

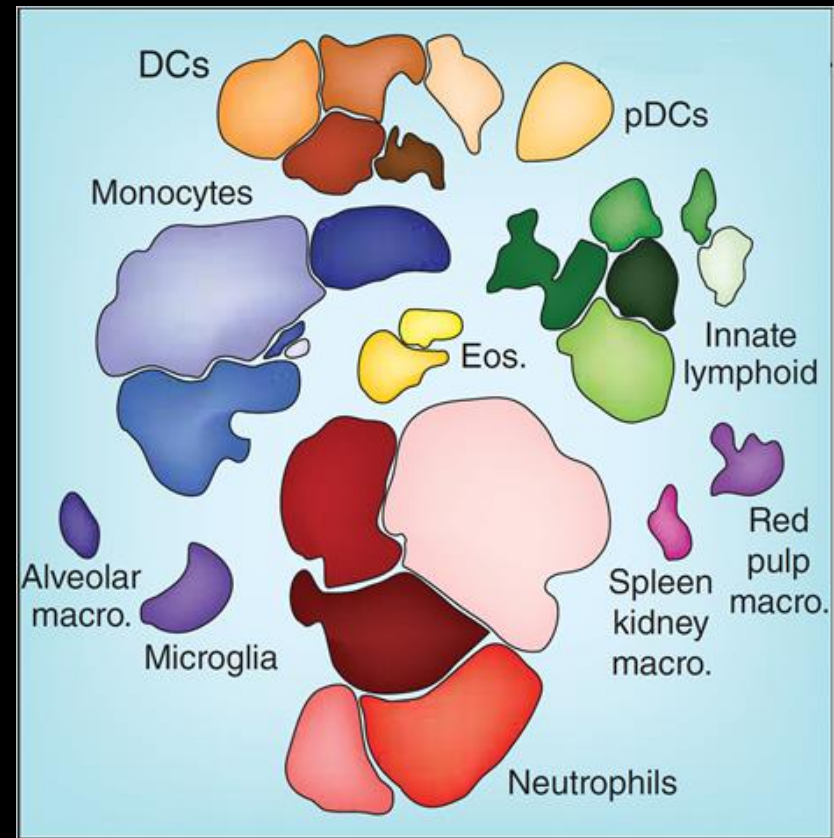
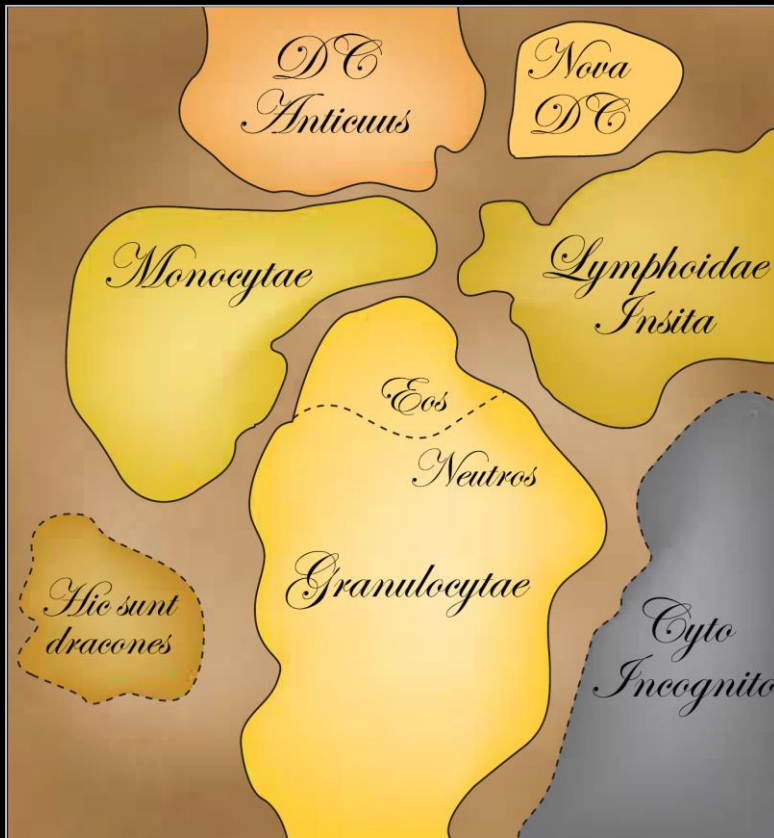
# Methods for Discovery & Characterization of Cell Subsets

*Jonathan Irish, Ph.D.*  
Vanderbilt University, Nashville, TN

Assistant Professor of Cancer Biology  
& Pathology, Microbiology & Immunology

Australasian Cytometry Society

10 October 2015



## Disclosures for Jonathan Irish, Vanderbilt University

Co-founder & board

Clinical research

Speaking honorarium

Invited speaker

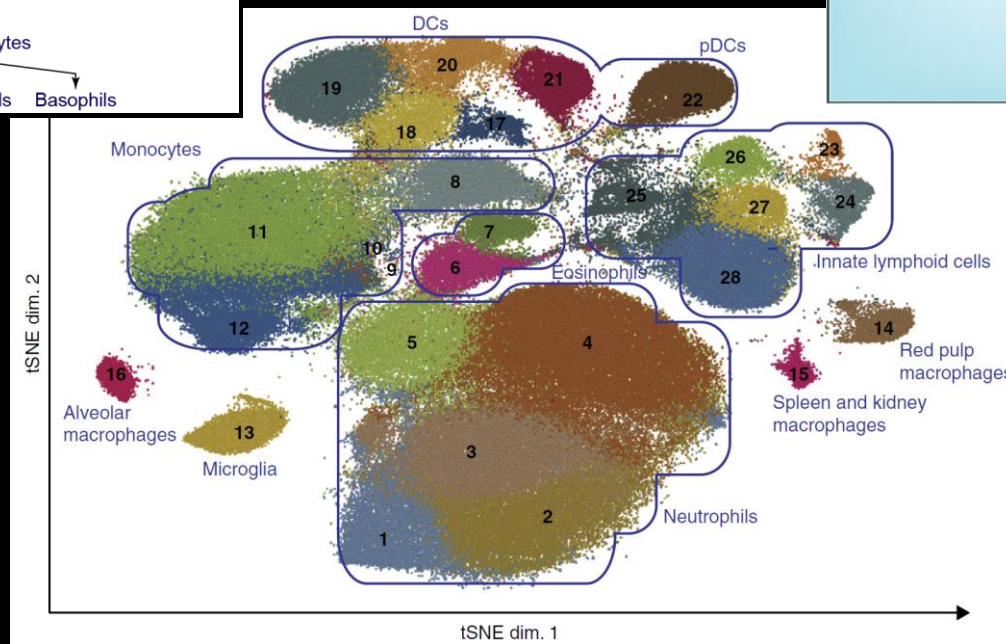
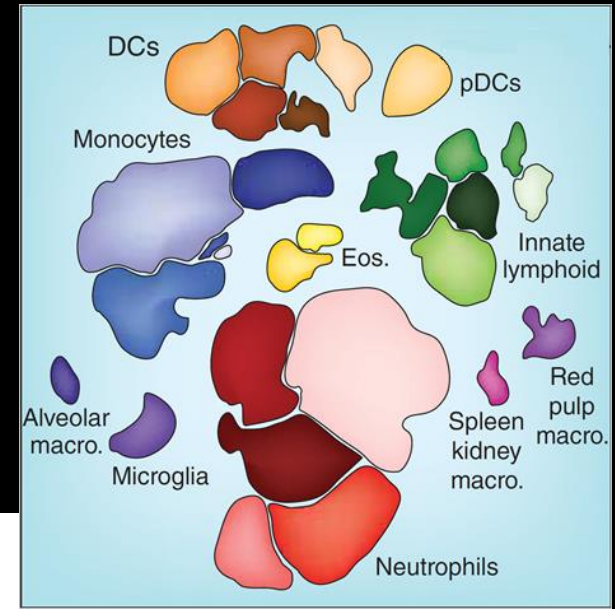
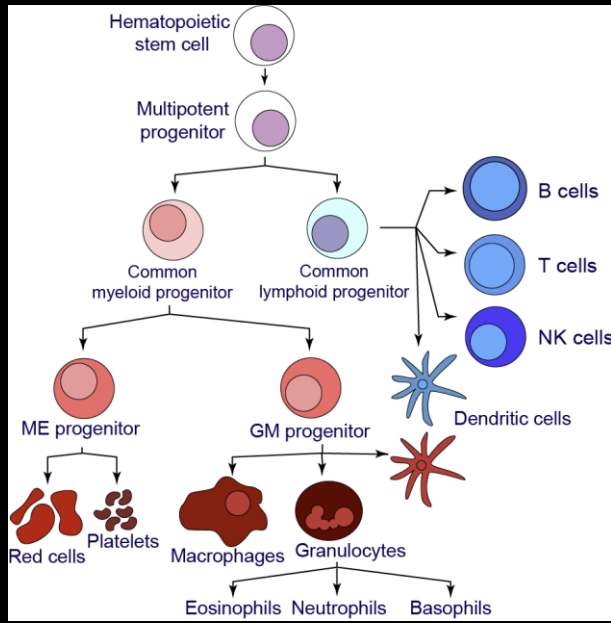
Cytobank

Incyte, Karyopharm

Novartis

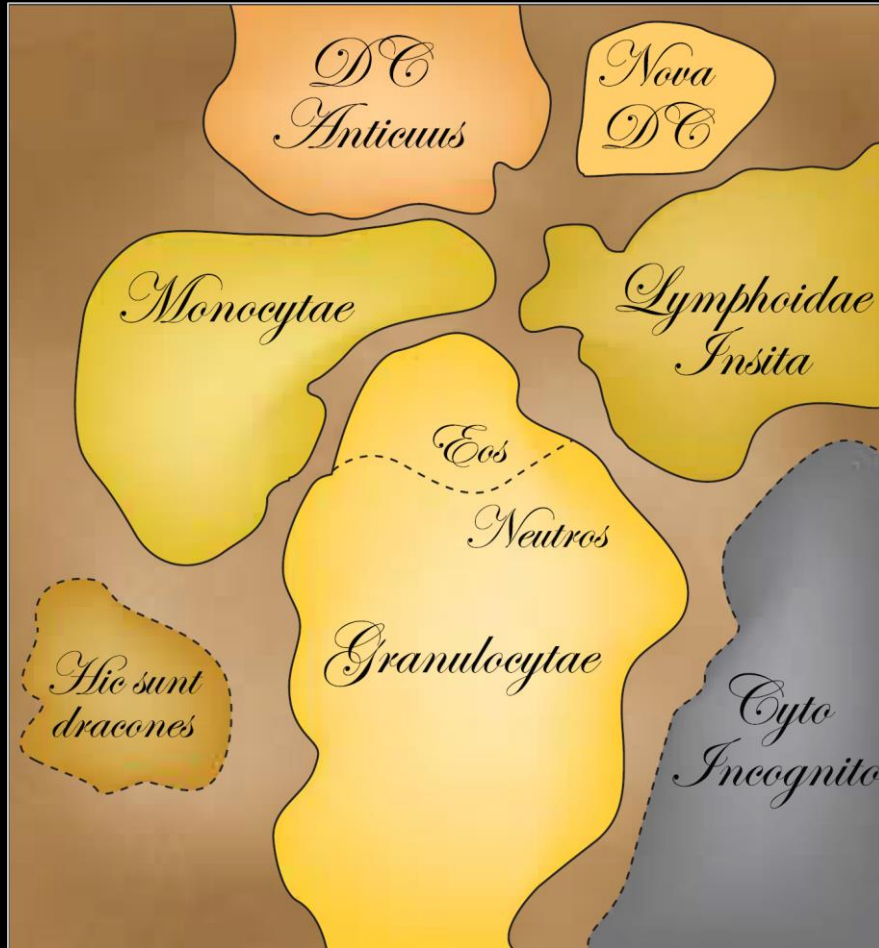
Fluidigm

# The Big Idea: Automatically Identify All Cell Types in Primary Tissues, Create Reference Models to Study Impact of Disease, Genetic Changes, etc.

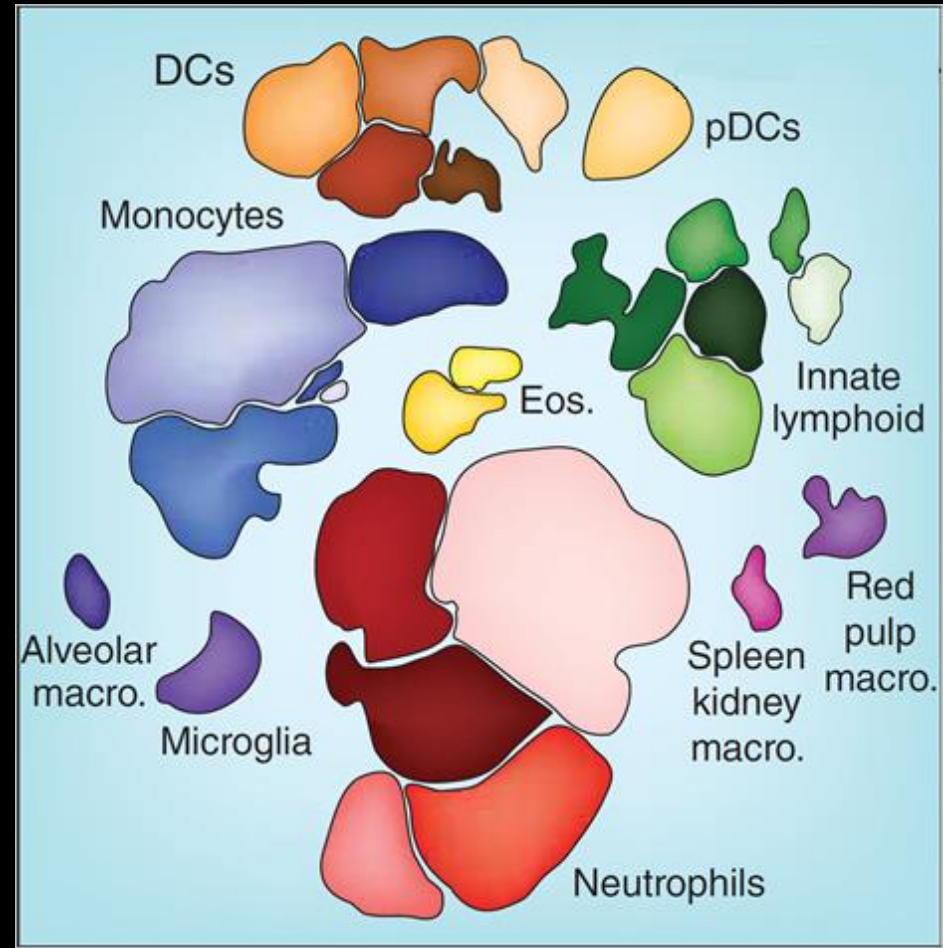


# Tools from Machine Learning + High Content Data: Comprehensive, Automatic Mapping of Cell Types

Classical map of the 'myeloid cell system'

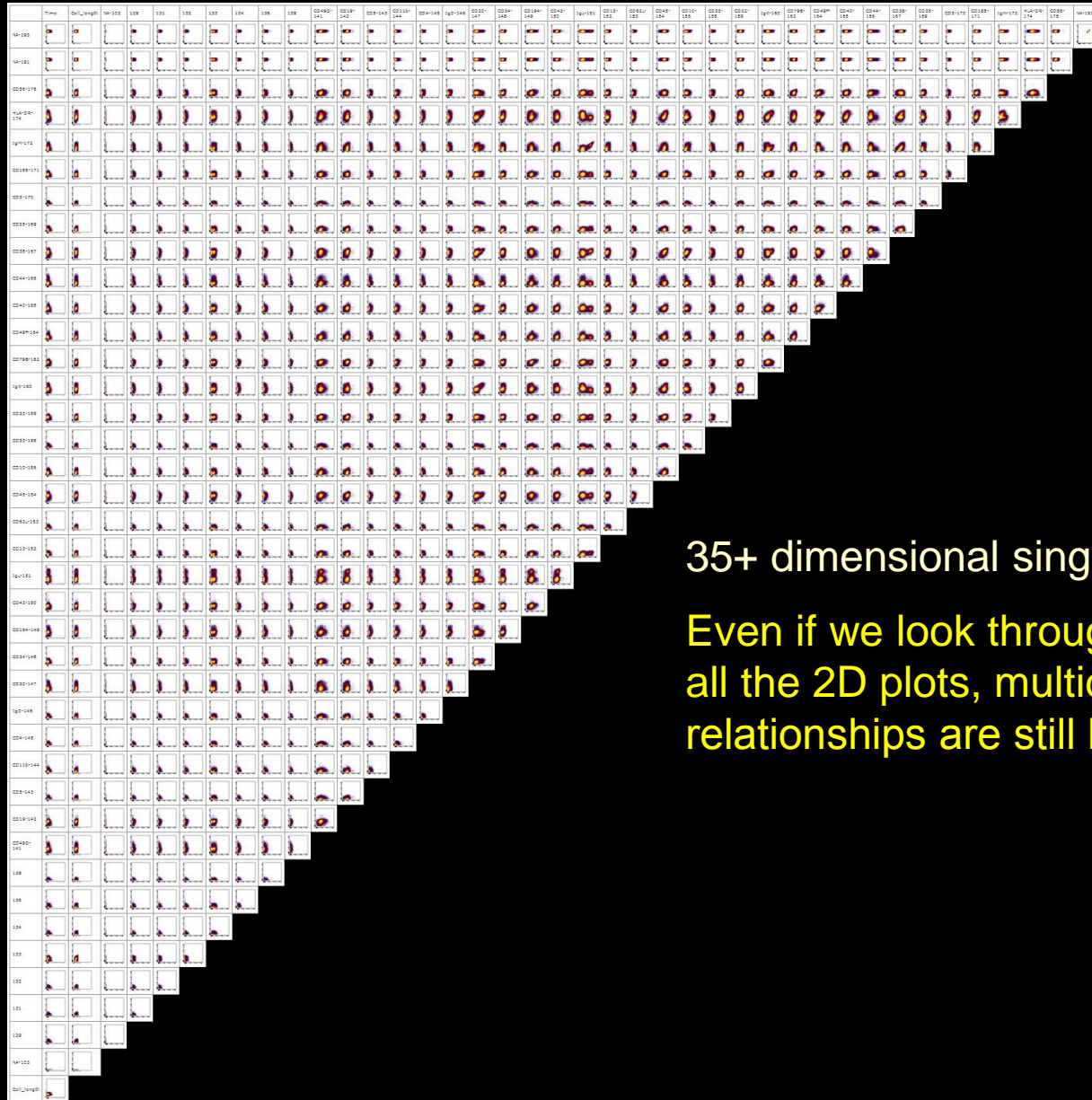


Modern map, computationally generated



Effective data analysis is critical  
to successful cytometry

# We Now Make Billions of Multi-D Single Cell Measurements => Need for Machine Learning Tools & Human Readable Views



35+ dimensional single cell data:

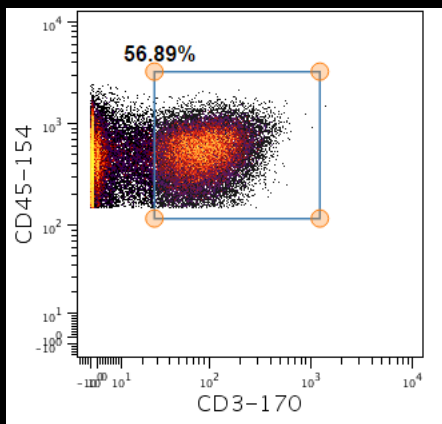
Even if we look through all the 2D plots, multidimensional relationships are still hidden...

# Unsupervised Analysis: Not Using Prior Knowledge To Guide the Analysis

Prior knowledge examples: Stem cells express CD34, AML cells express CD45

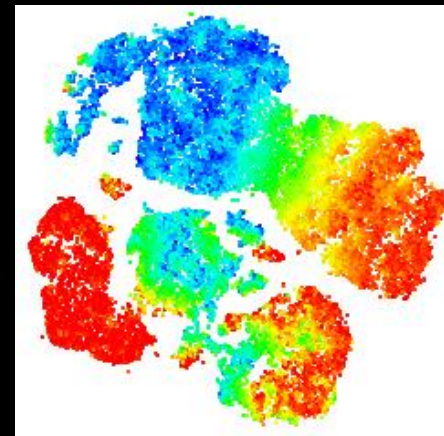
## Supervised Approaches

- Expert gating
- Gemstone
- Wanderlust
- Citrus

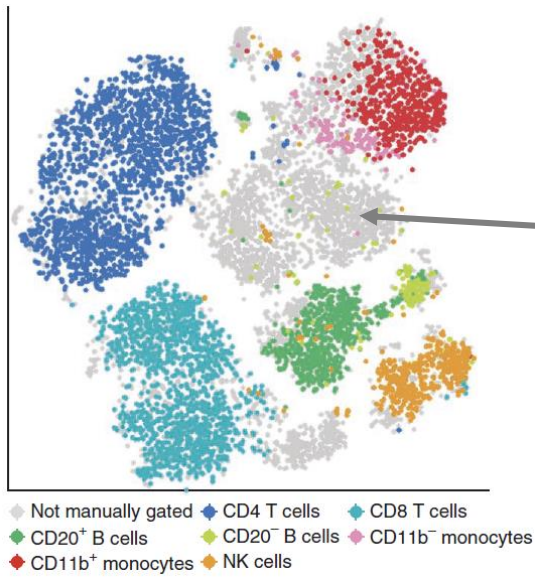


## Unsupervised Approaches

- Heatmap clustering
- SPADE
- viSNE
- Phenograph



# Traditional Gating Overlooks Many Cells in Primary Samples

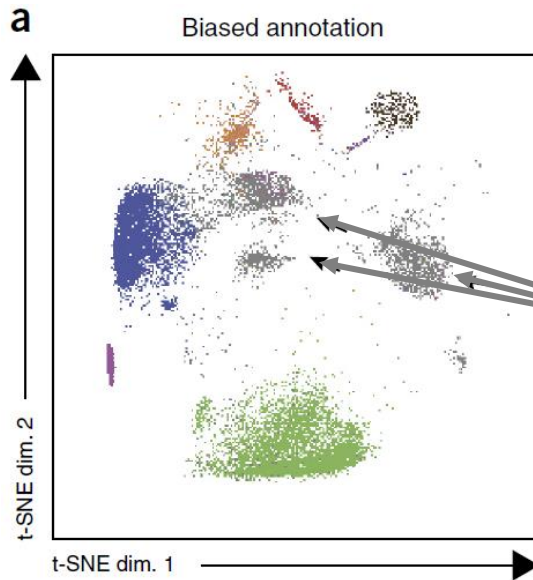


viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

El-ad David Amir<sup>1</sup>, Kara L Davis<sup>2,3</sup>, Michelle D Tadmor<sup>1,3</sup>, Erin F Simonds<sup>2,3</sup>, Jacob H Levine<sup>1,3</sup>, Sean C Bendall<sup>2,3</sup>, Daniel K Shenfeld<sup>1,3</sup>, Smita Krishnaswamy<sup>1</sup>, Garry P Nolan<sup>2,4</sup> & Dana Pe'er<sup>1,4</sup>

In all cases, the viSNE gate included cells that were not classified by the expert manually gated biaxial plots; these cells are labeled in gray in the viSNE map. Examination of the marker expression of these cells reveals that they are typically just beyond the threshold of one marker, but the viSNE classification is strongly supported on the expression of all other markers. For example, in **Figure 1d**, wherein cells are colored for CD11b marker expression, the cells in the gated region express the canonical monocyte marker CD33 (**Supplementary Fig. 1b**). However, only 47% of these cells were classified as monocytes by the manual gating (**Fig. 1b**).

nature  
biotechnology  
2013

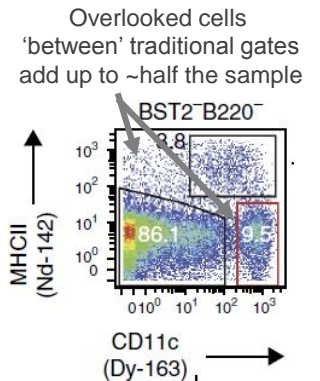


## High-dimensional analysis of the murine myeloid cell system

Burkhard Becher<sup>1,4,5</sup>, Andreas Schlitzer<sup>1,5</sup>, Jinmiao Chen<sup>1,5</sup>, Florian Mair<sup>2</sup>, Hermi R Sumatoh<sup>1</sup>, Karen Wei Weng Teng<sup>1</sup>, Donovan Low<sup>1</sup>, Christiane Ruedl<sup>3</sup>, Paola Riccardi-Castagnoli<sup>1</sup>, Michael Poidinger<sup>1</sup>, Melanie Greter<sup>2</sup>, Florent Ginhoux<sup>1</sup> & Evan W Newell<sup>1</sup>

Notably, whereas traditional biased gating strategies allowed for identification of only  $54.7 \pm 2.6\%$  (mean  $\pm$  s.e.m.,  $n = 3$  mice) of lung myeloid cells (different DC subsets, macrophages, monocytes, neutrophils), the automatic, computational approach identified nearly 100% of the cells ( $96.6 \pm 1.0\%$  (mean  $\pm$  s.e.m.,  $n = 3$  mice) accounted for by 14 predominant clusters).

nature  
immunology  
2014





# Major Steps in Most Single Cell Biology Workflows

Data collection	1) Panel design 2) Data collection
Data processing	3) Cell event parsing 4) Scale transformation
Distinguishing initial populations	5) Live single cell gating 6) Focal population gating
Revealing cell subsets	7) Feature selection 8) Dimensionality reduction 9) Identify cell clusters 10) Cluster refinement
Characterizing cell subsets	11) Feature comparison 12) Model populations 13) Learn cell identity 14) Statistical testing

How much can be automated?

Where do computers outperform humans?

How do we select tools and use them well?

# Teaching Computers To Spot Useful Patterns : Grouping Cells by Selected Features (e.g. Protein Expression)



HD cytometry!!



Woah, that's a lot of data...



Computational tools



Biological knowledge

# Many Great Tools Exist, But Key Gaps Remain

**Table 1 – A modular machine learning workflow for unsupervised high-dimensional single cell data analysis**

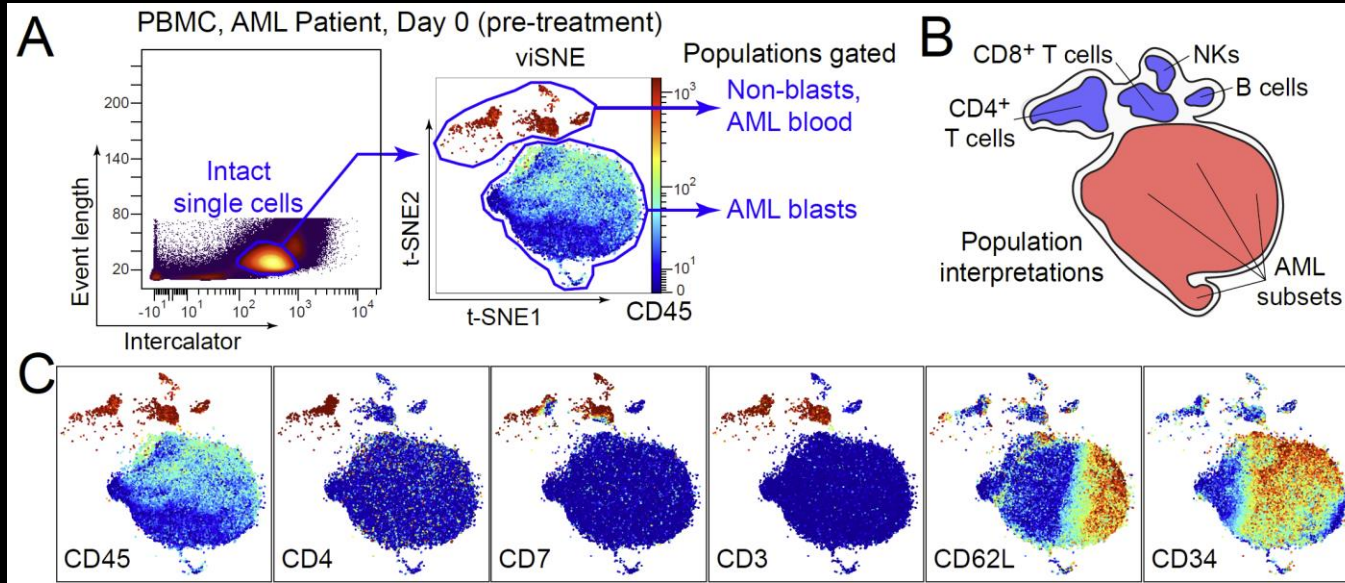
Analysis step	Traditional	Additional methods <sup>§</sup>	Method here
Data collection	1) Panel design	Human expert	-
	2) Data collection	Human expert	-
Data processing	3) Cell event parsing	Instrument software	Bead normalization and event parsing [31]
	4) Scale transformation	Human expert	Logicle [36]
Distinguishing initial populations	5) Live single cell gating	Biaxial gating + human expert	No event restriction, AutoGate [48]
	6) Focal population gating		viSNE + human expert (Figure 1) <sup>†</sup>
Revealing cell subsets	7) Select features	Human expert	Statistical threshold [40]
	8) Reduce dimensions or transform data	N/A	Heat plots [49], SPADE [12], t-SNE [50], viSNE [9], ISOMAP [23], LLE [25], PCA in R/flowCore [51]
	9) Identify clusters of cells	Human expert	SPADE, k-medians, R/flowCore, flowSOM [52], Misty Mountain [13], JCM [26], Citrus [14], ACCSENSE [53], DensVM [24], AutoGate
	10) Cluster refinement	Human expert	Citrus, DensVM, R/flowCore
Characterizing cell subsets	11) Feature comparison	Select biaxial single cell views	viSNE, SPADE, Heatmaps [34, 40], Histogram overlays [34, 40], Violin or box and whiskers plots [51]
	12) Model populations	N/A	JCM, PCA
	13) Learn cell identity	Human expert	-
	14) Statistical testing	Prism, Excel	R/flowCore

<sup>§</sup>Methods with broad application (e.g. R/flowCore) are listed minimally at select steps based on particular strengths or published applications.

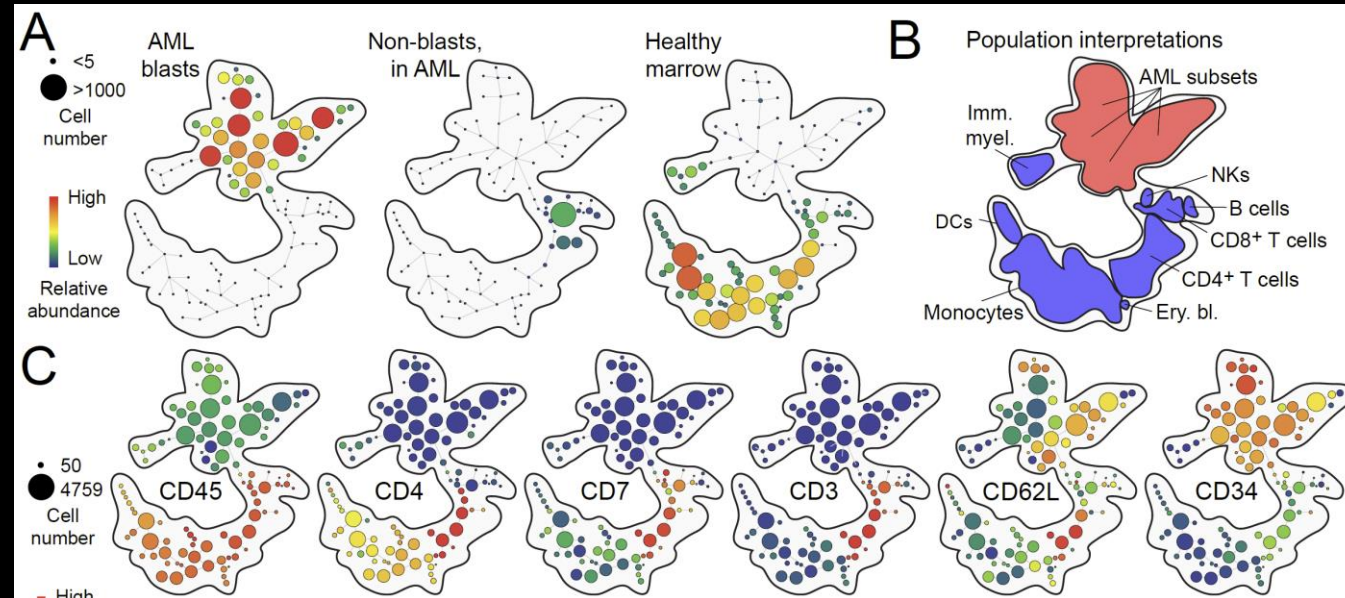
<sup>†</sup>Denotes the primary approach used at each step in the sequential analysis workflow shown here.

A major gap in the field is in true learning of cell identity

# Key Analysis Concepts: Dimensionality Reduction, Transformation, Clustering, Modeling, Visualization, & Integration



**viSNE**  
Amir et al.  
*Nature biotech* 2013

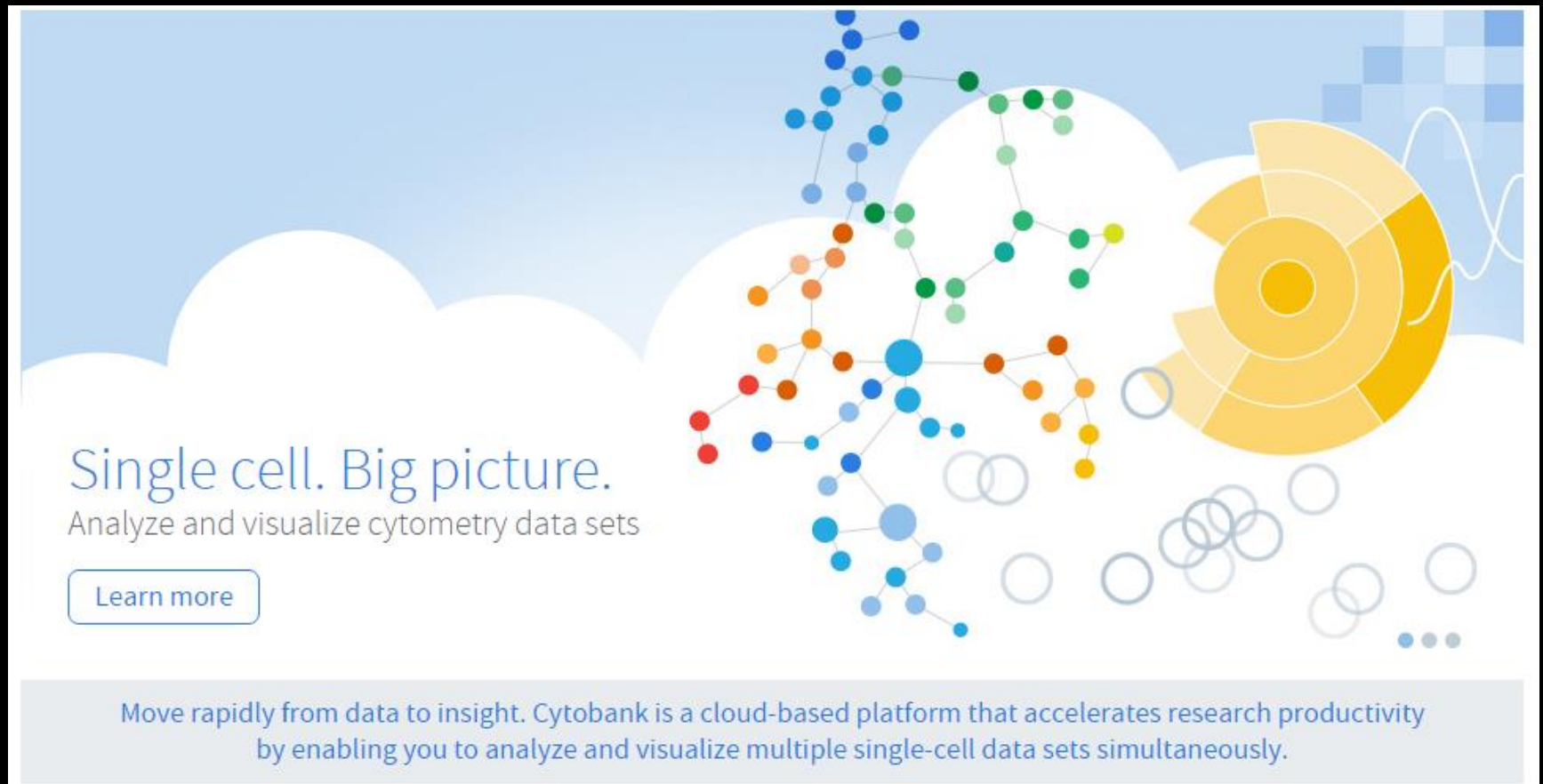


**SPADE**  
Qiu et al.  
*Nature biotech* 2011

# We Will Use Cytobank Software for viSNE & SPADE

Cytobank ([www.cytobank.org](http://www.cytobank.org)) is a commercial tool for web-based data storage, annotation, analysis, and visualization

30-day free trial with viSNE & SPADE: <https://premier.cytobank.org/signup>

The advertisement features a light blue background with white cloud-like shapes at the bottom. On the right side, there is a complex network graph with nodes in various colors (blue, green, orange, red) and a circular sunburst chart with yellow and orange segments. The text is positioned on the left side of the banner.

Single cell. Big picture.  
Analyze and visualize cytometry data sets

[Learn more](#)

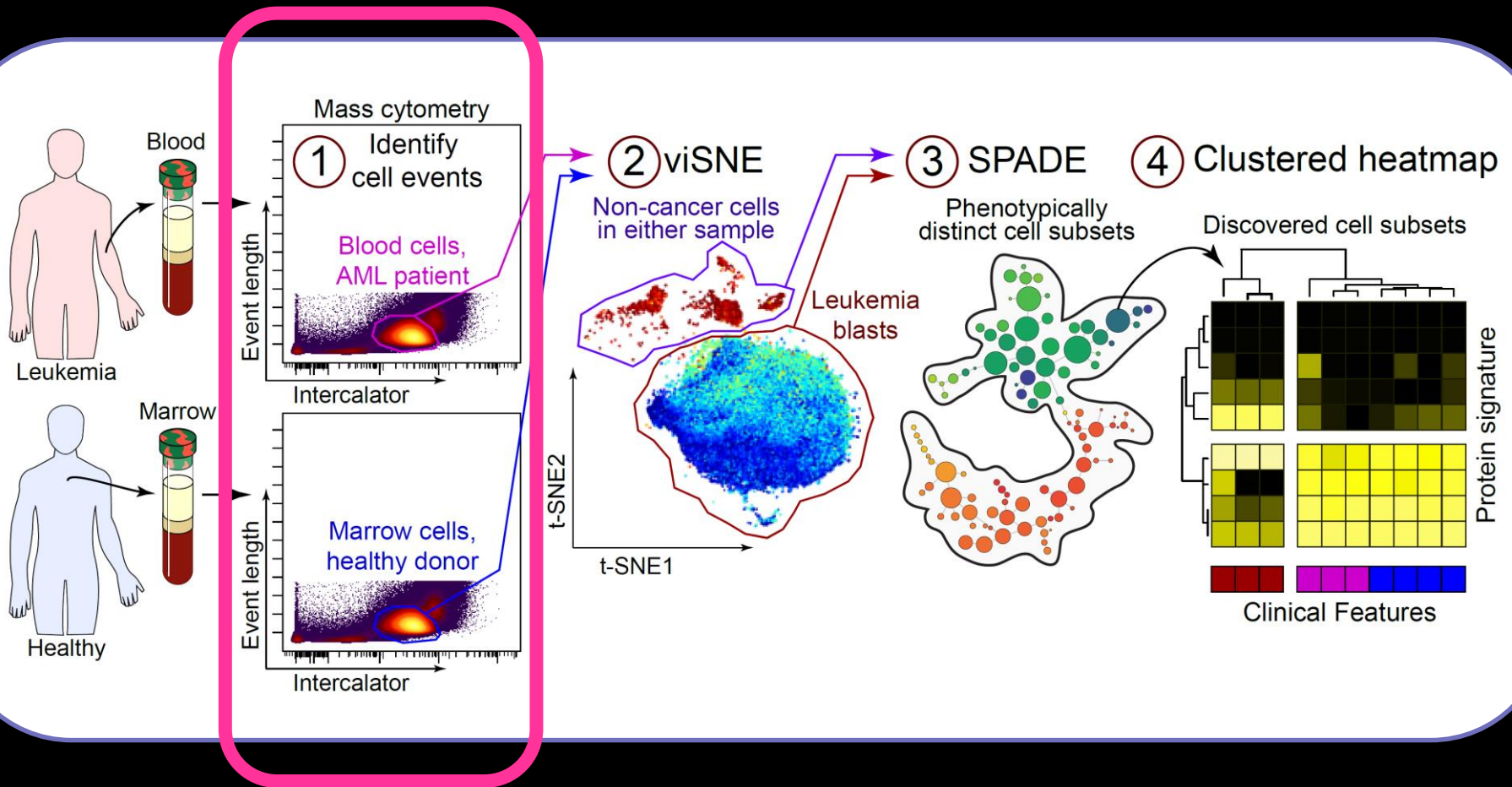
Move rapidly from data to insight. Cytobank is a cloud-based platform that accelerates research productivity by enabling you to analyze and visualize multiple single-cell data sets simultaneously.

# Discussion Questions Covered in Today's Course

- 1) What are key differences between tools (viSNE, SPADE, PCA)? What is the difference between transforming, clustering, and modeling data? What type of modeling are we doing (if any)?
- 2) What does non-linear vs. linear analysis mean? Does the data's scale matter for analysis (arcsinh5, arcsinh15, linear)?
- 3) What do viSNE and SPADE settings do (viSNE iterations, SPADE downsampling & node #)? When should they be changed?
- 5) How does one compare new samples with a prior analysis? How do we test tools with expert gating?
- 6) What are some "red flags" indicating problems? What does a good viSNE or SPADE analysis run look like?

Onward, to the analysis!

# Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

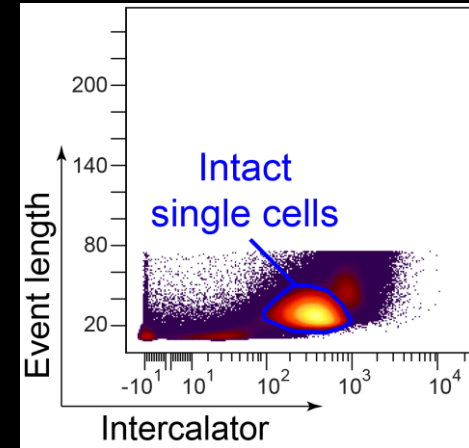




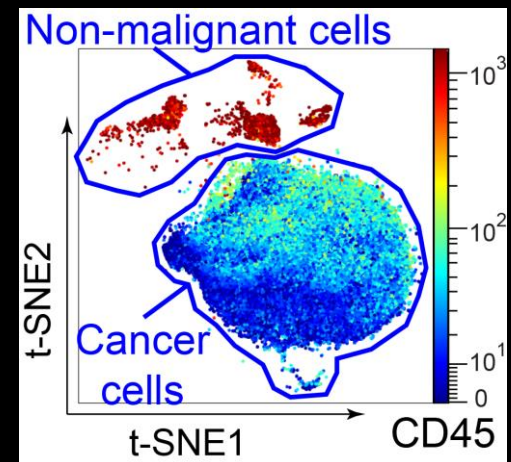
# Single Cell Biology Workflow

Data collection	1) Panel design 2) Data collection
Data processing	3) Cell event parsing 4) Scale transformation
Distinguishing initial populations	5) Live single cell gating 6) Focal population gating
Revealing cell subsets	7) Feature selection 8) Dimensionality reduction 9) Identify cell clusters 10) Cluster refinement
Characterizing cell subsets	11) Feature comparison 12) Model populations 13) Learn cell identity 14) Statistical testing

## Expert gating



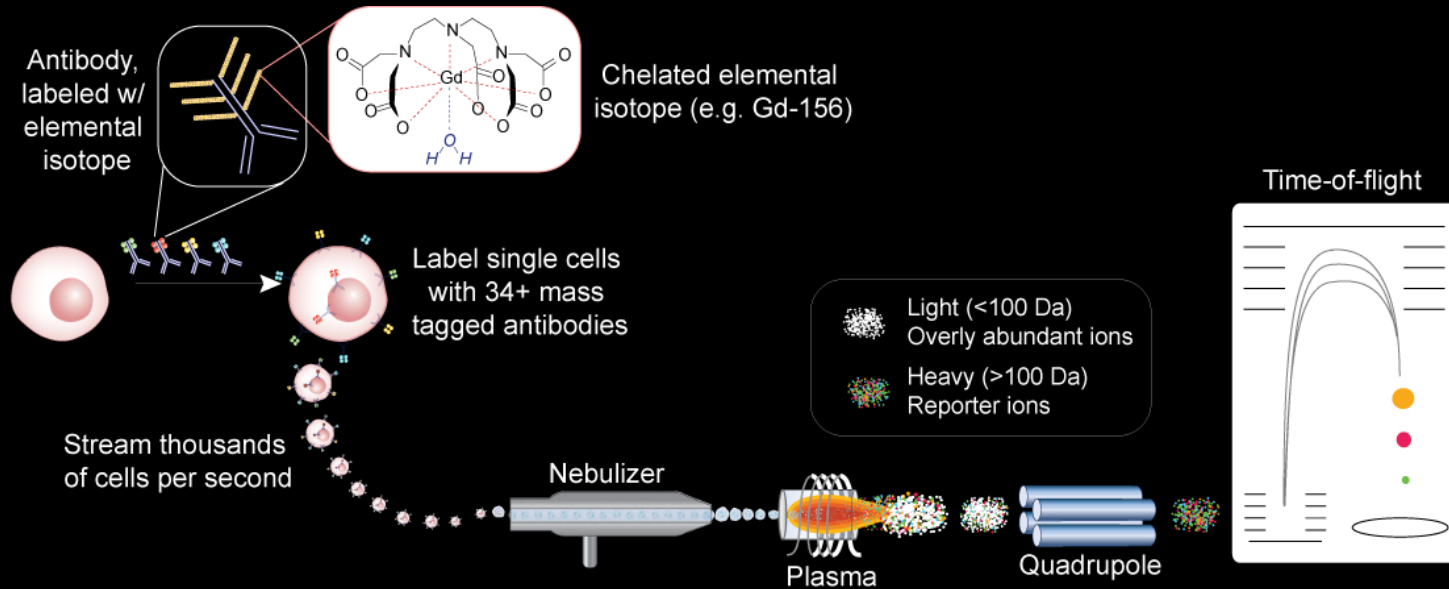
## viSNE + expert gating



# Mass Cytometry: 35+ Dimensional Analytical Cytometry



Vanderbilt Mass Cytometer



# Mass Cytometry Data Pre-Processing

Data Acquisition  
(IMD -> FCS)

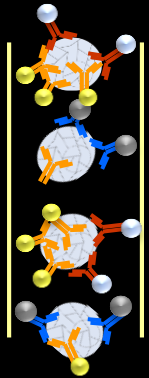
Concatenation

Normalization

Debarcoding

Transformation

Analysis



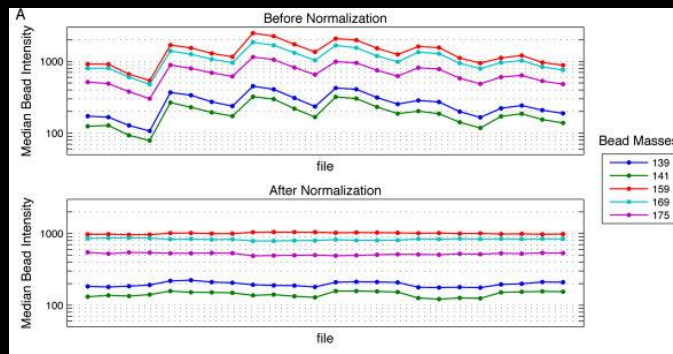
FCS File 1

+

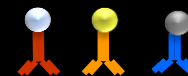
FCS File 2

+

FCS File 3



Bead Masses



$\text{arcsinh}(x/15)$

Fink et al, *Cytometry Part A* 2013

## Resources:

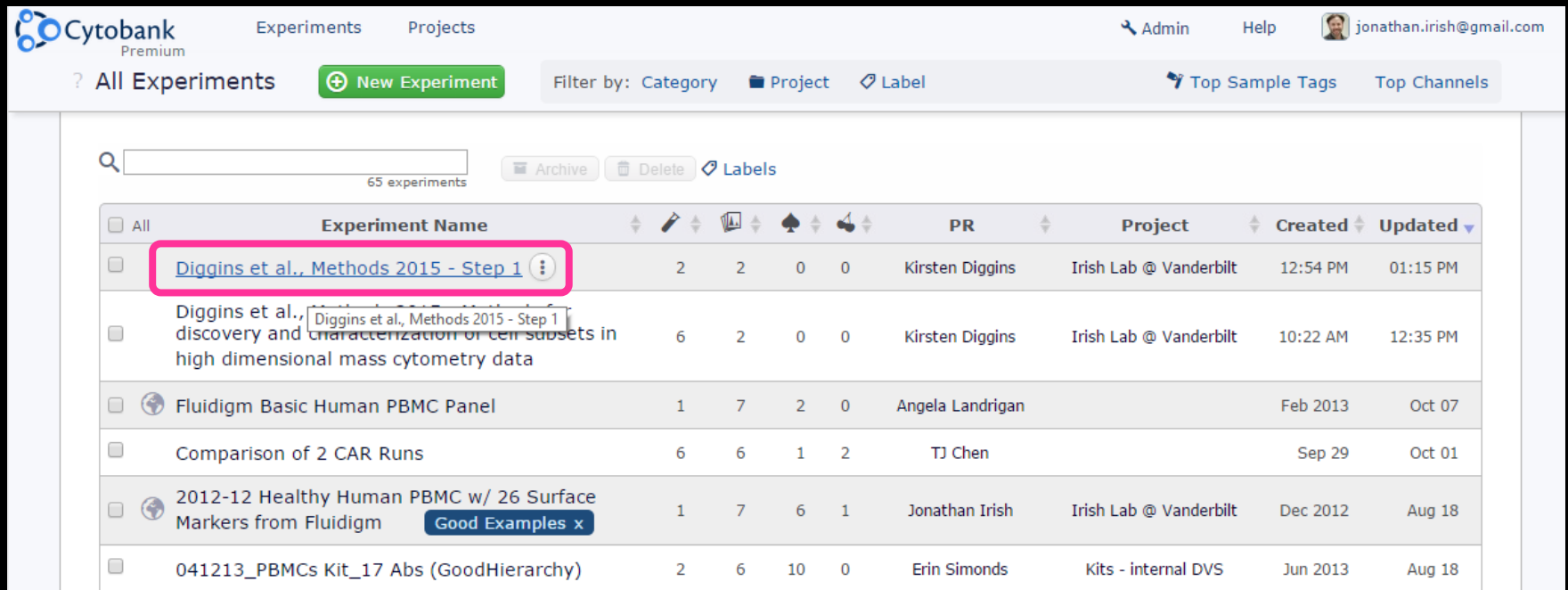
- **Concatenation:** downloadable tool from Cytobank (<http://support.cytobank.org/help/kb/cytobank-utilities/concatenating-fcs-files>)
- **Normalization:** *Cytometry Part A* [Volume 83A, Issue 5](#), pages 483-494, 19 MAR 2013 DOI: 10.1002/cyto.a.22271 <http://onlinelibrary.wiley.com/doi/10.1002/cyto.a.22271/full#fig6>
- **Barcoding:** Bodenmiller et al, *Nature Biotechnology* 2012 (<http://www.nature.com/nbt/journal/v30/n9/full/nbt.2317.html>)

# Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 1”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Gate for “Intact cells”

- Navigate to Diggins et al., Methods 2015 – Step 1
- Click the experiment name to load it



The screenshot shows the Cytobank Premium interface. At the top, there are navigation tabs for 'Experiments' and 'Projects', and a user profile for 'jonathan.irish@gmail.com'. Below the navigation, there are filters for 'All Experiments', 'New Experiment', and 'Filter by: Category, Project, Label'. A search bar is present with '65 experiments' results. The main content is a table of experiments with columns for 'Experiment Name', 'PR', 'Project', 'Created', and 'Updated'. The first row, 'Diggins et al., Methods 2015 - Step 1', is highlighted with a pink box. A tooltip is visible over the experiment name, showing the full name: 'Diggins et al., Methods 2015 - Step 1'.

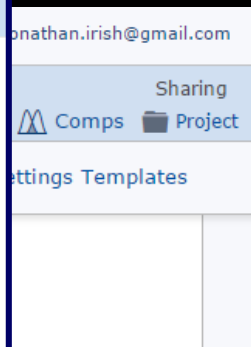
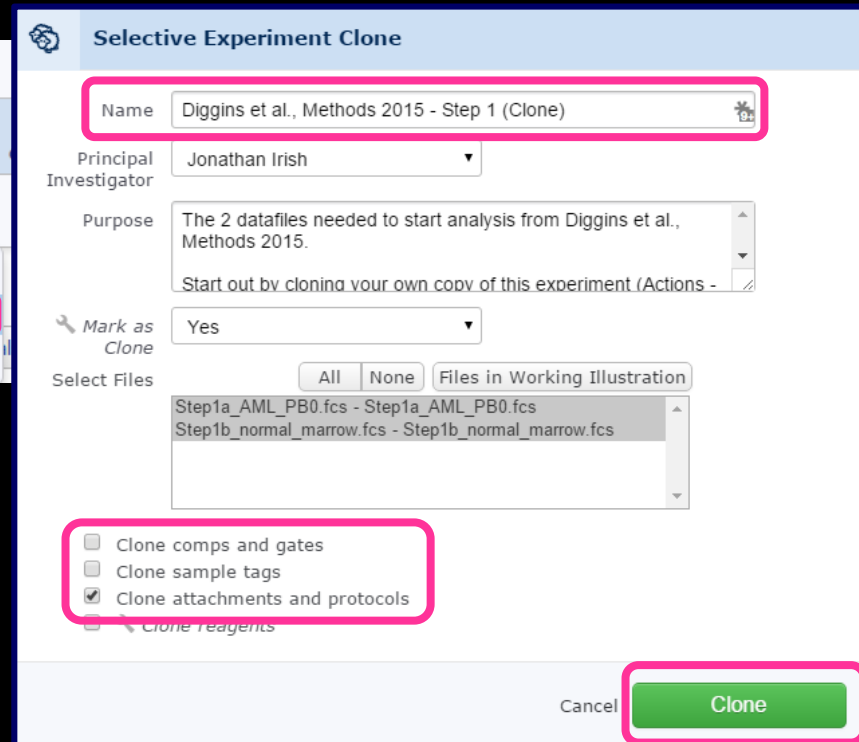
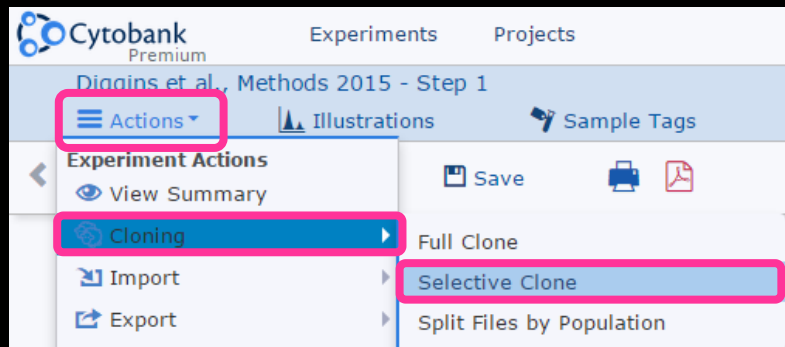
Experiment Name	PR	Project	Created	Updated
Diggins et al., Methods 2015 - Step 1	Kirsten Diggins	Irish Lab @ Vanderbilt	12:54 PM	01:15 PM
Diggins et al., discovery and characterization of cell subsets in high dimensional mass cytometry data	Kirsten Diggins	Irish Lab @ Vanderbilt	10:22 AM	12:35 PM
Fluidigm Basic Human PBMC Panel	Angela Landrigan		Feb 2013	Oct 07
Comparison of 2 CAR Runs	TJ Chen		Sep 29	Oct 01
2012-12 Healthy Human PBMC w/ 26 Surface Markers from Fluidigm	Jonathan Irish	Irish Lab @ Vanderbilt	Dec 2012	Aug 18
041213_PBMCs Kit_17 Abs (GoodHierarchy)	Erin Simonds	Kits - internal DVS	Jun 2013	Aug 18

# Steps 1-3: Getting Started & Preparing for viSNE

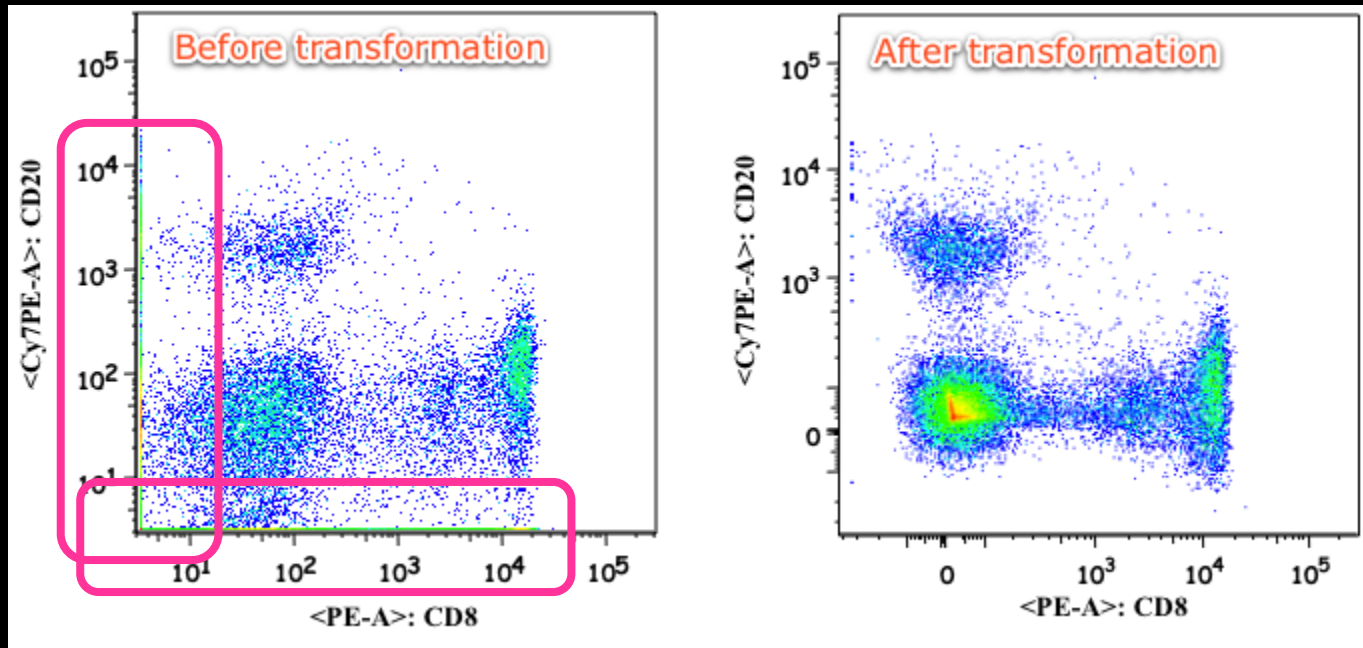
Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 1”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Gate for “Intact cells”

- Make a clone using Actions => Cloning => Selective Clone
- Edit the name, uncheck all except “clone attachments and protocols”, press Clone button.



Have you ever noticed two peaks within the cells that are biologically 100% negative for a marker?



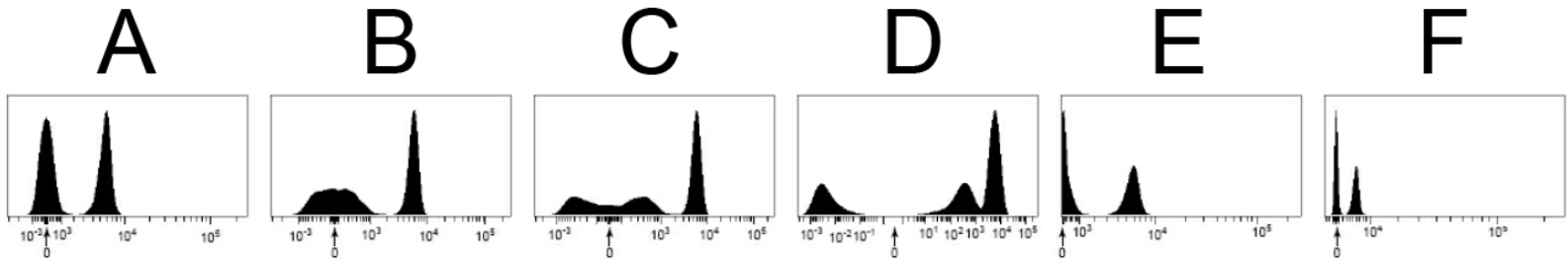
<http://www.flowjo.com/v76/en/displaytransformwhy.html>

Results from bad scaling (poor transformation) and it can be an issue for computational analysis.

Scaling is important in both mass and fluorescence cytometry.

# Scaling Matters for Measuring Distance

A 50:50 mix of + and - events stained only for PerCP-Cy5.5 is shown using different scales.



	A	B	C	D	E	F
Min:	-3000	-3000	-3000	-3000	1	-3000
Cofactor:	2500	500	150	1	2500	10,000
Result:	Good	Poor	Bad	Very Bad	Bad	Poor
Issue:		 Under-transformed			Off scale	Over-transformed

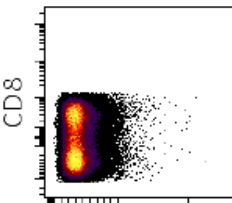
$$\operatorname{arcsinh}(x) \text{ with cofactor } c = \ln\left(\frac{x}{c} + \sqrt{1 + \left(\frac{x}{c}\right)^2}\right)$$

For fluorescent flow cytometry data a biexponential or arcsinh transformation corrects the scale near zero.

Since computational analysis techniques compare distance similar to what a person does when looking at a plot, these techniques can identify artificial populations near zero (see C and D) if data are not appropriately transformed prior to analysis.

# Inappropriate Scaling Can Lead to False Population Discovery

Rendering Controls



CD8

FSC-A

Sample  
Specimen\_001\_A\_34\_17\_CD4+ CD8- fc

Use left/right arrow keys to flip

Plot Controls

Plot Type  
Density

y-Channel  
CD8

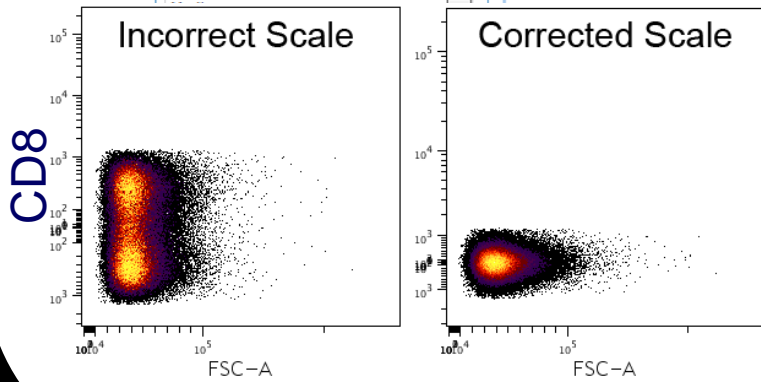
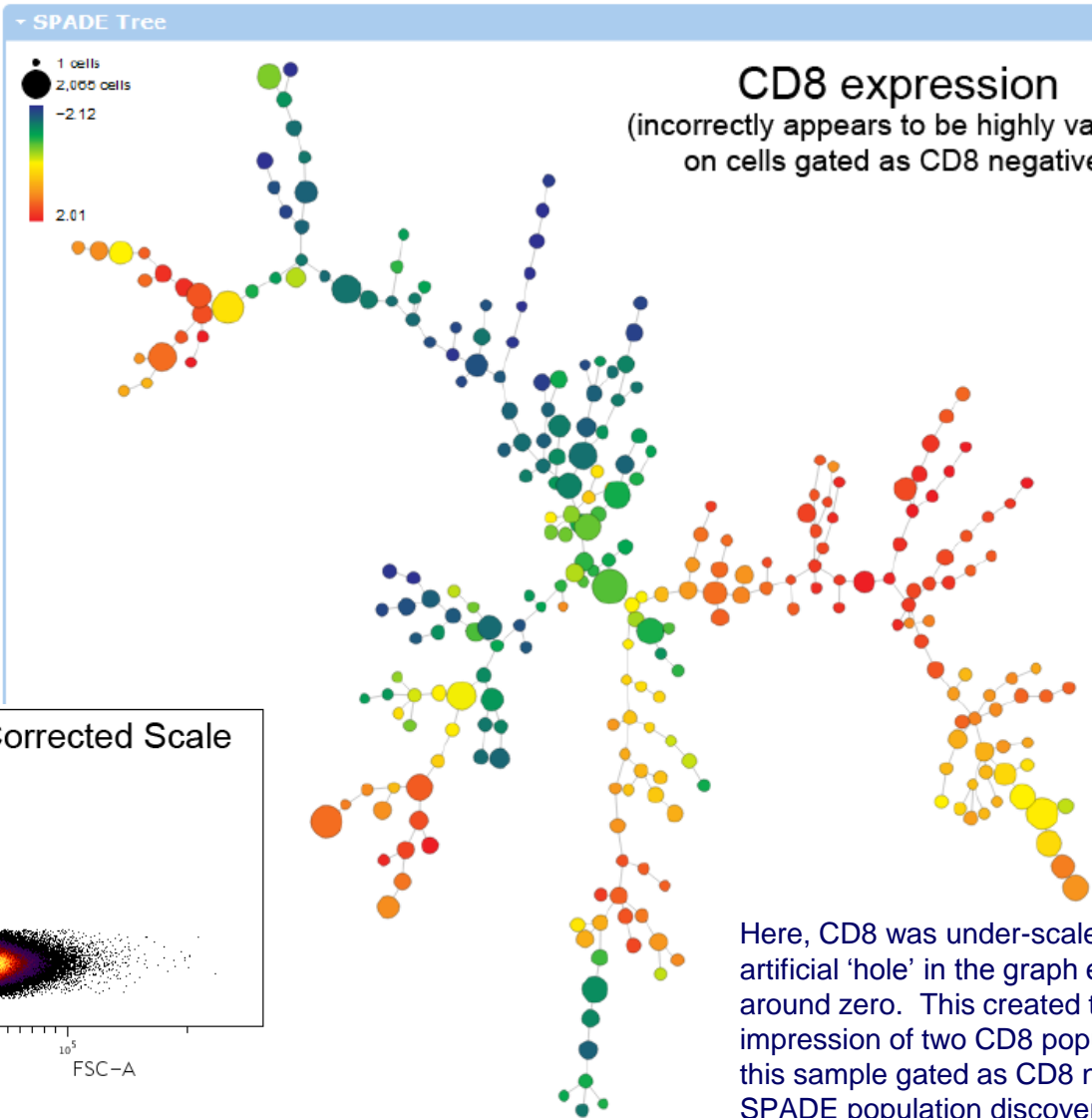
x-Channel  
FSC-A

Compensation  
File Internal Compensation

SPADE Tree Controls

Select All Nodes

Coloring Attribute



Here, CD8 was under-scaled so that an artificial 'hole' in the graph existed around zero. This created the false impression of two CD8 populations in this sample gated as CD8 negative. SPADE population discovery treated this as significant.



# Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 1”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Gate for “Intact cells”

- Open the Scales and set the scale argument to 15 for all channels, then press Apply.
- Choose “OK” when the popup asks if you really want to do it.

The screenshot shows the Cytobank Premium interface for the experiment 'Diggins et al., Methods 2015 - Step 1 (Clone)'. The 'Scales' tab is active, and the 'Experiment Scales' configuration window is open. The 'Scale Argument' is set to 15, and the 'Apply' button is highlighted. Below the configuration fields is a table of channel settings.

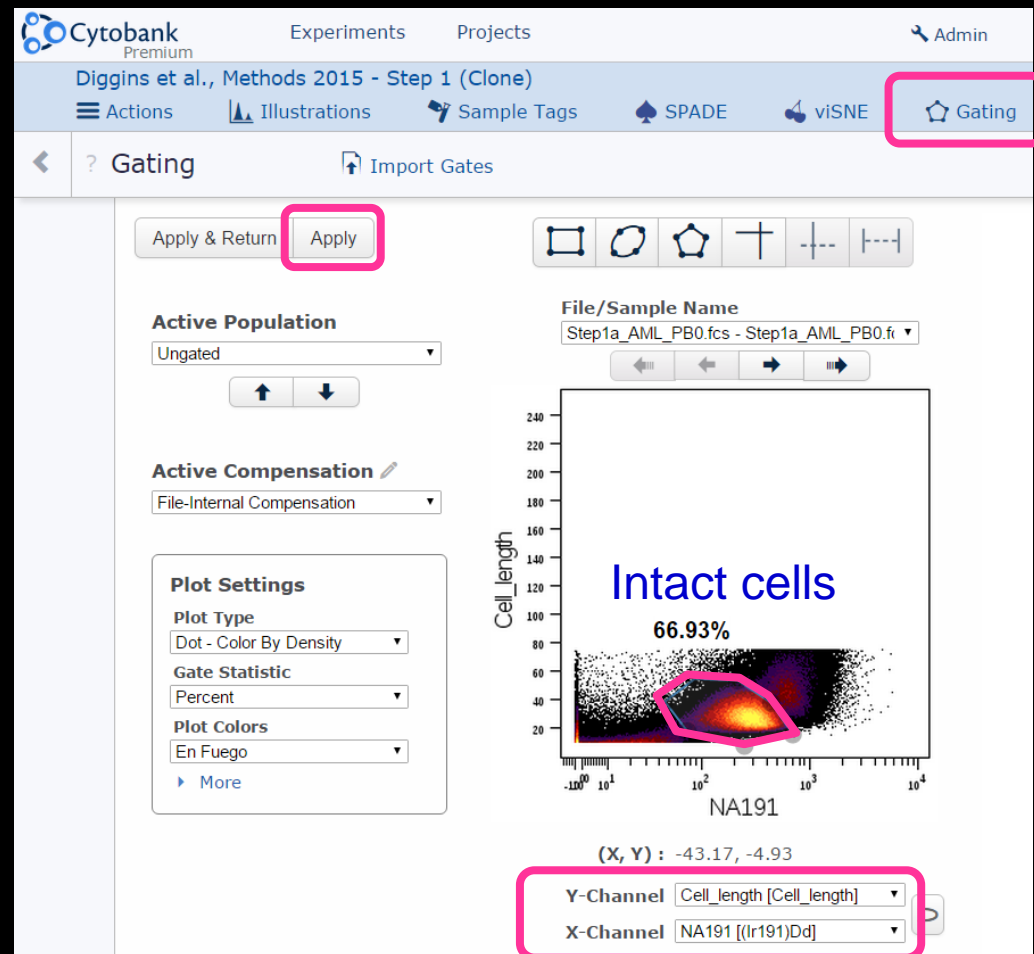
Channel	Scale	Scale Argument	Minimum	Maximum
Barium [(Ba137)Dd]	Arcsinh	5	-5.0	12000.0
Barium [(Ba138)Dd]	Arcsinh	5	-5.0	12000.0
CD15-164 [(Dy164)Dd]	Arcsinh	5	-5.0	12000.0

# Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 1”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Gate for “Intact cells”

- Go into Gating, change the y-axis to “Cell Length” (Event Length) and the x-axis to NA191 (Intercalator).
- Draw a polygon gate like the one below and call it “Intact cells”.
- Apply the gate.



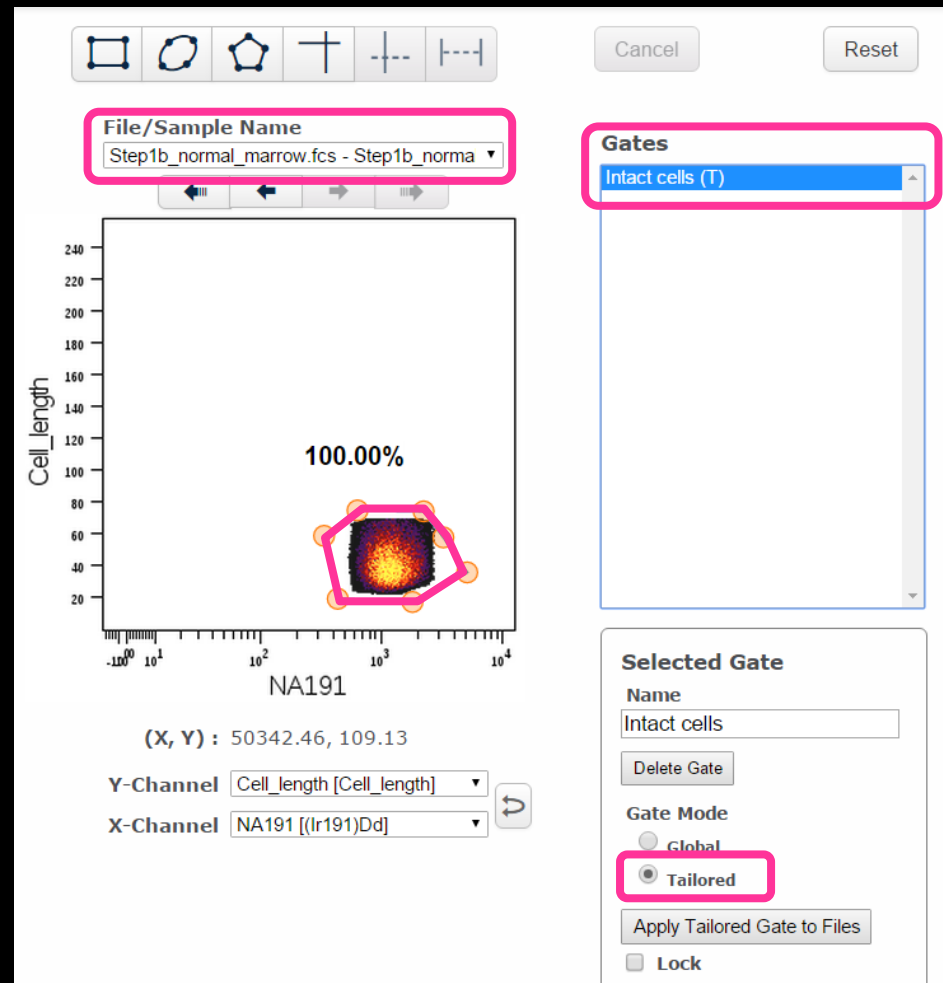
# Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:

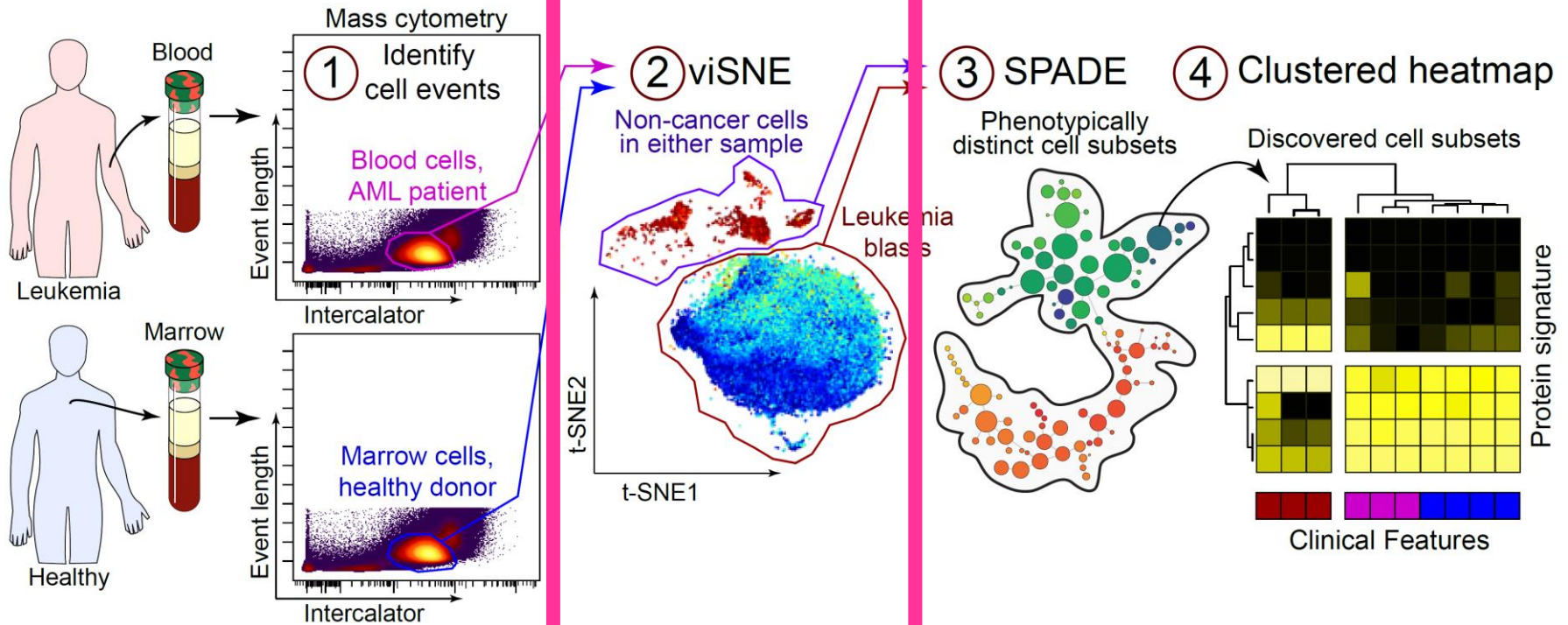
- 1) Clone experiment “Diggins et al., Methods 2015 – Step 1”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Gate for “Intact cells”

- Click on the “Intact cells” gate in the list of gates to the right.
- Make the gate a tailored gate by clicking the “Tailored” radio circle.
- Select the other file (Step1b\_normal\_marrow)
- Move the gate’s vertices to include 100% of normal marrow events.
- Make sure to to Apply the gate. Optional: “Check gate”.

Note: we’re doing this because intercalator levels tend to vary from run to run (in contrast with antibody staining).



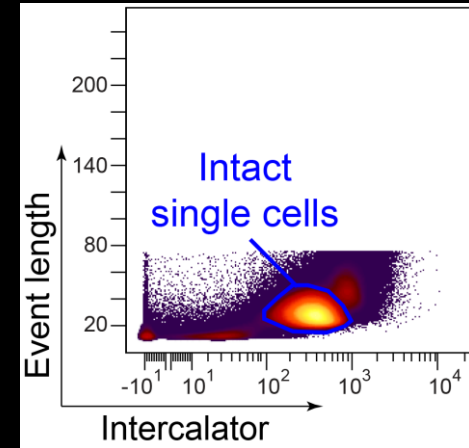
# Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity



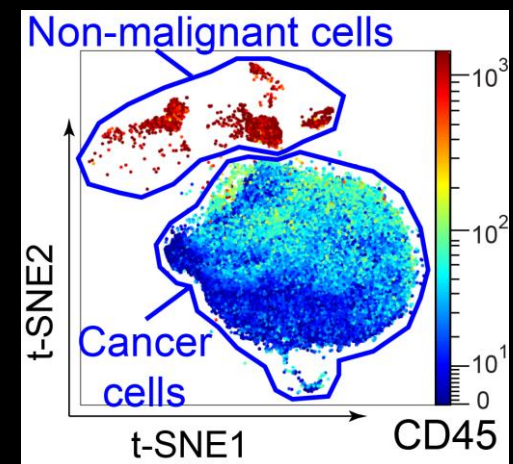
# Single Cell Biology Workflow

Data collection	<ol style="list-style-type: none"> <li>1) Panel design</li> <li>2) Data collection</li> </ol>
Data processing	<ol style="list-style-type: none"> <li>3) Cell event parsing</li> <li>4) Scale transformation</li> </ol>
Distinguishing initial populations	<ol style="list-style-type: none"> <li>5) Live single cell gating</li> <li>6) Focal population gating</li> </ol>
Revealing cell subsets	<ol style="list-style-type: none"> <li>7) Feature selection</li> <li>8) Dimensionality reduction</li> <li>9) Identify cell clusters</li> <li>10) Cluster refinement</li> </ol>
Characterizing cell subsets	<ol style="list-style-type: none"> <li>11) Feature comparison</li> <li>12) Model populations</li> <li>13) Learn cell identity</li> <li>14) Statistical testing</li> </ol>

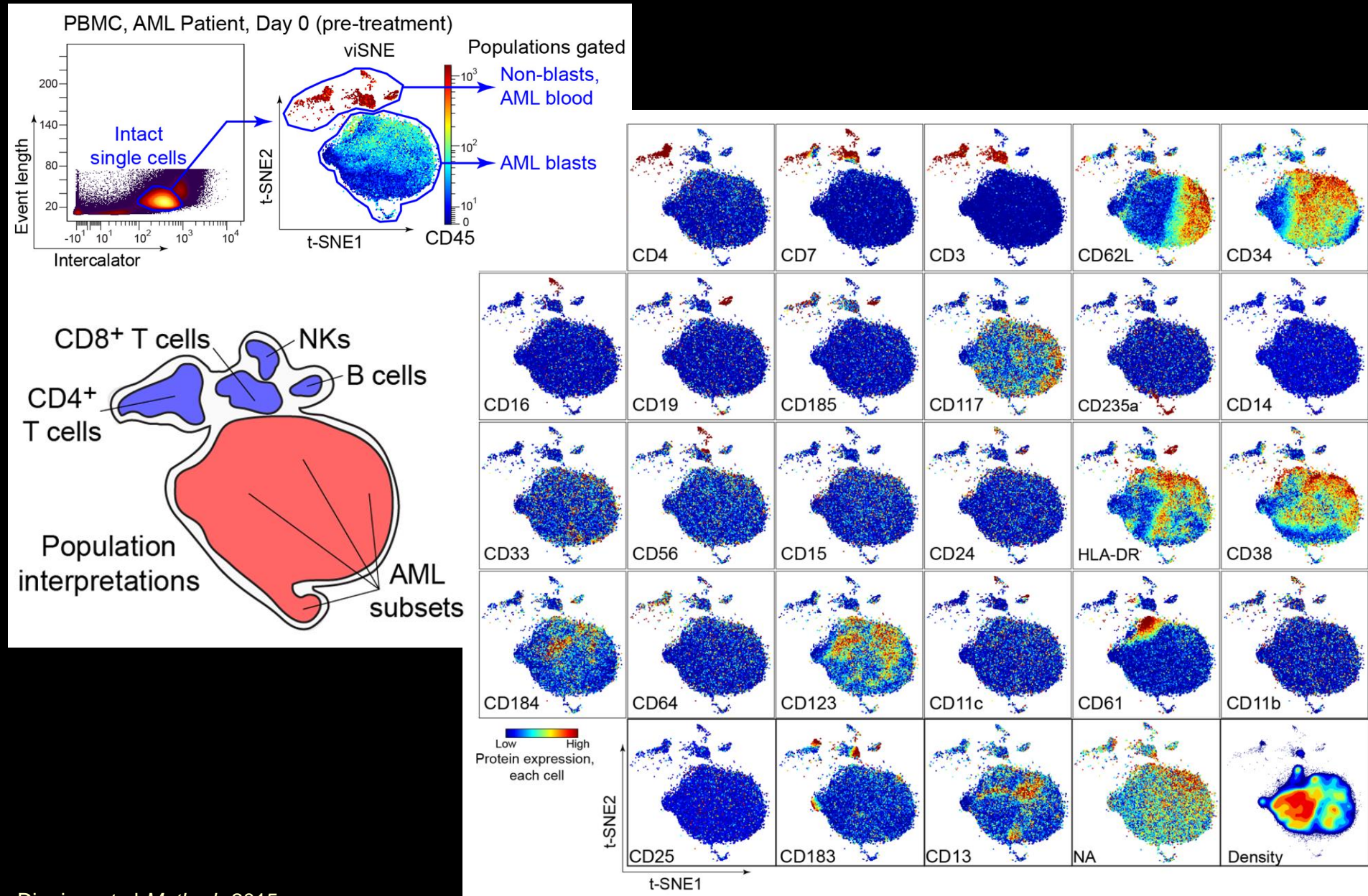
## Expert gating



## viSNE + expert gating



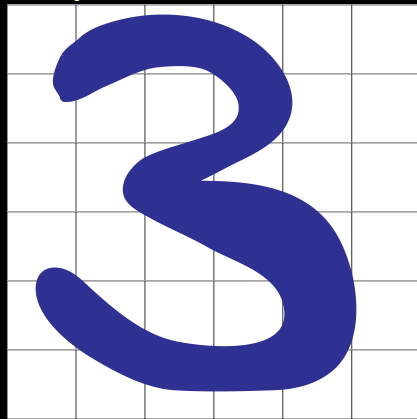
# Current Goal: Get to Figure 1 from Diggins et al. with viSNE



# Stochastic Neighbor Embedding (SNE)

- SNE used for image recognition
- 60,000 handwritten greyscale images
- 28x28 pixels each

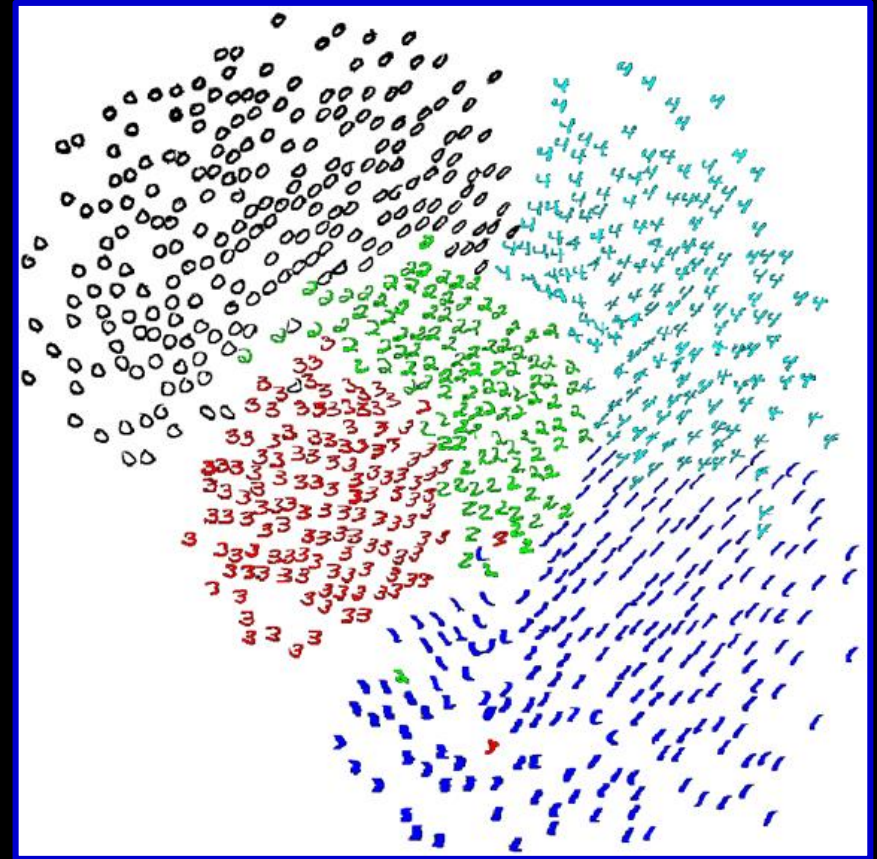
Example: 6x6 Pixel Image



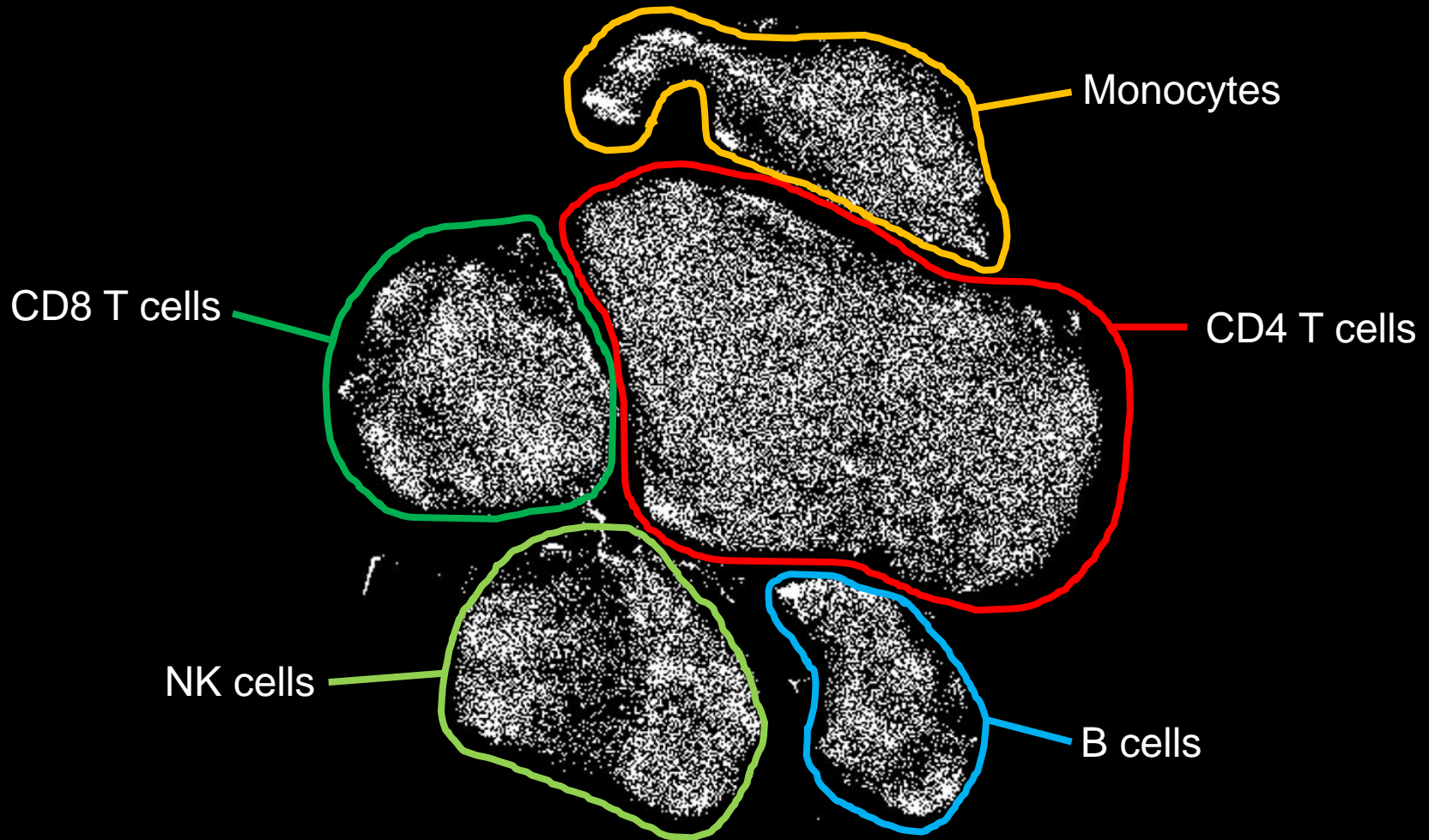
Vectorize (1x36)



tSNE on all pixels



# viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity

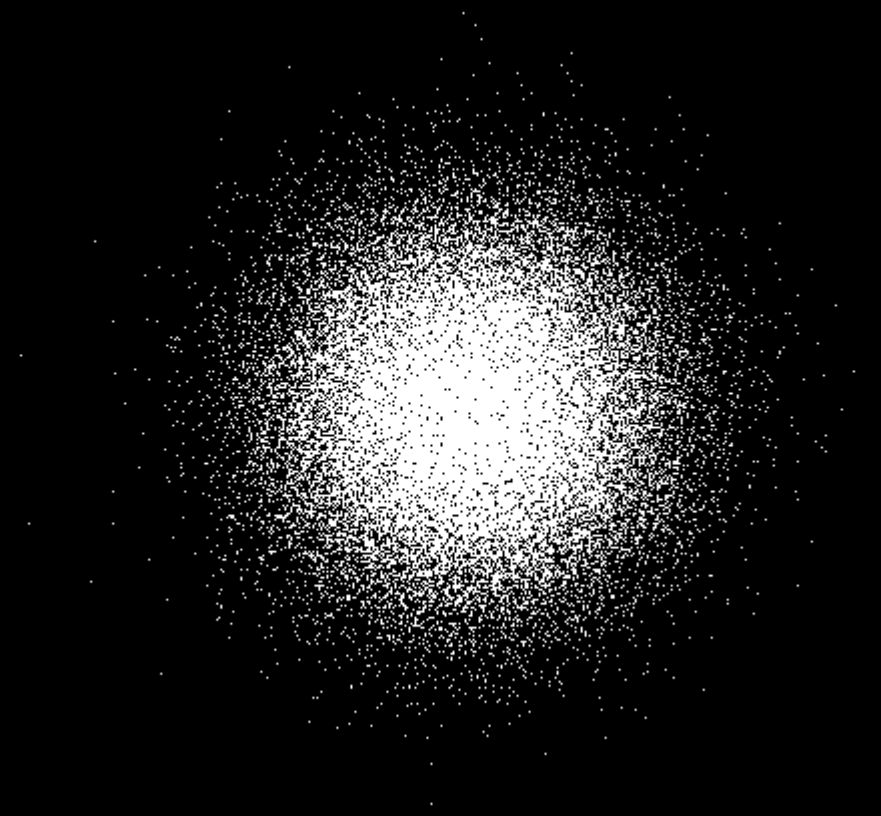


Healthy human blood, mass cytometry,  
26 markers measured, viSNE analysis tool



# viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity

Healthy human blood, mass cytometry, 26D viSNE analysis

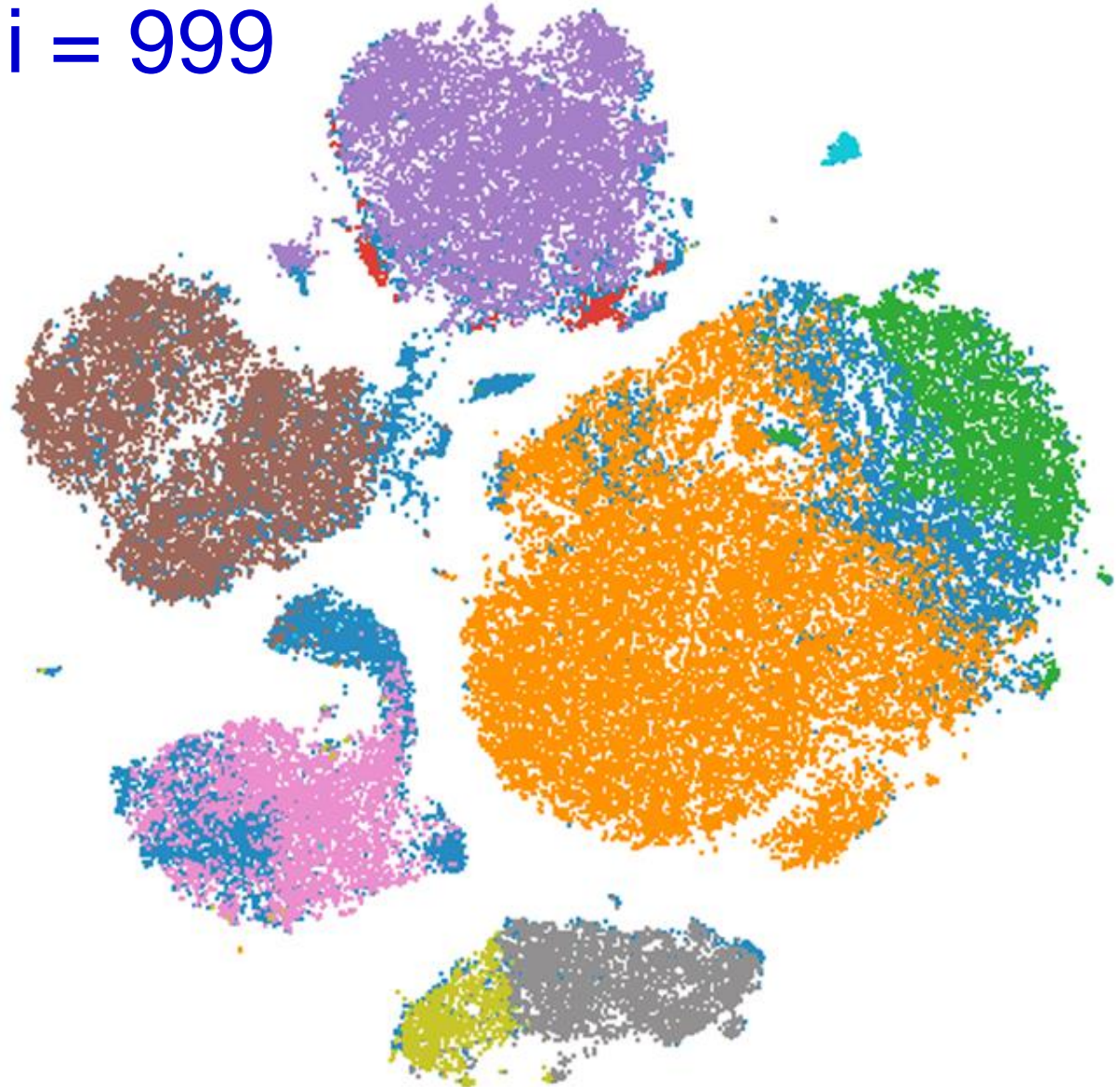


# Viewing Expert Gates with viSNE Reveals Cyto Incognito

$i = 999$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

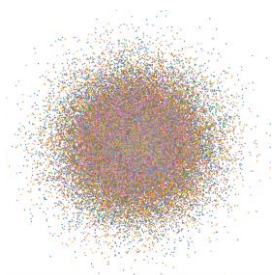


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 0$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

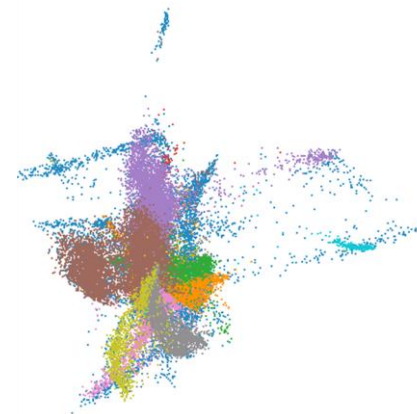


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 45$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

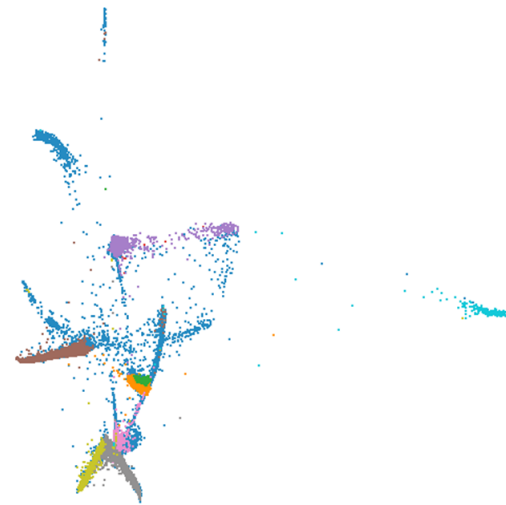


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

i = 88

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

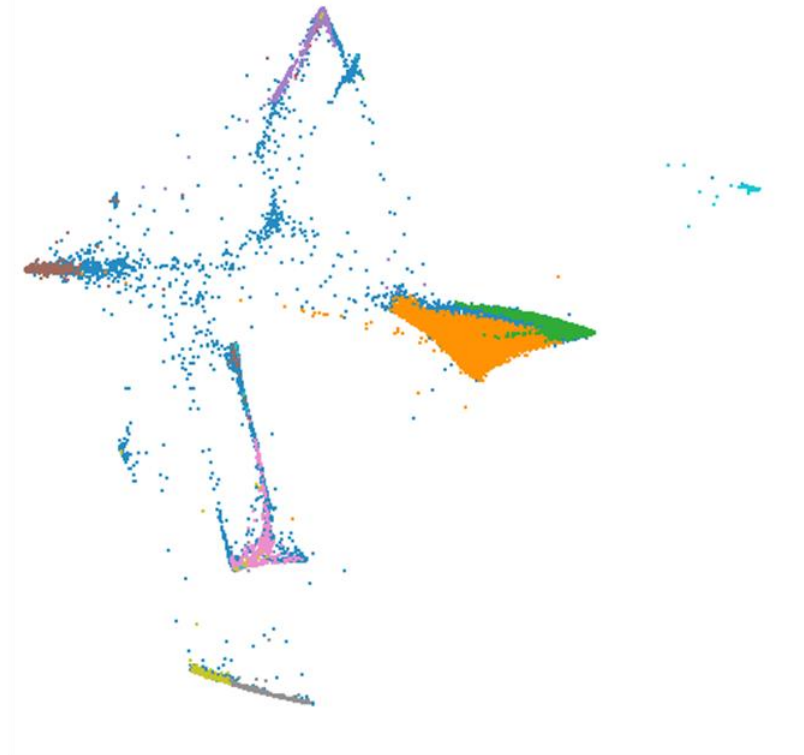


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 198$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

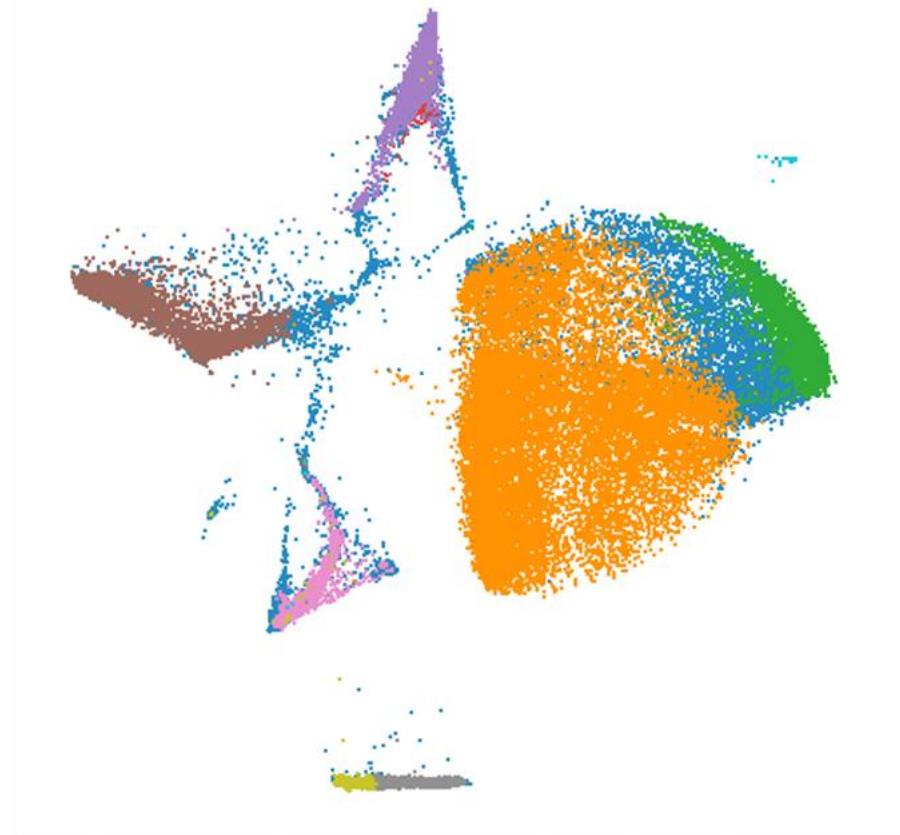


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 262$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

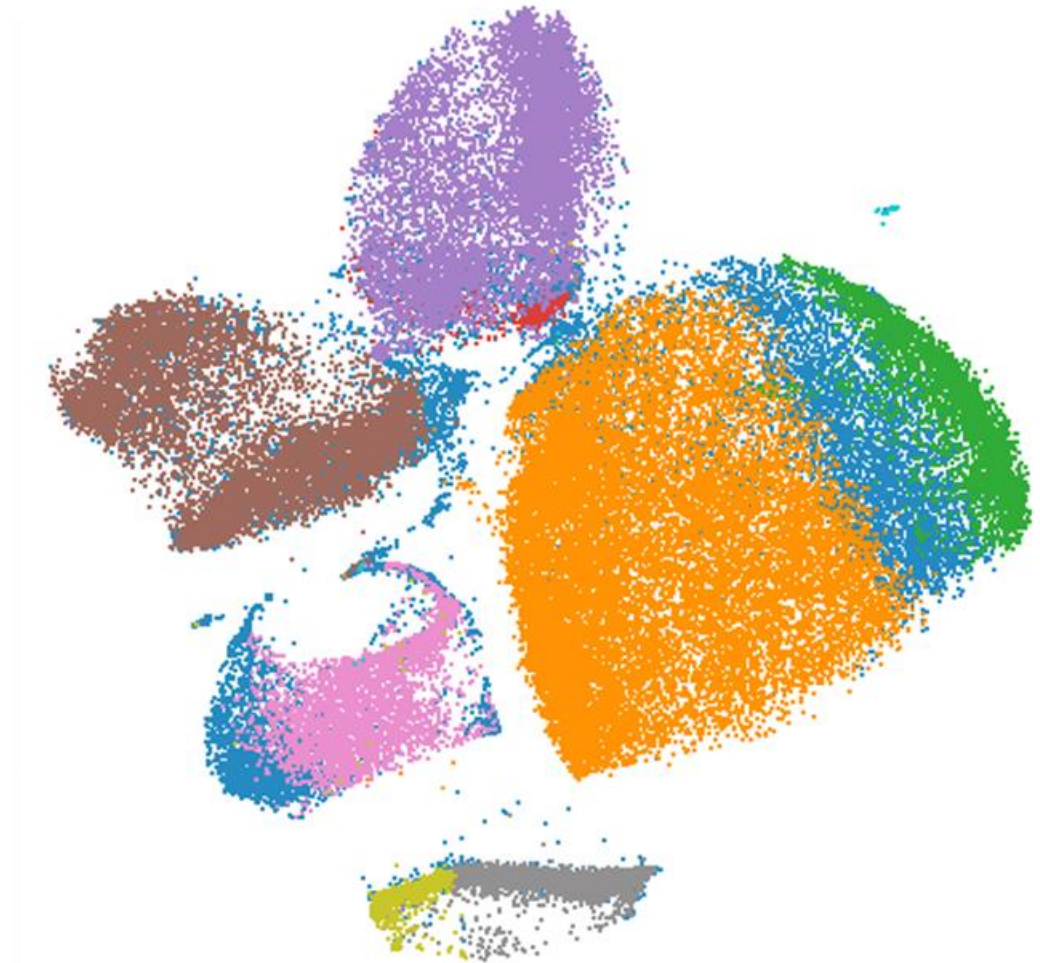


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 289$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs



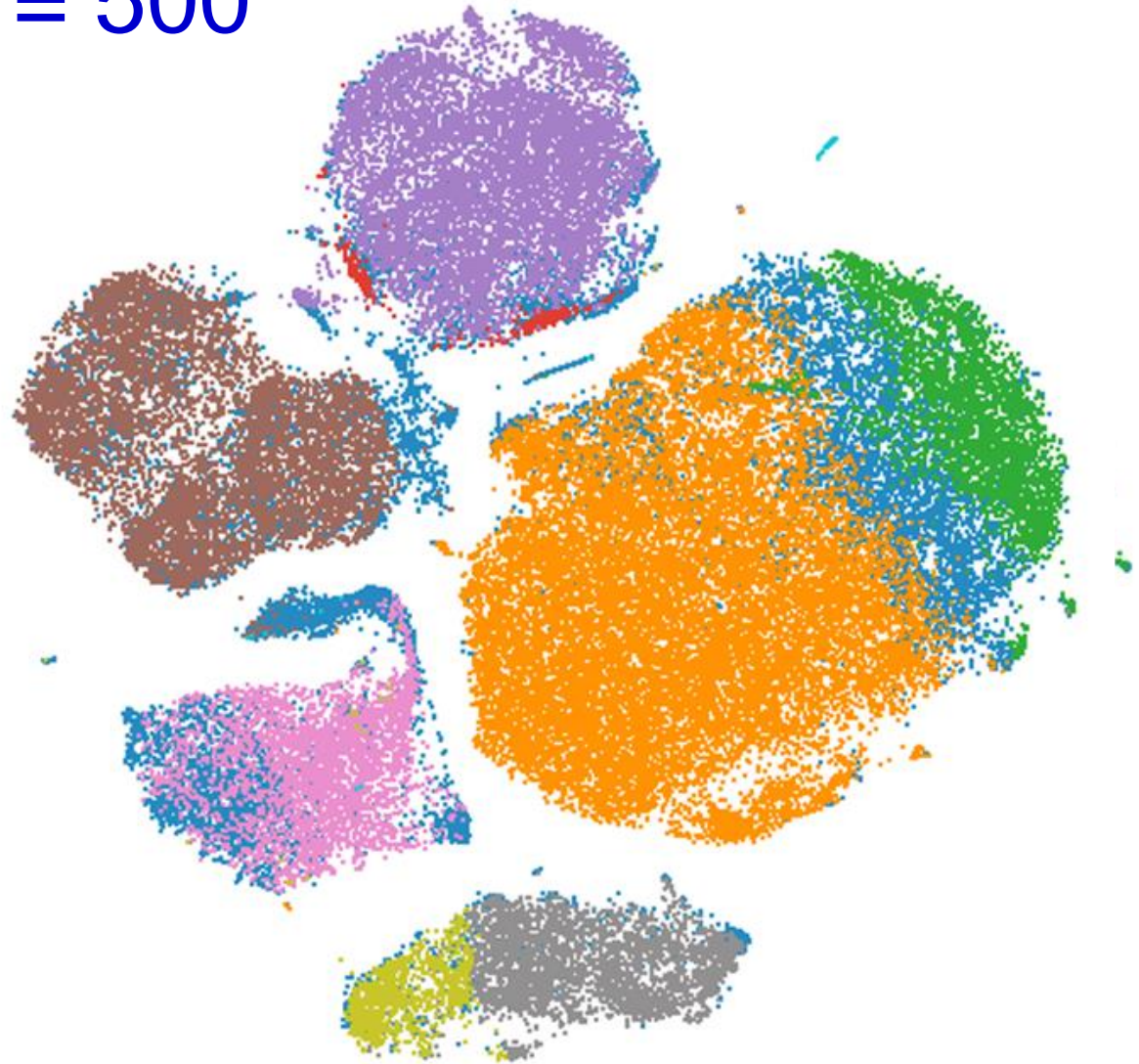


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 500$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

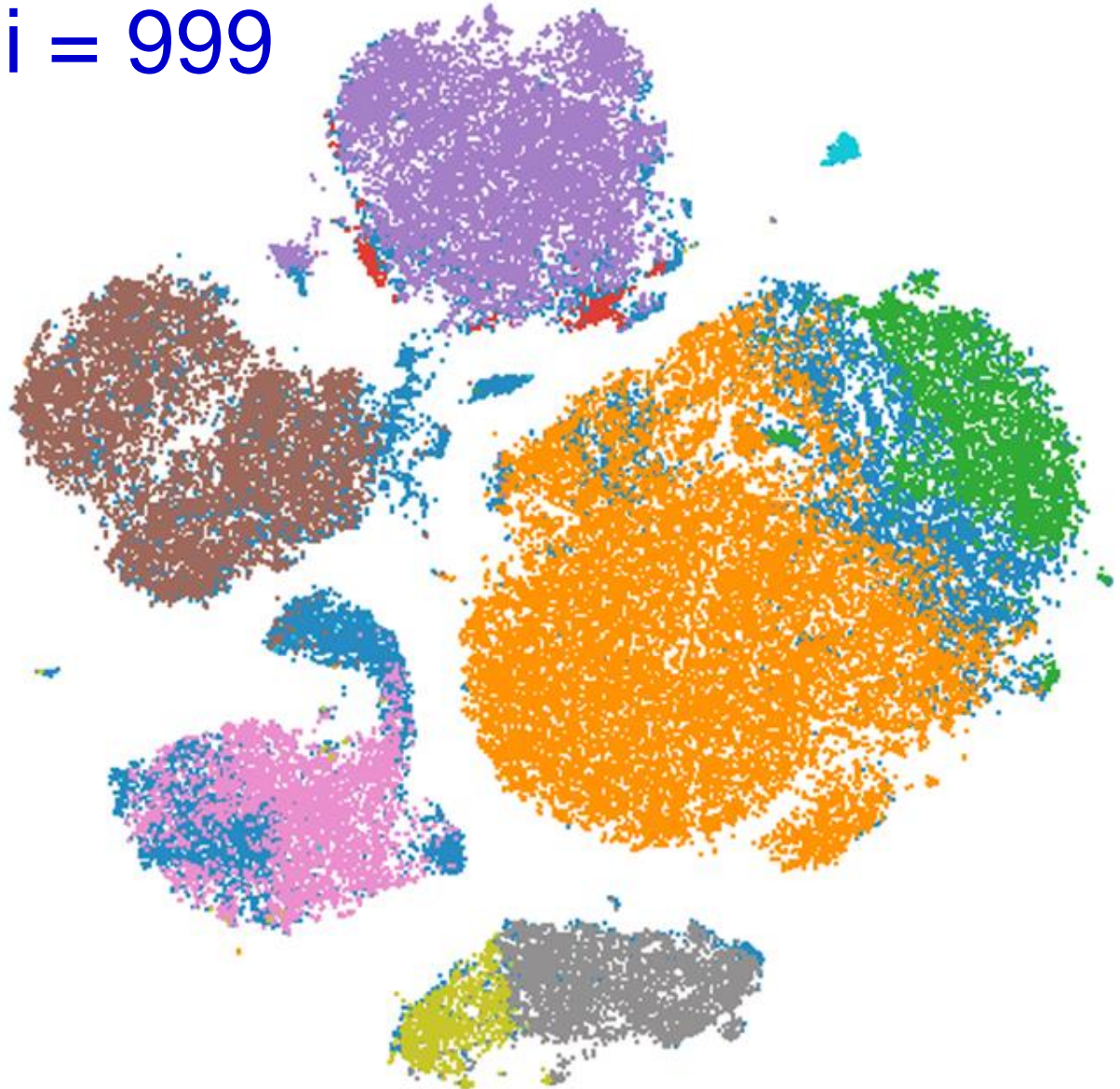


# Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

$i = 999$

Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

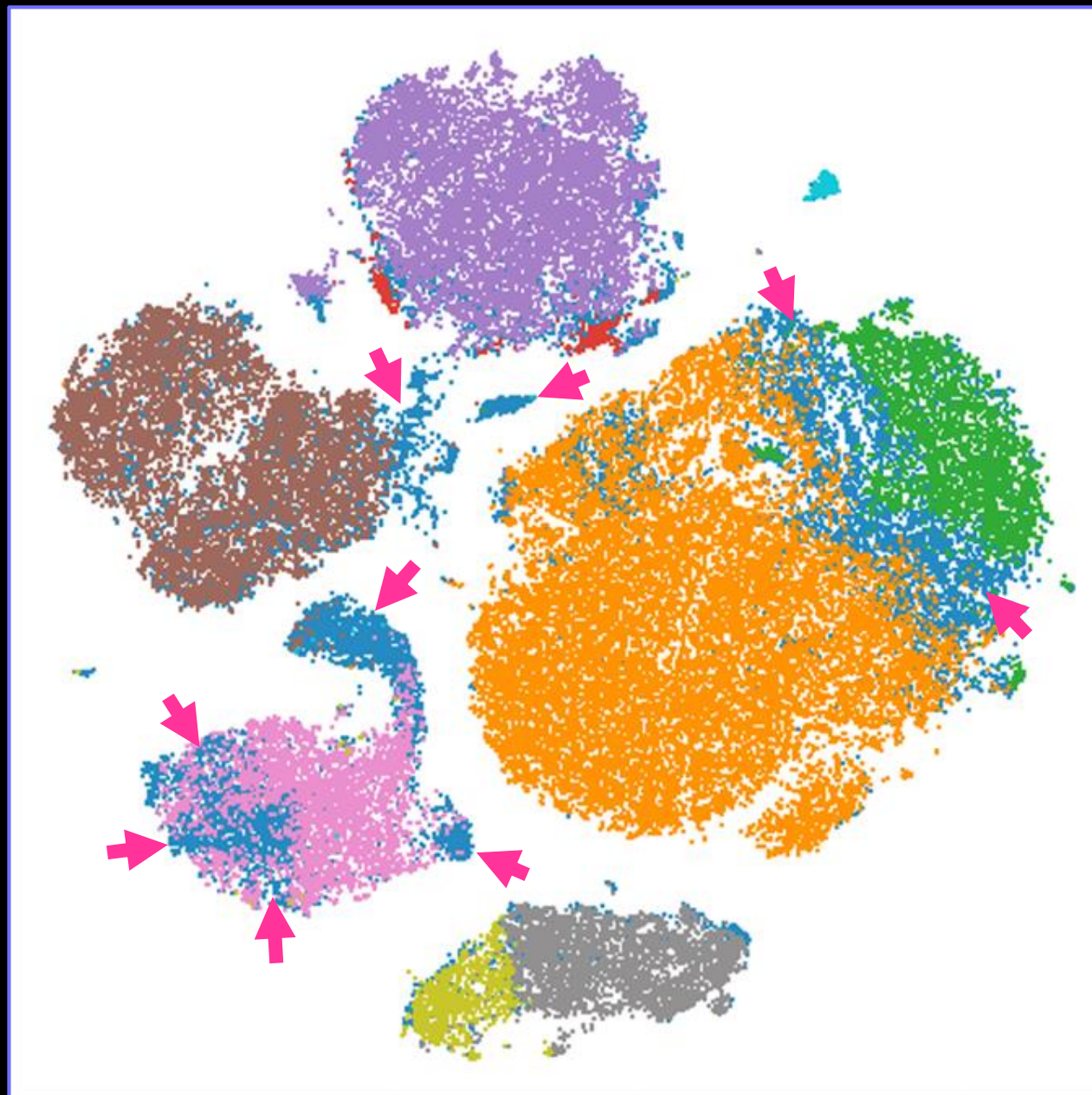


# Viewing Expert Gates with viSNE Reveals Cyto Incognito

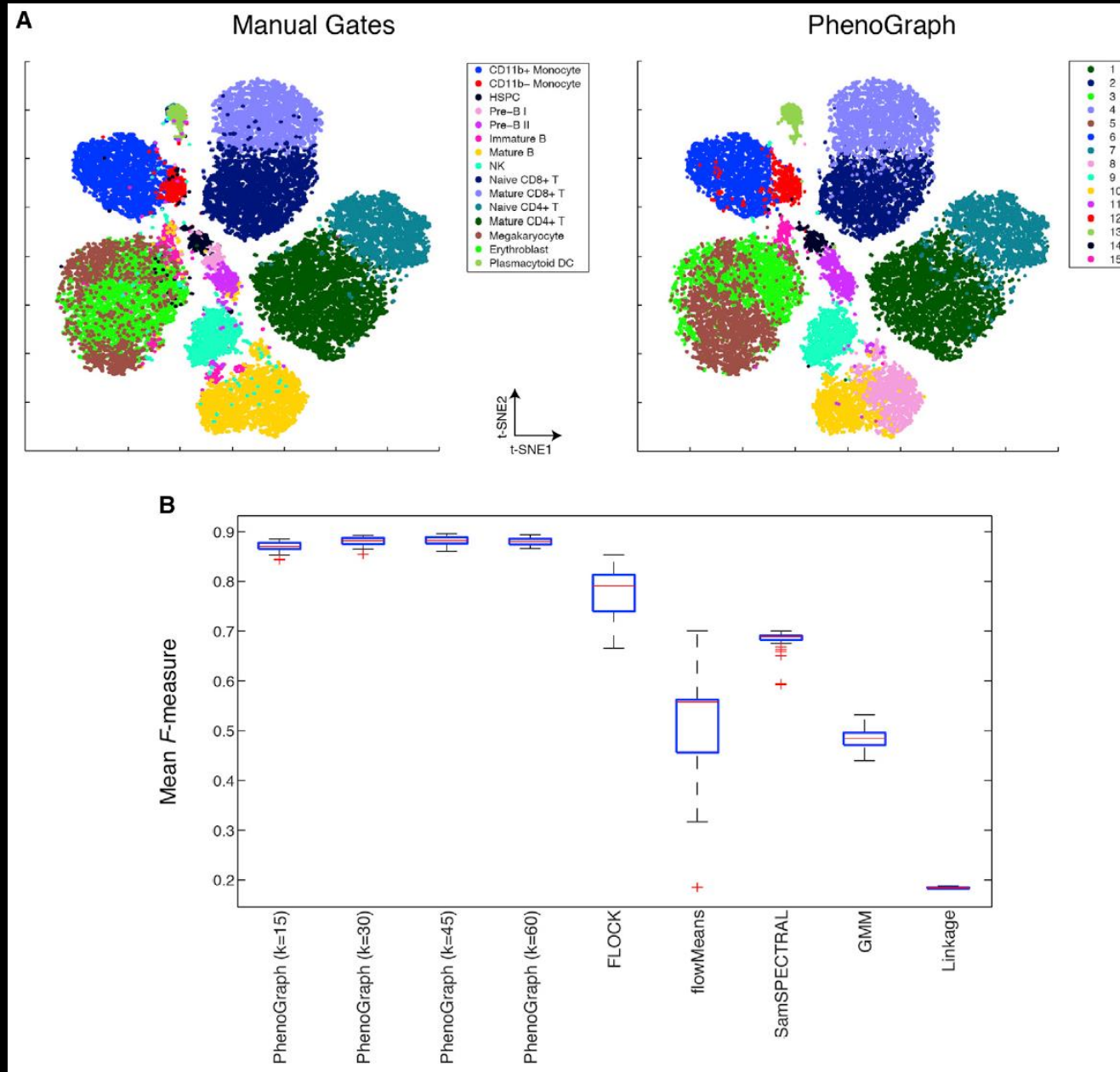
Healthy human blood,  
mass cytometry,  
26D viSNE analysis

- Ungated
- CD45RA+ Naive CD4+
- CD45RA- CD4+ T cells
- CD45RA- Memory CD8+ T cells
- CD45RA+ Naive CD8+ T cells
- CD16+ NK cells
- CD14+ CD33+ Monocytes
- IgM+ B cells
- IgM- B cells
- CD123+ pDCs

→ **Cyto incognito**  
(Cells overlooked or  
hidden in expert gating)



# Phenograph Adds Fast Clustering & Meta-Analysis to viSNE



# Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

4) Run viSNE.

- Pick viSNE => New viSNE Analysis
- Name it “viSNE of AML and healthy marrow”

The screenshot displays the Cytobank Premium web interface. The top navigation bar includes 'Experiments', 'Projects', 'Admin', 'Help', and a user profile for 'jonathan.irish@gmail.com'. The main header shows the current experiment: 'Diggins et al., Methods 2015 - Step 1 (Clone)'. A toolbar contains various analysis tools: 'Actions', 'Illustrations', 'Sample Tags', 'SPADE', 'viSNE', 'Gating', 'Scales', 'Comps', and 'Private'. The 'viSNE' tool is selected, and a dropdown menu is open, showing 'New viSNE Analysis' and 'View All Analyses'. A modal dialog box is overlaid on the screen, titled 'The page at https://premium.cytobank.org says:'. It contains a text input field with the value 'viSNE of AML and healthy bone marrow' and two buttons: 'OK' and 'Cancel'. The 'OK' button is highlighted with a red box.

# Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

## 4) Run viSNE.

- Pick Intact Cells from the Population selection area.
- Select both files to include in the viSNE analysis.

Cytobank Premium  
Experiments Projects  
Diggins et al., Methods 2015 - Step 1 (Clone)  
Actions Illustrations Sample Tags SPADE viSNE  
viSNE Setup Copy Analysis Settings  
Analysis Name: viSNE of AML and healthy bone marrow  
viSNE Inputs:  
Populations: - click the Population name to select FCS File subsets  
\* At least one Population must be chosen to run the analysis.  
Populations ?  
Nothing selected  
Population - FCS File Event Count  
Intact cells  
Ungated  
No selected populations.  
← Click on a Population

Choose Your FCS Files Cancel  
Done  
Select FCS Files for Intact cells. Type filter words  
Press Done when finished.  
Showing all 2 FCS Files  
Select: All None  
Selected 2: Step1a\_AML\_PB0.fcs - Step1a\_AML\_PB0.fcs, Step1b\_normal\_marrow.fcs - Step1b\_normal\_marrow.fcs  
Step1a\_AML\_PB0.fcs - Step1a\_AML\_PB0.fcs  
Step1b\_normal\_marrow.fcs - Step1b\_normal\_marrow.fcs  
Done

# Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

## 4) Run viSNE.

- Choose Equal subsampling and leave the “events per population” at the default of 50,000 (so 100,000 total events).

The screenshot shows the 'Populations' tab in Cytobank. It displays a table with two selected populations. The 'Event Count' column shows 228,457 for the first population and 74,367 for the second. The 'Equal Sampling' column shows 50,000 for both. Below the table, the 'Total Events Selected' is 302,824, and the 'to sample' value is 100,000. The 'Event Sampling' section shows 'Equal' selected, with 'Desired Total Events' set to 100,000 and 'Events Per Population' set to 50,000.

Population - FCS File	Event Count	Equal Sampling
Intact cells		
Step1a_AML_PB0.fcs - Step1a_AML_PB0.fcs	228457	50000
Step1b_normal_marrow.fcs - Step1b_normal_marrow.fcs	74367	50000
Ungated		

Total Events Selected: 302824      100000 to sample

Event Sampling: ?

Proportional  Equal

Desired Total Events:  (must be between 90 and 800000)

Events Per Population:

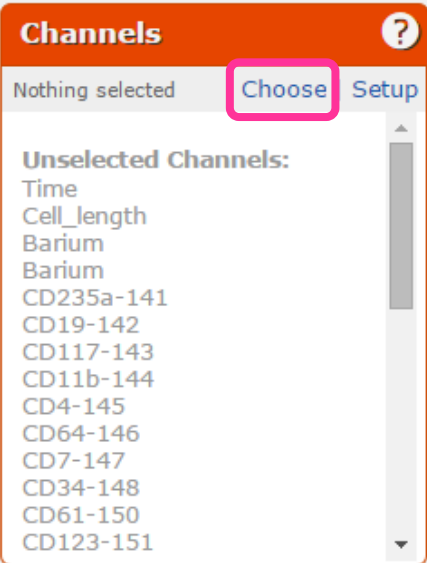
# Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

## 4) Run viSNE.

- Scroll down, and in the channel selector, pick the channels you want to be used in the comparison analysis (to “make the map”).
- Pick “all 27 markers”. Search on “CD” and “select all” (selects 26), then clear the selection and add HLA-DR to the list. Click “Done” and “Run viSNE Analysis”

**Channels:** - click Choose to select channels  
*\* At least one Channel must be chosen to run the analysis.*

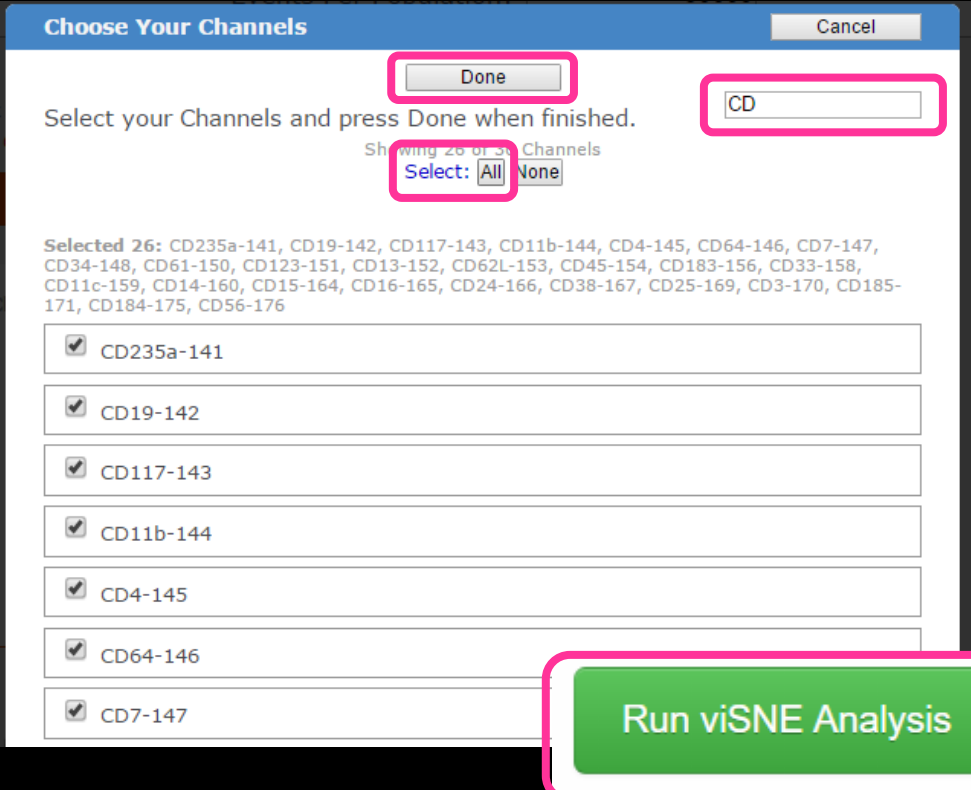


**Channels** ?

Nothing selected **Choose** Setup

**Unselected Channels:**

- Time
- Cell\_length
- Barium
- Barium
- CD235a-141
- CD19-142
- CD117-143
- CD11b-144
- CD4-145
- CD64-146
- CD7-147
- CD34-148
- CD61-150
- CD123-151



**Choose Your Channels** Cancel

**Done**

Select your Channels and press Done when finished.

Showing 26 of 27 Channels

Select: **All** None

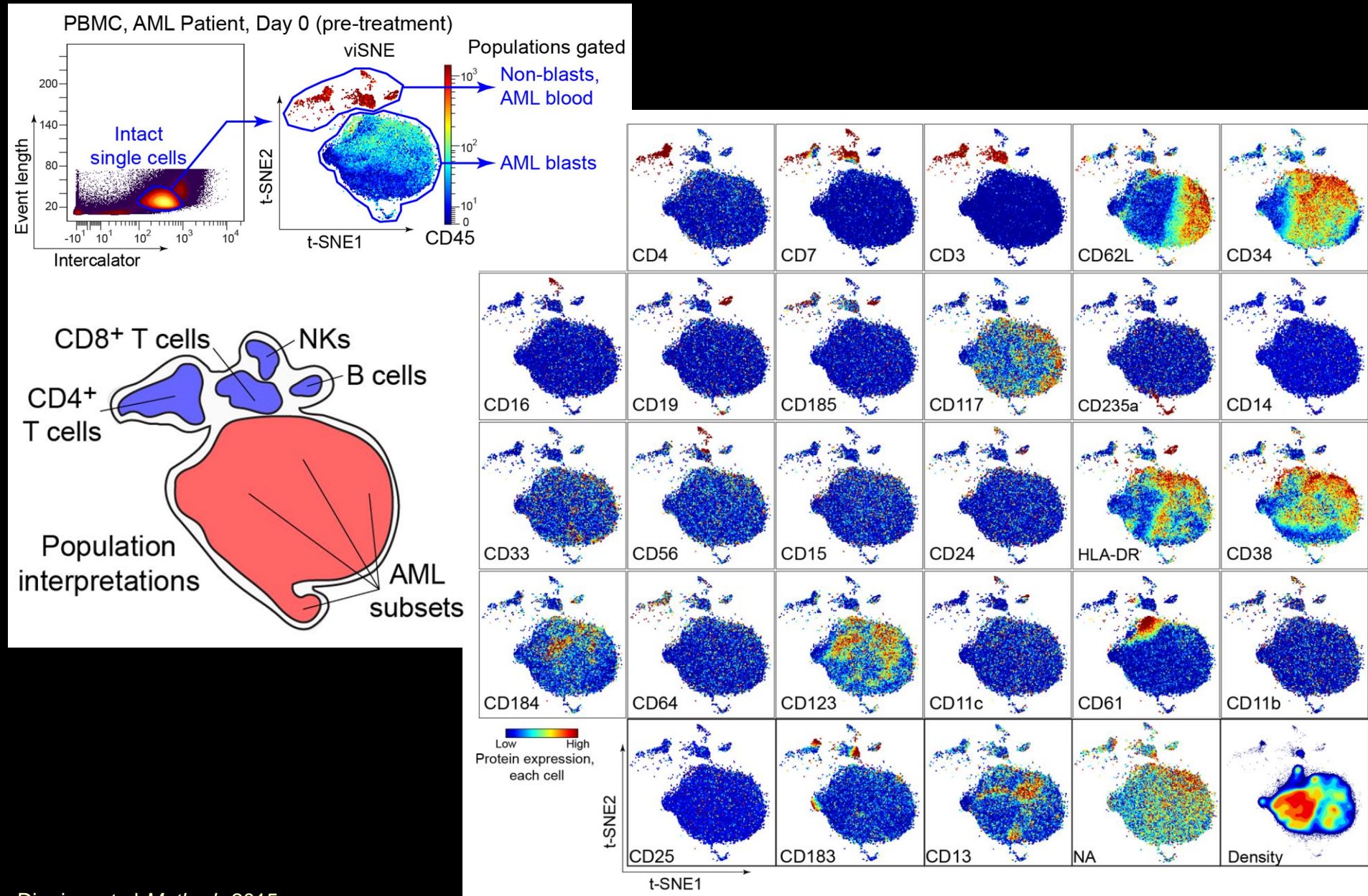
**Selected 26:** CD235a-141, CD19-142, CD117-143, CD11b-144, CD4-145, CD64-146, CD7-147, CD34-148, CD61-150, CD123-151, CD13-152, CD62L-153, CD45-154, CD183-156, CD33-158, CD11c-159, CD14-160, CD15-164, CD16-165, CD24-166, CD38-167, CD25-169, CD3-170, CD185-171, CD184-175, CD56-176

- CD235a-141
- CD19-142
- CD117-143
- CD11b-144
- CD4-145
- CD64-146
- CD7-147

**Run viSNE Analysis**



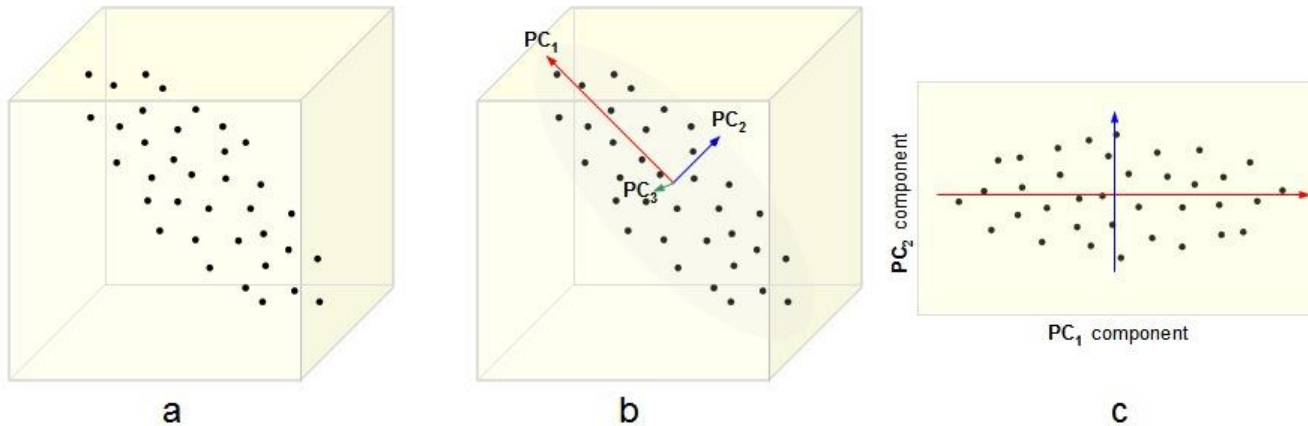
# Next: View All Channels to Make Figure 1 from Diggins et al.



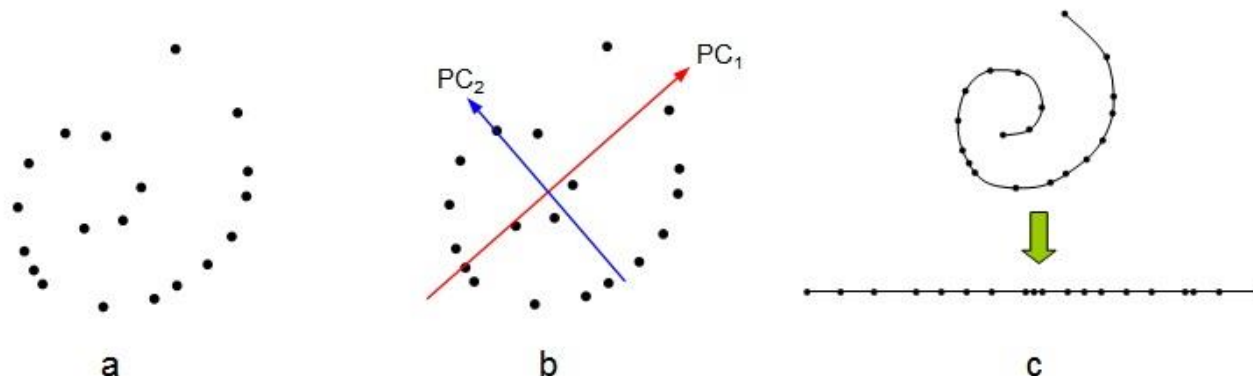
While viSNE runs...

A brief interlude about topology

# Is There 'Inherent Topology' in the Data?



An illustration of PCA. **a)** A data set given as 3-dimensional points. **b)** The three orthogonal Principal Components (PCs) for the data, ordered by variance. **c)** The projection of the data set into the first two PCs, discarding the third one.

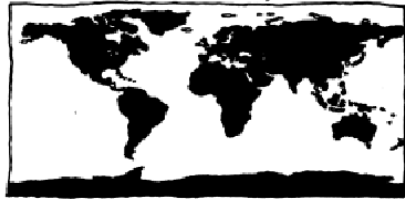


Effects of dimensionality reduction on an inherently non-linear data set. **a)** The original data given as a two-dimensional set. **b)** PCA identifies two PCs as contributing significantly to explain the data variance. **c)** However, the inherent topology (connectivity) of the data helps identify the set as being one-dimensional, but non-linear.

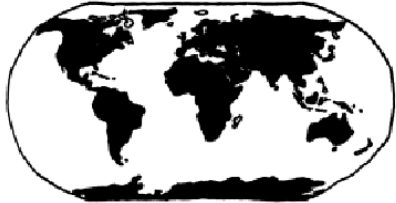
# Mapping Topology Matters When Considering "Travel"



PLATE CARRÉE  
(EQUIRECTANGULAR)



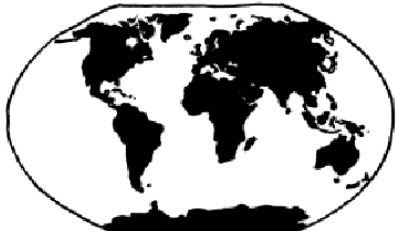
ROBINSON



WATERMAN BUTTERFLY



WINKEL-TRIPPEL



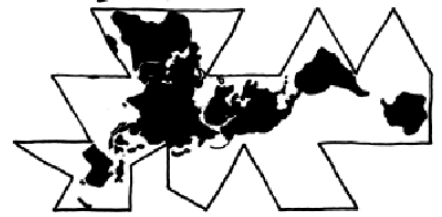
GOODE HOMOLOXINE



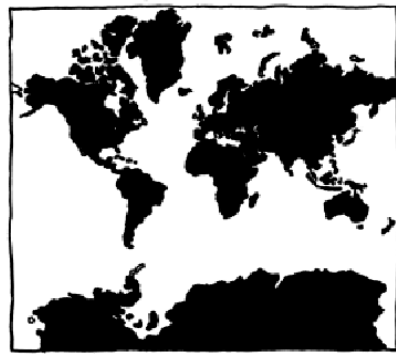
HOOB-DYER



DYMAXION



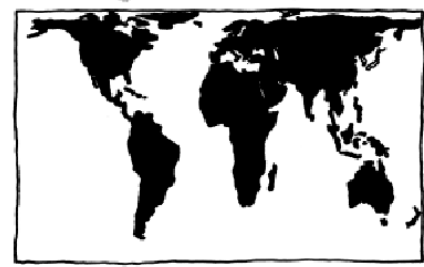
MERCATOR



VAN DER GRINTEN



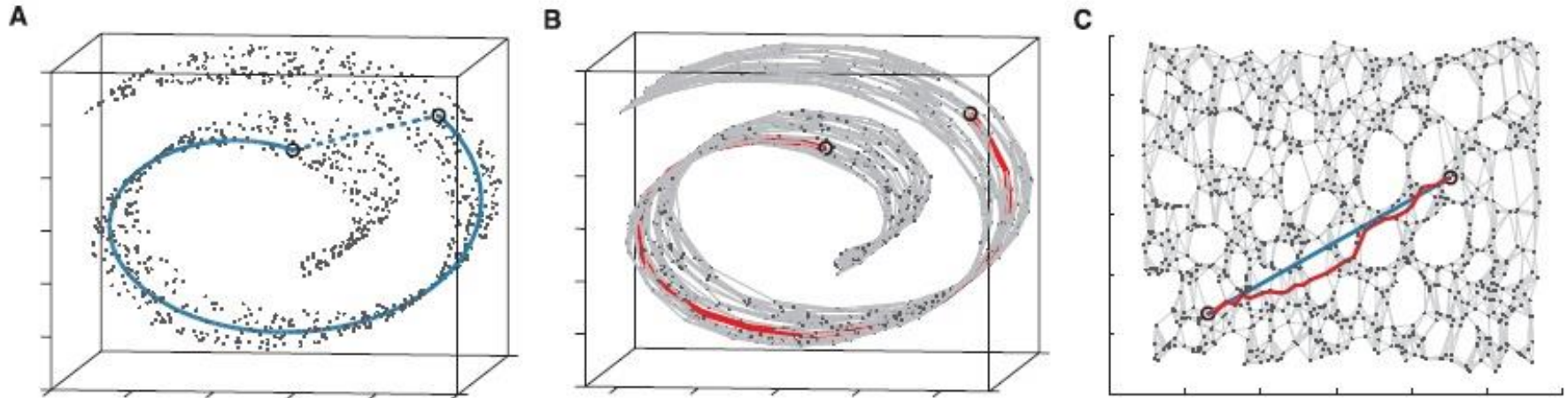
GALL-PETERS



PEIRCE QUINCUNCIAL

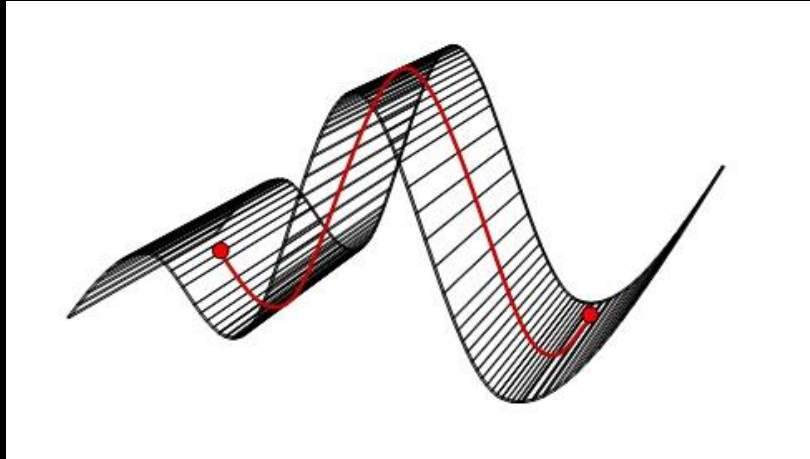


# Multidimensional Scaling Can “Flatten” or “Unroll” Multidimensional Progressions / Topologies for Linear Analysis



Isomap at work on the "swiss roll" data set. **a)** The input data are given as three-dimensional, but **are really two-dimensional in nature**. **b)** Each point in the data set is connected to its neighbors to form a neighborhood graph, overlaid in light grey. Geodesic distances are approximated by computing **shortest paths along the neighborhood graph** (red). **c)** MDS applied to the geodesic distances has the effect of "unrolling" the swiss roll into its natural, two-dimensional parameterization. The neighborhood graph has been overlaid for comparison. Now, the Euclidean distances (in blue) approximate the original geodesic distances (in red).

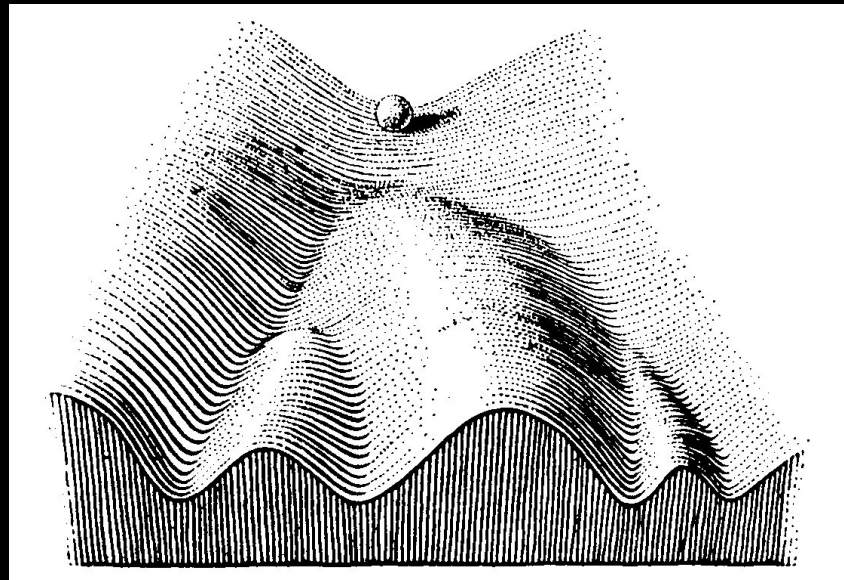
# Geodesic Distance Measurements May Better Capture Developmental Progressions



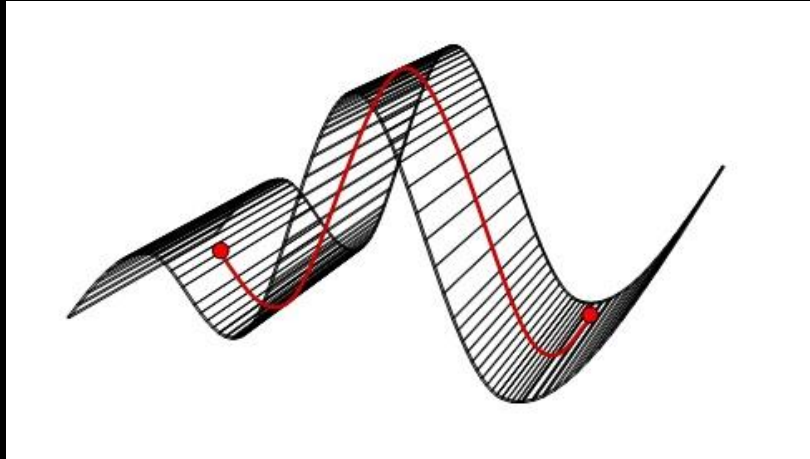
**Geodesic distance.** The geodesic distance between the two red points is the length of the geodesic path, which is the shortest path between the points, that lies on the surface.

Lydia E. Kavradi, *Geometric Methods in Structural Computational Biology*

“**Creode** is a neologism coined by the biologist C.H. Waddington to represent the developmental pathway followed by a cell as it grows to form part of a specialized organ. Combining the Greek roots for "necessary" and "path," the term was inspired by the property of regulation.”

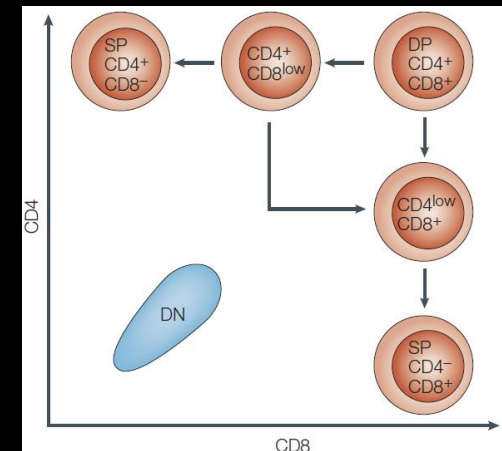
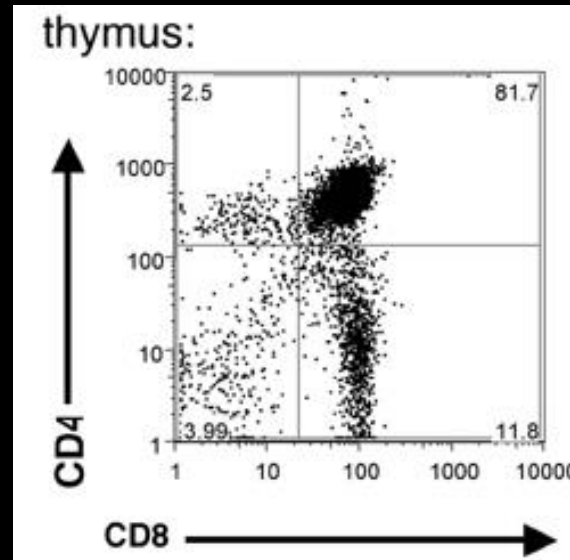
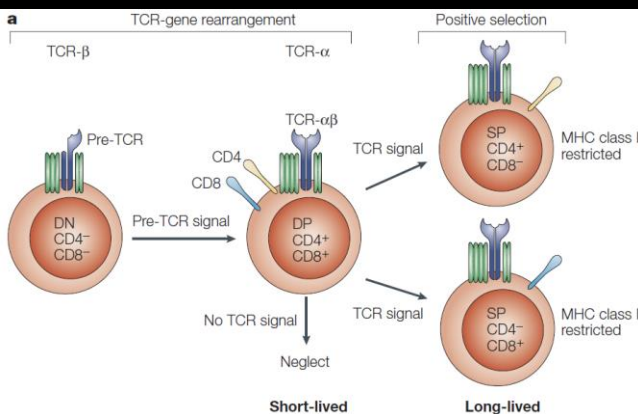


# Geodesic Distance Measurements May Better Capture Developmental Progressions



**Geodesic distance.** The geodesic distance between the two red points is the length of the geodesic path, which is the shortest path between the points, that lies on the surface.

Lydia E. Kavraki, *Geometric Methods in Structural Computational Biology*



Bosslet, *Nature Rev. Immuno.* 2004

OK, back to work...

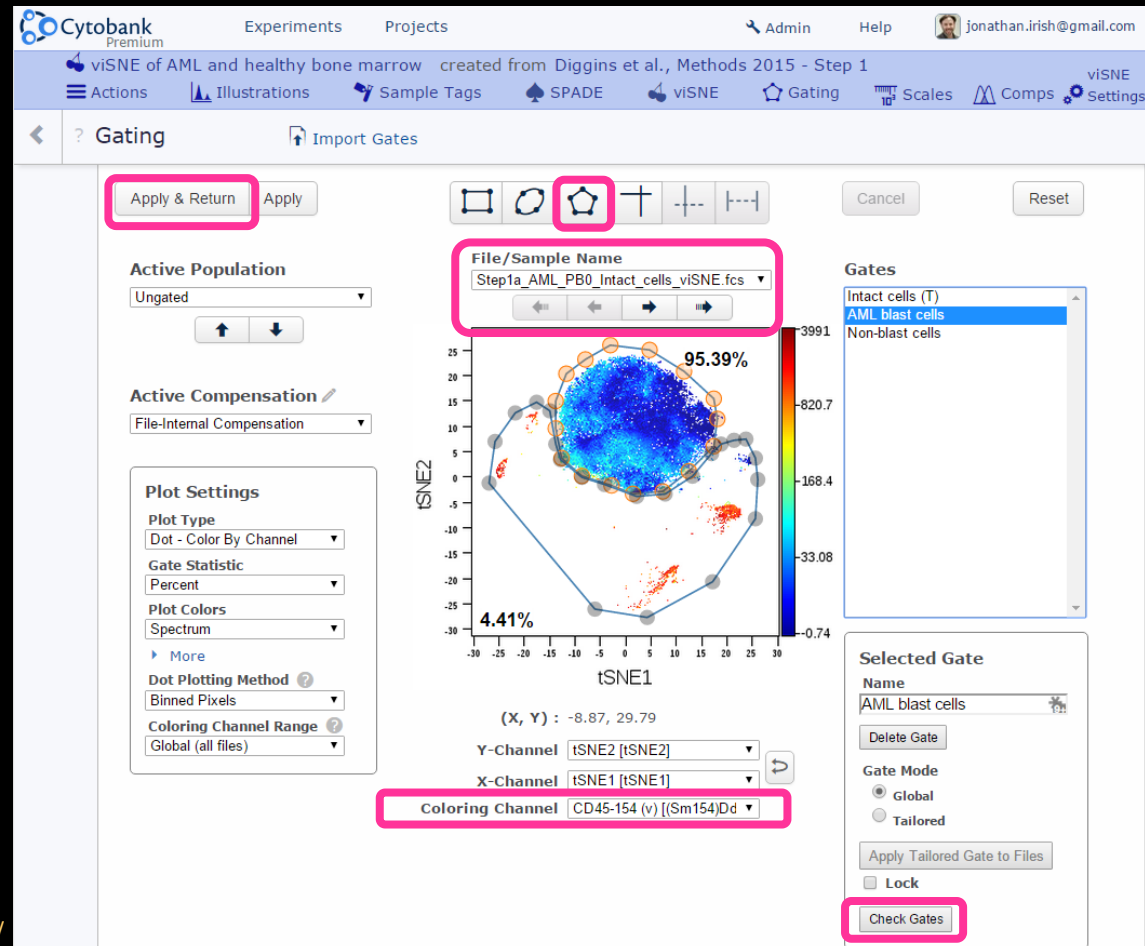


# Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

- 5) Gate on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)
- 6) Export gated populations as 4 new FCS files & discard the empty one.
- 7) Concatenate the 3 remaining files into a merged file.

- Change the z-axis coloring channel to CD45.
- Using the knowledge the AML blasts are CD45<sup>lo</sup>, draw gates for “AML blast cells” and “non-blast cells”.
- Use “Check Gates” button to make sure gates work well on both files. Healthy marrow should have <3% cells in the AML blast gate.
- Apply & Return.



# Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

- 5) Gate on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)
- 6) Export gated populations as 4 new FCS files & discard the empty one.
- 7) Concatenate the 3 remaining files into a merged file.

- Set up an illustration to export the viSNE gated data: Pick 2 Populations (“AML blast cells” & “Non-blast cells”). Pick one Channel (CD45). Activate “FCS files”. Update & Save.

The screenshot shows the Cytobank Premium interface for an experiment titled "viSNE of AML and healthy bone marrow". The "Fcs Files" tab is selected, showing a list of files. A "Choose Your Populations" dialog box is open, showing a list of populations with "AML blast cells" and "Non-blast cells" selected. The "Save" button is highlighted in the top navigation bar.

**Populations:**

Selected	Unselected
AML blast cells	Ungated
Non-blast cells	Intact cells

**Fcs Files:**

Selected	Unselected
Step1a_AML_PB0_Intact_cells_viSNE.fcs	Step1b_normal_marrow_Intact_cells_viSNE.fcs
Step1a_AML_PB0_Intact_cells_viSNE.fcs	Step1b_normal_marrow_Intact_cells_viSNE.fcs

**Channels:**

Selected	Unselected
CD45-154 (v) - Panel 1	Time - Panel 1
	Cell_length - Panel 1
	CD123-151 (v) - Panel 1
	CD62L-153 (v) - Panel 1
	NA191 - Panel 1
	NA193 - Panel 1
	CD19-142 (v) - Panel 1
	CD11b-144 (v) - Panel 1
	CD4-145 (v) - Panel 1

**Choose Your Populations Dialog:**

Select your Populations and press Done when finished.

Sort: [Up] [Down] [Reset order]

Showing: All 4 Populations  
Selected 2: AML blast cells, Non-blast cells

<input type="checkbox"/>	Ungated
<input type="checkbox"/>	Intact cells
<input checked="" type="checkbox"/>	AML blast cells
<input checked="" type="checkbox"/>	Non-blast cells

# Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

- 5) Gate on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)
- 6) Export gated populations as 4 new FCS files & discard the empty one.
- 7) Concatenate the 3 remaining files into a merged file.

- Export viSNE gated data: Actions => Cloning => Split Files by Population

The screenshot shows the Cytobank Premium interface for an experiment titled "viSNE of AML and healthy bone marrow". The "Actions" menu is open, and the "Cloning" option is selected, which has opened a sub-menu where "Split Files by Population" is highlighted. The interface also shows a table of "Fcs Files" with two selected files and a "Channels" panel with one selected channel (CD45-154) and a list of unselected channels.

Fcs Files
Step1a_AML_PB0_Intact_t_cells_viSNE.fcs - Step1a_AML_PB0_Intact_t_cells_viSNE.fcs
Step1b_normal_marrow_Intact_cells_viSNE.fcs - Step1b_normal_marrow_Intact_cells_viSNE.fcs

Channels
CD45-154 (v) - Panel 1
<b>Unselected Channels:</b>
Time - Panel 1
Cell_length - Panel 1
CD123-151 (v) - Panel 1
CD62L-153 (v) - Panel 1
NA191 - Panel 1
NA193 - Panel 1
CD19-142 (v) - Panel 1
CD11b-144 (v) - Panel 1
CD4-145 (v) - Panel 1

# Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

- 5) Gate on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)
  - 6) Export gated populations as 4 new FCS files & discard the empty one.
  - 7) Concatenate the 3 remaining files into a merged file.
- Name the experiment “Diggins et al. - Export for SPADE” and click “Create Experiment”

Cytobank Premium

Experiments Projects

Created from viSNE of AML and healthy bone marrow created from Diggins et al., Methods 2015 - Step 1

Actions Illustrations Sample Tags SPADE viSNE Gating

### Split Files by Population

Split Files by Population will create a new experiment where each FCS file is comprised of only the events in each population in the Working Illustration.

[Learn More](#)

\* Experiment Name

\* Purpose

Comments

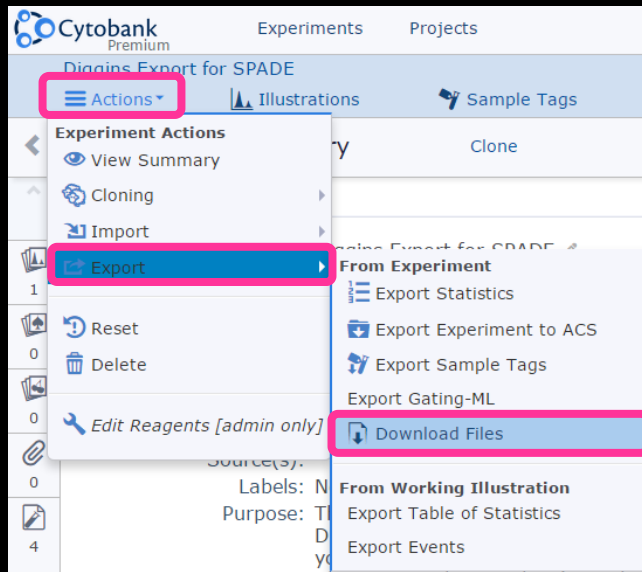
# Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

- 5) Gate on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)
- 6) Export gated populations as 4 new FCS files & discard the empty one.
- 7) Concatenate the 3 remaining files into a merged file.

- Download the files and concatenate (merge) into 1 file.

<http://blog.cytobank.org/2013/10/29/new-fcs-file-concatenation-tool/>



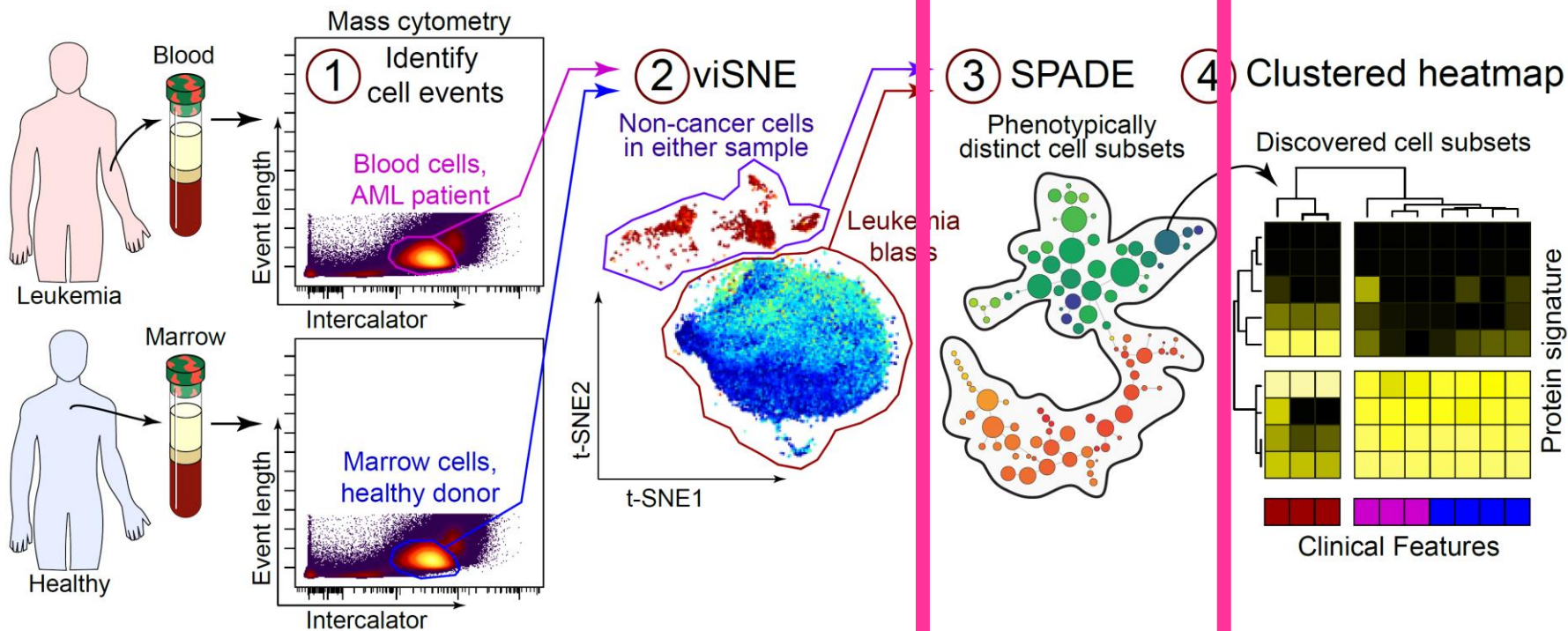
The screenshot shows the FCS File Concatenation Tool interface. The tool is divided into four main sections:

- 1. Choose FCS files to concatenate:** This section contains an 'Add' button, a 'Remove' button, and a 'Remove All' button. An arrow points to the 'Add' button with the annotation '1. Add files to concatenate'.
- 2. Edit the output file channel names:** This section contains a table with columns for '#', 'Short Channel Name', and 'Long Channel Name'. An arrow points to the table with the annotation '2. Edit channel names if necessary'.
- 3. Customize the output file:** This section contains checkboxes for 'Add file number channel' and 'Order by time (sBTIM)'. It also has input fields for 'Plate ID', 'Plate Name', 'Well ID', and 'Tube Name'. An arrow points to the 'Add file number channel' checkbox with the annotation '3. Select a save location and filename'.
- 4. Choose output location:** This section contains input fields for 'Output Path' (G:\Chrome downloads) and 'File Name' (concat.fcs), along with 'Browse' and 'Merge Files' buttons. An arrow points to the 'Merge Files' button with the annotation '4. Click to merge files'.

At the bottom of the interface, there are links for 'Terms and Conditions', 'How do I use this tool?' (http://support.cytobank.org/cytobank-utilities), and '© 2013, Cytobank Inc. - Send feedback to support@cytobank.org'. There is also an 'Output Console' button.

<https://premium.cytobank.org/cytobank/experiments/44415/>

# Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

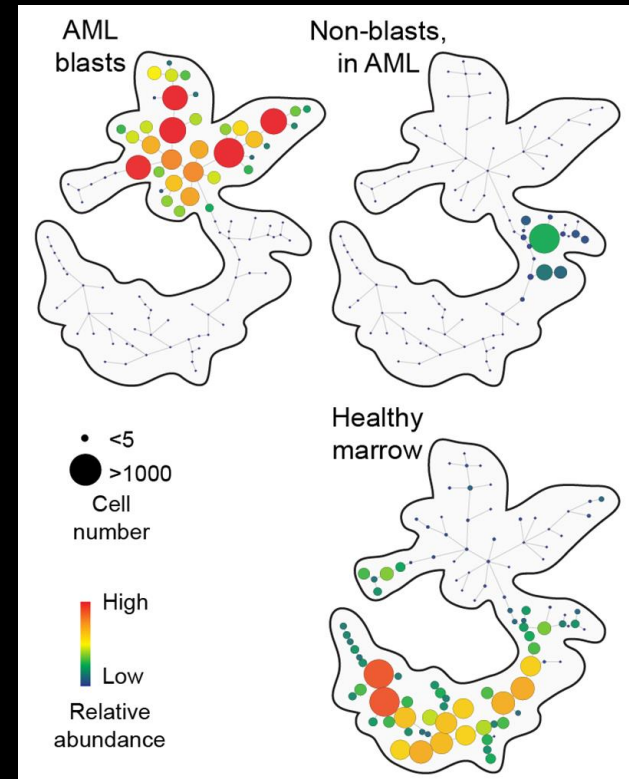


# Single Cell Biology Data Analysis

Data collection	<ol style="list-style-type: none"> <li>1) Panel design</li> <li>2) Data collection</li> </ol>
Data processing	<ol style="list-style-type: none"> <li>3) Cell event parsing</li> <li>4) Scale transformation</li> </ol>
Distinguishing initial populations	<ol style="list-style-type: none"> <li>5) Live single cell gating</li> <li>6) Focal population gating</li> </ol>
Revealing cell subsets	<ol style="list-style-type: none"> <li>7) Feature selection</li> <li>8) Dimensionality reduction</li> <li>9) Identify cell clusters</li> <li>10) Cluster refinement</li> </ol>
Characterizing cell subsets	<ol style="list-style-type: none"> <li>11) Feature comparison</li> <li>12) Model populations</li> <li>13) Learn cell identity</li> <li>14) Statistical testing</li> </ol>

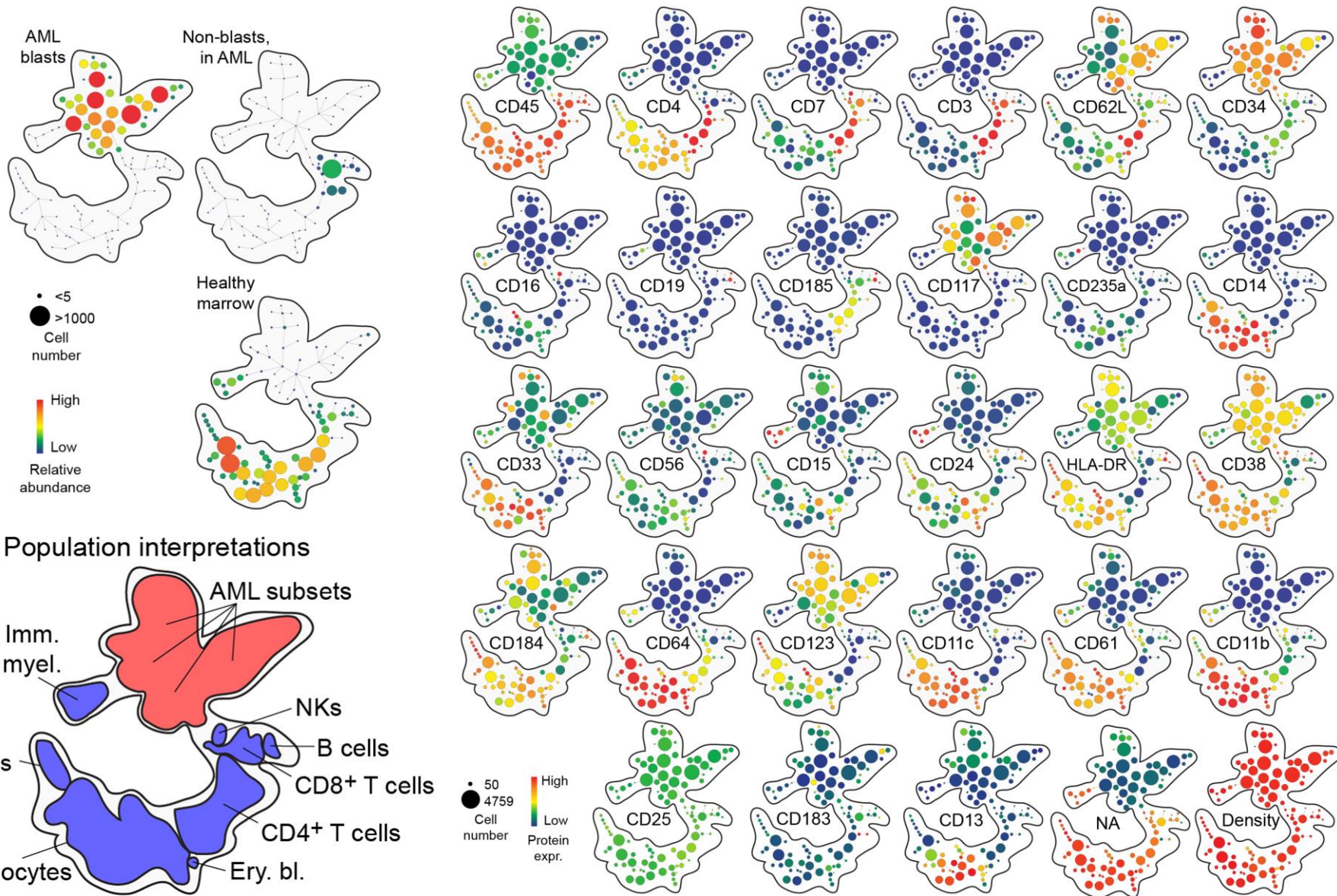
Feature selection

SPADE: cell subset clustering



Grouped populations of cells

# Current Goal: Make Figure 2 from Diggins et al. with SPADE





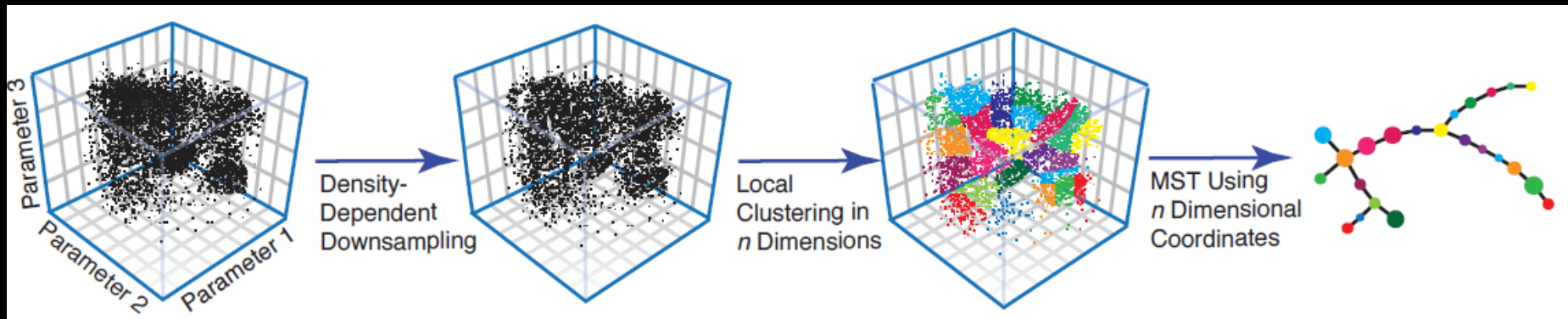
# SPADE Extracts Population Hierarchies from Multi-D Space

Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE

Peng Qiu<sup>1,2</sup>, Erin F Simonds<sup>3</sup>, Sean C Bendall<sup>3</sup>, Kenneth D Gibbs Jr<sup>3</sup>, Robert V Bruggner<sup>3</sup>, Michael D Linderman<sup>4</sup>, Karen Sachs<sup>3</sup>, Garry P Nolan<sup>3</sup> & Sylvia K Plevritis<sup>1</sup>

VOLUME 29 NUMBER 10 OCTOBER 2011 NATURE BIOTECHNOLOGY

SPADE method:



1) Get Data

2) Downsample  
(preserves rare subsets)

3) Cluster  
(group by similarity)

4) Project in 2D  
(minimum spanning tree)

SPADE stands for 'Spanning-tree Progression Analysis of Density-normalized Events'. A graphical interface and updates for fluorescence datasets are implemented on Cytobank.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)

---



SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)

---



SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)

---



SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)

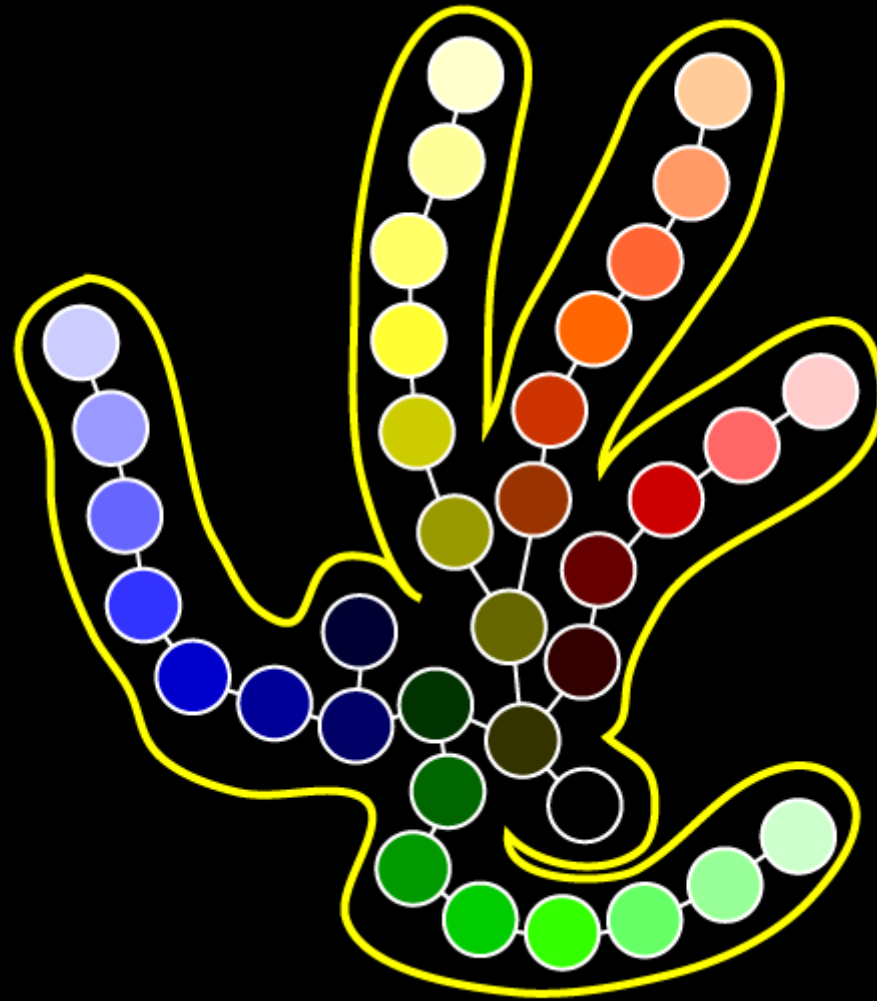
---



SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)

---



SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# SPADE Trees Depict Multidimensional Similarity (not necessarily developmental relationships)



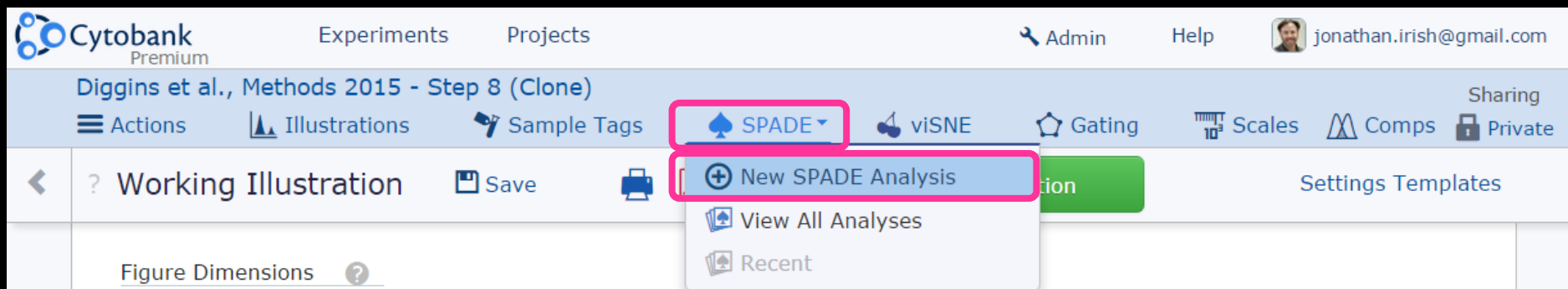
SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

# Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 8”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Run SPADE.

- Navigate to Diggins et al., Methods 2015 – Step 8.
- Click the experiment name to load it.
- Clone the experiment, as before. Actions => Cloning => Selective Clone.
- Click Scales to edit the scales and set the argument to 15 for all channels. Click Apply and then close this window.
- The cells have already been gated, so no gating is needed.
- Choose SPADE => New SPADE Analysis & give it a name.



The screenshot shows the Cytobank Premium web interface. The top navigation bar includes 'Experiments' and 'Projects' tabs, and a user profile for 'jonathan.irish@gmail.com'. The main header displays the experiment name 'Diggins et al., Methods 2015 - Step 8 (Clone)'. Below this, a toolbar contains various analysis tools: 'Actions', 'Illustrations', 'Sample Tags', 'SPADE', 'viSNE', 'Gating', 'Scales', 'Comps', and 'Private'. The 'SPADE' button is highlighted with a pink box, and its dropdown menu is open, showing options: 'New SPADE Analysis', 'View All Analyses', and 'Recent'. The 'New SPADE Analysis' option is also highlighted with a pink box. Other interface elements include a 'Working Illustration' section with 'Save' and 'Print' icons, and a 'Settings Templates' link.



# Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 8”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Run SPADE.

- Set SPADE to use 100 nodes & 1% downsampling.
- Select just the 2 t-SNE channels.
- Click the green bar to run the analysis.

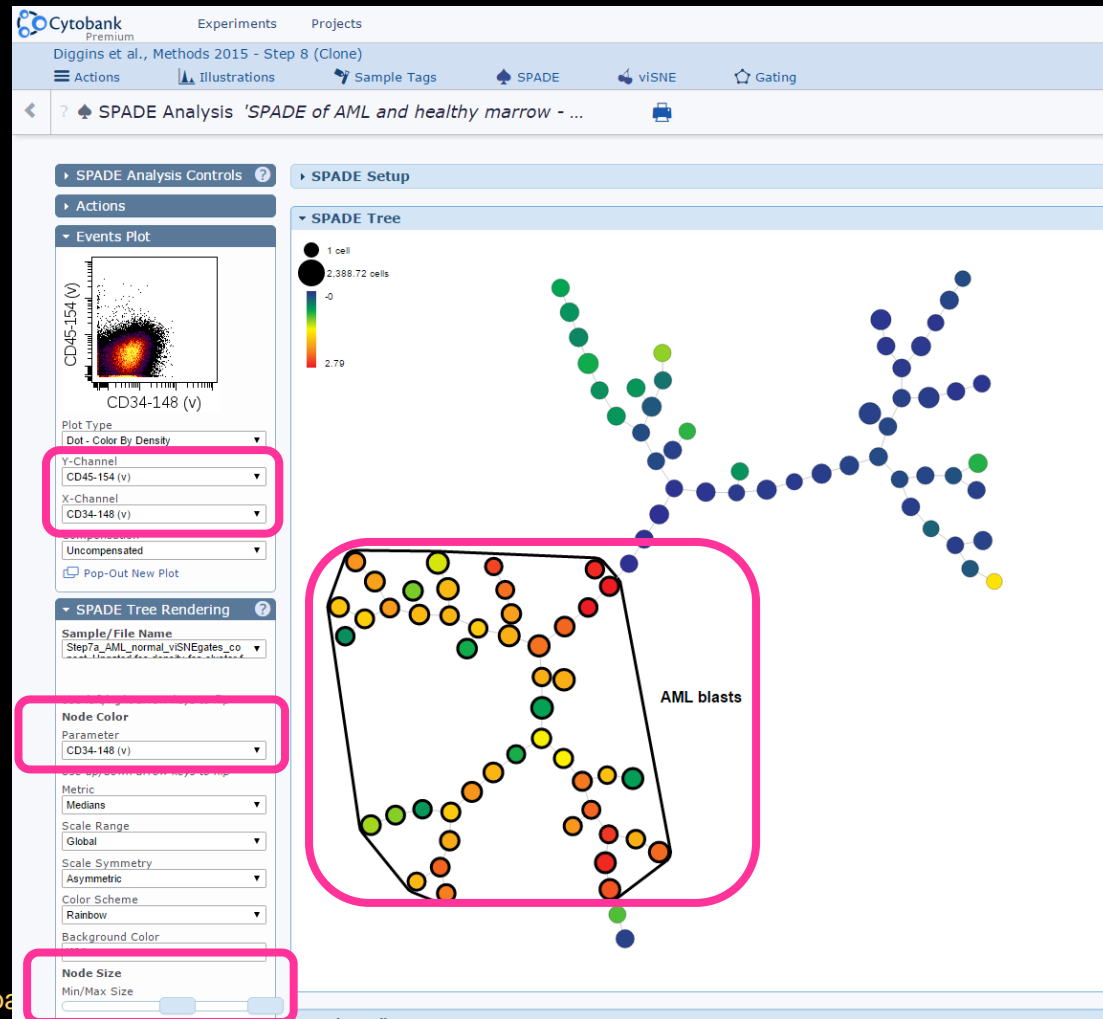
The screenshot displays the Cytobank SPADE Analysis interface. The top navigation bar includes 'Cytobank Premium', 'Experiments', 'Projects', and 'Admin'. The main header shows the experiment name 'Diggins et al., Methods 2015 - Step 8 (Clone)' and various action buttons like 'Actions', 'Illustrations', 'Sample Tags', 'SPADE', 'viSNE', and 'Gating'. The current analysis is titled 'SPADE Analysis 'SPADE of AML and normal marrow''. The interface is divided into two main panels: 'SPADE Analysis Controls' and 'SPADE Setup'. The 'SPADE Analysis Controls' panel includes a 'Compensation' dropdown set to 'File-Internal Compensation', a 'Target Number of Nodes' input field set to '100', and a 'Downsampled Events Target' section with 'Percent' selected and a value of '1'. The 'SPADE Setup' panel has a 'Settings' section with a 'Population' list (currently 'Ungated') and a 'Clustering Channels' list. The 'Clustering Channels' list shows 'tSNE1' and 'tSNE2' selected, with a list of 'Unselected Clustering Channels' including 'Time', 'Cell\_length', 'Barium', and various CD markers. A green bar at the bottom of the interface contains the text: 'When you are done setting up your SPADE parameters, click to run'.

# Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 8”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Run SPADE.

- Explore the data in the SPADE viewer.
- Make your node size larger. Set the x/y axes to CD45 and CD34. Set the Node Color Parameter to CD34.
- See if you can find the AML blasts and bubble them.
- View various Parameters.

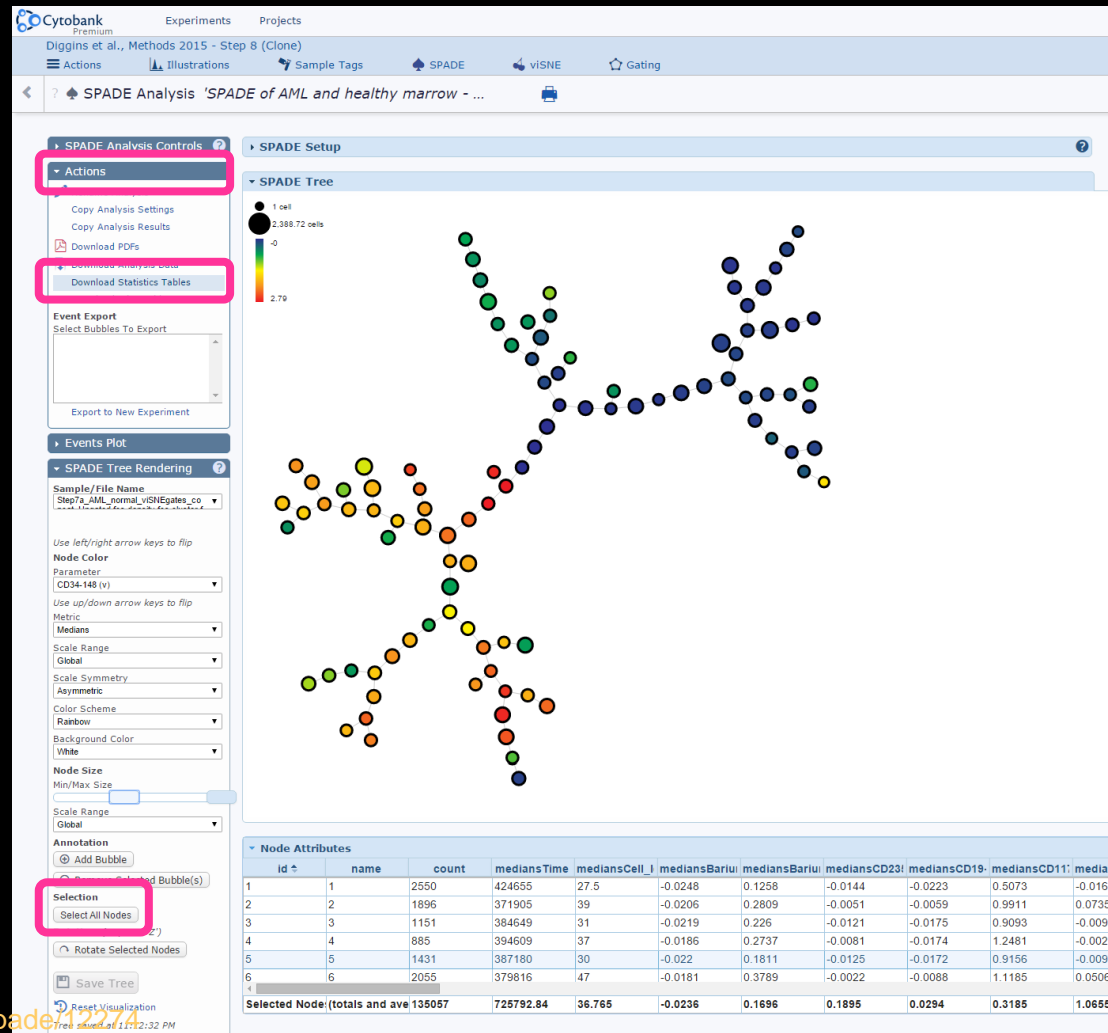


# Steps 8 and 9: SPADE Clustering on t-SNE Channels

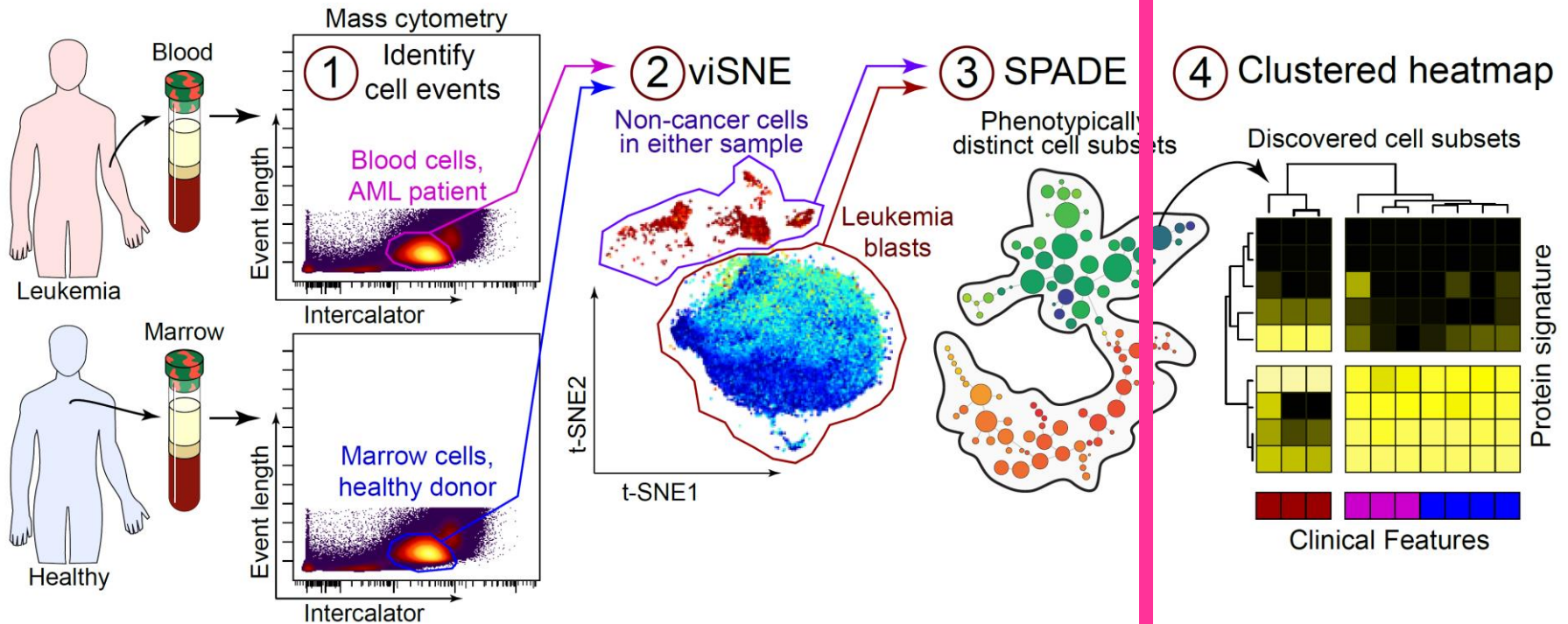
Workflow summary:

- 1) Clone experiment “Diggins et al., Methods 2015 – Step 8”
- 2) Set the Scales so that the scale argument is 15 for all channels.
- 3) Run SPADE.

- Scroll way down and click “Select All Nodes”.
- In the “Actions” panel, Choose “Download Statistics Table”
- In the ZIP file, you will find a folder called “By Sample” that has spreadsheets of median expression by node for all samples. These can be sorted, filtered, and clustered.



# Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

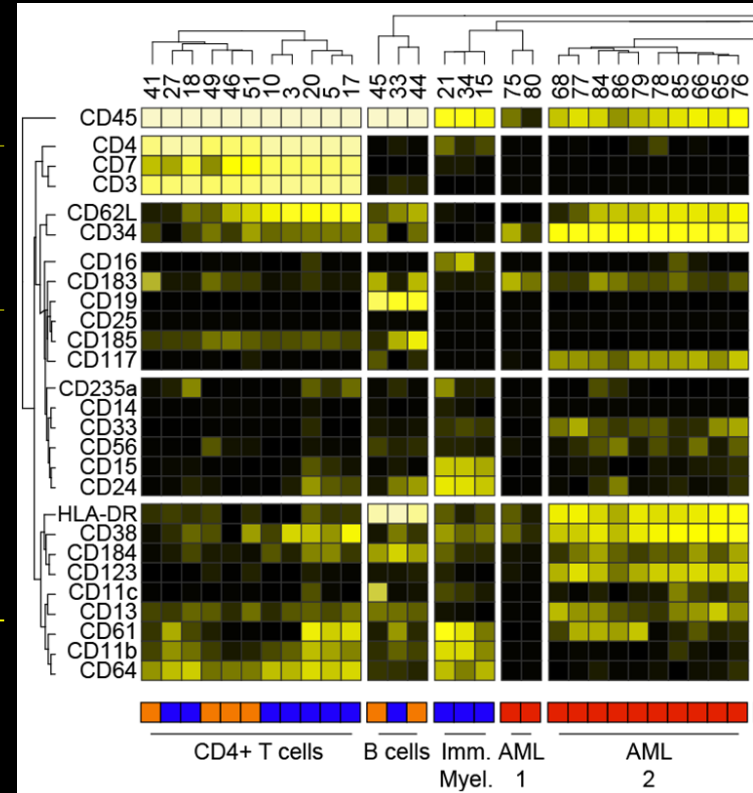


# Single Cell Biology Data Analysis

Data collection	1) Panel design 2) Data collection
Data processing	3) Cell event parsing 4) Scale transformation
Distinguishing initial populations	5) Live single cell gating 6) Focal population gating
Revealing cell subsets	7) Feature selection 8) Dimensionality reduction 9) Identify cell clusters 10) Cluster refinement
Characterizing cell subsets	11) Feature comparison 12) Model populations 13) Learn cell identity 14) Statistical testing

Grouped populations of cells

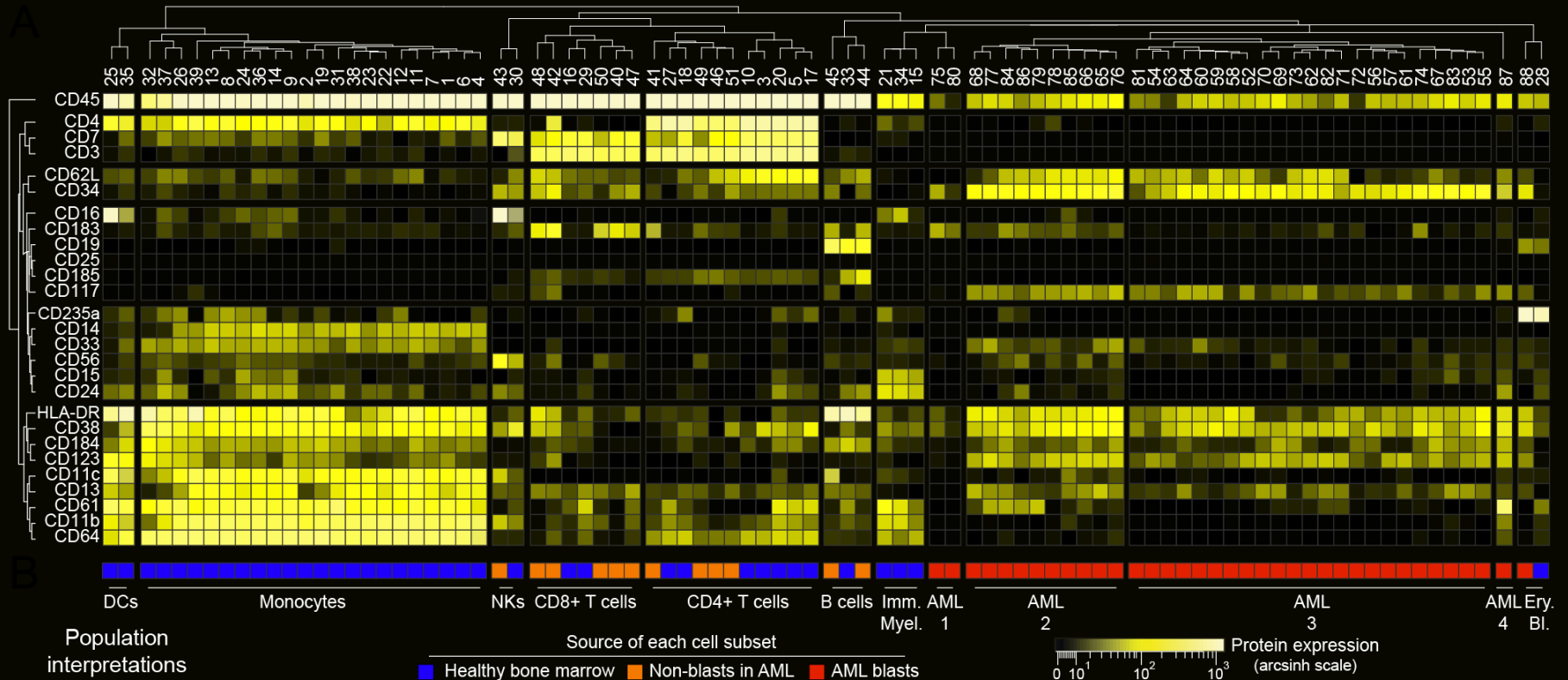
Heatmap: clustering populations



Compare and visualize populations across datasets

# Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

## 3. Cluster SPADE Nodes and Display as Heatmap of Medians

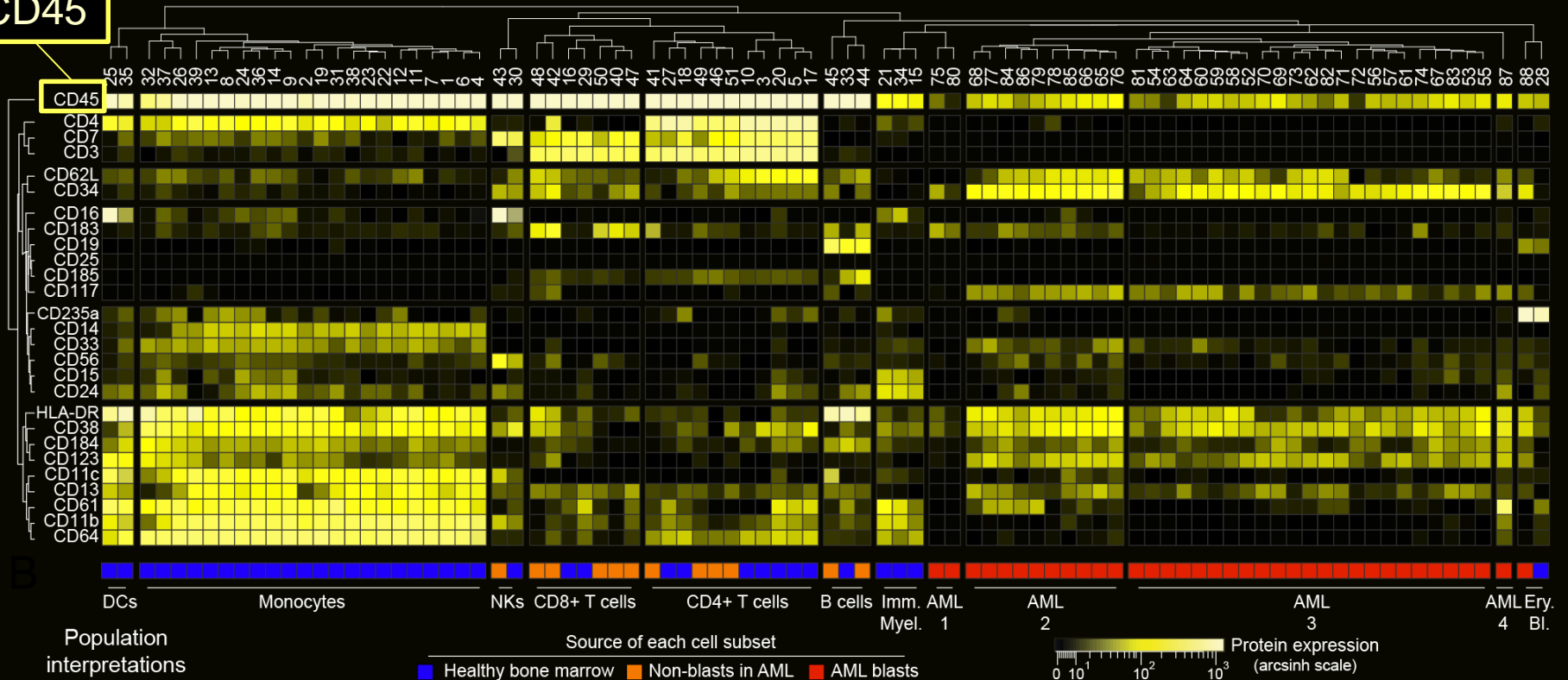


Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.

# Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

## 3. Cluster SPADE Nodes and Display as Heatmap of Medians

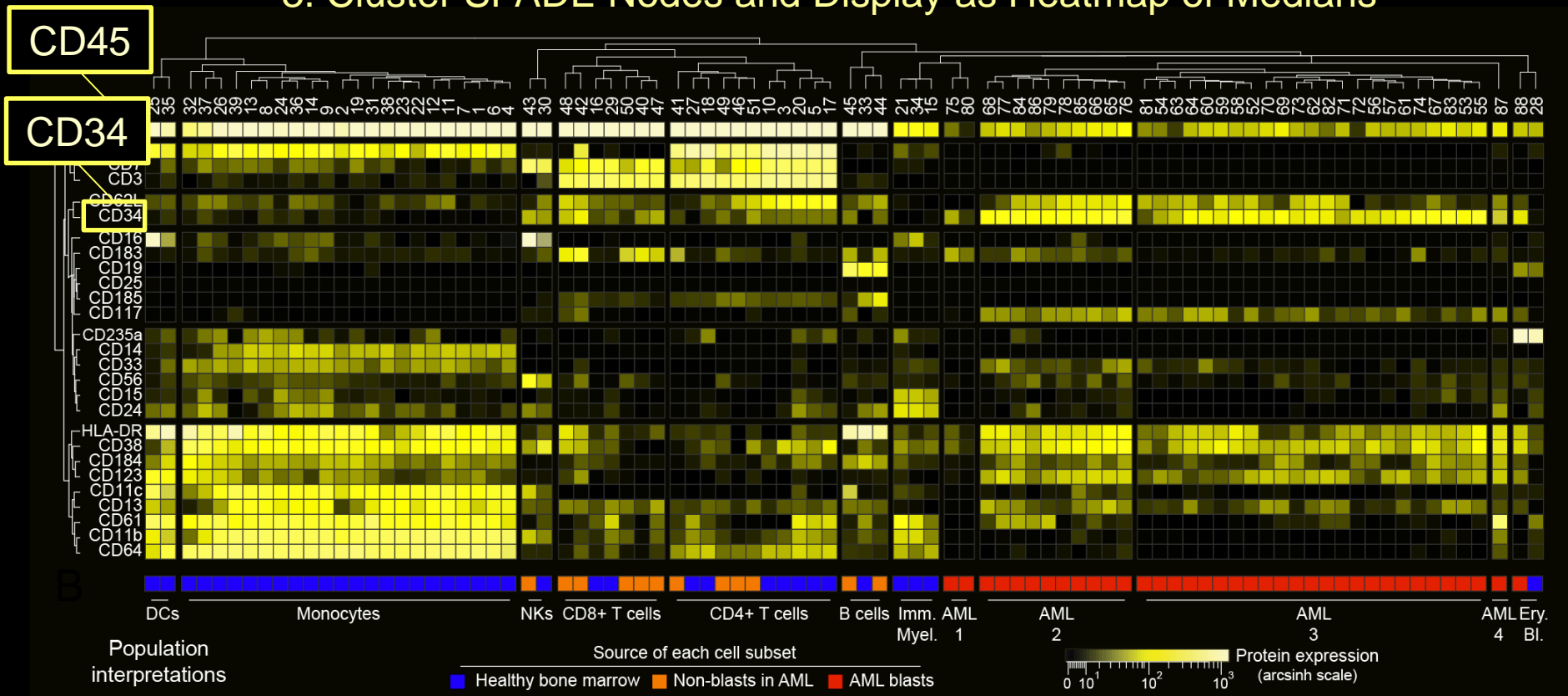
CD45



Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.

# Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

## 3. Cluster SPADE Nodes and Display as Heatmap of Medians

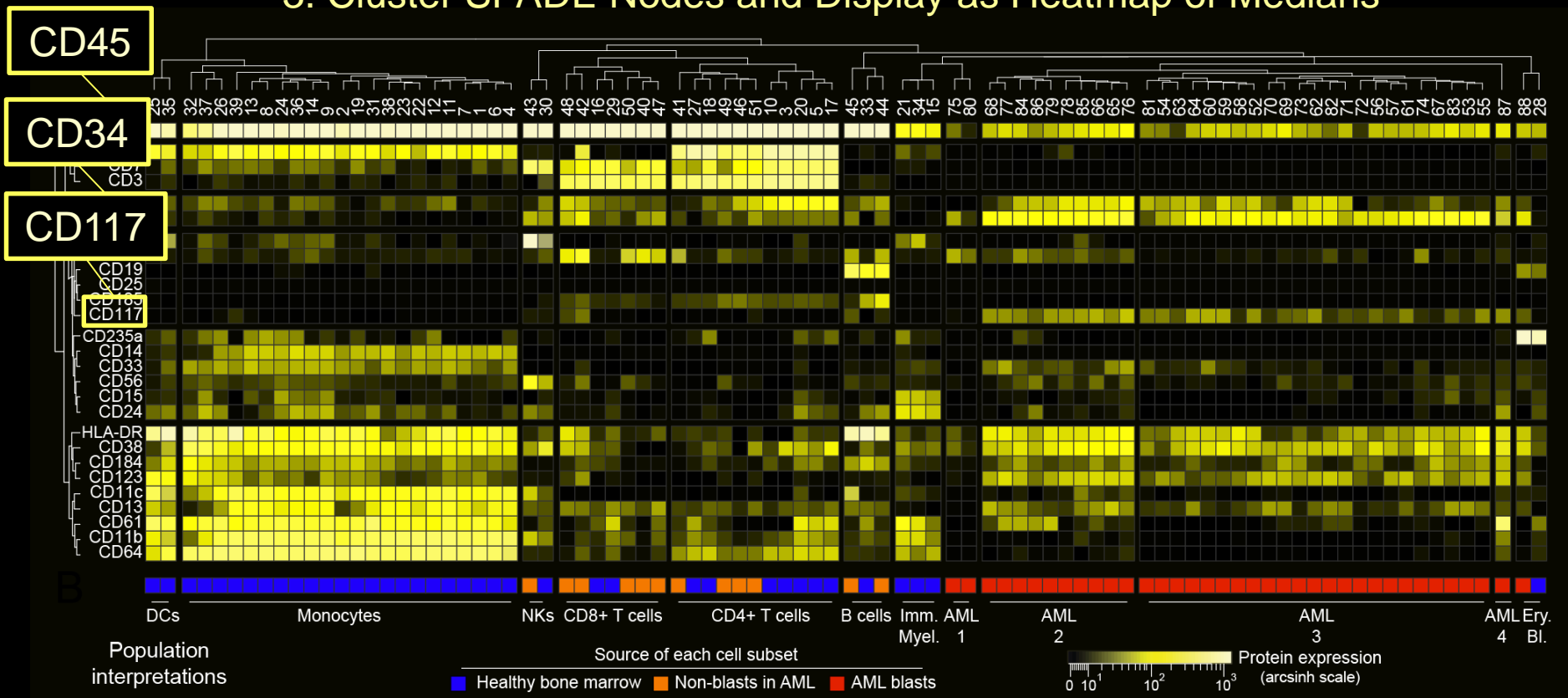


Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.



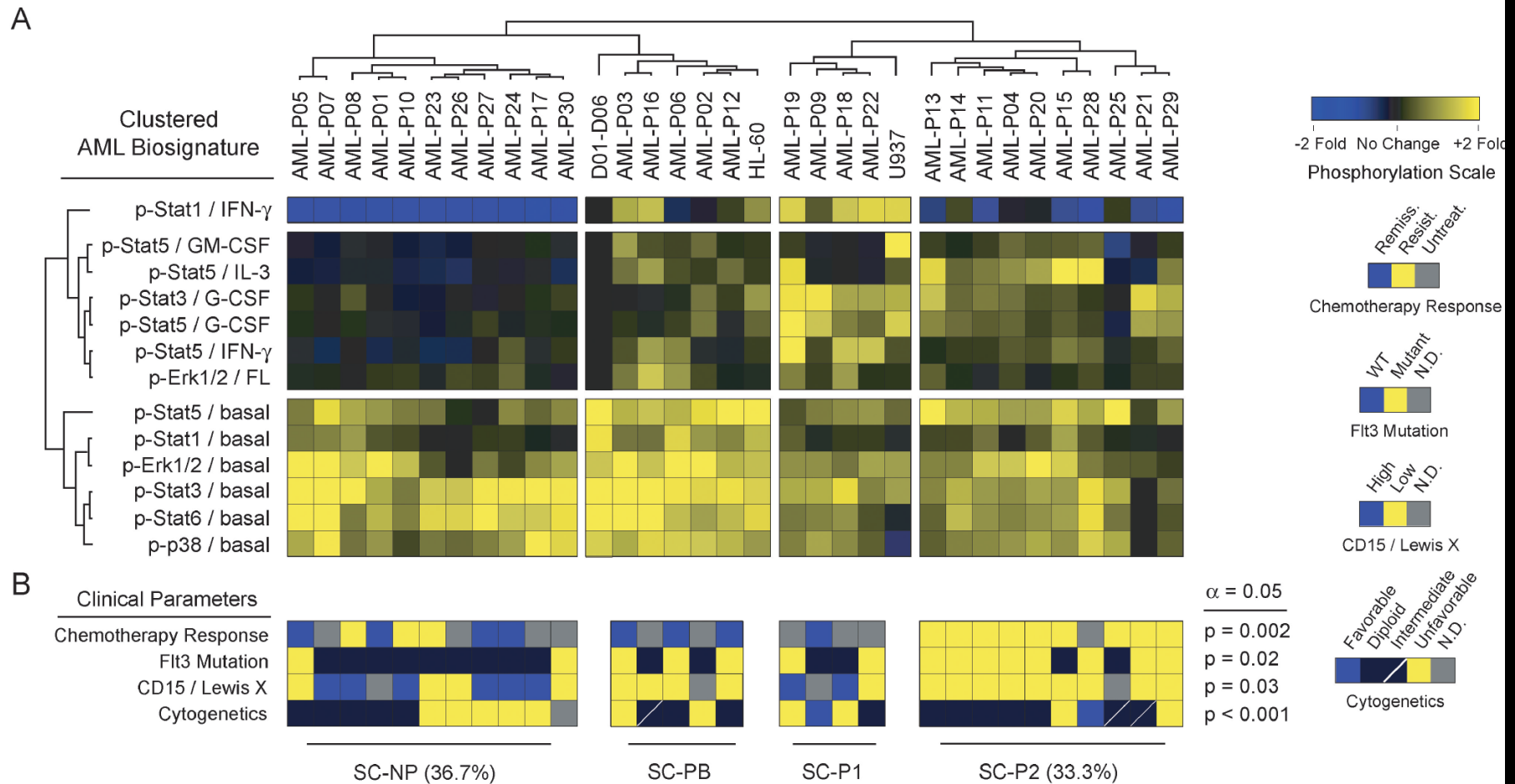
# Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

## 3. Cluster SPADE Nodes and Display as Heatmap of Medians



Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.

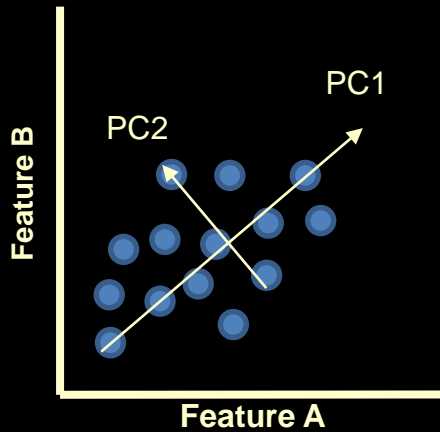
# Heatmaps Also Visualize Other Data Types (e.g. Stratified Clinical Outcomes) and Compare Across Analysis Runs



Heatmaps visualize across integrated data types, e.g. clinical outcomes, cytogenetics, & signaling profiles

Other analysis tools

# Principal Component Analysis



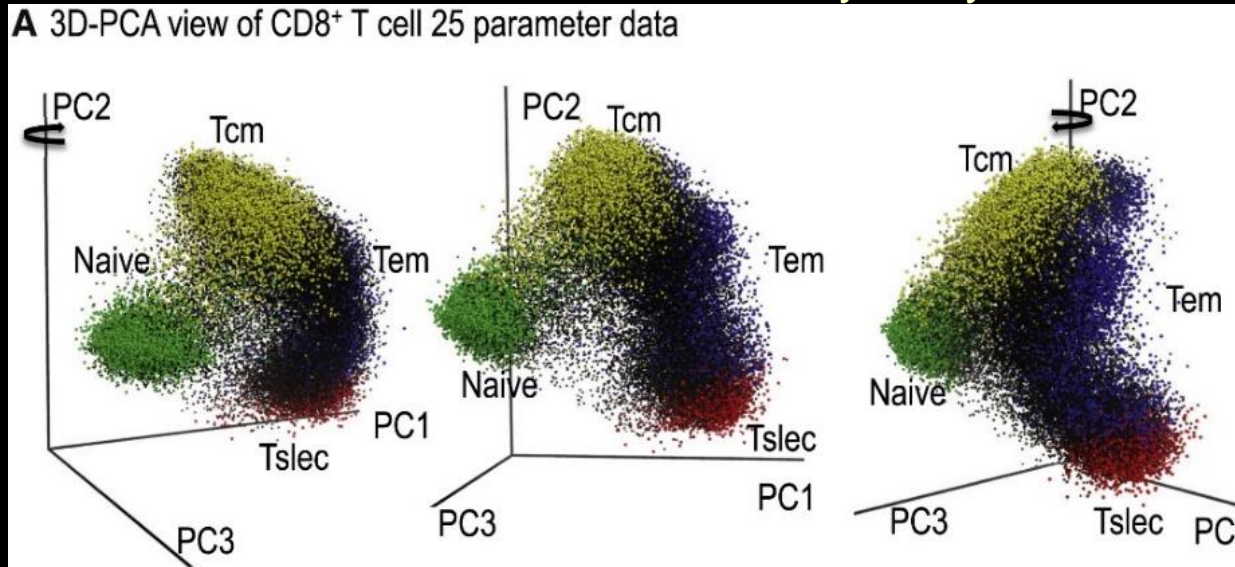
	Feature A	Feature B	Feature C
Principal Component 1	$X_{A1}A$	$X_{B1}B$	$X_{C1}C$
Principal Component 2	$X_{A2}A$	$X_{B2}B$	$X_{C2}C$
Principal Component 3	$X_{A3}A$	$X_{B3}B$	$X_{C3}C$

$$PC1 = X_{A1}A + X_{B1}B + X_{C1}C$$

$$PC2 = X_{A2}A + X_{B2}B + X_{C2}C$$

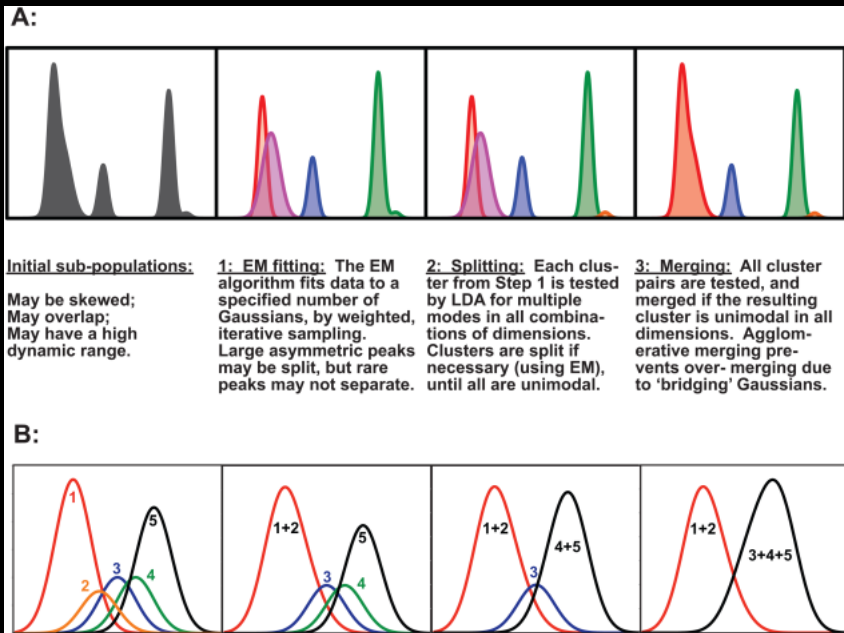
$$PC3 = X_{A3}A + X_{B3}B + X_{C3}C$$

## PCA used to Reduce Dimensionality of CyTOF Data

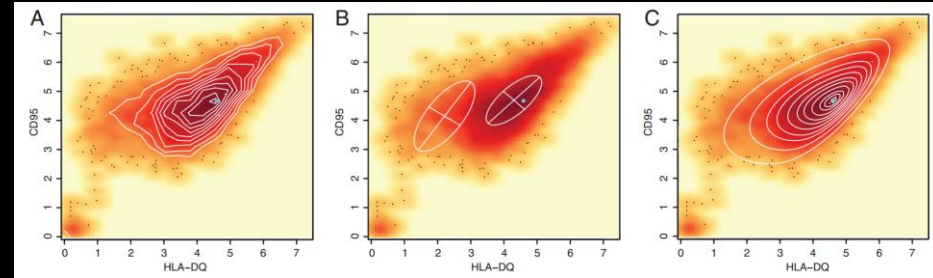


# Mixture Modeling

## SWIFT



## FLAME

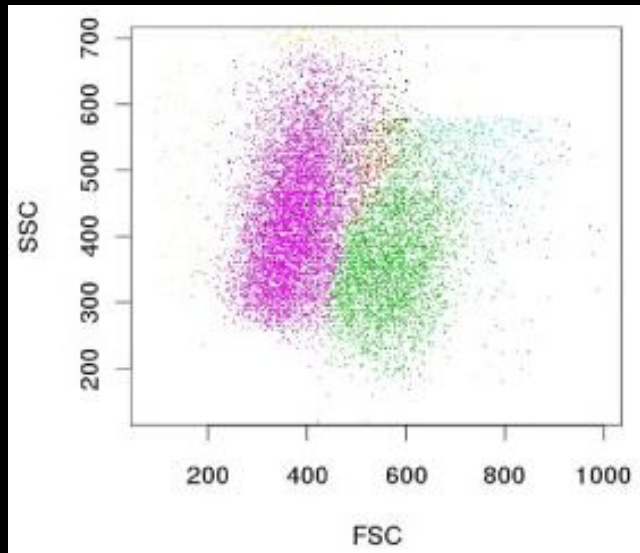
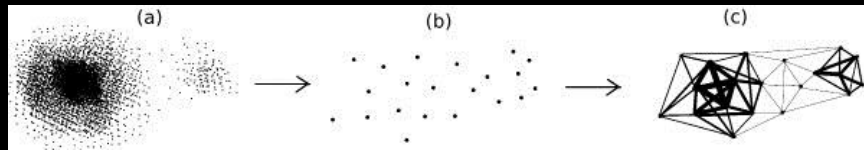


Pyne et al, 2009 *PNA*

Mosmann et al, 2014 *Cytometry A*

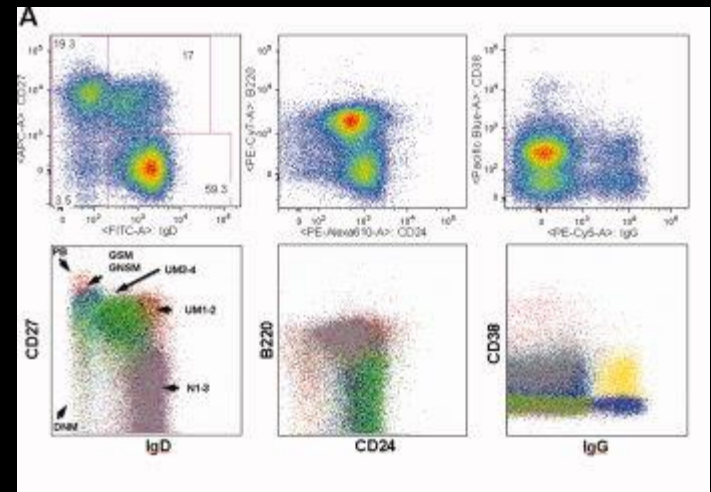
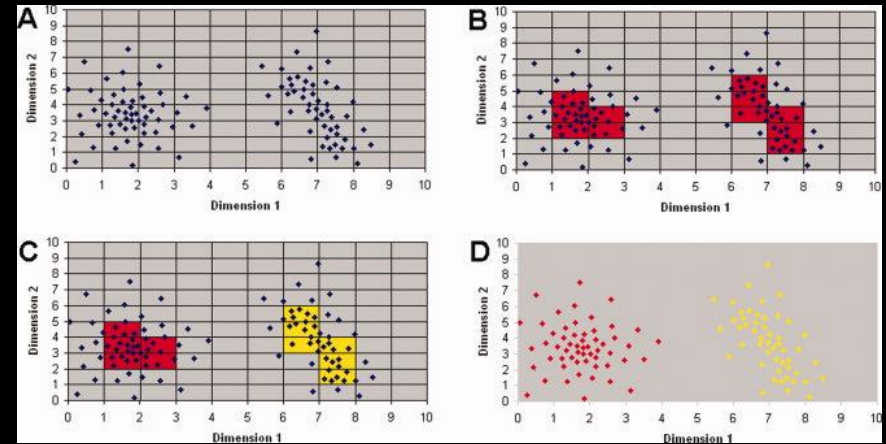
# Automated Clustering and Population Identification Methods Based on Density

## SamSpectral



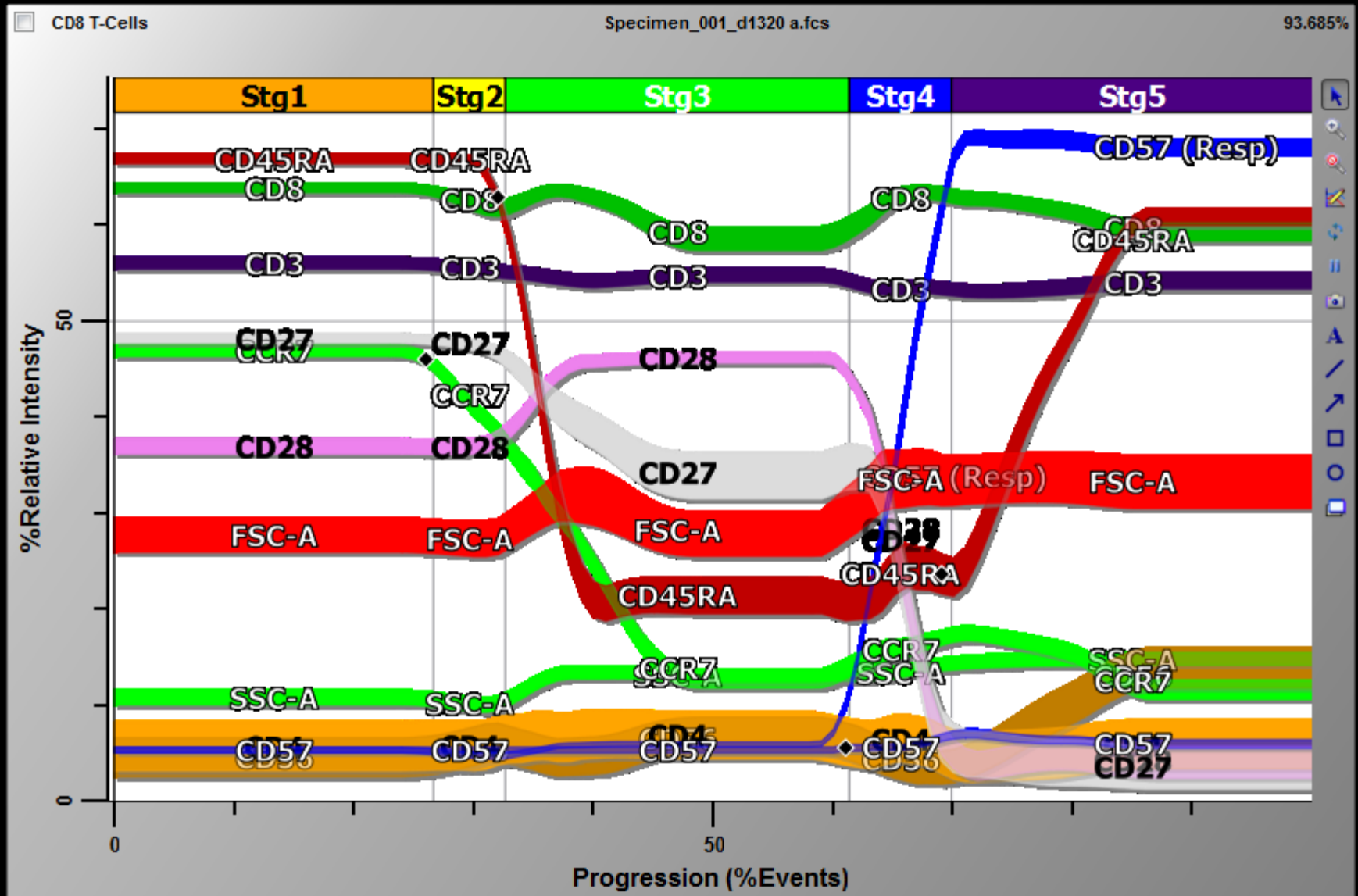
Zare et al, 2010 *BMC Bioinformatics*

## FLOCK

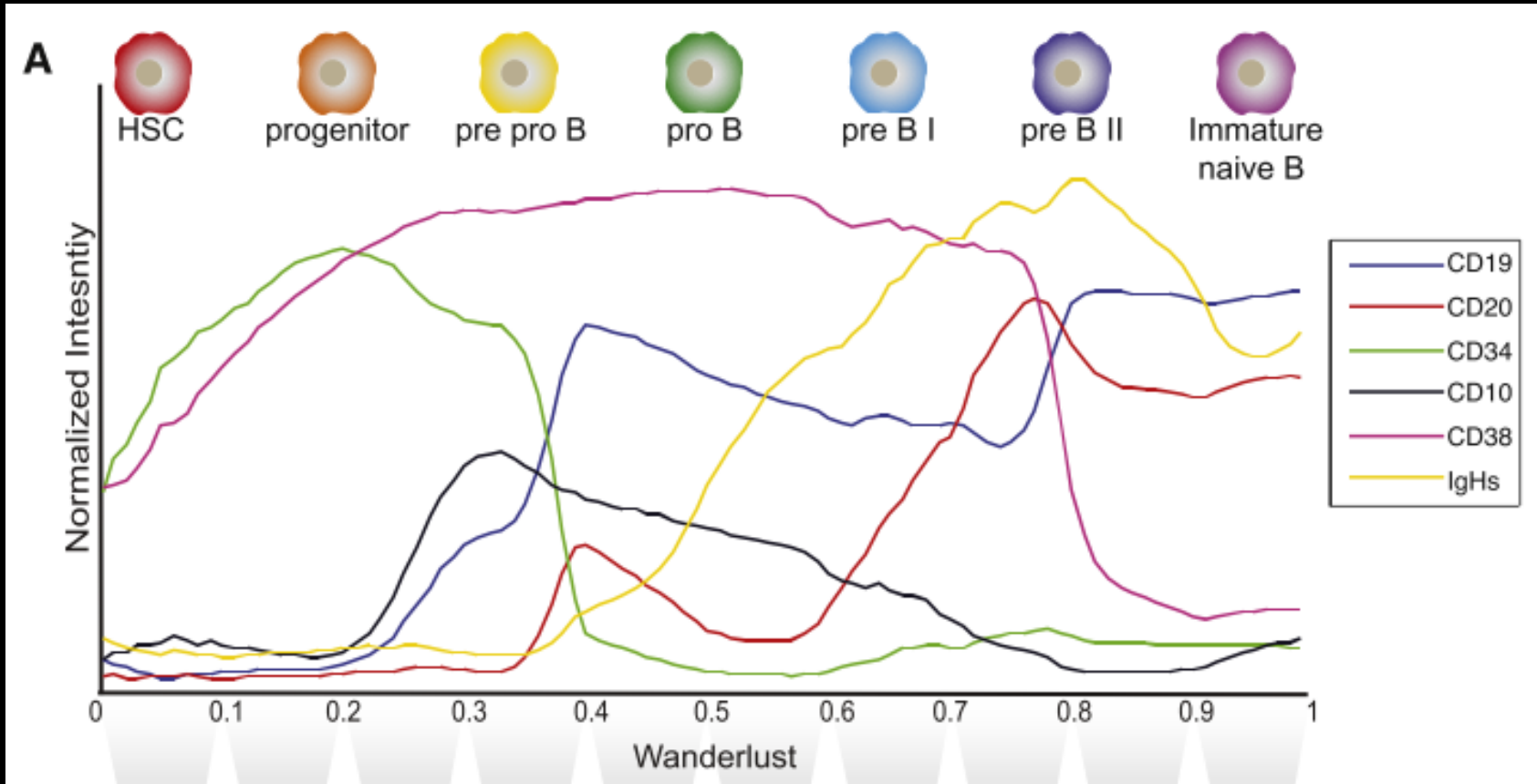


Qian et al, 2010 *Cytometry B Clin Cytom*

# Gemstone Uses Supervised Analysis to Identify Progressions

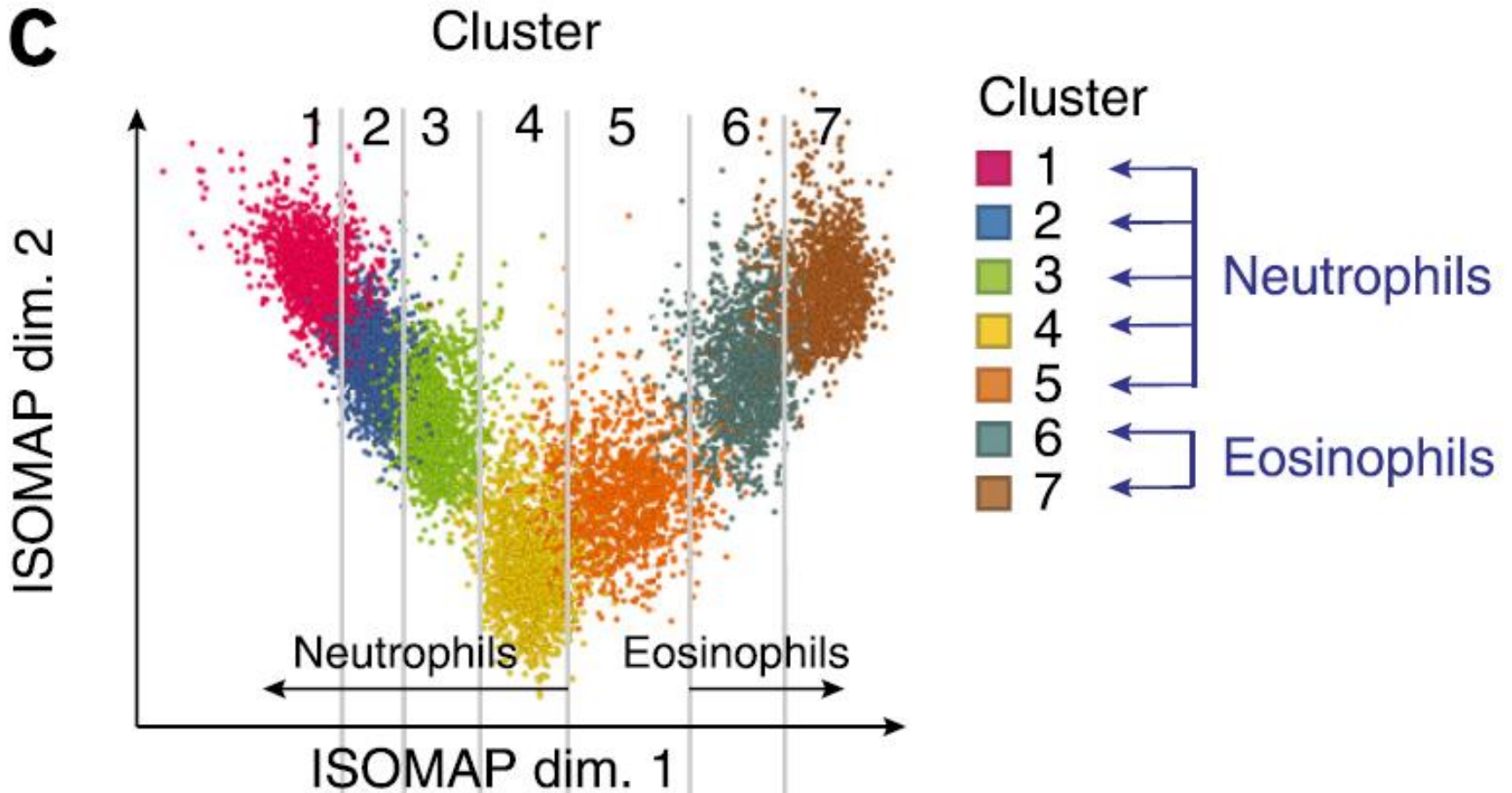


# Wanderlust Identifies Phenotypic Progression





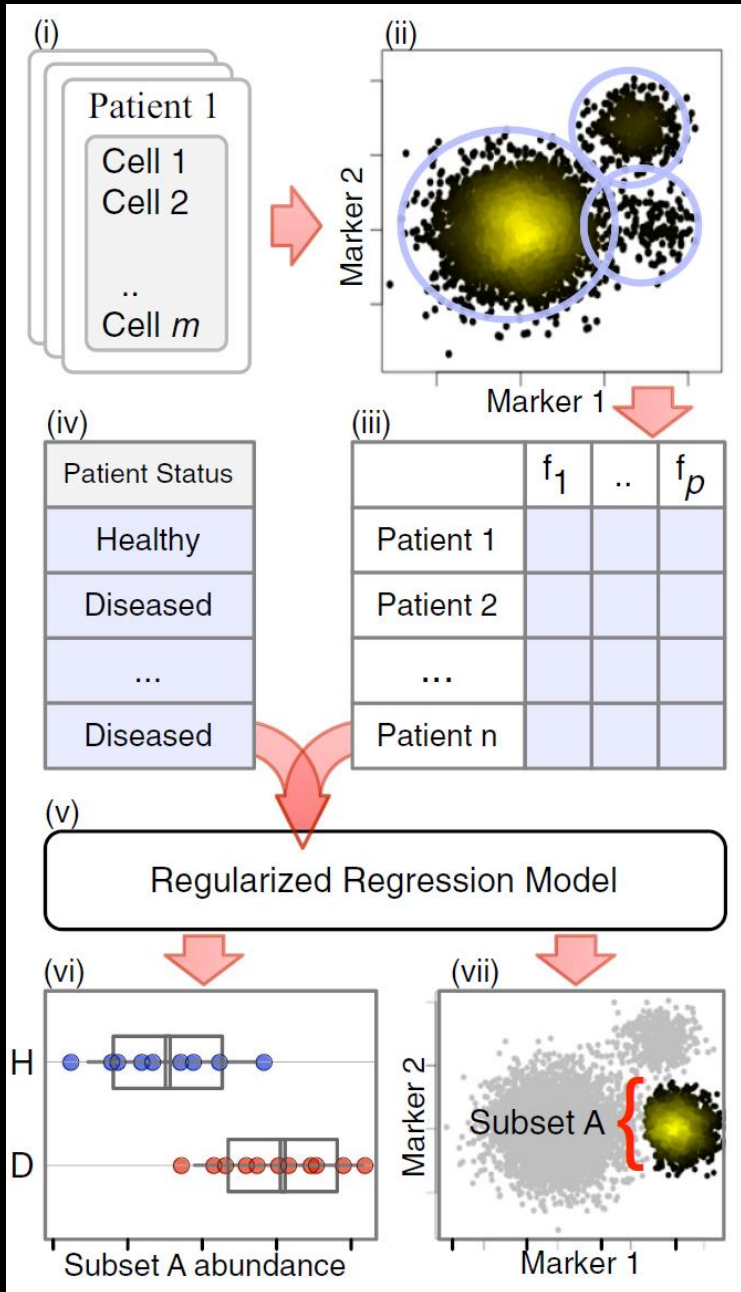
# ISOMAP guided analysis



(c) ISOMAP dimensionality reduction

to compare overall phenotypic relatedness of populations of neutrophil-like and eosinophil-like cells<sup>31</sup>. Top, cells color-coded by DensVM cluster number are plotted by their scores for ISOMAP dimensions 1 and 2. Binned median expression of defining markers (middle) and the tissue composition (percentage of each cluster as a fraction of total granulocytes from each tissue, bottom) of cells along this phenotypic progression defined by ISOMAP dimension 1 and DensVM clusters 1–7 are plotted.

# Citrus: Supervised Population Finding

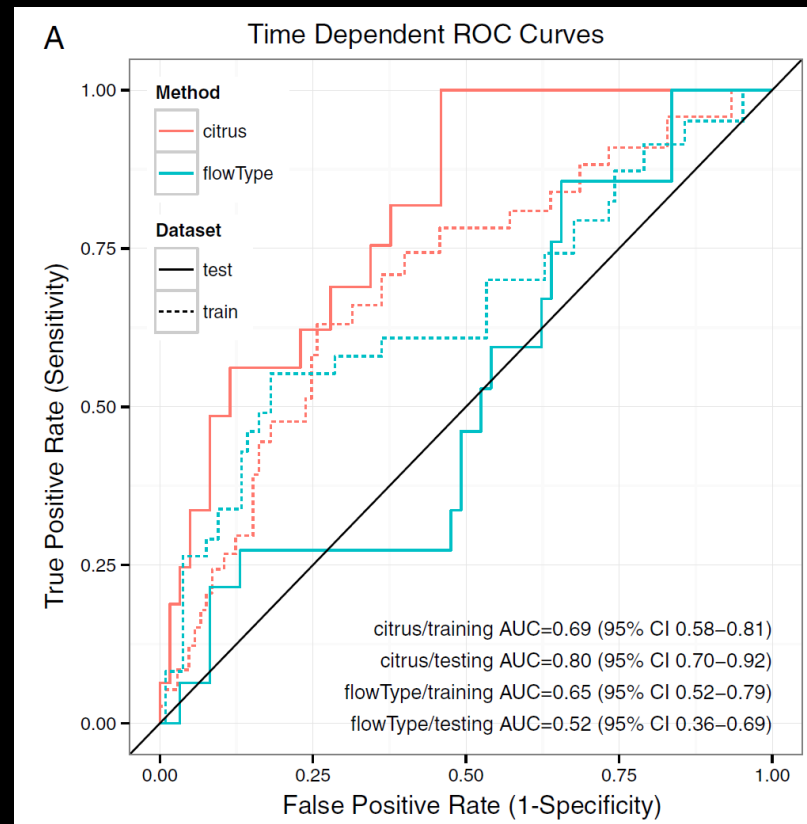


## Automated identification of stratifying signatures in cellular subpopulations

Robert V. Bruggner<sup>a,b</sup>, Bernd Bodenmiller<sup>c</sup>, David L. Dill<sup>d</sup>, Robert J. Tibshirani<sup>e,f,1</sup>, and Garry P. Nolan<sup>b,1</sup>

<sup>a</sup>Biomedical Informatics Training Program, Stanford University Medical School, Stanford, CA 94305; <sup>b</sup>Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, and Departments of <sup>c</sup>Computer Science, <sup>e</sup>Health Research and Policy, and <sup>f</sup>Statistics, Stanford University, Stanford, CA 94305; and <sup>d</sup>Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland

Contributed by Robert J. Tibshirani, May 14, 2014 (sent for review February 12, 2014)



Final notes & conclusions...

# Computational Tools + HD Cytomics Together Are Powering A New Era in Clinical Oncology & Immunity

## 1) Need pre-treatment prognosis & prediction

If diagnostic scheme does not provide actionable information, fix it.  
Who will benefit from expensive cell based therapies (~\$500,000 ea.)?  
In the absence of mutations, clinical response can be predicted by cell profiling.

## 2) Need to monitor treatment longitudinally

See early whether patient responded / adjust treatment, as needed.  
Monitor whether treatment is still required.

## 3) Need to check multiple biomarkers with one test

As with genetic tests, multiplexing biomarkers will give more information per sample, catch the unexpected, and cost less than repeated testing

## 4) Need to monitor biomarkers on all cell types

PD-L1 is a great example – expressed by many cell types & can be activated.

## 5) Need to characterize all cell types to monitor cancer

Evolving cancer cells adopt unexpected phenotypes.

# Conclusions: Data Analysis & Mapping Cell Identity

## Workflow summary:

- 1) viSNE with minimal pre-gating
- 2) SPADE, works especially well on t-SNE axes
- 3) Heatmap to compare with other data, do statistical tests

1) A modular workflow allows comparison of different tools at each step. Synthetic channels are useful (e.g. t-SNEs).

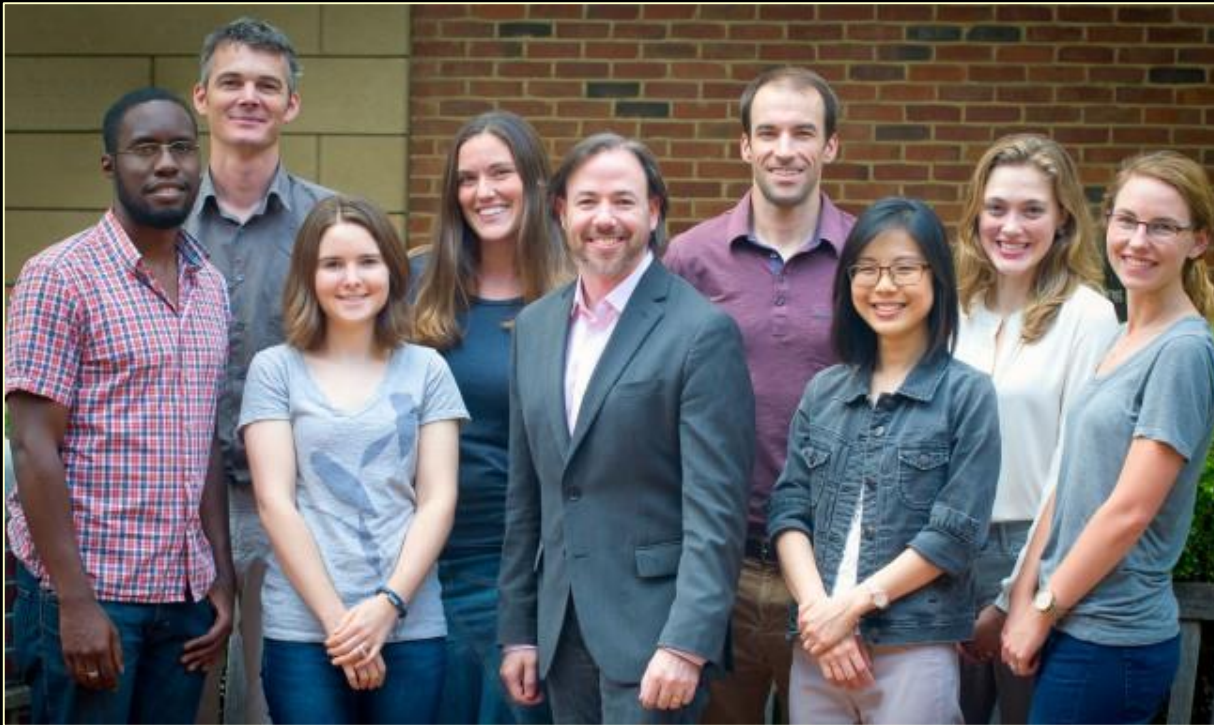
2) Non-linear transformation may help; tool selection depends on 'what works', biology, & data shape.

3) Many outstanding tools are available (see Diggins et al. for reference list). Still need tools that learn.

# Discussion Questions Covered in Today's Course

- 1) What are key differences between tools (viSNE, SPADE, PCA)? What is the difference between transforming, clustering, and modeling data? What type of modeling are we doing (if any)?
- 2) What does non-linear vs. linear analysis mean? Does the data's scale matter for analysis (arcsinh5, arcsinh15, linear)?
- 3) What do viSNE and SPADE settings do (viSNE iterations, SPADE downsampling & node #)? When should they be changed?
- 5) How does one compare new samples with a prior analysis? How do we test tools with expert gating?
- 6) What are some "red flags" indicating problems? What does a good viSNE or SPADE analysis run look like?

# Acknowledgements & Thank you!



L to R: Deon Doxie, Mikael Roussel, Kirsten Diggins, Cara Wogsland, Jonathan Irish, Brent Ferrell, Nalin Leelatian, Hannah Polikowsky, Allison Greenplate

## Oslo Collaborators

June Myklebust  
Kanutte Huse

## VU Collaborators

Madan Jagasia  
Michael Savona  
Jeff Sosman  
Doug Johnson  
Pierre Massion  
Rebecca Ihrle  
Ann Richmond

NIH/NCI

R00 CA143231 (Irish Lab)  
F31 CA199993 (Greenplate)  
T32 CA009592 (Doxie)  
R25 CA136440 (Diggins)  
F31 CA203383 (Diggins)  
K12 CA090625 (Ferrell)  
VISP (Leelatian)

Vanderbilt-Ingram Cancer Center (VICC)  
Ambassadors, Hematology Helping Hands

Vanderbilt University Discovery

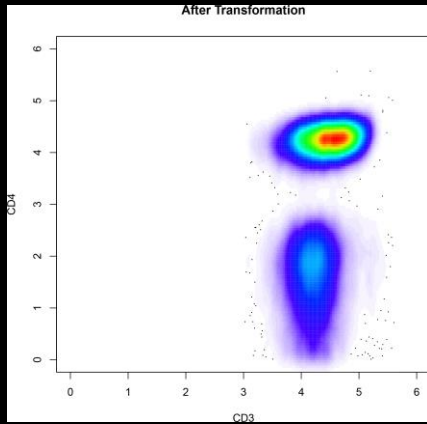
Norwegian Cancer Society (Huse)  
Swiss Foundation Nuovo-Soldati (Roussel)

**R/BIOCONDUCTOR**

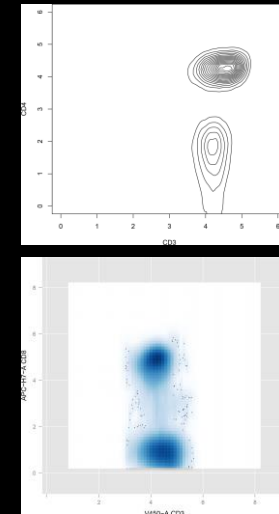
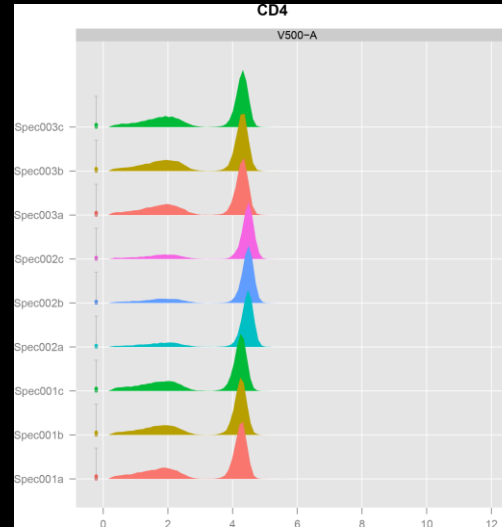


# R Flow Cytometry Data Analysis

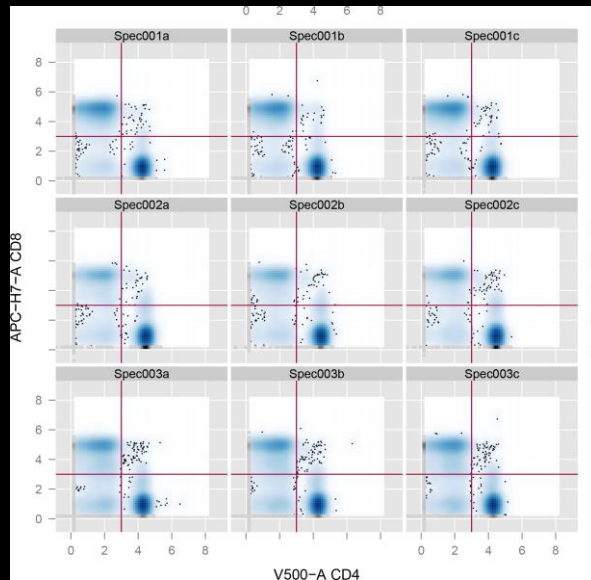
## Arcsinh Transformation of Data



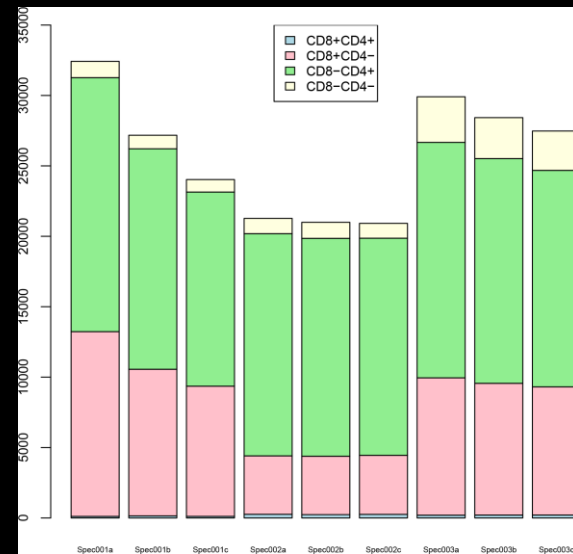
## Visualization



## Gate T Cells on CD4 and CD8 Expression



## Compare T cell Subset Frequencies Between Individuals



# What is R/Bioconductor?

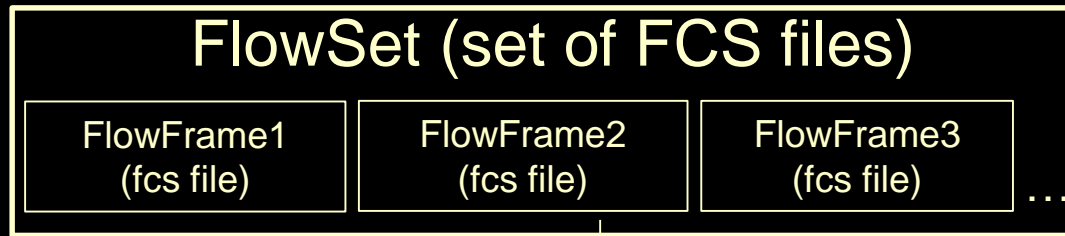
- **R**: statistical and graphical programming language/environment
- **Bioconductor**: provides open source packages for bioinformatics analysis in R

The screenshot shows the RStudio website's download page. At the top, there is a navigation bar with 'Products', 'Resources', 'Pricing', 'About Us', and 'Blog'. The main heading is 'Download RStudio'. Below this, there is a description of RStudio as a set of integrated tools. A callout box asks 'Do you need support or a commercial license?' with a link to 'Check out our commercial offerings'. Another callout says 'Let's stay in touch. Give us your email and we'll keep you in the loop.' with an email input field and a 'Submit' button. The 'Download RStudio Desktop v0.98.1087 — Release Notes' section states that RStudio requires R 2.11.1 or higher. At the bottom, there is a table of installers for all platforms.

Installers	Size	Date	MD5
RStudio 0.98.1087 - Windows XP/Vista/7/8	45 MB	2014-10-30	a14d1a2dc0e4867ffac254c947f326d9
RStudio 0.98.1087 - Mac OS X 10.6+ (64-bit)	38.4 MB	2014-10-30	dc033590b5129fa374ba2cb54a16595
RStudio 0.98.1087 - Debian 6+/Ubuntu 10.04+ (32-bit)	53 MB	2014-10-30	a8724f9d2159365336aba6a3334c1646
RStudio 0.98.1087 - Debian 6+/Ubuntu 10.04+ (64-bit)	54.9 MB	2014-10-30	dcee52d57d58cde13cd89fc632e5758f
RStudio 0.98.1087 - Fedora 13+/openSUSE 11.4+ (32-bit)	53.4 MB	2014-10-30	80ed866421d526927bad94c853ab9bb0
RStudio 0.98.1087 - Fedora 13+/openSUSE 11.4+ (64-bit)	55 MB	2014-10-30	dd0e8ef2665d3687935388e8df73fa1

The screenshot shows the Bioconductor website's home page. At the top, there is a navigation bar with 'Home', 'Install', 'Help', 'Developers', and 'About'. The main heading is 'About Bioconductor'. Below this, there are several sections: 'Install' (Get started with Bioconductor), 'Learn' (Master Bioconductor tools), 'Use' (Create bioinformatic solutions with Bioconductor), and 'Develop' (Contribute to Bioconductor). There is also a 'News' section and a 'Support' section. The 'Support' section includes a link to 'Read the posting guide' and 'bio-devel mailing list (for package authors)'. The 'Events' section includes 'Next Generation Data Analysis' and 'BioC Europe 2015'. The 'Tweets' section shows a tweet from Bioconductor.

# Anatomy of an FCS File in R



## FlowFrame (fcs file)

### Exprs

- Fluorescence intensity matrix

Cell	Parameter1	Parameter2
1	MFI_1-1	MFI_1-2
2	MFI_2-1	MFI_2-2

> exprs(FlowFrame)

### Parameters

- Annotated data frame  
- Info about parameters (stains)

rowNames	\$P1,\$P2,\$P3...
varLabels	Name, desc...max range
varMetadata	labelDescription

> parameters(FlowFrame)

### Description

-List of information from instrument

\$FCSversion
[1] "3"
\$\$BEGINANALYSIS
[1] "0"
\$\$ENDANALYSIS
[1] "0"
...

> description(FlowFrame)

# Flow Analysis in R

- flowCore: access data in fcs files
- flowQ: quality control for flow data
- flowStats: statistical analysis of flow data
- flowViz: visualization of flow data
- flowClust: mixture-modeling to find clusters
- flowMerge: merge clusters from mixture modeling
- flowUtils: parse Gating-ML files
- flowFP: transforms high dimensional data using modeling to facilitate data analysis

# flowCore

- R package flowCore
  - Read fcs files into R and perform basic analysis functions
  - View and access intensity matrix

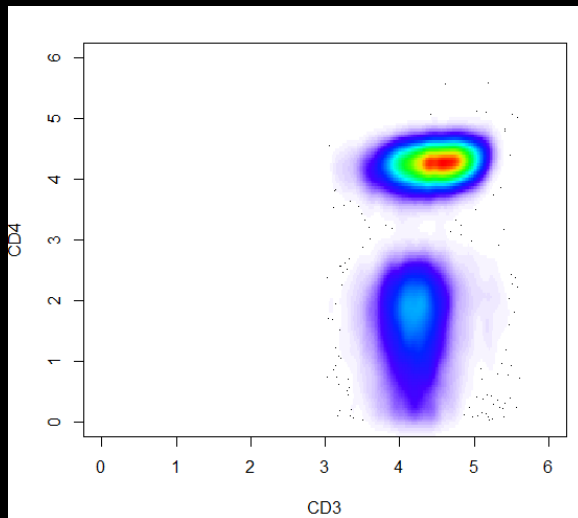
```
> head(exprs(tcCellData$Spec001a))
      FSC-A FSC-H   FSC-W   SSC-A SSC-H   SSC-W FITC-A PerCP-Cy5-5-A APC-A Alexa Fluor 700-A APC-H7-A V450-A
[1,] 74313.46 64310 75730.17 26153.92 28082 61036.37 345.28      1647.36 1310.4      688.5    100.8 3174.22
[2,] 54602.26 45362 78885.72 25446.72 16436 101464.84 236.08      2462.72 2231.1     1161.9    846.9 8924.06
[3,] 67478.18 49500 89338.38 15140.32 17640 56249.21 143.52      2442.96 2241.0     1305.0    687.6 7780.16
[4,] 83260.04 64328 84823.56 31904.08 21015 99493.97 295.36      2546.96 2165.4     1104.3    538.2 8496.02
[5,] 54679.66 43581 82225.88 14768.00 15760 61410.89 260.00      1998.88 2108.7      945.9    346.5 9673.54
[6,] 107701.24 83691 84337.73 30740.32 23988 83983.55 101.92      5040.88 3944.7     2043.0    1160.1 5296.38
      V500-A   PE-A PE-Cy7-A Time
[1,] 6639.540   66.33  102.96 0.4
[2,] 7260.280 1548.36 10053.45 0.7
[3,] 7010.180 1472.13  9365.40 2.1
[4,] 7343.100 1545.39  5906.34 2.1
[5,] 9341.439 1074.15  3357.09 3.0
[6,] 1669.520 1062.27 16261.74 3.2
>
```

```
> print(tcCellData$Spec001a)
flowFrame object 'Spec001a'
with 32423 cells and 16 observables:
  name desc range minRange maxRange
$P1   FSC-A FSC-A 262144      0 262144
$P2   FSC-H FSC-H 262144      0 262144
$P3   FSC-W FSC-W 262144      0 262144
$P4   SSC-A SSC-A 262144      0 262144
$P5   SSC-H SSC-H 262144      0 262144
$P6   SSC-W SSC-W 262144      0 262144
$P7   FITC-A CD57 262144     -111 262144
$P8   PerCP-Cy5-5-A CD28 262144     -111 262144
$P9   APC-A CD27 262144     -111 262144
$P10  Alexa Fluor 700-A CD56 262144     -111 262144
$P11  APC-H7-A CD8 262144     -111 262144
$P12  V450-A CD3 262144      0 262144
$P13  V500-A CD4 262144     -111 262144
$P14  PE-A CCR7 262144     -111 262144
$P15  PE-Cy7-A CD45RA 262144     -111 262144
$P16  Time Time 262144      0 262144
250 keywords are stored in the 'description' slot
```

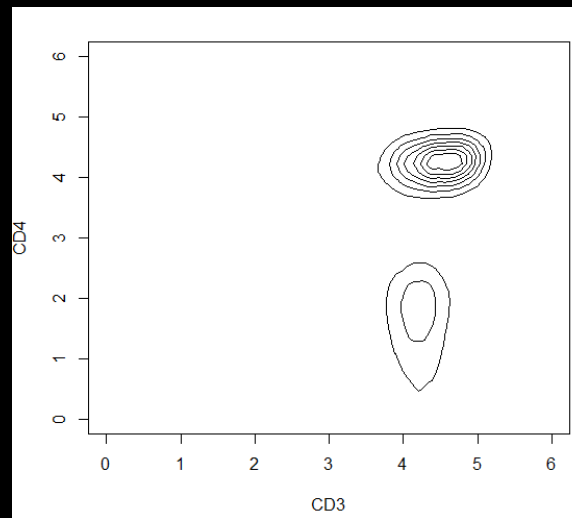
# flowViz

- Contour plot, density plot, scatter plot, trellis plot, and histograms for flow data

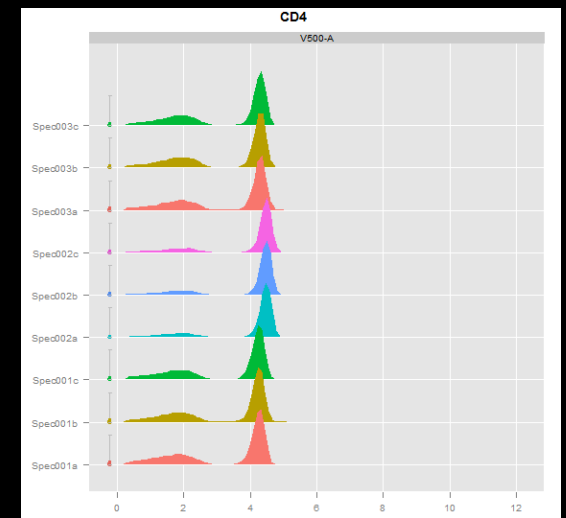
Density Dot (flowPlot(...))



Contour Plot (contour(...))



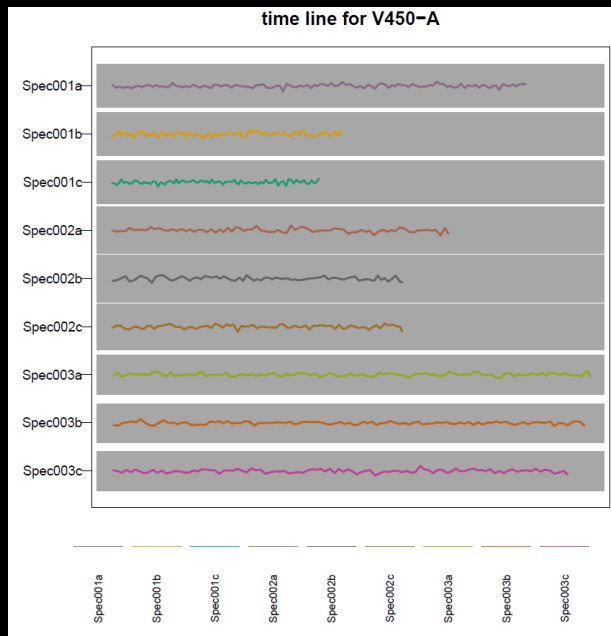
Histograms (densityplot(...))



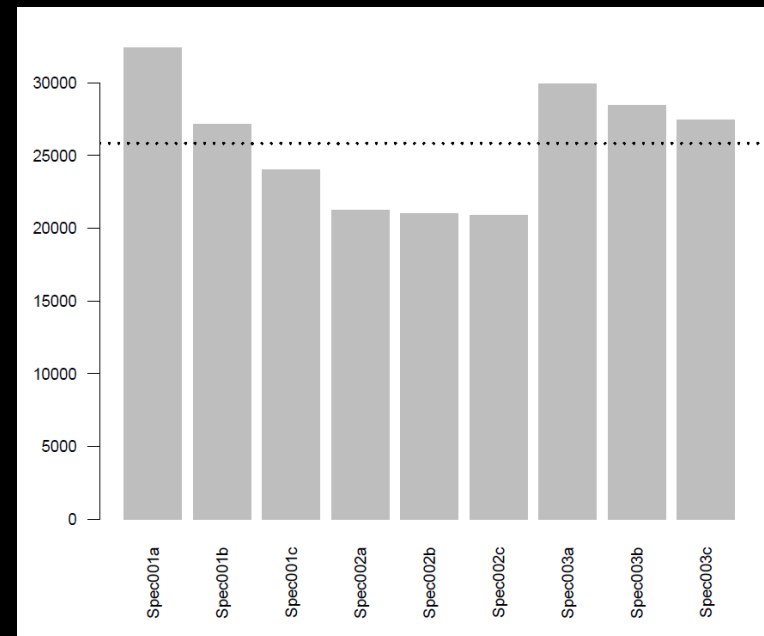
# flowQ

- Compare number of cells between files
- Check for margin events
- Check for time anomalies
- Data normalization
- KL Divergence

## Timeline



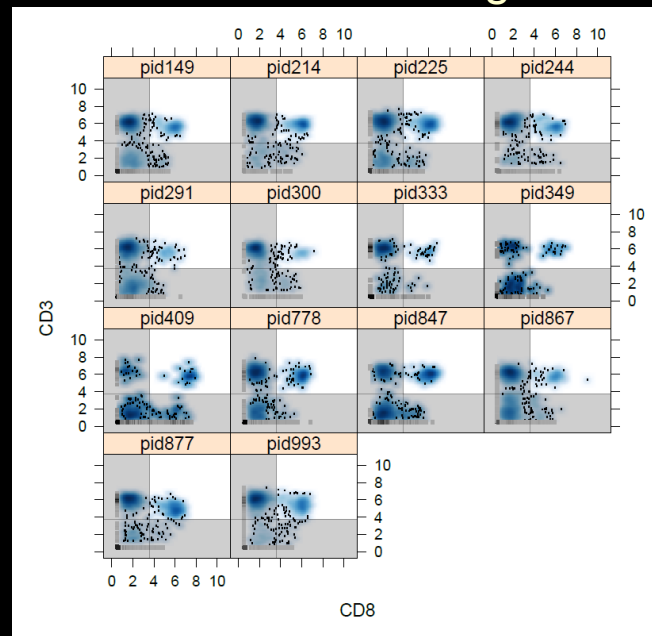
## Cell Number



# flowStats

- Functions for statistical analysis of flow data
  - Probability binning + Chi-squared test
  - Create filters for high density regions (auto-gating)
  - Quadrant gating
  - Methods for data normalization

## Quadrant Gating



## Normalization

