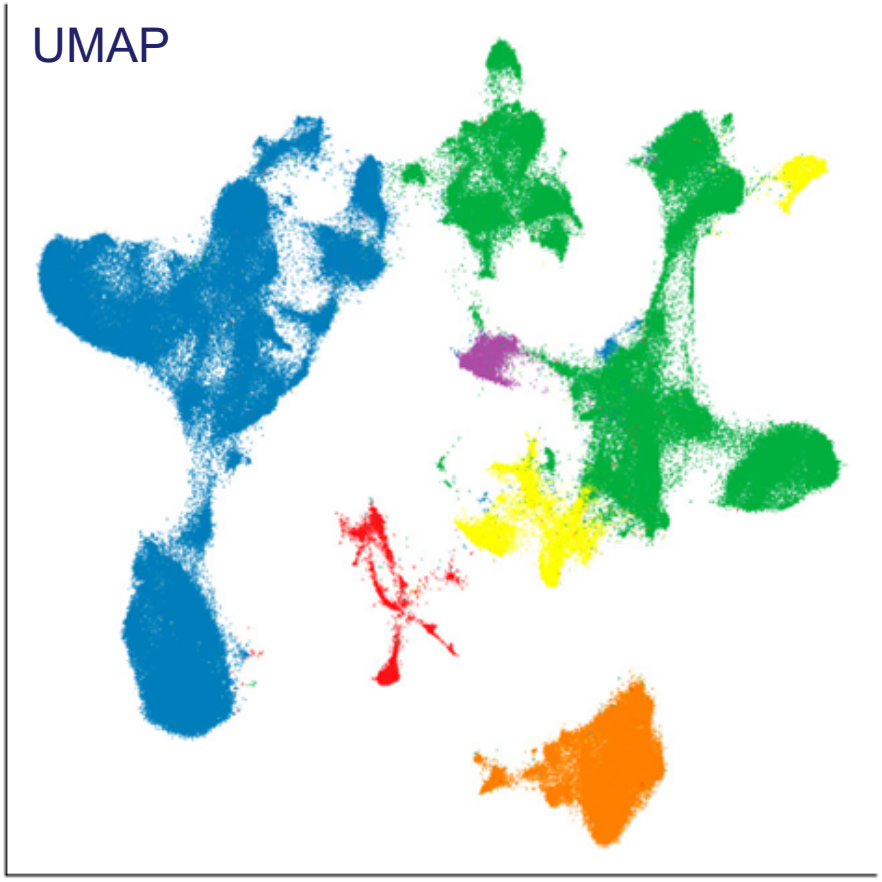
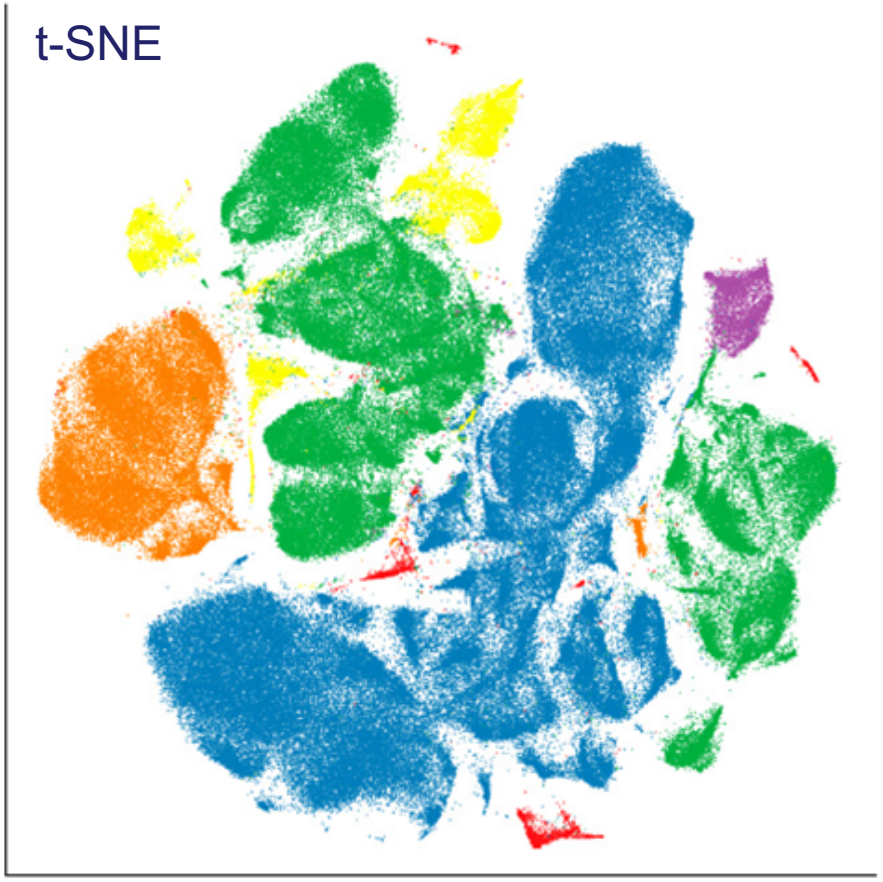
Data Science Methods for Quantitative Biology





Becht et al. 2018

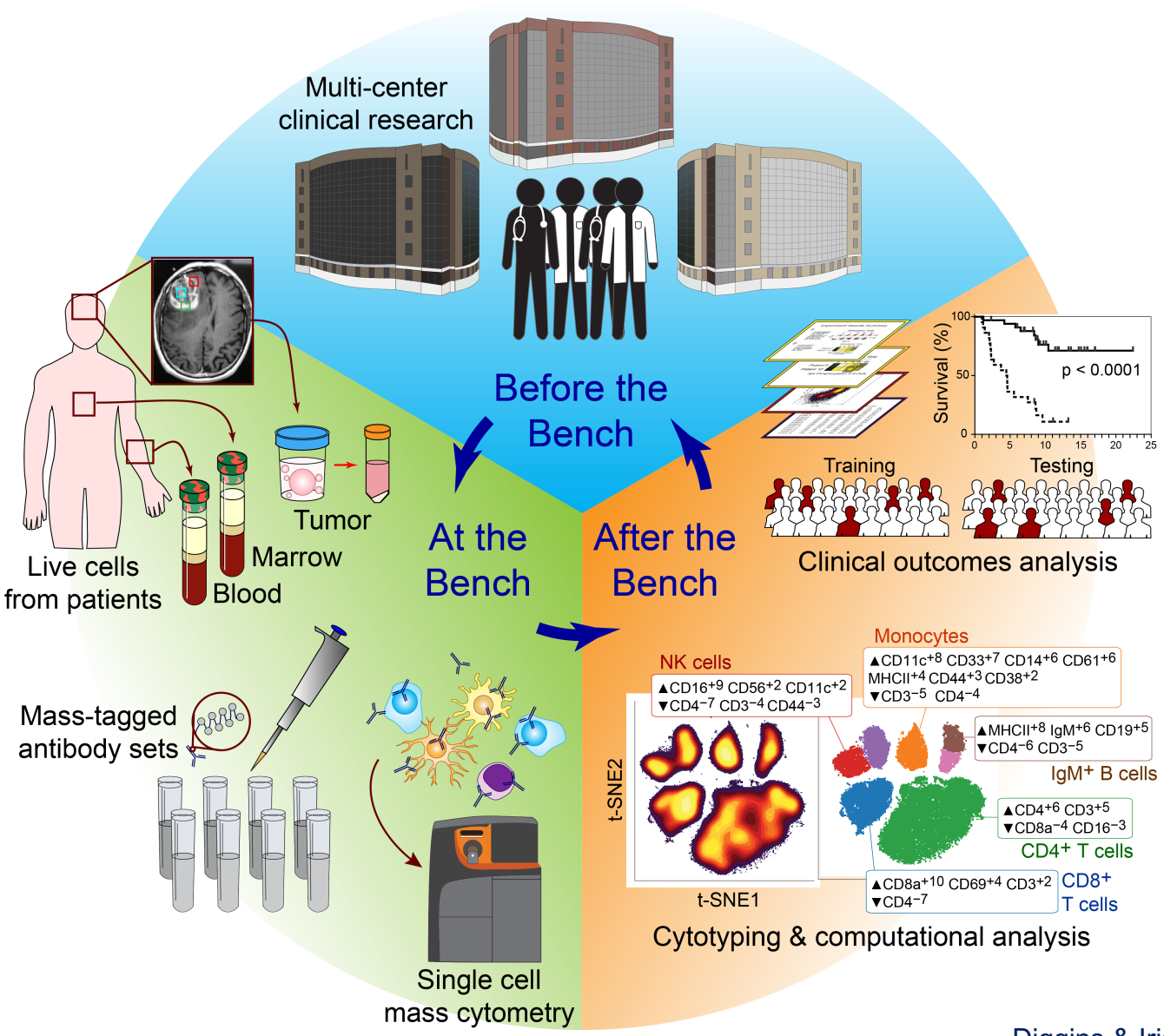
Jonathan Irish & Sierra Barone

Associate Professor
Cell & Developmental Biology
Pathology, Microbiology & Immunology

Data Science Research Assistant
Irish Lab, Cancer & Immunology Core

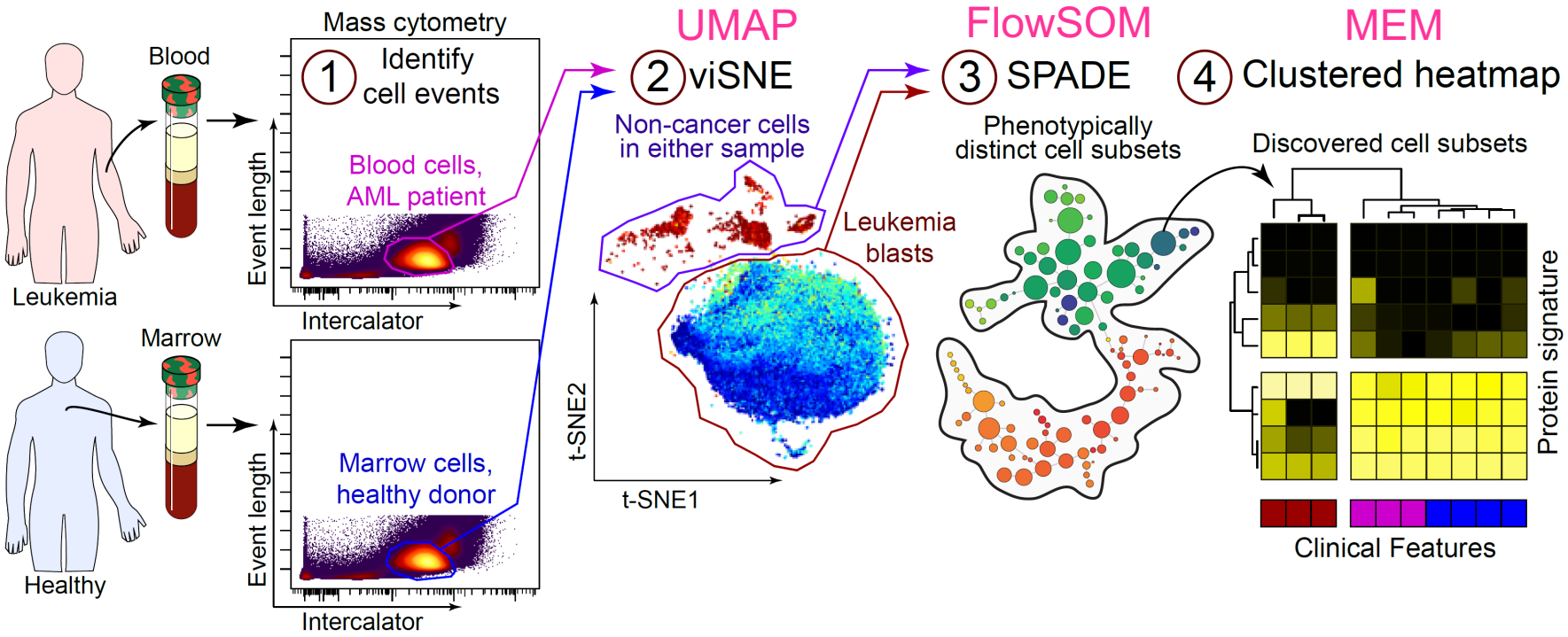
Disclosures:
Co-founded Cytobank (Beckman Coulter)
Mass Cytometry Center of Excellence w/ Fluidigm
Clinical research w/ Incyte, Janssen, Pharmacyclics

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



Diggins & Irish et al., *in review*

Machine Learning Is a Key Skillset for Biologists (And the Tools Are Rapidly Evolving)



Protocol: Diggins et al., *Current Protocols in Cytometry* 2017
Data files: https://flowrepository.org/id/FR-FCM-ZZKZ
More great tools: Saeys et al, *Nature Reviews Immunology* 2016

Original research:
Diggins et al., *Methods* 2015
Diggins et al., *Nature Methods* 2017

Typical workflow and goal: learn & label cytotypes (cell identities), reveal and assess unexpected & abnormal cells

Need: human reference data (more examples) with annotations

Effective data analysis is critical in biology, and this means working *with* computational tools that reveal and model patterns in data

Discussion Questions We Expect to Cover

1) What are key differences between tools (t-SNE/viSNE, SPADE, UMAP, FlowSOM, PCA, MEM, Citrus, etc.)? 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 all the settings do (e.g., t-SNE iterations, perplexity, SPADE downsampling & node #)? When should they be changed?

4) How does one compare new samples with a prior analysis? How do we test tools with expert gating?

5) What are some "red flags" indicating problems? What does a good t-SNE, UMAP, FlowSOM, or other analysis run look like?

But first: what is data science?

Irish lab view of data science:

Systematically varying analytical elements in order to test a hypothesis

(Varied analytical elements might be different data types, data sub-samples, different initial assumptions, contrasting analytical tools, input parameters, etc.)

It's relatively new that datasets are robust enough to enable mining & exploration.

Rumsfeldian Data Science

Known knowns: What do you know about your system?

Known unknowns: What do you know remains to be learned?

Unknown unknowns: What don’t you know you don’t know?

Donald Rumsfeld (Feb 12, 2002): Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones.

Socratic Data Science

Known knowns: What do you know about your system?

Known unknowns: What do you know remains to be learned?

Unknown unknowns: What don't you know you don't know?

Unknown knowns: What don't you know, but think you do?
i.e. Which 'priors' are incorrect?

If you fear incorrect priors, unsupervised analysis may be able to help.

Socrates according to Plato's *Apology*: I am wiser than this man, for neither of us appears to know anything great and good; but he fancies he knows something, although he knows nothing; whereas I, as I do not know anything, do not fancy I do. In this trifling particular, then, I appear to be wiser than he, because I do not fancy I know what I do not know.

Defining Your System

1) Elements, the studied units of the system.

- Patients, cells, images, pixels, transcripts, genomes, peptides.
- We will envision elements as "rows" in a spreadsheet.

2) Features, the things measured for each element.

- Clinical outcomes, phospho-proteins, pixel density, nucleotides.
- We will envision features as "columns" in a spreadsheet.
- Feature selection may rely on hypotheses, rules, or prior knowledge.

3) Scales, the type & range of the measurements for each feature.

- Categorical, linear, log & base, arcsinh & cofactor.
- -150 to 262,144; 1 to 10,000; 0 to 50; 1 to 100; 0 to 1; NR, PR, CR.
- Will largely explore the data without units until we create reports.

4) Prior knowledge, the things assumed to be known for the system.

- Organization of elements (groups, order, etc.), feature relationships.
- Supervised analysis explicitly uses prior knowledge.
- Unsupervised analysis looks for patterns without prior knowledge.

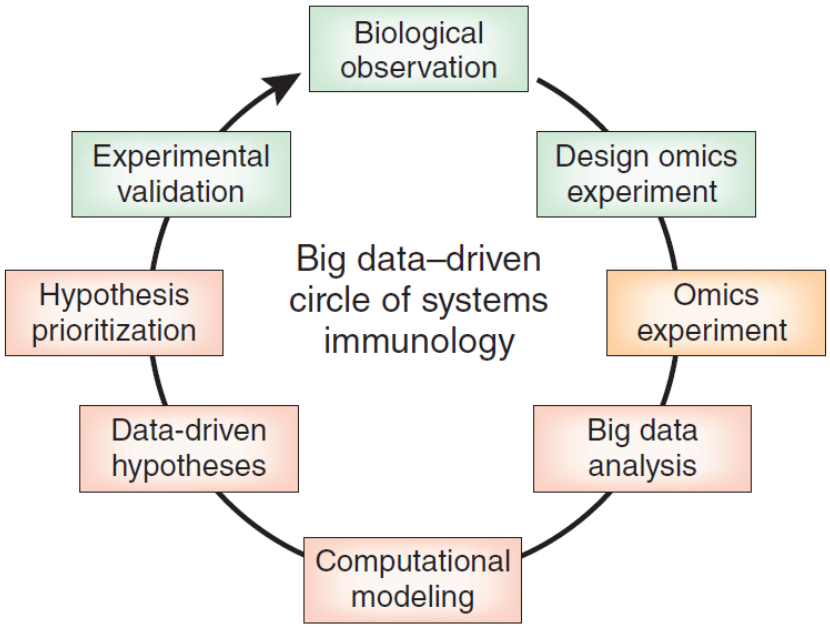
The data type for today: cytometry
(quantitative single cell measurements)

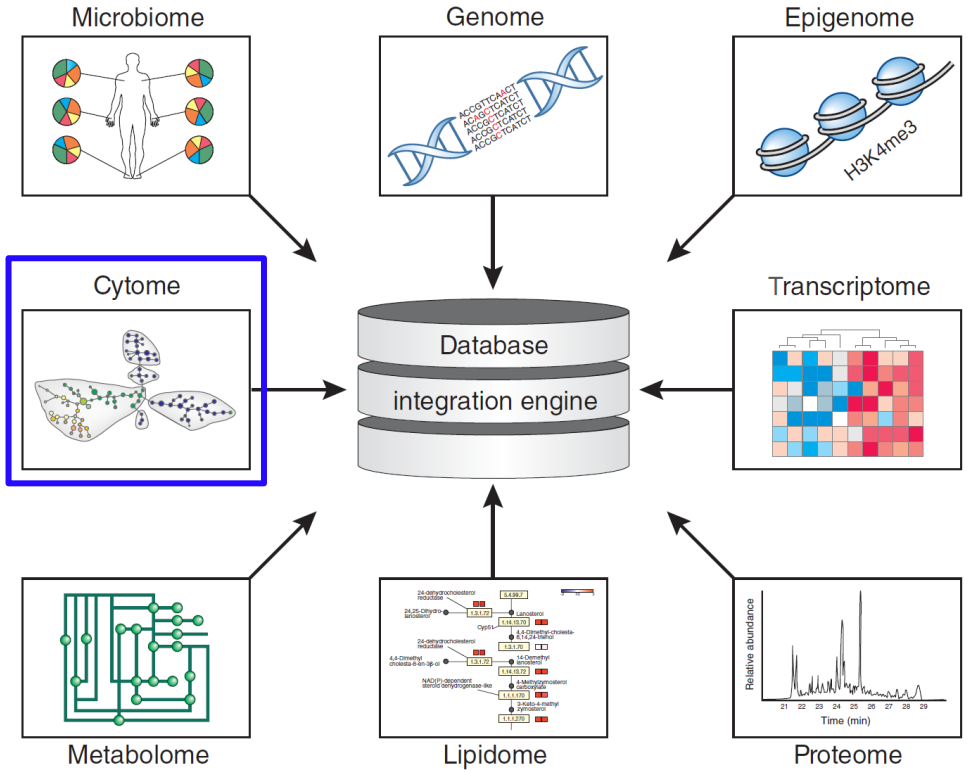
Cytomics: The 'Omics of Cells & Cell Identity

Teaching 'big data' analysis to young immunologists

NATURE IMMUNOLOGY
VOLUME 16 NUMBER 9 SEPTEMBER 2015

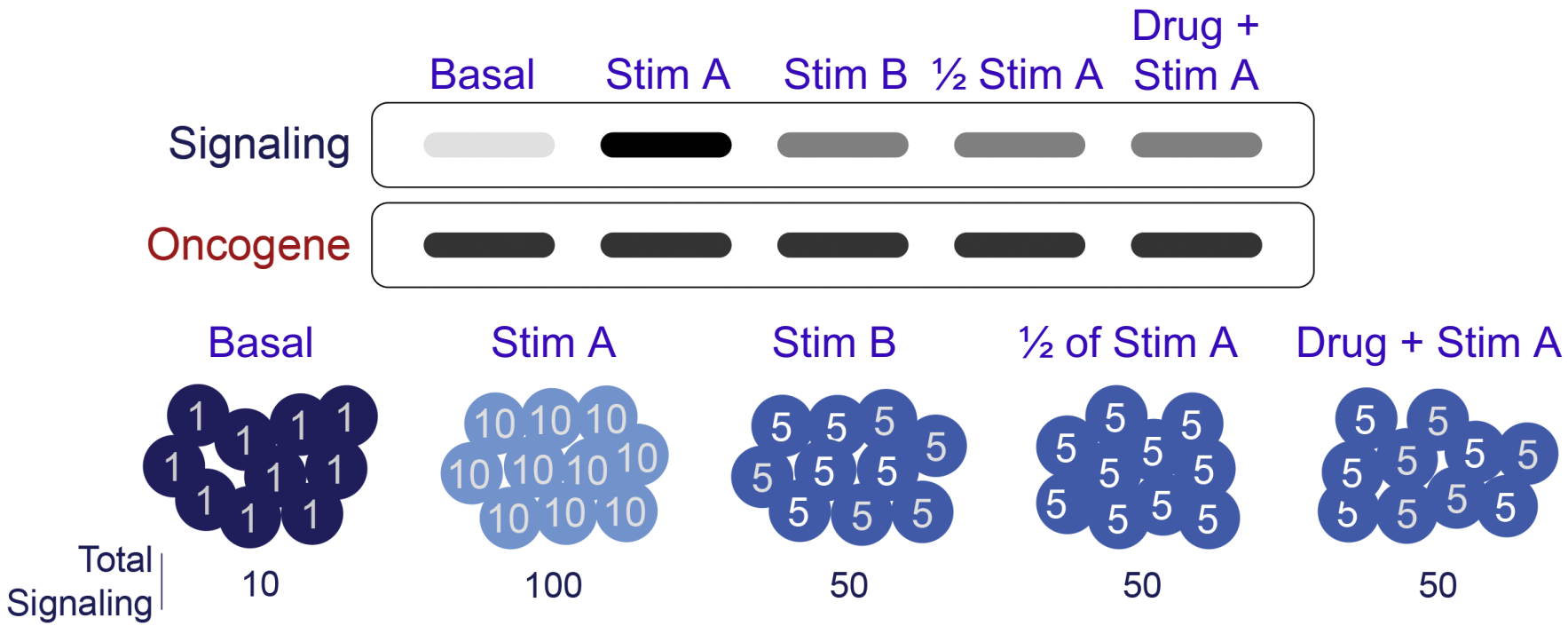
Joachim L Schultze





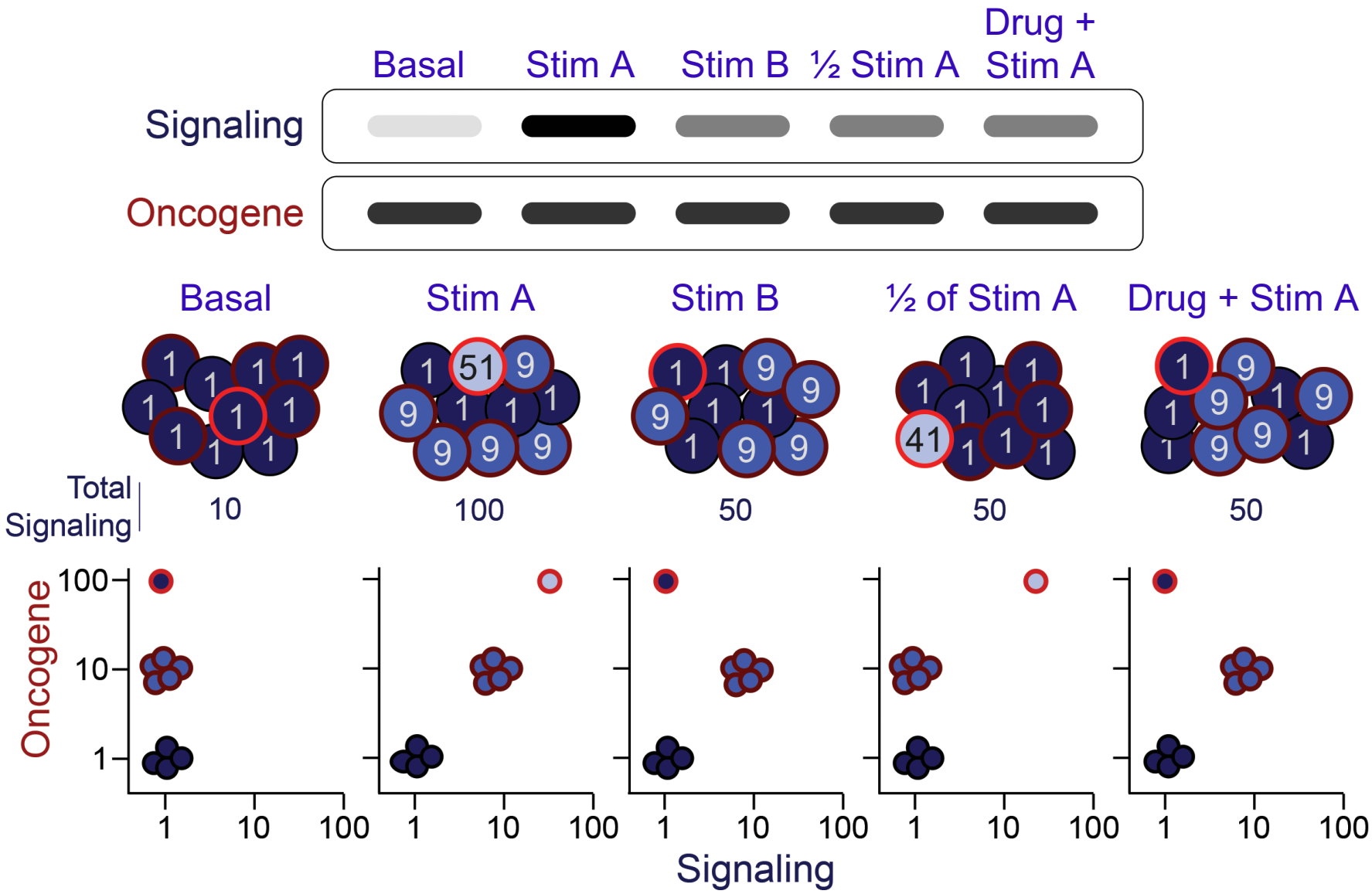
Schultze JL, *Nature Immunology* 2015

Single Cell Biology: Which Cell, How Much?



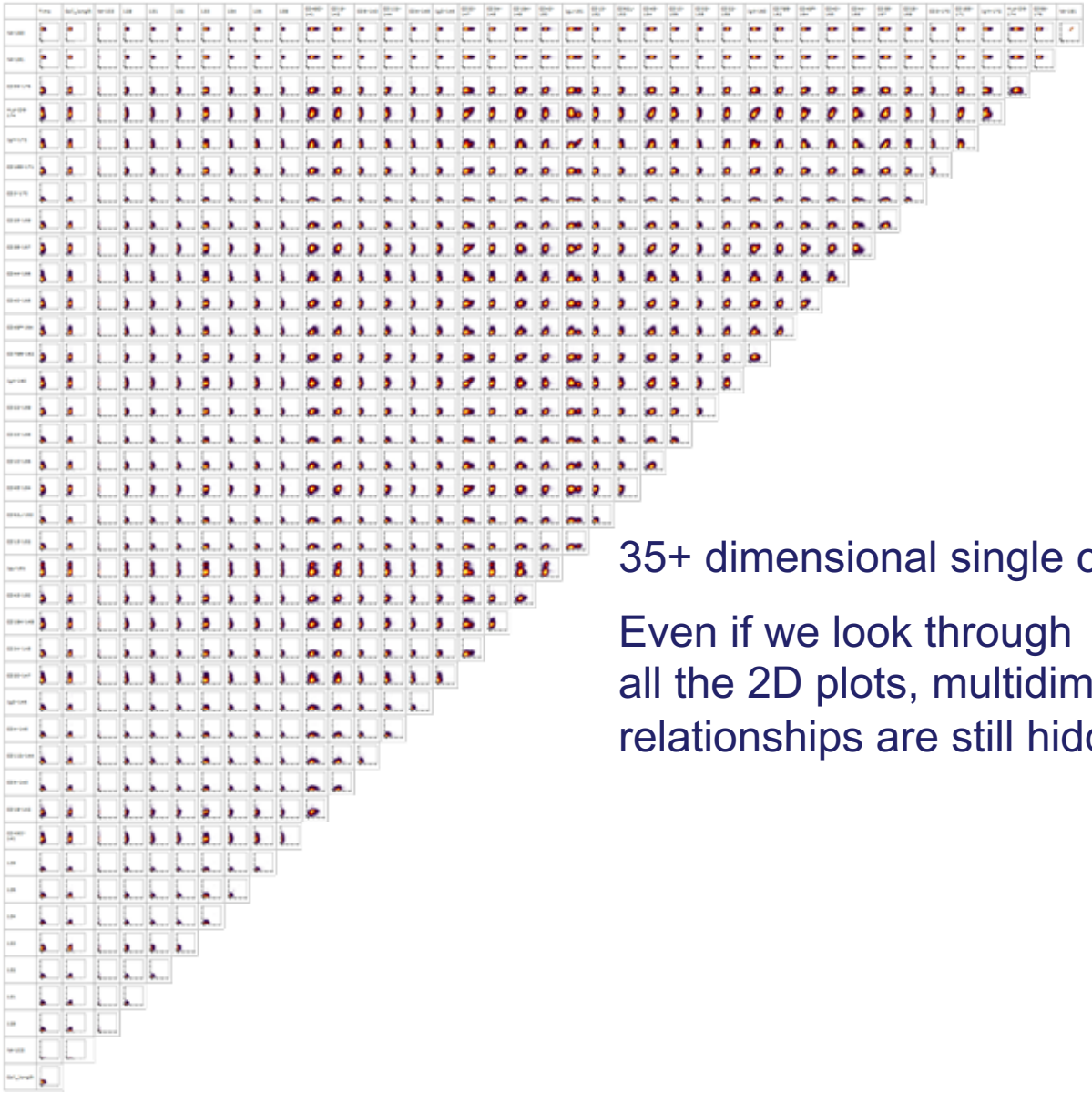
Doxie & Irish, *Current Topics in Microbiology & Immunology* 2014
& Krutzik et al., *Clinical Immunology* 2004

Single Cell Biology: Which Cell, How Much?



Doxie & Irish, *Current Topics in Microbiology & Immunology* 2014
& Krutzik et al., *Clinical Immunology* 2004

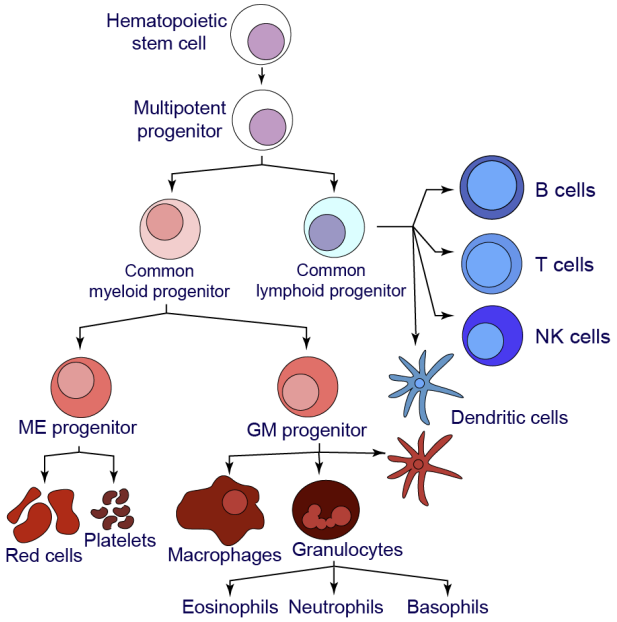
We Now Make Billions of Multi-D Single Cell Measurements
=> Need for Machine Learning Tools & Human Readable Views

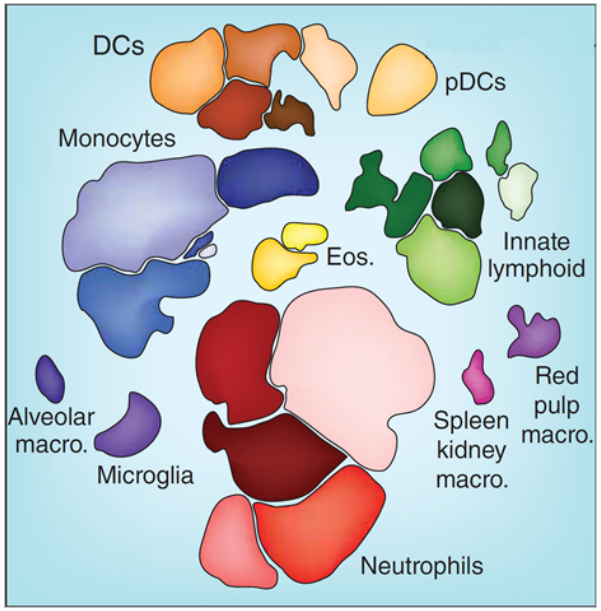


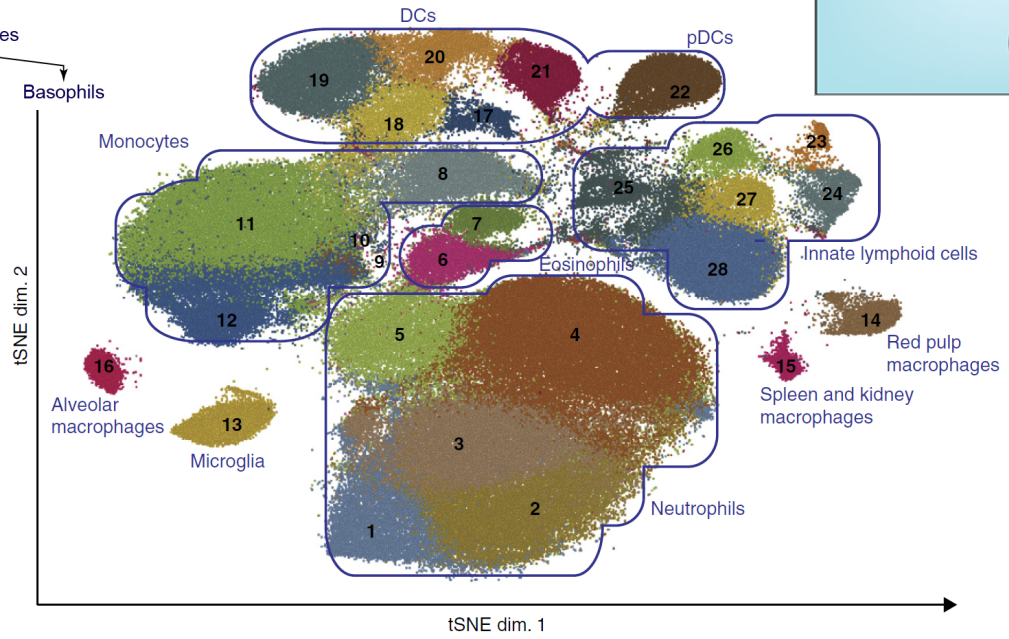
35+ dimensional single cell data:

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

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

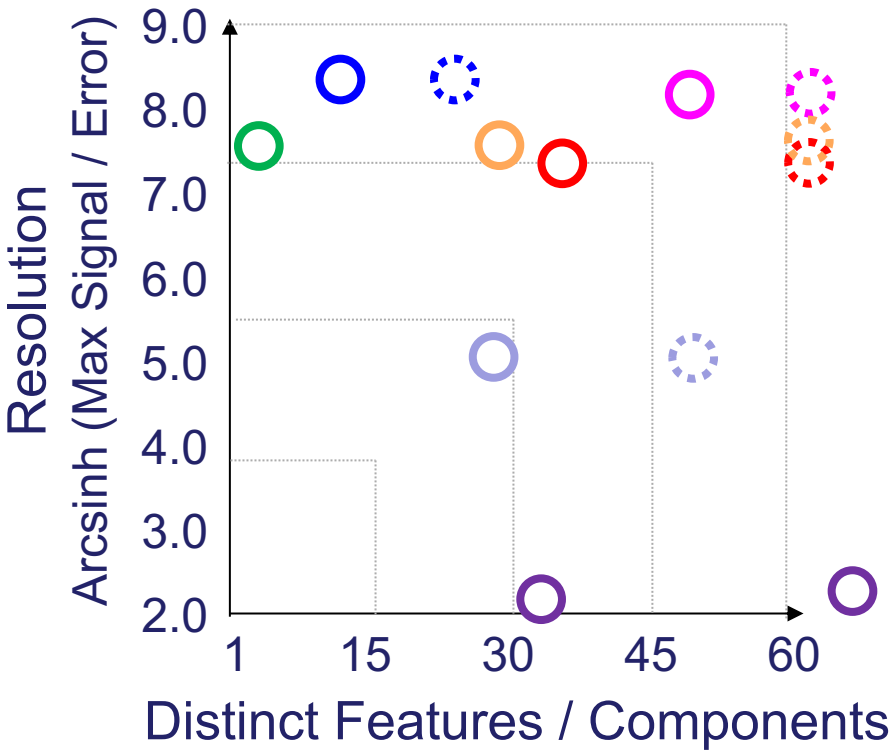






Irish, *Nature Immunology* 2014
Based on Becher et al., *Nature Immunology* 2014

Cytometry Probes & Instruments Keep Improving (ca. 2019)



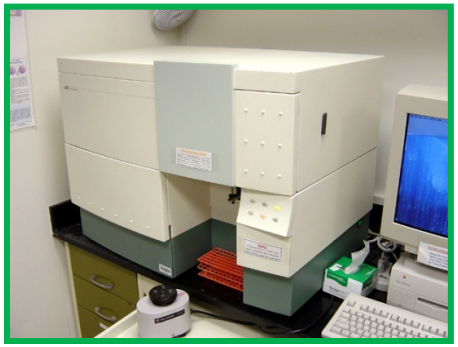
Symphony
27 probes (50)
262,144 / 3000
Res: 5.2

Aurora
25 probes (?)
4,194,300 / 6,000
Res: 7.2

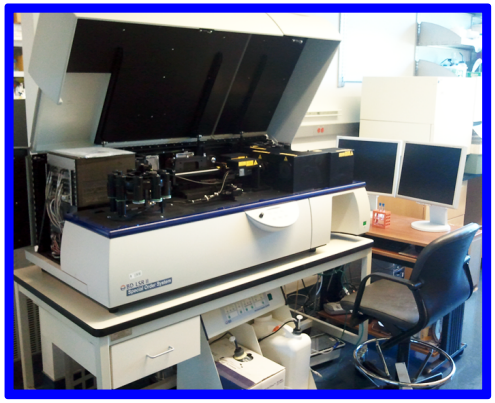
scRNA-seq
30K genes
(35 PCAs?)
15 / 3
Res: 2.3

50 / 1
Res: 4.6

CyTOF 3.0
45 probes (?)
12,000 / 5
Res: 8.3



FACSCalibur
4 probes
10,000 / 10
Res: 7.6



LSR ii
14 probes (25)
262,144 /
150
Res: 8.2

CyTOF 1.0
36 probes (50)
12,000 / 15
Res: 7.2

Mistry et al., *FEBS J* 2018; "Resolution" calculated from typical published data; dashed lines indicate theoretical or unpublished values.

Mass Cytometry Balances Signal, Throughput, & Cost

Mass cytometry: Imaging or flow cytometry method of multiplexed single cell analysis. Standard mass cytometry panels detect 37+ features per cell using pre-validated antibodies. The dynamic range is >10,000 intensity units per feature and a small flow-based mass cytometry dataset might include 1.2 million cells from 12 samples collected at a rate of 500 cells/second (~40 min instrument time) for a total cost of ~$4,500 ($0.004 per cell), including personnel time/effort.

Diggins et al., *Bench to Bedside to Bytes, in review*
Mass cytometry vs. other single cell technologies: Mistry et al., *FEBS J* 2018

Some of the literature from "big seq" is hilariously inaccurate when it comes to flow cytometry…

From: "Single cell RNA sequencing to explore immune cell heterogeneity", Nat Rev Immunology 2017

	FACS	CyTOF	qPCR	Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2)	Fluidigm C1	Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2)	Massively parallel approaches (Drop-seq, InDrop)
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500	48–96	500–2,000	5,000–10,000
Cost	$0.05 per cell	$35 per cell	$1 per cell	$3–6 per well	$35 per cell	$3–6 per well	$0.05 per cell
Sensitivity	Up to 17	Up to 40	10–30 genes	7,000–10,000 genes	6,000–9,00 genes for cell	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells

Wait, so this review says each experiment costs… ?!
FACS: $50,000 ($0.05 x 1,000,000 cells)
CyTOF: $35,000,000 ($35.00 x 1,000,000 cells)

mass cytometry); FACS, fluorescence-activated cell sorting; CR, quantitative PCR; SCRB-seq, single-cell RNA barcoding and sequencing; STRT-seq, single-cell tagged reverse transcription sequencing.

Mass Cytometry Dissects Cellular Mechanisms of Cancer Immune Response

Article

Cell

Systemic Immunity Is Required for Effective Cancer Immunotherapy

Spitzer et al., *Cell* 2017

Uses mass cytometry to characterize essential role of peripheral blood $CD4^+$ T cells in immunotherapy response

ARTICLE

doi:10.1038/nature22079

T-cell invigoration to tumour burden ratio associated with anti-PD-1 response

Huang et al., *Nature* 2017

Uses mass cytometry to reveal peripheral blood CD8 T cells associated with anti-PD-1 immunotherapy responses

Article

Cell

Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade

Wei et al., *Cell* 2017

Uses mass cytometry to characterize similar & distinct tumor-infiltrating immune cell subsets (mostly T cells) following immunotherapies

Cancer Immunology Is Powered by Cytometry

Dr. Allison, 70, is chairman of immunology at the University of Texas M.D. Anderson Cancer Center. He did the work recognized by the Nobel committee while working the University of California at Berkeley and Memorial Sloan Kettering Cancer Center in New York.

Dr. Honjo, 76, is a longtime professor at Kyoto University, where he did his landmark work. Previously, he did research at Osaka University, the University of Tokyo and the National Institutes of Health in Washington.

The Nobel Prize for Physiology and Medicine was awarded to James P. Allison, left, and Tasuku Honjo on Monday for their work on cancer research. Credit Jonathan Nackstrand / Agence France-Presse — Getty Images; New York Times 2018

Referenced Comparisons of Single Cell Techniques

	Mass cytometry	Fluorescence cytometry	scRNA-seq
Detection method	Time-of-flight of ionized heavy elemental isotopes	Emitted light and light scatter	Nucleic acid sequence
Example probes	Metal-conjugated antibodies, metal-conjugated small molecules	Fluorochrome-conjugated antibodies, fluorescent molecules	Poly-A targeted oligonucleotides, antibody-conjugated oligonucleotides [144]
Example cellular features and targets measured	Proteins, phospho-proteins, chromatin modifications [124], RNA transcripts [131–133], platinum drug uptake [128,129], cell cycle status and DNA synthesis [130], cell size [125], apoptosis [48], and viability [27]	Proteins, phospho-proteins, chromatin modifications [145], RNA transcripts [146], fluorescent drug uptake [147], metabolism and redox state [148,149], cell cycle status and DNA synthesis [150], cell size and granularity, apoptosis [151] and viability [152]	Poly-adenylated RNA transcripts, CITE-seq probes [144]
Minimum cells per sample needed at start of protocol	50 000 cells	50 000 cells	200 cells [153][a]
Cell capture rate	30–60% [31,32]	> 95%	5–65% [154]
Target capture	> 95%	> 95%	10–40% [154–156]
Example of analyzed cell events per study	2 000 000 cells per study	20 000 000 cells per study	5000 cells per study [157]
Resolution, arcsinh (max/error)[b]	~ 6–9 e.g., arcsinh (12 000/5)	~ 5–9 e.g., arcsinh (262 144/150), arcsinh (4 194 300/6000)[c]	~ 2–3 e.g., arcsinh (10/1)
Cell throughput	Up to 10^4 cells·s^{-1} [33]	Up to 10^5 cells·s^{-1} [31]	Up to 10^4 cells·s^{-1} [155]
Features/available channels	50/200 channels	30/64 channels[c]	~ 30 000/N/A
Degree of crosstalk between parameters	3% (range: 0–8.6%) [55]	19.5% (4.9–51.1%) [56]	N/A
Accessibility	Major research institutions	> 20 features: major research institutions; ~ 4 color common	Major research institutions
Total cost	$0.004-$0.01/cell [31]	< $0.001/cell	$0.05-$3.00/cell [158]

Beyond the Message: Advantages of Snapshot Proteomics with Single-Cell Mass Cytometry in Solid Tumors
Mistry et al., *FEBS Journal* 2018

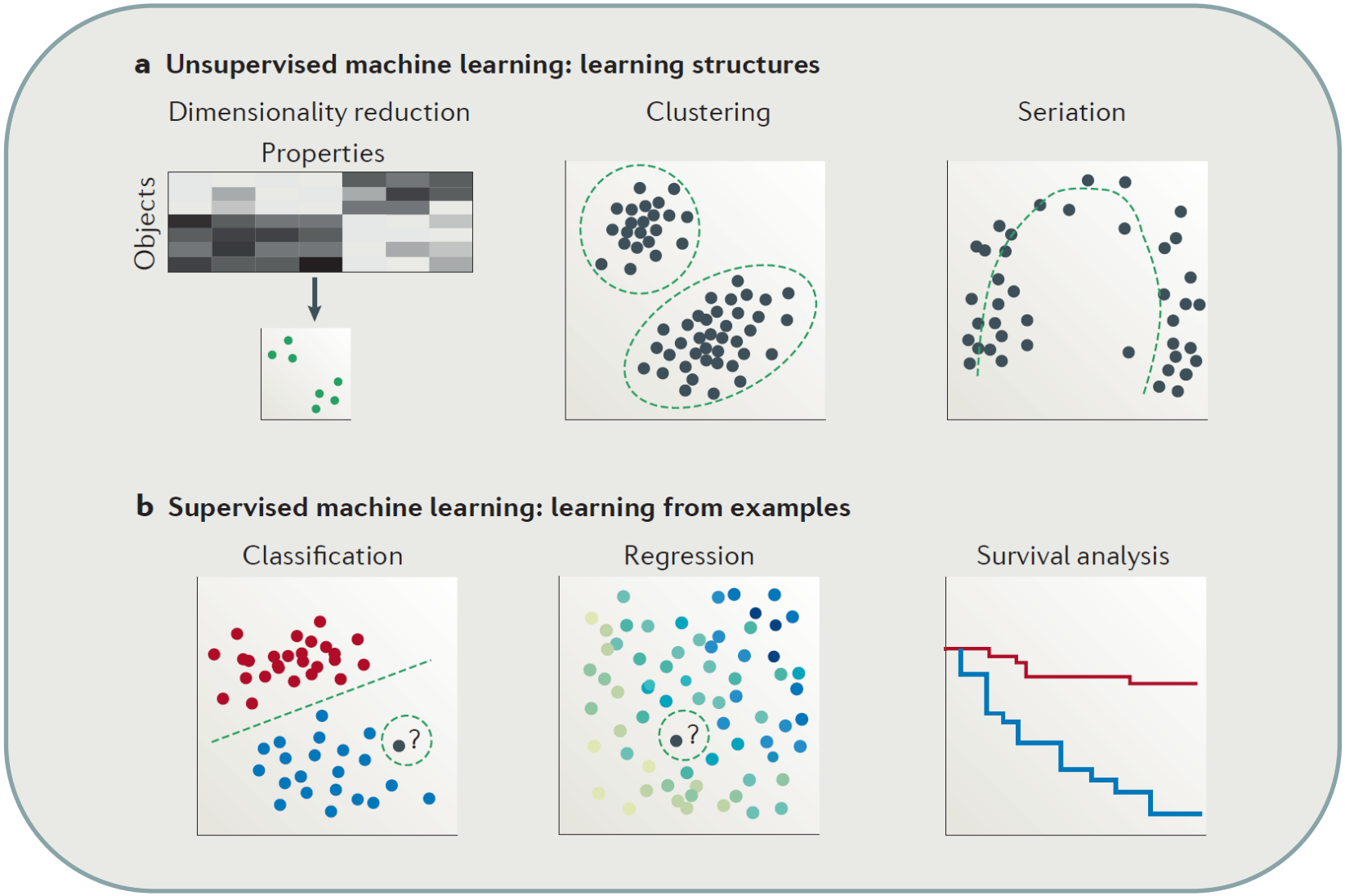
Effective data analysis is critical in clinical research, & this now means working *with* computational tools that reveal and model patterns across data types

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/

Computational Cytometry: Data Science Tool Types



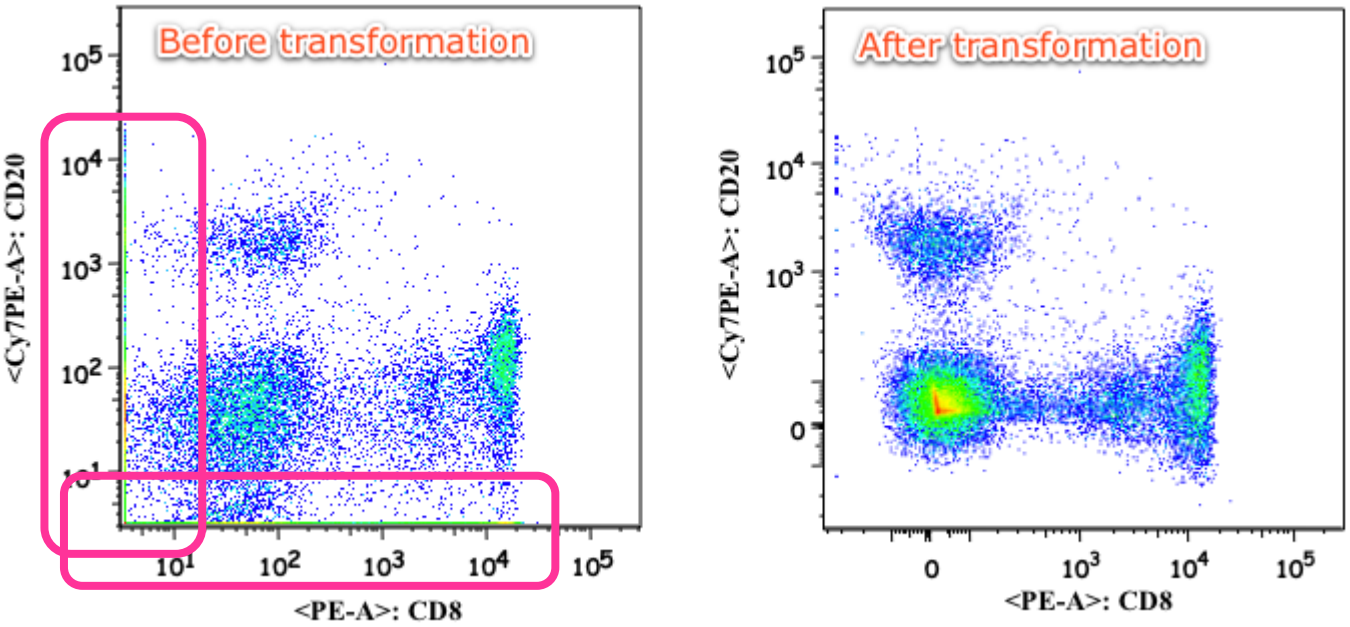
Saeys Y, Van Gassen S, and Lambrecht BN, *Nature Reviews Immunology* 2016

Before we get to ‘desert’, some math ‘veggies’:

Scales matter: poorly or variably scaled data can destroy an analysis, most issues arise near zero

(pre-processing & normalization can also be critical)

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

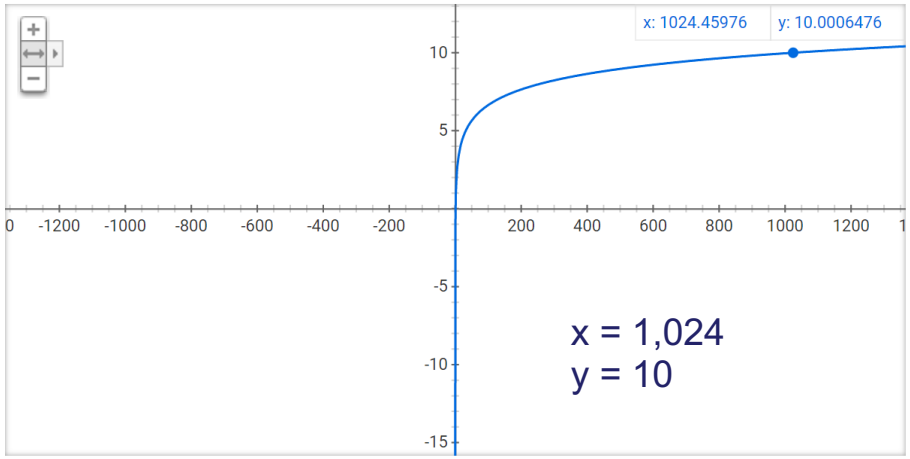
Results from bad scaling (poor transformation)
and it can be an issue for computational analysis.

Scaling is important in both mass and fluorescence cytometry.

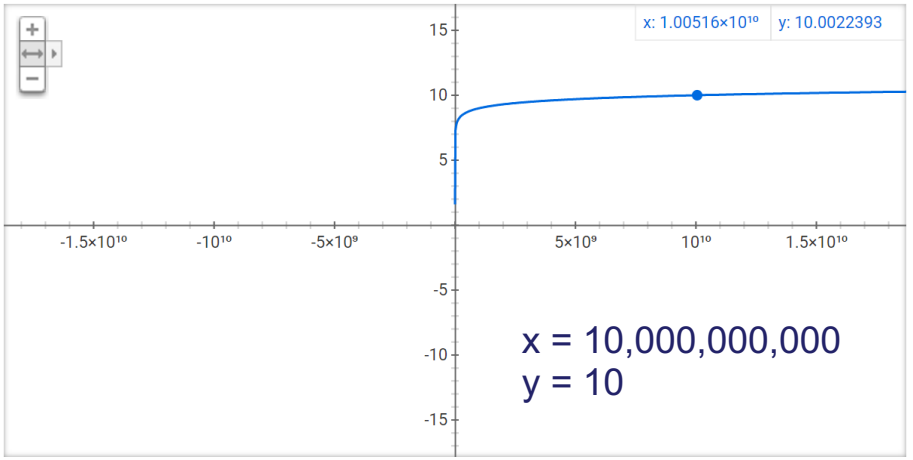
Examples of Four Common Data Scales

$\log_2(x)$ $\log_{10}(x)$ arcsinh(x/5) arcsinh(x/150)

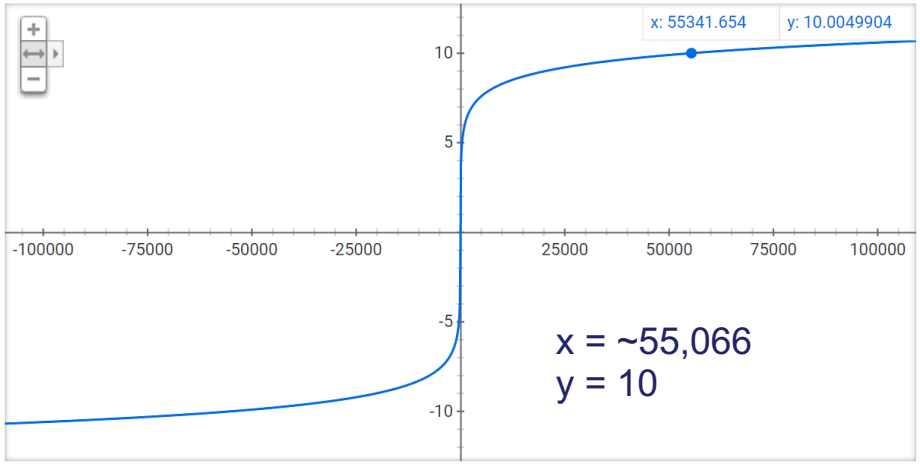
Graph for ln(x)/ln(2)



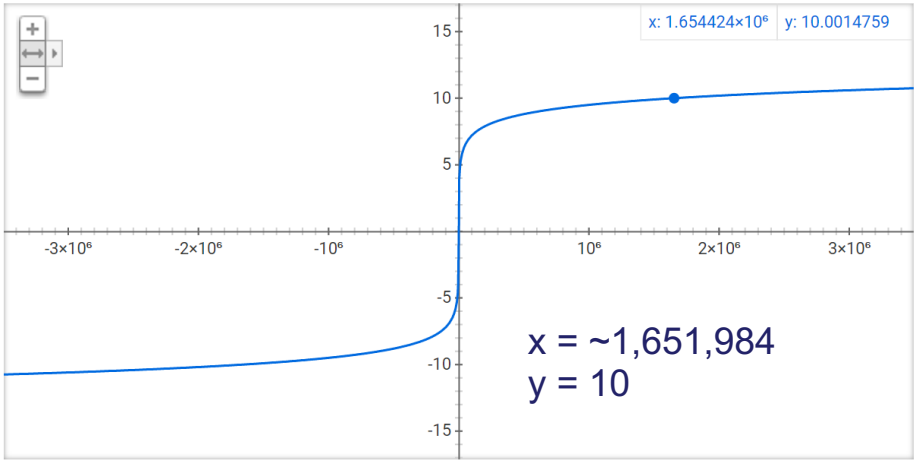
Graph for ln(x)/ln(10)



Graph for ln(x/5+sqrt(1+(x/5)^2)) = arcsinh(x/5)



Graph for ln(x/150+sqrt(1+(x/150)^2)) = arcsinh(x/150)

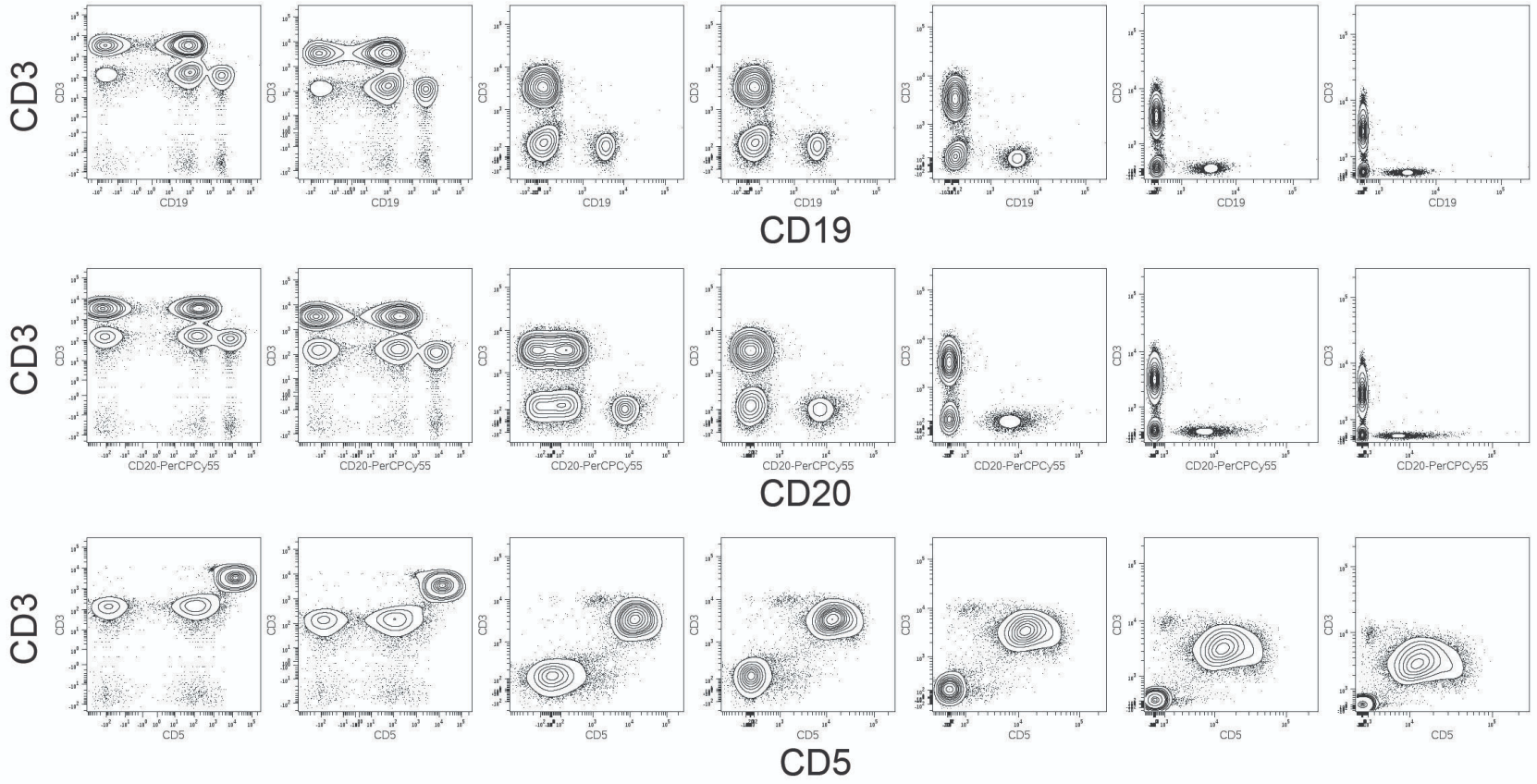


Scaling Matters for Measuring Distance (Fluorescence Flow)

Healthy human PBMC, intact cells gate

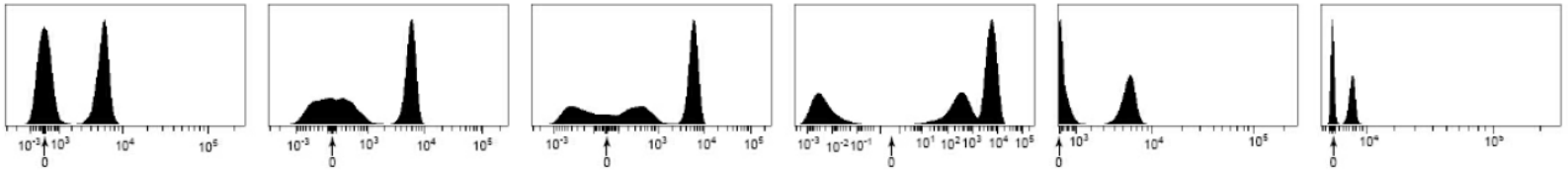
arcsinh(intensity / argument)

CD3 arg	1	5	150	150	300	750	1500
CD19 arg	1	5	150	150	300	750	1500
CD20 arg	1	5	150	500	1500	3000	6000
CD5 arg	1	5	150	500	1500	3000	6000



Scaling Matters for Measuring Distance (Compensation Beads)

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

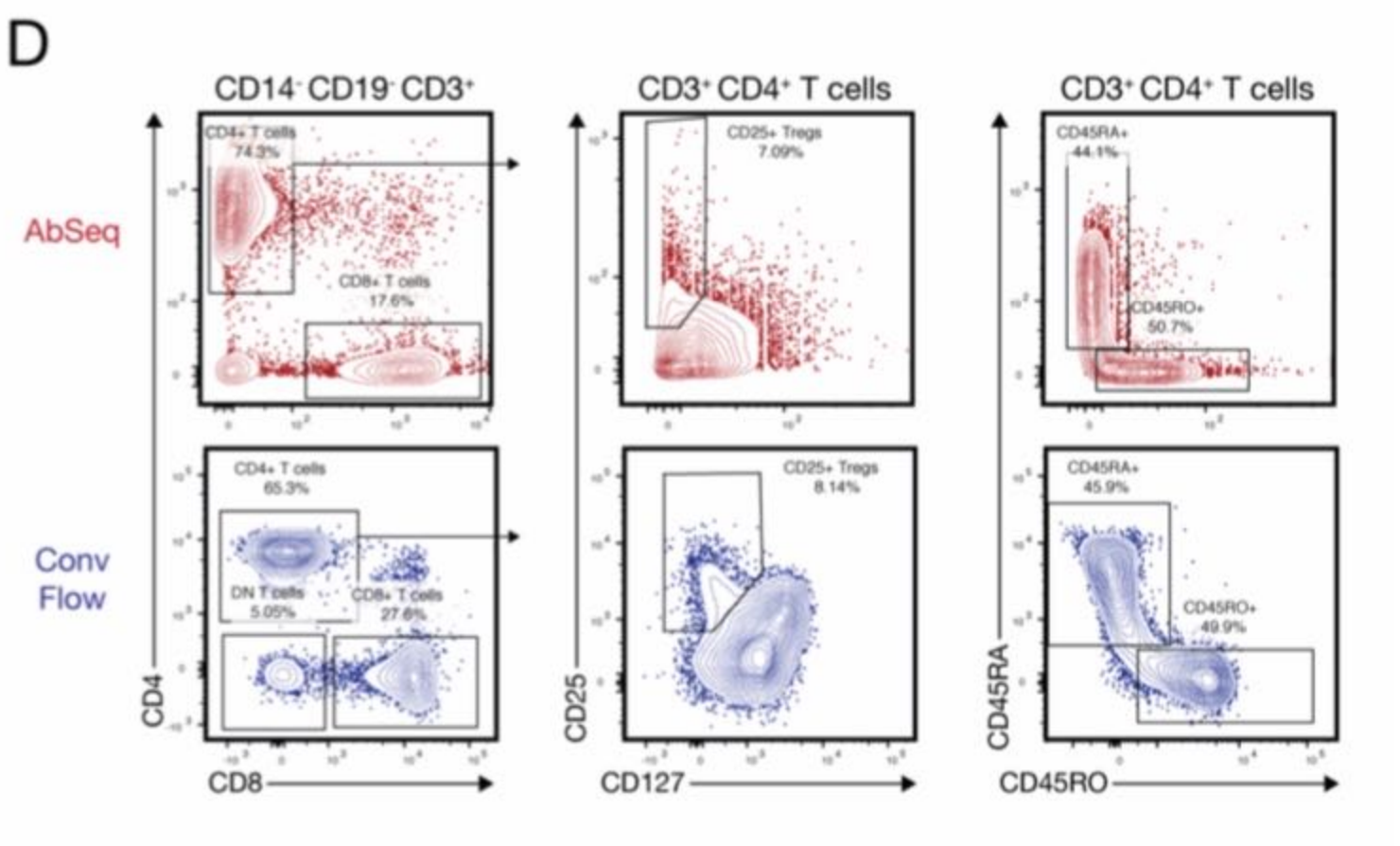
arcsinh(x) with cofactor c = $\ln\left(\frac{x}{c} + \sqrt{1 + \left(\frac{x}{c}\right)^2}\right)$

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

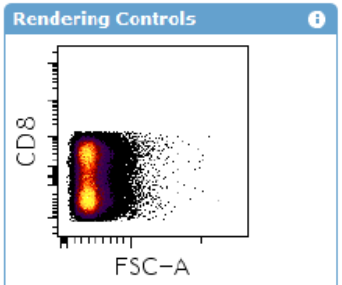
Since computational analysis techniques compare distance similar to what a person does when looking at a plot, these techniques can identify artificial populations near zero (see C and D) if data are not appropriately transformed prior to analysis.

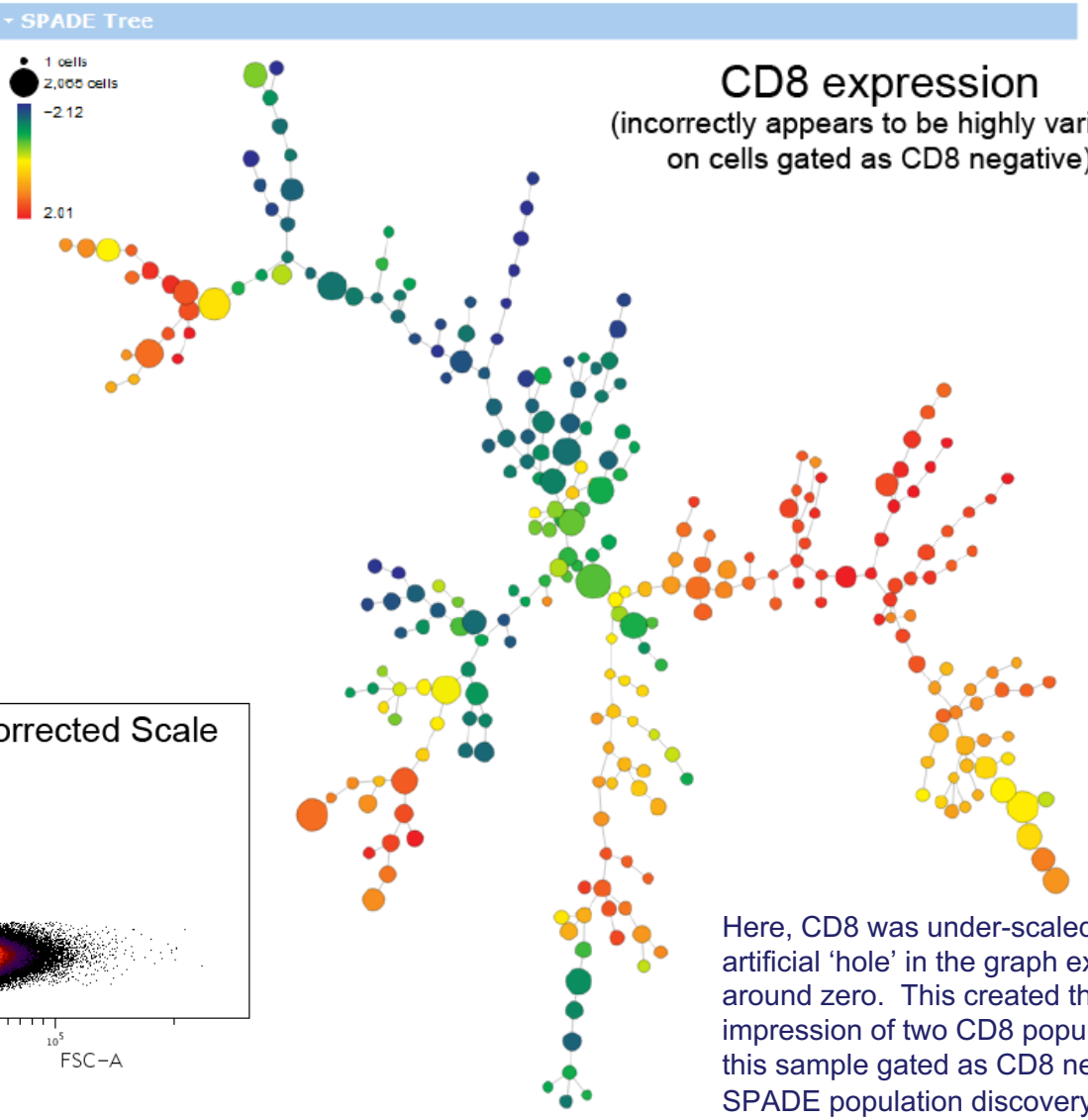
More information: https://my.vanderbilt.edu/irishlab/protocols/scales-and-transformation/
http://www.flowjo.com/v76/en/displaytransformwhy.html

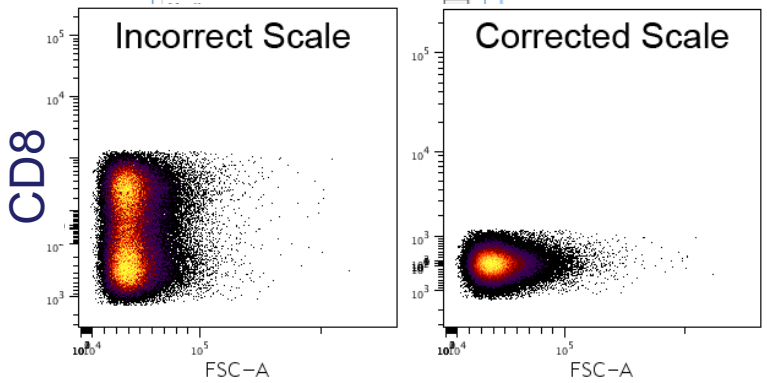
Scaling Matters for Measuring Distance (RNA seq vs. Fluorescence Flow)



Inappropriate Scaling Can Lead to False Population Discovery



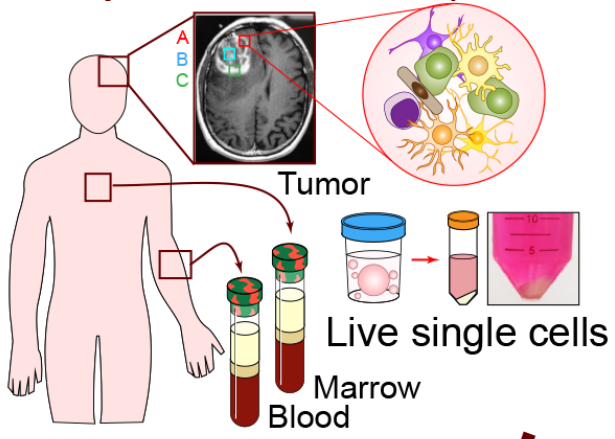




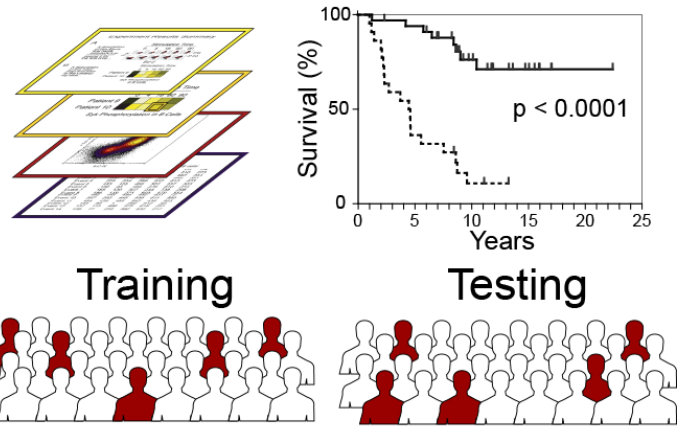
Here, CD8 was under-scaled so that an artificial 'hole' in the graph existed around zero. This created the false impression of two CD8 populations in this sample gated as CD8 negative. SPADE population discovery treated this as significant.

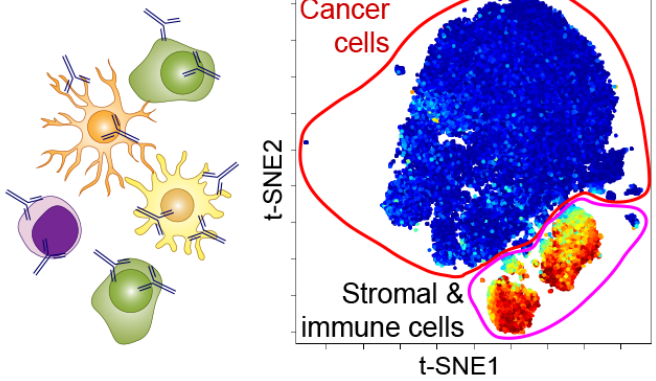
Elements of Translational Data Science

Sample patients over time
at key clinical decision points

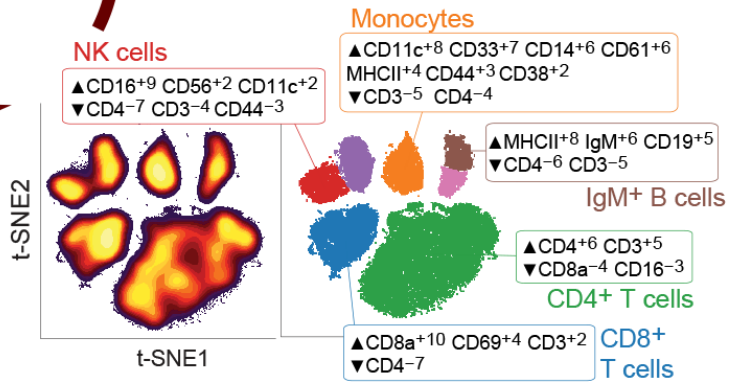


Integrate systems biology data
& clinical outcomes to guide treatment





Measure cell identity, signaling,
biomarkers, and functional responses



Apply machine learning tools
to reveal patterns & identify groups

Diggins et al., *in review*

Key Topic Areas, Terms, and Cytometry Workflows

1: Field Changes: Data Science & Latest Tools

2: History: Non-linear, PCA, Trajectories, Supervised

3: Dimensionality Reduction: t-SNE & UMAP

4: Clustering: SPADE (k-means), KNN, FlowSOM, Citrus

5: Enriched Features: MEM, ΔMEM, RMSD

6: Cytometry:
- 2004: Expert => Expert => Heatmap (Irish/Nolan)
- 2011: Expert => SPADE => Heatmap (Bendall/Qiu)
- 2013: t-SNE => Expert (viSNE/Pe'er, Van Der Maaten)
- 2014: t-SNE => DensVM => Heatmap (Newell)
- 2015: t-SNE => SPADE => Heatmap (Diggins/Irish)
- 2015: "KNN" (=> t-SNE) => Heatmap (Phenograph)
- 2015: FlowSOM (Van Gassen/Saeys)
- 2017: t-SNE => SPADE => MEM (Diggins/Irish)
- 2018: UMAP => Expert (Newell, McInnes)
- 2019: UMAP => FlowSOM => MEM (Barone/Irish)

Unsupervised Analysis: Not Using Prior Knowledge To Guide the Analysis

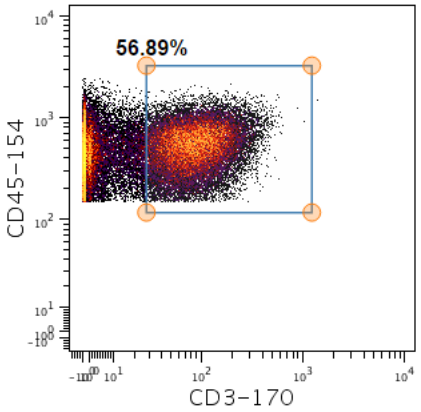
Prior knowledge examples: Stem cells express CD34,
these samples were from patients that responded to drug

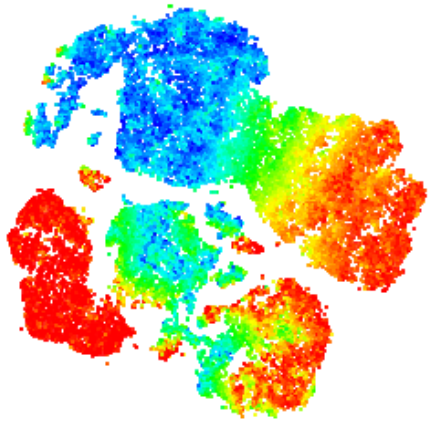
Supervised Approaches

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

Unsupervised Approaches

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



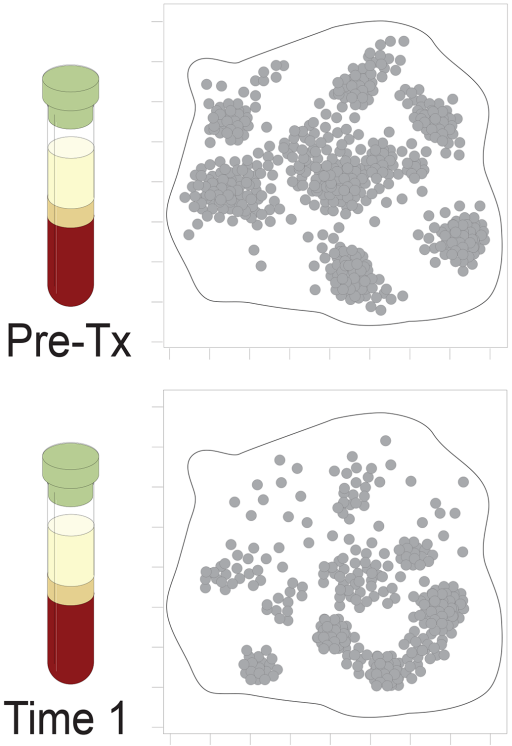


See Table 1 of Diggins et al., *Methods* 2015 for list of unsupervised tools

Clinical Trial Monitoring: What Do We Need to Know? Automate Four Key Readouts vs. Clinical Outcomes

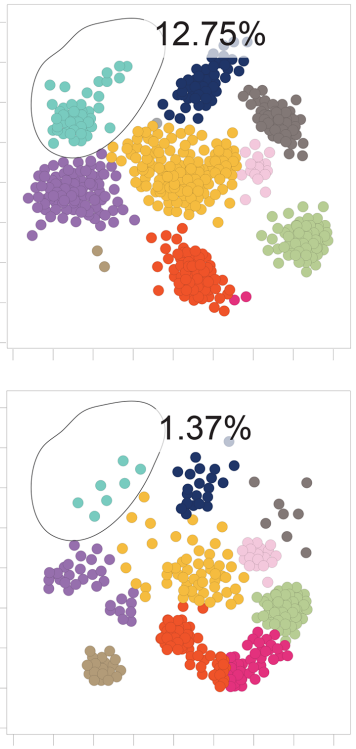
Features of Dynamic Populations

1 Systems Plasticity



Earth Mover's Distance on t-SNE or UMAP

2 Population abundance



Traditional gating or cluster frequency

3 Signature features

Pre-therapy

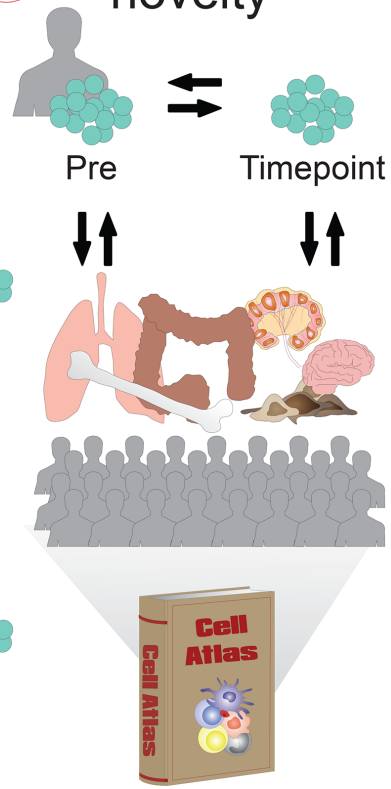
▲HLADR^{+2} CCR5^{+1} CD38^{+1} CD33^{+1}
▼CD8^{-8} CD45RO^{-6} CD3^{-6} CD4^{-4} CD45^{-2} CCR4^{-1} CCR7^{-1} CD28^{-1} CD27^{-1}

Time point 1

▲HLADR^{+2} CD38^{+1} CD45RA^{+1}
▼CD8^{-8} CD4^{-6} CD3^{-6} CD45RO^{-5} CCR5^{-2} CD45^{-2} CD28^{-2} CD20^{-1} CCR4^{-1} CD27^{-1}

Marker Enrichment Modeling (MEM)

4 Population novelty

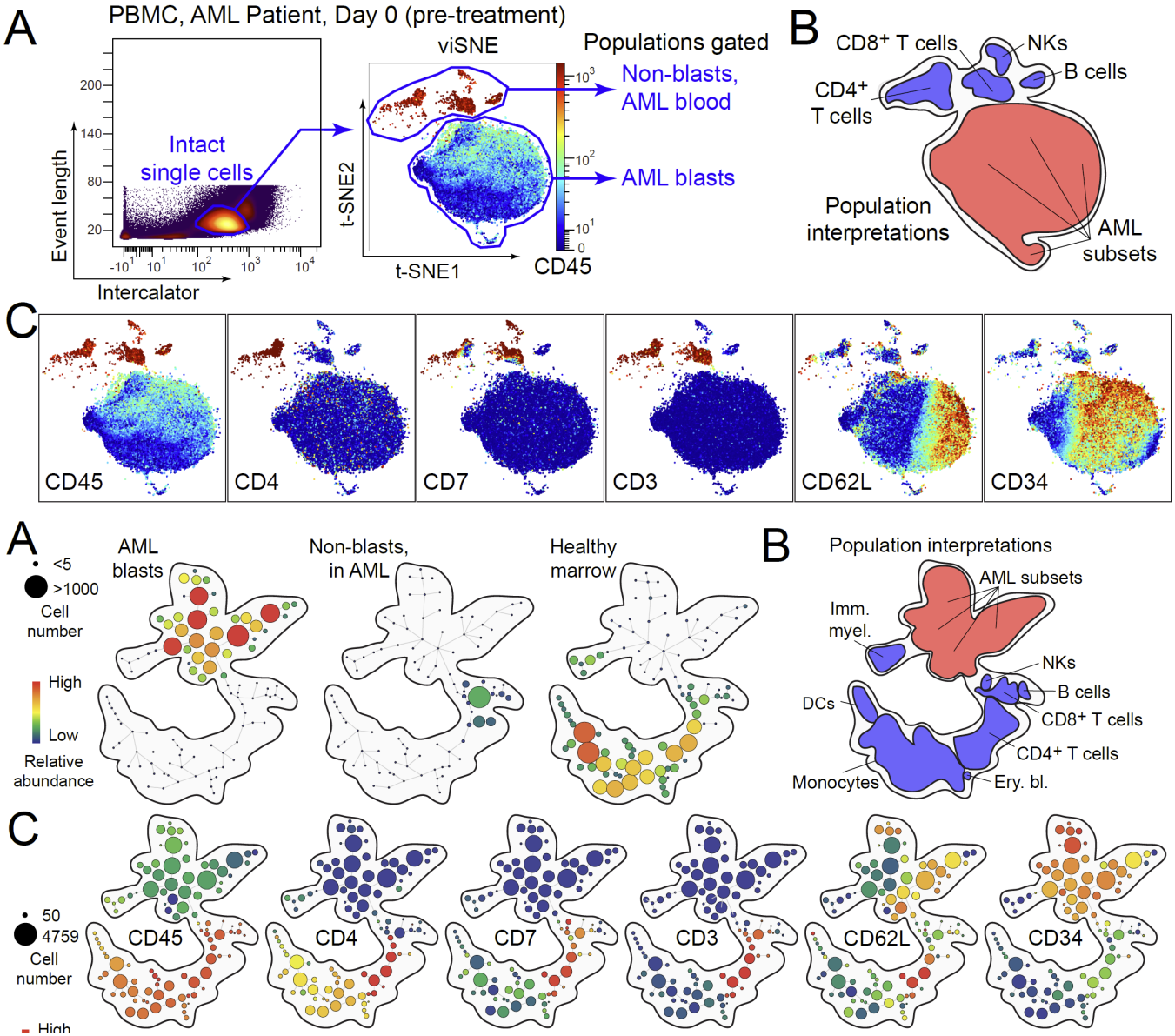


ΔMEM vs. Timepoint or Cell Atlas

How we quantified

Greenplate et al., *Cancer Immunology Research 2019*

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



viSNE

Amir et al.

Nature biotech 2013

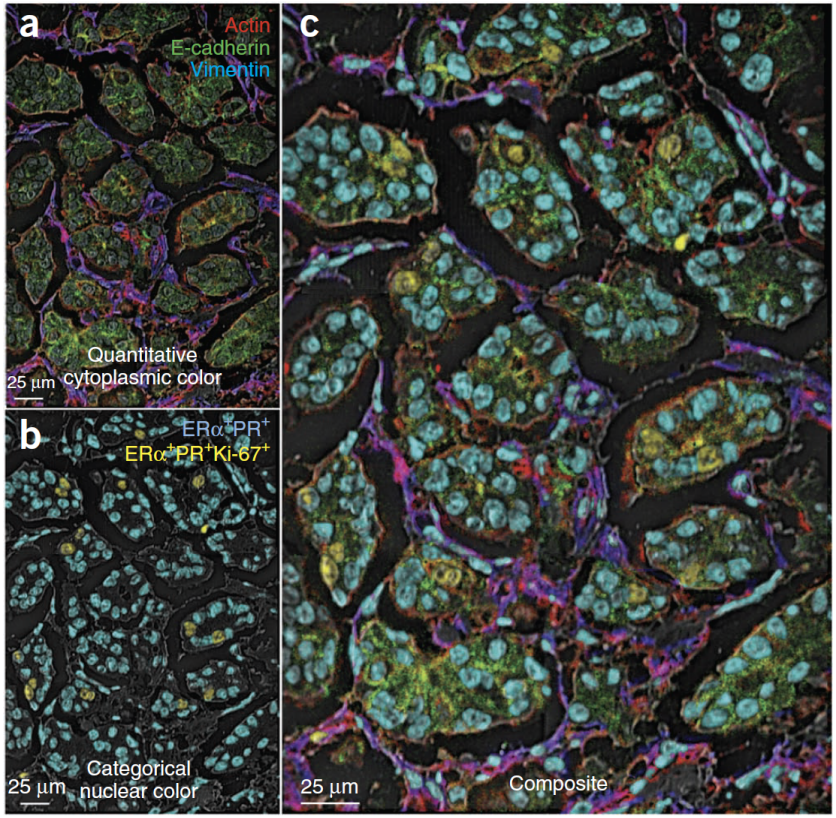
SPADE

Qiu et al.

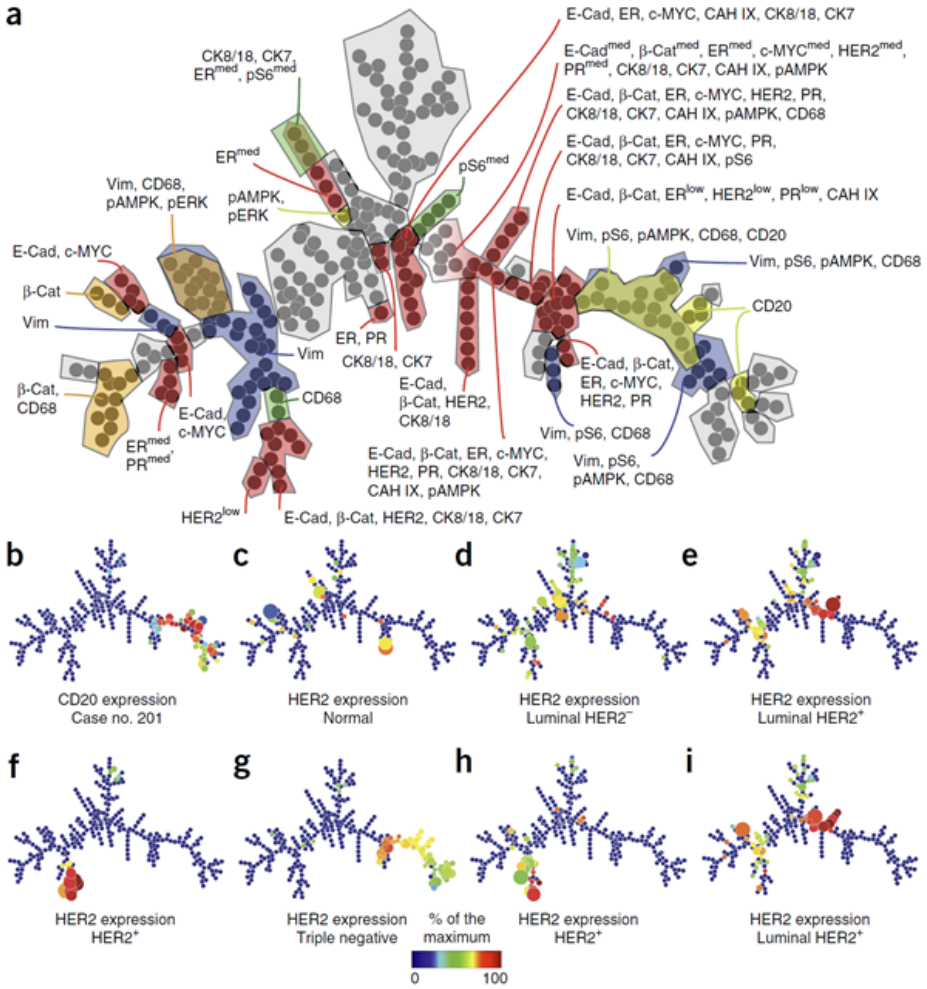
Nature biotech 2011

Diggins et al., *Methods* 2015

Cytomics (Cell Identity): Powered by Multiple Platforms: Imaging Mass Cytometry of Breast Cancer



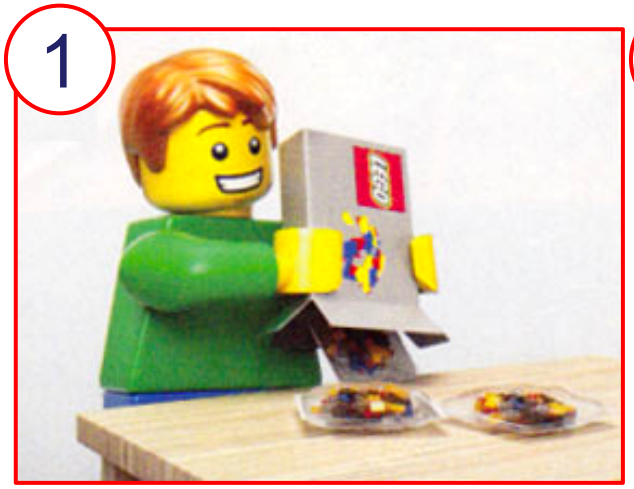
Example MIBI breast cancer histology
Angelo et al., *Nature Medicine* 2014



Analysis of IMC from 20+ breast cancer using SPADE
Giesen et al., *Nature Methods* 2014

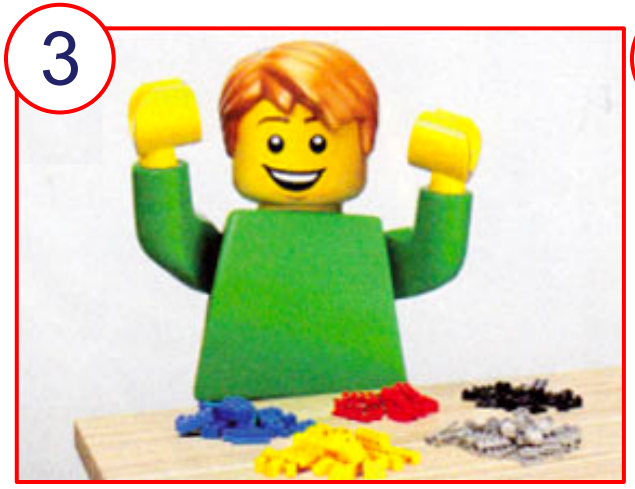
t-SNE was a game changer for single cell

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

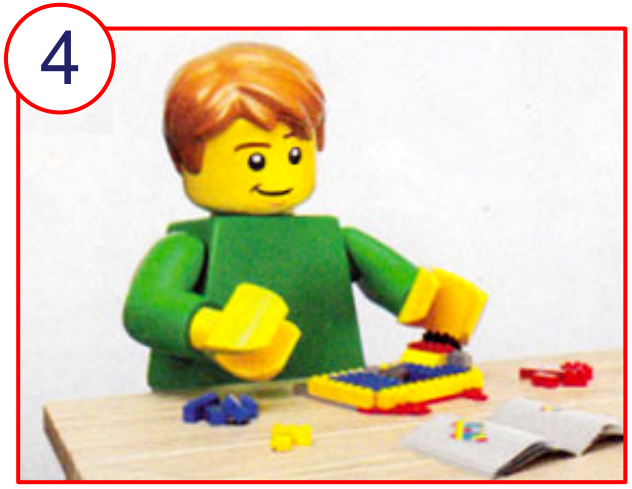


HD cytometry!!

Woah, that's a lot of data…



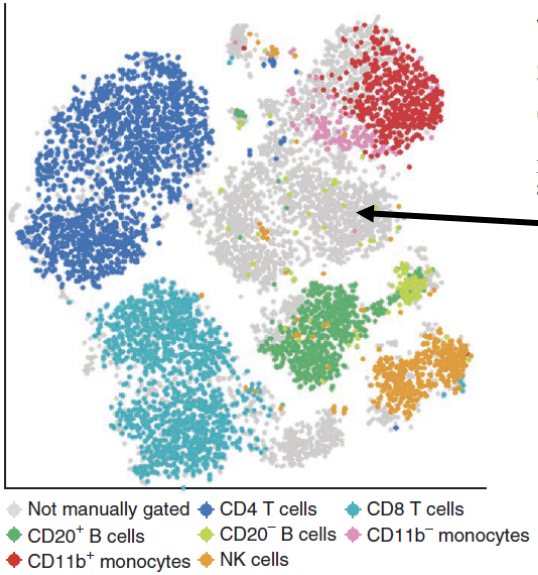
Computational tools



Biological knowledge

Adapted from LEGO 70803 Cloud Cuckoo Palace instructions 2014

Traditional Gating Overlooks Many Cells in Primary Samples



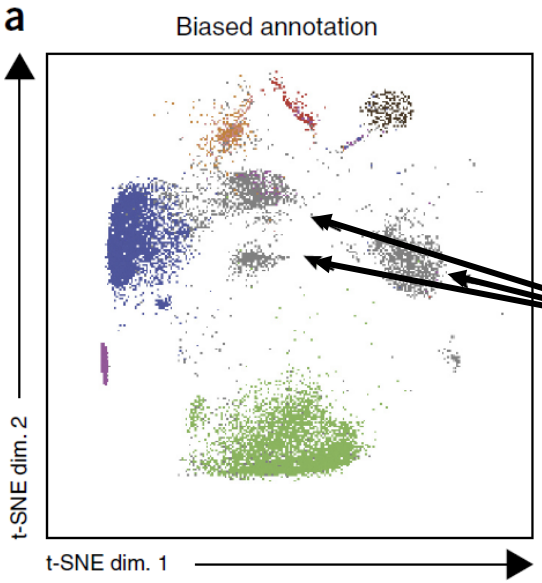
nature biotechnology

2013

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

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

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 based 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 immunology

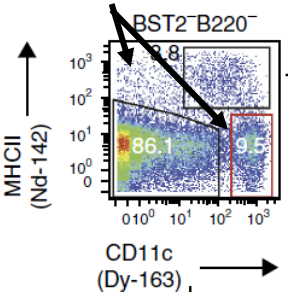
2014

High-dimensional analysis of the murine myeloid cell system

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

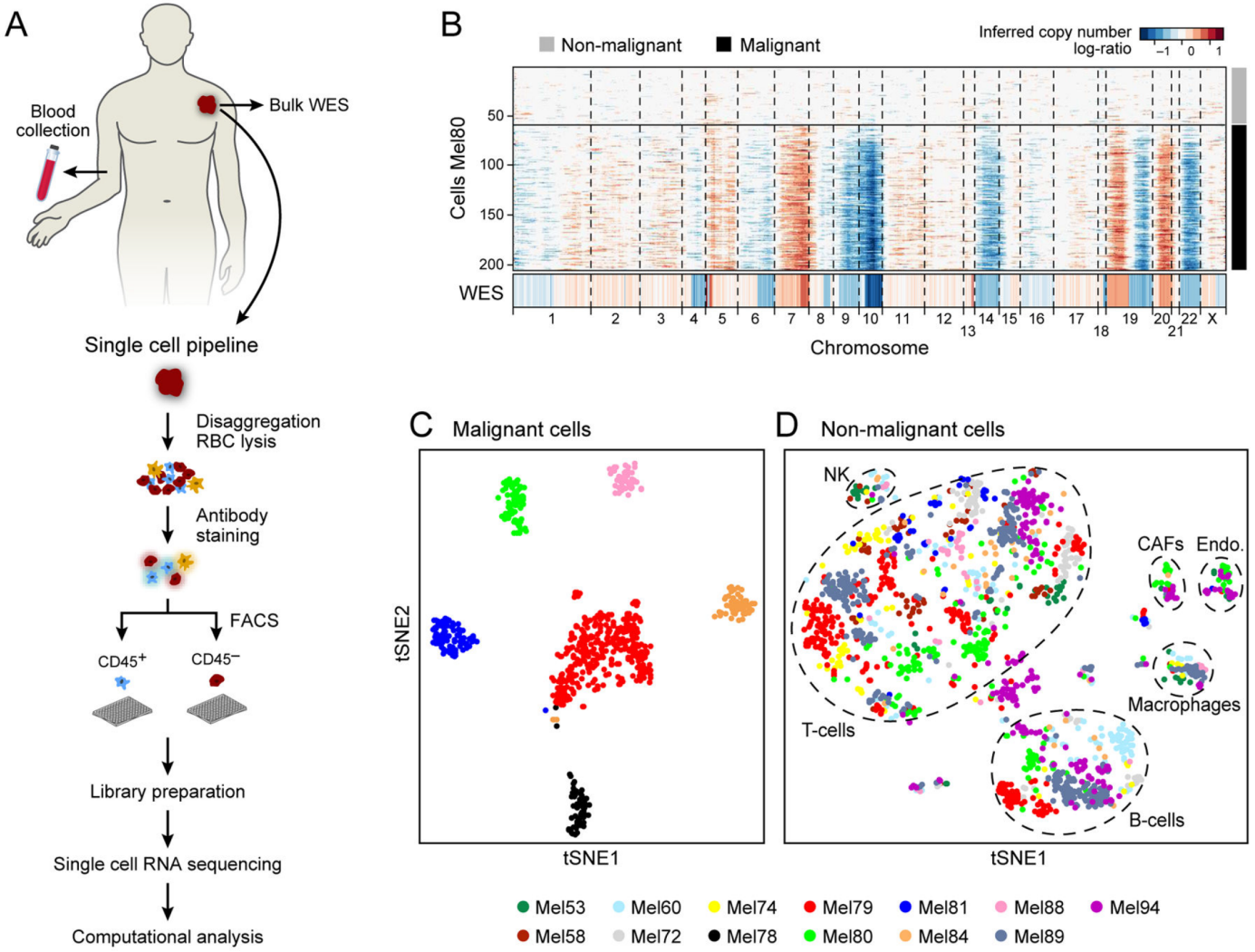
Notably, whereas traditional biased gating strategies allowed for identification of only 54.7 ± 2.6% (mean ± s.e.m., n = 3 mice) of lung myeloid cells (different DC subsets, macrophages, monocytes, neutrophils), the automatic, computational approach identified nearly 100% of the cells (96.6 ± 1.0% (mean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

Overlooked cells 'between' traditional gates add up to ~half the sample



Amir et al., *Nature Biotechnology* 2013; Becher et al., *Nature Immunology* 2014

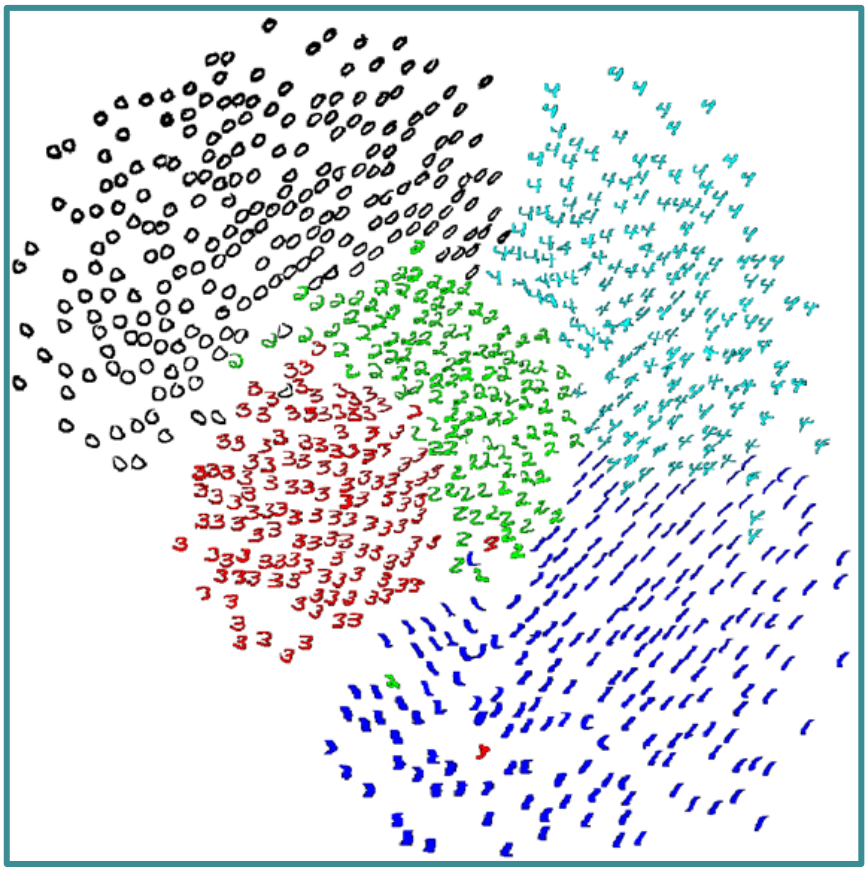
Cytomics (Cell Identity): Powered by Multiple Platforms
Melanoma Cell Diversity Based on scRNA-seq Data



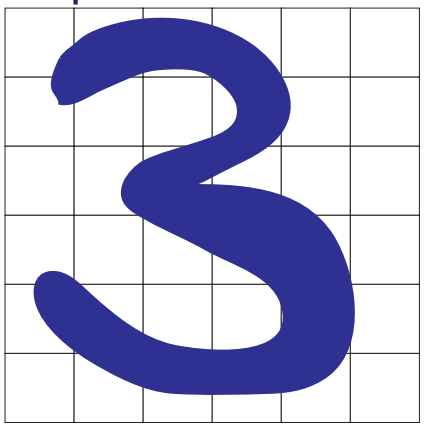
Tirosh et al., *Science* 2016

Stochastic Neighbor Embedding (SNE)

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



Example: 6x6 Pixel Image

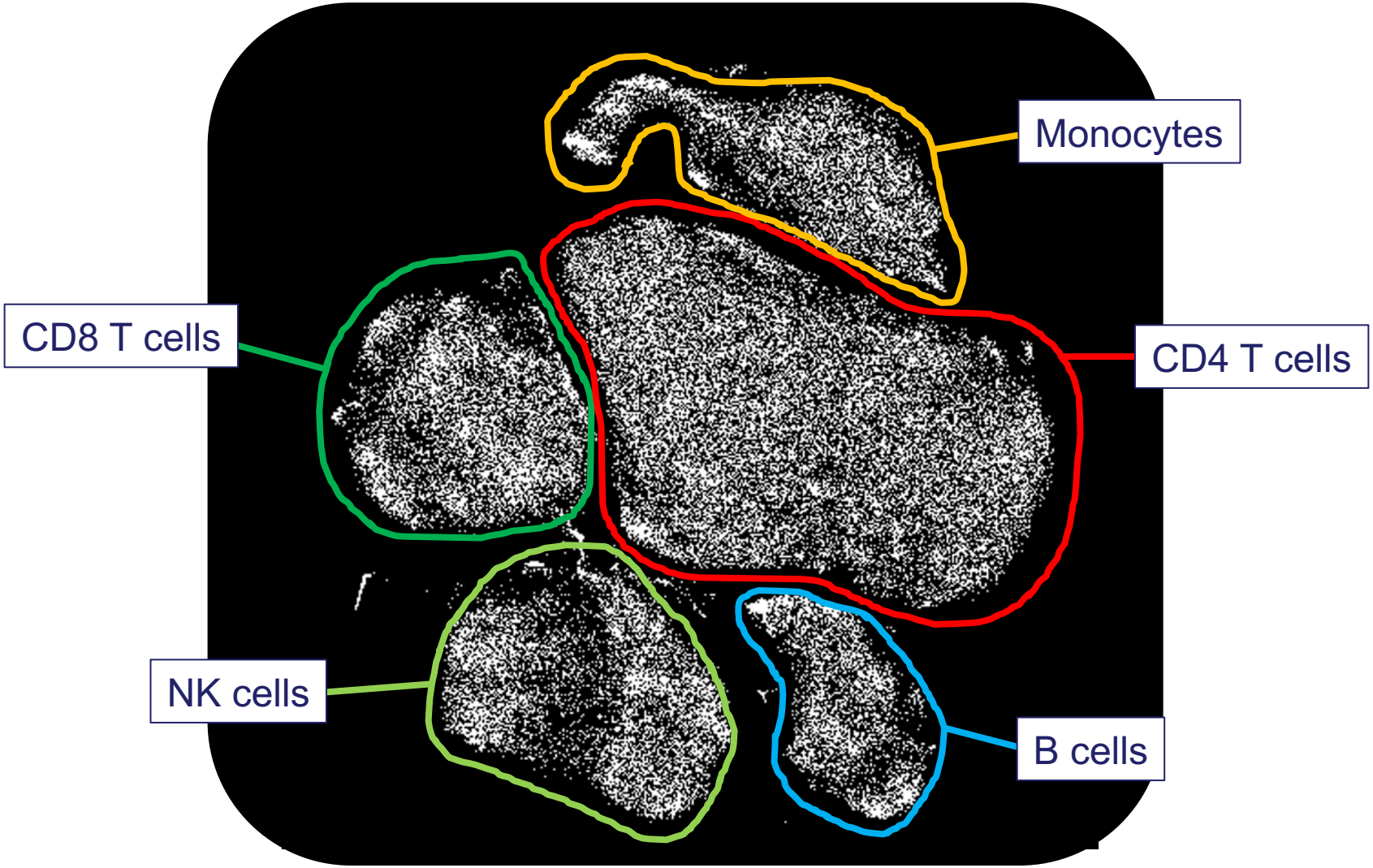


Vectorize (1x36)

tSNE on all pixels

Hinton et al., "*Advances in neural information processing systems*." 2002.

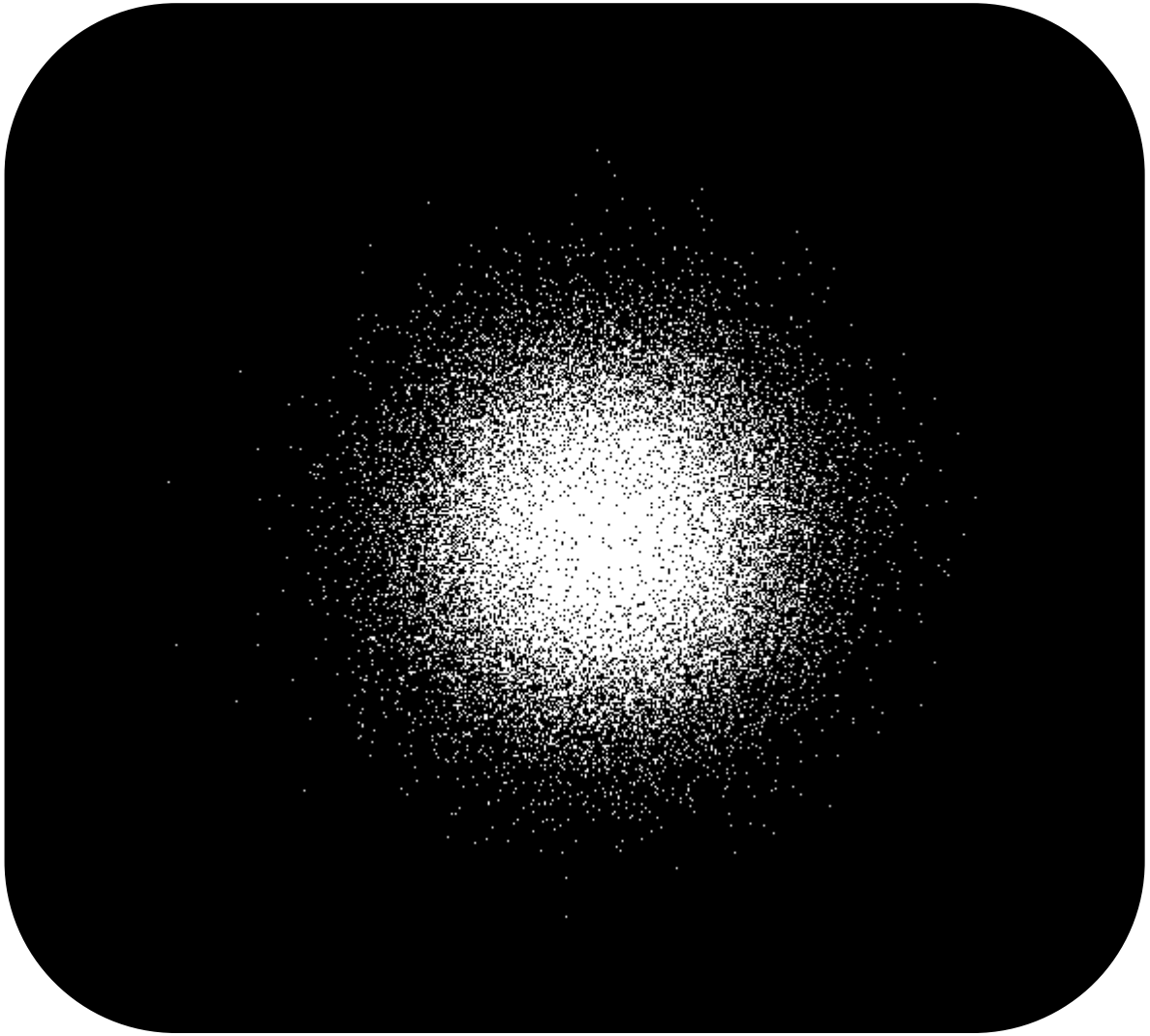
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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)

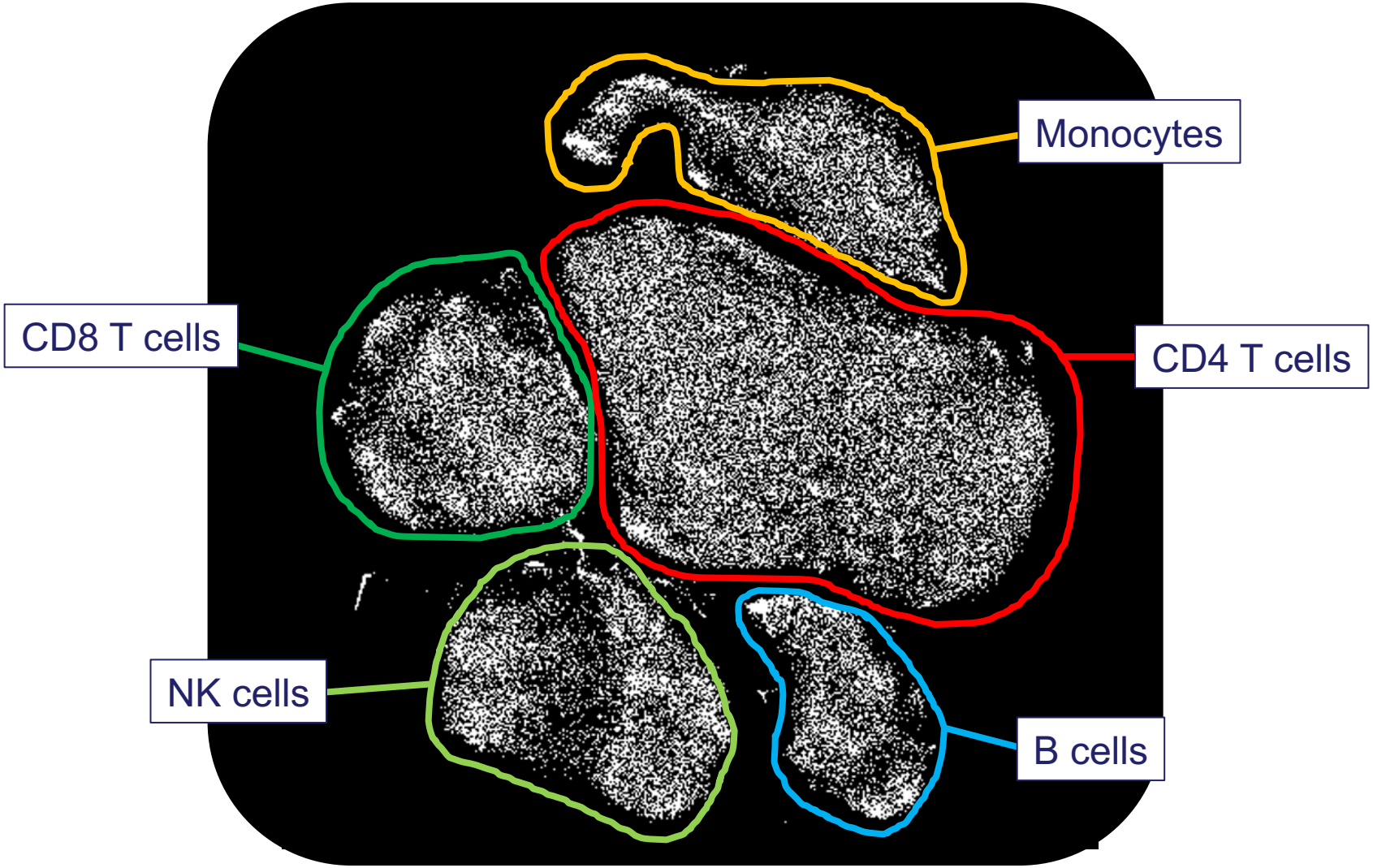
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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)

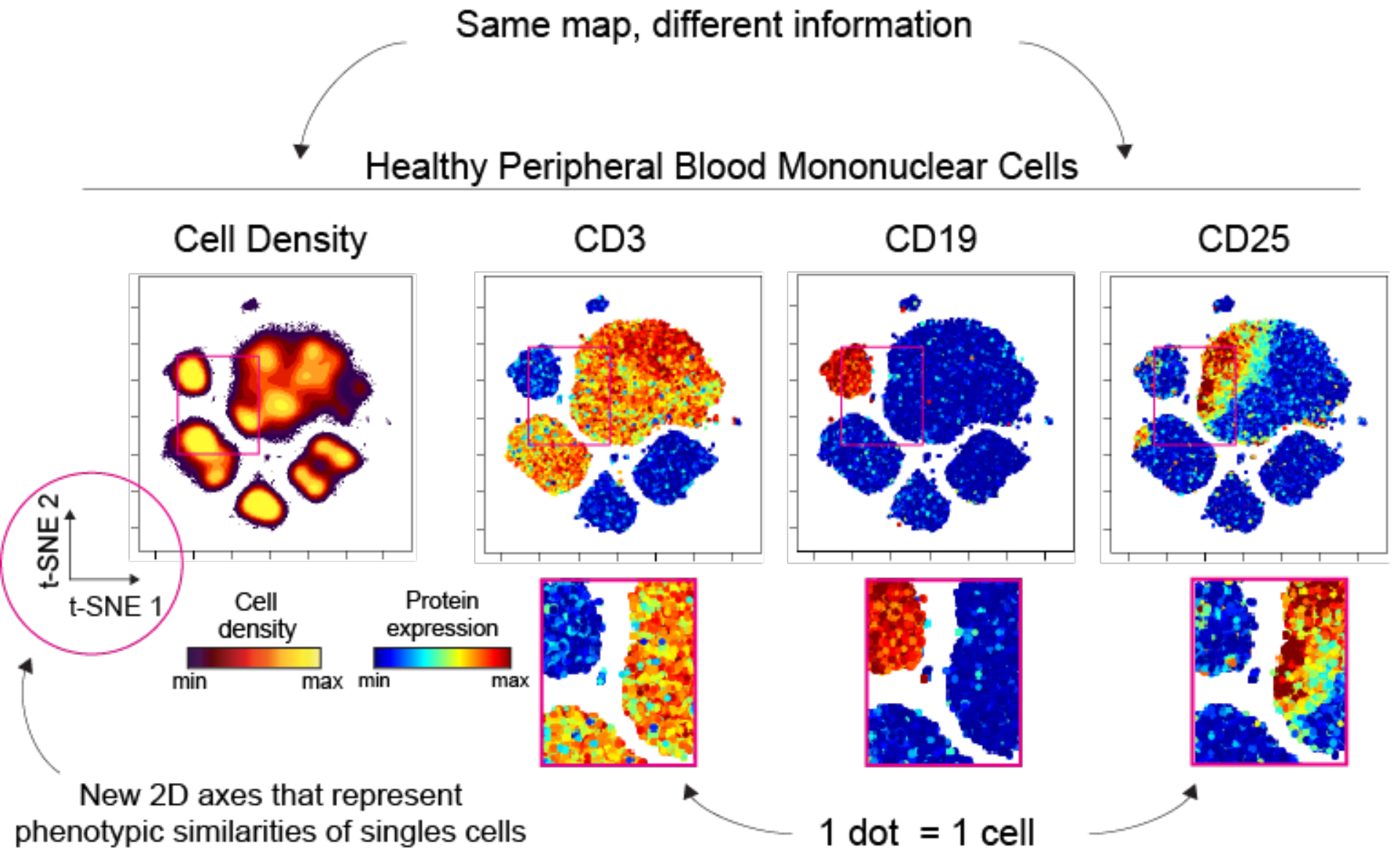
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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

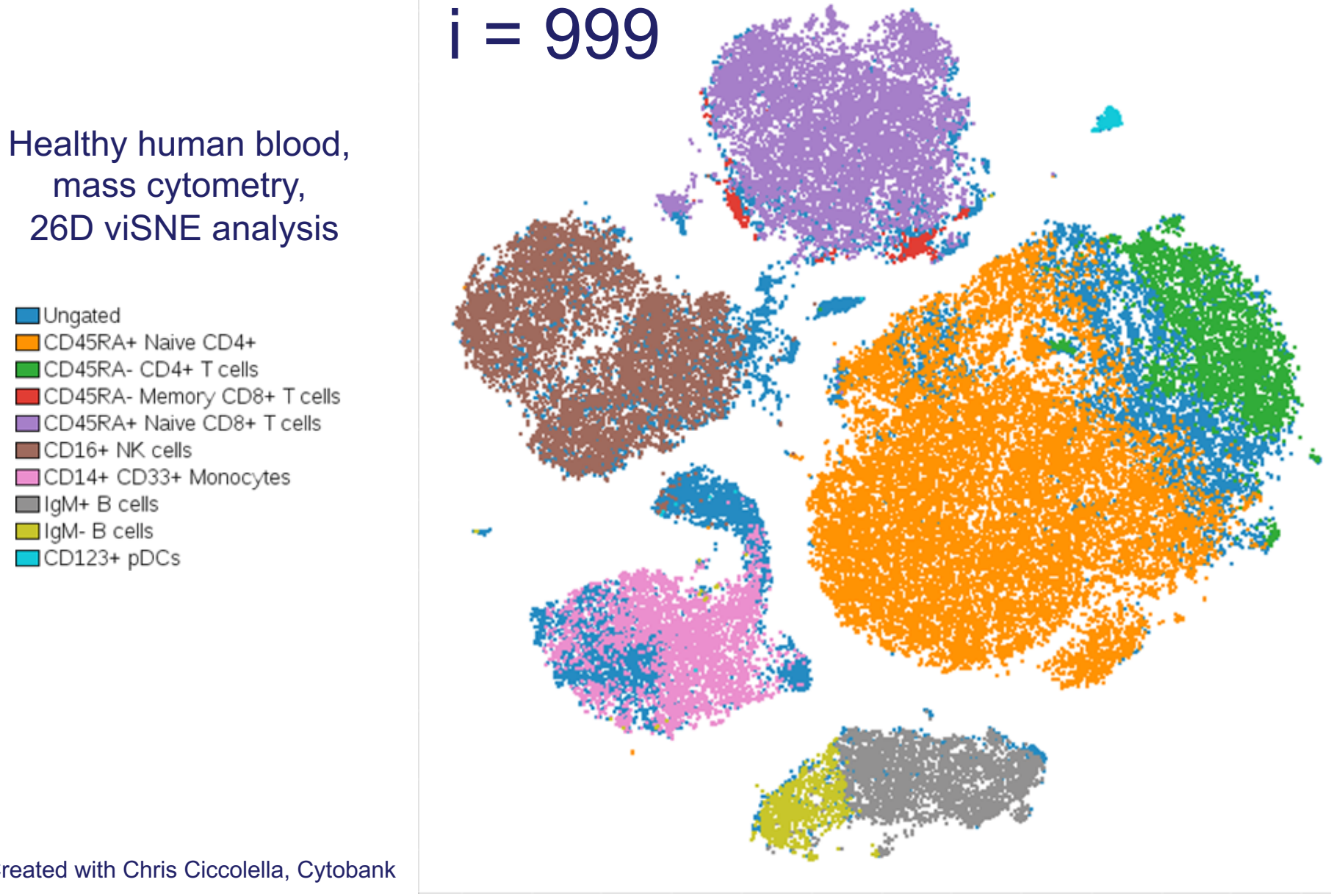
Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)

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



Viewing Expert Gates with viSNE Reveals Cyto Incognito

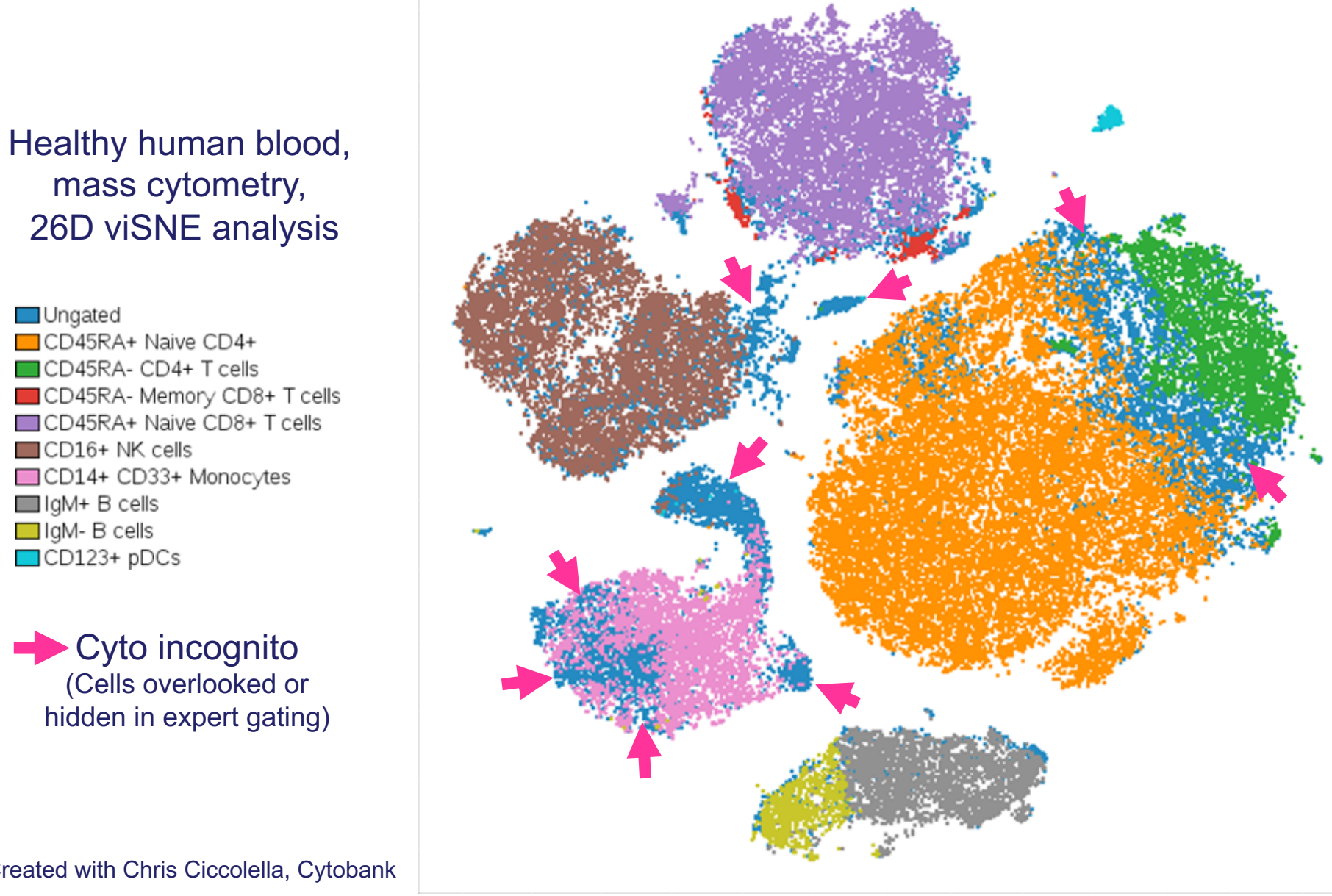
Healthy human blood,
mass cytometry,
26D viSNE analysis



Created with Chris Ciccolella, Cytobank

Viewing Expert Gates with viSNE Reveals Cyto Incognito

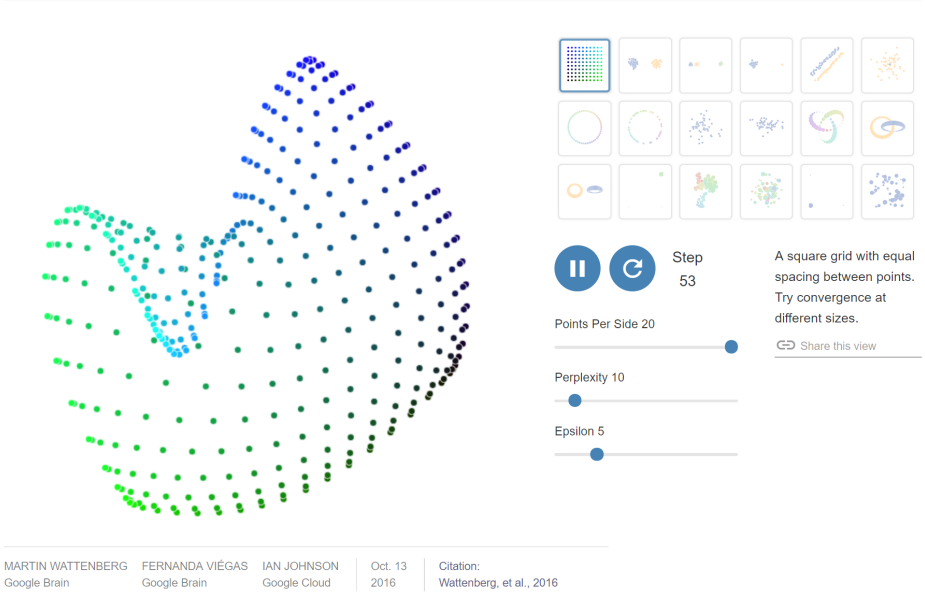
Healthy human blood,
mass cytometry,
26D viSNE analysis

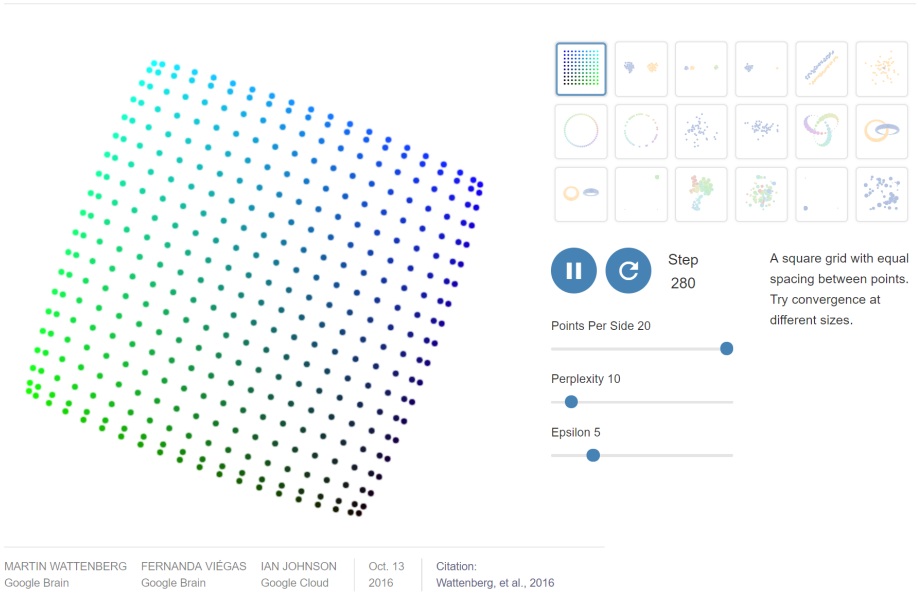


Created with Chris Ciccolella, Cytobank

t-SNE 2D Examples with Animations and Settings

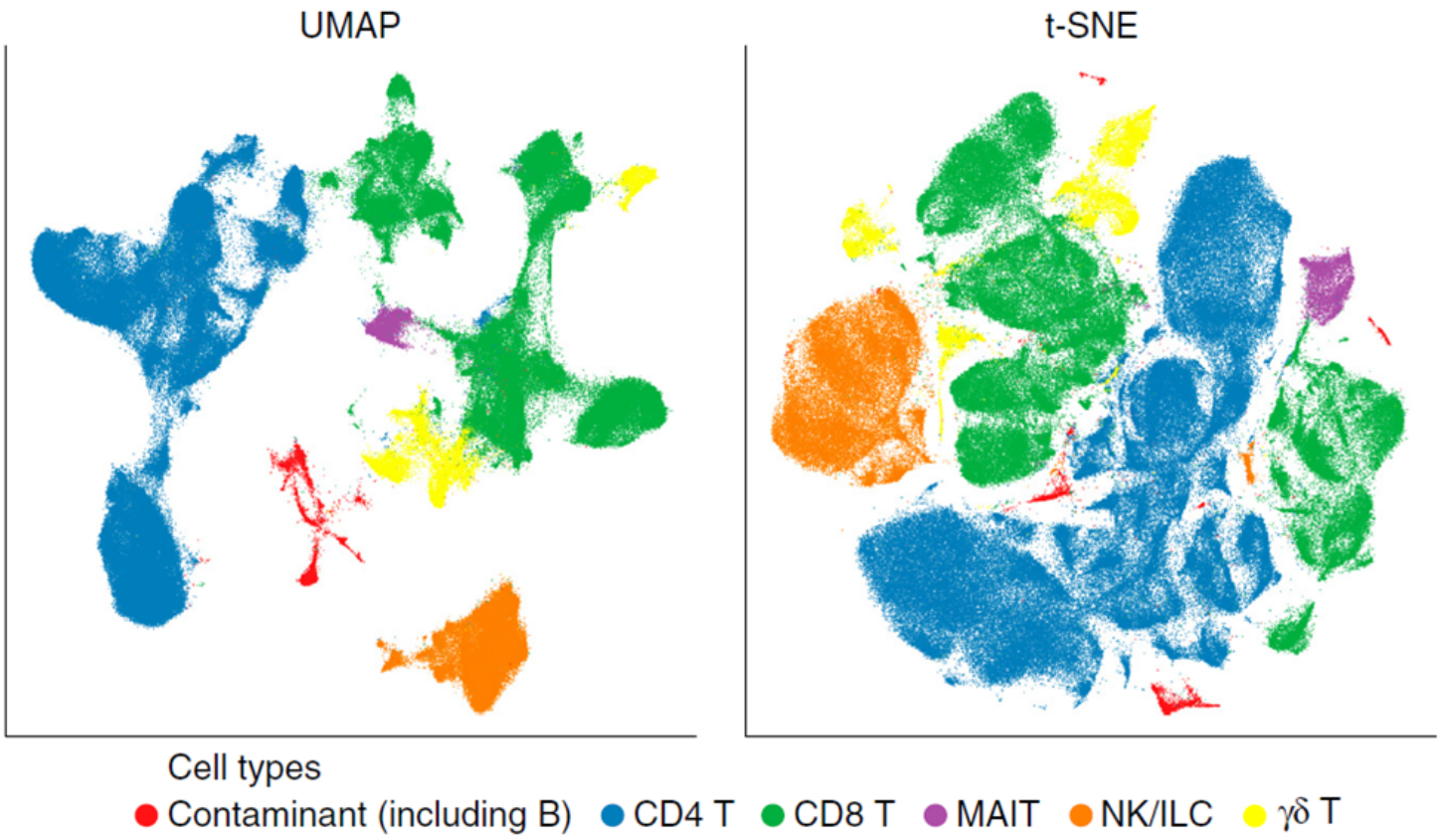
http://distill.pub/2016/misread-tsne/





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, γδ T cells, and contaminating cells

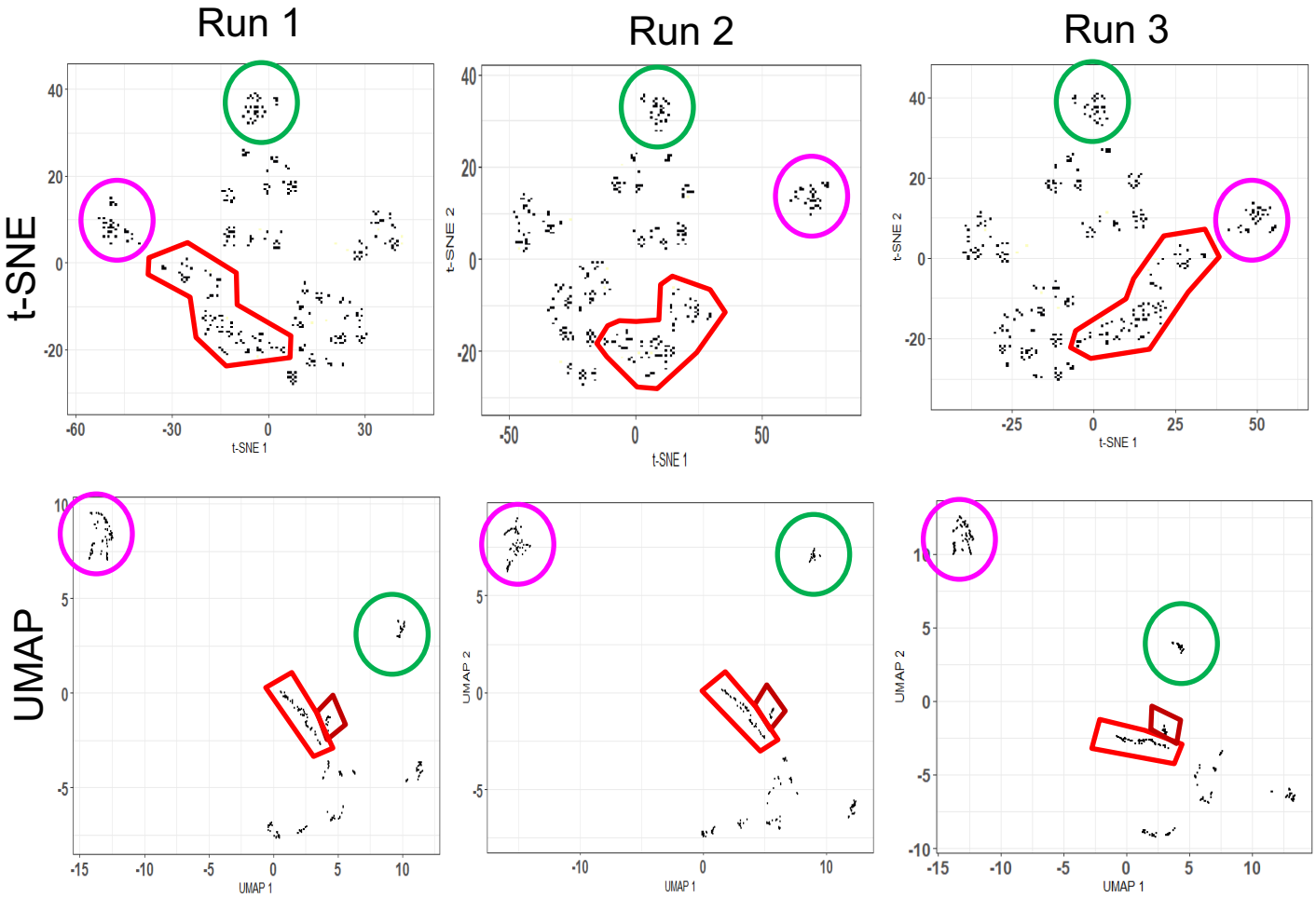


Dataset covering 35 samples originating from 8 distinct human tissues enriched for T and natural killer (NK) cells, of more than >300,000 cell events with 39 protein targets (Wong et al. dataset).

Becht et al., *Nature Biotechnology* 2018

Multiple Runs of t-SNE vs. UMAP on a Patient Dataset (n = 339)

Gandelman et al., cGVHD Patient Dataset



In the t-SNE plots, the relative relationship of the major islands ("global structure") alters between runs; t-SNE focuses on local structure

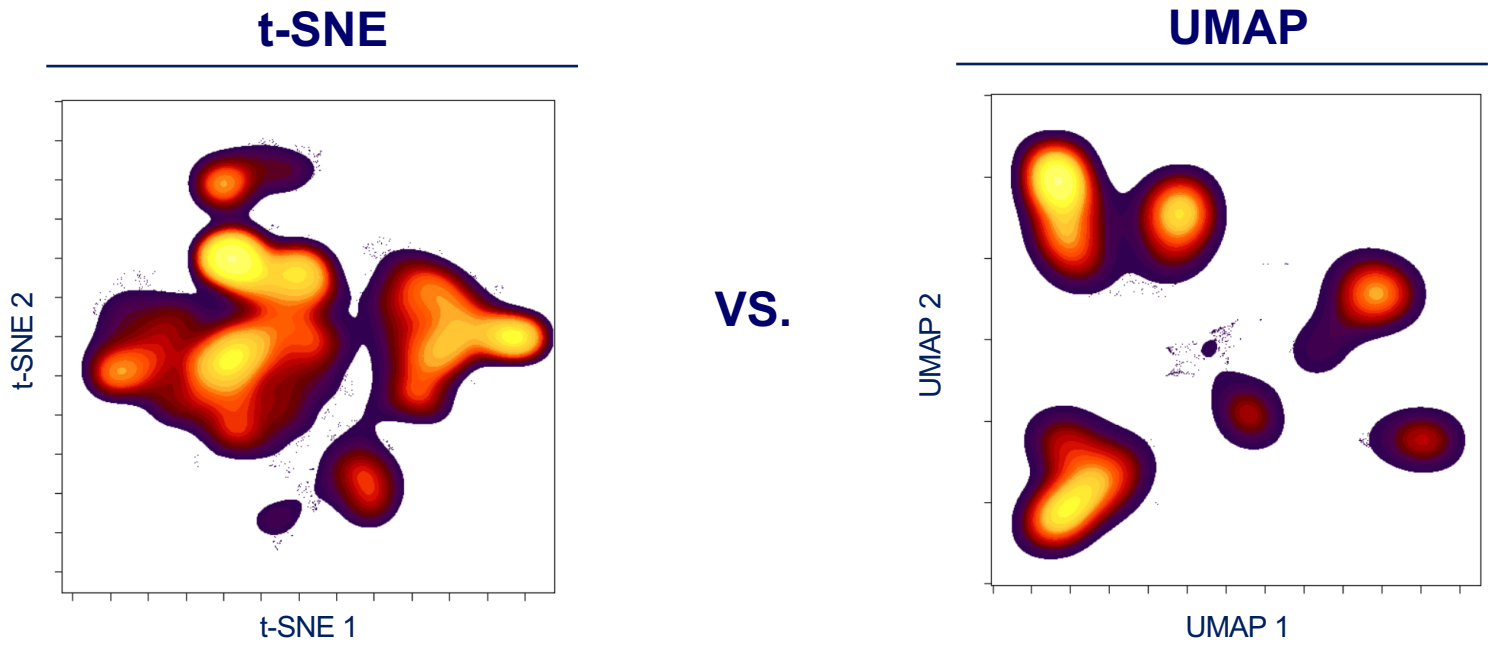
Relative island position ("global structure") is more stable & reflects original measurements in UMAP

Principle Component Analysis (PCA) is linear and deterministic, meaning that it strictly preserves global structure (and can overlook significant local structures / paths / trajectories)

UMAP (Uniform Manifold Approximation and Projection) is Another Dimensionality Reduction Tool

Superior run times

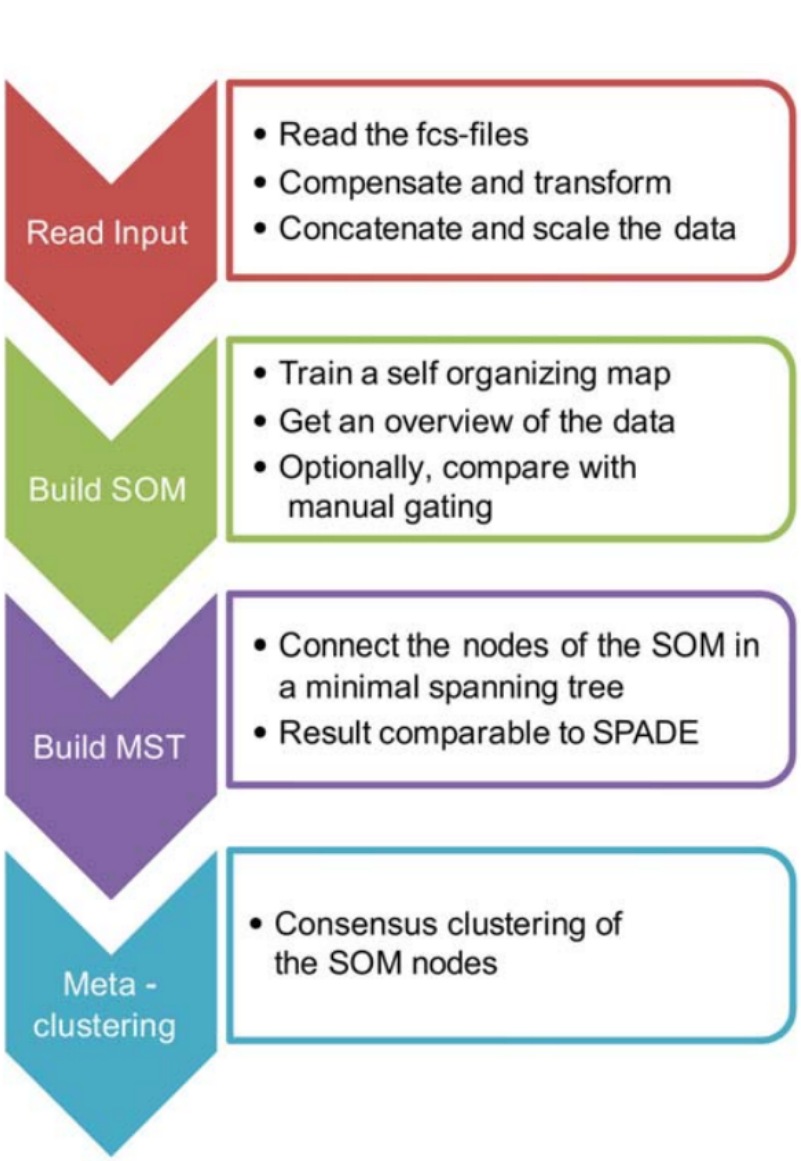
Emphasis on both global and local structure in the data

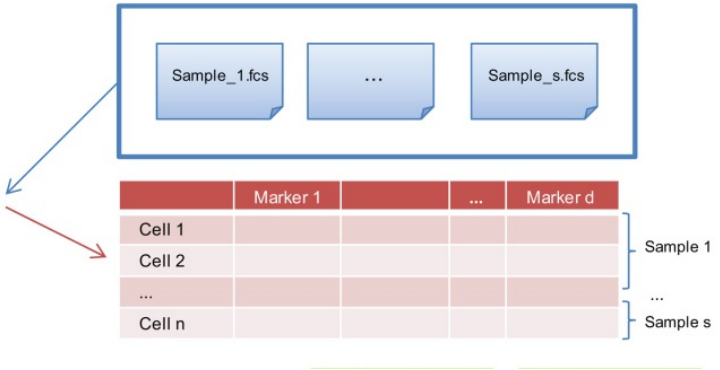


UMAP: McInnes et al., *arXiv* 2018; Becht et al., *Nature Biotech* 2019

Data used for t-SNE and UMAP: human PBMC

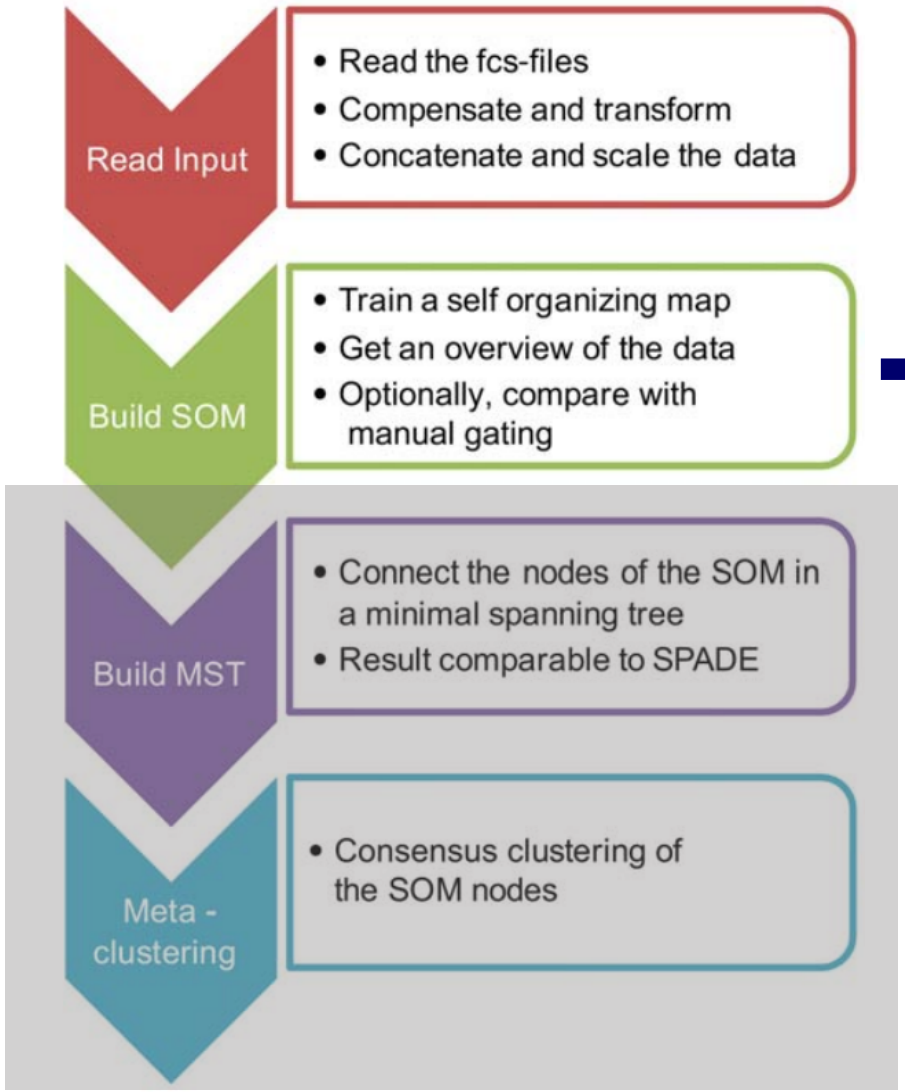
Clustering with FlowSOM: Self-organizing Maps





Van Gassen et al., *Cytometry A* 2015
See also, FlowSOM does well in a comparison of clustering tools: Weber & Robinson, *Cytometry A* 2017

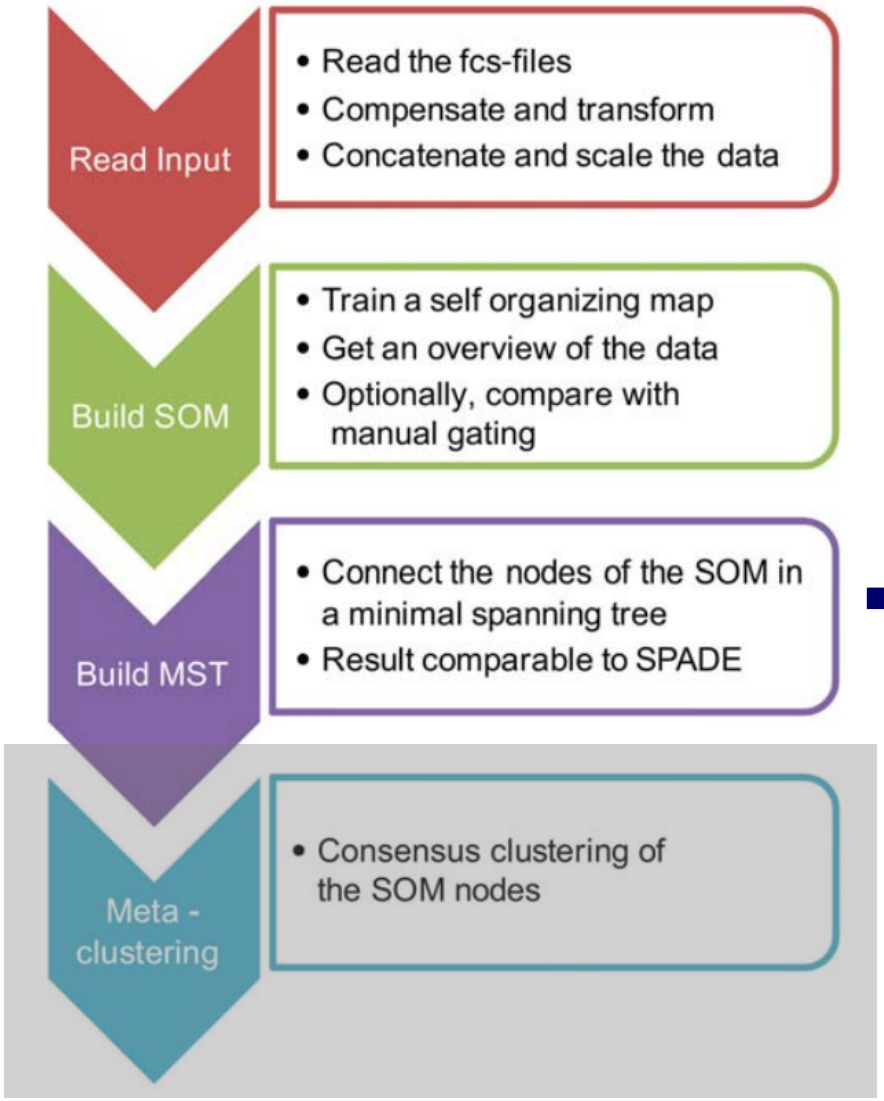
Clustering with FlowSOM: Self-organizing Maps



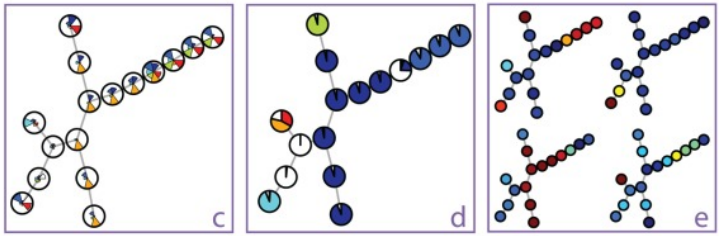
A SOM is created by assigning data points/cells to nodes based on their multidimensional phenotypes, updating this map repeatedly until each cell is assigned to a node with the most similar cells

Van Gassen et al., *Cytometry A* 2015
See also, FlowSOM does well in a comparison of clustering tools: Weber & Robinson, *Cytometry A* 2017

Clustering with FlowSOM: Self-organizing Maps

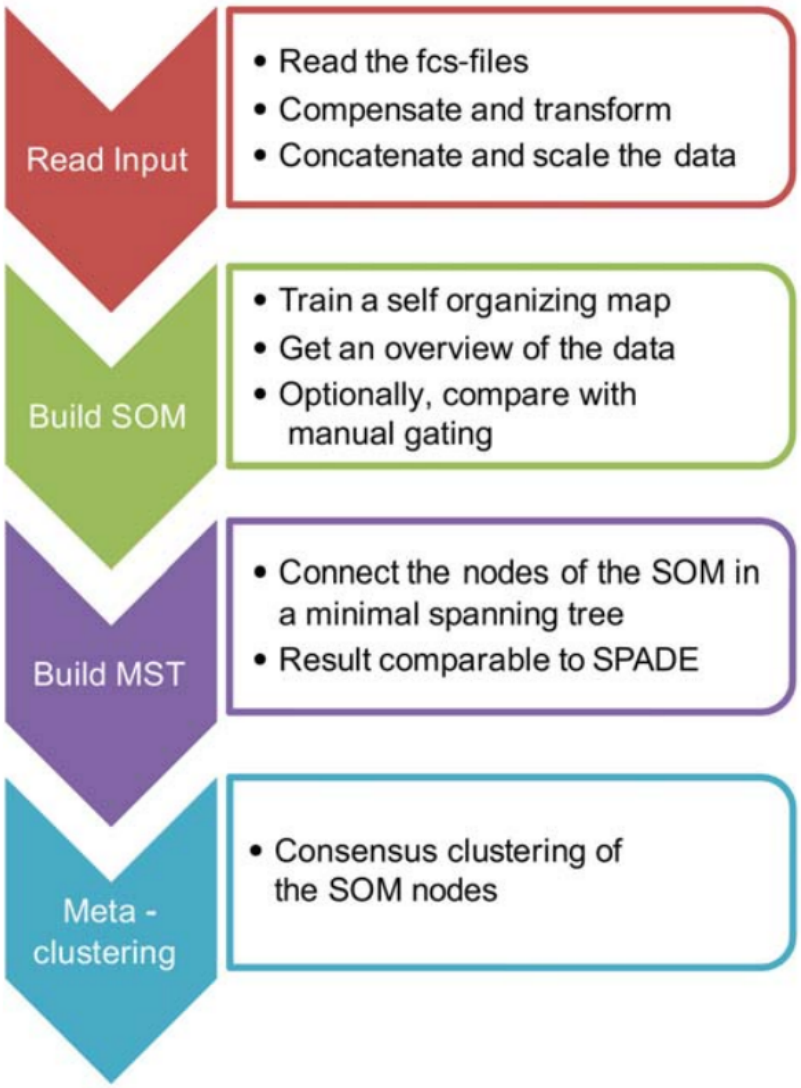


The next step is to arrange the nodes along a minimal spanning tree (MST), so that nodes that are most similar are closest on the tree
not used in our visualization



Van Gassen et al., *Cytometry A* 2015
See also, FlowSOM does well in a comparison of clustering tools: Weber & Robinson, *Cytometry A* 2017

Clustering with FlowSOM: Self-organizing Maps

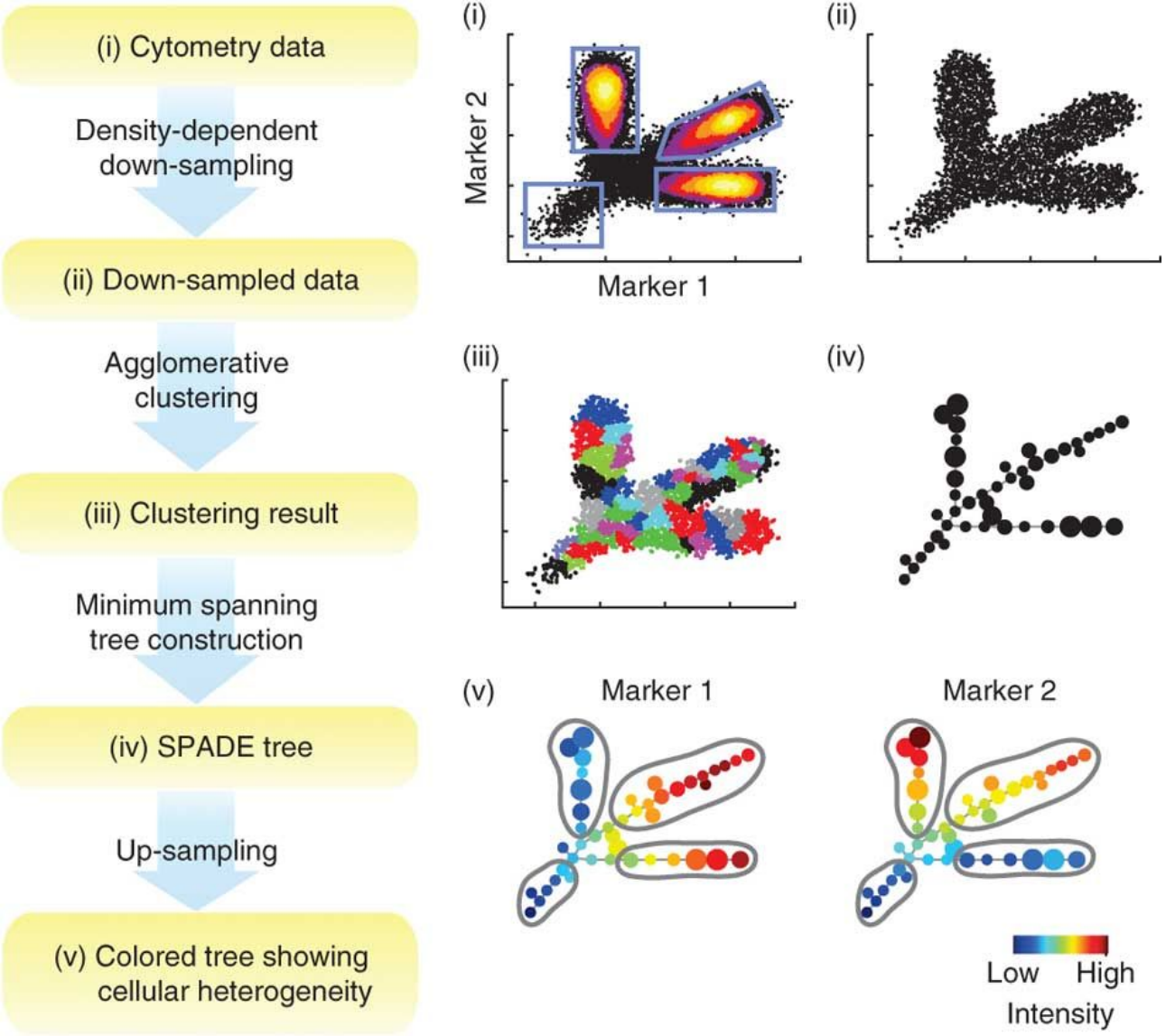


Finally, similar nodes are combined based on the number of desired clusters defined by the user. This desired number can be based on prior knowledge or a specific goal (i.e. minimizing intracluster variance)

Van Gassen et al., *Cytometry A* 2015

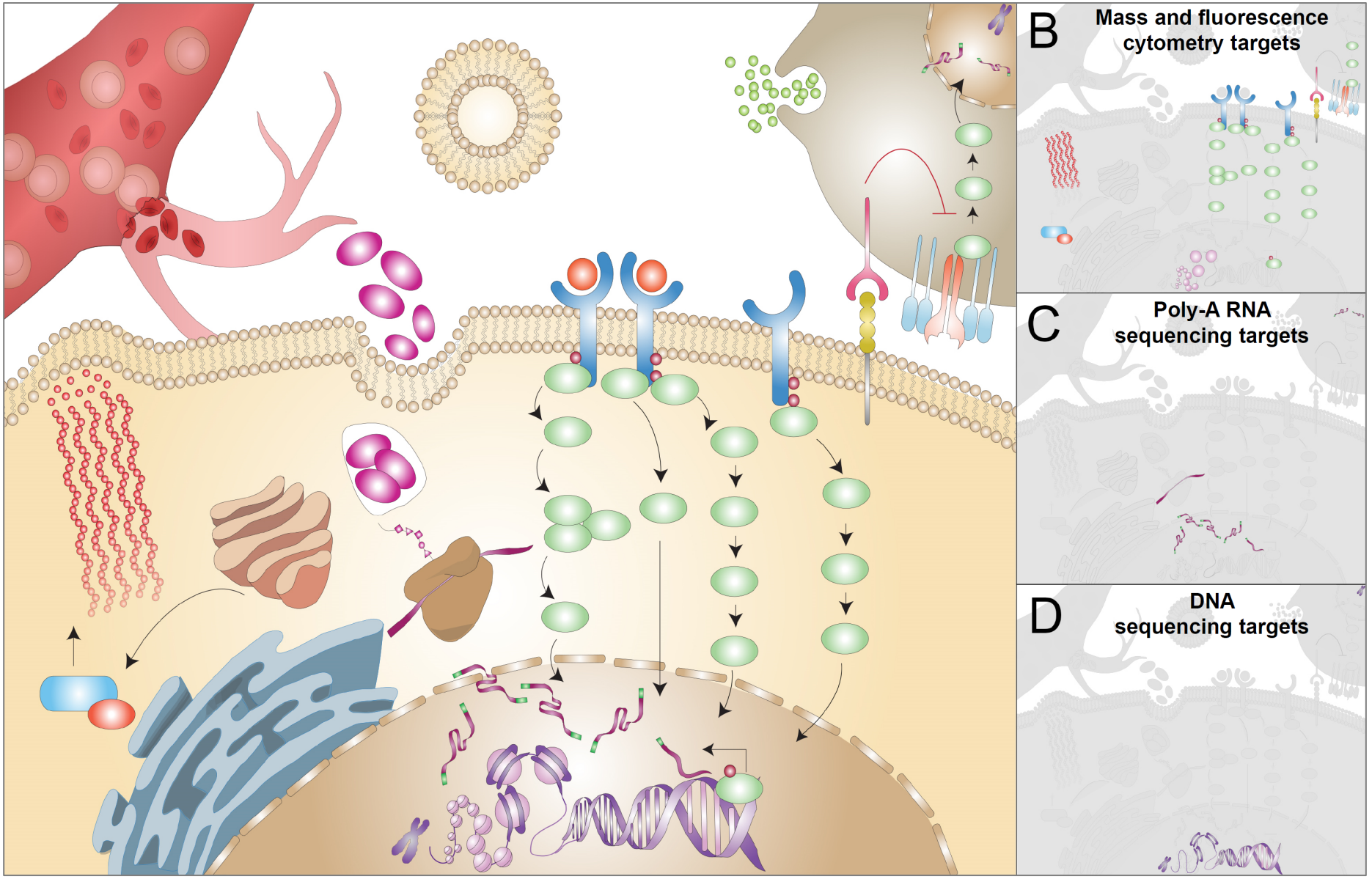
See also, FlowSOM does well in a comparison of clustering tools: Weber & Robinson, *Cytometry A* 2017

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



Qui et al., *Nature Biotechnology* 2011

Challenge to the Field: Multiplex Across Cell Functions

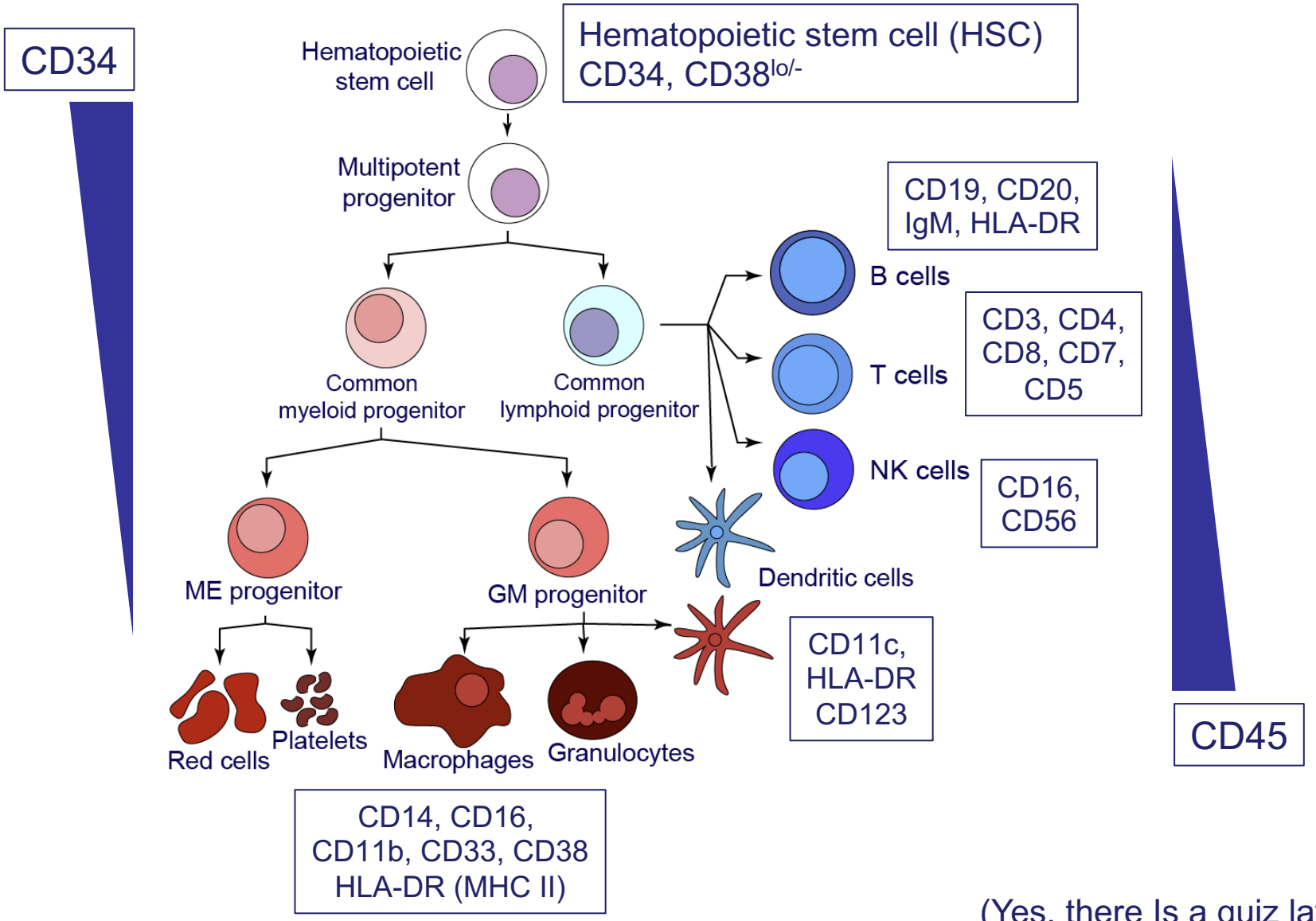


Beyond the Message: Advantages of Snapshot Proteomics with Single-Cell Mass Cytometry in Solid Tumors
Mistry et al., *FEBS Journal* 2018

MEM summarizes a population's special features
and is used in workflows "at the end"
(in place of box and whisker plots or heatmaps)

[So MEM complements tools from other steps, including
t-SNE, SPADE, Citrus, FlowSOM, SCAFFOLD, Phenograph]

Human Bone Marrow Hematopoiesis & “Famous” Cell Identity Markers



(Yes, there Is a quiz later)

Despite advances, no computational tools learn & label cell identity,
a human must “stare and compare” using expert knowledge

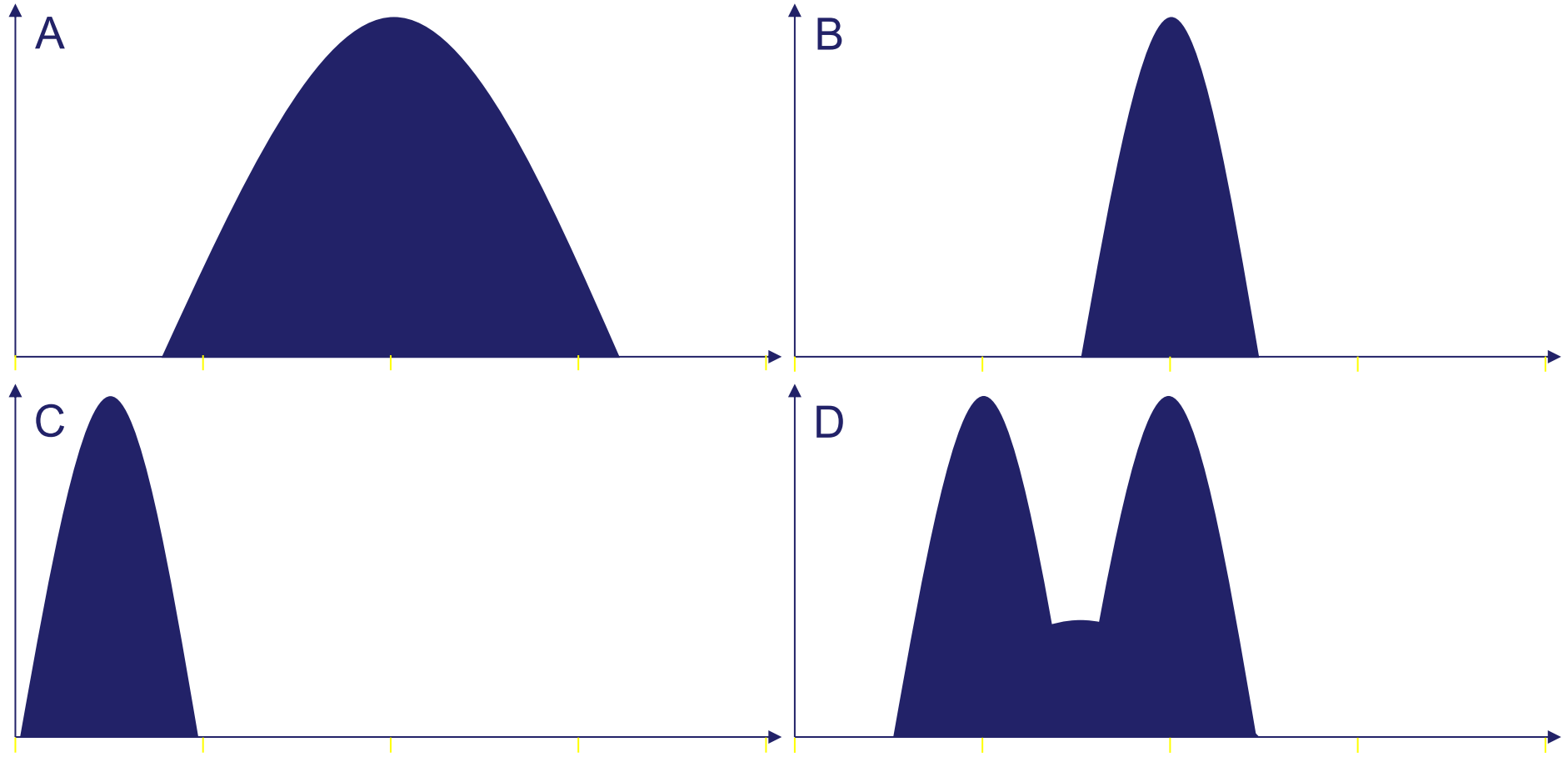
Diggins et al., *Methods* 2015

Populations are often labeled by metaphors of function
(“cancer stem cells”, “central memory T cells”)
or incomplete labels based on a few features (e.g. “PD-1+ CD8 T cells”).

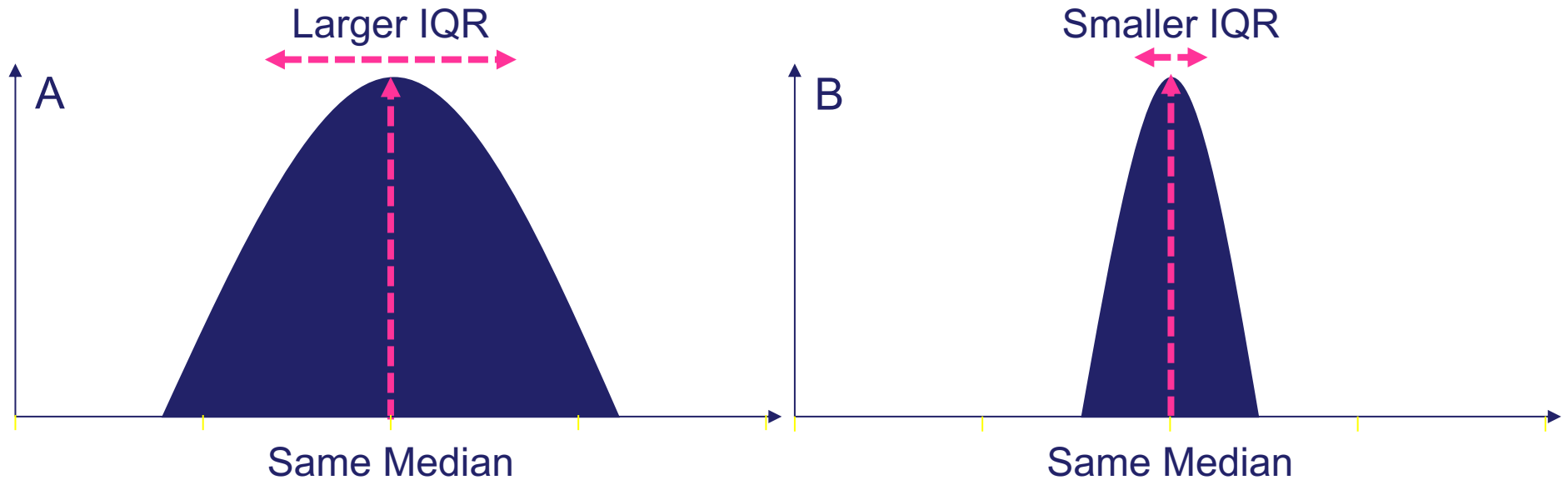
We need an unbiased way to label & identify cells
(regardless of how they are found)

Enrichment Tracks Feature Exclusivity In a Subset

A, B, C, and D are 4 subsets where Protein X was measured.
In which subset is Protein X most distinct? (Which would be easiest to gate?)



Median (50%) and Interquartile Range (25%-75%) Represent Key Features of Distributions



Core idea in MEM: given two protein distributions with equal medians, a smaller interquartile range (IQR) indicates greater enrichment

Not captured by median & IQR are other elements of shape (skewness, symmetry, # peaks, outliers, etc.)

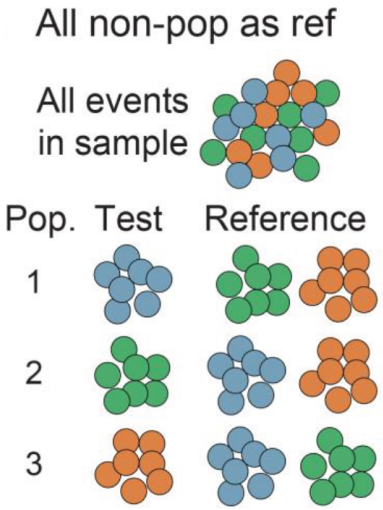
Diggins et al., *Nature Methods* 2017

MEM Quantifies Relative Enrichment By Combining Magnitude & Interquartile Range

$$MEM = |MAG_{test} - MAG_{ref}| + \frac{IQR_{ref}}{IQR_{test}} - 1$$

Linear transformation to -10 to +10
(d20 scale, cause that's how we roll)

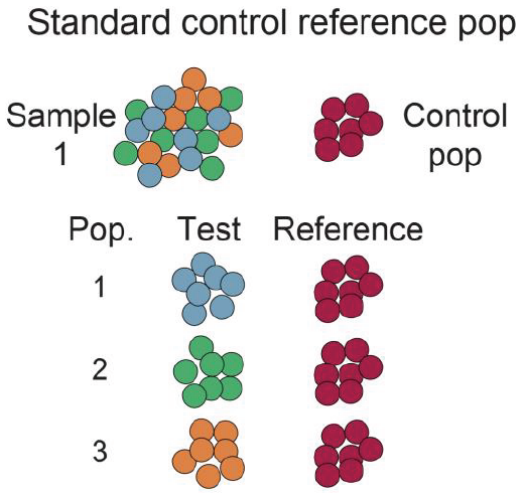
If $MAG_{test} - MAG_{ref} < 0$, MEM = -MEM



MEM label (CD19^{+} cells)

▲HLADR^{+10} CD20^{+9} CD19^{+7} IgM^{+5} C
CD45RA^{+3} CXCR4^{+2} CD47^{+2} CD33

▼CD7^{-2}



Diggins et al., *Nature Methods* 2017

Quiz Time: What Are These Cell Subsets & What Is This Tissue?

Stem cells (HSCs)

▲CD34^{+6} CD33^{+4} CD15^{+3} CD38^{+3} MHCII^{+3} CXCR4^{+2}
▼CD44^{-5} CD45^{-5} CD7^{-3} 0.07%

Natural killer cells

▲CD16^{+9} CD7^{+6} CD38^{+5} CD56^{+4} CD161^{+4} CD45RA^{+3} CD8^{+2} CD11b^{+2} CD47^{+2} 5.27%

Progenitors

▲MHCII^{+10} CD33^{+7} CD38^{+5} CD123^{+3} CD117^{+3} CD19^{+2} CD34^{+2} CD13^{+2} CD14^{+2} CXCR4^{+2}
▼CD45^{-3} CD15^{-2} 0.002%

CD8^{+} T cells

▲CD8^{+8} CD7^{+5} CD3^{+3} CD45RA^{+3} CXCR4^{+2} 9.25%

Early myeloid cells

▲MHCII^{+9} CD33^{+8} CD38^{+5} CD4^{+3} CD15^{+2} CD14^{+2}
▼ CD45^{-2} CD7^{-2} 0.02%

CD4^{+} T cells

▲CD4^{+7} CD7^{+5} CD3^{+5} CD47^{+2} CD45RA^{+2} 8.12%

Monocytes

▲CD33^{+10} CD14^{+8} CD11b^{+7} MHCII^{+5} CD4^{+4} CD11c^{+4} CD38^{+4} CD13^{+3}
▼ CXCR4^{-2} CD47^{-2} 10.57%

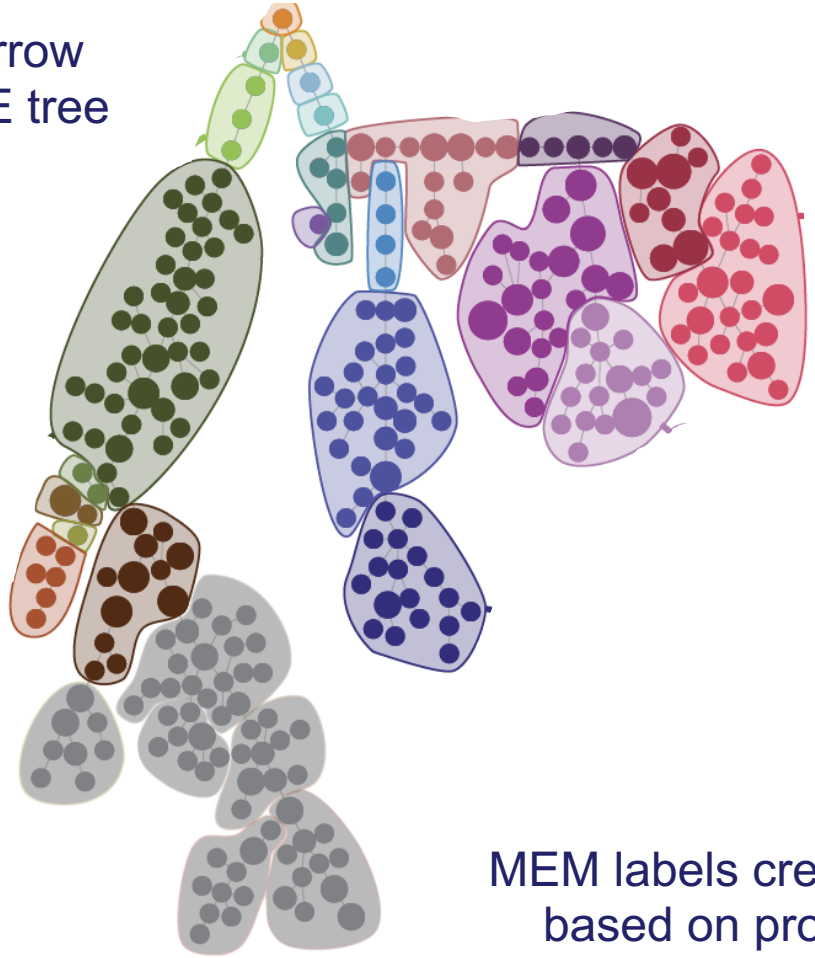
B cells

▲MHCII^{+10} CD20^{+9} CD19^{+7} IgM^{+5} CD34^{+3} CD45RA^{+3} CXCR4^{+2} CD47^{+2} CD33^{+2}
▼CD7^{-2} 2.44%

Data from healthy human bone marrow, Bendall et al., *Science* 2011
Diggins et al., *Nature Methods* 2017

Marker Enrichment Modeling Automatically Labels Cell Types in Human Bone Marrow Using -10 to +10 Enrichment Values

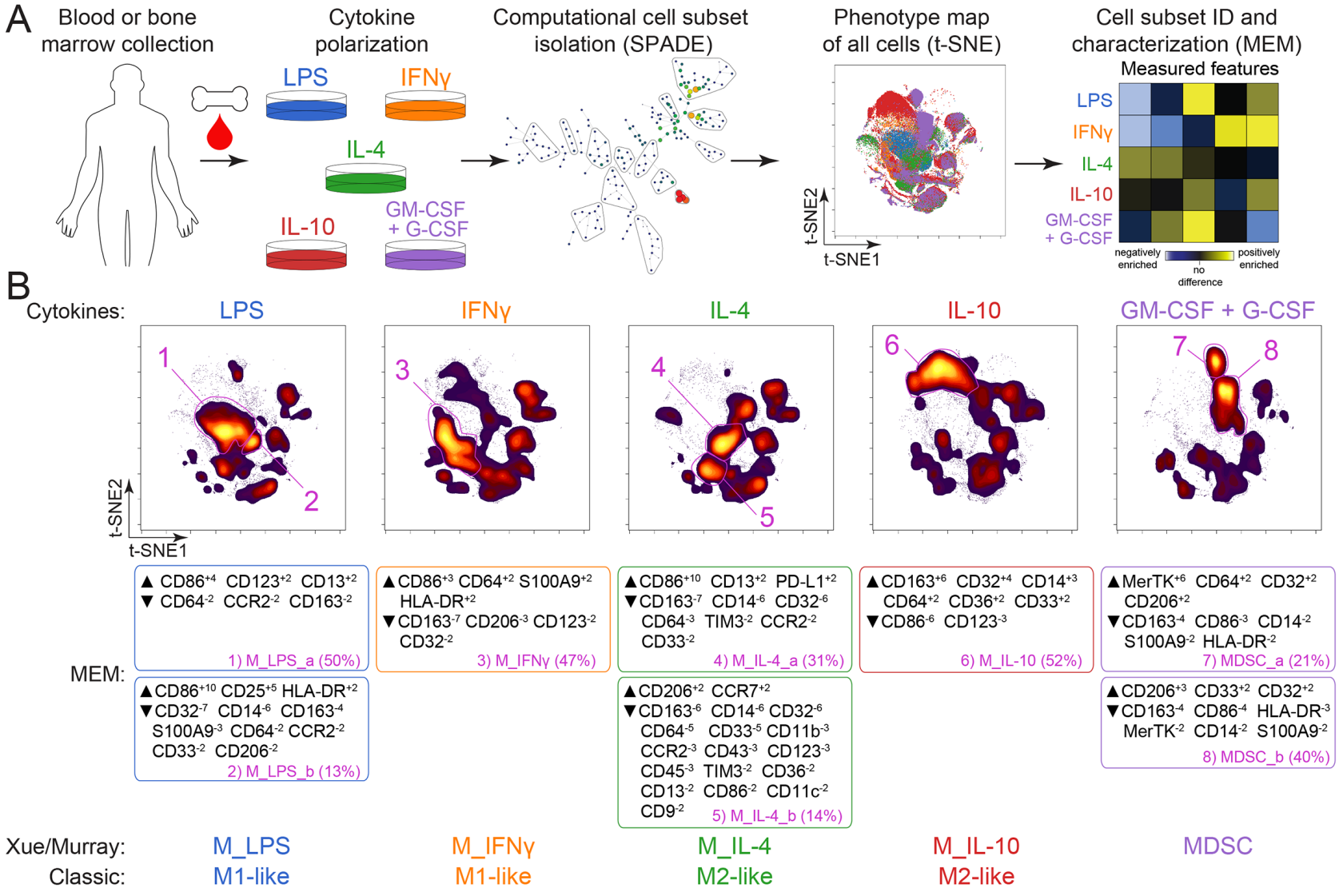
Cells from bone marrow grouped in a SPADE tree



MEM labels created automatically based on protein enrichment

Diggins et al., *Nature Methods* 2017
Data from Bendall et al., *Science* 2011

Functional Profiling: Polarization Capacity of Monocytes

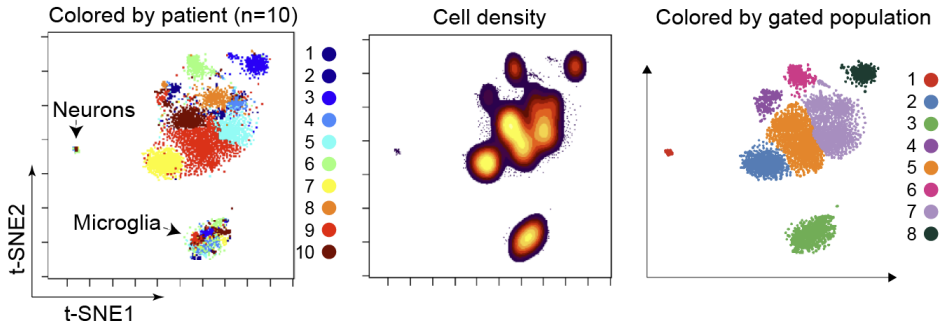


http://flowrepository.org/id/FR-FCM-Z2Z8

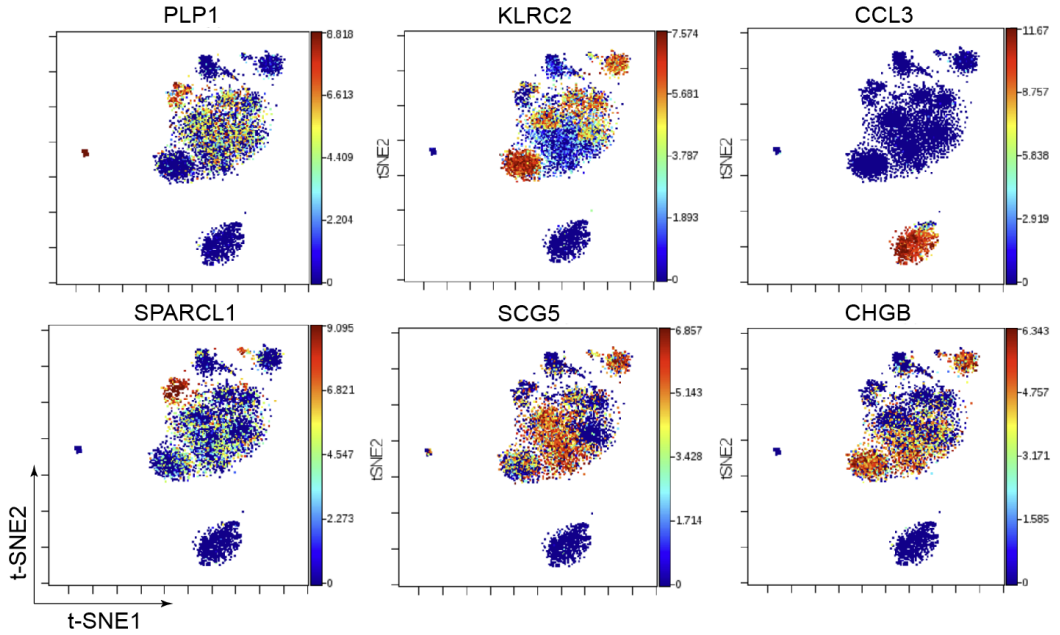
Roussel et al., *Methods in Molecular Biology* 2019, *in press*
Original work: Roussel et al., *Journal of Leukocyte Biology 2017*

MEM Classification of Cell Subsets in IDH-A Brain Tumors Based on scRNA-seq Transcript Expression

IDH-A brain tumor cells from 10 patients
Gating for 8 cell types: t-SNE using 500 most variable transcripts, gating on local density



Per-cell transcript expression, top 6 most positively enriched in gated populations



MEM Label, Population 3:

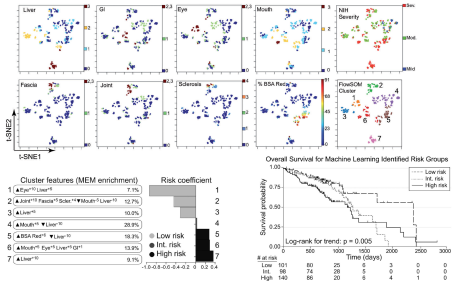
▲SAT1^{+6} CCL3^{+5} CCL4^{+5} CD74^{+5} HLA-DRA^{+5} RGS1^{+5} SPP1^{+5}

▼GFAP^{-8} OMG^{-8} APOD^{-7} KLRC2^{-7} SCG5^{-7} UCHL1^{-7}

(ID'd as microglia)

Data from leukemia patient blood, Venteicher et al., *Science* 2017
MEM for scRNA-seq: Diggins et al., *in preparation*

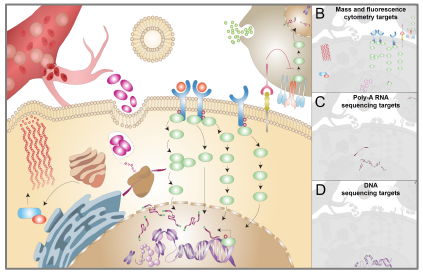
Adapting Advances in Machine Learning & Single Cell Biology for Immuno-Oncology



Haematologica 2018

1. Gandelman et al., Machine Learning Reveals Chronic Graft-Versus-Host Disease Phenotypes and Stratifies Survival After Stem Cell Transplant for Hematologic Malignancies. ***Haematologica*** 2018 PMC6312024.

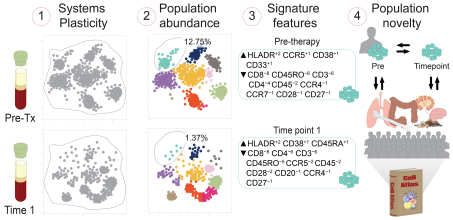
   Machine learning for patient medical phenotypes including t-SNE, FlowSOM, and MEM. Workflow inspired RAPID algorithm shown here for solid tumor cells.



FEBS J 2018

2. Mistry et al., Beyond the Message: Advantages of Snapshot Proteomics with Single-Cell Mass Cytometry in Solid Tumors. ***FEBS J*** 2018 Dec 13. PMID: 30549207 DOI: 10.1111/febs.14730.

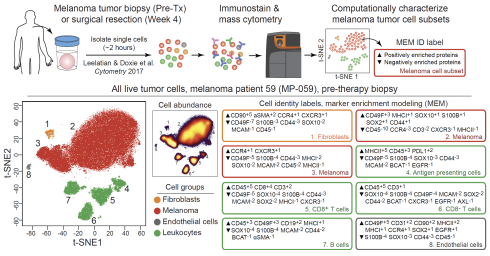
   Reviews single cell mass cytometry with an emphasis on solid tumor and tissue research. Provides a detailed comparison with complementary single cell technologies, including single cell RNA-sequencing and quantitative multiplex imaging.



Cancer Immunology Research 2019

3. Greenplate et al., Computational immune monitoring reveals abnormal double negative T cells present across human tumor types. ***Cancer Immunology Research*** 2019 Jan;7(1):86-99. PMC6318034.

   New computational workflow for longitudinal single cell tumor immunology. Revealed abnormal immune cells present in multiple tumor types. Will be applied to solid tumors to track changes in immune microenvironment.

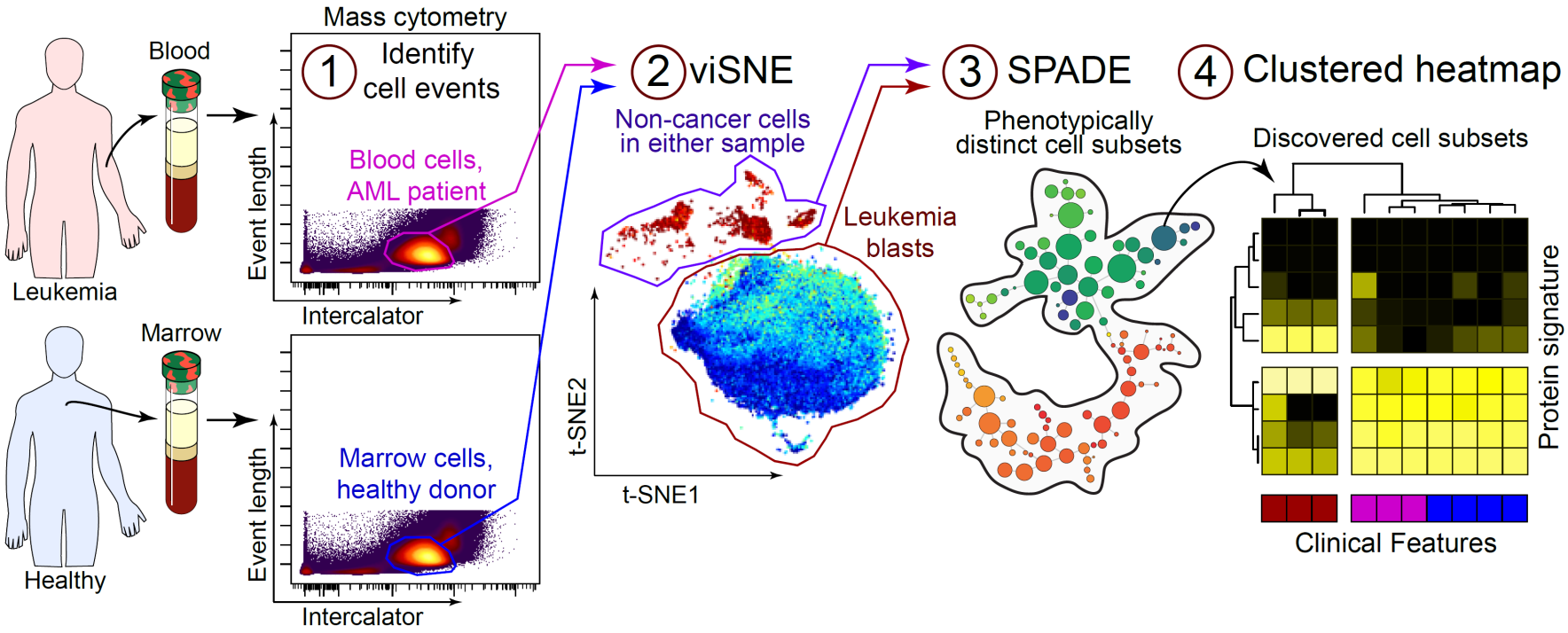


Pigment Cell & Melanoma Research 2018

4. Doxie et al., BRAF and MEK inhibitor therapy eliminates Nestin-expressing melanoma cells in human tumors. ***Pigment Cell Melanoma Research***. 2018 Jun 28. PubMed PMID: 29778085; PubMed Central PMCID: PMC6188784.

   Compared biopsies over time of melanoma patients on treatment with mass cytometry to reveal *in vivo* cellular changes on treatment following targeted therapy.

Machine Learning Tools Can Automate Sub-Population (Cluster / Subtype) Characterization

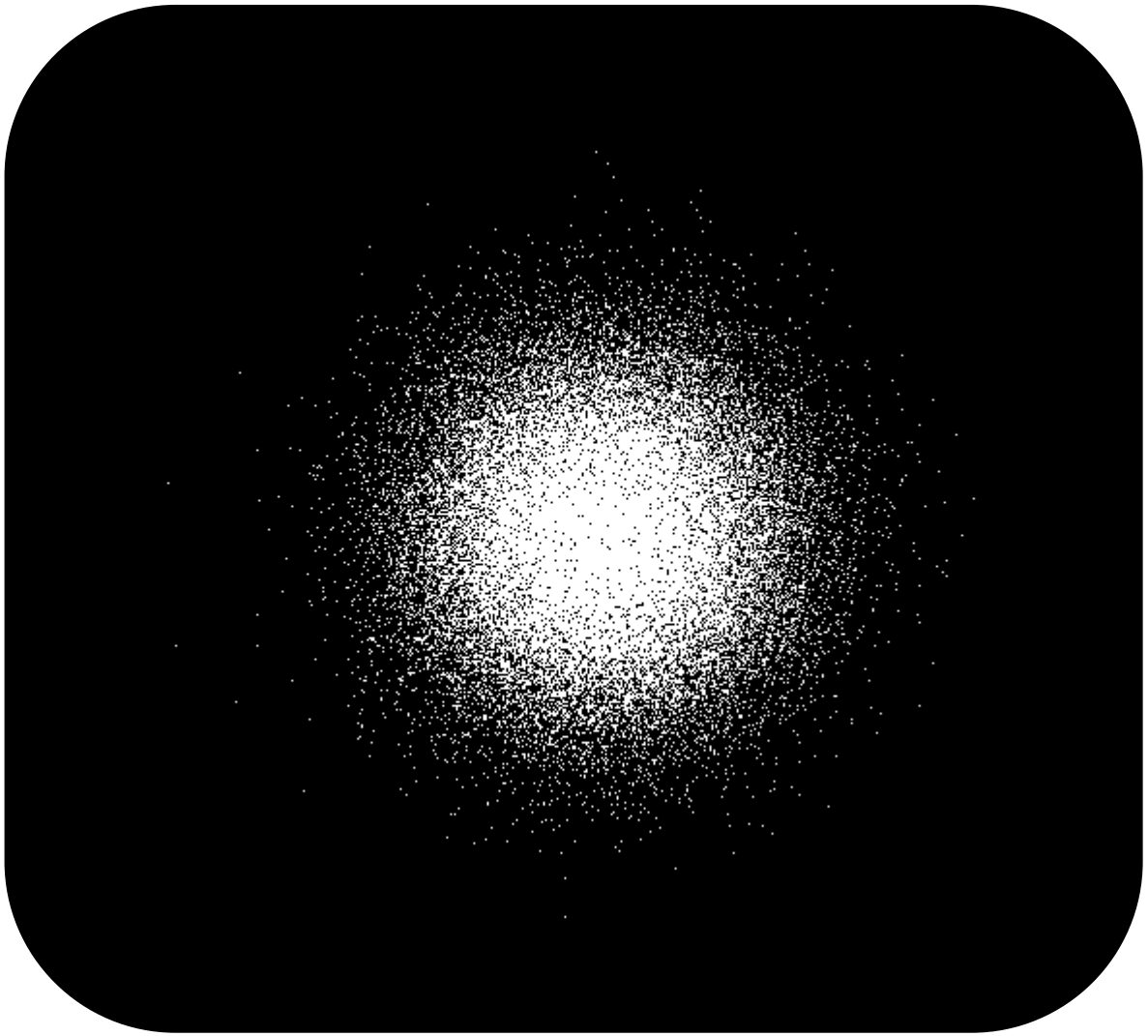


Protocol: Diggins et al., *Current Protocols in Cytometry* 2017
Data files: https://flowrepository.org/id/FR-FCM-ZZKZ
More great tools: Saeys et al, *Nature Reviews Immunology* 2016

Workflow: Diggins et al., *Methods* 2015
viSNE/t-SNE: Amir et al., *Nat Biotech* 2013
SPADE: Qiu et al., *Science* 2011

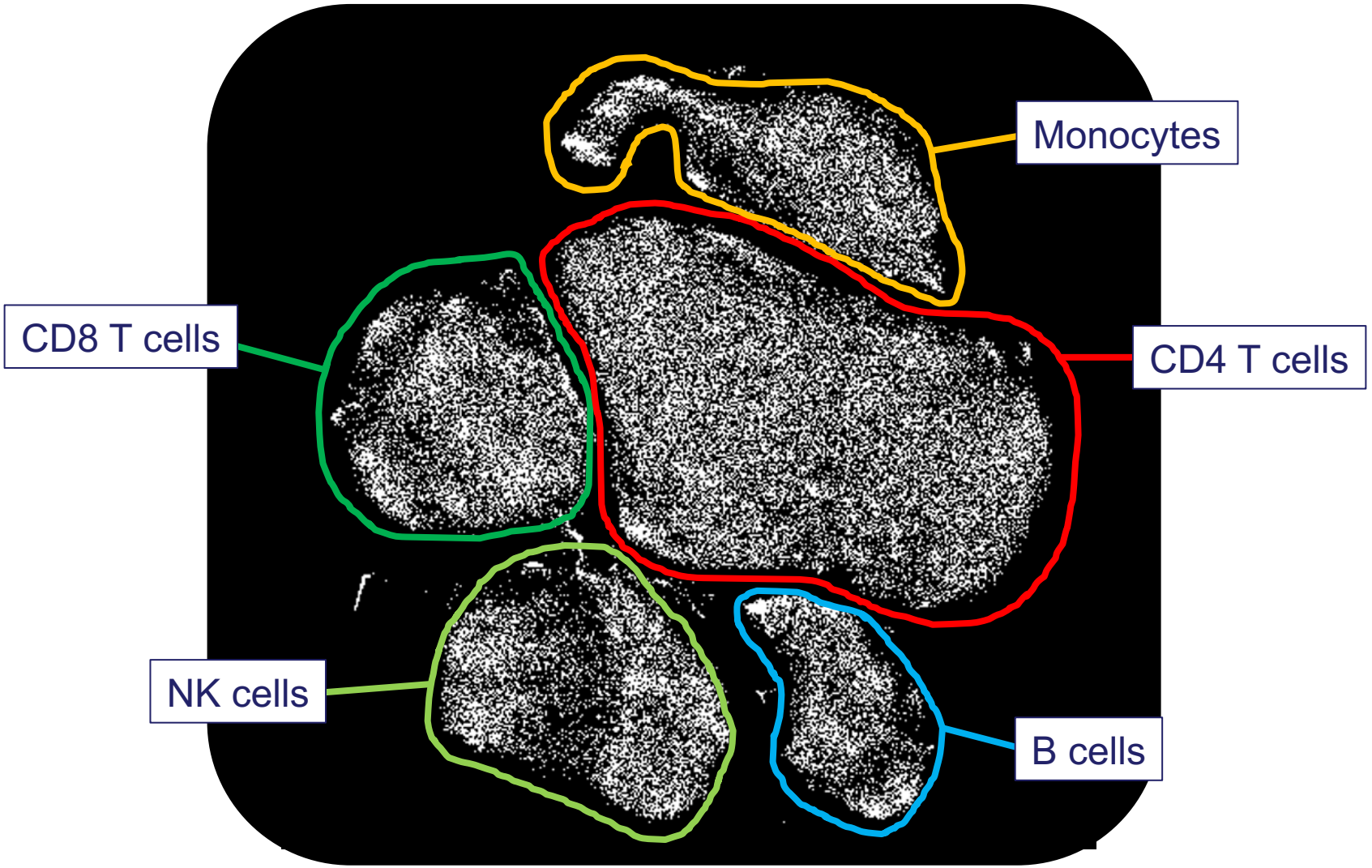
For sub-populations: quantify the known, reveal the unexpected, characterize the abnormal, & evaluate any associated risk

Key Algorithm Concept: t-SNE Visualizes (in 2D) Close Phenotypic Relationships Using Many Markers



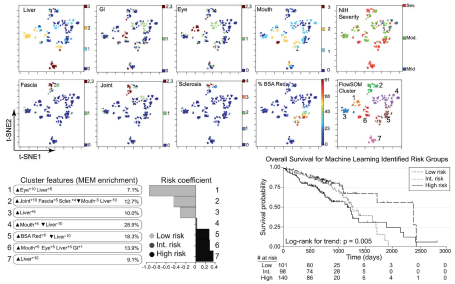
Healthy human blood, mass cytometry, 26 markers measured, viSNE / t-SNE analysis tool
(Animation by Cytobank: each movie frame shows one iteration of t-SNE algorithm)

Key Algorithm Concept: t-SNE Visualizes (in 2D) Close Phenotypic Relationships Using Many Markers



Healthy human blood, mass cytometry, 26 markers measured, viSNE / t-SNE analysis tool
(Animation by Cytobank: each movie frame shows one iteration of t-SNE algorithm)

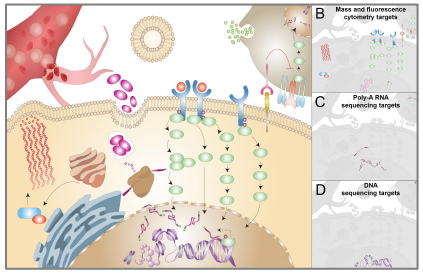
Adapting Advances in Machine Learning & Single Cell Biology for Immuno-Oncology



Haematologica 2018

1. Gandelman et al., Machine Learning Reveals Chronic Graft-Versus-Host Disease Phenotypes and Stratifies Survival After Stem Cell Transplant for Hematologic Malignancies. ***Haematologica*** 2018 PMC6312024.

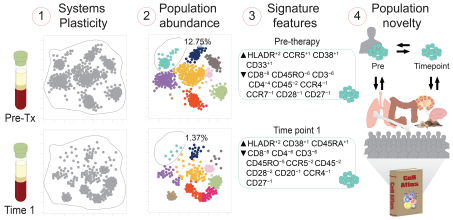
   Machine learning for patient medical phenotypes including t-SNE, FlowSOM, and MEM. Workflow inspired RAPID algorithm shown here for solid tumor cells.



FEBS J 2018

2. Mistry et al., Beyond the Message: Advantages of Snapshot Proteomics with Single-Cell Mass Cytometry in Solid Tumors. ***FEBS J*** 2018 Dec 13. PMID: 30549207 DOI: 10.1111/febs.14730.

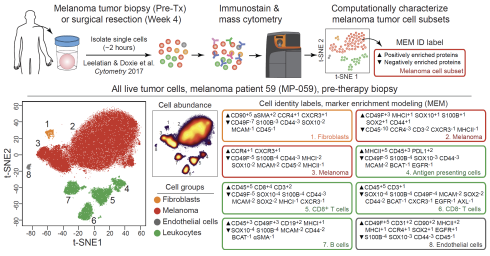
   Reviews single cell mass cytometry with an emphasis on solid tumor and tissue research. Provides a detailed comparison with complementary single cell technologies, including single cell RNA-sequencing and quantitative multiplex imaging.



Cancer Immunology Research 2019

3. Greenplate et al., Computational immune monitoring reveals abnormal double negative T cells present across human tumor types. ***Cancer Immunology Research*** 2019 Jan;7(1):86-99. PMC6318034.

   New computational workflow for longitudinal single cell tumor immunology. Revealed abnormal immune cells present in multiple tumor types. Will be applied to solid tumors to track changes in immune microenvironment.



Pigment Cell & Melanoma Research 2018

4. Doxie et al., BRAF and MEK inhibitor therapy eliminates Nestin-expressing melanoma cells in human tumors. ***Pigment Cell Melanoma Research***. 2018 Jun 28. PubMed PMID: 29778085; PubMed Central PMCID: PMC6188784.

   Compared biopsies over time of melanoma patients on treatment with mass cytometry to reveal *in vivo* cellular changes on treatment following targeted therapy.

cGVHD Arises from Immune Dysregulation, Is Clinically Heterogeneous, and Is Scored "Multidimensionally"

Areas of Potential Organ Involvement ("Dimensions"):

1. Liver
2. Gastrointestinal (GI)
3. Eye
4. Mouth
5. Fascia
6. Joint
7. Skin sclerosis
8. Skin redness
9. Lung

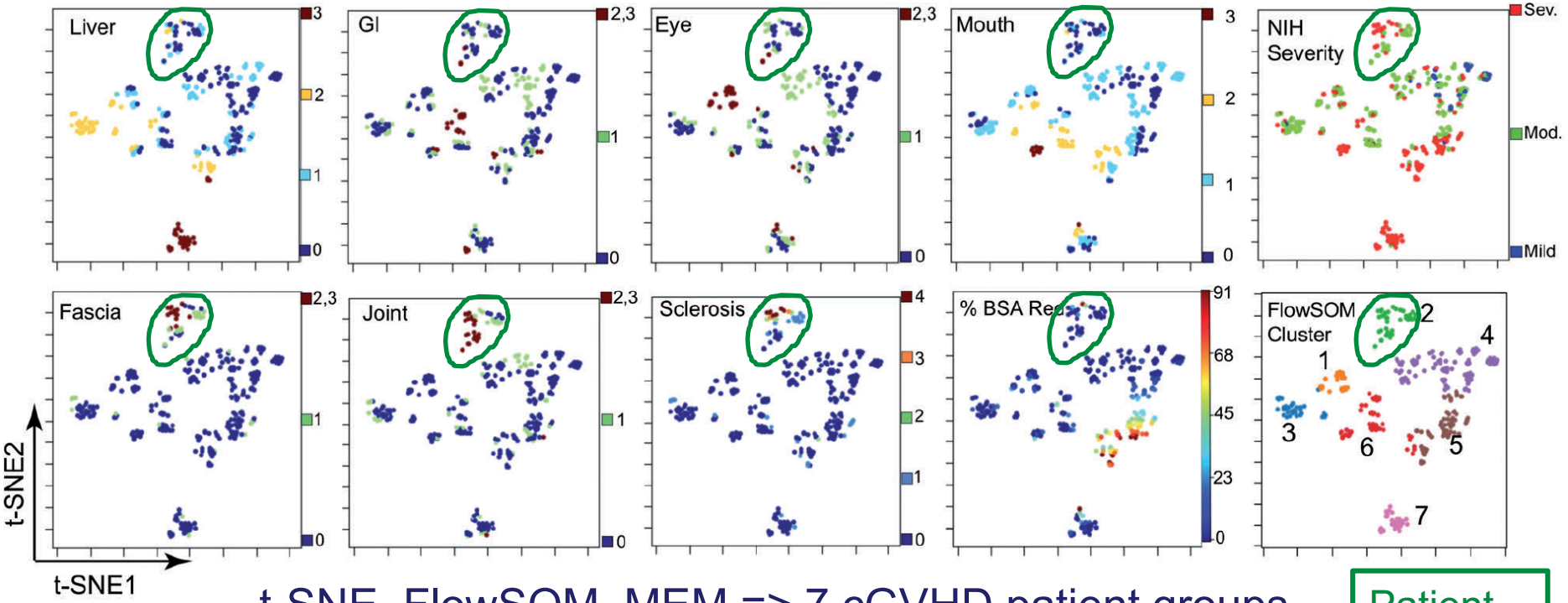
Each is scored (e.g. 0 to 3)

Sum of scores = NIH severity

Gandelman et al., *Haematologica* 2018

Data Science Strategies from Single Cell Analysis Reveal cGVHD Patient Groups Using Medical Phenotypes

Dots = 339 cGVHD patients
t-SNE = 8 dimensional medical phenotype
FlowSOM = identifies 7 cGVHD patient groups



t-SNE, FlowSOM, MEM => 7 cGVHD patient groups

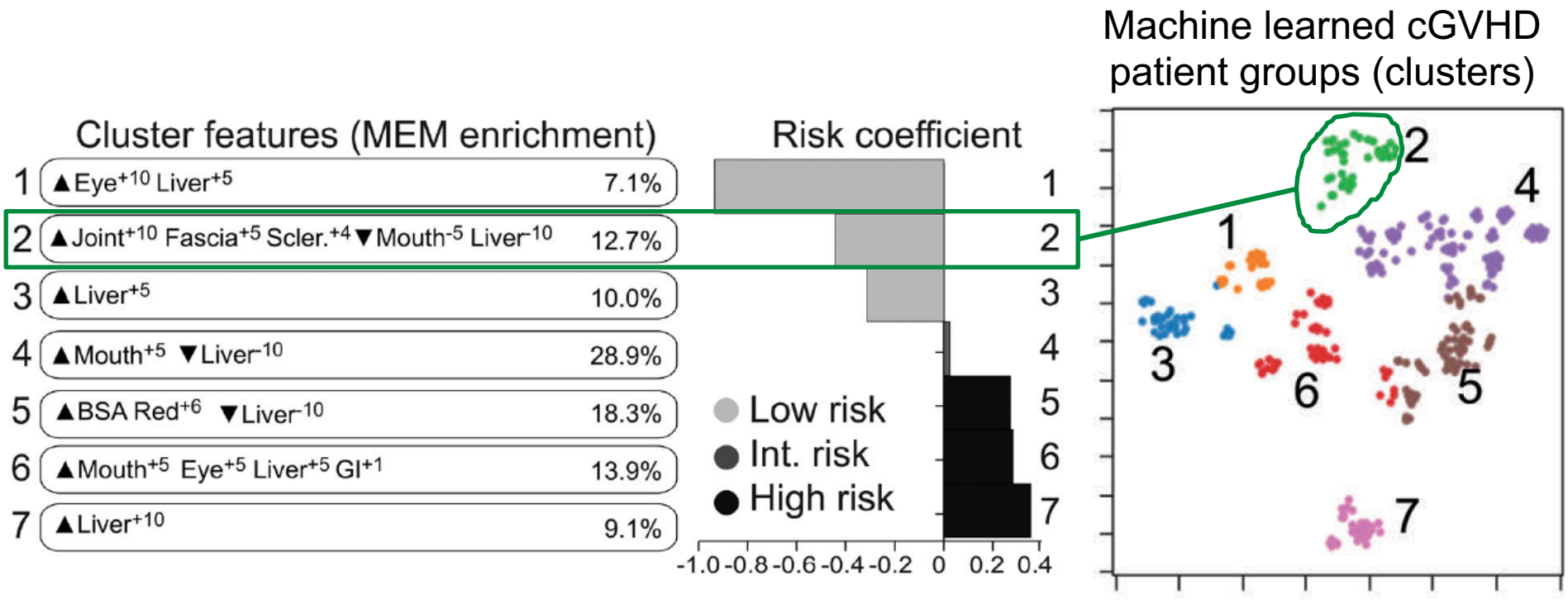
Patient Cluster 2

Chronic graft-*versus*-host disease (cGVHD)
t-SNE on 8 organ domain scores (e.g., liver involvement from 0 to 3)

Gandelman et al., *Haematologica* 2018

Data Science Strategies from Single Cell Analysis Reveal cGVHD Patient Groups Using Medical Phenotypes

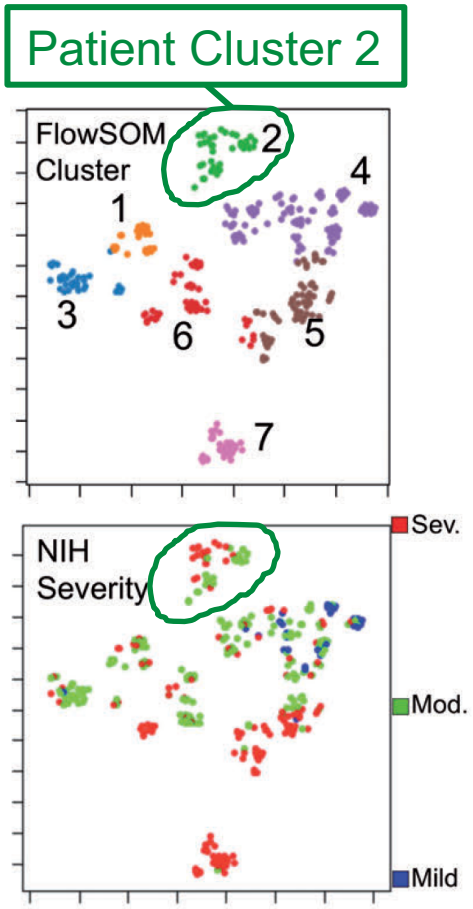
Dots = 339 cGVHD patients
t-SNE = 8 dimensional medical phenotype
FlowSOM = identifies 7 cGVHD patient groups

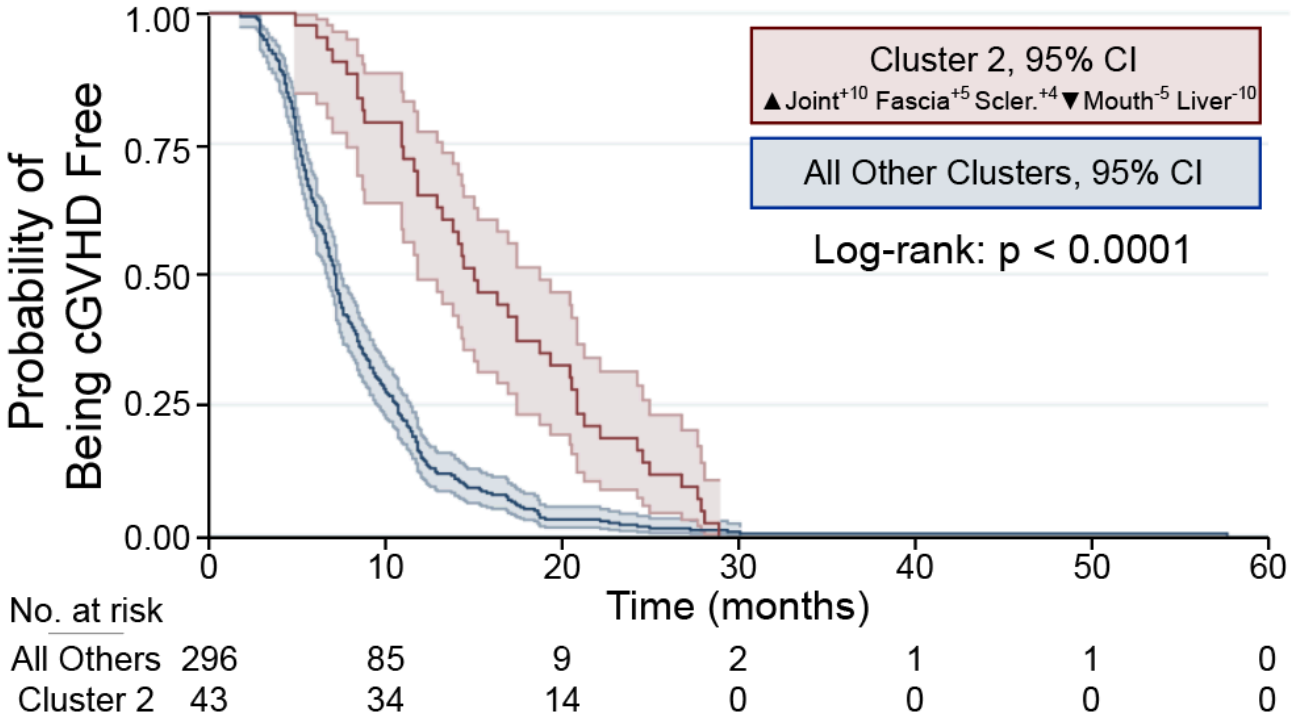


Chronic graft-*versus*-host disease (cGVHD)
t-SNE on 8 organ domain scores (e.g., liver involvement from 0 to 3)

Gandelman et al., *Haematologica* 2018

Cluster 2 Was a Distinct Subtype of cGVHD Patients with Longer Time from Stem Cell Transplant to cGVHD





Phenotype of cGVHD Patient Cluster 2 ("MEM label": +10 / -10)

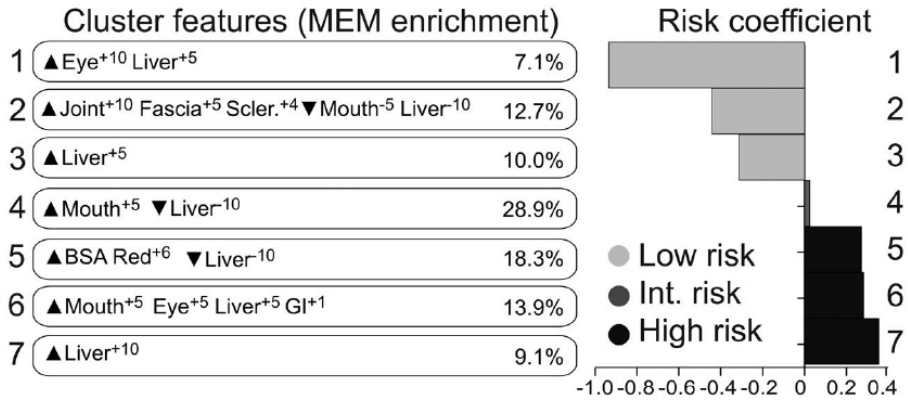
2 ▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$ ▼Mouth^{-5} Liver^{-10} 12.7%

cGVHD patient data: http://flowrepository.org/id/FR-FCM-ZYSU

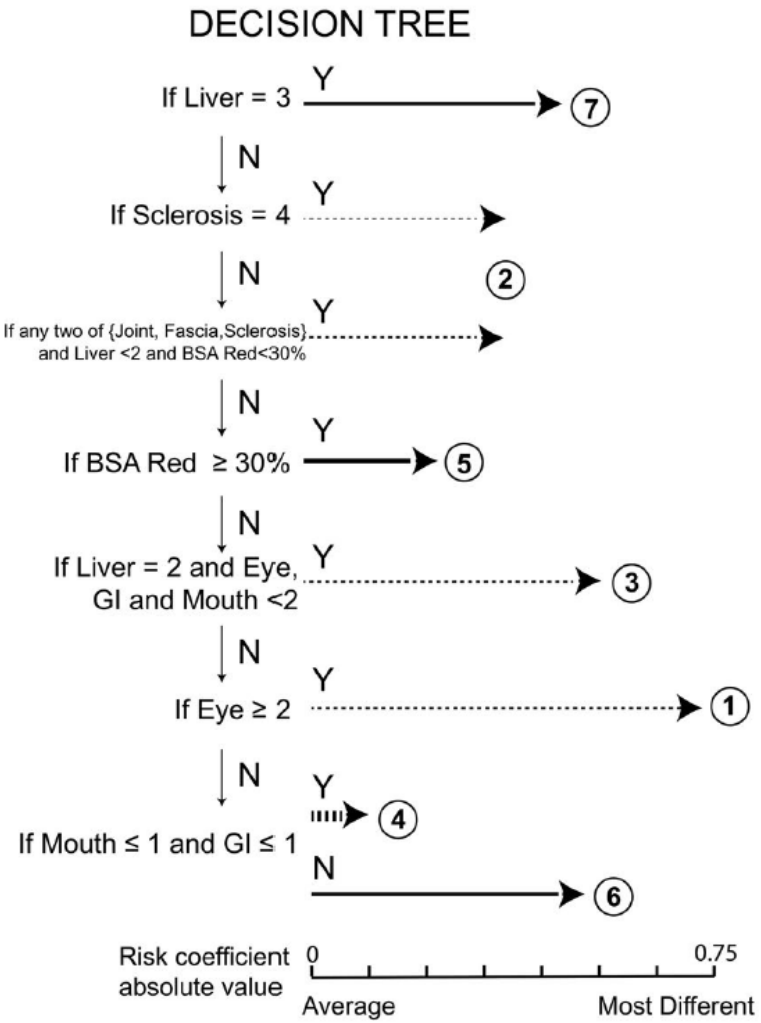
Gandelman et al., *Haematologica* 2018

Once Groups Are Revealed by Machine Learning, Clinically Accessible Strategies May Detect Them

7 Machine-Identified cGVHD patient groups



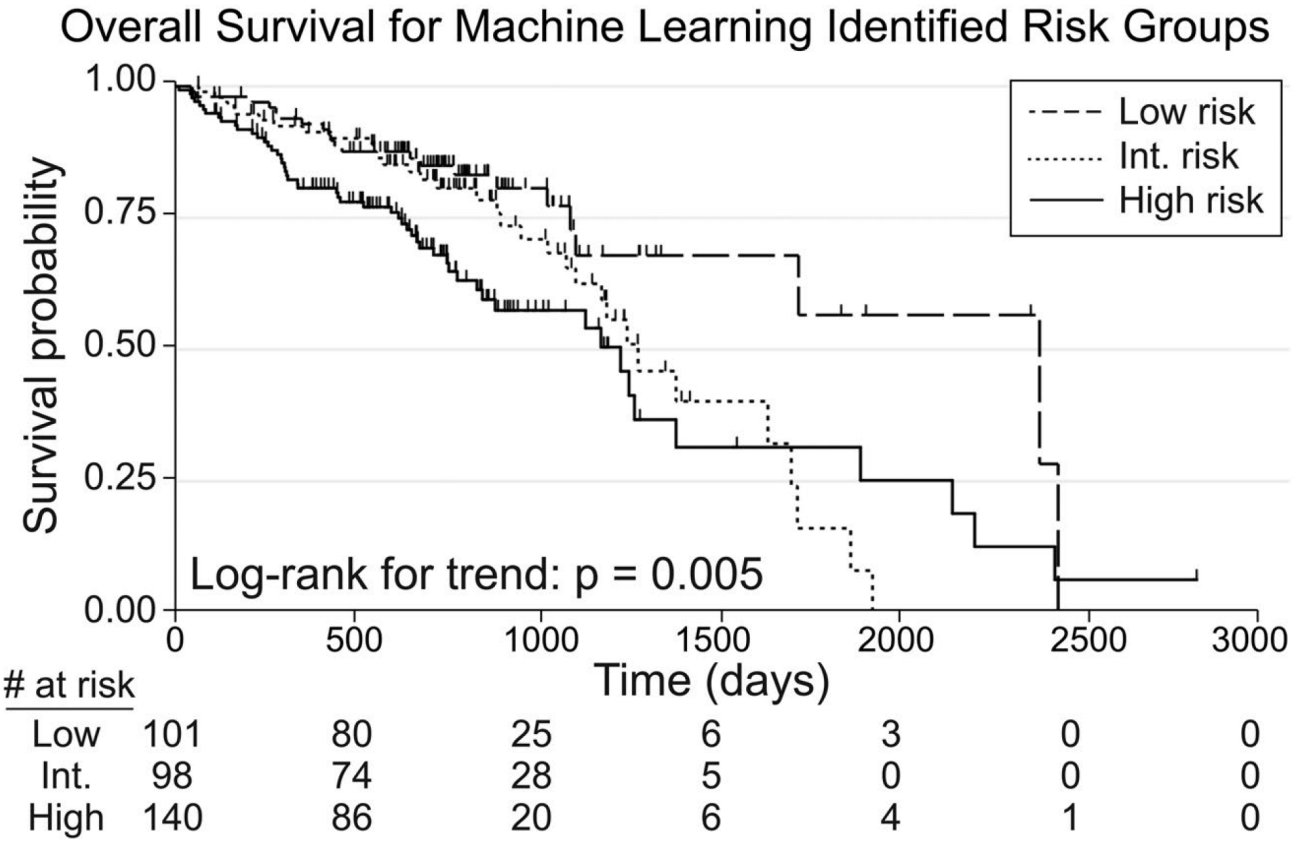
7 Comparable Patient Groups Identified by Simple Yes/No Question Series



cGVHD patient data: http://flowrepository.org/id/FR-FCM-ZYSU

Gandelman et al., *Haematologica* 2018

Decision Tree cGVHD Patient Groups Are Also Risk Stratified for Overall Survival



Low risk: HR = 2.79,
95% CI: 1.58-4.91; p<0.001

High risk: HR = 2.65,
95% CI: 1.42-4.94; p<0.0001

Decision tree cGVHD patient groups (clusters)

	Risk	Freq.	
1	-0.76	8.6%	Low
2	-0.34	10.0%	
3	-0.53	8.6%	
4	-0.06	33.6%	Int.
5	+0.21	17.4%	High
6	+0.49	11.8%	
7	+0.46	10.0%	

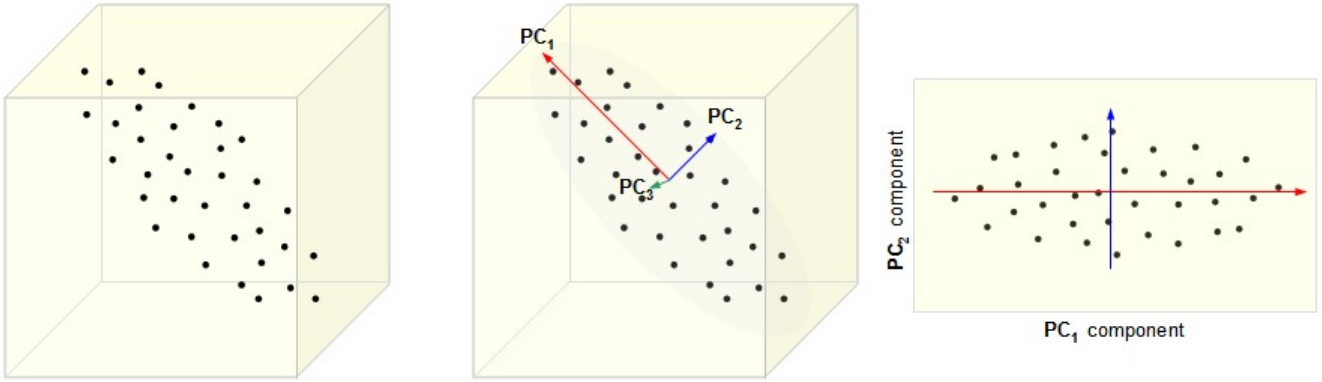
Chronic graft-*versus*-host disease (cGVHD)
t-SNE on 8 organ domain scores (e.g., liver involvement from 0 to 3)

Gandelman et al., *Haematologica* 2018

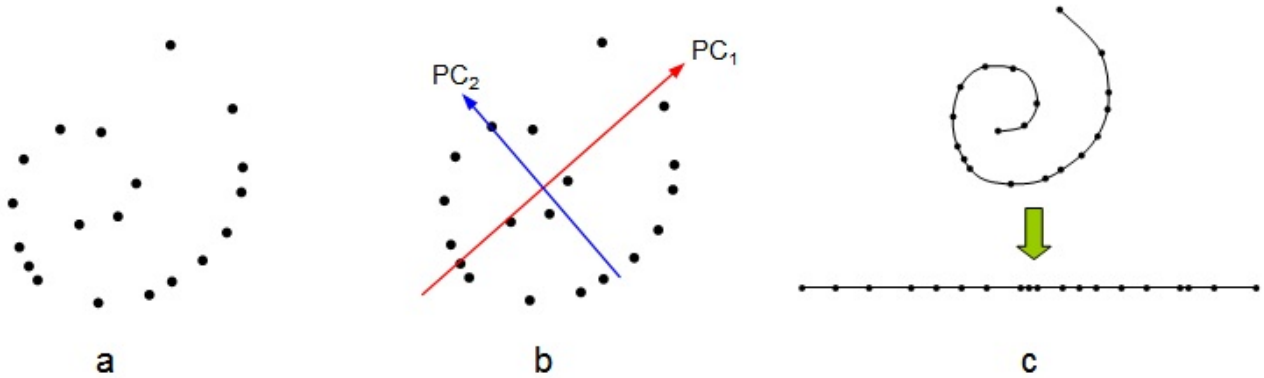
More Tools

PCA, Citrus, Trajectories

PCA is a Linear Dimensionality Reduction Tool

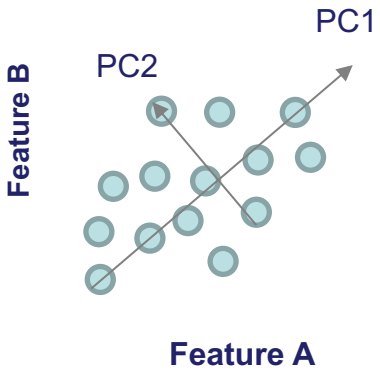


An illustration of PCA. **a)** A data set given as 3-dimensional points. **b)** The three orthogonal Principal Components (PCs) for the data, ordered by variance. **c)** The projection of the data set into the first two PCs, discarding the third one.



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.

Principal Component Analysis



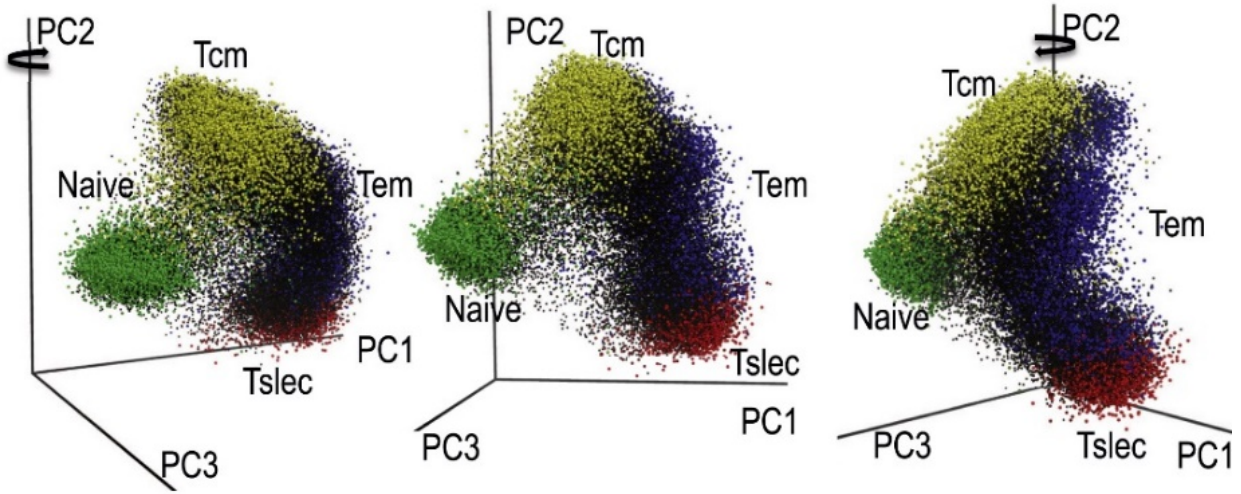
$PC1 = X^{A1}A + X^{B1}B + X^{C1}C$

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

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

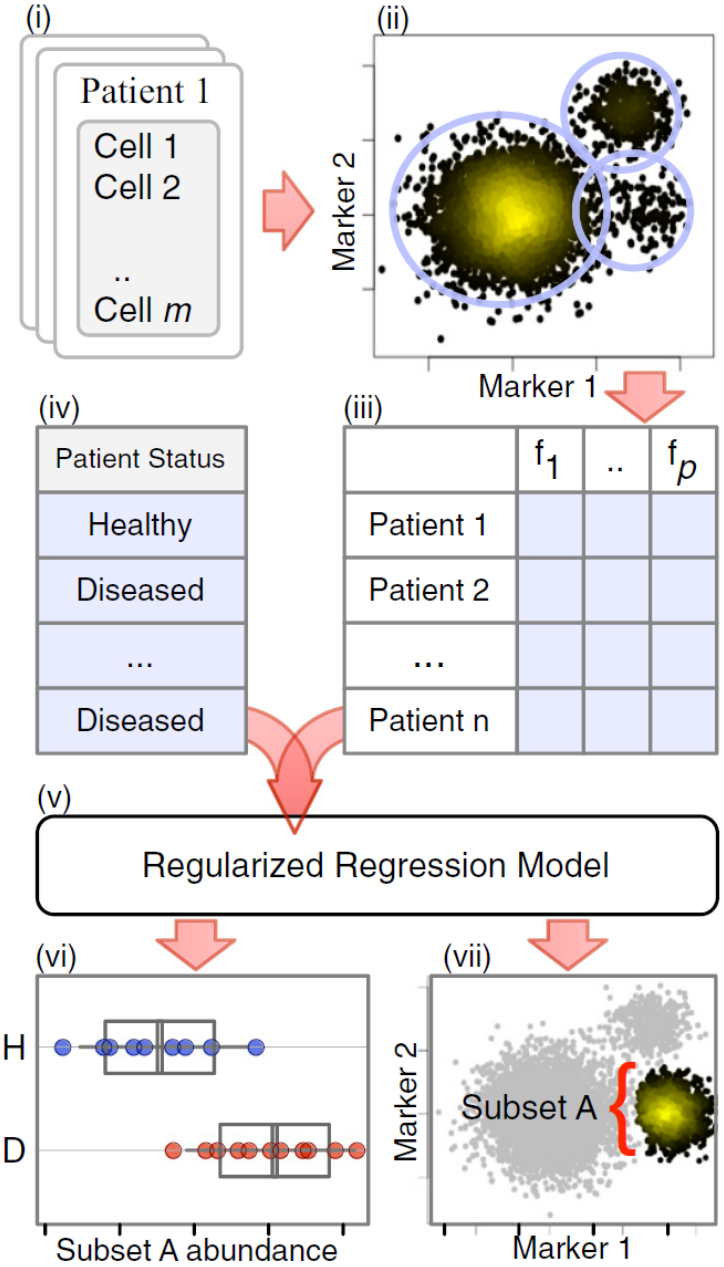
PCA used to Reduce Dimensionality of CyTOF Data

A 3D-PCA view of $CD8^+$ T cell 25 parameter data



Newell et al 2012, *Immunity*

Citrus: Supervised Population Finding

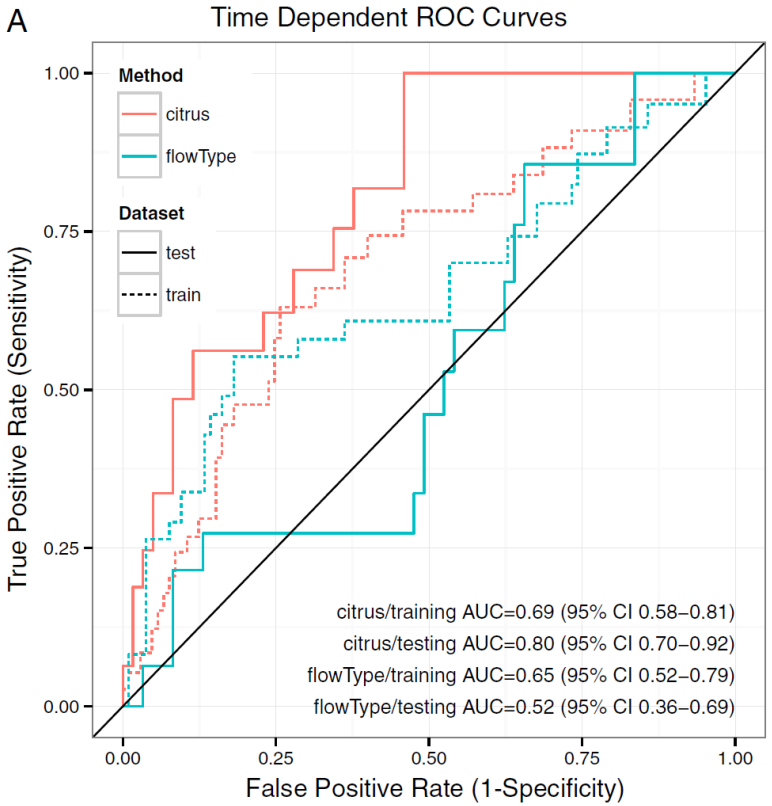


Automated identification of stratifying signatures in cellular subpopulations

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

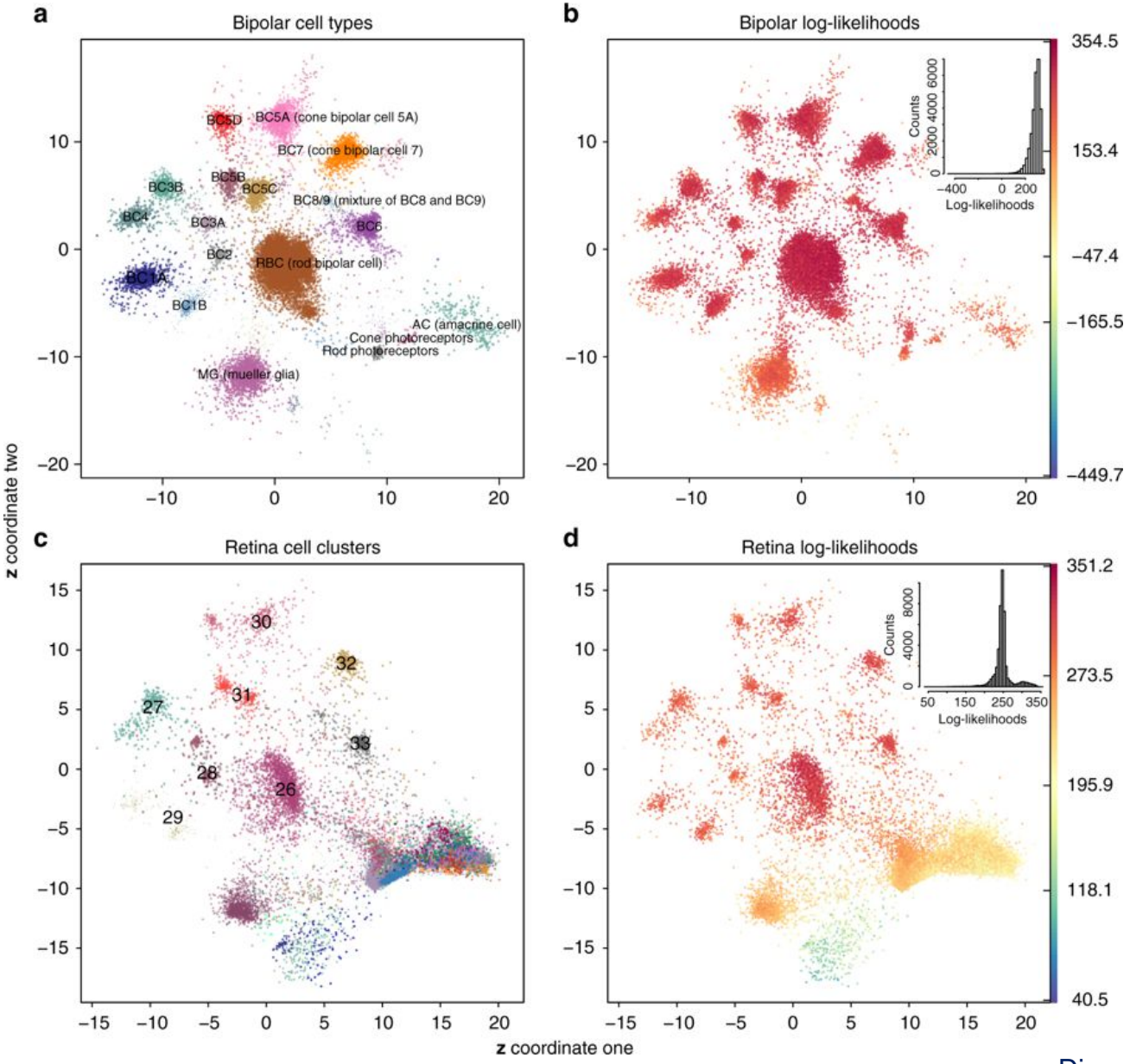
[a]Biomedical Informatics Training Program, Stanford University Medical School, Stanford, CA 94305; [b]Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, and Departments of [d]Computer Science, [e]Health Research and Policy, and [f]Statistics, Stanford University, Stanford, CA 94305; and [c]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)



Bruggner et al., *PNAS* 2014

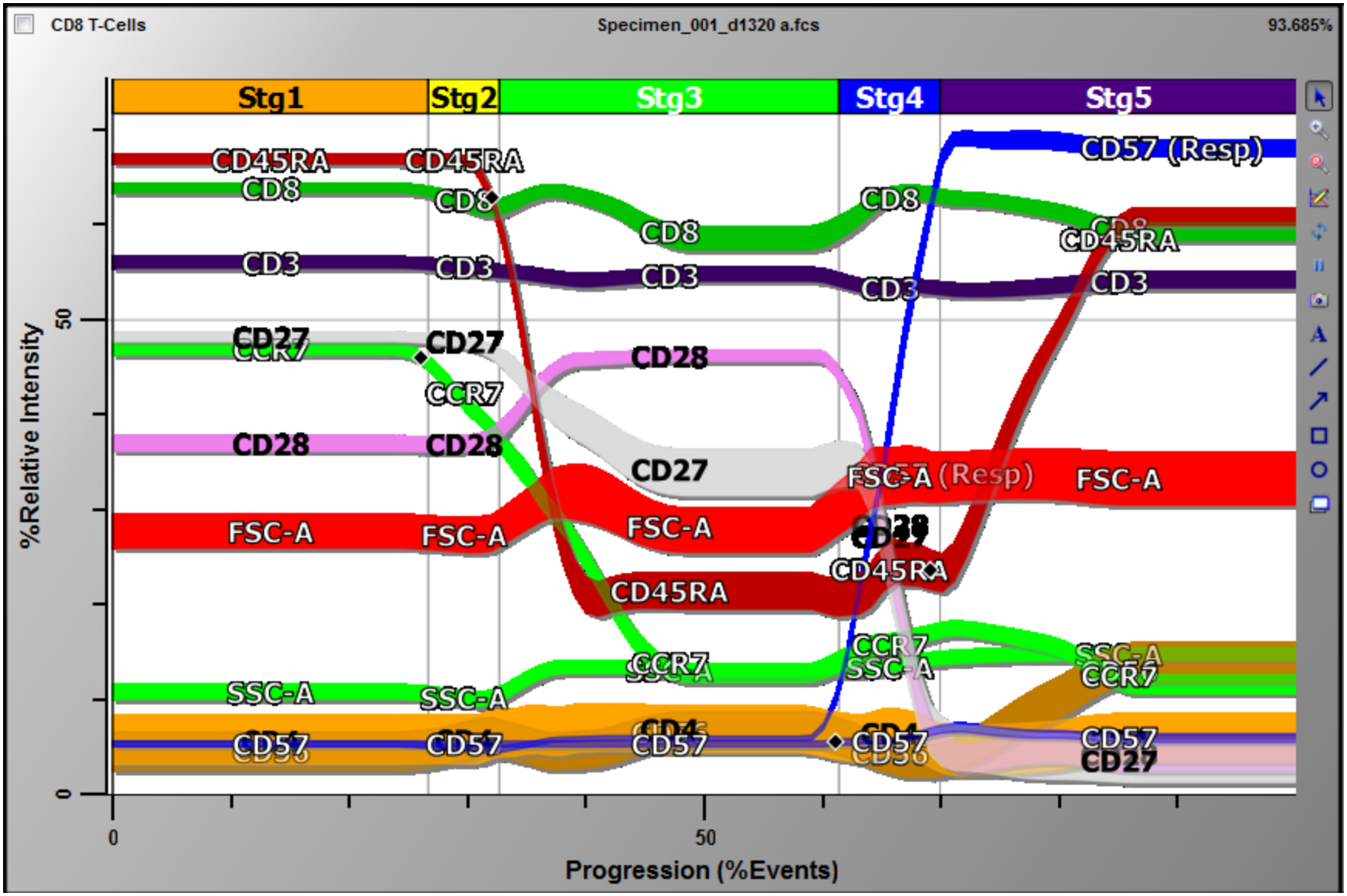
In 2018, Ding et al. Used scvis & *Probability* to Characterize / Identify Cells (scRNA-seq)



Learning a probabilistic mapping function from the bipolar data and applying the function to the independently generated mouse retina dataset. **a** scvis learned two-dimensional representations of the bipolar dataset, **b** coloring each point by the estimated log-likelihood, **c** the whole mouse retina dataset was directly projected to a two-dimensional space by the probabilistic mapping function learned from the bipolar data, and **d** coloring each point from the retina dataset by the estimated log-likelihood

Ding et al., *Nature Communications* 2018

Gemstone Uses Supervised Analysis to Identify Progressions

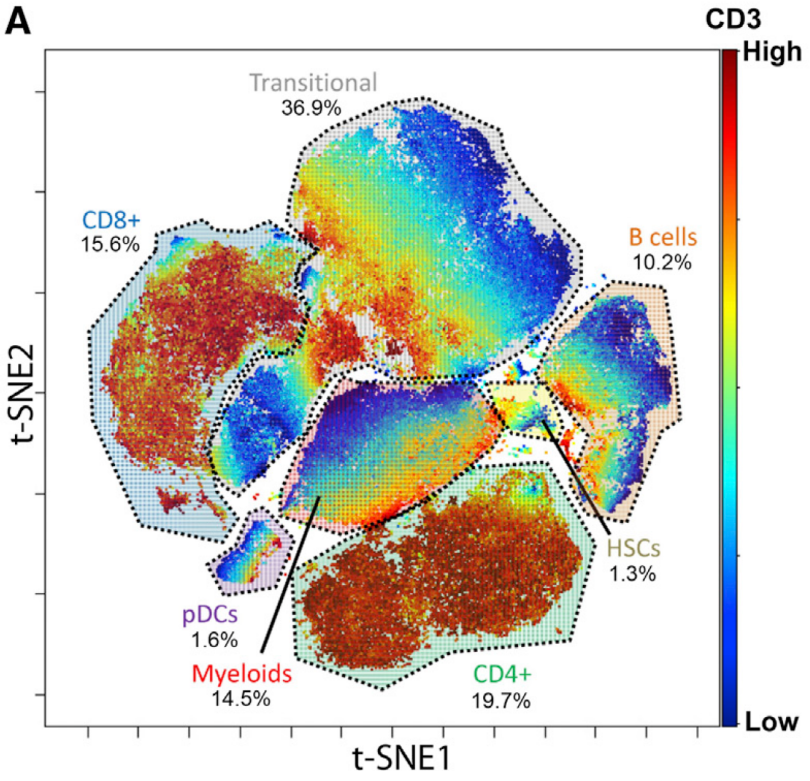


Verity Software House (www.vsh.com)

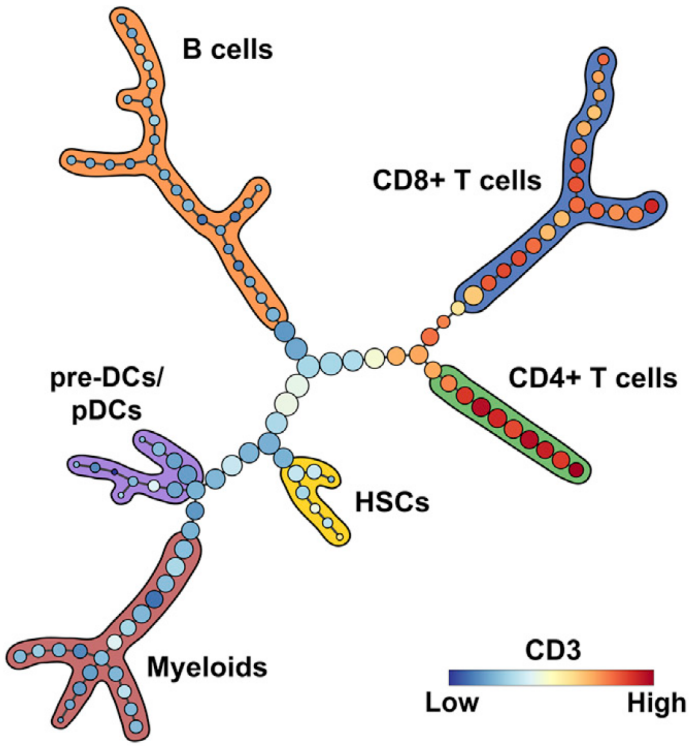
Inokuma, *JIM* 2013

In 2018, Ken Lau's Group Developed p-Creode to Infer Continua in Single Cell Data (e.g., human bone marrow, CyTOF)

t-SNE

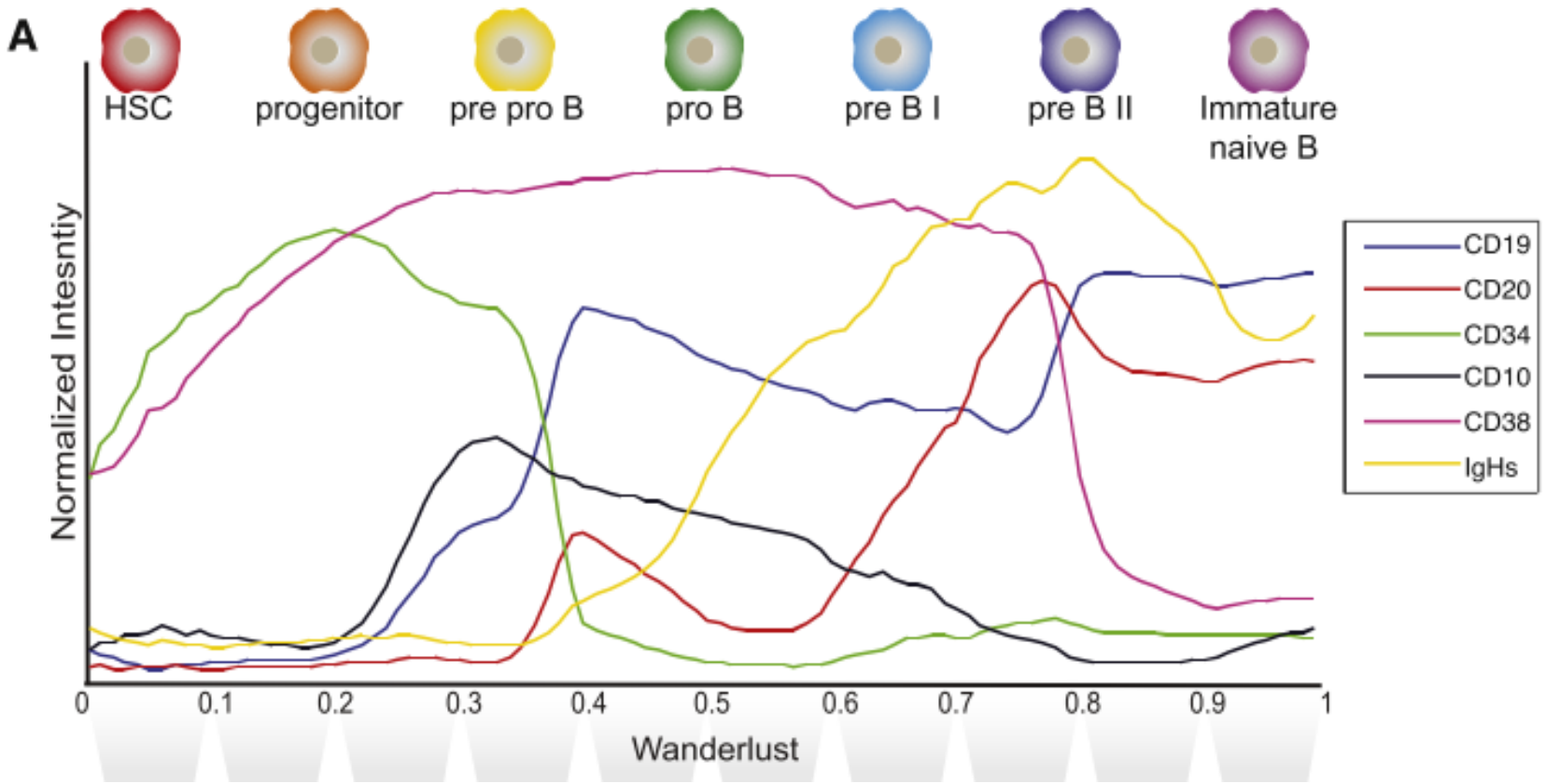


p-Creode



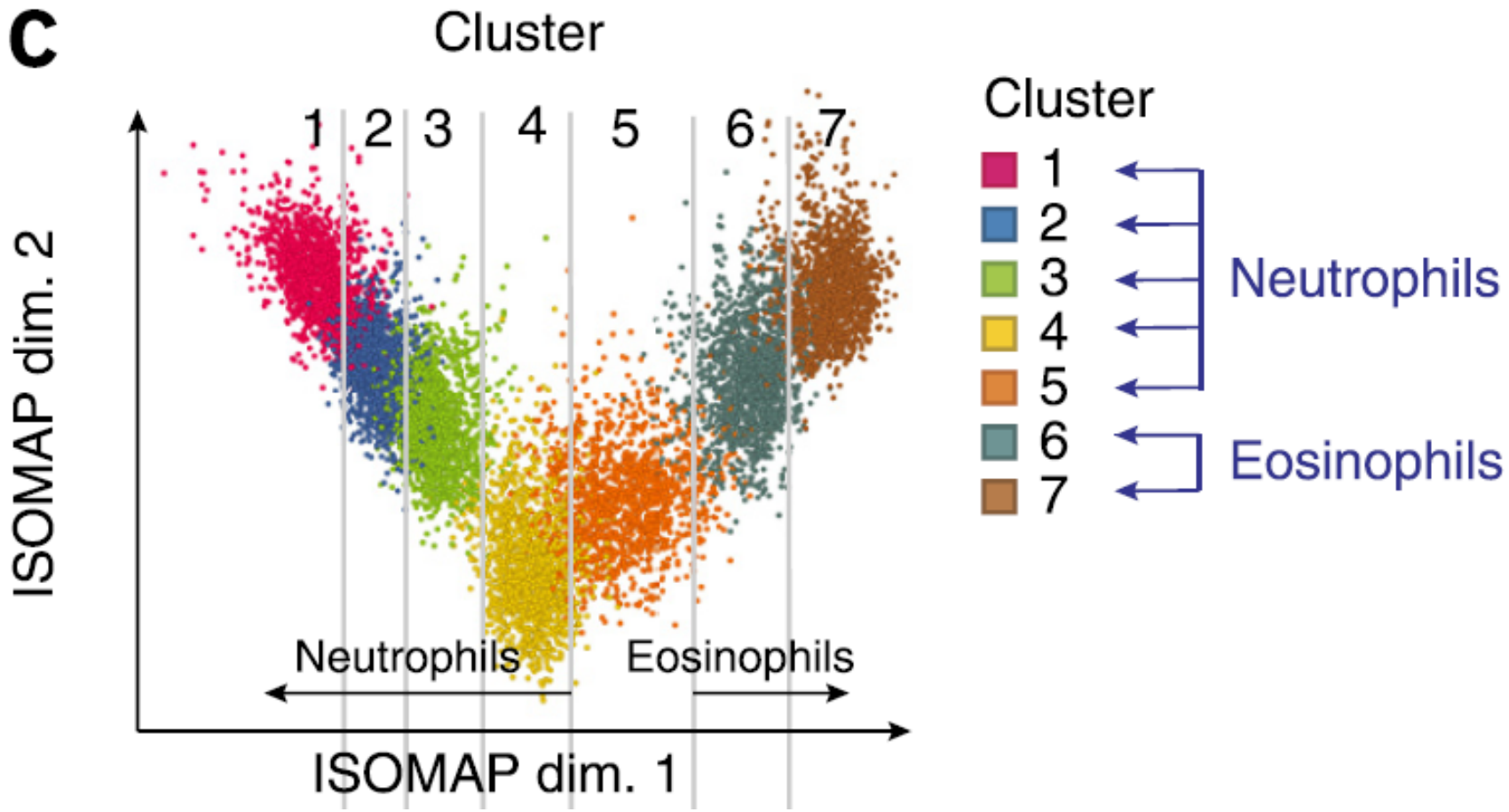
Herring et al., *Cell Systems* 2018

Wanderlust Identifies Phenotypic Progression



Bendall, *Cell* 2014

ISOMAP guided analysis

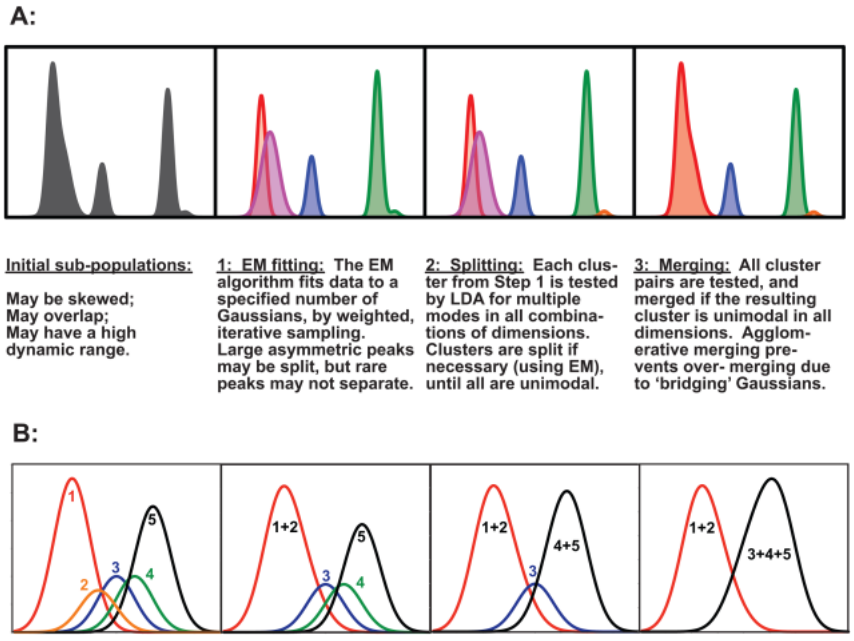


(**c**) ISOMAP dimensionality reduction to compare overall phenotypic relatedness of populations of neutrophil-like and eosinophil-like cells[31]. Top, cells color-coded by DensVM cluster number are plotted by their scores for ISOMAP dimensions 1 and 2. Binned median expression of defining markers (middle) 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.

Becher et al., *Nature Immunology* 2014

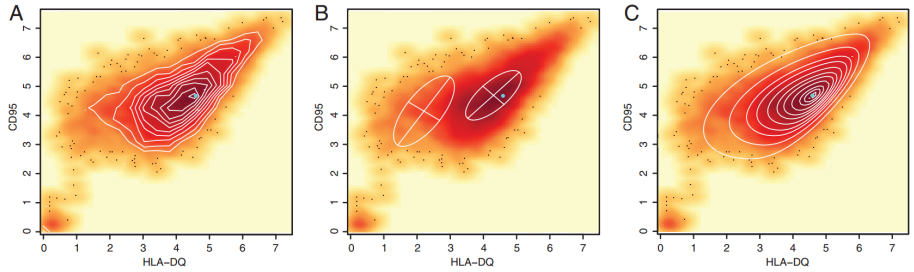
Mixture Modeling

SWIFT



Mosmann et al, 2014 *Cytometry A*

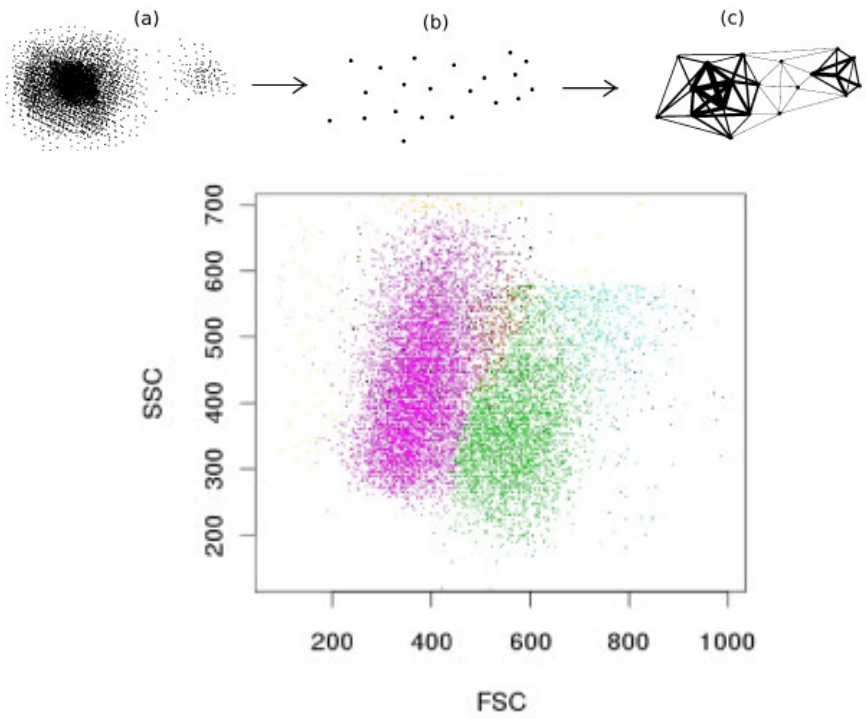
FLAME



Pyne et al, 2009 *PNA*

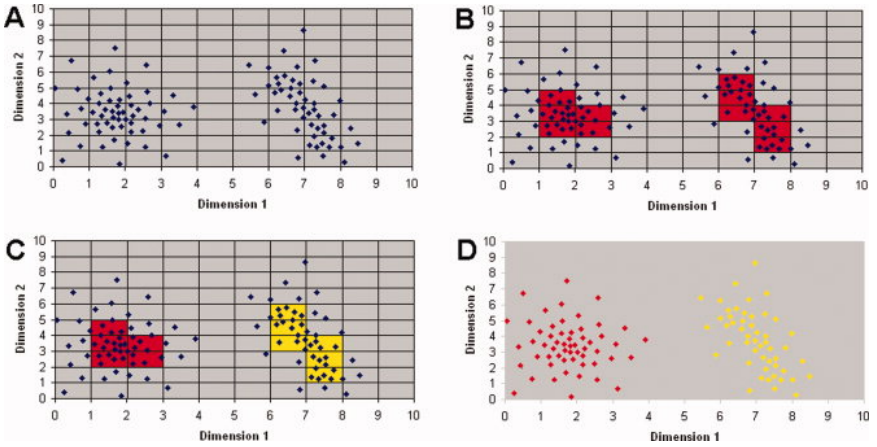
Automated Clustering and Population Identification Methods Based on Density

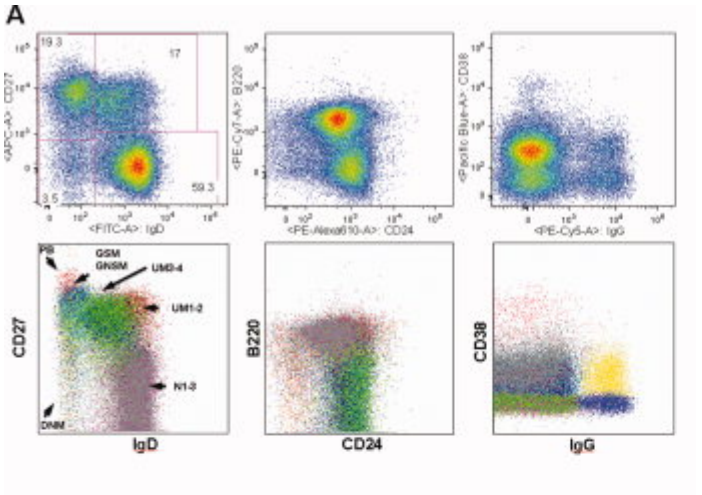
SamSpectral



Zare et al, 2010 *BMC Bioinformatics*

FLOCK

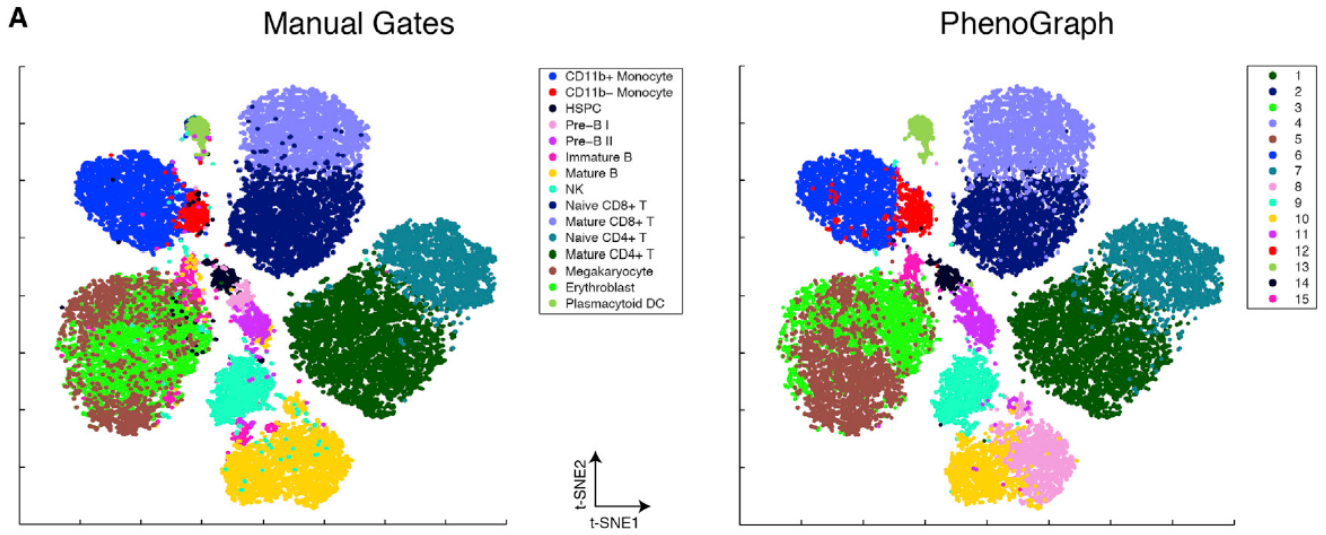


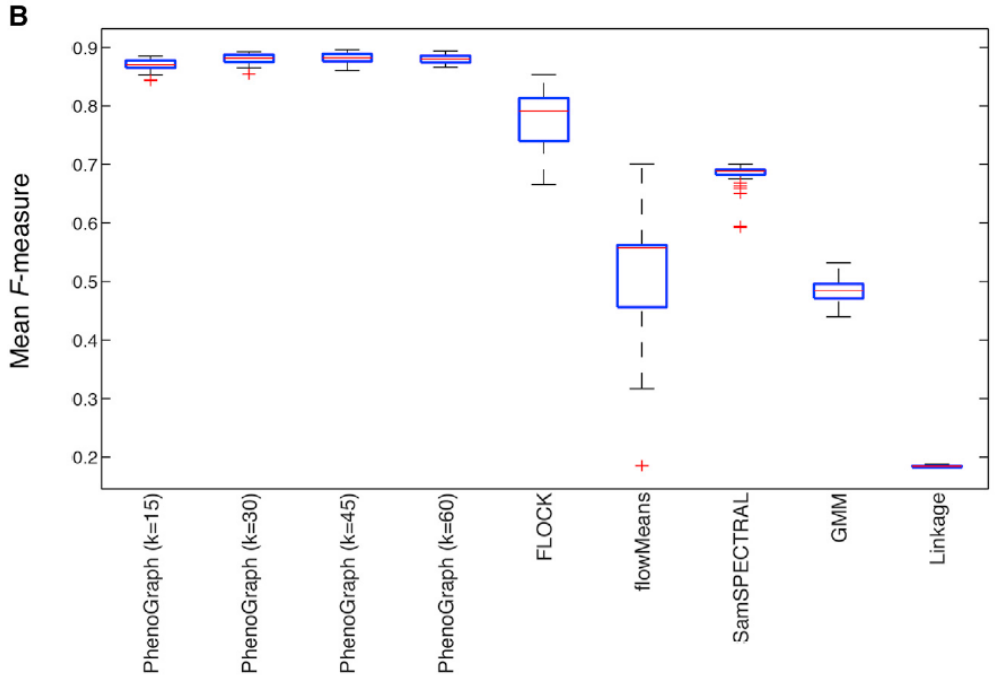


Qian et al, 2010 *Cytometry B Clin Cytom*

Phenograph

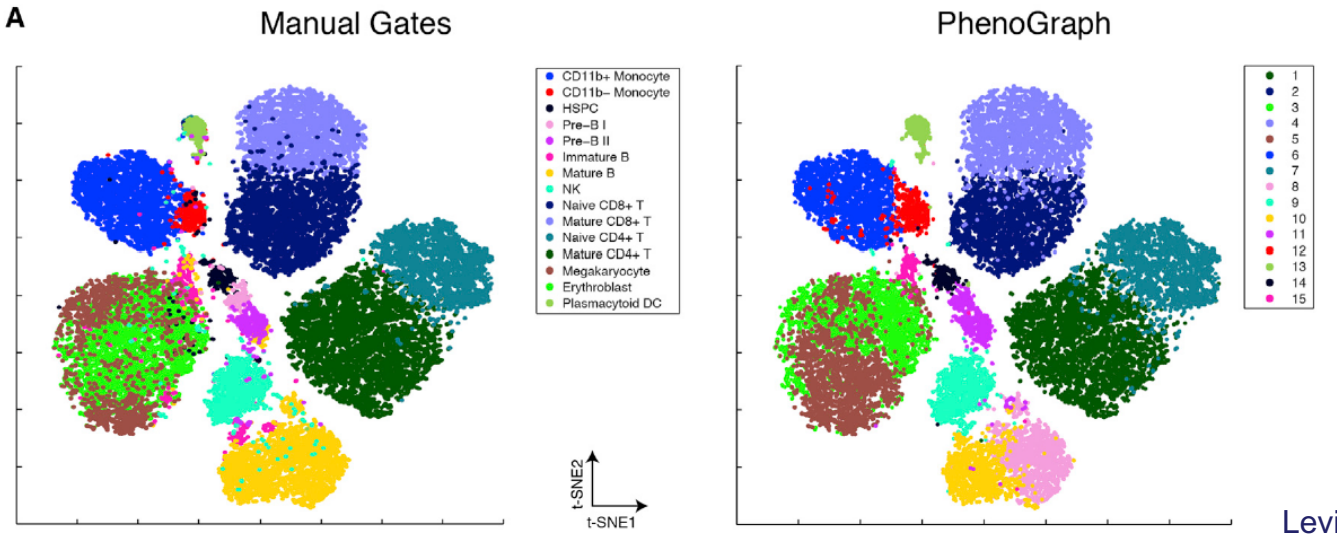
Phenograph Adds Fast Clustering & Meta-Analysis to viSNE





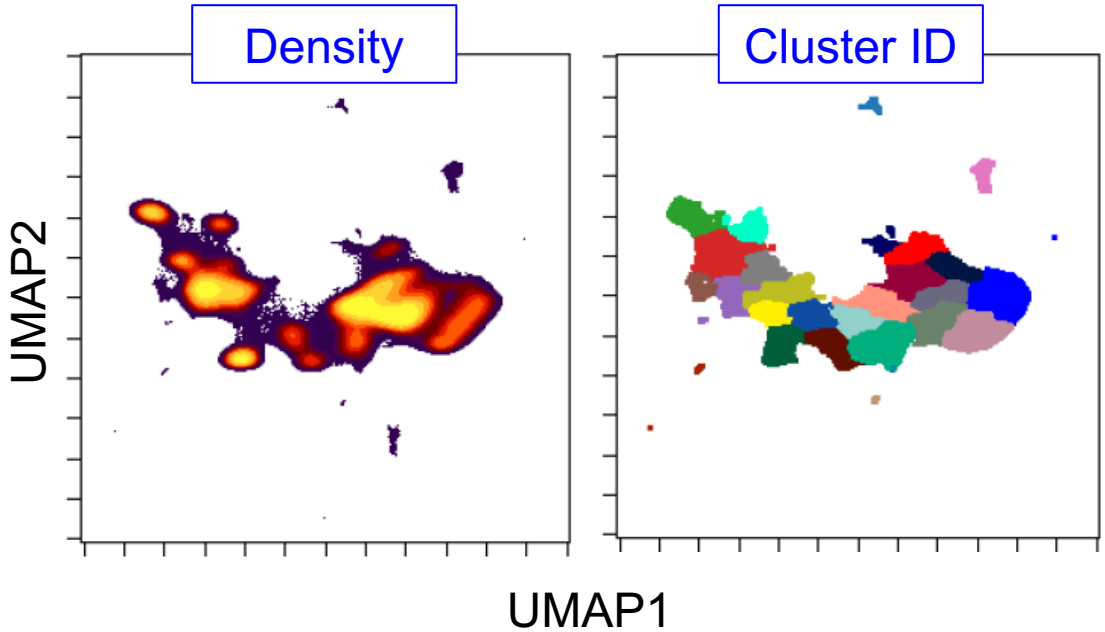
Levine et al., *Cell* 2015

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



Levine et al., *Cell* 2015

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



Diggins et al., *Methods* 2015

UMAP

ANALYSIS

nature biotechnology

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht[1], Leland McInnes[2], John Healy[2], Charles-Antoine Dutertre[1], Immanuel W H Kwok[1], Lai Guan Ng[1], Florent Ginhoux[1] & Evan W Newell[1,3]

Advances in single-cell technologies have enabled high-resolution dissection of tissue composition. Several tools for dimensionality reduction are available to analyze the large number of parameters generated in single-cell studies. Recently, a nonlinear dimensionality-reduction technique, uniform manifold approximation and projection (UMAP), was developed for the analysis of any type of high-dimensional data. Here we apply it to biological data, using three well-characterized mass cytometry and single-cell RNA sequencing datasets. Comparing the performance of UMAP with five other tools, we find that UMAP provides the fastest run times, highest reproducibility and the most meaningful organization of cell clusters. The work highlights the use of UMAP for improved visualization and interpretation of single-cell data.

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com

John Healy
Tutte Institute for Mathematics and Computing
jchealy@gmail.com

James Melville
jlmelville@gmail.com

December 7, 2018

https://arxiv.org/abs/1802.03426

Abstract

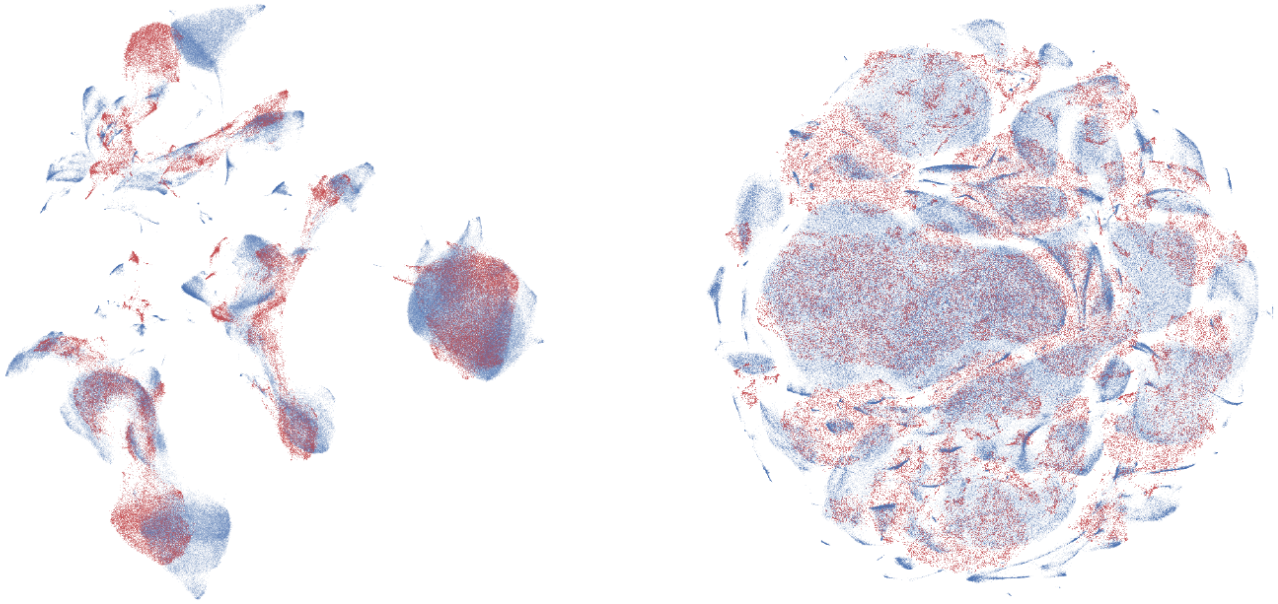
UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that applies to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning.

Comments: Reference implementation available at this http URL
Subjects: **Machine Learning (stat.ML)**; Computational Geometry (cs.CG); Machine Learning (cs.LG)
Cite as: **arXiv:1802.03426 [stat.ML]**
(or **arXiv:1802.03426v2 [stat.ML]** for this version)

Submission history

From: Leland McInnes [view email]
[v1] Fri, 9 Feb 2018 19:39:33 UTC (958 KB)
[v2] Thu, 6 Dec 2018 18:54:07 UTC (7,966 KB)

Flow Cytometry Data That Looks Like a Blob on a t-SNE Plot Appears to Have Structure on a UMAP Plot



(a) UMAP (b) t-SNE

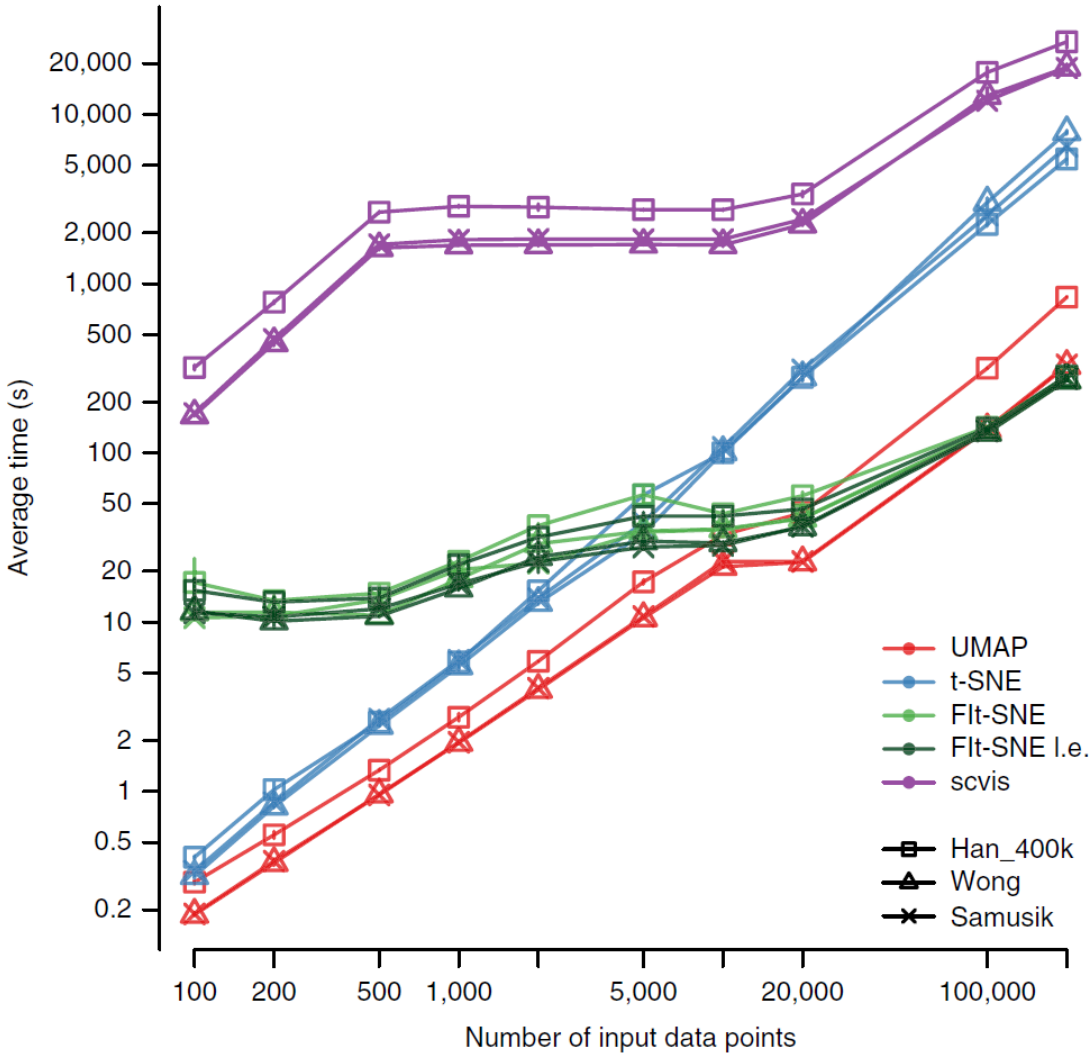
Figure 3: Procrustes based alignment of a 10% subsample (red) against the full dataset (blue) for the flow cytometry dataset for both UMAP and t-SNE.

In Greek mythology, **Procrustes** (Ancient Greek: Προκρούστης *Prokroustes*) or "the stretcher [who hammers out the metal]", also known as **Prokoptas** or **Damastes**(Δαμαστής, "subduer"), was a rogue smith and bandit from Attica who attacked people by stretching them or cutting off their legs, so as to force them to fit the size of an iron bed.

The word "Procrustean" is thus used to describe situations **where different lengths or sizes or properties are fitted to an arbitrary standard**.

McInnes et al., *Rxiv* 2018

UMAP & Fit-SNE are Much Faster than Traditional t-SNE



20,000 seconds = 5.5 hours

5,000 seconds = ~1.4 hours
(similar to our run times)

500 seconds = 8.3 minutes

CyTOF datasets are ~100,000 to 50,000,000 cell data points

[scRNA-seq is 1,000 to 100,000]

Figure 3 Run times of five dimensionality reduction methods for inputs of varying sizes. The average run time of three random subsamples is represented, with vertical bars representing s.d. after log-transforming the run times.

Becht et al., *Nature Biotechnology* 2018

Now, McInnes et al., UMAP Preserves Local and Global Structure

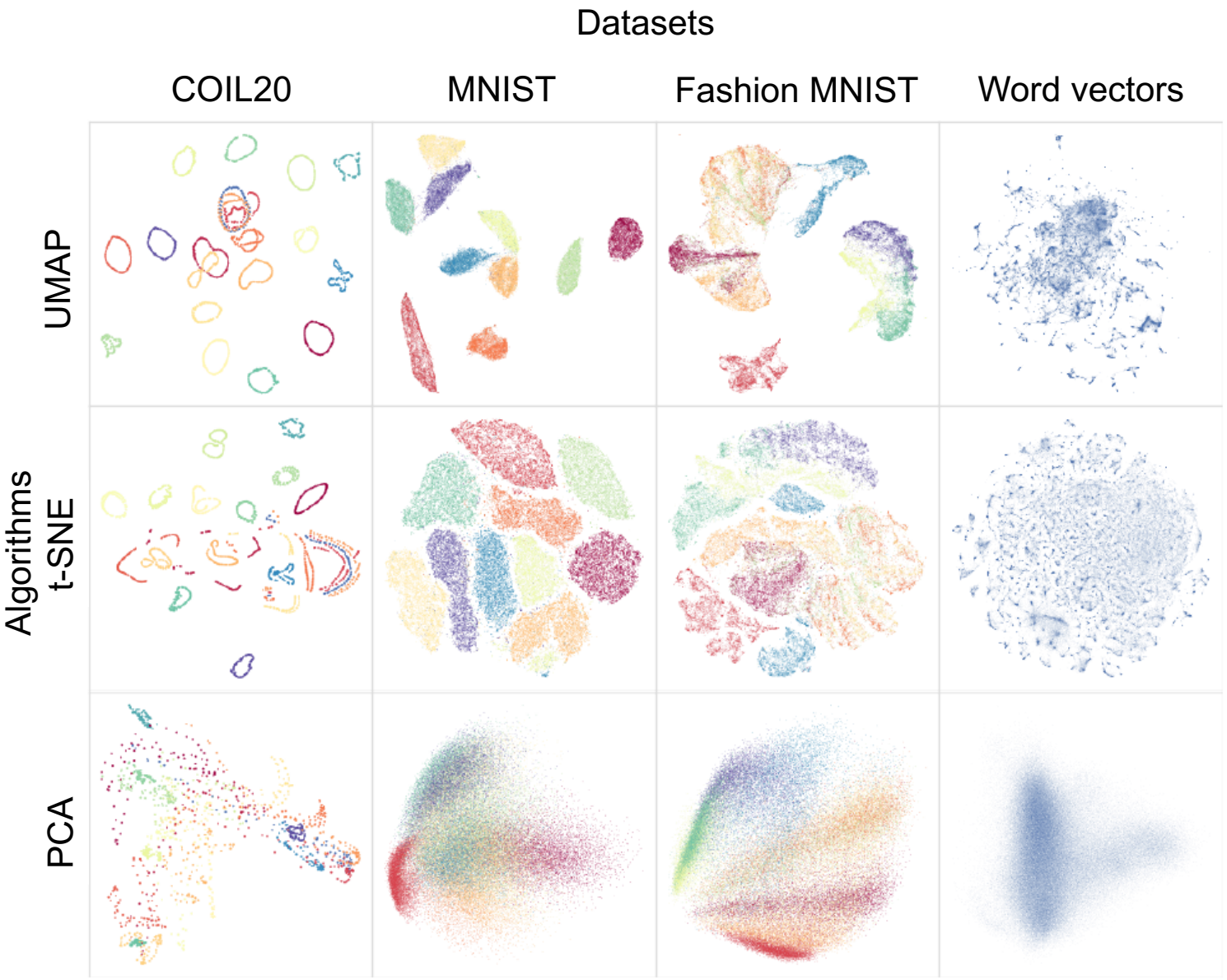


Figure 2: A comparison of several dimension reduction algorithms. We note that UMAP successfully reflects much of the large scale global structure that is well represented by Laplacian Eigenmaps and PCA (particularly for MNIST and Fashion-MNIST), while also preserving the local fine structure similar to t-SNE and LargeVis.

McInnes et al., *Rxiv* 2018

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, γδ T cells, and contaminating cells

(b) color-coding the tissues of origin; t-SNE separated cell populations according to their tissue of origin more often than UMAP. UMAP instead ordered events according to their origin within each major cluster, roughly from 1) cord blood and peripheral blood mononuclear cells, to liver and spleen, and to tonsils on the one end to skin, gut and lung on the other. **Continua not apparent in t-SNE.**

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 (the Wong dataset).

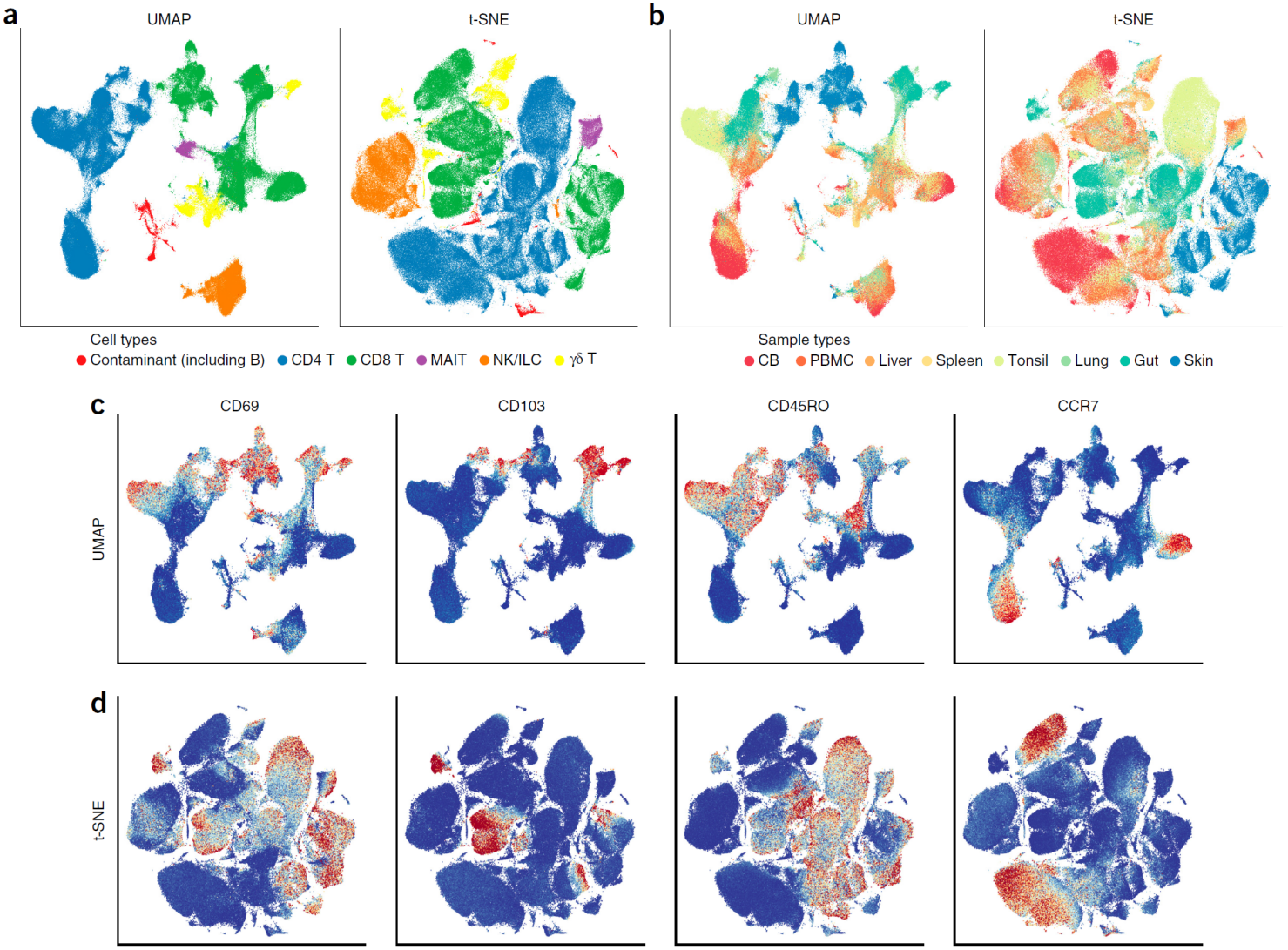


Figure 1 UMAP embeds local and large-scale structure of the data. UMAP and t-SNE projections of the Wong *et al.* dataset colored according to (**a**) broad cell lineages, (**b**) tissue of origin, and for (**c**) UMAP and (**d**) t-SNE, the expression of CD69, CD103, CD45RO and CCR7. For **c** and **d**, blue denotes minimal expression, beige intermediate and red high. MAIT, mucosal-associated invariant T cell; ILC, innate lymphoid cell; CB, cord blood; PBMC, peripheral blood mononuclear cell.

Becht et al., *Nature Biotechnology* 2018

Becht et al., UMAP Captures Developmental Trajectories

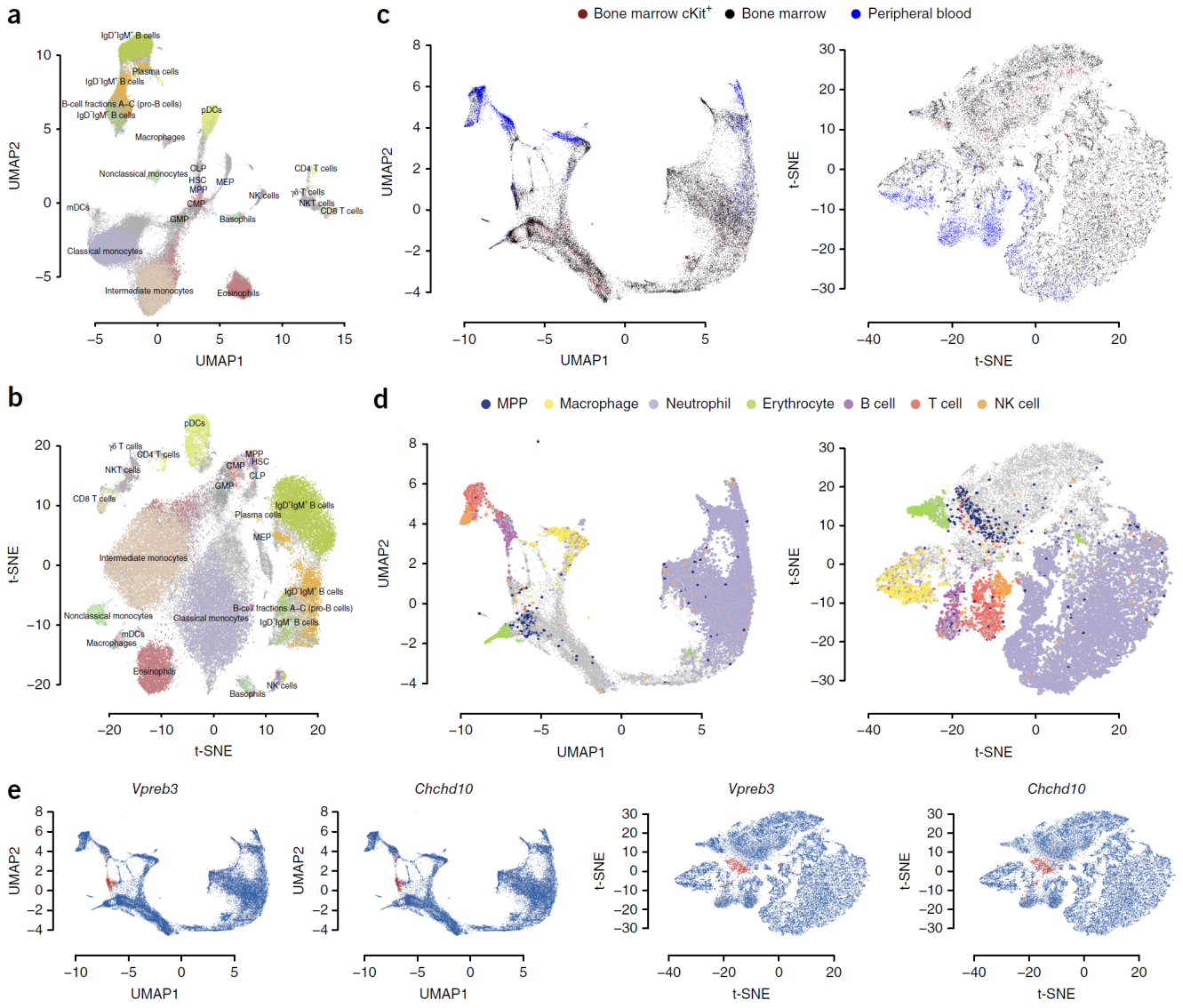


Figure 2 UMAP embeddings of bone marrow and blood samples recapitulate hematopoiesis. (**a**) UMAP and (**b**) t-SNE projection of the Samusik_01 dataset. Events are color-coded according to manual gates provided by the authors of the dataset. (**c**,**d**) UMAP and t-SNE projections of the Han dataset, color-coded by (**c**) tissue of origin or (**d**) cell populations. (**e**) Expression of the V-set pre-B cell surrogate light chain 3 (*Vpreb3*) and *Chchd10* genes on the UMAP and t-SNE projections of the Han dataset. Blue denotes minimal expression, beige intermediate and red high. pDC, plasmacytoid dendritic cell; mDC, myeloid dendritic cell; NKT, natural killer T.

Becht et al., *Nature Biotechnology* 2018

Speed is Nothing without Accuracy; UMAP is Also Accurate

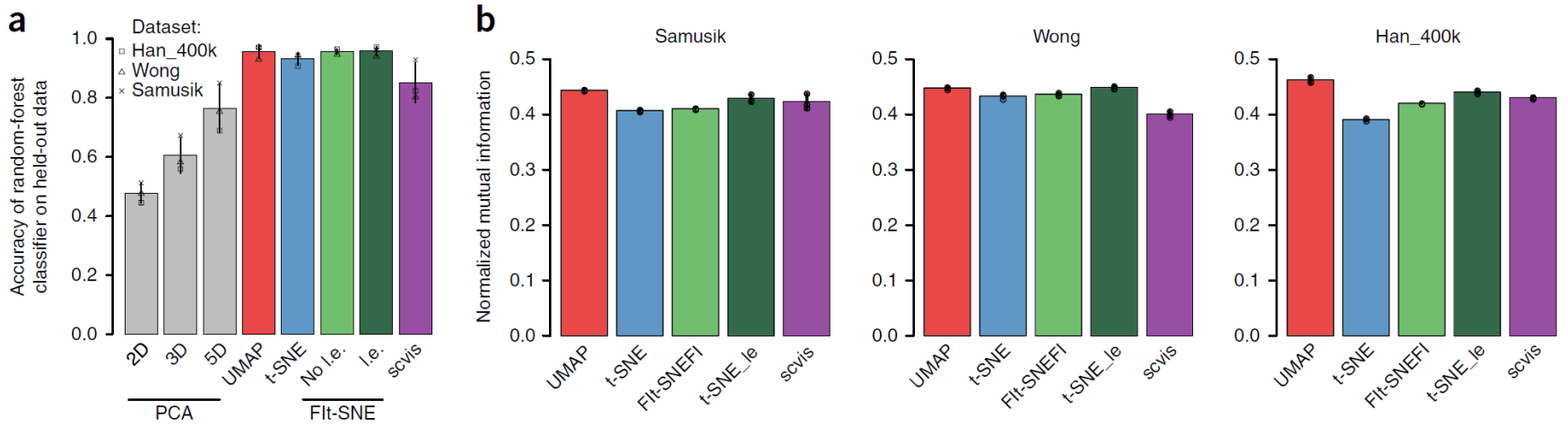
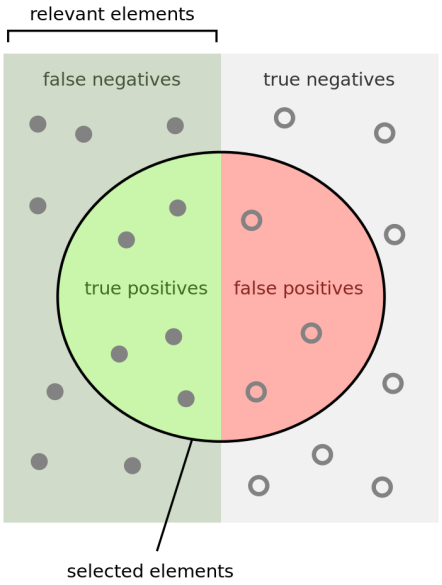


Figure 4 Analysis of local data structure in embeddings produced by each algorithm. (**a**) Accurate classification rate on held-out data of random-forest classifiers predicting Phenograph cluster labels using embedded coordinates as input. The average across the three datasets is shown, with vertical bars representing s.d. (**b**) Average normalized mutual information of *k*-means clustering ($k = 100$) performed on the embeddings of data subsamples and *k*-means clustering ($k = 100$) performed on total datasets. The average across the three random subsamples of size 200,000 is shown, with vertical bars representing s.d.



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

F_1 score = accuracy

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Becht et al., *Nature Biotechnology* 2018

UMAP Preserves “Large-Scale Structure” That t-SNE Ignores (Large = Position of Islands; Fine = Position of Cells in an Island)

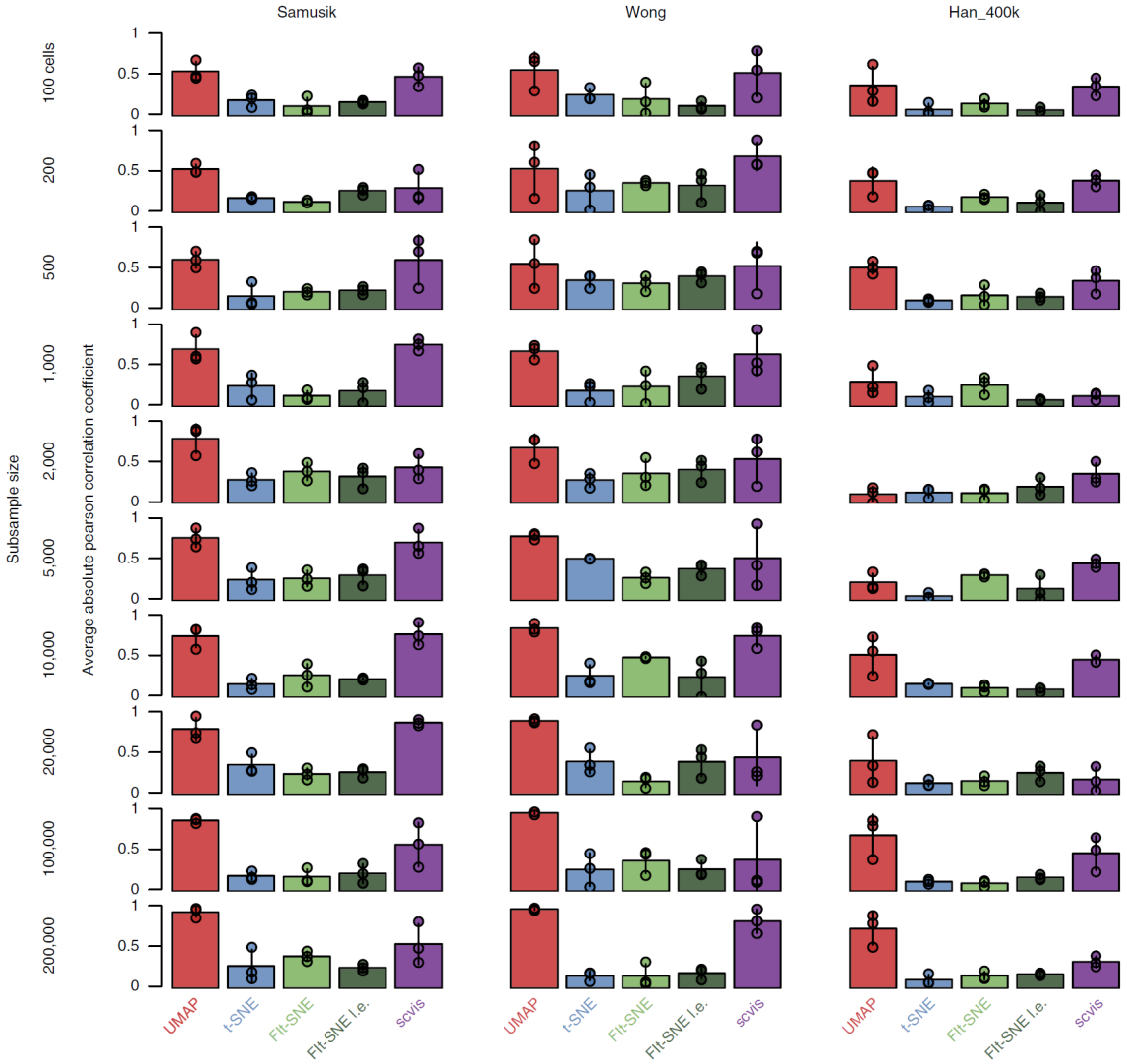
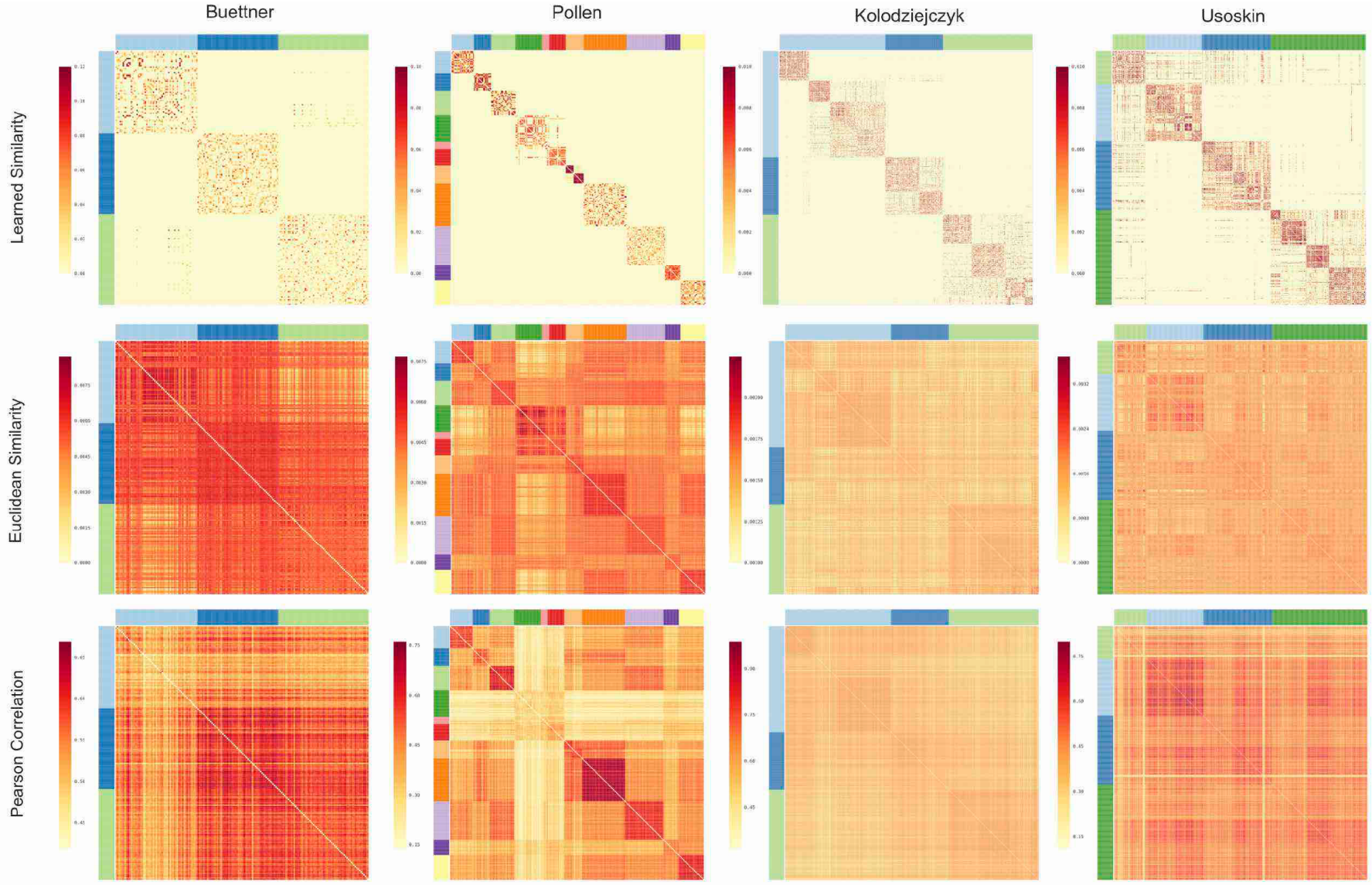


Figure 6 Reproducibility of large-scale structures in embeddings. Bar plots represent the average unsigned Pearson correlation coefficient of the points' coordinates in the embedding of subsamples versus in the embedding of the full dataset, thus measuring the correlation of coordinates in subsamples versus in the embedding of the full dataset, up to symmetries along the graph axes. Bar heights represent the average across three replicates and vertical bars the corresponding s.d.

Becht et al., *Nature Biotechnology* 2018

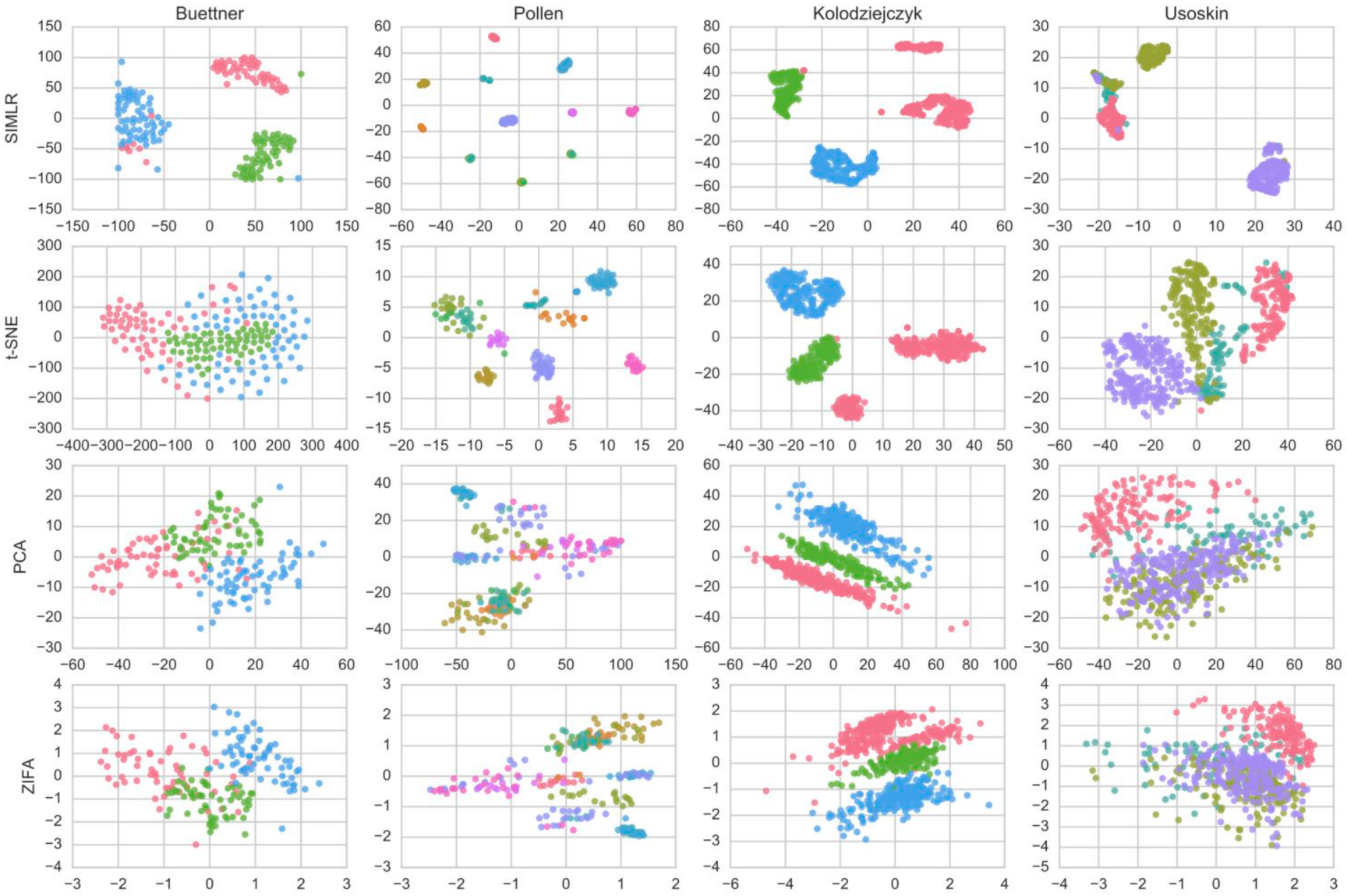
SIMLR

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



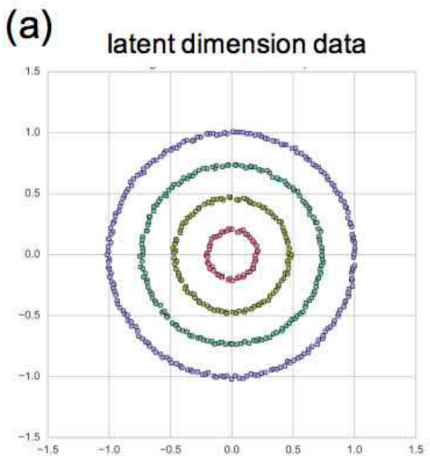
Wang et al., *Nature Methods* 2017

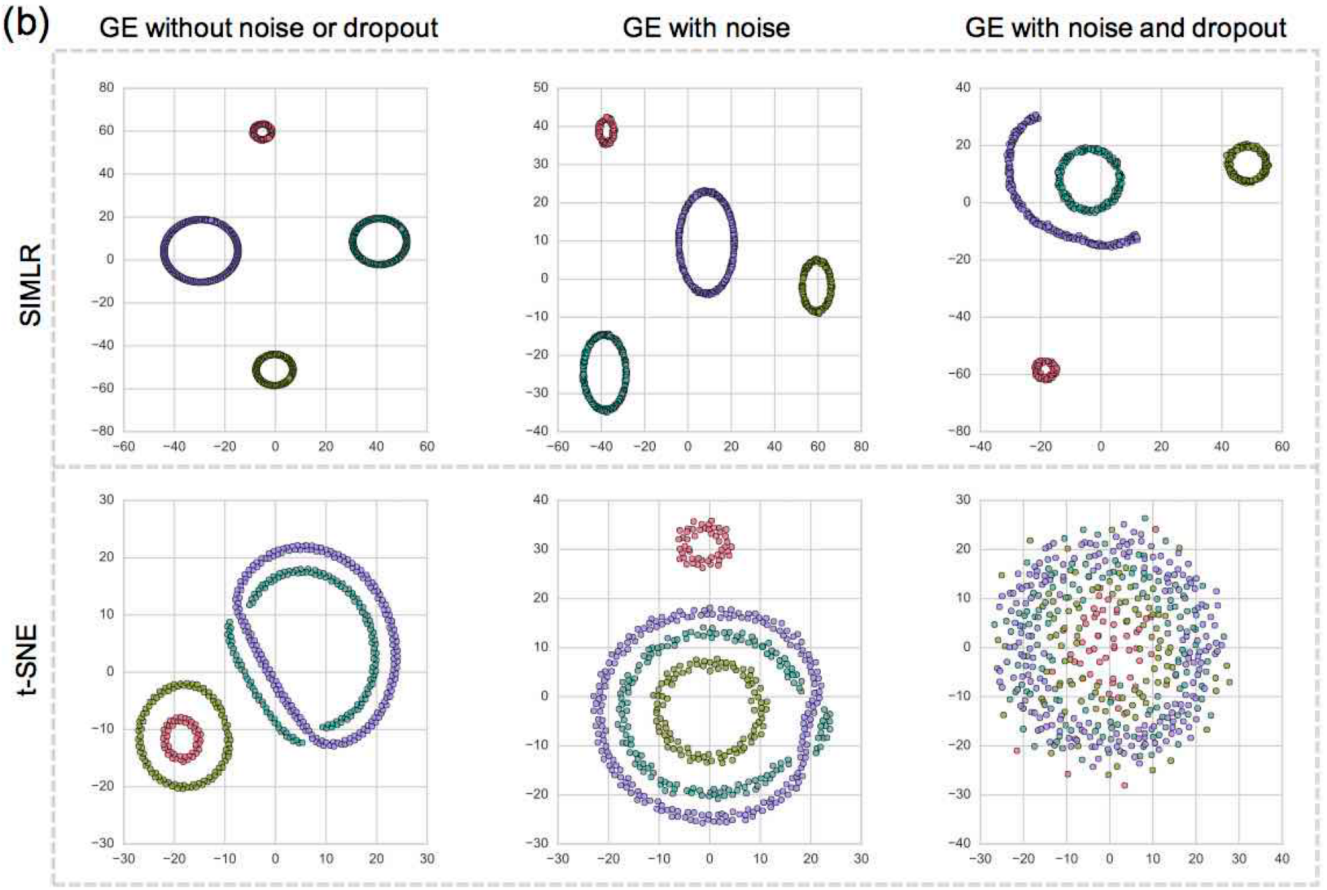
SIMLR vs. t-SNE vs. PCA on Four scRNA-seq Datasets



Wang et al., *Nature Methods* 2017

Analysis of “Toy” Gene Expression Data +/- Noise and Data Dropout

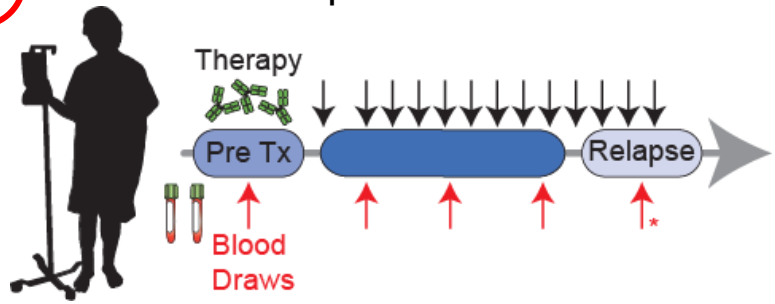




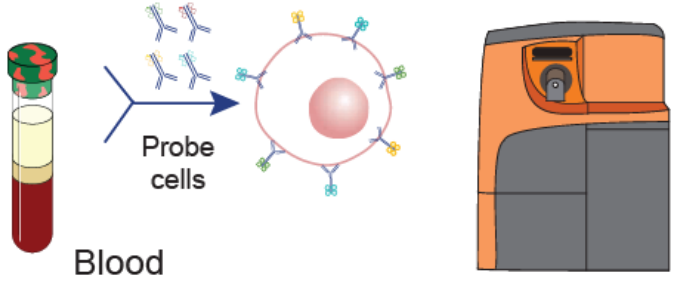
Wang et al., *Nature Methods* 2017

Human Immune Monitoring: T cell Focus

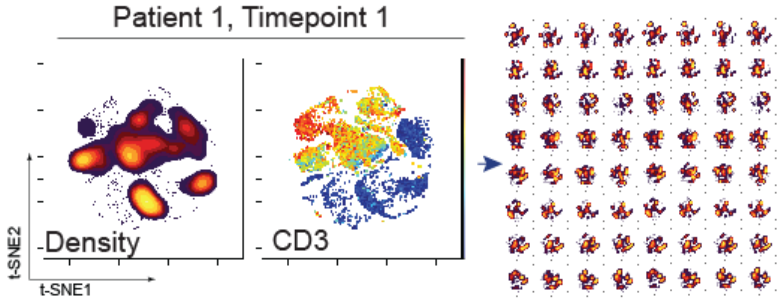
1 Live cells from patients



2 High dimensional, single cell measurements



3 Analyze systems immune changes



Tag Isotope #	Target
141 Pr	ICOS
142 Nd	CD19
143 Nd	TIM3
144 Nd	CCR5
145 Nd	CD4
146 Nd	CD64
147 Sm	CD20
148 Nd	CD38
149 Sm	CCR4
150 Ns	CD43
151 Eu	CD14
152 Sm	TCRgd
153 Eu	CD45RA
154 Sm	CD45
156 Gd	CXCR3
158 Gd	CD33
159 Tb	CCR7
160 Gd	CD28
161 Dy	CD32
162 Dy	CD69
163 Dy	HLA-DR
164 Dy	CD45RO
165 Ho	CD16
166 Er	CD44
167 Er	CD27
168 Er	CD8
169 Tm	CD25
170 Er	CD3
171 Yb	CXCR5
172 Yb	CD57
174 Yb	PD-1
175 Lu	PD-L1
176 Yb	CD56
191/193 Ir	Nucleic Acid
195 Pt	Viability

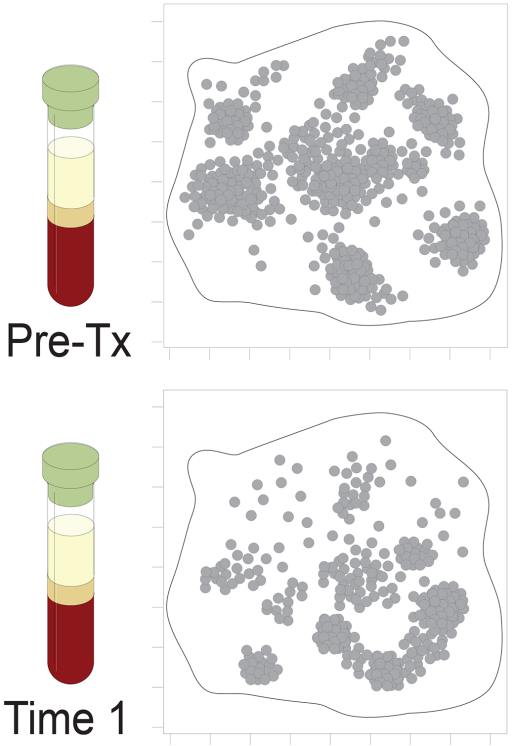
Custom conjugates
Commercially available

Greenplate et al., *Cancer Immunology Research* 2019

Clinical Trial Monitoring: What Do We Need to Know? Automate Four Key Readouts vs. Clinical Outcomes

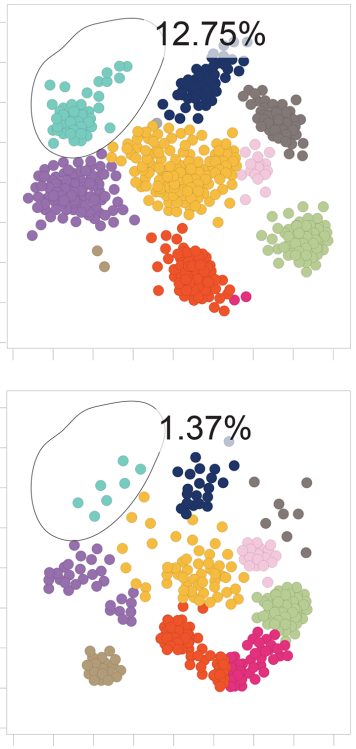
Features of Dynamic Populations

1 Systems Plasticity



Earth Mover's Distance on t-SNE or UMAP

2 Population abundance



Traditional gating or cluster frequency

3 Signature features

Pre-therapy

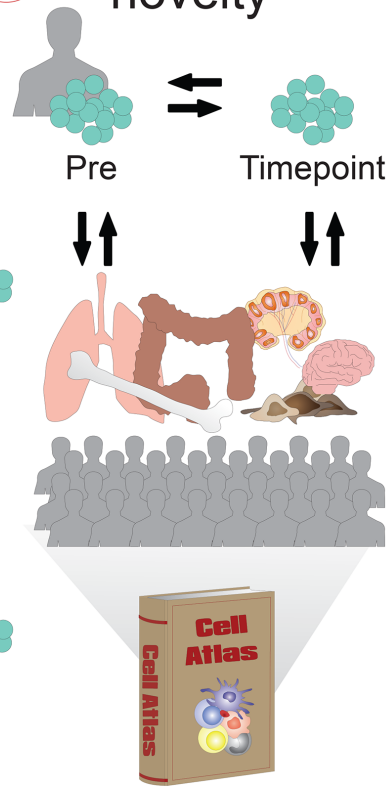
▲HLADR^{+2} CCR5^{+1} CD38^{+1} CD33^{+1}
▼CD8^{-8} CD45RO^{-6} CD3^{-6} CD4^{-4} CD45^{-2} CCR4^{-1} CCR7^{-1} CD28^{-1} CD27^{-1}

Time point 1

▲HLADR^{+2} CD38^{+1} CD45RA^{+1}
▼CD8^{-8} CD4^{-6} CD3^{-6} CD45RO^{-5} CCR5^{-2} CD45^{-2} CD28^{-2} CD20^{-1} CCR4^{-1} CD27^{-1}

Marker Enrichment Modeling (MEM)

4 Population novelty



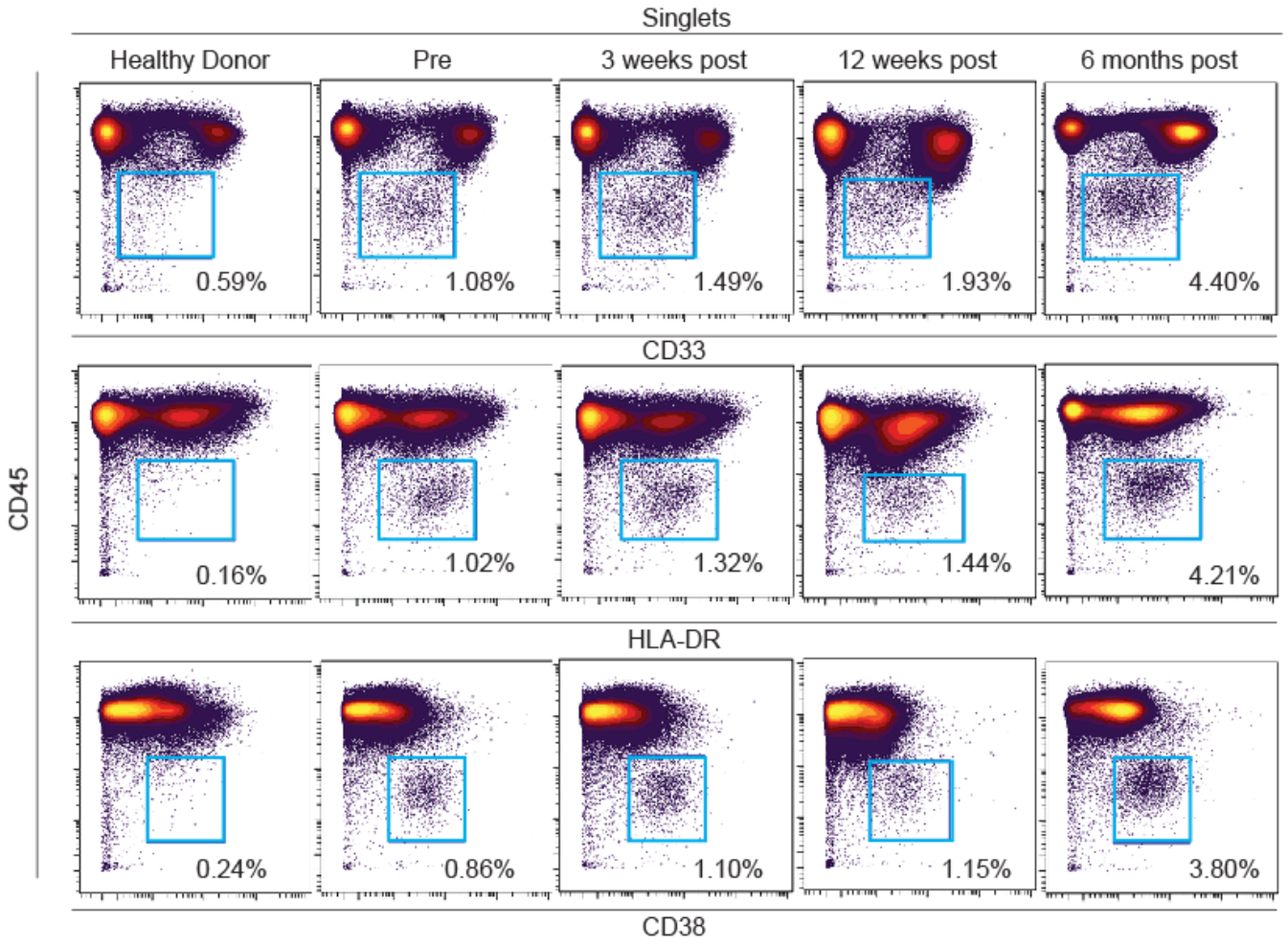
ΔMEM vs. Timepoint or Cell Atlas

How we quantified

Greenplate et al., *Cancer Immunology Research 2019*

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



Mass cytometry data (CyTOF)

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

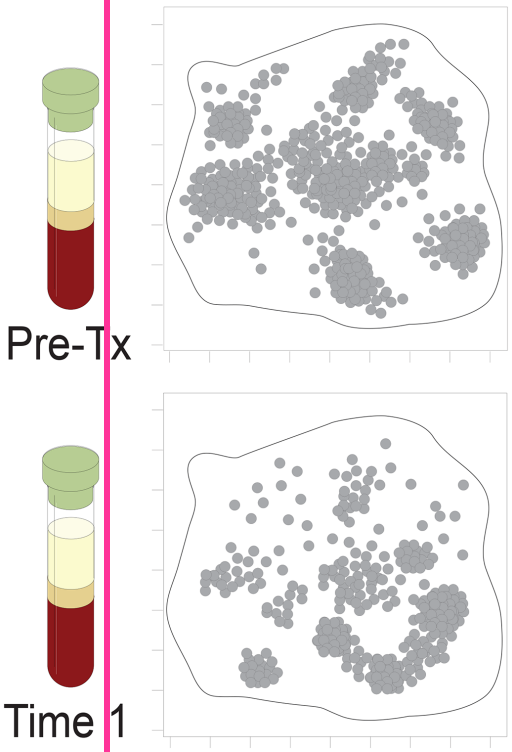
Melanoma data: https://flowrepository.org/id/FR-FCM-ZYDG

Greenplate et al., *Cancer Immunology Research* 2016

Clinical Trial Monitoring: What Do We Need to Know? Automate Four Key Readouts vs. Clinical Outcomes

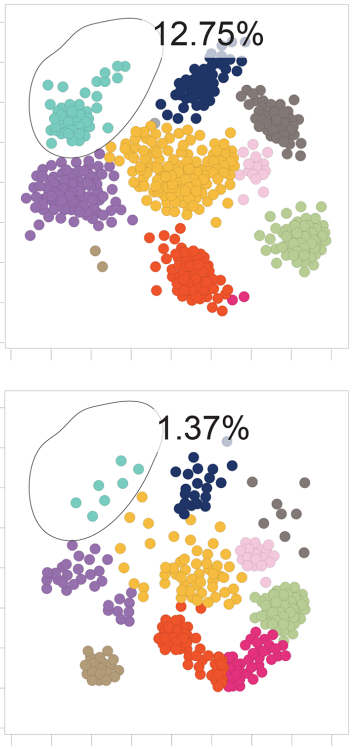
Features of Dynamic Populations

1 Systems Plasticity



Earth Mover's Distance on t-SNE or UMAP

2 Population abundance



Traditional gating or cluster frequency

3 Signature features

Pre-therapy

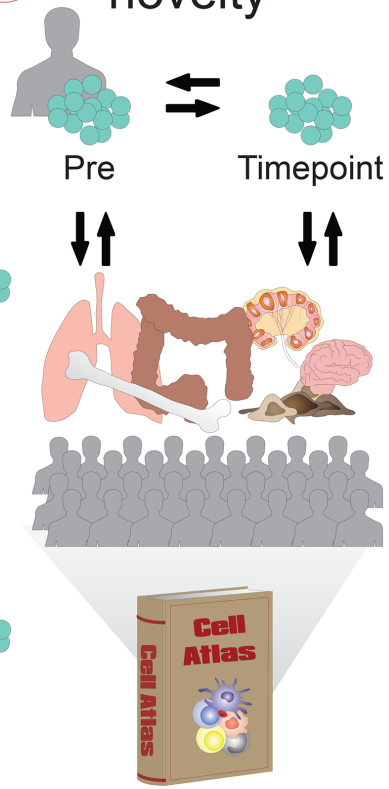
▲HLADR^{+2} CCR5^{+1} CD38^{+1} CD33^{+1}
▼CD8^{-8} CD45RO^{-6} CD3^{-6} CD4^{-4} CD45^{-2} CCR4^{-1} CCR7^{-1} CD28^{-1} CD27^{-1}

Time point 1

▲HLADR^{+2} CD38^{+1} CD45RA^{+1}
▼CD8^{-8} CD4^{-6} CD3^{-6} CD45RO^{-5} CCR5^{-2} CD45^{-2} CD28^{-2} CD20^{-1} CCR4^{-1} CD27^{-1}

Marker Enrichment Modeling (MEM)

4 Population novelty



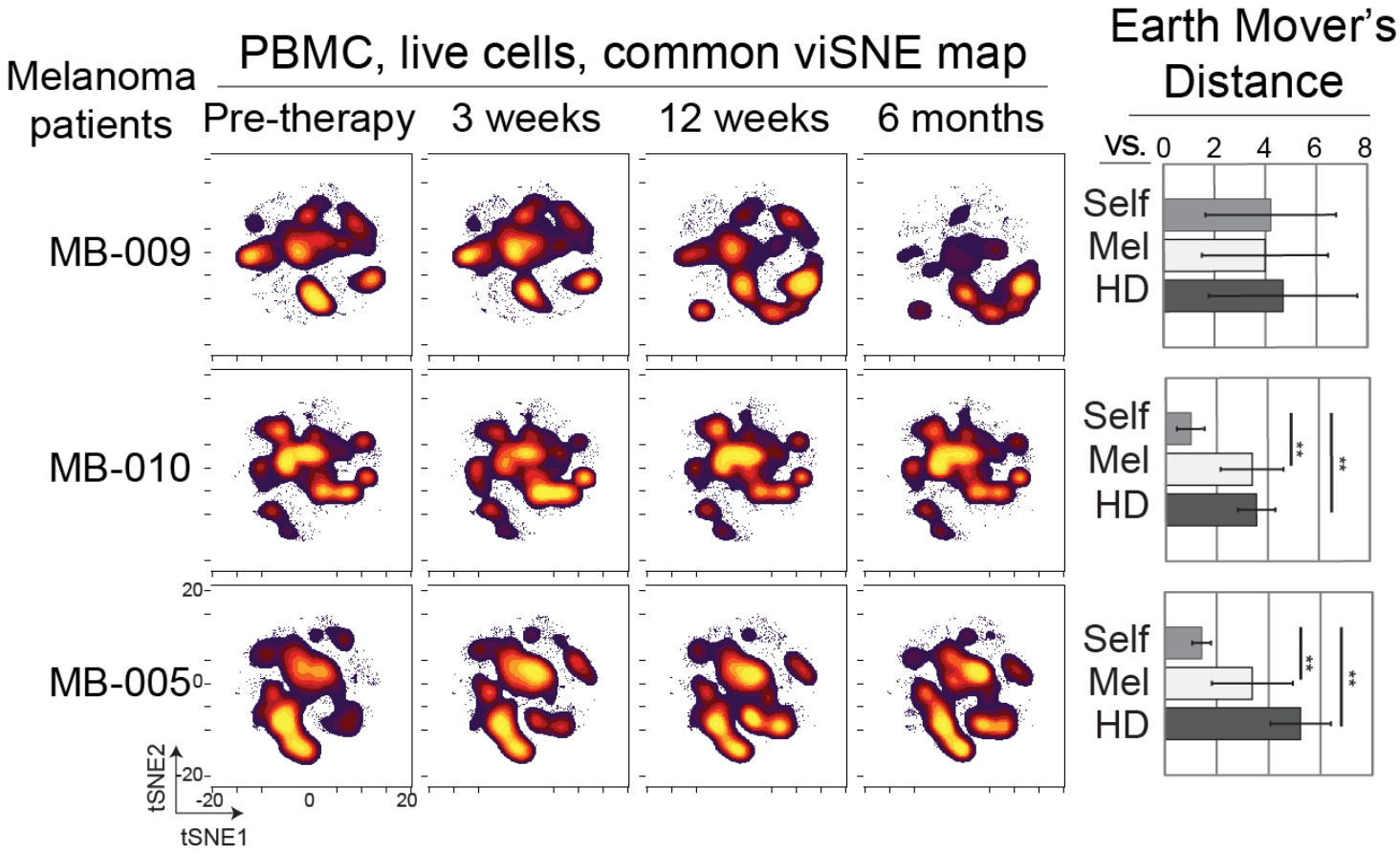
ΔMEM vs. Timepoint or Cell Atlas

How we quantified

Greenplate et al., *Cancer Immunology Research 2019*

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



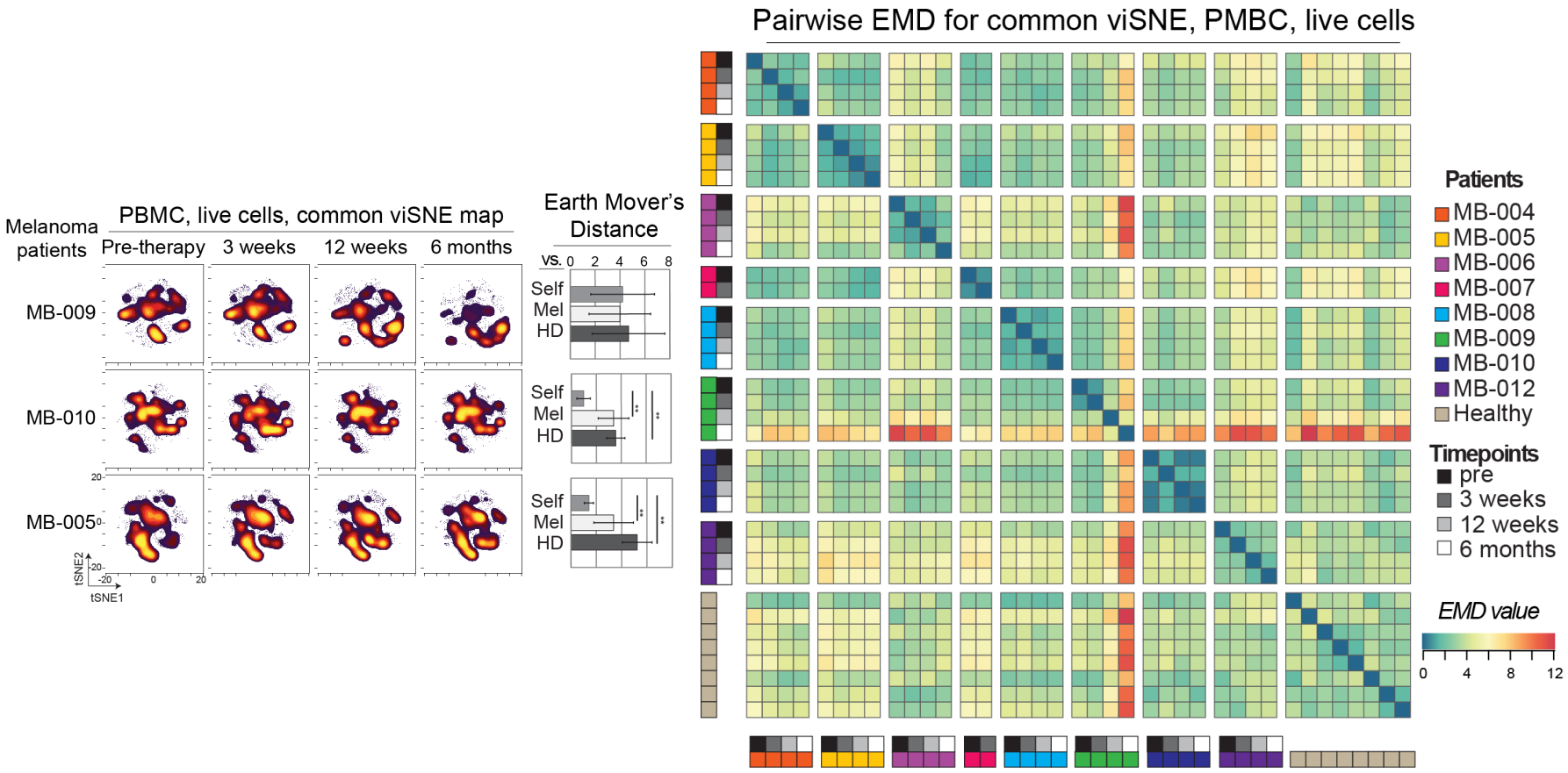
Systems immune monitoring reveals an unexpected pattern in MB-009
Individuals can be their own significantly stable baseline

Melanoma data: https://flowrepository.org/id/FR-FCM-ZYDG

Greenplate et al., *Cancer Immunology Research* 2019
MB-009 Case Study from Greenplate et al., *CIR* 2016

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

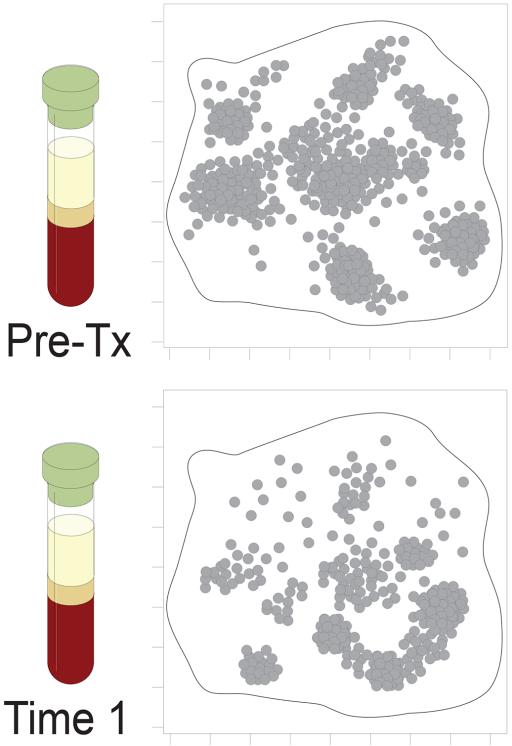
Melanoma data: https://flowrepository.org/id/FR-FCM-ZYDG

Greenplate et al., *Cancer Immunology Research* 2019
MB-009 Case Study from Greenplate et al., *CIR* 2016

Clinical Trial Monitoring: What Do We Need to Know? Automate Four Key Readouts vs. Clinical Outcomes

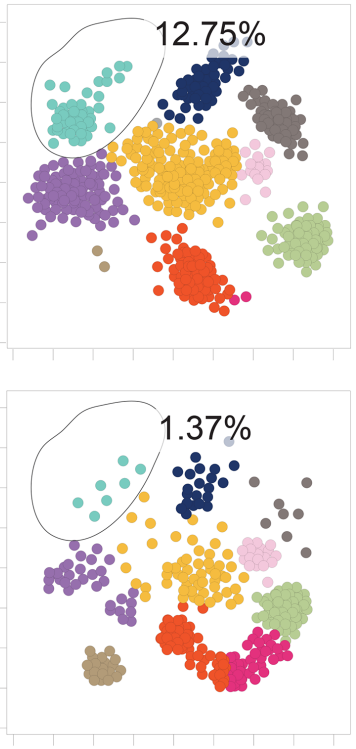
Features of Dynamic Populations

1 Systems Plasticity



Earth Mover's Distance on t-SNE or UMAP

2 Population abundance



Traditional gating or cluster frequency

3 Signature features

Pre-therapy

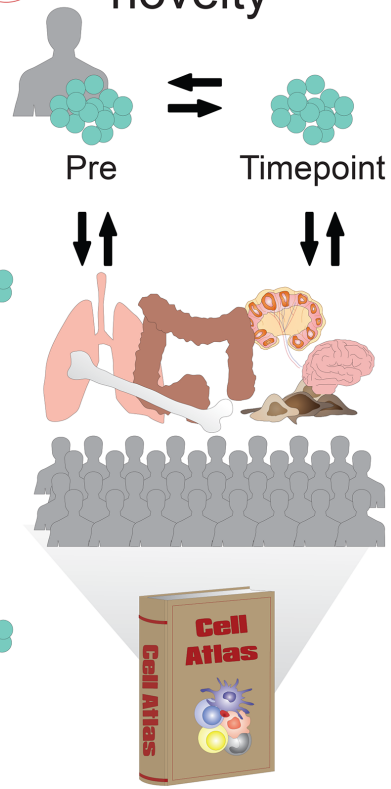
▲ $HLADR^{+2}$ $CCR5^{+1}$ $CD38^{+1}$ $CD33^{+1}$
▼ $CD8^{-8}$ $CD45RO^{-6}$ $CD3^{-6}$ $CD4^{-4}$ $CD45^{-2}$ $CCR4^{-1}$ $CCR7^{-1}$ $CD28^{-1}$ $CD27^{-1}$

Time point 1

▲ $HLADR^{+2}$ $CD38^{+1}$ $CD45RA^{+1}$
▼ $CD8^{-8}$ $CD4^{-6}$ $CD3^{-6}$ $CD45RO^{-5}$ $CCR5^{-2}$ $CD45^{-2}$ $CD28^{-2}$ $CD20^{-1}$ $CCR4^{-1}$ $CD27^{-1}$

Marker Enrichment Modeling (MEM)

4 Population novelty



ΔMEM vs. Timepoint or Cell Atlas

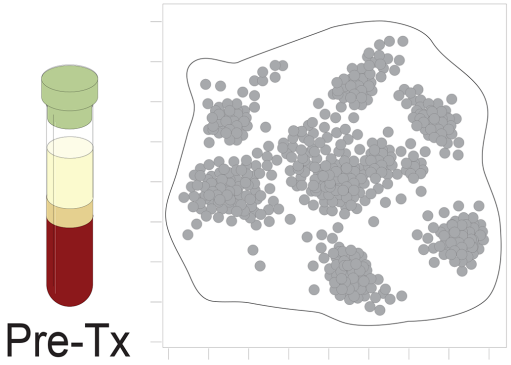
How we quantified

Greenplate et al., *Cancer Immunology Research 2019*

Clinical Trial Monitoring: What Do We Need to Know? Automate Four Key Readouts vs. Clinical Outcomes

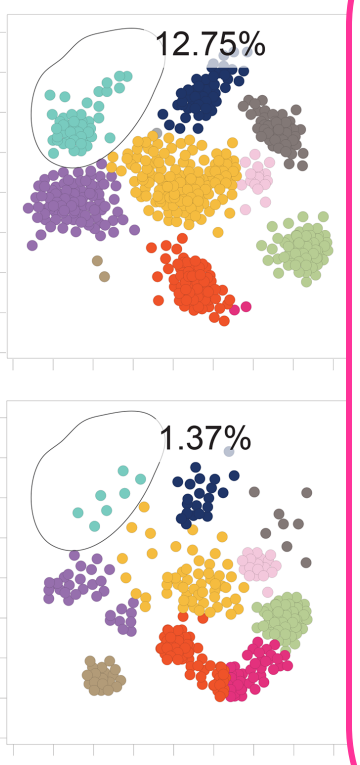
Features of Dynamic Populations

1 Systems Plasticity



Earth Mover's Distance on t-SNE or UMAP

2 Population abundance



Traditional gating or cluster frequency

3 Signature features

Pre-therapy

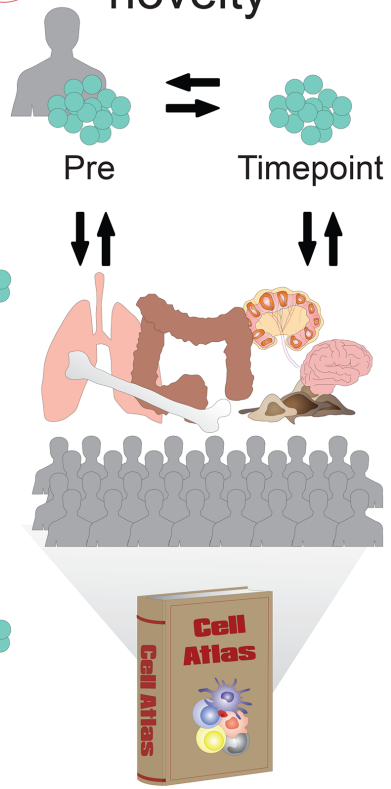
▲HLADR^{+2} CCR5^{+1} CD38^{+1} CD33^{+1}
▼CD8^{-8} CD45RO^{-6} CD3^{-6} CD4^{-4} CD45^{-2} CCR4^{-1} CCR7^{-1} CD28^{-1} CD27^{-1}

Time point 1

▲HLADR^{+2} CD38^{+1} CD45RA^{+1}
▼CD8^{-8} CD4^{-6} CD3^{-6} CD45RO^{-5} CCR5^{-2} CD45^{-2} CD28^{-2} CD20^{-1} CCR4^{-1} CD27^{-1}

Marker Enrichment Modeling (MEM)

4 Population novelty

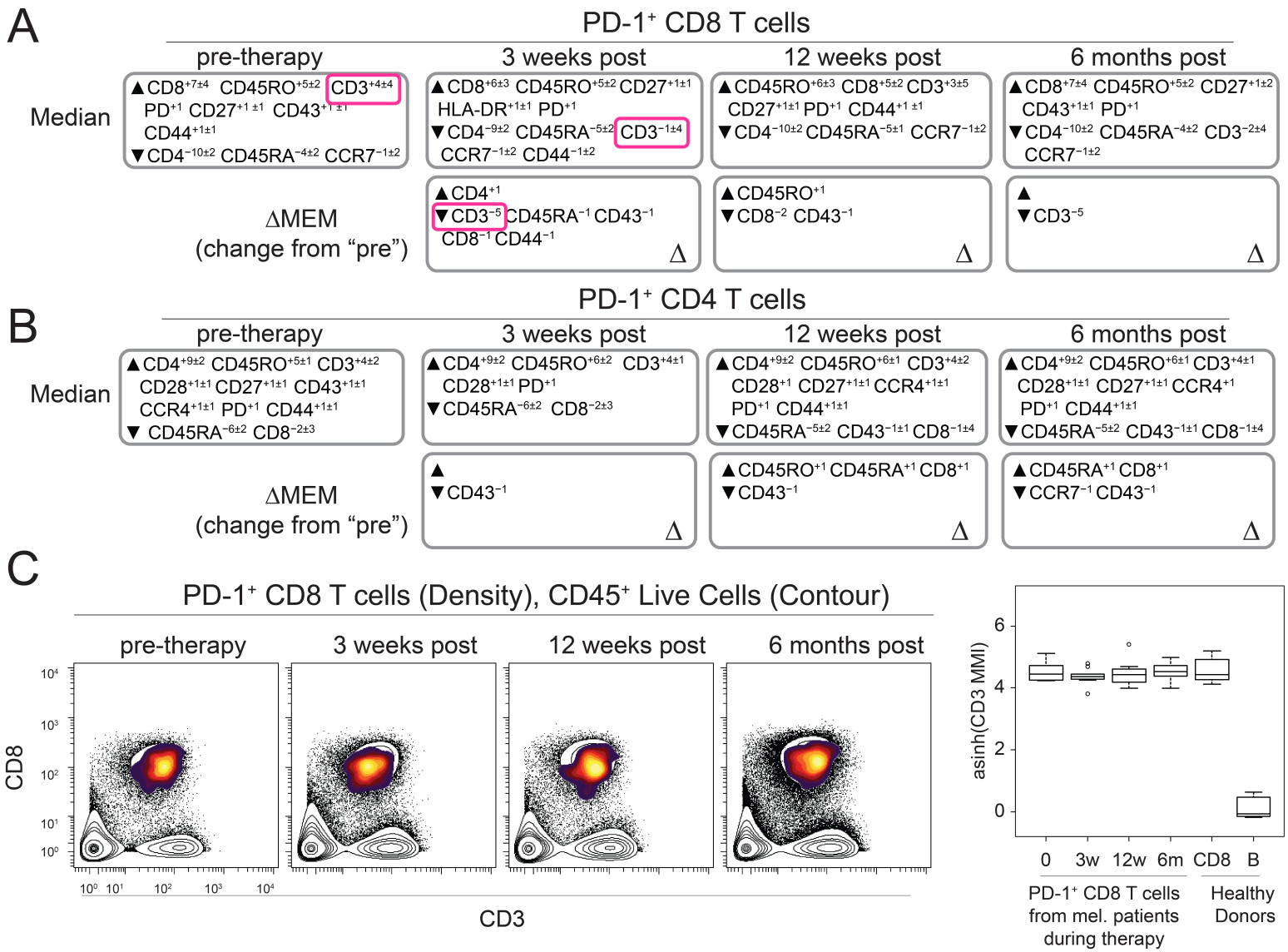


ΔMEM vs. Timepoint or Cell Atlas

How we quantified

Greenplate et al., *Cancer Immunology Research 2019*

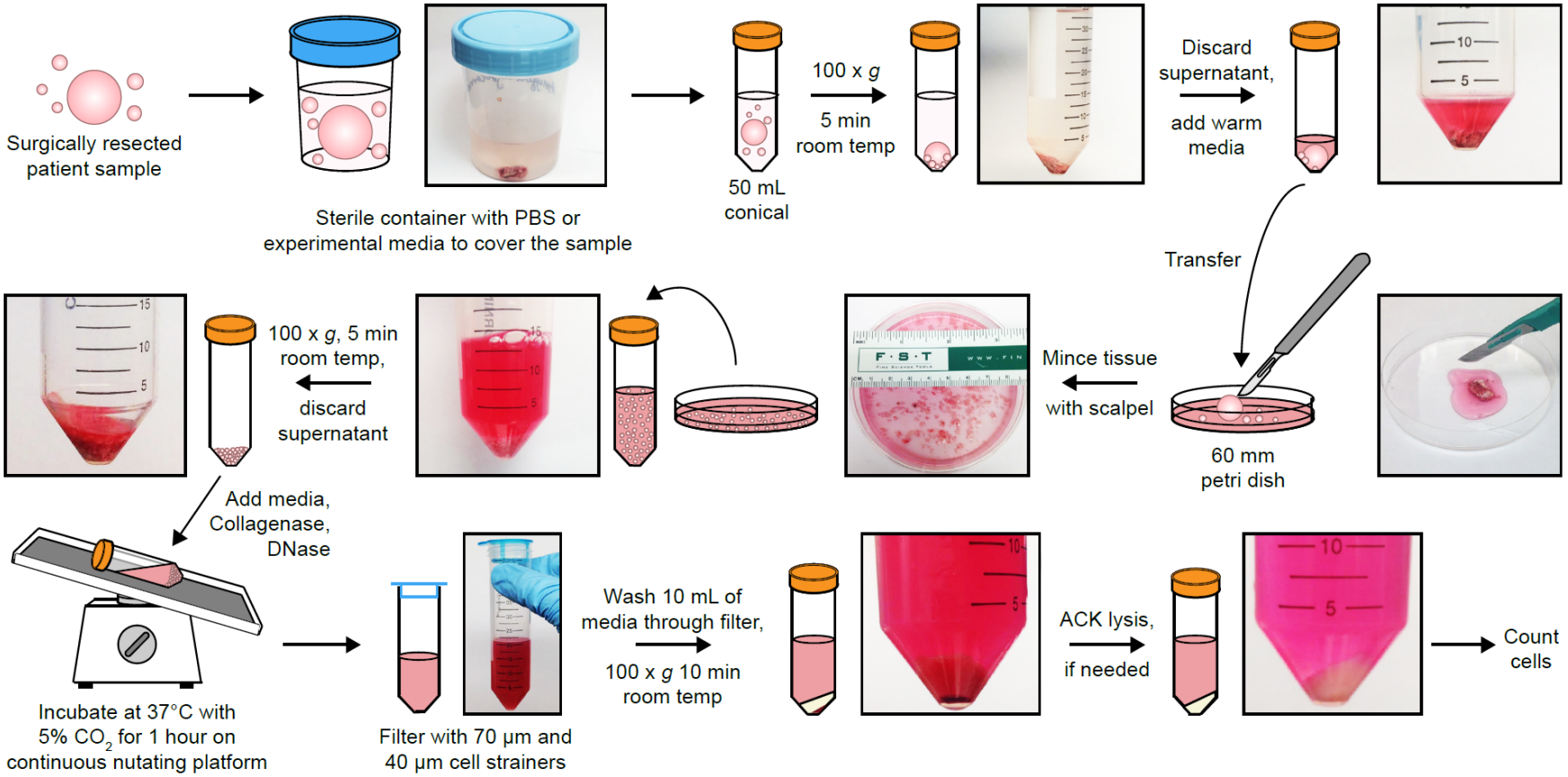
ΔMEM Reveals CD8⁺ Specific Decrease in Per-Cell CD3 in Melanoma Patient PBMC at 3 Weeks after α-PD-1 Therapy



Melanoma data: https://flowrepository.org/id/FR-FCM-ZYDG

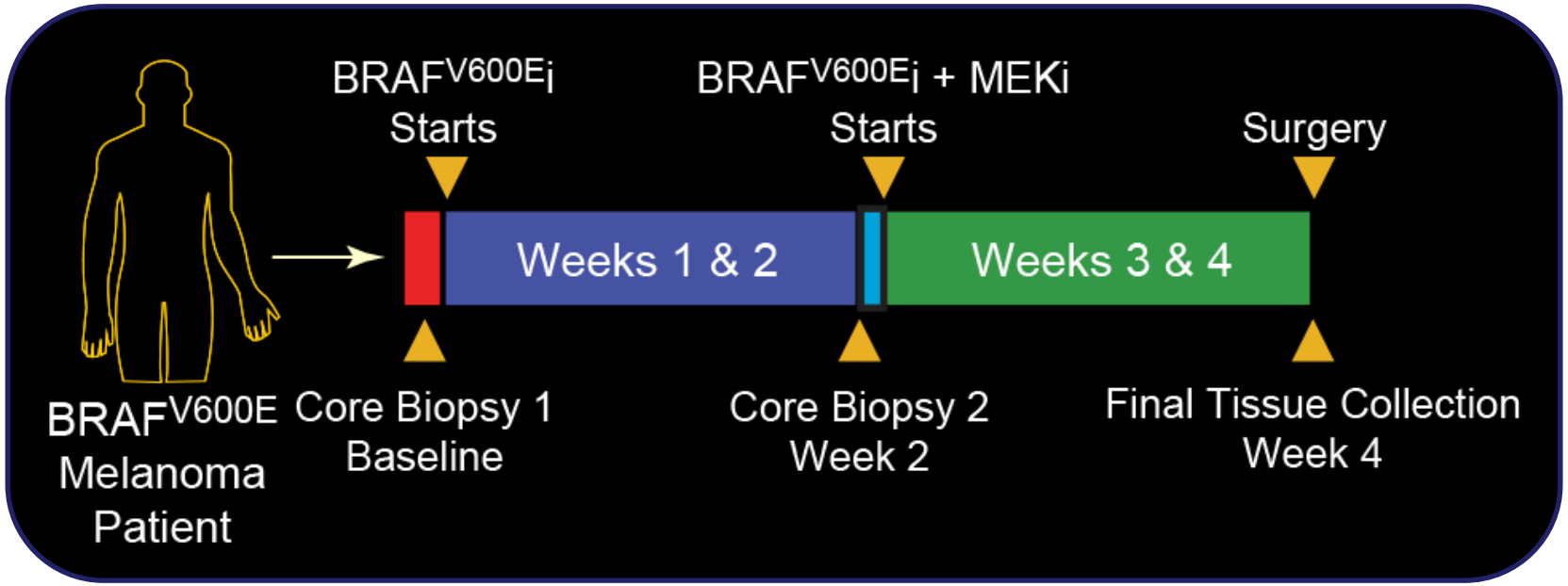
Greenplate et al., *Cancer Immunology Research* 2019

Tissue Biopsies Go from the Operating Room to the Lab and Viable Single Cells Isolated for Functional Studies



Protocol: Leelatian et al., *Current Protocols in Molecular Biology* 2017
Original research: Based on Leelatian and Doxie et al., *Cytometry B* 2017
CIC services: https://my.vanderbilt.edu/cancerimmunology/

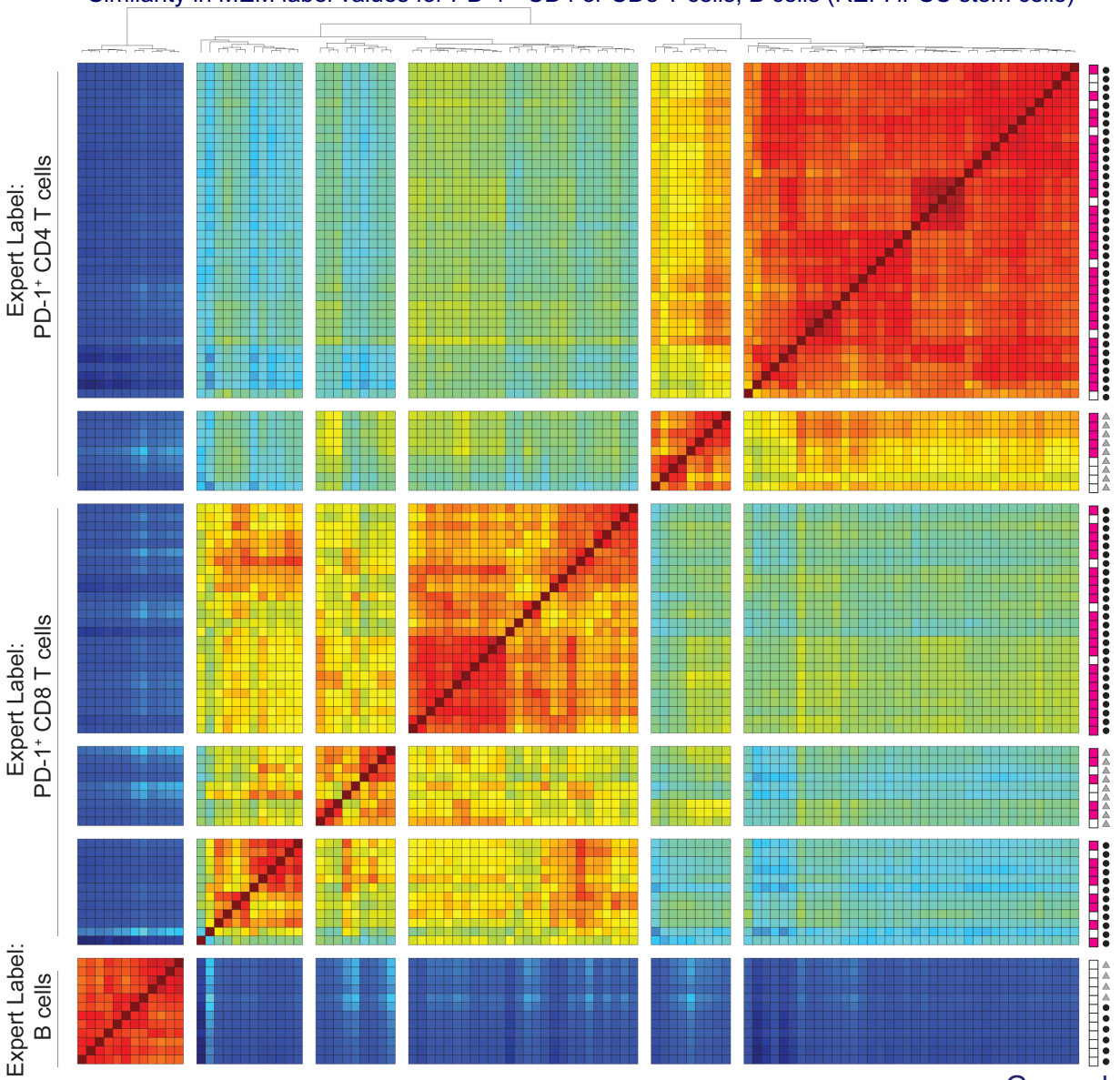
Identify and Characterize Tumor & Immune Cells That Regress or Persist During Clinical Trial Therapies

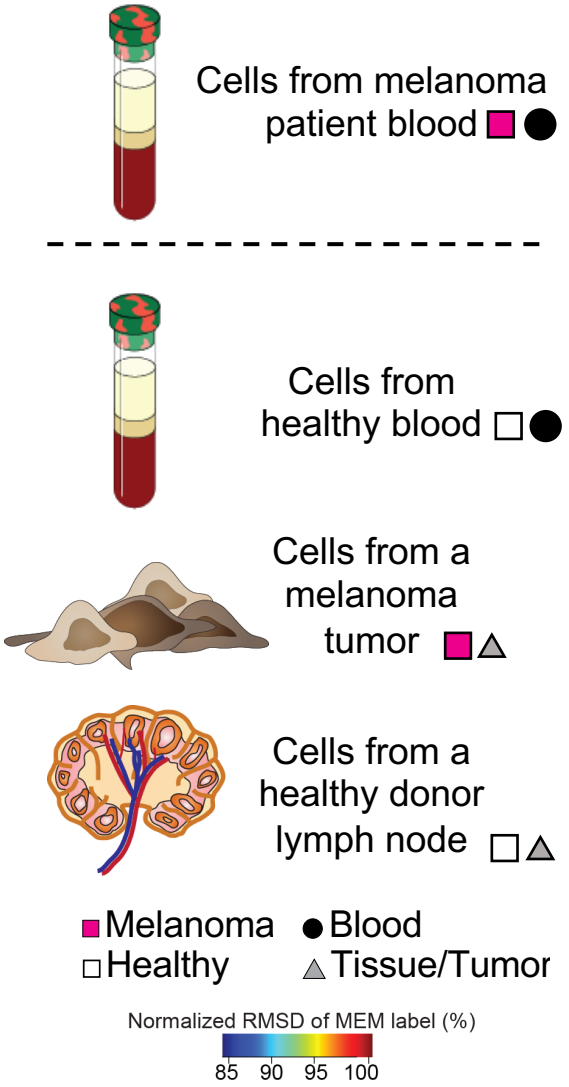


Melanoma patient tissues are brought directly from surgery to the lab for mass cytometry + machine learning single cell analysis

Doxie et al., *Pigment Cell & Melanoma Research* 2018
Greenplate et al., *Cancer Immunology Research* 2019

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



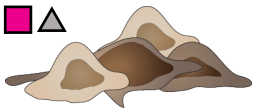


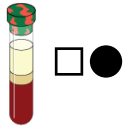
Greenplate et al., *Cancer Immunology Research* 2019

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

Methods: Diggins et al., *Nature Methods* 2017; *Curr Prot Cyt* 2018

Blood vs. Tumor: CXCR3 Is Enriched on PD-1^{+} CD4 and CD8 Melanoma TIL; CD57 Is Gained on Tumor Infiltrating PD-1^{+} CD8





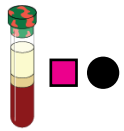
PD-1^{+} **CD8** T cells ΔMEM

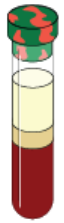
Melanoma tumor	Healthy blood
▲CD43^{+3} CD57^{+2} CCR4^{+2} CXCR3^{+2} CD27^{+1} CD44^{+1} ▼CD38^{-2} CD69^{-2} CD45RO^{-2} Δ	▲CCR4^{+2} CD45RO^{+1} ▼CD57^{-2} Δ

PD-1^{+} **CD4** T cells ΔMEM

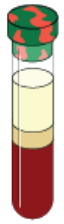
Melanoma tumor	Healthy blood
▲CD4^{+3} CD43^{+2} CXCR3^{+2} CCR5^{+1} CD44^{+1} ▼CD69^{-2} CD45RO^{-2} CD38^{-1} ICOS^{-1} Δ	No difference Δ

ΔMEM labels each subset vs. cognate in melanoma blood

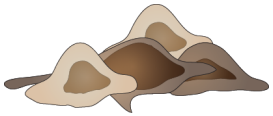




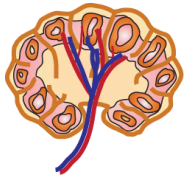
Cells from melanoma patient blood



Cells from healthy blood



Cells from a melanoma tumor



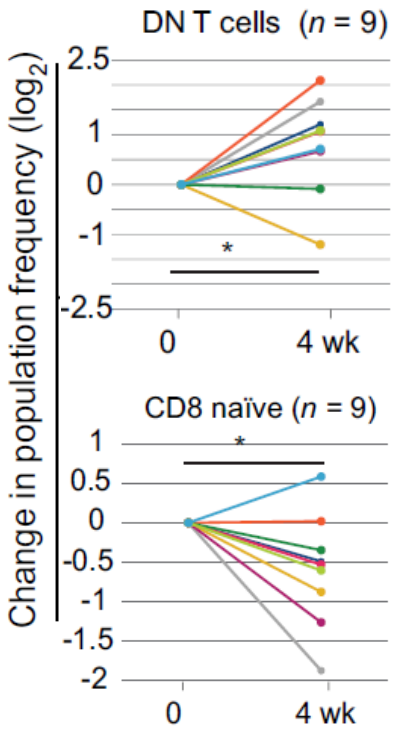
Cells from a healthy donor lymph node

Melanoma ● Blood
□ Healthy △ Tissue/Tumor

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

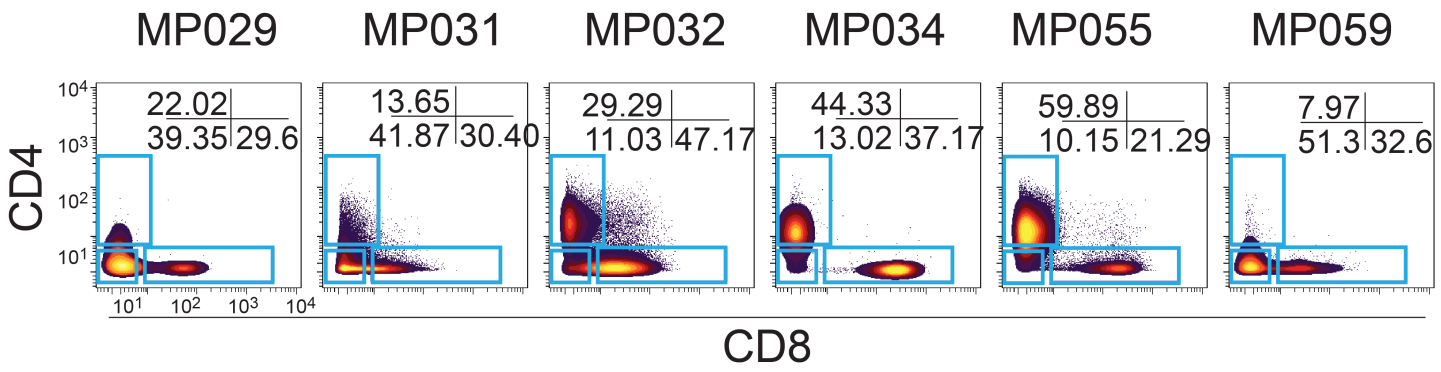
Greenplate et al., *Cancer Immunology Research* 2019
Methods: Diggins et al., *Nature Methods* 2017; *Curr Prot Cyt* 2018

Abnormal CD4$^-$ CD8$^-$ Double Negative (DN) T cells Are Enriched in Melanoma Tumors Following BRAFi + MEKi

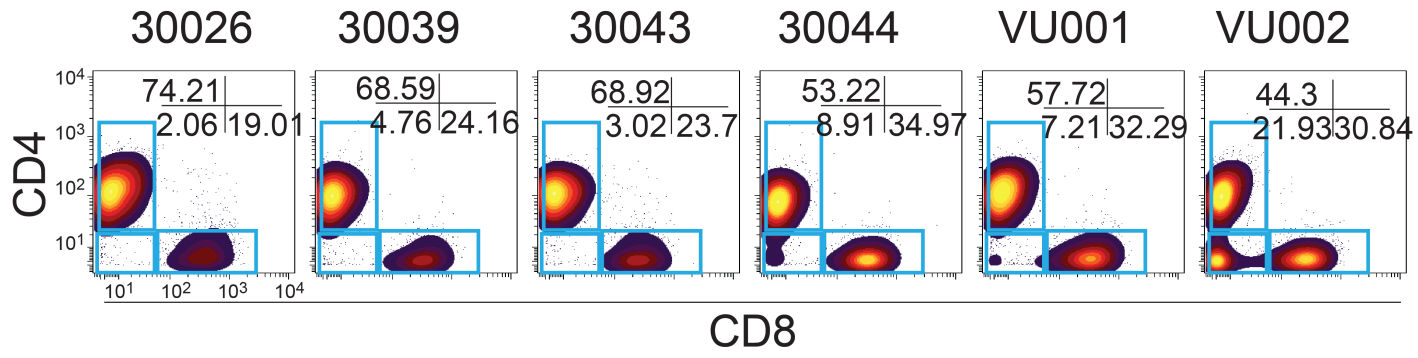


(DN T cells here were γδ negative)

Melanoma patient biopsy, all CD3+ tumor-infiltrating T cells (4 weeks after BRAF + MEK inhibitor therapy start)



Comparison point: activated T cells from healthy blood (Staphylococcal enterotoxin B stimulated)

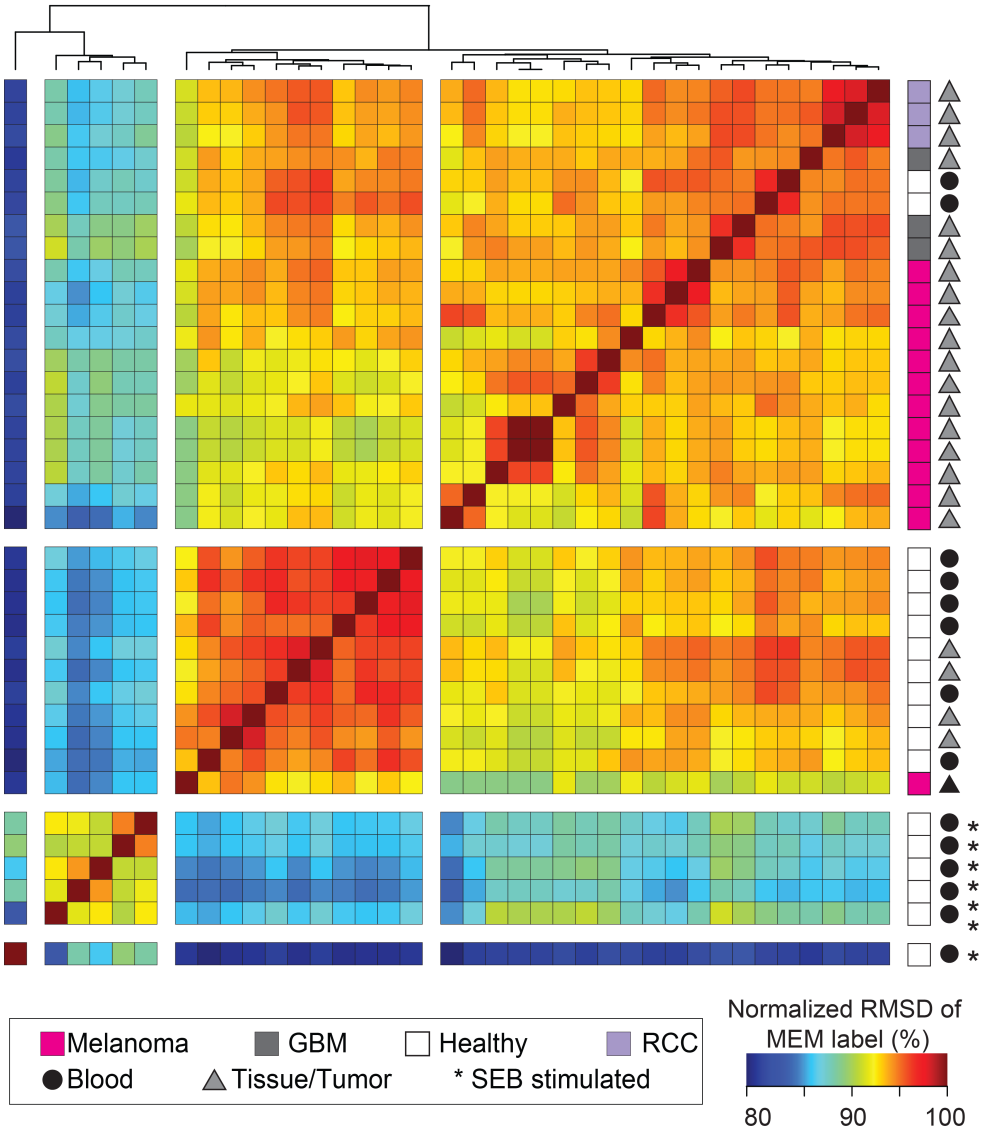


Data files: http://flowrepository.org/id/FR-FCM-ZYCC

Greenplate et al., *Cancer Immunology Research* 2019

Similar Abnormal Tumor-infiltrating CD4$^-$ CD8$^-$ DN T cells Are Observed Across Diverse Tumor Types

Comparing MEM labels for DN T cells across human tissues



Median MEM Labels iPCs as reference

RCC Tumor

▲CD45$^{+7\pm1}$ CD45RO$^{+2\pm0}$ CD3$^{+2\pm0}$ CD44$^{+2\pm1}$
▼CD57$^{-5\pm0}$ CD56$^{-3\pm0}$ CCR7$^{-1\pm0}$ PD1$^{-1\pm0}$
CD45RA$^{-1\pm1}$

GBM Tumor

▲CD45$^{+8\pm1}$ CD45RO$^{+3\pm1}$ CD3$^{+2\pm0}$ CD69$^{+2\pm1}$
CD44$^{+2\pm1}$
▼CD57$^{-5\pm2}$ CD56$^{-3\pm1}$ PDL1$^{-2\pm0}$ CCR5$^{-1\pm1}$

Melanoma Tumor

▲CD45$^{+7\pm3}$ CD3$^{+3\pm1}$ CD45RO$^{+2\pm3}$ CD44$^{+2\pm3}$
▼CD57$^{-6\pm2}$ CCR4$^{-3\pm1}$ CD56$^{-2\pm1}$ PDL1$^{-2\pm1}$

Non-malignant LN

▲CD45$^{+8\pm1}$ CD3$^{+4\pm0}$ CD44$^{+2\pm1}$ CD45RA$^{+1\pm2}$
▼CD57$^{-5\pm1}$ CCR4$^{-3\pm0}$ CD56$^{-3\pm0}$ PDL1$^{-2\pm0}$
CCR7$^{-1\pm0}$ CXCR3$^{-2\pm0}$

Healthy Donor PBMC

▲CD45$^{+10\pm1}$ CD3$^{+4\pm1}$ CD44$^{+2\pm1}$ CD45RA$^{+1\pm2}$
▼ CD57$^{-5\pm2}$ CCR4$^{-3\pm0}$ CD56$^{-3\pm0}$ PDL1$^{-2\pm0}$
CCR7$^{-1\pm0}$ CD28$^{-1\pm1}$ CXCR3$^{-1\pm1}$ CCR5$^{1\pm1}$

Activated T cells PBMC

▲CD45$^{+10\pm1}$ CD45RO$^{+6\pm2}$ CD69$^{+5\pm1}$ CD4$^{+4\pm2}$
CD3$^{+3\pm1}$ CCR7$^{+3\pm2}$ HLADR$^{+2\pm1}$ CD8$^{+2\pm1}$
CD27$^{+2\pm2}$ CD25$^{+2\pm3}$
▼CD57$^{-5\pm1}$ CD16$^{-1\pm0}$

Greenplate et al., *Cancer Immunology Research* 2019

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

Methods: Diggins et al., *Nature Methods* 2017; *Curr Prot Cyt* 2018

Big Picture Topics for Today

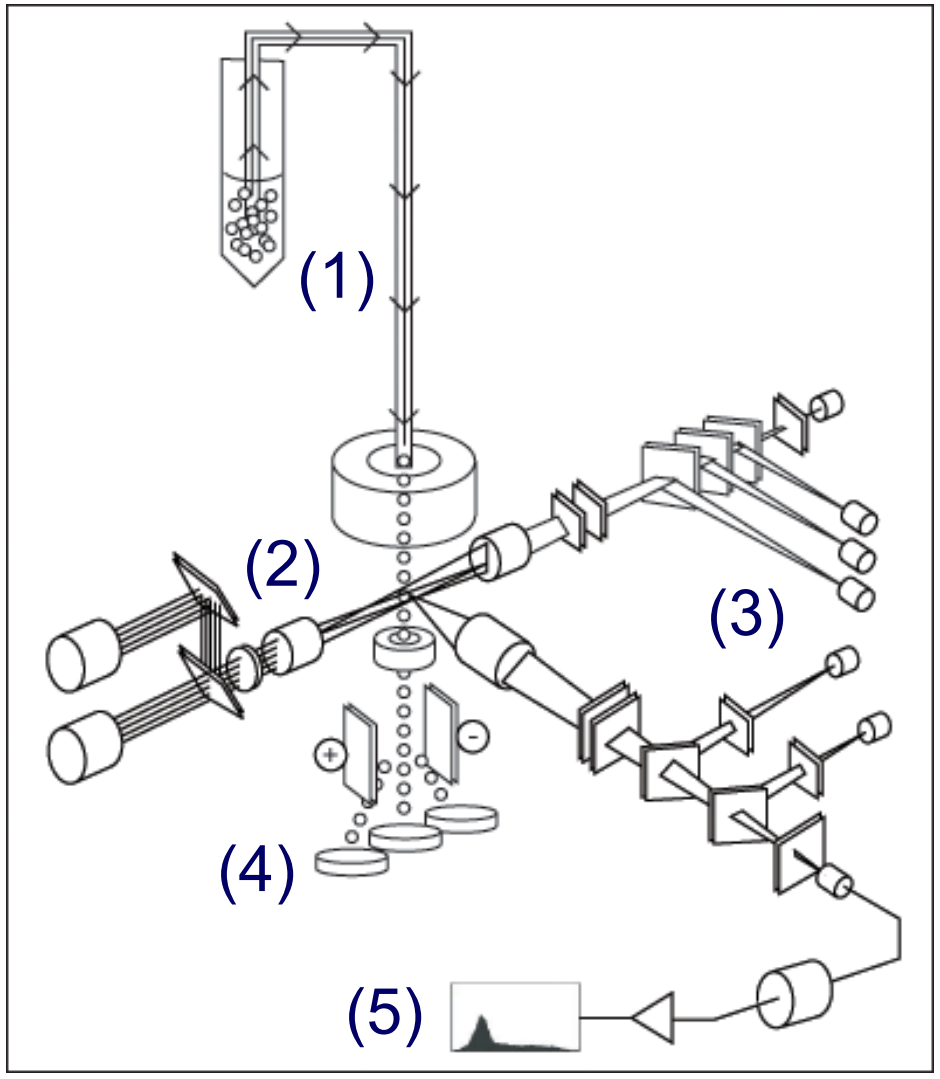
1) How are cytometry data structured & visualized?

2) Qualitative vs. quantitative single cell analysis

3) Supervised & unsupervised analysis tools

4) Overview of analysis methods & data science

Flow Cytometry & FACS (ca. 1968 - 2010)



(1) Analyzes cells in single file
- Mixed populations of cells right out of the body can be studied
- Pressurized stream of droplets containing one cell each.

(2) LASERs excite fluorophore dyes
- Dyes within the cells used to measure proteins & cell features.

(3) Light from fluorophores is filtered & reflected to detectors
- Each detector measures a different cell property.

(4) Optional sorting of cells by magnets

(5) Computational analysis is beginning to play a major role

Flow Cytometry Quantifies Cell Biology

- Differentiation state
- DNA or RNA content & copy #
- Cell cycle stages
- Proliferation
- Oncogene expression
- Mutant proteins
- Tumor suppressor activity
- Apoptosis
- Membrane & cytoskeleton
- Redox state
- Tumor antigens
- Signaling activity
- Endogenous fluorescence

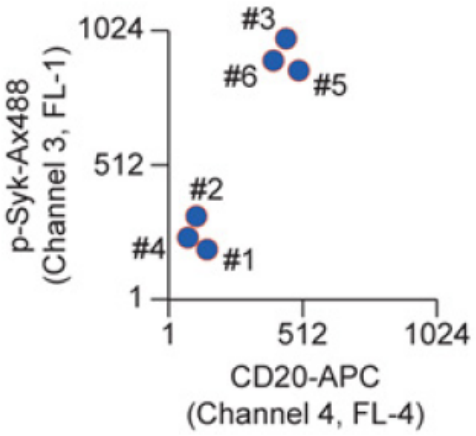
Table 1 | **Determining phenotypes of individual cancer cells**

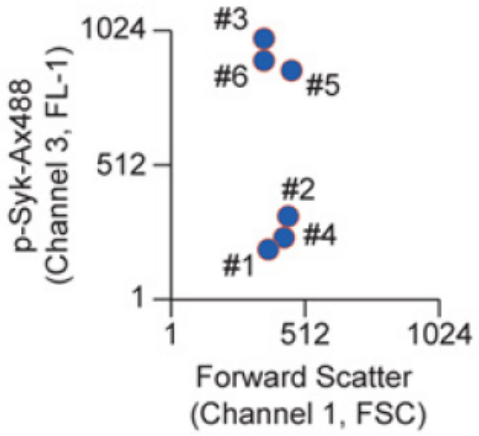
Cell property*	Example flow-cytometry method	References
Differentiation and lineage determination	Antibodies against KIT, CD34 (stem cells), CD38 or CD20, and other CD antigens	29–32
DNA content (aneuploidy, DNA fragmentation)	Propidium iodide, ethidium monoazide or 7-actinomycin D staining of DNA	30,33
RNA content (quiescence)	Pyronin Y staining of RNA	30
Cell-cycle stage	Antibodies against cyclin D, cyclin A, cyclin B1 or cyclin E; phosphorylated form of histone H3 (M phase)	30,34,35
Proliferation	Bromodeoxyuridine staining of DNA replication; antibodies against proliferating cell nuclear antigen; antibodies against Ki67; carboxyfluorescein diacetate succinimidyl ester dye	30,31,36,37
Oncogene expression	Antibodies against BCL2, MYC or Ras	31,38–40
Mutations	Antibodies against mutant p53 or $HRAS^{V12}$	41,42
Tumour-suppressor activity	Antibodies against p53 or p21 (also known as WAF1) promoter activity based on expression of green fluorescent protein (p53R–GFP system)‡; antibodies against the phosphorylated form of p53‡	23,41
Apoptosis	Antibodies against caspase 3 cleavage products	44
Cell-membrane changes	AnnexinV staining for extracellular phosphatidylserine exposure, which occurs on apoptotic cells	44
Redox state	Dichlorofluorescein diacetate staining, which is a measure of oxidation; monobromobimane staining, which is a measure of glutathione; lipophilic fluorochrome dihexaoxacarbocyanine iodide staining, which is a measure of mitochondrial membrane potential	44–46
Tumour antigens	Antibodies against B- or T-cell receptor idiotype; tetramers against tumour antigen-specific T cells (for example, against tyrosinase)	5,47,48
Signalling activity	Antibodies against phosphorylated signal transducer and activator of transcription 5, extracellular-regulated kinases 1 and 2, and many others; indo-1 staining for Ca^{2+} flux; antibodies against interleukin 12, interferon-γ or other cytokines	4,48–50

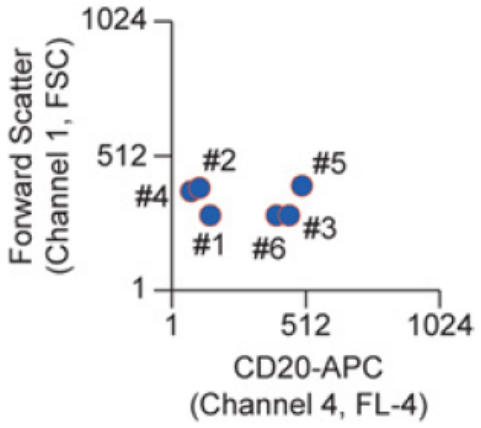
Doxie & Irish 2014; Irish, Kotecha, & Nolan 2006

Example Four-Parameter Cytometry Data (Event List)

	Channel 1	Channel 2	Channel 3	Channel 4
Cell Event #1	400	290	5	50
Cell Event #2	425	301	18	45
Cell Event #3	402	292	912	503
Cell Event #4	422	303	14	40
Cell Event #5	430	310	892	510
Cell Event #6	402	282	903	499
Long Channel Name	**Forward Scatter**	**Side Scatter**	**Fluorescent Channel 1 (FL-1)**	**Fluorescent Channel 4 (FL-4)**
What biologists call this channel:	Forward Scatter (FSC)	Side Scatter (SSC)	phospho-Syk-Alexa488 (p-Syk-Ax488)	CD20-APC

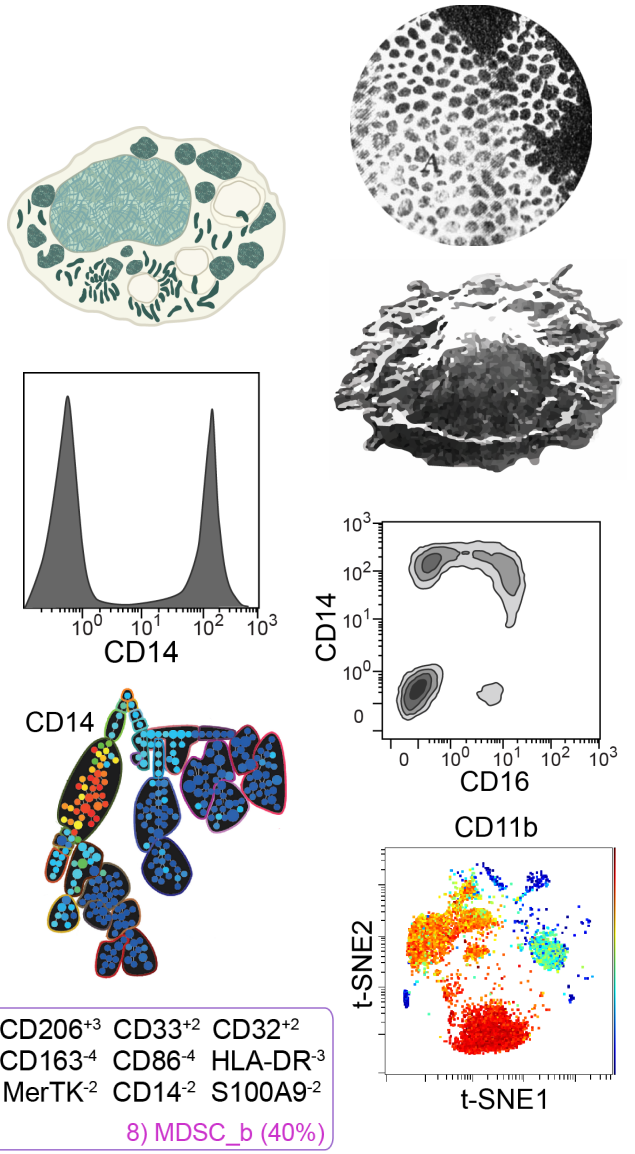






New Technology Reveals & Characterizes New Cells

Date	Approach	Dimensions (D) Per Cell & Speed	
1665*	Light microscopy	Low	Low
1908**	Light microscopy	Low	Low
1946	Scanning EM	Low	Low
1989	Flow cytometry identification	Low	1K cells/s
2001	Flow cytometry subsetting	4D	2 – 50K cell/s
2011	Mass cytometry + SPADE	32D	500 cell/s
2014	Mass cytometry + t-SNE / viSNE	38D	500 cell/s
(now)	Flow or Imaging MC + UMAP, FlowSOM, MEM	38D	500 cell/s



* Robert Hooke describes 'cells' in *Micrographia: or Some Physiological Descriptions of Miniature Bodies Made by Magnifying Glasses*

** Élie Metchnikoff characterizes mononuclear phagocytes: Lectures on the Comparative Pathology of Inflammation, Pasteur Institute in 1891, Nobel Prize in 1908 w/ Ehrlich.

Adapted from
Roussel et al., *Human Innate Immunity* 2016

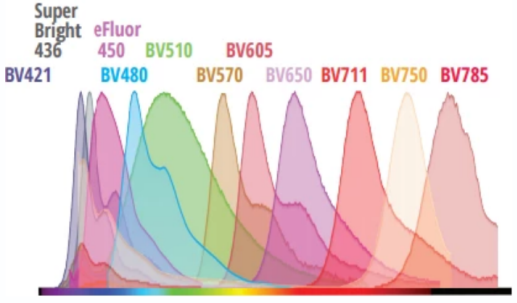
Spectral Flow Cytometry Can Separate ‘Overlapping’ Probes

More Choice, Greater Flexibility, Easier Setup

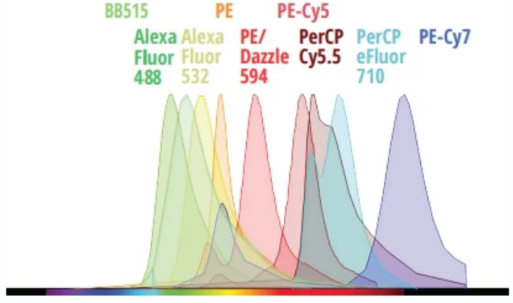
The optical design combined with the unmixing capability in SpectroFlo® software allows greater fluorochrome choice, panel flexibility, and easy setup without having to change filters. The three laser configuration provides outstanding multi-parametric data for a wide array of applications. Markers and fluorochromes in a 24-color panel designed for identification of circulating cell subsets in human peripheral blood are summarized in the table below:

SPECIFICITY	FLUOROCHROME	SPECIFICITY	FLUOROCHROME	SPECIFICITY	FLUOROCHROME
CCR7	Brilliant Violet 421™	CD11c	BD Horizon™ BB515	CD27	APC
CD19	Super Bright 436	CD45RA	Alexa Fluor® 488	CD123	Alexa Fluor® 647
CD16	eFluor® 450	CD3	Alexa Fluor® 532	CD127	BD Horizon™ APC R700
TCR γ/δ	BD Horizon™ BV480	CD25	PE	HLA DR	APC/Fire™ 750
CD14	Brilliant Violet 510™	IgD	PE/Dazzle™ 594		
CD8	Brilliant Violet 570™	CD95	PE-Cy™5		
CD1c	Brilliant Violet 605™	CD11b	PerCP-Cy™5.5		
PD-1	Brilliant Violet 650™	CD38	PerCP-eFluor® 710		
CD56	Brilliant Violet 711™	CD57	PE-Cy™7		
CD4	Brilliant Violet 750™				
CD28	Brilliant Violet 785™				

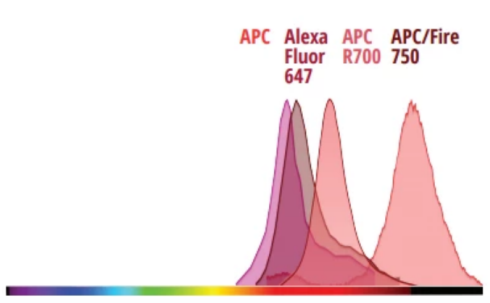
The 24-Color Panel Includes Many Highly Overlapping Dyes:



Violet Excited Dyes Emission Spectra

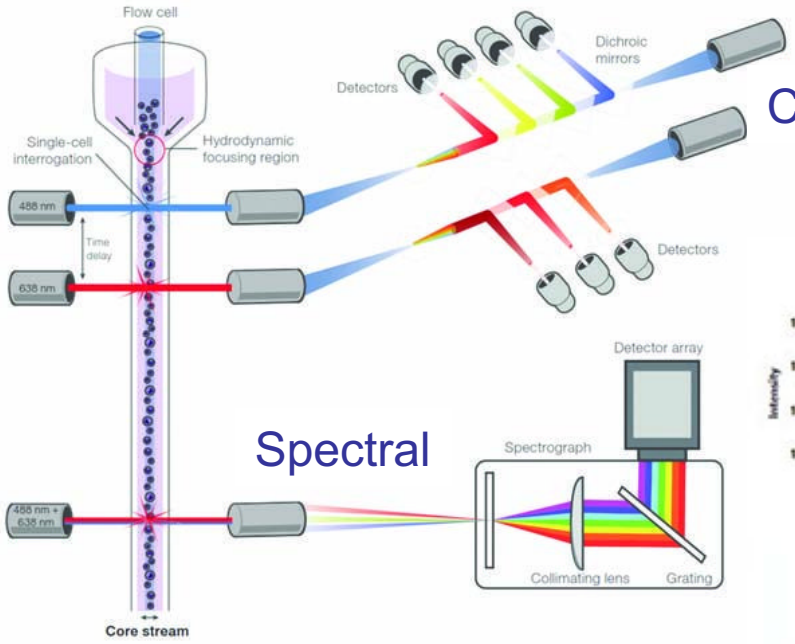


Blue Excited Dyes Emission Spectra



Red Excited Dyes Emission Spectra

By Detecting “Across the Spectrum” (vs. “Channels” of Light) Spectral Flow Distinguishes Probes with Overlapping Peaks

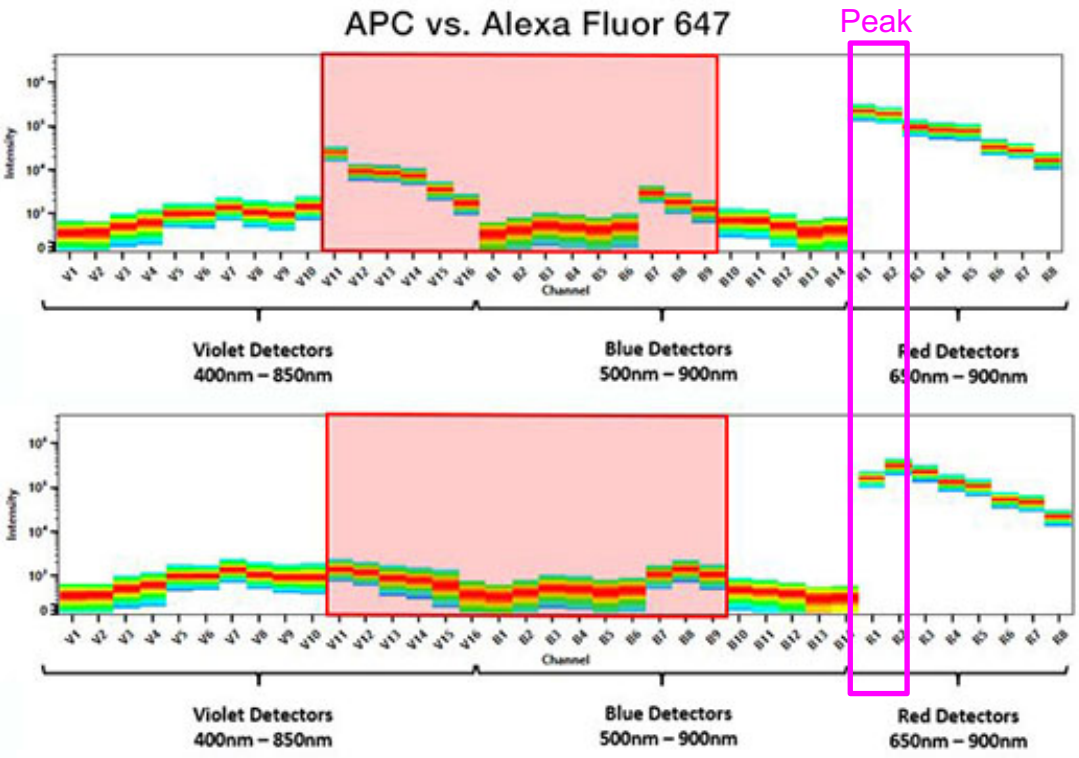


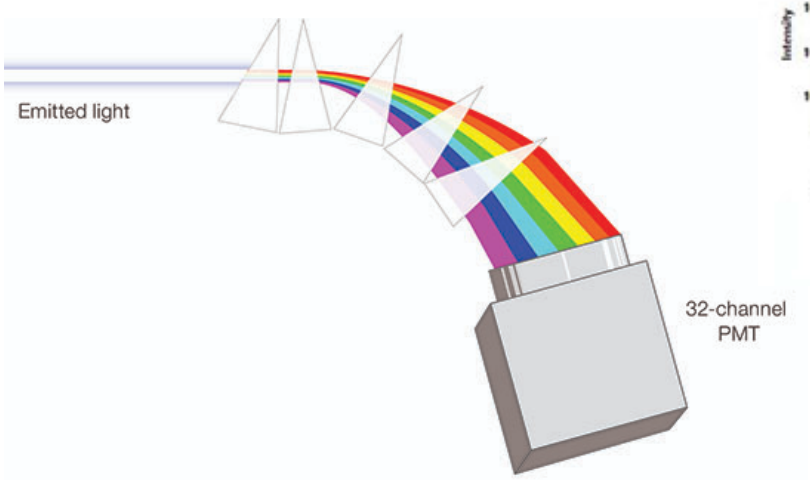
Conventional

Spectral

Themo Fisher, adapted from John Nolan et al. 2013

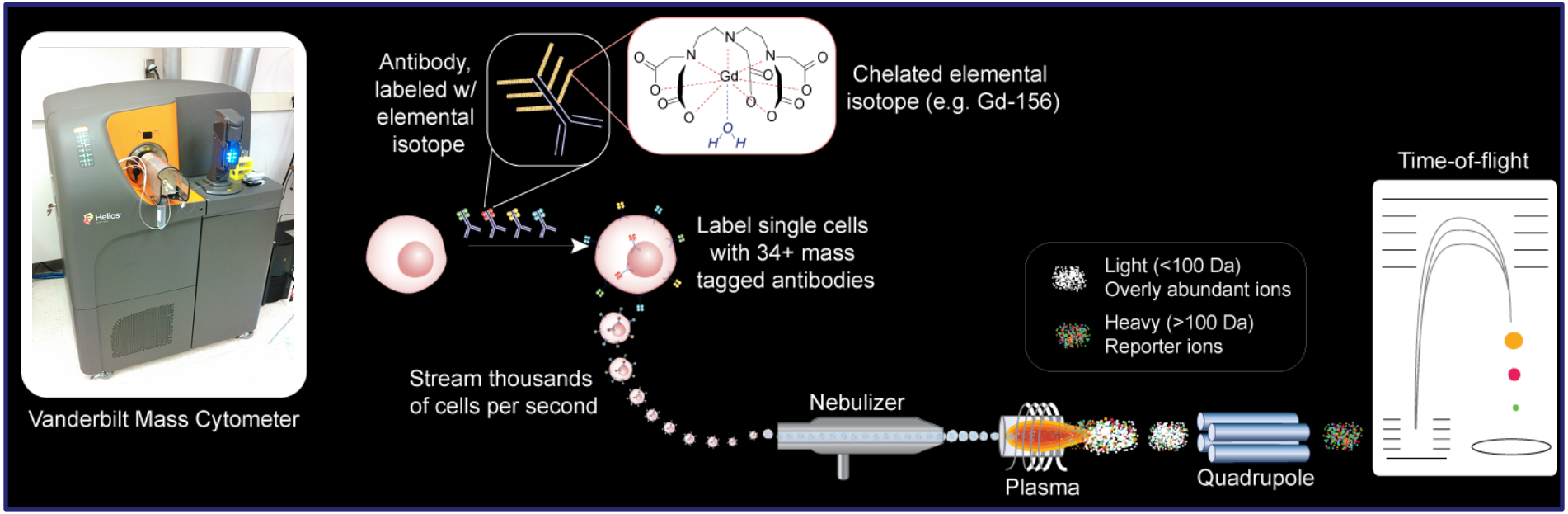
Probes with similar “peaks” are resolved



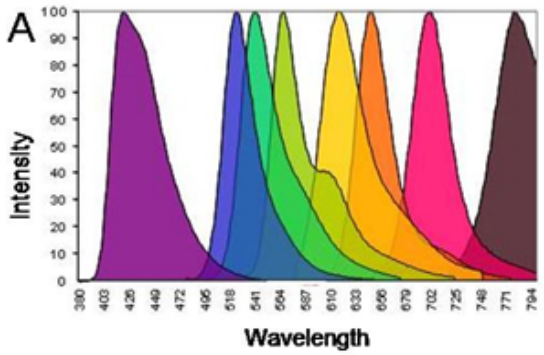


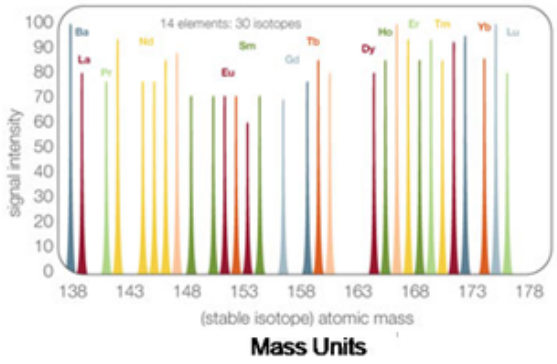
Conventional detector measures only peaks, uses dyes with distinct peaks

Mass Cytometry Uses Isotopically Pure Metals As Probes



Mass cytometry: standard panels detect 35+ features per cell using pre-validated antibodies. The dynamic range is >10,000 intensity units per feature. A small dataset might include 1.2 million cells from 12 samples collected at a rate of 500 cells/second (~40 min instrument time) for a total cost of ~$4,500 ($0.004 per cell), including reagents & personnel.





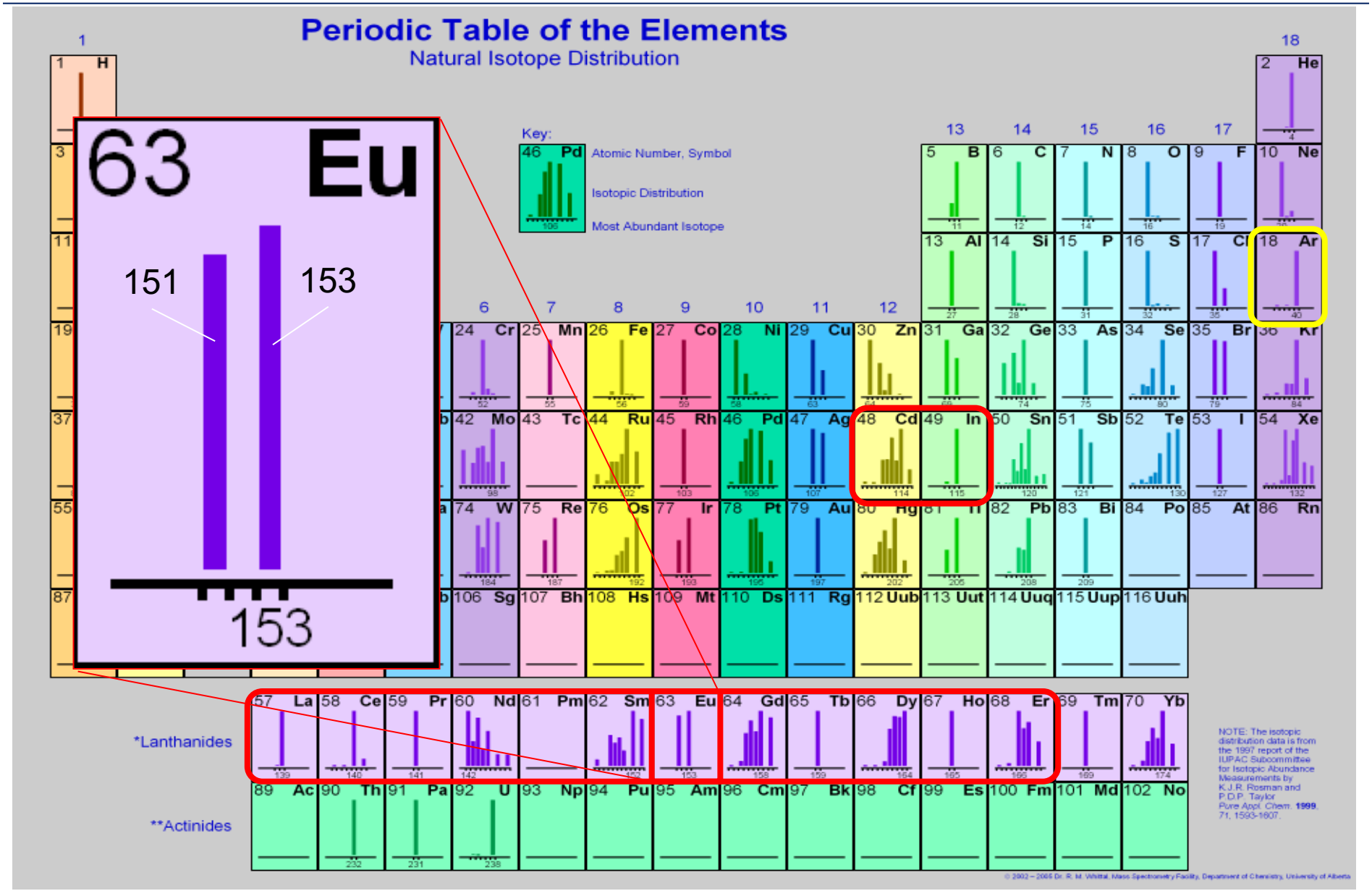
Experiment Protocols: Leelatian et al., *Methods in Molecular Biology* 2015
Leelatian et al., *Current Protocols in Molecular Biology* 2017

Analysis Protocols: Diggins et al., *Methods* 2015
Diggins et al., *Current Protocols in Cytometry* 2018

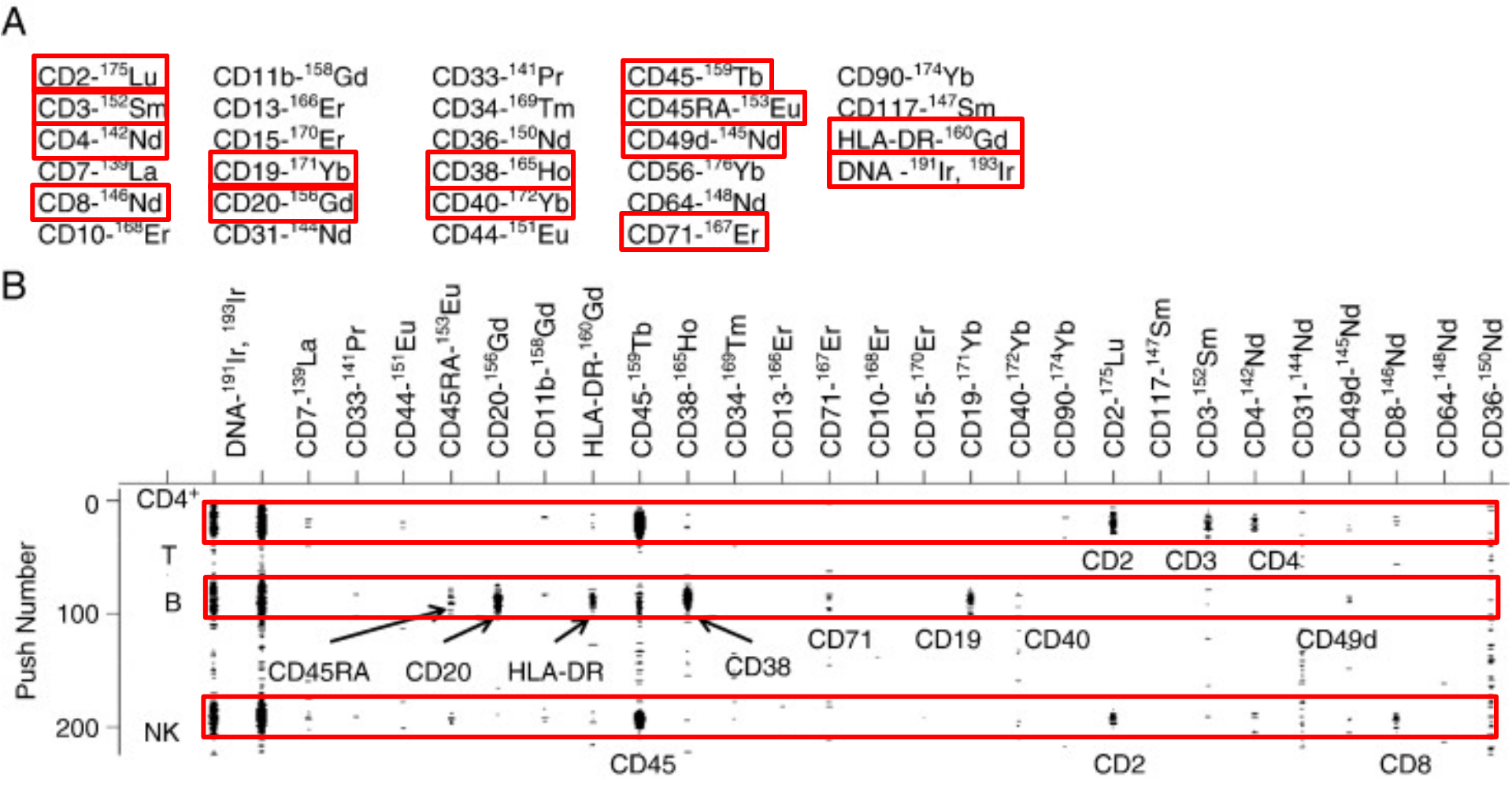
Reviewed in: Spitzer & Nolan, *Cell* 2016

Adapted from Bendall et al., *Science* 2011

Mass Cytometry Uses Isotopically Pure Metals As Probes

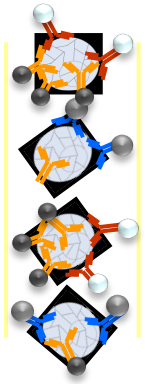


Mass Cytometry Uses Mass Spectrum, But Splits It Into Channels Because the Probes Overlap Very Little

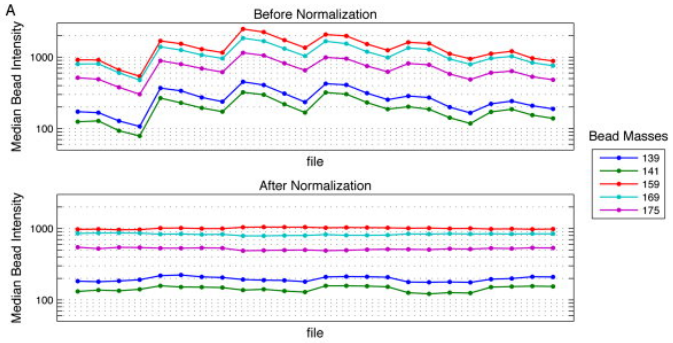


Modern Cytometry Includes Internal Batch Controls

Data Acquisition (IMD -> FCS) → Concatenation → Normalization → Debarcoding → Transformation → Analysis



FCS File 1
+
FCS File 2
+
FCS File 3



arcsinh(x/5)

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)

Mass Cytometry Balances Signal, Throughput, & Cost

Mass cytometry: Imaging or flow cytometry method of multiplexed single cell analysis. Standard mass cytometry panels detect 37+ features per cell using pre-validated antibodies. The dynamic range is >10,000 intensity units per feature and a small flow-based mass cytometry dataset might include 1.2 million cells from 12 samples collected at a rate of 500 cells/second (~40 min instrument time) for a total cost of ~$4,500 ($0.004 per cell), including personnel time/effort.

Diggins et al., *Bench to Bedside to Bytes, in review*
Mass cytometry vs. other single cell technologies: Mistry et al., *FEBS J* 2018

Some of the literature from "big seq" is hilariously inaccurate when it comes to flow cytometry…

From: "Single cell RNA sequencing to explore immune cell heterogeneity", Nat Rev Immunology 2017

	FACS	CyTOF	qPCR	Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2)	Fluidigm C1	Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2)	Massively parallel approaches (Drop-seq, InDrop)
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500	48–96	500–2,000	5,000–10,000
Cost	$0.05 per cell	$35 per cell	$1 per cell	$3–6 per well	$35 per cell	$3–6 per well	$0.05 per cell
Sensitivity	Up to 17	Up to 40	10–30 genes	7,000–10,000 genes	6,000–9,00 genes for cell	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells

Wait, so this review says each experiment costs… ?!
FACS: $50,000 ($0.05 x 1,000,000 cells)
CyTOF: $35,000,000 ($35.00 x 1,000,000 cells)

mass cytometry); FACS, fluorescence-activated cell sorting; CR, quantitative PCR; SCRB-seq, single-cell RNA barcoding and sequencing; STRT-seq, single-cell tagged reverse transcription sequencing.

Mass Cytometry Dissects Cellular Mechanisms of Cancer Immune Response

Article

Cell

Systemic Immunity Is Required for Effective Cancer Immunotherapy

Spitzer et al., *Cell* 2017

Uses mass cytometry to characterize essential role of peripheral blood $CD4^+$ T cells in immunotherapy response

ARTICLE

doi:10.1038/nature22079

T-cell invigoration to tumour burden ratio associated with anti-PD-1 response

Huang et al., *Nature* 2017

Uses mass cytometry to reveal peripheral blood CD8 T cells associated with anti-PD-1 immunotherapy responses

Article

Cell

Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade

Wei et al., *Cell* 2017

Uses mass cytometry to characterize similar & distinct tumor-infiltrating immune cell subsets (mostly T cells) following immunotherapies

Cancer Immunology Is Powered by Cytometry

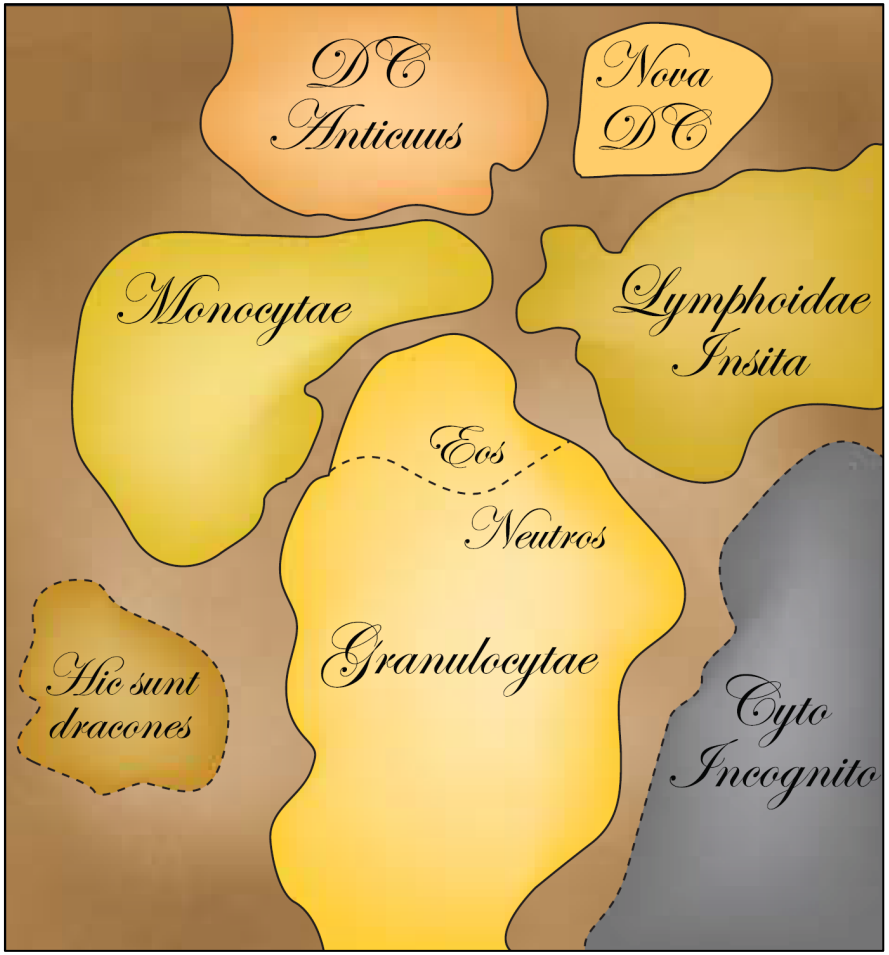
Dr. Allison, 70, is chairman of immunology at the University of Texas M.D. Anderson Cancer Center. He did the work recognized by the Nobel committee while working the University of California at Berkeley and Memorial Sloan Kettering Cancer Center in New York.

Dr. Honjo, 76, is a longtime professor at Kyoto University, where he did his landmark work. Previously, he did research at Osaka University, the University of Tokyo and the National Institutes of Health in Washington.

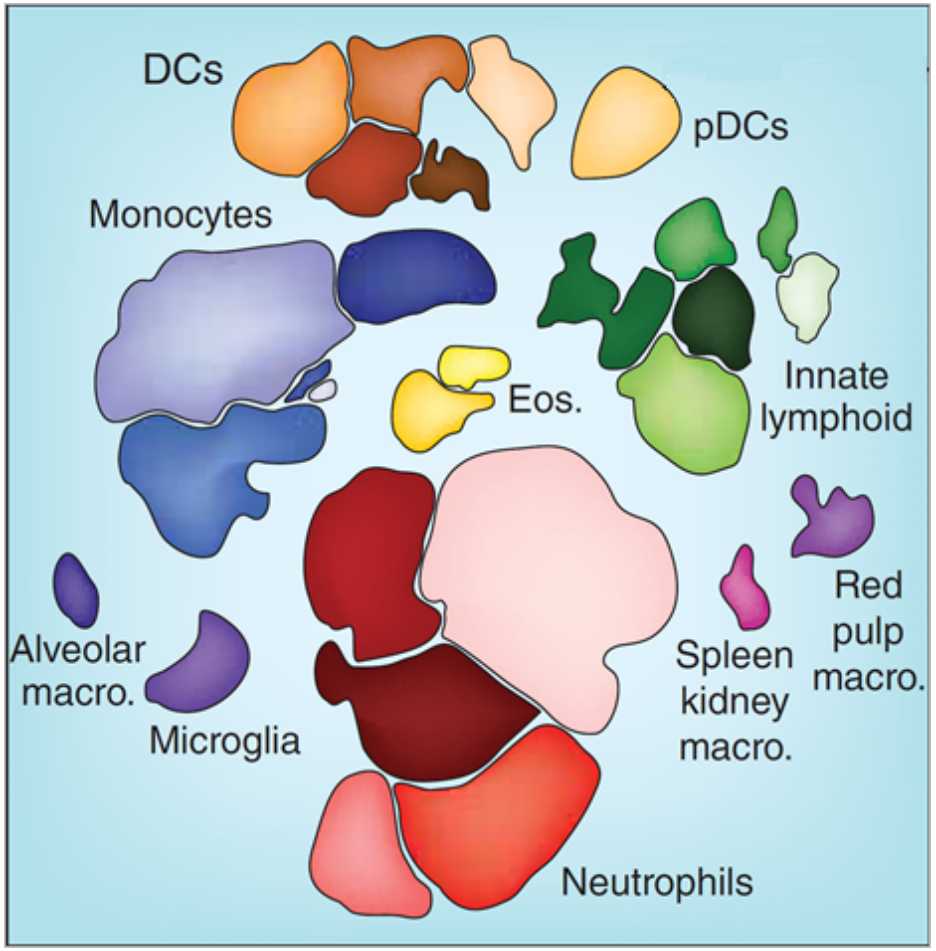
The Nobel Prize for Physiology and Medicine was awarded to James P. Allison, left, and Tasuku Honjo on Monday for their work on cancer research. Credit Jonathan Nackstrand / Agence France-Presse — Getty Images; New York Times 2018

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

Classical map of the 'myeloid cell system'

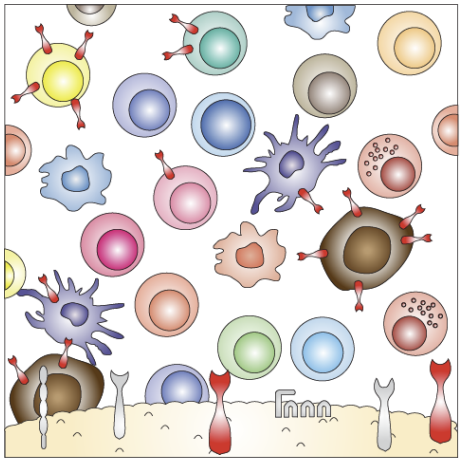
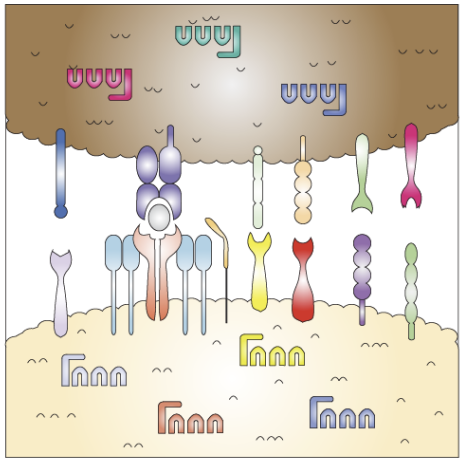
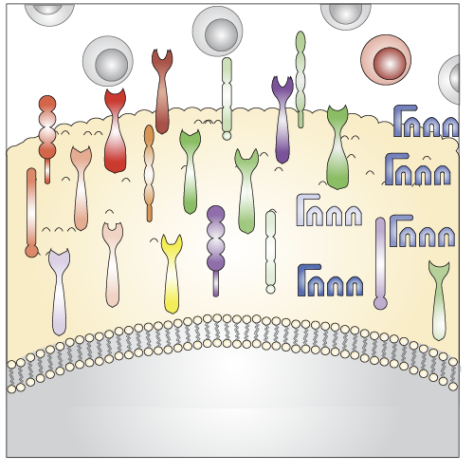
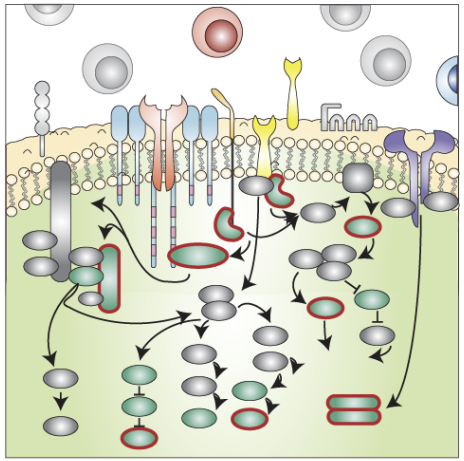


Modern map, computationally generated



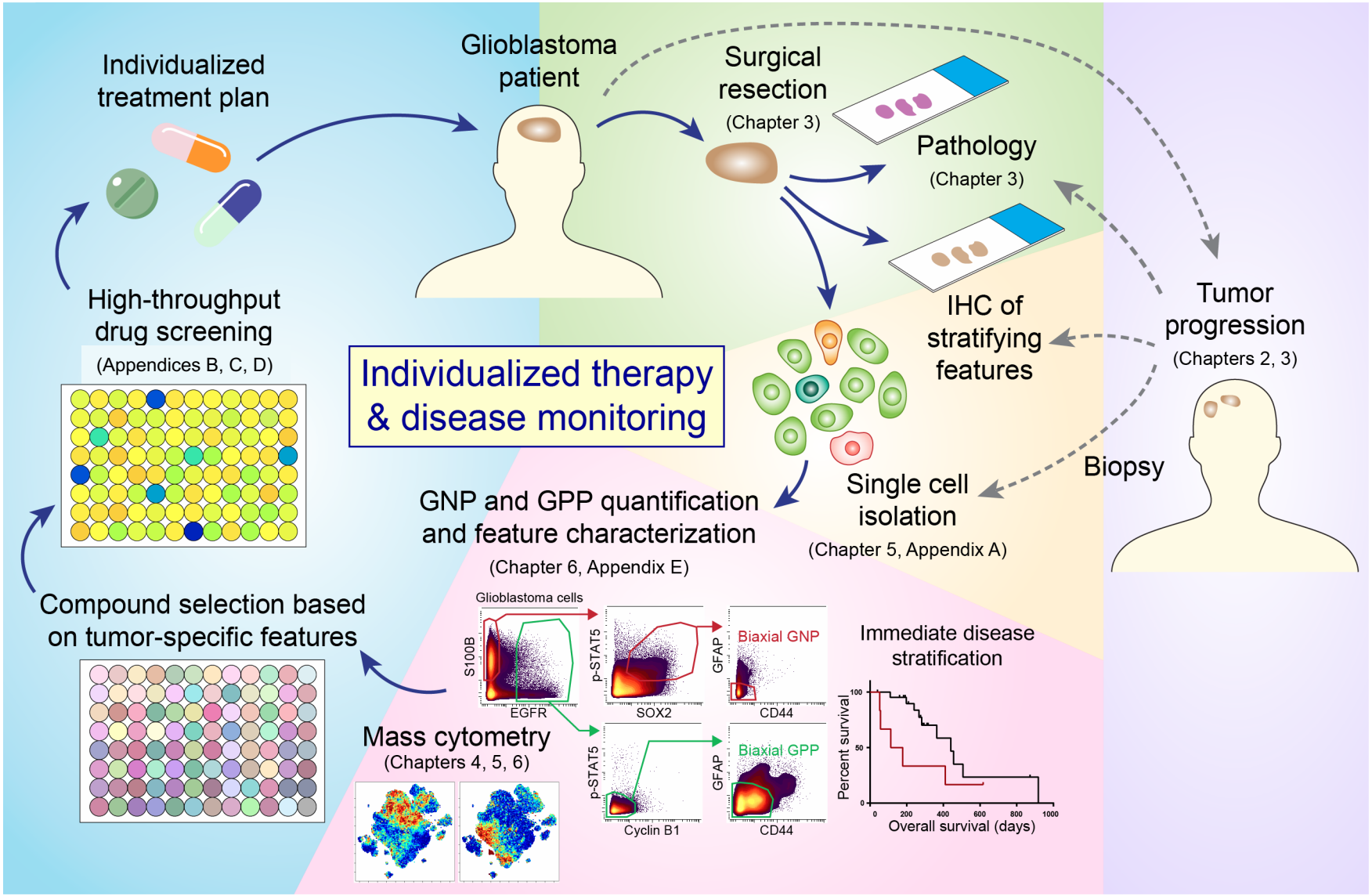
Irish, *Nature Immunology* 2014
Based on Becher et al., *Nature Immunology* 2014

Applications of Mass Cytometry in Cancer Biology

Microenvironment	Cell:cell interactions	Immunophenotype	Signaling & function
			
Track key biomarkers on all cell types; find cytokine producers	Monitor immune checkpoint proteins on T cells, APCs, & cancer cells	Measure differentiation; deep phenotype using fewer cells	Dissect signaling changes; characterize mechanisms of treatment response

Greenplate et al., *Euro J Cancer* 2016

Future Clinical & Research Workflow Pictured by Nalin Leelatian



Nalin Leelatian, MD PhD (Dissertation)

Central Questions & Goals of Systems Biology

1. How do we monitor dynamic cellular systems, track fates, reveal new cells, & dissect mechanisms in humans?

2. What molecules confer & maintain cell identity?
   Can we pinpoint cells and shift their functional identity?

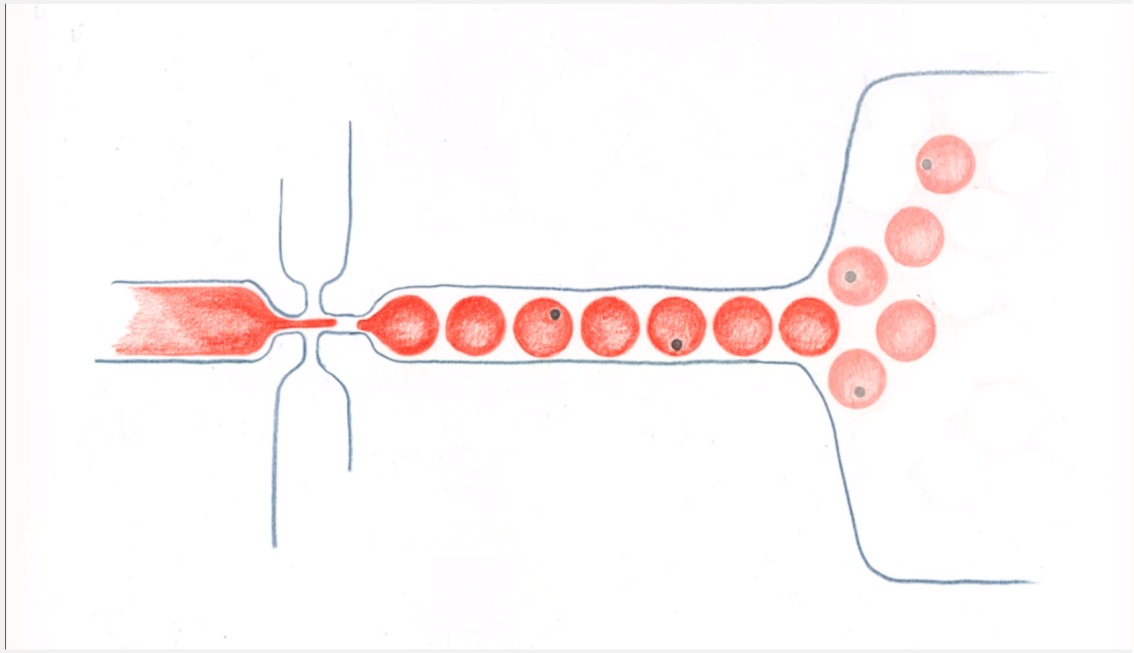
3. Do computers see patterns in the data that are useful?
   When is supervised learning useful vs. misleading?
   Can we build analysis tools that handle novelty well?

How and what you choose to measure
is critical to project success,
and every technology has limits and biases

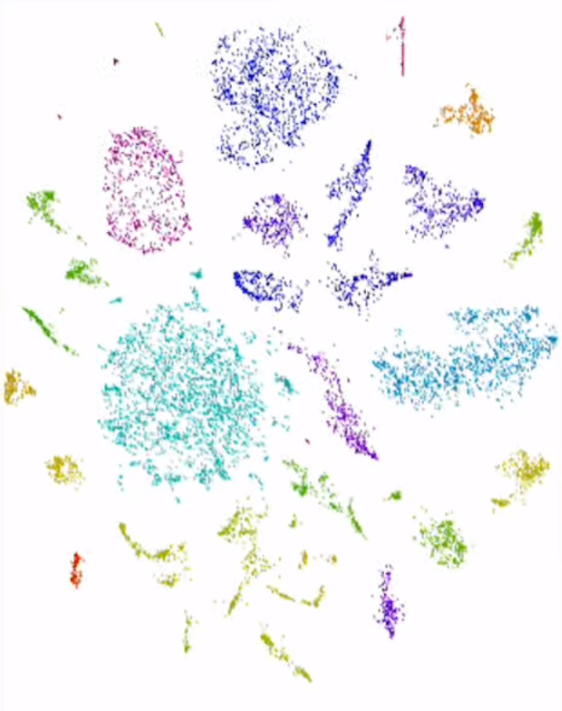
Gather as much information as possible.
Later, you can choose what to use.

Feature selection = data science hypothesis

t-SNE on RNA-seq (DropSEQ data)



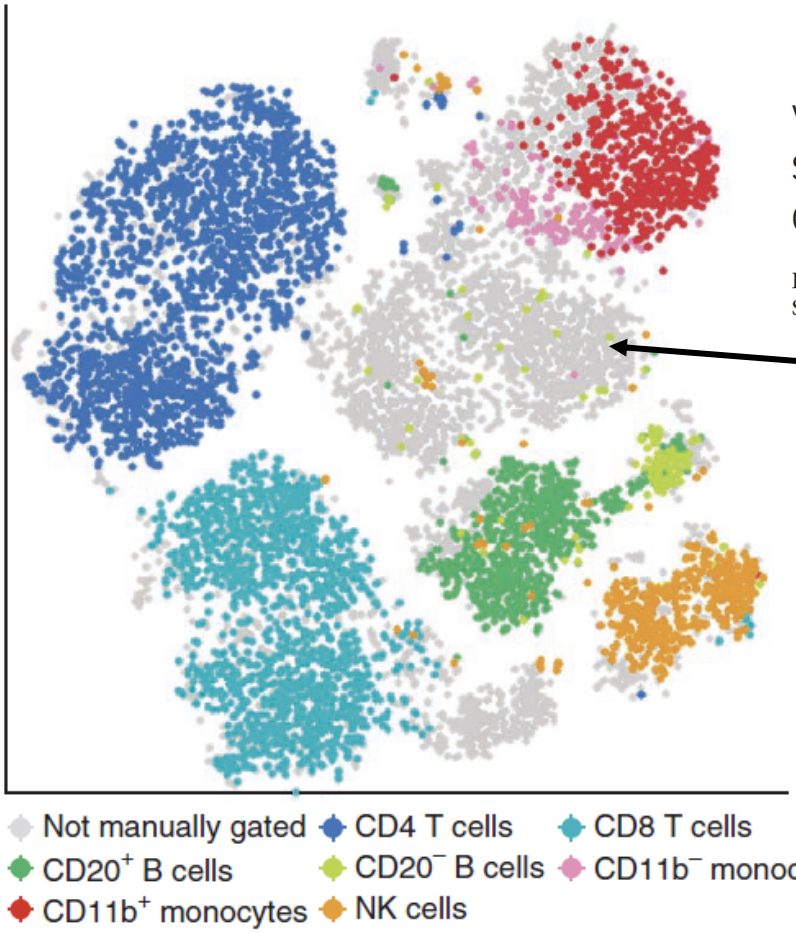
Smoothie vs. Fruit Salad/ Cell, May 21, 2015 (Vol. 161, Issue 5)



https://www.youtube.com/watch?v=XAsmHKfKHmc

About 4 min in, see an animation of t-SNE

t-SNE can Help to Identify Cells Otherwise Lost by Expert Identification



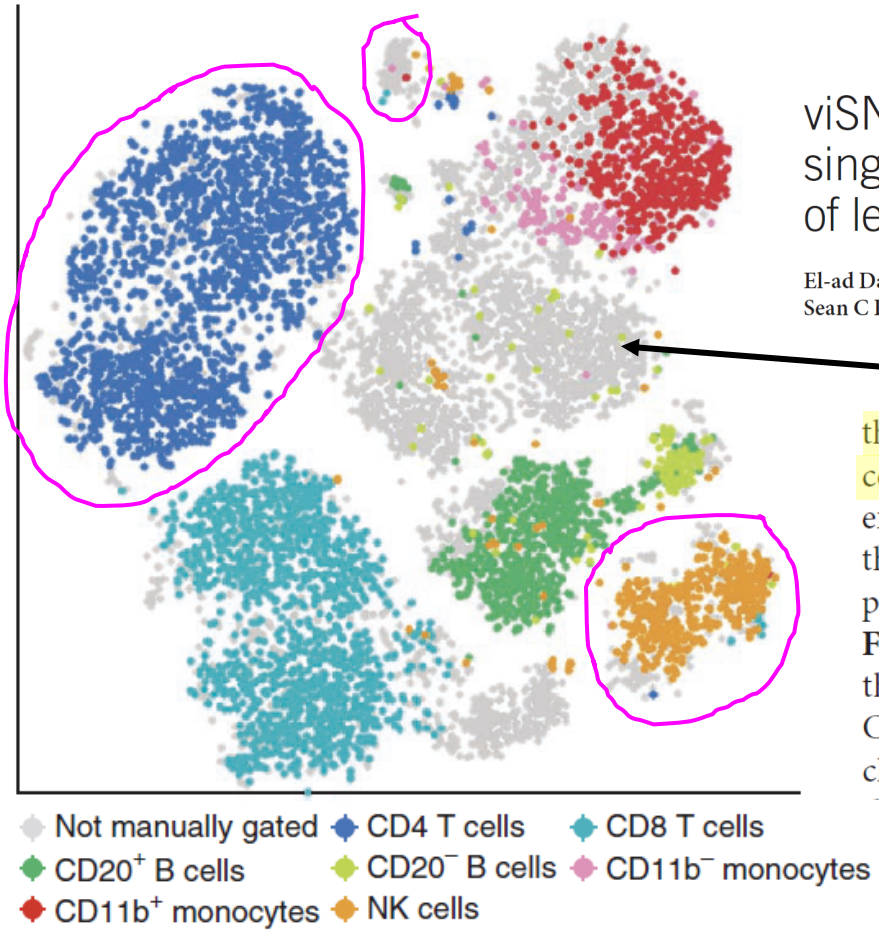
viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

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

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 based 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**).

Amir et al., *Nature Biotechnology* 2013

Experts can use t-SNE Axes to Select Cells of Interest



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

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

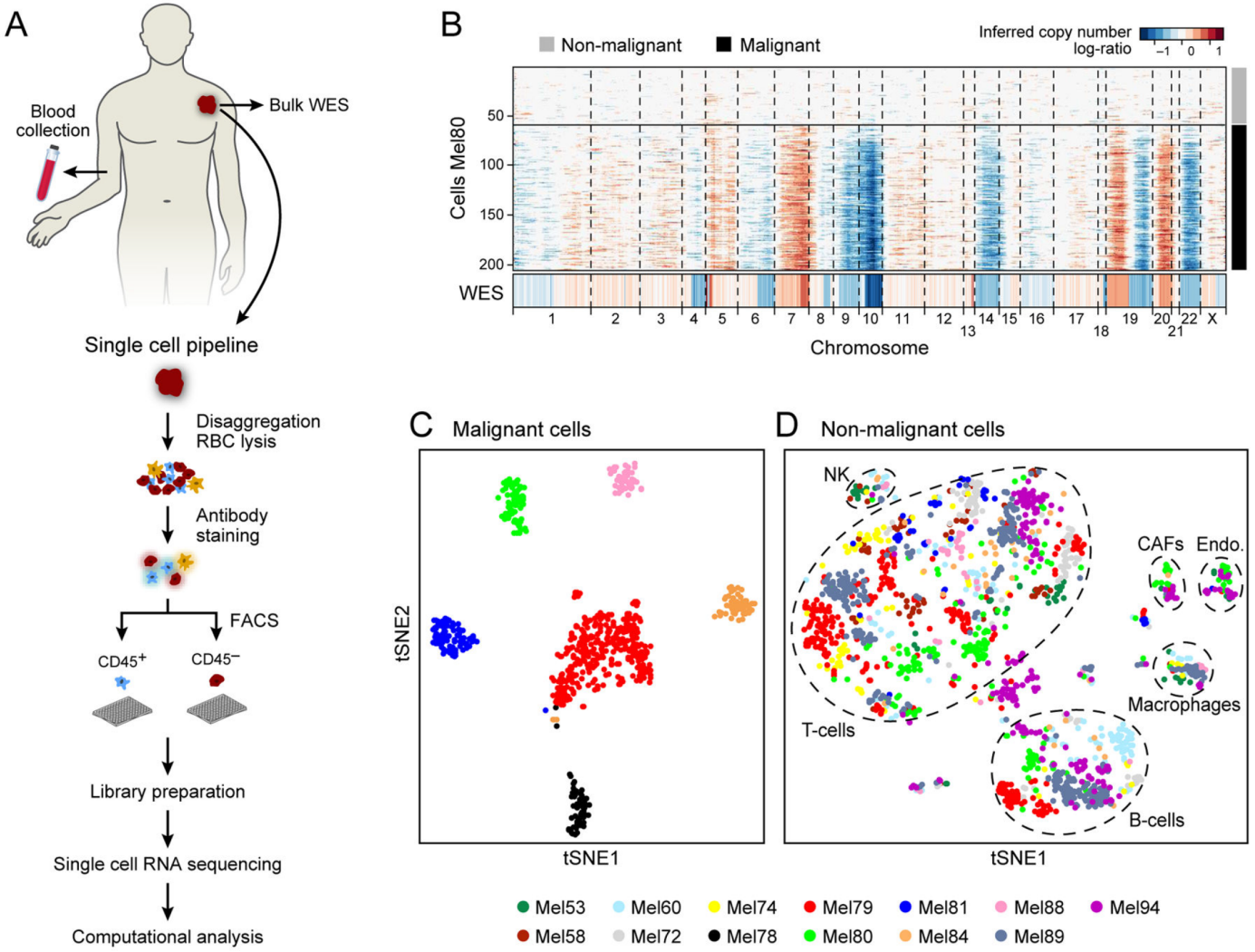
In all cases, the viSNE gate included cells that were not classified by the expert manually gated biaxial plots; these cells are labeled in gray in the viSNE map. Examination of the marker expression of these cells reveals that they are typically just beyond the threshold of one marker, but the viSNE classification is strongly supported based 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**).

Amir et al., *Nature Biotechnology* 2013

How Does Abundance / Density on a t-SNE Map Relate to Cell Identity?

Irish lab, unpublished

t-SNE Has Been Used Extensively in scRNA-seq (Can Reveal Issues with Data)



Tirosh et al., *Science* 2016