Analysis of bulk RNA-seq deconvolution using BayesPrism

Notation and Formula

We denoted $P \in R^{K \times S}$ as the unobservable truth of cell type proportions in the samples, and its estimation $\hat{P}$ is the output of deconvolution analysis.

The element of the matrix B, B_{gs}, represents expression levels of the g^{th} gene of the s^{th} sample. The element of matrix A, A_{gk}, shows the g^{th} reference gene expression level for the k^{th} cell type. Each element of $\hat{P}$ is the estimated proportion for the k-th cell type of the s^{th} sample. Note, g = 1, . . . , G is the index of genes, s = 1, . . . , S is the index of samples, and k = 1, . . . , K is the index of cell types. We denote the fitted bulk gene expression matrix as $\hat{B} \in R^{G \times S}$, derived from the following formula: $\hat{B} = A\hat{P}$

The ground truth data, denoted as $P \in R^{K \times S}$, specifies the actual cell type proportions for the corresponding bulk RNA-seq samples. The elements are the true K cell type proportions across S samples. This dataset is essential in the evaluation of the deconvolution model.

The ground truth could be obtained using multiple approaches as follows:

- Flow cytometry analyses, providing quantifiable cell-type fractions through cell sorting based on surface markers.
- Paired bulk and single-cell RNA-seq datasets, where cell type proportions are inferred from the high-resolution single-cell data.

Project Objective

This report proposes the execution of a comprehensive experiment involving two distinct datasets: PBMC and pancreas, based on the insights presented by CybersortX (Newman et al., 2019) and BayesPrism (Chu, 2022). The primary goal is to take advantage of the capabilities of both CybersortX and BayesPrism by using single-cell data as a reference and bulk data as input.

In this experimental framework, the two software packages will compute cell type proportions (denoted as $\hat{P}$) for each dataset independently. CybersortX will generate a signature based on the calculated proportions, and this signature will then be used to multiply $\hat{P}$, resulting in an estimated bulk data matrix (denoted as $\hat{B}$).

The study's analytical focus extends beyond the simple computation of P and B. The next step is to carefully examine the estimated proportion matrix and bulk matrix at both the cellular and sample levels. The goal of this multifaceted analysis is to elicit nuanced insights into the interplay between cell types and the overall composition of the samples. This method allows for a thorough understanding of the biological characteristics embedded in the datasets, providing useful information for subsequent interpretations and inferences.

Data Introduction

1. PBMC

The first dataset (ground truth determined by flow cytometry) is available on the CIBERSORTx website by selecting *Menu* → *Download* → *"NSCLC PBMCs Single Cell RNA-Seq (Fig. 2ab) (zip)"* (CIBERSORTx Download Page). In this folder:

- scRNA-seq reference data (*"Fig2ab-NSCLC PBMCs scRNAseq refsample.txt"*): matrix of 16476 genes and 1054 cells.
- signature matrix created by CIBERSORTx (*"Fig2ab-NSCLC PBMCs scRNAseq sigmatrix.txt"*): matrix of 2498 reference gene expression levels and 6 cell types.
- bulk data (*"Fig2b-WholeBlood RNAseq.txt"*): matrix of 5851 genes and 12 samples
- cell type of interest: "T.cells.CD8", "Monocytes", "T.cells.CD4", "B.cells", "NK.cells"; we discard "NKT.cells" which is omitted by CybersortX.

2. Pancreas

- scRNA-seq reference data (*"pancreas scref_v2.txt"*): matrix of 25525 genes and 410 cells.
- signature matrix created by CIBERSORTx: matrix of 3484 reference gene expression level and 9 cell types.
- bulk data ("*pancreas bulk_v2.txt") :* matrix of 25525 genes and 7 samples.
- The ground truth (*pancreas truth.txt"*): matrix of 10 cell types and 6 cell samples.
- cell type of interest: "delta", "alpha", "gamma", "ductal", "acinar", "beta", "PSC", "endothelial"; we discard "mast" which is omitted by CybersortX.

Data Format

Bulk.data: The sample-by-gene raw count matrix of bulk RNA-seq expression. rownames are bulk sample IDs, while colnames are gene names/IDs.

scRNA.data: The cell-by-gene raw count matrix of bulk RNA-seq expression. rownames are bulk cell IDs, while colnames are gene names/IDs.

cell.type.labels: A character vector of the same length as number of rows of scRNA data to denote the cell type of each cell in the reference.

cell.state.labels: is a character vector of the same length as number of rows of scRNA data to denote the cell state of each cell in the reference. This is only for input of the BayesPrism package. In our example, we are not studying tumor data, thus we ignore this input by setting "key=NULL" in prism function.

Data Preprocess: Filter outlier genes

Highly expressed genes, such as those related to ribosomal proteins and mitochondrial functions, can disproportionately impact the gene expression distribution, introducing bias in the inference

process. This bias may hinder the effective distinction of cell types, leading to significant spurious variation. To enhance accuracy and reliability, the author recommends excluding these genes during the deconvolution process.

We can visualize outlier genes in the single-cell RNA sequencing (scRNA-seq) reference, gaining insights through mean expression levels and cell type specificity scores computed across all cell types. This visualization identifies potential sources of variation within the scRNA-seq reference.

A parallel analysis on bulk RNA-seq data calculates the average expression of each gene across all cell types, unveiling outlier genes in both scRNA-seq and bulk RNA-seq datasets. Using the PBMC dataset as an example, resulting plots highlight outliers in both data types, visually illustrating the impact of these genes on the overall gene expression landscape. Here is the plot showcasing outlier genes within the PBMC dataset.

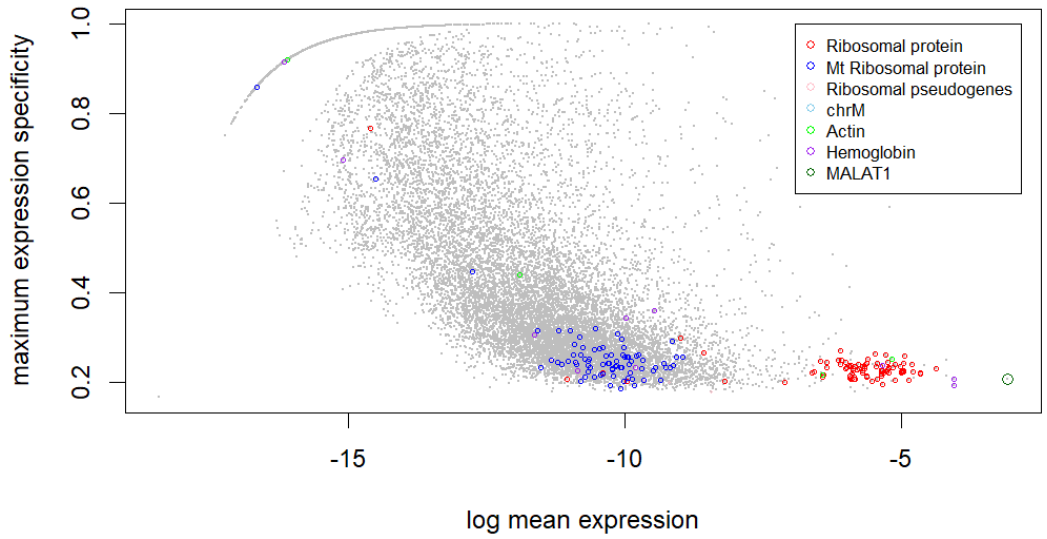

Figure 1 outlier genes in scRNA-seq (PBMC)

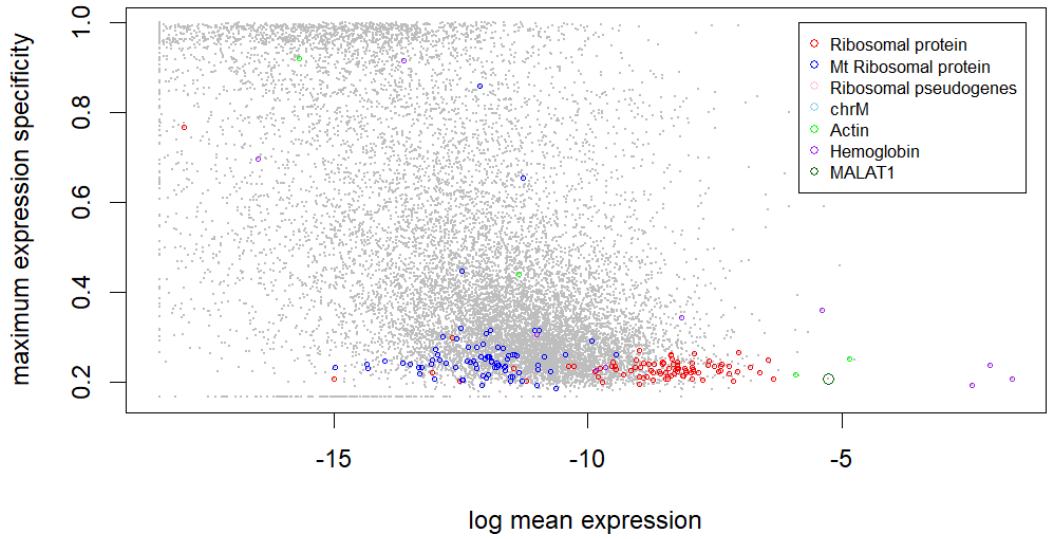

Figure 2 outlier genes in bulk data (PBMC)

Evaluation matrix

The formula for Mean Square Error at the cell-type-level is:

$$MSE_{cell} = \frac{1}{S}\sum_{s=1}^{S}(\hat{p}_{ks} - p_{ks})^2 \tag{2}$$

The Pearson Correlation coefficient is:

$$r_{cell} = \frac{\sum_{s=1}^{S}(\hat{p}_{ks} - \overline{\hat{p}_{k\cdot}})(x_{ks} - \overline{p_{k\cdot}})}{\sqrt{\sum_{s=1}^{S}(\hat{p}_{ks} - \overline{\hat{p}_{k\cdot}})^2}\sqrt{\sum_{s=1}^{S}(p_{ks} - \overline{p_{k\cdot}})^2}} \tag{3}$$

For sample-level analysis, we compare the columns of $\hat{\mathbf{P}}$ and **P** using similar metrics. The formulas are as follows:

$$MSE_{sample} = \frac{1}{K}\sum_{k=1}^{K}(\hat{p}_{ks} - p_{ks})^2 \tag{4}$$

$$r_{sample} = \frac{\sum_{k=1}^{K}(\hat{p}_{ks} - \overline{\hat{P}_{\cdot s}})(p_{ks} - \overline{p_{\cdot s}})}{\sqrt{\sum_{k=1}^{K}(\hat{p}_{ks} - \overline{\hat{p}_{\cdot s}})^2}\sqrt{\sum_{k=1}^{K}(p_{ks} - \overline{p_{\cdot s}})^2}} \tag{5}$$

For the overall matrix evaluation, the MSE is given by:

$$MSE_{overall} = \frac{1}{K \times S}\sum_{k=1}^{K}\sum_{s=1}^{S}(\hat{p}_{ks} - p_{ks})^2 \tag{6}$$

And the overall Pearson Correlation coefficient is calculated as:

$$r_{overall} = \frac{\sum_{k=1}^{K}\sum_{s=1}^{S}(\hat{p}_{ks} - \overline{\hat{p}_{\cdot\cdot}})(p_{ks} - \overline{p_{\cdot\cdot}})}{\sqrt{\sum_{k=1}^{K}\sum_{s=1}^{S}(\hat{p}_{ks} - \overline{\hat{p}_{\cdot\cdot}})^2}\sqrt{\sum_{k=1}^{K}\sum_{s=1}^{S}(p_{ks} - \overline{p_{\cdot\cdot}})^2}} \tag{7}$$

Analysis PBMC data

We can find the ground truth (denoted as $P \in R^{K\times S}$) for the dataset labeled "Ground truth whole blood (txt)" is available on the CybersortX website. This matrix comprises the true proportions of eight cell types across 12 samples. However, both CybersortX (abbreviated as Cyb) and BayesPrism (abbreviated as BP) packages provide only estimated fractions for select cell types, owing to the limitations of matrix size and other factors.

In our analysis, we focus solely on the common cell types in P, P_{Cyb} and P_{BP}:

"T.cells.CD8," "Monocytes," "T.cells.CD4," "B.cells," and "NK.cells." Notably, both packages exhibit a tendency to overestimate cell proportions. "B.cells" and "NK.cells" demonstrate commendable performance in both packages. BayesPrism excels in estimating the fractions of "T.cells.CD4," "B.cells," and "NK.cells," while CybersortX produces superior results for "T.cells.CD8."

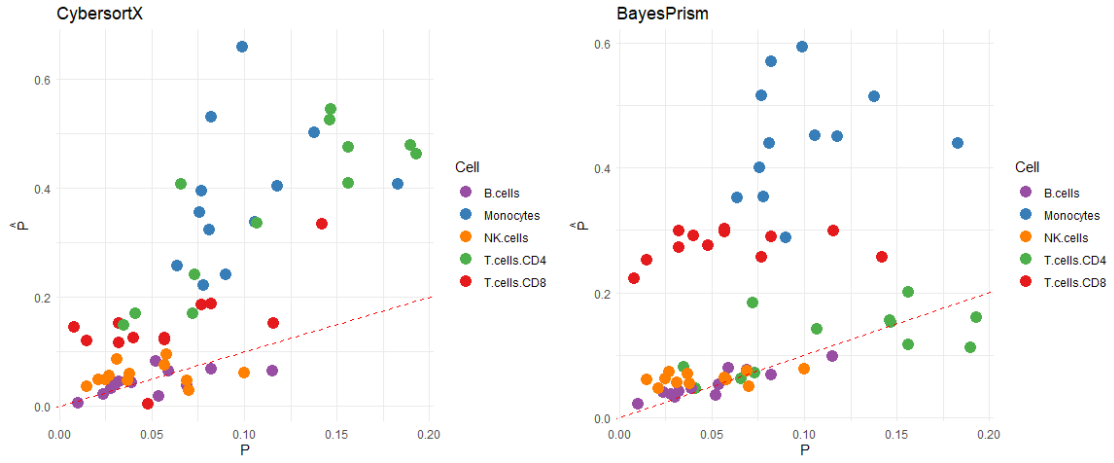

Figure 3 Scatterplot of estimated fraction for PBMC dataset

We could further exam the result by looking into cell type level and sample level of Mean Squared Error (MSE) and Pearson Correlation coefficient (r). The analysis of the fraction estimation results for pancreas cell types using CybersortX and BayesPrism reveals distinct patterns in their performance. In the fraction estimation analysis for PBMC (Table 1) using CybersortX and BayesPrism, distinctive patterns emerge across individual cell types and sample-level metrics. For individual cell types, both methods exhibit varying degrees of performance measured by MSE and correlation coefficient. CybersortX generally outperforms BayesPrism in estimating cell proportions for T cells CD8, Monocytes, and NK cells, as evidenced by lower MSE values and higher correlation coefficients. On the other hand, BayesPrism excels in accurately estimating T cells CD4 and B cells, demonstrating lower MSE and higher correlation values. At the sample level, both CybersortX and BayesPrism show fluctuating performance across different samples. Notably, sample W070517001160 stands out with near-perfect correlation and low MSE for both methods, while other samples exhibit varying degrees of accuracy. In the overall assessment, CybersortX outperforms BayesPrism in terms of both MSE and correlation for cell type fraction estimation.

In the bulk estimation analysis for PBMC (Table 2), a similar trend is observed with varying performance across samples for both CybersortX and BayesPrism. Notably, the overall assessment shows that CybersortX outperforms BayesPrism in terms of both MSE and correlation for bulk estimation.

The fraction estimation analysis for PBMC indicates that CybersortX generally performs better than BayesPrism in accurately estimating cell type proportions. This superiority is evident across individual cell types and sample-level metrics. The bulk estimation results align with this trend, reinforcing CybersortX's overall better performance in capturing the cellular composition of PBMC samples. The findings highlight the importance of considering both individual cell types and overall metrics when evaluating the performance of fraction estimation methods.

Tables for fraction estimation (PBMC):

	T cells CD8	Monocytes	T cells CD4	B cells	NK cells
MSE_cell_CybersortX	0.0489232	0.0108031	0.3527025	0.8617802	0.1286645
MSE_cell_BayesPrism	0.0956598	0.4292237	0.6896613	0.0022281	0.0715204
r_Cell_CybersortX	0.2930551	0.9484372	0.0001611	0.0005210	0.5519921
r_Cell_BayesPrism	0.8281485	0.0007675	0.0009481	0.7757775	0.8554698

	MSE_samp_CybersortX	MSE_samp_BayesPrism	r_samp_CybersortX	r_samp_BayesPrism
W070517001156	0.0601781	0.0397064	0.0286312	0.0149859
W070517001157	0.0681008	0.0288575	0.0373642	0.0256765
W070517001159	0.8431436	0.1639830	0.0058729	0.2570479
W070517001160	0.9803383	0.9783675	0.8933019	0.7537479
W070517001161	0.0361908	0.0183825	0.0514986	0.0336790
W070517001162	0.0442414	0.0196201	0.0448988	0.0293331
W070517102034	0.2438570	0.9437744	0.6307584	-0.1494216
W070517102035	0.9545480	0.9480204	0.9591039	0.7651818
W070517102036	0.0574810	0.0241669	0.0289304	0.0399560
W070517102037	0.0470502	0.0212165	0.0294107	0.0349161
W070517102038	0.4539980	-0.1385551	0.7761115	0.7368194
W070517102051	0.6104912	0.8605750	0.9675696	0.9396141

	MSE_all	r_all
overall_CybersortX	0.0361489	0.0358905
overall_BayesPrism	5.7225018	8.0941073

Table 1 Tables for fraction estimation (PBMC)

Tables for bulk estimation (PBMC):

	MSE_samp_CybersortX	MSE_samp_BayesPrism	r_samp_CybersortX	r_samp_BayesPrism
W070517001156	1.319546e+05	6.685077e+04	4.939735e+04	2.341469e+04
W070517001157	1.656799e+05	3.379625e+04	2.583211e+04	1.506268e+04
W070517001159	9.985770e-01	9.941452e-01	9.766753e-01	9.710344e-01
W070517001160	9.981295e-01	9.984838e-01	9.764437e-01	9.561642e-01
W070517001161	7.177169e+04	3.690347e+04	1.064546e+05	5.915099e+04
W070517001162	3.711239e+04	2.965028e+04	6.117898e+04	4.557737e+04
W070517102034	9.880961e-01	9.922675e-01	9.948591e-01	9.718925e-01
W070517102035	9.881290e-01	9.936189e-01	9.951285e-01	9.673582e-01
W070517102036	1.282312e+05	4.160056e+04	6.122457e+04	7.825943e+04
W070517102037	1.105148e+05	1.449805e+04	4.802010e+04	7.390563e+04
W070517102038	9.709818e-01	9.441979e-01	9.920839e-01	9.950818e-01
W070517102051	9.774458e-01	9.562618e-01	9.910188e-01	9.948166e-01

	MSE_all	r_all
overall_CybersortX	7.126782e+04	5.506904e+04
overall_BayesPrism	1.993000e-04	2.028000e-04

Table 2 Tables for bulk estimation (PBMC)

Analysis pancreas data

Similar to previous analysis for PBMC dataset, the ground truth (denoted as $P \in R^{K \times S}$) for the dataset labeled " *pancreas truth.txt*" has true proportions of twelve cell types across 7 samples. Here we focus solely on the eight common cell types in P, P_{Cyb} and P_{BP}: "delta", "alpha", "gamma", "ductal", "acinar", "beta", and "PSC".

In general, when comparing the performance of CybersortX and BayesPrism on the PBMC dataset, CybersortX continues to exhibit superior performance, particularly in the estimation of "acinar," "delta," "gamma," "PSC," and "endothelial" cell types. BayesPrism, on the other hand, notably falters in the estimation of "delta" but showcases improved performance for the "beta" cell type. Notably, both packages encounter challenges in accurately estimating the "alpha" cell type.

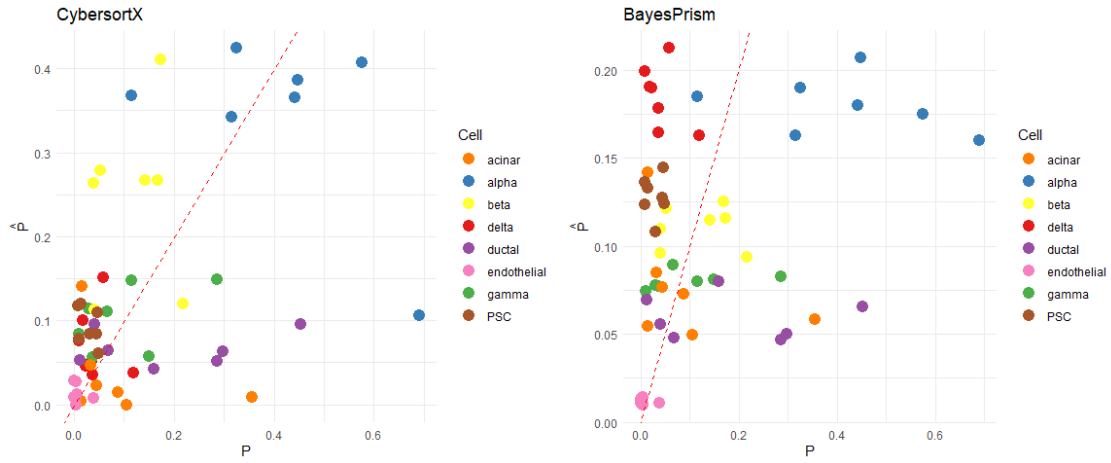

Figure 4 Scatterplot of estimated fraction for pancreas dataset

In the examination of individual cell types detailed in Table 3, CybersortX demonstrates a tendency toward higher Mean Squared Error (MSE) values, indicating a greater deviation from the ground truth proportions. This is particularly noticeable for ductal and acinar cell types. Conversely, BayesPrism consistently produces lower MSE values across various cell types, suggesting a generally more accurate estimation of cell proportions. Furthermore, BayesPrism consistently outperforms CybersortX in terms of correlation coefficients, reflecting a stronger linear relationship with the ground truth. Noteworthy variations in performance at the sample level are observed, with T2D2 displaying elevated MSE in BayesPrism.

In summary, the analysis highlights the nuanced performance of CybersortX and BayesPrism in fraction estimation for pancreas cell types. While CybersortX tends to exhibit higher MSE values, indicating greater deviation from the ground truth, BayesPrism consistently demonstrates lower MSE and higher correlation coefficients, reflecting a more accurate estimation of cell proportions. The sample-level analysis unveils variability in performance across specific samples. Overall, the findings suggest that the choice between CybersortX and BayesPrism depends on the specific priorities of the analysis, considering the trade-off between precision and bias in cell proportion estimation.

Tables for fraction estimation (pancreas):

	delta	alpha	gamma	ductal	acinar	beta	PSC	endothelial
MSE_cell_CybersortX	0.0223751	0.0039217	-0.3504644	0.8275831	0.0878031	0.0648132	0.4883967	0.6504096
MSE_cell_BayesPrism	0.0079501	0.0061516	-0.0591843	0.4869869	0.0397966	0.0364939	0.0624887	0.0478234
r_Cell_CybersortX	0.0161686	0.0218611	0.2385266	0.6906455	0.0048048	0.0282010	0.5723815	0.5810399
r_Cell_BayesPrism	0.0102612	0.0052469	-0.2649209	0.3779061	0.0001860	0.0003521	0.8724120	0.9039611

	MSE_samp_CybersortX	MSE_samp_BayesPrism	r_samp_CybersortX	r_samp_BayesPrism
H3	0.0256791	0.1589608	0.4808062	0.0243618
H4	0.0331811	0.0134633	0.9390541	0.0042664
H6	-0.3412411	0.0042530	0.0199598	0.3529117
T2D1	-0.0886969	0.4817391	0.0161727	0.0073134
T2D2	0.0414228	0.8689000	0.3474176	0.0066053
T2D3	0.0532611	0.0304258	0.5883135	0.5964767
T2D4	0.4289909	0.0083264	0.0274130	0.7611620

	MSE_all	r_all
overall_CybersortX	0.0236682	0.0208802
overall_BayesPrism	4.3794409	2.0879726

Table 3 Tables for fraction estimation (pancreas)

Tables for bulk estimation (pancreas):

	MSE_samp_CybersortX	MSE_samp_BayesPrism	r_samp_CybersortX	r_samp_BayesPrism
H3	4.840579e+05	4.903743e-01	7.871882e-01	1.698733e+05
H4	6.553671e+05	2.506696e+05	9.784685e-01	6.396799e-01
H6	4.577488e-01	4.974920e+04	3.081751e+05	8.104825e-01
T2D1	3.783529e-01	7.943034e-01	1.400007e+05	8.470927e+04
T2D2	8.746864e+05	9.643570e-01	7.232645e-01	6.732727e+04
T2D3	9.957440e+05	5.571523e+05	8.851967e-01	9.173045e-01
T2D4	5.841716e-01	1.435953e+05	2.775806e+05	9.449658e-01

	MSE_all	r_all
overall_CybersortX	4.052902e+05	3.173796e+05
overall_BayesPrism	8.010000e-05	4.930000e-05

Table 4 Tables for bulk estimation (PBMC)

Conclusion

This paper details a structured experiment utilizing CybersortX and BayesPrism to analyze two diverse datasets: PBMC and pancreas. The primary objective was to estimate cell type proportions and gain insights into the overall biological composition of the samples, employing single-cell data as a reference and bulk data as input.

To enhance the accuracy of subsequent deconvolution processes, the initial data preprocessing emphasized filtering out highly expressed genes related to ribosomal proteins and mitochondrial functions. Additionally, the visualization of outlier genes in both single-cell and bulk RNA-seq datasets provided crucial insights into potential sources of variation.

Considering computational efficiency, it's noteworthy that CybersortX stands out due to its avoidance of Gibbs Sampling, making it much preferred in terms of computational time. The estimated fraction on the CybersortX website, obtained through "Run CIBERSORTx," typically takes a few minutes. In contrast, the BayesPrism package's fraction calculation, executed on an R script on Compute Canada, requires several hours. Importantly, attempting to run such large datasets on local R software may lead to program interruption due to insufficient memory.

The subsequent analyses on PBMC and pancreas datasets uncovered distinct patterns in CybersortX and BayesPrism performance. In PBMC, CybersortX excelled in estimating cell type proportions, with significant differences observed between individual cell types. This trend was corroborated in the bulk estimation results, underscoring CybersortX's efficacy in capturing the cellular composition of PBMC samples.

In the pancreas dataset, CybersortX maintained its superiority over BayesPrism, particularly in estimating proportions for specific cell types. Although CybersortX displayed higher MSE values, BayesPrism showcased lower MSE and higher correlation coefficients, indicative of a more accurate estimation of cell proportions. This nuanced performance was evident at both the cell and sample levels.

In summary, the findings suggest that CybersortX and BayesPrism possess distinct advantages and disadvantages, with CybersortX generally outperforming BayesPrism in estimating cell type proportions. The choice between the two methods should be guided by specific analysis priorities, considering the balance of precision and bias in cell proportion estimation. This report contributes valuable insights into the tools' strengths and limitations, aiding researchers in making informed decisions based on the goals of their analyses. Additionally, the consideration of computational efficiency underscores the practical implications of choosing between these methods in large-scale analyses.

Reference

Chu, T., Wang, Z., Pe’er, D. et al. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 3, 505–517 (2022). https://doi.org/10.1038/s43018-022-00356-3

Newman, A.M., Steen, C.B., Liu, C.L. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 37, 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2