demonstrates that the three best performing are LIGER [Wel+19], Seurat Integration [Stu+19a] and Harmony [Kor+19]). The principle underlying this co-localisation validation is further described in [Lon+21].

Recent endeavours encompass *SpaOTsc*, by [CN20], and *Monkeybread*, by [Ber+23]. Both methods are user-friendly frameworks for the analysis of the spatial cellular layout and automated inference of co-localisation patterns, and have already been successfully applied to identify the niches making up the TME, providing enhanced insights into the underlying intricate biology of the tumour.

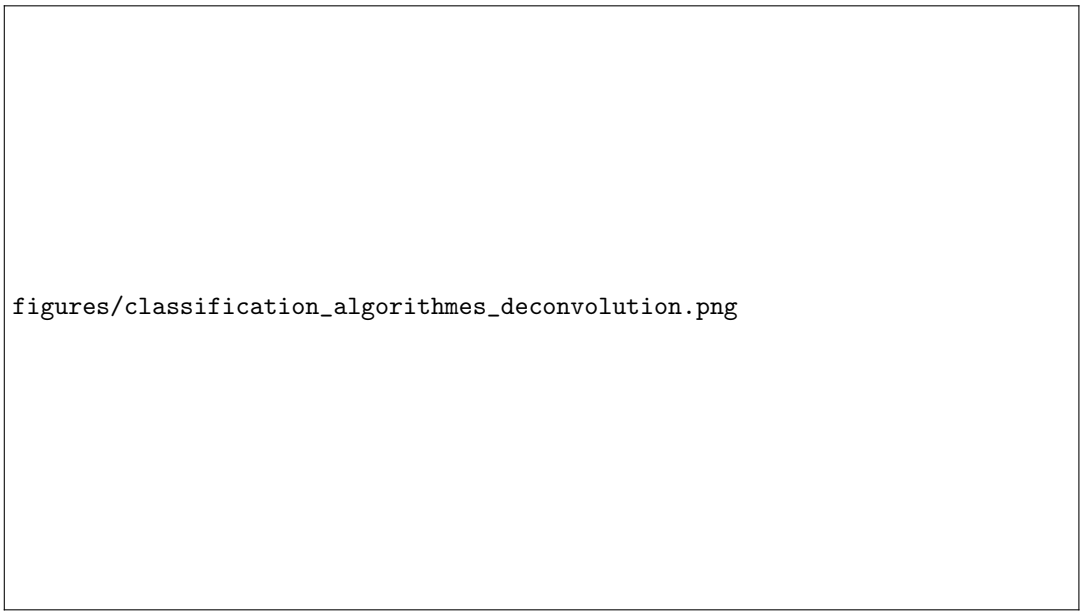
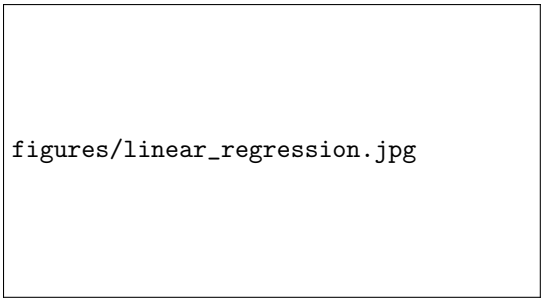
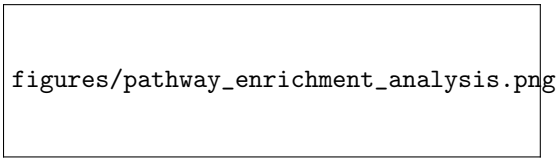


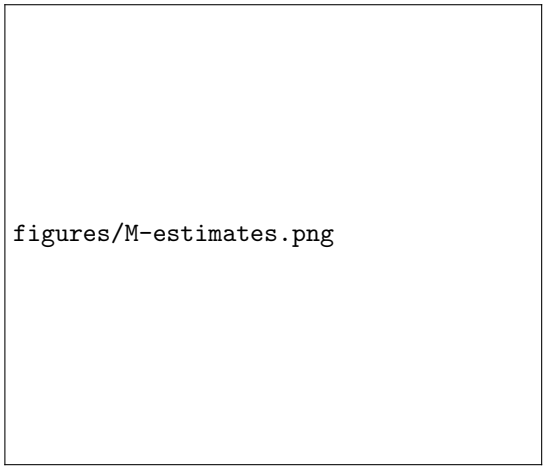
Figure 6. General classification of partial-based deconvolution algorithms.



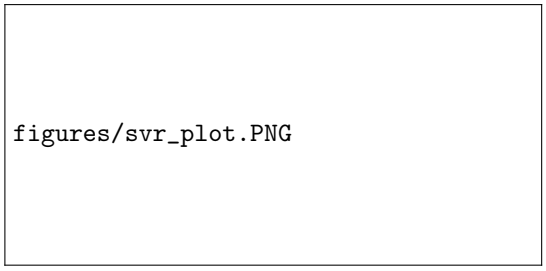
(a) LLS principle. Here, we present briefly the methodology with a simplified univariate regression framework, including an intersection term β_0. Reproduced from [MAL21, Fig. 2].



(c) Overview of three pathway enrichment analysis methods. Over-representation techniques focus on investigating whether a given gene list displays any pathways that are more prevalent than expected by chance when compared to a reference set. (B) In ranking-based methods, the whole gene set is examined to determine whether genes associated with the same pathway exhibit a tendency to cluster at either the top or the bottom of the ordered list of the universe of quantified genes. Such methods return an enrichment score reflecting the amplitude and the sense of the variation induced by the phenotype (C) Topology-based strategies incorporate scores that gauge both gene absolute positions and gene pairwise interactions (up to our knowledge, none of the marker-based methods we reviewed integrate this feature). Reproduced from [ZR23, Fig. 1].

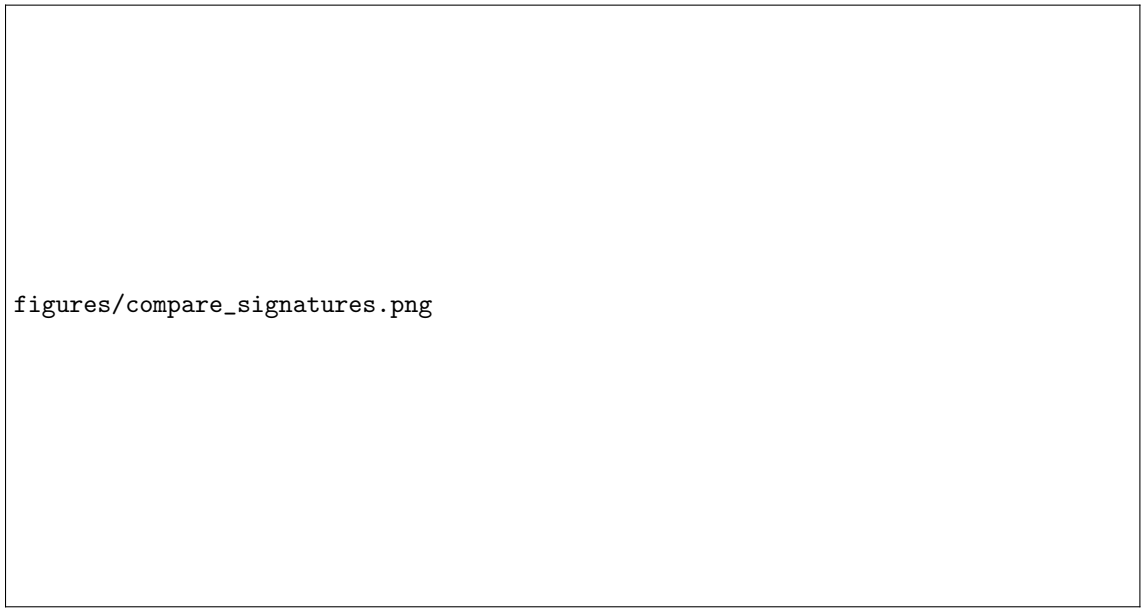


(b) Common influential functions.The weight function distributions for Huber's robust estimator and Tukey's bisquare (or biweight) compared with least squares estimation, in which each observation is assigned the same weight, no matter its contribution to the residuals errors. Reproduced from [Wri09, Fig. 1]



(d) Illustration of support vector regression (SVR).ξ and ξ^* are *slack variables* controlling the upper and lower error margins, respectively. Together, slack variables enable to define boundary decision lines, all points lying outside of the $\epsilon-$ tube making up the set of "support vectors" (red circles). $\nu-$SVR is a recent approach, in which the so-called hyper-parameter controls the amount of SVs (for instance, in the right picture, half of the genes lie beyond the confidence boundaries). Interestingly, only the set of SVs is required to predict cellular ratios, avoiding as such overfitting. Reproduced from [New+15, Supplementary Fig. 1].

(a) Workflow for bulk deconvolution methods. Reproduced from [Avi+18, Fig. 1].



(b) Guidelines for the selection of a deconvolution algorithm. The *overall performance* metric quantifies the correlation between the numerically inferred fractions with the initial parameters used within the generative model of the benchmark. The *background prediction* is a proxy of of the inclined of a deconvolution method to forecast the presence of a cellular classification, even when absent in the mixture. Reproduced from [Stu+19b, Table. 2].

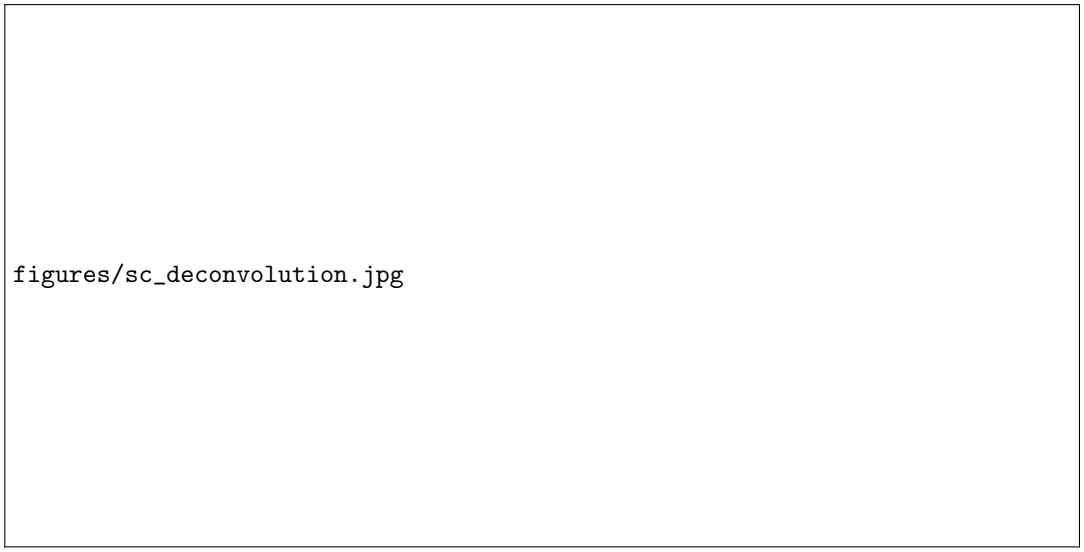
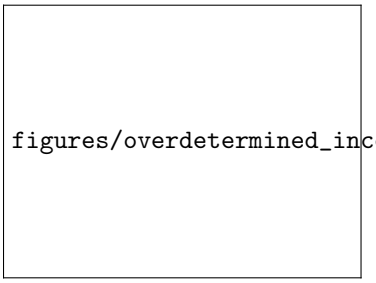
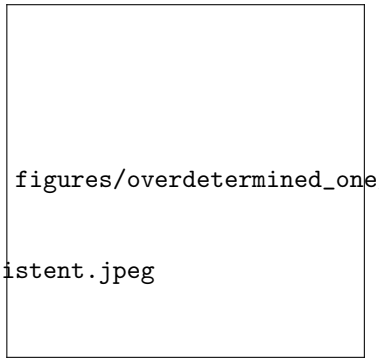


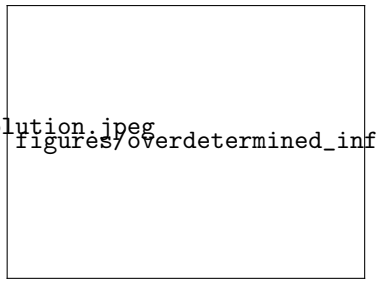
Figure 9. Illustration of a spatial deconvolution algorithm principle, with `stereoscope`.A deconvolution algorithm is used to model and infer the mixture composition of cell populations at a specific capture site using signatures derived from single-cell datasets. `stereoscope` precisely employs a convolution of Negative Binomials to model the mixture of cell types within a captured side. Reproduced from [Kho+21, Fig. 1].



(a) An inconsistent system, with three non concurrent lines: you can only find a set of solutions that verifies simultaneously two equations over three.



(b) A consistent system, with three converging lines intersecting each other at an unique point. The system has only one solution, but exhibits redundancy since you can rewrite one equation as a linear combination of the others.



(c) A trivially consistent system, displaying an infinity of solutions, all matching the identity function. Note interestingly that adding the unit-simplex constraint guarantees however an unique solution, namely equi-balanced cellular ratios: $p_1 = p_2 = \frac{1}{2}$.

Figure 10. Over-determined (three equations against only two unknowns) linear system, adapted from System of equations.