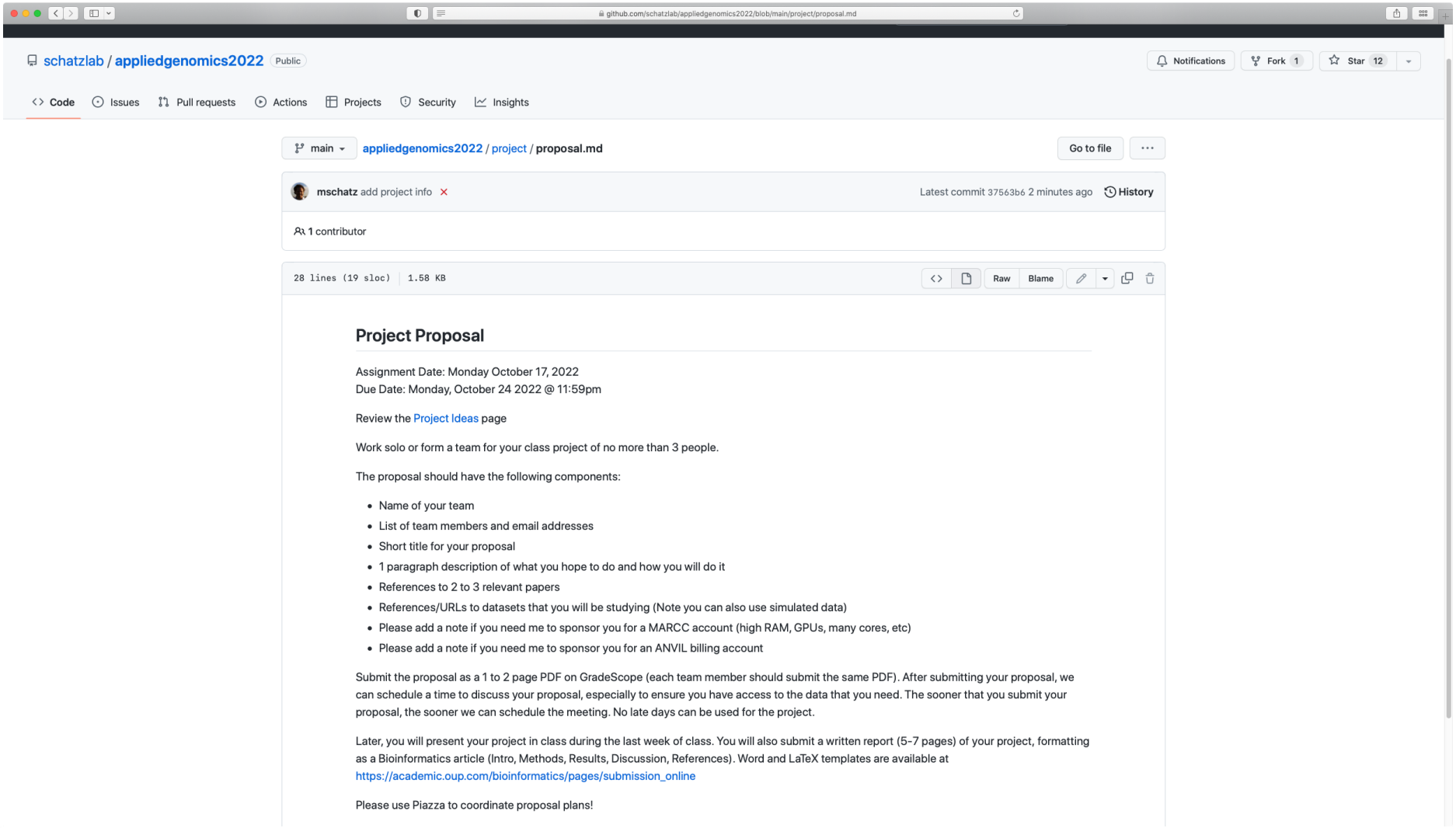
Lecture 16. scRNAseq and Gene Regulation

Michael Schatz

October 24, 2022

Applied Comparative Genomics

Project Proposal
Due Monday Oct 24 by 11:59pm



schatzlab / appliedgenomics2022 Public

Notifications Fork 1 Star 12

Code Issues Pull requests Actions Projects Security Insights

main appliedgenomics2022 / project / proposal.md Go to file

mschatz add project info Latest commit 37563b6 2 minutes ago History

1 contributor

28 lines (19 sloc) 1.58 KB Raw Blame

Project Proposal

Assignment Date: Monday October 17, 2022
Due Date: Monday, October 24 2022 @ 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)
- Please add a note if you need me to sponsor you for an ANVIL billing account

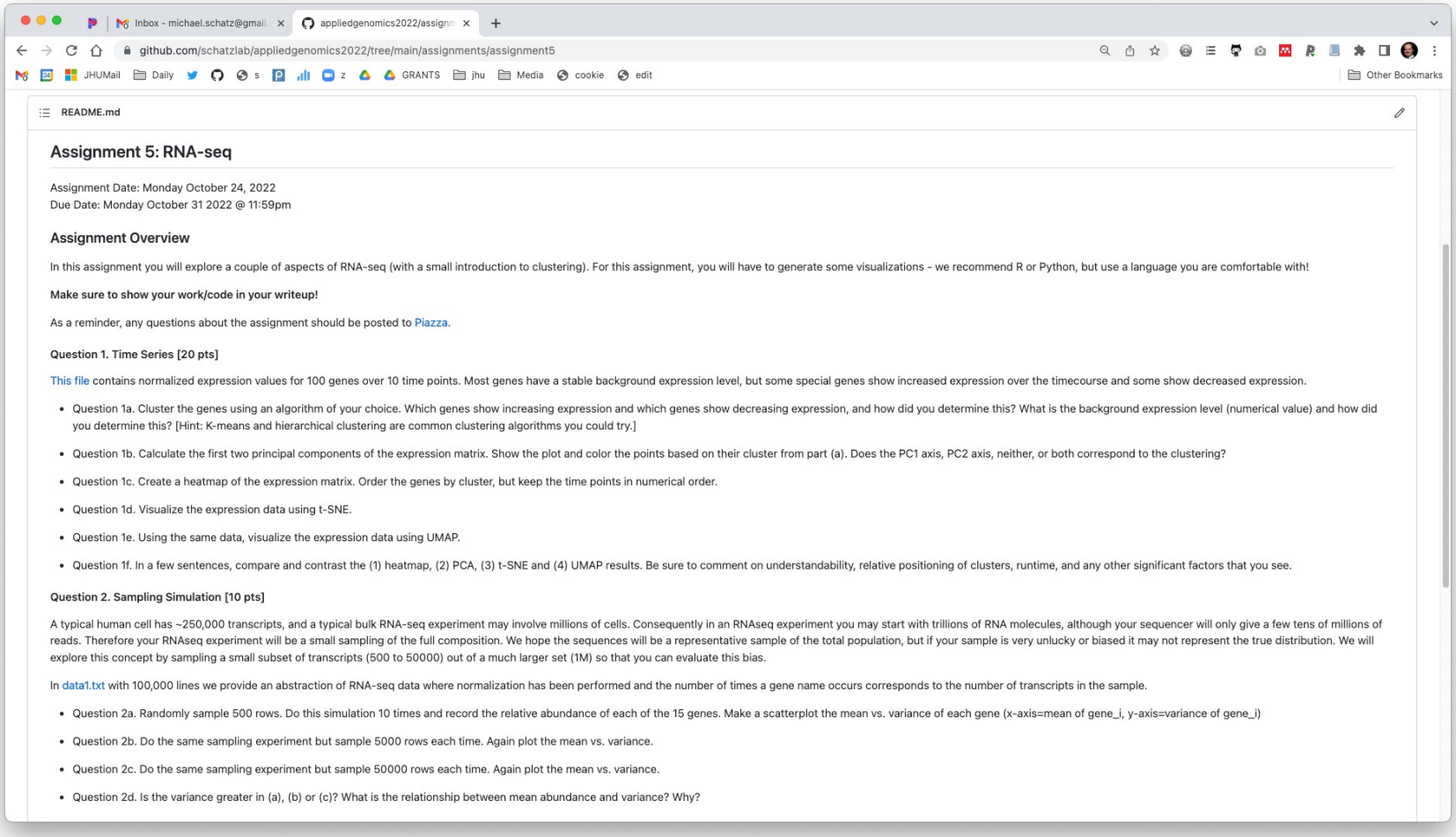
Submit the proposal as a 1 to 2 page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we 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/appliedgenomics2022/tree/main/project/proposal.md
Check Piazza for questions!

Assignment 5: Due Monday Oct 31



README.md

Assignment 5: RNA-seq

Assignment Date: Monday October 24, 2022
Due Date: Monday October 31 2022 @ 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, any questions about the assignment should be posted 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 timecourse and some show decreased expression.

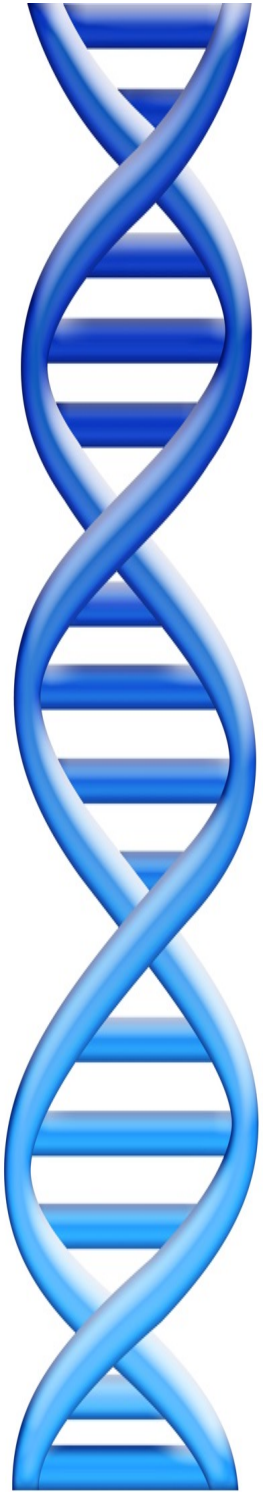
- Question 1a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical 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.

Question 2. Sampling Simulation [10 pts]

A typical human cell has ~250,000 transcripts, and a typical bulk RNA-seq experiment may involve millions of cells. Consequently in an RNAseq experiment you may start with trillions of RNA molecules, although your sequencer will only give a few tens of millions of reads. Therefore your RNAseq experiment will be a small sampling of the full composition. We hope the sequences will be a representative sample of the total population, but if your sample is very unlucky or biased it may not represent the true distribution. We will explore this concept by sampling a small subset of transcripts (500 to 50000) out of a much larger set (1M) so that you can evaluate this bias.

In data1.txt with 100,000 lines we provide an abstraction of RNA-seq data where normalization has been performed and the number of times a gene name occurs corresponds to the number of transcripts in the sample.

- Question 2a. Randomly sample 500 rows. Do this simulation 10 times and record the relative abundance of each of the 15 genes. Make a scatterplot the mean vs. variance of each gene (x-axis=mean of gene_i, y-axis=variance of gene_i)
- Question 2b. Do the same sampling experiment but sample 5000 rows each time. Again plot the mean vs. variance.
- Question 2c. Do the same sampling experiment but sample 50000 rows each time. Again plot the mean vs. variance.
- Question 2d. Is the variance greater in (a), (b) or (c)? What is the relationship between mean abundance and variance? Why?

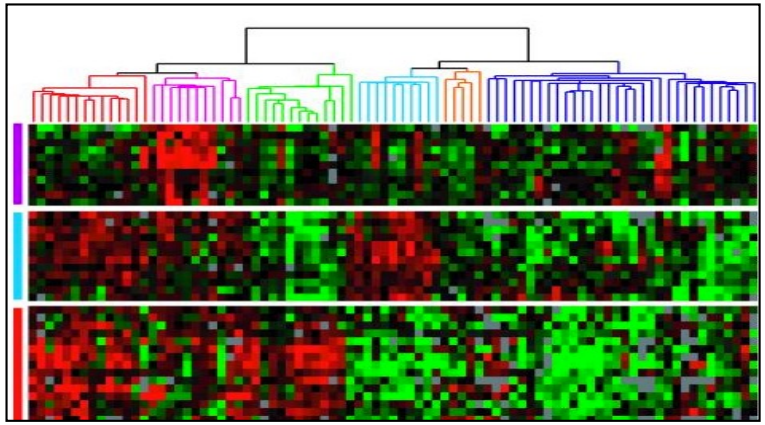


Outline

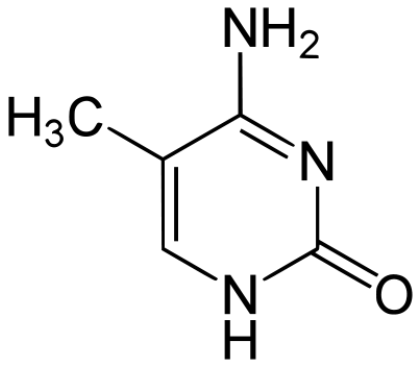
1. Bulk RNA-seq mini-refresher
2. Why single cells?
3. Gene Regulation
    1. Methylation
    2. Transcription factors
    3. Chromatin, enhancers, insulators, TADs

*-seq in 4 short vignettes

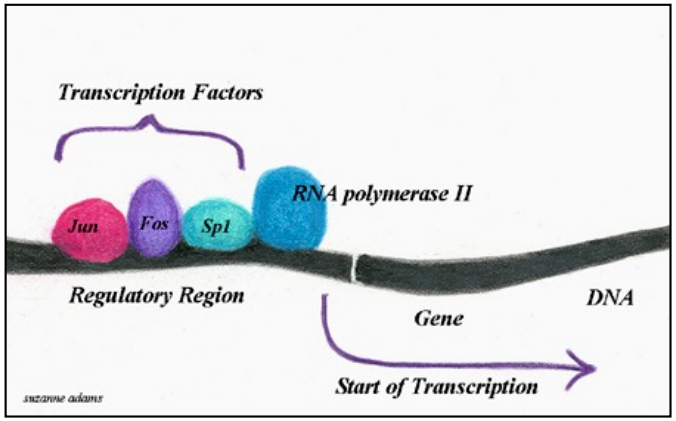
RNA-seq



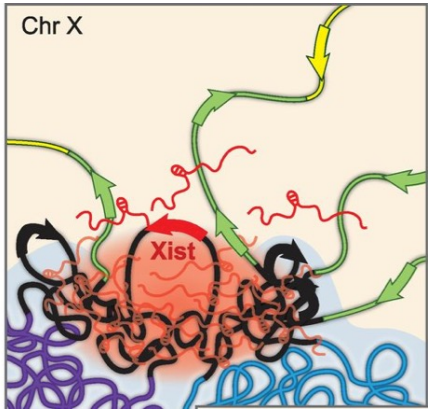
Methyl-seq



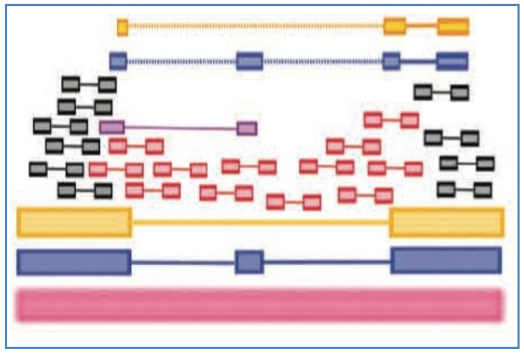
ChIP-seq



Hi-C



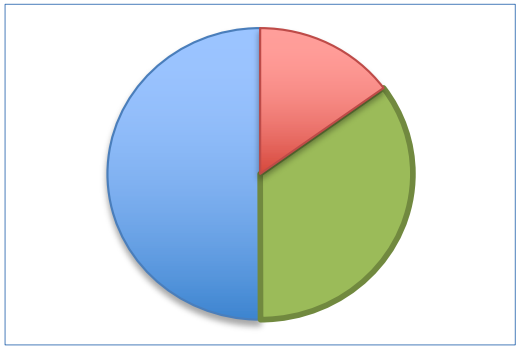
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

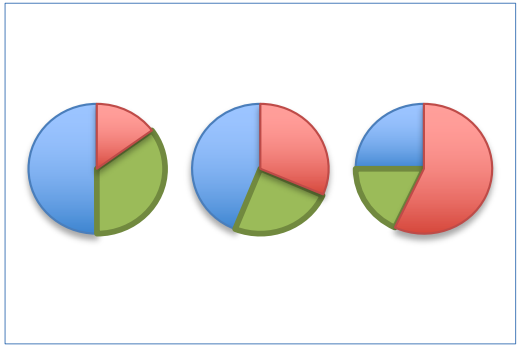
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



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

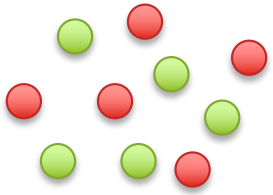
Solution: Replicates, replicates, and more replicates

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

Population Heterogeneity

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

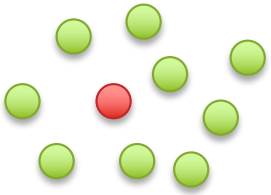
Experiment 1: 50/50



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

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

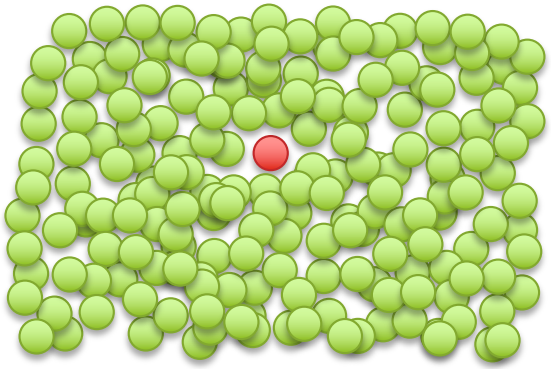
Experiment 2: 1/10



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

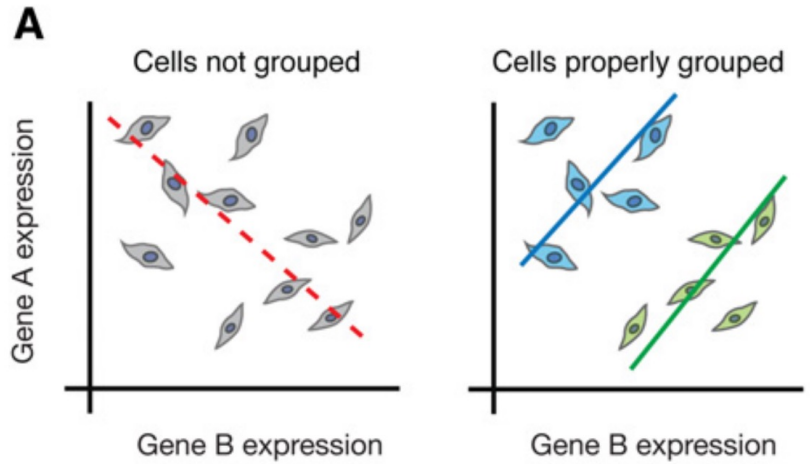
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

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



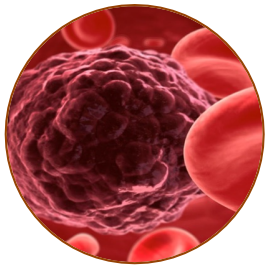
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

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)

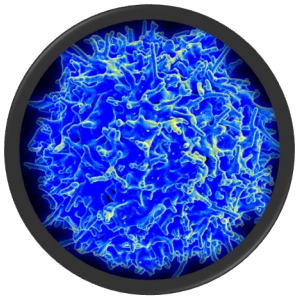
(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity

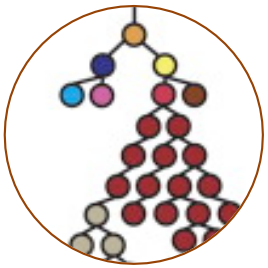


Circulating tumor cells

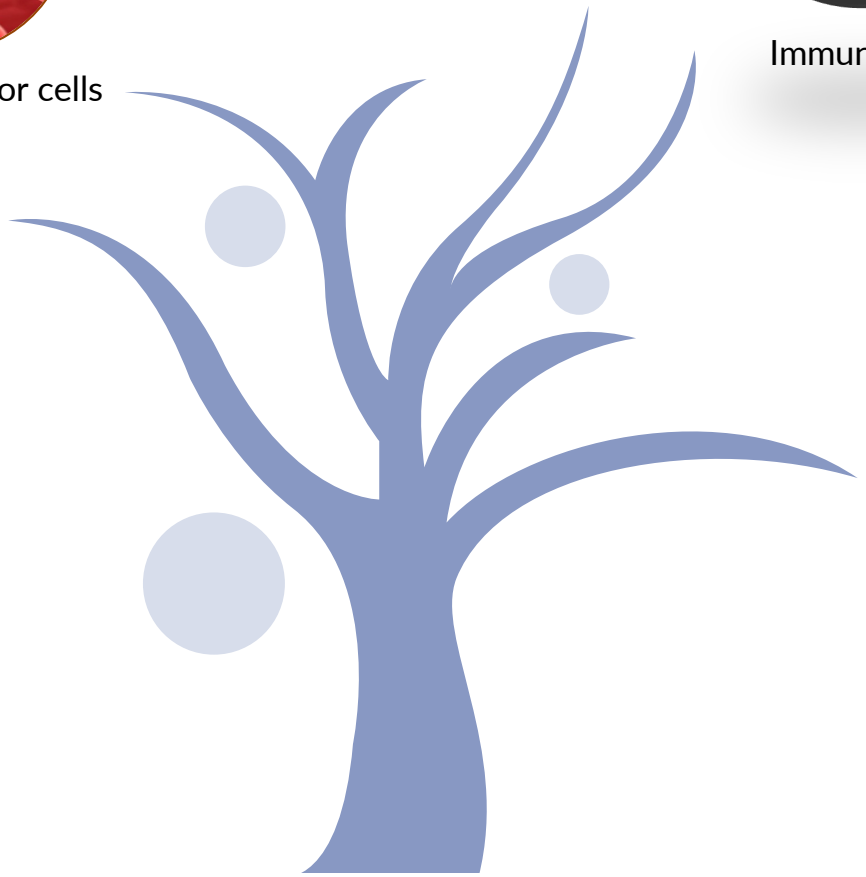
Neuronal mosaicism

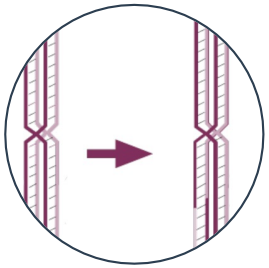


Immune Cells



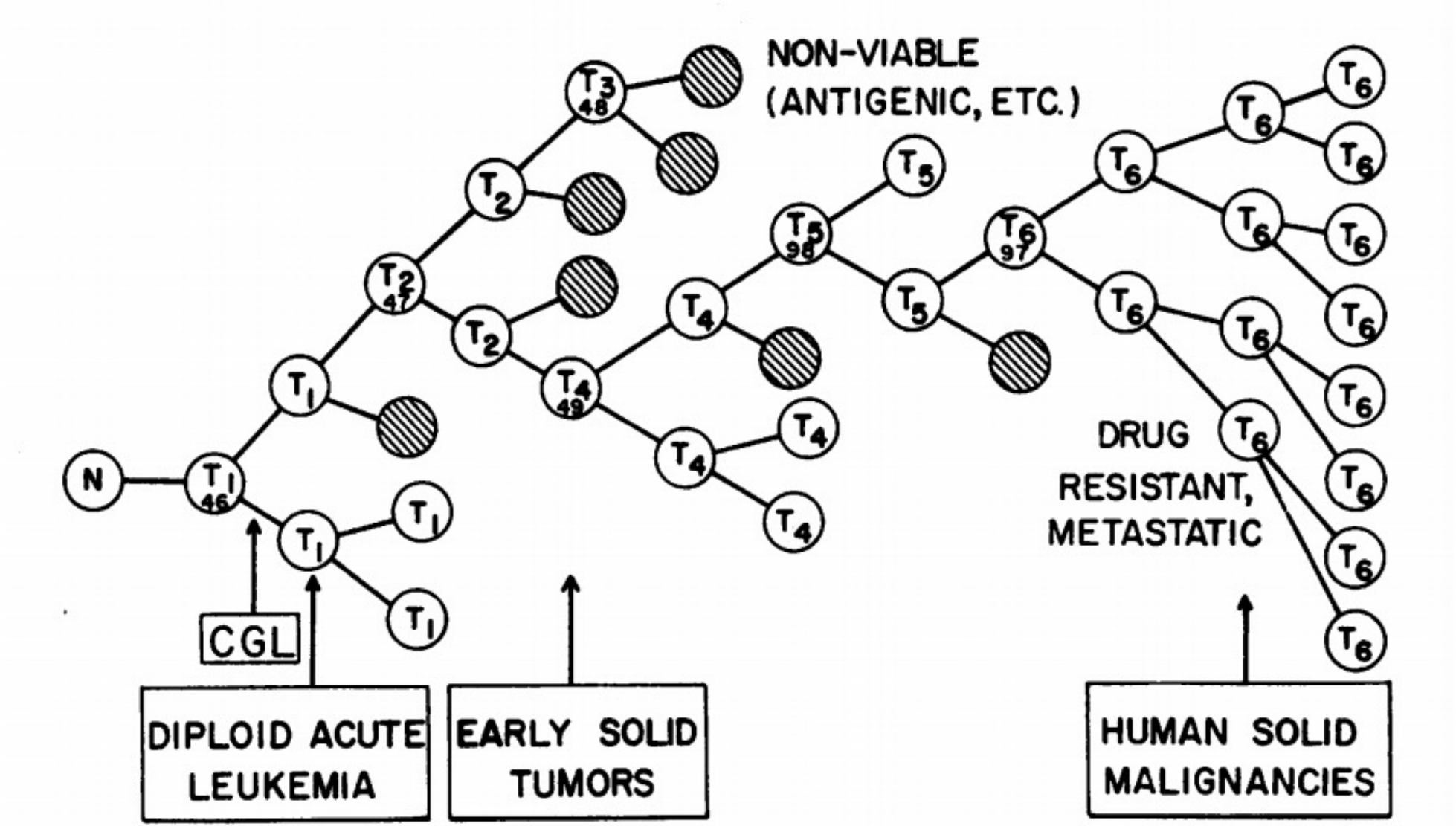
Tumor evolution





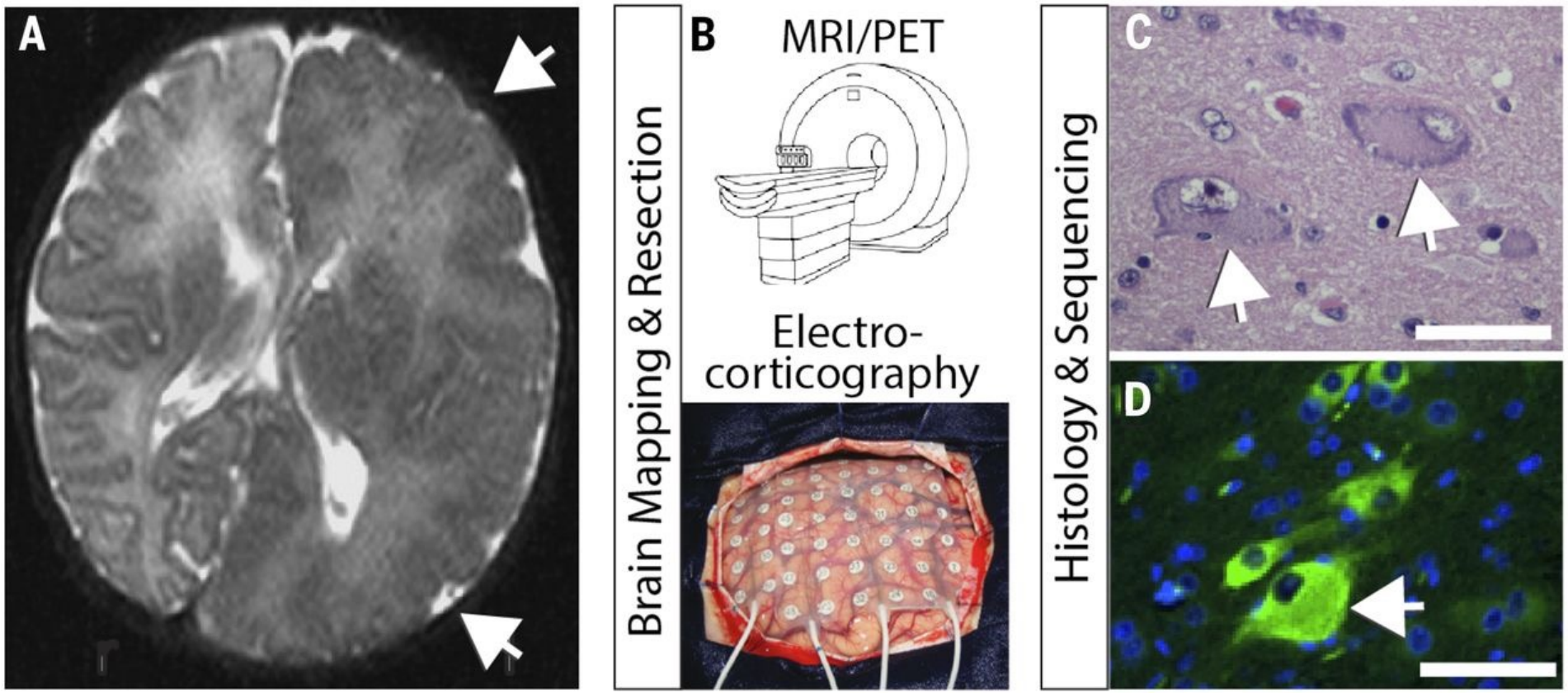
Recombination/crossover in germ cells

Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

Peter C. Nowell (1976) *Science.* 194(4260):23-28 DOI: 10.1126/science.959840

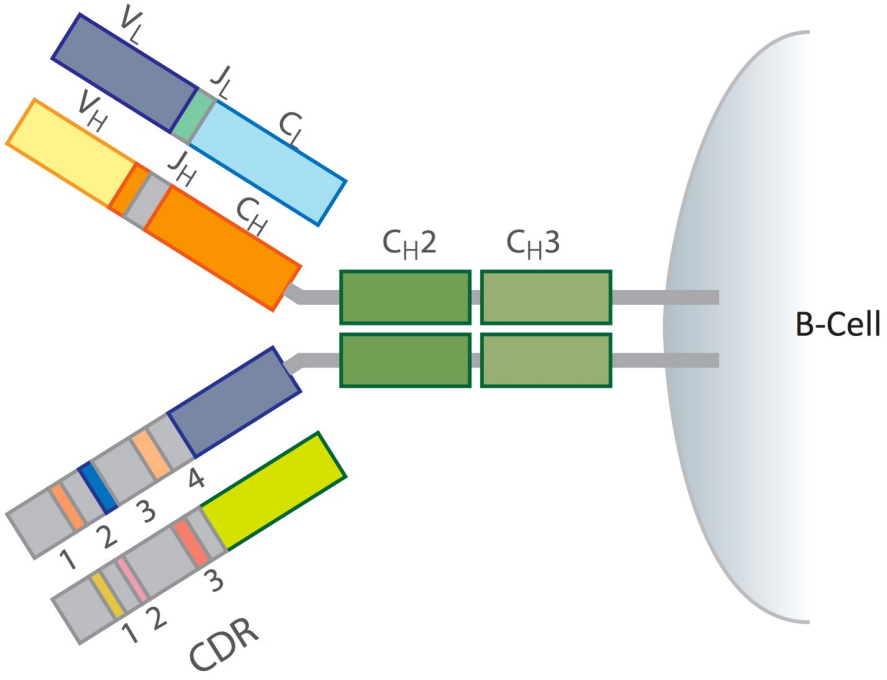


An example of brain somatic mosaicism that leads to a focal overgrowth condition.

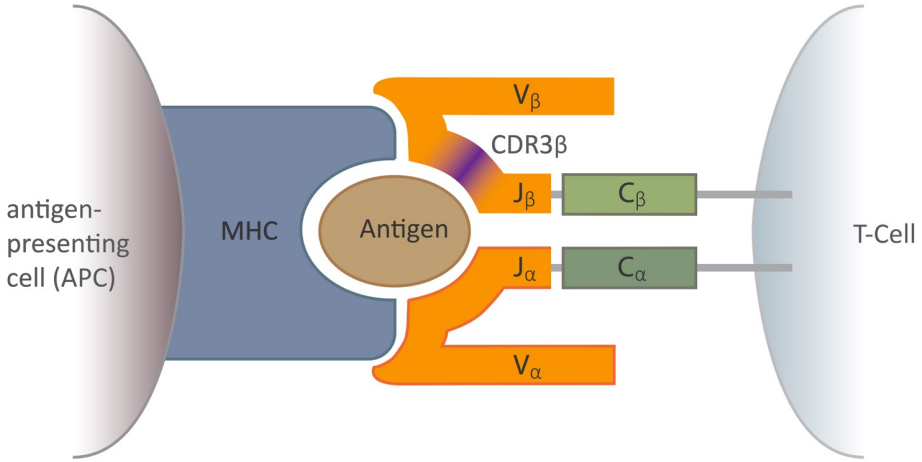
(**A**) Axial brain MRI of focal overgrowth from a 2-month-old child with intractable epilepsy and intellectual disability. (**B**) Brain mapping using high-resolution MRI is followed by surgical resection of diseased brain tissue. (**C**) Histological analysis with hematoxylin/eosin showing characteristic balloon cells consisting of large nuclei, distinct nucleoli, and glassy eosinophilic cytoplasm. (**D**) After surgery, the patient showed clinical improvement.

Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. McConnell et al (2017) Science. doi: 10.1126/science.aal1641

Immunology



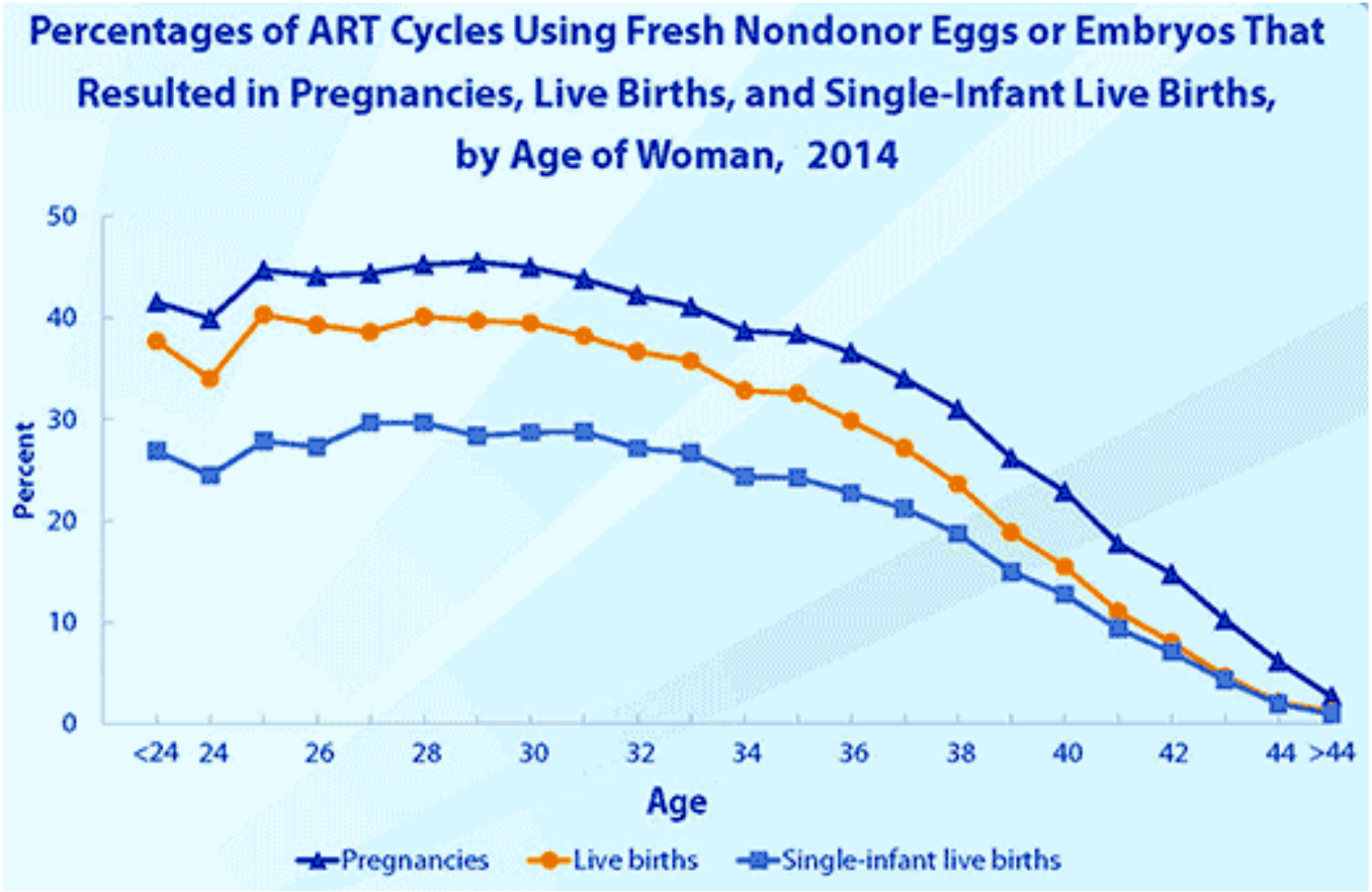
- Massive diversity rivaled only by germ cells
- Somatic recombination

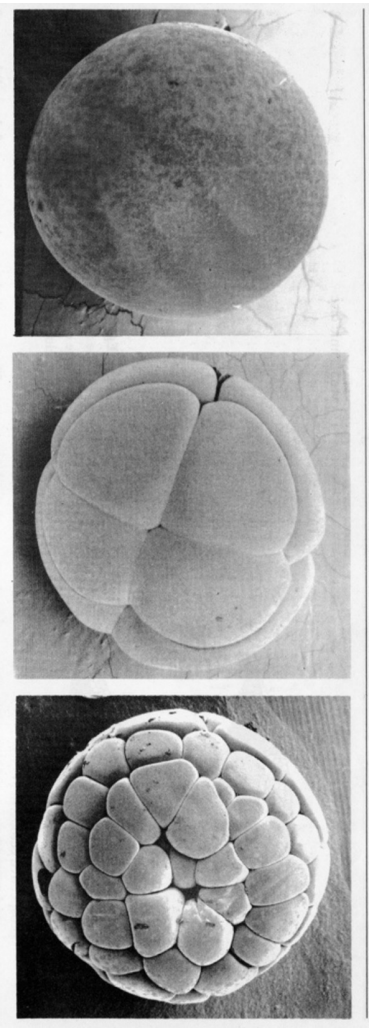


- B cells – antibody generation
- T cells – antigen response

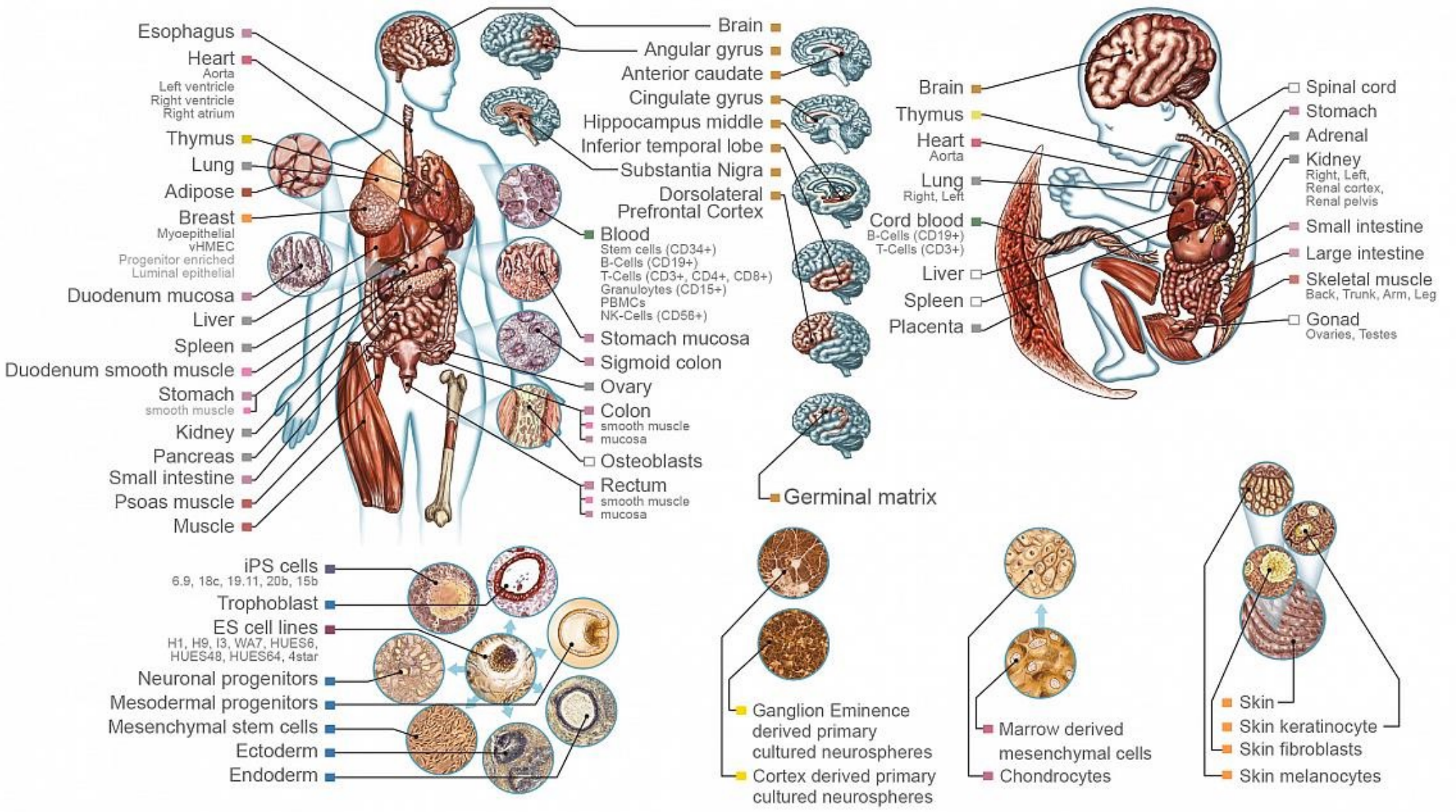
Single cell research. Illumina.

In-vitro Fertilization



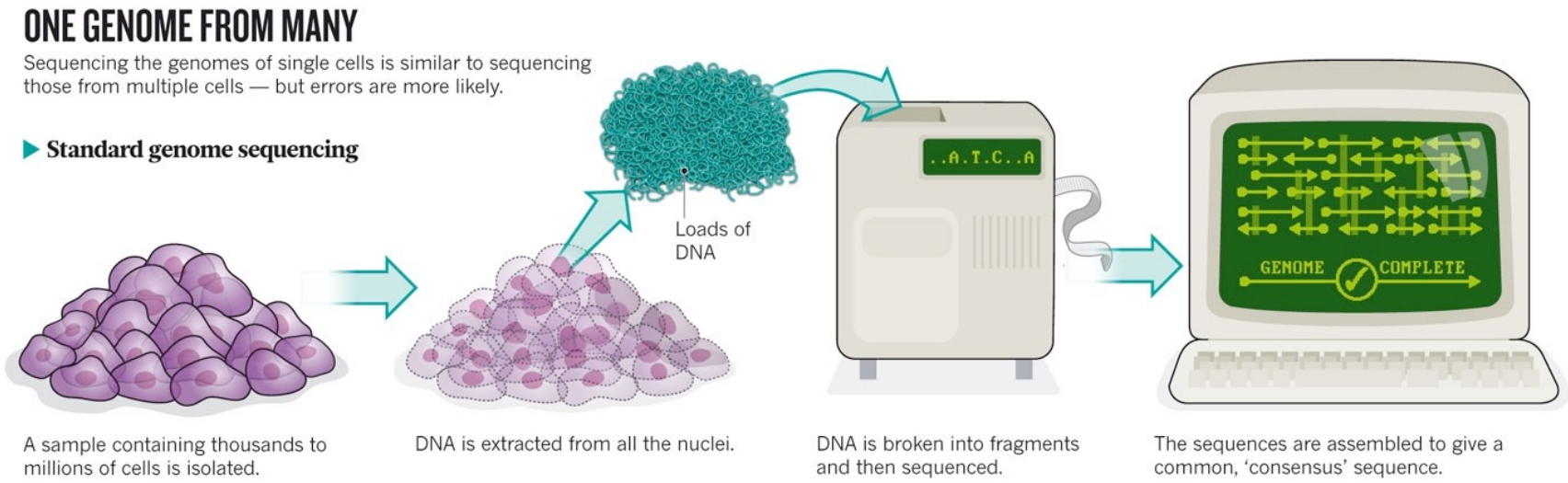


Sources of (Cellular) Heterogeneity



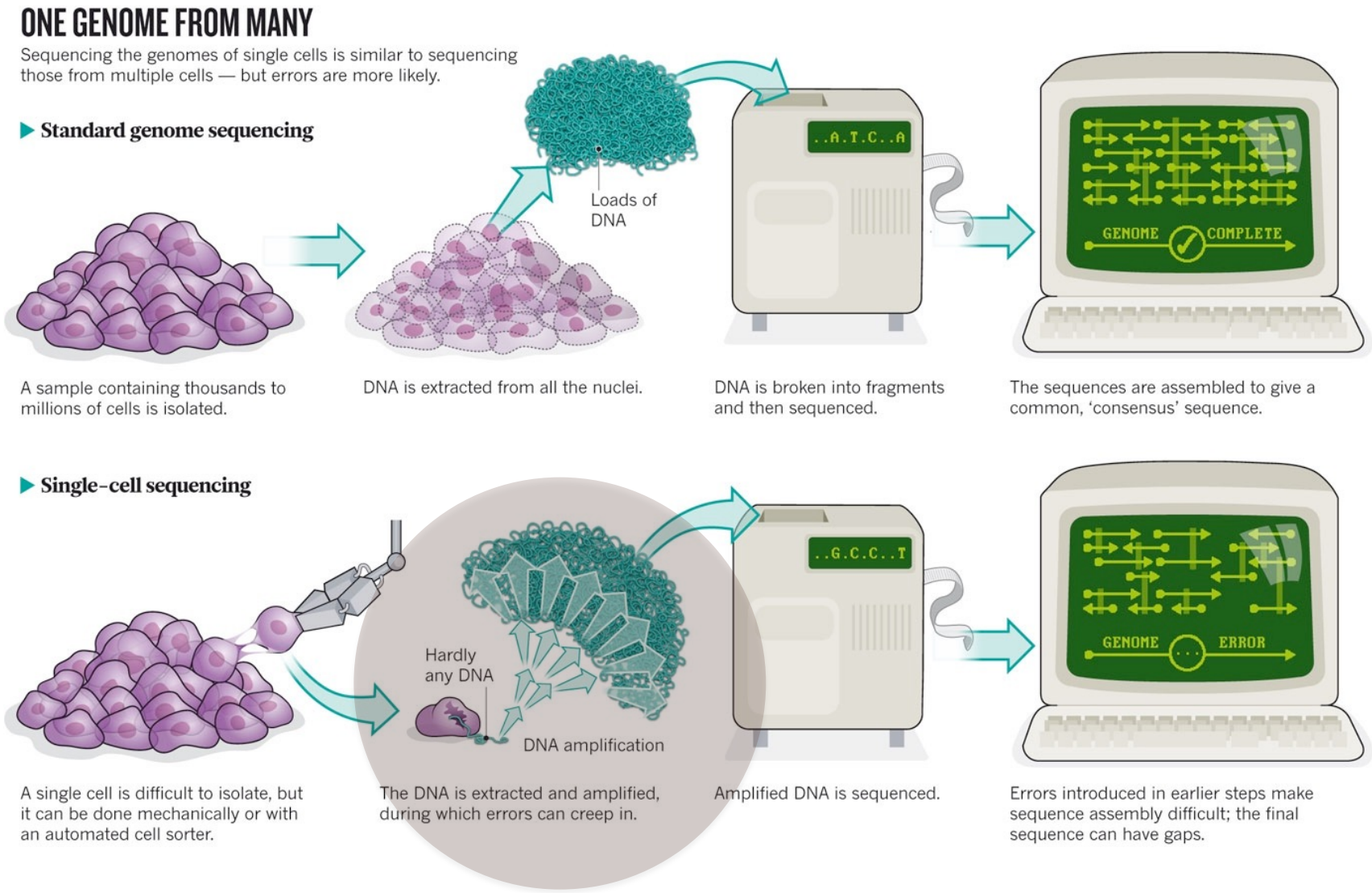
Roadmap Epigenomics Consortium

Single-cell vs. bulk sequencing



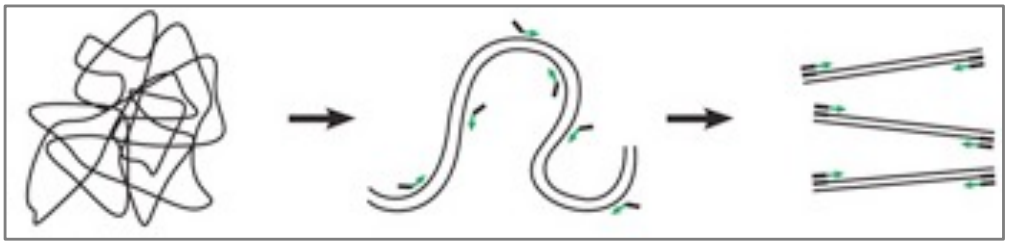
Brian Owens, Nature News 2012

Single-cell vs. bulk sequencing

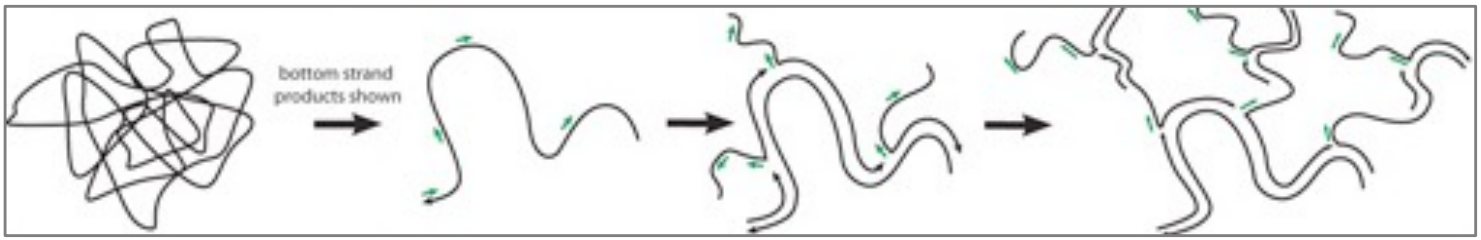


Brian Owens, Nature News 2012

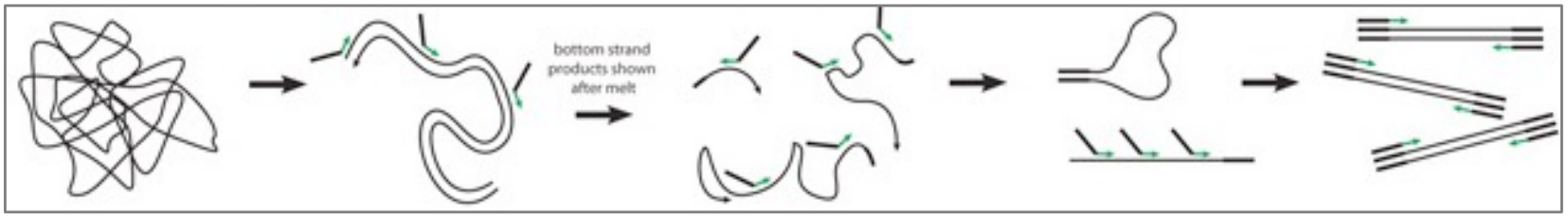
Whole Genome Amplification Techniques



DOP-PCR: Degenerate Oligonucleotide Primed PCR
Telenius et al. (1992) Genomics



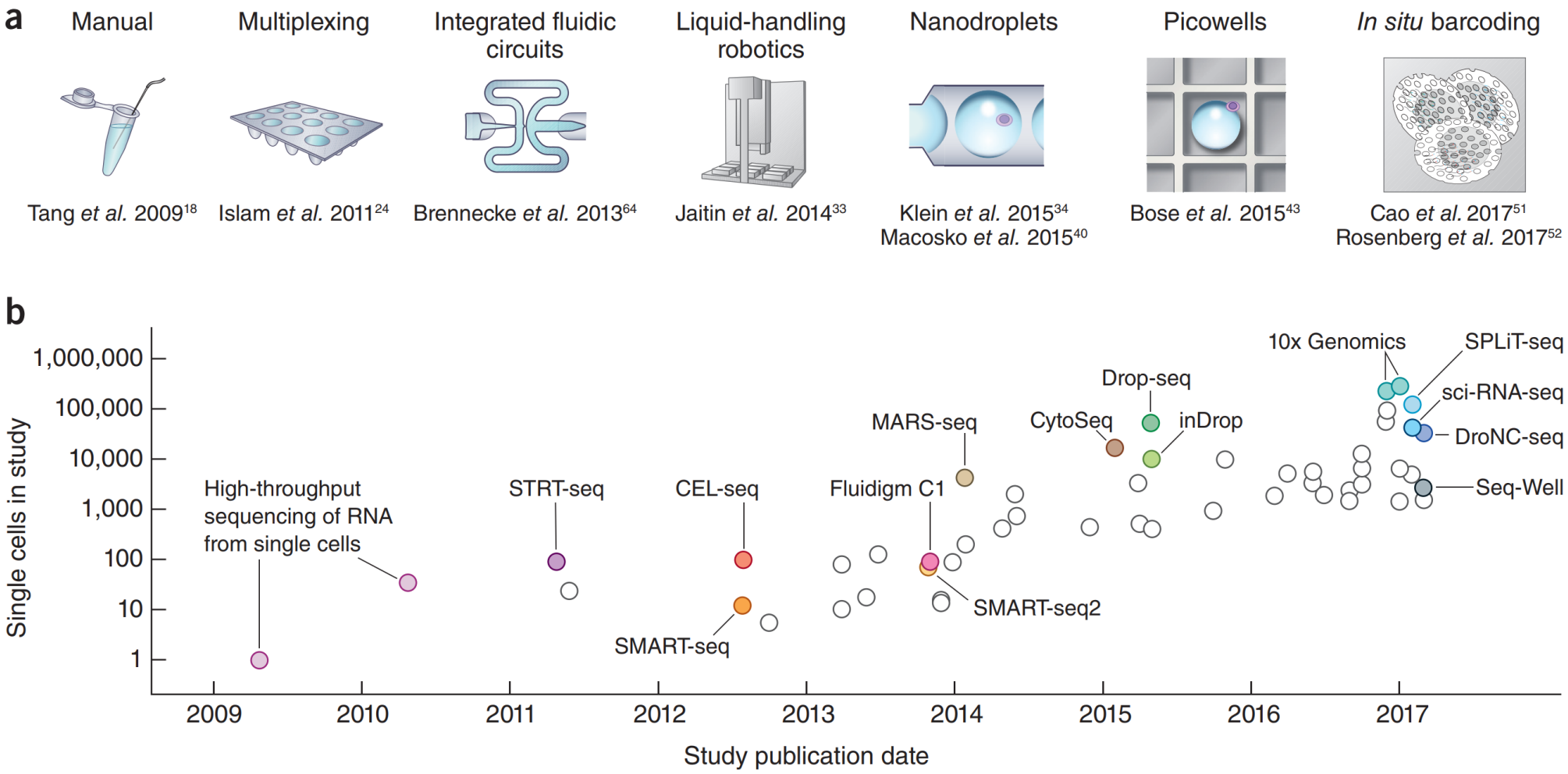
MDA: Multiple Displacement Amplification
Dean et al. (2002) PNAS



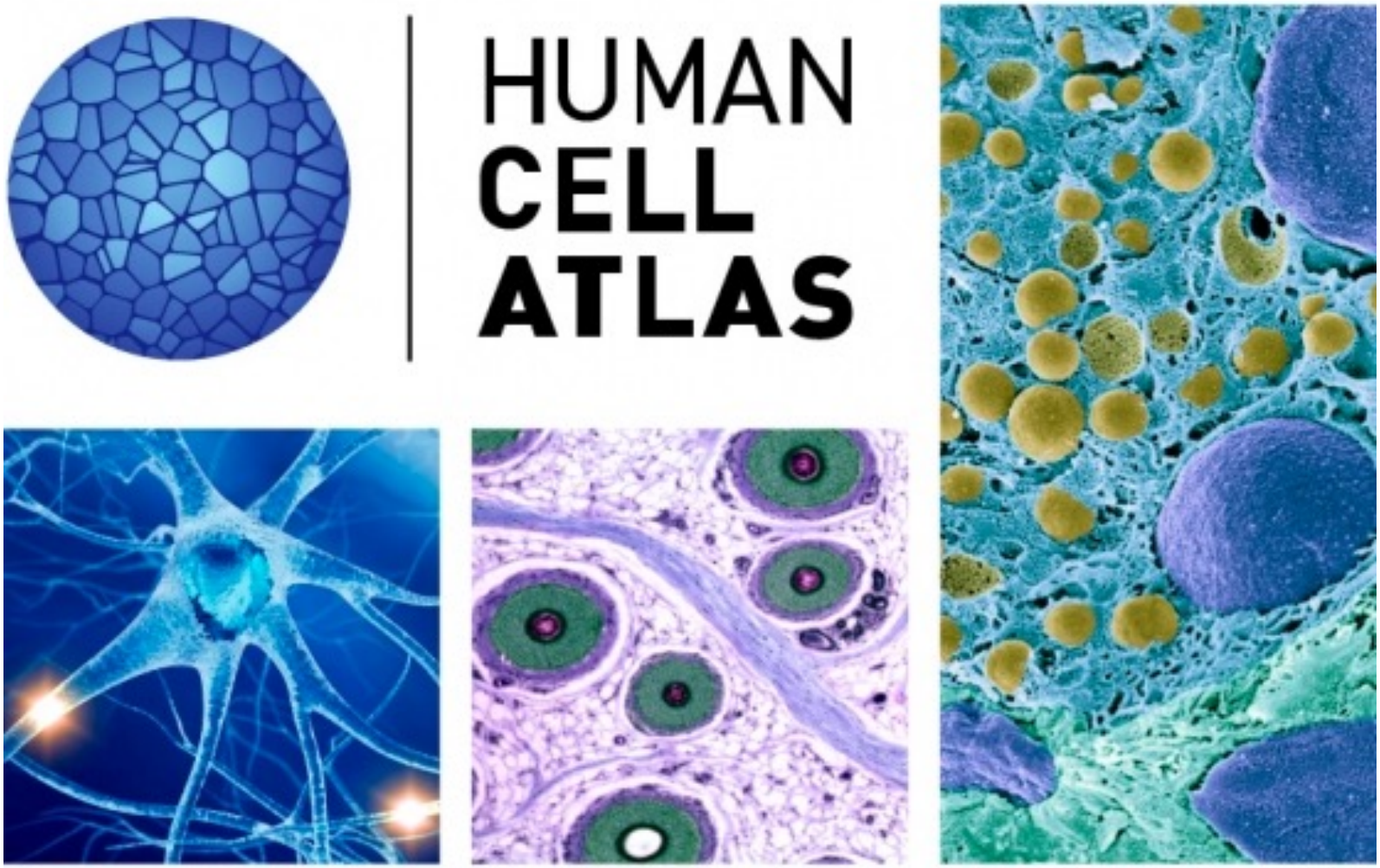
MALBAC: Multiple Annealing and Looping Based Amplification Cycles
Zong et al. (2012) Science

Paul Blainey, FEMS Microbiol Rev. 2013

A decade of single-cell RNA-seq



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



https://www.humancellatlas.org/

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**

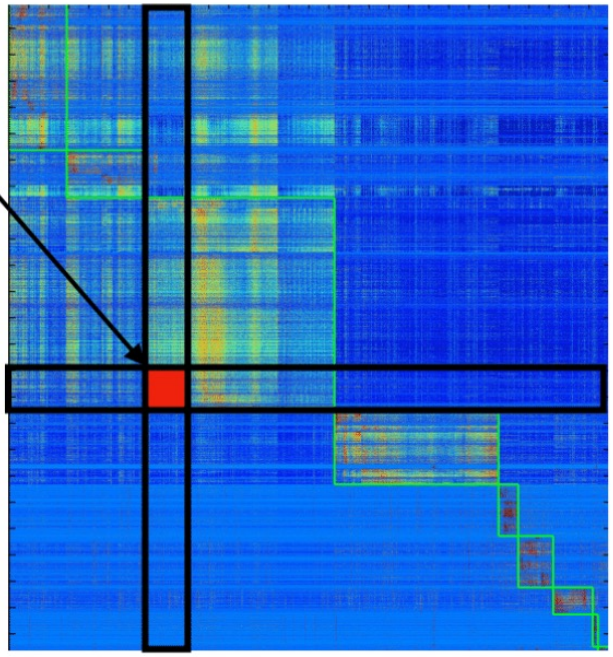
sequencing

Single Cells (samples)

cell j

Genes (features)

gene i



Gene expression matrix

computationally explore complex biological systems

Martin Zhang

Clustering Refresher

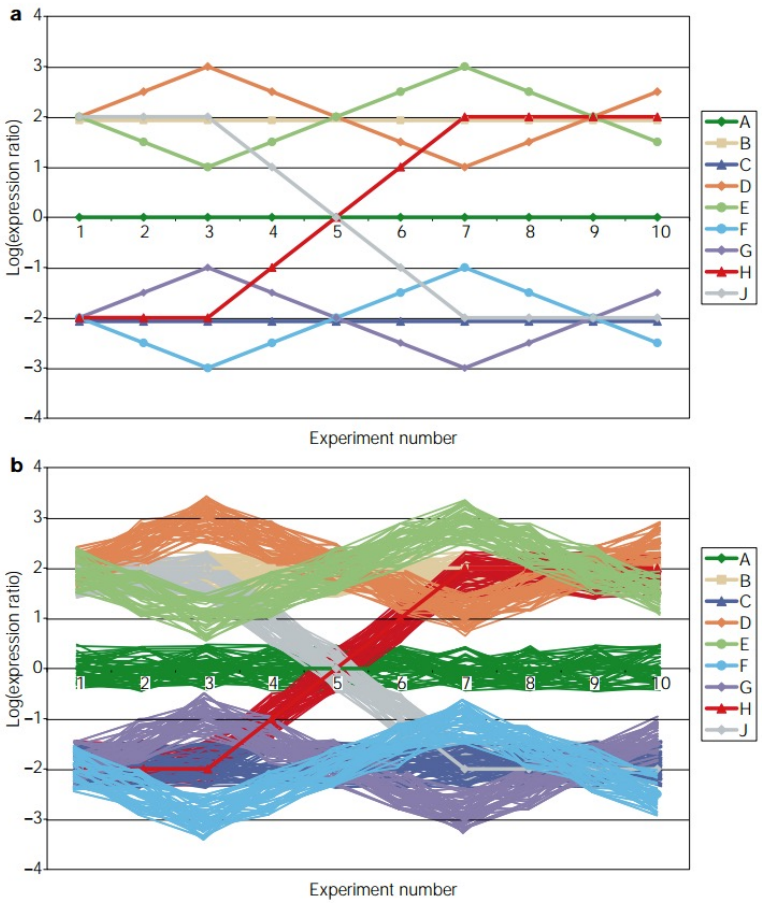
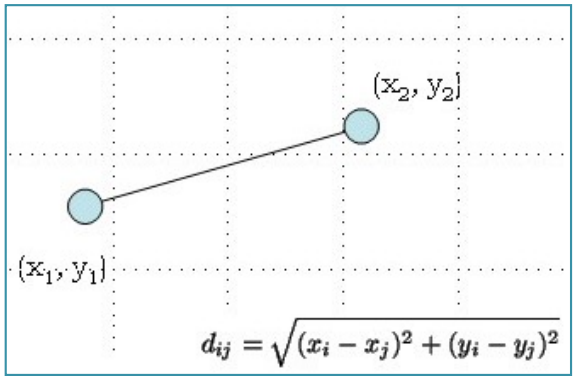
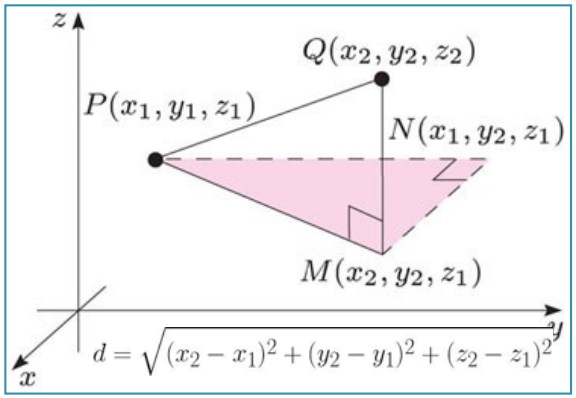


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments.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

Euclidean Distance

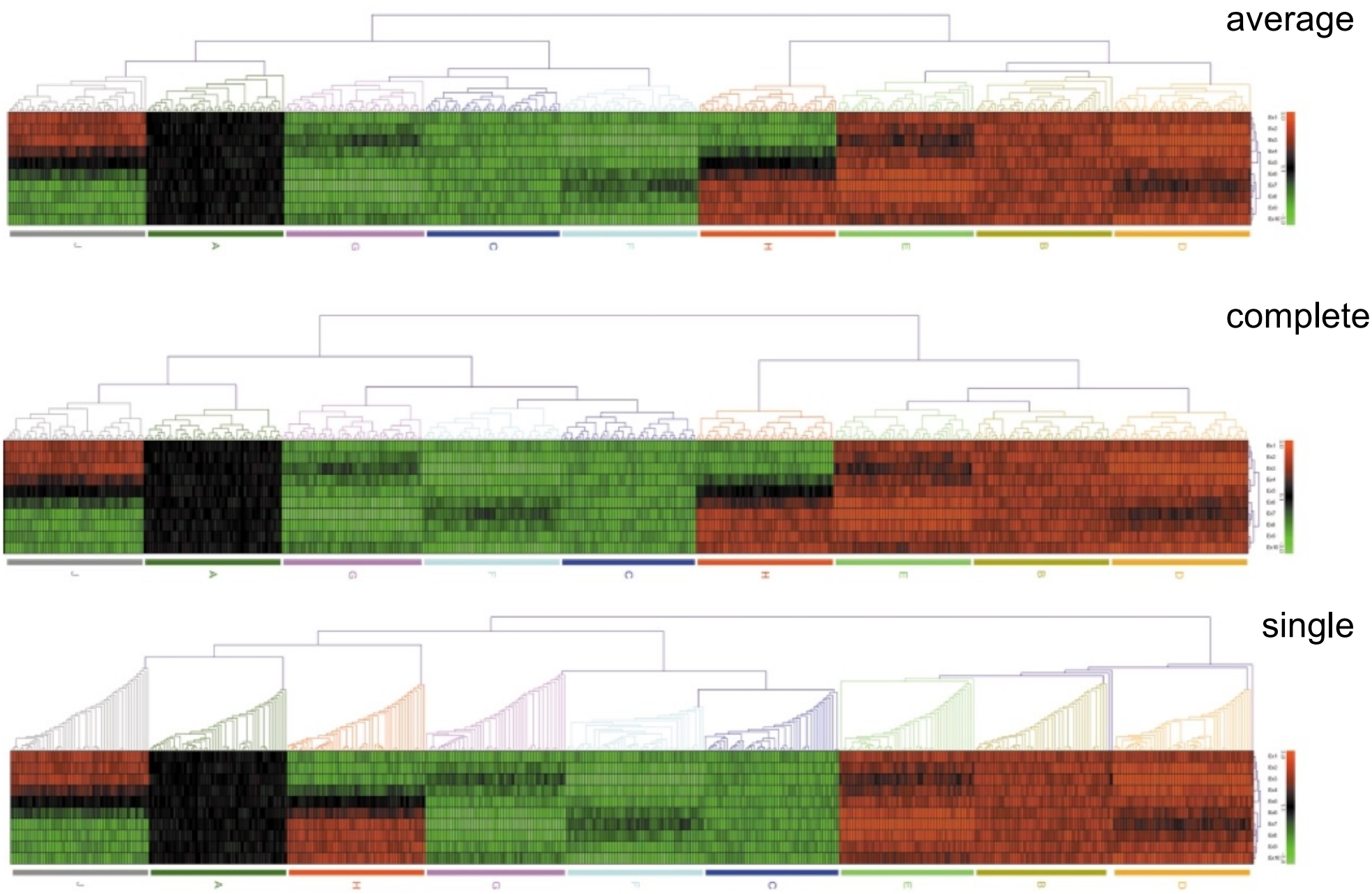




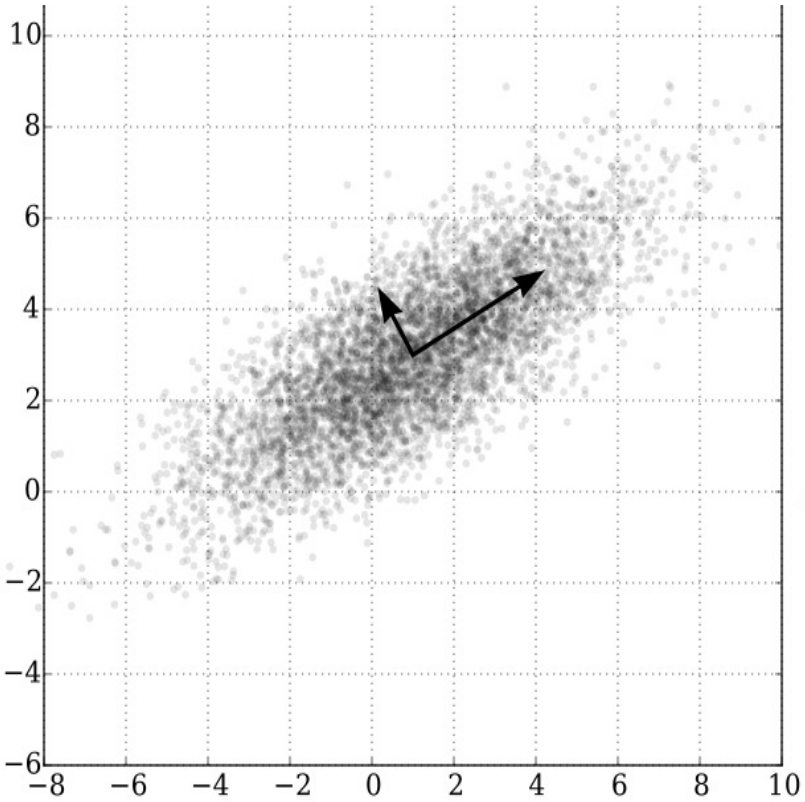
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

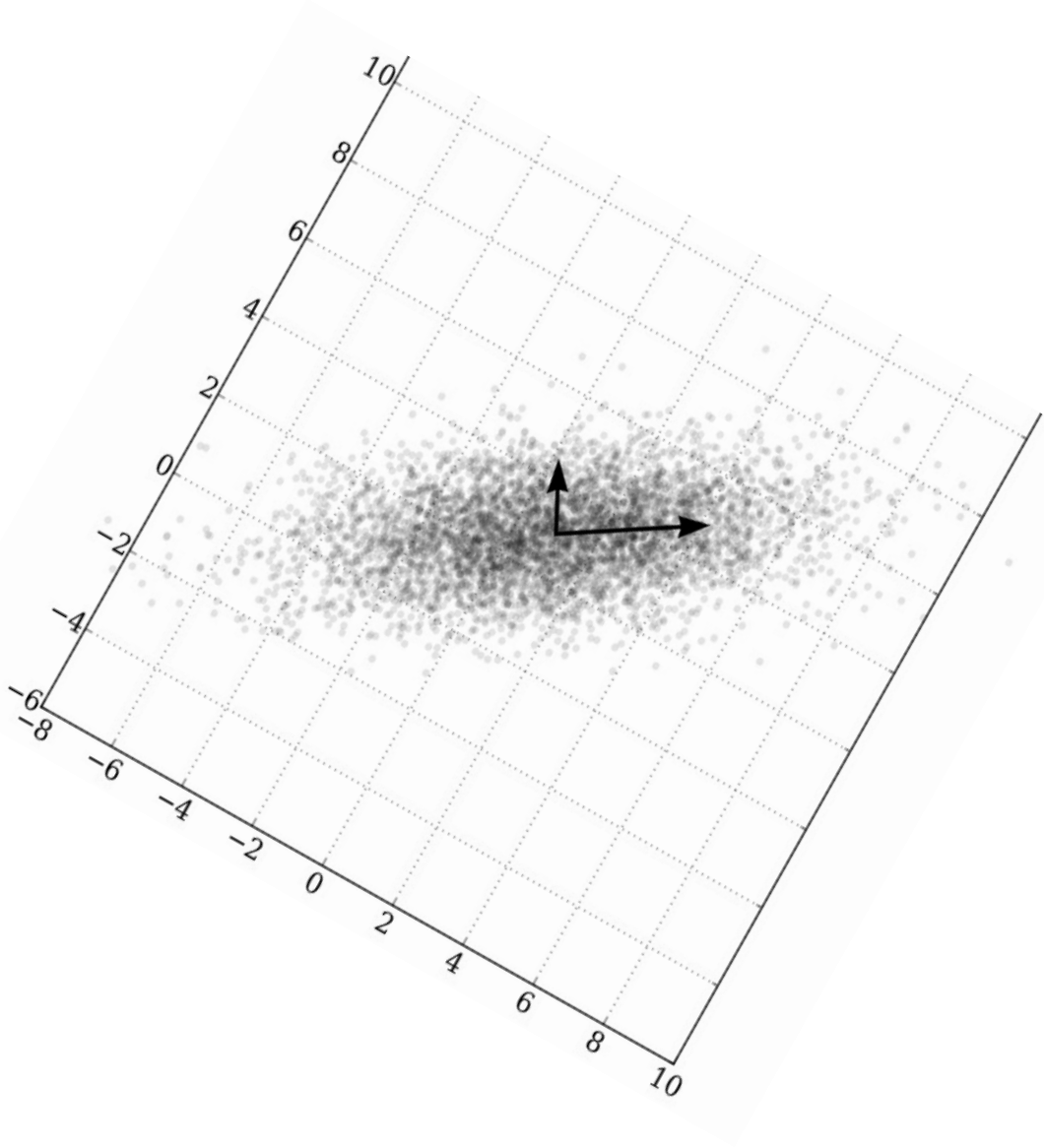
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



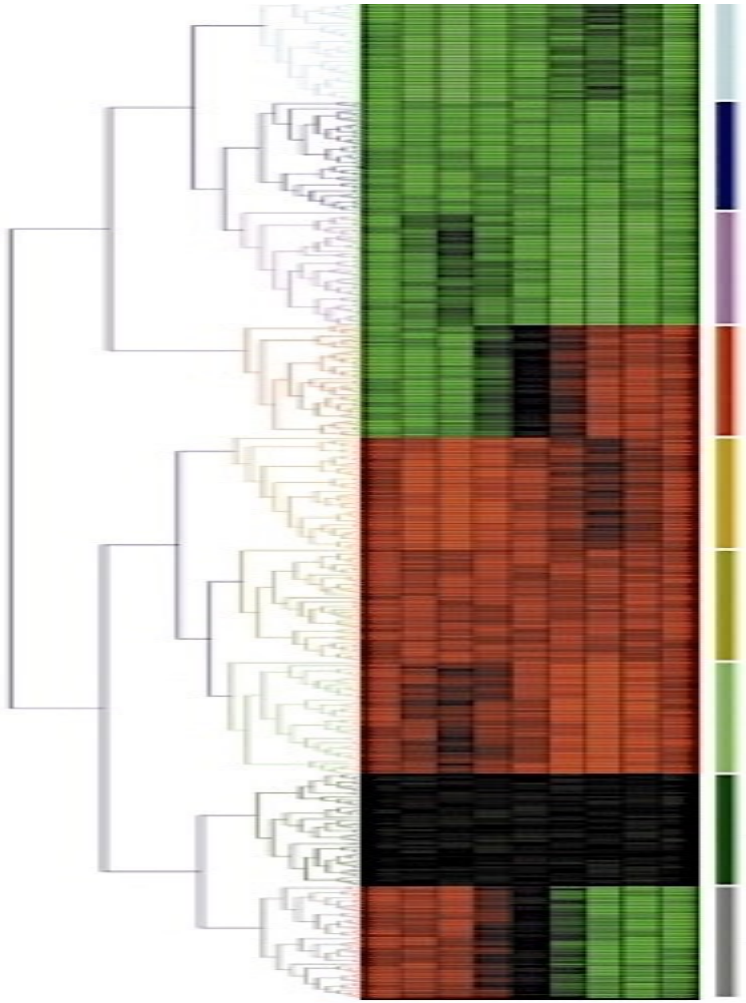
Principle Components Analysis (PCA)

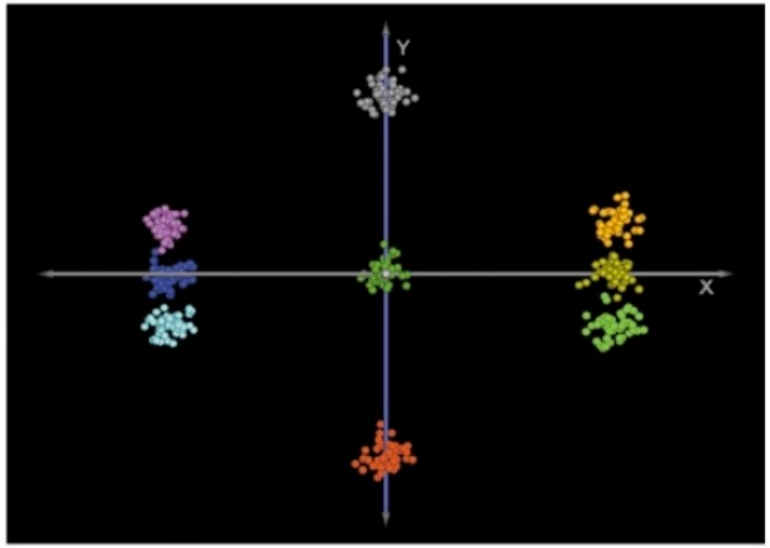




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





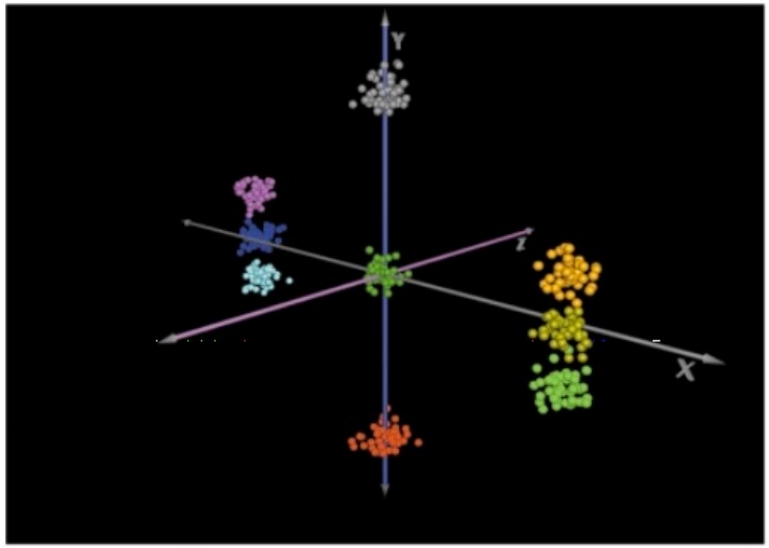
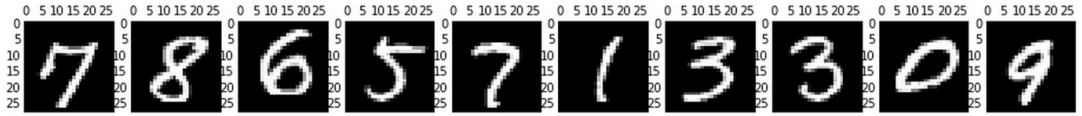
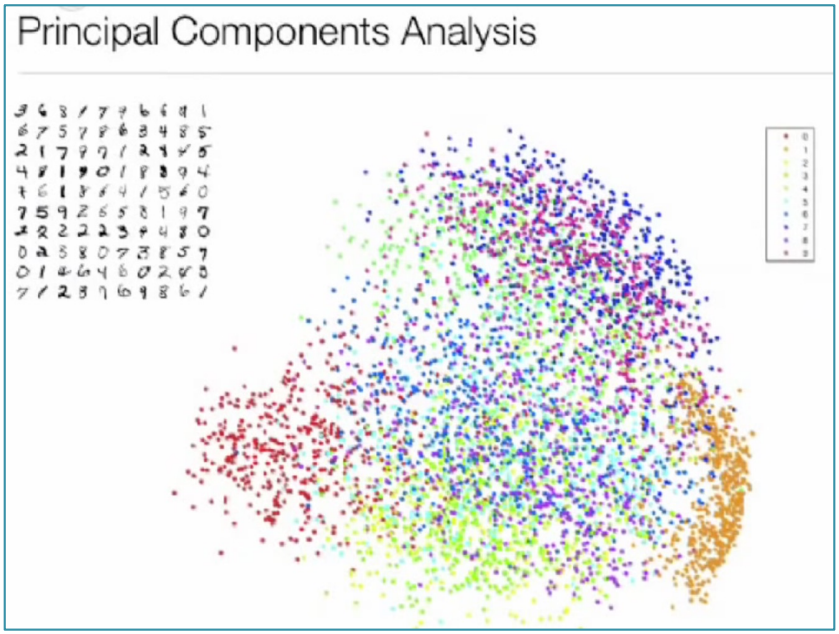
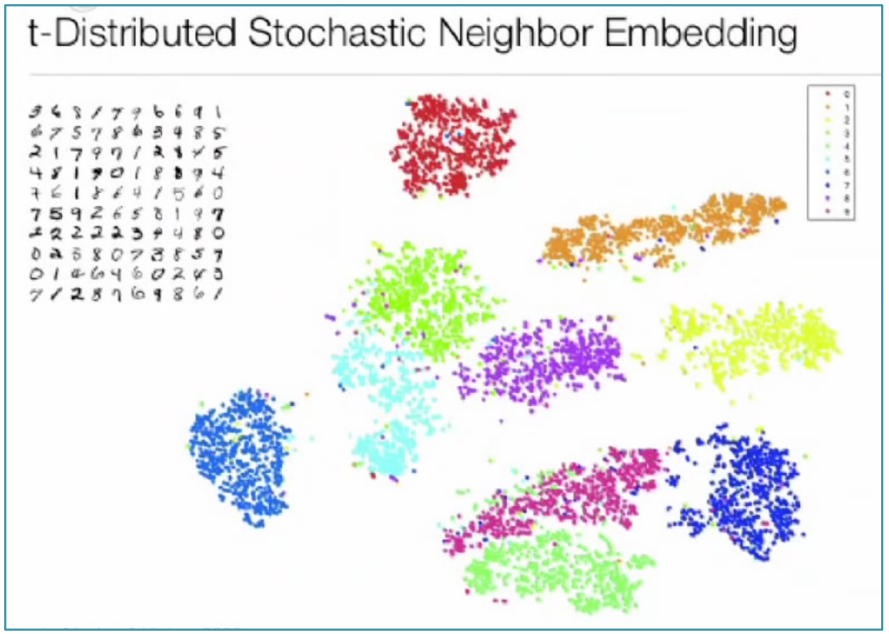


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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

PCA and t-SNE







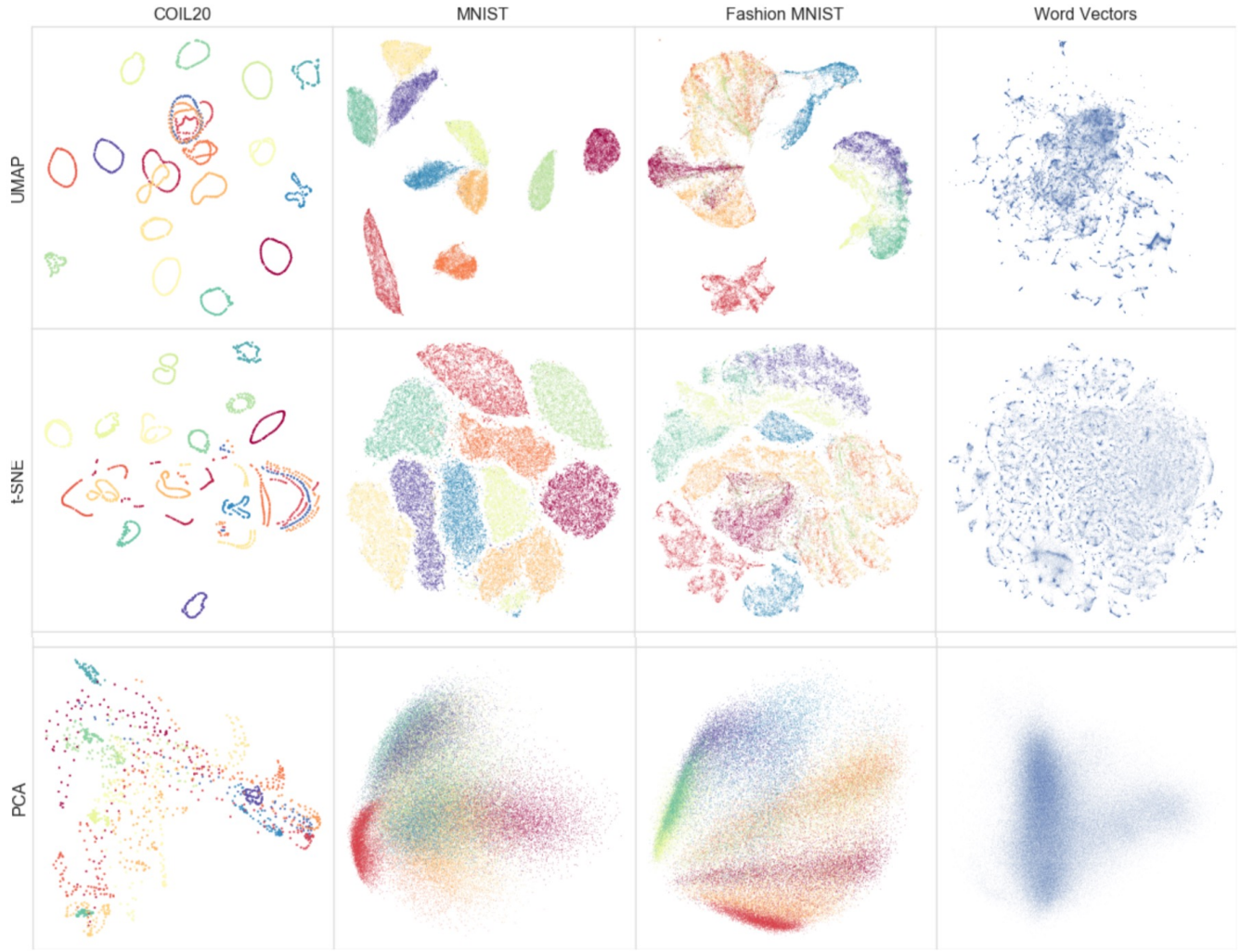
t-distributed Stochastic Neighborhood Embedding

- 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

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

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

scSummary

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