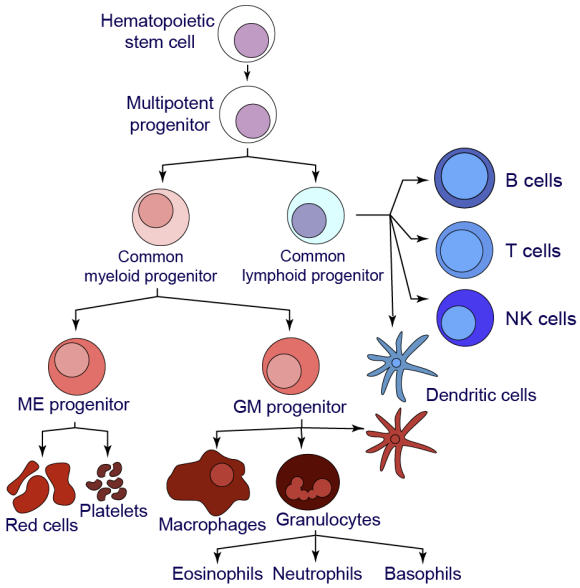
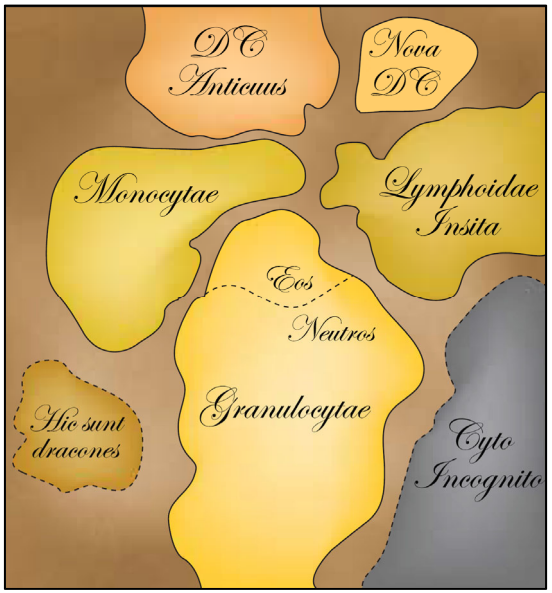
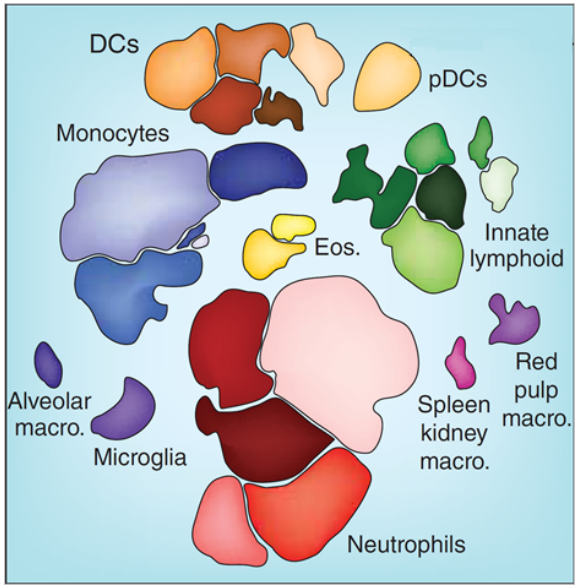
Combining Tools in a Modular Mass Cytometry Bioinformatics and Data Analysis Workflow







Diggins et al., *Methods* 2015

Mass Cytometry Training Course, London 2019, Day 4

Jonathan Irish, Ph.D.
Assistant Professor of Cell & Developmental Biology
Pathology, Microbiology & Immunology

Disclosures:
Co-founder & board at Cytobank Inc.
Mass Cytometry Center of Excellence w/ Fluidigm
Clinical research w/ Incyte, Janssen, Pharmacyclics

Key Topic Areas and Terms for This Course

1: Field Changes: Data Science & Latest Tools

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

3: Dimensionality Reduction: t-SNE, UMAP, FIt-SNE

4: Clustering: SPADE, KNN, FlowSOM, Citrus

5: Enriched Features: MEM, ΔMEM, RMSD

6: Cytometry:
- 2004: [Anything] => Heatmap (Irish/Nolan)
- 2011: SPADE => [Anything] (Bendall/Qiu)
- 2013: t-SNE => [Anything] (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 => [Anything] (Van Gassen/Saeys)
- 2017: [Anything] => MEM (Diggins/Irish)
- 2018: UMAP => [Anything] (Newell, McInnes)
- 2019: UMAP => FlowSOM => MEM (Barone/Irish)

Discussion Questions Covered in This Course

1) What are key differences between tools (t-SNE/viSNE, SIMLR, SPADE, UMAP, FlowSOM, PCA, MEM, Citrus, FIt-SNE)? 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)?

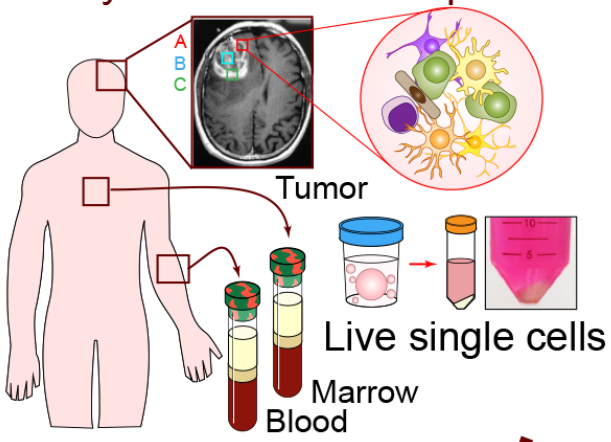
3) What do all the settings do (e.g., t-SNE iterations, perplexity, SPADE downsampling & node #)? When should they be changed?

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

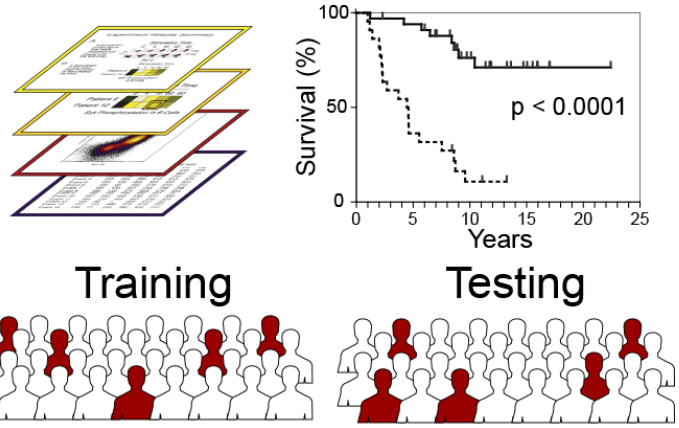
5) What are some "red flags" indicating problems? What does a good viSNE or SPADE analysis run look like?

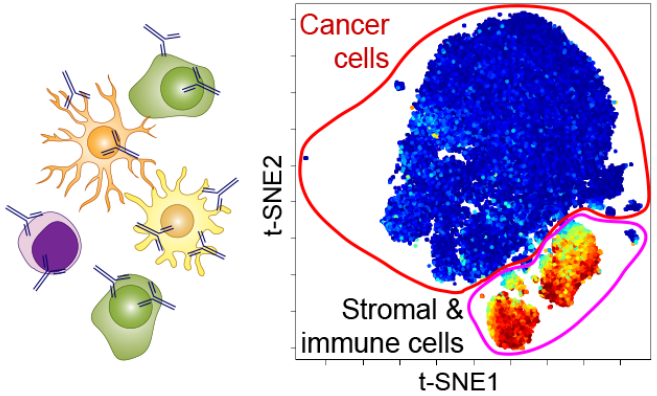
Elements of Translational Data Science

Sample patients over time
at key clinical decision points

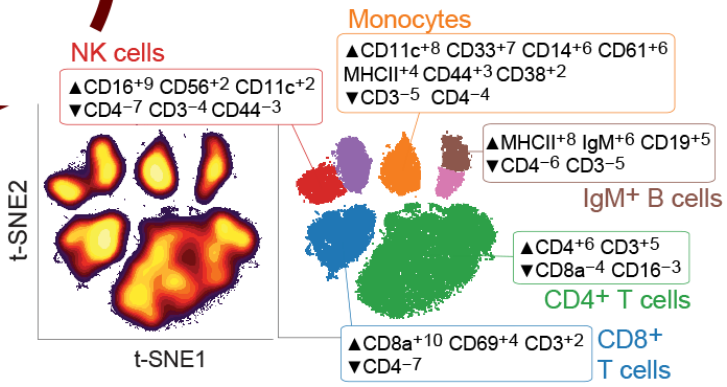


Integrate systems biology data
& clinical outcomes to guide treatment





Measure cell identity, signaling,
biomarkers, and functional responses



Apply machine learning tools
to reveal patterns & identify groups

Diggins et al., *in review*

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?

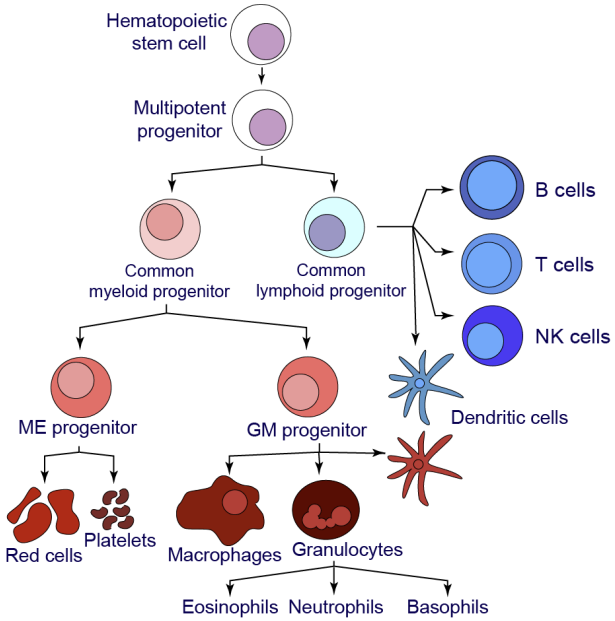
Unknown unknowns: What don't you know you don't know?

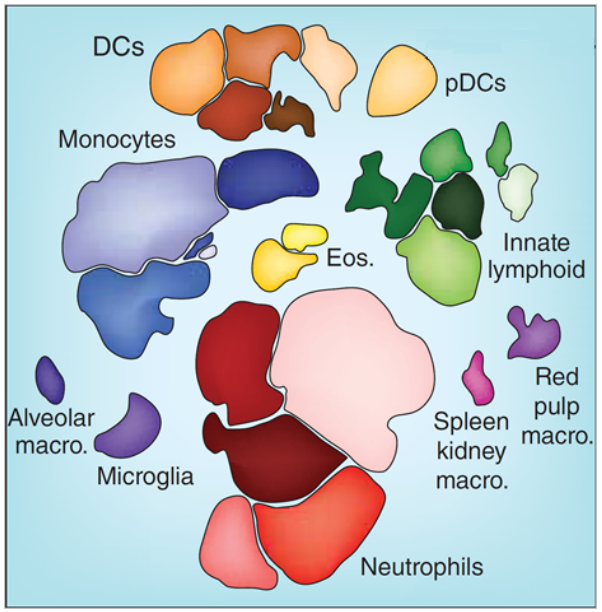
Unknown knowns: What don't you know, but think you do?
i.e. Which 'priors' are incorrect?

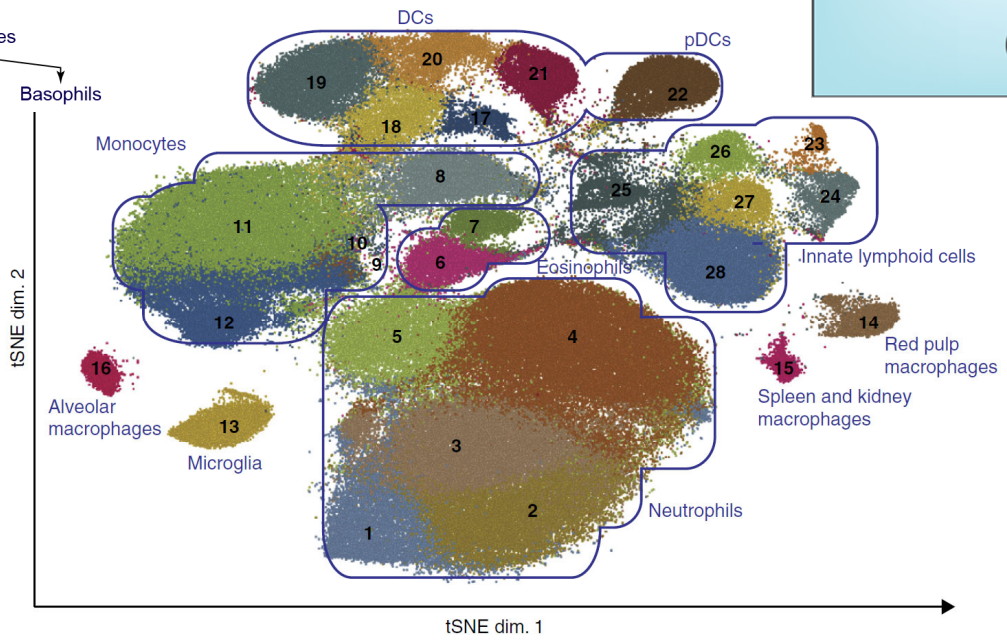
If you fear incorrect priors, unsupervised analysis may be able to help.

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.

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





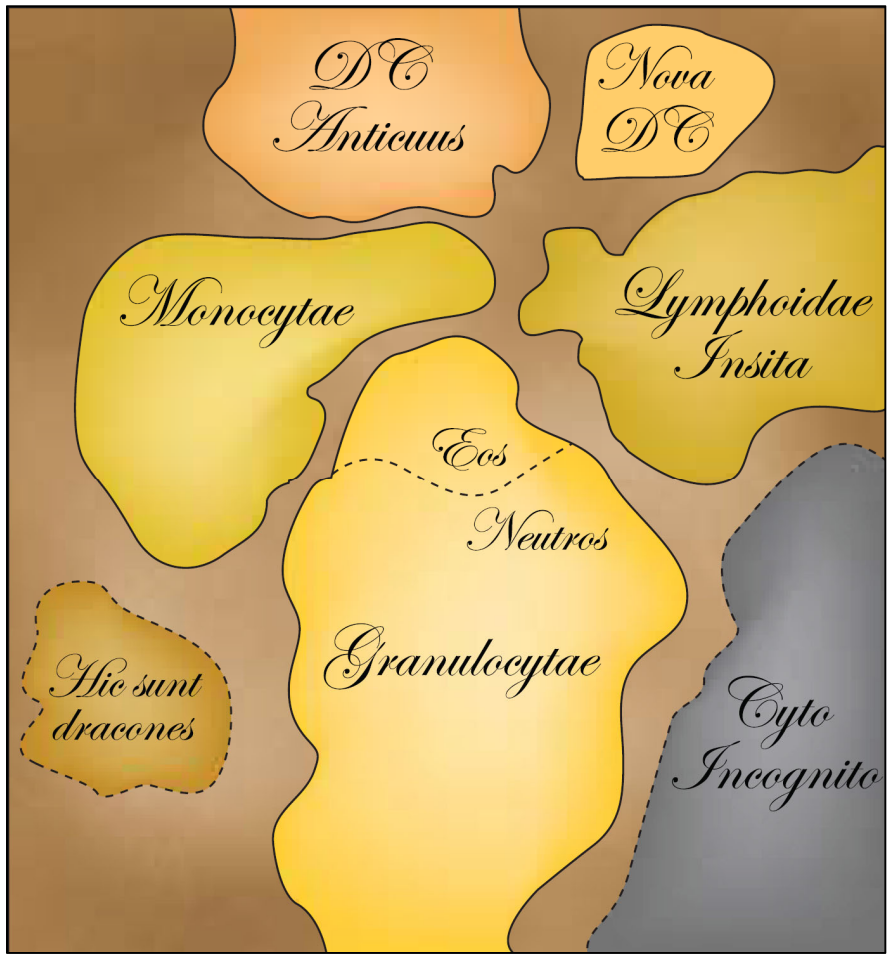


Irish, *Nature Immunology* 2014

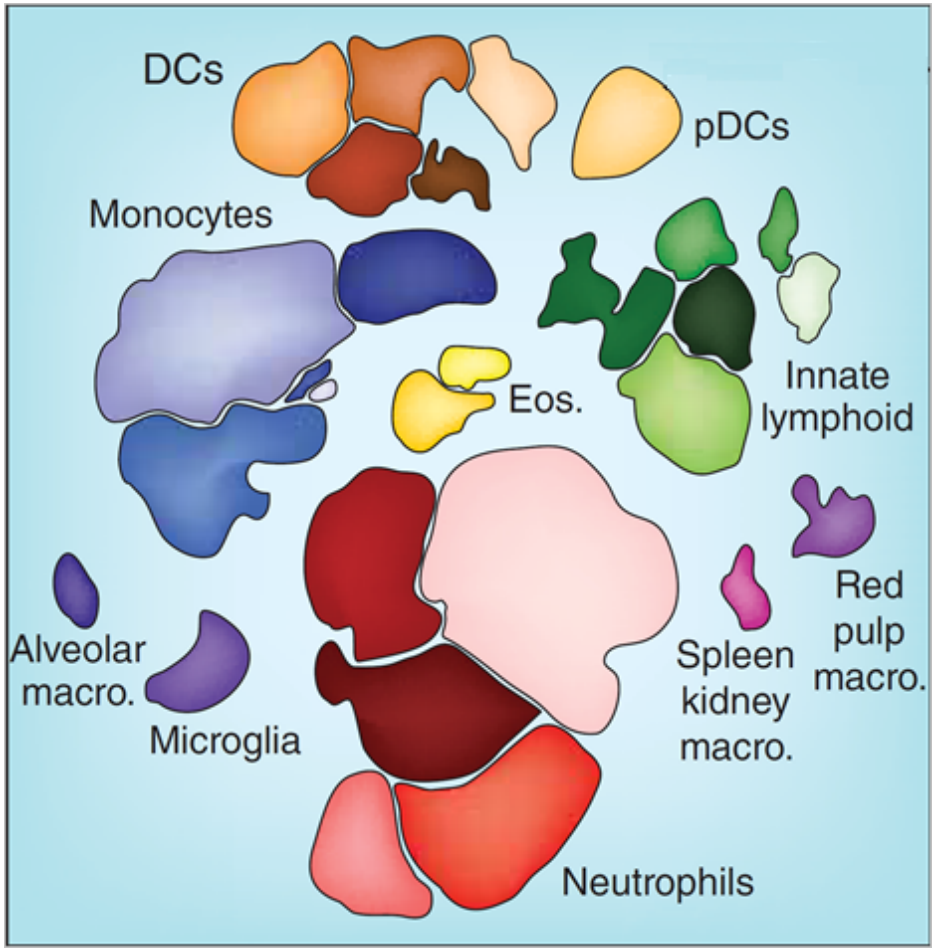
Based on Becher et al., *Nature Immunology* 2014

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

Classical map of the ‘myeloid cell system’

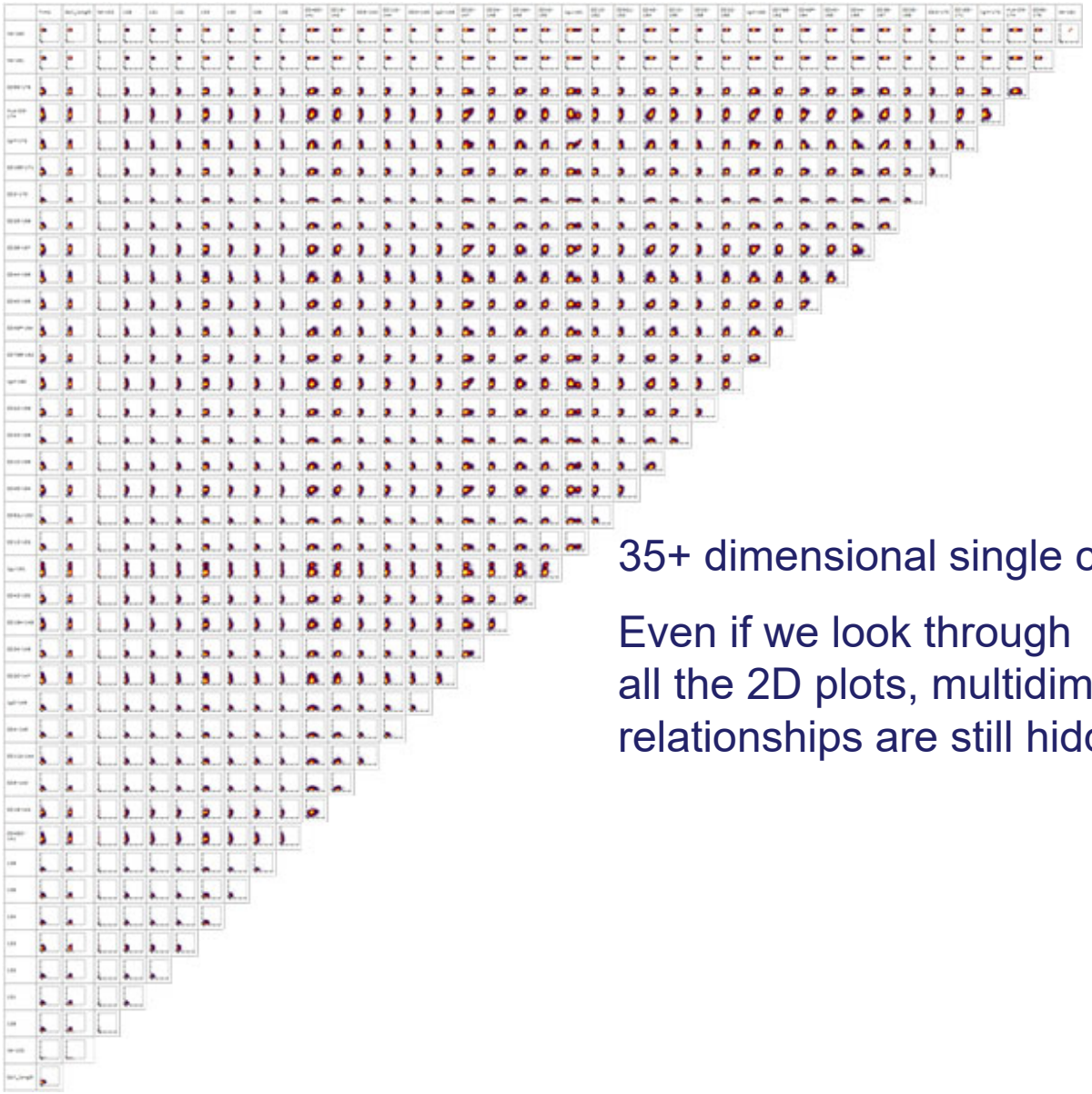


Modern map, computationally generated



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

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…

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

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.

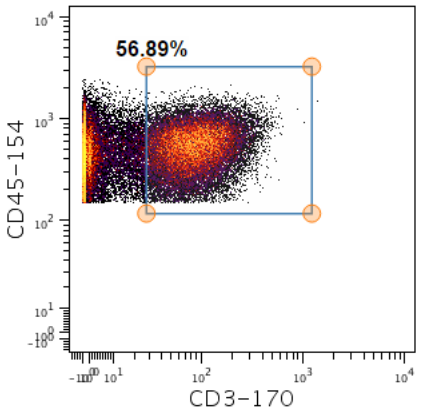
- Organization of elements (groups, order, etc.), feature relationships.
- Supervised analysis explicitly uses prior knowledge.
- Unsupervised analysis looks for patterns without prior knowledge.

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

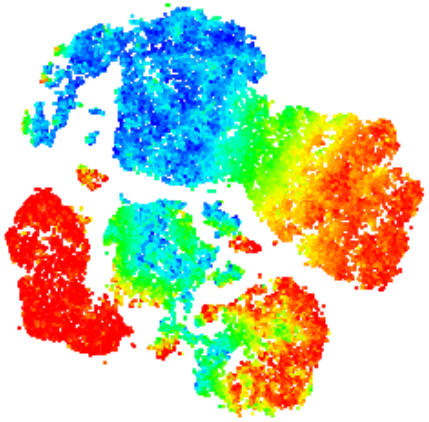
Supervised Approaches

- Expert gating
- Gemstone
- Wanderlust
- Citrus



Unsupervised Approaches

- Heatmap clustering
- SPADE
- viSNE
- Phenograph



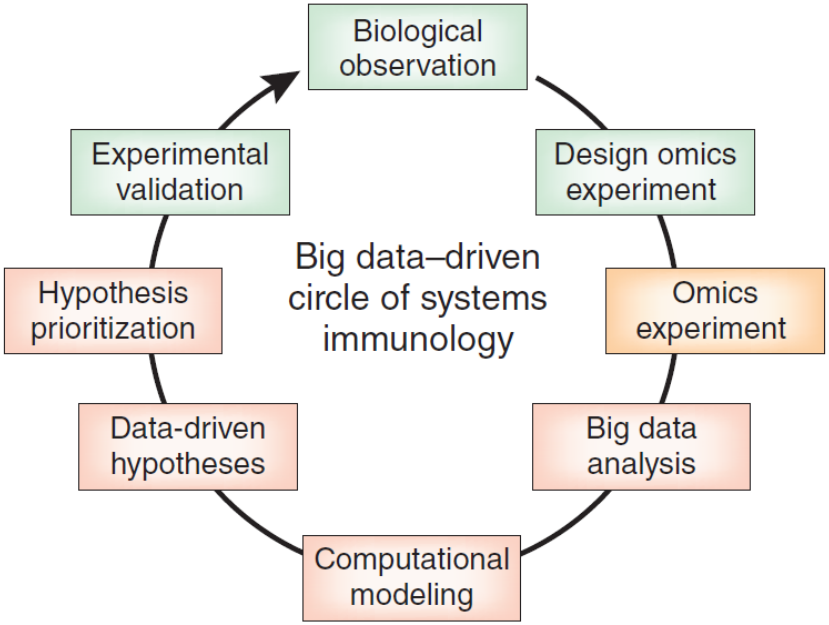
See Table 1 of Diggins et al., *Methods* 2015 for list of unsupervised tools

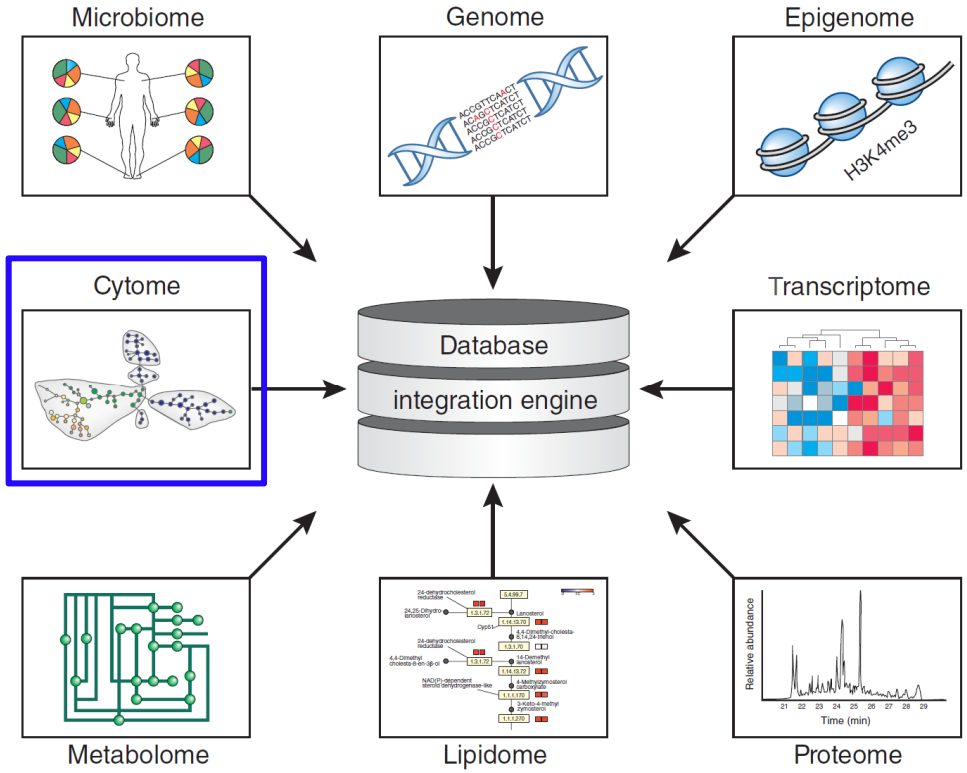
Cytomics: The 'Omics of Cells & Cell Identity

Teaching 'big data' analysis to young immunologists

NATURE IMMUNOLOGY
VOLUME 16 NUMBER 9 SEPTEMBER 2015

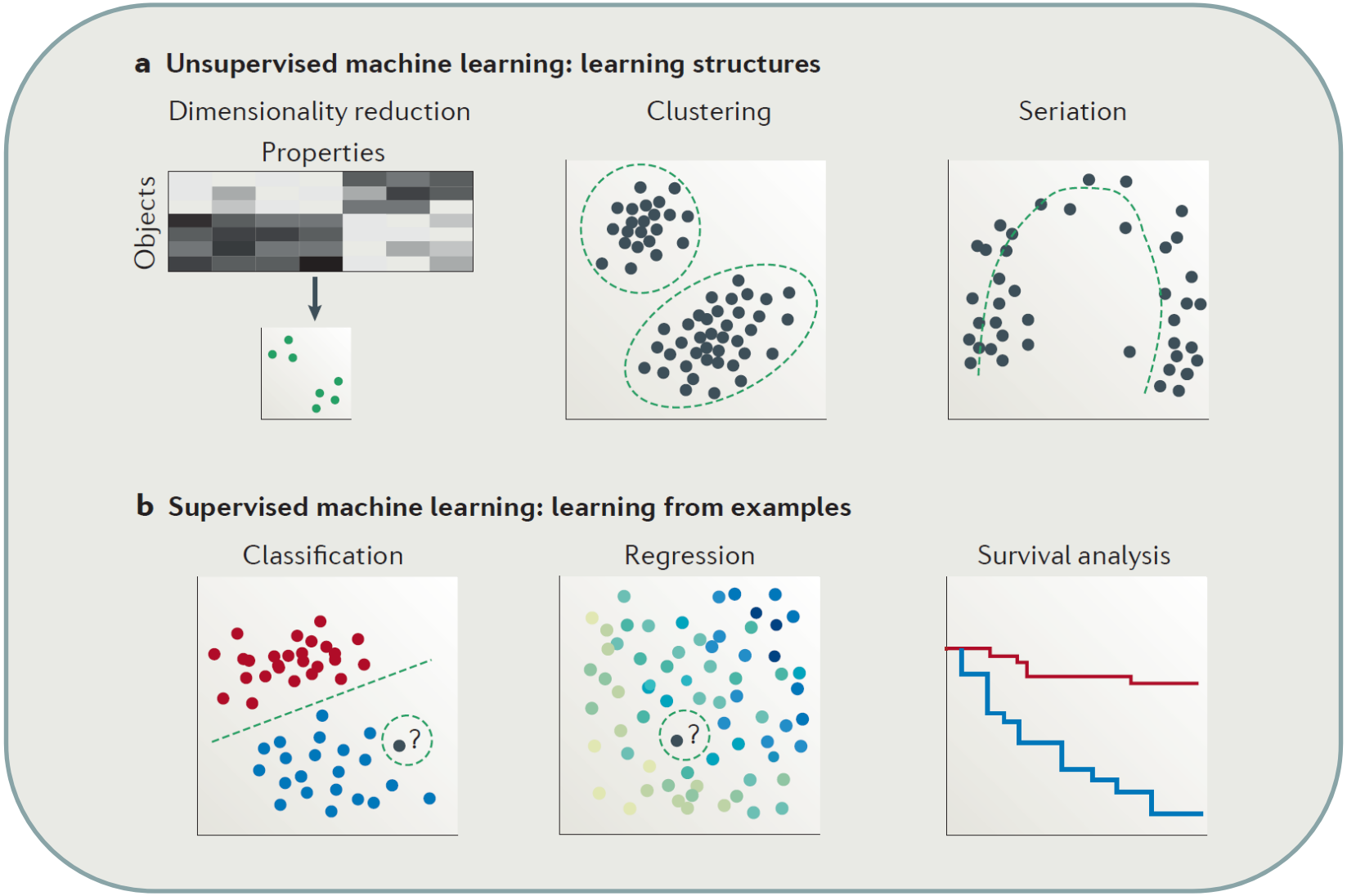
Joachim L Schultze





Schultze JL, *Nature Immunology* 2015

Computational Flow Cytometry: There Is Help for Data Analysis



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

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

Key Topic Areas and Terms for This Course

1: Field Changes: Data Science & Latest Tools

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

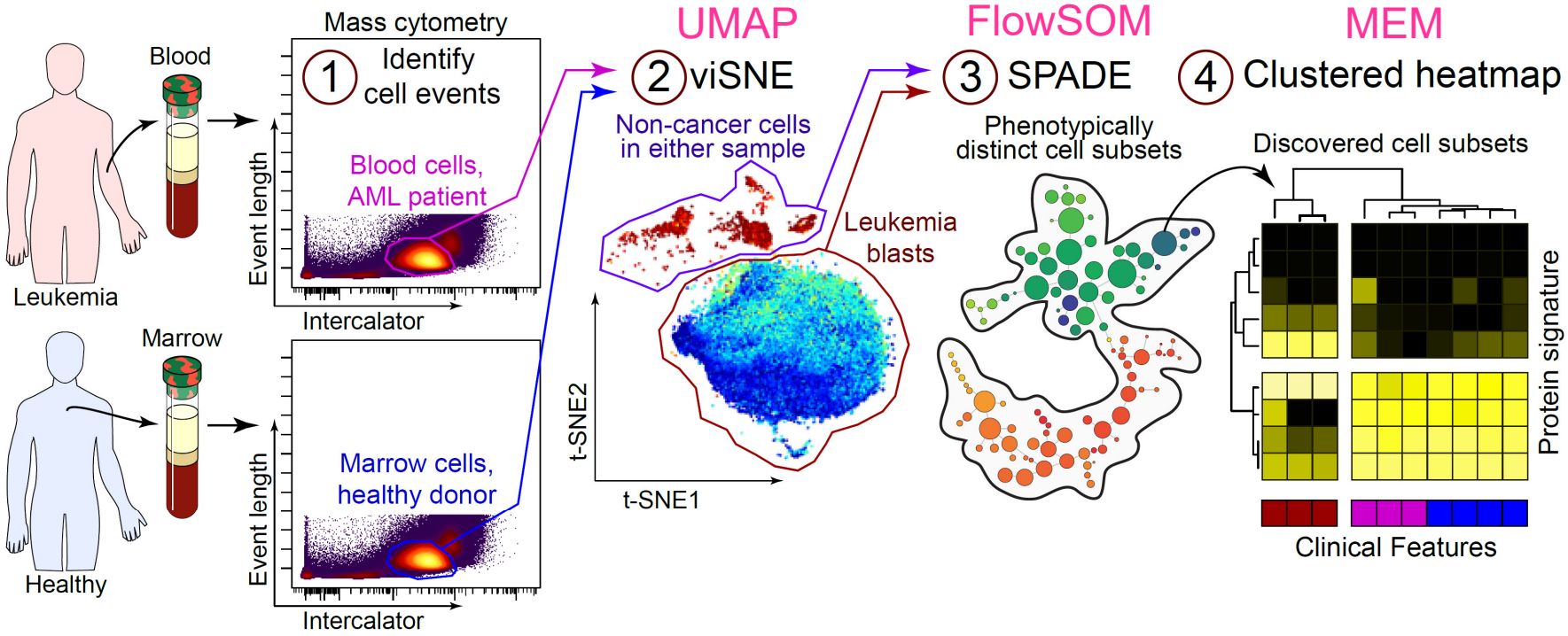
3: Dimensionality Reduction: t-SNE, UMAP, FIt-SNE

4: Clustering: SPADE, KNN, FlowSOM, Citrus

5: Enriched Features: MEM, ΔMEM, RMSD

6: Cytometry:
- 2004: [Anything] => Heatmap (Irish/Nolan)
- 2011: SPADE => [Anything] (Bendall/Qiu)
- 2013: t-SNE => [Anything] (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 => [Anything] (Van Gassen/Saeys)
- 2017: [Anything] => MEM (Diggins/Irish)
- 2018: UMAP => [Anything] (Newell, McInnes)
- 2019: UMAP => FlowSOM => MEM (Barone/Irish)

Machine Learning Cell Identity Is Now Possible (And 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

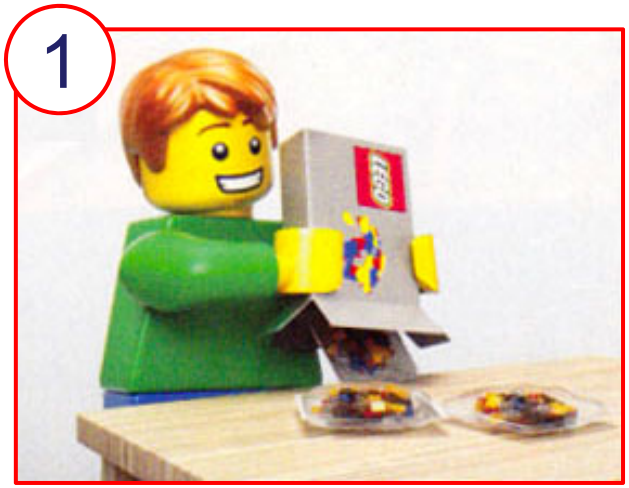
Diggins et al., *Methods* 2015
Diggins et al., *Nature Methods* 2017

Goal is to create computational tools that learn & label cytotypes and can cope with unexpected, abnormal cells

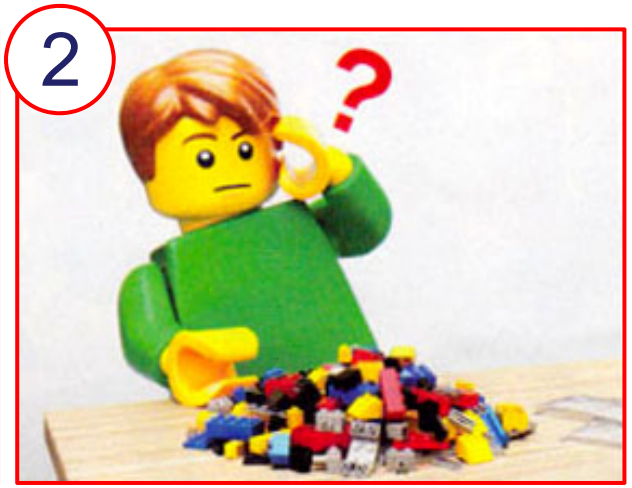
Need: human reference data (more examples) with annotations

t-SNE was a game changer for single cell

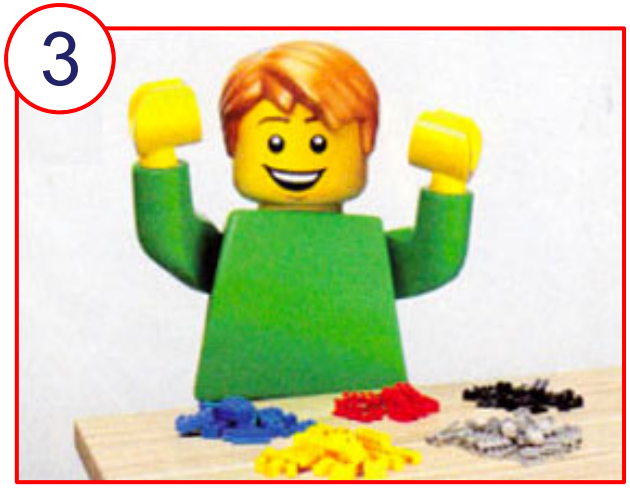
Teaching Computers To Spot Useful Patterns : Grouping Cells by Selected Features (e.g. Protein Expression)



HD cytometry!!



Woah, that's a lot of data…

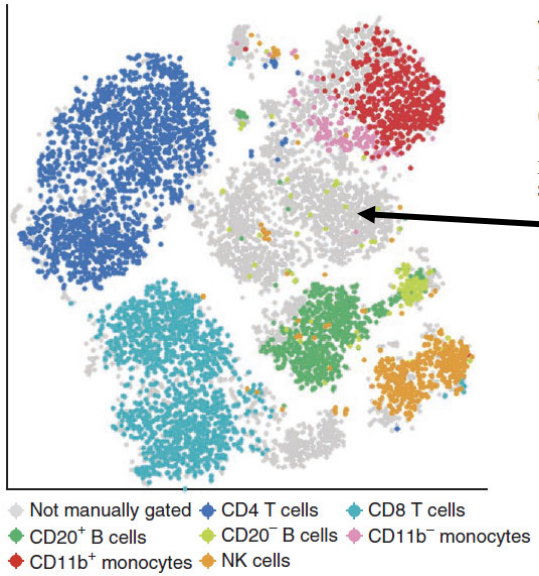


Computational tools

Biological knowledge

Adapted from LEGO 70803 Cloud Cuckoo Palace instructions 2014

Traditional Gating Overlooks Many Cells in Primary Samples



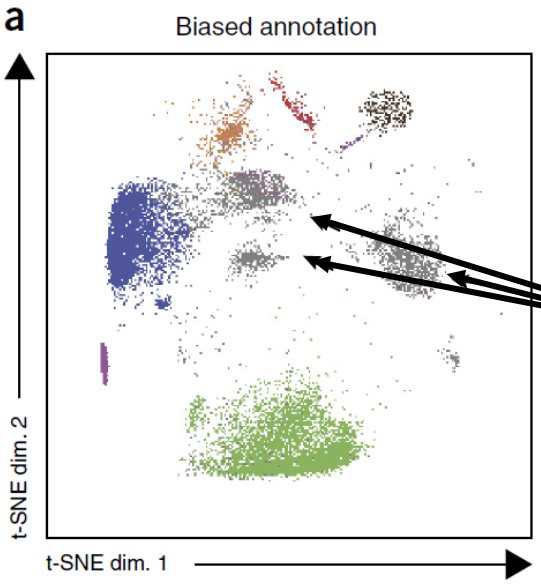
viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

nature biotechnology

2013

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**).



High-dimensional analysis of the murine myeloid cell system

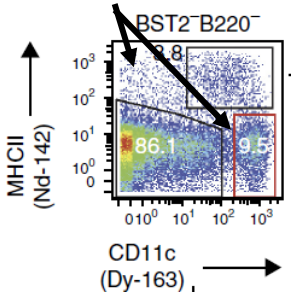
nature immunology

2014

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]

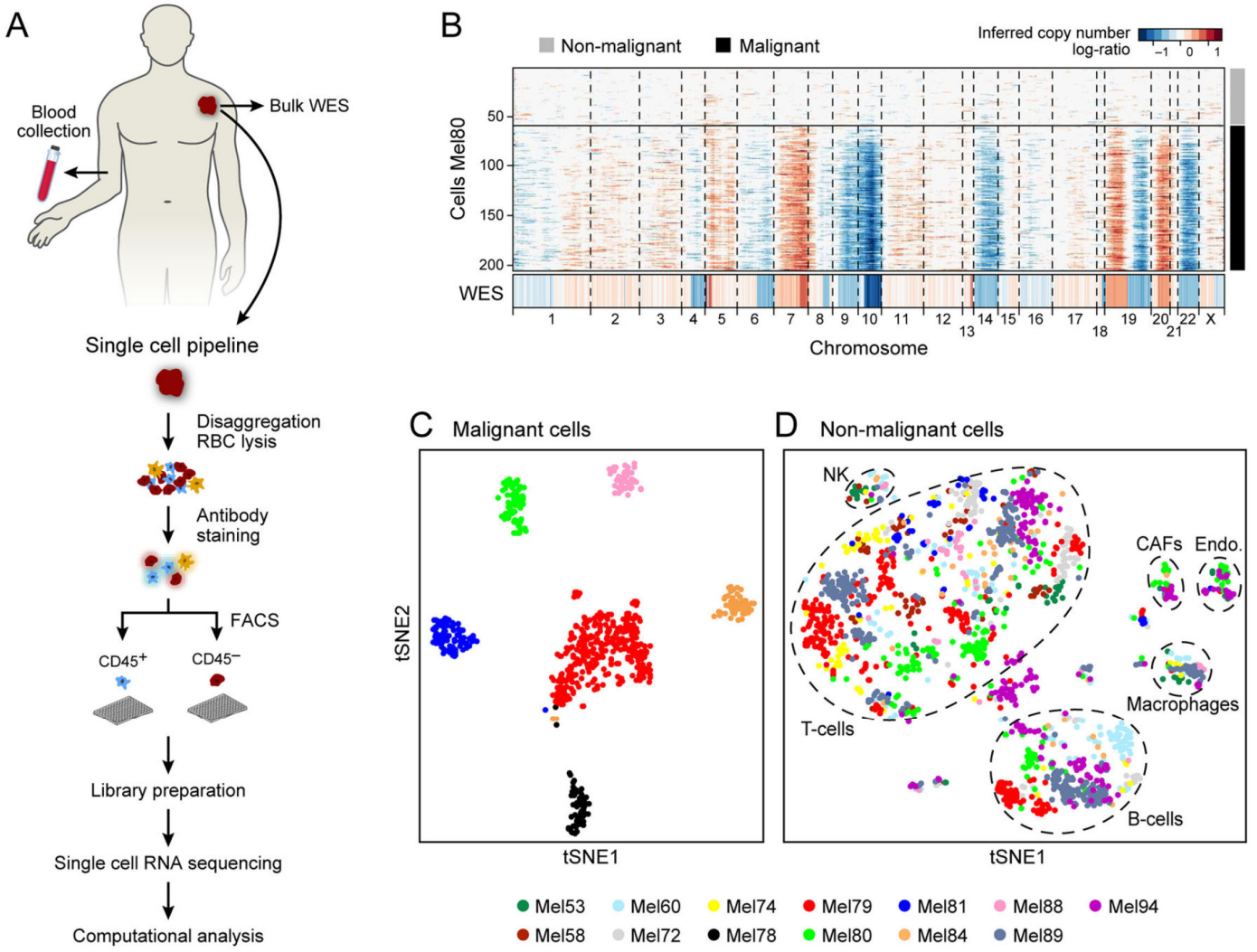
Notably, whereas traditional biased gating strategies allowed for identification of only 54.7 ± 2.6% (mean ± s.e.m., n = 3 mice) of lung myeloid cells (different DC subsets, macrophages, monocytes, neutrophils), the automatic, computational approach identified nearly 100% of the cells (96.6 ± 1.0% (mean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

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



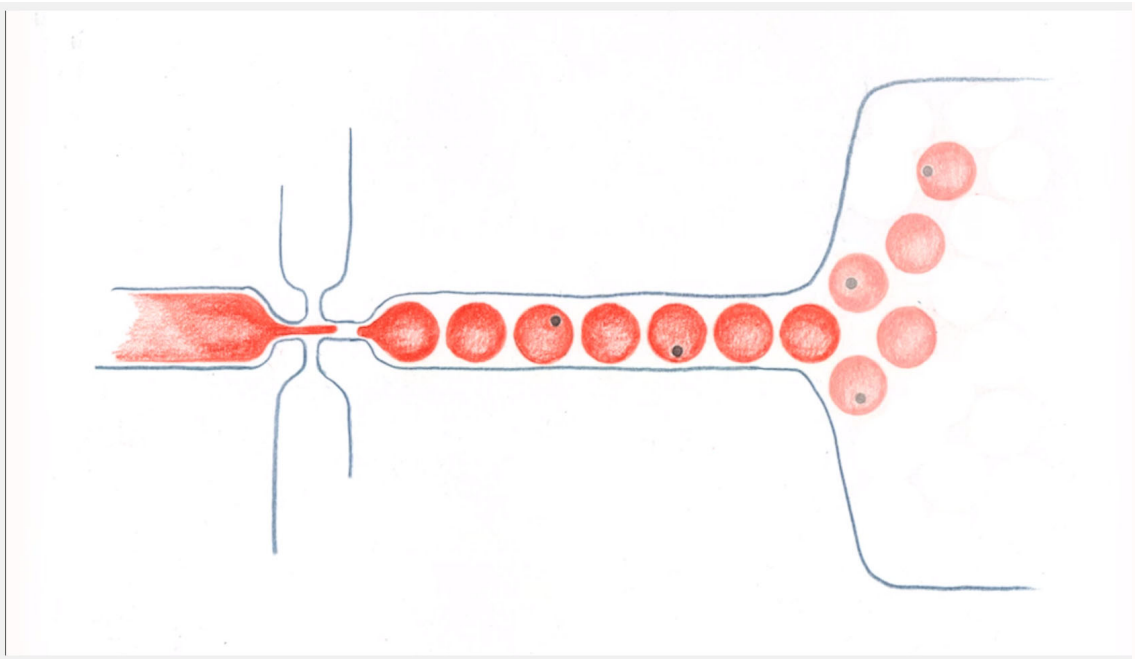
Amir et al., *Nature Biotechnology* 2013; Becher et al., *Nature Immunology* 2014

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

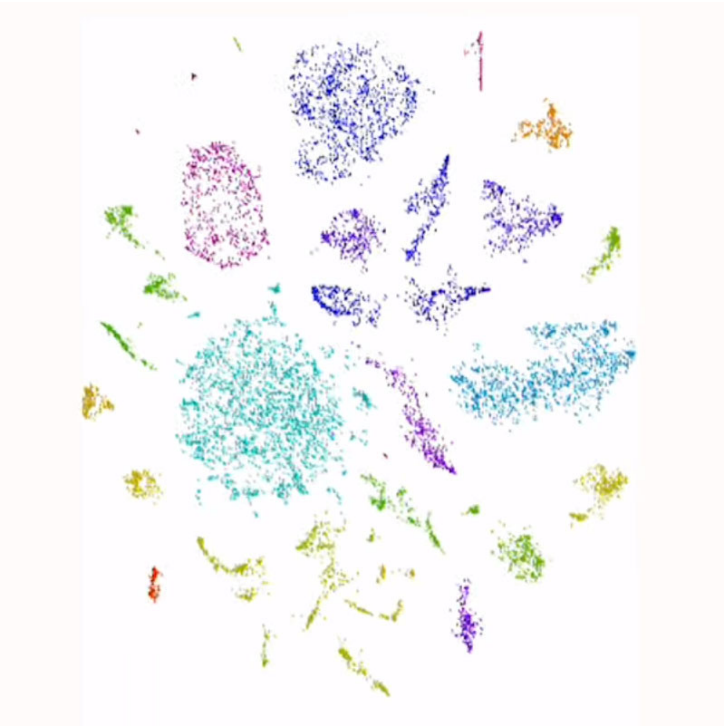


Tirosh et al., *Science* 2016

t-SNE on RNA-seq (DropSEQ data)

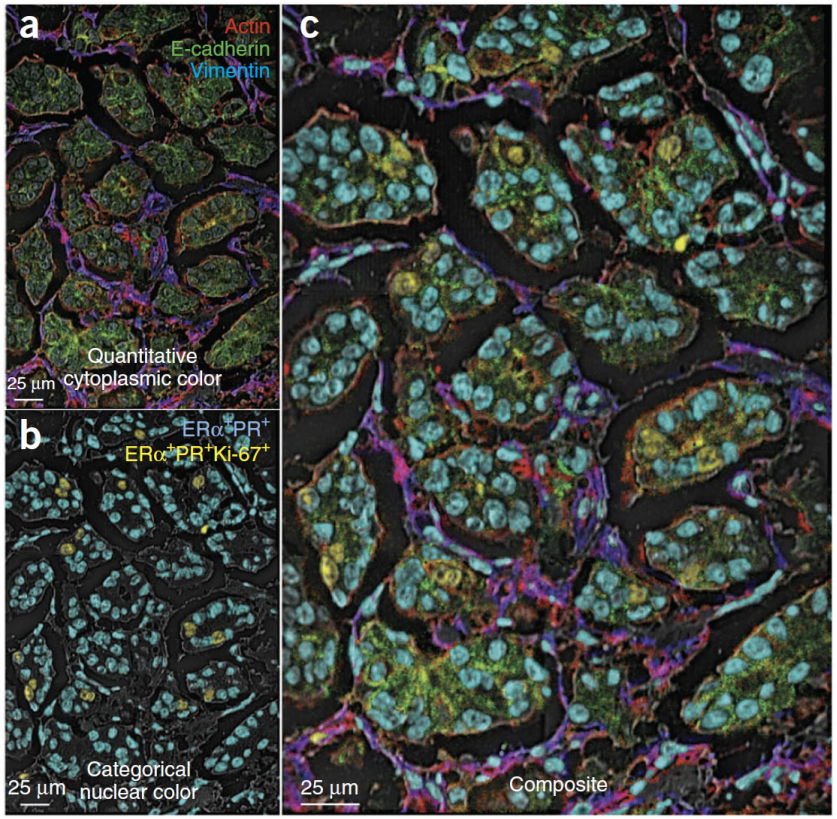


Smoothie vs. Fruit Salad/ Cell, May 21, 2015 (Vol. 161, Issue 5)

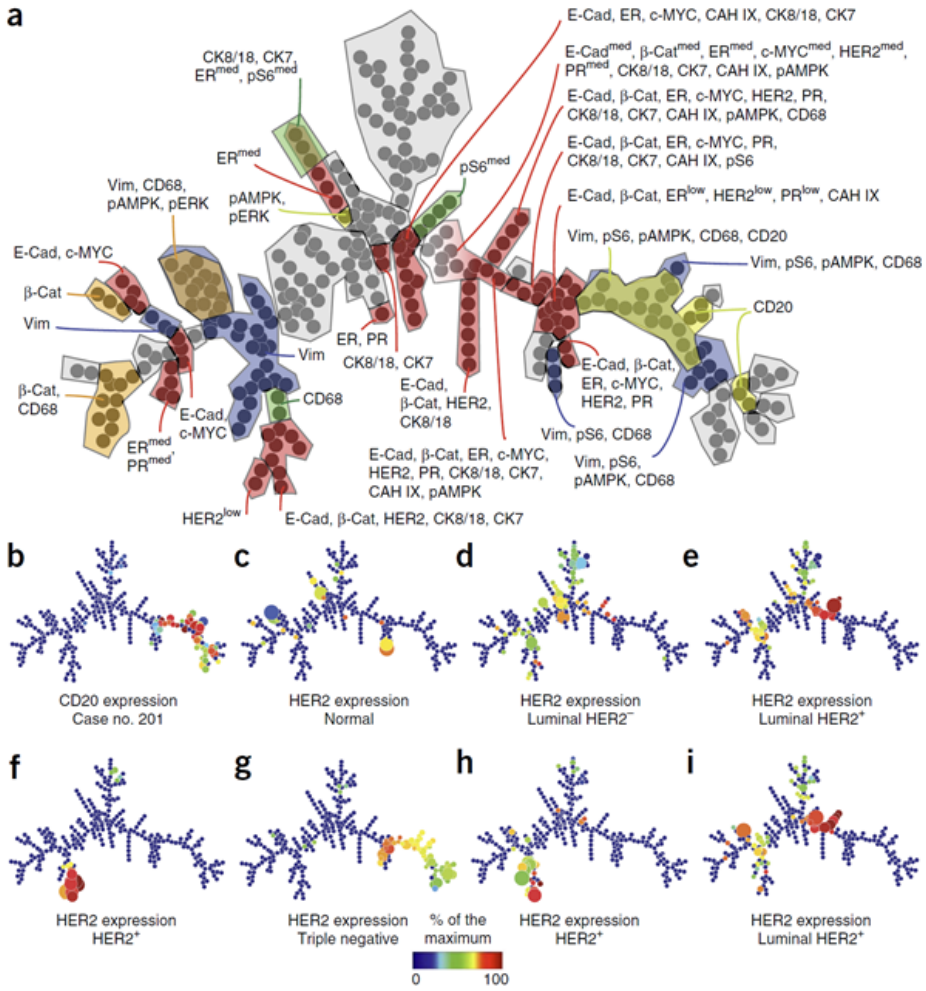


https://www.youtube.com/watch?v=XAsmHKfKHmc
About 4 min in, see an animation of t-SNE

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

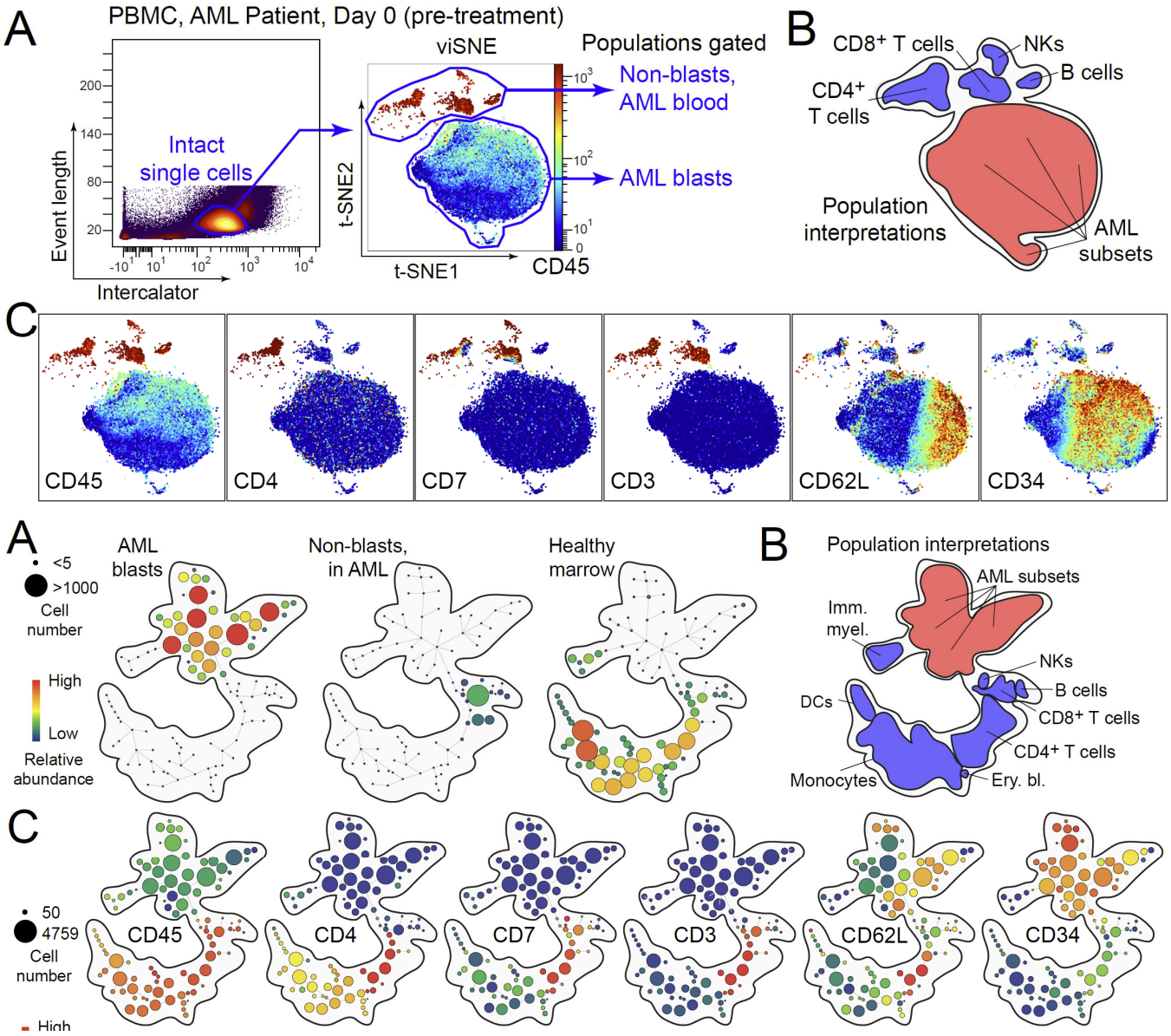


Example MIBI breast cancer histology
Angelo et al., *Nature Medicine* 2014



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

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



viSNE

Amir et al.
Nature biotech 2013

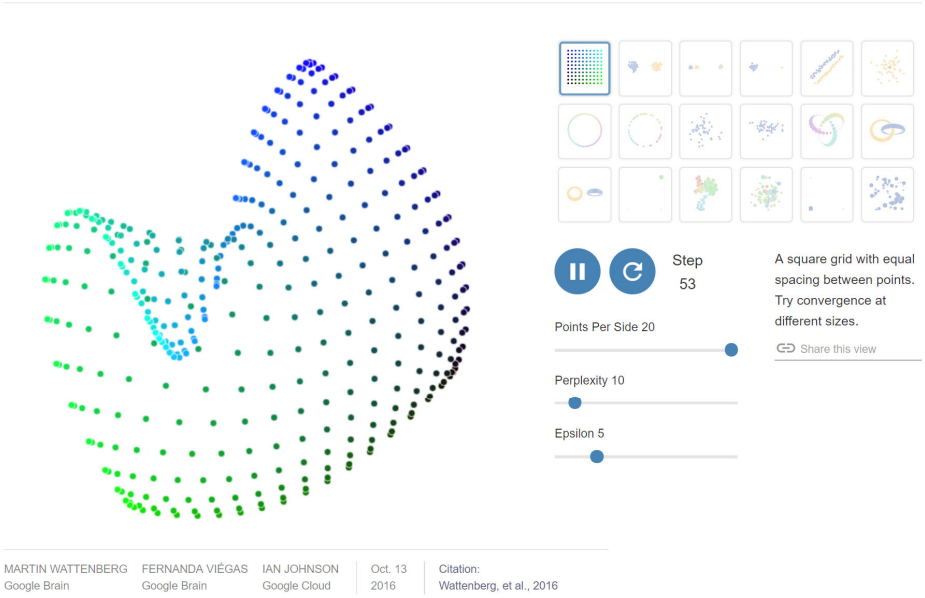
SPADE

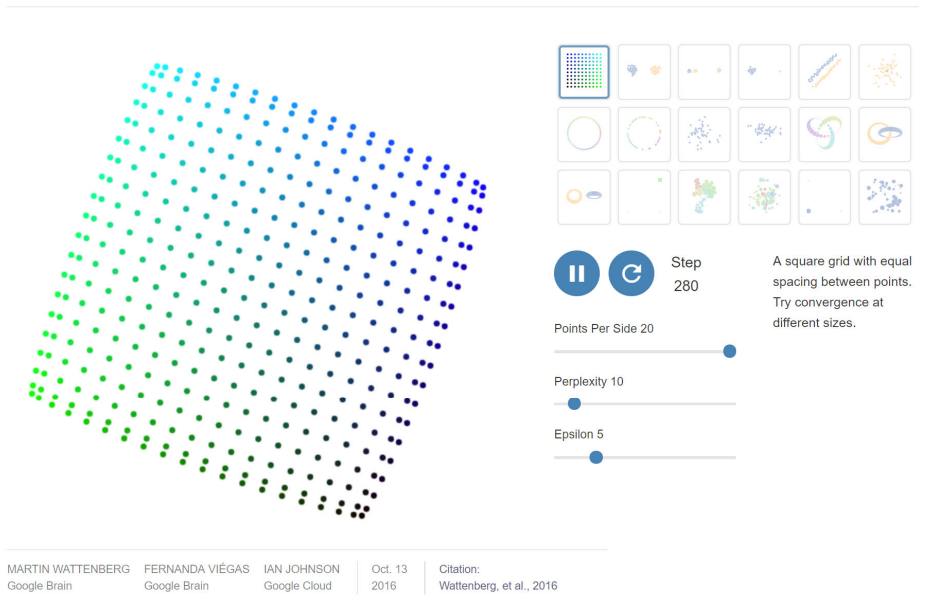
Qiu et al.
Nature biotech 2011

Diggins et al., *Methods* 2015

t-SNE 2D Examples with Animations and Settings

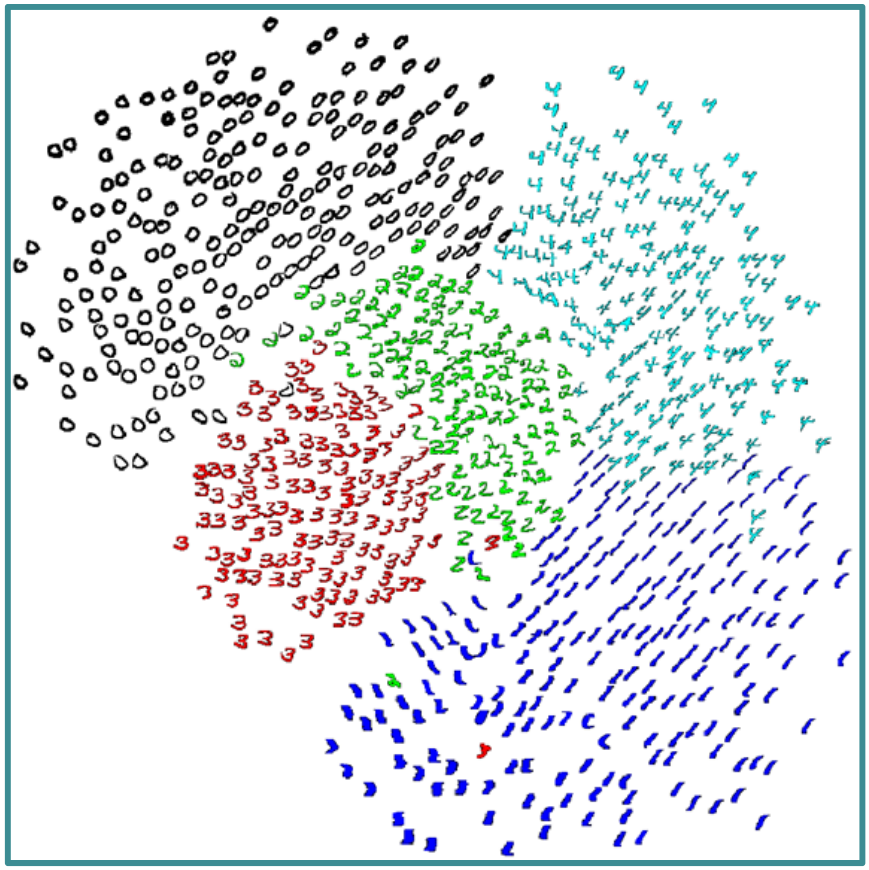
http://distill.pub/2016/misread-tsne/



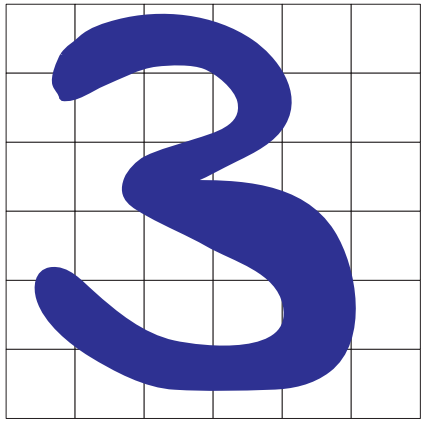


Stochastic Neighbor Embedding (SNE)

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



Example: 6x6 Pixel Image

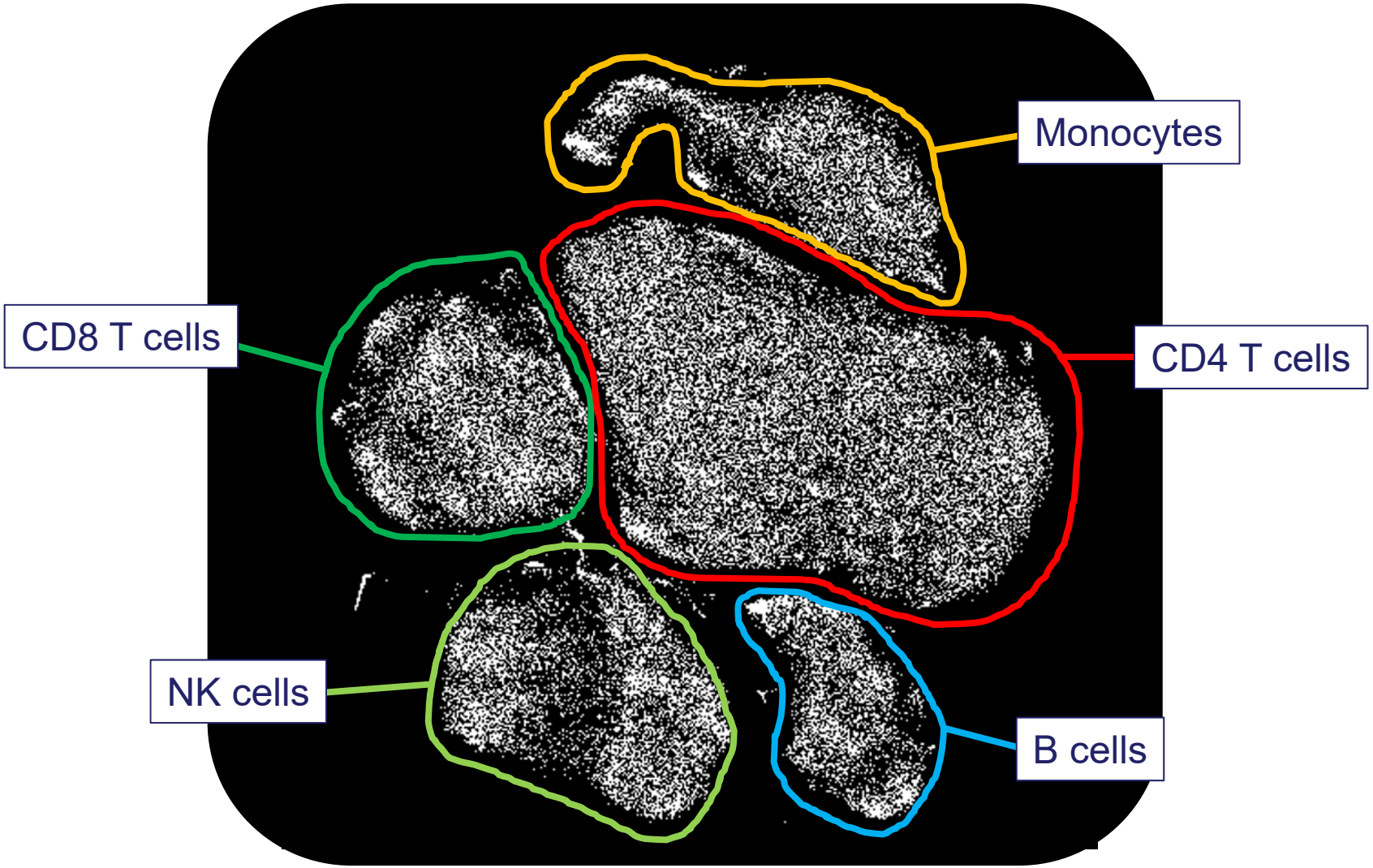


Vectorize (1x36)

tSNE on all pixels

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

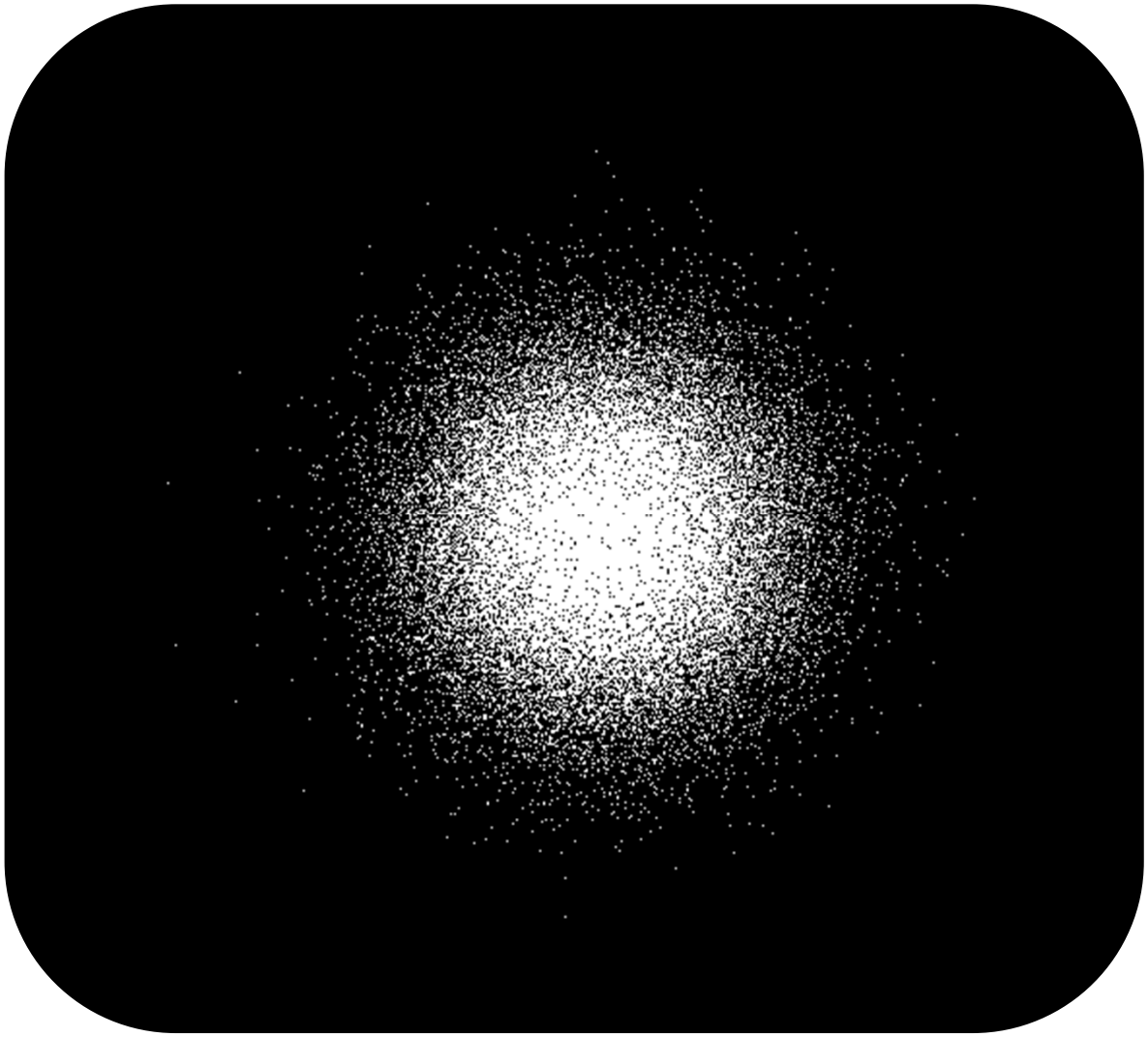
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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)

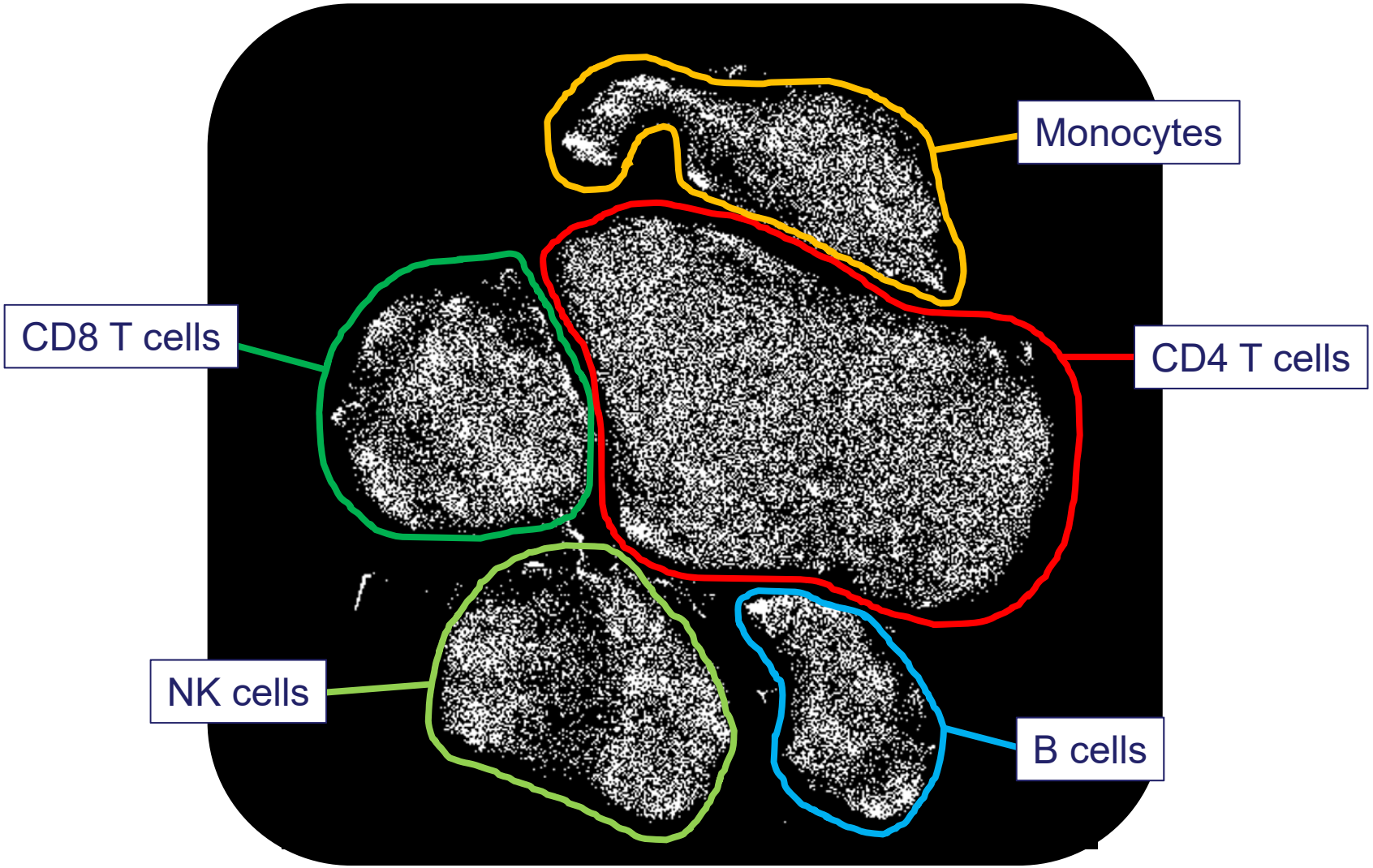
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity



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

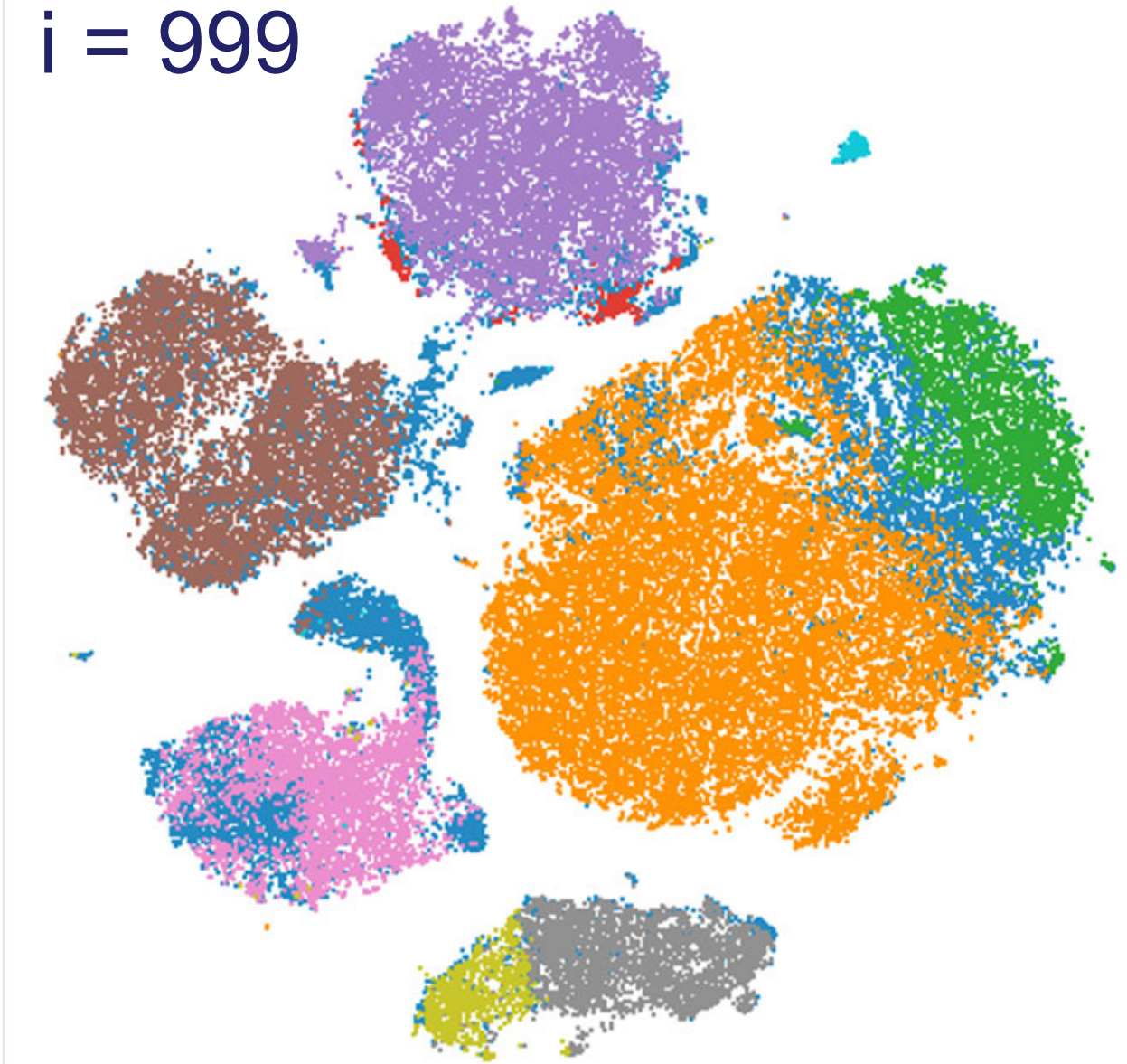


Healthy human blood, mass cytometry,
26 markers measured, viSNE analysis tool

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

Viewing Expert Gates with viSNE Reveals Cyto Incognito

Healthy human blood,
mass cytometry,
26D viSNE analysis



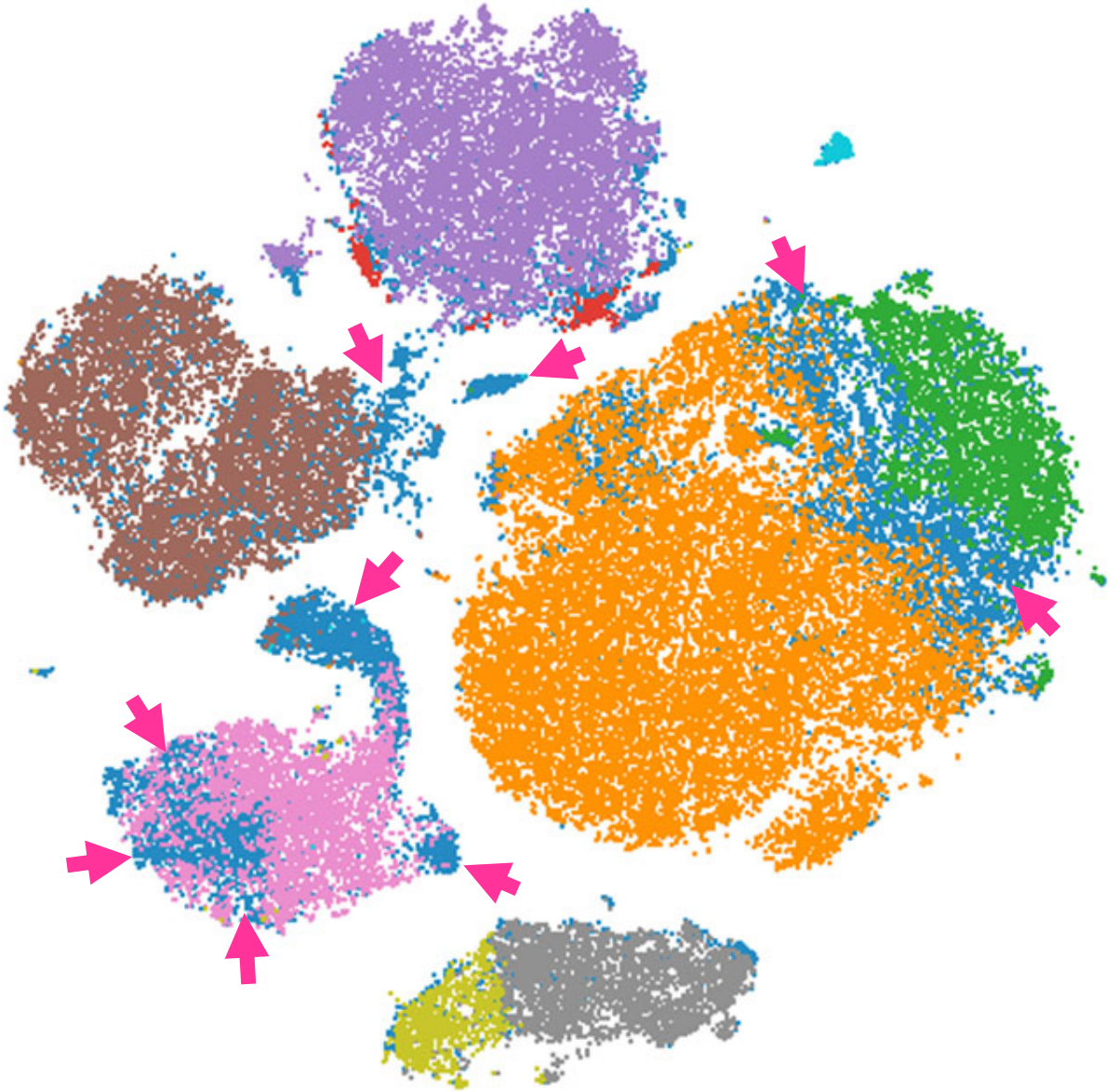
Created with Chris Ciccolella, Cytobank

Viewing Expert Gates with viSNE Reveals Cyto Incognito

Healthy human blood,
mass cytometry,
26D viSNE analysis

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

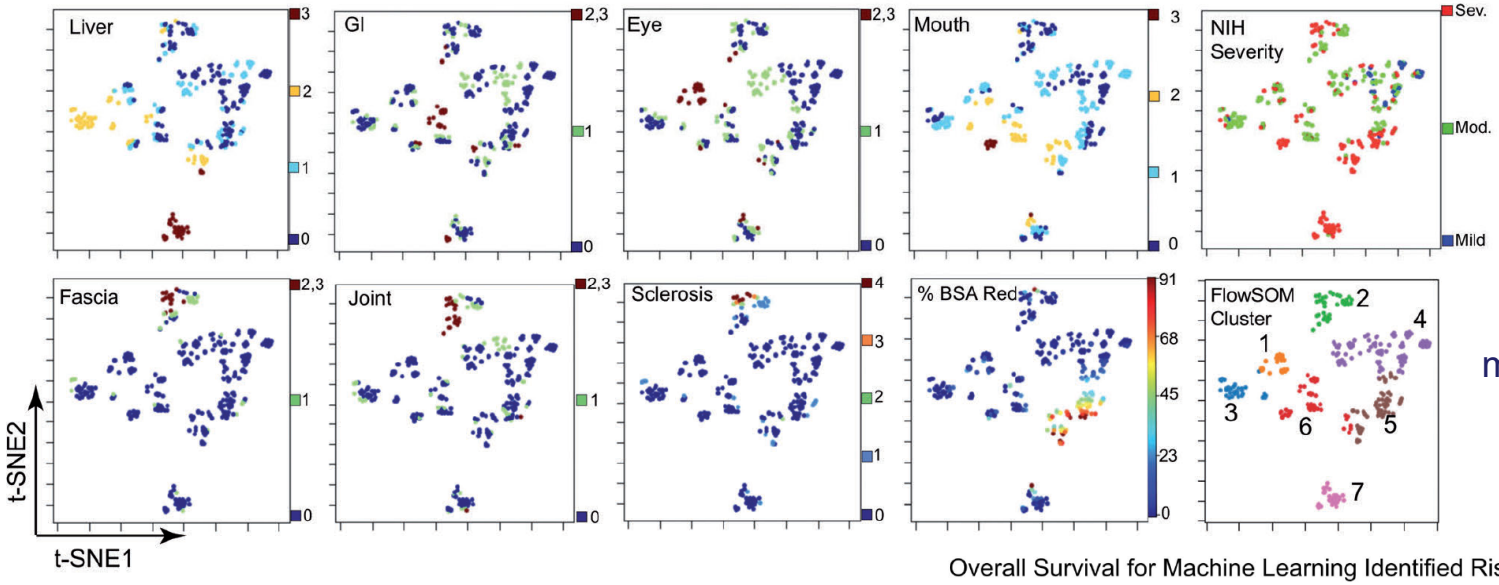
→ Cyto incognito
(Cells overlooked or
hidden in expert gating)



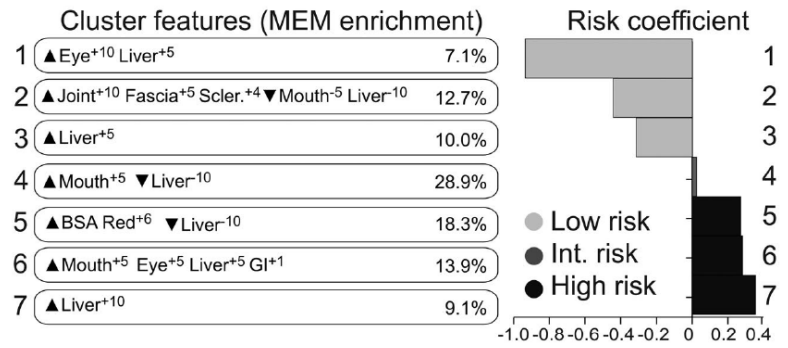
Created with Chris Ciccolella, Cytobank

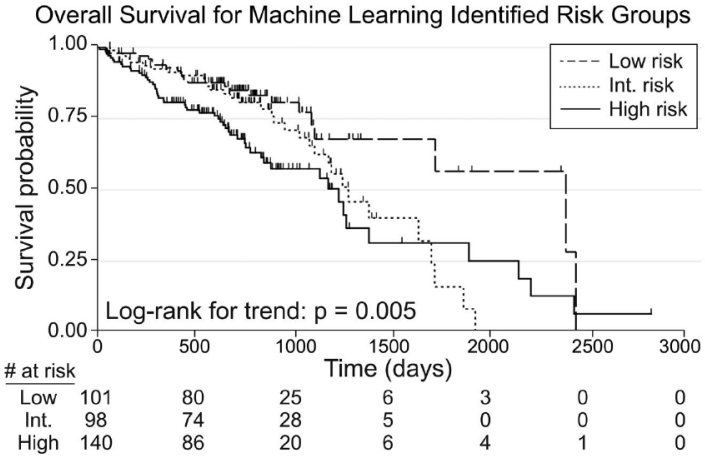
Data Science Strategies Are Widely Applicable, E.g., Reveal cGVHD Patient Groups by Medical Phenotype

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



Dots = patients
t-SNE on 8D medical phenotype



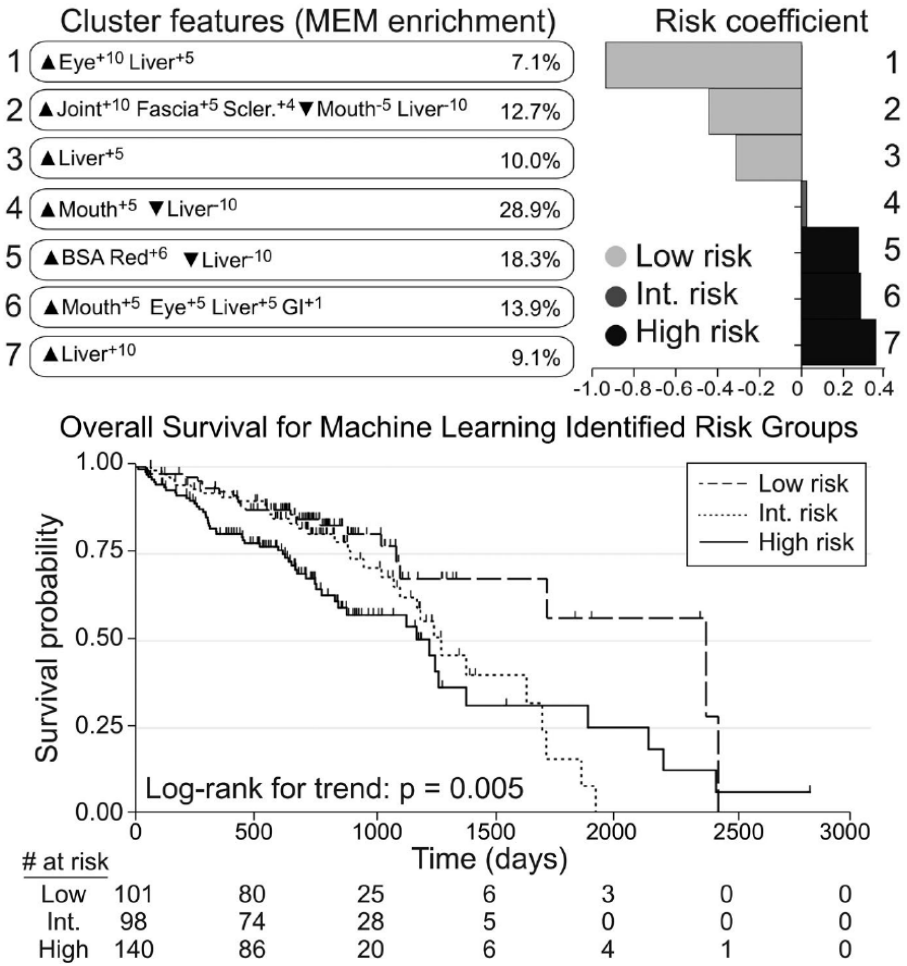


Chronic graft-*versus*-host disease (cGVHD)
t-SNE on 8 organ domain scores (e.g., liver involvement from 0 to 3)

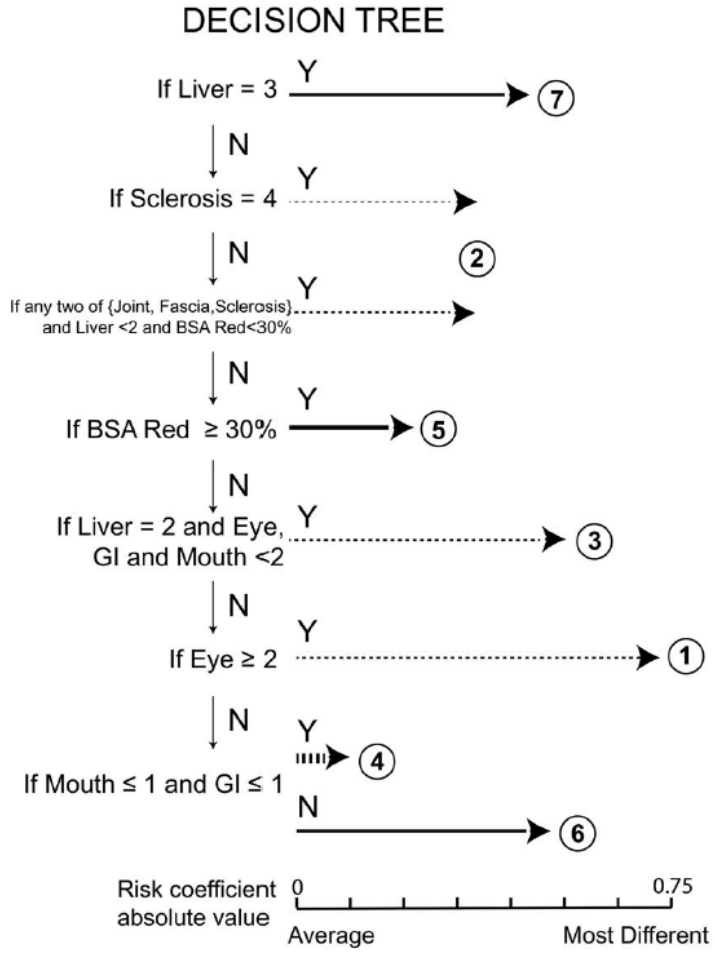
Gandelman et al., *Haematologica* 2018

Computational Strategies Are Widely Applicable, E.g., Reveal cGVHD Patient Groups by Medical Phenotype

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



Clinically compatible decision tree
ID's patient groups with ≈ MEM labels, risk



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

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

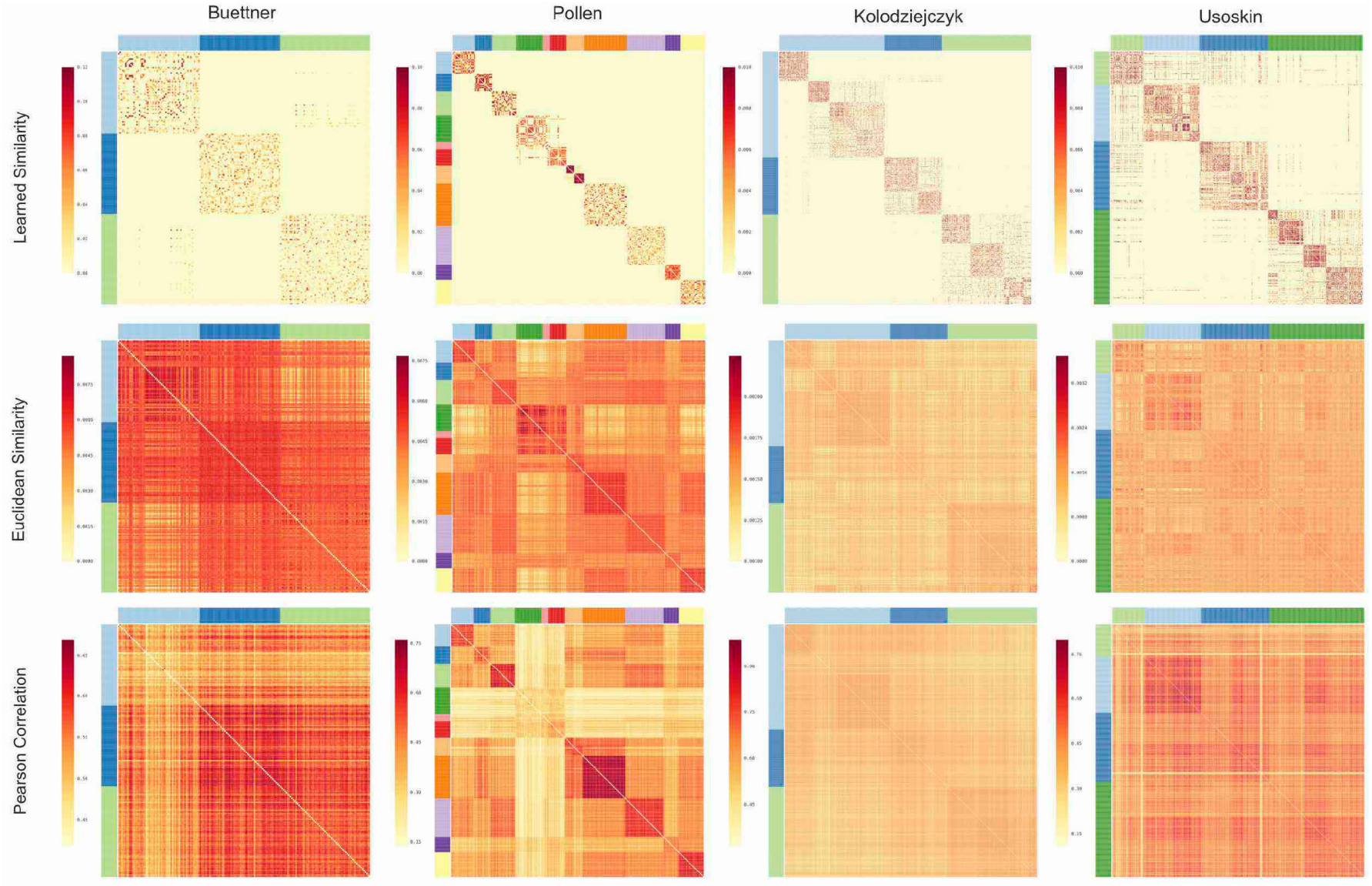
Gandelman et al., *Haematologica* 2018

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

Irish lab, unpublished

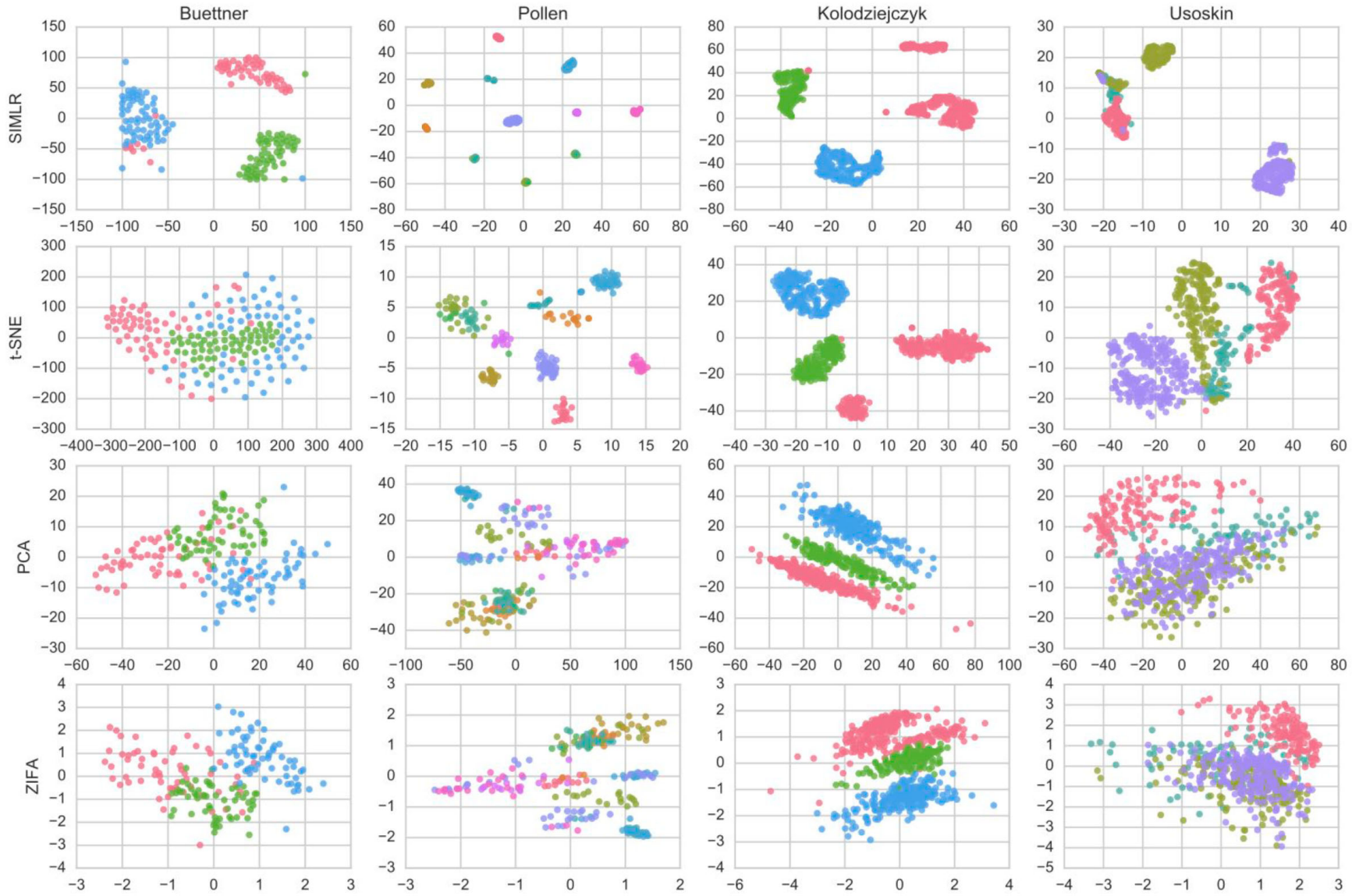
Other Related and Non-Related Tools

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



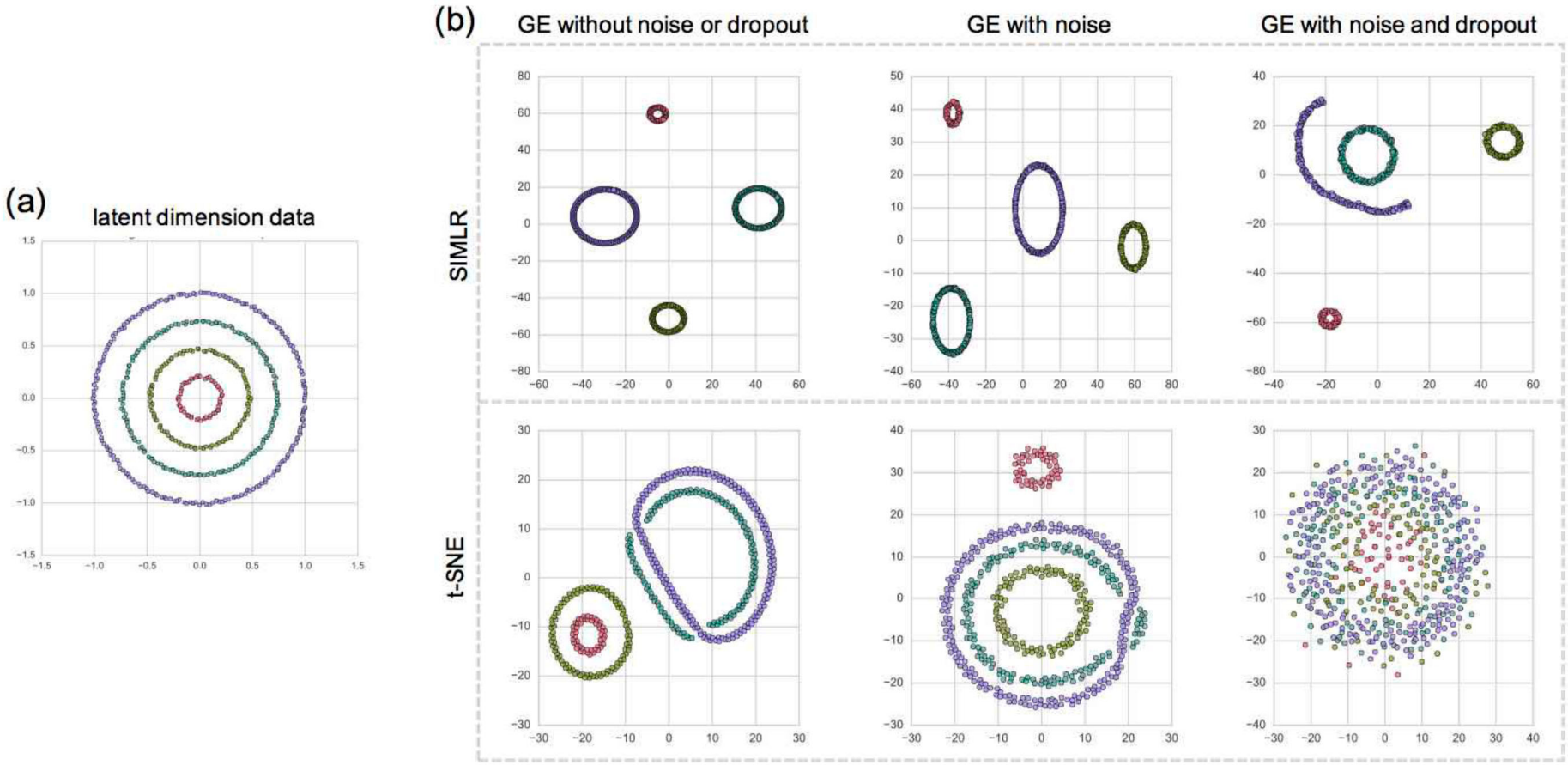
Wang et al., *Nature Methods* 2017

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



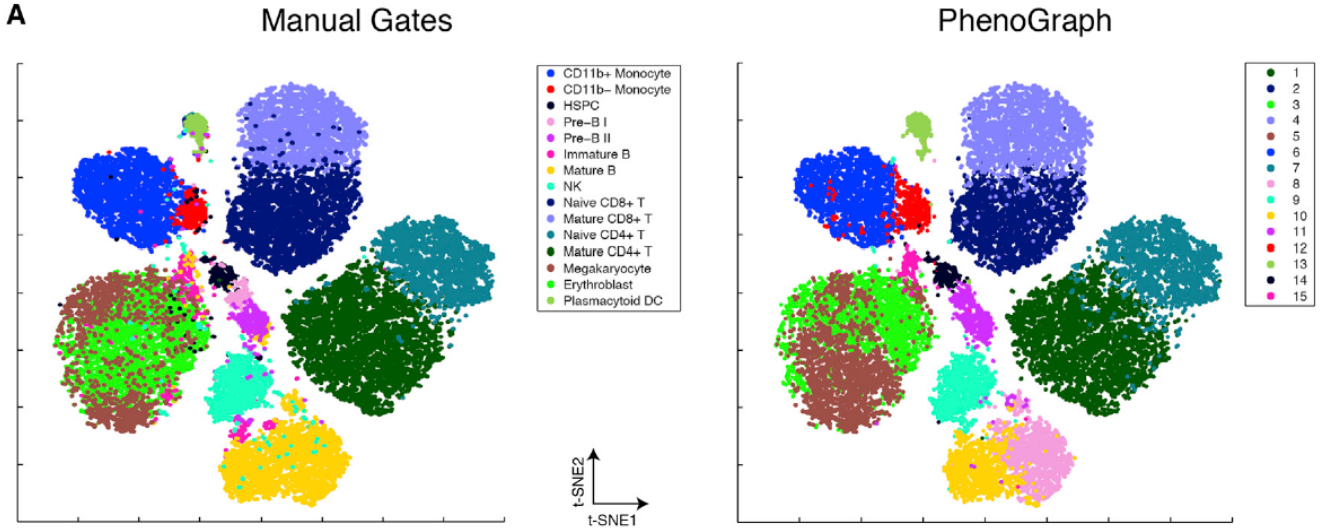
Wang et al., *Nature Methods* 2017

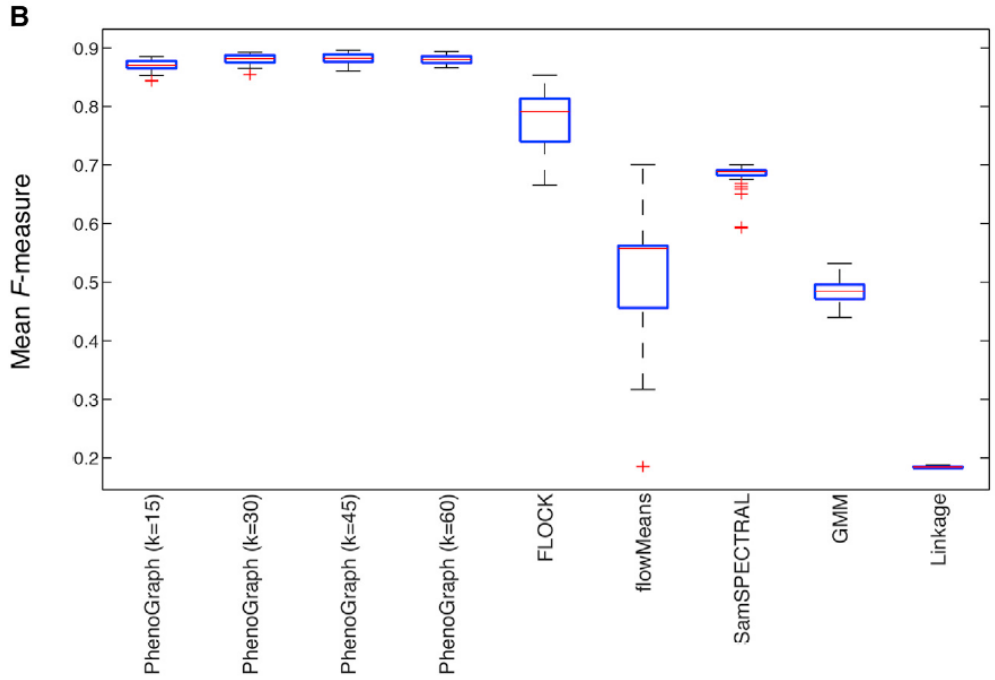
Analysis of “Toy” Gene Expression Data +/- Noise and Data Dropout



Wang et al., *Nature Methods* 2017

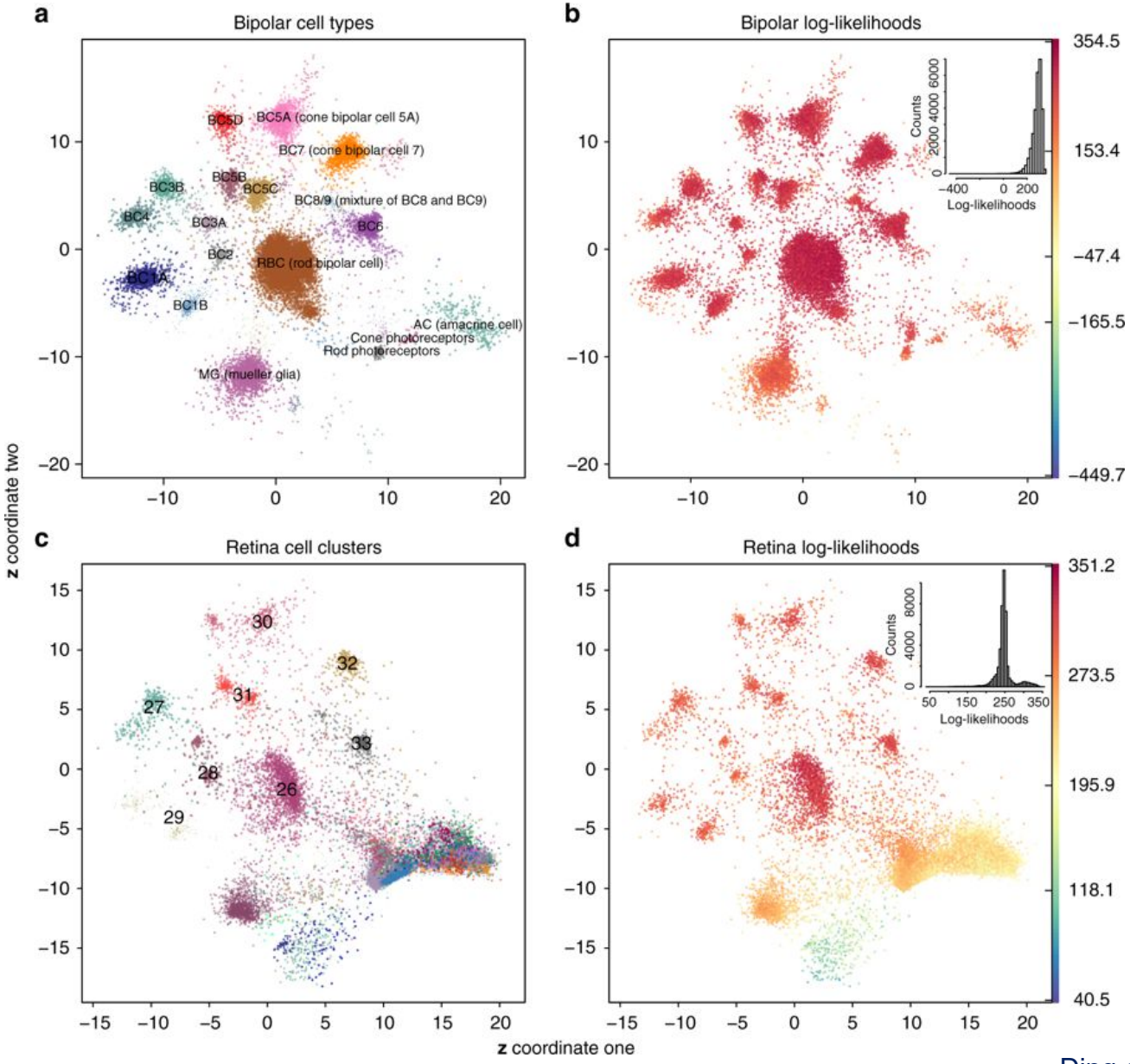
Phenograph Adds Fast Clustering & Meta-Analysis to viSNE





Levine et al., *Cell* 2015

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

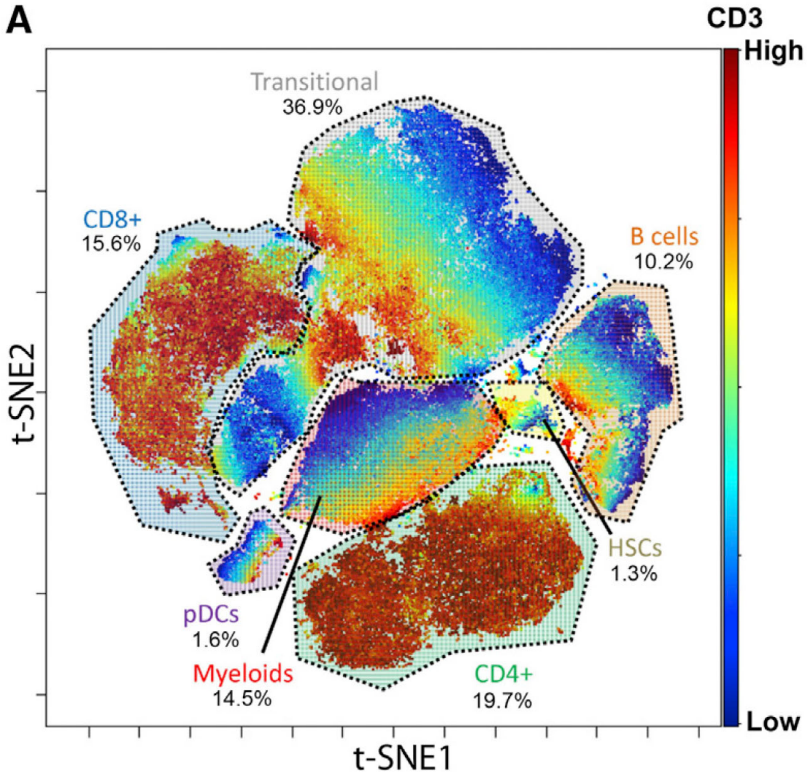


Learning a probabilistic mapping function from the bipolar data and applying the function to the independently generated mouse retina dataset. **a** scvis learned two-dimensional representations of the bipolar dataset, **b** coloring each point by the estimated log-likelihood, **c** the whole mouse retina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

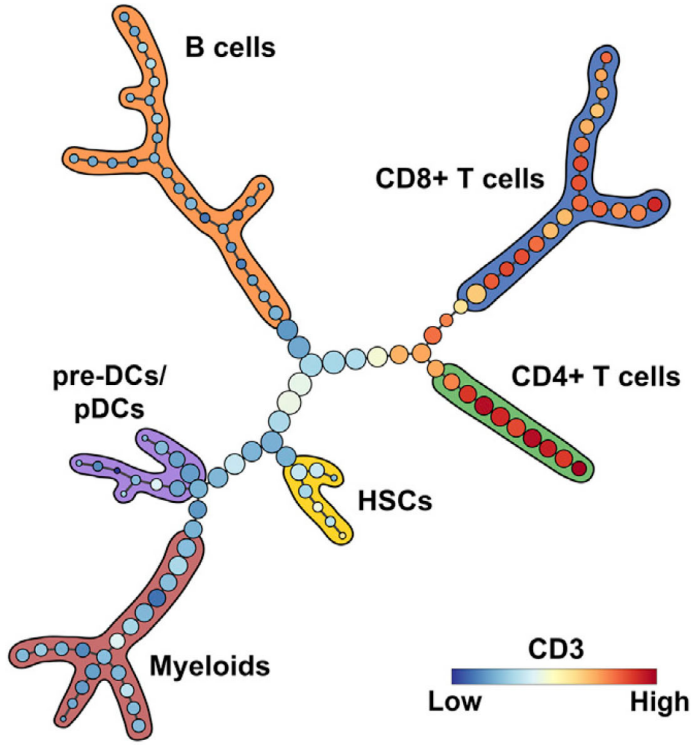
Ding et al., *Nature Communications* 2018

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

UMAP and FIt-SNE are Major Improvements

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

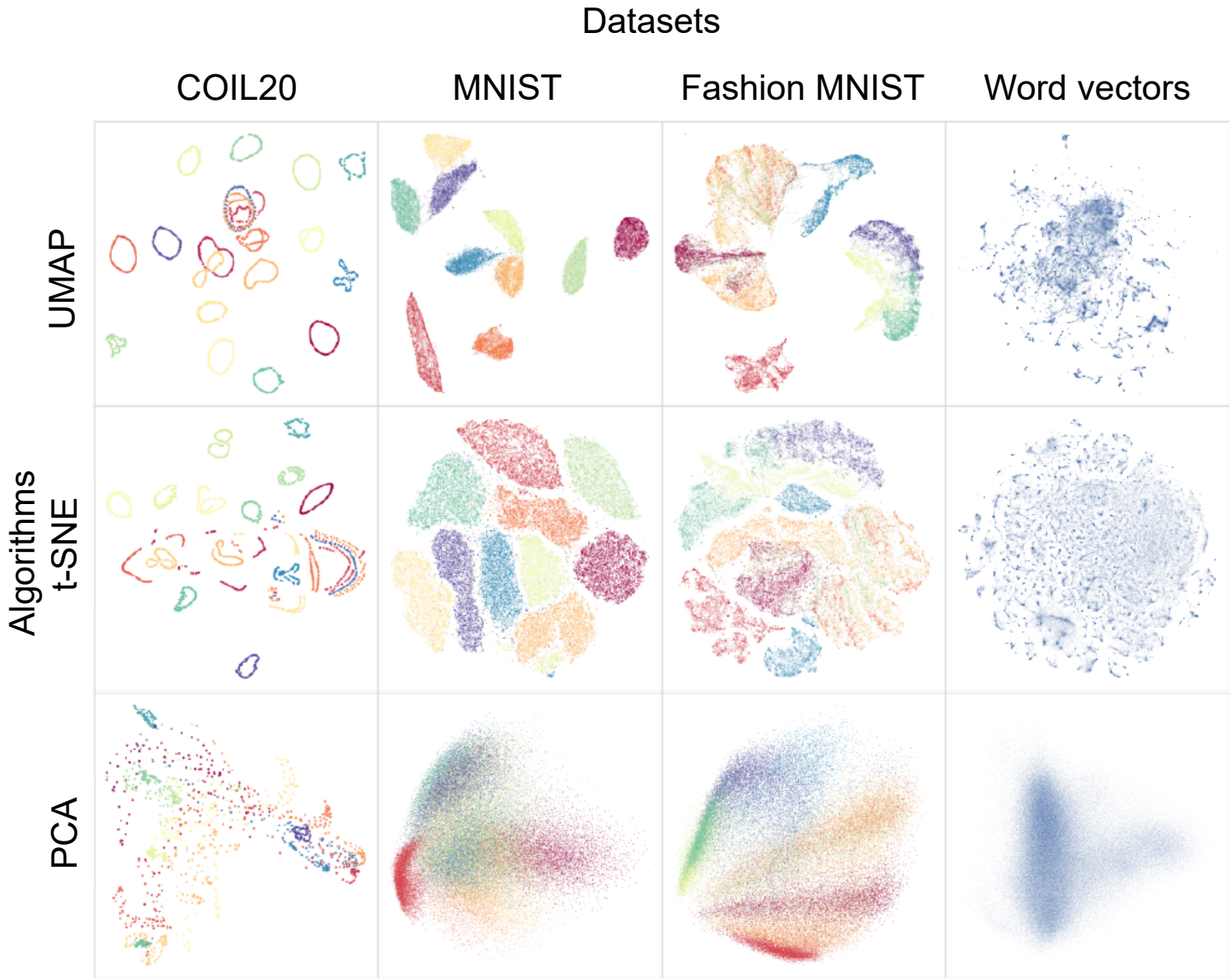


Figure 2: A comparison of several dimension reduction algorithms. We note that UMAP successfully reflects much of the large scale global structure that is well represented by Laplacian Eigenmaps and PCA (particularly for MNIST and Fashion-MNIST), while also preserving the local fine structure similar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

McInnes et al., *Rxiv* 2018

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

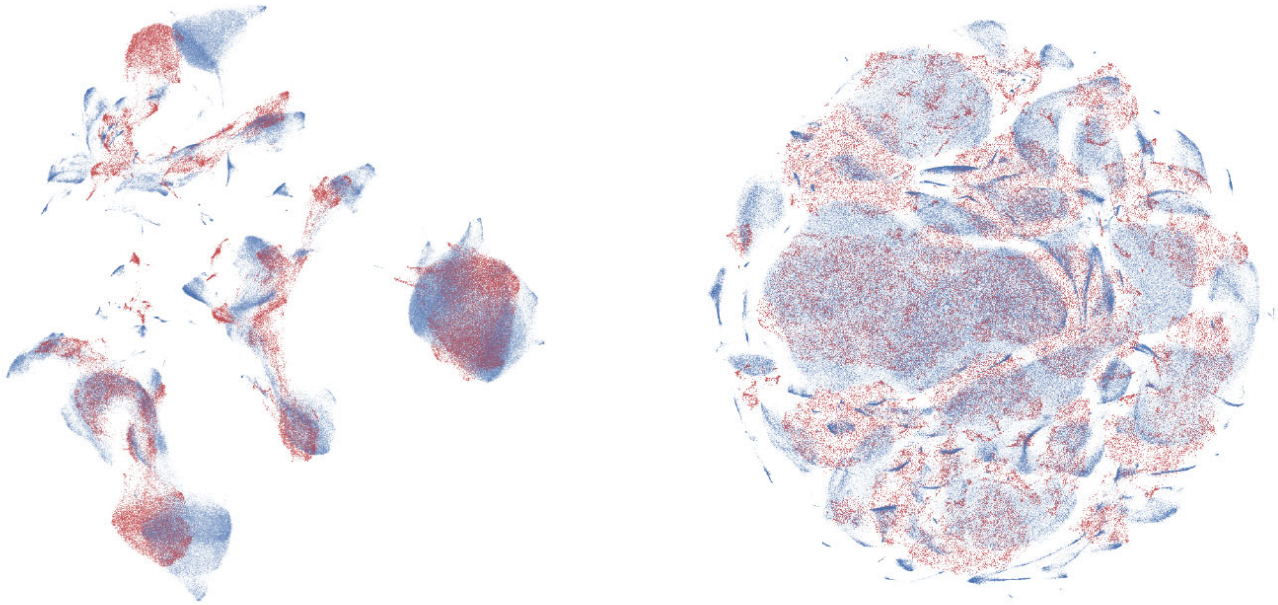
UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that applies to real world data. The 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

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).

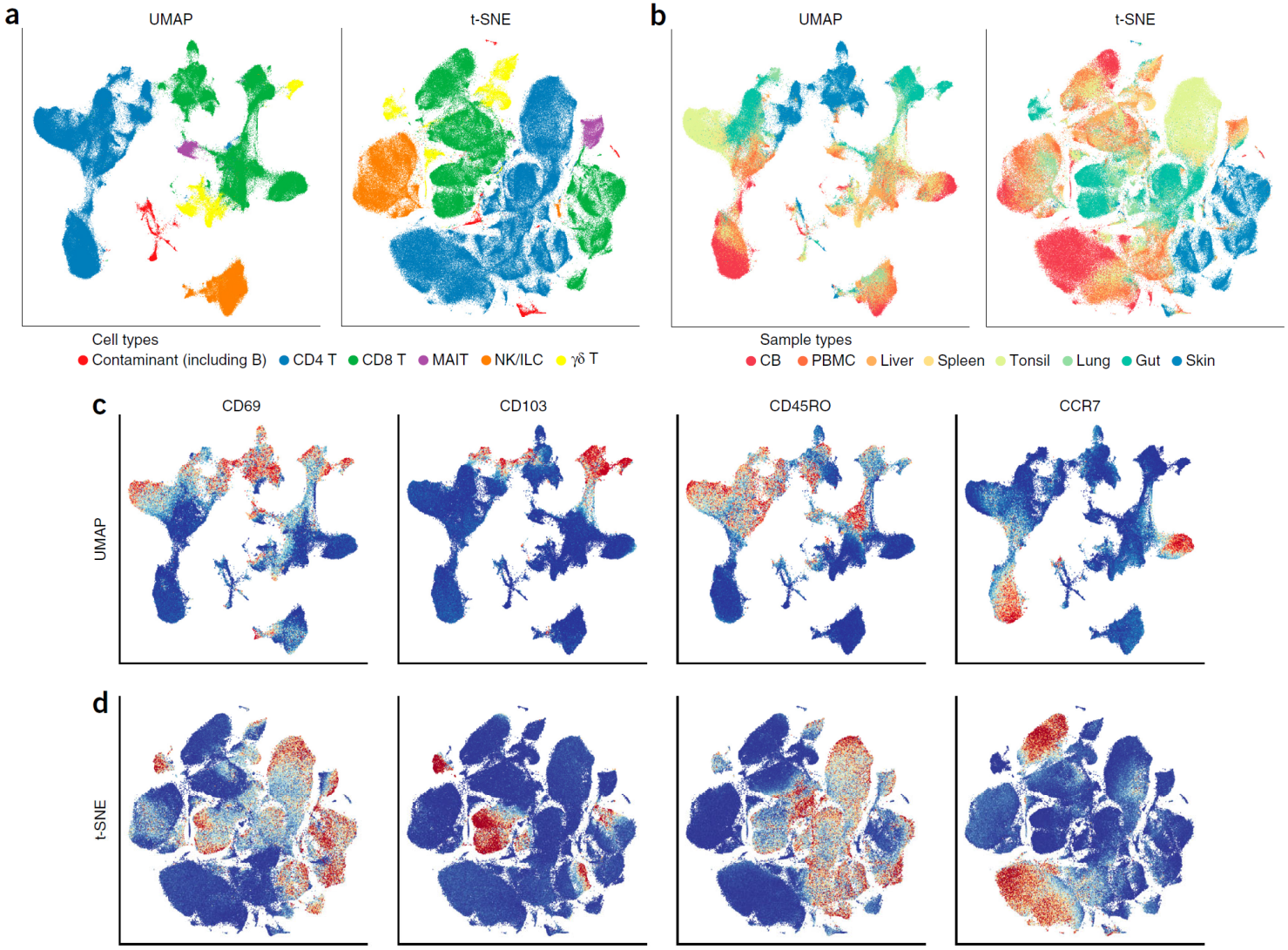


Figure 1 UMAP embeds local and large-scale structure of the data. UMAP and t-SNE projections of the Wong *et al.* dataset colored according to (**a**) broad cell lineages, (**b**) tissue of origin, and for (**c**) UMAP and (**d**) t-SNE, the expression of CD69, CD103, CD45RO and CCR7. For **c** and **d**, 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

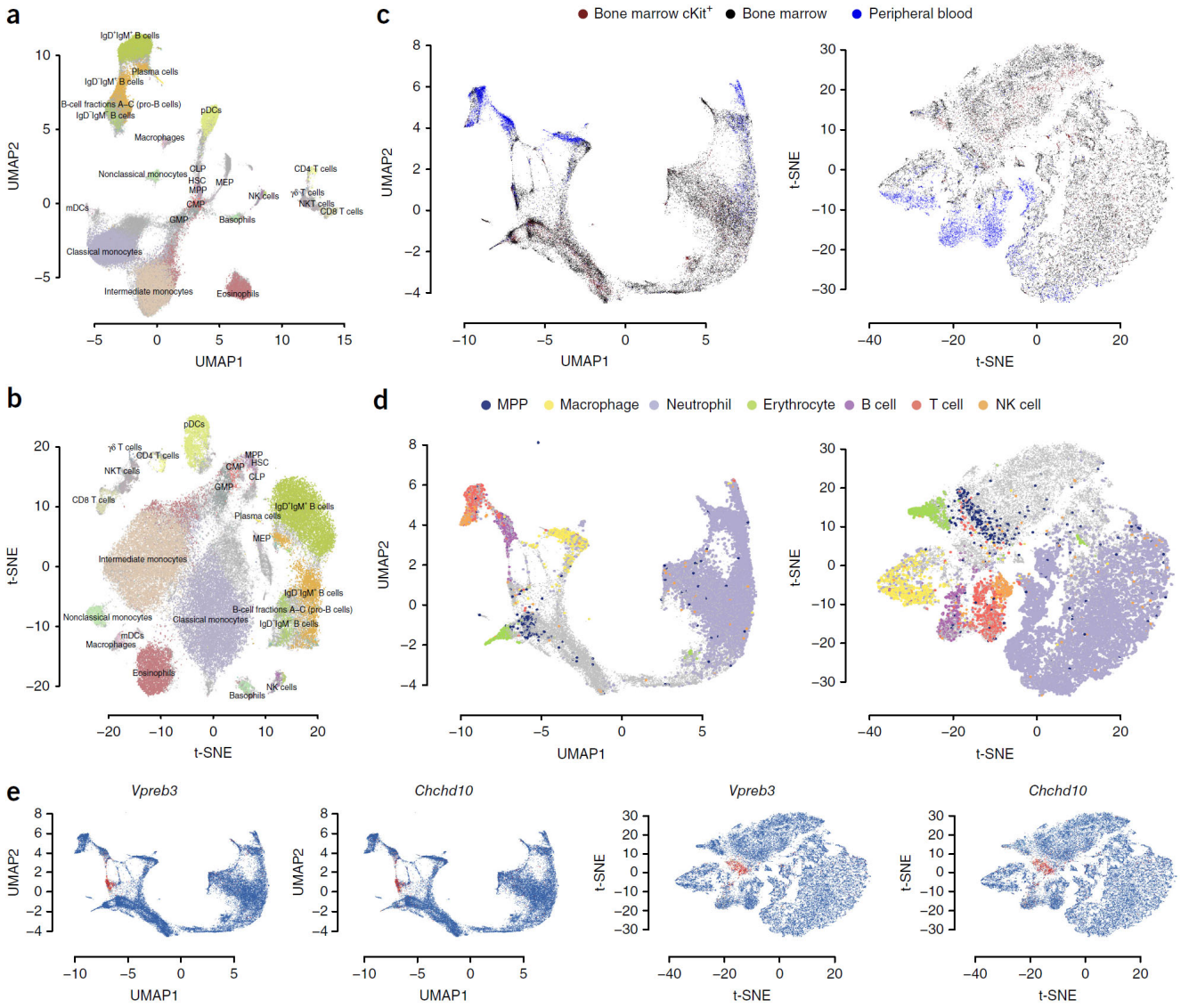
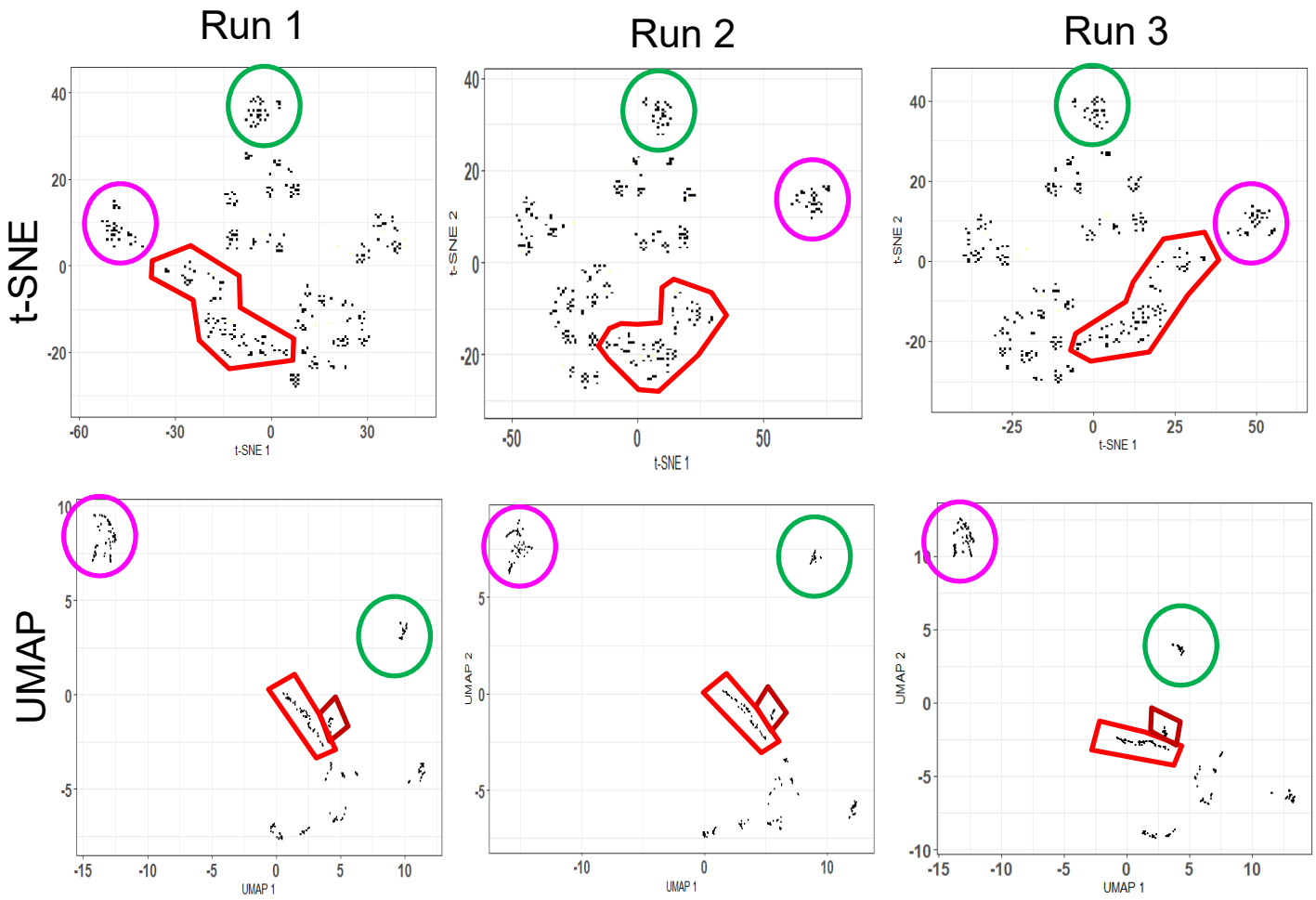


Figure 2 UMAP embeddings of bone marrow and blood samples recapitulate hematopoiesis. (**a**) UMAP and (**b**) t-SNE projection of the Samusik_01 dataset. Events are color-coded according to manual gates provided by the authors of the dataset. (**c**,**d**) UMAP and t-SNE projections of the Han dataset, 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

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

Gandelman et al., cGVHD Patient Dataset

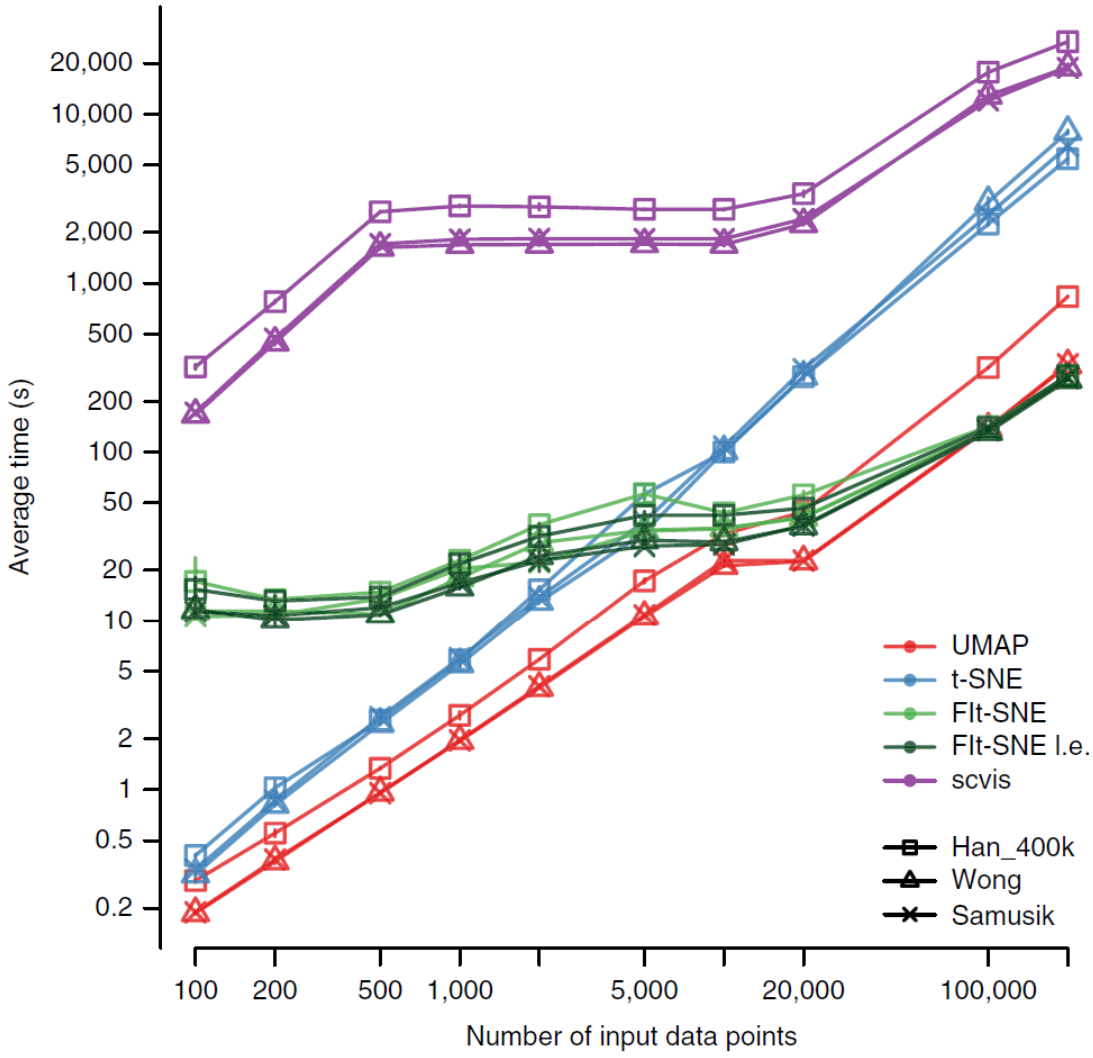


In the t-SNE plots, the relative relationship of the major islands ("global structure") alters between runs; t-SNE focuses on local structure

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 & 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

Speed is Nothing without Accuracy; UMAP is Also Accurate

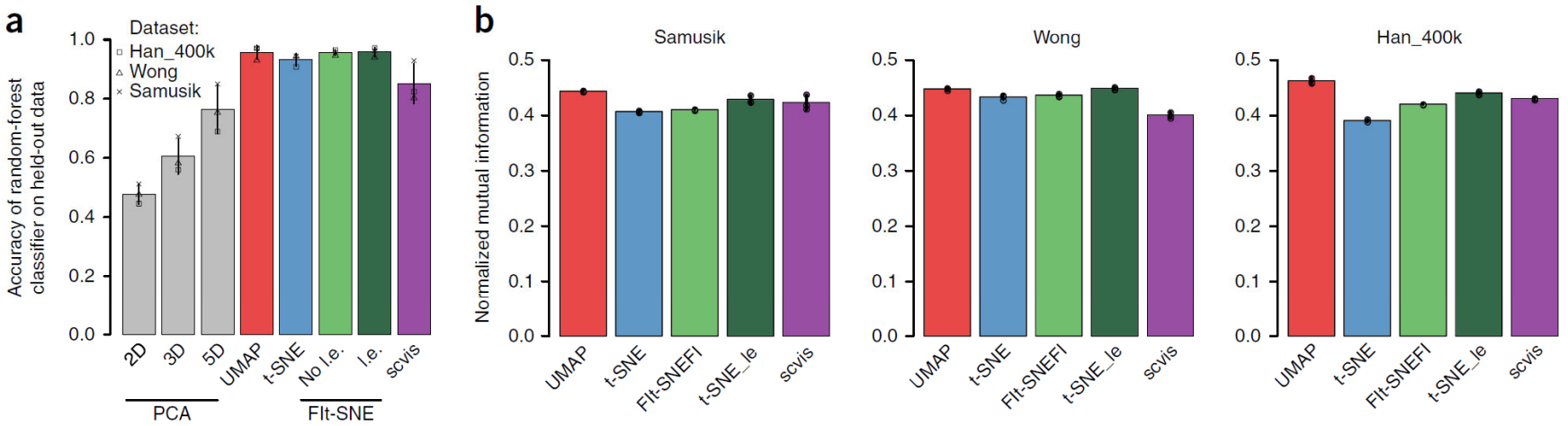
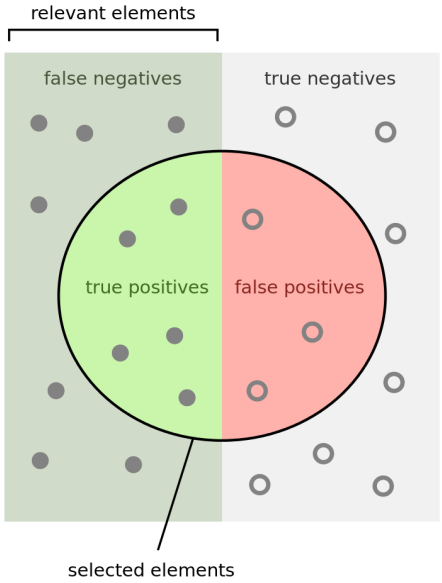


Figure 4 Analysis of local data structure in embeddings produced by each algorithm. (**a**) Accurate classification rate on held-out data of random-forest classifiers predicting Phenograph cluster labels using embedded coordinates as input. The average across the three datasets is shown, with vertical 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)

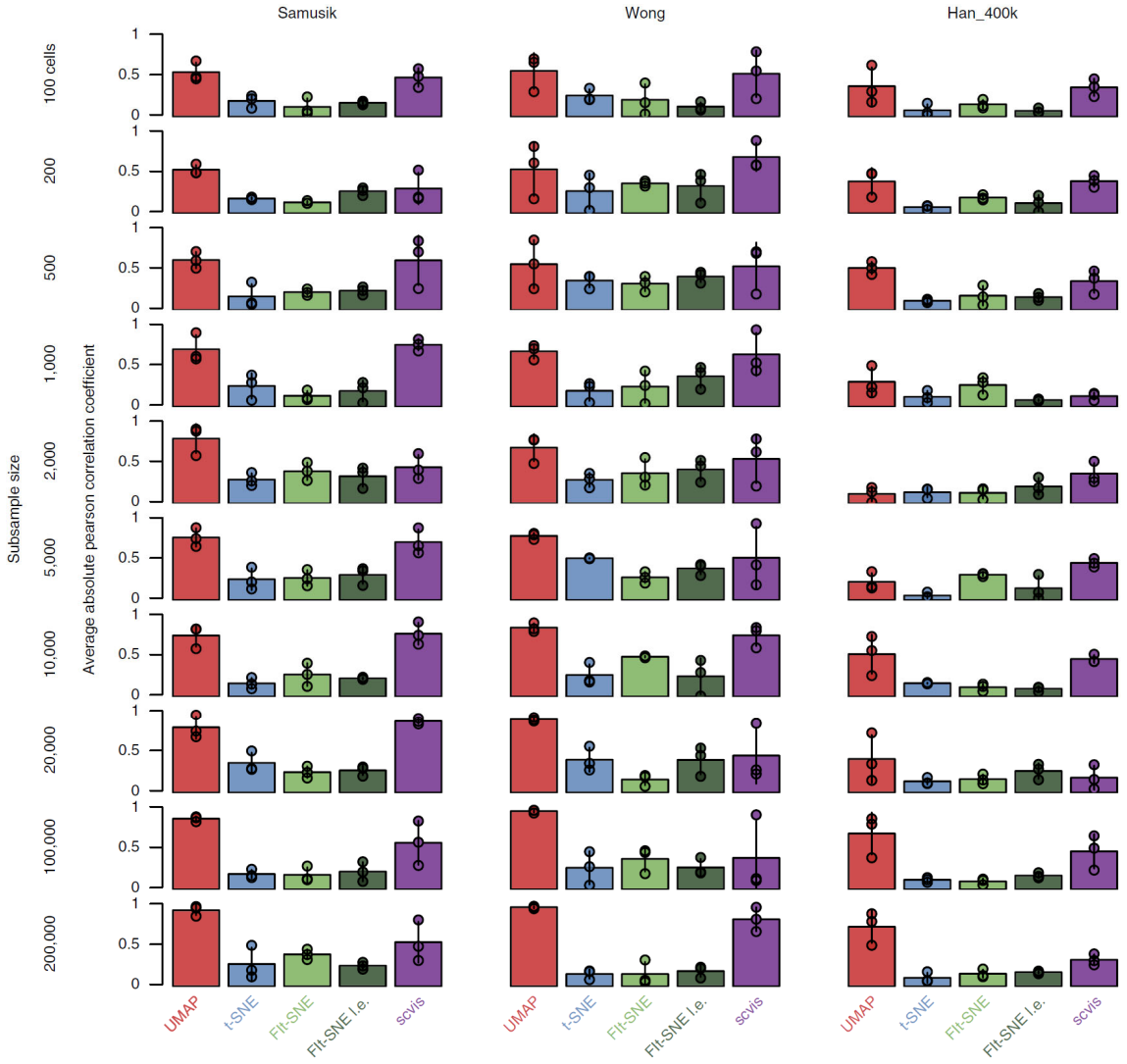


Figure 6 Reproducibility of large-scale structures in embeddings. Bar plots represent the average unsigned Pearson correlation coefficient of the points' coordinates in the embedding of subsamples versus in the embedding of the full dataset, thus measuring the correlation of coordinates in subsamples 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.

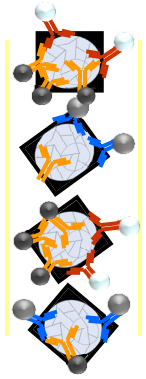
Becht et al., *Nature Biotechnology* 2018

Poorly scaled data disrupts analysis,
most issues arise near zero

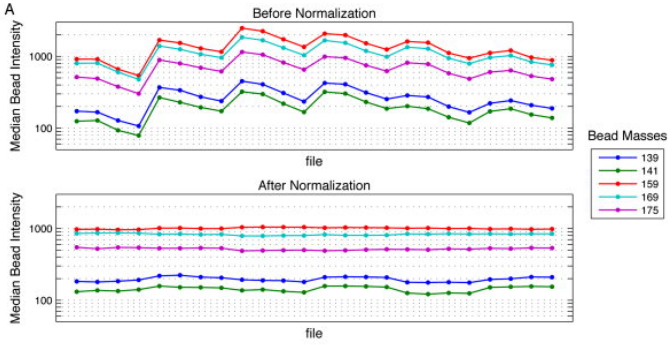
Pre-processing & normalization are also critical

Mass Cytometry Data Pre-Processing

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



FCS File 1
+
FCS File 2
+
FCS File 3



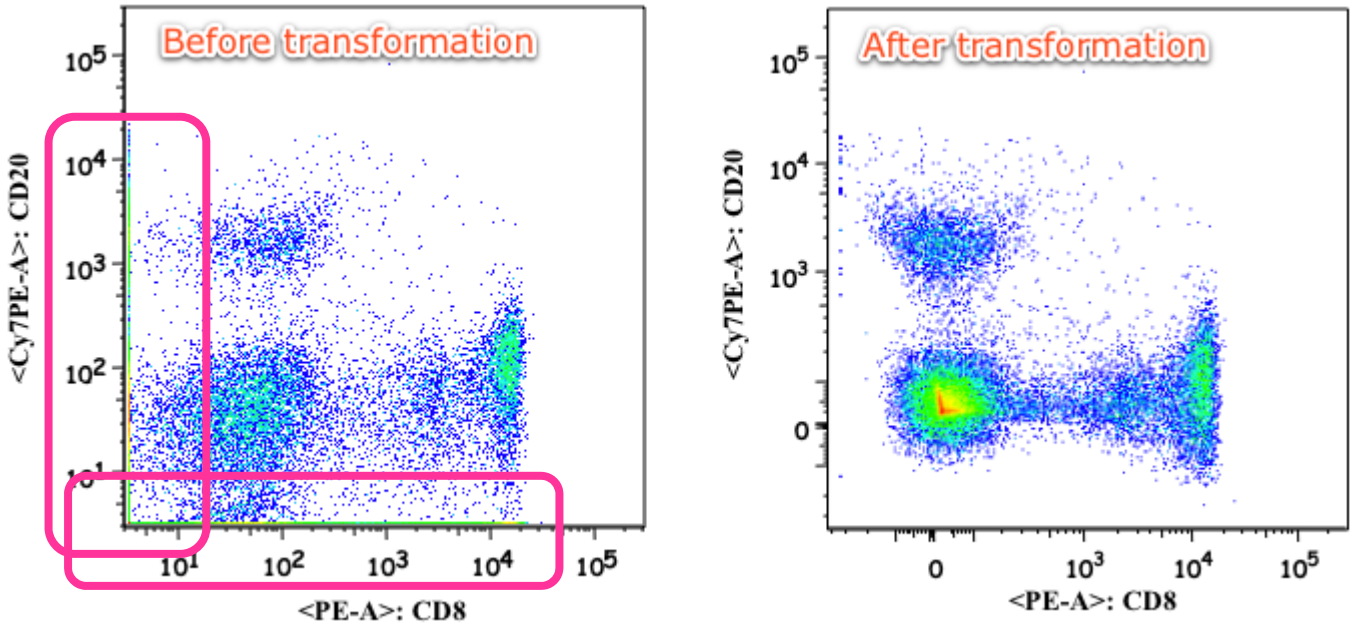
arcsinh(x/5)

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)

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

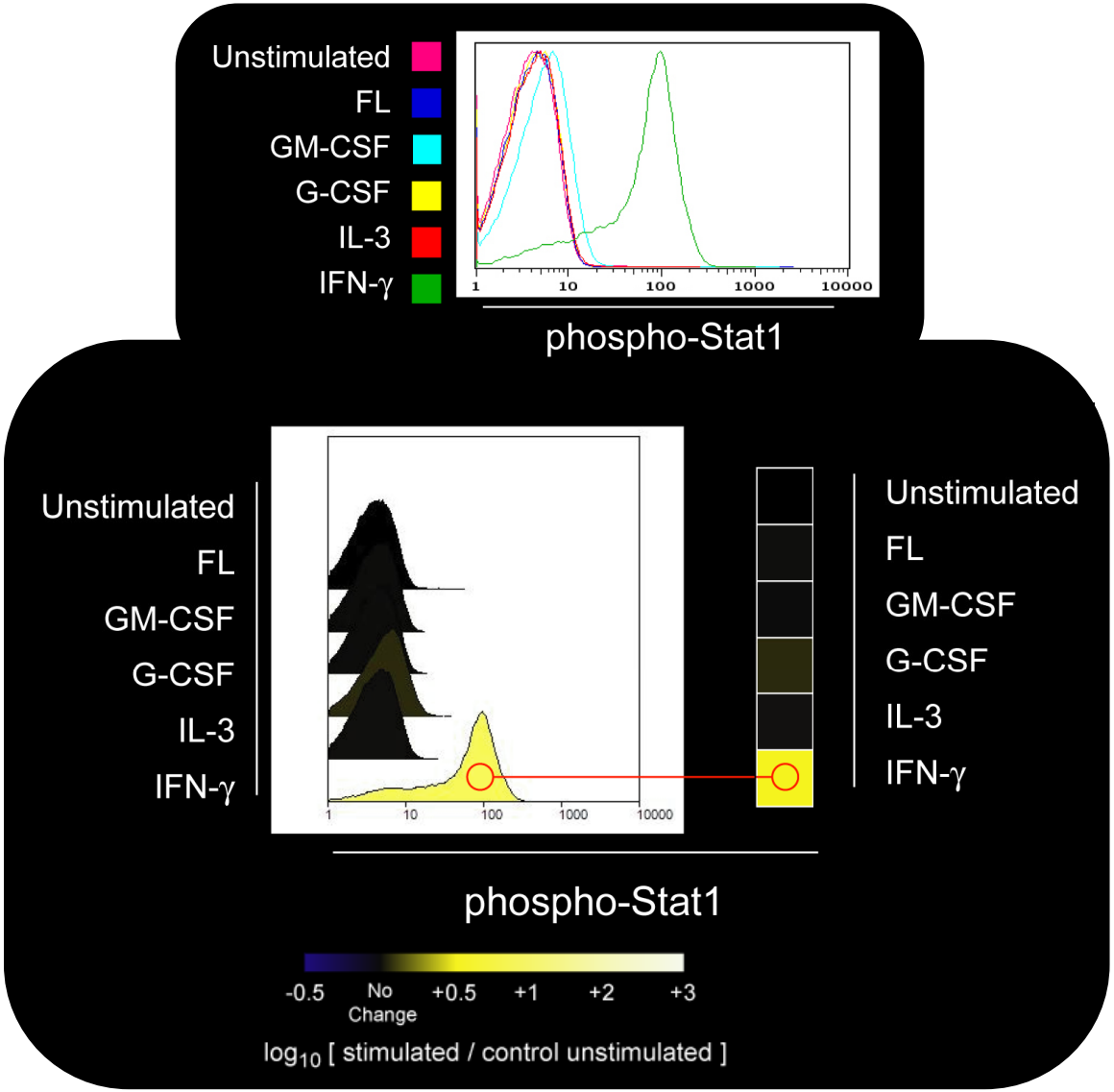


http://www.flowjo.com/v76/en/displaytransformwhy.html

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

Scaling is important in both mass and fluorescence cytometry.

Example Analysis Technique: Heatmaps and Histogram Overlays Use Fold Change



Fold change on a $\log_{10}$ scale:

$x' = \log_{10}(x \text{ MFI} / \text{control MFI})$

$x' = \log_{10}(x \text{ MFI}) - \log_{10}(\text{control MFI})$

Fold change on log-like scales:

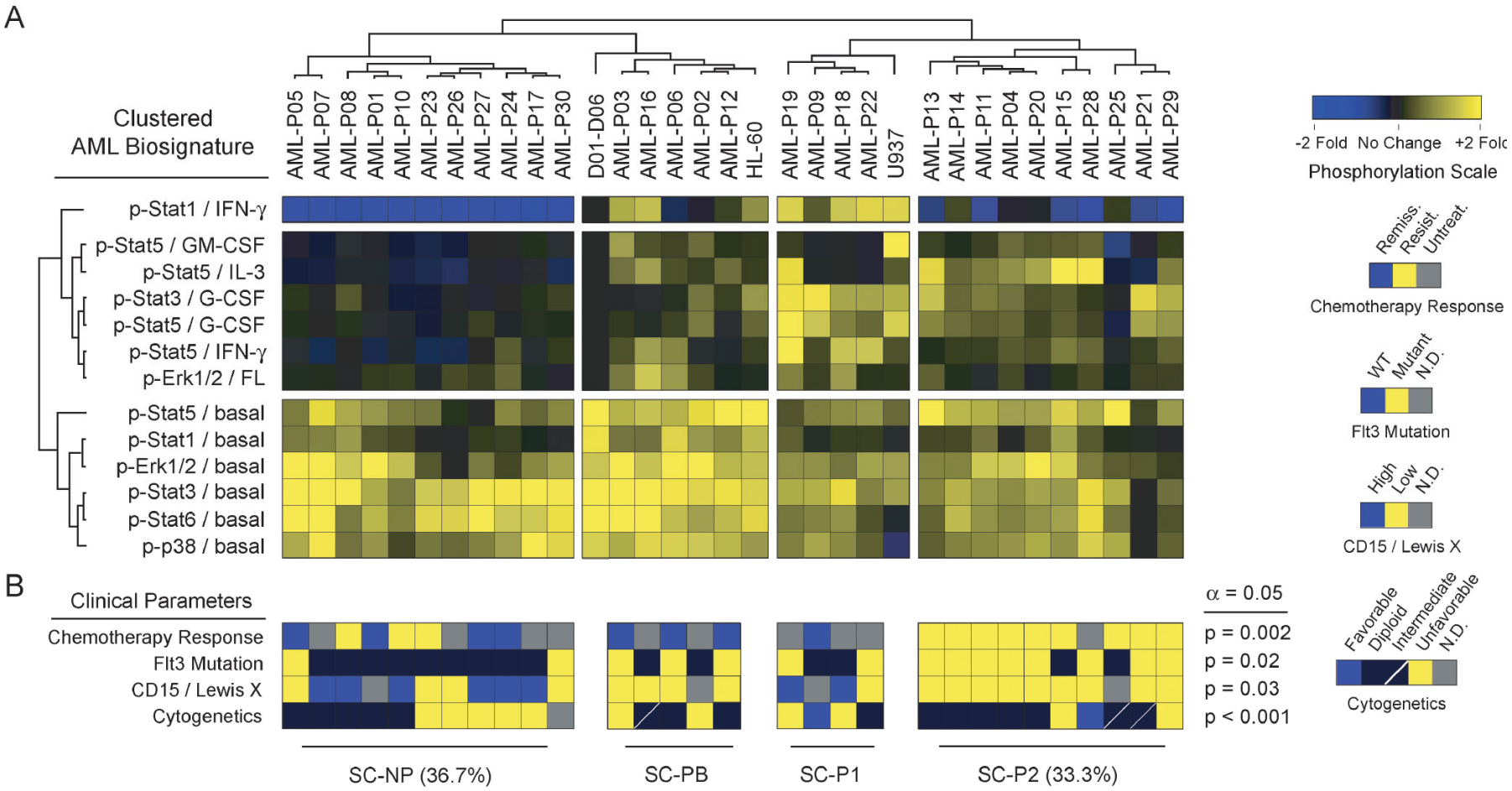
$x' = \text{scale}(x \text{ MFI}) - \text{scale}(\text{control MFI})$

MFI	Fold	Log_{10}	Asinh_{150}
4	1.0	0	0
4	1.0	0	0
4	1.0	0	0
7	1.8	0.2	0.02
4	1.0	0	0
100	25.0	1.4	0.60
-10	?	(error)	-0.09

$$\text{Asinh}_c(x) = \ln\left(\frac{x}{c} + \sqrt{\left(\frac{x}{c}\right)^2 + 1}\right)$$

http://en.wikipedia.org/wiki/Inverse_hyperbolic_function

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



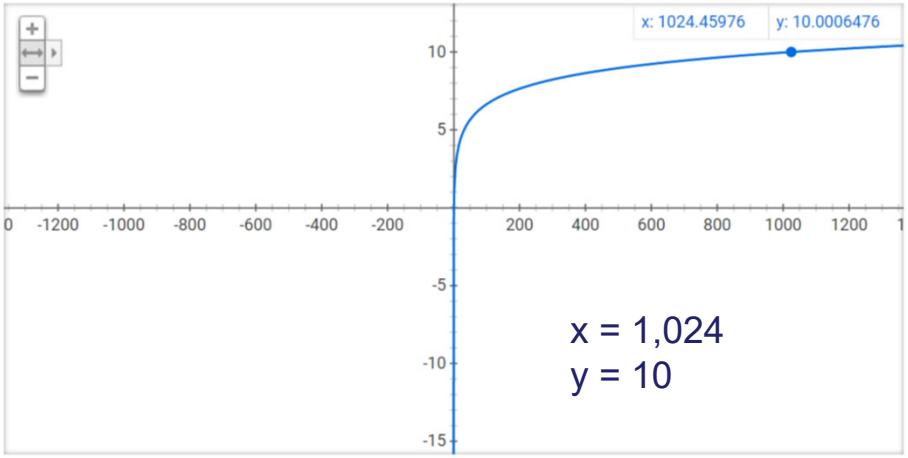
Heatmaps visualize across integrated data types,
e.g. clinical outcomes, cytogenetics, & signaling profiles

Irish, et al *Cell* 2004

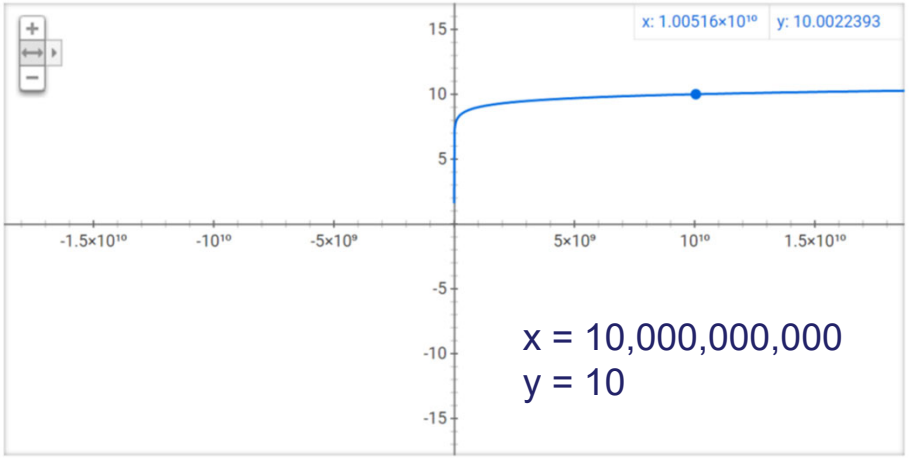
Examples of Four Common Data Scales

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

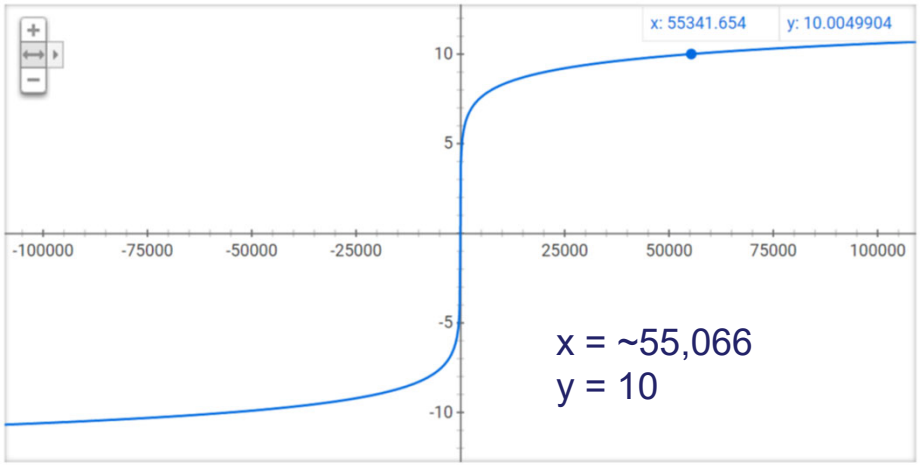
Graph for ln(x)/ln(2)



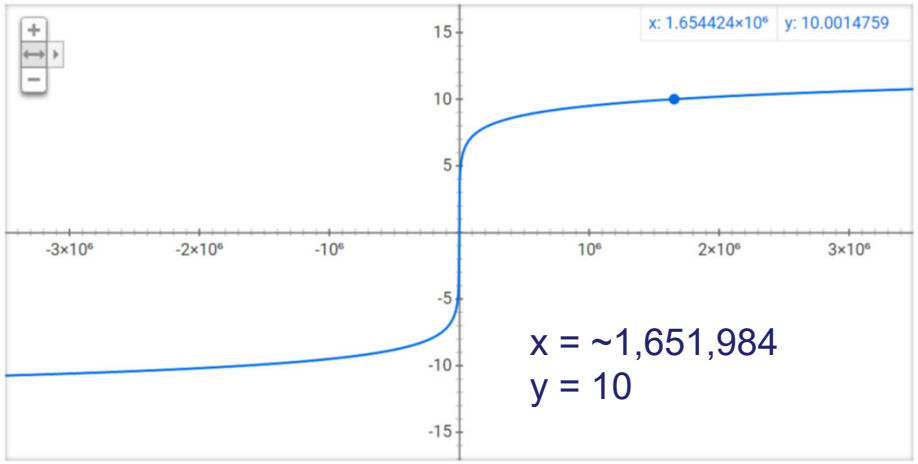
Graph for ln(x)/ln(10)



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

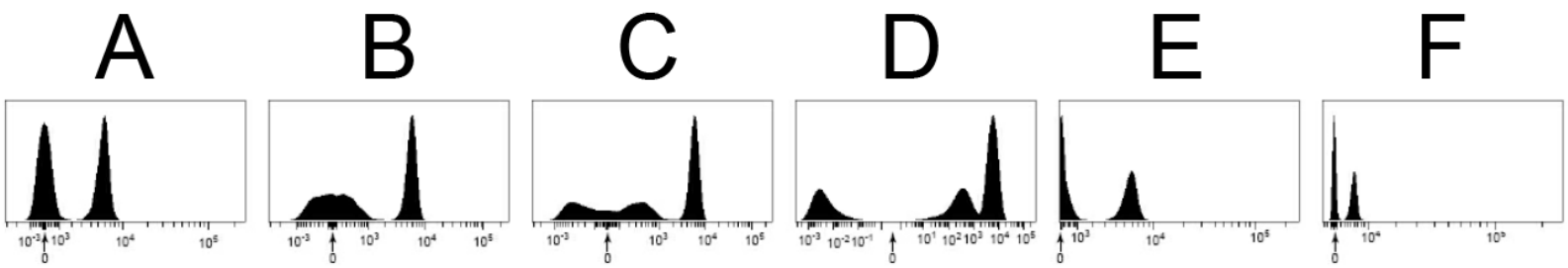


Graph for ln(x/150+sqrt(1+(x/150)^2)) = arcsinh(x/150)



Scaling Matters for Measuring Distance

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

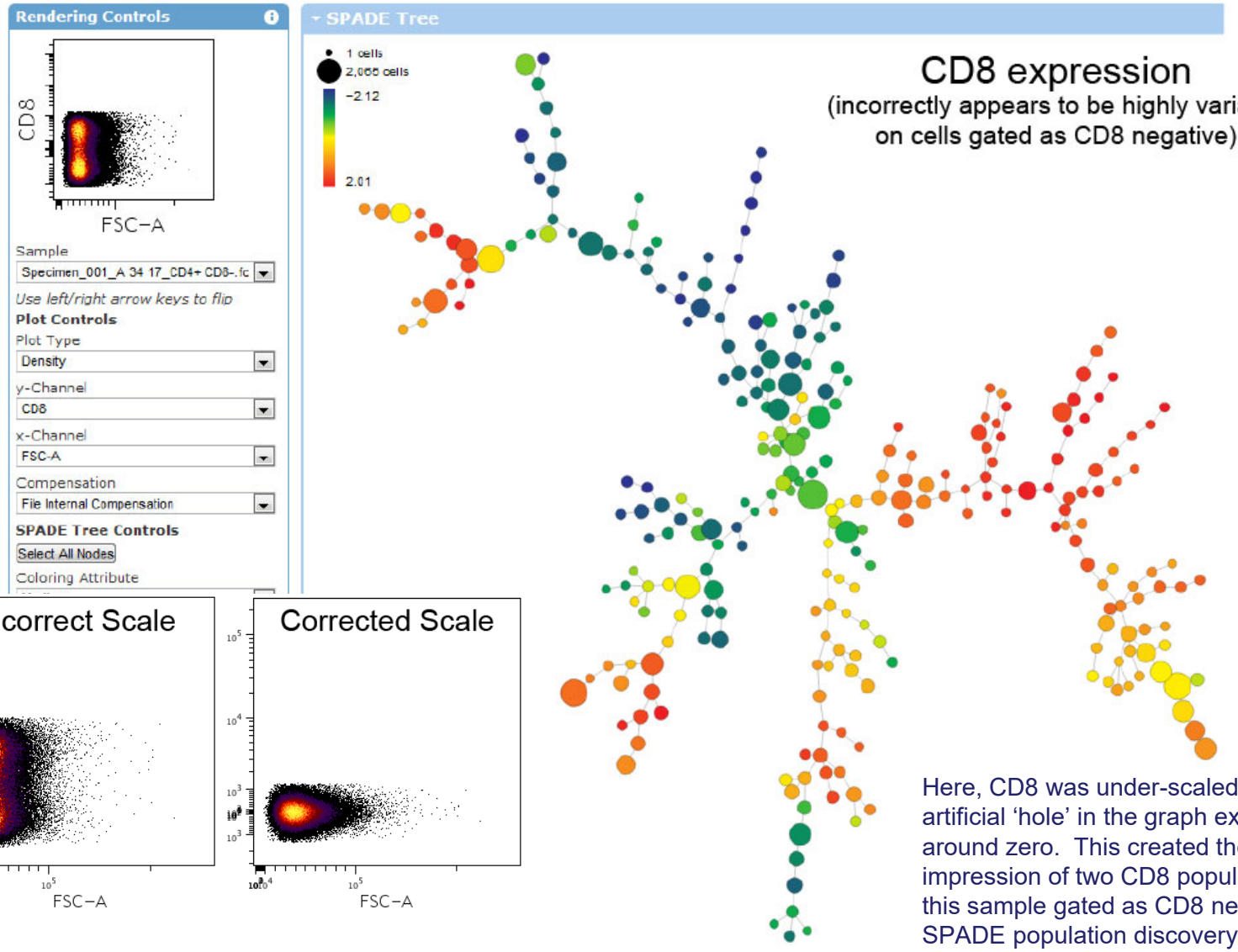
$$\text{arcsinh(x) with cofactor c} = \ln\left(\frac{x}{c} + \sqrt{1 + \left(\frac{x}{c}\right)^2}\right)$$

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

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.

Key Topic Areas and Terms for This Course

1: Field Changes: Data Science & Latest Tools

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

3: Dimensionality Reduction: t-SNE, UMAP, FIt-SNE

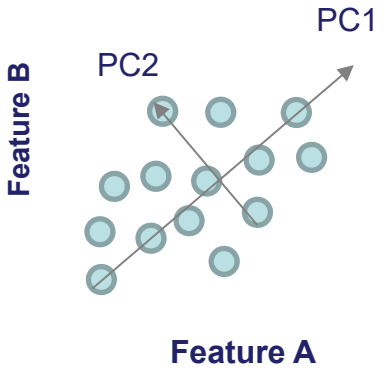
4: Clustering: SPADE, KNN, FlowSOM, Citrus

5: Enriched Features: MEM, ΔMEM, RMSD

6: Cytometry:
- 2004: [Anything] => Heatmap (Irish/Nolan)
- 2011: SPADE => [Anything] (Bendall/Qiu)
- 2013: t-SNE => [Anything] (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 => [Anything] (Van Gassen/Saeys)
- 2017: [Anything] => MEM (Diggins/Irish)
- 2018: UMAP => [Anything] (Newell, McInnes)
- 2019: UMAP => FlowSOM => MEM (Barone/Irish)

More Tools!

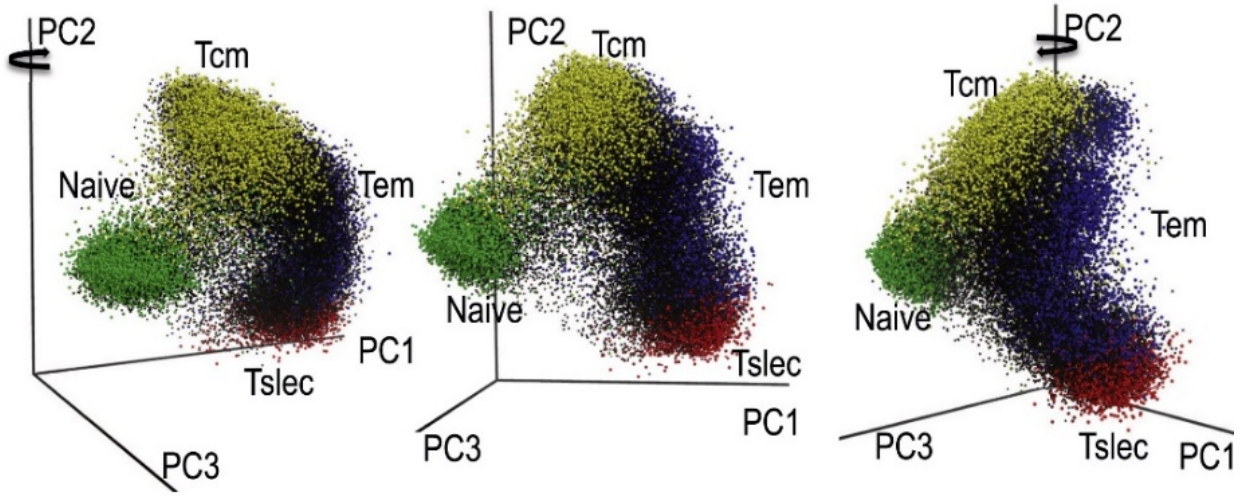
Principal Component Analysis



$$PC1 = X_{A1}A + X_{B1}B + X_{C1}C$$
$$PC2 = X_{A2}A + X_{B2}B + X_{C2}C$$
$$PC3 = X_{A3}A + X_{B3}B + X_{C3}C$$

PCA used to Reduce Dimensionality of CyTOF Data

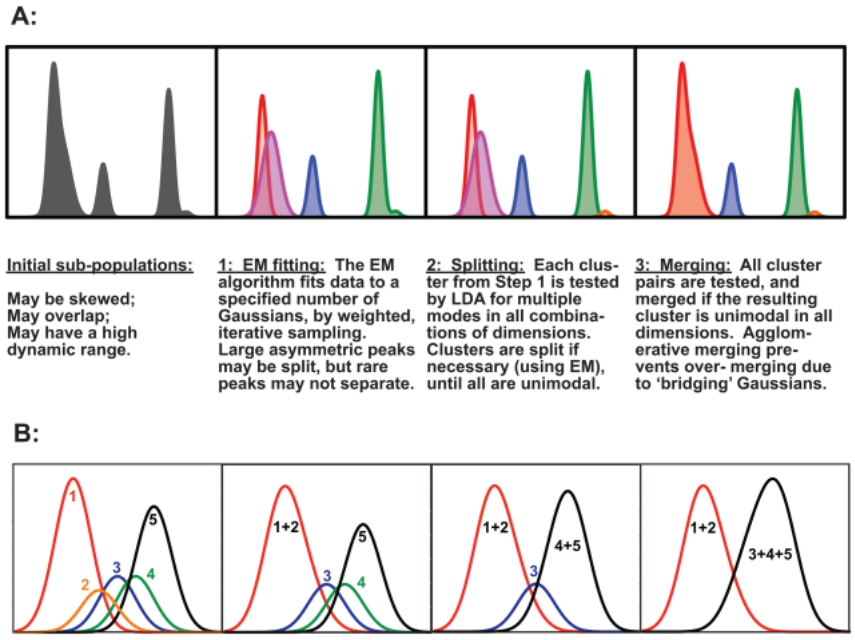
A 3D-PCA view of CD8⁺ T cell 25 parameter data



Newell et al 2012, *Immunity*

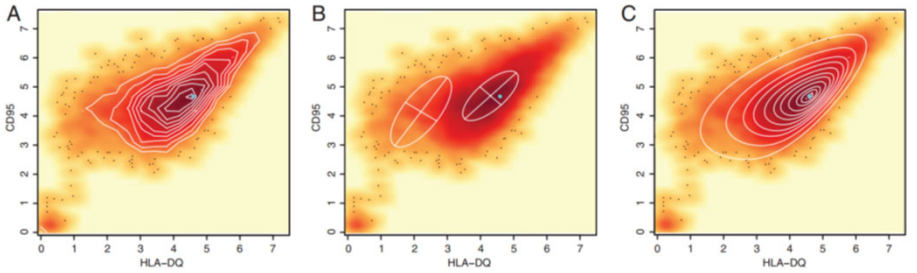
Mixture Modeling

SWIFT



Mosmann et al, 2014 *Cytometry A*

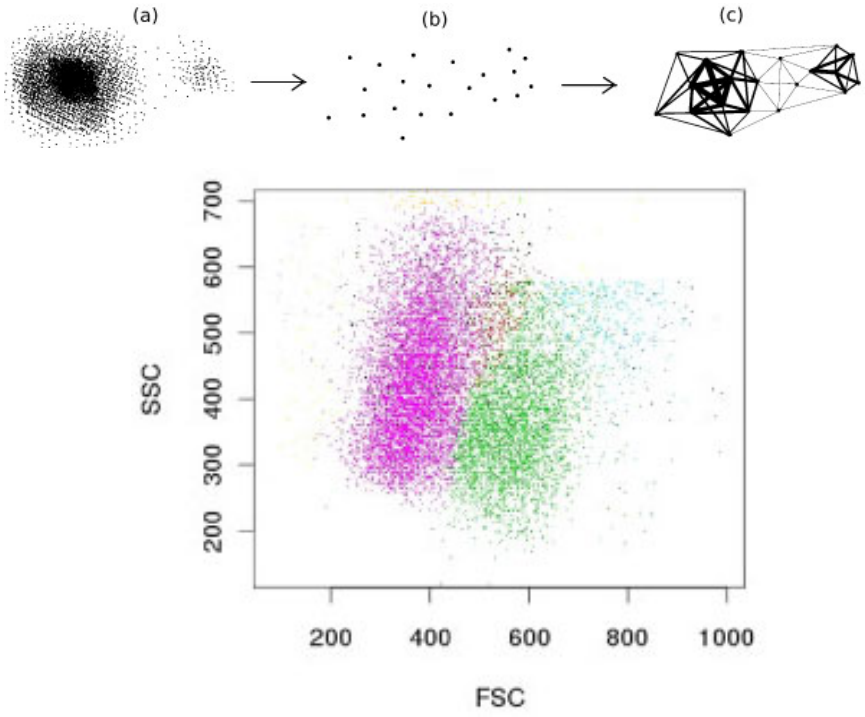
FLAME



Pyne et al, 2009 *PNA*

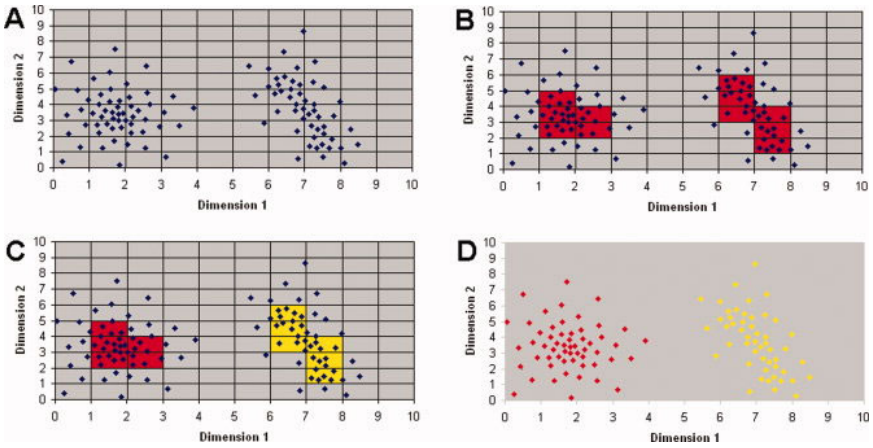
Automated Clustering and Population Identification Methods Based on Density

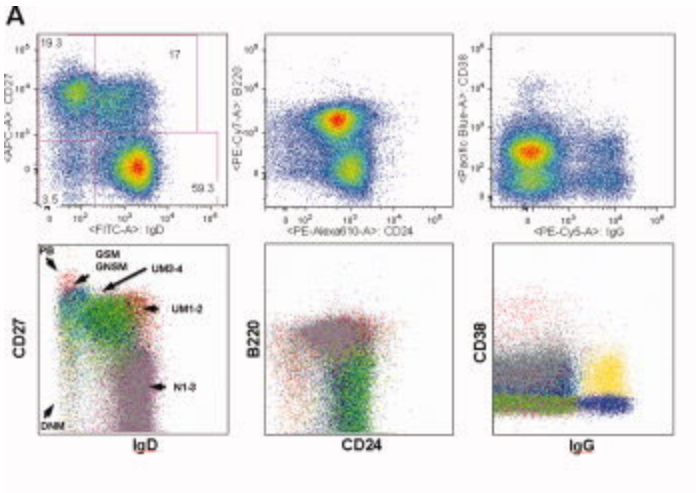
SamSpectral



Zare et al, 2010 *BMC Bioinformatics*

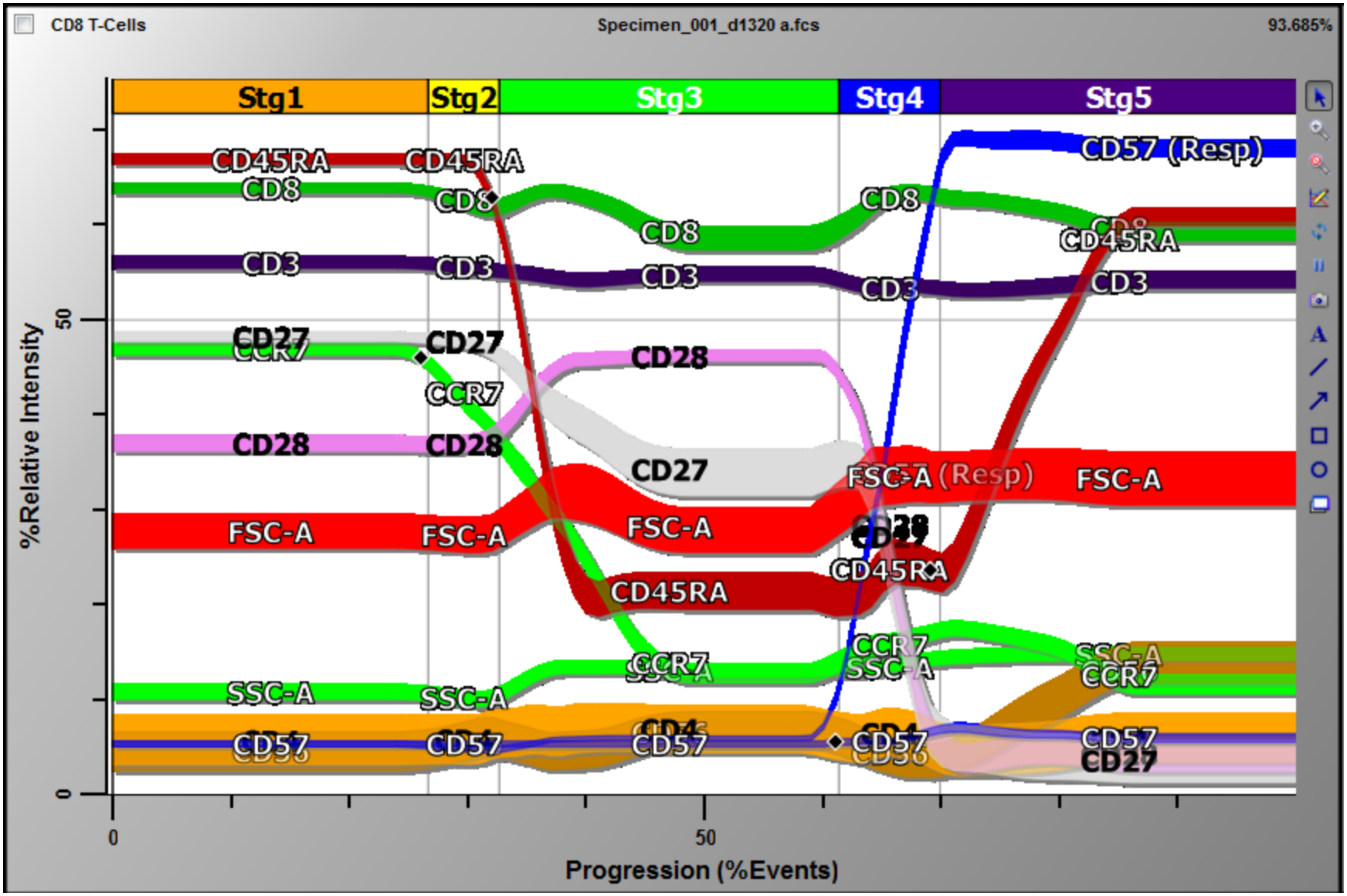
FLOCK





Qian et al, 2010 *Cytometry B Clin Cytom*

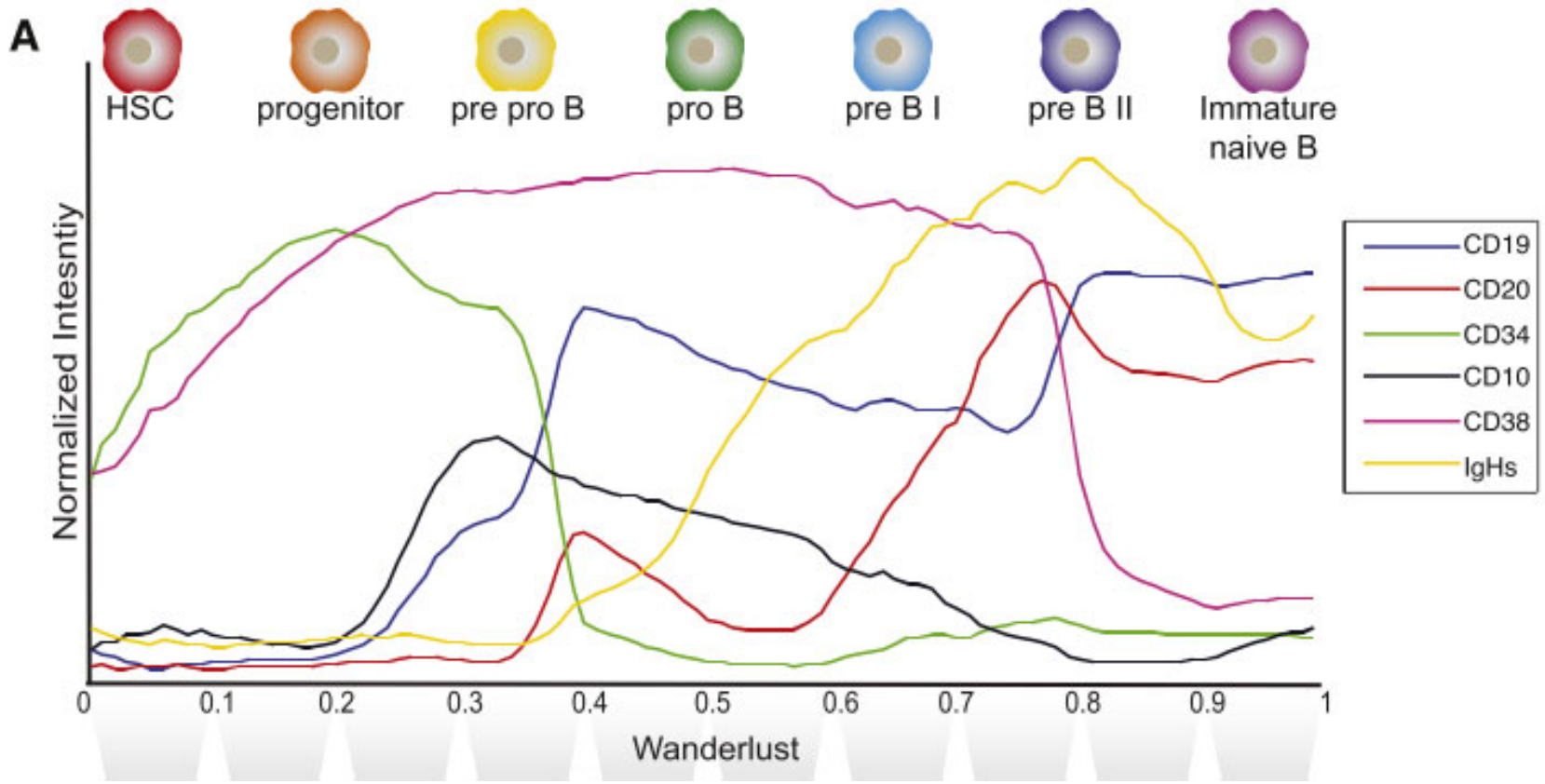
Gemstone Uses Supervised Analysis to Identify Progressions



Verity Software House (www.vsh.com)

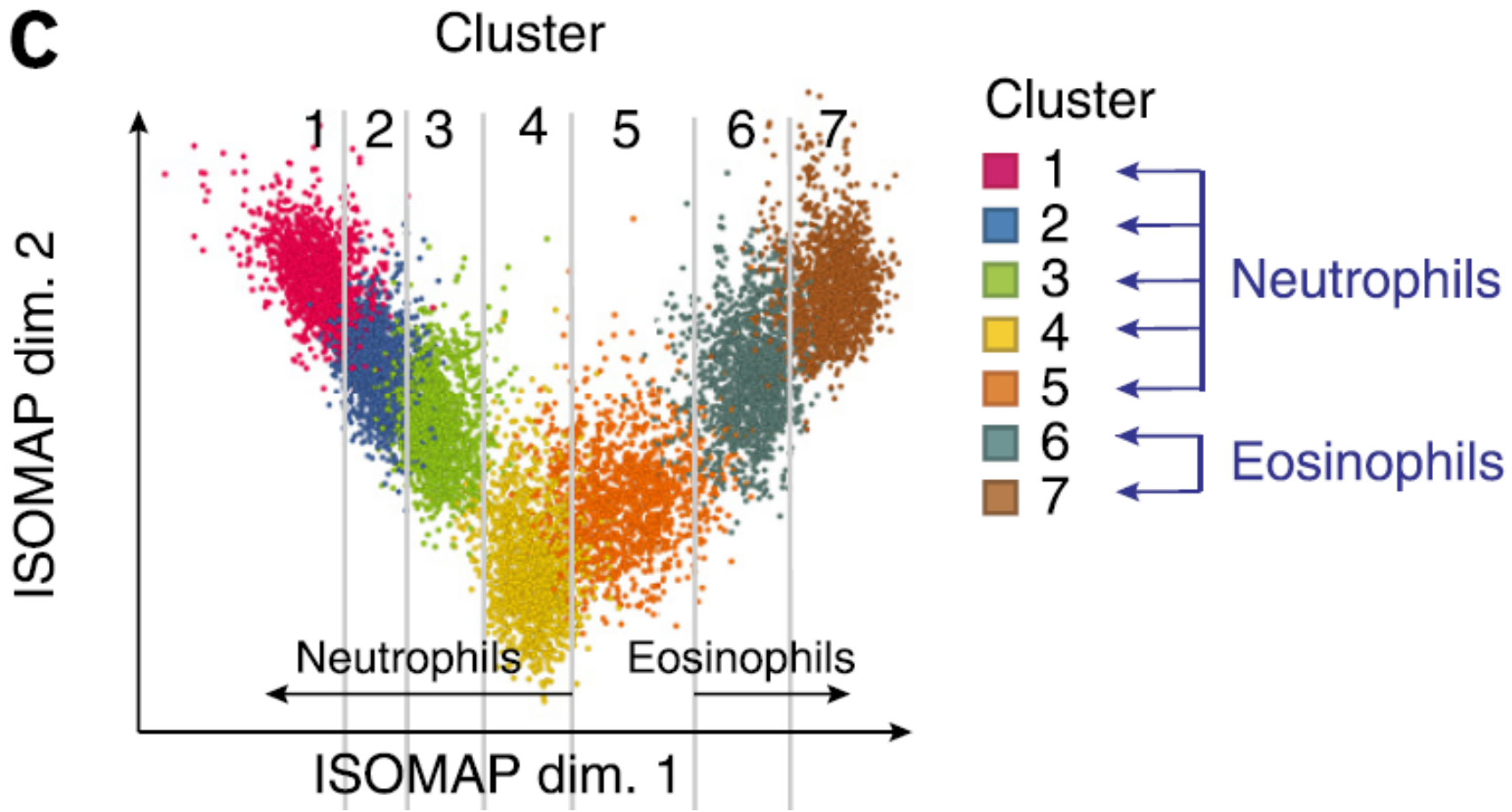
Inokuma, *JIM* 2013

Wanderlust Identifies Phenotypic Progression



Bendall, *Cell* 2014

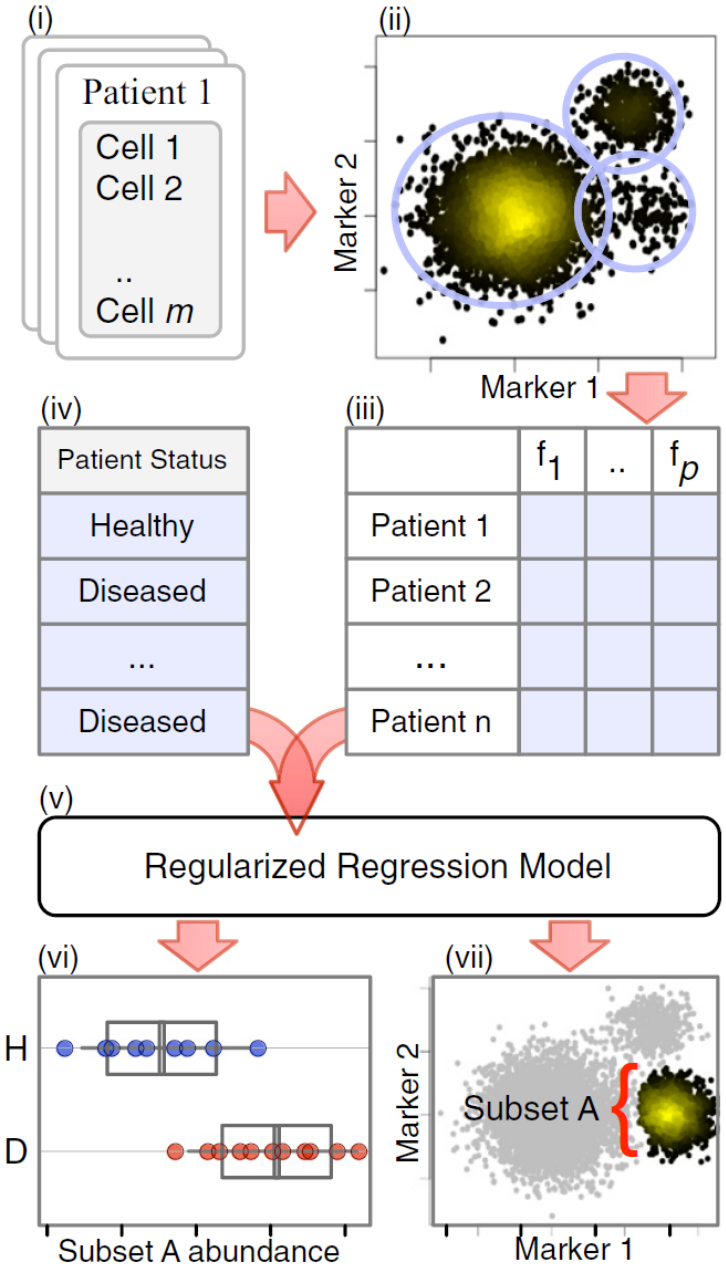
ISOMAP guided analysis



(**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

Citrus: Supervised Population Finding

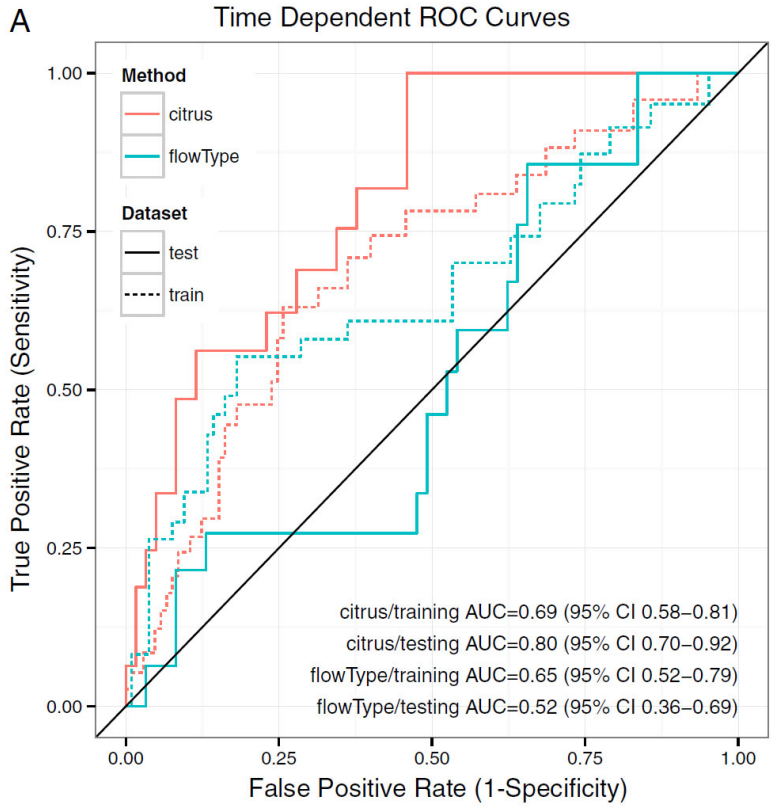


Automated identification of stratifying signatures in cellular subpopulations

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

Final notes & conclusions…

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

1) Need pre-treatment prognosis & prediction

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

2) Need to monitor treatment longitudinally

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

3) Need to check multiple biomarkers with one test

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

4) Need to monitor biomarkers on all cell types

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

5) Need to characterize all cell types to monitor cancer

Evolving cancer cells adopt unexpected phenotypes.

Conclusions: Data Analysis & Mapping Cell Identity

Workflow summary:

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

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

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

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

Acknowledgements & Thank You!

Recent Irish lab Graduate Students

Kirsten Diggins
PhD, Benaroya, Seattle, WA

Deon Doxie
PhD, Emory, Atlanta, GA

Nalin Leelatian
MD/PhD, Yale, New Haven, CT

Allison Greenplate
PhD, U Penn, Philadelphia, PA

Cara Wogsland
PhD, U Bergen, Bergen, Norway

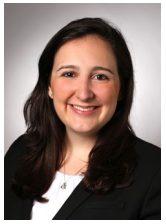
Current Irish Lab

Todd Bartkowiak
PhD Postdoc
K00 Fellow

Caroline Roe
CIC/MCCE
Program Manager

Madeline Hayes
Human Tissue
Res. Assistant

Sierra Barone
Data Science
Res. Assistant



Jocelyn Gandelman
MD, UCSF

Ben Reisman
MD/PhD Student

Chemical Biology Collaborators

+ Ben Reisman & Jordan Froese, PhD

VUMC & VU Collaborators

Madan Jagasia, GVHD, transplant
Pierre Massion, lung cancer
Jeff Rathmell, immune metabolism
Meena Madhur, hypertension immuno.
Brent Ferrell, myeloid cell signaling
Michael Savona, MDS, leukemia
Vivian Gama, neural stem cells
Vito Quaranta+, small cell lung cancer
Ann Richmond, melanoma research
Rebecca Ihrie, brain tumors, stem cells
Ken Lau, epithelial biology
Brian Bachmann, chemical biology
Gary Sulikowski, natural product chem.
Judith Woodfolk (UVA), allergy immuno.
and more… (thank you!)

International Collaborators

Mikael Roussel
MD/PhD, Asst. Prof., Rennes, France

June Myklebust
PhD, Asst. Prof.
Oslo, Norway

Kanutte Huse
PhD, Postdoc
Oslo, Norway

Shahram Kordasti
MD/PhD, Senior Lecturer, KCL & CRUK, London

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

Future Proofing Your Experiments & Files

(1) Tag FCS files (fill in a unique sample, tube, or file name)

- Use text tags that explain what makes each sample unique
- Example: Patient-J01_AtDiag_Panel1
  This might mean: Patient J01, Sampled at diagnosis, Staining Panel 1

(2) Label measured channels

- Format: Target-Detector
- Examples:

CD3-PerCP-Cy5.5	= CD3 on PerCP-Cy5.5
CD4-PacBlue	= CD4 on Pacific Blue
CD10-Gd156	= CD10 on Gadolinium 156
p-STAT5-Ax488	= phosphorylated STAT1 on Alexa488
p-p53-S15-Ax647	= phosphorylated p53 at serine 15 on Alexa647

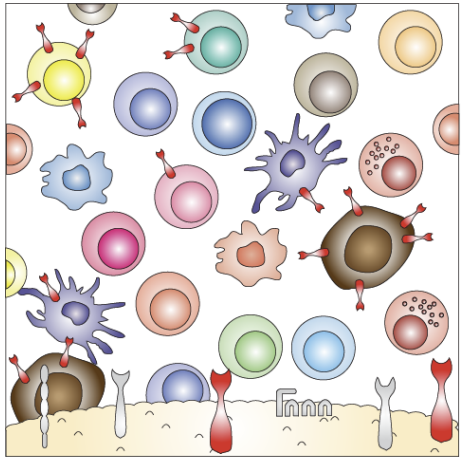
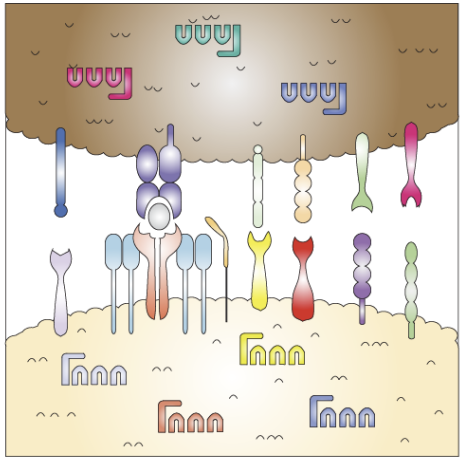
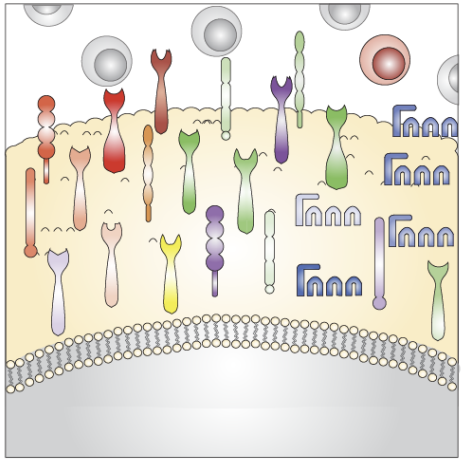
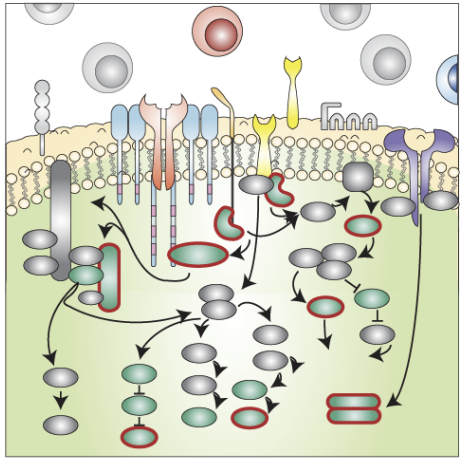
(3) Make sure scales and compensation work before collecting data

- Collect compensation controls & store a comp matrix in files
- Be aware of where peaks fall on the scale before and after compensation

(4) Encode clinical sample IDs (don't use sensitive patient information)

- LP-J001 = First lymphoma patient from "Dataset J"

Tracking, Targeting, and Changing Cell Identity
(If you liked this: CYTO 2019 is in Vancouver from June 22-26)

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

Critical opportunities to improve clinical research:

1) Increase knowledge gained from each case, each biopsy, each trial
2) Accelerate discovery to treatment, learn *in vivo* from humans

Greenplate et al., *Euro J Cancer* 2016

List of detectable single cell features @ level (CyTOF & p-flow): Doxie & Irish, *Curr Topics in Micro & Immuno* 2014

Methods for comparing analysis tools: Diggins et al., *Methods* 2015

Tracking, Targeting, and Changing Cell Identity
(If you liked this: CYTO 2019 is in Vancouver from June 22-26)

1. Cell functional assays using living cells from humans.
   Are the DN T cells suppressive? Clonal? Were they ever CD4s/CD8s?

2. Tractable, immunocompetent, human cell systems.
   Cell signaling interactions in neural development, cell identity, cancer.

3. Selectively shifting cell identity (or at least key features).
   In situ: kill or differentiate cancer cells, upregulate MHC, renew T cells.

4. Unsupervised approaches; annotated reference cells.

5. June 22-26, 2019 – CYTO / ISAC meeting, come see:
   - Single cell signaling profiles stratify glioblastoma patient clinical risks (Jonathan Irish)
   - Training protocol for mass cytometry phospho-flow (Caroline Roe)
   - Neuroimmunology in brain tumors and stem cell niches (Todd Bartkowiak)
   - Next generation barcoding for single cell screening (Ben Reisman),
   - Single cell gene editing of BCR signaling mechanisms (Kanutte Huse),
   - Human immune monitoring in sepsis (Aida Meghraoui).

Cytometry: Measure ~Anything in a Cell

- 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, cytokines
- 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

Adapted from Irish, Kotecha, and Nolan, *Nature Reviews Cancer* 2006
Revised for CyTOF in Doxie & Irish, *Current Topics in Microbiology* 2014

Many Great Tools Exist, But Key Gaps Remain

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

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

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

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

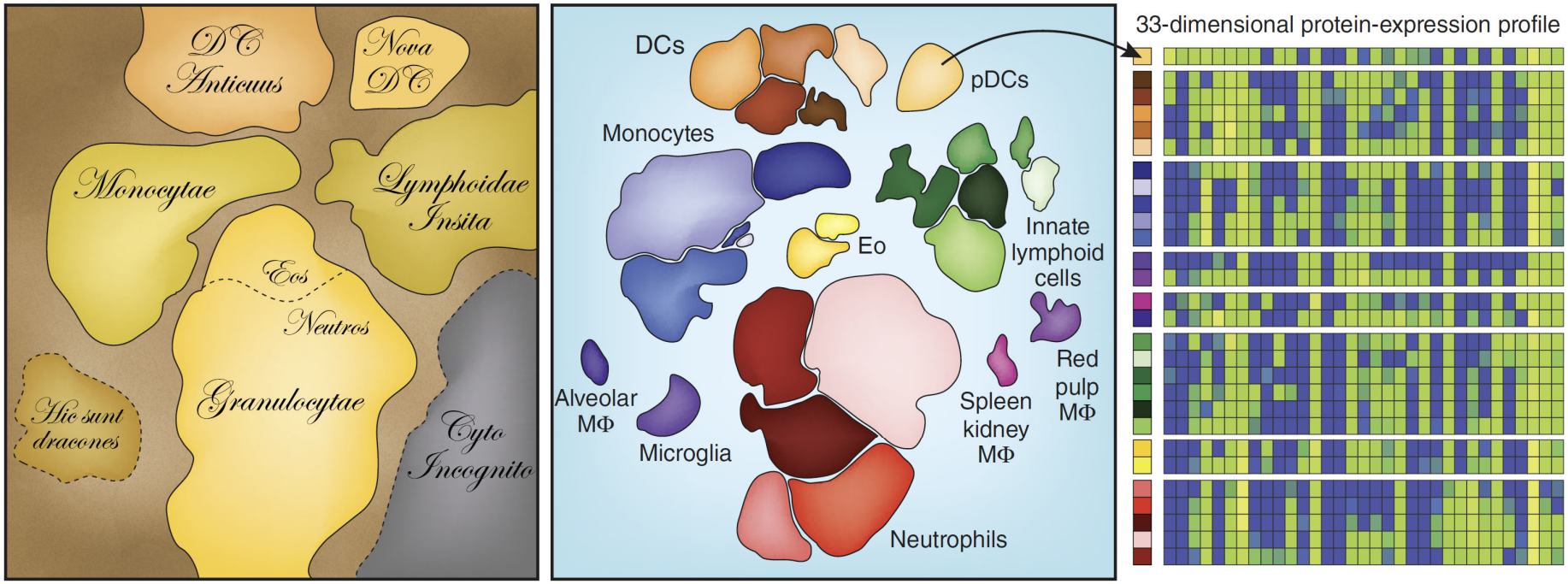
A major gap in the field is in true learning of cell identity

Diggins et al., *Methods* 2015

Unsupervised Analysis: Automatic, Comprehensive, and Less Biased

Classic myeloid system

Modern quantitative map of cellular phenotypes



Classically (left), the myeloid cell system has been poorly described due to inherent heterogeneity and disagreement on key markers, nomenclature and population boundaries. Traditional analysis approaches identify major populations using canonical markers and may leave unexplored cell subsets with unexpected phenotypes ('Cyto Incognito'). In contrast, machine learning (middle) provides a more comprehensive view that automatically resolves a two-dimensional map of cell types from high-dimensional single-cell data, which can be read by humans. Phenotypic distances between cells are measured and cells are arranged so that proximity indicates similarity. Cell groups from various tissues and conditions are characterized by distinct protein features, as in the heat maps used by Becher *et al*. (right)[5]. In this heat map, color represents the average expression of one of the 33 proteins (columns) in one of the 28 automatically identified cell populations (rows). For example, the expression profile for plasmacytoid DCs (pDCs) is presented in the top row. Applied to myeloid system cells from eight tissues, this approach reveals previously underappreciated populations of tissue-resident macrophages and identifies phenotypically distinct subpopulations of heterogeneous DC, monocyte, neutrophil and eosinophil (Eo) populations. Unexpected phenotypes that would probably have been overlooked in a focused, classic study are revealed. For example, *Csf2rb*$^{-/-}$ mice lack tissue-resident natural killer cells, a subgroup of the innate lymphoid population. This observed absence of innate lymphoid cells is picked up by a mass cytometry panel designed to study the myeloid compartment.

Irish, *Nature Immunology* 2014

Becher et al., *Nature Immunology* 2014