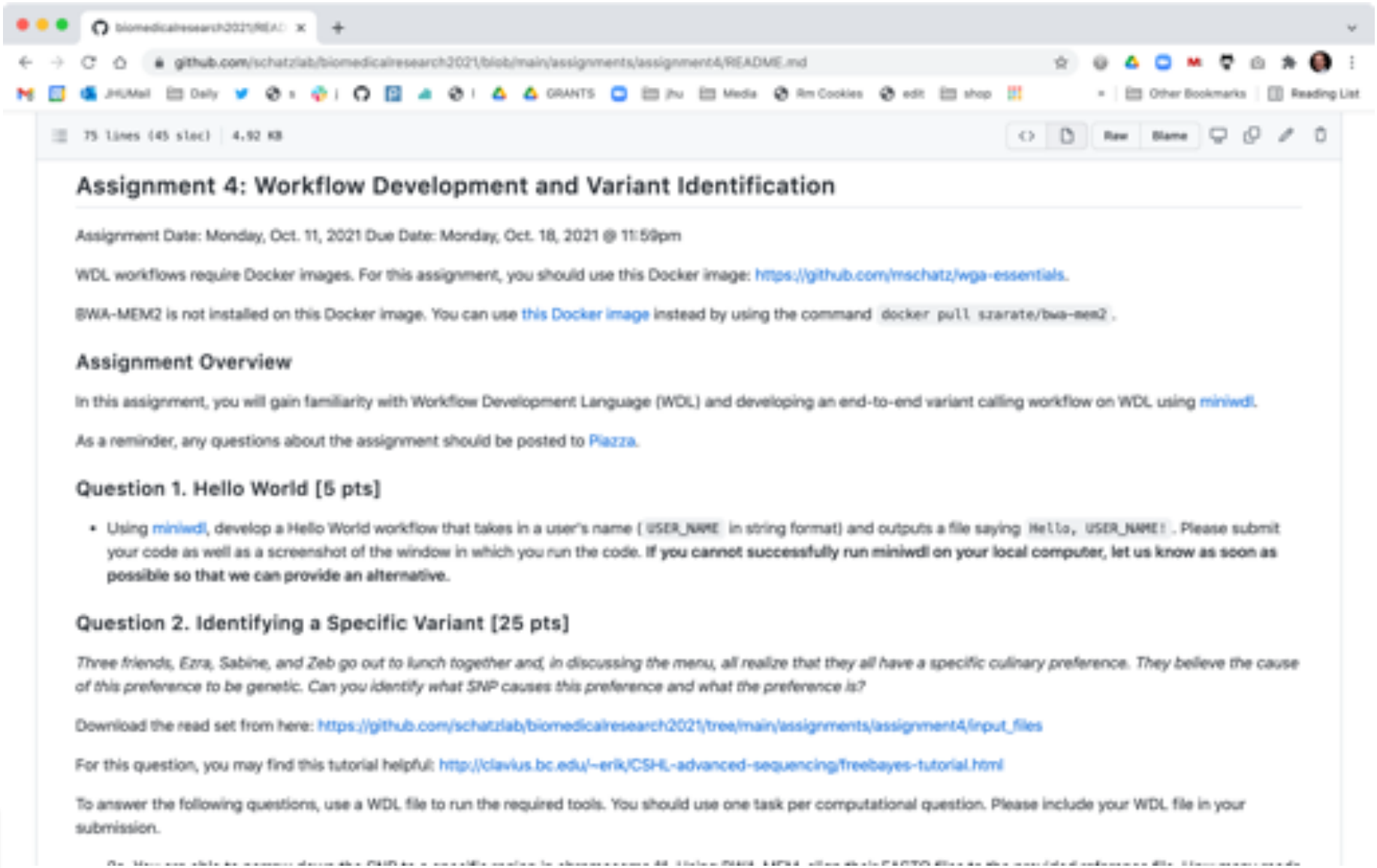
Lecture 16. Projects + scRNAseq

Michael Schatz

October 25, 2021

Advanced Biomedical Research

Assignment 4: WDLs
Due Oct 18 @ 11:59pm



Assignment 4: Workflow Development and Variant Identification

Assignment Date: Monday, Oct. 11, 2021 Due Date: Monday, Oct. 18, 2021 @ 11:59pm

WDL workflows require Docker images. For this assignment, you should use this Docker image: https://github.com/mschatz/wga-essentials.

BWA-MEM2 is not installed on this Docker image. You can use this Docker image instead by using the command `docker pull szarate/bwa-mem2`.

Assignment Overview

In this assignment, you will gain familiarity with Workflow Development Language (WDL) and developing an end-to-end variant calling workflow on WDL using miniwdl.

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. Hello World [5 pts]

- Using miniwdl, develop a Hello World workflow that takes in a user's name (`USER_NAME` in string format) and outputs a file saying `Hello, USER_NAME!`. Please submit your code as well as a screenshot of the window in which you run the code. **If you cannot successfully run miniwdl on your local computer, let us know as soon as possible so that we can provide an alternative.**

Question 2. Identifying a Specific Variant [25 pts]

Three friends, Ezra, Sabine, and Zeb go out to lunch together and, in discussing the menu, all realize that they all have a specific culinary preference. They believe the cause of this preference to be genetic. Can you identify what SNP causes this preference and what the preference is?

Download the read set from here: https://github.com/schatzlab/biomedicalresearch2021/tree/main/assignments/assignment4/input_files

For this question, you may find this tutorial helpful: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

To answer the following questions, use a WDL file to run the required tools. You should use one task per computational question. Please include your WDL file in your submission.

https://github.com/schatzlab/biomedicalresearch2021

Updated schedule!

16.	Mo 10/25	Functional Analysis 5: Single Cell Genomics	* Ginkgo: interactive analysis and assessment of single-cell copy-number variations (Garvin et al, 2015, Nature Methods) * The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells (Thapnell et al, Nature Biotech, 2014) * Eleven grand challenges in single-cell data science (Lähnemann et al, Genome Biology, 2020)	Assignment 5 & Preliminary Project Report
17.	We 10/27	Gene Regulation		
18.	Mo 11/1	Midterm Review		
19.	We 11/3	Midterm Exam		Take home exam
20.	Mo 11/8	Human Evolution	* An integrated map of genetic variation from 1,092 human genomes (1000 Genomes Consortium, 2012, Nature) * Analysis of protein-coding genetic variation in 60,706 humans (Lek et al, 2016, Nature) * A Draft Sequence of the Neandertal Genome (Green et al. 2010, Science) * Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals (Vernot et al. 2016, Science)	Project Interim Report
21.	We 11/10	Human Genetic Diseases	* Genome-Wide Association Studies (Bush & Moore, 2012, PLOS Comp Bio) * The contribution of de novo coding mutations to autism spectrum disorder (Iossifov et al, 2014, Nature)	
22.	Mo 11/15	Cancer Genomics	* The Hallmarks of Cancer (Hanahan & Weinberg, 2000, Cell) * Evolution of Cancer Genomes (Yates & Campbell, 2012, Nature Reviews Genetics) * Comprehensive molecular portraits of human breast tumours (TCGA, 2012, Nature)	Project Presentations Scheduling
23.	We 11/17	Microbiome and Metagenomics	* Kraken: ultrafast metagenomic sequence classification using exact alignments (Wood and Salzberg, 2014, Genome Biology) * Chapter 12: Human Microbiome Analysis (Morgan and Huttenhower)	Project Report Assignment
	Mo 11/22	Thanksgiving Break		
	We 11/24	Thanksgiving Break		
24.	Mo 11/29	Project Presentations		
25.	We 12/1	Project Presentations		
26.	Mo 12/6	Project Presentations	Last Day of class	
	Mo 12/20	Final Project Report Due!		

https://github.com/schatzlab/biomedicalresearch2021

Assignment 5: RNAseq
Due Nov 1 @ 11:59pm

Assignment 5: RNA-seq

Assignment Date: Monday, Oct. 25, 2021
Due Date: Monday, Nov. 1, 2021 @ 11:59pm

Assignment Overview

In this assignment, you will explore a couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, post any questions about this assignment to Piazza.

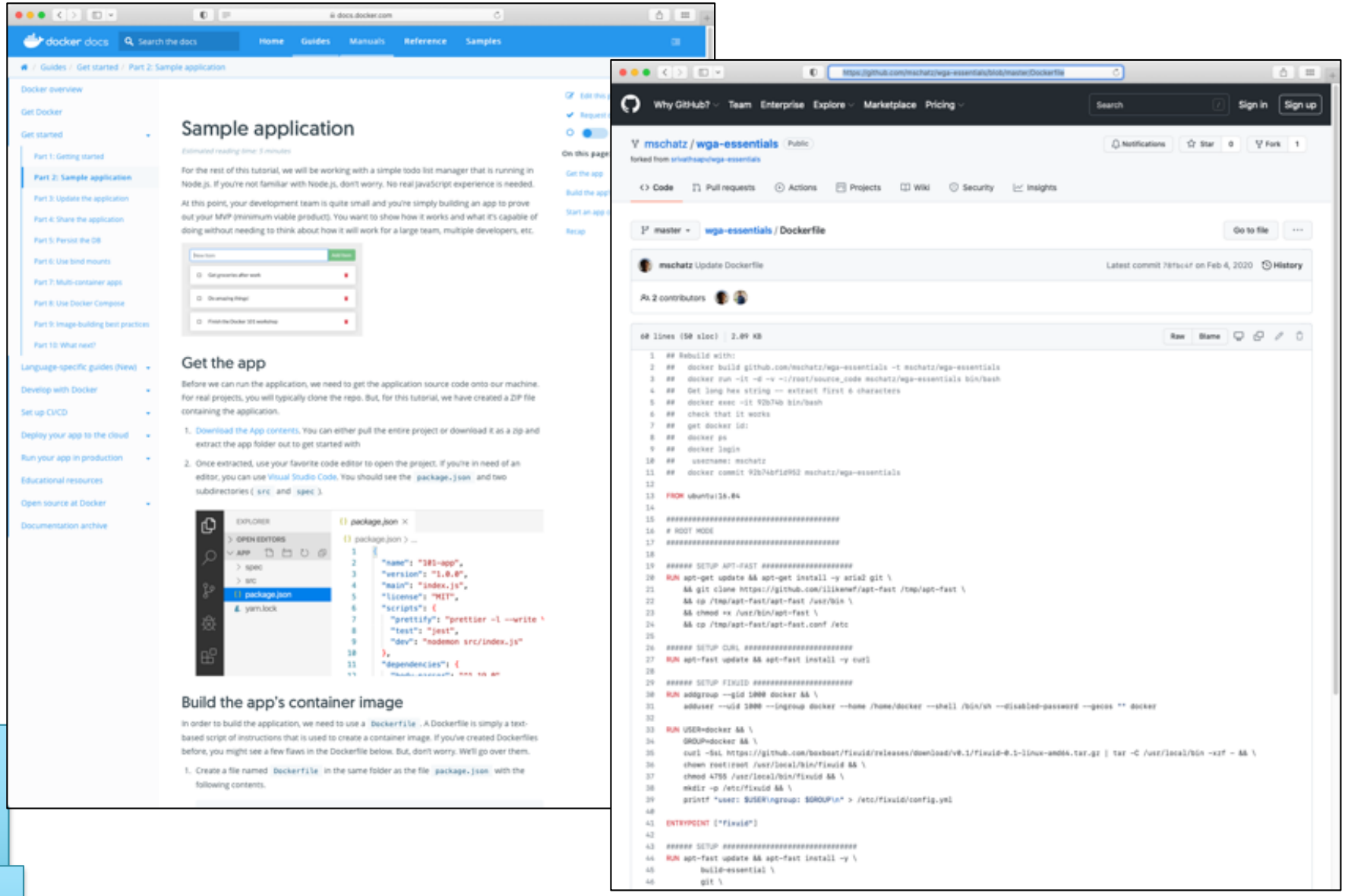
Question 1. Time Series (20 pts)

This file contains normalized expression values for 100 genes over 10 time points. Most genes have a stable background expression level, but some special genes show increased expression over the time course and some show decreased expression.

- Question 1a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical clustering are common clustering algorithms you could try.]
- Question 1b. Calculate the first two principal components of the expression matrix. Show the plot and color the points based on their cluster from part (a). Does the PC1 axis, PC2 axis, neither, or both correspond to the clustering?
- Question 1c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.
- Question 1d. Visualize the expression data using t-SNE.
- Question 1e. Using the same data, visualize the expression data using UMAP.
- Question 1f. In a few sentences, compare and contrast the (1) heatmap, (2) PCA, (3) t-SNE, and (4) UMAP results. Be sure to comment on understandability, relative positioning of clusters, runtime, and any other significant factors that you see.

https://github.com/schatzlab/biomedicalresearch2021

Docker containers



https://docs.docker.com/get-started/02_our_app/
https://github.com/mschatz/wga-essentials/blob/master/Dockerfile

Class Project!
Proposal Due Nov 1

Project Proposal

Assignment Date: Monday Oct 25, 2021
Due Date: Monday, November 1 2021 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project of no more than 3 people.

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc) or AnVIL account (WDL-based, T2T)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

https://github.com/schatzlab/biomedicalresearch2021/blob/main/project/proposal.md

Project Ideas

Here are a few selected projects organized by theme, although you are also free to present your own project ideas. A good project is one that:

1. Has a well defined goal
2. Has a well defined method proposed for solving it; and
3. Has appropriate data and resources available.

If any of these are not available, your project will not be successful, especially in the limited time remaining for class. A successful model for the class project is to apply a technique developed in a different context to the biological problem of your choice. Another successful model is to identify an important paper of interest and then try to improve apon their method in some way (faster, less memory, more sensitive, more precise, novel species, novel data sets, etc).

T2T-based analysis: T2T-main and T2T-variants

1. Short-read structural variant analysis: Use Parliament2 to call SVs in some of the 1000 genomes samples. Compare the result using GRCh38 to CHM13 and compare short and long read SV calls.
2. Evaluate DeepVariant with CHM13 using short read data from a few samples. Compare to both GRCh38 and CHM13, and compare DeepVariant to GATK results. Investigate the regions that show the largest differences
3. Evaluate exome sequencing with the CHM13 genome by aligning several samples to both GRCh38 and CHM13. Include a few samples that also have whole genome data available.
4. RNA-seq analysis: Use StringTie2 to analyze short read RNAseq data. Align the reads to both GRCh38 and CHM13 and investigate differences. RNAseq reads are available from (Lappalainen et al, 2013, Nature). Stretch goal is to investigate eQTLs
5. Functional genomics analysis: Apply Basenji to both GRCh38 and CHM13 and compare results. If needed, focus on one chromosome

CS Theory and Systems

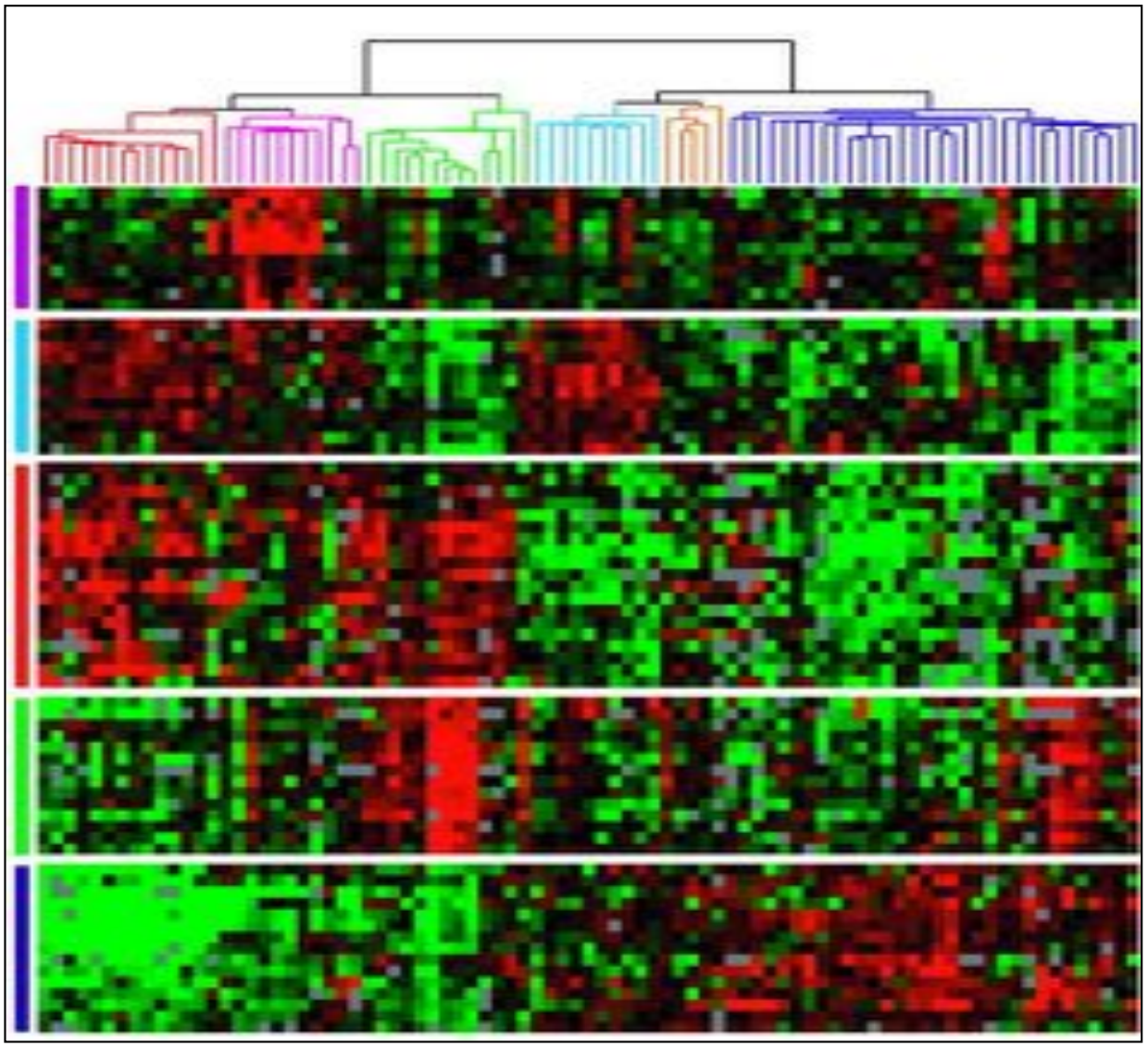
1. Adapt one or more learned data structures for genomics data: Learned Data Structures
2. Accelerate an important genomics pipeline using GPUs or cloud computing and use that to study a larger dataset Rail-RNA
3. Implement a genomics processing pipeline using WebAssembly and/or Objective-C/Swift/Android fastq.bio
4. Develop a novel visualization of genomics data (especially from 23-and-me reports or single cell data): Circos
5. Apply deep learning techniques to a problem in genomics Primer
6. Develop a novel fastq/BAM compression scheme for long term storage (which may require a large precomputed dictionary and/or extensive compute)

Or your own idea!

This should be more than you are already doing, but can be a novel twist to a dataset/idea you are already using. If you have a research idea but not the right data, let me know and I'll help you find some.

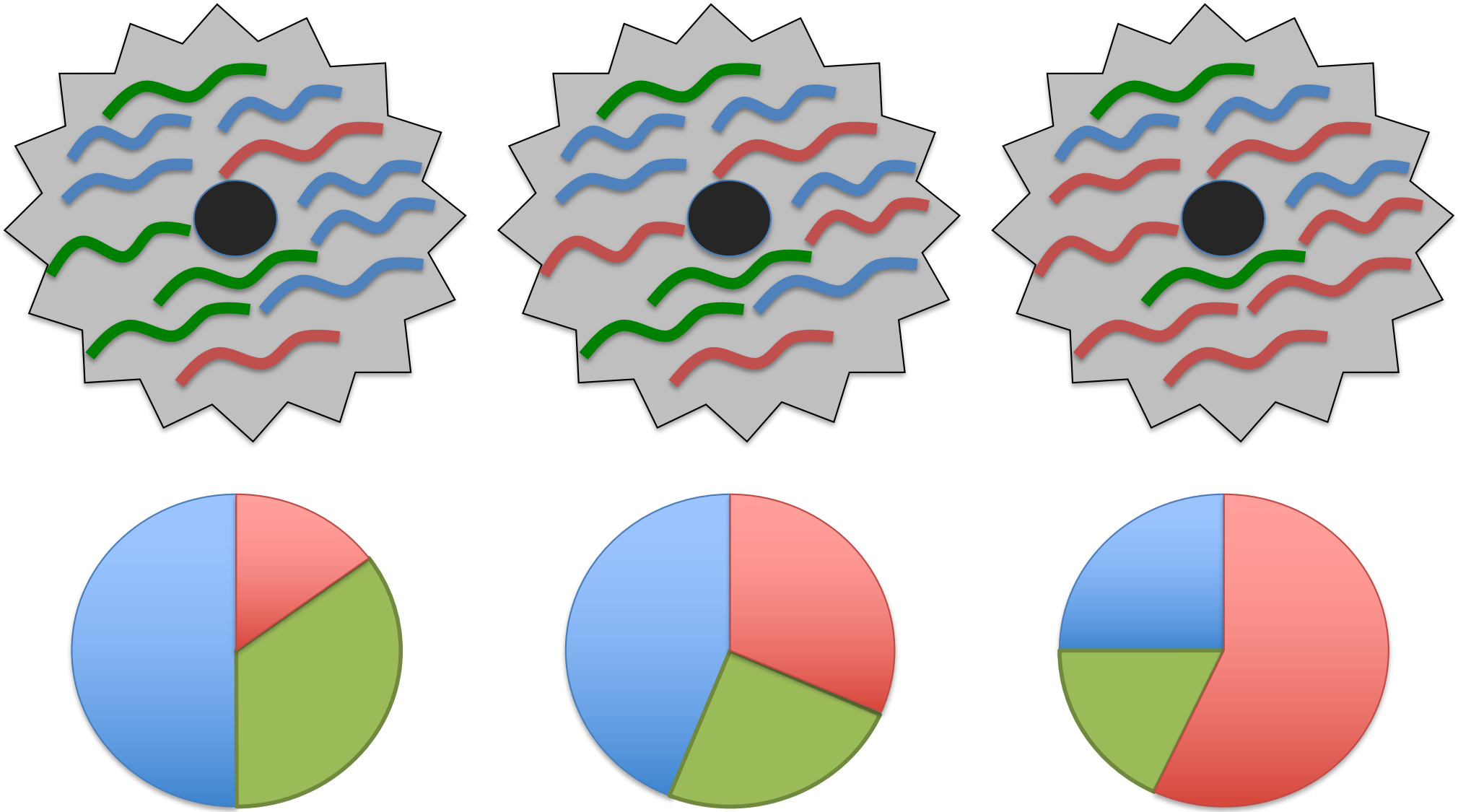
https://github.com/schatzlab/biomedicalresearch2021/blob/main/project/ideas.md

RNA-seq

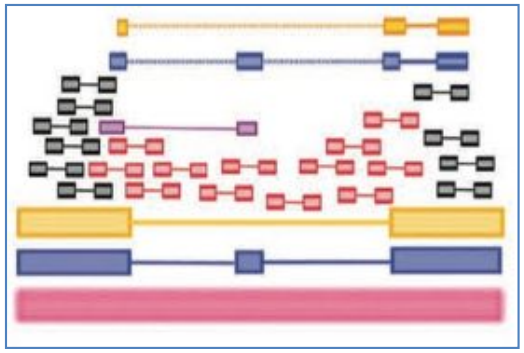


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

RNA-seq Overview



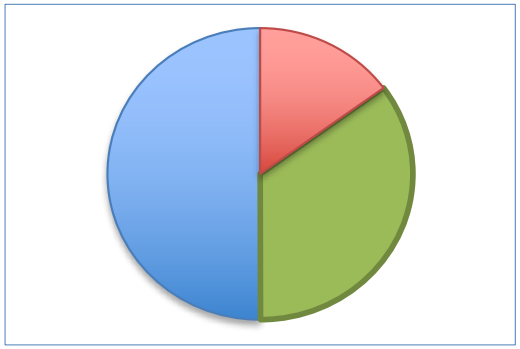
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

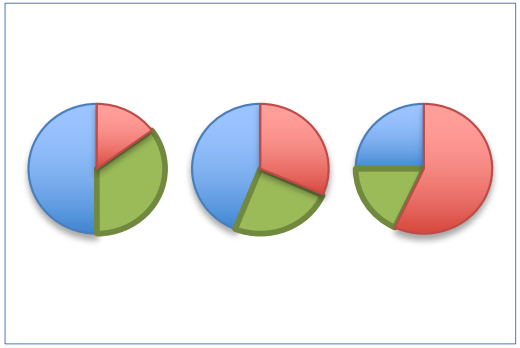
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

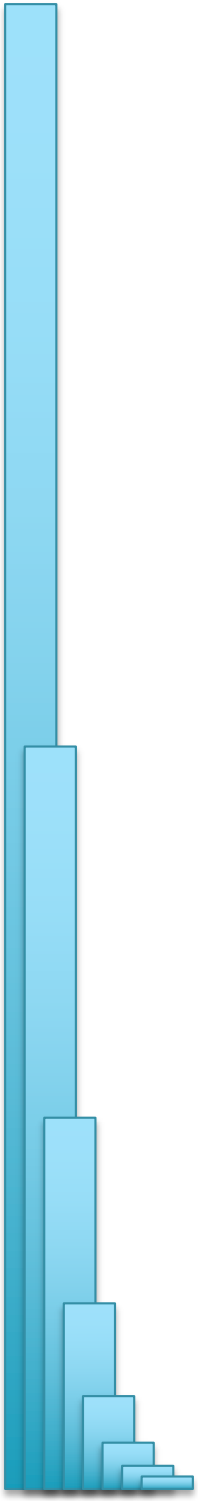
Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics.* doi:10.1093/bioinformatics/btt688

Annotation Summary

- Three major approaches to annotate a genome
  1. Alignment:
     - Does this sequence align to any other sequences of known function?
     - Great for projecting knowledge from one species to another
  2. Prediction:
     - Does this sequence statistically resemble other known sequences?
     - Potentially most flexible but dependent on good training data
  3. Experimental:
     - Lets test to see if it is transcribed/methylated/bound/etc
     - Strongest but expensive and context dependent

- Many great resources available
  - Learn to love the literature and the databases
  - Standard formats let you rapidly query and cross reference
  - Google is your number one resource ☺



Unsupervised Learning
aka Clustering

Clustering Refresher

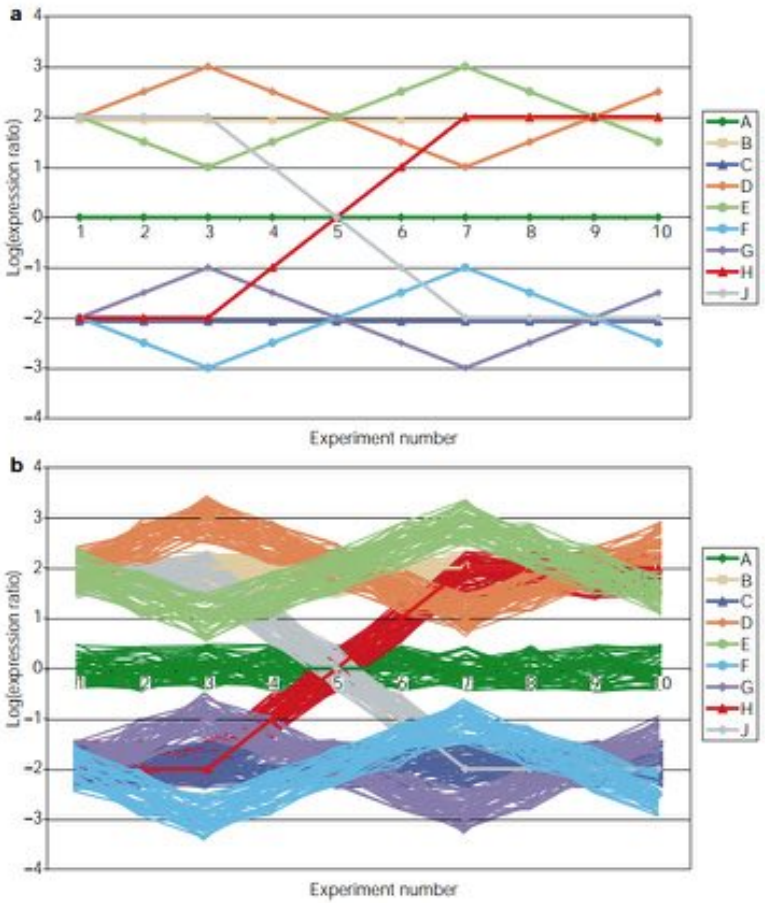
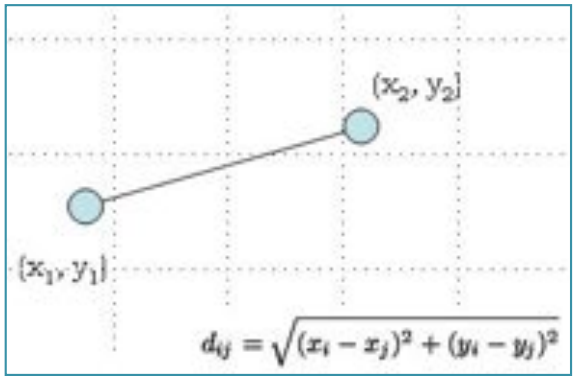
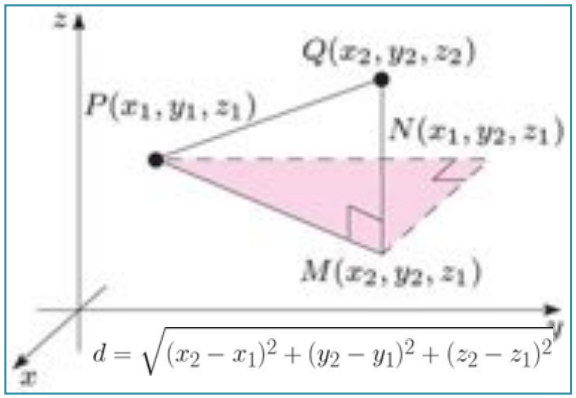


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments. **b** | For each expression pattern, 50 additional genes were generated, representing variations on the basic patterns.

Euclidean Distance

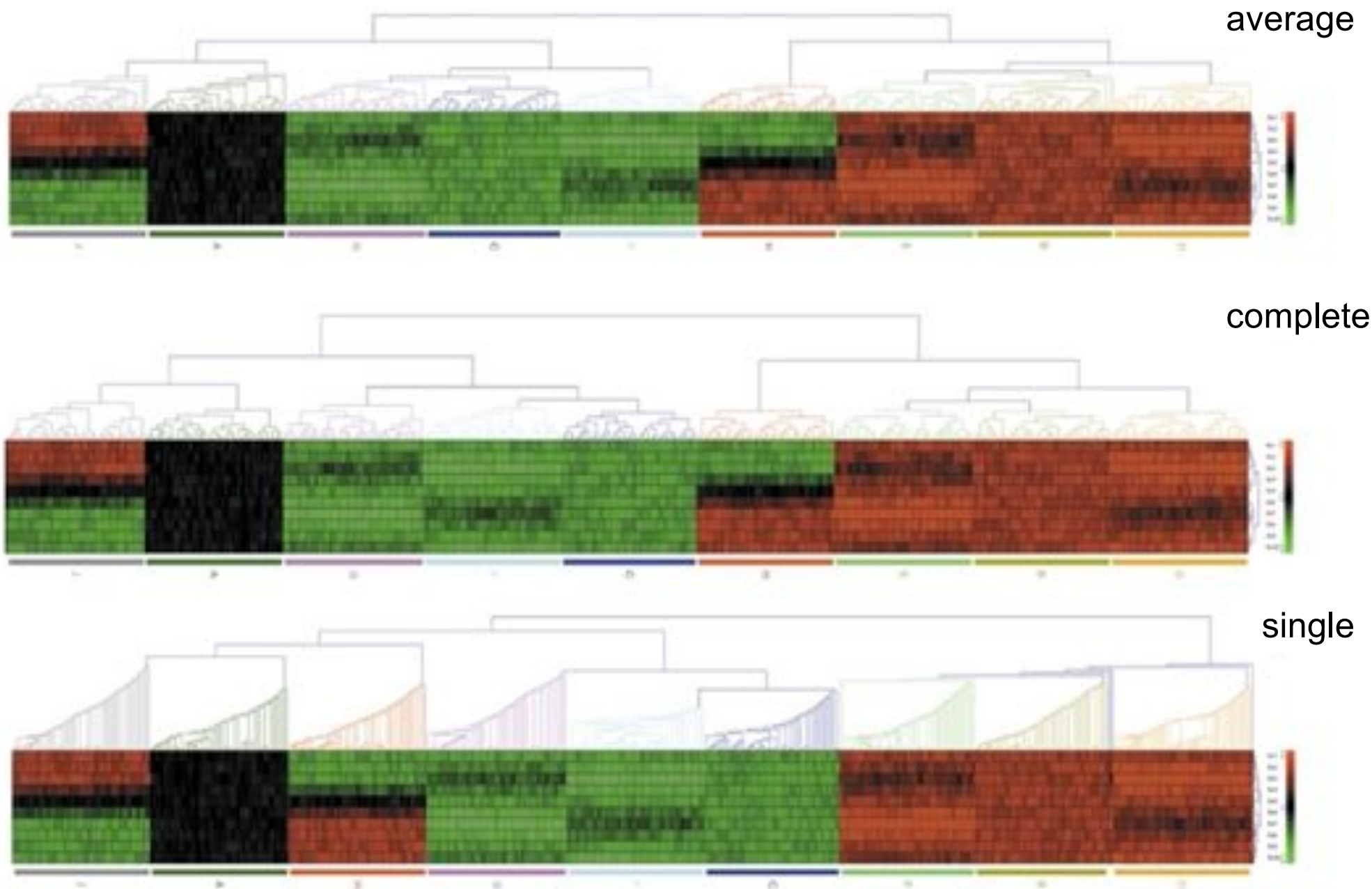




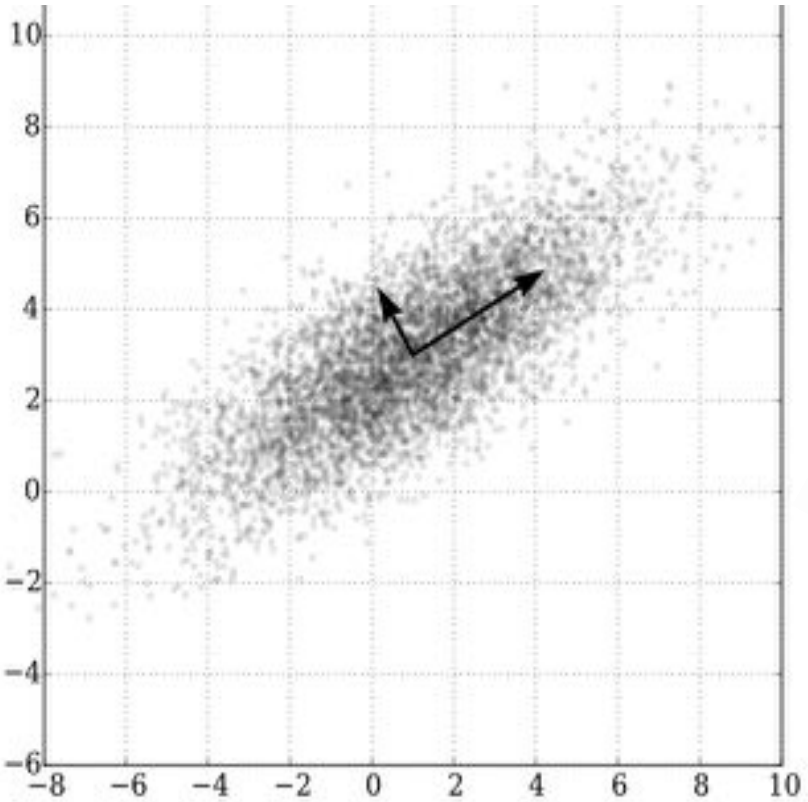
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

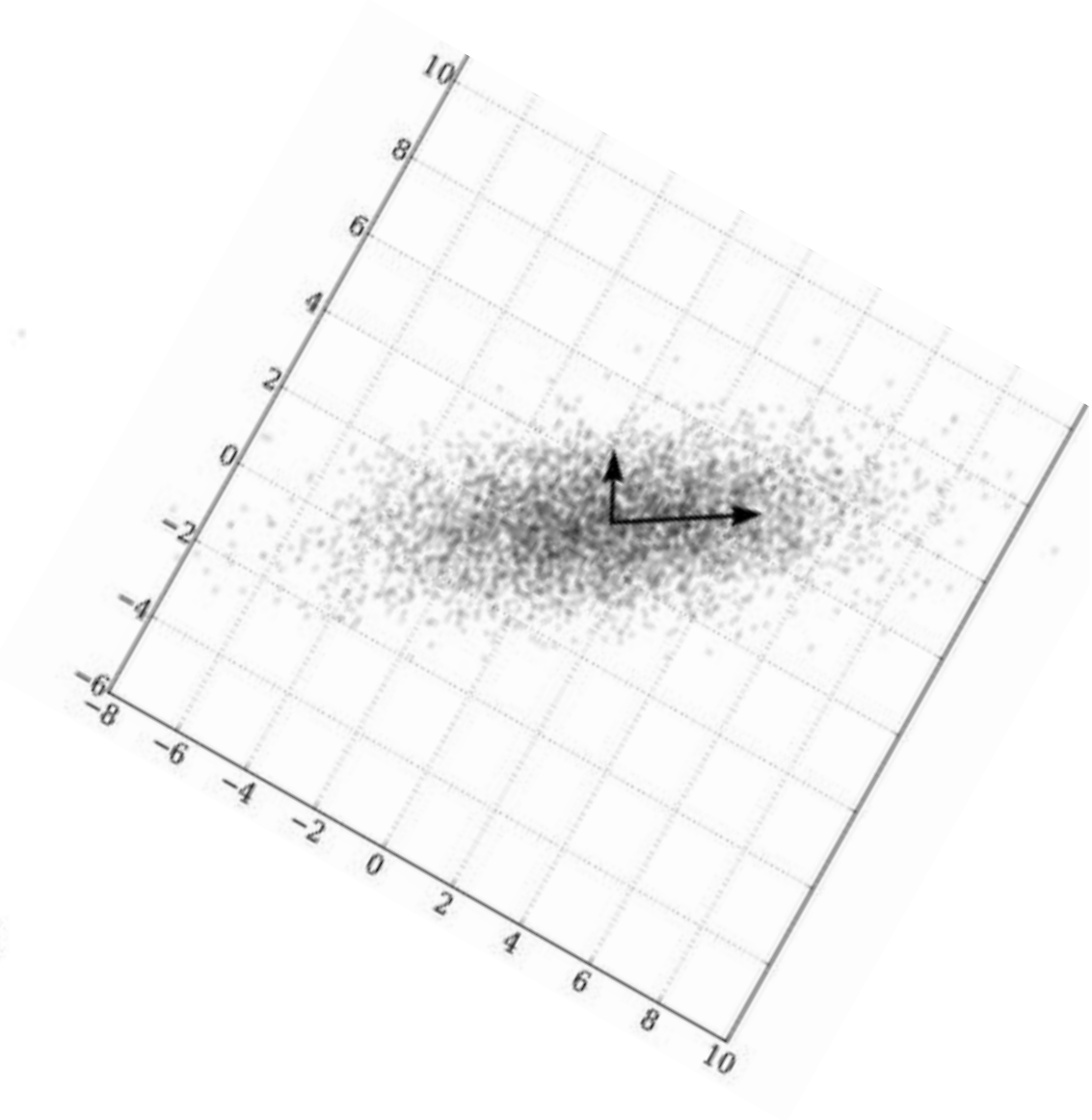
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



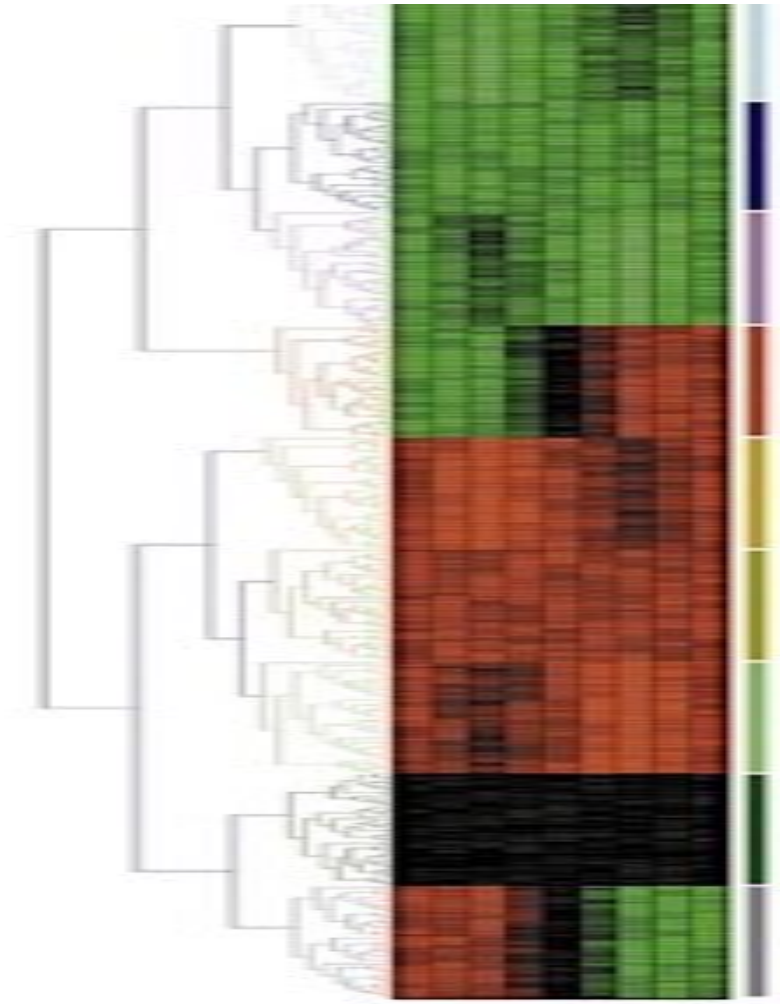
Principle Components Analysis (PCA)

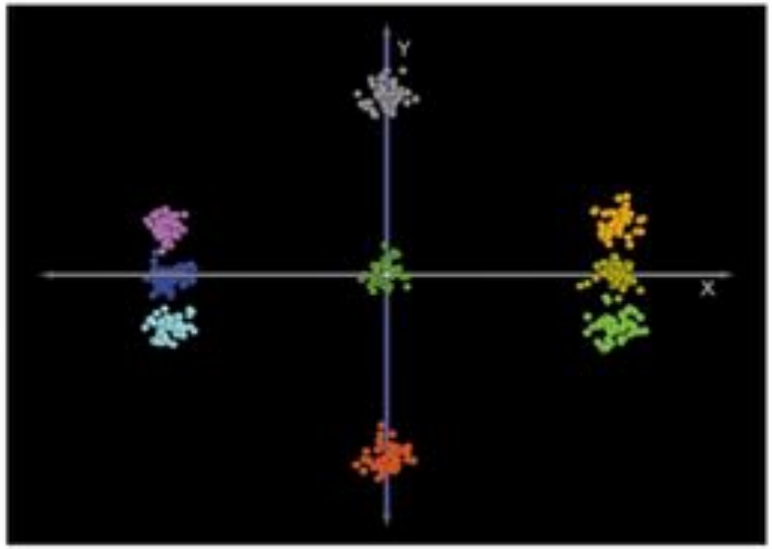




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





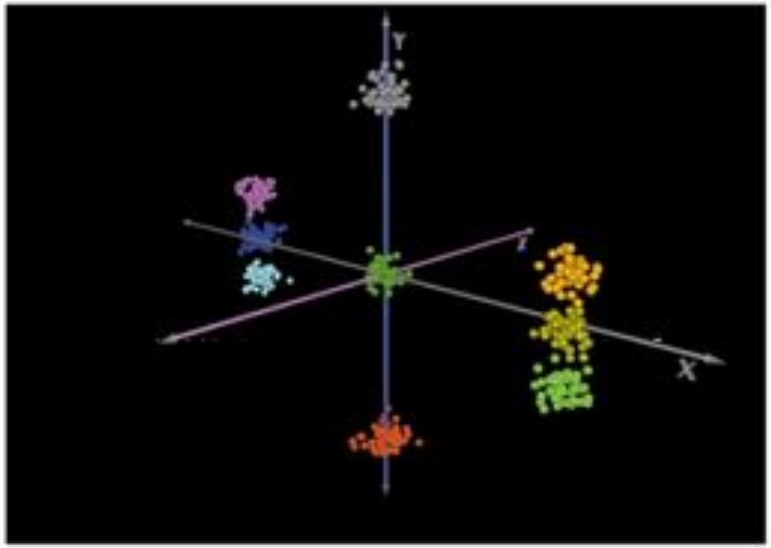
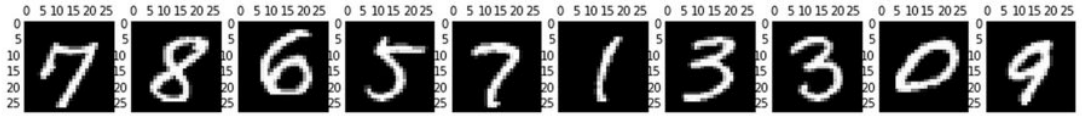
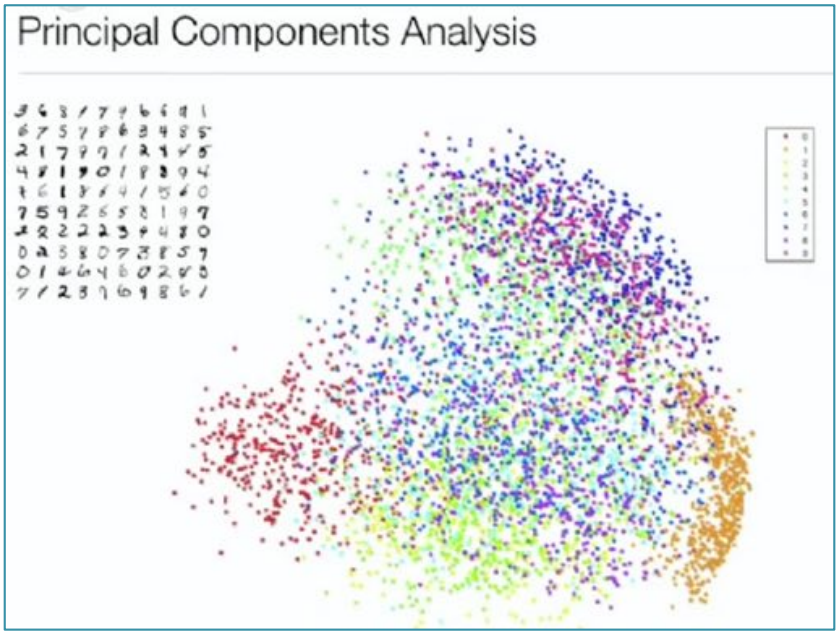
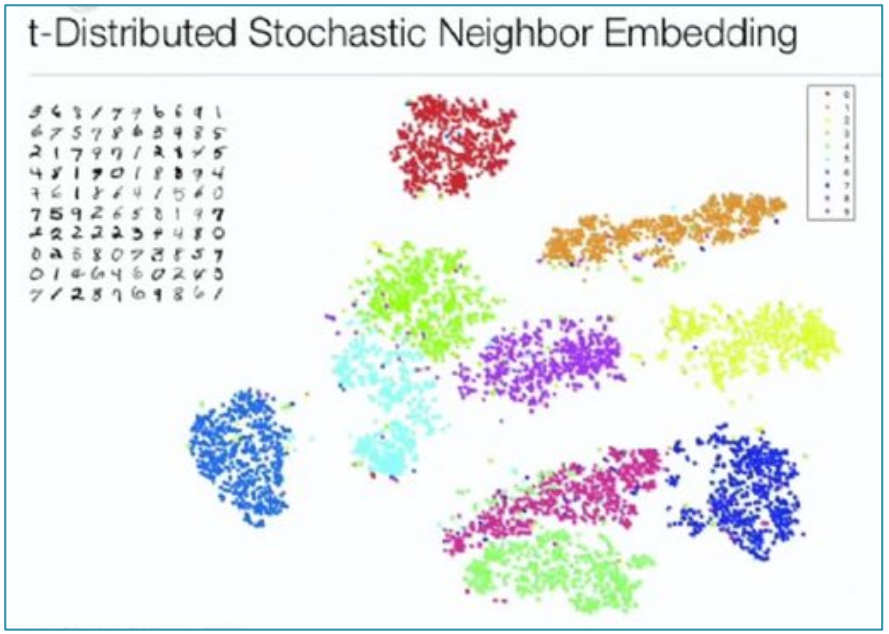


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

PCA and t-SNE







t-distributed Stochastic Neighborhood Embedding

- Non-linear dimensionality reduction technique: distances are only locally meaningful
- Rather than Euclidean distances, for each point fits a Gaussian kernel to fit the nearest N neighbors (perplexity) that define the probabilities that two points should be close together
- Using an iterative spring embedding system to place high probability points nearby

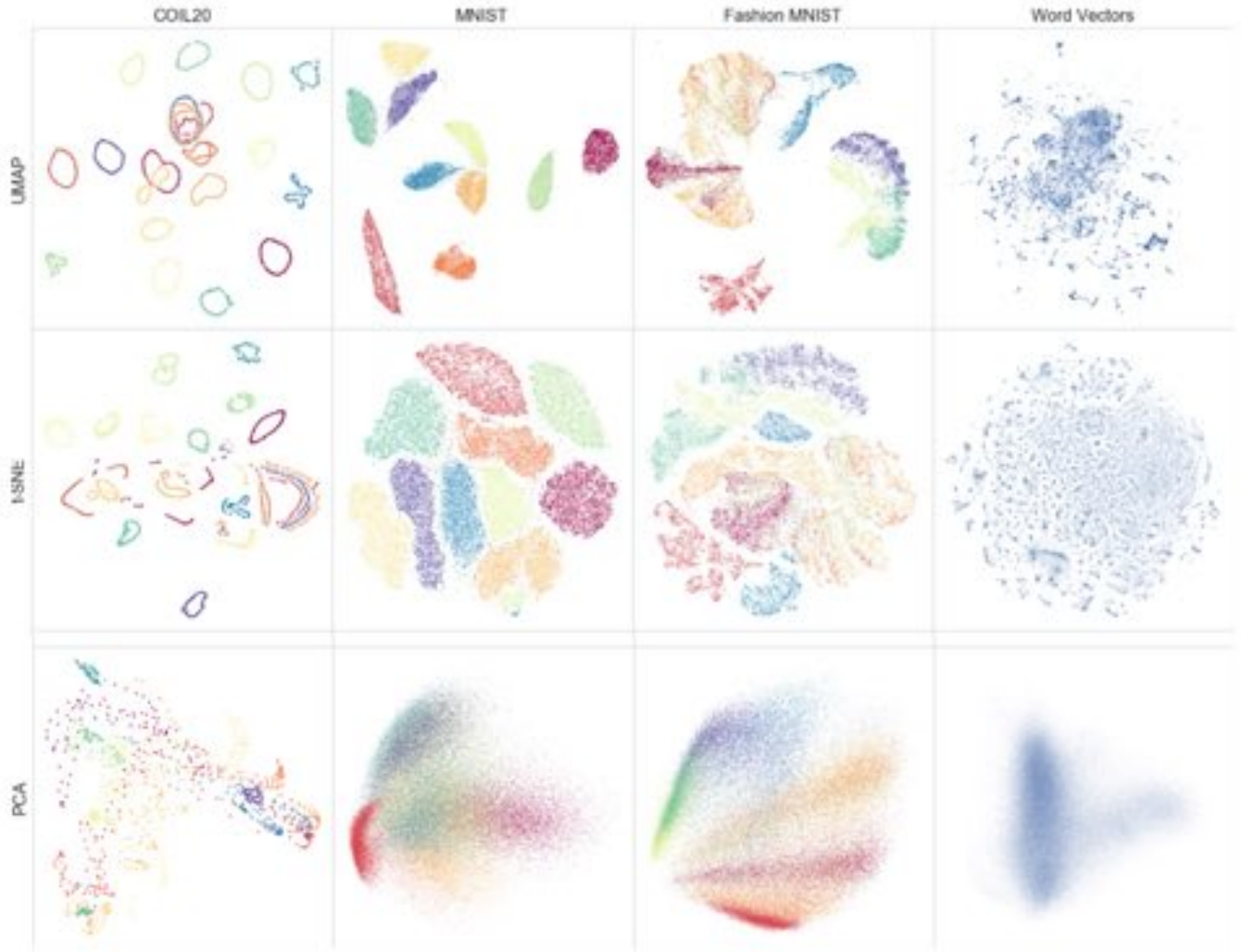
Visualizing Data Using t-SNE

van der Maaten & Hinton (2008) Journal of Machine Learning Research. 9: 2579–2605.

https://www.youtube.com/watch?v=RJVL80Gg3lA

https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

UMAP



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
McInnes et al (2018) arXiv. 1802.03426
https://www.youtube.com/watch?v=nq6iPZVUxZU
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

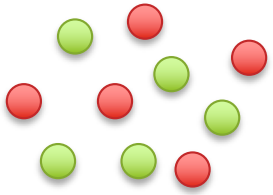
Single Cell Analysis

1. Why single cells?
2. scRNAseq

Population Heterogeneity

Red cells express twice the abundance of “brain” genes compared to green cells

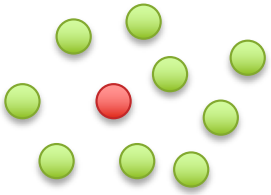
Experiment 1: 50/50



Compared to a control sample of pure green cells, this sample will show:

50% 2x + 50% 1x
= 1.5x over expression of brain genes

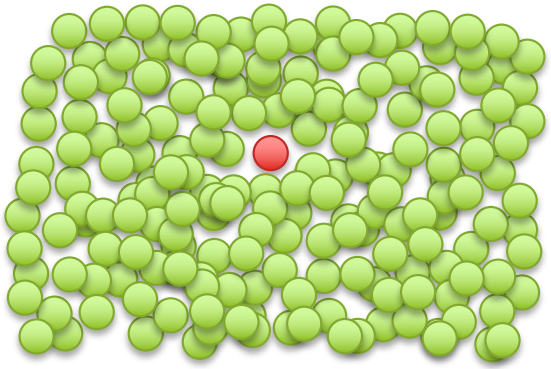
Experiment 2: 1/10



Compared to a control sample of pure green cells, this sample will show:

10% 2x + 90% 1x
= 1.1x over expression of brain genes

Experiment 3: 1/1000



Compared to a control sample of pure green cells, this sample will show:

0.1% 2x + 99.1% 1x
= 1.001x over expression of brain genes

The limitations of averages

	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**

The limitations of averages

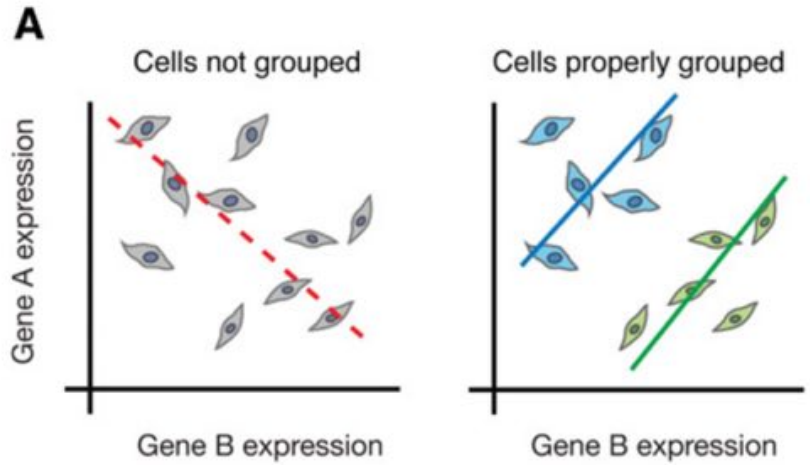
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

Example of Simpson's paradox:
Trend of the overall average may reverse the trends of each constituent group

In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included. (Based on real analysis of kidney stone treatments)

The paradox of averages



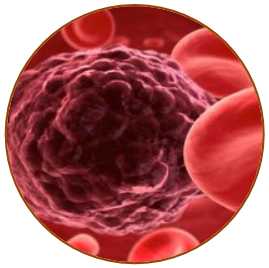
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

Example of Simpson's paradox:
Trend of the overall average may reverse the trends of each constituent group

In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included. (Based on real analysis of kidney stone treatments)

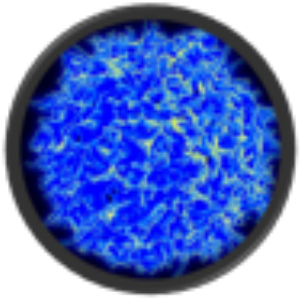
(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity



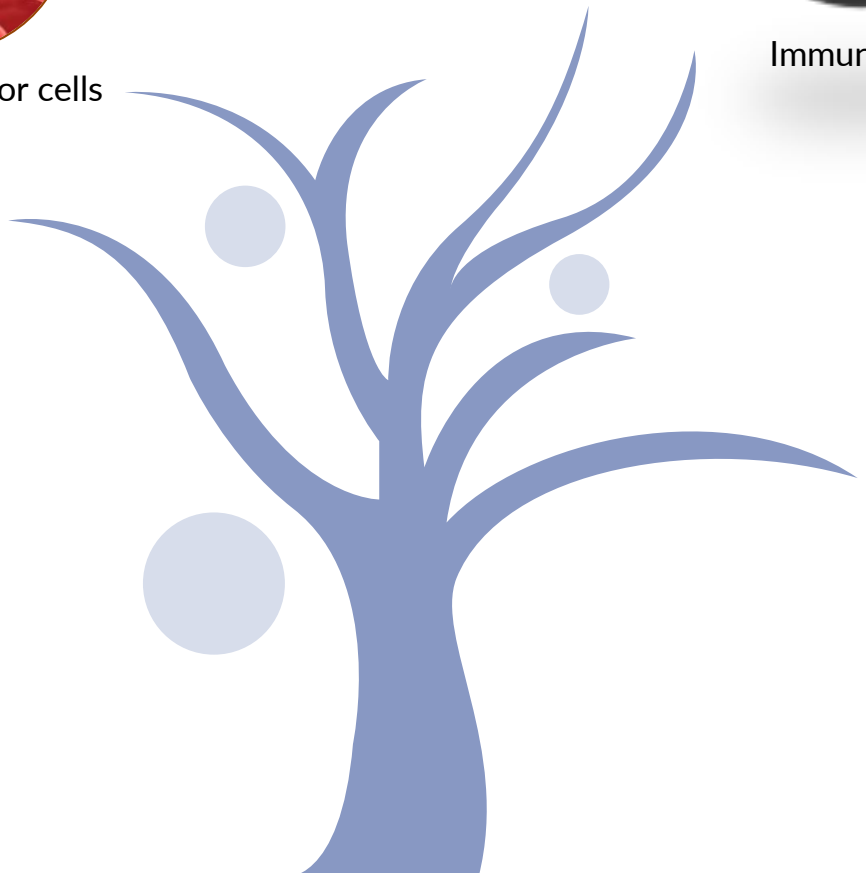
Circulating tumor cells

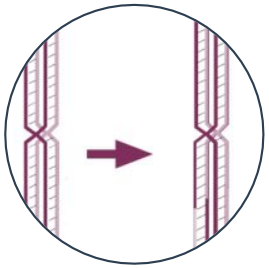
Neuronal mosaicism



Immune Cells

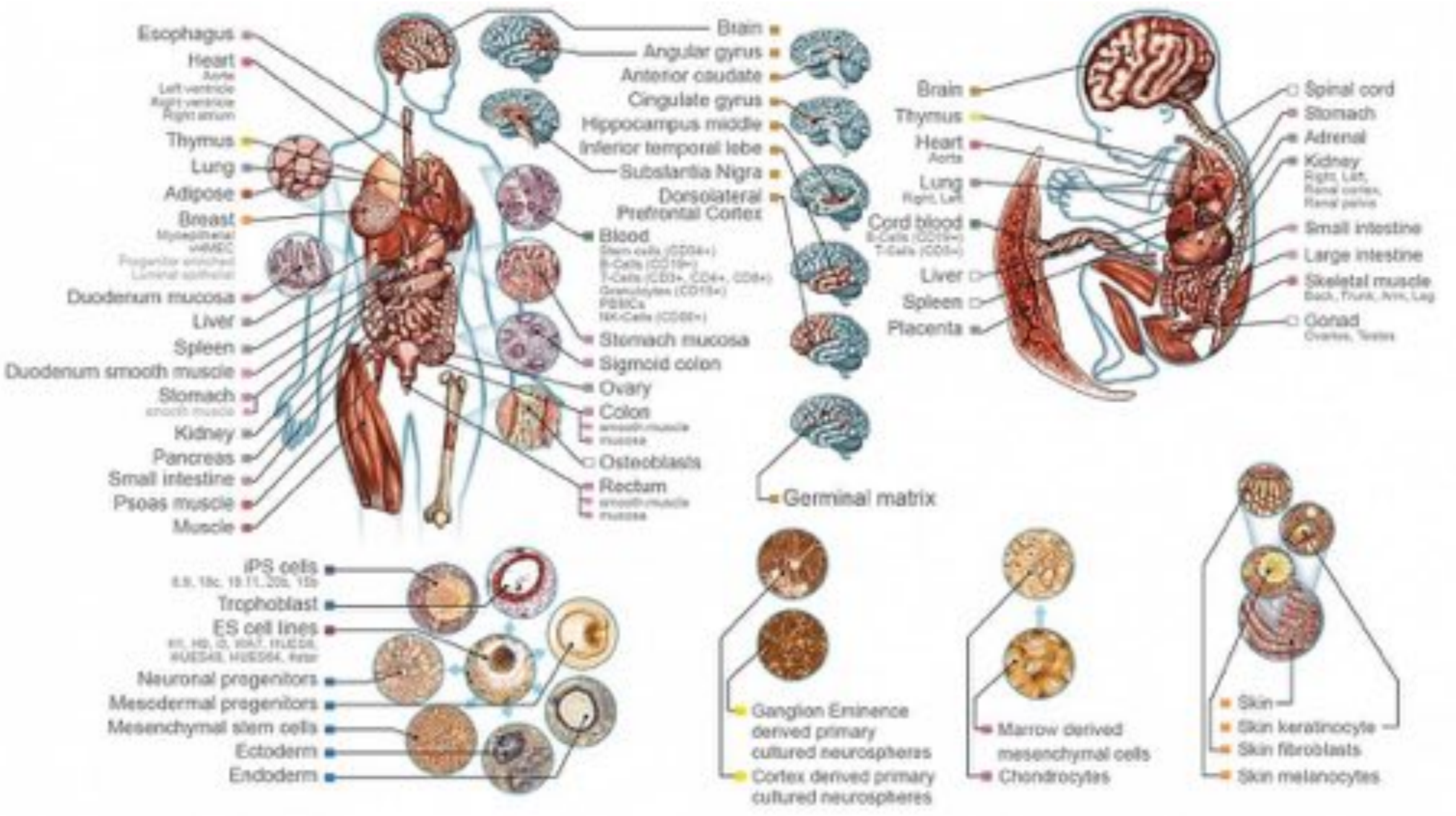
Tumor evolution



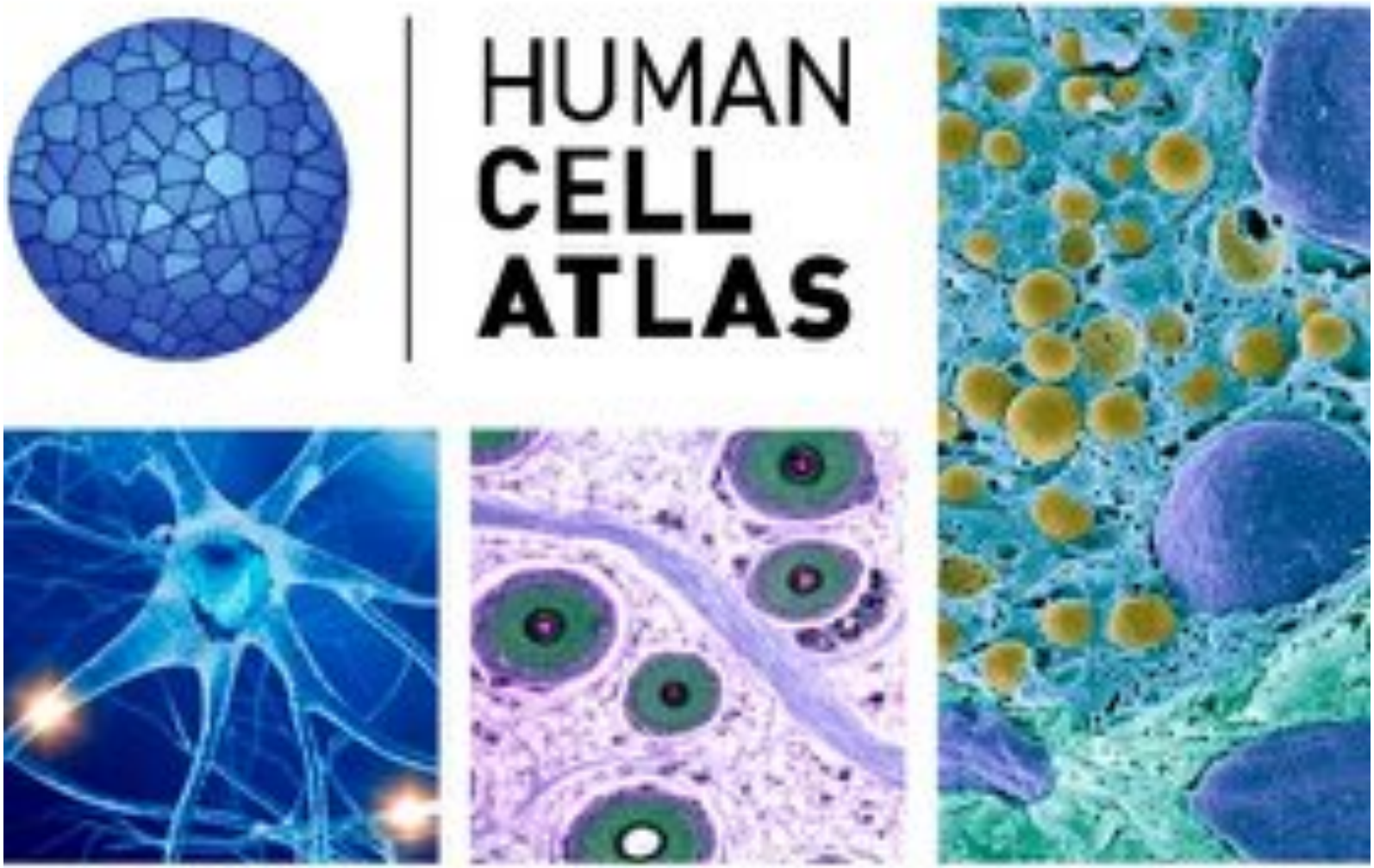


Recombination/crossover
in germ cells

Sources of (Cellular) Heterogeneity



Roadmap Epigenomics Consortium



https://www.humancellatlas.org/

Single Cell Analysis

1. Why single cells?
2. scRNAseq

Single-cell RNA sequencing, "the bioinformatician's microscope"

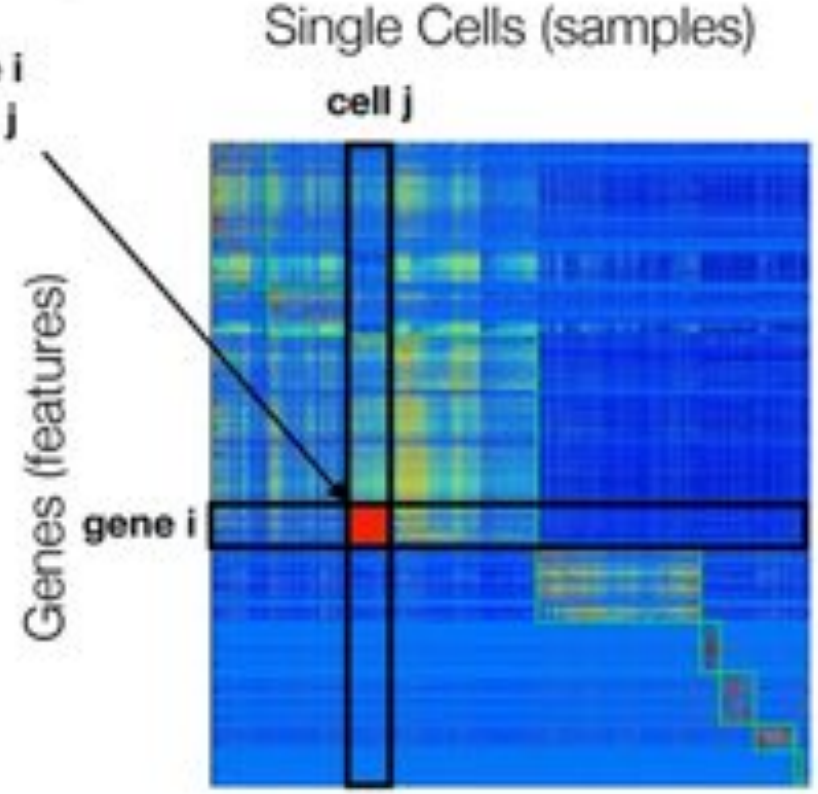
— a snapshot of the underlying biology in a data matrix.

Brain cells

Biological sample

number of times **gene i** was expressed in **cell j**

sequencing

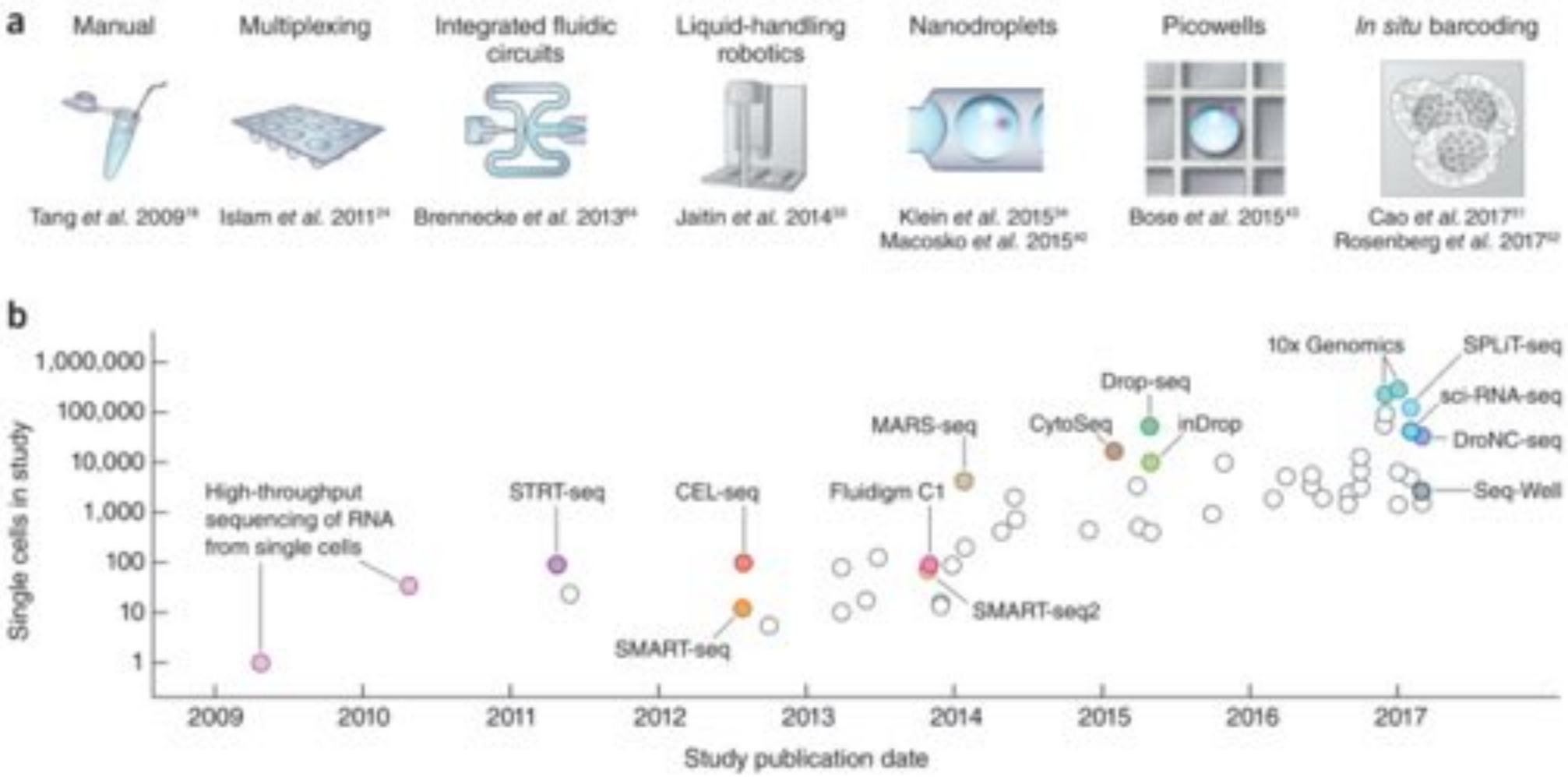


Gene expression matrix

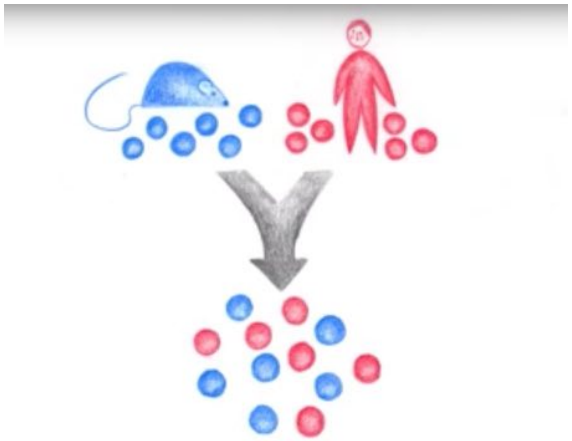
computationally explore complex biological systems

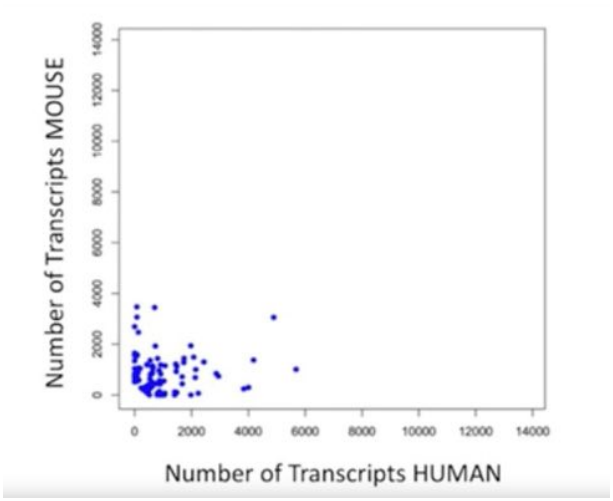
Martin Zhang

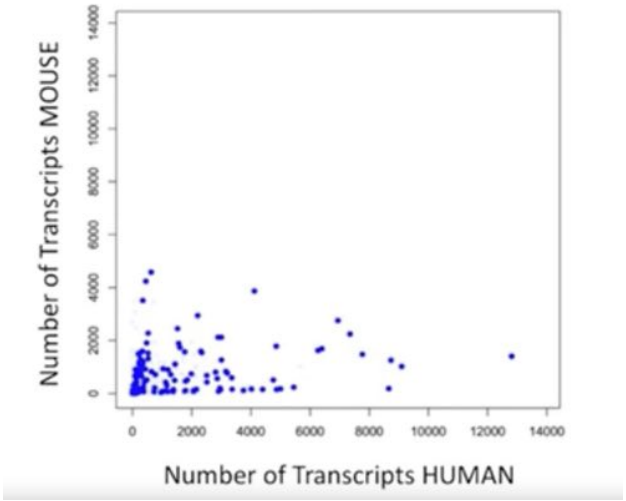
A decade of single-cell RNA-seq

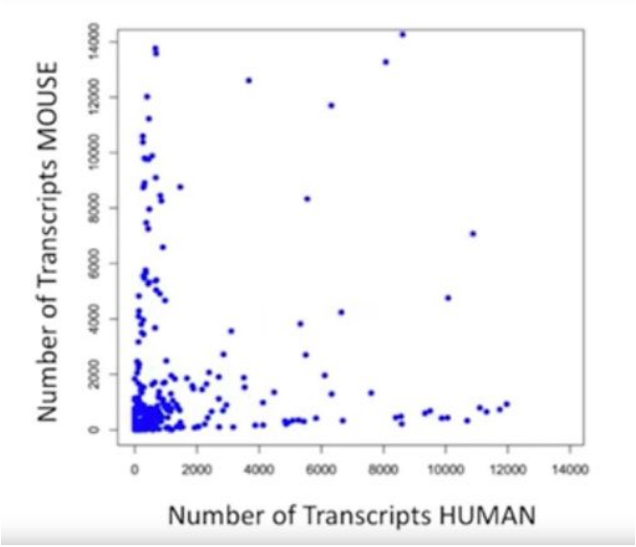


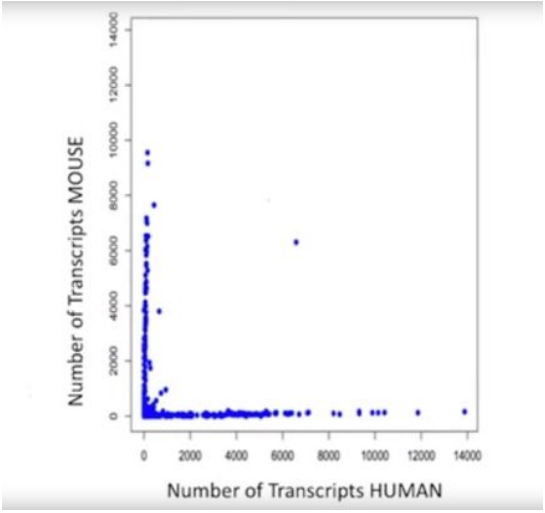
Svensson et al. 2018. DOI:10.1038/nprot.2017.149

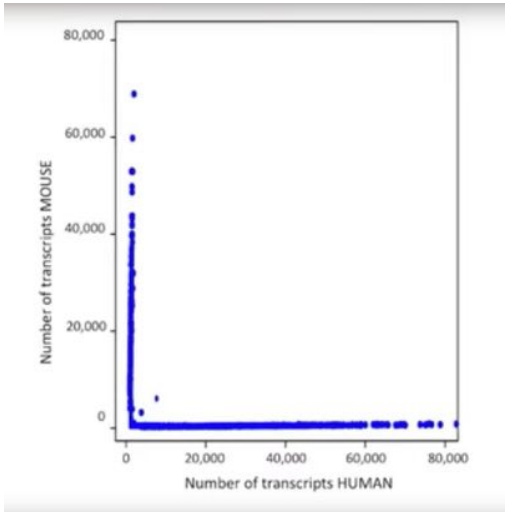






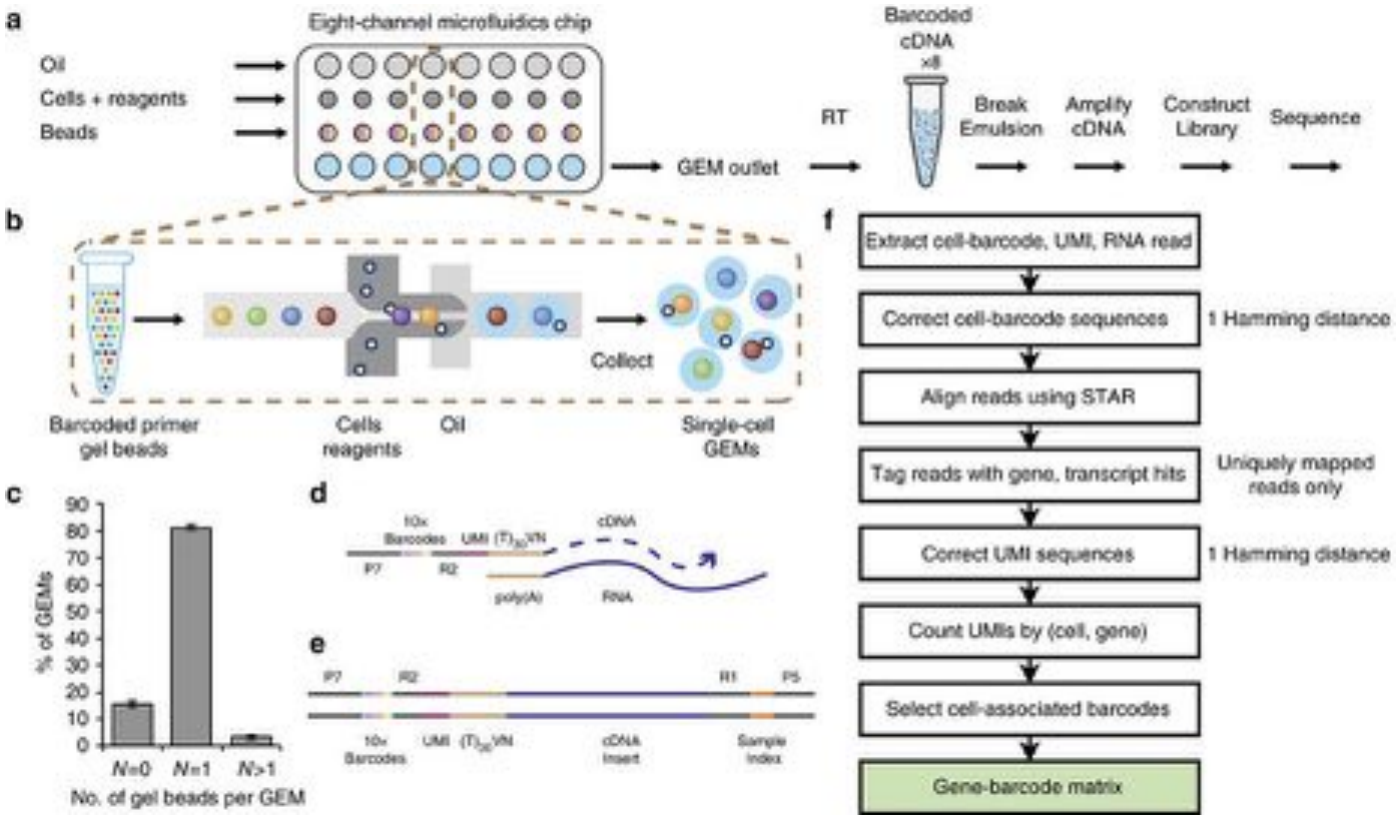






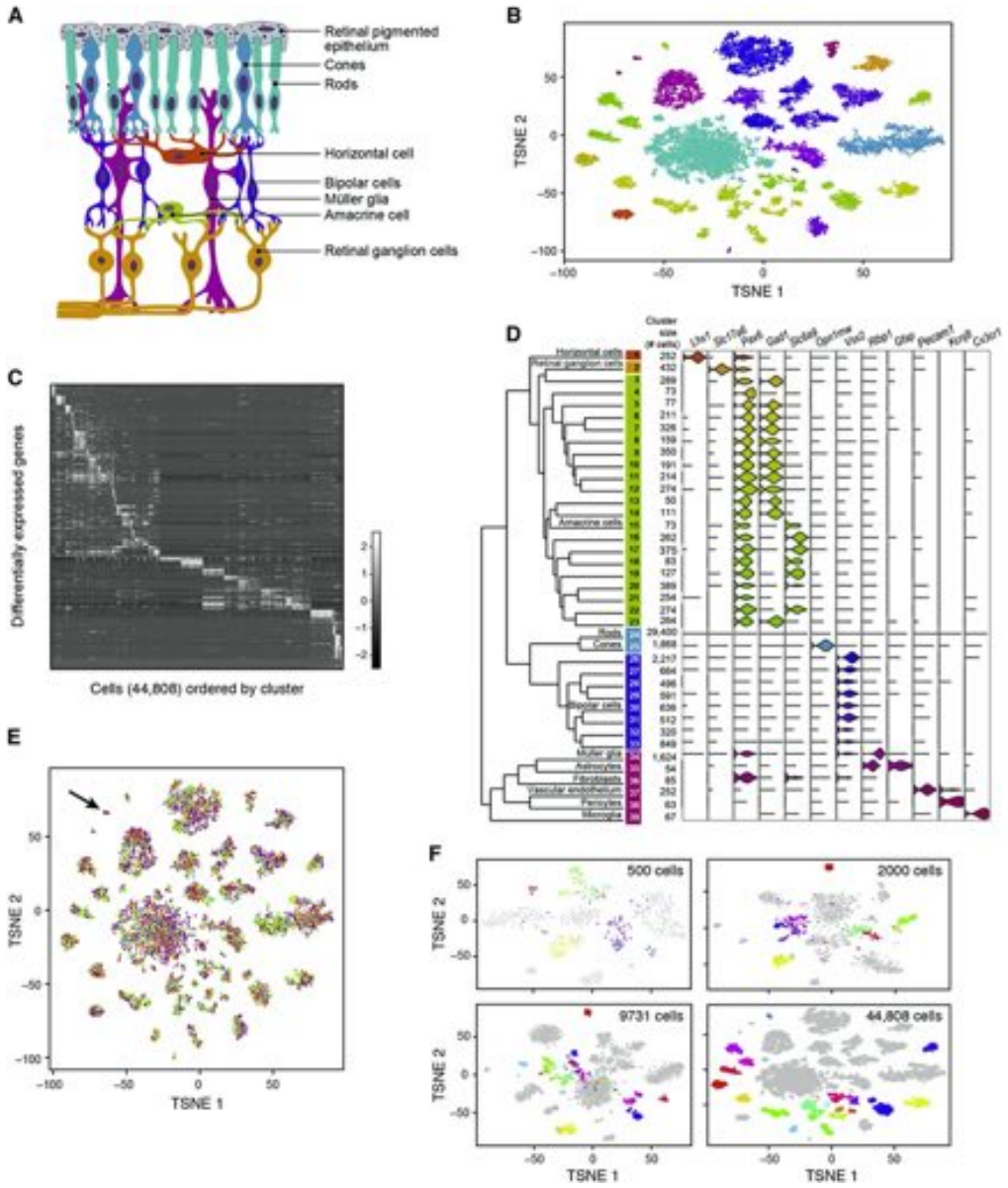
Drop-seq: Droplet barcoding of single cells
https://www.youtube.com/watch?v=vL7ptq2Dcf0

10x GENOMICS



Up to 1M cells in a single analysis

Massively parallel digital transcriptional profiling of single cells
Zheng et al (2017) Nature Communication. doi:10.1038/ncomms14049



Key Results

(a) schematic of known cell populations in retina

(b) 44,808 Drop-Seq profiles clustered into 39 retinal cell populations using tSNE

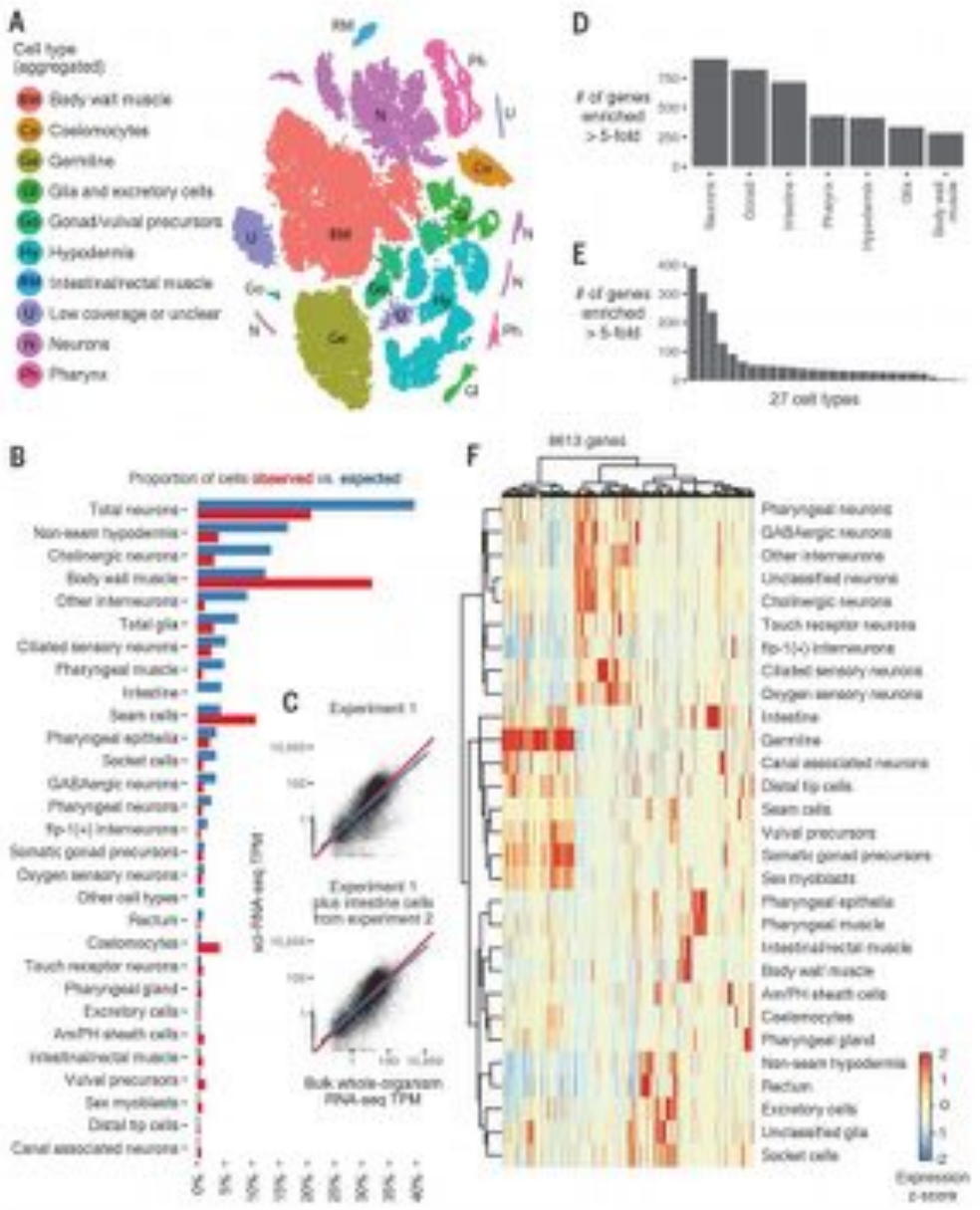
(c) Differentially expressed genes in each cluster

(d) Different cell types can be recognized using marker genes

(e) replicates well

(f) robust to down sampling

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets
Macosko et al (2015) Cell. https://doi.org/10.1016/j.cell.2015.05.002



Key Results

Profile every cell of C. elegans larva using combinatorial indexing

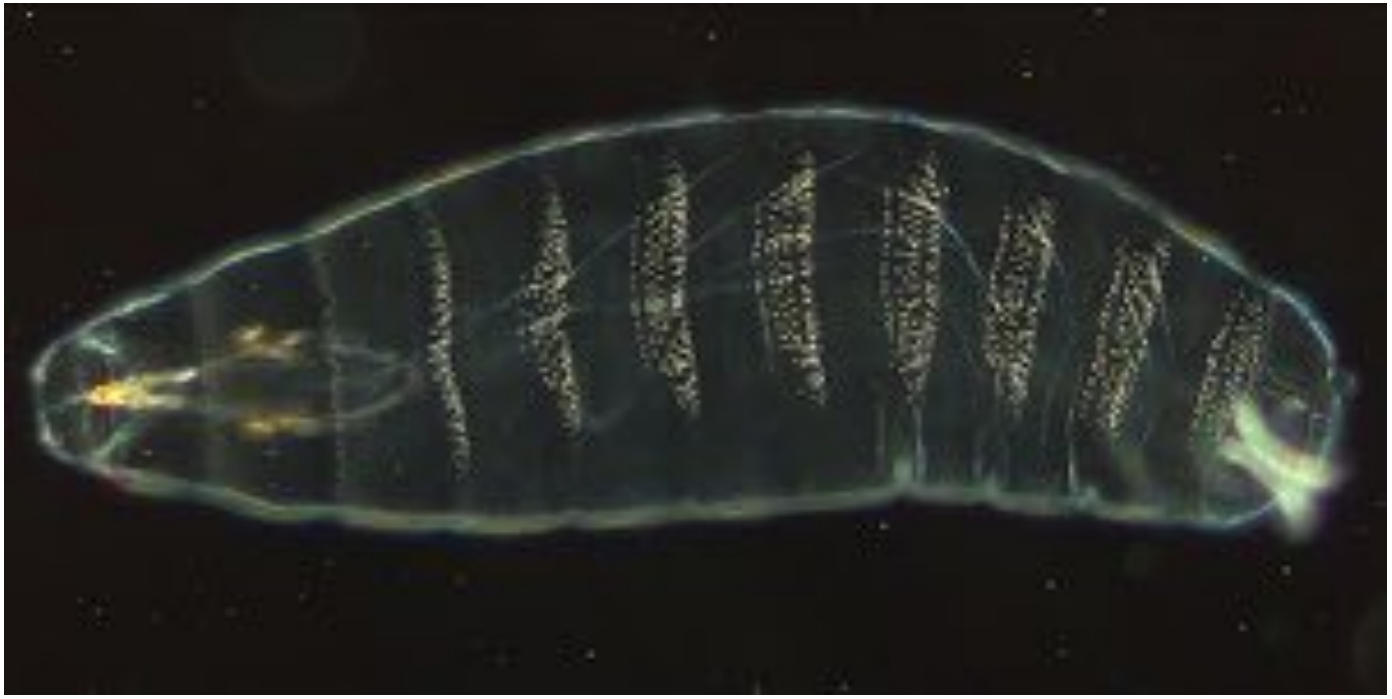
(a) t-SNE visualization of clusters

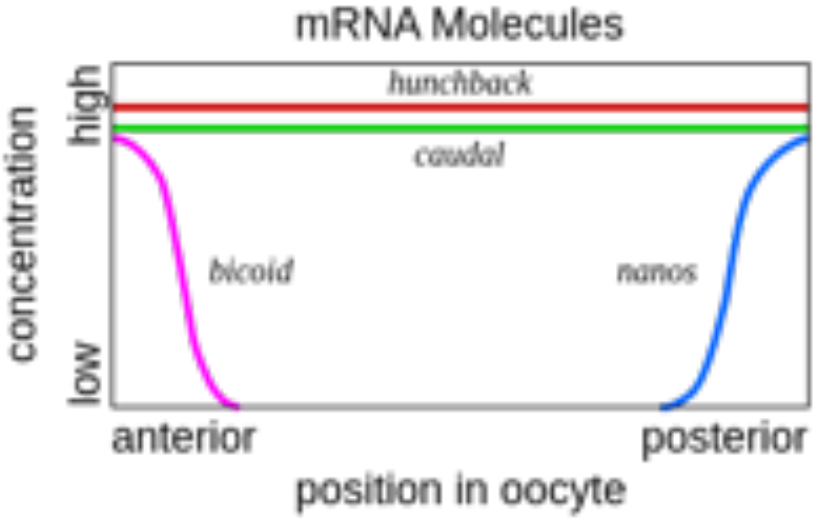
(b) Proportion of cells observed vs expected match well (including cells that only occur once or twice in the animal)

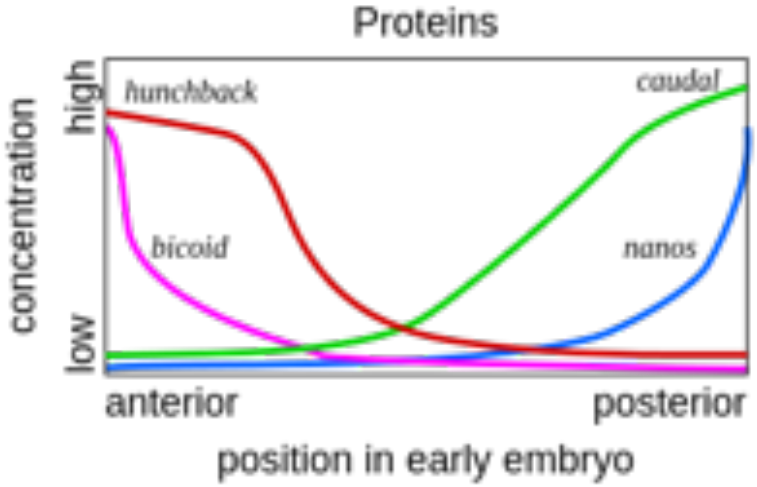
(c) Good correlation between single cell and bulk analysis of selected cell types

(d-f) Analysis of key genes per cell type

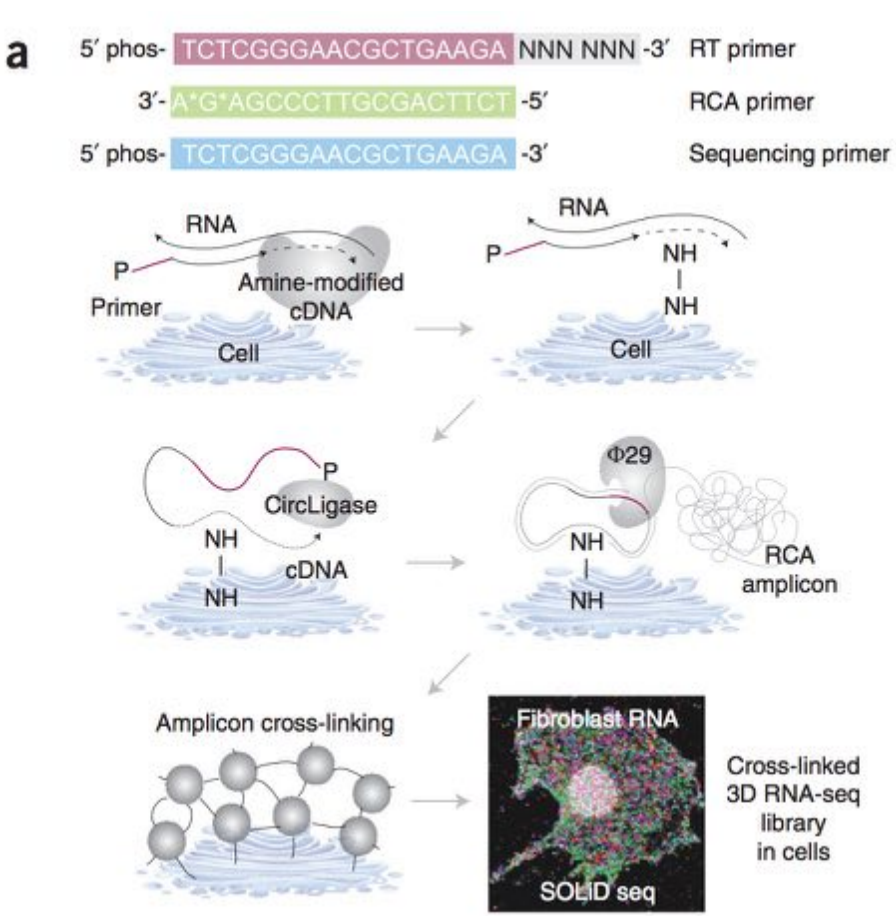
Comprehensive single-cell transcriptional profiling of a multicellular organism
Cao et al (2017) Science. 357:661-557

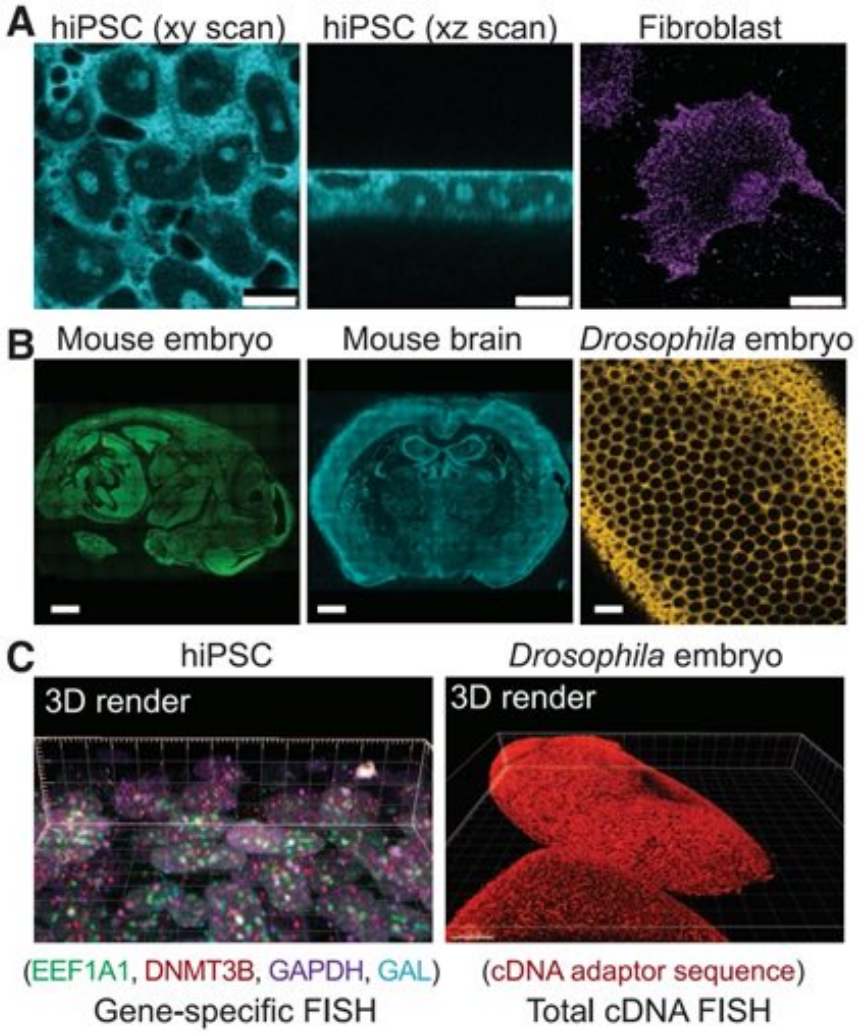






https://en.wikipedia.org/wiki/Drosophila_embryogenesis





Highly multiplexed subcellular RNA sequencing in situ ("FISSEQ")

Lee et al (2014) Science. doi: 10.1126/science.1250212

Summary

Single cell analysis is a powerful tool to study heterogeneous tissues

- Overcomes fundamental problems that can arise when averaging
- scRNA analysis used to identify novel cell types, understand the progression from one cell type to another across development or disease
- Many other sc-assays in development, expect 1000s to 1Ms of cells in essentially any assay

Major challenges

- Very sparse amplification and few reads per cell
- Find large CNVs, identify major cell types; hard to find small variants or perform differential expression
- Allelic-dropout and unbalanced amplification hides or distorts information
- Use statistical approaches to smooth results based on prior information or other cells from the same cell type

- Need new ways to process and analyze millions of cells at a time