D7.1

Application of software enabling computational deconvolution of bulk RNA-sequencing data to immune cell profiles of patient samples

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D7.1 / 1.0
Work package contributing to the deliverable	WP7
Due date	May 2021 – M29
Actual submission date	2nd June, 2021

Responsible organisation	UGent
Editor	Pieter Mestdagh
Dissemination level	PU
Revision	1.0

Abstract	Computational deconvolution of bulk RNA-sequencing data to infer cell type composition of a sample is challenging. Benchmarking of various computational deconvolution tools revealed various data processing parameters that impact deconvolution accuracy and revealed the importance of a complete reference matrix. As a complete reference matrix is often not available, an algorithm was designed that can handle missing cell types. This algorithm can be applied to establish the immune cell repertoire of primary tumor biopsies without prior knowledge of the full spectrum of cell types in the biopsy.
Keywords	Deconvolution, transcriptomics, immune cell

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

UGent

Contributors

Pieter Mestdagh (UGENT)

Francisco Avila Cobos (UGENT)

Disclaimer

Executive Summary

Computational deconvolution of bulk RNA-sequencing data to infer cell type composition of a sample is challenging. First, a benchmarking of various computational deconvolution tools revealed various data processing parameters that impact deconvolution accuracy and revealed the importance of a complete reference matrix. Next, as a complete reference matrix is often not available, a framework that can handle missing cell types has been implemented. It can be applied to establish the immune cell repertoire of primary tumor biopsies without prior knowledge of the full spectrum of cell types in the biopsy.

The procedure starts from the assumption that the reference matrix is not complete (ie not representative of all cell types in the mixture) and, from the available bulk RNA-sequencing data, will search for genes that serve as markers for the missing cell type(s). Having established these genes, their expression levels in the (aggregated) missing cell type(s) are estimated and added to the new reference matrix that can now be applied for computational deconvolution using a suitable method. We have benchmarked the performance of this algorithm using pseudo-bulk mixtures (= with known composition) from single cell RNA-sequencing data. This procedure can be applied to establish the immune cell repertoire from bulk RNA-sequencing data of (pediatric) tumor samples biopsies.

Table of Content

Chapter 1 Benchmarking deconvolution workflows 1
Chapter 2 Identifying markers for missing cell types 5
Chapter 3 Validating and benchmarking the novel workflow 6
Chapter 4 Application of novel deconvolution workflow to pediatric tumor datasets 8
Chapter 5 Summary and Conclusion 10
Chapter 6 List of Abbreviations 11
Chapter 7 Bibliography 12

List of Figures

Figure 1. Schematic representation of the benchmarking study. 2
Figure 2. Combined impact of data normalization and methodology on the deconvolution results. . 3
Figure 3. Effect of cell type removal on the deconvolution results for the PBMCs dataset (100-cell mixtures; linear scale). 4
Figure 4. Investigating those genes with poor fit in a scenario where one cell type (beta) is missing in the reference matrix but actually present in the mixtures. 5
Figure 5. Validation and benchmarking of the novel workflow using pseudo-bulk mixtures from 10x scRNA-seq data of peripheral blood mononuclear cells (PBMCs) where CD56+ natural killer cells were artificially removed from the reference matrix. 6
Figure 6. t-distributed stochastic neighbor embedding (t-SNE) representation of the combination of Dong and Kildisiute datasets. 9

List of Tables

Table 1. Description of the datasets collected for each tumor type. ES = Ewing sarcoma; MB = Medulloblastoma; NB = Neuroblastoma; HB = Hepatoblastoma. 8

Chapter 1 Benchmarking deconvolution workflows

Before establishing a deconvolution workflow (ie. software) to determine the immune cell repertoire of pediatric tumor samples, we first performed a thorough benchmarking study of various algorithms and several data processing parameters. Many computational methods have been developed to infer cell type proportions from bulk transcriptomics data. However, an evaluation of the impact of data transformation, pre-processing, marker selection, cell type composition and choice of methodology on the deconvolution results was lacking in literature. Using five single-cell RNA-sequencing (scRNA-seq) datasets, we generated pseudo-bulk mixtures to evaluate the combined impact of these factors (Figure 1, Avila Cobos et al., Nature Communications, 2020). We found that both bulk deconvolution methodologies and those that use scRNA-seq data as reference perform best when applied to data in linear scale and the choice of normalization has a dramatic impact on some, but not all methods (Figure 2). Overall, methods that use scRNA-seq data have comparable performance to the best performing bulk methods whereas semi-supervised approaches show higher error values.

We then selected nnls and CIBERSORT as representative top-performing bulk deconvolution methods and assessed the impact of removing a specific cell type by comparing the absolute RMSE values between the ideal scenario where the reference matrix contains all the cell types present in the pseudo-bulk mixtures and the RMSE values obtained after removing one cell type at a time from the reference. We then focussed on those cases where the median absolute RMSE values between the results using the complete reference matrix and all other scenarios where a cell type was removed, increased at least 2-fold. In the PBMC dataset, removing CD19+, CD34+, CD14+ or NK cells all had an impact on the computed T-cell proportions (between a three and six-fold increase in the median absolute RMSE values, Figure 3). Similar results were observed in other datasets, and none of the method and normalization combinations was able to provide accurate cell type proportion estimates when the reference was missing a cell type. From these analysis, we concluded that failure to include cell types in the reference that are present in a mixture leads to substantially worse results, and none of the available deconvolution procedures is capable of properly dealing with such a scenario. Nevertheless, we expect such scenarios to occur frequently when dealing with tumor biopsies. In order to establish a complete reference matrix for a given (pediatric) tumor type, single cell RNA sequencing data of representative (bulk) samples of that tumor can provide insights in the cellular composition and guide selection of marker genes to establish a comprehensive reference matrix. However, in case of (extensive) tumor heterogeneity or differential stromal composition, the number of single cell RNA-sequencing datasets required to capture the entire spectrum of cell type compositions for a given tumor type could become (very) large. Therefore, a computational deconvolution algorithm that can cope with missing cell types would be highly beneficial.

For applications aiming to establish the fraction of the most important immune cell subtypes in a tumor biopsy (i.e. immune cell deconvolution), a reference matrix containing immune cell markers only, combined with a deconvolution workflow capable of handling missing cell types (namely the tumor fraction and other component(s)), could potentially be applied to any bulk (pediatric) tumor RNA-sequencing dataset. We therefore decided to develop a dedicated deconvolution workflow that is not affected by missing cell types in the reference.

The software that was developed to benchmark deconvolution methodologies and generate pseudobulk mixtures has been published on Github: github.com/favilaco/deconv_benchmark

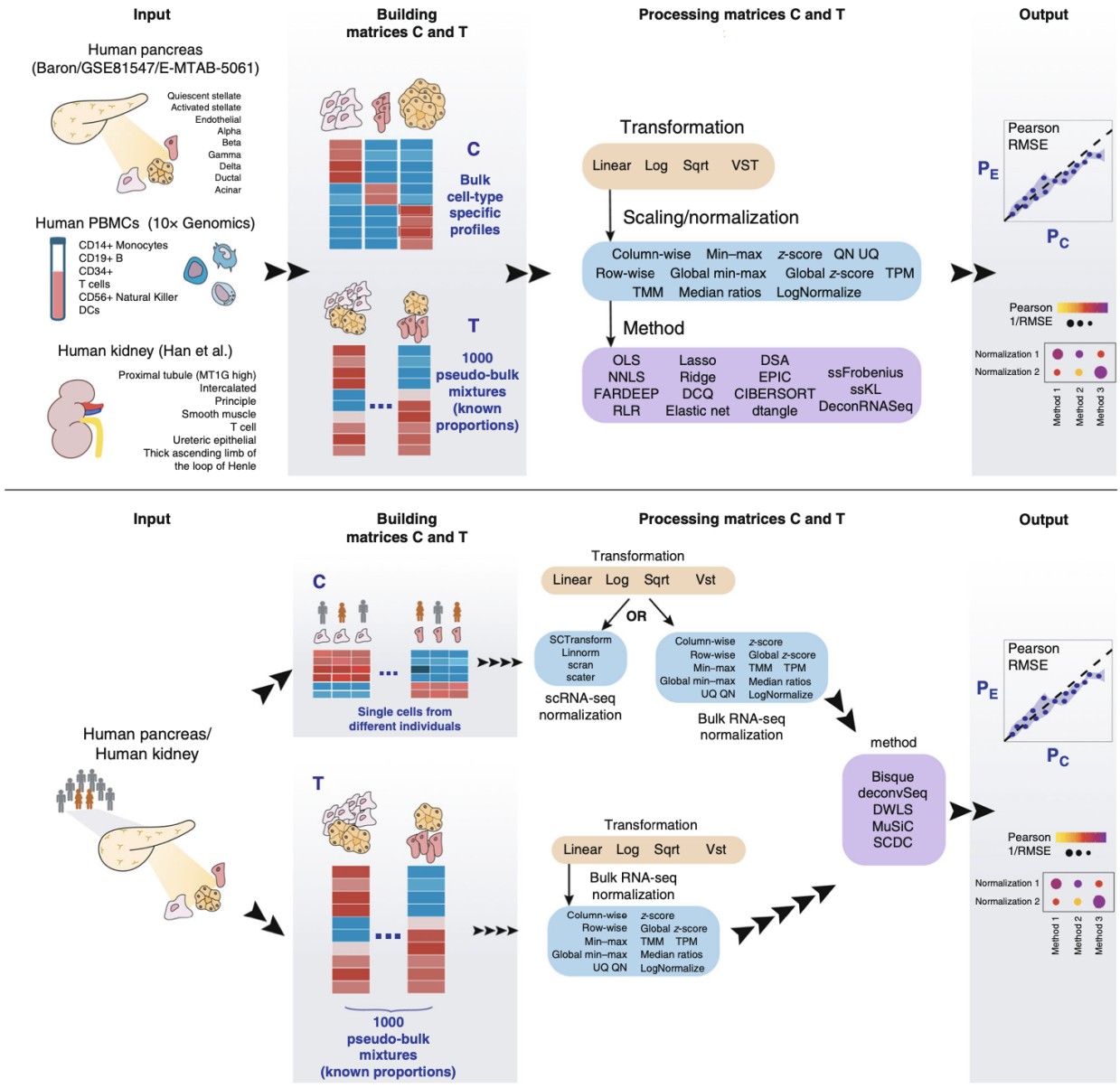

Figure 1. Schematic representation of the benchmarking study.

Top panel: workflow for bulk deconvolution methods. Bottom panel: workflow for deconvolution methods using scRNA-seq data as reference. In both cases the deconvolution performance is assessed by means of Pearson correlation and root-mean-square error (RMSE).

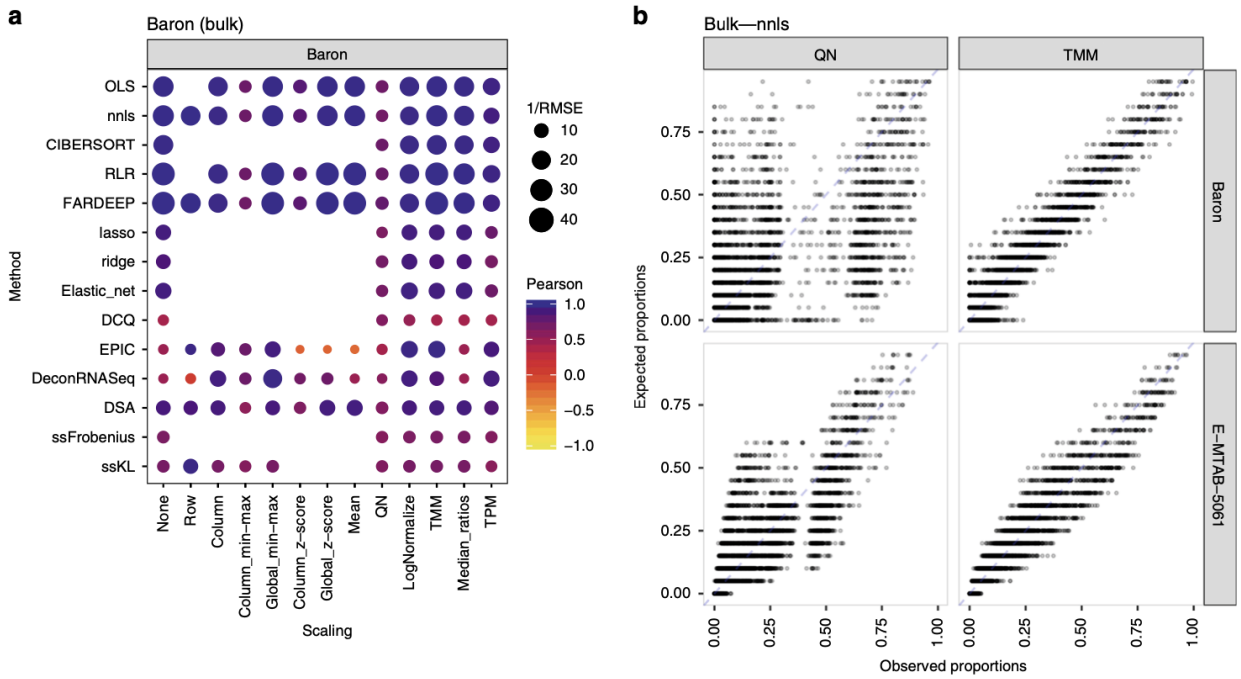

Figure 2. Combined impact of data normalization and methodology on the deconvolution results.

RMSE and Pearson correlation values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures in linear scale (pool size = 100 cells per mixture) and the output proportions from the different bulk deconvolution methods (a). The darker the blue and the higher the area of the circle represents higher Pearson and lower RMSE values, respectively. (b) Scatter plot showing the impact of the normalization strategy (TMM versus quantile normalization (QN)) comparing the expected proportions (y-axis) and the results obtained through computational deconvolution using nnls (x-axis).

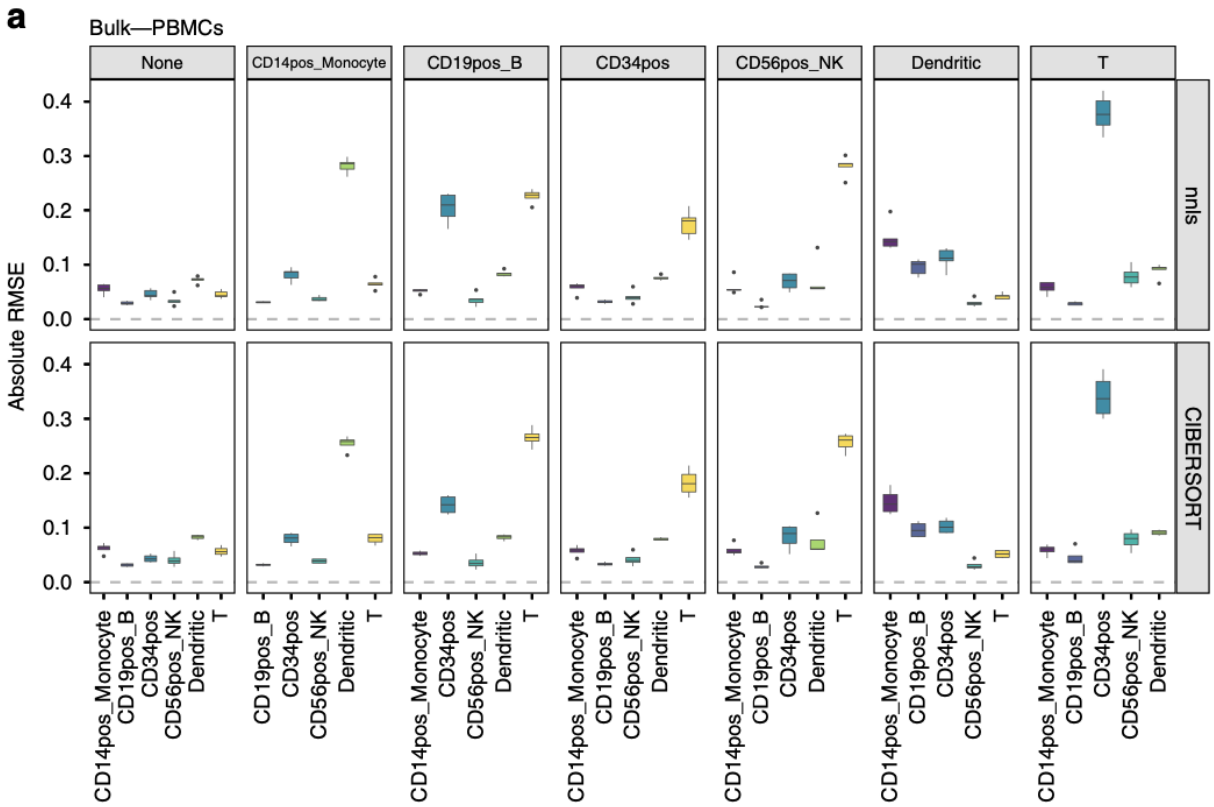

Figure 3. Effect of cell type removal on the deconvolution results for the PBMCs dataset (100-cell mixtures; linear scale).

Results using bulk deconvolution methods (nnls and CIBERSORT). Each gray column represents a specific cell type removed. Each data point conforming a boxplot represents a different scaling/normalization strategy used.

Chapter 2 Identifying markers for missing cell types

The first challenge when designing a computational deconvolution workflow that can handle missing cell types is to identify marker genes for those missing cell types. To this end we hypothesized that genes deviating from the general linear model (= with poor fit) would both marker genes for the unknown cell type and markers (among the known cell types) that we failed to include in the original reference matrix (Figure 4).

To find those genes with poor fit (outliers, genes with high leverage, etcetera) we used several quantitative metrics: Cook's distance; studentized Difference in Fits (DFFITS); studentized residuals and covariance ratios. In short, most of these metrics intrinsically delete one observation (gene) at a time from the reference matrix, re-fit the regression model on remaining (n−1) observations and examine how much all of the fitted values change when the ith observation is deleted.

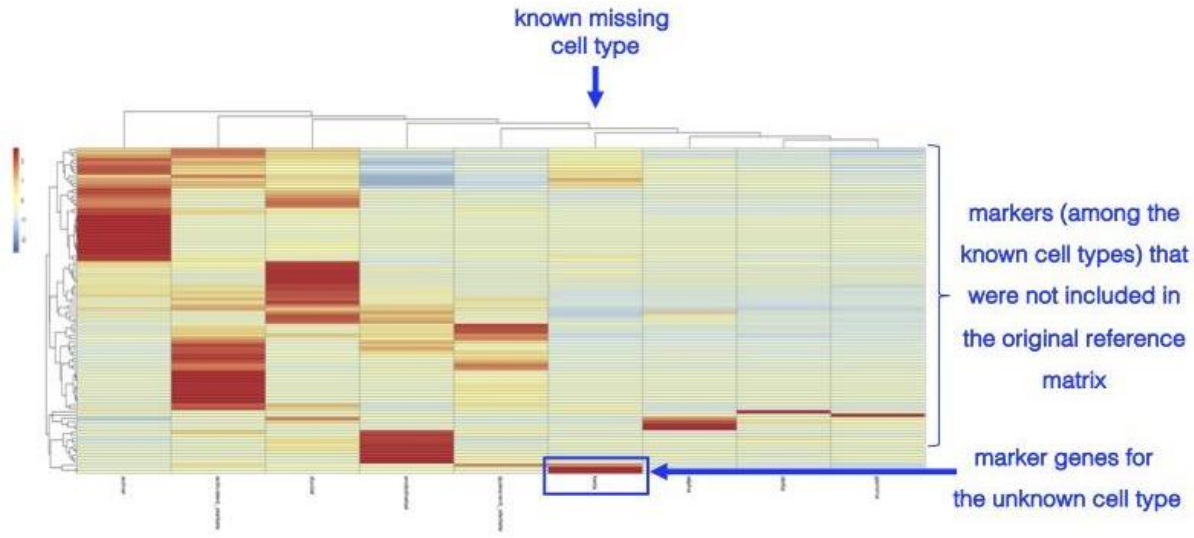

Figure 4. Investigating those genes with poor fit in a scenario where one cell type (beta) is missing in the reference matrix but actually present in the mixtures.

By going back to a full reference matrix (= all cell types present, including beta cells), we identified the two entities we initially hypothesized: a set of genes that were found to be markers for the unknown cell type (blue square, bottom part of the heatmap) and another set that were found to be markers that were not initially included in the reference matrix.

Finally, to obtain expression values for the entire "unknown" column we re-purposed ISOpureR (Anghel *et al.*, 2015), which was originally designed to decompose a patient's particular tumour mRNA abundance profile into its cancer and healthy profile components. Instead, by using a reduced version of both the heterogeneous mixtures and the incomplete reference matrix, we obtained the unknown-specific expression profile that is finally appended to the initial reference matrix and is used in a second step with conventional computational deconvolution methods such as CIBERSORT (Newman *et al.*, 2015).

Chapter 3 Validating and benchmarking the novel workflow

We defined as “complete” the scenario where the reference matrix contains all cell types present in the mixtures and “incomplete” where a cell type was missing in the reference. Furthermore, we labelled as “new” our proposed approach described in Chapter 2. These three reference matrices were used as input for CIBERSORT.

Furthermore, we included EPIC, another state-of-the-art computational deconvolution method which attempts to include the proportion of an unknown cell type present in the mixture by using markers of non-malignant cells that are not expressed in cancer cells.

The results of our analyses are shown in Figure 5 and depict a better performance (lower RMSE and higher Pearson correlation values) of our proposed framework (panel III) compared to situations where the reference is incomplete or where EPIC is used (panels II, V and VI). Panel III also shows a similar performance compared to scenarios where the reference matrix is complete (panels I and IV).

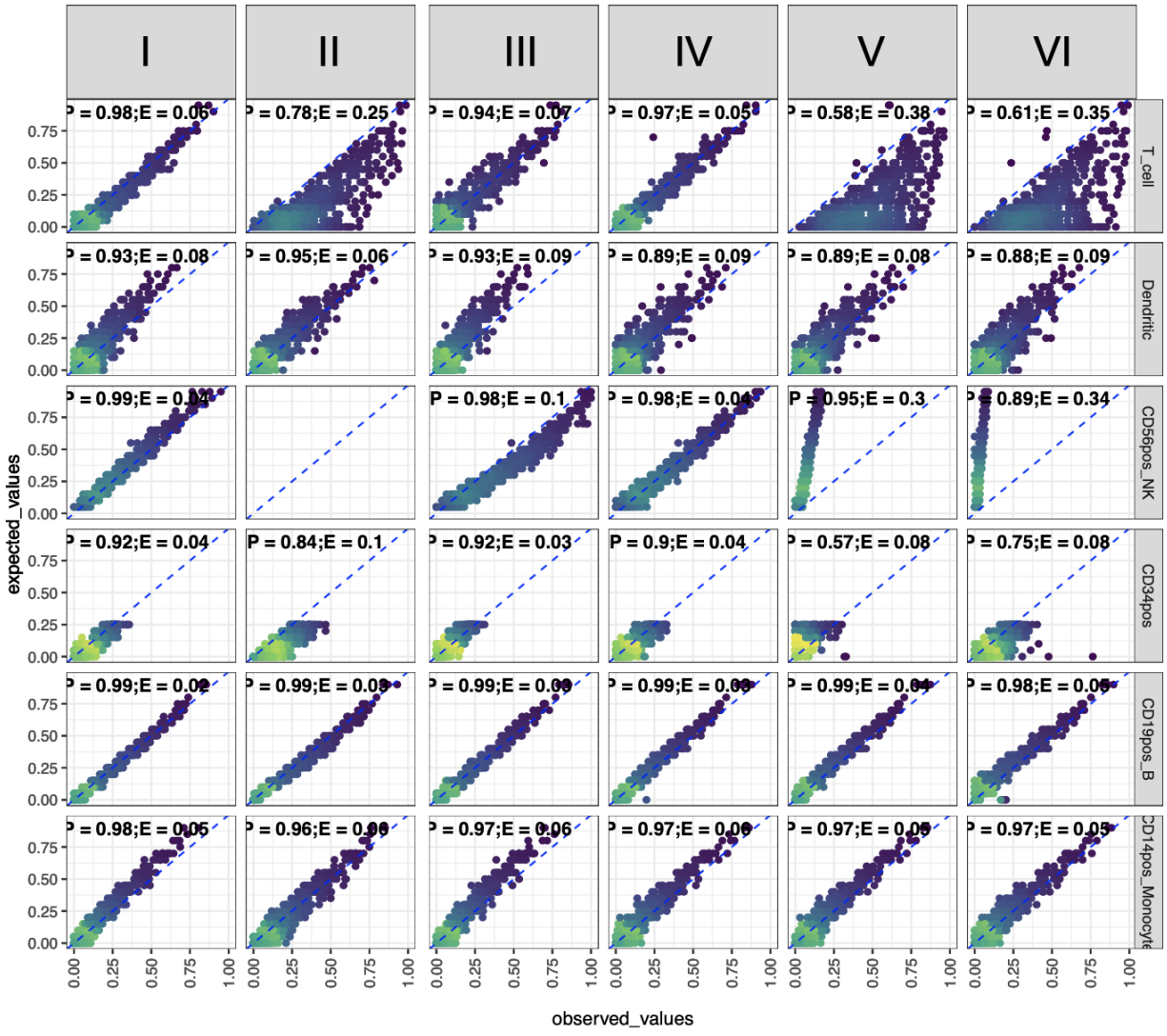

Figure 5. Validation and benchmarking of the novel workflow using pseudo-bulk mixtures from 10x scRNA-seq data of peripheral blood mononuclear cells (PBMCs) where CD56+ natural killer cells were artificially removed from the reference matrix.

Each row represents a given cell type and depicts a scatter plot between the expected (known) cell type proportions and the observed proportions (= output from the computational deconvolution). Panels I to III: “complete”, “incomplete” and “new” reference matrices as input for CIBERSORT; Panels IV to VI: EPIC with three different reference matrices: “complete”, “incomplete” and using the set of genes with poor fit described in Chapter 2. P = Pearson correlation; RMSE = root mean square error.

Chapter 4 Application of novel deconvolution workflow to pediatric tumor datasets

Having established a novel deconvolution workflow to establish the immune cell repertoire using bulk RNA-sequencing data of a tumor sample, we started to build and apply dedicated models for pediatric cancer datasets, collected within iPC. An overview of these available datasets is provided in Table 1. Note that these datasets are also applied for deliverable 3.1, aiming to apply deconvolution by matrix factorization.

	Number bulk datasets (number of samples)		**Number of single cell datasets (number of cells)**	
	patient	adult	patient	control
ES	5 (432)			1 (96)
MB	7 (1,743)	2 (696)	25 (7,745)	
NB	4 (804)		28 (146,800)	9 (63,776)
HB	3 (134)	1 (373)		

Table 1. Description of the datasets collected for each tumor type. ES = Ewing sarcoma; MB = Medulloblastoma; NB = Neuroblastoma; HB = Hepatoblastoma.

In order to establish dedicated models for neuroblastoma, we first started with the integration of two different 10x scRNA-seq datasets from Dong *et al.* (2020) and Kildisiute *et al.* (2021). Figure 6 shows the results of this integration.

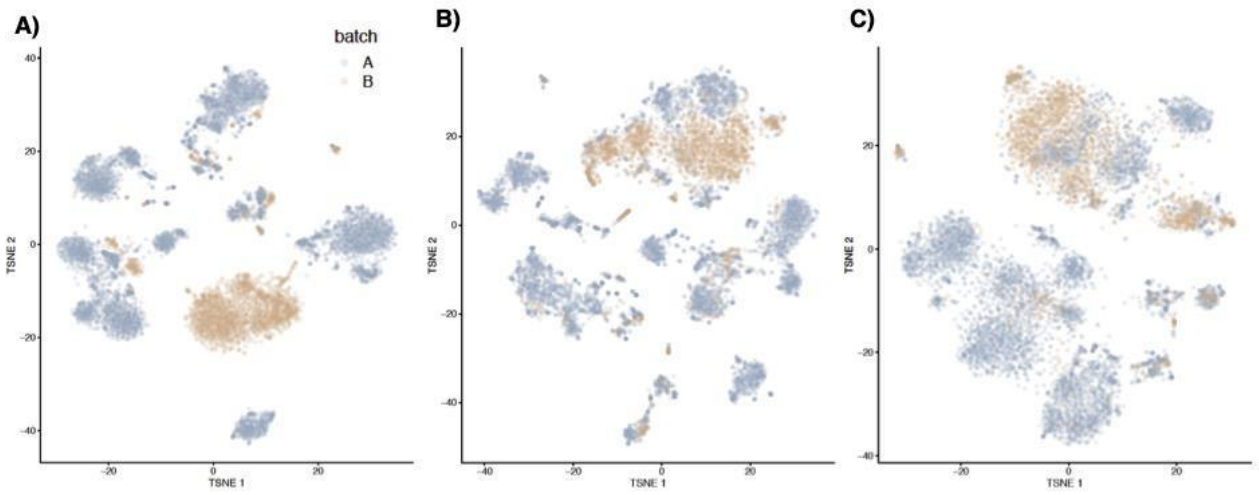

Figure 6. t-distributed stochastic neighbor embedding (t-SNE) representation of the combination of Dong and Kildisiute datasets.

Figure 6A: a clear batch effect is present by merely trying to cluster these datasets as they are. This is alleviated by using fast mutual nearest neighbors correction (fastMNN; https://github.com/LTLA/batchelor/blob/master/R/fastMNN.R) on two different alternatives: top 5000 genes with the largest biological components of their variance (Figure 6B) or a set of informative genes found by the expression entropy model (S-E method) from ROGUE (Liu et al. (2020)) (Figure 6C).

In parallel, we have also downloaded and processed 10x scRNA-seq data from PBMCs.

Our ongoing research has two parallel branches that built upon the framework described in Chapter 2: a) considering only the scRNA-seq data from PBMCs, the unknown component would account for the neuroblastoma tumor signal; b) considering the integrated scRNA-seq neuroblastoma datasets, the unknown component would account for both the immune component and other cell types that may have not been represented in these but can potentially be present in the bulk RNA-seq neuroblastoma samples.

Finally, in collaboration with Pavel Sumazin (iPC partner), we are evaluating our proposed framework on pediatric acute myeloid leukemia (AML) data.

Chapter 5 Summary and Conclusion

Benchmarking of computational deconvolution workflows revealed the importance of a complete reference matrix for accurate deconvolution performance, irrespective of the algorithm. Other general guidelines to maximize the deconvolution performance would be keeping their input data in linear scale and use a stringent marker selection strategy that focuses on differences between the first and second cell types with highest expression values.

Finally, as more scRNA-seq datasets become available in the near future, its aggregation (while carefully removing batch effects) will increase the robustness of the reference matrices being used in the deconvolution and will fuel the development of methodologies similar to SCDC (Dong et al.), which allows direct usage of more than one scRNA-seq dataset at a time.

As single cell RNA-sequencing data for pediatric tumors is still scarce, and what is available most likely does not represent the entire spectrum of heterogeneity between pediatric patient tumor samples, we decided to develop a method that can handle missing cell types in the reference matrix. Here, the ultimate goal is to determine the immune profile of individual pediatric tumor samples using bulk RNA-sequencing data. To this end, we established and benchmarked a deconvolution workflow capable of handling missing cell types (namely the tumor fraction and other component(s)) without severe impact on overall deconvolution accuracy for cell types present in the reference matrix. This method is being applied to large series of bulk RNA-sequencing data from pediatric patient samples available within iPC to establish immune profiles and correlate these to various clinical parameters.

Chapter 6 List of Abbreviations

Abbreviation	Translation
scRNA-seq	single-cell RNA-sequencing
RMSE	root mean square error
P	Pearson correlation
DFFITS	studentized Difference in Fits
PBMCs	Peripheral Blood Mononuclear Cells

Chapter 7 Bibliography

Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020 Nov 6;11(1):5650.

Anghel C, Quon G, Haider S, Nguyen F, Deshwar A, Morris Q, Boutros P. ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. BMC Bioinformatics, 2015 May 14;16:156.

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May;12(5):453-7. doi: 10.1038/nmeth.3337.

Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017 Nov 13;6:e26476.

Dong M, Thennavan A, Urrutia E, Li Y, Perou C, Zou F, Jiang Y. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 416–427

Dong et al. (2020): Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma

Kildisiute et al. (2021); Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma cancer cell

Liu et al (2020): An entropy-based metric for assessing the purity of single cell populations