Elucidata | Case Study

Polly: Explore More

Accelerate cell-type annotation of scRNA-seq Data with the power of curated datasets

OVERVIEW

To enable effective targeting of disease mechanics, knowledge of cellular heterogeneity and dynamics is essential. In the data processing protocols of scRNA-seq experiments, cell type identification is a vital step for subsequent analysis. While the human body is estimated to contain ~ 100 trillion cells, identifying distinct cell types from cluster-based experiments remains a challenge.

Despite emerging advances in annotation methods of single-cell experiments, a consensus amongst R&D teams is that manual annotation of cell types is often time-consuming and suffers from limited reproducibility. To overcome these limitations, we demonstrate a reproducible bioinformatics solution to identify cell types in scRNA-seq datasets representing liver tissues from Polly's Liver OmixAtlas.

CHALLENGES

- Semi-structured, raw scRNA-seq data from public repositories are difficult to retrieve and integrate together for cell-type and cell-function annotation exercises.
- Performing a literature search for cell-type markers from each cluster can be a time-consuming and error-prone process.

APPROACH

CELL TYPE ANNOTATION OF scRNA-seq DATA ON POLLY

Analysis of scRNA-seq datasets generally starts with dimensionality reduction and clustering. Further, assigning identities to the cells in each of the clusters generated, a process known as annotation is a crucial step in scRNA-seq data analysis.

The cluster-based automatic annotation method using marker gene database as reference is one of the several types of strategies that have been reported to aid annotation efforts.

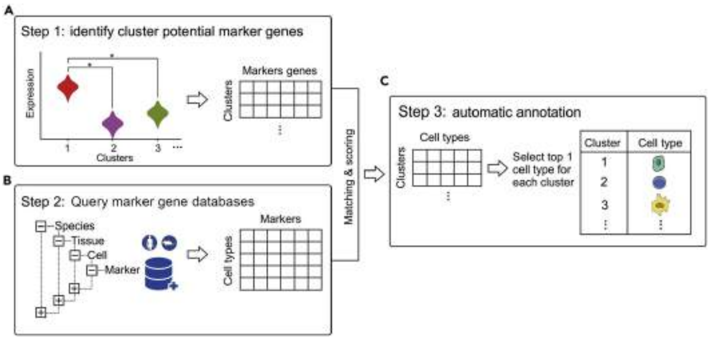

A
Step 1: identify cluster potential marker genes
Expression
1
2
3
...
Clusters
Markers genes
Clusters
B
Step 2: Query marker gene databases
Species
Tissue
Cell
Marker
Markers
Cell types
Matching & scoring
C
Step 3: automatic annotation
Cell types
Clusters
Select top 1 cell type for each cluster
Cluster
Cell type
1
2
3

Figure 1. Workflow for Automated cell-type annotation in single cell transcriptomics data. A) Identification of marker genes from pre-computed cluster. B) Clustermole is used to query marker gene databases which uses the selected markers to find overlap with manually curated databases. C) The cell type with highest overlap is selected as the cell type for that cluster.

Annotating scRNA-seq clusters with corresponding cell types is a challenging task.

Build and implement custom pipelines on Polly and harness the power of curated datasets to execute successful annotation excercises.

A validated , robust pipeline was built that semi-automates the marker gene database annotation method and successfully identifies cell types of in cell clusters.

Curated omics data and metadata play a crucial role in building robust pipelines that assist the automation of cell-type annotations. The curated scRNA-seq dataset used for this study was easily queried by its GSE_ID and retrieved from Polly's Liver OmixAtlas. Datasets of interest can be queried by GSE_ID, dataset_ID, and publication identifier on Polly.

The dataset **GSE124395** contained healthy liver cells from 9 donors lacking known cell-type annotation (**Aizarani N., Saviano A., Sagar et. al**.). A pipeline built on Polly was run on this dataset that predicted the cell types in the present clusters based on the **marker gene database annotation method** illustrated below.

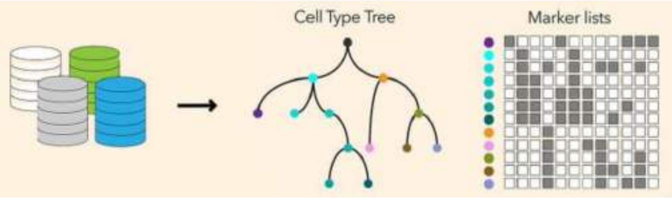
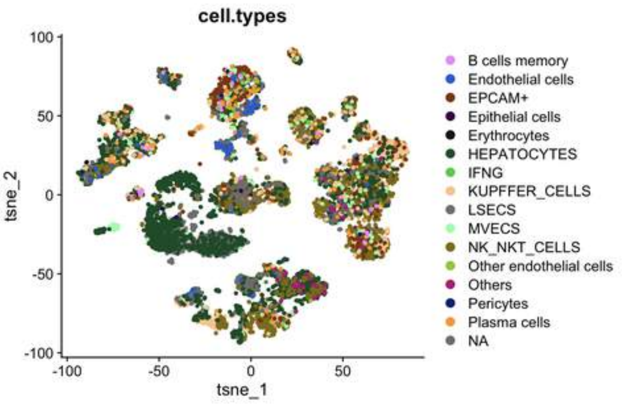
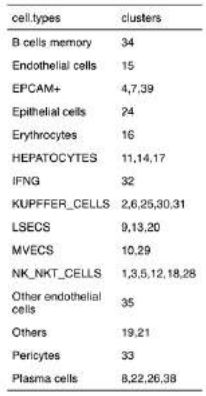


Figure 2. Marker gene database-based annotation takes advantage of cell type atlases. Literature- and scRNA-seq analysis-derived markers have been assembled into reference cell type hierarchies and marker lists. In this approach, basic scoring systems are used to ascribe cell types at the cluster level in the query dataset.

RESULTING CELL TYPE ANNOTATIONS USING MARKER GENES





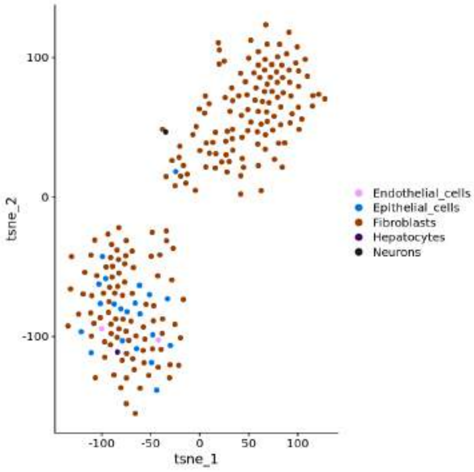
cell.types	clusters
B cells memory	34
Endothelial cells	15
EPCAM+	4,7,39
Epithelial cells	24
Erythrocytes	16
HEPATOCYTES	11,14,17
IFNG	32
KUPFFER_CELLS	2,6,25,30,31
LSECS	9,13,20
MVECS	10,29
NK_NKT_CELLS	1,3,5,12,18,28
Other endothelial cells	35
Others	19,21
Pericytes	33
Plasma cells	8,22,26,38

Figure 3. t-SNE plot showing annotated cell types in a liver scRNAseq dataset using **clustermole** on Polly. The pipeline was successful in annotating all the cell types present in the experimental clusters.

On Polly, it is possible to build and run custom pipelines on thousands of curated datasets of interest. Combining datasets to derive meaningful biological insights is easy with Polly.

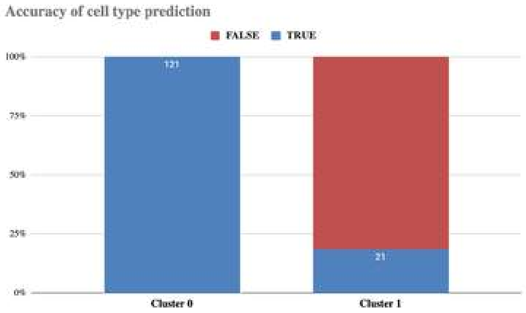
VALIDATION OF THE PIPELINE

The dataset GSE99989 was used to validate the accuracy of the pipeline. The dataset contained cells separated by two clusters. As we can see from the t-SNE plot below, one of the clusters (top) is well defined by a single cell type belonging to the fibroblast population. However, the adjoining clusters contain heterogeneous cell types. On looking at the plot, the cluster defined by the fibroblast population is most likely to be correctly predicted by the pipeline.



CELL TYPE ANNOTATION OF THE VALIDATION DATASET

Following the pipeline mentioned above, markers were identified in each of the two clusters. The cells were identified using the top 50 most significant marker genes. A larger gene set was chosen so as to account for the smaller size of the cluster. A small gene set will not be a significant representation for a small-sized cluster.



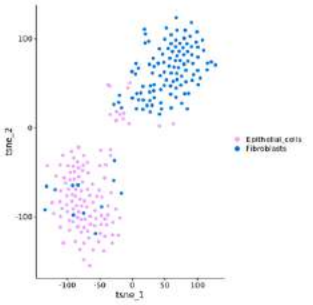


Figure 4. (T) Predicted cell-types.

Figure 5. (L) Cluster level accuracy shows 100% accuracy in cluster 0 defined by a single cell type population.

VALIDATION DATA USING KNOWN CELL TYPES MINED FROM LITERATURE

The dataset **GSE115469** was used to further validate the accuracy of the pipeline. The scRNA-seq dataset used was a normal liver dataset comprising 8439 cells taken from 5 patients.

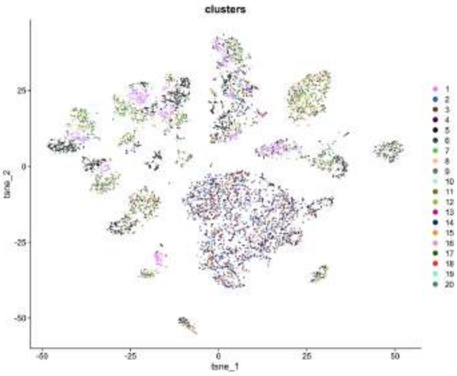
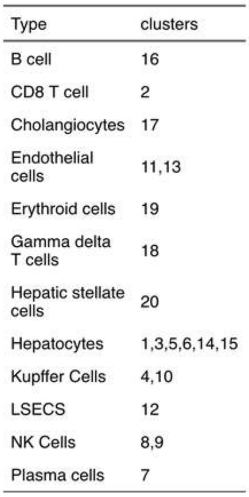


Figure 6. The dataset GSE115469 contained various cell types segregated by 20 pre-computed clusters.

CELL TYPE ANNOTATION OF THE VALIDATION DATASET

Using the same pipeline, markers from each cluster in the dataset were identified. Using the top 25 markers from each cluster, single-cell types were annotated. Based on this, 12 cell type populations were identified and further compared against the original experimentally identified cell types for validation.



Type	clusters
B cell	16
CD8 T cell	2
Cholangiocytes	17
Endothelial cells	11,13
Erythroid cells	19
Gamma delta T cells	18
Hepatic stellate cells	20
Hepatocytes	1,3,5,6,14,15
Kupffer Cells	4,10
LSECS	12
NK Cells	8,9
Plasma cells	7

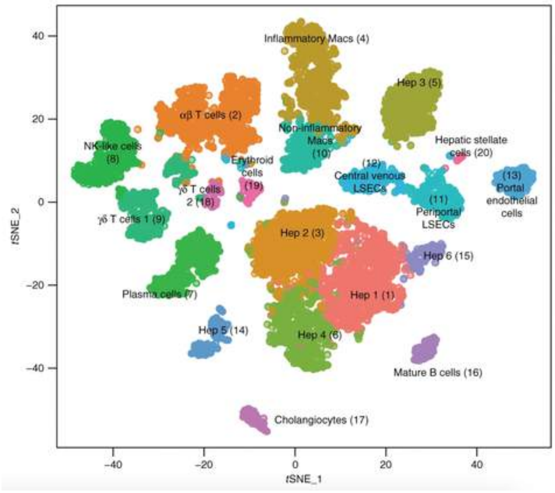


Figure 7. (L) The pipeline built on Polly revealed the biological cell types for each cluster (Left). t-SNE plot showing cell type populations in the normal liver dataset. (R) The cluster number for each cell population is indicated in parentheses. On comparing the table on the left, with the tSNE plot on the right, cell types are correctly identified by the pipeline as described above. Hence, we can confirm that this pipeline works on annotating single cell clusters that are well defined. We were able to clearly annotate homogeneous clusters with their corresponding cell types.