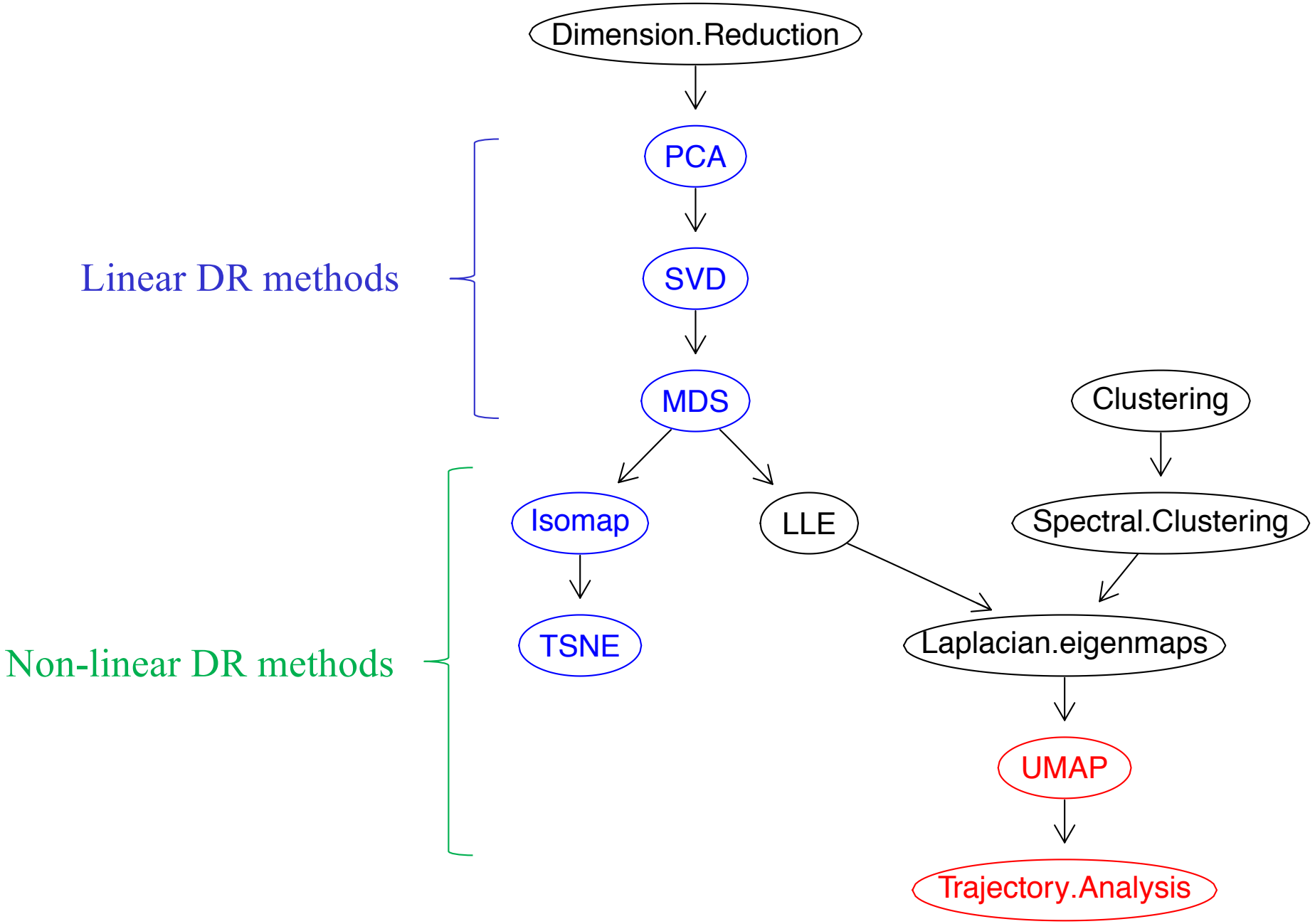
Dimension Reduction Methods:
From PCA to TSNE and UMAP

Maxwell Lee

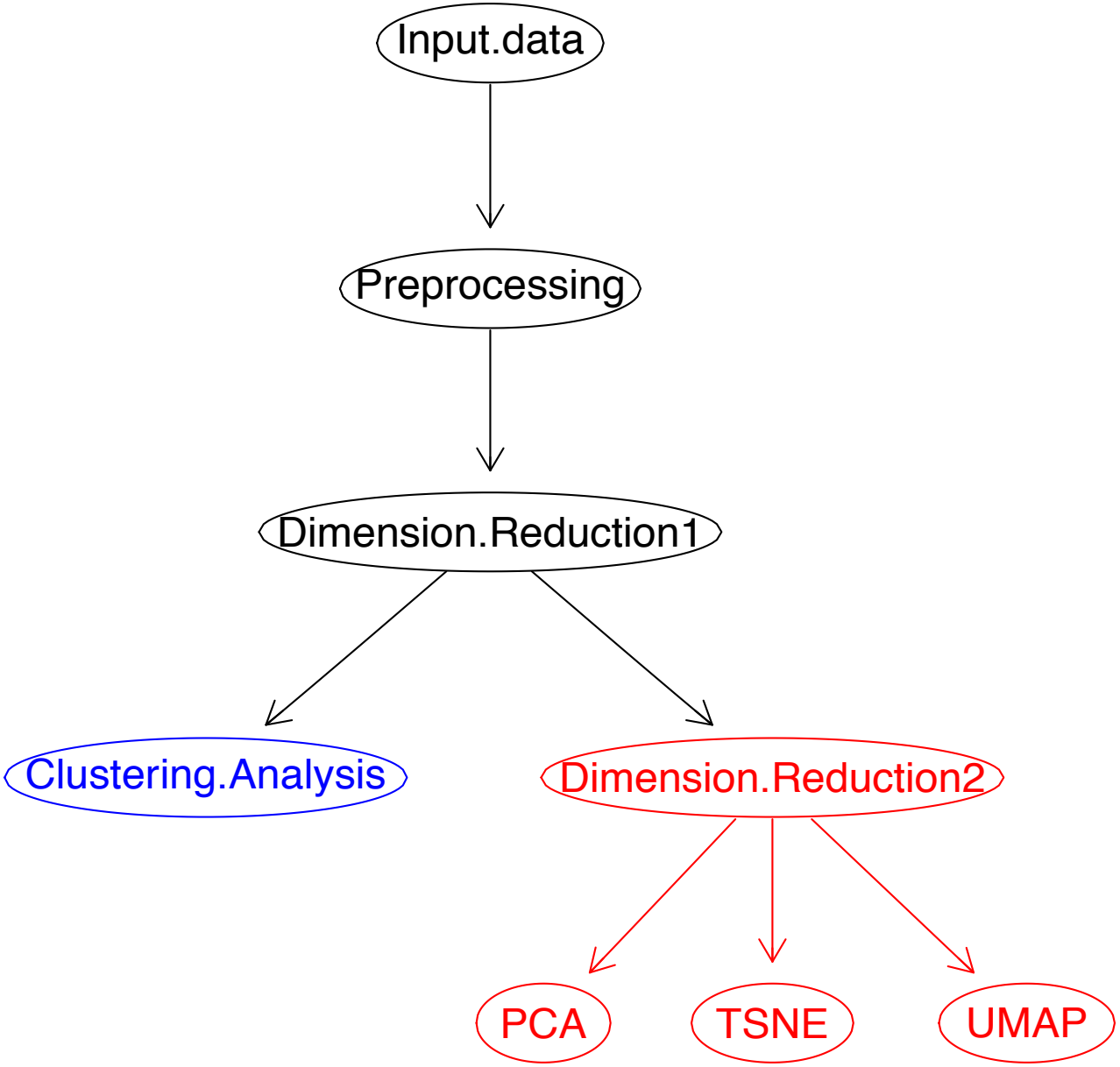
High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

May 28, 2020

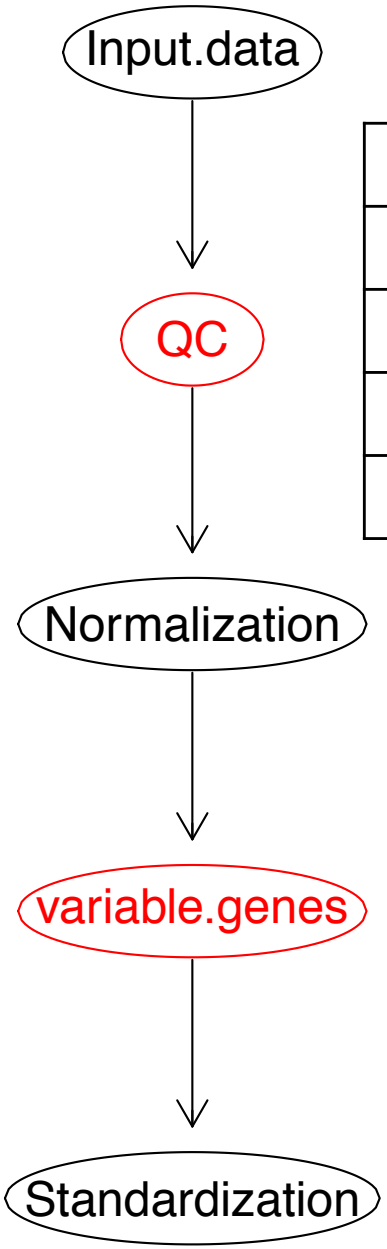
Road Map for Dimension Reduction Methods



Flow Chart of ScRNAseq Analyses with Seurat Package



Preprocessing Steps in Seurat Package

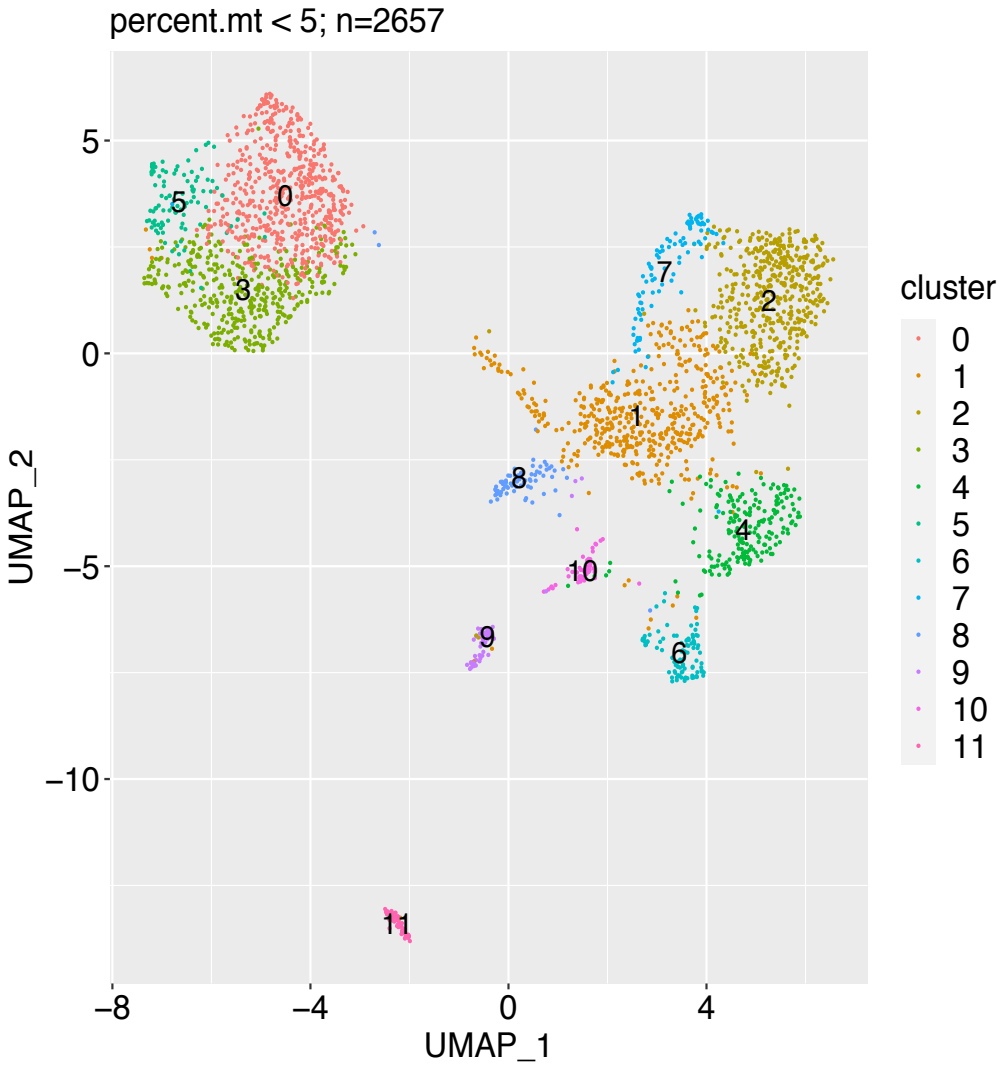


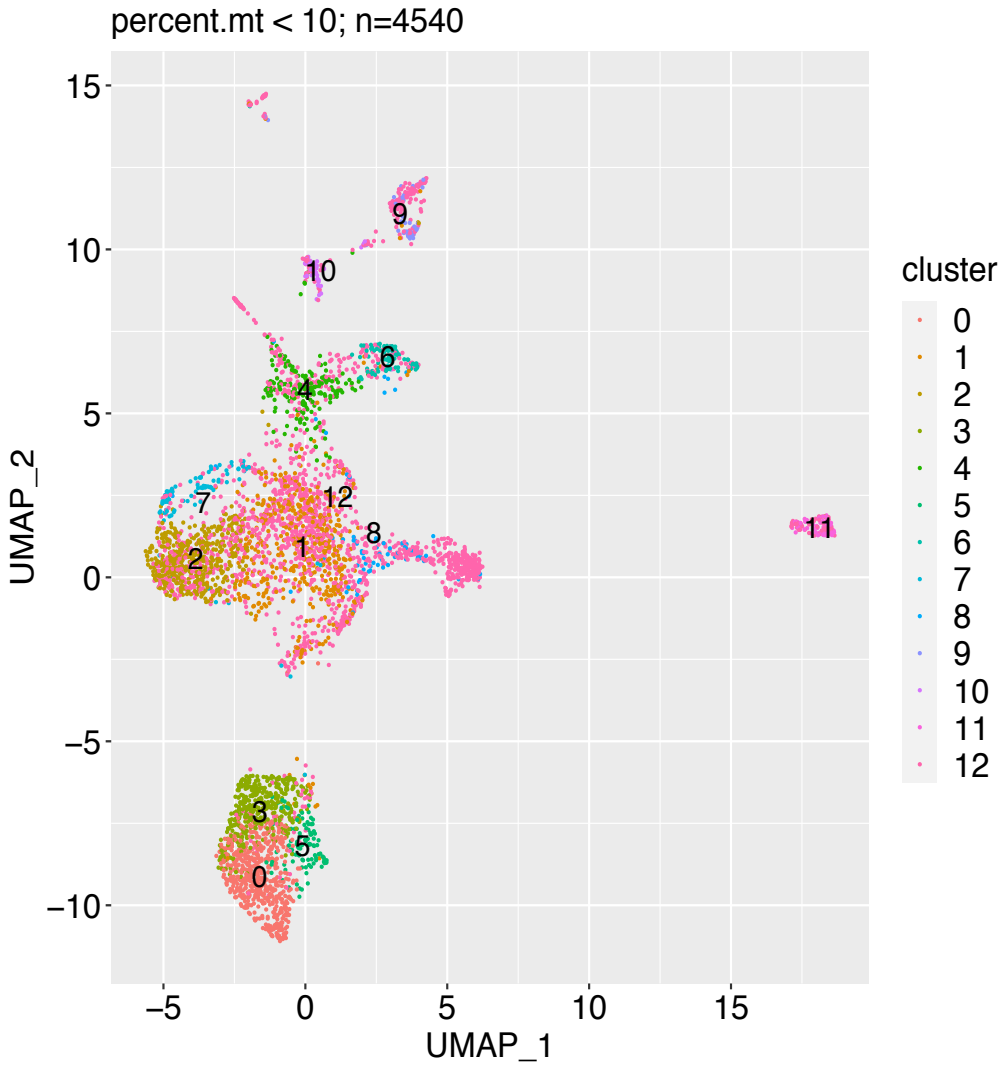
Preprocessing	function	Description
QC	Select cells	percent.mt < 5%
Normalization	Normalizing cells	TP10K
Variable genes	Most variable genes	nfeatures = 2000
Standardization	Standardization across cells	z score

Effects of Using Percent of Mitochondrial Gene Cutoff on UMAP

Clustering and Dimension Reduction 2

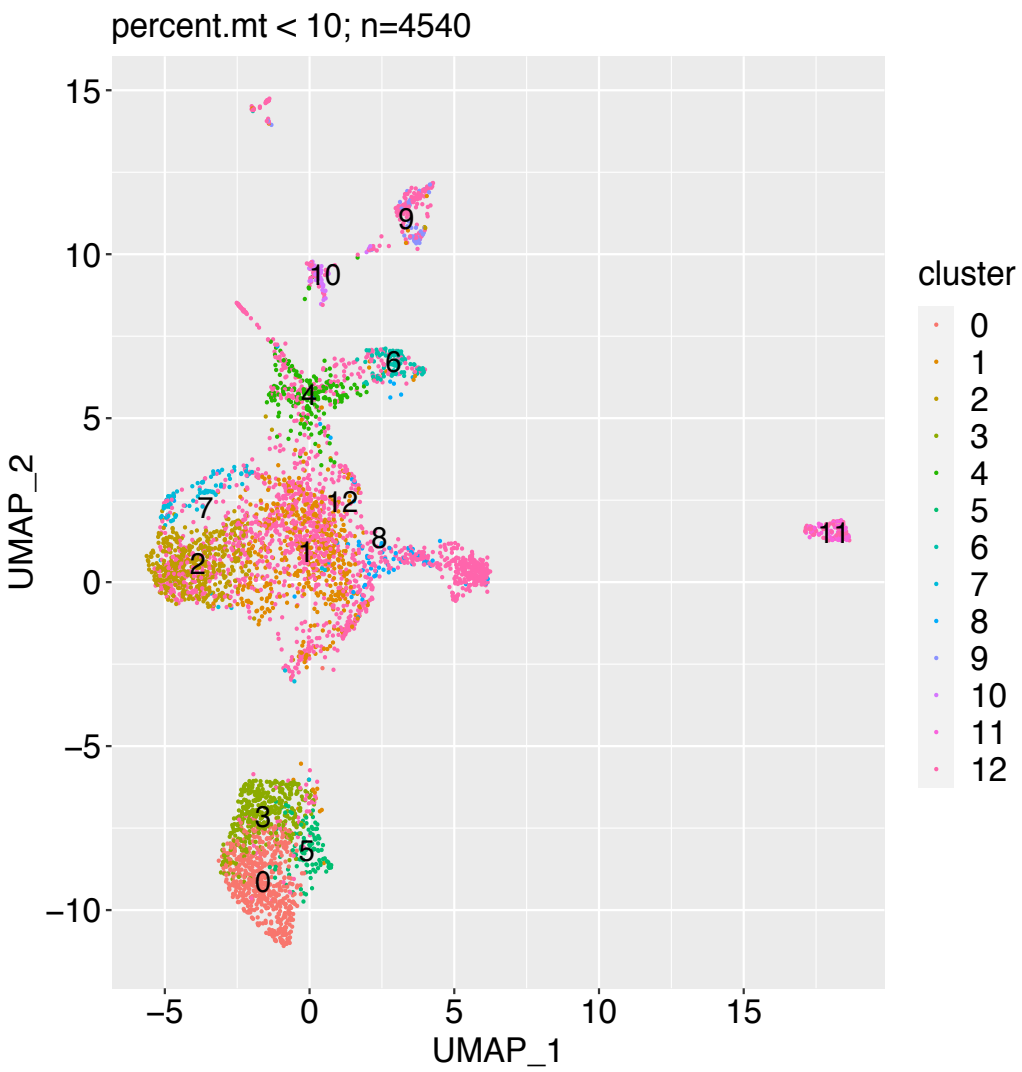
Dimension Reduction 2
Clusters 0-11 are identical to the left plot (percent.mt < 5)
Cluster 12 has the additional cells (percent.mt between 5 and 10)



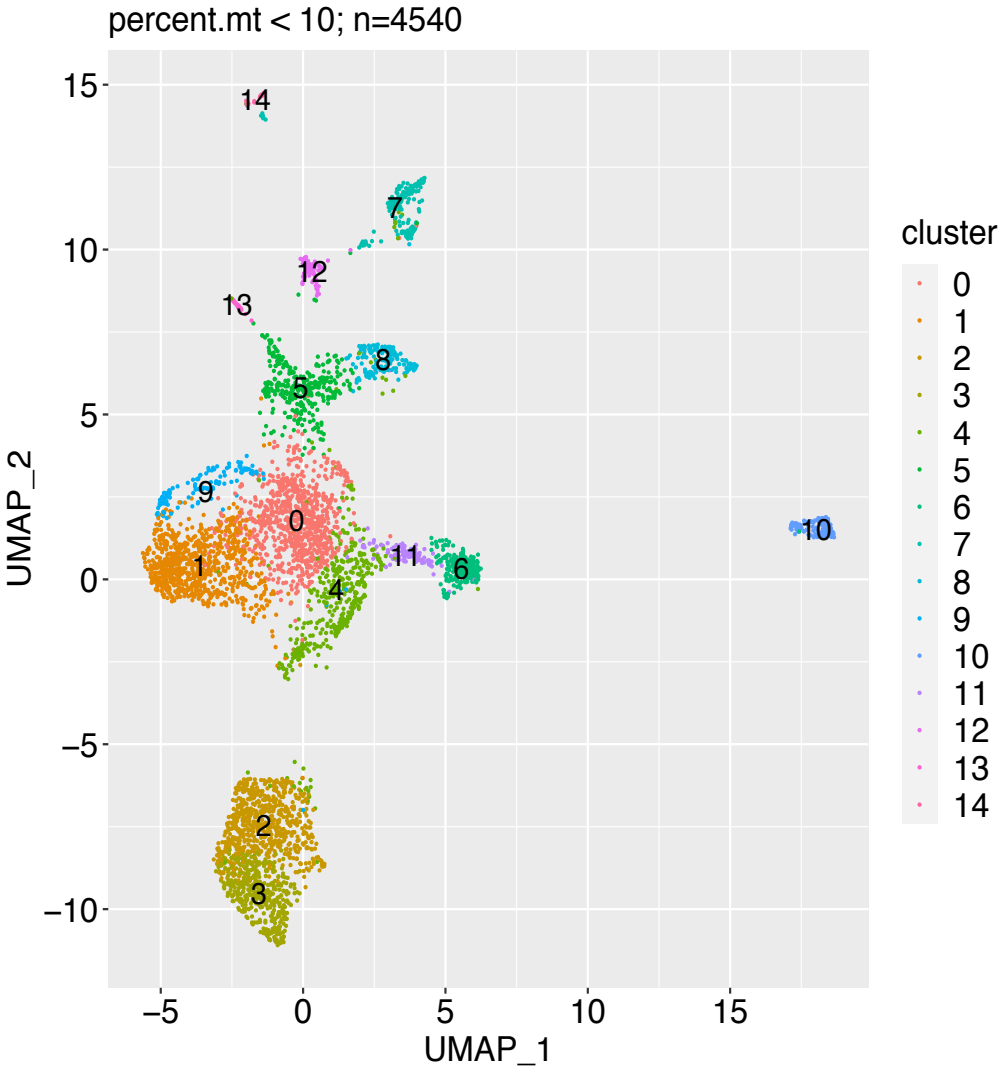


How Is UMAP Affected by New Clustering Analysis?

Dimension Reduction 2

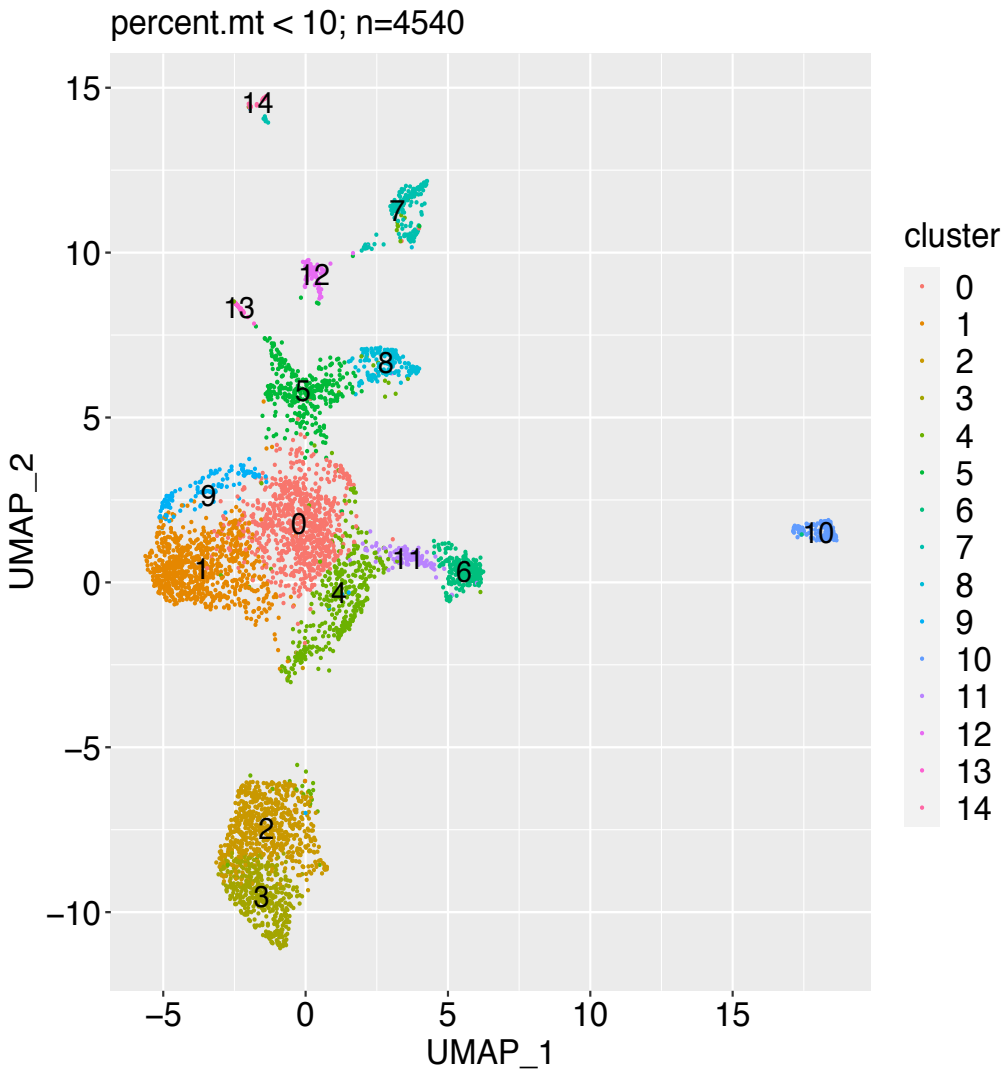


Clustering and Dimension Reduction 2

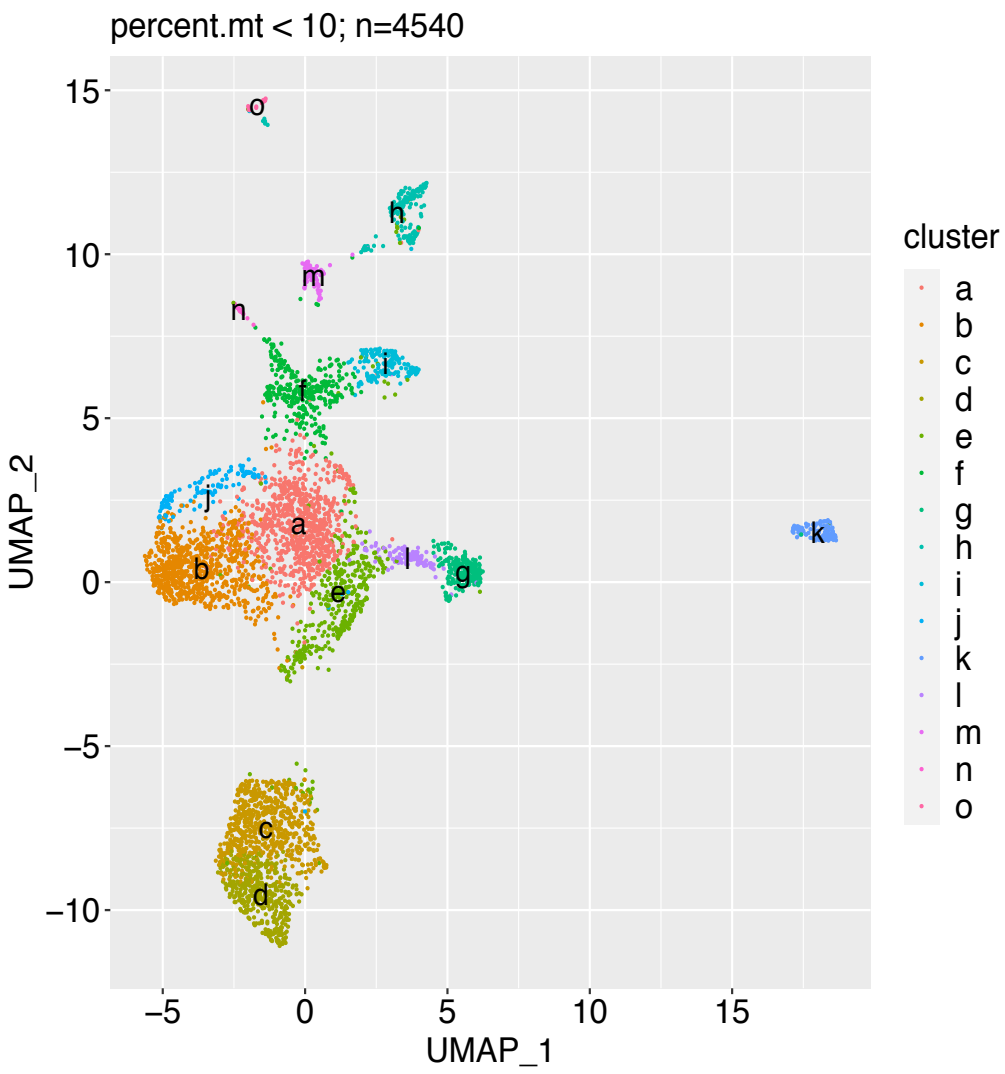


How Is UMAP Affected by New Clustering Analysis?

Clustering and Dimension Reduction 2

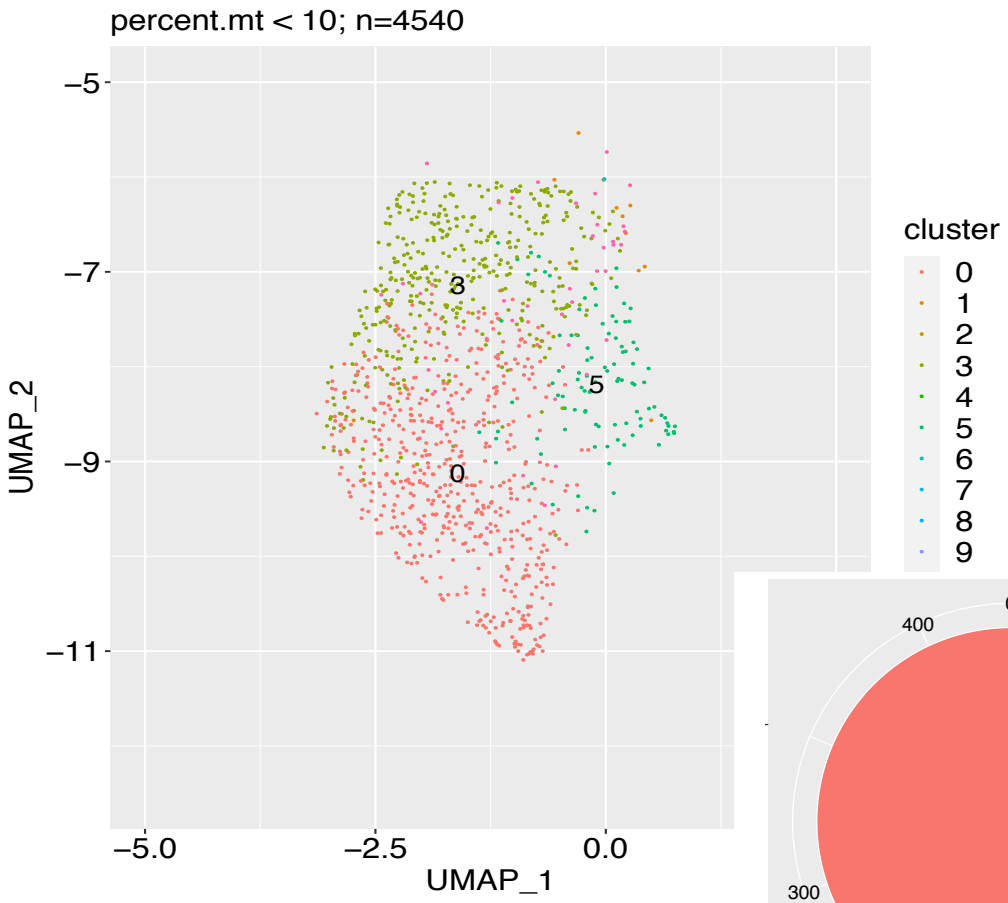


Clustering and Dimension Reduction 2

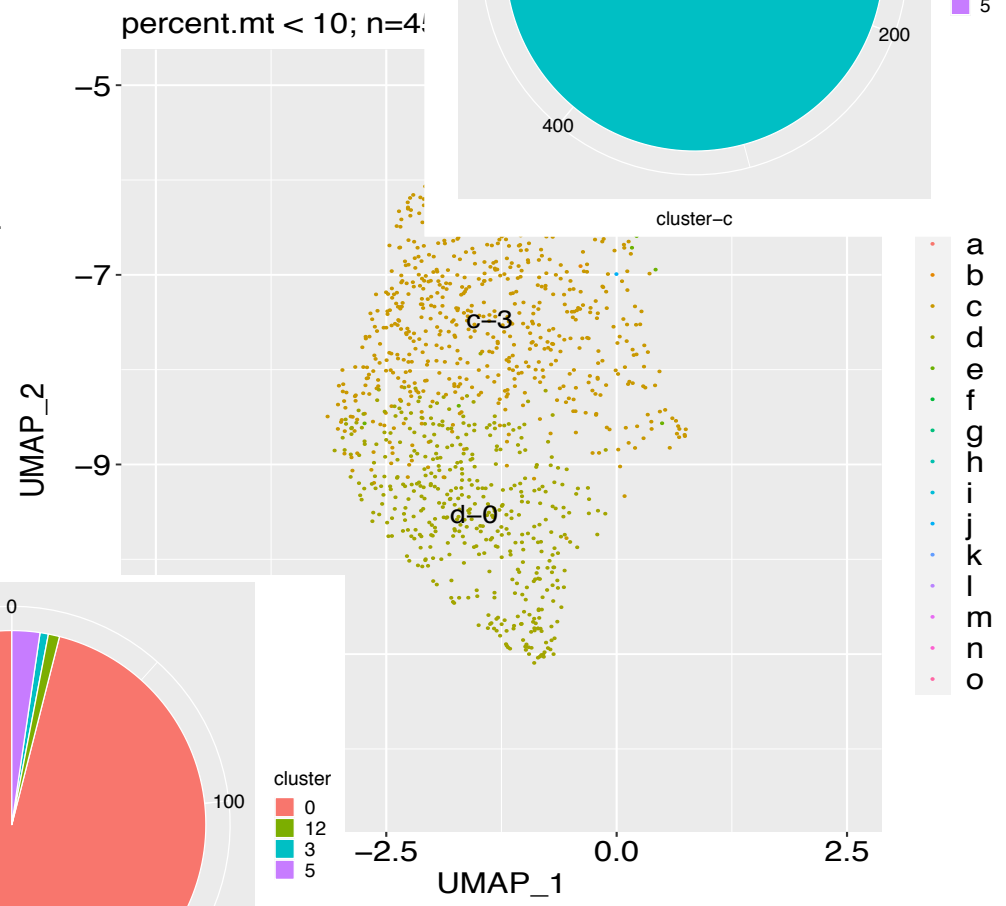


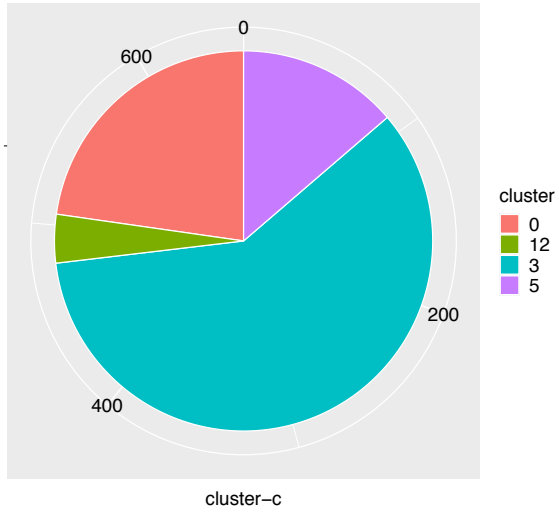
How Is UMAP Affected by New Clustering Analysis?

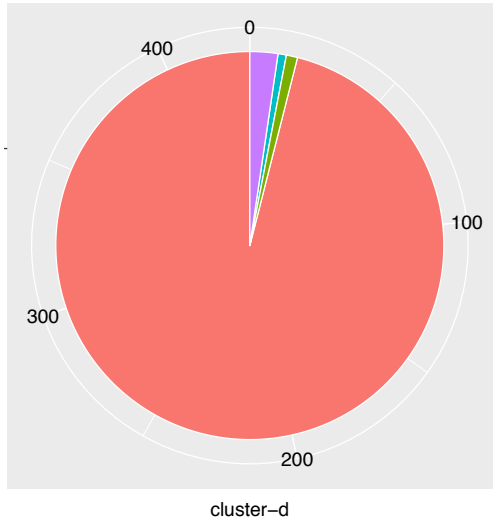
Clustering and Dimension Reduction 2



Clustering and D
Label with alphal



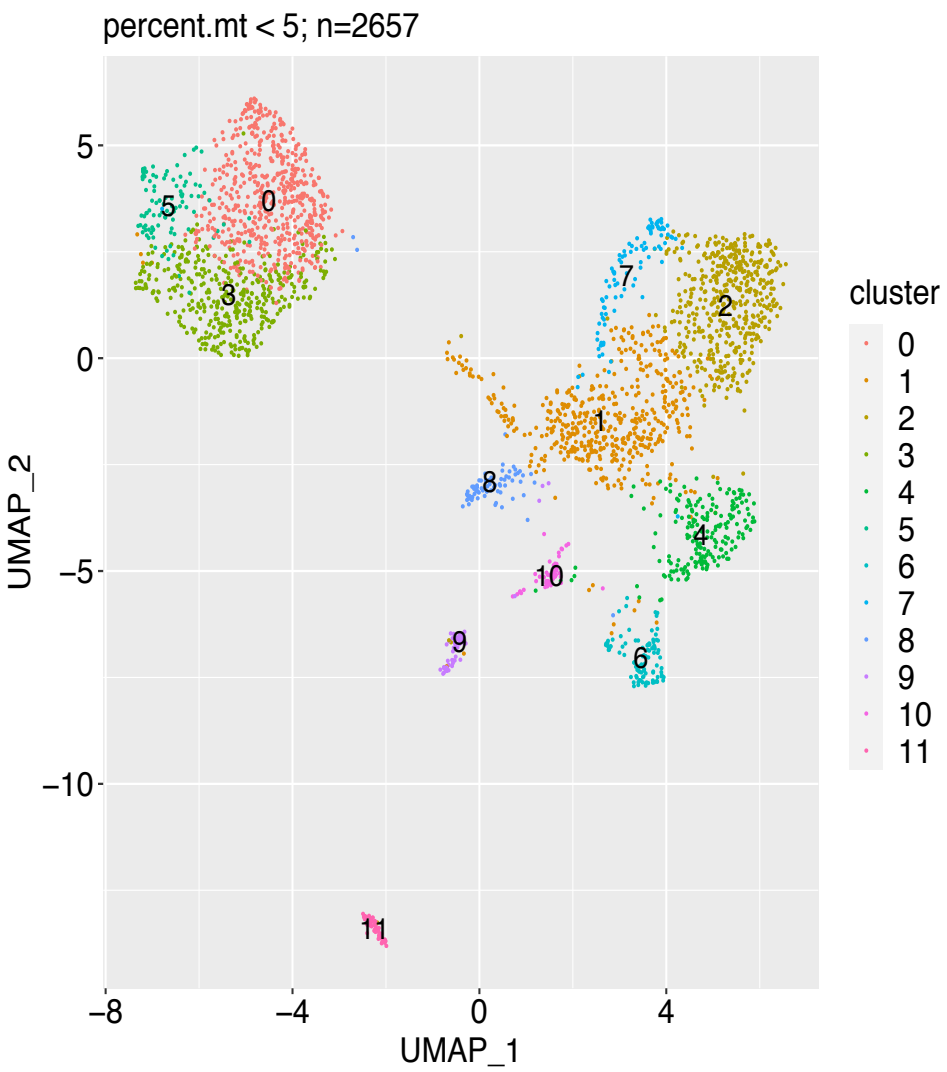




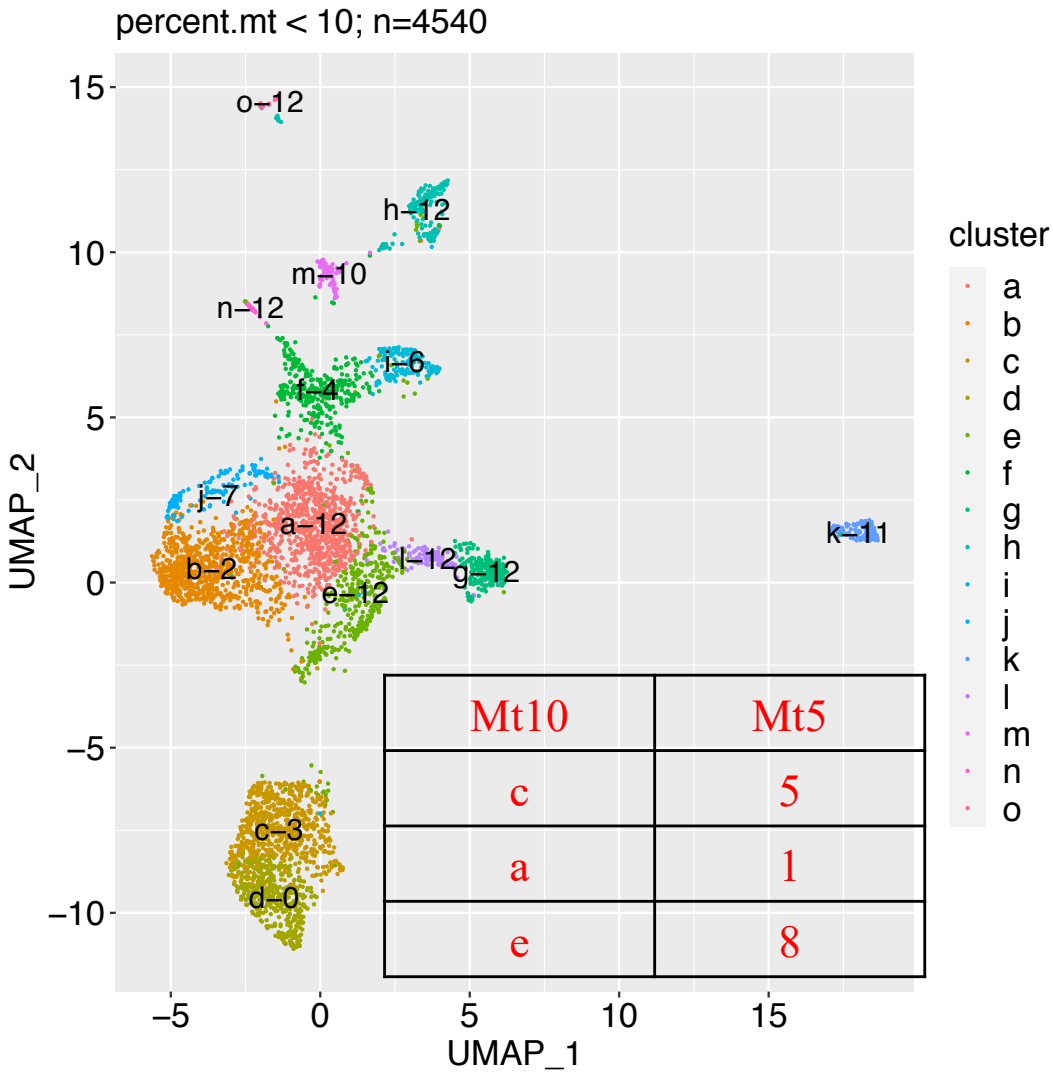
Numbers outside of the pie chart are numbers of cells

How Is UMAP Affected by New Clustering Analysis?

Clustering and Dimension Reduction 2

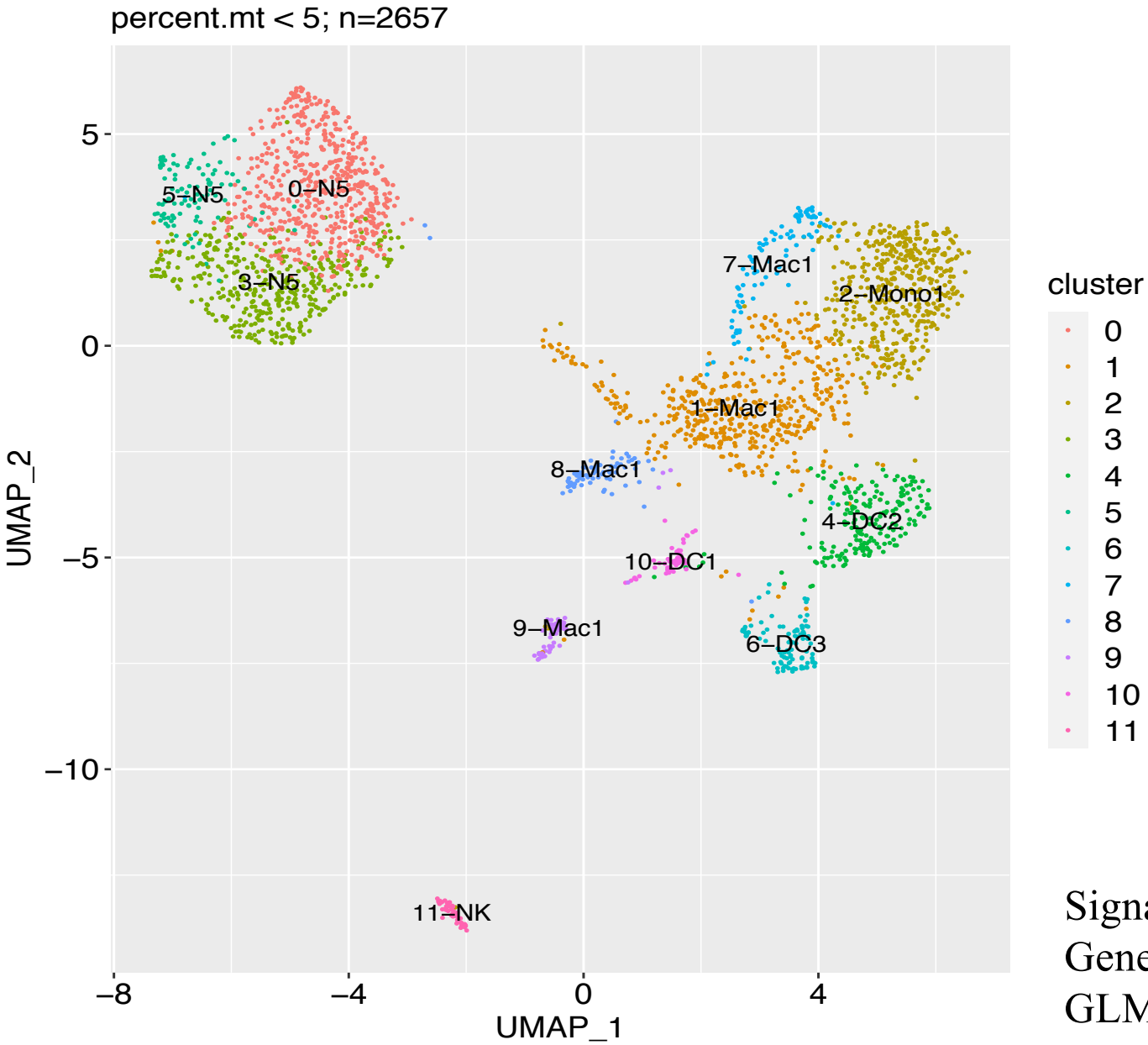


Clustering and Dimension Reduction 2
Label with alphabet and its mapping

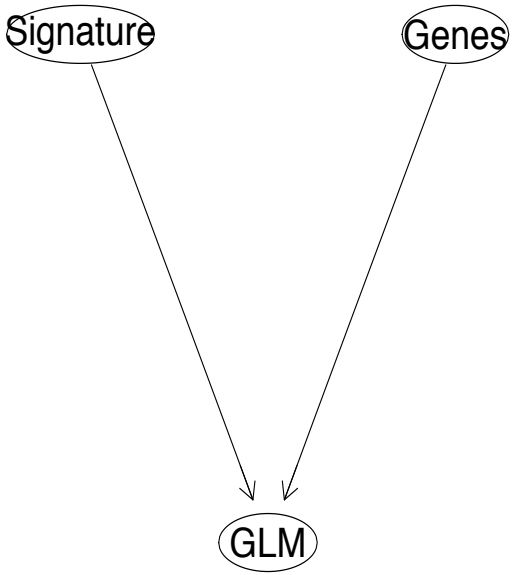


Mt10	Mt5
c	5
a	1
e	8

Tissue Subtype of Clusters with Percent.mt5



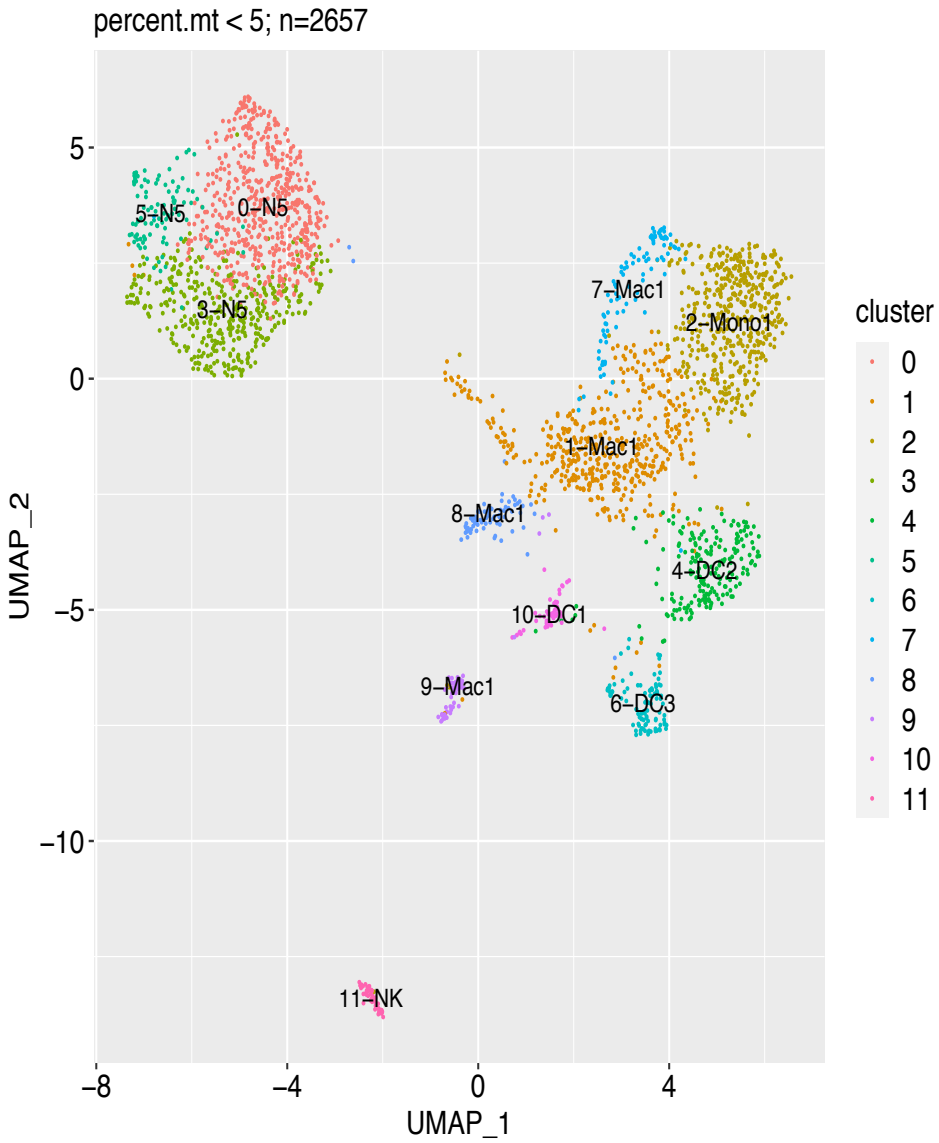
DC: dendritic cells
Mac: macrophage
Mono: monocyte
N: neutrophils
NK: natural killer cells

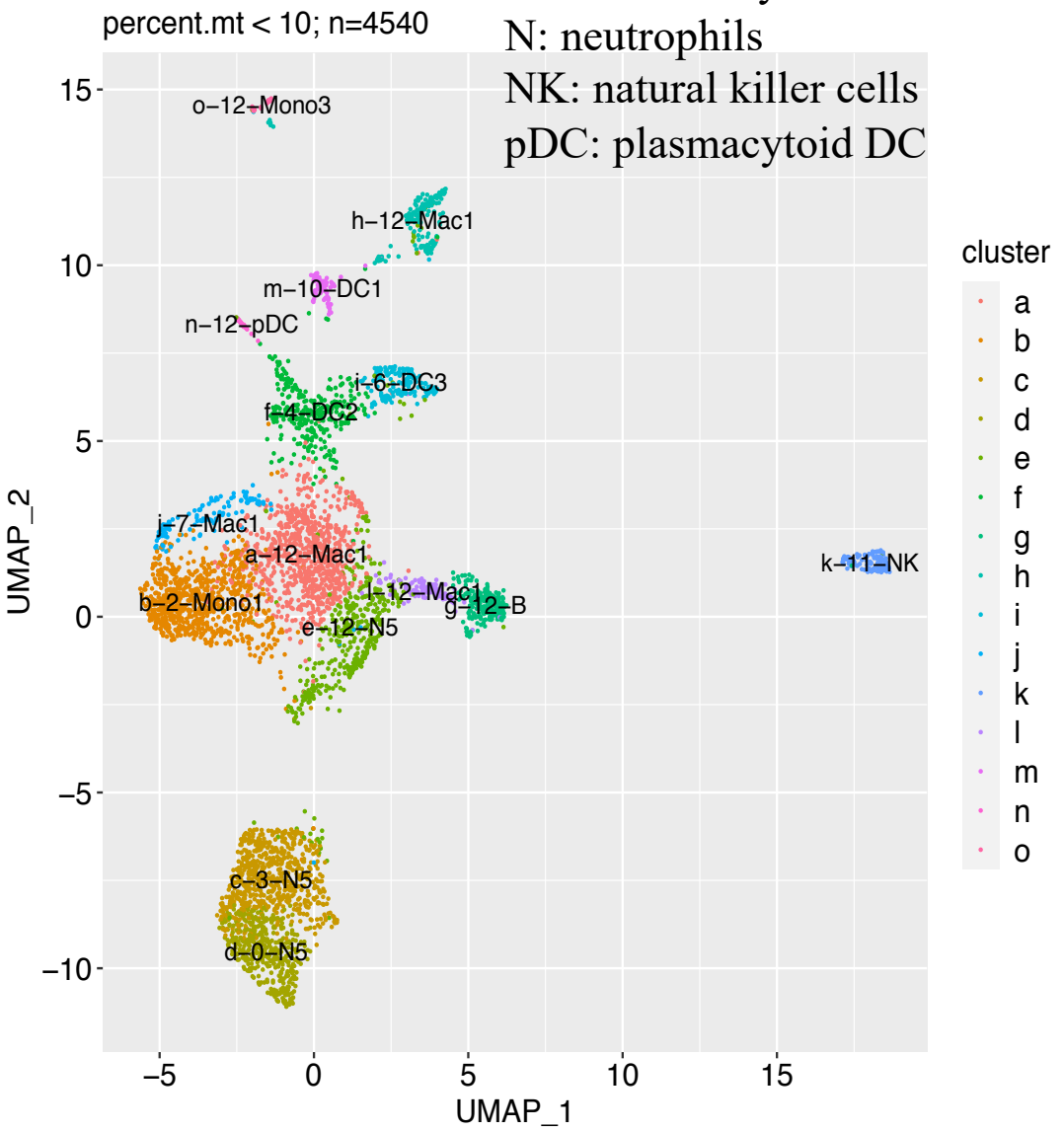


Signature: gene signatures
Genes: genes from Cibersort
GLM: generalized linear model

Rapolas Zilionis..Allon Klein 2019 Immunity 50, 1317-1334

Tissue Subtype of Clusters: Percent.mt5 vs Percent.mt10



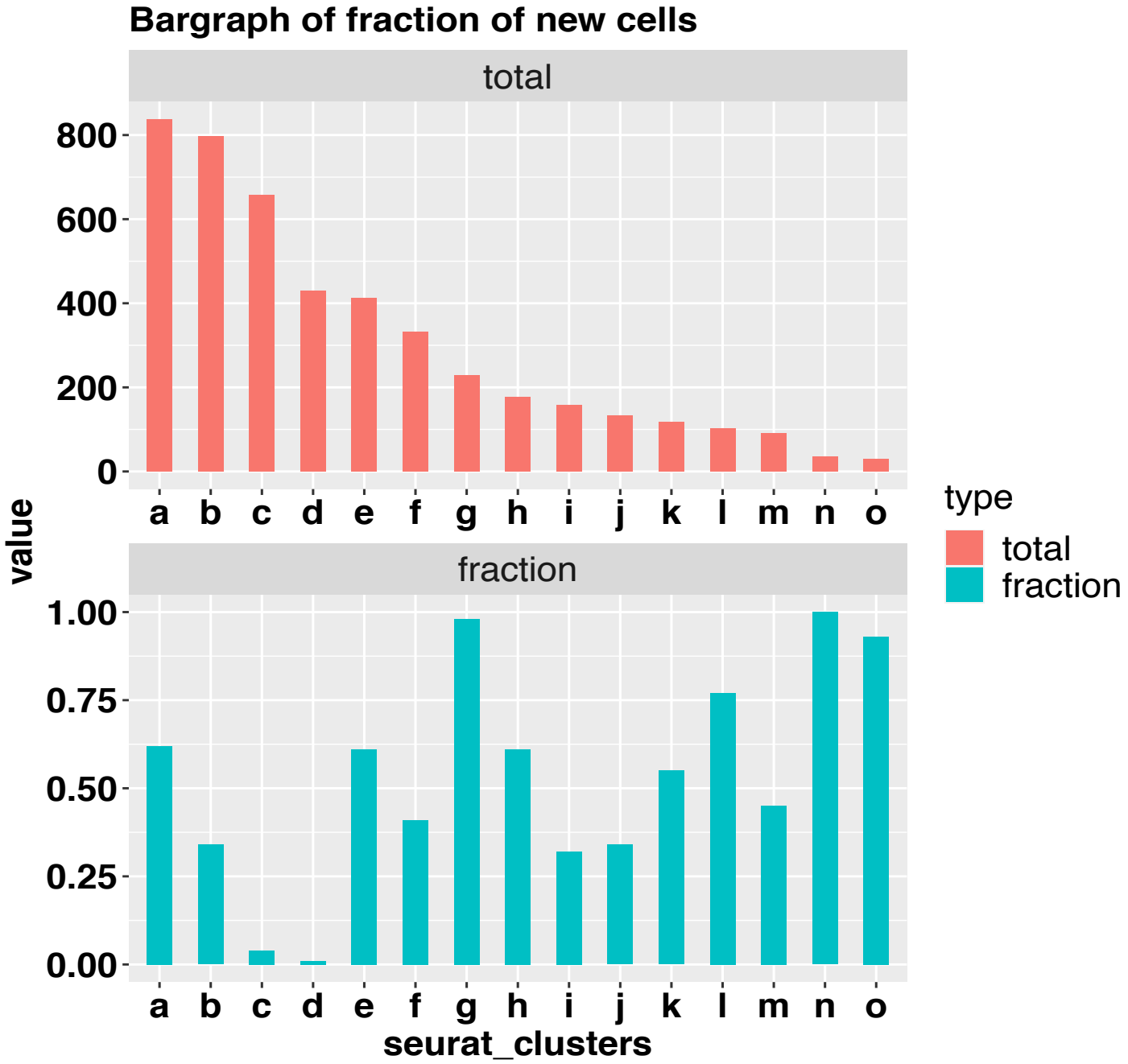


B: B cells
DC: dendritic cells
Mac: macrophage
Mono: monocyte
N: neutrophils
NK: natural killer cells
pDC: plasmacytoid DC

Comparison of Cluster Tissue Subtypes

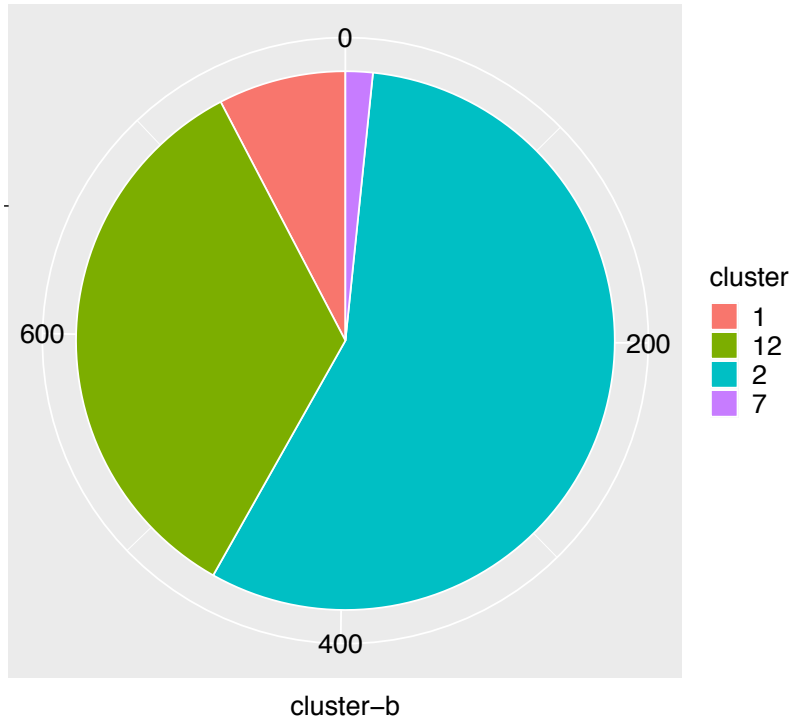
mt10.tissue	cluster	mt5.tissue
Mac1	a-12	
Mono1	b-2	Mono1
N5	c-3	N5
N5	d-0	N5
N5	e-12	
DC2	f-4	DC2
B	g-12	
Mac1	h-12	
DC3	i-6	DC3
Mac1	j-7	Mac1
NK	k-11	NK
Mac1	l-12	
DC1	m-10	DC1
pDC	n-12	
Mono3	o-12	

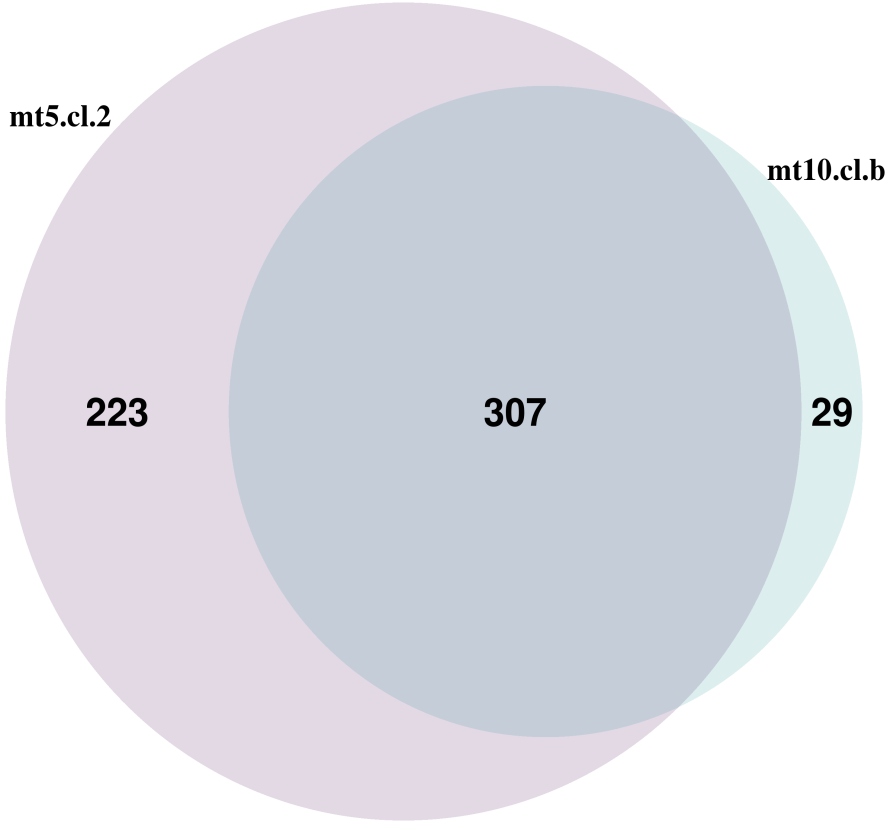
Fraction of the Cells with Percent.mt Larger Than 5



Mono1: Cluster b (Mt10) vs Cluster 2 (Mt5) Specific Genes

mt5.cl.2: mt5 cluster 2
mt10.cl.b: mt10 cluster b

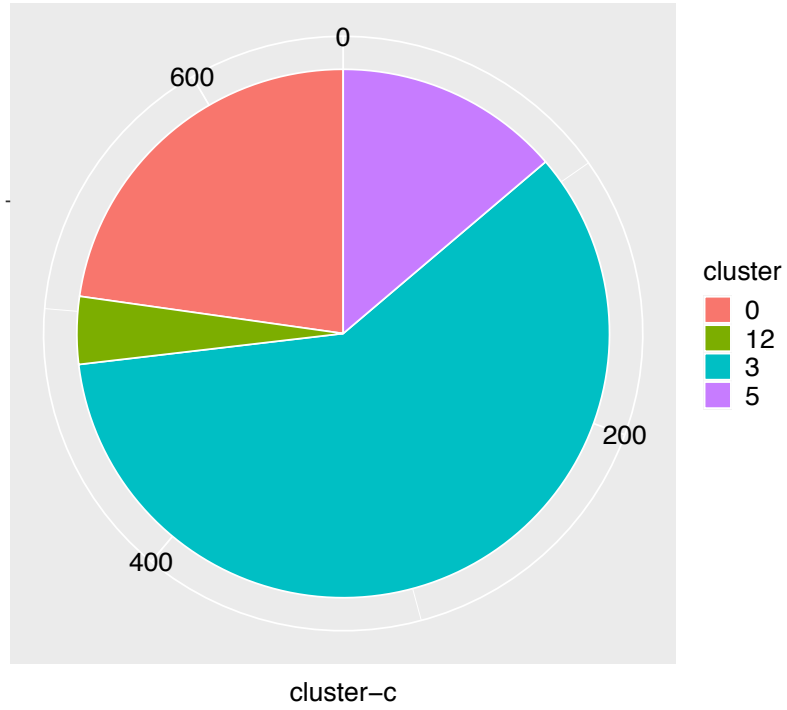


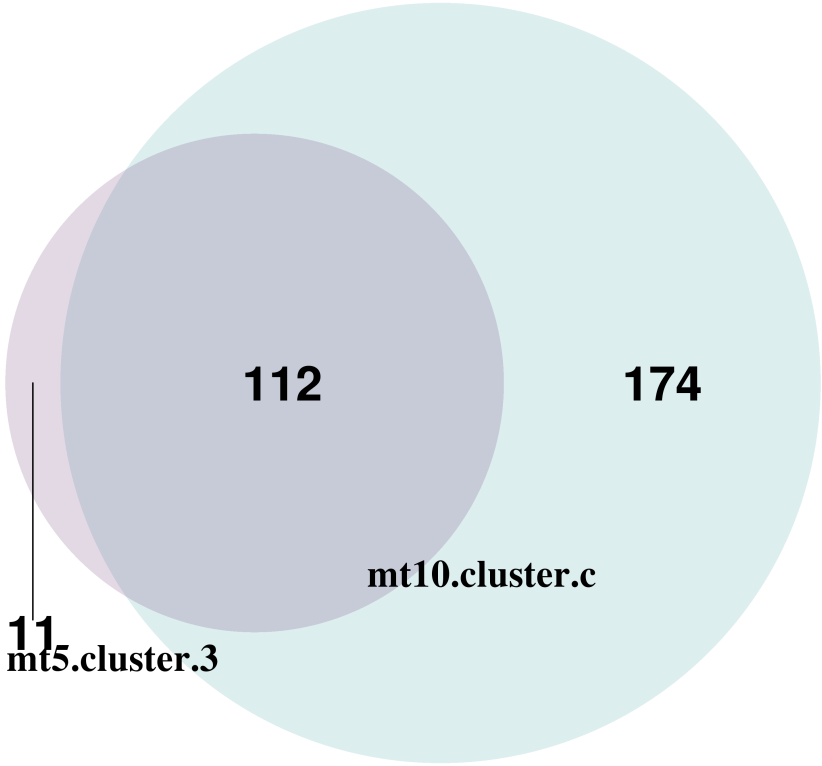


Numbers outside of the pie chart
are numbers of cells

Cluster-specific genes
Common: 307 genes
Unique to mt5 cluster 2: 223 genes
Unique to mt10 cluster b: 29 genes

N5: Cluster c (Mt10) vs Cluster 3 (Mt5) Specific Genes

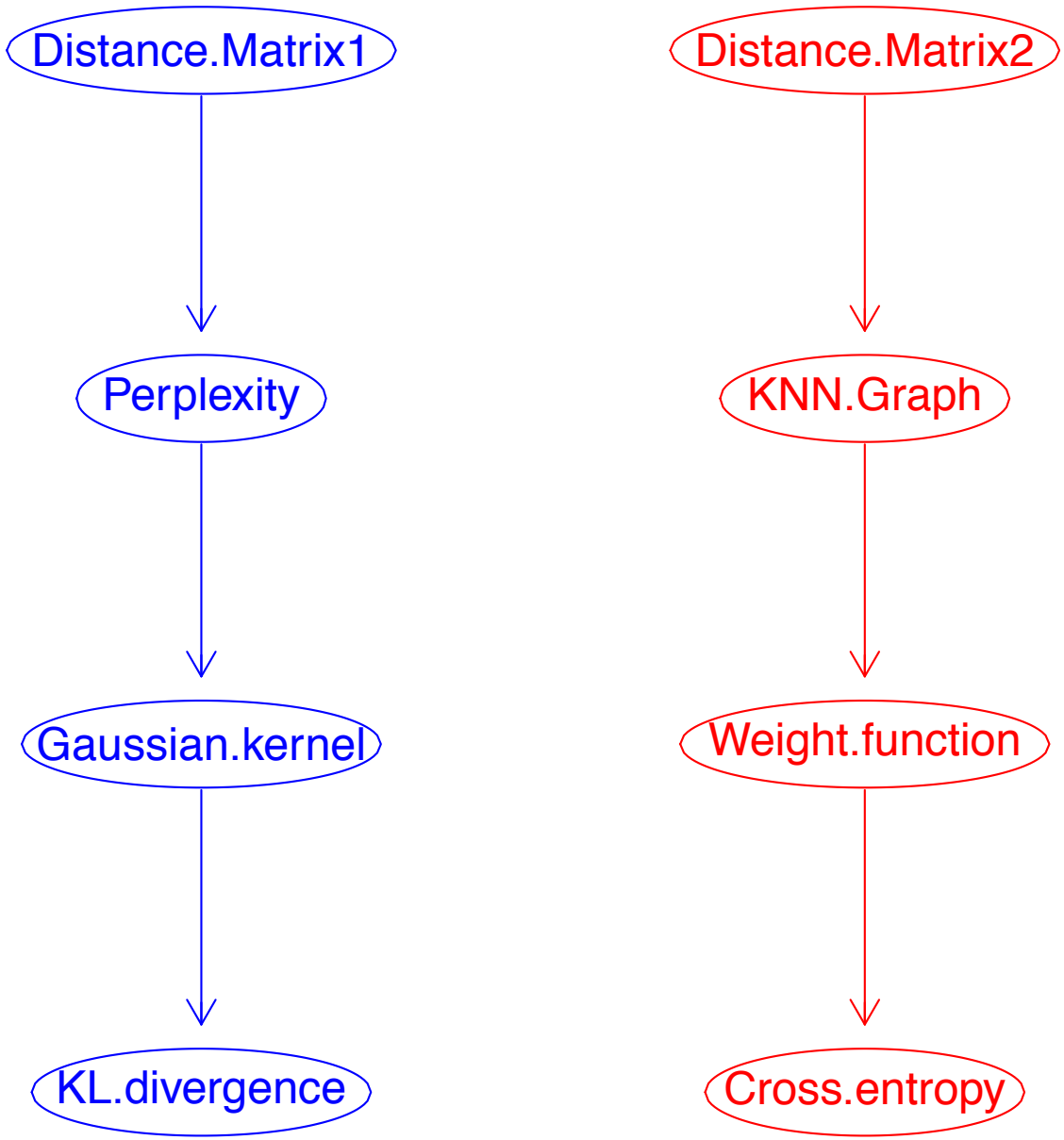




Numbers outside of the pie chart are numbers of cells

Cluster-specific genes
Common: 112 genes
Unique to mt5 cluster 0: 11 genes
Unique to mt10 cluster d: 174 genes

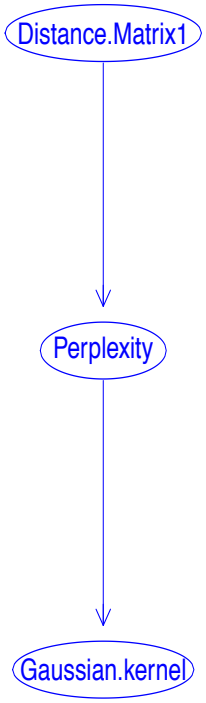
TSNE vs. UMAP


Distance.Matrix1
Perplexity
Gaussian.kernel
KL.divergence
Distance.Matrix2
KNN.Graph
Weight.function
Cross.entropy

TSNE vs. UMAP

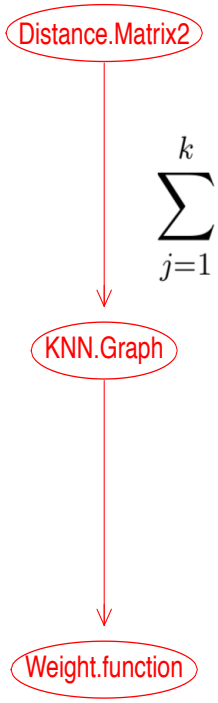
$$Perp(P_i) = 2^{H(P_i)}$$

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$



$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$



$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right) = \log_2(k)$$

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

$$B = A + A^\top - A \circ A^\top$$

ρ_i: shortest distance of x_i neighbors

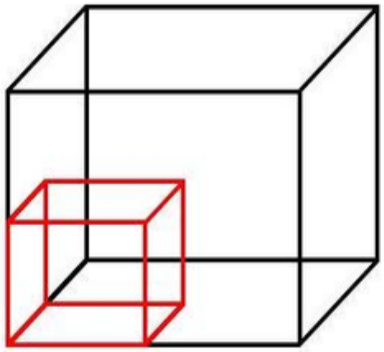
Euclidean Distance and Other Distance Metrics

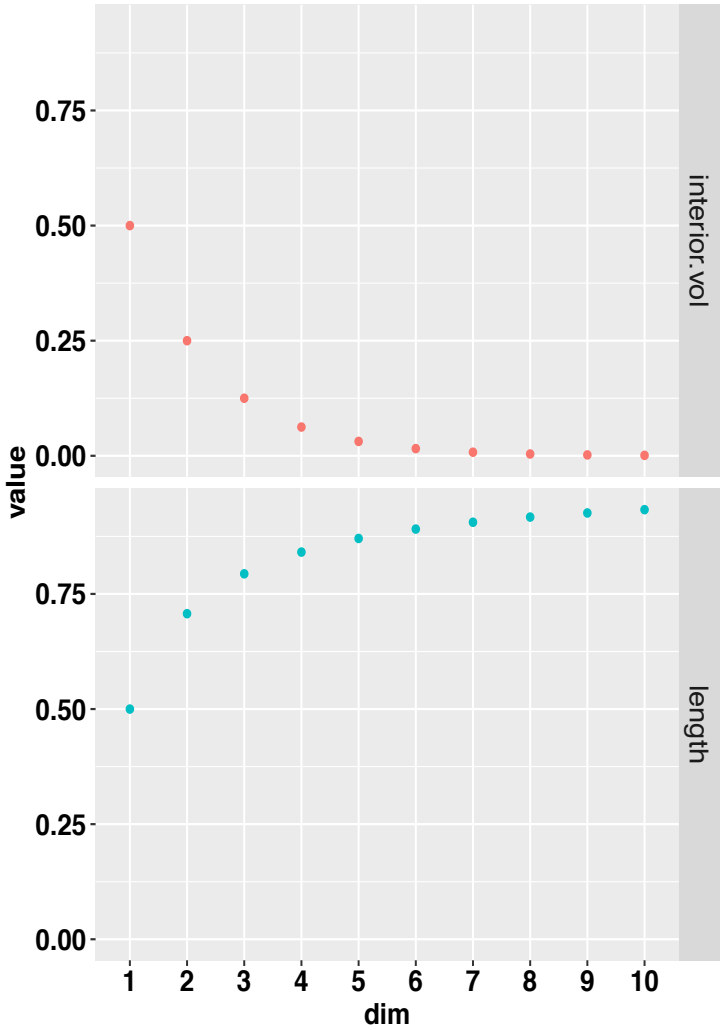
- Euclidean distance vs geodesic distance
- Euclidean distance vs Mahalanobis distance
- Curse of dimensionality

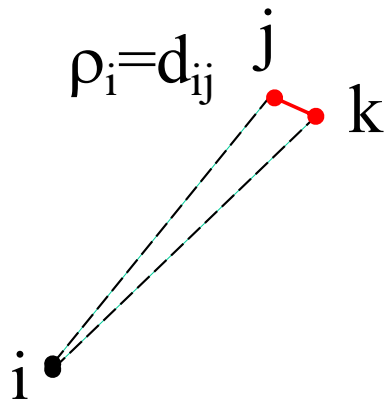
Curse of Dimensionality

(I) 50% of each dimension is sufficient to cover 25% of a 2-dimensional space

(II) 50% of each dimension is only sufficient to cover 12.5% of a 3-dimensional space







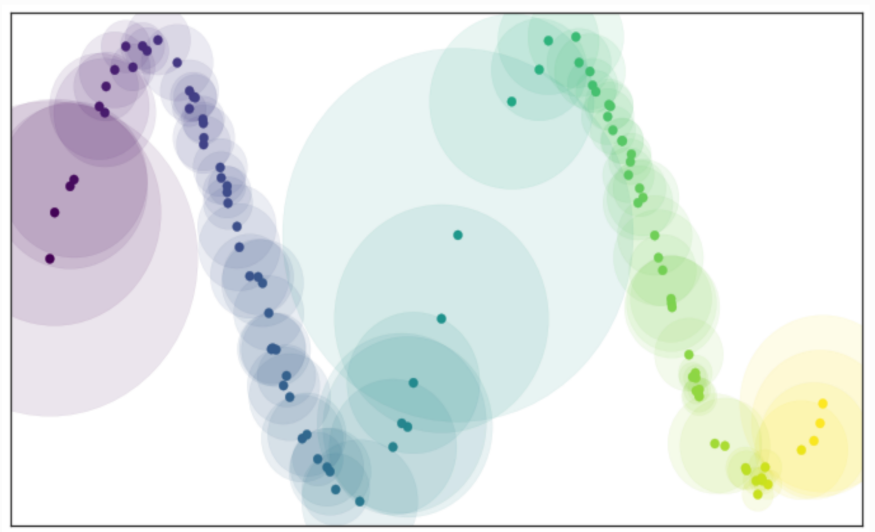
type
- interior.vol
- length

Uniform Manifold Approximation and Projection (UMAP)

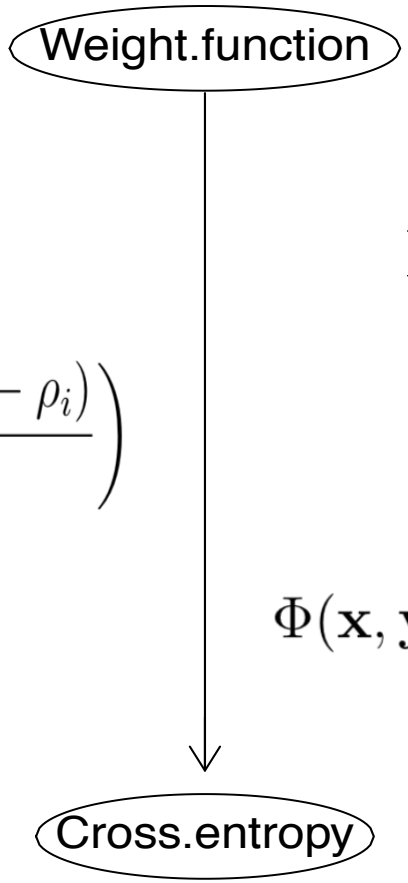
$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

$$B = A + A^\top - A \circ A^\top$$

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right) = \log_2(k)$$



Uniform Manifold Approximation and Projection (UMAP)



High-dimension

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

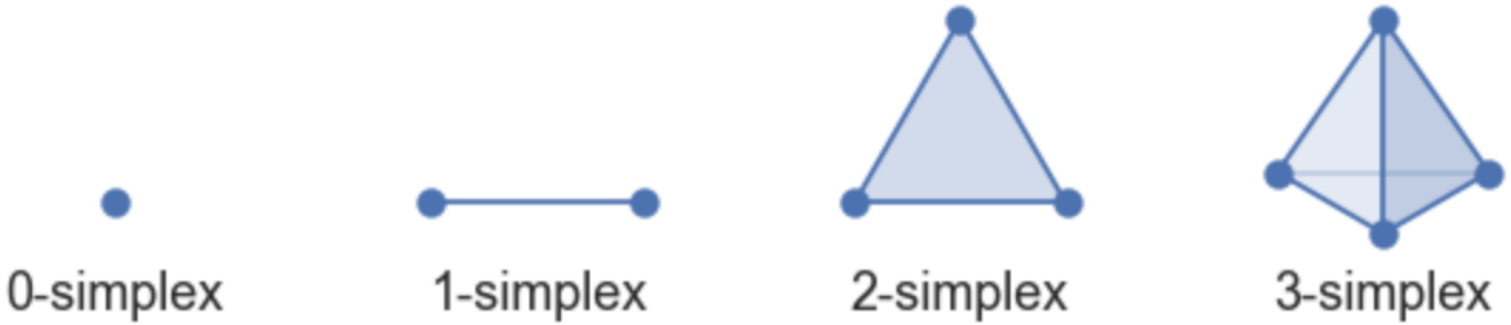
$$B = A + A^\top - A \circ A^\top$$

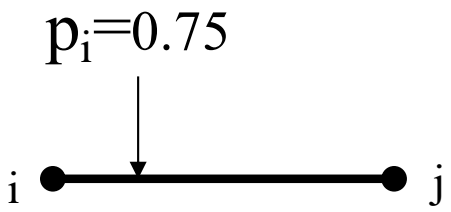
Low-dimension

Laplacian Eigenmaps

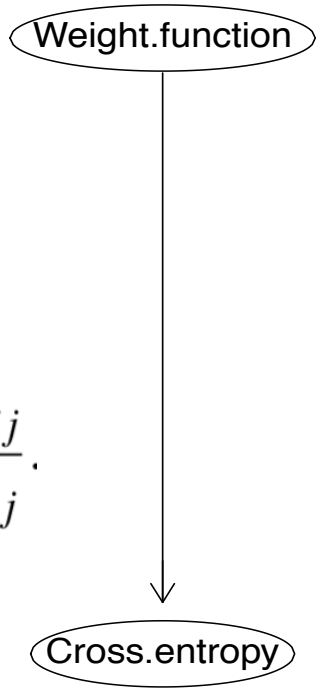
$$\Phi(\mathbf{x}, \mathbf{y}) = \left(1 + a(\|\mathbf{x} - \mathbf{y}\|_2^2)^b\right)^{-1}$$

Fuzzy Simplicial Sets





Uniform Manifold Approximation and Projection (UMAP)



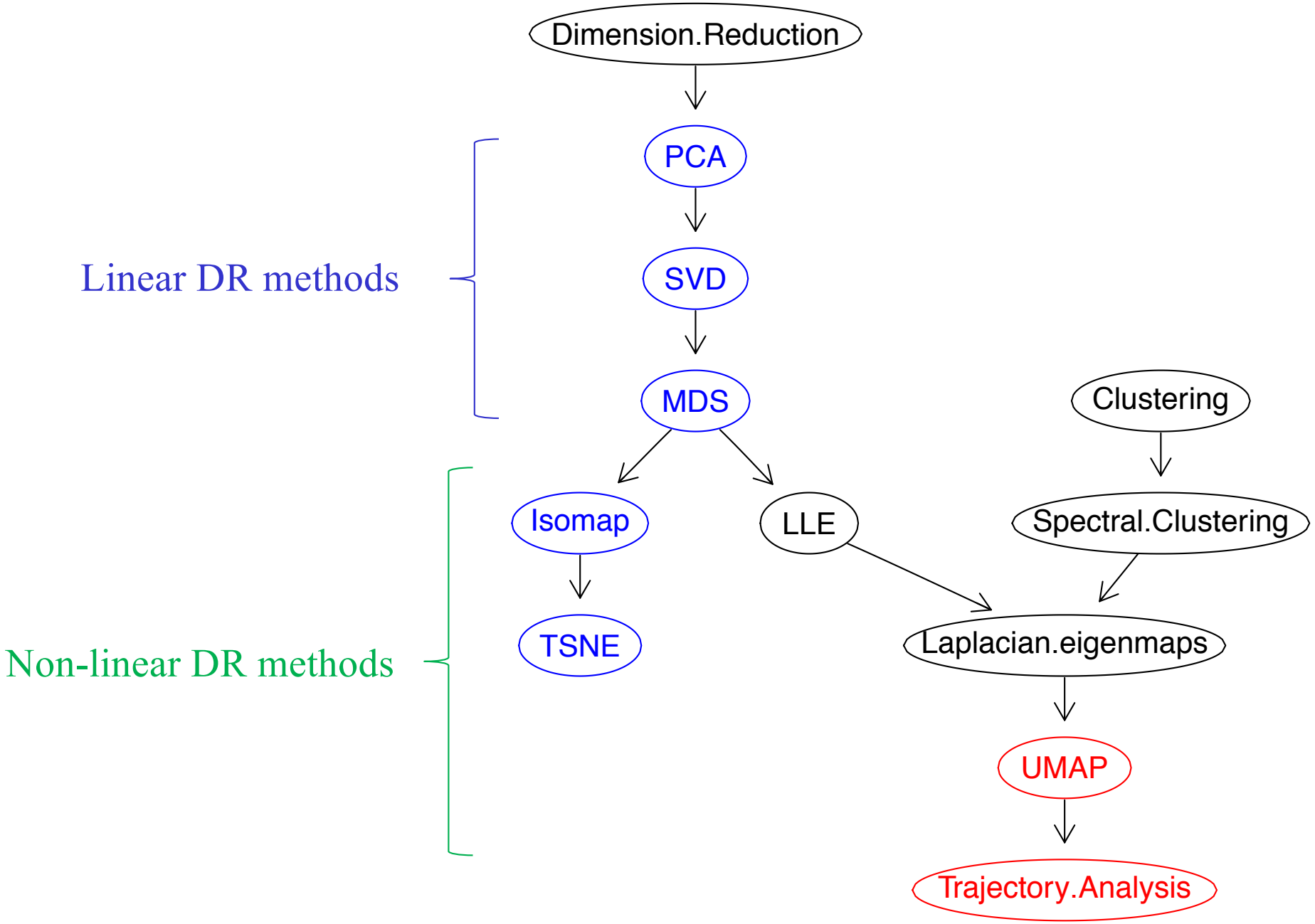
TSNE cost function

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

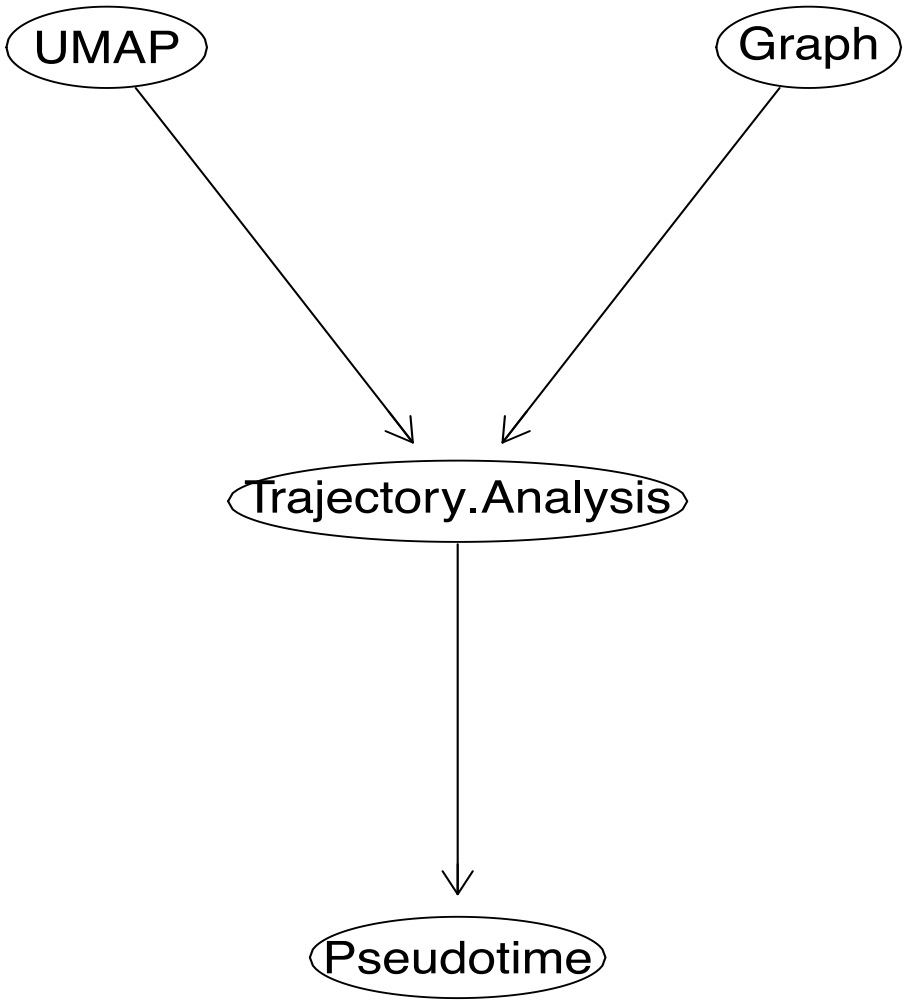
UMAP cost function

$$C((A,\mu),(A,\nu)) = \sum_{a \in A} \mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right) + (1-\mu(a)) \log \left(\frac{1-\mu(a)}{1-\nu(a)} \right)$$

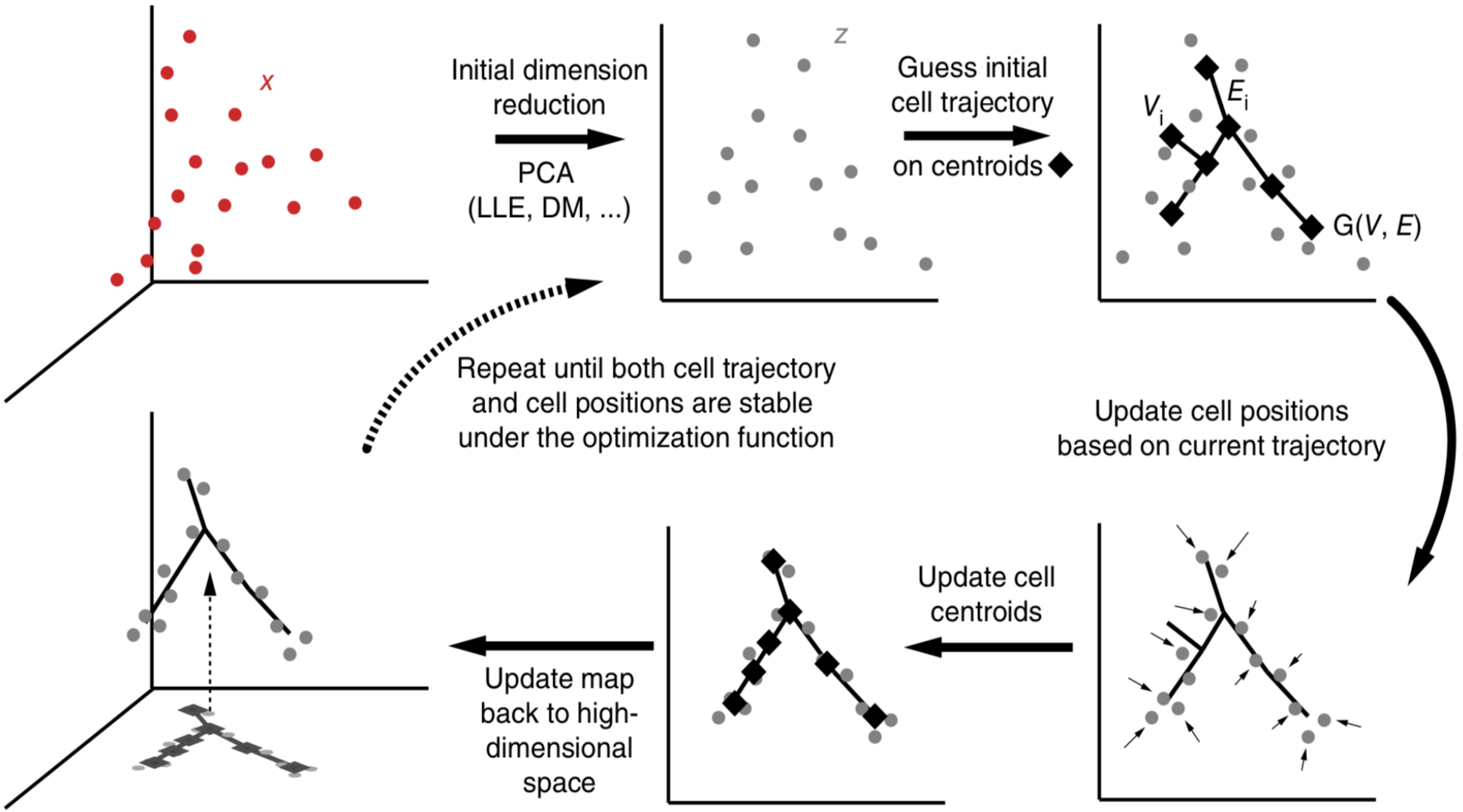
Road Map for Dimension Reduction Methods



Outline of Trajectory Analysis

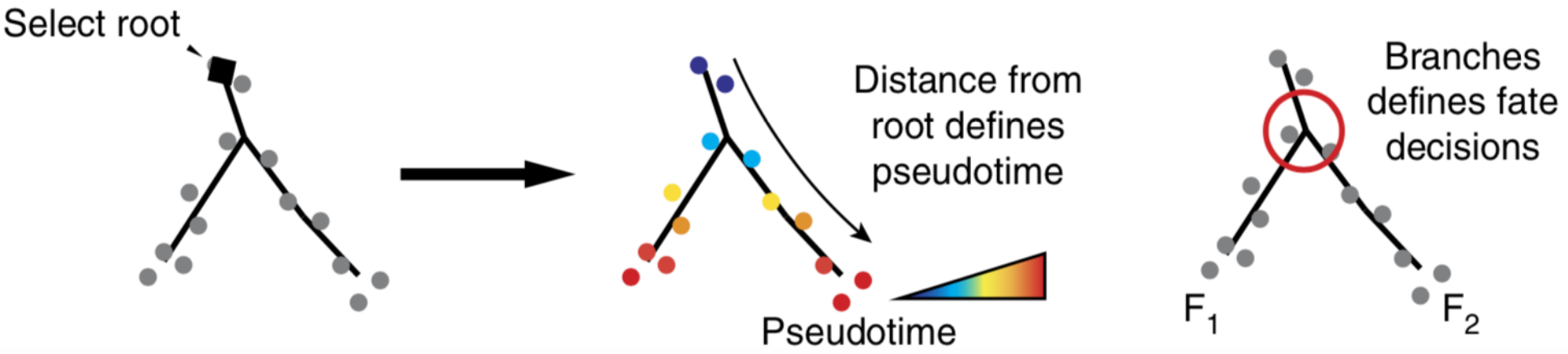


Trajectory Analysis (Monocle II)



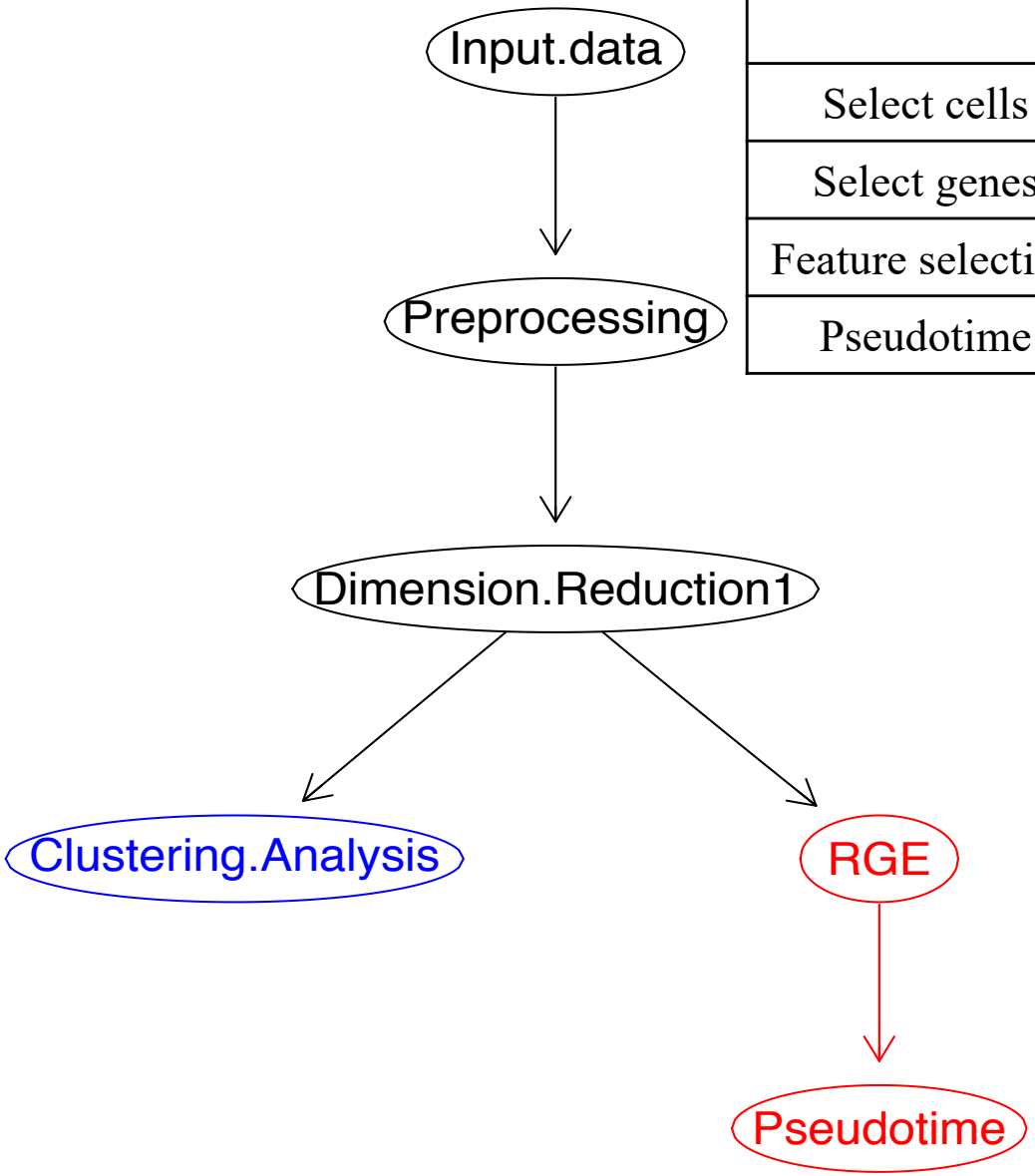
Xiaojie Qiu..Cole Trapnell Nat Methods 14:979-982 2017

Pseudotime (Monocle II)



Xiaojie Qiu..Cole Trapnell Nat Methods 14:979-982 2017

Flowchart of Trajectory Analysis with Monocle Package



	function	Description
Select cells	Remove low quality cells	
Select genes	Most variable genes	
Feature selection	Tree topology	Select subspace
Pseudotime	Select root	

RGE: reversed graph embedding

Road Map for Dimension Reduction Methods