A Python Pipeline for scRNA-seq

Created by the Data and Analytic Research Environment (DARE) working group

Start Using the Discovery Environment Now!

de.cyverse.org

DATA & ANALYSIS

- Free data storage
- 100s open-source scientific apps
- Containers and notebooks
- Visualize & interact with data

YOUR WORKSPACE

- Manage and share data
- Perform analysis with your own or community datasets
- Do large-scale science from your web browser
- Build and customize apps

COLLABORATION

- Secure, shared workspace for your team
- Reproducibility
- Manage the data lifecycle
- Make data more FAIR
- Open science

LEARN, TEACH & TRAIN

- Tutorials and documentation
- Webinars
- Workshops
- Teach using CyVerse

COMMUNITY

- Join 95K+ users
- In-app chat support
- Find publicly available data
- Share data and analyses
- Deploy your own CyVerse

Sign Up

user.cyverse.org

Learn More

www.cyverse.org

CYVERSE® The Open Science Workspace for Collaborative, Data-driven Discovery

University of Arizona Data and Analytic Research Environment (DARE) Working Group

Mission

- Develop a user-friendly computing and analysis environment to facilitate cross-collaborative research

- Support current extramurally funded research that have highly multivariate and complex data integration

- Develop novel methods for analysis and computing – to gain prominence and demonstrate expertise nationally and internationally.

Funding: UAHS 4.1 (Personalized Defense); UAHS 5.3 (Health Analytics Powerhouse); Precision Aging Network (U19 AG065169)

Concept of Team Science WorkSpace

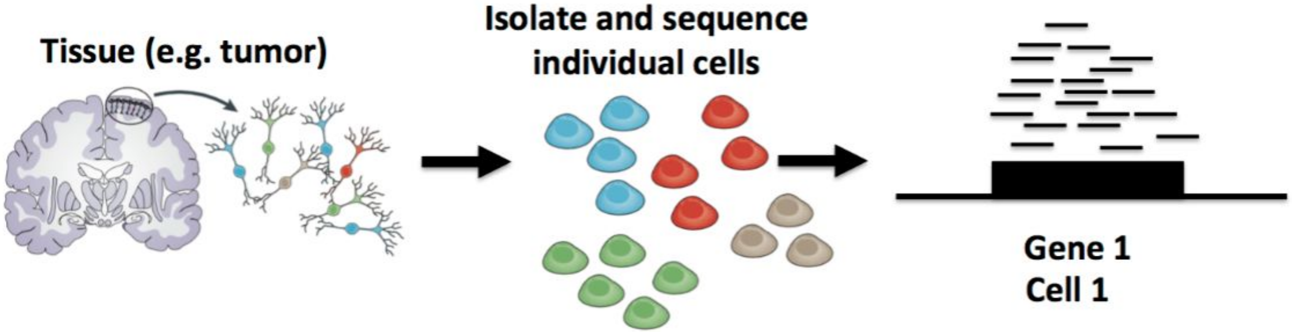
- Use GitBook for project documentation

- Secure large scale data storage and retrieval built on the CyVerse Data Store and Data Commons (a distributed and federated data grid)

- Reproducible computation and data analysis, visualization, and reproducible tools built in Rstudio, RShiny, and Jupyter notebooks through the CyVerse Discovery Environment (DE)

- Containerized workflows based on Docker and Singularity, with features that allow teams to share and collaborate harnessing a powerful computer infrastructure

Single-cell RNA-Seq (scRNA-Seq)



Tissue (e.g. tumor)

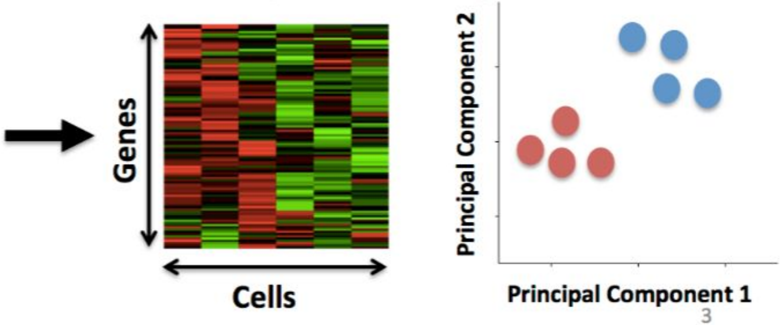
Isolate and sequence individual cells

Gene 1
Cell 1

Read Counts

	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			

Compare gene expression profiles of single cells



https://speakerdeck.com/stephaniehicks/welcome-to-the-world-of-single-cell-rna-sequencing?slide=3

Quality Control and Normalization

ScRNA-Seq data provides insight into cellular processes at a higher resolution than traditional bulk RNA sequencing. However, this higher resolution comes with the cost of increased noise and bias. Therefore, quality control and normalization are critical steps for a rigorous analysis. Common examples of quality control and normalization include:

- Total counts (sequencing depth)
- Number of genes
- Ribosomal fraction per cell
- Mitochondrial fraction per cell
- Predicted doublets
- Cell level normalization
- Gene level normalization

Predicted Doublets

ScRNA-Seq can produce multiplets, where two or more cells receive the same barcode. This can account for several percent of the transcriptomes and confound the downstream analysis. Multiplets can

- Appear as distinct cell types
- Bridge cell states
- Interfere with differential gene expression
- Interfere with gene regulatory networks

Scrublet (included in scanpy) is a method to predict and remove multiplets.

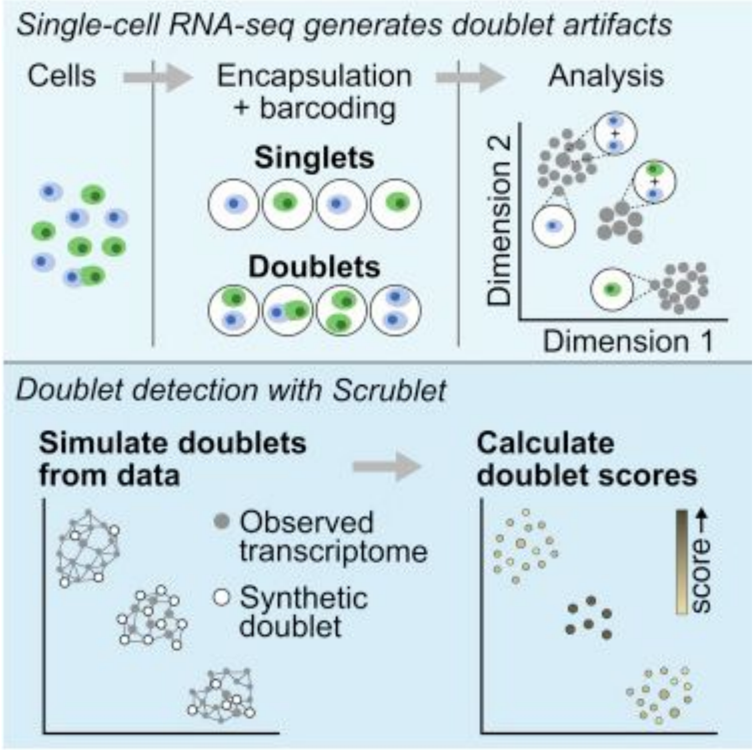
Scrublet

Scrublet is a computation library of statistical methods used for detecting doublets. Scrublet utilizes a two step process for detecting doublets

1. Creating synthetic doublet observations

2. Measuring the relatedness between the true cells and the synthetic doublets

Cells that are more related to the synthetic doublets are removed as part of the quality control.



Single-cell RNA-seq generates doublet artifacts
Cells
Encapsulation + barcoding
Analysis
Singlets
Doublets
Dimension 2
Dimension 1
Doublet detection with Scrublet
Simulate doublets from data
Calculate doublet scores
Observed transcriptome
Synthetic doublet
score

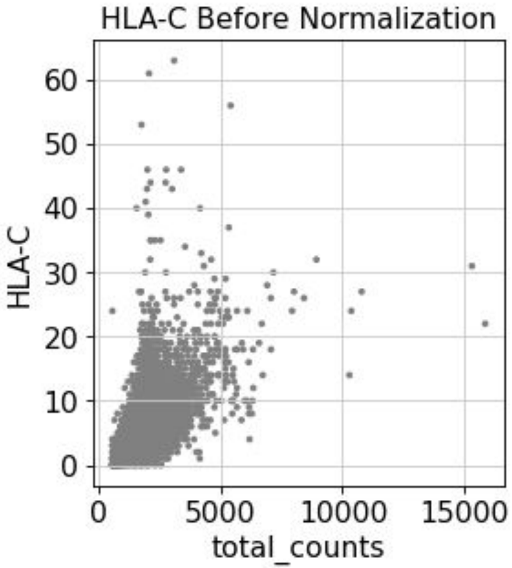
Wolock, S. L., Lopez, R., & Klein, A. M. (2019).

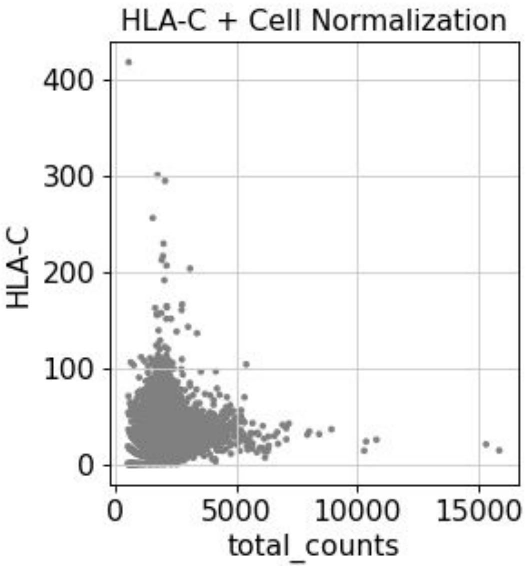
Normalization

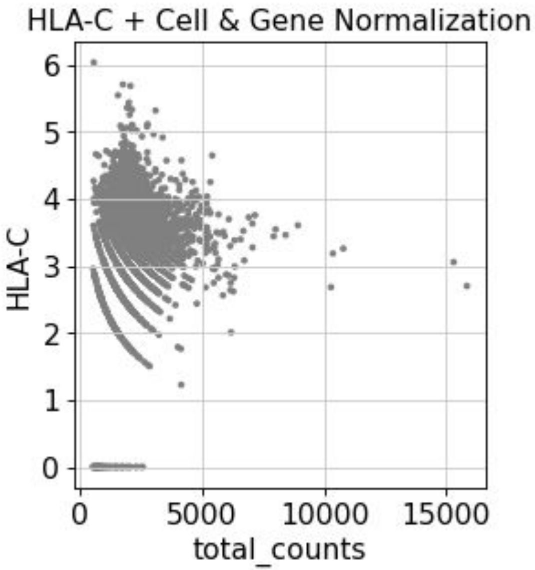
Normalization can be broken into two categories.

1. **Cell level normalization**
   - **a.** Gene expression can be linearly dependent on sequencing depth, causing unwanted variation.
   - **b.** To mediate this effect, we divide gene expression by the sequencing depth and multiply by 10,000. Therefore, each cell is normalized to a library size of 10,000.
2. **Gene level normalization**
   - a. Outliers in gene expression are common in scRNA-seq data. These outliers can cause issues in downstream analysis.
   - b. We add 1 to all gene expressions to mediate outliers and then take the natural logarithm.

Normalization



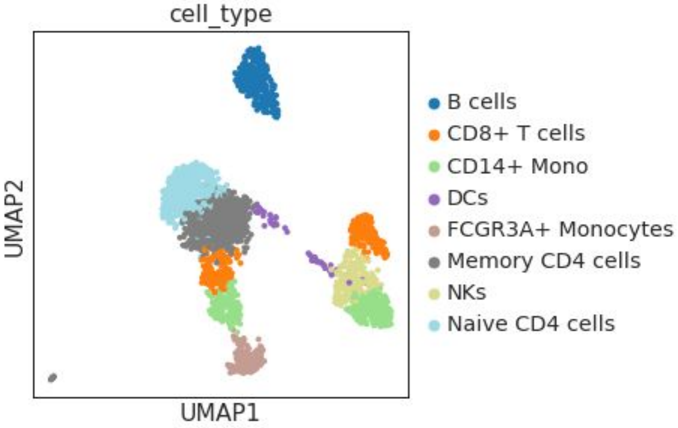




Clustering

Cell types may be identified manually using alternative methods to unsupervised clustering. However, these methods can be expensive, require expert knowledge, are not feasible for rare cell types, or are slow for extensive scRNA-seq methods. Therefore, researchers often turn to unsupervised clustering methods to identify cell subpopulations. Clustering of scRNA-seq data are divided into two crucial steps

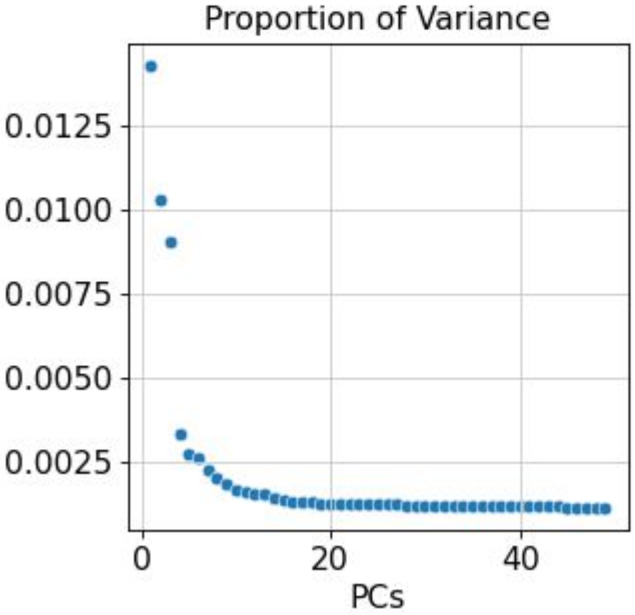
1. **Dimensionality reduction**
2. **Clustering**



Dimensionality Reduction (Multivariate)

More is not always better! High dimensional data suffer from the “curse of dimensionality.” Researchers must use statistical methods to reduce the number of dimensions used in the clustering step. Common dimensionality reduction techniques include

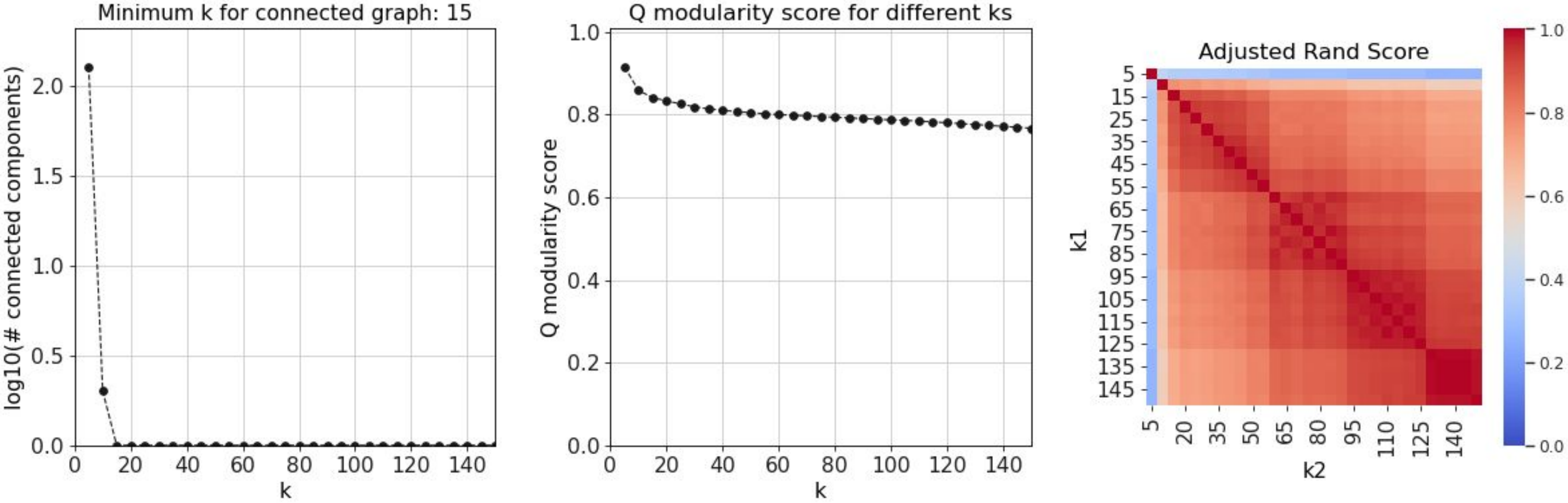
- Principal component analysis (PCA)
- Single value decomposition (SVD)
- t-SNE
- UMAP



Clustering

Many clustering algorithms exist for scRNA-seq data. Graph-based methods are fast and scalable with relatively few assumptions. Scanpy includes PhenoGraph, which combines k-nearest neighbors and Louvain community detection to cluster cells.

- Do not need to specify the number of clusters
- PhenoGraph only has one parameter (k)
- The optimal k is data-dependent
- Choosing the optimal parameter is challenging


Minimum k for connected graph: 15
log10(# connected components)
0.0
0.5
1.0
1.5
2.0
k
0
20
40
60
80
100
120
140
Q modularity score for different ks
Q modularity score
0.0
0.2
0.4
0.6
0.8
1.0
k
0
20
40
60
80
100
120
140
Adjusted Rand Score
k1
5
15
25
35
45
55
65
75
85
95
105
115
125
135
145
k2
5
20
35
50
65
80
95
110
125
140
1.0
0.8
0.6
0.4
0.2
0.0

Wrap up

ScRNA-seq can provide transcriptional insights at single-cell resolution. Our CyVerse ShunPykeR app offers researchers the tools needed to streamline their analysis. Our pipeline includes critical analysis steps, including

- Quality control
- Normalization
- Dimensionality reduction
- Clustering

Future directions

Currently, our pipeline featuring scanpy contains the most up-to-date and widely used analysis methods for scRNA-seq data. Our team is developing novel approaches to tackle challenging computational and analytical problems within this field. Our cutting edge research includes

- Benchmarking of current methods
- Parameter tuning
- Multi-omics data integration

References

1. Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, *8*(4), 281-291.