scscape Pipeline Overview

ScRNAseq in the Cloud

MDIBL Comparative Genomics and Data Science Core

MDI Biological Laboratory

Pipeline

01 Make Seurat

02 Normalize QC

03 Doublet Finder

04 Merge

05 PCA

06 Integration

07 Neighbors Clusters Markers

08 Plotting

MDI Biological Laboratory

By Sample

01

Make Seurat

MDI Biological Laboratory

Make Seurat – MEX format

features.tsv.gz | barcodes.tsv.gz | matrix.mtx.gz

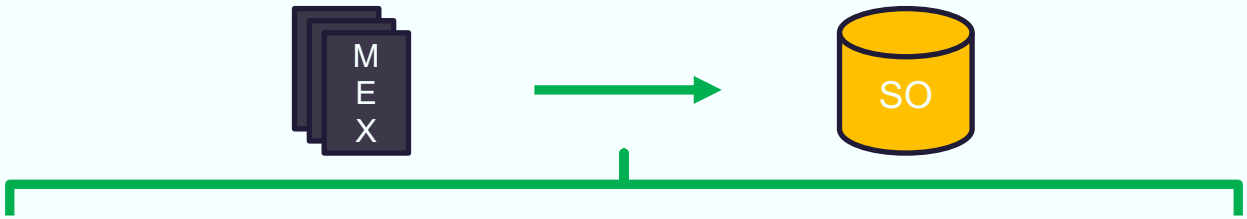
ENSMUSG00000100764	Gm29155	AAACCTGCAAGCTGTT-1	%%MatrixMarket matrix coordin
ENSMUSG00000100635	Gm29157	AAACCTGCACAGCGTC-1	%metadata_json: {"software_ve
ENSMUSG00000100480	Gm29156	AAACCTGTCGGATGTT-1	33596 1760 2893894
ENSMUSG00000051285	Pcmtd1	AAACGGGCAGGATTGG-1	35 1 1
ENSMUSG00000097797	Gm26901	AAACGGGCATCGATGT-1	39 1 1
ENSMUSG00000103067	Gm30414	AAAGATGAGCGTGTCC-1	54 1 2
ENSMUSG00000026312	Cdh7	AAAGATGCATTACGAC-1	71 1 1
ENSMUSG00000039748	Exo1	AAAGATGGTTTGTGTG-1	81 1 1
ENSMUSG00000104158	Becn2	AAAGATGTCACATGCA-1	86 1 1
ENSMUSG00000057363	Uxs1	AAAGATGTCAGCCTAA-1	132 1 7
ENSMUSG00000047216	Cdh19	AAAGATGTCCTTGCCA-1	156 1 1
ENSMUSG00000101253	Gm29088	AAAGCAAAGAGTACAT-1	177 1 15
ENSMUSG00000038702	Dsel	AAAGCAACATAGACTC-1	185 1 1
ENSMUSG00000101453	Gm28189	AAAGTAGCACCAACCG-1	248 1 1

Make Seurat – MEX format

features.tsv.gz		barcodes.tsv.gz	matrix.mtx.gz
ENSMUSG00000000001	Gene01	AAACCTGCAAGCTGTT-1	%%MatrixMarket matrix coordin
ENSMUSG00000000002	Gene02	AAACCTGCACAGCGTC-1	%metadata_json: {"software_ve
ENSMUSG00000000003	Gene03	AAACCTGTCGGATGTT-1	33596 1760 2893894
ENSMUSG00000000004	Gene04	AAACGGGCAGGATTGG-1	1 1 1
ENSMUSG00000000005	Gene05	AAACGGGCATCGATGT-1	4 1 3
ENSMUSG00000000006	Gene06	AAAGATGAGCGTGTCC-1	8 1 2
ENSMUSG00000000007	Gene07	AAAGATGCATTACGAC-1	14 1 1
ENSMUSG00000000008	Gene08	AAAGATGGTTTGTGTG-1	12 2 1
ENSMUSG00000000009	Gene09	AAAGATGTCACATGCA-1	1 2 1
ENSMUSG00000000010	Gene10	AAAGATGTCAGCCTAA-1	4 2 7
ENSMUSG00000000011	Gene11	AAAGATGTCCTTGCCA-1	4 3 1
ENSMUSG00000000012	Gene12	AAAGCAAAGAGTACAT-1	7 3 15
ENSMUSG00000000013	Gene13	AAAGCAACATAGACTC-1	2 3 1
ENSMUSG00000000014	Gene14	AAAGTAGCACCAACCG-1	8 3 1

Cell AAACCTGTCGGATGTT-1 has 15 counts of Gene07.

Make Seurat



Coarse Filtering of Data

- Remove Cell with fewer than a certain number of genes
  - Default set to 200
- Remove Genes with expressed in fewer than a certain number of cells
  - Default set to 3
- Option to manually remove certain genes.
  - ie. unannotated genes.
  - This was not set for us

By Sample

02

Normalize QC

MDI Biological Laboratory

Normalize QC

Mitochondrial Percent

- High Mitochondrial Percentage is indicative of poor quality or dying cells.
  - Default Set to 10% Maximum

Cell Cycle Scoring

- These scores are calculated for both G2M and S Phase and are subsequently regressed out.

Thresholds

- Minimum and maximum gene-count and read-count thresholds can be set.
  - Defaults:
    - Gene and read count minimum set at 10^{th} percentile.
    - Gene and read count maximum not set.

MDI Biological Laboratory

Normalize QC

Why

- Account for cell-to cell difference
  - Varying read count per cell
- Correct technical biases
  - Adjust for sequencing depth differences
- Reveal true biological differences
- Improves data quality

Log Normalize

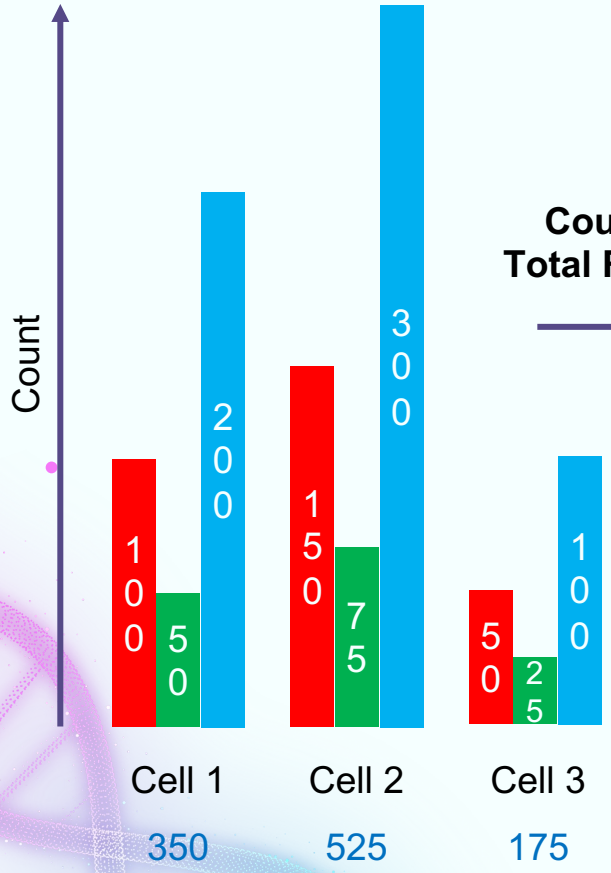
- Simple and Fast
- Applies log transformation
- Struggles with low or zero count genes

SCTransform

- More advanced and computationally intensive
- Creates a model for each gene, looking at gene expression and variability.
- Preforms well for low and zero count genes.
- Specifically designed for Single Cell.

MDI Biological Laboratory

Normalize



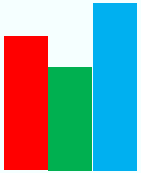
Count / Total Reads

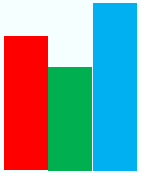
	Cell1	Cell2	Cell3
Gene A	100/350	150/525	50/175
Gene B	50/350	75/525	25/175
Gene C	200/350	300/525	100/175

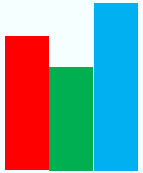
Multiply By Scale Factor

	Cell1	Cell2	Cell3
Gene A	(100/350*10000)+1	(150/525*10000)+1	(50/175*10000)+1
Gene B	(50/350*10000)+1	(75/525*10000)+1	(25/175*10000)+1
Gene C	(200/350*10000)+1	(300/525*10000)+1	(100/175*10000)+1

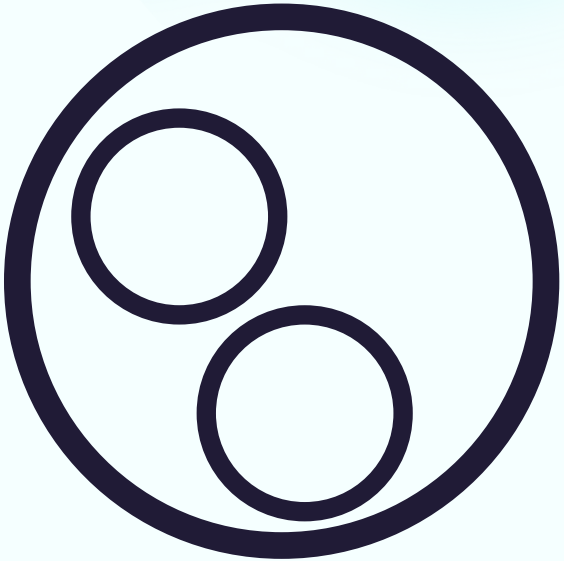
natural log transformation







MDI Biological Laboratory



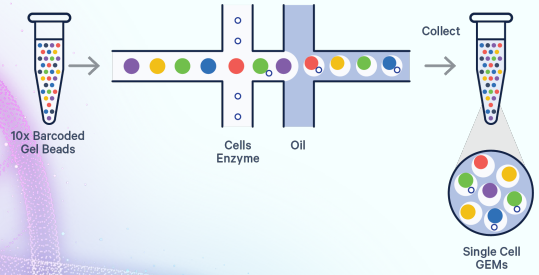
03

Doublet Finder

Doublet Finder

What is a Doublet

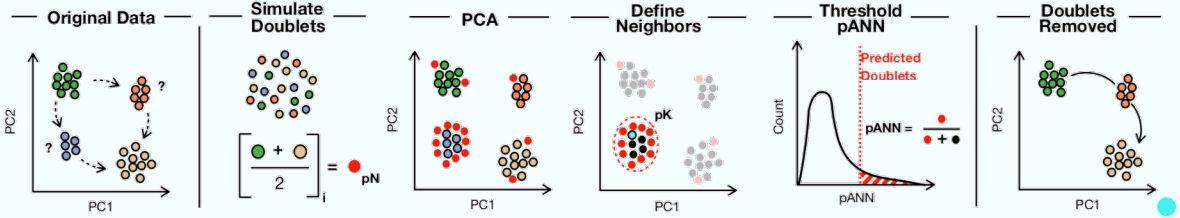
- The 10x chromium technology is a droplet-based meaning
- 2 cells may sneak into the same droplet getting assigned the same barcode.



Doublet Finder Overview

1. Generate artificial doublets from existing ScRNAseq data.
2. Pre-process combined rea and artificial data.

3. Performs PCA and use the PC distance matrix to find each cell's proportion of artificial k nearest neighbors (paNN)
Rank order and threshold pANN values according to expected doublet number.

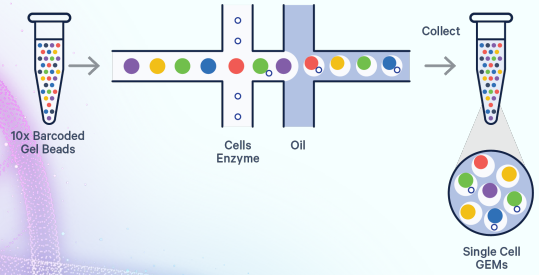


MDI Biological Laboratory

Doublet Finder

What is a Doublet

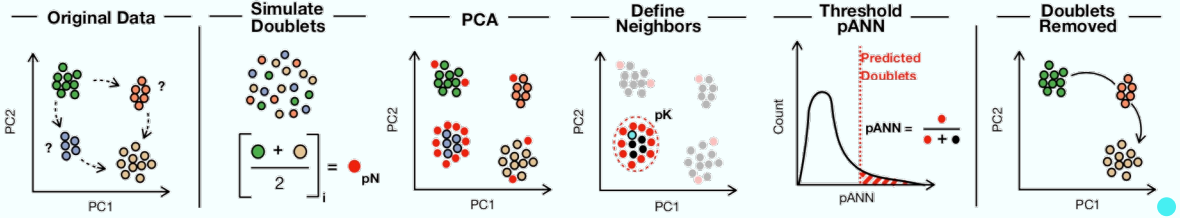
- The 10x chromium technology is a droplet-based meaning
- 2 cells may sneak into the same droplet getting assigned the same barcode.



Doublet Finder Overview

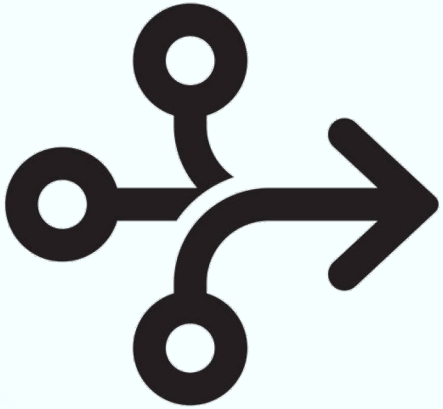
1. Generate artificial doublets from existing ScRNAseq data.
2. Pre-process combined rea and artificial data.

3. Performs PCA and use the PC distance matrix to find each cell's proportion of artificial k nearest neighbors (paNN)
Rank order and threshold pANN values according to expected doublet number.



MDI Biological Laboratory

By
Group



04

Merge

MDI Biological Laboratory

Merge

Merge

- Now that QC has been completed for each sample, we can combine them into various analysis groups for the remainder of the pipeline.

Normalize

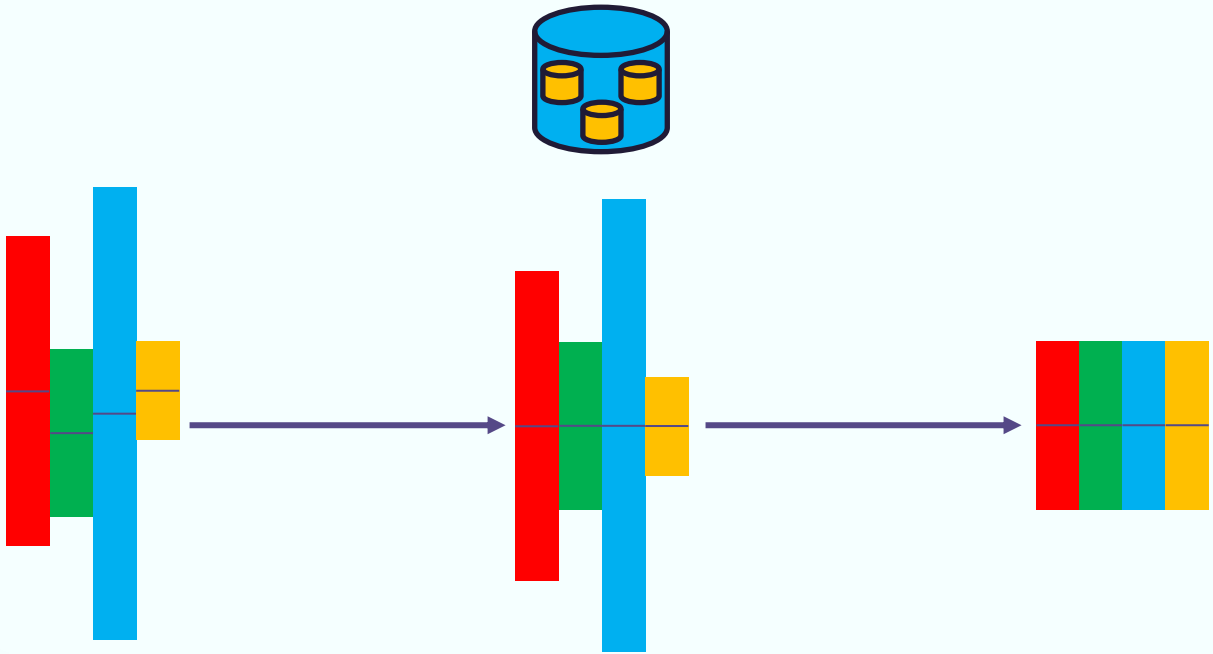
- Normalize with all cells together.
- Still have the same options previously discussed.

Scale

- Regress out unwanted sources of variation in the data to help to focus on biological signal.
  - Cell cycle genes
  - MT %
  - Number of Genes or Reads detected per cell.
  - Etc
- Puts all genes on the same scale.
  - Prevents highly expressed genes from dominating the anaysis

MDI Biological Laboratory

Merge -- Scale

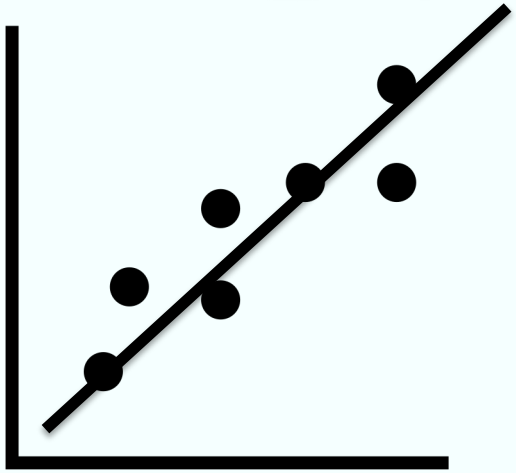


Pre-Scale/Center

Centered

Centered and Scaled

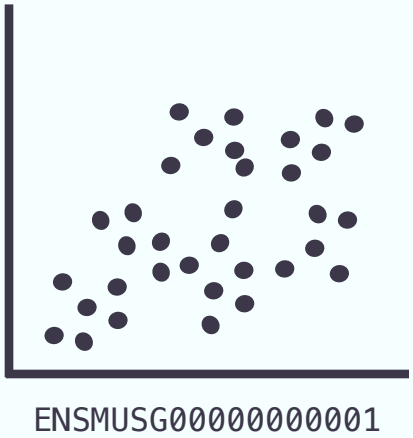
MDI Biological Laboratory

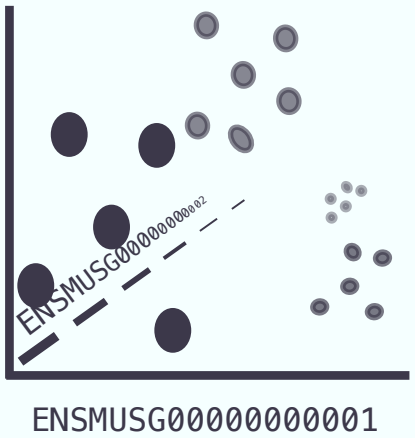


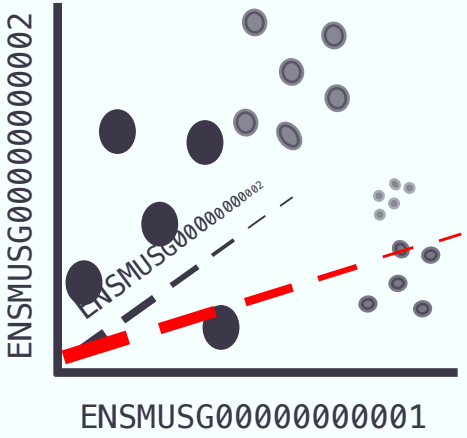
05

PCA

Comparing Samples in Varying Dimensions



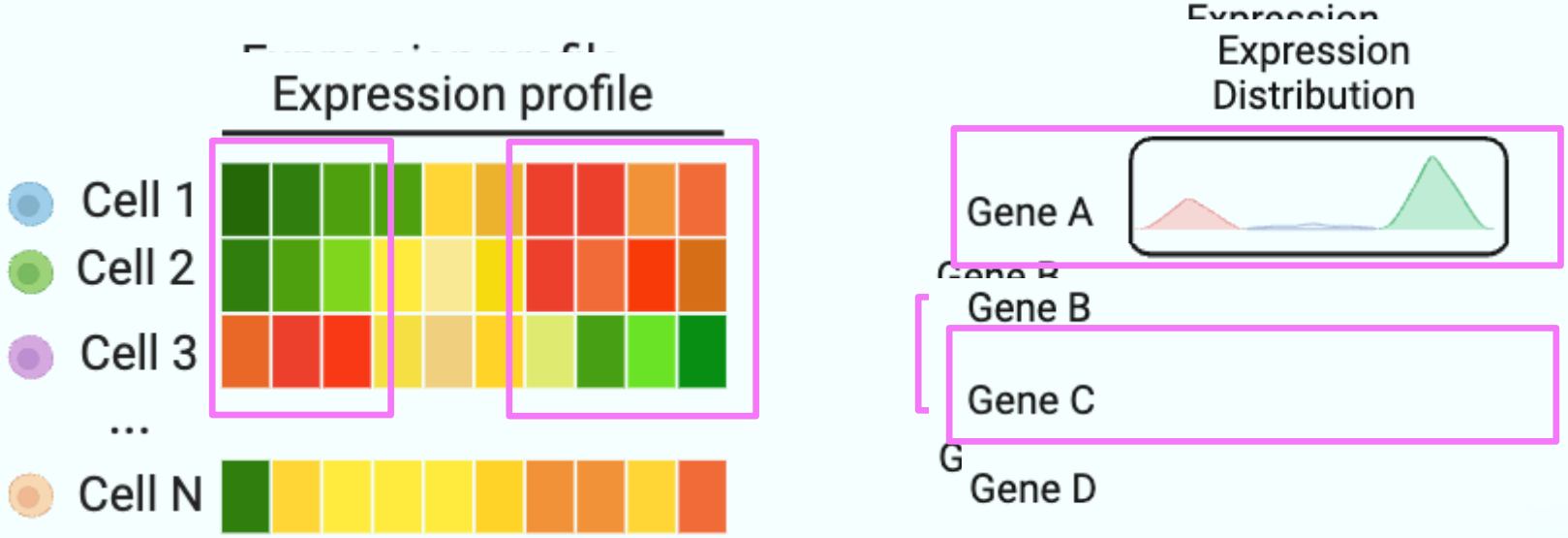




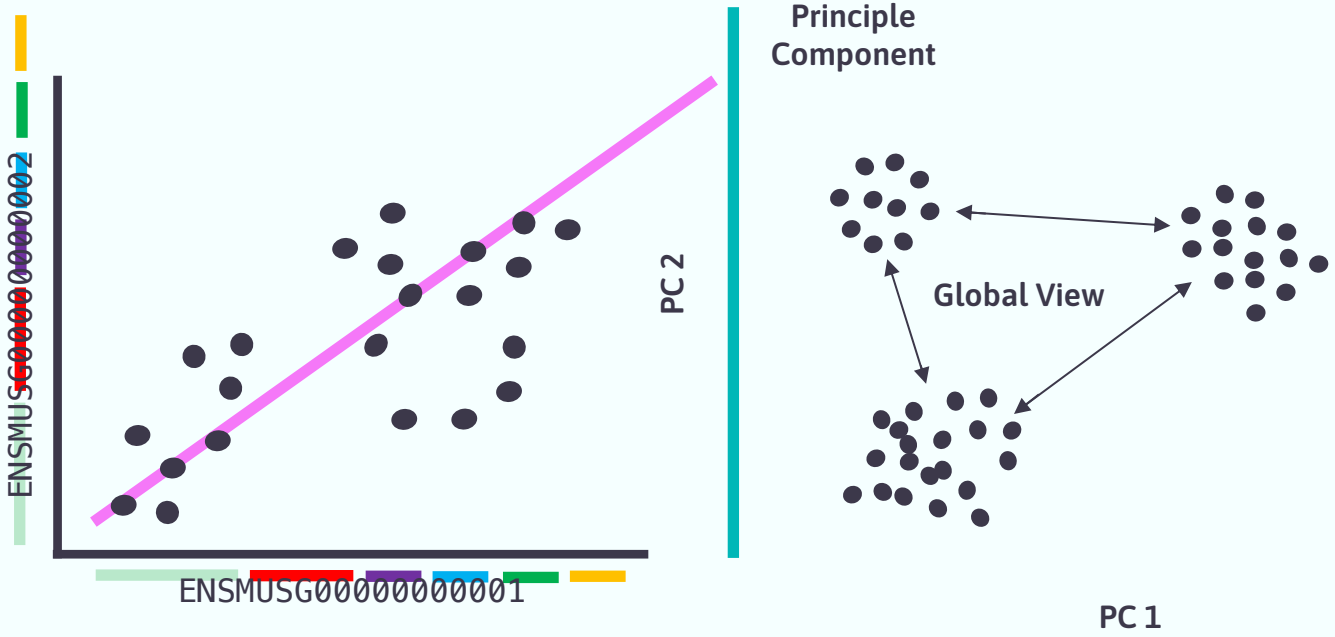
DIMENSIONAL REDUCTION

Variable Features and Covariance

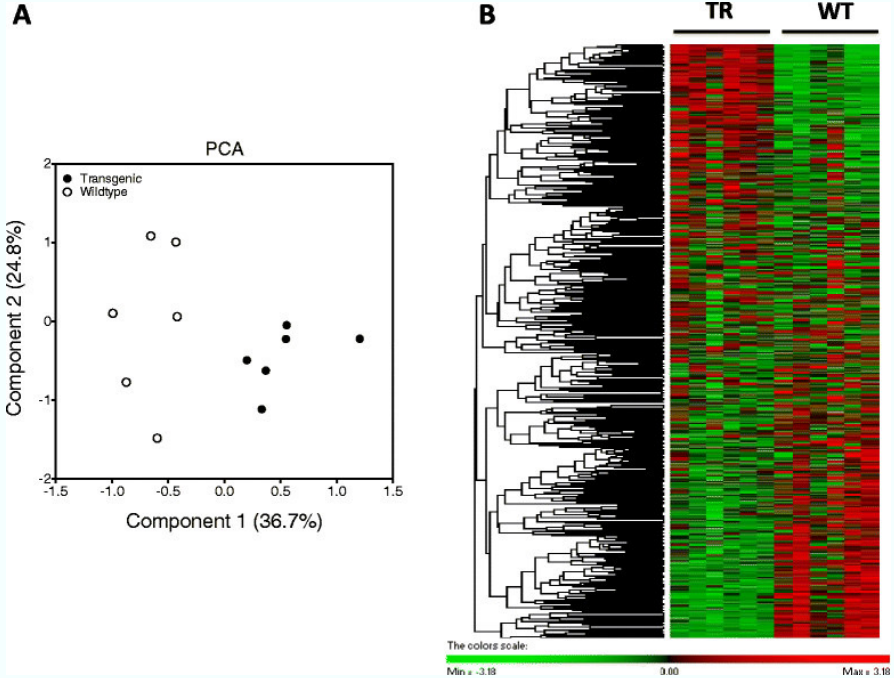
Variable Features and Covariance

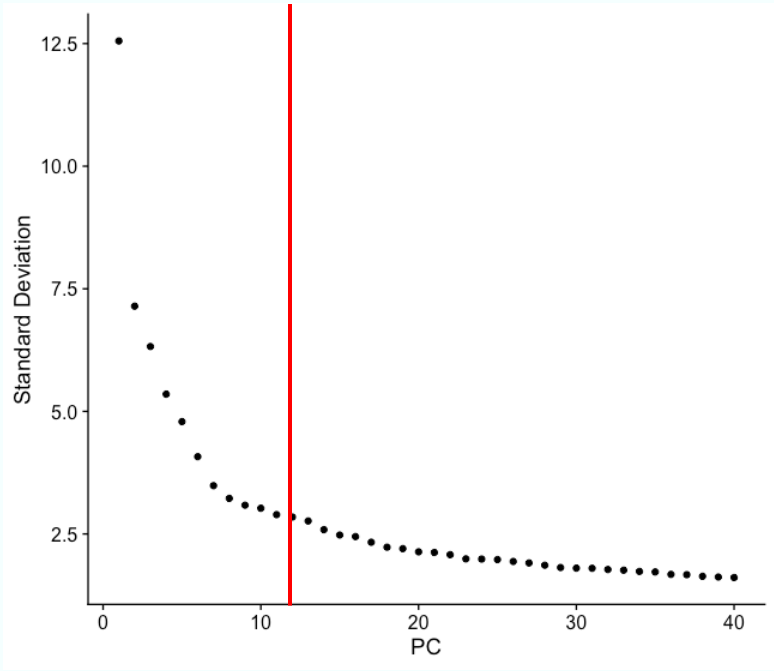


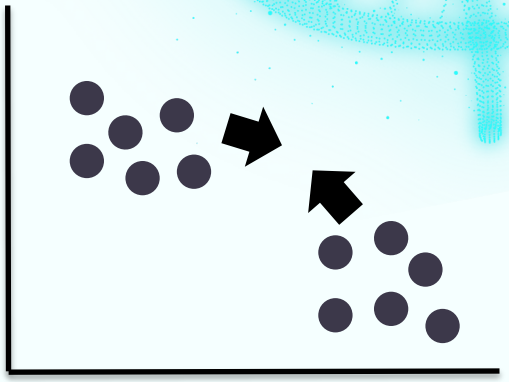
Principle Components



Principle Component Thresholding



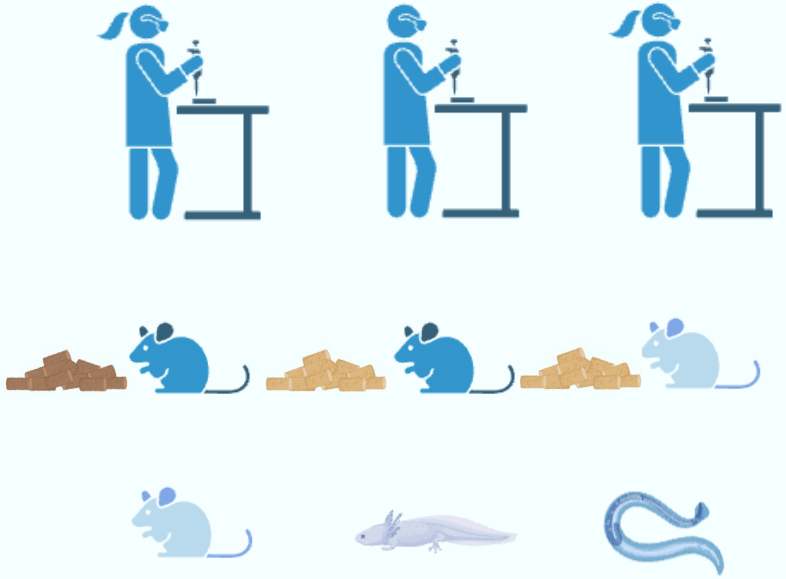


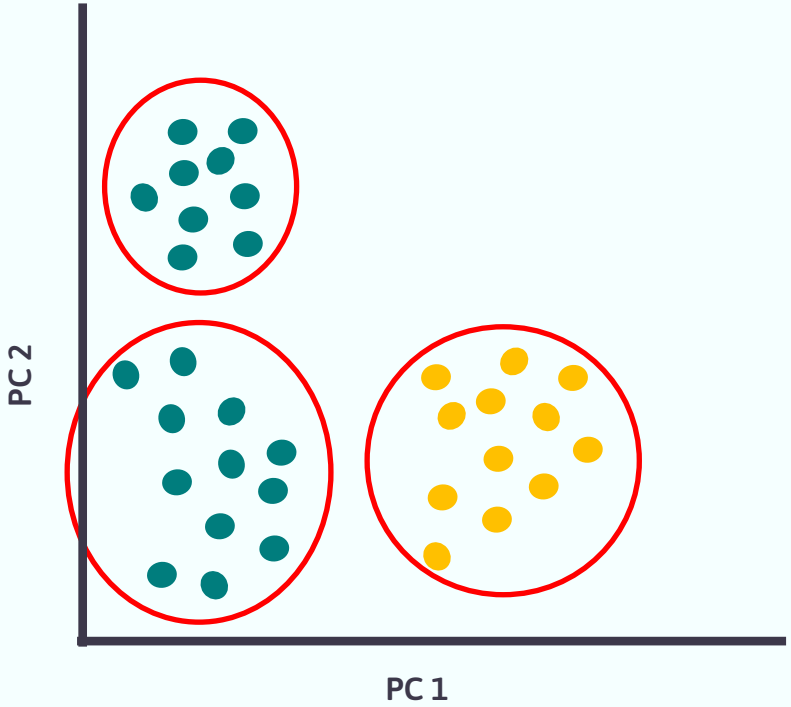


06

Integration

Integration Use Cases

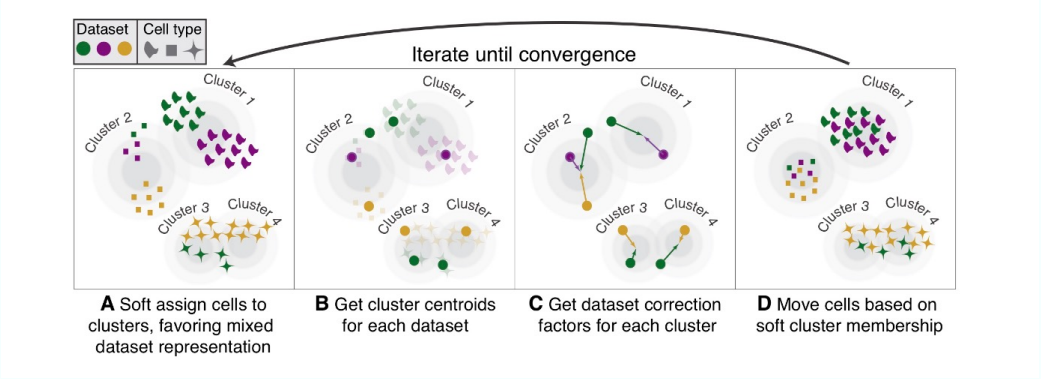


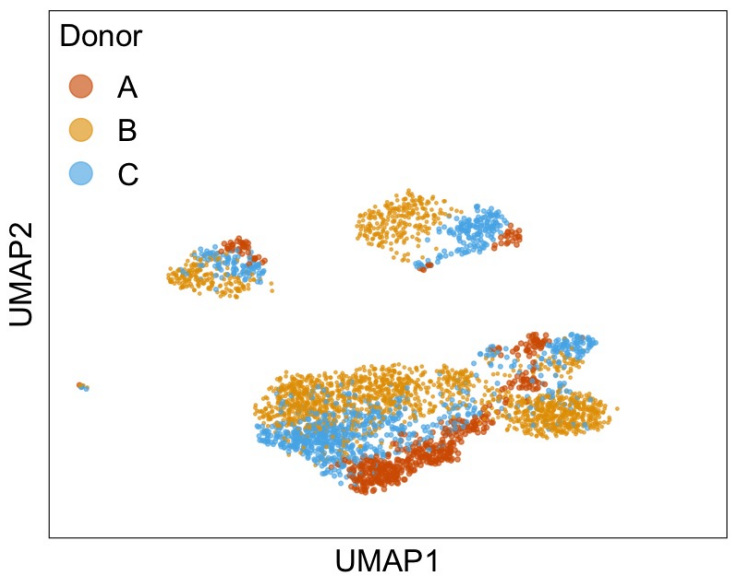


Harmony

- **Harmony** : An algorithm that projects cells into clusters based on their cell identity rather than dataset specific conditions.

- Harmony applies a transformation to the principal component values. The algorithm then determines if there is a balanced quantity of cells from each batch within the clusters. Each cell is then evaluated to see how much its batch identity influences its PC coordinates. The cells position is corrected by shifting it towards the centroid of its cluster.





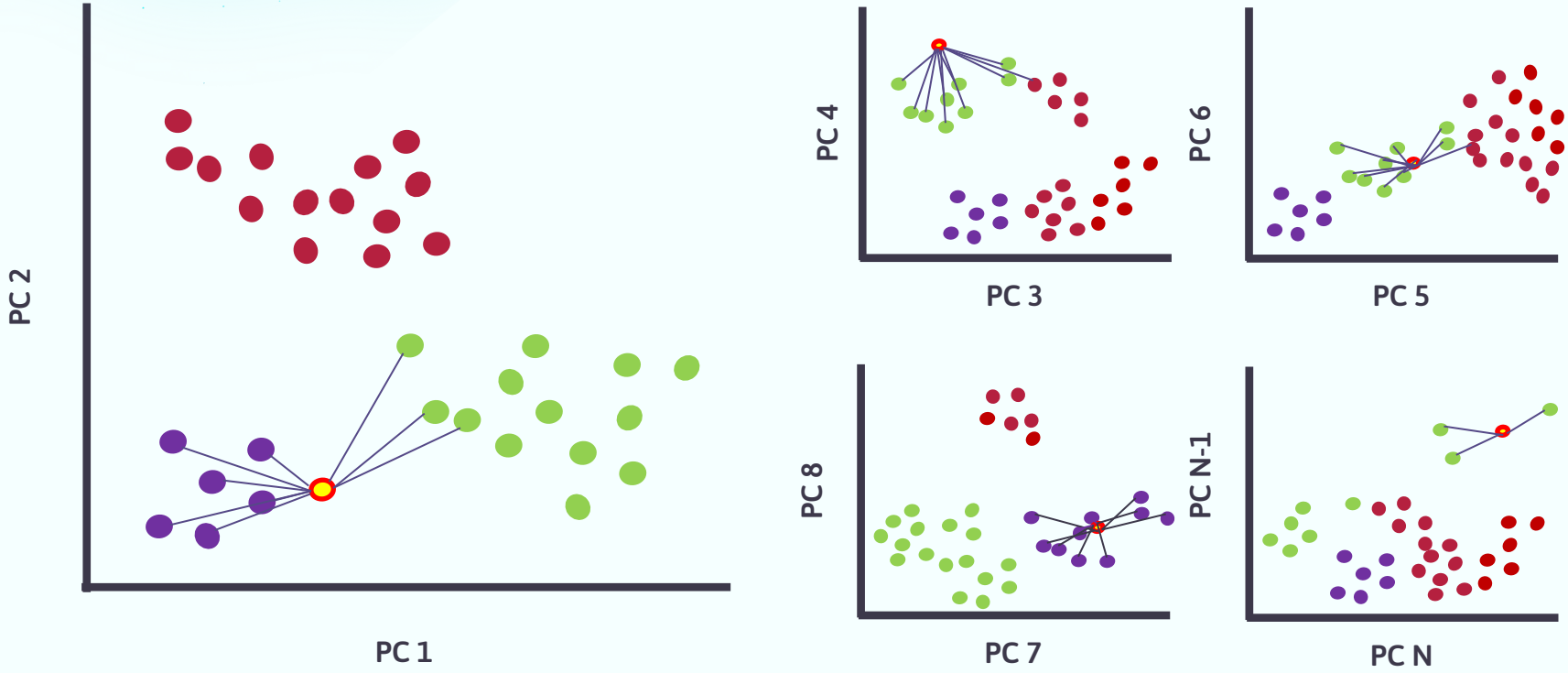
By
Group

07

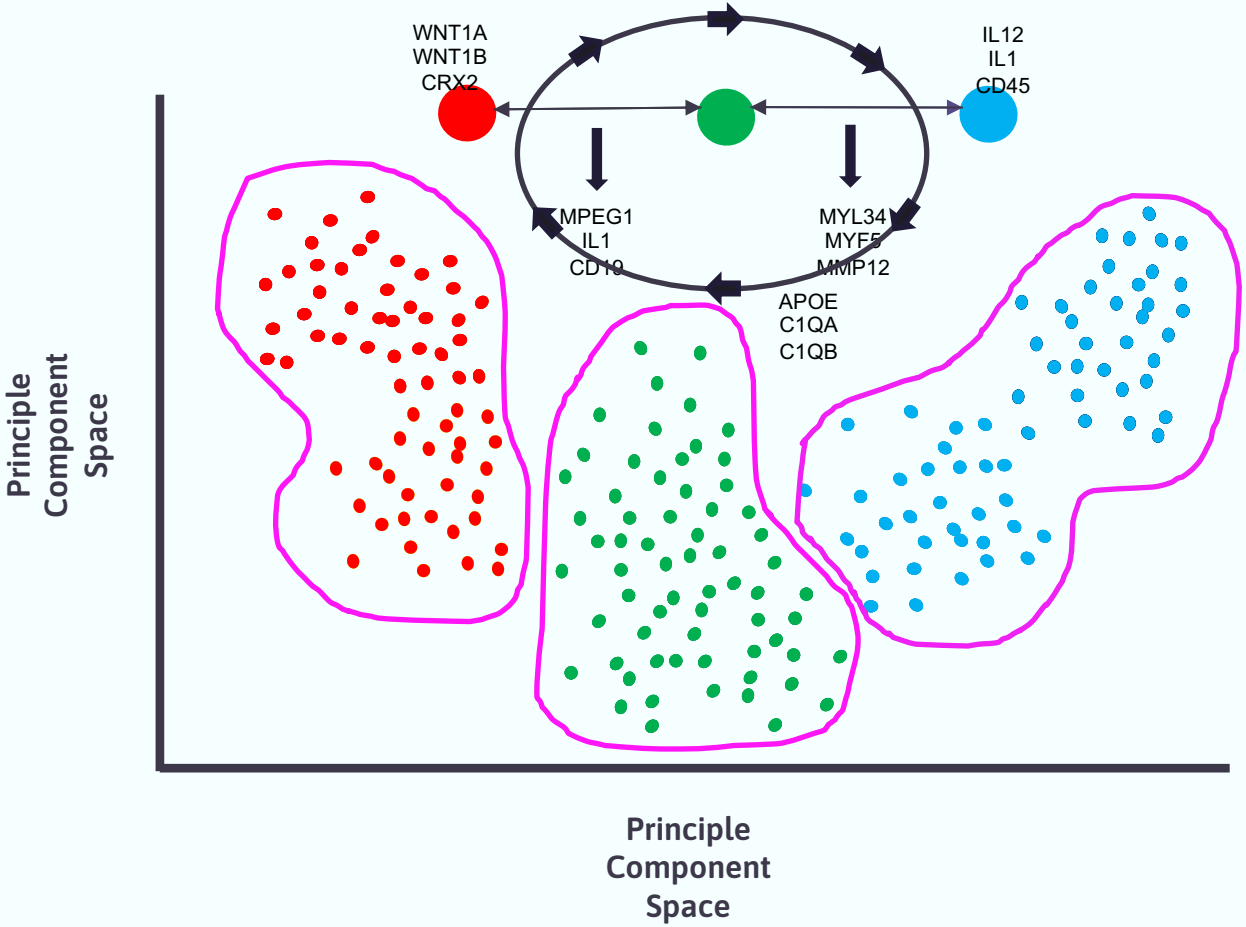
Neighbors Clusters Markers

MDI Biological Laboratory

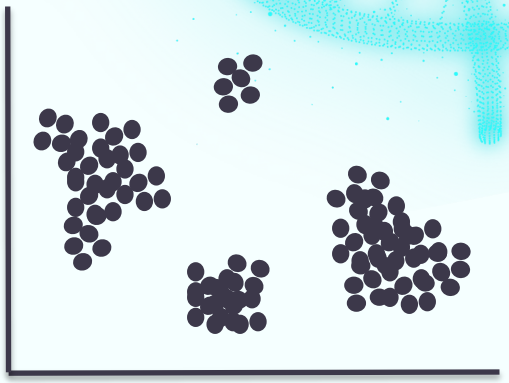
K Nearest Neighbors



Louvain Clustering and Marker Calculation



By
Group

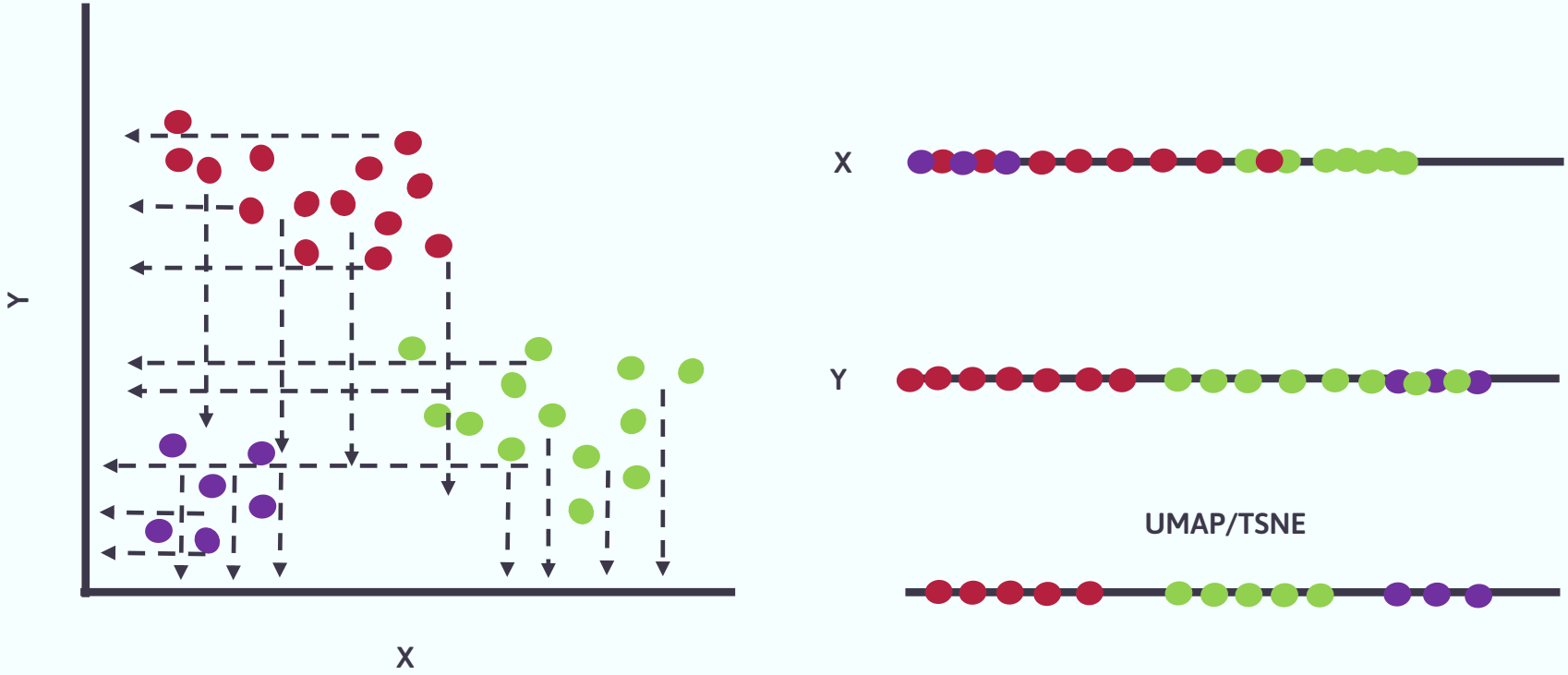


08

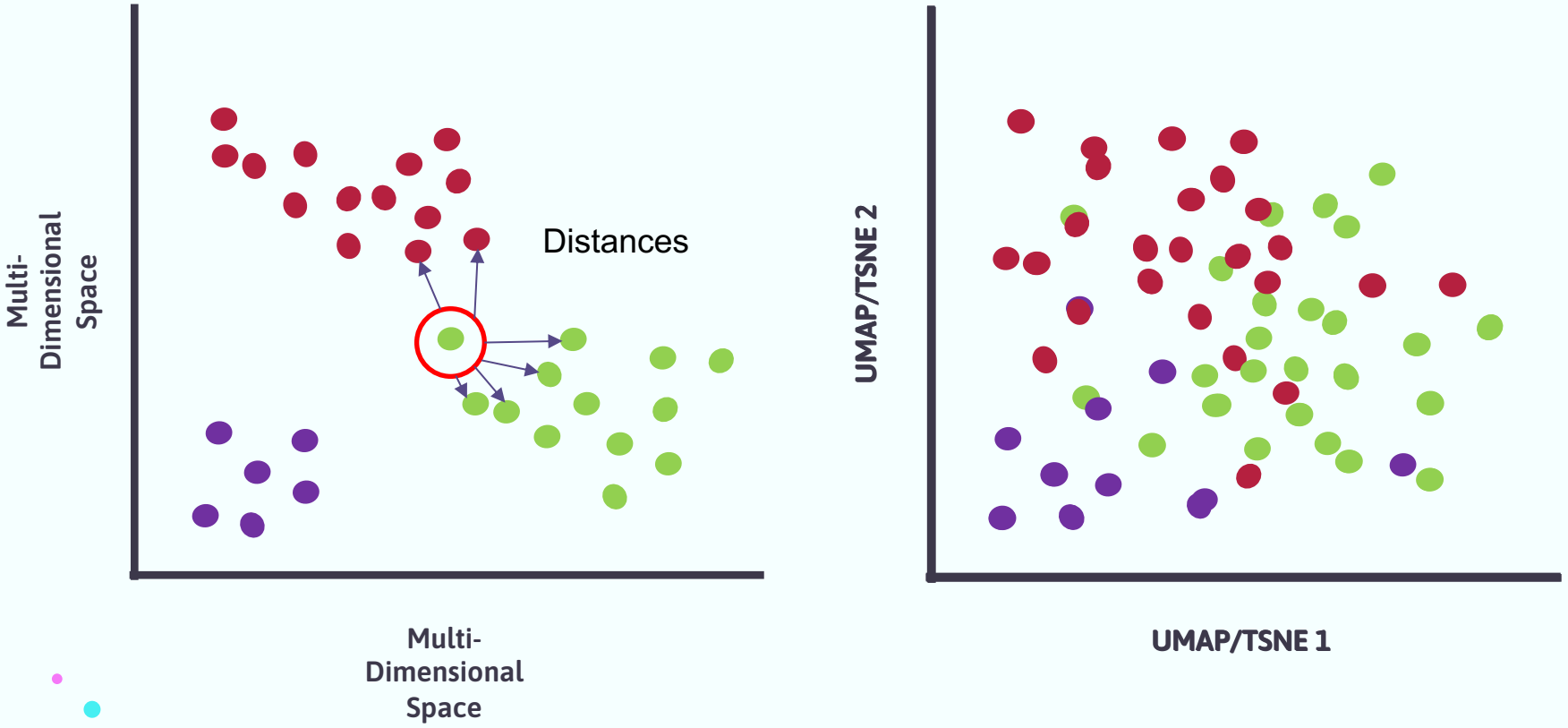
Plotting

MDI Biological Laboratory

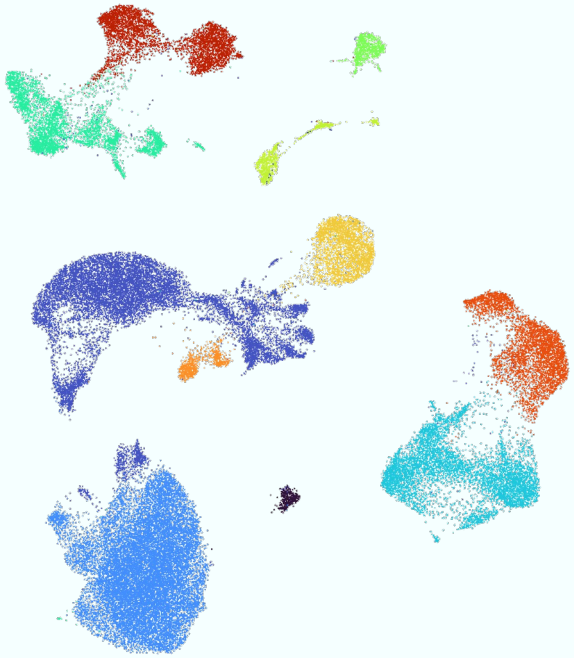
Dimensional Reduction



UMAP and T-SNE



Questions?



MDI Biological Laboratory