Lecture 15. RNAseq and scRNAseq

Michael Schatz

March 15, 2021

Applied Comparative Genomics

Project Proposal: Due March 15

Project Proposal

Assignment Date: Monday March 8, 2021
Due Date: Monday, March 15 2021 @ 11:59pm

Review the Project Ideas page

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

The proposal should have the following components:

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

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!

Assignment 5: Due Wed Mar 17

140 lines (89 sloc) 8.82 KB Raw Blame

Assignment 5: BWT and RNA-seq

Assignment Date: Wednesday, Mar. 3, 2021
Due Date: Wednesday, Mar. 17 , 2021 @ 11:59pm

Assignment Overview

In this assignment you will write a simple BWT encoder and decoder, and 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. BWT Encoding [10 pts]

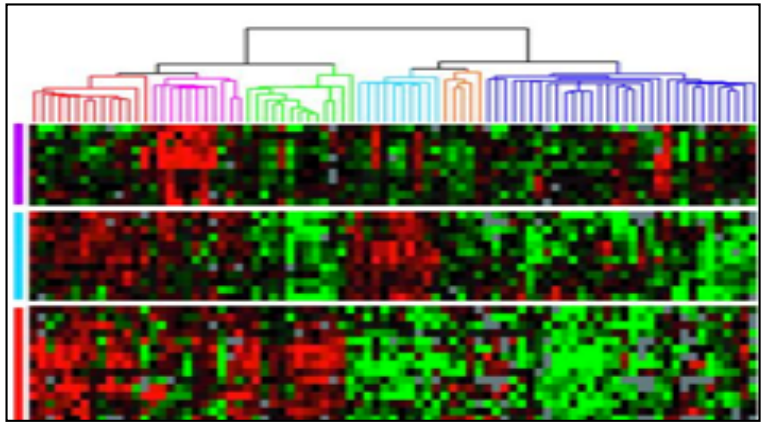
In the language of your choice, implement a BWT encoder and encode the string below. Linear time methods exist for computing the BWT, although for this assignment you can use the simple method based on standard sorting techniques. Your solution does *not* need to be an optimal algorithm and can use O(n^2) space and O(n^2 lg n) time.

Here is the recommended pseudo code (make sure to submit your code as well as the encoded string):

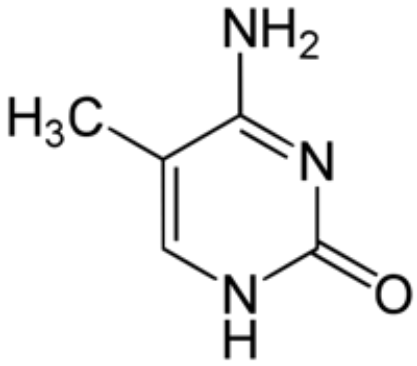
```
computeBWT(string s)
  ## add the magic end-of-string character
  s = s + "$"

  ## build up the BWM from the cyclic permutations
  ## note the ith cyclic permutation is just "s[i..n] + s[0..i]"
  StringList rows = []
  for (i = 0; i < length(s); i++)
    rows.append(cyclic_permutation(s, i))

  ## just use the builtin sort command
```

*-seq in 4 short vignettes

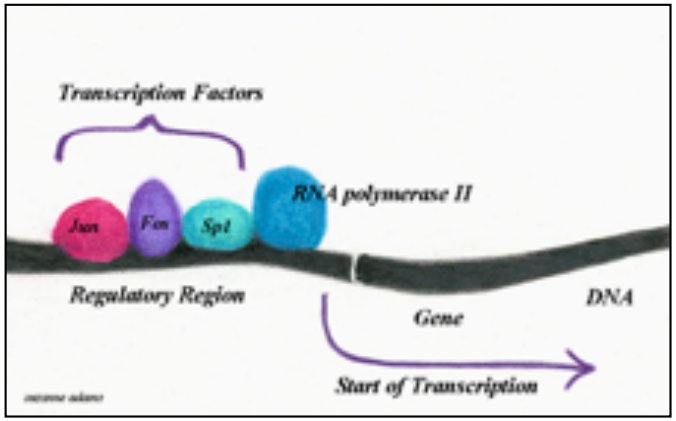
RNA-seq



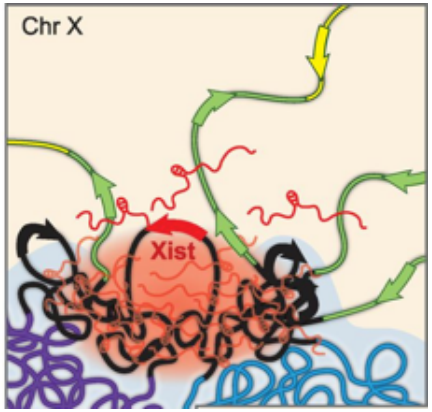
Methyl-seq



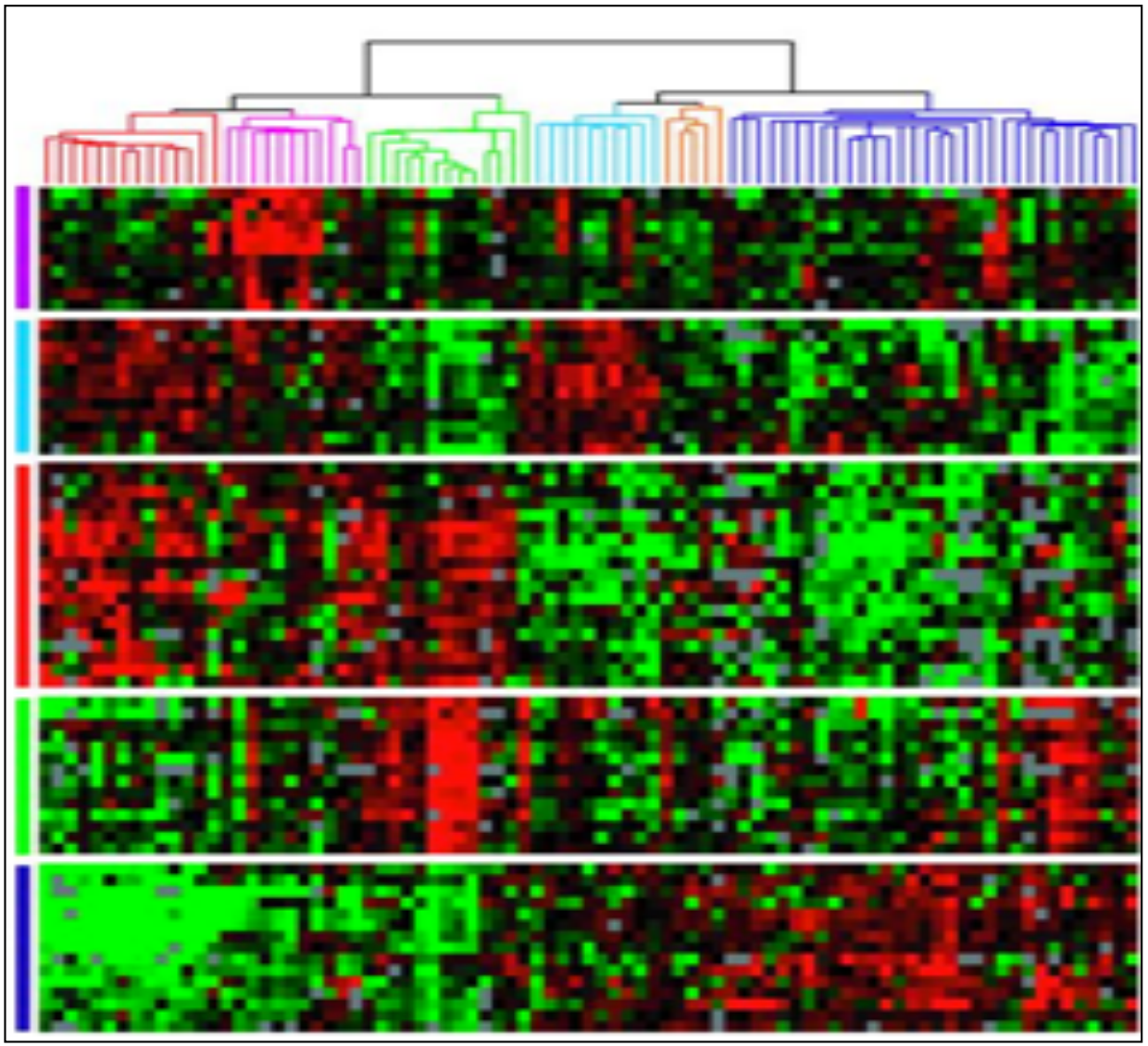
ChIP-seq



Hi-C

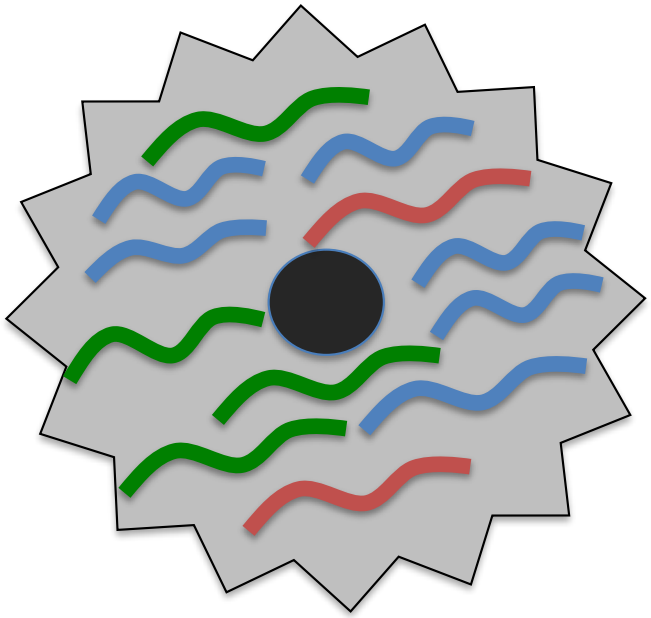


RNA-seq

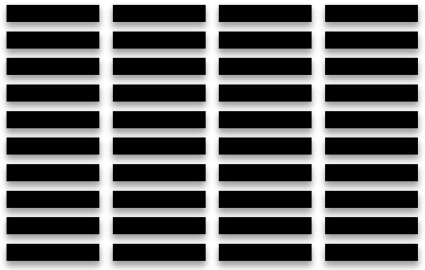


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

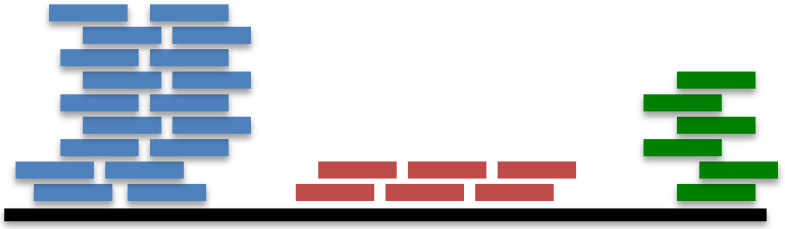
RNA-seq Overview



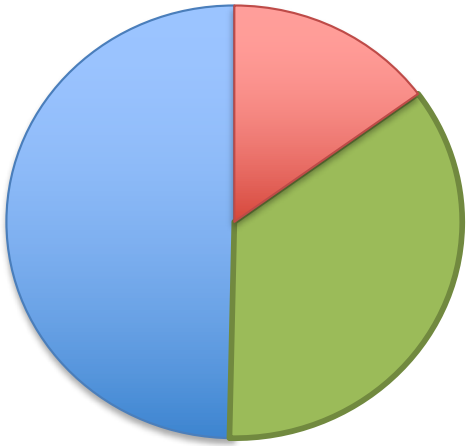
Sequencing



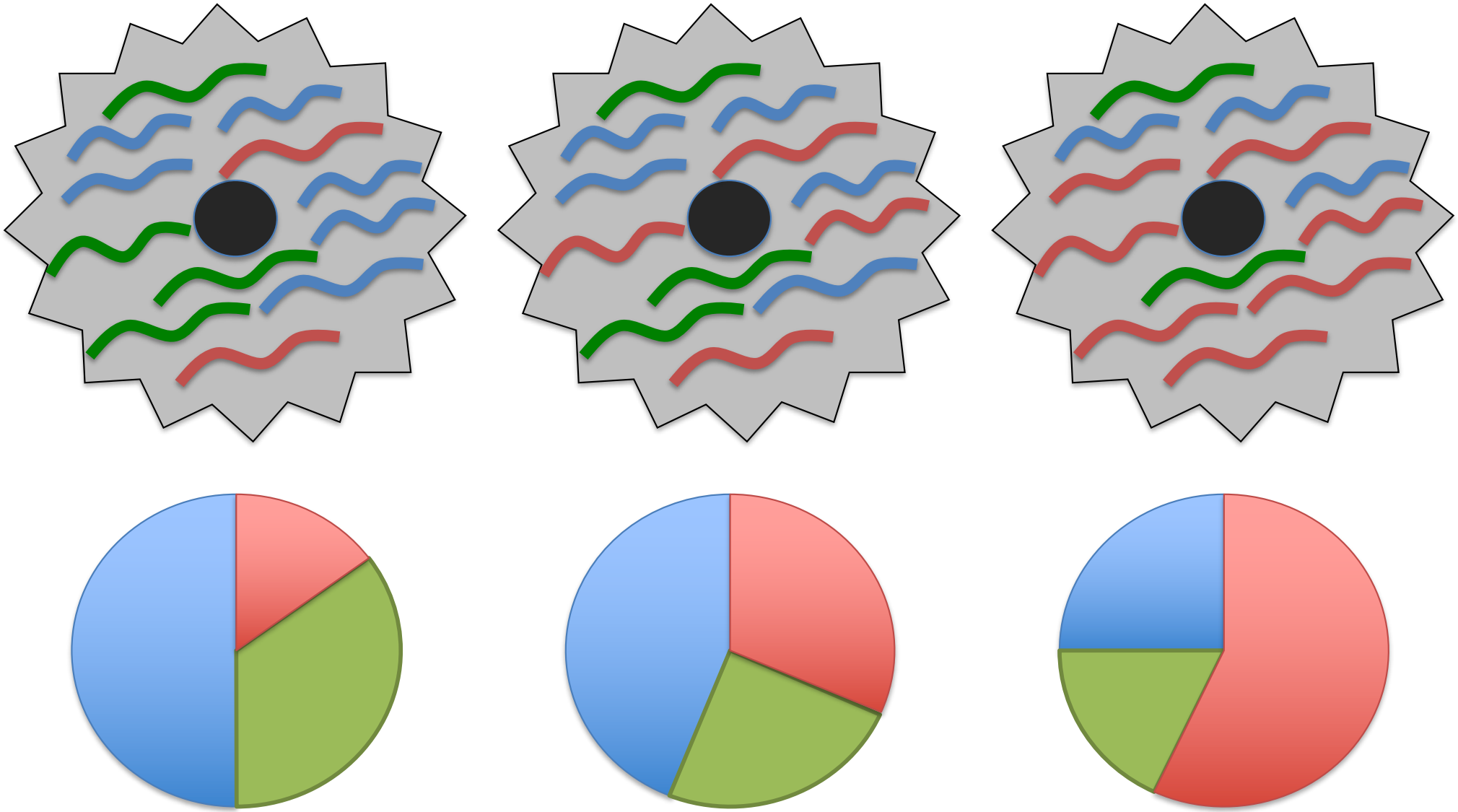
Mapping
& Assembly



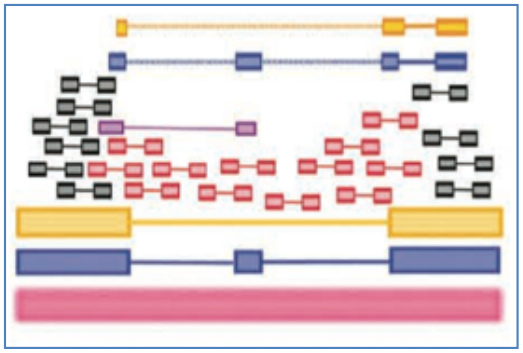
Quantification



RNA-seq Overview



RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

RNA-Seq Approaches

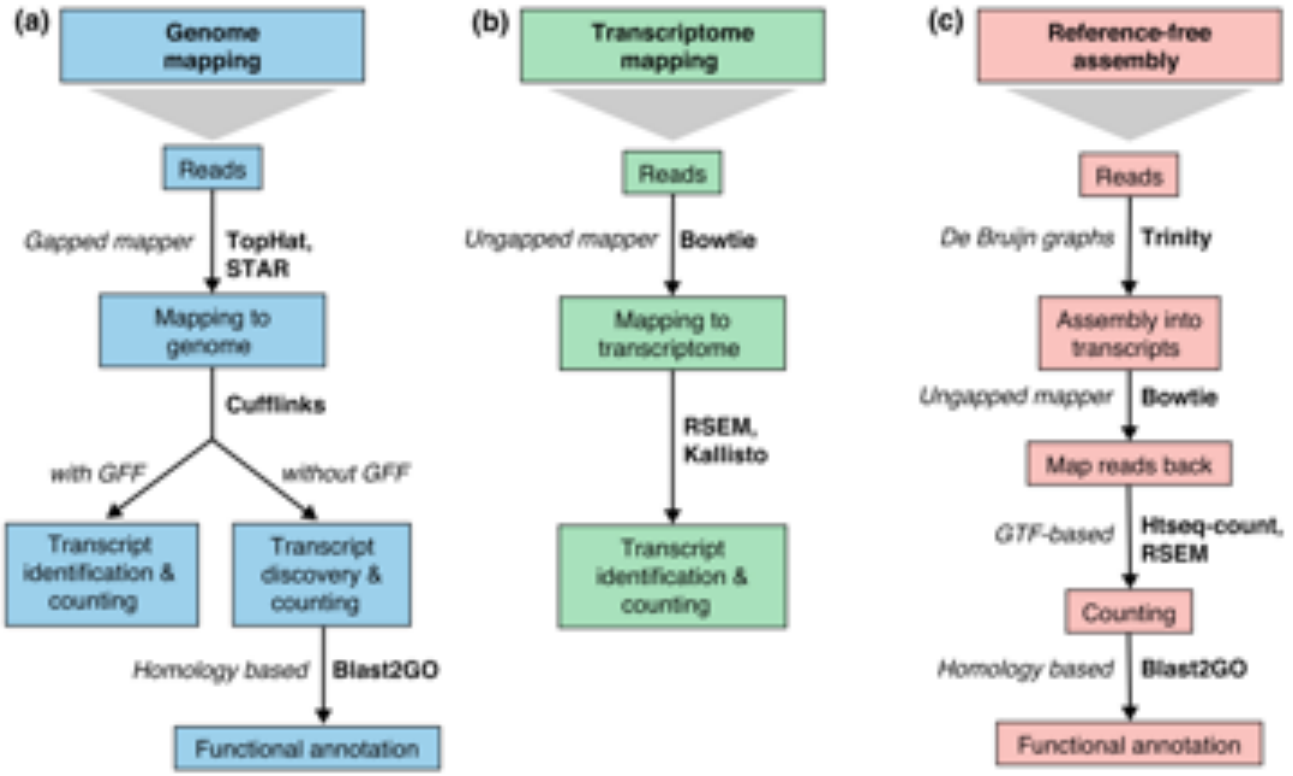


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in (**b**) followed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used at each analysis step are indicated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

A survey of best practices for RNA-seq data analysis
Conesa et al (2016) Genome Biology. doi 10.1186/s13059-016-0881-8

RNA-Seq Approaches

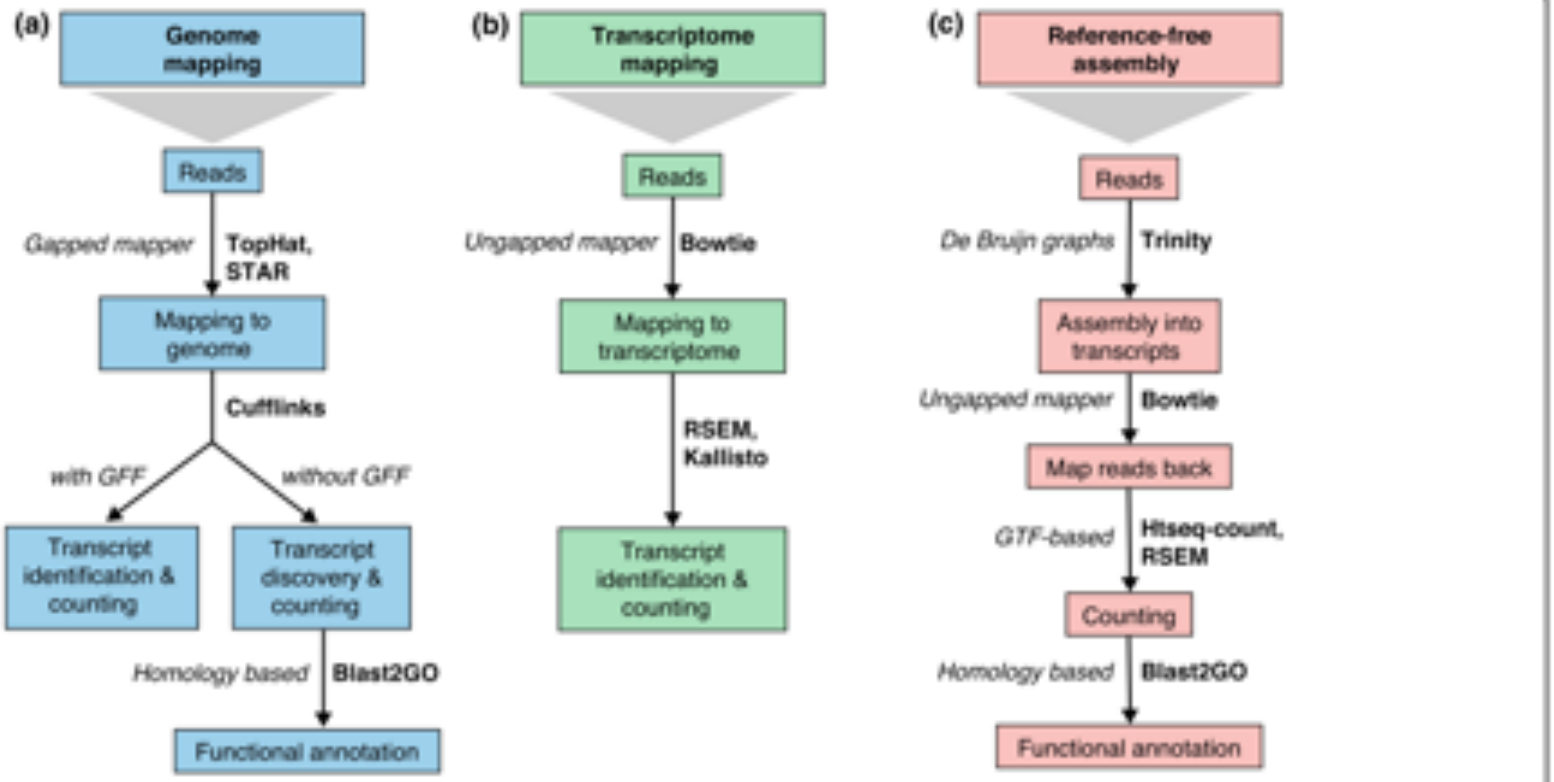


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and re…vel) transcript discovery and quantification can proceed with or without an ann… no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads n… contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further a… owed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used a… ated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

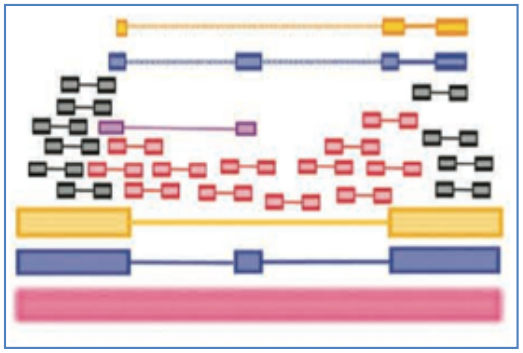
Which approach should we use?

It depends….

A survey of best practices for RNA-seq data analysis

Conesa et al (2016) Genome Biology. doi 10.1186/s13059-016-0881-8

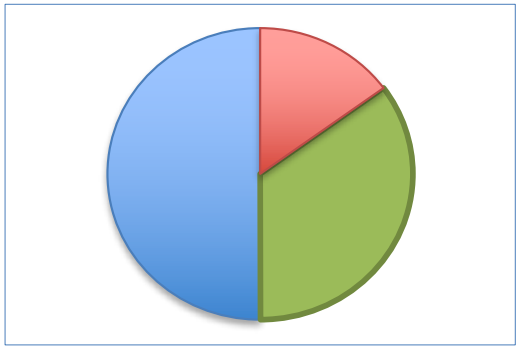
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

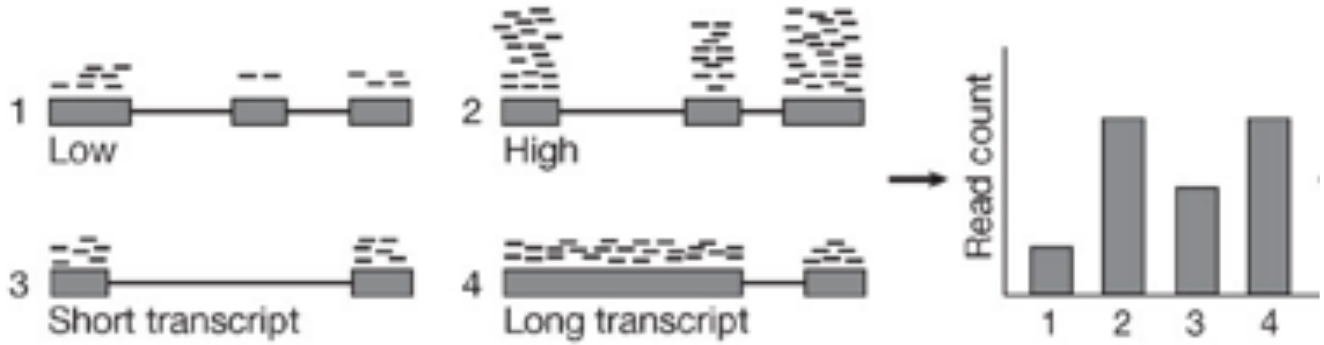
Solution: Use a spliced aligner, and assemble isoforms

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

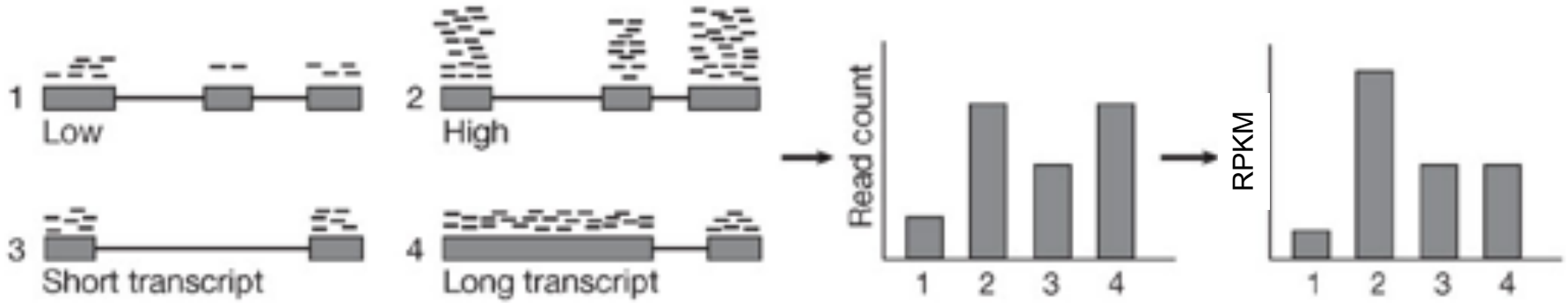


Counting Reads that align to a gene DOESN'T work!

- 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)

RPKM, FPKM, TPM



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

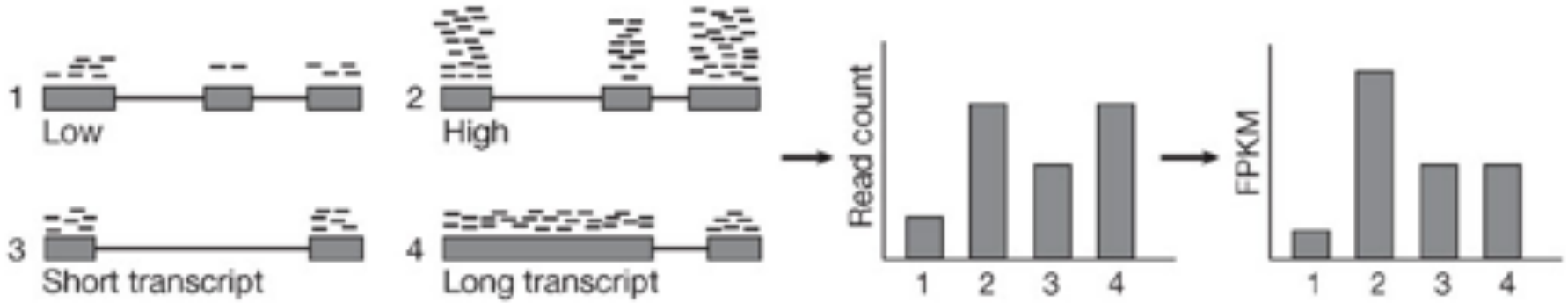
- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

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

(Count reads aligned to gene) / (length of gene in kilobases) / (# millions of read mapped)

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

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)

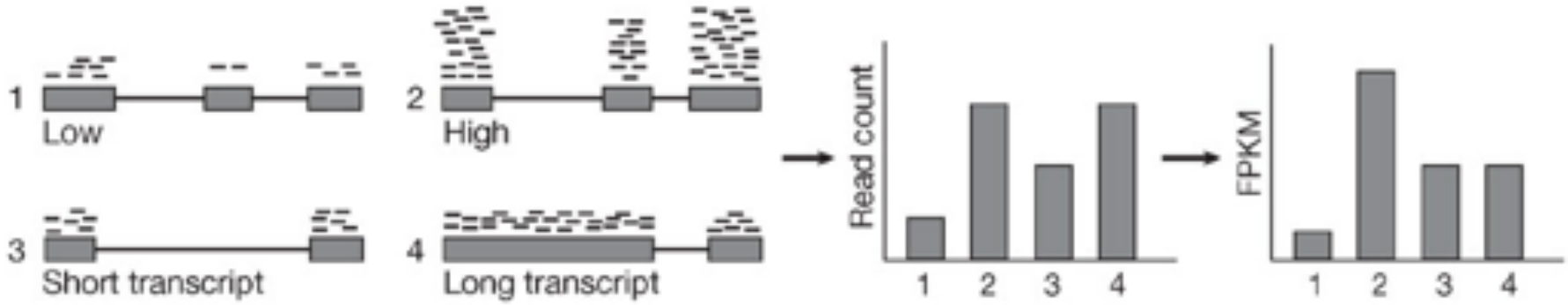
=> Wait a second, reads in a pair arent independent!

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

$\Rightarrow$ Does a much better job with short exons & short genes by boosting coverage

$\Rightarrow$ Wait a second, FPKM depends on the average transcript length!

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!

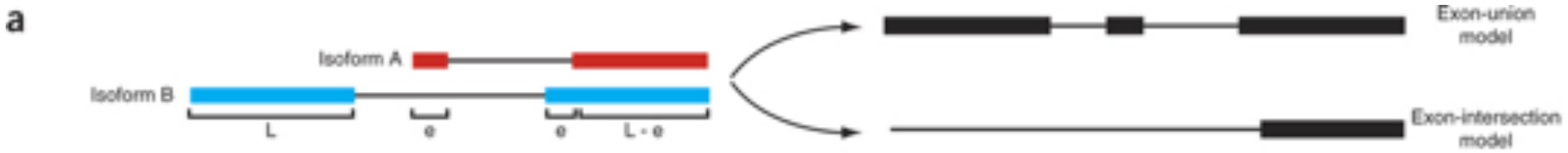
3. TPM: Transcripts Per Million (Li et al, 2011)

⇒ 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

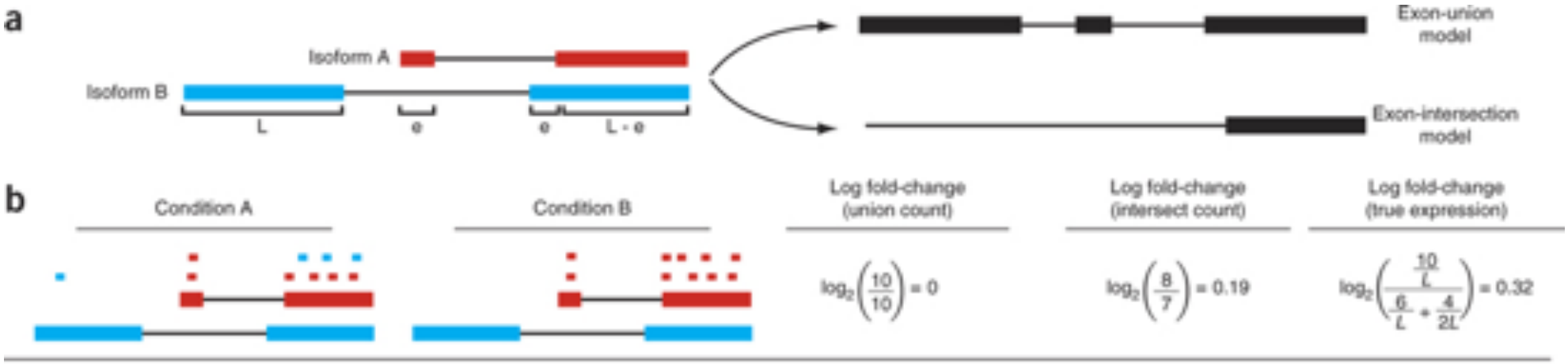
$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

Gene or Isoform Quantification?



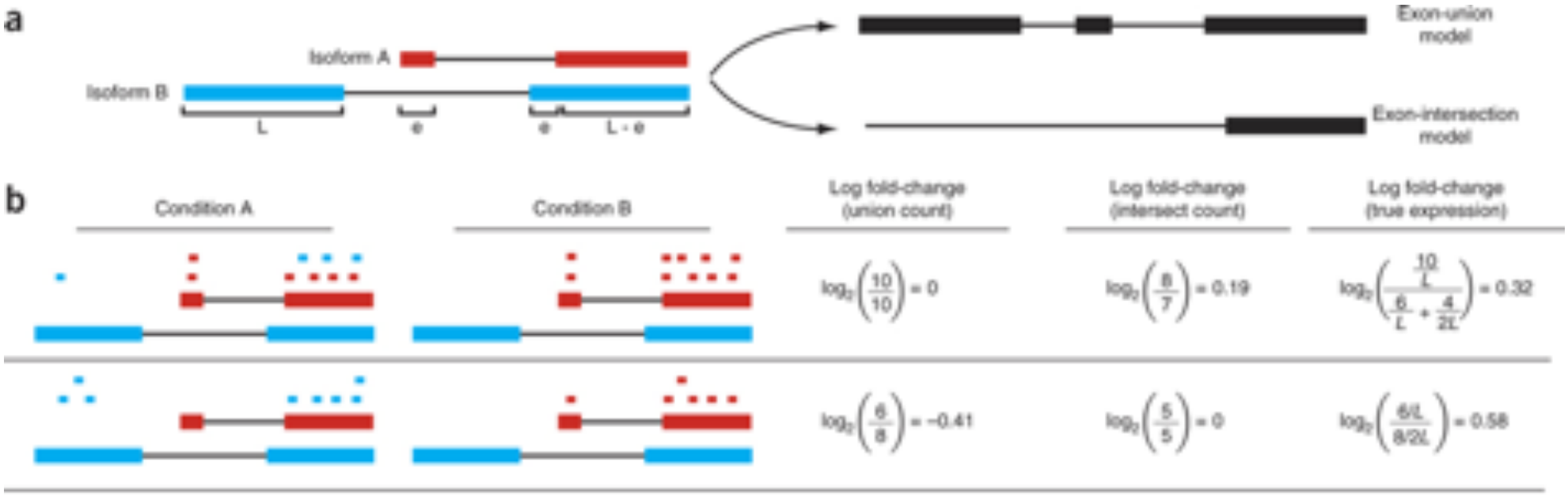
Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

Gene or Isoform Quantification?



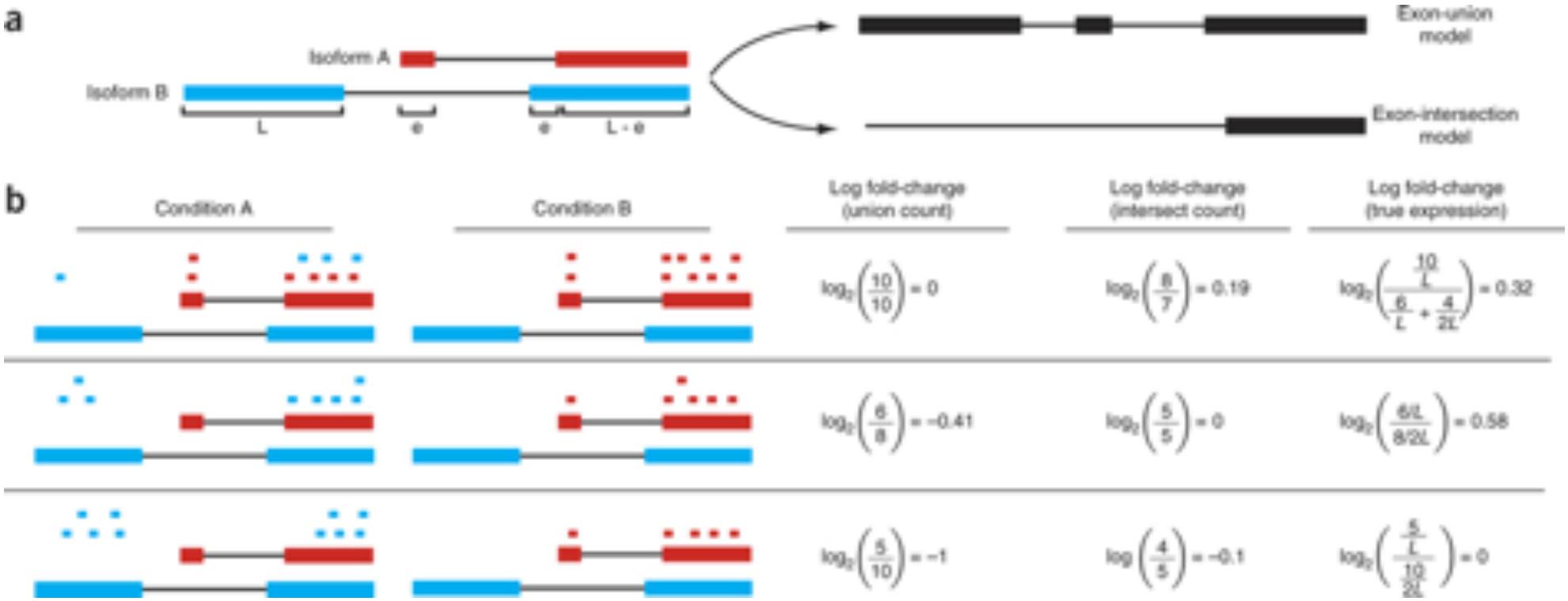
Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

Gene or Isoform Quantification?



Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

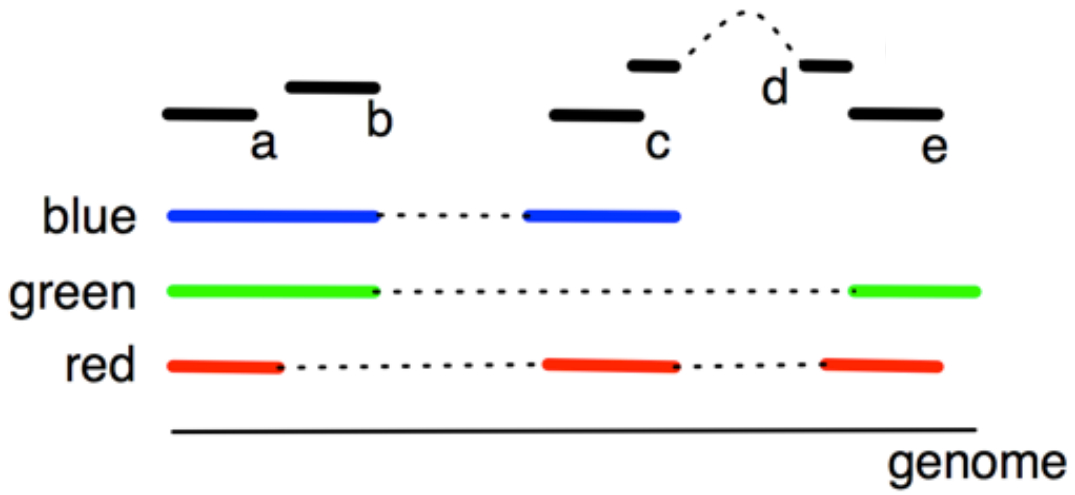
Gene or Isoform Quantification?



Key point : The length of the actual molecule from which the fragments derive is crucially important to obtaining accurate abundance estimates.

Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

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

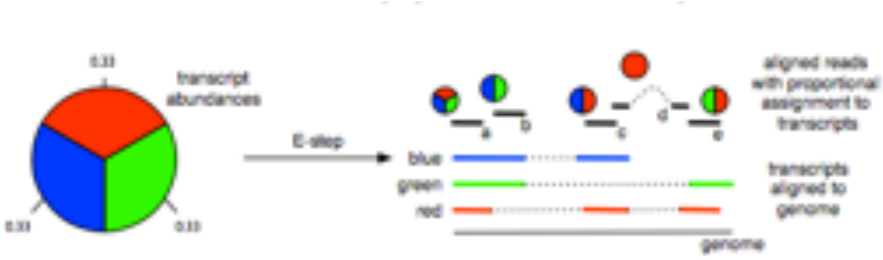
There are five reads (a,b,c,d,e) mapping to the gene.

- 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



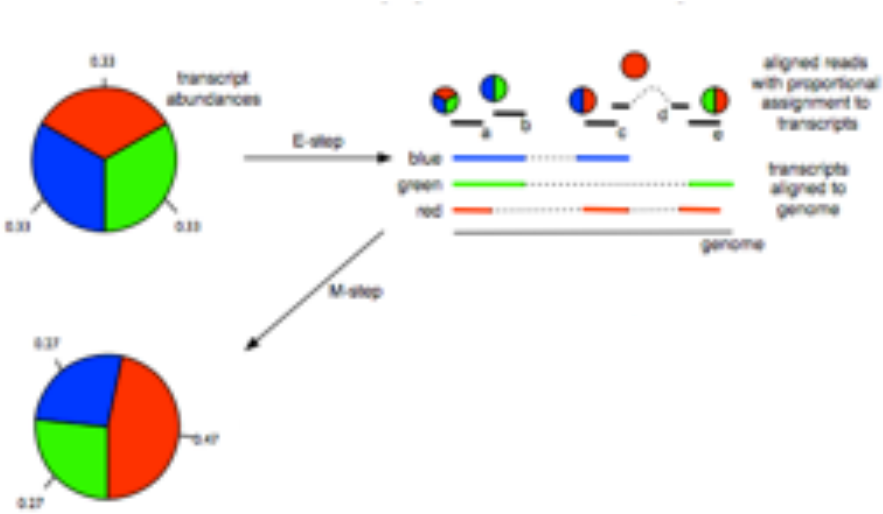
The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

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)

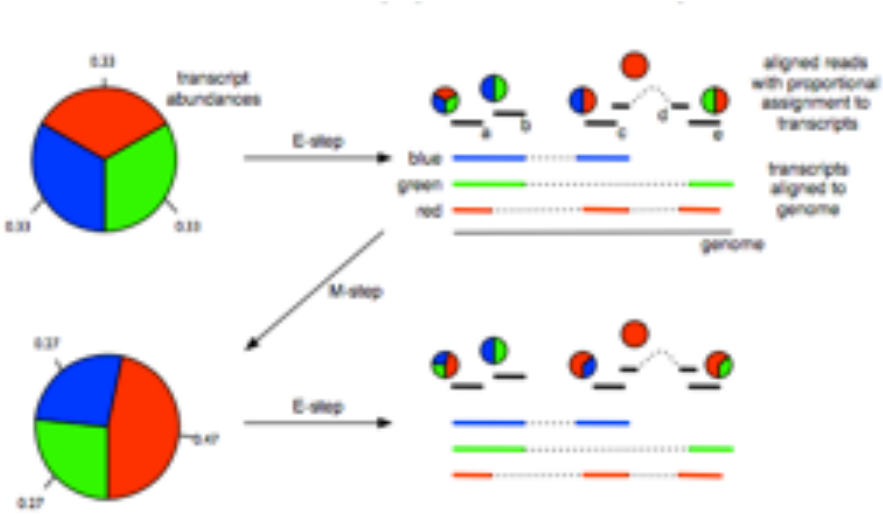
There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

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.

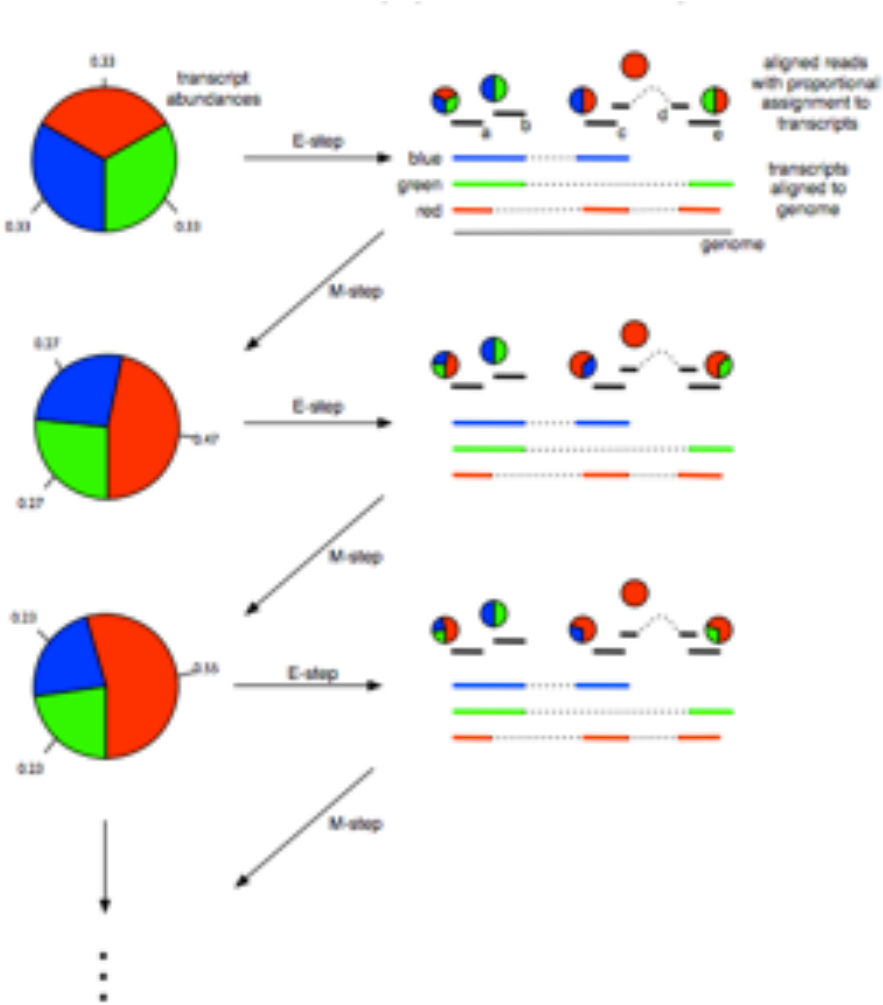
During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

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.

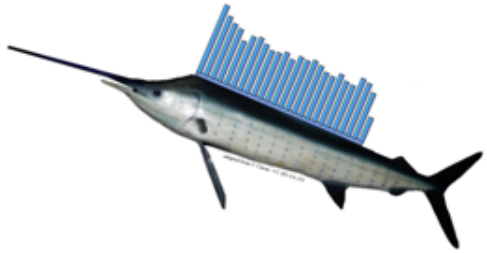
During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

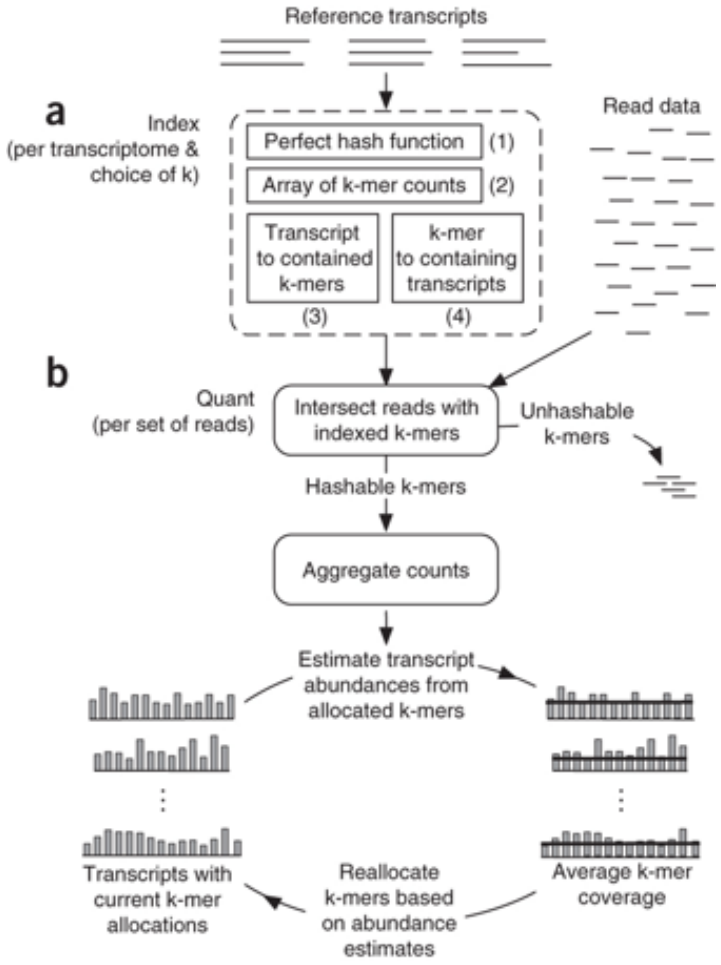
Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

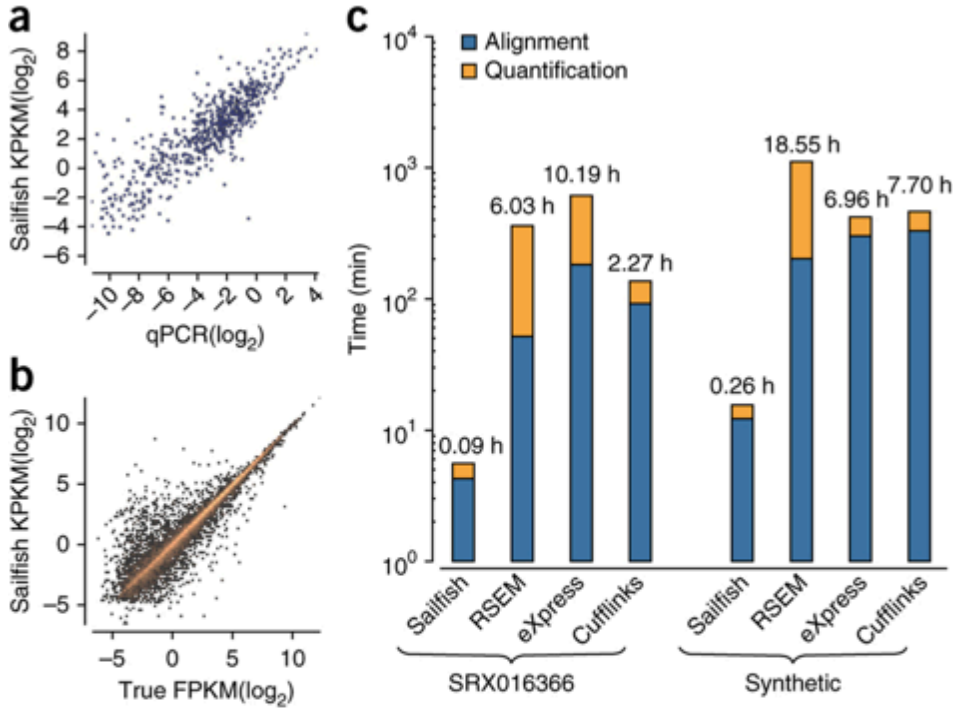
Repeat until convergence!

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

Sailfish: Fast & Accurate RNA-seq Quantification





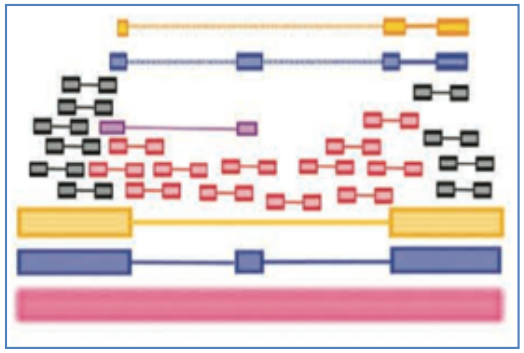


d

	Human brain tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.85	0.82	0.85	0.85	0.96	0.96	0.95	0.94
Spearman	0.84	0.80	0.85	0.85	0.76	0.77	0.77	0.75
RMSE	–	–	–	–	6.06	8.90	8.83	10.05
medPE	–	–	–	–	6.50	12.48	14.06	12.42

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms
Patro et al (2014) Nature Biotechnology 32, 462–464 doi:10.1038/nbt.2862

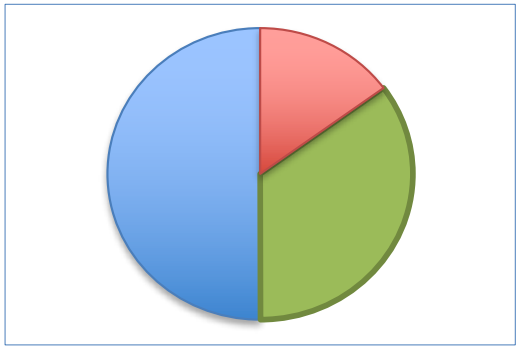
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

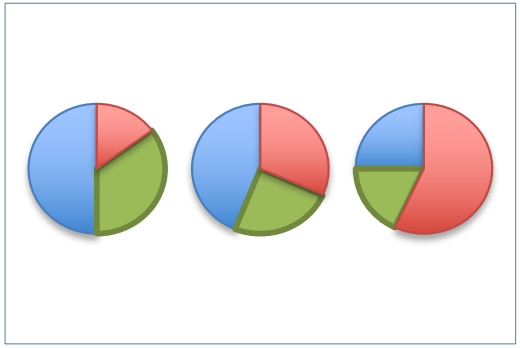
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

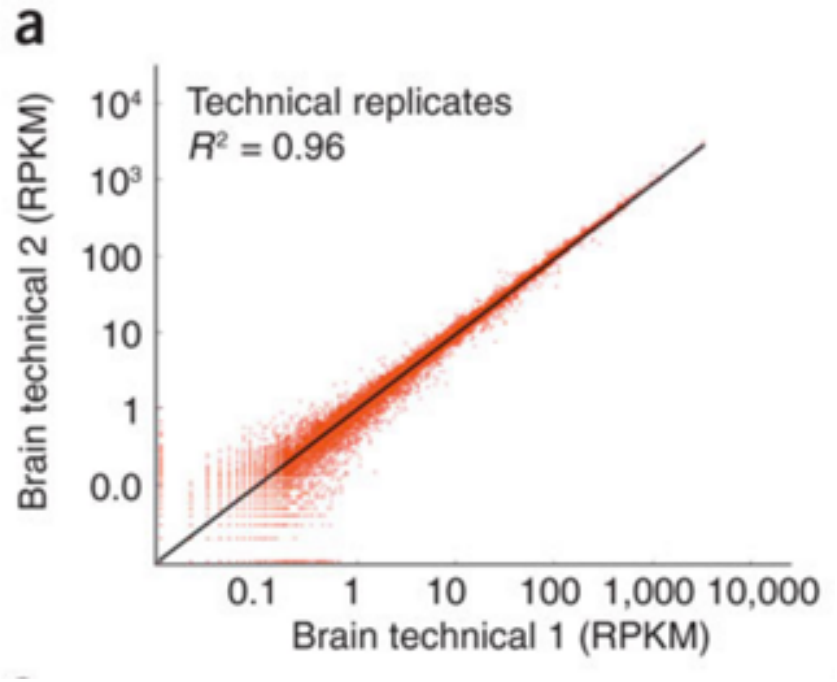
Solution: Infer underlying abundances (e.g. TPM)

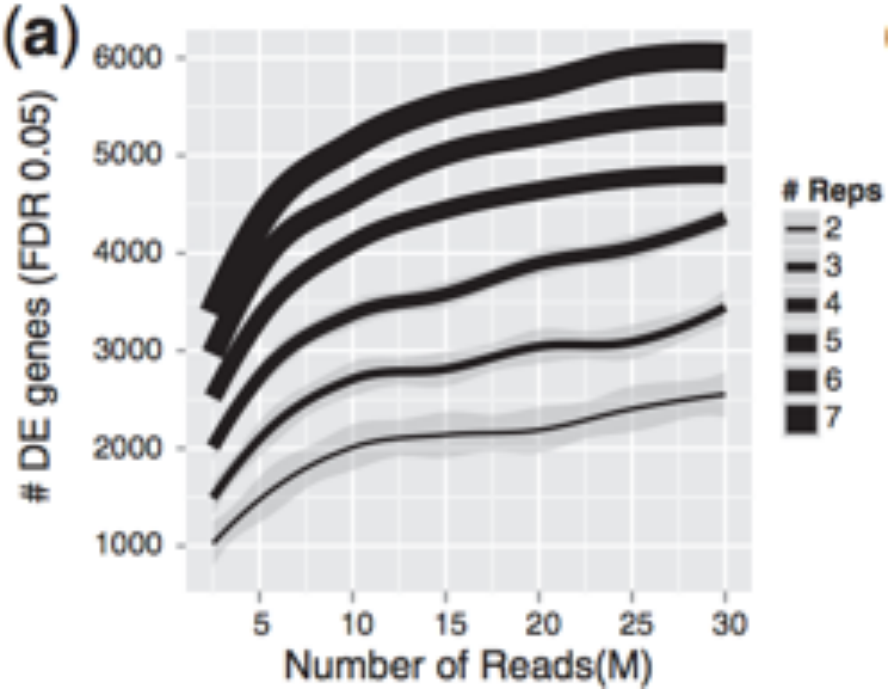
Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

How Many Replicates?



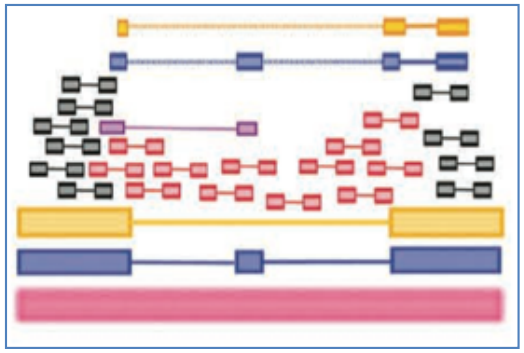


Why don't we have perfect replicates?

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

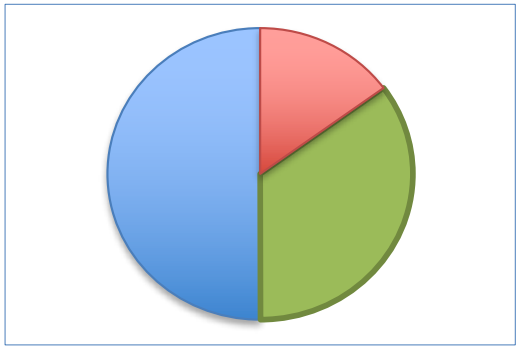
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

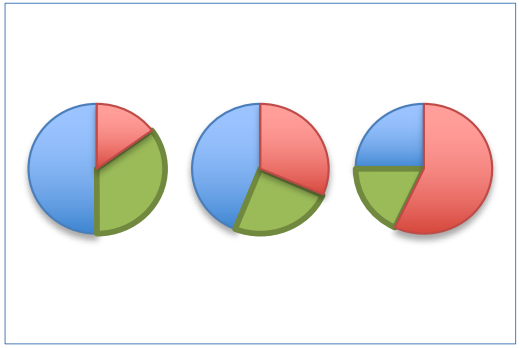
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



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

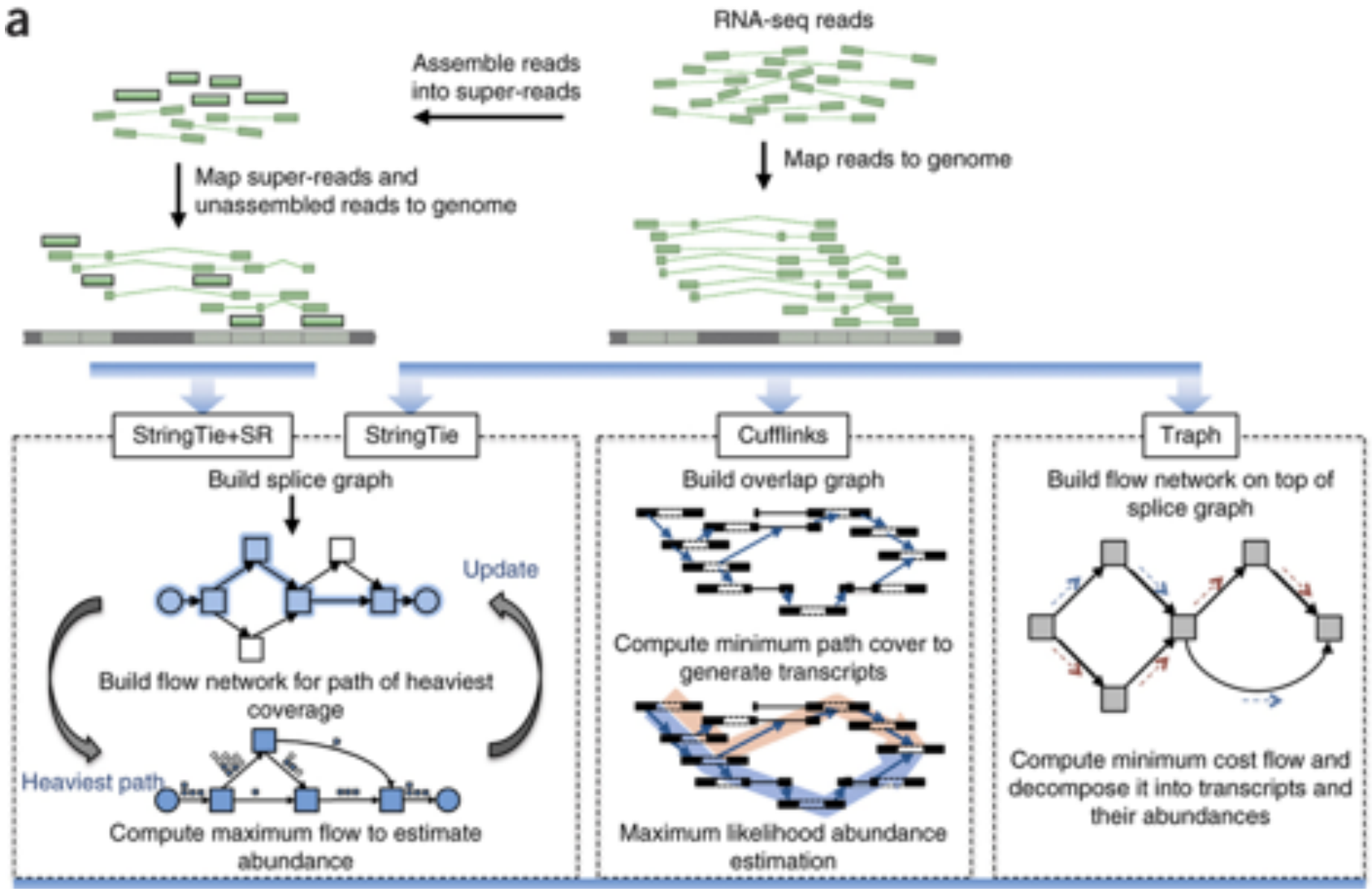


Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

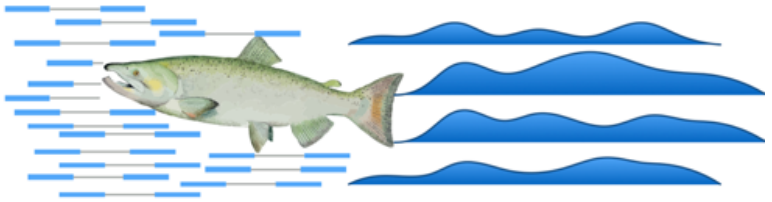
RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics.* doi:10.1093/bioinformatics/btt688

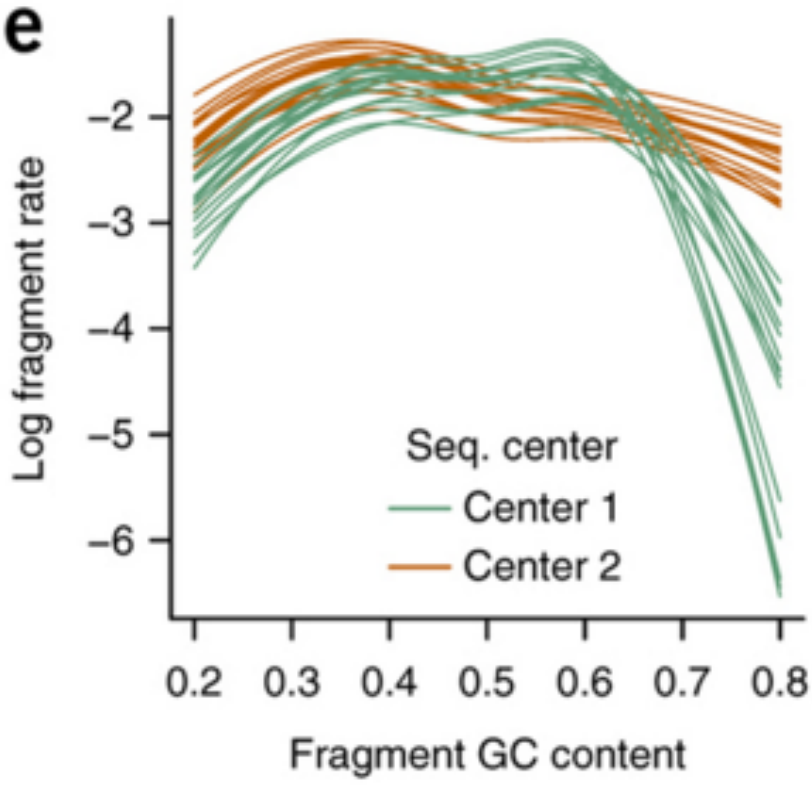
Isoform Quantification Approaches



StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
Pertea M, et al. (2015) Nature Biotechnology. doi: 10.1038/nbt.3122.

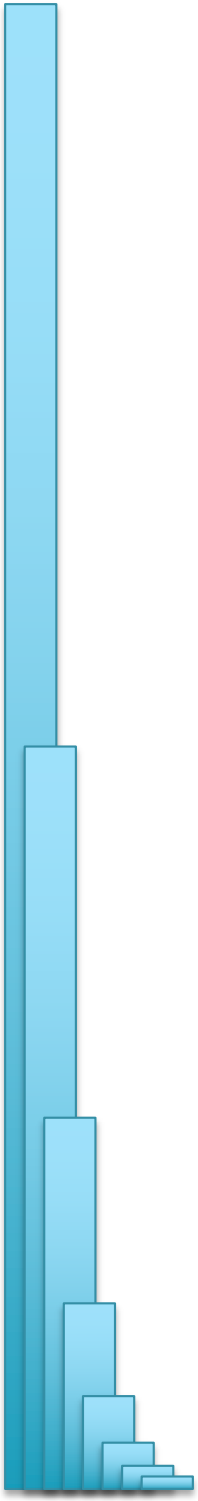
Salmon: The ultimate RNA-seq Pipeline?





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



Unsupervised Learning
aka Clustering

Clustering Refresher

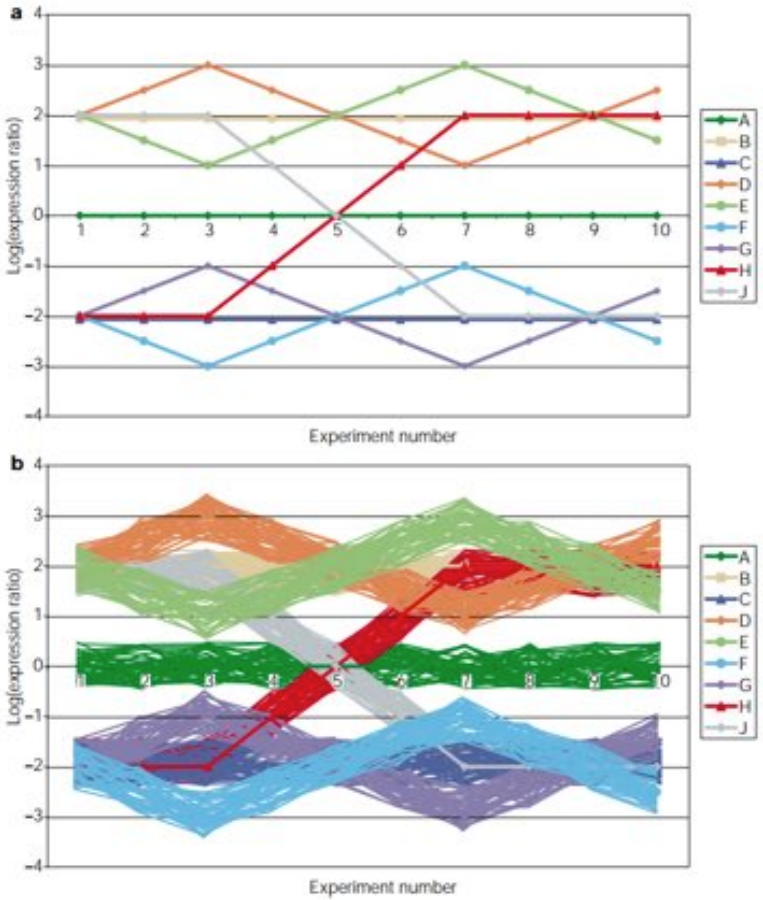
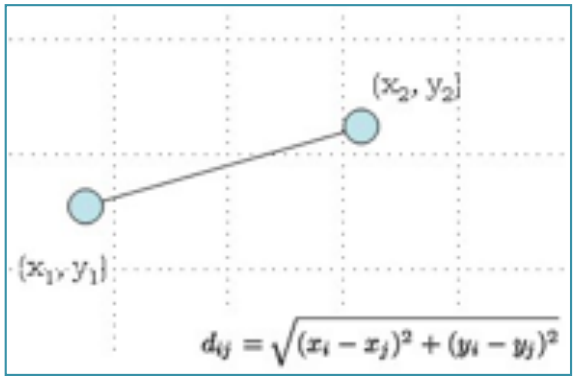
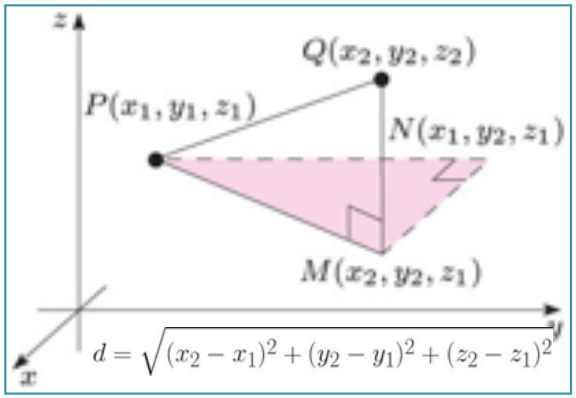


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments.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

Euclidean Distance

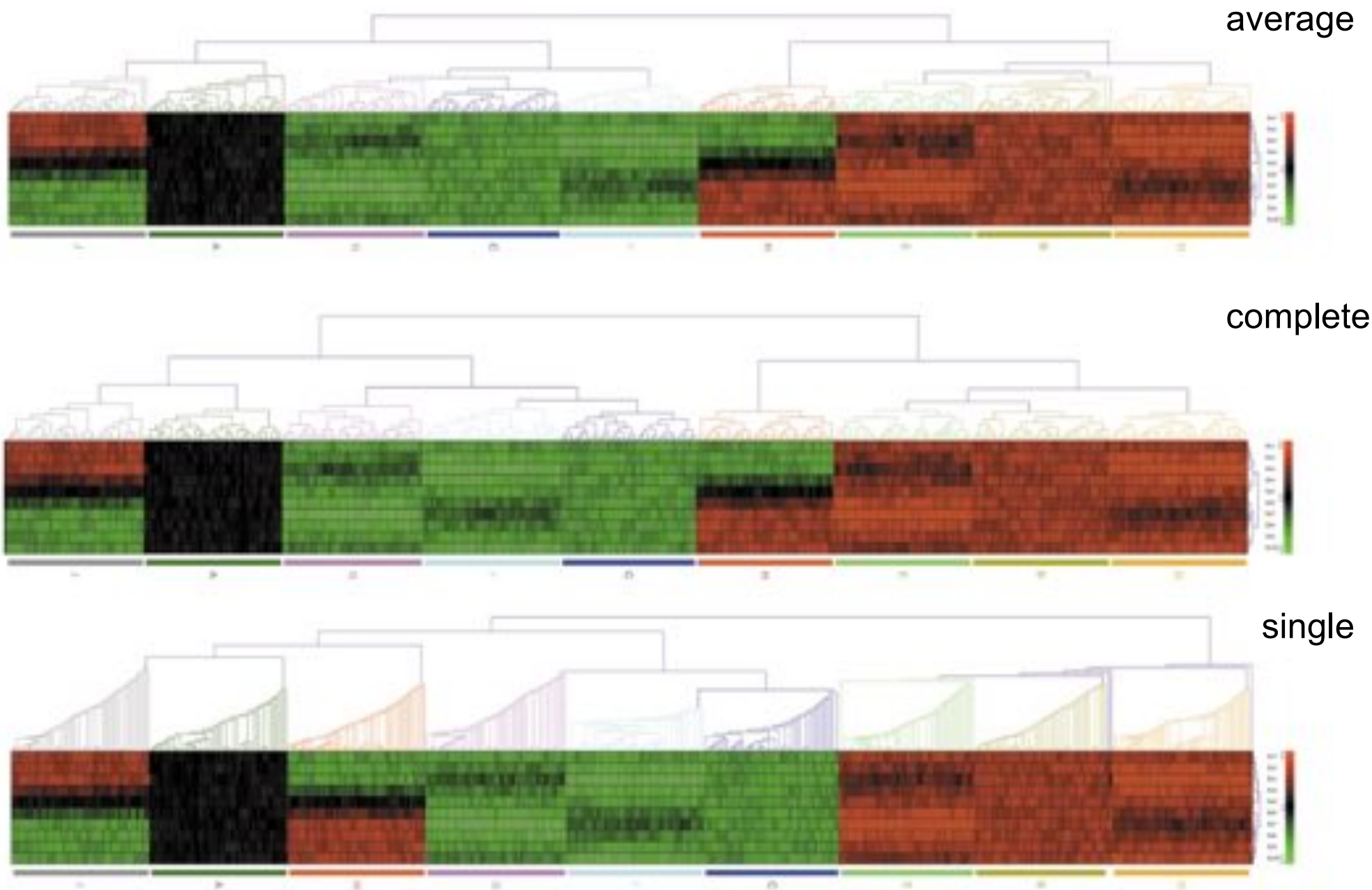




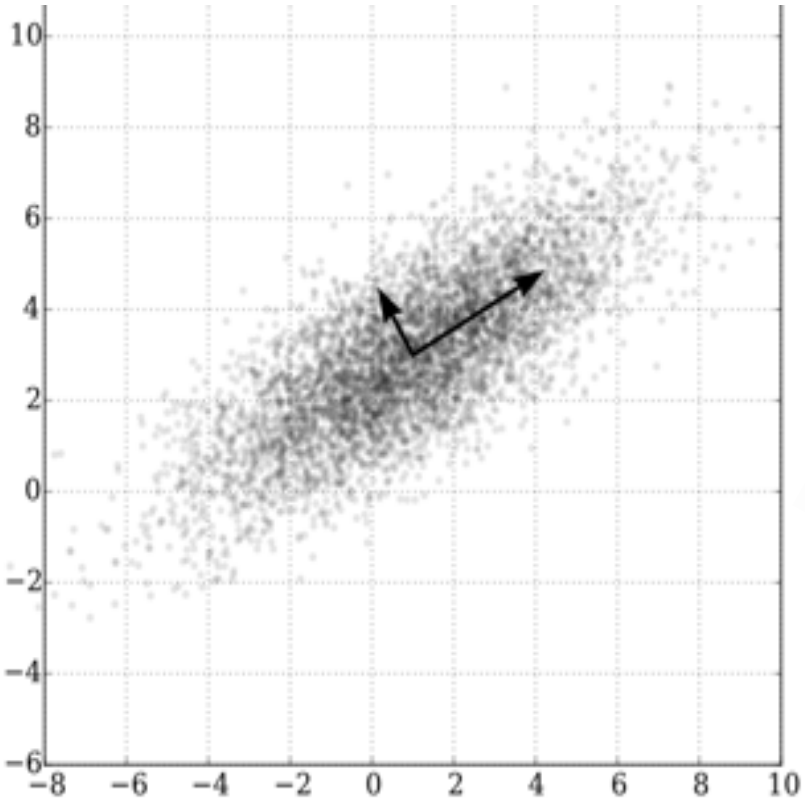
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

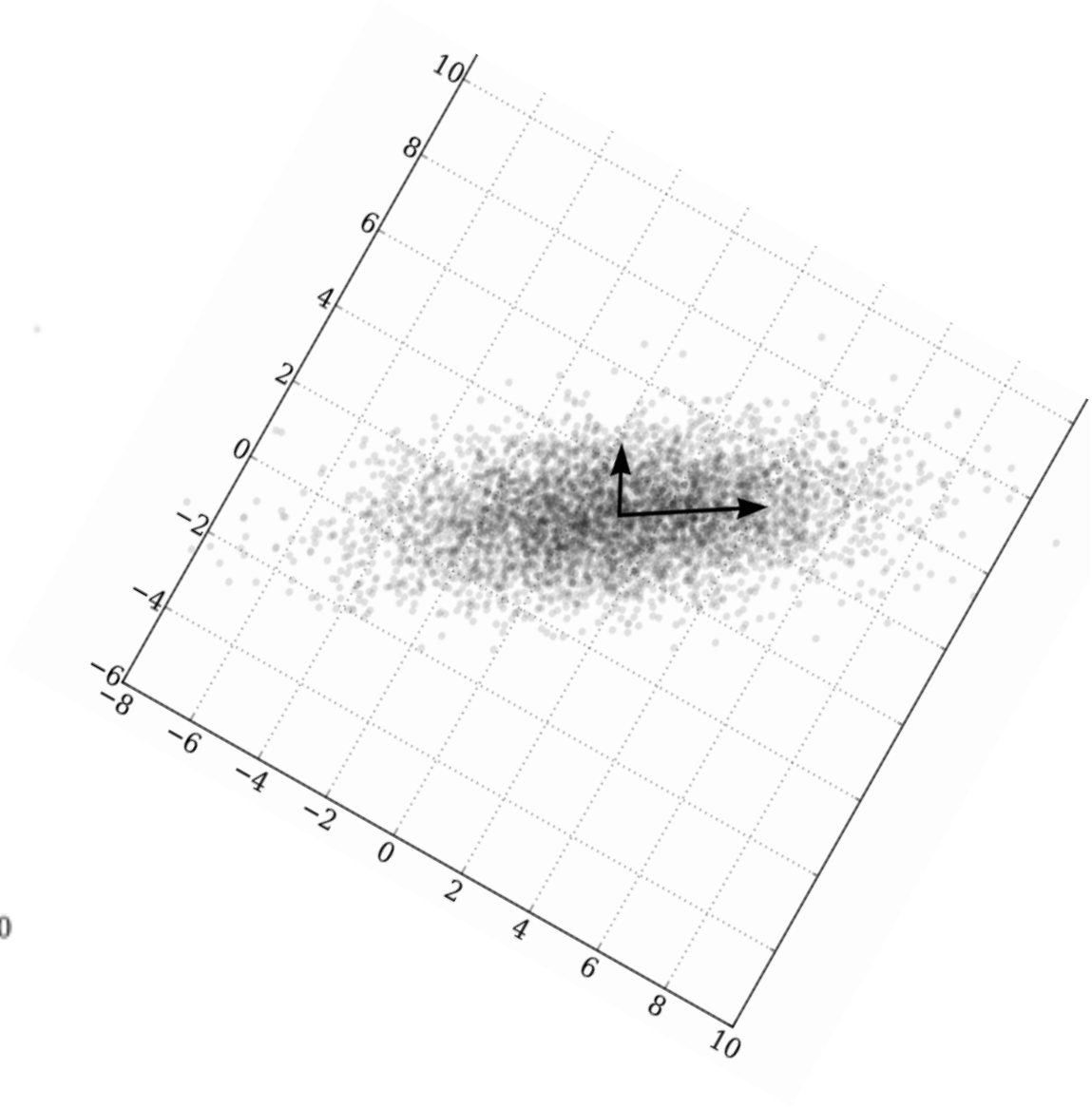
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



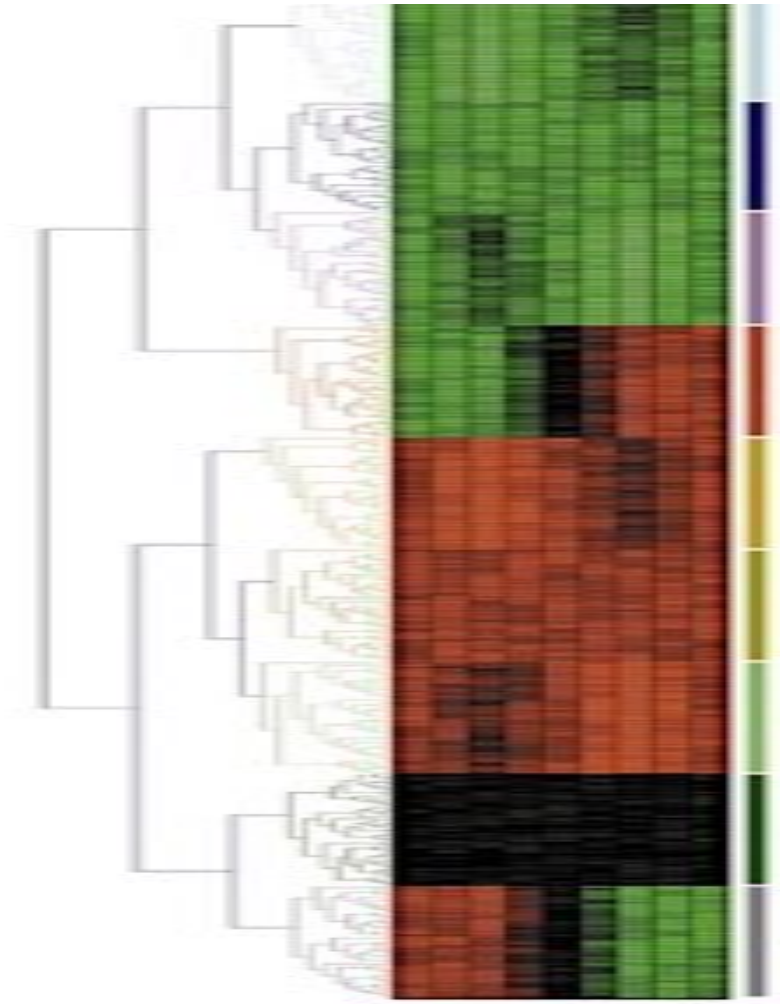
Principle Components Analysis (PCA)

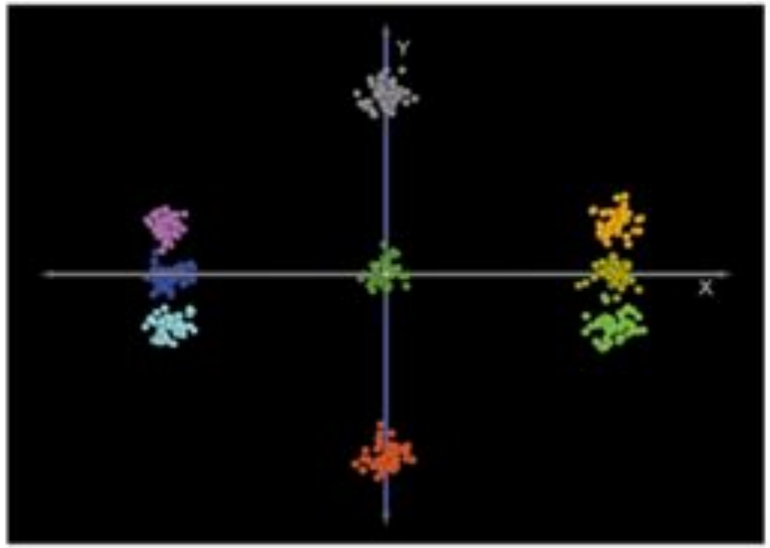




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





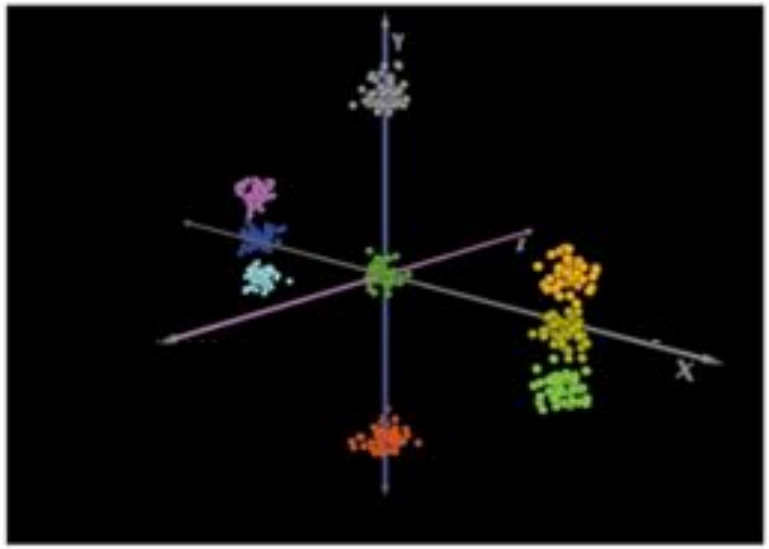
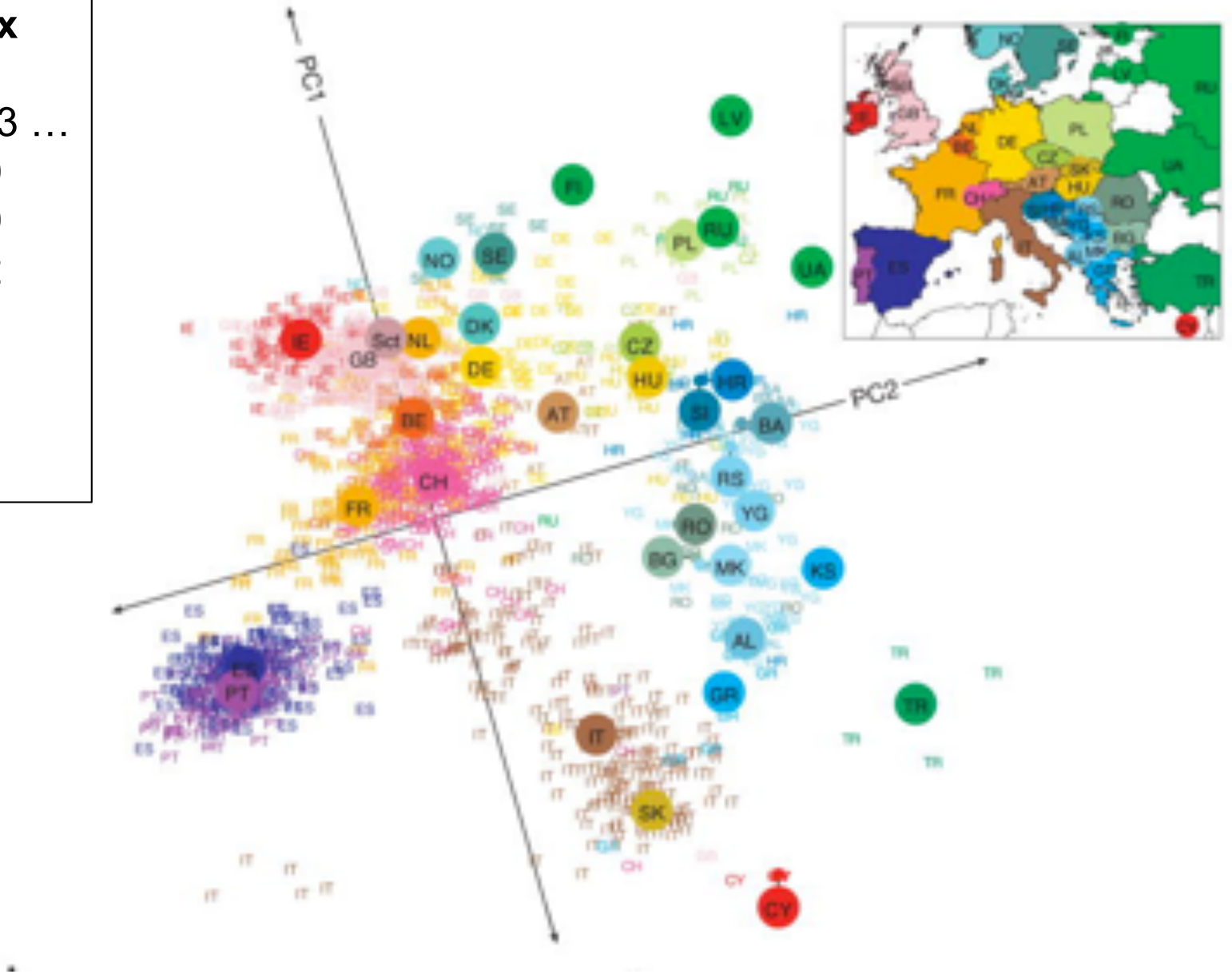


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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

Genotype Matrix

	P1	P2	P3	…
SNP1	0	1	0	
SNP2	1	0	0	
SNP3	0	0	2	
…				

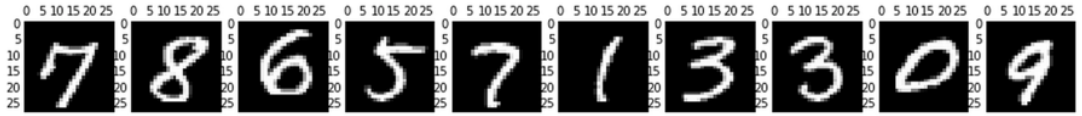
0 = hom ref
1 = het ref/alt
2 = hom alt

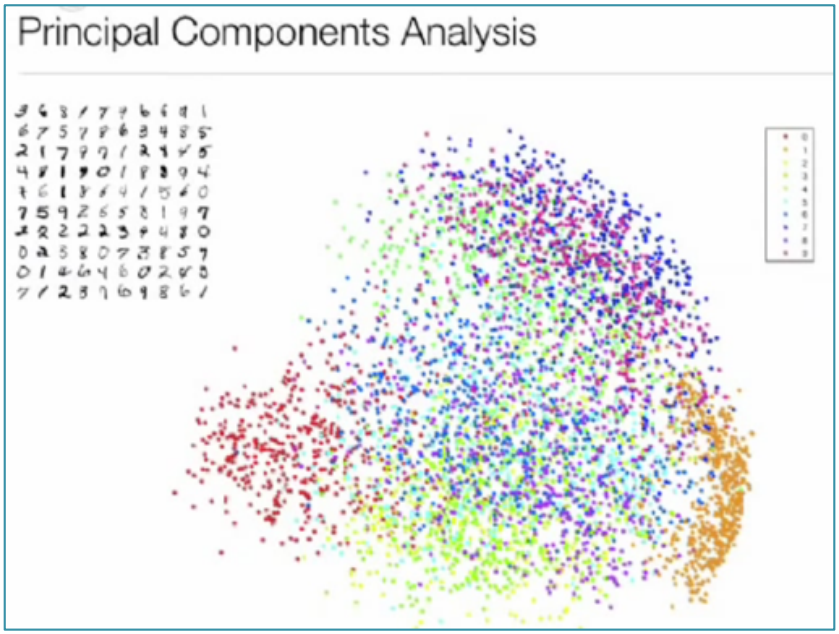


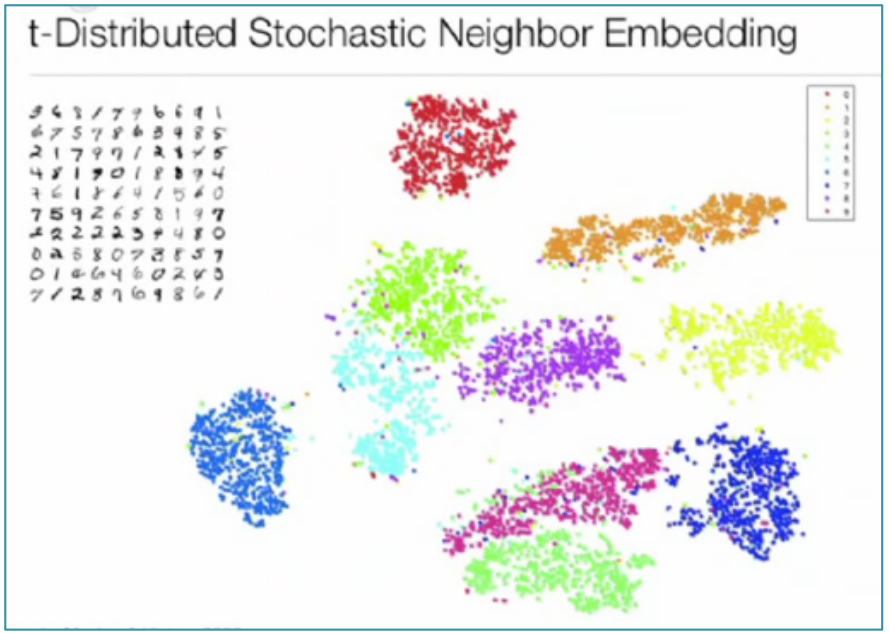
Genes mirror geography within Europe

Novembre et al (2008) Nature. doi: 10.1038/nature07331

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

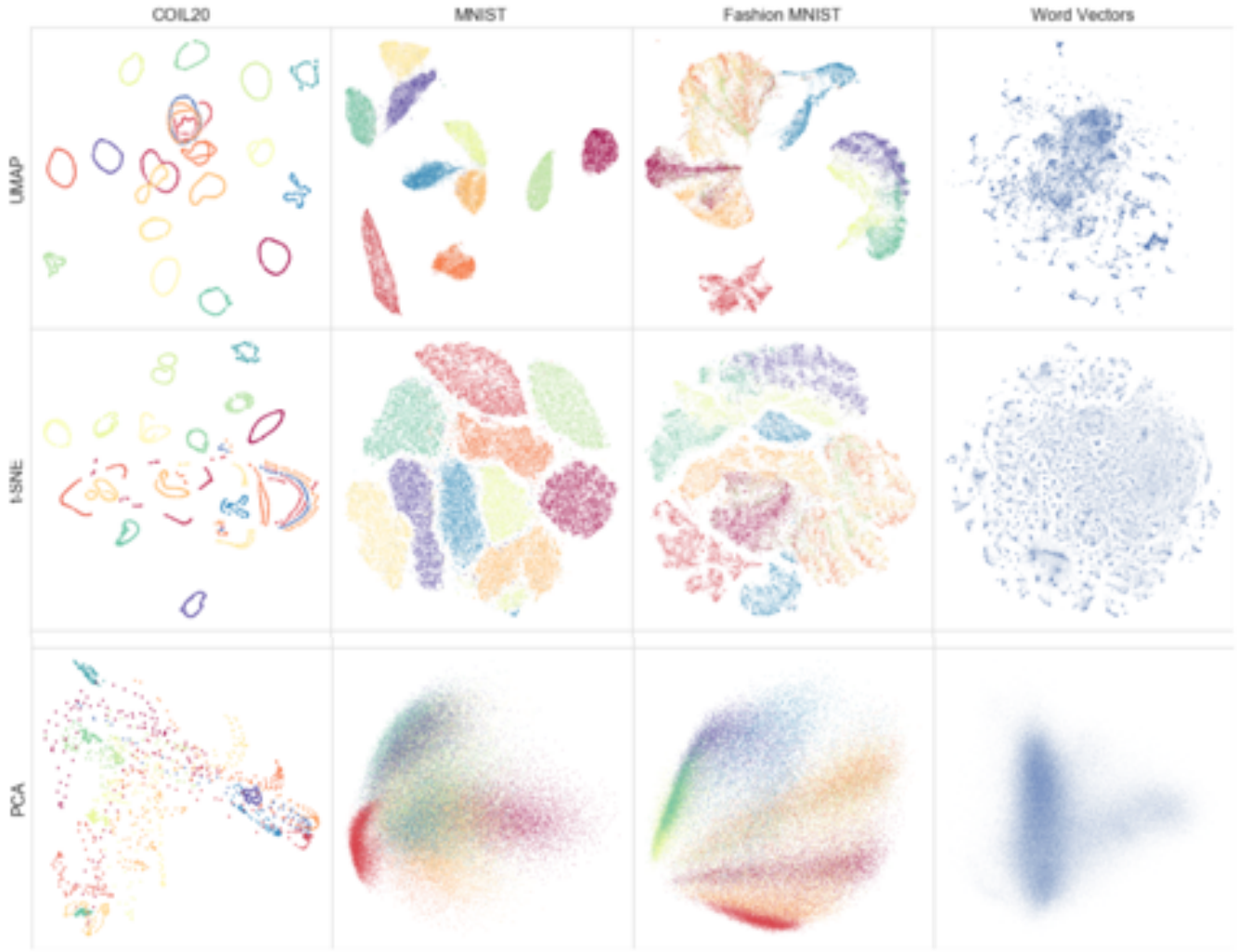
Visualizing Data Using t-SNE

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

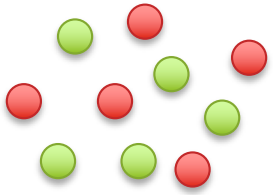
1. Why single cells?
2. scRNA and other assays

3. scDNA as Bonus Slides

Population Heterogeneity

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

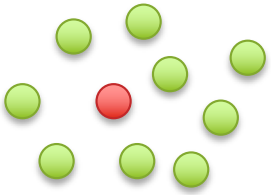
Experiment 1: 50/50



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

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

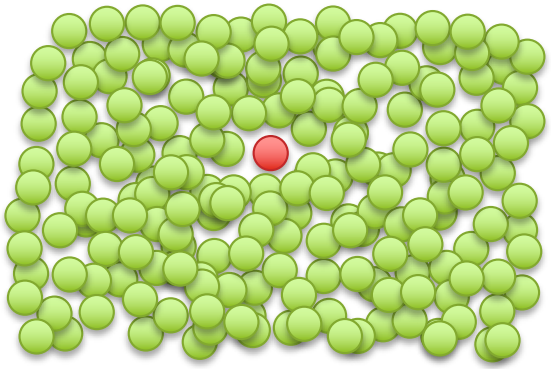
Experiment 2: 1/10



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

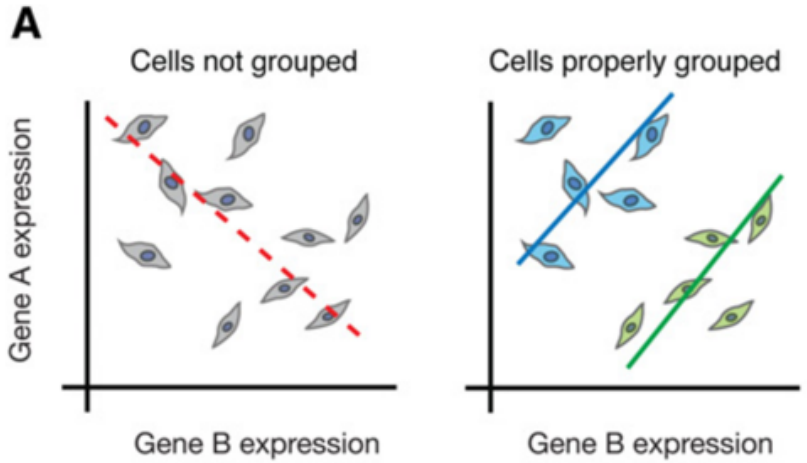
	Drug A	Drug B
Overall Response	78% (273/350)	**83% (289/350)**
Male Response	**93% (81/87)**	87% (234/270)
Female Response	**73% (192/263)**	69% (55/80)

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



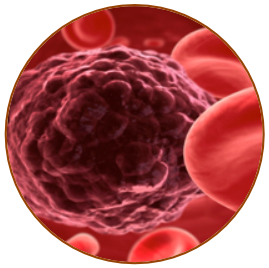
What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

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)

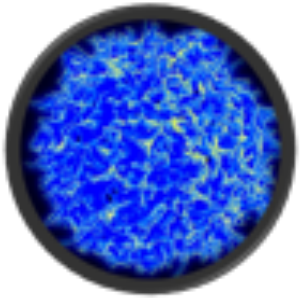
(Trapnell, 2015, Genome Research)

Sources of (Genomic) Heterogeneity

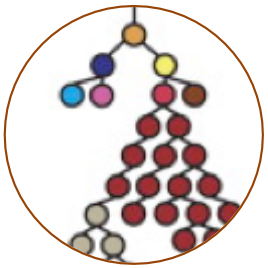


Circulating tumor cells

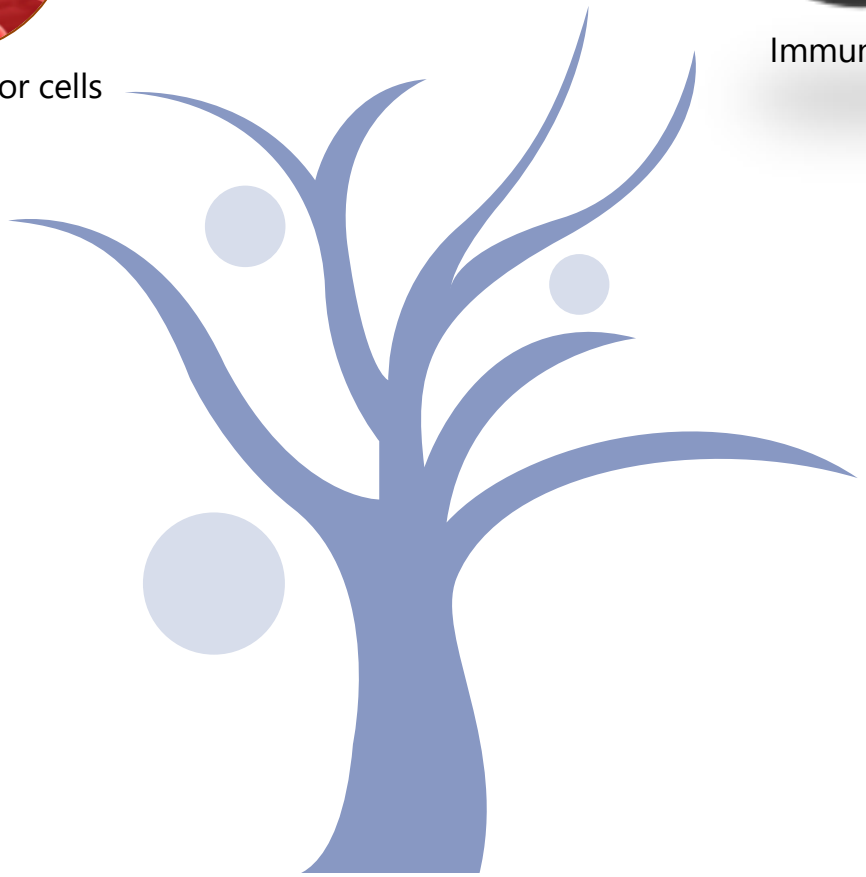
Neuronal mosaicism

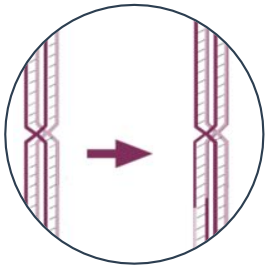


Immune Cells



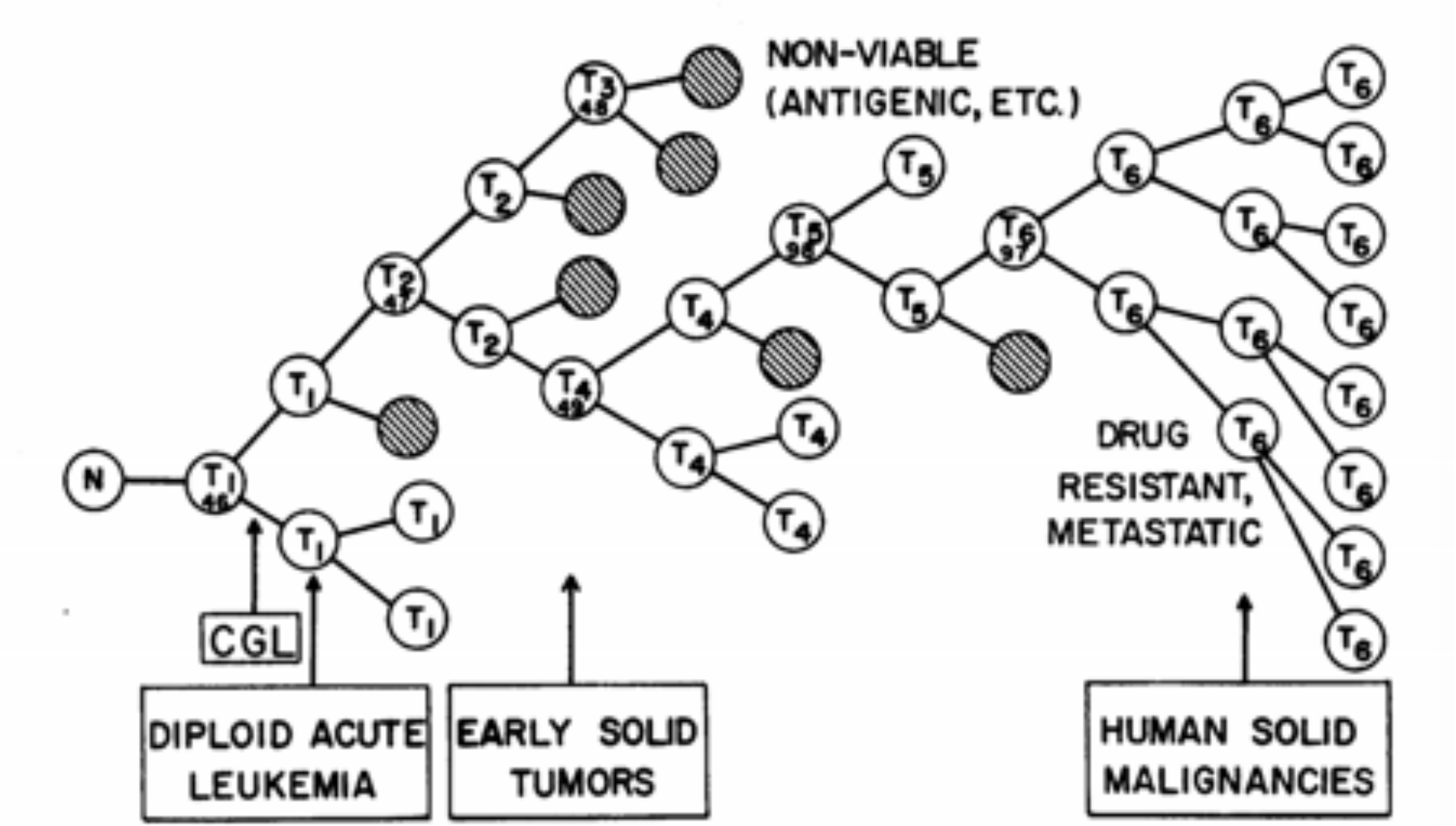
Tumor evolution





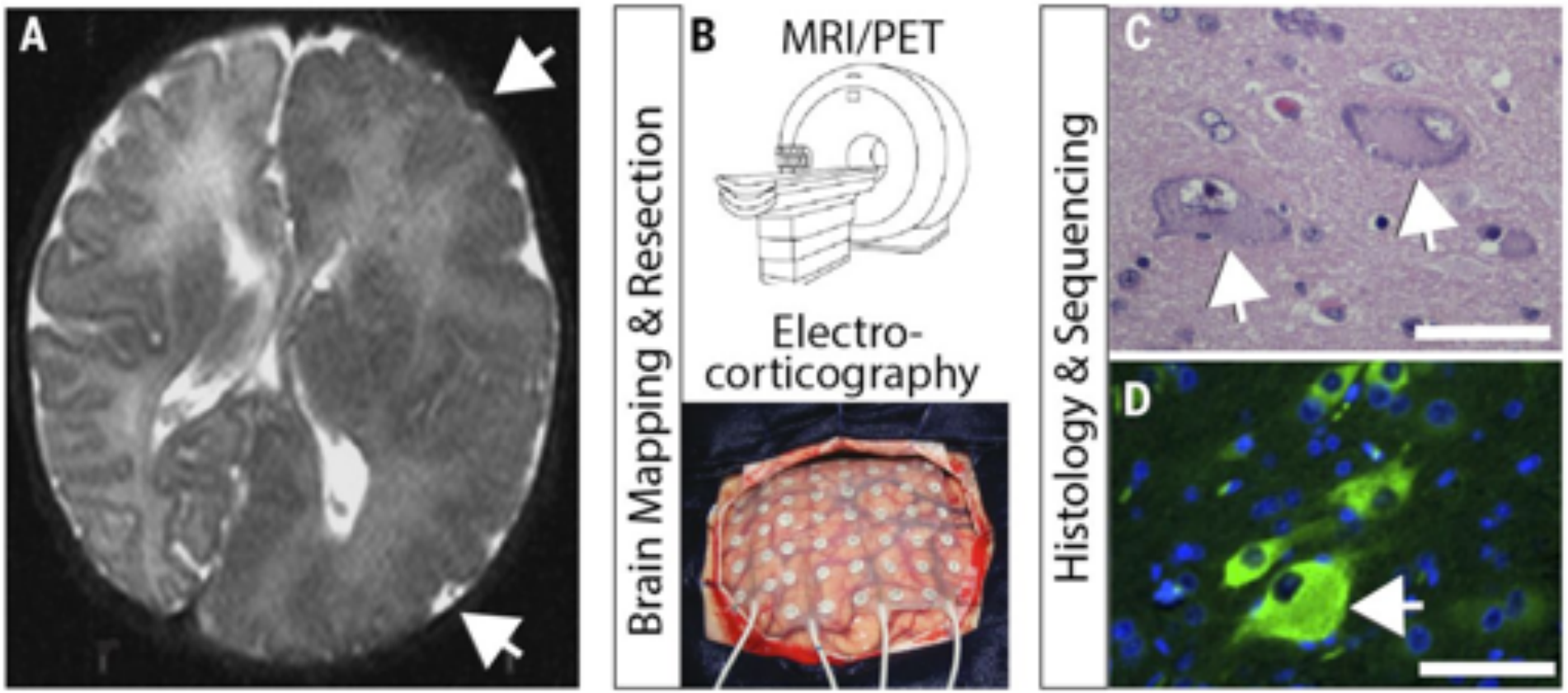
Recombination/crossover in germ cells

Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

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

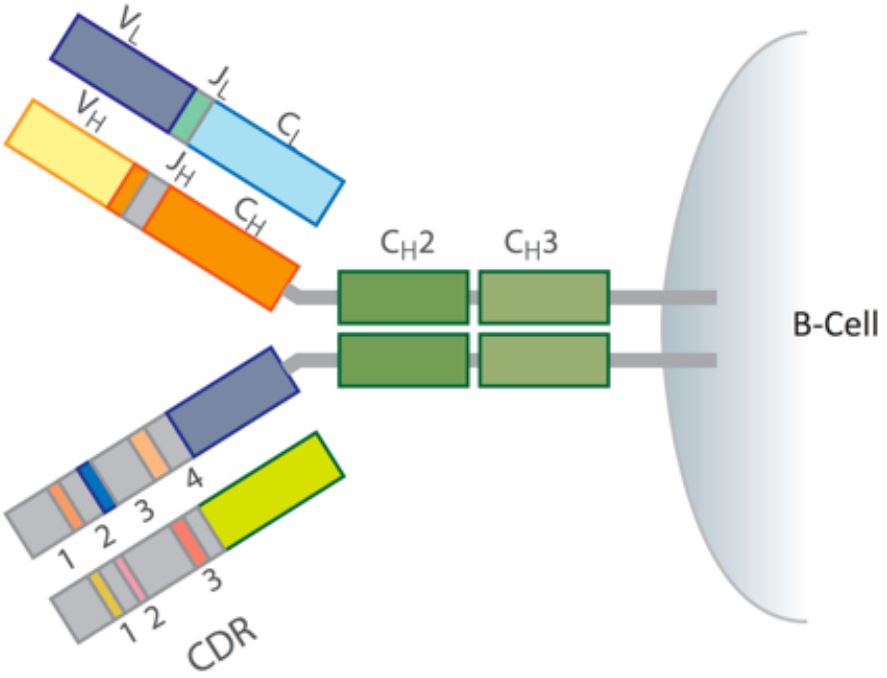


An example of brain somatic mosaicism that leads to a focal overgrowth condition.

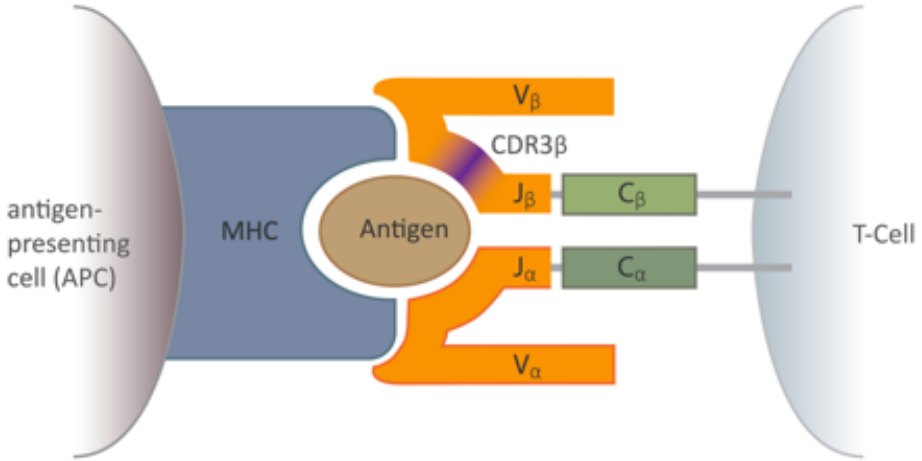
(**A**) Axial brain MRI of focal overgrowth from a 2-month-old child with intractable epilepsy and intellectual disability. (**B**) Brain mapping using high-resolution MRI is followed by surgical resection of diseased brain tissue. (**C**) Histological analysis with hematoxylin/eosin showing characteristic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



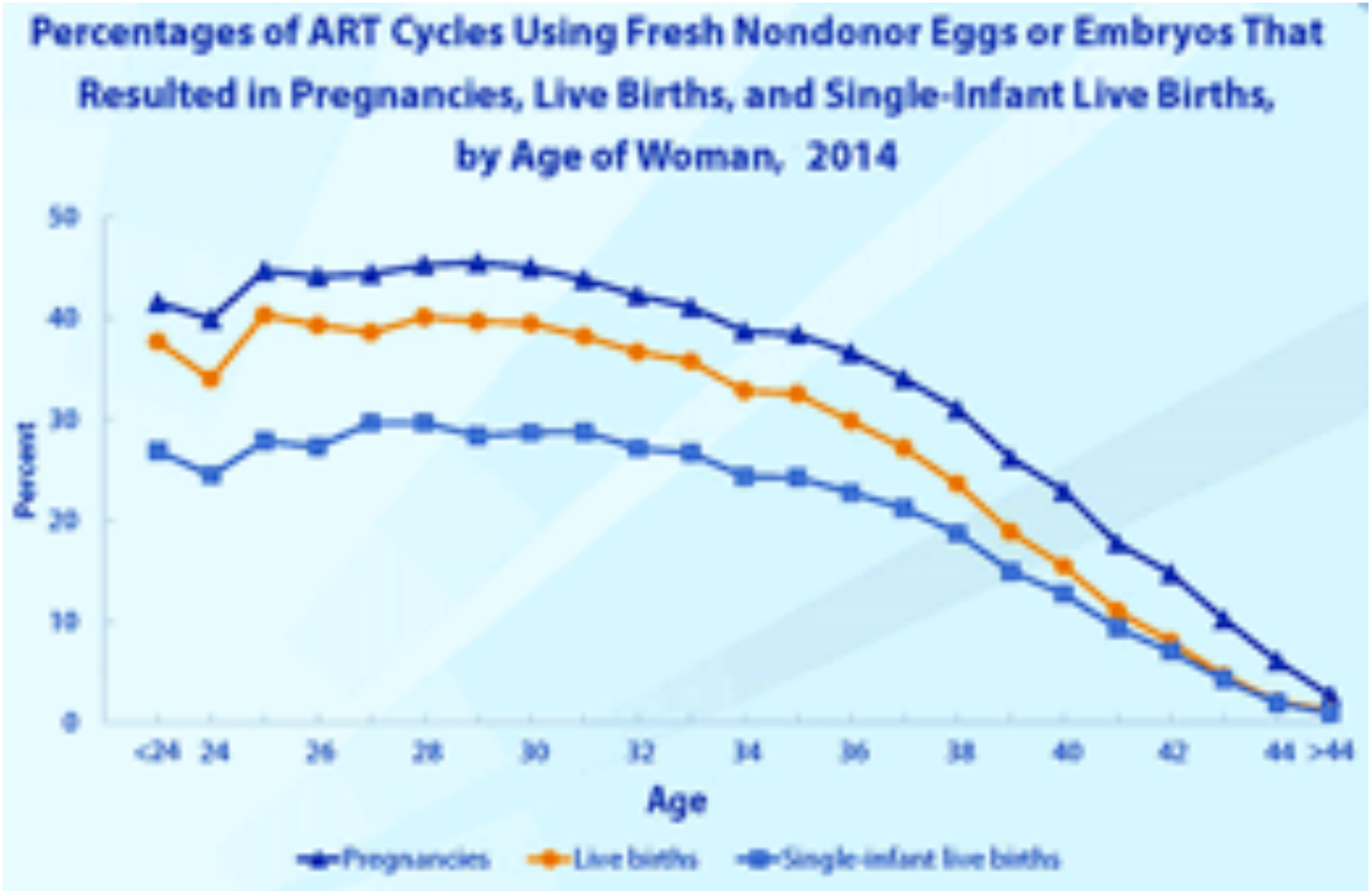
- **Massive diversity rivaled only by germ cells**
- **Somatic recombination**

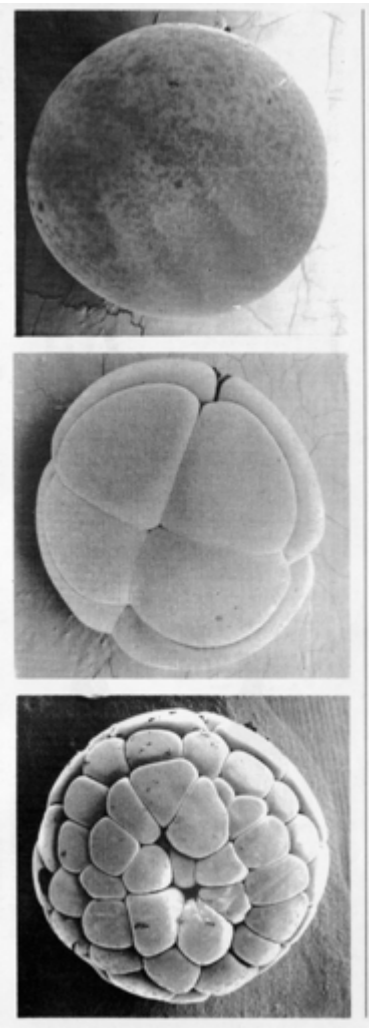


- **B cells – antibody generation**
- **T cells – antigen response**

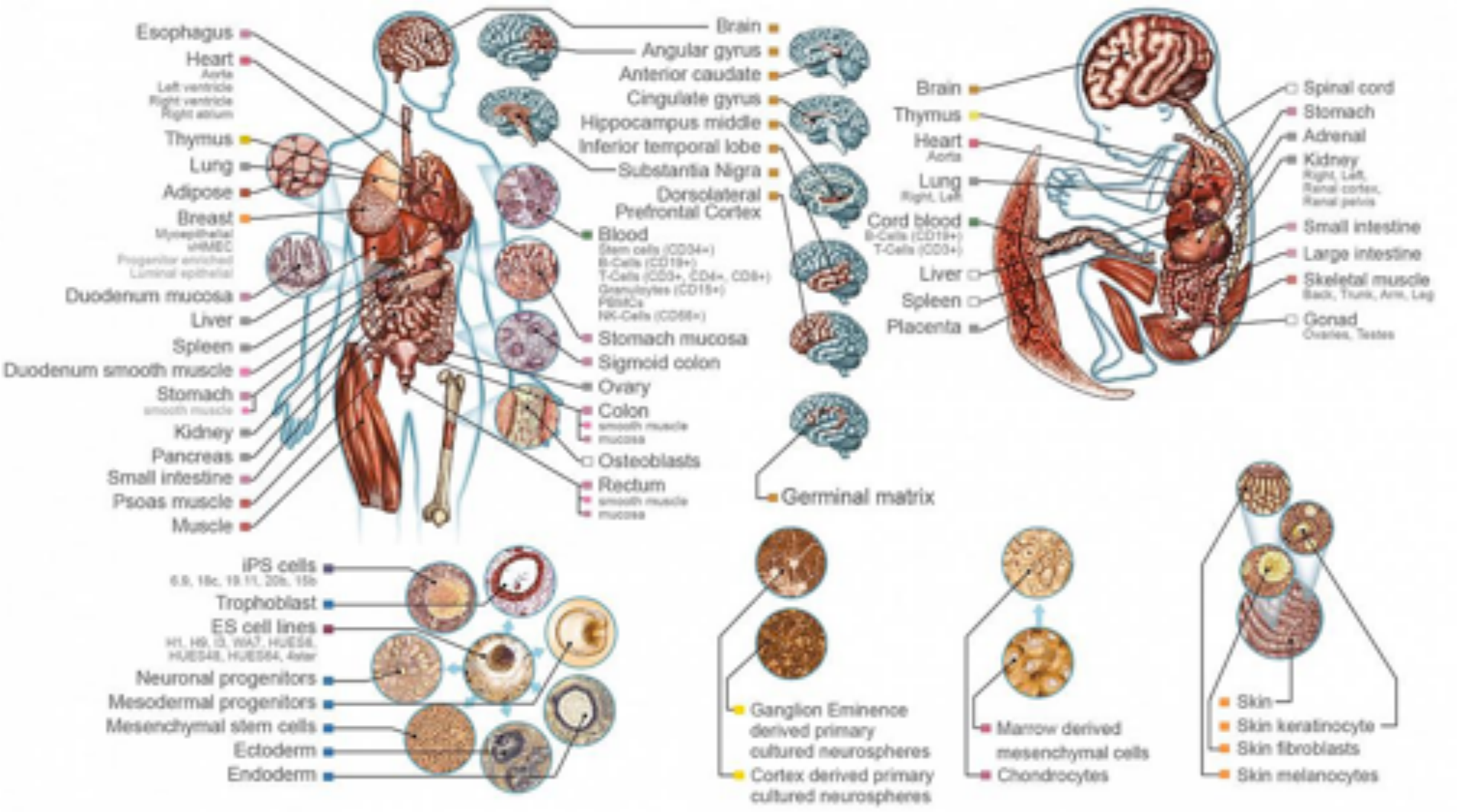
Single cell research. Illumina.

In-vitro Fertilization

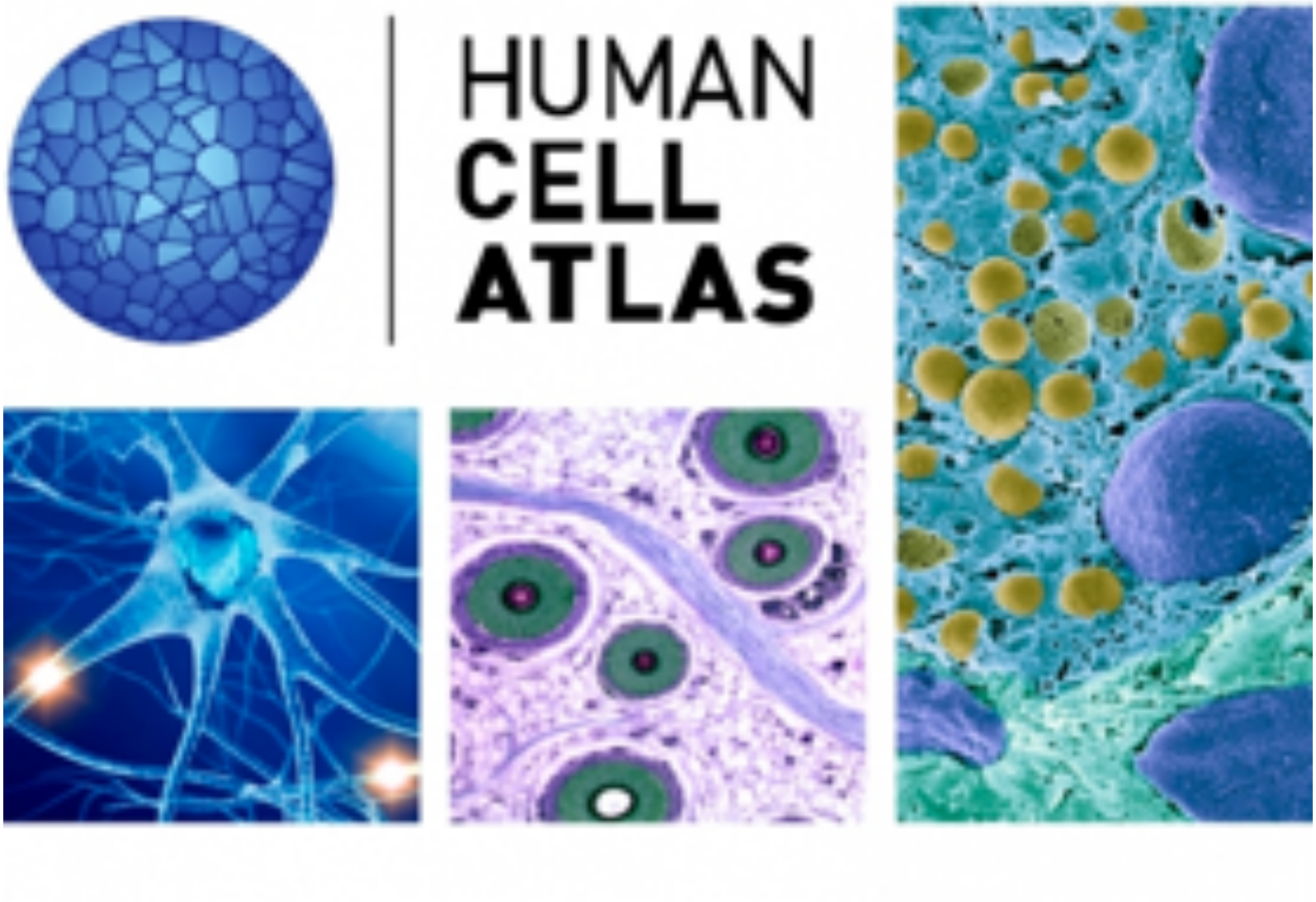




Sources of (Cellular) Heterogeneity



Roadmap Epigenomics Consortium



https://www.humancellatlas.org/

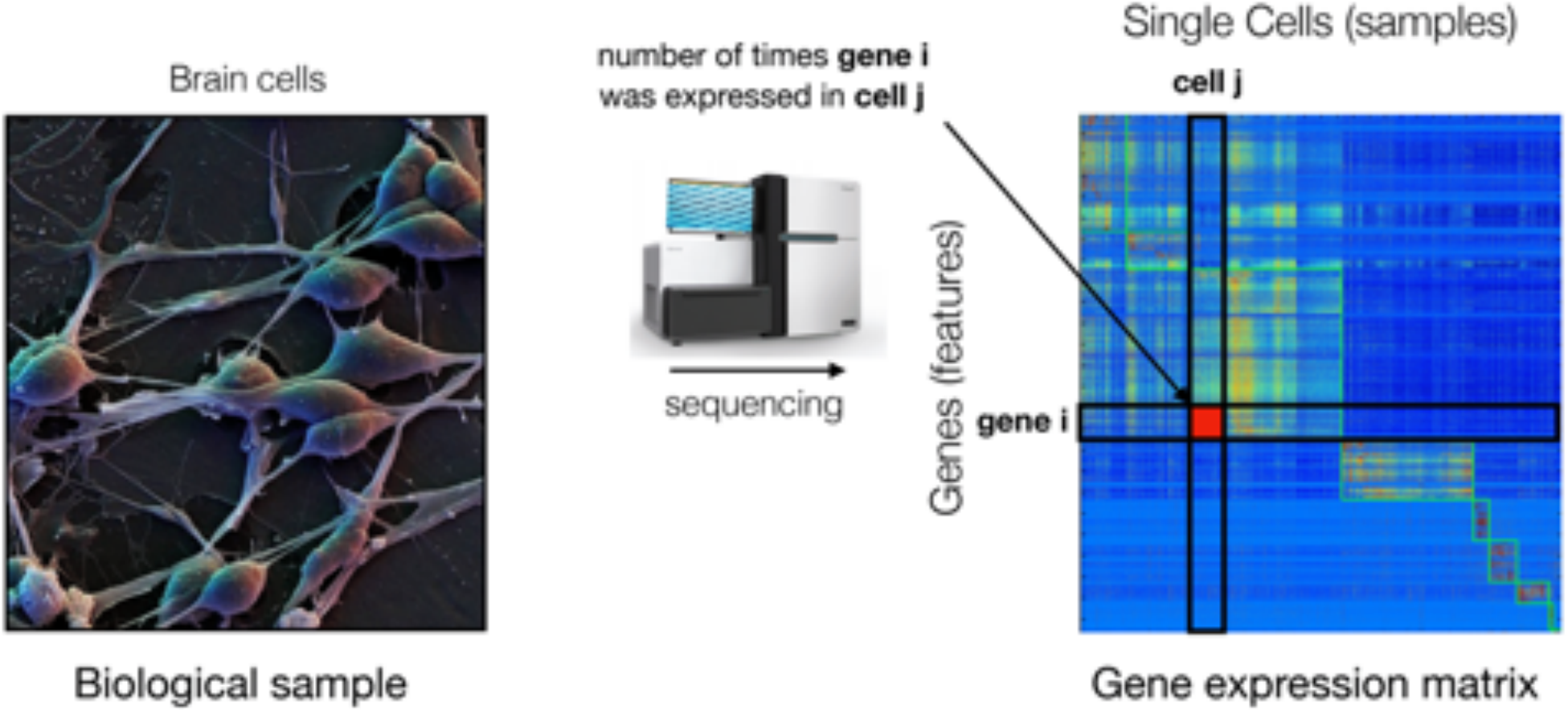
Single Cell Analysis

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

3. scDNA as Bonus Slides

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

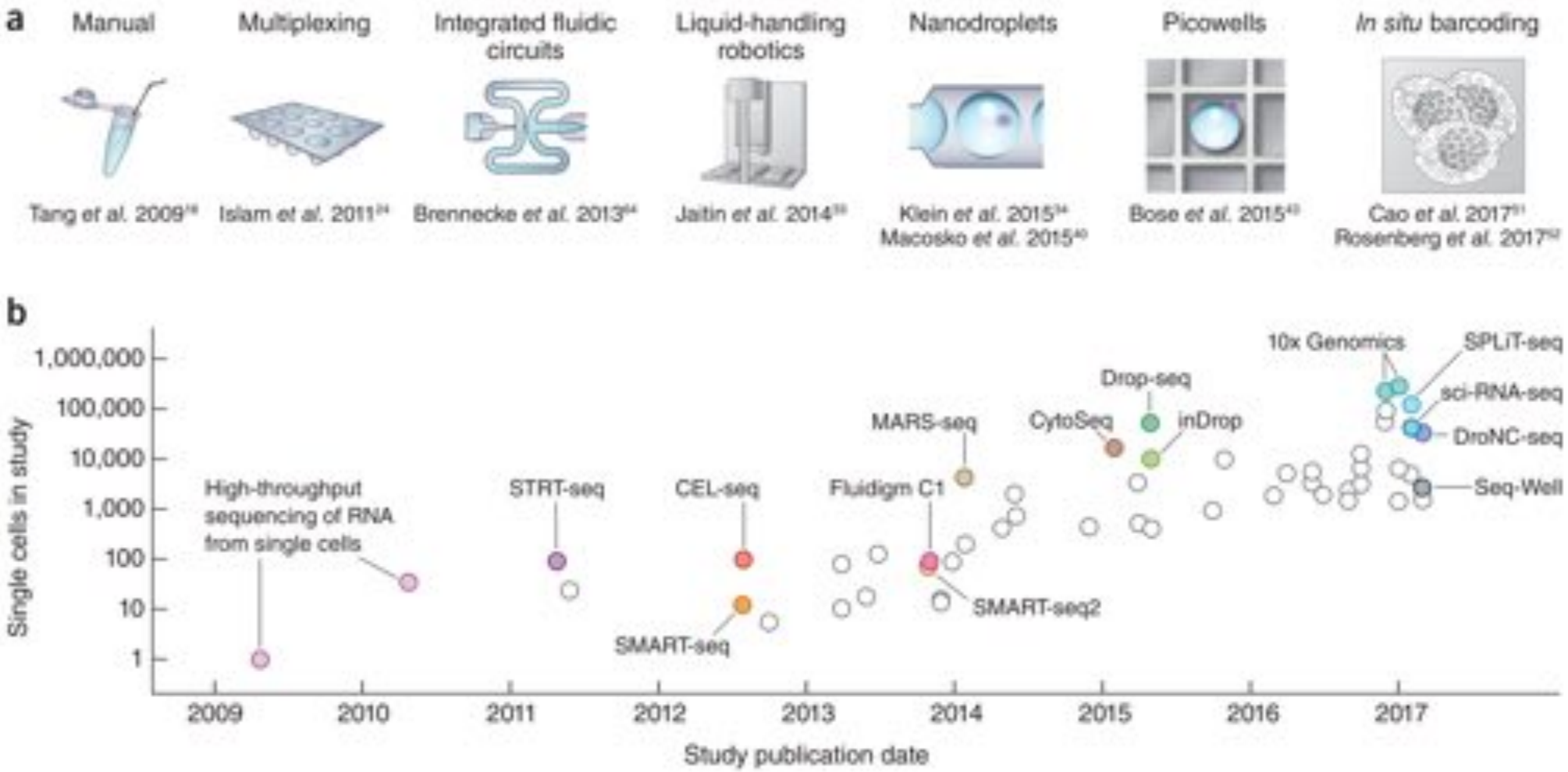
— a snapshot of the underlying biology in a data matrix.



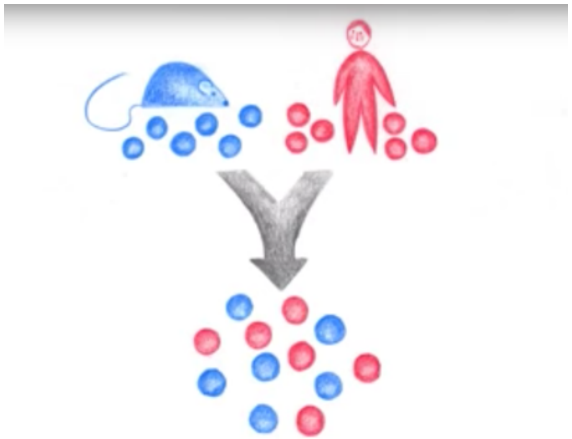
computationally explore complex biological systems

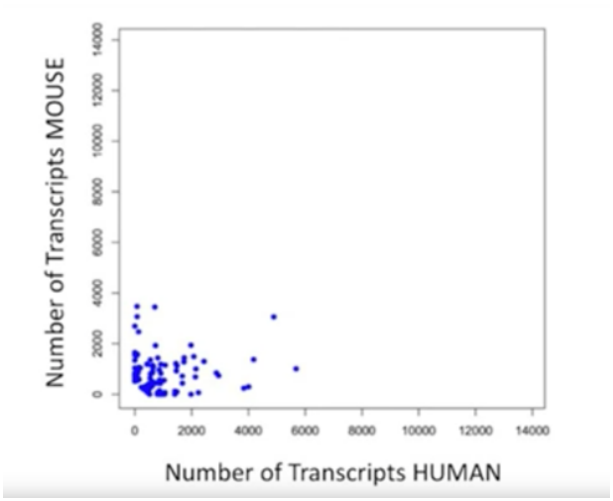
Martin Zhang

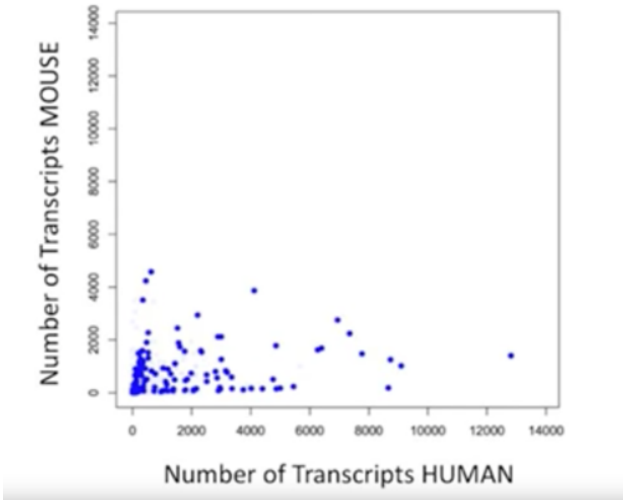
A decade of single-cell RNA-seq

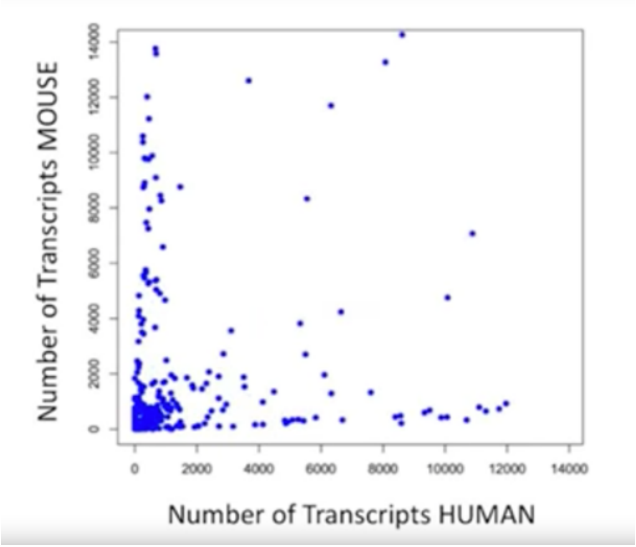


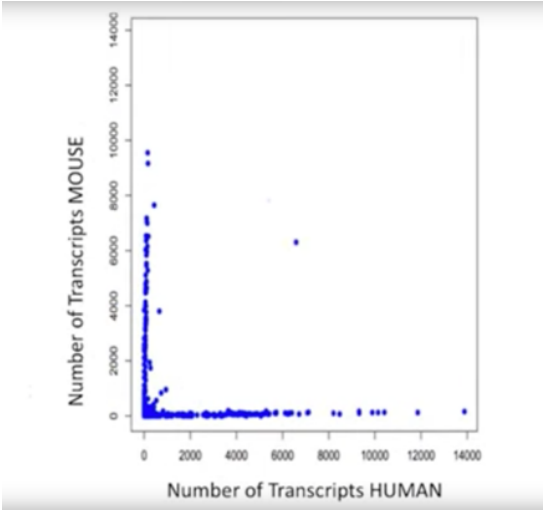
Svensson et al. 2018. DOI:10.1038/nprot.2017.149

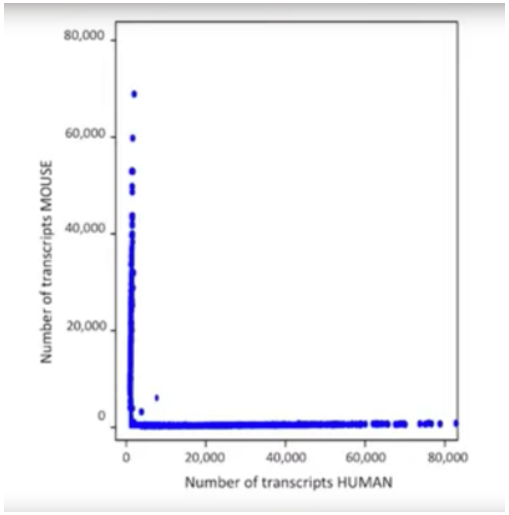








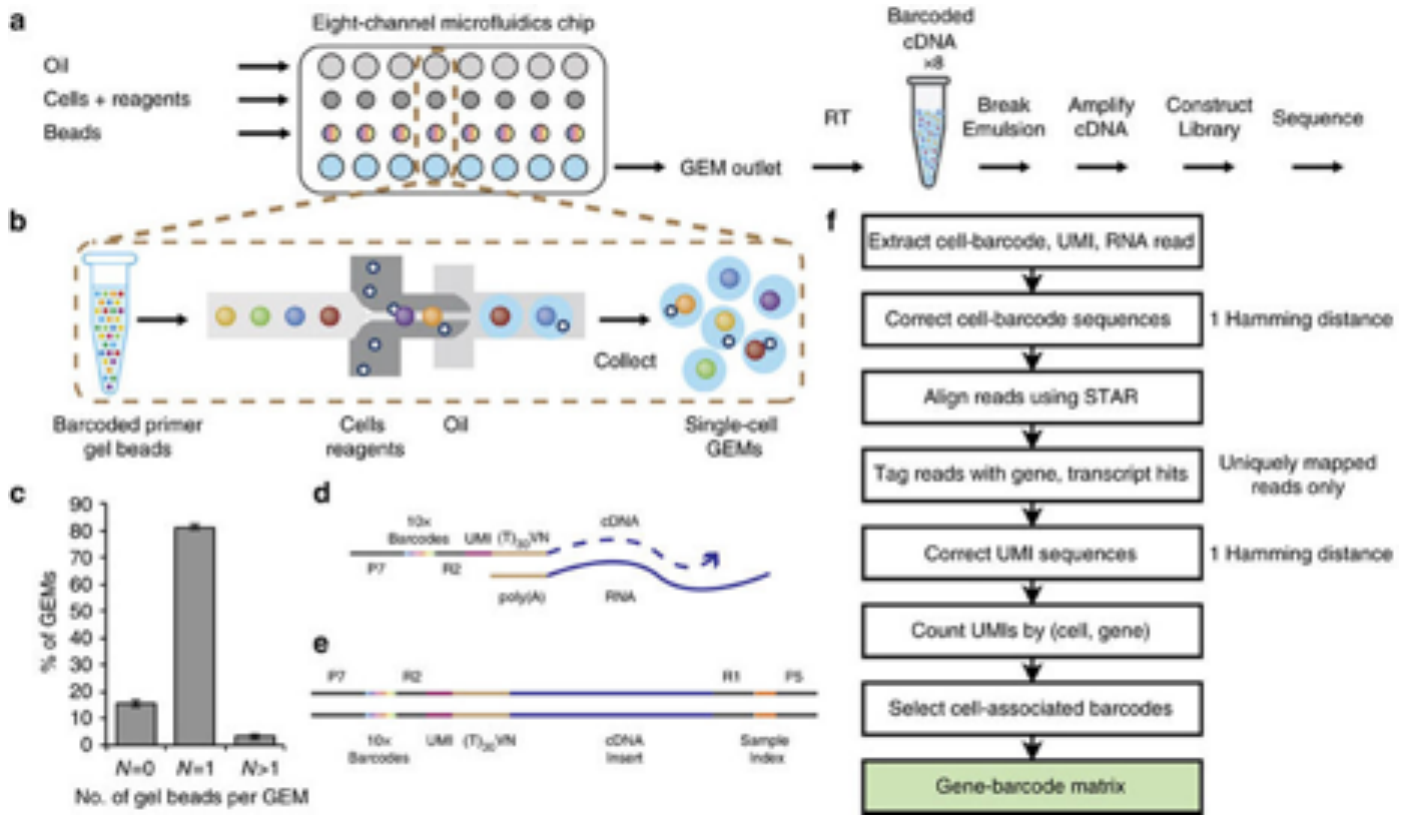




Drop-seq: Droplet barcoding of single cells

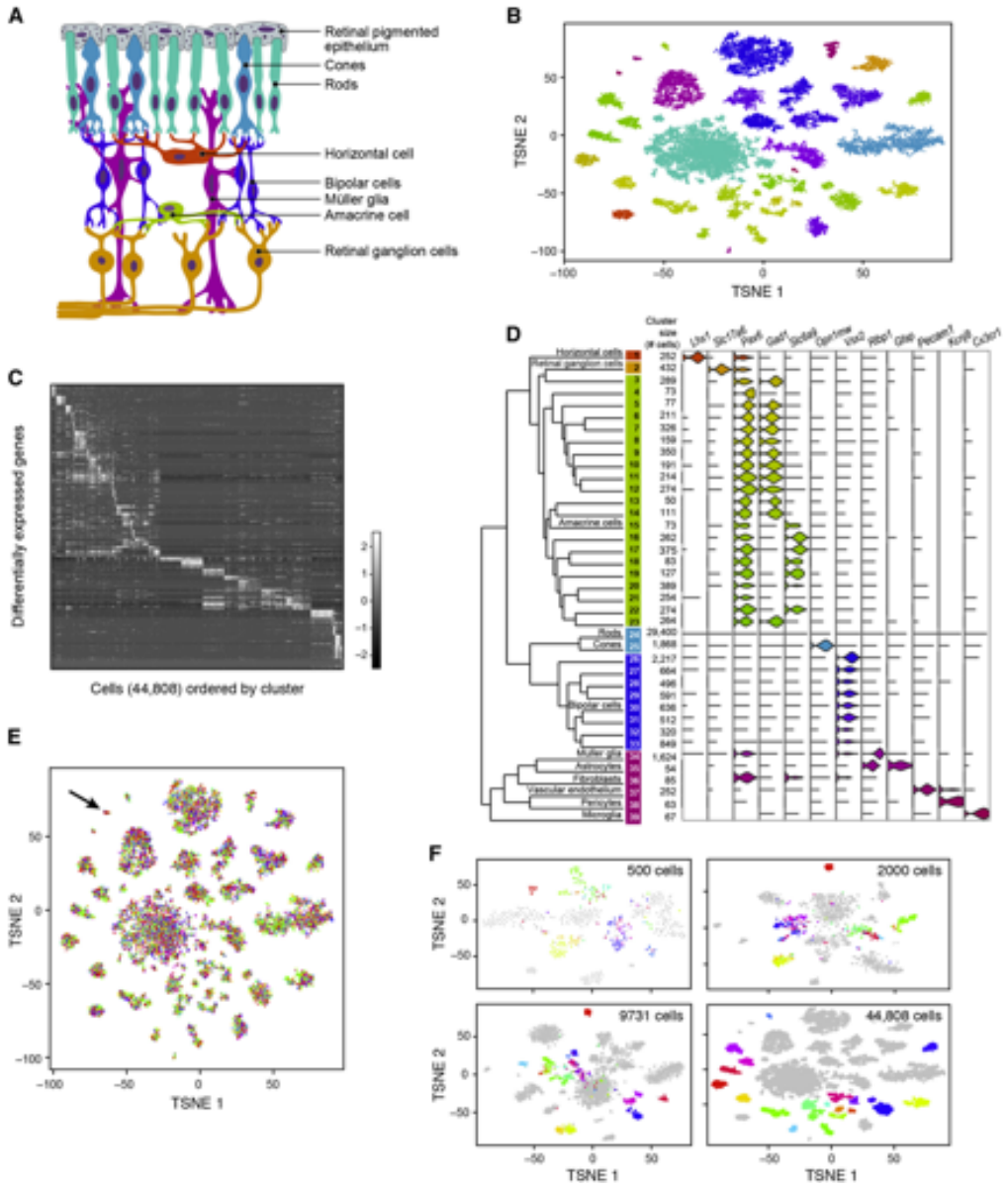
https://www.youtube.com/watch?v=vL7ptq2Dcf0

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

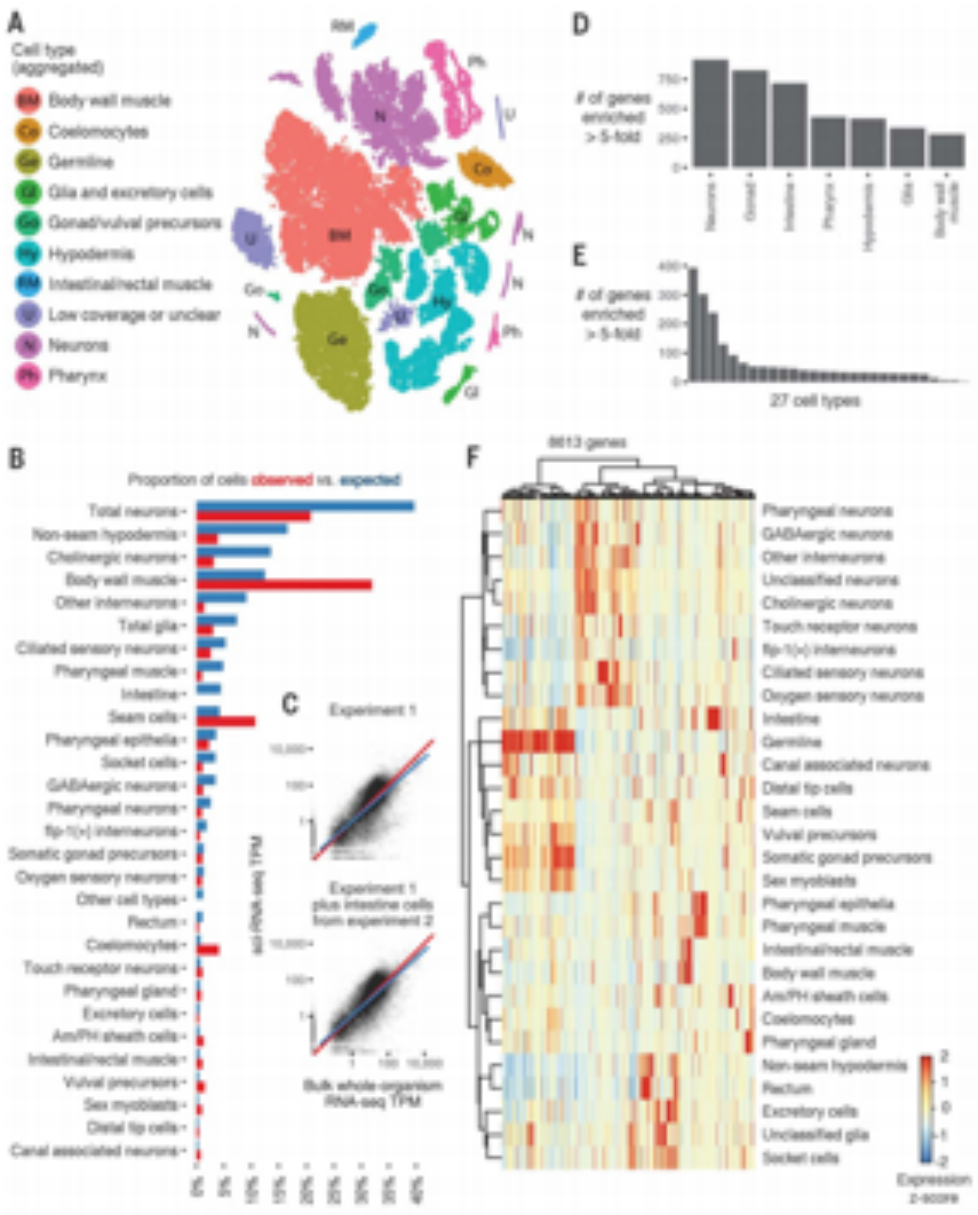
(c) Differentially expressed genes in each cluster

(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

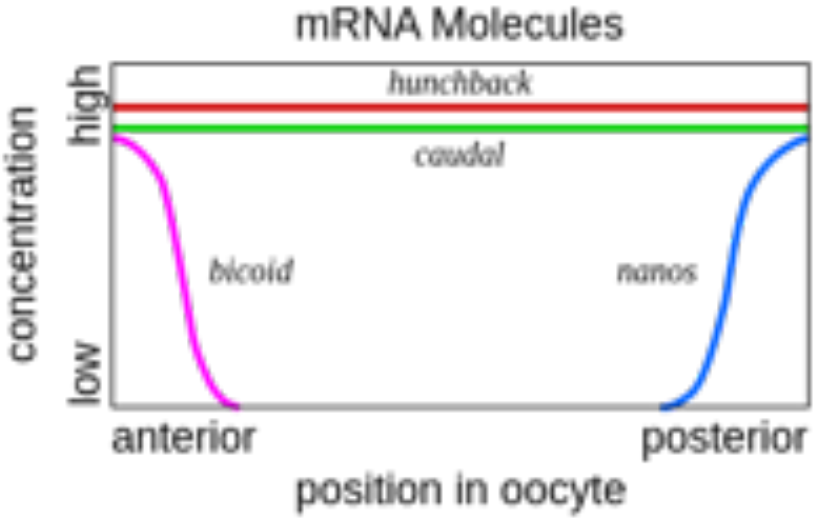
(a) t-SNE visualization of clusters

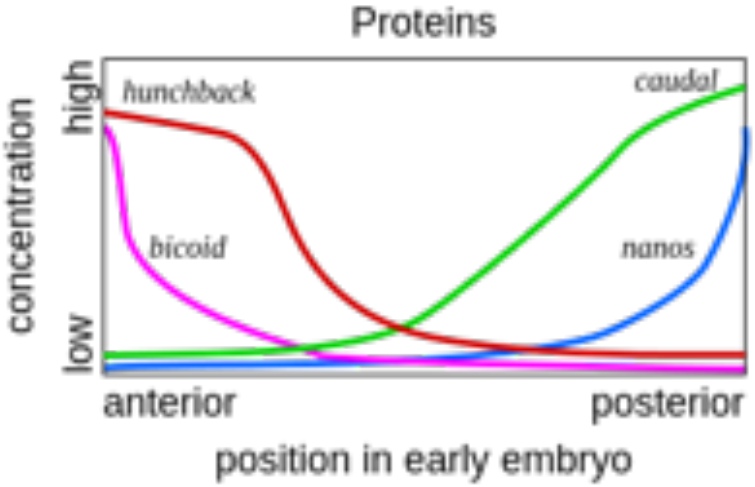
(b) Proportion of cells observed vs expected match well (including cells that only occur once or twice in the animal)

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

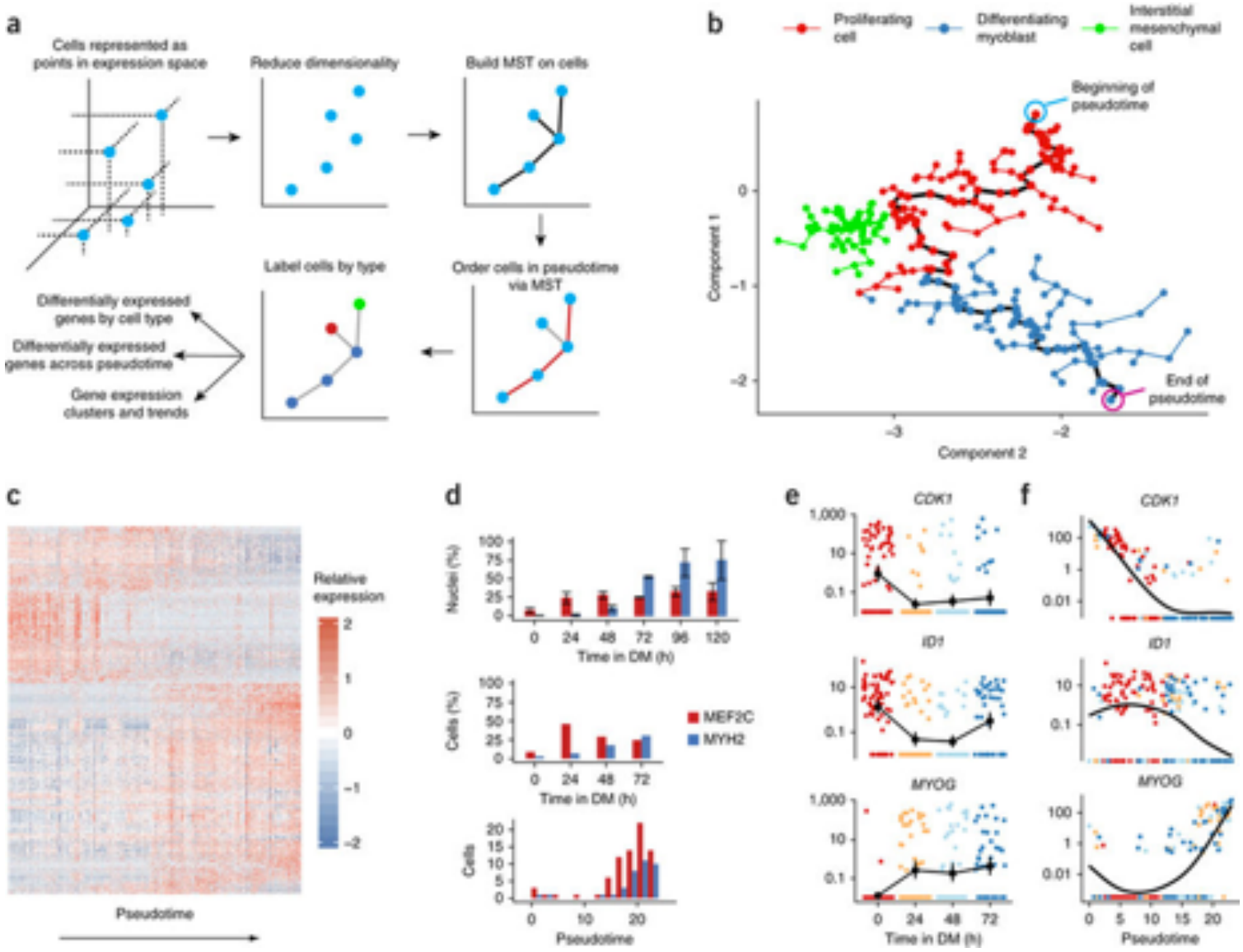
(d-f) Analysis of key genes per cell type

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



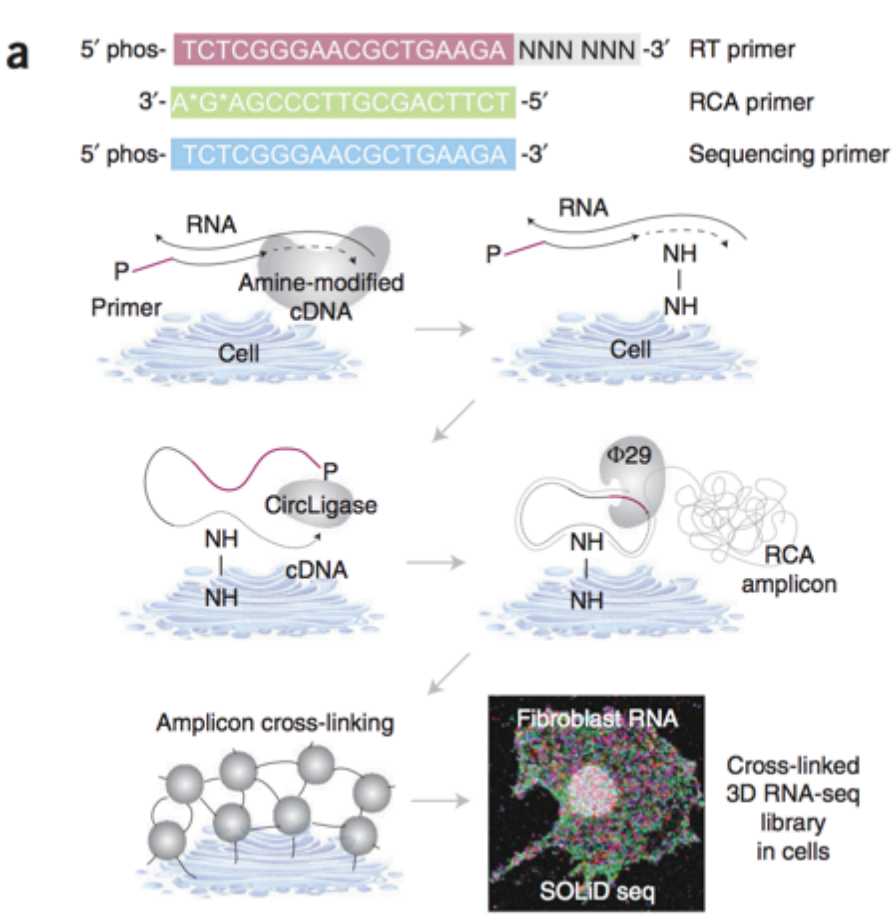


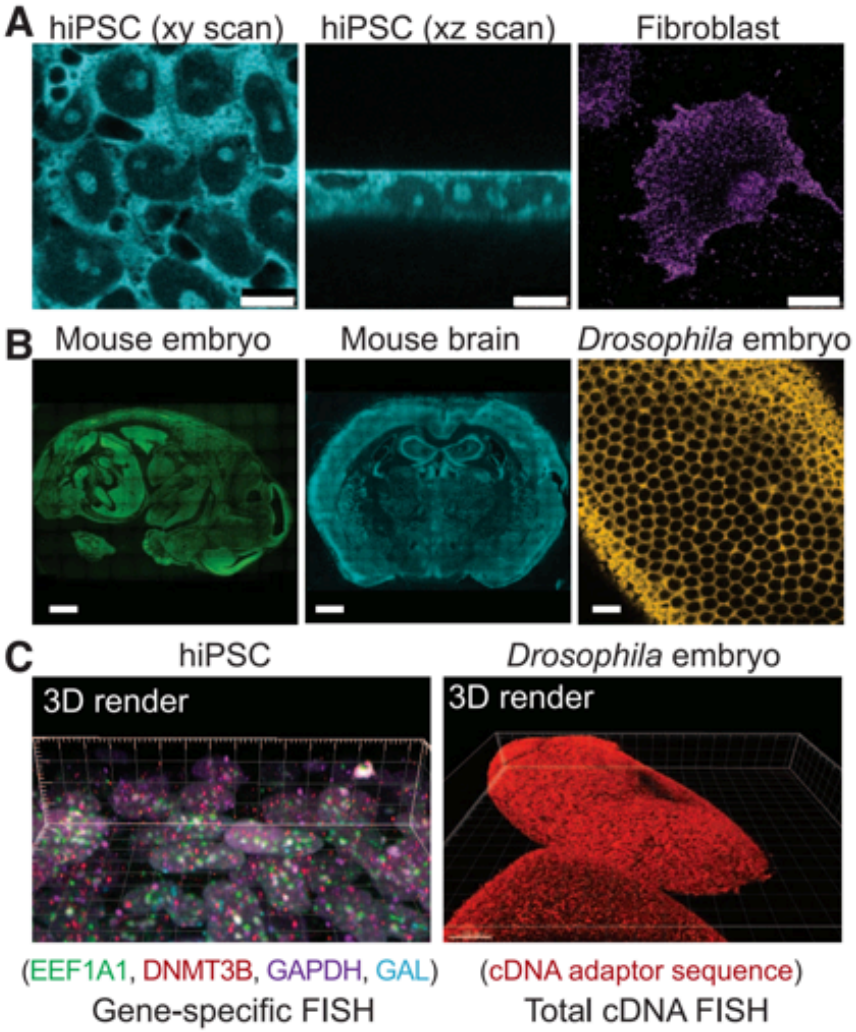
https://en.wikipedia.org/wiki/Drosophila_embryogenesis



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





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

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

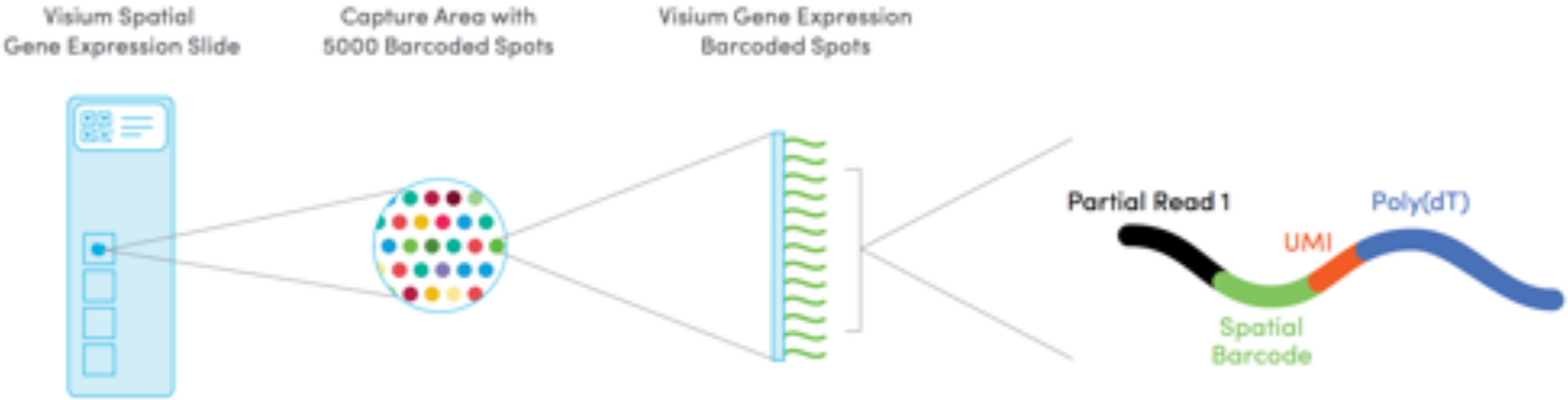


Figure 1. Here we show the composition of the Visium Spatial Gene Expression slide. Each slide contains four Capture Areas with approximately 5000 barcoded spots, which in turn contain millions of spatially-barcoded capture oligonucleotides. Tissue mRNA is released and binds to the barcoded oligos, enabling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

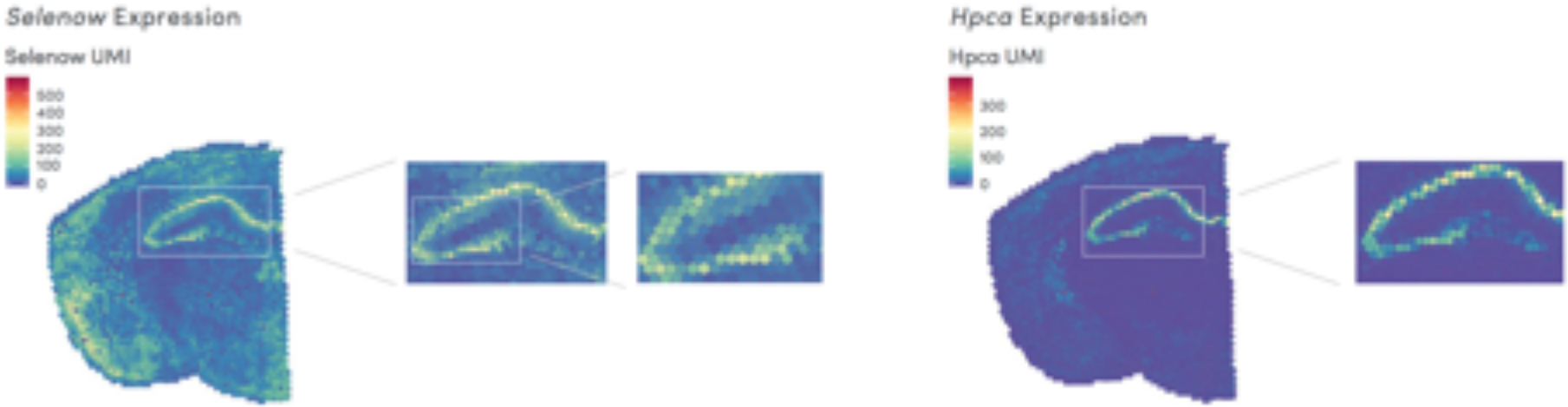


Figure 2. This is a coronal mouse brain section with overlaid spatial gene expression information. The spots correspond to localized mRNA of *Selenow* and *Hpca*, both known to have predominant hippocampal expression.

10xgenomics.com/spatial-gene-expression

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

Single Cell Analysis

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

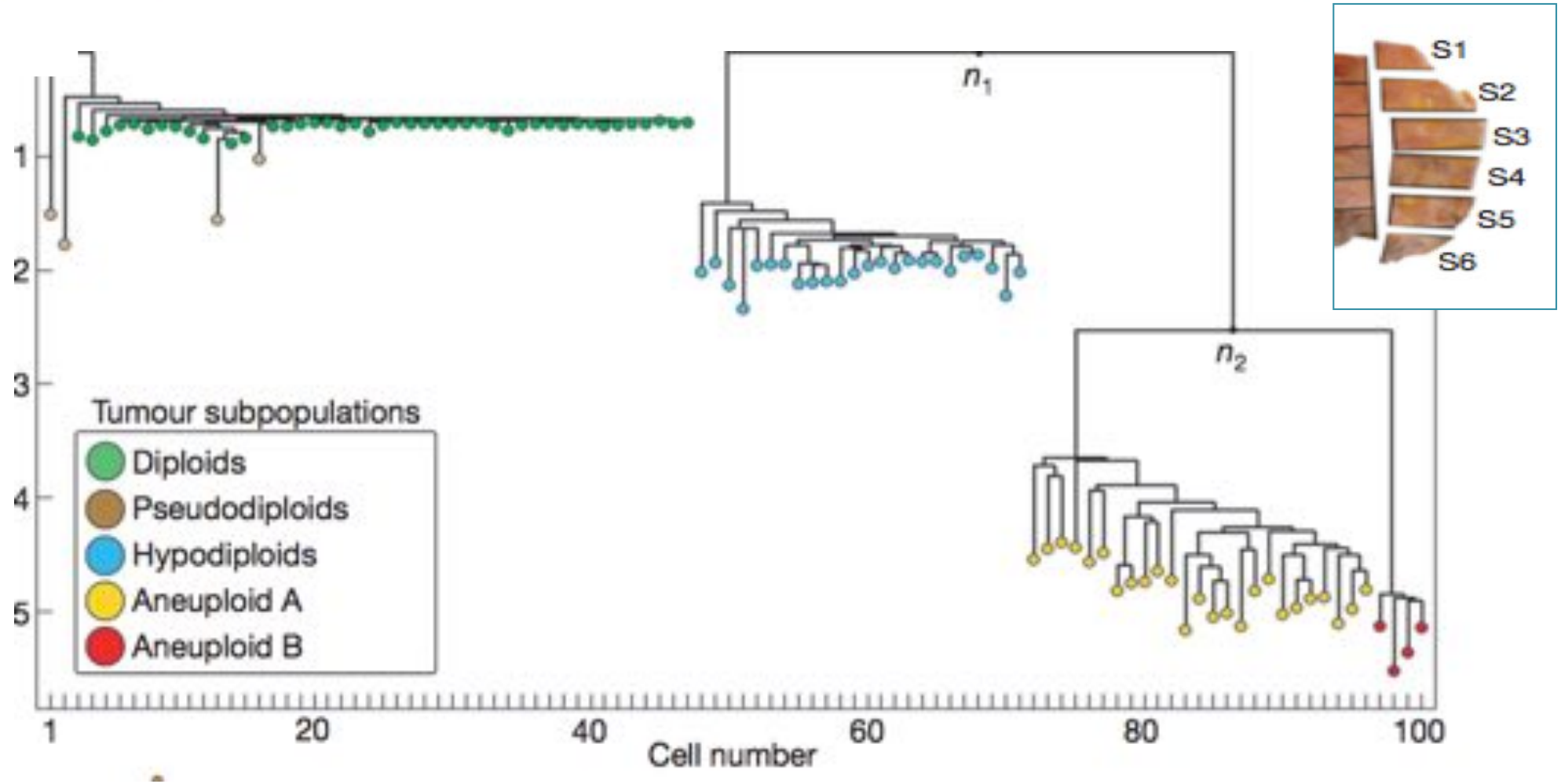
3. scDNA as Bonus Slides

LETTER

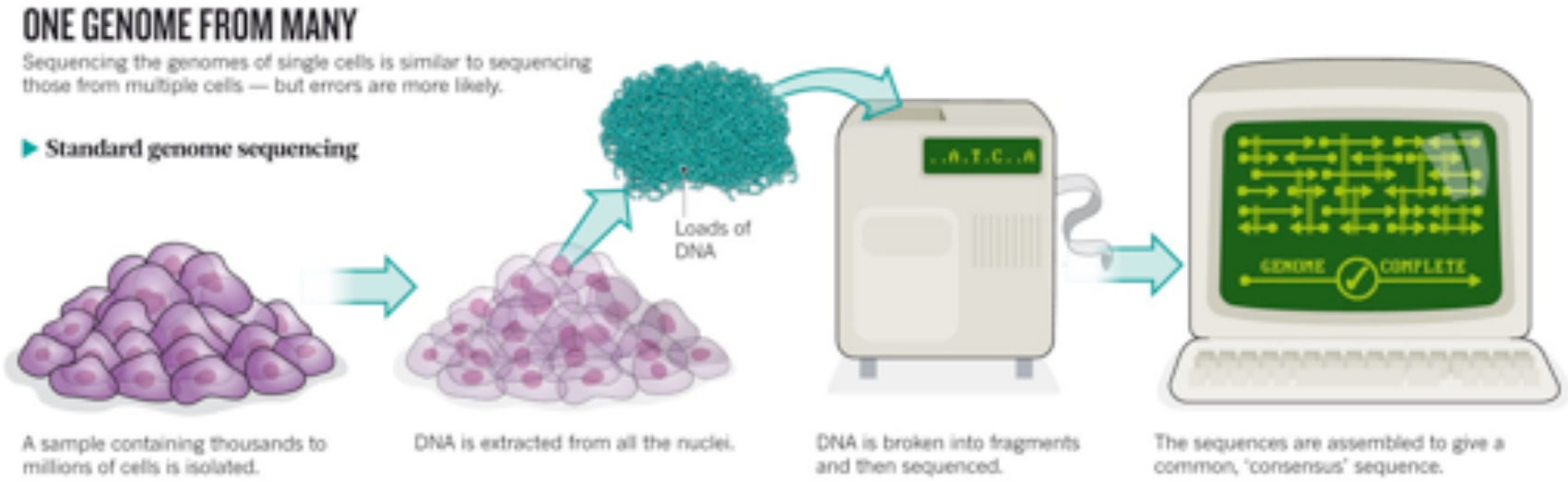
doi:10.1038/nature09807

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]

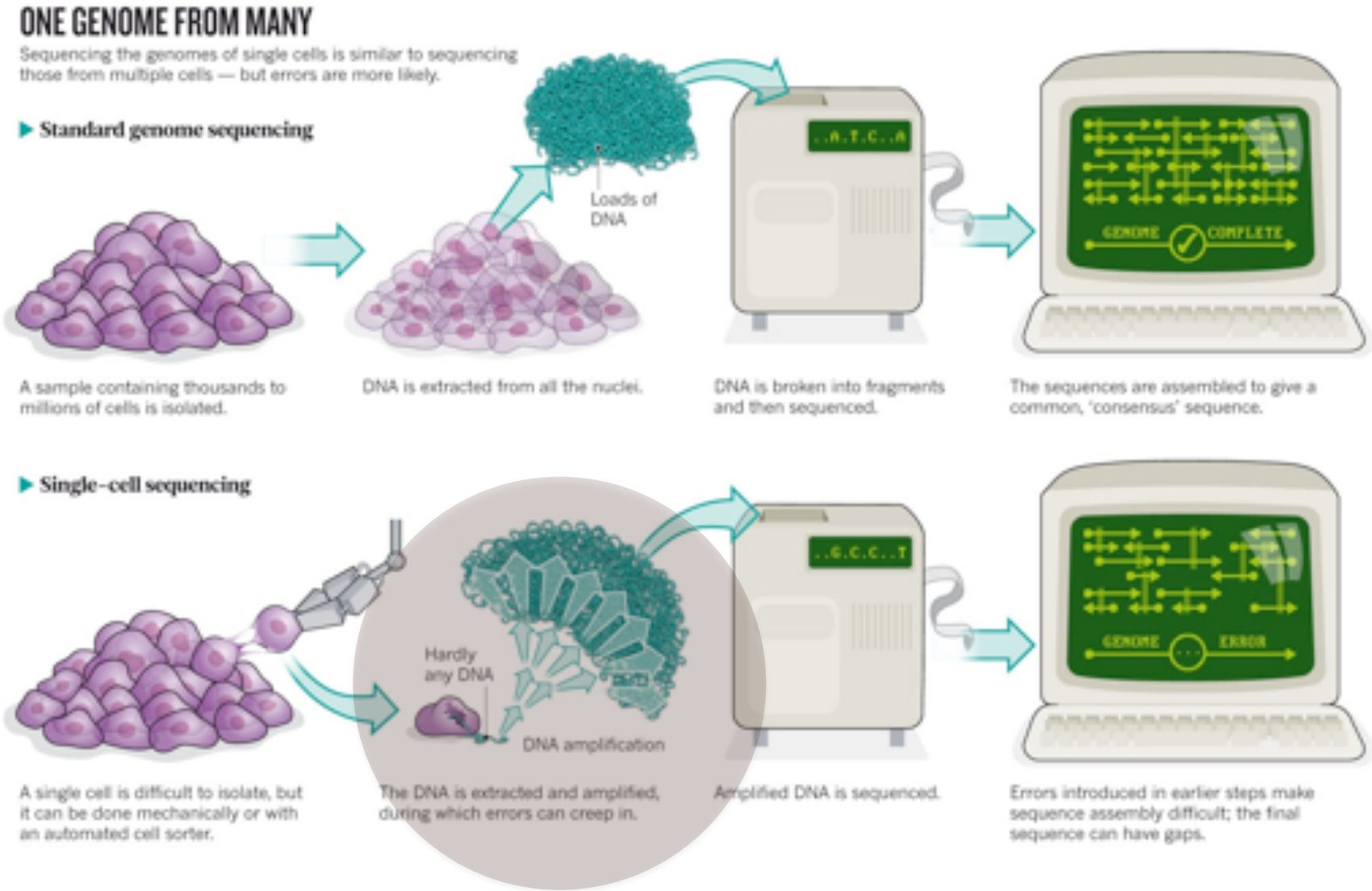


Single-cell vs. bulk sequencing



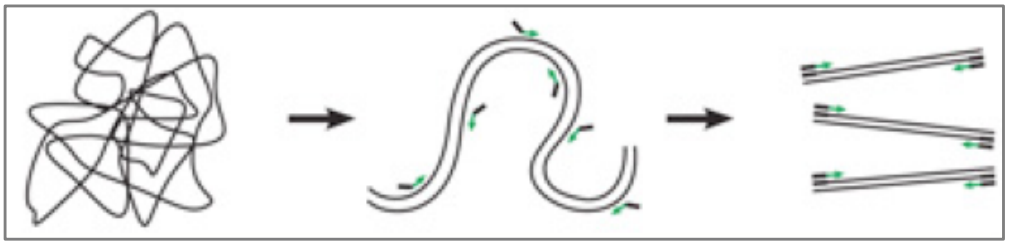
Brian Owens, Nature News 2012

Single-cell vs. bulk sequencing

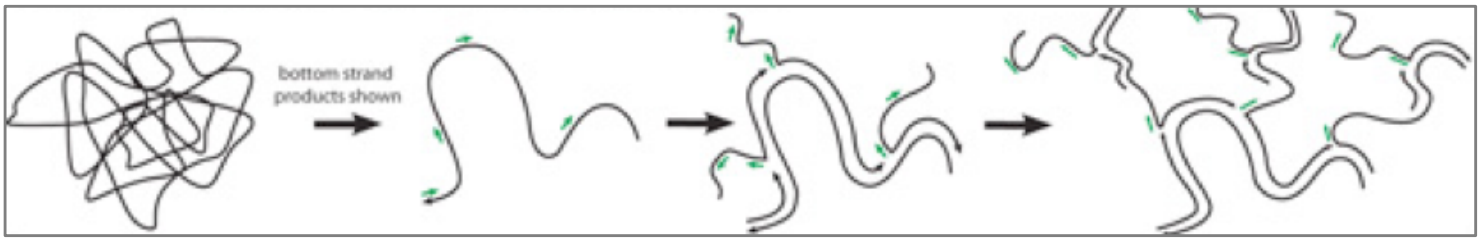


Brian Owens, Nature News 2012

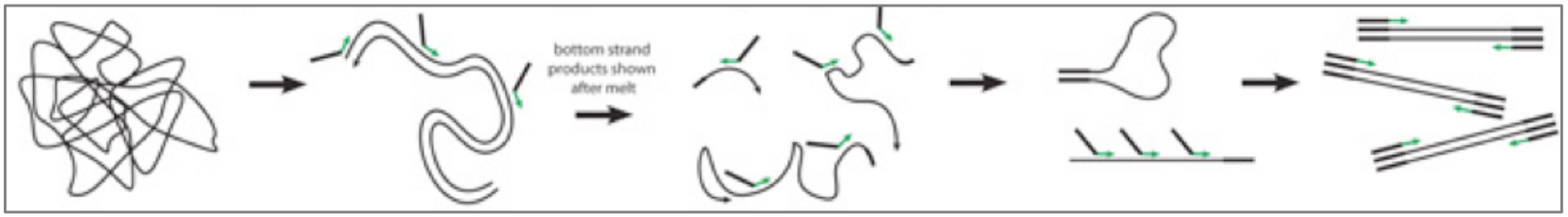
Whole Genome Amplification Techniques



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

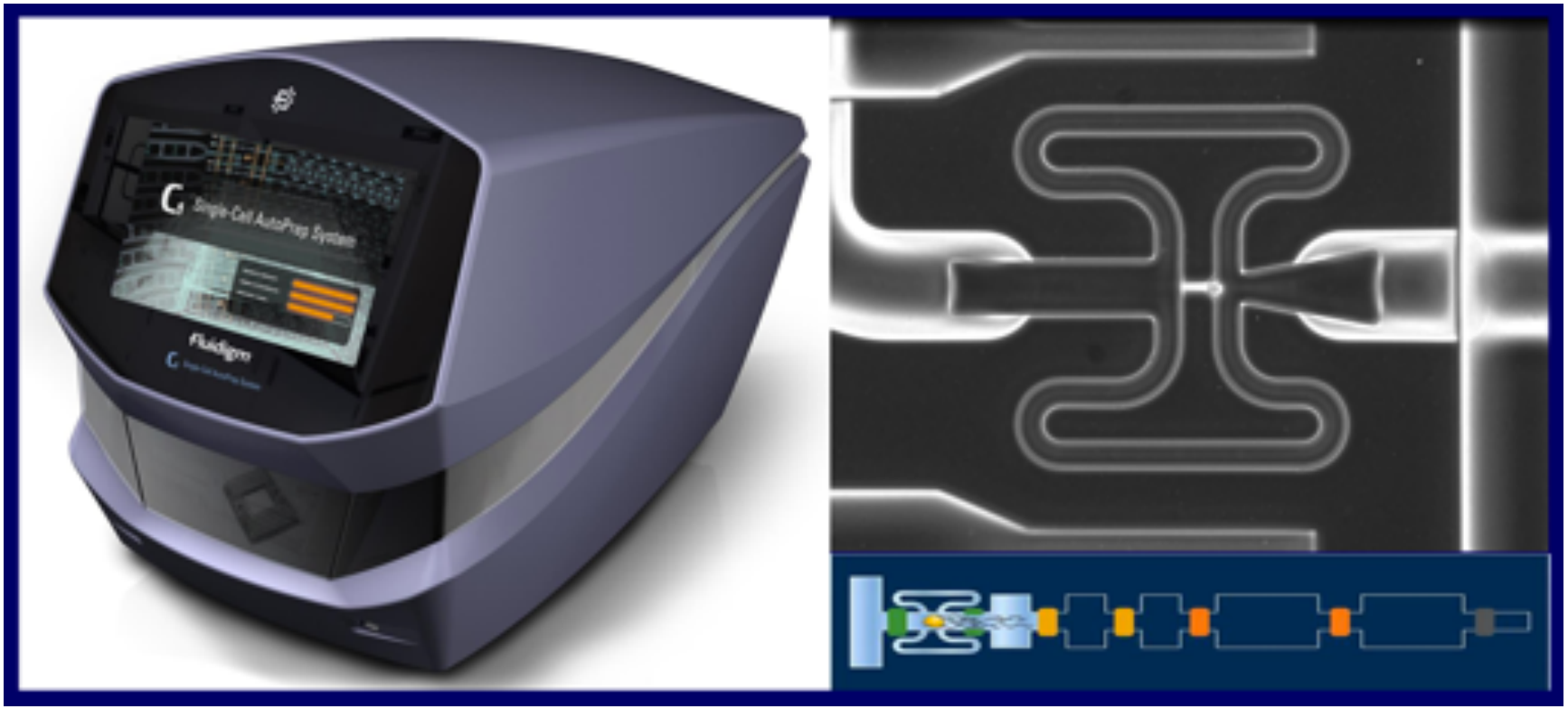


MDA: Multiple Displacement Amplification
Dean et al. (2002) PNAS



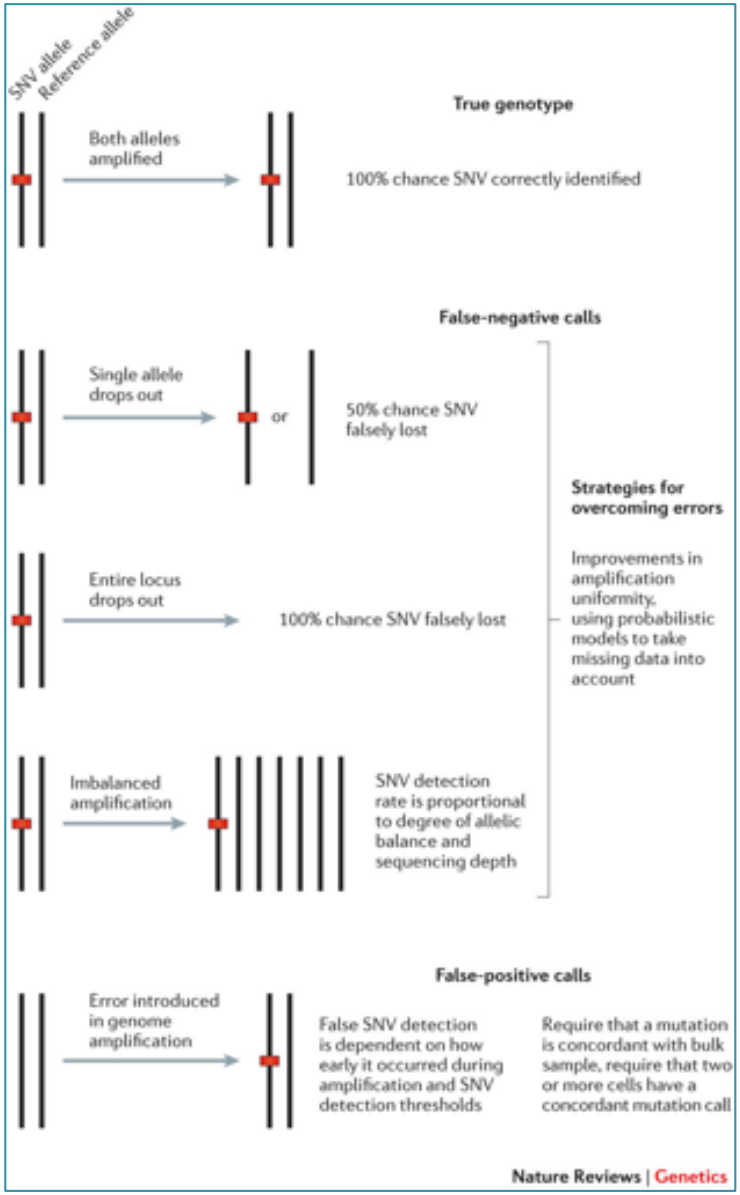
MALBAC: Multiple Annealing and Looping Based Amplification Cycles
Zong et al. (2012) Science

Paul Blainey, FEMS Microbiol Rev. 2013

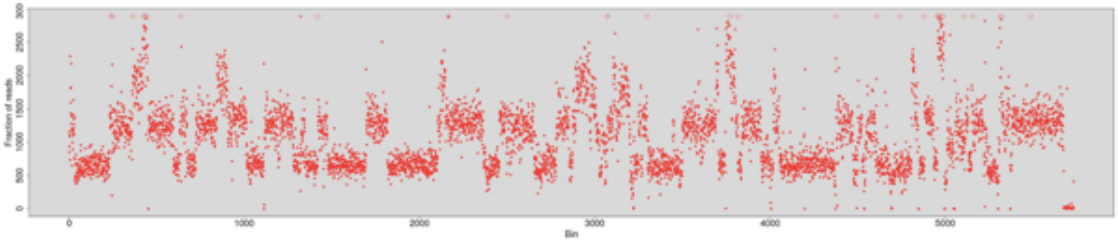


Fluidigm C1

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

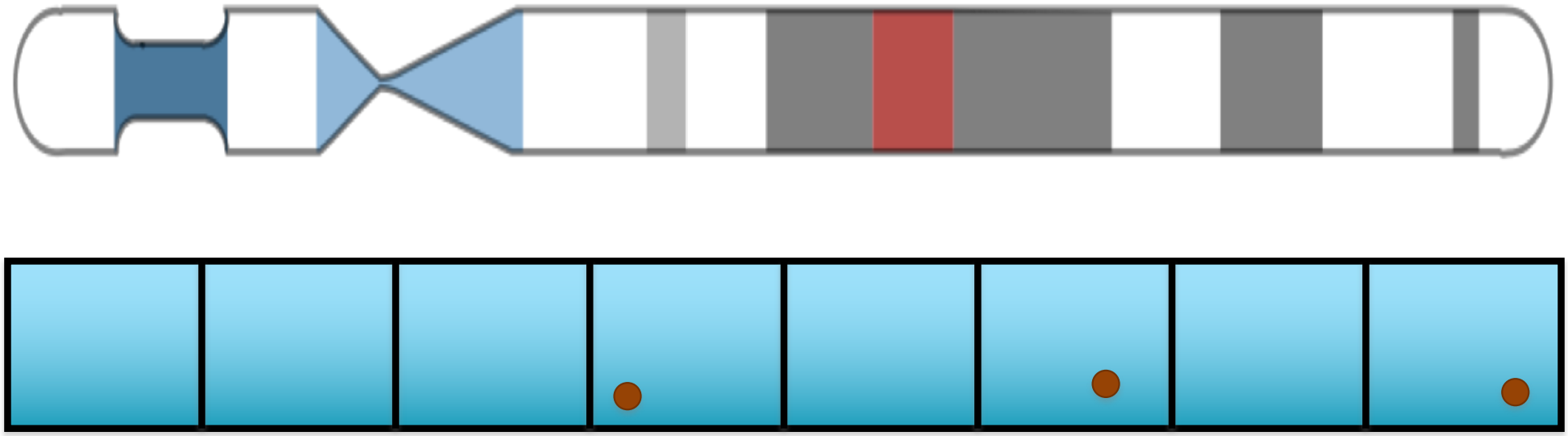
Coverage is very sparse and noisy

-> 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

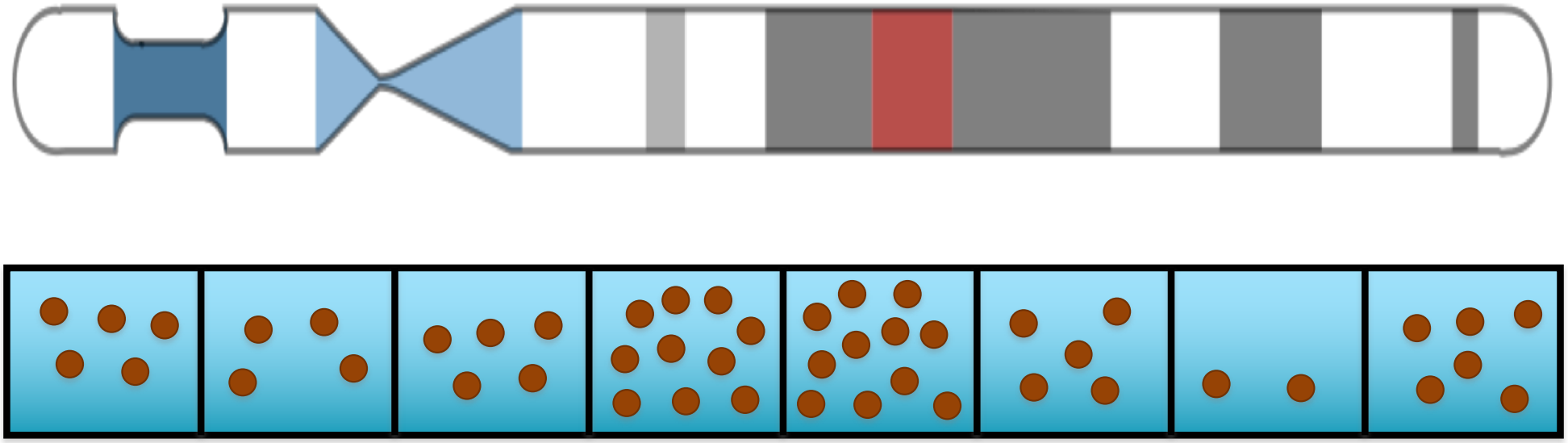


Single Cell CNV analysis

- 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

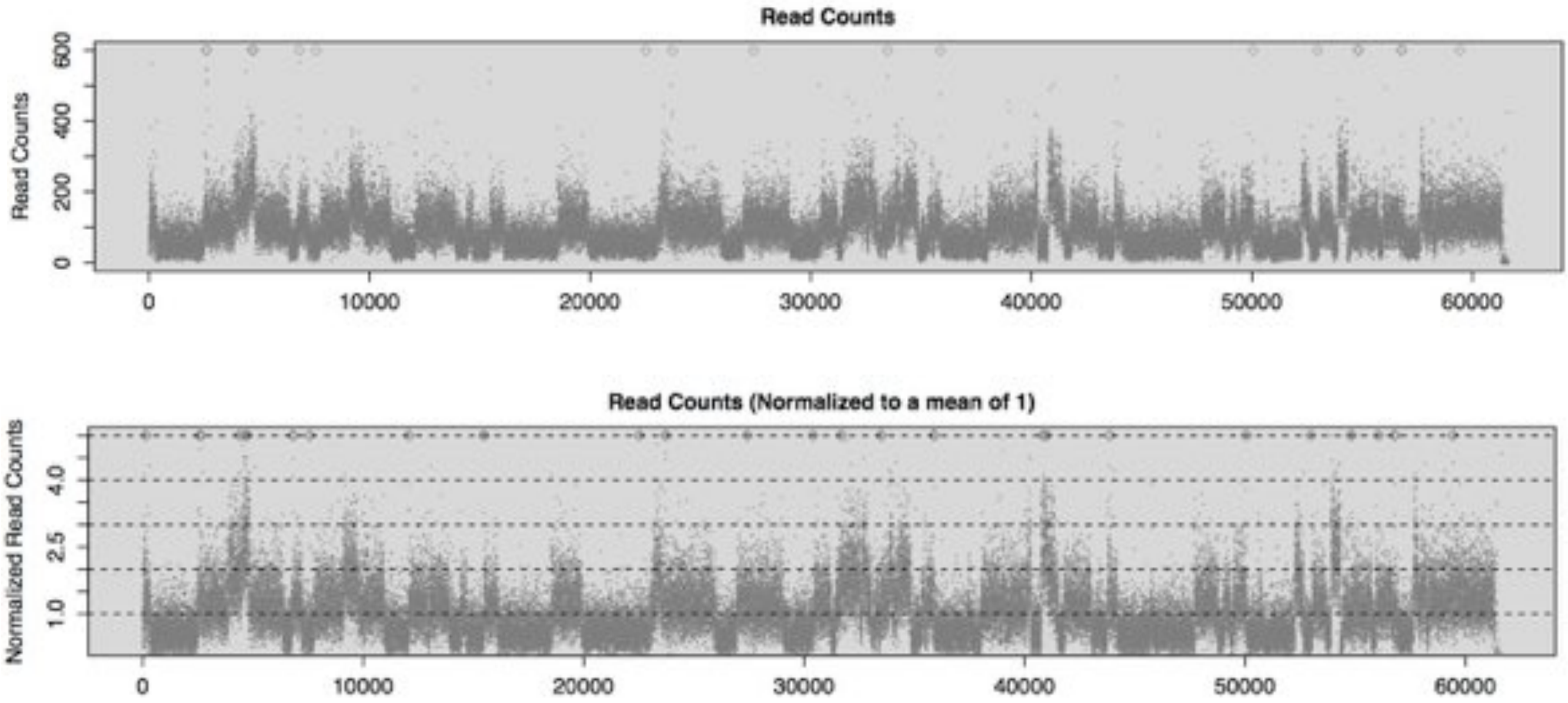
5	4	5	10	11	5	2	5

Single Cell CNV analysis

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

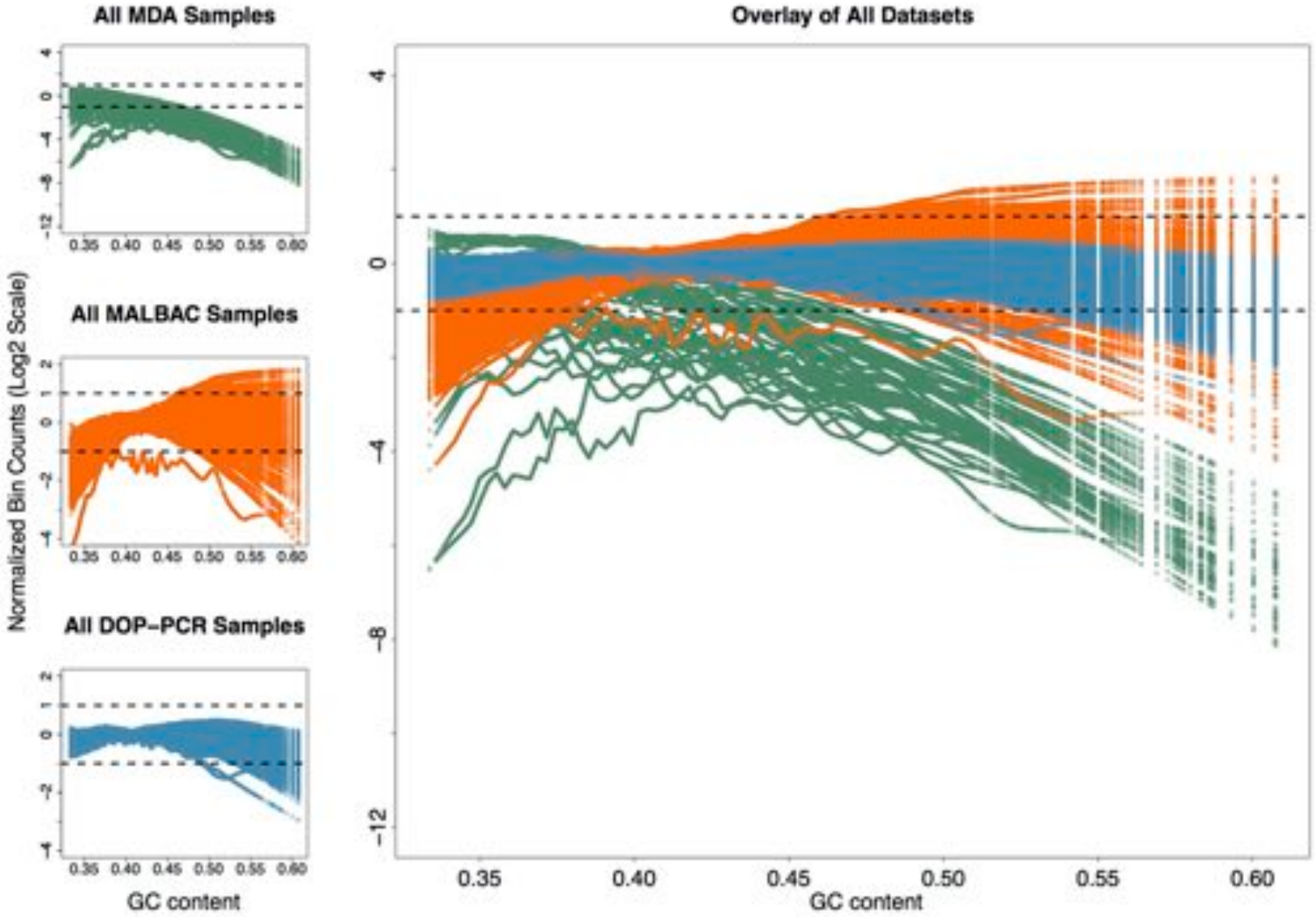
Use uniquely mappable bases to establish bins

2) Normalization



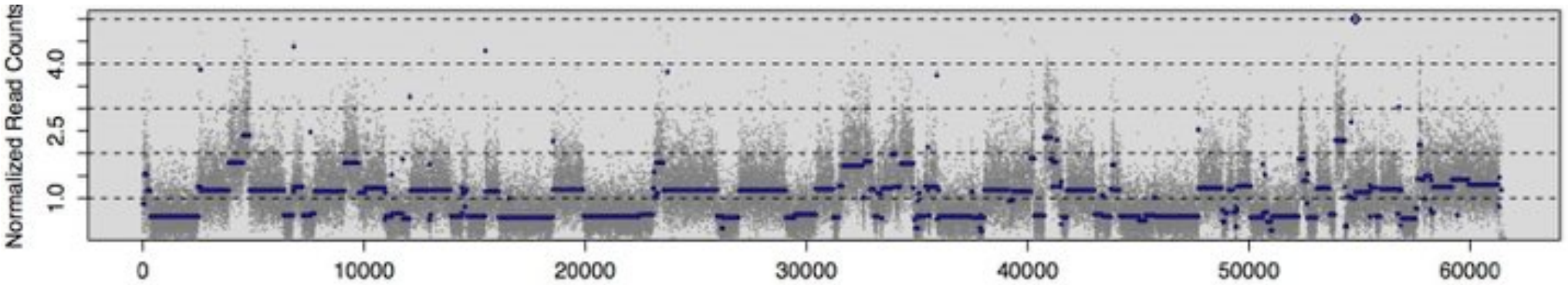
Also correct for mappability, GC content, amplification biases

GC Bias

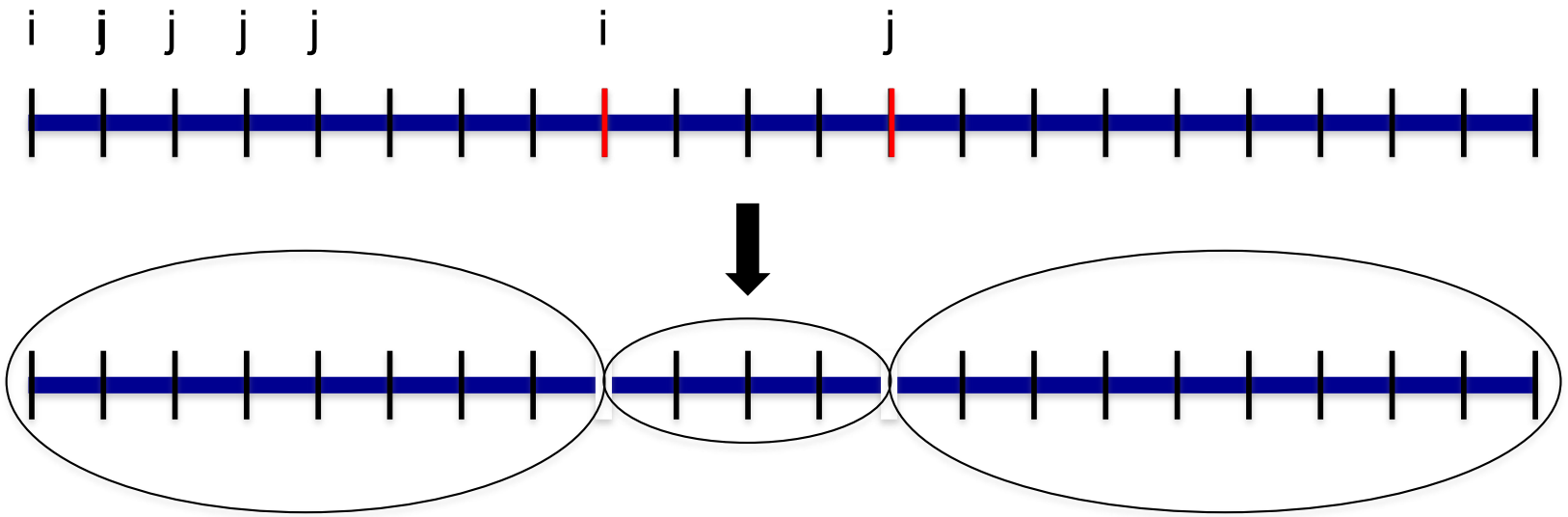


Garvin and Aboukhalil et al., Nature Methods, 2015

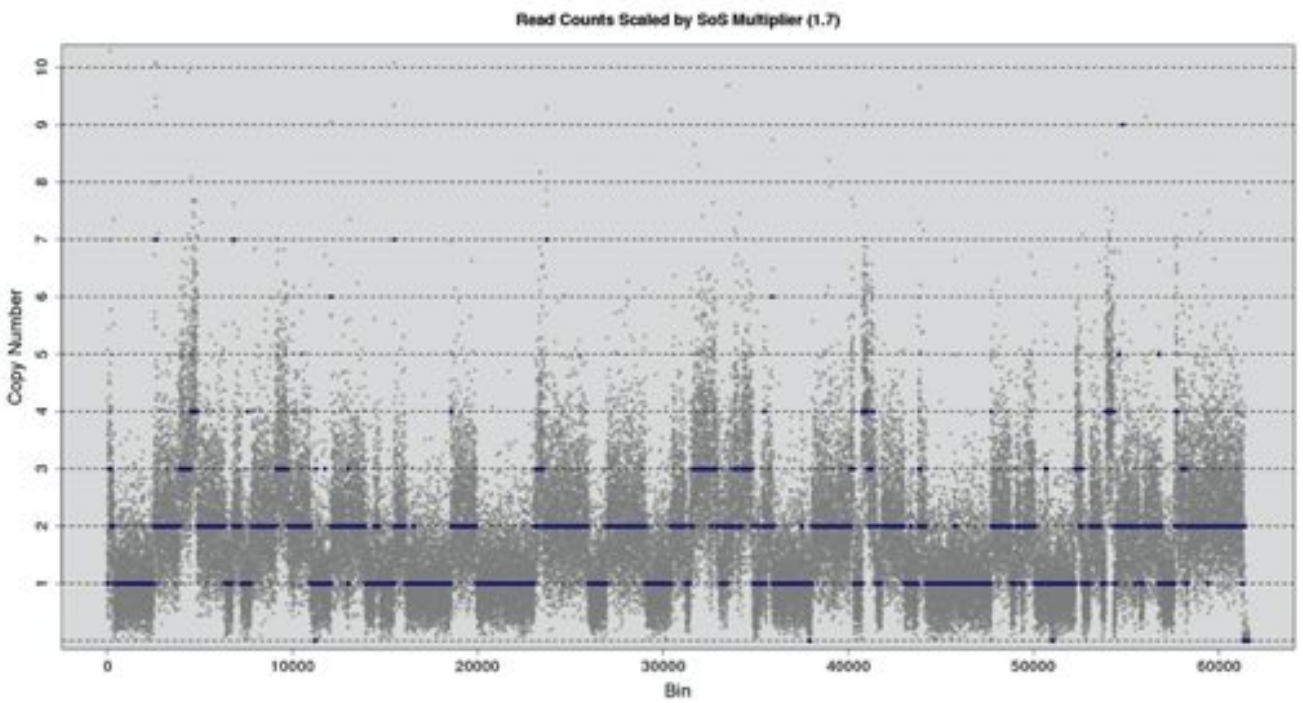
3) Segmentation



Circular Binary Segmentation (CBS)

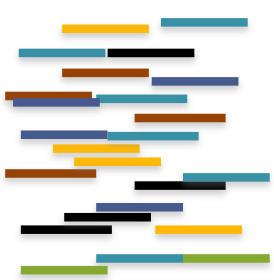


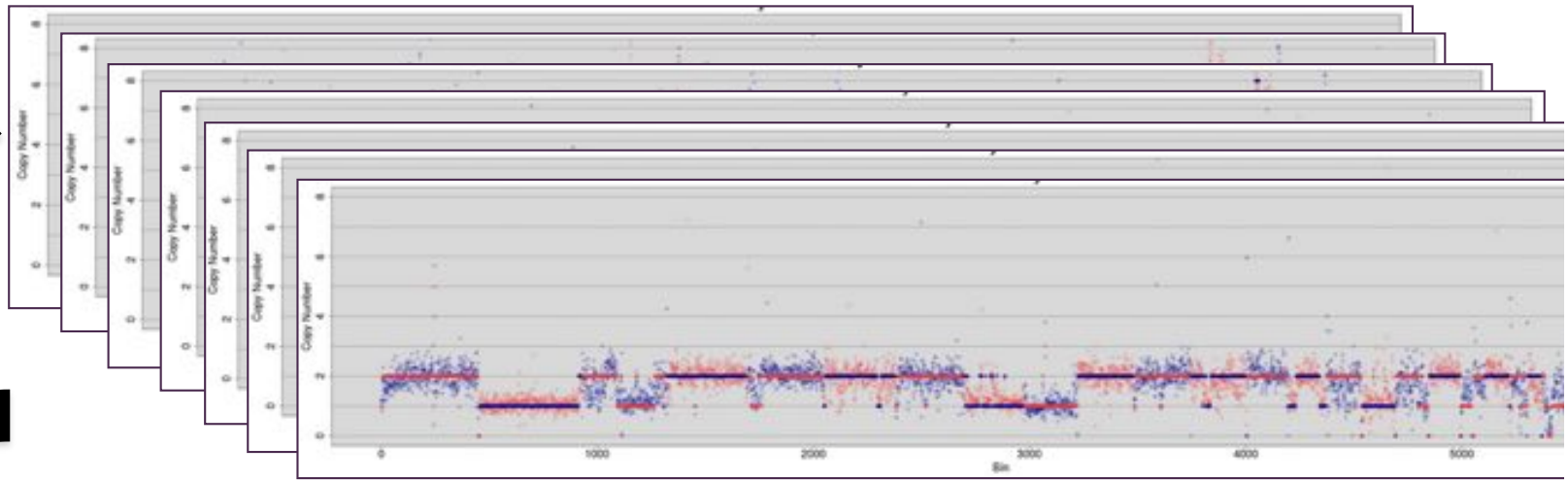
4) Estimating Copy Number

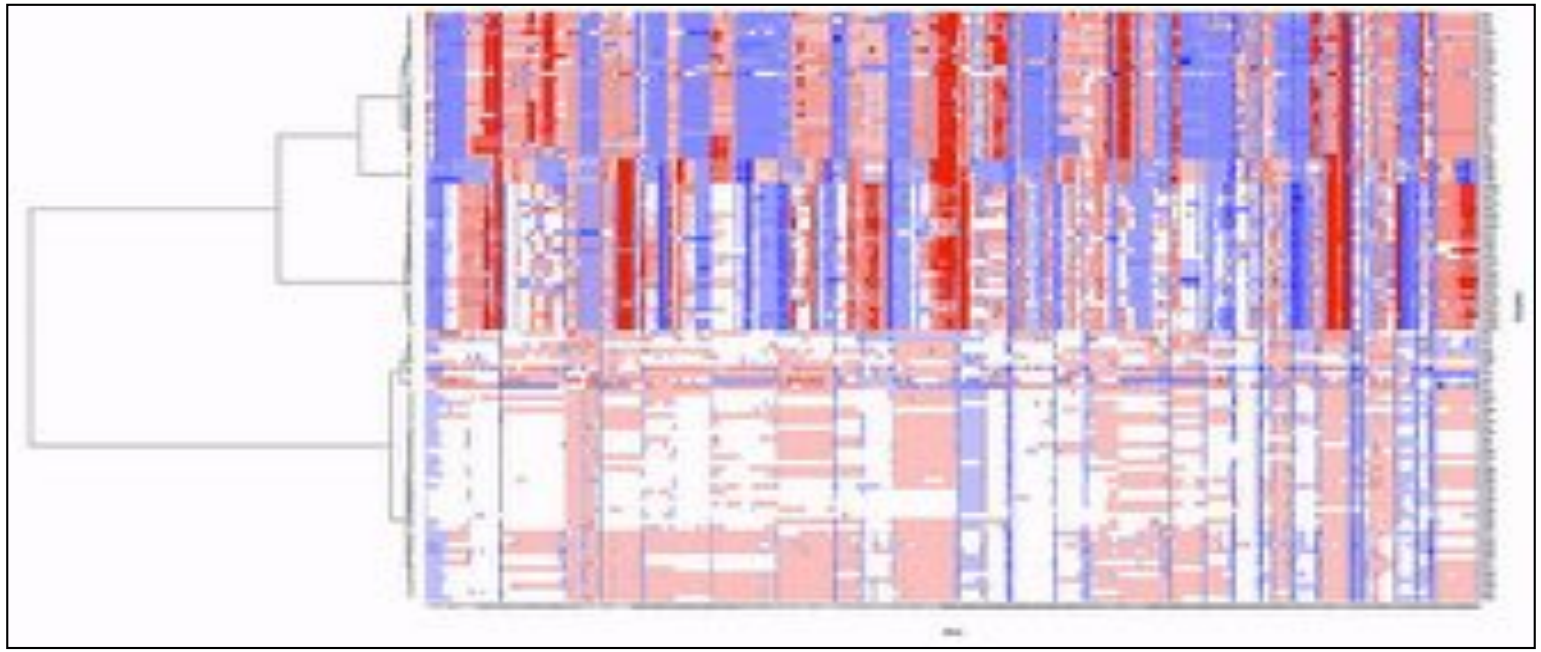


$$CN = argmin\left\{\sum_{i,j}(\hat{Y}_{i,j} - Y_{i,j})^2\right\}$$

5) Cells to Populations

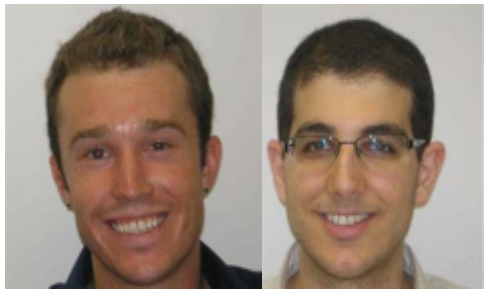


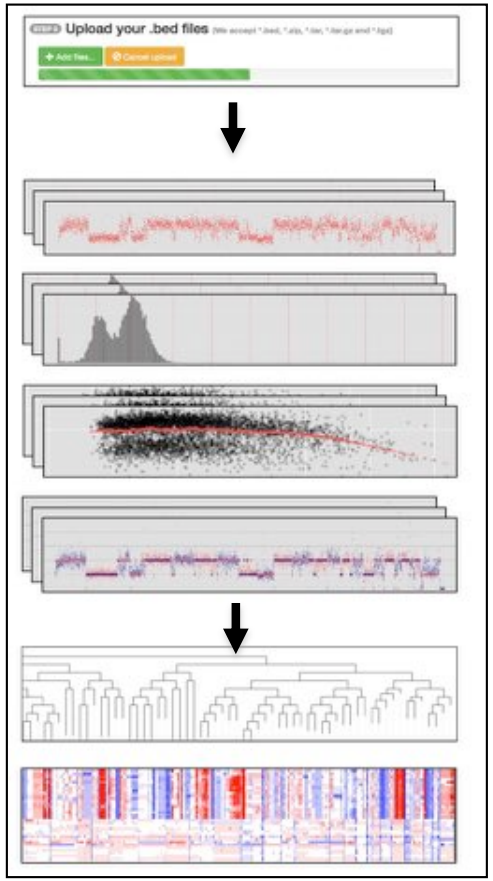




Gingko

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