Lecture 15. Single Cell Analysis

Michael Schatz

March 27, 2017

JHU 601.749: Applied Comparative Genomics

Project Proposal!
Due March 15

Project Proposal

Assignment Date: March 7, 2018
Due Date: Thursday, March 15, 2017 @ 11:59pm

Review the Project Ideas page

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

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 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

Submit the proposal as a single page PDF on blackboard. After submitting your proposal, we will 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!

HW6:
Due March 29

Assignment 6: RNA-seq and differential expression

Assignment Date: Thursday, March 15, 2018
Due Date: Thursday, March 29, 2018 @ 11:59pm

Assignment Overview

In this assignment, you will analyze gene expression data and learn how to make several kinds of plots in the environment of your choice. (We suggest Python or R.) **Make sure to show your work/code in your writeup!** As before, any questions about the assignment should be posted to Piazza.

Question 1. Time Series [10 pts]

This file contains pre-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.

a. 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.]

b. 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?

c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.

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 million to billions 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 (1000 to 5000) out of a much larger set (1M) so that you can evaluate this bias.

In this file with 1,000,000 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 sequenced.

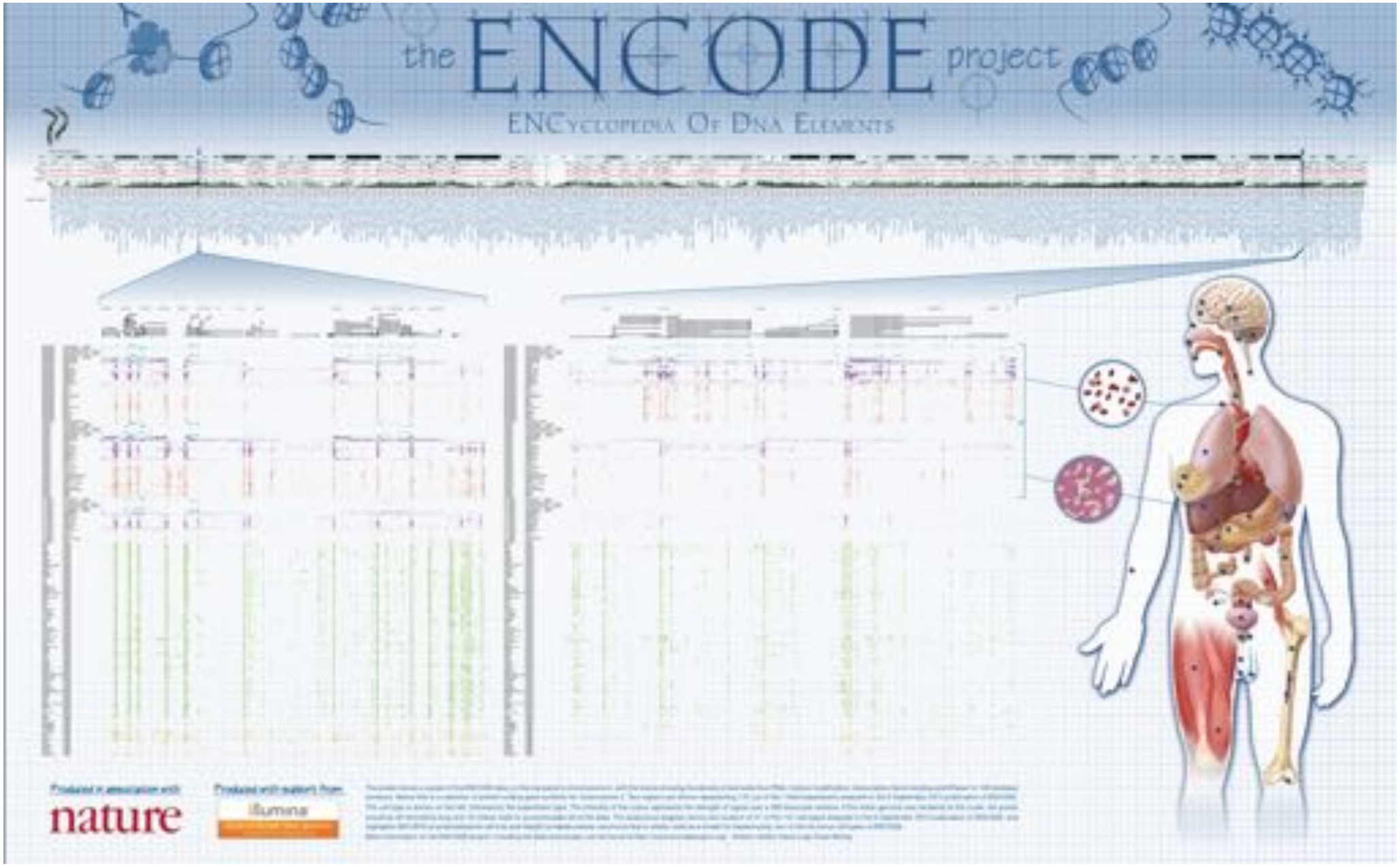
a. Randomly sample 1000 rows. Do this simulation 10 times and record the relative abundance of each of the 15 genes. Plot the mean vs. variance.

b. Do the same sampling experiment but sample 5000 rows each time. Again plot the mean vs. variance.

c. Is the variance greater in (a) or (b)?, and explain why. What is the relationship between abundance and variance?

d. Suppose you had received data where the number of times a gene name occurs corresponds to the number of reads mapped to that gene. In a few sentences explain how would you normalize the data, and what additional information would you need? [Hint: why is read count not enough?]

ENCODE Data Sets



1,640 data sets total over 147 different cell types

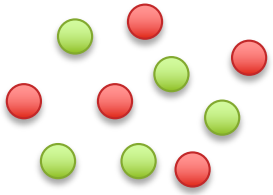
Single Cell Analysis

1. Why single cells?
2. scDNA
3. scRNA and other assays

Population Heterogeneity

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

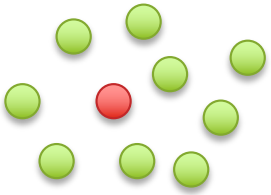
Experiment 1: 50/50



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

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

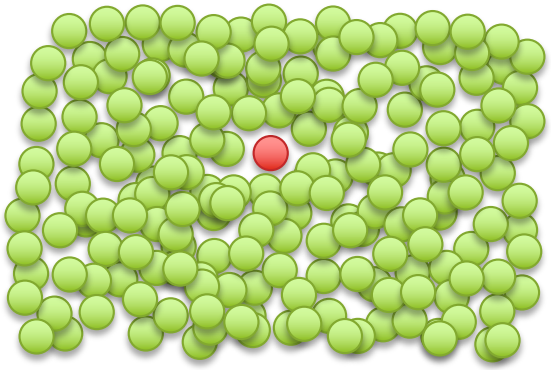
Experiment 2: 1/10



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

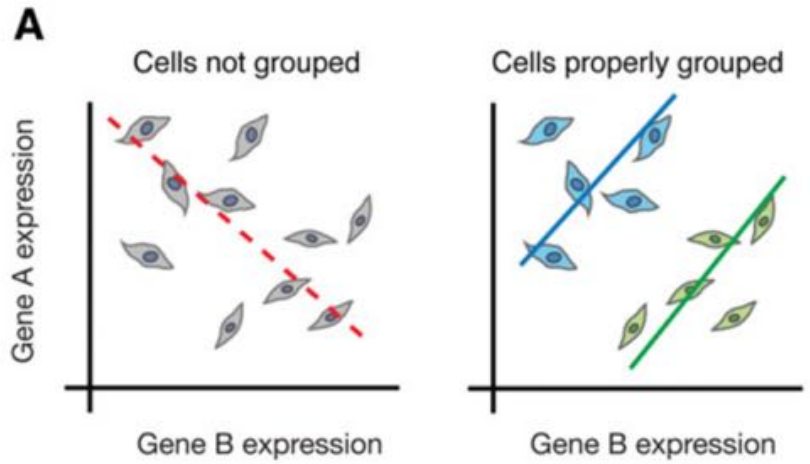
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

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



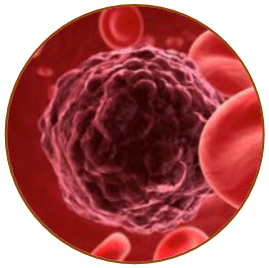
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

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)

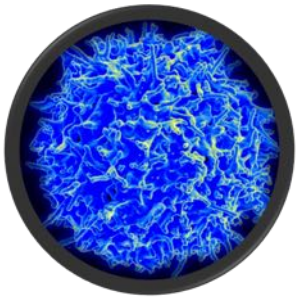
(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity

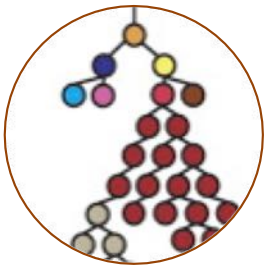


Circulating tumor cells

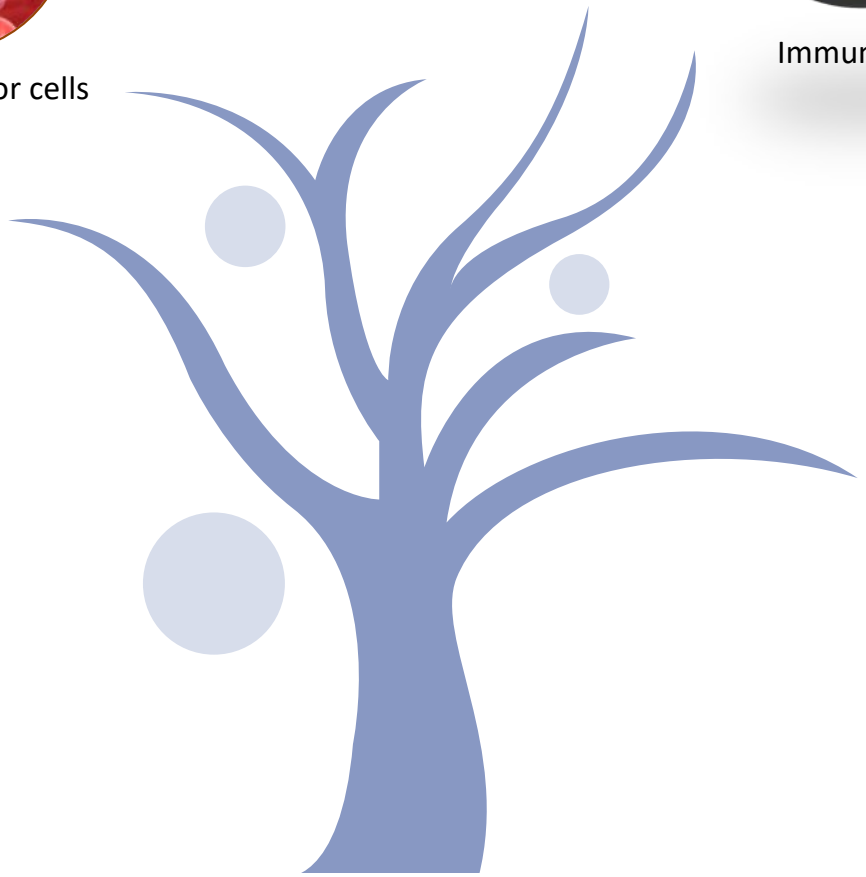
Neuronal mosaicism

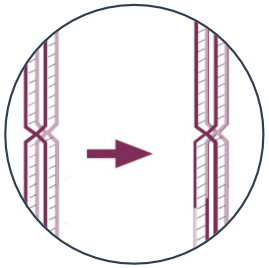


Immune Cells



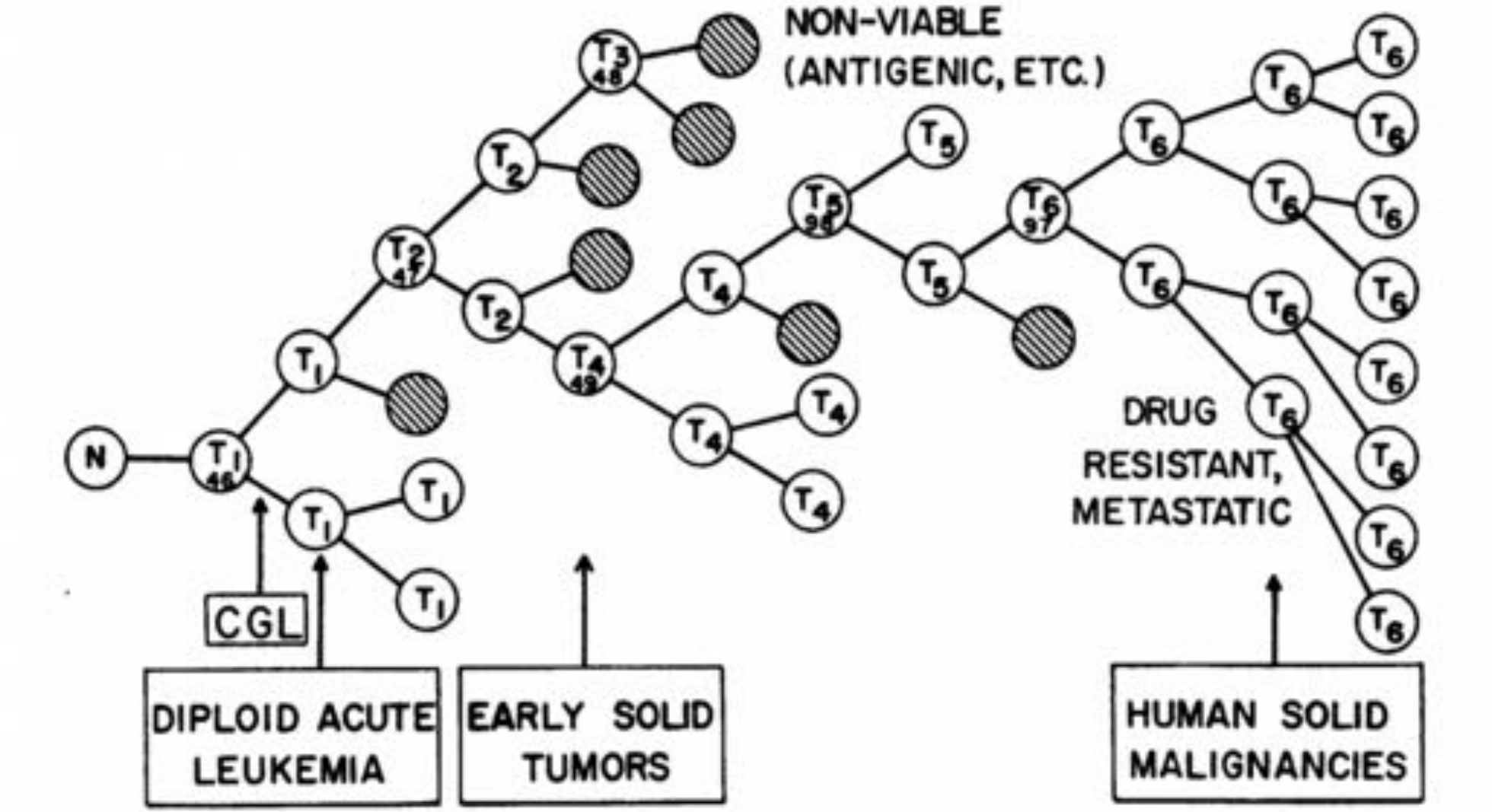
Tumor evolution





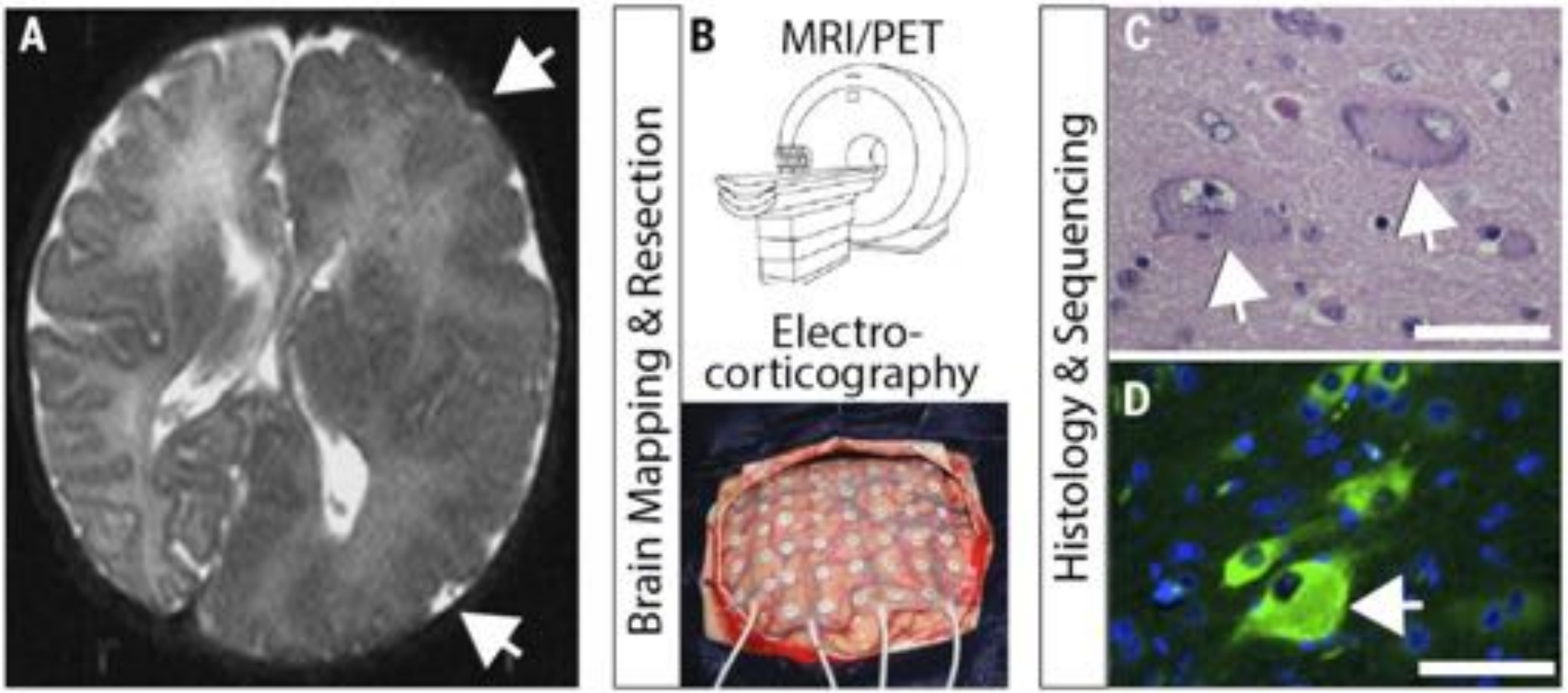
Recombination/crossover
in germ cells

Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

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

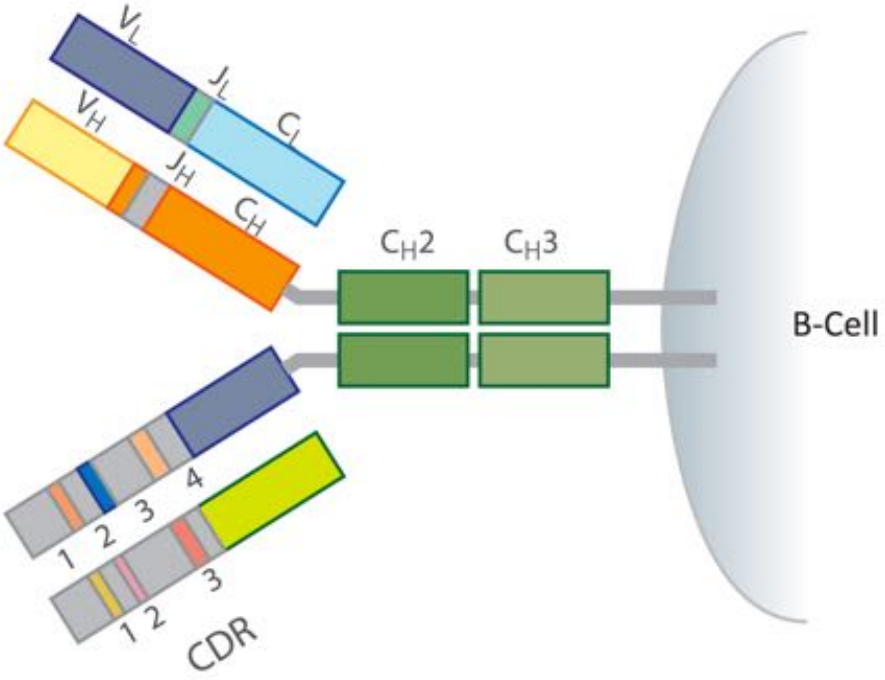


An example of brain somatic mosaicism that leads to a focal overgrowth condition.

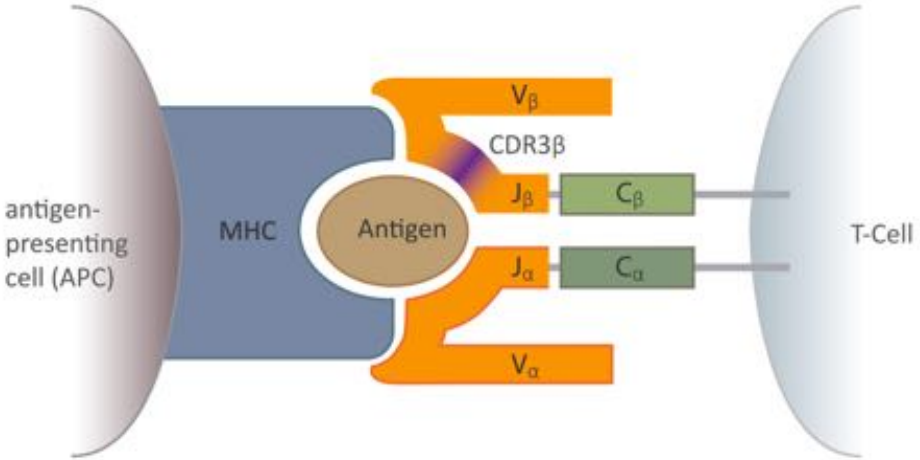
(**A**) Axial brain MRI of focal overgrowth from a 2-month-old child with intractable epilepsy and intellectual disability. (**B**) Brain mapping using high-resolution MRI is followed by surgical resection of diseased brain tissue. (**C**) Histological analysis with hematoxylin/eosin showing characteristic 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



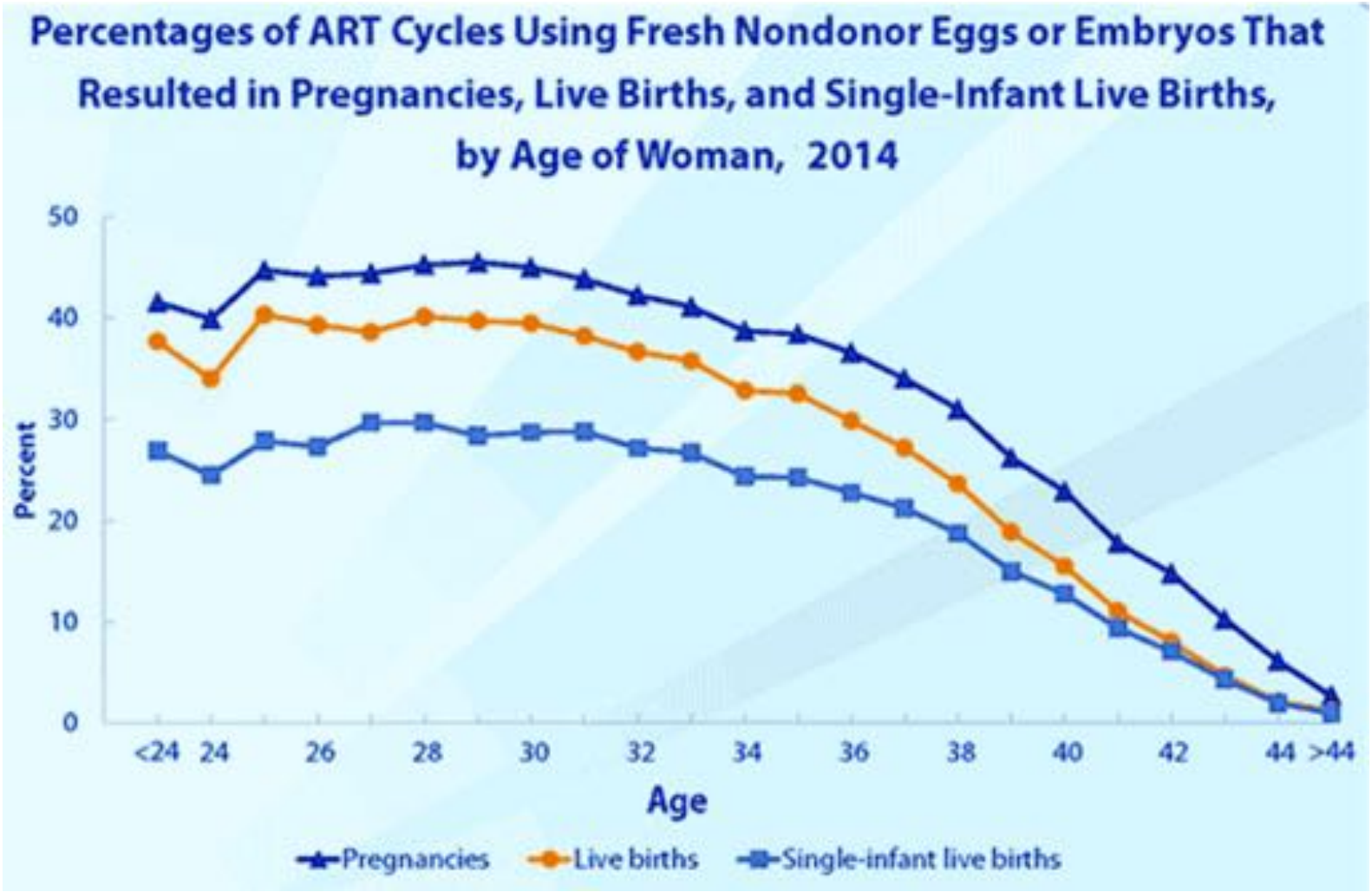
- **Massive diversity rivaled only by germ cells**
- **Somatic recombination**

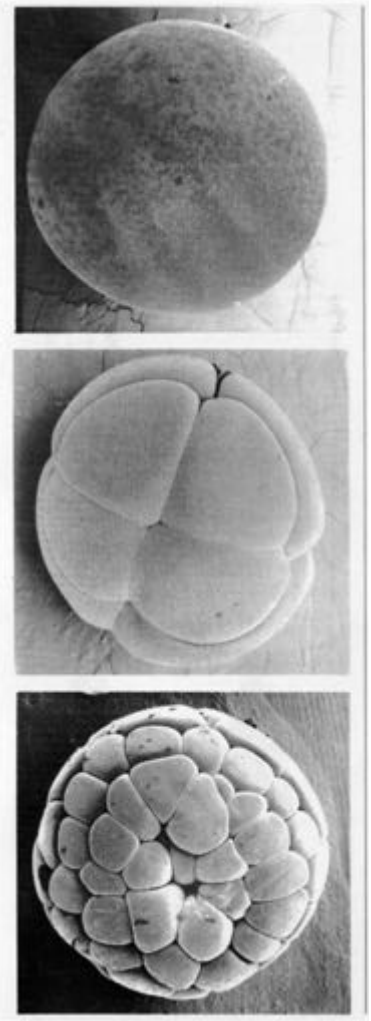


- **B cells – antibody generation**
- **T cells – antigen response**

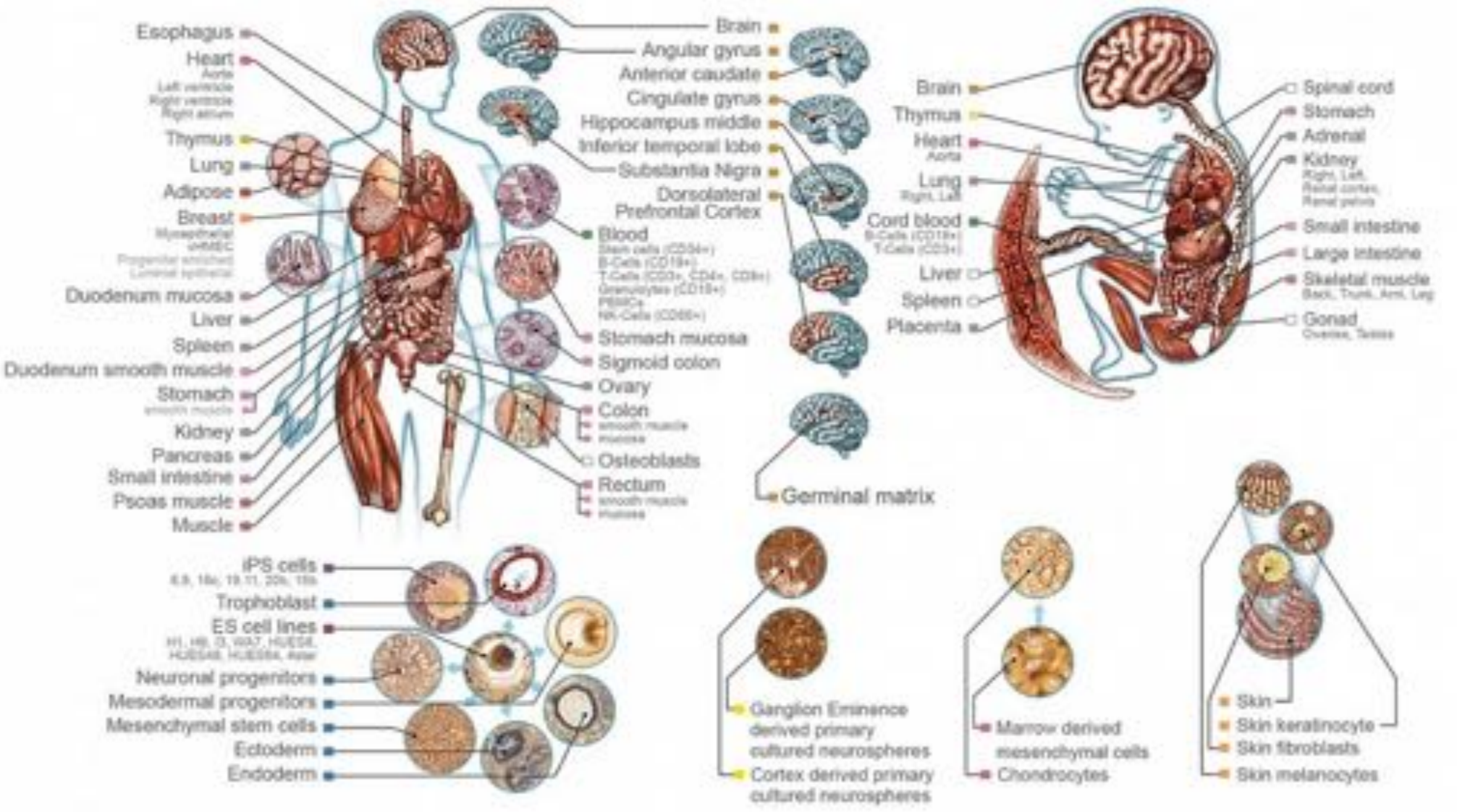
Single cell research. Illumina.

In-vitro Fertilization

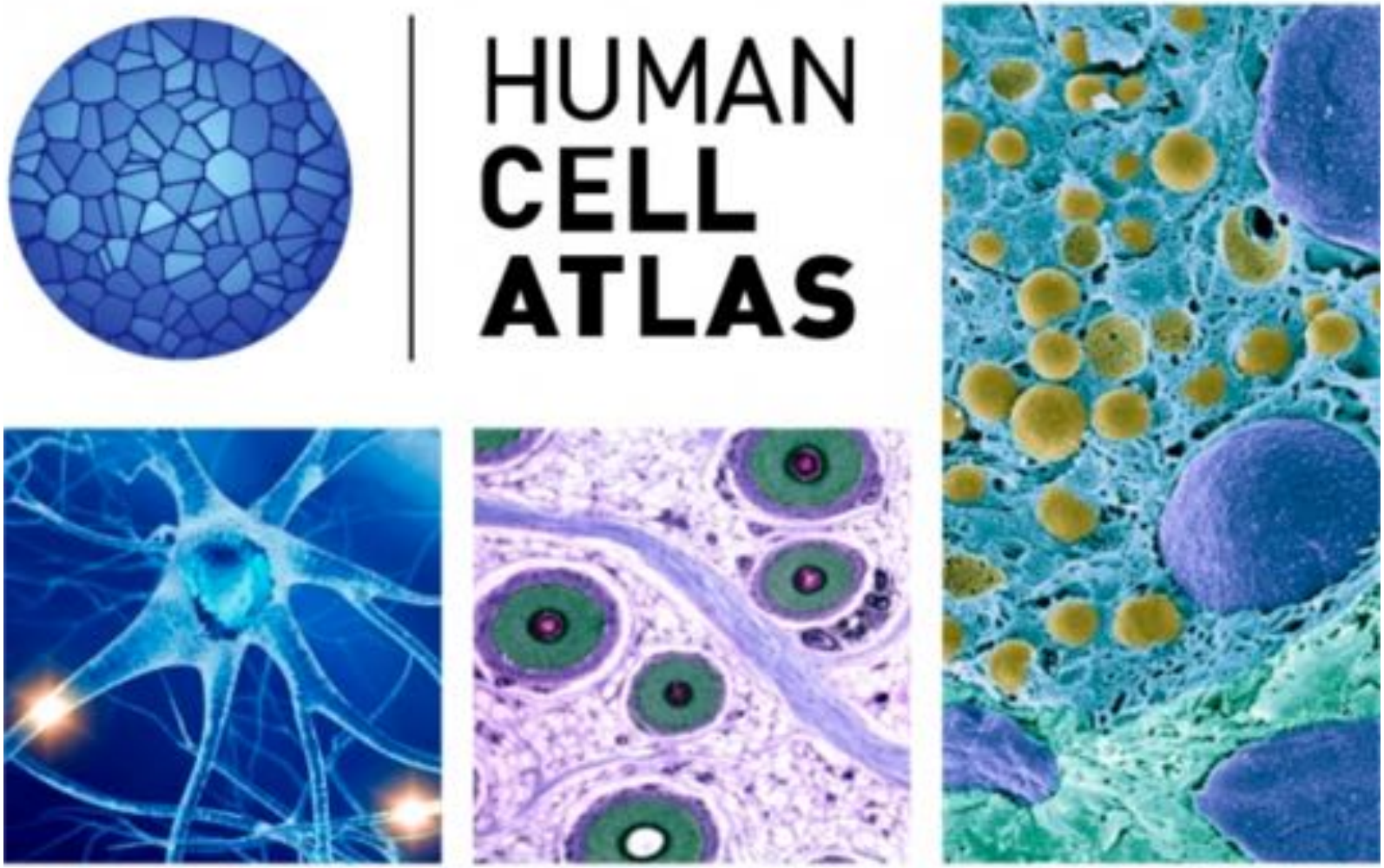




Sources of (Cellular) Heterogeneity



Roadmap Epigenomics Consortium



https://www.humancellatlas.org/

Clustering Refresher

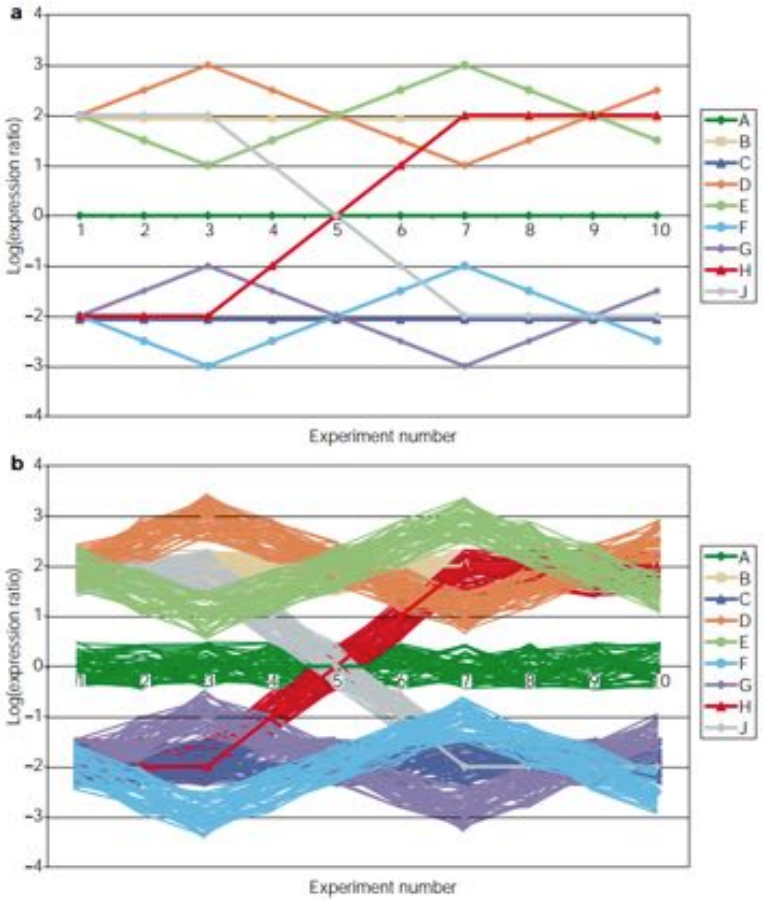
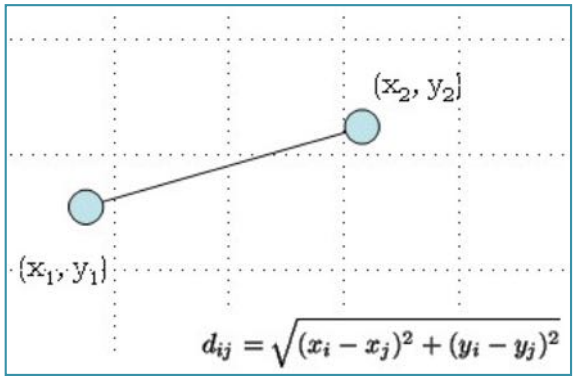
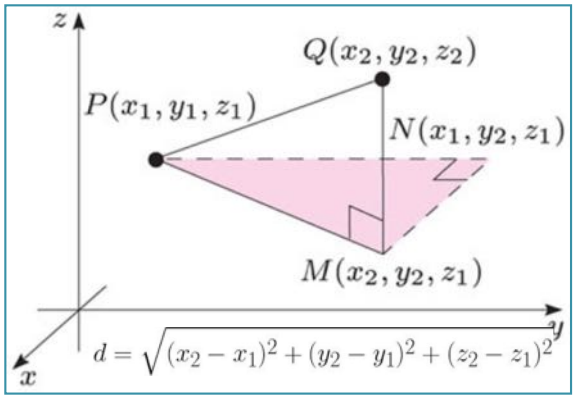


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments.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

Euclidean Distance

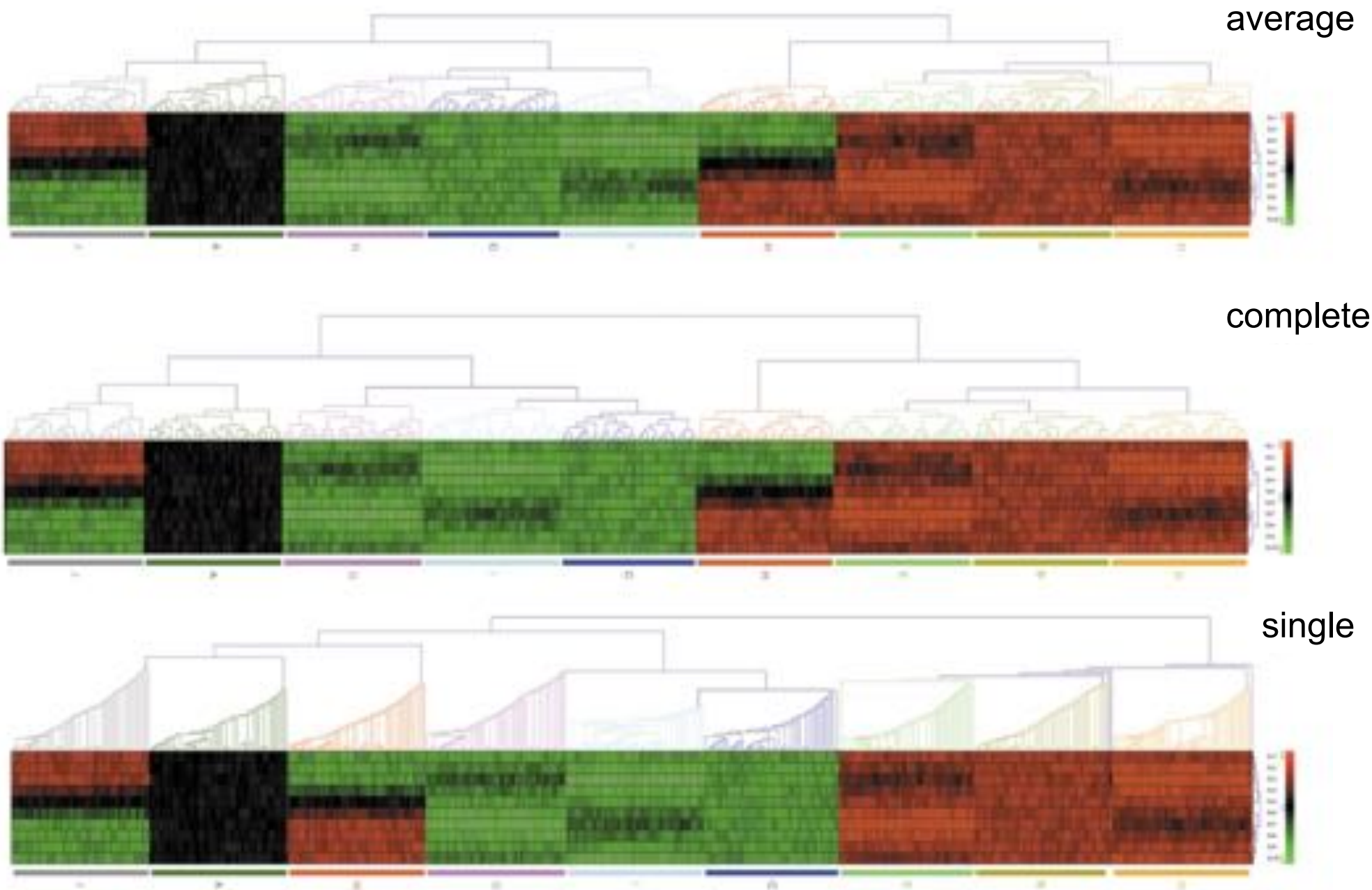




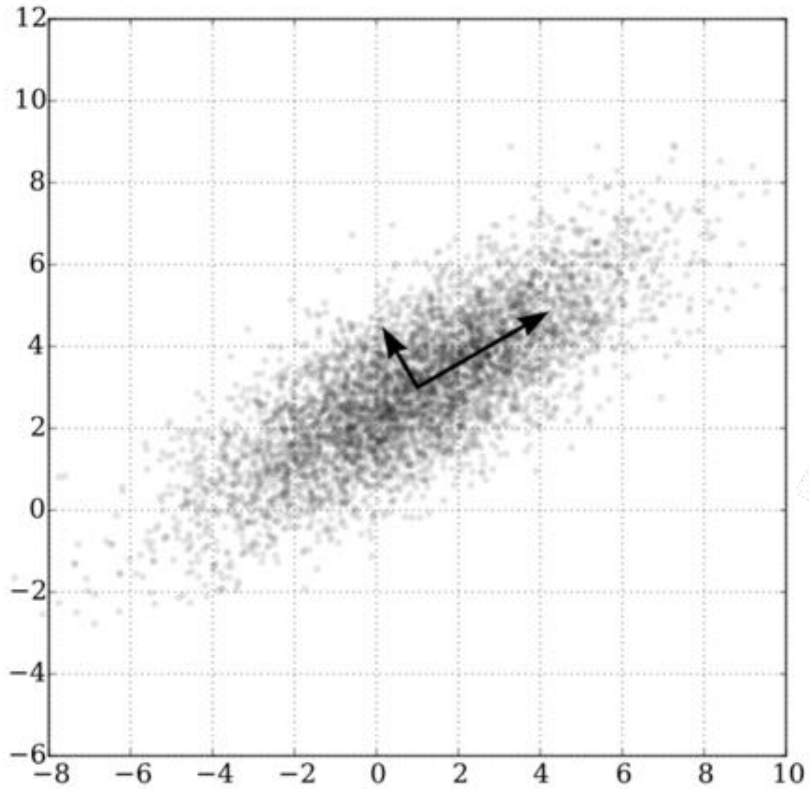
$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

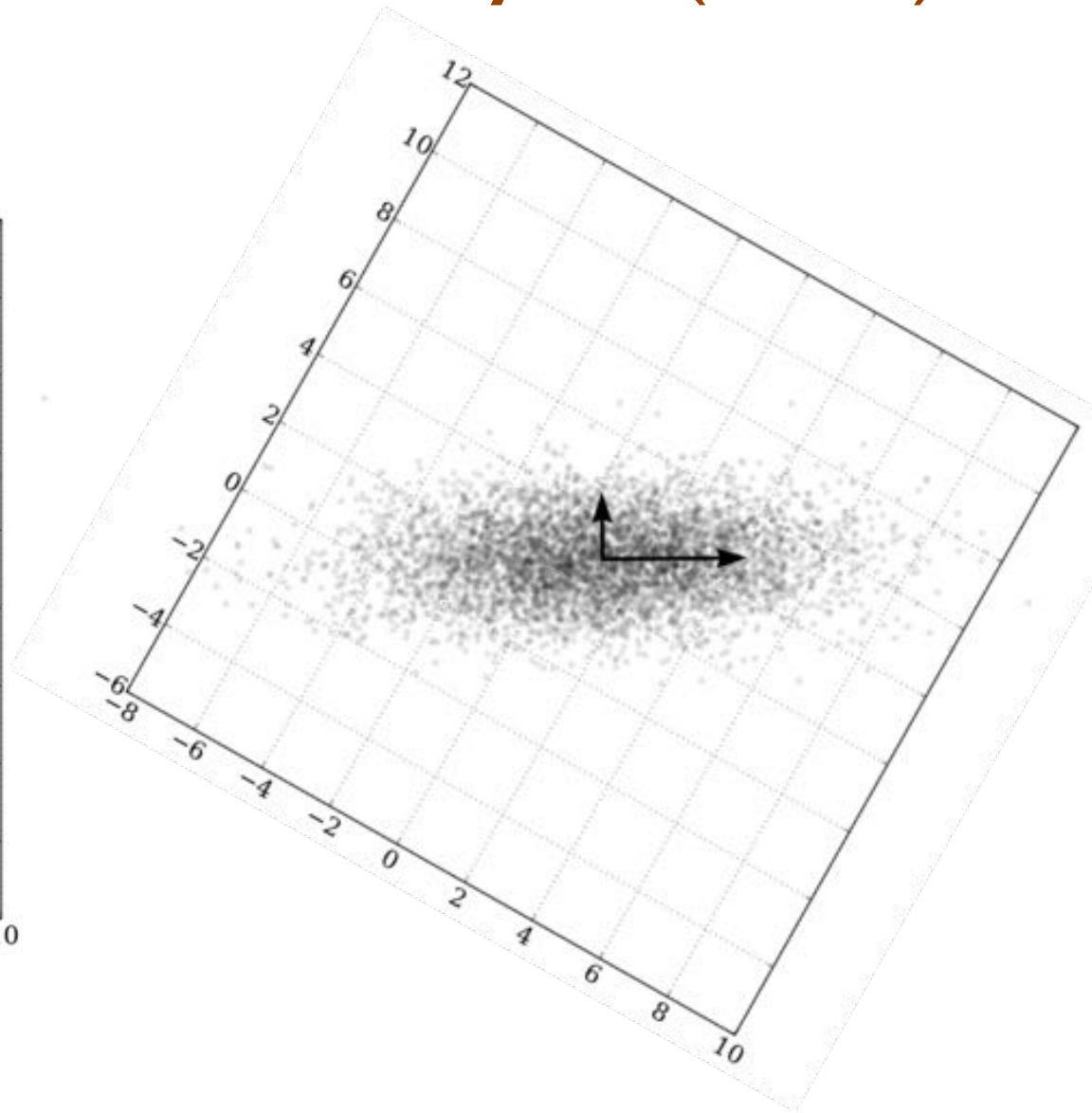
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



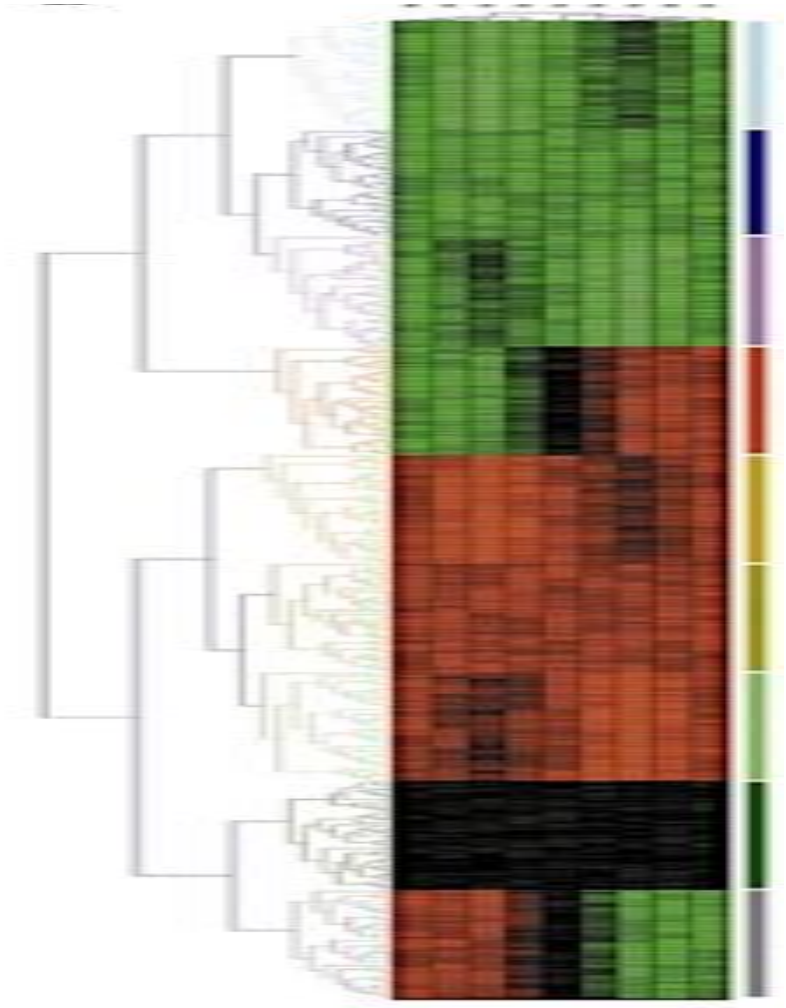
Principle Components Analysis (PCA)

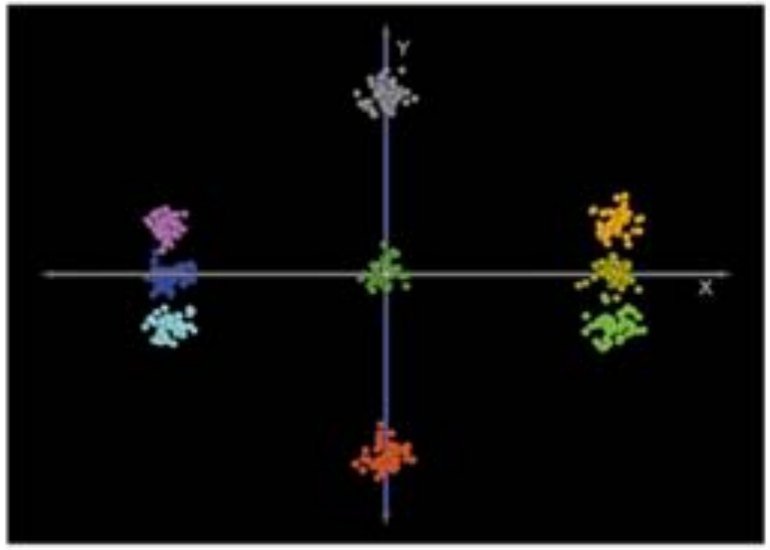




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





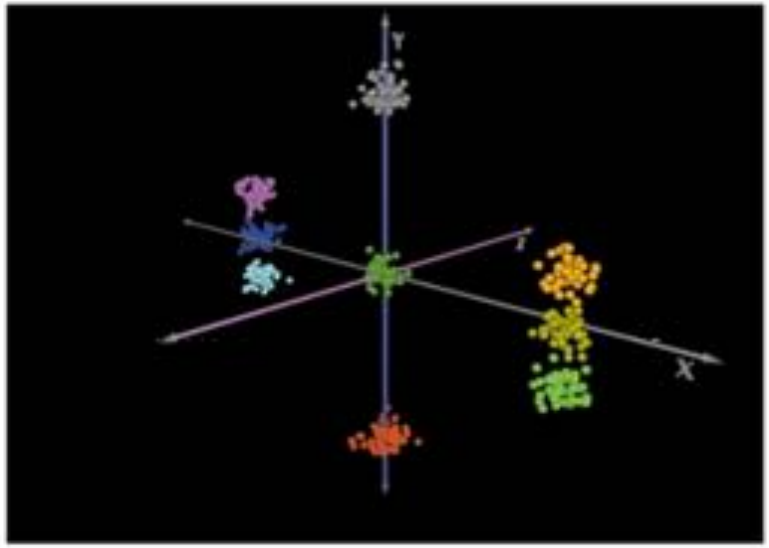
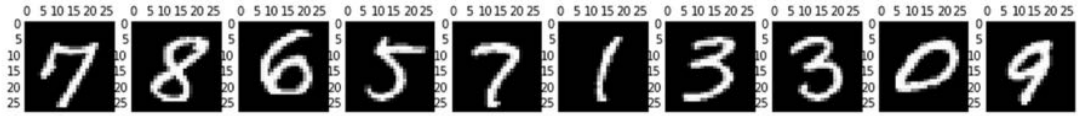
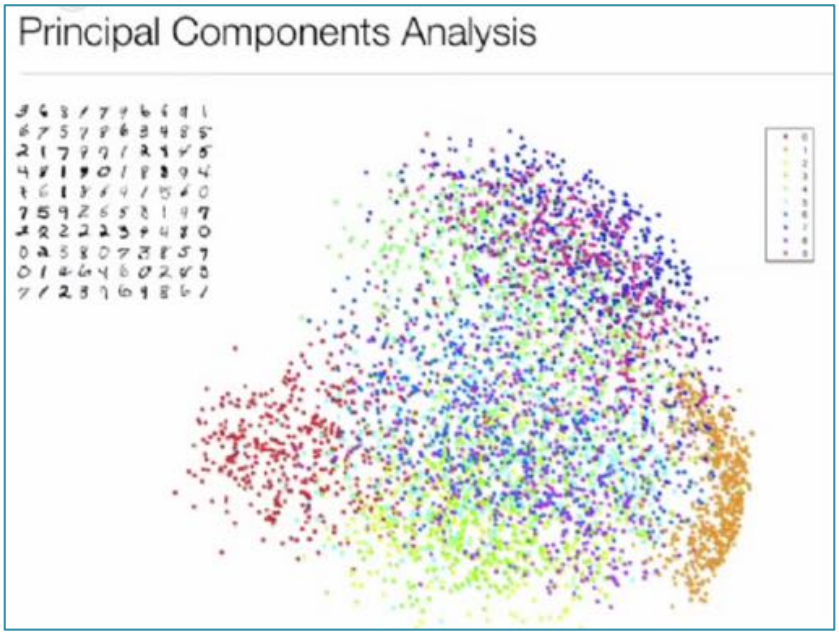
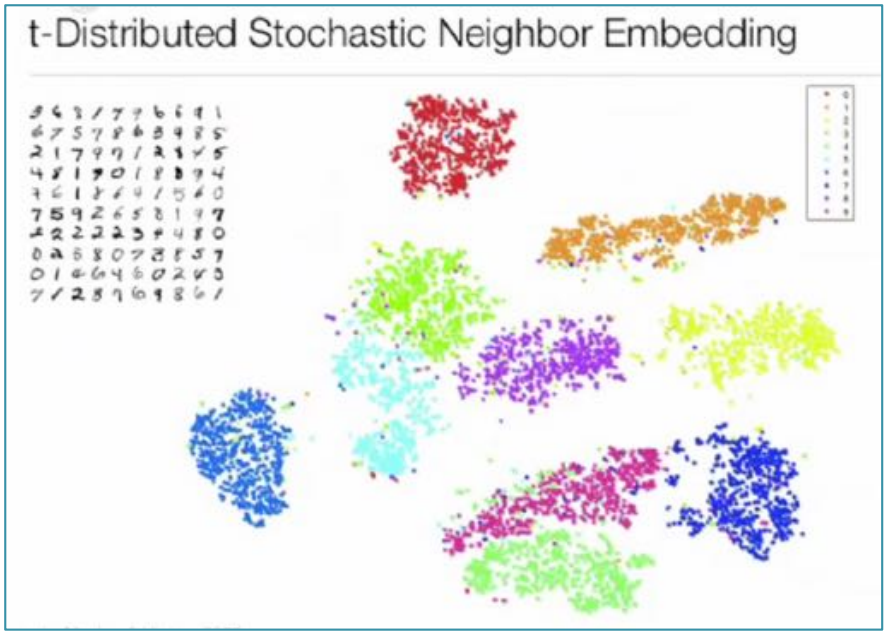


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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

Single Cell Analysis

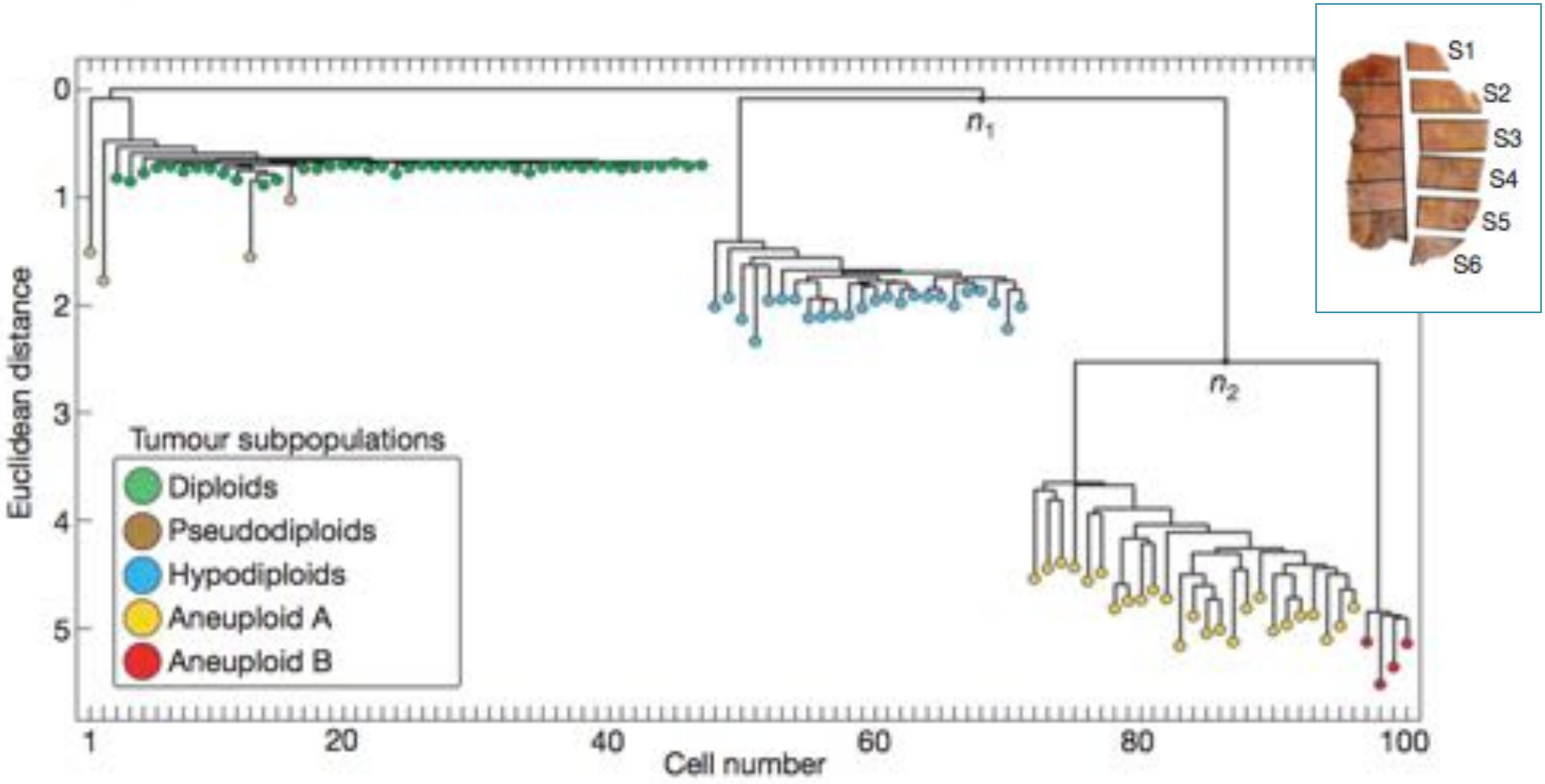
1. Why single cells?
2. scDNA
3. scRNA and other assays

LETTER

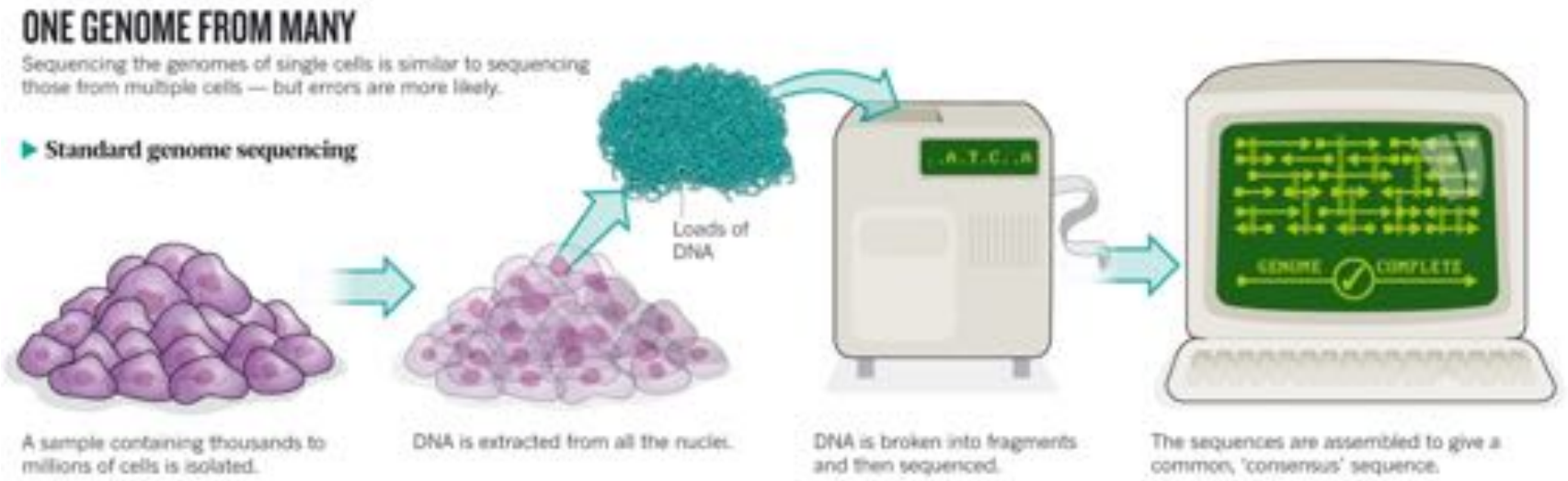
doi:10.1038/nature09807

Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]

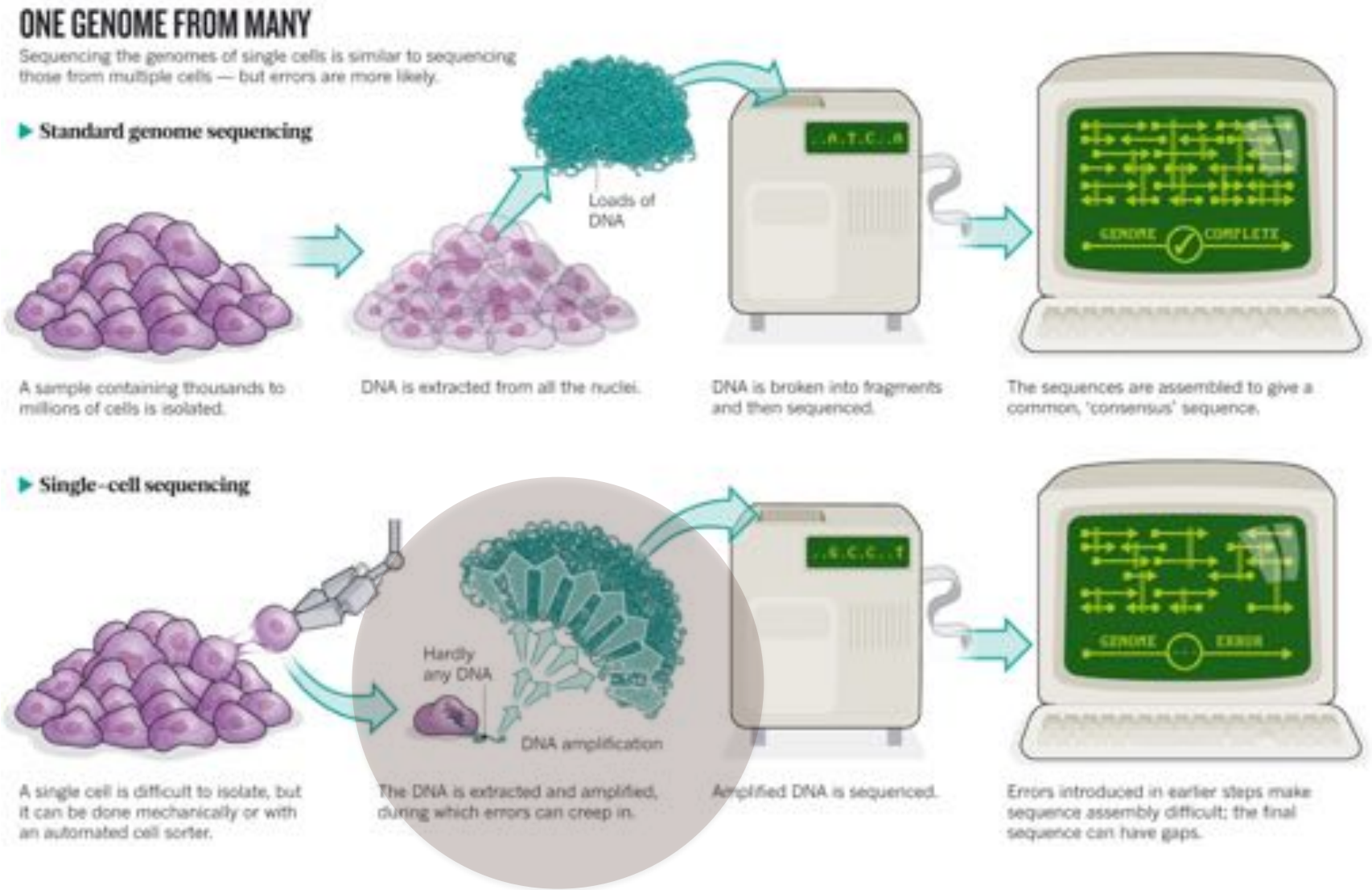


Single-cell vs. bulk sequencing



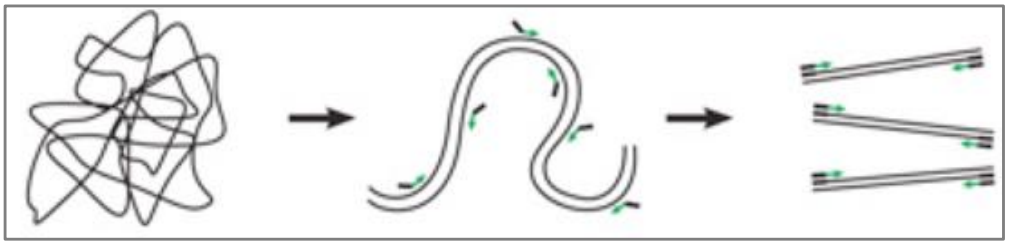
Brian Owens, Nature News 2012

Single-cell vs. bulk sequencing

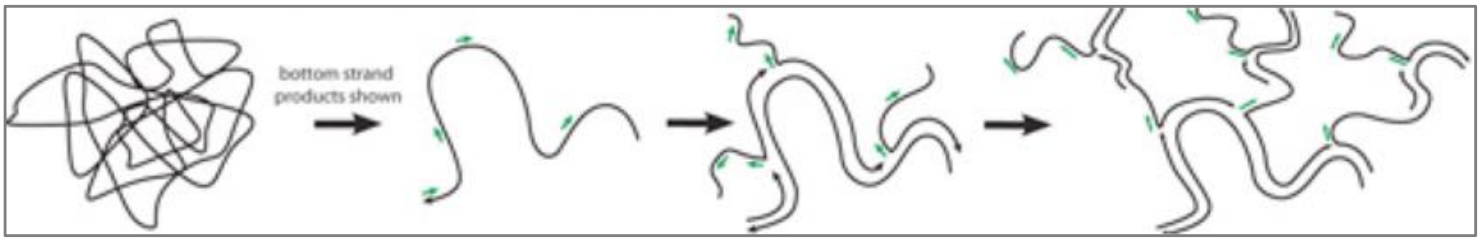


Brian Owens, Nature News 2012

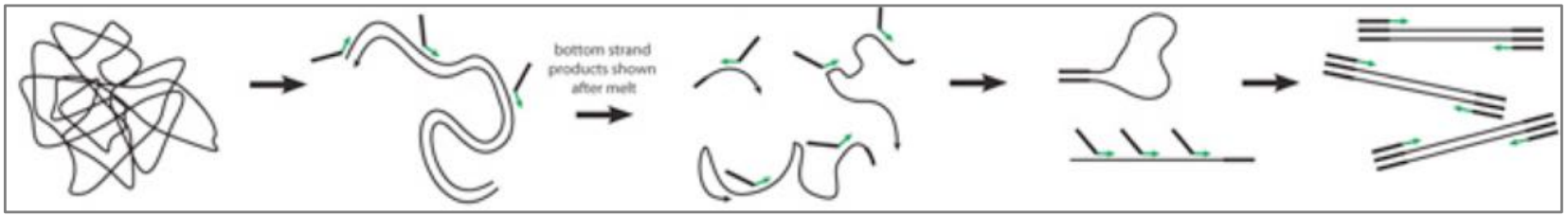
Whole Genome Amplification Techniques



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



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

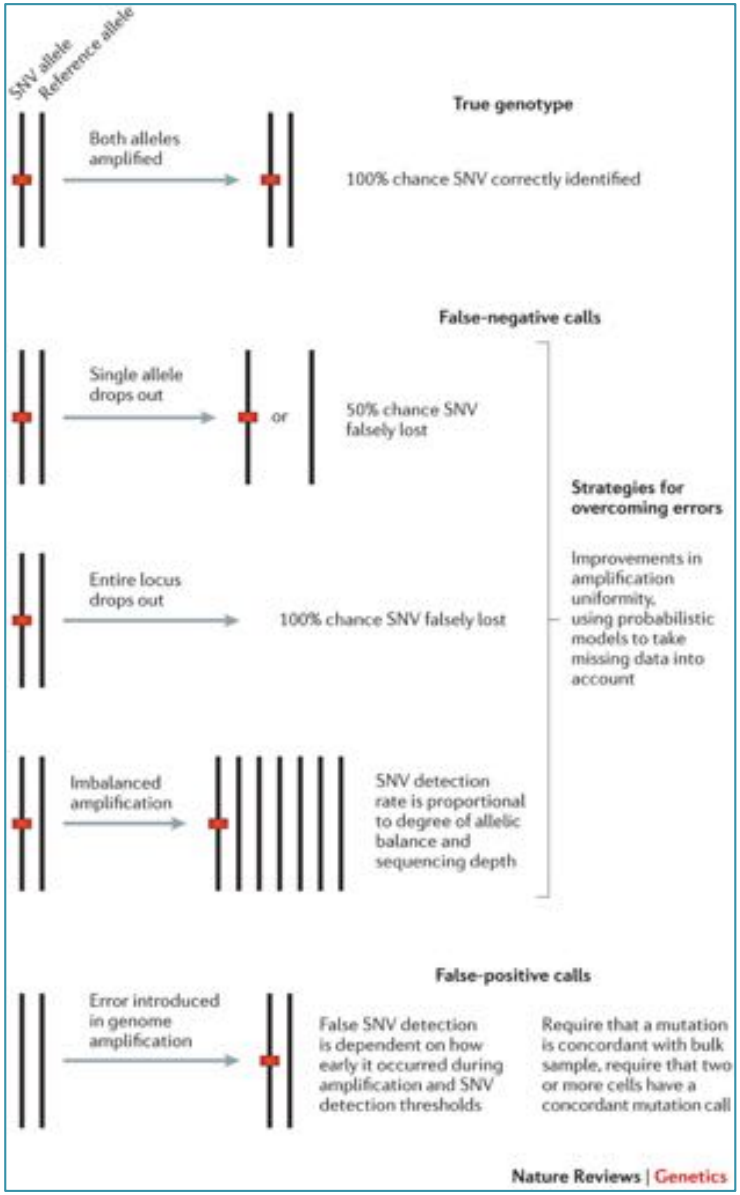


MALBAC: Multiple Annealing and Looping Based Amplification Cycles
Zong et al. (2012) Science

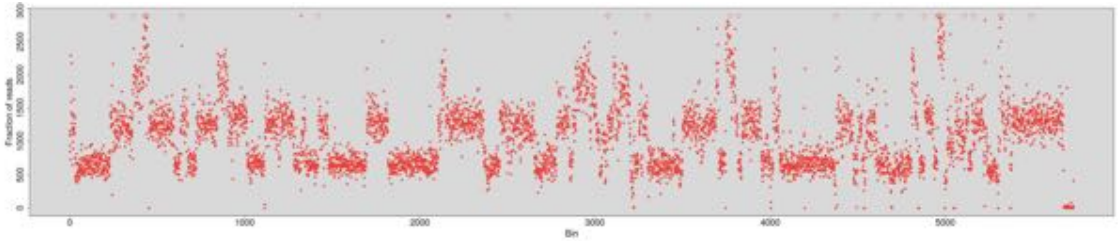
Paul Blainey, FEMS Microbiol Rev. 2013

Fluidigm C1

Benchtop automated single-cell isolation and preparation system(lysis and pre-amplification) for genomic analysis. The C1 System provides an easy and highly reproducible workflow to process ***96 single cells*** for DNA or RNA analysis.



scCNVs



Potential for biases at every step

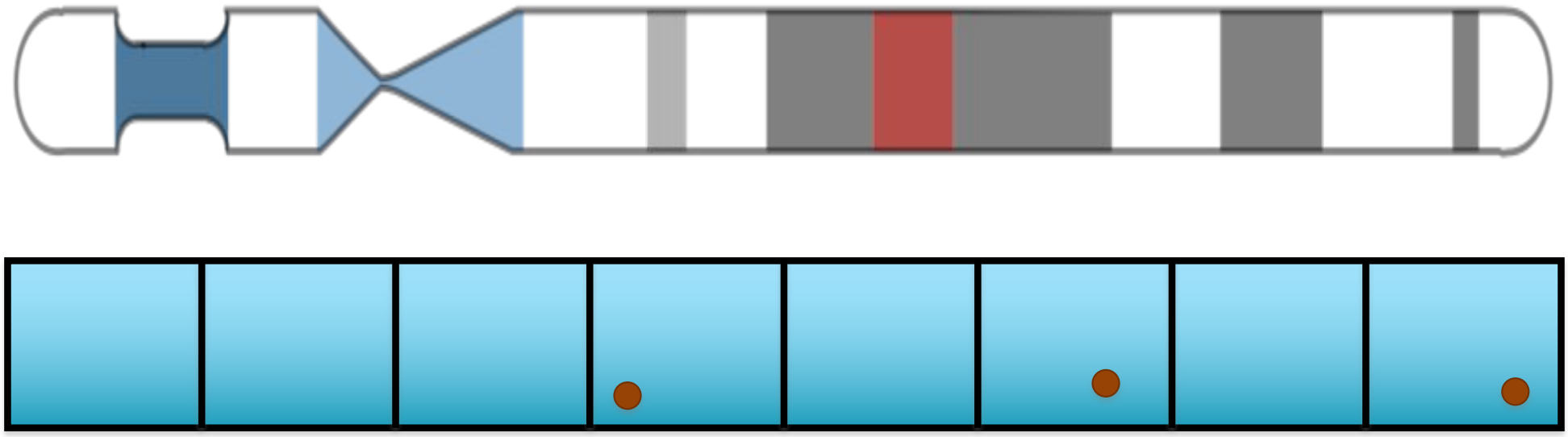
- WGA: Non-uniform amplification
- Library Preparation: Low complexity, read duplications, barcoding
- Sequencing: GC artifacts, short reads
- Computation: mappability, GC correction, segmentation, tree building

Coverage is very sparse and noisy

-> requires special processing

Single-cell genome sequencing: current state of the science
Gawad et al (2016) Nature Reviews Genetics. doi:10.1038/nrg.2015.16

1) Binning

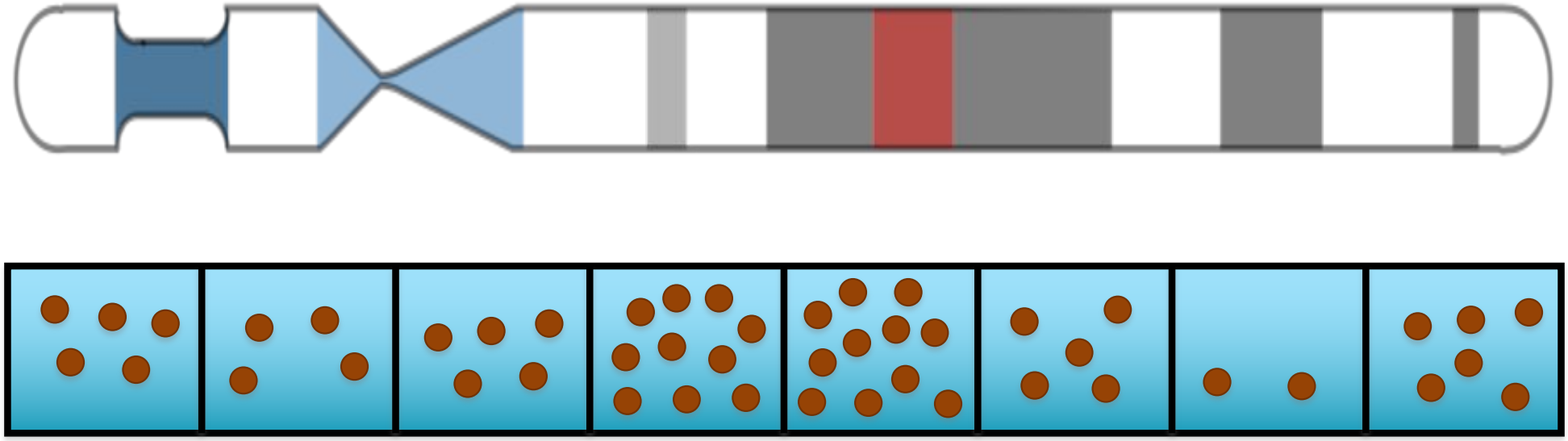


Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1) Binning



Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1) Binning

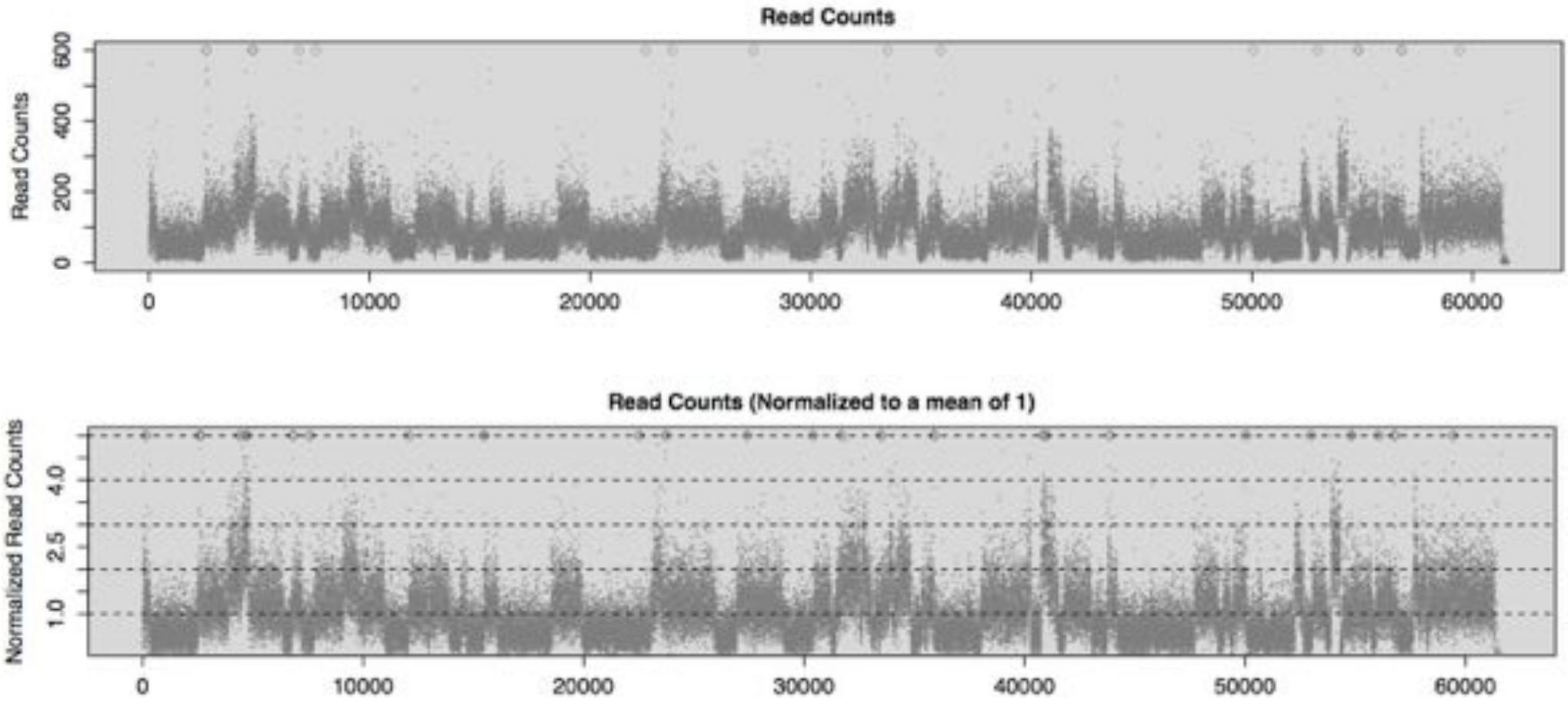
5	4	5	10	11	5	2	5

Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

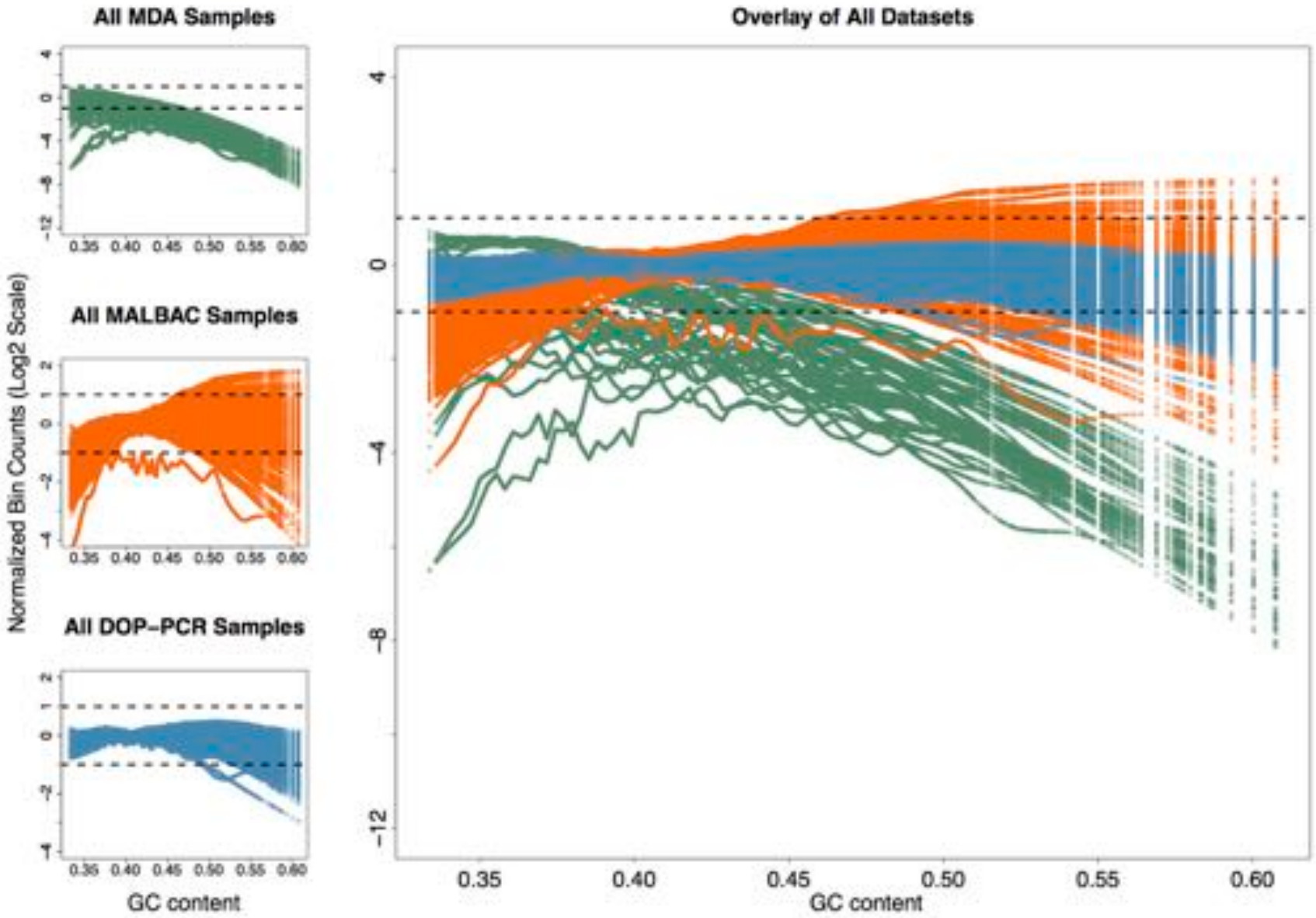
Use uniquely mappable bases to establish bins

2) Normalization



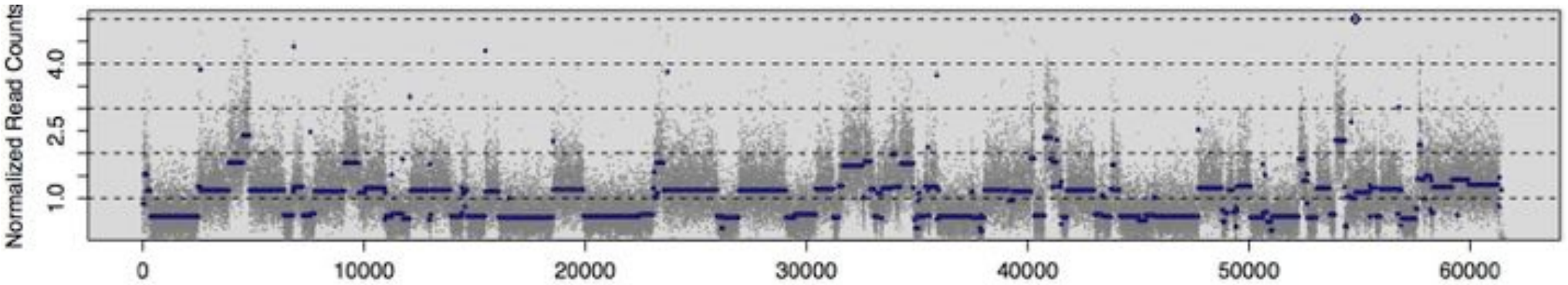
Also correct for mappability, GC content, amplification biases

GC Bias

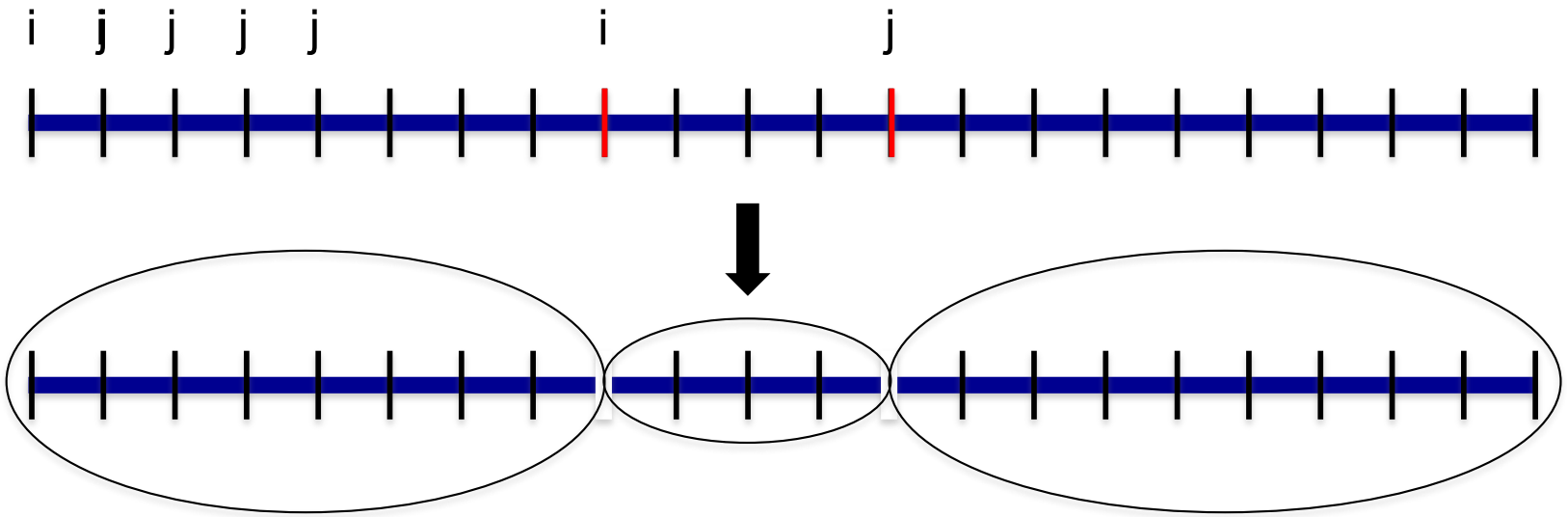


Garvin and Aboukhalil et al., Nature Methods, 2015

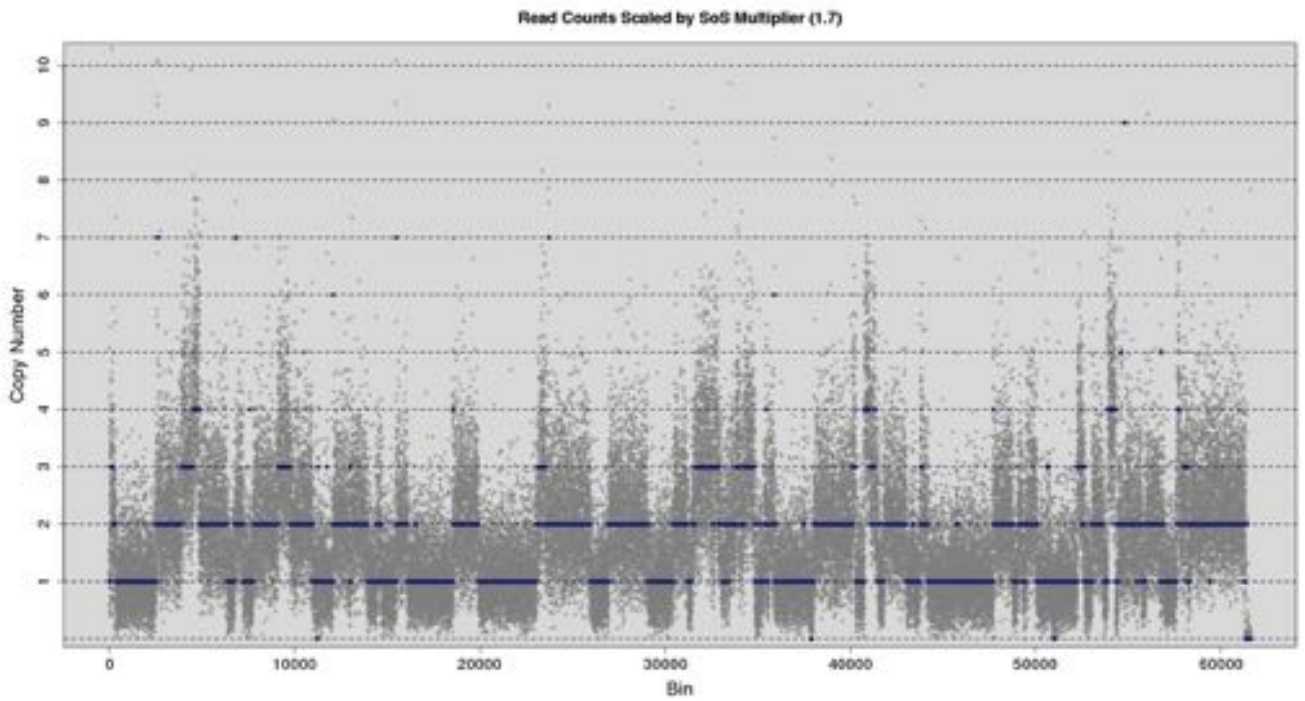
3) Segmentation



Circular Binary Segmentation (CBS)

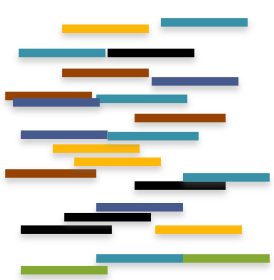


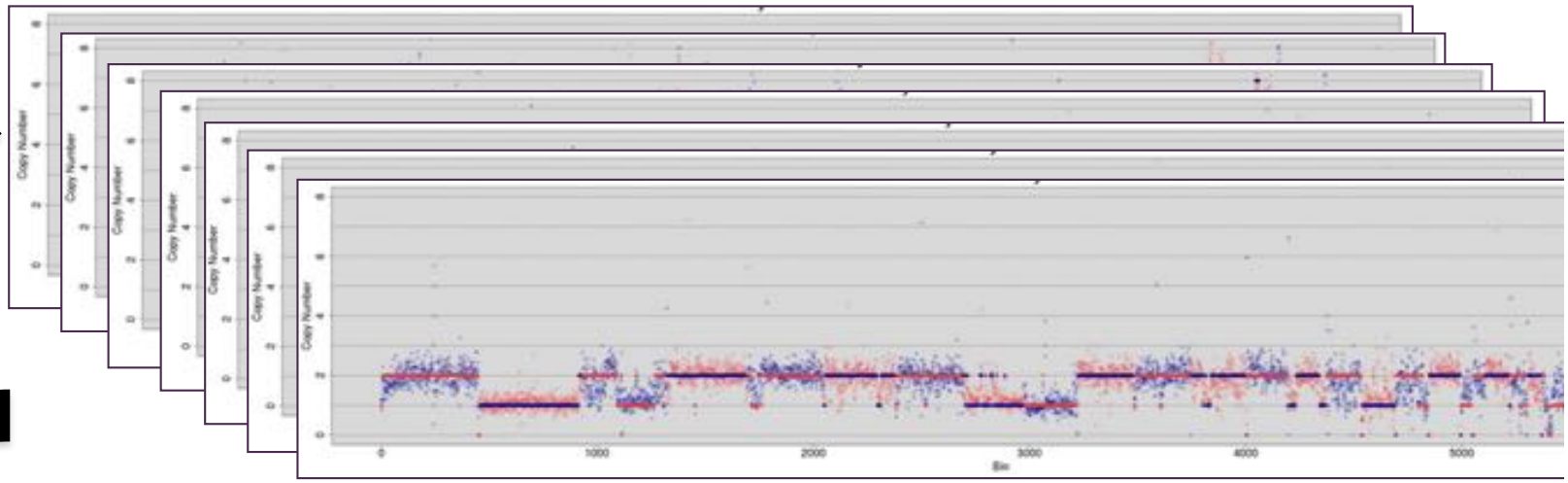
4) Estimating Copy Number

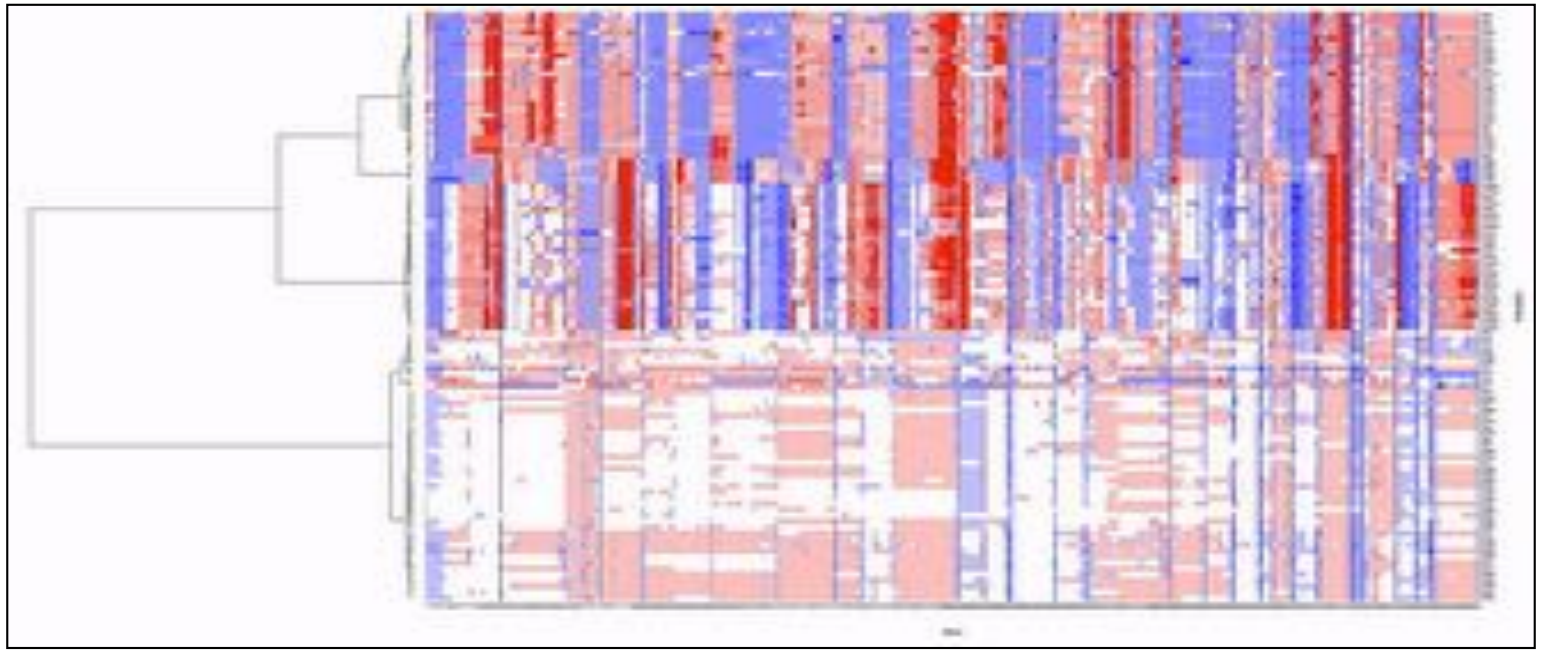


$$CN = argmin\left\{\sum_{i,j}(\hat{Y}_{i,j} - Y_{i,j})^2\right\}$$

5) Cells to Populations

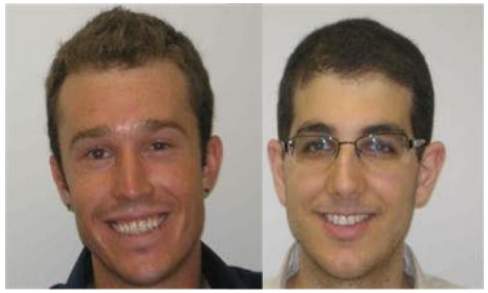


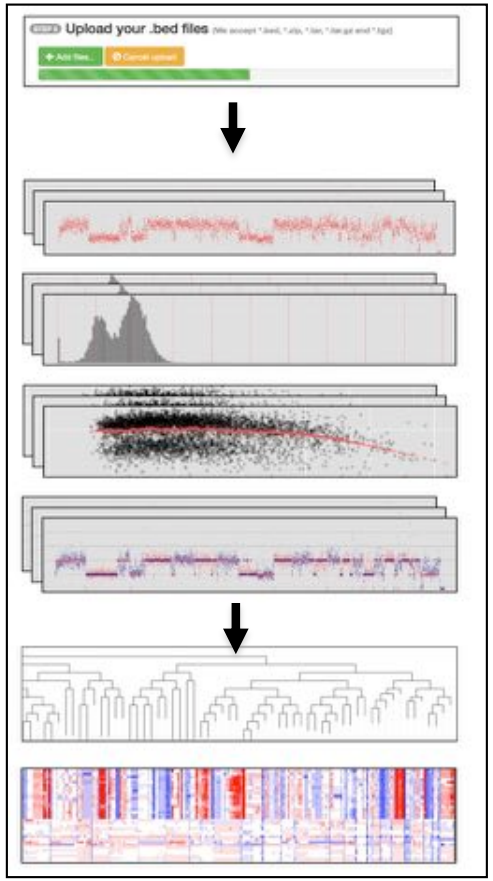




Gingko

http://qb.cshl.edu/ginkgo





Interactive Single Cell CNV analysis & clustering

- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

Compare MDA, DOP-PCR, and MALBAC

- DOP-PCR shows superior resolution and consistency

Available for collaboration

- Analyzing CNVs with respect to different clinical outcomes
- Extending clustering methods, prototyping scRNA

Interactive analysis and assessment of single-cell copy-number variations.
Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC (2015)
Nature Methods doi:10.1038/nmeth.3578

10x GENOMICS®

scCNV-Seq

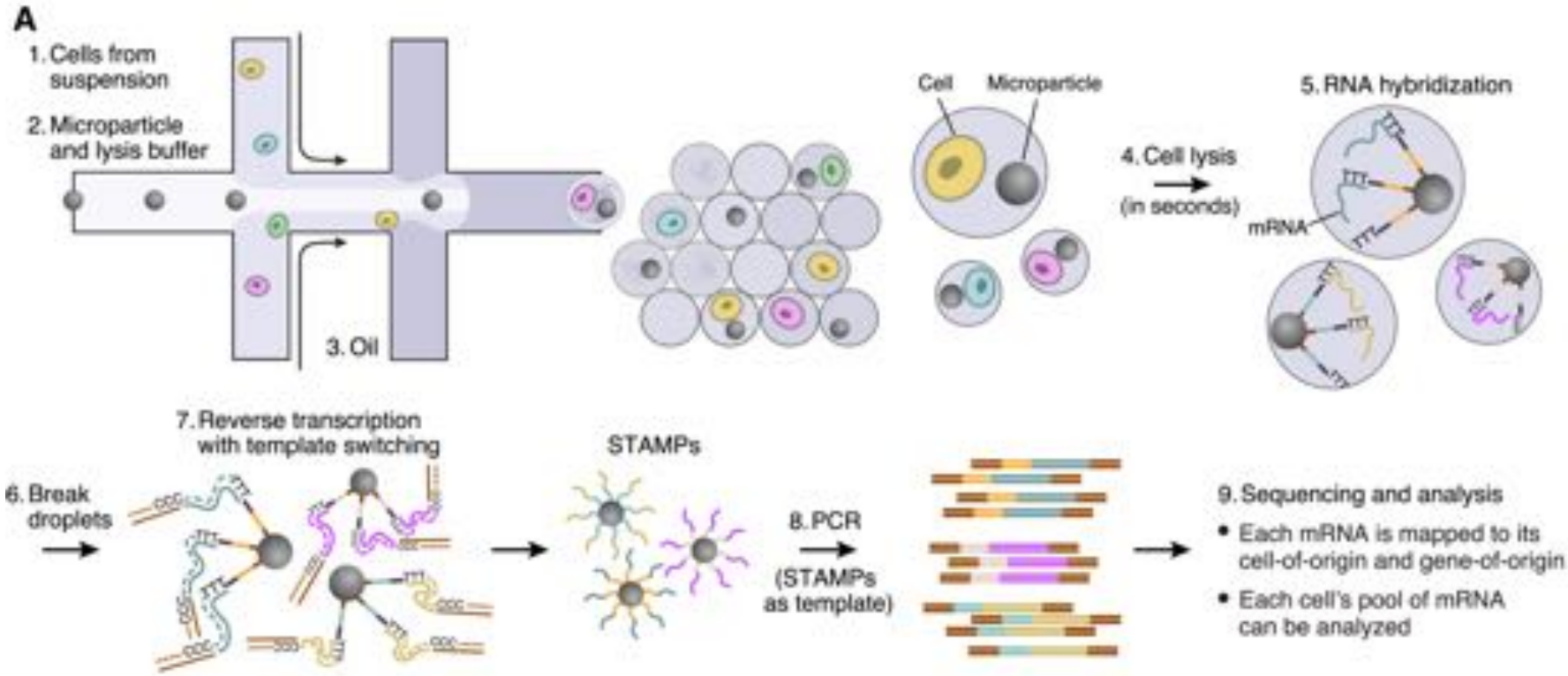
Single Cell CNV-Seq

- Reveal genomic heterogeneity
- Understand clonal evolution
- Determine pathogenesis and cancer progression
- Scalable from 100s-1000s of cells
- Single-cell CNV calling
- Call CNVs down to 100kb resolution
- CNV-Seq specific software pipeline

Single Cell Analysis

1. Why single cells?
2. scDNA
3. scRNA and other assays

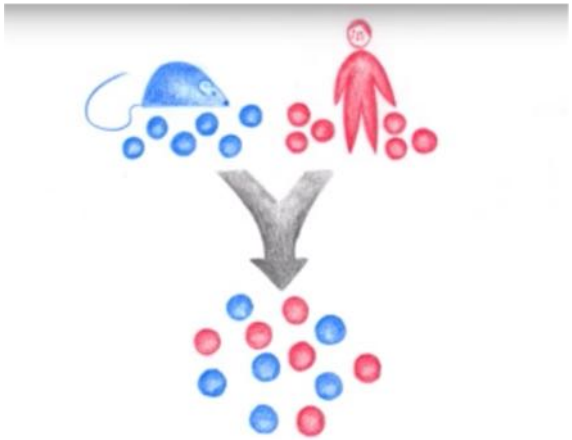
Introducing Drop-Seq

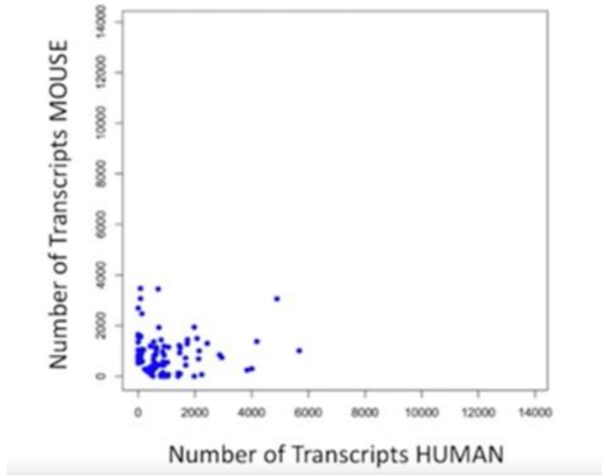


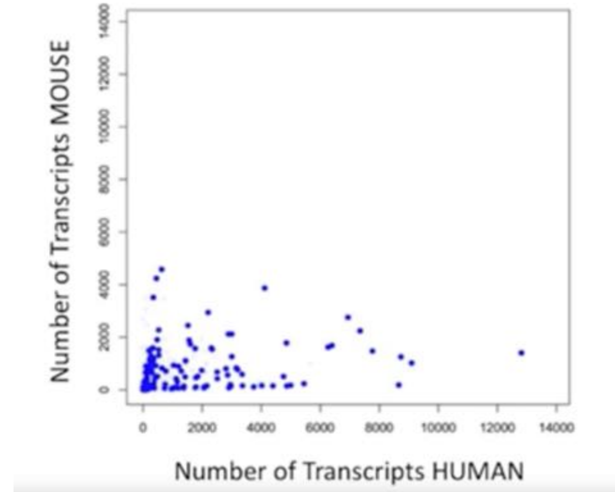
Many of the same technical challenges apply:

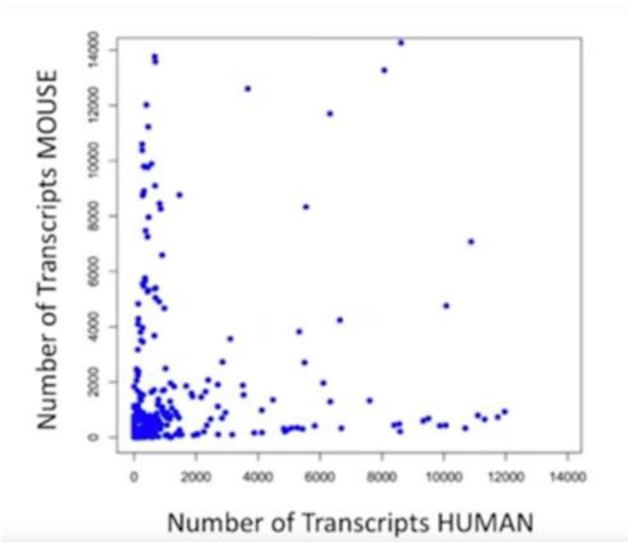
- Allelic dropout, non-uniform amplification, few reads per cell (~50k / cell).
- Remarkably, cell type identity can often be determined with this many reads
- Use statistics to smooth out uneven coverage across cells.

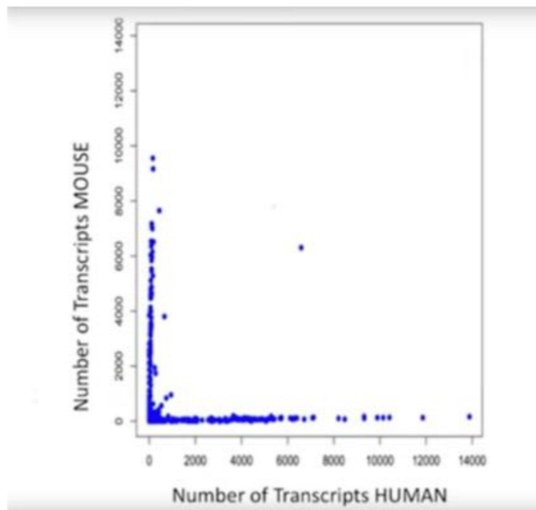
Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets
Macosko et al (2015) Cell. https://doi.org/10.1016/j.cell.2015.05.002

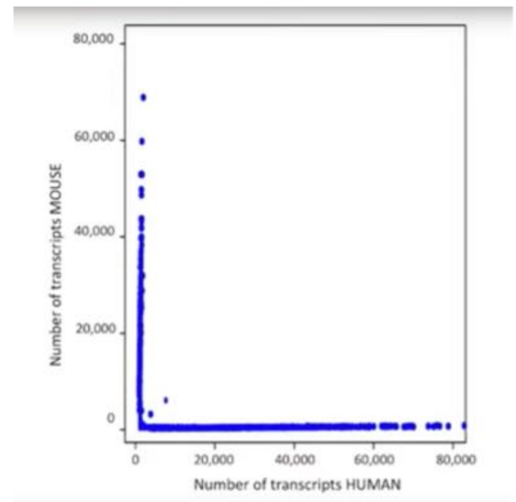




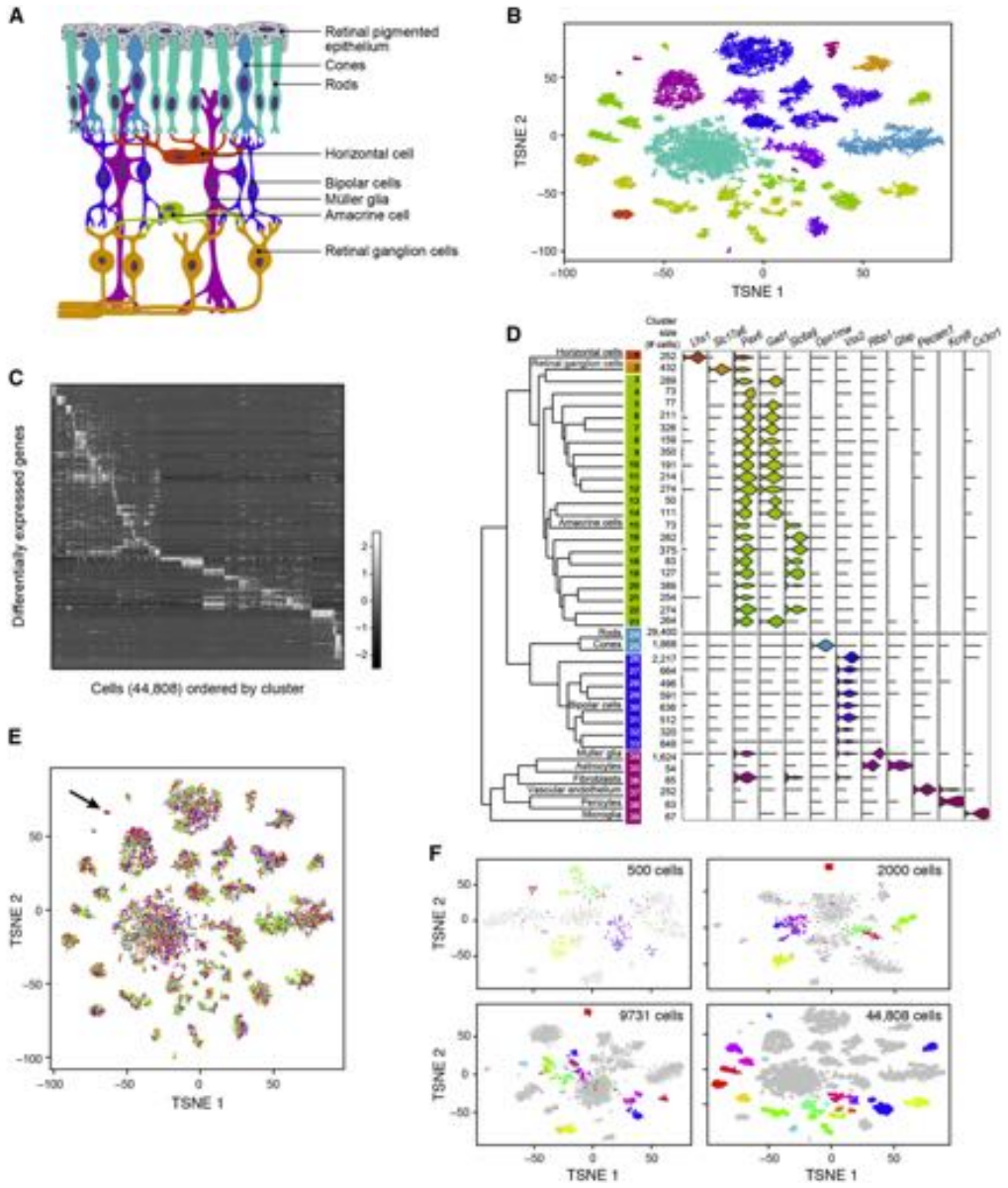








Drop-seq: Droplet barcoding of single cells
https://www.youtube.com/watch?v=vL7ptq2Dcf0



Key Results

(a) schematic of known cell populations in retina

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

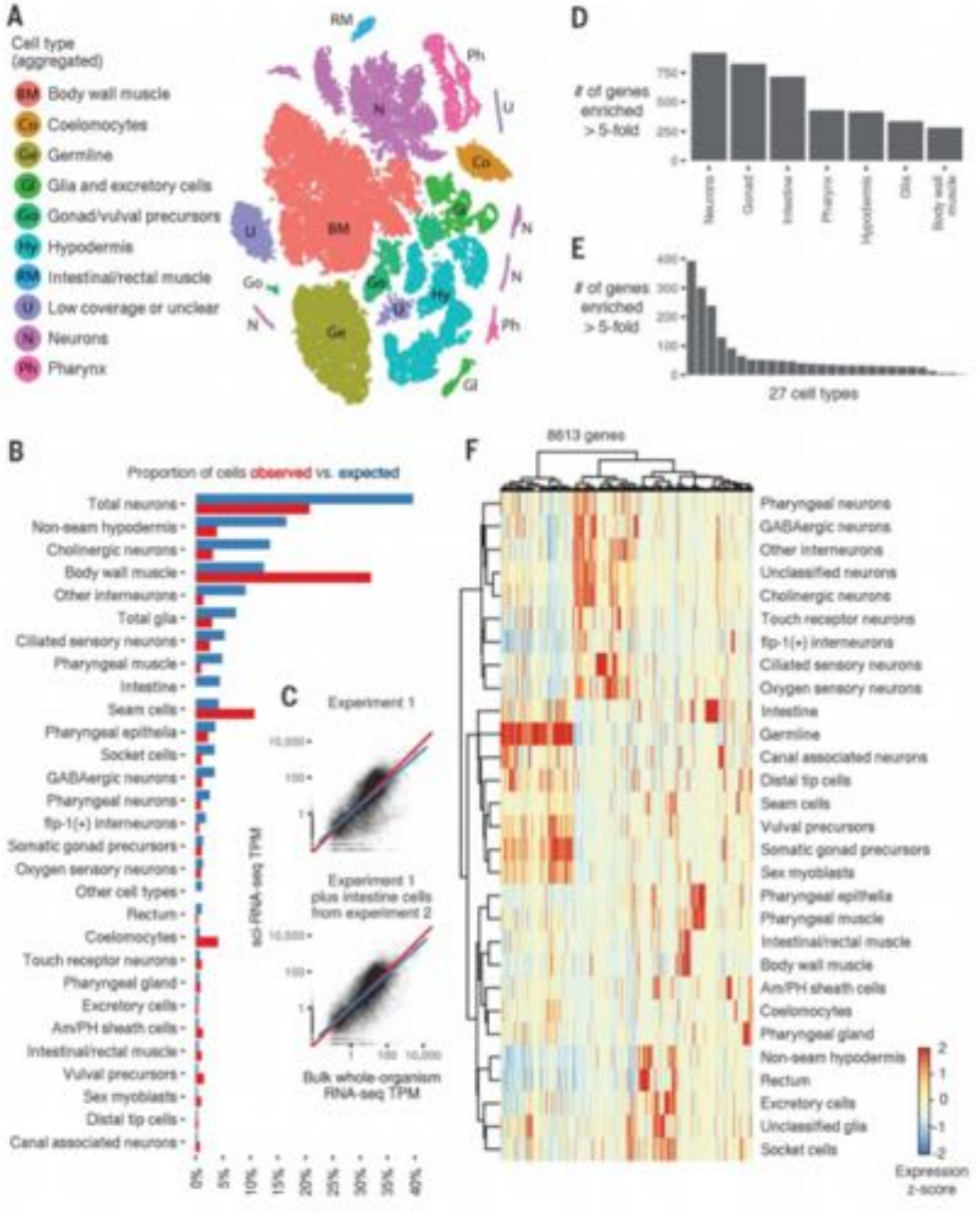
(c) Differentially expressed genes in each cluster

(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

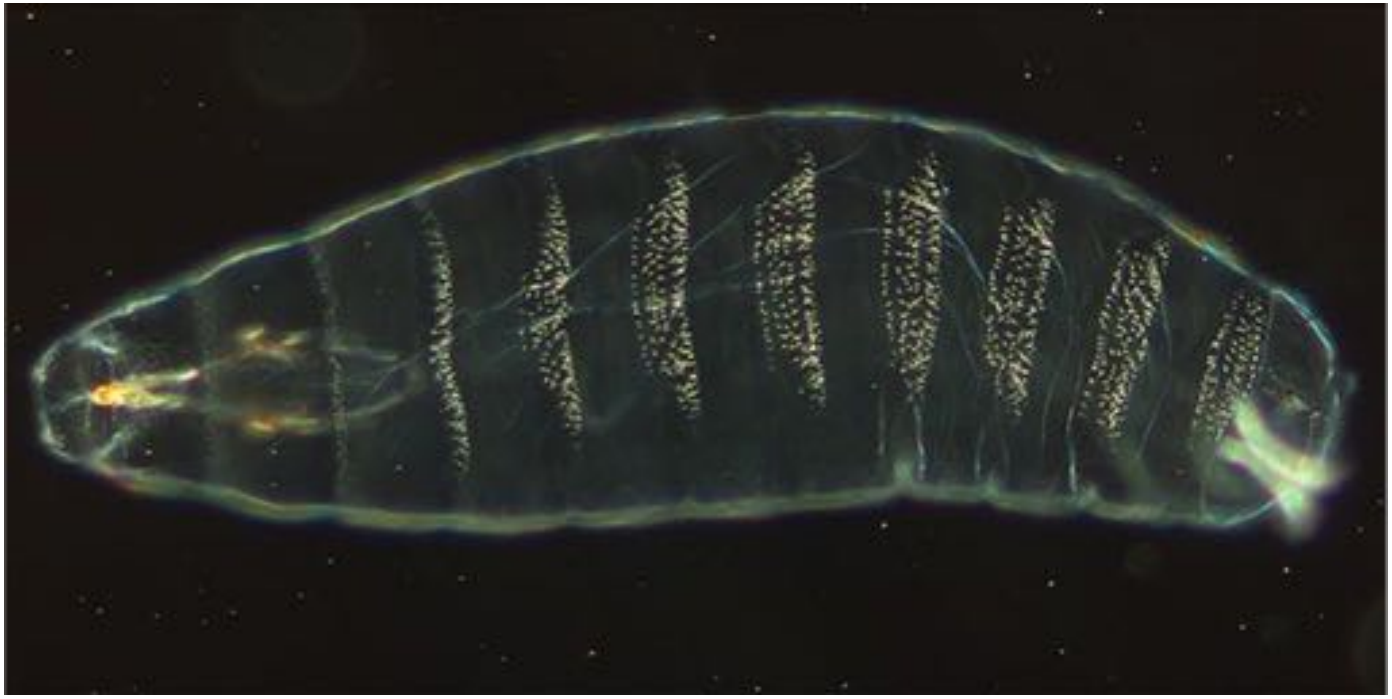
(a) t-SNE visualization of clusters

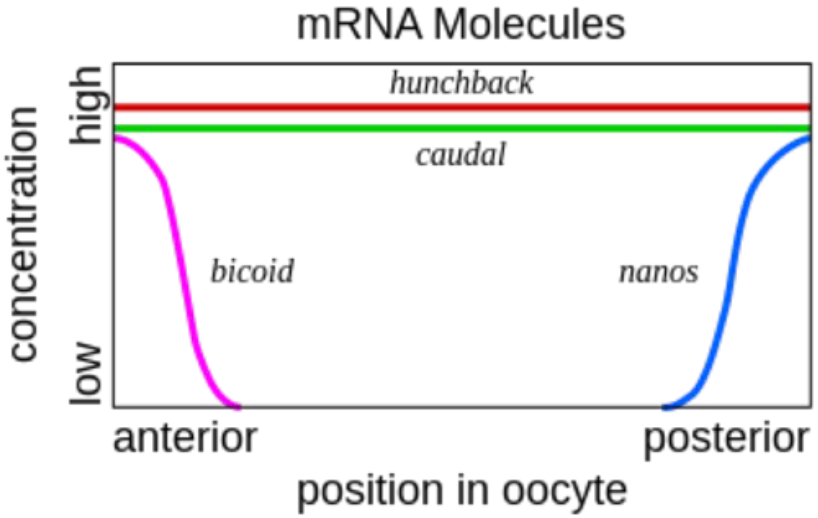
(b) Proportion of cells observed vs expected match well (including cells that only occur once or twice in the animal)

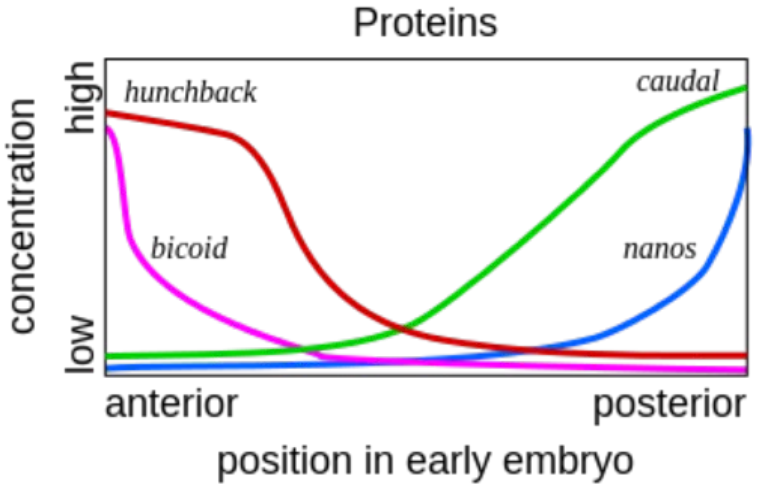
(c) Good correlation between single cell and bulk analysis of selected cell types

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

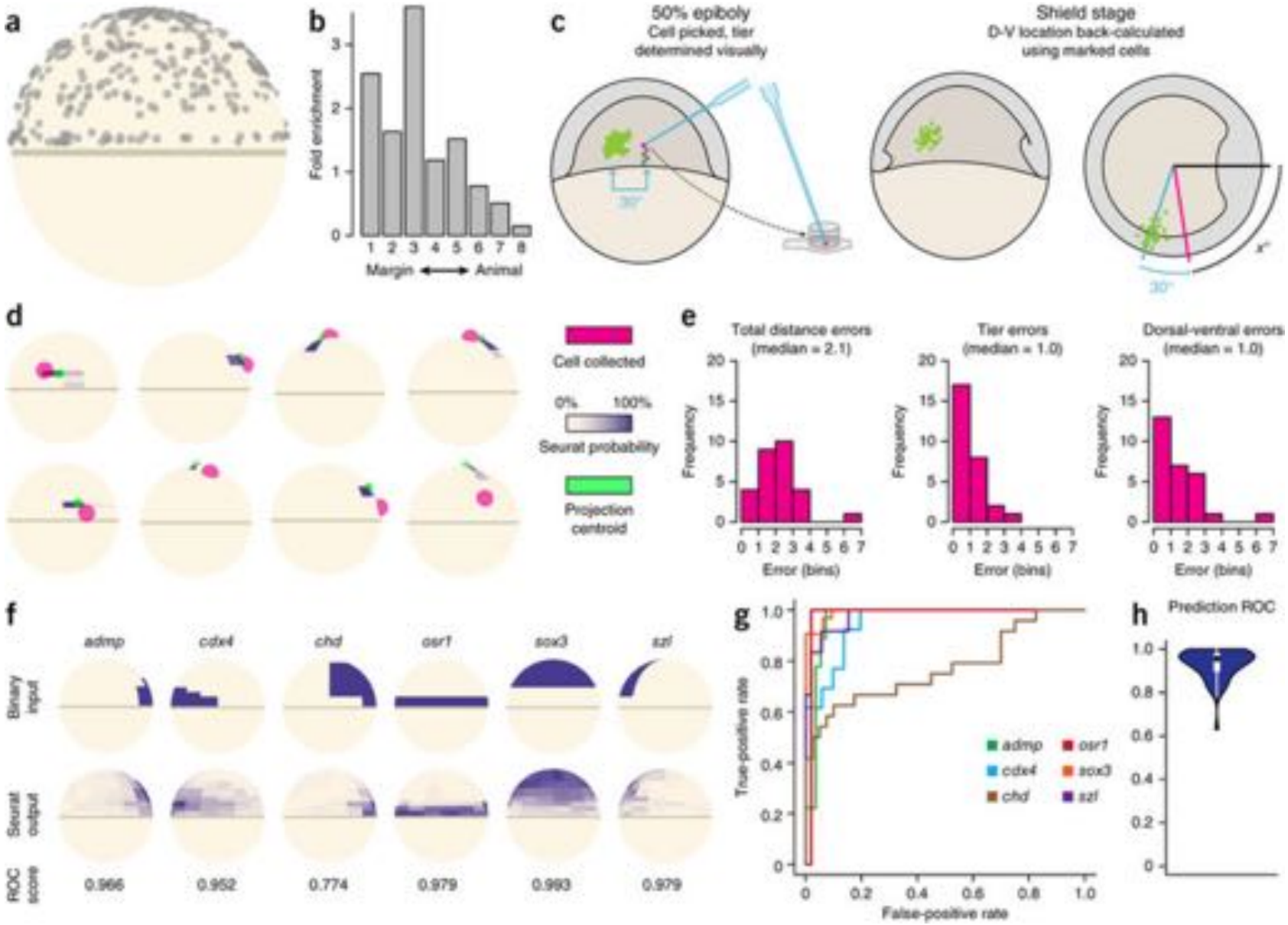
Comprehensive single-cell transcriptional profiling of a multicellular organism
Cao et al (2017) Science. 357:661-557



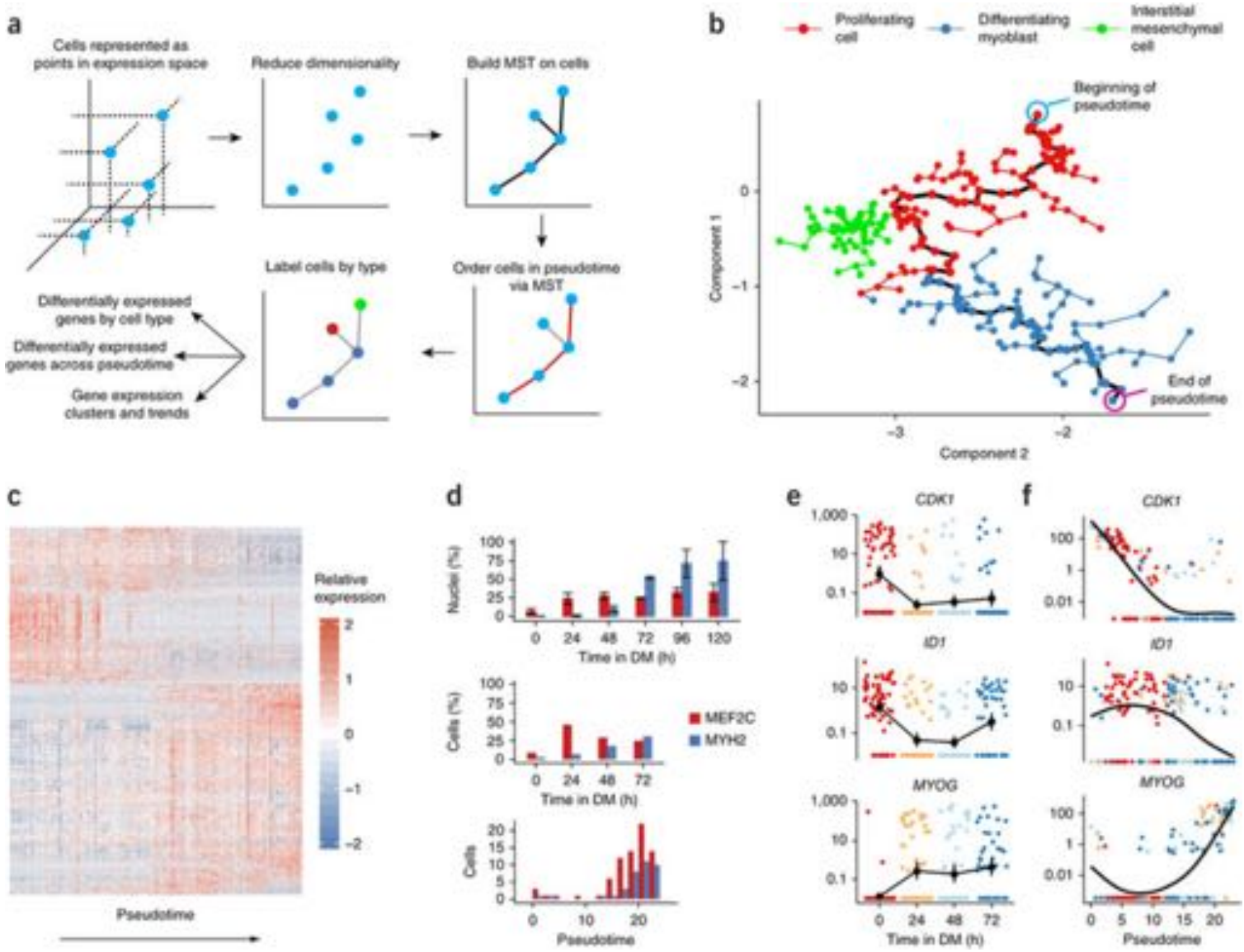




https://en.wikipedia.org/wiki/Drosophila_embryogenesis



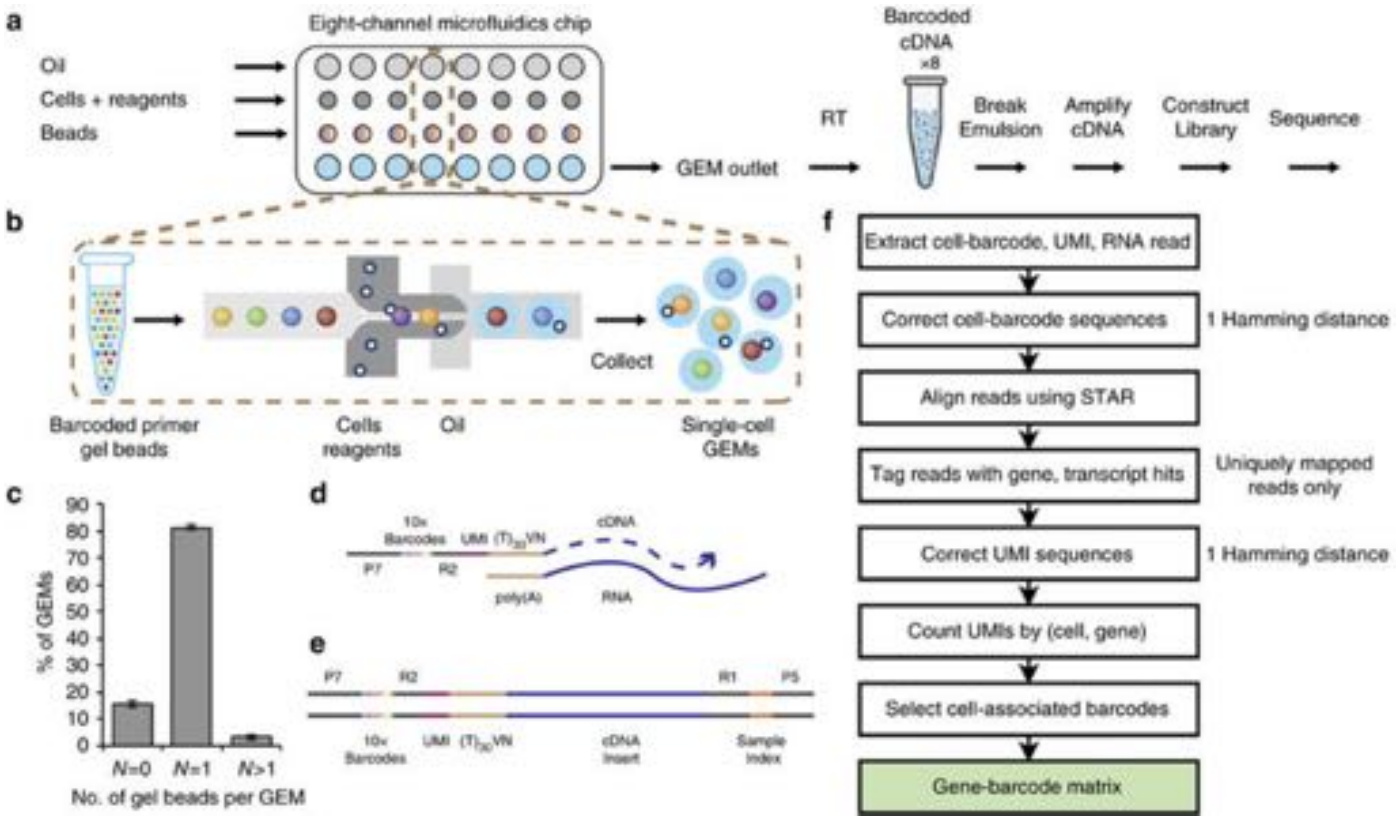
Spatial reconstruction of single-cell gene expression data ("Seurat")
Satija et al (2015) Nature Biotechnology. doi:10.1038/nbt.3192



The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells ("Monocle")

Trapnell et al (2014) Nature Biotechnology. doi:10.1038/nbt.2859

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

10X GENOMICS® scATAC-Seq

Single Cell ATAC-Seq

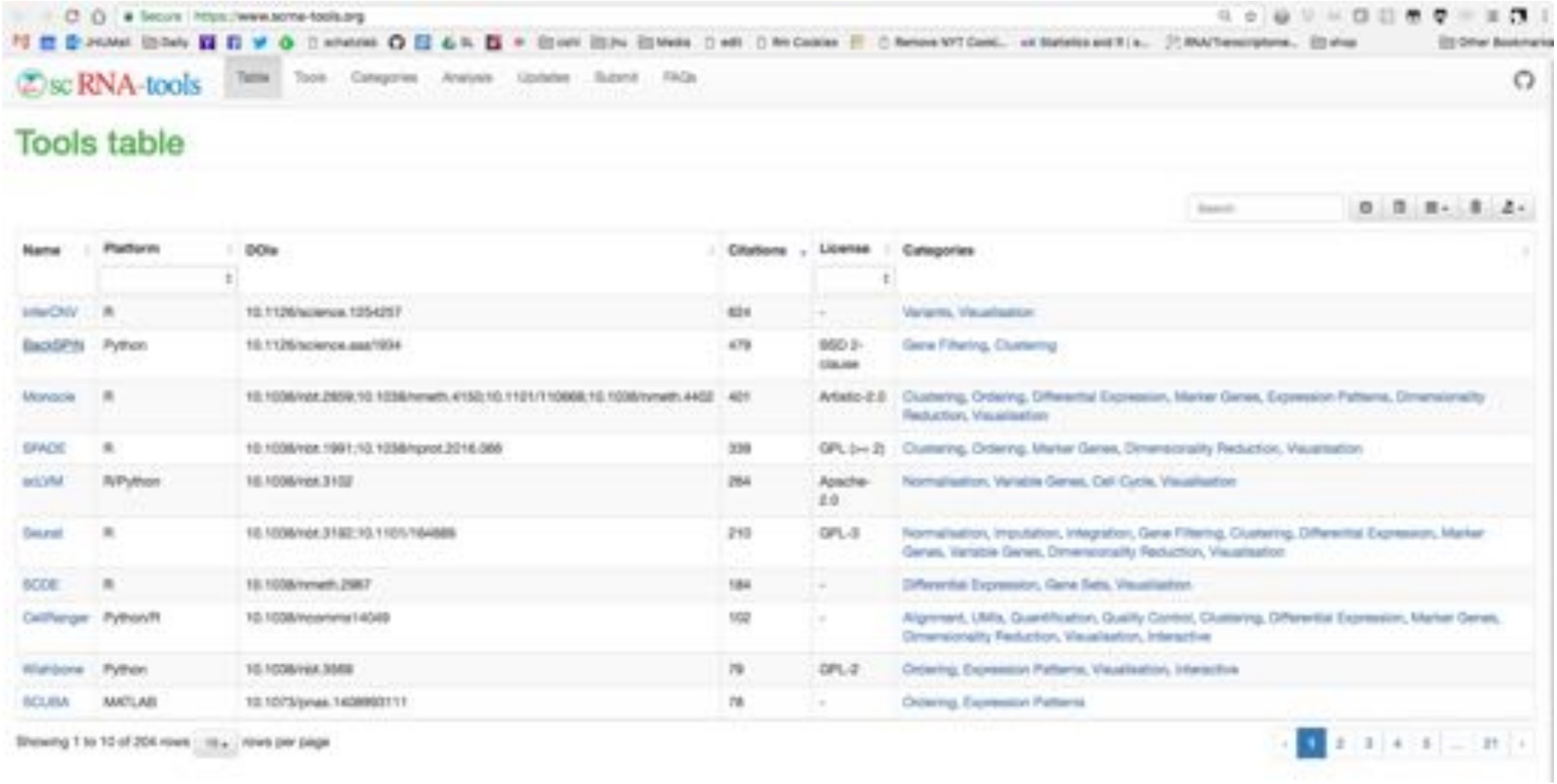
- Interrogate epigenomics at single-cell resolution
- Define cell types and states
- Investigate regulatory mechanisms
- Scalable from 1000s of cells
- High cell capture efficiency
- High transpososome capture sensitivity
- ATAC-Seq specific software pipeline

10x GENOMICS® sc-Feature Barcoding

Single Cell Feature Barcoding

- Reveal protein abundance and gene expression from the same cell
- Understand diverse CRISPR perturbations at single-cell level
- Feature barcoding reagents and protocols
  - Custom antibody conjugation
  - Preferred partners for pre-conjugated antibodies
- Scalable from 100s-1000s of cells
- Interactive visualization in Loupe cell browser
- CNV-Seq specific software pipeline

scRNA Analysis Tools: 204 and counting….



Single Cell Analysis Summary

Single cell analysis is a powerful tool to study heterogeneous tissues

- Overcomes fundamental problems that can arise when averaging
- scCNV analysis used for understanding tumor progression, other mutational processes
- 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 100s to 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