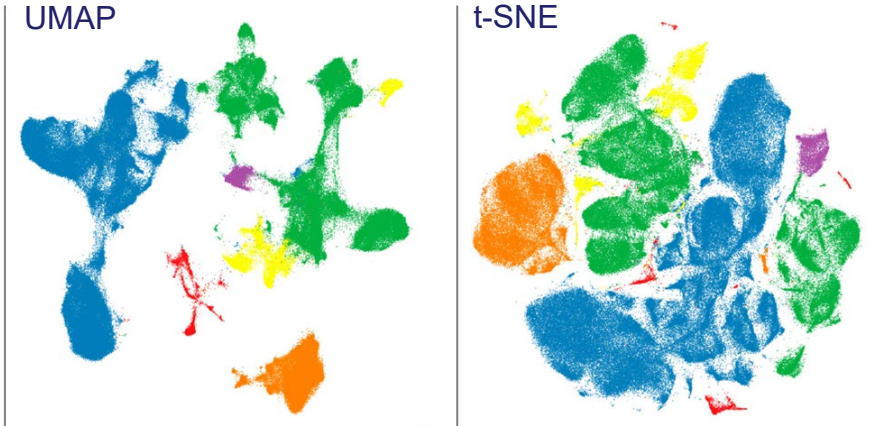


Analysis II: Tailoring Cytometry
Data Science Workflows (90 min)

Starting Poll: <https://www.menti.com/>

Poll Code: 7939 8162

Analysis II: Tailoring Cytometry Data Science Workflows



Becht et al. 2018



Course slides & webapps on CytoLab: <https://cytolab.github.io/>



Jonathan Irish

*Associate Professor
Cell & Developmental Biology
Vanderbilt University*

jonathan.irish@vanderbilt.edu



Cass Mayeda

*Web Applications
Research Assistant
Vanderbilt University*

cass.mayeda@vanderbilt.edu



Josef Spidlen

*Senior Director R&D
BD Biosciences*

josef.spidlen@bd.com



Nicolas Loof

*Senior Application Scientist
Multi-omics Specialist
BD Biosciences*

nicolas.loof@bd.com

Systems Immune Monitoring & Tailoring Workflows

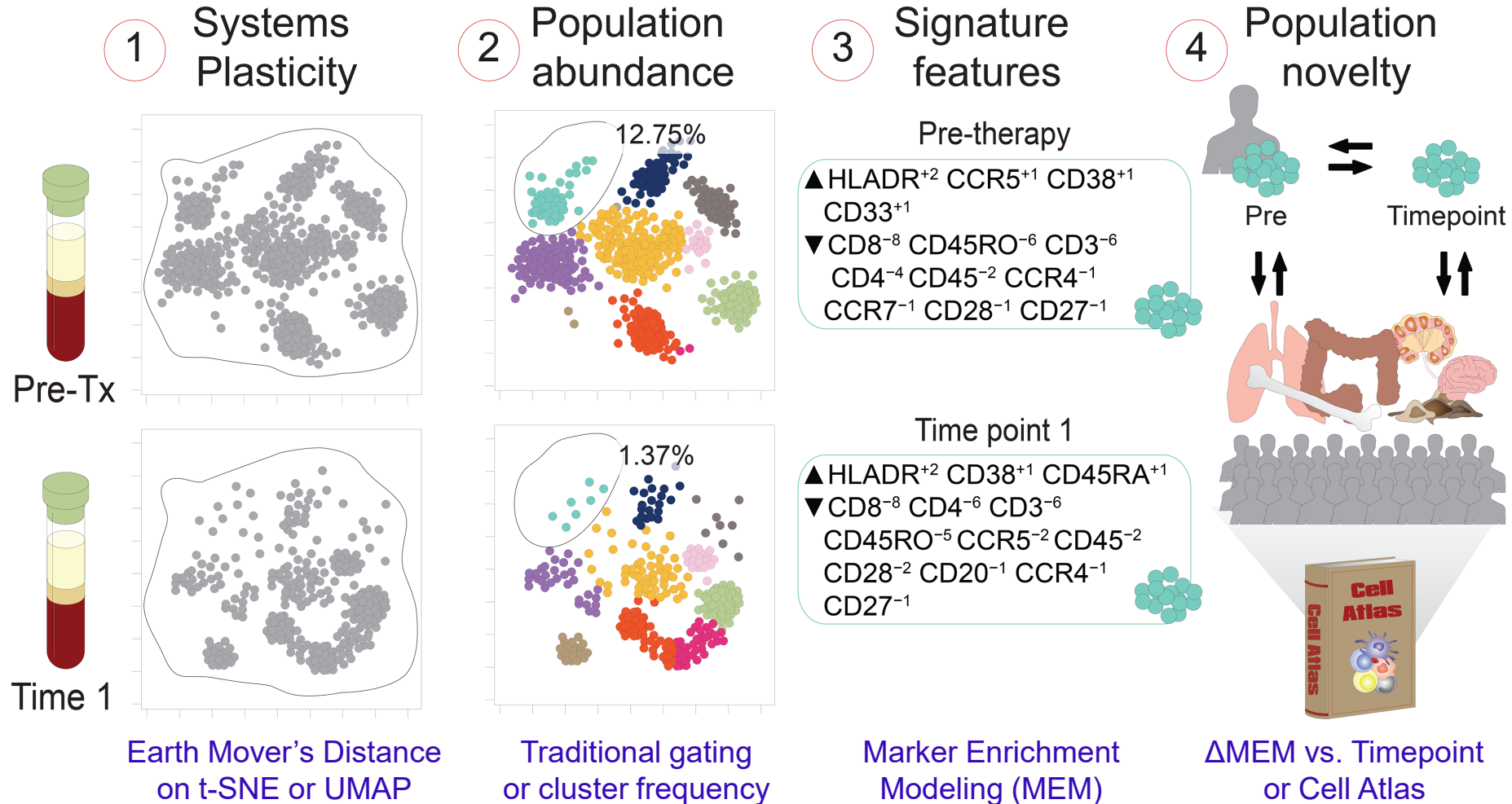
Samples Over Time Reveal Immune System Dynamics

Comparisons with Earth Mover's Distance,
Root Mean Square Deviation (RMSE),
and Change in MEM label (Δ MEM)

Clinical Trial Monitoring: What Do We Need to Know?

Automate Four Key Readouts vs. Clinical Outcomes

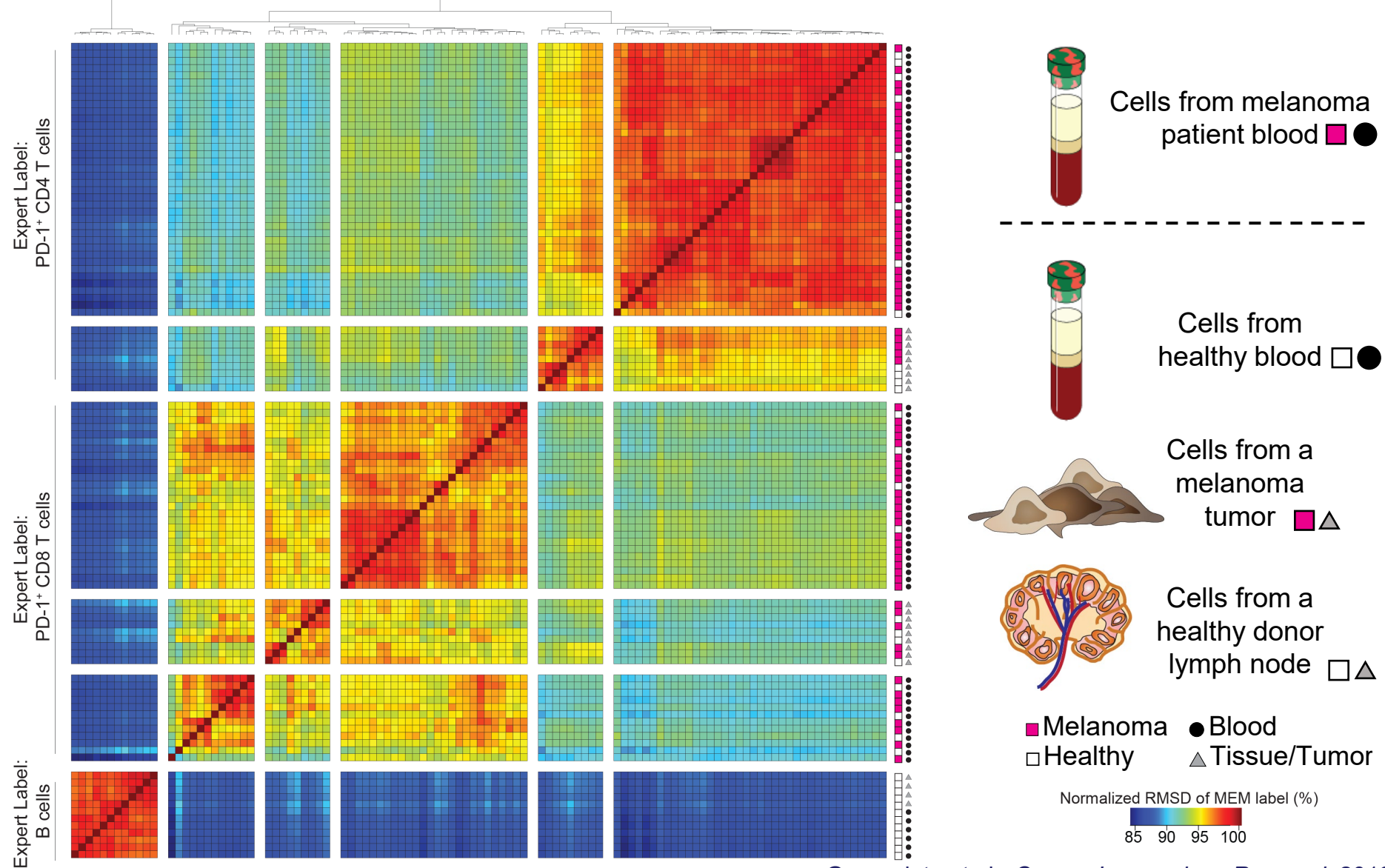
Features of Dynamic Populations



How we quantified

Distinct Phenotypes of PD-1⁺ CD8⁺ T cells in Melanoma Tumors Revealed by Quantitatively Comparing MEM Text Labels

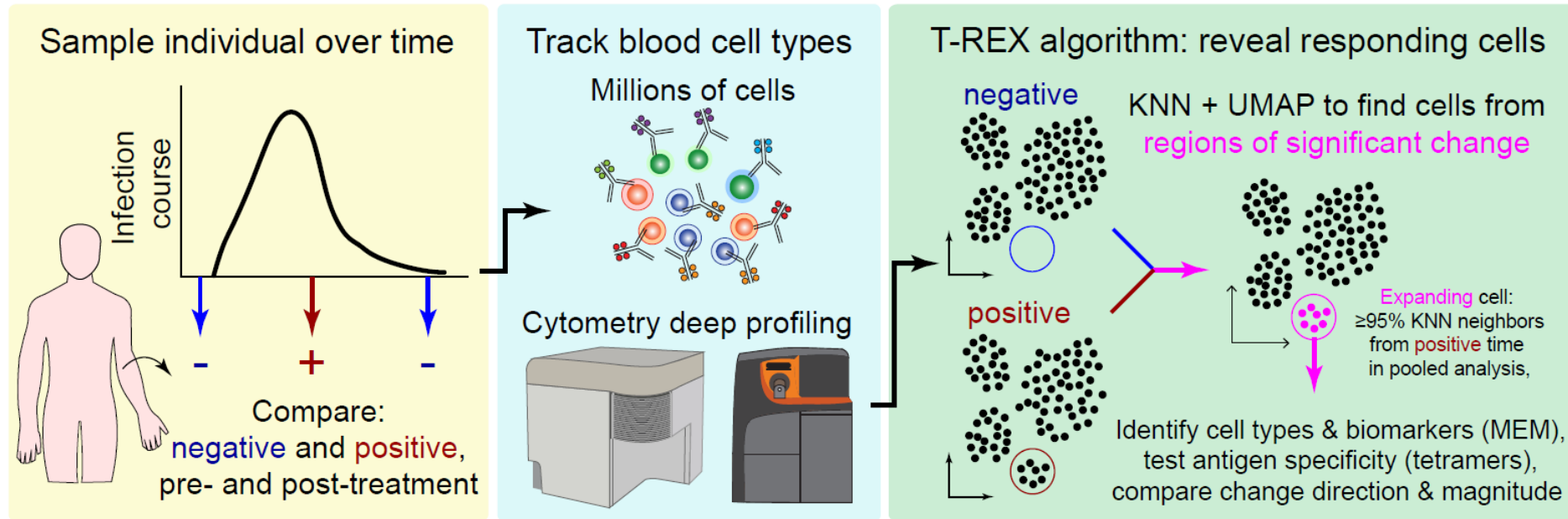
Similarity in MEM label values for PD-1⁺ CD4 or CD8 T cells, B cells (REF: iPSC stem cells)



Data files: <http://flowrepository.org/id/FR-FCM-ZYCC>

Greenplate et al., *Cancer Immunology Research* 2019
Methods: Diggins et al., *Nature Methods* 2017; *Curr Prot Cyt* 2018

RAPID & T-REX Are Both Unsupervised, RAPID: Continuous Outcomes vs. T-REX: Categorical Groups



T-REX (Tracking Responders EXpanding) identifies phenotypic hotspots undergoing great change between conditions (e.g., +/- infection)

Code: <https://github.com/cytolab/t-rex>
Manuscript: <https://elifesciences.org/articles/64653>



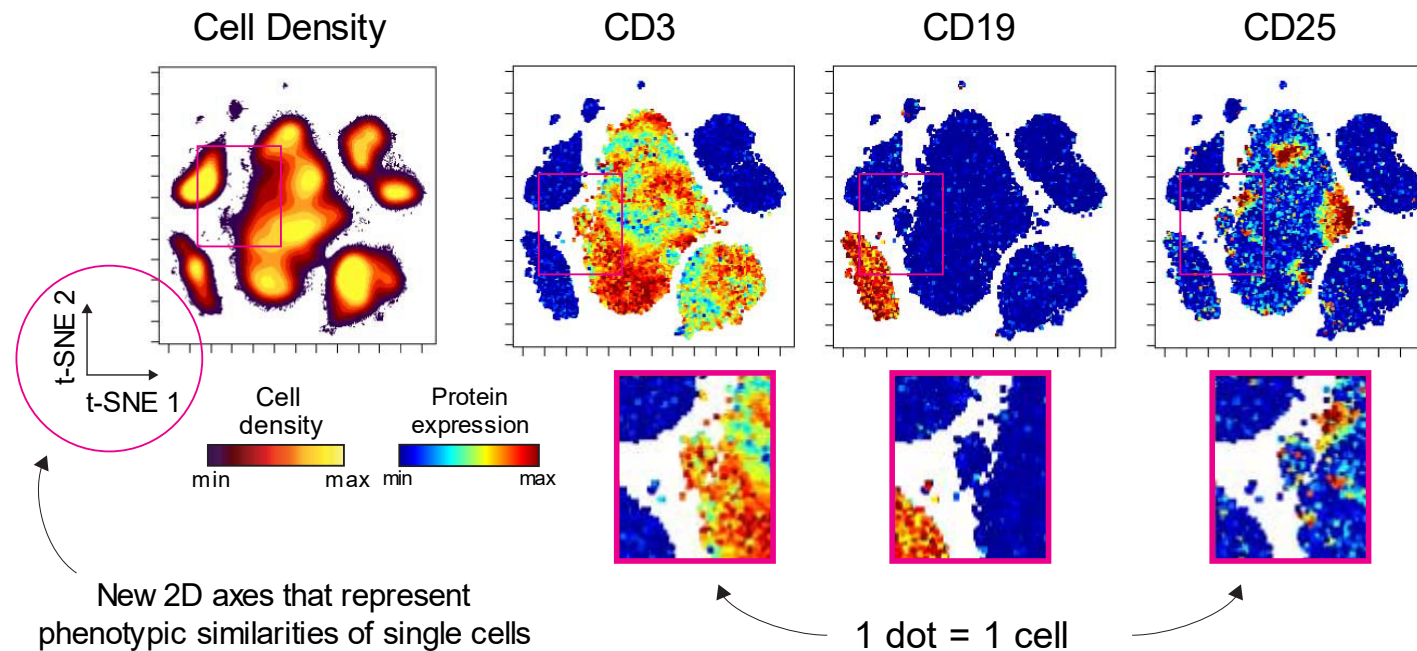
Running the Workflow on PBMC

Dots = 50,000 cells

t-SNE = 25 measured protein features (25D)

Identification of 7 canonical cell types (CD4+ T cells, CD8+ T cells, NK cells, Monocytes, Dendritic Cells, IgM+ B cells, IgM- B cells)

Healthy Peripheral Blood Mononuclear Cells



Let's Analyze PBMC Data!

<https://cytolab.shinyapps.io/PBMC/>

This web app is running R code live.

Data Science Tutorial on Human Blood Cells

Welcome to a data science tutorial on healthy human peripheral blood mononuclear cells (PBMCs). Here you will apply t-SNE, FlowSOM, and MEM algorithms on the data, and learn how changing different settings impacts your results.

The dataset is from [Diggins et al., Nature Methods 2017](#), and contains around 50,000 cells each measured for 25 different proteins. Viewing the first few cells in spreadsheet form, the data looks like the following:

	CD19	CD117	CD11b	CD4	CD8	CD20	CD34	CD61	CD123	CD45RA	CD45	CD10	CD33	CD11c	CD14	CD69	CD15	CD16	CD44	CD38	CD25	CD3	IgM	HLA-DR	CD56
cell 1	-3	-6	11	132	12	-8	-8	-7	-5	101	284	-2	-8	2	-06	-8	.7	-6	46	-1	10	71	-9	10	-.09
cell 2	-3	-4	-6	204	4.6	-2	-6	-.03	-8	-1	400	-7	-7	-6	1	-1	-8	-2	222	10	3	99	-5	-9	-.04
cell 3	1.4	-5	-5	145	2.4	-2	-4	-5	-6	25	360	-6	5	-.08	-8	-3	-2	-4	320	24	18	50	-6	-8	5

For this tutorial, we've taken a random sample of 5,000 cells from the 50,000 to run analyses on. If you'd like a larger or smaller sample size, you have the option to change that in the following menu. Alternatively if you'd like to reset your session, you can use the clear session button.

SAMPLE SIZE

1) Build a map with t-SNE

In the first exercise we select protein features and use the t-SNE algorithm to build a map of cell phenotypes. t-SNE or t-distributed stochastic neighbor embedding, looks at all the cell features and over several iterations embeds cells with similar expression patterns close to each other. The result is a 2 dimensional map of phenotypic similarity, simplified from 25 dimensions.

With the default settings we see the 50,000 cells arranged in major islands corresponding to phenotypically distinct immune cell types, namely CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes.

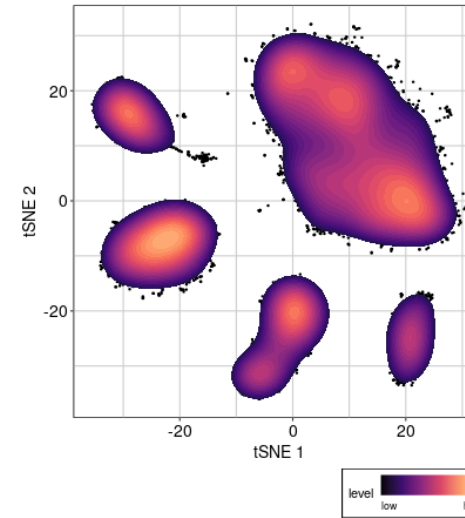
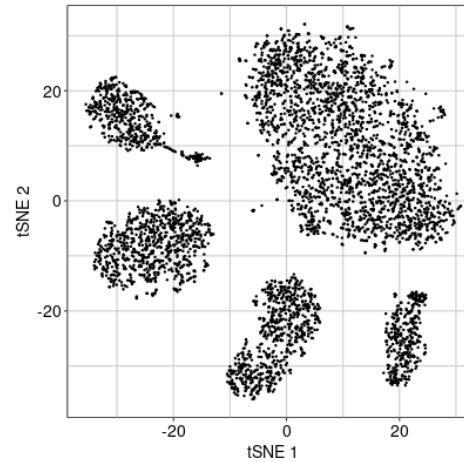
SEED
43

PERPLEXITY
0 30 120

CHANNELS

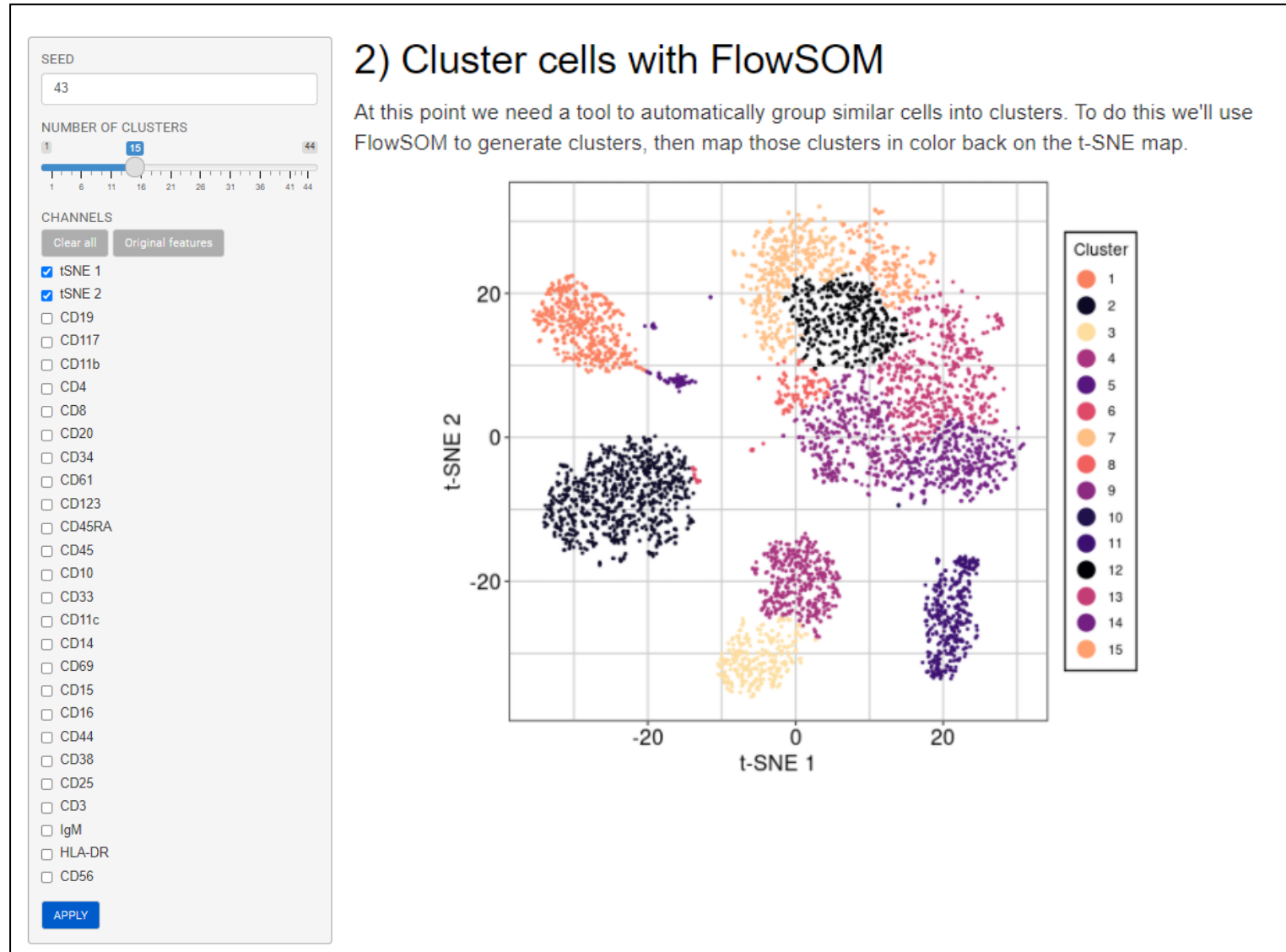
- CD19
- CD117
- CD11b
- CD4
- CD8
- CD20
- CD34
- CD61
- CD123
- CD45RA
- CD45
- CD10
- CD33
- CD11c
- CD14
- CD69
- CD15
- CD16
- CD44
- CD38
- CD25
- CD3
- IgM
- HLA-DR
- CD56

APPLY



2) Cluster cells with FlowSOM

At this point we need a tool to automatically group similar cells into clusters. To do this we'll use FlowSOM to generate clusters, then map those clusters in color back on the t-SNE map.



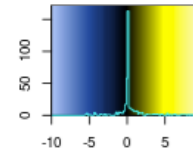
Referenceless?

APPLY

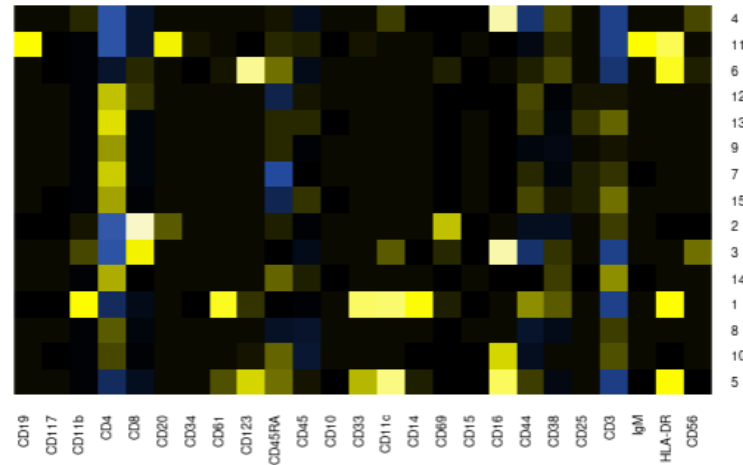
3) Describe clusters with MEM

Once we've grouped cells into clusters, how can we identify what kind of cells are in each cluster? You can look up marker expression values in the "spreadsheet" view of the data, or run an algorithm like MEM which calculates features that are enriched within the various groupings. A MEM label often provides enough information to infer identity if it is a known cell type, or guess at its biological significance. A methods paper explaining the MEM algorithm and going through examples is available in [Current Protocols in Cytometry](#).

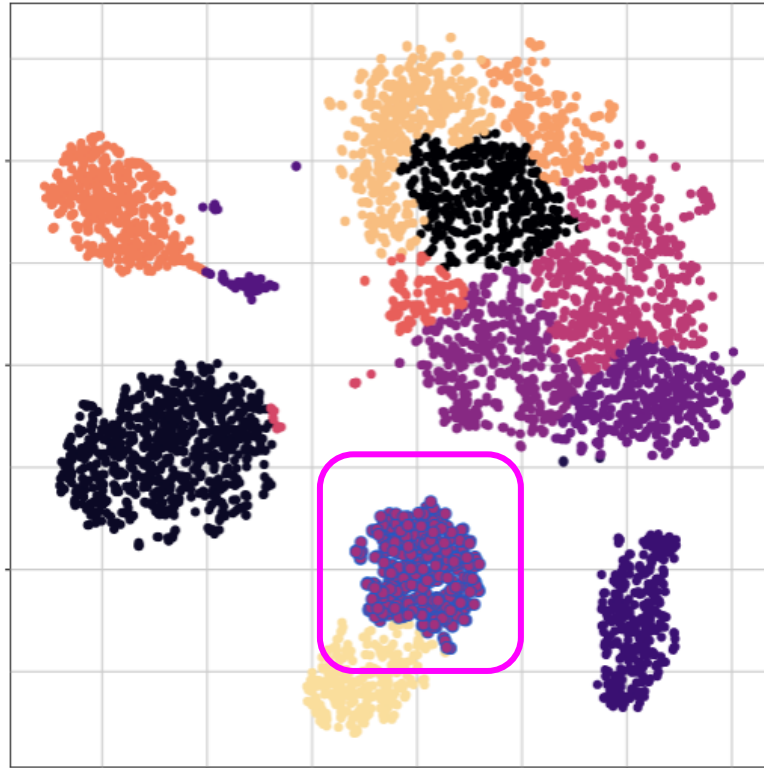
Here MEM outputs a heatmap of the relative expression of each protein organized by cluster.



MEM Heatmap



Number of cells: 5000

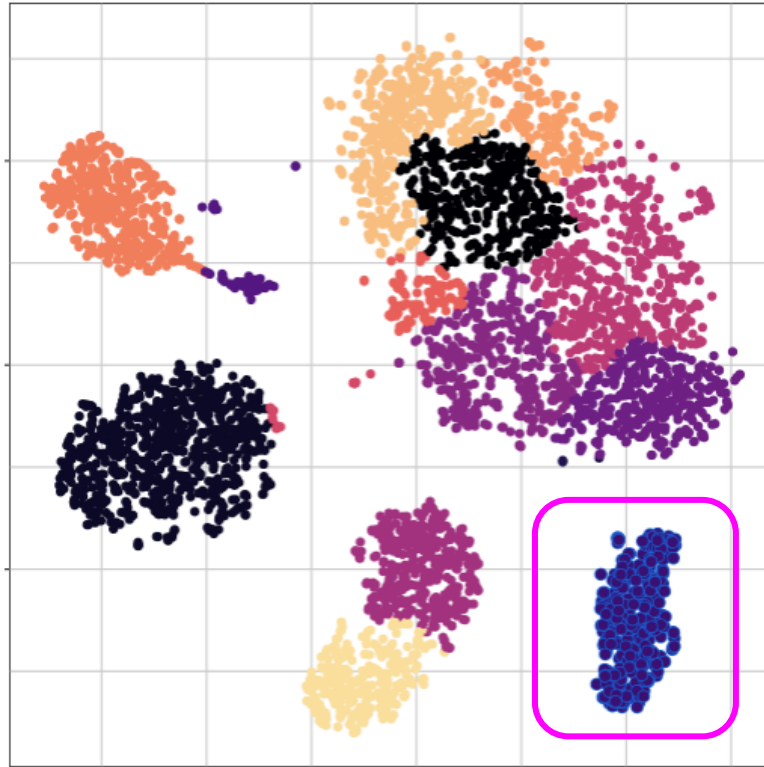


You can also explore a particular cluster by clicking on it in the plot to the left, and reading the MEM label that's generated below.

Cluster: 4 8 % of sample

- ▲ CD16 ⁺⁹ CD11b ⁺¹ CD11c ⁺¹
CD38 ⁺¹ CD56 ⁺¹
- ▼ CD4 ⁻⁵ CD3 ⁻⁴ CD44 ⁻³ CD8 ⁻¹
CD45 ⁻¹

Number of cells: 5000



You can also explore a particular cluster by clicking on it in the plot to the left, and reading the MEM label that's generated below.

Cluster: 11 8 % of sample

- ▲ HLA-DR ⁺⁷ CD19 ⁺⁵ IgM ⁺⁵ CD20 ⁺⁴
CD45RA ⁺¹ CD38 ⁺¹
- ▼ CD4 ⁻⁵ CD3 ⁻⁴ CD8 ⁻¹ CD44 ⁻¹

1) Build a map with t-SNE

In the first exercise we select protein features and use the t-SNE algorithm to build a map of cell phenotypes. t-SNE or t-distributed stochastic neighbor embedding, looks at all the cell features and over several iterations embeds cells with similar expression patterns close to each other. The result is a 2 dimensional map of phenotypic similarity, simplified from 25 dimensions.

With the default settings we see the 50,000 cells arranged in major islands corresponding to phenotypically distinct immune cell types, namely CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes.

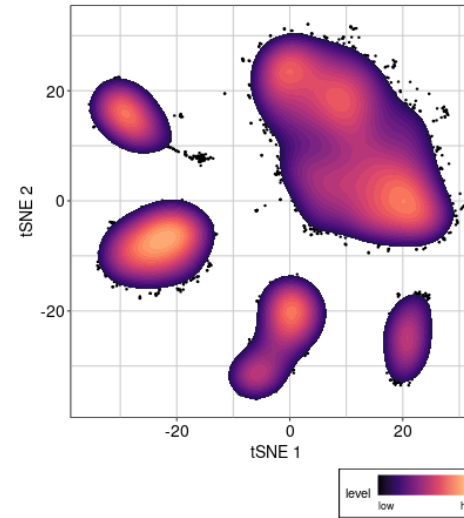
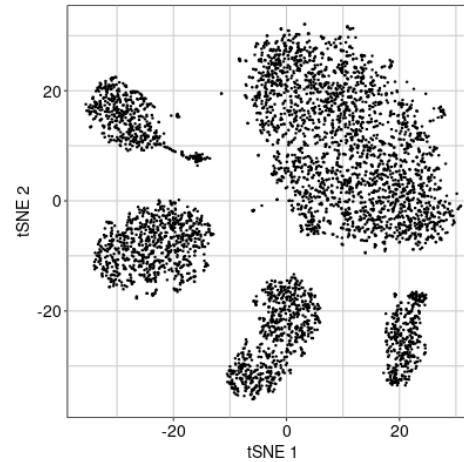
SEED
43

PERPLEXITY
0 30 120

CHANNELS

- CD19
- CD117
- CD11b
- CD4
- CD8
- CD20
- CD34
- CD61
- CD123
- CD45RA
- CD45
- CD10
- CD33
- CD11c
- CD14
- CD69
- CD15
- CD16
- CD44
- CD38
- CD25
- CD3
- IgM
- HLA-DR
- CD56

APPLY



SEED
84

PERPLEXITY
0 30 120

CHANNELS

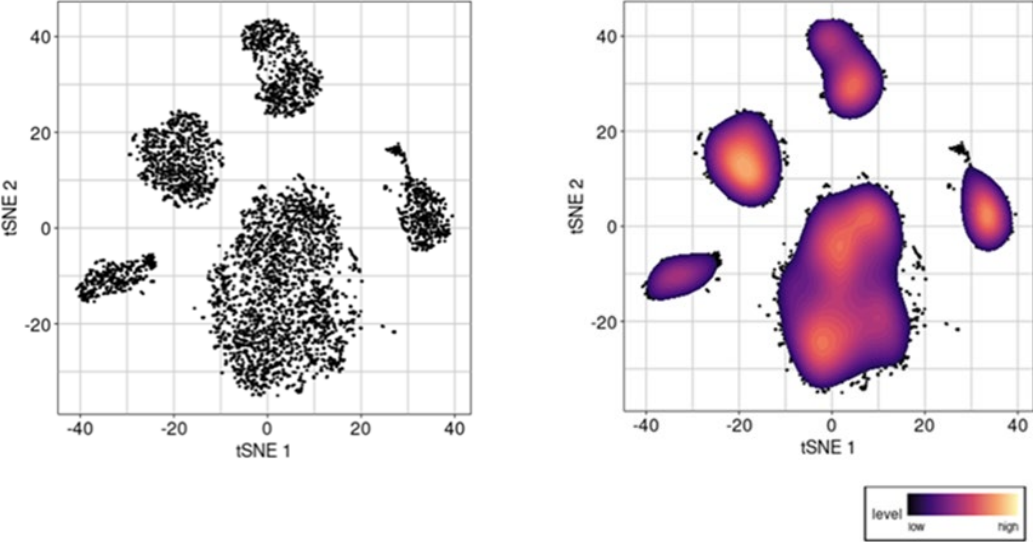
- CD19
- CD117
- CD11b
- CD4
- CD8
- CD20
- CD34
- CD61
- CD123
- CD45RA
- CD45
- CD10
- CD33
- CD11c
- CD14
- CD69
- CD15
- CD16
- CD44
- CD38
- CD25
- CD3
- IgM
- HLA-DR
- CD56

APPLY

1) Build a map with t-SNE

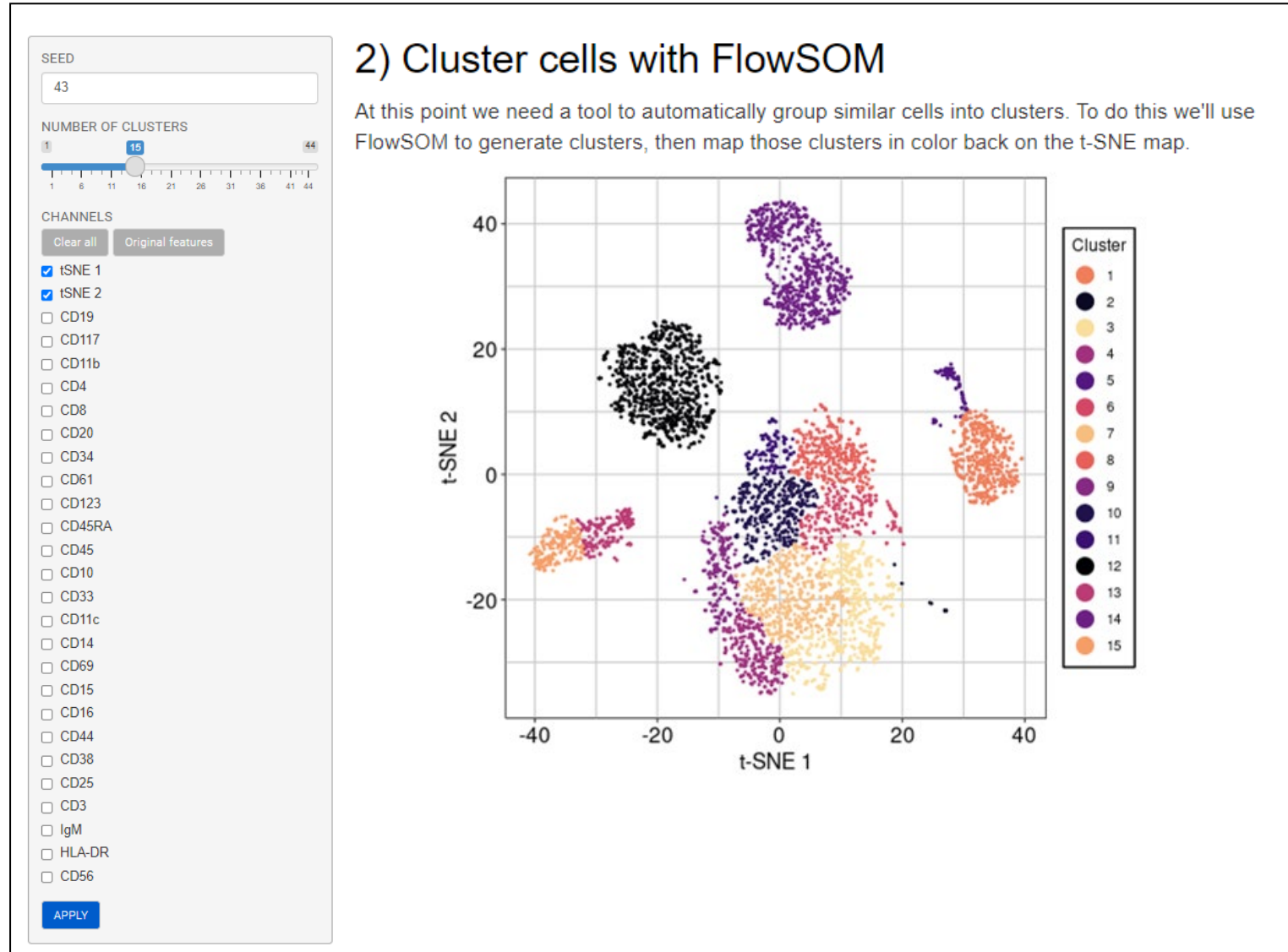
In the first exercise we select protein features and use the t-SNE algorithm to build a map of cell phenotypes. t-SNE or t-distributed stochastic neighbor embedding, looks at all the cell features and over several iterations embeds cells with similar expression patterns close to each other. The result is a 2 dimensional map of phenotypic similarity, simplified from 25 dimensions.

With the default settings we see the 50,000 cells arranged in major islands corresponding to phenotypically distinct immune cell types, namely CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes.



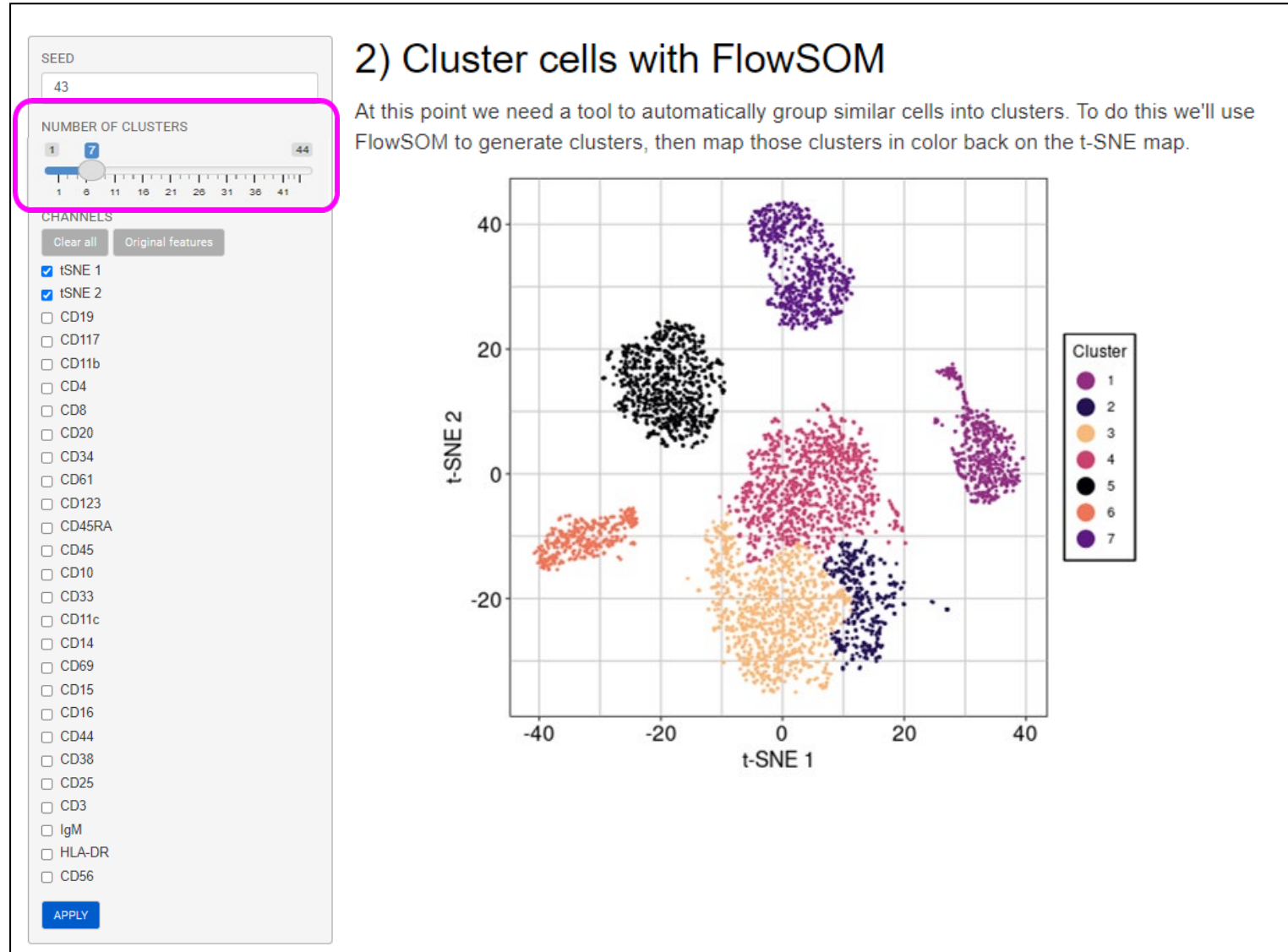
2) Cluster cells with FlowSOM

At this point we need a tool to automatically group similar cells into clusters. To do this we'll use FlowSOM to generate clusters, then map those clusters in color back on the t-SNE map.

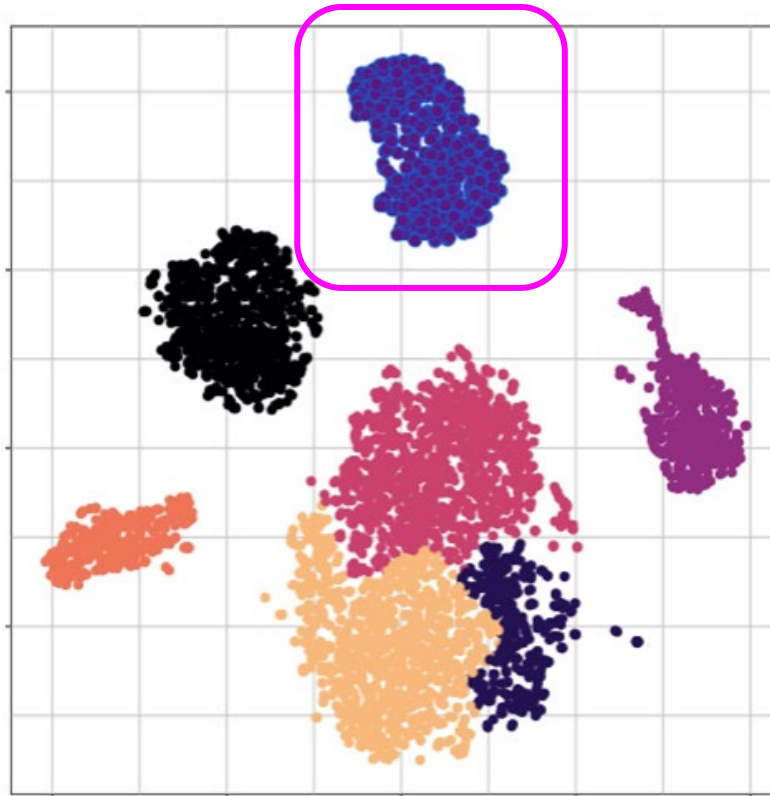


2) Cluster cells with FlowSOM

At this point we need a tool to automatically group similar cells into clusters. To do this we'll use FlowSOM to generate clusters, then map those clusters in color back on the t-SNE map.



Number of cells: 5000



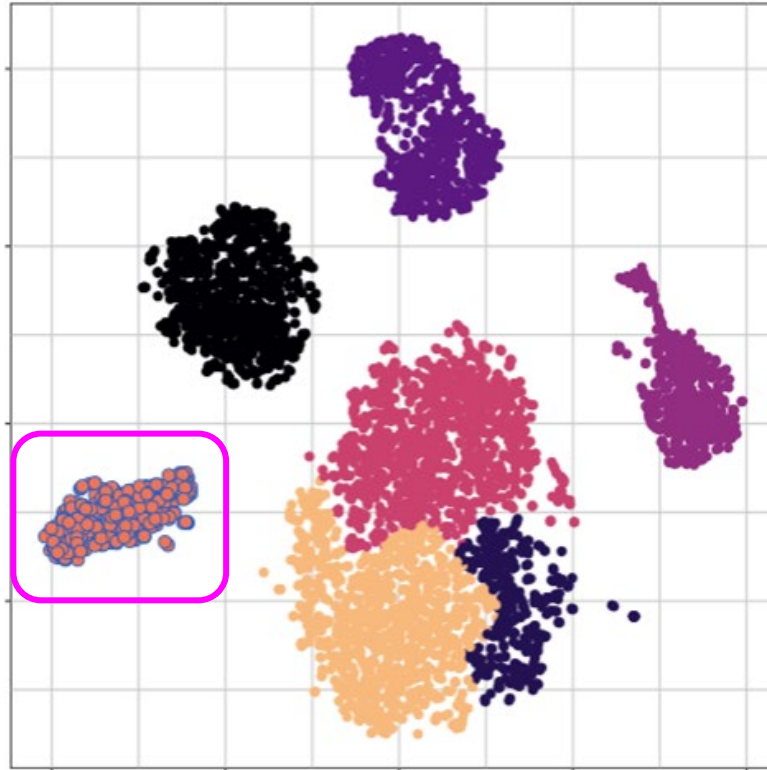
You can also explore a particular cluster by clicking on it in the plot to the left, and reading the MEM label that's generated below.

Cluster: 7 14 % of sample

▲ CD16 ⁺⁹ CD56 ⁺² CD11b ⁺¹
CD11c ⁺¹ CD38 ⁺¹

▼ CD4 ⁻⁶ CD44 ⁻⁴ CD3 ⁻⁴ CD45 ⁻¹

Number of cells: 5000



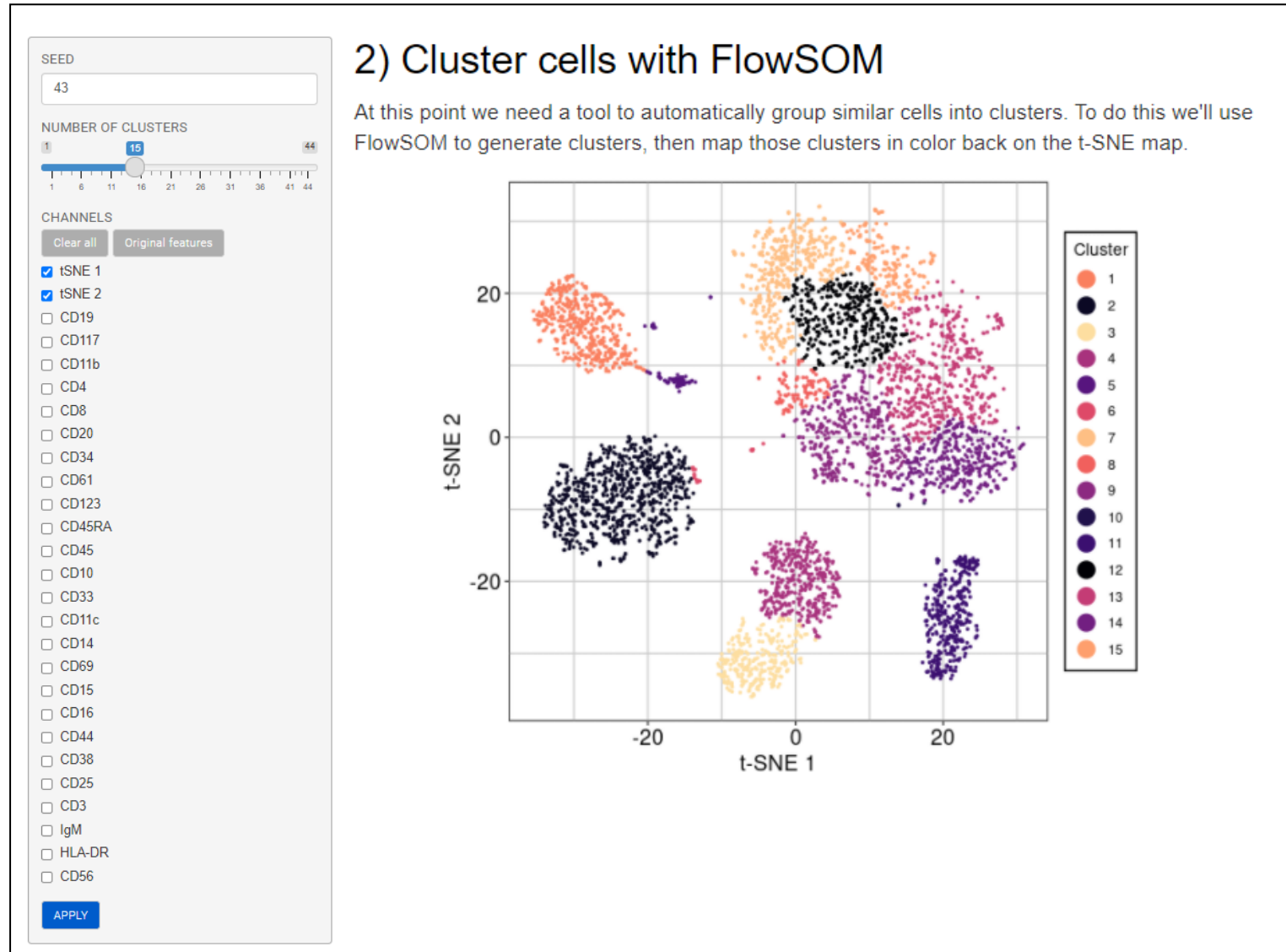
You can also explore a particular cluster by clicking on it in the plot to the left, and reading the MEM label that's generated below.

Cluster: 6 8 % of sample

▲ HLA-DR ⁺⁷ CD19 ⁺⁵ IgM ⁺⁵
CD20 ⁺⁴ CD45RA ⁺¹ CD38 ⁺¹
▼ CD4 ⁻⁵ CD3 ⁻⁴ CD8 ⁻¹ CD44 ⁻¹

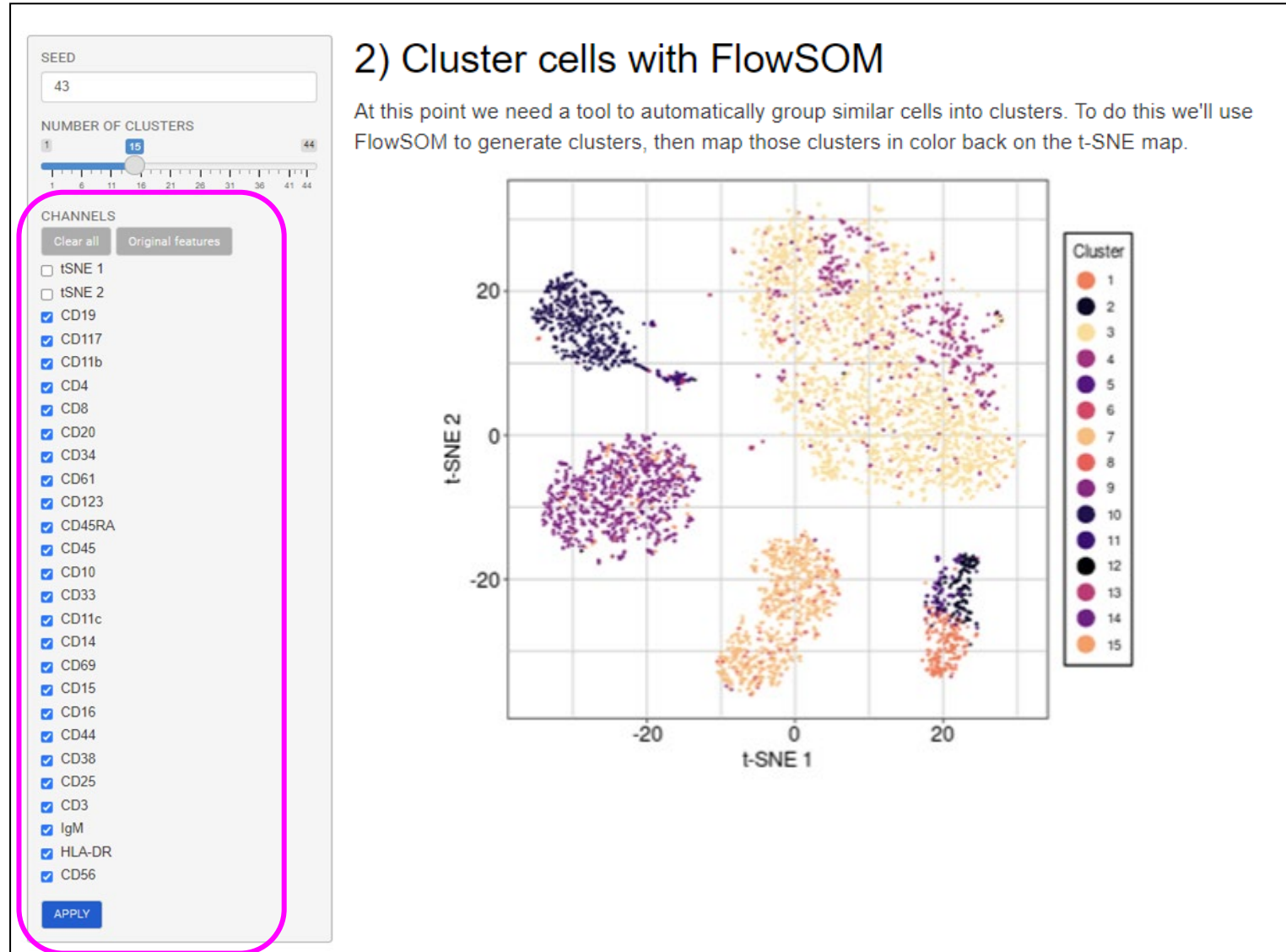
2) Cluster cells with FlowSOM

At this point we need a tool to automatically group similar cells into clusters. To do this we'll use FlowSOM to generate clusters, then map those clusters in color back on the t-SNE map.

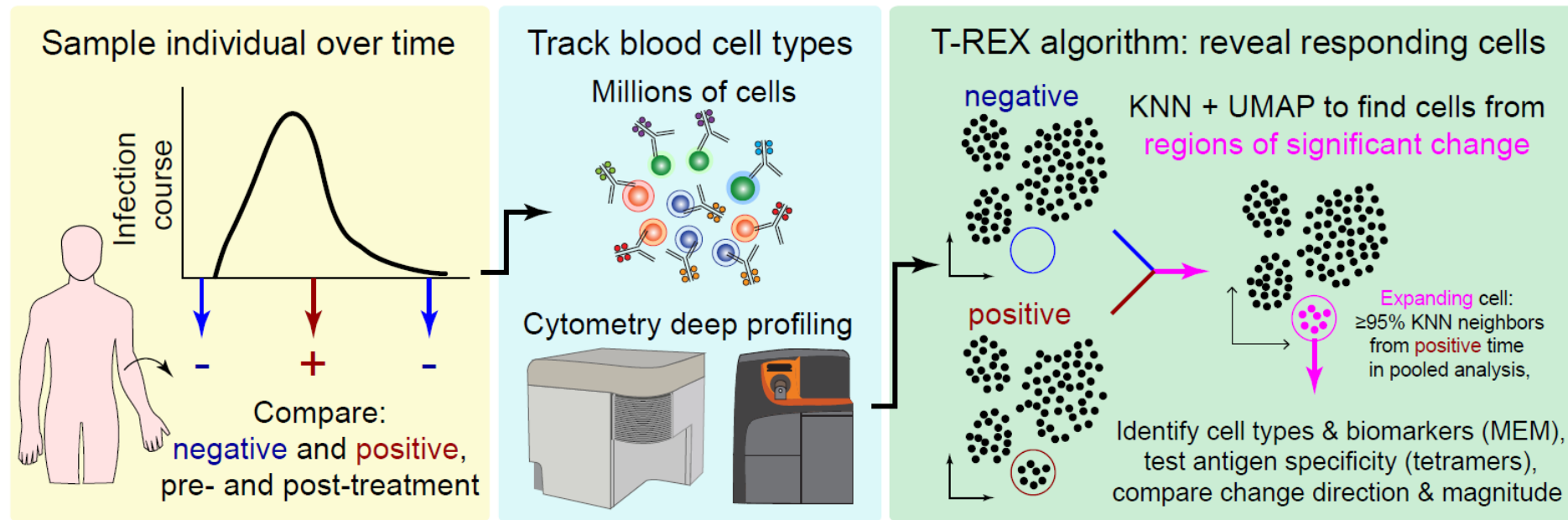


2) Cluster cells with FlowSOM

At this point we need a tool to automatically group similar cells into clusters. To do this we'll use FlowSOM to generate clusters, then map those clusters in color back on the t-SNE map.



T-REX: Compare Two Samples to Identify Things Enriched in Either One; e.g., Reveal Rare, Virus-Specific Immune Cells



New algorithm: T-REX (Tracking Responders EXpanding)

Code: <https://github.com/cytolab/t-rex>
Manuscript: <https://elifesciences.org/articles/64653>



Data Science Workflow Using T-REX

Revealing very rare cells or cells changing significantly

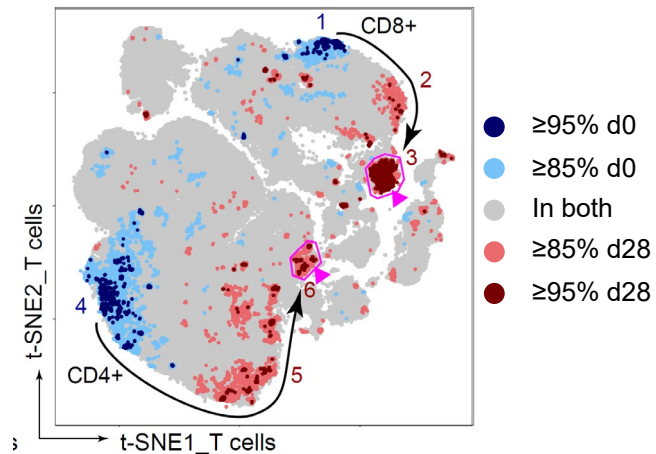
Dimensionality reduction

- t-SNE or UMAP

Clustering

- KNN on all cells
- DBSCAN on cells in regions changing significantly

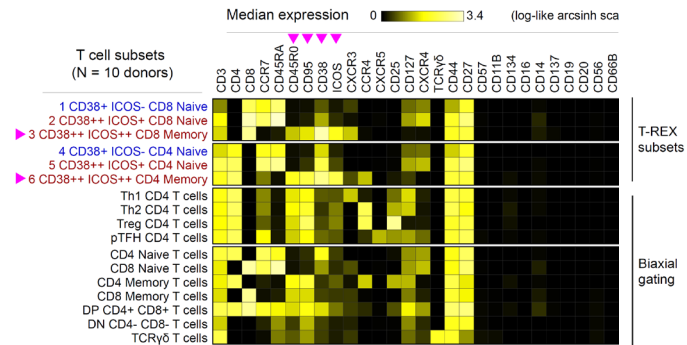
T cells, N = 10 donors,
T-REX Day 0 vs. Day 28



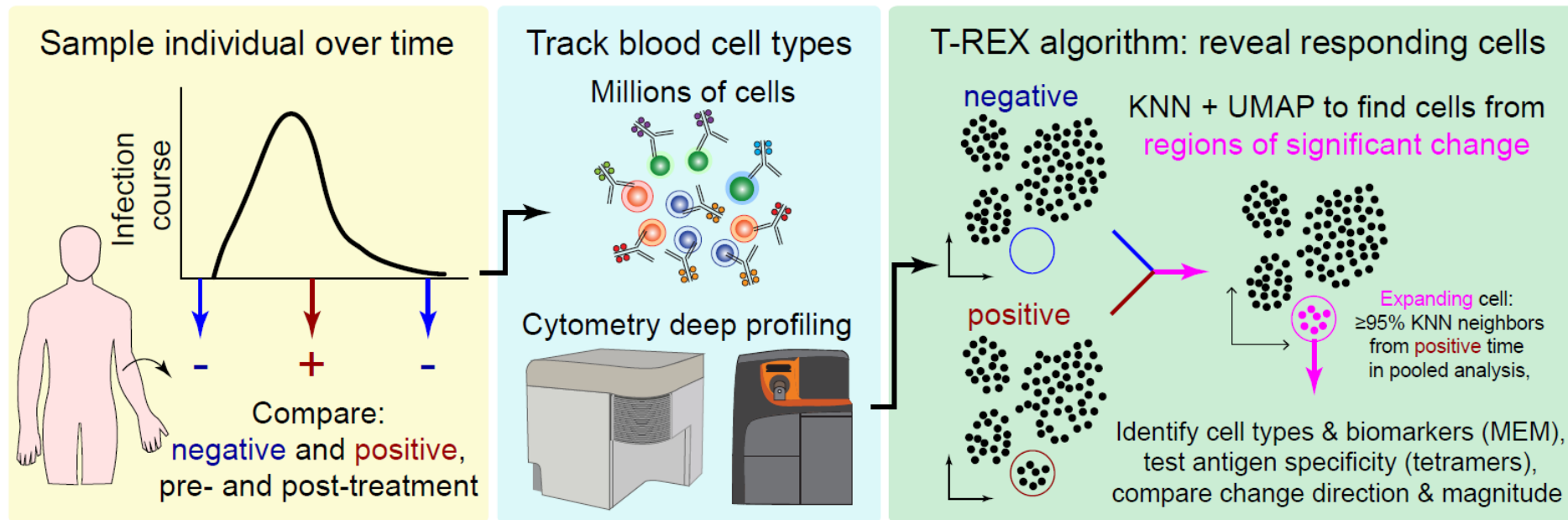
Characterizing cells expanding or contracting

Learn cell identity

- MEM

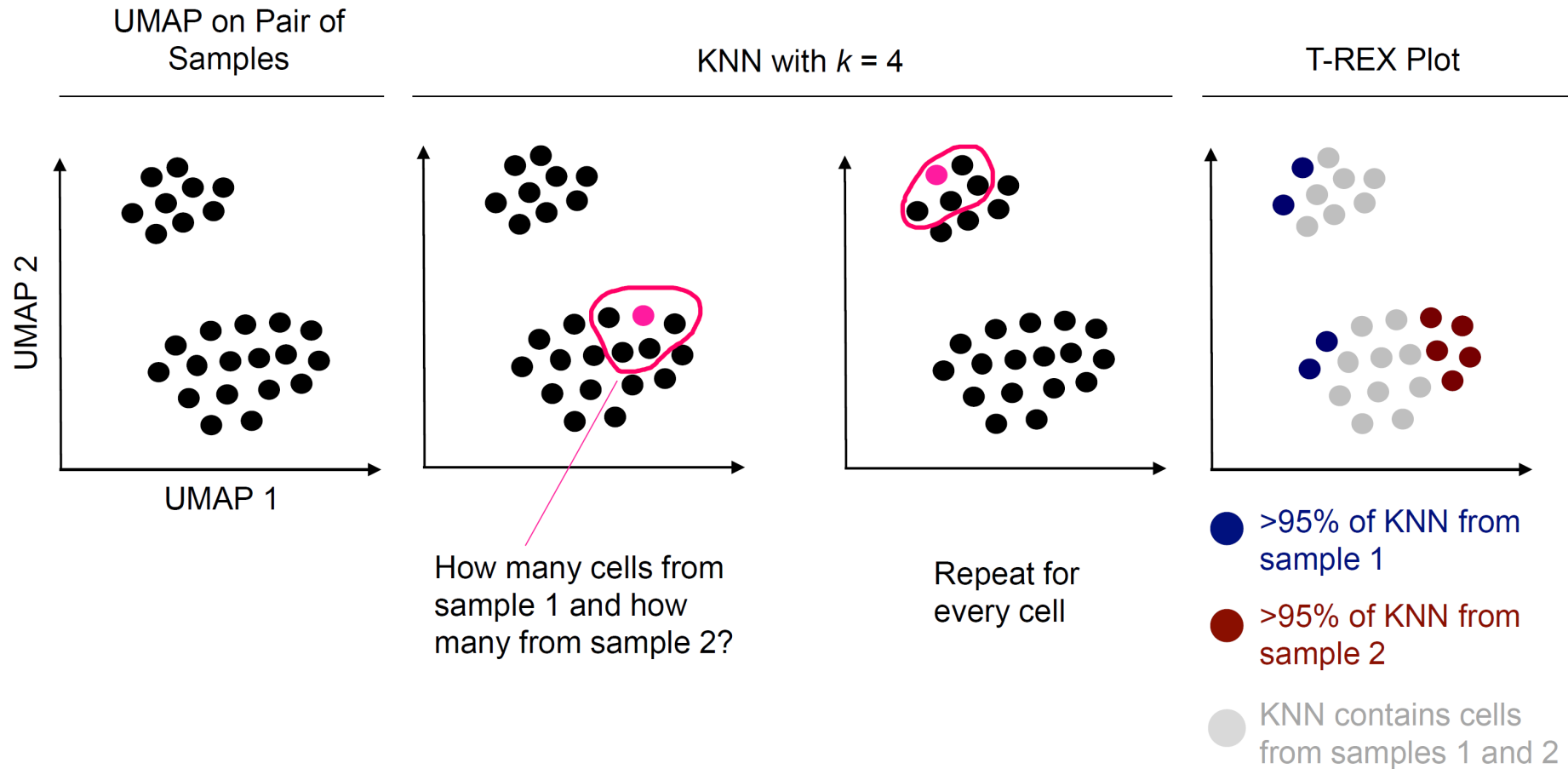


Key Ideas & Findings in Today's Talk

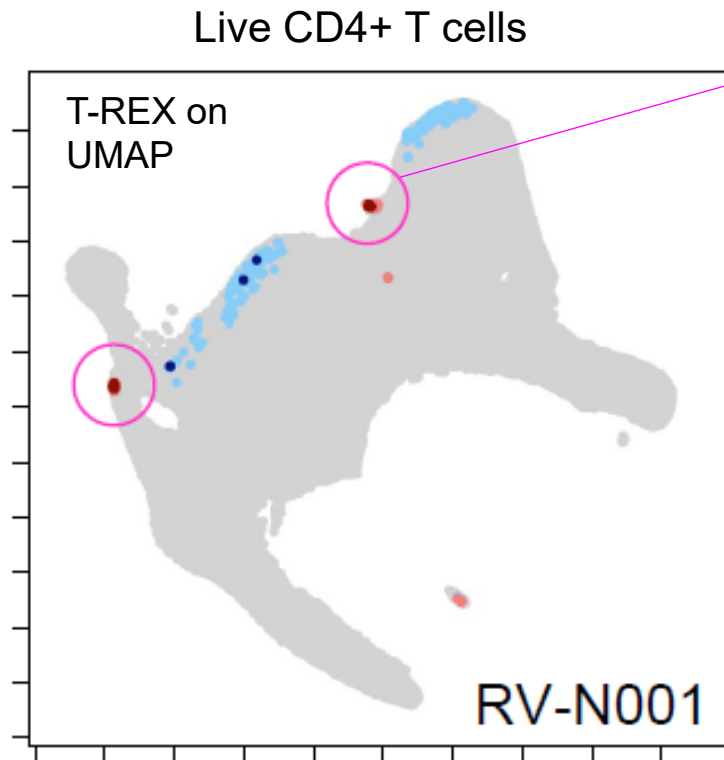


- Idea 1: T-REX automatically reveals virus-specific T cells in rhinovirus & SARS-CoV-2 vaccine response (without the need for tetramers, sorting, or sequencing)
- Idea 2: Approach focuses on extreme change & can summarize disease, therapy, or perturbation response (direction & magnitude of change; rhinovirus, COVID-19, cancer therapy, compound screening)
- Finding: Mass cytometry + T-REX characterized SARS-CoV-2 vaccine-induced memory CD4 and CD8 T cells (phenotype: CD38⁺⁺ ICOS⁺⁺ CD45R0⁺ PD-1⁺ Ki-67⁺ CXCR5⁻)
- Finding: Phenotype of SARS-CoV-2 vaccine responding T cells closely matched rhinovirus-specific T cells

T-REX Algorithm Uses K-Nearest Neighbors (KNN) to Characterize Each Cell's Immediate Phenotypic Neighborhood



T-REX: Tracking Responders EXpanding, Every Cell Is Characterized in a Search for Hotspots of Change



MHCII tetramers marking rhinovirus specific CD4 T cells were not used to make the UMAP, instead used to show: Change hotspots were enriched for virus-specific T cells

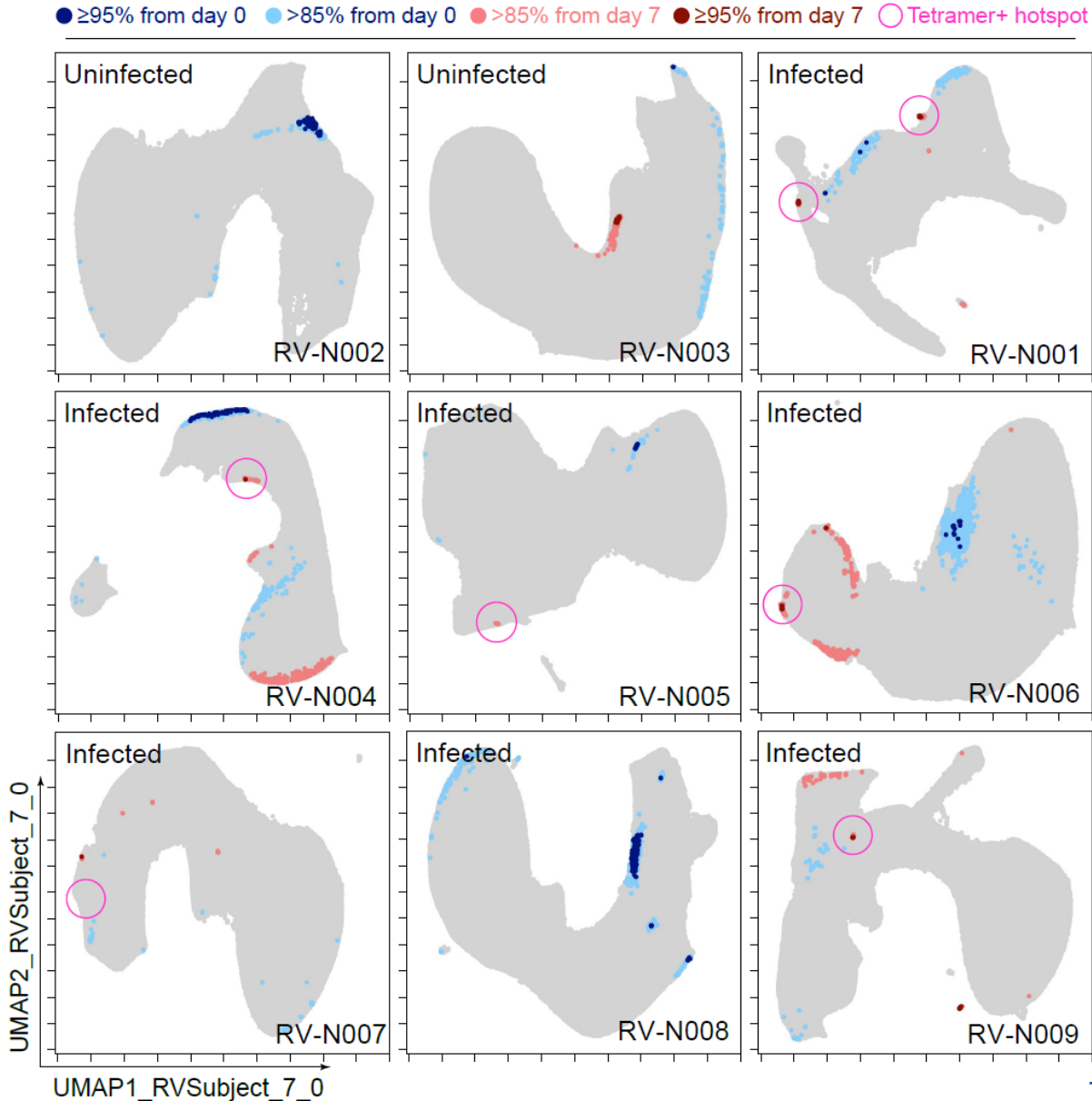
Color: cells in that phenotypic neighborhood are mostly from one sample

Dark red = cells mostly from day 7 (expanding) 

- >95% from d0
- >95% from d7
- >75% from d0
- >75% from d7
- >5% of neighbors tetramer+

CD4 T cells, Day 0 vs. Day 7,
individual infected with rhinovirus (RV-N001)
no cell enrichment, Aurora data, $\sim 3 \times 10^6$ cells

In Analysis of a Rhinovirus Challenge Cohort, T-REX Revealed Virus-Specific Cell Phenotypes



CD4 T cells, Day 0 vs. Day 7,
individuals infected with rhinovirus
no cell enrichment, Cytex Aurora data

In 5 of 7 infected individuals, **expansion hotspots** were enriched for **virus-specific cells**



The phenotype of rhinovirus-specific memory CD4+ T cells calculated by MEM:
CCR5+ ICOS+ CD38+ PD-1+ CXCR5-

Gating based on this MEM phenotype =>
enriched for tetramer+ cells
(without gating on tetramers):

**Indicated we could sort cells (FACS)
based on T-REX MEM labels**

T-REX revealed virus-specific T cells without tetramers



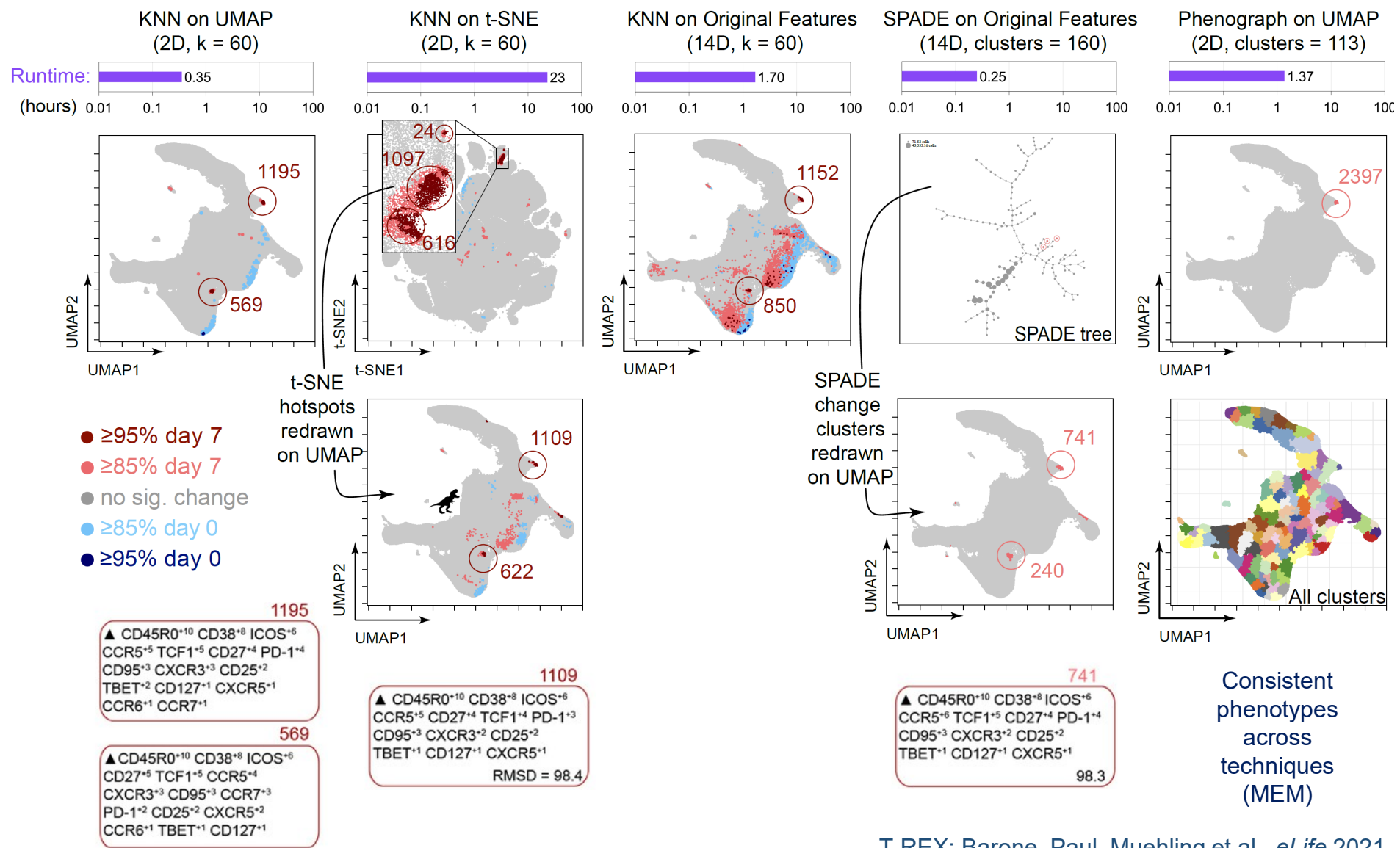
Would this approach work with other clustering algorithms?

Is it 'OK' to do KNN on UMAP axes as parameters?

(Perhaps: all embeddings are wrong, but some are useful...)

T-REX Worked with Other Algorithms to Identify Comparable Cells, But KNN on UMAP or t-SNE Outperformed KNN on Original Features

Methods that identified at least one >85% change cluster (T-REX hotspot of change)



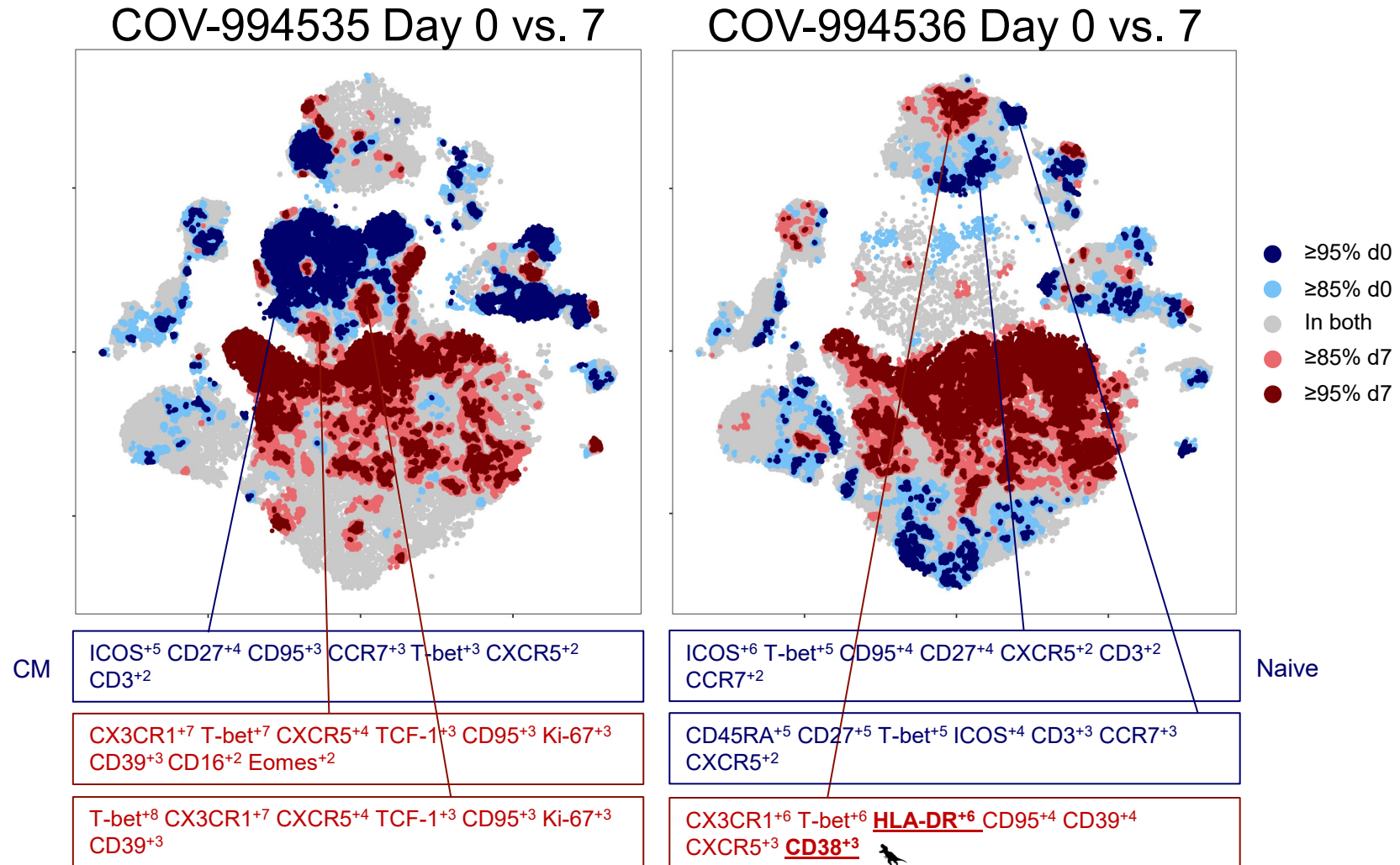
T-REX revealed virus-specific T cells without tetramers

Also found to work for:

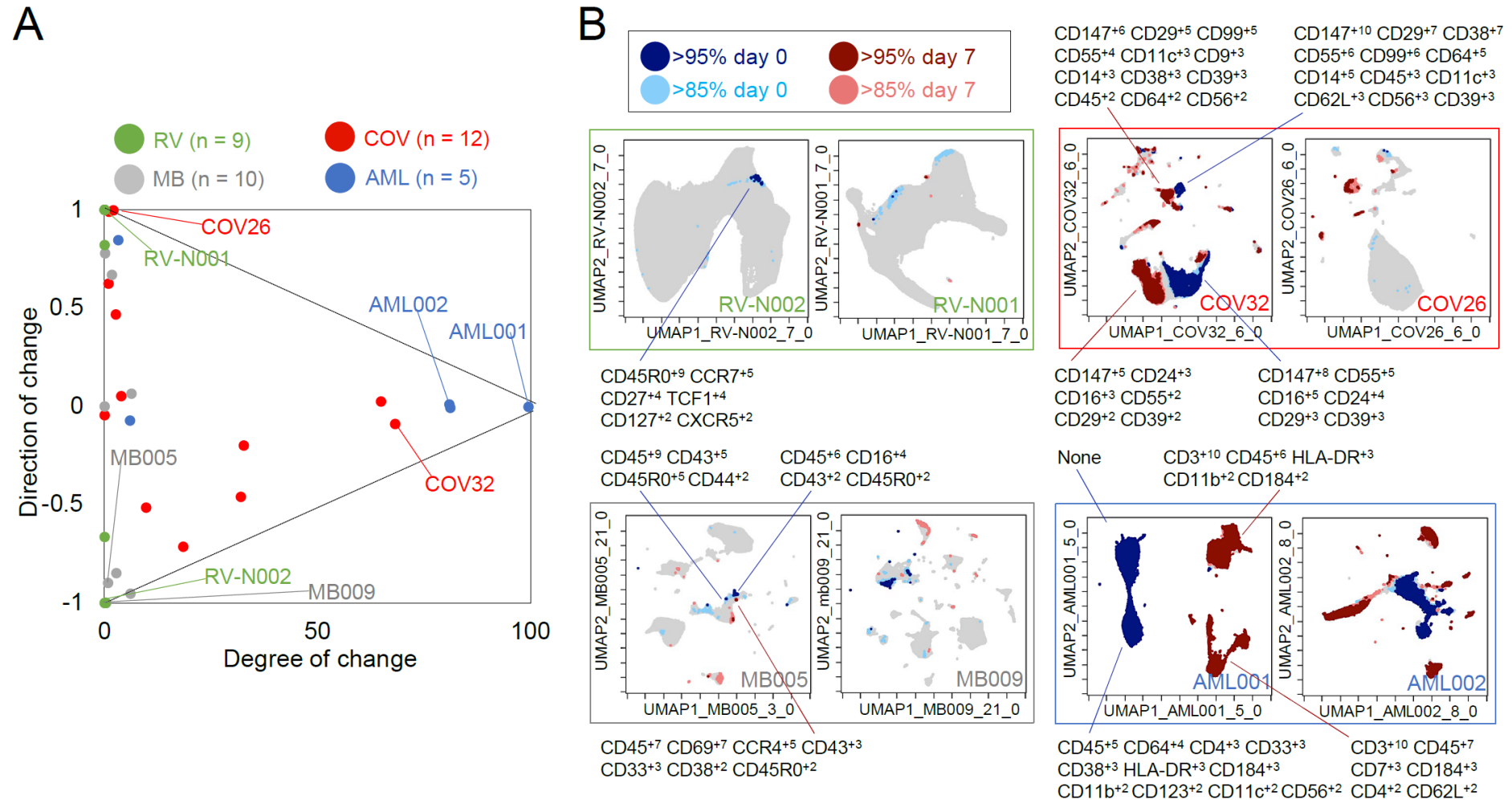
- a range of k-values (k = 60 was optimal)
- post-infection as the comparison point to day 7
- data from a range of cytometers, studies, and labs
- COVID-19, melanoma immunotherapy response, AML

(see the manuscript for this & more!) 

Massive Immune Change, Common Shifts in Expanding Cell Subsets Observed Between Day 0 and Day 7 in COVID-19



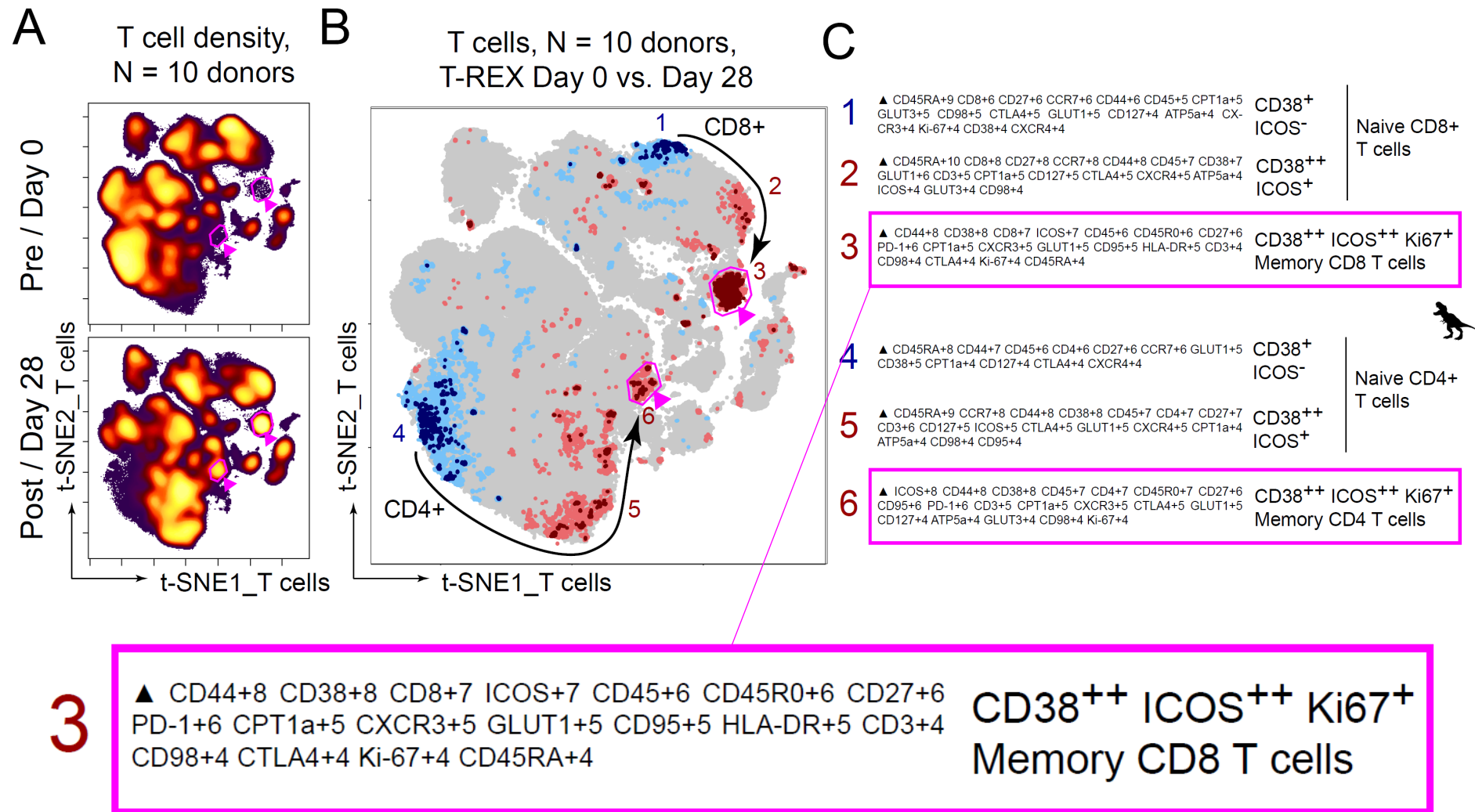
Half of COVID-19 Patients Displayed Immune Changes Comparable to AML Patients with a Complete Response to Chemotherapy



T-REX revealed virus-specific T cells without tetramers
& characterized massive immune changes in COVID-19

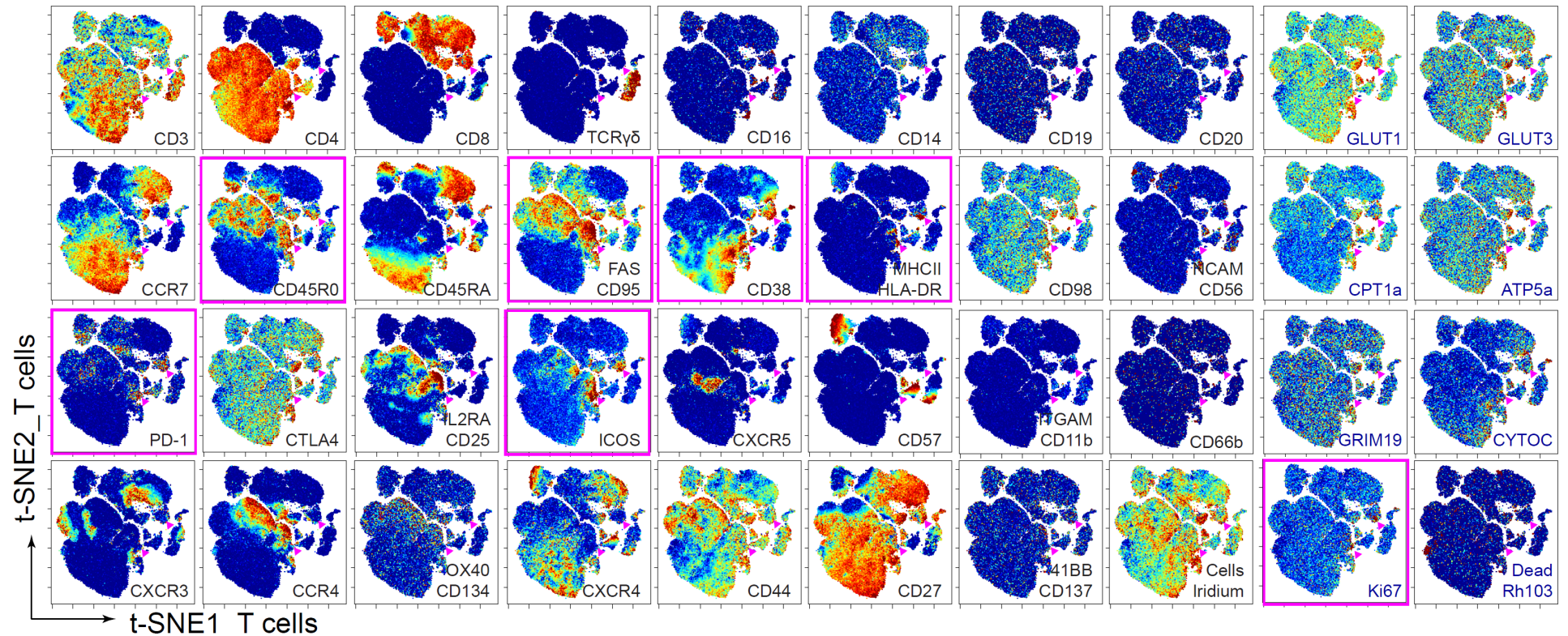
Would it also work to characterize SARS-CoV-2 vaccine response? 

T-REX Reveals Memory CD4 & CD8 T Cell Phenotypes Expanding following BNT162b2 SARS-CoV-2 RNA Vaccine

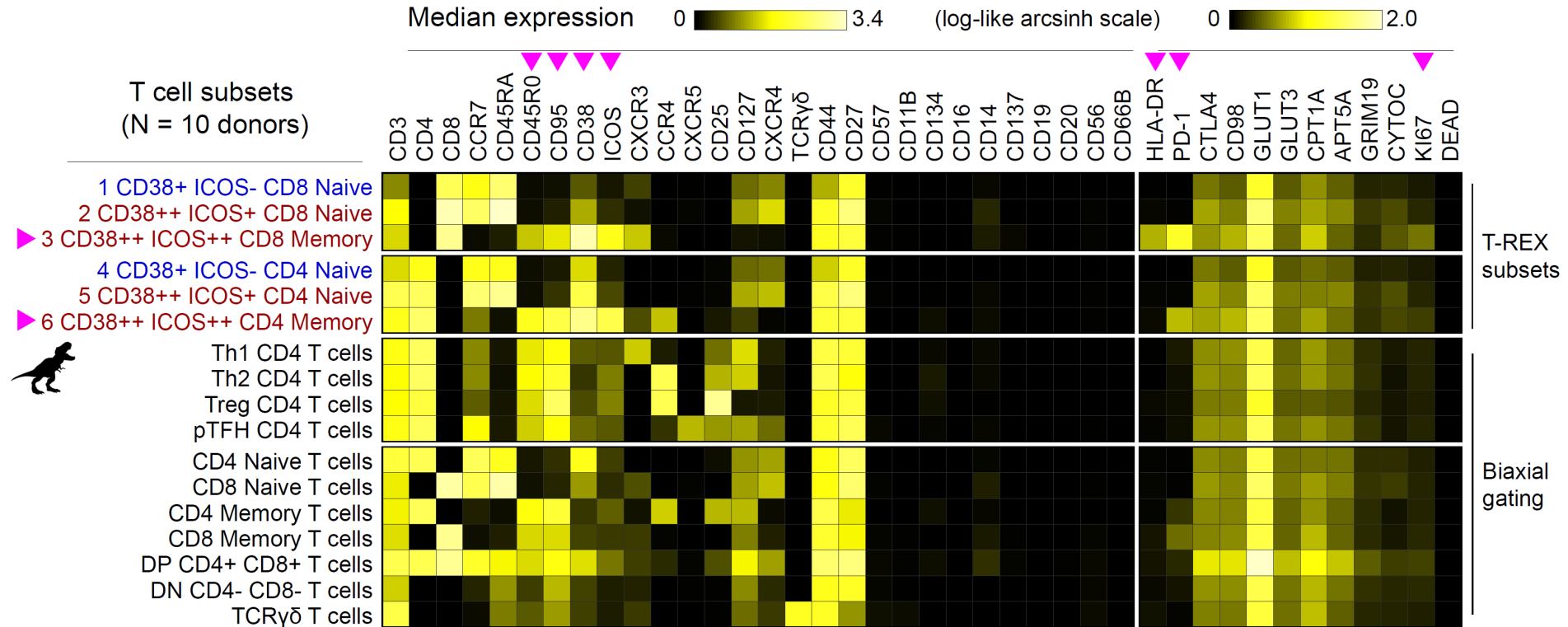


Mass Cytometry Phenotyping of ICOS+ CD38+ PD-1+ Ki-67+ CXCR5- Memory CD4 & CD8 T Cells following SARS-CoV-2 Vaccination

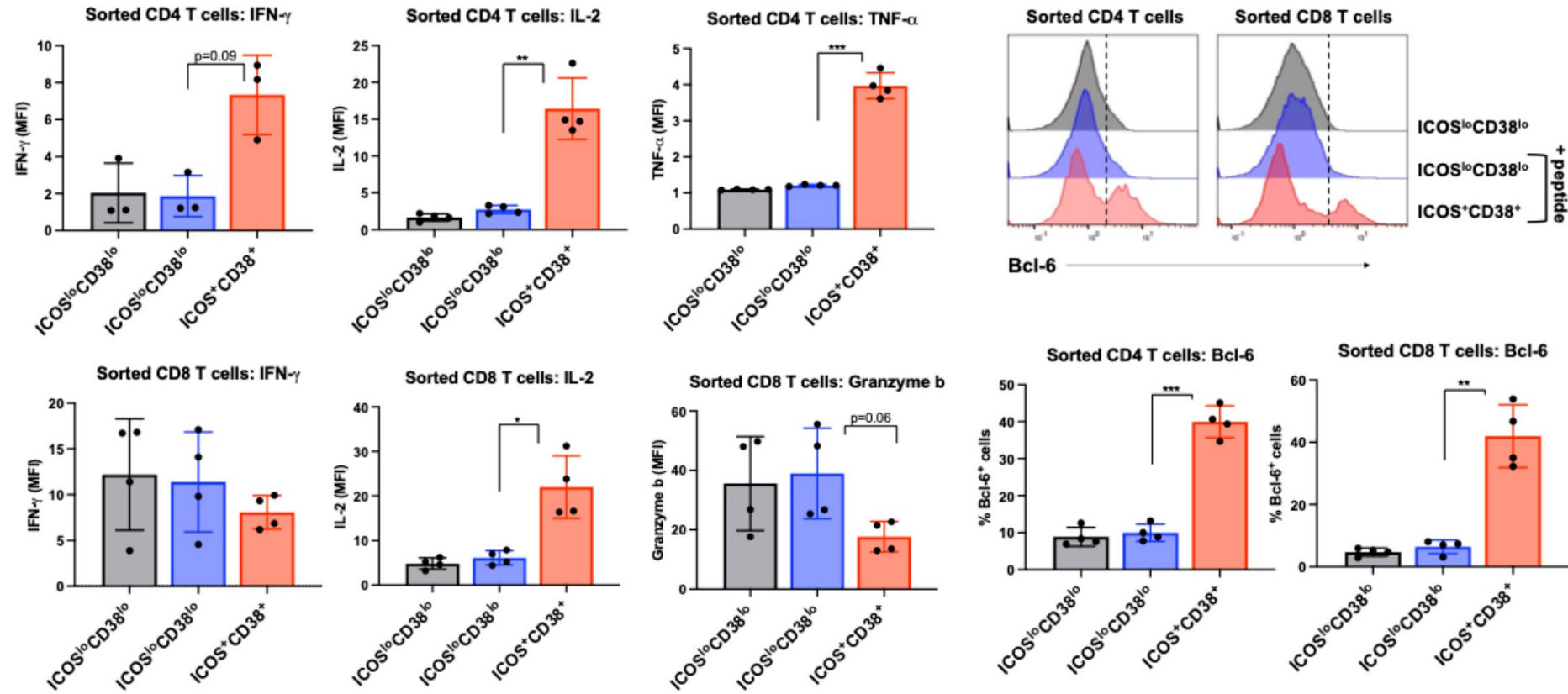
T cell mass cytometry panel on merged post-vaccine data (Day 28, N = 10)



Mass Cytometry Phenotyping of ICOS+ CD38+ PD-1+ CXCR5- Memory CD4 & CD8 T Cells following SARS-CoV-2 Vaccination



Sorting T cells on T-REX MEM Phenotype (ICOS⁺⁺ CD38⁺⁺) Confirms Specific SARS-CoV-2 Spike Peptide Reactivity



T_{FH}/T_{FC}? Only half of these cells were BCL-6+, and the cells from T-REX were CXCR5-



T-REX revealed virus-specific T cells without tetramers,
characterized massive immune changes in COVID-19,
& identified a SARS-CoV-2 reactive non-canonical
memory T cell that expands by day 28 following RNA vaccination

Check out the pre-print for more, including plasmablasts, B cell LIBRA-seq,
and a breakthrough case who did NOT generate the ICOS+ CD38+ T cells.



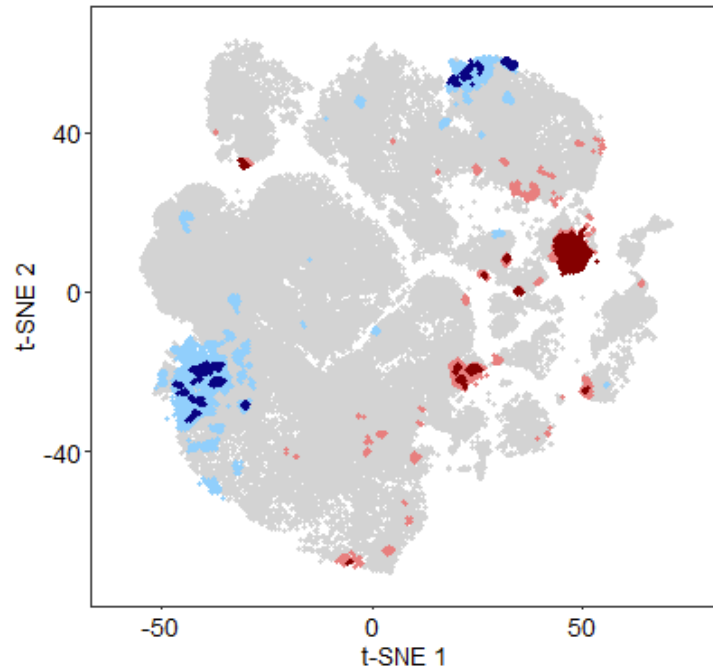
Let's Analyze Using T-REX!

<https://cytolab.shinyapps.io/TREX/>

This web app is running R code live.

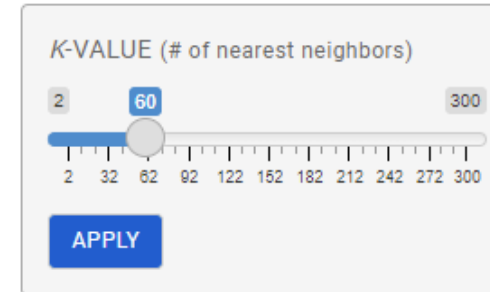
1) Identify populations of expansion and contraction with T-REX

CD3 T cells, COVID vaccine
Day 0 vs. Day 28 (N = 10)

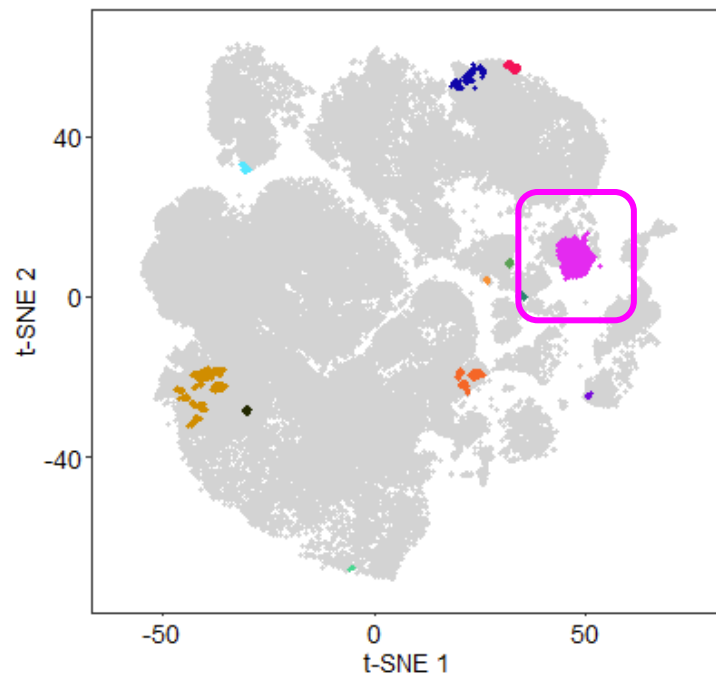


CUTOFF

- $\geq 95\%$ from day 0
- 85-95% from day 0
- from day 0 and 28
- 85-95% from day 28
- $\geq 95\%$ from day 28



2) Cluster with DBSCAN, and examine MEM labels



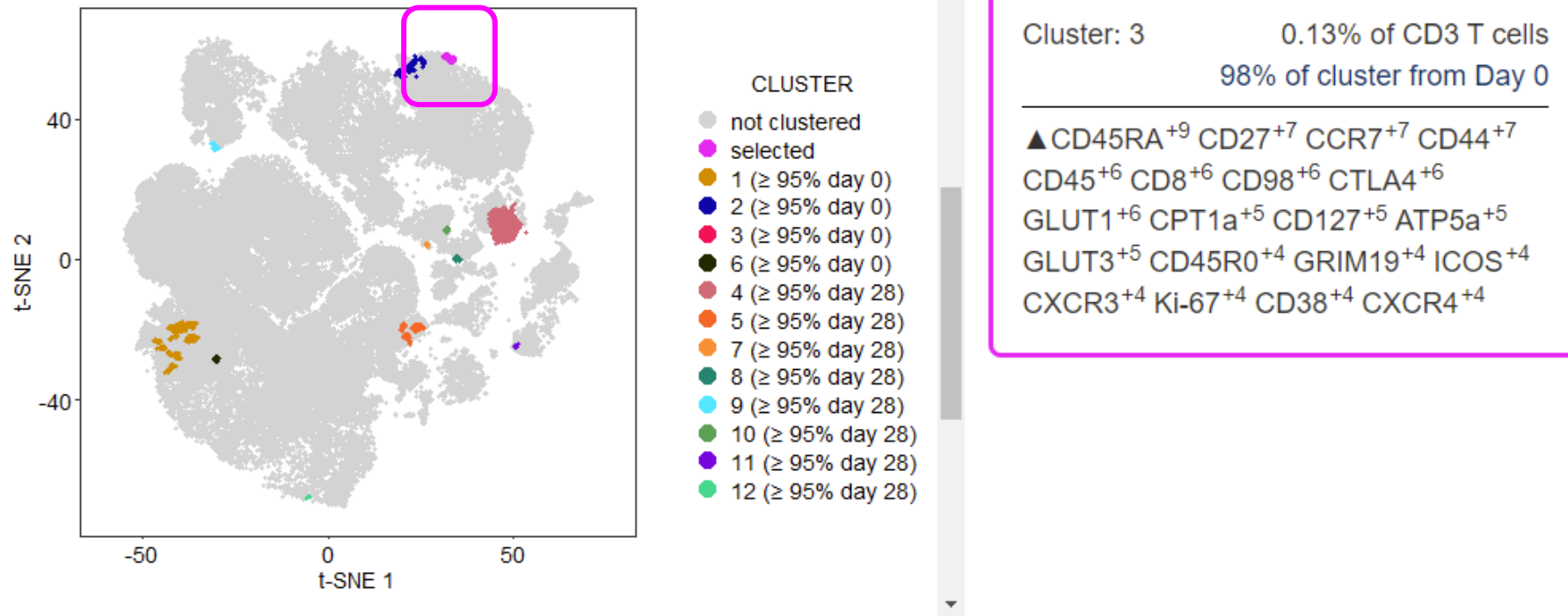
- CLUSTER
- not clustered
 - selected
 - 1 ($\geq 95\%$ day 0)
 - 2 ($\geq 95\%$ day 0)
 - 3 ($\geq 95\%$ day 0)
 - 6 ($\geq 95\%$ day 0)
 - 4 ($\geq 95\%$ day 28)
 - 5 ($\geq 95\%$ day 28)
 - 7 ($\geq 95\%$ day 28)
 - 8 ($\geq 95\%$ day 28)
 - 9 ($\geq 95\%$ day 28)
 - 10 ($\geq 95\%$ day 28)
 - 11 ($\geq 95\%$ day 28)
 - 12 ($\geq 95\%$ day 28)

Cluster: 4 3.7% of CD3 T cells

98% of cluster from Day 28

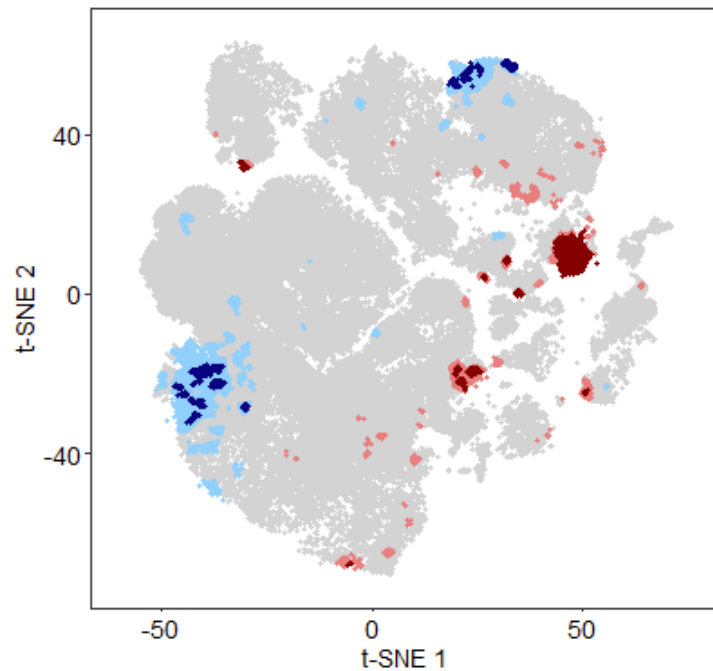
▲ CD38⁺⁹ CD45⁺⁷ CD8⁺⁷ CD45R0⁺⁷
ICOS⁺⁷ CD44⁺⁷ CD27⁺⁶ CXCR3⁺⁶
GLUT1⁺⁶ CD95⁺⁶ PD-1⁺⁶ CPT1a⁺⁵
CD98⁺⁵ Ki-67⁺⁵ HLA-DR⁺⁵ CD3⁺⁴
CTLA4⁺⁴ CYTOC⁺⁴

2) Cluster with DBSCAN, and examine MEM labels



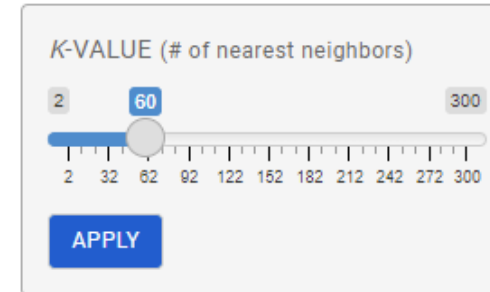
1) Identify populations of expansion and contraction with T-REX

CD3 T cells, COVID vaccine
Day 0 vs. Day 28 (N = 10)



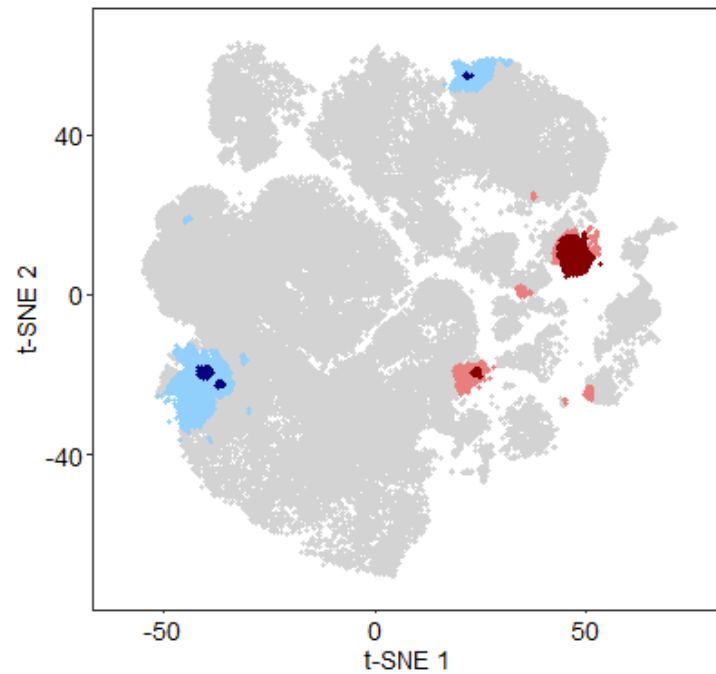
CUTOFF

- $\geq 95\%$ from day 0
- 85-95% from day 0
- from day 0 and 28
- 85-95% from day 28
- $\geq 95\%$ from day 28



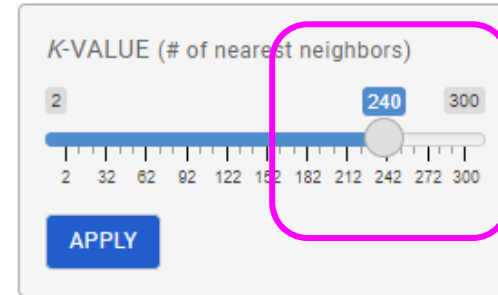
1) Identify populations of expansion and contraction with T-REX

CD3 T cells, COVID vaccine
Day 0 vs. Day 28 (N = 10)



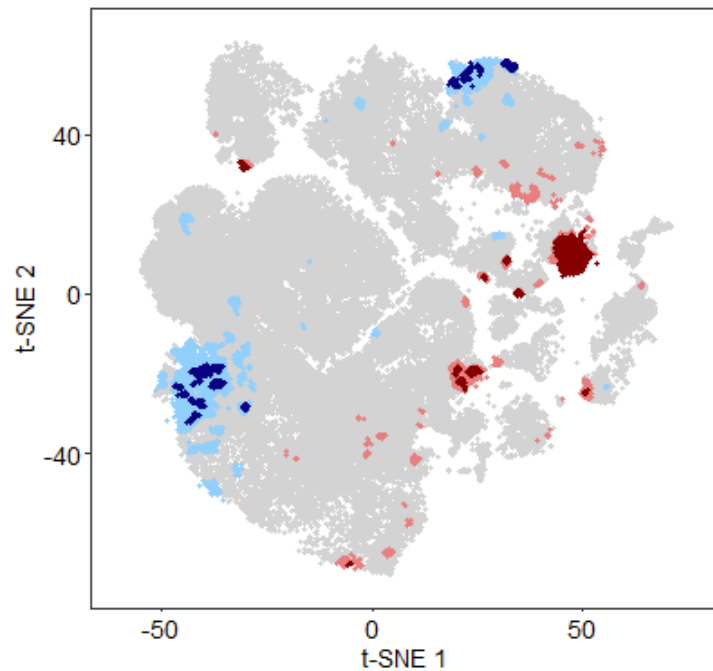
CUTOFF

- $\geq 95\%$ from day 0
- 85-95% from day 0
- from day 0 and 28
- 85-95% from day 28
- $\geq 95\%$ from day 28



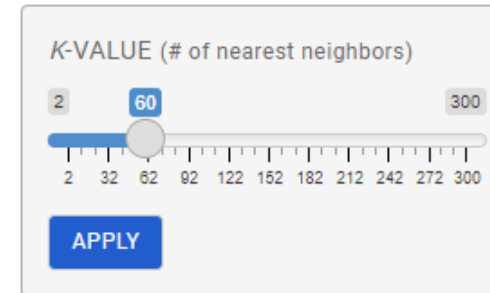
1) Identify populations of expansion and contraction with T-REX

CD3 T cells, COVID vaccine
Day 0 vs. Day 28 (N = 10)



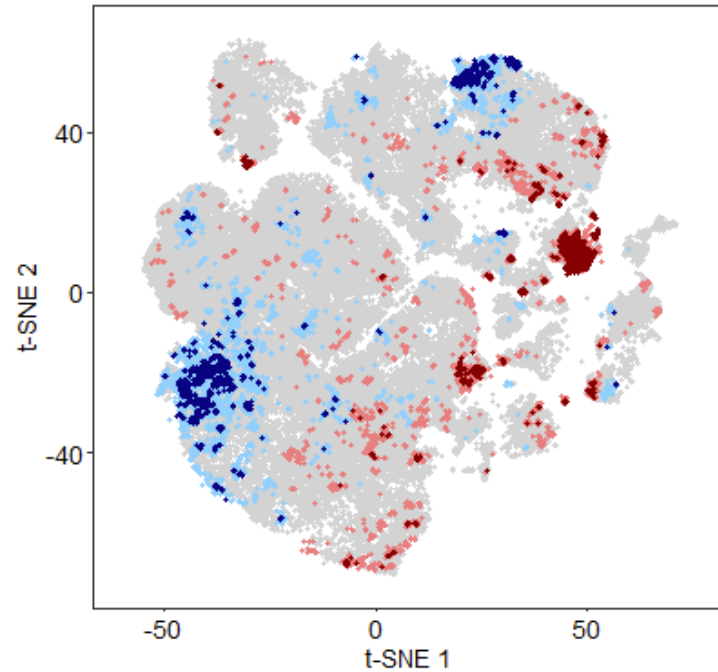
CUTOFF

- $\geq 95\%$ from day 0
- 85-95% from day 0
- from day 0 and 28
- 85-95% from day 28
- $\geq 95\%$ from day 28



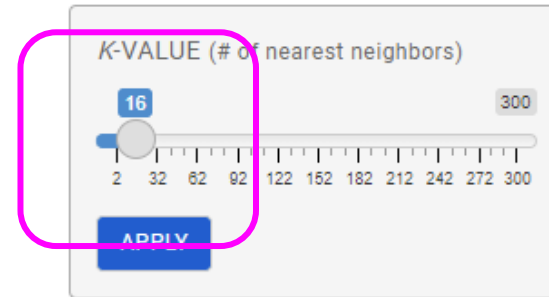
1) Identify populations of expansion and contraction with T-REX

CD3 T cells, COVID vaccine
Day 0 vs. Day 28 (N = 10)



CUTOFF

- $\geq 95\%$ from day 0
- 85-95% from day 0
- from day 0 and 28
- 85-95% from day 28
- $\geq 95\%$ from day 28



T-REX & COVID-19 Acknowledgements

Irish Lab at Vanderbilt University + Cancer & Immunology Core



Stephanie Medina
PhD Student



Amanda Kouaho
VU Undergraduate



Sierra Barone Lima
Data Science
Program Coordinator



Todd Bartkowiak
PhD Postdoc
K00 Fellow



Caroline Roe
CIC/MCCE
Senior Research
Specialist



Madeline Hayes
Lab Development
Program
Coordinator

University of Virginia Collaborators (Rhinovirus T-REX)



Alberta Paul



Lyndsey Muehling



Judith Woodfolk

Vanderbilt University Medical Center (SARS-CoV-2 vaccine response)



Kevin Kramer



Erin Wilfong



Kelsey Voss



Rachel Bonami



Ivelin Georgiev



Jeff Rathmell

Southeastern Brain Tumor Foundation (Ihrle & Irish), NIH/NCI : R00 CA143231 (Irish), R01 CA226833 (Bachmann & Irish), **HIDI TIPS (Vanderbilt)** U01 CA196405 (Massion), CSBC U54 CA217450 (Quaranta), **U01 AI125056 (Woodfolk)**, **R01 HL136664 (Rathmell)**, F31 CA199993 (Greenplate), T32 CA009592 & R25 GM062459 (Doxie), R25 CA136440 (Diggins), K12 CA090625 (Ferrell), P30 CA68485 (Vanderbilt-Ingram Cancer Center)

Acknowledgements & Thank You!

Past Lab Members

Grad & Med Students

Deon Doxie (Emory)
Cara Wogsland (UiB & BergenBio)
Kirsten Diggins (Benaroya Institute)
Allison Greenplate (U Penn)
Nalin Leelatian (Yale & Vanderbilt)
Jocelyn Gandelman (UCSF)

Postdocs

Kanutte Huse (Oslo University)
Mikael Roussel (CHU Rennes)
P. Brent Ferrell, Jr. (Vanderbilt)
Ashley Wu (Vanderbilt)

Undergrads & Staff

Sierra Barone Lima (Data Science)
Hannah Polikowsky (Vanderbilt)
Alejandra Rosario-Crespo (U Puerto Rico)
Daniel McClanahan (Twitter & LBL)
Daniel Liu (Univ Wisc., Madison)
Nathan Wasserman (Univ Florida, Miami)

Visiting Scholars

Shahram Kordasti (King's College London)
Faustine L'homme (CHU Rennes)
Monica Hellesøy (Univ. Bergen)
Aida Meghraoui-Kheddar (Paris & Nice)
Eleni Syrimi (Univ. Birmingham, UK)
Laura Ferrer Font (Malaghan, NZ)

Irish Lab, Vanderbilt University



Stephanie
Medina
PhD Student
Cell & Dev Biology



Claire
Cross
PhD Student
Chem & Phys Bio



Hannah
Thirman
PhD Student
Chem & Phys Bio



Niraj
Rama
Vanderbilt
Undergraduate



Amanda
Kouaho
Vanderbilt
Undergraduate



Todd
Bartkowiak
PhD Postdoc
K00 Fellow



Caroline
Roe
Director
CIC Core



Madeline
Hayes
Program
Coordinator



Cass
Mayeda
Web Applications
Research Assistant



Jonathan
Irish
Principal
Investigator

Collaborations

Glioblastoma & Neural Stem Cells

Rebecca Ihrie
Vivian Gama
Asa Brockman
Bret Mobley
Lola Chambless
Reid Thompson

Chemical Biology

Brian Bachmann &
Lab for Biosynthetic Studies

John Porco (BU)
Lauren Brown (BU)

Human Immunology Discovery Initiative (HIDI)

Jeff Rathmell
Jim Connelly
Saara Kaviany

Viral Immunology

Judith Woodfolk (UVA)
Lyndsey Muehling (UVA)
Glenda Canderan (UVA)

Quantitative & Systems Biology Center

Vito Quaranta & many more
(U54 CA217450)

Center for Extracellular Vesicles

Alissa Weaver
Heather Pua
Andries Zijlstra

Michael David Greene Brain Cancer Fund (Thompson, Ihrie, Irish), Southeastern Brain Tumor Foundation (Ihrie & Irish), Ivy Foundation (Ihrie & Irish), R01 NS118580 (Ihrie & Ess), R00 CA143231 (Irish), R01 CA226833 (Bachmann & Irish), U01 CA196405 (Massion), CSBC U54 CA217450 (Quaranta), R01 HL136664 (Rathmell), U01 AI125056 (Woodfolk), U01 TR002625 (Porco), Cancer & Immunology Core, Flow Cytometry Core, HIDI TIPS (Vanderbilt), P30 CA68485 (Vanderbilt-Ingram Cancer Center)



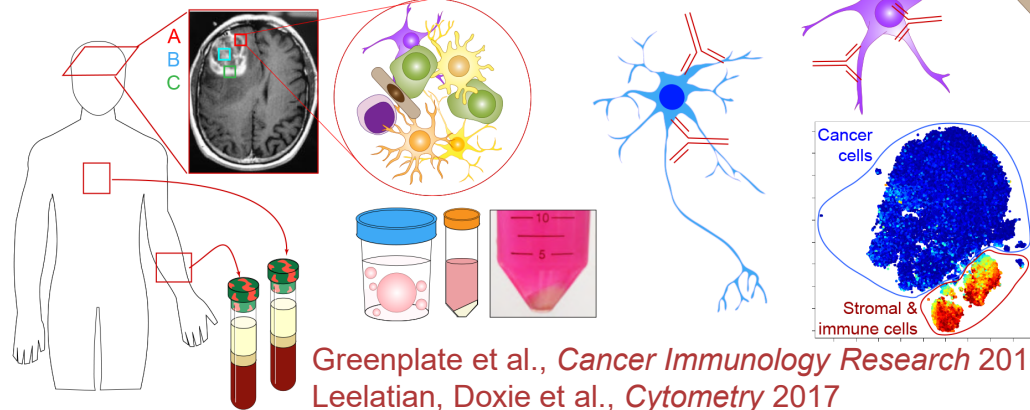
Irish Lab @ Vanderbilt University

Single cell biology for precision medicine

Irish lab website:

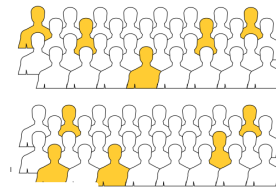
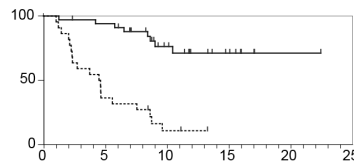
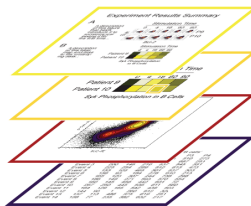


Human immune & tumor tissues



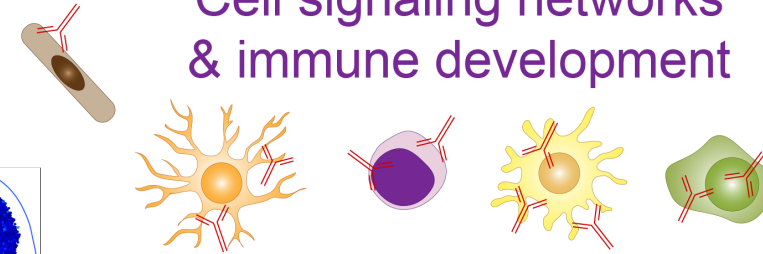
Greenplate et al., *Cancer Immunology Research* 2016
Leelatian, Doxie et al., *Cytometry* 2017
Doxie et al., *Pigment Cell & Melanoma Research* 2018
Leelatian, Sinnaeve et al., *eLife* 2020 ([RAPID](#))

Precision medicine & screening



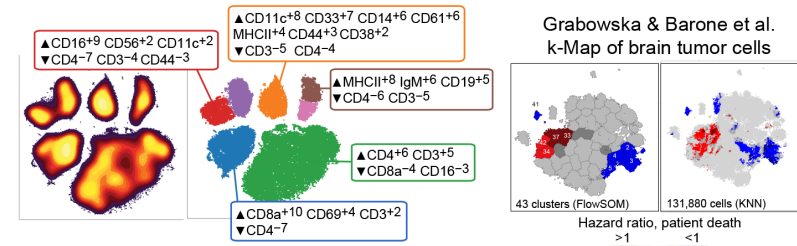
Ferrell et al., *PLoS One* 2016
Earl, Ferrell et al., *Nature Communications* 2018
Greenplate et al., *Cancer Immunology Research* 2019
Kramer, Wilfong, Voss et al., *bioRxiv* 2021 (COVID vaccine response)

Cell signaling networks & immune development



Polikowsky et al., *J Immunology* 2015
Roussel et al., *J Leukocyte Biology* 2017
Huse et al., *Cytometry* 2018
Bartkowiak et al., 2021 (in prep)

AI & machine learning



Diggins et al., *Methods* 2015
Diggins et al., *Nature Methods* 2017 ([MEM](#))
Gandelman et al., *Hematologica* 2018
Barone, Paul, Muehling et al., *eLife* 2020 ([T-REX](#))

Cell & Developmental Biology

Cancer Biology

Chemical & Physical Biology

Molecular Pathology & Immunology

Vanderbilt University in Nashville, TN

Vanderbilt University



Irish Lab



VU & Medical Center



Jonathan Irish, Ph.D.
Vanderbilt University, Nashville, TN
Cell & Developmental Biology
Pathology, Microbiology & Immunology

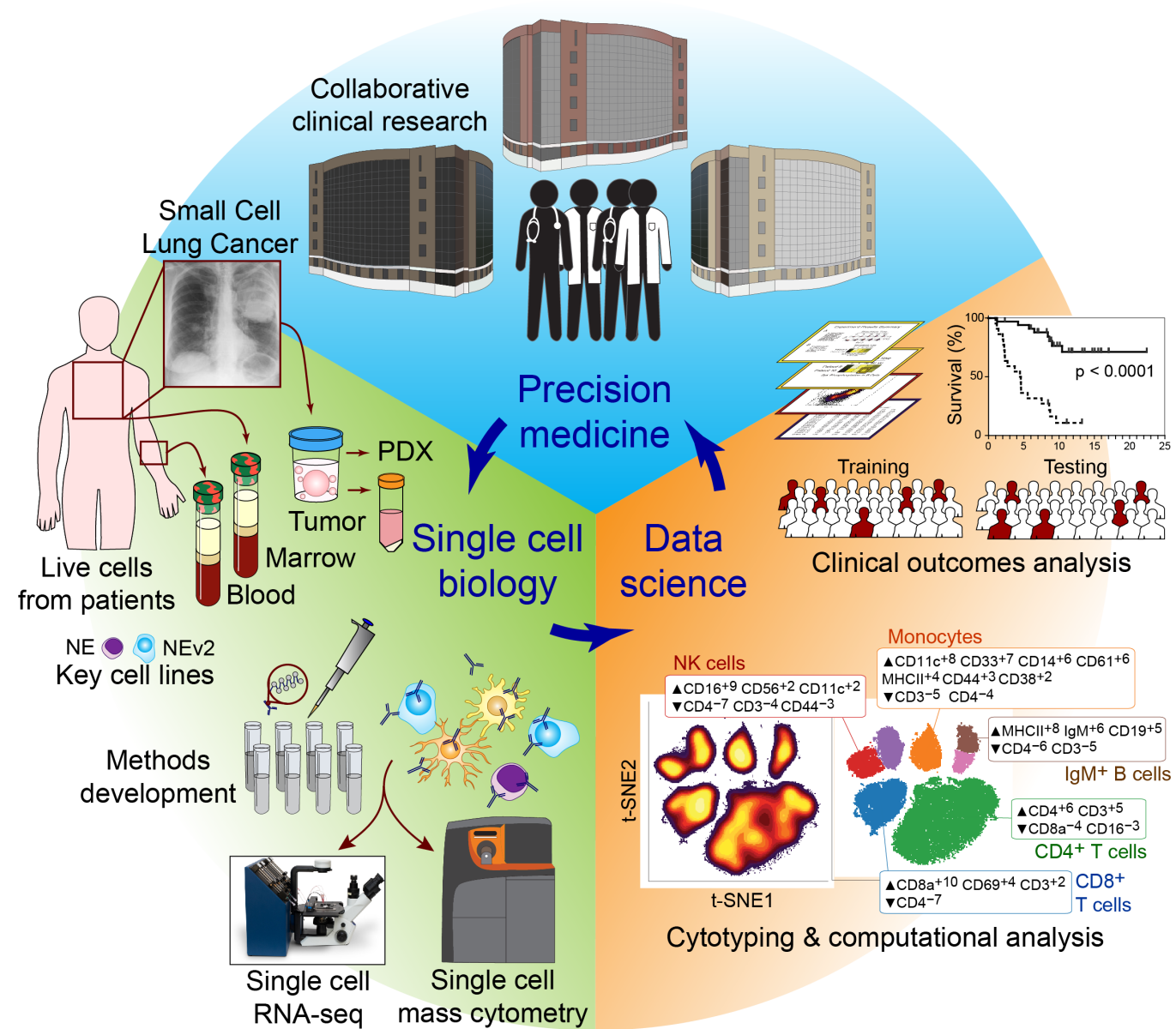
Nashville, Tennessee – Music City USA

 @JonathanIrish

2022

Introduction to Data Science and Computational Tools

Goal: Systematically Dissect Cellular Mechanisms Across Time, Treatments, Tissues, & Tumor Types



Imagine Finding Pieces of a Jigsaw Puzzle...

Flow Cytometry

Puzzle

Setup	<p>Manual review (scaling, single cell gating, compensation, batch correction)</p>	<p>Manual review (make sure all the pieces are from the same puzzle)</p>
Organization	<p>t-SNE, UMAP, PCA (simplify the problem by organizing the data)</p>	<p>Group pieces (find corners, edges, pieces with distinct colors)</p>
Grouping	<p>FlowSOM, SPADE, gating (split cells into cell types like T cells or monocytes)</p>	<p>Assemble parts (connect similar pieces, create distinct shapes)</p>
Interpretation	<p>Heatmaps, MEM, RMSD (analyze group features, learn cell identities)</p>	<p>Interpret picture (see both the pieces and the whole picture)</p>

Effective data analysis is critical in clinical research,
& this now means working *with* computational tools
that reveal and model patterns across data types

Tools from one area can be applied in others
(economics, math, **patients**, **cells**, pixels, ...)

Data science workshop can be self-taught:

<https://github.com/cytolab/>

Unsupervised Analysis: Not Using Prior Knowledge To Guide the Analysis

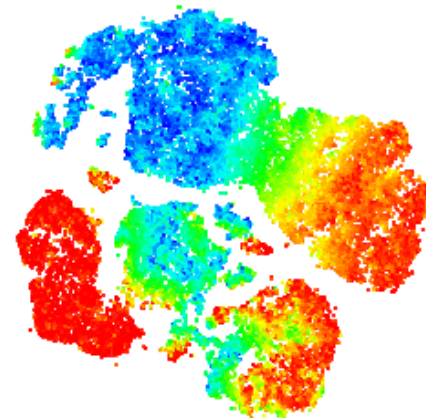
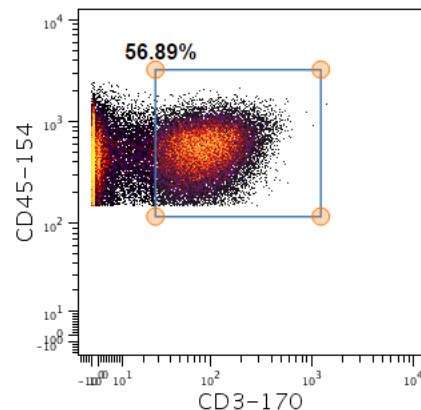
Prior knowledge examples: Stem cells express CD34, these samples were from patients that responded to drug

Supervised Approaches

- Expert gating
- Citrus
- CellCNN (neural network)
- Wanderlust

Unsupervised Approaches

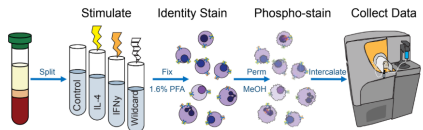
- Most heatmap clustering
- SPADE, FlowSOM
- t-SNE / viSNE, UMAP
- Phenograph



Flow Cytometry Workflow from Data Collection to Deep Analysis

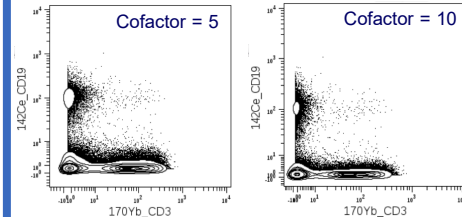
Data collection

- 1) Panel design
- 2) Data collection



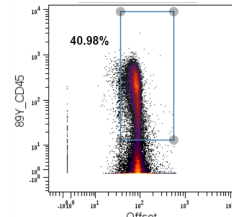
Data processing

- 3) Normalization
- 4) Concatenation
- 5) Scale transformation



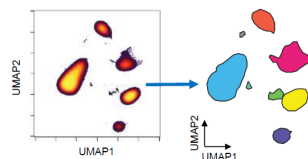
Distinguishing initial populations

- 6) Live single cell gating
- 7) Focal population gating



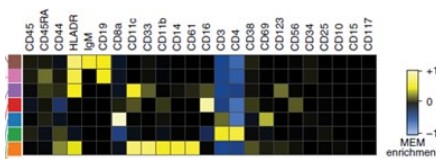
Revealing cell subsets

- 8) Feature selection
- 9) Dimensionality reduction
- 10) Identify cell clusters



Characterizing cell subsets

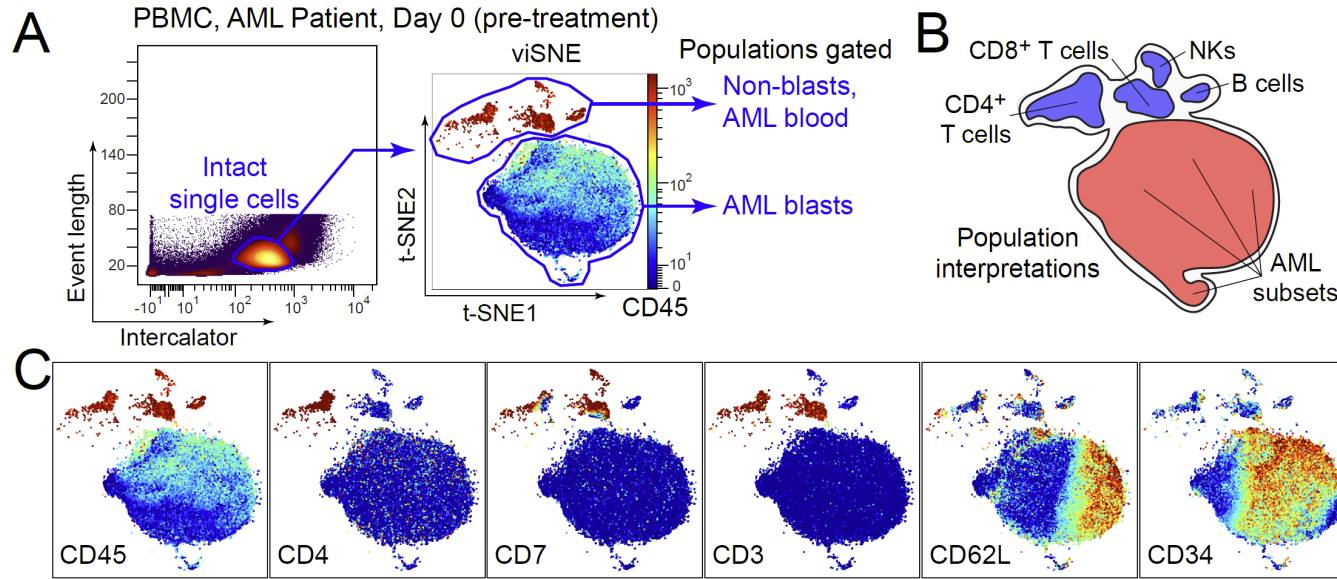
- 11) Feature comparison
- 12) Model populations
- 13) Learn cell identity
- 14) Statistical testing



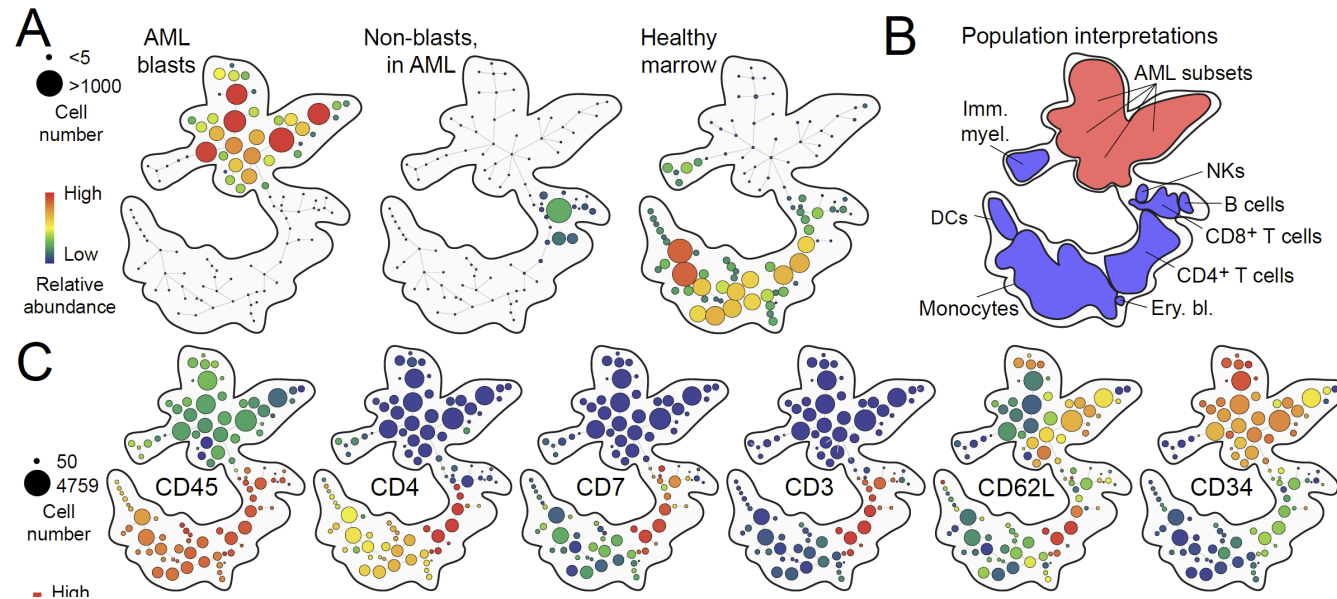
How much can be automated?

How do we select tools and use them well?

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

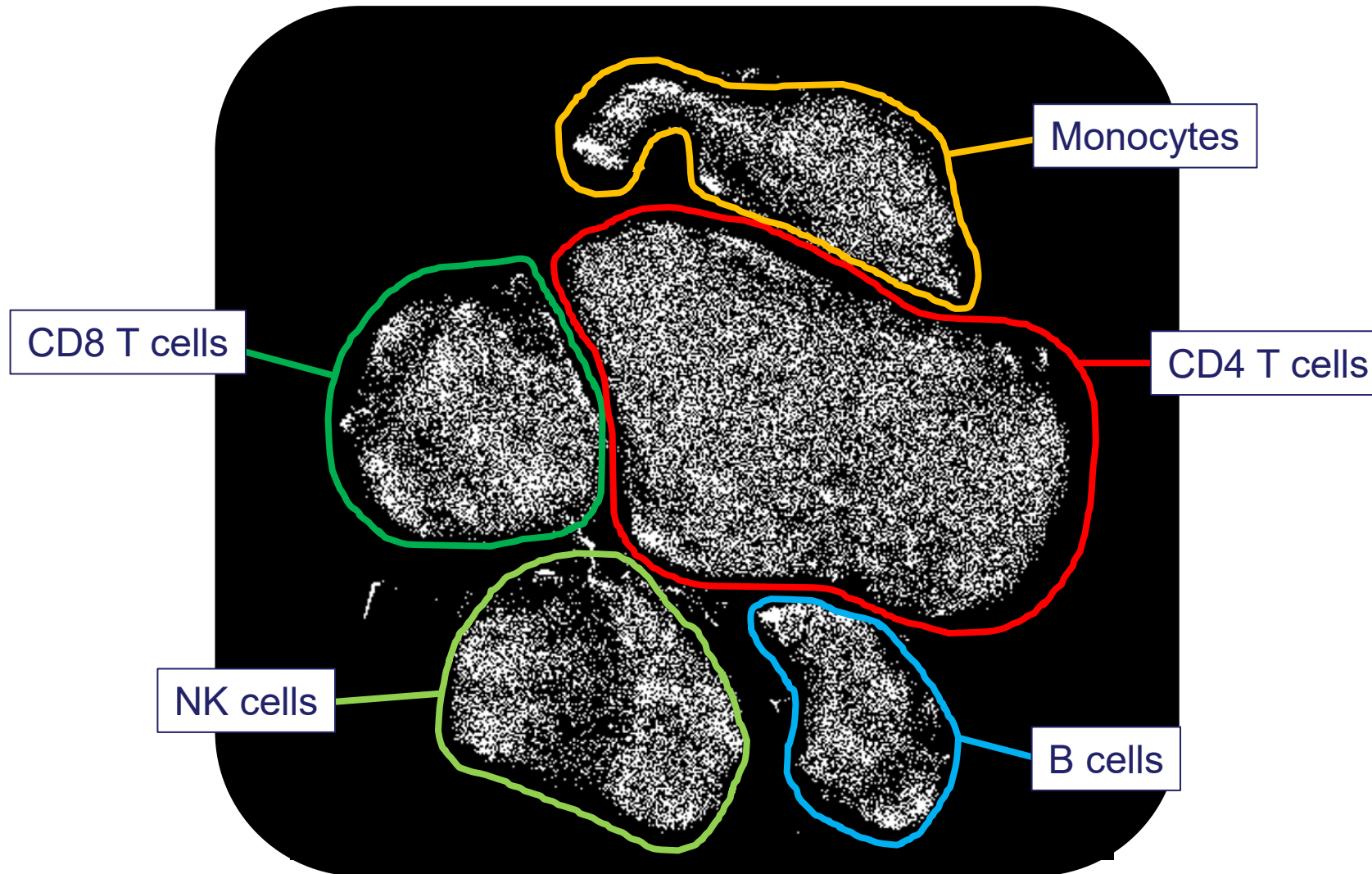


viSNE
Amir et al.
Nature biotech 2013



SPADE
Qiu et al.
Nature biotech 2011

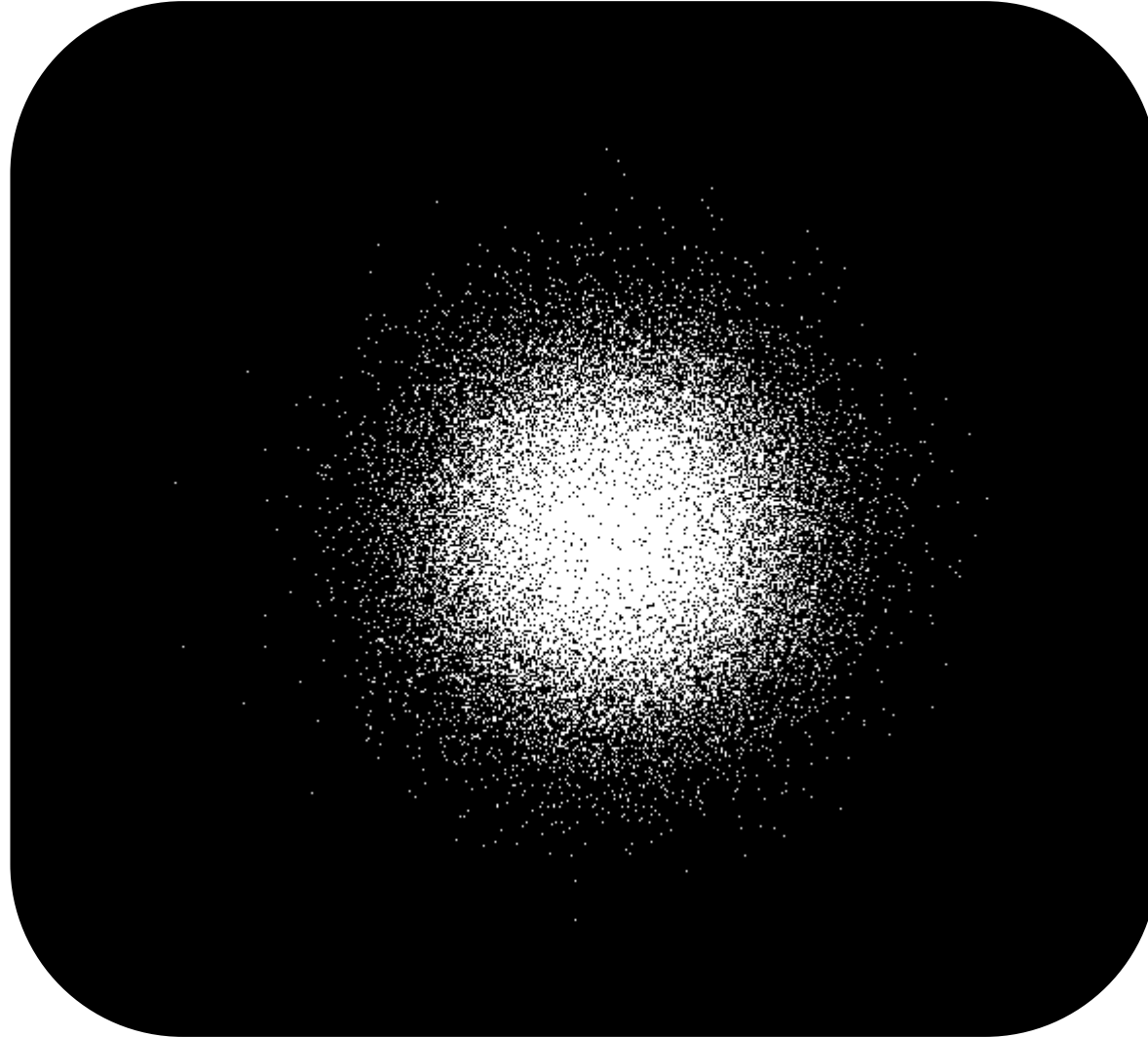
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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

Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)

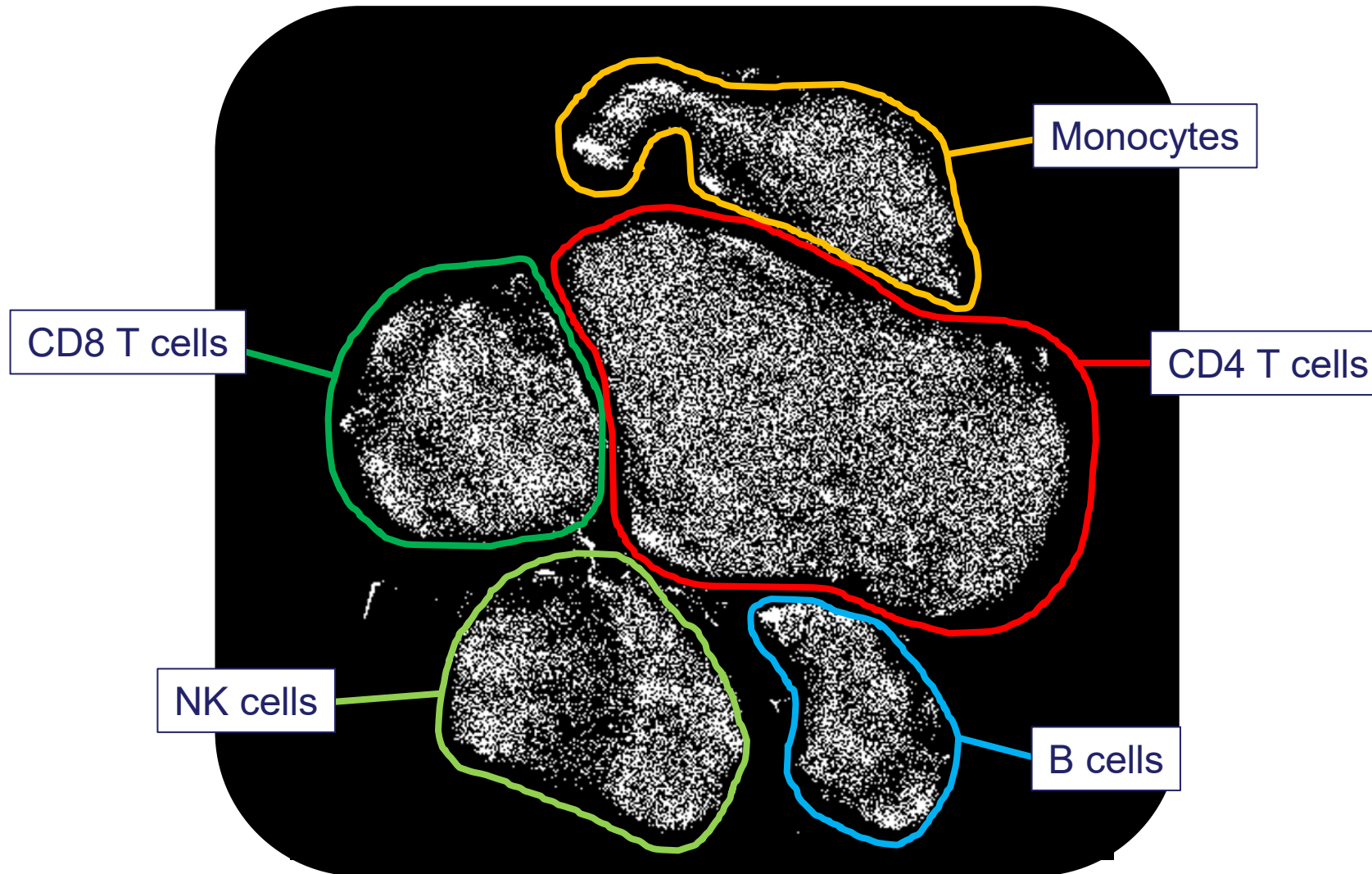
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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

Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)

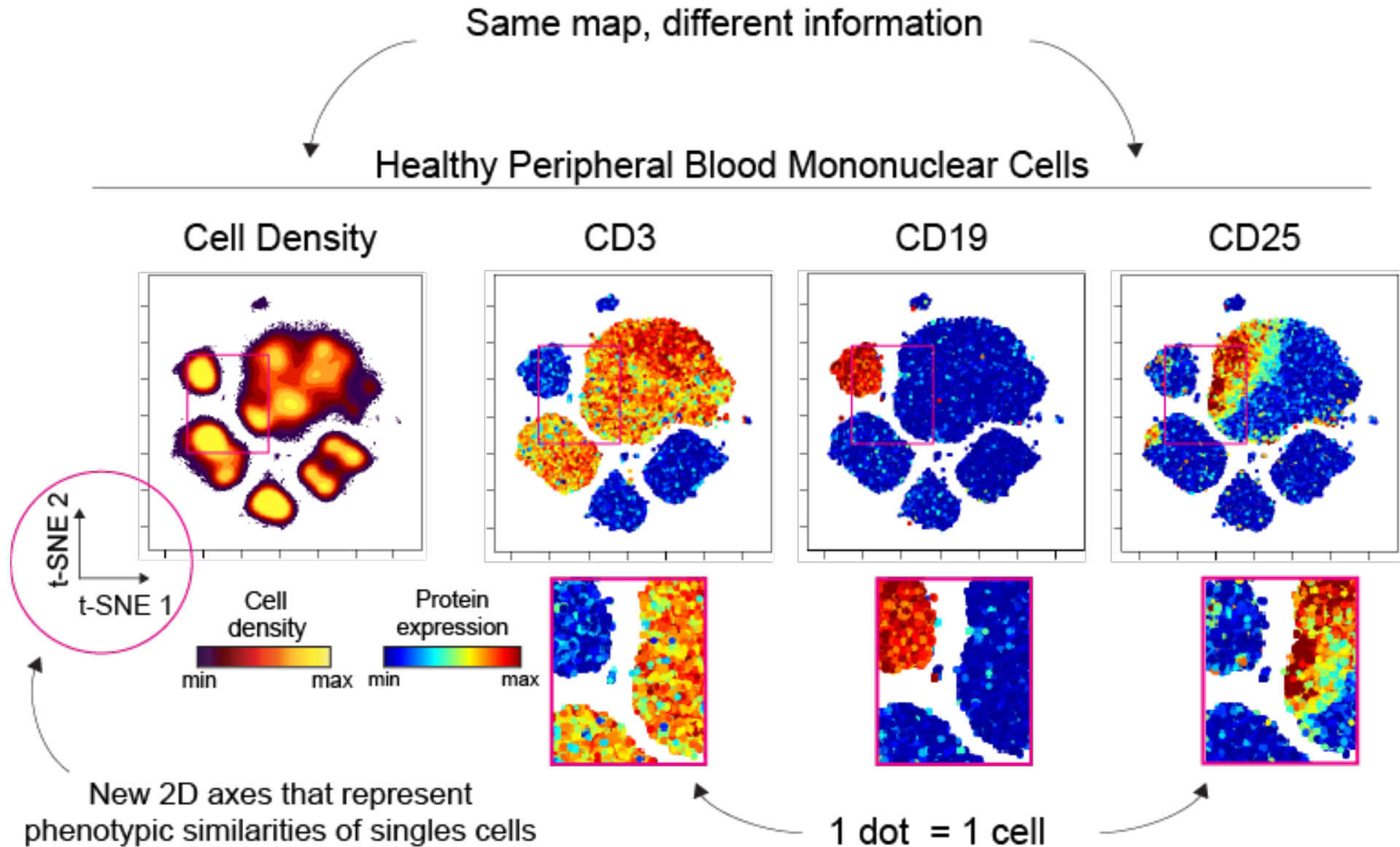
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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

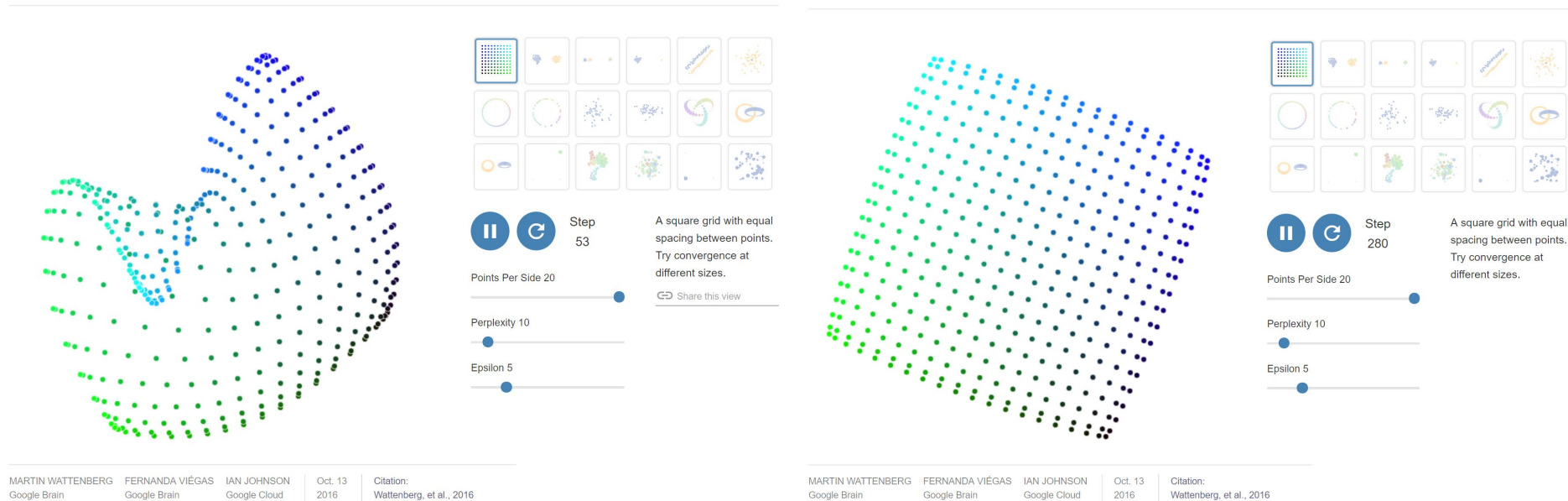
Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)

t-SNE Analysis Allows 2D Visualization of High Dimensional Single Cell Data

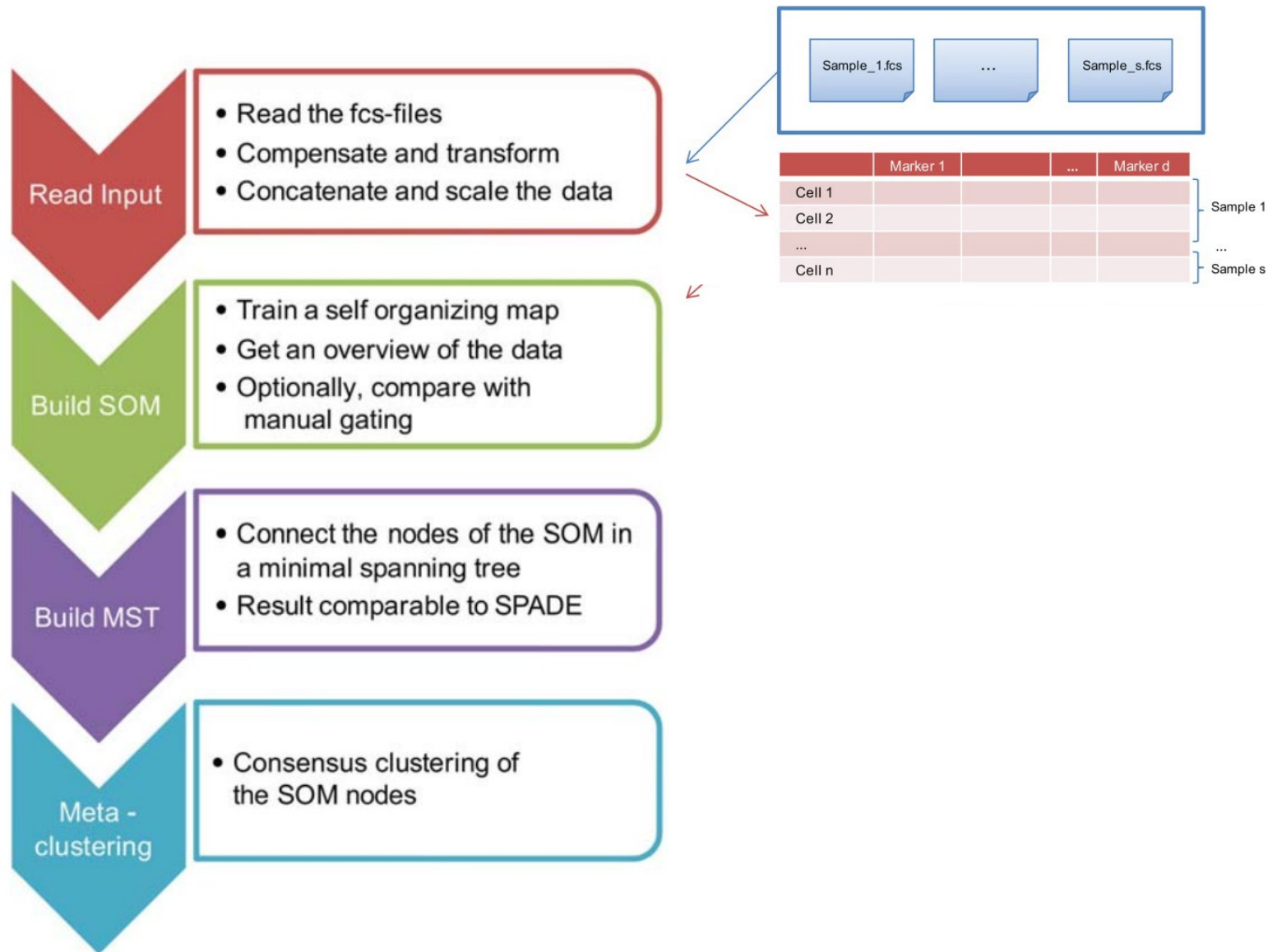


t-SNE 2D Examples with Animations and Settings

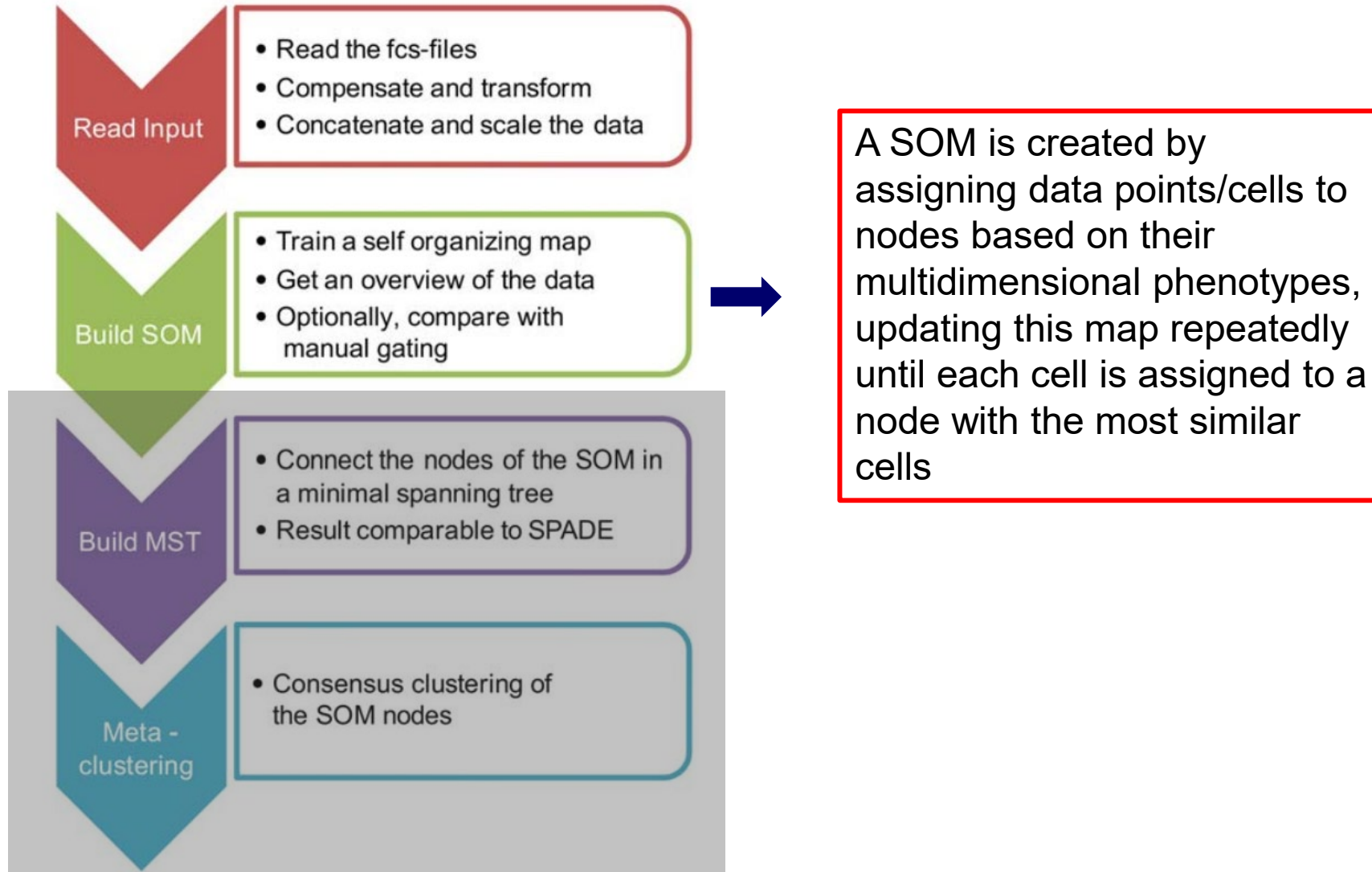
<http://distill.pub/2016/misread-tsne/>



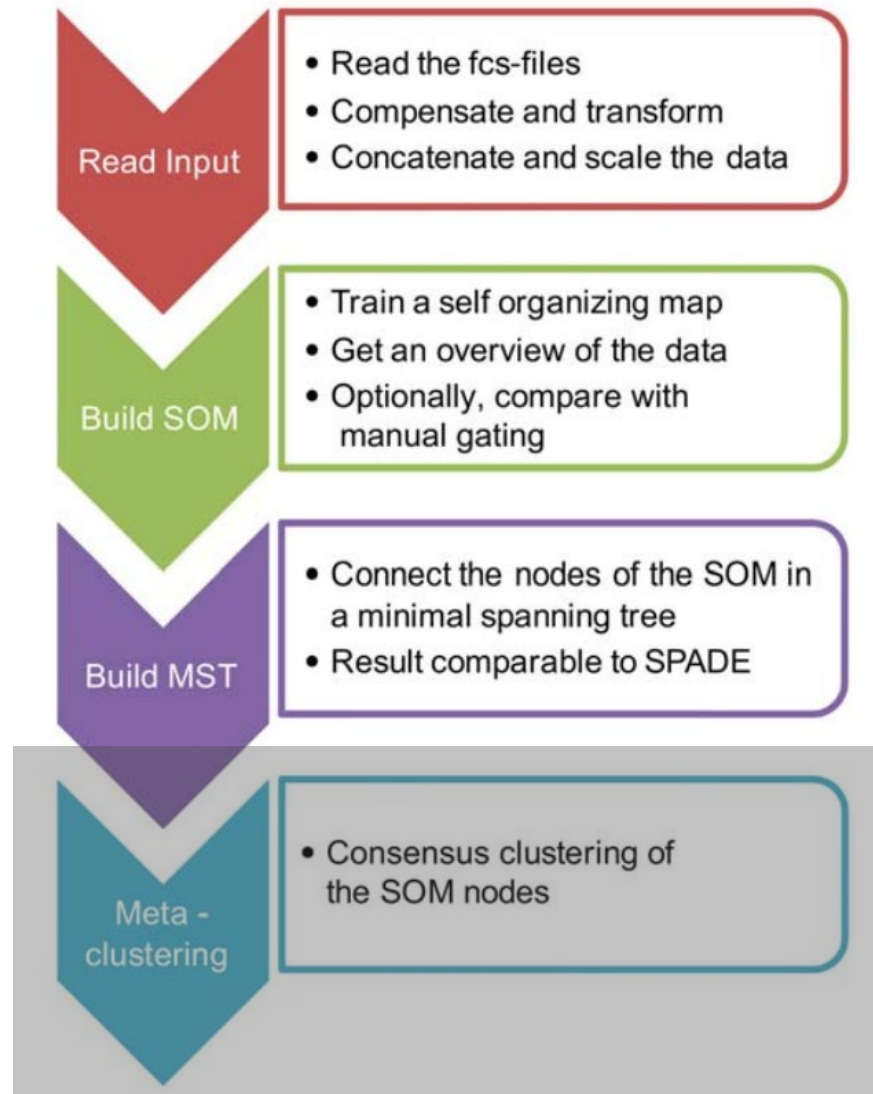
Clustering with FlowSOM: Self-organizing Maps



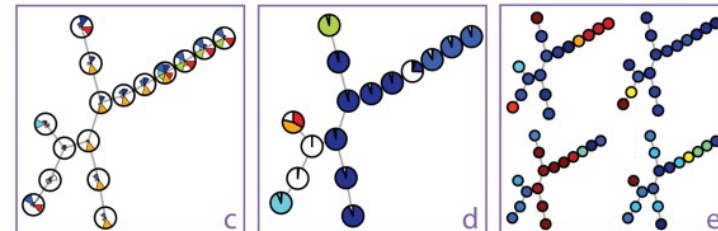
Clustering with FlowSOM: Self-organizing Maps



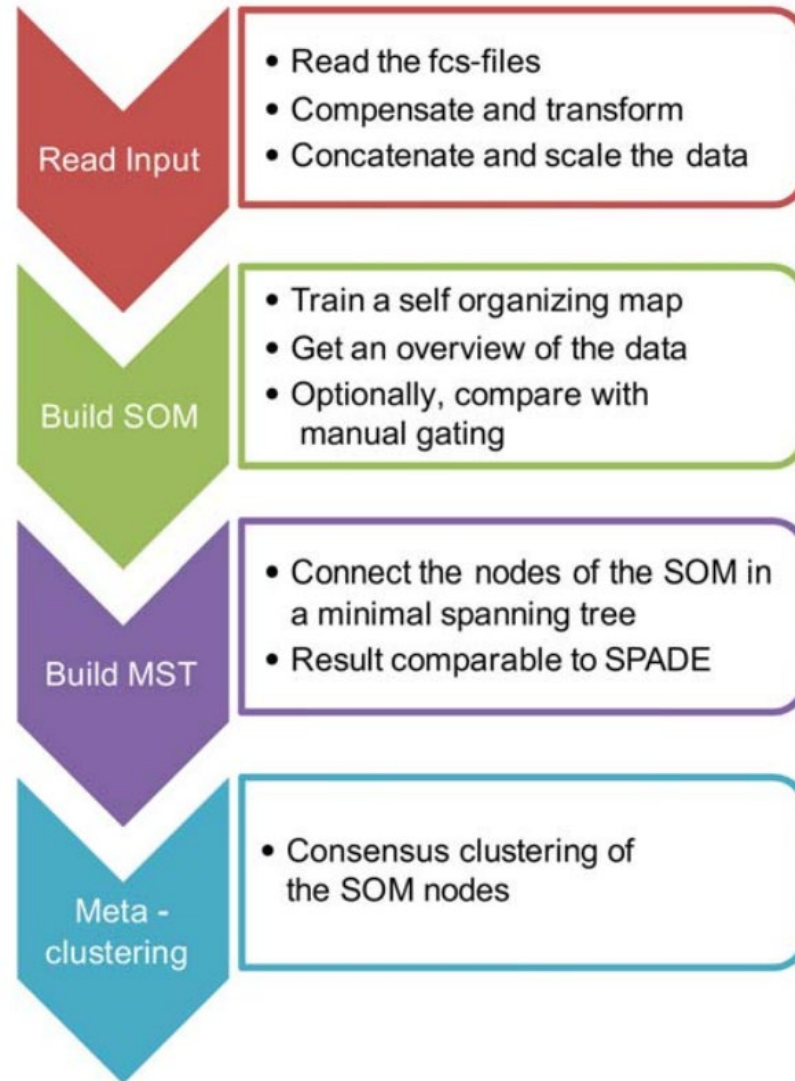
Clustering with FlowSOM: Self-organizing Maps



The next step is to arrange the nodes along a minimal spanning tree (MST), so that nodes that are most similar are closest on the tree
not used in our visualization

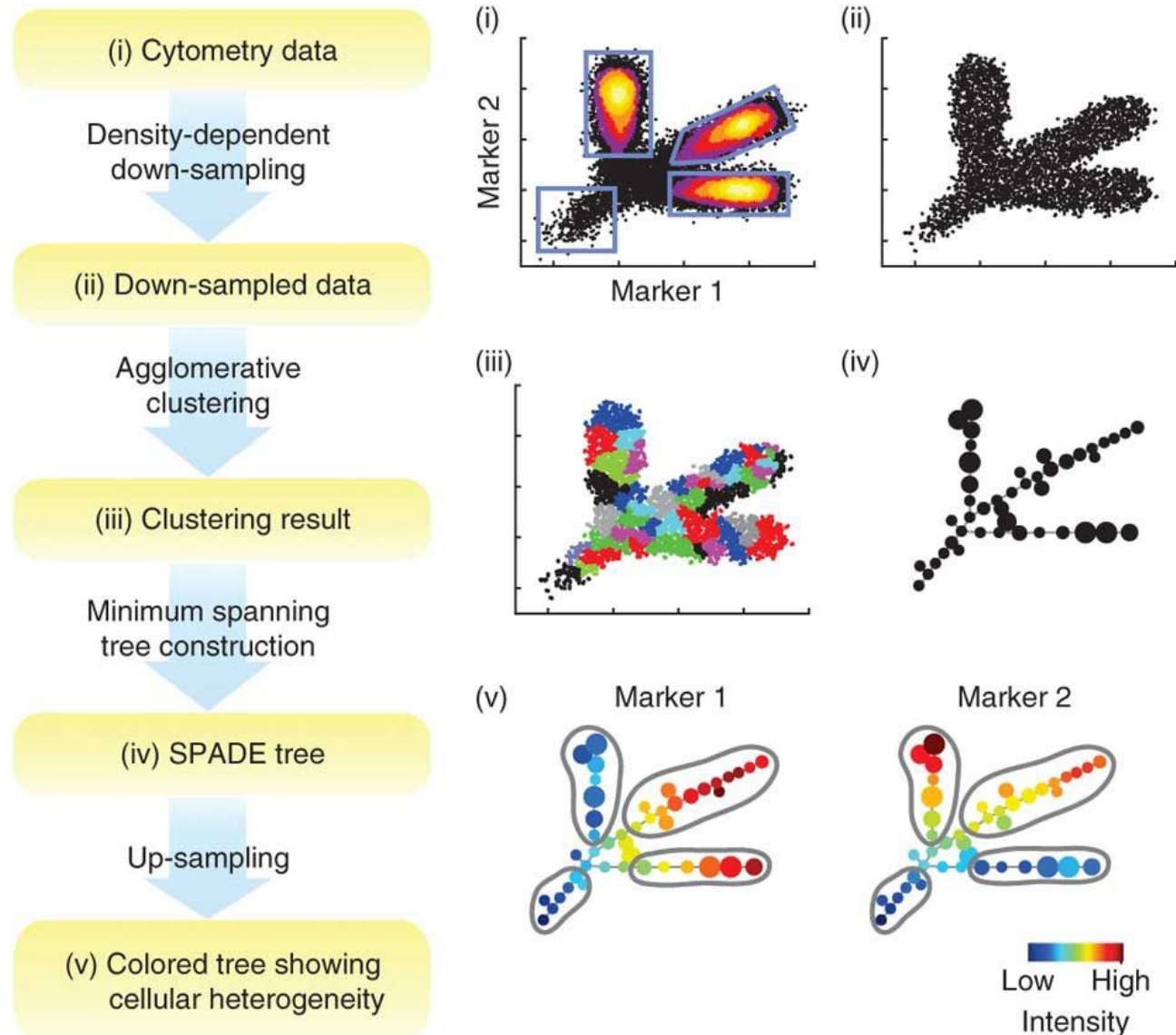


Clustering with FlowSOM: Self-organizing Maps

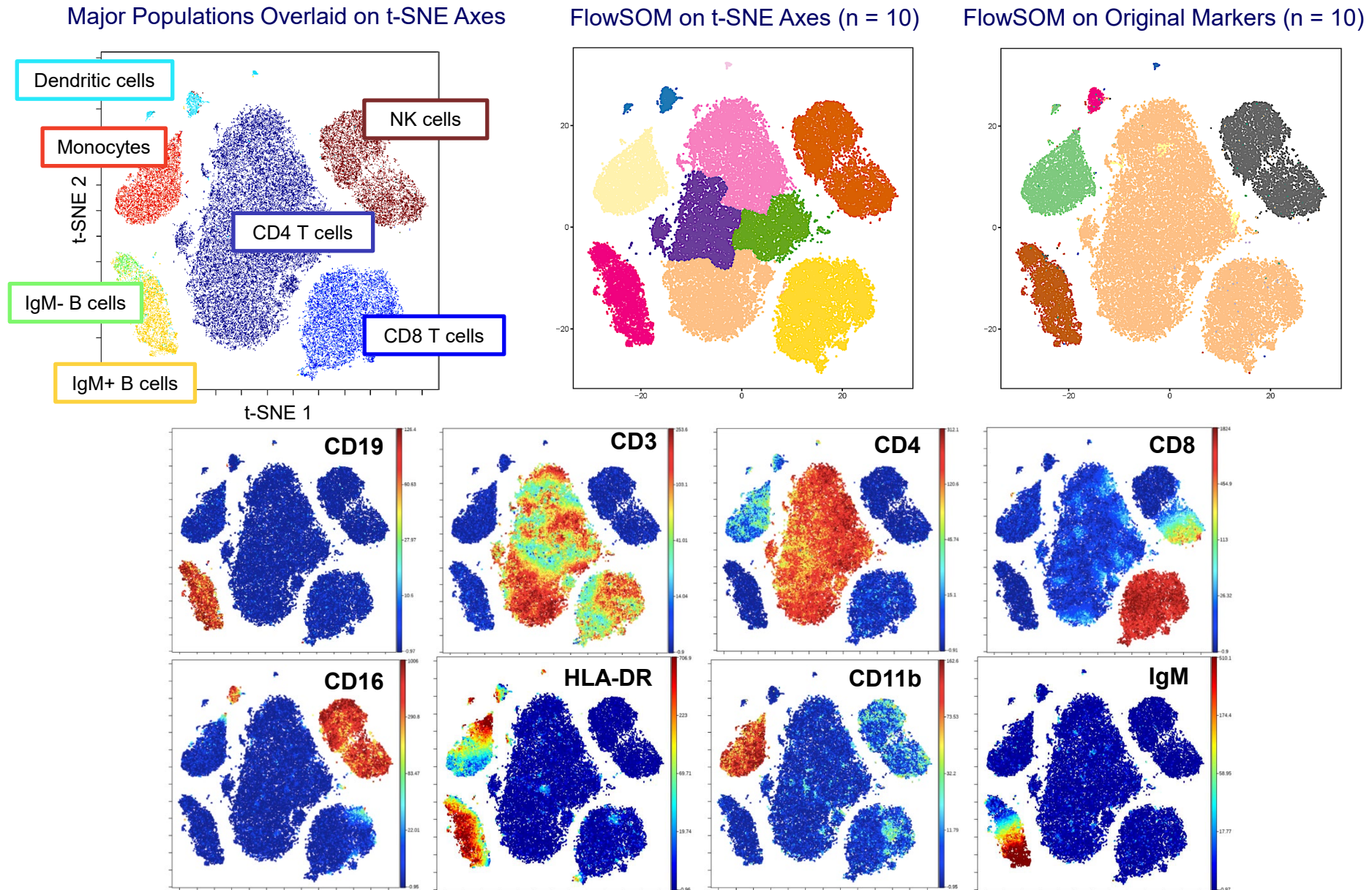


Finally, similar nodes are combined based on the number of desired clusters defined by the user. This desired number can be based on prior knowledge or a specific goal (i.e. minimizing intracluster variance)

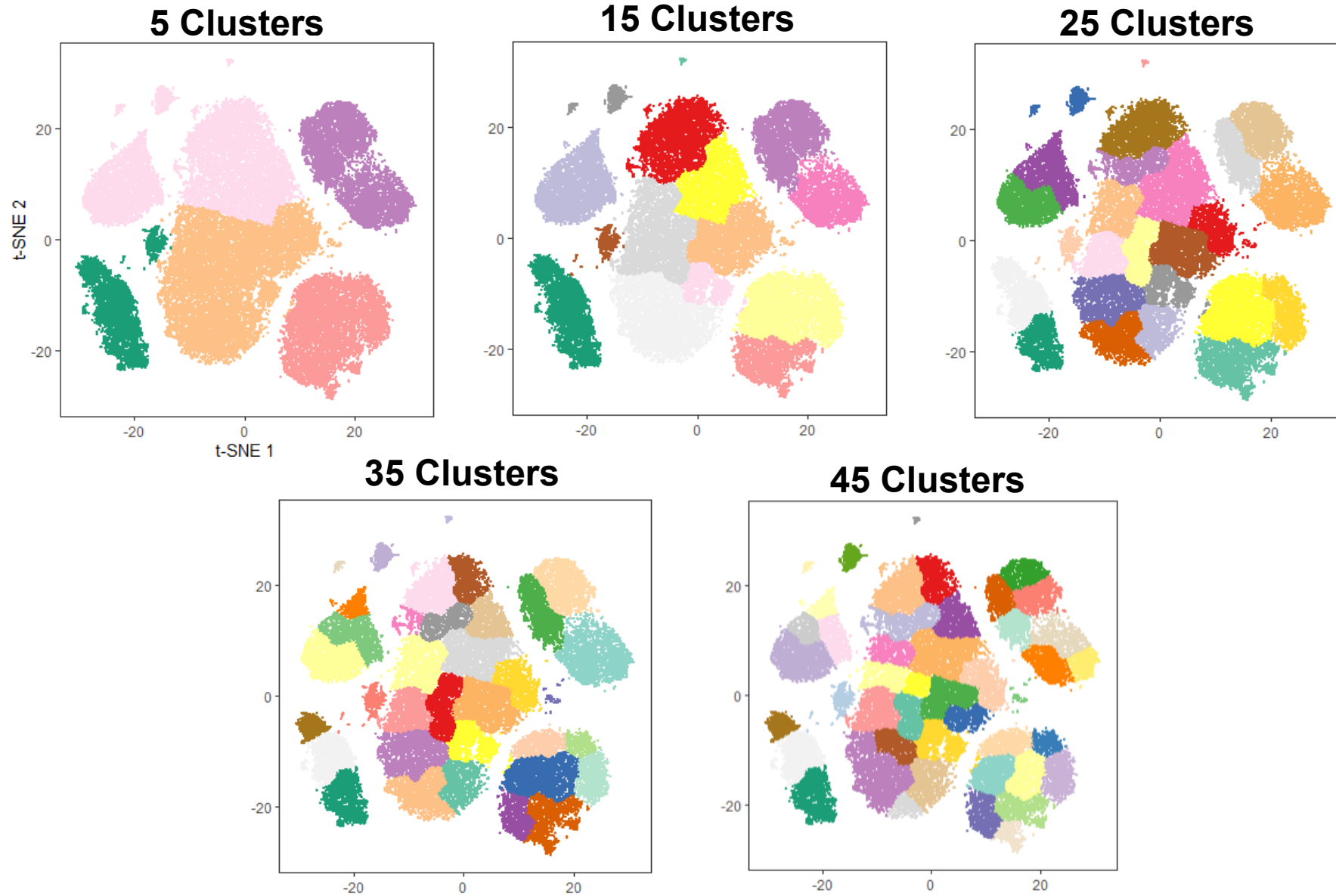
Spanning-Tree Progression Analysis of Density-Normalized Events (SPADE) is an Alternative Clustering Tool



FlowSOM Clusters are Dependent on Input Parameters

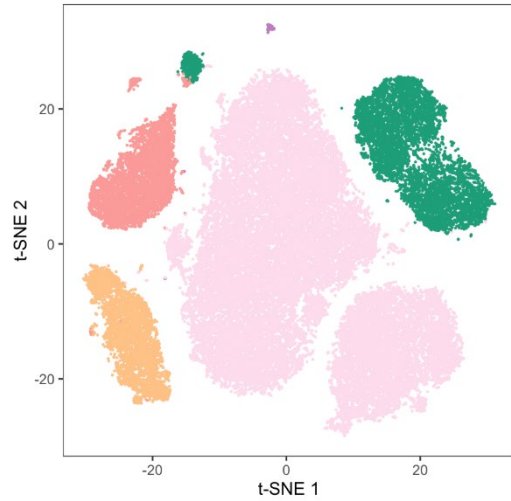


FlowSOM Requires that Users Choose a Number of Clusters

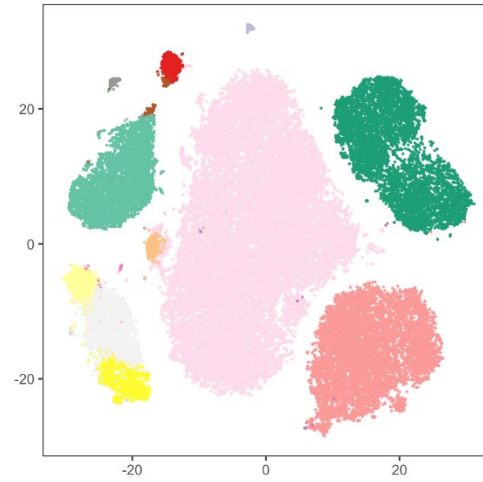


FlowSOM Clusters are Dependent on Input Parameters

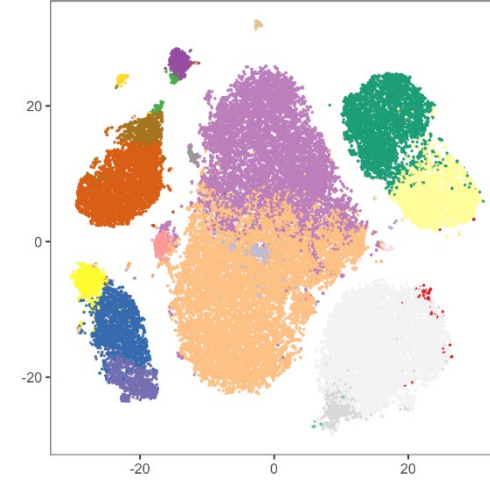
5 Clusters



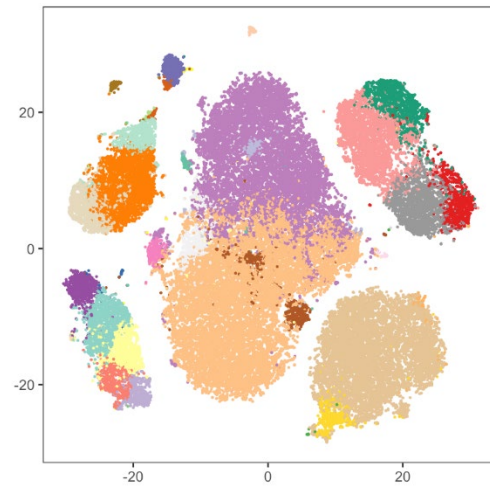
15 Clusters



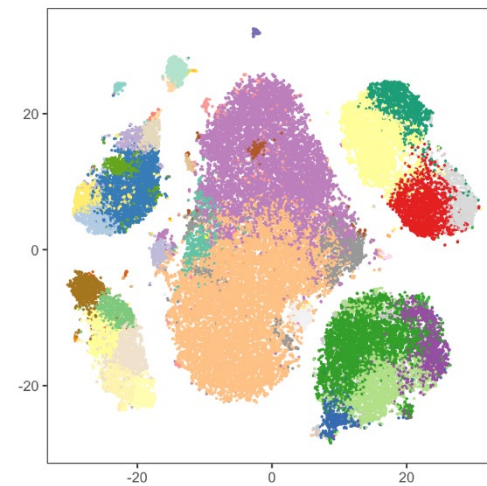
25 Clusters



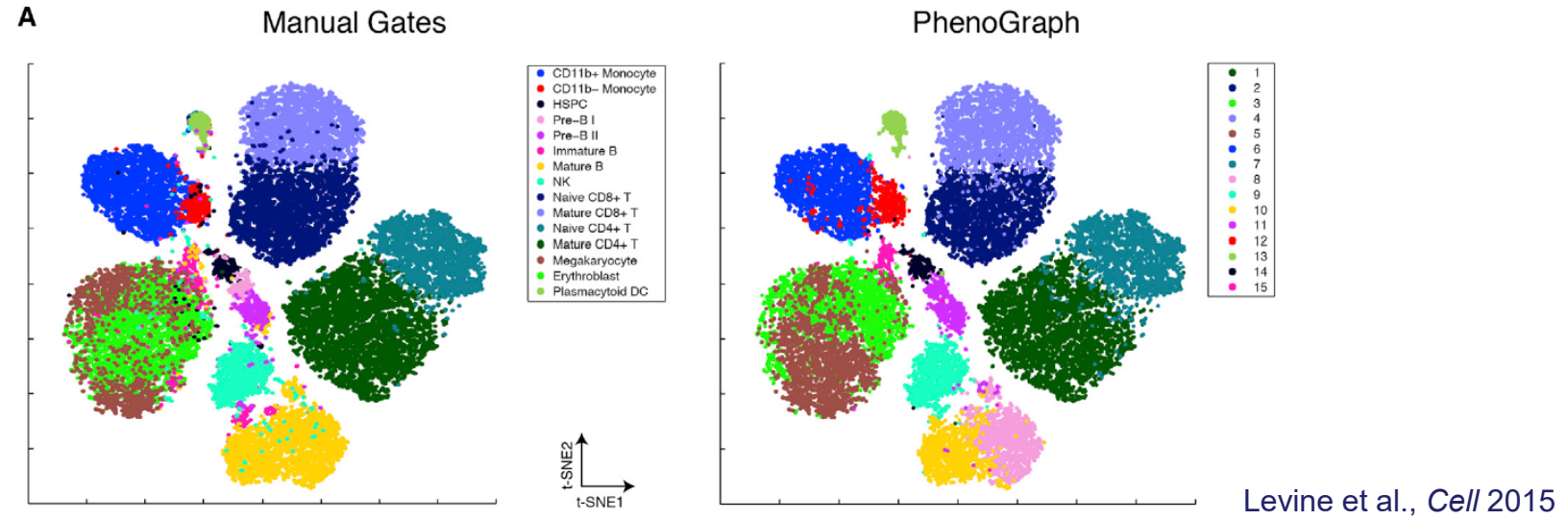
35 Clusters



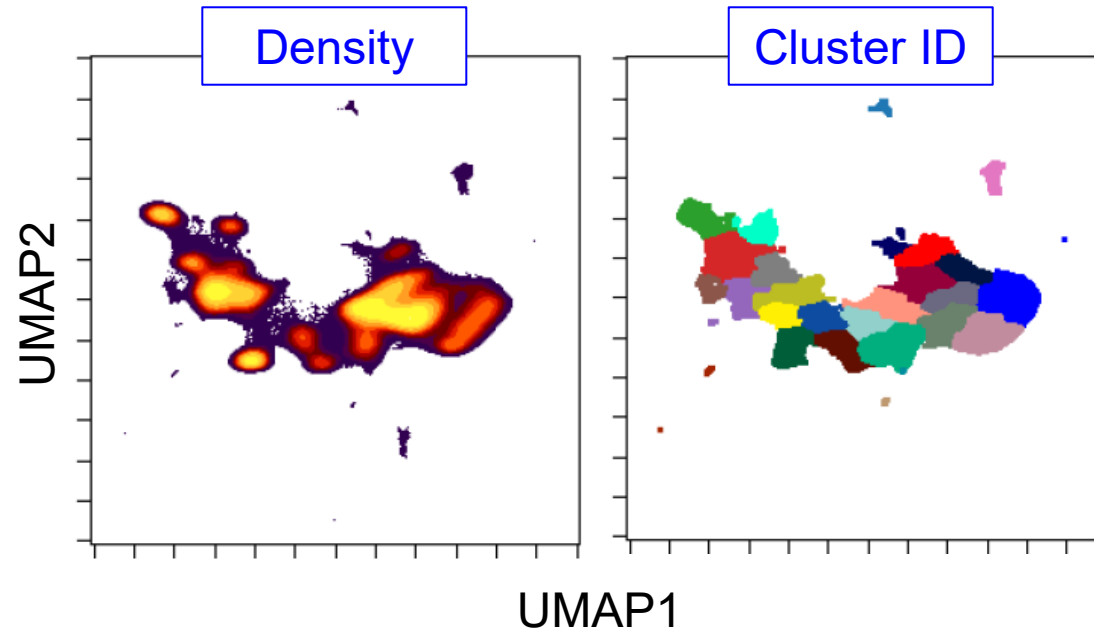
45 Clusters



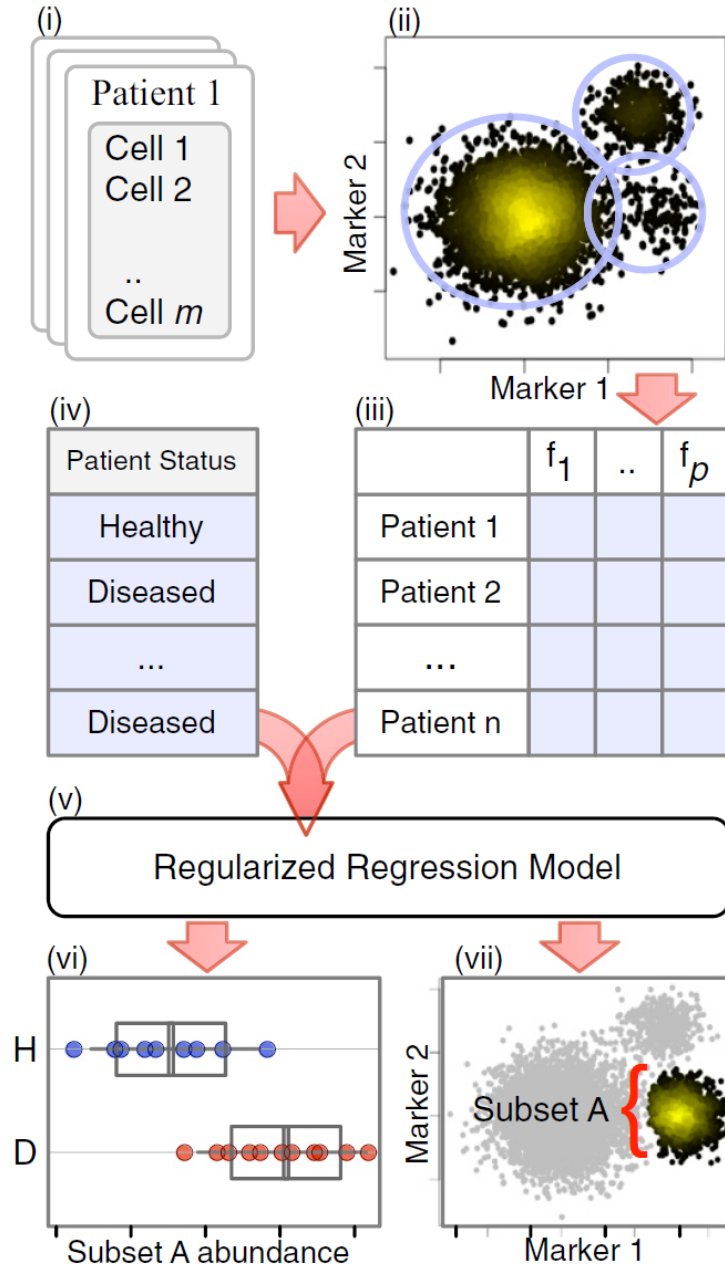
Phenograph: Clustering 35 Features => t-SNE (Not the Reverse)



Diggins: t-SNE or UMAP on Features => Clustering on 2 axes



Citrus: Supervised Population Finding

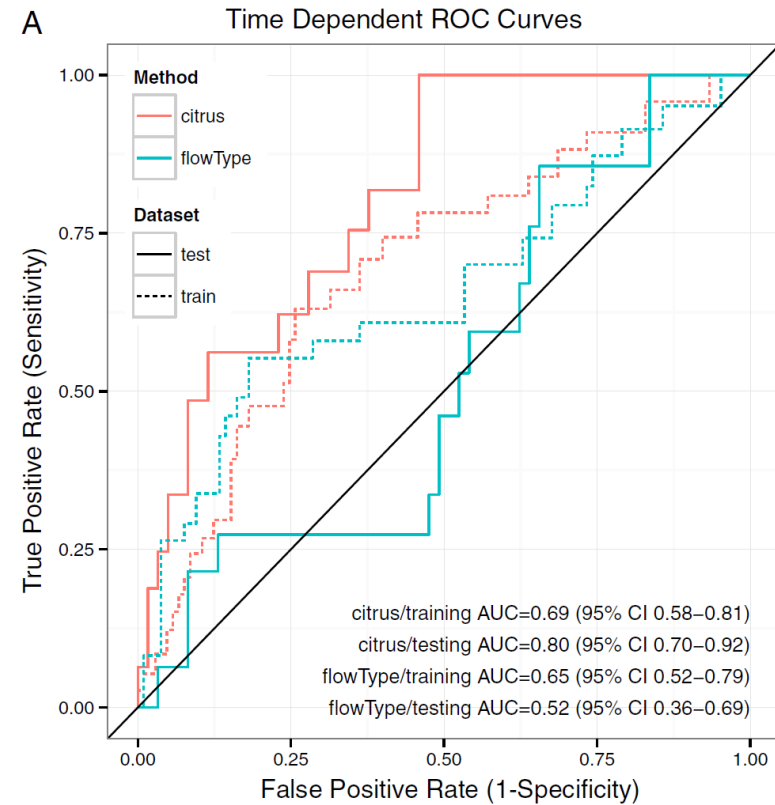


Automated identification of stratifying signatures in cellular subpopulations

Robert V. Bruggner^{a,b}, Bernd Bodenmiller^c, David L. Dill^d, Robert J. Tibshirani^{e,f,1}, and Garry P. Nolan^{b,1}

^aBiomedical Informatics Training Program, Stanford University Medical School, Stanford, CA 94305; ^bBaxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, and Departments of ^cComputer Science, ^eHealth Research and Policy, and ^fStatistics, Stanford University, Stanford, CA 94305; and ^dInstitute 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)



Citrus & RAPID Connect Cell Clusters to Clinical Outcomes, RAPID is Designed for Unsupervised Analysis of Survival

Citrus

Bruggner, Tibshirani, et al., PNAS 2014

RAPID

eLife 2020

Finding cell clusters	Unsupervised (hierarchical clustering, cells may be in 2+ clusters)	Unsupervised (various: FlowSOM, KNN, t-SNE + FlowSOM)
Determining number of cell clusters to seek	Unsupervised (must be >5% of sample)	Unsupervised (seeks few clusters w/ low internal variation)
Modeling cluster features	Supervised, multivariate (lasso regularized logistic regression, nearest shrunken centroid)	Unsupervised, univariate (median or MEM, simply a statistical description of cluster)
Splitting patients into groups	Supervised, happens at start (expert knows cut points, assigns patients to groups)	Unsupervised, happens at end (cluster abundance as cut point, Cox model of hazard)

Data Science Workflow using RAPID

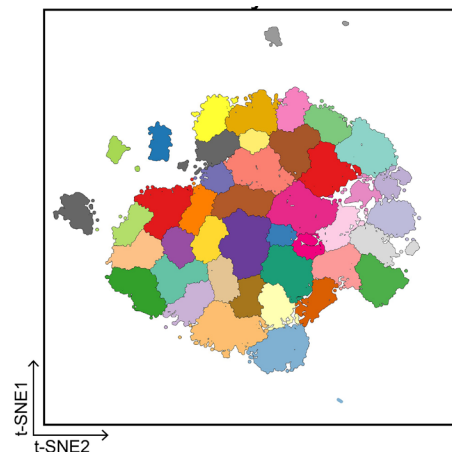
Revealing cell subsets

Dimensionality reduction

- t-SNE or UMAP

Clustering

- FlowSOM optimized



Characterizing cell subsets

Learn cell identity

- MEM

Statistical Testing

- Risk Assessment

Feature Comparison

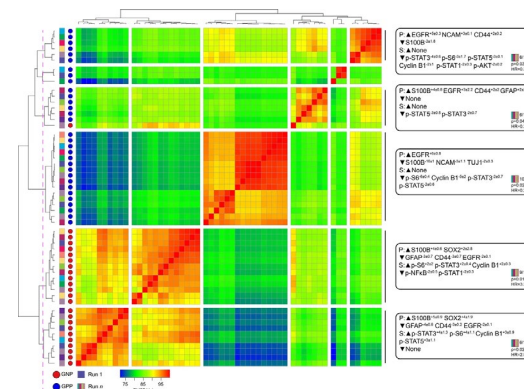
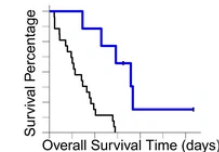
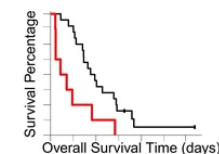
- RMSD

▲ Feature1⁺⁵ FeatureX⁺³

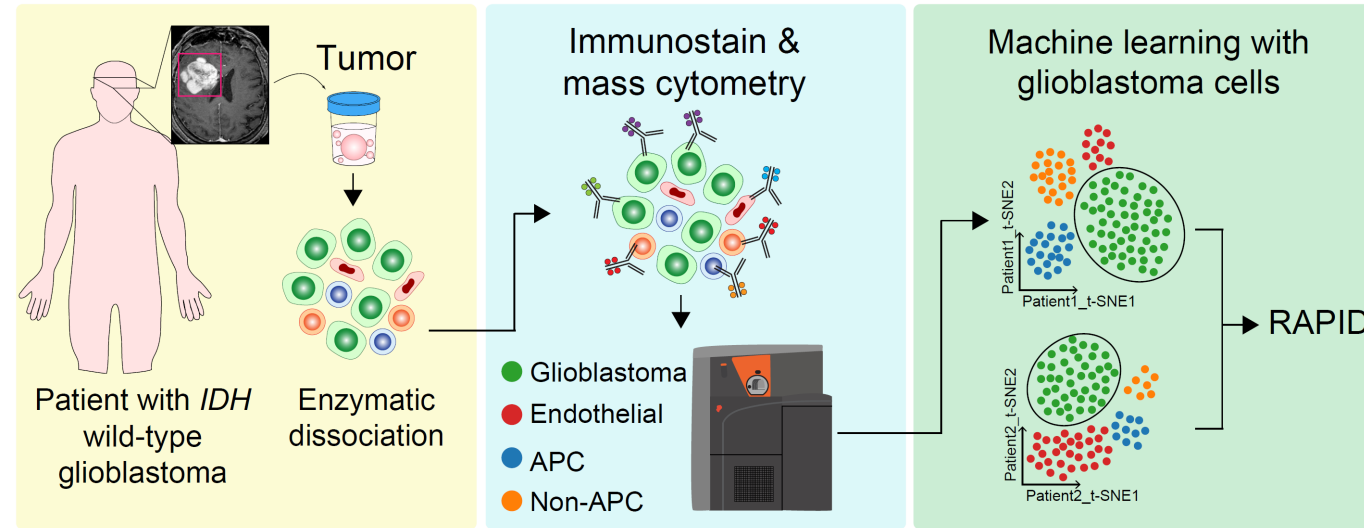
▼ Feature2⁻⁷ FeatureY⁻²

▲ Feature2⁺⁵ FeatureY⁺³

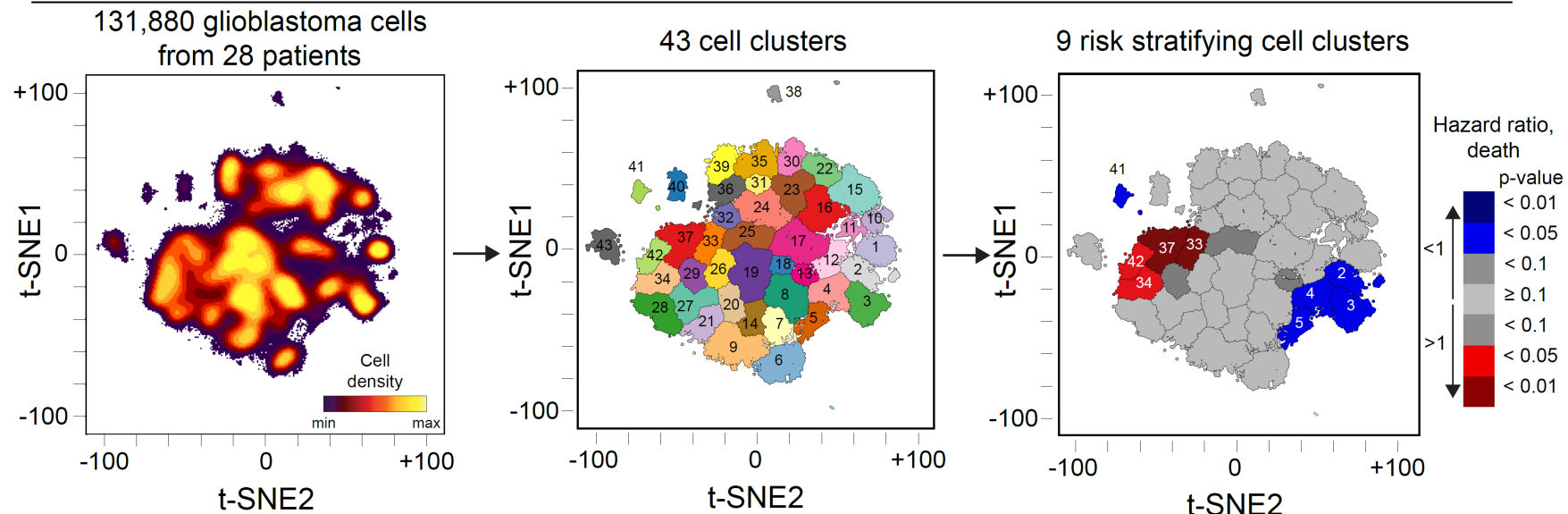
▼ Feature1⁻⁷ FeatureX⁻²



RAPID Maps Clinical Outcomes Onto Clusters (in t-SNE, UMAP, 2D image, original features, PCA, etc.)

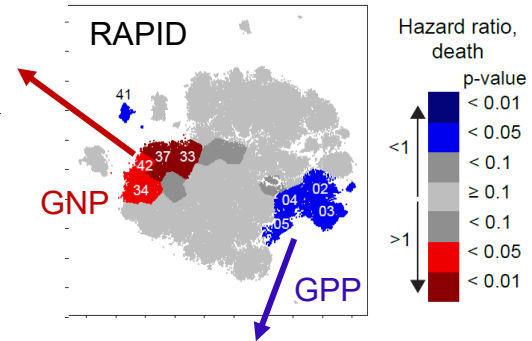
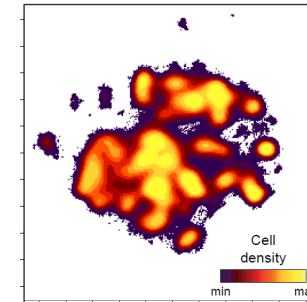
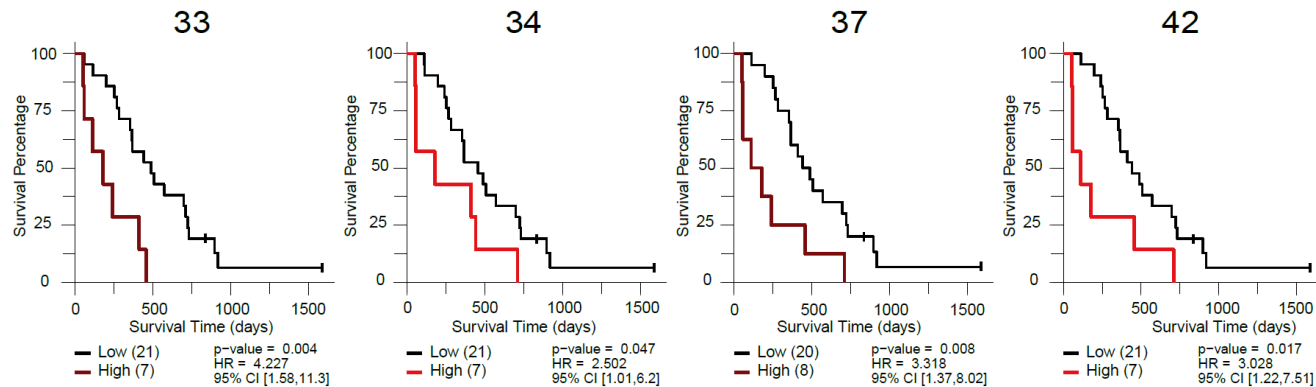


Risk Assessment Population IDentification (RAPID) Maps Outcome onto t-SNE

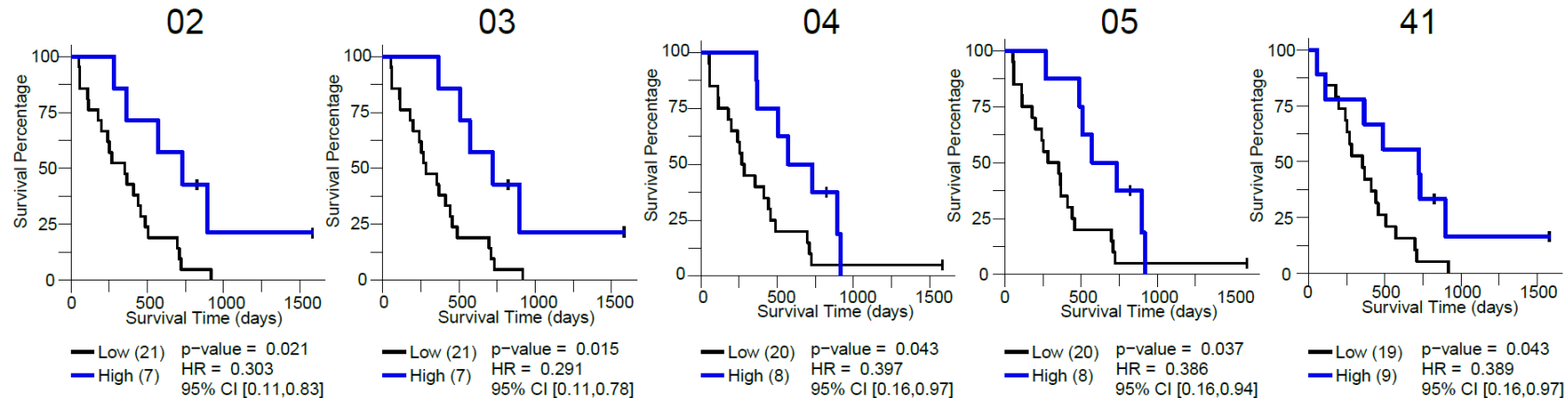


RAPID Revealed Phenotypically Distinct Risk Stratifying Glioblastoma Cell Clusters

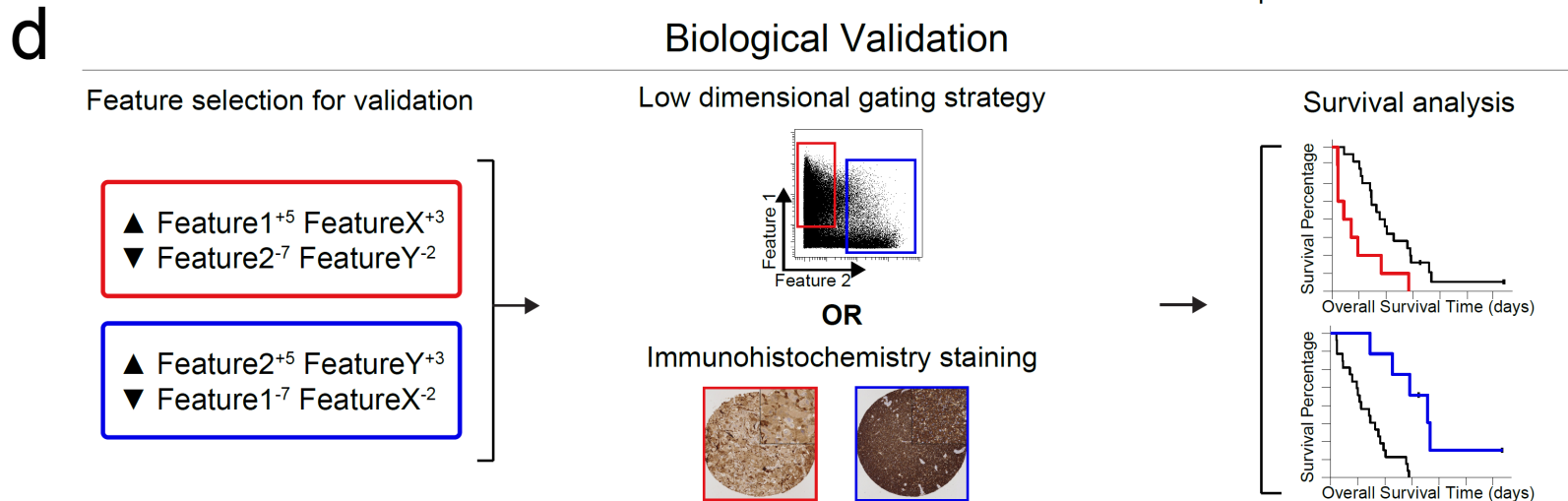
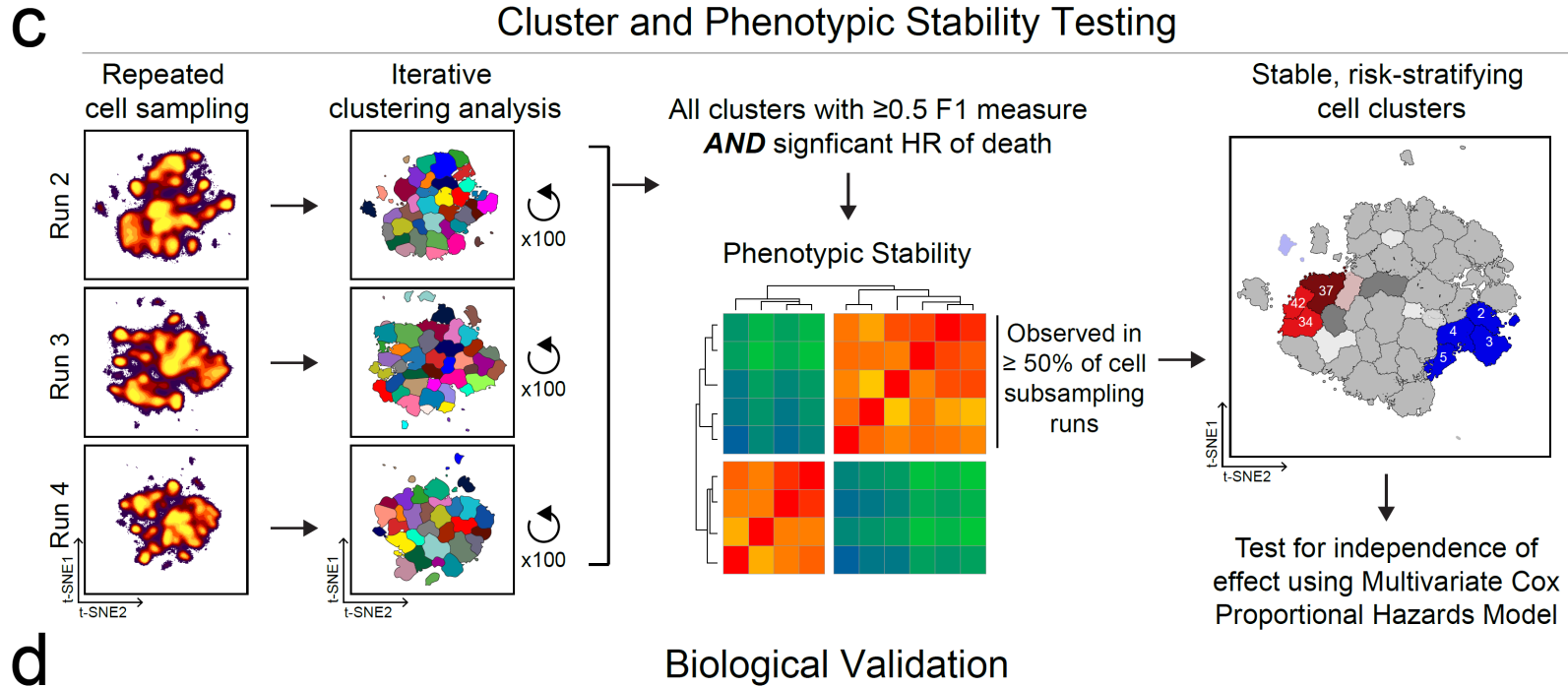
Poor survival of Glioma **Negative Prognostic (GNP)** high patients



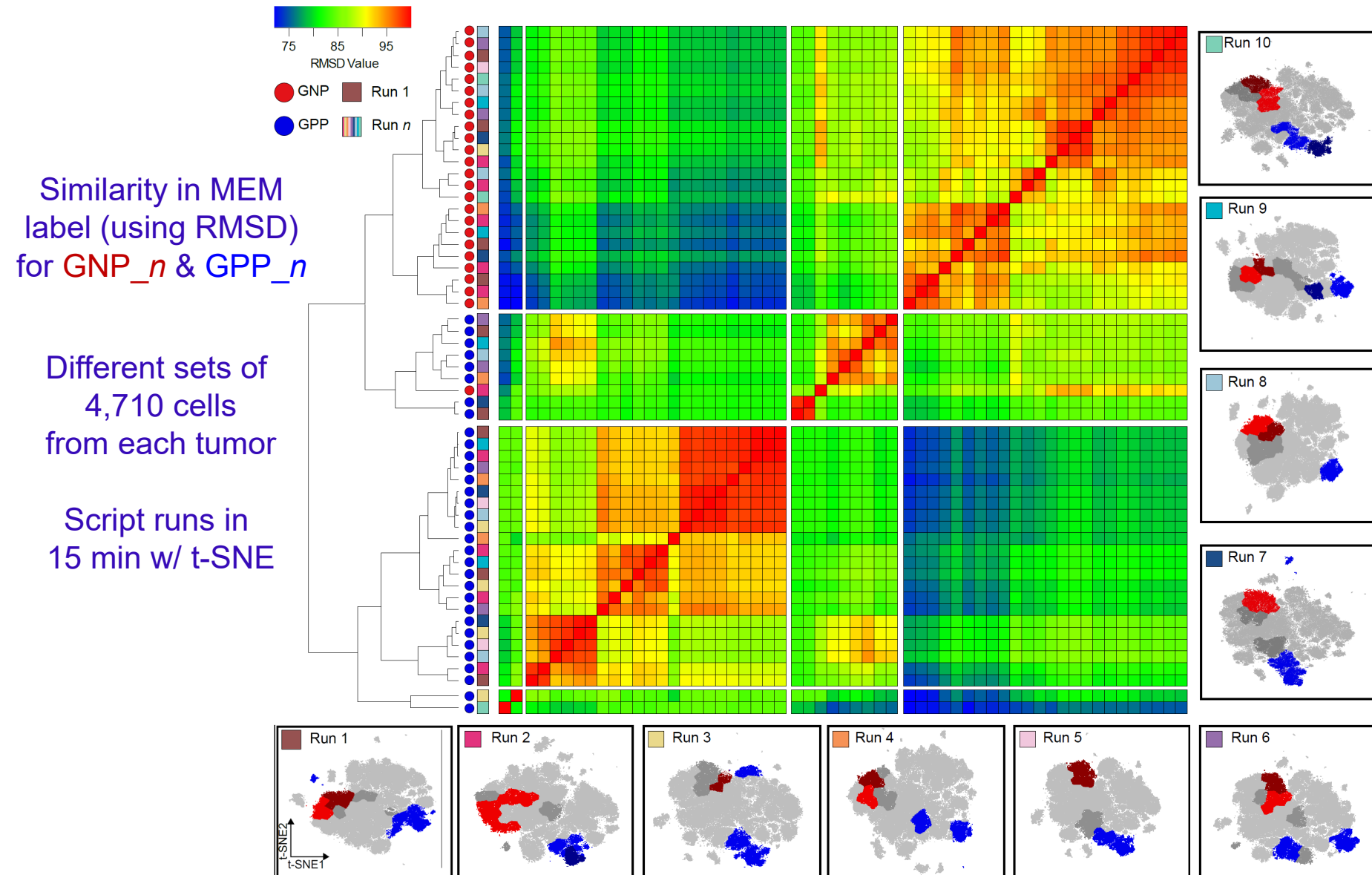
Better survival of Glioma **Positive Prognostic (GPP)** high patients



Statistical & Biological Validation Are Essential Parts of Algorithm & Study Design

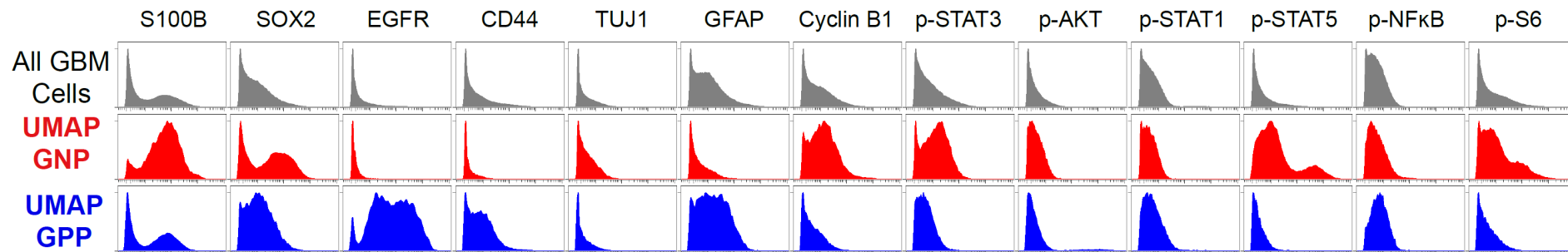
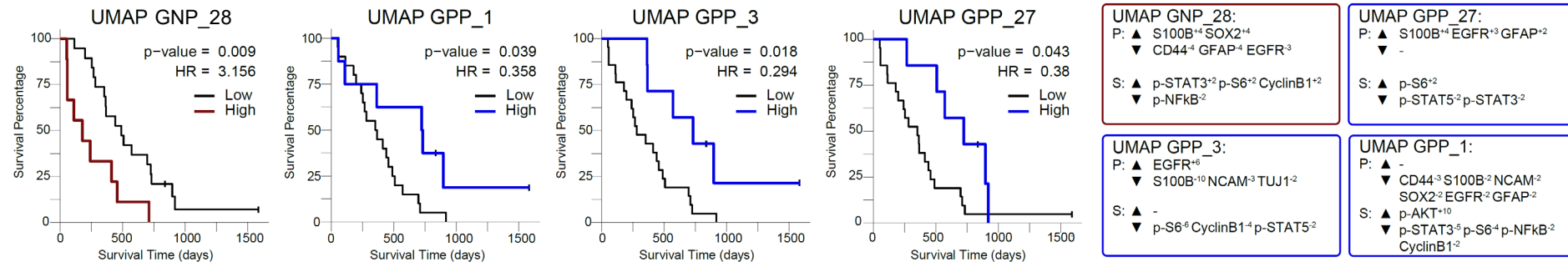
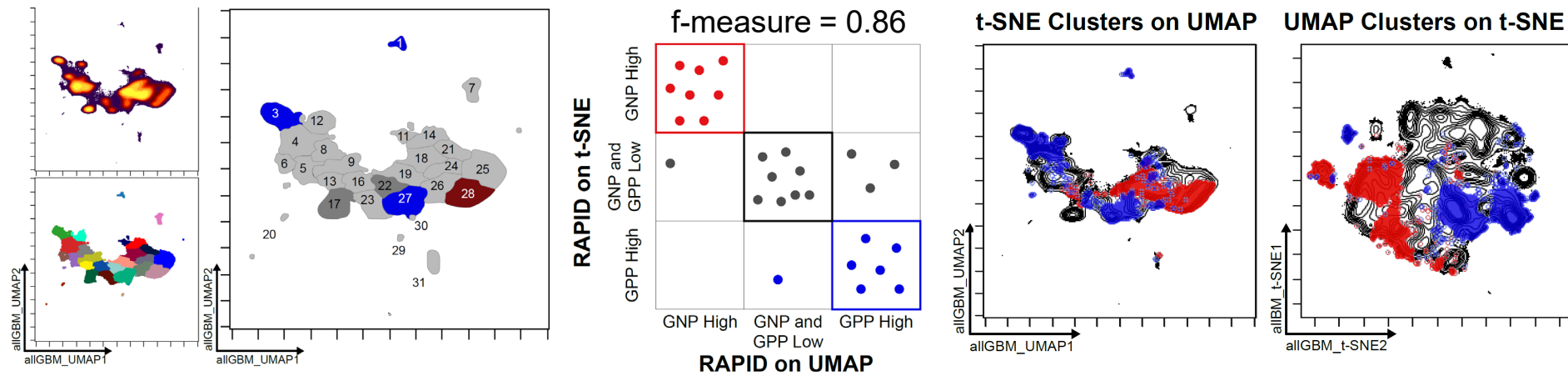


Re-Running RAPID +9X with Different Cells from the Same Tumors Gave Similar GNP & GPP Phenotypes and Risk Stratification



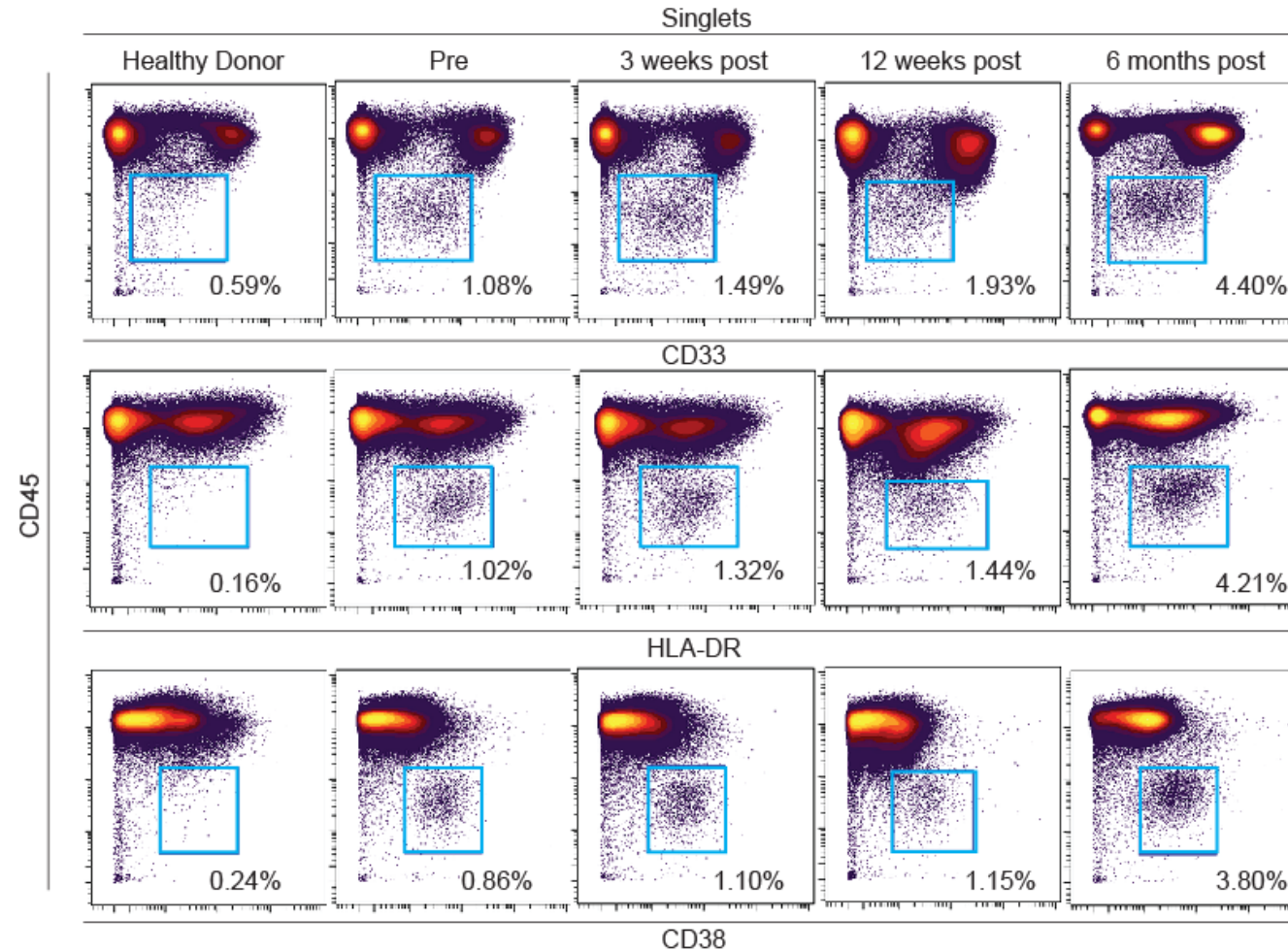
Re-Running RAPID with UMAP Instead of t-SNE

Gave Similar GNP & GPP Phenotypes and Risk Stratification



A Case Study: Systems Immune Monitoring with Mass Cytometry Reveals A Clinically Significant Rare Cell Subset

MDS in Melanoma Patient Revealed During α -PD-1 Therapy



Healthy donor looks similar to melanoma in 2D views

At Pre-Tx, MDS blasts were not detected by standard CBC

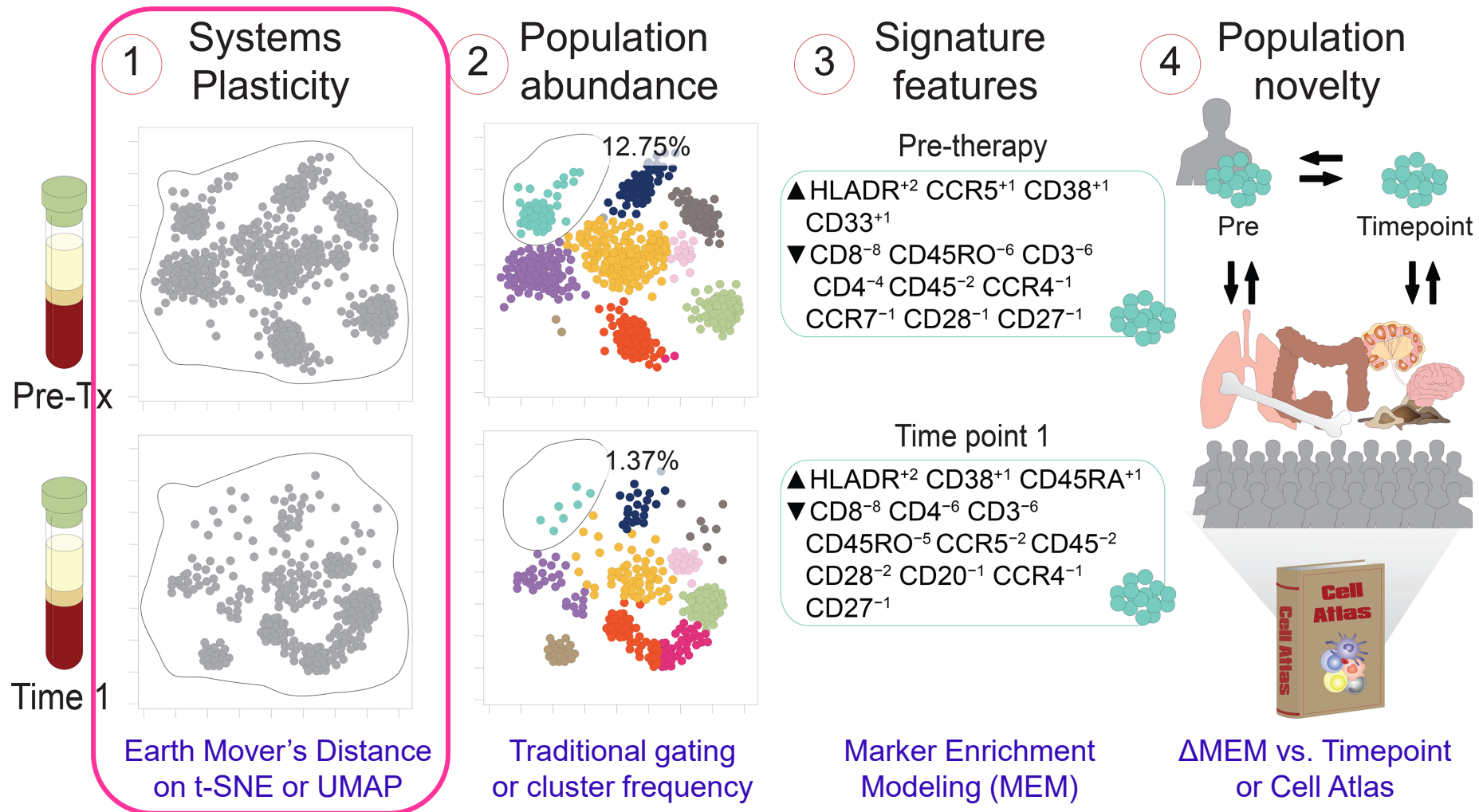
High dimensional panel allowed review of PD-1 on MDS blasts w/ existing data

Mass cytometry data (CyTOF)

Clinical Trial Monitoring: What Do We Need to Know?

Automate Four Key Readouts vs. Clinical Outcomes

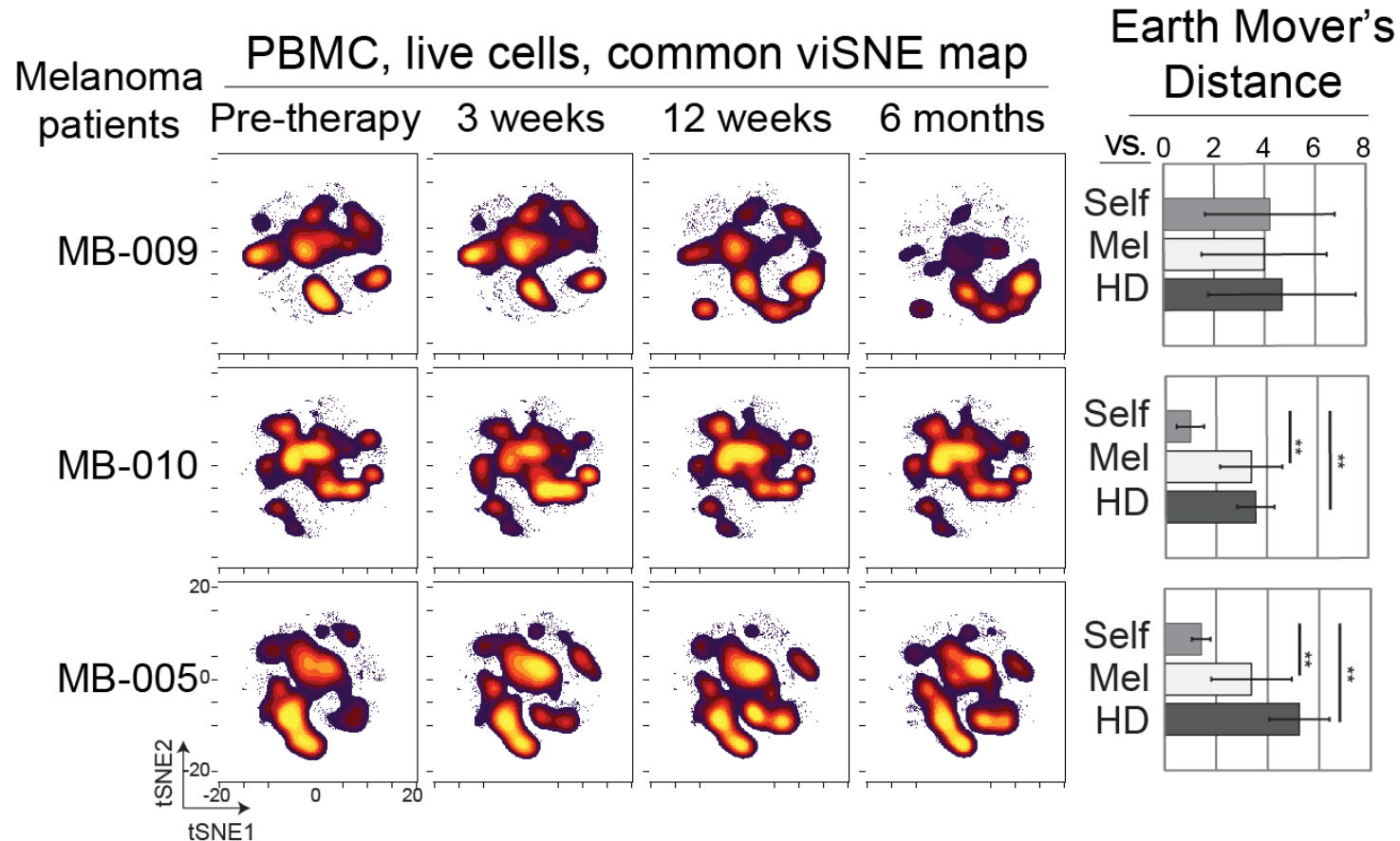
Features of Dynamic Populations



How we quantified

Plasticity / Stability: Earth Mover's Distance Quantifies Change Over Time Within a t-SNE Analysis

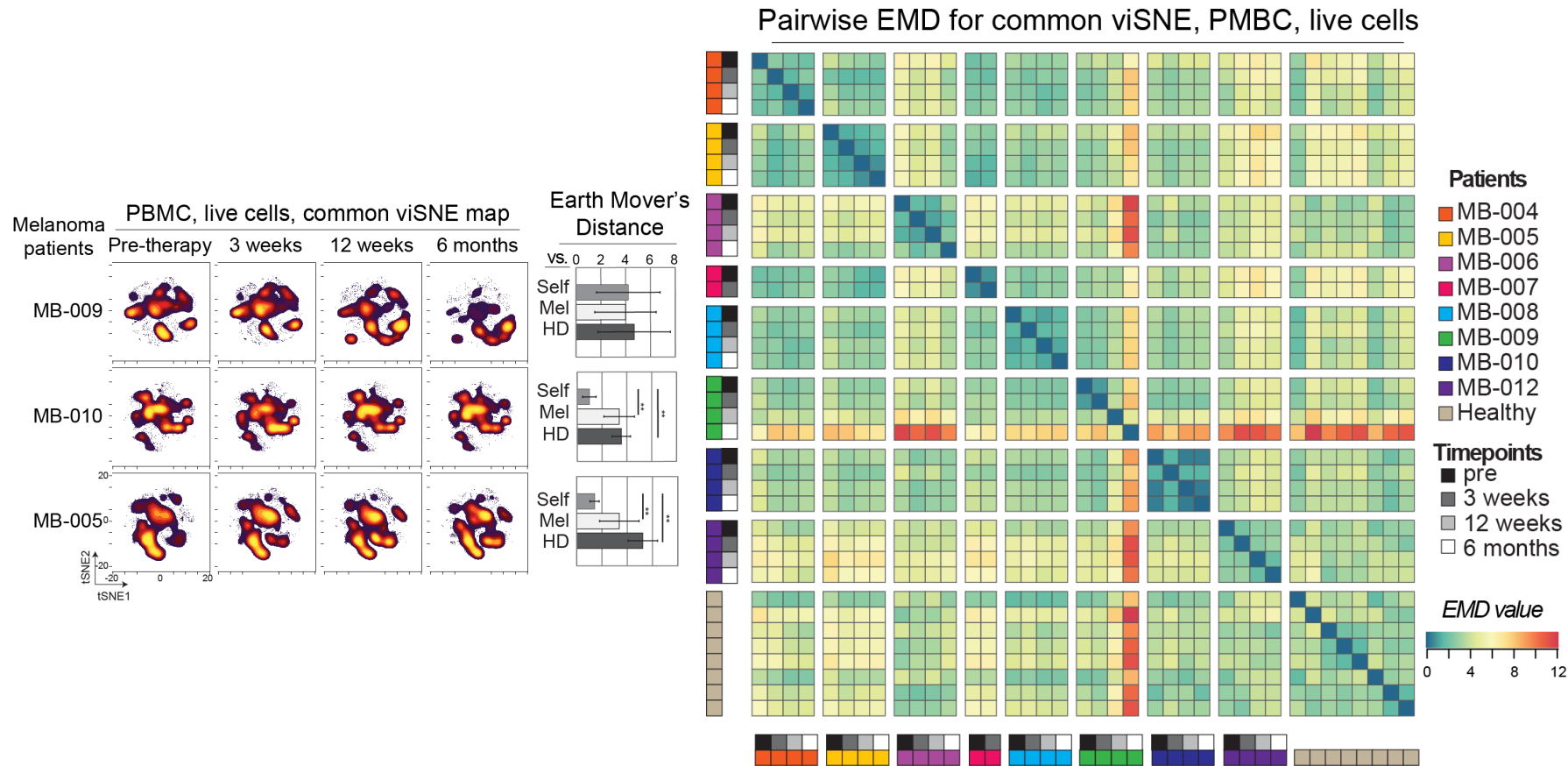
Melanoma Patients Treated with α -PD-1 Therapy, Monitored by Mass Cytometry



Systems immune monitoring reveals an unexpected pattern in MB-009
Individuals can be their own significantly stable baseline

Plasticity / Stability: Earth Mover's Distance Quantifies Change Over Time Within a t-SNE Analysis

Melanoma Patients Treated with α -PD-1 Therapy, Monitored by Mass Cytometry

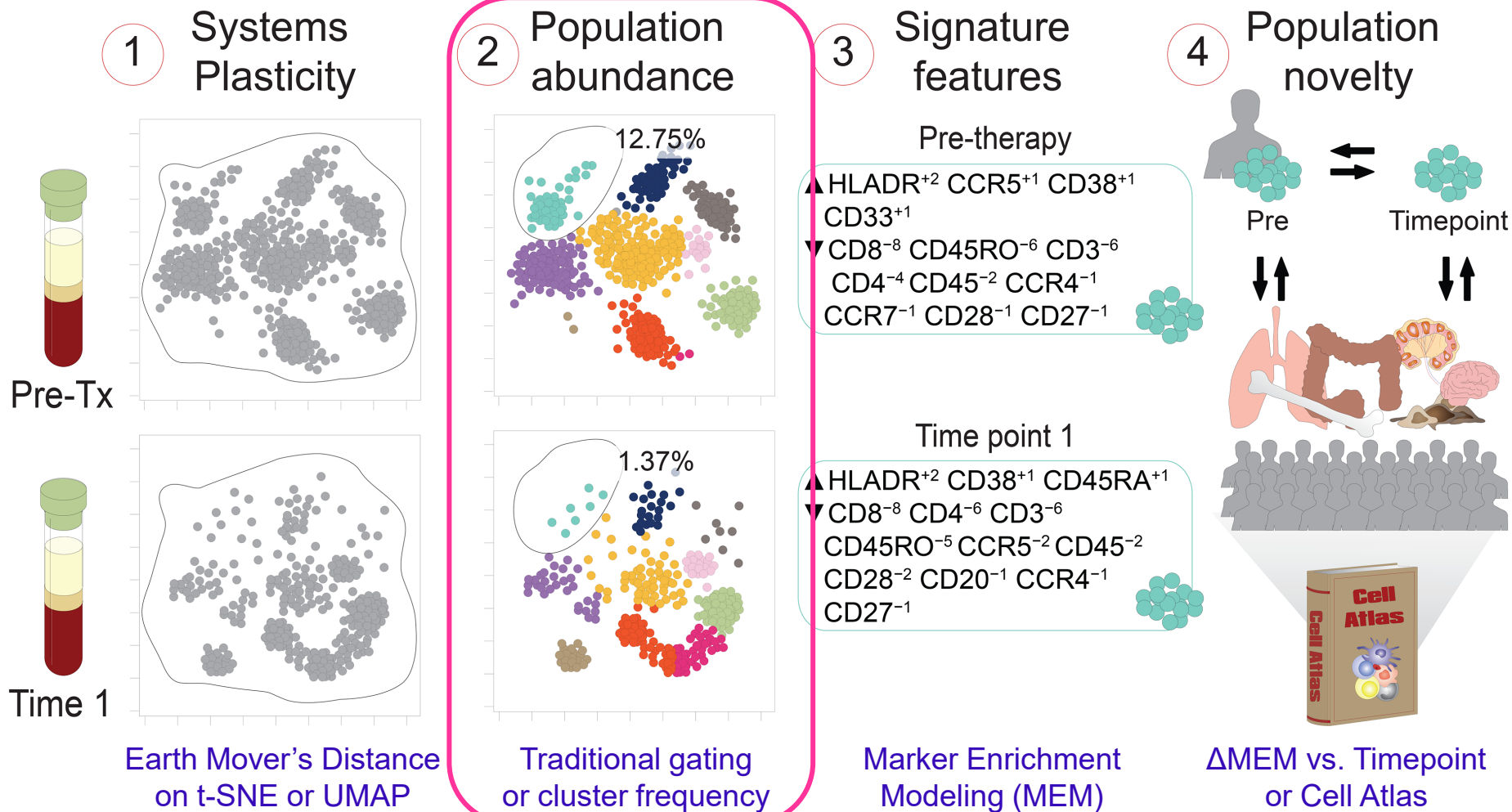


Systems immune monitoring reveals an unexpected pattern in MB-009

Clinical Trial Monitoring: What Do We Need to Know?

Automate Four Key Readouts vs. Clinical Outcomes

Features of Dynamic Populations

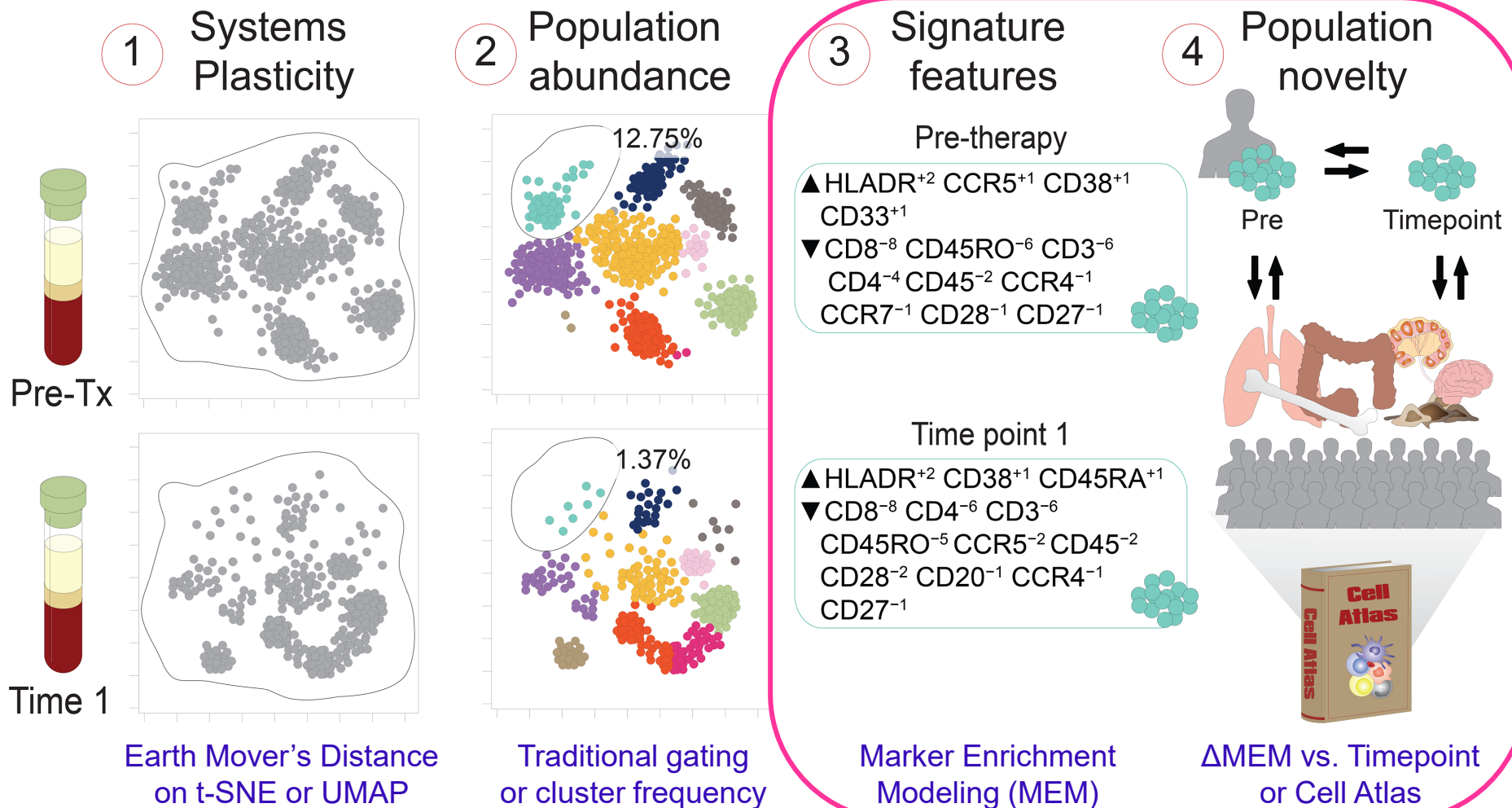


How we quantified

Clinical Trial Monitoring: What Do We Need to Know?

Automate Four Key Readouts vs. Clinical Outcomes

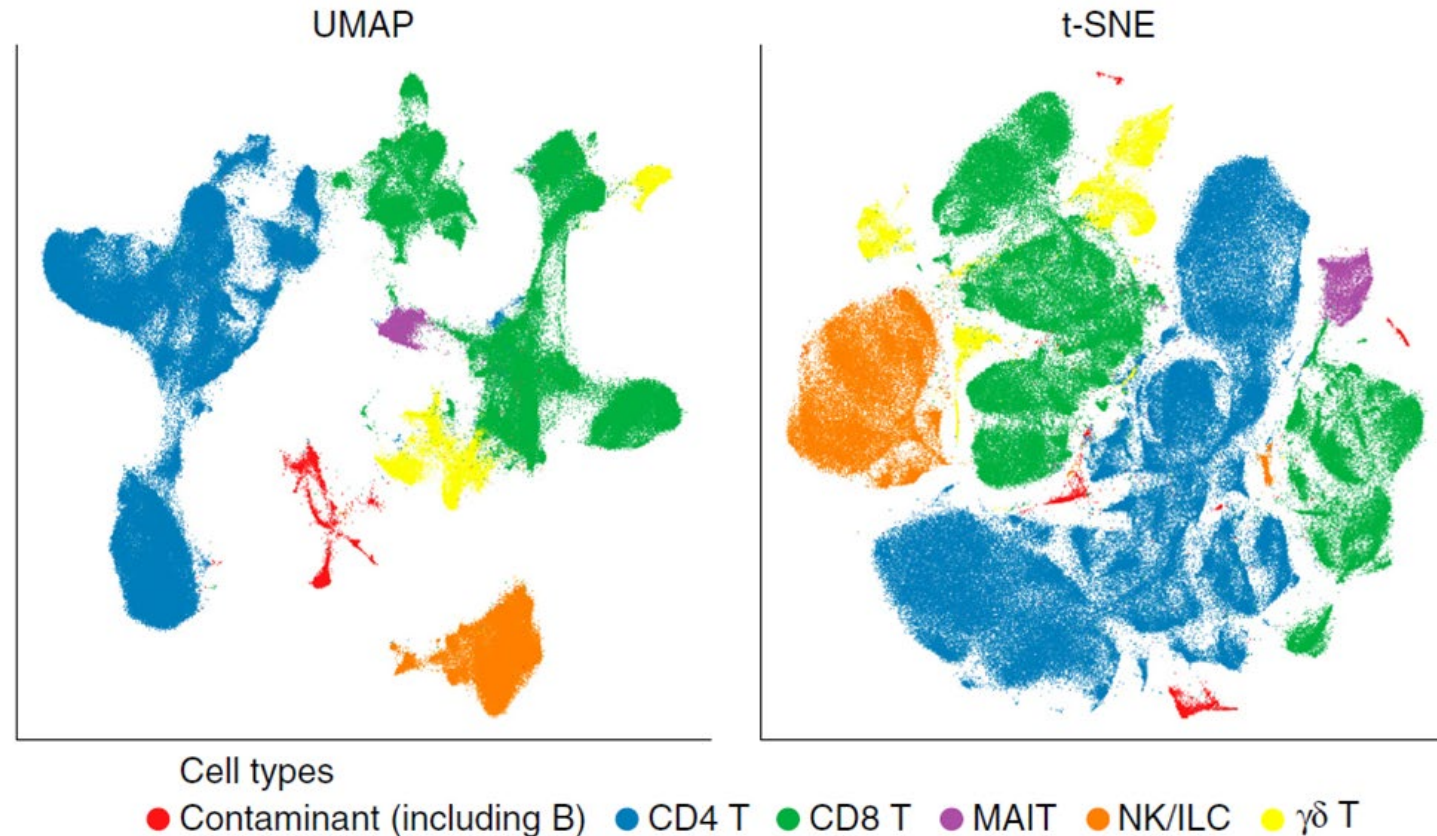
Features of Dynamic Populations



How we quantified

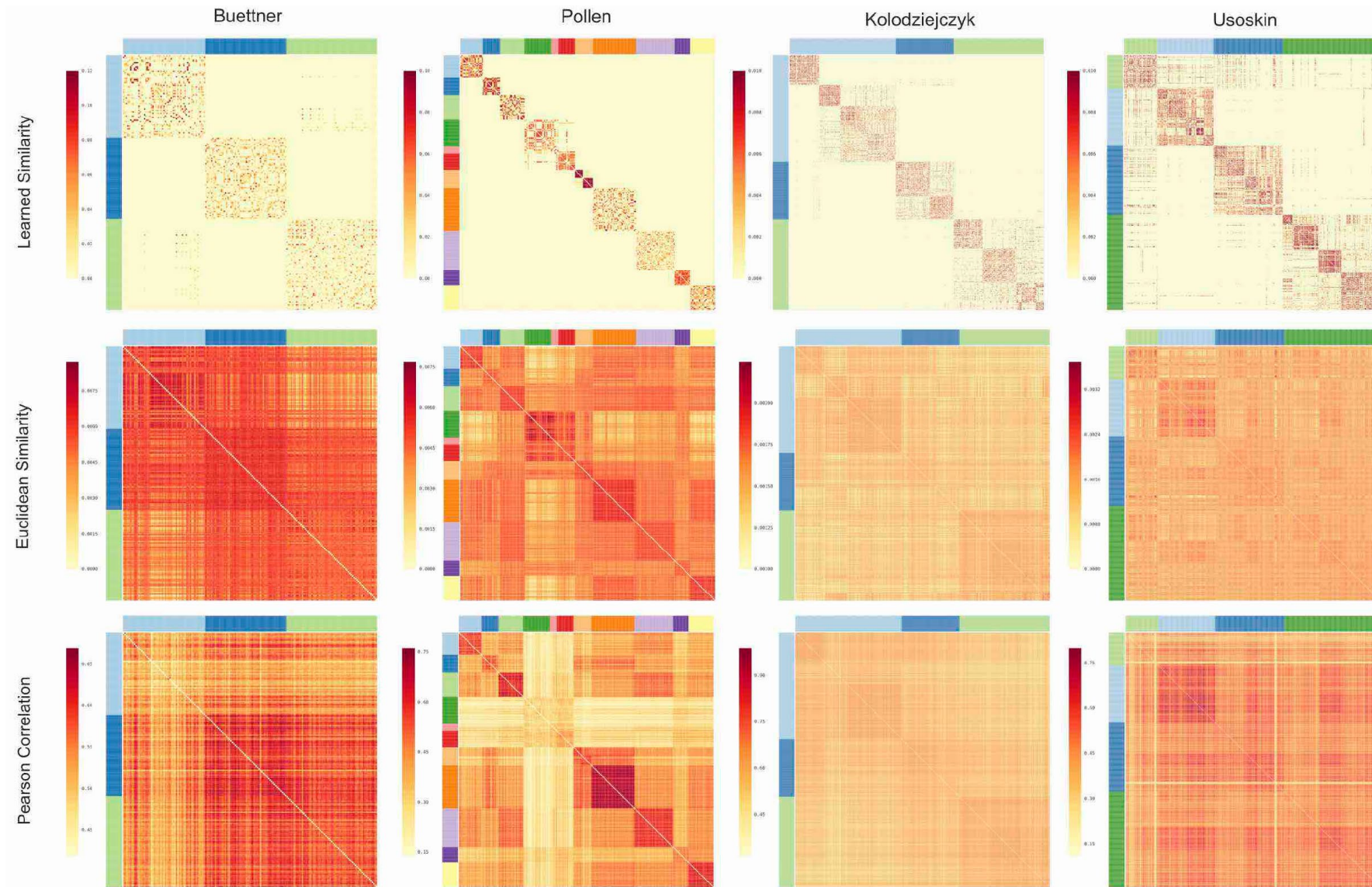
Becht et al., UMAP Preserves Local and Global Structure (Analysis of Tissue T Cells; Color = Expert Knowledge / Source)

(a) UMAP better split CD8 T cells, $\gamma\delta$ T cells, and contaminating cells



Dataset covering 35 samples originating from 8 distinct human tissues enriched for T and natural killer (NK) cells, of more than >300,000 cell events with 39 protein targets (Wong et al. dataset).

Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning (SIMLR)



Resources

Normalization

<https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.22271>

Gaussian Gating

<http://cytoforum.stanford.edu/download/file.php?id=242&sid=37e5ec0a3dedb53865bbbcb6a023c316>

t-SNE

<https://www.nature.com/articles/nbt.2594>

Opt-SNE

<https://www.biorxiv.org/content/10.1101/451690v3.full>

UMAP

<https://www.nature.com/articles/nbt.4314>

FlowSOM

<https://www.ncbi.nlm.nih.gov/pubmed/25573116>

SPADE

<https://www.nature.com/articles/nbt.1991>

Phenograph

<https://www.sciencedirect.com/science/article/pii/S0092867415006376>

MEM

<https://www.nature.com/articles/nmeth.4149>

RAPID

<https://elifesciences.org/articles/56879>

T-REX

<https://elifesciences.org/articles/64653>

“A Beginner’s Guide to Analyzing and Visualizing Mass Cytometry Data”

<https://www.jimmunol.org/content/200/1/3>

Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data

<https://www.ncbi.nlm.nih.gov/pubmed/27992111>

Contact Info

Jonathan Irish

Principal Investigator, Associate Professor

Cell & Developmental Biology Pathology, Microbiology & Immunology

Email: jonathan.irish@vanderbilt.edu

Cass Mayeda

Web Apps Research Assistant

Email: cass.mayeda@vanderbilt.edu