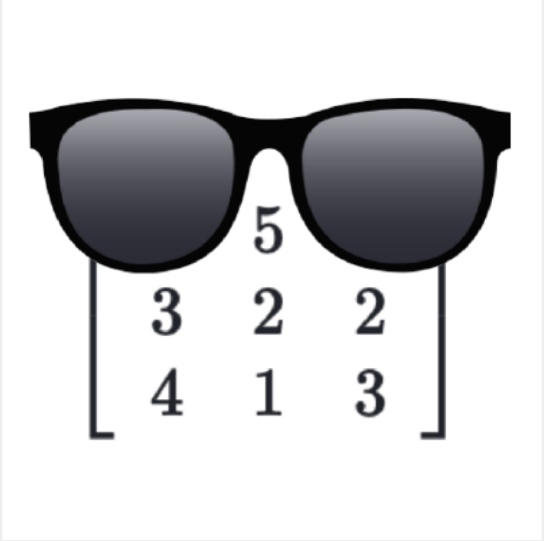
***normjam**: a gentle overview*

normalization refers to the ***within sample*** removal of ***technical*** variability from a scRNA-seq dataset, while maintaining ***biological*** variability



deep ganguli &
rahul satija

un-normalized
un-annotated

genes

cells

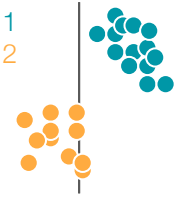
tsne 1

tsne 2

it is difficult to do *anything* with the data until it's normalized

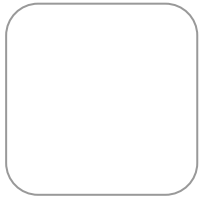
normalized
annotated

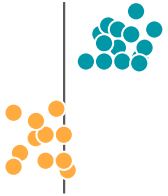
cell type 1
cell type 2



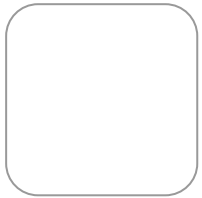
after normalization, it's often easier to to do *biology,* e.g., define cell types

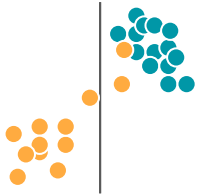
normalization
algorithm 1





normalization
algorithm N

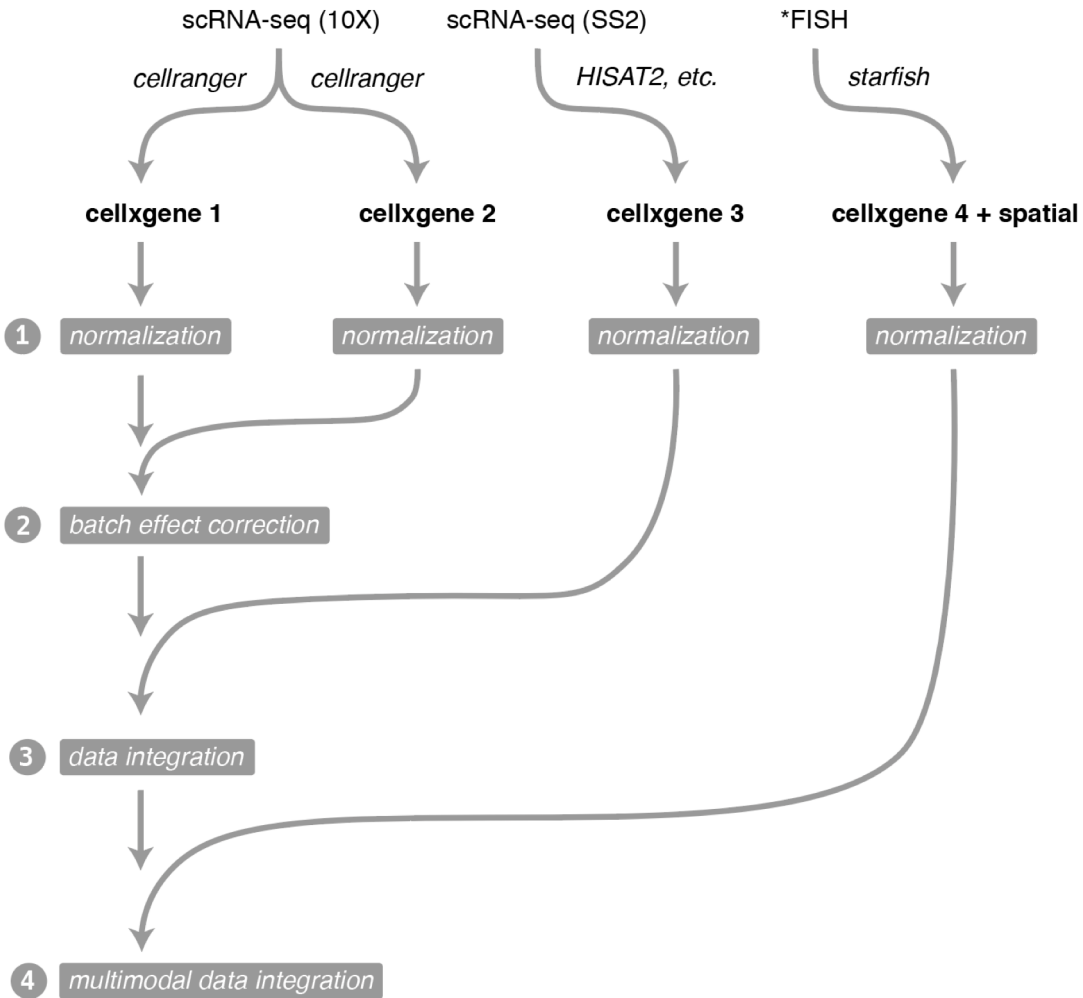




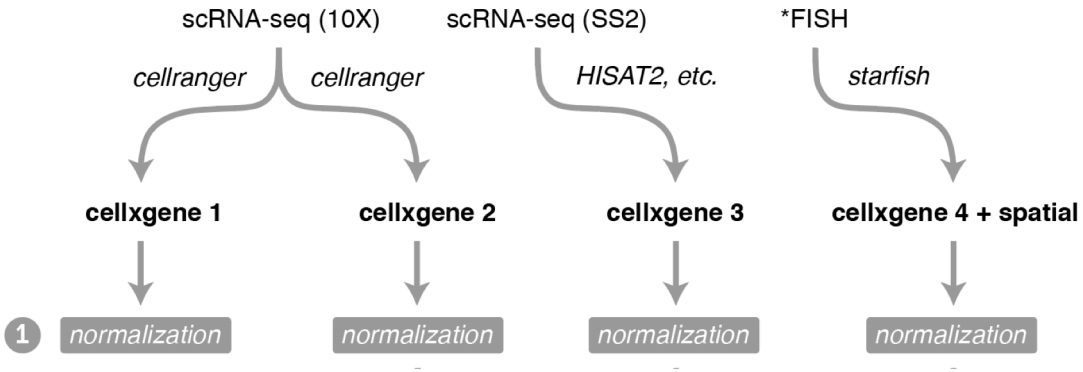
there are many different normalization algorithms. how to evaluate? how to test statistical assumptions?

one small step

- how does error propagate?
- do we want monolithic or modular algorithms?
- can we generalize our thinking across modalities?



today's talks



scTransform - Christoff Hafemeister
glmPCA- Will Townsend
scVI - Romain Lopez

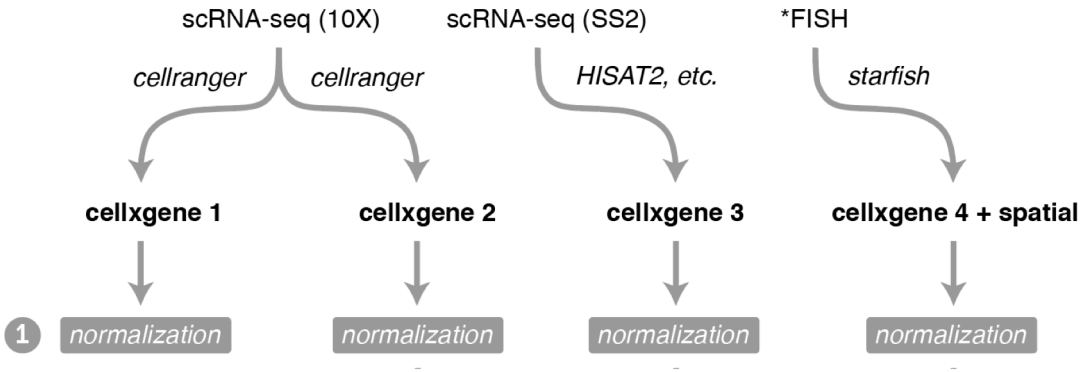
panel discussion

BASICS/scran - Catalina Vallejos
scNorm - Rhonda Bacher
BISCUIT - Sandhya Prabhakar

panel discussion

today's talks

- how do you define normalization? why is it important?
- how do you normalize?
- how do you demonstrate success?
- where does your method break?
- how should we spend tomorrow working?



scTransform - Christoff Hafemeister
glmPCA- Will Townsend
scVI - Romain Lopez

panel discussion

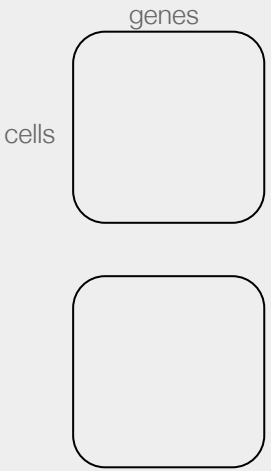
BASICS/scran - Catalina Vallejos
scNorm - Rhonda Bacher
BISCUIT - Sandhya Prabhakar

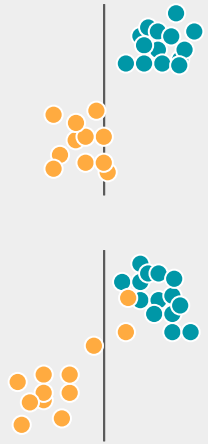
panel discussion

today's discussions

normalization algorithm 1

normalization algorithm N





there are many different normalization algorithms. how to evaluate? how to test statistical assumptions?

today's discussions

what is the right distribution for technical variability? - josh/stephanie

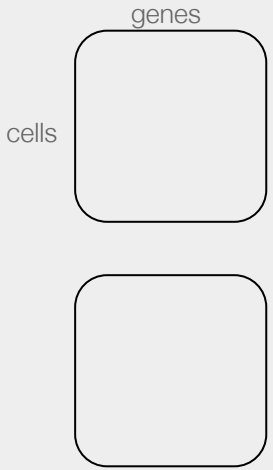
what transform should we use for variance stabilization? - rahul/val

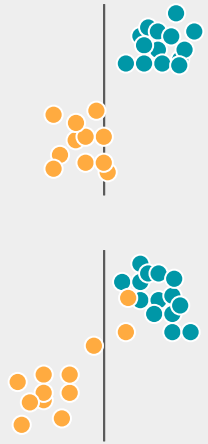
how do we define success? - rafael/peter

normalization algorithm 1

⋮

normalization algorithm N





there are many different normalization algorithms. how to evaluate? how to test statistical assumptions?

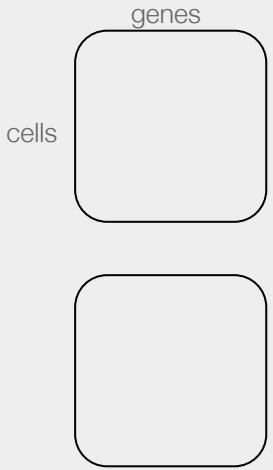
today's discussions

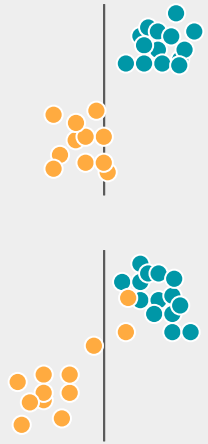
what are useful potential biological/negative controls? - catalina/angela

should we benchmark? how? - dana/ambrose

normalization algorithm 1

normalization algorithm N





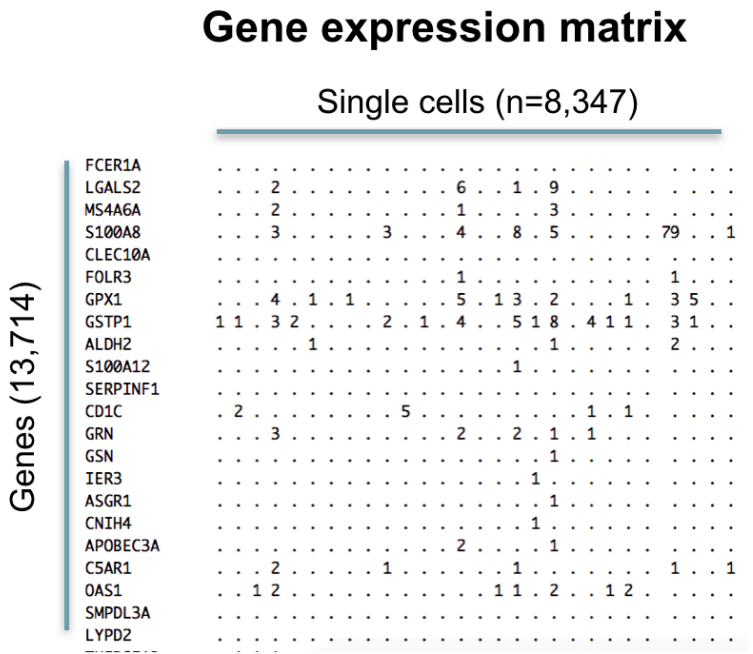
there are many different normalization algorithms. how to evaluate? how to test statistical assumptions?

tomorrow we jam

today we record & synthesize everyone's thoughts on how to jam

Wednesday, November 20th	
8:30 - 9:00 am	Breakfast
9:00 - 9:15 am	Opening - *Deep Ganguli and Rahul Satija*
9:15 - 10:45 am	Working Session 1
10:45 - 11:00 am	Break
11:00 - 11:15 am	Report out from Session 1
11:15 - 12:45 pm	Working Session 2
12:45 - 1:45 pm	Lunch
1:45 - 2:00 pm	Report out from Session 2
2:00 - 3:30 pm	Working Session 3
3:30 - 3:45 pm	Break
3:45 - 4:00 pm	Report out from Session 3
4:00 - 5:30 pm	Group discussion
5:30 - 6:00 pm	Break and walk over to restaurant
6:00 - 7:00 pm	Happy hour
7:00 - 9:00 pm	Dinner and social activities

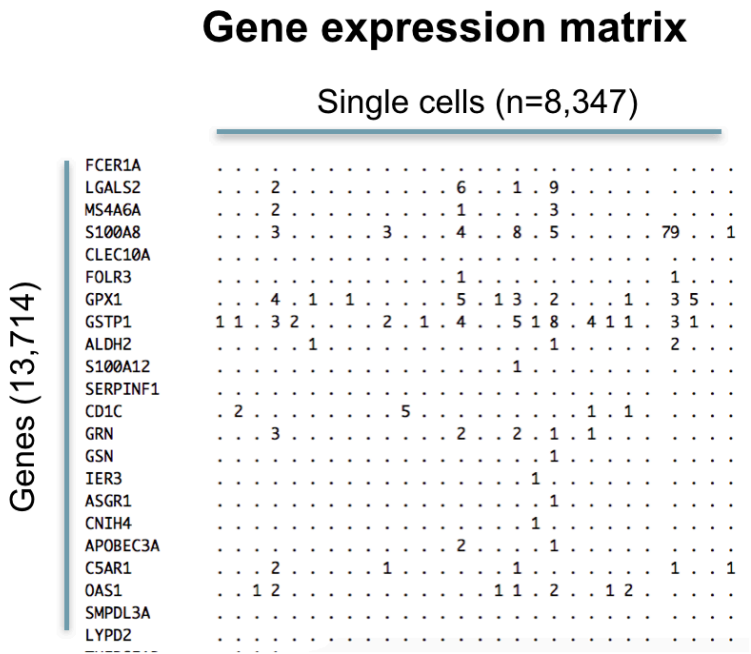
What defines variance in scRNA-data?



1. **"Technical" Sources**

2. **"Biological" Sources**

What defines variance in scRNA-data?



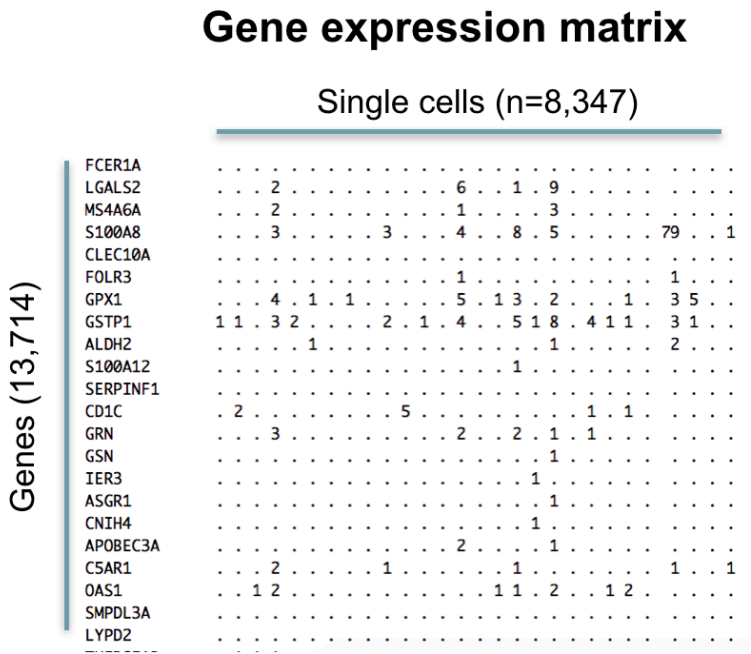
1. "Technical" Sources

- Stochastic Loss (cell-specific rate)

2. "Biological" Sources

- Total RNA content (cell-specific size)

What defines variance in scRNA-data?



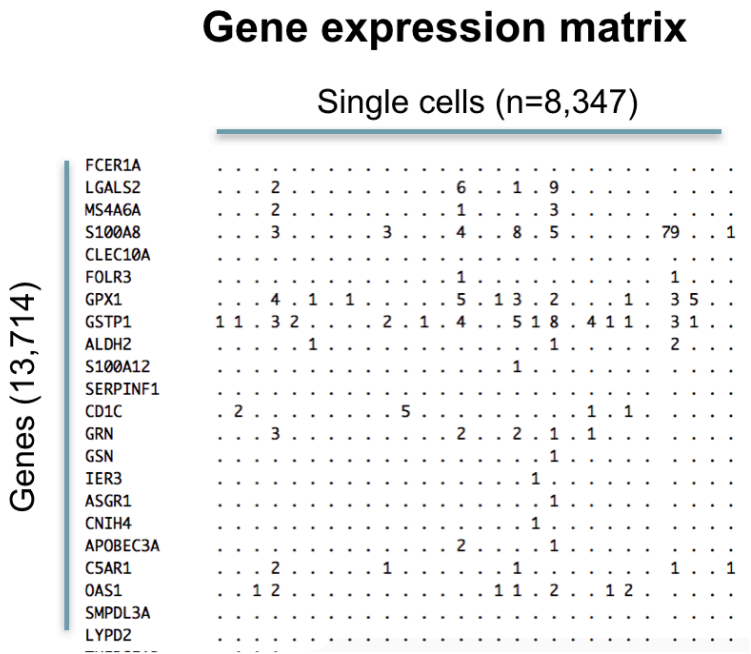
1. "Technical" Sources

- Stochastic Loss (cell-specific rate)
- Amplification/PCR (gene-specific)

2. "Biological" Sources

- Total RNA content (cell-specific size)

What defines variance in scRNA-data?



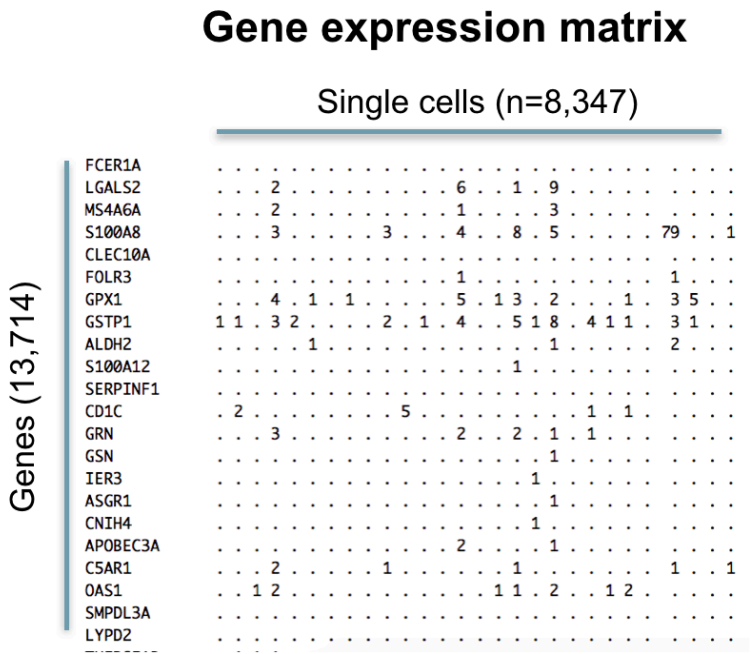
1. "Technical" Sources

- Stochastic Loss (cell-specific rate)
- Amplification/PCR (gene-specific)

2. "Biological" Sources

- Total RNA content (cell-specific size)
- 'Homogeneous' noise (baseline)
- Variation in cell state

What defines variance in scRNA-data?



1. "Technical" Sources

- Stochastic Loss (cell-specific rate)
- Amplification/PCR (gene-specific)

2. "Biological" Sources

- Total RNA content (cell-specific size)
- 'Homogeneous' noise (baseline)
- Variation in cell state

We normalize data to order to focus our analyses on specific sources of variation

What defines a normalization method?

What are the underlying assumptions for technical vs. biological noise?

What data transformations are applied?

How are features selected/weighted for downstream analysis?

Are all cells normalized in the same way?

Are all genes normalized in the same way?

The 'standard' log-normalization approach

What are the underlying assumptions for technical vs. biological noise?

* All cells have the same underlying 'size'

What data transformations are applied?

* Rescale values based on size-factors
* Logarithmic transformation (and pseudocount addition)
* Z-score transformation (variance stabilization)

How are features selected/weighted for downstream analysis?

* All variable features have equal weight
* All non-variable features have zero weight

Are all cells normalized in the same way?

* Yes

Are all genes normalized in the same way?

* Yes

Improved methods for normalization

What are the underlying assumptions for technical vs. biological noise?

* All cells within a cluster have the same underlying 'size' (BISCUIT/scran)
* Construct GLM to predict gene expression from total UMI count (SCNorm, SCTransform (SCT), glmpca,…)

What data transformations are applied?

* Logarithmic transformation (lognormal distribution; BISCUIT)
* Rescale values after modeling mean-variance relationship (scran)
* Residuals of statistical model (raw/Pearson/Deviance; SCT, SCNorm)
* Fully probabilistic (likelihood-based) framework (scVI, glmpca)

How are features selected/weighted for downstream analysis?

* Features weighted by residual biological variance (SCT, glmpca, scVI)

Are all cells normalized in the same way?

* Perform pre-clustering prior to normalization (BISCUIT/scran)

Are all genes normalized in the same way?

* Different strategies for highly and lowly expressed genes (SCNorm, SCT)

The NormJam 'Zeitgeist'

We should explore alternatives to log-transformation

* Log transformation is intuitive, commonly used in genomics, reduces sensitivity to extreme values, highlights robust trends.

* However, pseudocount addition is suboptimal for sparse data

* Additionally, may dampen biological variation

The NormJam 'Zeitgeist'

Lowly and highly expression genes require different normalization strategies

* Both the underlying noise model, and the relationship between technical noise and gene expression, are not consistent across gene groups.

* Normalization errors in highly expressed genes are common and deleterious

* How should we overcome this (binning? Regularization?)

Pre-clustering is likely beneficial for normalization

* Likely helps to distinguish between biological and technical noise, especially when very different cell types are present

* Clear evidence is lacking, but should be conducive to benchmarking

* How should we pre-cluster? Iterative normalization/clustering?

The NormJam 'Zeitgeist'

Lowly and highly expression genes require different normalization strategies

* Both the underlying noise model, and the relationship between technical noise and gene expression, are not consistent across gene groups.

* Normalization errors in highly expressed genes are common and deleterious

* How should we overcome this (binning? Regularization?)

Pre-clustering is likely beneficial for normalization

* Likely helps to distinguish between biological and technical noise, especially when very different cell types are present

* Clear evidence is lacking, but should be conducive to benchmarking

* How should we pre-cluster? Iterative normalization/clustering?

The NormJam 'Zeitgeist'

It is difficult to trust biological benchmarks

* Typical benchmark: Normalize dataset with different methods, calculate silhouette distances based on known annotations

* Challenge: Extremely reliant on annotations

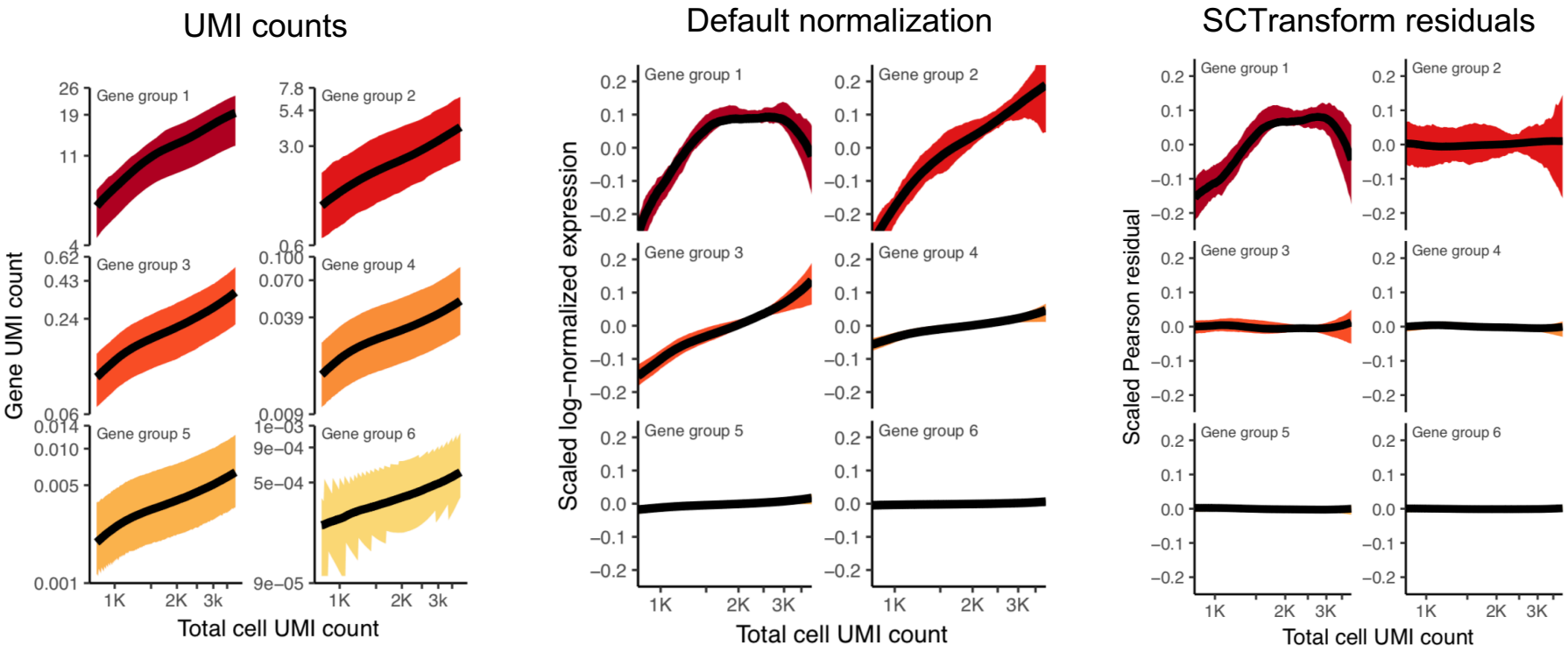
* Challenge: Gives almost no weight to rare subpopulations

* Challenge: Unclear how to benchmark on continuous manifolds

* Solution: Metrics based on objective (non-biological) criteria?

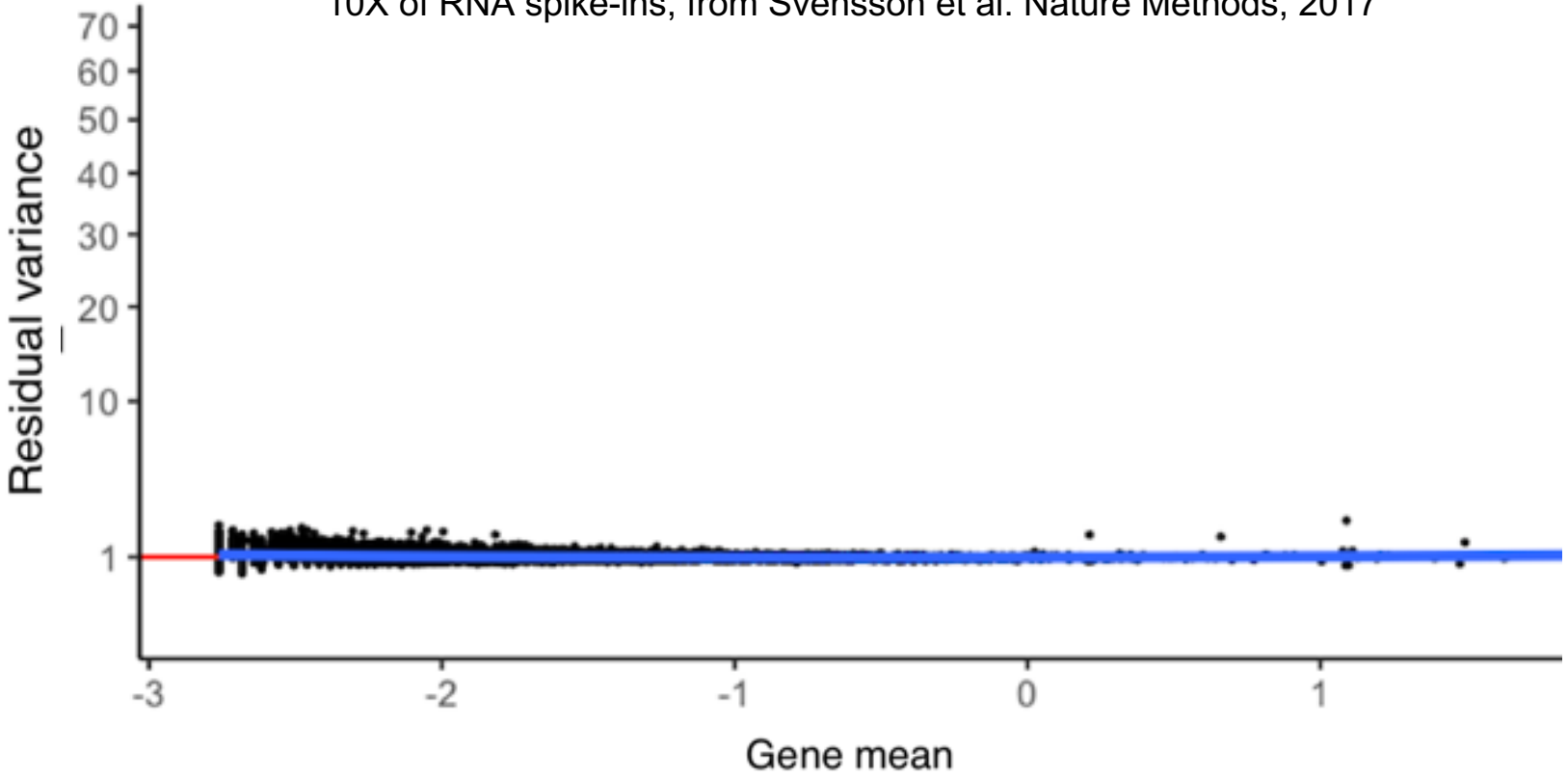
* Solution: Datasets with complementary 'ground truth', where annotations are derived (in-part) from secondary sources of (**accurate**) information

The NormJam 'Zeitgeist'



The NormJam 'Zeitgeist'

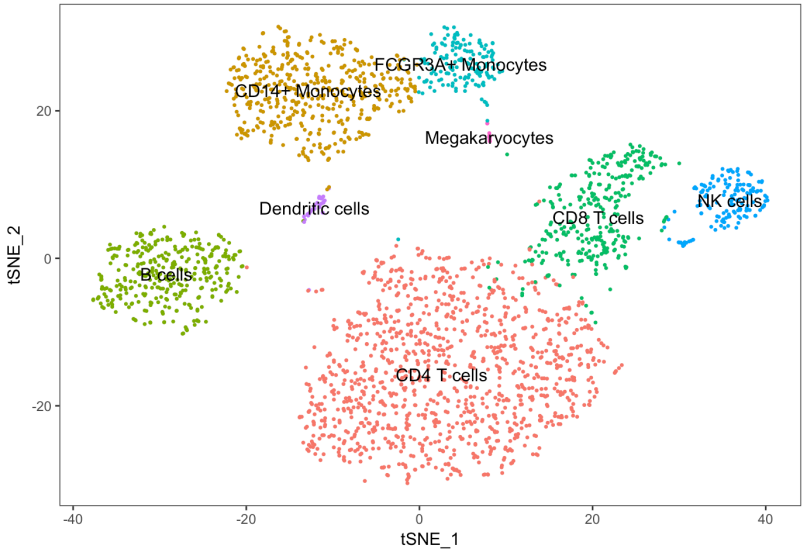
10X of RNA spike-ins, from Svensson et al. Nature Methods, 2017



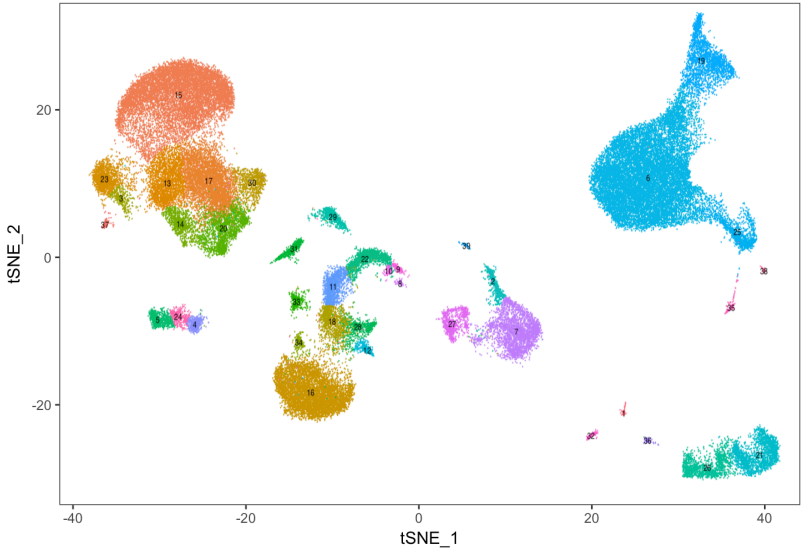
'Ground truth' datasets

Human PBMC, ~3,000 cells
(10X Genomics v1)

Unsupervised analysis
Default normalization



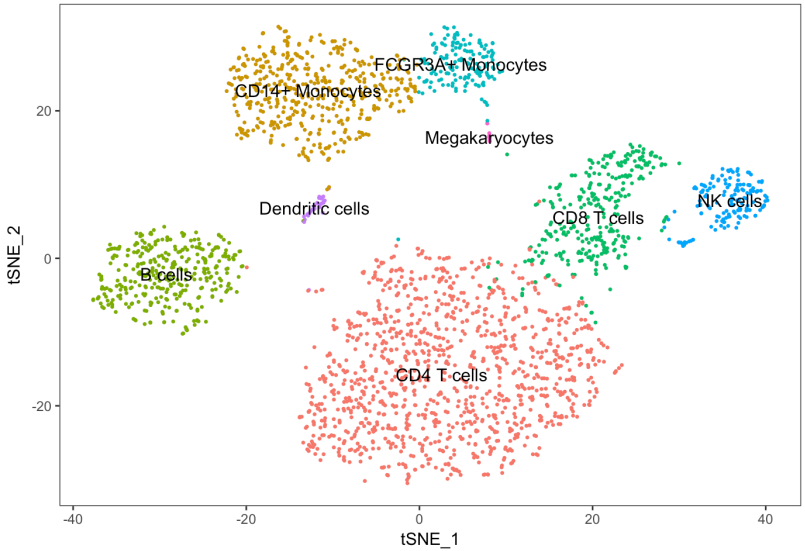
Reference dataset
(~300,000 human PBMCs)



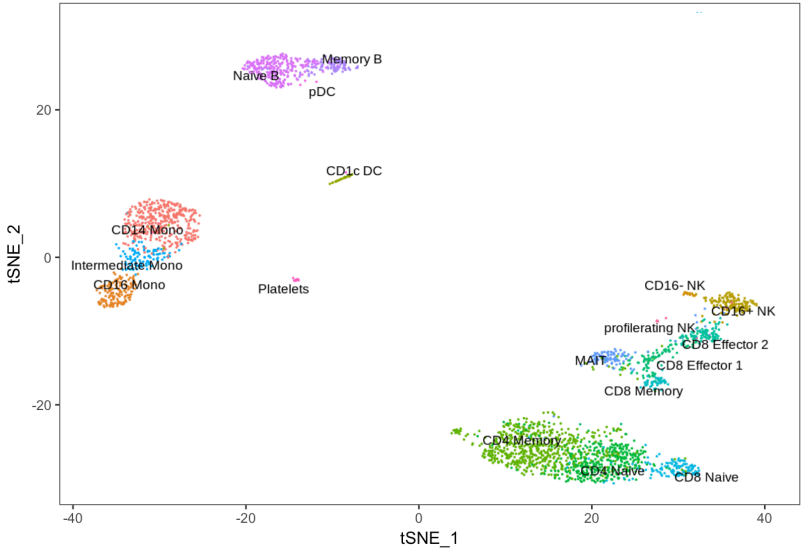
‘Ground truth’ datasets

Human PBMC, ~3,000 cells
(10X Genomics v1)

Unsupervised analysis
Default normalization



Supervised analysis
SCTransform normalization



Challenges in benchmarking

Dataset 2: ~31,000 human bone marrow cells, 25 antibodies (CITE-seq)
(Stuart*, Butler* et al, Cell 2019)
* Clusters were identified using both protein and RNA data. Modalities are consistent, but protein data is far more robust
* Yields high-resolution annotations, even when compared to HCA.
* Challenge is to recreate clustering results using RNA alone

Dataset 3: ~40,000 emerging cells from mouse endoderm gut tube
(Nowotschin et al., Nature 2019)
* Data exhibits extremely deep sequencing
* For each cell, we downsampled the depth ~ Unif(0.05, 1)
* Challenge is to reconstruct the original relationships between cells from the downsampled data
* Trajectories represent both pseudotime and pseudospace

Possible tasks for Day 2

1. **How do we model technical noise in scRNA-seq data**

   * Test different distributions on data from 'homogeneous' populations. Examine goodness-of-fit and minimize free parameters.

   * Can we identify cases that clearly **violate** particular models?

   * Consensus here would be important milestone for scRNA-seq analysis

2. **Benchmark the relative value of pre-clustering on normalization**

   * Easy to include, even for methods which otherwise lack pre-clustering

   * How robust is improvement to perturbations in clustering parameters?

   * Can we improve robustness? (soft-clustering, iterative approach, etc.)

Possible tasks for Day 2

1. **How do we model technical noise in scRNA-seq data**

   * Test different distributions on data from 'homogeneous' populations. Examine goodness-of-fit and minimize free parameters.

   * Can we identify cases that clearly **violate** particular models?

   * Consensus here would be important milestone for scRNA-seq analysis

2. **Benchmark the relative value of pre-clustering on normalization**

   * Easy to include, even for methods which otherwise lack pre-clustering

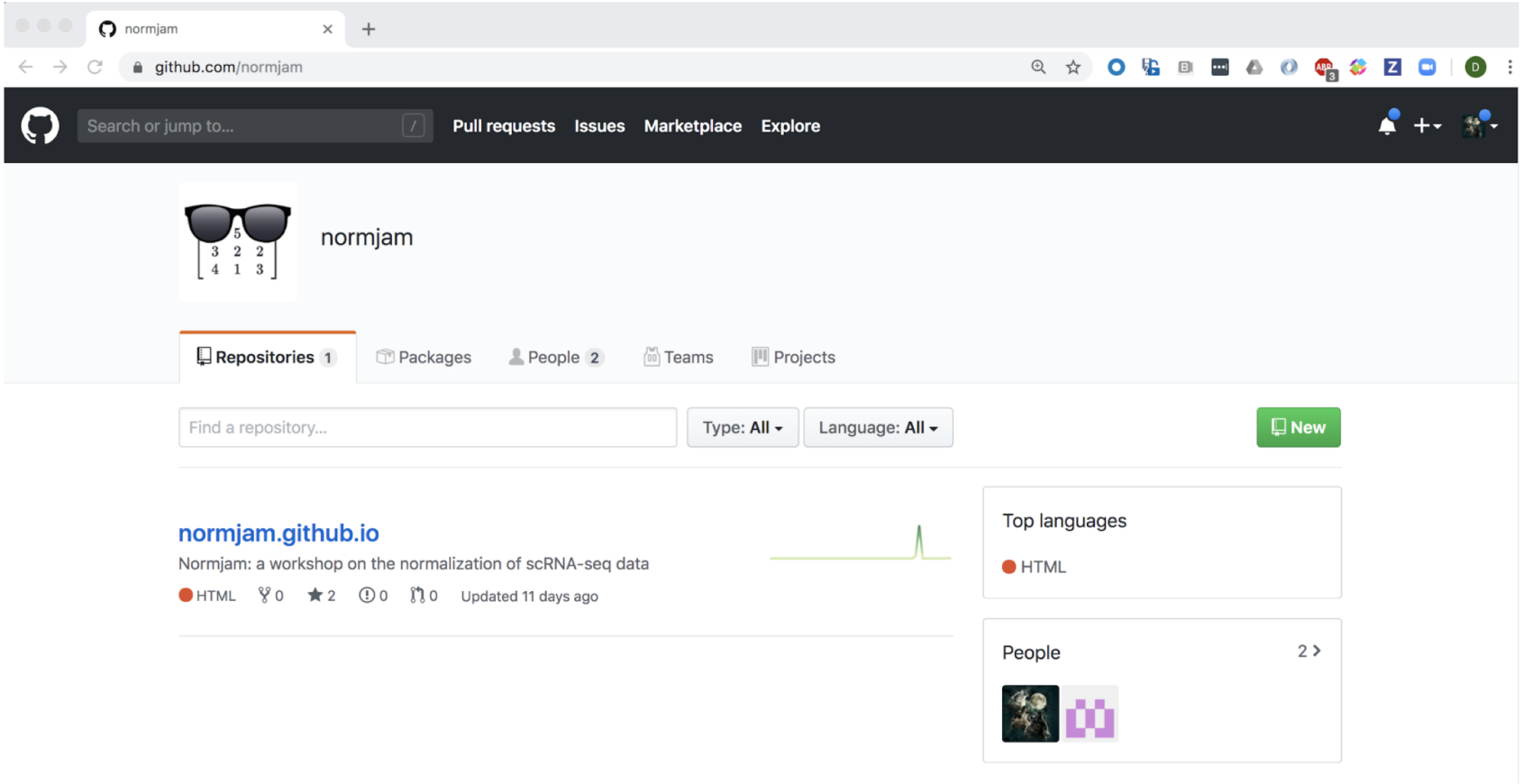
   * How robust is improvement to perturbations in clustering parameters?

   * Can we improve robustness? (soft-clustering, iterative approach, etc.)

life after normjam

“we’re jammin, and I hope this jam is gonna last” - Bob Marley

put code/markdown here: https://github.com/normjam



think about how we want to keep jammin and communicate our findings