Cell type labeling of chronic lymphocytic leukemia single-cell RNA-sequencing data

Robert Bosek, Marian Schön

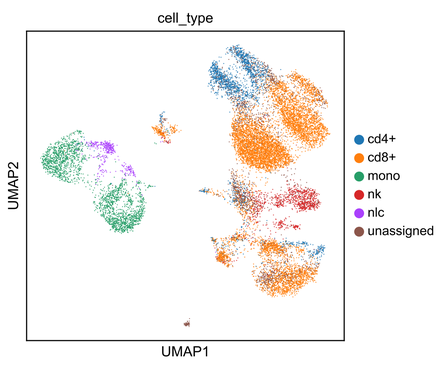
Universität Regensburg

Functional Genomics Regensburg

The motivation for this project is to understand typical and current best-practices for preprocessing and downstream analysis of scRNA-seq data. Therefore we apply a scanpy tutorial pipeline provided by [1] to a CLL dataset of PBMC transcriptomic data provided by [2] with the aim to assign cell type labels to expression profiles. We give insight into methods and results for different steps in the analysis workflow and point out alternative ways of assigning cell type labels.
Our results show, that the annotation of expression profiles with cell type labels depends on the strategy and thus is not distinct. Furthermore it also depends on the usage of additional data, like cell type marker gene sets.

Motivation

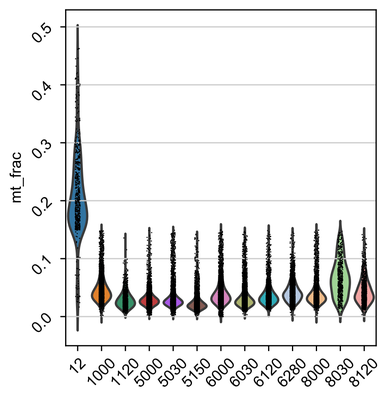
- Understand current best-practices in single-cell RNA-sequencing (scRNA-seq) and scanpy pipeline[1]
- Understand scRNA-seq dataset[2]
- Apply scanpy pipeline on chronic lymphocytic leukemia (CLL) data
- Evaluate cell type labeling by pointing out alternative labeling strategies



CLL Dataset

The scRNA-seq data was published by Rendeiro et al., in scope of a study with 8 CLL-patients, who were set on medication. Peripheral blood mononuclear cell (PBMCs) samples were taken at different points in time in order to monitor the impact of the treatment.

12 of these PBMC-samples were used for single-cell RNA-sequecing (scRNA-seq) following a 10xGenomics protocol. The raw expression data was preprocessed using CellRanger and analysed by **seurat** pipeline.

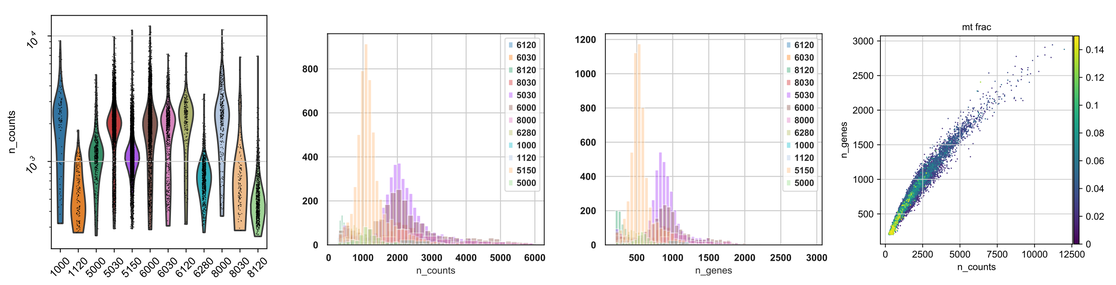


The data contains **43738 celltype labeled single-cell expression profiles with 22021 genes** expressed. Following data exploration results and results from testing the pipeline, profiles with 'malignant' (62,4%) and 'not annotated' (0,1%) celltype-labels were excluded, leaving **15784 profiles** in the dataset.

Quality control

To assess the quality of scRNA-seq data, the quality of the expression profiles is measured in regard to the amount of molecule counts (**n_counts**), amount of genes expressed (**n_genes**), fraction of mitochondrial gene counts expressed (**mt_frac**).

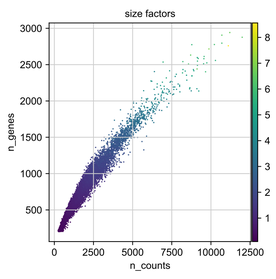
Molecule counts as well as genes expressed can be informative for finding empty droplet or duplicate profiles, high fractions of mitochondrial gene expression can be indicative for cell stress. These information should be regarded together to deduce thresholds.



Our thresholds correspond to Rendeiro et. al, keeping only profiles with:
200 < n_genes < 3000 and **mt_frac > 0.15**
Also genes expressed in less than 20 cells are removed, resulting in data with **15777 profiles and 12608 genes** and n_counts ranging from 253 to 11982.

Normalization and further preprocessing

Normalization in the pipeline is done **by size factor estimation**, which takes into account that cell sizes can differ substantially and so do the expression counts in related profiles.

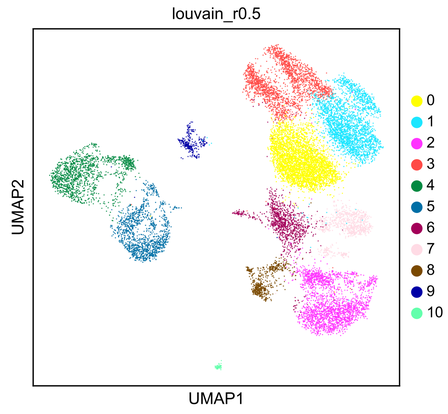


Further preprocessing typically includes:

- **Batch correction**, to adjust data for batch effects (technical variation)
- **Cell cycle scoring**, to adjust data for biological variation
- Calculation of **highly variable genes (HVG)**, to reduce dimensionality of data

Data visualization suggests that the data was already batch corrected and cell cycle scored. The scanpy method to calculate HVGs was not verifyable, thus these three preprocessing steps have been skipped.

Visualisation and Clustering



Louvain clustering can be computed in different resolutions, higher resolutions result in more clusters. We choose a **resolution (r = 0.5)** that results in slightly more clusters as expected celltypes.

UMAP-Embeddings are based on **all features**, as HVGs were not computed. Marker genes for each cluster are all genes that are **significantly, differentially expressed using t-test** with Benjamini-Hochberg p-value correction method and **confidence interval of 0.95**.

Celltype marker gene selection

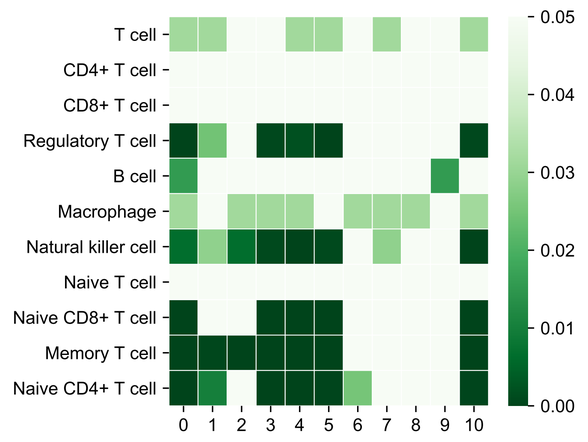
PBMCs are immune cells resulting from differentiation of hematopoietic stem cells released from the bone marrow into the blood stream (Hematopoiesis).

Cluster of differentiation (CD) marker are common marker in immunophenotyping. They have been determined by proteomic approaches, using antibodies to cell surface proteins to assign celltypes to single cells. Due to the proteomic origin, CD marker have to be well-considered in transcriptomic approaches.
Furthermore many transcriptomic marker databases were built on information from microarray data analysis, just as MSigDB used by Rendeiro et al.

CellMarker Database provides manually curated cell type marker genes from microarray as well as scRNA-seq experiments. Therefore we decided upon CellMarker as our marker source and only use **cell type marker genes for PBMCs originating from scRNA-seq experiments**.

Cell type annotation

Two ways of assigning specific cell types to clusters using differential expression of genes are gene set enrichment analysis (GSEA) and gene ontology analysis (GOA).

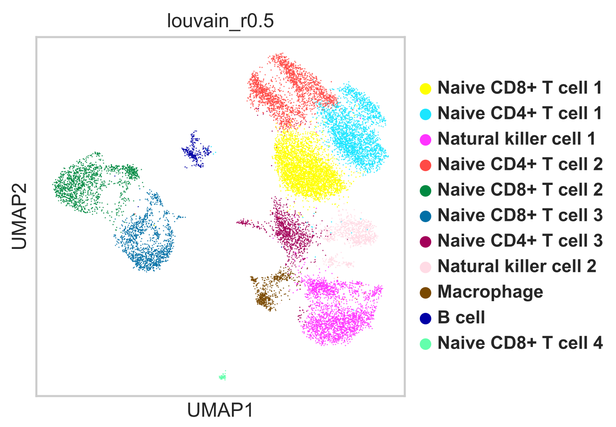


GSEA detects if a set of cell type marker genes is enriched in a set cluster marker genes. The provided method in the tutorial pipeline only intersects a set of cluster markers and a set of cell type markers and is not sufficient for this approach.

We implemented a method to calculate **one sided Fisher's exact test** for each cluster and each cell type. Corresponding p-values are color-coded in the heatmap.

In GOA functional terms (ontologies) are retrieved from GeneOntology Database for given list of genes. Using these ontologies, cell types can be assigned to each cluster. This method of annotation needs to be done manually and is very subjective regarding the complexicity of hematopoesis. The GO-results and plots were not reproducible, thus we did not annotate using GOA.

Comparison of annotations



proportion of cell types

cell type	label	annot
cd8+	0.53	0.40
cd4+	0.13	0.35
nk	0.06	0.19
macro	0	0.04
B	0	0.02
mono	0.14	0
nlc	0.02	0
unassigned	0.12	0

References

[1] M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Mol. Syst. Biol.* 15.6 (June 2019), e8746.
[2] A. F. Rendeiro et al. "Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib response in CLL". In: *Nat Commun* 11.1 (Jan. 2020), p. 577.