Single-cell RNA-seq clustering and dimensionality reduction

The goal of this exercise is to use different dimensionality reduction methods for visualisation and clustering.

The data example is 5 000 peripheral blood mononuclear cells It should be already loaded and preprocessed at the start of your Galaxy history.

Exercise

For each of PCA, tSNE and UMAP, perform the following steps:

- dimensionality reduction
- clustering
- visualisation

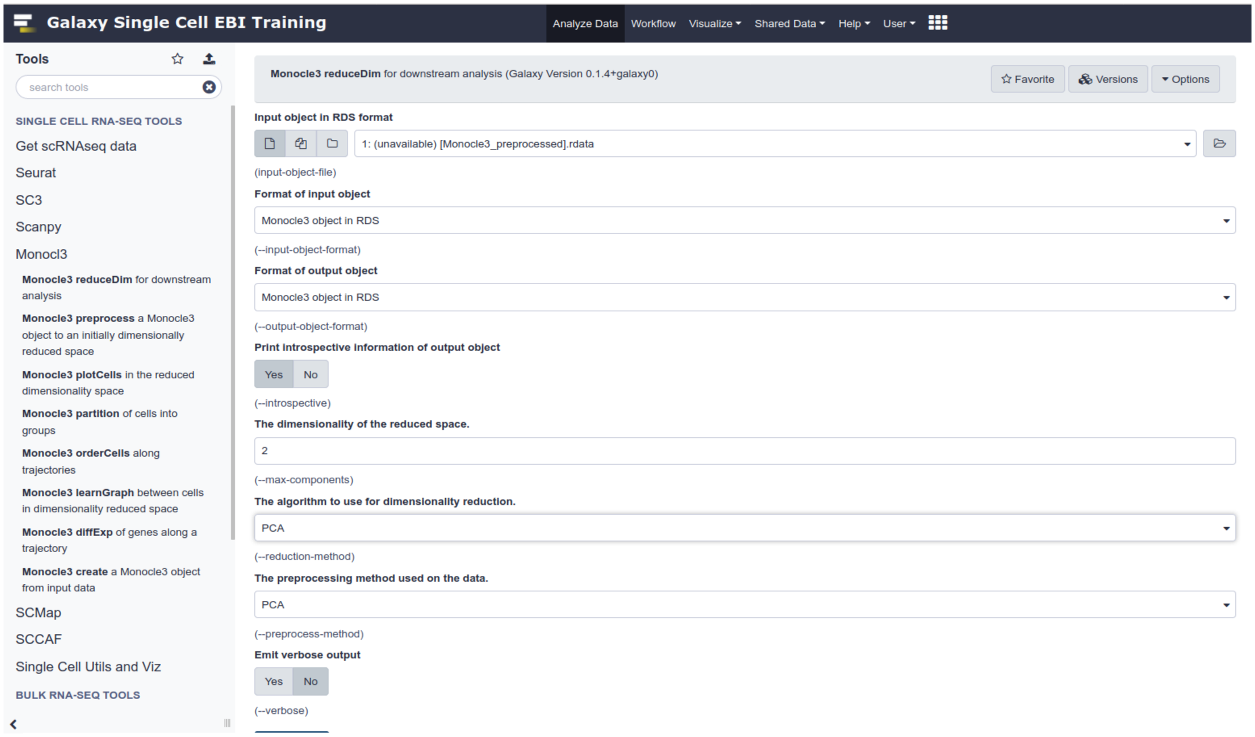
Compare the results.

- What are key differences between the plots?
- How many clusters are generated?
- Can you think of cases where one of the visualisations would be a better choice?

Afterwards you will see top markers assigned to each of the cluster. Can you recognise any cell types from the top markers?

Dimensionality reduction

Use **Monocl3 reduceDim** tool under **Monocl3** section. Use the initial prepared data as input and select a dimensionality reduction algorithm (PCA, tSNE or UMAP)


Galaxy Single Cell EBI Training
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
SINGLE CELL RNA-SEQ TOOLS
Get scRNAseq data
Seurat
SC3
Scanpy
Monocl3
Monocle3 reduceDim for downstream analysis
Monocle3 preprocess a Monocle3 object to an initially dimensionally reduced space
Monocle3 plotCells in the reduced dimensionality space
Monocle3 partition of cells into groups
Monocle3 orderCells along trajectories
Monocle3 learnGraph between cells in dimensionality reduced space
Monocle3 diffExp of genes along a trajectory
Monocle3 create a Monocle3 object from input data
SCMap
SCCAF
Single Cell Utils and Viz
BULK RNA-SEQ TOOLS
Favorite
Versions
Options

Monocle3 reduceDim for downstream analysis (Galaxy Version 0.1.4+galaxy0)

Input object in RDS format
1: (unavailable) [Monocle3_preprocessed].rdata
(input-object-file)

Format of input object
Monocle3 object in RDS
(--input-object-format)

Format of output object
Monocle3 object in RDS
(--output-object-format)

Print introspective information of output object
[✓ Yes] No
(--introspective)

The dimensionality of the reduced space.
2
(--max-components)

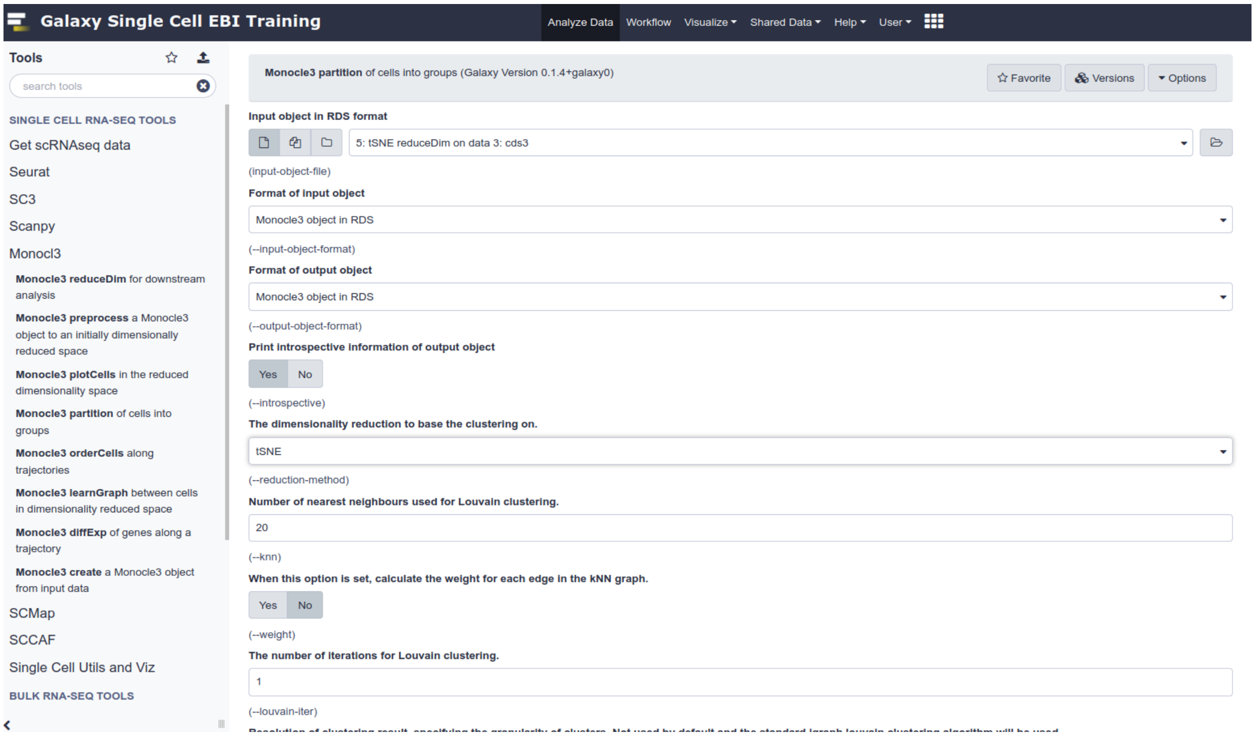
The algorithm to use for dimensionality reduction.
PCA
(--reduction-method)

The preprocessing method used on the data.
PCA
(--preprocess-method)

Emit verbose output
Yes [✓ No]
(--verbose)

Clustering

Use **Monocl3 partition** tool under **Monocl3** section. Use the result of the dimensionality reduction as input and select the same dimensionality reduction.


Galaxy Single Cell EBI Training
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
SINGLE CELL RNA-SEQ TOOLS
Get scRNAseq data
Seurat
SC3
Scanpy
Monocl3
Monocle3 reduceDim for downstream analysis
Monocle3 preprocess a Monocle3 object to an initially dimensionally reduced space
Monocle3 plotCells in the reduced dimensionality space
Monocle3 partition of cells into groups
Monocle3 orderCells along trajectories
Monocle3 learnGraph between cells in dimensionality reduced space
Monocle3 diffExp of genes along a trajectory
Monocle3 create a Monocle3 object from input data
SCMap
SCCAF
Single Cell Utils and Viz
BULK RNA-SEQ TOOLS
Favorite
Versions
Options

Monocle3 partition of cells into groups (Galaxy Version 0.1.4+galaxy0)

Input object in RDS format
5: tSNE reduceDim on data 3: cds3
(input-object-file)

Format of input object
Monocle3 object in RDS
(--input-object-format)

Format of output object
Monocle3 object in RDS
(--output-object-format)

Print introspective information of output object
[✓ Yes] No
(--introspective)

The dimensionality reduction to base the clustering on.
tSNE
(--reduction-method)

Number of nearest neighbours used for Louvain clustering.
20
(--knn)

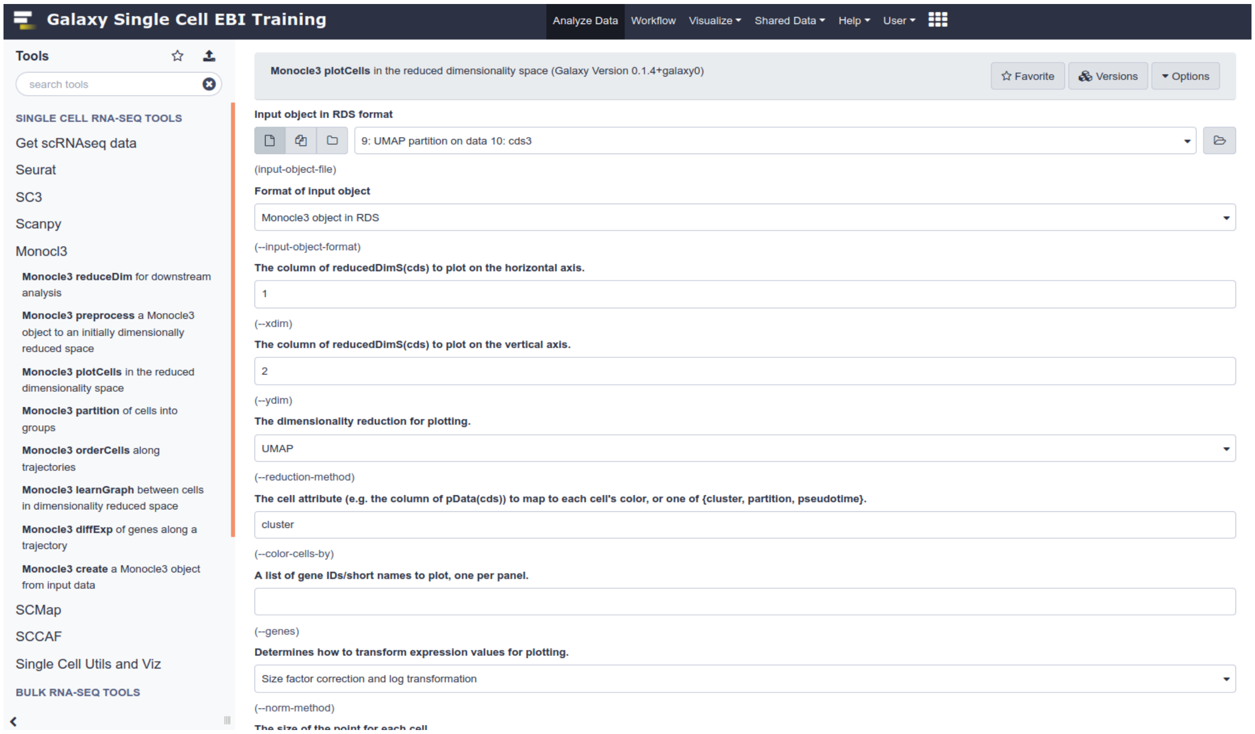
When this option is set, calculate the weight for each edge in the kNN graph.
Yes [✓ No]
(--weight)

The number of iterations for Louvain clustering.
1
(--louvain-iter)

Resolution of clustering result, specifying the granularity of clusters. Not used by default and the standard igraph louvain clustering algorithm will be used

Generating plots

Use **Monocl3 plotCells** tool under **Monocl3** section. The input is the result of the clustering step. Select the appropriate dimensionality reduction and set cell attribute to “cluster”.


Galaxy Single Cell EBI Training
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
SINGLE CELL RNA-SEQ TOOLS
Get scRNAseq data
Seurat
SC3
Scanpy
Monocl3
Monocle3 reduceDim for downstream analysis
Monocle3 preprocess a Monocle3 object to an initially dimensionally reduced space
Monocle3 plotCells in the reduced dimensionality space
Monocle3 partition of cells into groups
Monocle3 orderCells along trajectories
Monocle3 learnGraph between cells in dimensionality reduced space
Monocle3 diffExp of genes along a trajectory
Monocle3 create a Monocle3 object from input data
SCMap
SCCAF
Single Cell Utils and Viz
BULK RNA-SEQ TOOLS
Favorite
Versions
Options

Monocle3 plotCells in the reduced dimensionality space (Galaxy Version 0.1.4+galaxy0)

Input object in RDS format
9: UMAP partition on data 10: cds3
(input-object-file)

Format of input object
Monocle3 object in RDS
(--input-object-format)

The column of reducedDimS(cds) to plot on the horizontal axis.
1
(--xdim)

The column of reducedDimS(cds) to plot on the vertical axis.
2
(--ydim)

The dimensionality reduction for plotting.
UMAP
(--reduction-method)

The cell attribute (e.g. the column of pData(cds)) to map to each cell's color, or one of {cluster, partition, pseudotime}.
cluster
(--color-cells-by)

A list of gene IDs/short names to plot, one per panel.

(--genes)

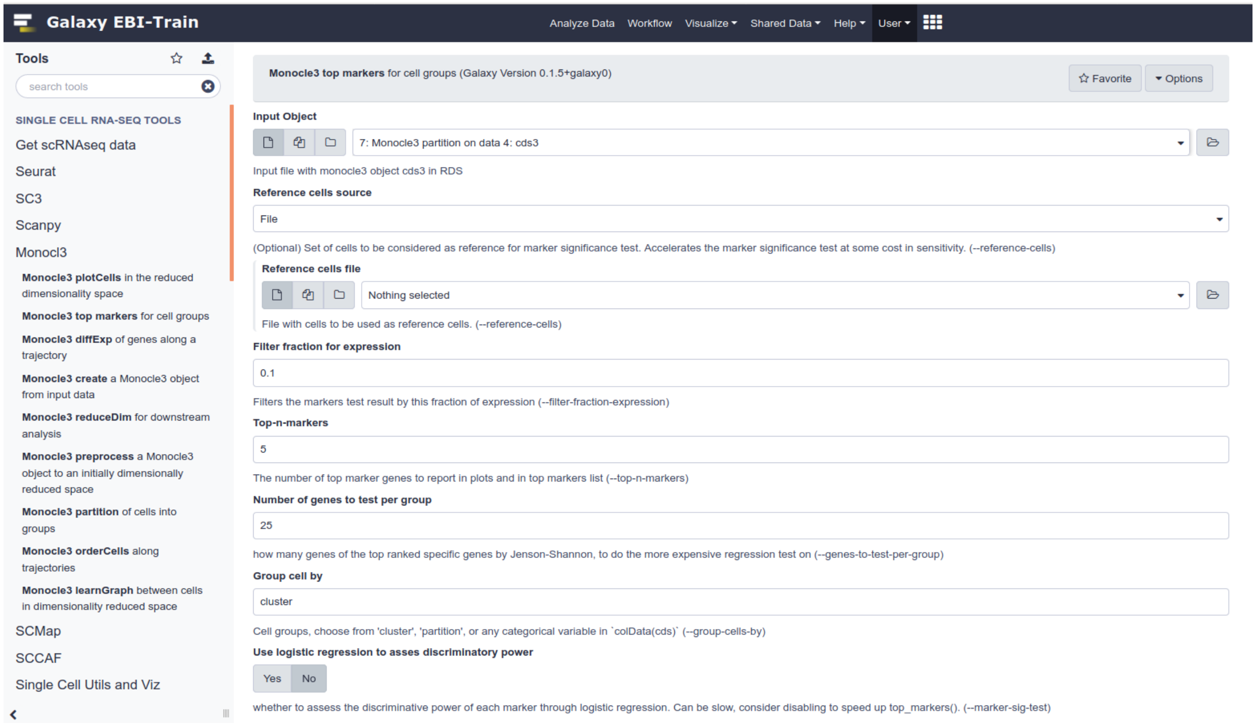
Determines how to transform expression values for plotting.
Size factor correction and log transformation
(--norm-method)

The size of the point for each cell

Marker identification

Use **Monocl3 top markers** tool. The input is the result of the UMAP clustering step. It will not work with other types. It will execute well just with the default options, but you can change the number of top markers per cluster to be listed, genes to be tested and the fraction of cells that need to be expressing a gene in order for it to be considered.

Once you launch it, it will start 2 jobs - one with a table as a result and the other with the corresponding plot as a result.


Galaxy EBI-Train
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
SINGLE CELL RNA-SEQ TOOLS
Get scRNAseq data
Seurat
SC3
Scanpy
Monocl3
Monocle3 plotCells in the reduced dimensionality space
Monocle3 top markers for cell groups
Monocle3 diffExp of genes along a trajectory
Monocle3 create a Monocle3 object from input data
Monocle3 reduceDim for downstream analysis
Monocle3 preprocess a Monocle3 object to an initially dimensionally reduced space
Monocle3 partition of cells into groups
Monocle3 orderCells along trajectories
Monocle3 learnGraph between cells in dimensionality reduced space
SCMap
SCCAF
Single Cell Utils and Viz
Favorite
Options

Monocle3 top markers for cell groups (Galaxy Version 0.1.5+galaxy0)

Input Object

7: Monocle3 partition on data 4: cds3

Input file with monocle3 object cds3 in RDS

Reference cells source

File

(Optional) Set of cells to be considered as reference for marker significance test. Accelerates the marker significance test at some cost in sensitivity. (--reference-cells)

Reference cells file

Nothing selected

File with cells to be used as reference cells. (--reference-cells)

Filter fraction for expression

0.1

Filters the markers test result by this fraction of expression (--filter-fraction-expression)

Top-n-markers

5

The number of top marker genes to report in plots and in top markers list (--top-n-markers)

Number of genes to test per group

25

how many genes of the top ranked specific genes by Jenson-Shannon, to do the more expensive regression test on (--genes-to-test-per-group)

Group cell by

cluster

Cell groups, choose from 'cluster', 'partition', or any categorical variable in `colData(cds)` (--group-cells-by)

Use logistic regression to asses discriminatory power

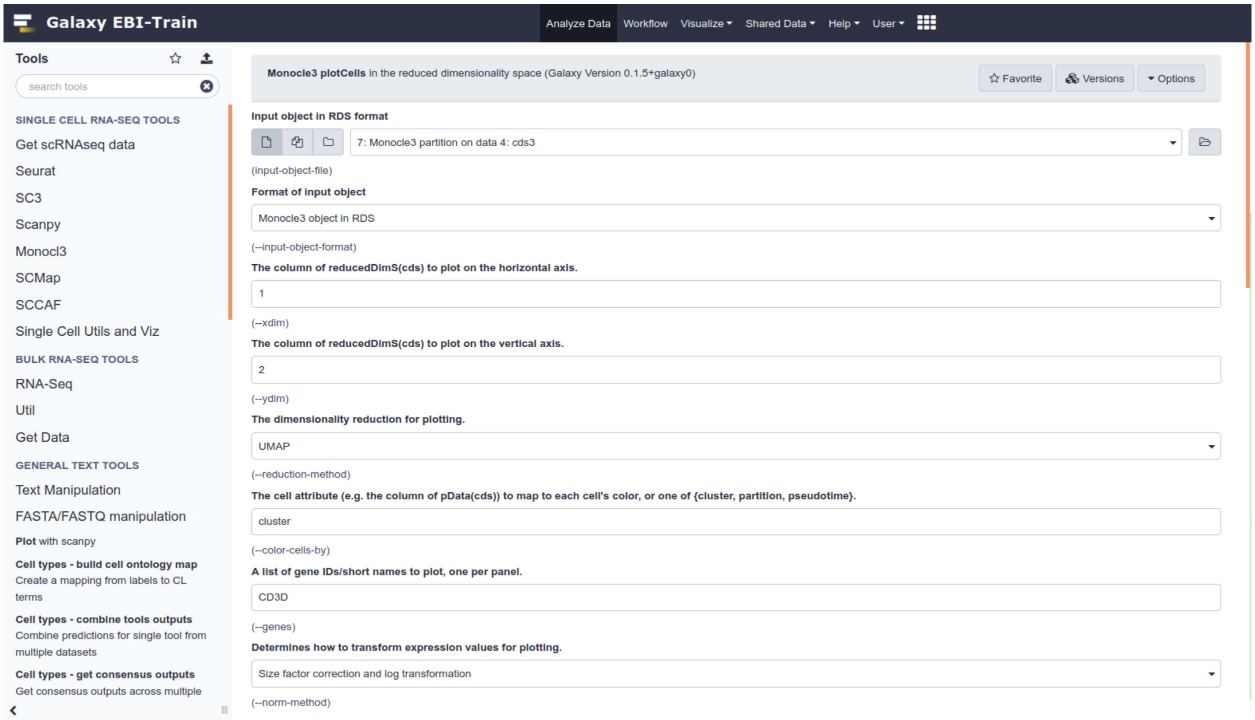
Yes [✓ No]

whether to assess the discriminative power of each marker through logistic regression. Can be slow, consider disabling to speed up top_markers(). (--marker-sig-test)

Plotting specific markers

You can use the **Monocl3 plotCells** tool to plot the expression levels of a specific gene. As the input use the UMAP clustering. Put the name of a gene in the gene ID cell.

Check out some markers from the previous step that seems interesting!


Galaxy EBI-Train
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
search tools
SINGLE CELL RNA-SEQ TOOLS
Get scRNAseq data
Seurat
SC3
Scanpy
Monocl3
SCMap
SCCAF
Single Cell Utils and Viz
BULK RNA-SEQ TOOLS
RNA-Seq
Util
Get Data
GENERAL TEXT TOOLS
Text Manipulation
FASTA/FASTQ manipulation
Plot with scanpy
Cell types - build cell ontology map
Create a mapping from labels to CL terms
Cell types - combine tools outputs
Combine predictions for single tool from multiple datasets
Cell types - get consensus outputs
Get consensus outputs across multiple
Favorite
Versions
Options

Monocle3 plotCells in the reduced dimensionality space (Galaxy Version 0.1.5+galaxy0)

Input object in RDS format
7: Monocle3 partition on data 4: cds3
(input-object-file)

Format of input object
Monocle3 object in RDS
(--input-object-format)

The column of reducedDimS(cds) to plot on the horizontal axis.
1
(--xdim)

The column of reducedDimS(cds) to plot on the vertical axis.
2
(--ydim)

The dimensionality reduction for plotting.
UMAP
(--reduction-method)

The cell attribute (e.g. the column of pData(cds)) to map to each cell's color, or one of {cluster, partition, pseudotime}.
cluster
(--color-cells-by)

A list of gene IDs/short names to plot, one per panel.
CD3D
(--genes)

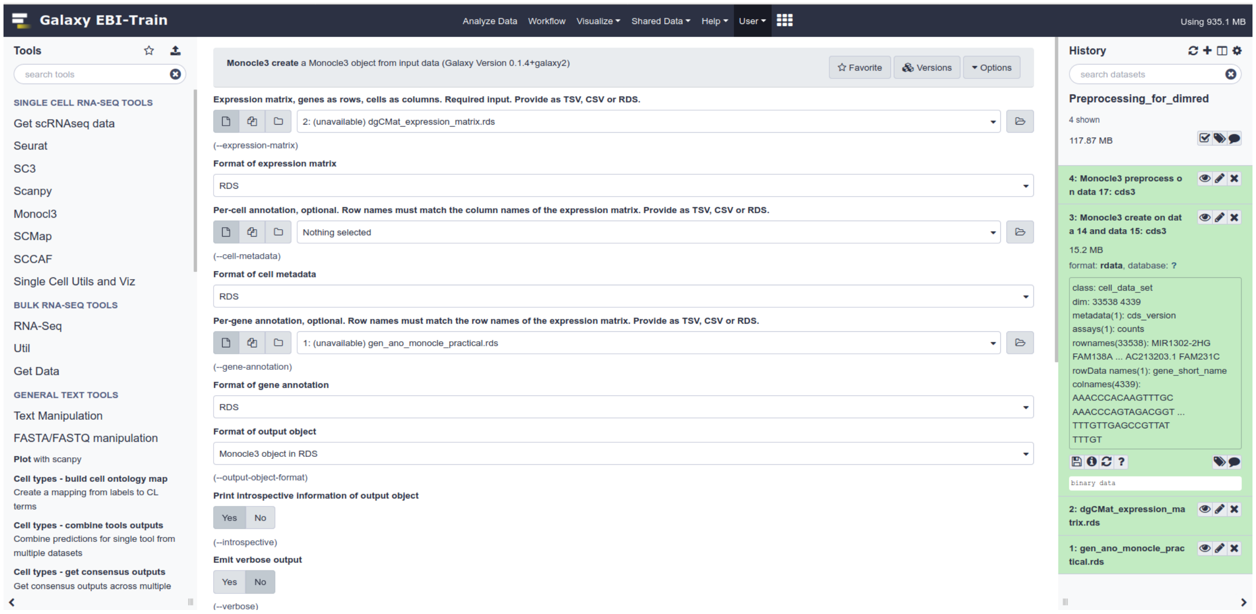
Determines how to transform expression values for plotting.
Size factor correction and log transformation
(--norm-method)

Preprocessing (optional)

If you have data you would like to use as input to the pipeline above, the preprocessing pipeline is currently located at the EBI-Train server here.

The raw data you need for it is a count matirx - expression of genes by cells, and an annotation data frame in R made of a column of gene names. Both of those should be found within this repository.

First you must create the Monocle object with **Monocl3 create**. The inputs are the expression matrix file and the gene annotation file. If there is cell metadata you would like to add, you can add that information too.



Monocle3 create a Monocle3 object from input data (Galaxy Version 0.1.4+galaxy2)

Expression matrix, genes as rows, cells as columns. Required input. Provide as TSV, CSV or RDS.
2: (unavailable) dgCMat_expression_matrix.rds
(--expression-matrix)

Format of expression matrix
RDS

Per-cell annotation, optional. Row names must match the column names of the expression matrix. Provide as TSV, CSV or RDS.
Nothing selected
(--cell-metadata)

Format of cell metadata
RDS

Per-gene annotation, optional. Row names must match the row names of the expression matrix. Provide as TSV, CSV or RDS.
1: (unavailable) gen_ano_monocle_practical.rds
(--gene-annotation)

Format of gene annotation
RDS

Format of output object
Monocle3 object in RDS
(--output-object-format)

Print introspective information of output object
[✓ Yes] No
(--introspective)

Emit verbose output
Yes [✓ No]
(--verbose)

Preprocessing_for_dimred
4 shown
117.87 MB

4: Monocle3 preprocess on data 17: cds3

3: Monocle3 create on data 14 and data 15: cds3
15.2 MB
format: rdata, database: ?

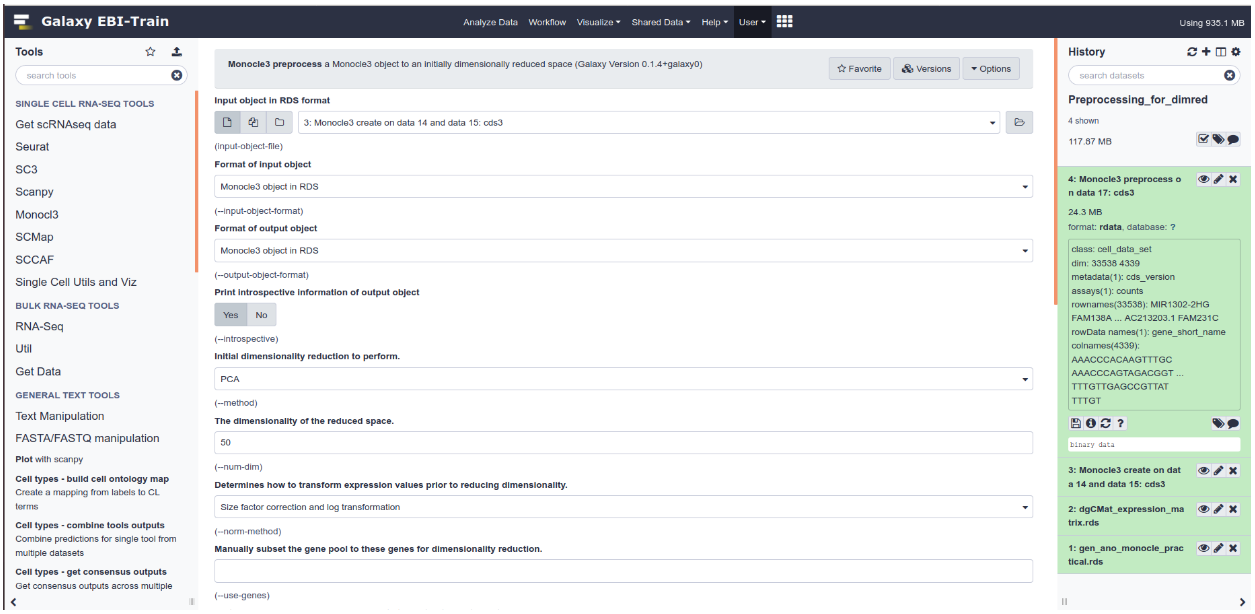
```
class: cell_data_set
dim: 33538 4339
metadata(1): cds_version
assays(1): counts
rownames(33538): MIR1302-2HG
FAM138A ... AC213203.1 FAM231C
rowData names(1): gene_short_name
colnames(4339):
AAACCCACAAGTTTGC
AAACCCAGTAGACGGT ...
TTTGTTGAGCCGTTAT
TTTGT
```

binary data

2: dgCMat_expression_matrix.rds

1: gen_ano_monocle_practical.rds

Use the **Monocl3 preprocess** on the resulting object. You can change the number of principal components it uses in order to speed it up or reduce the memory size, but normally that isn't ecessary.



Monocle3 preprocess a Monocle3 object to an initially dimensionally reduced space (Galaxy Version 0.1.4+galaxy0)

Input object in RDS format
3: Monocle3 create on data 14 and data 15: cds3
(input-object-file)

Format of input object
Monocle3 object in RDS
(--input-object-format)

Format of output object
Monocle3 object in RDS
(--output-object-format)

Print introspective information of output object
[✓ Yes] No
(--introspective)

Initial dimensionality reduction to perform.
PCA
(--method)

The dimensionality of the reduced space.
50
(--num-dim)

Determines how to transform expression values prior to reducing dimensionality.
Size factor correction and log transformation
(--norm-method)

Manually subset the gene pool to these genes for dimensionality reduction.

(--use-genes)

Preprocessing_for_dimred
4 shown
117.87 MB

4: Monocle3 preprocess on data 17: cds3
24.3 MB
format: rdata, database: ?

```
class: cell_data_set
dim: 33538 4339
metadata(1): cds_version
assays(1): counts
rownames(33538): MIR1302-2HG
FAM138A ... AC213203.1 FAM231C
rowData names(1): gene_short_name
colnames(4339):
AAACCCACAAGTTTGC
AAACCCAGTAGACGGT ...
TTTGTTGAGCCGTTAT
TTTGT
```

binary data

3: Monocle3 create on data 14 and data 15: cds3

2: dgCMat_expression_matrix.rds

1: gen_ano_monocle_practical.rds