Profiling Cellular Ecosystems at Single-Cell Resolution and at Scale with EcoTyper

Chloé B. Steen[1,2,3], Bogdan A. Luca[1,4], Ash A. Alizadeh[3,5,6,7], Andrew J. Gentles[1,4,7], and Aaron M. Newman[1,5,7]

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA; [2]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway; [3]Division of Oncology, Department of Medicine, Stanford University, Stanford, CA, USA; [4]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA, USA; [5]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA; [6]Division of Hematology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, CA, USA; [7]Stanford Cancer Institute, Stanford University, Stanford, CA 94305, USA.

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

Summary

Tissues are comprised of diverse cell types and cellular states that organize into distinct ecosystems with specialized functions. EcoTyper is a collection of machine learning tools for large-scale delineation of cellular ecosystems and their constituent cell states from bulk, single-cell, and spatially resolved gene expression data. In this chapter, we provide a primer on EcoTyper and demonstrate its use for the discovery and recovery of cell states and ecosystems from healthy and diseased tissue specimens.

1 Introduction

Recent advances in multiplexed imaging, single-cell sequencing, and computational deconvolution have revealed new insights into the cellular diversity and modularity of complex tissues [1,2]. For example, it is now well-established that individual cells can exhibit context-dependent phenotypic states that organize into functionally distinct cellular ecosystems [3]. Such ecosystems, commonly referred to as communities or ecotypes [4-6], are thought to play critical roles during tissue development, homeostasis, repair, and disease. In cancer, immune and stromal-rich ecotypes have been linked to heterogeneity in patient outcomes, including differential response to therapy [4,6]. Understanding how ecotypes shape clinical trajectories in cancer and other diseases has significant potential for the discovery of more effective diagnostics and therapeutic targets.

EcoTyper is a suite of machine learning tools for characterizing transcriptionally defined cell states and their co-association patterns (i.e., ecotypes) from complex tissue specimens [4,6]. It performs two key tasks: (i) the discovery of cell states and ecotypes from bulk or single-cell expression data, and (ii) the recovery of previously defined cell states and ecotypes from bulk, single-cell, or spatially resolved expression data (**Figure 1**).

We recently applied EcoTyper to illuminate the cell state and ecotype landscape of human carcinoma – the most common human malignancy – and diffuse large B cell lymphoma, the most prevalent blood cancer [4,6]. A companion website, available at https://ecotyper.stanford.edu/, includes the following functions and features:

1. Recovery of carcinoma and lymphoma-specific cell states and ecotypes in user-provided expression datasets.
2. Interactive exploration and visualization of gene expression signatures, survival associations, and predicted ligand-receptor pairs for carcinoma and lymphoma-specific cell states and ecotypes.
3. Raw data associated with both publications.

Here we demonstrate how to run EcoTyper on a personal computer, server, or high-performance computing cluster. By illustrating key functions and commands using hands-on examples, we show how EcoTyper can be used for the discovery (**Sections 3.1** and **3.2**) and recovery (**Sections 3.3 to 3.5**) of cell states and ecotypes from bulk and single-cell expression data. For those primarily interested in the recovery of predefined carcinoma or lymphoma-specific cell states/ecotypes, see **Sections 3.3** to **3.5** (**Figure 1A**).

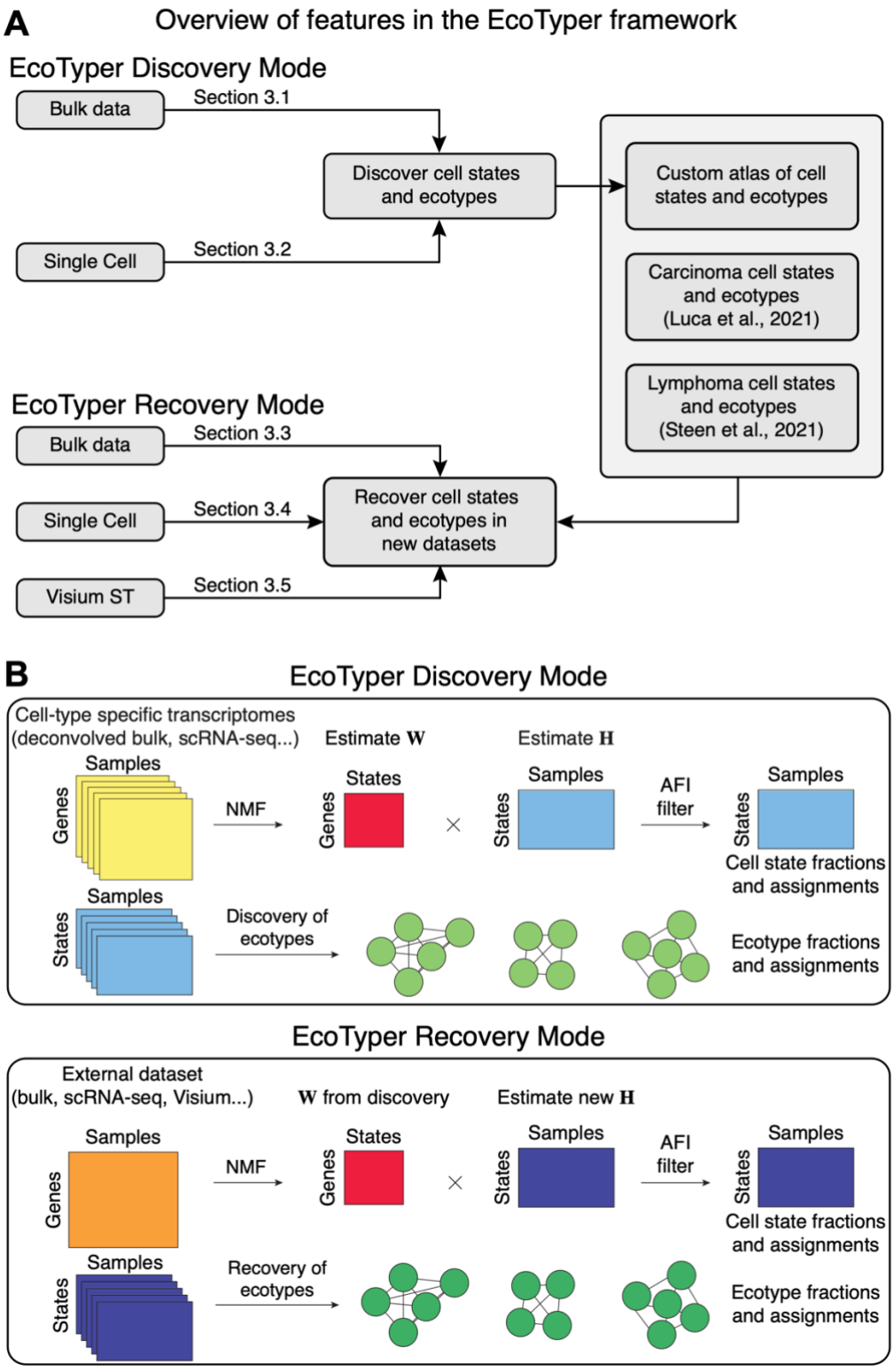


Figure 1. The EcoTyper Framework. (A) Key EcoTyper functions and their corresponding sections in this chapter. ST, spatial transcriptomics. (B) Schematic depiction of the EcoTyper analytical pipeline, including the discovery and recovery of cell states and ecotypes from diverse transcriptomic platforms. For further information, see Luca et al. (2021).

2 Materials

EcoTyper is written in R. The latest source code can be found at https://github.com/digitalcytometry/ecotyper or https://ecotyper.stanford.edu/. All instructions necessary to install EcoTyper and its dependencies are available from GitHub (see also **Notes 1, 2 and 3**).

3 Methods

3.1 Discovery of Cell States and Ecotypes from Bulk Data

In this section, we describe how to apply EcoTyper to the discovery of cell states and ecotypes from bulk tissue expression data profiled by RNA-seq or microarrays (**Figure 1**). An example bulk RNA-seq discovery dataset consisting of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) tumor samples from The Cancer Genome Atlas (TCGA) [7] is used to illustrate each step of the workflow. Once EcoTyper has been downloaded and installed (**Section 2**), the example dataset can be found in *example_data/bulk_lung_data.txt* together with a sample annotation file *example_data/bulk_lung_annotation.txt*.

This section involves the following key steps:

1. *In silico* purification: Imputation of cell type-specific gene expression profiles from a discovery dataset of bulk tissue transcriptomes using CIBERSORTx [8] (See **Notes 2** and **4**).
2. Cell state discovery: Identification and quantitation of cell type-specific states using the output of step 1.
3. Ecotype discovery: Co-assignment of cell states (from step 2) into ecotypes.

The above steps are described in **Sections 3.1.1** through **3.1.6** below (see also **Figure 2A**). The remainder of **Section 3.1** reviews the relevant commands, input files, and configuration settings.

3.1.1 Cell Type Fraction Estimation

To identify cell states within a discovery expression dataset, EcoTyper starts with CIBERSORTx (see **Note 2**), a machine learning method for digital cytometry that can determine cell-type-specific fractions and expression profiles from bulk tissue transcriptomes [8]. CIBERSORTx is applied in two sequential phases. In the first phase, cell type fractions are determined using the *CIBERSORTx Fractions* module [8]. This step requires a signature matrix comprised of reference profiles that discriminate major cell types within a tissue type of interest. For example, a signature matrix for a solid

malignancy would ideally include key immune cell lineages (e.g., B cells, CD8 T cells, CD4 T cells, NK cells, monocytes/macrophages, dendritic cells, etc.), stromal subsets (fibroblasts and endothelial cells), and cancer cells. Such matrices are straightforward to derive from purified cell type-specific expression profiles or scRNA-seq data, as described elsewhere [8,9]. Instructions and vignettes for creating signature matrices are also available from the CIBERSORTx website (https://cibersortx.stanford.edu/).

Table 1. Built-in Signature Matrices and Cell Type Mapping Schemes for Carcinoma EcoTyper and Lymphoma EcoTyper. For details about the derivation and application of both signature matrices, see Newman et al. (2019) [8], Luca et al. (2021) [4], and Steen et al. (2021) [6].

Signature Matrix	Cell Type	Carcinoma EcoTyper	Lymphoma EcoTyper
LM22	B cells naive	B cells	B cells
	B cells memory	B cells	B cells
	Plasma cells	PCs	PCs
	T cells CD8	T cells CD8	T cells CD8
	T cells CD4 naive	T cells CD4	T cells CD4
	T cells CD4 memory resting	T cells CD4	T cells CD4
	T cells CD4 memory activated	T cells CD4	T cells CD4
	T cells follicular helper	T cells CD4	T cells follicular helper
	T cells regulatory (Tregs)	T cells CD4	Tregs
	T cells gamma delta	T cells CD8	T cells CD8
	NK cells resting	NK cells	NK cells
	NK cells activated	NK cells	NK cells
	Monocytes	Monocytes and Macrophages	Monocytes and Macrophages
	Macrophages M0	Monocytes and Macrophages	Monocytes and Macrophages
	Macrophages M1	Monocytes and Macrophages	Monocytes and Macrophages
	Macrophages M2	Monocytes and Macrophages	Monocytes and Macrophages
	Dendritic cells resting	Dendritic cells	Dendritic cells
	Dendritic cells activated	Dendritic cells	Dendritic cells
	Mast cells resting	Mast cells	Mast cells
	Mast cells activated	Mast cells	Mast cells
	Neutrophils	Neutrophils	Neutrophils
	Eosinophils	Excluded	Excluded
TR4	Immune cells	Immune cells	Immune cells
	Endothelial cells	Endothelial cells	Endothelial cells
	Epithelial cells	Epithelial cells	Excluded
	Fibroblasts	Fibroblasts	Fibroblasts

In recent publications demonstrating the utility of EcoTyper [4,6], we applied two well-validated signature matrices (LM22 and TR4, **Table 1**) covering the major immune, stromal, and cancer cell populations found in human carcinomas and lymphomas [8]. Both signature matrices are packaged with EcoTyper and can be applied as covered in **Section 3.1.8**, *Cell type fractions* (see also **Table 1**). To interrogate other cell types, users can apply CIBERSORTx (or an alternative deconvolution method with comparable functionality) to design their own signature matrix and then obtain cell type fractions prior to running EcoTyper.

3.1.2 Cell Type Expression Purification

In the second phase, once cell type fractions are determined, CIBERSORTx is applied to impute cell type-specific gene expression profiles from the discovery dataset. For this purpose, EcoTyper employs the *CIBERSORTx Impute High-Resolution Cell Expression* module ("CIBERSORTx HiRes").

As input, CIBERSORTx HiRes requires bulk expression data and cell type fractions obtained from the previous step (*step 1, Cell type fraction estimation* in **Figure 2A**). The output consists of a purified gene expression matrix for each cell type, with genes as rows and samples as columns (see **Note 5**). These cell-type specific gene expression matrices serve as input for cell state identification.

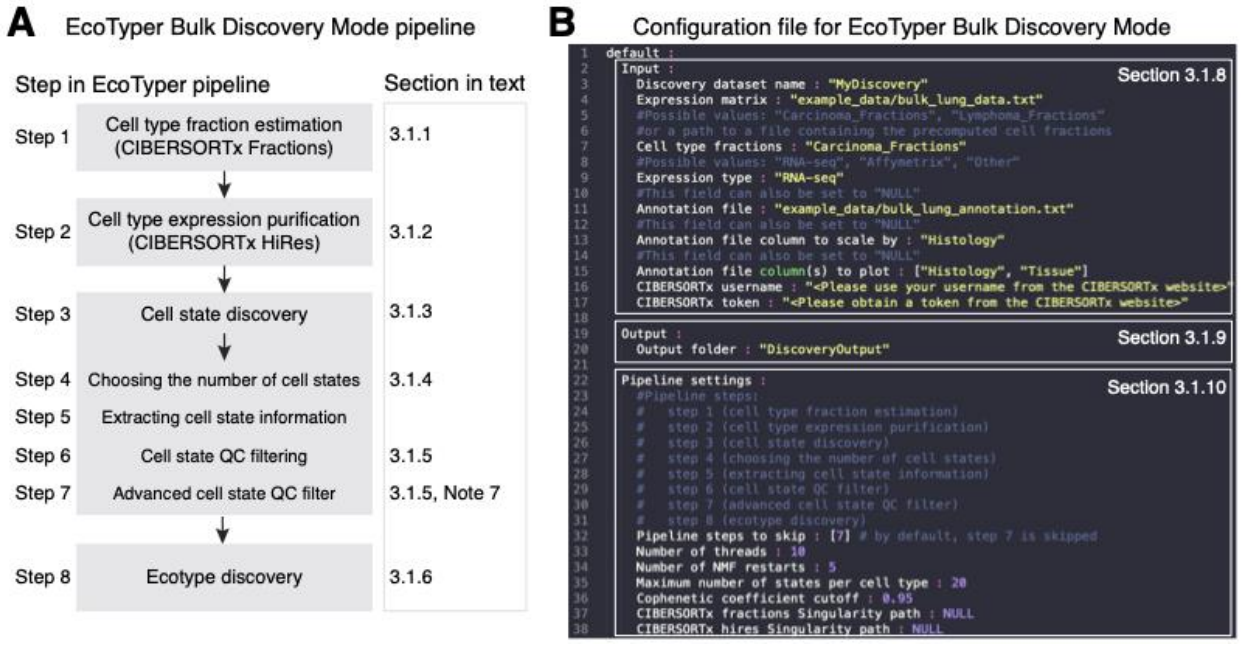


Figure 2. Overview and Configuration of Bulk Discovery Mode. (A) Typical EcoTyper workflow for the discovery of cell states and ecotypes from bulk gene expression data. Key steps are indicated along with their corresponding sections in this chapter. (B) Example configuration file for bulk discovery mode.

3.1.3 Cell State Discovery

Given a collection of cell-type-specific gene expression profiles (GEPs), EcoTyper performs cell state identification and quantification (**Figure 1B**). Cell states are identified through simultaneous deconvolution and clustering of cell-type-specific GEPs via non-negative matrix factorization (NMF). NMF is combined with specialized preprocessing and postprocessing steps to automatically select the number of states per cell type while minimizing false positives using a novel adaptive filtering approach (**Note 6**; **Figure 1B**). Importantly, NMF defines cell states and determines their relative fractional abundance in each sample, allowing for downstream analyses of cell state composition, including ecotype detection (e.g., **Sections 3.1.6** and **3.1.13**).

3.1.4 Determining the Number of Cell States

For each set of cell-type-specific GEPs, EcoTyper applies NMF across a range of ranks (2–20, by default) to determine the number of transcriptionally-defined cell states. To select the most appropriate rank (i.e., the number of cell states), we developed a heuristic based on the cophenetic coefficient (CC) (*step 4, Choosing the number of cell states* in **Figure 2A**). The rank at which the CC starts decreasing is typically selected manually, however this rank may be suboptimal when the CC exhibits a multi-modal shape [4]. Therefore, the number of cell states is automatically determined based on the first occurrence for which the CC drops below 0.95 (by default) and having been above this level for at least two consecutive ranks. As this threshold is predetermined and may not be ideal in every case, we recommend that users visually inspect the plots created by EcoTyper (see **Section 3.1.12**) and revise the threshold accordingly, with the goal of selecting a CC threshold that captures the first substantial drop in CC across the majority of cell types.

3.1.5 Cell State Quality Control

EcoTyper applies gene-level standardization (i.e., normalizing the mean of each gene to zero and standard deviation to one) to improve the sensitivity of cell state discovery with NMF. However, as NMF requires a non-negative input, EcoTyper applies a *posneg* transformation to enforce the non-negativity requirement of NMF [4]. Since the posneg transformation can lead to the identification of spurious cell states driven by features with more negative values than positive ones, we devised an adaptive false positive index (AFI) (**Figure 1B**) to eliminate likely false positives while maximizing sensitivity (Figure S1 in Luca et al. 2021 [4]). EcoTyper automatically filters states with AFI ≥ 1, and this is done in *step 6, Cell state QC filtering* of the EcoTyper pipeline (**Figure 2A**). An additional filtering step is also performed in *step 7, Advanced cell state QC filter* (**Figure 2A**), with further details provided in **Note 7**.

3.1.6 Ecotype (Cellular Community) Discovery

Following quality control, EcoTyper analyzes all remaining cell states to identify statistically robust patterns of co-occurrence across samples in the discovery dataset. These cellular communities constitute ecosystem subtypes or “ecotypes” (**Figure 1B**). To identify ecotypes, EcoTyper calculates the degree of overlap between each pair of cell states across samples using the Jaccard index combined with statistical filtering. Ecotypes are subsequently identified by hierarchical clustering and silhouette width maximization. Ecotypes with less than 3 cell states are eliminated from further analysis by default.

3.1.7 Configuring the EcoTyper Run

The script for cell type and ecotype discovery from bulk data is *EcoTyper_discovery_bulk.R*.

This script takes as input a configuration file in YAML format, which has three parts: (i) Input, (ii) Output, and (iii) Pipeline settings (**Figure 2B**). In the following section, we describe the contents of this file and provide instructions to select the most appropriate settings for a given application. The configuration file used for the example discovery dataset is *config_discovery_bulk.yml*.

3.1.8 Input Section of Configuration File

The *Input* section of the configuration file is used for setting parameters tailored to the input data. Each field is described in detail below.

Discovery dataset name

The *Discovery dataset name* field is the identifier used by EcoTyper to internally save and retrieve information about the cell states/ecotypes defined in the discovery dataset. Alphanumeric characters and ‘_’ are accepted for this field. For the example dataset used in this section, this field is set to “MyDiscovery” (**Figure 2B**); otherwise, it should be customized by the user.

Expression matrix

The *Expression matrix* field refers to the path of a tab-delimited file where the discovery dataset is located. The expression matrix must be formatted with genes as rows and samples as columns and should be normalized to transcripts-per-million (TPM) for bulk RNA-seq and non-logarithmic (exponential) space for microarrays. It should have gene symbols as the first column and gene expression for each sample in the remaining

columns. Column (sample) names should be unique (see also **Note 8**). The example dataset used in this section is *example_data/bulk_lung_data.txt*.

Cell type fractions

The *Cell type fractions* field refers to the source data for cell type fractions in the discovery dataset. If the major cell types expected in a given discovery dataset are the same as those analyzed with Carcinoma EcoTyper [4] (**Table 1**), this field should be set to *Carcinoma_Fractions* (case sensitive) and EcoTyper will invoke CIBERSORTx to automatically estimate fractions for these populations in *step 1* of the workflow (**Figure 2A, Section 3.1.1**). Similarly, if the cell types analyzed with Lymphoma EcoTyper [6] are appropriate (**Table 1**), the user can set this field to *Lymphoma_Fractions* and CIBERSORTx will be automatically applied to these cell types.

If neither of these cases apply, the user must provide a path to a tab-delimited file containing the fractions for the major cell types in their discovery dataset. This file can be obtained using CIBERSORTx, another deconvolution tool, or an independent source of cell type fractions, outside of the EcoTyper pipeline. The first column should contain sample names that match the column names in the discovery dataset. The remaining columns should contain the fractions for each cell type, with no missing values. These fractions should sum to 1 for each row. For an example of a cell type fraction file, see *example_data/bulk_fractions_example.txt*.

Expression type

The *Expression type* field is only relevant if cell type fractions are estimated automatically by EcoTyper (i.e., *Carcinoma_Fractions* or *Lymphoma_Fractions* are provided in the *Cell type fractions* field of the configuration file, as described above). *Expression type* specifies the platform used to generate the data provided in the expression matrix. The accepted values are “RNA-seq” for bulk RNA-seq data, “Affymetrix” for data profiled using Affymetrix microarray platforms, and “Other” for data from non-Affymetrix microarray platforms. Once set, EcoTyper determines the appropriate parameters for the *CIBERSORTx Fractions* module.

Annotation file (optional)

A path to an annotation file may be provided in the *Annotation file* field. This file must contain a first column called ID with the same names as the columns of the expression matrix. Additional columns can be used for defining sample batches (see subsection *Annotation file column to scale by* below) and for plotting color bars in the heat map output (see *Annotation file column(s) to plot* below). If not provided, this field must be

set to “NULL”. The example annotation file for the lung cancer discovery dataset is *example_data/bulk_lung_annotation.txt*.

Annotation file column to scale by (optional)

To discover cell states and ecotypes across 16 types of human carcinoma while minimizing tumor-type-specific variation, we standardized genes to a mean of zero and standard deviation of one within each type of carcinoma [4]. The field *Annotation file column to scale by* optionally allows users to specify a column name in the annotation file by which the samples will be grouped when performing standardization. The example discovery dataset has samples from LUAD and LUSC, as specified in the “Histology” column of the annotation file. To identify cell states preferentially shared by LUAD and LUSC, we will use the “Histology” column to perform standardization on the example dataset.

Of note, independently standardizing different sample types in the discovery dataset is a consideration that depends on the purpose of the analysis. For example, if users are instead interested in prioritizing cell states and ecotypes that are specific to each tumor type, this argument can be set to “NULL”. In this case, standardization will be applied once across all samples in the discovery cohort (as was done for Lymphoma EcoTyper [6], in which a single cancer type was analyzed). The same will happen if the annotation file is not supplied.

Annotation file column(s) to plot

The *Annotation file column(s) to plot* field specifies which columns in the annotation file to use for custom color bar(s) in the heat map outputs.

CIBERSORTx username and token

The *CIBERSORTx username* and *CIBERSORTx token* fields should respectively contain the CIBERSORTx username and token necessary to run the CIBERSORTx executable. Once registered at https://cibersortx.stanford.edu, the token can be obtained by request from *Menu* > *Download*. For installation information, see **Note 2**.

3.1.9 Output Section of Configuration File

The *Output* section contains the *Output folder* field, which specifies the path where the final output will be saved. This folder will be created if it does not exist. For the lung cancer discovery dataset, the output folder is set to “DiscoveryOutput”.

3.1.10 Pipeline Settings

This section determines how the EcoTyper bulk discovery workflow will be run.

Pipeline steps to skip

The *Pipeline steps to skip* option allows the user to omit key steps of the workflow (**Figure 2A**). Please note that this option is only intended when the pipeline has already been run once and parameter adjustments are being made. For example, if the cophenetic coefficient threshold used in *step 4* (**Figure 2A**) needs to be modified, the user should skip *steps 1-3* and re-run from *step 4*.

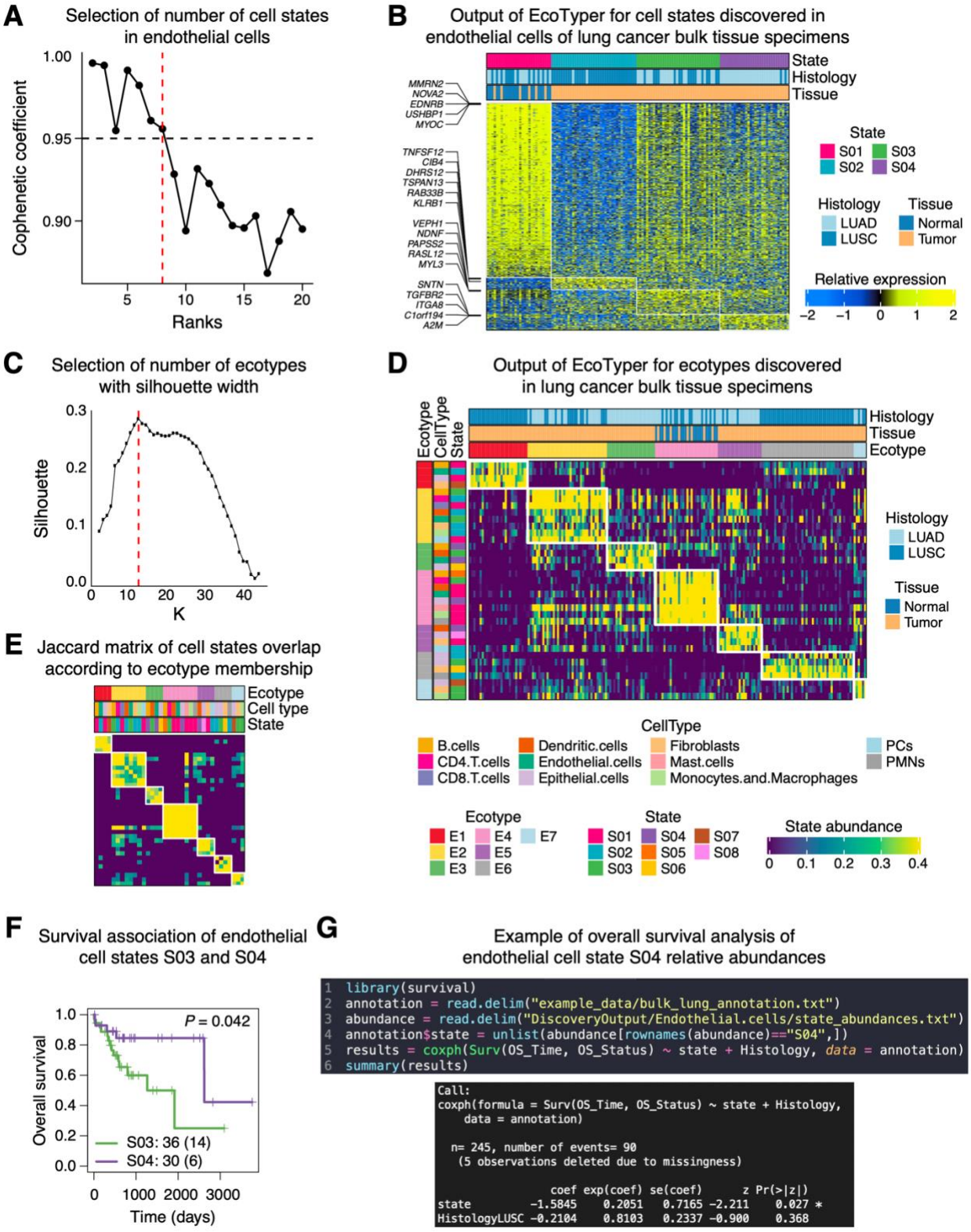
Number of threads

This fields specifies the number of threads (i.e., CPU cores) allocated to EcoTyper.

Number of NMF restarts

The NMF approach used by EcoTyper can give slightly different results each time owing to stochastic initialization. To obtain a stable solution, EcoTyper runs NMF multiple times with different seeds and the solution that best explains the discovery data is chosen. The parameter *Number of NMF restarts* specifies the number of NMF restarts for each rank and cell type. Since this is a very time-consuming process, we recommend 5 restarts for exploratory purposes and at least 50 restarts for publication-quality results.

Figure 3. Bulk Discovery Mode Applied to Non-Small Cell Lung Cancer. (A) Cophenetic coefficient plot used for the selection of cell state numbers, shown here for lung cancer epithelial cells deconvolved from the bulk discovery dataset. The black dashed line indicates the cophenetic coefficient threshold set by default, and the red dashed line denotes the corresponding matrix rank (i.e., number of cell states) selected. For further details, refer to **Section 3.1.4**. (B) Heat map showing endothelial cell states identified by EcoTyper from the bulk discovery dataset. Selected marker genes are indicated on the left side of the heat map. Color bars above the heat map denote the dominant cell state to which each sample is assigned and the corresponding tissue type. LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; Normal, adjacent normal; Tumor, LUAD or LUSC. (C) Silhouette plot showing the optimal number of ecotypes (red line) identified in the bulk discovery dataset. (D) Heat map of inferred ecotype abundances in the bulk discovery dataset. Rows are cell states (n = 35), columns are samples (n = 190), and both axes are ordered by cell state co-association patterns (i.e., ecotypes). The color bars above the heat map denote the ecotype label and tissue (as in panel B). (E) Jaccard matrix of cell state overlap patterns according to ecotype membership. (F) Kaplan-Meier plot showing differences in overall survival for lung cancer patients in the bulk discovery dataset assigned to endothelial cell states S03 (n = 36) and S04 (n = 30). Statistical significance was calculated by a two-sided log-rank test. (G) Example R script for analyzing the association between the abundance of endothelial state S04 and overall survival in the bulk discovery dataset. Cox proportional hazards regression was used to calculate statistical significance and adjust for histological subtype.




```r
library(survival)
annotation = read.delim("example_data/bulk_lung_annotation.txt")
abundance = read.delim("DiscoveryOutput/Endothelial.cells/state_abundances.txt")
annotation$state = unlist(abundance[rownames(abundance)=="S04",])
results = coxph(Surv(OS_Time, OS_Status) ~ state + Histology, data = annotation)
summary(results)
```

```
Call:
coxph(formula = Surv(OS_Time, OS_Status) ~ state + Histology,
    data = annotation)

  n= 245, number of events= 90
   (5 observations deleted due to missingness)

                coef exp(coef) se(coef)      z Pr(>|z|)
state       -1.5845    0.2051   0.7165 -2.211    0.027 *
HistologyLUSC -0.2104    0.8103   0.2337 -0.900    0.368
```

Maximum number of states per cell type

The maximum number of states per cell type is set to 20 by default.

Cophenetic coefficient cutoff

As described in **Section 3.1.4**, the cophenetic coefficient (CC) is used to select the number of cell states. This field indicates the CC cutoff (range of 0 to 1) used for determining the number of states in *step 4* (**Figure 2A**). By default, the cutoff is set to 0.95, and in general, the lower the cutoff, the more clusters are identified. Users can select a more appropriate cutoff after inspecting CC plots obtained after the first run. This is done by re-running the pipeline and skipping *steps 1-3* (**Figure 2A**).

CIBERSORTx Fractions/HiRes Singularity path

These two fields specify the paths to the Singularity containers (.SIF files) for the CIBERSORTx Fractions and HiRes modules. These fields are relevant when running EcoTyper on servers or high-performance computing clusters that support Singularity instead of Docker as a container platform. If these paths are provided, cell fraction estimation (*step 1*, **Figure 2A**) and expression purification (*step 2*, **Figure 2A**) will be performed using Singularity. If set to “NULL”, these steps will be performed using Docker.

3.1.11 Run Bulk Discovery Mode

After editing the configuration file (*config_discovery_bulk.yml*), the script can be run as indicated below. Please note that running this script on the example dataset might take ~2 hours with 10 threads. Although EcoTyper can be run on the example dataset using a typical laptop with ~16GB RAM and at least 10 CPU cores, larger datasets may require a server or cluster. In general, we recommend using a system with at least 10-20 cores and 32GB of RAM. For details on how to parallelize the code and run on a cluster, see **Note 9**.

```
Rscript EcoTyper_discovery_bulk.R -c config_discovery_bulk.yml
```

3.1.12 Overview of Output Files

When run in bulk discovery mode, EcoTyper outputs a directory containing the following key files and subdirectories:

1. *rank_plot.pdf*: This plot, shown in **Figure 3A** for the example discovery dataset, consists of cophenetic coefficient (CC) plots calculated for each cell type (calculated for *step 4, Choosing the number of cell states*, **Figure 2A**). The

horizontal dotted line indicates the CC threshold specified in the configuration file (=0.95 by default). The vertical dotted red line indicates the number of states automatically selected based on the CC cutoff.

As indicated above (**Section 3.1.4**), we recommend that users inspect this file to ensure that automatic CC selection provides sensible results. For example, a high cutoff may result in less than two states inferred for a given cell type after applying the AFI filter (**Section 3.1.5**). This may indicate that the CC threshold is too high. In this case, users can lower it and re-run EcoTyper from *step 4* (**Figure 2A**; see also **Notes 10 and 11**).

2. The following output files are stored in subdirectories labeled by cell type:

    1. *state_assignment_heatmap.pdf*: A heat map showing the expression of genes that have the highest $\log_2$ fold change in each cell state versus the remaining cell states (i.e., cell-state-specific marker genes). The columns represent samples in the discovery dataset and the rows represent the marker genes for each cell state. Selected marker genes are shown on the left side of the heat map. The heat map is ordered according to cell state assignment, and the top color bar shows the dominant state for a given sample (see *state_assignment.txt* below). In addition, the heat map shown in **Figure 3B** includes two color bars on top corresponding to *Tissue* and *Histology*. These are parameters that have been provided in the configuration file field *Annotation file column(s) to plot*.
    2. *state_assignment.txt*: This file, consisting of samples as columns and cell states as rows, contains the assignment of samples in the discovery dataset to the cell state with the highest inferred abundance (i.e., the dominant cell state for each sample). If the cell state with the highest abundance in a given sample was filtered out in an earlier quality control step (see **Section 3.1.5**), the sample is considered unassigned and is not included in this output file.
    3. *state_abundances.txt*: This file, consisting of samples as columns and cell states as rows, contains the relative abundance of each cell state in each sample. In contrast to the *state_assignment.txt* file where only samples assigned to a cell state are included, all samples included in the discovery dataset are included here. This file can be used for downstream analyses where tissues are represented as mixtures of cell states (e.g., when evaluating continuous associations with overall survival; see **Section 3.1.13**).
    4. *gene_info.txt*: the $\log_2$ fold change of state-specific genes prior to the quality filtering steps performed in *step 6* and *7* (**Figure 2A**). Rows are

genes are columns are cell states. Because marker genes are identified prior to filtering steps, the columns are labeled *IS01-IS0X*, for *InitialState*. Each gene is assigned as a marker gene of a given state (shown in *InitialState*) based on the maximum fold change (shown in column *MaxFC*). The cell state labels after filtering (*S01-S0X*) are provided in the column *State*.

5. *heatmap_data.txt*: The normalized gene expression data underlying the output heat map shown in *state_assignment_heatmap.*
6. *heatmap_top_ann.txt*: The annotation file provided by the user after being (i) restricted to the samples assigned to cell states and (ii) populated with cell state labels, both pre- (column *initialState*) and post-quality control filtering (column *State*). Only samples assigned to a dominant cell state are included in this file.

Of note, only samples assigned to cell states remaining after the quality control in *steps 6* and *7* (**Figure 2A**) are included in the output files *state_assignment.txt* and *state_assignment_heatmap.pdf*. The samples not assigned to one of these states are considered unassigned and are not included in these output files. In contrast, the file *state_abundances.txt* includes all samples in the discovery dataset.

3. EcoTyper also provides an output directory called *Ecotypes*, consisting of the following files:
    1. *nclusters_jaccard.pdf*: A plot depicting the number of clusters (or “ecotypes”) obtained by clustering the Jaccard index matrix and using the average silhouette width maximization method for cluster number selection (**Figure 3C**). This is the number of ecotypes prior to filtering those with <3 cell states.
    2. *ecotypes.txt*: The cell state composition of each ecotype (the set of cell states comprising each ecotype).
    3. *ecotype_abundance.txt*: The relative abundance inferred for each ecotype across all samples in the discovery dataset.
    4. *ecotype_assignment.txt*: The assignment of samples in the discovery dataset to ecotypes. A *Q* -value is calculated to assess the probability of ecotype assignment for each sample, by comparing the cell state abundances of a sample in a given ecotype versus the cell state abundances of the same sample in all other ecotypes (two-sided unpaired t-test with unequal variance corrected by the Benjamini-Hochberg method). Each sample is then assigned to the ecotype with the highest ecotype abundance if: (i) its corresponding *Q*-value (provided in column *AssignmentQ*) is less than 0.25 and (ii) the sample is assigned to at least one cell state contributing to the ecotype. Otherwise, the sample is

unassigned. The samples not assigned to any ecotype are filtered out from this file. Additional information from the user-provided annotation file is also included.

5. *heatmap_assigned_samples_viridis.pdf*: A heat map of cell state fractions inferred by EcoTyper across the samples assigned to ecotypes (**Figure 3D**). If an annotation file has been provided and the field *Annotation file column(s) to plot* was set, annotation variables will be shown as color bars over the heat map. For example, in the example dataset analyzed here, ecotype 4 (E4) is highly enriched in samples obtained from adjacent normal tissue (**Figure 3D**). Similarly, ecotypes 1 and 6 are enriched in squamous cell carcinoma samples, while other ecotypes are enriched in adenocarcinoma samples.
6. *jaccard_matrix.pdf*: A heat map of the Jaccard index matrix, after filtering ecotypes with less than 3 cell states (**Figure 3E**).

3.1.13 Example Downstream Analyses

Output files produced by bulk discovery mode may be used for a variety of downstream analyses.

For example, the *state_assignment.txt* files can be used for enrichment analyses and comparison with other sample labels that are available for the discovery dataset. In **Figure 3B**, samples assigned to endothelial cell state S01 are largely from adjacent normal tissue (shown in blue in the “Tissue” color bar above the heat map), while the samples assigned to endothelial cell states S02-S04 are exclusively derived from tumor tissue (shown in tan in the “Tissue” color bar). Similarly, endothelial cell state S02 is highly enriched for samples of lung squamous cell carcinoma while endothelial cell state S04 is highly enriched for lung adenocarcinoma samples.

Another example is to interrogate the cell states for associations with clinical outcome. **Figure 3F** shows a Kaplan-Meier plot and log-rank test comparing the overall survival associations between lung cancer patients in the discovery dataset assigned to endothelial cell states S03 and S04 (*state_assignment.txt* file). Patients assigned to endothelial cell state S04 show significantly longer survival time compared to patients assigned to endothelial cell state S03 (log rank $P = 0.042$). The survival follow-up data to conduct this analysis is provided in the annotation file of the example dataset.

For downstream analyses including all samples in the input dataset, one can use the output files *state_abundances.txt* or *ecotype_abundance.tx*t. Such analyses can model cell state or ecotype abundance as a continuous variable using Cox proportional hazards regression (*coxph* from the R package *survival* [10]). To calculate continuous survival associations for EcoTyper-derived states and ecotypes, the relative abundance

estimates can be provided as explanatory variables to the model. In **Figure 3G**, we show an example script where endothelial cell state S04 is provided as a continuous explanatory variable to *coxph()*, adjusting for the categorical variable “Histology”, which specifies whether a sample is LUAD or LUSC. This analysis shows that endothelial cell state S04 is significantly associated with longer overall survival time after adjustment for histological subtype (P = 0.027).

Other examples of relevant downstream analyses, such as gene set enrichment analyses, ligand-receptor interactions, and associations with immunotherapies, can be found in Luca et al. (2021) [4] and Steen et al. (2021) [6].

3.2 Discovery of Cell States and Ecotypes from scRNA-seq Data

In the previous section, we reviewed how to perform cell state and ecotype discovery from bulk tissue specimens. Here we describe how to discover cell states and ecotypes directly from scRNA-seq data (**Figure 1**). Prior to analysis with EcoTyper, users must perform preprocessing of scRNA-seq data to remove low quality cells, to annotate major cell types, and to integrate across technical batches (e.g., using Seurat or Scanpy). To illustrate the concepts, commands, and files covered in this section, we have included an example discovery dataset consisting of a down-sampled colorectal cancer scRNA-seq atlas [11] (*example_data/scRNA_CRC_data.txt*) along with a sample annotation file (*example_data/scRNA_CRC_annotation.txt*). Key steps are depicted schematically in **Figure 4A**. As several steps are the same as those covered in **Section 3.1**, here we emphasize steps that are specific to scRNA-seq discovery mode.

3.2.1 Extract Cell Type Specific Genes from scRNA-seq Input Matrix (Optional)

To limit cell state discovery to states that are cell type-specific, rather than ubiquitous, and to improve the performance of cell state and ecotype recovery in bulk data (**Section 3.3**), EcoTyper starts (by default) with the removal of genes in scRNA-seq data that are not over-expressed in a given cell type, relative to other cell types. Given considerations of computational efficiency and balanced representation, up to 500 cells per cell type are randomly selected and used for this step. Differential expression is then calculated between each cell type and the remaining cell types. Genes with a Q-value > 0.05 (two-sided Wilcoxon test with Benjamini-Hochberg correction) are omitted from further analysis. This operation is performed by default in *step 1, Extract cell type specific genes* of scRNA-seq discovery mode (**Figure 4A**).

3.2.2 Cell State Discovery

Round One

Similar to bulk data (**Section 3.1.3**), EcoTyper applies NMF to identify cell states from scRNA-seq data. However, to combat the dropout and sparsity inherent to scRNA-seq data, EcoTyper first constructs a cross-correlation matrix for each cell type and then applies NMF to each cross-correlation matrix in order to identify cell states. For efficiency, EcoTyper randomly samples up to 2,500 single-cell transcriptomes per cell type to build each cross-correlation matrix. As described in **Section 3.1.3**, NMF should be run ~5 times for exploratory purposes and up to 50 times for publication-quality results. This procedure is performed in *step 2, Cell state discovery on correlation matrices* of EcoTyper Bulk Discovery Mode (**Figure 4A**; see also **Section 3.2.7**).

Number of Cell States per Cell Type

The number of states per cell type is determined as described in **Section 3.1.4**.

Round Two

Although the use of cross-correlation matrices can overcome sparsity and dropout in scRNA-seq data, the resulting NMF models are dependent on the number of input cells and cannot be used to recover cell states in new data. To address this issue, EcoTyper performs a second round of NMF, in which (i) the creation of cross-correlation matrices is bypassed and (ii) NMF is only applied to cell state-specific marker genes in the scRNA-seq data. In this way, NMF will be conditioned to discover the same cell states as identified in round one, while also generating models that enable recovery of cell states (and ecotypes) in new datasets.

To accomplish this, EcoTyper first identifies marker genes for each cell state by $\log_2$ fold change. Marker genes are ranked by $\log_2$ fold change within each cell state and the top K/N markers per cell state are selected for the second round of NMF, where K = 1,000 (or the total number of genes if <1,000) and N = the number of cell states (i.e., if 5 cell states are detected, the top 200 marker genes per state are selected). In this case, NMF is applied to the original scRNA-seq data after (i) standardizing each marker gene to a mean of zero and standard deviation of one and (ii) applying posneg transformation to the resulting matrix in order to enforce non-negativity. In this step (*step 5* in **Figure 4A**), EcoTyper applies NMF once using the rank selected in *step 3* (**Figure 4A**) for each cell type. Quality control filtering performed in *step 7* (**Figure 4A**) proceeds as described for bulk data with one minor modification (**Note 12**).

3.2.3 Ecotype Discovery

Ecotype detection from scRNA-seq data is the same as described in **Section 3.1.6**. As this step involves the identification of cell state co-association patterns across tissue samples, we recommend analyzing scRNA-seq data from at least 10 distinct samples to obtain the most reliable results.

3.2.4 Configuring the EcoTyper Run

The script for cell type and ecotype discovery from scRNA-seq data is *EcoTyper_discovery_scRNA.R*. It has three parts: (i) Input, (ii) Output, and (iii) Pipeline settings (**Figure 4B**). Given similarities with the bulk data configuration file (**Sections 3.1.7-3.1.10**), we focus on aspects unique to scRNA-seq discovery mode below. The configuration file used for the example discovery dataset is *config_discovery_scRNA.yml*.

3.2.5 Input Section of Configuration File

Expression matrix

The *Expression matrix* field specifies the path to a tab-delimited file containing the scRNA-seq discovery dataset, with genes as rows and cells as columns. We recommend TPM, counts-per-million (CPM), or SCTransform [12] normalization space without log adjustment. Users should perform their own quality control before applying EcoTyper (e.g., to filter low-quality cells, doublets, etc.); however, we do not recommend pre-filtering the matrix for variable genes as this will be done internally by default (**Section 3.2.1**). The matrix should have gene symbols in the first column and gene expression for each cell in the remaining columns. Column (cell) names should be unique. In this section, we use *example_data/scRNA_CRC_data.txt*.

Annotation file

A path to an annotation file, such as the example available in *example_data/scRNA_CRC_annotation.txt*, should be provided in the field *Annotation file*. This file must contain at least three columns: ID, CellType, and Sample. The “ID” column must have the same names (e.g., cell barcodes) as the columns of the expression matrix. The “CellType” column denotes the cell type annotation of each cell. The “Sample” column indicates the sample identifier for each cell (e.g., “Tumor 1”) and used for ecotype discovery. We recommend at least 10 tissue samples for de novo discovery of cell states and ecotypes.

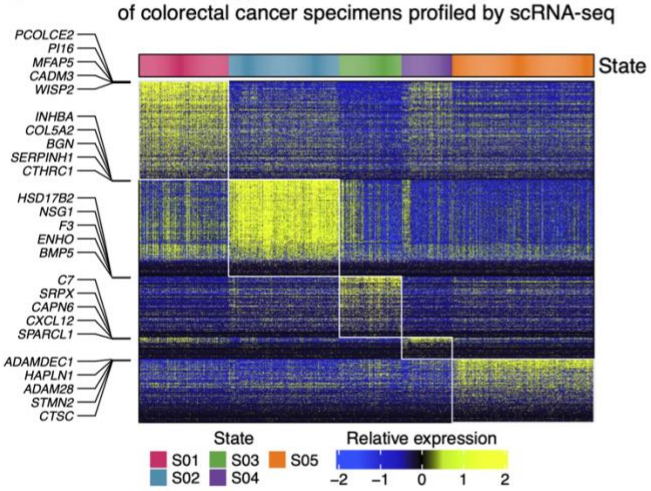
A EcoTyper Single Cell Discovery Mode pipeline

Step in EcoTyper pipeline		Section in text
Step 1	Extract cell-type specific genes from scRNA-seq input matrix	3.2.1
Step 2	Cell state discovery on correlation matrices	3.2.2
Step 3	Choosing the number of cell states	3.1.4 and 3.2.2
Step 4	Extracting cell state information	
Step 5	Cell state re-discovery in expression matrices	3.2.2
Step 6	Extracting cell state information	
Step 7	Cell state QC filtering	3.2.2
Step 8	Ecotype discovery	3.2.3

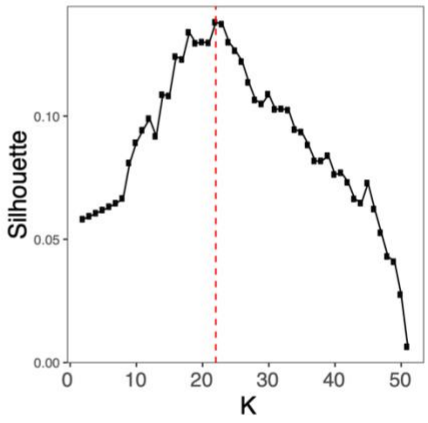
B Configuration file for EcoTyper Single Cell Discovery Mode

```
default :
  Input :                                                    Section 3.2.5
    Discovery dataset name : "discovery_scRNA_CRC"
    Expression matrix : "example_data/scRNA_CRC_data.txt"
    Annotation file : "example_data/scRNA_CRC_annotation.txt"
    Annotation file column to scale by : NULL
    Annotation file column(s) to plot : []

  Output :                                                   Section 3.2.6
    Output folder : "DiscoveryOutput_scRNA"

  Pipeline settings :                                        Section 3.2.7
    #Pipeline steps:
    #   step 1 (extract cell type specific genes)
    #   step 2 (cell state discovery on correrlation matrices)
    #   step 3 (choosing the number of cell states)
    #   step 4 (extracting cell state information)
    #   step 5 (cell state re-discovery in expression matrices)
    #   step 6 (extracting information for re-discovered cell states)
    #   step 7 (cell state QC filter)
    #   step 8 (ecotype discovery)
    Pipeline steps to skip : []
    Filter non cell type specific genes : True
    Number of threads : 10
    Number of NMF restarts : 5
    Maximum number of states per cell type : 20
    Cophenetic coefficient cutoff : 0.975
    #The p-value cutoff used for filtering non-significant overlaps in the
    #jaccard matrix used for discovering ecotypes in step 8.
    #Default: 1 (no filtering).
    Jaccard matrix p-value cutoff : 1
```

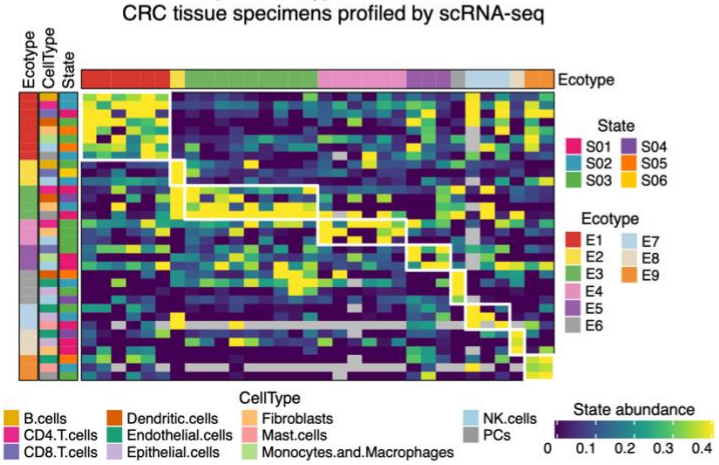
C Output of EcoTyper for cell states discovered in fibroblasts of colorectal cancer specimens profiled by scRNA-seq



D Selection of number of ecotypes with silhouette width



E Output of ecotypes discovered in CRC tissue specimens profiled by scRNA-seq



F Comparison of EcoTyper-derived fibroblast states with author-annotated fibroblast states (Lee et al.)

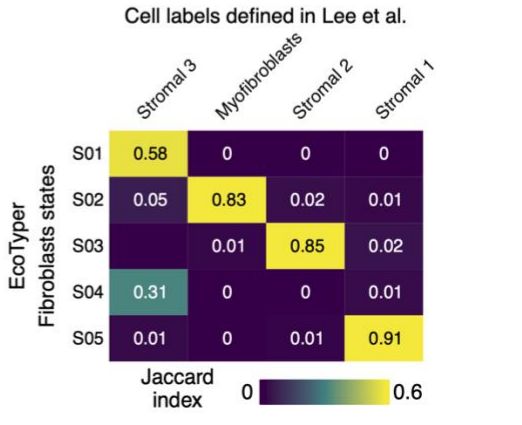


Figure 4. Single-Cell RNA-Seq Discovery Mode and Its Application to Colorectal Cancer. (A) Typical EcoTyper workflow for the discovery of cell states and ecotypes from scRNA-seq data. Key steps are indicated along with their corresponding sections in this chapter. (B) Configuration file for scRNA-seq discovery mode using an example discovery dataset (*n* = 13,780 single-cell transcriptomes from 23 tumor and 10 normal tissue specimens obtained from 23 colorectal cancer patients; Lee et al., 2020). (C) Heat map of fibroblast states identified by EcoTyper in the scRNA-seq discovery dataset. Selected cell state marker genes are indicated on the left side of the heat map. The color bar above the heat map denotes the dominant cell state to which each single-cell transcriptome was assigned. (D) Silhouette plot showing the optimal number of ecotypes identified in scRNA-seq data (dashed red line). (E) Heat map showing ecotype abundances inferred by EcoTyper from the scRNA-seq discovery dataset. Rows are cell states (*n* = 34) columns are patient samples assigned to ecotypes (*n* = 32), and both axes ordered by ecotypes. The color bar above the heat map denotes the dominant ecotype to which each sample is assigned. (F) Jaccard matrix comparing EcoTyper-derived fibroblast states with author-annotated fibroblast states (Lee et al., 2020) in the scRNA-seq discovery dataset.

The annotation file can contain any number of additional columns for use as meta-data. These columns can be used for defining sample batches and for plotting color bars in the heat map output (see **Section 3.1.8**).

3.2.6 Output Section of Configuration File

The Output section contains a single field, *Output folder*, which specifies the path where the final output will be saved. This folder will be created if it does not exist. For the example dataset in this section, set this field to “DiscoveryOutput_scRNA”.

3.2.7 Pipeline Settings

Filter non-cell-type-specific genes

This parameter controls whether to apply the filter for cell type specific genes as outlined in **Section 3.2.1**. We recommend setting this field to “TRUE” to limit cell state discovery to cell type-specific states.

Jaccard matrix p-value cutoff

Ecotype identification is performed by clustering a Jaccard index matrix that quantifies the sample overlap between each pair of states. Prior to ecotype identification, the Jaccard matrix values corresponding to pairs of states for which the sample overlap is not significant are set to 0. The value provided in the *Jaccard matrix p-value cutoff* field specifies the p-value cutoff above which cell state overlaps are considered non-significant. When the number of samples in the scRNA-seq dataset is modest (i.e., <50), such as in the current example, we recommend disabling this filter by setting the p-value cutoff to 1. However, if the dataset has at least 50 samples, we encourage

users to set this cutoff to an alpha value that enforces nominal statistical significance (e.g., 0.05 or lower).

The remaining pipeline settings are the same as for bulk discovery (see **Section 3.1.10**).

3.2.8 Run Single Cell Discovery Mode

After editing the configuration file (*config_discovery_scRNA.yml*), the scRNA-seq discovery script can be executed by running the line of code below. This script may take 24-48 hours to run on 10 threads. Of note, EcoTyper may not be runnable on the example data from this tutorial using a typical laptop with 16GB of memory or less. Therefore, we recommend executing the script on a server or high-performance computing cluster with >50-100GB of RAM (see **Note 9**).

```
Rscript EcoTyper_discovery_scRNA.R -c config_discovery_scRNA.yml
```

3.2.9 Overview of Output Files

The output files generated by scRNA-seq discovery mode are the same as those described for bulk discovery mode (**Section 3.1.12**), with the following exception:

In contrast to cell-type specific sample-level gene expression profiles obtained from bulk samples, which may represent mixtures of cell states, each single cell is only assigned to one specific cell state. Therefore, the main output file for each cell type is the file *state_assignment.txt*, which contains the cell state assignments of single cells assigned to clusters that passed quality control filters. Cells are columns and cell states are rows. Output files from the example dataset are shown in **Figure 4C-E**.

3.2.10 Example Downstream Analysis

The scRNA-seq dataset analyzed in this section includes author-supplied cell type annotations. In **Figure 4F**, we show the overlap between fibroblast cell states identified by EcoTyper (**Figure 4C**) and fibroblast states defined by Lee et al. for the same cells. The overlap is calculated using the Jaccard index. While EcoTyper states S01, S02, S03 and S05 show substantial overlap with author-defined states, state S04 represents a novel state. Despite similarities in cell state definitions, one advantage of using EcoTyper for cell state discovery is the ability to recover cell states and ecotypes in external single-cell and bulk expression datasets. Examples of cell state and ecotype recovery are described below.

3.3 Recovery of Cell States and Ecotypes from Bulk Samples

EcoTyper comes pre-loaded with the resources necessary for reference-guided recovery of previously defined carcinoma and lymphoma-specific cell states and ecotypes [4,6] in user-provided bulk expression data (see **Note 1**). Please note that the recovery procedure described here can also be applied to user-defined cell states and ecotypes (**Sections 3.1 and 3.2**).

This section includes two example datasets in the *example_data* directory:

1. A subset of bulk LUAD transcriptomes from TCGA, available in *bulk_lung_data.txt* with sample annotation file *bulk_lung_annotation.txt*. Of note, this is the same example dataset used in **Section 3.1**.
2. Bulk DLBCL tumor transcriptomes [13] available in *bulk_lymphoma_data.txt* along with annotation file *bulk_lymphoma_annotation.txt.*

3.3.1 Run Recovery from Bulk Data

The script used to perform recovery in bulk data is *EcoTyper_recovery_bulk.R* and takes the following arguments:

-d/--discovery: The name of the discovery dataset used for defining cell states. Users can specify *Carcinoma* or *Lymphoma* (case sensitive) to recover cell states and ecotypes that were previously defined across carcinomas and in lymphoma, respectively [4,6]. Alternatively, for user-defined cell states and ecotypes (**Sections 3.1** and **3.2**), the name of the discovery dataset is the value provided in the *Discovery dataset name* field of the configuration file used for running cell state discovery (**Section 3.1.8**).

-m/--matrix: Path to the input expression matrix. The expression matrix should be in the TPM or FPKM space for bulk RNA-seq data and non-logarithmic (exponential) space for microarrays. It should have gene symbols in the first column and gene expression for each sample in the remaining columns. Column (sample) names should be unique (see also **Note 8**).

-a/--annotation: Path to a tab-delimited annotation file. This argument is optional. If provided, the annotation file should contain a column called ID with the same values as the columns of the expression matrix. Additionally, this file can contain any number of meta-data columns that can be used for plotting color bars in the heat map outputs (see argument *-c/--columns*).

-c/--columns: A comma-separated list of column names from the annotation file (see argument *-a/--annotation*) to be plotted as color bars in the heat map outputs. By default, heat maps contain a color bar denoting the cell state label each cell is assigned to. The column names indicated by this argument will be added to this color bar.

-t/--threads: Number of threads to be used for running the analysis; the default is 10.

-o/--output: Output folder. The output folder will be created if it does not already exist.

3.3.2 Overview of Output Files

For each cell type analyzed, EcoTyper creates a directory containing the following files: *state_assignment.txt*, *state_abundances.txt*, *heatmap_data.txt*, *heatmap_top_ann.txt*, *ecotype_abundance.txt* and *heatmap_assigned_samples_viridis.pdf*, which are identical to the files described in **Section 3.1.12**, with the following exception: the *state_assignment_heatmap.pdf* file displays heat maps of the discovery dataset (left) and the user-provided bulk dataset (right).

These output files can then be used for downstream analyses, such as the ones described in **Section 3.1.13**.

3.4 Recovery of Cell States and Ecotypes from Single-Cell RNA-seq Data

In this section, we show how cell states and ecotypes can be recovered in scRNA-seq data. Importantly, this type of analysis can validate a cell state and ecotype atlas at the single-cell level, though technical caveats, including dropout, sparsity, and dissociation-induced artifacts, may influence results [4,8].

This section can be run with two example datasets provided in the EcoTyper *example_data* directory:

1. A down-sampled colorectal cancer scRNA-seq atlas [11] (*scRNA_CRC_data.txt*) and sample annotation file (*scRNA_CRC_annotation.txt*). Of note, this is the same example dataset used in **Section 3.2**.
2. A down-sampled B cell lymphoma scRNA-seq atlas [6] (*scRNA_lymphoma_data.txt*) and sample annotation file (*scRNA_lymphoma_annotation.txt*).

3.4.2 Run Recovery from scRNA-seq Data

The script used to perform recovery in scRNA-seq data is *EcoTyper_recovery_scRNA.R*. It takes the same input arguments as the script described in **Section 3.3.1**, with the following additions:

-z/--z-score: Boolean flag indicating whether the significance of cell state recovery should be calculated (default is FALSE). This optional procedure allows users to determine whether cell states are significantly recovered in a given dataset. This procedure is detailed in Luca et al. (2021) [4] (*Significance of cell state recovery* in STAR Methods). Of note, the significance calculation can be very slow, as the NMF model is applied 30 times per cell type to generate null distributions.

-s/--subsample: An integer specifying the maximum number of cells per cell type. Cell types exceeding this number will be randomly down-sampled. For values <50, no down-sampling will be performed. Default = –1 (no down-sampling).

3.4.3 Overview of Output files

For each cell type analyzed, EcoTyper outputs a directory containing text files with cell state details (*state_assignment.txt*), along with output plots (*state_assignment_heatmaps.pdf*). These files are identical to those described in **Section 3.2.9**, with the exception that the output plot displays the expression of cell state marker genes in the discovery dataset (left) and the scRNA-seq dataset (right), with the latter smoothed to mitigate the impact the impact of scRNA-seq dropout.

3.5 Recovery of Cell States and Ecotypes from Spatial Transcriptomics Data

As described in **Sections 3.3** and **3.4**, EcoTyper provides a framework to interrogate cell states and ecotypes in external datasets, whether they are obtained from bulk or single-cell cell expression profiling platforms. This also extends to spatial transcriptomics data, for example as obtained with the 10x Visium platform (10x Genomics). In this section, we show how to perform cell state and ecotype recovery from Visium data. A publicly available Visium profile of a breast cancer specimen is used as an example (https://www.10xgenomics.com/resources/datasets/human-breast-cancer-block-a-section-1-1-standard-1-0-0).

3.5.1 Overview of Input files

The Visium input files required for EcoTyper are the filtered feature-barcode matrices *barcodes.tsv.gz*, *features.tsv.gz* and *matrix.mtx.gz* in the format provided by 10x

Genomics and *the tissue_positions_list.csv* file produced by the 10x Genomics Space Ranger pipeline, the latter of which provides the spatial position of each spot barcode. The example files for this section are available in *example_data/VisiumBreast.*

3.5.2 Configuration File

The script that performs cell type and ecotype discovery from bulk Visium data is *EcoTyper_recovery_visium.R*. This script takes as input a configuration file in YAML format. The example configuration file used here is *config_discovery_scRNA.yml*. It consists of three parts: (i) Input, (ii) Output, and (iii) Pipeline settings (**Figure 5A**). Here we focus on aspects that differ from **Sections 3.1.7** and **3.2.4**.

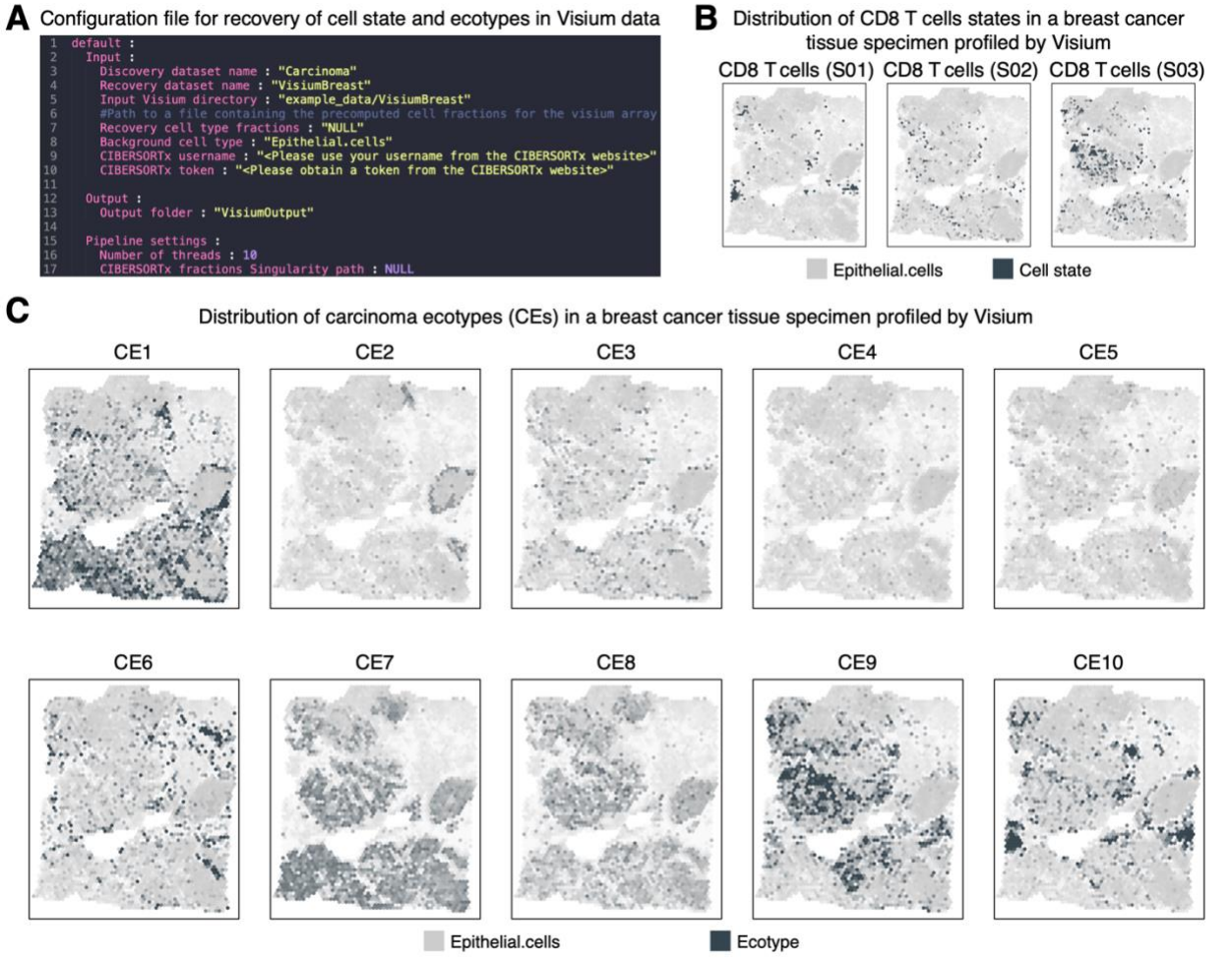


Figure 5. Application of EcoTyper to Spatial Transcriptomics Data. (A) Example configuration file for running EcoTyper on Visium data. (B, C) Heat maps showing the relative abundances of (B) CD8 T cell states and (C) carcinoma ecotypes (CEs) recovered by Carcinoma EcoTyper [4] in a Visium breast cancer profile.

Input Visium directory

The four input files enumerated above must be stored in one directory. The path to that directory is specified in this field. For the example dataset used in this section, the path is *example_data/VisiumBreast*.

Recovery cell type fraction

This field specifies the path to a file containing the cell type fraction estimations for each spot on the Visium array. This field is ignored when the *discovery dataset name* field has been set to *Carcinoma* or *Lymphoma*, or when the discovery has been performed as described in **Section 3.1.8** *Cell type fractions* using *Carcinoma_Fractions* or *Lymphoma_Fractions*. However, if users have determined custom cell type fractions in **Section 3.1.8**, this field must point to a tab-delimited text file containing cell type fractions for each spot in the Visium array. The cell type labels must match those used for cell state/ecotype discovery and the cell type fractions must sum to 1 for each spot. Users are free to select any suitable deconvolution method for this purpose, including cell2location [14], RCTD [15], or CIBERSORTx [8]. An example file is "*example_data/visium_fractions_example.txt*". However, for the example Visium array used in this section, this field should be set to "NULL".

Background cell type

If the tissue specimen analyzed is a tumor specimen, this field specifies the label of the malignant cell type being analyzed. This field is used for plotting a gray background in the resulting output plot, with the intensity of gray depicting the abundance of the malignant cell type. The field is ignored when the *discovery dataset name* field has been set to *Carcinoma* or Lymphoma, or when the discovery has been performed as described in **Section 3.1.8** *Cell type fractions*, using *Carcinoma_Fractions* or *Lymphoma_Fractions*. In these cases, the background cells are automatically considered to be originating from *Epithelial.cells* or *B.cells*, the malignant cell populations in carcinomas and B cell lymphomas, respectively. Otherwise, this field may specify a column name in the file provided to the *Recovery cell type fractions* field. If no column name is specified, all spots in the Visium dataset will be shown in gray.

3.5.3 Run Recovery from Spatial Transcriptomics Data

After editing the configuration file, the command line for recovering cell states and ecotypes in Visium data can now be run. Please note that if spot-level deconvolution results are not provided by the user, this script will invoke CIBERSORTx for cell type fraction estimation, which can take up to two hours to run.

```
Rscript EcoTyper_recovery_visium.R -c config_recovery_visium.yml
```

3.5.4 Output Files

The following output files are generated:

- "*state_abundances.txt*": The columns are spot IDs, X/Y coordinates for each spot in the Visium array, and the inferred cell state abundances across all cell types.
- "*CellTypeLabel_spatial_heatmaps.pdf*", where CellTypeLabel is the annotation given to each cell type: Plots showing cell state abundances for each cell type. The intensity of charcoal represents cell state abundance. The intensity of gray represents the fraction of the background population (**Section 3.5.2**; **Figure 5B**).
- "*ecotype_abundances.txt":* The relative abundance inferred for each ecotype across all spots in the Visium array.
- "*ecotype_abundances.pdf*": Plots depicting inferred ecotype abundances. The intensity of charcoal represents ecotype abundance. The intensity of gray represents the fraction of the background population (**Section 3.5.2**; **Figure 5C**).

3.5.5 Downstream Analysis

Among various downstream analyses, users can determine (i) the distance of each cell state or ecotype to predefined geographic regions (e.g., tumor or stromal regions), (ii) patterns of spatial organization or aggregation, and (iii) potential cell state/ecotype interaction patterns, including heterotypic ligand-receptor pairings, as demonstrated in Luca et al. (2021) and Steen et al. (2021) [4,6].

3.6 Conclusion

EcoTyper is the first dedicated machine learning framework for cell state and ecotype discovery from bulk and single-cell gene expression data. Once defined, cell states and ecotypes can be recovered in external datasets and across platforms, including bulk, single-cell, and spatially resolved gene expression data. As such, we anticipate that EcoTyper will prove useful for understanding the cellular organization of complex tissues in health and disease, with implications for the discovery of new candidate biomarkers and therapeutic targets.

4 Notes

1. The EcoTyper website enables the exploration and recovery of cell states and ecotypes previously identified in carcinoma and lymphoma [4,6]. With the exception of melanoma, where carcinoma cell states have been previously observed and validated [4], users studying other tissue types or disease states

should either (i) carefully validate their results obtained with Carcinoma or Lymphoma EcoTyper (**Sections 3.3** and **3.4**), or (ii) perform a custom discovery of cell states and ecotypes (**Sections 3.1** and **3.2**). Importantly, the EcoTyper website does not provide the latter functionality; discovery modules are only available via the EcoTyper source code, as described in this chapter.

2. In some instances, such as discovery from bulk expression data (**Section 3.1**) or recovery from Visium spatial transcriptomics data (**Section 3.5**), EcoTyper uses CIBERSORTx by default [8]. To run CIBERSORTx, the following resources are needed: (i) Docker or Singularity and (ii) the CIBERSORTx executables available from https://cibersortx.stanford.edu/ (*Menu* > *Download*). Users must follow the instructions on the Download section of the CIBERSORTx website to download the Docker images and obtain the token necessary for running them. If Singularity is used, the Docker images need to be converted to Singularity Image Files (SIF). A link to instructions can be found in the EcoTyper GitHub repository, under the "Frequently Asked Questions" section.

3. Some EcoTyper functions are computationally intensive, especially the cell state discovery modules described in **Sections 3.1** and 3.2. Therefore, EcoTyper is designed as a collection of modular command-line R scripts that can be run in parallel on a multi-processor server or a high-performance computing cluster (also, see **Note 9**).

4. EcoTyper Bulk Discovery Mode may also be run starting from cell-type specific expression matrices, obtained either through FACS-sorting cell populations of interest and then performing bulk tissue expression profiling of each cell population, or by performing *in silico* purification using a deconvolution method comparable to CIBERSORTx.

5. CIBERSORTx High Resolution mode employs an adaptive noise filter that eliminates cell-type-specific genes that cannot be reliably estimated [8]. The more abundant a cell type is, the more genes are imputed.

6. In *step 3, Cell State Discovery* (**Figure 2A**), EcoTyper identifies cell states with NMF. If, for a given cell type, more than 1,000 genes are available (e.g., >1,000 genes imputed by CIBERSORTx High Resolution mode; **Section 3.1.2**), the top 1,000 genes with highest relative dispersion across samples are used for NMF (*Cell state discovery* in STAR Methods in Luca et al, 2021 [4]). Otherwise, all genes are used as input. If <50 genes are imputed for a given cell type, that cell type is not used for cell state identification [4].

7. When run in bulk discovery mode, EcoTyper calculates a “drop out score” to identify poor quality cell states based on marker genes that preferentially show a pattern of low variance and high expression across samples. The drop out score is used for cell state filtering in *step 7, Advanced cell state QC filter* (**Figure 2A**). Calculation of the dropout score is described in Luca et al., 2021 (*Cell state quality control* in STAR Methods) [4].

8. Column names should not contain special characters that are modified by the R function *make.names*. For example, names should not contain digits at the beginning or special characters, including space, tab or dash.

9. EcoTyper can be modified to run on a high-performance cluster by overriding the *pipeline/lib/multithreading.R* library. Currently the library provides two functions, *PushToJobQueue* which adds a command line call to the job queue, and *RunJobQueue*, which waits for all the jobs in the queue to finish. The default implementation of these functions uses the R function *mclapply* to perform computations on multiple threads. Users can re-write these two functions according to the requirements of their cluster infrastructure. A primitive example of how this can be achieved on a high-performance computing cluster built on the SLURM infrastructure is provided in the EcoTyper GitHub repository, under the “Frequently Asked Questions” section.

10. The plots displaying the cophenetic coefficient indicate the number of cell states obtained before applying the filters for low-quality states in *steps 6* and *7* (**Figure 2A**). Therefore, the results will likely contain fewer cell states once filtering is complete.

11. The cophenetic coefficient plot shown in **Figure 3A** might look slightly different than the one obtained by a user. This is because some EcoTyper steps, including the NMF algorithm initialization step, depend on stochastic initialization, and may show modest variation across runs. For stable results, we recommend at least 5 NMF restarts for exploratory analyses and 50 restarts when running EcoTyper for publication-quality results.

12. The “drop out score” (**Note 7**) is not applied when the input is scRNA-seq data.

Acknowledgments

We would like to thank E. Brown for providing critical feedback on the manuscript. This work was supported by grants from the American Association in Cancer Research (C.B.S., 19-40-12-STEE), the National Cancer Institute (A.M.N., R01CA255450 and R00CA187192; A.J.G., U54CA209971 and U24CA224309; A.A.A., R01CA233975), the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) (A.M.N.), the Donald E. and Delia B. Baxter Foundation (A.M.N.)., the Virginia and D.K. Ludwig Fund for Cancer Research (A.A.A., A.M.N.)., the Stinehart-Reed Foundation (A.A.A., A.M.N.), the Bakewell Foundation (A.A.A.), the SDW/DT and Shanahan Family Foundations (A.A.A.) and the Fund for Cancer Informatics (A.J.G.).

References

1. Armingol E, Officer A, Harismendy O, Lewis NE (2021) Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet 22 (2):71-88. doi:10.1038/s41576-020-00292-x

2. Elmentaite R, Dominguez Conde C, Yang L, Teichmann SA (2022) Single-cell atlases: shared and tissue-specific cell types across human organs. Nat Rev Genet. doi:10.1038/s41576-022-00449-w

3. Keidar Haran T, Keren L (2022) From genes to modules, from cells to ecosystems. Mol Syst Biol 18 (5):e10726. doi:10.15252/msb.202110726

4. Luca BA, Steen CB, Matusiak M, Azizi A, Varma S, Zhu C, Przybyl J, Espin-Perez A, Diehn M, Alizadeh AA, van de Rijn M, Gentles AJ, Newman AM (2021) Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell 184 (21):5482-5496 e5428. doi:10.1016/j.cell.2021.09.014

5. Nirmal AJ, Maliga Z, Vallius T, Quattrochi B, Chen AA, Jacobson CA, Pelletier RJ, Yapp C, Arias-Camison R, Chen YA, Lian CG, Murphy GF, Santagata S, Sorger PK (2022) The Spatial Landscape of Progression and Immunoediting in Primary Melanoma at Single-Cell Resolution. Cancer Discov 12 (6):1518-1541. doi:10.1158/2159-8290.CD-21-1357

6. Steen CB, Luca BA, Esfahani MS, Azizi A, Sworder BJ, Nabet BY, Kurtz DM, Liu CL, Khameneh F, Advani RH, Natkunam Y, Myklebust JH, Diehn M, Gentles AJ, Newman AM, Alizadeh AA (2021) The landscape of tumor cell states and ecosystems in diffuse large B cell lymphoma. Cancer Cell 39 (10):1422-1437 e1410. doi:10.1016/j.ccell.2021.08.011

7. Tatlow PJ, Piccolo SR (2016) A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci Rep 6:39259. doi:10.1038/srep39259

8. 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 37 (7):773-782. doi:10.1038/s41587-019-0114-2

9. Steen CB, Liu CL, Alizadeh AA, Newman AM (2020) Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods Mol Biol 2117:135-157. doi:10.1007/978-1-0716-0301-7_7

10. Therneau TM, Grambsch PM (2000) The cox model. In: Modeling survival data: extending the Cox model. Springer, pp 39-77

11. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, Vanhecke J, Verbandt S, Hong H, Min JW, Kim N, Eum HH, Qian J, Boeckx B, Lambrechts D, Tsantoulis P, De Hertogh G, Chung W, Lee T, An M, Shin HT, Joung JG, Jung MH, Ko G, Wirapati P, Kim SH, Kim HC, Yun SH, Tan IBH, Ranjan B, Lee WY, Kim TY, Choi JK, Kim YJ, Prabhakar S, Tejpar S, Park WY (2020) Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet 52 (6):594-603. doi:10.1038/s41588-020-0636-z

12. Hafemeister C, Satija R (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20 (1):296. doi:10.1186/s13059-019-1874-1

13. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, Lawrence MS, Roemer MGM, Li AJ, Ziepert M, Staiger AM, Wala JA, Ducar MD, Leshchiner I, Rheinbay E, Taylor-Weiner A, Coughlin CA, Hess JM, Pedamallu CS, Livitz D, Rosebrock D, Rosenberg M, Tracy AA, Horn H, van Hummelen P, Feldman AL, Link BK, Novak AJ, Cerhan JR, Habermann TM, Siebert R, Rosenwald A, Thorner AR, Meyerson ML, Golub TR, Beroukhim R, Wulf GG, Ott G, Rodig SJ, Monti S, Neuberg DS, Loeffler M, Pfreundschuh M, Trumper L, Getz G, Shipp MA (2018) Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med 24 (5):679-690. doi:10.1038/s41591-018-0016-8

14. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, Elmentaite R, Lomakin A, Kedlian V, Gayoso A, Jain MS, Park JS, Ramona L, Tuck E, Arutyunyan A, Vento-Tormo R, Gerstung M, James L, Stegle O, Bayraktar OA (2022) Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol 40 (5):661-671. doi:10.1038/s41587-021-01139-4

15. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, Irizarry RA (2022) Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol 40 (4):517-526. doi:10.1038/s41587-021-00830-w