Single-cell RNA-sequencing data analysis pipeline

Progress report

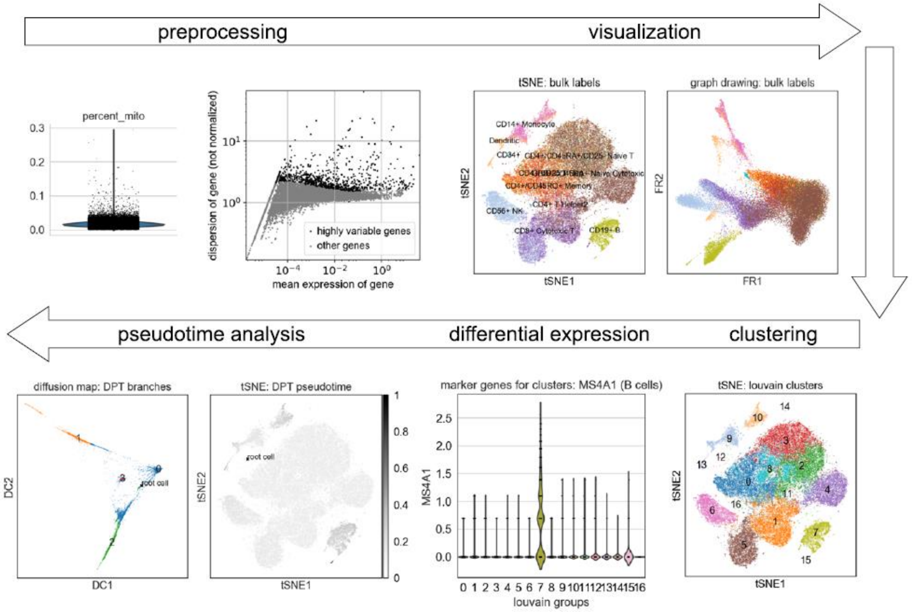
Robert Bosek
Institute for functional genomics
FACULTY OF MEDICINE

Universität Regensburg

Tasks

Apply scanpy pipeline provided in scope of „Current best practices in single-cell RNA-seq analysis" (Luecken & Theis 2019) to scRNA-seq Data from a CLL study (Reindero 2019)

Tasks



Picture from ‘SCANPY: large-scale single-cell gene expression data analysis’

1. Understand current best practices in single-cell RNA-seq workflow (Luecken & Theis 2019)

2. Adapt scRNA-seq pipeline to CLL Data

3. Try to include flexible decision metrics in each step, data-dependent where necessary

The Data

peripheral blood mononuclear cells (PBMCs) from 7 chronic lymphocytic leukemia (CLL) patients in Central Hospital of Southern Pest, Budapest, Hungary

patients were treated with Ibrutinib (bruton tyrosine kinase inhibitor), which provides effective treatment by inducing lymphocytosis and impacting cell-cell cohesion

samples of each patient were taken at 8 time-points during treatment (0, 1, 2, 3, 8, 30, 120/150, 240 days into treatment) and phenotyped by flow cytometry

also Droplet-based single-cell RNA-sequencing with a subsample

The Data

peripheral blood mononuclear cells (PBMCs) from 7 chronic lymphocytic leukemia (CLL) patients in Central Hospital of Southern Pest, Budapest, Hungary

patients were treated with Ibrutinib (bruton tyrosine kinase inhibitor), which provides effective treatment by inducing lymphocytosis and impacting cell-cell cohesion

samples of each patient were taken at 8 time-points during treatment (0, 1, 2, 3, 8, 30, 120/150, 240 days into treatment) and phenotyped by flow cytometry

also droplet-based single-cell RNA-sequencing with a subsample

The Data

peripheral blood mononuclear cells (PBMCs) from 7 chronic lymphocytic leukemia (CLL) patients in Central Hospital of Southern Pest, Budapest, Hungary

patients were treated with Ibrutinib (bruton tyrosine kinase inhibitor), which provides effective treatment by impacting cell-cell cohesion and inducing lymphocytosis

samples of each patient were taken at 8 time-points during treatment (0, 1, 2, 3, 8, 30, 120/150, 240 days into treatment) and phenotyped by flow cytometry

also Droplet-based single-cell RNA-sequencing with a subsample

→ **this celltype-annotated subsample is our dataset**

scRNA-seq analysis pipeline

contains several steps of data preprocessing and visualization as well as several downstream analysis steps (clustering, cluster annotation using marker genes, trajectory inference, differential gene expression and more)

to save time for the initial testing and adapting of the pipeline we used a subset of our data, which contains 10% of cells for each celltype taken randomly

preprocessing

to reduce complexity and prepare the dataset for downstream analysis

reduce dimensions by sorting out genes and cells without sufficient information in the following steps

preprocessing

quality control:
define metrics for cell and gene quality control
→ **with metrics find thresholds for processing the data**

Rendeiro et al (2019) excluded cells with:
- less than 200 expressed genes (indicative of no cell in droplet)
- more than 3000 expressed genes (indicative of cell duplicates)
- and cells with more than 15% mitochondrial genes expressed (indicative of cell stress)

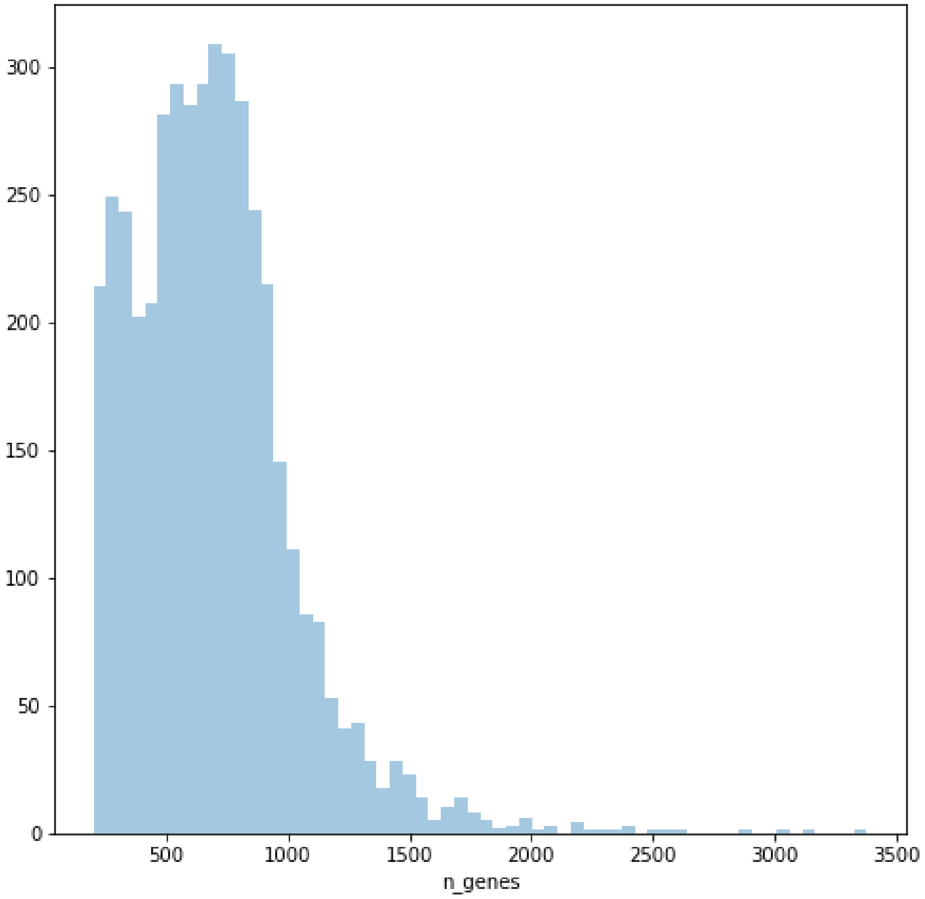
Luecken & Theis (2019) suggest to exclude genes that are expressed in less than 20 cells

Robert Bosek
Institute for functional genomics
FACULTY OF MEDICINE

preprocessing

quality control:
find metrics for cell and gene quality control like:

- total number of molecule counts (n_counts)
- total number of expressed genes
- mitochondrial gene expression count fractions

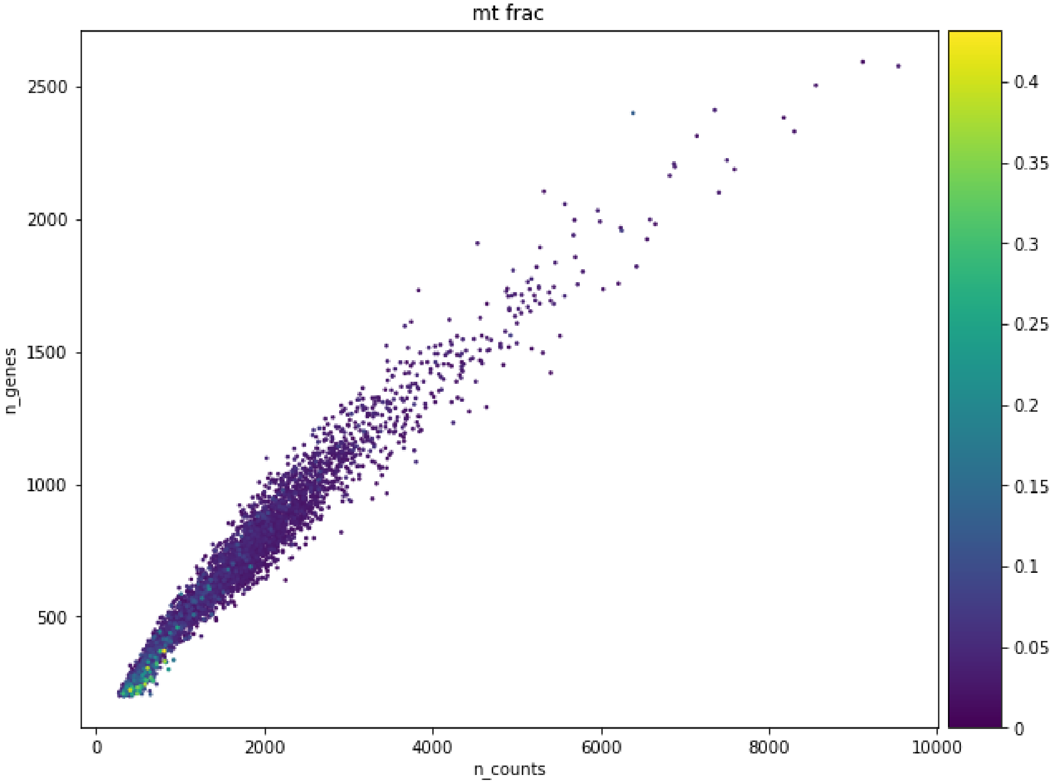


distribution plot of the total number of expressed genes

preprocessing

quality control:
find metrics for cell and gene quality control like:

- total number of molecule counts (n_counts)
- total number of expressed genes
- mitochondrial gene expression count fractions



scatter plot of cells the with their total number of molecule counts vs. total number of genes and a color gradient showing the mitochondrial gene expression count fractions

Can we automate the pipeline with more flexible thresholds?

preprocessing

normalization:
problem: scRNA-seq does not capture every single mRNA molecule in a cell

cells do not have the equal amount of molecules expressed and differ in size which also impacts the amount of molecules expressed

to normalize expression values we can estimate cell-specific factors, proportional to the true number of molecules of a cell

→ **divide measured counts by the size factor for each cell**

preprocessing

data correction and integration methods

batch correction: tries to handle systematic errors between different samples

cell cycle scoring: uses a list of genes to classify cells by their cell cycle phase to use cell cycle effect scores for further data correction

considering natural variation we skiped these preprocessing steps for now (random vs. systematic error)

Might come back to these methods if we want to integrate different datasets later on!

preprocessing

feature selection:

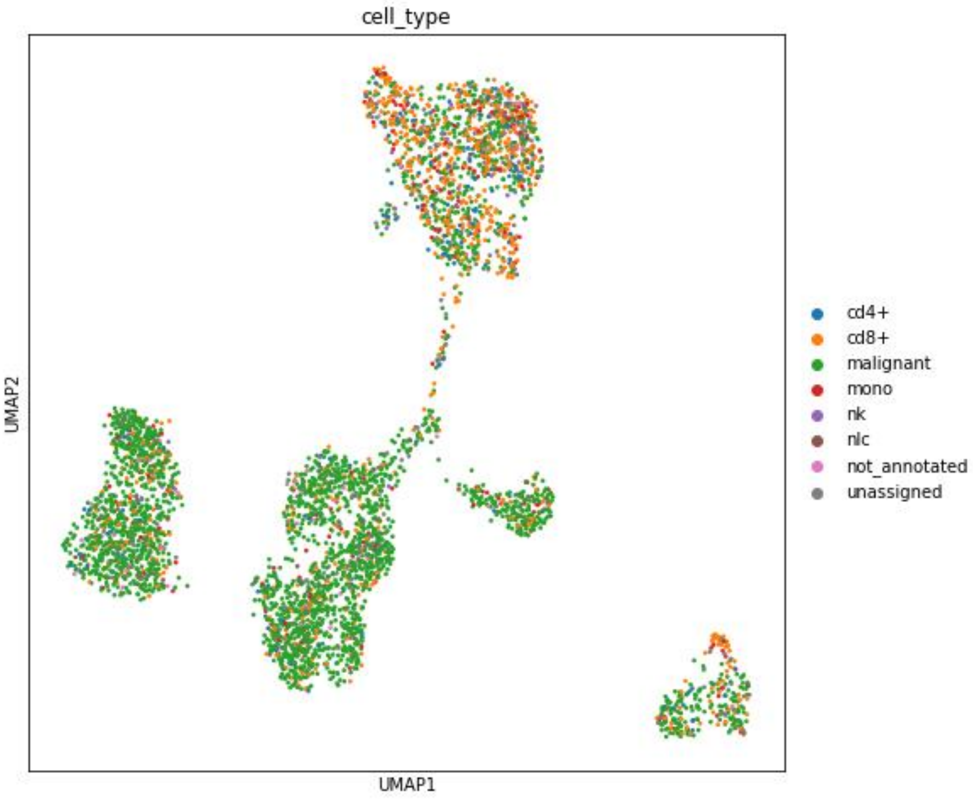
select highly variable genes (HVG) to further reduce dimensionality and only include most informative genes

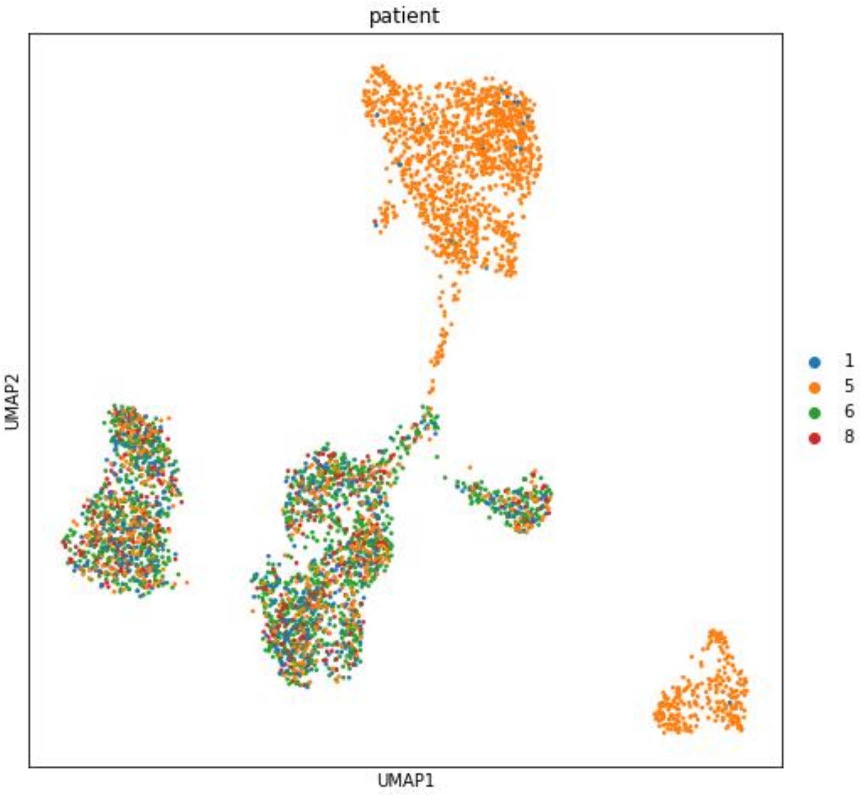
Luecken and Theis suggest to select between 1000 and 5000 HVGs

→ For first tests we selected 4000 HVGs on the data subset

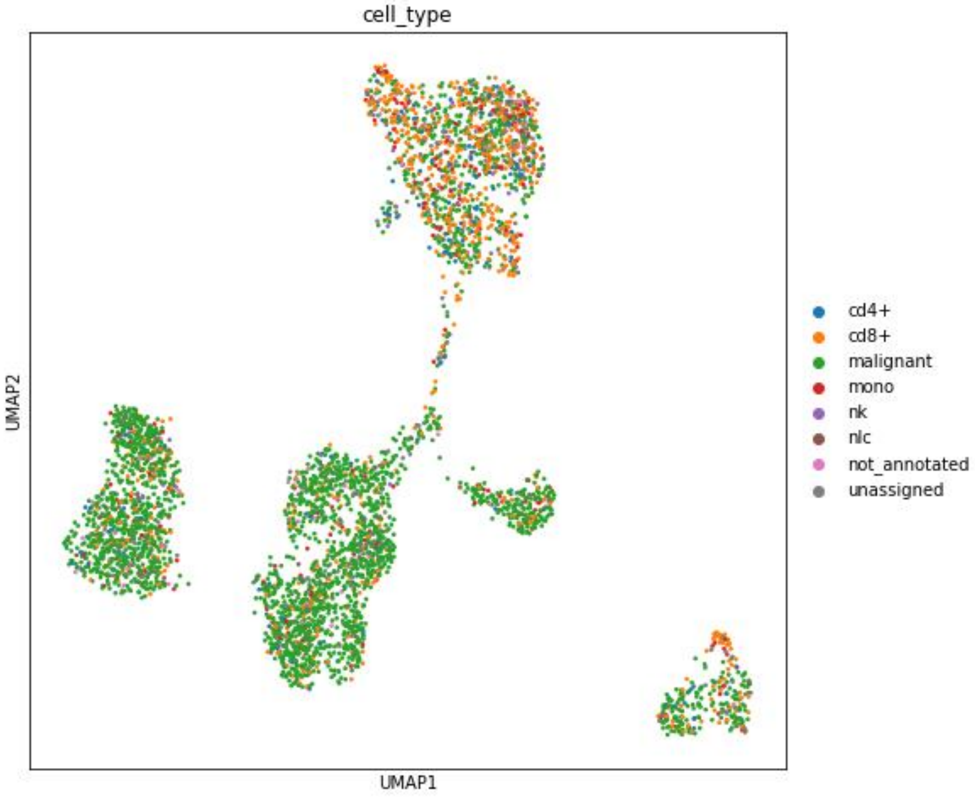
visualization

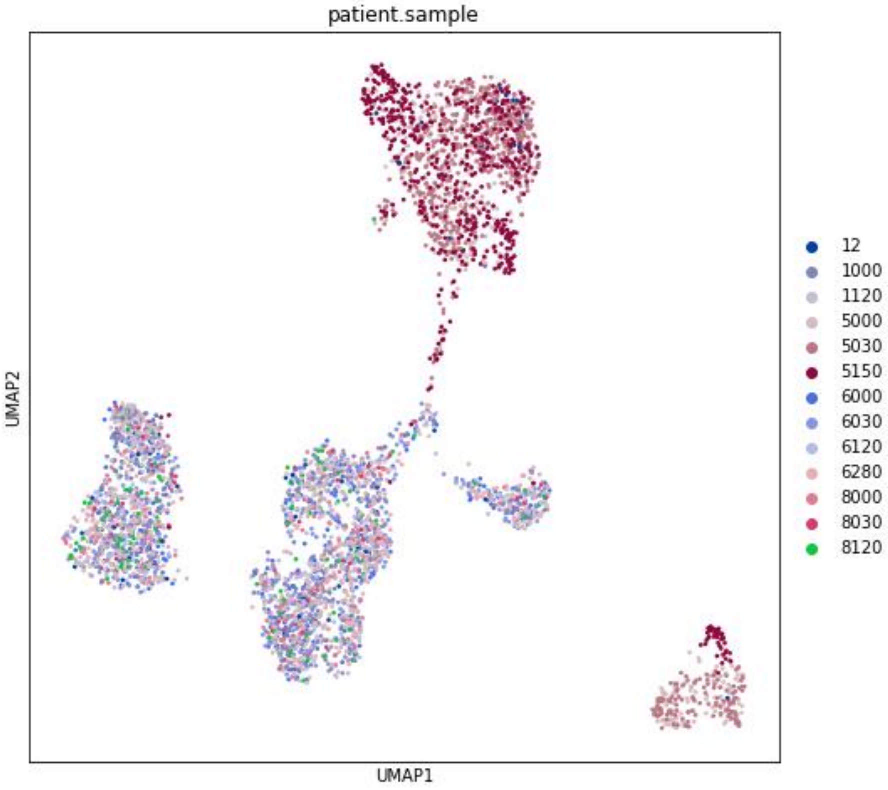
Further methods of dimensionality reduction to printable 2D





visualization





work to do

clustering:

- find good sets of marker genes for PBMCelltypes
- clustering and cluster annotation, if necessary with more general celltypes and subclustering afterwards

more downstream analysis:

- Compositional analysis, trajectory inference and pseudotime analysis
- Slingshot and non-batch corrected slingshot, monocle2, DPT, gene expression dynamics, metastable states, partistion based graph abstraction, differential gene expression, gene set analysis

for further application keep in mind:

- data integration methods
- automation of celltype annotation by comparing top ranked cluster genes to a database of marker genes

Sources

Wolf, F., Angerer, P. & Theis, F. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19,** 15

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, *15*(6)

Rendeiro, A.F. (2019) Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib drug response in chronic lymphocytic leukemia