Lessons Learned and New Perspectives From Dean and Provost to aspiring Data Scientist

Jeremy Teitelbaum
October 18, 2018
University of Wisconsin – Madison

Outline

Some personal thoughts on administration: Likes

- Opportunity to be a force for good
- Work with a diverse group of people (intellectual diversity within the unversity; skilled non-academics; more normal gender diversity)
- Challenging (in fact, often intractable) problems
- Teamwork
- Use different skills than in research (interpersonal skills and empathy; communication)

Dislikes

- Less control over your time
- Constant need to manage up and down
- Stress
- Frustration

Thoughts on academia

- Academic careers are long. Some people get stuck and get unhappy, then make everyone around them unhappy. Math particularly impermeable.
- Academic institutions are far too tolerant of bad behavior; people work around it and manage it.
- Lack of diversity and failure to make progress on it taints the enterprise. Deep problems and no progress on them.

Jackson Labs

- Founded in 1929 in Bar Harbor, Maine.
- World center for mouse genetics. Supplier of laboratory mice to the world – 3M per year shipped world wide.
- Mouse genome database contains enormous amount of information on mice and their genetic profiles.
- 26 Nobel Prizes associated with JAX.
- New CT facility has 350 employees focused on precision medicine and cancer.
- About 90 million in NIH funding and 256M per year in mouse business.

JAX, cont'd

- 30 faculty: 10:1 staff to faculty ratio. Huge IT infrastrcture, imaging and sequencing base staffed by Ph.D.'s.
- LOTS of jobs. Very diverse group of postdocs and research scientists including biologists, physicsts, computer scientists and some mathematicians.
- Very different than a university because they are GROWING and HIRING all the time.
- JAX, or places like it, worth thinking about for math Ph.D.'s if they want to do science as an alternative to NSA or Finance jobs outside academia.

1. Diagnosis and the Human Phenotype Ontology

A human mendelian disease is a disorder that is attributable to a mutation in a single gene. Such diseases are inherited according to the classical laws of mendelian genetics. *Sickle cell anemia*, caused by the change of a single nucleotide on chromosome 11, is an example of such a disease.

The website Online Mendelian Inheritance in Man summarizes information about "Mendelian Diseases" It identifies over 5000 conditions for which the cause is a change in a single gene. Of the over 30000 genes in humans, 3500 genes have known mutations that cause disease.

Many of these diseases cause a constellation of abnormalities making them difficult to diagnose.

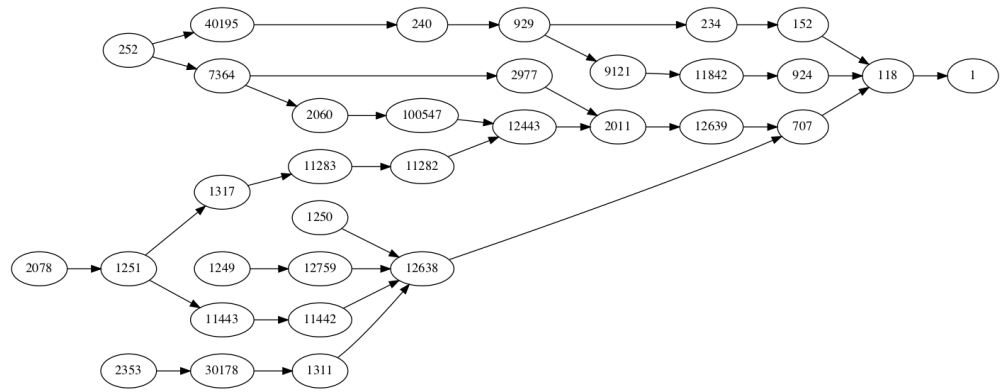
The Human Phenotype Ontology (HPO) is a directed acyclic graph *HPO* nodes are 'symptoms.' More specific symptoms are child nodes of more general ones. The ontology is carefully curated to reflect standard medical terminology.

The HPO was initially constructed by Peter Robinson and his group at Charite hospital in Berlin. Robinson is now at JAX.

A disease d is characterized by a finite set of symptoms (nodes of the HPO) and therefore by its spanning subtree $HPO(d) \subset HPO$.

Angleman's Disease

1	All
40195	Decreased head circumference
2060	Abnormality of the cerebrum
11282	Abnormality of hindbrain morphology
11283	Abnormality of the metencephalon
152	Abnormality of head or neck
12443	Abnormality of brain morphology
924	Abnormality of the skeletal system
2078	Truncal ataxia
1311	Abnormal nervous system electrophysiology
2977	Aplasia/Hypoplasia involving the central nervous system
929	Abnormality of the skull
9121	Abnormal axial skeleton morphology
1317	Abnormality of the cerebellum
2353	EEG abnormality
11442	Abnormality of central motor function
11443	Abnormality of coordination
11842	Abnormality of skeletal morphology
100547	Abnormality of forebrain morphology
7364	Aplasia/Hypoplasia of the cerebrum
707	Abnormality of the nervous system
12759	Neurodevelopmental abnormality
2011	Morphological abnormality of the central nervous system
12638	Abnormality of nervous system physiology
12639	Abnormality of nervous system morphology
1249	Intellectual disability
1250	Seizures
1251	Ataxia
30178	Abnormality of central nervous system electrophysiology
234	Abnormality of the head
240	Abnormality of skull size
118	Phenotypic abnormality
252	Microcephaly



A doctor examining a patient identifies a list of symptoms S, which in turn determine a spanning subtree $HPO(S)$.

Problem

Given a symptom subtree $HPO(S)$, find a set of diseases (subtrees $HPO(D)$), scored by some type of likelihood measure, that are consistent with the symptoms.

Complications:

- Doctor may give more general version of a symptom, rather than the most specific one.
- Doctor may miss some symptoms, or they may be rare even among people with a disease.
- Doctor may add unrelated symptoms.

Semantic Similarity

In 1995 Resnick proposed a similarity measure for ontologies based on the "information content" of a node.

- Given a node (symptom) m, let $\mathcal{A}(m)$ be the set of diseases that are associated with that node – meaning some descendant of that node is a specific symptom of the disease.
- Assign weights to the edges of the DAG by setting the weight of $n \to m$ to be $\log |\mathcal{A}(m)| - \log |\mathcal{A}(n)|$.
- The information content $IC(n)$ of a node n is the sum of the weights of the edges from the node to the root (along any path).

The Resnick similarity between two sub-DAG's X and Y is $(H(X,Y)+H(Y,X))/2$ where

$$H(X,Y)=\frac{1}{|X|}\sum_{x\in X}\max_{y\in Y} IC(z(x,y))$$

where z is the common ancestor of x and y with the greatest IC.

A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology, in *BMC Bioinformatics 2018*, lists six other 'distance measures' between subtrees of a DAG, and adds a new one. There are many others.

More sophisticated models under development:

1. incorporate information about the relative frequencies of symptoms in patients with the disease
2. attempt to incorporate information about genetic variants into the diagnosis.

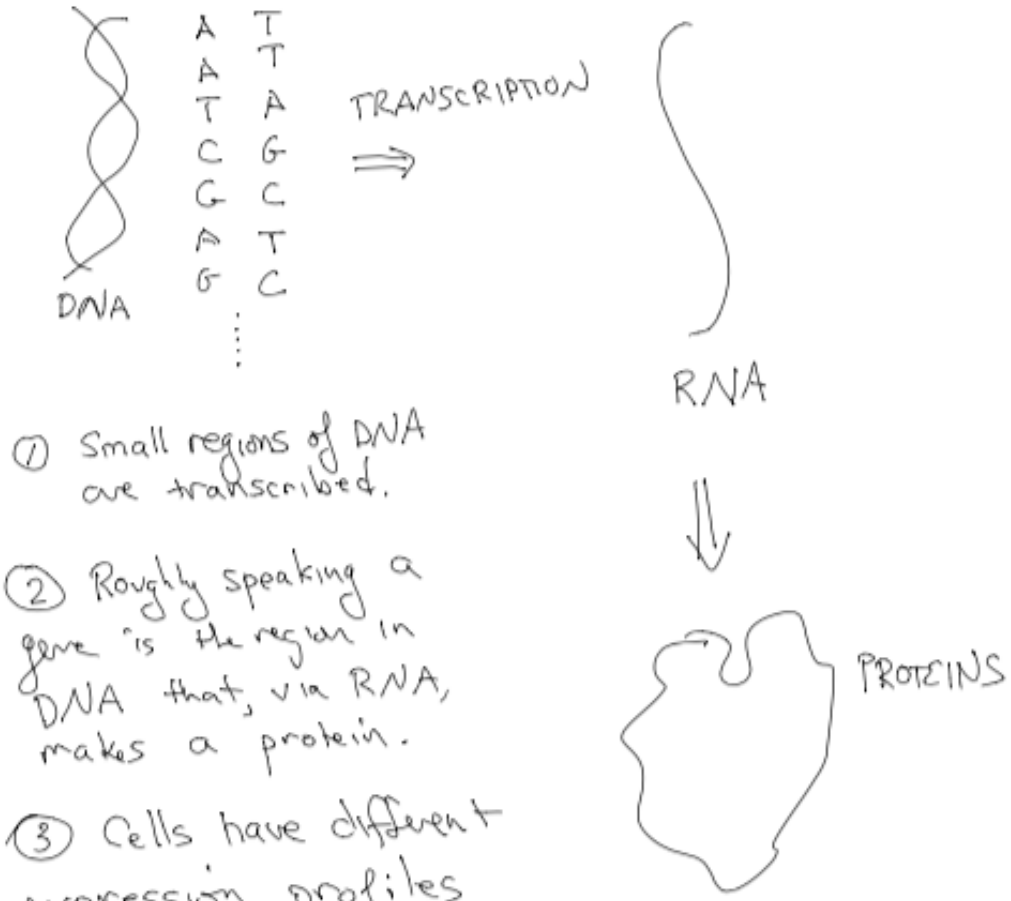
Common validation technique is to test against simulated data.

Problem

Formally characterize the differences, strengths, weaknesses of these different approaches.

2. Analysis of single cell RNA experiment data

Two seconds on human molecular genetics.



RNA-seq

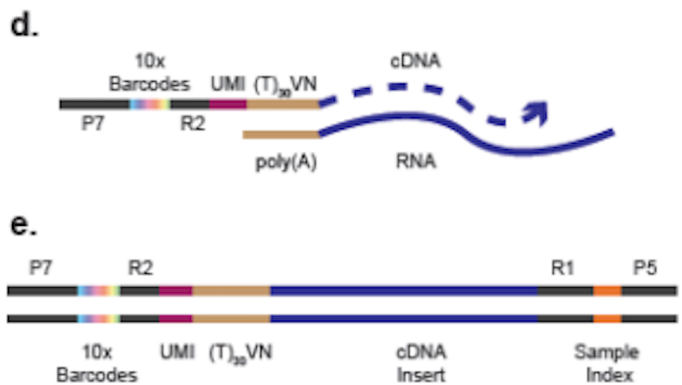
- In 'Bulk RNA sequencing' (RNA-seq) researchers take a sample of a particular tissue type, sequence the RNA in the sample by cutting it up into pieces and aligning the pieces against a reference genome.
- They count the fragments of RNA that line up against a particular gene and use that as a measure of the relative activation level of that particular gene AVERAGED OVER THE CELLS IN THE SAMPLE.
- Typical experiments:
  - Take cells before and after treatment with a drug and look to see if the expression profile has changed.
  - Compare expression profile of normal and cancer cells and look for genes that are amplified or repressed in cancer; these might suggest biochemical pathways for drug targeting.

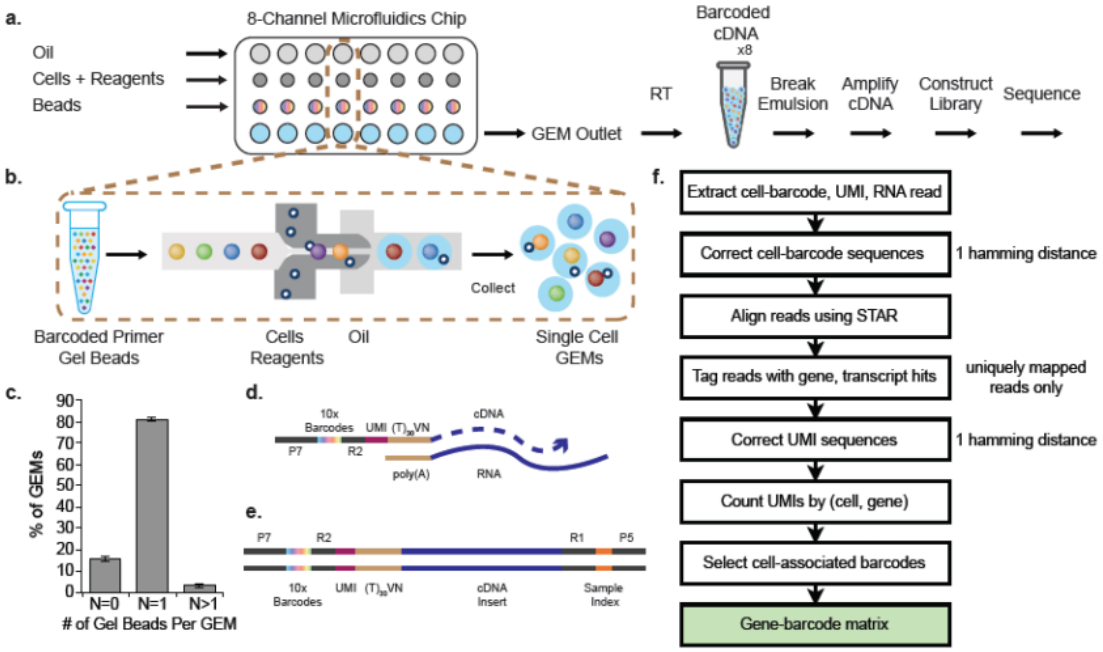
- Typical experiment might have 3 or 4 controls, 3 or 4 treatment cases, and 30000 genes. So for Bulk RNA sequencing the problem is to identify which of those 30000 genes differ between the controls and treatment cases.
- Although this technology is perhaps only 10 years old, it is now standard and there are established techniques for this.

Single Cell RNA-Seq

In single cell RNA-seq, *individual cells* are captured to obtain a cell by cell expression profile.

- Cells are caught in droplets and RNA in each droplet gets a unique sequence identifying the droplet attached; plus each RNA molecule gets a unique sequence identifying the molecule.
- Then these pieces are amplified (replicated many times). The resulting 'library' gets sequenced.
- In this way one can count the number of RNA molecules from each gene in each cell.
- Some nice, elementary use of error correcting codes takes place here.'





10x Genomics droplet high-throughput single cell platform

Single Cell Data

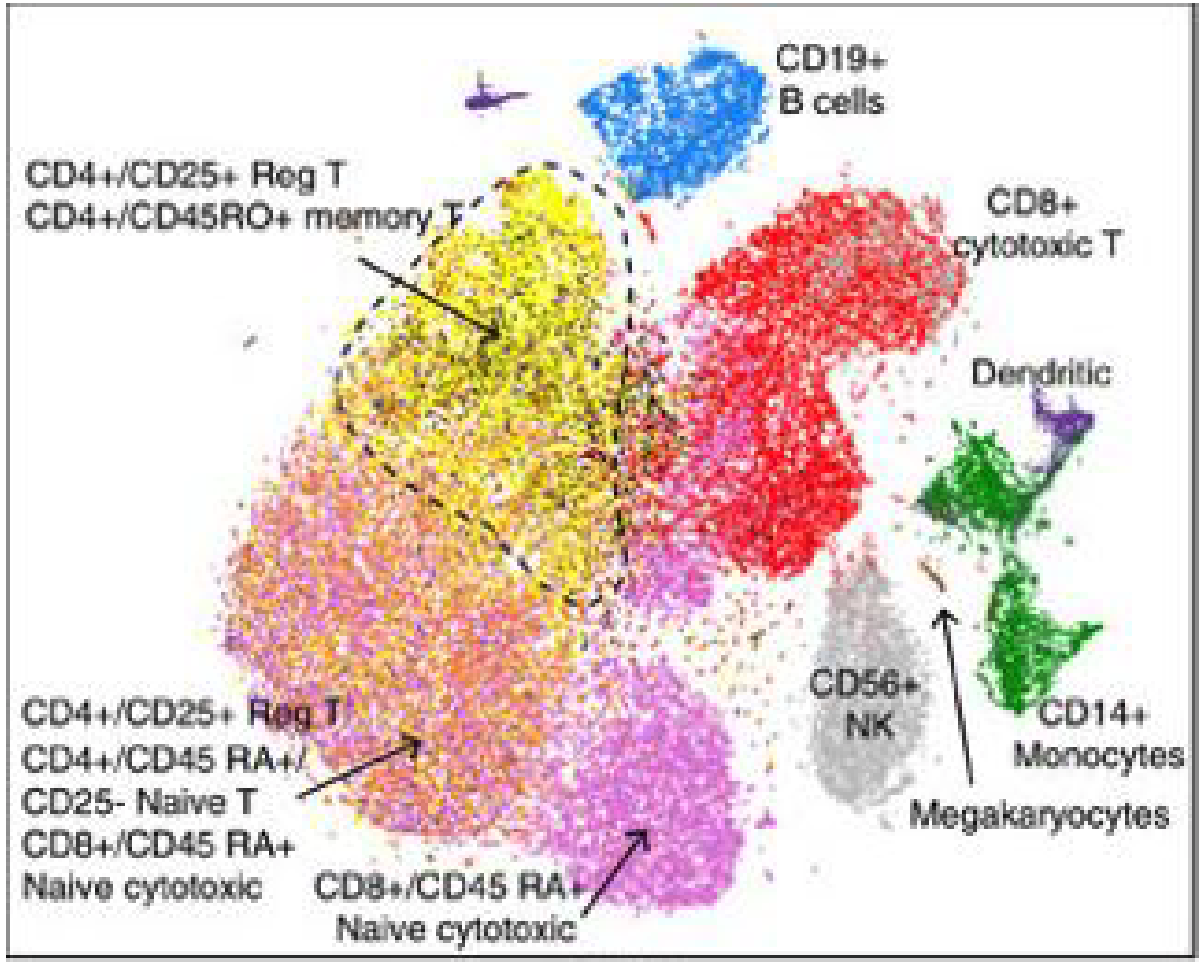
The output from a single cell experiment is an $N \times K$ sparse integer matrix. Here N, the number of cells, could range from a few hundred to a few million, and K, the number of genes, is on the order of 30000. The (i, j)-th entry of the matrix is the activation level of gene j in cell i.

Goals for these experiments are:

- Obtain fine structure among classes of cells based on their expression profiles.
- Understand developmental history and reconstruct development.
- Understand tumor heterogeneity and its relationship to chemotherapy response (*)
- Many other things....

Outline of Analysis

1. Clean up the data by throwing out cells with few detected genes and genes that hardly ever show up.
2. Normalize the data by cell to account for the different levels of sequencing success.
3. Do a first round of dimensionality reduction by identifying genes that show a lot of variance
4. Use a second round of dimensionality reduction to two or three dimensions (PCA, tSNE, or UMAP)
5. Cluster the results
6. Look at the gene profiles of each cluster to try to find genetic markers for each cluster.
7. Figure out the biological implications.



Immune cell types in blood from *Massively parallel digital transcriptional profiling of single cells* by Zheng, et. al., Nature Comm. 8, 2017.

Challenges with scRNA data

Sparsity

There is considerable activity in the literature over some basic questions about how to analyze single cell data. Much of this comes from the sparsity of the data.

1. How to distinguish between genes that have low (or zero) expression and genes that show up with zero expression for technical reasons?
2. How to properly normalize the data by cell when there is so much variation in the total count per cell?
3. How to identify subpopulations within the data when the subpopulations could be distinguished by differential expression of only a very small number of the 30000 genes being profiled?

More questions:

1. What is the proper statistical model for this data? I count more than 10 proposals that have been published in different studies.
2. How sensitive are the common dimensionality reduction algorithms used for clustering to the different choices one makes for normalization and to the possible dropout of some values?
3. How to integrate this data with 'bulk' sequencing data or other types of data.

Very recent: random matrix theory applied to single cell data

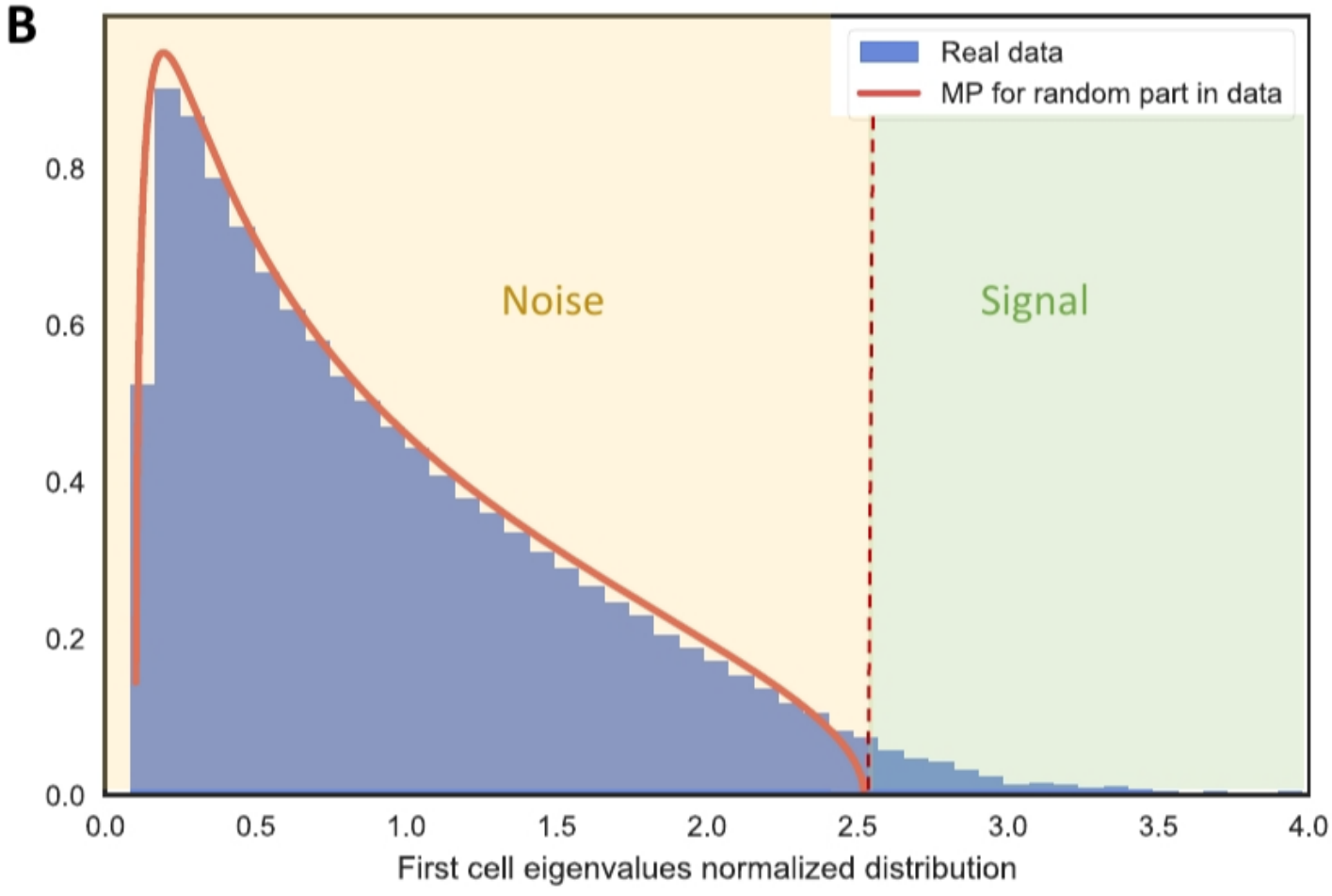
The paper *Quasi-universality in single-cell sequencing data*, by Aparicio, Bordyuh, Blumberg, and Rabadan, looks at the spectra of matrices arising from single cell experiments through the lens of random matrix theory.

One formulates a 'null hypothesis' that the matrix of data is random, and and can identify a signal by identifying the the extent that the eigenvalue distribution of the correlation matrix differs from the universal distribution that one expect.

The sparsity of the data matrices is an obstacle to this and needs to be accounted for.

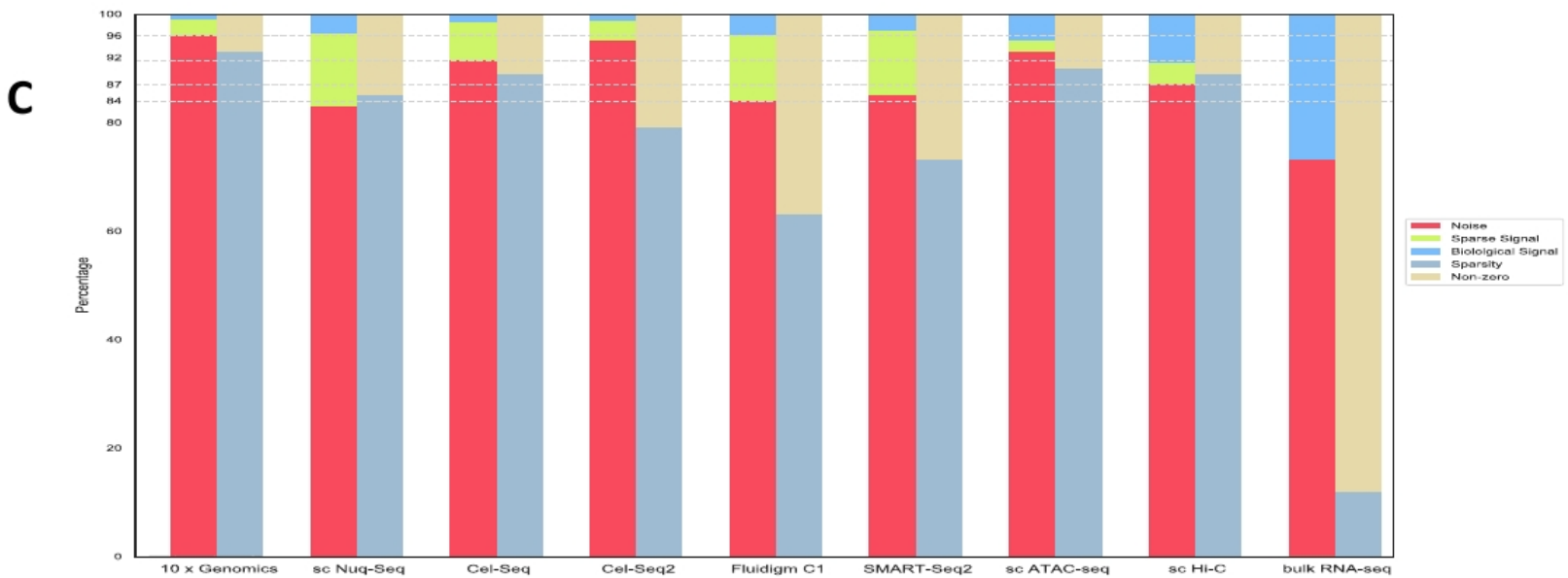
More on random matrices and single cell data

From the paper cited above.



Structure of single cell data from random matrix perspective

From the paper cited above.



Some thoughts about graduate education in math

Goal is to train successful research mathematicians, but students need to protect their economic mobility as a hedge against exploitation. Not clear what long-term future of academic jobs will be.

- (Not a radical idea) Insist on knowledge of programming.
- Include some statistics in the core curriculum for mathematicians.
- Find a way to include group projects in the curriculum.
- Math in general is pretty good about time-to-degree but in academia in general students have to get through fast.