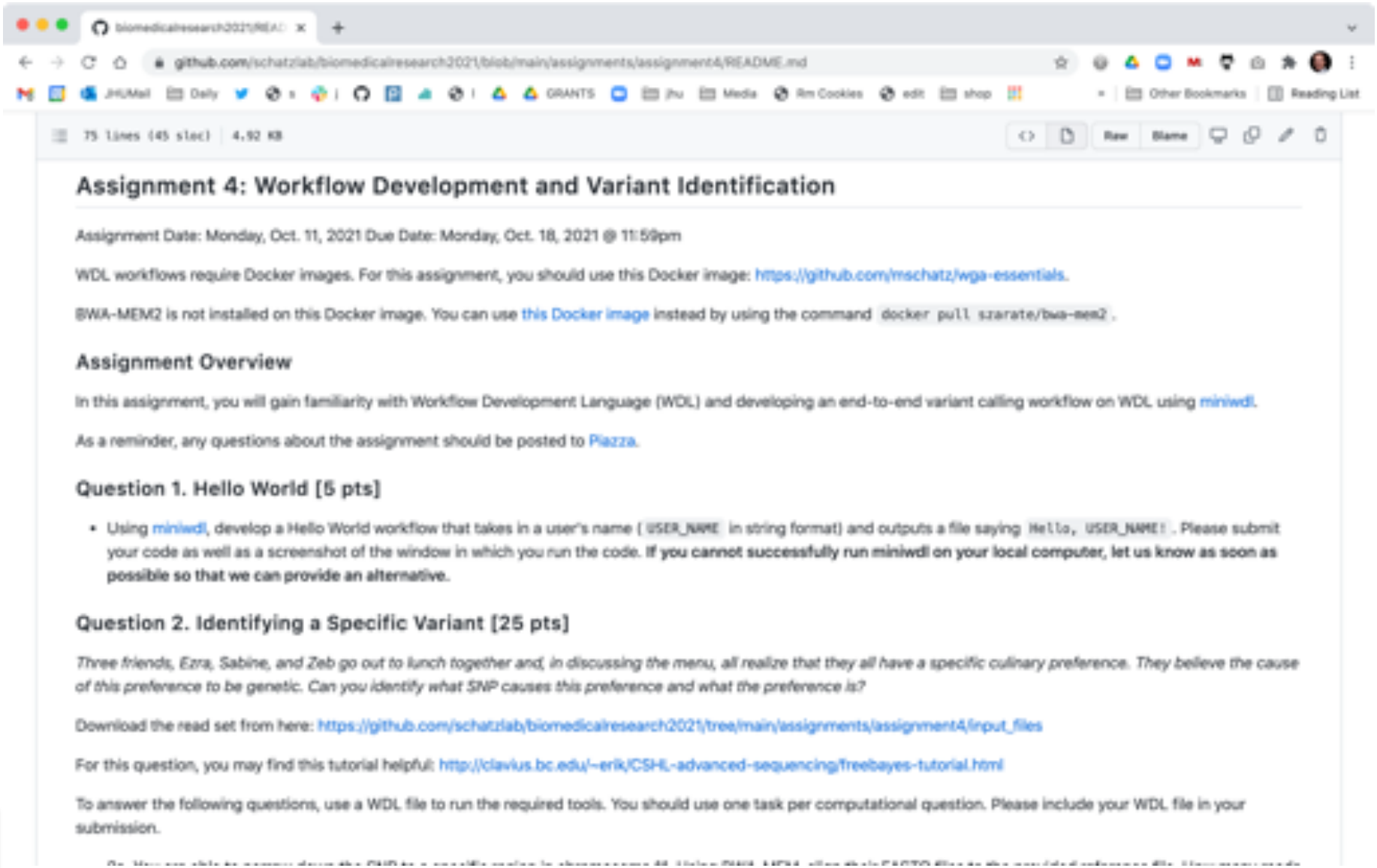
Lecture 15. RNAseq + scRNAseq

Michael Schatz

October 20, 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

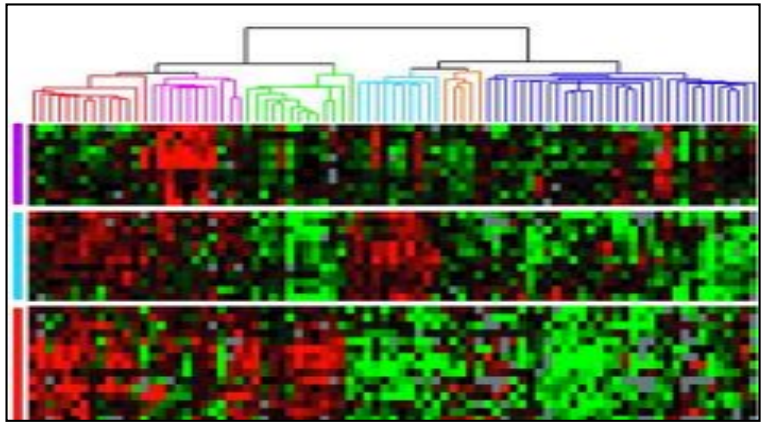
For this question, you may find this tutorial helpful: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

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.

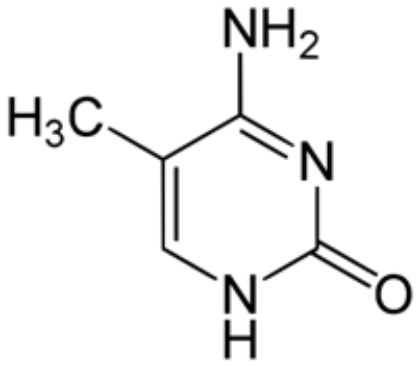
https://github.com/schatzlab/biomedicalresearch2021

*-seq in 4 short vignettes

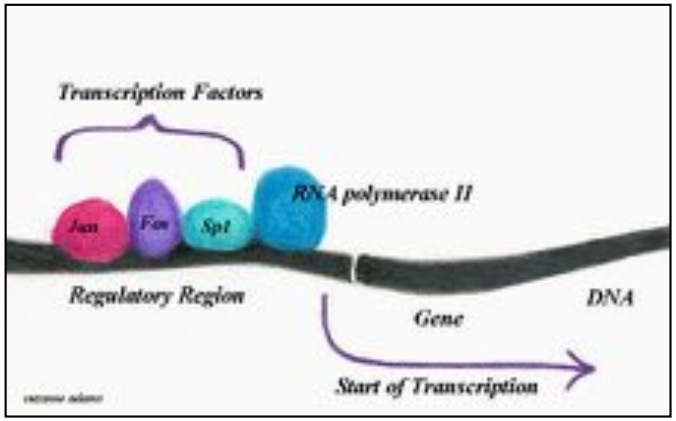
RNA-seq



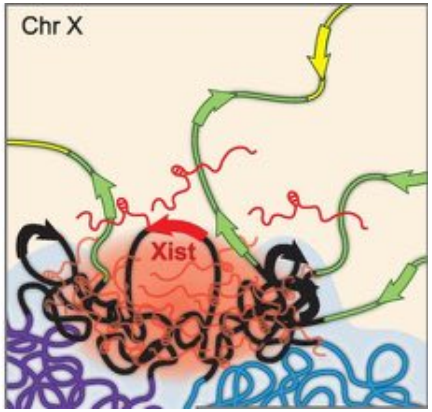
Methyl-seq



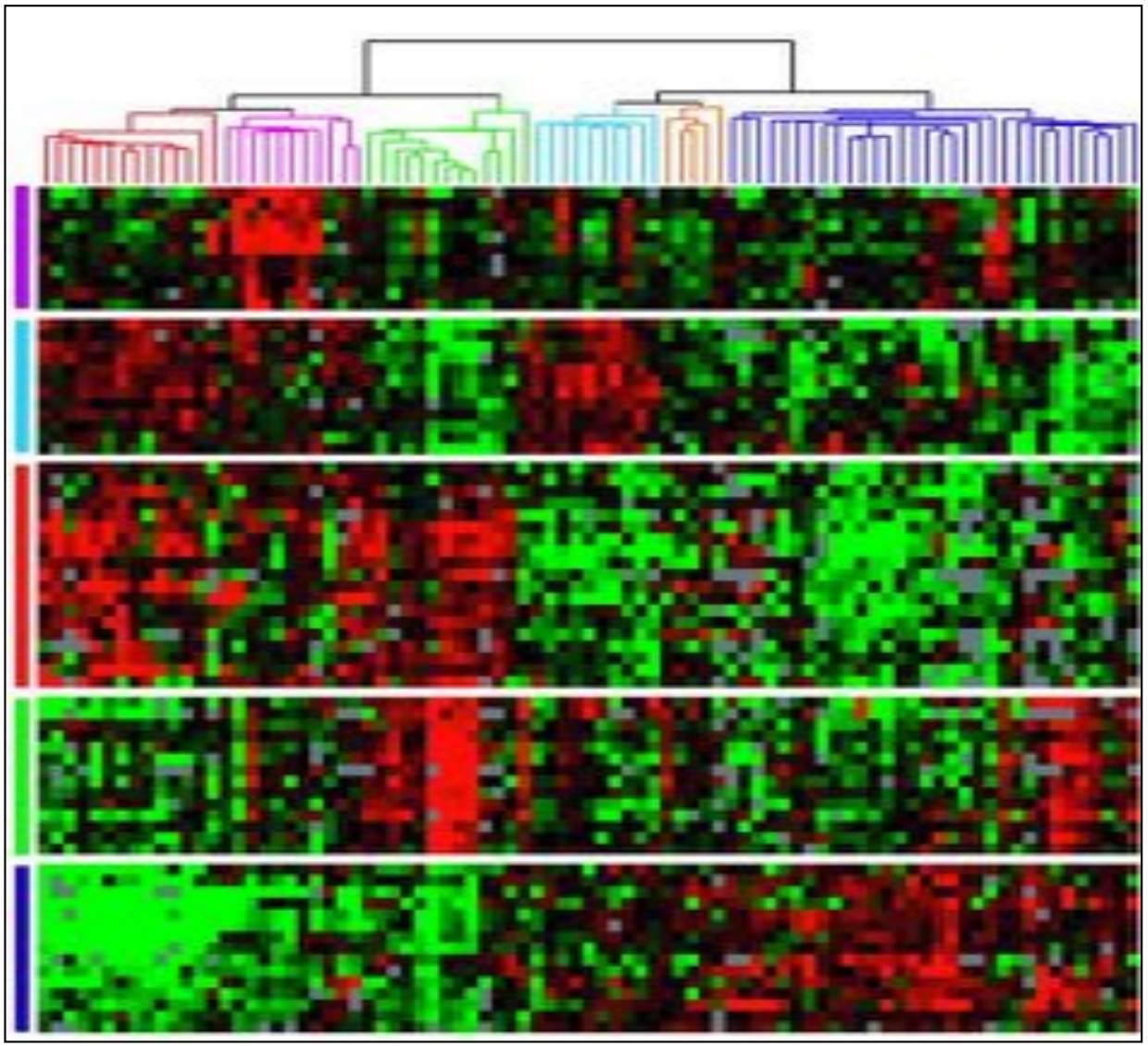
ChIP-seq



Hi-C

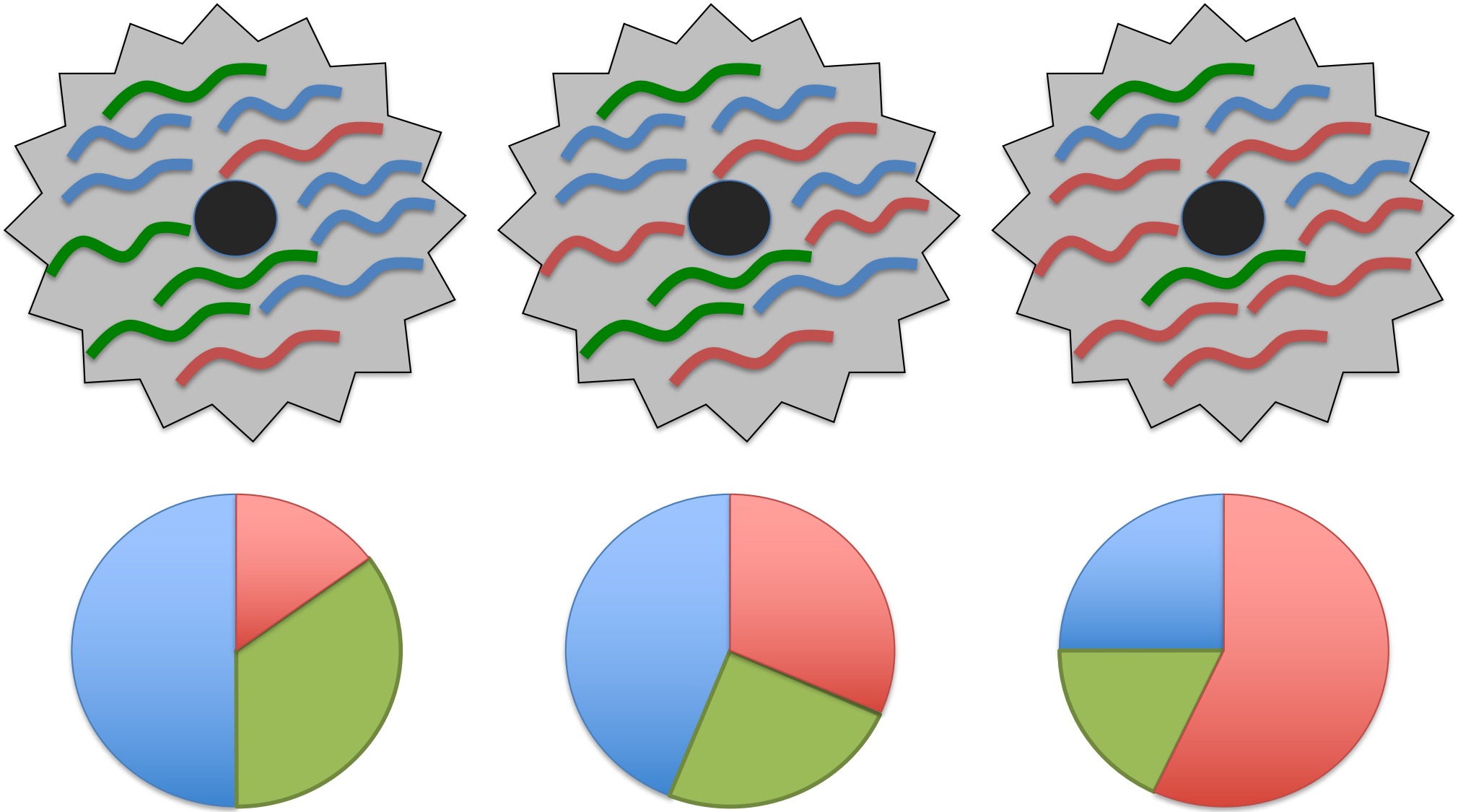


RNA-seq

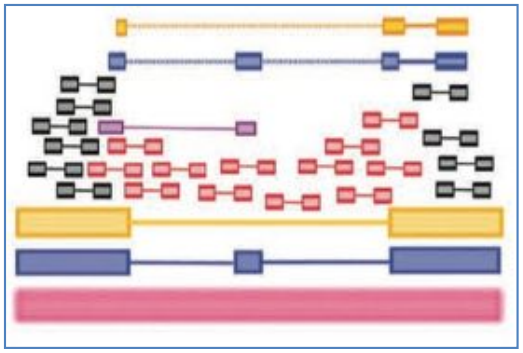


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

RNA-seq Overview

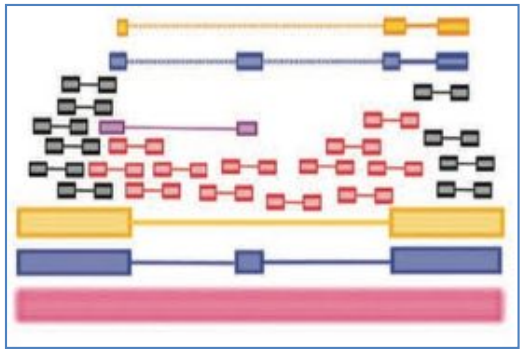


RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

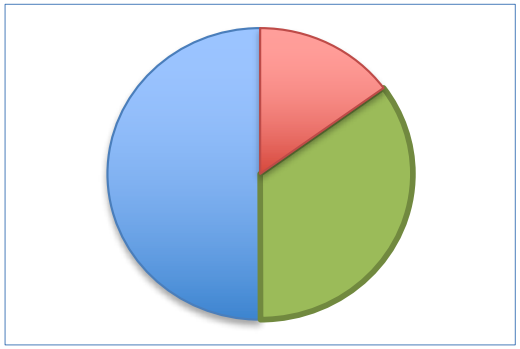
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

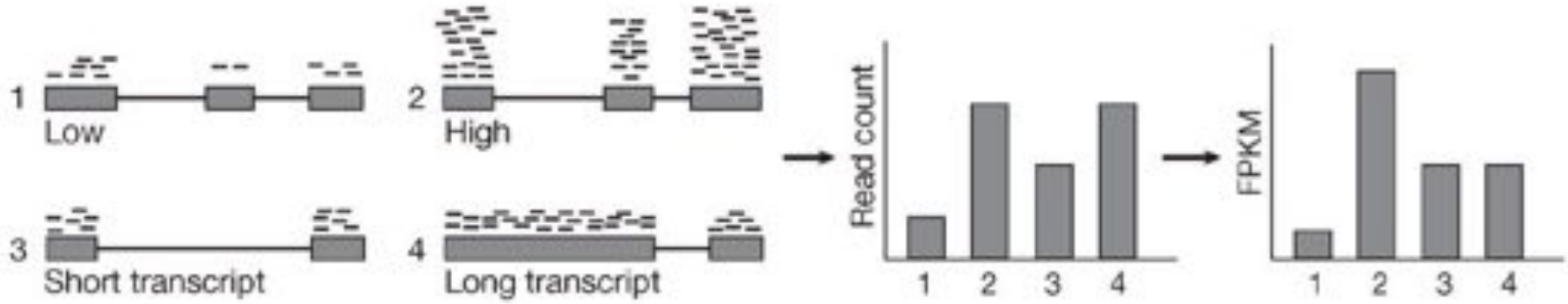
Solution: Use a spliced aligner, and assemble isoforms

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



Challenge 2: Read Count != Transcript abundance

RPKM, FPKM, TPM



Counting Reads that align to a gene DOESN'T work!

- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)

=> Wait a second, reads in a pair arent independent!

2. FPKM: Fragments Per Kilobase of Exon Per Million Reads Mapped (Trapnell et al, 2010)

=> Wait a second, FPKM depends on the average transcript length!

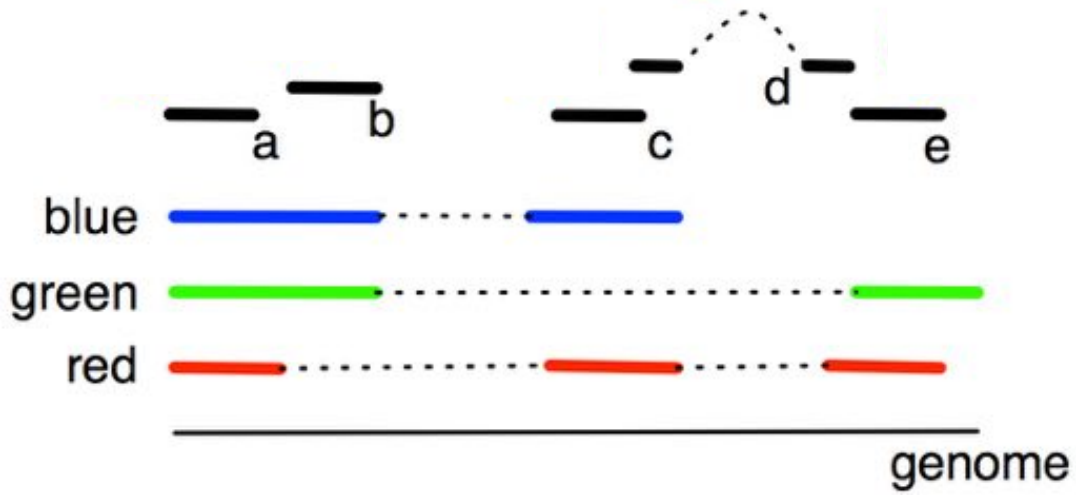
3. TPM: Transcripts Per Million (Li et al, 2011)

⇒ If you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen of type i, given the abundances of the other transcripts in your sample

=> Recommend you use TPM for all analysis, easy to compute given FPKM

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length.
Our initial expectation is all 3 isoforms are equally expressed

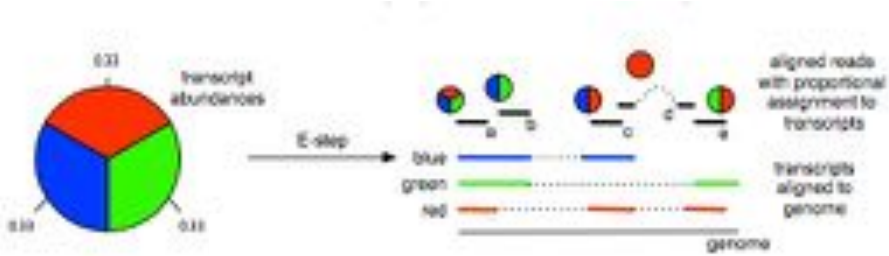
There are five reads (a,b,c,d,e) mapping to the gene.

- Read a maps to all three isoforms
- Read d only to red
- Reads b,c,e map to each of the three pairs of isoforms.

What is the most likely expression level of each isoform?

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



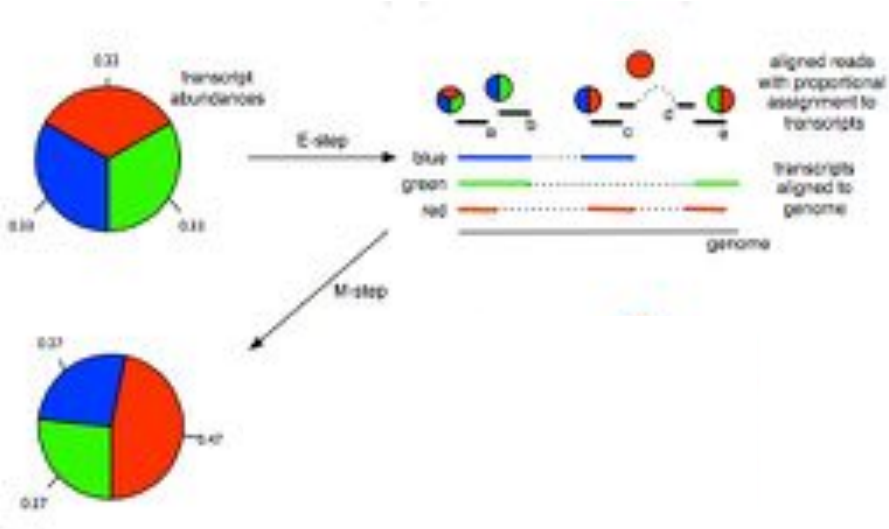
The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

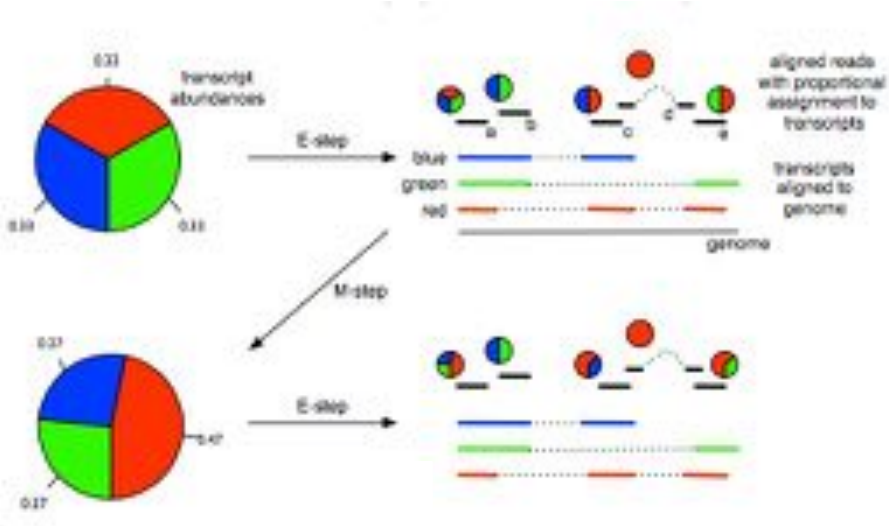
There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

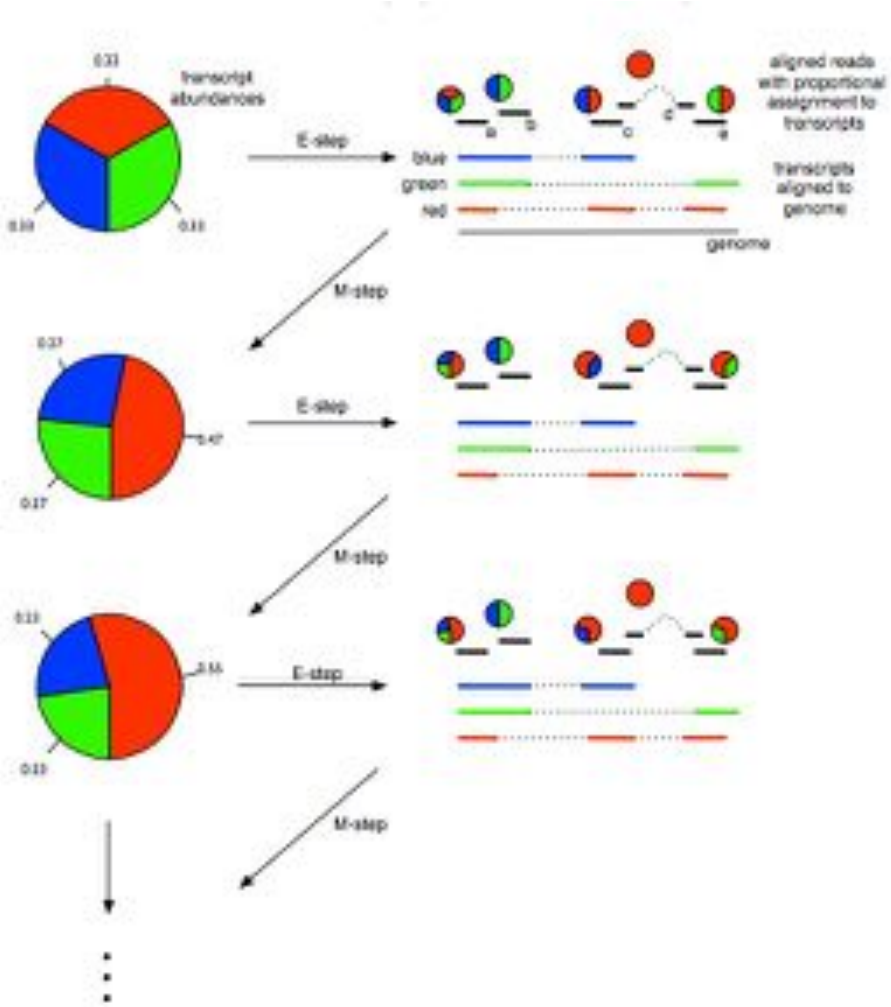
During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

Repeat until convergence!

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

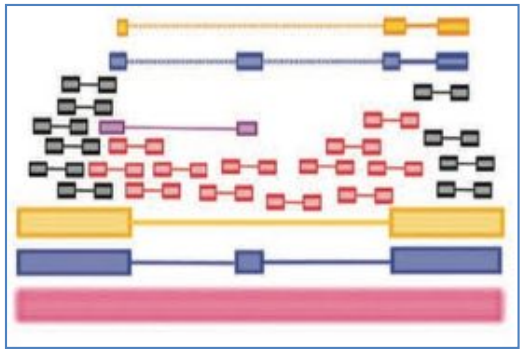
Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:

red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

Repeat until convergence!

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

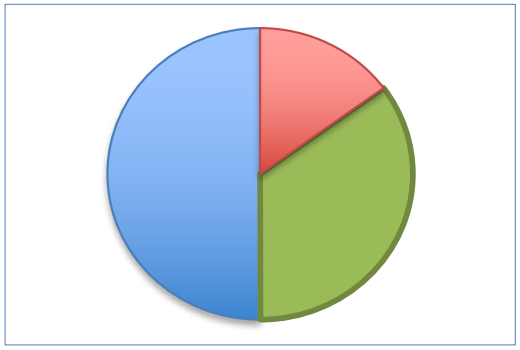
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

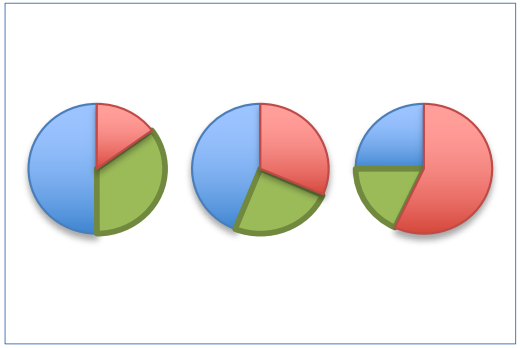
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

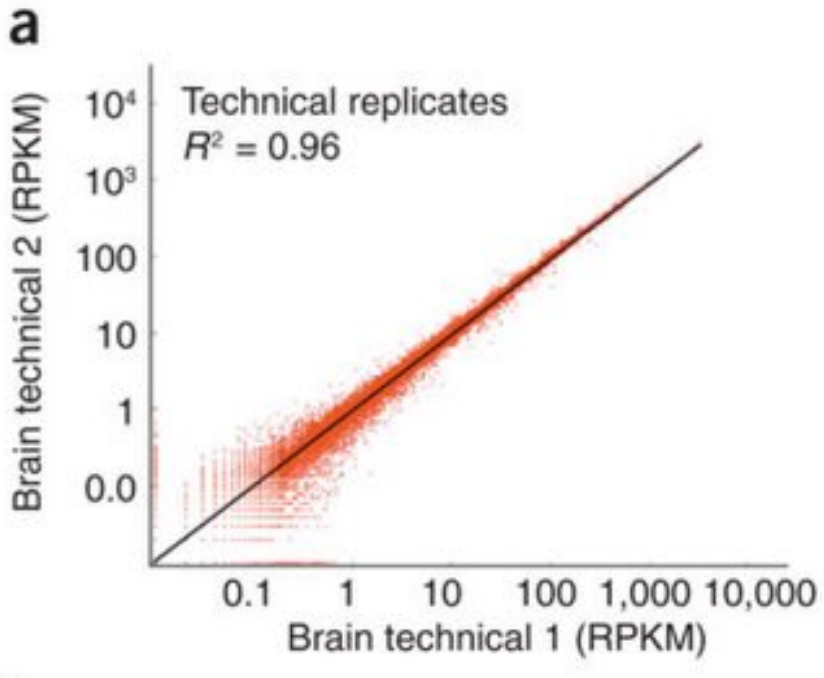
Solution: Infer underlying abundances (e.g. TPM)

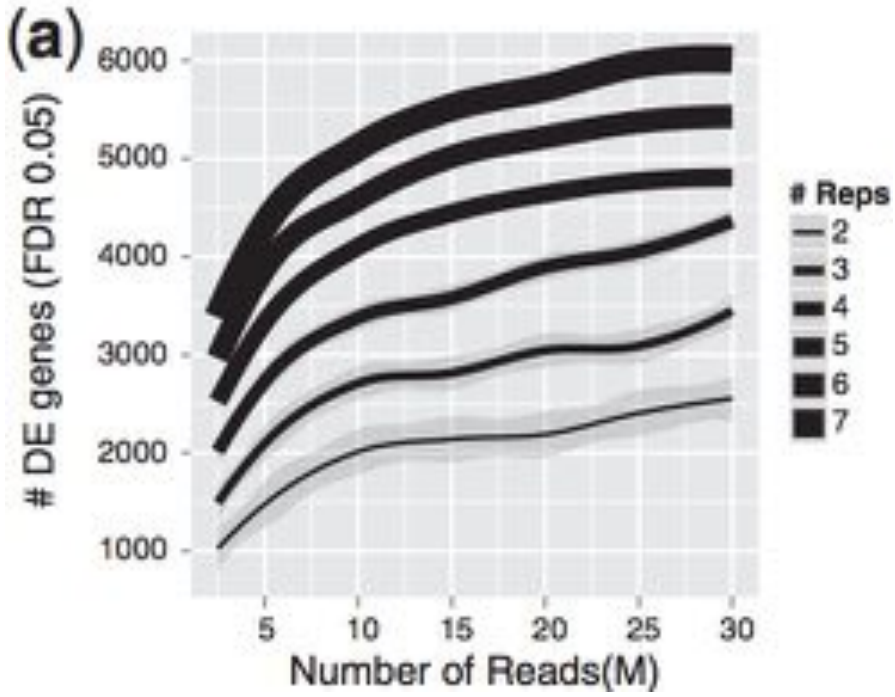
Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

How Many Replicates?



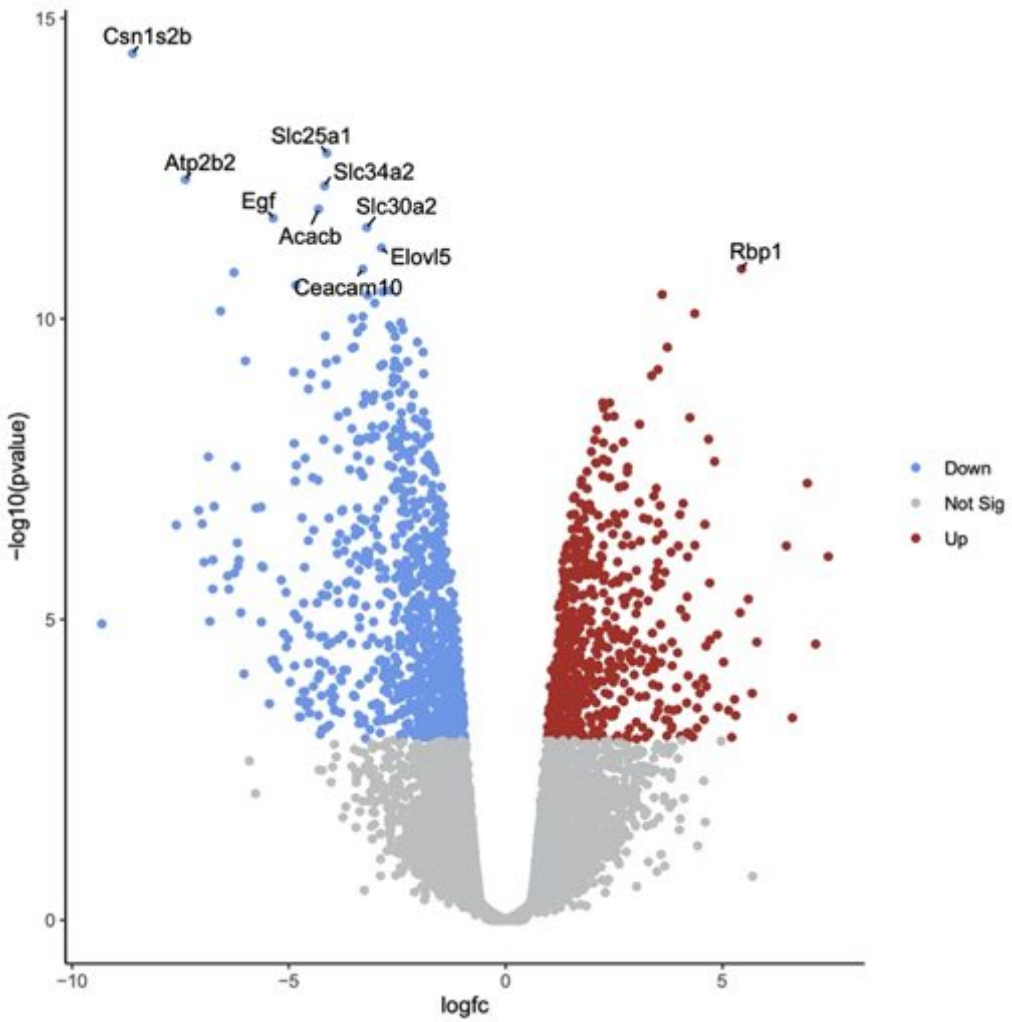


Why don't we have perfect replicates?

Mapping and quantifying mammalian transcriptomes by RNA-Seq
Mortazavi et al (2008) Nature Methods. 5, 62-628

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

Volcano Plots



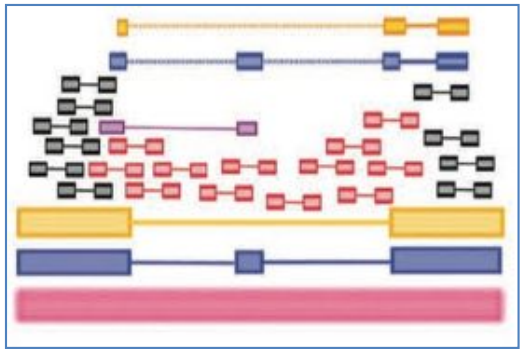
Volcano plots are commonly used to display the results of RNA-seq or other omics experiments.

- Scatterplot that shows statistical significance (P value) versus magnitude of change (fold change).
- Enables quick visual identification of genes with large fold changes that are also statistically significant.
- The most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top.

Visualization of RNA-Seq results with Volcano Plot

https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

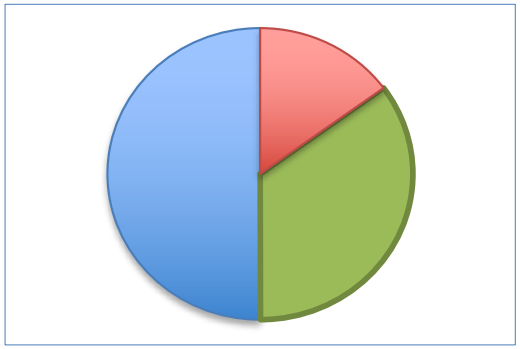
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

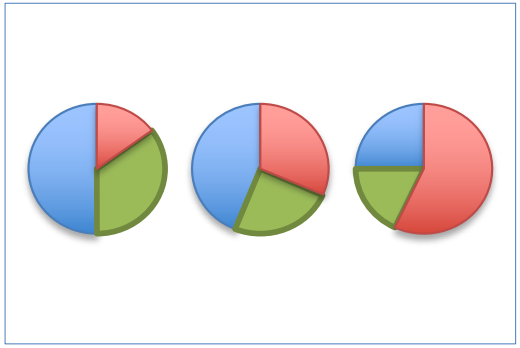
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



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

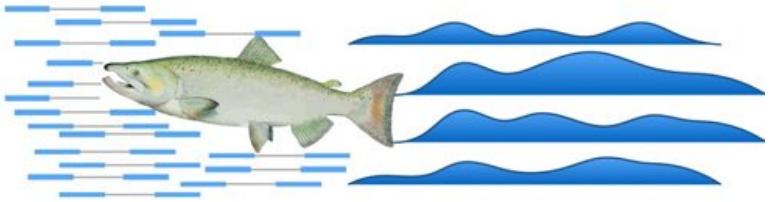


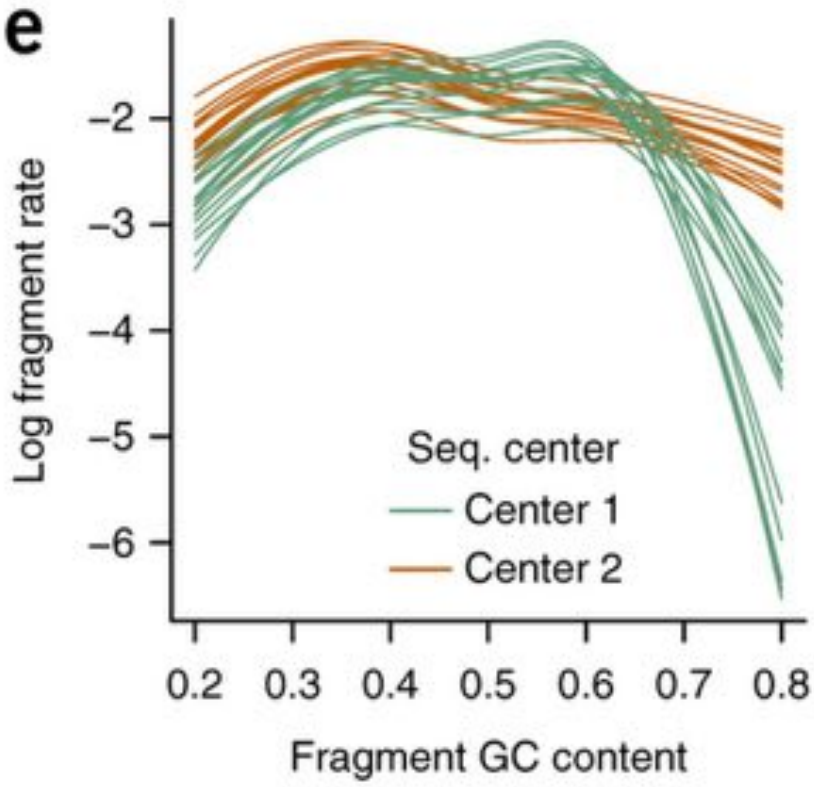
Challenge 3: Transcript abundances are stochastic

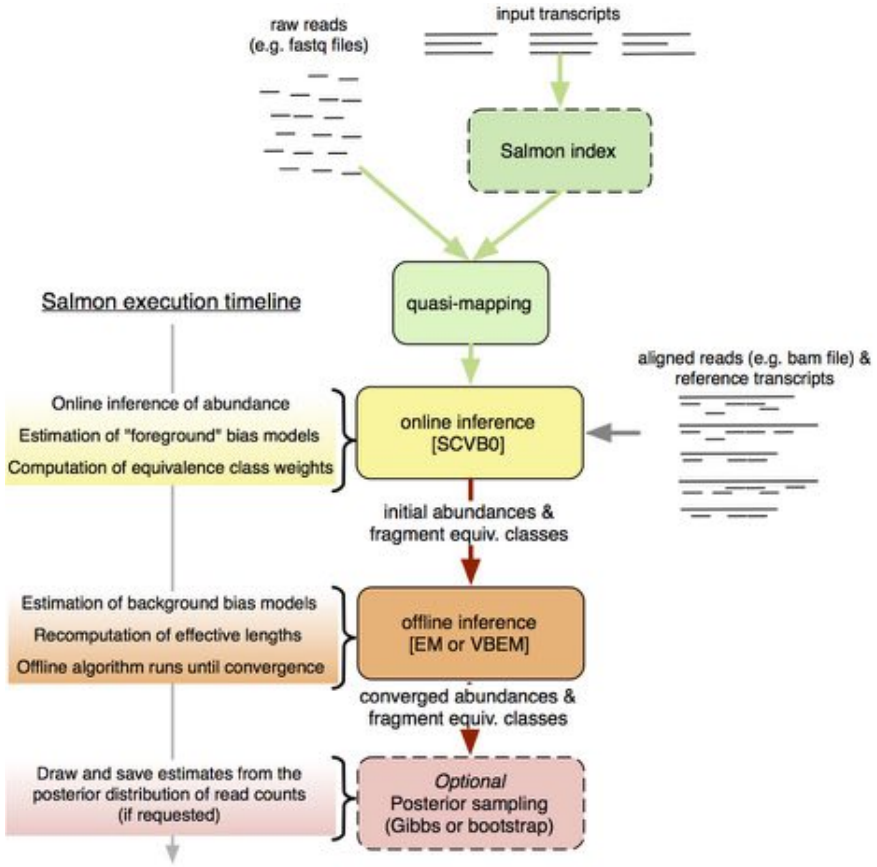
Solution: Replicates, replicates, and more replicates

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

Salmon: The ultimate RNA-seq Pipeline?







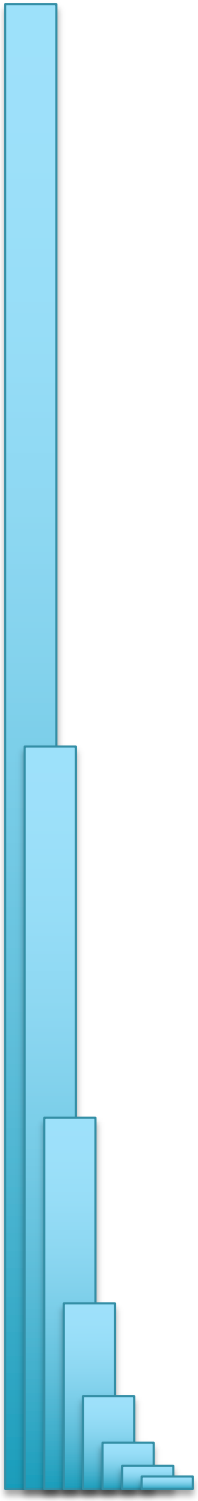
Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation
Love et al (2016) Nature Biotechnology 34, 1287–1291 (2016) doi:10.1038/nbt.3682

Salmon provides fast and bias-aware quantification of transcript expression
Patro et al (2017) Nature Methods (2017) doi:10.1038/nmeth.4197

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

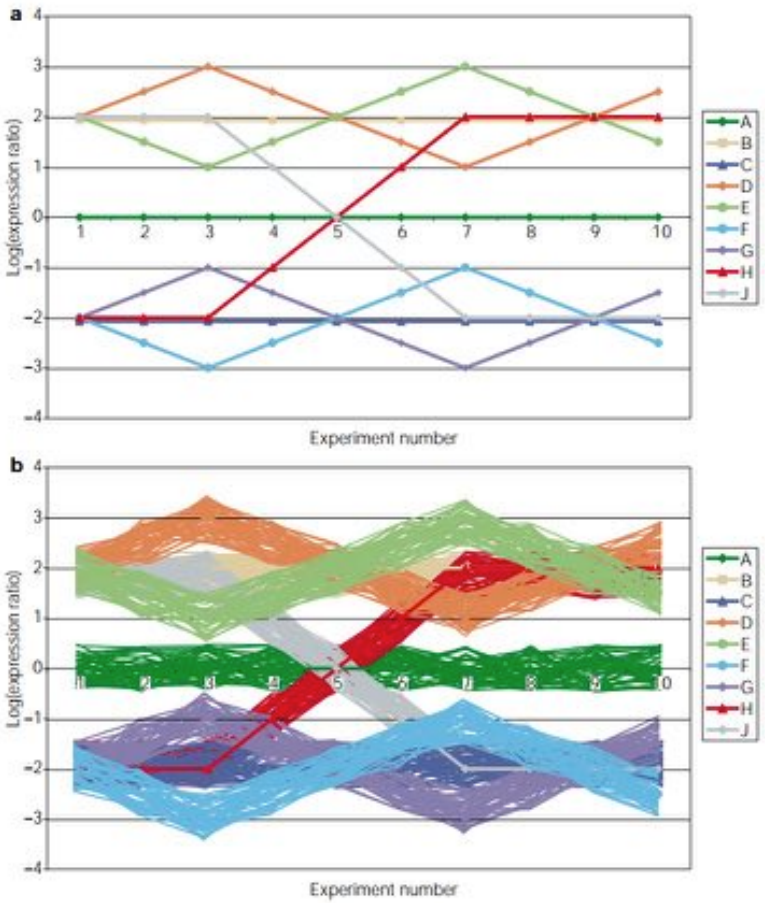
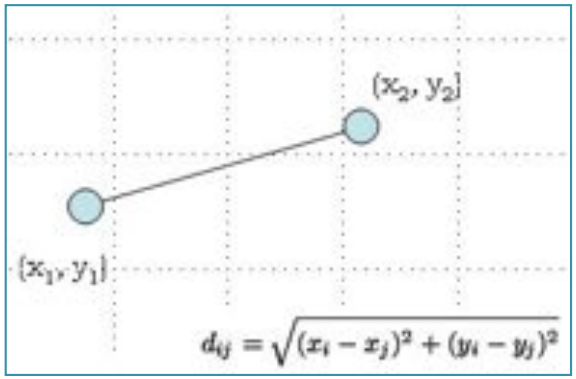
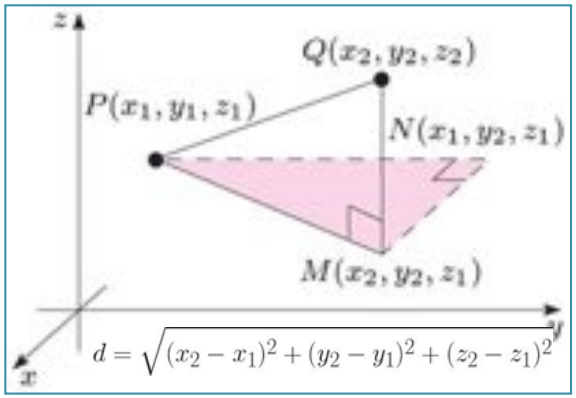


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments.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

Euclidean Distance

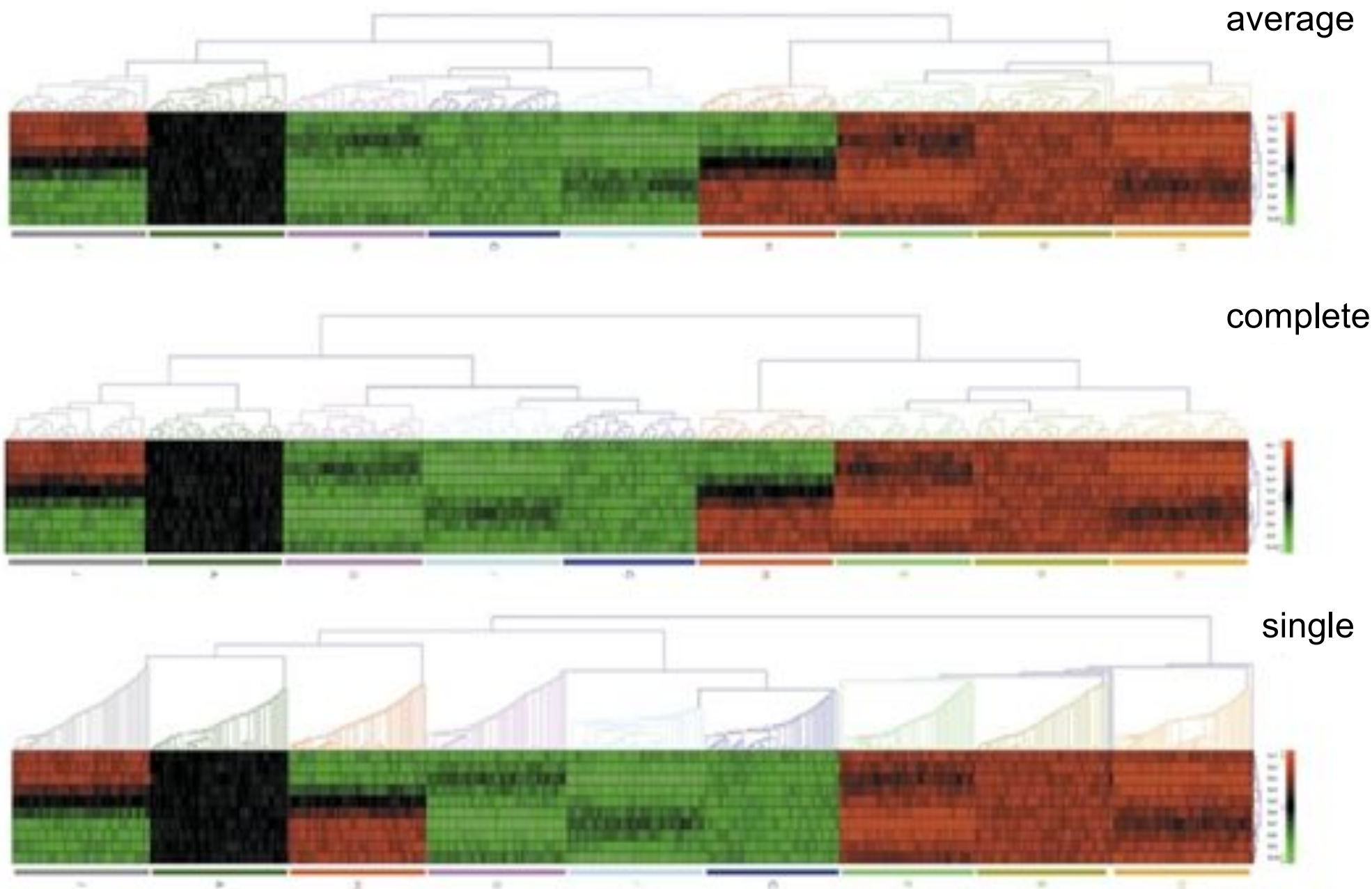




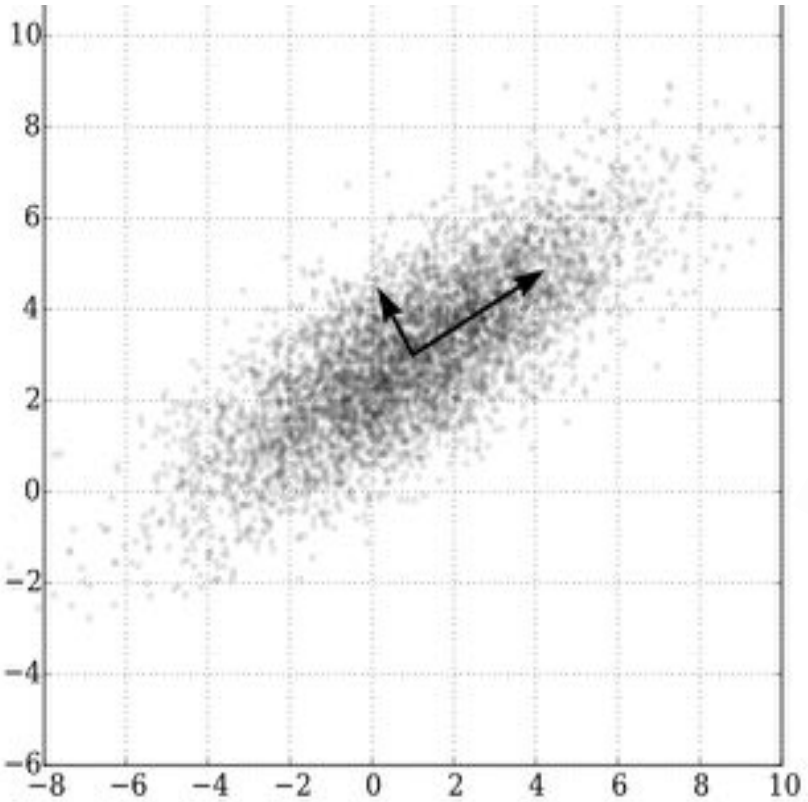
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

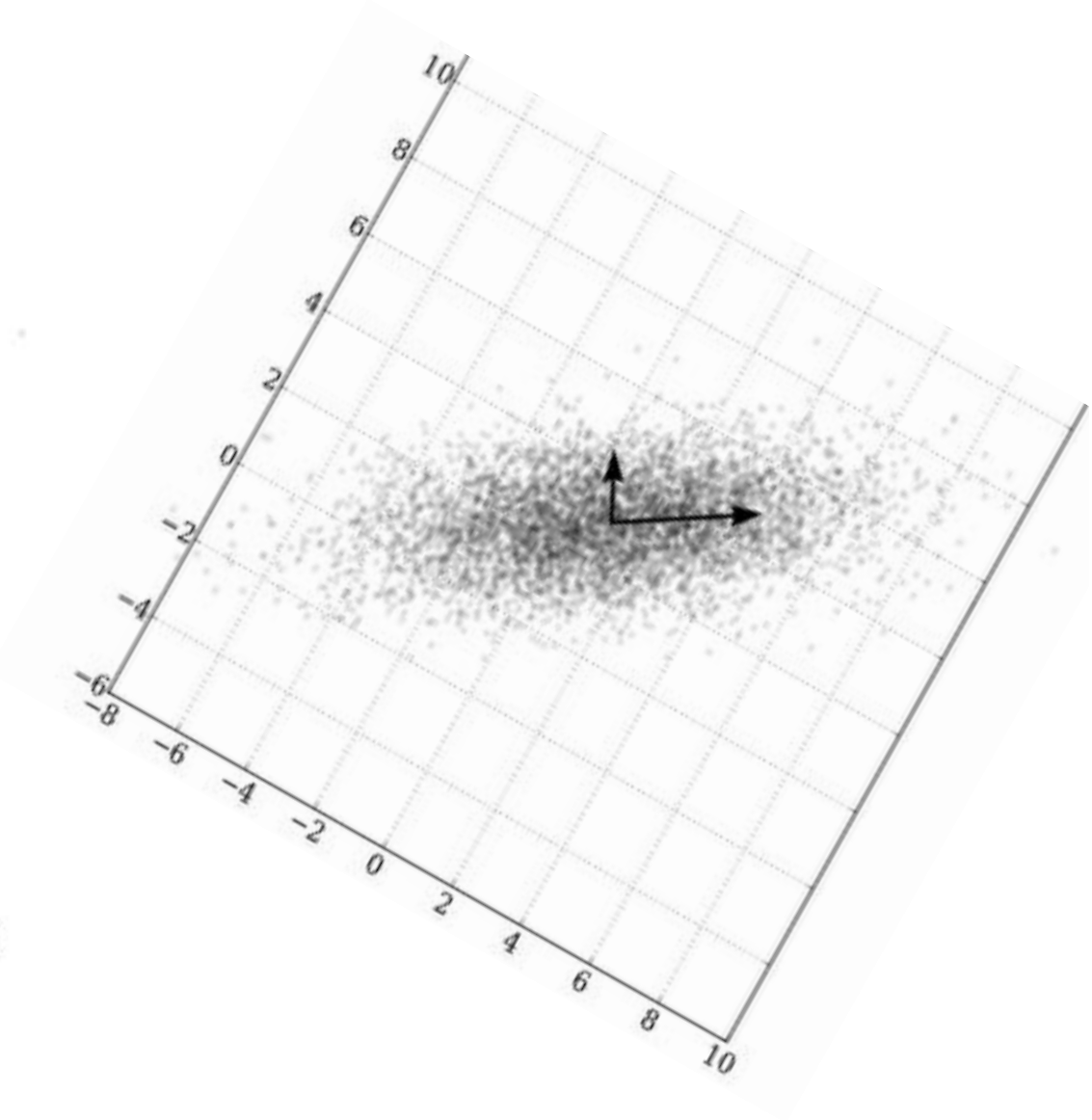
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



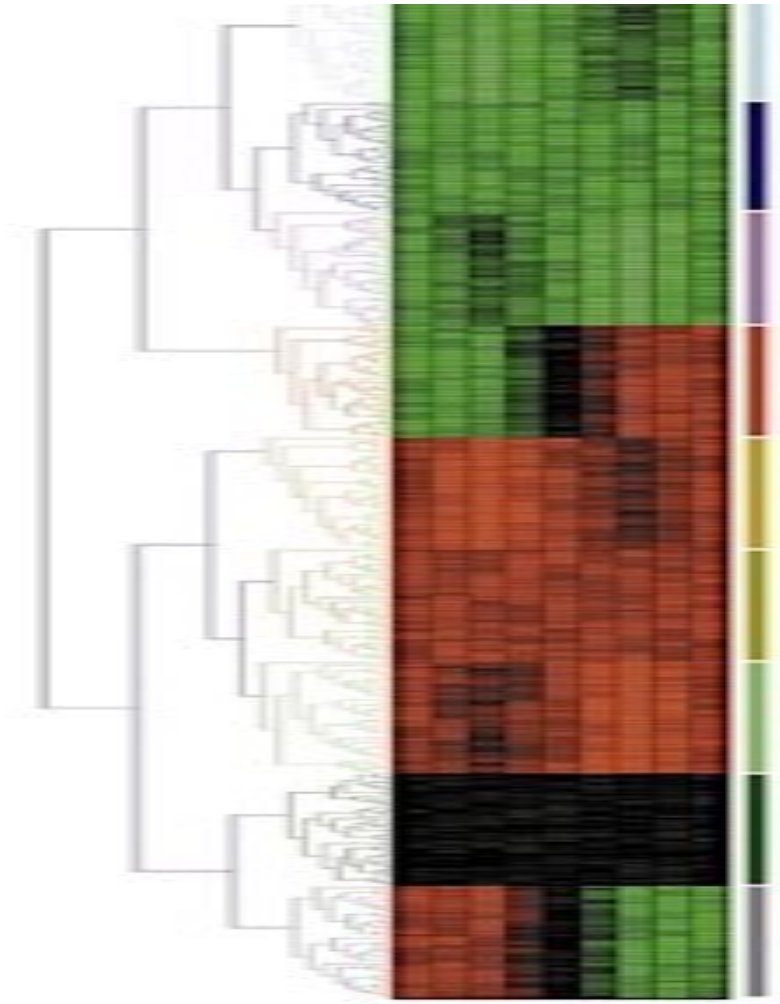
Principle Components Analysis (PCA)

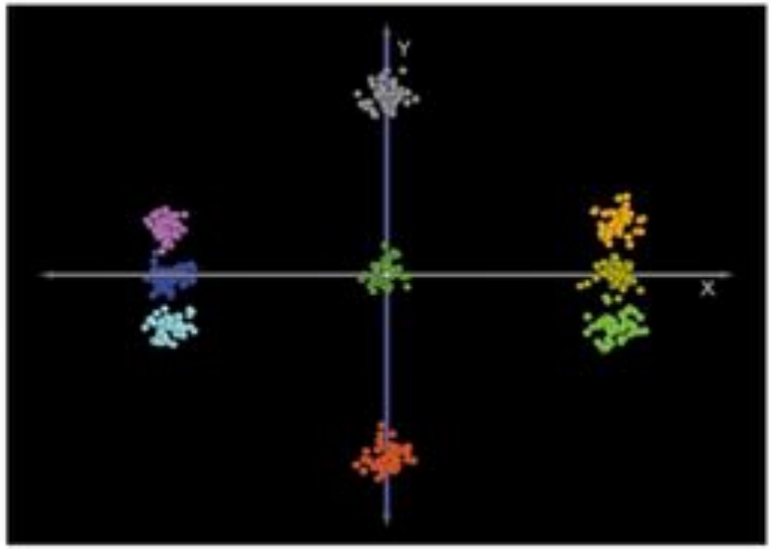




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





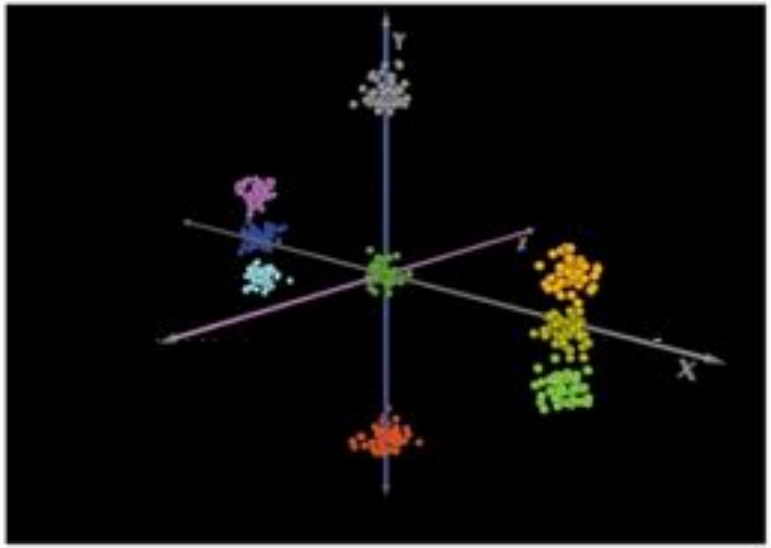
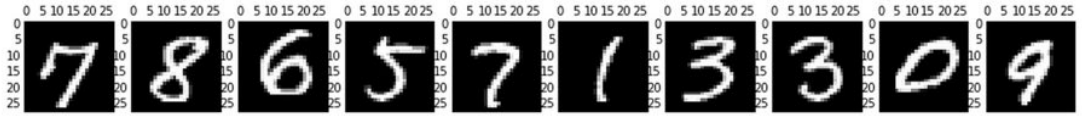
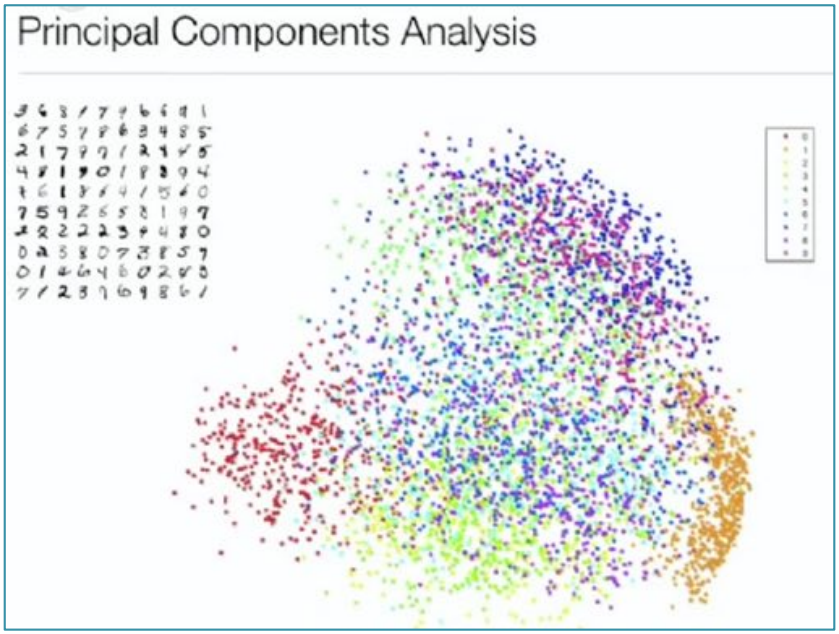
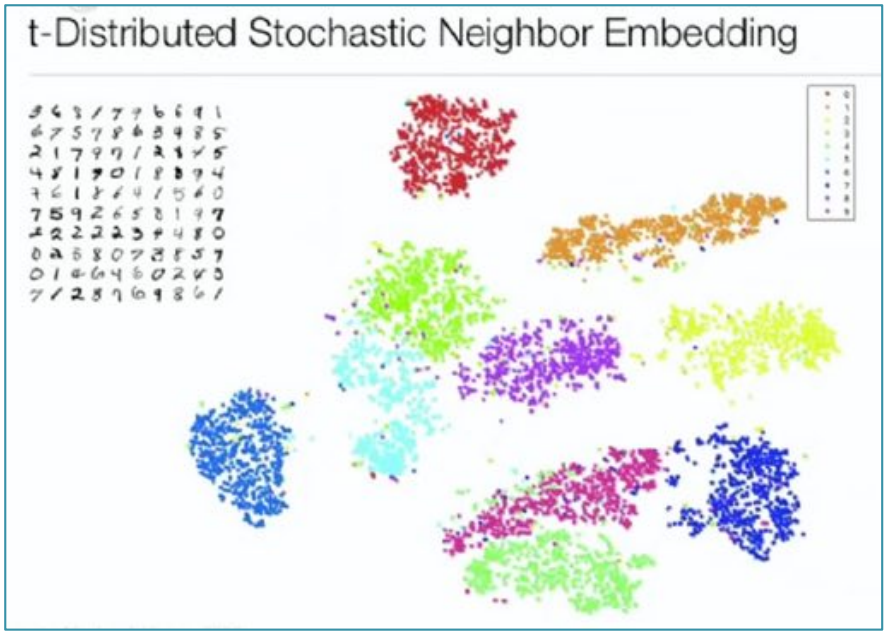


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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

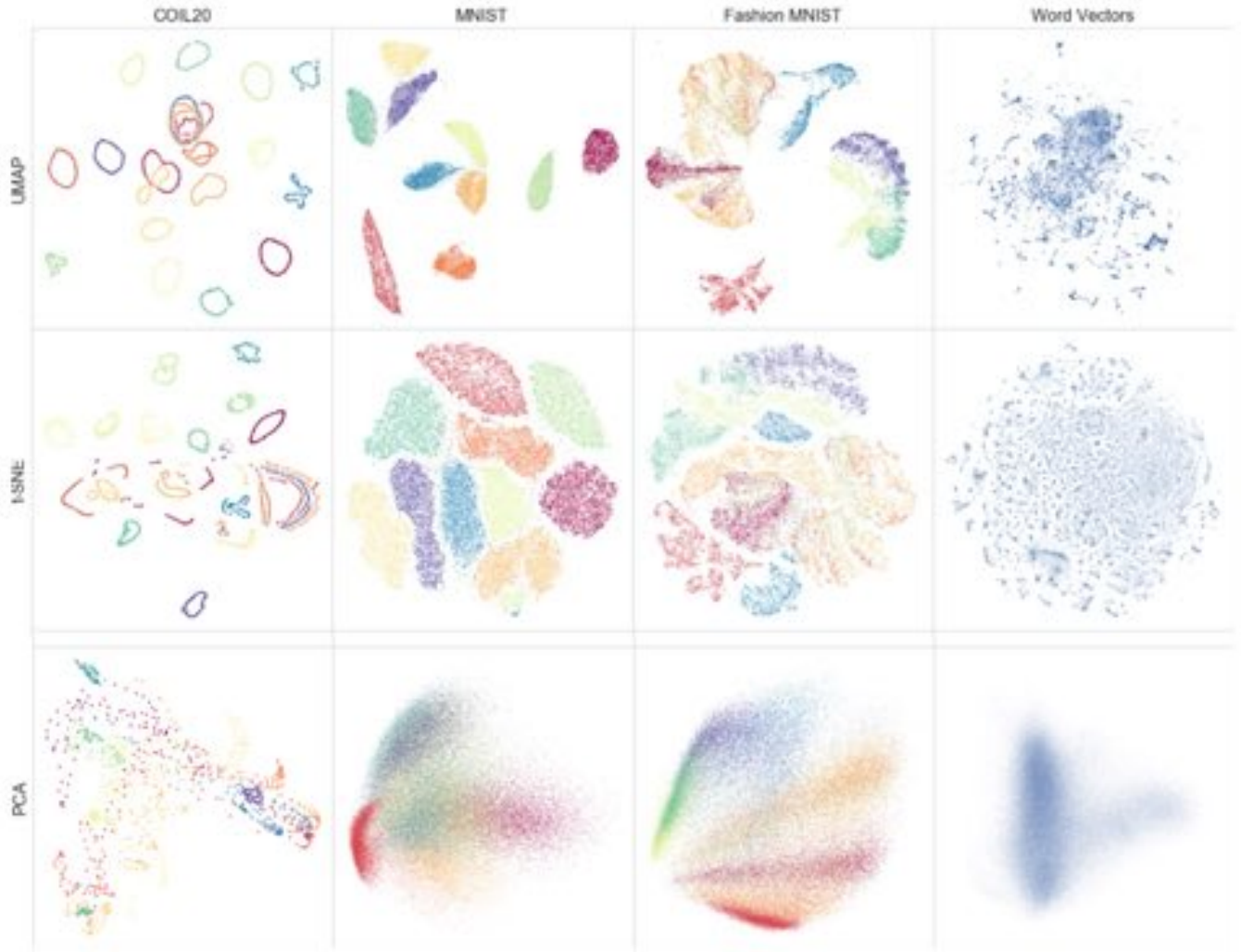
Visualizing Data Using t-SNE

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

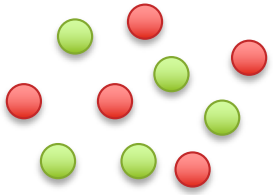
Single Cell Analysis

1. Why single cells?
2. scRNAseq

Population Heterogeneity

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

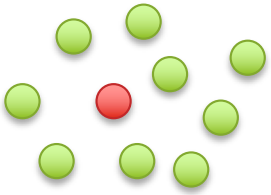
Experiment 1: 50/50



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

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

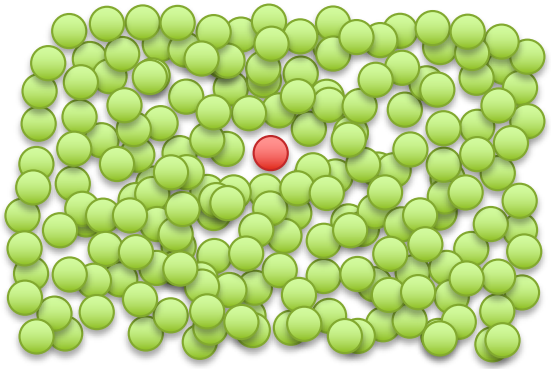
Experiment 2: 1/10



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

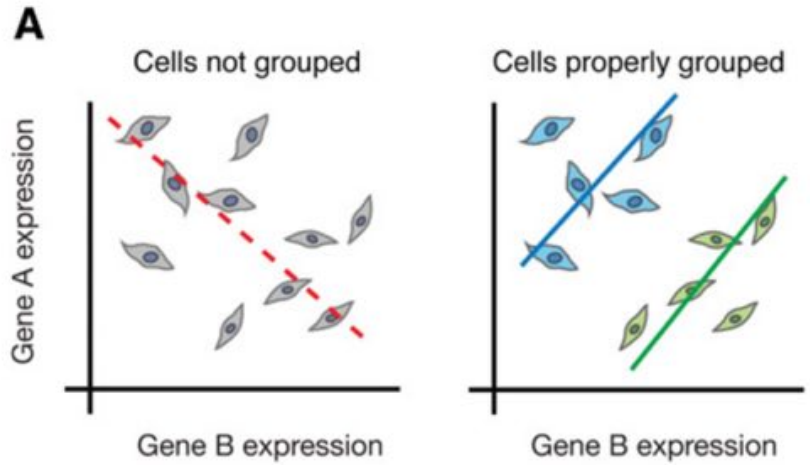
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

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



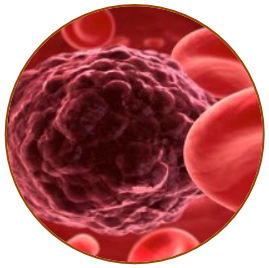
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

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)

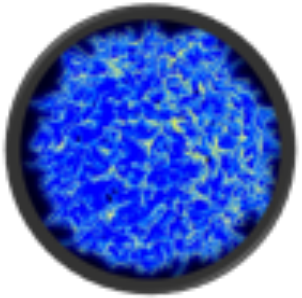
(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity



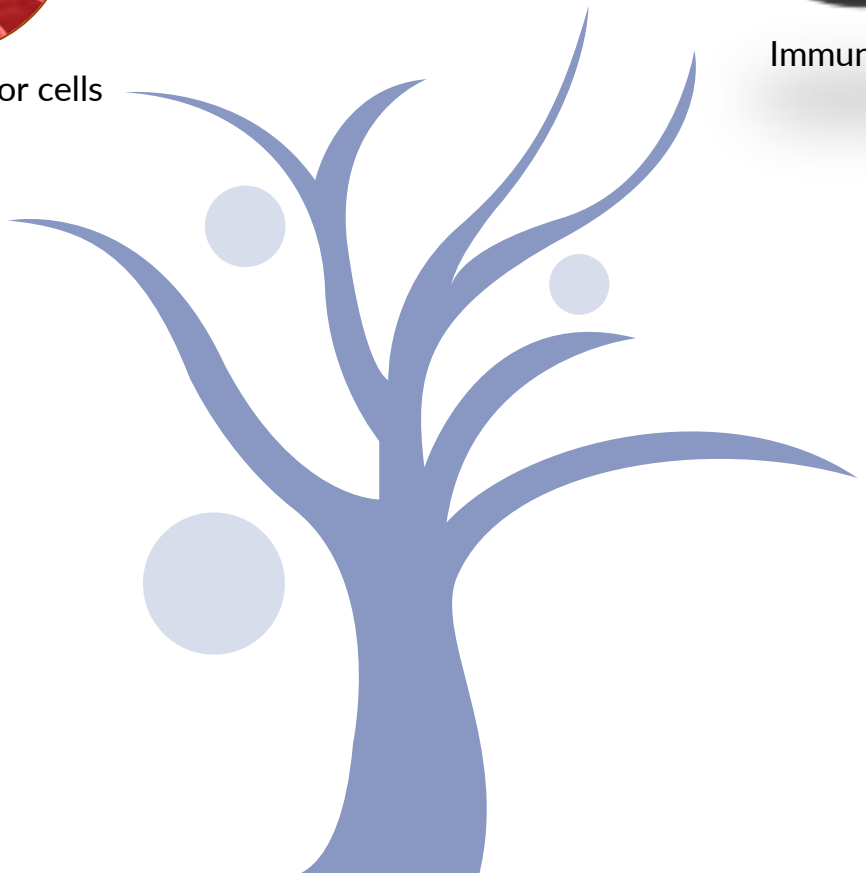
Circulating tumor cells

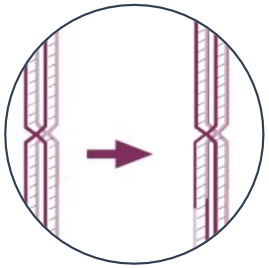
Neuronal mosaicism



Immune Cells

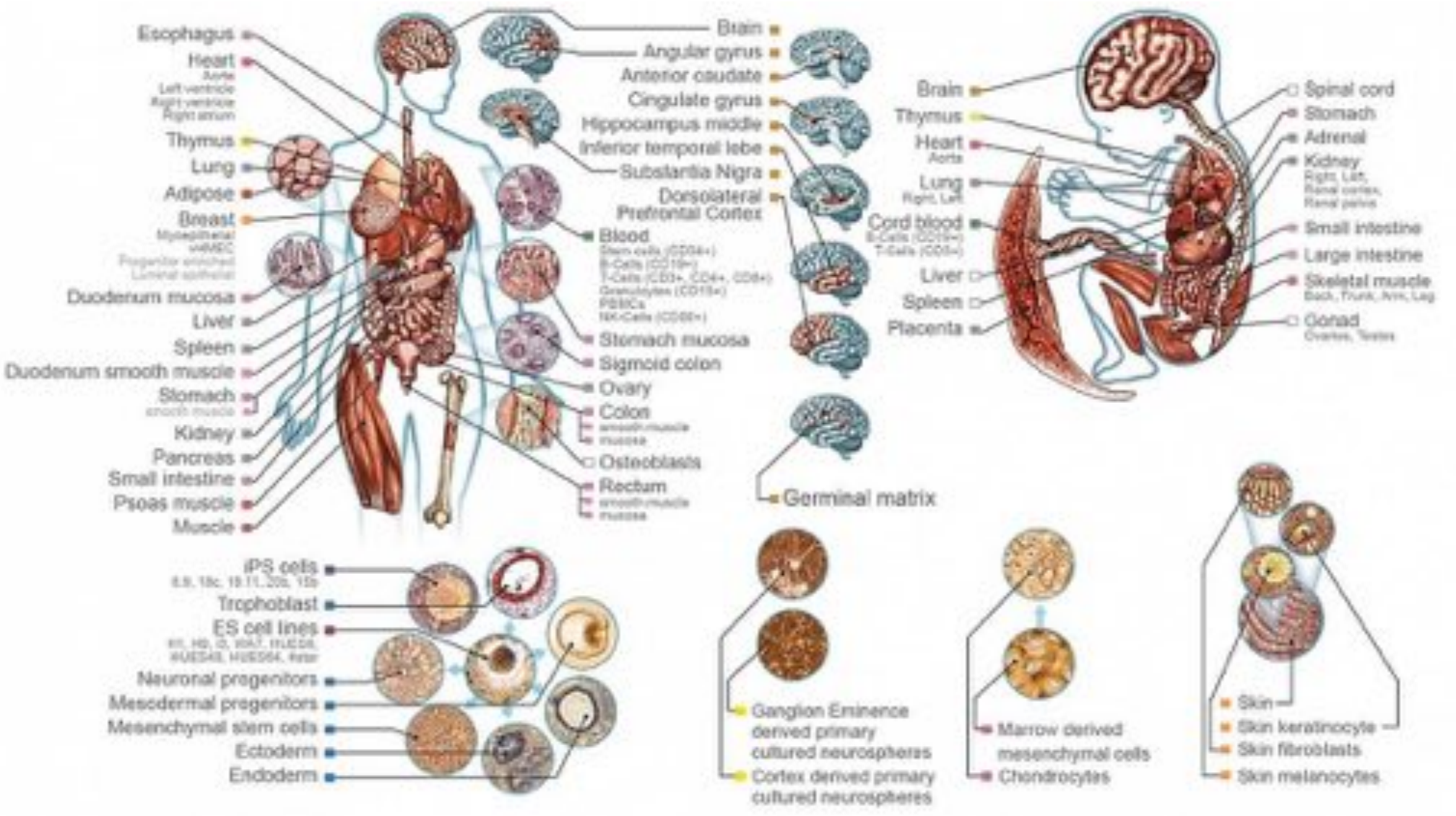
Tumor evolution



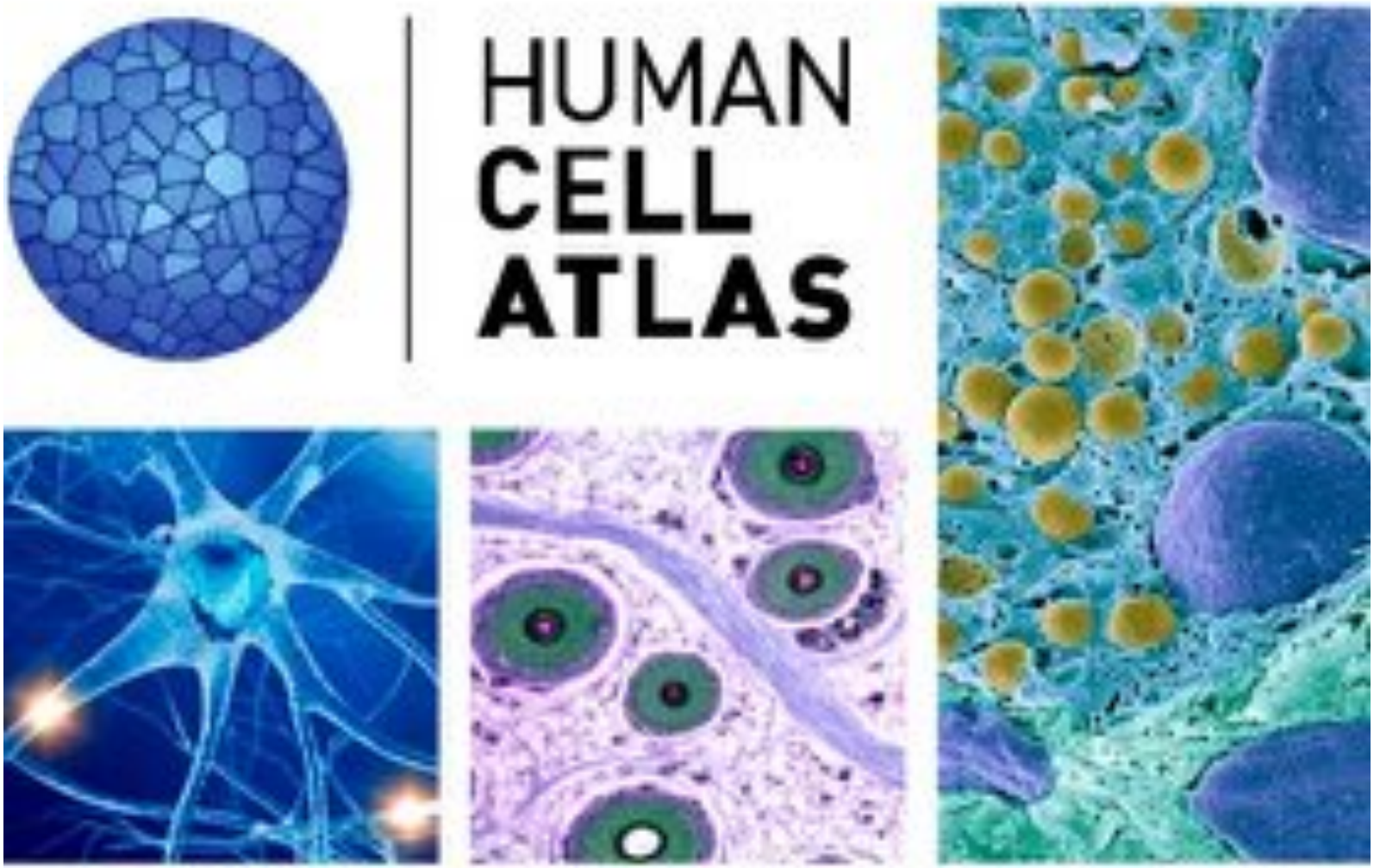


Recombination/crossover
in germ cells

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"

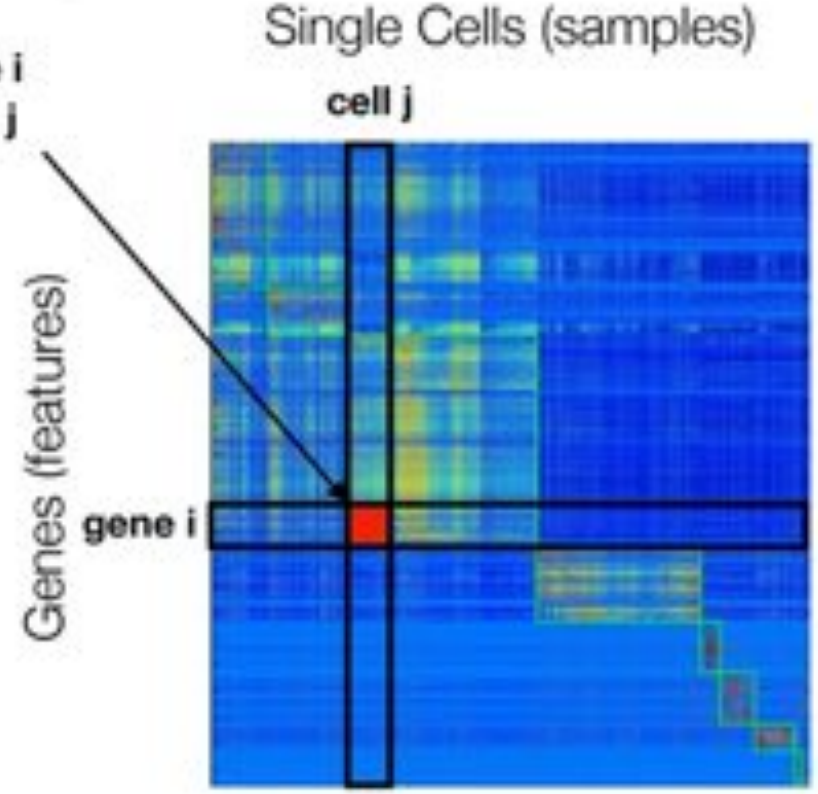
— a snapshot of the underlying biology in a data matrix.

Brain cells

Biological sample

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

sequencing

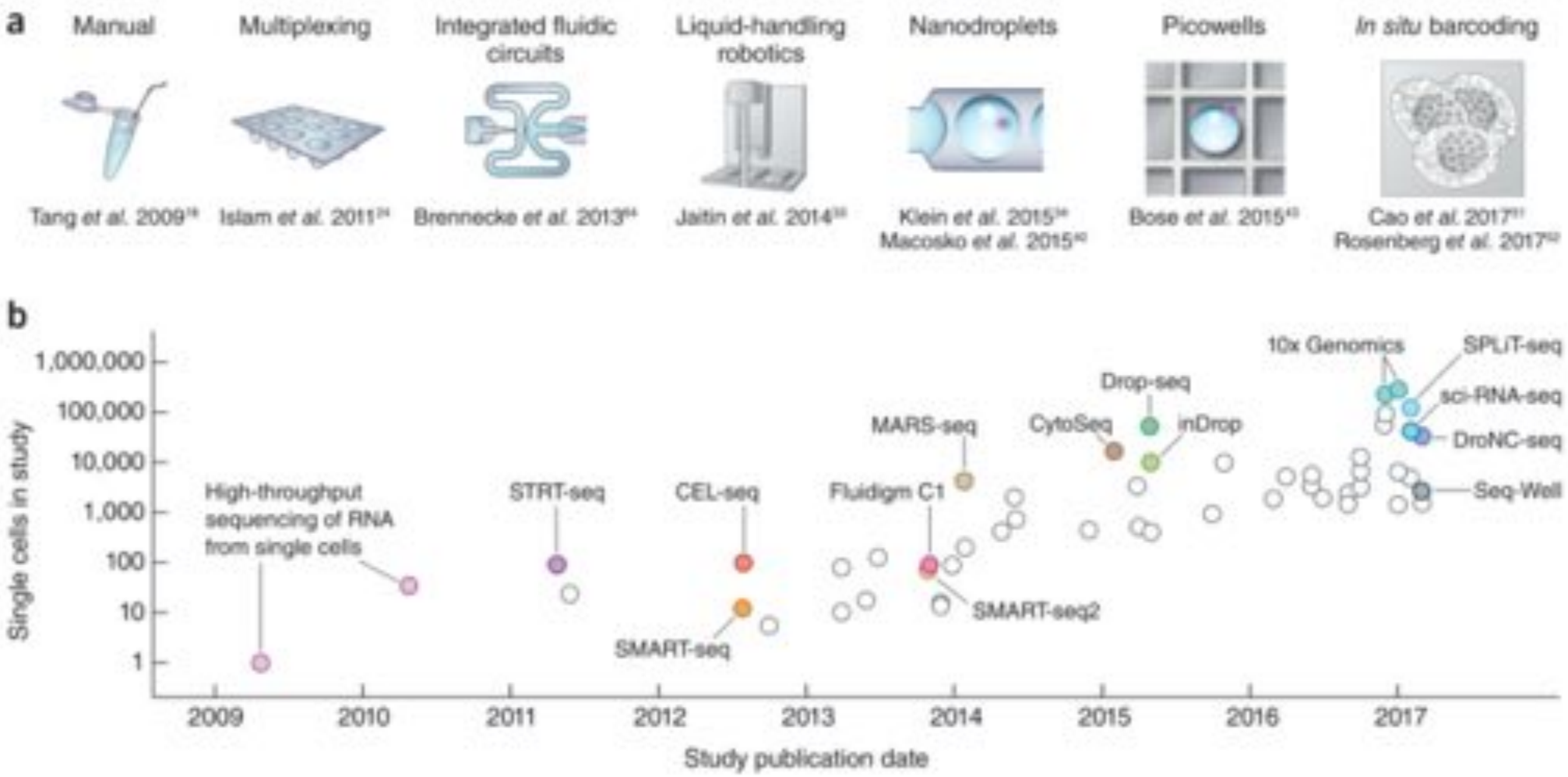


Gene expression matrix

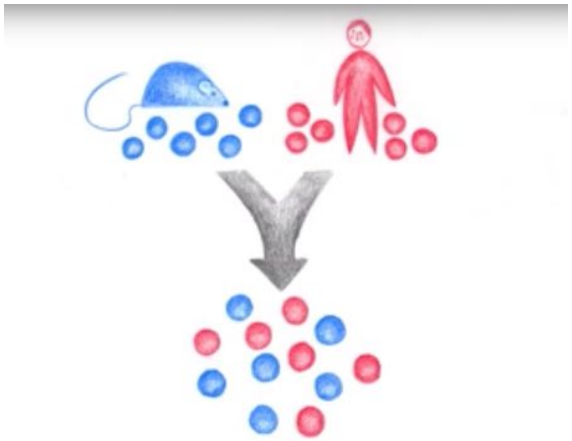
computationally explore complex biological systems

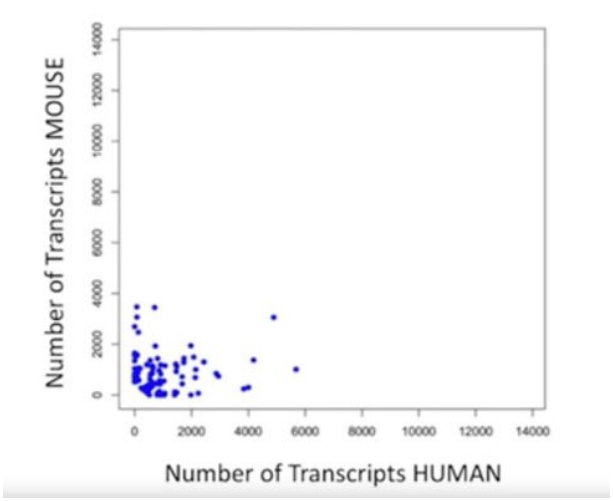
Martin Zhang

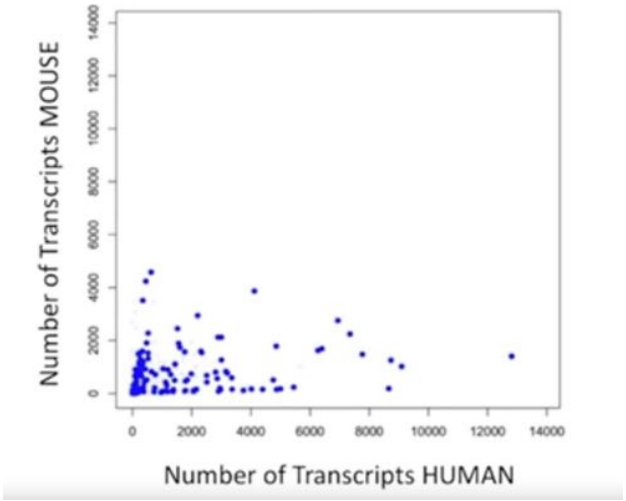
A decade of single-cell RNA-seq

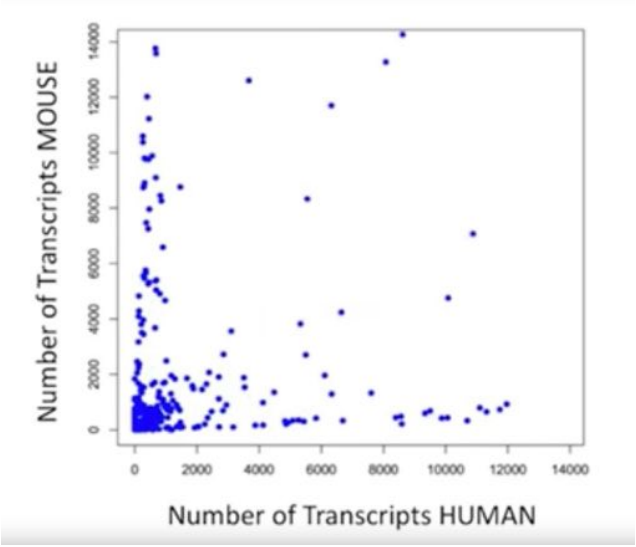


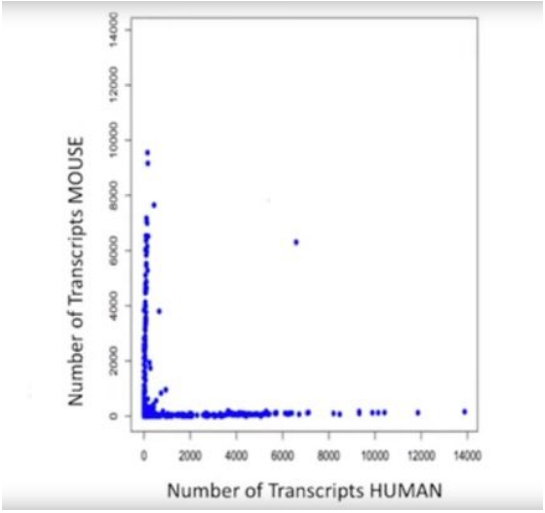
Svensson et al. 2018. DOI:10.1038/nprot.2017.149

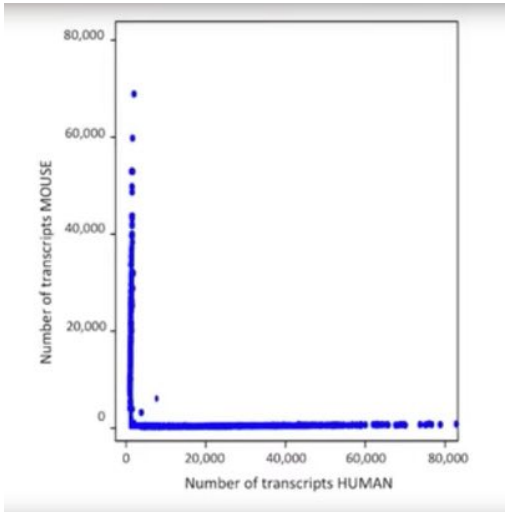






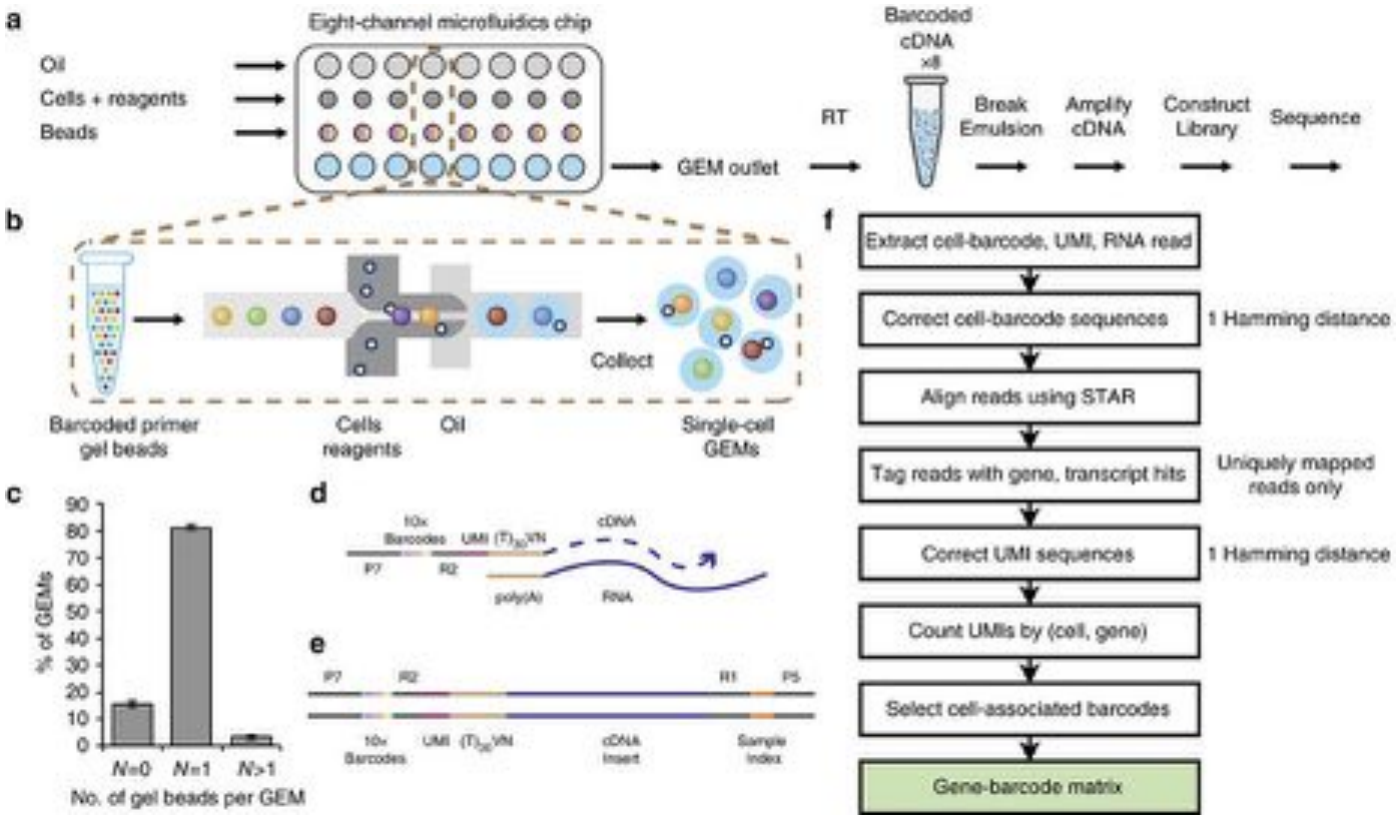






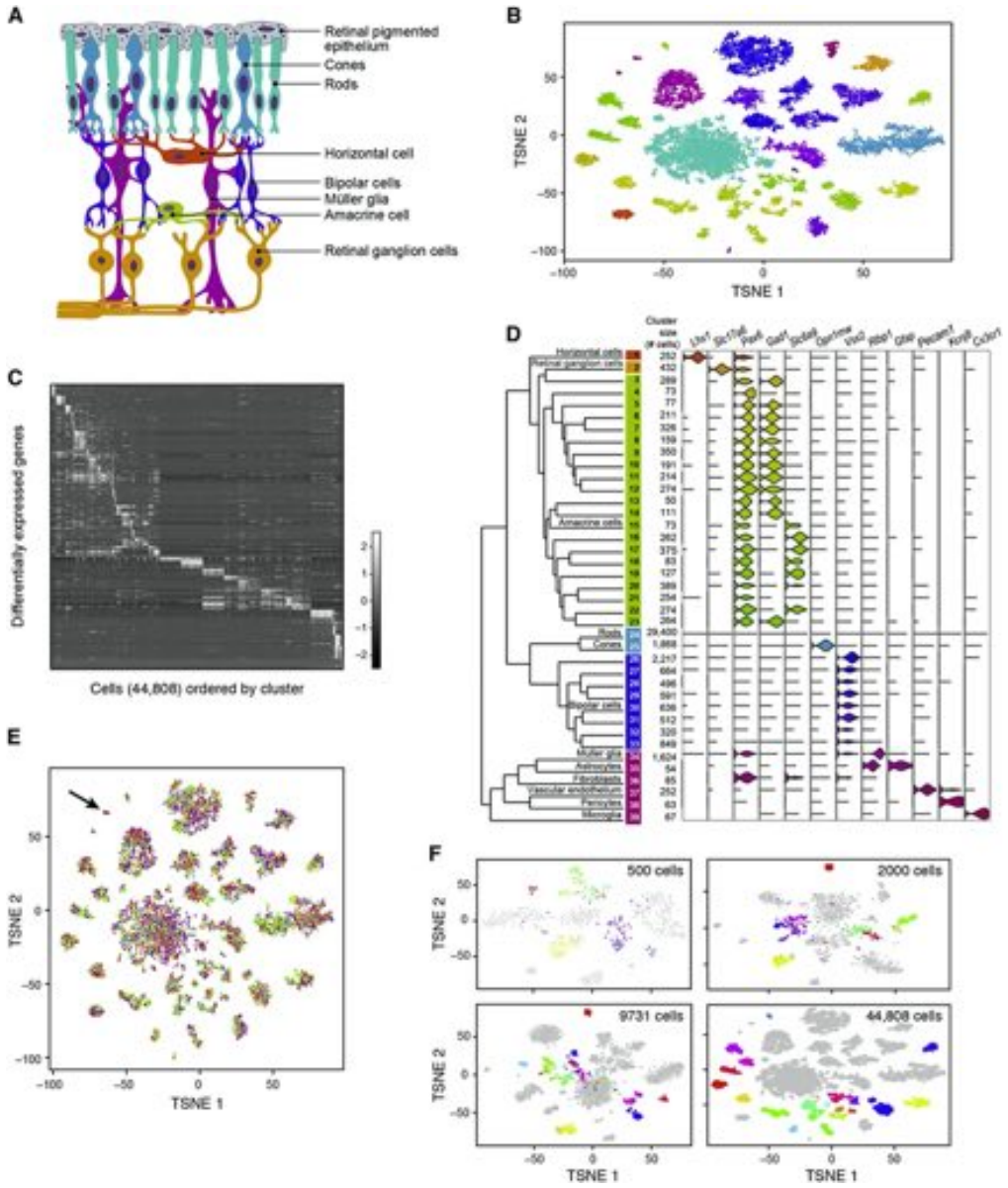
Drop-seq: Droplet barcoding of single cells
https://www.youtube.com/watch?v=vL7ptq2Dcf0

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

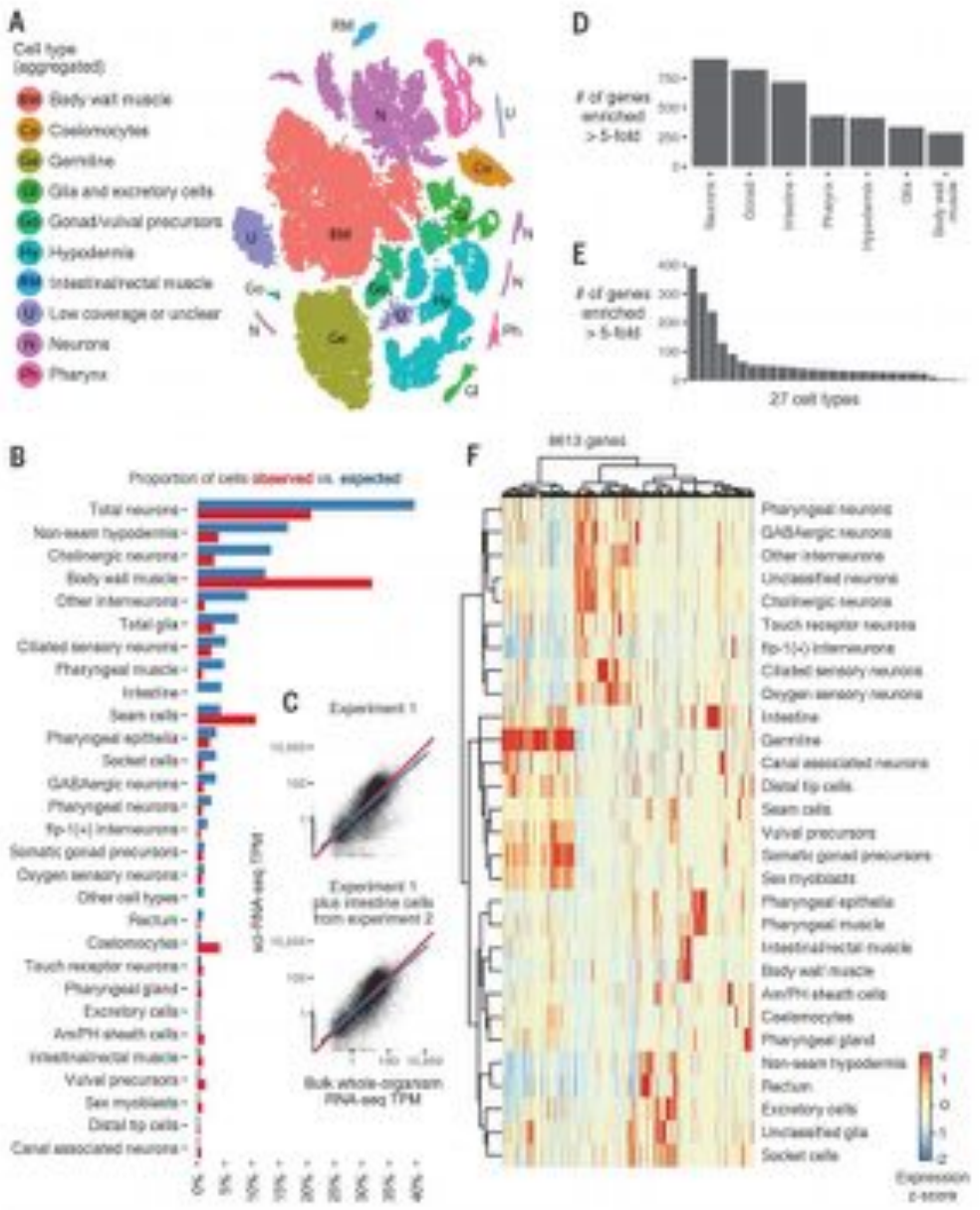
(c) Differentially expressed genes in each cluster

(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

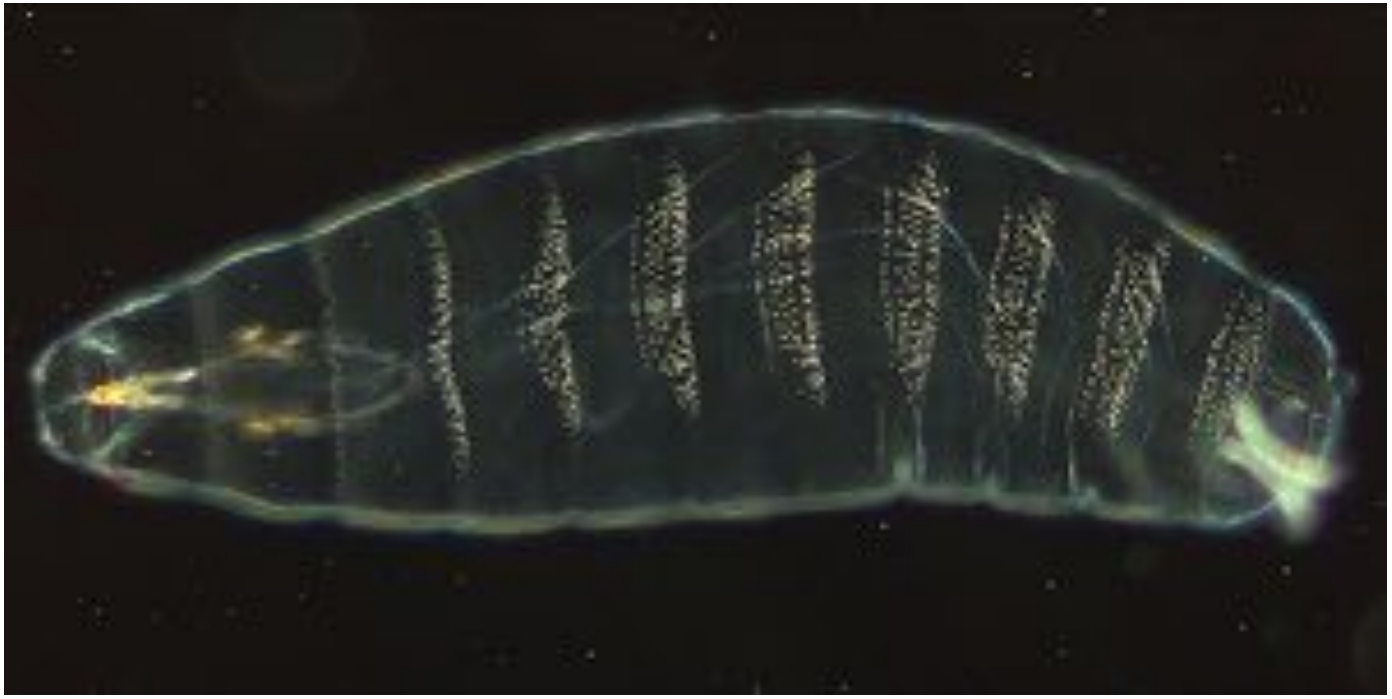
(a) t-SNE visualization of clusters

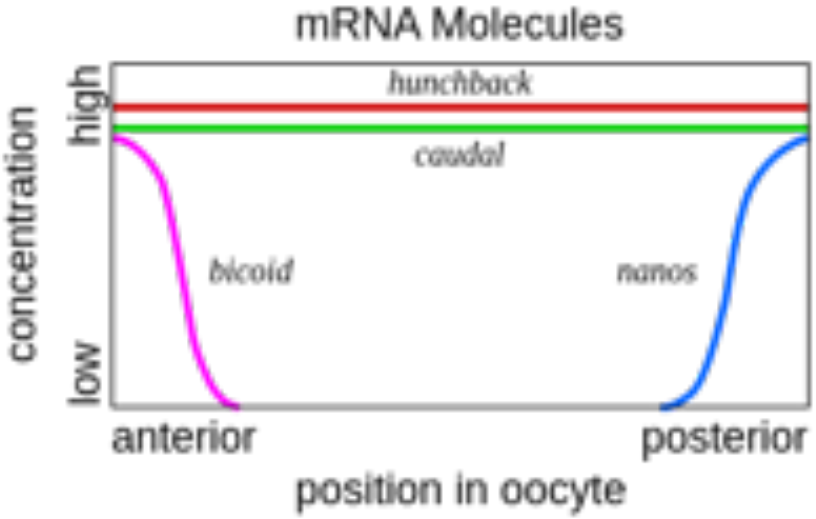
(b) Proportion of cells observed vs expected match well (including cells that only occur once or twice in the animal)

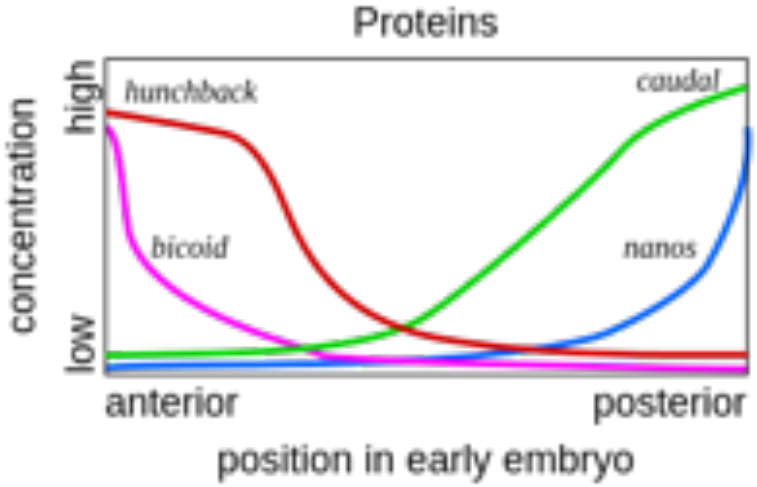
(c) Good correlation between single cell and bulk analysis of selected cell types

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

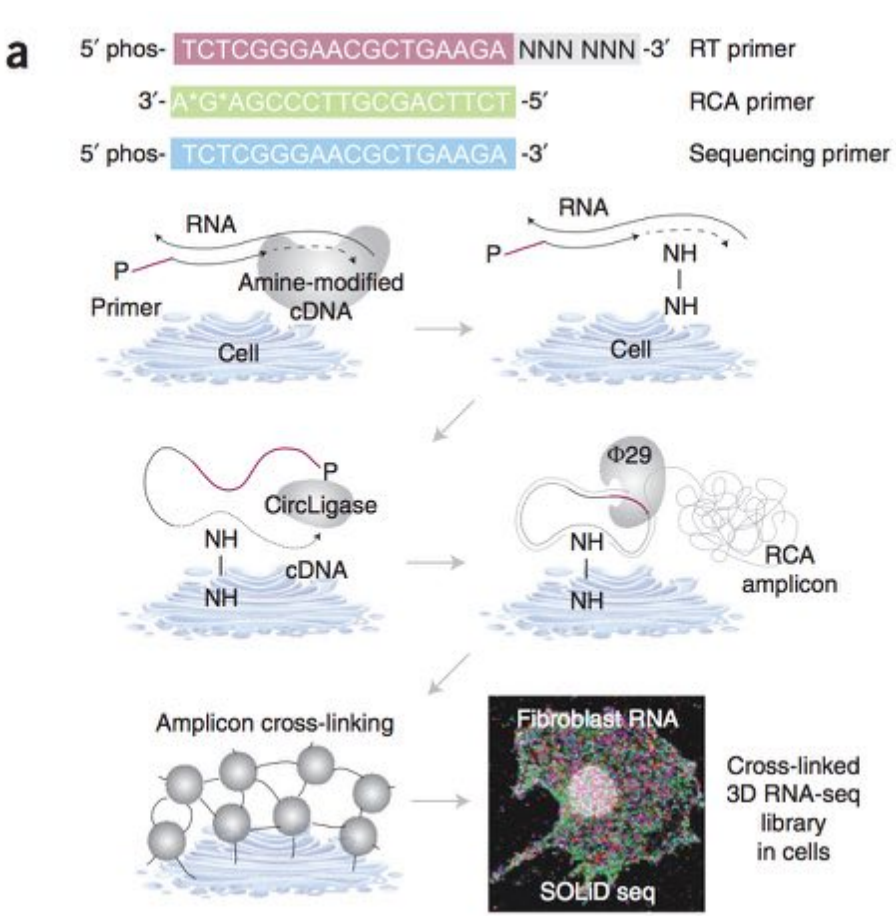
Comprehensive single-cell transcriptional profiling of a multicellular organism
Cao et al (2017) Science. 357:661-557

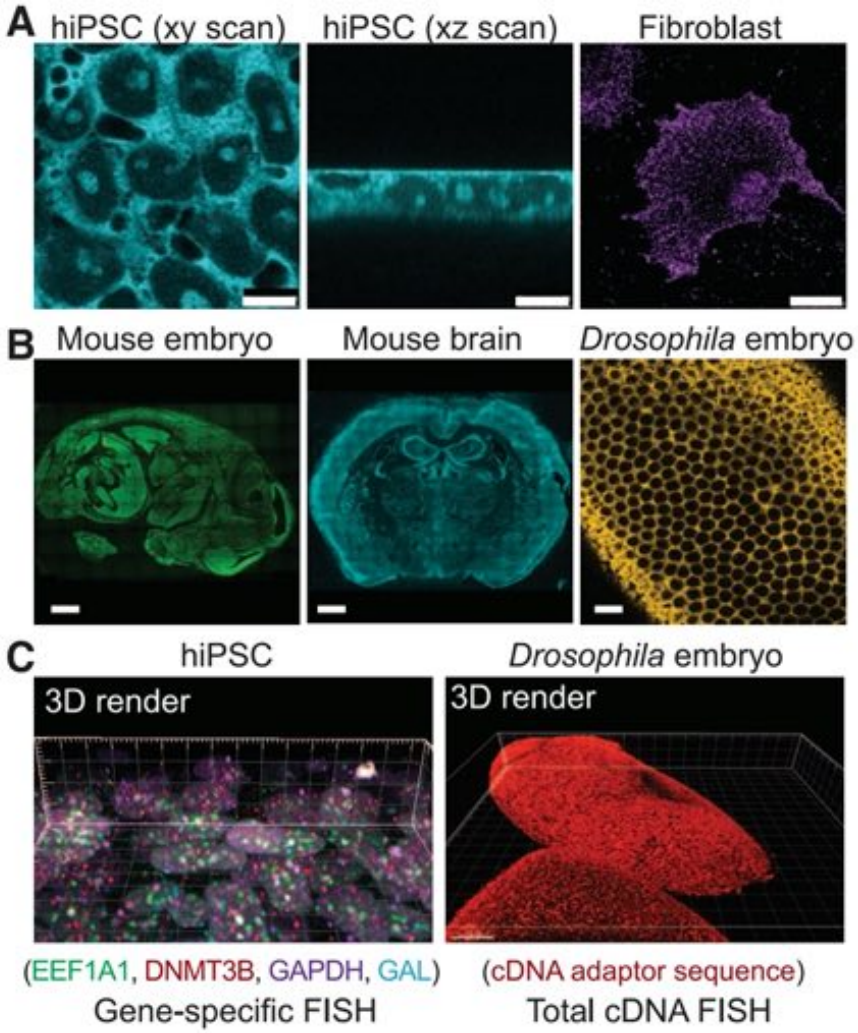






https://en.wikipedia.org/wiki/Drosophila_embryogenesis





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