Benchmarking Single-cell RNA-Seq Pipelines

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

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); 18th Mile TRIF

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

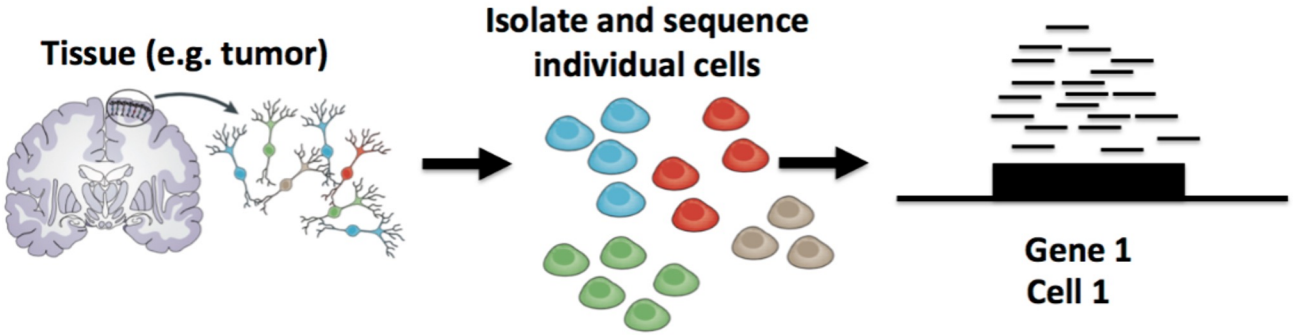
Learn More
www.cyverse.org

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

Concept of Team Science WorkSpace

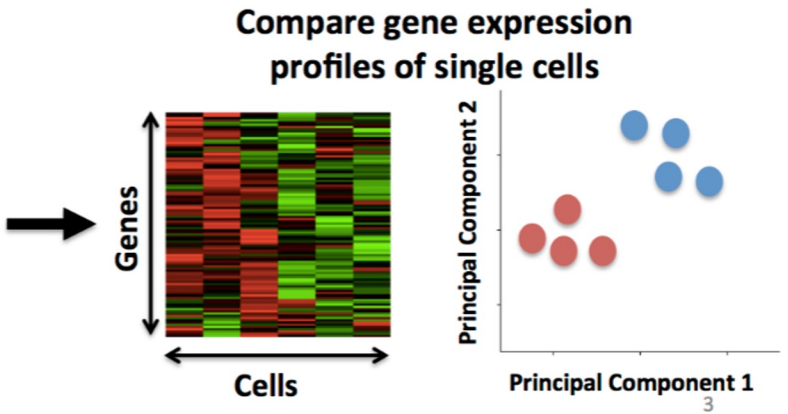
- Use GitBook for project documentation
- Secure large scale data storage and retrieval built on the CyVerse Data Store and Data Commons (a distributed and federated data grid)
- Reproducible computation and data analysis, visualization, and reproducible tools built in Rstudio, RShiny, and Jupyter notebooks 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)



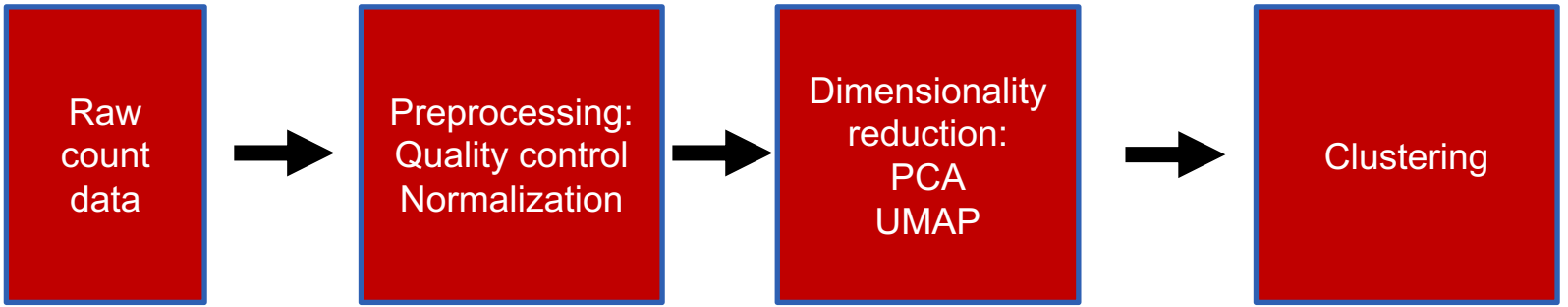
Read Counts

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



https://www.rna-seqblog.com/

scRNA-seq Pipelines



Examples: **Seurat** (R); Scanpy (Python)

Benchmarking And User-Defined Parameters

- User-defined parameters: Parameters in an algorithm that can be changed by the user and can affect the overall results of the algorithm.

- Benchmarking: To help select "smart" user-defined parameters or for new method development as a way of testing individual elements of an algorithm as distinct from the whole.

Run unsupervised clustering analysis using PhenoGraph

Next, we will filter out cells that we think are either dying cells or empty droplets. Both are characterized by low library size, dying cells additionally have relatively high percentages of mitochondrial RNA. Rather than using cutoffs, we remove cells by cluster. This prevents us from taking out cells that belong to a biologically relevant cluster, despite having e.g. low library size or high mitochrondrial RNA. It furthermore ensures that we take out all cells with a phenotype similar to what we think are cells that should be taken out, even if they happen not to exceed a possible cutoff we would have otherwise chosen. We can furthermore remove clusters that we are not interested in, such as doublets and contaminants.

For clustering, we use PhenoGraph. This is a highly robust graph-based clustering algorithm that was designed for single cell data. Your choice of k (number of nearest neighbors to use for graph construction) can affect the number of clusters and their size.

❗ How to select a number for neighbour cells (k) for fairly robust clustering.

Calculate clustering characteristics for a range of ks (5 to 155), by a step of 5.

Inspect the following clustering metrics:

- Minimum number of k for a connected graph.
- Changes in Q-modularity score (see PhenoGraph paper) for different ks. The Q-modularity usually drops quickly in the beginning (due to swiftly increasing connectedness when increasing a low k), and only slightly decreases after a certain point. However, in the case of homogeneous data with little structure, you might see a constant rapid decay in Q score. The plot below is not needed for a good choice of k, but is informative about the modularity of your data.
- Similarity/difference in cluster assignments between the different ks. We will use the Rand index to this end. The Rand index algorithm compares to clusterings by testing for every pair of cells, whether or not they were both clustered together or separately in both clusterings. E.g. if two cells were not in the same cluster in clustering 1, but were clustered together in clustering 2, this will decrease the Rand score. Alternatively, if they were in separate clusters in both cases, or in the same cluster in both cases, this will increase the score. The Rand index ranges from 0 to 1, with 1 indicating identical cluster assignments. We will calculate the Rand indices for all pairs of k and visualize the Rand indices using a heatmap, with our chosen range of k in the rows and columns, and the color indicating similarity in cluster assignments (red=high similarity, blue=low similarity). We can use the heatmap to find a region of ks where the clustering is fairly similar between runs (e.g. an index around 0.8 or higher). Within that 'robust region', we should choose a low k to retain as much resolution as possible.

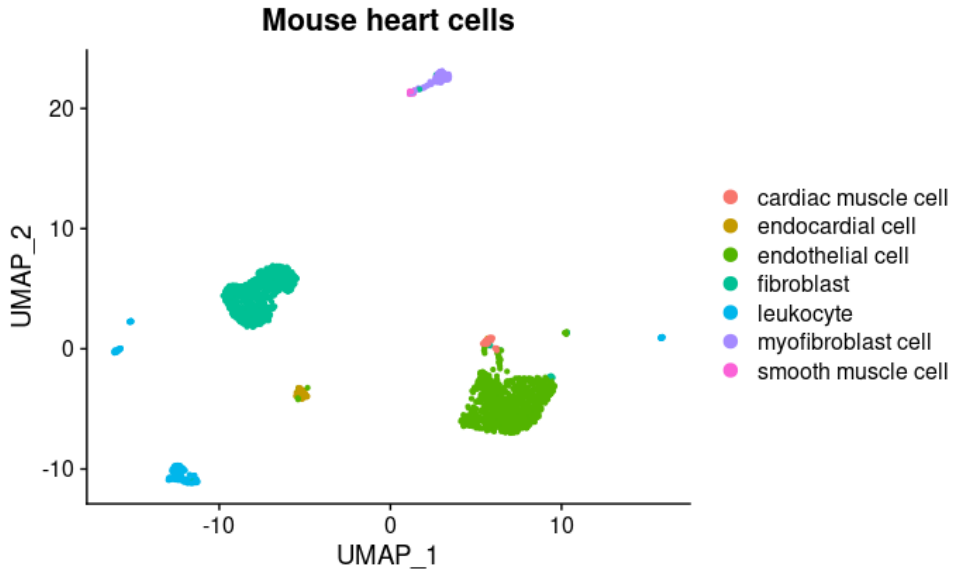
⏳ This step might take a while so be patient

```
# calculate clustering characteristics for a range of ks (5 to 155), by a step of 5.
ks = np.arange(5,155,5)
cluster_chars = [calc_clustering_characteristics(k, adata.obsm['X_pca']) for k in ks];
```

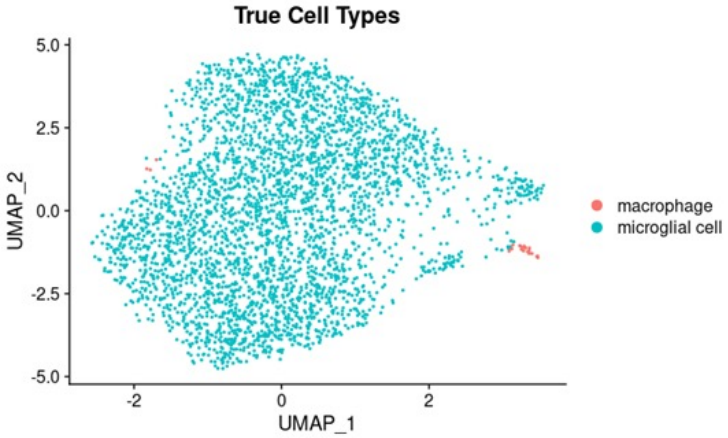
Clustering

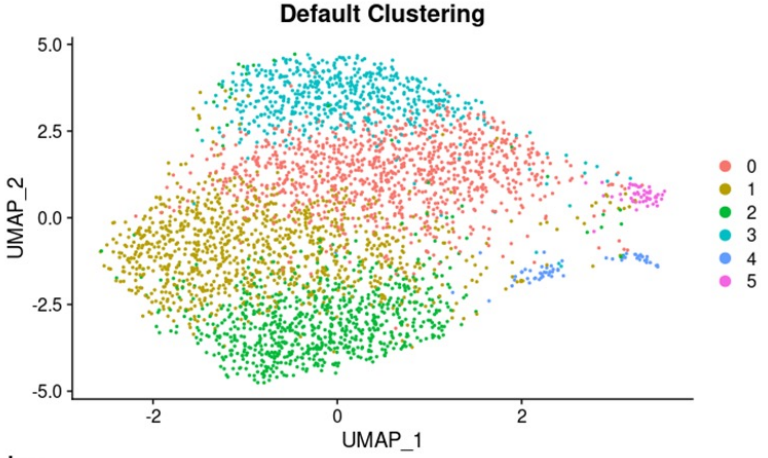
Researchers often turn to unsupervised clustering methods to identify cell subpopulations. After dimensionality reduction, we can perform clustering analysis.

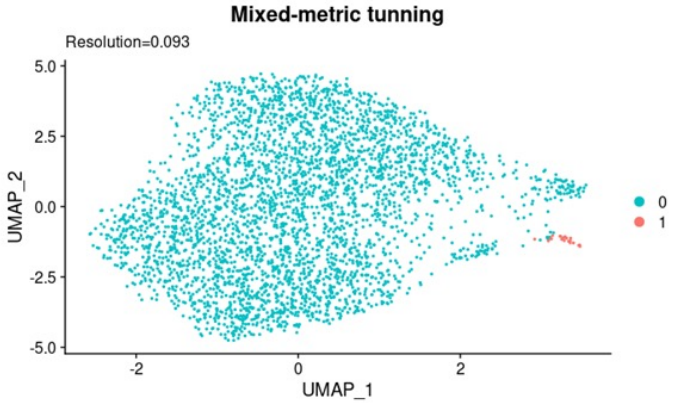
- Results can vary between clustering methods
- Clustering algorithms contain user-defined parameters that must be carefully selected



Motivating Example

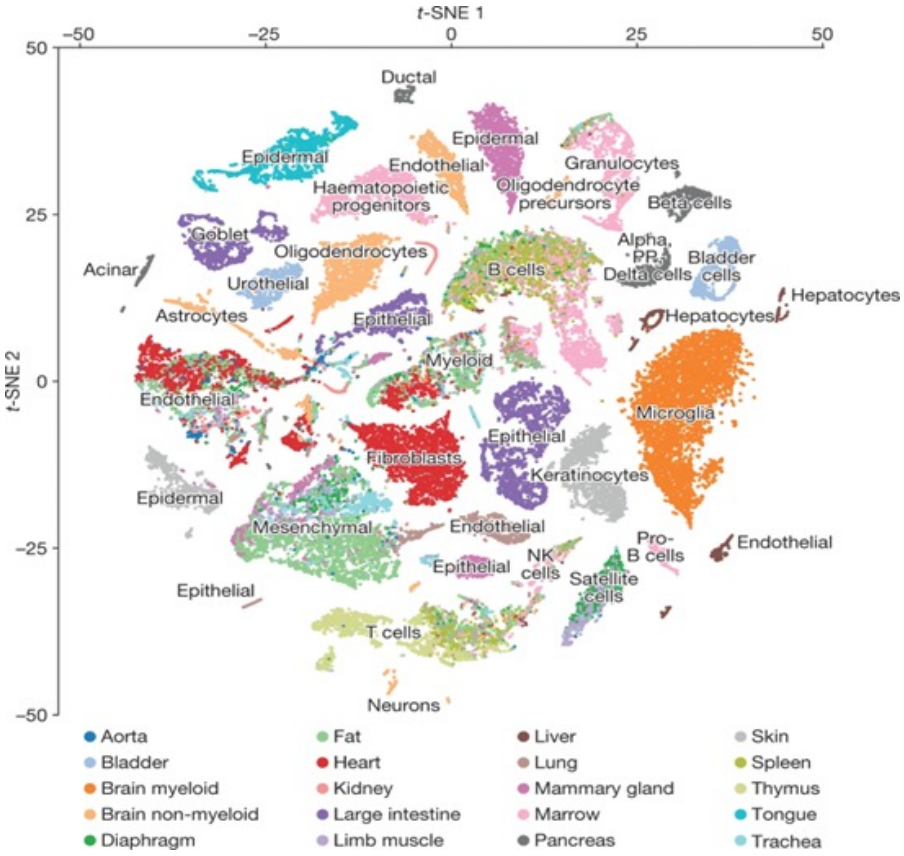




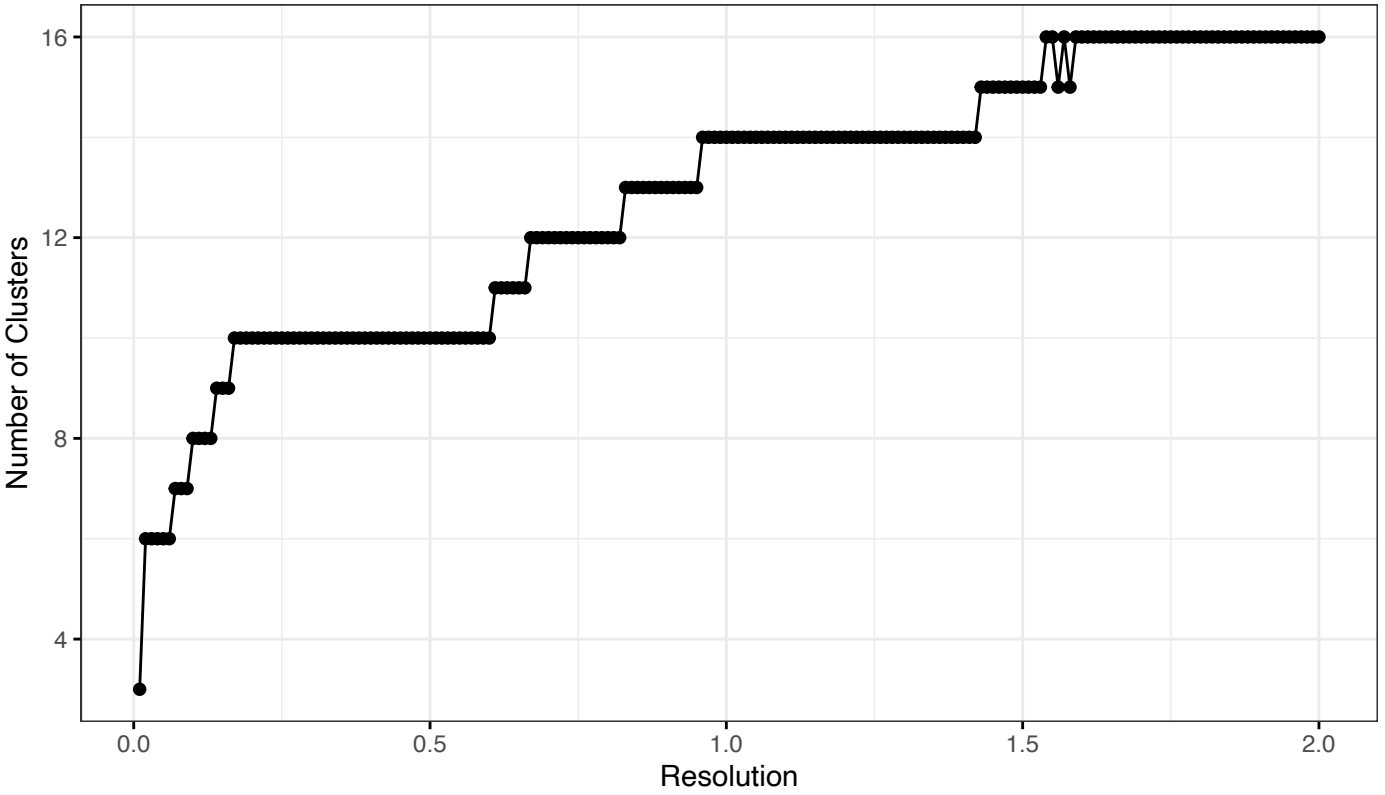


Mouse Atlas

- 20 mouse organs sequenced
- Varying number of cell types in each data set
- FACS sorted

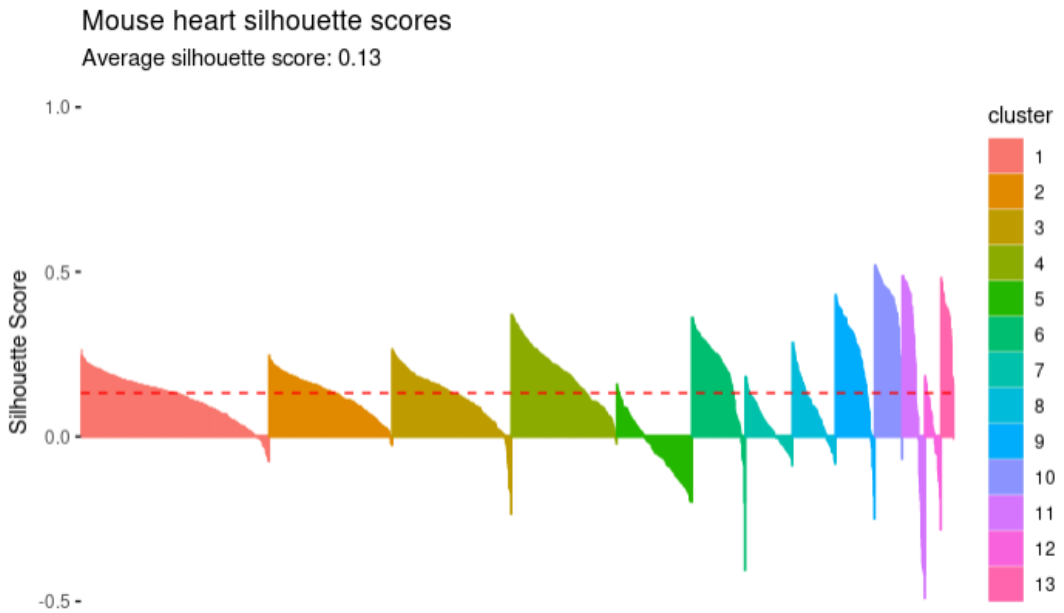


Relationship between Resolution and Number of Clusters



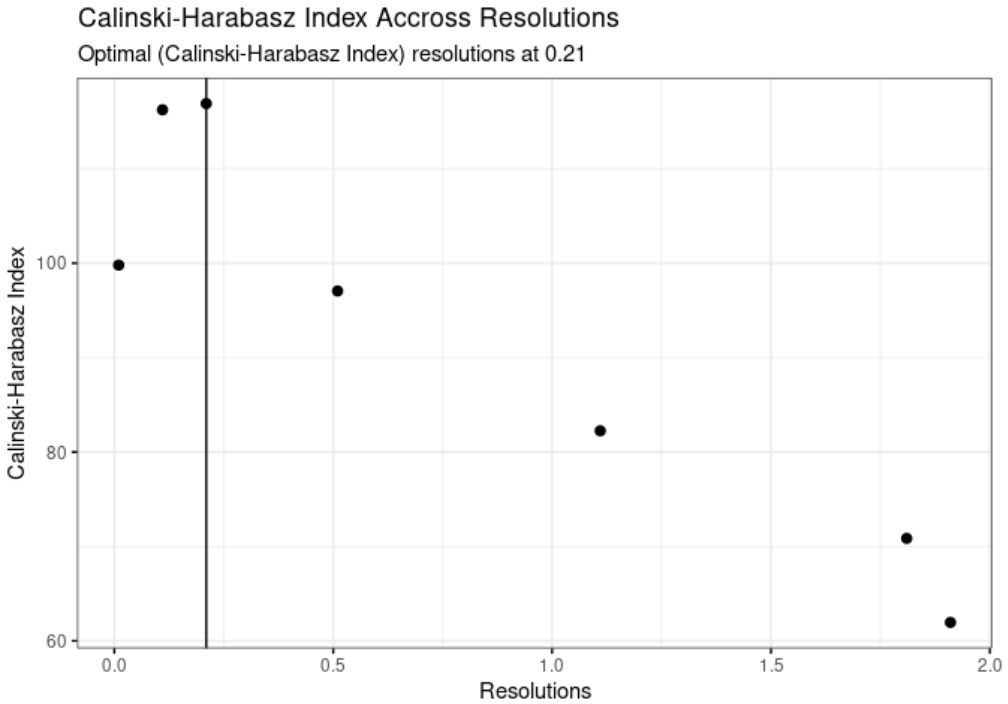
Methods to Tune Parameters: Average Silhouette Score

- How similar each cell is to other cells in its cluster relative to cells in other clusters.
- Metric ranging from -1 to 1 for each cell in the data set.
- Higher scores are better
- Average of all scores can be used to determine performance.



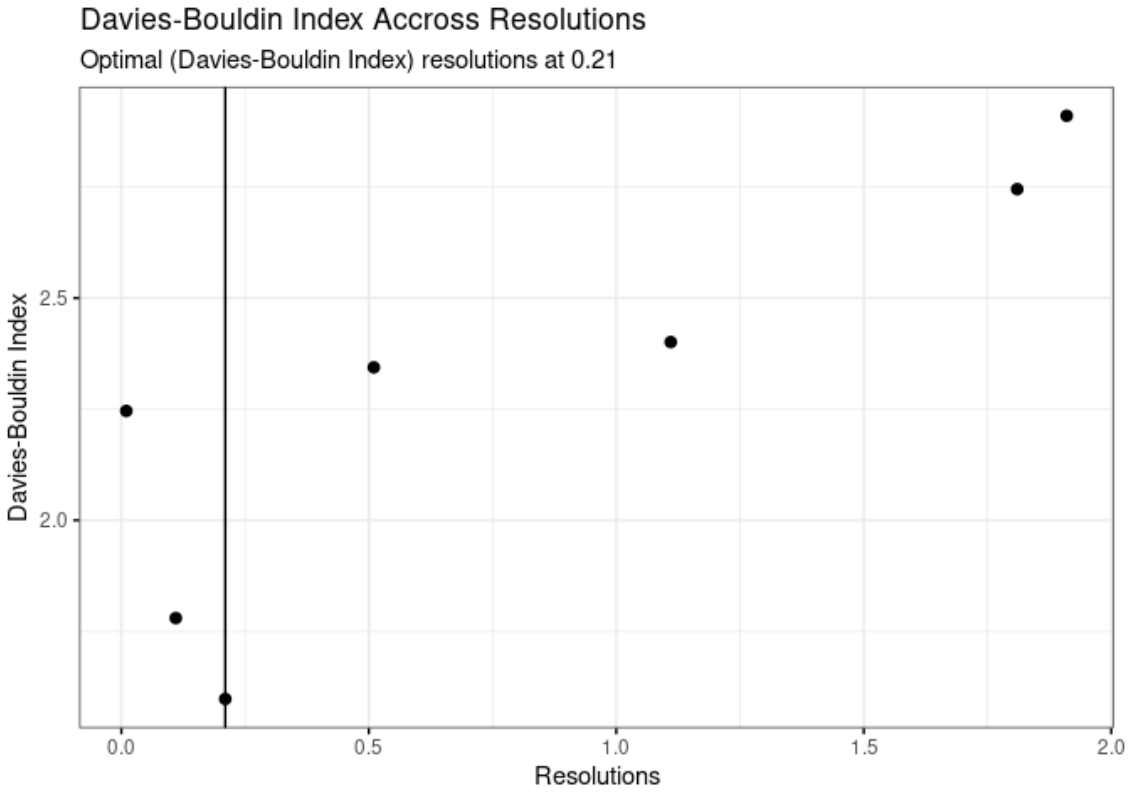
Methods to Tune Parameters: Calinski-Harabasz Index

- Also known as the variance ration criterion.
- Measures the between cluster variance compared to the within cluster. variance between cells
- Takes values greater than 0.
- Larger values are better.

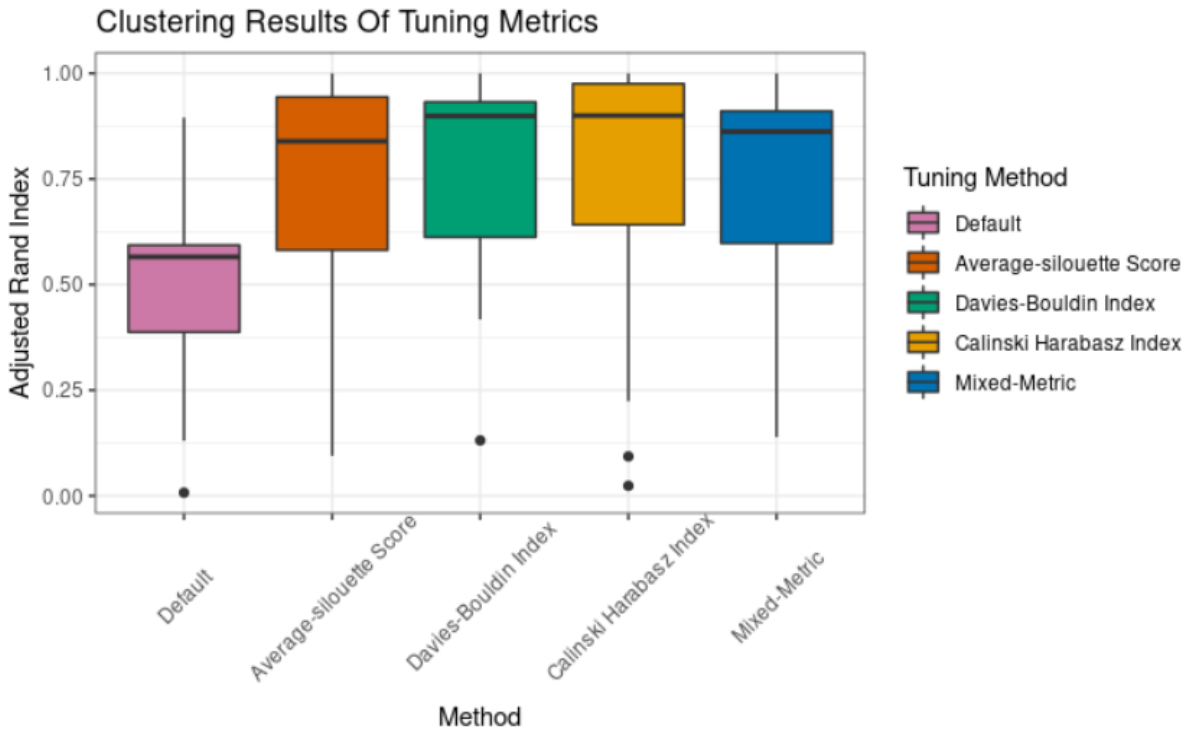


Methods to Tune Parameters: Davies-Bouldin Index

- Defined as the average similarity measure of each cluster to its most similar cluster.
- Minimum score is 0.
- Lower is better.



Default Parameters Are Not Optimal



Summary

- ScRNA-seq is a powerful technology that can allow researchers to unbiasedly classify cell types based on gene expression.
- Benchmarking metrics can help chose which clustering algorithm is best for your experiment.
- Additionally, benchmarking metrics can help tune user-defined parameters of clustering algorithms.
- The clusTuneR package can be utilized with the Seurat pipeline to help choose the best algorithm settings.
- clusTuneR will be available to the public soon.

Future directions

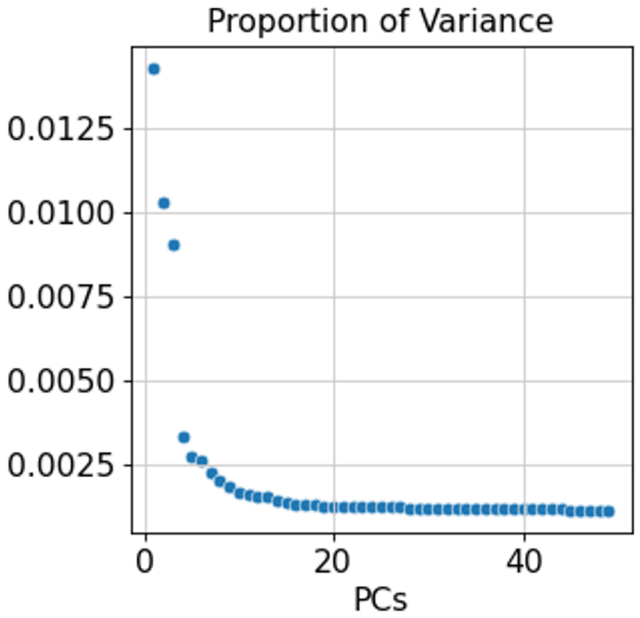
Our team is developing novel approaches to tackle challenging computational and analytical problems within this field. Our cutting-edge research includes

- Multi-omics data integration
- Spatial transcriptomics data analysis

Dimensionality Reduction

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)
- t-SNE
- UMAP



Dimensionality Reduction & Clustering Results