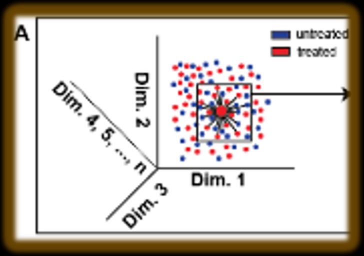
Nearest neighborhood-based comparisons across biological conditions in single cell data

2 February 2018

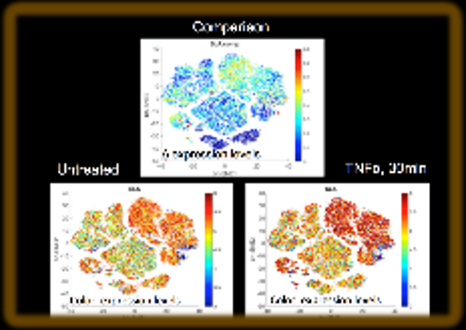
Tyler J Burns, PhD
AG Mei at DRFZ

Outline

Building per-cell k-nearest neighborhoods in high-D space

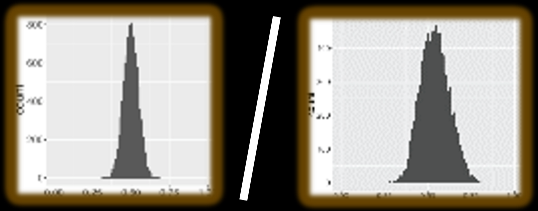


Making single-cell comparisons across t-SNE maps

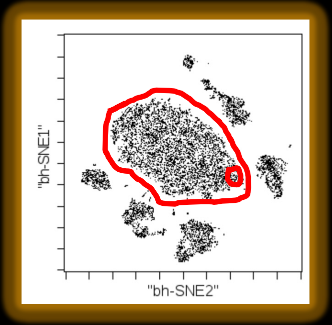


Establishing an evaluation metric for data quality

m =

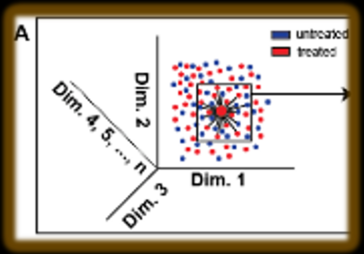


Evaluating the fidelity of lower-dimensional embeddings

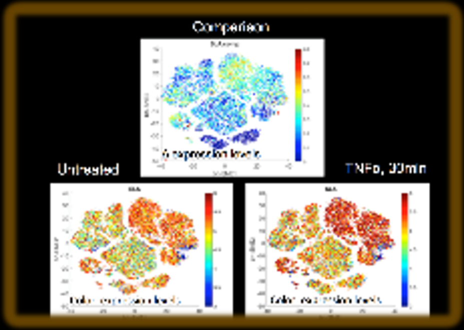


Outline

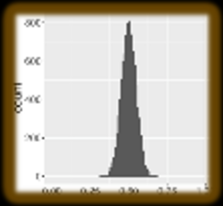
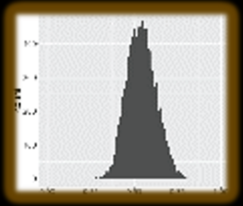
Building per-cell k-nearest neighborhoods in high-D space



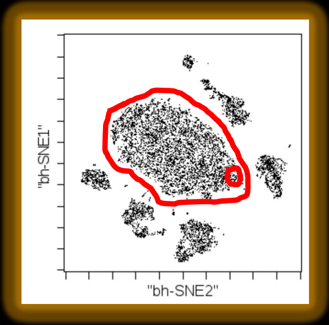
Making single-cell comparisons across t-SNE maps



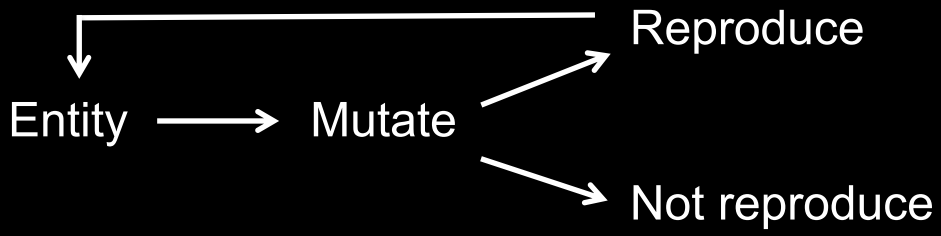
Establishing an evaluation metric for data quality

m =  / 

Evaluating the fidelity of lower-dimensional embeddings



Biodiversity exists between organisms and between cells



Entity

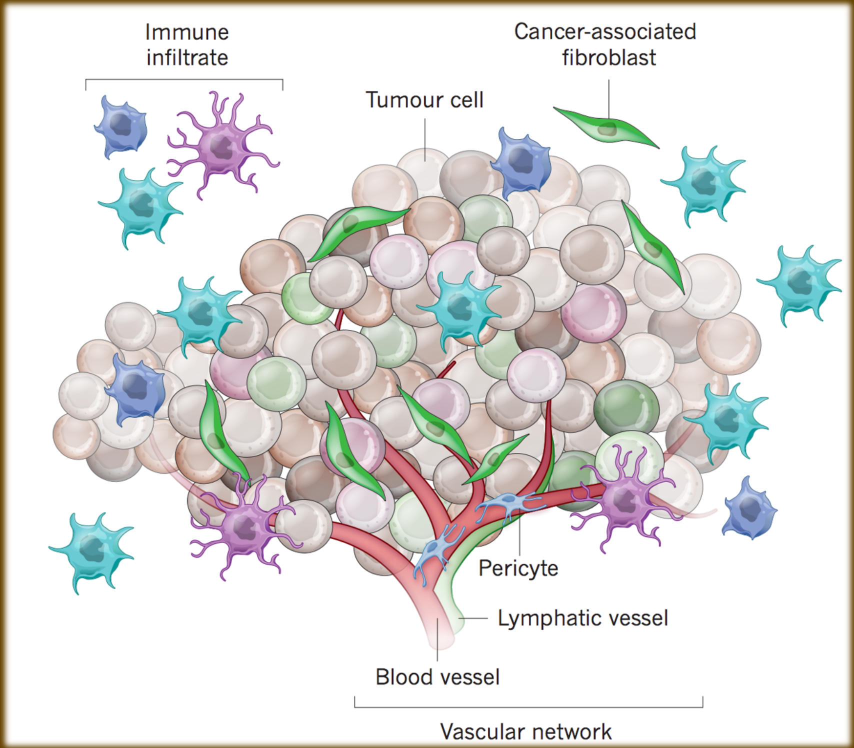
Mutate

Reproduce

Not reproduce

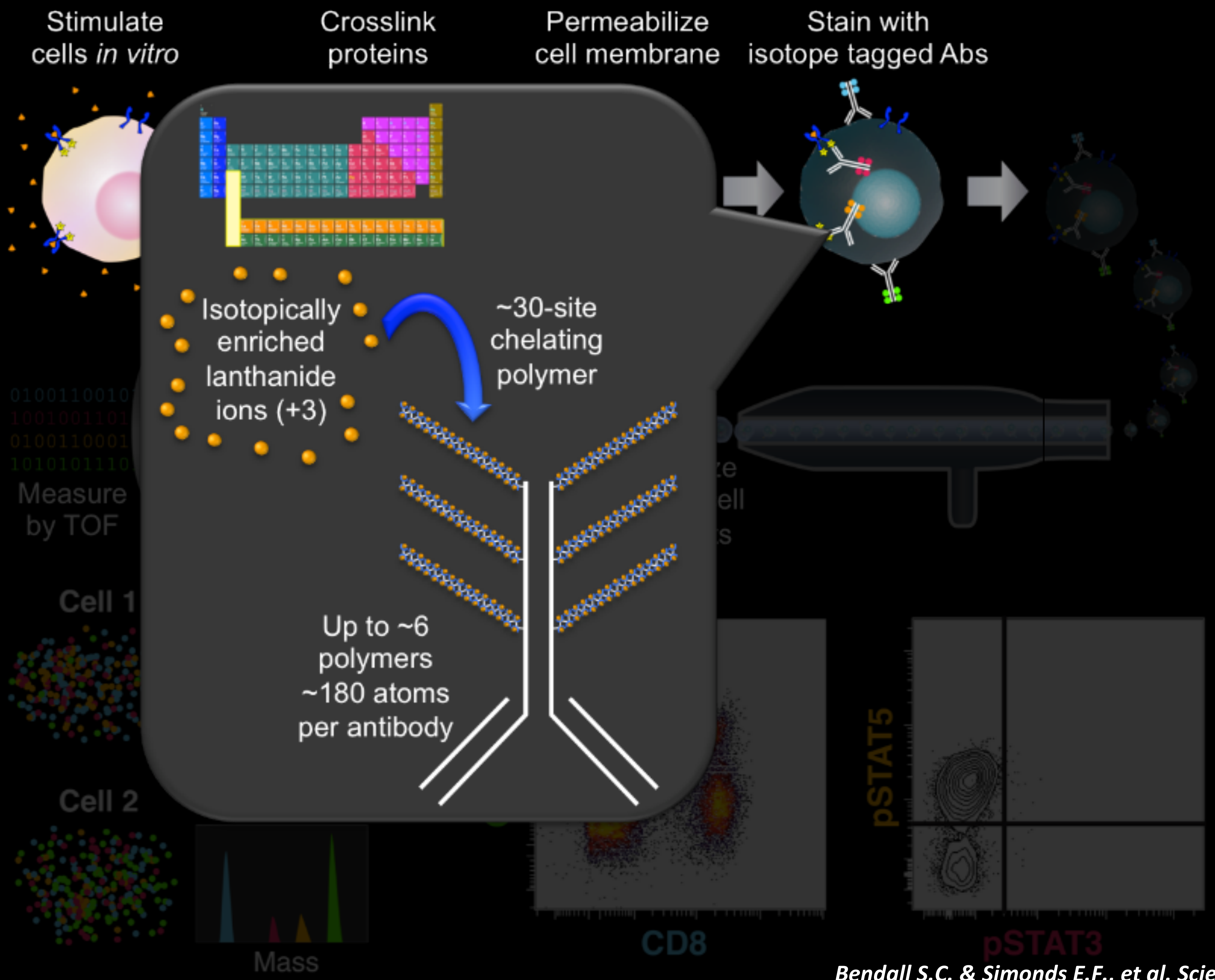
Organismal biodiversity

Single cell biodiversity



Slide adapted from Sean Bendall

Mass cytometry is a powerful technique for single-cell analysis



Stimulate cells *in vitro*

Crosslink proteins

Permeabilize cell membrane

Stain with isotope tagged Abs

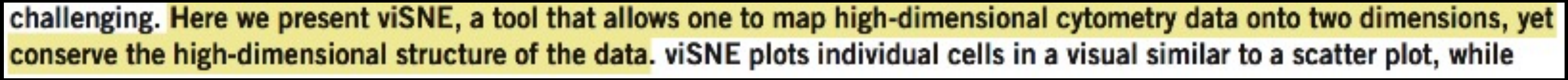
Isotopically enriched lanthanide ions (+3)

~30-site chelating polymer

Up to ~6 polymers ~180 atoms per antibody

Bendall S.C. & Simonds E.F., et al. Science (2011)

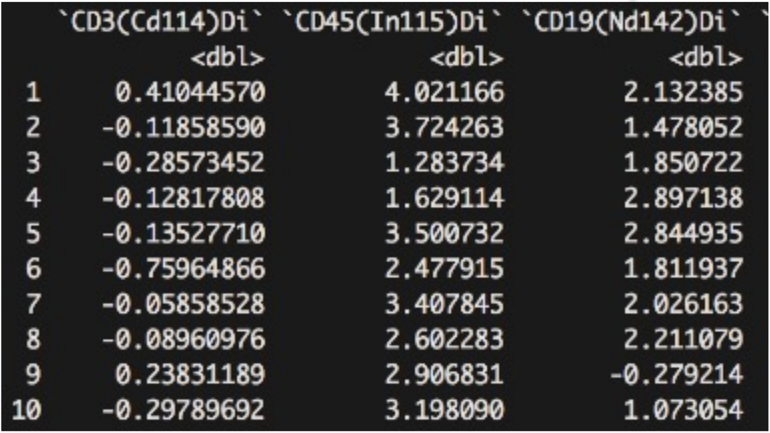
Dimension reduction algorithms (eg. t-SNE) map high-dimensional data to two dimensions



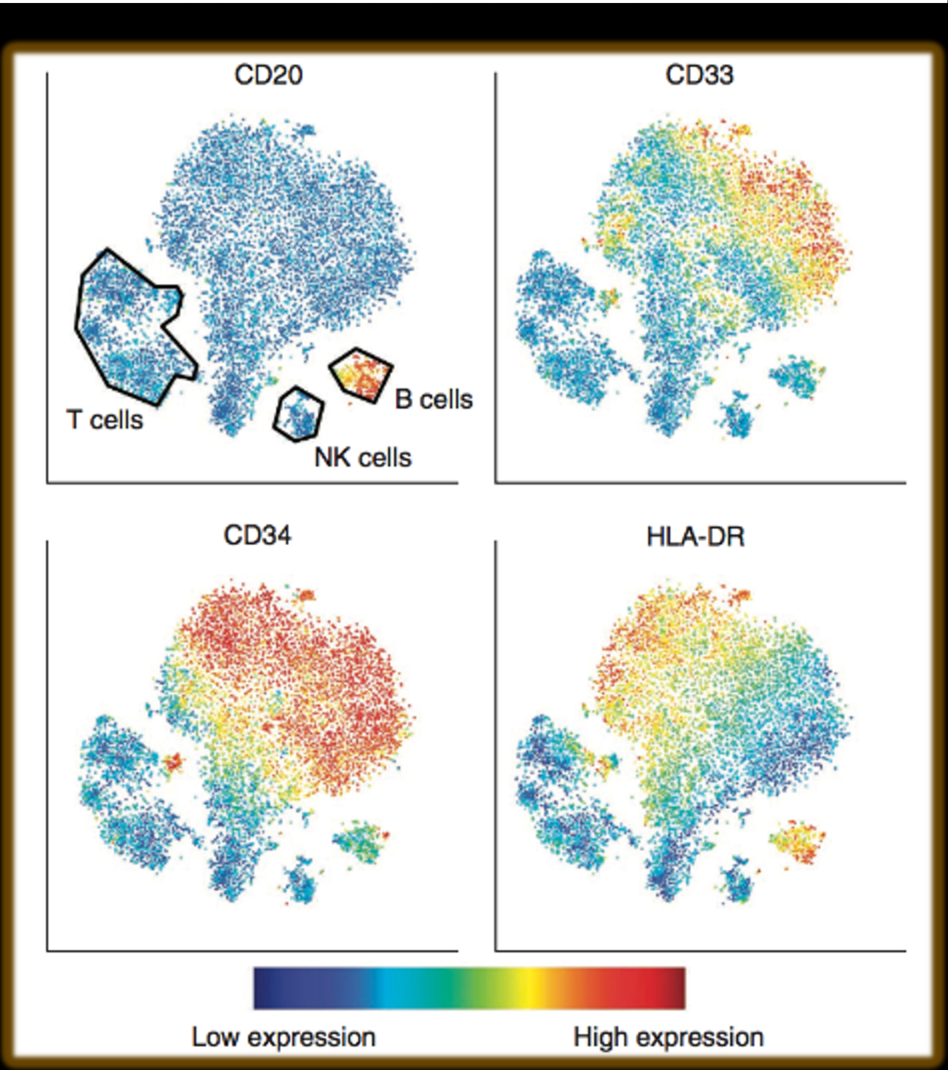
challenging. Here we present viSNE, a tool that allows one to map high-dimensional cytometry data onto two dimensions, yet conserve the high-dimensional structure of the data. viSNE plots individual cells in a visual similar to a scatter plot, while

Features (30-50)

Cells (10^4 – 10^6+)

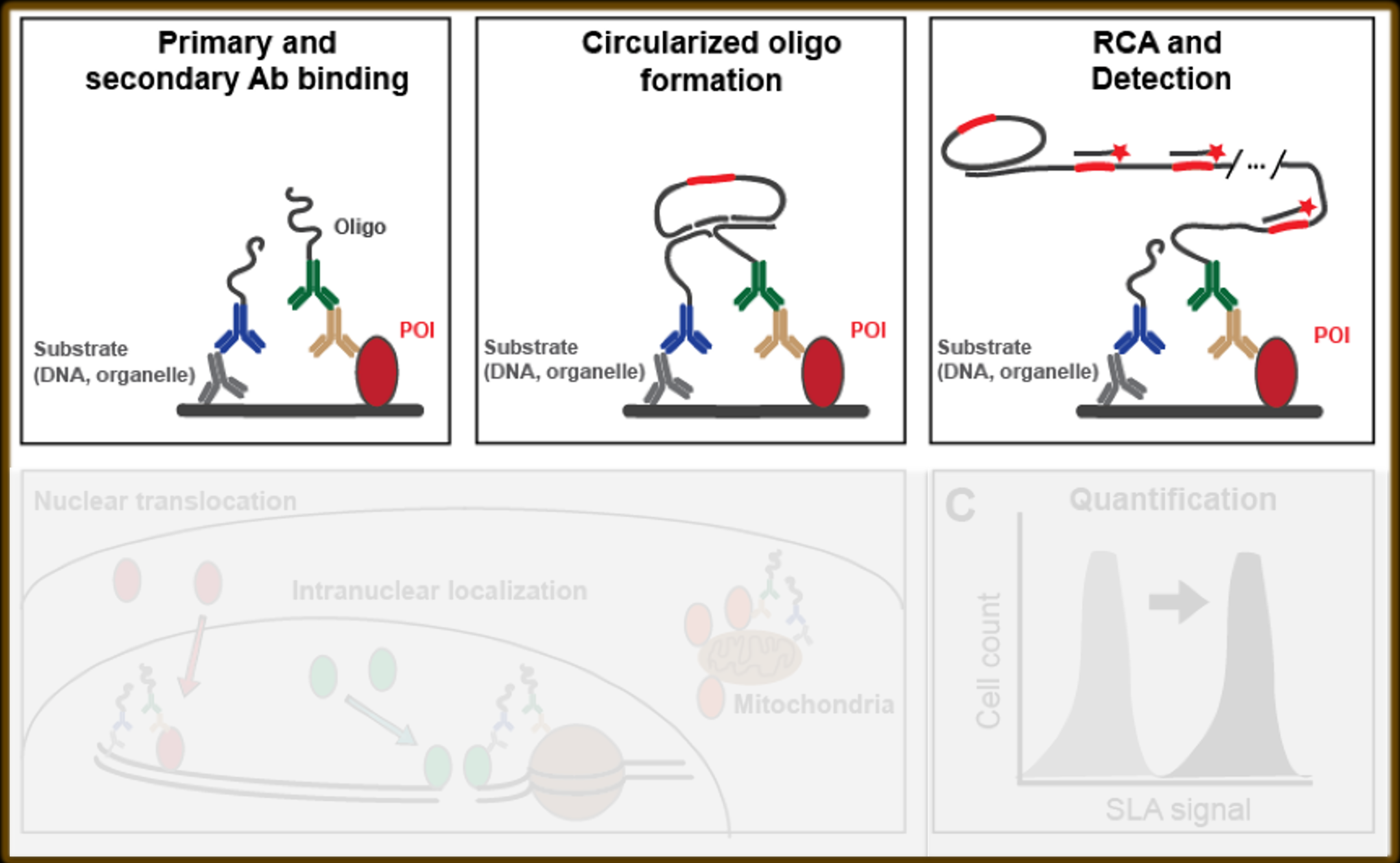


```
   `CD3(Cd114)Di` `CD45(In115)Di` `CD19(Nd142)Di` `
            <dbl>           <dbl>           <dbl>
 1     0.41044570        4.021166        2.132385
 2    -0.11858590        3.724263        1.478052
 3    -0.28573452        1.283734        1.850722
 4    -0.12817808        1.629114        2.897138
 5    -0.13527710        3.500732        2.844935
 6    -0.75964866        2.477915        1.811937
 7    -0.05858528        3.407845        2.026163
 8    -0.08960976        2.602283        2.211079
 9     0.23831189        2.906831       -0.279214
10    -0.29789692        3.198090        1.073054
```



Amir *et al, Nat. Biotech* 2013

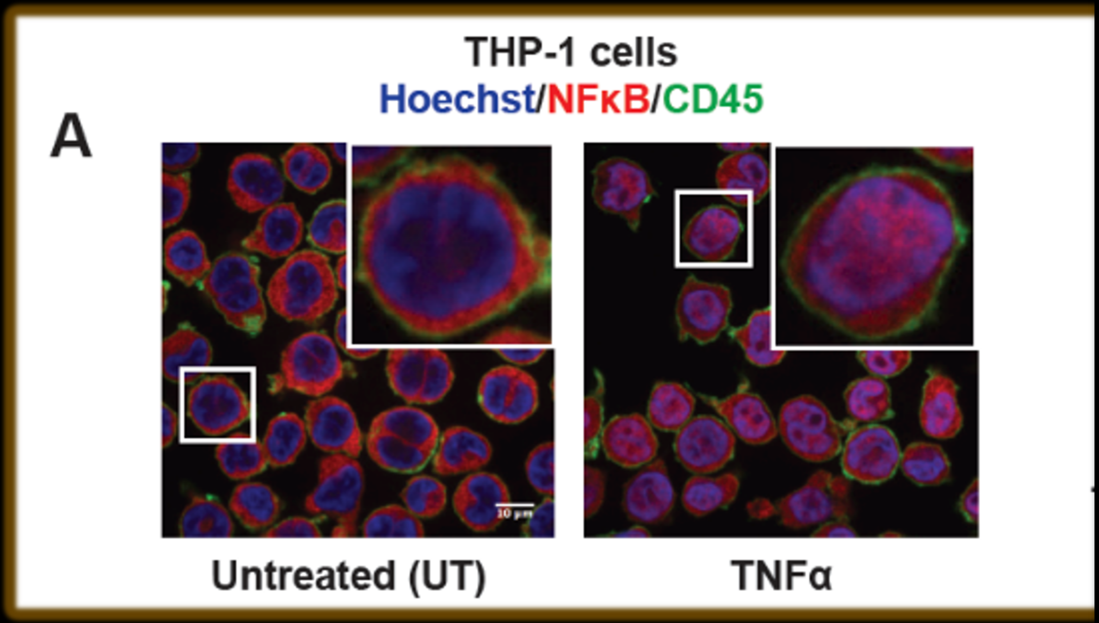
Subcellular Localization Assay brings visual-spatial information to flow and mass cytometry


Primary and secondary Ab binding
Oligo
POI
Substrate (DNA, organelle)
Circularized oligo formation
POI
Substrate (DNA, organelle)
RCA and Detection
POI
Substrate (DNA, organelle)
Nuclear translocation
Intranuclear localization
Mitochondria
C
Quantification
Cell count
SLA signal

Burns et al, *Cytometry* 2017

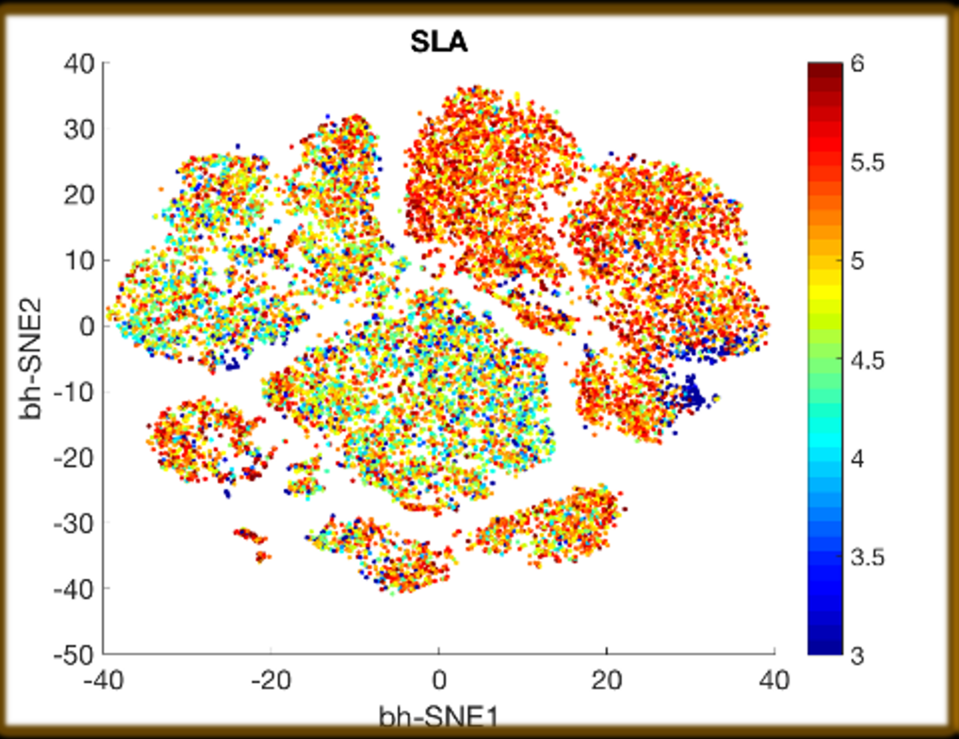
Nuclear import of NF-kB can be visualized with flow cytometry

Confocal microscopy

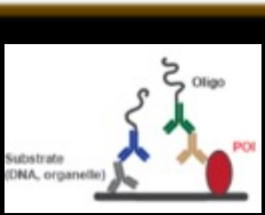


SLA applied to mass cytometry requires comparison of colored t-SNE maps

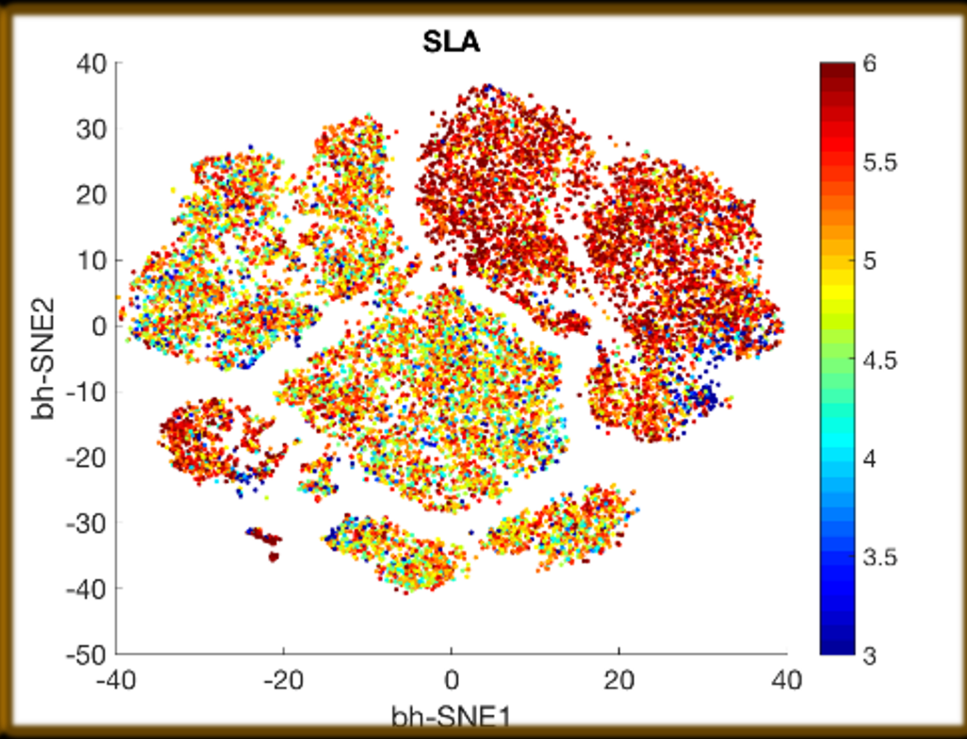
Untreated



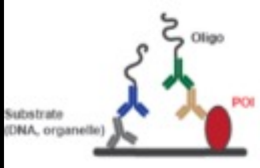
Color:
Nuclear NF-kB



TNFα, 30min

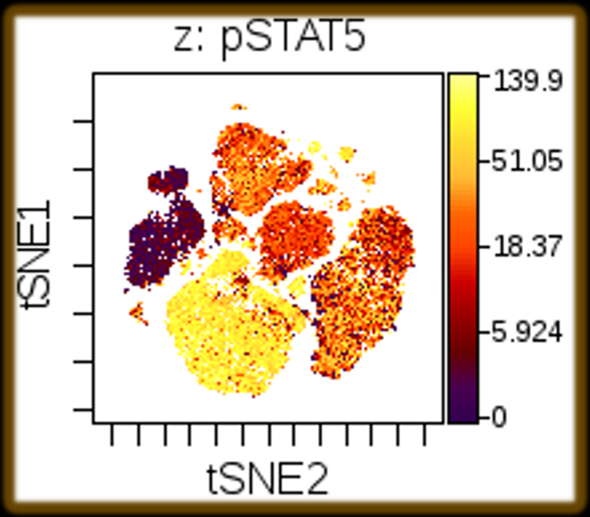


Color:
Nuclear NF-kB



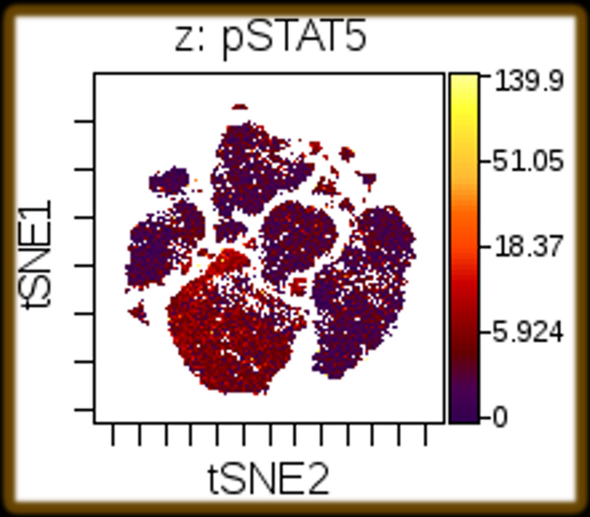
One solution: Pixel color value subtraction of t-SNE maps

IL-2

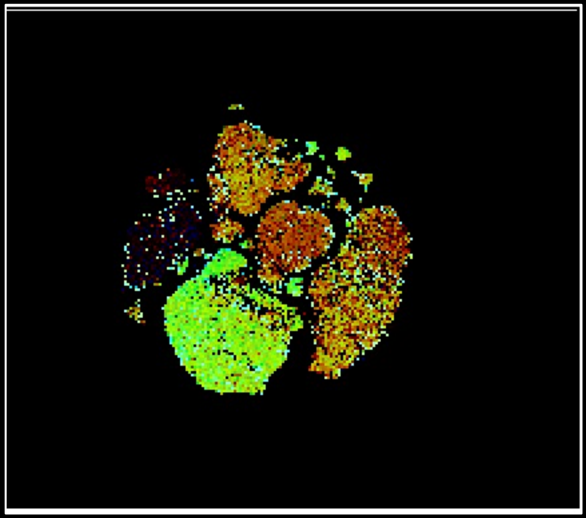


-

Untreated



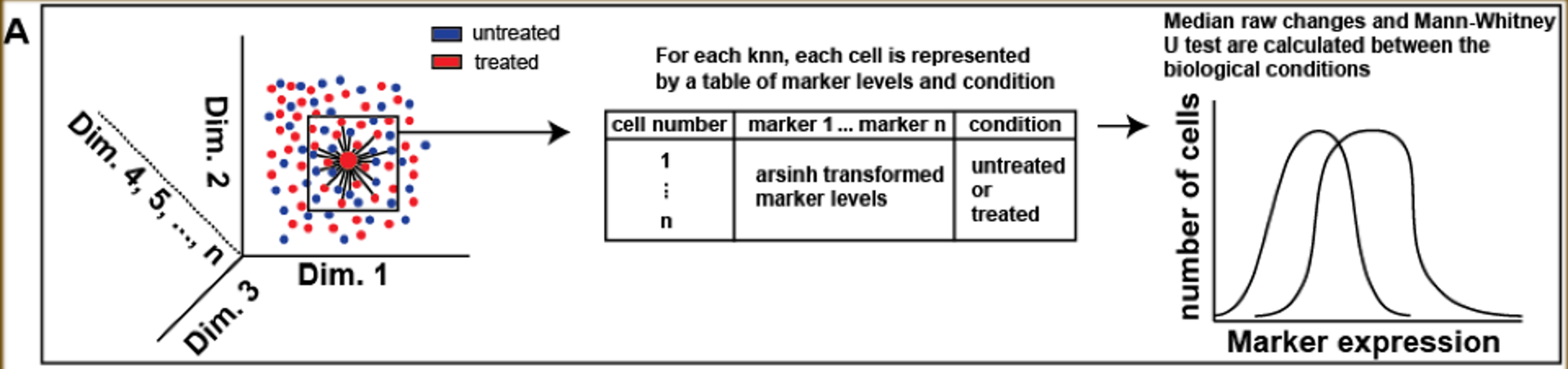
=



Each pixel = (red 1-255, green 1-255, blue 1-255)
subtract image 2 from image 1 pixel by pixel

Yellow = significant increase(yellow – black)
Green = moderate increase(yellow – red)
Red = small increase(red – black)
Black = no difference (any – any)

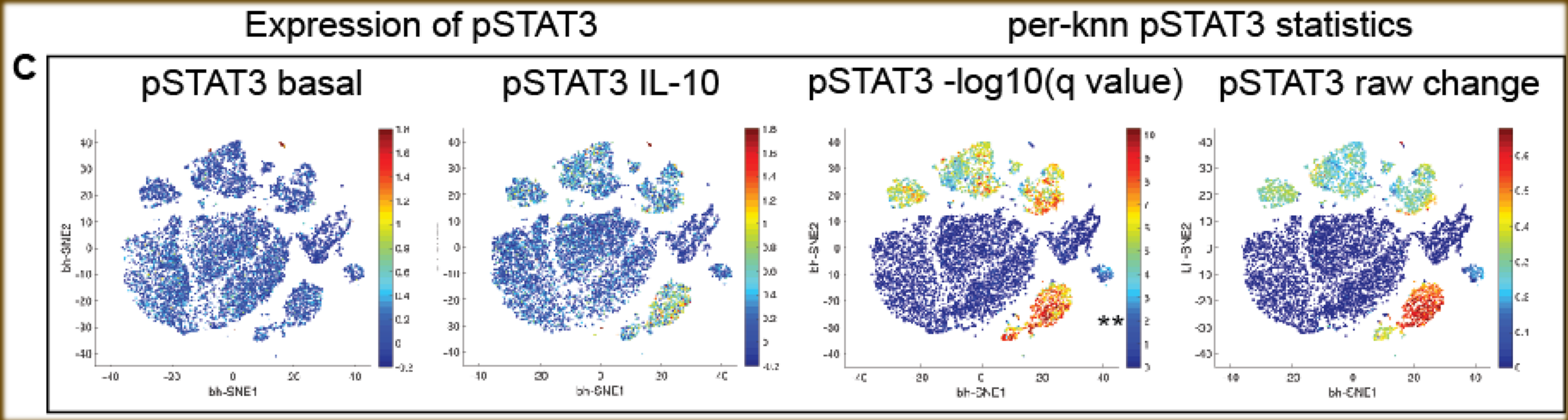
My solution: **S**mooth **C**omparisons **O**ver nearest **Ne**ighbors (SCONE)



For each knn, each cell is represented by a table of marker levels and condition

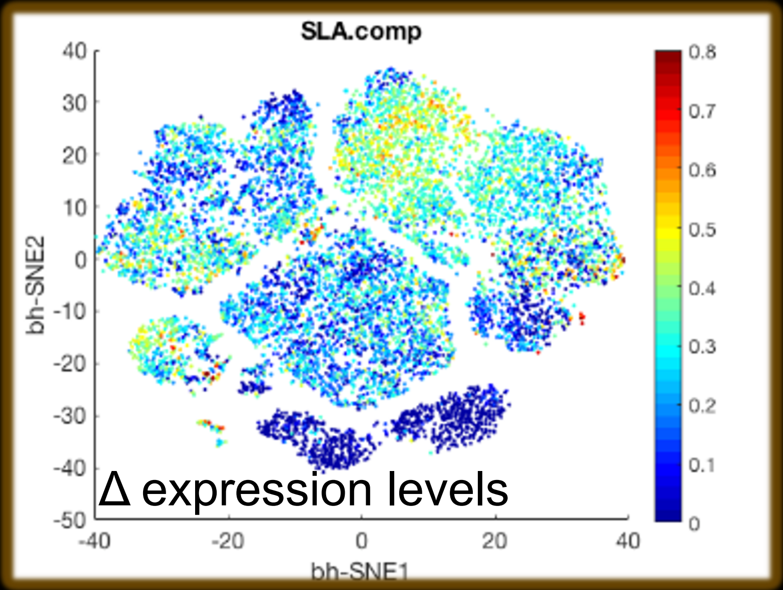
cell number	marker 1 ... marker n	condition
1 ⋮ n	arsinh transformed marker levels	untreated or treated

Median raw changes and Mann-Whitney U test are calculated between the biological conditions



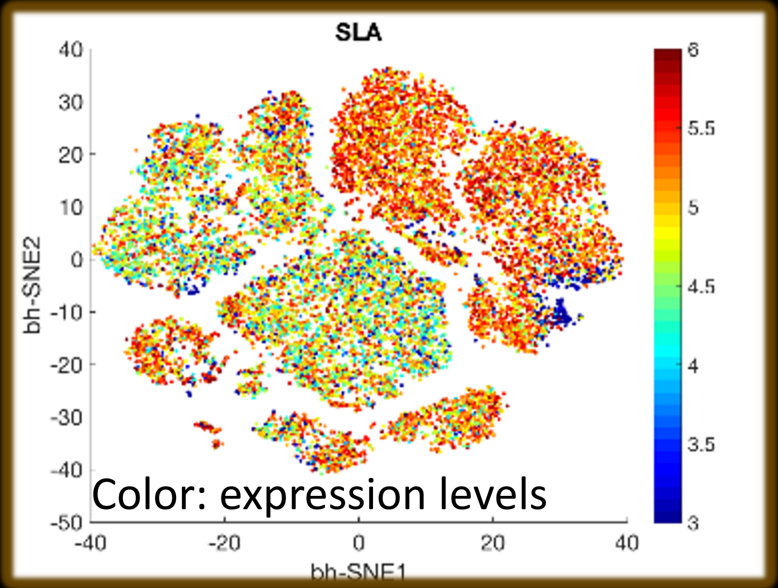
SCONE visualizes nuclear *import* of NF-κB

Comparison



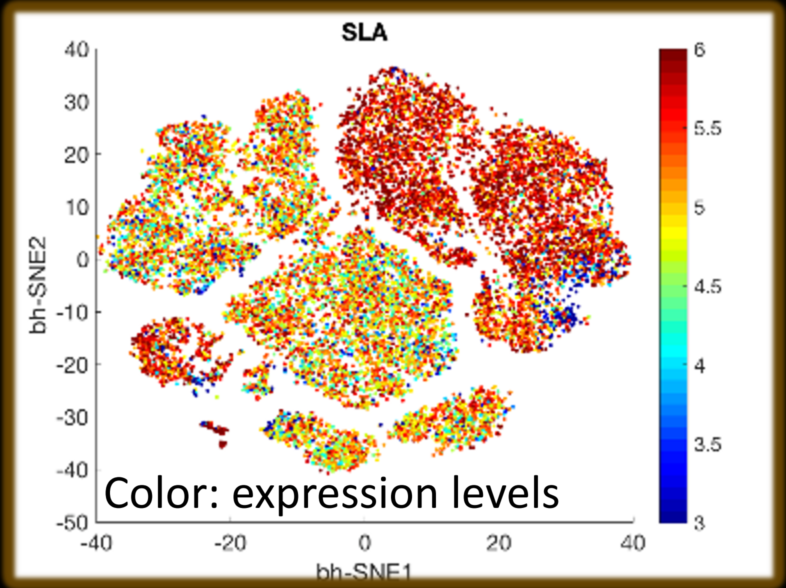
Δ expression levels

Untreated



Color: expression levels

TNFα, 30min

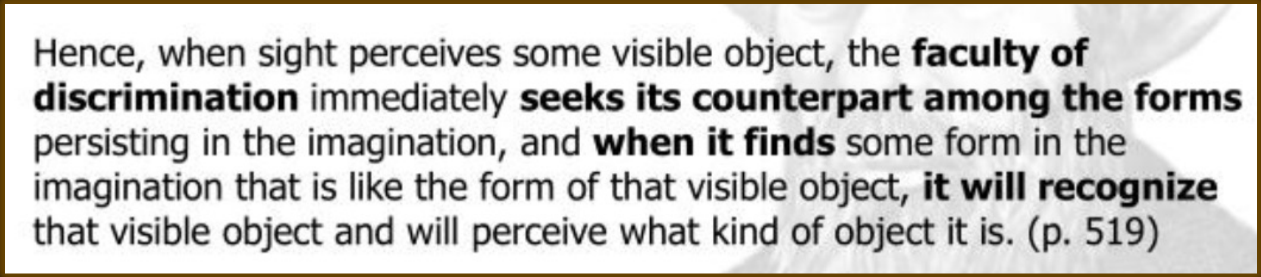


Color: expression levels

The idea of nearest neighbor analysis

Ibn Al-Haytham (Alhazen), 965-1040

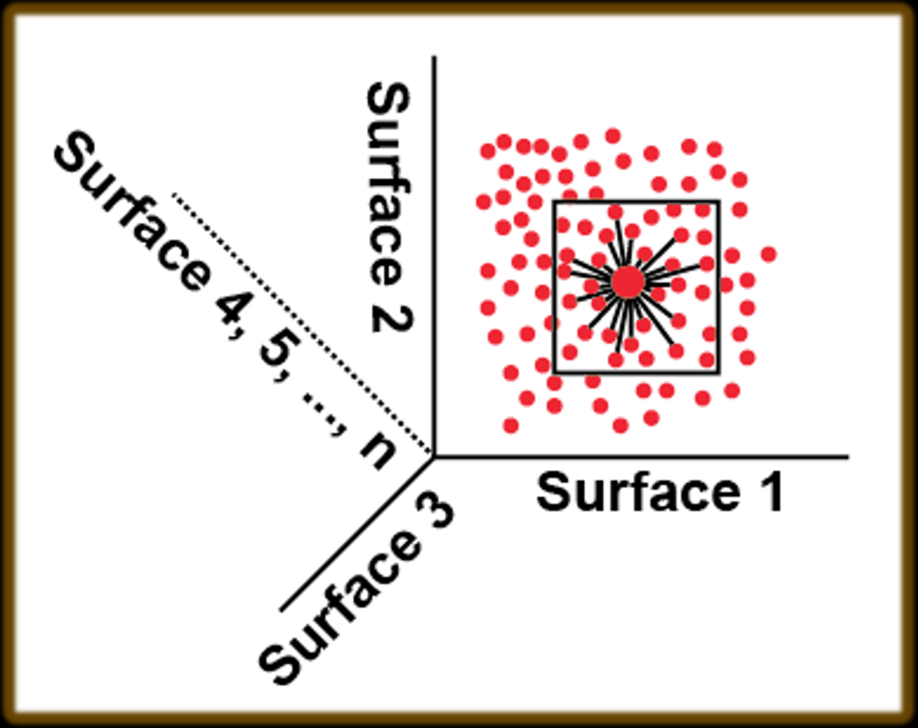
- X-Shift, Samusik et al, *Nat. Meth* 2016 (KNN density estimation)
- Phenograph, Levine et al, *Cell* 2015 (KNN graph clustering)
- One-SENSE, Chang et al, *J Immuno* 2015 (validation of 1D t-SNE)
- KNN smoothing, Wagnar et al, *BiorXiv* 2017



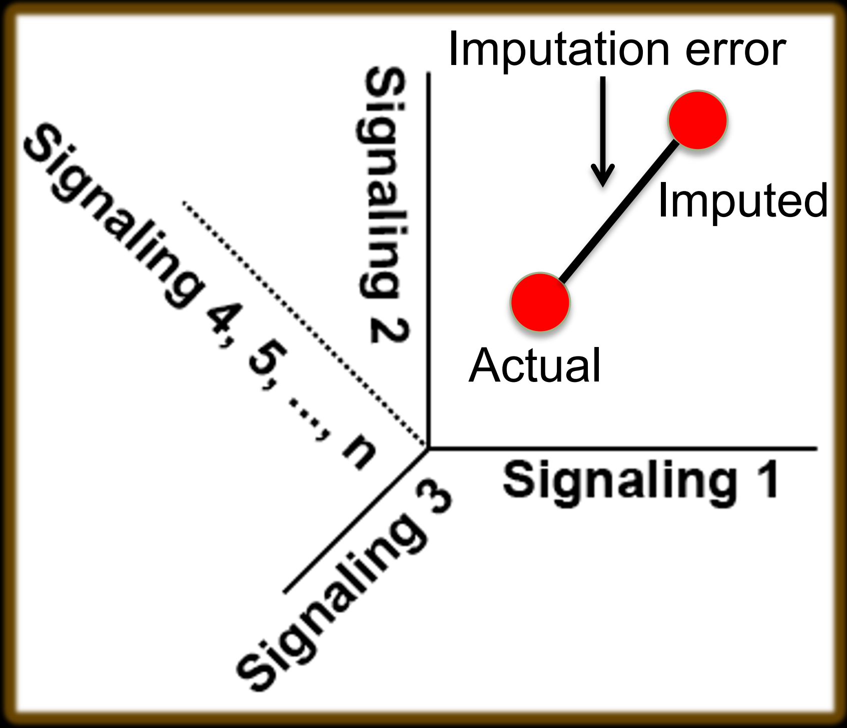
Hence, when sight perceives some visible object, the **faculty of discrimination** immediately **seeks its counterpart among the forms** persisting in the imagination, and **when it finds** some form in the imagination that is like the form of that visible object, **it will recognize** that visible object and will perceive what kind of object it is. (p. 519)

Finding k objectively: optimize imputation of functional markers

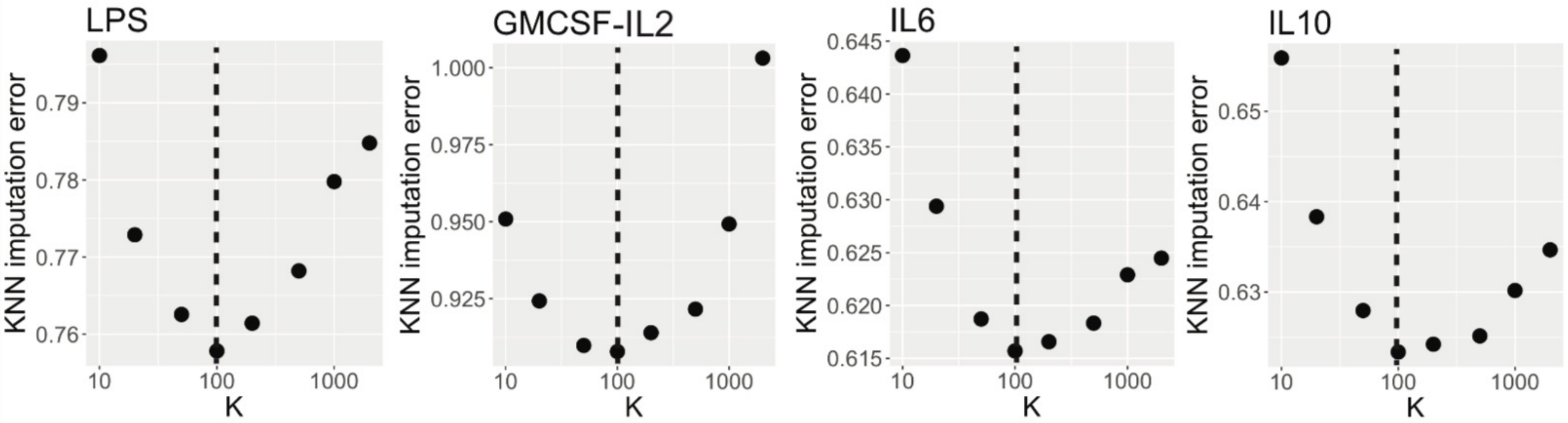
KNN of cell in surface marker space



Cell in signaling space



Global imputation error across different values of k is convex



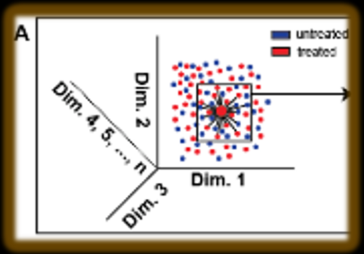
ideal k = n/100 **ideal k = n/100** **ideal k = n/100** **ideal k = n/100**

Dataset: Fragiadakis *et al*, *Anesthesiology* (2015)
Donor: healthy human
Cell type: whole blood
Cell number (n): 10,000

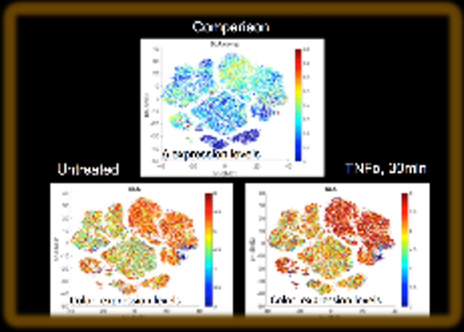
n = number of cells in dataset

Outline

Building per-cell k-nearest neighborhoods in high-D space

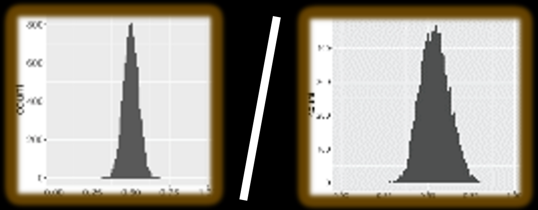


Making single-cell comparisons across t-SNE maps

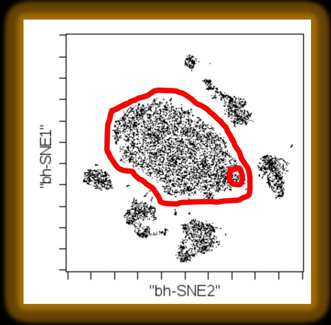


Establishing an evaluation metric for data quality

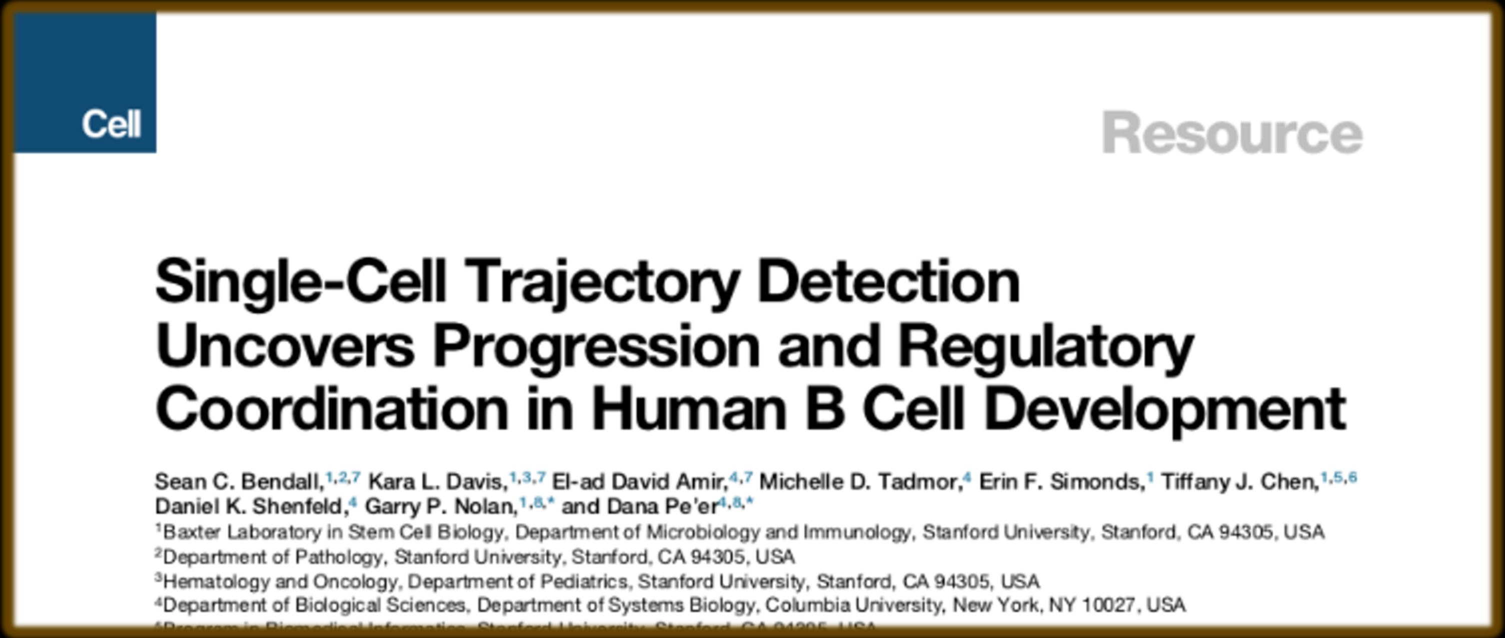
m =



Evaluating the fidelity of lower-dimensional embeddings



Use case: continuous B cell developmental trajectory



Cell

Resource

Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development

Sean C. Bendall,[1,2,7] Kara L. Davis,[1,3,7] El-ad David Amir,[4,7] Michelle D. Tadmor,[4] Erin F. Simonds,[1] Tiffany J. Chen,[1,5,6] Daniel K. Shenfeld,[4] Garry P. Nolan,[1,8,*] and Dana Pe'er[4,8,*]

[1] Baxter Laboratory in Stem Cell Biology, Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA

[2] Department of Pathology, Stanford University, Stanford, CA 94305, USA

[3] Hematology and Oncology, Department of Pediatrics, Stanford University, Stanford, CA 94305, USA

[4] Department of Biological Sciences, Department of Systems Biology, Columbia University, New York, NY 10027, USA

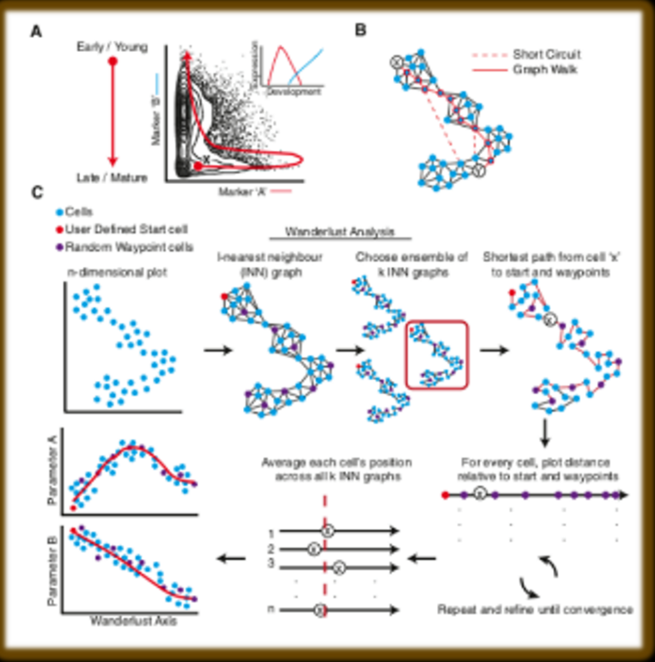
[5] Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA

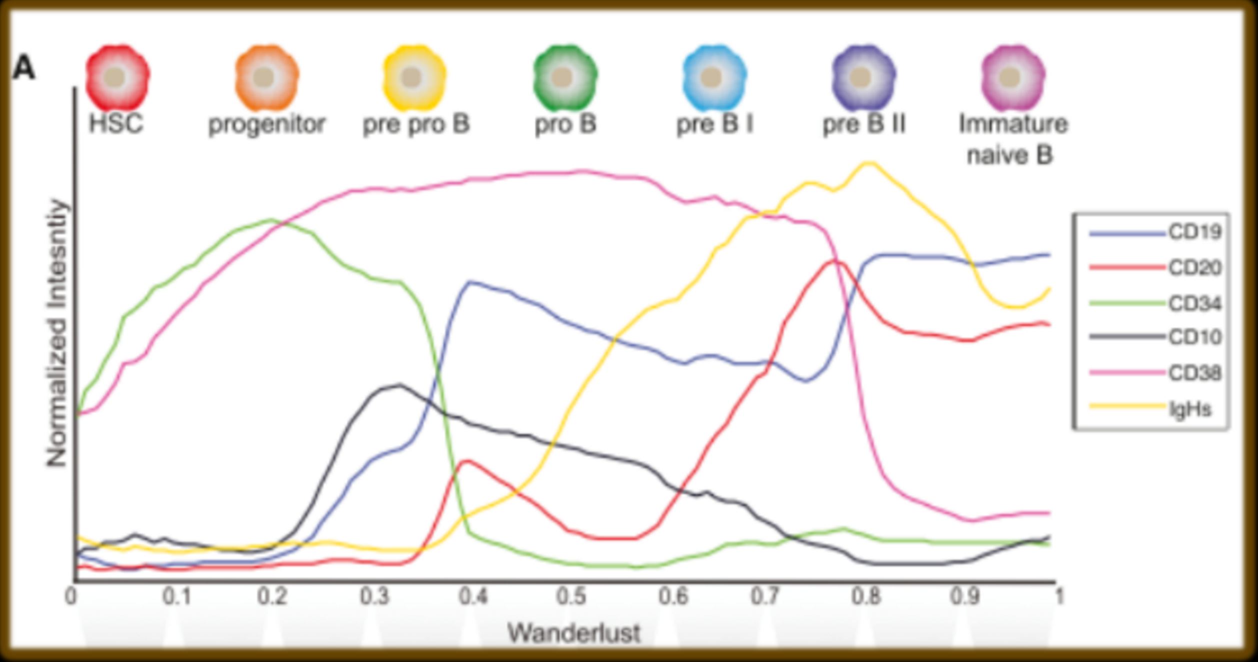
- Cells: B cell precursors manually gated (by expert – Kara Davis, DO) from healthy human bone marrow
- Stimulation conditions: untreated, IL-7
- Goals:
  - Visualize an IL-7 responsive subset along the B cell trajectory

Wanderlust finds a developmental trajectory in single cell data

Cell alignment by time

Reveals developmental trajectories

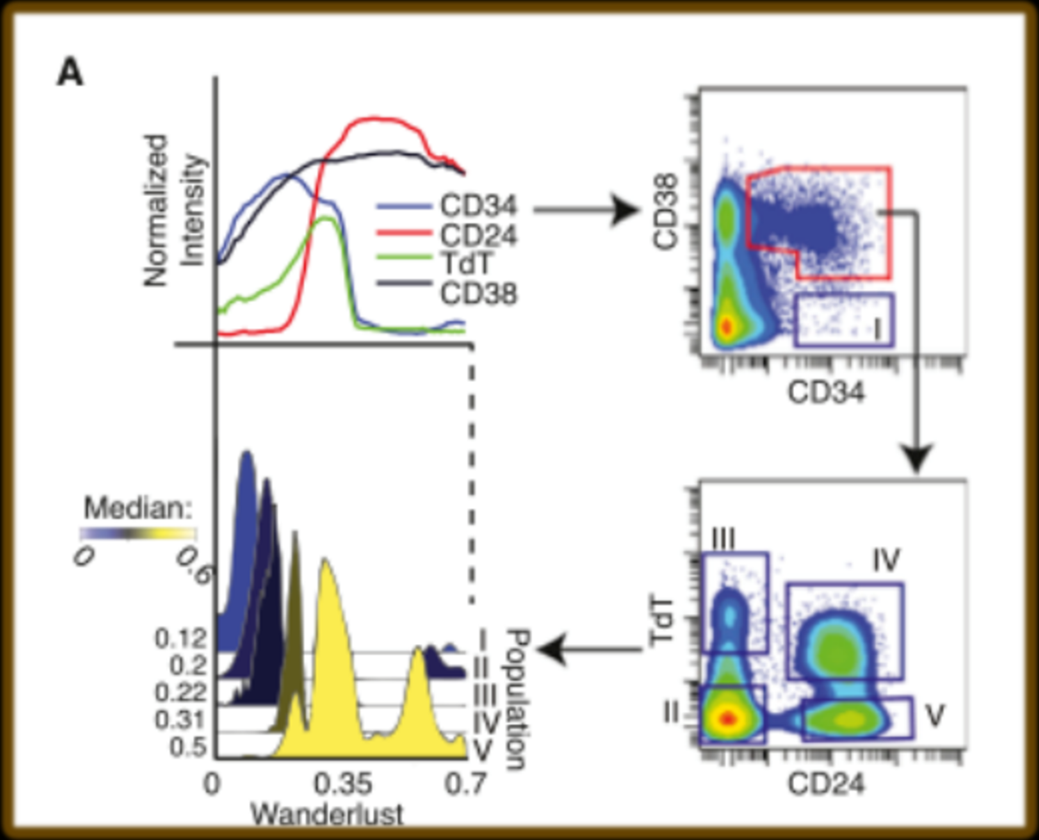

A
Early / Young
Late / Mature
Marker 'B'
Marker 'A'
Expression
Development
B
Short Circuit
Graph Walk
C
Cells
User Defined Start cell
Random Waypoint cells
Wanderlust Analysis
n-dimensional plot
l-nearest neighbour (lNN) graph
Choose ensemble of k lNN graphs
Shortest path from cell 'x' to start and waypoints
For every cell, plot distance relative to start and waypoints
Repeat and refine until convergence
Average each cell's position across all k lNN graphs
1
2
3
n
Parameter A
Parameter B
Wanderlust Axis



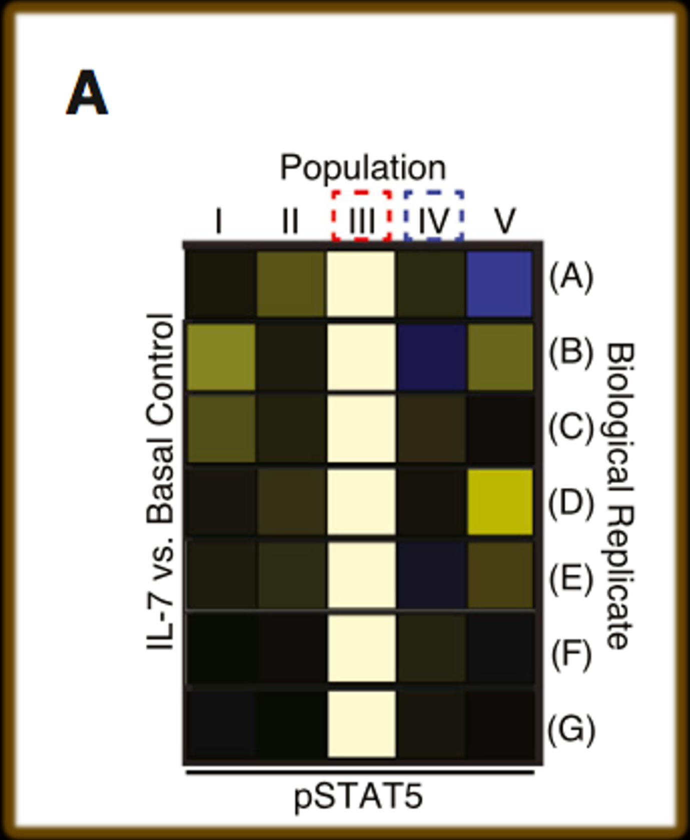
Bendall,
Davis,
Cell 2014

Wanderlust discovered an IL7-pSTAT5 responsive subset

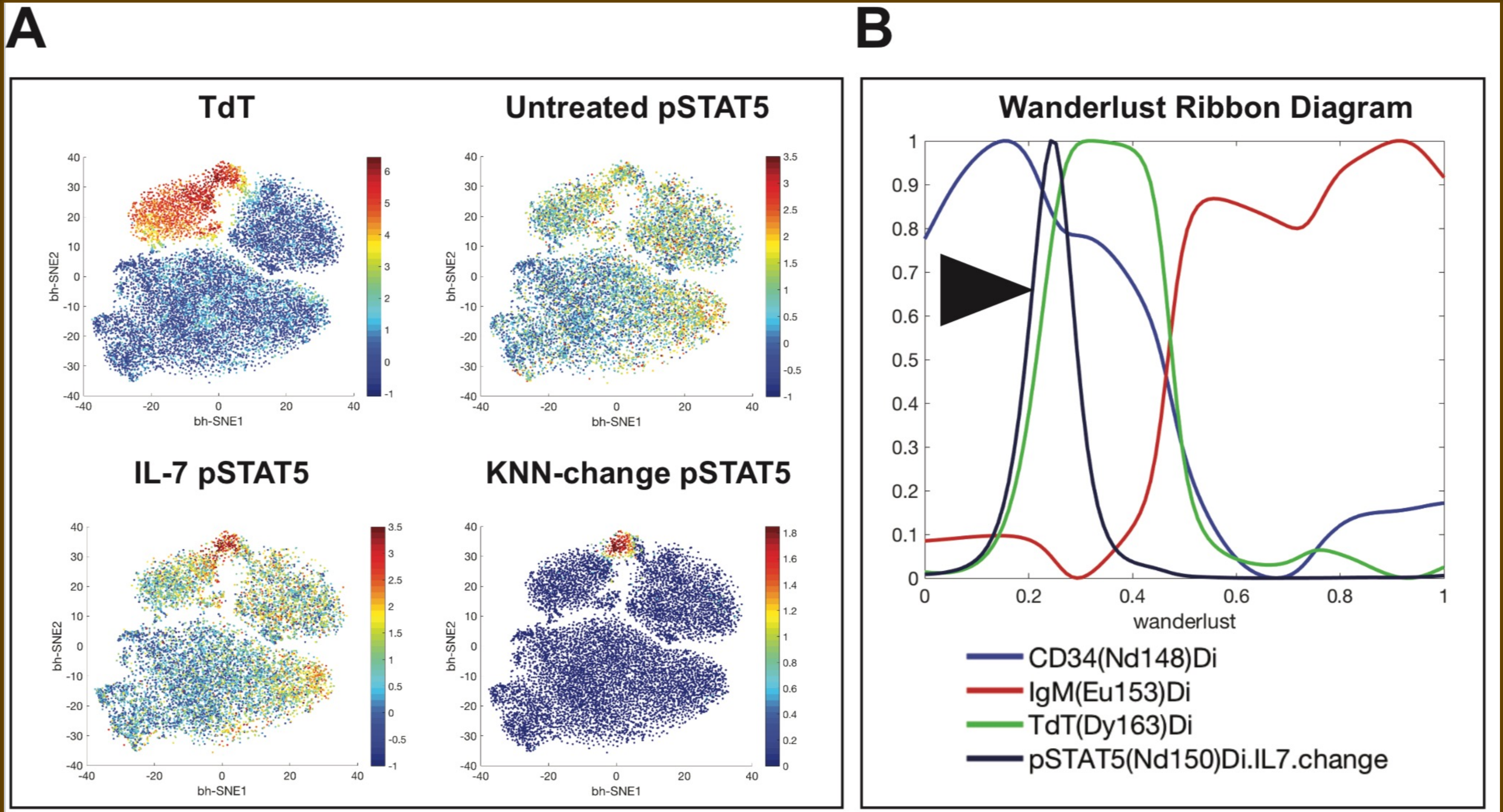
Manual gating



Functional testing



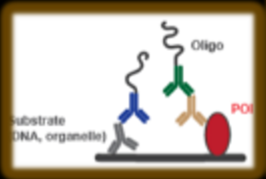
IL7-pSTAT5 responsive subset resides between two “coordination points”

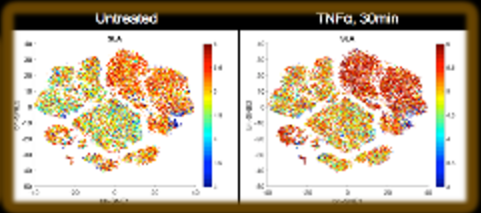


Dataset: Bendall, Davis, Amir *et al, Cell* (2014)
Donors: healthy human
Cells: B cell precursors gated from bone marrow
Cell number: 20,000

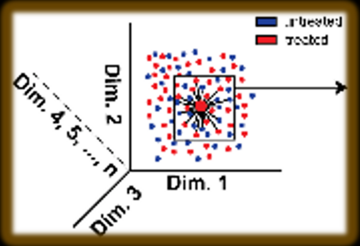
Summary 1

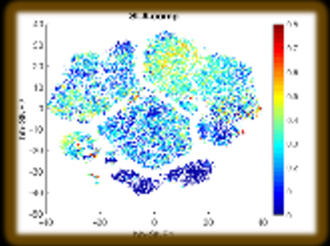
SLA method revealed t-SNE comparison problem



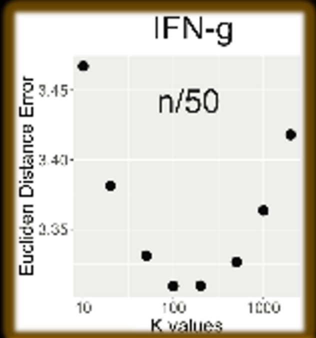


t-SNE comparison problem solved with K-nearest neighbors

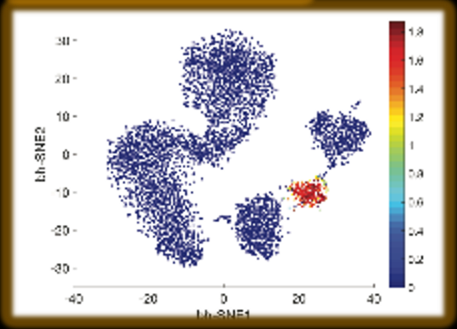




K is selected by minimizing the KNN-imputation error for functional markers

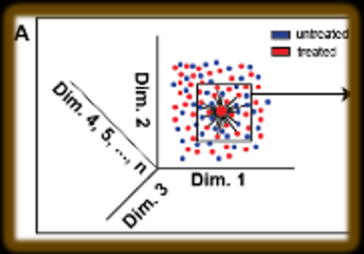


IL-7 responsive population and density estimation shown at single cell level

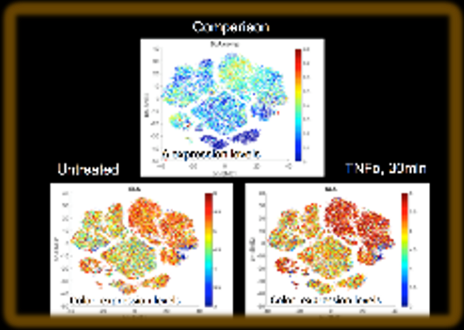


Outline

Building per-cell k-nearest neighborhoods in high-D space

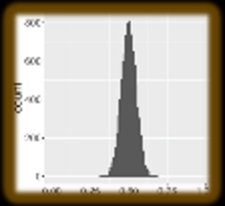


Making single-cell comparisons across t-SNE maps

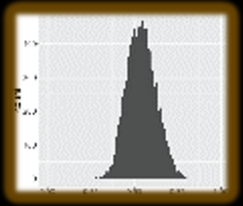


Establishing an evaluation metric for data quality

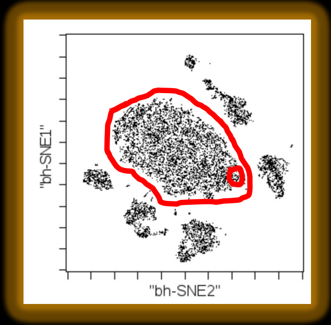
m =



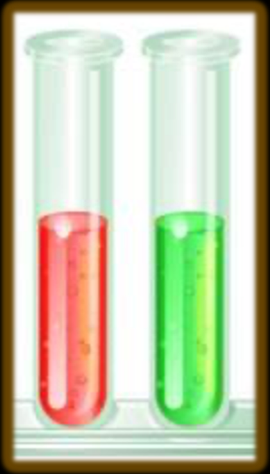
/



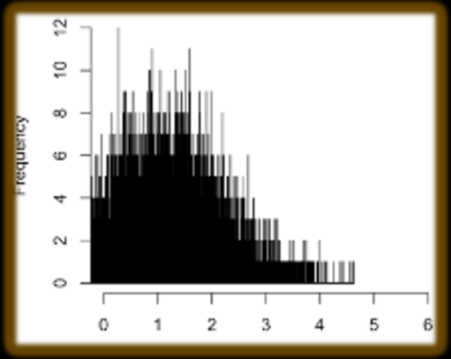
Evaluating the fidelity of lower-dimensional embeddings



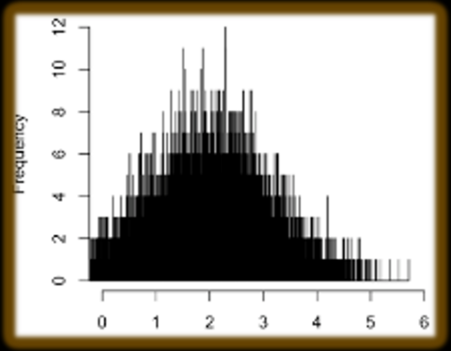
Does population-defining marker space "shift" due to technical artifact between tubes?



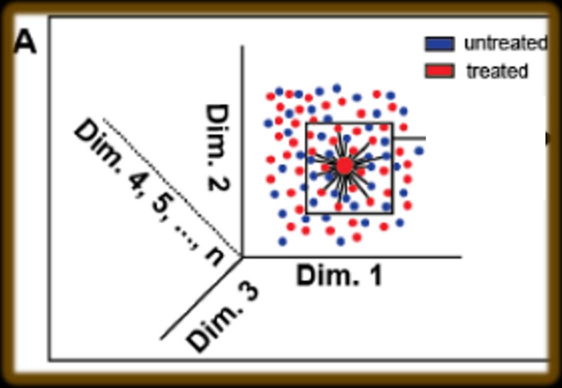
Tube 1



Tube 2

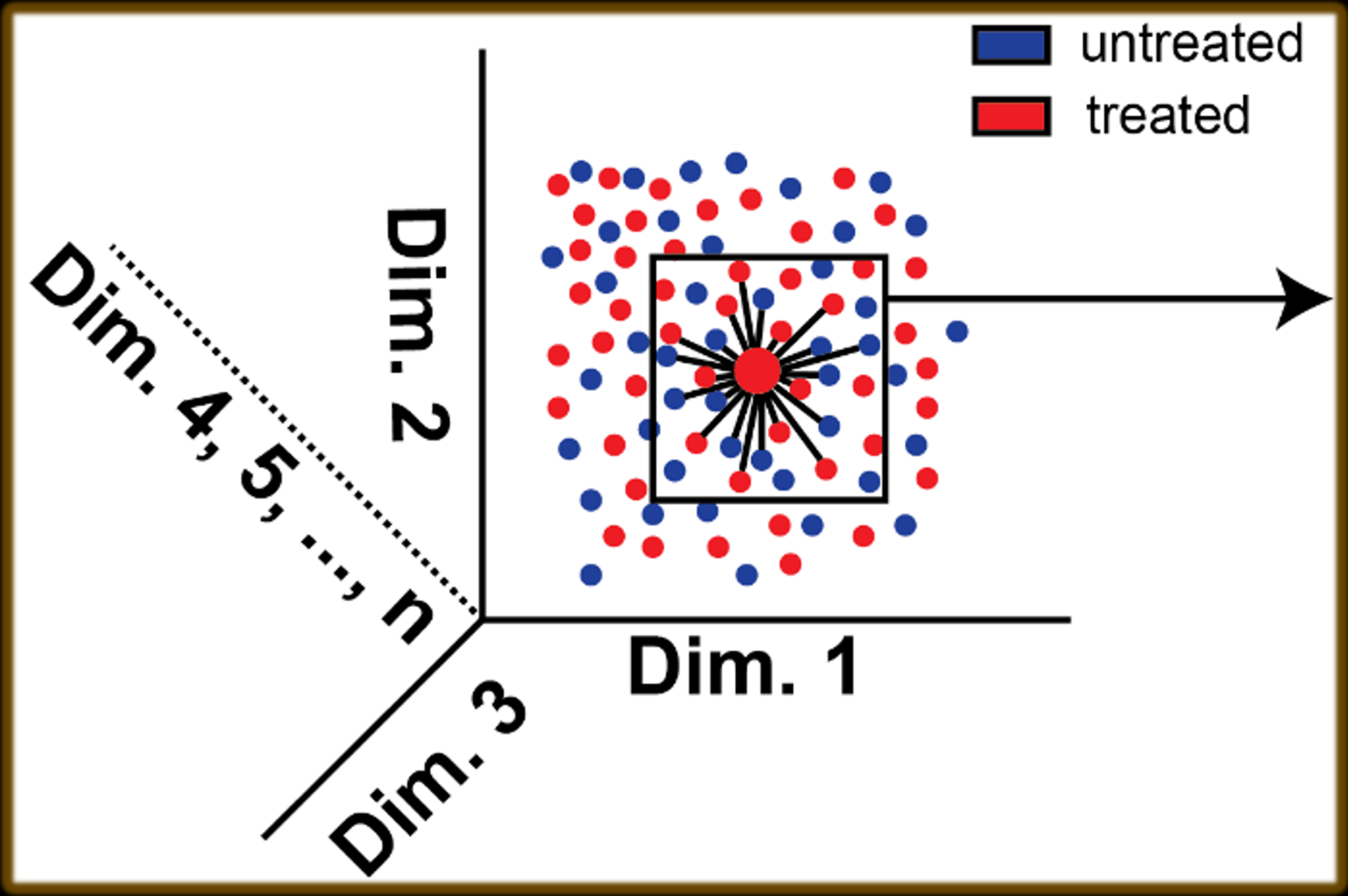


CD33
(not supposed
to change)

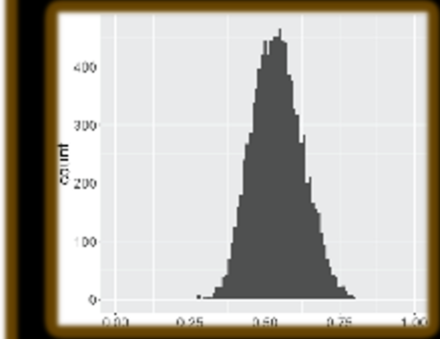


?

How to test for marker “shift” due to technical artifact? Use KNN.



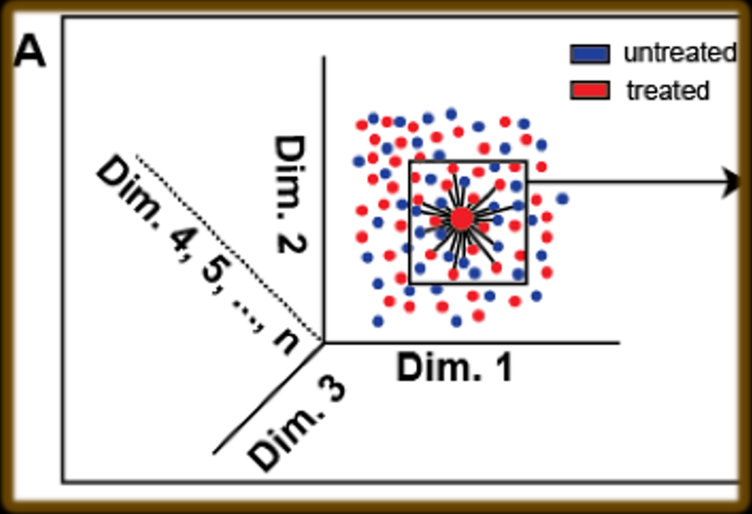
For each KNN calculate the fraction belonging to “red” condition



But what do we benchmark the SD to?

A coin toss distribution represents "perfect" manifold overlap

Sample from each KNN size 100, for 10,000 KNN

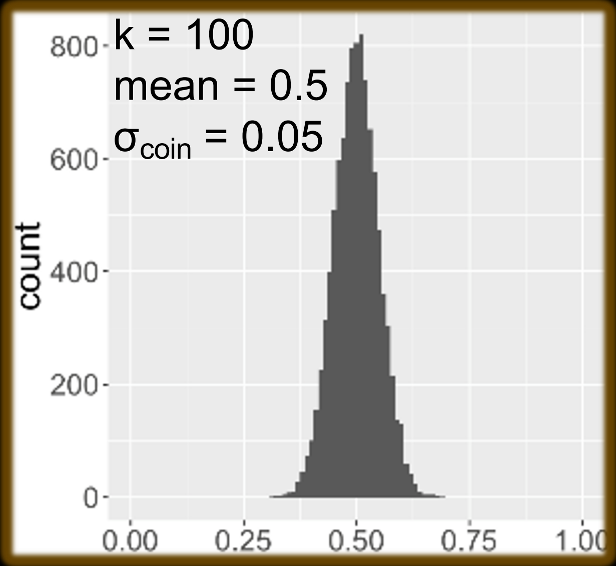


Flip a coin
100 times,
repeat
10,000 times

Compare to simulated coin toss

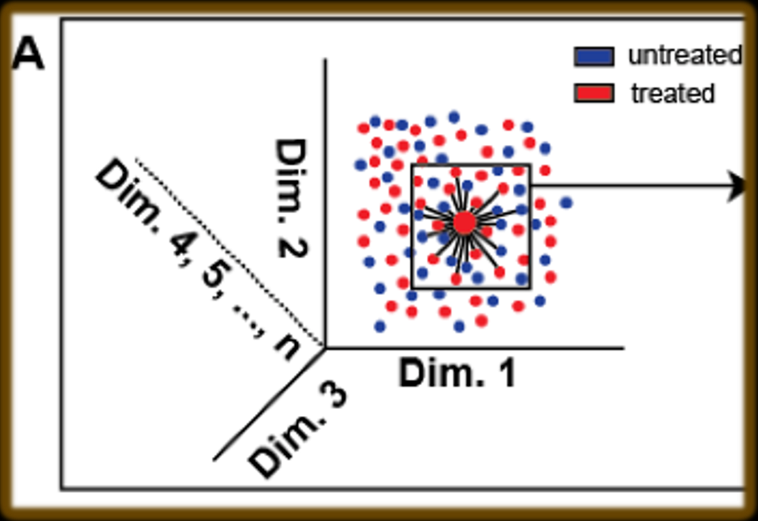
Do *k* flips,
repeat *n* times

$$\sigma_{coin} = \frac{0.5}{\sqrt{k}}$$



Fraction heads

Evaluation metric: manifold overlap score to quantify global tube-to-tube technical variation



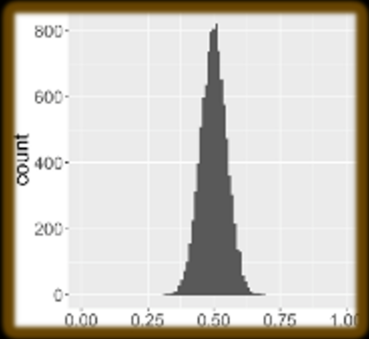
The fraction of the KNN that is red

$$\alpha_n(x_i, x_b) = \frac{count_n(x_i)}{count_n(x_i) + count_n(x_b)}$$

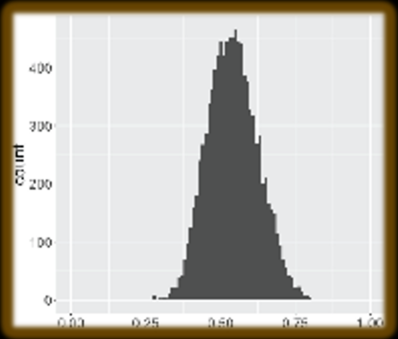
“Fraction red” for all KNN in the dataset, one for each cell

$$\alpha(x_i, x_b) = \{\alpha_1(x_i, x_b), \alpha_2(x_i, x_b), \alpha_3(x_i, x_b), \alpha_4(x_i, x_b), \dots, \alpha_n(x_i, x_b)\}$$

coin toss



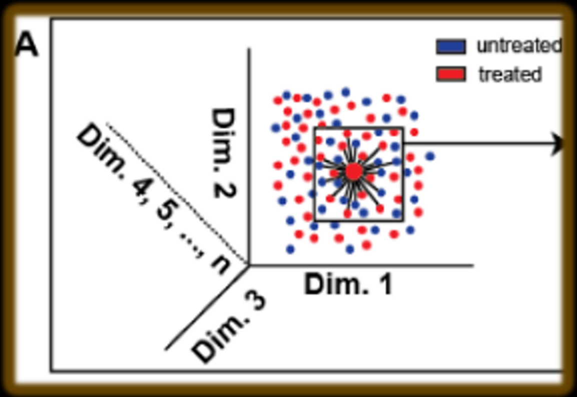
fraction red



SD of fair coin toss distribution, divided by SD of “fraction red” distribution

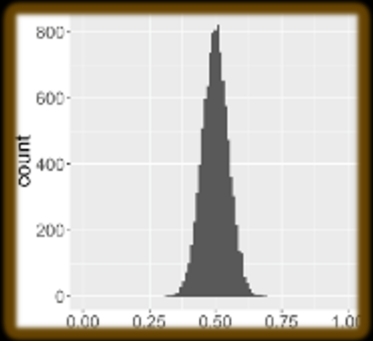
$$m = \frac{\sigma_{coin}}{\sigma(\alpha(x_i, x_b))}$$

Normalization can improve manifold overlap score



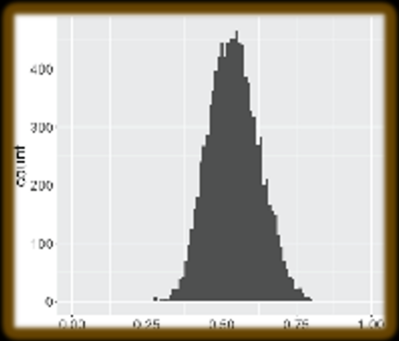
m =

coin toss



/

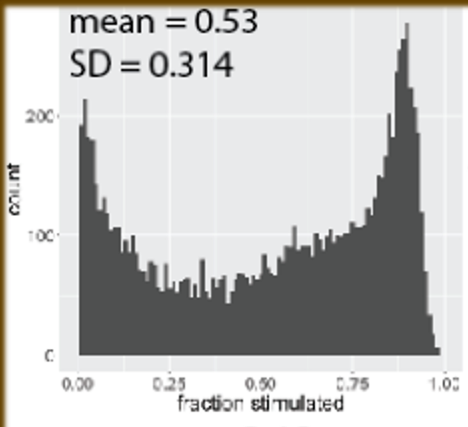
fraction red



Bodenmiller, Zunder *et al, Nat Biotech* 2012
Untreated vs GM-CSF

Bendall, Davis *et al, Cell* 2014
Untreated vs IL-7

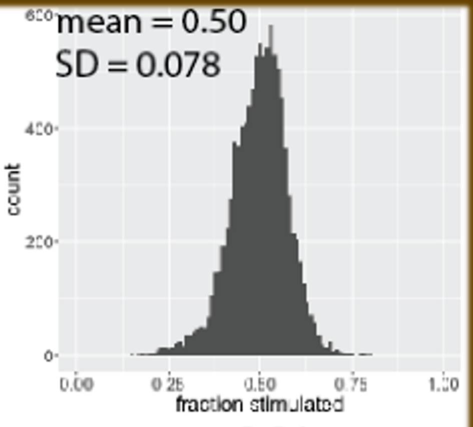
Before normalization



mean = 0.53
SD = 0.314

m = 0.16

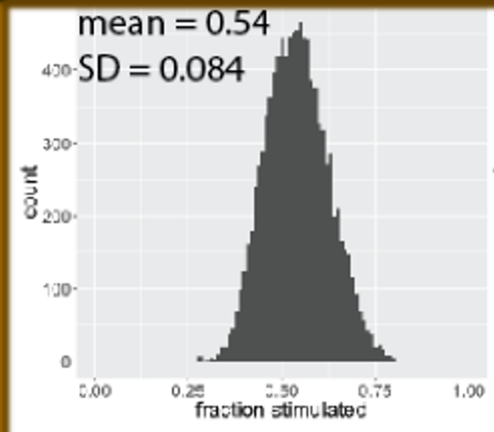
After normalization



mean = 0.50
SD = 0.078

m = 0.64

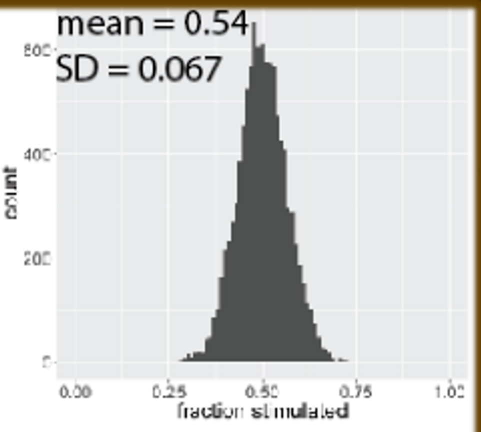
Before normalization



mean = 0.54
SD = 0.084

m = 0.60

After normalization

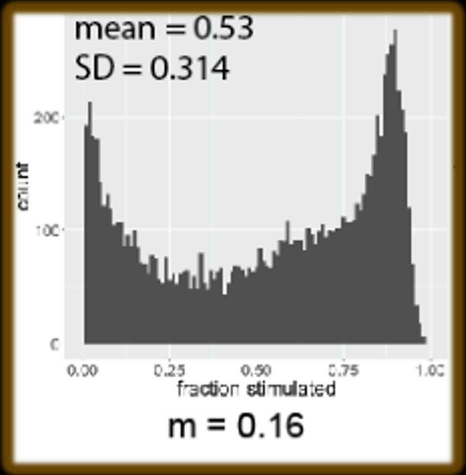


mean = 0.54
SD = 0.067

m = 0.75

Higher m score: better-defined functional subsets

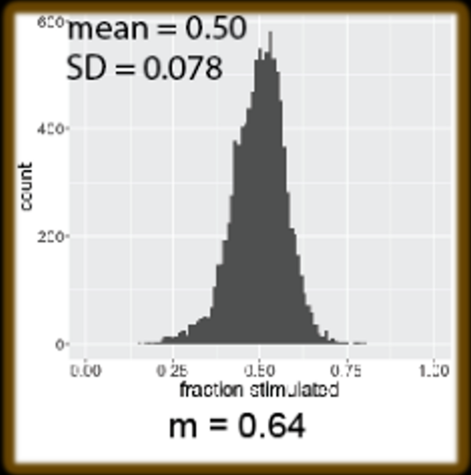
Before normalization



mean = 0.53
SD = 0.314

m = 0.16

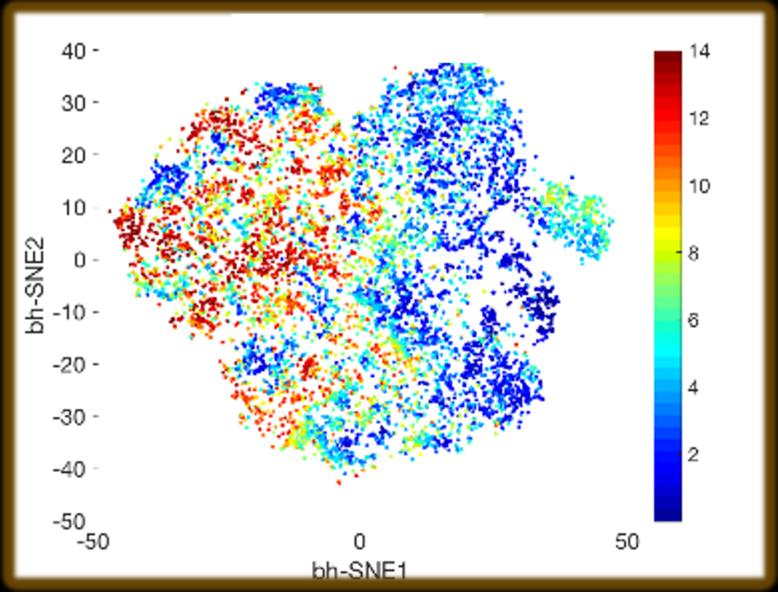
After normalization



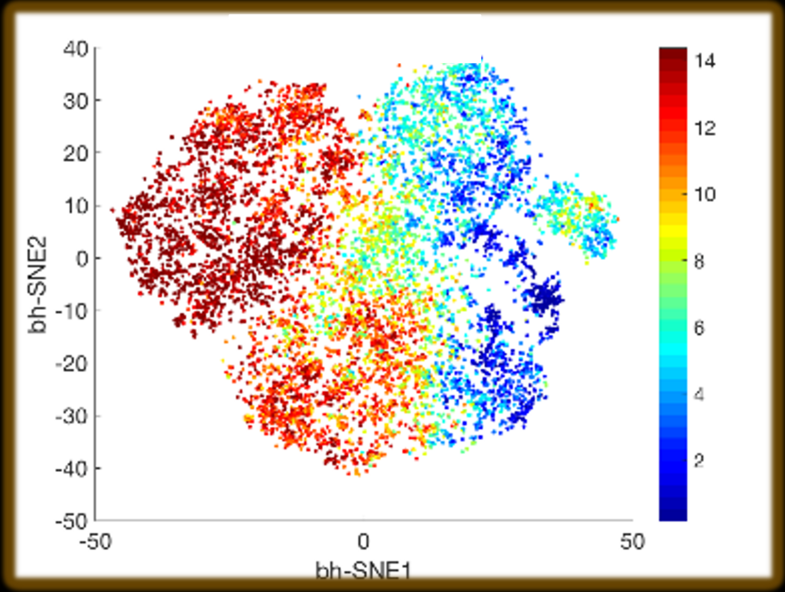
mean = 0.50
SD = 0.078

m = 0.64

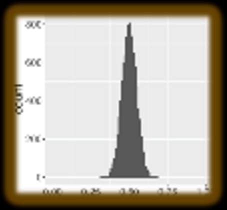
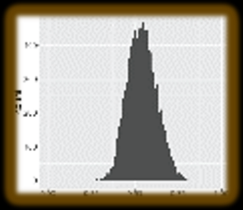
IFNα - pSTAT5 –log(q value)



IFNα - pSTAT5 –log(q value)



Summary 2

m =  / 

- KNN architecture can be used to assess global tube-to-tube technical variation
- Normalization of data brings knn ratios closer to 50%, and does not alter functional information

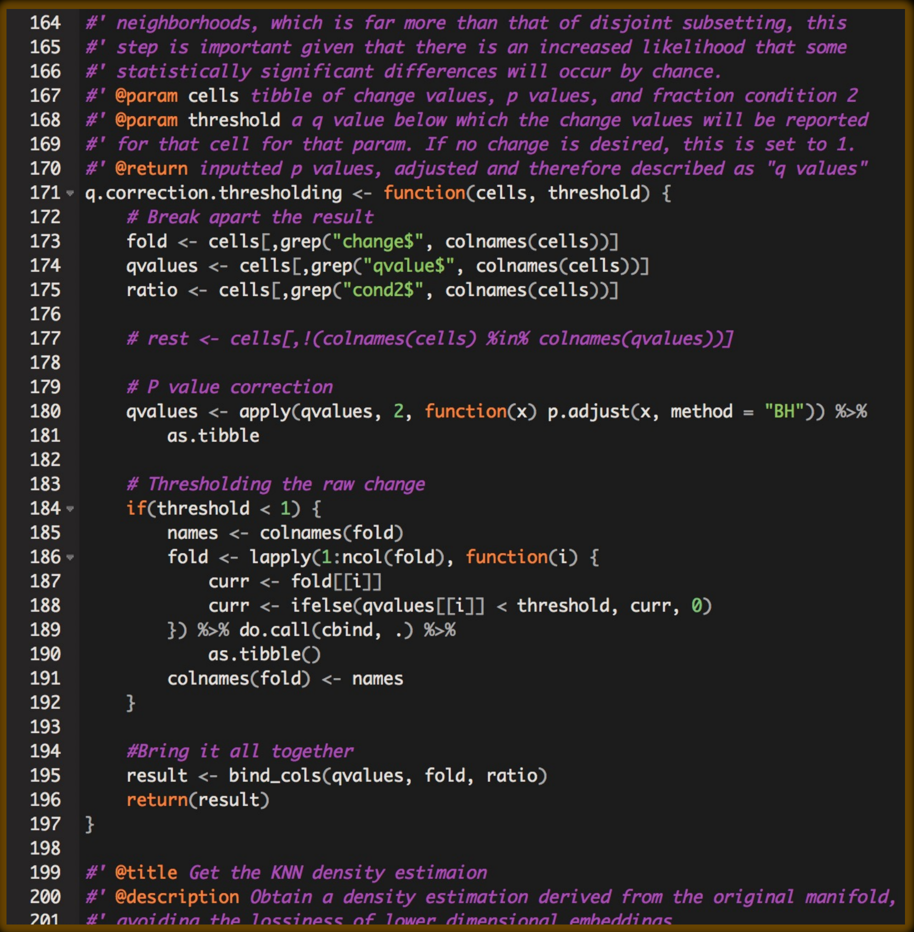
- Applications: replicate variation, donor-donor variation, optimizing normalization methods…

Other questions that KNN can be used to answer

- Does one’s panel contain any redundant markers?
- How much information do you lose by doing a low dimensional embedding (and which is the best?)
  - Flow-CAP for low-D embeddings
- What is the Shannon entropy of a CyTOF dataset (quantify heterogeneity, esp for cancer)

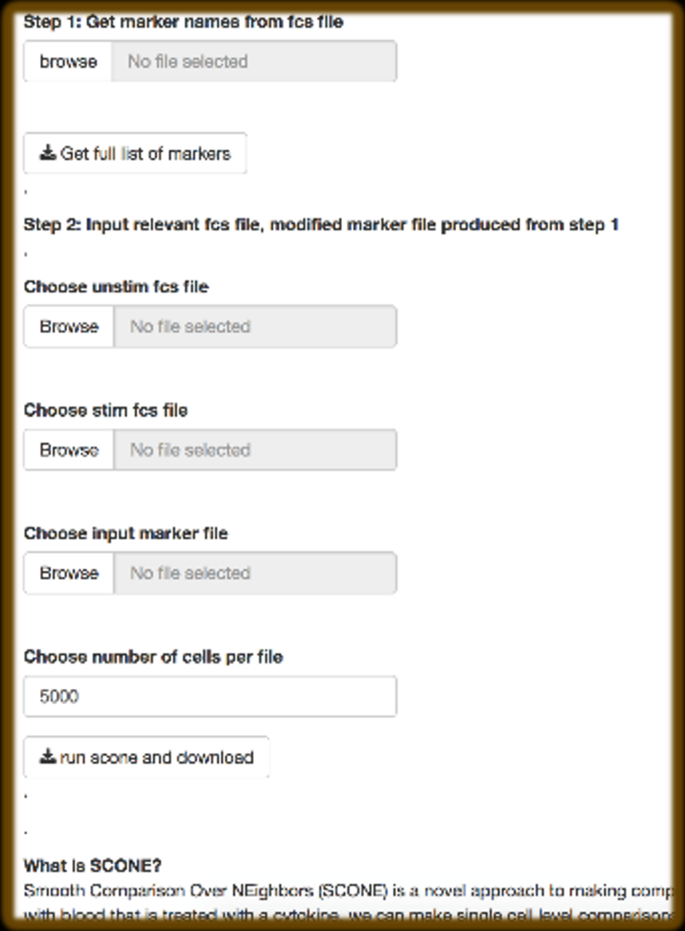
You should try this out yourself!

Bioconductor: Sconify



```
#' neighborhoods, which is far more than that of disjoint subsetting, this
#' step is important given that there is an increased likelihood that some
#' statistically significant differences will occur by chance.
#' @param cells tibble of change values, p values, and fraction condition 2
#' @param threshold a q value below which the change values will be reported
#' for that cell for that param. If no change is desired, this is set to 1.
#' @return inputted p values, adjusted and therefore described as "q values"
q.correction.thresholding <- function(cells, threshold) {
    # Break apart the result
    fold <- cells[,grep("change$", colnames(cells))]
    qvalues <- cells[,grep("qvalue$", colnames(cells))]
    ratio <- cells[,grep("cond2$", colnames(cells))]

    # rest <- cells[,!(colnames(cells) %in% colnames(qvalues))]

    # P value correction
    qvalues <- apply(qvalues, 2, function(x) p.adjust(x, method = "BH")) %>%
        as.tibble

    # Thresholding the raw change
    if(threshold < 1) {
        names <- colnames(fold)
        fold <- lapply(1:ncol(fold), function(i) {
            curr <- fold[[i]]
            curr <- ifelse(qvalues[[i]] < threshold, curr, 0)
        }) %>% do.call(cbind, .) %>%
            as.tibble()
        colnames(fold) <- names
    }

    #Bring it all together
    result <- bind_cols(qvalues, fold, ratio)
    return(result)
}

#' @title Get the KNN density estimaion
#' @description Obtain a density estimation derived from the original manifold,
#' avoiding the lossiness of lower dimensional embeddings
```

www.sconify.org



Step 1: Get marker names from fcs file

.

Step 2: Input relevant fcs file, modified marker file produced from step 1

.

Choose unstim fcs file

Choose stim fcs file

Choose input marker file

Choose number of cells per file

5000

.

.

What is SCONE?

Smooth Comparison Over NEighbors (SCONE) is a novel approach to making comp
with blood that is treated with a cytokine, we can make single cell level comparisons

github.com/tjburns08

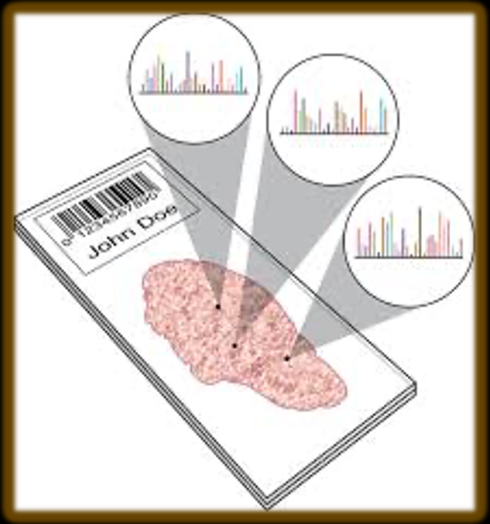
email: burns.tyler@gmail.com

Burns *et al*, *Cytometry* 2017(2) (in review)

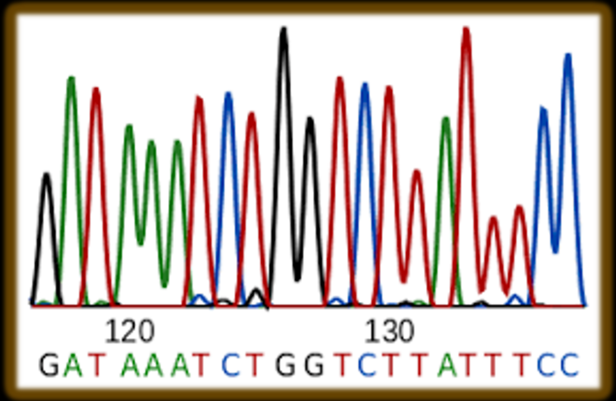
High parameter single cell analysis is becoming more available (and popular) in biomedicine

High-dim cytometry

High-dim imaging



Single cell sequencing



Acknowledgement

DRFZ BERLIN
Deutsches Rheuma-Forschungszentrum
Ein Institut der Leibniz-Gemeinschaft

Henrik Mei Pawel Durek **Axel Schulz** Andreas Grützkau

Silke Stanislawiak **Sabine Baumgart** Marie Urbicht
Christina Schäfer Sarah Gillert

Heike Hirseland **Tyler Burns** Lisa Budzinski Edward Rullmann Julia Schulze

Cytodiagnostics, Canada

Ben Pacheco

Miltenyi BioTec

Christian Dose, Susanne Krauthäuser

Stanford

Michael Leipold

Holden Maecker, Mark Davis, Garry Fathman

Scailyte

Manfred Claassen, Daniel Sonnleithner

Freie Universität Berlin

Prof. Dr. Susanne Hartmann
Institut für Immunologie
Dr. Svenja Steinfelder
Institut für Immunologie

IMMANUEL DIAKONIE

Prof. Dr. Andreas Krause
Innere Medizin, Rheumatologie und Klinische Immunologie
Prof. Dr. Andreas Michalsen
Naturheilkunde

DFG Deutsche Forschungsgemeinschaft

DGKL Deutsche Vereinte Gesellschaft für Klinische Chemie und Laboratoriumsmedizin e.V.
Nachwuchsakademie 2017

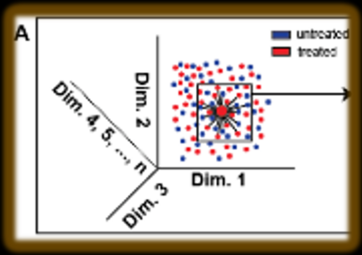
GerMaNet German Mass Cytometry Network

Be The Cure FOR RHEUMATOID ARTHRITIS

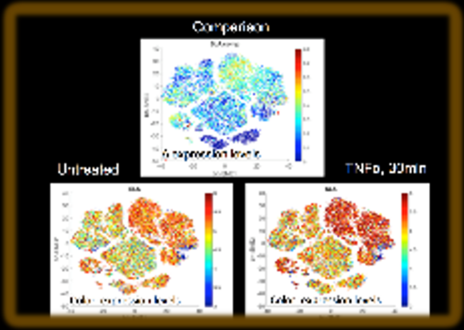
SYS INFLAME

Outline

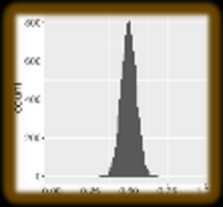
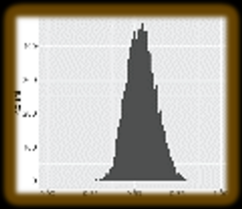
Building per-cell k-nearest neighborhoods in high-D space



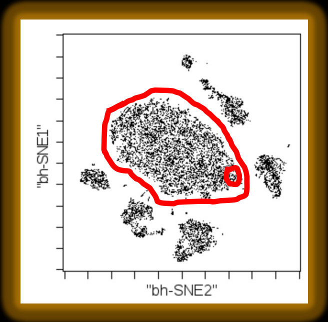
Making single-cell comparisons across t-SNE maps



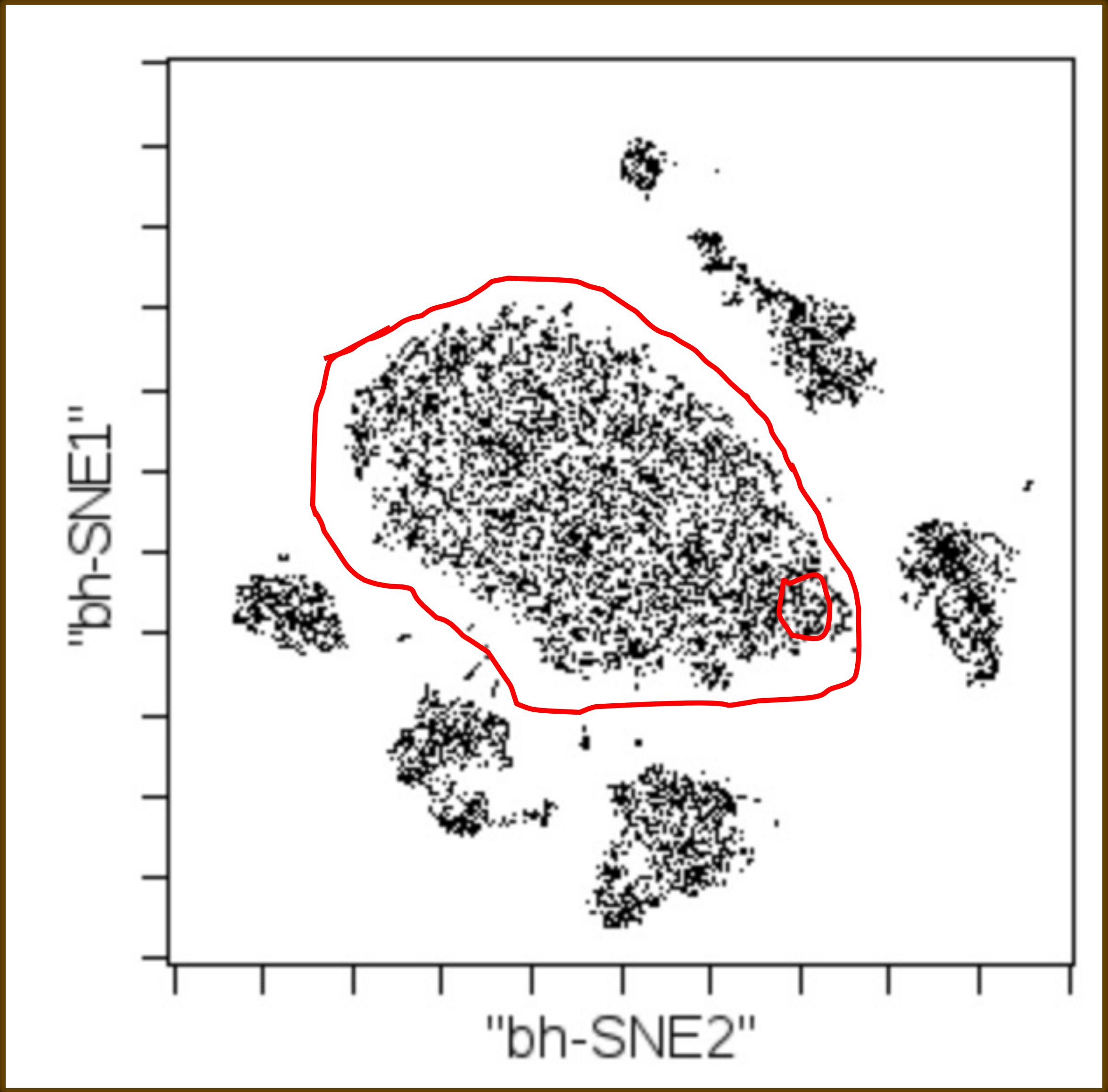
Establishing an evaluation metric for data quality

m =  / 

Evaluating the fidelity of lower-dimensional embeddings



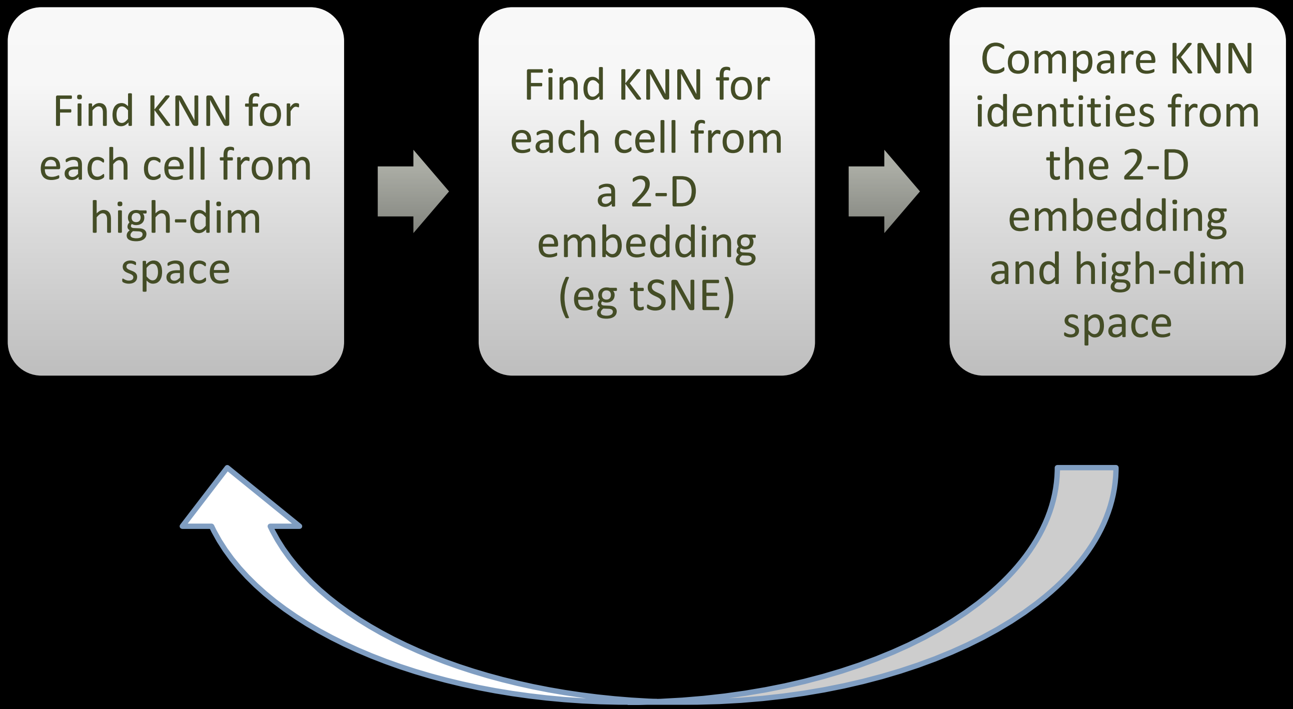
How precise is a t-SNE map? (should we gate/cluster it?)



Gate around an Island?

Gate within an Island?

KNN to determine fidelity of lower dimensional embeddings



Find KNN for each cell from high-dim space

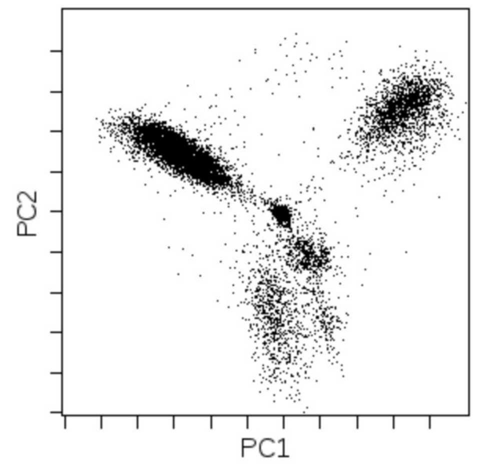
Find KNN for each cell from a 2-D embedding (eg tSNE)

Compare KNN identities from the 2-D embedding and high-dim space

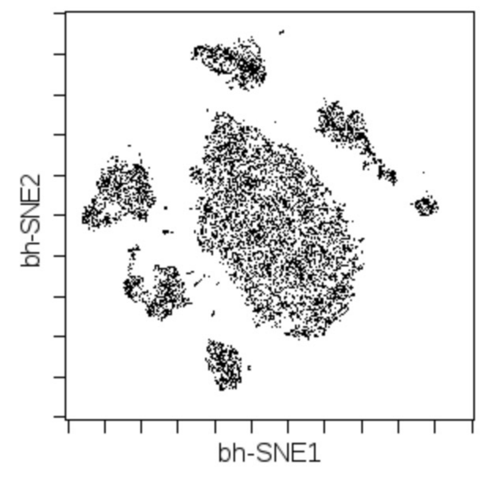
Repeat across a wide range of values for K

Two low dim embeddings: t-SNE vs PCA

- PCA
  - Seeks to explain the variance of data
  - Can only pick up linear structure
  - Consistent: same result every time
  - Very fast run time

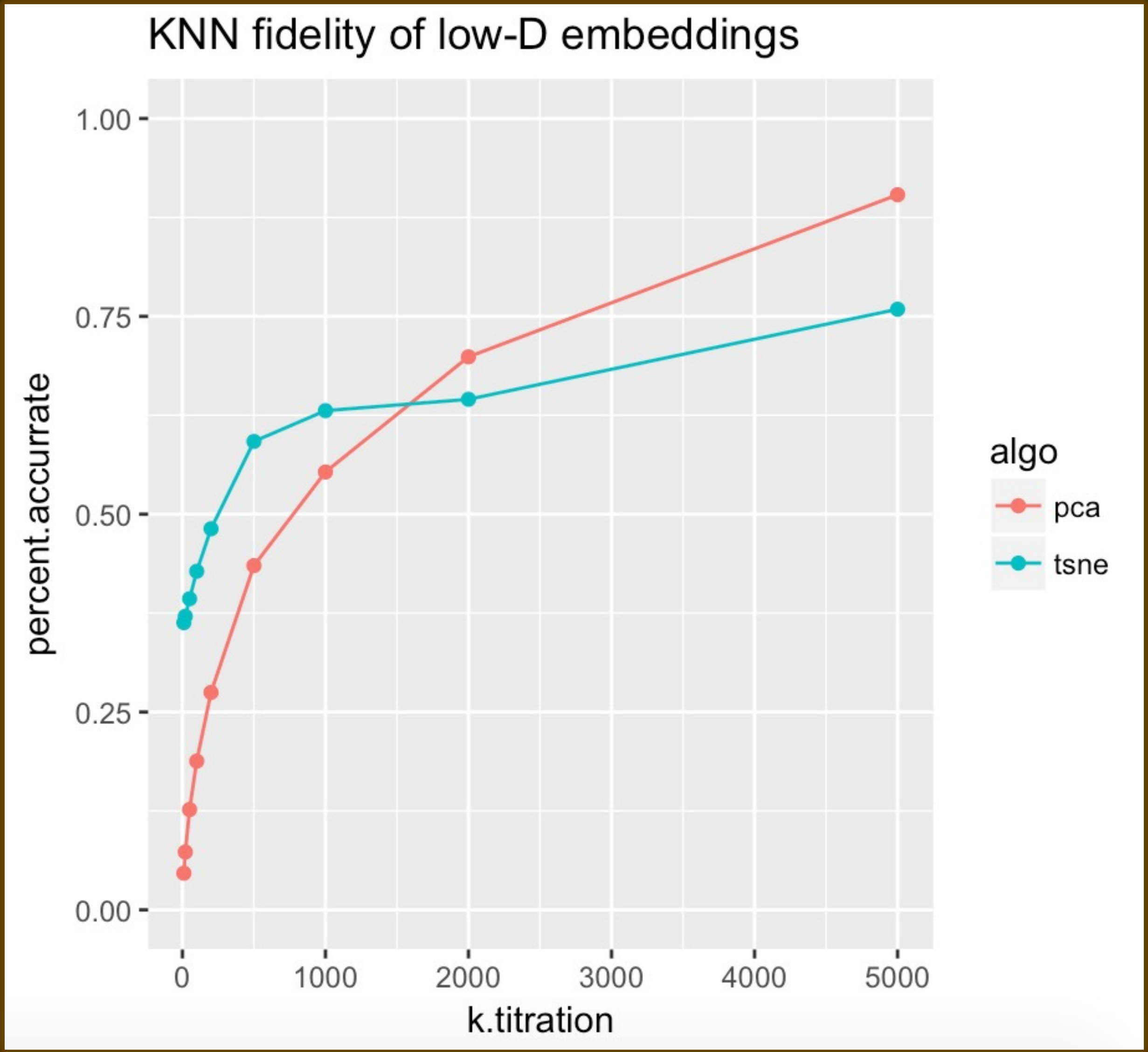


- t-SNE
  - Seeks to preserve local structure
  - Can pick up non-linear structure
  - Inconsistent: different result every time
  - Very slow run time



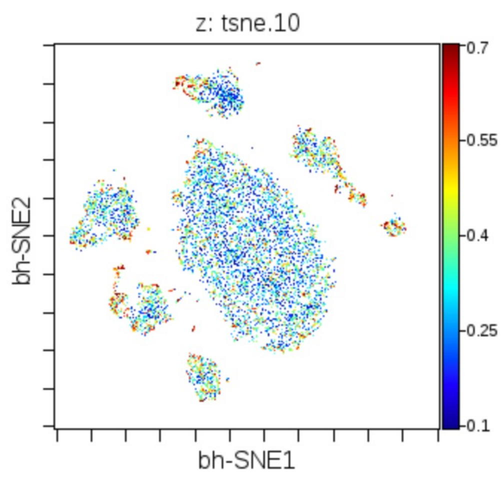
Data from Fragidakis *et al* *Anesthesiology* 2015

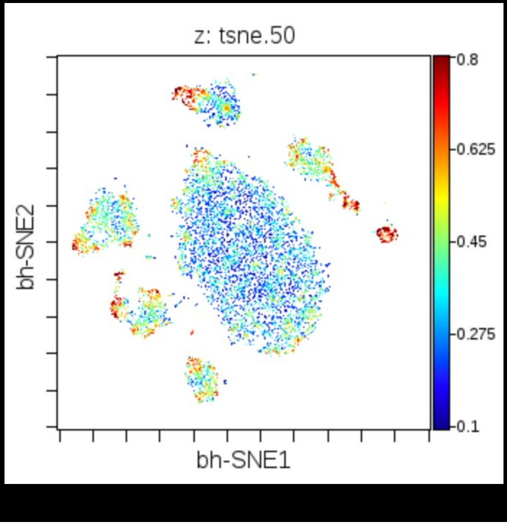
Global fidelity of lower dimensional embeddings: tSNE vs PCA

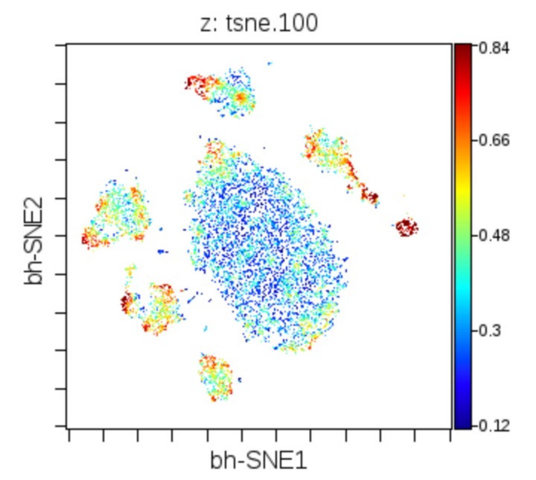


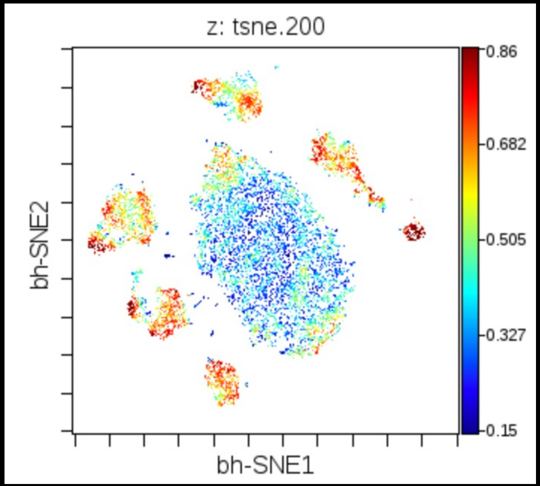
Data:
Fragidakis *et al*
Anesthesiology 2015
Cells: whole blood
Cell number: 10,000

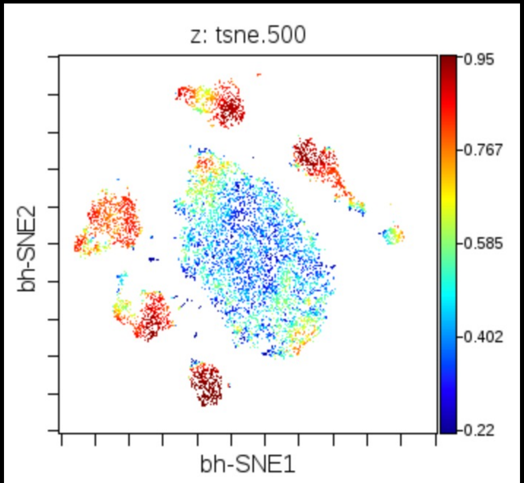
Fidelity of lower dimensional embeddings is region-specific

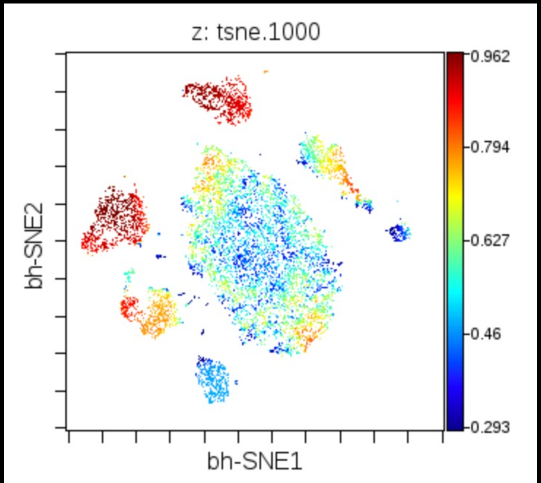




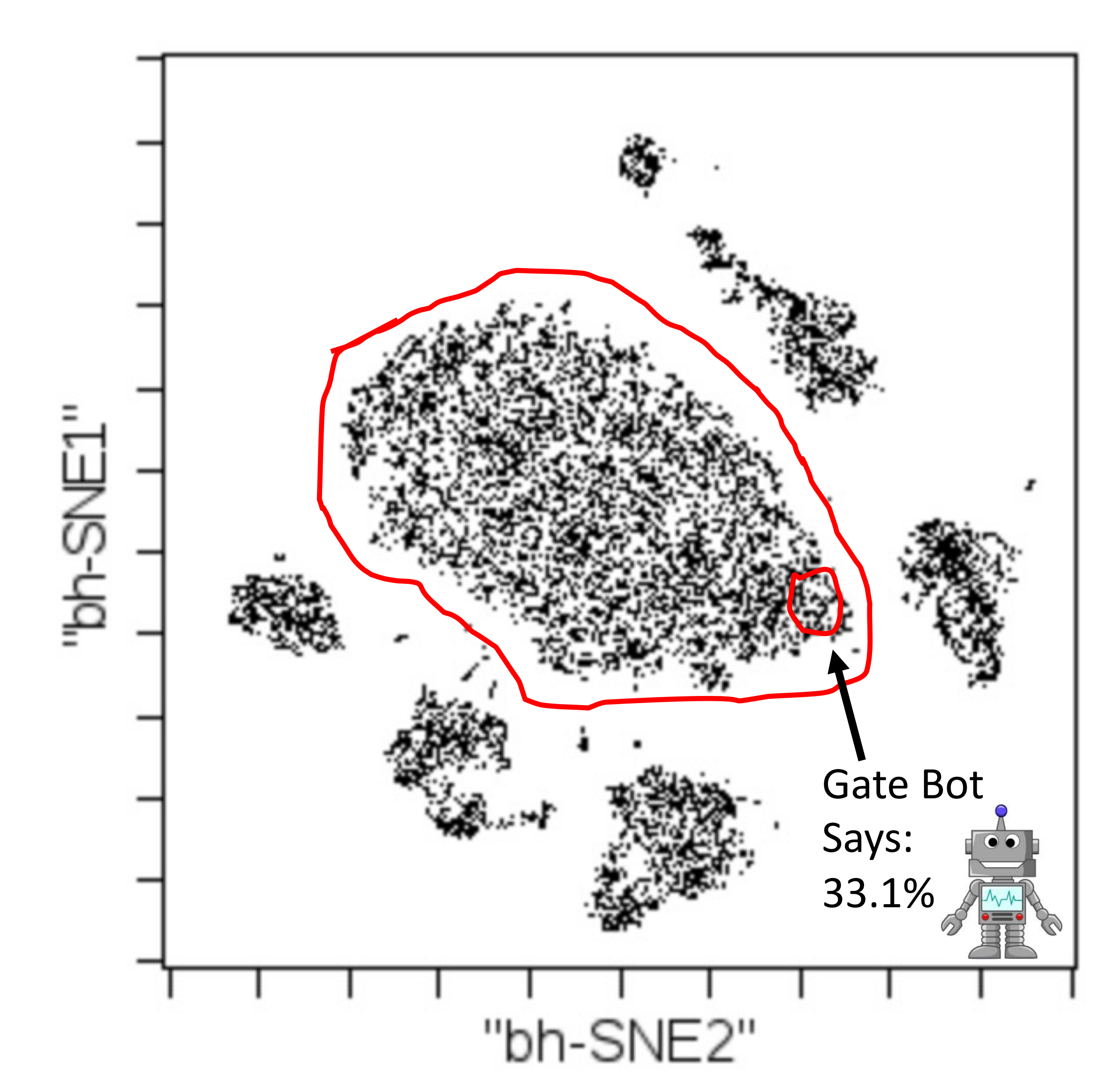








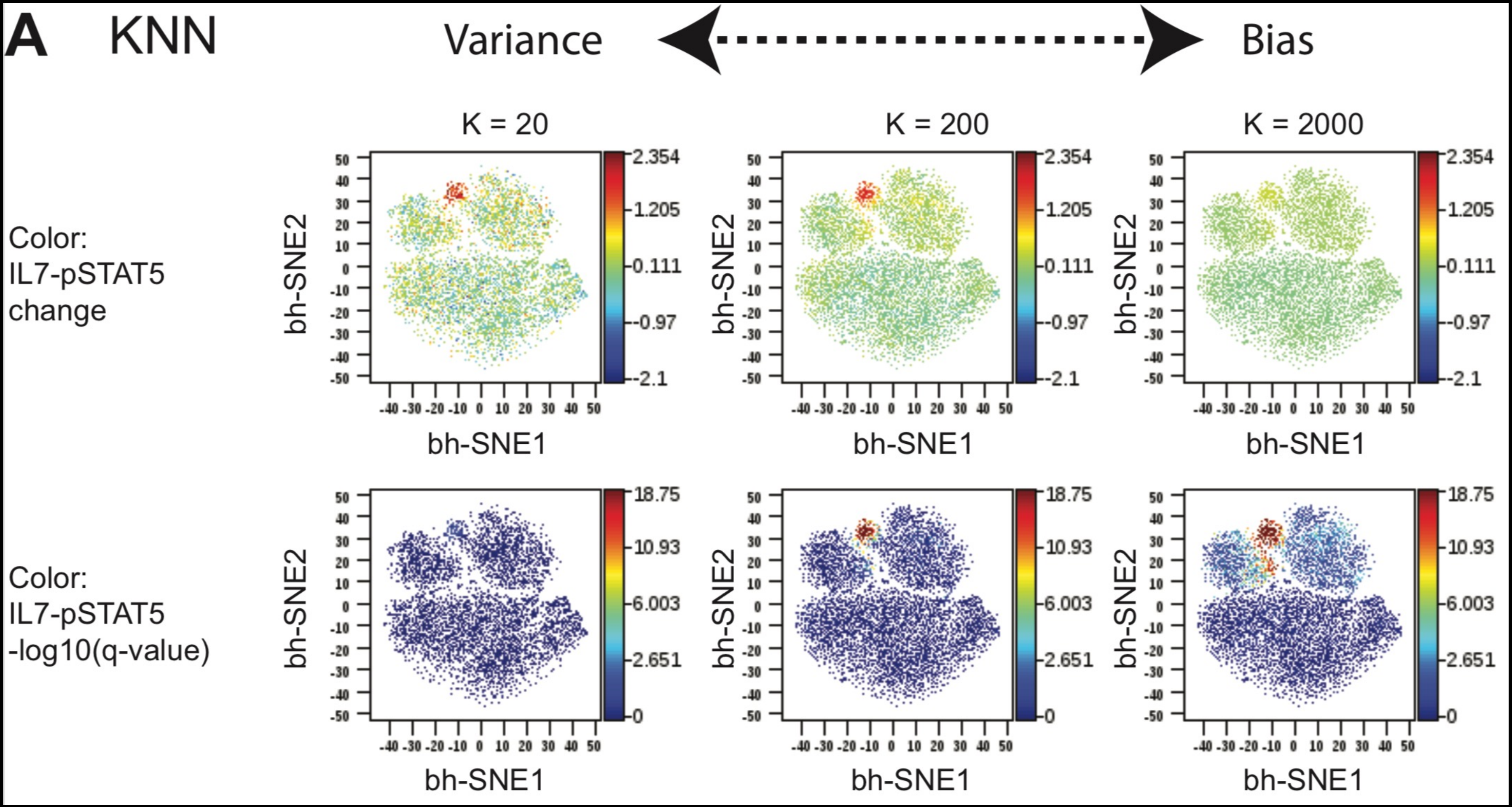
Future direction: toward a tool for people who want to gate their t-SNE maps



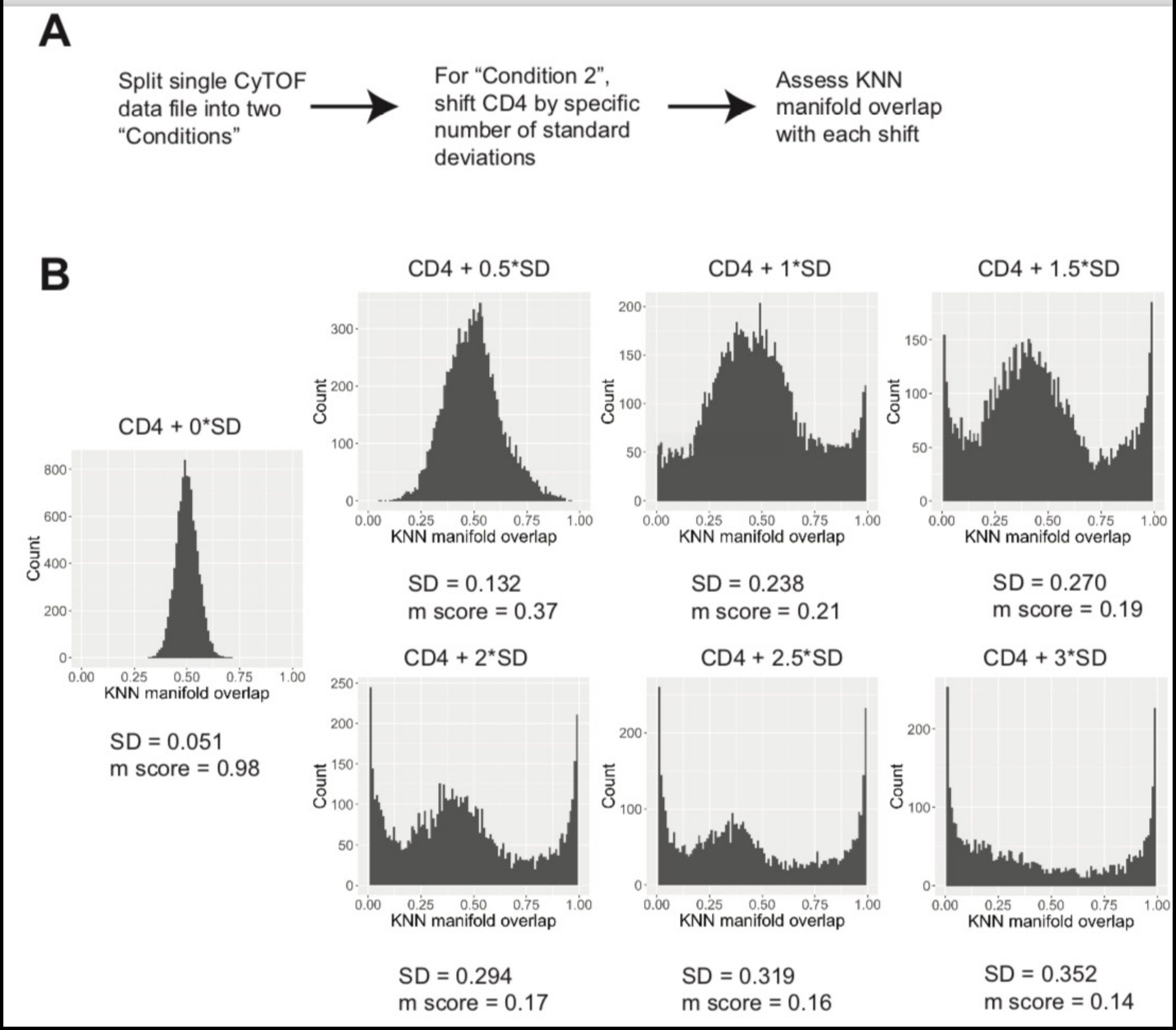
Step 1: draw a gate (or cluster)

Step 2: computer outputs % accuracy compared to high-d space

Visual of choice of K: bias-variance tradeoff



Synthetically altering data: the sensitivity of KNN



Split single CyTOF data file into two "Conditions" → For "Condition 2", shift CD4 by specific number of standard deviations → Assess KNN manifold overlap with each shift

CD4 + 0*SD: SD = 0.051, m score = 0.98

CD4 + 0.5*SD: SD = 0.132, m score = 0.37

CD4 + 1*SD: SD = 0.238, m score = 0.21

CD4 + 1.5*SD: SD = 0.270, m score = 0.19

CD4 + 2*SD: SD = 0.294, m score = 0.17

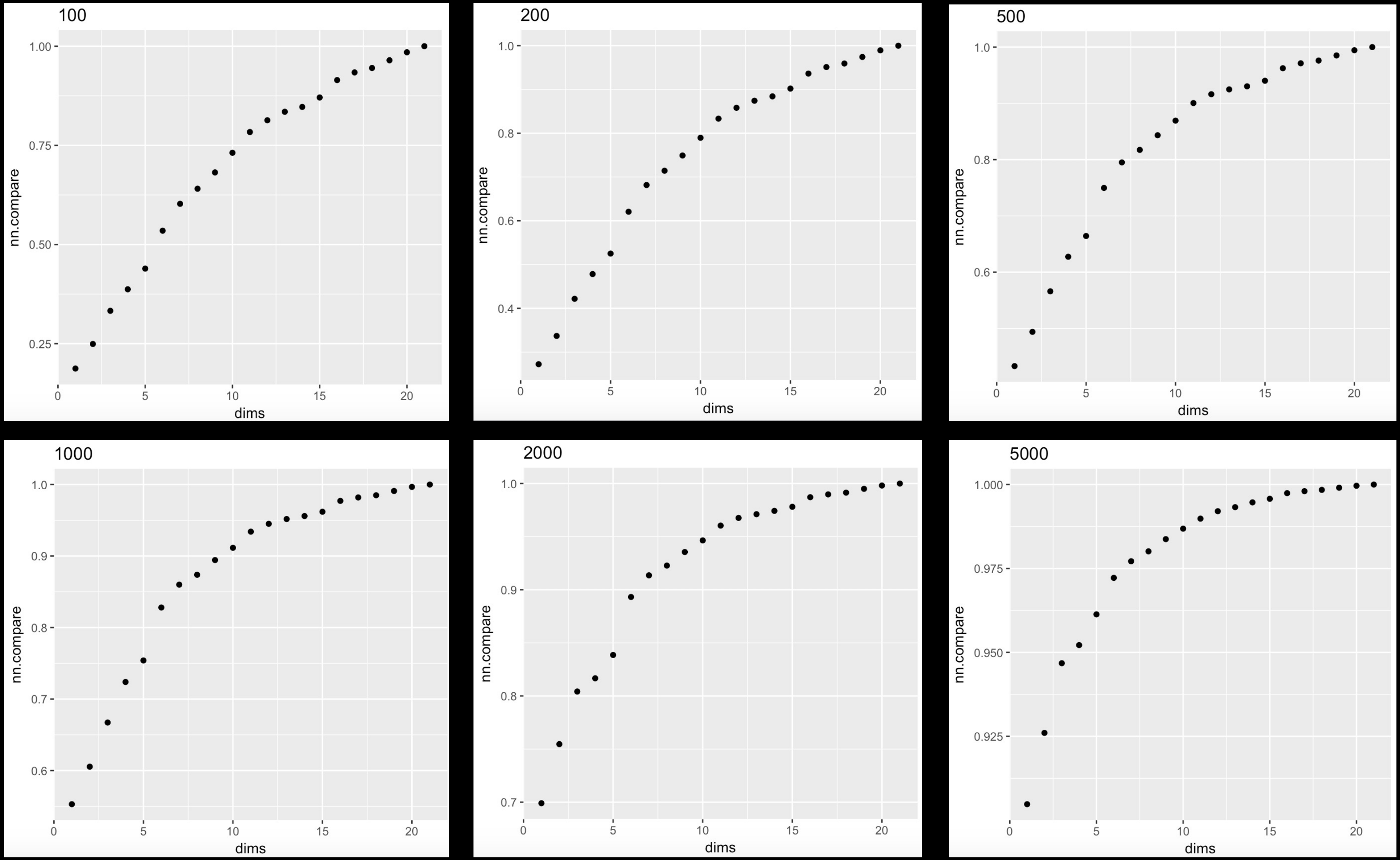
CD4 + 2.5*SD: SD = 0.319, m score = 0.16

CD4 + 3*SD: SD = 0.352, m score = 0.14

Where does KNN fit into a data analysis pipeline

- Initial stages of research:
  - Get an understanding of what your dataset has
    - What markers are relevant
    - How dramatic are the ”differences”
    - Does the data need to be normalized and scaled
    - Are there regions where sparsity increases (eg that could point to negative selection)
  - Use this information to determine the appropriate scaled-up analysis:
    - How many ”clusters” should we expect
    - Where should we expect (and NOT expect) differences

Information loss contains an elbow point

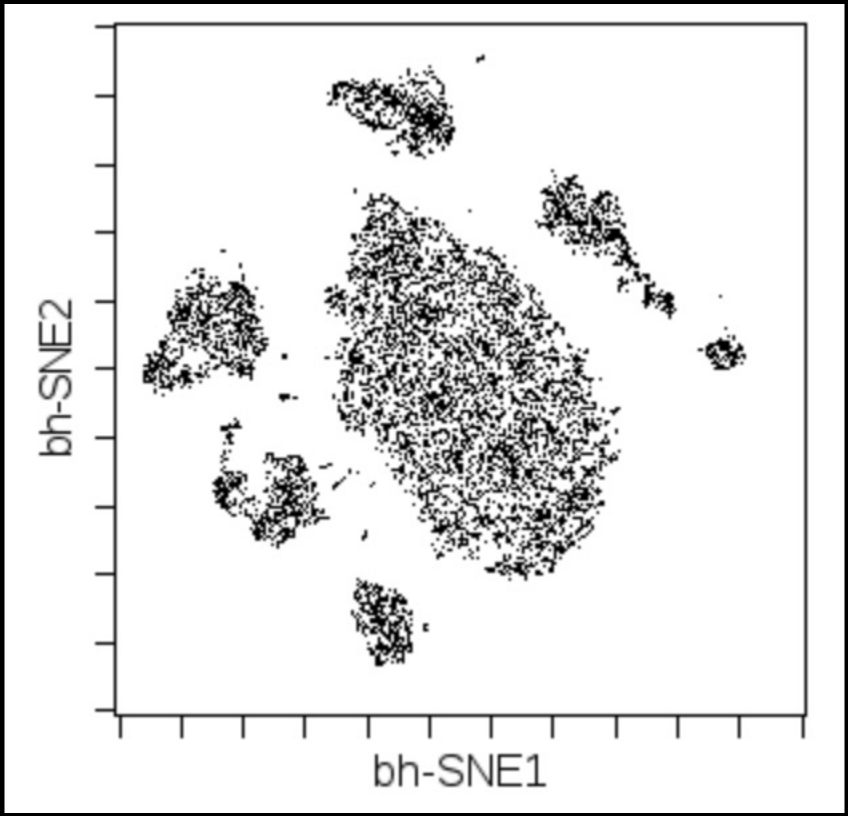


Shared KNN between PCA and original space

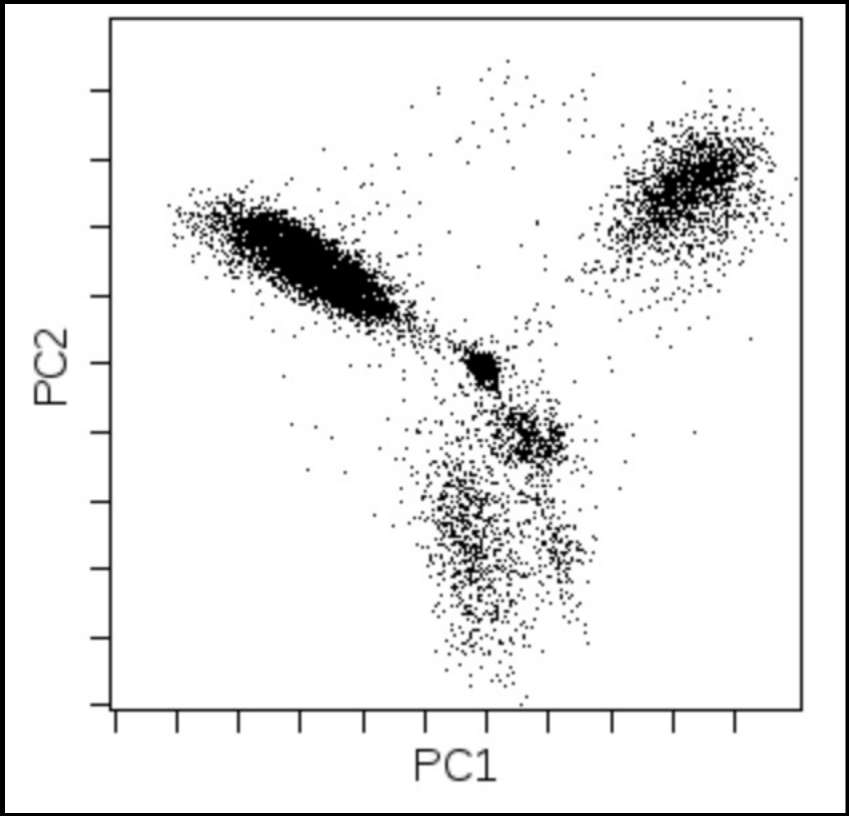
Number of principal components to take KNN from

What t-SNE and PCA look like

t-SNE

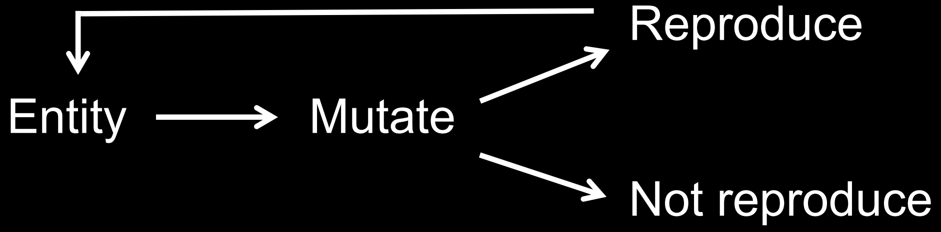


PCA



Fragidakis *et al*
Anesthesiology 2015
Cells: whole blood
Cell number: 10,000

Single cell analysis: the big picture


Reproduce

Entity

Mutate

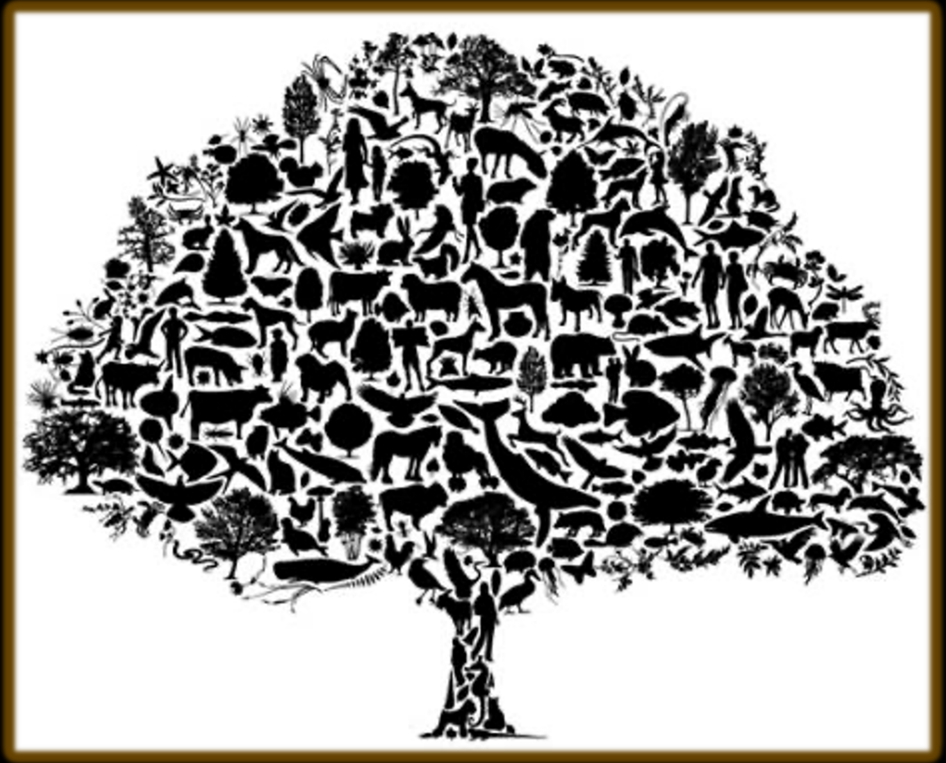
Not reproduce

Normal biology = emergent order

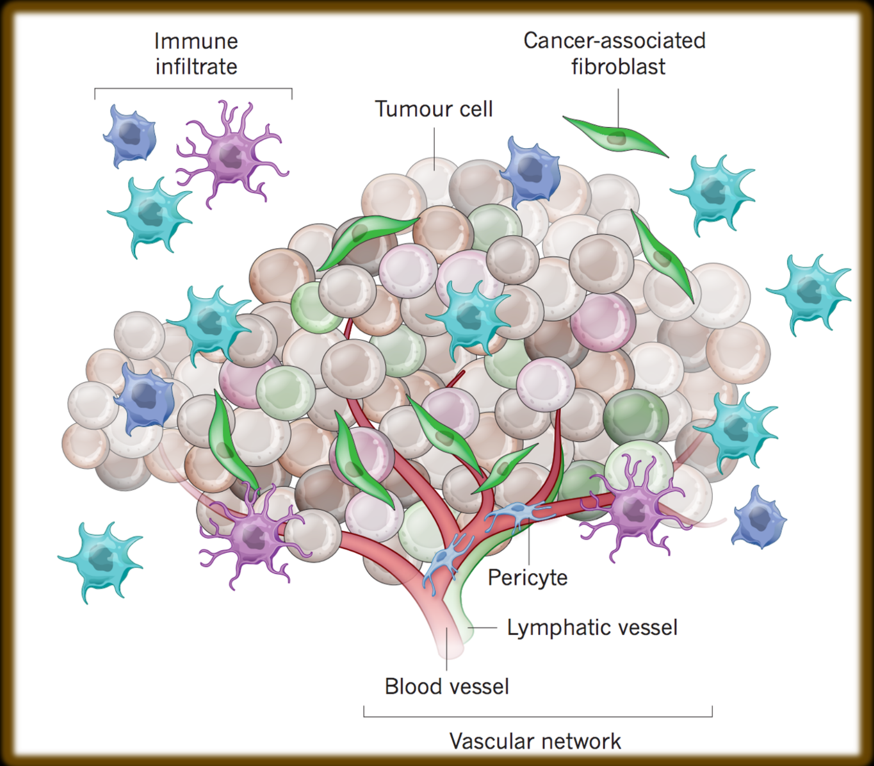
Disease biology = emergent order

Single cell analysis = uncover emergent order

Organismal biodiversity



Single cell biodiversity



Questions?