Lecture 16. Single Cell Analysis

Michael Schatz

March 25, 2020

JHU 601.749: Applied Comparative Genomics

Project Proposal: Due Mon Mar 23

Project Proposal

Assignment Date: Monday March 9, 2020
Due Date: Monday, March 16 2020 @ 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:

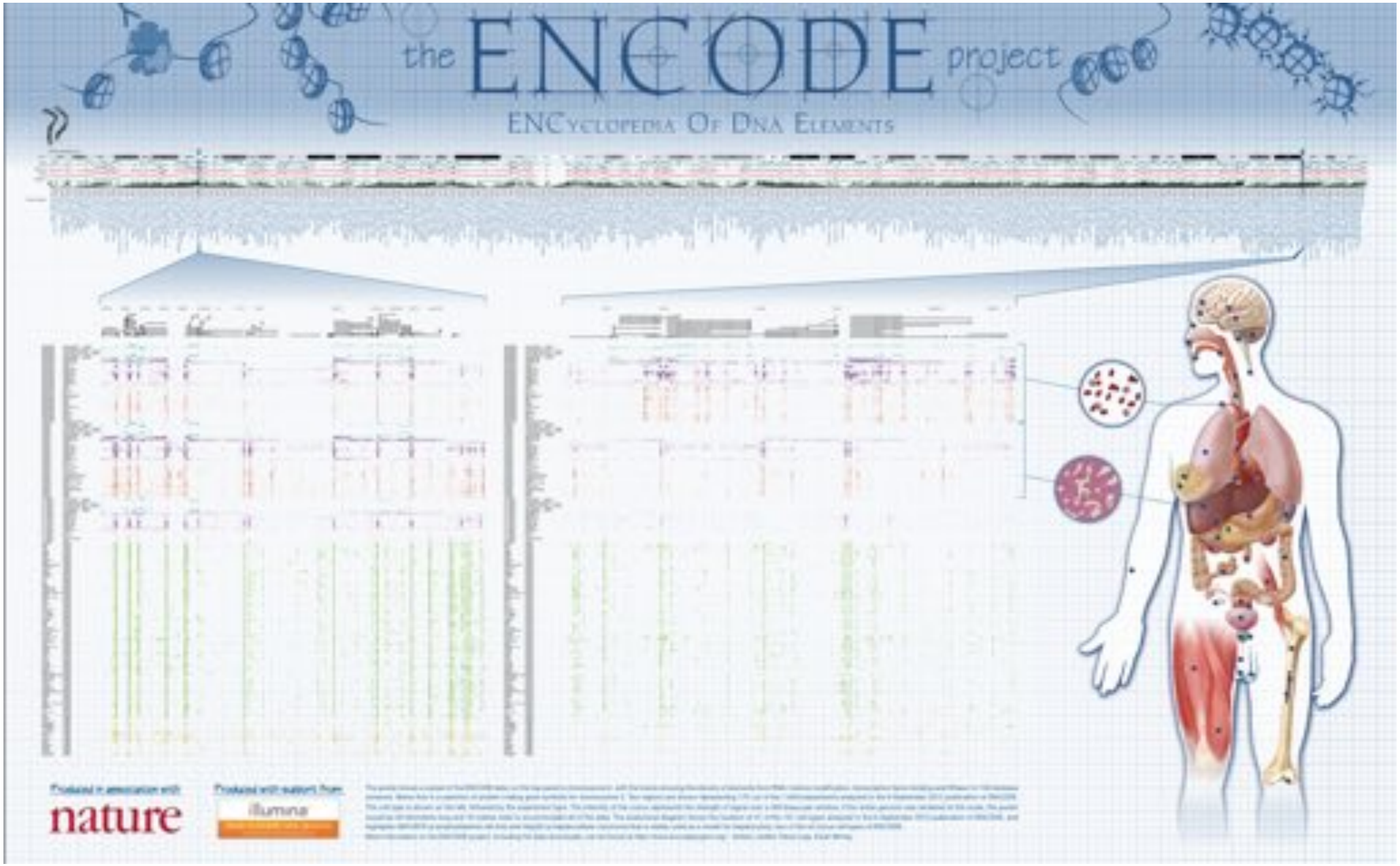
- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)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

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 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!

ENCODE Data Sets



1,640 data sets total over 147 different cell types

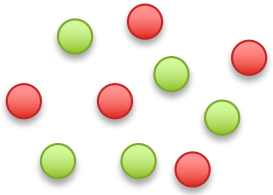
Single Cell Analysis

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

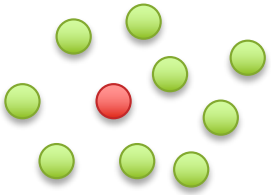
Experiment 1: 50/50



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

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

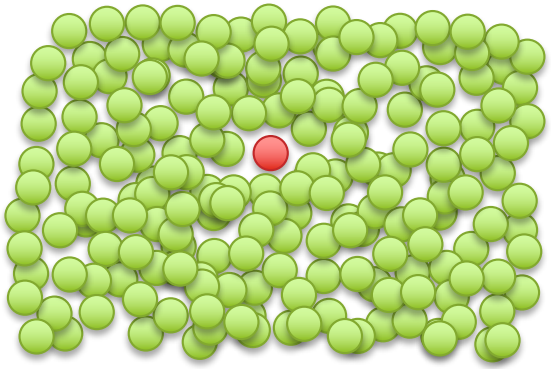
Experiment 2: 1/10



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

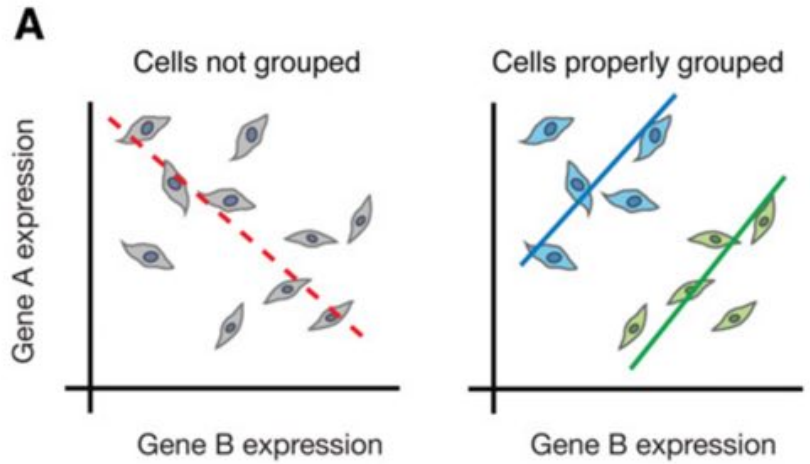
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

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



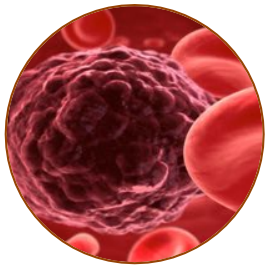
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

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)

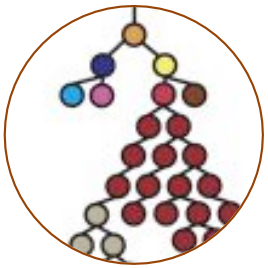
Sources of (Genomic) Heterogeneity



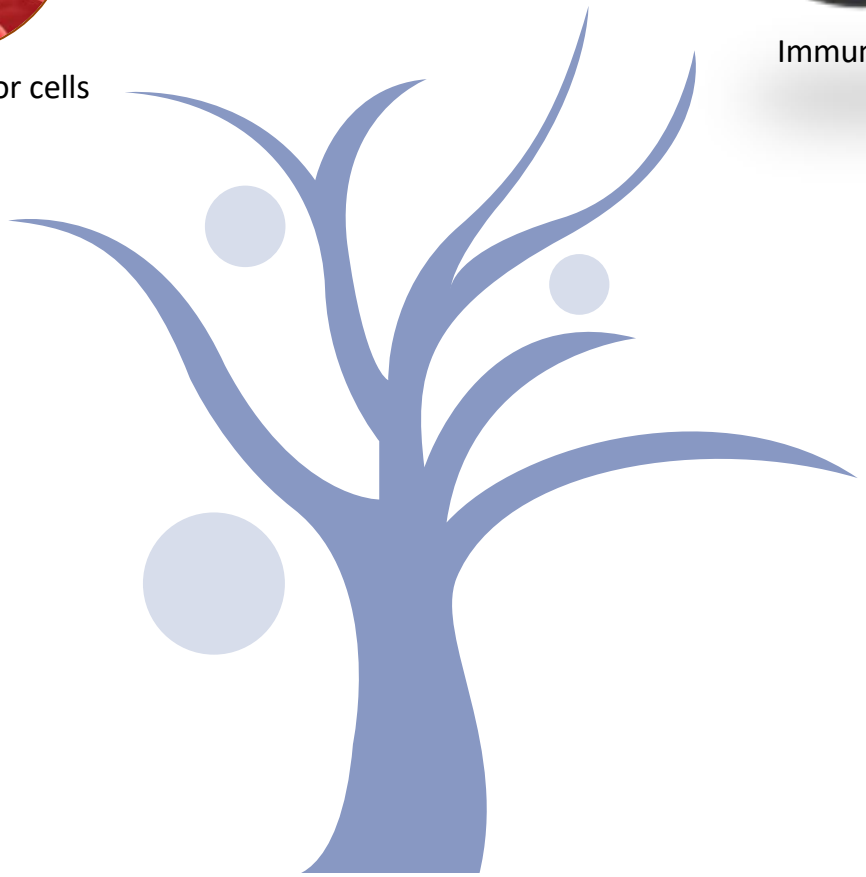
Circulating tumor cells

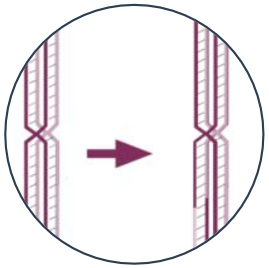
Neuronal mosaicism

Immune Cells



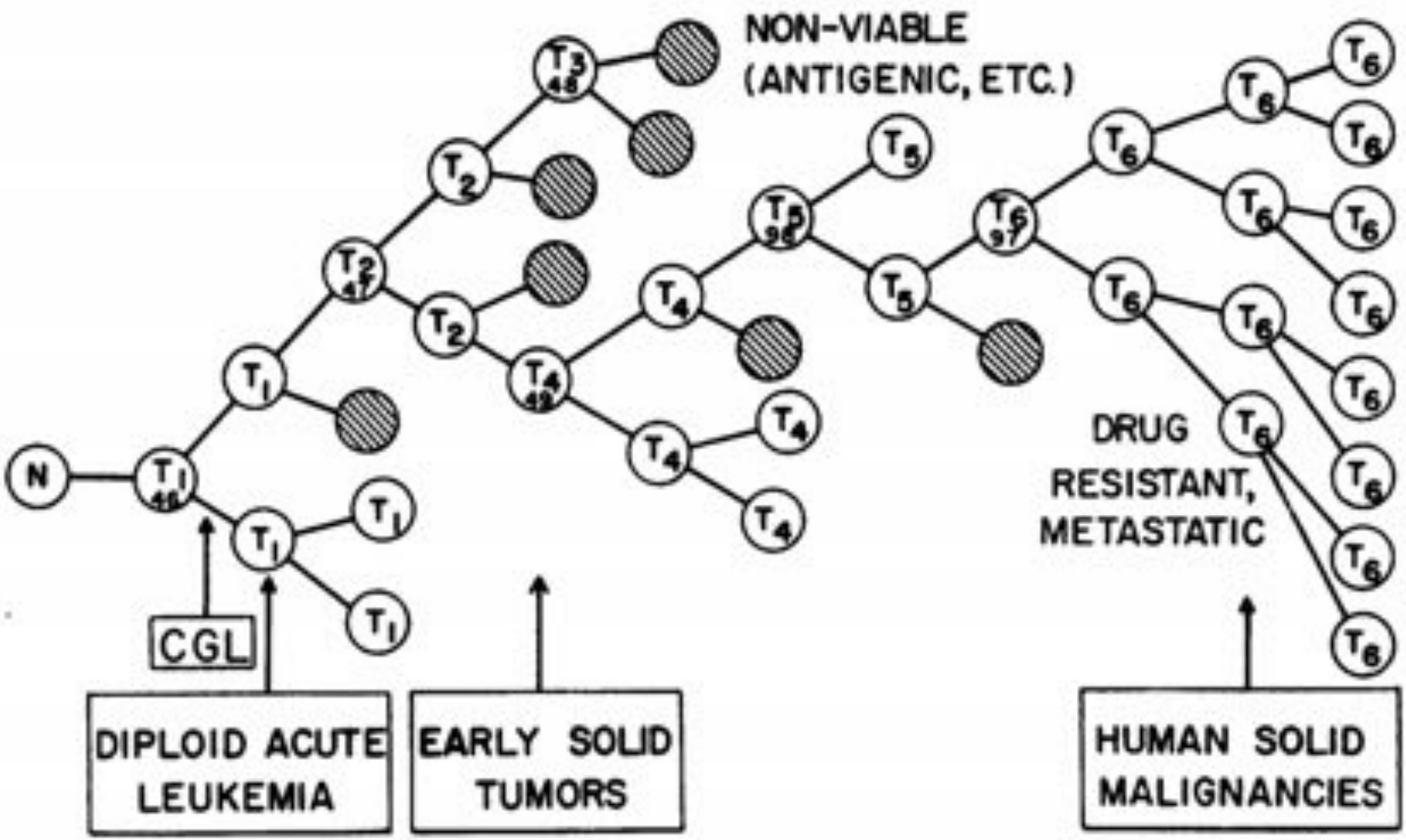
Tumor evolution





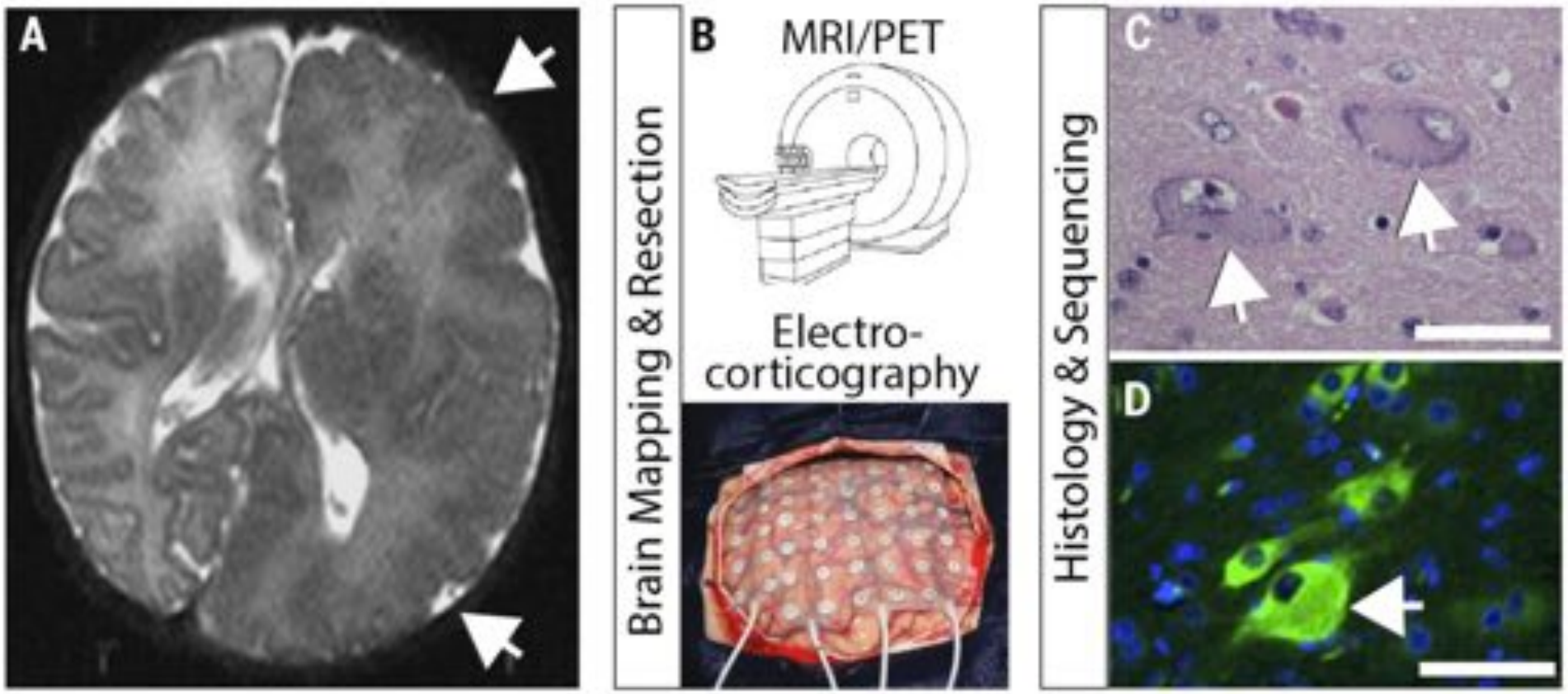
Recombination/crossover
in germ cells

Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

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

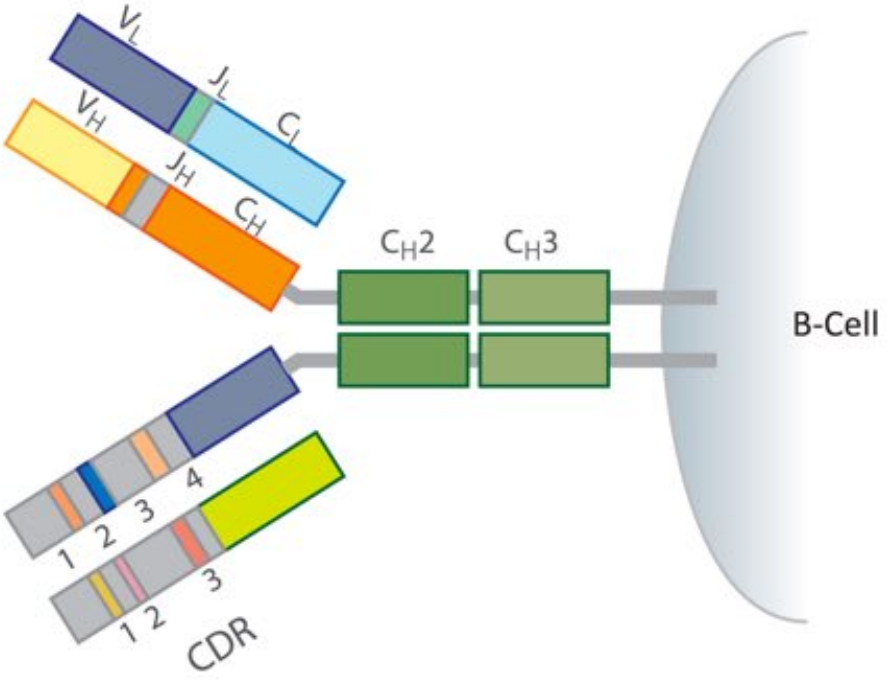


An example of brain somatic mosaicism that leads to a focal overgrowth condition.

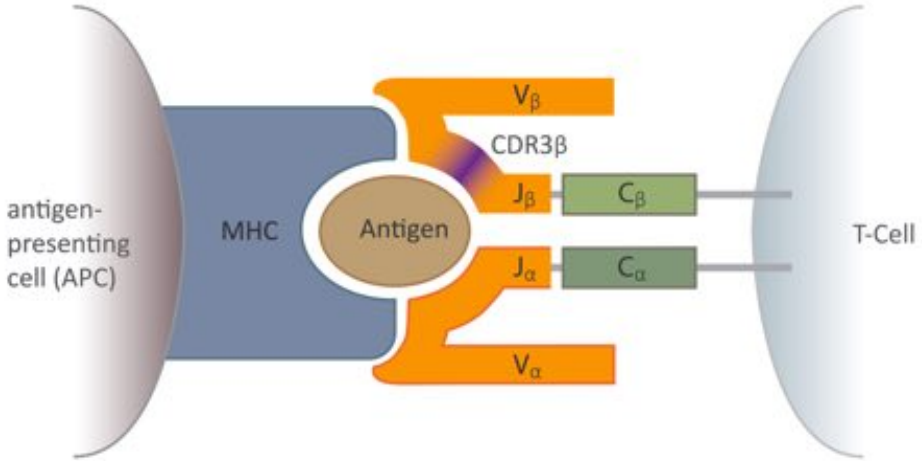
(**A**) Axial brain MRI of focal overgrowth from a 2-month-old child with intractable epilepsy and intellectual disability. (**B**) Brain mapping using high-resolution MRI is followed by surgical resection of diseased brain tissue. (**C**) Histological analysis with hematoxylin/eosin showing characteristic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



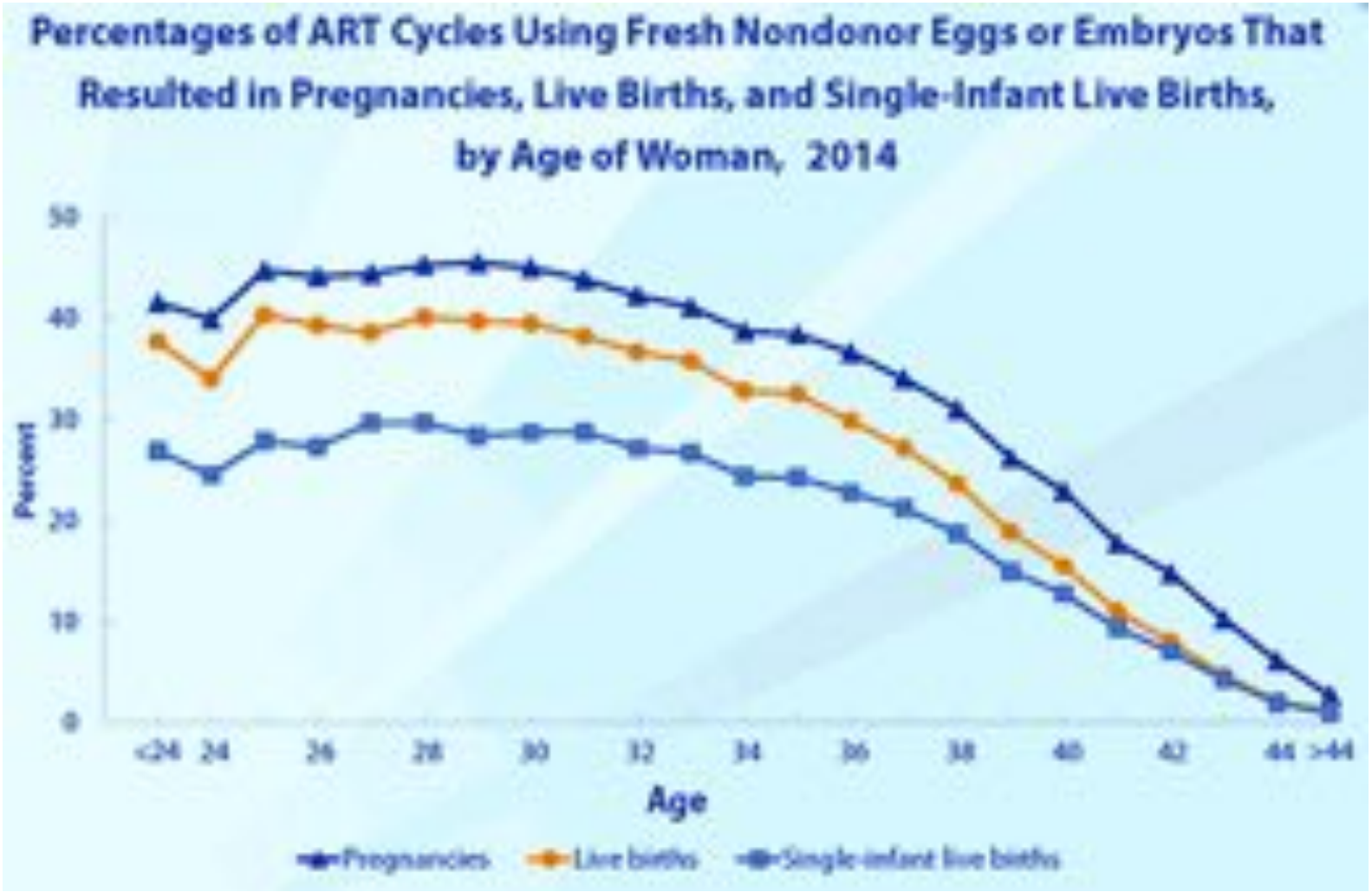
- **Massive diversity rivaled only by germ cells**
- **Somatic recombination**

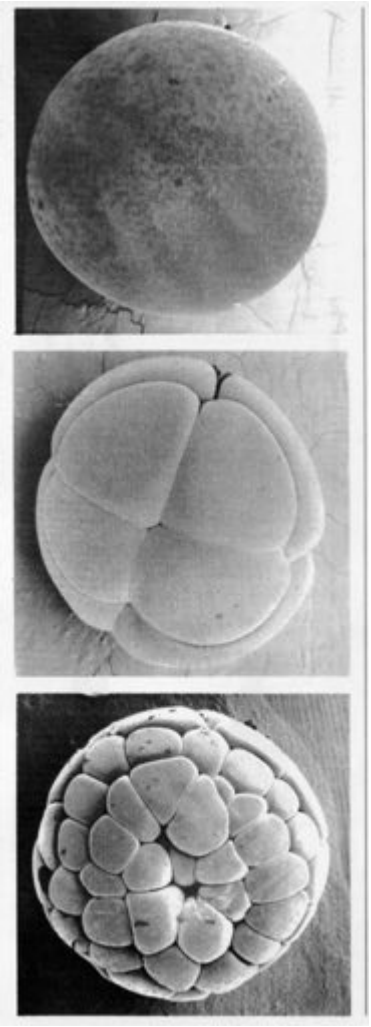


- **B cells – antibody generation**
- **T cells – antigen response**

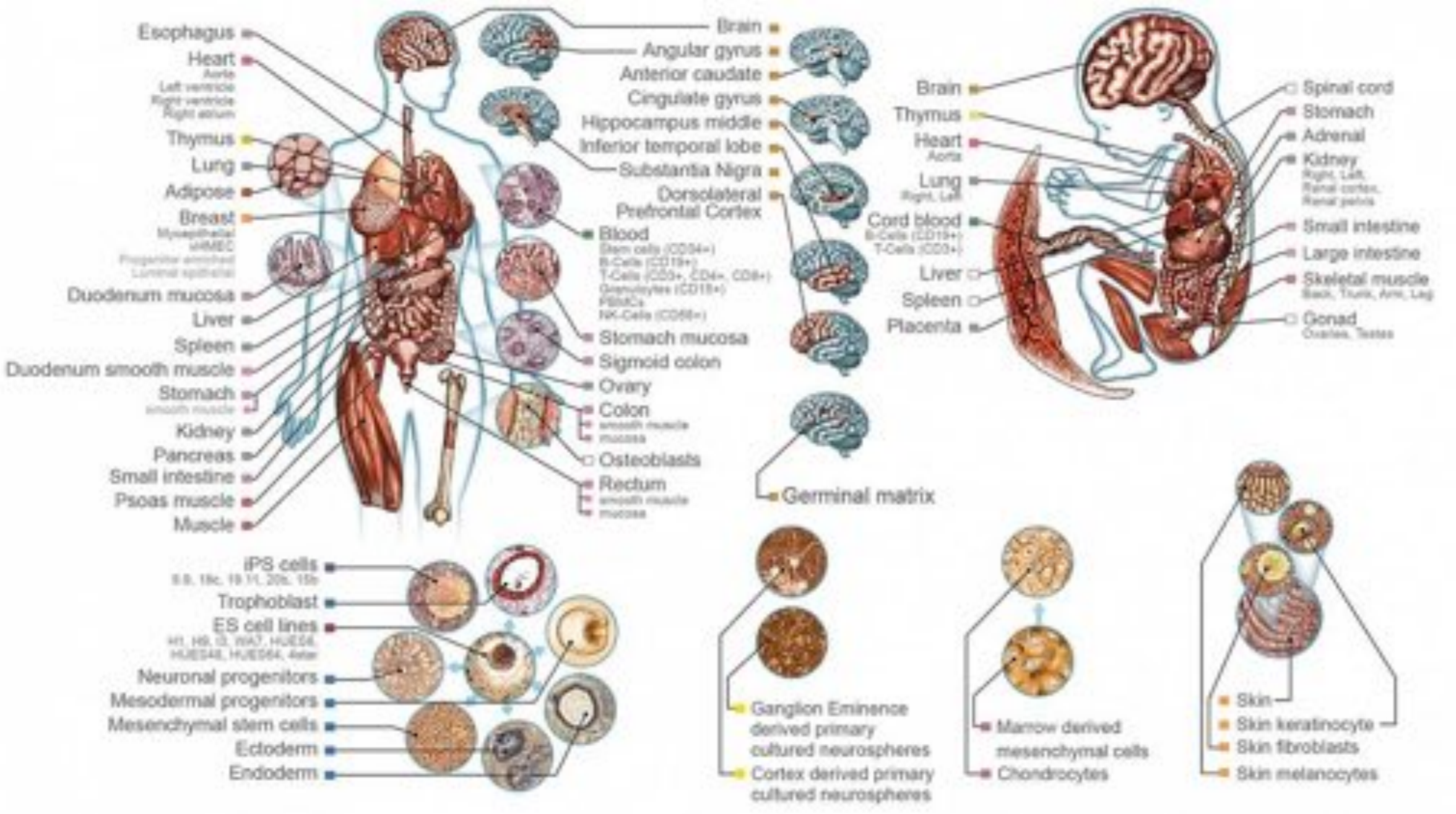
Single cell research. Illumina.

In-vitro Fertilization

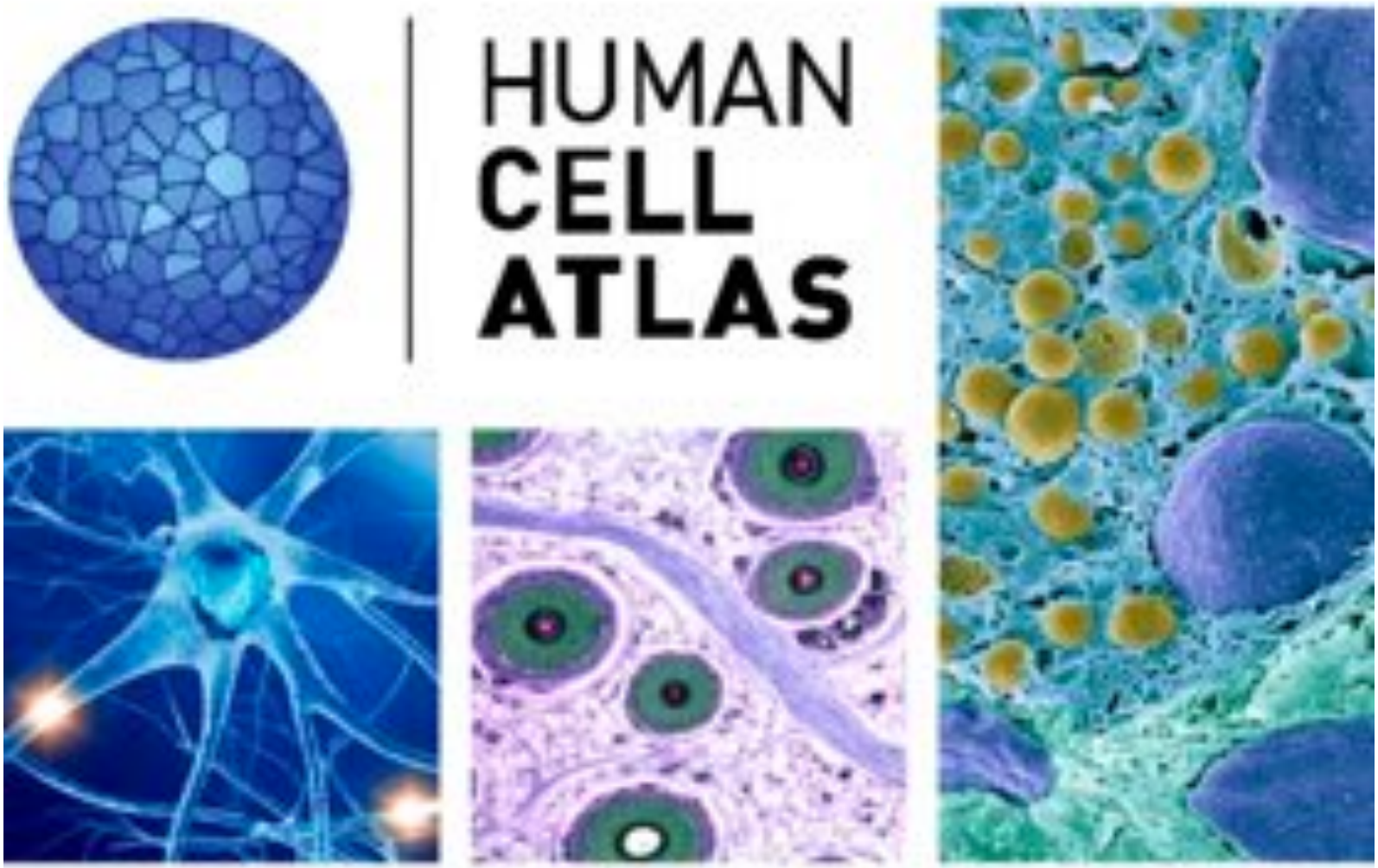




Sources of (Cellular) Heterogeneity



Roadmap Epigenomics Consortium



https://www.humancellatlas.org/

Single Cell Analysis

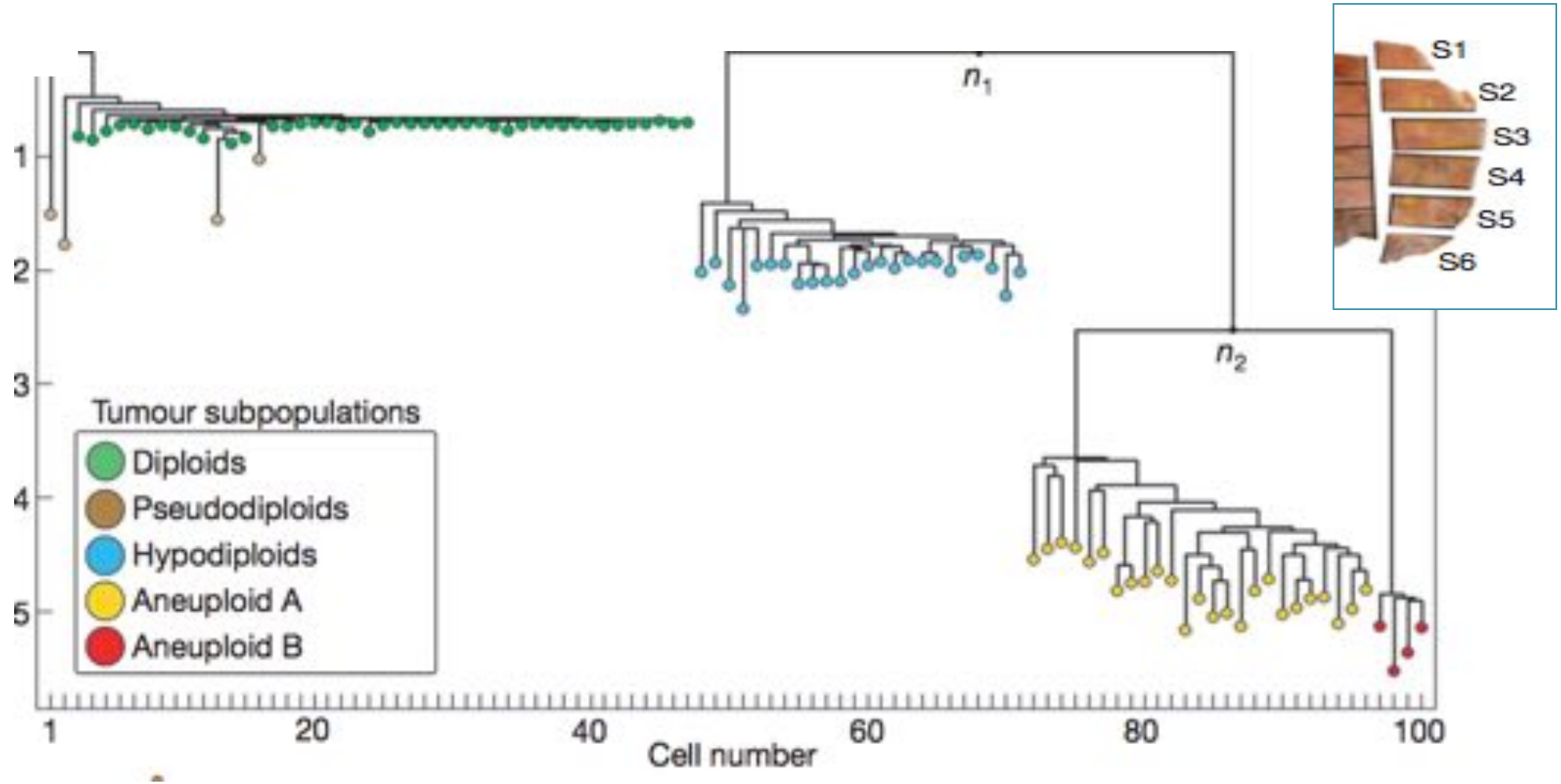
1. Why single cells?
2. scDNA
3. scRNA and other assays

LETTER

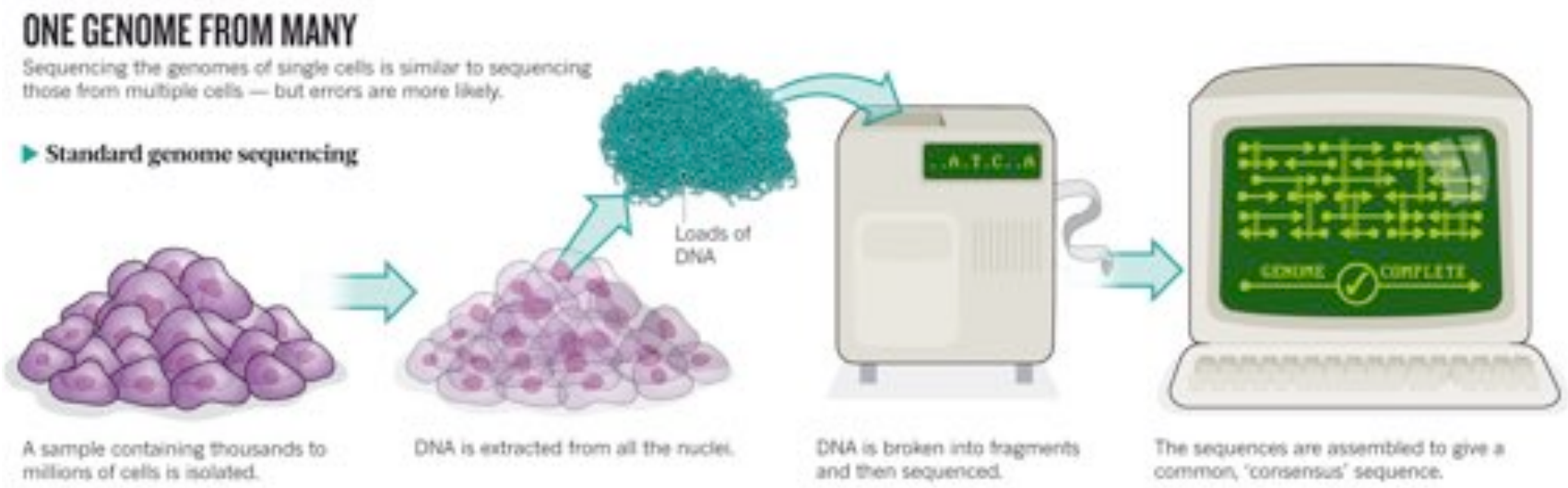
doi:10.1038/nature09807

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]

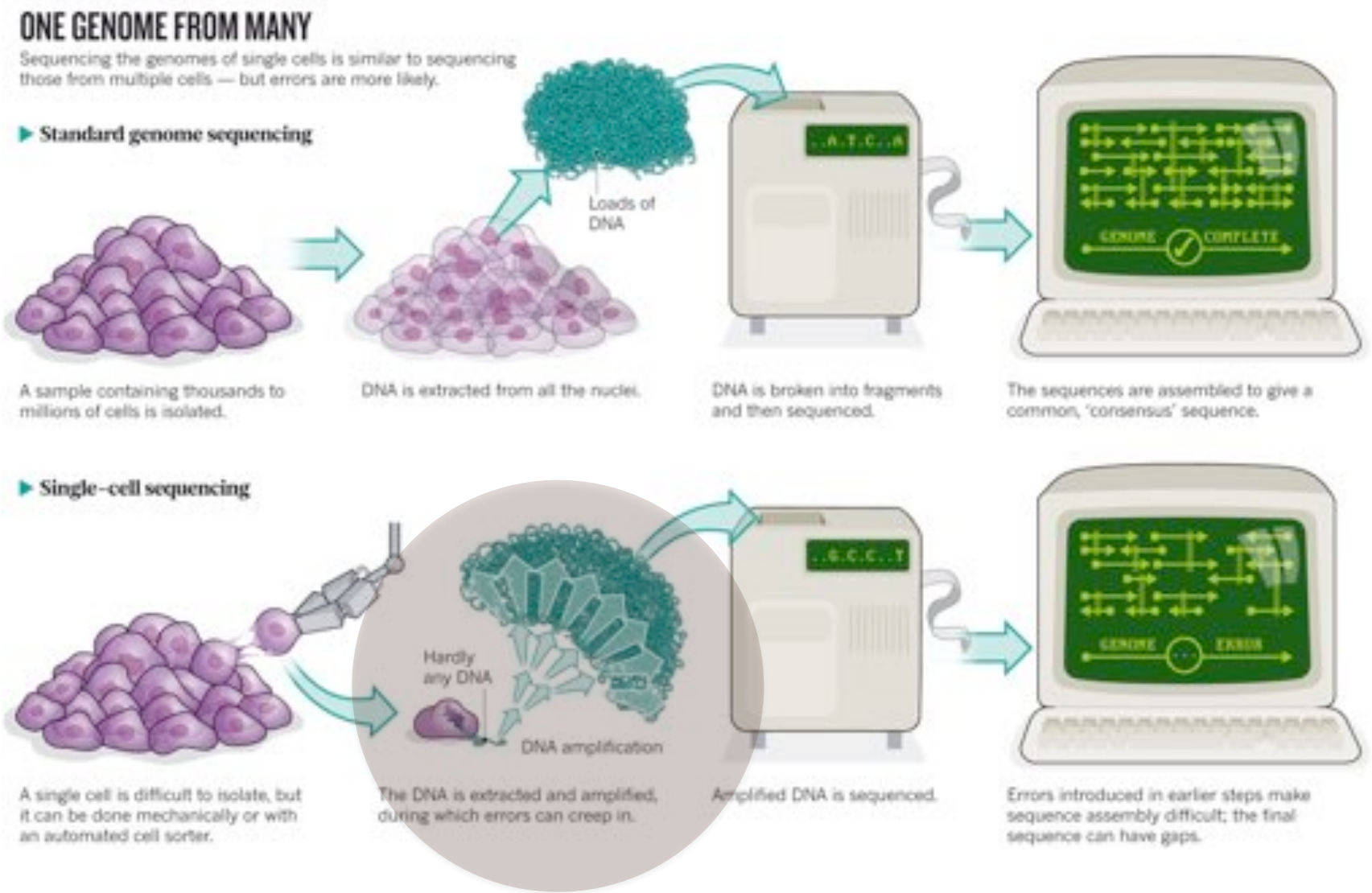


Single-cell vs. bulk sequencing



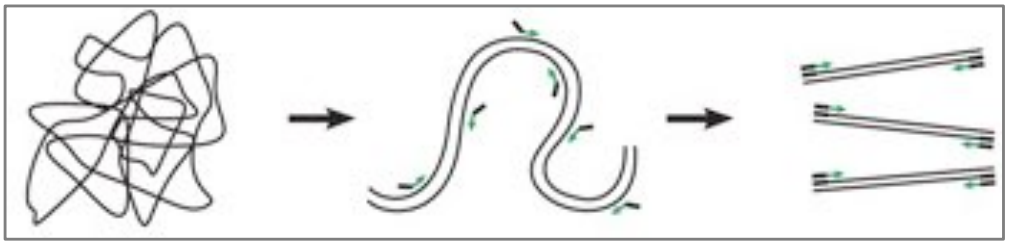
Brian Owens, Nature News 2012

Single-cell vs. bulk sequencing

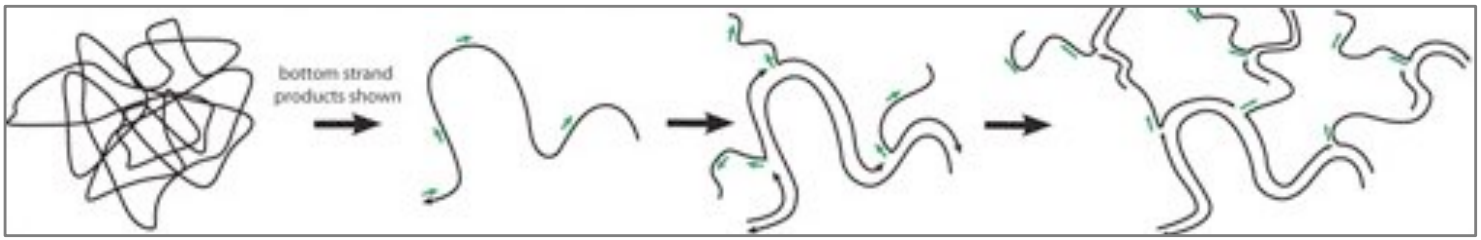


Brian Owens, Nature News 2012

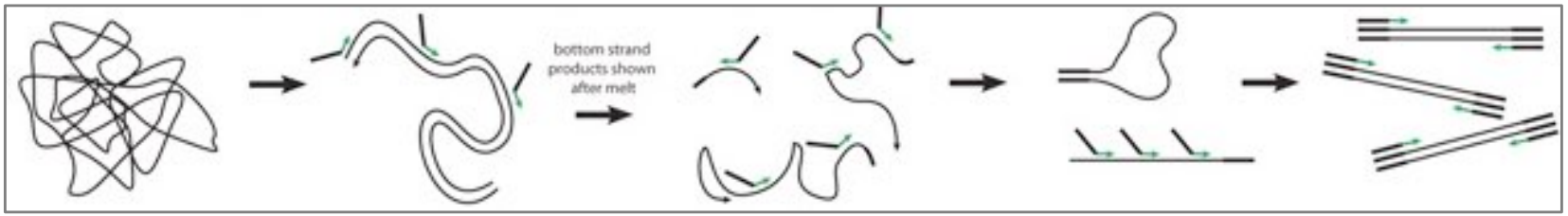
Whole Genome Amplification Techniques



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



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

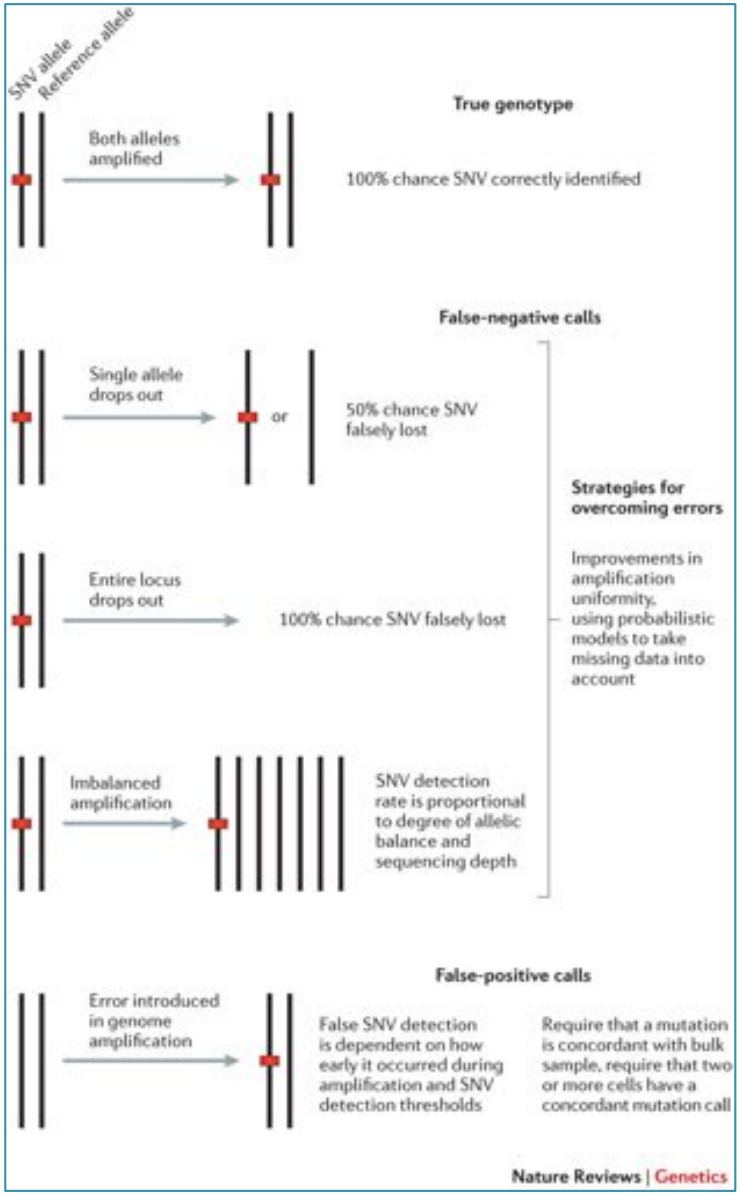


MALBAC: Multiple Annealing and Looping Based Amplification Cycles
Zong et al. (2012) Science

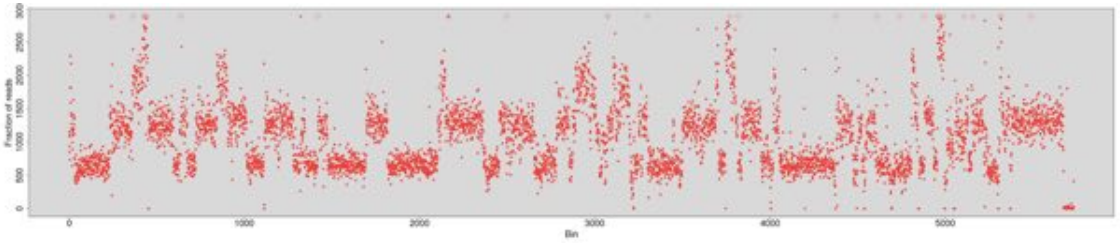
Paul Blainey, FEMS Microbiol Rev. 2013

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

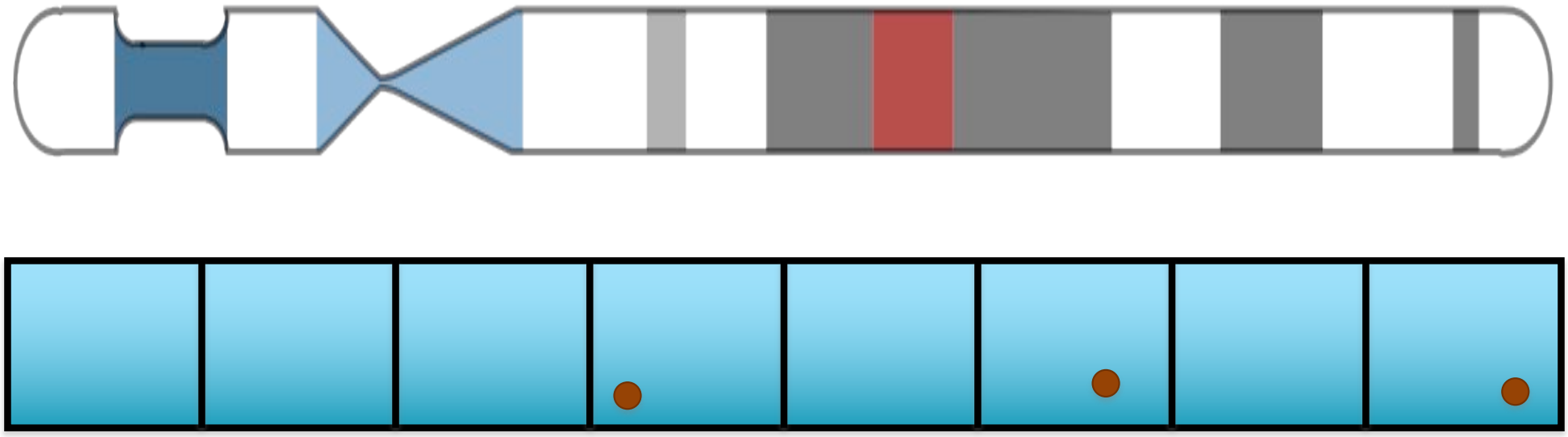
Coverage is very sparse and noisy

-> 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

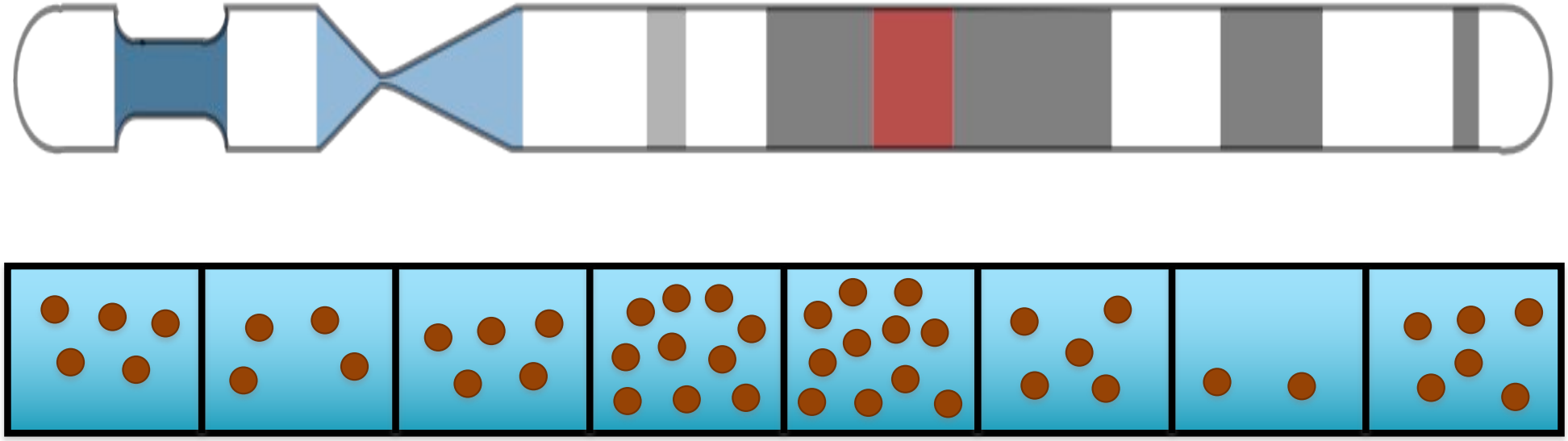


Single Cell CNV analysis

- 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

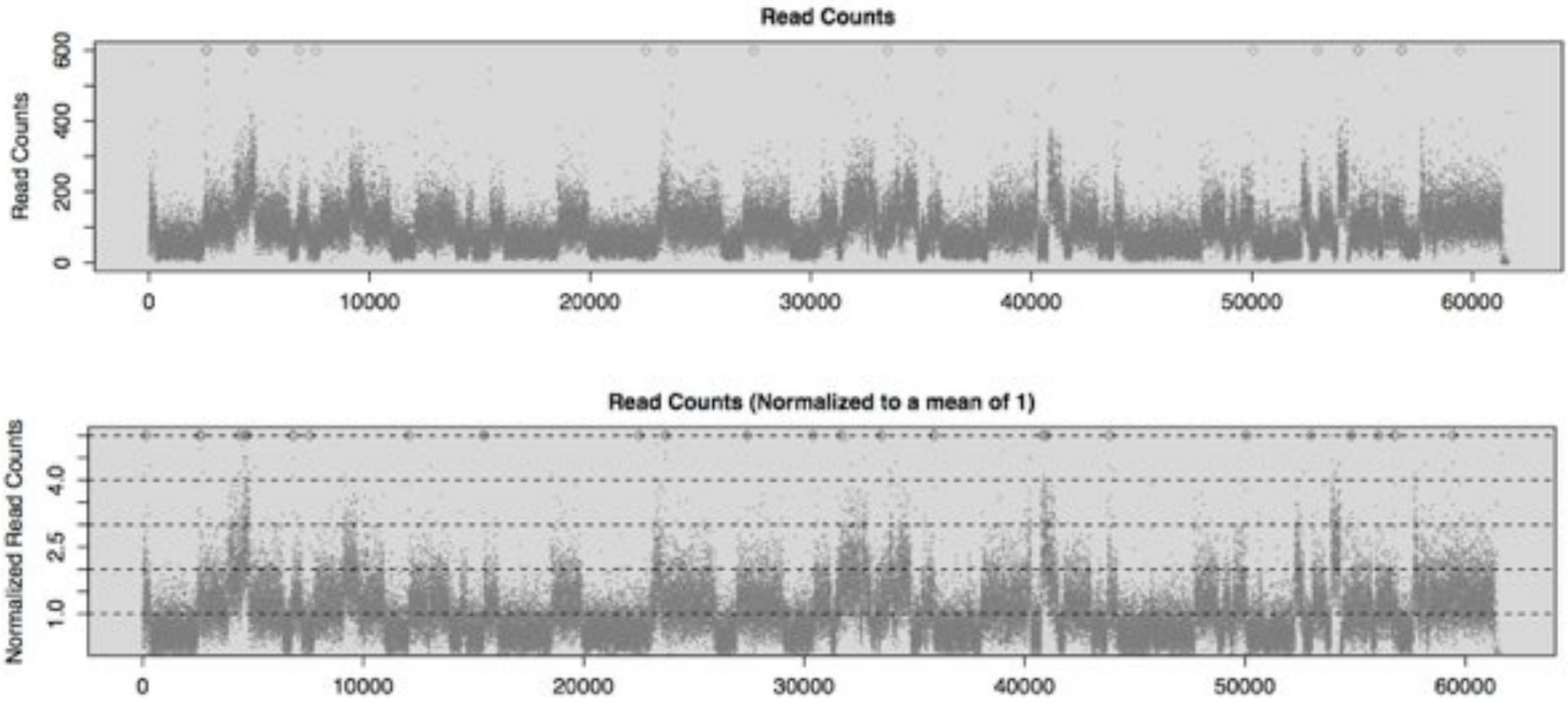
5	4	5	10	11	5	2	5

Single Cell CNV analysis

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

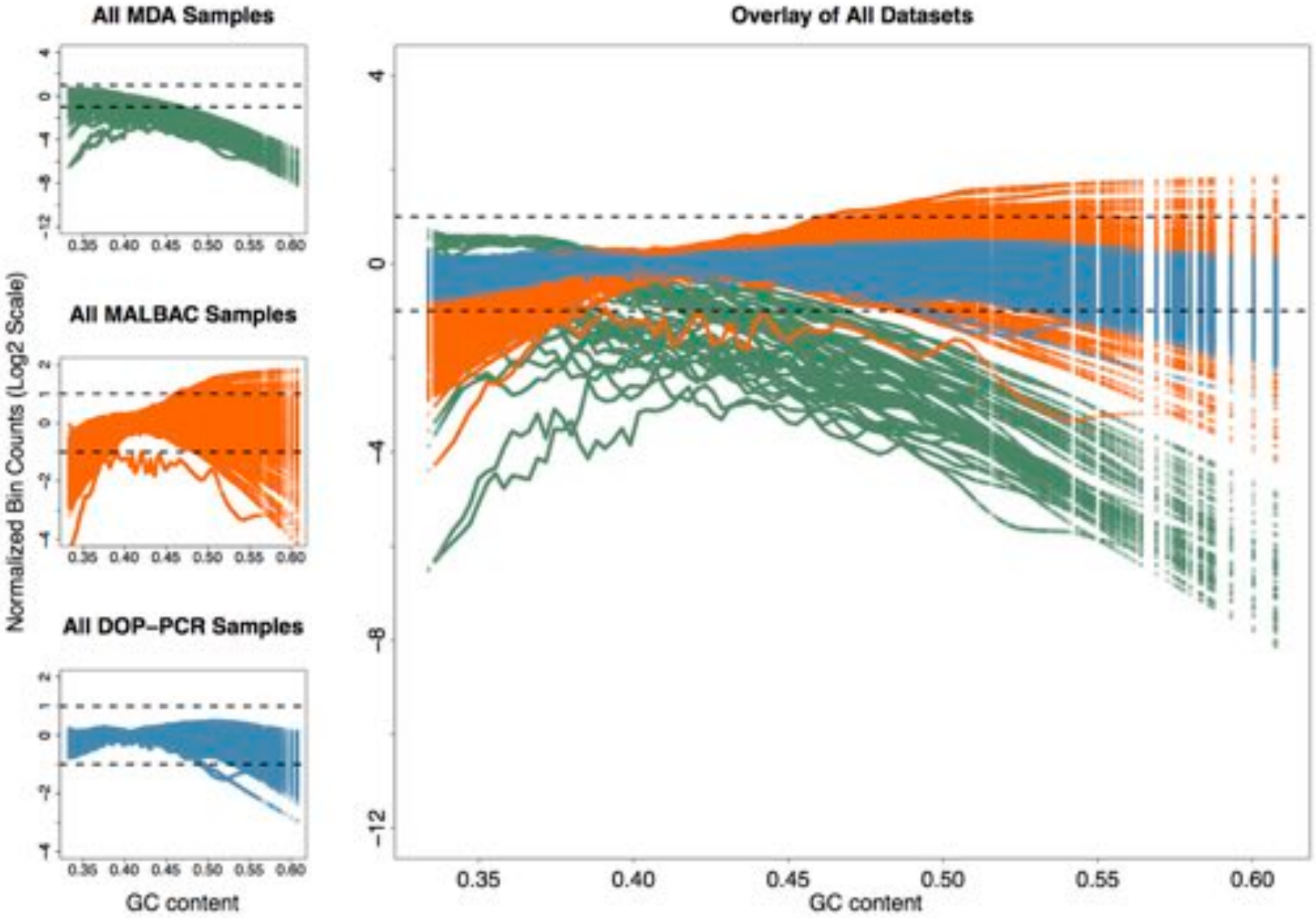
Use uniquely mappable bases to establish bins

2) Normalization



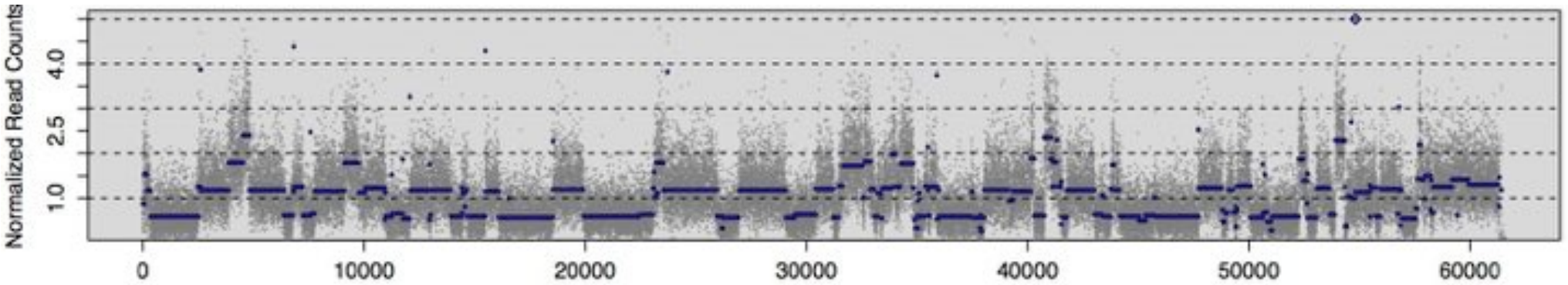
Also correct for mappability, GC content, amplification biases

GC Bias

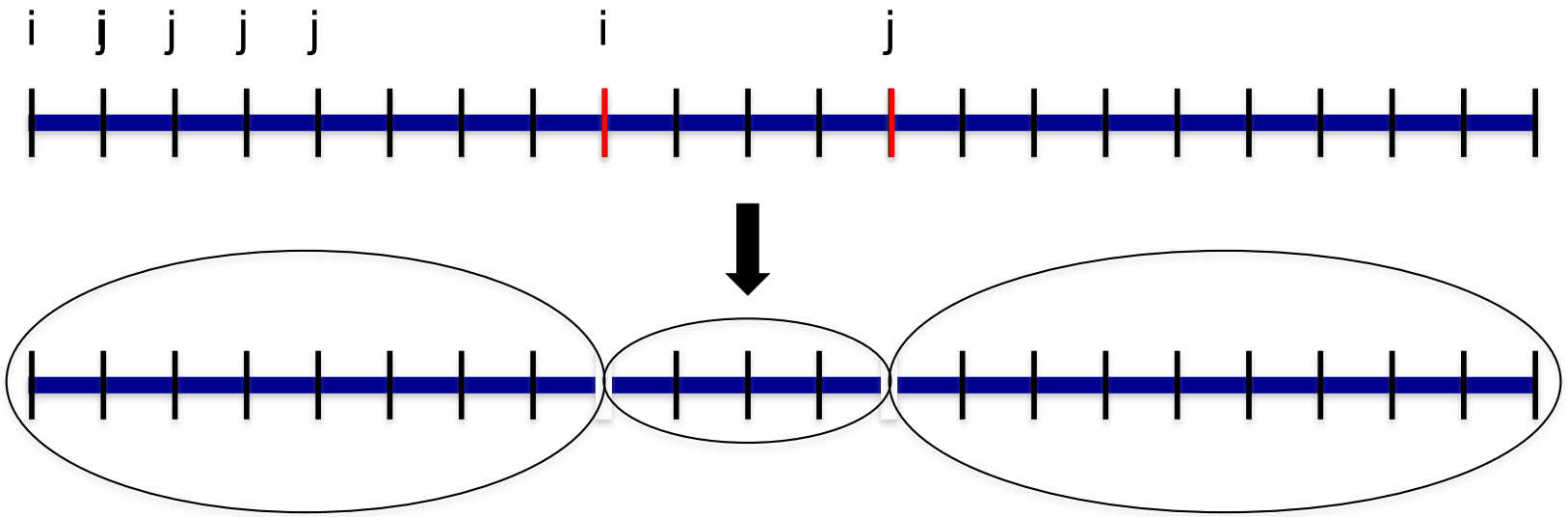


Garvin and Aboukhalil et al., Nature Methods, 2015

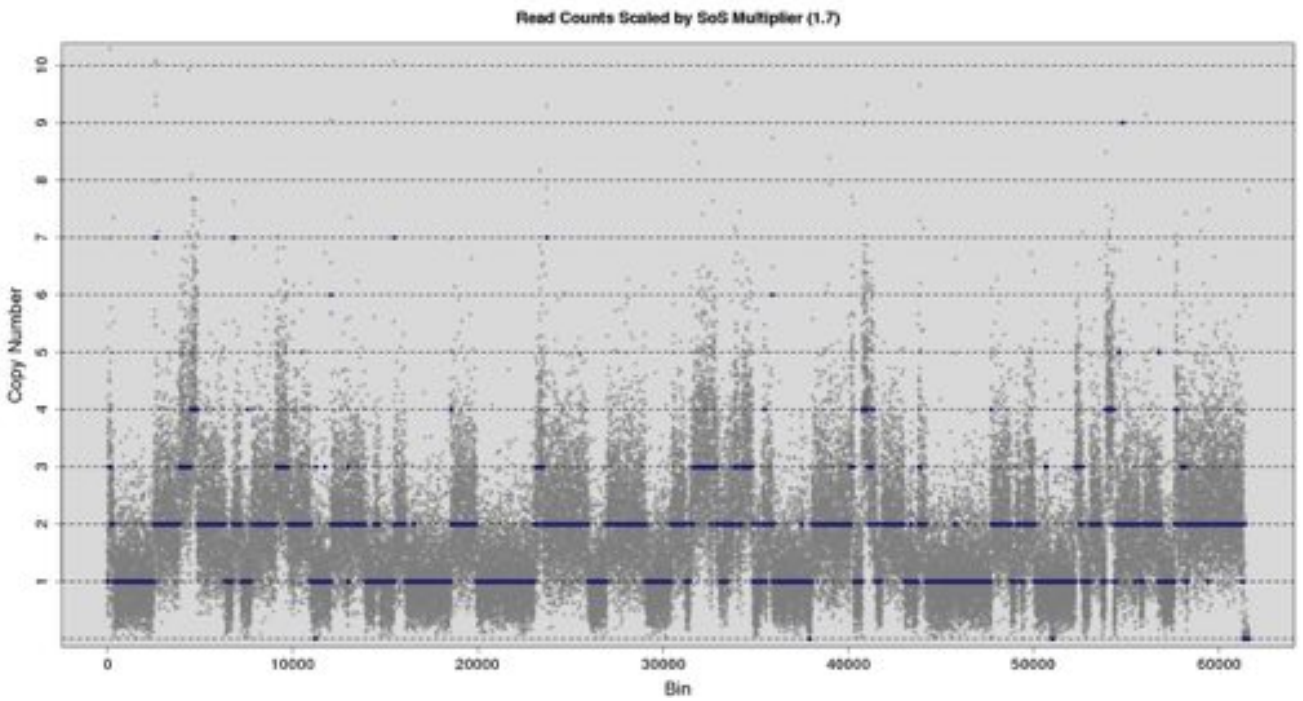
3) Segmentation



Circular Binary Segmentation (CBS)

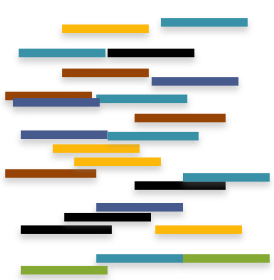


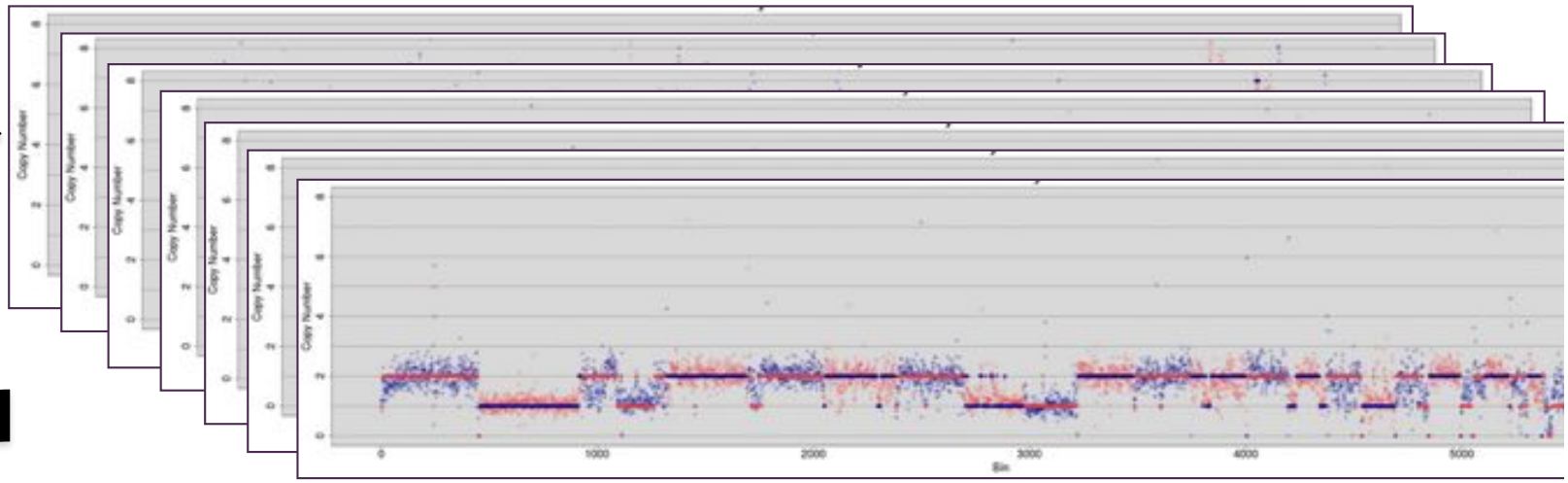
4) Estimating Copy Number

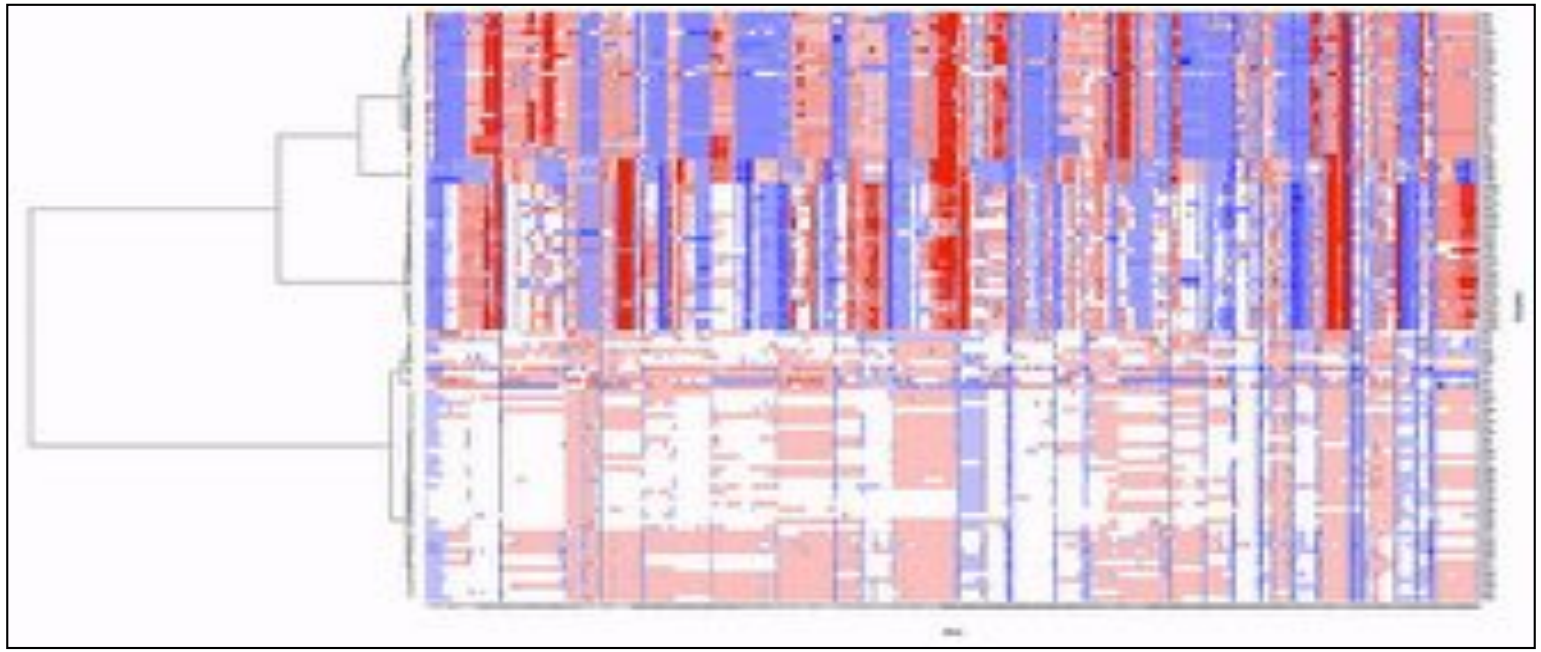


$$CN = argmin\left\{\sum_{i,j}(\hat{Y}_{i,j} - Y_{i,j})^2\right\}$$

5) Cells to Populations

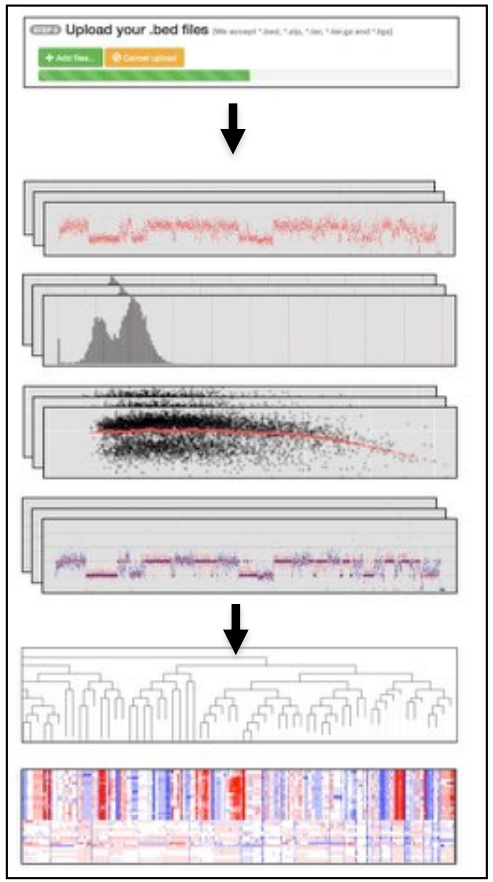






Gingko

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

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

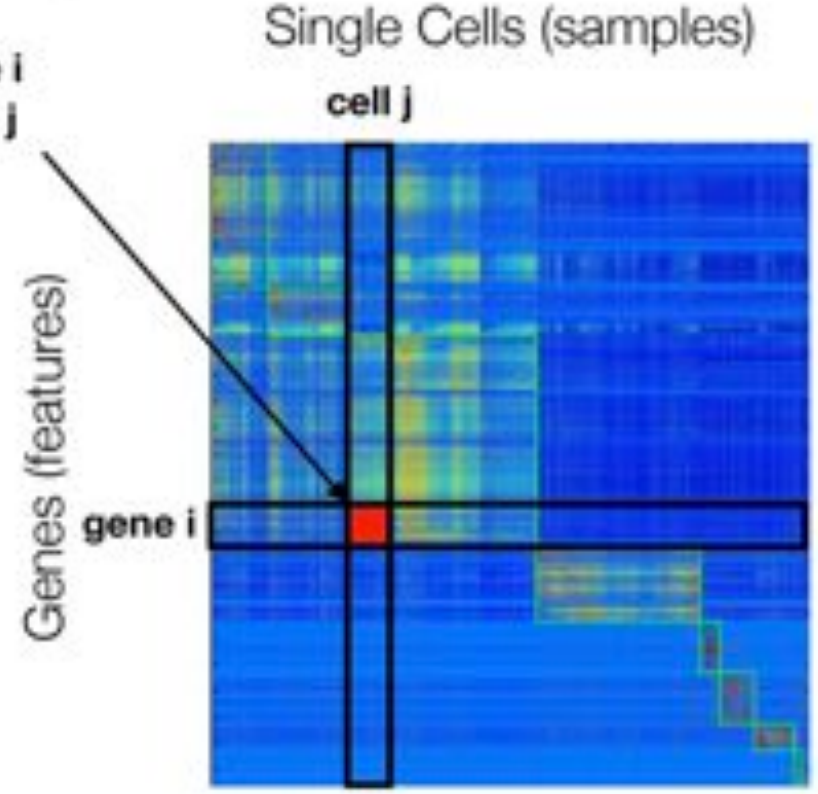
— a snapshot of the underlying biology in a data matrix.

Brain cells

Biological sample

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

sequencing

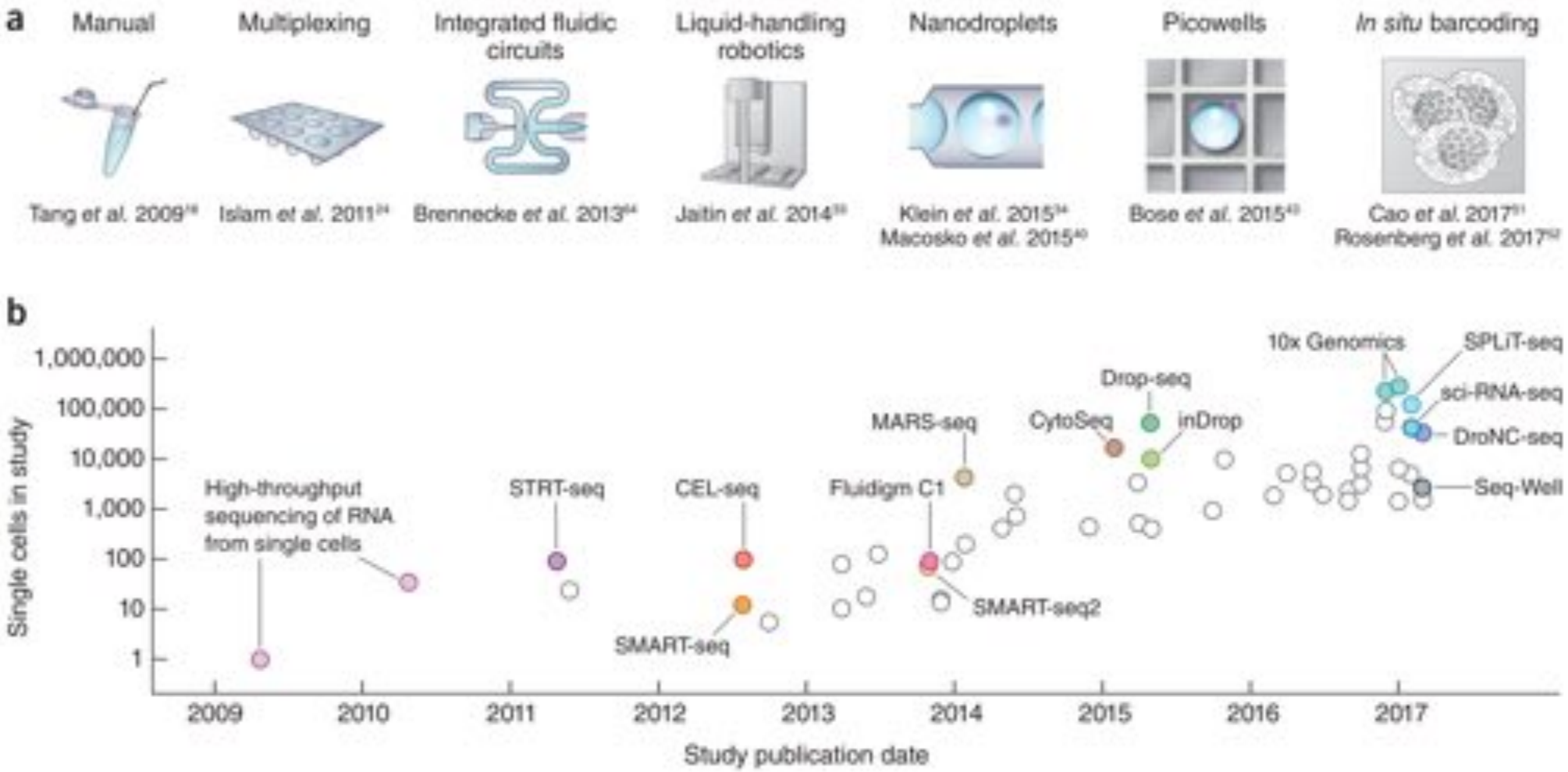


Gene expression matrix

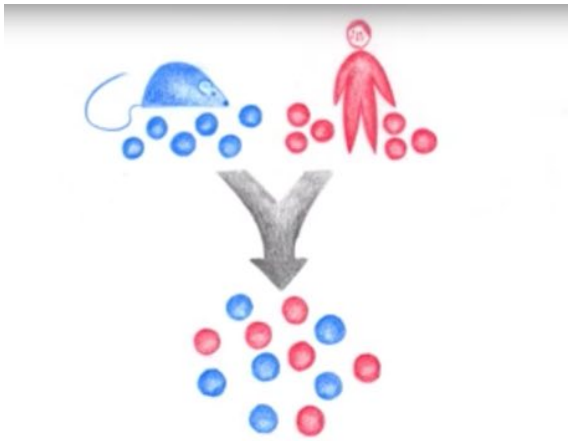
computationally explore complex biological systems

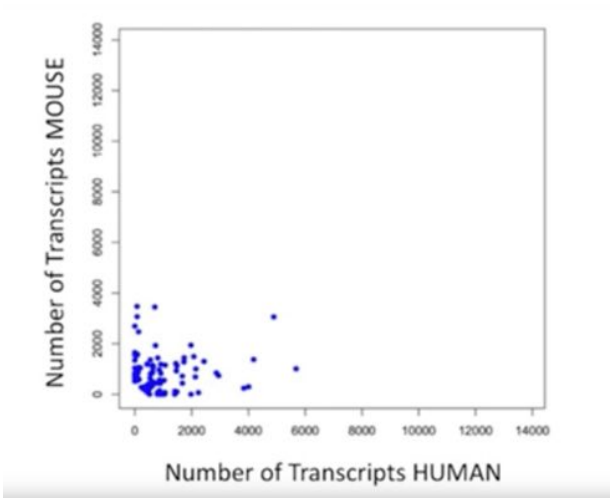
Martin Zhang

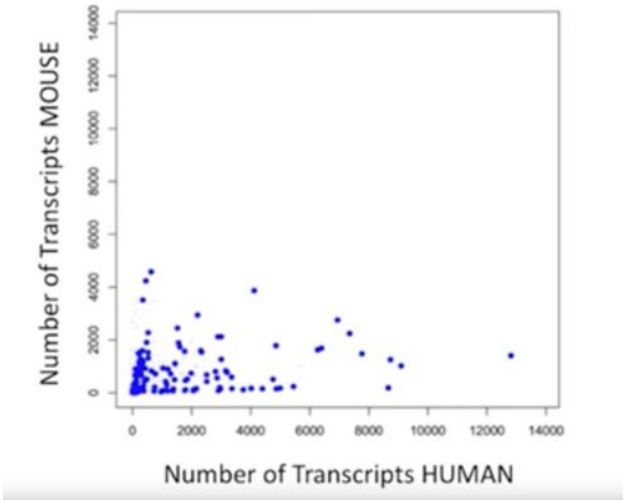
A decade of single-cell RNA-seq

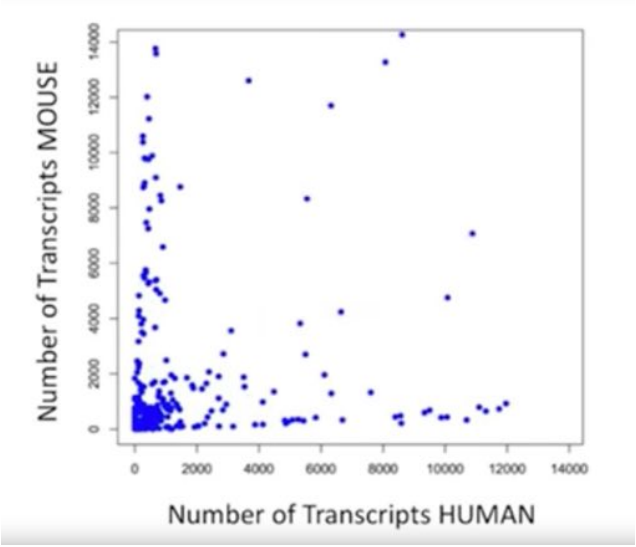


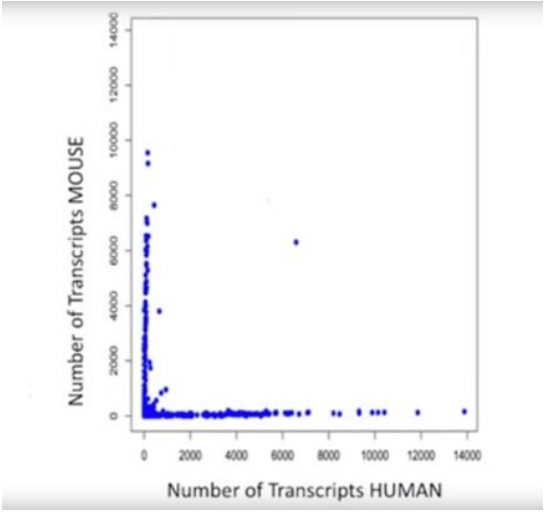
Svensson et al. 2018. DOI:10.1038/nprot.2017.149

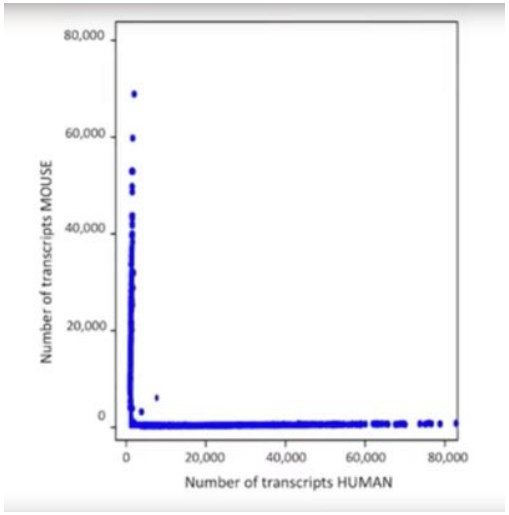






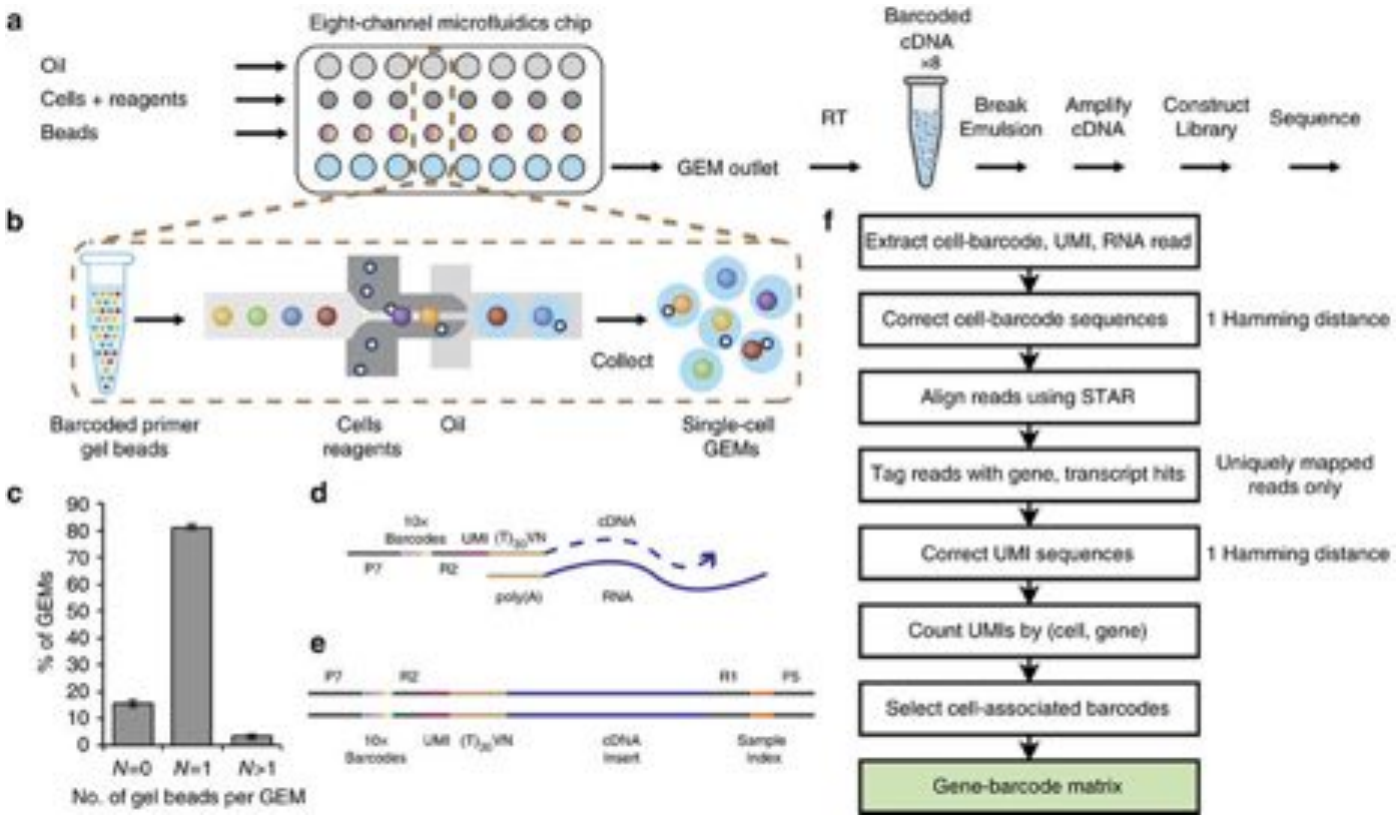






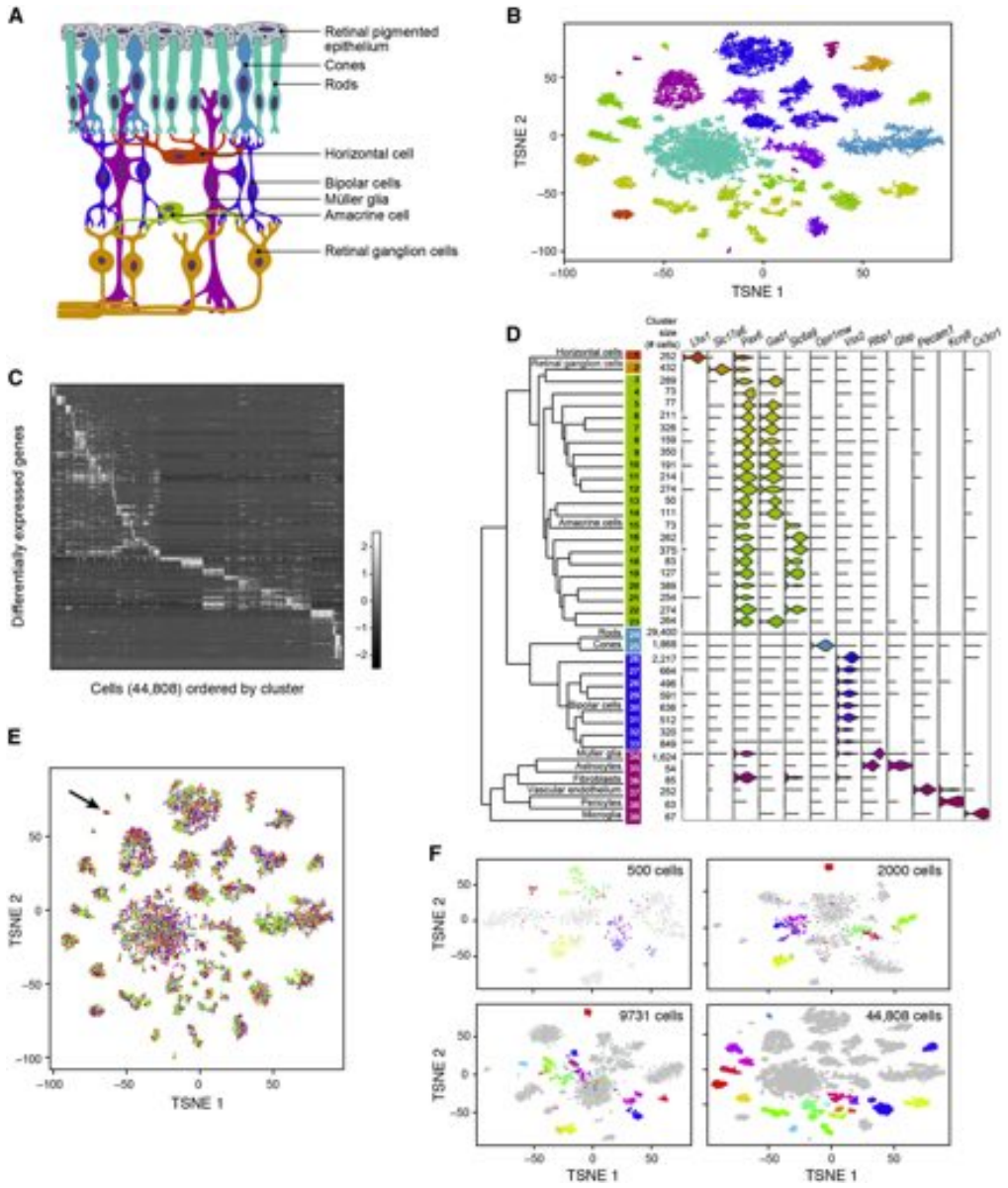
Drop-seq: Droplet barcoding of single cells
https://www.youtube.com/watch?v=vL7ptq2Dcf0

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

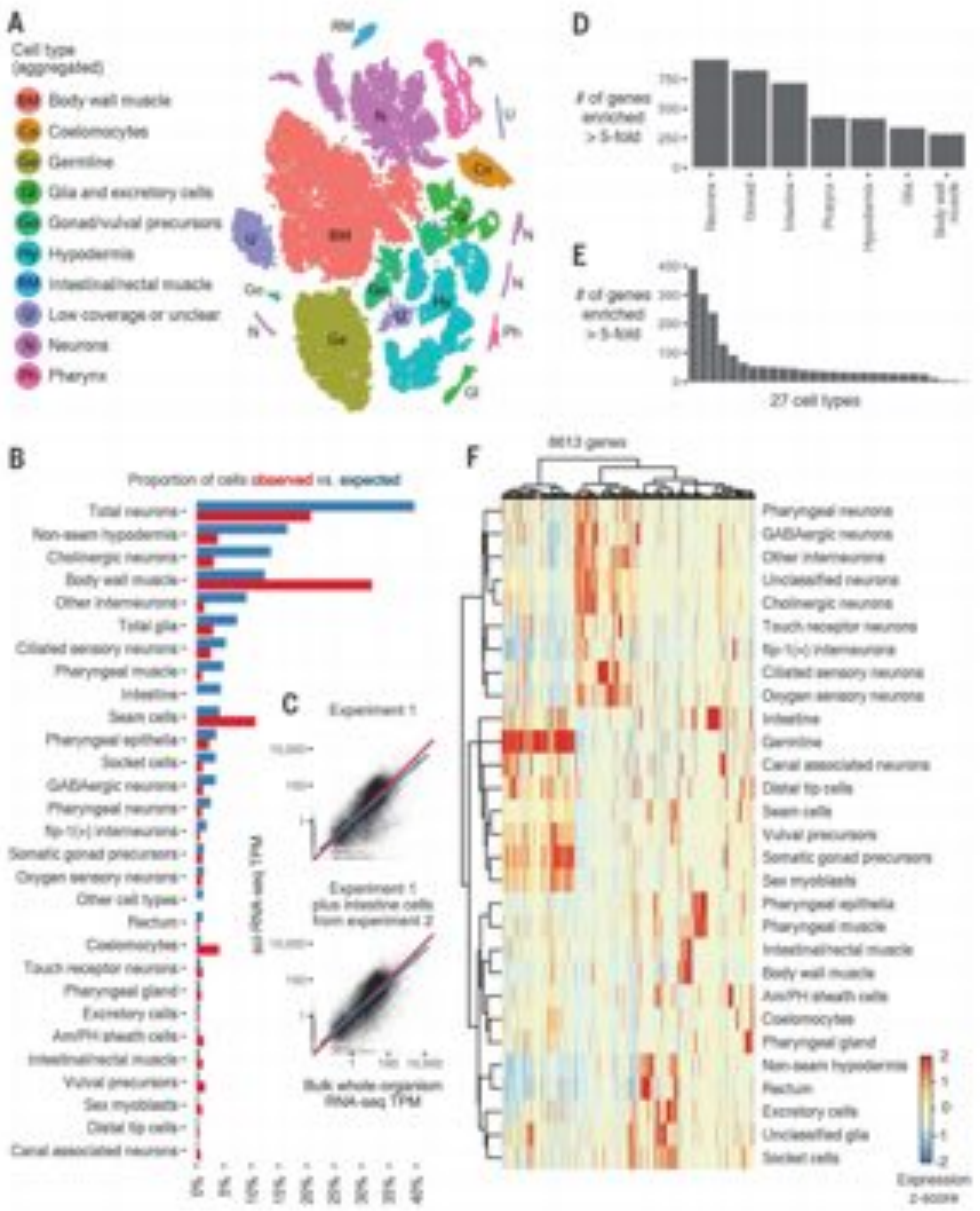
(c) Differentially expressed genes in each cluster

(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

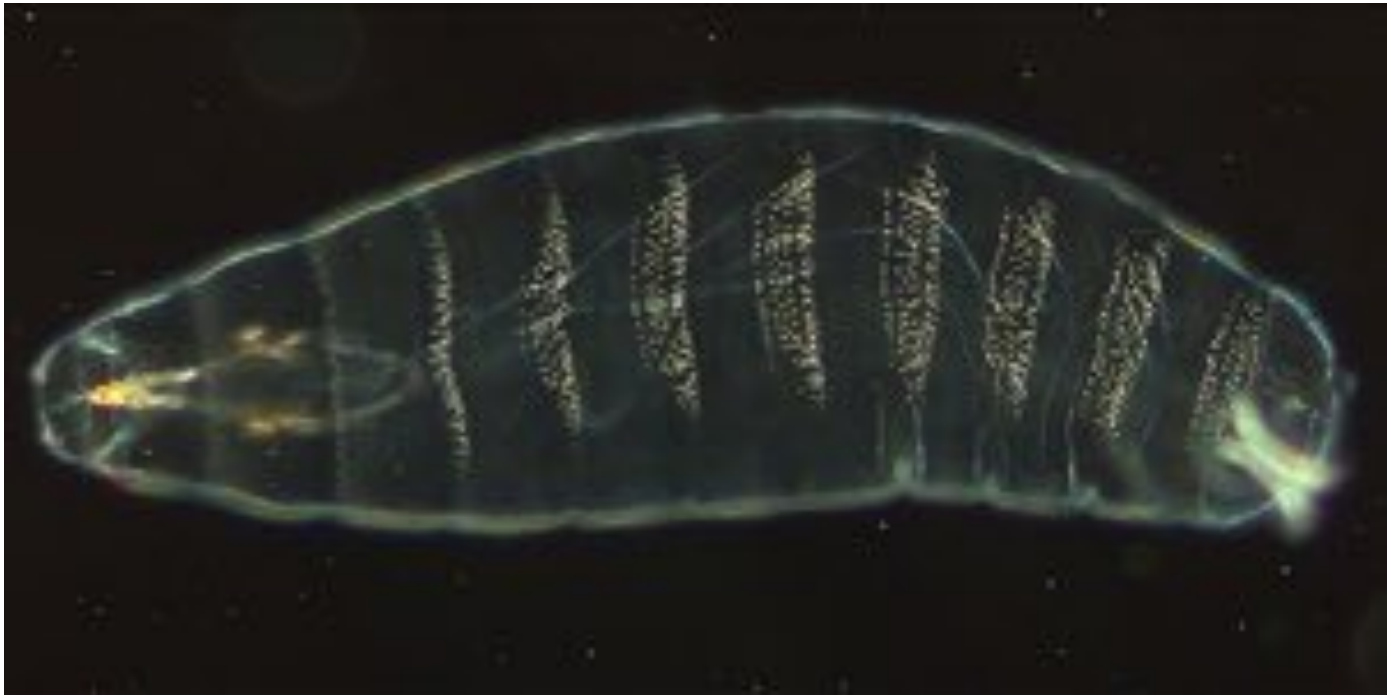
(a) t-SNE visualization of clusters

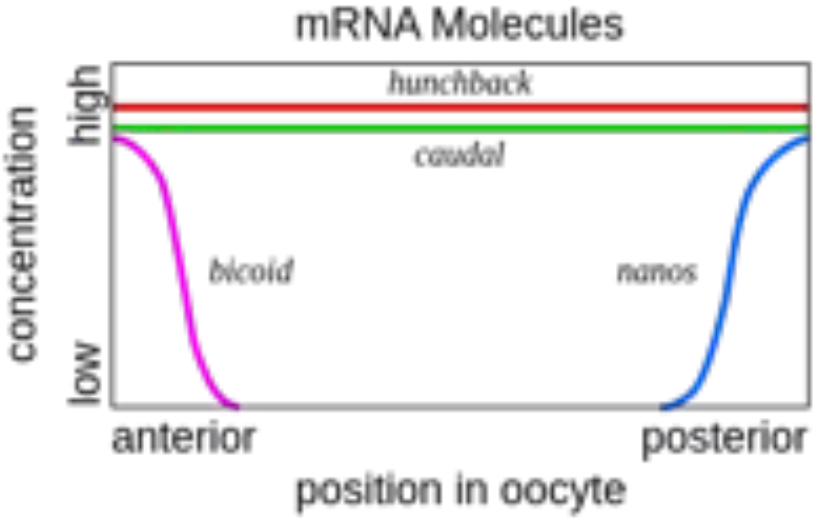
(b) Proportion of cells observed vs expected match well (including cells that only occur once or twice in the animal)

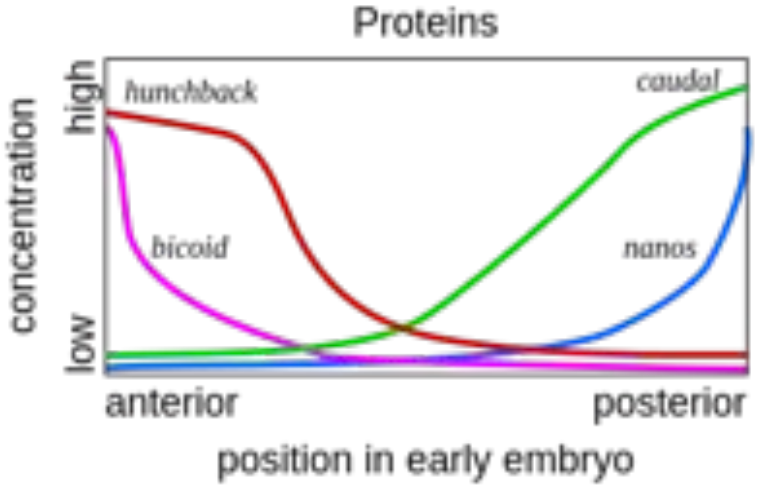
(c) Good correlation between single cell and bulk analysis of selected cell types

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

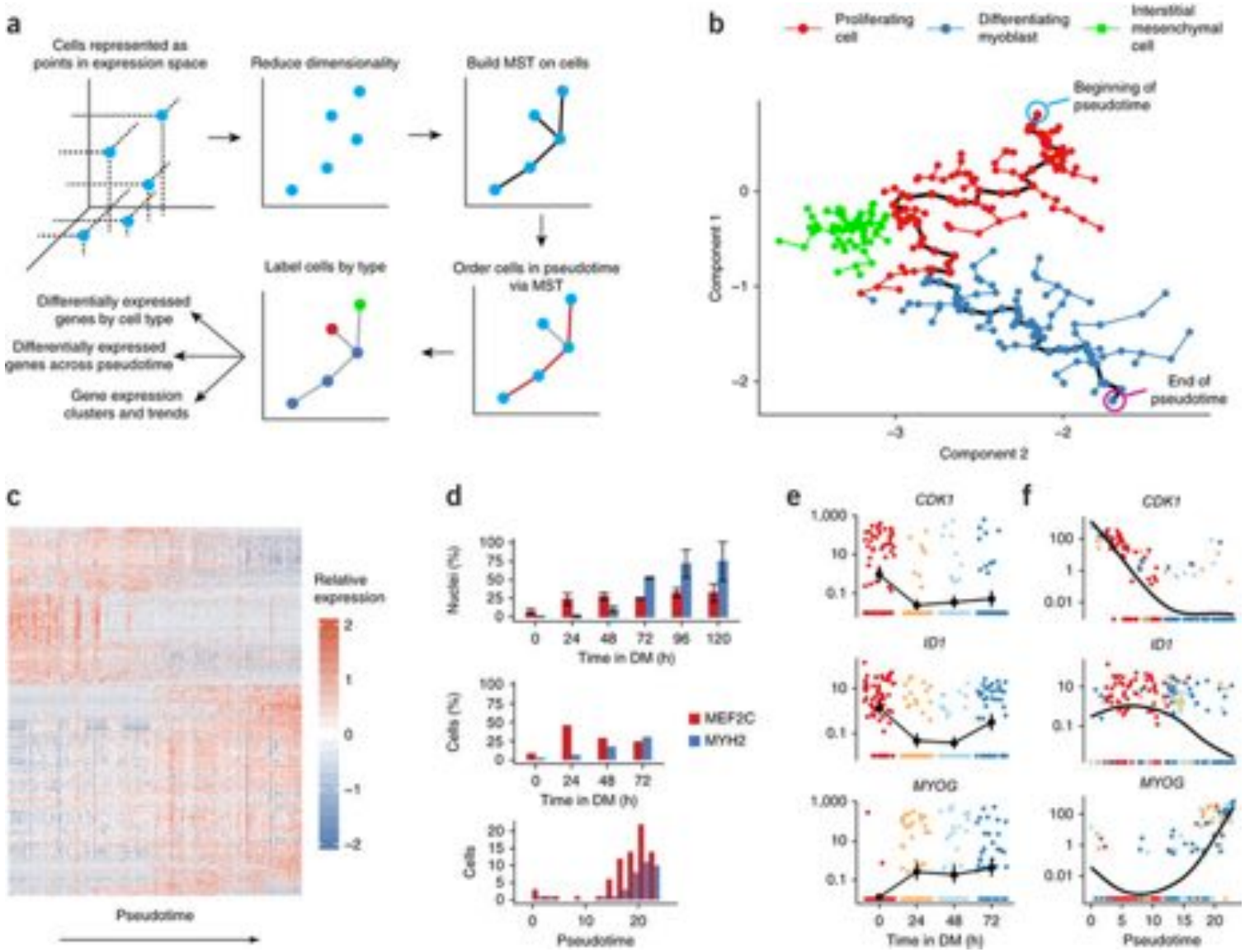
Comprehensive single-cell transcriptional profiling of a multicellular organism
Cao et al (2017) Science. 357:661-557





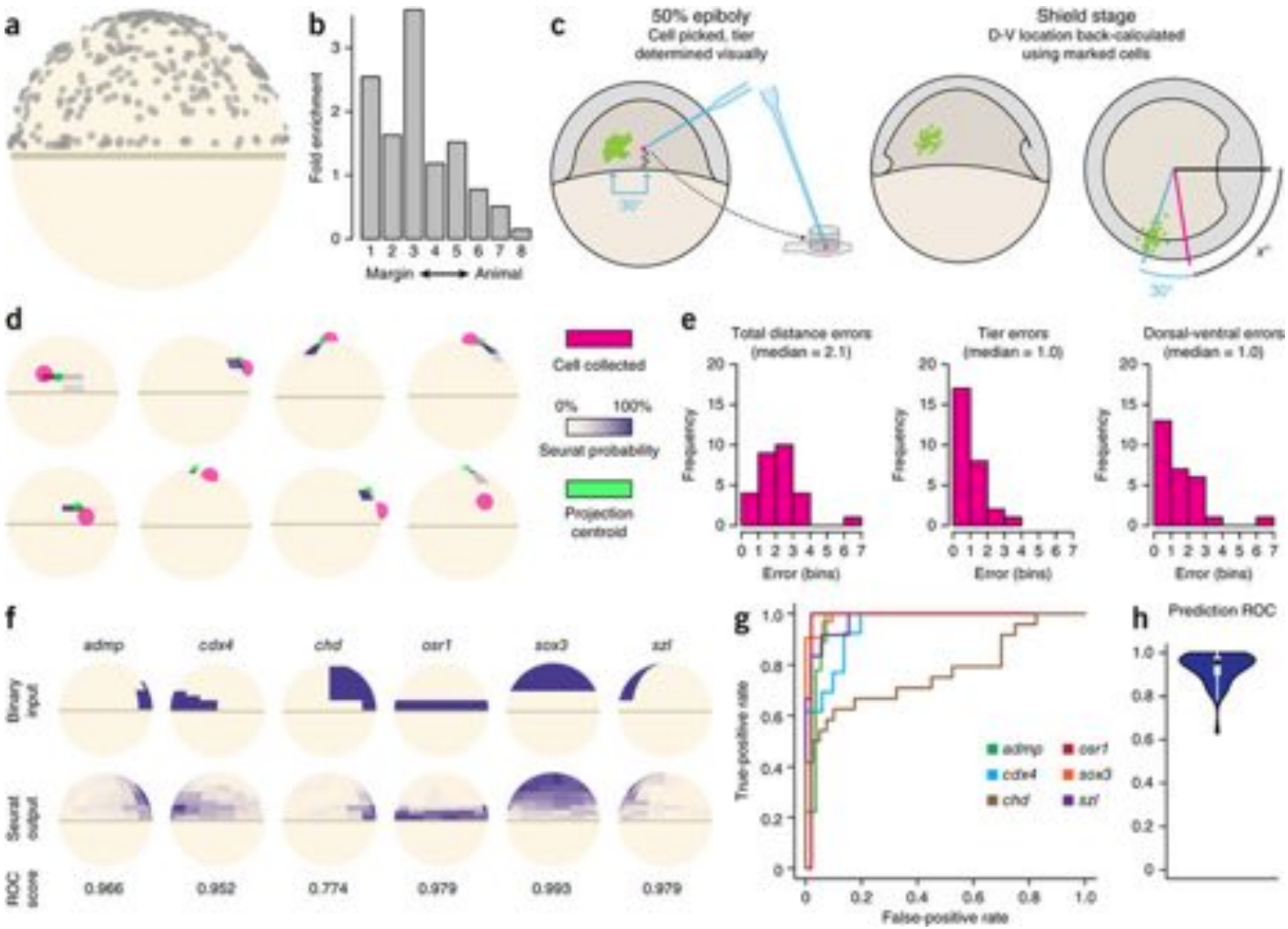


https://en.wikipedia.org/wiki/Drosophila_embryogenesis



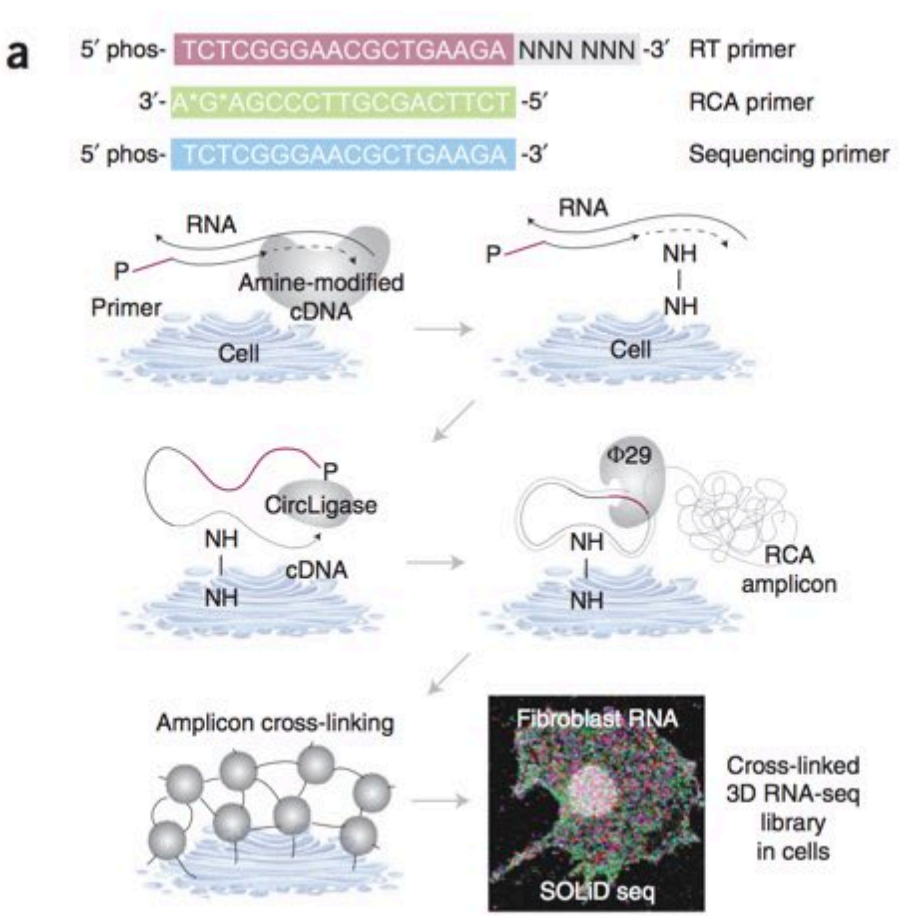
The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells ("Monocle")

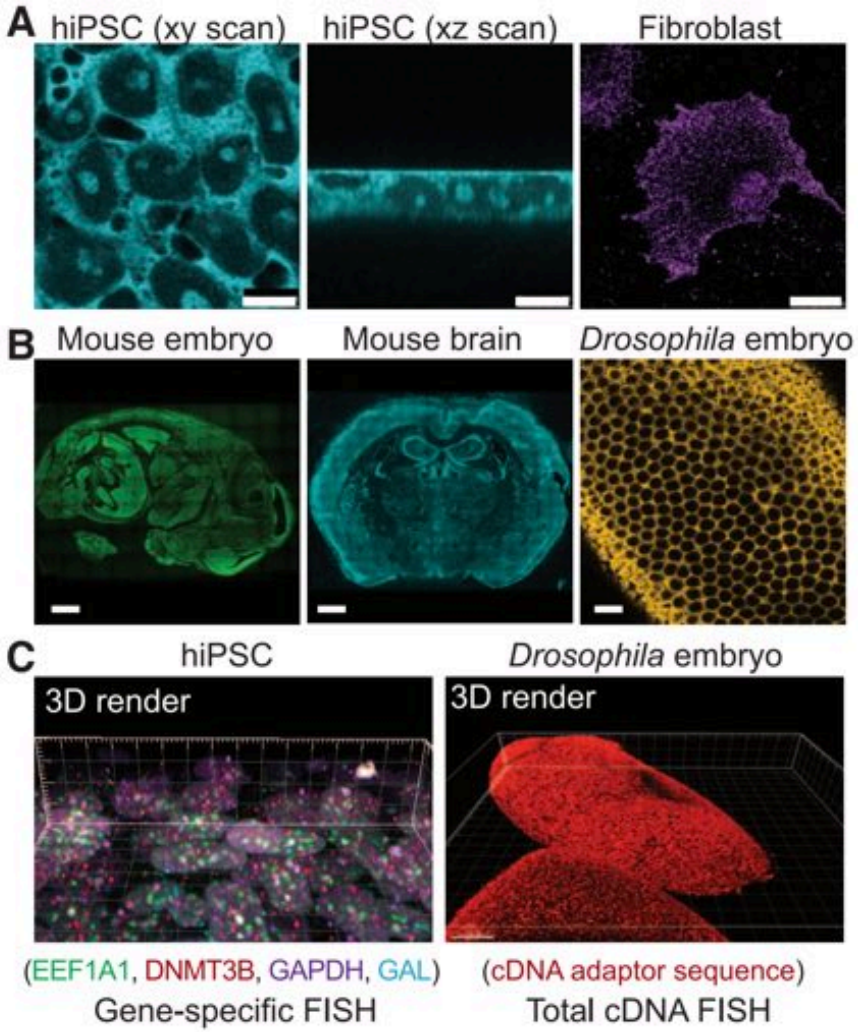
Trapnell et al (2014) Nature Biotechnology. doi:10.1038/nbt.2859



Spatial reconstruction of single-cell gene expression data ("Seurat")

Satija et al (2015) Nature Biotechnology. doi:10.1038/nbt.3192





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

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

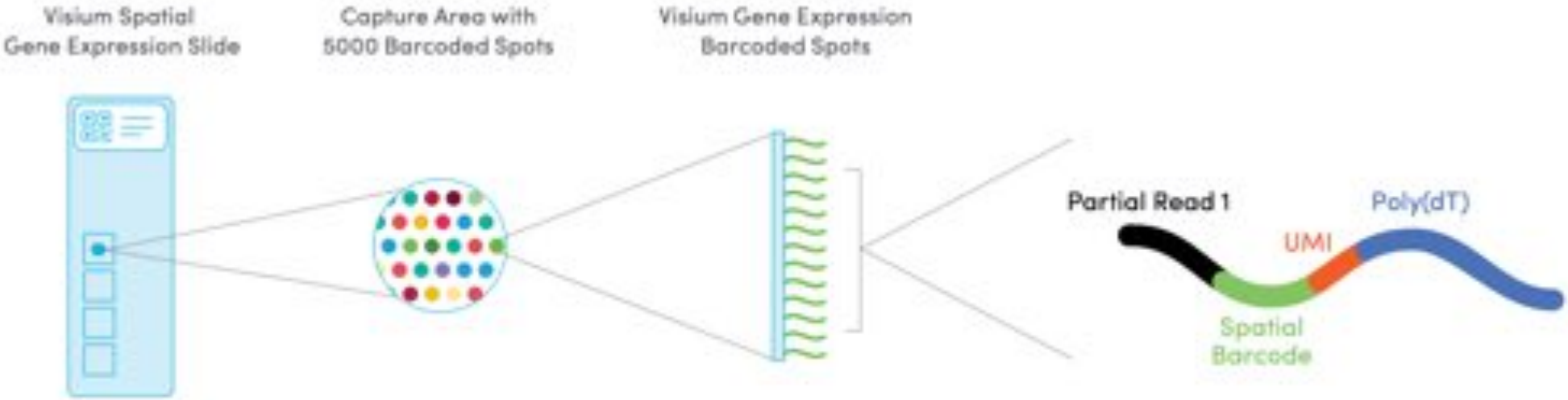


Figure 1. Here we show the composition of the Visium Spatial Gene Expression slide. Each slide contains four Capture Areas with approximately 5000 barcoded spots, which in turn contain millions of spatially-barcoded capture oligonucleotides. Tissue mRNA is released and binds to the barcoded oligos, enabling capture of gene expression information.

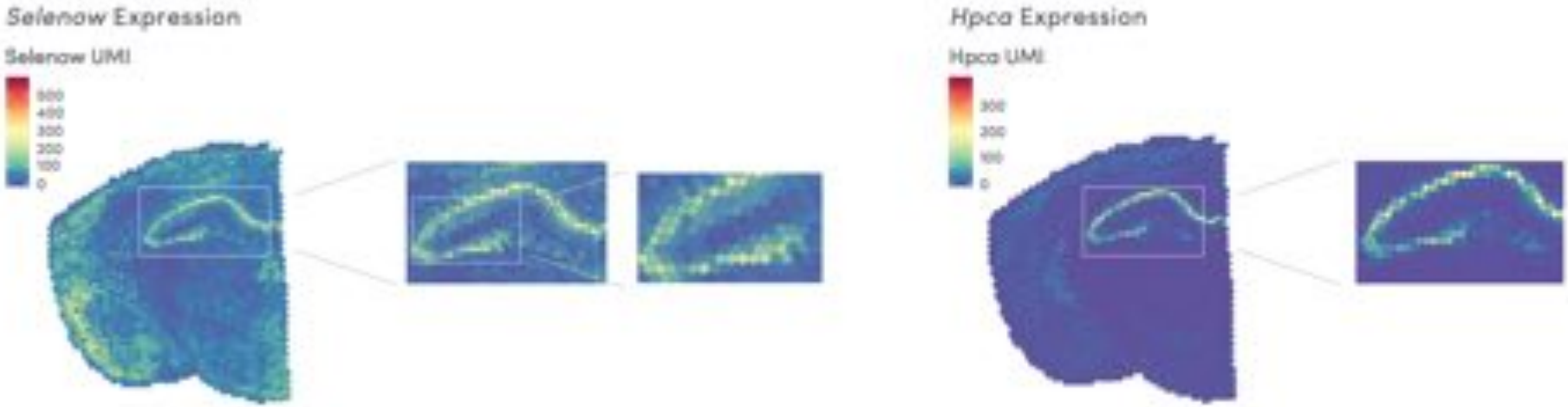


Figure 2. This is a coronal mouse brain section with overlaid spatial gene expression information. The spots correspond to localized mRNA of *Selenow* and *Hpca*, both known to have predominant hippocampal expression.

10xgenomics.com/spatial-gene-expression

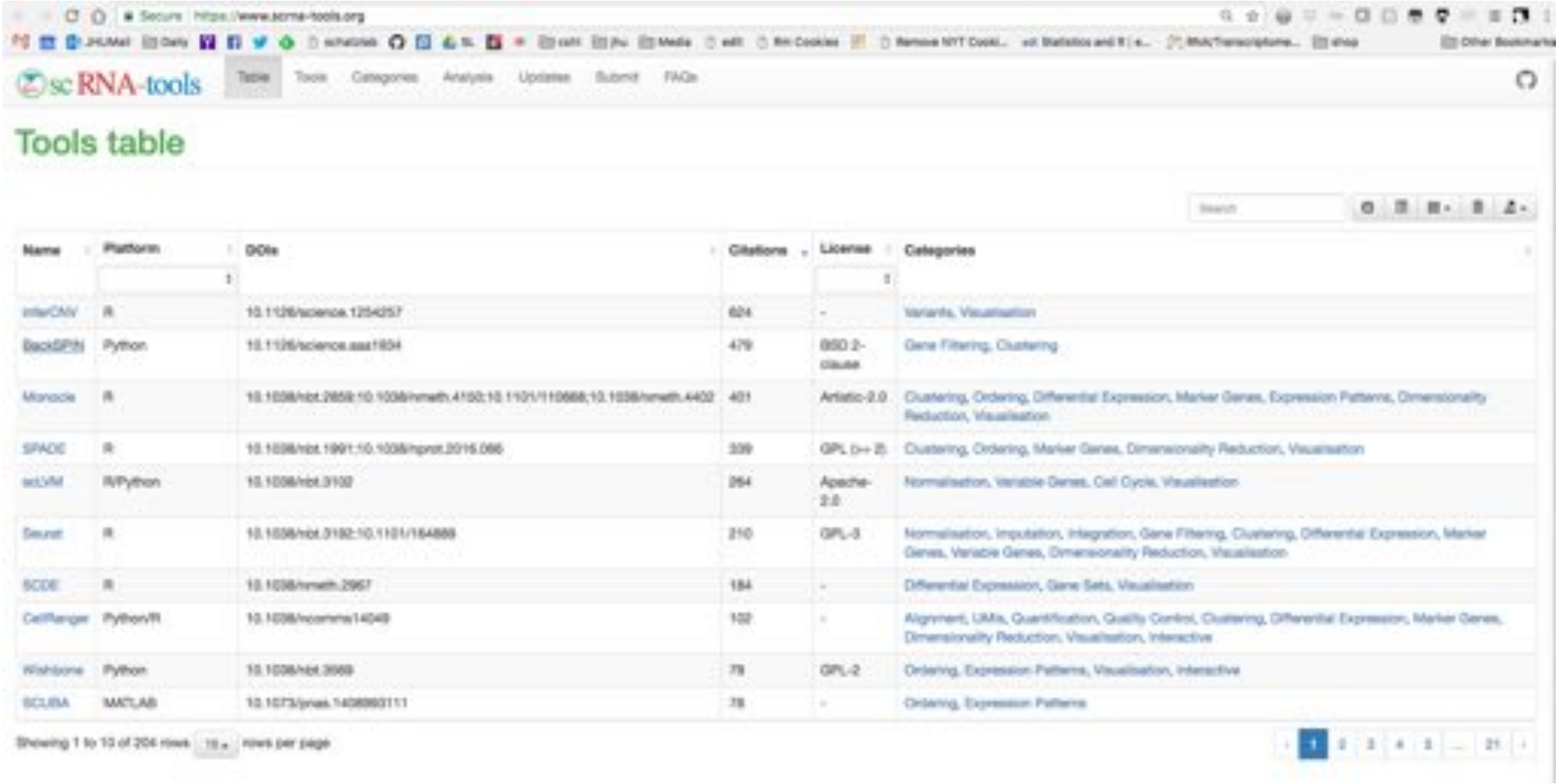
10x GENOMICS®

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

scRNA Analysis Tools: 607 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 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