L'impact négatif des matrices de référence incomplètes sur la performance de la déconvolution des fréquences cellulaires à partir de l'expression génique

Kalidou BA [1,2], Rodolphe THIÉBAUT [1,2,3], Xavier HINAUT[4,5,6] & Boris HEJBLUM [1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH, U1219, F-33000 Bordeaux, France*
[2] *Vaccine Research Institute, F-94000 Créteil, France*
[3] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*
[4] *INRIA Bordeaux Sud-Ouest, France*
[5] *LaBRI, Bordeaux INP, CNRS, UMR 5800*
[6] *Institut des Maladies Neurodégénératives, Université de Bordeaux, CNRS, UMR 5293*

Résumé. La déconvolution cellulaire désigne l'estimation des fréquences des populations cellulaires à partir des mesures de l'expression des gènes dans un échantillon biologique. Bien que de nombreuses approches supervisées aient été proposées pour résoudre ce problème (telles que `CibersortX` – Newman et al. (2019), ou `scaden` – Menden et al. (2020)), leurs bonnes performances dépendent essentiellement de la matrice des signatures d'expression génique de référence par population. Cette matrice encode les profils d'expression génique de référence des différents types cellulaires, à partir de connaissances préalables et de jeux de données externes. Toutefois, le cas où certaines populations cellulaires présentes dans l'échantillon sont manquantes dans la matrice de référence n'a reçu qu'une attention limitée dans la profusion d'algorithmes de déconvolution proposés, en particulier au vu de sa réalité pratique. Nous quantifions le manque de robustesse des méthodes de déconvolution de l'état de l'art, à la fois dans des simulations numériques et aussi à l'aide de jeux de données réelles. Nos simulations se basent sur une distribution multivariée (soit de Poisson soit Gaussienne) au plus proche de données réelles extraites de la littérature. Nos résultats démontrent que les performances de déconvolution restent relativement inchangées tant que la matrice de référence inclut la grande majorité des populations cellulaires présentes dans l'échantillon, mais qu'à l'inverse les performances de toutes les méthodes de déconvolution se détériorent rapidement à mesure que la matrice de référence devient de plus en plus incomplète. Cependant l'impact des populations cellulaires manquantes dans la matrice de référence dépend de leur fréquence réelle dans l'échantillon.

Mots-clés. Déconvolution cellulaire, RNA-Seq, Matrices de signatures de référence, Types cellulaires, Décomposition matricielle, Optimisation.

Abstract. Cellular deconvolution refers to the estimation of cellular population frequencies from gene expression measurements in a biological sample. While numerous supervised approaches have been proposed (such as `CibersortX` or `scaden`), their good performance critically depends on the reference signature matrix. This matrix encodes the gene expression profiles of the different cell types from external prior knowledge. However, addressing the common scenario of missing cellular populations from the reference matrix has received limited attention compared to the profusion of proposed deconvolution algorithms. We assess the lack of robustness of the state-of-the-art deconvolution methods in both simulations and benchmarking real data. Our simulations designs, based on either a Poisson or a Gaussian multivariate distribution, are validated against real data from the literature. Results from simulations and multiple real datasets, demonstrate that deconvolution performance remains relatively unaffected as long as the reference matrix includes most cellular populations present in the sample. Conversely, performance rapidly deteriorates for all deconvolution methods as the reference matrix becomes increasingly incomplete. Moreover, the impact of missing cell populations in the reference signature matrix depends on their actual frequency in the sample.

Keywords. Cellular deconvolution, RNA-seq, Reference signature matrix, Cell type, Matrix decomposition, Optimisation.

1 Introduction

La déconvolution cellulaire désigne l'estimation des abondances cellulaires à partir de profils d'expression génique spécifiques à chaque type cellulaire grâce à des données de séquençage de l'ARN en masse (*bulk RNA-seq data*) (Avila et al., 2020). L'abondance et la fréquence des populations cellulaires dans un tissu peuvent varier selon différents facteurs, tels que le stade de développement, l'état pathologique ou encore l'influence de l'environnement. La déconvolution des signaux d'expression génique mélangés permet alors de déterminer la contribution spécifique de chaque type de cellule à l'expression globale des gènes. C'est un problème difficile, en raison de la complexité et de l'hétérogénéité des tissus biologiques, de la grande dimension des données d'expression génique, de l'incertitude sur les signatures de référence d'expression génique ainsi que sur les types cellulaires pertinents à inclure, et de la présence de bruit technique dans les mesures RNA-seq (Wang et al., 2019).

Les approches supervisées s'appuient principalement sur les signatures de référence disponibles pour faire correspondre les profils d'expression génique observés selon un modèle de mélange linéaire. Il est impératif de connaître précisément la matrice des signatures d'expression de référence. En effet, celle-ci joue un rôle majeur dans la déconvolution. Cette matrice est essentiellement une représentation des profils d'expression génique pour des types de cellules connus (Shen-Orr and Gaujoux, 2013). La précision, la pertinence, l'adéquation et l'exhaustivité de la matrice de référence par rapport aux données étudiées sont essentielles pour obtenir de bons résultats par ces méthodes de déconvolution. Un problème important et fréquent dans la déconvolution cellulaire est celui des populations manquantes dans la matrice de signature de référence. En effet, en l'absence d'une référence correspondante pour un type de cellule particulier présent dans l'échantillon biologique, le modèle d'estimation se retrouve mal spécifié et les résultats sont alors biaisés (plus ou moins largement). La normalisation induite par les modèles de mélange implique que l'absence de populations cellulaires dans la matrice de référence affecte aussi l'estimation des autres types de cellules connus. Cette limitation rend difficile l'utilisation de ces méthodes de déconvolution dans des contextes où les populations cellulaires présentes sont inconnues (par exemple dans des échantillons sanguins), et empêche l'identification de populations cellulaires rares ou nouvelles avec ces méthodes.

Afin de mieux comprendre les implications d'une matrice de référence de signature incomplète sur la performance des méthodes de déconvolution, nous examinons un ensemble de 16 méthodes extraites d'une liste exhaustive d'une trentaine de méthodes sélectionnées sur la base des éléments suivants : i) la possibilité de personnaliser *a priori* les informations contenues dans la matrice de référence ; ii) l'absence de restrictions sur le nombre de types cellulaires identifiés et la robustesse de la détection des populations de faible abondance ; et iii) l'intégration de la contrainte de non-négativité et la standardisation des abondances estimées. La phase de sélection préliminaire a été orientée par le nombre de citations sur PubMed, en tenant compte des dates de publication. Nous avons veillé à représenter le large éventail de méthodologies proposées pour traiter la déconvolution, y compris des méthodes linéaires, quadratiques, probabilistes et d'apprentissage automatique. Par la suite, nous nous sommes concentrés exclusivement sur les méthodes démontrant les meilleures performances dans l'estimation des populations cellulaires largement étudiées dans la littérature et pertinentes dans l'étude du système immunitaire à partir du sang circulant (telles que les cellules B, les cellules T CD4+, les cellules T CD8+, les neutrophiles et les plasmocytes).

2 Méthodes

2.1 Méthodes de déconvolution

Nous avons évalué plusieurs méthodes de déconvolution cellulaire basées sur des matrices de signature de référence prédéfinies, chacune employant des approches et des techniques spécifiques pour la déconvolution des types de cellules. Dans cet ensemble, `CibersortX` (Steen et al., 2020), `scaden` (Menden et al., 2020) et `Fardeep` (Hao et al., 2019) se distinguent par leur capacité à estimer avec précision les proportions de types cellulaires de manière robuste, grâce à l'utilisation de techniques d'apprentissage automatique. Plus précisément, ils utilisent des techniques d'apprentissage automatique telles que la régression vectorielle de support, les réseaux neuronaux d'apprentissage profond et un algorithme des moindres carrés *trimmés* adaptatifs. En particulier, les méthodes de déconvolution basées sur l'apprentissage profond telles que `scaden` intègrent une phase de simulation de données artificielles pour optimiser le réseau, et `Fardeep` est robuste en présence de bruit en supprimant les valeurs aberrantes avant la déconvolution en utilisant l'idée des moindres carrés trimmés. `AutoGeneS` (Aliee and Theis, 2021) et `DeconRNASeq` (Gong and Szustakowski, 2013) imposent des contraintes de non-négativité et intègrent respectivement les approches des moindres carrés et de la programmation quadratique. `LinDeconSeq` (Li et al., 2020) utilise des méthodes de notation de la spécificité et de linéarité mutuelle pour la sélection des gènes marqueurs, puis applique une régression linéaire robuste pondérée. `EPIC` (Racle and Gfeller, 2020), spécialement conçue pour la déconvolution avec des populations cellulaires inconnues, effectue une régression par les moindres carrés. Les cadres bayésiens ont été adoptés par `BayesPrism` (Chu et al., 2022) et `CDSeq` (Kang et al., 2019), ce dernier mettant en œuvre un modèle bayésien hiérarchique. `DCQ` (Zeev et al., 2014) et `ABIS` (Monaco et al., 2019) utilisent respectivement la régularisation *elastic-net* et le modèle linéaire robuste. QProg (sans contraintes) et QProgwc (Gong et al., 2011) (avec contrainte de non-négativité et de somme à 1) utilisent la régression quantile, prenant en compte l'hétérogénéité spécifique au type de cellule dans les données d'expression génique. `LR` applique la régression linéaire régularisée, tandis que `RLS` (Sharma et al., 2019) se concentre sur la sélection de régions de données d'expression génique qui présentent des modèles similaires entre les échantillons.

2.2 Simulation des données

Nous avons effectué des simulations numériques pour étudier les performances des méthodes présentées ci-dessus. Un grand nombre de populations cellulaires a été considéré dans la matrice de référence W. Pour couvrir un large panel, nous avons simulé nos données selon deux distributions : gaussienne d'une part (dont les paramètres sont obtenus par une approximation gaussienne de la matrice de signature *LM22* proposée par (Chen et al., 2018)) et Poisson d'autre part (qui permet de simuler des données de comptage à l'instar des données RNA-Seq réelles). Afin de générer de la corrélation, inhérente entre les proportions des populations cellulaires (de par leur nature compositionnelle), nous avons défini une matrice de variance-covariance uniforme pour la distribution gaussienne, et adopté la stratégie proposée par (Barbiero and Ferrari, 2015) pour la distribution de Poisson. Les proportions de types de cellules de tous les individus en vrac P sont générées avec une distribution de Dirichlet. Enfin, un bruit indépendant ϵ est ajouté pour générer les données d'expression génique $X(X = WP+\epsilon)$. La Table 1 résume ces choix. La Figure 1A présente un exemple de ces simulations, comparé aux données réelles.

Nombre de gènes	$n = 500$
Nombre de population de cellules	$K = 100$
Number d'échantillon	$N = 50$
Matrice des proportions	$P_{i,*} \sim \mathcal{DP}(K, \alpha_k)$, où $\alpha_k \in [1, ..., K]$, $\sum_{j=1}^{K} P_{i,j} = 1$ et $\forall i \in [1, ..., N]$ $P_{i,j} < 1, \ \forall \ i, j$
Avec une distribution Gaussienne	$W \sim e^{\mathcal{N}(\mu, \Sigma)} + 6$, avec $\mu \sim \mathcal{U}(-1, \ 0.5)$ $\epsilon \sim \mathcal{N}(0, \sigma)$ où $\sigma \sim \mathcal{U}(1, 2)$
Avec une distribution de Poisson	$W \sim \mathcal{P}(\lambda_k, \Sigma)$, avec $\lambda_k \sim \mathcal{U}(10, 25)$ $\epsilon \sim \mathcal{P}(\theta_k)$ où $\theta_k \sim \mathcal{U}(0, 0.3)$

Table 1 : Configuration des simulations numériques

Trois scénarios ont ensuite été définis pour étudier les performances des différentes méthodes de déconvolution en fonction du nombre de populations cellulaires dans la matrice de référence W. Le scénario « *randomly* » consiste à extraire aléatoirement (sans remise) un pourcentage de populations cellulaires définient dans la matrice de référence W. Les scénarios « *lowest* » et « *highest* » correspondent à l'extraction à partir de la matrice de référence les populations de cellules ayant une abondance moyenne soit faible, soit élevée, respectivement.

2.3 Métriques d'évaluation

La précision des différentes méthodes de déconvolution est évaluée à l'aide des mesures de l'erreur quadratique moyenne relative ($RRMSE$) et du coefficient de corrélation intraclasse (ICC) entre les proportions estimées et les véritables proportions.

Le $RRMSE$ fournit une mesure normalisée et interprétable de l'erreur de prédiction, qui facilite des comparaisons équitables entre différents modèles, ensembles de données et échelles, même lorsque l'abondance réelle est faible.

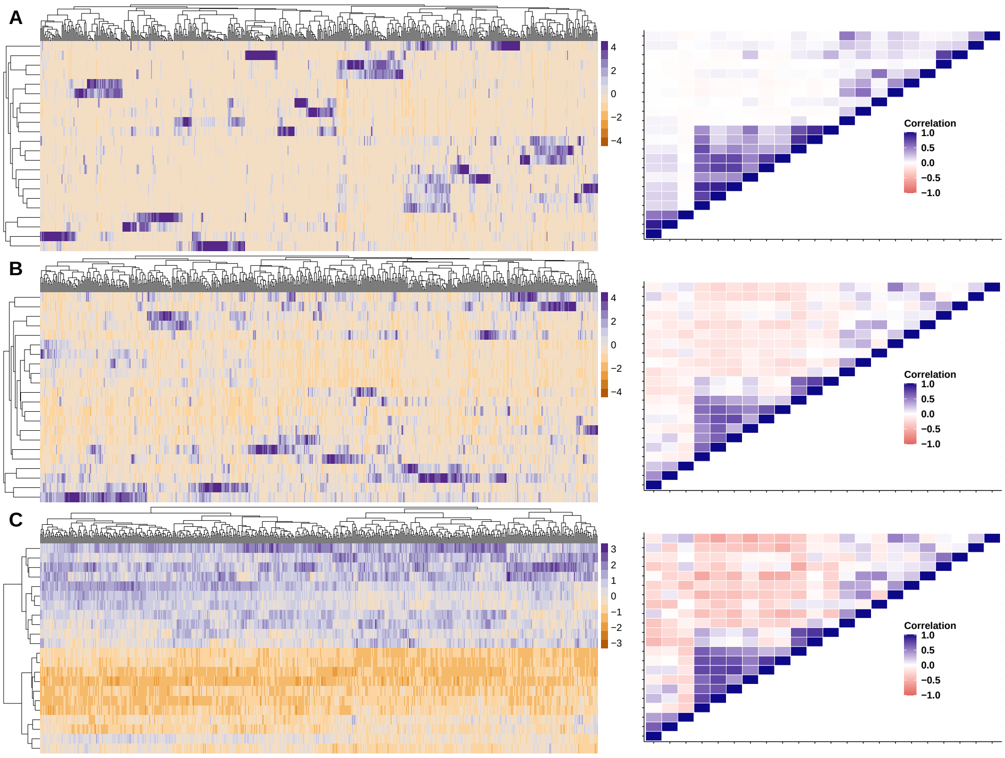

FIGURE 1 : Analyse comparative de matrices de référence simulées issues de distributions gaussiennes **(B)** et de Poisson **(C)**, ainsi que de la matrice de signature de référence *LM22* **(A)**. Les lignes représentent les populations de cellules (K = 22) et les colonnes représentent les gènes (N = 547). À droite, les cartes thermiques illustrent les corrélations par paire entre les populations de cellules, les valeurs R-carré étant utilisées comme mesure de l'association. Les corrélations les plus fortes sont représentées en bleu, tandis que les corrélations les plus faibles sont représentées en rouge.

L'ICC (compris entre 0 et 1) est une alternative au coefficient de corrélation de Pearson (ce dernier étant sensible à la présence de valeurs aberrantes) et mesure la fiabilité et la cohérence des mesures, largement utilisé pour les interprétations cliniques. Plus il est proche de 1, meilleure est la performance. Nous utilisons la définition de l'ICC_3 proposée par (PE and JL, 1979), car nous sommes intéressés par l'évaluation d'un ensemble fixe de treize méthodes de déconvolution. Ces deux indicateurs se calculent ainsi pour une population spécifique :

$$RRMSE = \frac{\sqrt{\frac{\sum_{j=1}^{J}(\hat{f}_j - f_j)^2}{J}}}{\sigma_f} \qquad ICC = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2} \tag{1}$$

où j est l'indice de l'échantillon et J le nombre total d'échantillons biologique dans un jeu de données. $\hat{f}_j$ est la proportion estimée de la population de cellules de l'échantillon j et f_j est la vérité terrain de l'échantillon j. σ_r^2, σ_c^2 et σ_e^2 correspondent respectivement aux variances de l'écart par rapport à la moyenne pour l'échantillon j, au biais de la mesure k et à la composante résiduelle.

3 Résultats

En considérant toutes les populations cellulaires lors de la déconvolution, nous avons observé que 5 des 16 méthodes, `scaden`, `EPIC`, `DCQ`, `BayesPrism`, et `RLS`, avaient des ICC moyen inférieurs à $0,5$; contrairement aux autres méthodes qui fournissaient des estimations bien meilleures. Ces résultats pourraient s'expliquer par le fait que des méthodes telles que `scaden` n'ont pas été optimisées pour traiter plus d'une vingtaine de populations cellulaires.

Indépendamment de la stratégie de simulation adoptée et de la méthode de déconvolution utilisée, plus le nombre de populations cellulaires identifiées dans la matrice de signature est faible, plus les ICC observées entre les véritable valeurs et les prédictions de la déconvolution sont faibles. Cette dégradation des performances était plus importante lorsque les populations cellulaires manquantes étaient très fréquentes dans le mélange, comme le montrent les résultats obtenus avec les 50 populations cellulaires les plus abondantes. La performance moyenne de l'ICC était supérieure à $0,65$ pour le scénario "lowest", alors qu'elle était estimée supérieure à $0,5$ avec les 50 populations cellulaires les moins abondantes dans le scénario "highest". Dans le scénario "lowest", un grand nombre ($> 20\%$) de population doit être considéré comme inconnu pour avoir un impact significatif sur la performance. Ces constatations se reflètent également dans le $RRMSE$ qui augmente à chaque niveau d'incomplétude de la matrice de référence. Ces résultats sont détaillés dans les Figures 2 et 3.

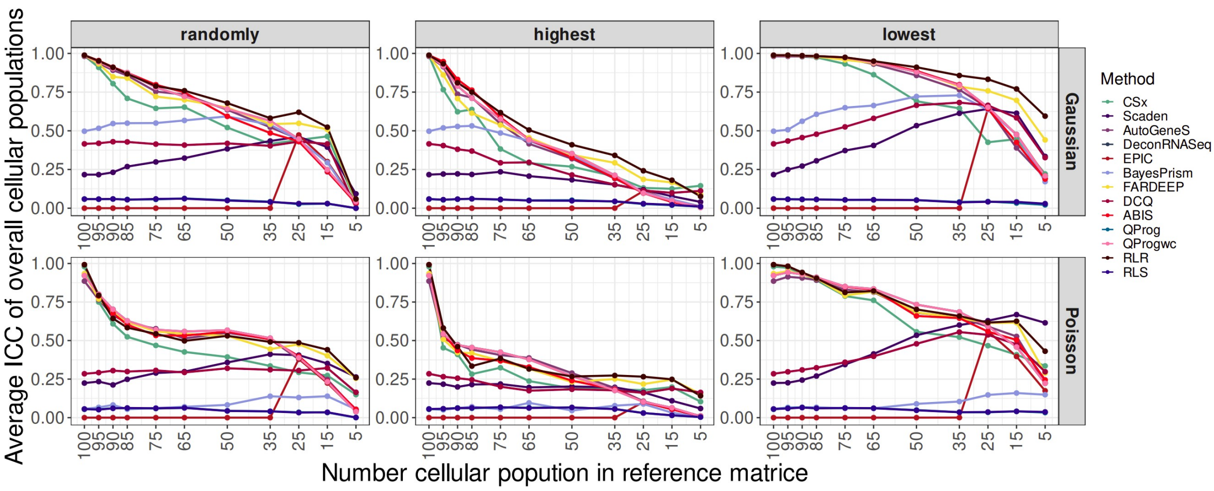

FIGURE 2 : Coefficient de corrélation intraclasse (ICC) moyen des populations cellulaires globales

4 Discussion

Nos résultats de simulation ont permis de mieux comprendre l'impact d'une matrice de référence incomplète. Nous avons constaté que les performances des méthodes de déconvolution se détérioraient à mesure que le nombre de populations cellulaires manquantes

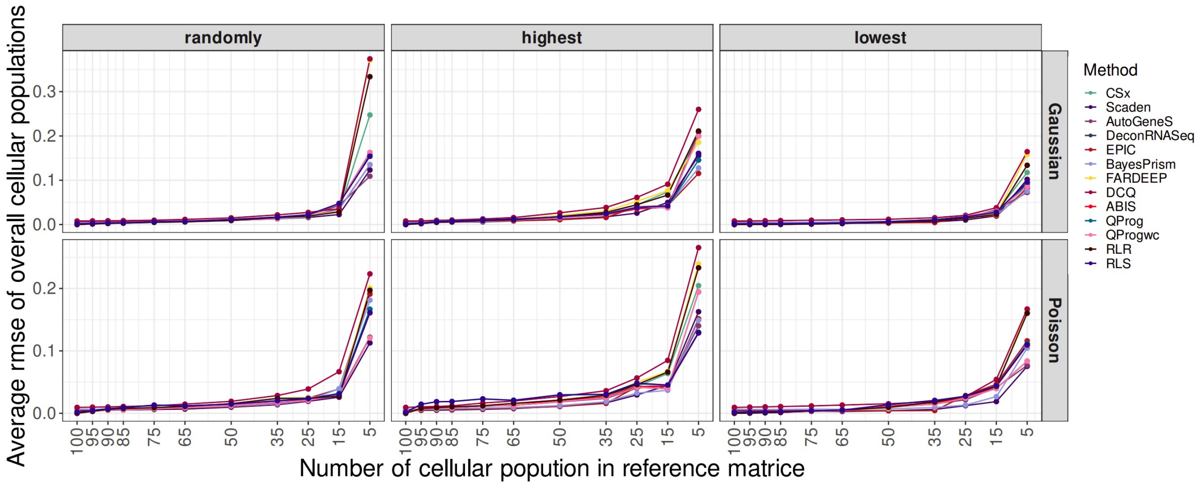

FIGURE 3 : Erreur quadratique moyenne relative (RRMSE) des populations cellulaires globales.

dans la matrice de référence augmentait, d'autant plus que les populations cellulaires manquantes étaient fréquentes dans le mélange. Cela souligne la sensibilité des méthodes de déconvolution à la disponibilité de matrices de signatures de référence complètes.

Il convient de noter que certaines méthodes de déconvolution ont montré des limites lorsqu'elles étaient confrontées à un grand nombre de populations cellulaires, comme nous l'avons observé dans nos simulations. Des méthodes comme `scaden`, `EPIC`, `DCQ`, `BayesPrism` et `RLS` ont eu du mal à fournir des estimations précises lorsque la matrice de référence contenait un grand nombre de populations cellulaires.

En conclusion, notre étude met en évidence la forte dépendance entre la performance des méthodes de déconvolution cellulaire et la qualité et l'exhaustivité des matrices de signature de référence. Les résultats soulignent la nécessité de développer des algorithmes de déconvolution plus robustes, capables de traiter des tissus divers et complexes. Les recherches futures devraient se concentrer sur l'amélioration de l'adaptabilité et de la précision des méthodes de déconvolution, en particulier en présence de matrices de référence incomplètes. En outre, les efforts visant à normaliser les matrices de référence et à établir les meilleures pratiques pour la construction des matrices de référence peuvent améliorer la reproductibilité et la comparabilité des études de déconvolution dans différents contextes biologiques. En résumé, la déconvolution cellulaire est un outil puissant pour disséquer les compositions tissulaires complexes, mais son efficacité dépend de la qualité des données de référence et des capacités de la méthode de déconvolution choisie.

Références

H Aliee and FJ Theis. Autogenes : Automatic gene selection using multi-objective optimization for rna-seq deconvolution. *Cell Systems*, 12(7):706–715, 2021.

CF Avila, J Alquicira-Hernandez, JE Powell, P Mestdagh, and PK De. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11 (1):1–4, 2020.

A Barbiero and PA Ferrari. Simulation of correlated poisson variables. *Applied Stochastic Models in Business and Industry*, 31:669–680, 2015.

B Chen, MS Khodadoust, CL Liu, AM Newman, and AA Alizadeh. Profiling tumor infiltrating immune cells with cibersort. *Methods in Molecular Biology*, 1711:243–259, 2018.

T Chu, Z Wang, D Pe'er, and CG Danko. Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature Cancer*, 3(4):505–517, 2022.

T Gong and JD Szustakowski. Deconrnaseq : a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–5, 2013.

T Gong, N Hartmann, I S Kohane, V Brinkmann, F Staedtler, M Letzkus, S Bongiovanni, and J D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLOS One*, 6(11):e27156, 2011.

Y Hao, M Yan, BR Heath, YL Lei, and Y Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLOS Computational Biology*, 15(5):e1006976, 2019.

Q Kang, K andMeng, I Shats, DM Umbach, M Li, Y Li, X Li, and L Li. Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Computational Biology*, 15(12):e1007510, 2019.

H Li, A Sharma, W Ming, X Sun, and H Liu. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC Genomics*, 21:652, 2020.

K Menden, M Marouf, S Oller, A Dalmia, DS Magruder, K Kloiber, P Heutink, and S Bonn. Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 2020.

G Monaco, B Lee, W Xu, S Mustafah, YY Hwang, C Carré, N Burdin, L Visan, M Ceccarelli, M Poidinger, A Zippelius, J Pedro de Magalhães, and A Larbi. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell Reports*, 26(6):1627–1640, 2019.

AM Newman, CB Steen, CL Liu, AJ Gentles, AA Chaudhuri, F Scherer, MS Khodadoust, MS Esfahani, BA Luca, D Steiner, and M Diehn. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

Shrout PE and Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86:420–8, 1979.

J Racle and D Gfeller. Epic: A tool to estimate the proportions of different cell types from bulk gene expression data. *Methods in molecular biology*, 15:233–248, 2020.

A Sharma, E Merritt, X Hu, A Cruz, C Jiang, H Sarkodie, Z Zhou, J Malhotra, GM Riedlinger, and S De. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. *Cell Reports*, 29(8):2164–2174, 2019.

SS Shen-Orr and R Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 25(5):571–8, 2013.

CB Steen, CL Liu, AA Alizadeh, and AM Newman. Profiling cell type abundance and expression in bulk tissues with cibersortx. *Methods in Molecular Biology*, pages 135–157, 2020.

T Wang, B Li, CE Nelson, and S Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC Bioinformatics*, 20(1):40, 2019.

A Zeev, S Yael, D Eyal, Zohar BI, V Liran, KS Hadas, M Tal, M Ella, M Michal, GV Irit, and A Ido. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology*, 10(2):720, 2014.