sc-RNA-seq annotation

Mary O'Neill, Ph.D.
Director, Single Cell Genomics
ONeillMB@uw.edu
@ONeillMB1

Anh Vo, M.S.
Bioinformatician
athuyvo@uw.edu

Clustering

What is Clustering in scRNA-seq?

- **Clustering** groups cells based on similarity in gene expression profiles, identifying distinct populations or cell types.
- The goal is to identify **biologically meaningful groups** in high-dimensional scRNA-seq data (e.g., cell types, states, or developmental stages).

Types of Clustering

- **Unsupervised Clustering**: No prior labels; groups cells based on intrinsic similarities.
- **Supervised Clustering**: Uses known labels (e.g., cell types) to guide the clustering process.

Common Clustering Approaches

- **Graph-based Clustering**: Treats cells as nodes in a graph, using algorithms like Louvain or Leiden for community detection.
- **K-means**: Partitions cells into K clusters.
- **Hierarchical Clustering**: Builds a dendrogram to group cells based on similarity.
- **Density-based Clustering**: Identifies clusters as regions of high density in the data (e.g., DBSCAN).

Annotation

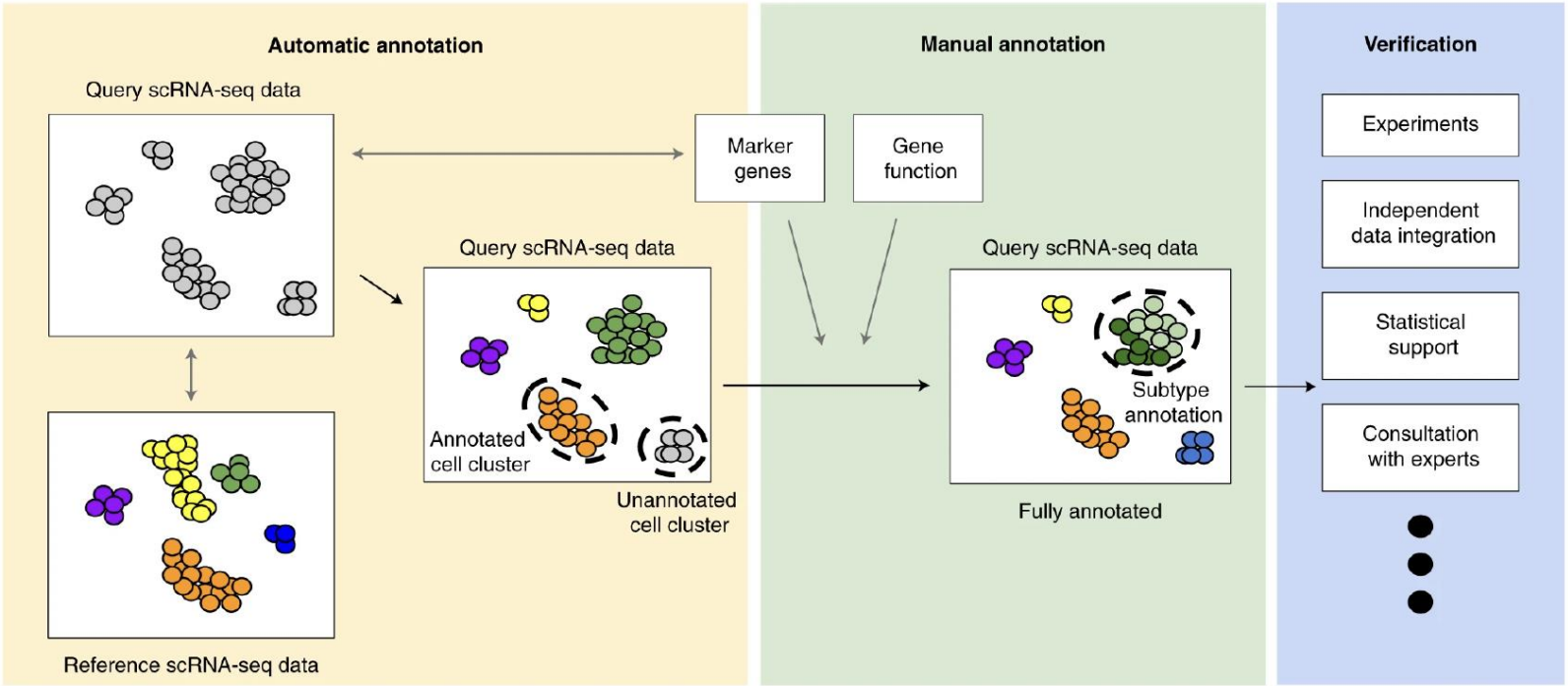
What is Annotation in scRNA-seq?

- **Annotation** is the process of assigning biological meaning to clusters of cells, usually based on their gene expression profiles.
- The goal is to map clusters to known **cell types**, **states**, or **lineages** based on expression patterns and marker genes.

Why is Annotation Important?

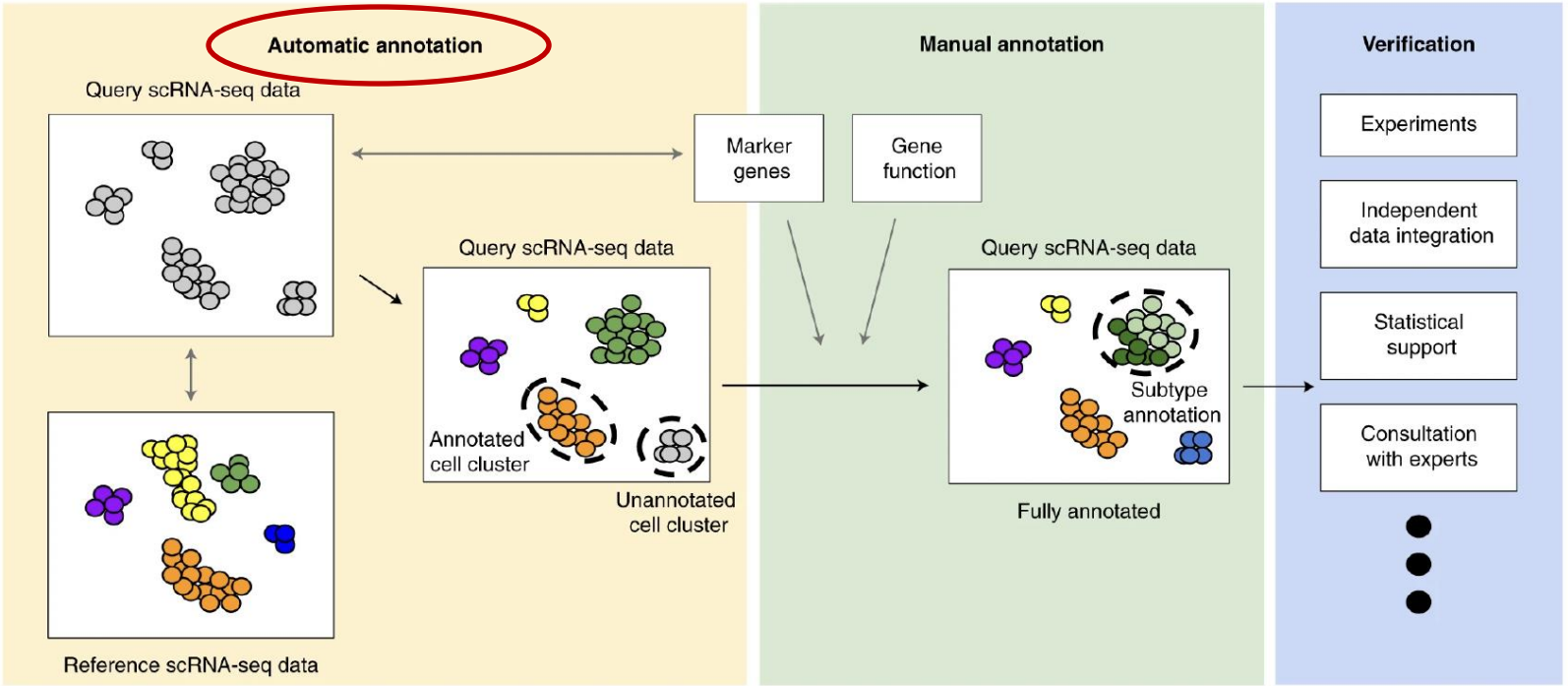
- Provides biological context for identified clusters.
- Helps **interpret cellular diversity** in tissues or developmental stages.
- Facilitates **comparative analysis** across conditions or diseases.

A proposed three step workflow



Clarke et al 2021 (Nature Protocols)

A proposed three step workflow



Where do I even start?

Clarke et al 2021 (Nature Protocols)

README Code of conduct MIT license

awesome-single-cell

List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc. Contributions welcome...

Citation

DOI 10.5281/zenodo.1117762

https://github.com/seandavi/awesome-single-cell

Non-exhaustive list of programs aimed at cell type identification

Cell type identification and classification

- ceLLama - [R/Python] - ceLLama is a streamlined automation pipeline for cell type annotations using local large-language models (LLMs).
- cellassign - [R] - Automated, probabilistic assignment of scRNA-seq to known types. `cellassign` automatically assigns single-cell RNA-seq data to known cell types across thousands of cells accounting for patient and batch specific effects. Information about a priori known markers for cell types is provided as input to the model. cellassign then probabilistically assigns each cell to a cell type, removing subjective biases from typical unsupervised clustering workflows. bioRxiv
- CHETAH - [R] - CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. CHETAH (CHaracterization of cEll Types Aided by Hierarchical clustering) is an accurate cell type identification algorithm that is rapid and selective, including the possibility of intermediate or unassigned categories. Evidence for assignment is based on a classification tree of previously available scRNA-seq reference data and includes a confidence score based on the variance in gene expression per cell type. For cell types represented in the reference data, CHETAH's accuracy is as good as existing methods. Its specificity is superior when cells of an unknown type are encountered, such as malignant cells in tumor samples which it pinpoints as intermediate or unassigned. bioRxiv
- CIPR - [R] - (Cluster Identity PRedictor-pronounced cy-per). A Shiny web applet (and R-package) that helps annotating the cluster identities in single-cell RNA-sequencing (scRNA-seq) experiments. The algorithm compares gene expression signature of experimental clusters with known reference datasets. In addition to 7 reference datasets implemented in CIPR (2 from mouse and 5 from human), users can upload custom high-throughput reference data for specialized studies. The CIPR pipeline can be further tailored to different analytical contexts by excluding irrelevant reference subsets and low-variance reference genes from the analysis. The manuscript describing CIPR and comparing its performance against other similar software was published in BMC Bioinformatics. CIPR's fast and computationally efficient calculations and graphical outputs will facilitate scRNA-seq analysis where the user wants to try different clustering parameters iteratively and examine the cluster identities. Source code for the Shiny and R-package implementations are available on GitHub.
- easybio - [R] - easybio is an R pacakge for cell type annotation using the CellMarker2.0 database. bioRxiv
- Garnett - [R] - Garnett is a software package that facilitates automated cell type classification from single-cell expression data. Garnett works by taking single-cell data, along with a cell type definition (marker) file, and training a regression-based classifier. Once a classifier is trained for a tissue/sample type, it can be applied to classify future datasets from similar tissues. In addition to describing training and classifying functions, this website aims to be a repository of previously trained classifiers. Supervised Classification Enables Rapid Annotation of Cell Atlases
- scANVI - [python] - single-cell ANnotation using Variational Inference (scANVI) is a semi-supervised variant of scVI designed to leverage any available cell state annotations — for instance when only one data set in a cohort is annotated, or when only a few cells in a single data set can be labeled using marker genes. Harmonization and Annotation of Single-cell Transcriptomics data with Deep Generative Models
- SignacX - [R] - Signac classifies the cellular phenotype for each individual cell in scRNA-seq data using neural networks trained with sorted bulk gene expression data from the Human Primary Cell Atlas. Signac can: map cells from one data set to another, classify non-human single cell data, identify novel cell types, and classify single cell data across many tissues, diseases and technologies. Cell type classification and discovery across diseases, technologies and tissues reveals conserved gene signatures and enables standardized single-cell readouts
- singleCellNet - [R] - A near-universal step in the analysis of single cell RNA-Seq data is to hypothesize the identity of each cell. Often, this is achieved by finding cells that express combinations of marker genes that had previously been implicated as being cell-type specific, an approach that is not quantitative and does not explicitly take advantage of other single cell RNA-Seq studies. SingleCellNet, which addresses these issues and enables the classification of query single cell RNA-Seq data in comparison to reference single cell RNA-Seq data. bioRxiv
- SingleR - [R] - SingleR leverages reference transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells independently. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology (2019)
- scCATCH - [R] - A single cell cluster-based annotation package from cluster marker genes identification to cluster annotation based on evidence-based score by matching the identified potential marker genes with known cell markers in tissue-specific cell taxonomy reference database (CellMatch) Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience (2020)
- DeepSort - [python] - A reference-free cell-type annotation tool for single-cell RNA-seq data using deep learning with a weighted graph neural network, which is learned based on the most comprehensive single-cell transcriptomics atlases involving 764,741 cells across 88 tissues of human and mouse. bioRxiv
- ImmClassifier - [R,python,Docker] - A cell type annotation algorithm that employs a knowledge-based approach to annotating cells based on their underlying ontology and multitudes of previously-published data. By encoding immune cell hierarchy in a neural network, ImmClassifier is able to identify fine-grained cell types with high accuracy. By running in Docker the tool is platform-agnostic. bioRxiv
- Celltypist - [Python] - Celltypist is an automated cell type annotation tool for scRNA-seq datasets on the basis of logistic regression classifiers optimized by the stochastic gradient descent algorithm. Celltypist provides several different models for predictions, with a current focus on immune sub-populations, in order to assist in the accurate classification of different cell types and subtypes.
- scPRINT - [python] - scPRINT is pretrained on 50M cells to predict multiple cell labels de novo, from any single cell RNAseq profile. scPRINT: pre-training on 50 million cells allows robust gene network predictions

Non-exhaustive list of programs aimed at cell type identification

Cell type identification and classification

- ceLLama - [R/Python] - ceLLama is a streamlined automation pipeline for cell type annotations using local large-language models (LLMs).
- cellassign - [R] - Automated, probabilistic assignment of scRNA-seq to known types. `cellassign` automatically assigns single-cell RNA-seq data to known cell types across thousands of cells accounting for patient and batch specific effects. Information about a priori known markers for cell types is provided as input to the model. cellassign then probabilistically assigns each cell to a cell type, removing subjective biases from typical unsupervised clustering workflows. bioRxiv
- CHETAH - [R] - CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. CHETAH (CHaracterization of cEll Types Aided by Hierarchical clustering) is an accurate cell type identification algorithm that is rapid and selective, including the possibility of intermediate or unassigned categories. Evidence for assignment is based on a classification tree of previously available scRNA-seq reference data and includes a confidence score based on the variance in gene expression per cell type. For cell types represented in the reference data, CHETAH's accuracy is as good as existing methods. Its specificity is superior when cells of an unknown type are encountered, such as malignant cells in tumor samples which it pinpoints as intermediate or unassigned. bioRxiv
- CIPR - [R] - (Cluster Identity PRedictor-pronounced cy-per). A Shiny web applet (and R-package) that helps annotating the cluster identities in single-cell RNA-sequencing (scRNA-seq) experiments. The algorithm compares gene expression signature of experimental clusters with known reference datasets. In addition to 7 reference datasets implemented in CIPR (2 from mouse and 5 from human), users can upload custom high-throughput reference data for specialized studies. The CIPR pipeline can be further tailored to different analytical contexts by excluding irrelevant reference subsets and low-variance reference genes from the analysis. The manuscript describing CIPR and comparing its performance against other similar software was published in BMC Bioinformatics. CIPR's fast and computationally efficient calculations and graphical outputs will facilitate scRNA-seq analysis where the user wants to try different clustering parameters iteratively and examine the cluster identities. Source code for the Shiny and R-package implementations are available on GitHub.
- easybio - [R] - easybio is an R pacakge for cell type annotation using the CellMarker2.0 database. bioRxiv
- Garnett - [R] - Garnett is a software package that facilitates automated cell type classification from single-cell expression data. Garnett works by taking single-cell data, along with a cell type definition (marker) file, and training a regression-based classifier. Once a classifier is trained for a tissue/sample type, it can be applied to classify future datasets from similar tissues. In addition to describing training and classifying functions, this website aims to be a repository of previously trained classifiers. Supervised Classification Enables Rapid Annotation of Cell Atlases
- scANVI - [python] - single-cell ANnotation using Variational Inference (scANVI) is a semi-supervised variant of scVI designed to leverage any available cell state annotations — for instance when only one data set in a cohort is annotated, or when only a few cells in a single data set can be labeled using marker genes. Harmonization and Annotation of Single-cell Transcriptomics data with Deep Generative Models
- SignacX - [R] - Signac classifies the cellular phenotype for each individual cell in scRNA-seq data using neural networks trained with sorted bulk gene expression data from the Human Primary Cell Atlas. Signac can: map cells from one data set to another, classify non-human single cell data, identify novel cell types, and classify single cell data across many tissues, diseases and technologies. Cell type classification and discovery across diseases, technologies and tissues reveals conserved gene signatures and enables standardized single-cell readouts
- singleCellNet - [R] - A near-universal step in the analysis of single cell RNA-Seq data is to hypothesize the identity of each cell. Often, this is achieved by finding cells that express combinations of marker genes that had previously been implicated as being cell-type specific, an approach that is not quantitative and does not explicitly take advantage of other single cell RNA-Seq studies. SingleCellNet, which addresses these issues and enables the classification of query single cell RNA-Seq data in comparison to reference single cell RNA-Seq data. bioRxiv
- SingleR - [R] - SingleR leverages reference transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells independently. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology (2019)
- scCATCH - [R] - A single cell cluster-based annotation package from cluster marker genes identification to cluster annotation based on evidence-based score by matching the identified potential marker genes with known cell markers in tissue-specific cell taxonomy reference database (CellMatch) Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience (2020)
- DeepSort - [python] - A reference-free cell-type annotation tool for single-cell RNA-seq data using deep learning with a weighted graph neural network, which is learned based on the most comprehensive single-cell transcriptomics atlases involving 764,741 cells across 88 tissues of human and mouse. bioRxiv
- ImmClassifier - [R,python,Docker] - A cell type annotation algorithm that employs a knowledge-based approach to annotating cells based on their underlying ontology and multitudes of previously-published data. By encoding immune cell hierarchy in a neural network, ImmClassifier is able to identify fine-grained cell types with high accuracy. By running in Docker the tool is platform-agnostic. bioRxiv
- Celltypist - [Python] - Celltypist is an automated cell type annotation tool for scRNA-seq datasets on the basis of logistic regression classifiers optimized by the stochastic gradient descent algorithm. Celltypist provides several different models for predictions, with a current focus on immune sub-populations, in order to assist in the accurate classification of different cell types and subtypes.
- scPRINT - [python] - scPRINT is pretrained on 50M cells to predict multiple cell labels de novo, from any single cell RNAseq profile. scPRINT: pre-training on 50 million cells allows robust gene network predictions

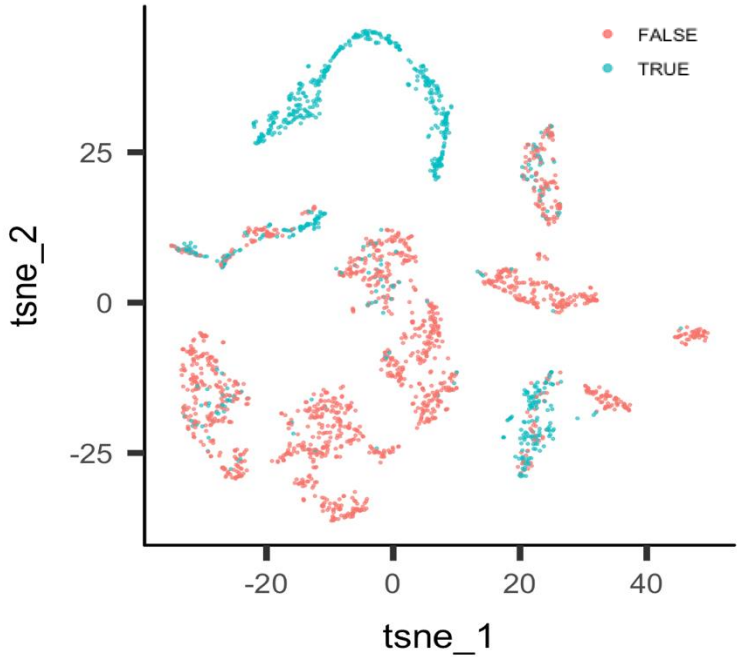
What makes this a hard problem?

Single-cell data is **sparse.**

Many “specific” marker genes aren’t at least at whole animal scale.

Some highly expressed genes are “leaky” (ambient RNA).

Albumin expressed?



The Tabula Muris Consortium, Nature, 2018

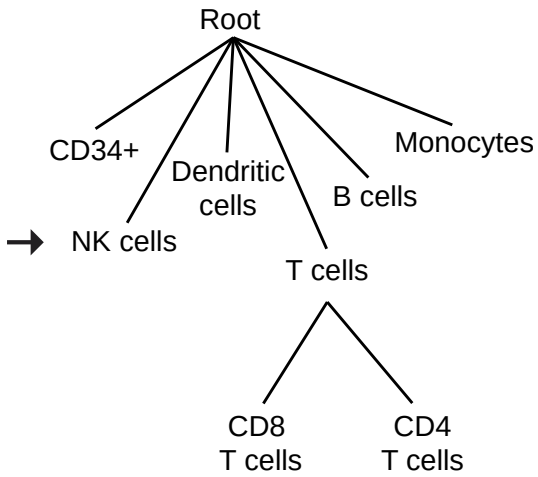
Slide from Hanah Pliner

The Garnett marker file defines a hierarchy of cell types

Define cell markers

```
>CD34+
expressed: CD34, THY1, ENG, KIT, PROM1

>NK cells
expressed: NCAM1, FCGR3A

>Monocytes
expressed: CD14, FCGR1A, CD68, S100A12

>B cells
expressed: CD19, MS4A1, CD79A

>T cells
expressed: CD3D, CD3E, CD3G

>CD4 T cells
expressed: CD4, FOXP3, IL2RA, IL7R
subtype of: T cells

>CD8 T cells
expressed: CD8A, CD8B
subtype of: T cells

>Dendritic cells
expressed: IL3RA, CD1C, BATF3, THBD,
CD209
```

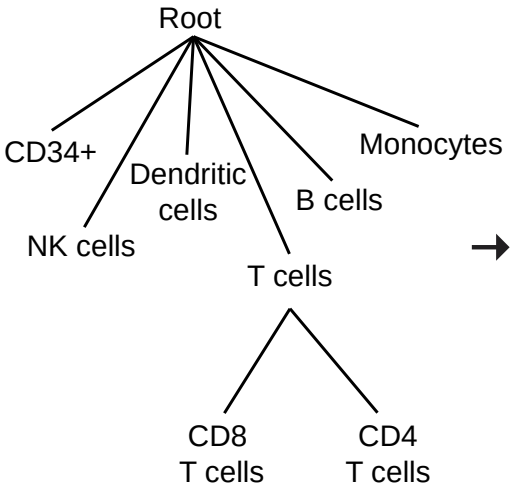
Generate cell type hierarchy



Slide from Hanah Pliner

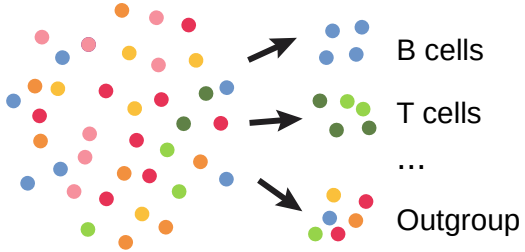
Garnett trains a multi-level classifier to (sub)type cells

Generate cell type hierarchy

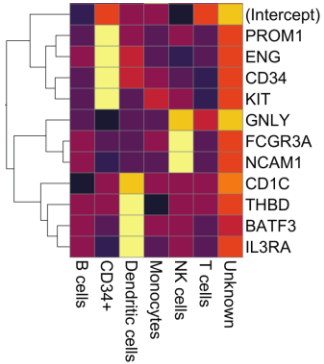


Train at each node:

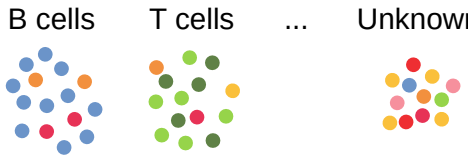
1. Find representative cells for child nodes using markers



2. Train multinomial classifier (elastic net regression)



3. Classify cells at permissive threshold



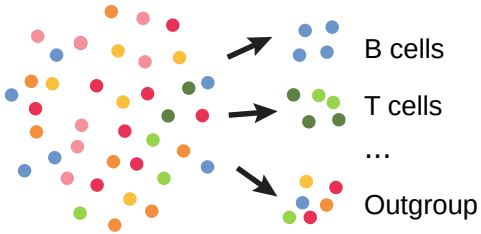
4. Repeat for nodes with further children

Slide from Hanah Pliner

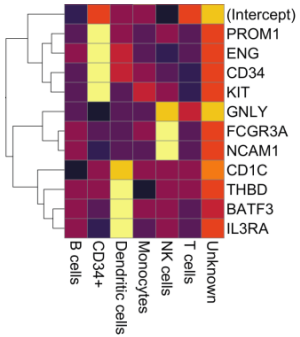
The classifier can be used on new datasets

Train at each node:

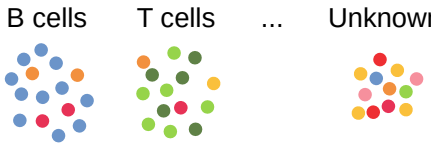
1. Find representative cells for child nodes using markers



2. Train multinomial classifier (elastic net regression)

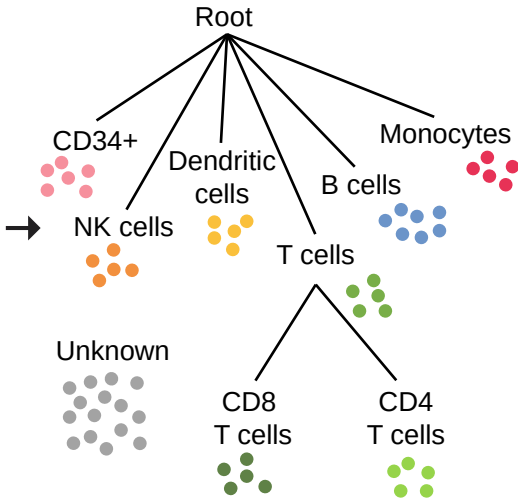


3. Classify cells at permissive threshold



4. Repeat for nodes with further children

Hierarchically classify cells at strict threshold

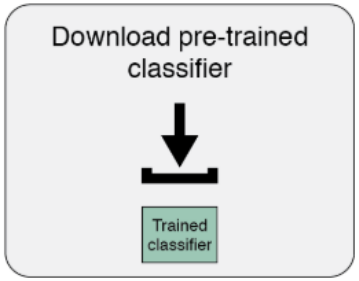


Slide from Hanah Pliner

Pre-trained classifiers

Garnett | Documentation | Publications | Pre-trained Classifiers | GitHub | Interactive

Currently available pre-trained classifiers:

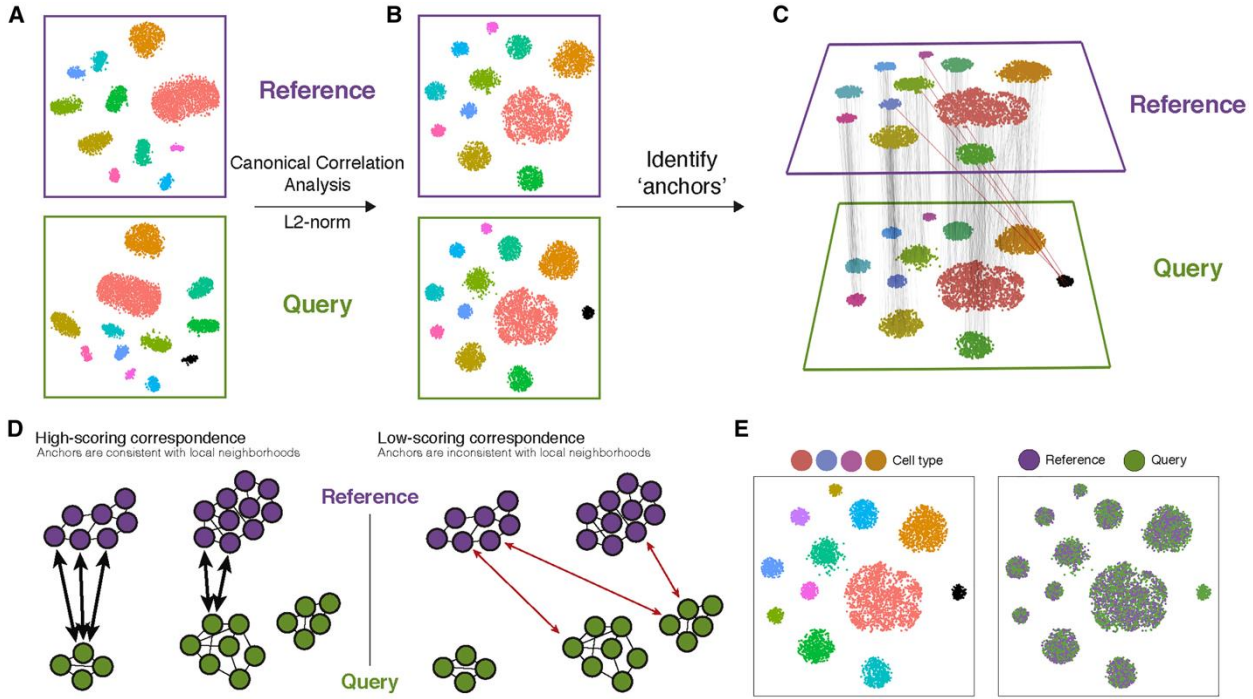


Classifier	Marker file	Species	Tissue	Contributer	Training data source	Publication
hsLung	hsLung_markers.txt	Human	Lung	Hannah Pliner	Lambrechts et. al.	Pliner et. al.
hsPBMC	hsPBMC_markers.txt	Human	PBMC	Hannah Pliner	10x Genomics	Pliner et. al.
mmLung	mmLung_markers.txt	Mouse	Lung	Hannah Pliner	Han et. al.	Pliner et. al.
ceWhole	ceWhole_markers.txt	C. elegans	Whole	Hannah Pliner	Cao et. al.	Pliner et. al.

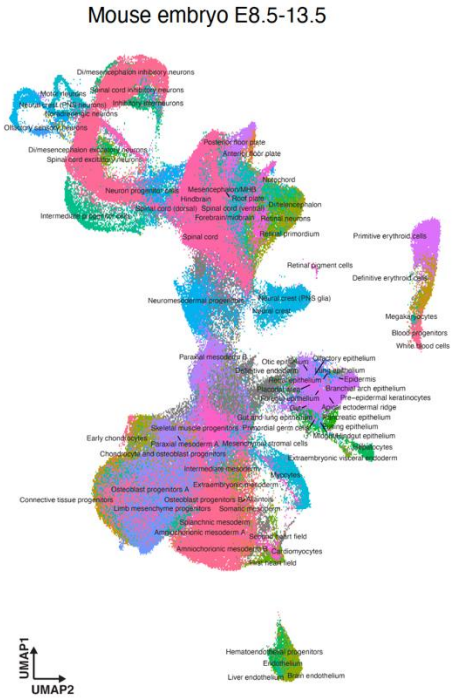
Help build this resource! See here, to learn how to submit your own classifier!

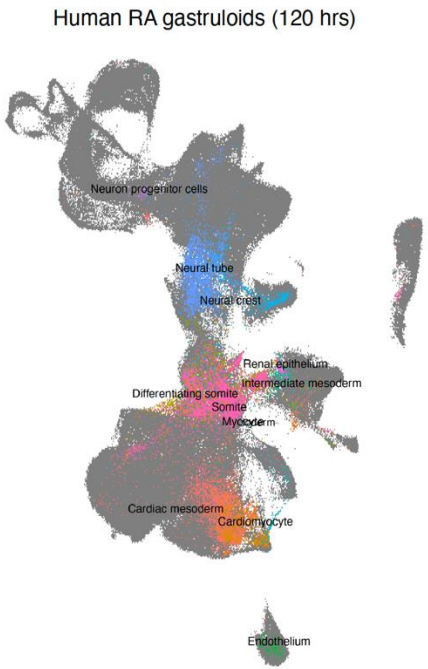
Slide from Hanah Pliner

Integration & Label Transfer



An integration example





* Integration by Seurat V3 (Hamazaki, Yang et al., 2024)

Web Tools

- Reference-based web tools.
- Minimal requirement for programming.
- Available to a limited set of organisms and organs.

	Azimuth	Tabula Sapiens	MapmyCells
Species	Human(11) and mouse (1)	Human	Human and mouse
Organs	PBMC, Motor cortex, pancreas, fetal, kidney, bone marrow, lung, adipose, tonsil, heart, liver	24 organs	Whole brain and middle temporal gyrus
Assay	RNA(12) and ATAC(2)	RNA	RNA
Interface	Web portal, Seurat	Google collab	Web portal
Link	https://azimuth.hubmapconsortium.org/	https://tabula-sapiens.sf.czbiohub.org/annotateuserdata	https://knowledge.brain-map.org/mapmycells/process/

Slide from Wei Yang

Automatic annotation - summary

• A well-annotated reference is required for automatic annotation.

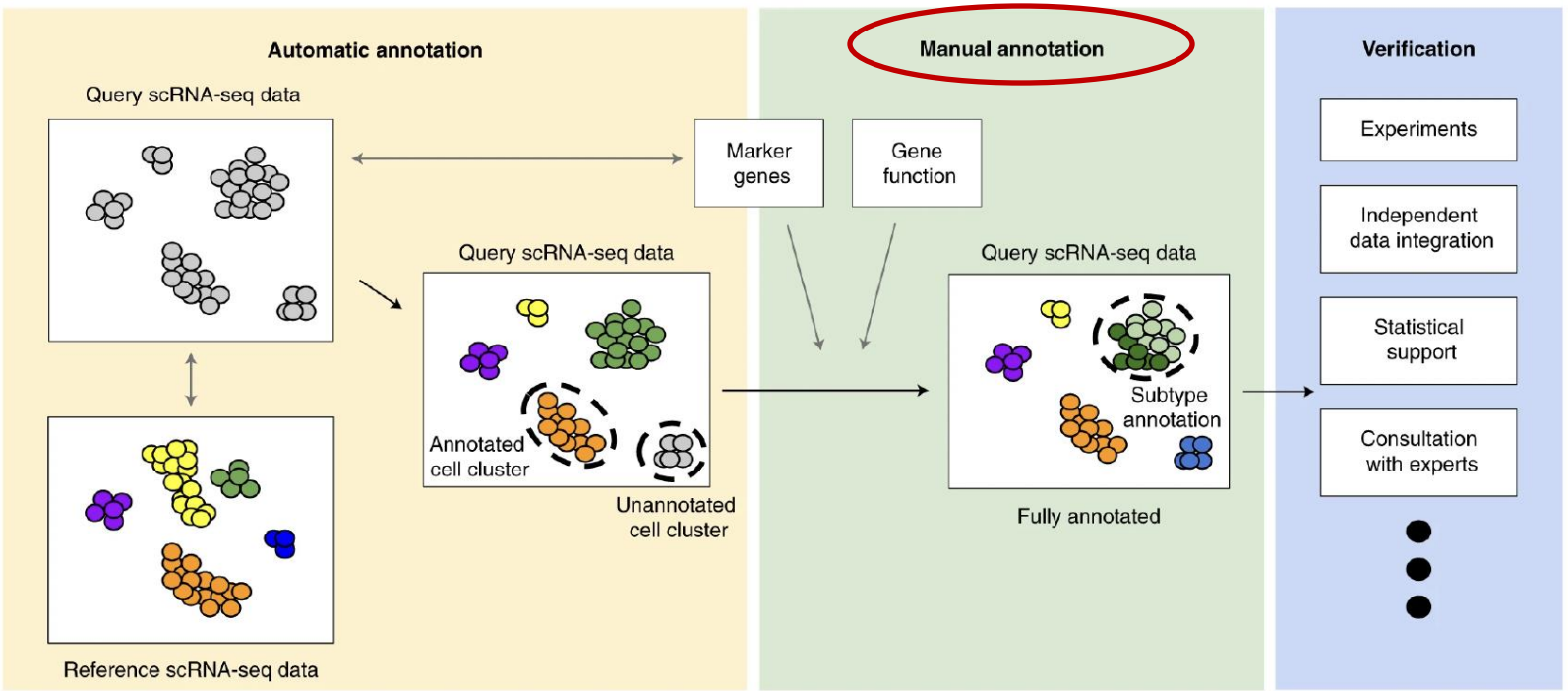
• Many algorithms are available for integration and label transfer. The main consideration is computational resources (e.g. run time, memory).

• The quality of integration may be different between same-species and cross-species integration.

• Web-tools are less computational demanding options but have limited sets of reference for now.

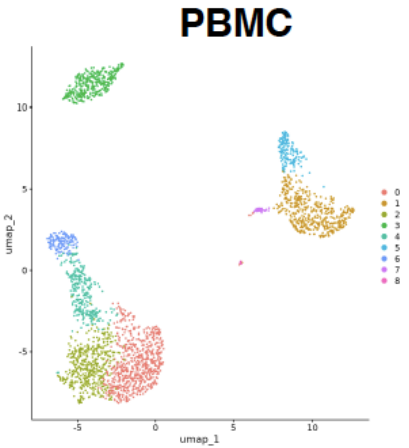
Slide from Wei Yang

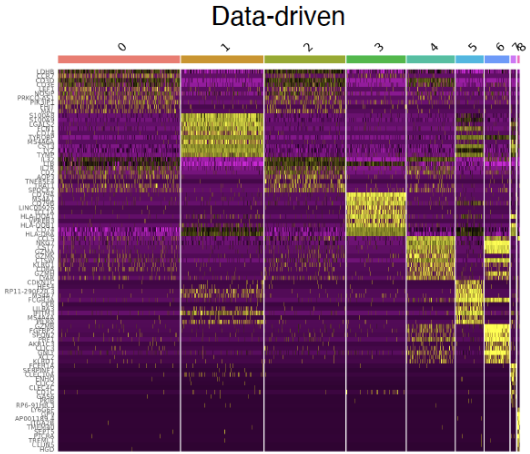
A proposed three step workflow

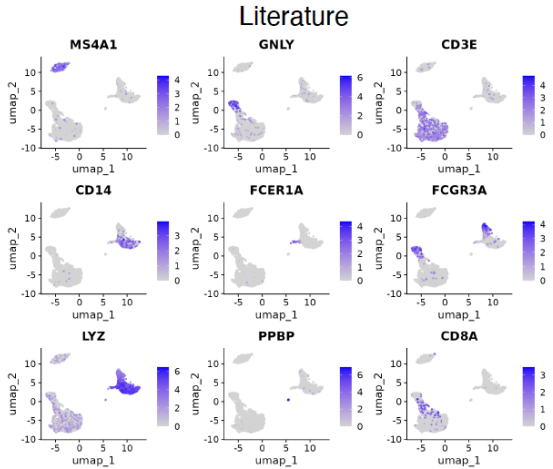


Clarke et al 2021 (Nature Protocols)

Manual Annotation – Marker Genes



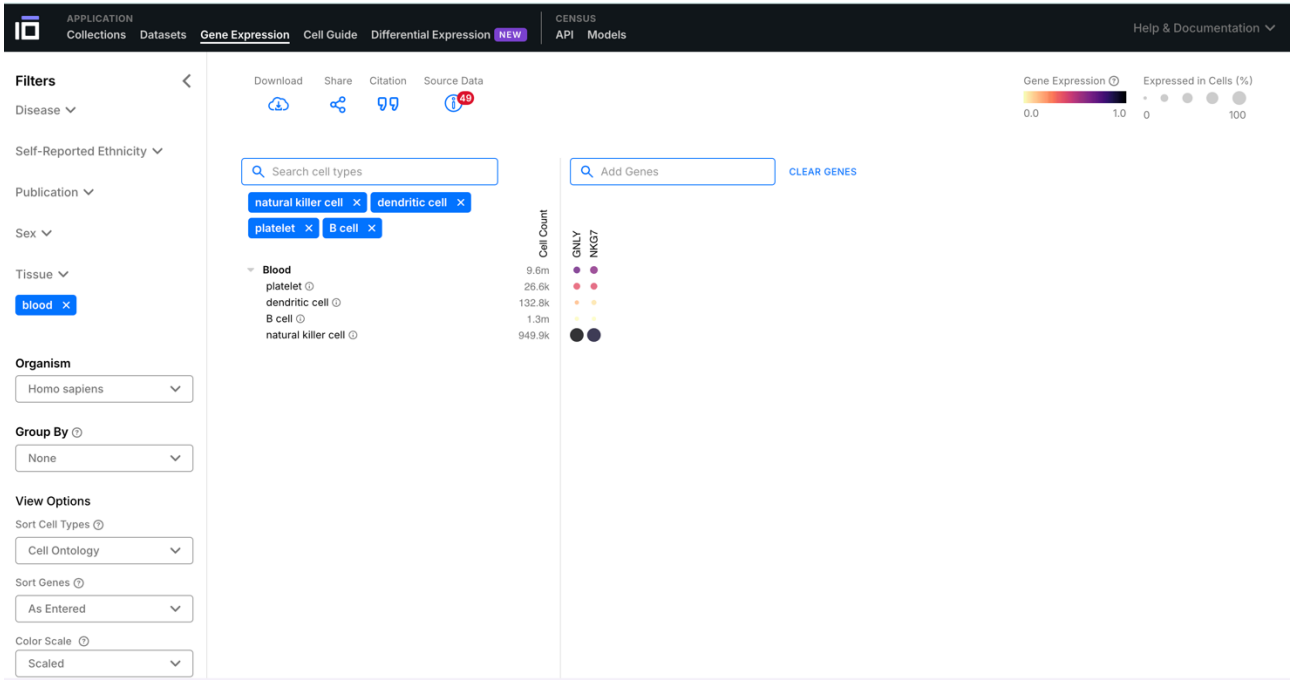




Slide from Wei Yang

Manual Annotation – Web Tools

https://cellxgene.cziscience.com/gene-expression



Slide from Wei Yang

Manual Annotation – Web Tools

https://panglaodb.se/index.html

Database statistics		
	Mus musculus	*Homo sapiens*
Samples	1063	305
Tissues	184	74
Cells	4,459,768	1,126,580
Clusters	8,651	1,748

PanglaoDB Home Search Datasets Tools Papers FAQ/Help About

Filter

Show cell type: NK cells (98) get tsv file add marker

Gene expression markers for NK cells

Vote(s)	Species	Official gene symbol	UI	Sensitivity (human)	Sensitivity (mouse)	Specificity (human)	Specificity (mouse)	Marker count	Cell type	Germ layer	Organ	Aliases	Product description	Disease	Action
19	Mm Hs	TRDC	0.011	0.351	NA	0.018	NA	4	NK cells	Mesoderm	Immune system		T cell receptor delta constant		flag
8	Mm Hs	NKG7	0.056	0.946	1	0.088	0.04	3	NK cells	Mesoderm	Immune system	GMP-17	natural killer cell granule protein 7		flag
5	Hs	KLRF1	0.003	0.257	NA	0.015	NA	2	NK cells	Mesoderm	Immune system	CLEC5C,NKp80	killer cell lectin like receptor F1		flag
3	Mm Hs	KLRD1	0.02	0.595	0.915	0.026	0.011	3	NK cells	Mesoderm	Immune system	CD94	killer cell lectin like receptor D1		flag
3	Hs	GNLY	0.013	0.851	NA	0.064	NA	2	NK cells	Mesoderm	Immune system	NKG5,LAG-2,D2S69E,TLA519,LAG2	granulysin		flag
3	Mm Hs	NCR1	0.004	NA	0.798	NA	0	1	NK cells	Mesoderm	Immune system	NK-p46,NKP46,CD335,LY94	natural cytotoxicity triggering receptor 1		flag
2	Mm Hs	GZMA	0.026	0.905	0.787	0.046	0.014	3	NK cells	Mesoderm	Immune system	HFSP,CTLA3	granzyme A		flag
1	Mm Hs	HOPX	0.033	0.514	NA	0.082	0.022	10	NK cells	Mesoderm	Immune system	LAGY,OB1,NECC1,SMAP31	HOP homeobox		flag
1	Mm Hs	ITGAM	0.025	NA	0.011	0.003	0.03	8	NK cells	Mesoderm	Immune system	CD11b,CR3A,CD11B	integrin subunit alpha M	Y	flag
1	Mm Hs	TGFB1	0.11	0.122	NA	0.043	NA	3	NK cells	Mesoderm	Immune system	CED,TGFbeta,TGFB,DPD1	transforming growth factor beta 1	Y	flag
1	Mm Hs	GZMB	0.017	0.514	0.67	0.05	0.004	7	NK cells	Mesoderm	Immune system	CCPI,CGL-1,CSP-B,CGL1,CTSGL1,SECT,CSPB	granzyme B		flag
1	Mm	KLRE1	0.005	NA	0.862	NA	0.001	1	NK cells	Mesoderm	Immune system		killer cell lectin-like receptor family E member 1		flag

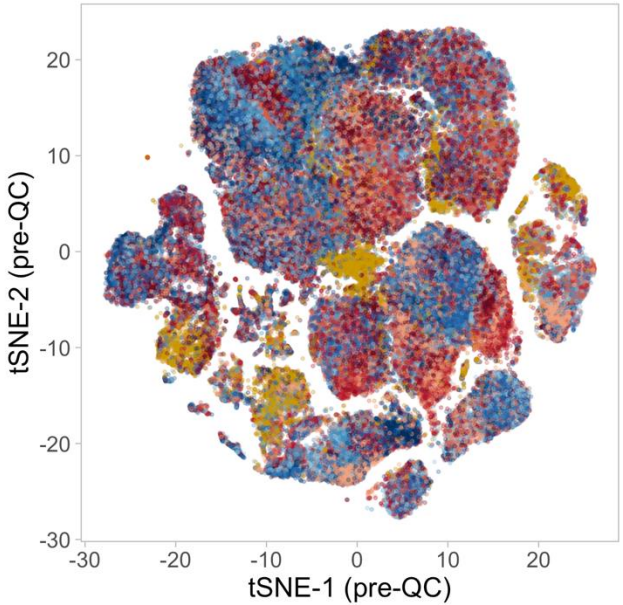
Slide from Wei Yang

Manual annotation - summary

- In manual annotation, we are looking at marker genes that could distinguish a cluster from the rest.
- Marker genes can be identified in a data-driven way. This is very useful for de novo discovery of de novo annotation of cell types/states. However, some marker genes could be introduced by technical factors (e.g. ribosomal and histone).
- Maker genes can also be identified from literature. However, the expression level of such marker genes might be different in different studies (i.e. profiling methods, organisms, locations).
- Web tools are available to explore gene expression in a large number of cell types. The annotation of these webtools stay at a broad level.

Slide from Wei Yang

Sidenote 1: doublets often form clusters



What are Doublets

- Doublets occur when two cells are captured in the same droplet or well during library preparation, resulting in mixed transcriptional profiles.
- Doublets can mimic rare cell types or artifacts, leading to false biological conclusions.

Sources of Doublets

- *Technical Doublets:* Occur during droplet or microwell encapsulation (common in droplet-based methods).
- *Biological Doublets:* Result from cell-cell interactions (e.g., syncytia, aggregates).

Impact on Data Analysis

- Misinterpretation of clustering and differential expression analysis.
- Overestimation of cell population heterogeneity.

Some Doublet Detection Software

Doublet Detecting Software	QC Filtering Required	Requires Pre-clustering	Doublet Detecting Method
DoubletDecon	✗	✓	Deconvolution based on clusters provided.
DoubletDetection	✗	✗	Iterative boost classifier to classify doublets.
DoubletFinder	✓	✗	Identify ideal cluster size and call expected number of droplets with highest number of simulated doublet neighbors as doublets.
scDblFinder	✗	✗	Gradient boosted trees trained with number neighboring doublets and QC metrics to classify doublets
Scds	✗	✗	**cxds**: Uses genes pairs that are typically not expressed in the same droplet to rank droplets based on co-expression of all pairs. **bcds**: Uses highly variable genes and simulated doublets to train a binary classification algorithm and return probability of droplet being a doublet.
Scrublet	✗	✗	Identifies the number of neighboring simulated doublets for each droplet and uses bimodal distribution of scores to classify singlets and doublets.
Solo	✗	✗	Simulates doublets and fits a two-layer neural network.

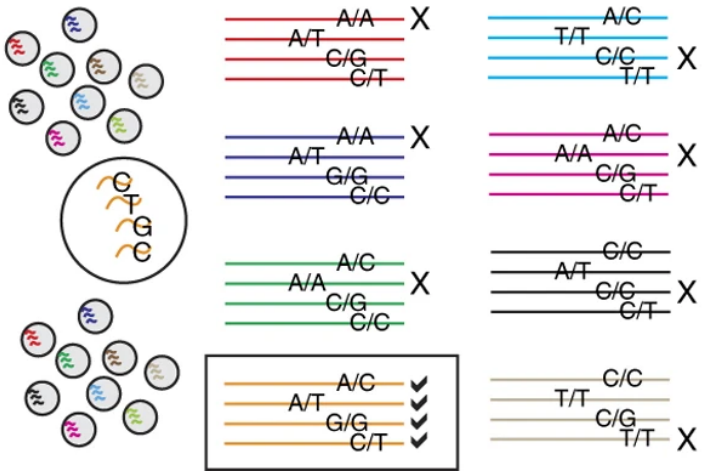
Source: Demuxafy; https://demultiplexing-doublet-detecting-docs.readthedocs.io/en/latest/DoubletDetectingSoftwares.html

Sidenote 2: multiplexing can help identify doublets

Method	Applications	Advantages	Limitations
Genetic Demultiplexing	Multi-donor datasets	No extra reagents needed	Requires genetic variation
Hashing with Antibodies	General purpose	Cost-effective for many samples	Antibody spillover potential
Hashing with Oligos	High-throughput datasets	Efficient and robust	Requires optimized protocols
Barnyard (Mixed Species)	Cross-species studies	Simple and robust	Limited to interspecies mixtures

Genetic demultiplexing

Pooling individuals with known genotypes allows demultiplexing cells based on genetics



~ 92% of doublets can be identified based on genetics

(12 individuals per library)

Kang et al. 2018 (Nature Biotechnology)

Tutorial

Background on the datasets

10X Genomics, PBMCs

Cell Ranger · pbmc_1k_v3 · Peripheral blood mononuclear cells (PBMCs) from a healthy donor

SUMMARY ANALYSIS

Estimated Number of Cells

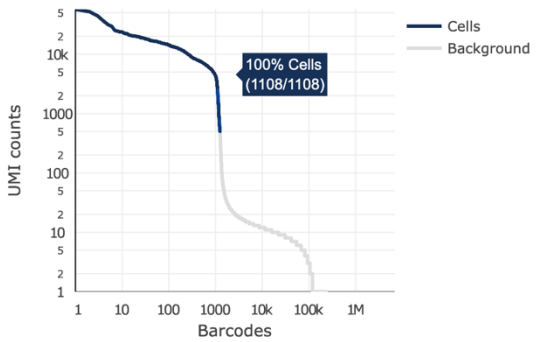
1,222

Mean Reads per Cell	Median Genes per Cell
54,502	1,919

Sequencing

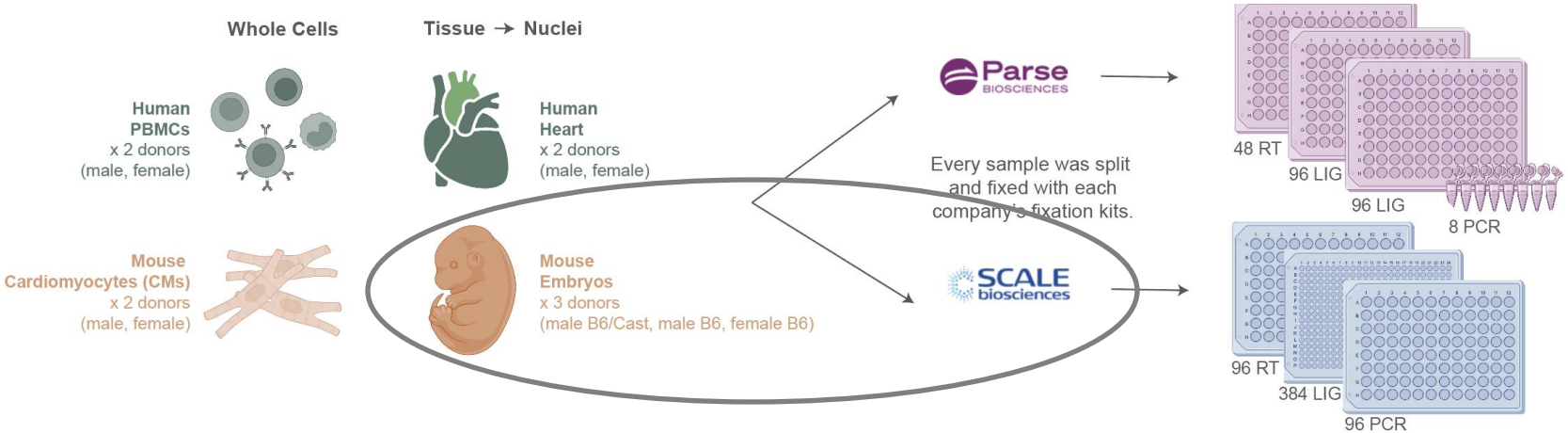
Number of Reads	66,601,887
Valid Barcodes	97.4%
Sequencing Saturation	70.8%
Q30 Bases in Barcode	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in Sample Index	91.1%
Q30 Bases in UMI	92.7%

Cells



Estimated Number of Cells	1,222
Fraction Reads in Cells	94.9%
Mean Reads per Cell	54,502
Median Genes per Cell	1,919
Total Genes Detected	18,391
Median UMI Counts per Cell	6,628

Combinatorial-indexing benchmarking (generated *in-house*)



Tutorial: mixed sample of two mouse embryos from different genetic backgrounds that have been genetically demultiplexed.