DeCovarT: Robust deconvolution of cell mixture in transcriptomic samples by leveraging cross-talk between genes

Bastien CHASSAGNOL[1,2], Yufei LUO[1], Antoine BICHAT[1], Gregory NUEL[2] and Etienne BECHT[1]
[1] Les Laboratoires Servier, 50 Rue Carnot, 92150, Suresnes, France
[2] LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), 4 place Jussieu, 75005, Paris, France

Corresponding author: bastien.chassagnol@upmc.fr

Abstract ***Motivation:*** *Transcriptomic analyses have increasingly contributed to our understanding of the biological processes involved in the pathophysiology of complex diseases. However, conventional bulk analyses ignore the intrinsic complexity of biological samples, by averaging measurements across multiple distinct cell populations. Furthermore, single-cell sequencing is a destructive and burdensome process, requiring in particular the dissociation of tissues into individual cells. To leverage historical RNAseq analysis or to study fibrous tissues, computational deconvolution methods are still required to unravel the heterogeneous composition of biological samples. However, the performance of these algorithms is hampered by their strong assumption that gene expression is independent of each other, which prevents them from disentangling closely related cell types.*

Results: *We developed a new deconvolution algorithm, DeCovarT, that accounts for the network structure of each purified transcriptomic cell profile. Briefly, we hypothesise that transcriptomic interactions could be modelled by a multivariate Gaussian distribution, parametrised by a sparse precision matrix whose non null inputs represent direct connections between the genes. Finally, we assume that each mixture profile could be reconstructed by a linear combination of each purified transcriptomic cell profile, characterised by their plugged-in covariance estimated beforehand by the gLasso algorithm. The cell ratios are then simply the weighted parameters, inferred in our paper by a constrained and reparametrised version of the Levenberg–Marquardt algorithm, which globally optimises the resulting convolution of Gaussian distributions.*

We highlight that not only the distance between centroids (namely the mean differences between gene expression), but also the structure of the transcriptomic crosstalk, were relevant for selecting the minimal subset of genes to unambiguously characterise each cell population. Using this integrated approach, we obtained a mean correlation of 0.998 (95 % CI 0.995–0.999) from the RNA sequencing data of 35 whole blood samples over ... cell populations, comparable to other standard deconvolution methods, such as deconRNAseq, CIBERSORT or MuSiC, while taking advantage of additional gene sets that were overlooked by other methods. We were also able to distinguish closely related cell populations with similar mean expression profiles but divergent transcriptomic structure.

Keywords cellular deconvolution, gLasso, generative model, bulk RNA Sequencing, tumor micro environment

1 Introduction

The analysis of the bulk transcriptome using high-throughput sequencing methods provided new insights on the mechanisms involved in the development of diseases. However, such methods obliterate the intrinsic heterogeneous nature of biological samples and thus reveal generally worthless to identify the causal sources of the variability observed between individuals.

Indeed, on par with the technical noise or the phenotypical environment, the cell composition plays a crucial role on the evolution of disease conditions. For instance, the tumoral micro environment encompasses a large variety of cell populations, whose interactions will directly impact the tumoral growth, cancer progression and henceforth the patient outcome (1). Furthermore, the expression

profiles of a given cell population can even significantly differ within the same individual, driven by signalling pathways which induce *cell motility* and *cell differentiation* [1].

Not accounting for changes of the cell composition as one of the biological drivers as a confounding signal in downstream analyses, particularly in differential analysis, is likely to result in a loss of *specificity* (genes wrongly identified as differentially expressed, while they only reflect an increase of the cell population naturally producing them) and *sensibility* (genes expressed by minor cell populations are amenable to be masked by the expression and high-variability of dominant cell populations), which in turn prevents from identifying the true causal drivers of the change of gene dysregulation [6].

We can set apart two groups of technics for studying cell heterogeneity. Physical methods, such as immunohistochemistry and flow cytometry, can only take profit of a small subset of phenotypic markers to disentangle cell populations, making them burdensome, low-throughput and costly3. Single-cell sequencing (scRNASeq) is also a promising avenue, but disassociation of the tissues to isolate single cells prior to sequencing make them inconvenient to analyse deeply intertwined and fibrous tissues. Finally, a bench of computational methods were developed for estimating fractions of cell types in bulk admixtures, but they underperform for discriminating closely related cell types (e.g., naïve vs. memory B cells), due to their strong assumption of no interactions between genes. We thus introduce DeCovarT (Deconvolution using the Transcriptomic Covariance), a computational approach that we claim to provide unbiased and less noisy estimates, including ratios of closely related cell populations.

DeCovarT requires a larger compendium of purified RNASeq datasets to estimate for each purified cell population a vector of the averaged transcriptomic expressions and a precision matrix, assumed to be sparse. With these inputs, we consider a generative model to rebuild the admixture of cell populations, assuming that the variability observed for each gene only stems from the stochastic nature of each cell population and its contribution to the global cell composition. To do so, we model the mixture by a convolution of multivariate Gaussian distributions, relaxing the assumption of Independence between observations (here, the individual expressions of transcripts), but keeping the assumption of Independence between covariates (here, the cell populations themselves), for identifiability and computational issues. Finally, the relative cell ratios were assigned the set of parameters that optimise the log-likelihood of the distribution, namely the MLE (maximum likelihood estimate).

We asserted the performance of our deconvolution algorithm by first characterising the network structure and the mean expression profile of the prevalent immune cell populations in blood vessels. ... genes enable to distinguish ... human hematopoietic cell phenotype, including ...

We then benchmarked DeCovarT on idealised mixtures with well-defined composition, and compared it with six GEP deconvolution methods —linear least squares regression (LLSR)4, quadratic programming (QP)5, PERT6, robust linear regression (RLR), MMAD7 and DSA8 (Supplementary Table 3). We next investigated more specifically the DeCovarT's ability to disentangle highly correlated cell types, for which it was nearly unfeasible to identify any DEG (differentially expressed gene). Eventually, we asserted its performance on real datasets, for which FACS measurements of cell types were paired with bulk RNASeq analyses page 4 and 5 of Cibersort, we observed significant improvement over other expression-based methods. Specially, some cell types, likely owing to multicollinearity, were more proned to "drop out" and system underestimation.

2 Material and Methods

2.1 RNA sequencing datasets and preprocessing

2.2 Optimisation of the signature matrix

2.2.1 Discard background noise

2.2.2 Estimate a sparse transcriptomic network

[6] [2] shows that most of the inter-variability of gene expression within healthy patients was brought by variations of the neutrophils population, the major population of blood samples, making up for up to 70 % of the nuclear-equipped cells

2.2.3 Multi-objective optimisation criterion for gene feature selection

2.3 Overview of DeCovarT

Most of transcriptomic deconvolution models, including ours, assume that the total mRNA extracted can be reconstructed by summing the individual contributions of each cell population present in the sample, weighted by its proportion. Formally, let $\boldsymbol{X} = (x_{gj}) \in \mathbb{R}_+^{G \times J}$ the signature matrix storing the purified transcriptomic profile of any of the J cell populations, supposed known, and $\boldsymbol{p} = (p_{ji}) \in [0,1]^{J \times N}$ the unknown relative proportions of cell populations in the N samples disentangled. The linear assumption can then be represented explicitly by the following formula Eq. 1:

$$\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{p} \tag{1}$$

We also consider that he use of a distribution of null average imply that no additional cell population, acting as a *latent variable*, is present in the bulk mixture.

Transcriptomic expression, either in raw counts or after normalisation, is always positive, while most deconvolution methods are not interested in retrieving the absolute frequencies of cell populations, but rather the relative ratios Eq. 2:

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

Without biological or technical noise, the deconvolution problem Eq. 1 is determined (unique solution) if the number of genes is equal or exceeds the number of cell types and no purified cellular expression profile can be rewritten as a linear combination of the other cell populations [7], as stated by the "Rouché-Capelli" theorem ([3]). However, in real use case, biological and technical confusing factors are likely to break the linear constraints in the system of G equations Eq. 1. When the number of genes exceeds those of populations, the problem is said *overdetermined*, which is generally handled by most deconvolution problems by *regression based methods*. The general idea of regression methods is to minimise the Euclidean distance between observed values, noted $\boldsymbol{y_i}$, and the values, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$, predicted by a set of co-variables Eq. 3 [8]:

$$\hat{p}_i^{\text{OLS}} = \arg\min_{\hat{p}_i} ||\hat{y}_i - y_i||^2 = \arg\min_{\hat{p}_i} ||\boldsymbol{X}\hat{p}_i - y_i||^2 = \sum_{g=1}^{G} \left(y_{gi} - \sum_{j=1}^{J} x_{gj} \hat{p_{ji}} \right) \tag{3}$$

The solution to Eq. 3 is given by:

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top y_i \tag{4}$$

An alternative method consists of deriving a probabilistic, *generative model* that represent the uncertainties on the observations by drawing them from probability density functions. The set of parameters, $\theta = (\boldsymbol{p}, \boldsymbol{\sigma})$, that captures the most variability in the real distribution is termed the *maximum likelihood estimate* (MLE). The MLE, from definition, maximises the likelihood or the log-likelihood of observed data, $\ell(\theta|\mathcal{D}) = \P_\theta(\boldsymbol{Y}|\boldsymbol{X})$, with $\theta = (\boldsymbol{p}, \boldsymbol{\sigma})$ the set of parameters, respectively the proportions and standard deviations, to estimate. Under some classical assumptions, enumerated in Theorem 2.1, it appears that there is a direct link between the estimate returned by linear regression and the one derived from maximum likelihood estimation. However, in next section, we consider a new generative model useful when the hypotheses of independence between the genes and exogeinity do not hold.

THEOREM 2.1 (**GAUSS-MARKOV THEOREM**). *If the following assumptions hold,*

***g exogeneity**: The cell type-specific expression profiles are considered fixed and not drawn from probability distributions. Accordingly, we suppose that the expression of each gene within a cell population is*

[7] Alternatively, non multicollinearity is guaranteed if the reference matrix $\boldsymbol{X}$ is invertible and of full rank J

[8] When the function linking the explanatory variables, here our purified expression profiles, to our outcome, here the bulk expression, is linear, we refer to it as OLS *ordinary least squares* method

constant. Biologically, these hypotheses imply that the cellular expression profiles are uncorrelated: $1 \leq i \leq J, \quad 1 \leq k \leq J, \forall i \neq k, \quad cov(x_{gi}, x_{gk}) = 0.$

kov *hypotheses* ***homoscedasticity****: The variance associated to the measured error is fixed for the set of genes, formally* $var(\epsilon_{gi}) = \sigma_i^2, \quad \forall g \in \widetilde{G}$. ***null expected value of the error term****:* $1 \leq g \leq G, \quad \mathbb{E}(\epsilon_{gi}) = 0$, *implying biologically there's no other generation source for the bulk expression data than the set of cell populations referenced in the signature matrix,* $\widetilde{J}$. *Statistically, this also implies that the variance of the measured gene expression does not depend on the predictors. The two previous assumptions are commonly met by using a white Gaussian distribution to model the residual distribution:* $\epsilon_{gi} \sim \mathcal{N}(0, \sigma_i)$

ndependence*:* $1 \leq i \leq G, \quad 1 \leq j \leq G, \forall i \neq j, \quad \mathbb{C}ov[(]\epsilon_i, \epsilon_j) = 0$, *which implies independence between the bulk measures:* $\mathbb{C}ov[(]y_i, y_j) = 0$ *using the weak exogeneity property.*

then, the corresponding MLE estimate is equal to the OLS estimate, solution of Eq. 3 that can be computed using the ***Normal equations****. Additionally, the MLE is unique (only one global maximum of the log-likelihood function) and BLUE (best linear unbiased estimator), i.e. the unbiased estimator with the lowest variance.*

Additionally, we supposed that the samples were uncorrelated between and within a biological condition, and that the proportions differ from a sample to another. Accordingly, we estimate independently the cellular ratios of an individual i under a given experimental condition an drop, for simplicity of redaction, the corresponding indexes.

2.3.1 Deconvolution model Our DeCovarT algorithm relaxes two assumptions from Theorem 2.1: the premise of *exogeneity*, assigning to each purified cell expression a random probability distribution instead of considering it fixed (DSection [4] and DeMixt [5] already came up with using stochastic distribution of the explanatory variables) and the independence assumption between the observations (instead, we assume that genes within a cell population interact to each other). To do so, we consider that the G-dimensional vector $\boldsymbol{X}_j$ describing the interconnected expression of each cell population j follows a multivariate Gaussian distribution: $\boldsymbol{X}_j \sim \mathcal{N}_G(\mu_j, \boldsymbol{\Sigma}_j)$, explicitly stated in Definition 2.2.

Definition 2.2 (**Multivariate Gaussian distribution**). *The multivariate Gaussian distribution of the multivariate variable* $\boldsymbol{X}$*, of size* G*, is given by:*

$$\mathrm{Det}(2\pi\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)\Sigma^{-1}(\boldsymbol{x} - \mu)^\top\right)$$

in which:

- $\mu = \boldsymbol{X}$ *is the* G*-dimensional mean vector*
- $\boldsymbol{\Sigma}$ *is a* $G \times G$ *positive-definite and full rank matrix, which guarantees that the distribution is identifiable and non-degenerate. We define* $\Theta = \Sigma^{-1}$*, the precision matrix.*

For the derivation of the likelihood of the distribution, we *plug-in* for each purified cell population the parameter $\zeta_j = (\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ that was estimated in Sec. 2.2, assuming that the biological and technical conditions underlying the phenotype of the sample in the purified cell distribution and the bulk mixture are similar. Accordingly, the only unknown parameters are the cellular ratios $\boldsymbol{p}_i$ for each sample i. We additionally consider that the ratios are likely to differ even within the same biological condition or individual and that the sequencing of the samples at the bulk level was performed independently. Accordingly, the estimation of the cellular ratios is similar for any sample and can be performed in parallel, we thus drop the sample index i for the sake of simplicity.

Then, considering the matricial relation Eq. 1 and assuming no additional error term and no correlation between the predictors (namely, no transcriptomic crosstalk between to distinct cell populations), the conditional probability of the bulk admixture, $\boldsymbol{Y_i}|\boldsymbol{X}$, given the individual mean and covariance parameters of each purified cellular expression, is given by a multivariate Gaussian distribution, stated explicitly in Eq. 5:

$$\boldsymbol{Y_i}|\boldsymbol{X} \sim \mathcal{N}_G(\boldsymbol{X}\boldsymbol{p_i}, \Sigma_i) \tag{5}$$

with $\Sigma_i = \sum_{j=1}^{J} p_{ij}^2 \Sigma_j$. Indeed, the *convolution product* of independent multivariate Gaussian variables can be readily computed from the *affine invariant* property of multivariate Gaussian distributions.

We can readily derive from the conditional distribution Eq. 5 its log-likelihood, injecting the previously estimated plug-in estimates of the mean and covariance of each purified cell type:

$$\ell_\theta(\boldsymbol{X}, \zeta) = C + \frac{1}{2}\log\left(\mathrm{Det}(\Theta)\right) - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x_i} - \mu)^\top \Theta(\boldsymbol{x_i} - \mu) \tag{6}$$

with $C = -\frac{G}{2}\log(2\pi)$ a constant.

$$\ell_{\boldsymbol{y}|\boldsymbol{X},\Sigma}(\boldsymbol{p}) = C + \log\left(\mathrm{Det}(\left(\sum_{j=1}^{J} p_j^2 \Sigma_j\right)^{-1}\right) - \frac{1}{2}\underbrace{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{p})^\top \left(\sum_{j=1}^{J} p_j^2 \Sigma_j\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{p})}_{\text{squared Mahalanobis distance}} \tag{7}$$

2.3.2 Package implementation

2.3.3 Other GEP deconvolution methods

3 Results

3.1 Impact of multicollinearity

To highlight the counter-intuitive effect of the multicollinearity between gene observations, we simply design a toy example with two genes and two cell proportions, the simplex constraint amounting to estimate a single free parameter. Interestingly, playing over the correlation between the two genes within a cell population and the distance between the two centroids (here, the two-dimensional mean expression vectors of each cell population), we illustrate with that really simple example that no direct relationship between the pairwise correlation between the transcripts and the overall quality estimation could be established. Rather, the level of overlap between two cell populations distributions was a much better proxy of the quality of the estimation: the less the two cell populations distributions overlie, the better the sensibility of the deconvolution. We also showcase that our method, accounting for the covariance structure, systematically performs better compared to a traditional least squares regression method that assumes independent observations.

3.2 Simulation on synthetic idealised mixtures

3.3 Performance on real biological use case

4 Conclusion

We introduced the first deconvolution algorithm that is based on a generative statistical model loosening the strong independence assumption of gene independence. Biologically, discarding this strong hypothesis makes sense, since sets of genes interplay together to perform intricate biological functions, in structures named pathways. Statistically, benchmarking our results in whole blood samples with known FACS counts against DeconRNASeq [6], CIBERSORT [7], MuSiC [8] and xCell [9] results in systematically better performance over a set of quality metrics. Interestingly, the correlation for monocytes was the least accurate for all methods used, confirm this statement?, cf deconvSeq. We are also able to leverage a new set of genes with close mean expression profiles but distinct co-expression patterns, which were traditionally discarded by other deconvolution methods. We hence believe that the higher flexibility of our deconvolution algorithm will make it relevant to increase the currently poor cell resolution of deconvolution methods, with an extended ability to discriminate highly correlated cell population.

4.1 Limitations

The sparse estimate of the precision matrix returned by the gLasso [10] algorithm is generally shrunk, entailing in practice that the non null partial correlations are generally underestimated. *Parameter shrinkage* is a common and well-documented issue of regularisation methods that penalise the complexity of the model. A way to circumvent this problem is to use the *support* returned by the penalisation method to refine the estimation by a canonical maximum likelihood strategy that would integrate the topological constraints induced by the null inputs of the precision matrix (an item set to zero means no direct edge connecting the two transcripts). Unfortunately, except in really specific configurations in which the graph is *chordal*, there is no single MLE solution that optimally capture the correlation structure of a set of observations.

The INDEED algorithm [11] is tailored to select markers between two biological conditions only and not among several groups. While we worked around the problem by performing an one-vs-all strategy, such an approach is quite controversial, since this heuristic strategy does not account for the specificity of each cell profile, the remaining expression profiles being averaged, (discrimination of minor cell populations with lower depth sequencing are likely to be confused with such strategy), nor enable to identify collectively the minimal subset of genes able to discriminate any cell population. Additionally, the way the Indeed algorithm performs to combine the close transcriptomic neighbourhood and the mean expression within a single metric function is derogatory, since the units are not on the same scale.

Rather, a multi-objective and global optimisation approach seems an excellent alternative to manage the trade-off of neighbourhood and mean gene expression discrimination at the scale of a cell population. Finally, while exploring globally the space of gene subsets is generally infeasible due to the combinatorial explosion of possibilities (2^G with G the number of genes retained after background filtering), a genetic and evolutionary approach, likewise to the AutoGeneS algorithm [12]), in which a population of candidate solutions are randomly modified (each solution is the bite wise or indicator function of a set of genes, a zero input representing a gene discarded) and iteratively optimised to finally return a collection of "Pareto-solutions" that represent on a two-dimensional plot the best trade off between the two metrics compared.

Finally, we could refine our generative model to integrate heterotypic interactions, namely accounting for modifications induced by a change of the environmental medium, like a the release of a signalling molecule or a dysregulation of the metabolic pathways induced by a genetic mutation, or a technical batch. A mixed linear-model could be used to account for known environmental and technical confusing factors. However, if no reference profile is available or the nature of the confusing variable is unknown, a better alternative would be to encompass any latent driver within a Bayesian framework, whose parameters of the prior distributions would be retrieved from the literature or FACS data on similar samples (for cell ratios) or from the plugged-in estimates of purified profiles extracted in similar environmental conditions (for the mean and covariance parameters). A Bayesian approach would also alleviate the difficulty of asserting the statistical relevance of the estimates, by directly returning from the simulation likelihood intervals.

4.2 Perspectives

We assumed a multivariate Gaussian distribution to model our purified transcriptome. However, the original outputs returned by RNASeq approaches are discrete integer counts of transcripts that are better captured with Poisson or negative Binomial approaches. We circumvent this problem by using TPM normalisation, followed by a log2 transformation. However, doing so, we do not directly estimate the cellular RNASeq fractions, since we assume that the bulk mixture is a linear combination of the individual transcriptomic contributions of the raw counts, and not of the log2 normalised. While deconvSeq [13] directly estimates the cell ratios on the original material, by modelling gene expressions by a negative binomial distribution, and Kassandra [14] adds artificial Poisson environmental and technical noise to the inputs of its ensemble LightGbm regression approach, none of them accounts for the interaction between genes, the multivariate factor making the estimation hardly challenging.

While for the sake of simplicity, we consider since the beginning that the outputs returned by the deconvolution algorithms were directly the relative cell ratios, this is generally a misleading shortcut. What we compute indeed is rather the fraction of RNASeq produced by a given cell type, rather than the cell proportion itself [9]. Indeed, the total amount extracted from a given cell is contingent on its size and on technical constraints that play on the efficiency of the RNA extraction. To yield the actual cellular ratio, we need to normalise the inferred cellular RNASeq fractions by dividing them by the expected total amount of RNA transcripts released within the cell population, as described in Eq. 8:

$$\hat{p_j^*} = K\frac{\hat{p_j}}{r_j}, \quad K = \frac{1}{\sum_{j=1}^{k}\frac{\hat{p_j}}{r_j}} \tag{8}$$

with r_j the average number of transcripts extracted per cell type and K the normalisation constant asserting the unit simplex constraint. Previous studies handle this biased effect by either normalising back the inferred RNASeq ratios based on extraction experiences prior to the deconvolution (EPIC algorithm in [15] and QuantiSeq [16]) or as an additional unknown parameter to infer (MMAD algorithm, in [17]).

Finally, in complex tissues, such as the TME, there is always a part of the cell composition and expression that is undescribed, notably tumoral cells that mutated from germinal lines display over-expression or amplification of the expression of mutagen driver genes. Simply adding an intercept noise to account for these uncovered cell populations reveals generally inadequate, since this amounts to consider that the contribution of the unknown cell content is similar for any gene. Some deconvolution algorithms tailored this issue, by hypothesising that the unexplained cell transcriptomic profile is the scalar multiplication of one (ISOLATE algorithm [18]) or an additive linear combination of a subset of the ground cell populations (NNML_{np} algorithm [19]) [10], by adding voluntarily some hyper-expression noise to avoid overfitting (the Kassandra algorithm uses an ensemble model refined on thousands of simulated mixtures, that reproduce high-leveraging confusing factors by adding globally a Poisson technical noise and for an *ad-hoc* fraction of genes an uniform constant factor) or by preventing aberrant genes whose expressions were largely altered by somatic mutations from confounding the deconvolution results (robust linear regression or support vector regression, as in Cibersort algorithm, are two approaches that reveal efficient to discard aberrant gene expressions prior to the estimation stage).

Times New Roman 10-point font for the text (Fig. 1 shows an example).

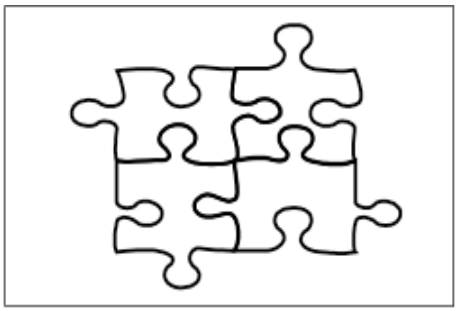

Fig. 1. Old JOBIM puzzle (end of the 20th century).

Acknowledgements

This study makes use of data generated by the Blueprint Consortium, whose proprietary use was granted for exploratory analysis to the company Les Laboratoires Servier. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu.

[9] [15]highlights in supplementary Figure 1, panel A, the expected mRNA amount in common immune cell populations: while lymphocytes and NK cells, all highly associated in terms of cell lineage, display the same mRNA amount, around 0.4pg, the quantity extracted from neutrophils is twice lower and 3 folds higher for monocytes.

[10] Considering that the unknown transcriptomic profile is a modified expression of one or more existing populations avoids that all the unexplained variability is captured by this unknown cell population

We would also like to thank the BBC team within my company who implemented the RNASeq pipeline that was used to preprocess and curate both purified and bulk RNA datasets. The code used for that section can be found on Servier-Github.

References

[1] Jason E. Shoemaker, Tiago JS Lopes, Samik Ghosh, Yukiko Matsuoka, Yoshihiro Kawaoka, and Hiroaki Kitano. CTen: A web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics*, 13(1):460, 2012.

[2] Adeline R. Whitney, Maximilian Diehn, Stephen J. Popper, Ash A. Alizadeh, Jennifer C. Boldrick, David A. Relman, and Patrick O. Brown. Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1896–1901, 2003.

[3] Alexander R. Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS One*, 4(7):e6098, 2009.

[4] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, 2010.

[5] Zeya Wang, Shaolong Cao, Jeffrey S. Morris, Jaeil Ahn, Rongjie Liu, Svitlana Tyekucheva, Fan Gao, Bo Li, Wei Lu, Ximing Tang, Ignacio I. Wistuba, Michaela Bowden, Lorelei Mucci, Massimo Loda, Giovanni Parmigiani, Chris C. Holmes, and Wenyi Wang. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience*, 9:451–460, 2018.

[6] Ting Gong and Joseph D. Szustakowski. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics (Oxford, England)*, 29(8):1083–1085, 2013.

[7] Aaron Newman, Chih Liu, Michael Green, Andrew Gentles, Weiguo Feng, Yue Xu, Chuong Hoang, Maximilian Diehn, and Ash Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12, 2015.

[8] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):380, 2019.

[9] Dvir Aran, Zicheng Hu, and Atul Butte. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, 18:220, 2017.

[10] Rahul Mazumder and Trevor Hastie. The Graphical Lasso: New Insights and Alternatives. *Electronic Journal of Statistics*, 6, 2011.

[11] Yiming Zuo, Yi Cui, Cristina Di Poto, Rency S. Varghese, Guoqiang Yu, Ruijiang Li, and Habtom W. Ressom. INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. *Methods (San Diego, Calif.)*, 111:12–20, 2016.

[12] Hananeh Aliee and Fabian J. Theis. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Systems*, 12(7):706–715.e4, 2021.

[13] Rose Du, Vince Carey, and Scott T. Weiss. deconvSeq: Deconvolution of cell mixture distribution in sequencing data. *Bioinformatics (Oxford, England)*, 35(24):5095–5102, 2019.

[14] Aleksandr Zaitsev, Maksim Chelushkin, Daniiar Dyikanov, Ilya Cheremushkin, Boris Shpak, Krystle Nomie, Vladimir Zyrin, Ekaterina Nuzhdina, Yaroslav Lozinsky, Anastasia Zotova, Sandrine Degryse, Nikita Kotlov, Artur Baisangurov, Vladimir Shatsky, Daria Afenteva, Alexander Kuznetsov, Susan Raju Paul, Diane L. Davies, Patrick M. Reeves, Michael Lanuti, Michael F. Goldberg, Cagdas Tazearslan, Madison Chasse, Iris Wang, Mary Abdou, Sharon M. Aslanian, Samuel Andrewes, James J. Hsieh, Akshaya Ramachandran, Yang Lyu, Ilia Galkin, Viktor Svekolkin, Leandro Cerchietti, Mark C. Poznansky, Ravshan Ataullakhanov, Nathan Fowler, and Alexander Bagaev. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell*, 40(8):879–894.e16, 2022.

[15] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6:e26476, 2017.

[16] Francesca Finotello, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, Zuzana Loncova, Wilfried Posch, Doris Wilflingseder, Sieghart Sopper, Marieke Ijsselsteijn, Thomas P. Brouwer, Douglas Johnson, Yaomin Xu, Yu Wang, Melinda E. Sanders, Monica V. Estrada, Paula Ericsson-Gonzalez, Pornpimol Charoentong, Justin Balko, Noel Filipe da Cunha Carvalho de Miranda, and Zlatko Trajanoski. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Medicine*, 11(1):34, 2019.

[17] 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. *Bioinformatics (Oxford, England)*, 30(5):682–689, 2014.

[18] Gerald Quon and Quaid Morris. ISOLATE: A computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics (Oxford, England)*, 25(21):2882–2889, 2009.

[19] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W. Zandstra. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLOS Computational Biology*, 8(12):e1002838, 2012.

A Model

A.1 The DeCoVart algorithm for deconvolution as constrained optimisation of a multivariate Gaussian convolution

PROPERTY A.1 (MATRIX CALCULUS). *Given two invertible squared matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$*, with respective inverses* $\boldsymbol{A}^{-1}$ *and* $\boldsymbol{B}^{-1}$*,* $A = \boldsymbol{A}(p)$ *and* $B = \boldsymbol{B}(p)$ *being functions of a scalar variable* p*, the following properties hold:*

$$(a)\ \frac{\partial \operatorname{Det}(\boldsymbol{A})}{\partial p} = \operatorname{Det}(\boldsymbol{A}) \operatorname{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p}\right) \quad (b)\ \frac{\partial \boldsymbol{U}\boldsymbol{A}\boldsymbol{V}}{\partial p} = \boldsymbol{U} \frac{\partial \boldsymbol{A}}{\partial p} \boldsymbol{V} \quad (c)\ \frac{\partial \boldsymbol{A}^{-1}}{\partial p} = -\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p} \boldsymbol{A}^{-1}$$

Notably, $\frac{\partial \log(\operatorname{Det}(\boldsymbol{A}))}{\partial p} = \operatorname{Tr}\left(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial p}\right)$ using the chain rule applied to a logarithmic function (alternative notation of the determinant of matrix: $\operatorname{Det}(\boldsymbol{A}) = |\boldsymbol{A}|$?). The function is well-defined, as $\boldsymbol{A}$ is positive-definite, hence its determinant is strictly positive.

A.2 Parameter estimation with simplex constraints and the Levenberg–Marquardt algorithm