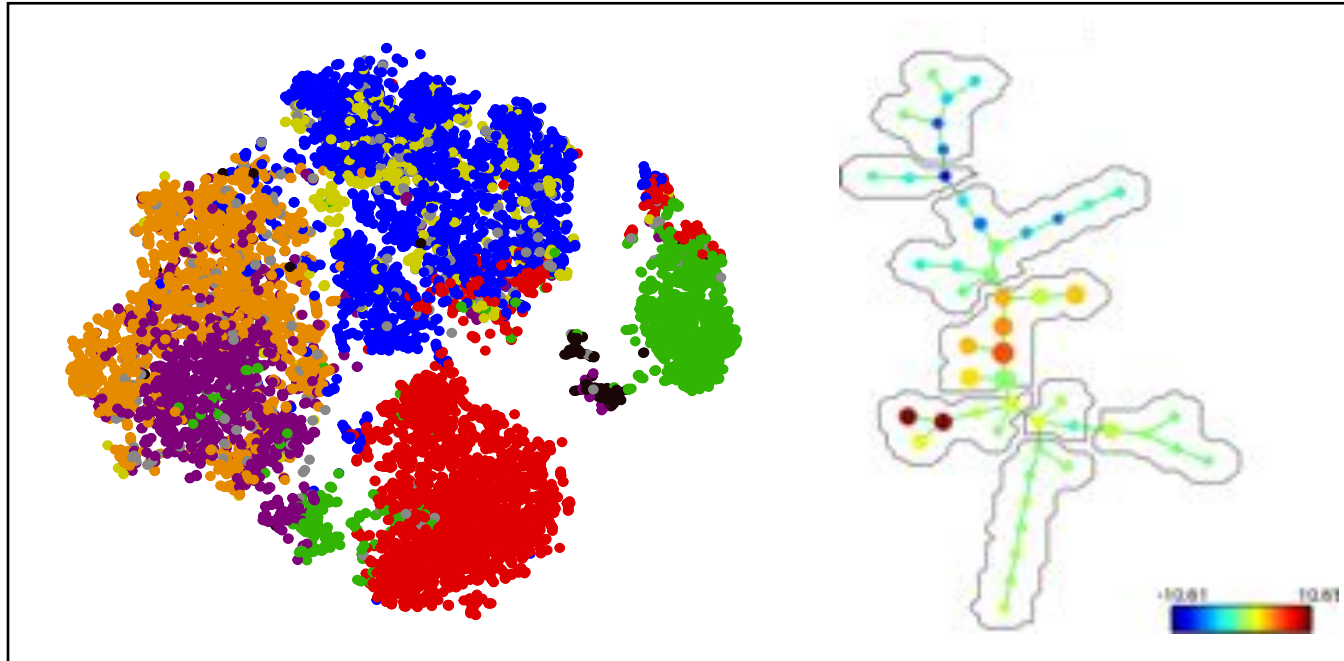


Intro to Computational Flow Workshop: viSNE, PhenoGraph, SPADE



Lisa Borghesi
Associate Professor of Immunology
Director, Unified Flow Core

Outline

viSNE, PhenoGraph, SPADE

how the algorithms work

maximizing comparable results across experiments

Overview of Analysis Workflow

FJ → export populations to cluster → algorithm

Go for it!

Cyt (viSNE, PhenoGraph) then SPADE

Systems Immunology

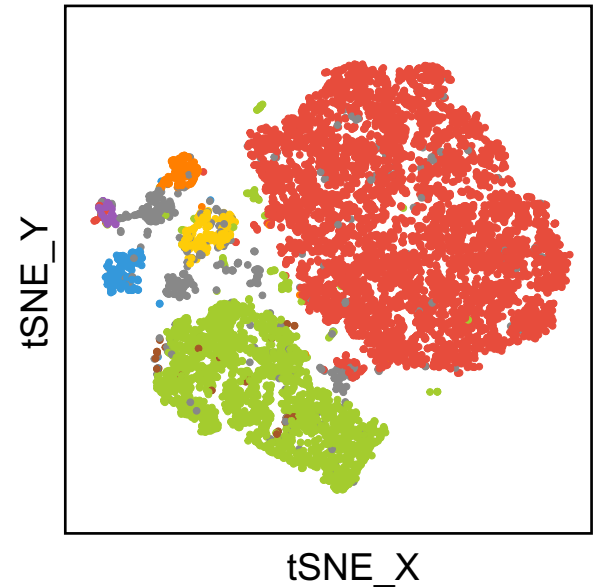
Allows for many simultaneous measurements to be made in a given sample, allowing the variation to be taken into account and used in order to make new discoveries

**Imagine 20+ color high dimensional
flow cytometry space...**

t-SNE – dimensionality reduction algorithm

Goal: find a low dimensional visualization that best reflects population structure in high dimensional space

→ colloquially, get a feel for how objects are arranged in data space



Laurens van der Maaten explains t-SNE (UCSD seminar) – fun and informative!!

<https://www.youtube.com/watch?v=EMD106bB2vY>

t-SNE vs viSNE

Van der Maaten

t-SNE: t-distributed stochastic neighbor embedding

bh-SNE: Barnes-Hut modification of t-SNE (computationally faster)

Dana Pe'er

viSNE: visualization of t-distributed stochastic neighbor embedding

t-SNE → bh-SNE

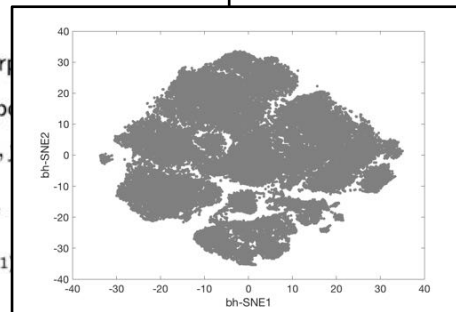
t-SNE algorithm:

begin

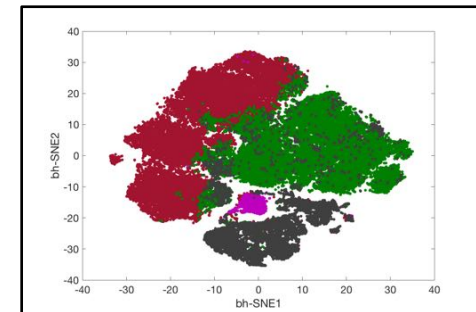
- 1 Compute pairwise affinities $p_{j|i}$ with per
- 2 Set $p_{ij} = \frac{p_{ij} + p_{ji}}{2n}$ (n - Number of data p
- 3 Sample initial solution $\mathcal{Y}^{(0)} = y_1, y_2, \dots$
for t=1 to T do
 - 1 Compute low-dimensional affinities q_{ij}
 - 2 Compute gradient $\frac{\delta C}{\delta y}$
 - 3 Set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta y} + \alpha(t)(\mathcal{Y}^{(t-1)}$

end

end



viSNE

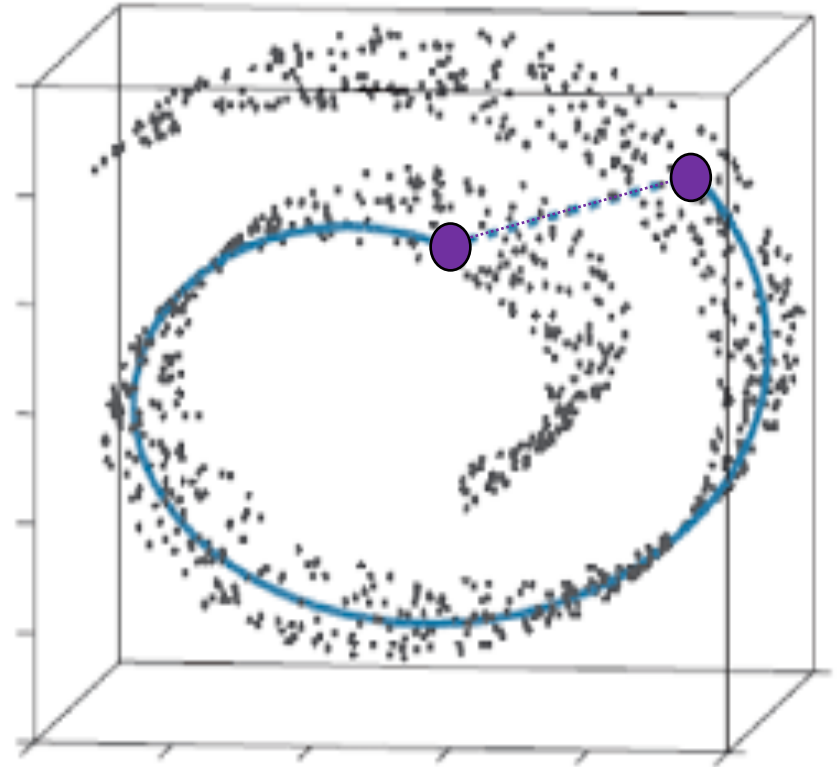


Why not just use PCA?

PCA



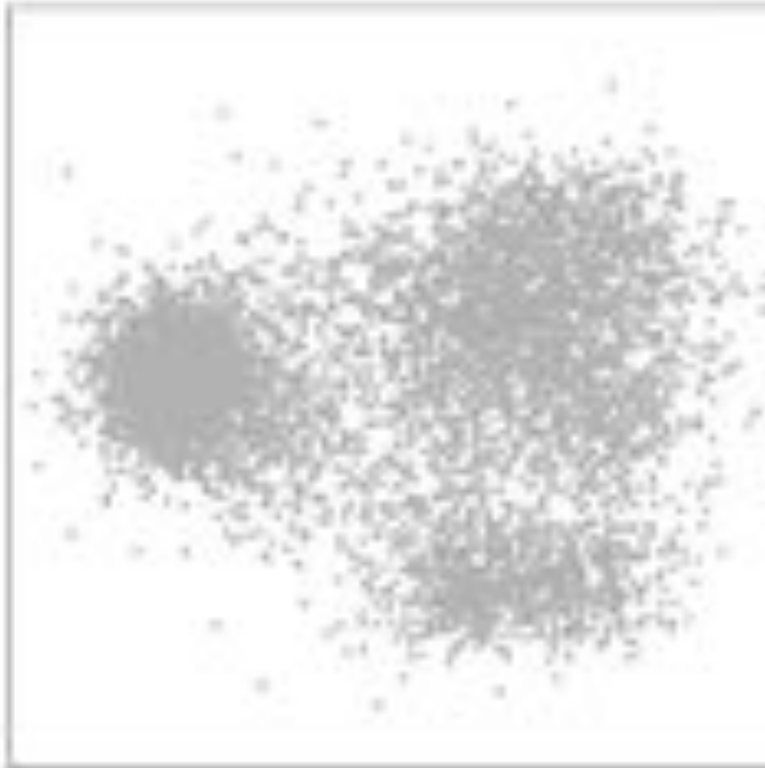
Swiss roll problem



- Principal Components Analysis (PCA) preserves large pairwise distances
- Euclidean distance between two points on the Swiss roll does not accurately reflect local structure

t-SNE preserves local distances and global distances

PCA



tSNE

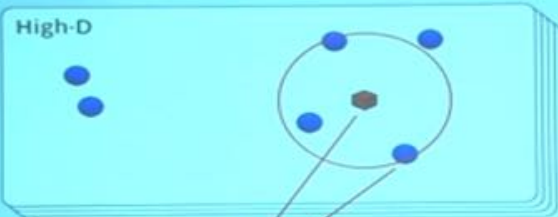


Amir 2013 Nature Biotech, Suppl.

t-SNE operation

Stochastic Neighbor Embedding

- Measure pairwise similarities between high-dimensional objects:



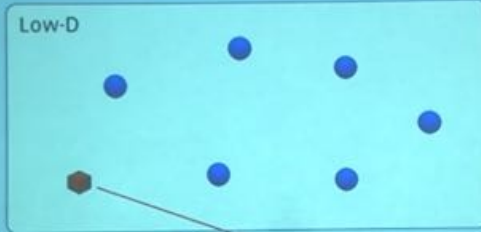
High-D

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_k \sum_{l \neq k} \exp(-\|x_k - x_l\|^2 / 2\sigma^2)}$$

11:57 / 1:06:37

Stochastic Neighbor Embedding

- Measure pairwise similarities between low-dimensional map points:



Low-D

$$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_k \sum_{l \neq k} \exp(-\|y_k - y_l\|^2)}$$

Subtitles/clo

12:58 / 1:06:37

High-D data space. Draw Gaussian bell (circle) around data point. Measure density of all other points relative to that Gaussian bell, and establish probability distribution that represents their similarity. Computes local densities to get a distribution of pairs of points.

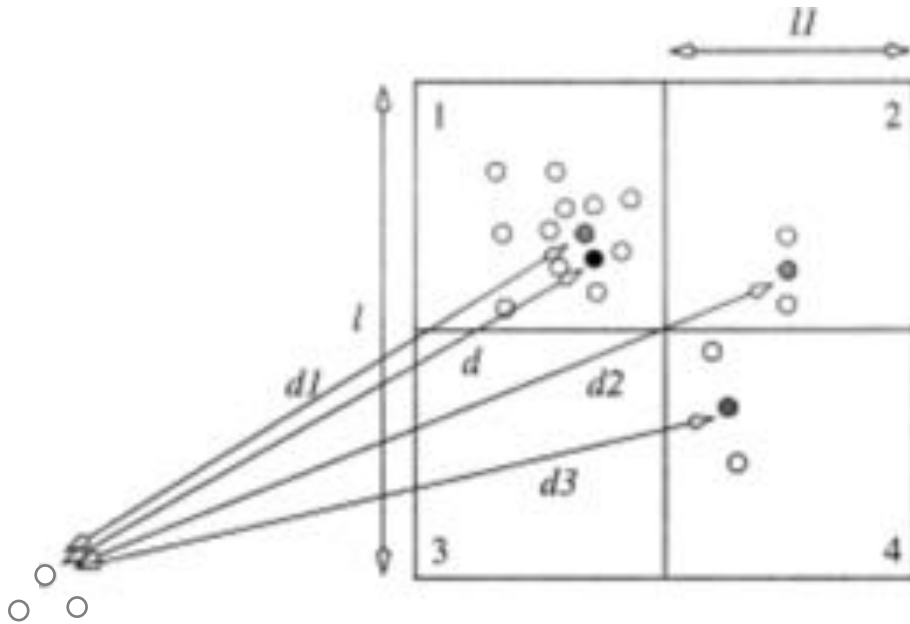
→ P_{ij}

Low-D 2D map. Repeat above.

→ Q_{ij}

Mathematically minimize $P||Q$ difference. Zero would be if two points were the same.

Barnes-Hut Modification of t-SNE



- Center of mass of domain
- Centers of mass of subdomains
- Source particle

if ($l/d < a$)

*compute direct force interaction
with the center of mass of domain.*

else

if ($l/d_1 < a$)

*compute direct force computation
with center of mass of subdomain 1*

else

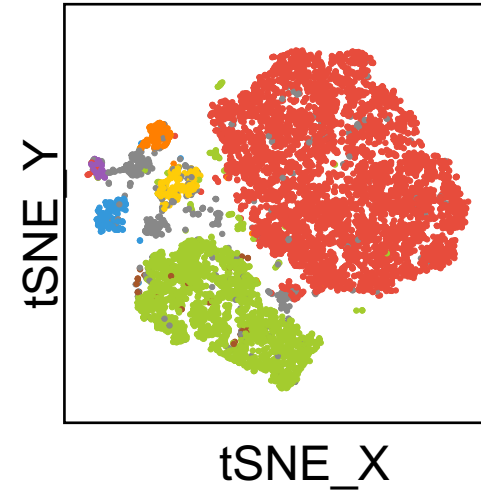
expand subdomain 1 further

Apply similar criteria to domains 2, 3, and 4

t-SNE

Advantages

- single cell information
- non-linear assumptions (as opposed to PCA)
- preserves local and global structure



Limitations

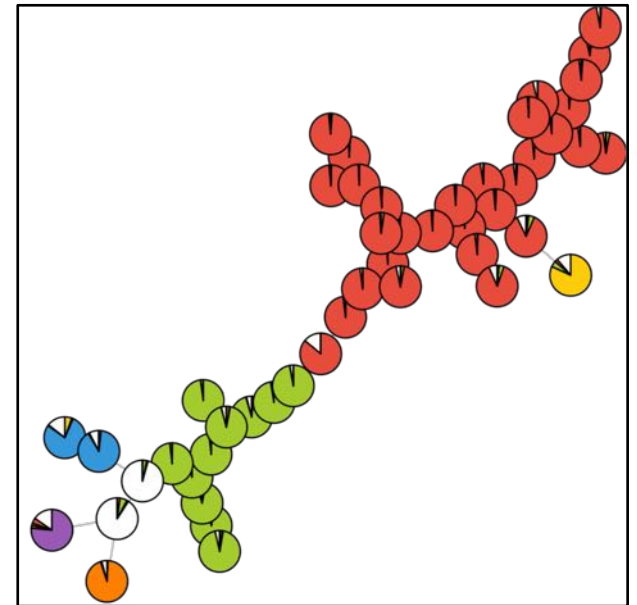
- computationally expensive; obligate downsampling means data are discarded
- plot axes are arbitrary and have no intrinsic meaning
- no population identification; follow up approaches required to assign identity to clusters and cells
- distance between clusters is not meaningful; no hierarchy

SPADE: Hierarchical clustering algorithm

spanning tree progression of density normalized events

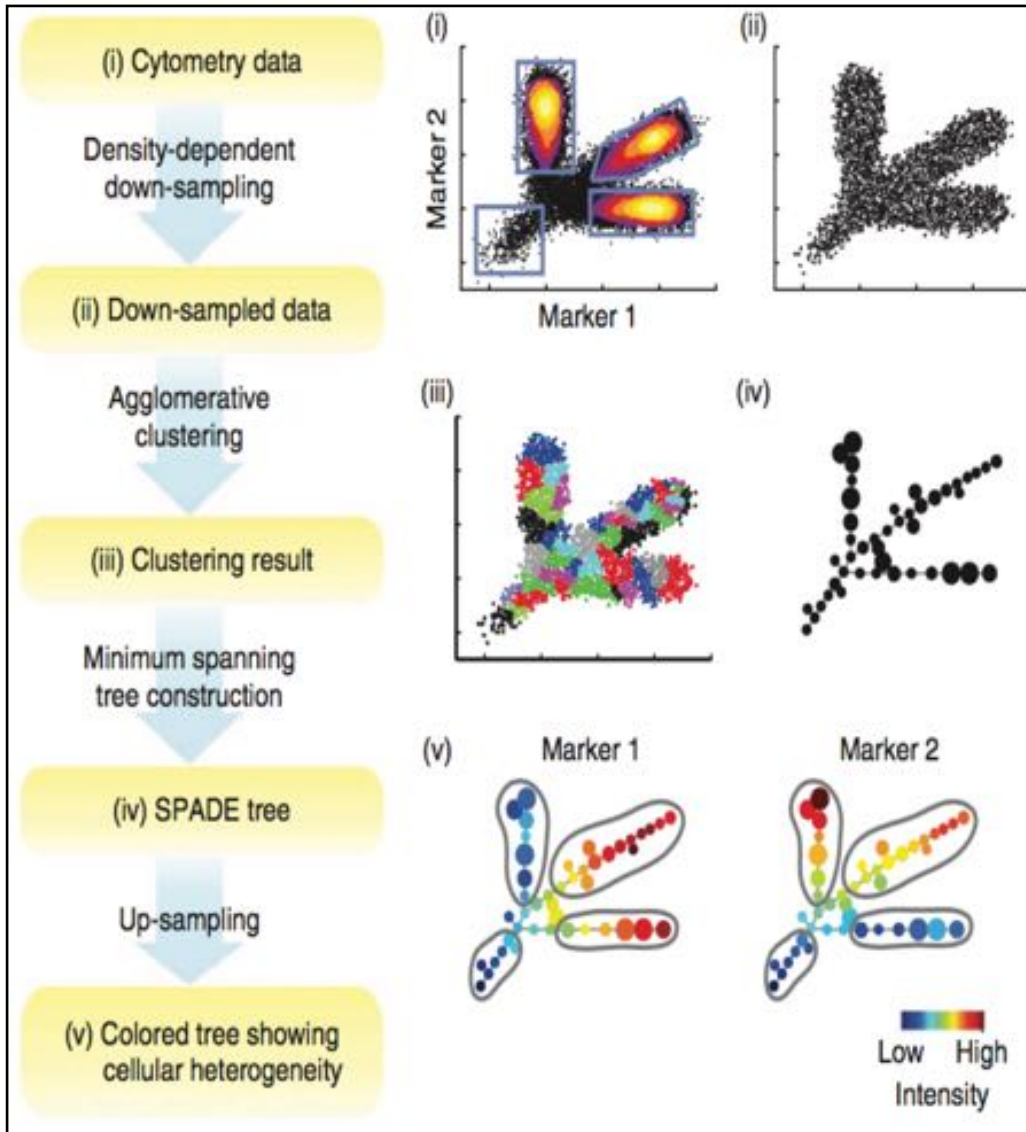
Goal: organize cells into a **hierarchy** using unsupervised approaches

→ colloquially, generate a tree of relationships



Output minimal spanning tree (MST) highlights the relationships between most closely related cell type clusters

SPADE



SPADE views **data as a cloud** of points (cells) where the dimensions = # markers

Density-dependent downsampling to equalize density in different parts of cloud, ensures rare cells not lost

Agglomerative clustering based on marker intensity

Connect clusters in minimal spanning tree that best reflects geometry of the original cloud

Upsampling, map each cell in the original data set to the clusters

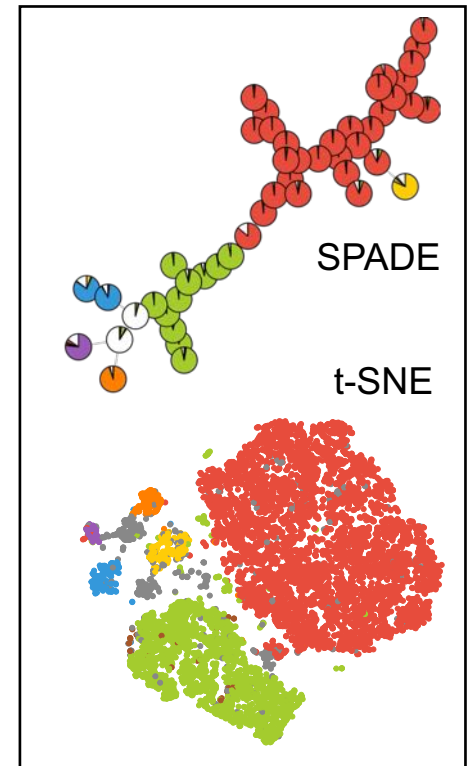
SPADE

Advantages

- rare pops preserved through density-dependent downsampling
- enables visualization of continuity of phenotypes
- can combine data sets that share common markers, and then co-map any markers unique to each data set (see orig. paper)

Limitations

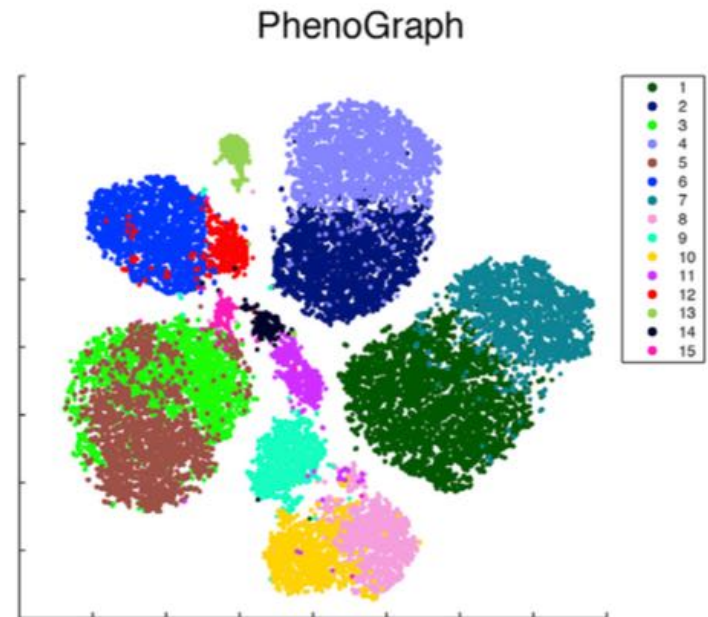
- loss of single cell information
- user chooses cluster number
- MSTs are non-cyclic and paths can be artificially split



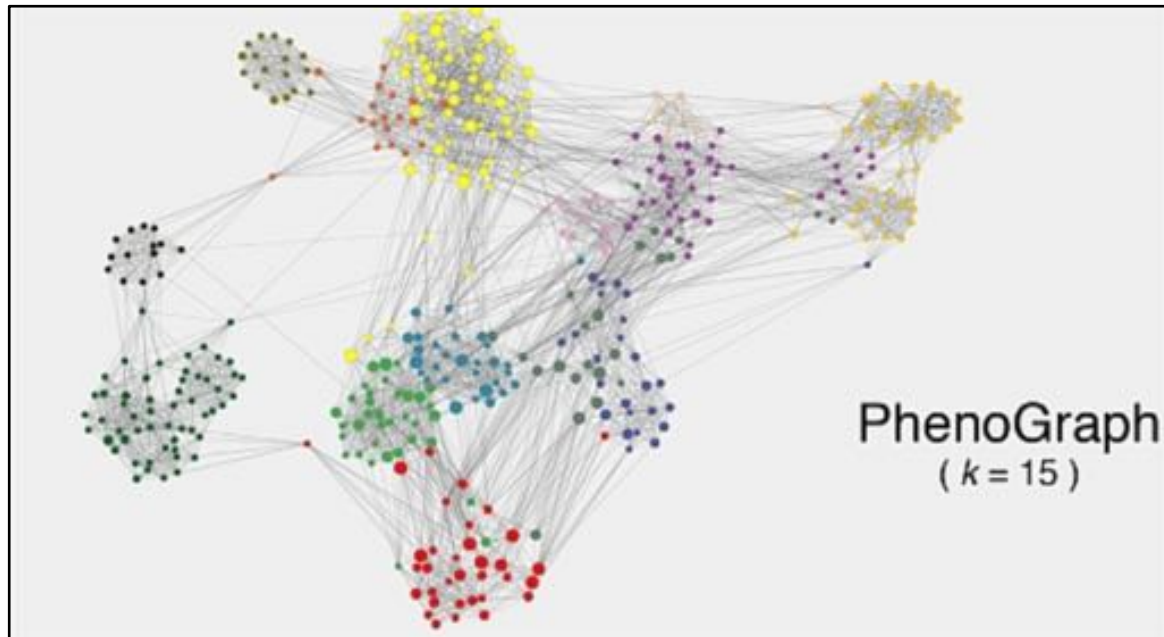
PhenoGraph

Goal: automated partitioning of high-dimensional single-cell data into subpopulations

→ colloquially, map nearest neighbors



PhenoGraph



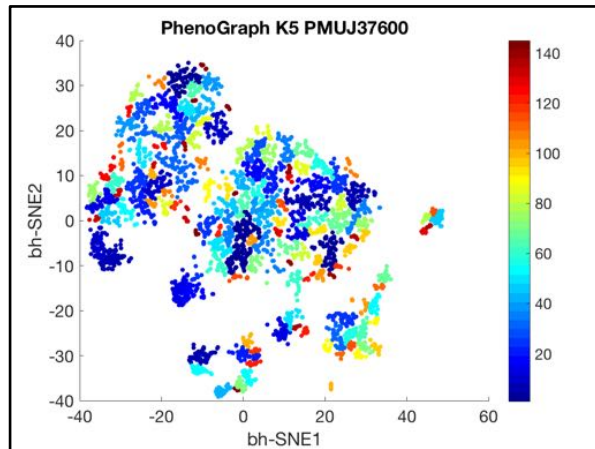
First order relationship – find the k nearest neighbors for each cell using Euclidean distance

Second order relationships – cells with shared neighbors should be placed near one another

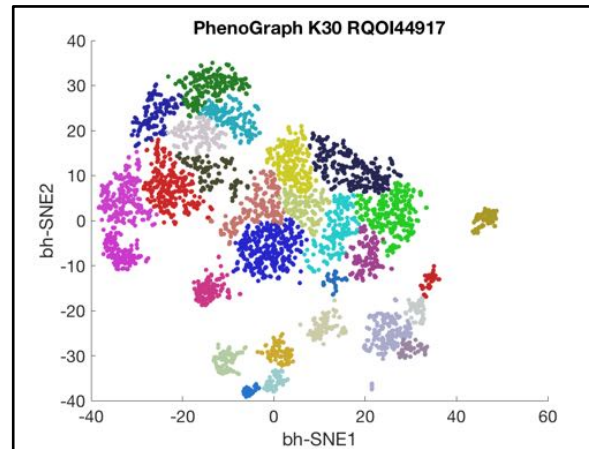
Third, identify communities – Louvain method that measures the density of edges inside communities to edges outside communities

PhenoGraph: Number of neighbors

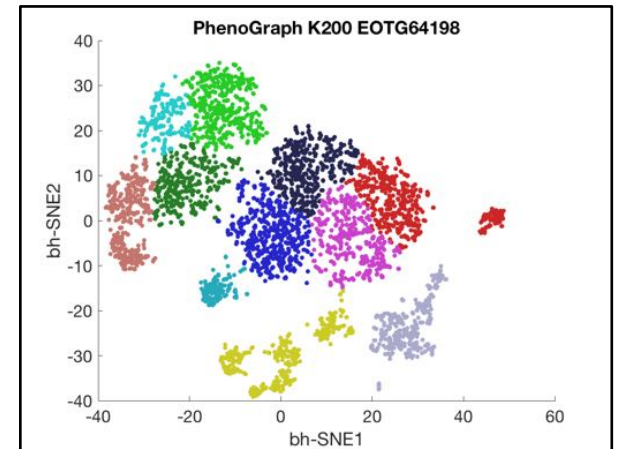
Neighbors = 5



Neighbors = 30

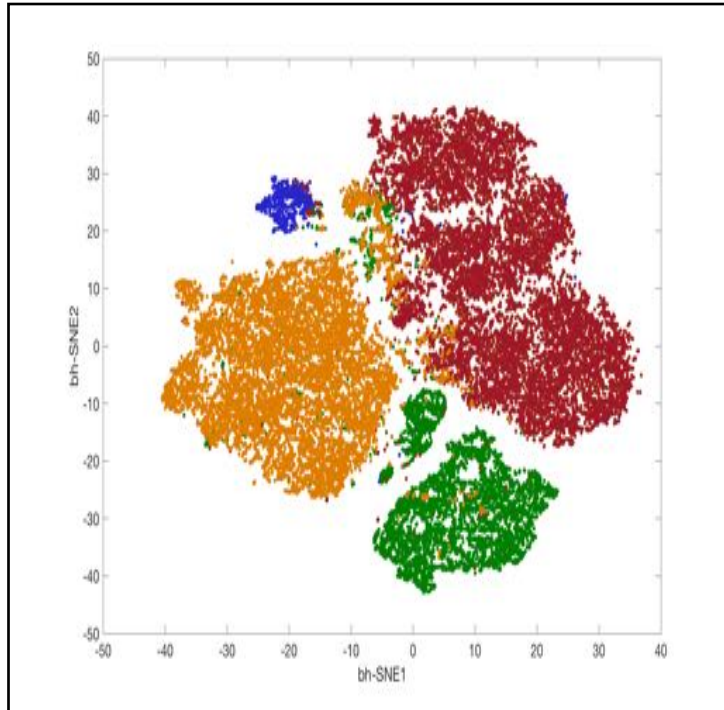


Neighbors = 200

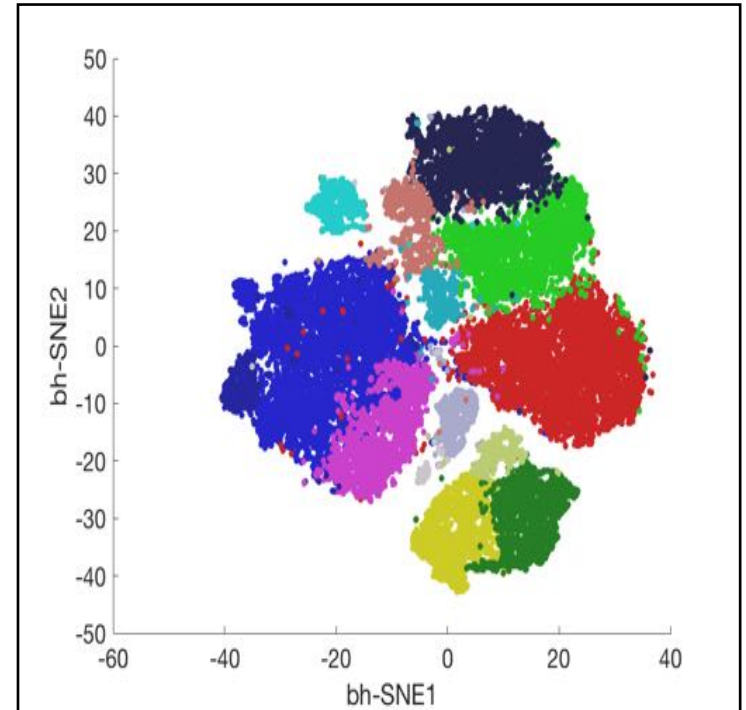


PhenoGraph – population discovery

Manual gate overlays



PhenoGraph



- Naive B
- ASC
- MBC (total)
- Ag-exper.

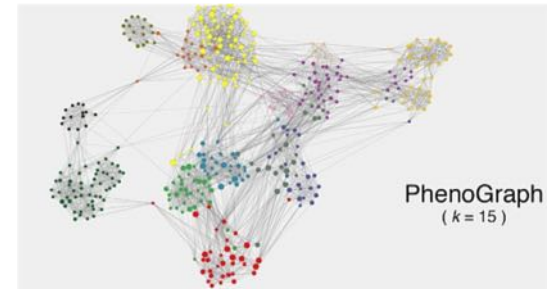
PhenoGraph

Advantages

- opportunity for population discovery
- can resolve subpopulations as rare as 1 in 2000 cells
- robust to cluster shape (e.g., need not be spherical)

Limitations

- user specifies number of neighbors
- ideal cluster number, or biologically relevant cluster number, is largely unconstrained



Pooling data across experiments

Standardize fluorescence intensities when performing experiments over time

**→ Use Rainbow Beads to adjust PMT voltages for
each fluorescence channel to target MFI values**

Leukemia (2012) 26:1986-2010

Nature Protocols (2012) 7:2067-2079

Outline

viSNE, PhenoGraph, SPADE

how the algorithms work

maximizing comparable results across experiments

Overview of Analysis Workflow

FJ → export populations to cluster → algorithm

Go for it!

Cyt (viSNE, PhenoGraph) then SPADE

Basic Workflow



Gate on live singlets, population to cluster

Ex. Live single CD45+ cells

Ex. Live single CD19+ B cells

Ex. Live single CD4+Foxp3+ Tregs

Goal: Multidimensional profiling of human skin T cells

→gate on dump^{neg}CD3^{pos}

Two specimens

1. Normal skin “lib 1mg”
2. Psoriasis (PsO_stelara)

Two compensation matrices

1. Normal
2. Psoriasis



Antibody Panel:

CD4 BUV 395

CD8 BUV 737

CD3 FITC = population to cluster

TCRab APC (A647)

TCRgd PerCP/Cy5.5

CD45RO BV510

CTLA4/CD152 PE-TxRed

CD69 AF700

CD103 PE-Cy7

Foxp3 PE

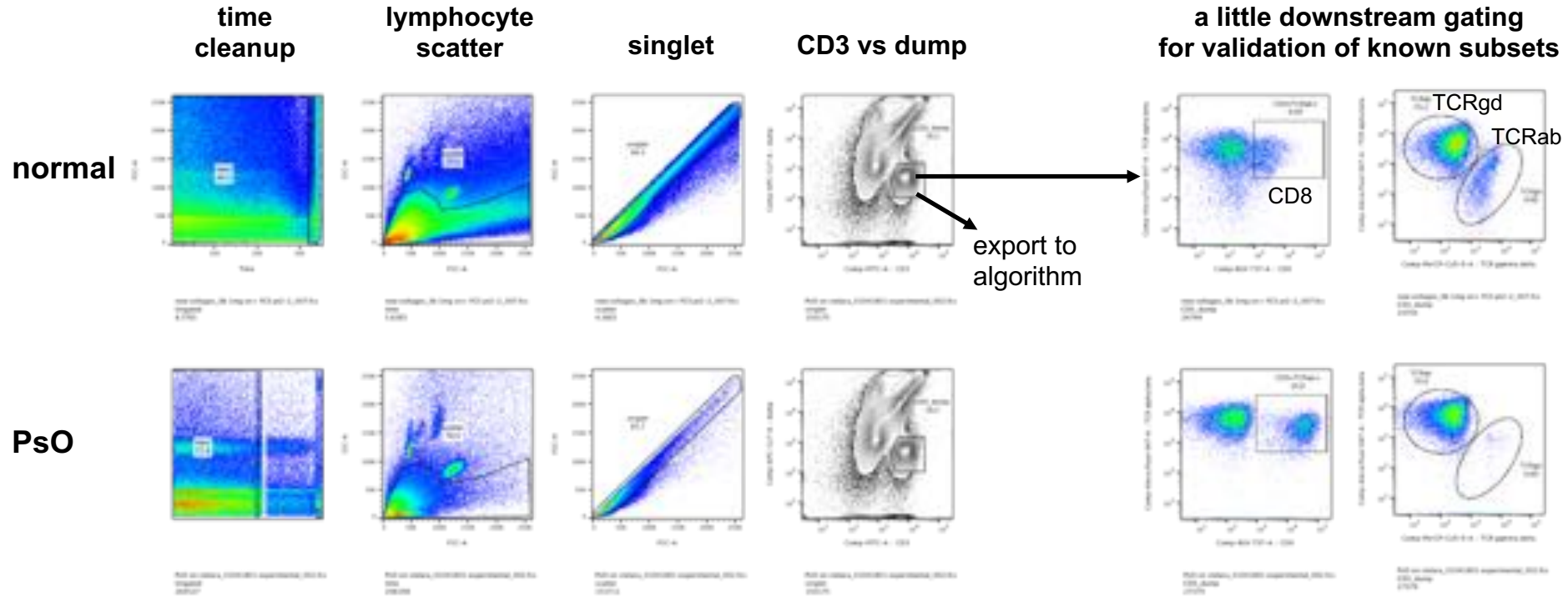
Tbet BV605

RORgt BV421

Dump: live/dead, CD11c, CD19, CD14 APCCy7

Go to FlowJo → import files, gate

FlowJo gating, live single CD3+ T cells



Export gated **population** that you want to cluster

The screenshot shows the FlowJo software interface. The 'Tools' menu is open, and the 'Export/Concatenate' option is circled in red with a red '2.' next to it. In the main workspace, three 'CD3_dump' population entries are highlighted in red, with a red '1.' next to the first one. The interface also shows a tree view on the left and a table of statistics on the right.

Name	Statistic	#Cells
PsO on stelara_01041801 experimental_002.fcs		264527
time	77.9	206198
scatter	74.5	153711
singlet	97.7	150175
CD3_dump	18.2	27379
TCRab	15.6	4265
TCRgd	70.8	19397
	0.60	164
new voltages_lib 1mg on+ FCS pt2-2_007.fcs		876893
time	64.2	563286
scatter	77.5	436479
singlet	94.3	411583
CD3_dump	6.01	24744
TCRab	8.99	2225
TCRgd	73.2	18111
	9.85	2438
tinea pt 12191701 001.fcs		106534
time	93.5	99654
scatter	64.9	64645
singlet	97.3	62911
CD3_dump	5.33	3351
	16.6	556

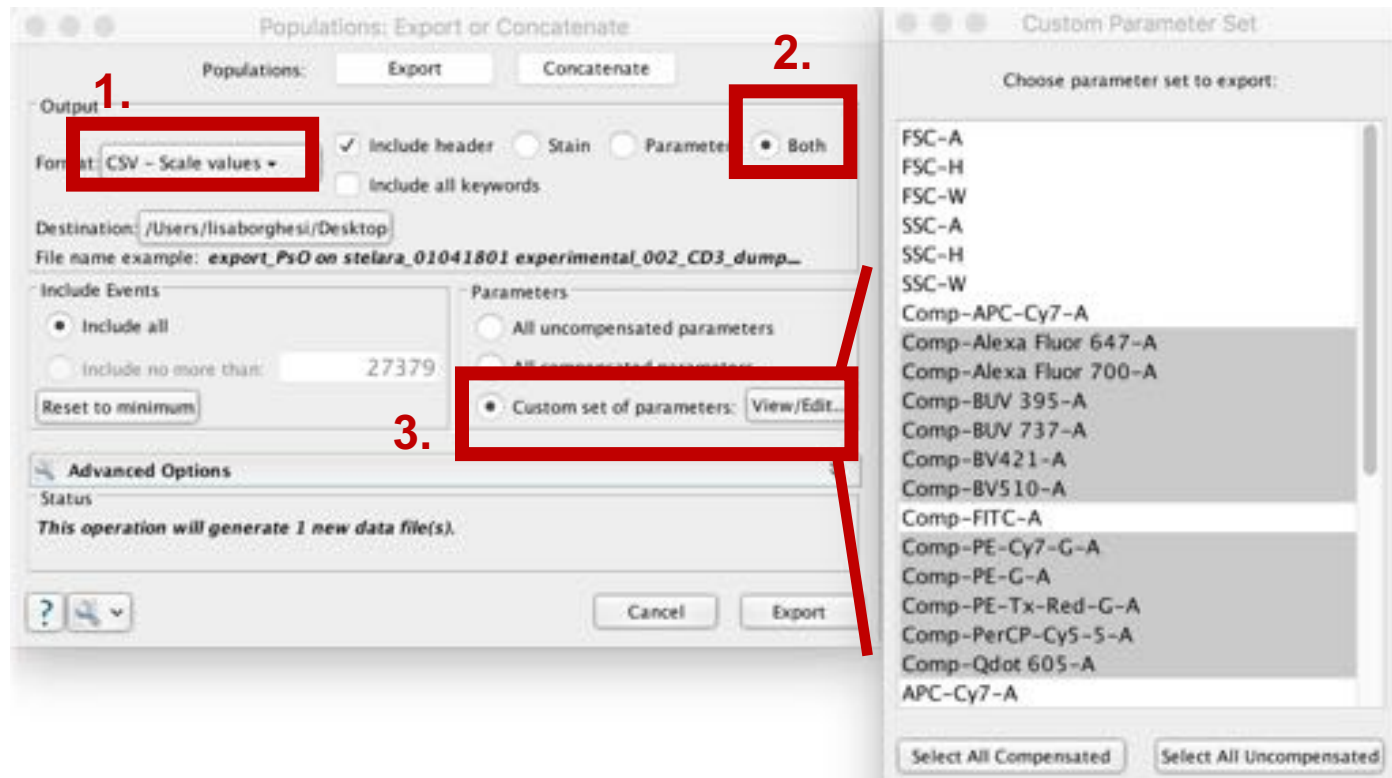
1. Highlight gates to export

2. Select Export →
Export/Concatenate
Populations

3. You'll get a pop-up
window (next slide)

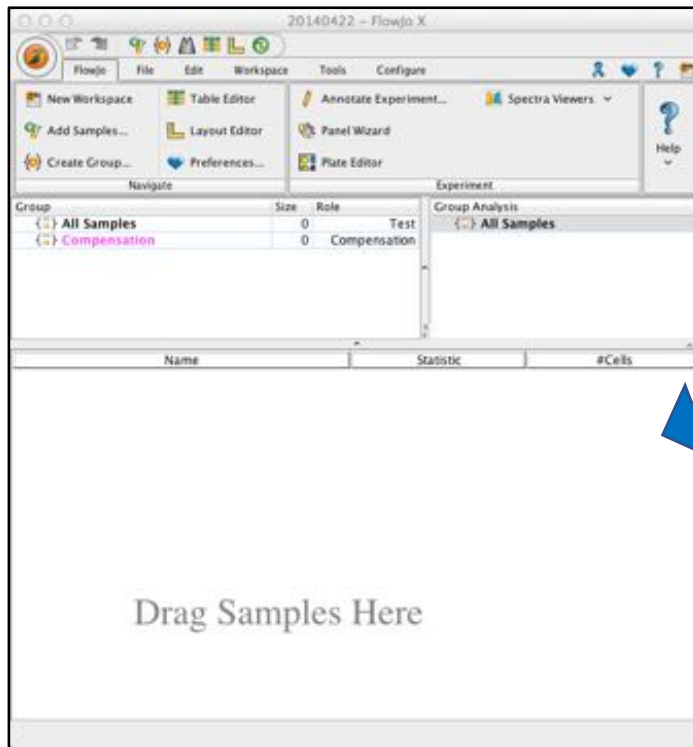
Export **custom parameters** that you want to cluster

1. Under Format choose “CSV – Scale values”
2. Select “both” stain & parameter
3. Select the compensated fluors you wish to export for clustering
 - leave behind: viability dye APC-Cy7 (you’ve already excluded dead cells)
 - leave behind: CD3 FITC (you’ve already gated on CD3)
 - leave behind: uncompensated parameters



Convert CSV → FCS

1. Open a new FlowJo workspace
2. Drop the CSV file onto the workspace
3. a new FCS file will appear in the same location where you saved the CSV file
 - conversion may take a minute or two so be patient



File.csv



File.fcs –
will appear in same
folder as your .csv

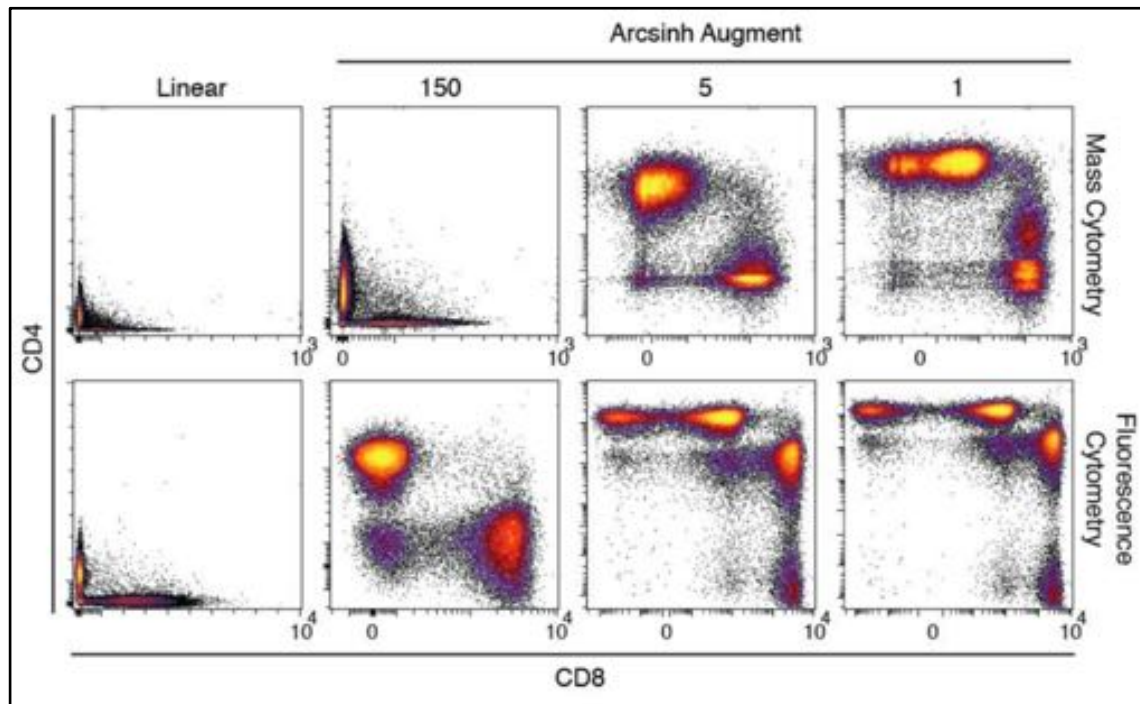
Launch Matlab/Cyt3

Steps for viSNE analysis in Matlab/Cyt3

1. **Import** fcs data gated on the population you wish to cluster
2. **Transform**, arcsinh 150
3. **Downsample** (i.e., subsample a portion of total events)
 - to reduce computational burden
 - to select a small subset of events for a quick first-pass analysis
 - to normalize events across comparative analyses
4. Invoke **bh-SNE**

Transformation, value 150 for flow cytometry data

- fcs data can have negative numbers due to compensation correction and instrument baseline correction. **Algorithms can't handle negative numbers.**
- hyperbolic arcsine (arcsinh) transformation is similar to biexponential transformation in FlowJo. See <http://docs.flowjo.com/d2/graphs-and-gating/gw-transform-overview/>



Bendall 2011 Science 332:687, Fig S2

Import gated populations into Cyt3 “SightOf”

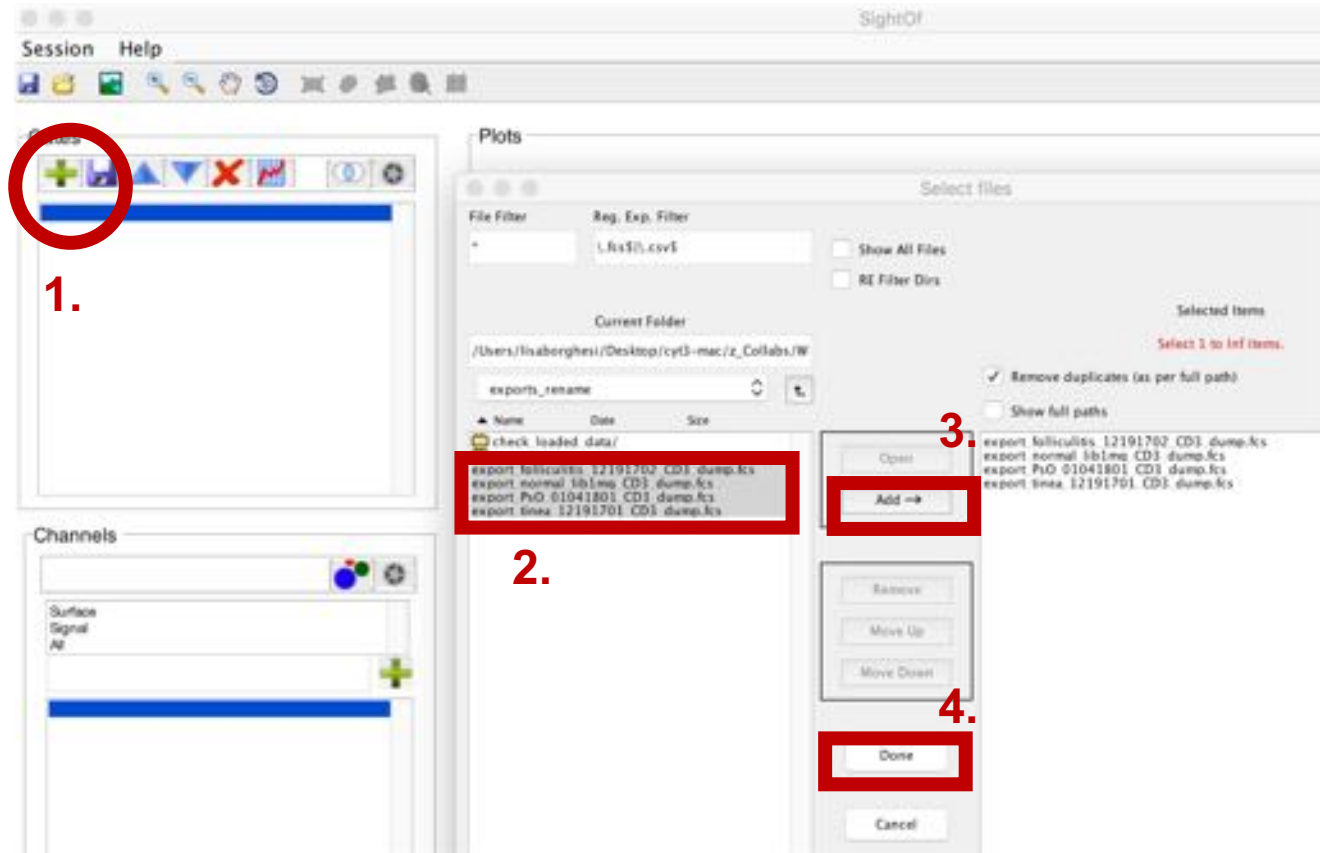
Cyt3

1. Click the “+” sign to import fcs files

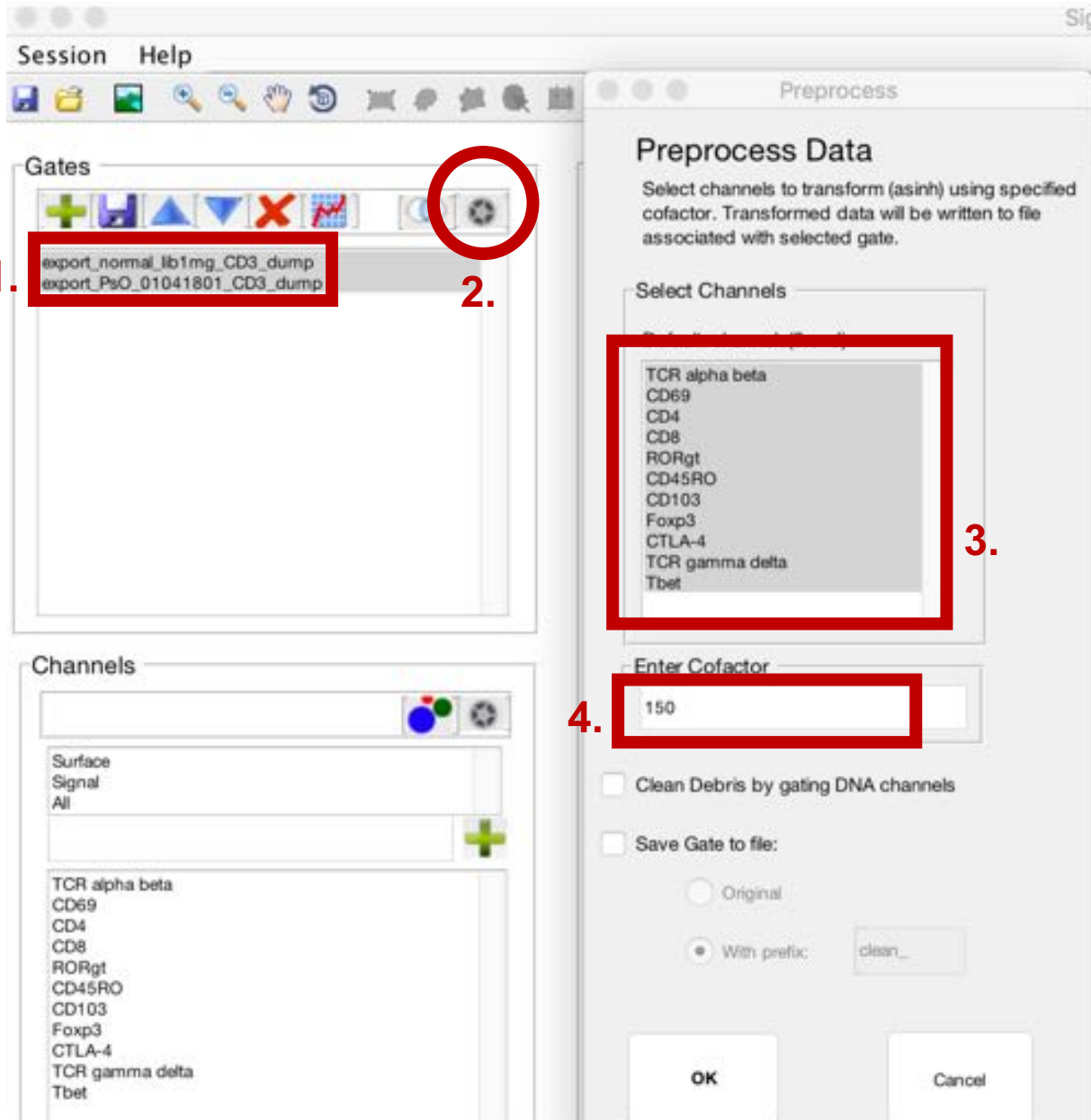
2. Navigate folder and select files

3. Select “Add”

4. Click “Done”



Pre-processing steps for viSNE analysis: Transform



1. Import fcs files

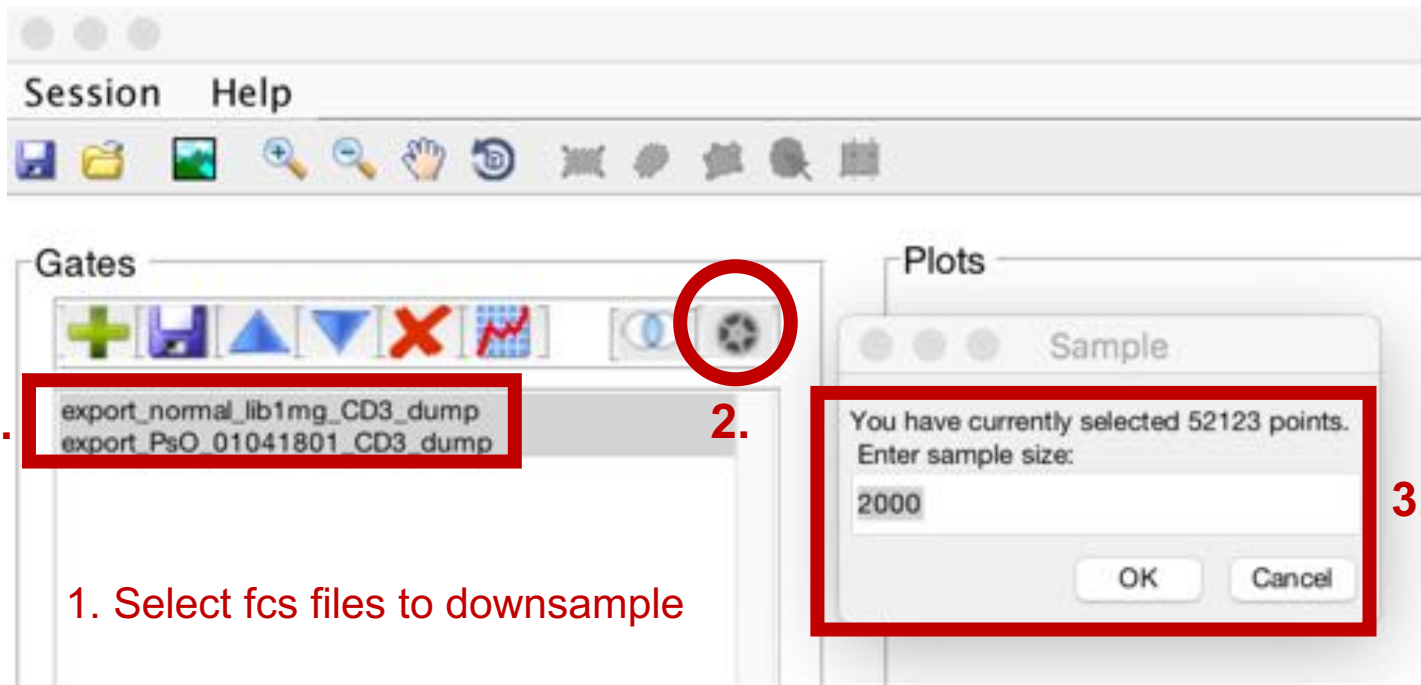
2. Select "navigator wheel" icon, and scroll to Transform

3. In popup window Select Channels to transform

4. Enter Cofactor 150 for flow cytometry (or 5 for mass cytometry)

5. Click "OK"

Pre-processing steps for viSNE analysis: Downsample



1. Select fcs files to downsample

2. Select “navigator wheel” icon, and scroll to “Subsample” (or “Subsample Each” if you have multiple samples to simultaneously downsample)

3. In popup window specify the number of events

4. Click “OK”. Then type a short prefix to be appended to sample name.

5. The new subsampled files you created should appear in the list

Invoke bh-SNE

The screenshot shows a software window with a menu bar (Session, Help) and a toolbar. Below the toolbar is a 'Gates' panel with a list of files. A red box highlights the first two items: 'dwn24Kexport_normal_ib1mg_CD3_dump' and 'dwn24Kexport_PsO_01041801_CD3_dump'. Below the 'Gates' panel is a 'Channels' panel with a list of markers. A red box highlights the list of markers: 'TCR alpha beta', 'CD69', 'CD4', 'CD8', 'RORgt', 'CD45RO', 'CD103', 'Foxp3', 'CTLA-4', 'TCR gamma delta', and 'Tbet'. A red circle highlights a gear icon in the 'Channels' panel. At the bottom of the 'Channels' panel, there are buttons for 'Histograms by Gates', 'Plot', and a share icon.

1. Select the (transformed, downsampled) fcs files to cluster

2. Select fluors to include in clustering

3. Under “lower navigator wheel” select bh-SNE

1. Select the (transformed, downsampled) fcs files to cluster

2. Select fluors to include in clustering

3. Under “lower navigator wheel” select bh-SNE

4. Check original Matlab window (remember, you are currently in Cyt window) for algorithm progress

→ Once algorithm has finished, the two new derived parameters will appear in lower pane: bh-SNE1 and bh-SNE2

View your bh-SNE plot



Gates



export_normal_lb1mg_CD3_dump

dwn24Kexport_normal_lb1mg_CD3_dump
dwn24Kexport_PsO_01041801_CD3_dump

5.

Channels

Surface
Signal
All

TCR alpha beta
CD69
CD4
CD8
RORgt
CD45RO
CD103
Foxp3
CTLA-4
TCR gamma delta
Tbet
cyt_placeholder_tmp
cyt_placeholder_tmp
cyt_placeholder_tmp
bh-SNE1
bh-SNE2

6.

Scatter

Plot

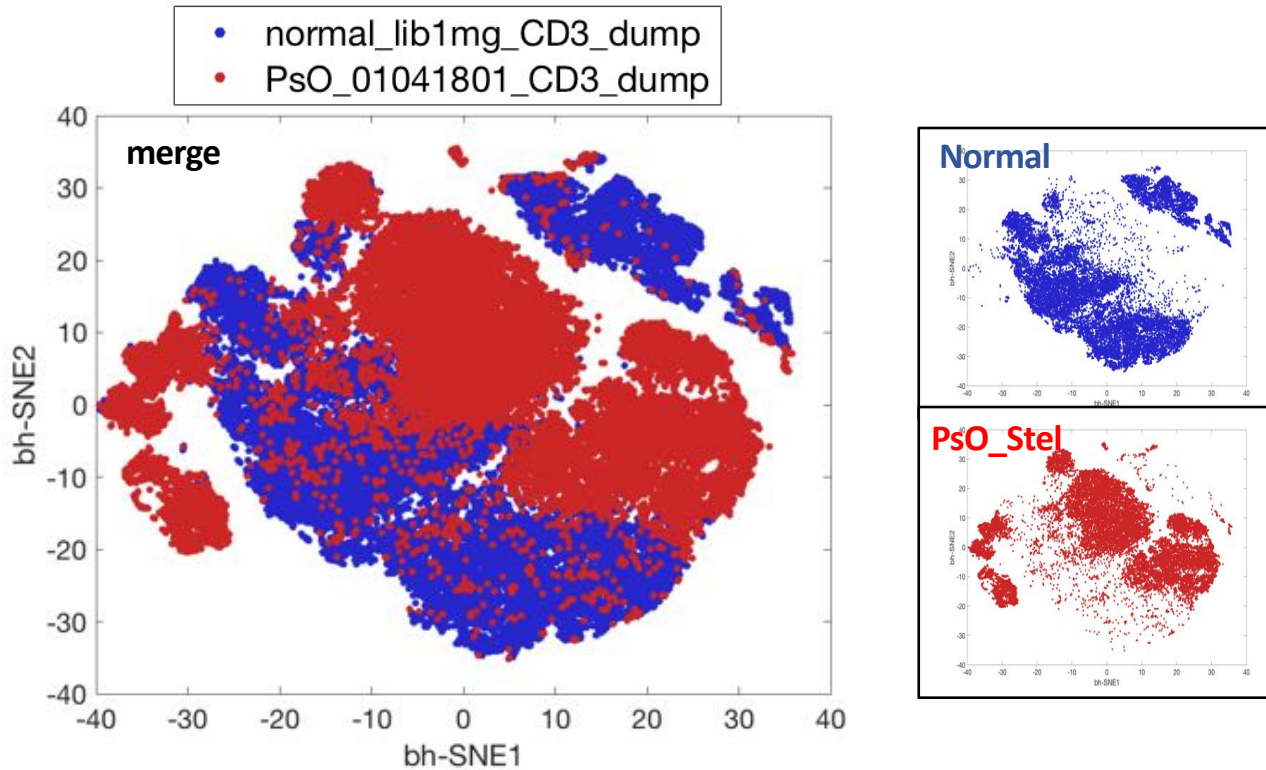
7.

5. Make sure the files you clustered are still selected.

6. Select the two new derived parameters bh-SNE1 and bh-SNE2

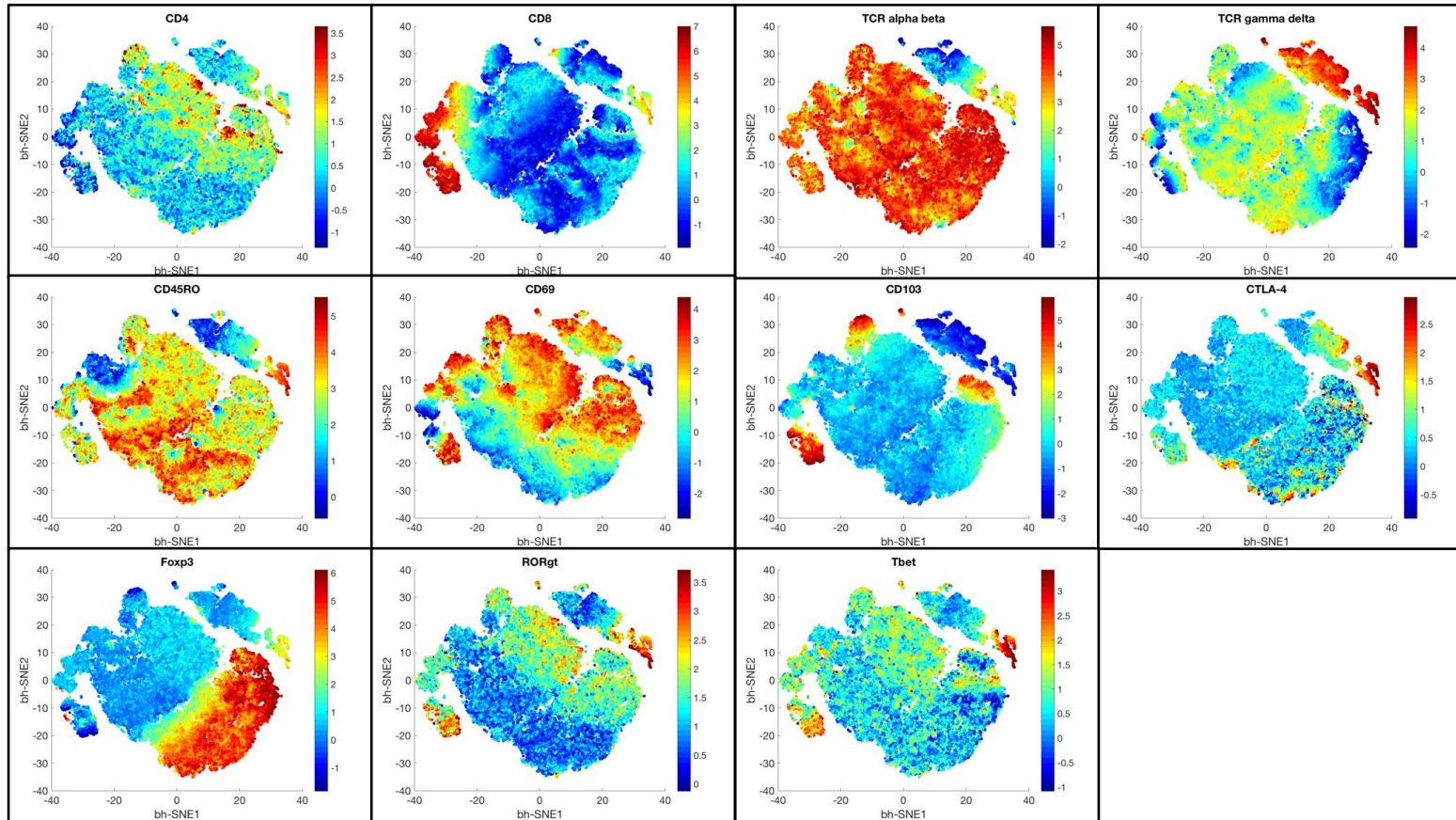
7. Select "Plot"

viSNE: normal vs PsO

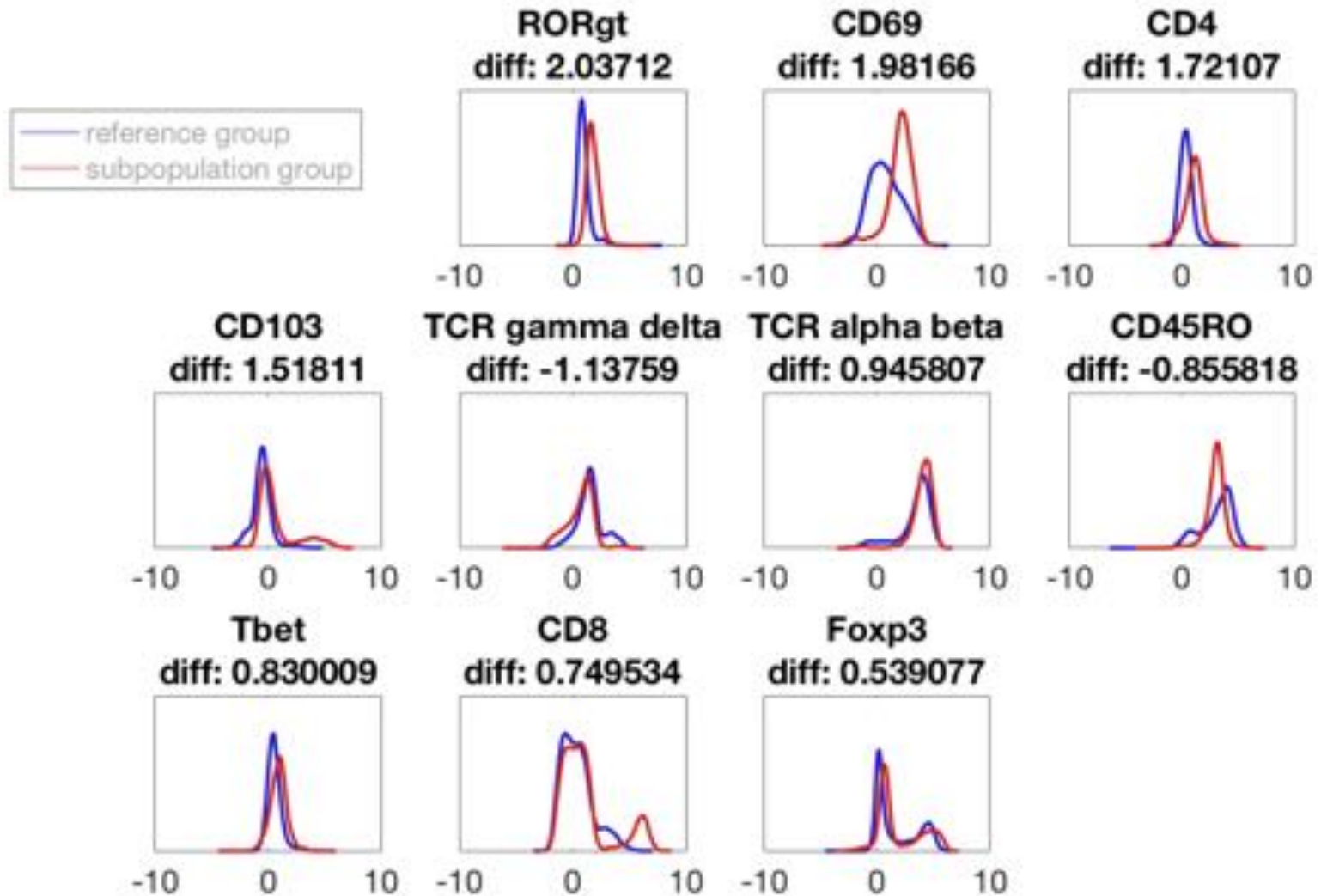


- downsample Normal, PsO_stel to 24K/ea
- cluster all markers except for pre-gated channels (CD3, dump)

Marker heatmaps

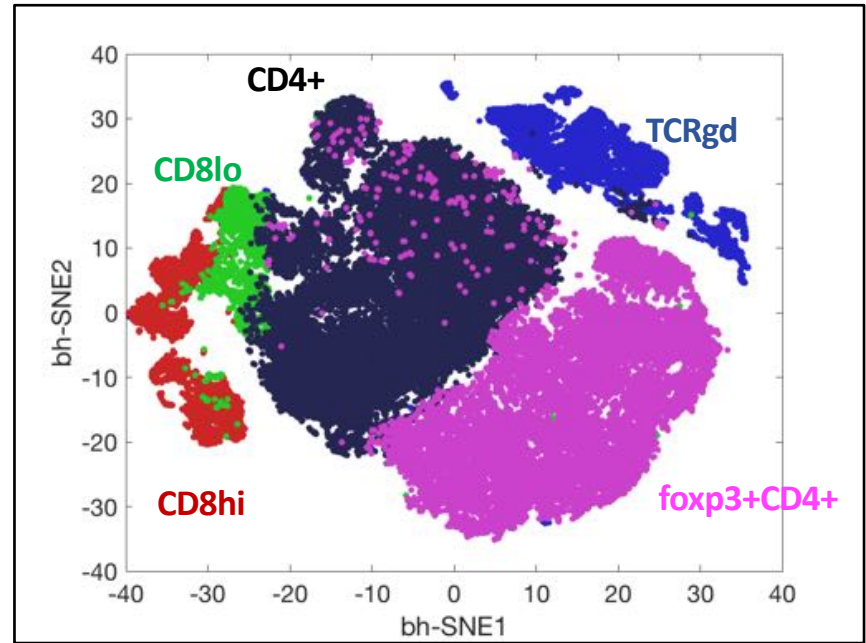
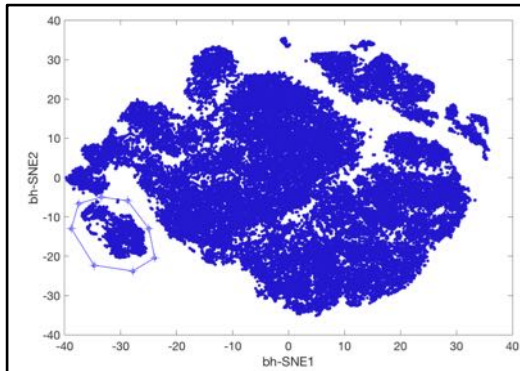


L1 statistic – difference between marker distributions

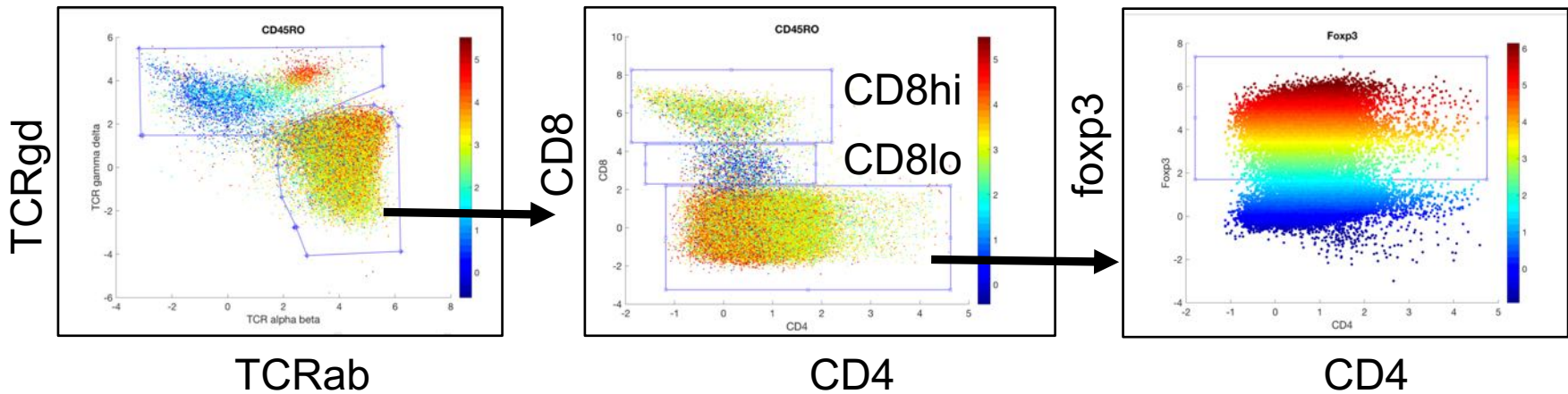


Manual gating overlays

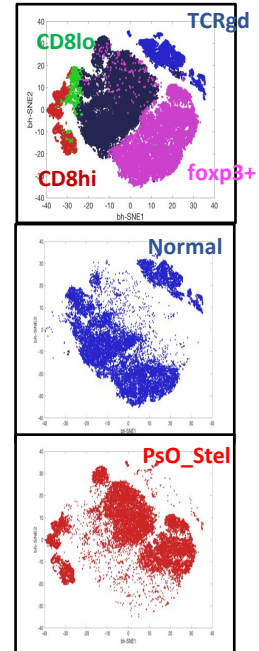
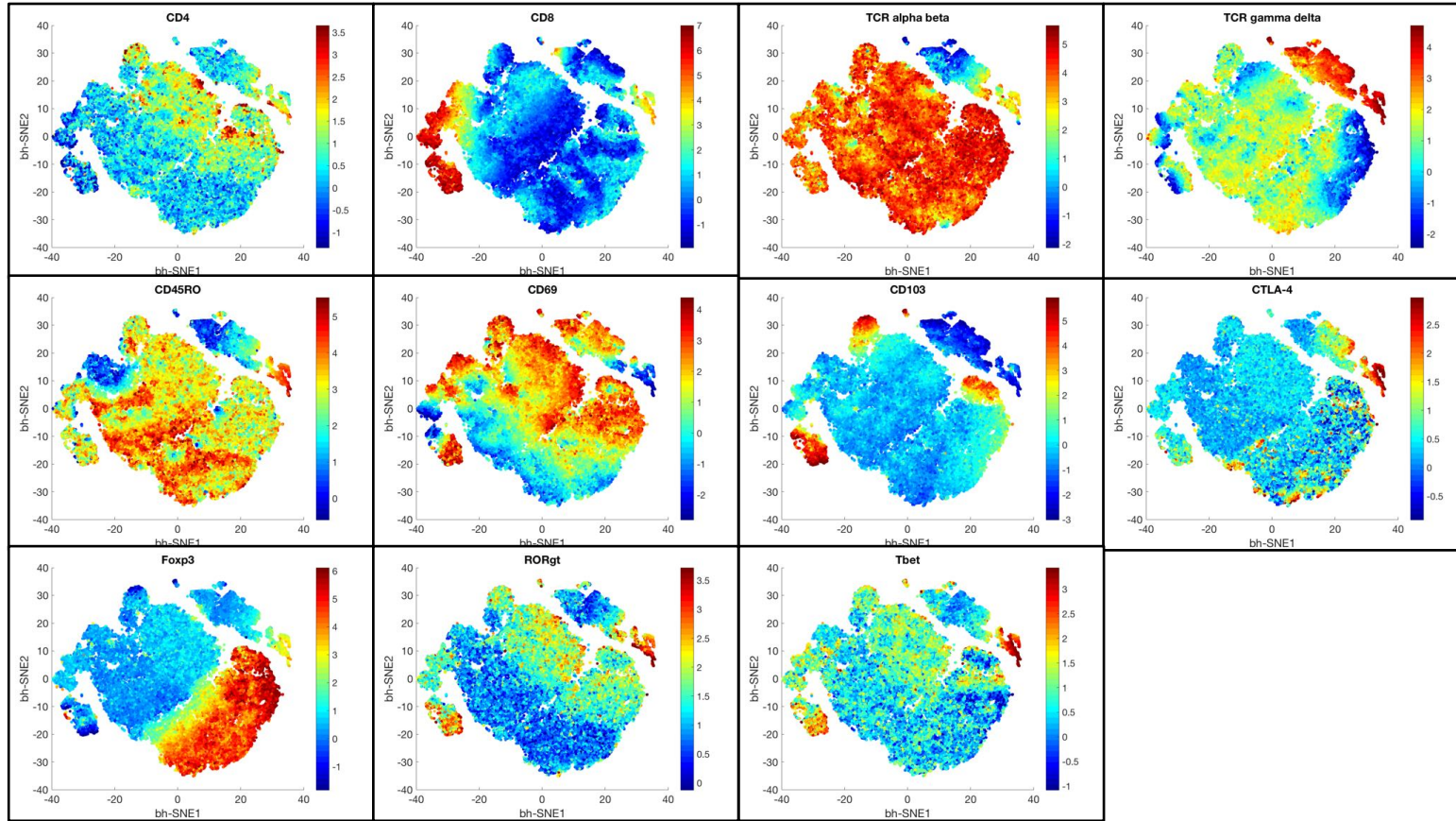
Gate on tSNE clusters



Gate on fluors



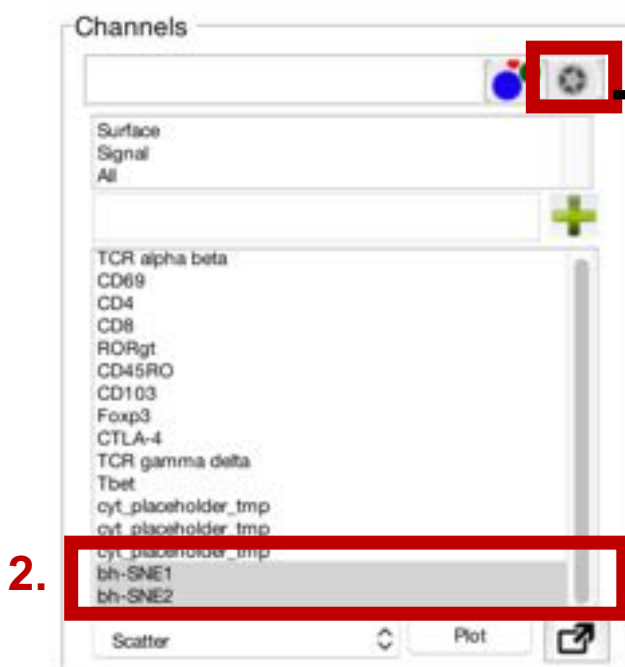
Summary: Multidimensional profiling of human skin T cells



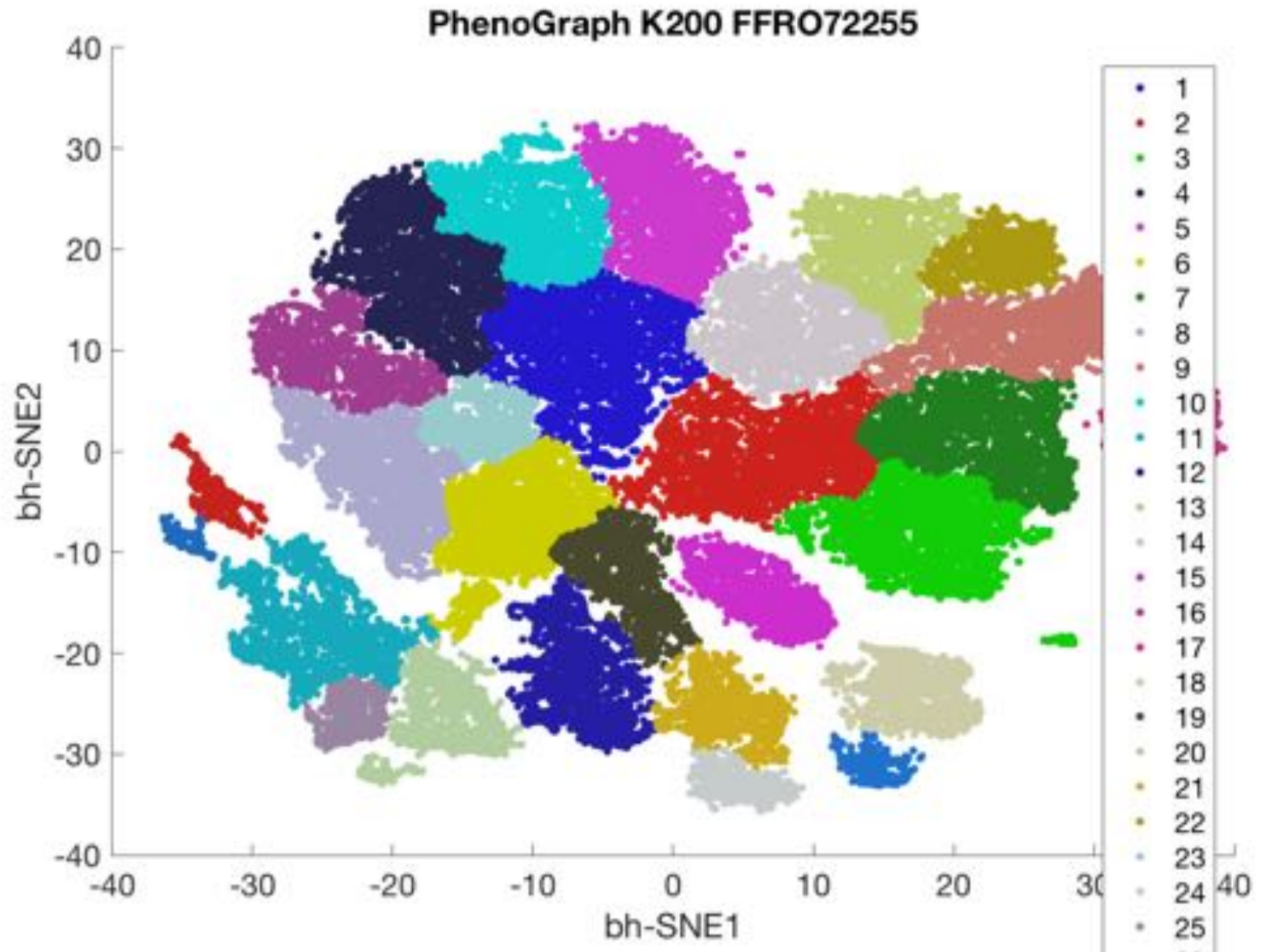
PhenoGraph



1. Make sure the files you clustered are still selected.
2. Select the two new derived parameters bh-SNE1 and bh-SNE2
3. Under navigation wheel select “PhenoGraph”



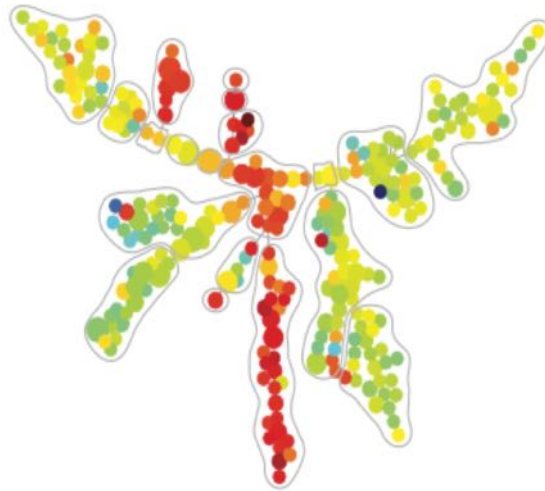
PhenoGraph – automated clustering



SPADE

SPADE developer provides great instructions

<http://pengqiu.gatech.edu/software/SPADE/>



Import gated populations into SPADE

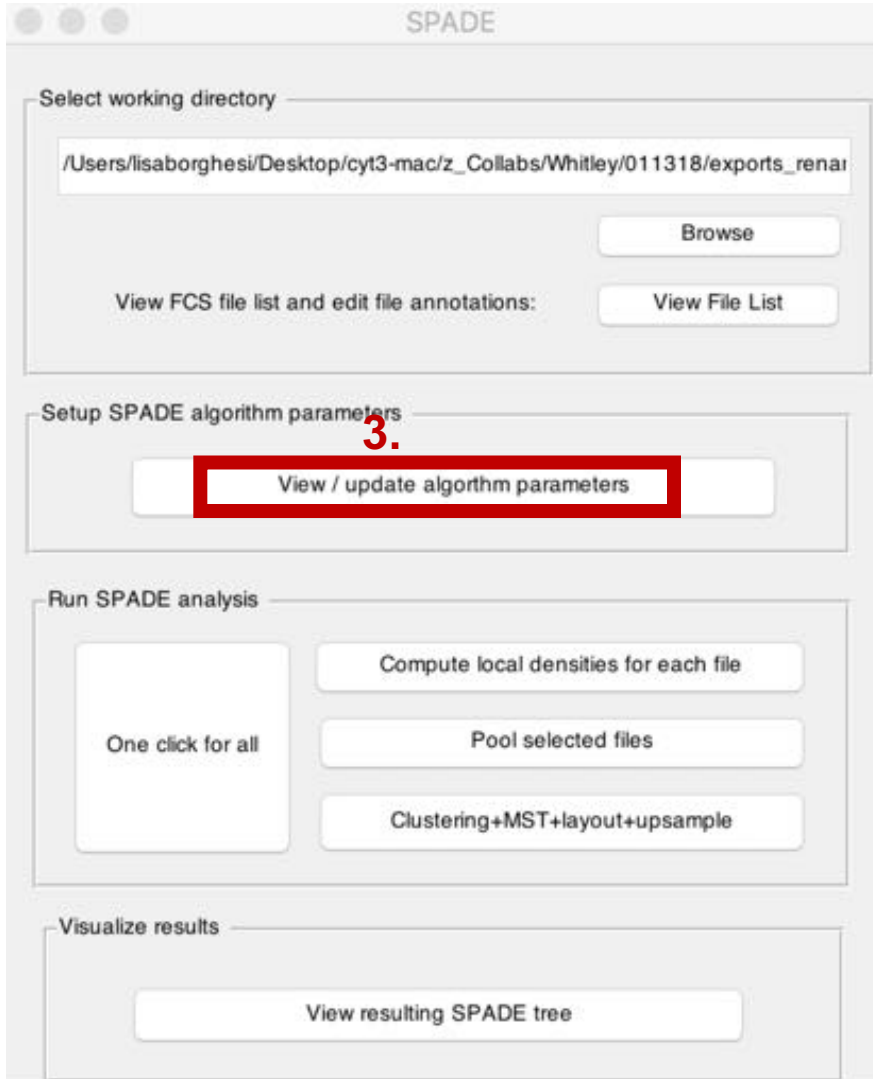
1. Click “Browse” and navigate to folder that contains gated fcs files
2. Select the folder and click “Close”

SPADE

The screenshot displays the SPADE software interface. On the left, the 'Select working directory' section shows a text field with a file path and a 'Browse' button highlighted with a red box and the number '1.'. Below this are buttons for 'View FCS file list and edit file annotations' and 'View File List'. The 'Setup SPADE algorithm parameters' section has a 'View / update algorithm parameters' button. The 'Run SPADE analysis' section includes a 'One click for all' button and three buttons: 'Compute local densities for each file', 'Pool selected files', and 'Clustering+MST+layout+upsample'. The 'Visualize results' section has a 'View resulting SPADE tree' button. On the right, the 'View/Edit file annotations' window is open, showing a table with two columns: 'filename' and 'file short annot...'. The table contains two rows of data. A 'Close' button in the bottom right corner of this window is highlighted with a red box and the number '2.'.

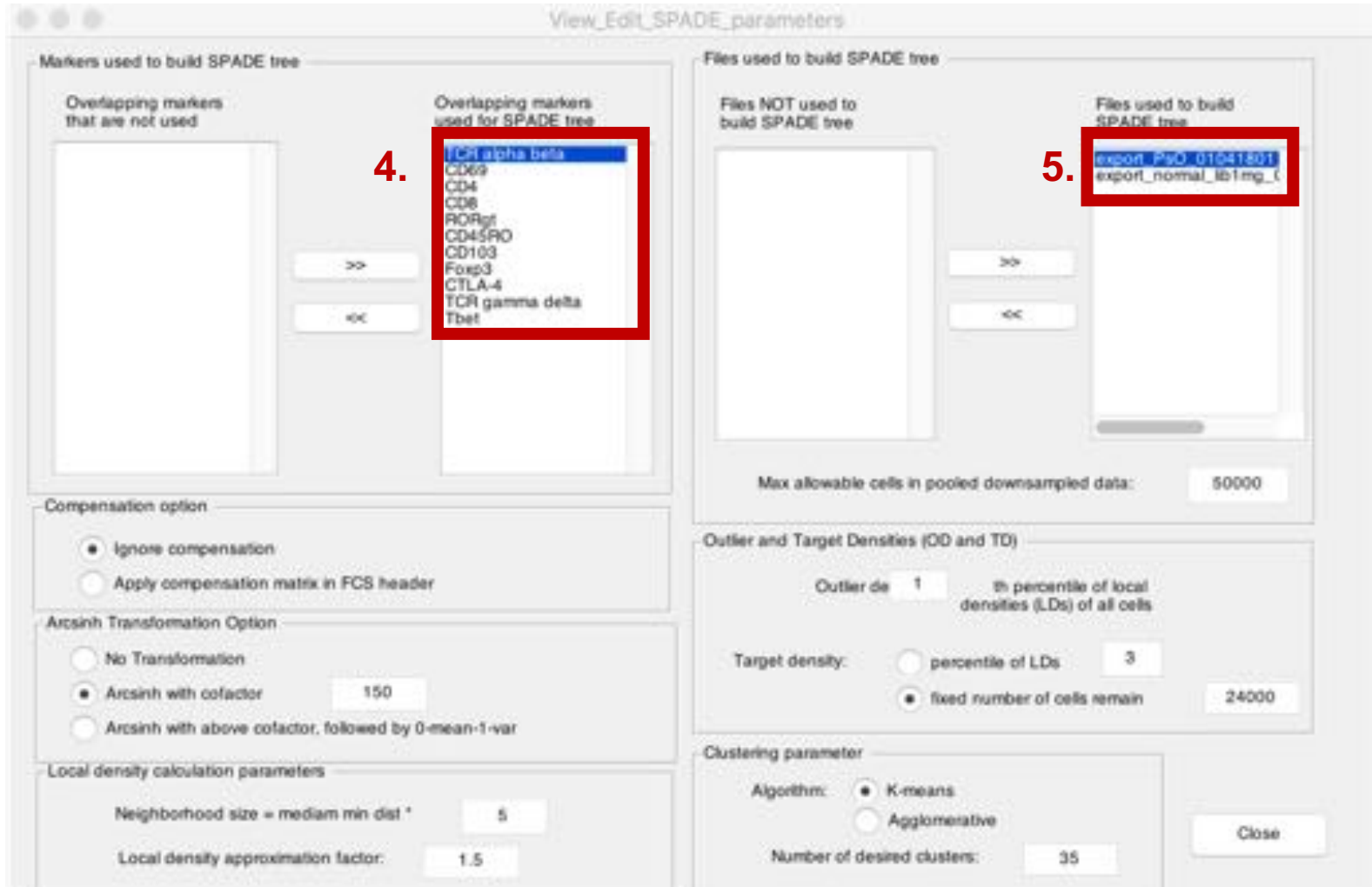
	filename	file short annot...
1	export_PsD_01041801_CD3_dump.fcs	export_PsD_0104...
2	export_normal_ib1mg_CD3_dump.fcs	export_normal_ib1...

Setup SPADE parameters



3. Click "View/update algorithm parameters"

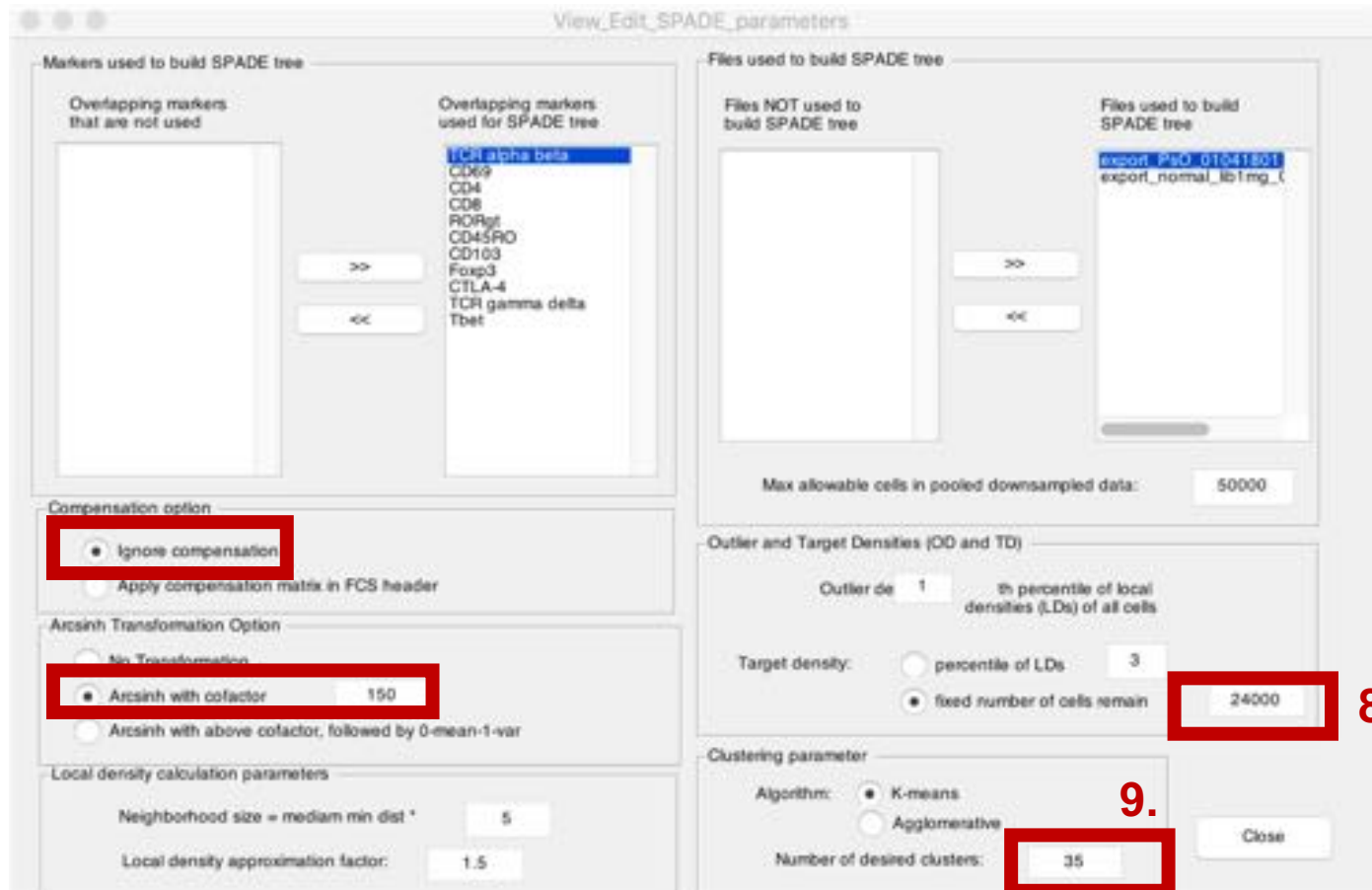
Setup SPADE parameters



4. Select markers for building SPADE tree (move left >> right)

5. Select files to use for building SPADE tree (ditto)

Setup SPADE parameters



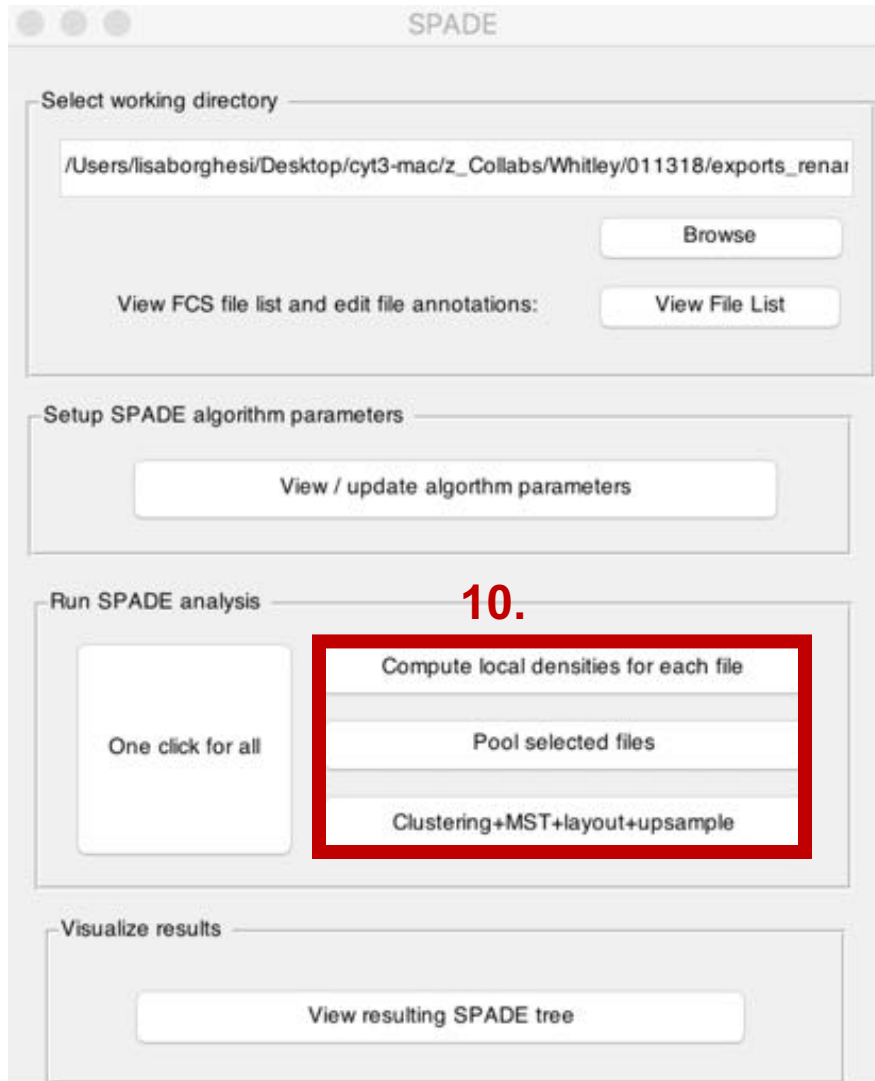
6. Select "Ignore compensation" since we are using compensated data from FlowJo

7. Arcsinh transform, cofactor 150

8. Assign target density such that a fixed number of cells survive the downsampling process

9. Set desired number of clusters

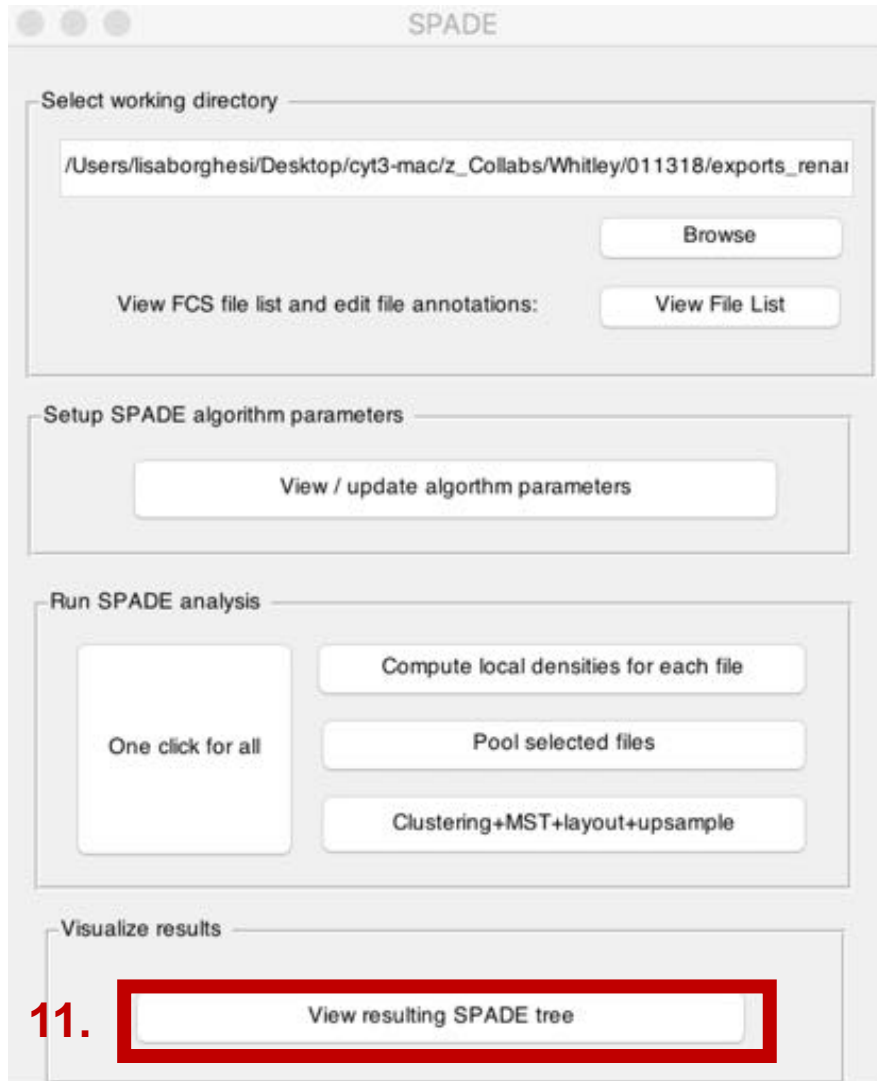
Run SPADE analysis



10. Today we'll sequentially walk through each step in the algorithm.

- a) Compute local densities for each file
→ Fast. Feedback is "100% Done"
- b) Pool selected files
→ Takes a bit longer. Feedback is "Done!"
- c) Clustering
→ Fast. Feedback is "Done"

Run SPADE analysis



11. View resulting SPADE tree

Color tree by marker

show tree in new window Arch Layout Spring RadioExpand

Show node/cluster index

Edit SPADE tree layout

Selected nodes:

Scale spin of selected nodes:

Rotate select nodes:

Change node size:

SPADE tree annotation

Auto Suggest Annotation View Annotations Remove

No show Show all Show selected Annotation bubble size: 5

overlay information by coloring nodes

select a marker: TCRab

Ref files for ratio: TCRab

add to ref remove from ref

Color definition: expr ratio cell freq

Color scheme: JET half JET Gray scale

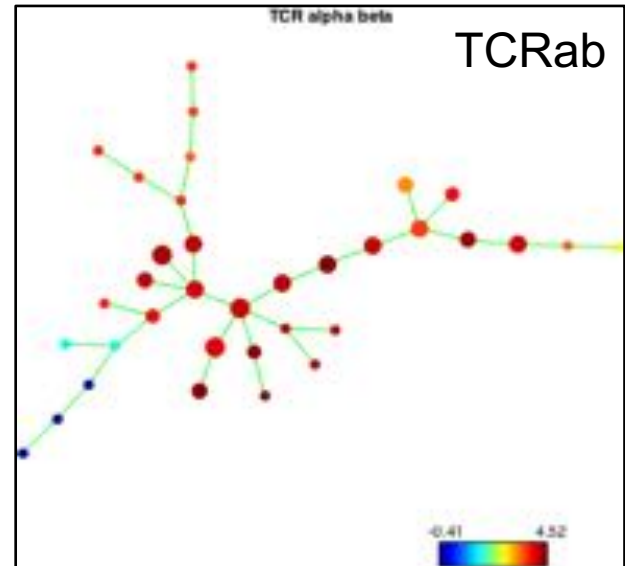
Export results to PS, GML, TXT, FCS

Tree colored by markers Tree colored by CellFreq

Tree to GML & TXT Node/Annot expr & CellFreq to TXT

Clustering result to FCS EarthMoverDist b/w samples Annot to cytobank

→ Circle size proportional to pop frequency



- Check the overlay box
- Select marker (e.g., TCRab)
- Select file (e.g., pooled)
- Select “expr” = median fluor intensity of cells in the cluster

Compare samples

Color options

overlay information by coloring nodes

Select a marker: Select a file Ref files for ratio

TCR alpha beta

CD69

CD4

CD8

RORgt

CD45RO

CD103

Foxp3

CTLA-4

TCR gamma delta

Tbet

FileInd

CellFreq

export_PsO_01041801_CD3_dump.

export_normal_lib1mq_CD3_dump.

export_normal_lib1mq_CD3_dump.

add to ref remove from ref

Color definition

expr

ratio

cell freq

Color scheme

JET

half JET

Gray scale

Export results to PS, GML, TXT, FCS formats

Tree colored by markers

Tree colored by CellFreq

Tree to GML & TXT

Node/Annot expr & CellFreq to TXT

Clustering result to FCS

EarthMoverDist b/w samples

Annot to cytobank

Or compare samples:

Expr = median fluor intensity of cells in that cluster

Ratio = the difference b/w two samples

Cell freq = freq. of cells in cluster

Compare samples

Color options

overlay information by coloring nodes

Select a marker:

- TCR alpha beta
- CD69
- CD4
- CD8
- RORgt
- CD45RO
- CD103
- Foxp3
- CTLA-4
- TCR gamma delta
- Tbet
- FileInd
- CellFreq

Select a file:

POOLED
export_PsO_01041801_CD3_dump!
export_normal_lib1mq_CD3_dump!

Ref files for ratio:

export_normal_lib1mq_CD3_dump!

add to ref remove from ref

Color definition:

expr
 ratio
 cell freq

Color scheme:

JET
 half JET
 Gray scale

Export results to PS, GML, TXT, FCS formats

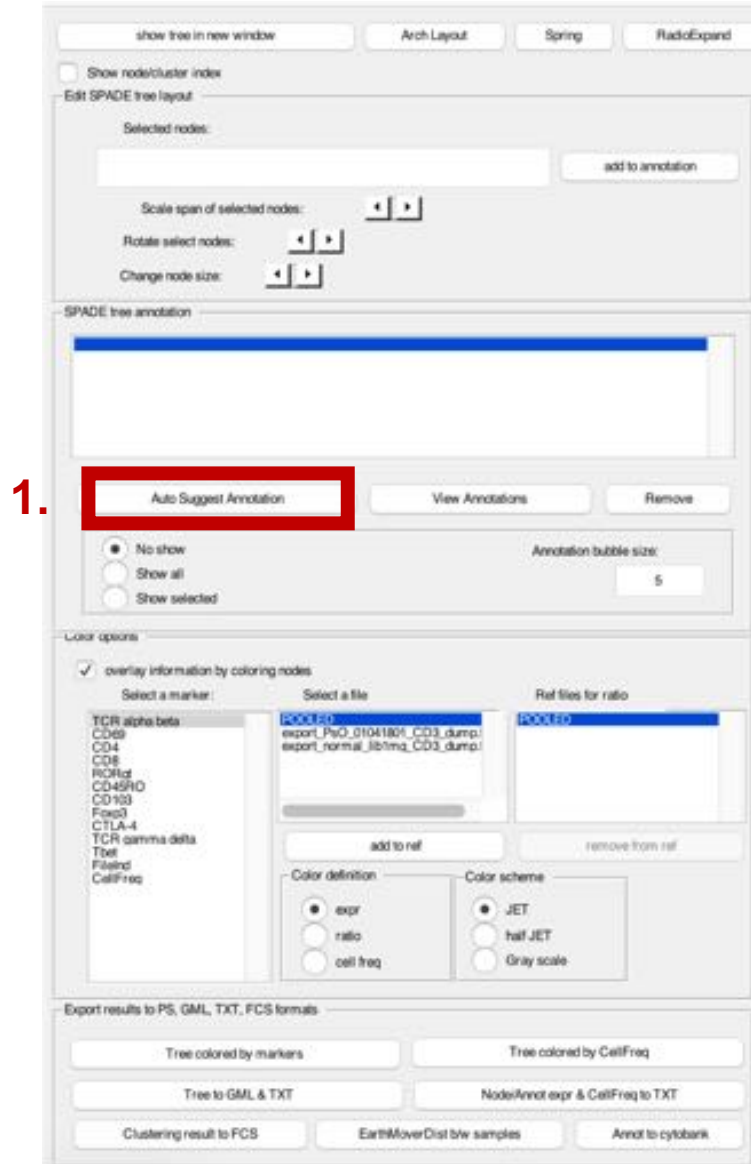
Tree colored by markers Tree colored by CellFreq

Tree to GML & TXT Node/Annot expr & CellFreq to TXT

Clustering result to FCS EarthMoverDist b/w samples Annot to cytobank

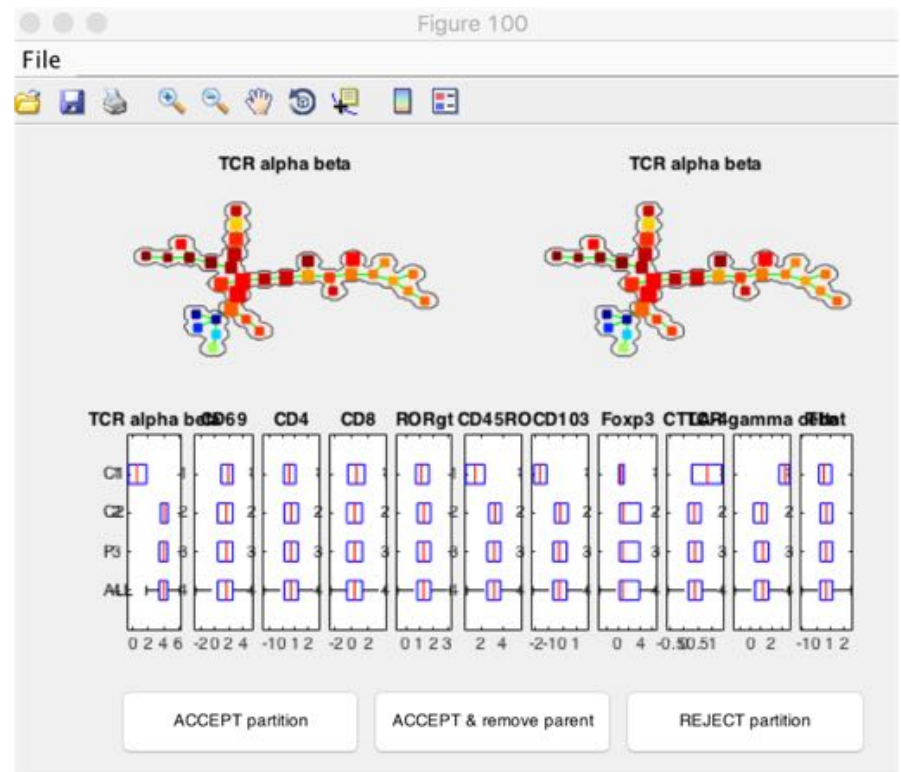
Export
SPADE trees
or files

Annotate SPADE tree (unbiased)

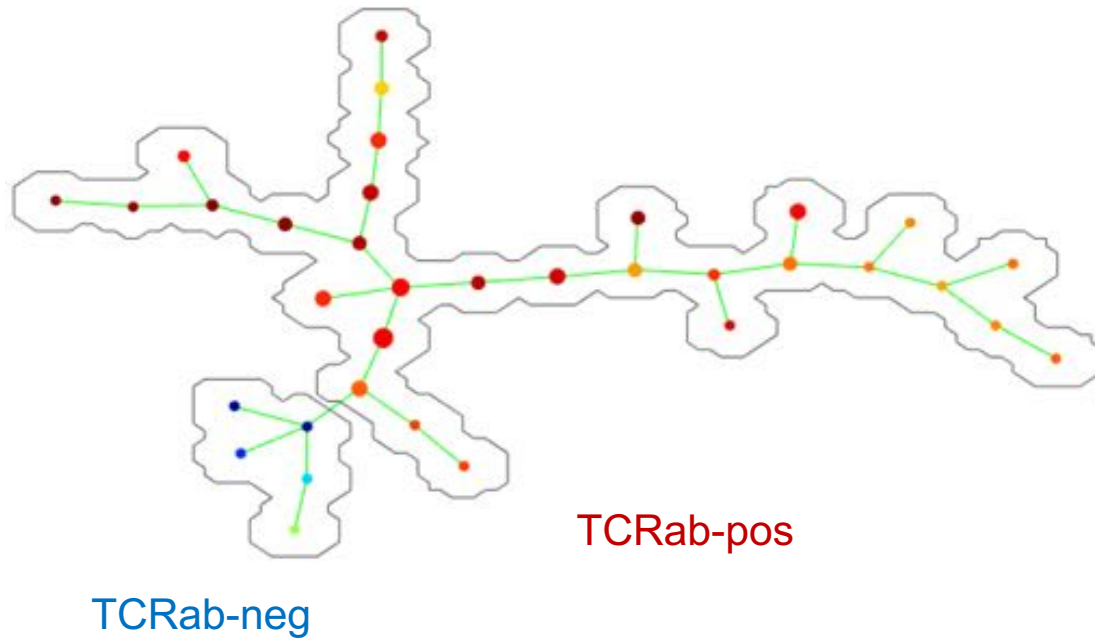


1. Select "Auto Suggestion Annotation"
2. In popup window, accept/reject proffered partition

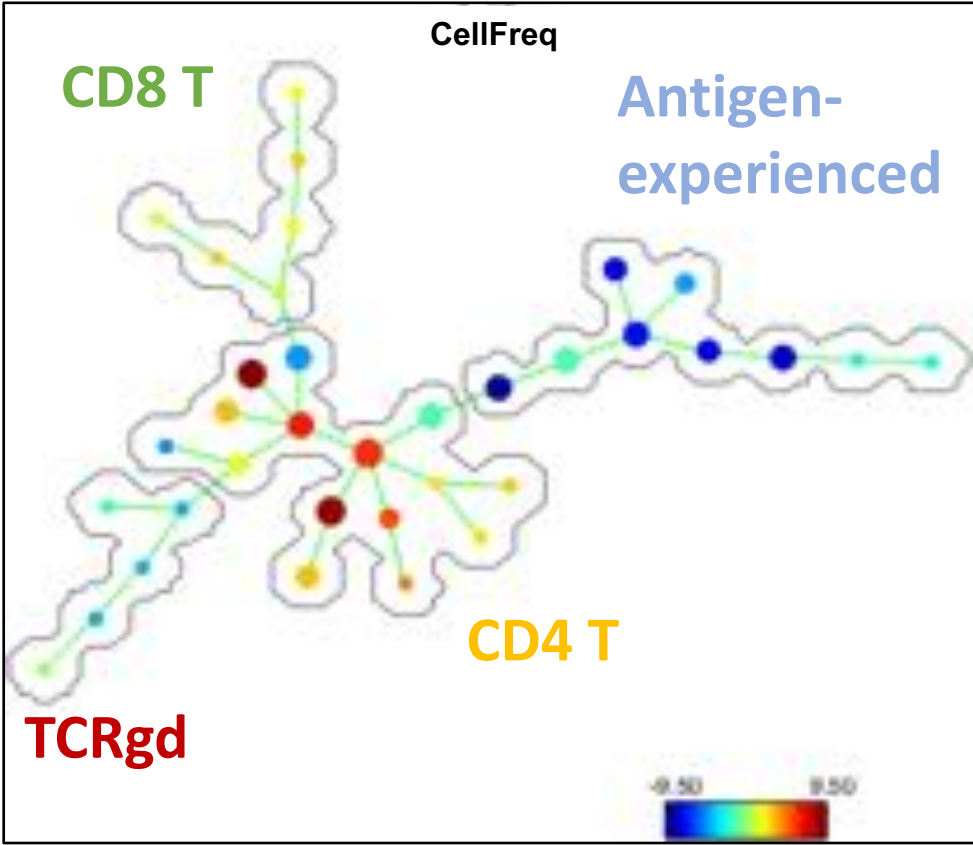
2.



Annotations (unbiased)



Changes in population frequency PsO_Stelara relative to Normal



Outline

viSNE, PhenoGraph, SPADE

how the algorithms work

maximizing comparable results across experiments

Overview of Analysis Workflow

FJ → export populations to cluster → algorithm

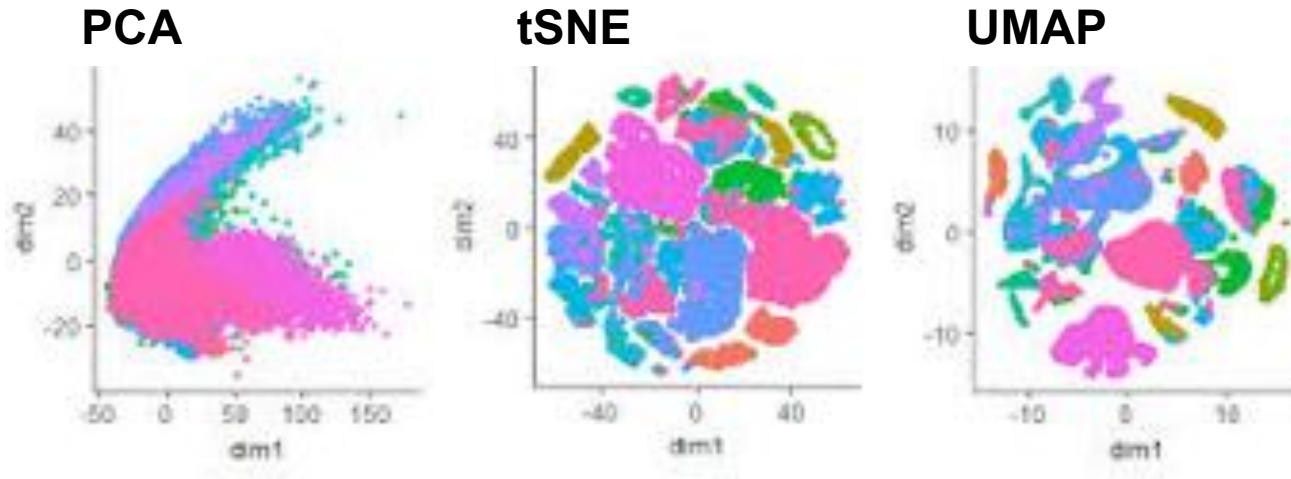
Go for it!

Cyt (viSNE, PhenoGraph) then SPADE

UMAP*

Uniform Manifold Approximation and Projection for Dimension Reduction

<https://arxiv.org/abs/1802.03426>



PCA – preserves **largest pairwise difference**, loss of local structure

tSNE – dimensionality reduction algorithm that preserves **local** and **global** structure but **NOT cluster distance**

UMAP – dimensionality reduction algorithm that preserves **local** and **global** structure, **including cluster distance**

*for MatLab savvy users who don't need a GUI; available in MathWorks as Add On

References

Reviews

1. Kimball AK, Oko LM, Bullock BL, Nemenoff RA, van Dyk LF, Clambey ET. A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data. *J Immunol.* 2018 Jan 1;200(1):3-22.
2. Saeys Y, Gassen SV, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol.* 2016 Jul;16(7):449-62.
3. Mair F, Hartmann FJ, Mrdjen D, Tosevski V, Krieg C, Becher B. The end of gating? An introduction to automated analysis of high dimensional cytometry data. *Eur J Immunol.* 2016 Jan;46(1):34-43.
4. Chester C & Maecker HT. *J Immunol.* 2015 Aug 1;195(3):773-9. doi: 10.4049/jimmunol.1500633. Algorithmic Tools for Mining High-Dimensional Cytometry Data. *J Immunol.* 2015 Aug 1;195(3):773-9.

Algorithms

1. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol.* 2011 Oct 2;29(10):886-91. **SPADE**
2. Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Amir el-AD, Davis KL, 3. Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. *Nat Biotechnol.* 2013 Jun;31(6):545-52. **viSNE**
3. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell.* 2015 Jul 2;162(1):184-97. **PhenoGraph**
4. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci U S A.* 2014 Jul 1;111(26):E2770-7. **CITRUS**
5. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol.* 2016 Jun;34(6):637-45. **Wishbone**
6. Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell.* 2014 Apr 24;157(3):714-25. **Wanderlust**

Acknowledgements

Unified Flow Core



Dewayne Falkner
Director of Operations



Aarika Yates MacIntyre
Senior Technologist



Ailing Liu
Applications Specialist

Center for Research Computing

Kim Wong

MBIS HSLS

Ansuman Chattopadhyay

Carrie Iwema

Srilakshmi Chaparala

Unified Core, Board

Mark Shlomchik

Fadi Lakkis

Mark Gladwin

Larry Moreland

John McDyer