Random Matrix Theory as a tool for analysis of single-cell sequencing data

after Aparicio, Bordyuh, Blumberg, and Rabadan

Jeremy Teitelbaum
UConn Department of Mathematics
JAX Visiting Scientist
January 22, 2019

Reference

Quasi-universality in single-cell sequencing data

Luis Aparicio, Mykola Bordyuh, Andrew J. Blumberg, Paul Rabadan

https://www.biorxiv.org/content/early/2018/10/05/426239

See also

An introduction to random matrices

Greg Anderson, Alice Guionnet, Ofer Zietouni

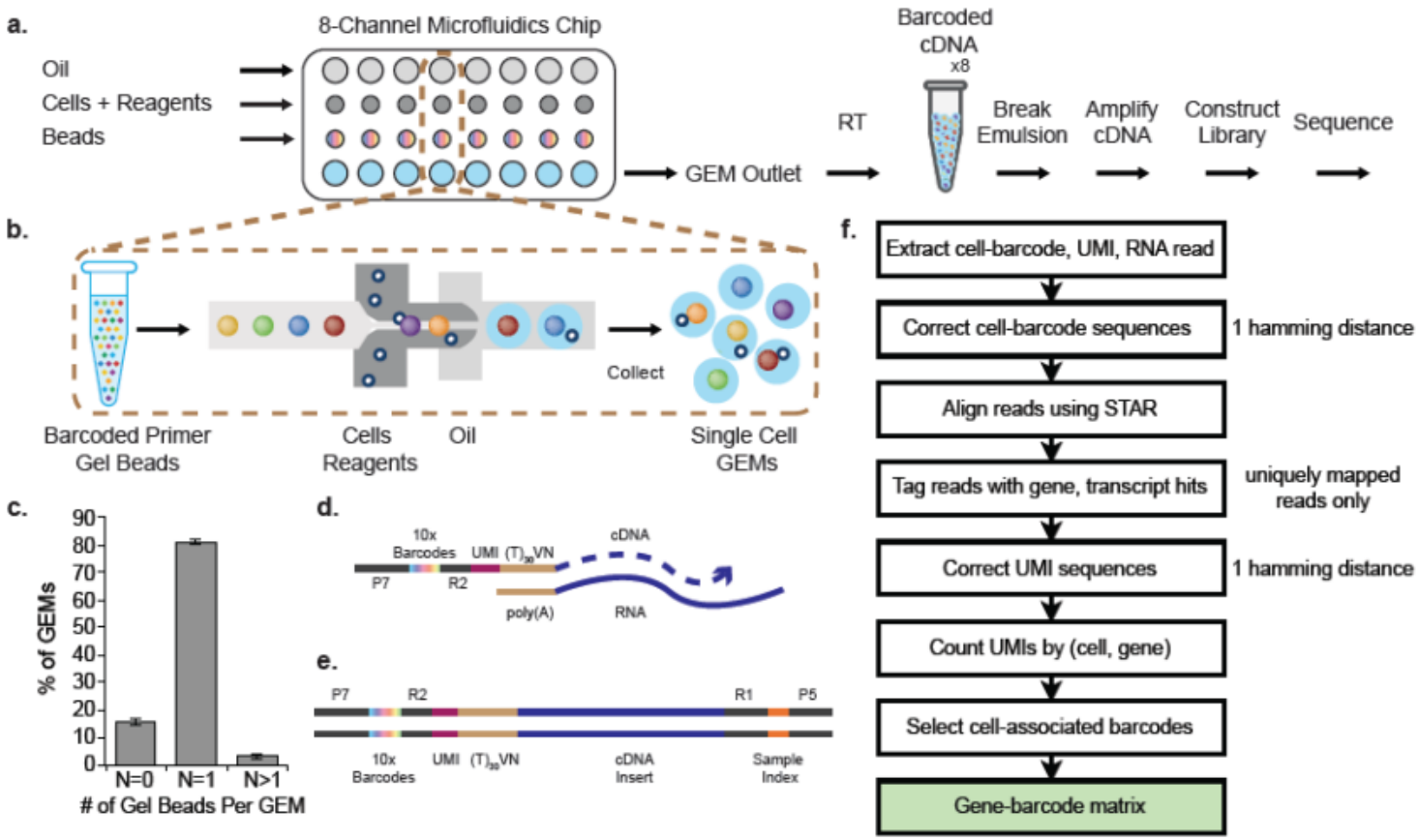
Cambridge University Press, 2011

Single-Cell RNA Data

High throughput single-cell methods use bar-coding to label and count individual RNA transcripts on a cell by cell basis. The process is "noisy" and there are many proposed ways to model the data statistically.

This talk describes ideas derived from the mathematical theory of random matrices to guide the analysis.

Brief overview of scRNA-seq on 10x Genomics Platform



See *Massively parallel digital transcriptional profiling of single cells* by Zheng, et. al. bioRxiv 10.1101/1065912.

Complicating factors with the process

- Relationship of cell barcodes to "real" cells is inexact, and most putative cells have very few associated transcripts.
- The transcript counts are based on the 3' end of the RNA, so the identification of a gene may be ambiguous and alternative forms of the transcript aren't visible.
- Much of the RNA is missed, so even with many transcripts, most genes have zero counts.

Single-Cell RNA Data 2

Preliminary preparation of output from a high-throughput single cell experiment:

- Set a threshold number of transcripts to decide that a bar-code comes from a cell
- Filter out seqences that don't map uniquely to the genome

An example

10x GENOMICS

Cell Ranger · JC18001 ·

SUMMARY ANALYSIS

Estimated Number of Cells

3,321

Mean Reads per Cell	Median Genes per Cell
110,894	4,477

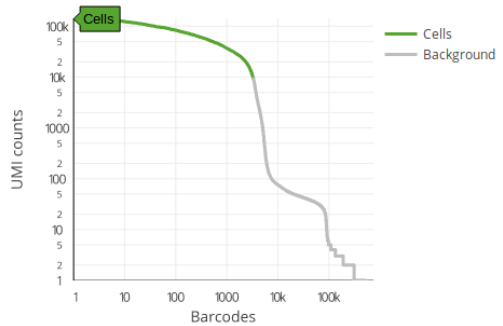
Sequencing

Number of Reads	368,281,552
Valid Barcodes	98.5%
Sequencing Saturation	51.1%
Q30 Bases in Barcode	98.1%
Q30 Bases in RNA Read	84.1%
Q30 Bases in Sample Index	96.6%
Q30 Bases in UMI	98.3%

Mapping

Reads Mapped to Genome	92.2%
Reads Mapped Confidently to Genome	89.5%
Reads Mapped Confidently to Intergenic Regions	5.1%
Reads Mapped Confidently to Intronic Regions	13.4%
Reads Mapped Confidently to Exonic Regions	71.0%
Reads Mapped Confidently to Transcriptome	67.2%
Reads Mapped Antisense to Gene	1.1%

Cells



Estimated Number of Cells	3,321
Fraction Reads in Cells	91.9%
Mean Reads per Cell	110,894
Median Genes per Cell	4,477
Total Genes Detected	23,105
Median UMI Counts per Cell	27,505

Sample

Name	JC18001
Description	
Transcriptome	hg19
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.1

Single-cell RNA Data 3

- Barcodes where the total UMI counts were less than 10% of the maximum are considered artifacts and dropped. This left about 3000 cells.
- About 2/3 of the sequences mapped confidently (i.e. more or less uniquely) to the transcriptome.
- Most of the genes in each cell showed zero counts.

Data

A large integer valued matrix (3000 cells by 30000 genes) most of whose entries are zero.

Goal

Identify patterns of gene expression in this data that characterize subpopulations of cells

Linear Analysis

A standard first step in the analysis of such data is to use a version of "principal component analysis" to reduce the dimension of the data to something comprehensible by humans.

2 steps

First, a linear step to reduce to, say, 30 or 50 dimensions, followed by:

- a graph based method such as TSNE to reduce to two dimensions, followed by a clustering algorithm, or
- a direct application of a clustering method to the 30 dimensional reduced matrix.

This talk will focus on the linear part.

Linear Analysis: A walkthrough

1. We begin with our matrix X, whose rows correspond to cells C and whose columns correspond to genes G.
2. We normalize the matrix, yielding a new matrix $\tilde{X}$.
   - 2.1 We divide each row by the total transcript count in that row (and maybe multiply by a million)
   - 2.2 We take the logarithm of of entry (actually, it's usually $\log_2(1 + x)$ zeros)
   - 2.3 We standardize each column so that the (log)-measurements of each gene in the normalized matrix have mean 0 and standard deviation 1.

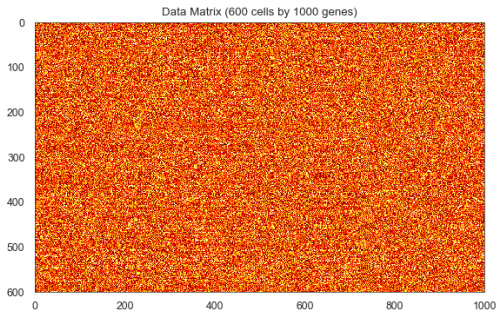
Walkthrough, cont'd

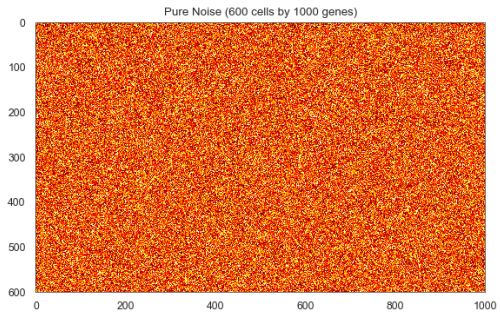
3. We compute the covariance matrix $W = \tilde{X}\tilde{X}^t/G$. W is a $C \times C$ matrix. *The diagonal entries entries are the variances of the gene expressions for each cell; the off diagonal entries are the covariances.*
4. We diagonalize the covariance matrix and obtain C eigenvectors and C associated eigenvalues. *The eigenvectors can be thought of as 'pseudo-cells' that capture some particular variance pattern in the genes.*

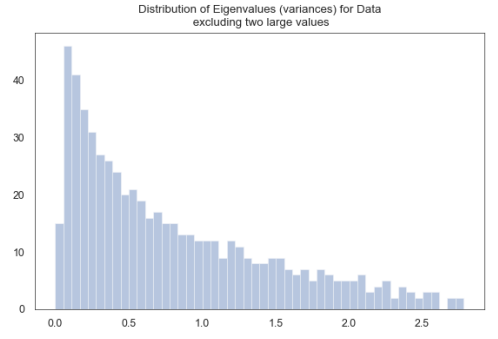
Walkthrough, cont'd

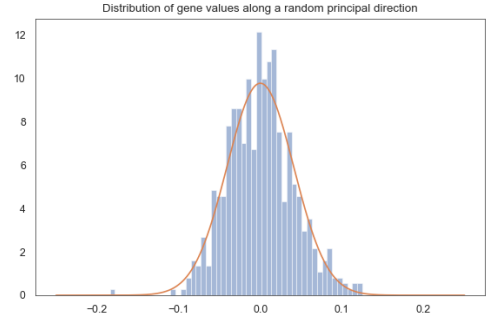
5. We look at the C eigenvalues and select a subset of, say, K large ones as corresponding to signal. *The eigenvectors corresponding to large eigenvalues capture significant variation among the expression data*
6. We project the cells into the subspace spanned by the K signal dimensions. This gives us a $C \times K$ matrix that hopefully contains all of the useful information about the data. *Projecting into this subspace focuses us on relevant information that will hopefully discriminate among different patterns in the data*

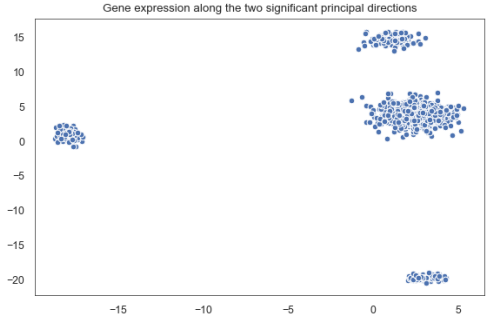
Walkthrough, cont'd

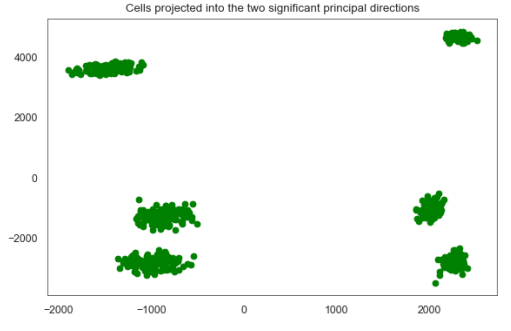












Insights from Random Matrix Theory

1. To distinguish "signal" from "noise", we need to have a clear understanding of noise.
2. Random matrix theory describes the noise.
3. Results originate with quantum physics and insights of Wigner. Now a major area of interest in probability.

Key results

- How are the eigenvalues (the "principal values") distributed when a matrix has no signal at all?
- How big is the largest eigenvalue/principal value that can be explained by noise?
- How are the principal directions distributed among the coordinate axes? Are there any preferred directions in the absence of signal?

The Marchenko-Pastur Distribution

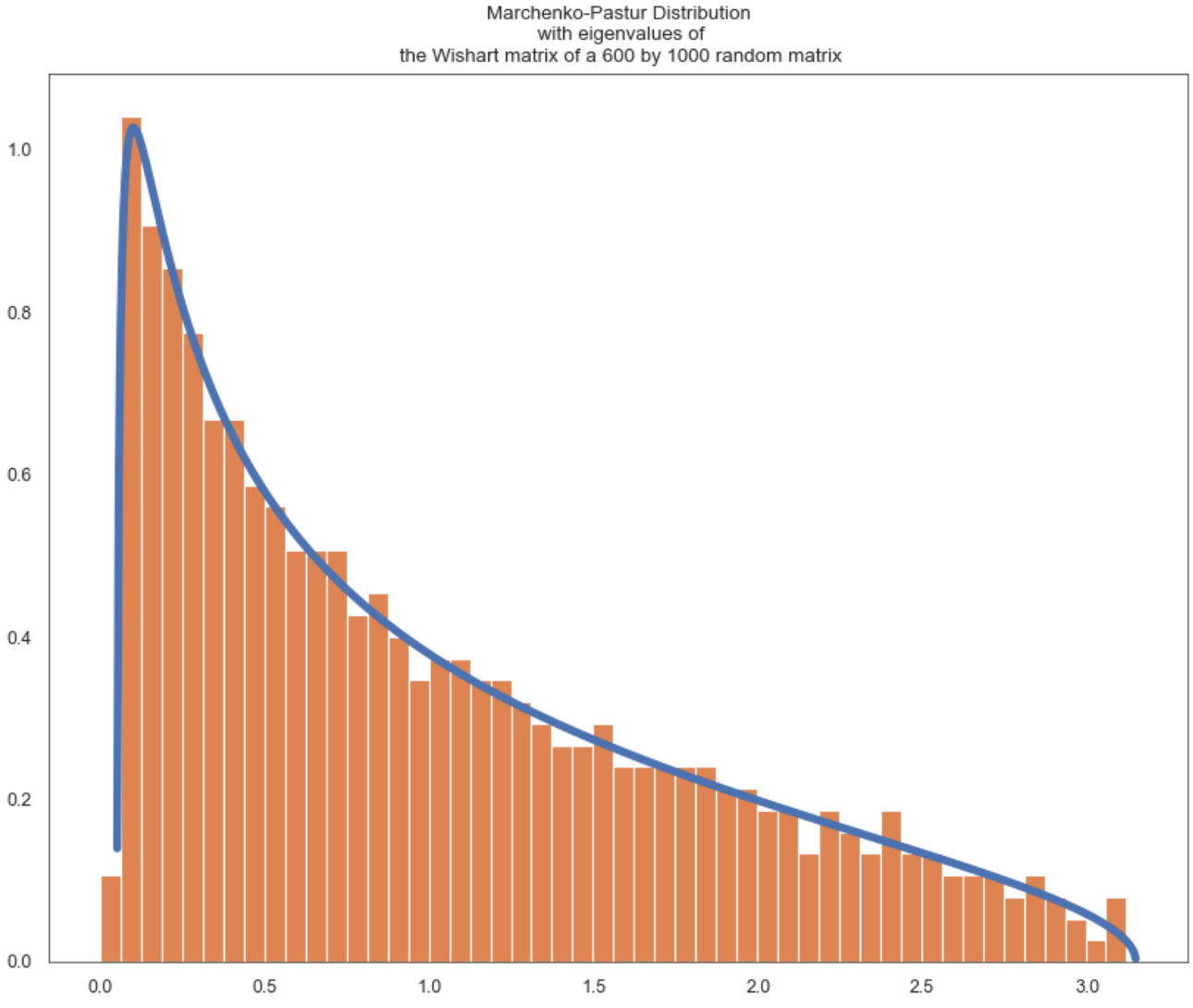
How are the eigenvalues distributed?

Let X be an $n \times k$ matrix whose entries are chosen at random from a distribution with mean zero and variance 1. (I omit some technical hypotheses). Let $W = XX^t/k$ be the associated covariance matrix, also called the **Wishart Matrix**. Suppose for simplicity that $n \leq k$. If we consider sequences of such matrices with increasing numbers of rows n and columns k, under the assumption that $n/k \to \lambda$ for some constant λ, then the limiting distribution of eigenvalues of W is given by the distribution function

$$MP(x) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}$$

where $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$.

The MP Distribution (an example)



A short digression on the beauty of random matrix theory

Consider the case of a large random $N \times N$ matrix with entries that are chosen independently at random from the standard normal distribution. Let $W = XX^T$.

Consider the sequence $tr(W), tr(W^2), tr(W^3), \ldots$.

$$1, 2, 5, 14, 42.1, 133, 432, 1440, \ldots$$

The first few numbers in this sequence $1, 2, 5, 14, 42$ are the beginning of the Catalan numbers:

$$1, 2, 5, 14, 42, 132, 429, 1430, ...$$

Random Matrices and Catalan Numbers

The MP distribution in the square matrix case reduces to a version of Wigner's semicircle distribution

$$f(x) = \frac{1}{2\pi x}\sqrt{x(4-x)}.$$

The moments of this distribution are the Catalan numbers C_n. Therefore, in the limiting case, the expected value of the trace of W^n is C_n.

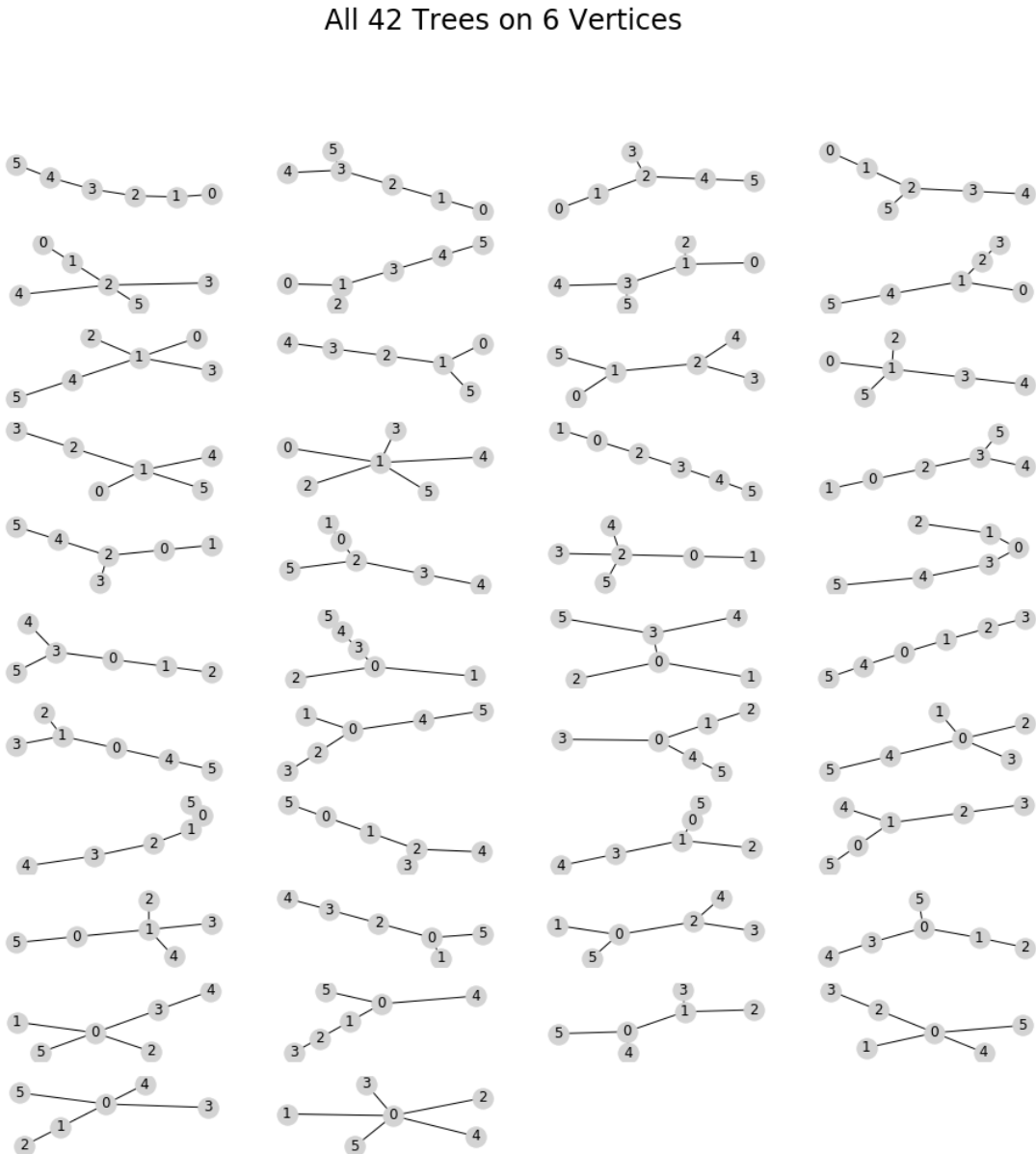
The Catalan Numbers

The n^{th} Catalan number counts:

1. the number of expressions containing n matched parentheses.
2. The number of (ordered) binary trees on $2n + 1$ vertices, $n - 1$ edges, and n leaves.
3. The number of (ordered) trees on $n + 1$ vertices.
4. The number of paths from bottom left to upper right through an $n \times n$ grid that stay below the diagonal.

Stanley's Enumerative Combinatorics, Volume 2, gives at least 60 different interpretations of these integers.

42 Trees on 6 vertices



The Tracy-Widom Distribution

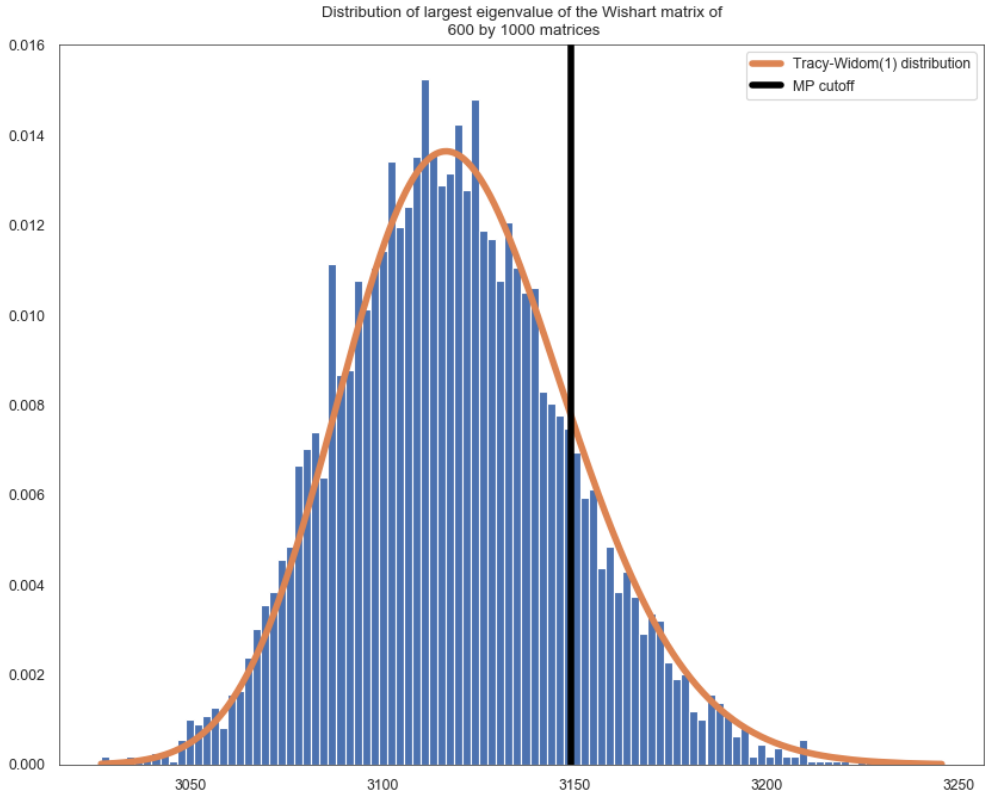
How big are the largest eigenvalues?

The MP distribution drops off at a largest critical eigenvalue λ_+. The largest eigenvalue of a random symmetric matrix is distributed around this critical value. That distribution is called the Tracy-Widom distribution.

Eigenvalues that lie "far out" in the Tracy-Widom distribution are likely attributable to signal.

This was worked out in the late 90's. Computing TW distribution is complicated, arising as the solution of an ODE.

The Tracy-Widom Distribution cont'd



Eigenvector delocalization

Are there preferred directions in the absence of signal?

RMT tells us that, for random matrices, the components of an eigenvector are essentially random. In particular:

- For a "pseudo-cell" eigenvector, the expression values of the different genes are random; there are no preferred directions.
- The expression values of a gene are normally distributed among the pseudo-cell eigenvectors.
- The variances of the genes are distributed by chi-square.

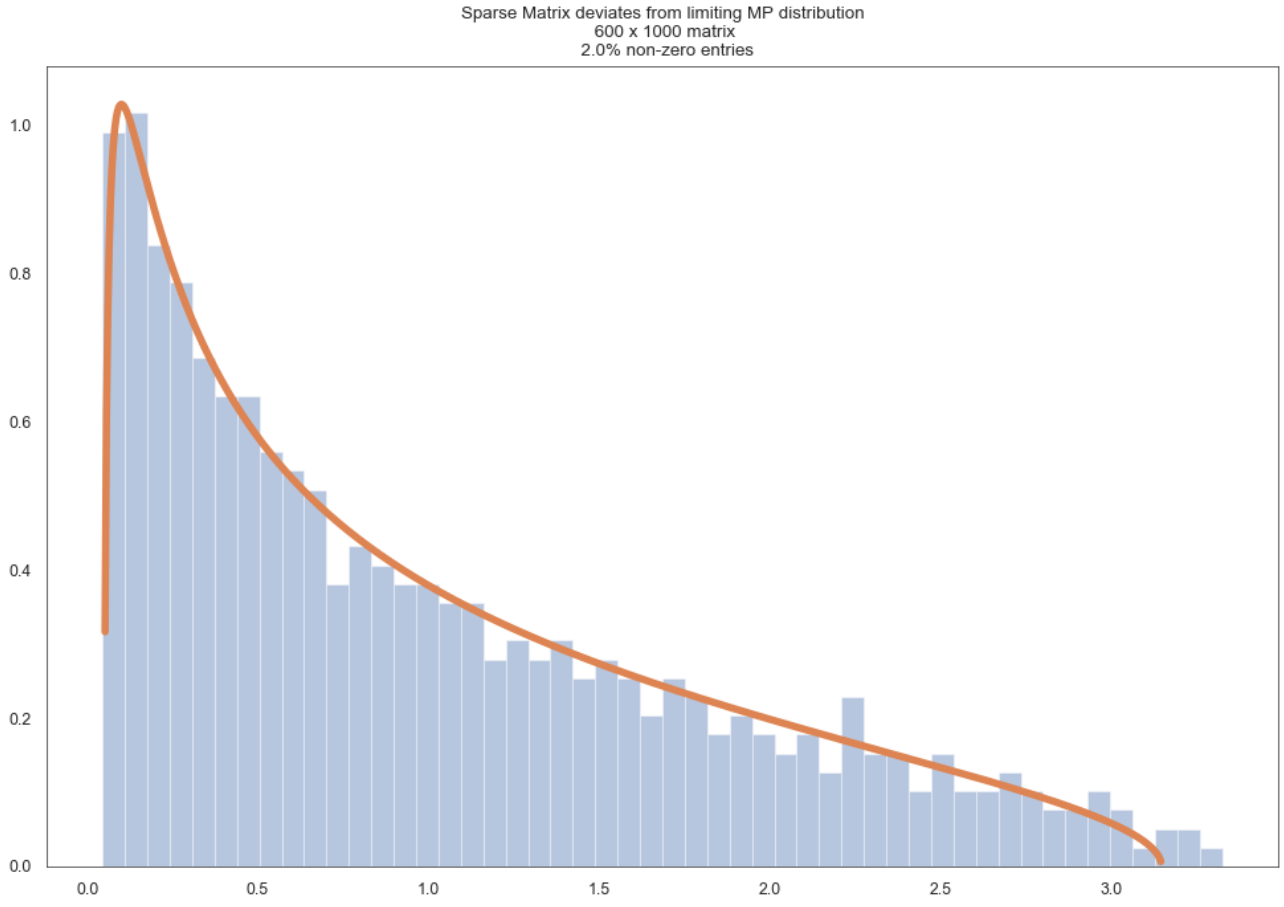
MP and TW in SC RNA

Statistical basis for analysis of covariance matrices – distinguish "signal eigenvectors" from noise.

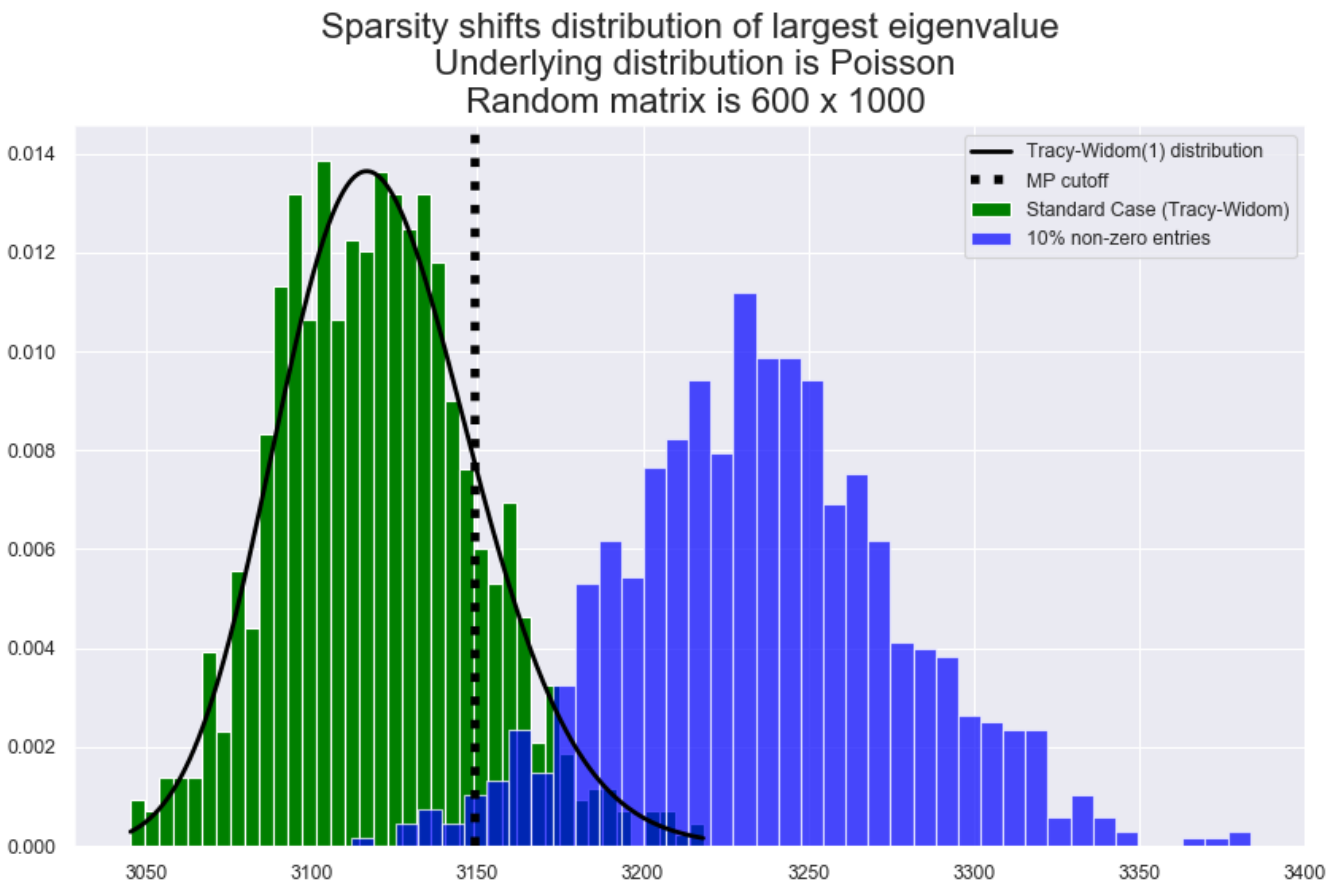
- MP distribution shows what the noise spectrum looks like.
- TW distribution sets limits on what constitutes an unusually large eigenvalue

Catch for SCRNA data: Sparse matrices have different statistics.

Sparsity distorts the MP distribution

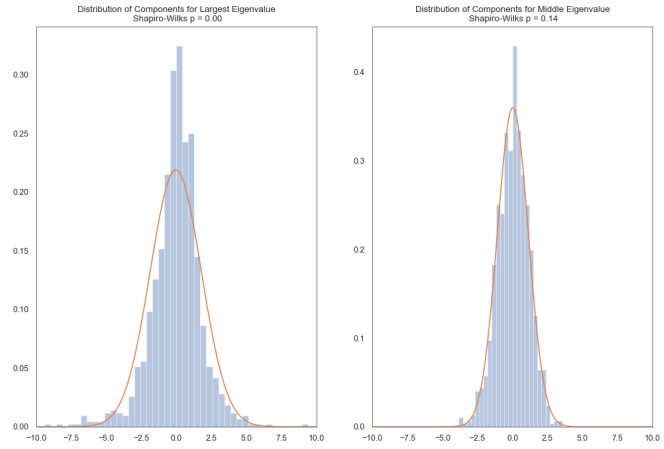


Sparsity shifts distribution of largest eigenvalue



A typical scRNA dataset is very sparse. The 10x example set of 1000 pbmc cells is a matrix that is 94% zeros.

Sparsity causes eigenvectors to make artificial choices of direction



Use randomization to separate sparsity effects from signal

1. Randomizing the expression values for each gene destroys any correlation.
2. The resulting "data" should have the statistics of a random matrix distorted by any signal coming purely from sparsity.
3. Fit a Marchenko-Pastur distribution to the actual data and compare with the randomized data.

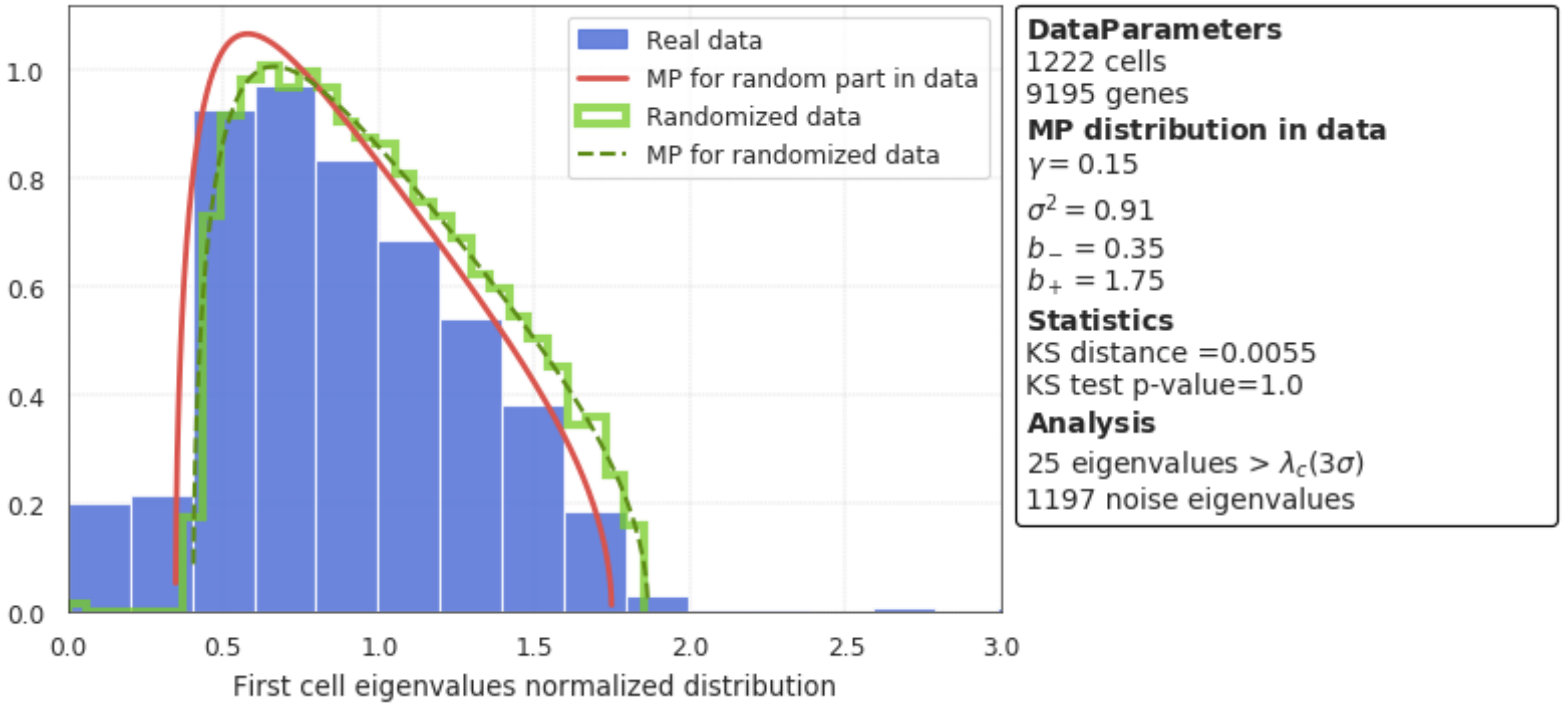
Randomization cont'd

The MP distribution is determined by its mean and variance. The key parameter γ is σ^2/μ^2 for the eigenvalues. The limits of the MP distribution are $(1 \pm \gamma)^2$.

In practice we want to identify boundary points u_- and u_+ so that the set of eigenvalues λ such that

$$u_- \leq \lambda \leq u_+$$

give mean and variance so that $(1 \pm \gamma)^2 \cong u_\pm$.



The green line is the MP distribution for the randomized data.

The red line is the MP distribution fitted to the data.

From the fitted MP distribution, one can estimate the 3σ limit for the TW distribution and identify "signal."

Application: Identifying 'bad' genes

- One consequence of random matrix theory is that the distribution of expression values for a gene in the absence of any signal should be gaussian.
- The algorithm in the paper uses this criterion to exclude "bad genes" from the analysis.
- The algorithm excludes genes whose expression values along the eigenvectors of the randomized matrix fail a normality test. This should exclude genes that are distorted by the sparsity of the matrix.

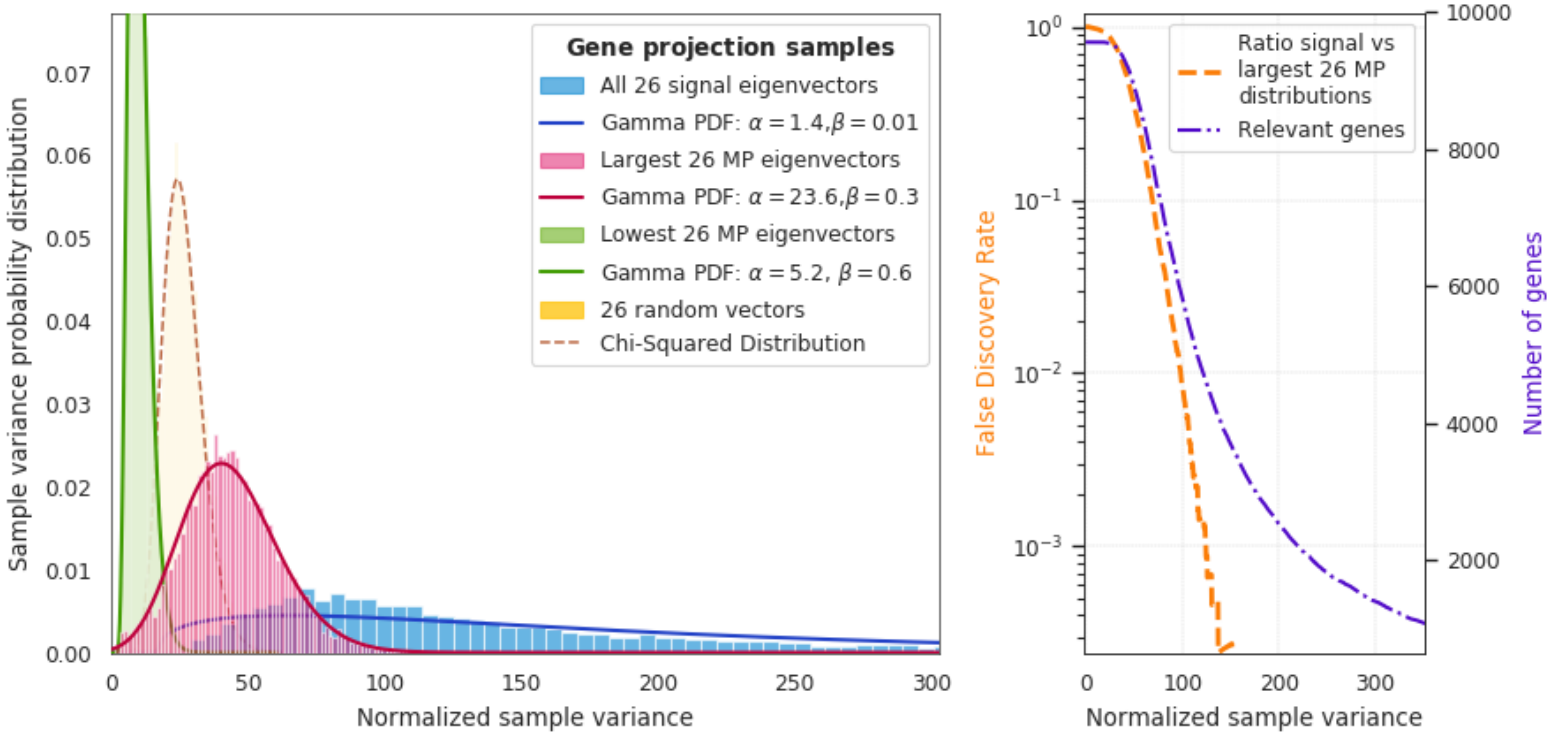
Application: Selecting significant genes

Fitting the MP distribution to the data allows one to identify:

- The eigenvectors above the critical value, corresponding to signal – say there are S of these.
- The S eigenvectors just below the critical value. These are the "least noisy" noise eigenvalues.

Then we can compare the variance of a gene in the signal directions with the variance in the "least noisy" directions.

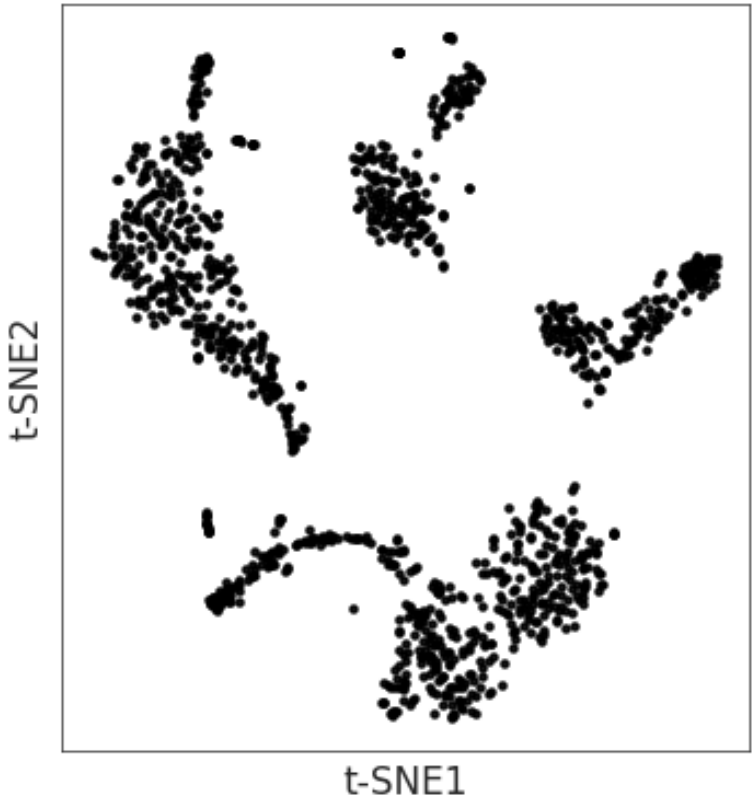
Variance Statistics



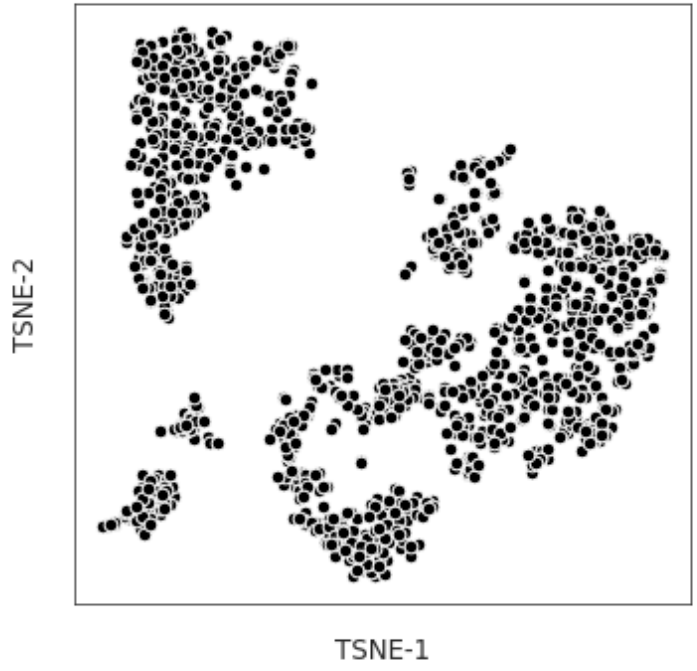
- The left hand graph shows that portion of the variance of a gene that comes from a particular part of the spectrum follows different distributions; with the 'random case' being chi-square.
- The right hand graph compares the variance of the genes in the signal space against the "least noisy" space.

What good is it?

The paper gives a number of examples to argue that selecting genes that have high variance by their "RMT" criterion gives sharper clustering and visualization results. Here is a naive example.



TSNE on RMT selected genes



TSNE on 50 largest eigenspaces

Thanks for listening!

Special thanks to:

Bill Flynn
Joshy George

The Chuang Group:

Jeff Chuang
Javad Noobakhsh
Ada Zhan
Zi-ming Zhao
Scott Adamson
Victor Wang
Patience Mukashyaka