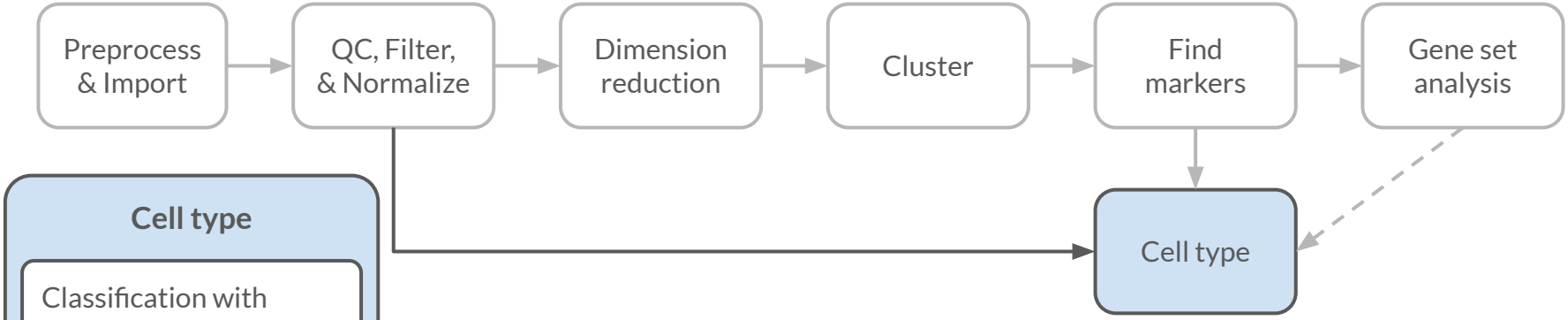
# Annotating Cell Types from Single-cell Data

The Data Lab




**Cell type**

Classification with known marker genes

Cell-typed reference dataset

Identify informative genes in reference `trainSingleR()`

Classify cells `classifySingleR()`

- Manual Classification
  - Look at expression patterns of marker genes for cell types of interest
  - Assign cell types to individual cells or clusters of cells
  - Take advantage of expert knowledge
- Automated Classification
  - Use previously annotated reference data
  - Many algorithms: we will focus on `SingleR`
  - Quality of results depends on quality and appropriateness of reference data
    - Examine and validate your results!

# Manual annotation

- Look at expression of known marker genes
- Label cells or clusters by expression thresholds
- Labor intensive, often requires expert knowledge


https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/tutorial-celltypes

# Where do you find marker genes?

- Literature and databases
  - commonly used marker genes and proteins
  - databases of marker genes
    - C8 dataset from MSigDB Cell type signature gene sets
    - PanglaoDB
    - CellMarker (flakey lately)

- Within your data
  - differentially expressed genes among clusters

# This is fine

*Although labour-consuming and inevitably subjective and arbitrary, manual annotation must be mentioned here, because this primitive slash-and-burn mode of research is still being adopted by many laboratories and biotech companies…*

– Wang, Ding, and Zou 2020

*The database of cell type markers was compiled by manual curation of thousands of published articles and abstracts, and by querying internet search engines with strings such as 'GENE1 is expressed in * cells'.*

– Franzén, Gan, and Björkegren 2019

# Automated annotation tools

- Dozens of methods
  - A recent review: Xie *et al.* (2021)
- Marker-based
  - Use an input database of cell type marker genes
  - Assign cell types based on which genes are expressed in each cell
  - CellAssign, scTyper, scType
- Reference-based
  - Use a cell-type annotated gene expression reference dataset
  - Assign cell types based on the closest match in the reference
  - SingleR, Seurat/Azimuth, scArches
- ChatGPT/LLMs
  - Already a number of papers/preprints looking at this: Hou & Ji (2024); Zeng and Du (2023)

# Why focus on SingleR

- It works well and is well-integrated with the pipeline/tools we use
  - We have not rigorously evaluated others, YMMV!
- Flexible
  - Reference data can be bulk data RNA-seq, single-cell RNA-seq, even bulk microarrays
- Great documentation:
  - http://bioconductor.org/books/release/SingleRBook/

# SingleR method

- “Train” on a labeled reference set
  - Identify most informative genes
    - differentially expressed among labels
    - method can vary depending on reference type
- Assign labels to query cells
  - For each cell (or group of cells), calculate the Spearman correlation to each reference sample
    - use only the markers identified previously
    - calculate  label scores based on correlations to reference samples with each label
    - top label score is the assigned label
  - Optionally: fine tune labels
    - select top labels
    - repeat above, using only markers relevant to the top label groups

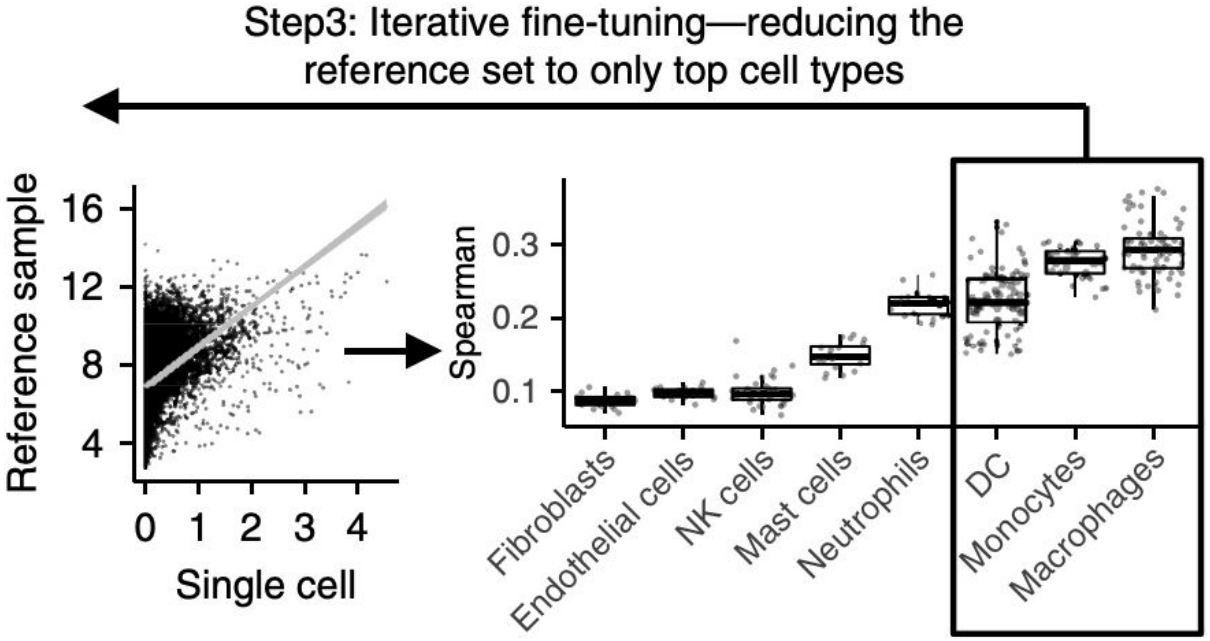
# SingleR method

Step 1:

Identifying variable genes among cell types in the reference set

Step 2:

Correlating each single-cell transcriptome with each sample in the reference set




Aran *et al.* (2019) https://doi.org/10.1038/s41590-018-0276-y

# Reference datasets

- Quality and appropriateness of the reference will strongly affect results
  - If your reference dataset does not have the cell types you are interested in, you won't find them!
  - Many methods will bias towards assigning labels, even if there is no good match

- Finding "good" reference data is one of the biggest challenges
  - Previous experiments, if you trust them
    - manual annotation isn't all bad!
  - Public data, if you can find it
    - `celldex` package: precompiled reference datasets ready for SingleR (Usage guide)
    - Azimuth reference datasets
  - Be aware that most public datasets are of normal tissue!
    - If you are working with cancer cells, you may have trouble finding good references for automated annotation

# Validating automatic annotations

"Don't blindly trust the robots on this one." –Ally

- Some methods provide diagnostic statistics and plots that may be useful
- Compare labels using different tools and references
- Visualize known markers to validate assignments