Empowering bulk RNA-seq deconvolution algorithms by integrating multiple transcriptomics datasets

Martina GALLINARO[1], Vincenzo ALFANO[2,3], Coline KERBAJ[2,3], Giulia MACCARONE[6],
Giovanni MALERBA[1], Marie-Laure PLISSONNIER[2,3], Mirjam ZEISEL[2,3], Massimo LEVRERO[2,3,4,5], Massimiliano COCCA[2,3]

1. Biology and Genetics Section, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona 37134, Italy
2. Cancer Research Center of Lyon (CRCL), UMR Inserm 1052 CNRS 5286 Mixte CLB, Université Lyon 1 (UCBL1), Lyon, France
3. The Lyon Hepatology Institute EVEREST, Lyon, France
4. Department of Hepatology, Croix Rousse hospital, Hospices Civils de Lyon, Service d'Hépato-Gastroentérologie, Lyon, France
5. Department of Internal Medicine - DMISM and the IIT Center for Life Nanoscience (CLNS), Sapienza University, Rome, Italy
6. Department of Molecular Medicine, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185, Roma, Italy

1 Background

Over the past two decades, various methods have emerged to infer cell type proportions from bulk transcriptomics data (i.e. deconvolution methods), including those using single-cell RNA sequencing as a reference scaffold. Developing these methods faces several challenges: building standardized reference datasets, standardizing cell type annotation and marker selection, and improving algorithm and signature atlas generalizability to new bulk sample conditions. Our goal is to implement a pipeline addressing these challenges. Using a single-cell RNA-seq reference panel, we aim to perform gene expression imputation at the cell type level, inspired by "genotype imputation" in GWAS [1]. Here we present our approach to create a standardized and generalized single-cell reference panel with consistent annotation to serve as ground truth for the deconvolution algorithm, to ultimately obtain cell type-level data from bulk gene expression.

2 Materials and Methods

Integration:

For the construction of the reference panel, we included 2 liver-based datasets: GSE149614 [2] and a subset of GSE243981 [3] (details in Table 1), to create a balanced resource without a focus on function/disease. The integration of the two datasets was performed using the Seurat/harmony pipeline [3] and resulted in a panel of 96.159 cells and 16 samples (Table 1).

Dataset ID	Description
GSE149614	71.915 cells, 3 non-viral tumour samples, 7 HBV or HCV related tumour samples
GSE243981	24.242 cells, 6 healthy samples

Table 1: Description of the datasets used for the integration step.

Annotation:

To maximize the standardization of our workflow we performed marker-based annotation of the integrated panel using the software sc-type [5] with a curated list of signatures from GSE149614, GSE243981, and a subset of the CellMarker 2.0 database [6].

The number of genes for each cell type signature is summarized in Table 2. The annotation follows a 2-step approach: 1) initial annotation by main cell type, 2) sub-type identification for each cell type. After the first annotation step, we compared the assigned labels with the labels in the original datasets, and only cells with matching annotations were selected for the second step.

Deconvolution test:

Deconvolution was executed using the b-VAE implementation provided by the bulk2space software (Fig. 1) [7], to generate single-cell-like expression data. Various types of bulk RNA-seq data were used to assess the resource capabilities (Table 3).

Cell type	n° of genes
Hepatocyte	95
T Cell	84
B Cell	23
Endothelial	27
Fibroblast	19
Myeloid	87

Table 2: Number of signature genes for each cell type.

Type	Tissue	N	Description
Bulk RNA-seq	Normal liver	2	GteX; internal HCC dataset
Bulk RNA-seq	Tumor liver	1	Internal HCC dataset
Bulk RNA-seq	PHH	1	Liver resection
Pseudo-Bulk RNA-seq	Normal liver	5	Liver based single sell dataset [8]
Pseudo-Bulk RNA-seq	Cirrhotic liver	5	Liver based single cell dataset [8]

Table 3: Bulk and pseudo-bulk RNA-seq data used in deconvolution tests.

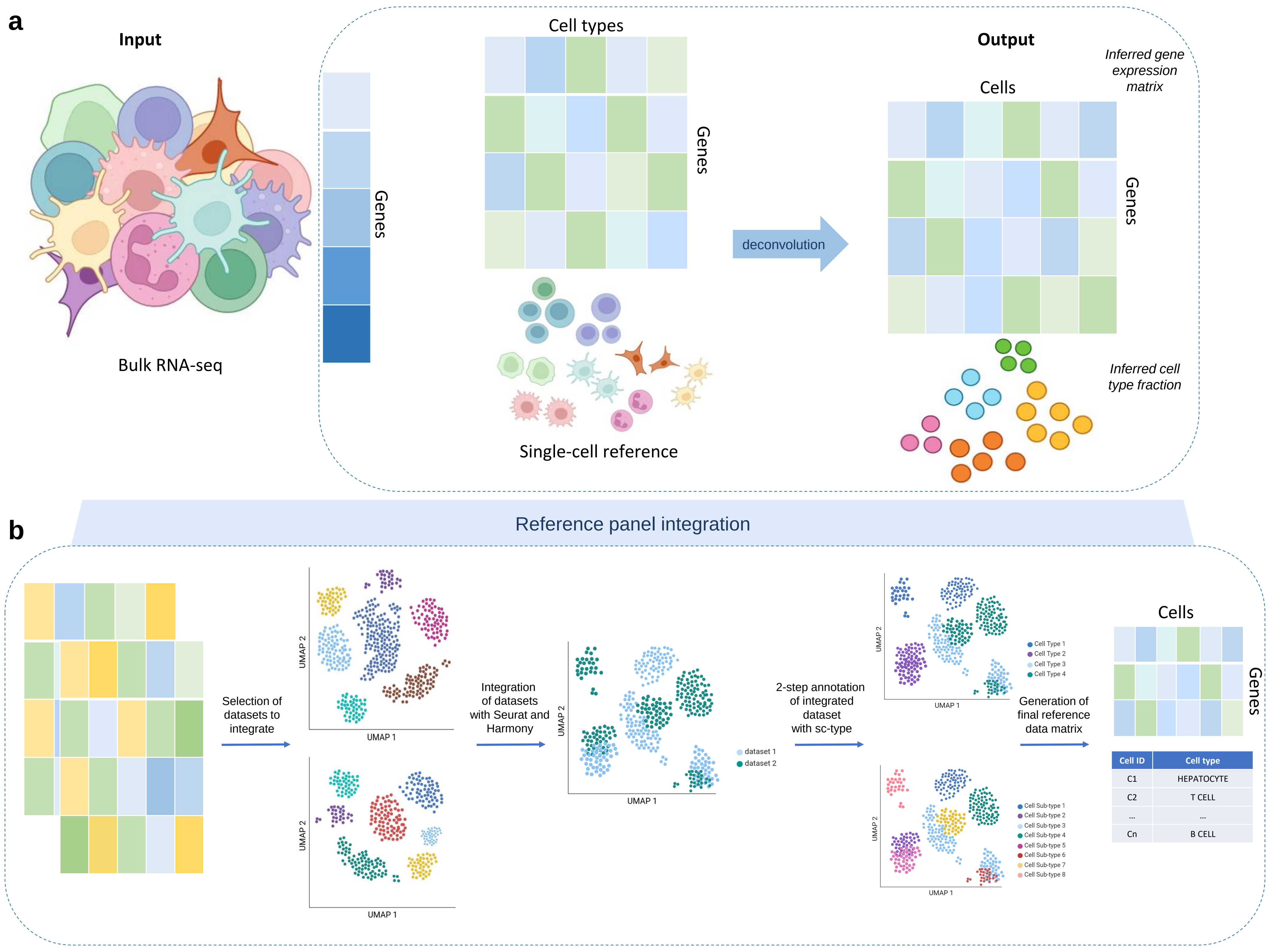

Fig 1: Pipeline workflow. **a)** Overview of the deconvolution pipeline: starting from a bulk RNA sample, a single cell panel is used as reference to characterize the heterogeneous tissue. After the deconvolution phase we obtain single-cell-like data in the form of an inferred gene expression matrix; **b)** Detail of the single cell reference panel creation procedure.

3 Results

Integration and annotation:

We successfully set up a reproducible workflow for single-cell data integration and annotation. The marker-based annotation approach resulted in a 91% concordance of the integrated dataset annotation with the original cell type annotation (Fig. 3, Table 4). We removed the 9% of discordantly annotated cells after the integration step to reduce the possible sources of noise for downstream analyses. UMAP plot of the annotated datasets are shown in Figure 2.

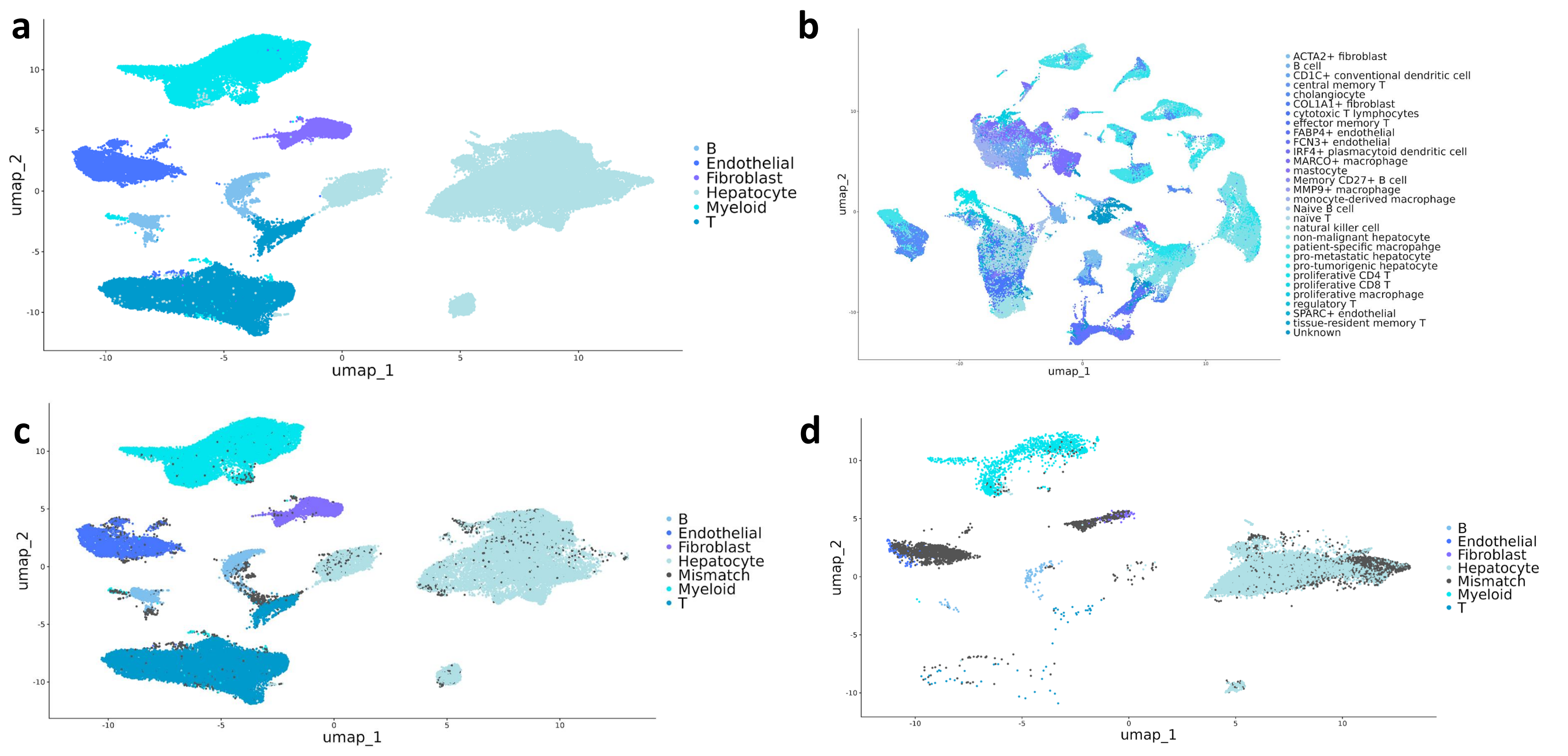

Fig 2: UMAP representation of **a)** the integrated reference panel with cell type annotation and **b)** cell sub-type annotation. **c)** UMAP representation of GSE149614 with cell type annotation and **d)** of GSE243981 with cell type annotation. Grey dots represent mismatched labels between the integrated panel and the original ones.

Cell type	Mislabelled percentage
T	13,81%
Myeloid	6,58%
Endothelial	1,02%
Fibroblast	10,89%
Hepatocyte	5,62%
B	19,02%

Table 4: Percentage of mislabeled cells for each cell type in the integrated reference panel.

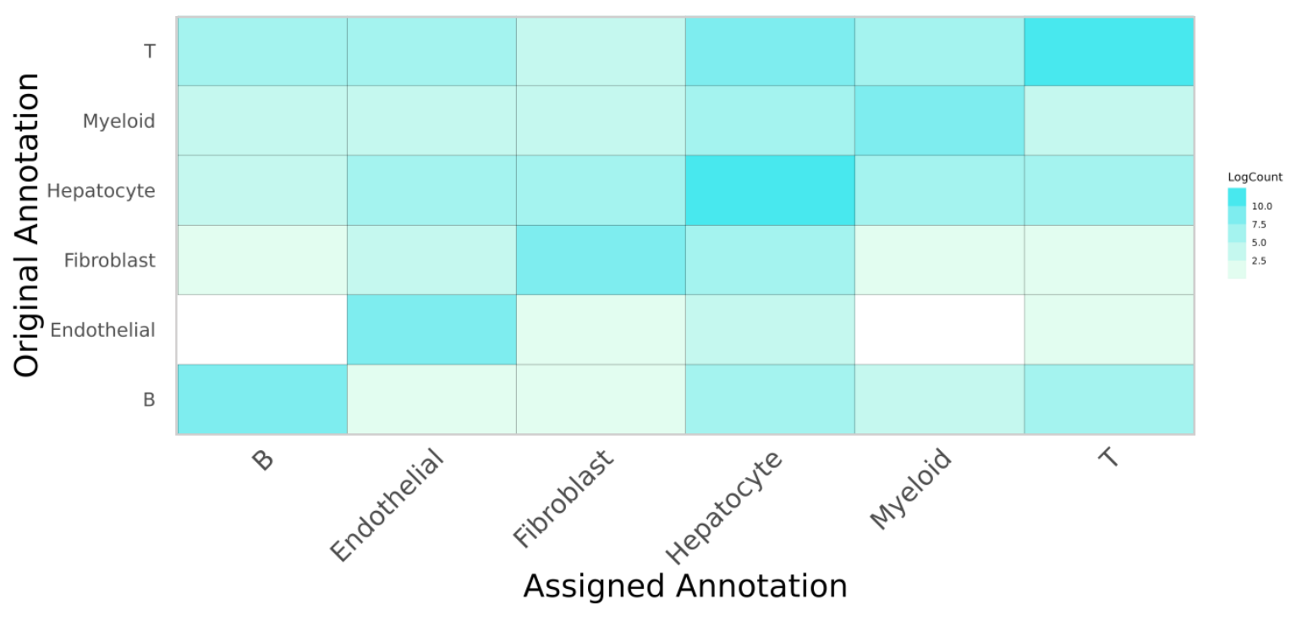

Fig 3: Heatmap of original annotations compared with annotations assigned by sc-type.

Deconvolution test:

Deconvolution results of liver bulk RNA-seq samples and PHH sample resulted in the expected transfer of the "Hepatocyte" label for all generated cells. To have a better grasp on the performance of our resource, we compared deconvolution results of 10 pseudo-bulk RNA-seq samples with their original assigned cell types. We achieved a high correlation between the original and transferred annotation labels (Fig. 4a). We used unsupervised clustering to assess the impact of the deconvolution process on the pseudo-bulk test samples. Figure 4c represents the original single-cell samples clustering, while Figure 4d shows the clustering results after the deconvolution: we can see that the two main clusters are preserved.

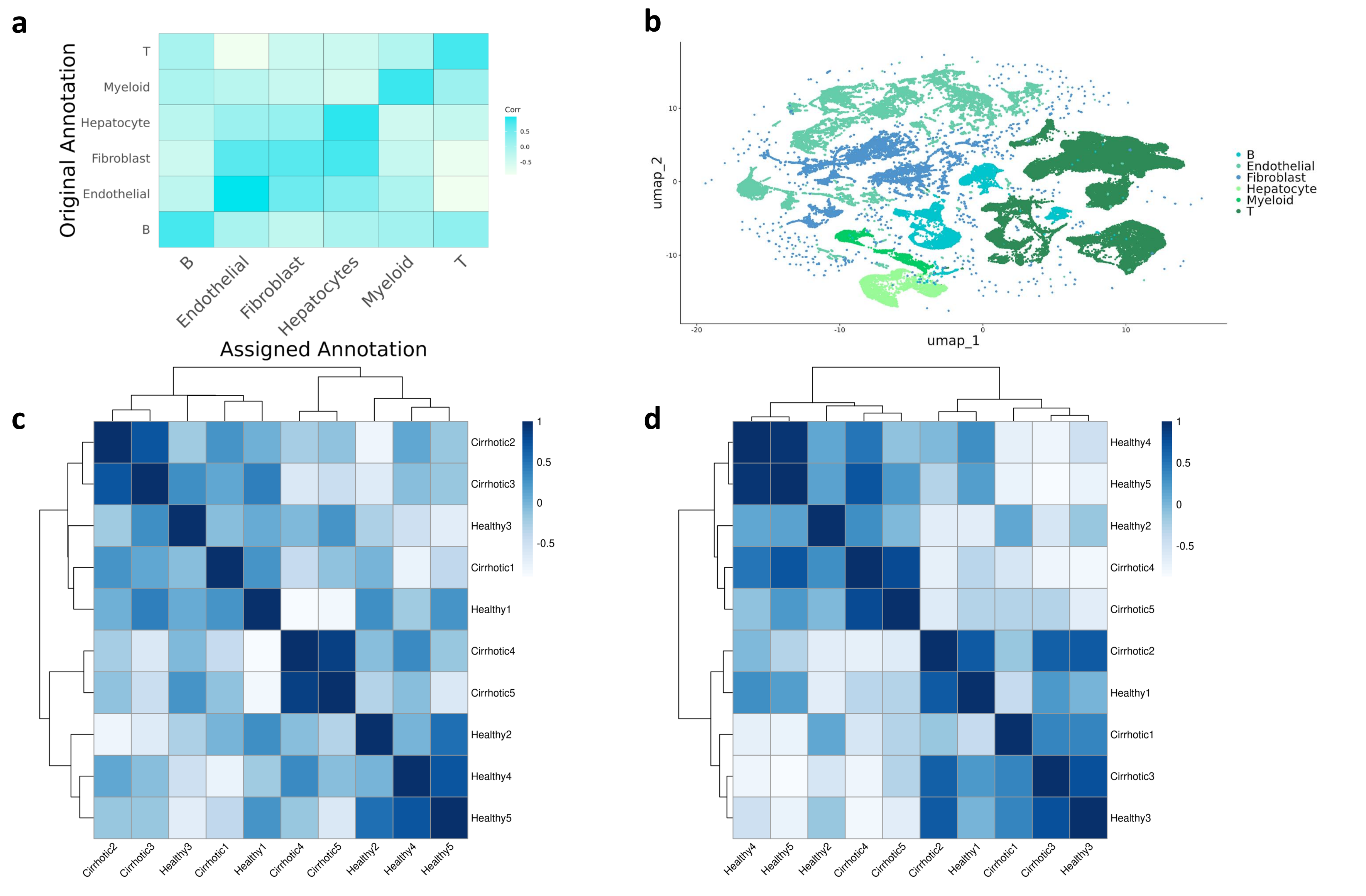

Fig 4: Deconvolution test results. **a)** correlation heatmap of the original single cell annotation compared with the label transfer performed by the deconvolution. **b)** UMAP representation of one deconvoluted pseudo-bulk test sample (Cirrhotic1). **c)** Unsupervised clustering result of the original single-cell sample-to-sample distance matrix. **d)** Unsupervised clustering result of the sample-to-sample distance matrix after deconvolution of pseudo-bulk RNA obtained from single-cell samples.

4 Discussion and Conclusions

Our integration and annotation workflow for the reference panel created a standardized and replicable resource. We observed that the common Seurat/harmony-integration approach still results in "unknown" cell line profiles when re-labeling our integrated dataset using a marker-based annotation strategy (Fig. 2c, 2d). This issue may arise from the normalization procedure, which can introduce artifacts while addressing batch effects during dataset integration. Comparing the integrated and re-annotated dataset with the original annotation, provided an additional quality control (QC) step, minimizing false positives. Preliminary tests on RNA-seq bulk datasets validate that our resource, using the chosen deconvolution method, accurately recovers major cell type labels in matched tissue samples. It also retrieves a significant proportion of cell type labels in pseudo-bulk RNA-seq samples of non-represented etiologies (Fig. 4). This suggests that a multi-purpose reference panel, not oriented towards a specific disease/function, could adapt to new bulk samples with varying conditions or phenotypes. Increasing the number of diverse samples will capture greater variability and offer a more comprehensive representation of different cell populations. This expansion will also improve the applicability of our reference panel to diverse bulk RNA samples. Moreover, we will apply the same 2-step annotation procedure to the deconvoluted bulk RNA-seq results, better characterizing the generated data for further single-cell-like analyses.

References

1. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels. Annu. Rev. Genomics Hum. Genet. 2018;19:73–96.
2. Lu Y, Yang A, Quan C, Pan Y, Zhang H, Li Y, et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. Nat. Commun. 2022;13:4594.
3. Andrews TS, Nakib D, Perciani CT, Ma XZ, Liu L, Winter E, et al. Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver. J. Hepatol. [Internet] 2024.
4. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 2019;16:1289–96.
5. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat. Commun. 2022;13:1246.
6. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51:D870–6.
7. Liao J, Qian J, Fang Y, Chen Z, Zhuang X, Zhang N, et al. De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution. Nat. Commun. 2022;13:6498.
8. Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu NT, et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. Nature 2019;575:512–8.

Corresponding authors: massimiliano.cocca@inserm.fr, martina.gallinaro@univr.it