Profiling cell type abundance and expression in bulk tissues with CIBERSORTx

Authors

Chloé B. Steen[1], Chih Long Liu[1,2], Ash A. Alizadeh[1,2,3,4] and Aaron M. Newman[2,5]

Affiliations:

[1]Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA, USA.

[2]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA.

[3]Center for Cancer Systems Biology, Stanford University, Stanford, CA, USA.

[4]Division of Hematology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA, USA.

[5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA.

Correspondence should be addressed to A.M.N. (amnewman@stanford.edu) or A.A.A. (arasha@stanford.edu)

Abstract

CIBERSORTx is a suite of machine learning tools for the assessment of cellular abundance and cell type-specific gene expression patterns from bulk tissue transcriptome profiles. With this framework, single-cell or bulk-sorted RNA sequencing data can be used to learn molecular signatures of distinct cell types from a small collection of biospecimens. These signatures can then be repeatedly applied to characterize cellular heterogeneity from bulk tissue transcriptomes without physical cell isolation. Here, we provide a detailed primer on CIBERSORTx and demonstrate its capabilities for high-throughput profiling of cell types and cellular states in normal and neoplastic tissues.

1. Introduction

Tissue composition is a major determinant of phenotypic variation and a key factor influencing disease outcomes. For example, in malignant tumors, dynamic interactions between heterogeneous subpopulations of cancer cells, stromal elements, and immune subsets differentially impact patient survival [1,2] and can be leveraged therapeutically [3]. Common assays for studying cellular heterogeneity, such as flow cytometry and immunohistochemistry, require fluorescently labeled antibodies to mark and distinguish diverse cell types of interest. Despite their utility, these approaches are limited by the availability of highly specific antibodies and by the number of cell types that can be simultaneously assessed. While single-cell RNA sequencing (scRNA-seq) has recently emerged as a powerful strategy for resolving cellular heterogeneity at high-resolution and without prior knowledge [4-7], it remains impractical for large-scale analyses and cannot be applied clinically to formalin-fixed, paraffin embedded (FFPE) tissue biopsies.

To complement these approaches, especially in settings where tissue is limited, fixed, or challenging to disaggregate, computational methods for dissecting cellular composition directly from bulk tissue gene expression profiles (GEPs) have been previously described [8-20]. These techniques for “digital cytometry” mathematically deconvolve RNA admixtures into their component cell types, allowing significant new insights into bulk tissue expression profiles [21]. Our group previously developed CIBERSORT, an *in silico* strategy that employs a machine learning model to enumerate cell composition from bulk tissue GEPs [16]. Like other deconvolution techniques [22,12,15,17], CIBERSORT relies on a specialized expression matrix of cell type-specific “barcode” genes, often called a “signature matrix,” which provides a reference atlas of known cellular signatures for the deconvolution procedure.

To improve the performance and versatility of digital cytometry, we recently extended CIBERSORT into a new computational framework called CIBERSORTx [23]. Unlike previous methods (**Table 1**), CIBERSORTx can (1) leverage scRNA-seq-derived reference profiles for bulk tissue dissection, (2) overcome technical variation across different platforms (e.g., scRNA-seq, bulk RNA-seq, microarrays) and tissue preservation techniques (e.g., fresh/frozen vs. FFPE), and (3) digitally “purify” cell-type-specific expression profiles from bulk tissues without physical cell isolation. In this chapter, we describe each of these new features and demonstrate how CIBERSORTx can be broadly applied to dissect cellular heterogeneity from complex tissues without antibodies, tissue dissociation, or viable cells.

2. Materials

CIBERSORTx is available as an online tool with a user-friendly interface that does not require prior bioinformatics training or programming experience

(http://cibersortx.stanford.edu). Its key functionalities are divided into three main components (**Figure 1**):

1. Creation of a custom signature matrix from scRNA-seq or bulk sorted RNA-seq (or microarray) data.

2. Estimation of cell type composition in bulk tissue GEPs.

3. Imputation of cell type-specific expression profiles from bulk tissue GEPs.

In the following sections, we describe each component in detail and provide guidance on how to design and execute a CIBERSORTx analysis. All datasets used in this chapter are available at http://cibersortx.stanford.edu, under "*Menu*" → "*Download*". The website also offers a series of dedicated tutorials covering a wide range of potential use cases and applications (under "*Menu*" → "*Tutorial*").

3. Methods

3.1 Construction of a custom signature matrix from scRNA-seq data

CIBERSORTx requires the use of marker gene reference profiles to enumerate cell subsets in bulk tissue samples. Commonly referred to as a "signature matrix," marker gene profiles can be derived from a variety of sources, including scRNA-seq data and sorted cell populations profiled by bulk RNA-seq or microarrays [21,23]. We previously described a microarray-derived signature matrix for profiling 22 functionally defined human immune cell types, called LM22 [16]. With scRNA-seq, it is now possible to create a signature matrix that captures all major cell subsets in a tissue without complex sorting experiments (see **Note 1**). This can enable large-scale investigation of novel or poorly understood phenotypic states in bulk tissue GEPs. In malignancies, cellular states of interest may include subpopulations of activated, resting, or exhausted T cells [24-26], cancer-associated fibroblasts [27], or cancer cells [28,29], including tumor initiating cells or cancer stem cells [30]. As further described below, scRNA-seq can also provide a powerful means of validating reference signatures through the use of mixture samples created from single-cell transcriptomes.

In the next section, we describe how to create a signature matrix from scRNA-seq data. Other platforms are described in detail elsewhere (see tutorials 6 and 7 at http://cibersortx.stanford.edu/tutorial.php).

3.1.1 Input file

In order to create a custom signature matrix from scRNA-seq data, CIBERSORTx requires a *single cell reference matrix file*. This file consists of a tab-delimited scRNA-

seq expression matrix saved as *.txt* or *.tsv*, where each row is a gene and each column is a single-cell transcriptome. The sum of the expression of all genes for a given cell should not be zero. Gene names must populate the first column of the file. CIBERSORTx will append unique numerical identifiers to redundant gene symbols, but we recommend that users remove any redundant gene names prior to uploading their file. In addition, each single cell must be assigned a cell phenotype by the user in row 1 (e.g., "CD8 T cell", "B cell"), and there should be at least 3 cells per phenotype. The same cell label should be used for each cell belonging to a particular phenotype, and periods should be avoided unless employed to separate the phenotype label from a numerical label. For example, users can label the cells, “CD8 T cell”, “CD8 T cell.1”, or “CD8 T cell.2”, but not “CD8 T.cell” or “CD8.T.cell”. Suffixes separated by a period will be removed during the construction of the signature matrix. Any cells without a cell label (i.e., unassigned) should be excluded before uploading the reference matrix file.

Importantly, clustering and *de novo* identification of cell types from scRNA-seq data are not currently supported by CIBERSORTx. As such, all cell labels must be provided by the user. There are several software tools available to identify cell phenotypes for generating such cell labels when starting from unlabeled scRNA-seq data [31-37] (also, see **Note 2**). For more information about the formatting of the scRNA-seq reference matrix file, see **Notes 3-5**.

In the following sections, we use a publicly available scRNA-seq dataset consisting of 7,688 human peripheral blood mononuclear cells (PBMCs) to illustrate the process of creating a single-cell-derived signature matrix with CIBERSORTx. This dataset is available for download from 10x Genomics (https://www.10xgenomics.com/) and from the CIBERSORTx website (“*Menu*” → “*Download*” → “*10x scRNA-Seq reference matrices (zip)*” → “*Fig2e-5PBMCs_scRNAseq_matrix.txt*”). By applying Seurat [31] along with prior knowledge of key marker genes, we previously identified eight major cell types in this dataset [23]: NK cells, monocytes, B cells, NKT cells, CD4 T cells, CD8 T cells, dendritic cells and megakaryocytes (**Figure 2A**).

3.1.2 File upload

Once the single-cell reference matrix file is formatted according to the instructions above, it may be uploaded by selecting “*Menu*” → “*Upload files*” and “+ *Add file*”. The user will then be able to select their input file, provide a title, and select the “*File Type*”. It is important to specify the correct file type to ensure that the input file appears in the appropriate drop-down menu when configuring the CIBERSORTx analysis. For the current application, select the file type, “*Single Cell Reference Matrix*”, under “*Cell Fractions Module*”.

3.1.3 Building the scRNA-seq signature matrix

Having uploaded the single-cell expression matrix, we can now select "*Menu*" → "*Run CIBERSORTx*". We choose the analysis module, "*1. Create Signature Matrix*", the "*Custom*" analysis mode, and "*scRNA-seq*" as our input data type (**Figure 2B**).

In the "*Single cell reference matrix file*" dropdown menu, our newly uploaded file is available for selection. We provide a name for the signature matrix we are about to create under "*Custom sig file name*". Providing a file name is required in order to identify the signature matrix in future analyses. See **Note 6** for addition details on parameter configuration. Once the single-cell reference file is uploaded and the CIBERSORTx job is configured, click "*Run*" to start creating the signature matrix.

The output of the job will contain (1) the new signature matrix saved as a *.txt* file with the file name provided by the user, (2) the reference sample and phenotypic classes files created by CIBERSORTx as an intermediate step to build the signature matrix, and (3) a heat map of the signature matrix that is organized to show patterns of differentially expressed genes (**Figure 2C**). The newly created signature matrix will be automatically available from the "*Signature matrix file*" dropdown menu for future jobs.

3.1.4 Validation of the scRNA-seq signature matrix

When creating a new signature matrix, it is critical to evaluate its reliability for resolving each cell subset in bulk tissue GEPs. Below we provide some suggestions on how users can validate the performance of a single-cell-derived signature matrix before applying it to real tissue specimens.

First, it is important to check whether known marker genes are present and highly expressed in the correct cell subsets. For example, *MS4A1*, a B cell marker gene that encodes CD20, is expected to be selected as a marker gene for the B-cell population in our running example. If the same cell subsets have been profiled elsewhere, either as bulk sorted populations or by scRNA-seq, one can also evaluate the generalizability of genes in the signature matrix for correctly classifying each cellular phenotype (e.g., see Newman et al, 2015 [16]).

Second, if scRNA-seq data are used to build a signature matrix, it is straightforward to characterize its performance using synthetic tissues created from single-cell transcriptomes. To ensure an unbiased assessment, these source scRNA-seq transcriptomes used for the creation of a synthetic tissue should be held-out from the creation of the signature matrix. Moreover, to avoid violating linearity assumptions, each single-cell transcriptome should be represented in non-log linear space prior to creating synthetic mixtures. By allowing for fine-grained control over the composition of each mixture, this strategy allows one to systematically evaluate both proportion estimation

and cellular detection limits without the cost and time associated with profiling new samples with associated ground-truth expectations of compositional representation.

Finally, the gold standard approach for validating a signature matrix is to compare deconvolution performance against orthogonal methods, such as flow cytometry or immunohistochemistry (see **Note 7**). This is done by imputing the cell fractions of a bulk GEP dataset with CIBERSORTx and by comparing the imputed cell proportions with known ground truth proportions in each sample (for details of this process, see section 3.2).

3.2 Impute cell fractions with CIBERSORTx

In this section, we demonstrate how to apply a scRNA-seq-derived signature matrix to resolve cellular composition in 12 whole blood RNA-seq profiles (GEO accession number: GSE127813). This dataset is available from the CIBERSORTx website by selecting "*Menu*" → "*Download*" → "*Expression Datasets (zip)*" ("*Fig2b-WholeBlood_RNAseq.txt*"). Importantly, for each of these 12 samples, ground truth cell proportions determined by flow cytometry and complete blood counts are available ("*Fig2b_ground_truth_whole_blood.txt*" under "*Expression dataset (zip)*" in "*Download*").

We have preprocessed the mixture dataset into transcript-per-million (TPM) space (also, see **Note 8**). Like signature matrices, all bulk mixture GEPs should be in linear space (not $\log_2$) without negative or missing values. In addition, for optimal performance, the signature matrix and mixture files should be normalized identically. The mixture file should be uploaded as described for the single-cell reference matrix file in section 3.1.2 (i.e., by selecting "*Upload Files*" under the "*Menu*" tab). When uploading the file, choose "*Mixture*" under "*All Modules*" as the file type.

Having generated a signature matrix in section 3.1.3 and uploaded our mixture file, we may now select "*Menu*" → "*Run CIBERSORTx*" and choose the "*2. Impute Cell Fractions*" analysis module and "*Custom*" analysis mode (**Figure 3A**). In the "*Signature matrix file*" drop-down menu, we select the file that we previously generated (section 3.1.3).

When configuring the analysis, we have the option of selecting "*Batch correction*". An important caveat with the precursor of CIBERSORTx is that it did not address platform-specific variation (e.g., between scRNA-seq and RNA-seq). In the next section, we describe how CIBERSORTx addresses this important issue.

3.2.1 Cross-platform deconvolution

Owing to technical variation between different platforms and between different tissue-preservation techniques (for example FFPE vs. fresh-frozen tissues), we have implemented a batch correction method within CIBERSORTx to allow the application of a signature matrix derived from one protocol to bulk mixtures GEPs derived from another protocol. Batch correction is available in two modes: (1) *bulk,* or *B-mode*, and (2) *single-cell,* or *S-mode*. A decision tree to help users identify the mode that is best suited for their analysis is provided in **Figure 3B**. **Table 2** lists examples of signature matrices and mixtures pairs that would require batch correction, and the type of batch correction that we recommend be applied. Deconvolving these datasets without batch correction may lead to cell types being misestimated due to uncorrected technical variation. For batch effects within the mixture or scRNA-seq datasets, see **Notes 9 and 10**.

B-mode batch correction

B-mode – or bulk-mode – batch correction is appropriate for deconvolution when a signature matrix is derived from bulk sorted cell populations or from scRNA-seq data generated by plate-based scRNA-seq protocols without unique molecular identifiers (UMIs) (e.g., SMART-Seq2, which is more similar to bulk RNA-seq). B-mode requires a minimum of three mixture GEPs, but we recommend at least 10 mixture samples for optimal performance. B-mode will adjust the mixture dataset so that it is in the same space as the signature matrix. The adjusted mixtures will then be used for estimating the cell fractions.

S-mode batch correction

S-mode – or single-cell mode – batch correction is tailored for single cell-derived signature matrices generated from droplet-based protocols or protocols that utilize UMIs, such as 10x Chromium [38], InDrop [5] or Drop-Seq [6]. Unlike protocols that profile full transcripts (e.g., bulk RNA-seq and SMART-Seq2), these technologies capture gene expression data from the 3’ or 5’ end of each transcript and employ UMIs to remove PCR duplications and to assign transcripts to individual cells. When running S-mode, an additional input file is required, which is the single-cell reference matrix that was used to build the signature matrix. The drop-down menu for selecting this file automatically appears when the user selects S-mode as an option. S-mode also requires a minimum of three samples in the mixture dataset, but at least 10 is recommended.

Unlike B-mode, which adjusts the mixture dataset, S-mode adjusts the signature matrix, which is then used to estimate the cell fractions. The resulting adjusted output will be available for download after running a CIBERSORTx job with batch correction.

3.2.2 Configuring a CIBERSORTx Fractions job

Since the signature matrix that we created in section 3.1.3 is derived from a UMI-based scRNA-seq dataset generated on the 10x Chromium platform, we select "*S-mode batch correction*". We then select the corresponding single-cell reference profile that we used for building the signature matrix.

CIBERSORTx provides an empirical p-value to evaluate deconvolution performance. The p-value is calculated by comparing imputed fractions in a given mixture dataset with fractions that would have been obtained by random chance. The higher the number of permutations, the more reliable the p-value estimate will be. For our analysis, we set the number of permutations to 100.

Having configured the CIBERSORTx Fractions jobs, press "*Run*".

Once the job is complete, results will appear in tabular format as a heatmap. A stacked bar plot representation showing the distribution of cell fractions for each mixture sample will also be produced.

3.2.3 Evaluation of estimated cell fractions

Since ground truth cell proportions are available for the 12 blood samples analyzed in Section 3.2.2 through the use of automated blood counting and flow cytometric immunophenotyping, we can directly assess the accuracy of our deconvolution results. Doing so can provide insights into the robustness of the signature matrix and the impact of batch correction. **Figure 3C** shows the distribution of Pearson correlations between the fractions inferred by CIBERSORTx and ground truth proportions for each cell type, both before and after S-mode batch correction (also, see **Note 11**). Clearly, batch correction provides a substantial improvement in the average Pearson correlation across all five cell types. For example, while B cells failed to be imputed without batch correction (**Figure 3D, left**), after S-mode batch correction, B cells are successfully imputed (**Figure 3D, right**).

3.3 Impute cell-type-specific gene expression with CIBERSORTx

In this section, we review how to use CIBERSORTx for imputing cell type-specific transcriptomes from bulk tissue GEPs without the need for physical cell sorting. To impute cell type-specific gene expression profiles, select the analysis module, "*3. Impute gene expression*" and "*Custom*" analysis mode (**Figure 4A**). Two distinct modes are available:

Group-mode: This approach infers a representative transcriptome profile for each cell type in the signature matrix from a group of bulk tissue transcriptomes (**Figure 4B**).

Group mode is useful for learning context-dependent changes in a cell's expression profile when a biological class (or grouping) is already known. For example, one can apply group-mode to study cell-type specific transcriptional differences between responders and non-responders to a therapy, or between tumor samples and the adjacent normal tissues.

High-resolution: This approach imputes cell-type specific GEPs from bulk tissues at sample-level resolution. As a result, the output of high-resolution mode is an expression matrix for each cell type with the same number of samples as the input mixtures (**Figure 4C**). High-resolution mode is useful for exploring cell type expression variation without prior knowledge of biological or functional groupings. It can be used, for example, to investigate cell type specific gene expression in tumors in relation to patient survival or for identifying novel transcriptional states.

3.3.1 Group-mode expression imputation

CIBERSORTx group-mode imputes cell type-specific gene expression profiles from a group of bulk tissue GEPs and will learn a single representative transcriptome profile for each cell type in the signature matrix across the set of mixture samples being considered. The required inputs for group-mode are a signature matrix and a mixture file in the same format as described in sections 3.1.1 and 3.2. Optionally, users may wish to merge the cell types in the signature matrix into broader phenotypic classes in order to reduce the number of distinct cell subsets to deconvolve (see **Note 12**). To do this, CIBERSORTx supports the use of a "*Merged class file*", which is a simple text file consisting of one row with the new class labels separated by tabs. The "*Merged class file*" described allows users to group cell types in the signature matrix into a broader set of phenotypic classes. This makes it easy to reduce the number of evaluable cell types in cases where the number of mixture samples is limiting. The order of the labels must match the ordering of cell types in the signature matrix, and a label must be provided for every cell type in the signature matrix. If a merged class file is not provided, all cell types in the signature matrix will be used.

The main output of a CIBERSORTx group-mode analysis is a gene expression matrix (non-log) with genes as rows and cell types as columns (**Figure 4B**). Importantly, CIBERSORTx employs an adaptive noise filter that eliminates unreliably estimated genes for each cell type. Both the filtered and unfiltered results are saved to file. The unfiltered results contain all of the genes that were present in the input mixture file, while the filtered results contain only genes that could be reliably imputed for a given cell type. The more abundant a cell type, the more genes are imputed for that cell type [23].

Tutorial 4 on the CIBERSORTx website provides detailed instructions for setting up a group-mode analysis (also, see **Note 13**).

3.3.2 High-resolution expression imputation

CIBERSORTx high-resolution mode derives cell type-specific gene expression profiles at the sample-level. Although the input requirements are the same as group-mode, the output is an expression matrix for each cell type rather than a single representative transcriptome profile per cell type (**Figure 4C**). Tutorial 7 on the CIBERSORTx website provides detailed instructions for running high-resolution mode. We summarize key steps below.

3.3.3 Input files

To illustrate CIBERSORTx high-resolution mode, we use a gene expression matrix of 150 diffuse large B-cell lymphoma (DLBCL) samples [39]. The file is available from the CIBERSORTx website by selecting "*Menu*" → "*Download*" and "*High Resolution GEPs - DLBCL (Supp. Fig. 11) (zip)*" ("*SuppFig11-DLBCL_CHOP_Lenz-arrays-bulktumors.MAS5.txt*"). Users may upload the file by selecting "*Menu*" → "*Upload Files*" and "*Mixture*" as the file type (for details on the file upload process, see section 3.1.2).

High-resolution mode is computationally intensive. Therefore, to reduce running times, we recommend that users provide a 'gene subset file' consisting of a list of at most 1,000 genes (see **Note 14** for full-transcriptome analyses). DLBCL can be divided into two major molecular subtypes that are strongly associated with distinct clinical outcomes [40]: Activated B-cell (ABC) DLBCL and Germinal Center B-cell (GCB) DLBCL. Using the file type, "*Gene subset – High-Resolution Mode*", we upload a list of genes that are known to be differentially expressed between the two classes and that are primarily associated with B cells. The corresponding 'gene subset file' is available in "*High Resolution GEPs - DLBCL (Supp. Fig. 11) (zip)*" ("*SuppFig11-DLBCL-GCBABC-genes.txt*").

3.3.4 Configure a high-resolution job

Once all required input files are uploaded, select "*Menu*" → "*Run CIBERSORTx*", then "*3. Impute Cell Expression*", "*Custom"*, and "*High-Resolution*".

In this example, we will use "*LM22 (22 immune cell types)"* [16], a signature matrix that is available by default in the signature matrix drop-down menu. Select "*LM10 merged into 10 major cell subsets*", which is available in the "*Merged classes file*" drop-down menu. LM10 groups the LM22 cell types into 10 broader phenotypes, thereby increasing the ratio between the number of samples and cell types, which will improve

deconvolution performance (see **Note 12** and Newman et al. [23]). For the mixture file, select the DLBCL GEP dataset uploaded in the previous section. Then select the ABC/GCB gene list uploaded under "*Gene Subset file*". Finally, check the box, "*Group genes by hierarchical clustering in output heat map*", and press "*Run*". Note that in some instances, it may be appropriate to apply batch correction when performing expression imputation. For details, see **Note 15**.

3.3.5 Output of high-resolution mode

Expression matrices: The main output of high-resolution mode is a set of expression matrices, one for each evaluated cell subset. Each matrix contains imputed expression profiles (non-log space) with genes as rows and samples as columns. These expression matrices can be downloaded from the CIBERSORTx job results page and used for downstream analyses elsewhere.

Heat maps: The output of high-resolution mode is also presented as a series of heat maps organized by cell type. Each heat map shows the imputed expression of the selected cell type across all of the samples (columns) and analyzed genes (rows). All samples and genes are presented in the same ordering as the original input mixture matrix. To highlight expression variation, each gene is $\log_2$ adjusted and mean-centered. Genes that have little to no expression variation across the cohort will be represented in black. If the option "*Group genes by hierarchical clustering*" has been selected, the genes are ordered by the results of the cluster analysis.

t-SNE plots: For each cell type with at least 20 imputed genes, the CIBERSORTx website will perform t-distributed stochastic neighbor embedding (t-SNE) to generate a two-dimensional projection of the imputed expression profiles. In the DLBCL example, only B-cells have enough imputed genes from the 142 gene-list to be rendered as a t-SNE plot. If users have pre-defined labels for their samples, they may upload a custom label file to highlight the samples in the t-SNE plot according to these labels. To highlight known GCB/ABC classes, select the pre-uploaded file "*supp. fig. 10 tSNE Class Labels*" under "*Select tSNE plot Class Labels"*.

3.3.6 When to run CIBERSORTx group-mode versus high-resolution mode:

Because high-resolution mode performs deconvolution of cell-type-specific gene expression profiles at the sample level, it is useful for discovering novel cellular states in a dataset [23]. However, high-resolution mode requires a large number of samples for better performance (see **Note 12**). In settings where users have prior knowledge of sample classes (e.g., responders and non-responders to a therapy, different molecular subtypes of a given cancer, etc.), cell type-specific differentially expressed genes can be identified in several ways:

1. Run CIBERSORTx high-resolution mode on the entire dataset, then identify differentially expressed genes between known classes using standard methods (e.g., as described in Newman et al. [23]).
2. Run CIBERSORTx high-resolution mode on the two classes separately, then combine the results by intersecting genes that are imputed in both classes. Of note, while this approach has higher power to identify differentially expressed genes than the above strategy, it requires more samples to run high-resolution mode on each class.
3. Run group-mode on each class separately, and then apply the following R script to identify statistically significant differentially expressed genes.

```
for (i in 1:ncol(geps1)){
      vBetaZ <- sapply(1:nrow(geps1), function(j) (geps1[j,i]-
geps2[j,i])/sqrt(stderr1[j,i]^2+stderr2[j,i]^2))
      ZPs <- 2*pnorm(-abs(vBetaZ))
      Zqvals <- p.adjust(ZPs, method="BH")
}
```

Where:

- `geps1` are the cell-type-specific GEPs outputted by CIBERSORTx group mode for sample class 1.
- `geps2` are the cell-type-specific GEPs outputted by CIBERSORTx group mode for sample class 2.
- `stderr1` are the standard errors for sample class 1 outputted by CIBERSORTx Group mode and saved to file as *CIBERSORTxGEP_[…]_GEPs_StdErrs.txt*.
- `stderr2` are the standard errors for sample class 2 outputted by CIBERSORTx Group mode and saved to file as *CIBERSORTxGEP_[…]_GEPs_StdErrs.txt*.
- `i` is the cell type number for which we are calculating differentially expressed genes between class 1 and class 2.
- `BH` stands for Benjamini-Hochberg, and is the method used to adjust the p-values for multiple hypothesis testing.

The script will output a list of genes that are significantly differentially expressed between the two classes after multiple hypothesis testing.

3.4 Conclusion

CIBERSORTx is a novel computational framework for high-throughput dissection of cellular heterogeneity in bulk tissue genomic profiles. Unlike previous methods, it handles cross-platform technical variation, enabling the use of scRNA-seq reference signatures for deconvolution, and can derive cell-type-specific gene expression profiles

at sample-level resolution. We anticipate that CIBERSORTx will complement existing experimental methods for studying complex tissues, with implications for the discovery of novel phenotypic states and the identification of more effective biomarkers and therapeutic targets.

4. Notes:

1. While the per-sample-cost of scRNA-seq experiments remains relatively high, CIBERSORTx leverages scRNA-seq data from a small number of samples to characterize the cellular heterogeneity of bulk GEPs from a large number of tissue samples. Users can also download publicly available scRNA-seq datasets to create custom signature matrices.

2. To facilitate the identification of cell types in a single-cell RNA-seq dataset, it is possible to use deconvolution or other reference-based methods to label the cells [41,22,12,15,23,17].

3. The scRNA-seq dataset should be in non-log space [42] with no missing values. Data from scRNA-seq experiments summarizing expression levels using unique molecular identifiers (UMIs) are acceptable. If the maximum expression value is <50, CIBERSORTx will assume a prior logarithmic transformation, and anti-log all expression values by 2^x.

4. CIBERSORTx will automatically normalize the input data such that the sum of all normalized reads is the same for each transcriptome. If a gene length-normalized expression matrix is provided (e.g., RPKM), then the signature matrix will be adjusted to TPM (transcripts per million). If a count matrix is provided, the signature matrix will be normalized to CPM (counts per million).

5. While increasing the number of cells per phenotype and the number of biological replicates can improve the quality of the signature matrix (up to ~20 cells per phenotype and up to 2-3 donor samples), simulation experiments and empirical observations suggest that as few as 3 cells per phenotype and as few as one donor sample can still generate reliable results [23]. For this reason, and to limit the amount of space and time necessary to run a CIBERSORTx job, we recommend that users restrict the number of cells to at most 5,000 when uploading the scRNA-seq reference profile to the CIBERSORTx website (currently a strict upper limit of 10,000 cells is allowed).

6. When creating a signature matrix with scRNA-seq data derived from droplet-based methods (e.g., 10x Genomics [38], InDrop [5] or Drop-seq [6]), the user may wish to change the single-cell-specific parameter, *“Min. Expression”,* under

"*Single Cell Input Options"*. This parameter refers to the minimum average $\log_2$ expression of a gene per cell phenotype (default = 0.75), and is used to filter low-level noise in plate-based scRNA-seq experiments (e.g., data generated by the SMART-Seq2 protocol) that may result from contamination, index swapping [43], or other sources. As this threshold can be too restrictive for data produced by droplet-based platforms, which capture a smaller number of genes, we generally recommend reducing this parameter to 0.50 or even 0 when processing droplet-based scRNA-seq data. Otherwise, the sparsity of the data may yield too few genes for creating a reliable signature matrix.

7. When working with solid tissues and comparing the performance of CIBERSORTx with technologies that require tissue dissociation (e.g., flow or mass cytometry), it is important to generate the gene expression profiles and perform fluorescence activated cell sorting (FACS) or scRNA-seq profiling on the same cell suspensions. If expression profiles are instead generated from intact non-dissociated tissues, CIBERSORTx results will not capture dissociation-induced differences in cell composition (e.g., loss of myeloid cells due to adherence to plastics, loss of plasma cells and plasmablasts due to their fragility, etc.), leading to biases that artificially degrade CIBERSORTx performance in the comparative analysis.

8. CIBERSORTx performs a feature selection within the deconvolution step, and typically does not use all genes in the signature matrix for this step. It is therefore generally acceptable if some genes in the signature matrix are missing from the mixture file.

9. When we refer to batches in the context of expression deconvolution, we refer to two input datasets, the signature matrix and the mixture dataset. However, within a mixture dataset, one may also observe technical variation due to batch effects. Technical batches may be present between mixture datasets and within mixture datasets. For example, a dataset may consist of two experiments performed at two different time points or by two different individuals. These differences may lead to technical variation in the gene expression data. If there are known batch effects in a mixture dataset and these cannot be removed using available methods [44-46], we recommend running CIBERSORTx on the datasets separately.

10. Batch effects between samples profiled by scRNA-seq may also occur, and these may affect the performance of the resulting signature matrix. There are available tools to remove batch effects in scRNA-seq datasets prior to construction of the signature matrix [47,31,48-50].

11. When comparing CD8 T cells to flow cytometry in our example, we pooled CD8 T cells with NKTs. We did this because (1) NKTs were not separately enumerated by flow cytometry, and (2) NKT cells express both *CD8A* and *CD3D* in the scRNA-seq dataset used to build the signature matrix.

12. CIBERSORTx Gene Expression Analysis Mode works best when the number of samples to deconvolve is much larger than the number of cell types in the signature matrix. A rule of thumb is to have at least four to five times as many mixture samples as cell types [23]. The "*Merged class file*" allows users to group cell types in the signature matrix into a broader set of phenotypic classes. This makes it easy to reduce the number of evaluable cell types in cases where the number of mixture samples is limiting.

13. As detailed in the online tutorial, users may also upload a file with known gene expression profiles as a ground truth file. If a ground truth file is uploaded and selected, CIBERSORTx will generate a variety of plots that can be used for quality control purposes. For example, this will allow users to check the concordance between imputed and ground truth GEPs.

14. For full transcriptome analyses, users are encouraged to request the CIBERSORTx executable ("*Menu*" → "*Download*").

15. When performing gene expression purification, the signature matrix should represent most of the cell types in a tissue. A signature matrix that does not include all major cell types is generally not recommended for expression imputation. However, when this is the case, users may apply batch correction for expression purification to down-weight the genes that are expressed on cell types not included in the signature matrix.

Acknowledgments

We would like to thank B. Chen, B. Nabet and M. Matusiak for assistance in beta-testing CIBERSORTx. This work was supported by grants from the 2019 AACR-AstraZeneca Lymphoma Research Fellowship (C.B.S., 19-40-12-STEE), the National Cancer Institute (A.M.N., R00CA187192; A.A.A., U01CA194389; A.A.A., R01CA188298), the Stinehart-Reed foundation (A.M.N., A.A.A.), the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) (A.M.N.), the Virginia and D.K. Ludwig Fund for Cancer Research (A.M.N., A.A.A.), the US Department of Defense (A.M.N., W81XWH-12-1-0498), anonymous donors (A.A.A., A.M.N.), the Shanahan and Bronzini Family Funds (A.A.A.), the V Foundation for Cancer Research (A.A.A.), the Leukemia and

Lymphoma Society (A.A.A.), and the Damon Runyon Cancer Research Foundation (A.A.A.), the American Society of Hematology (A.A.A.).

References

1. Maman S, Witz IP (2018) A history of exploring cancer in context. Nat Rev Cancer 18 (6):359-376. doi:10.1038/s41568-018-0006-7

2. Valkenburg KC, de Groot AE, Pienta KJ (2018) Targeting the tumour stroma to improve cancer therapy. Nat Rev Clin Oncol 15 (6):366-381. doi:10.1038/s41571-018-0007-1

3. Ribas A, Wolchok JD (2018) Cancer immunotherapy using checkpoint blockade. Science 359 (6382):1350-1355. doi:10.1126/science.aar4060

4. Fan HC, Fu GK, Fodor SPA (2015) Combinatorial labeling of single cells for gene expression cytometry. Science 347 (6222). doi:10.1126/science.1258367

5. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161 (5):1187-1201. doi:10.1016/j.cell.2015.04.044

6. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161 (5):1202-1214. doi:10.1016/j.cell.2015.05.002

7. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509 (7500):371-375. doi:10.1038/nature13173

8. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4 (7):e6098. doi:10.1371/journal.pone.0006098

9. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II, Wang W (2013) DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinformatics 29 (15):1865-1871

10. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, Waldner MJ, Bindea G, Mlecnik B, Galon J, Trajanoski Z (2015) Characterization of the

immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol 16:64. doi:10.1186/s13059-015-0620-6

11. Aran D, Hu Z, Butte AJ (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18 (1):220. doi:10.1186/s13059-017-1349-1

12. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One 6 (11):e27156. doi:10.1371/journal.pone.0027156

13. Kuhn A, Thu D, Waldvogel HJ, Faull RL, Luthi-Carter R (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Methods 8 (11):945-947. doi:10.1038/nmeth.1710

14. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, Signoretti S, Liu JS, Liu XS (2016) Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 17 (1):174. doi:10.1186/s13059-016-1028-7

15. Liebner DA, Huang K, Parvin JD (2014) MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics 30 (5):682-689. doi:10.1093/bioinformatics/btt566

16. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12 (5):453-457. doi:10.1038/nmeth.3337

17. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol 8 (12):e1002838. doi:10.1371/journal.pcbi.1002838

18. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q (2013) Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome medicine 5 (3):29

19. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ (2010) Cell type–specific gene expression differences in complex tissues. Nature methods 7 (4):287

20. Zhong Y, Wan Y-W, Pang K, Chow LM, Liu Z (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC bioinformatics 14 (1):89

21. Newman AM, Alizadeh AA (2016) High-throughput genomic profiling of tumor-infiltrating leukocytes. Curr Opin Immunol 41:77-84. doi:10.1016/j.coi.2016.06.006

22. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, Melton DA, Yanai I (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst 3 (4):346-360 e344. doi:10.1016/j.cels.2016.08.011

23. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. doi:10.1038/s41587-019-0114-2

24. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, Wang S, Rabasha B, Liu D, Zhang G, Margolais C, Ashenberg O, Ott PA, Buchbinder EI, Haq R, Hodi FS, Boland GM, Sullivan RJ, Frederick DT, Miao B, Moll T, Flaherty KT, Herlyn M, Jenkins RW, Thummalapalli R, Kowalczyk MS, Canadas I, Schilling B, Cartwright ANR, Luoma AM, Malu S, Hwu P, Bernatchez C, Forget MA, Barbie DA, Shalek AK, Tirosh I, Sorger PK, Wucherpfennig K, Van Allen EM, Schadendorf D, Johnson BE, Rotem A, Rozenblatt-Rosen O, Garraway LA, Yoon CH, Izar B, Regev A (2018) A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell 175 (4):984-997 e924. doi:10.1016/j.cell.2018.09.006

25. Li H, van der Leun AM, Yofe I, Lubling Y, Gelbard-Solodkin D, van Akkooi ACJ, van den Braber M, Rozeman EA, Haanen J, Blank CU, Horlings HM, David E, Baran Y, Bercovich A, Lifshitz A, Schumacher TN, Tanay A, Amit I (2019) Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell 176 (4):775-789 e718. doi:10.1016/j.cell.2018.11.043

26. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CG, Kazer SW, Gaillard A, Kolb KE,

Villani AC, Johannessen CM, Andreev AY, Van Allen EM, Bertagnolli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jane-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352 (6282):189-196. doi:10.1126/science.aad0501

27. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, Mylvaganam R, Rozenblatt-Rosen O, Rocco JW, Faquin WC, Lin DT, Regev A, Bernstein BE (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell 171 (7):1611-1624 e1624. doi:10.1016/j.cell.2017.10.044

28. Andor N, Simonds EF, Czerwinski DK, Chen J, Grimes SM, Wood-Bouwens C, Zheng GXY, Kubit MA, Greer S, Weiss WA, Levy R, Ji HP (2019) Single-cell RNA-Seq of follicular lymphoma reveals malignant B-cell types and coexpression of T-cell immune checkpoints. Blood 133 (10):1119-1129. doi:10.1182/blood-2018-08-862292

29. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW (2018) Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 9 (1):3588. doi:10.1038/s41467-018-06052-0

30. Giraddi RR, Chung CY, Heinz RE, Balcioglu O, Novotny M, Trejo CL, Dravis C, Hagos BM, Mehrabad EM, Rodewald LW, Hwang JY, Fan C, Lasken R, Varley KE, Perou CM, Wahl GM, Spike BT (2018) Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. Cell Rep 24 (6):1653-1666 e1657. doi:10.1016/j.celrep.2018.07.025

31. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36 (5):411-420. doi:10.1038/nbt.4096

32. Ji Z, Ji H (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res 44 (13):e117. doi:10.1093/nar/gkw430

33. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M (2017) SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 14 (5):483-486. doi:10.1038/nmeth.4236

34. Lin P, Troup M, Ho JW (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 18 (1):59. doi:10.1186/s13059-017-1188-0

35. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14 (10):979-982. doi:10.1038/nmeth.4402

36. Senabouth A, Lukowski SW, Alquicira Hernandez J, Andersen S, Mei X, Nguyen QH, Powell JE (2017) ascend: R package for analysis of single cell RNA-seq data. bioRxiv:207704. doi:10.1101/207704

37. Zurauskiene J, Yau C (2016) pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics 17:140. doi:10.1186/s12859-016-0984-y

38. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049. doi:10.1038/ncomms14049

39. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, Vose J, Bast M, Fu K, Weisenburger DD, Greiner TC, Armitage JO, Kyle A, May L, Gascoyne RD, Connors JM, Troen G, Holte H, Kvaloy S, Dierickx D, Verhoef G, Delabie J, Smeland EB, Jares P, Martinez A, Lopez-Guillermo A, Montserrat E, Campo E, Braziel RM, Miller TP, Rimsza LM, Cook JR, Pohlman B, Sweetenham J, Tubbs RR, Fisher RI, Hartmann E, Rosenwald A, Ott G, Muller-Hermelink HK, Wrench D, Lister TA, Jaffe ES, Wilson WH, Chan WC, Staudt LM, Lymphoma/Leukemia Molecular Profiling P (2008) Stromal gene signatures in large-B-cell lymphomas. N Engl J Med 359 (22):2313-2323. doi:10.1056/NEJMoa0802885

40. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403 (6769):503-511. doi:10.1038/35000501

41. Diaz-Mejia J, Meng E, Pico A, MacParland S, Ketela T, Pugh T, Bader G, Morris J (2019) Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data [version 1; peer review: 3 approved with reservations]. F1000Research 8 (296). doi:10.12688/f1000research.18490.1

42. Zhong Y, Liu Z (2011) Gene expression deconvolution in linear space. Nat Methods 9 (1):8-9; author reply 9. doi:10.1038/nmeth.1830

43. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, Conley SD, Chaib H, Red-Horse K, Longaker MT, Snyder MP, Krasnow MA, Weissman IL (2017) Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv:125724. doi:10.1101/125724

44. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31 (4):265-273. doi:10.1016/S1046-2023(03)00155-5

45. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (1):118-127. doi:10.1093/biostatistics/kxj037

46. Giordan M (2014) A Two-Stage Procedure for the Removal of Batch Effects in Microarray Studies. Stat Biosci 6 (1):73-84. doi:10.1007/s12561-013-9081-1

47. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C (2017) SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods 14 (6):584-586. doi:10.1038/nmeth.4263

48. Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36 (5):421-427. doi:10.1038/nbt.4091

49. Hie B, Bryson B, Berger B (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. doi:10.1038/s41587-019-0113-3

50. Johnson M, Purdom E (2017) Clustering of mRNA-Seq data based on alternative splicing patterns. Biostatistics 18 (2):295-307. doi:10.1093/biostatistics/kxw044

51. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, Schoeberl B, Raue A (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun 8 (1):2032. doi:10.1038/s41467-017-02289-3

52. Wang Z, Cao S, Morris JS, Ahn J, Liu R, Tyekucheva S, Gao F, Li B, Lu W, Tang X, Wistuba, II, Bowden M, Mucci L, Loda M, Parmigiani G, Holmes CC, Wang W (2018) Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. iScience 9:451-460. doi:10.1016/j.isci.2018.10.028

53. Wang X, Park J, Susztak K, Zhang NR, Li M (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun 10 (1):380. doi:10.1038/s41467-018-08023-x

54. Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, Yankovitz G, Mandelboim M, Iraqi FA, Amit I, Mayo L, Bacharach E, Gat-Viks I (2019) Cell composition analysis of bulk genomics using single-cell data. Nat Methods 16 (4):327-332. doi:10.1038/s41592-019-0355-5

Tables

Table 1. Overview of selected bioinformatics tools for studying cellular composition in tissue samples. The list is restricted to digital cytometry tools that either leverage scRNA-seq reference profiles for deconvolution, handle technical variation between datasets by batch correction, or perform gene expression imputation. The tool/method column refers to the name of the tool presented in each publication or the underlying methodology when no other name is available. Abbreviations: DSA Digital Sorting Algorithm, MMAD: Microarray Microdissection with Analysis of Differences, TIMER: Tumor IMmune Estimation Resource; MuSiC: MUlti-Subject SIngle Cell deconvolution; CPM: Cellular Population Mapping.

Reference	Tool/Method	scRNA-seq reference	Batch correction	Cell type-specific gene expression
Shen-Orr et al. [19]	csSAM	No	No	Yes
Zhong et al. [20]	DSA	No	No	Yes
Liebner et al. [15]	MMAD	No	No	Yes
Quon et al. [18]	ISOpure	No	No	Yes, 2 component sample-level cell type-specific gene expression
Li et al. [14]	TIMER	No	Yes	No
Baron et al. [22]	BSEQ-sc	Yes	No	Yes, cell type-specific differential expression
Schelker et al. [51]	CIBERSORT	Yes	No	No
Wang et al. [52]	DeMixT	No	No	Yes, 3 component sample-level cell type-specific gene expression
Wang et al. [53]	MuSiC	Yes	No	No
Frishberg et al. [54]	CPM	Yes	No	No
Newman et al. [23]	CIBERSORTx	Yes	Yes	Yes, *n* component sample-level cell type-specific gene expression

Table 2. Pairs of signature matrices and mixture datasets where CIBERSORTx batch correction is strongly recommended.

Signature matrix	Mixture dataset	Batch correction method
Microarray of sorted cell populations	Bulk RNA-seq of fresh/frozen tissues	B-mode
	NanoString nCounter of FFPE tissues	
RNA-seq of sorted cell populations	Bulk RNA-seq of FFPE tissues	B-mode
	Microarray of fresh/frozen tissues	
Plate-based scRNA-seq without UMIs (i.e., SMART-Seq2)	Bulk RNA-seq of fresh/frozen tissues	B-mode
	Microarray of fresh/frozen tissues	
	Bulk RNA-seq of FFPE tissues	
	NanoString nCounter of FFPE tissues	
Droplet-based scRNA-seq (i.e., 10x Chromium, Drop-seq)	Bulk RNA-seq of fresh/frozen tissues	S-mode
	Microarray of fresh/frozen tissues	
	Bulk RNA-seq of FFPE tissues	
	NanoString nCounter of FFPE tissues	

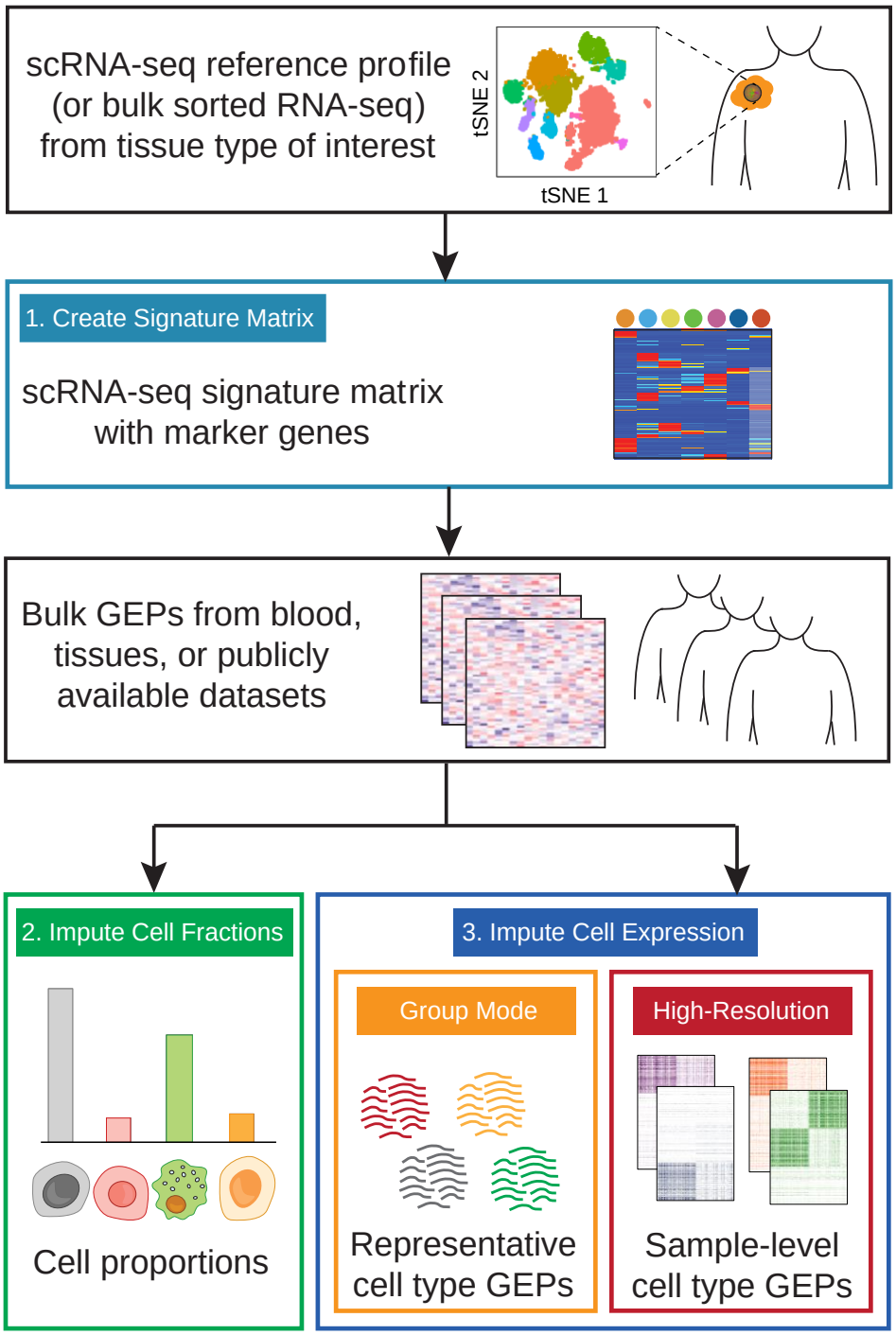

Figure 1: Overview of CIBERSORTx. Starting from reference profiles generated by scRNA-seq, bulk sorted RNA-seq, or microarrays, CIBERSORTx generates a deconvolution signature matrix, consisting of cell type-specific barcode genes (step 1), which is then repeatedly used to enumerate cell fractions (step 2) or impute cell-type-specific gene expression profiles (step 3) from bulk tissue GEPs. Gene expression imputation can be performed with group-mode, which results in a representative transcriptome profile for each cell type in the signature matrix, or high-resolution mode, which yields sample-level expression estimates for each cell type.

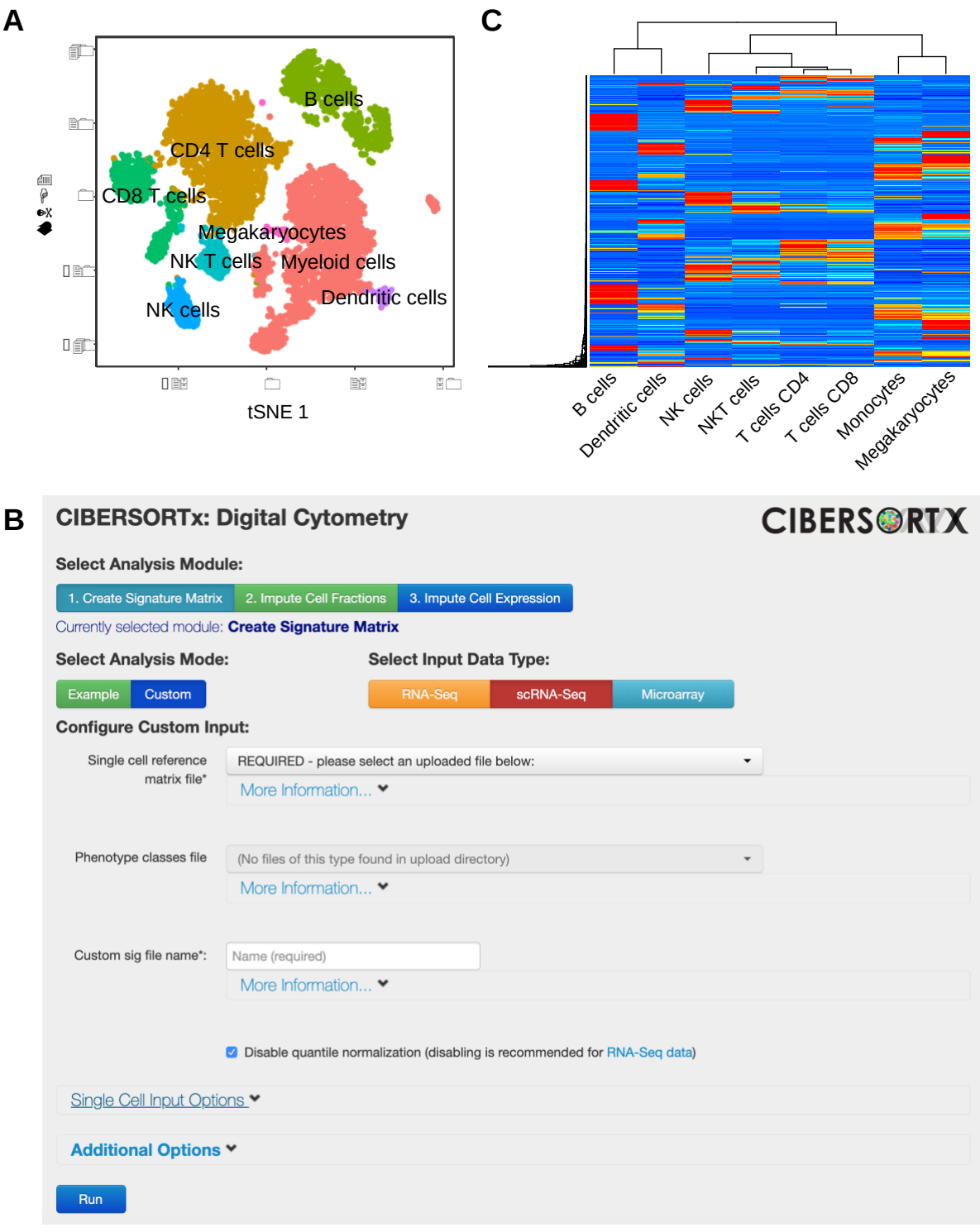

Figure 2: Building a signature matrix with scRNA-seq data. (A) t-SNE projection of PBMCs profiled by scRNA-seq (10x Chromium v2 platform with 5' chemistry). (B) Screen shot of the CIBERSORTx website showing the selection of scRNA-seq data as the input data type. (C) Heat map of the signature matrix generated by CIBERSORTx when applied to the dataset shown in panel A.

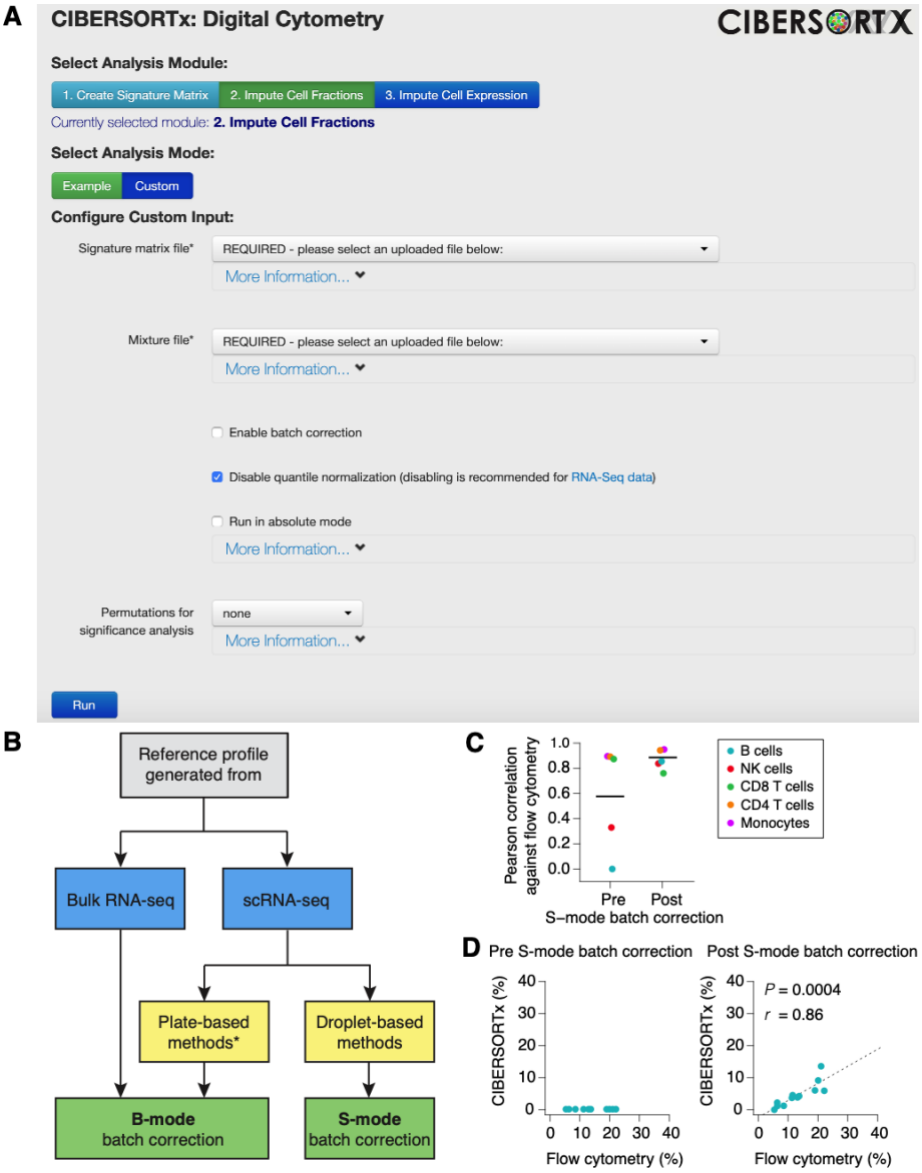

Figure 3: Imputation of cell fractions and cross-platform deconvolution. (A) Screenshot of the CIBERSORTx website showing the selection of "*2. Impute Cell Fractions*" module. (B) Decision tree to help CIBERSORTx users identify the batch correction mode that is best suited for their application: (1) single-cell reference mode ('S-mode') or (2) bulk reference mode ('B-mode'). *Plate-based methods refer to methods that show little or moderate technical variation compared to bulk RNA-seq, such as SMART-seq2. (C) Estimation of cell fractions in RNA-seq GEPs from 12 healthy whole blood samples using the PBMC signature matrix (from Figure 2) pre and post S-mode batch correction. The deconvolution performance is shown as Pearson correlations comparing the CIBERSORTx estimated fractions with the flow cytometry ground truth for five cell types (B cells, NK cells, CD8 T cells, CD4 T cells and monocytes). Medians are shown as horizontal lines. (D) Same as C but showing imputed proportions of B cells before (left) and after (right) S-mode batch correction. *Right:* Concordance between CIBERSORTx and flow cytometry was determined by Pearson correlation (*r*) and linear regression (dashed line).

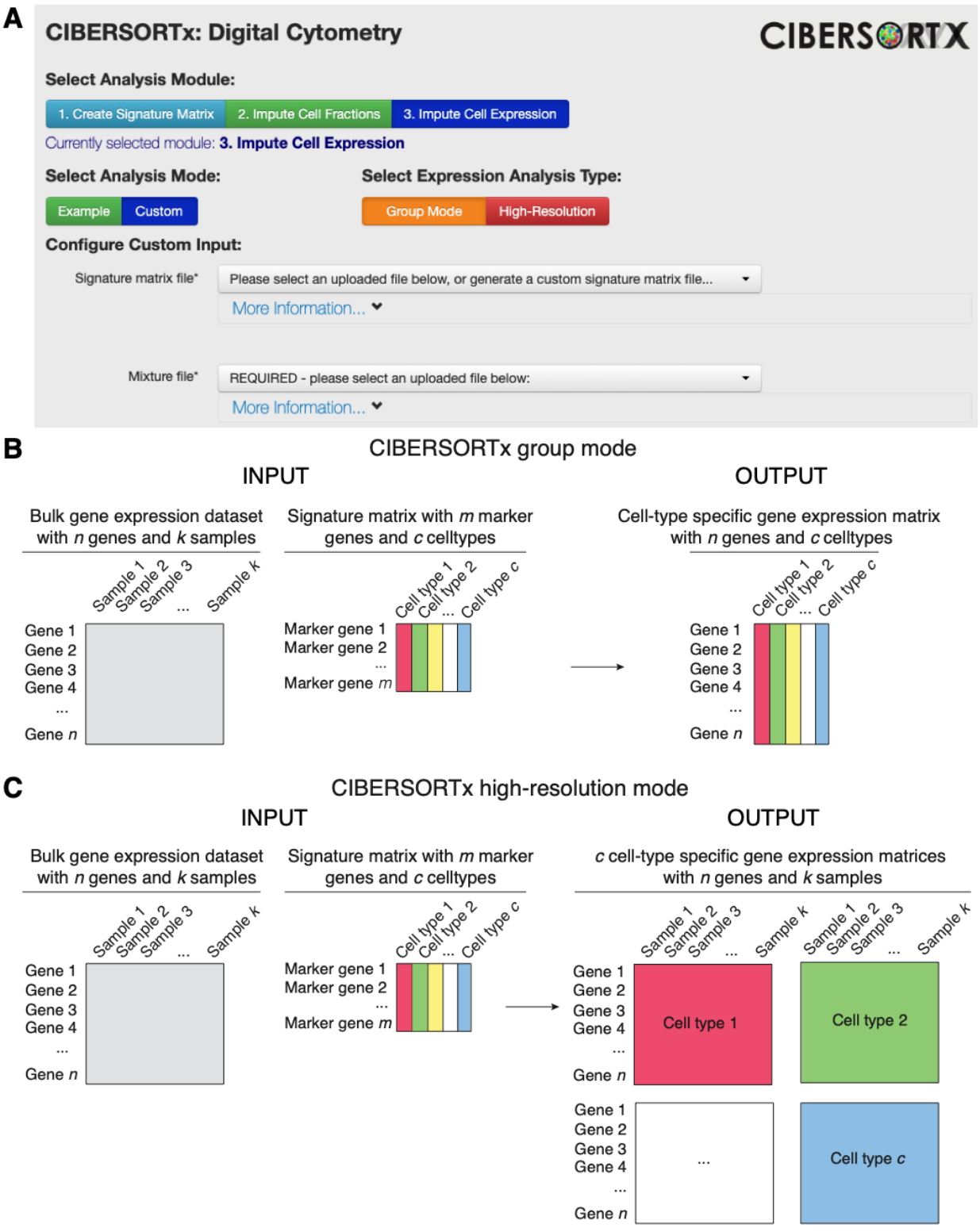

Figure 4: Cell type-specific gene expression imputation. (A) Screenshot of the CIBERSORTx website showing the selection of the "*3. Impute Cell Expression*" module. (B, C) Schematic overview of group mode (B) and high-resolution mode (C) expression imputation.