Analysis of gene expression at single-cell resolution

Karthik Shekhar

Broad Institute

Broad Goals

- Introduction to the rapidly expanding world of **single-cell transcriptomics**
- Focus less on specific software tools but more on **underlying concepts** - so down the line, you can make informed choices
- **Lab session** : Hands on exercise on analyzing single-cell heterogeneity in R

Agenda

- Single cell analysis - why?
- A short survey of scRNA-seq methods
- Quality comparison of different methods and power analysis
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R
- Some example applications
- Future

Agenda

- Single cell analysis - why?
- A short survey of scRNA-seq methods

Coffee Break 1

- Quality comparison of different methods and power analysis
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R

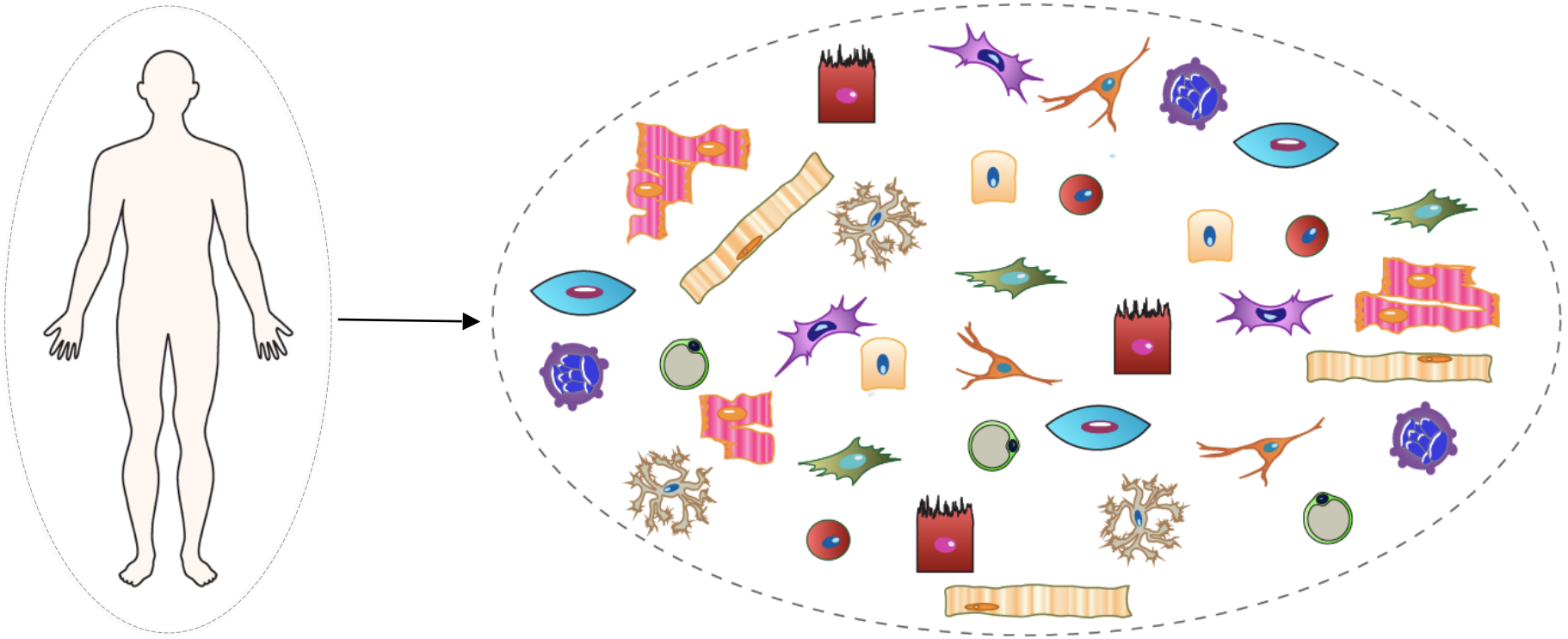
Coffee Break 2

- Some example applications
- Future

Agenda

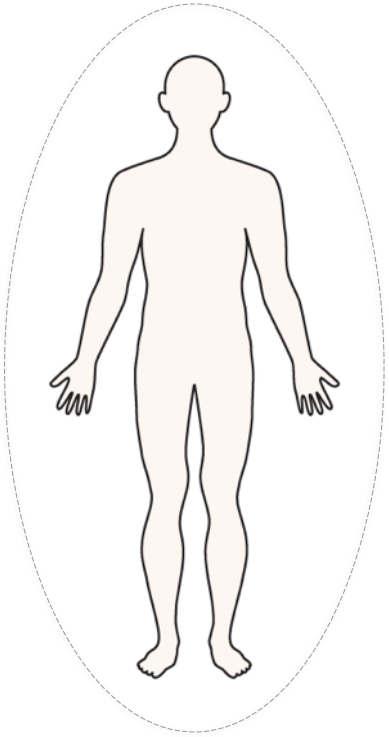
- **Single cell analysis - why?**
- A short survey of scRNA-seq methods
- Quality comparison of different methods and power analysis
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R
- Some example applications
- Future

Cells are the basic unit of life

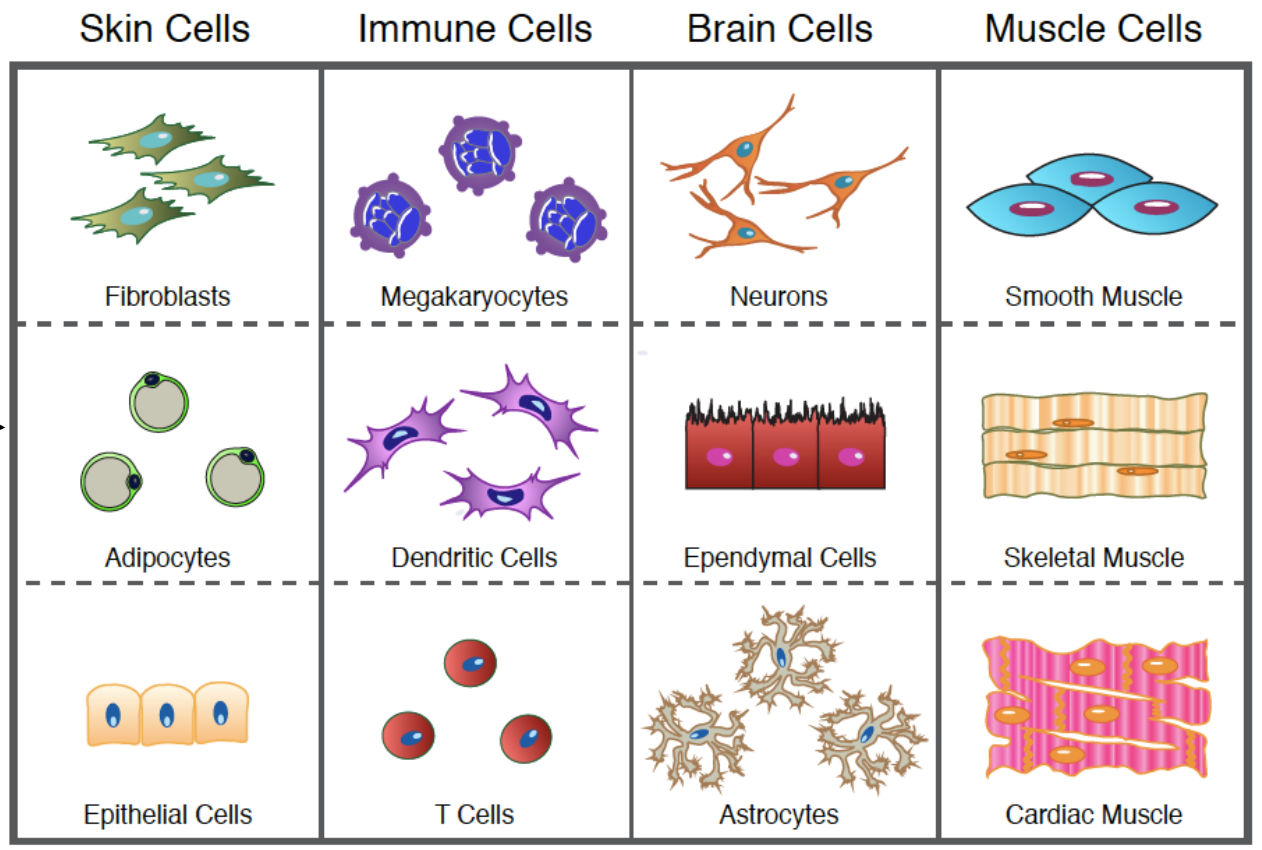


30 trillion cells

Cells are the basic unit of life

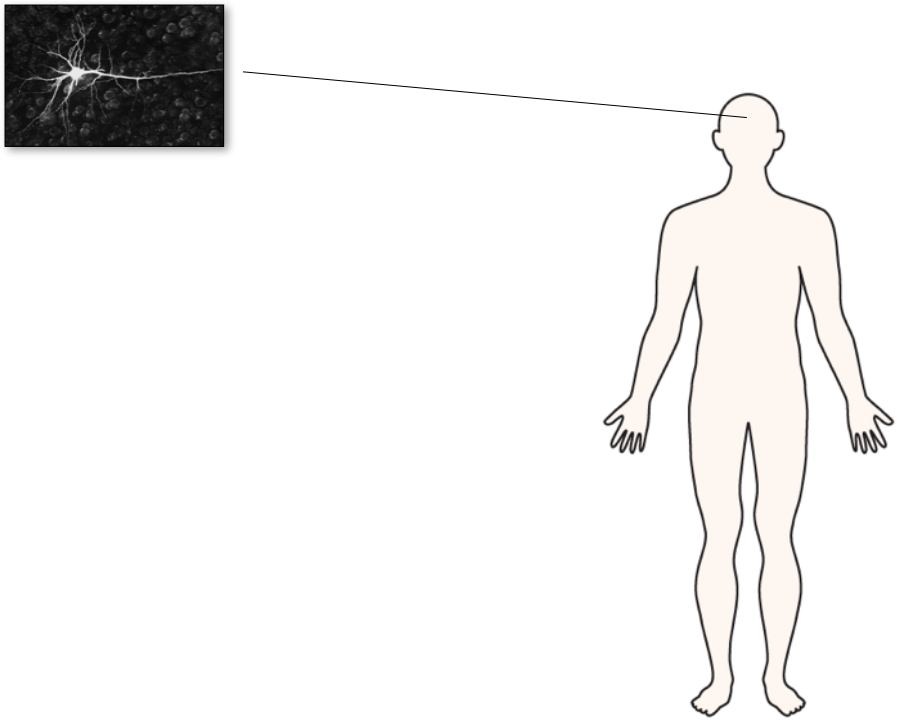


30 trillion cells



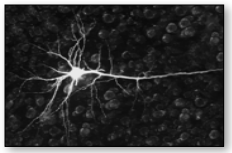
Cell identity encompasses many aspects…

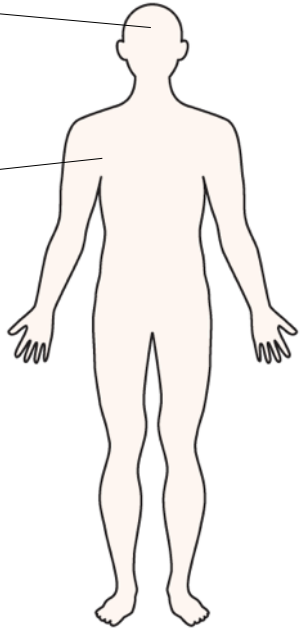
1. MORPHOLOGY



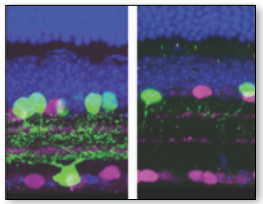
Cell identity encompasses many aspects…

1. MORPHOLOGY



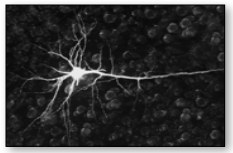


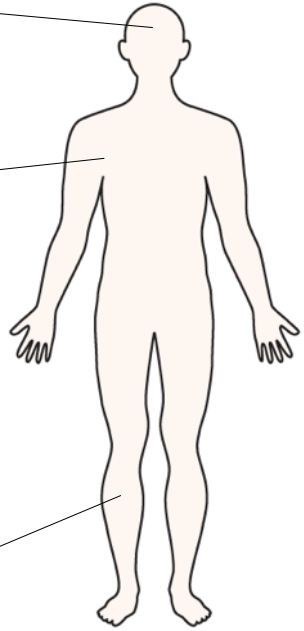
2. LOCATION



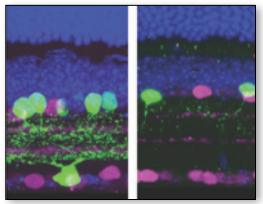
Cell identity encompasses many aspects…

1. MORPHOLOGY

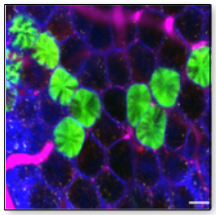




2. LOCATION

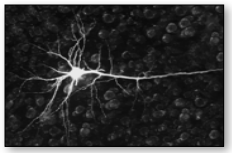


3. INTERACTIONS

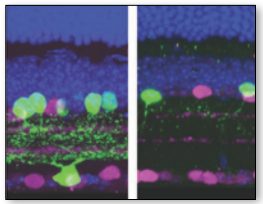


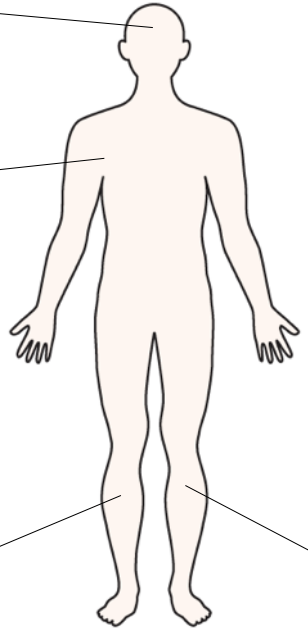
Cell identity encompasses many aspects…

1. MORPHOLOGY

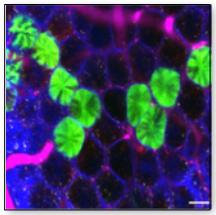


2. LOCATION

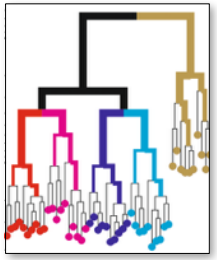




3. INTERACTIONS

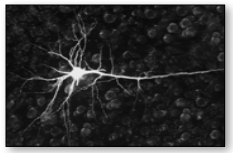


4. LINEAGE

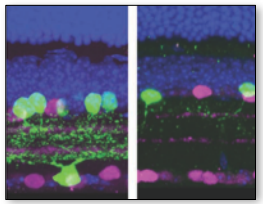


Cell identity encompasses many aspects…

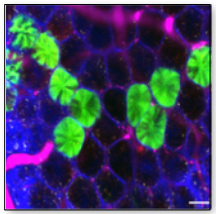
1. MORPHOLOGY

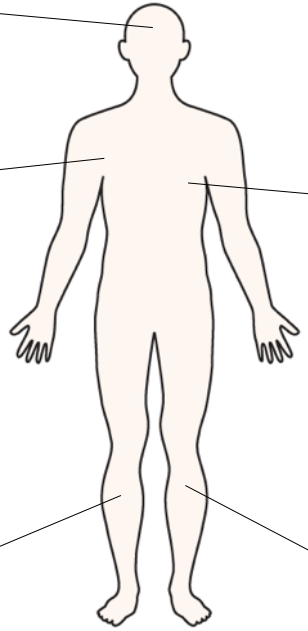


2. LOCATION

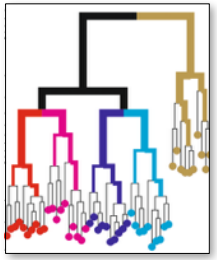


3. INTERACTIONS

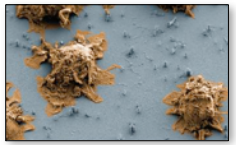


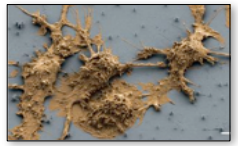


4. LINEAGE



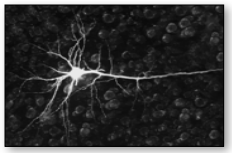
5. STATE



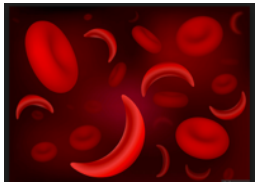


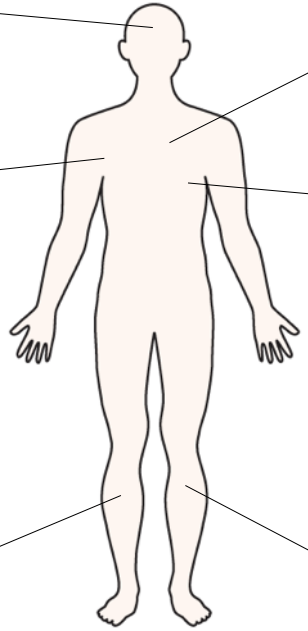
Cell identity encompasses many aspects…

1. MORPHOLOGY

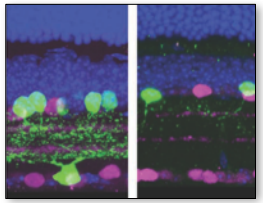


6. NORMAL/DISEASE

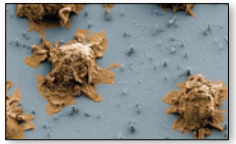


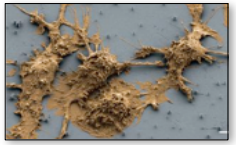


2. LOCATION

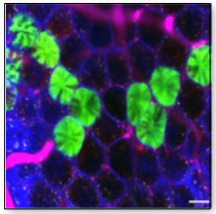


5. STATE

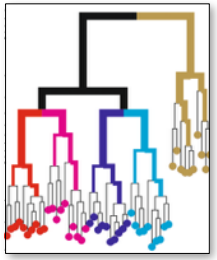




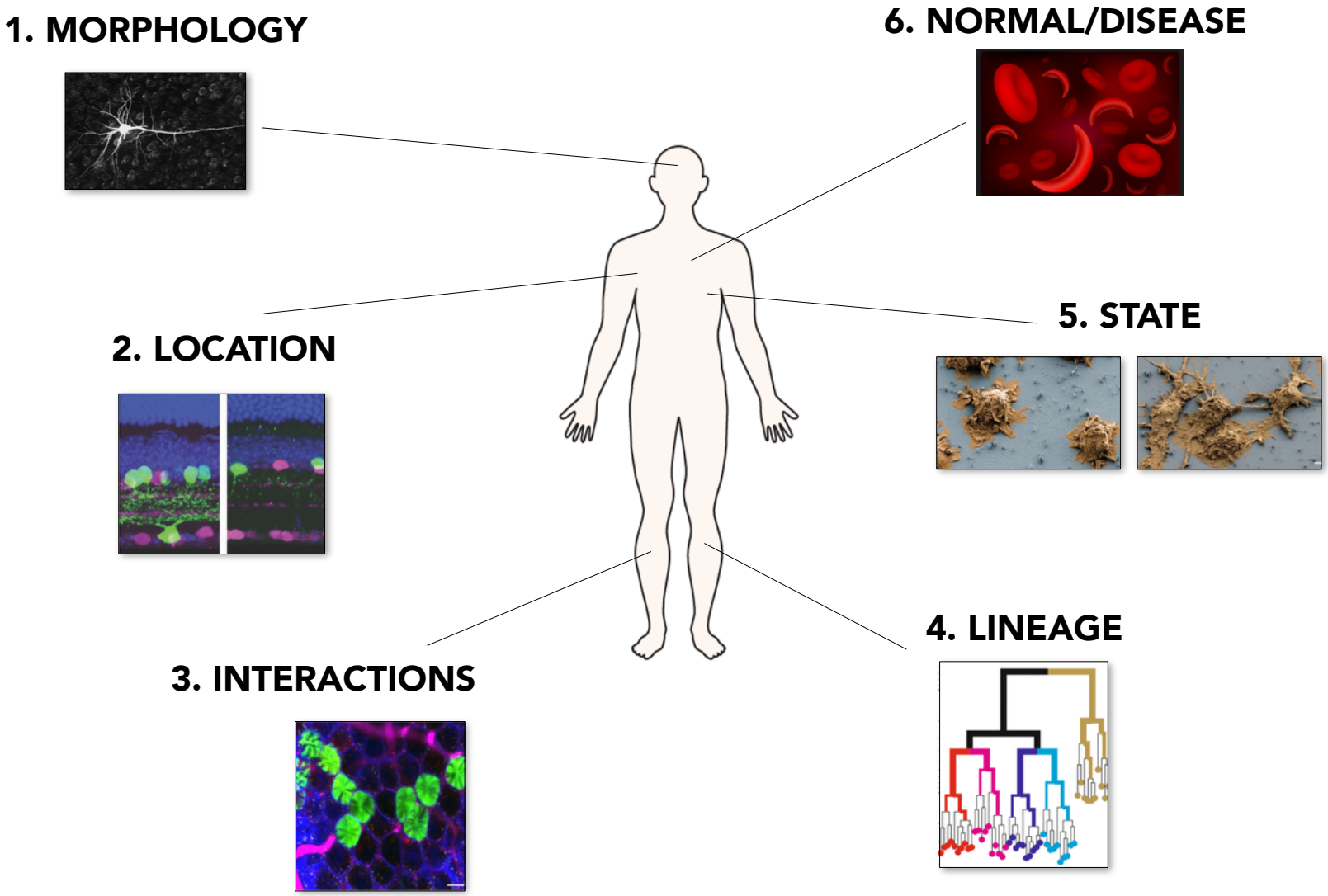
3. INTERACTIONS



4. LINEAGE

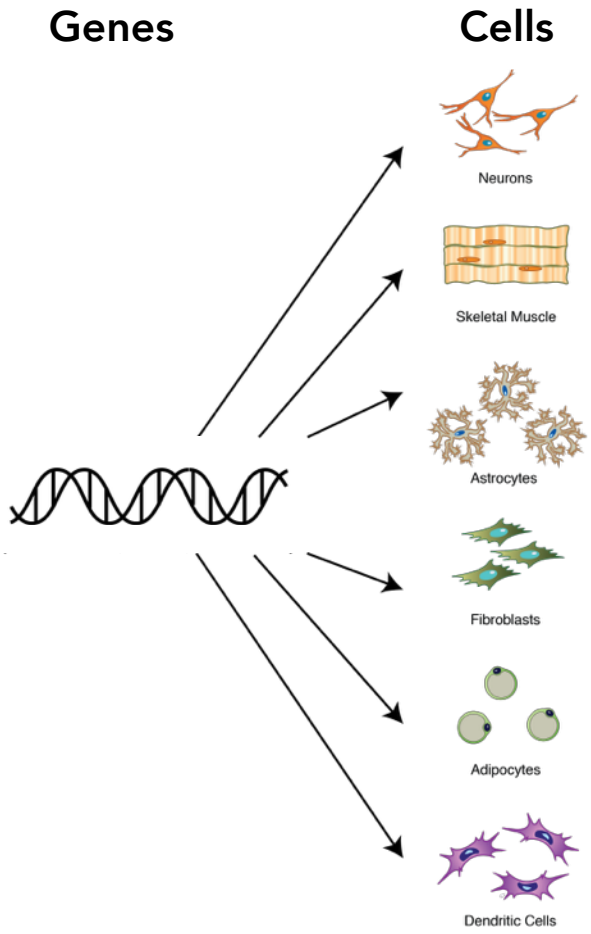


Cell identity encompasses many aspects…

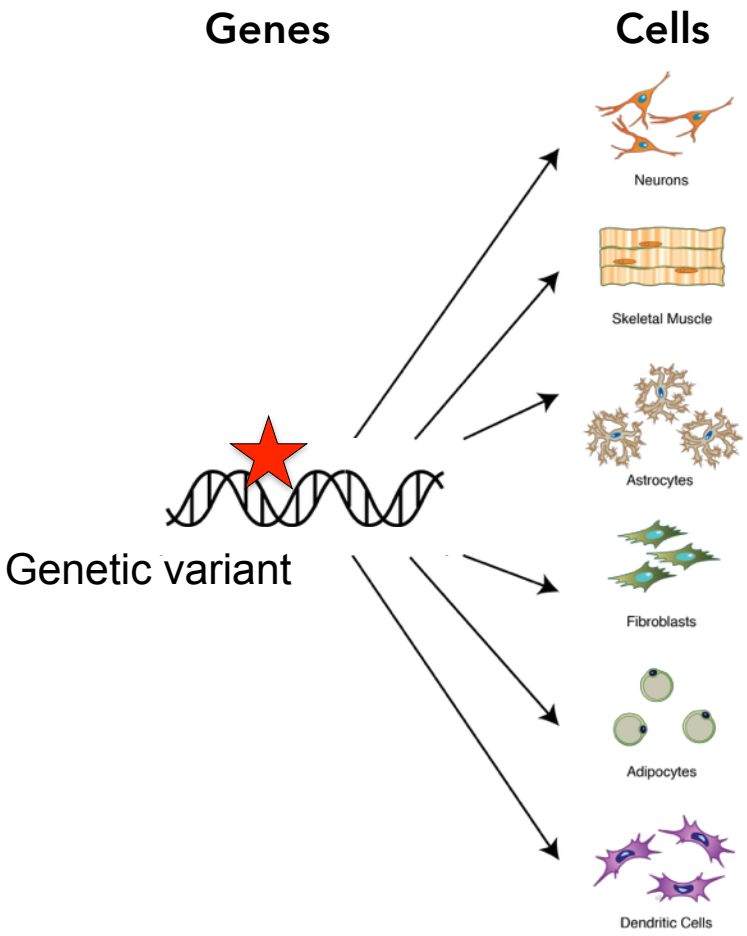


Each of these possesses a molecular correlate(s)

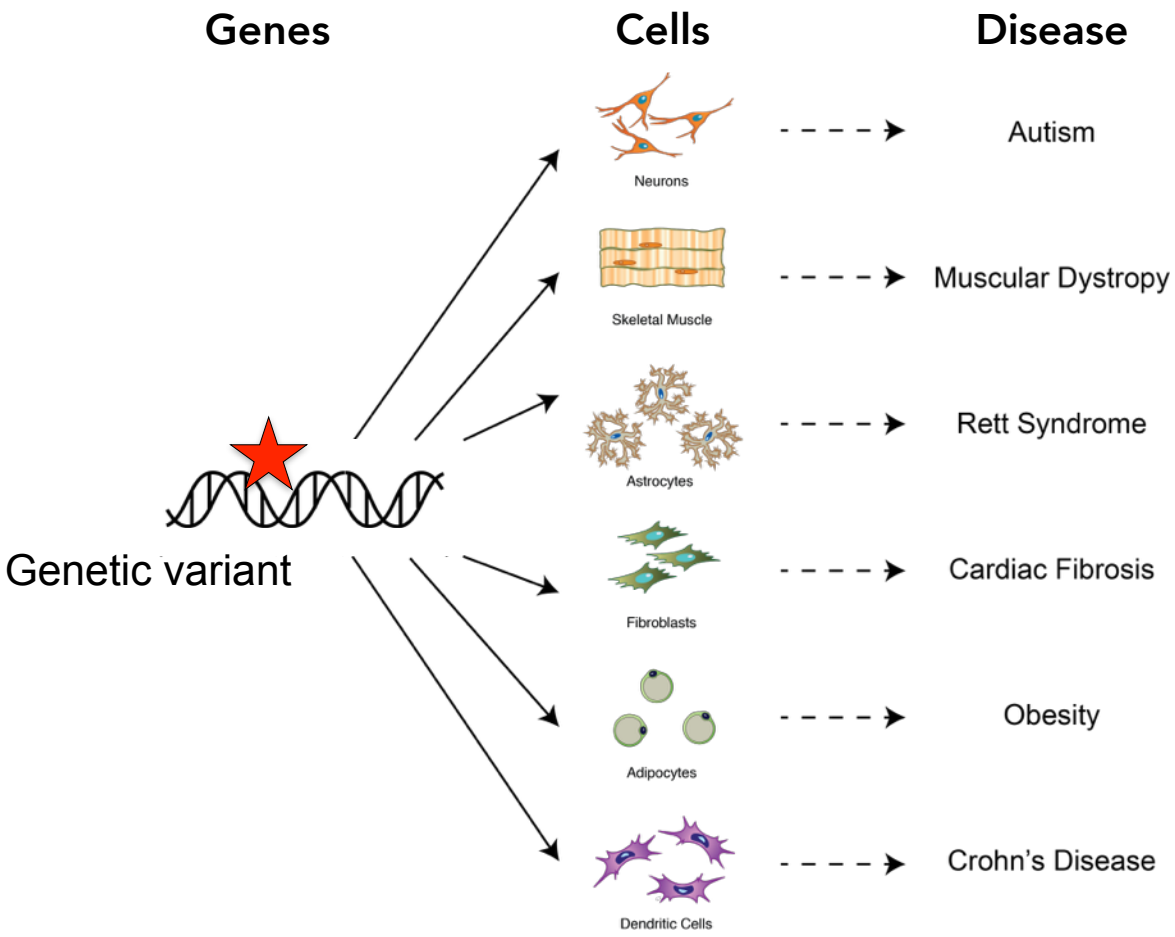
One genome, but many cells



One genome, but many cells



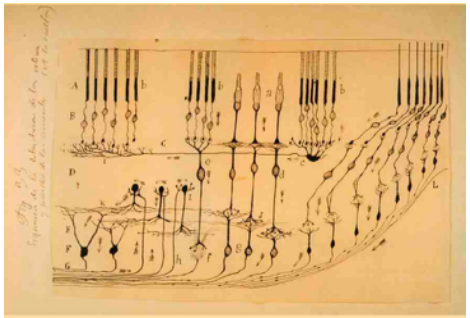
One genome, but many cells



The problem is that we do not know all of our cell types - what are they and what genes mark them

Methods to interrogate cellular diversity

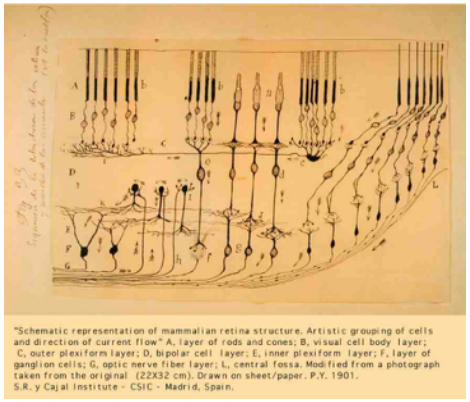
Light Microscopy (1850's)



"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901.
S.R. y Cajal Institute - CSIC - Madrid, Spain.

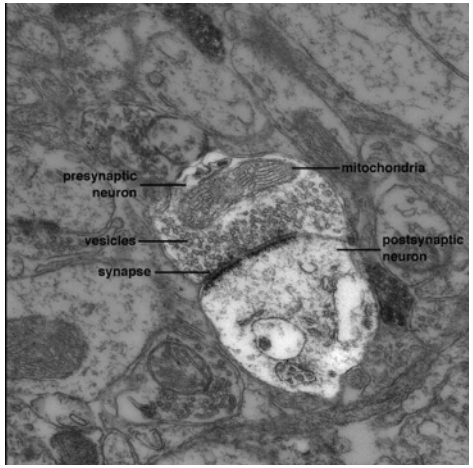
Methods to interrogate cellular diversity

Light Microscopy (1850's)



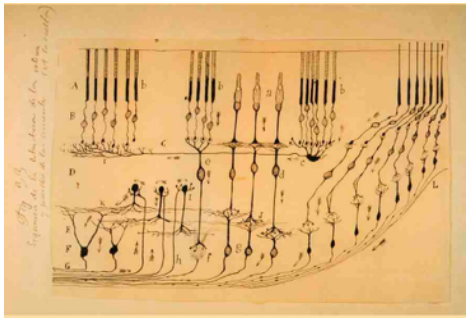
"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901.
S.R. y Cajal Institute - CSIC - Madrid, Spain.

Electron Microscopy (1940's)



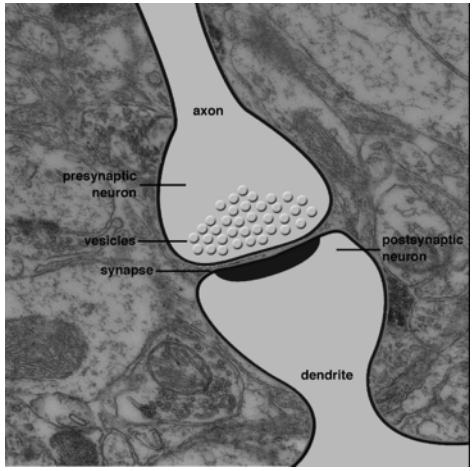
Methods to interrogate cellular diversity

Light Microscopy (1850's)



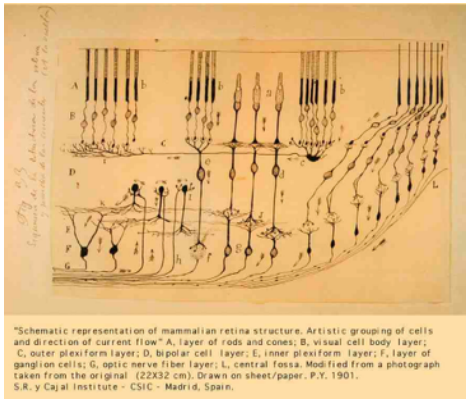
"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901.
S.R. y Cajal Institute - CSIC - Madrid, Spain.

Electron Microscopy (1940's)



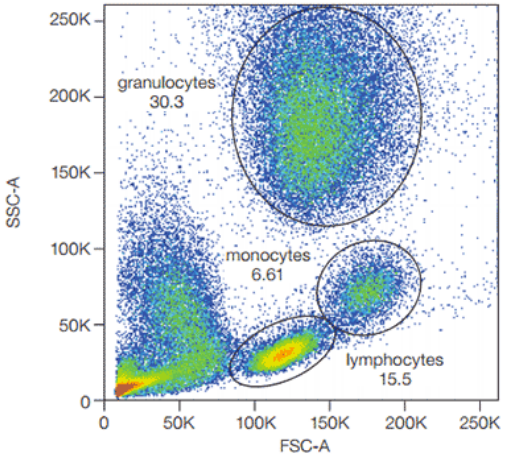
Methods to interrogate cellular diversity

Light Microscopy (1850's)

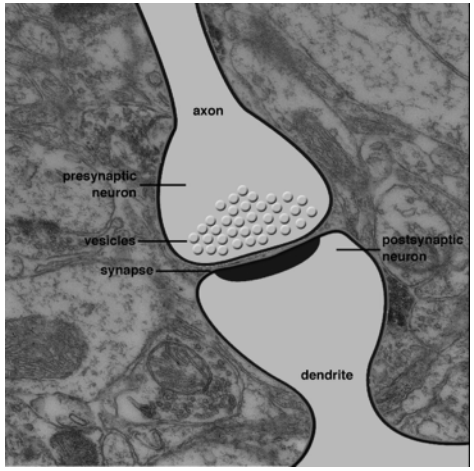


"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901. S.R. y Cajal Institute - CSIC - Madrid, Spain.

FACS (1970's)

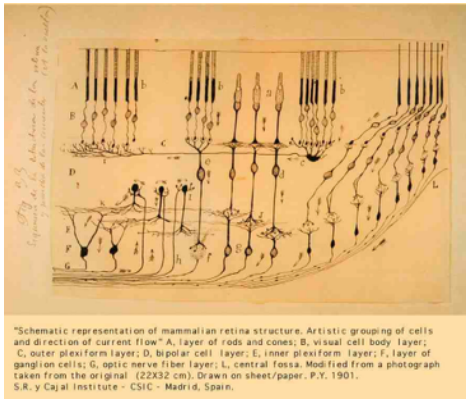


Electron Microscopy (1940's)



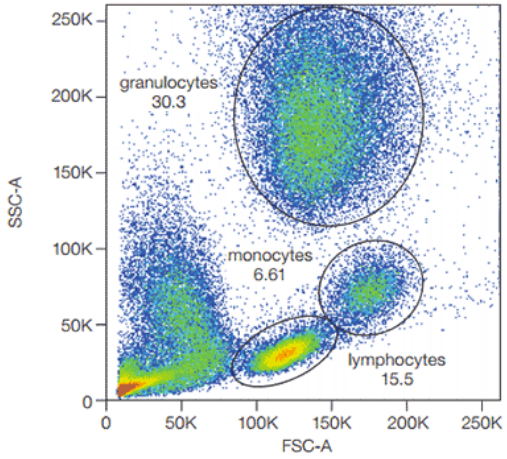
Methods to interrogate cellular diversity

Light Microscopy (1850's)

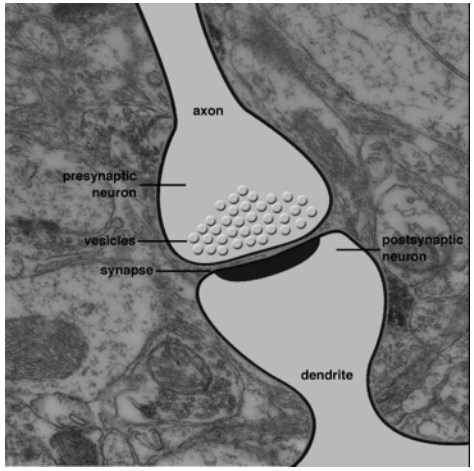


"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901. S.R. y Cajal Institute - CSIC - Madrid, Spain.

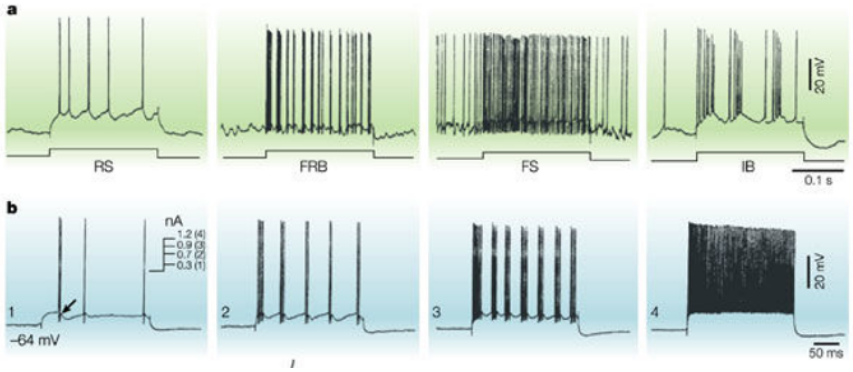
FACS (1970's)



Electron Microscopy (1940's)

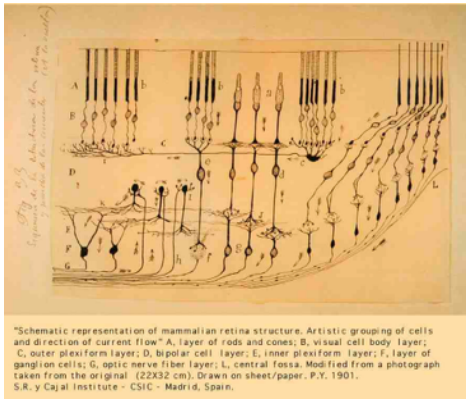


Electrophysiology (1970's)



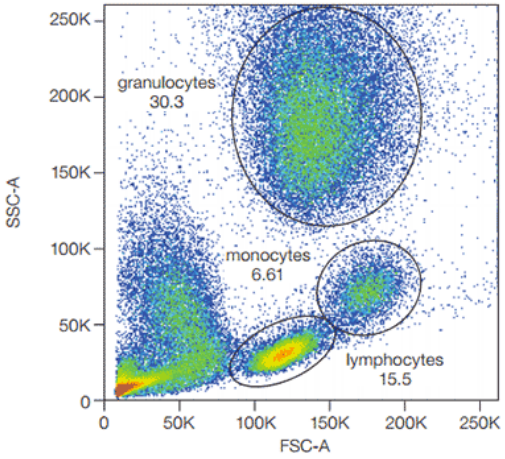
Methods to interrogate cellular diversity

Light Microscopy (1850's)

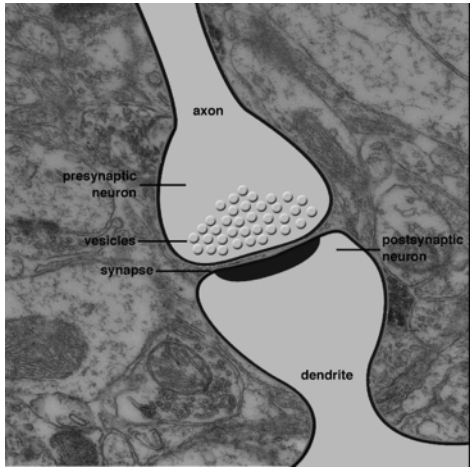


"Schematic representation of mammalian retina structure. Artistic grouping of cells and direction of current flow" A, layer of rods and cones; B, visual cell body layer; C, outer plexiform layer; D, bipolar cell layer; E, inner plexiform layer; F, layer of ganglion cells; G, optic nerve fiber layer; L, central fossa. Modified from a photograph taken from the original (22X32 cm). Drawn on sheet/paper. P.Y. 1901. S.R. y Cajal Institute - CSIC - Madrid, Spain.

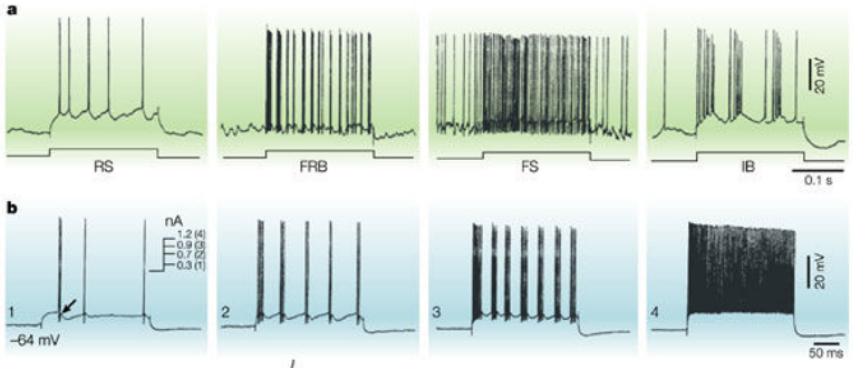
FACS (1970's)



Electron Microscopy (1940's)

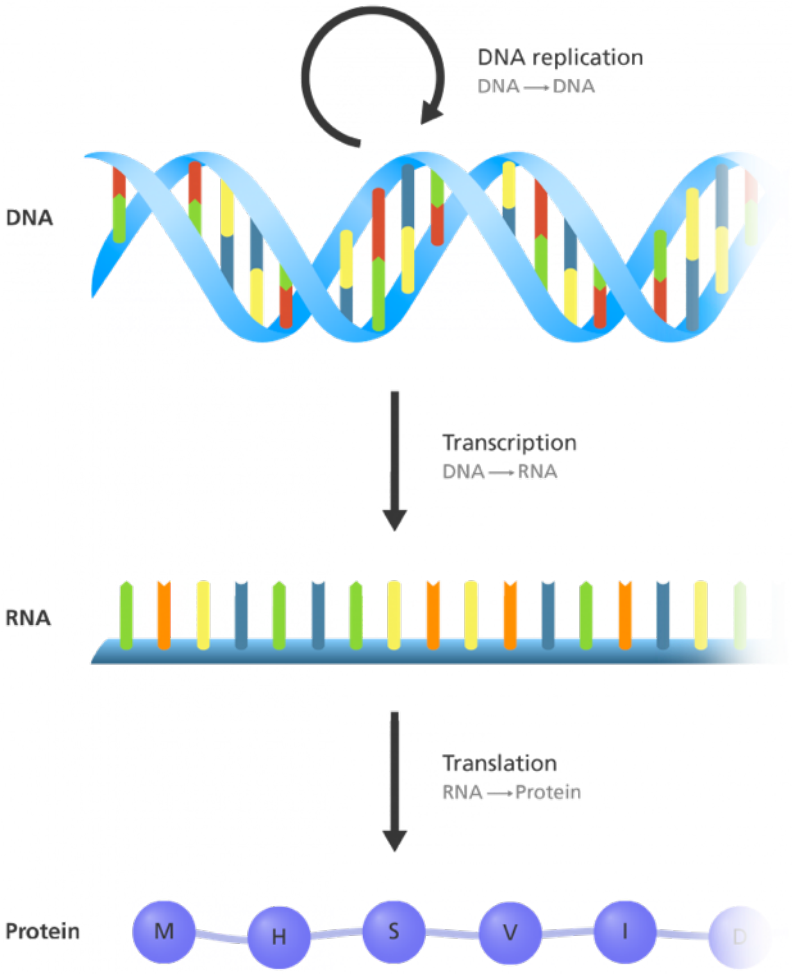


Electrophysiology (1970's)



Recent advances in "high throughput" counterparts (many cells, many features)

Measuring molecular features in a cell

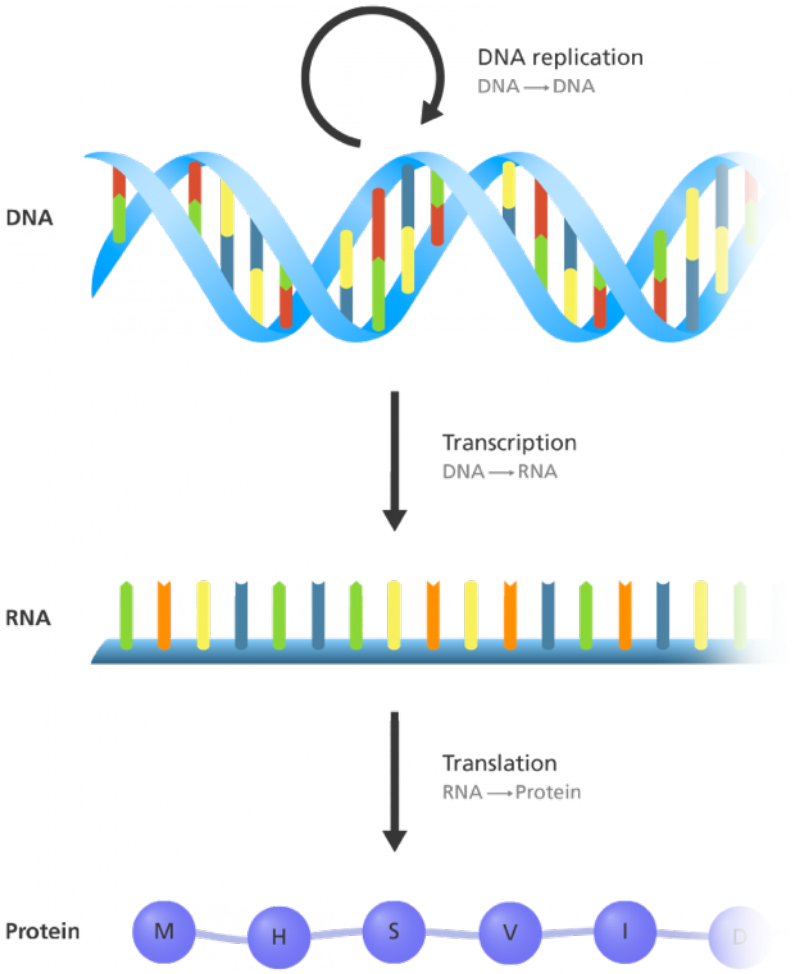


- Genome: SNPs, CNVs, …
- Epigenome: Accessibility, Histone modifications …
- TFs bound at any given site (ChIP-seq)

- mRNA + ncRNA + …
- mRNA modifications

- protein levels
- post translational modifications

Measuring molecular features in a cell



- Genome: SNPs, CNVs, …
- Epigenome: Accessibility, Histone modifications …
- TFs bound at any given site (ChIP-seq)

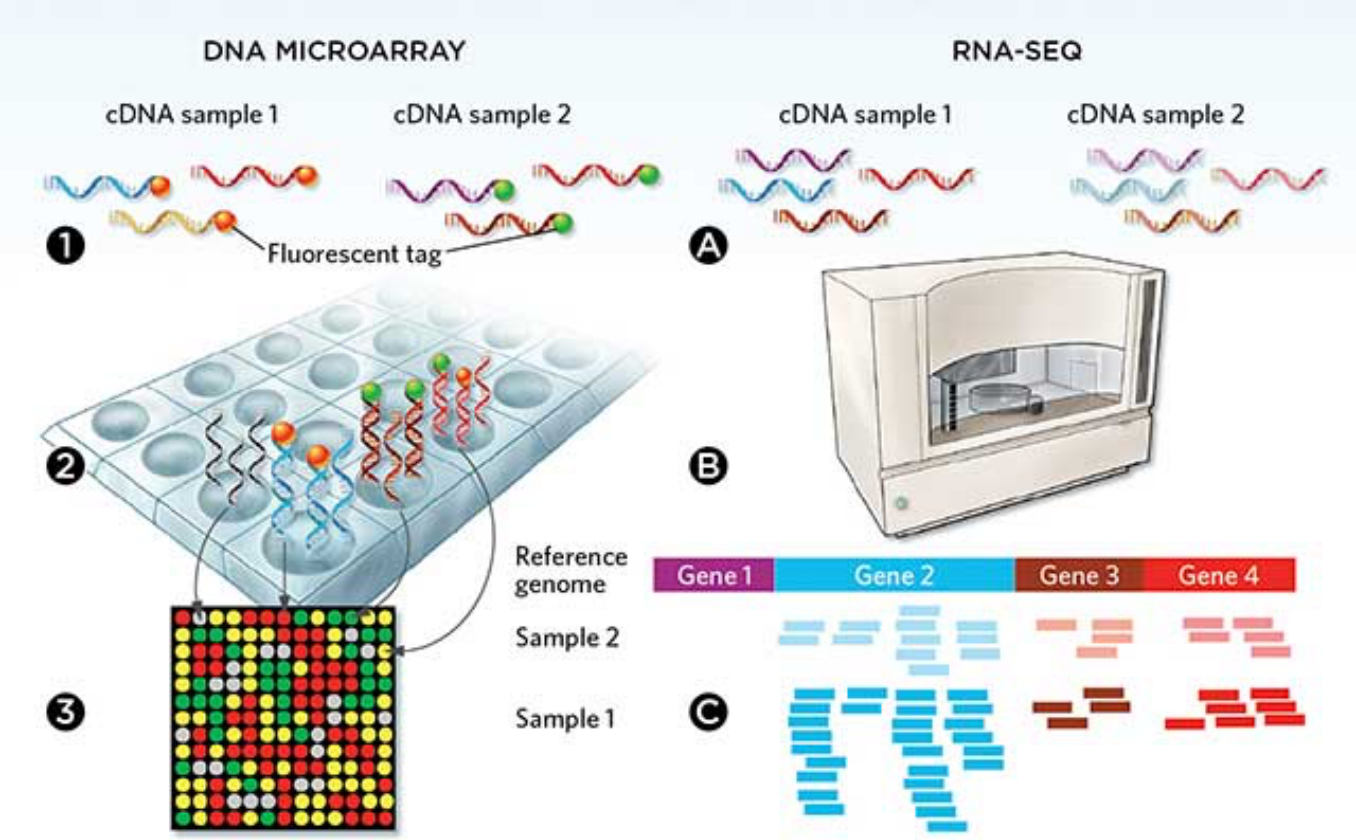
- mRNA + ncRNA + …
- mRNA modifications

- protein levels
- post translational modifications

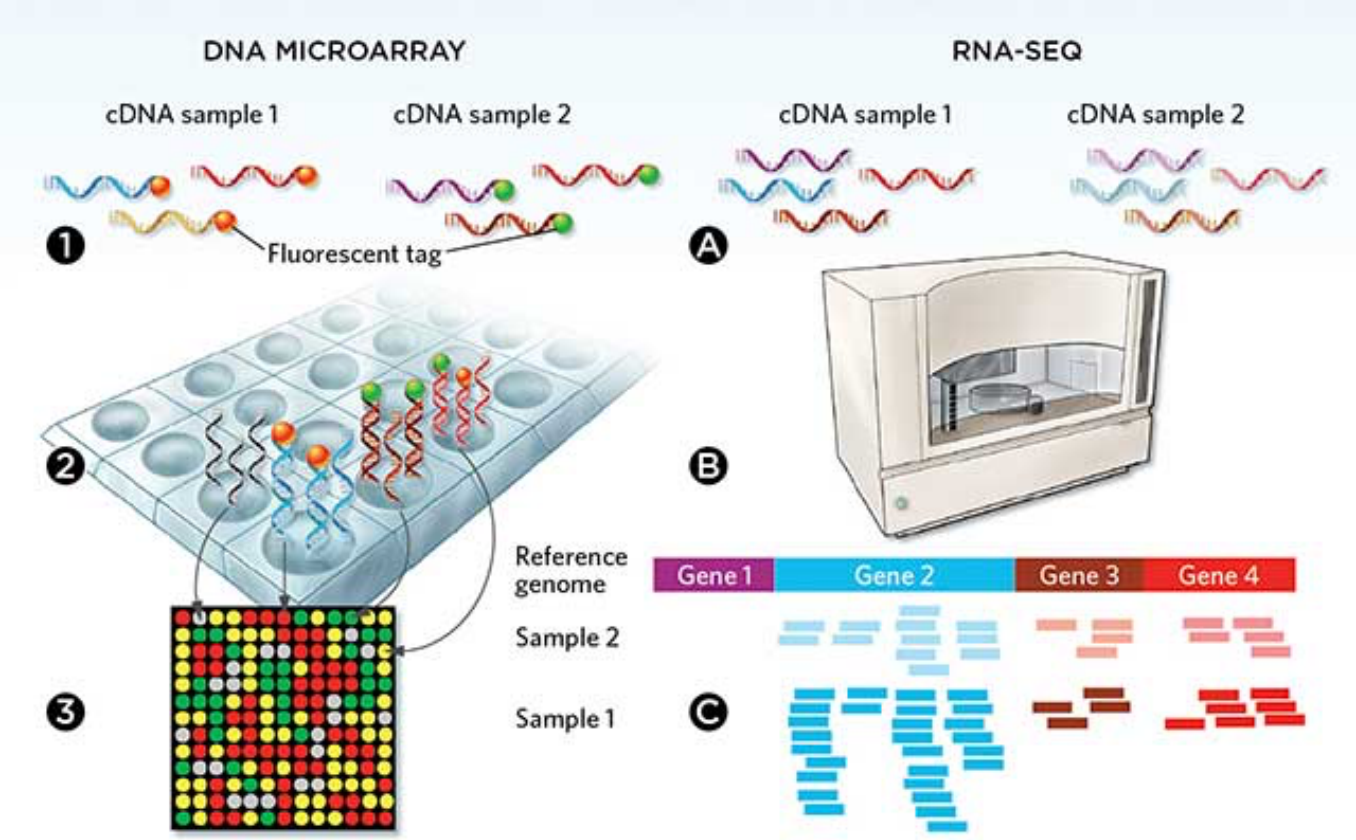
Currently transcriptome-wide gene expression is the most easily accessible and advanced!

Measuring genome wide expression

Measuring genome wide expression

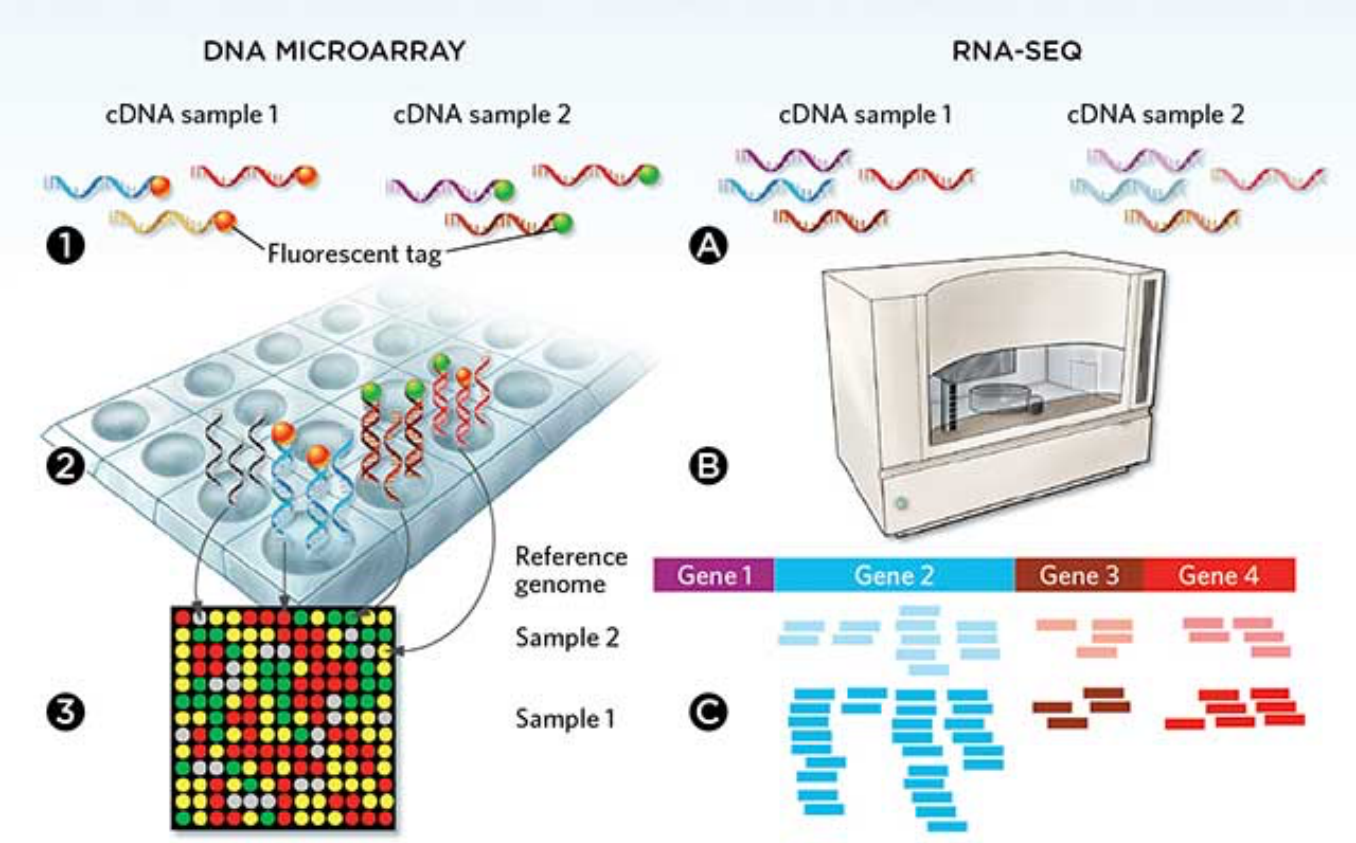


Measuring genome wide expression



Probe-dependent

Measuring genome wide expression

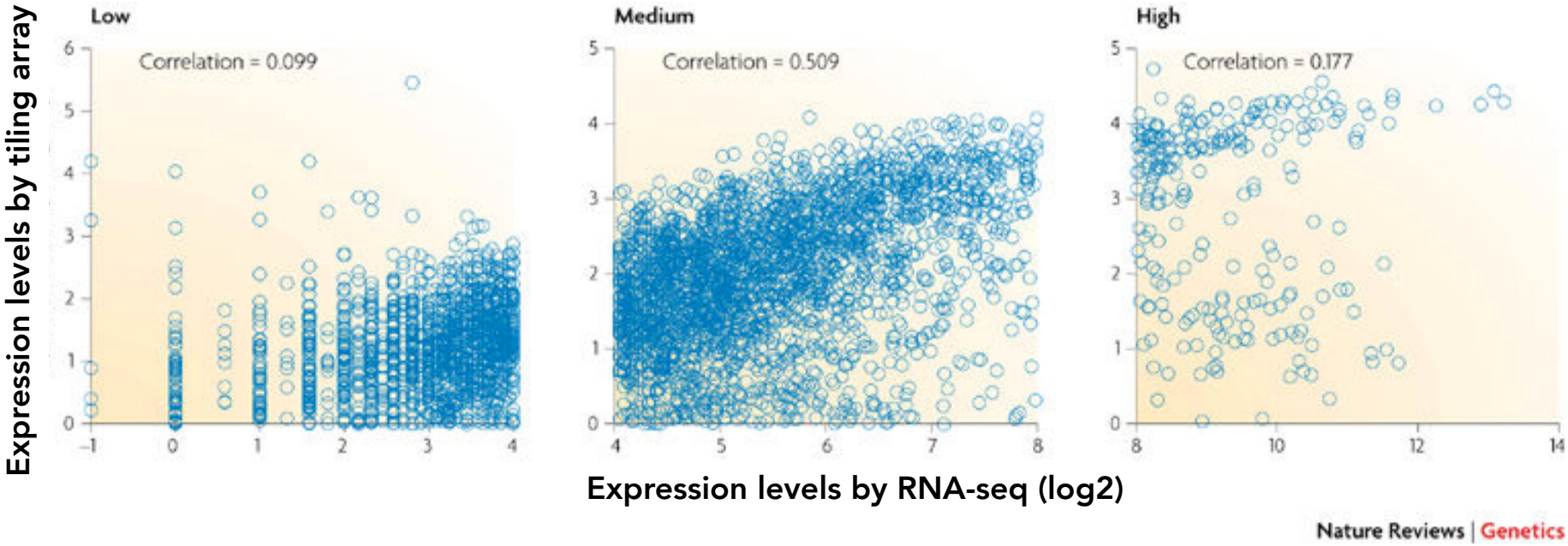


Probe-dependent

Probe-independent

RNA-seq vs microarrays

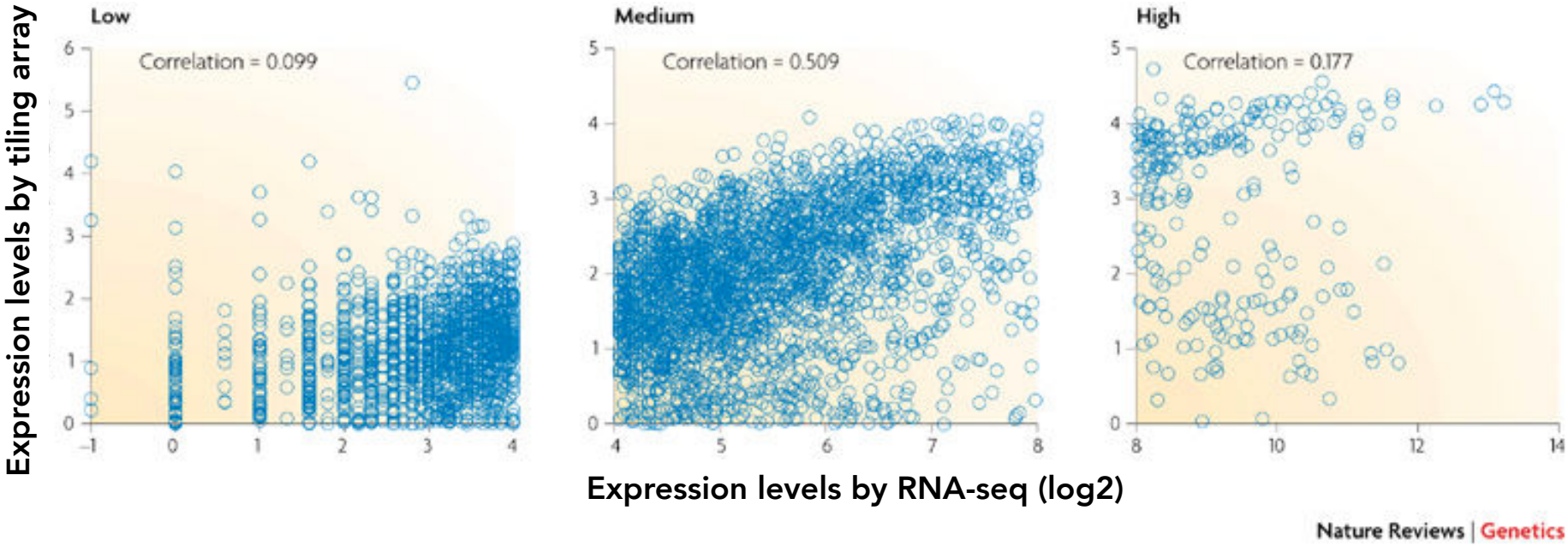
- Microarrays convert mRNA abundance into a fluorescence signal (**ANALOG**)
- In RNA-seq, we directly count the reads/molecules (**DIGITAL**)



Wang, Gerstein, and Snyder, *Nat. Rev. Genetics*, 2009

RNA-seq vs microarrays

- Microarrays convert mRNA abundance into a fluorescence signal (**ANALOG**)
- In RNA-seq, we directly count the reads/molecules (**DIGITAL**)



Which makes RNA-seq, in principle, a much more sensitive and accurate read out of gene expression!

Wang, Gerstein, and Snyder, *Nat. Rev. Genetics*, 2009

Typical RNA-seq workflow

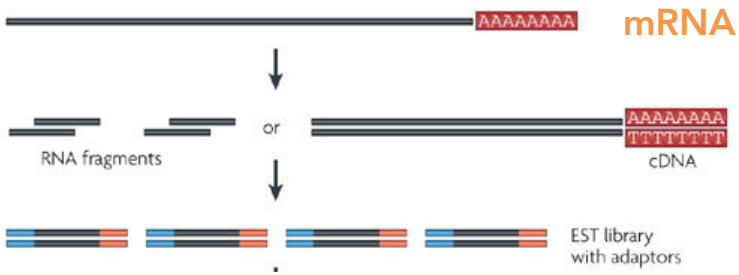
mRNA from a source

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Typical RNA-seq workflow

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Typical RNA-seq workflow



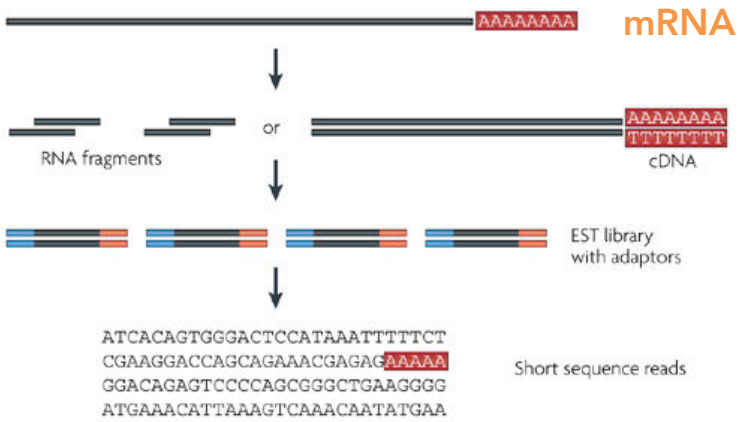
mRNA from a source

RT + + amplification

Add adapters to ends

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Typical RNA-seq workflow



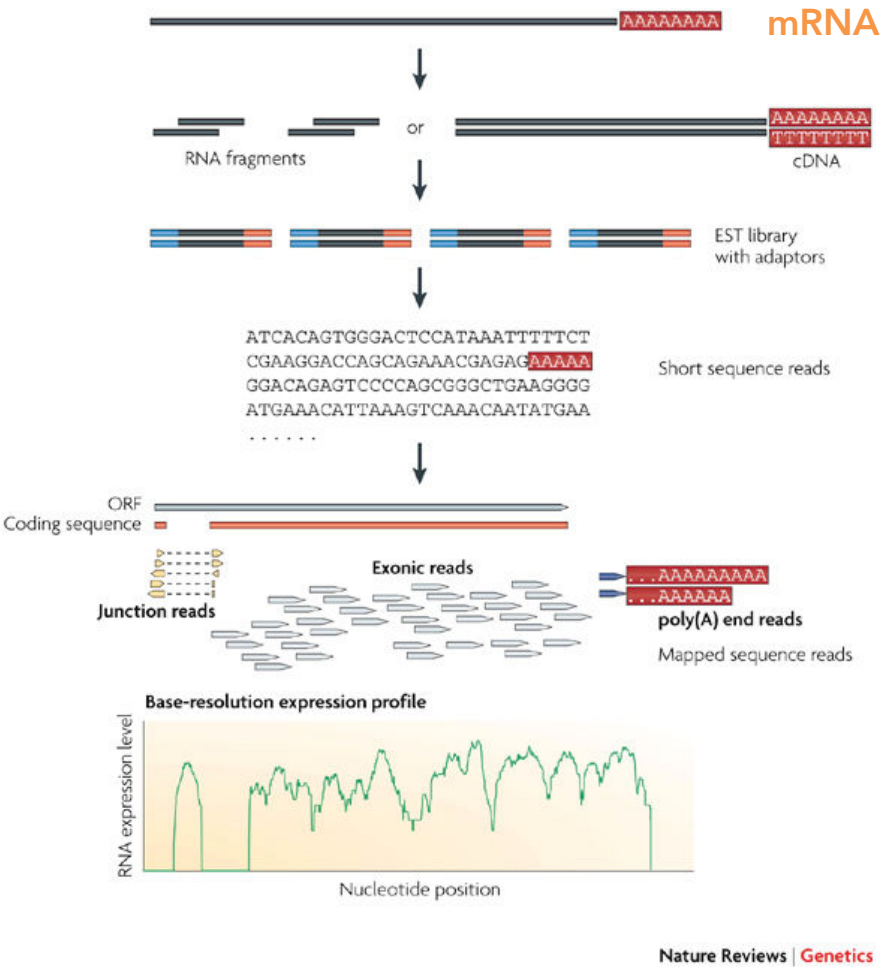
mRNA from a source

RT + + amplification

Add adapters to ends

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Typical RNA-seq workflow



mRNA from a source

RT + + amplification

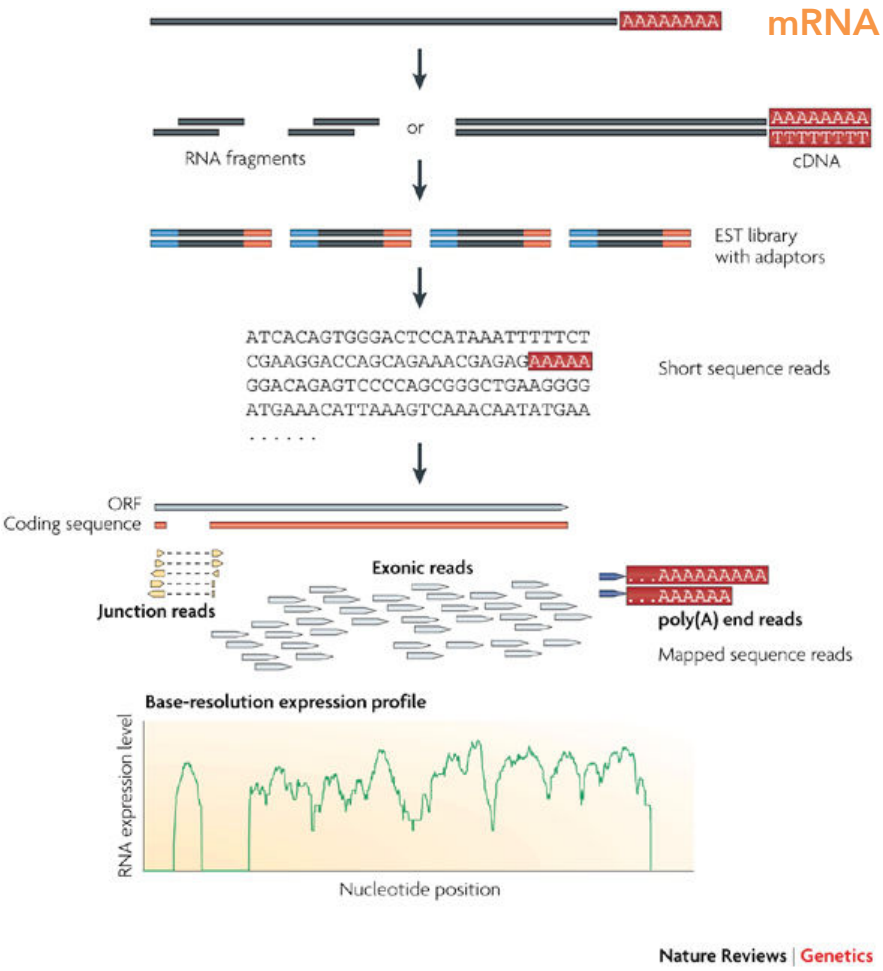
Add adapters to ends

Sequence fragments

Align reads to the reference
to quantify abundance of transcripts

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Typical RNA-seq workflow



mRNA from a source

RT + + amplification

Add adapters to ends

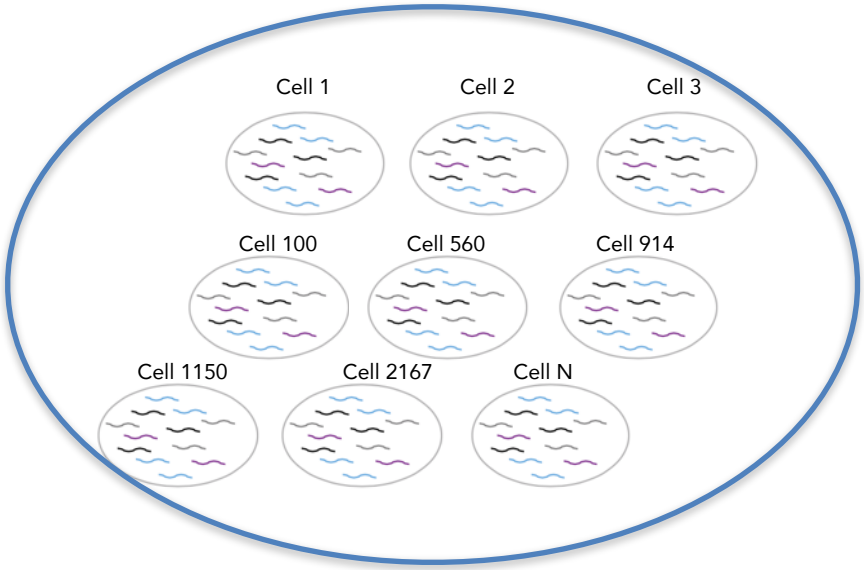
Sequence fragments

Align reads to the reference
to quantify abundance of transcripts

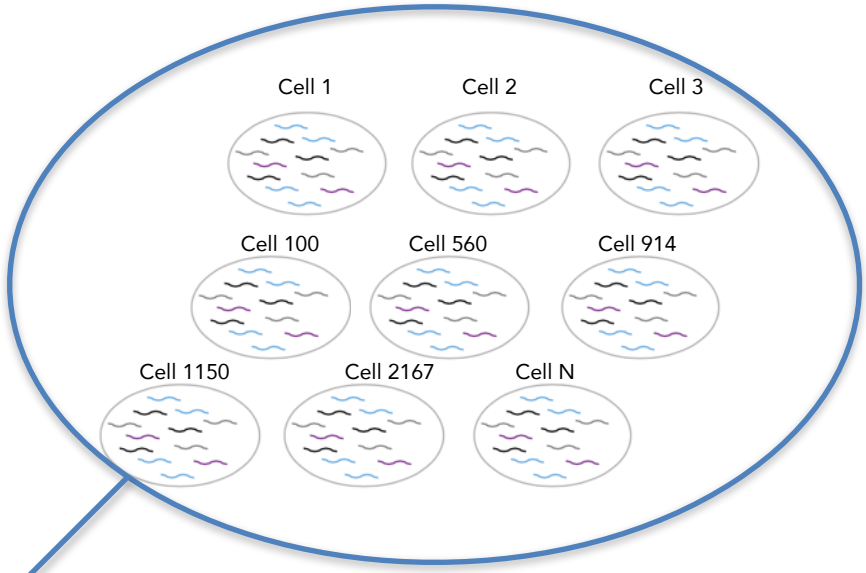
Questions : Can we do this at a single-cell level? How accurate is this measurement? How can we multiplex samples?

Wang, Gerstein, and Snyder, Nat. Rev. Genetics, 2009

Bulk vs. single cell gene expression



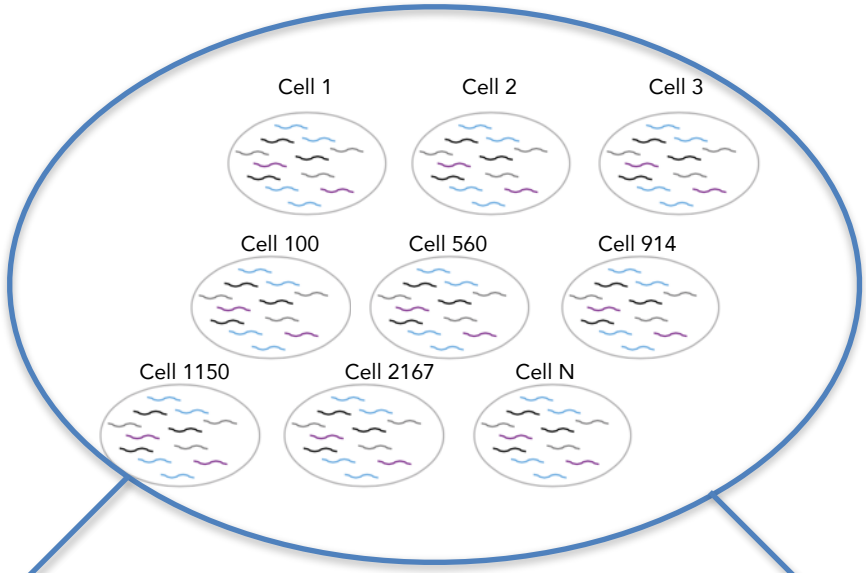
Bulk vs. single cell gene expression



	Exp
Gene 1	100
Gene 2	3.5
Gene 3	10
…	
Gene K	95
…	
Gene 25K	0.4

Population Average

Bulk vs. single cell gene expression



	Exp
Gene 1	100
Gene 2	3.5
Gene 3	10
…	
Gene K	95
…	
Gene 25K	0.4

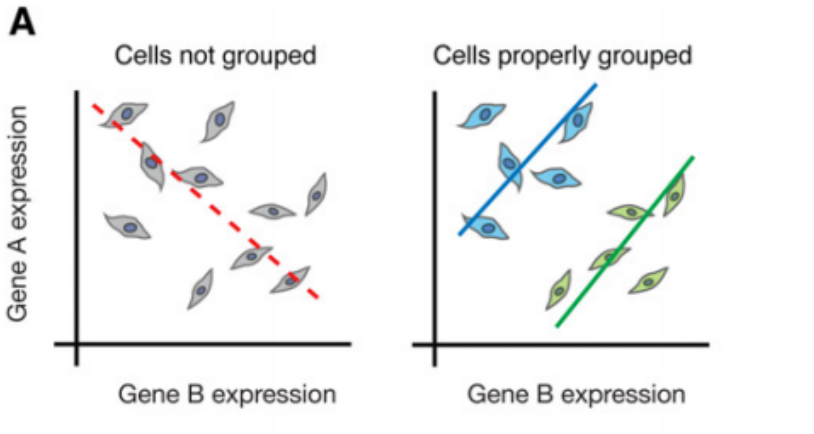
	Cell 1	Cell 2	Cell 3	….	Cell 5K
Gene 1	3	0	1		2
Gene 2	0	2	0		1
Gene 3	1	0	3		5
…					
Gene K	14	7	1		0
…					
Gene 25K	0	13	1		0

Population Average

Cellular resolution

Why do we need single cell resolution?

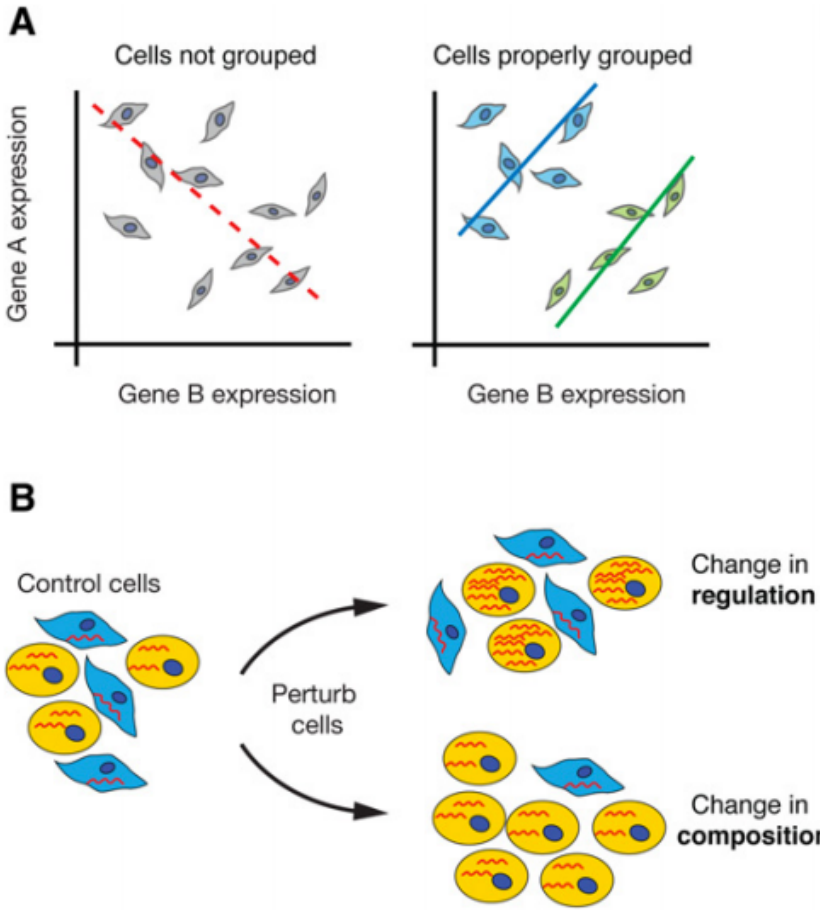
Simpson's paradox



Trapnell et al., *Genome Research*, 2016

Why do we need single cell resolution?

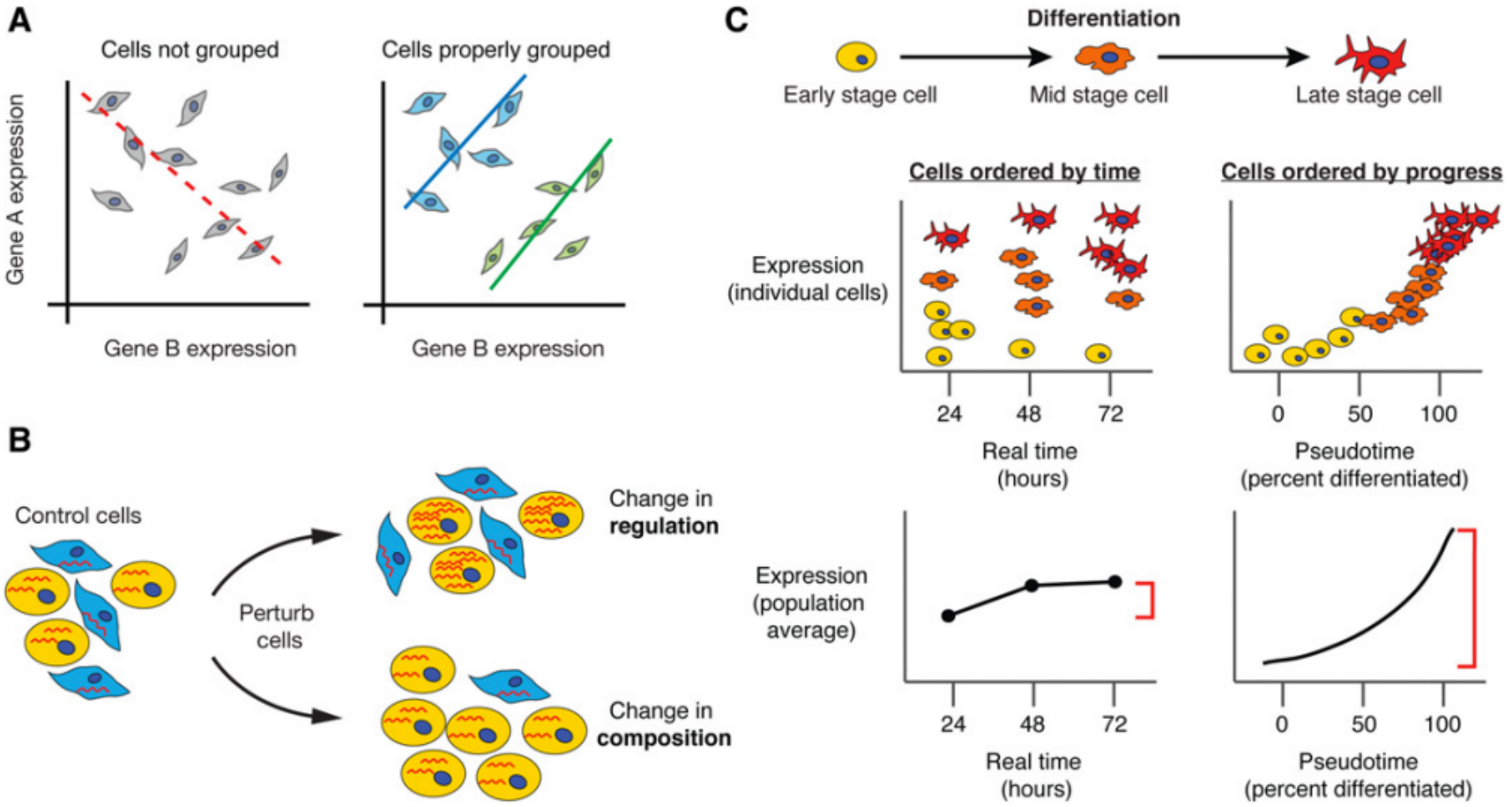
Simpson's paradox



Trapnell et al., *Genome Research*, 2016

Why do we need single cell resolution?

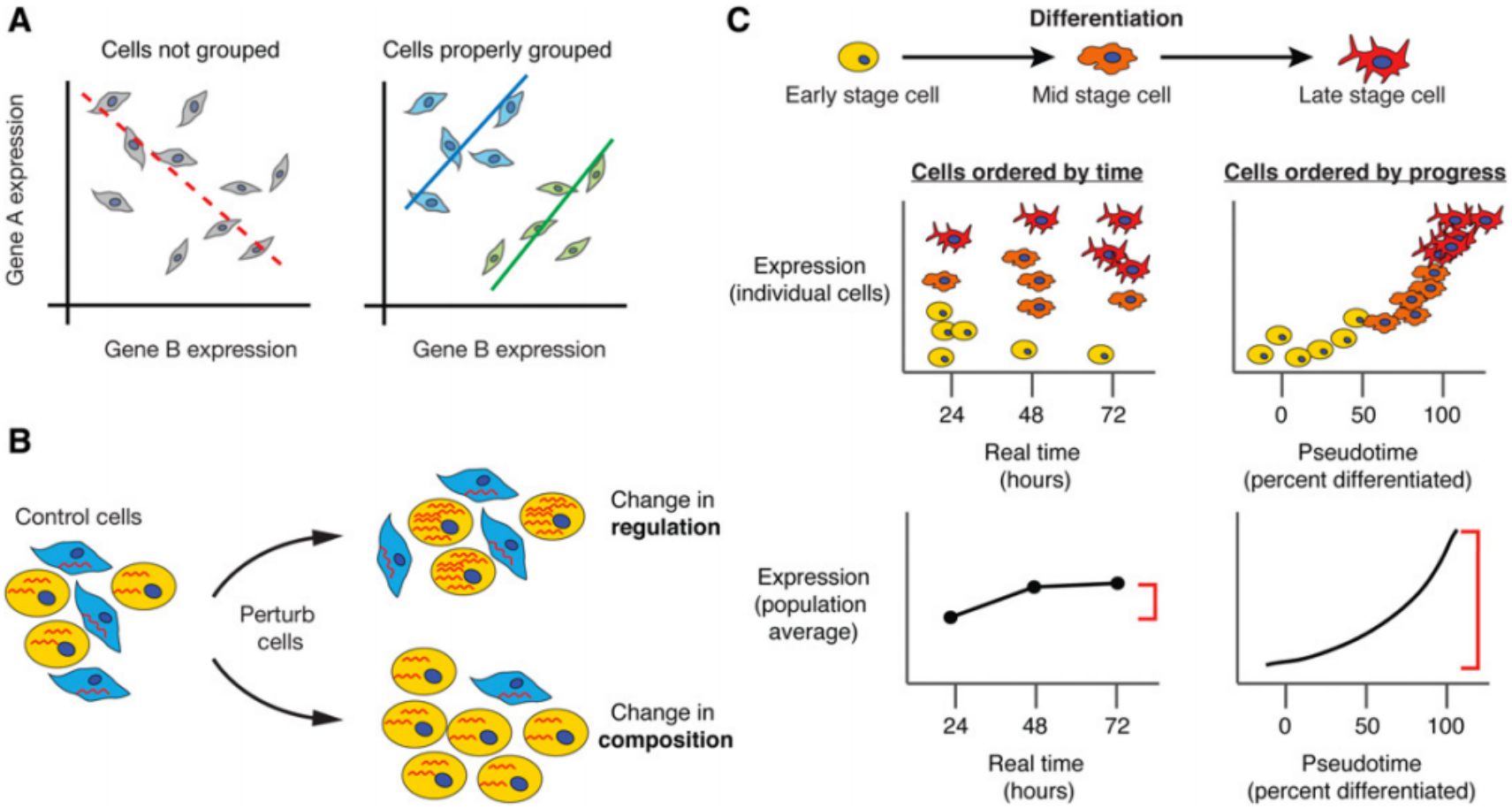
Simpson's paradox



Trapnell et al., *Genome Research*, 2016

Why do we need single cell resolution?

Simpson's paradox



Bulk expression data can be misleading as it averages across many distinct cell subpopulations!

Trapnell et al., *Genome Research*, 2016

scRNA-seq by the numbers

2013, 18 cells

LETTER

doi:10.1038/nature12172

Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells

Alex K. Shalek[1]*, Rahul Satija[2]*, Xian Adiconis[2], Rona S. Gertner[1], Jellert T. Gaublomme[1], Raktima Raychowdhury[2], Schraga Schwartz[2], Nir Yosef[2], Christine Malboeuf[2], Diana Lu[2], John J. Trombetta[2], Dave Gennert[2], Andreas Gnirke[2], Alon Goren[2,3], Nir Hacohen[2,4], Joshua Z. Levin[2], Hongkun Park[1,2] & Aviv Regev[2,5]

scRNA-seq by the numbers

2013, 18 cells

LETTER

doi:10.1038/nature12172

Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells

Alex K. Shalek[1]*, Rahul Satija[2]*, Xian Adiconis[2], Rona S. Gertner[1], Jellert T. Gaublomme[1], Raktima Raychowdhury[2], Schraga Schwartz[2], Nir Yosef[2], Christine Malboeuf[2], Diana Lu[2], John J. Trombetta[2], Dave Gennert[2], Andreas Gnirke[2], Alon Goren[2,3], Nir Hacohen[2,4], Joshua Z. Levin[2], Hongkun Park[1,2] & Aviv Regev[2,5]

2014, 1700 cells

ARTICLE

doi:10.1038/nature13437

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation

Alex K. Shalek[1,2,3]*, Rahul Satija[3]*, Joe Shuga[4]*, John J. Trombetta[3], Dave Gennert[3], Diana Lu[3], Peilin Chen[4], Rona S. Gertner[1,2], Jellert T. Gaublomme[1,2], Nir Yosef[2], Schraga Schwartz[2], Brian Fowler[4], Suzanne Weaver[4], Jing Wang[4], Xiaohui Wang[4], Ruihua Ding[1,2], Raktima Raychowdhury[3], Nir Friedman[3], Nir Hacohen[3,4], Hongkun Park[1,2,3], Andrew P. May[4] & Aviv Regev[3,7]

scRNA-seq by the numbers

2013, 18 cells

LETTER

doi:10.1038/nature12172

Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells

Alex K. Shalek[1]*, Rahul Satija[2]*, Xian Adiconis[2], Rona S. Gertner[1], Jellert T. Gaublomme[1], Raktima Raychowdhury[2], Schraga Schwartz[2], Nir Yosef[2], Christine Malboeuf[2], Diana Lu[2], John J. Trombetta[2], Dave Gennert[2], Andreas Gnirke[2], Alon Goren[2,3], Nir Hacohen[2,4], Joshua Z. Levin[2], Hongkun Park[1,2] & Aviv Regev[2,5]

2014, 1700 cells

ARTICLE

doi:10.1038/nature13437

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation

Alex K. Shalek[1,2,3]*, Rahul Satija[3]*, Joe Shuga[4]*, John J. Trombetta[3], Dave Gennert[3], Diana Lu[3], Peilin Chen[4], Rona S. Gertner[1,2], Jellert T. Gaublomme[1,2], Nir Yosef[2], Schraga Schwartz[2], Brian Fowler[4], Suzanne Weaver[4], Jing Wang[4], Xiaohui Wang[4], Ruihua Ding[1,2], Raktima Raychowdhury[2], Nir Friedman[3], Nir Hacohen[3,4], Hongkun Park[1,2,3], Andrew P. May[4] & Aviv Regev[3,7]

2015, 45,000 cells

Cell

Resource

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,7] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[1] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[1] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[3] John J. Trombetta,[4] David A. Weitz,[5,10] Joshua R. Sanes,[3] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

scRNA-seq by the numbers

2013, 18 cells

LETTER

doi:10.1038/nature12172

Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells

Alex K. Shalek[1]*, Rahul Satija[2]*, Xian Adiconis[2], Rona S. Gertner[1], Jellert T. Gaublomme[1], Raktima Raychowdhury[2], Schraga Schwartz[2], Nir Yosef[2], Christine Malboeuf[2], Diana Lu[2], John J. Trombetta[2], Dave Gennert[2], Andreas Gnirke[2], Alon Goren[2,3], Nir Hacohen[2,4], Joshua Z. Levin[2], Hongkun Park[1,2] & Aviv Regev[2,5]

2014, 1700 cells

ARTICLE

doi:10.1038/nature13437

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation

Alex K. Shalek[1,2,3]*, Rahul Satija[3]*, Joe Shuga[4]*, John J. Trombetta[3], Dave Gennert[3], Diana Lu[3], Peilin Chen[4], Rona S. Gertner[1,2], Jellert T. Gaublomme[1,2], Nir Yosef[2], Schraga Schwartz[2], Brian Fowler[4], Suzanne Weaver[4], Jing Wang[4], Xiaohui Wang[4], Ruihua Ding[1,2], Raktima Raychowdhury[3], Nir Friedman[3], Nir Hacohen[3,4], Hongkun Park[1,2,3], Andrew P. May[4] & Aviv Regev[3,7]

2015, 45,000 cells

Cell

Resource

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,*] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[1] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[1] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[3] John J. Trombetta,[3] David A. Weitz,[5,10] Joshua R. Sanes,[9] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

2016, 200,000 cells

Resource

Cell

Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens

Atray Dixit,[1,2,9] Oren Parnas,[1,9,10] Biyu Li,[1] Jenny Chen,[1,2] Charles P. Fulco,[1,4] Livnat Jerby-Arnon,[1] Nemanja D. Marjanovic,[1,3] Danielle Dionne,[1] Tyler Burks,[1] Raktima Raychowdhury,[1] Britt Adamson,[5] Thomas M. Norman,[5] Eric S. Lander,[1,4,8] Jonathan S. Weissman,[2,7] Nir Friedman,[1,8] and Aviv Regev[1,6,7,11,*]

scRNA-seq by the numbers

2013, 18 cells

LETTER

doi:10.1038/nature12172

Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells

Alex K. Shalek[1]*, Rahul Satija[2]*, Xian Adiconis[2], Rona S. Gertner[1], Jellert T. Gaublomme[1], Raktima Raychowdhury[2], Schraga Schwartz[2], Nir Yosef[2], Christine Malboeuf[2], Diana Lu[2], John J. Trombetta[2], Dave Gennert[2], Andreas Gnirke[2], Alon Goren[2,3], Nir Hacohen[2,4], Joshua Z. Levin[2], Hongkun Park[1,2] & Aviv Regev[2,5]

2014, 1700 cells

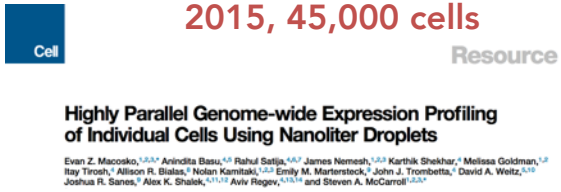
ARTICLE

doi:10.1038/nature13437

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation

Alex K. Shalek[1,2,3]*, Rahul Satija[3]*, Joe Shuga[4]*, John J. Trombetta[3], Dave Gennert[3], Diana Lu[3], Peilin Chen[4], Rona S. Gertner[1,2], Jellert T. Gaublomme[1,2], Nir Yosef[2], Schraga Schwartz[2], Brian Fowler[4], Suzanne Weaver[4], Jing Wang[4], Xiaohui Wang[4], Ruihua Ding[1,2], Raktima Raychowdhury[3], Nir Friedman[3], Nir Hacohen[3,5], Hongkun Park[1,2,3], Andrew P. May[4] & Aviv Regev[3,7]

2015, 45,000 cells



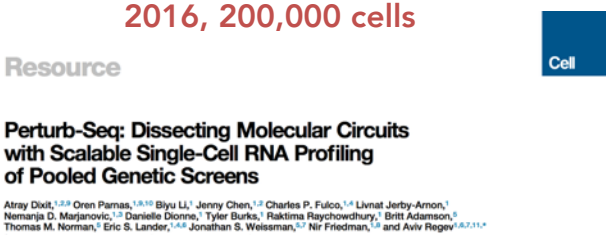
Cell

Resource

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,*] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[1] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[1] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[3] John J. Trombetta,[3] David A. Weitz,[5,10] Joshua R. Sanes,[3] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

2016, 200,000 cells



Resource

Cell

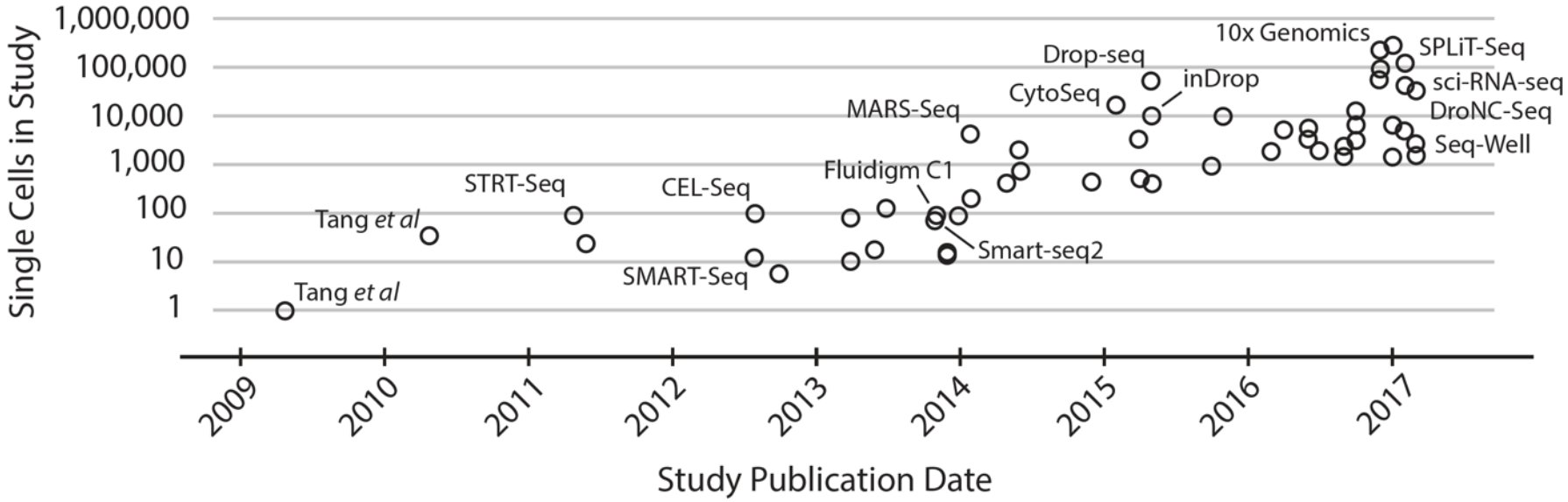
Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens

Atray Dixit,[1,2,9] Oren Parnas,[1,9,10] Biyu Li,[1] Jenny Chen,[1,2] Charles P. Fulco,[1,4] Livnat Jerby-Arnon,[1] Nemanja D. Marjanovic,[1,3] Danielle Dionne,[1] Tyler Burks,[1] Raktima Raychowdhury,[1] Britt Adamson,[5] Thomas M. Norman,[5] Eric S. Lander,[1,4,8] Jonathan S. Weissman,[5,7] Nir Friedman,[1,8] and Aviv Regev[1,6,7,11,*]

2017, 1.3 million cells (10X genomics)

Vastly improved protocols have made it possible to increase scRNA-seq cell throughput

Note : scRNA-seq is not "one" method. It refers to many protocols that differ in their strengths and limitations

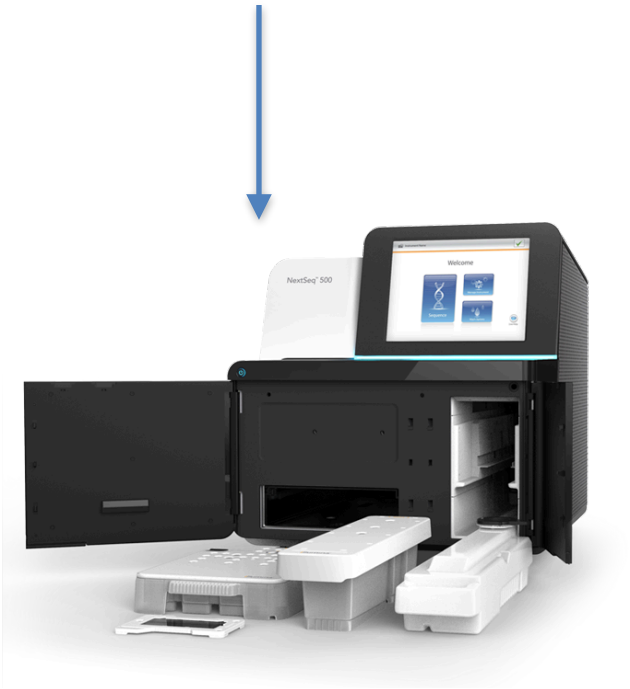


Svensson et al., 2017

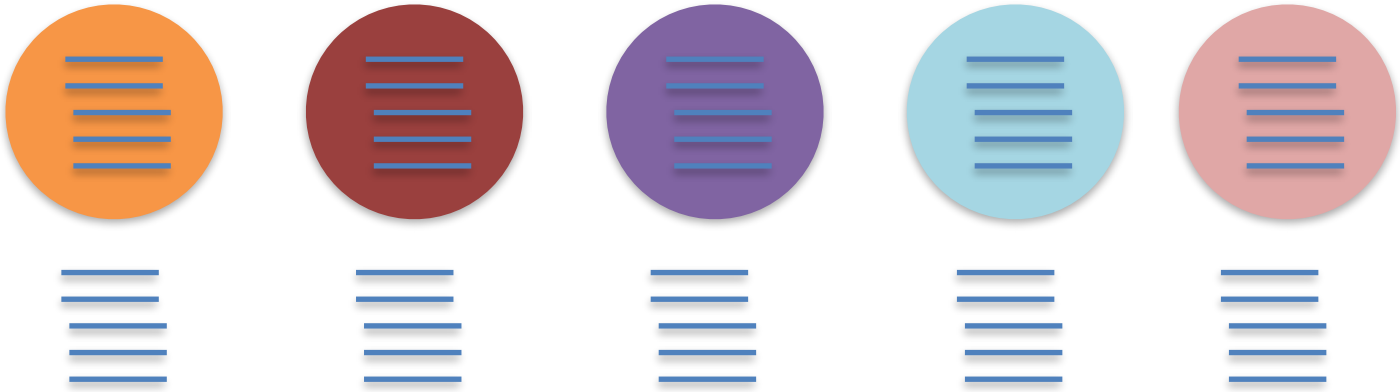
Agenda

- Single cell analysis - why?
- **A short survey of scRNA-seq methods**
- Quality comparison of different methods and power analysis
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R
- Some example applications
- Future

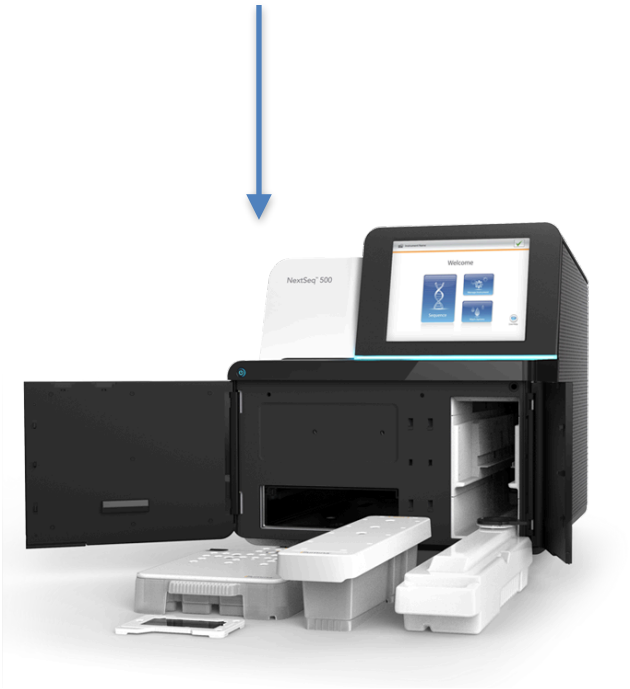
Tracking the cell-of-origin of individual transcripts



Tracking the cell-of-origin of individual transcripts



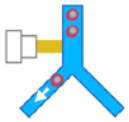
Tracking the cell-of-origin of individual transcripts



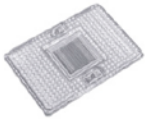
Single-cell RNA-seq pipeline

Single-cell RNA-seq pipeline

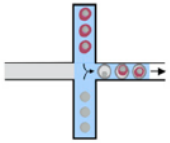
FACS



Micro-capture

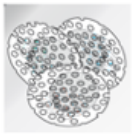


Droplet



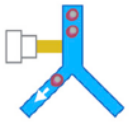
Nanowell

Split/pool barcoding

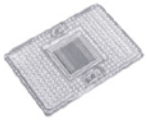


Single-cell RNA-seq pipeline

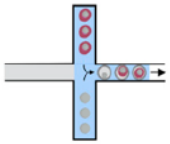
FACS



Micro-capture

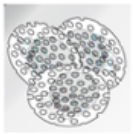


Droplet



Nanowell

Split/pool barcoding



PolyA vs random priming

3′/5′ end tagging vs full-length

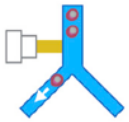
PCR vs *in vitro* transcription

…

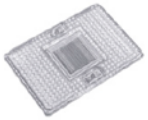
Single-cell RNA-seq pipeline

Cell capture and Barcoding → Library Prep → Sequencing

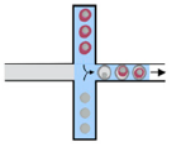
FACS



Micro-capture

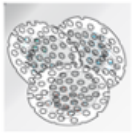


Droplet



Nanowell

Split/pool barcoding



PolyA vs random priming

3′/5′ end tagging vs full-length

PCR vs *in vitro* transcription

…

Miseq

~ 20M reads total

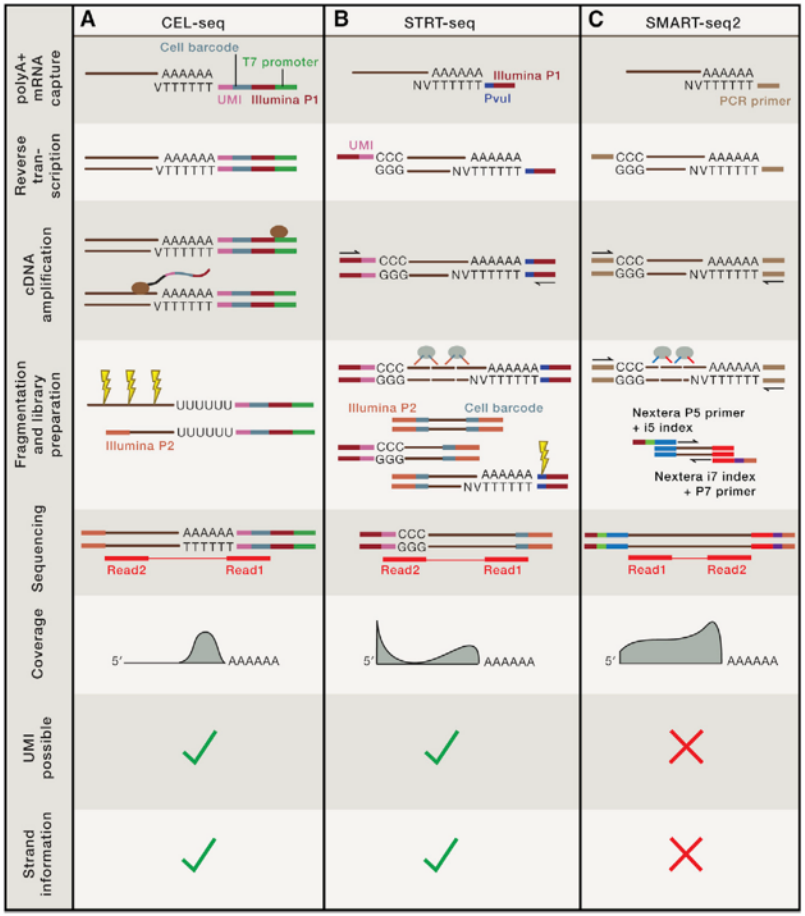
Nextseq

~ 500M reads total

HiSeq 4000

4 billion reads

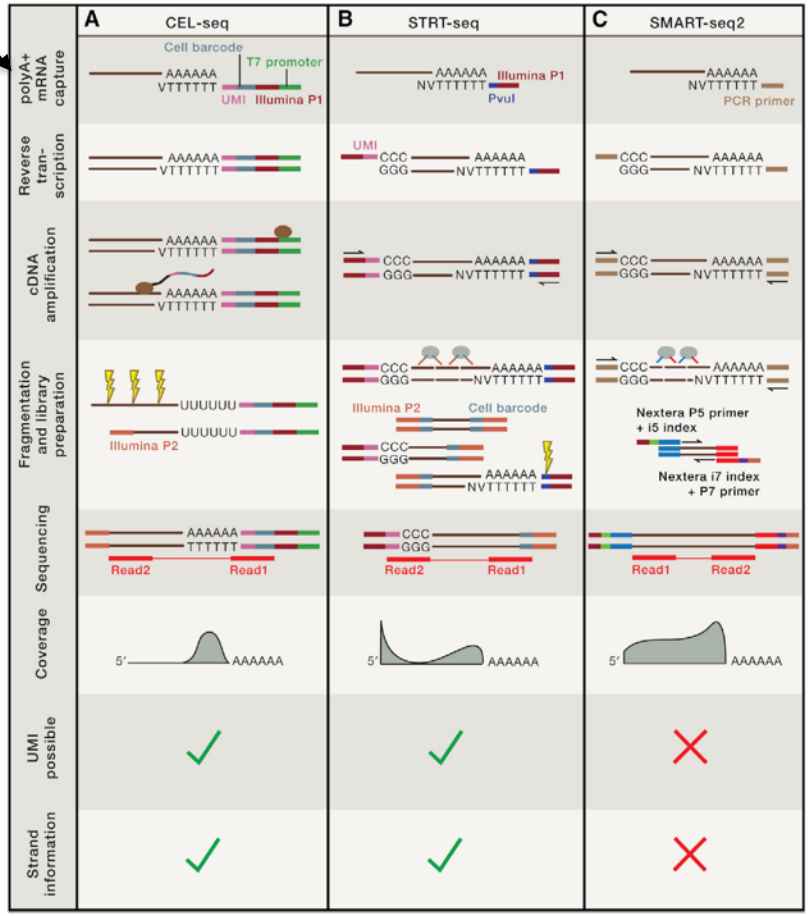
Three commonly used scRNA-seq library prep methods



Grun and van oudenaarden, *Cell, 2015*

Three commonly used scRNA-seq library prep methods

in vitro transcription,
3′ end tagging

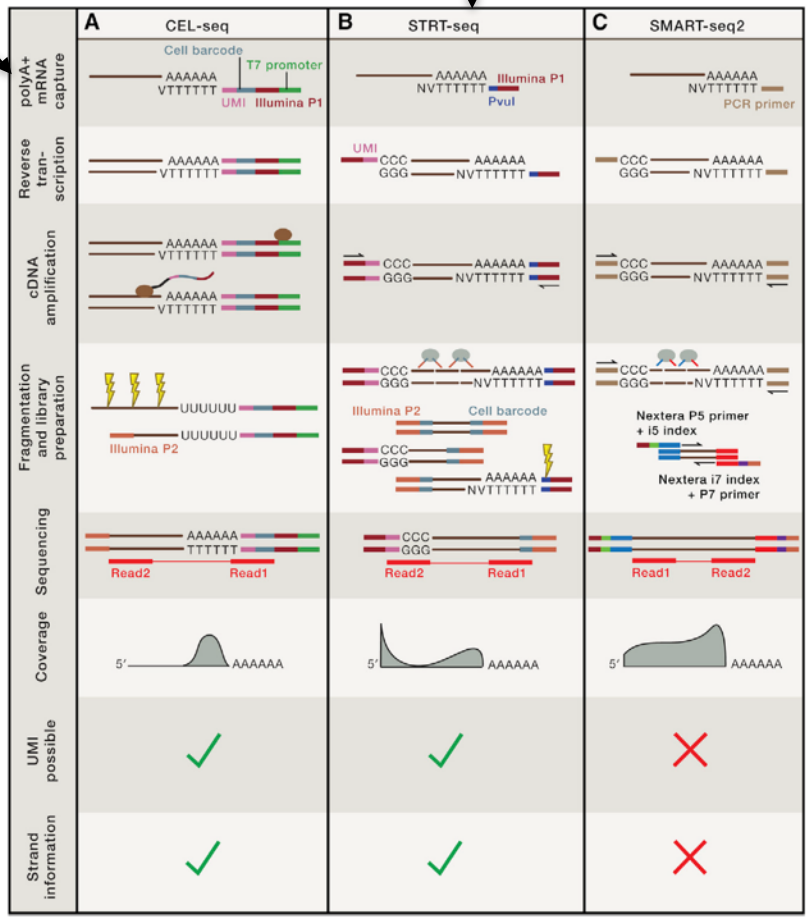


Grun and van oudenaarden, *Cell, 2015*

Three commonly used scRNA-seq library prep methods

in vitro transcription,
3′ end tagging

5′ end tagging



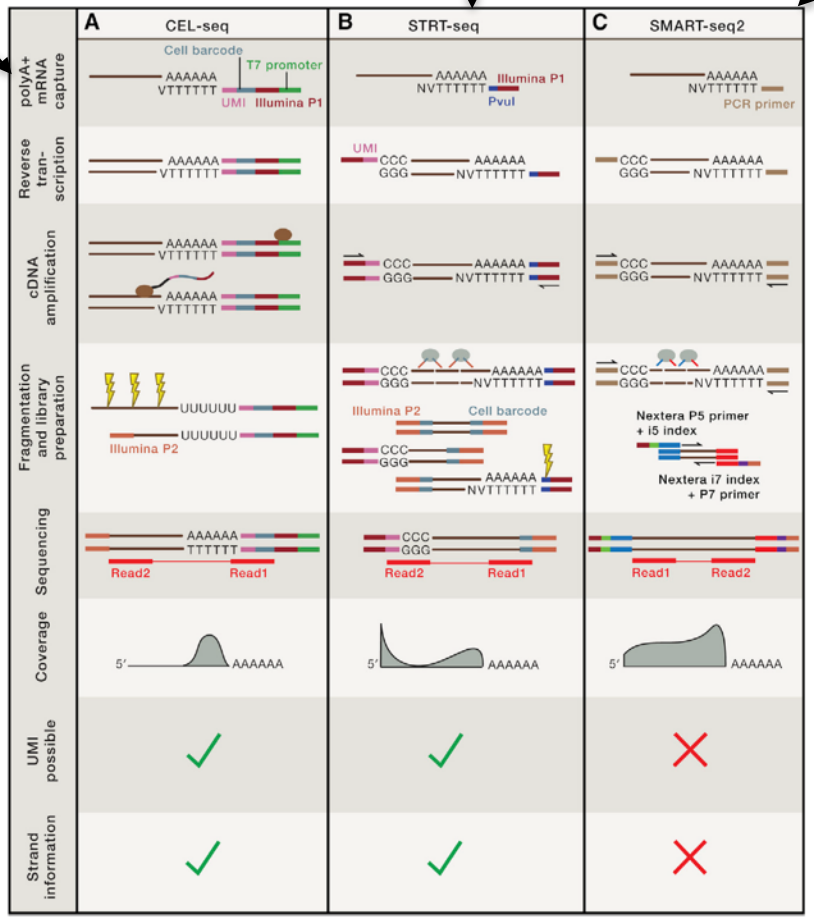
Grun and van oudenaarden, *Cell, 2015*

Three commonly used scRNA-seq library prep methods

in vitro transcription,
3′ end tagging

5′ end tagging

Full length coverage



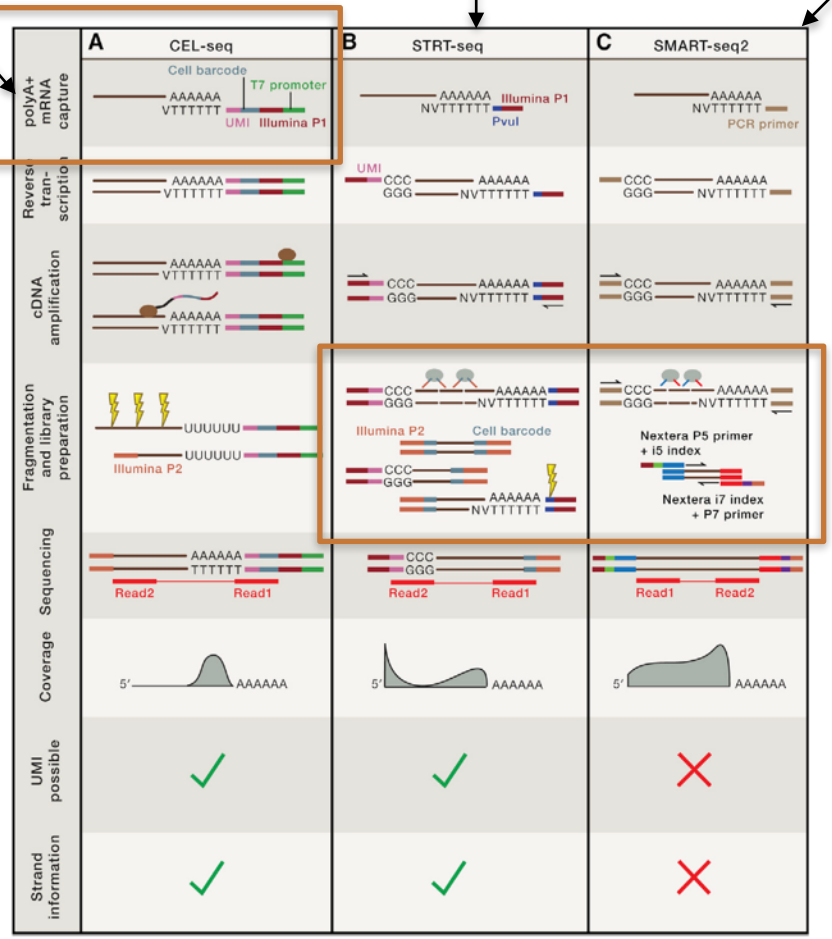
Grun and van oudenaarden, *Cell, 2015*

Three commonly used scRNA-seq library prep methods

in vitro transcription,
3′ end tagging

5′ end tagging

Full length coverage



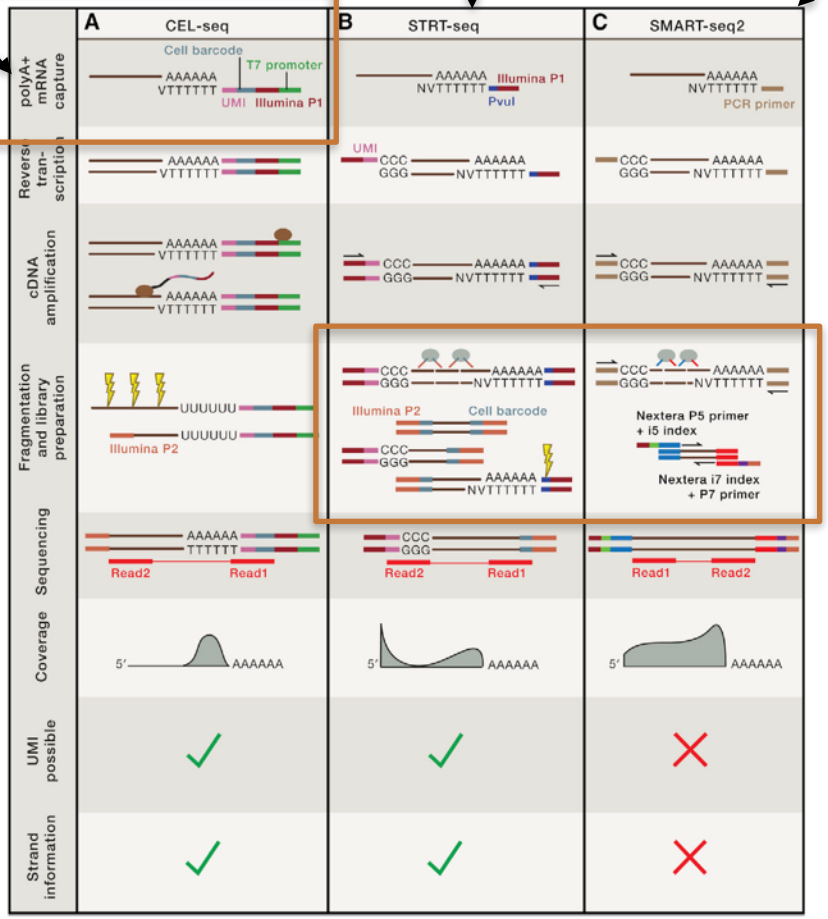
Grun and van oudenaarden, *Cell, 2015*

Three commonly used scRNA-seq library prep methods

in vitro transcription, 3′ end tagging

5′ end tagging

Full length coverage



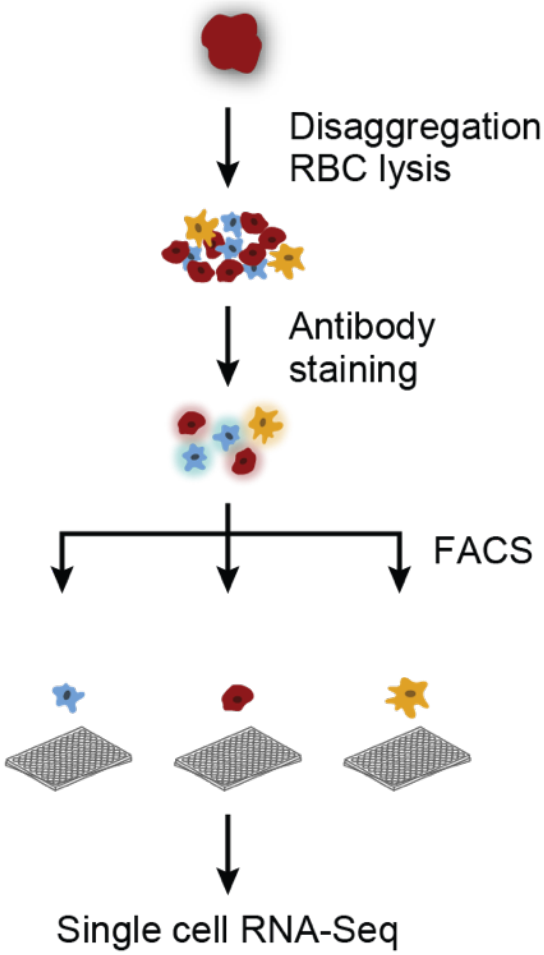
Each protocol has **advantages** and **limitations**. What one ends up using is often dictated by multiple features - the **biological context**, **cost**, **objective** etc.

Grun and van oudenaarden, *Cell, 2015*

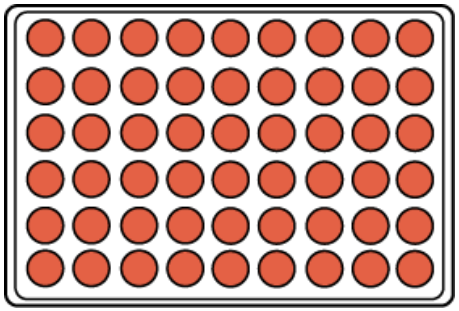
A survey of popular scRNA-seq platforms

- Plate-based RNA-seq (2011-)
- Droplet/nanowell-based RNA-seq (2015-)
- Combinatorial Indexing (2017-)

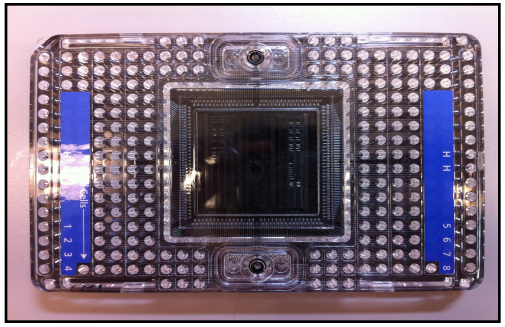
1. Plate-based scRNA-seq



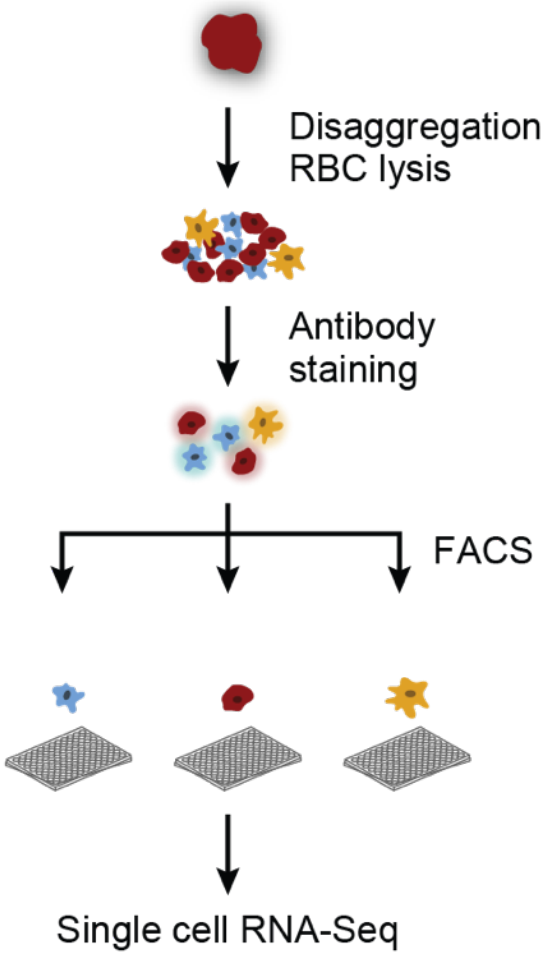
FACS sorting on 96/384-well plates



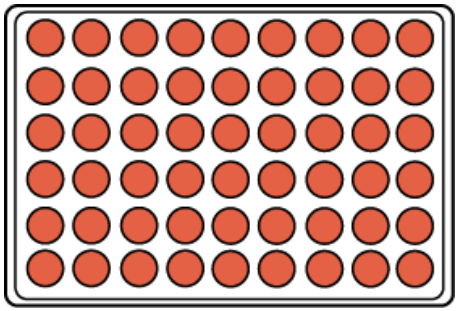
Fluidigm C1-autoprep system



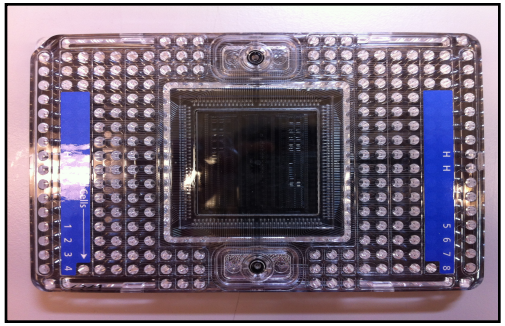
1. Plate-based scRNA-seq



FACS sorting on 96/384-well plates



Fluidigm C1-autoprep system

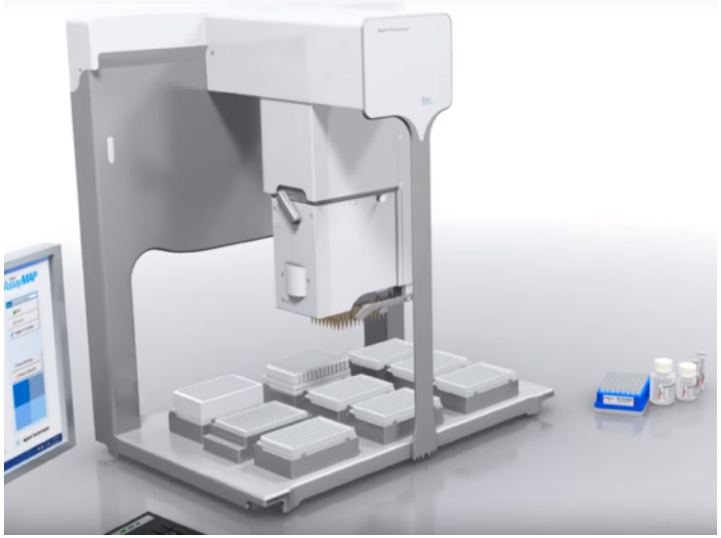


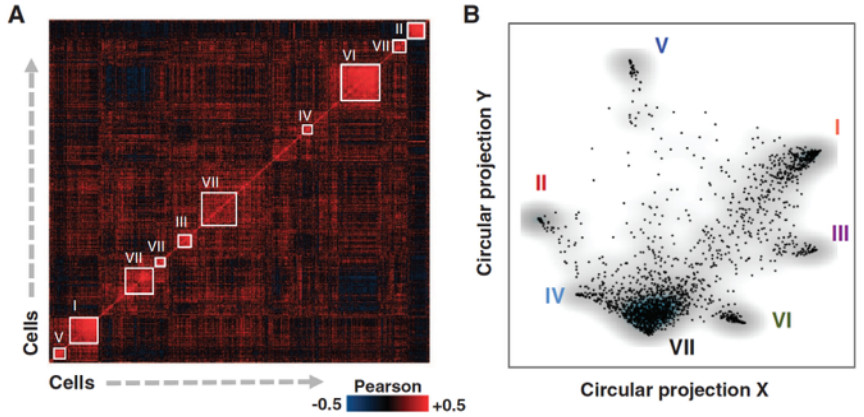
Some advantages like index sorting, and retrieval

But labor intensive, slow, and costly (~$12/cell)

Automation is possible through the use of robots

MARS-Seq adapted the CEL-Seq protocol with a liquid-handling robot to generate the first high-throughput scRNA-seq dataset of immune cells

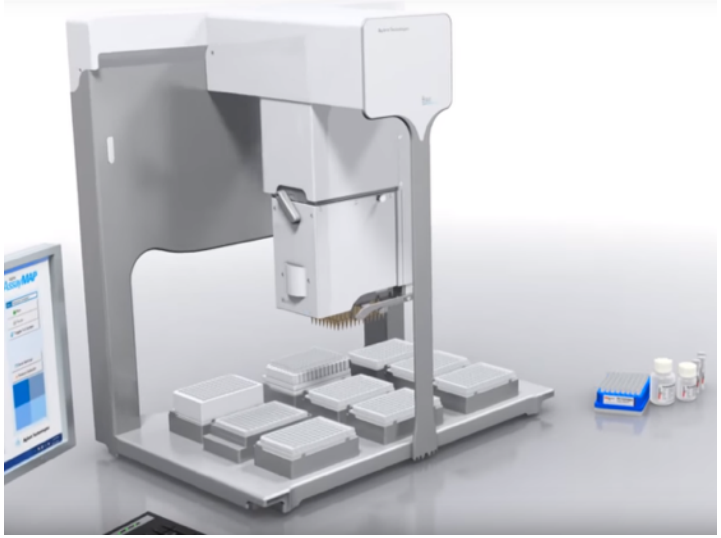


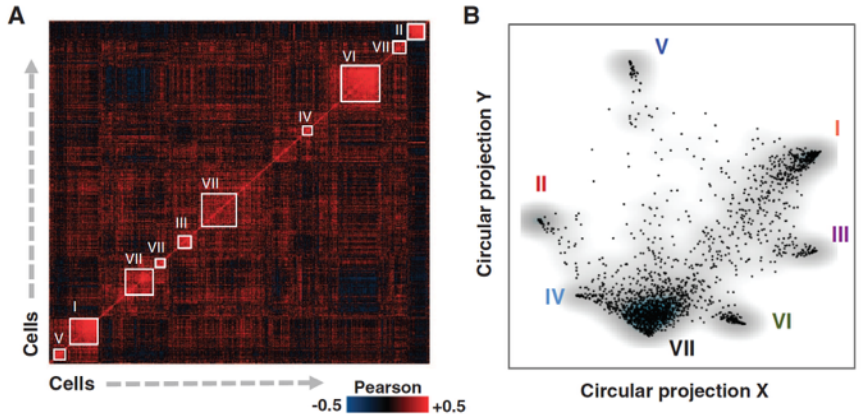


Jaitin et al., Science, 2014

Automation is possible through the use of robots

MARS-Seq adapted the CEL-Seq protocol with a liquid-handling robot to generate the first high-throughput scRNA-seq dataset of immune cells



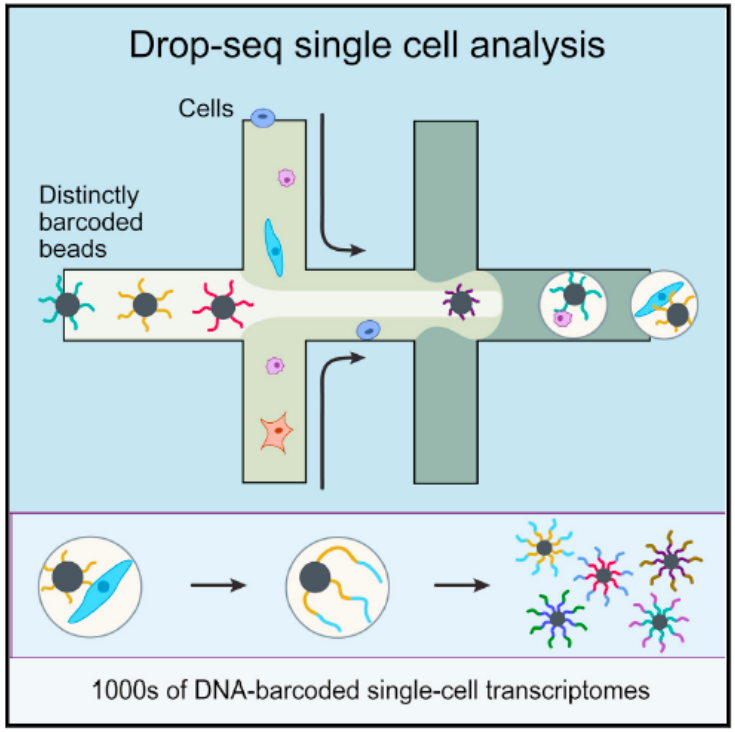


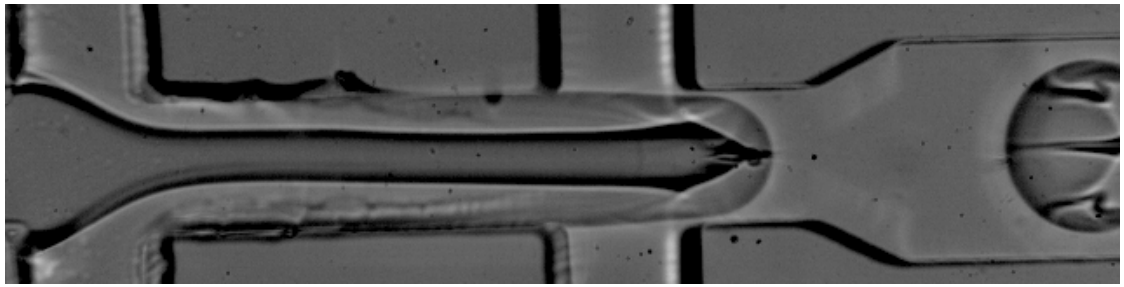
Reduction in costs : $12 / cell to $4 / cell

Jaitin et al., Science, 2014

2. Drop-seq: cells in drops with beads

A method for transcriptomic profiling (RNAseq) of thousands of individual cells at high speed and low cost (< **10 cents/cell**)

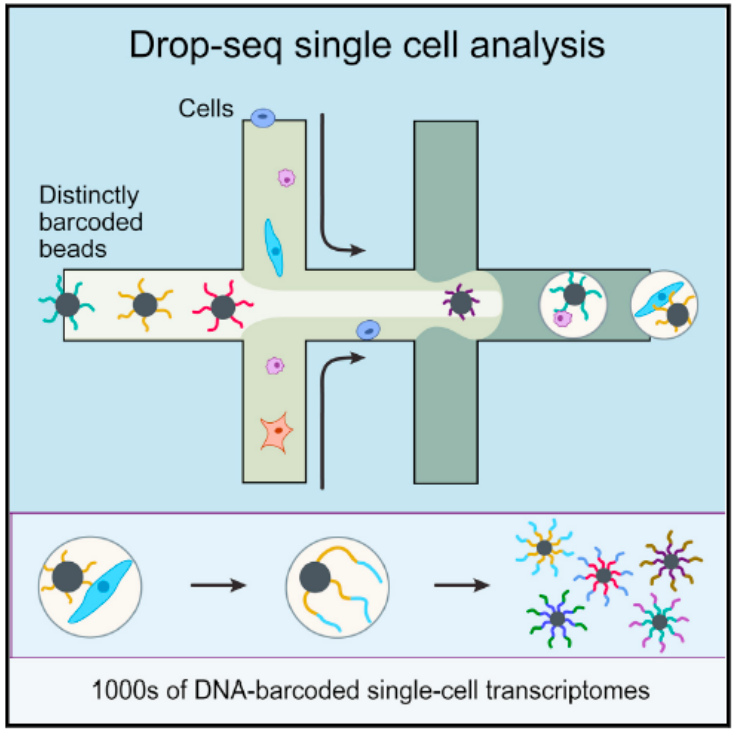


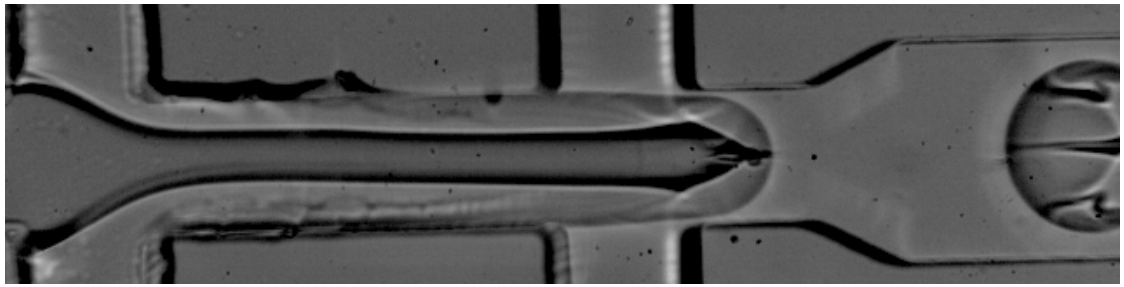


Macosko et al., 2015
inDrop - Klein et al., 2015
10X GemCode - Zheng et al., 2016

2. Drop-seq: cells in drops with beads

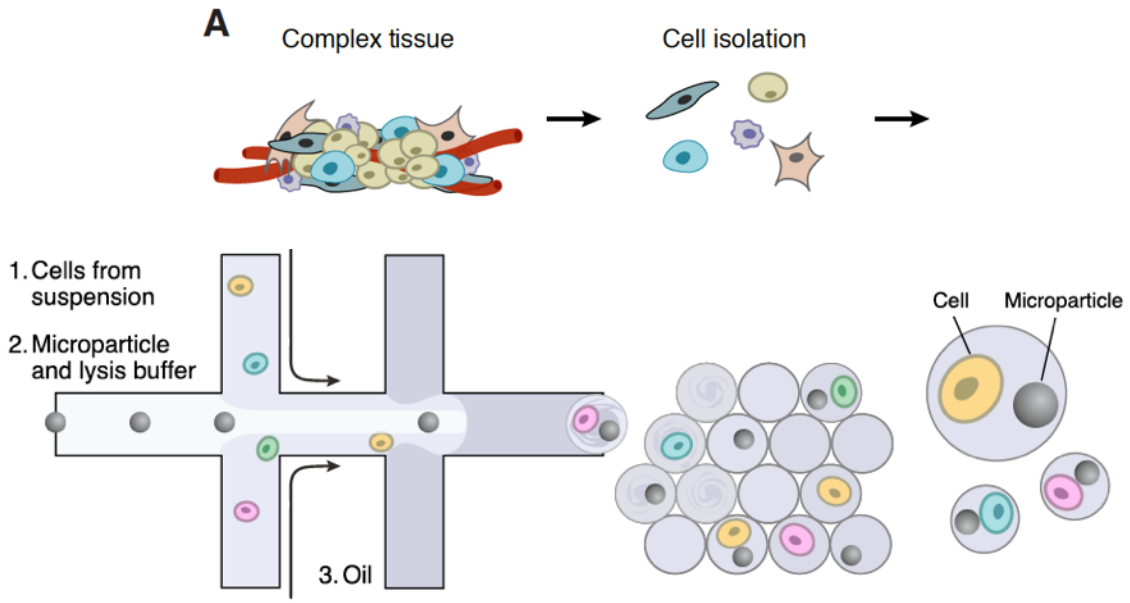
A method for transcriptomic profiling (RNAseq) of thousands of individual cells at high speed and low cost (< **10 cents/cell**)





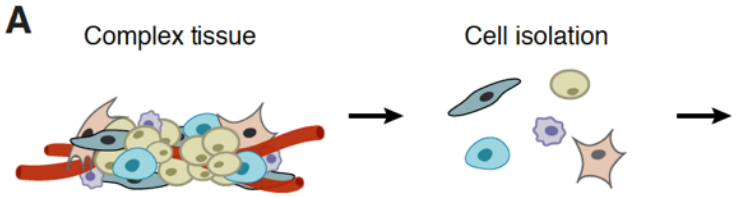
Macosko et al., 2015
inDrop - Klein et al., 2015
10X GemCode - Zheng et al., 2016

Drop-seq: Microfluidics



- Dissociate cells
- Pass cells and beads through a microfluidic device that encapsulates them in 1 nl oil droplets
- Droplets vastly outnumber beads or cells, so few droplets contain more than one cell or more than one bead (**double poisson loading**)
- Droplets form at a rate of ~100,000 per minute, or ~500 cell-bead pairs per minute

Drop-seq: Microfluidics

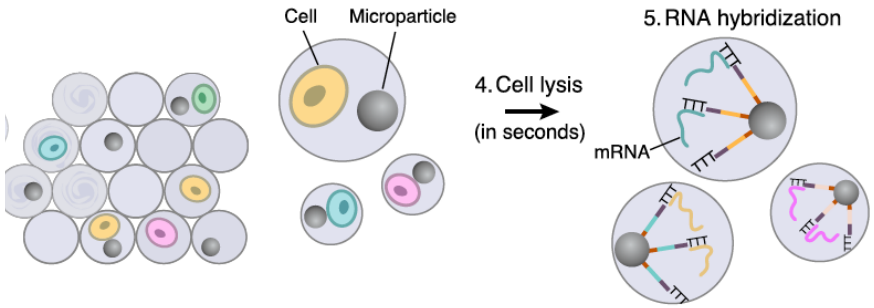


A Big Disadvantage!

Because of the double-Poisson loading to avoid 2 cells in the same droplet (doublets), <10% of the loaded cells are captured into libraries

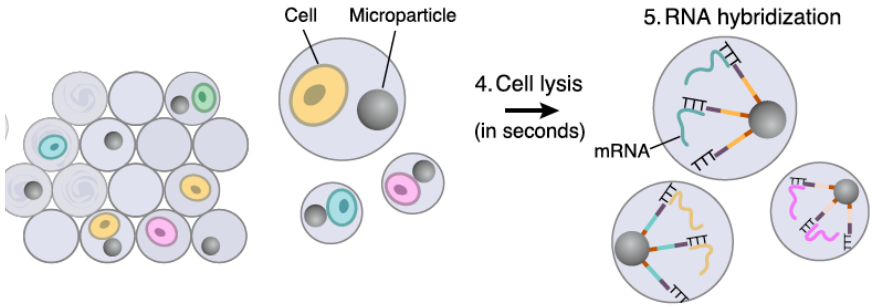
- Dissociate cells
- Pass cells and beads through a microfluidic device that encapsulates them in 1 nl oil droplets
- Droplets vastly outnumber beads or cells, so few droplets contain more than one cell or more than one bead (**double poisson loading**)
- Droplets form at a rate of ~100,000 per minute, or ~500 cell-bead pairs per minute

Drop-seq: Transcriptomics

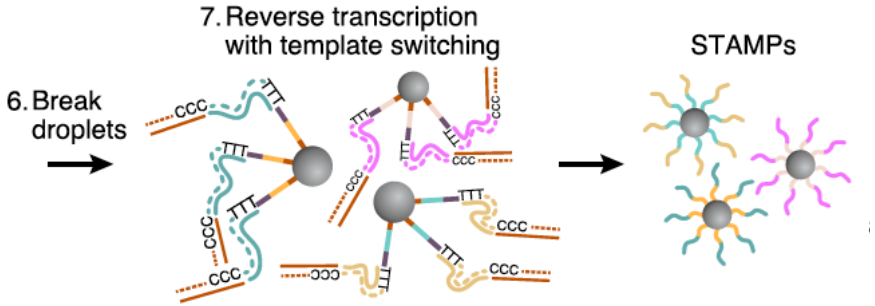


- The beads have oligo-dT containing nucleotides that capture mRNA when the cells lyse.

Drop-seq: Transcriptomics

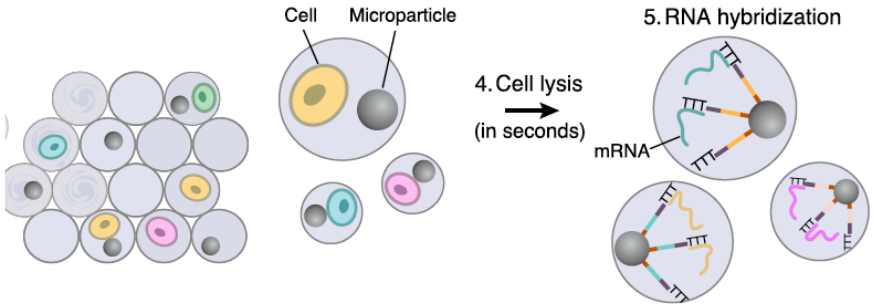


- The beads have oligo-dT containing nucleotides that capture mRNA when the cells lyse.

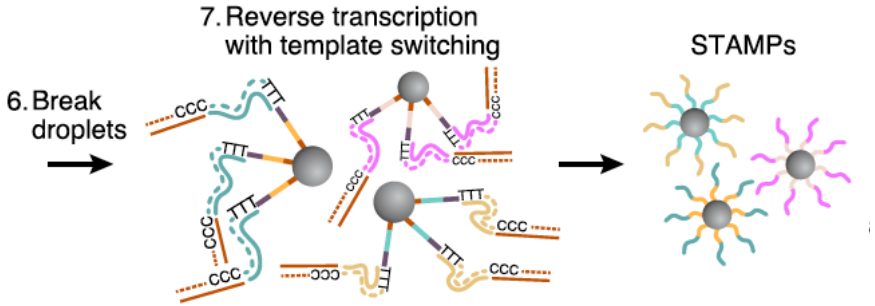


- The emulsion is broken, and cDNAs are generated
- The cDNAs are then amplified by PCR and sequenced in bulk.

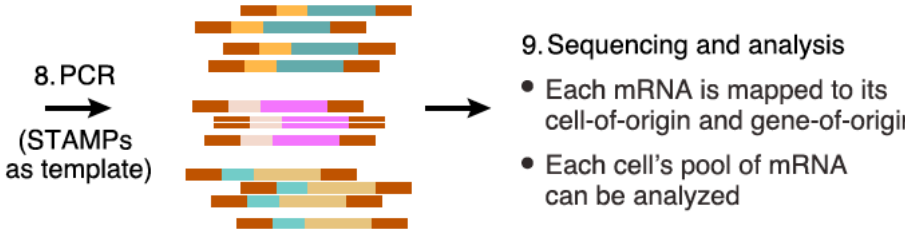
Drop-seq: Transcriptomics



- The beads have oligo-dT containing nucleotides that capture mRNA when the cells lyse.

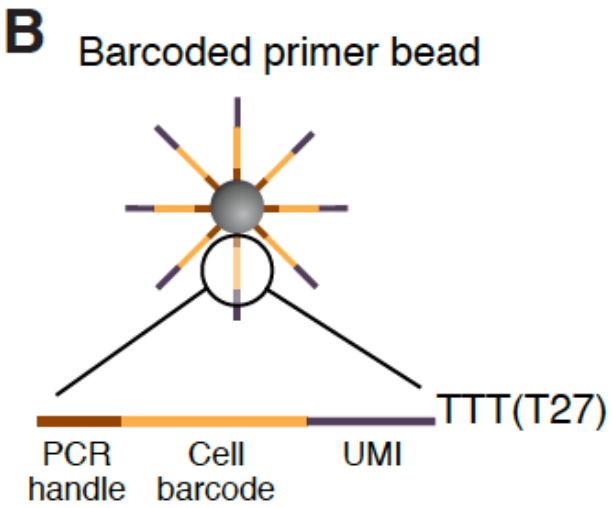


- The emulsion is broken, and cDNAs are generated
- The cDNAs are then amplified by PCR and sequenced in bulk.



- Library preparation cost: ~6 cents per cell, ~500X lower than other single cell methods available then

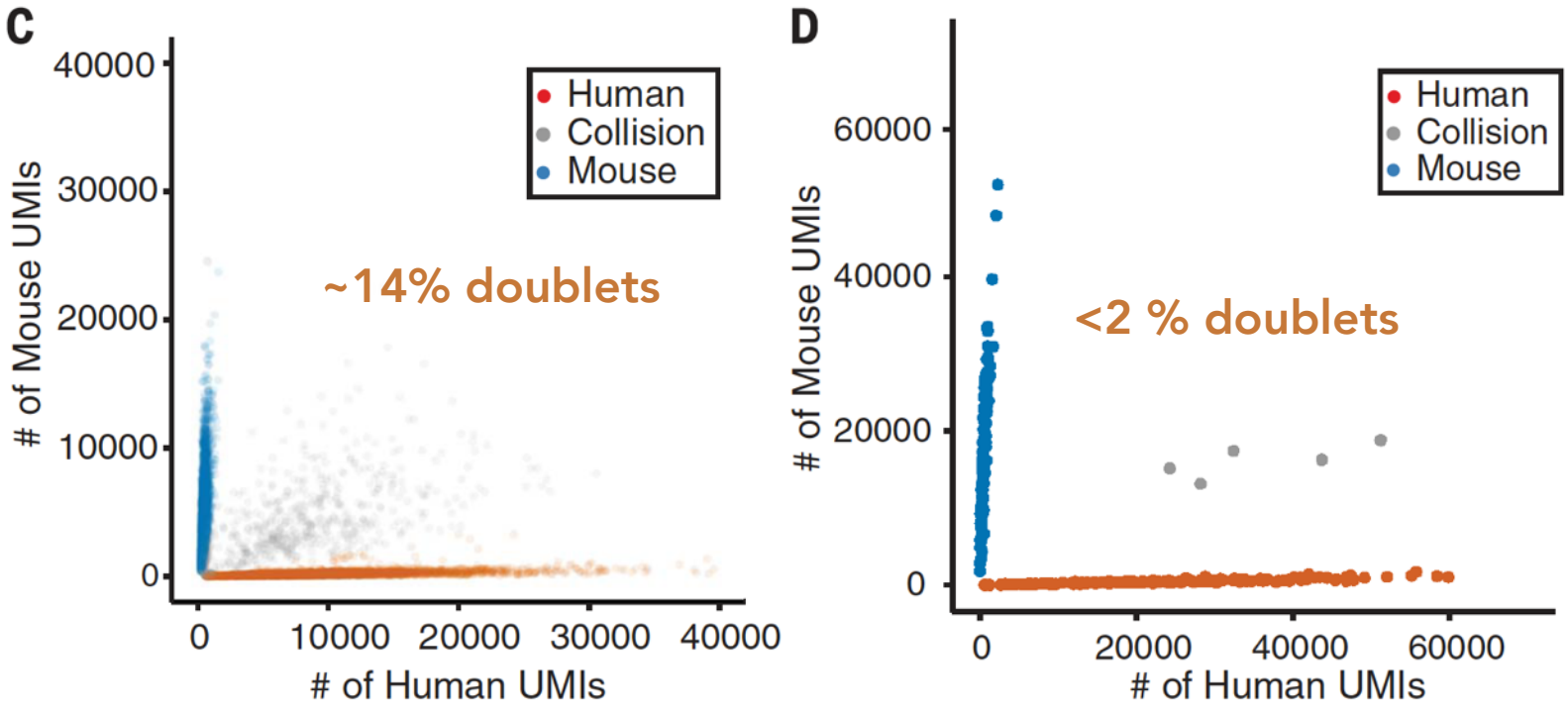
Drop-seq: The beads



- The oligonucleotides on the bead have "bar codes" that are amplified and sequenced along with the transcript.
- The "cell barcode" is a 12bp sequence that is the same for all oligos on each bead but different from those on other beads (4^12=10^6). It allows transcripts to be assigned to their cell of origin
- The "unique molecular identifier" (UMI) is an 8bp random sequence that gives each transcript from a given cell a unique identity (4^8=65K). Thus, even though PCR leads to biased amplification, each transcript can be counted only once.

Doublets

Because of the setup, it is possible that two or more cells can enter the same droplet. Studies estimate doublet frequency through a “mixed-species” experiment

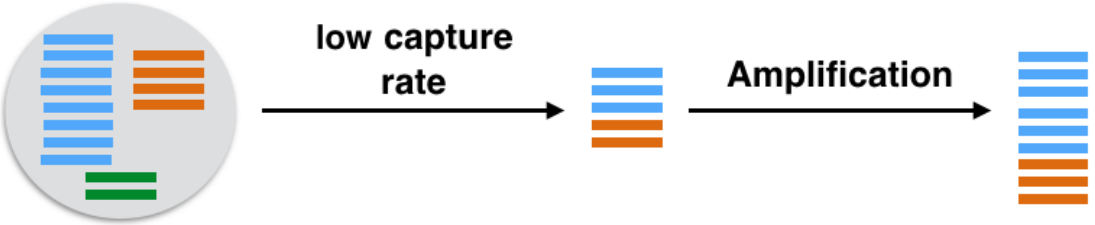


The doublet frequency is +vely correlated with throughput

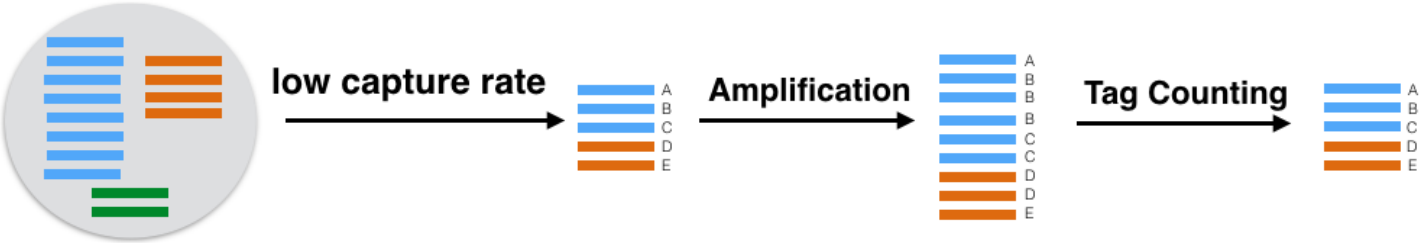
UMI : Unique Molecular Identifiers (or Random Molecular Tags)

Early labeling of mRNA molecules with random nucleotide tags enables amplification biases to be corrected

- Low input amount -> transcript dropout + PCR amplification bias



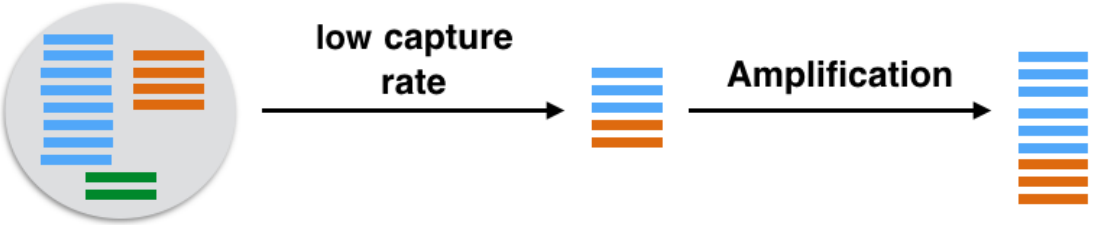
- Unique Molecular Identifiers (UMIs) can correct for PCR bias



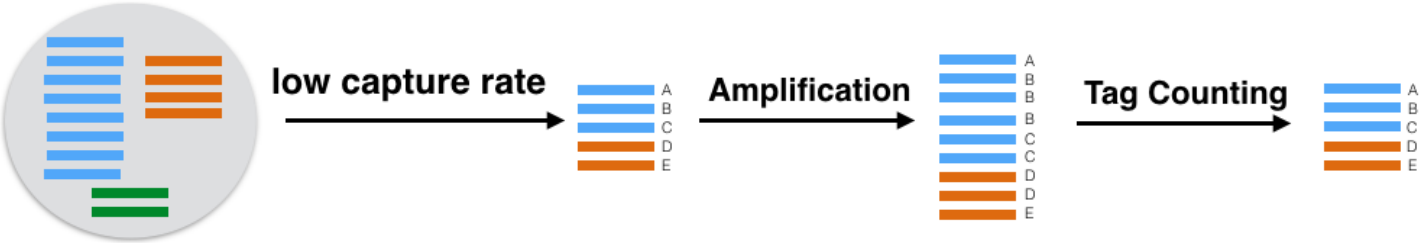
UMI : Unique Molecular Identifiers (or Random Molecular Tags)

Early labeling of mRNA molecules with random nucleotide tags enables amplification biases to be corrected

- Low input amount -> transcript dropout + PCR amplification bias

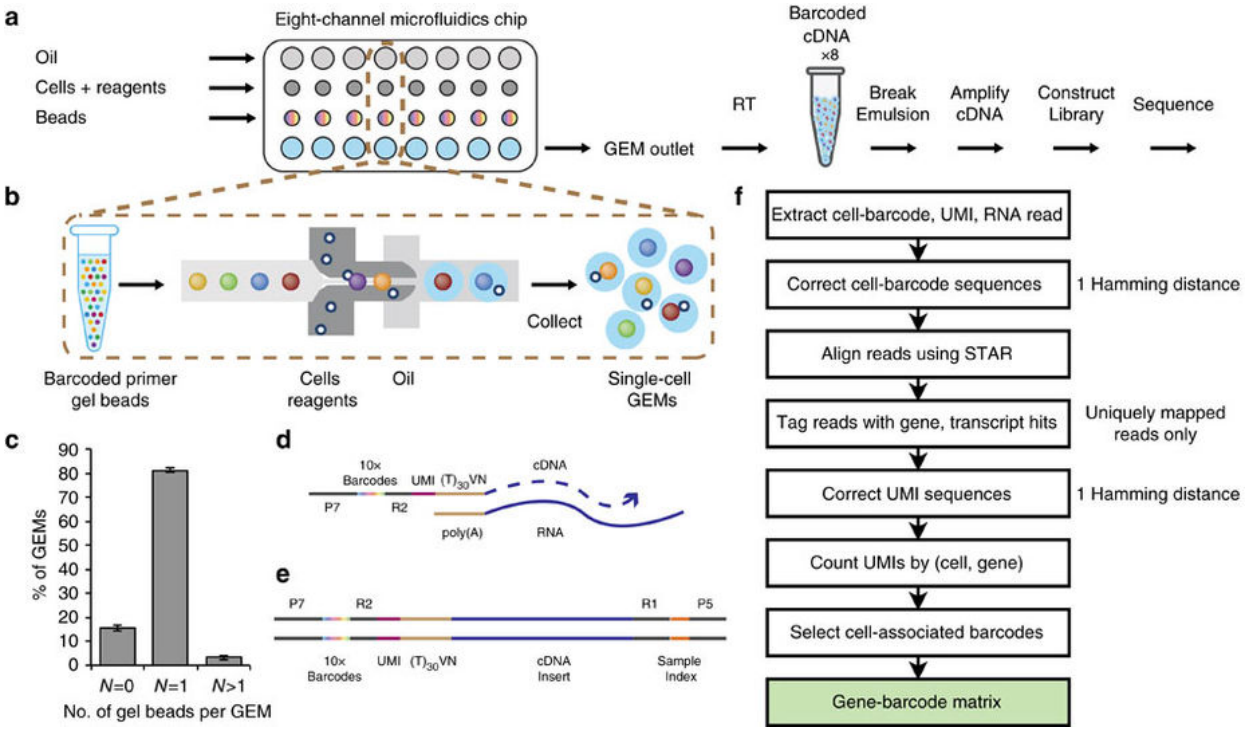


- Unique Molecular Identifiers (UMIs) can correct for PCR bias



Remember : UMIs do not correct for low-capture rates, which leads to an abundance of false negatives. Capture rates are estimated to 5-20% across various protocols

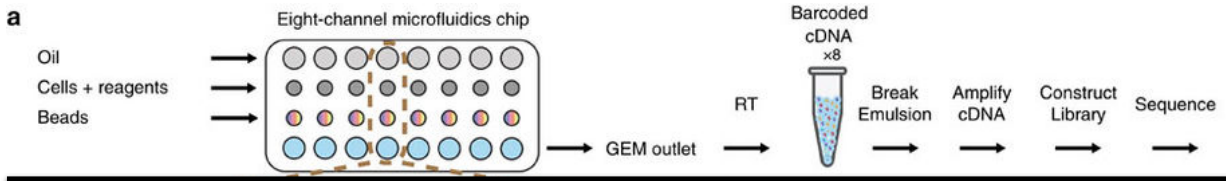
The 10X system: Bringing high throughput scRNA-seq to the masses



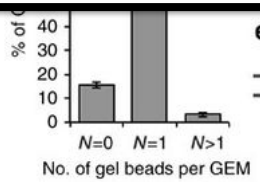
- In contrast to Drop-seq, where solid beads are used for RNA capture, 10X uses soft hydrogels containing oligos. These enable "**single Poisson loading**" leading to capture of >60% of input cells

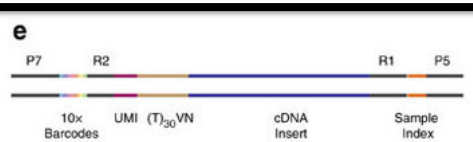
10X GemCode - Zheng et al., 2016

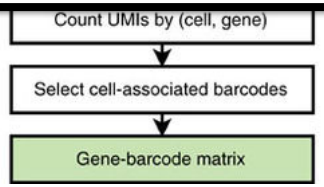
The 10X system: Bringing high throughput scRNA-seq to the masses



- **The 10X system allows 8 samples to be processed simultaneously, with a capture of ~2000-8000 cells per sample**
- **The doublet rate increases with the # of cells loaded**





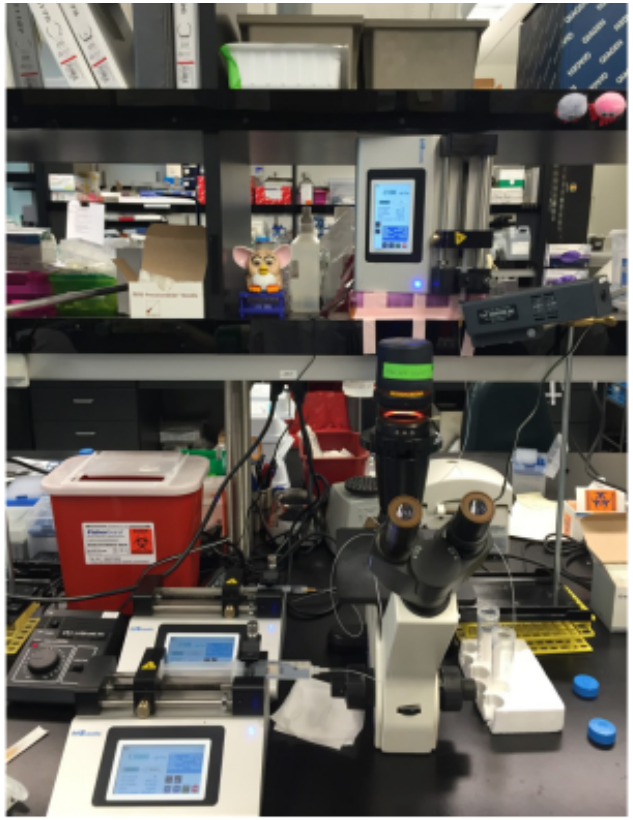


- In contrast to Drop-seq, where solid beads are used for RNA capture, 10X uses soft hydrogels containing oligos. These enable "**single Poisson loading**" leading to capture of >60% of input cells

10X GemCode - Zheng et al., 2016

Drop-seq vs 10X

Drop-seq setup

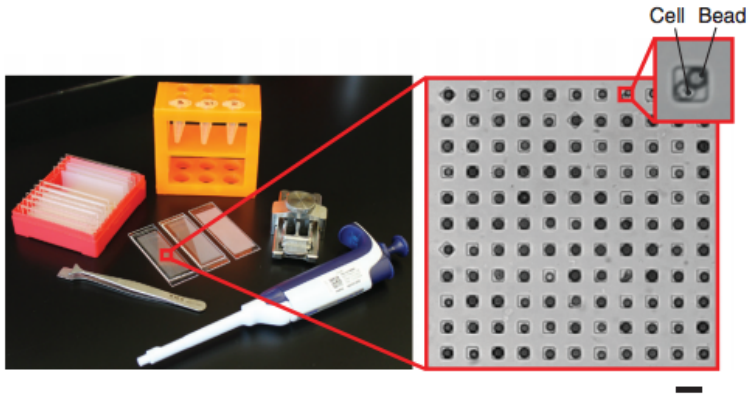


10X Chromium V2

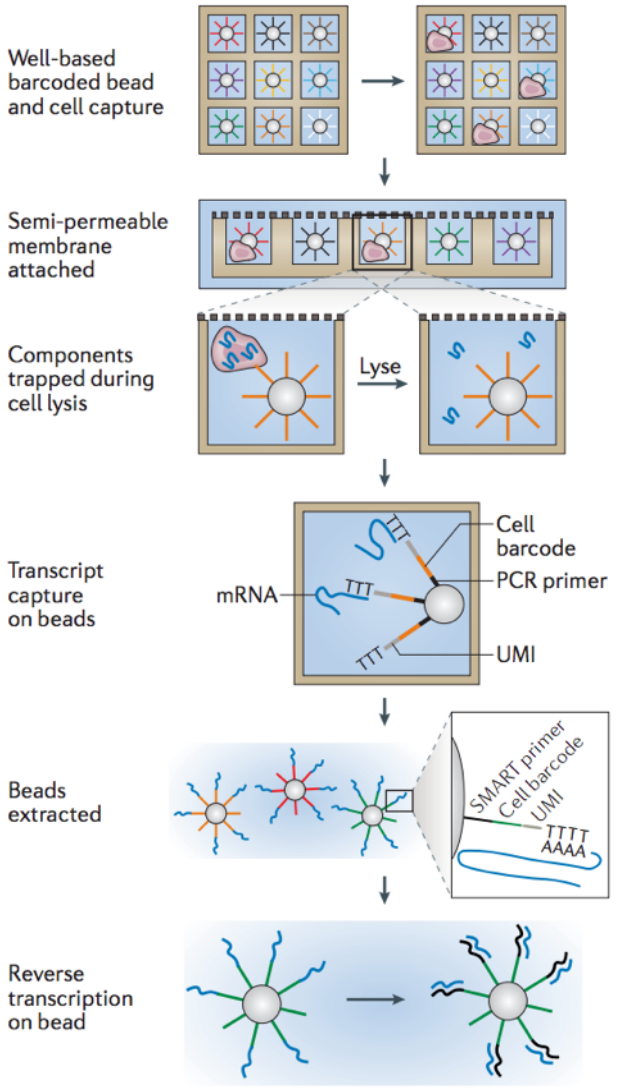
- ✓ Customizable/Controllable
- ✓ Cheap (< 10c /cell)
- \- Requires expertise and optimization
- \- Cell loss / serial processing

- ✓ Easy to use
- ✓ Capture efficiency / parallelized
- \- Expensive (20-30c / cell)
- \- Unhackable

Seq-Well : scRNA-seq in nano-wells

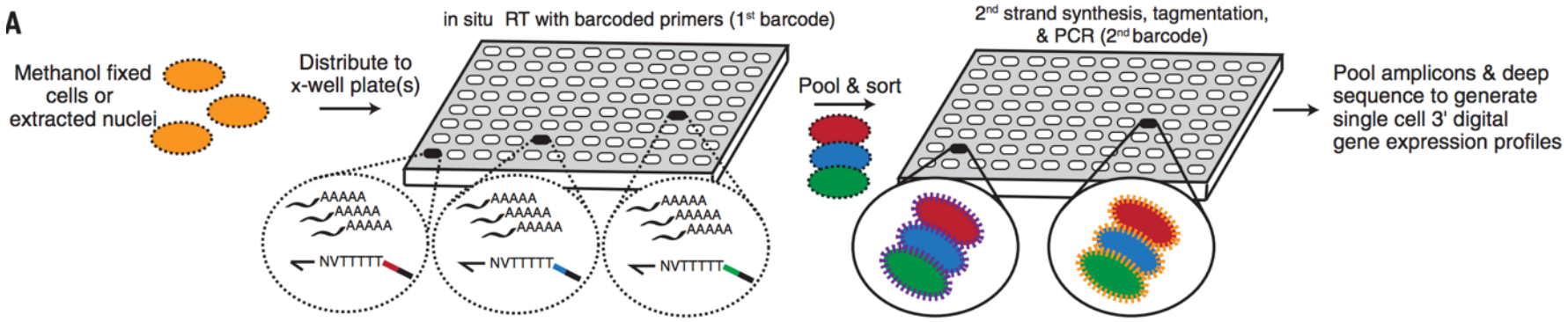


- Capture beads are loaded onto each well in the chip
- Cells are flown in at a dilution such that a well contains no more than one cell
- Platform permits imaging after staining with antibodies
- Higher capture efficiency compared to Drop-seq



Gierahn et al., *Nature Methods*, 2016

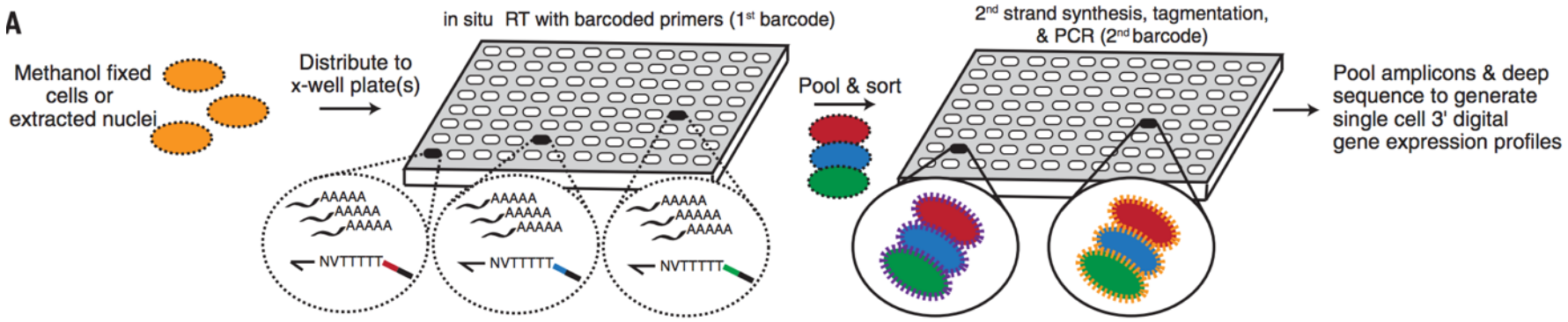
3. Scaling up via combinatorial indexing



Cao et al., *Science*, 2017

3. Scaling up via combinatorial indexing

Introduction of the first cell barcode
+ UMI

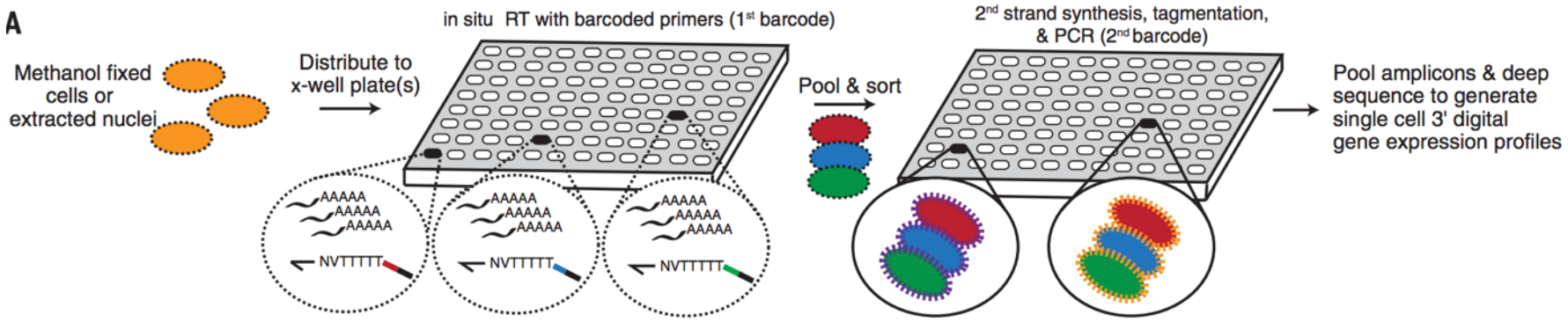


Cao et al., *Science*, 2017

3. Scaling up via combinatorial indexing

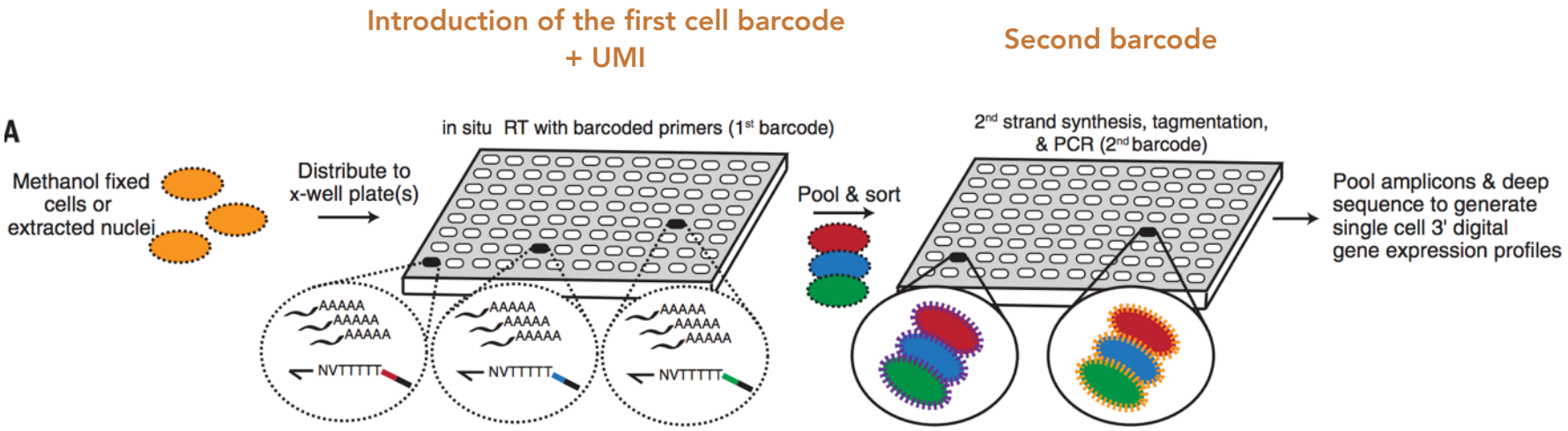
Introduction of the first cell barcode + UMI

Second barcode



Cao et al., *Science*, 2017

3. Scaling up via combinatorial indexing

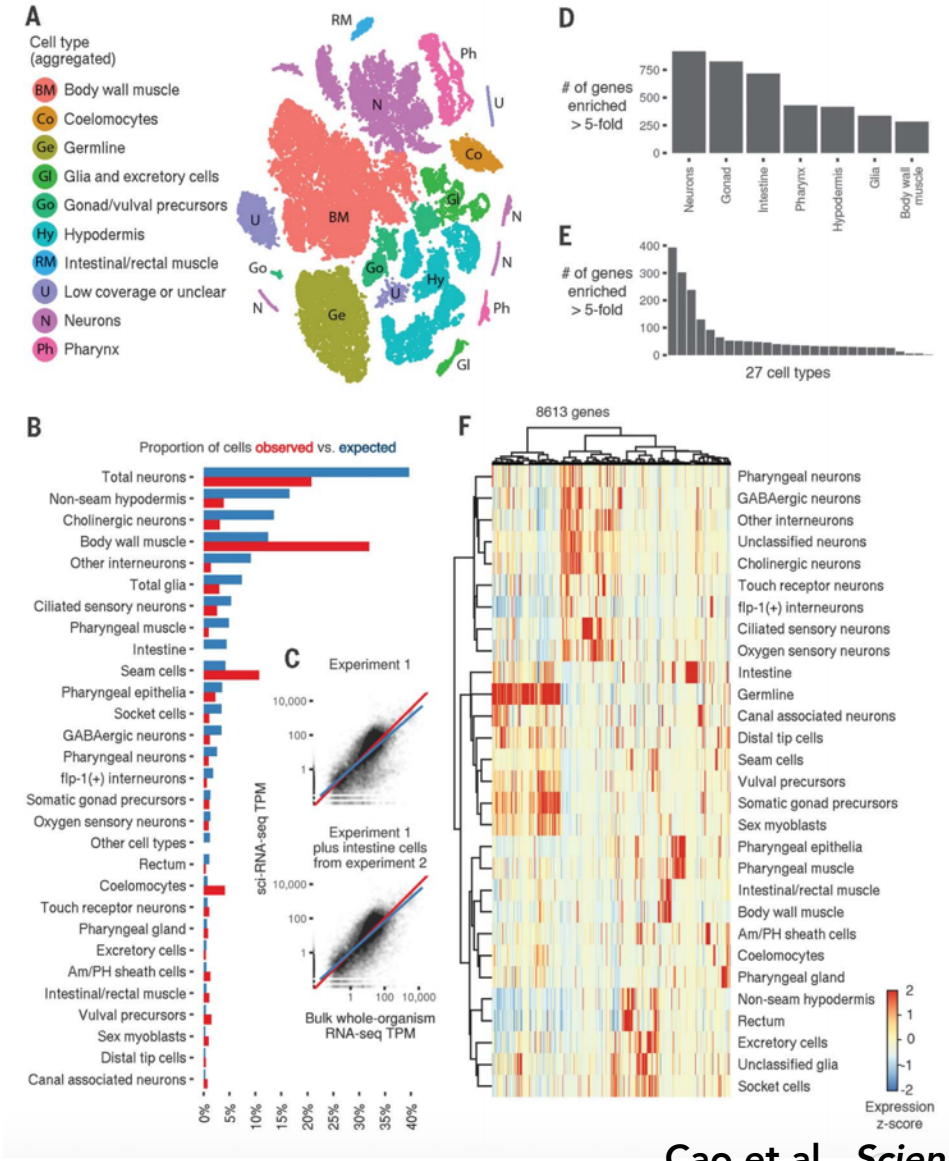


- Two rounds of indexing on 96-well plate enables 96^2 ~ 10,000 cells to be barcoded
- Cells that are together in the same well in both steps cannot be distinguished ("collision")
- Additional rounds of indexing can "exponentially" increase throughput
- Simplest available protocol and most economical (~5c/cell)

Cao et al., *Science*, 2017

Whole organism scRNA-seq

C. elegans

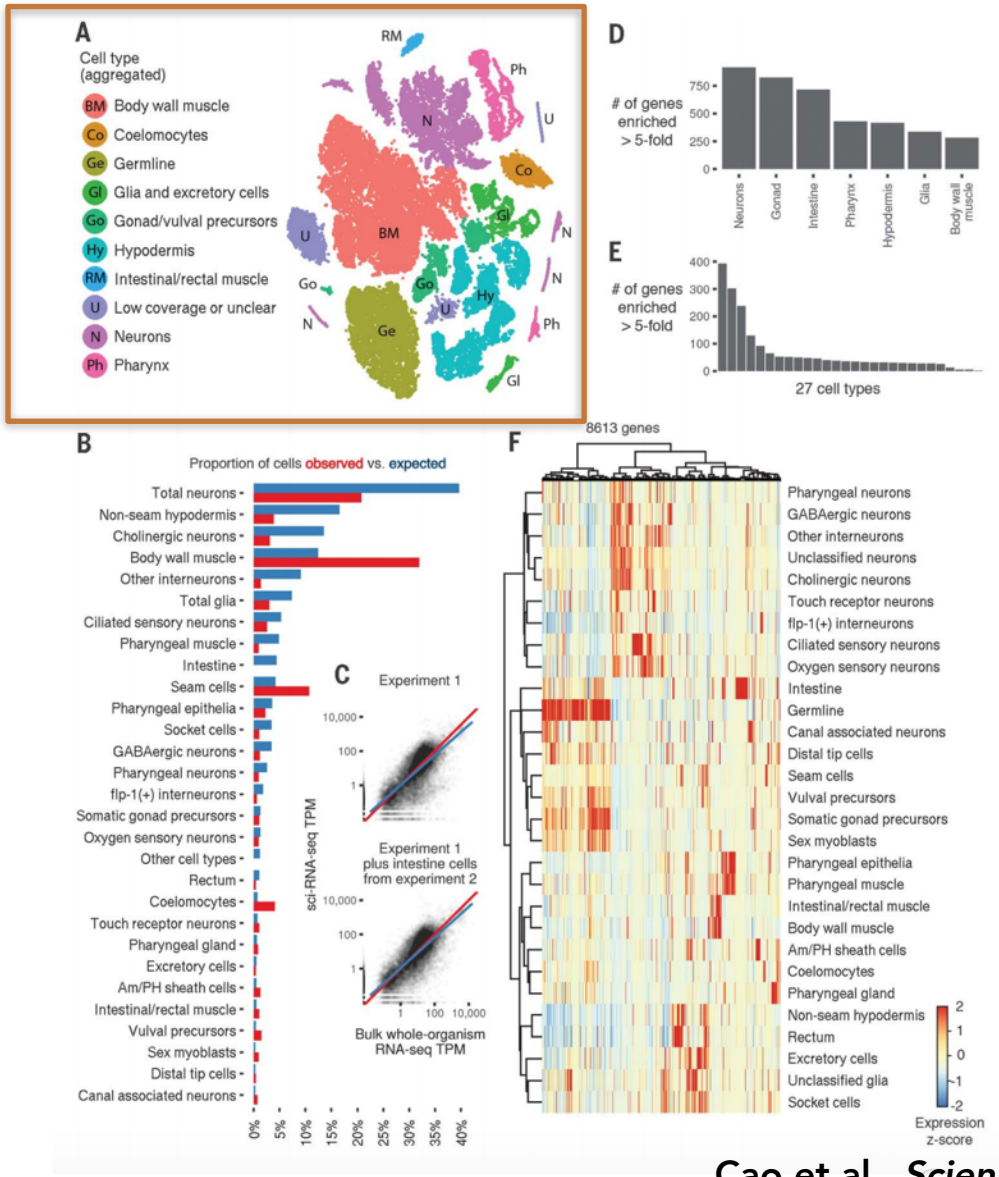


Cao et al., *Science*, 2017

Whole organism scRNA-seq

Marker-free identification of different tissues (43,000 cells)

C. elegans

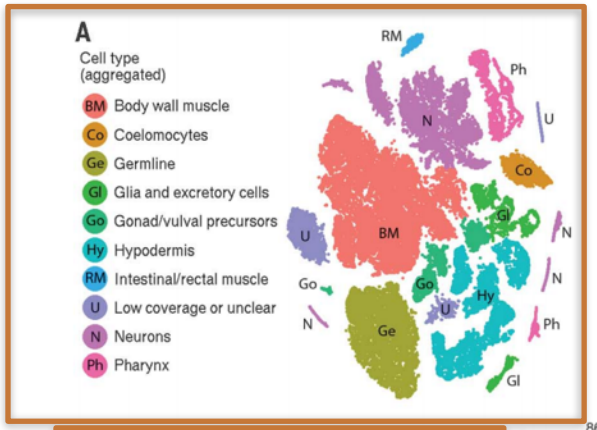


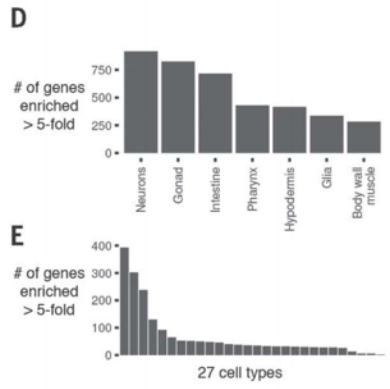
Cao et al., *Science*, 2017

Whole organism scRNA-seq

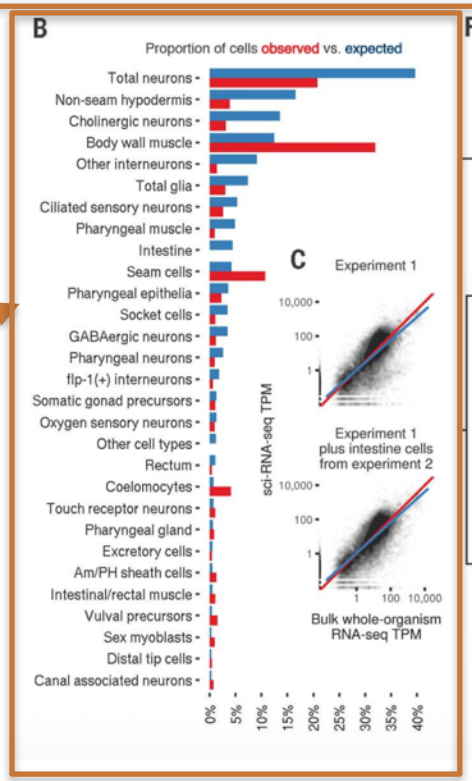
Marker-free identification of different tissues (43,000 cells)

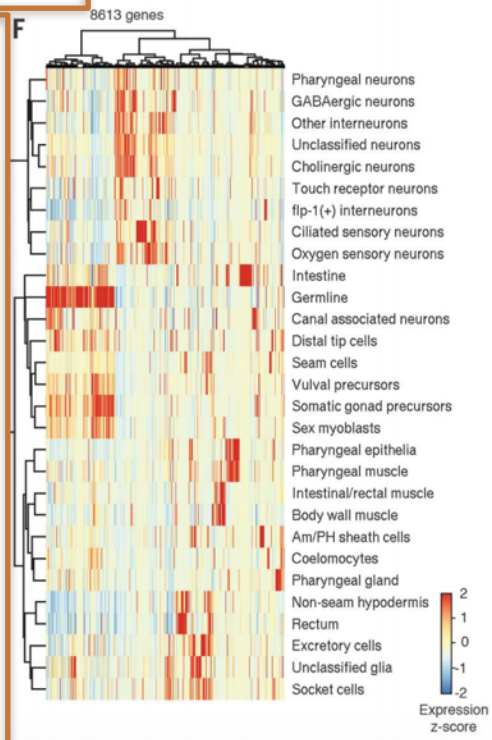
C. elegans





Expected vs. observed proportion of cell types





Cao et al., *Science*, 2017

Considerations for scRNA-seq

Choose protocol based on :

- Throughput (# of cells / reaction)
- Sample of origin
- Cost / Labor / Time limitations
- Gene body coverage - 5', 3' biased, or full-length?
- UMI vs no-UMI
- Sequencing depth / cell

Considerations for scRNA-seq

Choose protocol based on :

- Throughput (# of cells / reaction)
- Sample of origin
- Cost / Labor / Time limitations
- Gene body coverage - 5', 3' biased, or full-length?
- UMI vs no-UMI
- Sequencing depth / cell

For example :

- If I want to classify all cell types in a diverse tissue (e.g. brain), I **need high throughput**
- If I want to re-annotate the transcriptome and discover new isoforms, I need **full-length coverage**
- If I only have access to archival human samples, I will need to use a method that permits fixed cells (or nuclei)

More cells or more reads / cell

- Earlier scRNA-seq studies used to sequence each cell to > 10 million reads / cell - **this is now widely accepted as a ridiculous number!**

- **~50,000-100,000 reads/cell** is now widely regarded as sufficient for most applications. ~1M reads per cell effectively means saturation

- The modular nature of biology “guarantees” that key signals can be recovered at shallow sequencing depth*

Cleary et al., Cell, 2017

Experimental Design

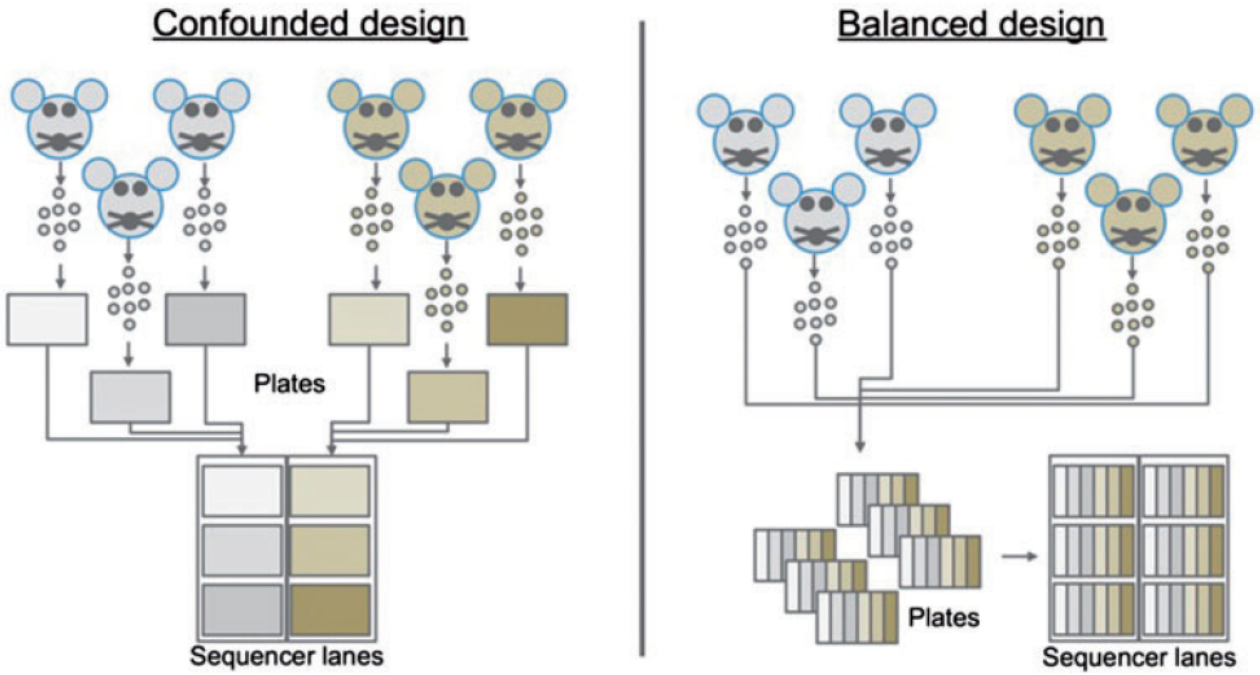
Sound experimental design : Replication, Randomization and Blocking

- R. A. Fisher, 1935

Experimental Design

Sound experimental design : Replication, Randomization and Blocking

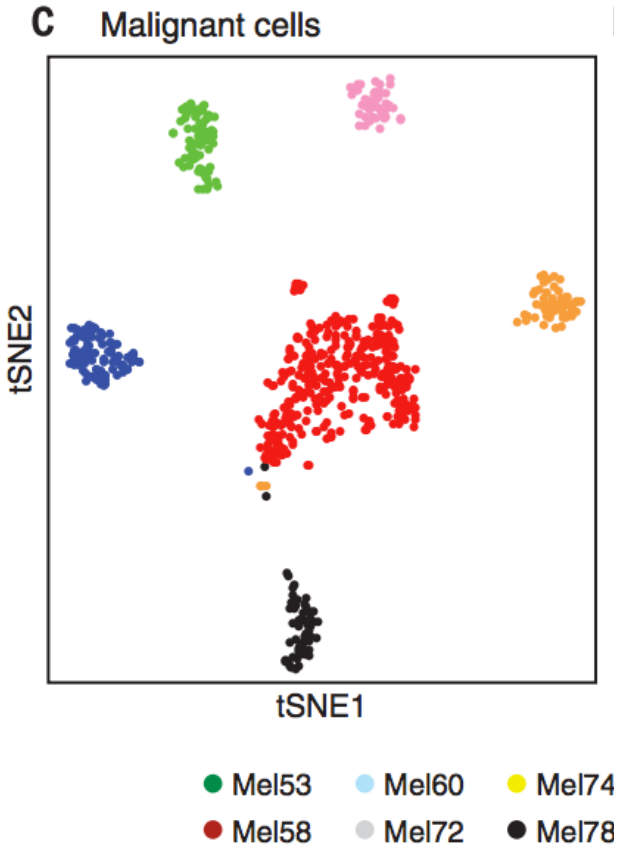
- R. A. Fisher, 1935



- For example, when analyzing tumor phenotypes in a patient process the tumor sample and a matched control on the same day, using the same reagents!
- Blocking is not always possible because of logistic limitations, in which case ensure that any biological conclusion is supported by multiple, independently collected samples

Checks and balances if blocking cannot be done

- **Metastatic Melanoma** : From the point of view of experimental design, a completely confounded study!

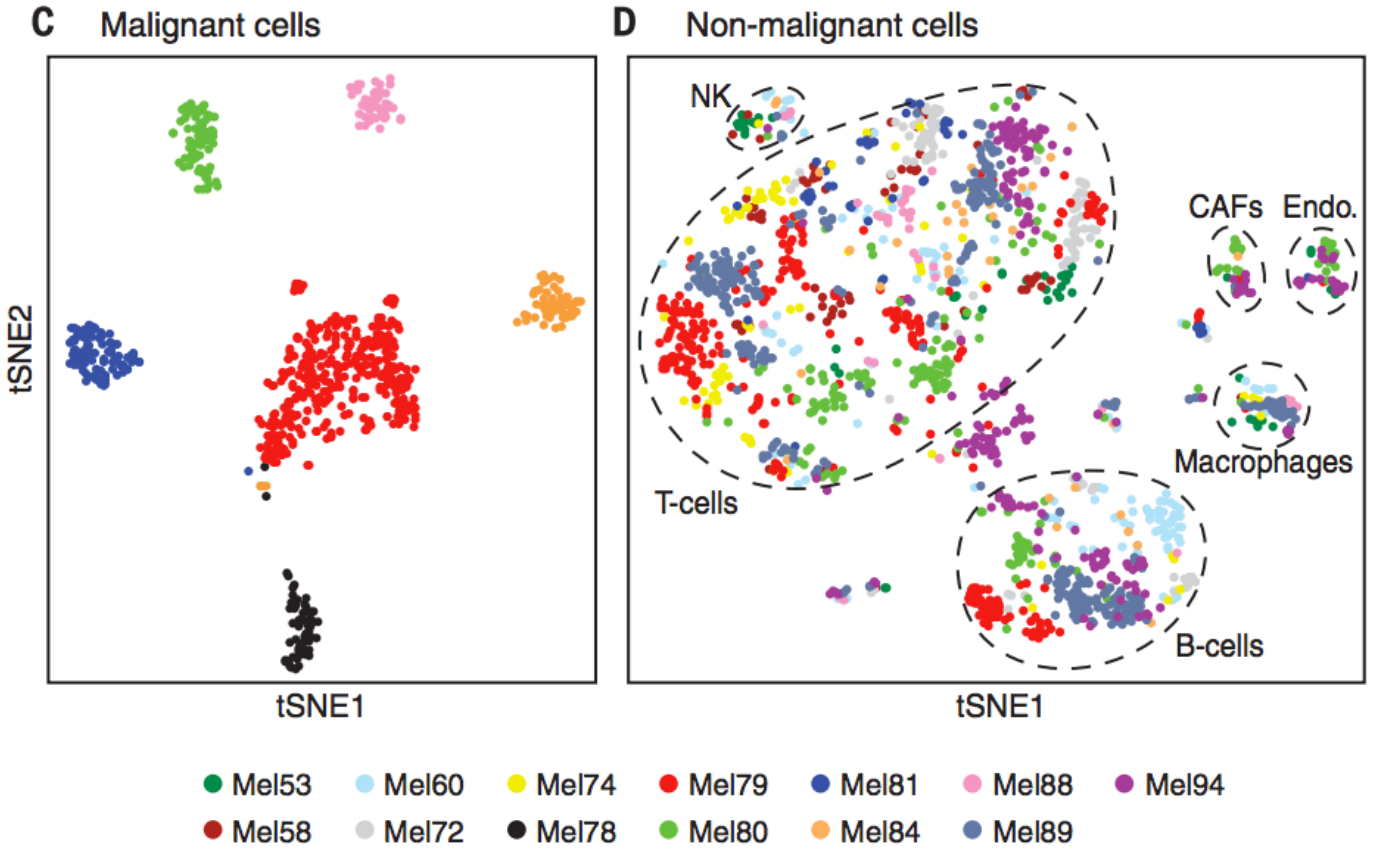


Tumor cells cluster by patient. By itself, this could be simply batch effects!

Tirosh et al., Science 2016

Checks and balances if blocking cannot be done

- **Metastatic Melanoma** : From the point of view of experimental design, a completely confounded study!



Tumor cells cluster by patient. By itself, this could be simply batch effects!

But non-malignant cells cluster by type, rather than patient!

Tirosh et al., Science 2016

Coffee Break

Agenda

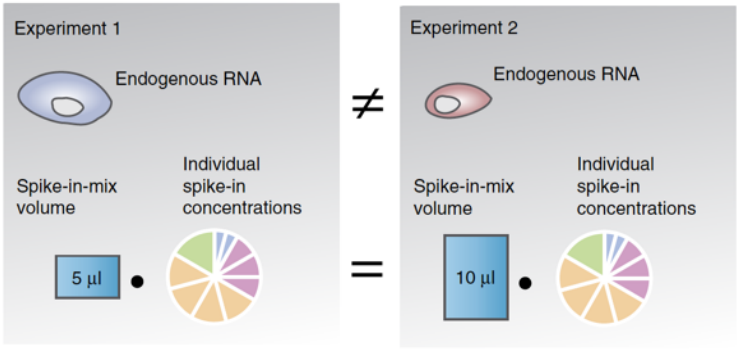
- Single cell analysis - why?
- A short survey of scRNA-seq methods
- **Quality comparison of different methods and power analysis**
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R
- Some example applications
- Future

Quantifying sensitivity and accuracy

- How efficiently can we detect mRNA molecules? Can we measure one molecule? (**Sensitivity**)
- Do we detect molecules in proportion to their abundance? (**Accuracy**)

Quantifying sensitivity and accuracy

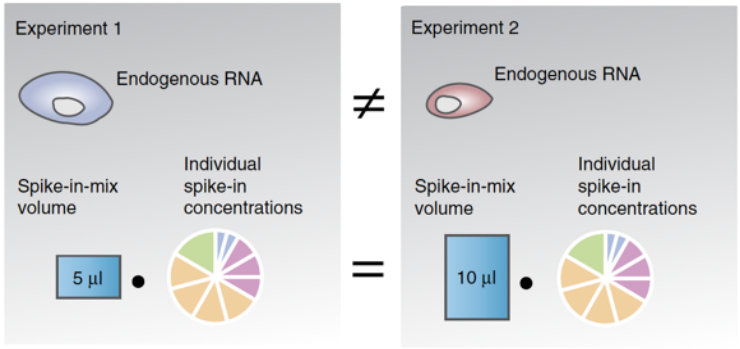
- How efficiently can we detect mRNA molecules? Can we measure one molecule? (**Sensitivity**)
- Do we detect molecules in proportion to their abundance? (**Accuracy**)



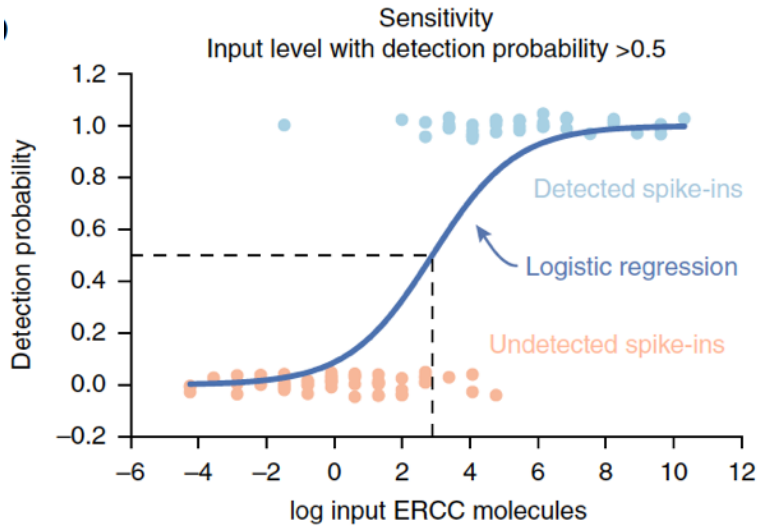
External RNA Controls Consortium (**ERCC**) spike-in mix : A standardized soup of 92 RNA molecules at varying concentration levels that are **known**

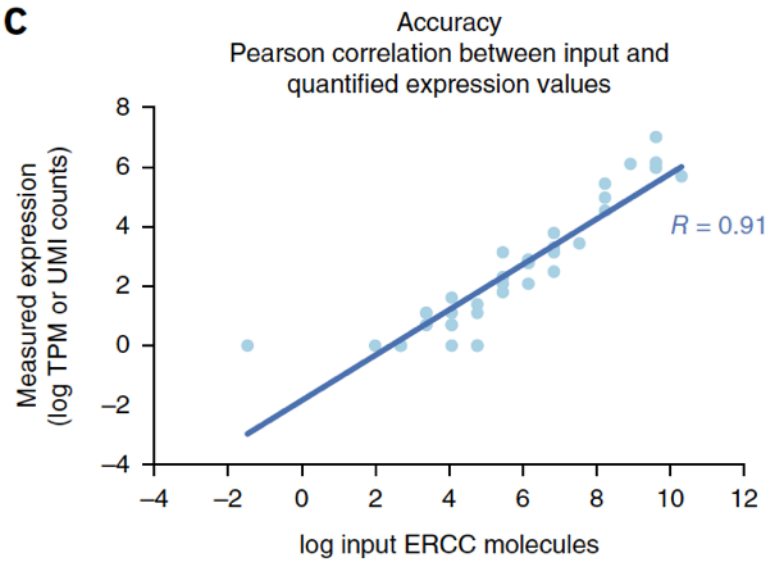
Quantifying sensitivity and accuracy

- How efficiently can we detect mRNA molecules? Can we measure one molecule? (**Sensitivity**)
- Do we detect molecules in proportion to their abundance? (**Accuracy**)



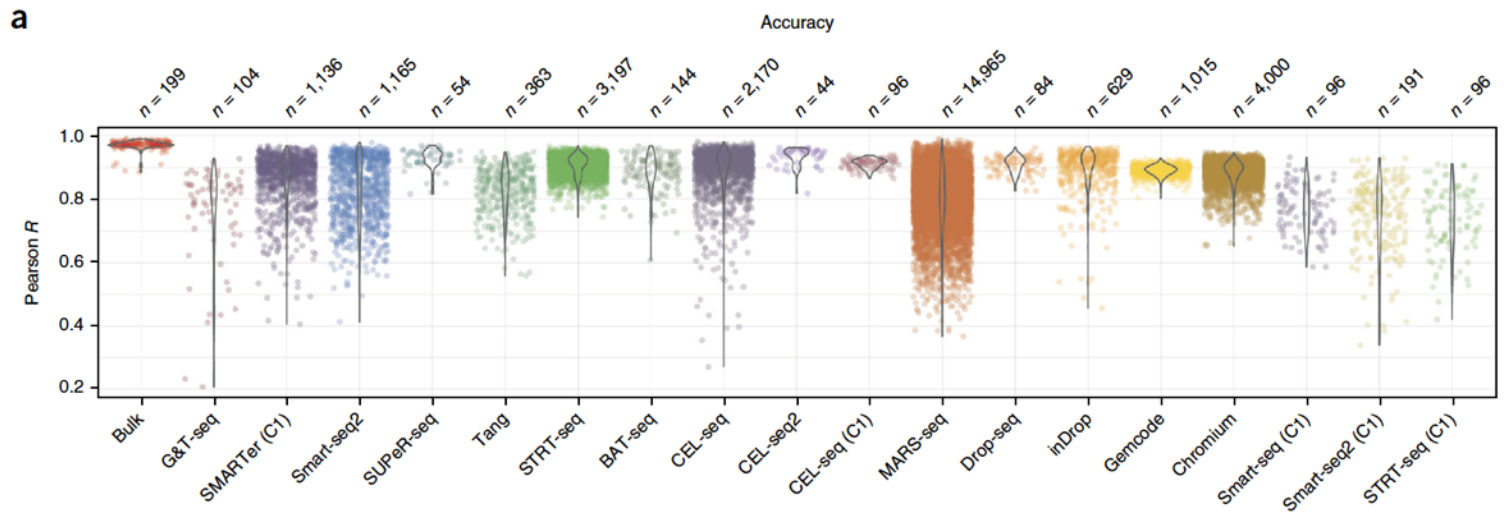
External RNA Controls Consortium (**ERCC**) spike-in mix : A standardized soup of 92 RNA molecules at varying concentration levels that are **known**





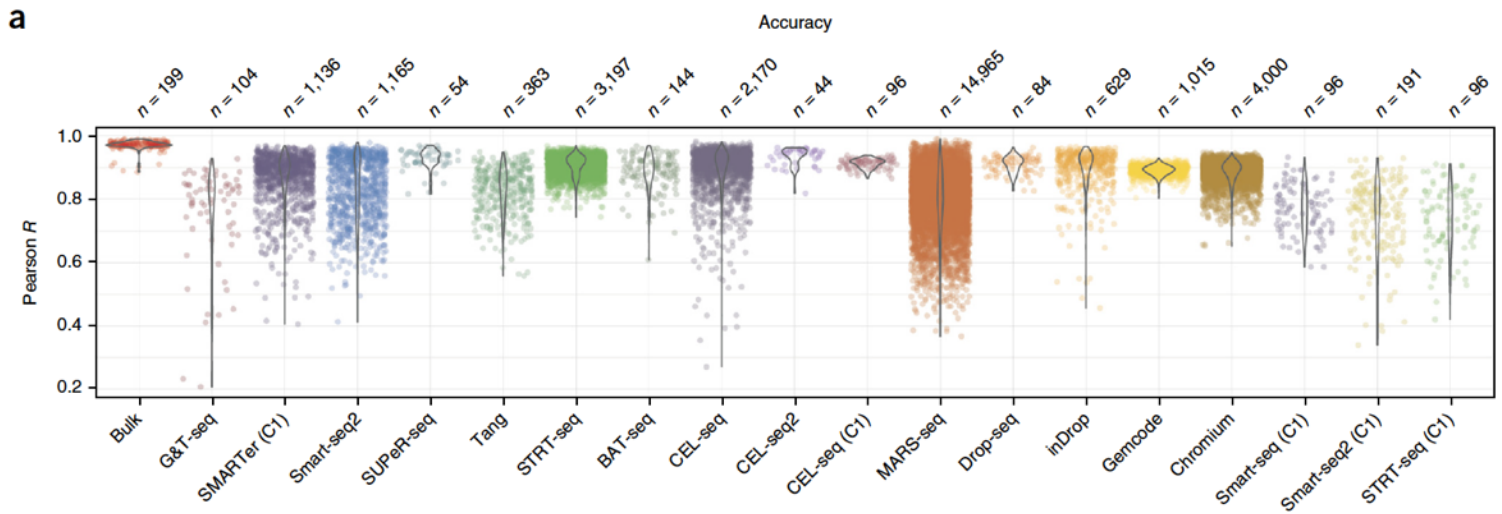
Svensson et al., *Nature Methods*, 2017

Comparing Sensitivity and Accuracy across RNA-seq protocols



Svensson et al., *Nature Methods*, 2017

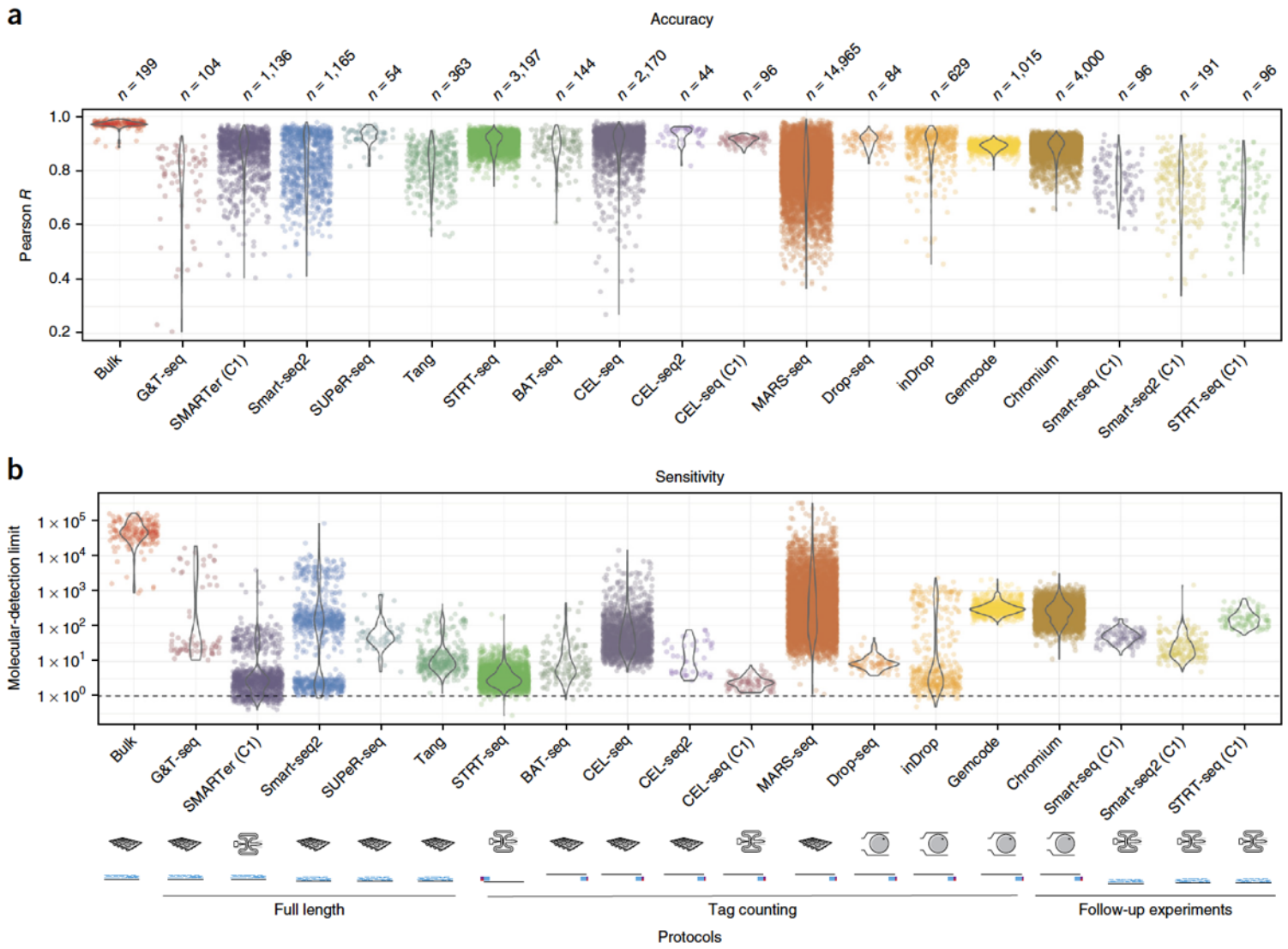
Comparing Sensitivity and Accuracy across RNA-seq protocols



Larger is better!

Svensson et al., *Nature Methods*, 2017

Comparing Sensitivity and Accuracy across RNA-seq protocols

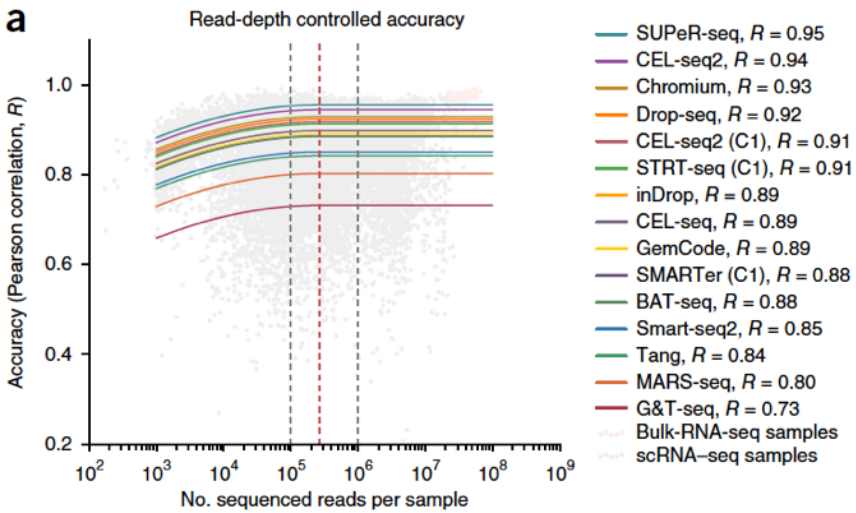


Larger is better!

Smaller is better!

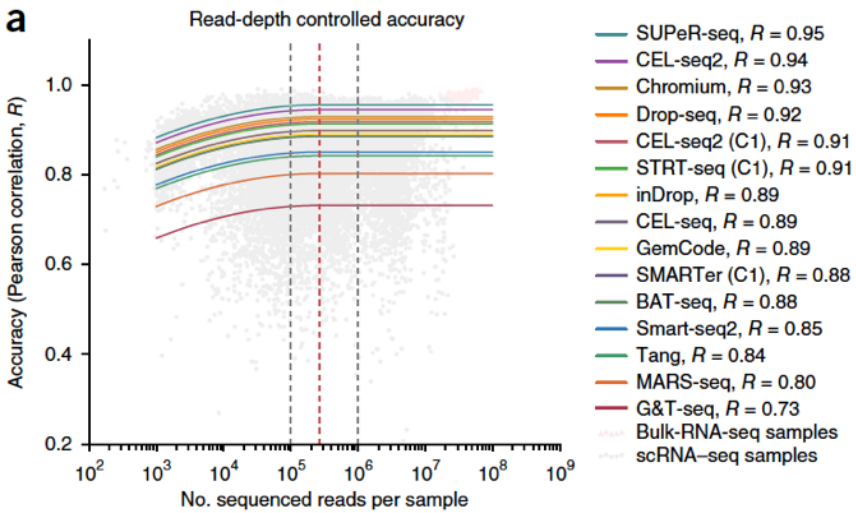
Svensson et al., *Nature Methods*, 2017

How does sequencing depth impact accuracy and sensitivity



Svensson et al., Nature Methods, 2017

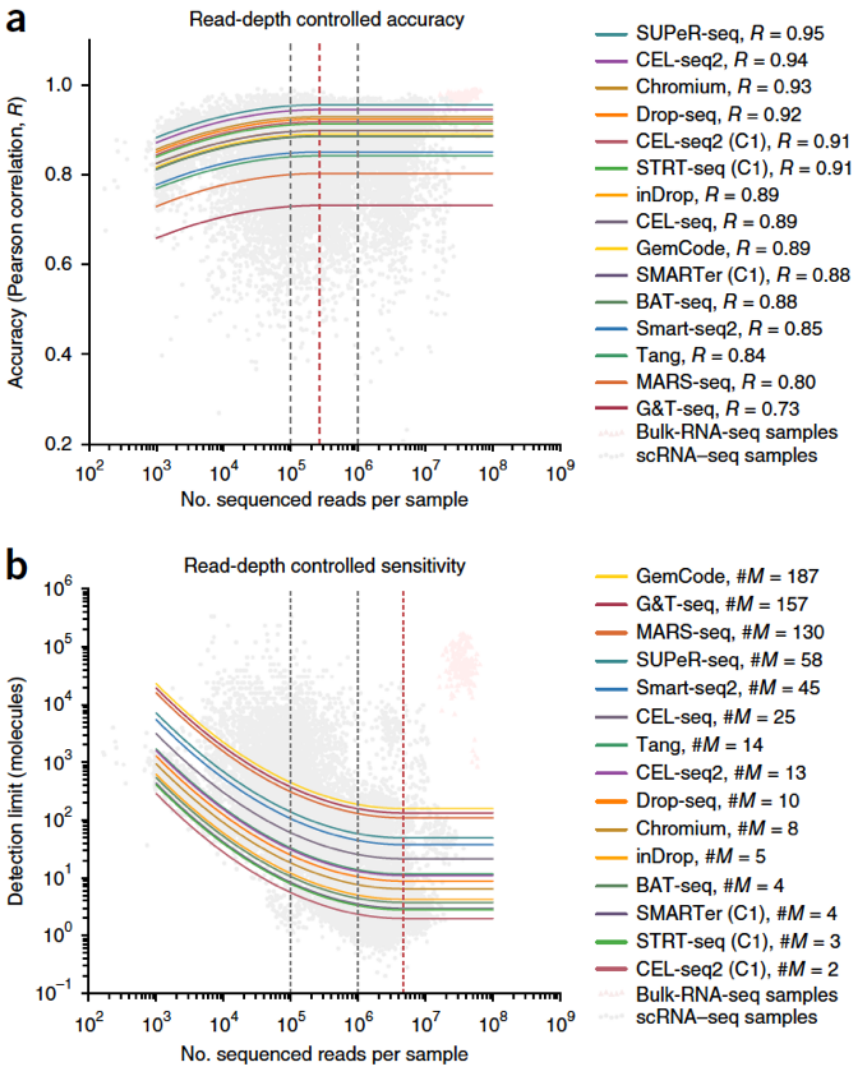
How does sequencing depth impact accuracy and sensitivity



For a given method, read depth does not impact accuracy beyond 50k reads/ cell

Svensson et al., Nature Methods, 2017

How does sequencing depth impact accuracy and sensitivity



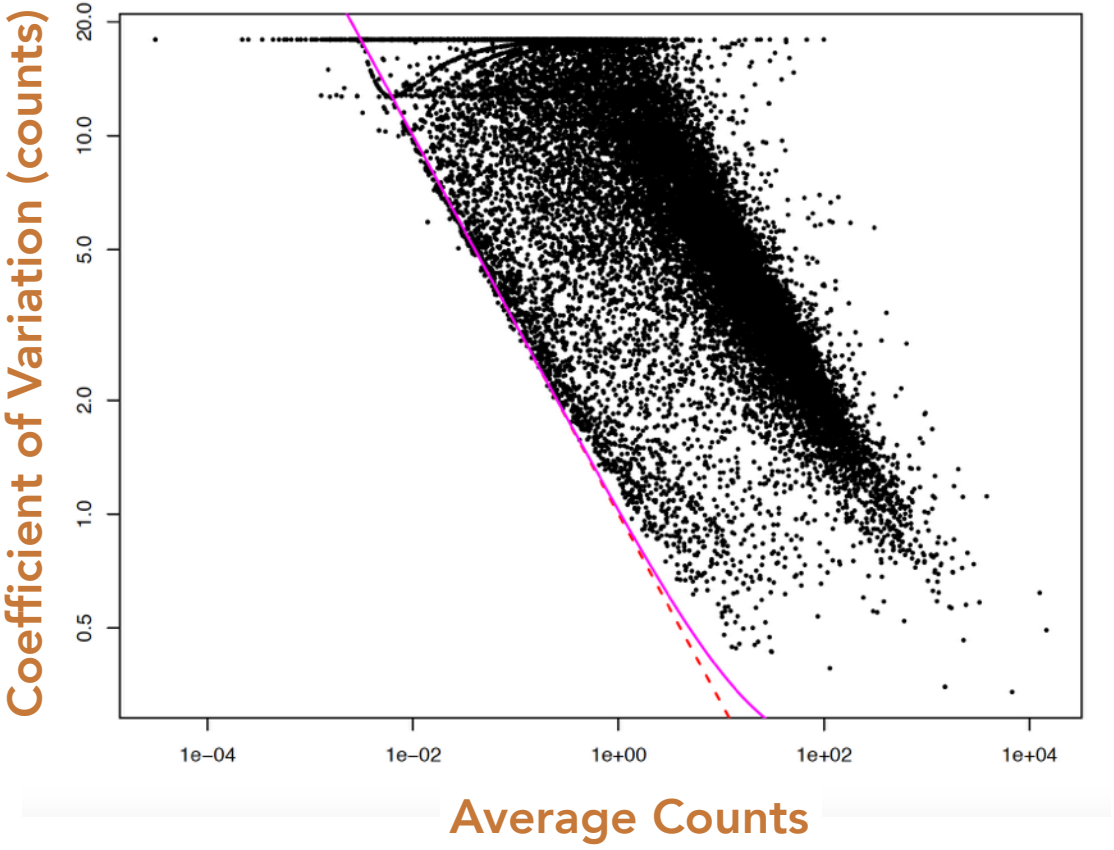
For a given method, read depth does not impact accuracy beyond 50k reads/ cell

Sensitivity is sharply impacted by sequencing depth!

Svensson et al., Nature Methods, 2017

The Impact of UMIs : Variability in gene expression

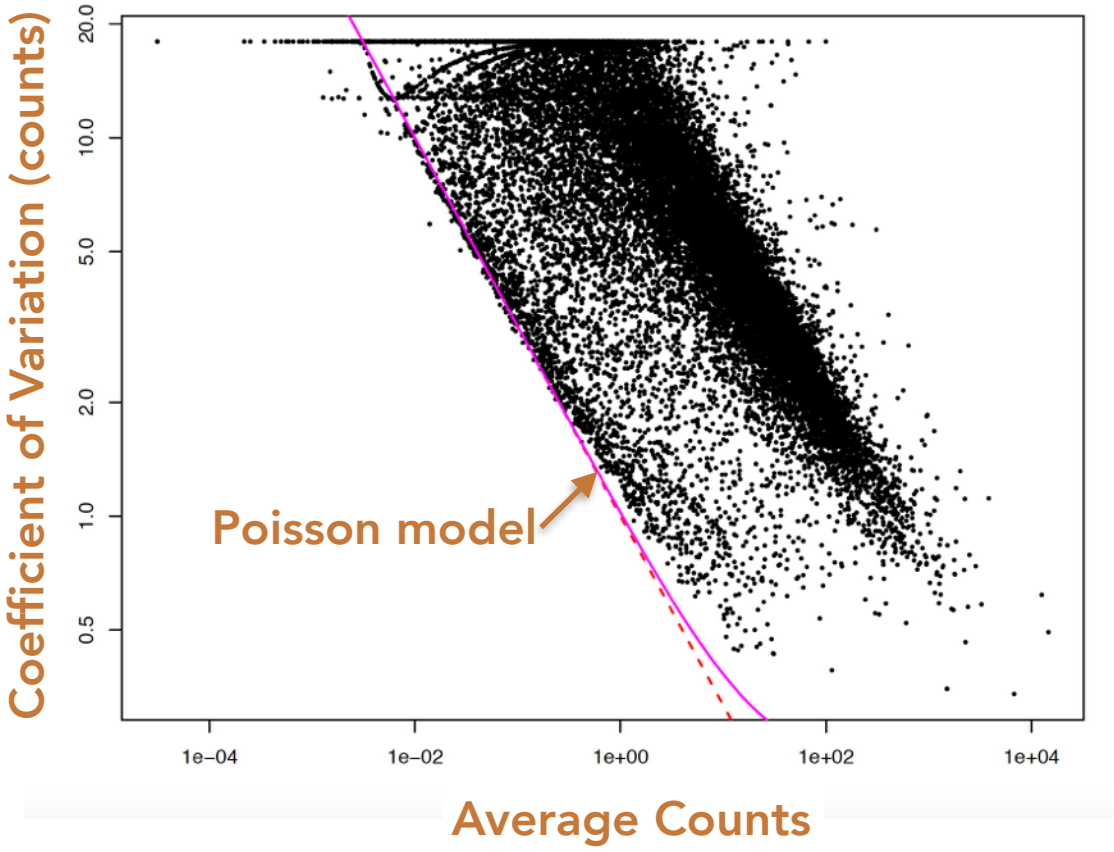
Fundamentally gene expression data consists of expression counts. The Poisson family of distributions is the most natural way to model such data.



Coefficient of variation = Standard Deviation / mean

The Impact of UMIs : Variability in gene expression

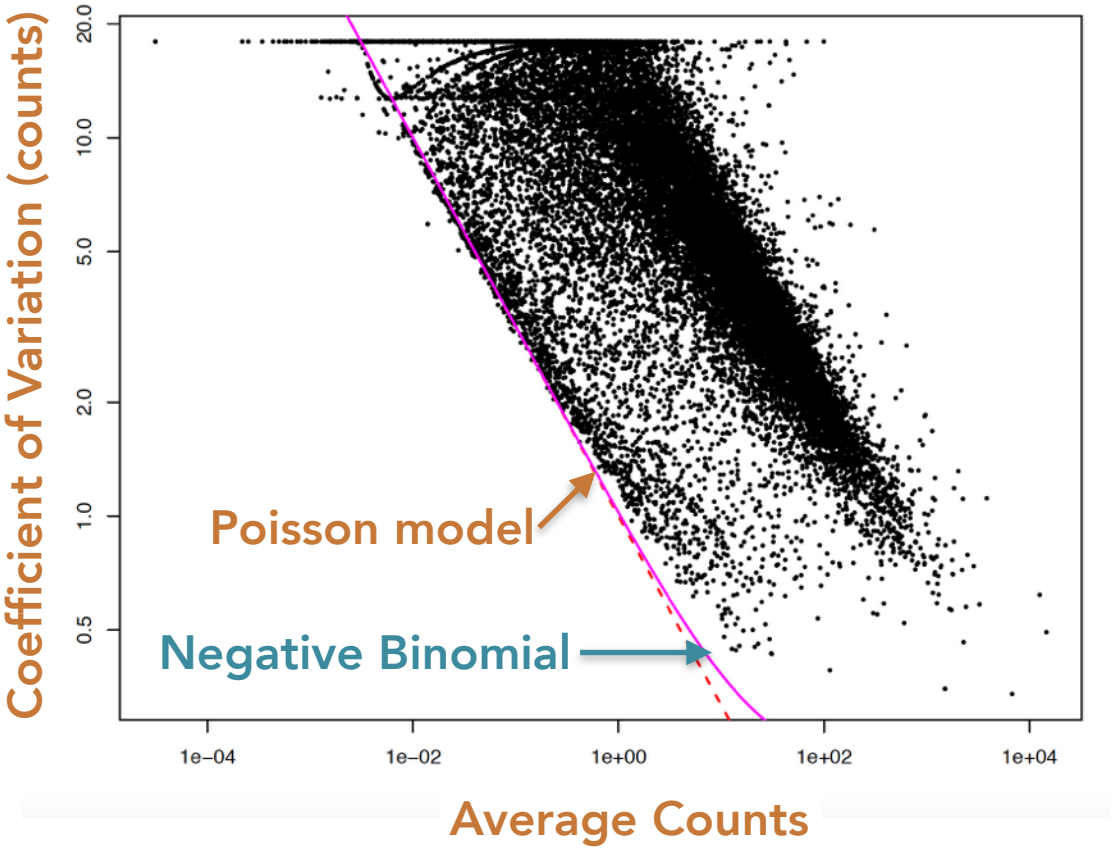
Fundamentally gene expression data consists of expression counts. The Poisson family of distributions is the most natural way to model such data.



Coefficient of variation = Standard Deviation / mean

The Impact of UMIs : Variability in gene expression

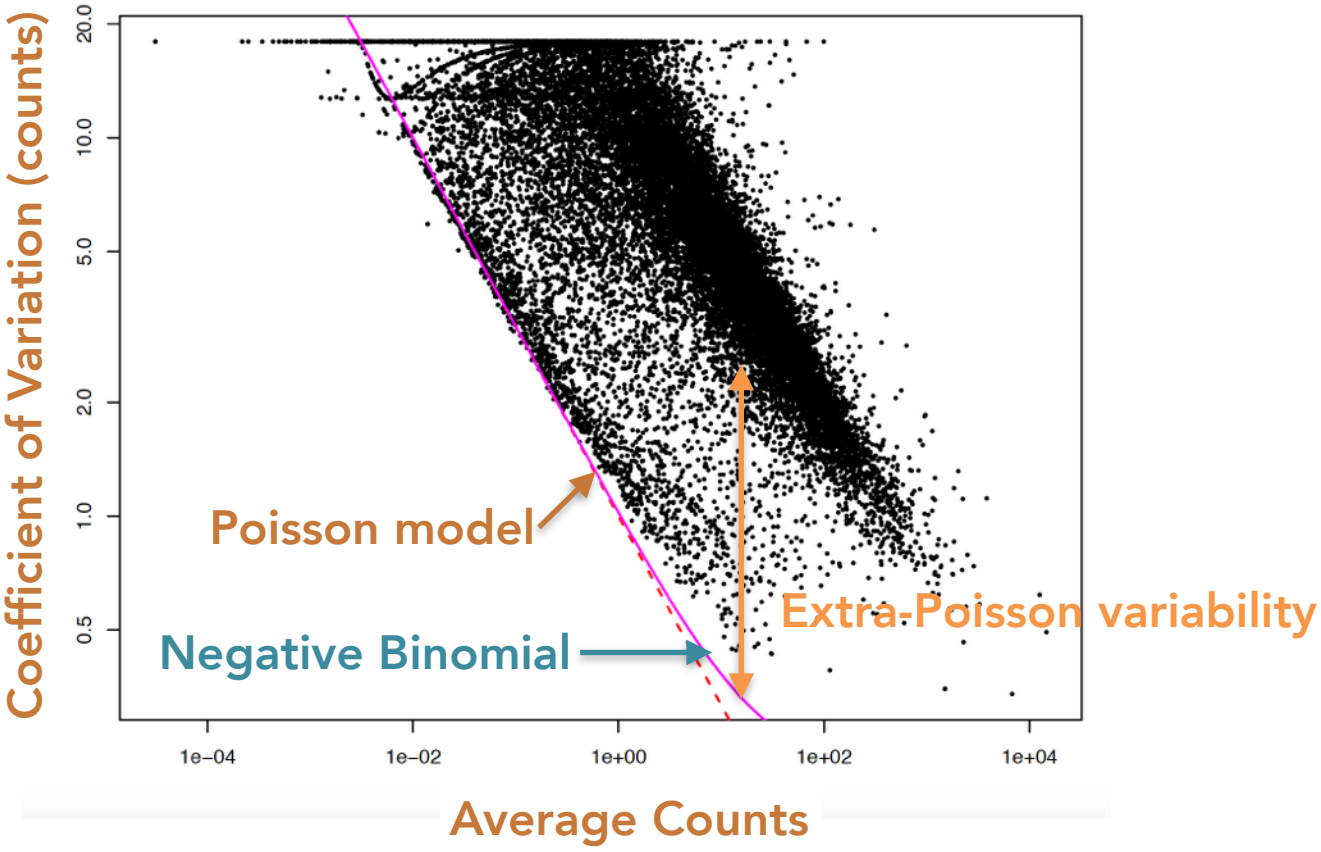
Fundamentally gene expression data consists of expression counts. The Poisson family of distributions is the most natural way to model such data.



Coefficient of variation = Standard Deviation / mean

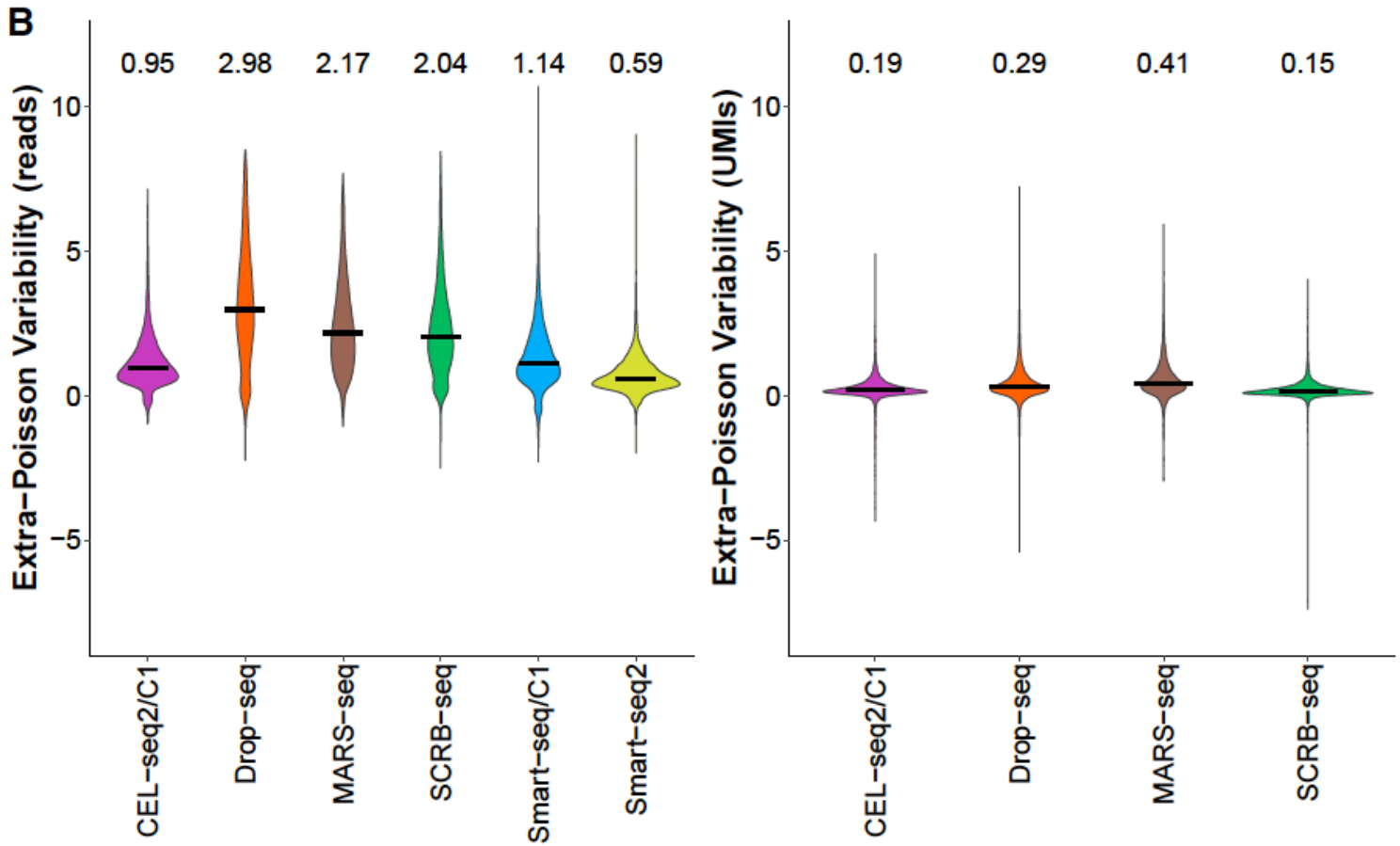
The Impact of UMIs : Variability in gene expression

Fundamentally gene expression data consists of expression counts. The Poisson family of distributions is the most natural way to model such data.



Coefficient of variation = Standard Deviation / mean

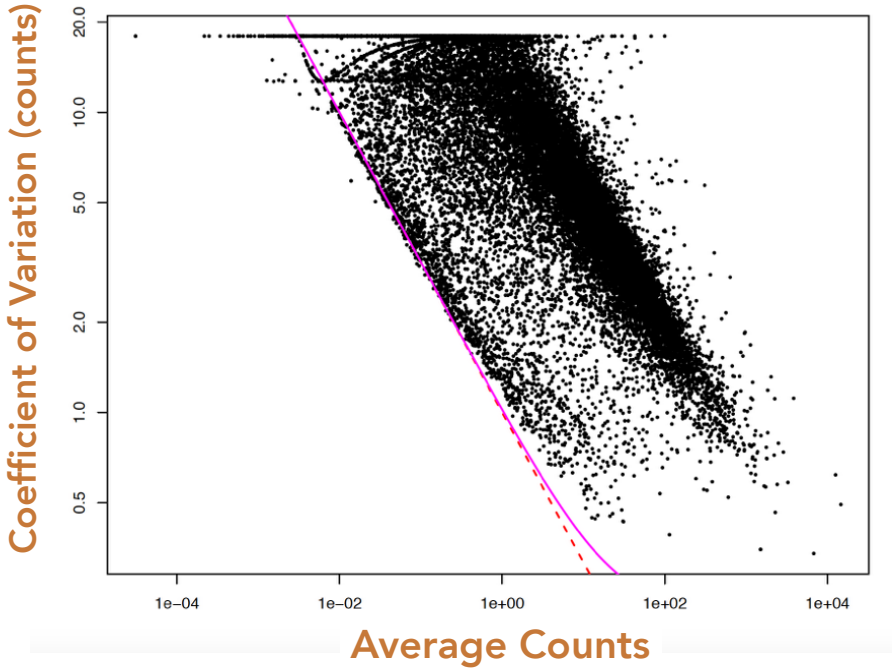
UMIs reduce extra Poisson variability



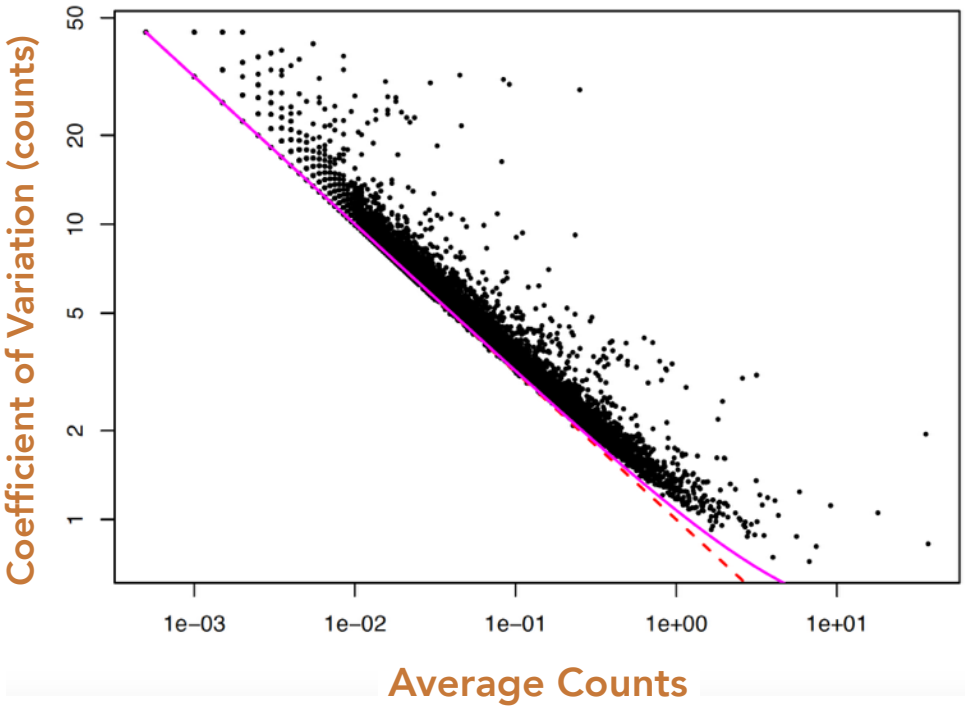
Zieganhain et al., *Cell Reports*, 2017

UMIs reduce extra Poisson variability

Smart-seq2 data (full-length, no UMIs)

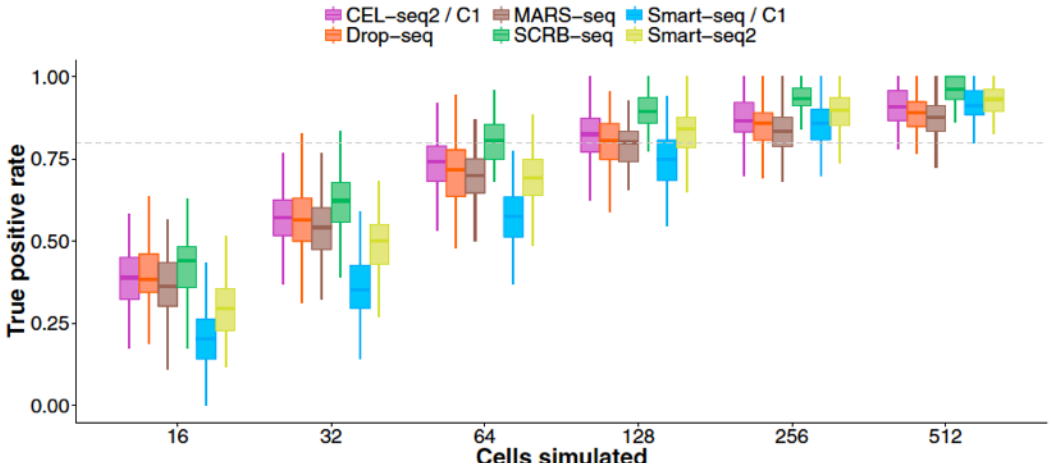


Drop-seq data (3′, with UMIs)



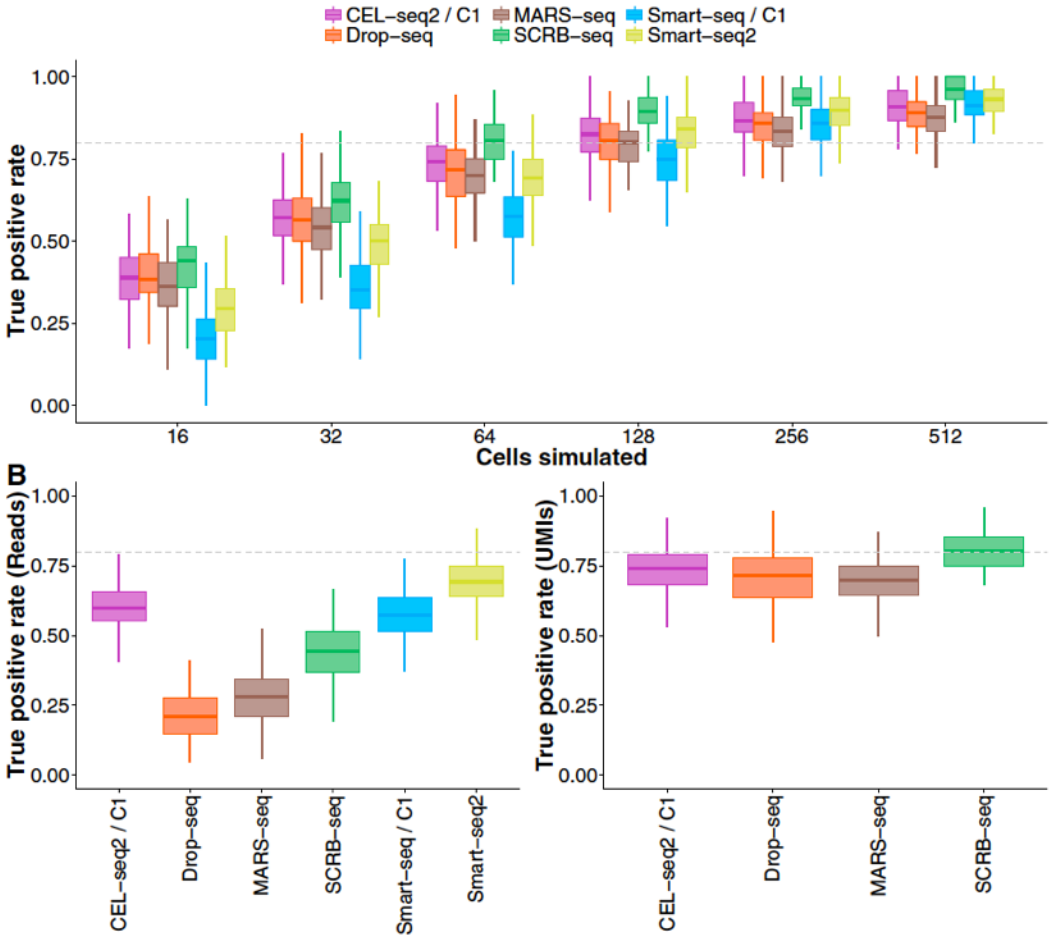
Power analysis of DE test

How many cells (samples) do I need to detect DE genes?



Power analysis of DE test

How many cells (samples) do I need to detect DE genes?



Also, as expected reducing amplification biases greatly increases the power to detect DE genes!

Cost-comparison of scRNA-seq methods

- For each protocol, how many cells do I need to achieve a TPR of 80% to detect DE genes at three different sequencing depths (1M, 0.5M and 0.25M)?
- In each case, what is the cost per cell?

Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040/13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

See also Figure 6.

[a]True positive rate and false discovery rate are based on simulations (Figure 6; Figure S9).

[b]Sequencing depth of one, 0.5, and 0.25 million reads.

[c]Assuming $5 per one million reads.

Zieganhain et al., *Cell Reports*, 2017

Cost-comparison of scRNA-seq methods

- For each protocol, how many cells do I need to achieve a TPR of 80% to detect DE genes at three different sequencing depths (1M, 0.5M and 0.25M)?
- In each case, what is the cost per cell?

Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040/13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

See also Figure 6.

[a]True positive rate and false discovery rate are based on simulations (Figure 6; Figure S9).

[b]Sequencing depth of one, 0.5, and 0.25 million reads.

[c]Assuming $5 per one million reads.

Zieganhain et al., *Cell Reports*, 2017

Cost-comparison of scRNA-seq methods

- For each protocol, how many cells do I need to achieve a TPR of 80% to detect DE genes at three different sequencing depths (1M, 0.5M and 0.25M)?
- In each case, what is the cost per cell?

Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040/13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

See also Figure 6.

[a]True positive rate and false discovery rate are based on simulations (Figure 6; Figure S9).

[b]Sequencing depth of one, 0.5, and 0.25 million reads.

[c]Assuming $5 per one million reads.

Beyond a point, sequencing depth stops affording additional advantage. Always better to sample more cells!

Zieganhain et al., *Cell Reports*, 2017

Agenda

- Single cell analysis - why?
- A short survey of scRNA-seq methods
- Quality comparison of different methods and power analysis
- **Overview of computational workflow**
  - **Preprocessing**
  - Secondary analysis in R
- Some example applications
- Future

Alignment and quantification

Typical Illumina runs (NextSeq) produce ~200 million short reads 30-70 bp long

Alignment and quantification

Typical Illumina runs (NextSeq) produce ~200 million short reads 30-70 bp long

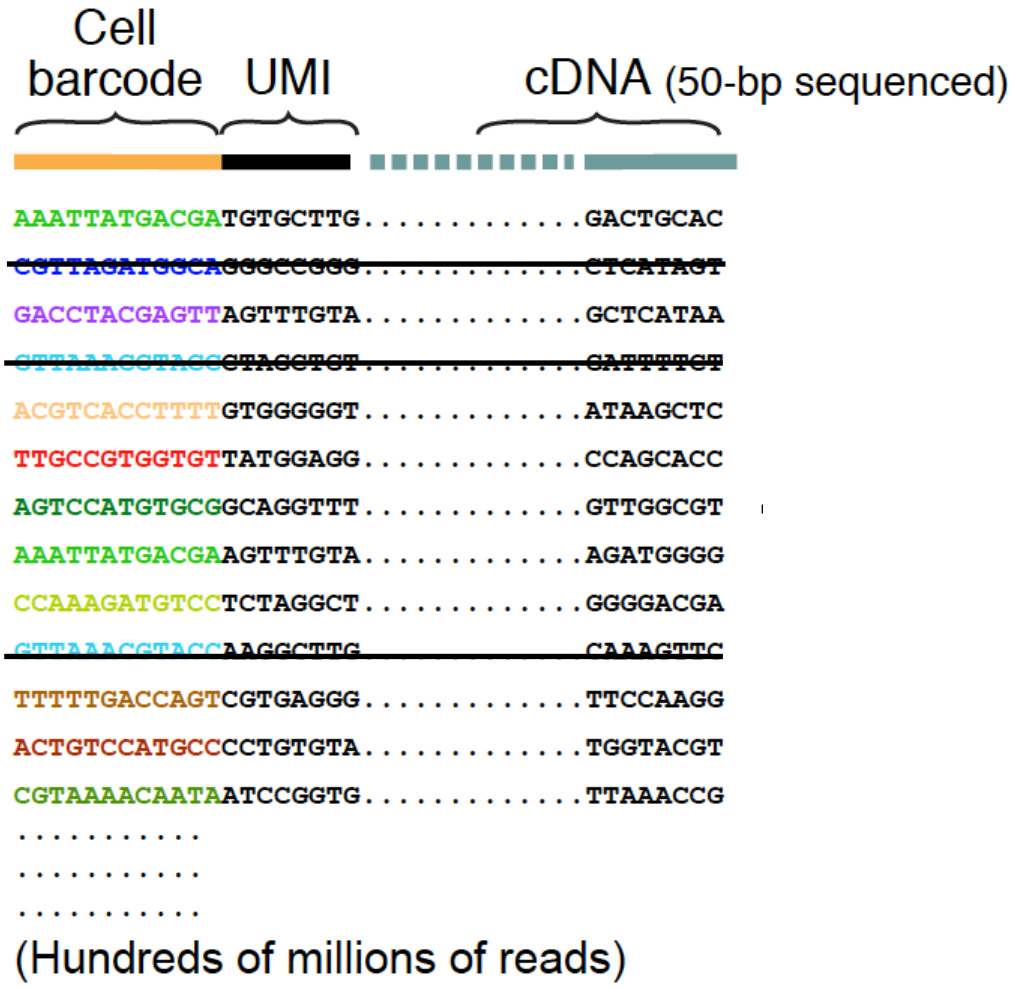
In high throughput methods (e.g. droplet sequencing), one read is reserved for the cell barcode and the UMI

Raw reads


```
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
GACCTACGAGTTAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
ACGTCACCTTTTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
AGTCCATGTGCGGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CCAAAGATGTCCTCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTTTTGACCAGTCGTGAGGG............TTCCAAGG
ACTGTCCATGCCCCTGTGTA............TGGTACGT
CGTAAAACAATAATCCGGTG............TTAAACCG
...........
...........
...........
```

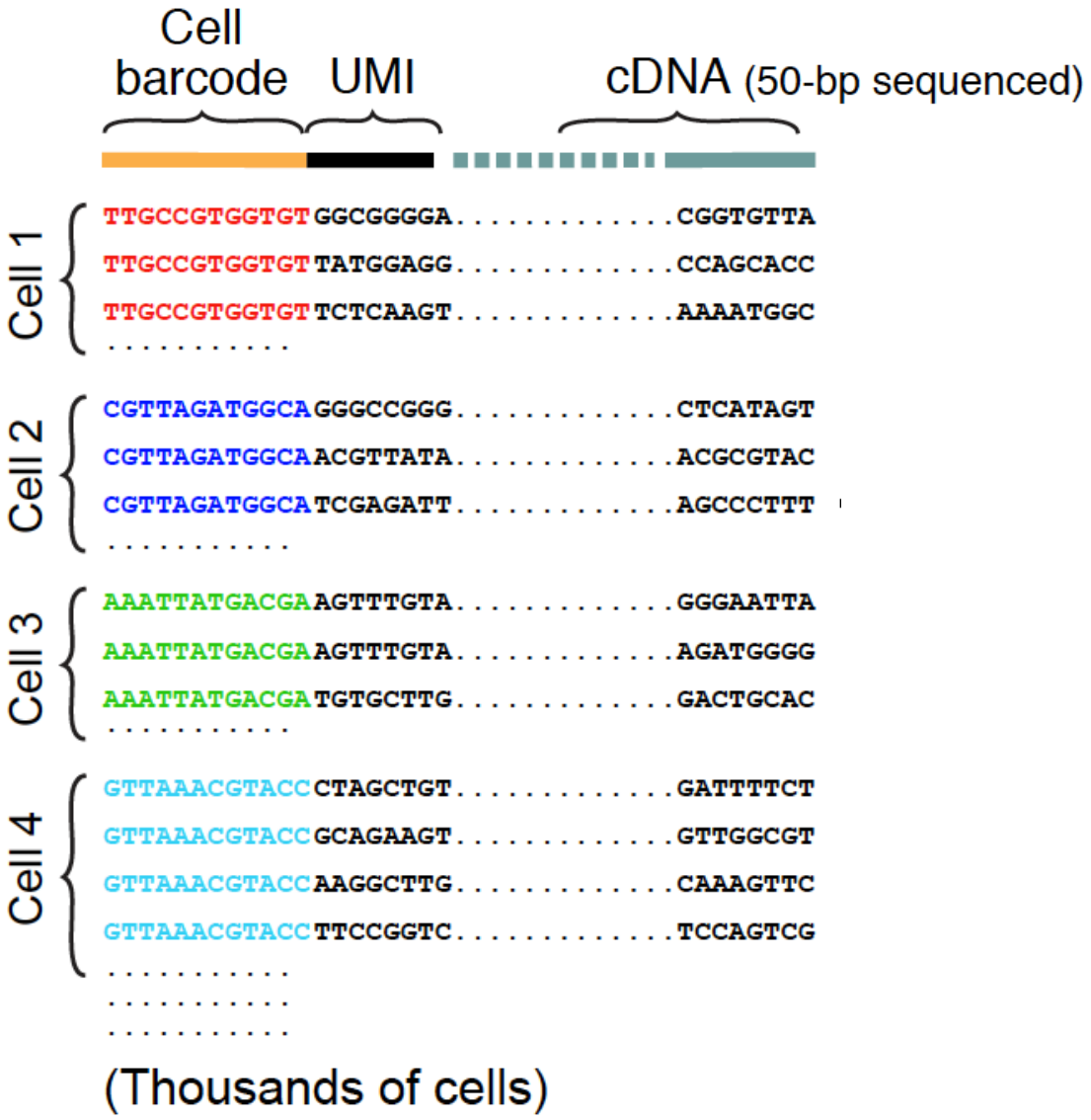
(Hundreds of millions of reads)

Raw reads



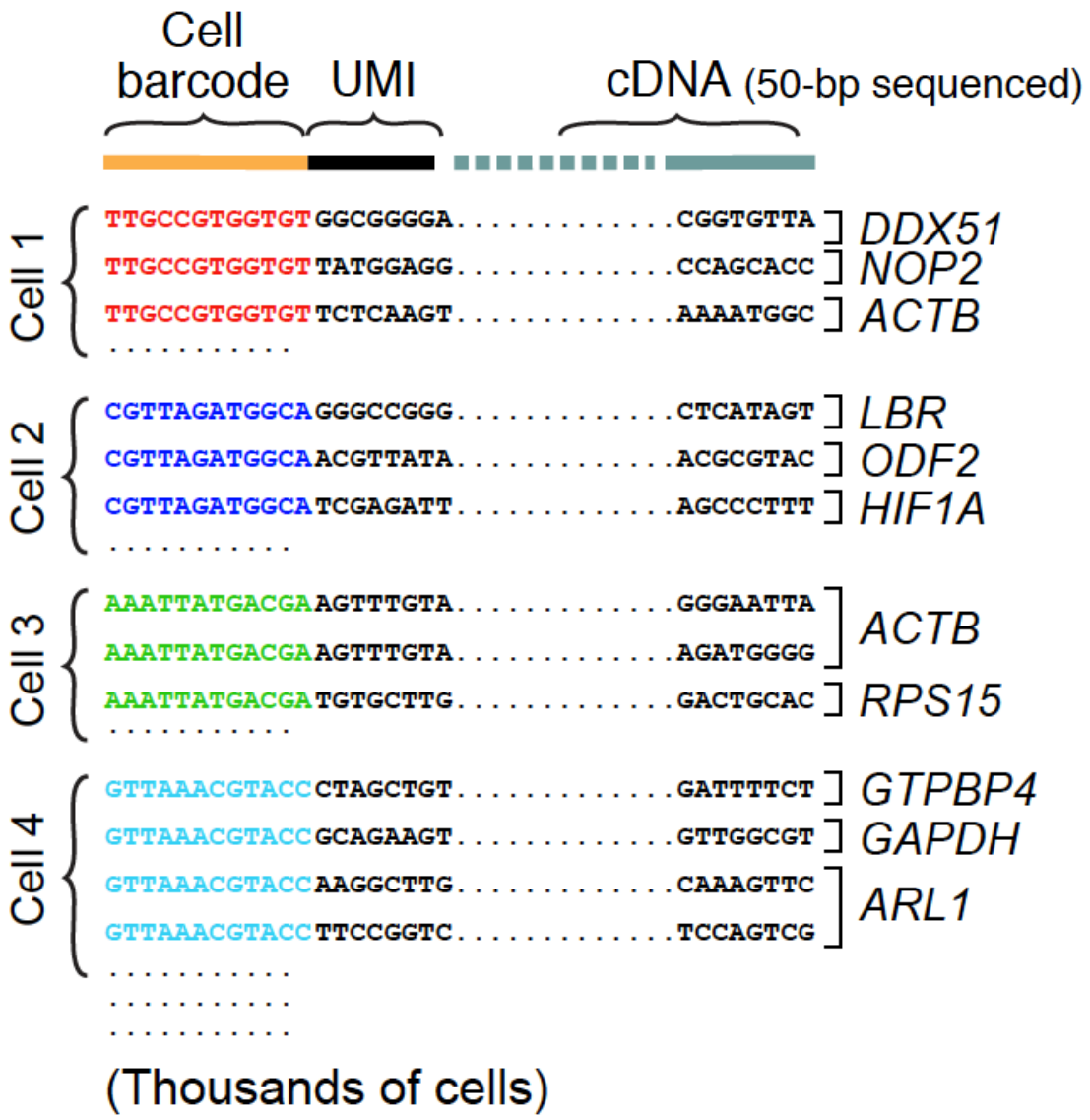
Remove reads with low phred quality at the barcodes

Group reads by cell barcode



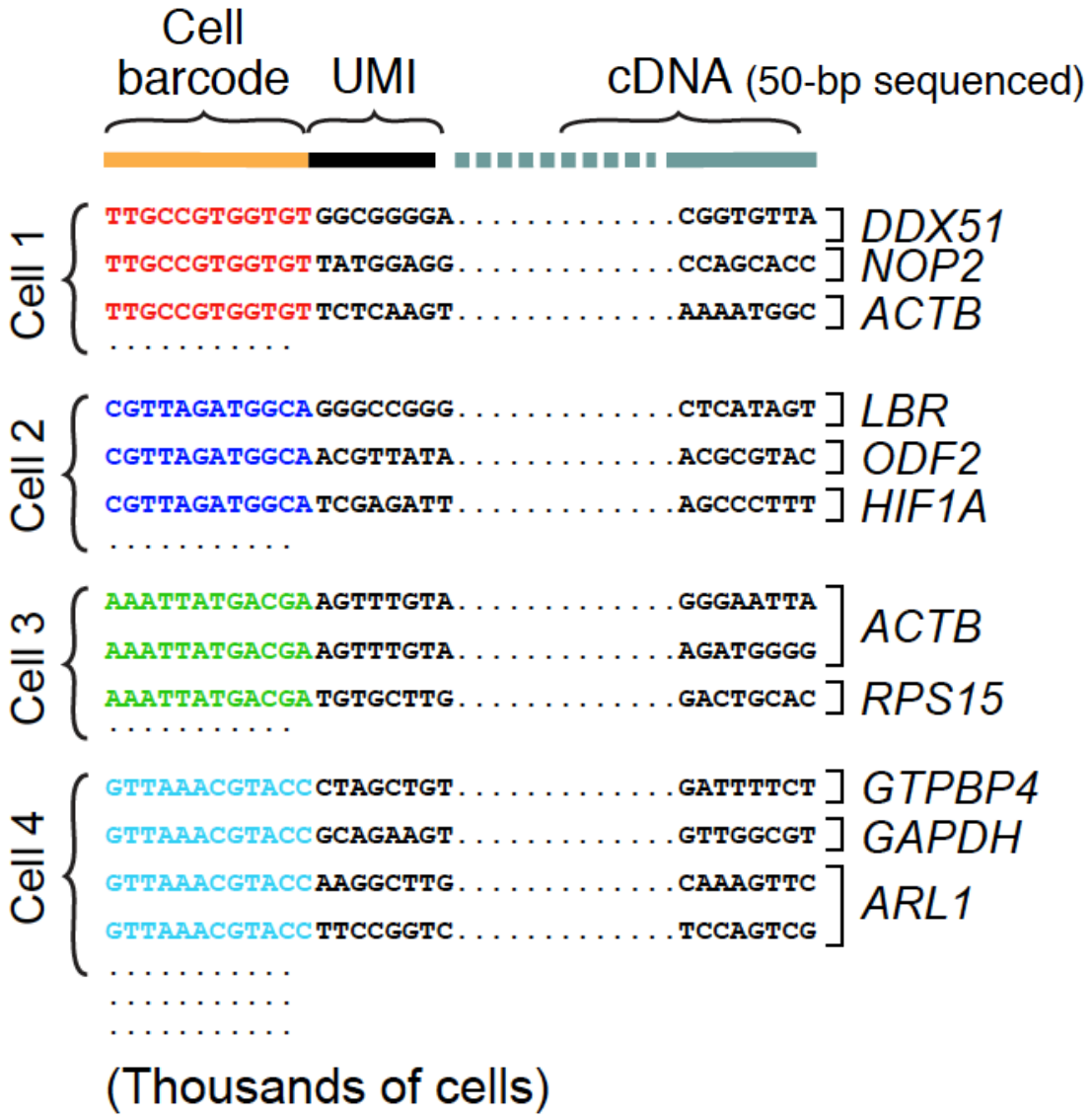
Some platforms incorporate "error-correcting" barcodes which makes the pipeline robust to sequencing errors

Align reads to the genome using a splice-aware aligner

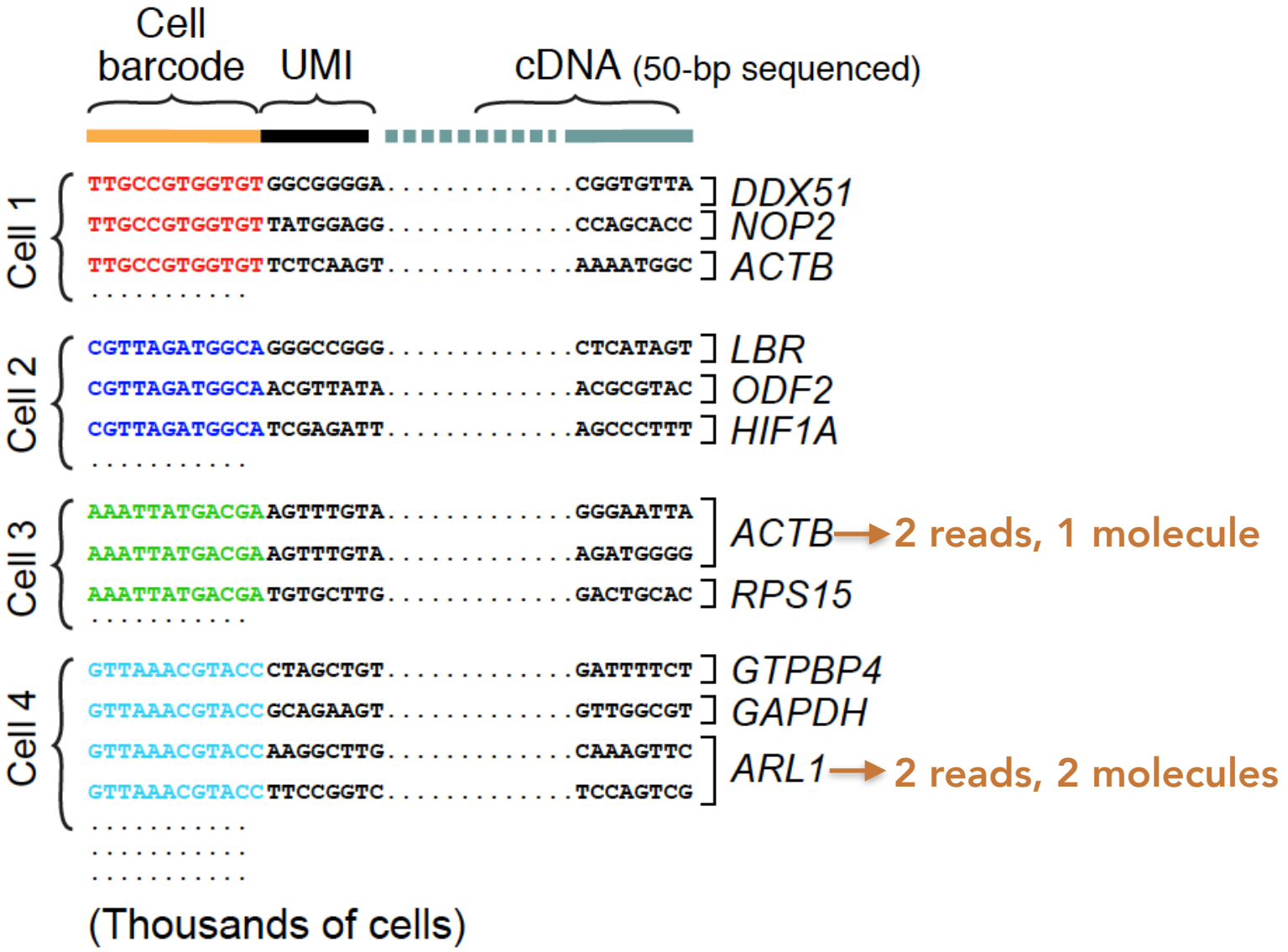


A lot of pipelines use the STAR aligner, which consumes a lot of memory, but is EXTREMELY fast

Collapse UMIs to count transcripts



Collapse UMIs to count transcripts



Gene expression matrix

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Gene expression matrix

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Could also be,

- Proteins
- Genomic Peaks
- SNPs
- …

Gene expression matrix

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Could also be,

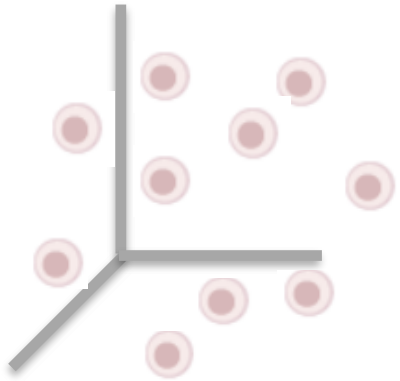
- Proteins
- Genomic Peaks
- SNPs
- …

I have a gene expression matrix, what do I do with it?

Agenda

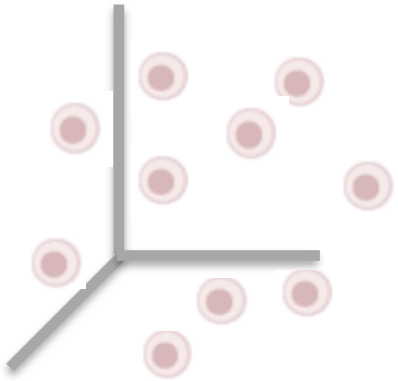
- Single cell analysis - why?
- A short survey of scRNA-seq methods
- Quality comparison of different methods and power analysis
- **Overview of computational workflow**
  - Preprocessing
  - **Secondary analysis in R**
- Some example applications
- Future

Every cell is a vector in gene expression space



Wagner et al., *Nature Biotechnology* 2016

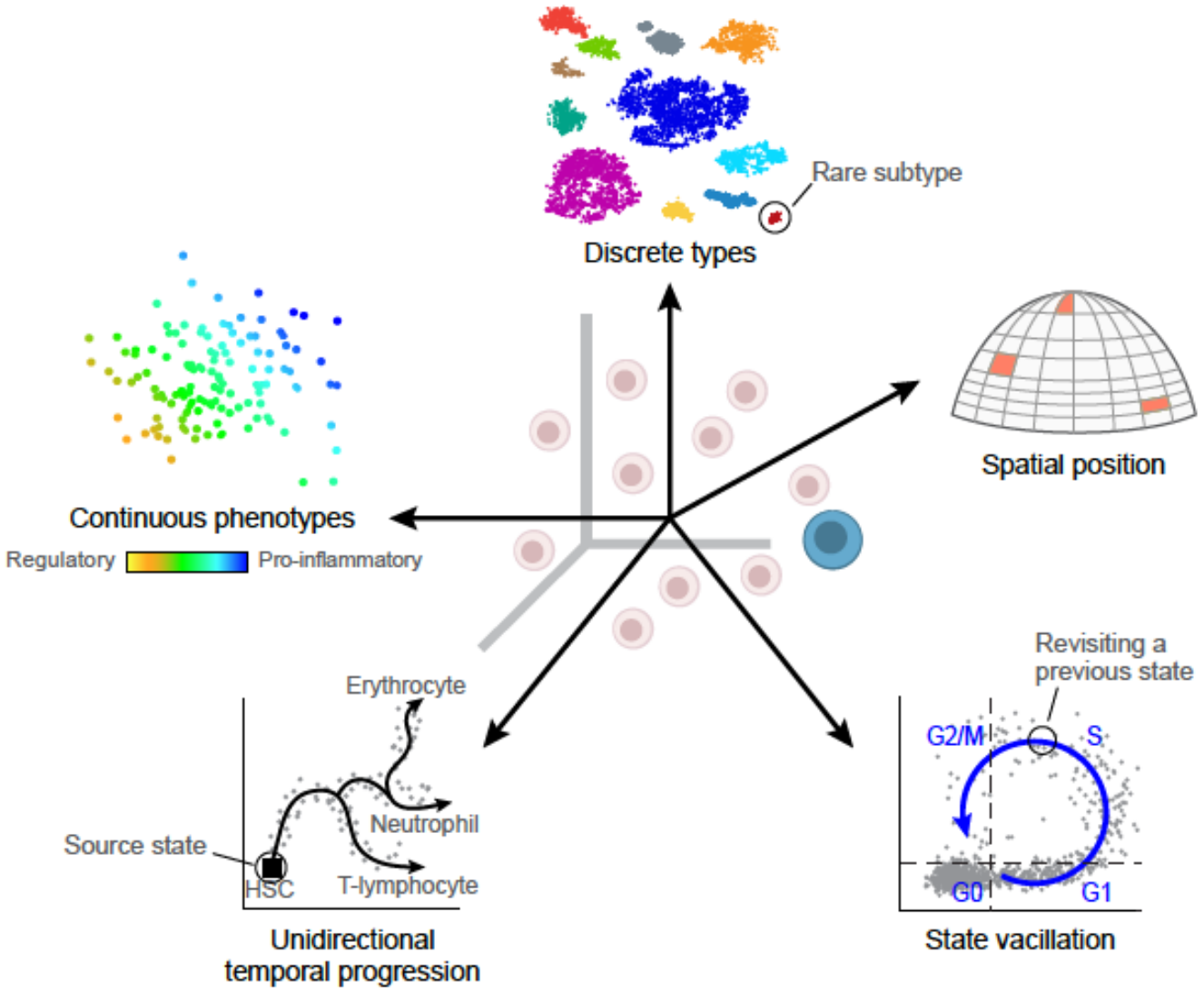
Every cell is a vector in gene expression space



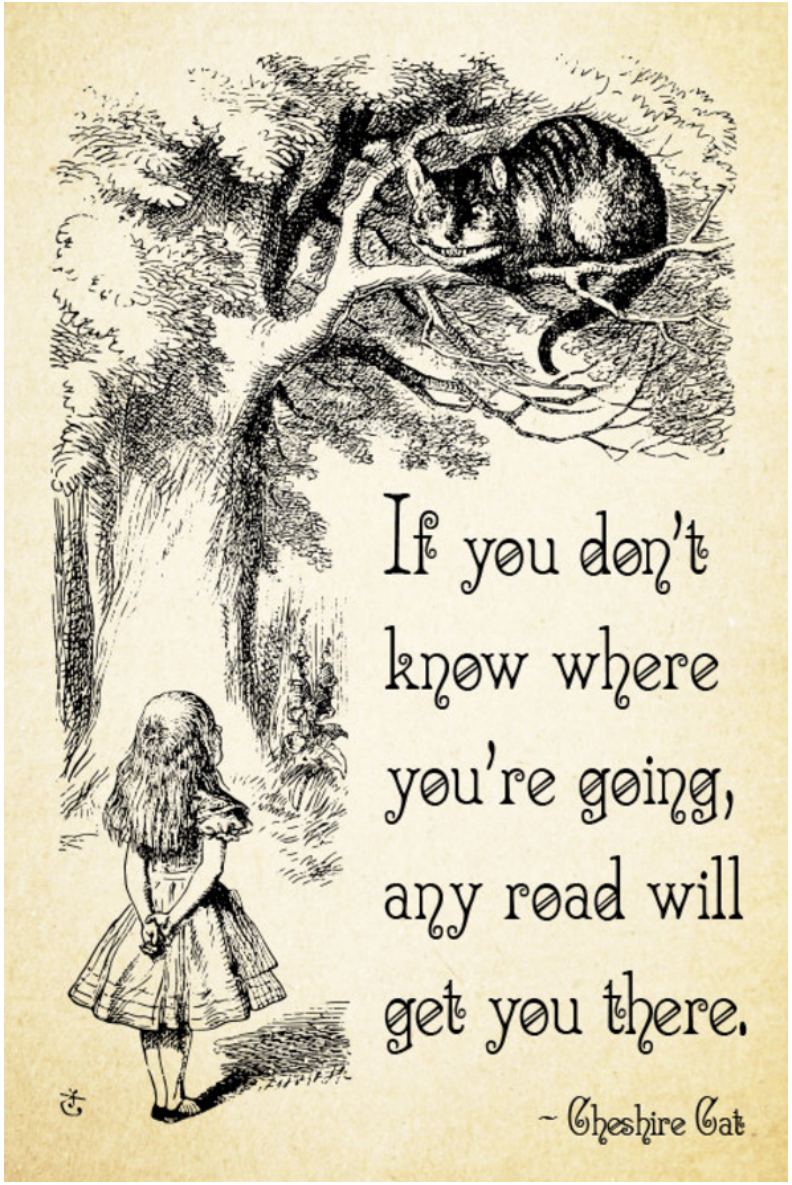
This is a 20,000 dimensional space!

Wagner et al., *Nature Biotechnology* 2016

A variety of questions can be asked …

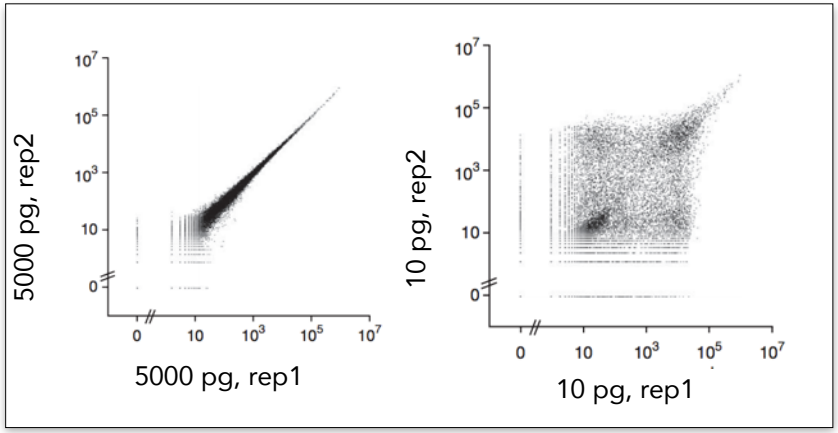


Wagner et al., *Nature Biotechnology* 2016


If you don't know where you're going, any road will get you there.
~ Cheshire Cat

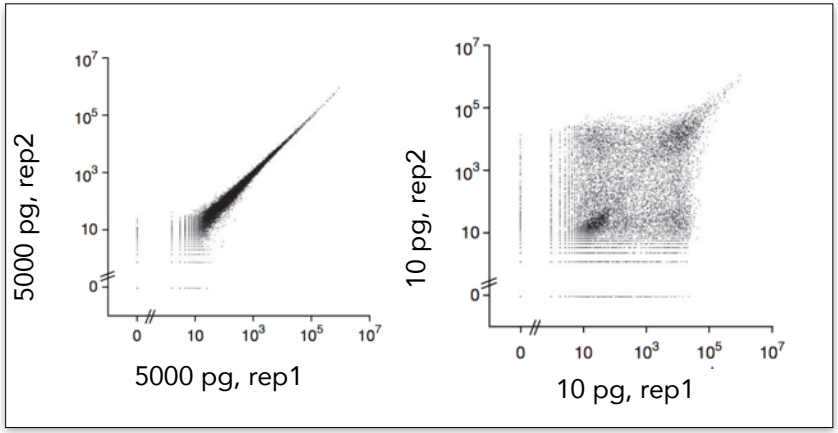
Technical/conceptual challenges in single-cell RNA-seq

Dropouts

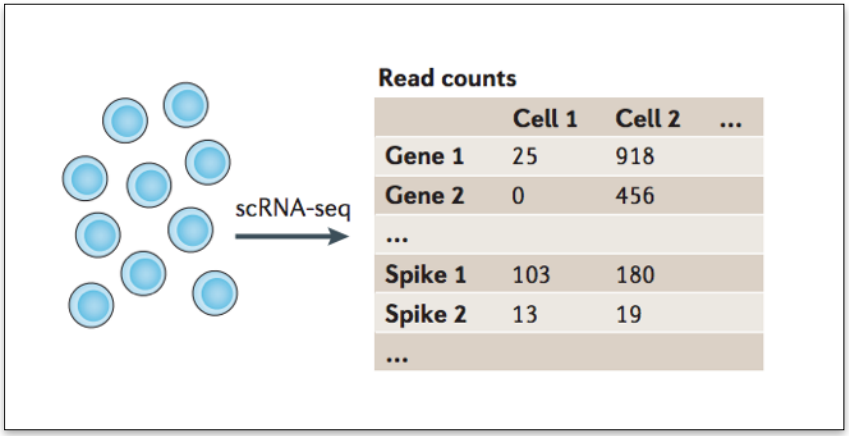


Technical/conceptual challenges in single-cell RNA-seq

Dropouts

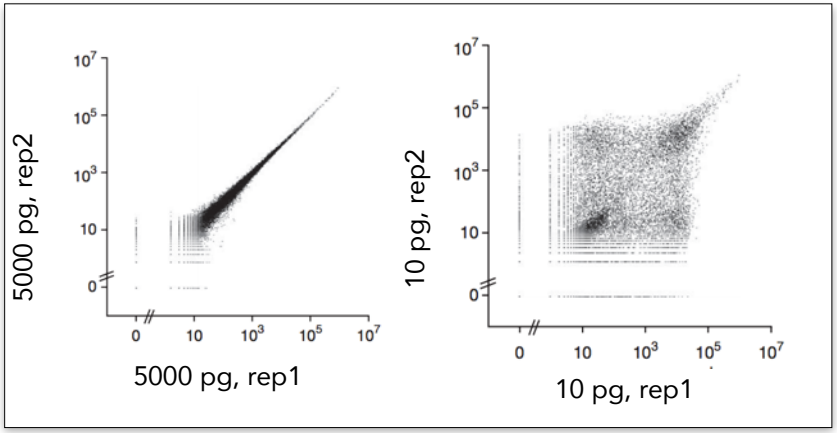


Variation in cell size and quality

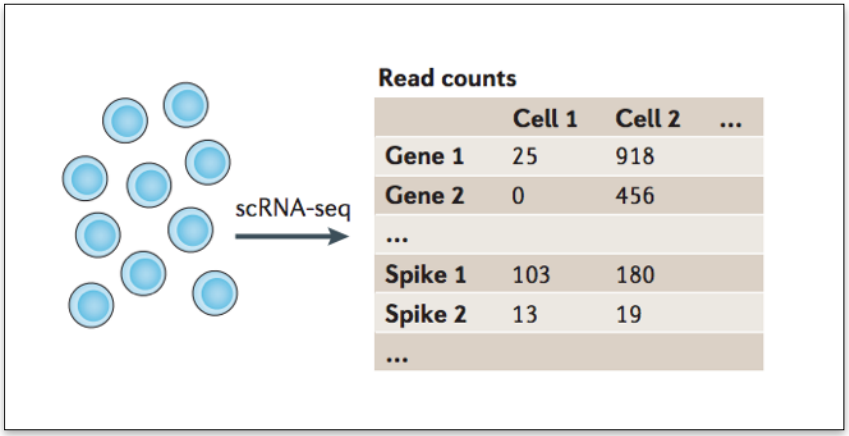


Technical/conceptual challenges in single-cell RNA-seq

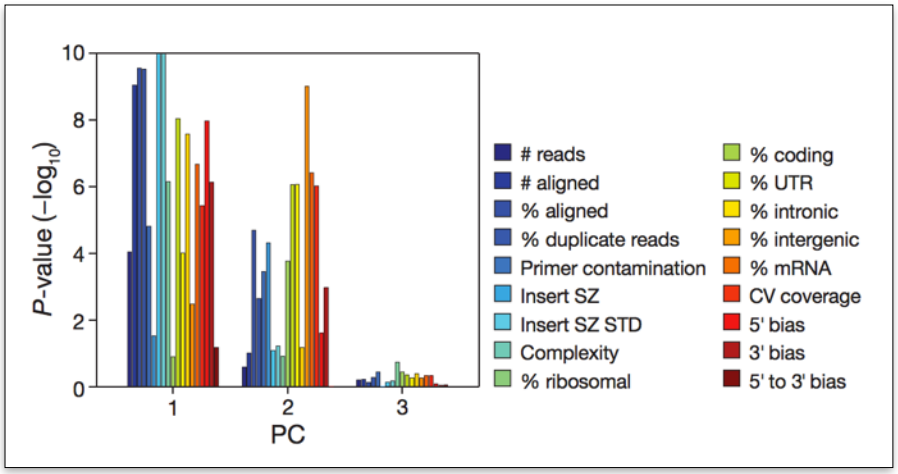
Dropouts



Variation in cell size and quality

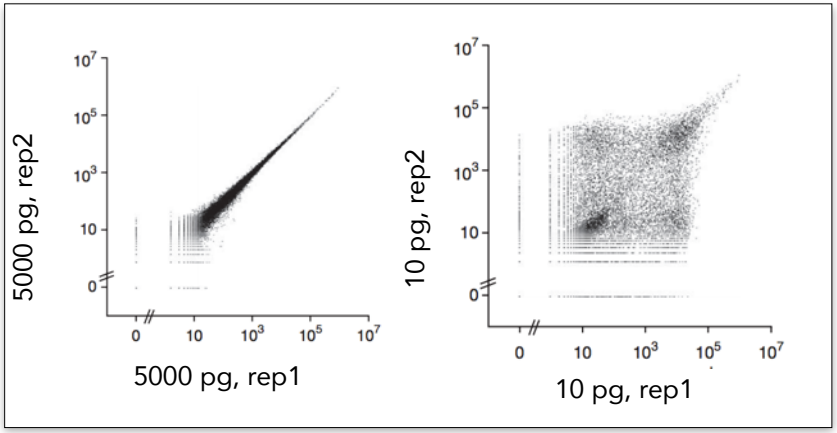


Variation dominated by technical factors

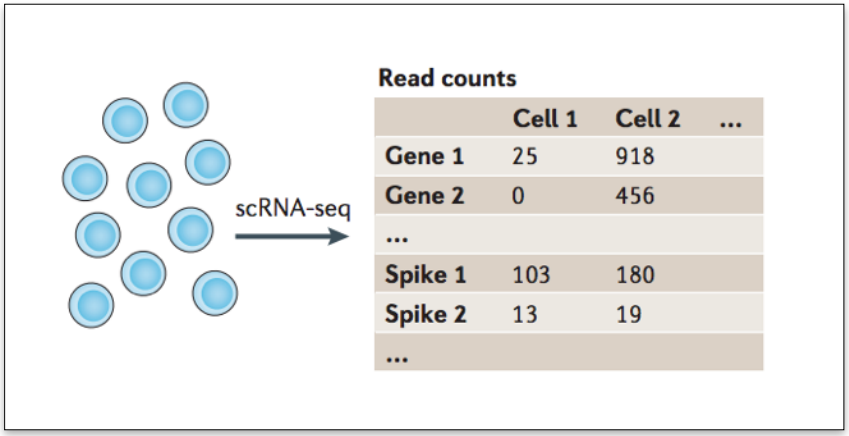


Technical/conceptual challenges in single-cell RNA-seq

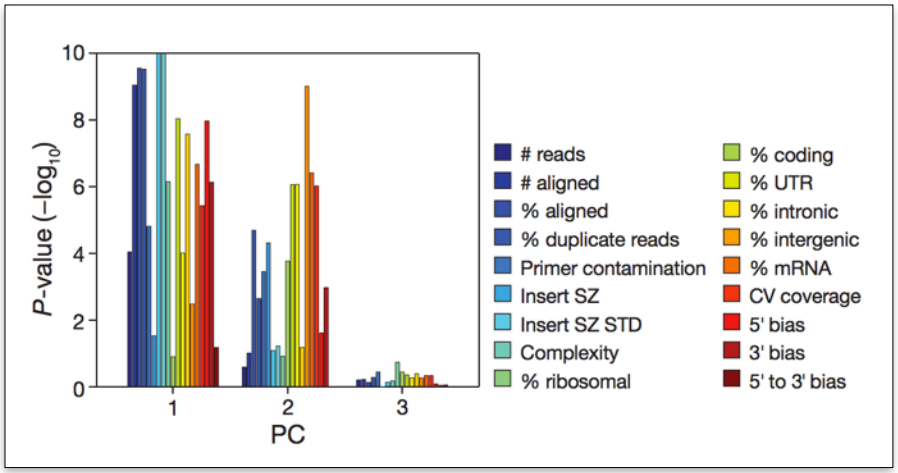
Dropouts



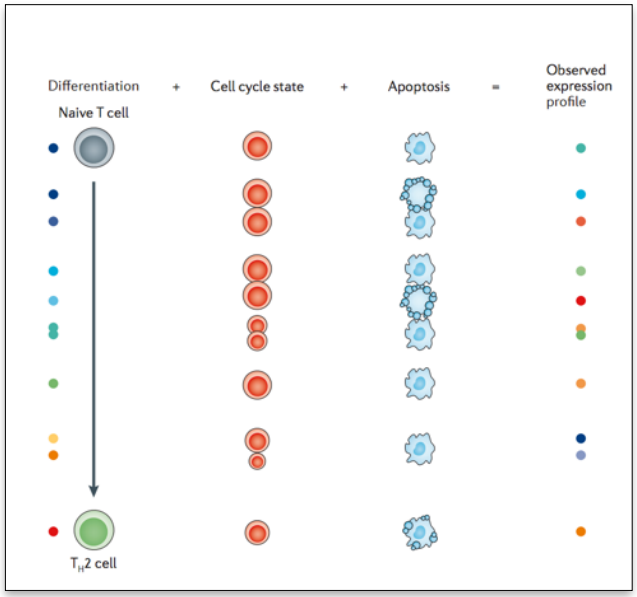
Variation in cell size and quality



Variation dominated by technical factors



Observed gene expression is a convolution



Making sense of variation

- **Fact 1** : For something to be informative, it needs to exhibit variation

Making sense of variation

- **Fact 1** : For something to be informative, it needs to exhibit variation

- **Fact 2** : Not everything that exhibits variation in real life, is informative

1. Identifying relevant, “highly variable” features

First filter out,

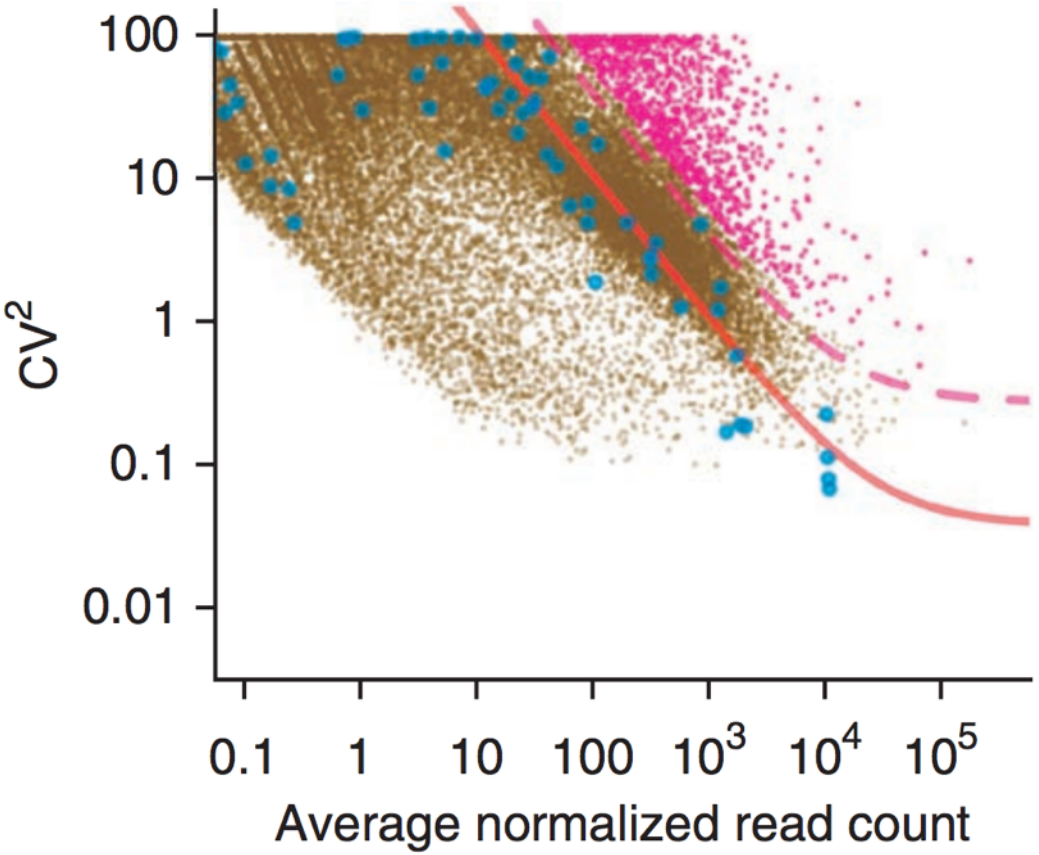
- Lowly expressed genes
- “Housekeeping” genes

1. Identifying relevant, "highly variable" features

First filter out,

- Lowly expressed genes
- "Housekeeping" genes

Typical practice to identify "highly" variable genes is to create a null model of statistical variation based on housekeeping or spike-in genes



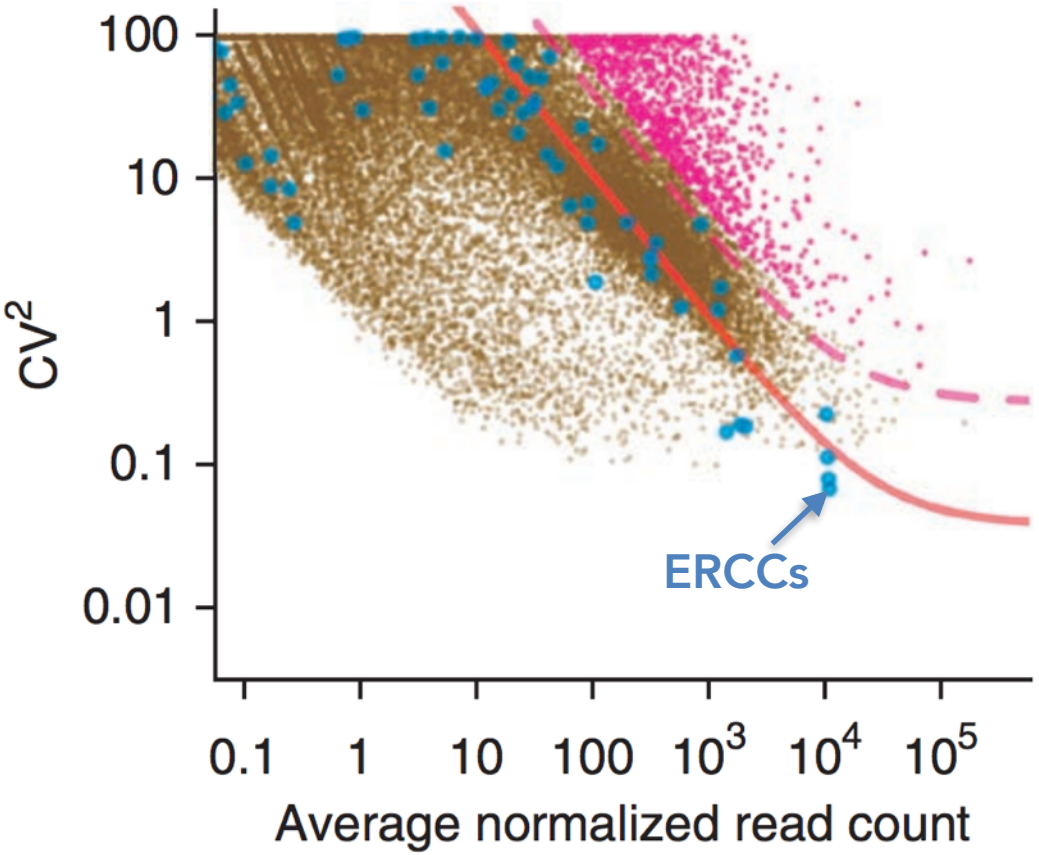
Brennecke et al., *Nature Methods*, 2013

1. Identifying relevant, "highly variable" features

First filter out,

- Lowly expressed genes
- "Housekeeping" genes

Typical practice to identify "highly" variable genes is to create a null model of statistical variation based on housekeeping or spike-in genes



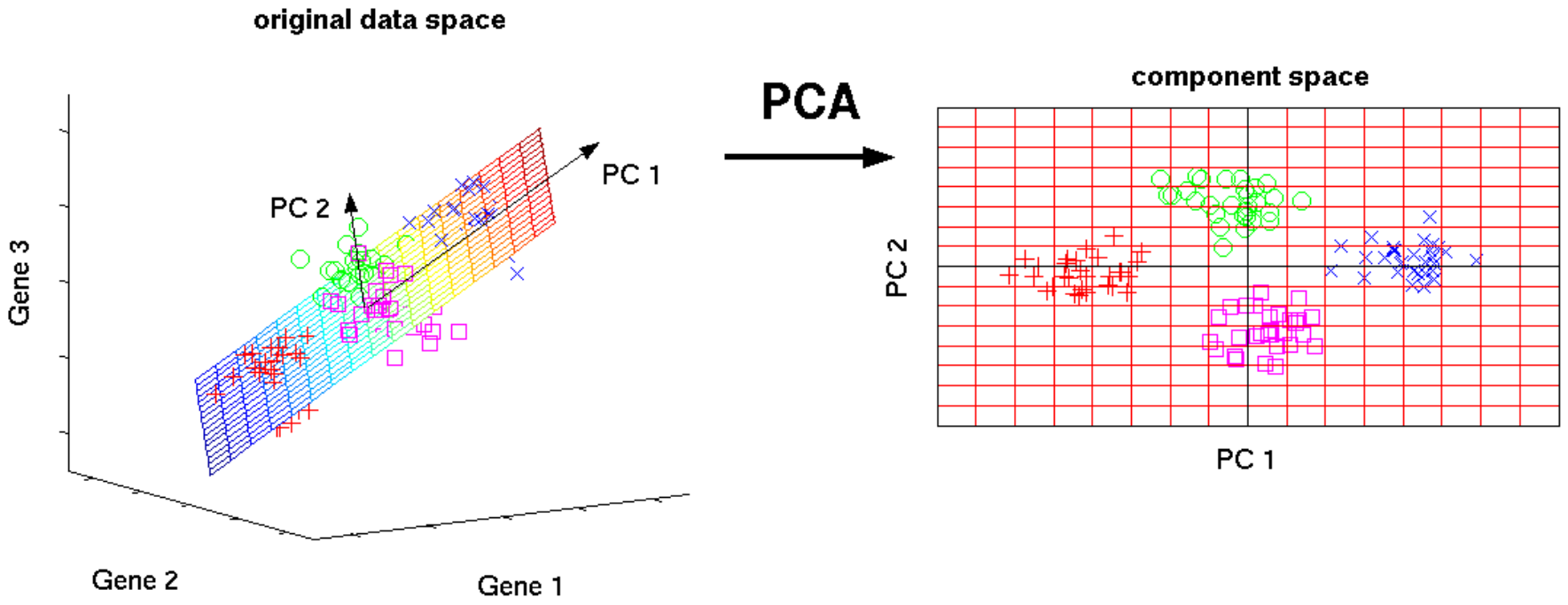
Brennecke et al., *Nature Methods*, 2013

2. Dimensionality reduction

- **Why? :** Genes do not act independently, but as coregulatory “modules”. E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by these modules

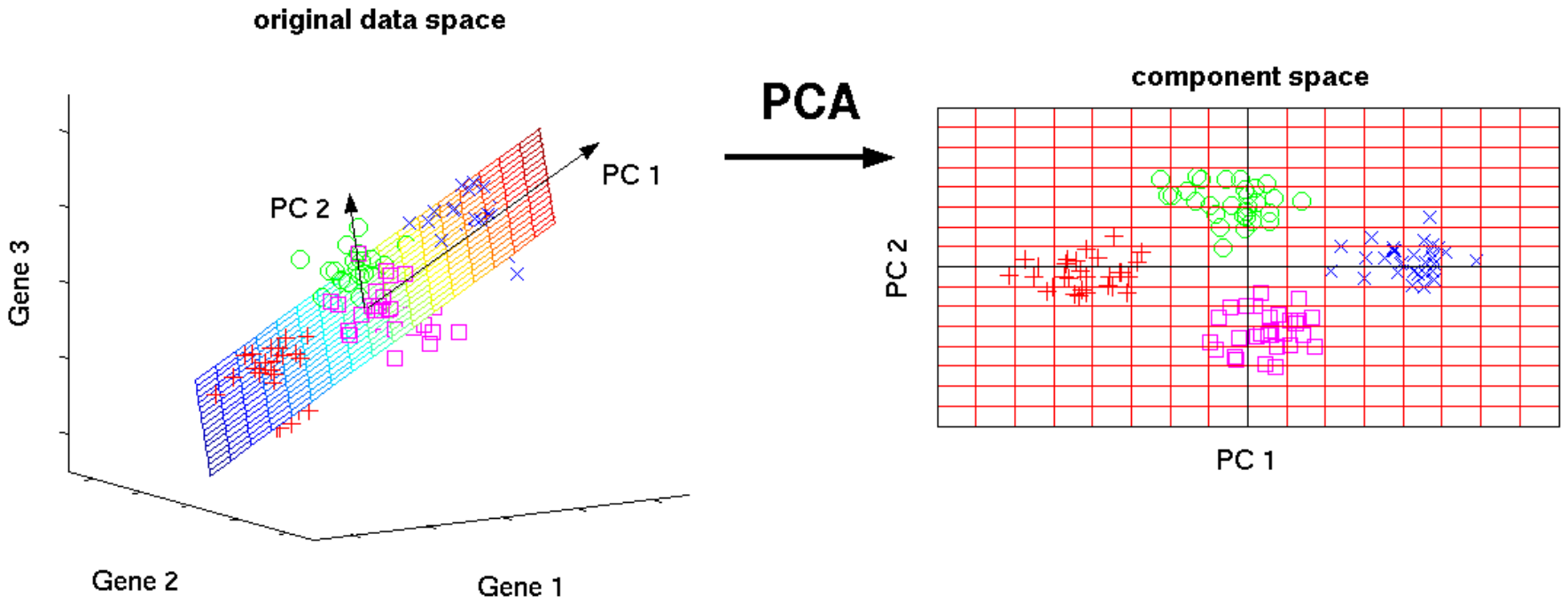
2. Dimensionality reduction

- **Why? :** Genes do not act independently, but as coregulatory "modules". E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by these modules



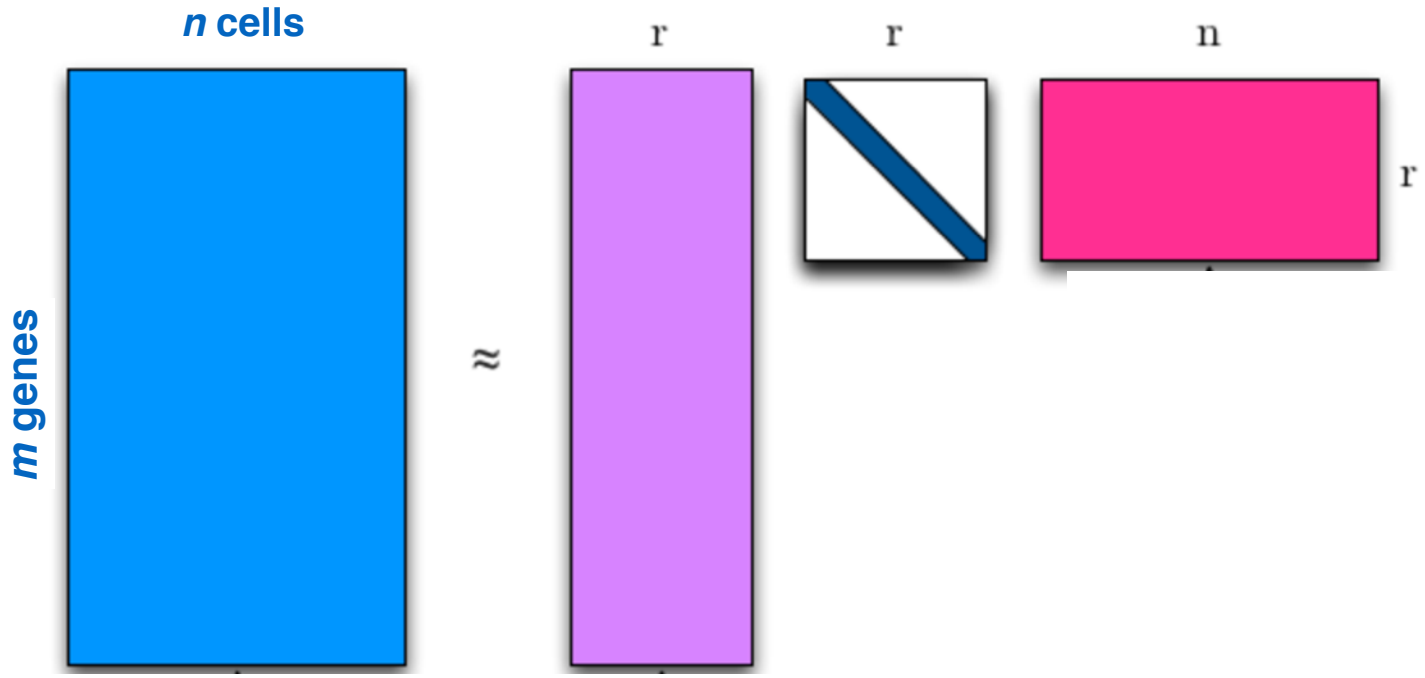
2. Dimensionality reduction

- **Why? :** Genes do not act independently, but as coregulatory "modules". E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by these modules

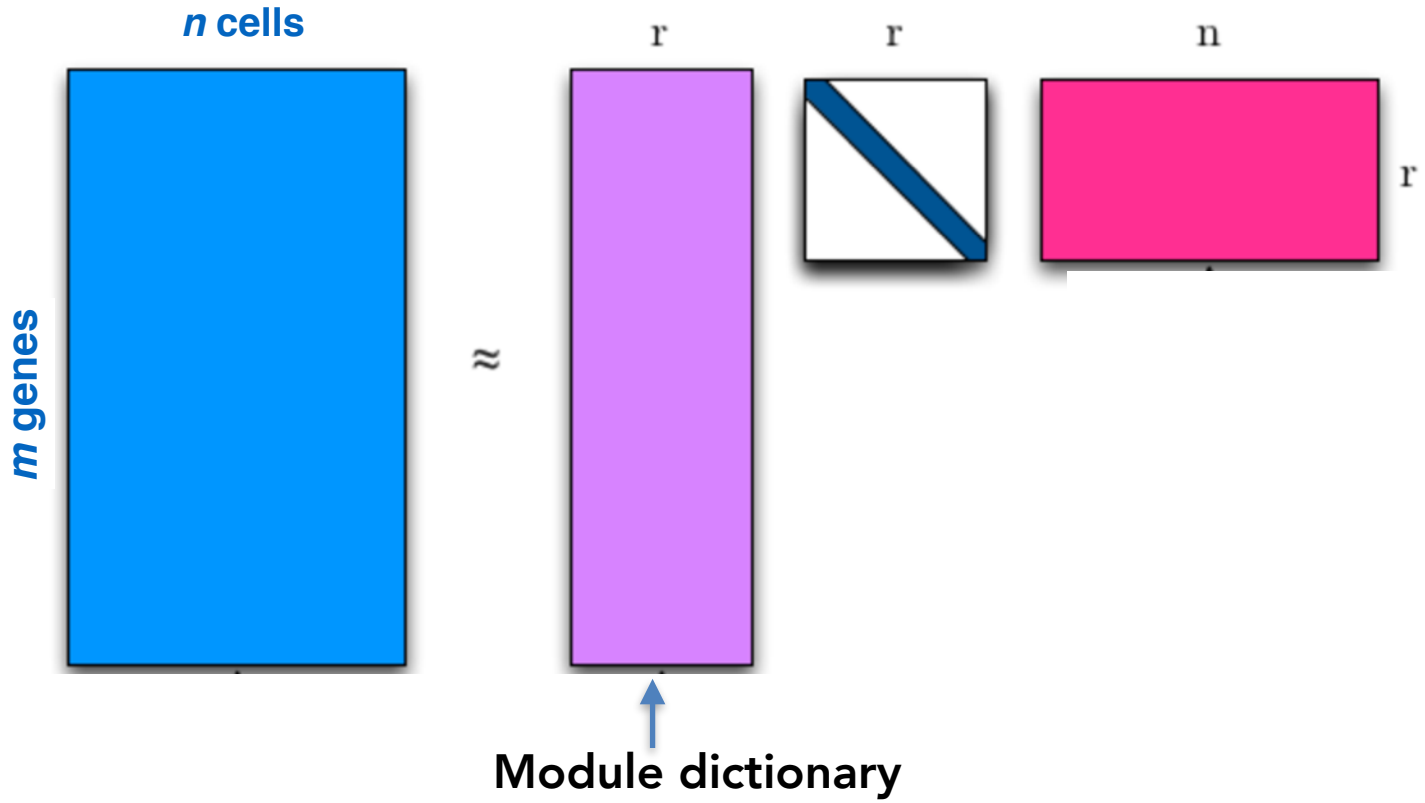


- Principal Component Analysis (PCA) is a **popular linear-method** to identify these modules

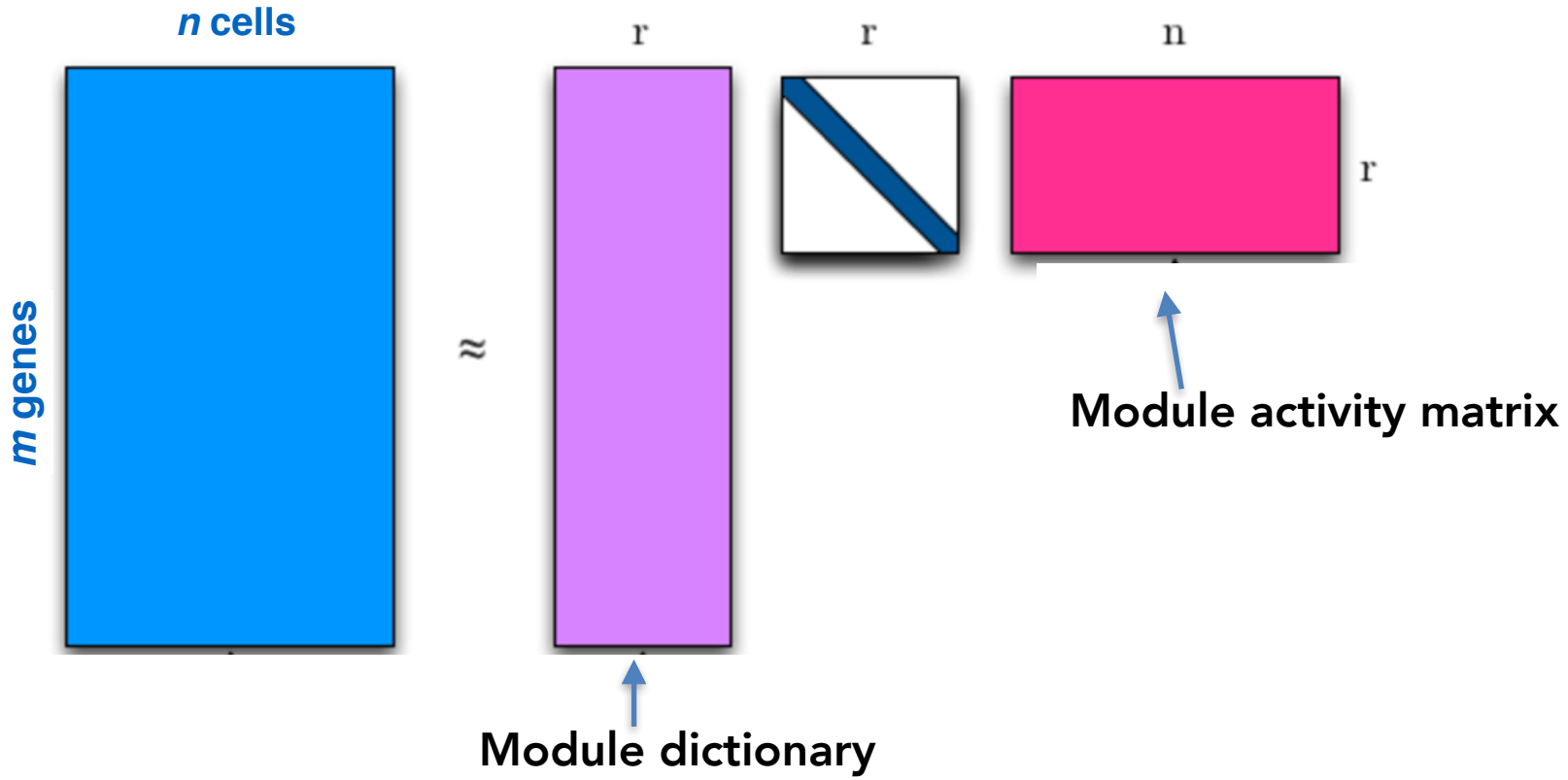
Dimensionality reduction is essentially a matrix factorization problem



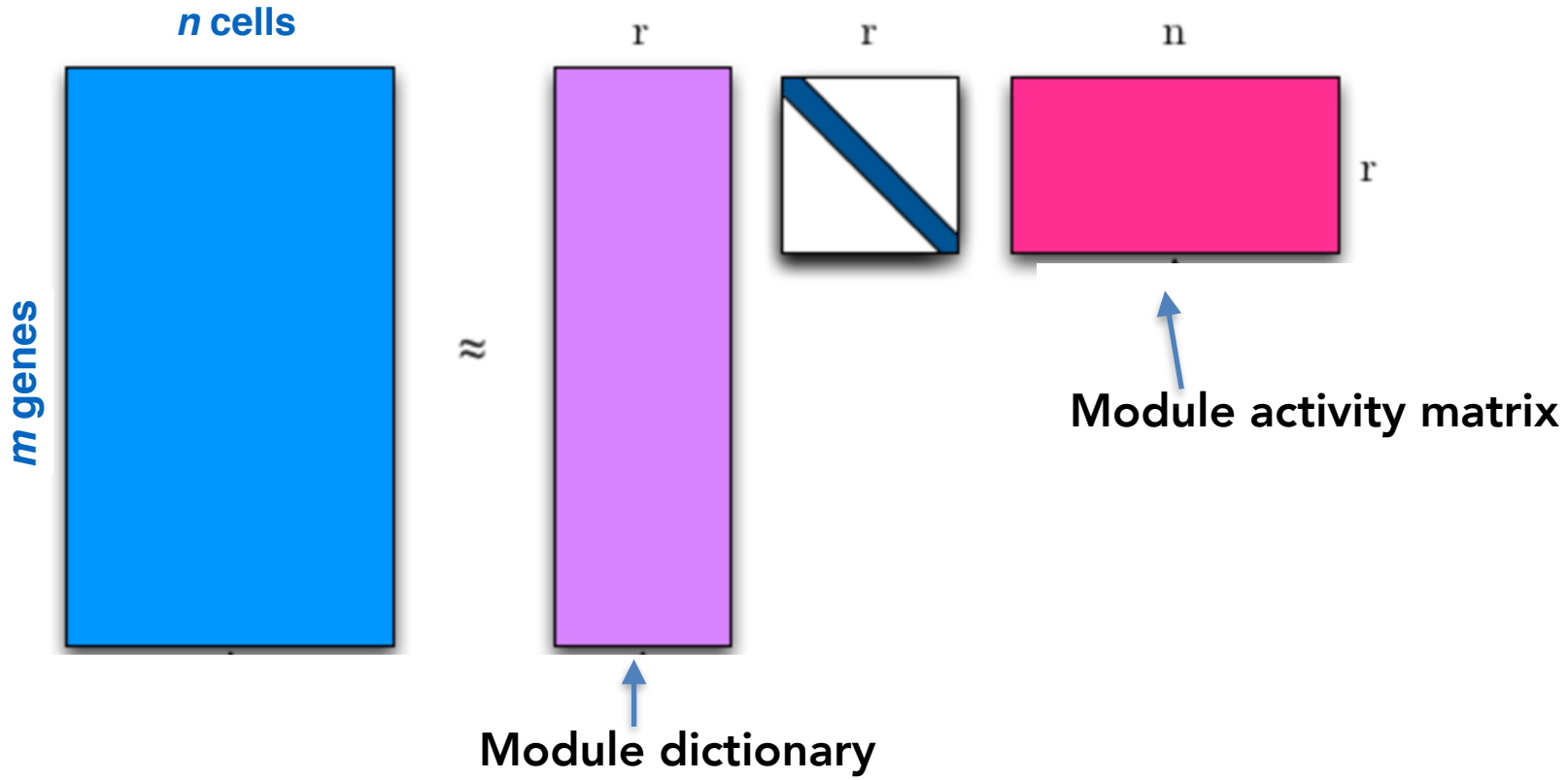
Dimensionality reduction is essentially a matrix factorization problem



Dimensionality reduction is essentially a matrix factorization problem

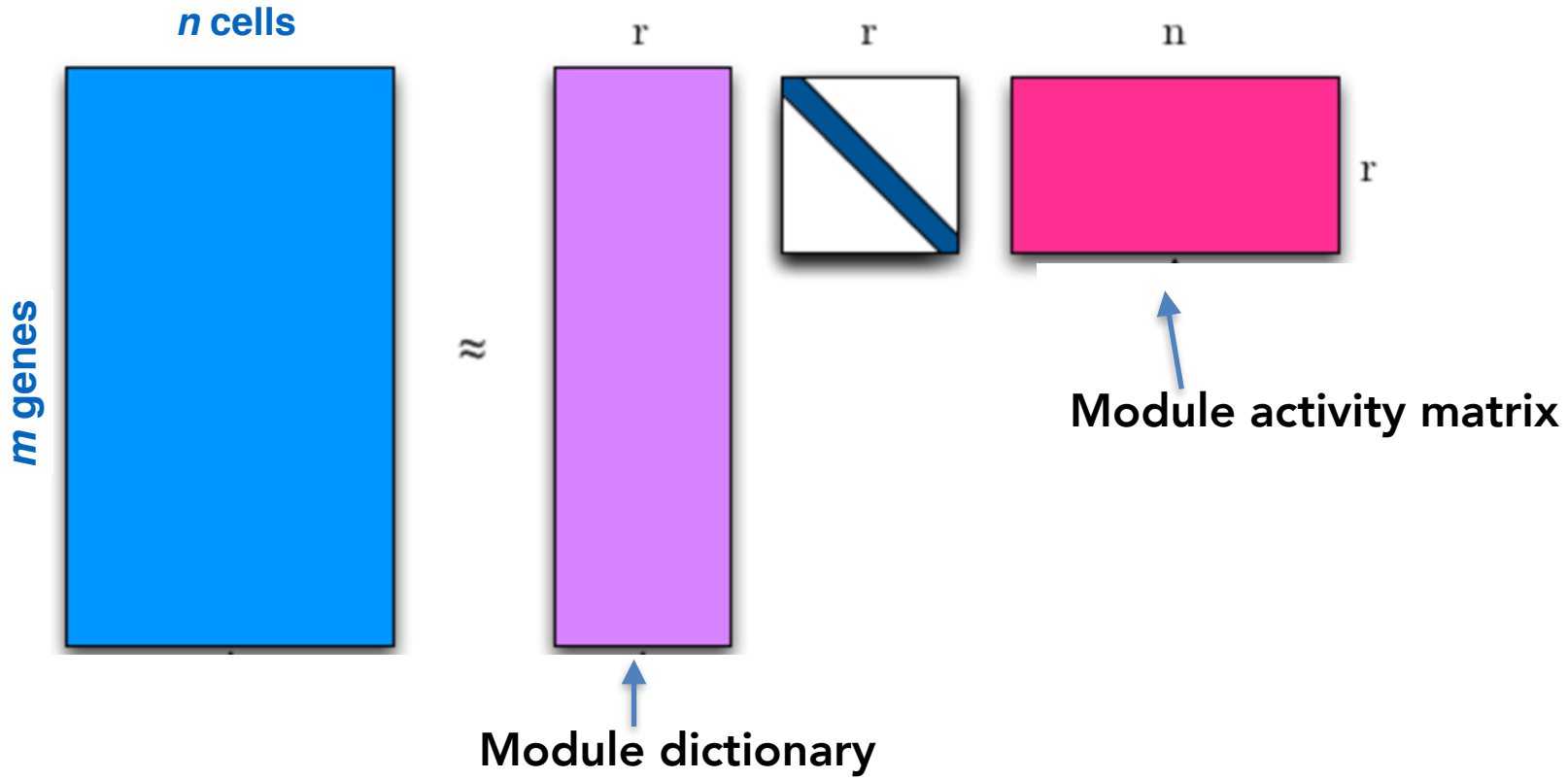


Dimensionality reduction is essentially a matrix factorization problem



- Dimensionality reduction results from the fact that r << m

Dimensionality reduction is essentially a matrix factorization problem



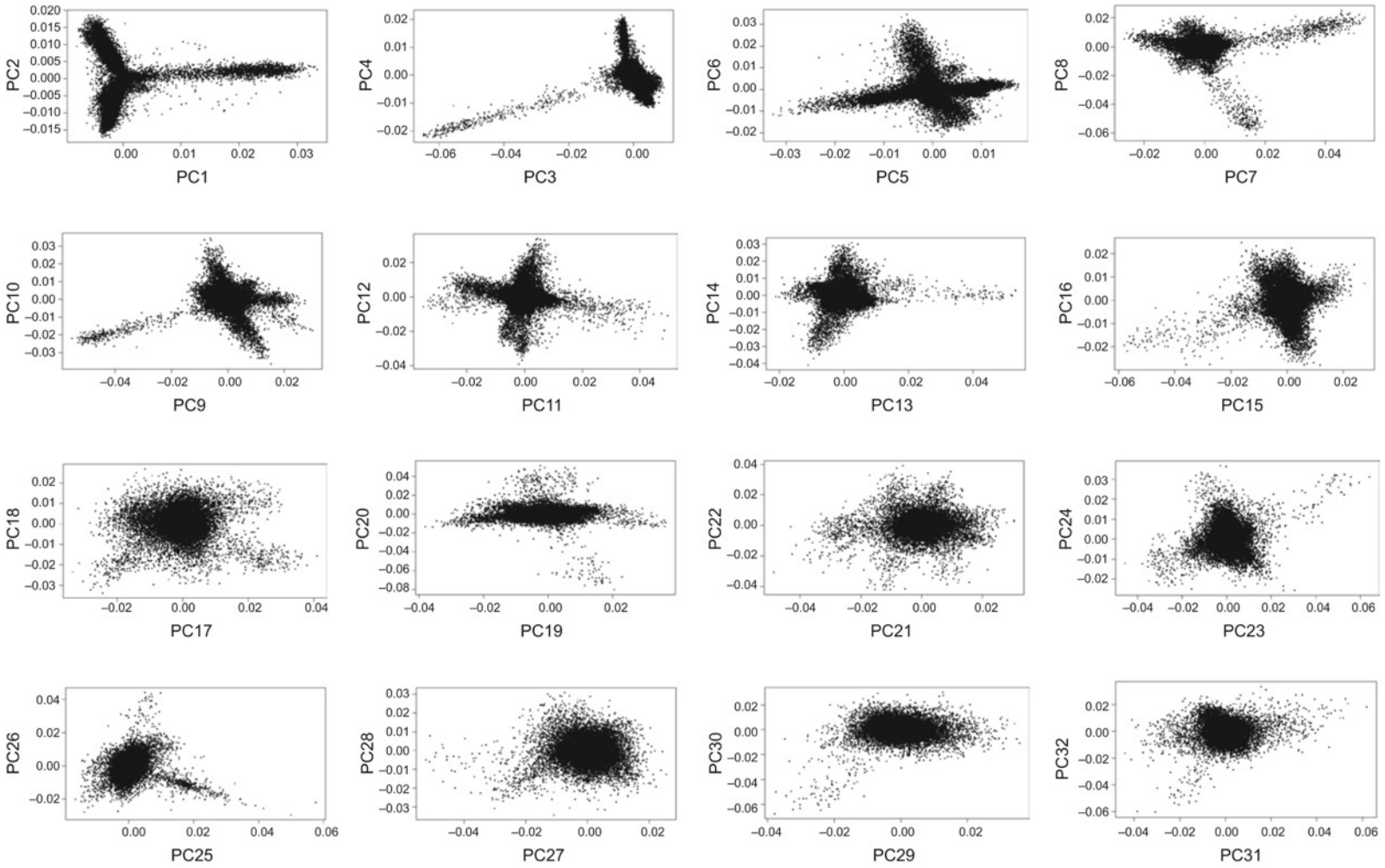
- Dimensionality reduction results from the fact that r << m
- Many methods to factorize matrices are available - **PCA/SVD**, Non-negative matrix factorization (**NMF**), Independent Component Analysis (**ICA**), Hierarchical Poisson Factorization (**HPF**), Latent Dirichlet Allocation (**LDA**) ..

3. Visualization

3. Visualization

- scRNA-seq data is **inherently high dimensional**
- We need visualization tools that "**distill**" key features of the data that we want to explore while remaining "**faithful**" to the overall structure
- PCA compresses the data in terms of **few PCs**, but often there can be more than 3 PCs that capture meaningful variation, making it difficult to visualize

PCs - Drop-seq data



Notice how lower PCs look more and more "spherical" - this loss of structure indicates that the variation captured by these PCs mostly reflects noise

t-distributed Stochastic Neighbor Embedding

- A non-linear embedding method that preserves local distances between data-points in the low-dimensional space

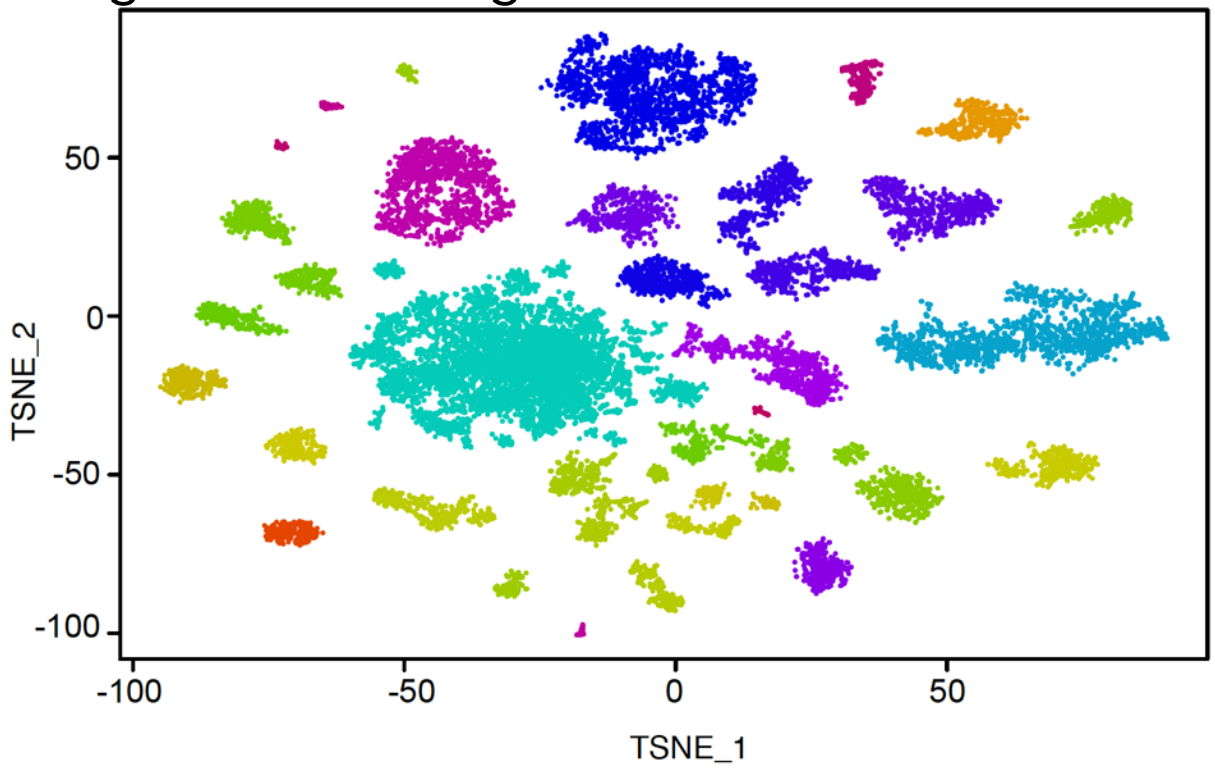
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

t-distributed Stochastic Neighbor Embedding

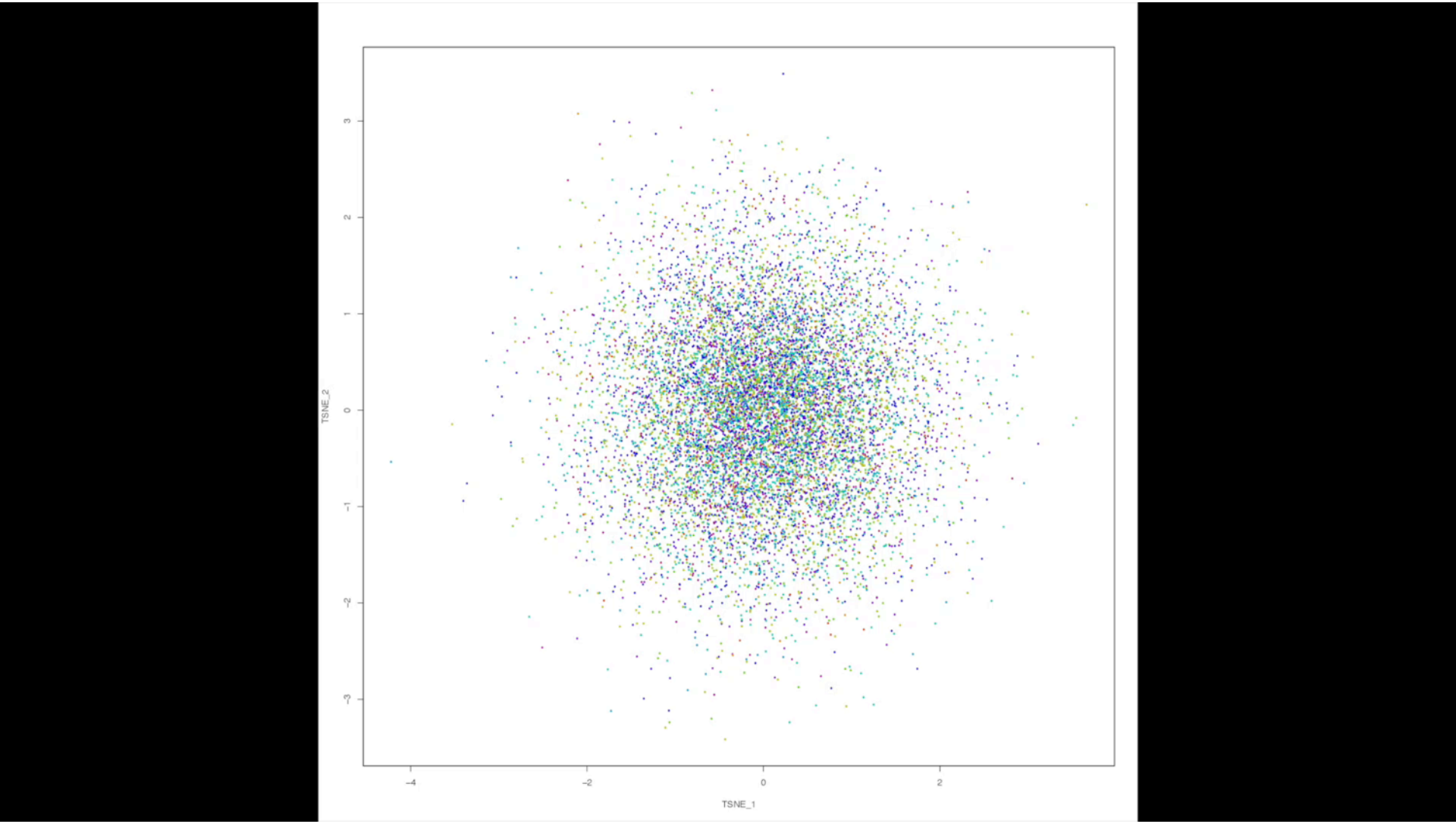
- A non-linear embedding method that preserves local distances between data-points in the low-dimensional space

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

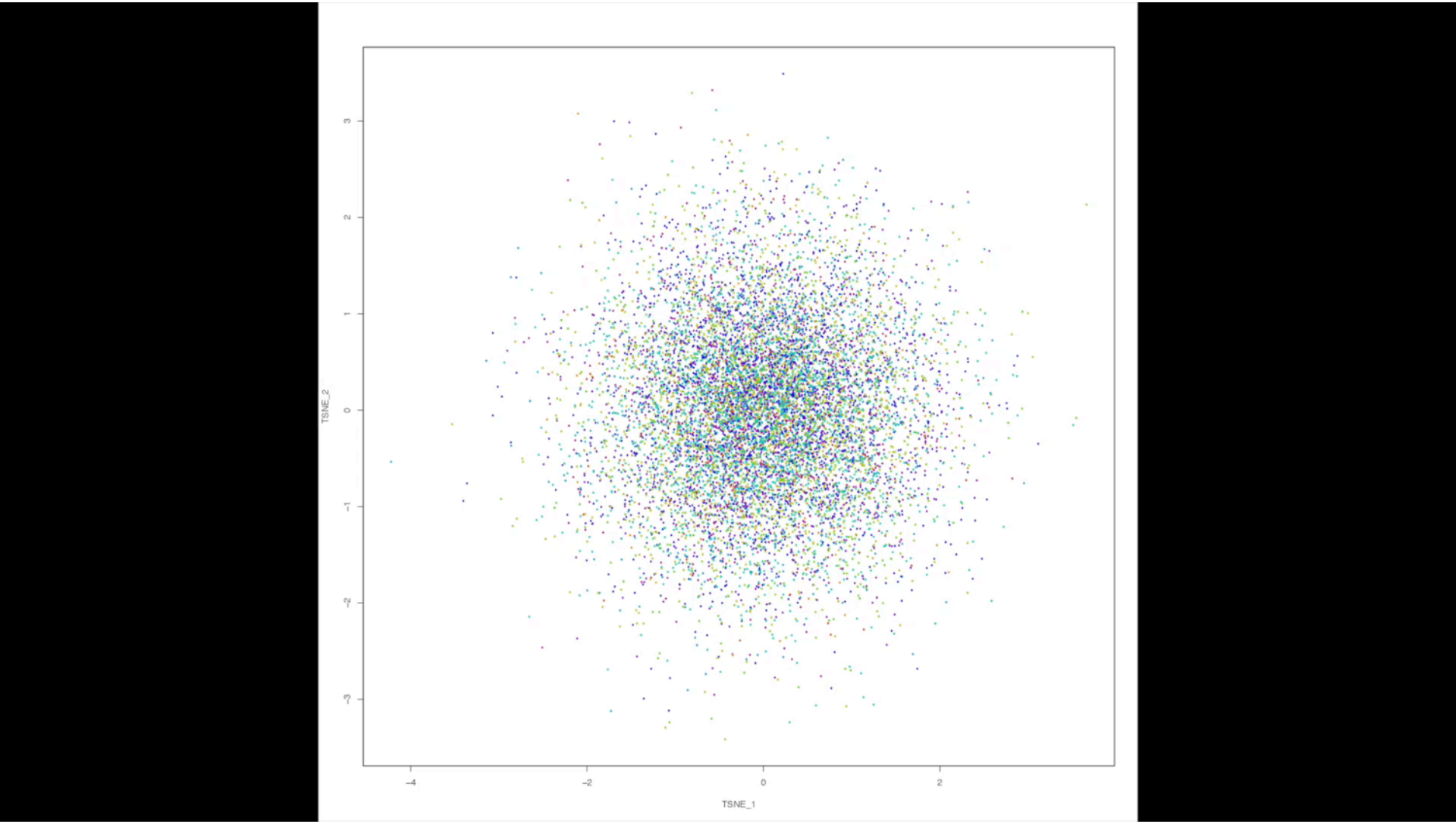
- t-SNE embedding of 45,000 retinal neurons sequenced using Drop-seq and clustered using the DBScan algorithm



Visualizing t-SNE

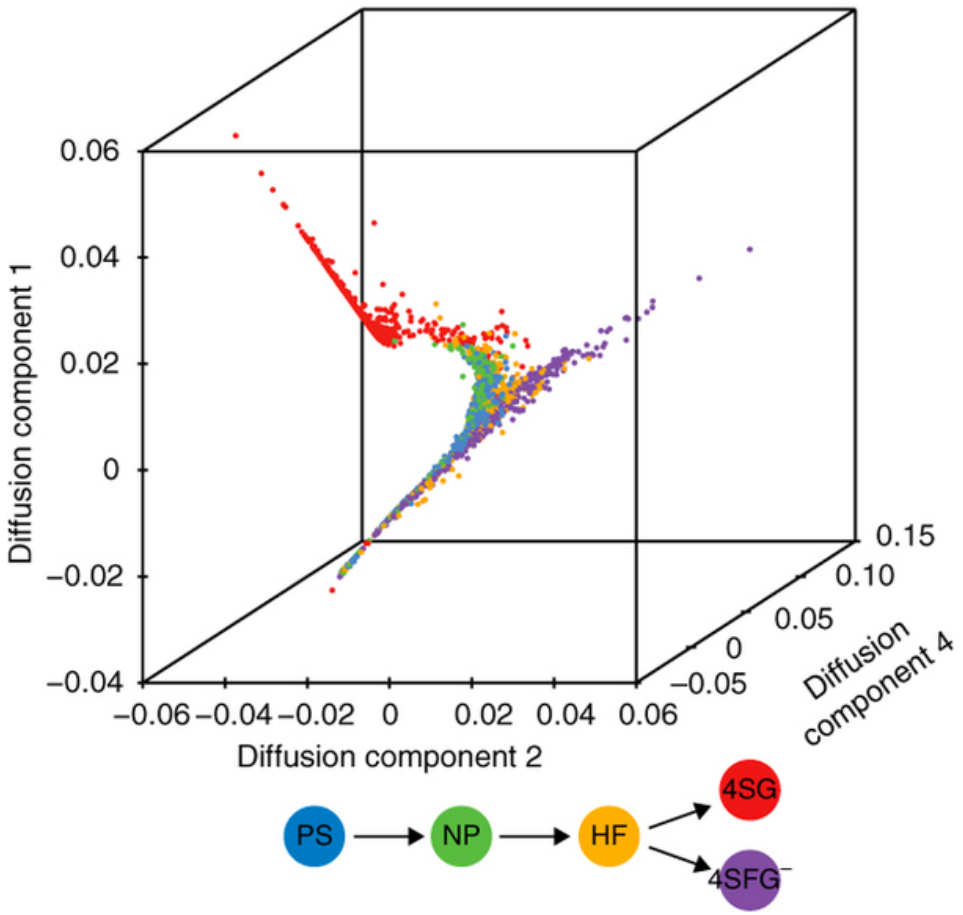


Visualizing t-SNE



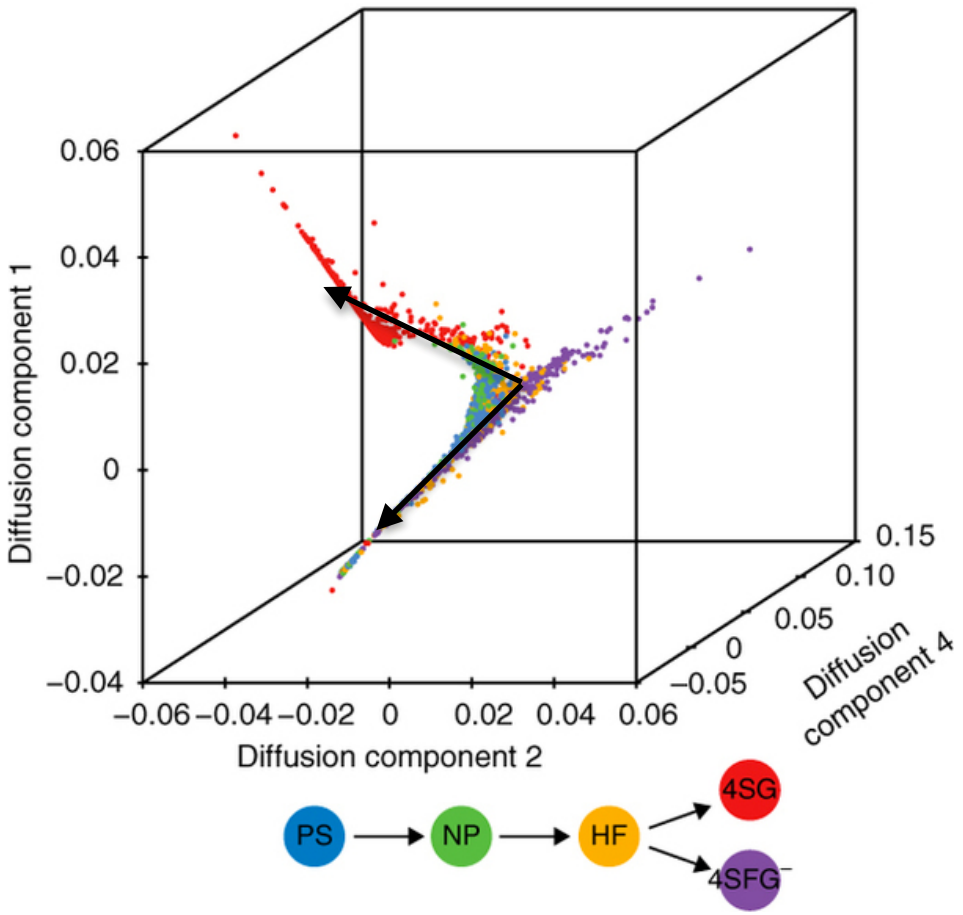
Visualizing Continuous Variation using Diffusion Maps

- Assumes that there exists a "continuous manifold" of states that cells are diffusing through (e.g. during differentiation).
- Diffusion maps construct a **random walk** (Markov process) on the data, and infer a low dimensional representation (DCs) that describes these paths



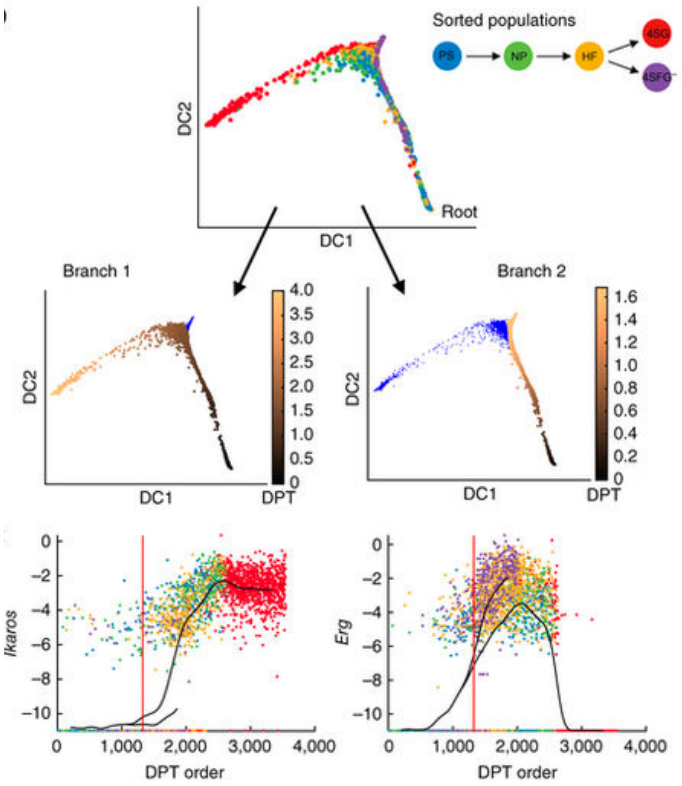
Visualizing Continuous Variation using Diffusion Maps

- Assumes that there exists a "continuous manifold" of states that cells are diffusing through (e.g. during differentiation).
- Diffusion maps construct a **random walk** (Markov process) on the data, and infer a low dimensional representation (DCs) that describes these paths



Visualizing Continuous Variation using Diffusion Maps

- Assumes that there exists a "continuous manifold" of states that cells are diffusing through (e.g. during differentiation).
- Diffusion maps construct a **random walk** (Markov process) on the data, and infer a low dimensional representation (DCs) that describes these paths

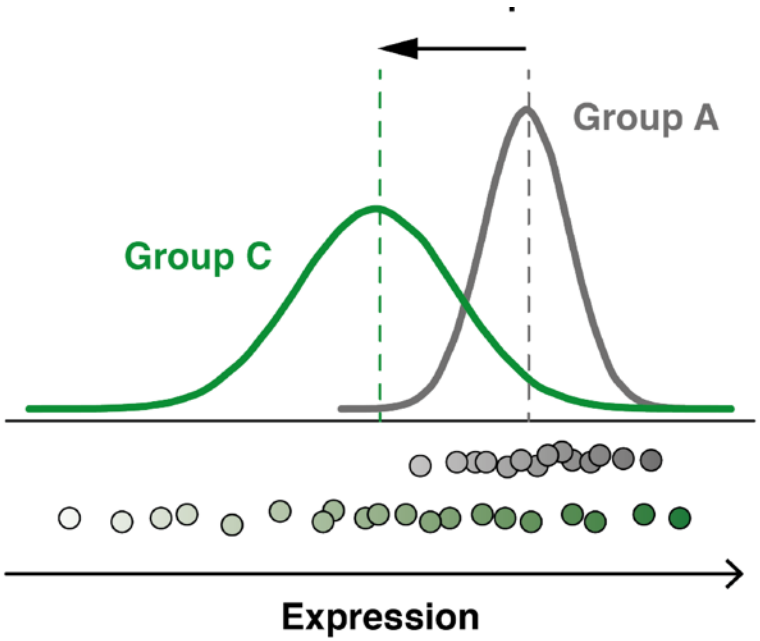


A host of other visualization methods exist

- Locally linear embedding
- Samon mapping
- Multi-kernel visualization
- Graph-abstraction
- IsoMap
- Force-directed layout
- ….

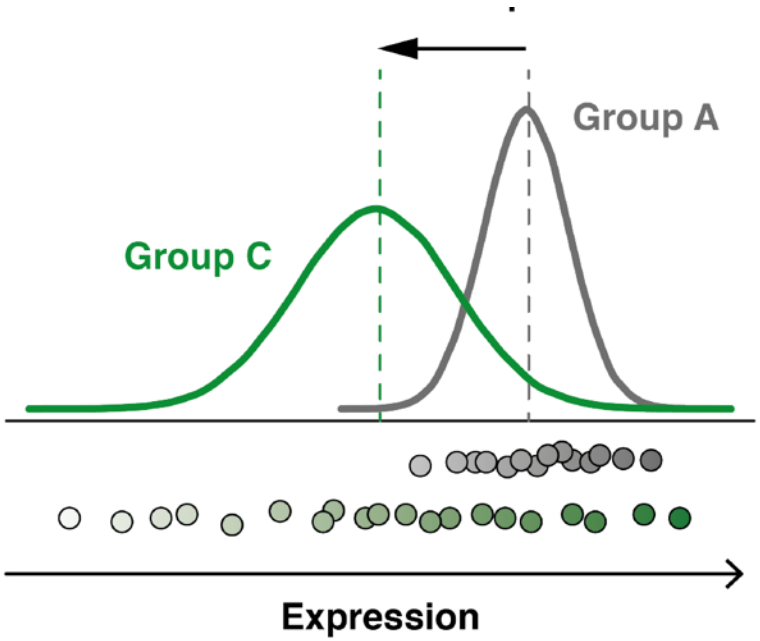
4. Differential expression analysis

Group A > Group B (p-value < 0.01)



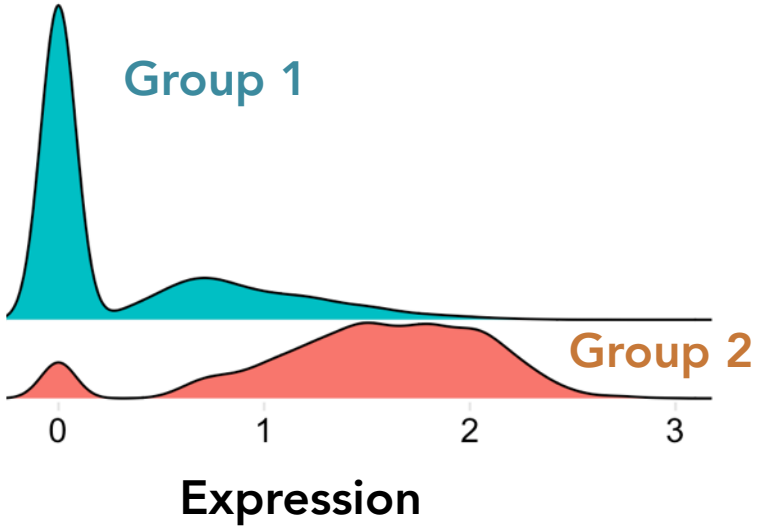
4. Differential expression analysis

Group A > Group B (p-value < 0.01)



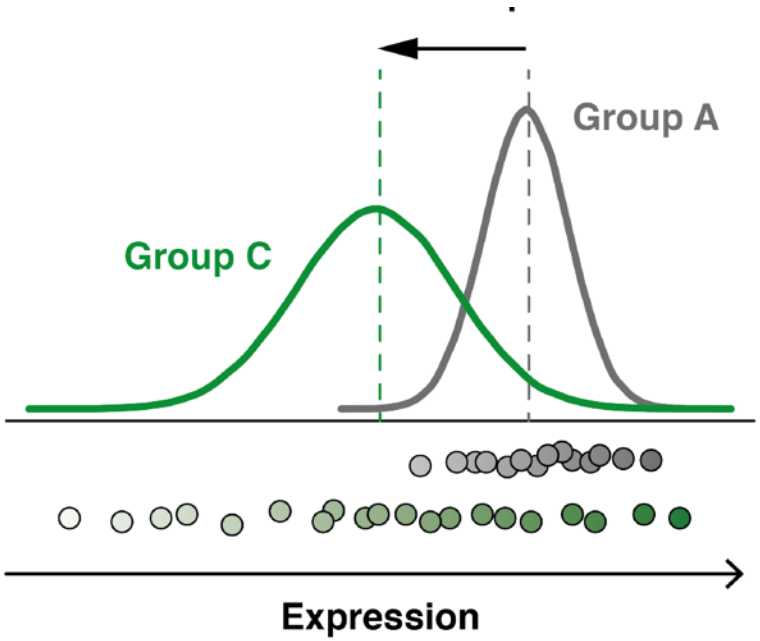
BUT

"Zero inflation" poses a challenge in single-cell data!



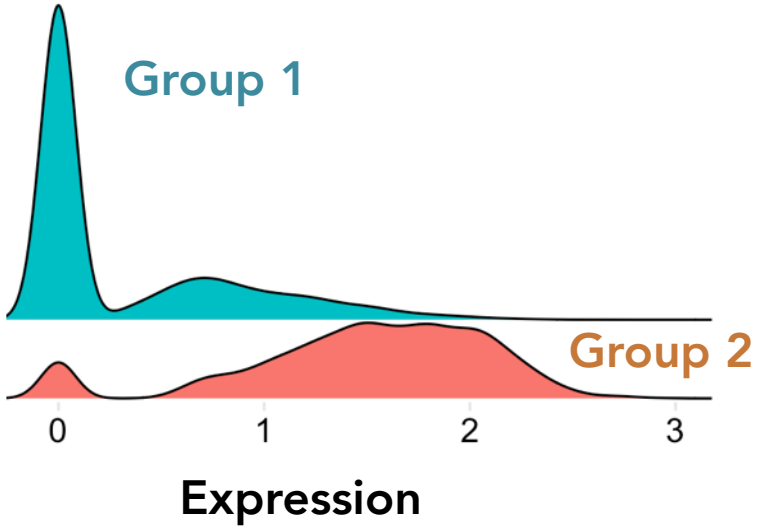
4. Differential expression analysis

Group A > Group B (p-value < 0.01)



BUT

"Zero inflation" poses a challenge in single-cell data!



Conventional statistical tests (e.g. "Student's t"), which assume a unimodal distribution can be underpowered in detecting true genes

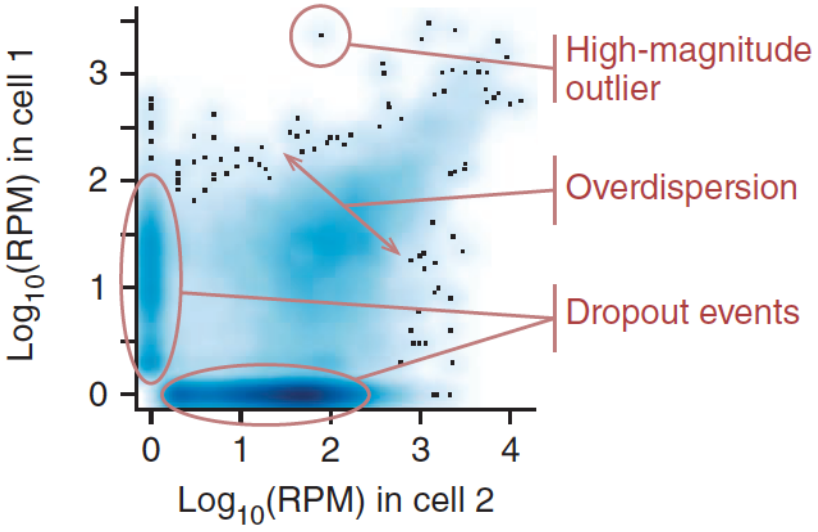
Differential expression analysis

Many of the DE methods developed for bulk RNA-seq (e.g. edgeR, DE-seq) have serious limitations when applied to scRNA-seq data because of dropouts, so apply with caution!

	Short name	Method	Software version	Input	Reference
	BPSC	BPSC	BPSC 0.99.0	CPM	[48]
	D3E	D3E	D3E 1.0	raw counts	[49]
	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	[14]
	DESeq2census	DESeq2	DESeq2 1.14.1	census counts	[14]
	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	[14]
	edgeRLRT	edgeR/LRT	edgeR 3.17.5	raw counts	[15, 41, 37]
	edgeRLRTcensus	edgeR/LRT	edgeR 3.17.5	census counts	[15, 41, 37]
	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.17.5, scran 1.2.0	raw counts	[15, 37, 42]
	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.17.5	raw counts	[15, 41, 37, 40]
	edgeRQLF	edgeR/QLF	edgeR 3.17.5	raw counts	[15, 38, 41]
	limmatrend	limma-trend	limma 3.30.13	raw counts	[57, 16]
	MASTcpm	MAST	MAST 1.0.5	log_2(CPM+1)	[50]
	MASTcpmDetRate	MAST - accounting for detection rate	MAST 1.0.5	log_2(CPM+1)	[50]
	MASTtpm	MAST	MAST 1.0.5	log_2(TPM+1)	[50]
	MASTtpmDetRate	MAST - accounting for detection rate	MAST 1.0.5	log_2(TPM+1)	[50]
	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts	[54]
	monocle	monocle	monocle 2.2.0	TPM	[44]
	monoclecensus	monocle	monocle 2.2.0	census counts	[44, 43]
	NODES	NODES	NODES 0.0.0.9010	raw counts	[47]
	ROTScpm	ROTS	ROTS 1.2.0	CPM	[55, 56]
	ROTStpm	ROTS	ROTS 1.2.0	TPM	[55, 56]
	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	[55, 56]
	SAMseq	SAMseq	samr 2.0	raw counts	[45]
	SCDE	SCDE	scde 1.99.4	raw counts	[51]
	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM	[52, 44]
	voomlimma	voom-limma	limma 3.30.13	raw counts	[57, 16]
	Wilcoxon	Wilcoxon test	stats (R v 3.3.1)	TMM-normalized TPM	[41, 46]

Soneson and Robinson, 2017

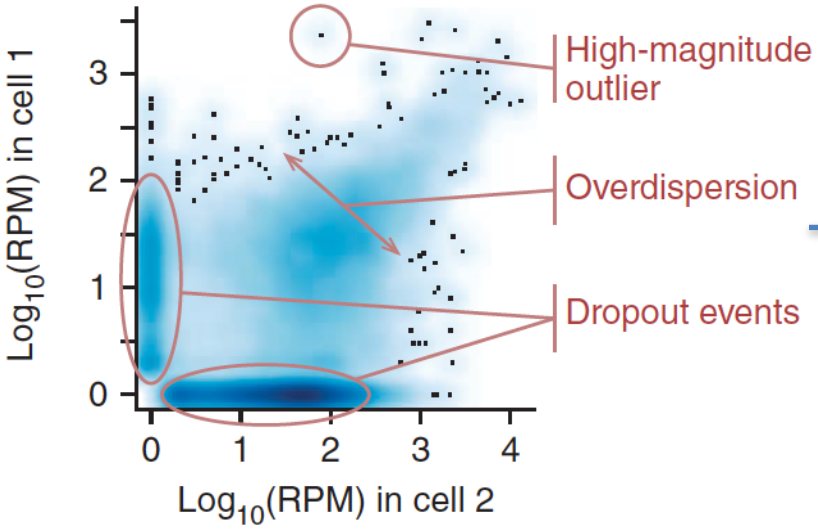
Single-cell differential Expression (SCDE)

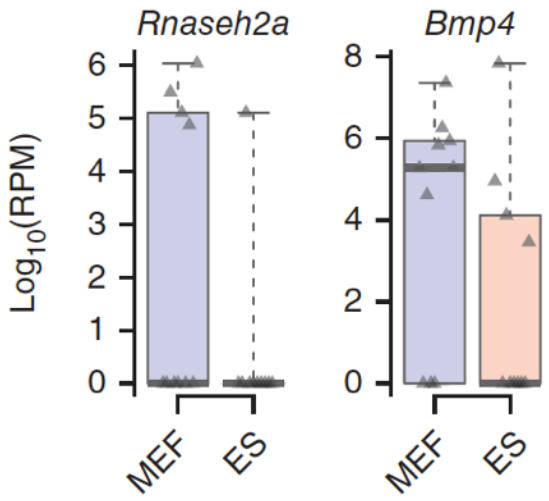


Kharchenko et al., 2014

Single-cell differential Expression (SCDE)

DE genes using conventional methods can include high magnitude outliers or dropout events

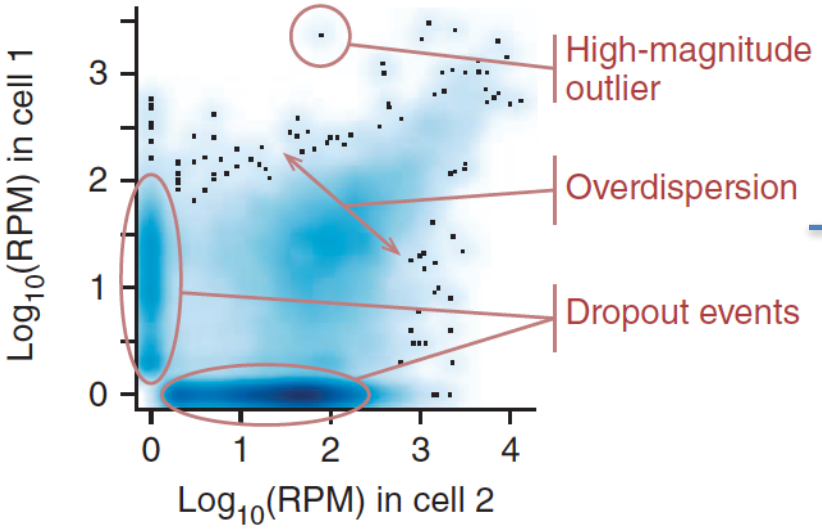


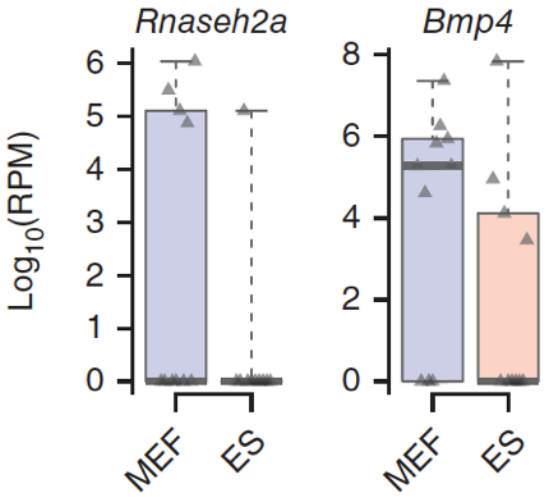


Kharchenko et al., 2014

Single-cell differential Expression (SCDE)

DE genes using conventional methods can include high magnitude outliers or dropout events





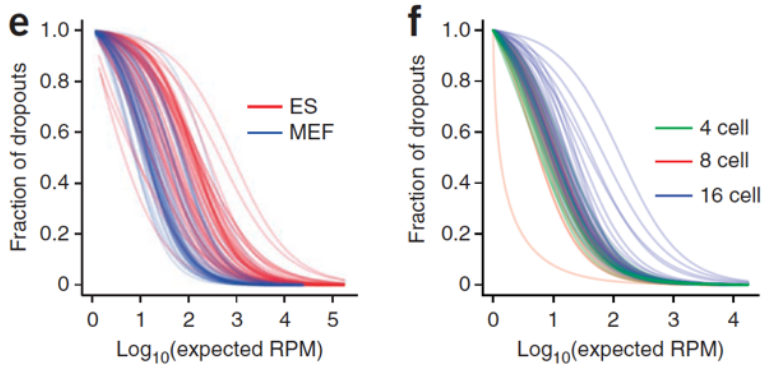
SCDE exchanges information between closely related cells to estimate dropout rates for every cell!

$$\begin{cases} r_1 \approx Poisson(\lambda_0) & \text{Dropout in } c_1 \\ \begin{cases} r_1 \approx NB(r_2) \\ r_2 \approx NB(r_1) \end{cases} & \text{Amplified} \\ r_2 \approx Poisson(\lambda_0) & \text{Dropout in } c_2 \end{cases}$$

Kharchenko et al., 2014

Single-cell differential Expression (SCDE)

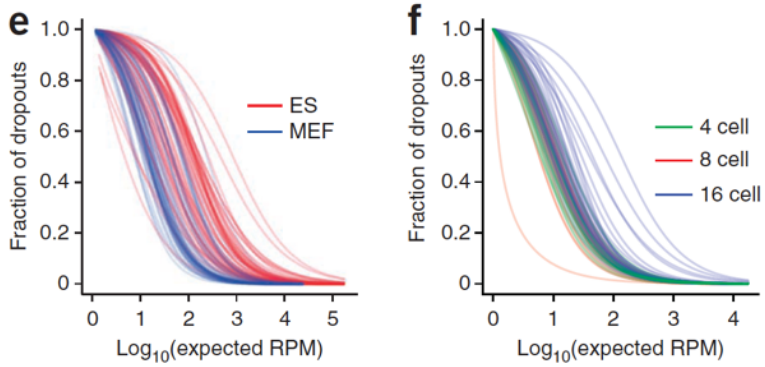
For every cell, a "dropout curve" is estimated



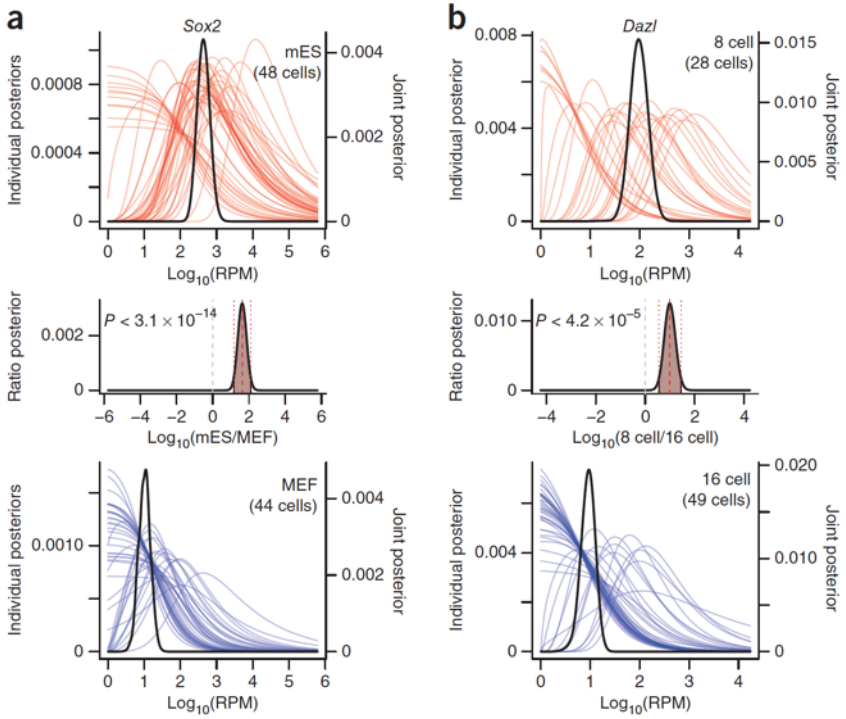
Kharchenko et al., 2014

Single-cell differential Expression (SCDE)

For every cell, a "dropout curve" is estimated

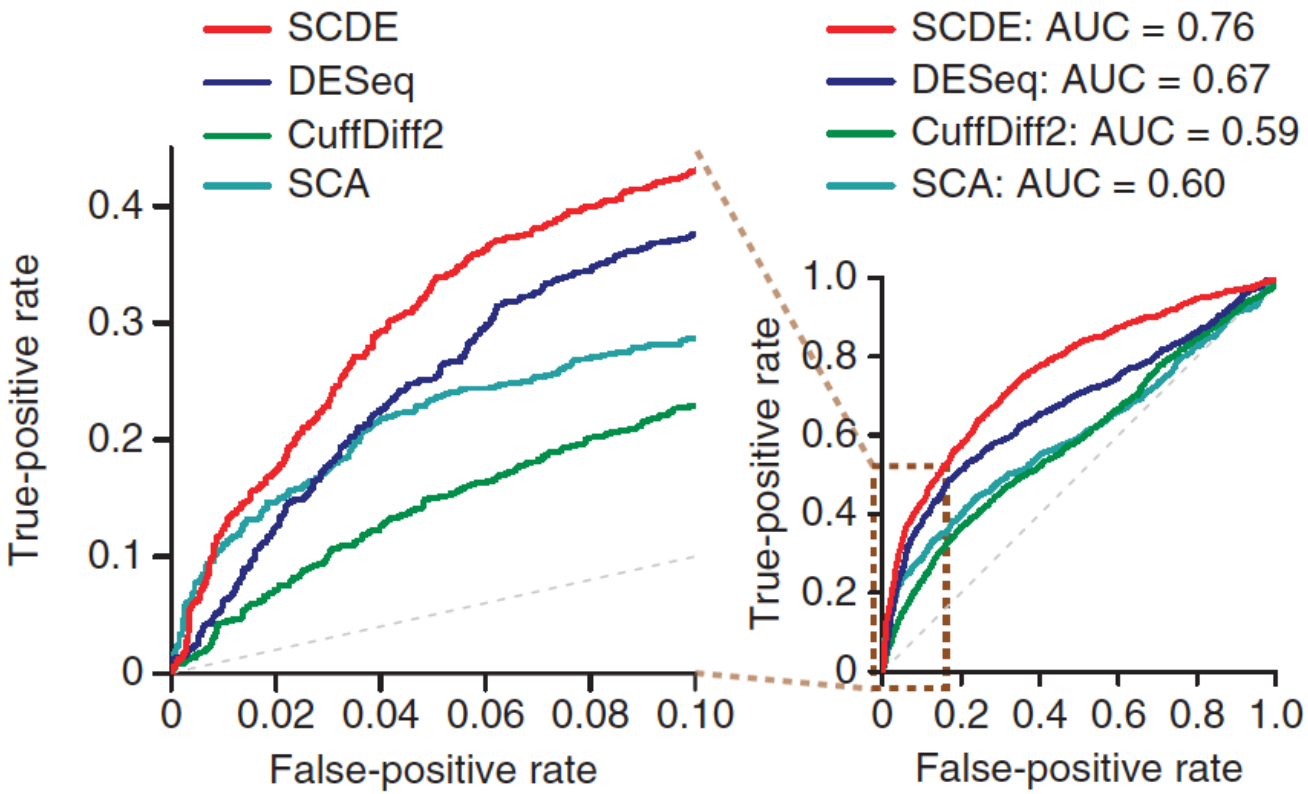


Which is used in a Bayesian framework to estimate posterior distributions for every gene in every cell



Kharchenko et al., 2014

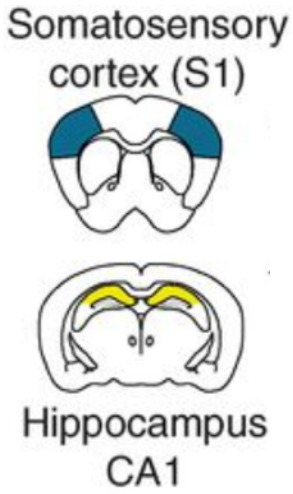
SCDE is much more sensitive and specific



One of the disadvantages of SCDE is its run-time, which does not scale well for large datasets. Newer methods like MAST (Finak et al., 2016) overcome this!

Pathway analysis

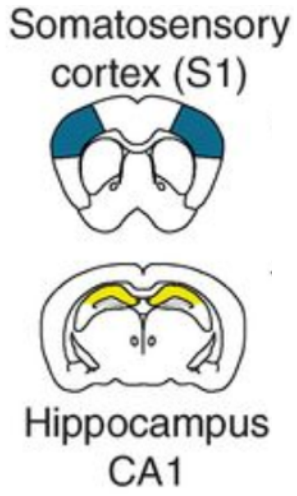
- More often than not, one is interested in seeing which biological processes drive heterogeneity
- Pathway and Gene OverDispersion Analysis (**PAGODA**), evaluates key Gene Ontology terms that drive variation in the data.

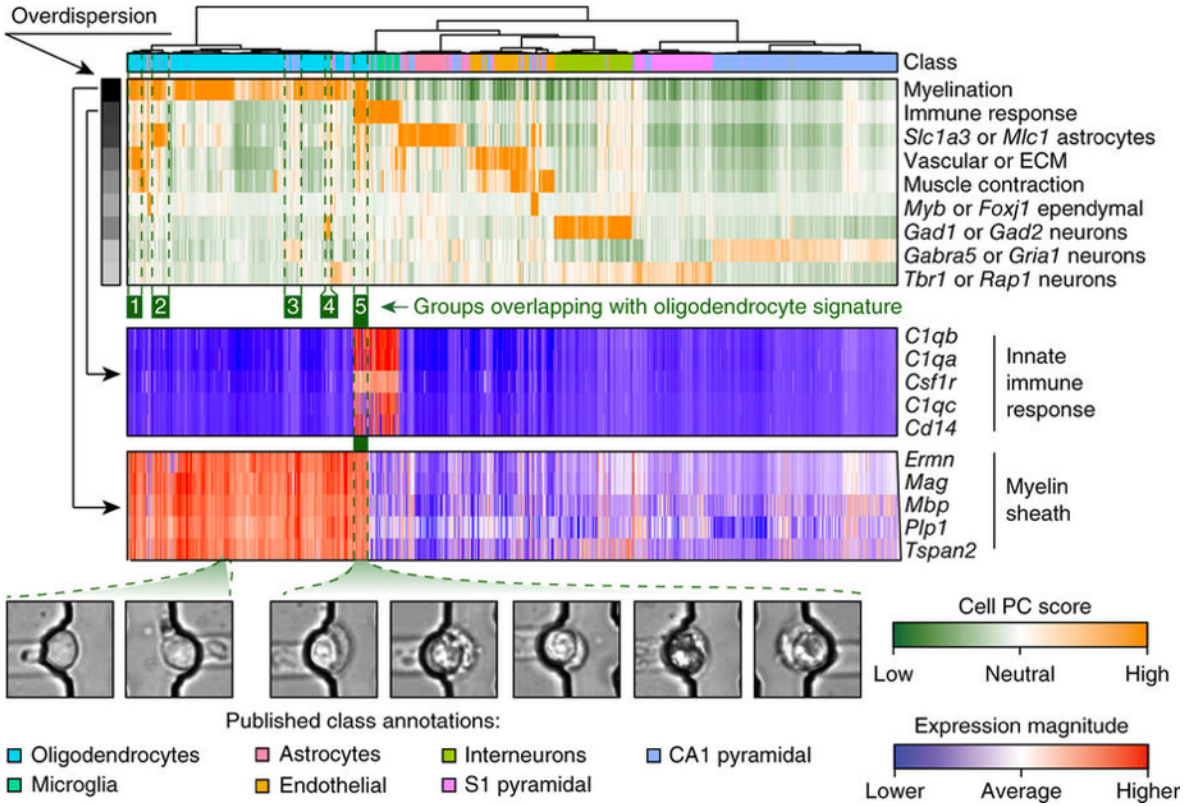


Fan et al., Nature Methods, 2016

Pathway analysis

- More often than not, one is interested in seeing which biological processes drive heterogeneity
- Pathway and Gene OverDispersion Analysis (**PAGODA**), evaluates key Gene Ontology terms that drive variation in the data.





Fan et al., Nature Methods, 2016

Single-cell tools (2016)

Table 2
Software packages for scRNA-seq analysis

Package name (language of implementation)	Functions	References
ZIFA (Python)	Dimensionality reduction algorithm that accounts for transcript dropouts in single-cell data	[66]
Monocle (R)	Mapping transcripts on differentiation cascades, and arranging single cells along a differentiation tree (pseudotime estimation)	[67]
scLVM (Python)	Dissecting cofounding sources of variation (e.g., differentiation vs. cell-cycle) in scRNA-seq data	[68]
SCDE (R)	Testing for differential expression in scRNA-seq data where technical artifacts abound (transcript dropouts, library quality variation)	[32]
BASiCS (R)	A Bayesian framework to normalize and assess biological/technical variation in scRNa-seq data	[69]
Seurat (R)	Spatial mapping of single-cell transcriptomes based on a preexisting spatial pattern of landmark genes. R package also includes wrapper functions for analysis and visualization	[75]
Pagoda (R)	Pathway and geneset overdispersion analysis	[70]
Sincell (R)	Assessment of cell-state hierarchies	[71]
RaceID (R)	Rare cell type identification in complex populations of single cells	[72]
Scuba (Matlab)	Extracting lineage relationships from single-cell data	[73]
Scater (R)	R-based package for quality control, visualization, and preprocessing	[74]

Villani and Shekhar, 2016

Awesome Single-Cell

https://github.com/seandavi/awesome-single-cell

README.md

awesome-single-cell

List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc. Contributions welcome...

Citation

DOI 10.5281/zenodo.1117763

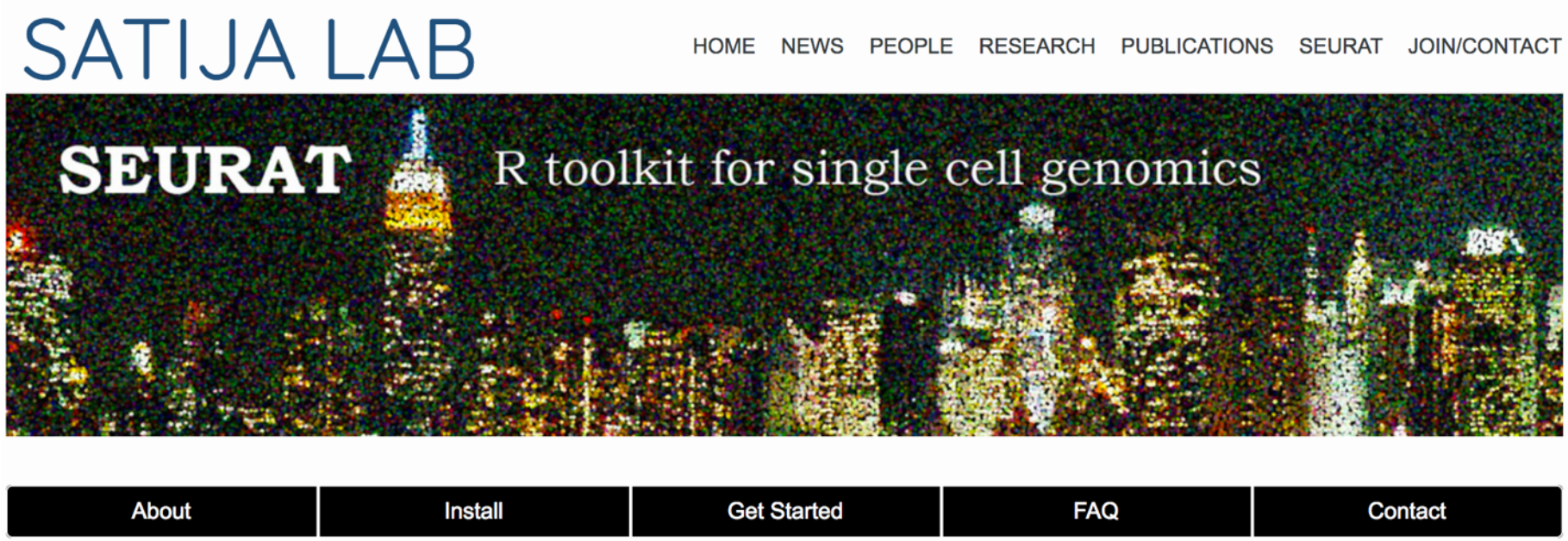
Software packages

RNA-seq

- anchor - [Python] - ⚓ Find bimodal, unimodal, and multimodal features in your data
- ascend - [R] - ascend is an R package comprised of fast, streamlined analysis functions optimized to address the statistical challenges of single cell RNA-seq. The package incorporates novel and established methods to provide a flexible framework to perform filtering, quality control, normalization, dimension reduction, clustering, differential expression and a wide-range of plotting.
- BackSPIN - [Python] - Biclustering algorithm developed taking into account intrinsic features of single-cell RNA-seq experiments.

More than ~80 software packages are available!

http://satijalab.org/seurat/



RStudio®

Coffee Break

Agenda

- Single cell analysis - why?
- A short survey of scRNA-seq methods
- Quality comparison of different methods and power analysis
- Overview of computational workflow
  - Preprocessing
  - Secondary analysis in R
- **Some example applications**
- Future

Applications: What can we learn?

1. Census and taxonomy - Molecular definitions of cell types

2. Anatomy and Physiology - Spatial structure of tissues

3. Cancer - Cancer stem cells? Role of the microenvironment?

4. Development - How do stem cells commit to fates?

Applications: What can we learn?

1. Census and taxonomy - Molecular definitions of cell types

2. Anatomy and Physiology - Spatial structure of tissues

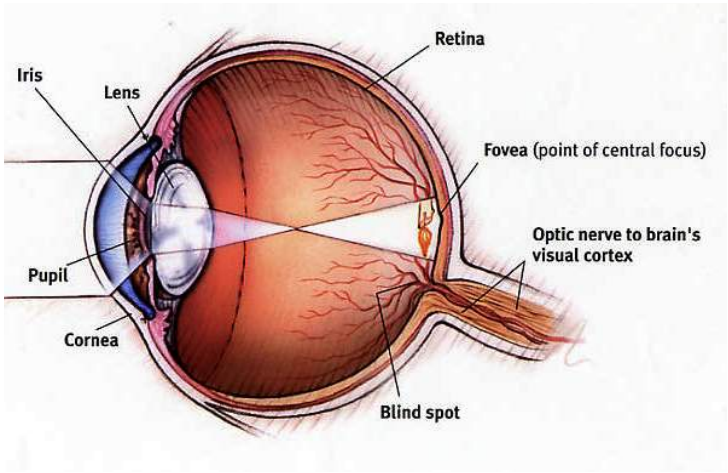
3. Cancer - Cancer stem cells? Role of the microenvironment?

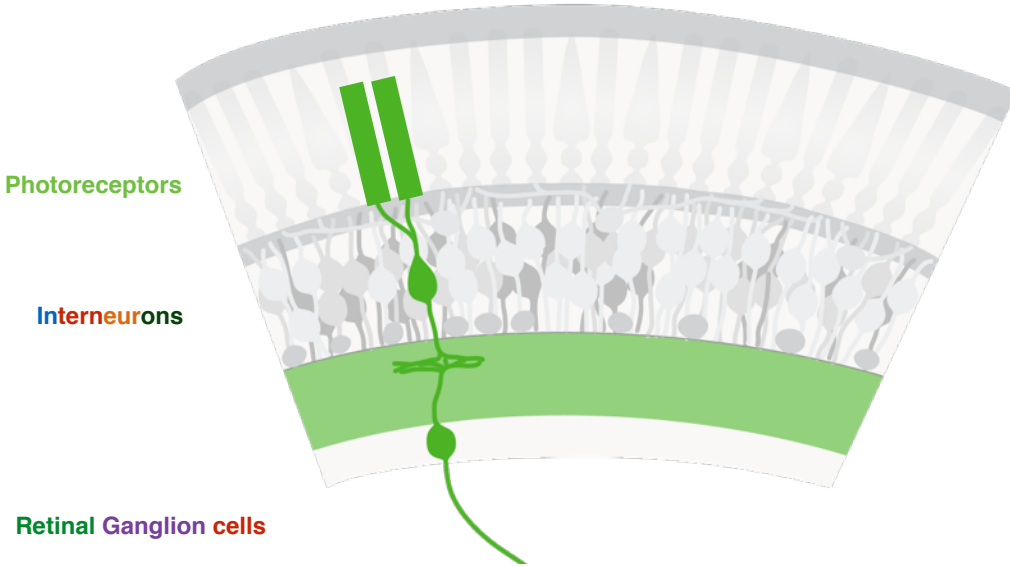
4. Development - How do stem cells commit to fates?

Census and taxonomy

- A complete list of molecularly specified cell types in a tissue
- What's a cell type? - how do we harmonize different aspects of cell identity

The retina: Light sensitive tissue in the eye that processes vision

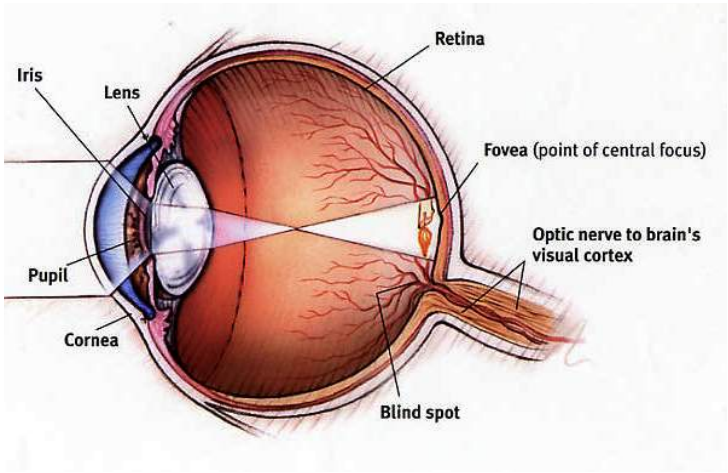


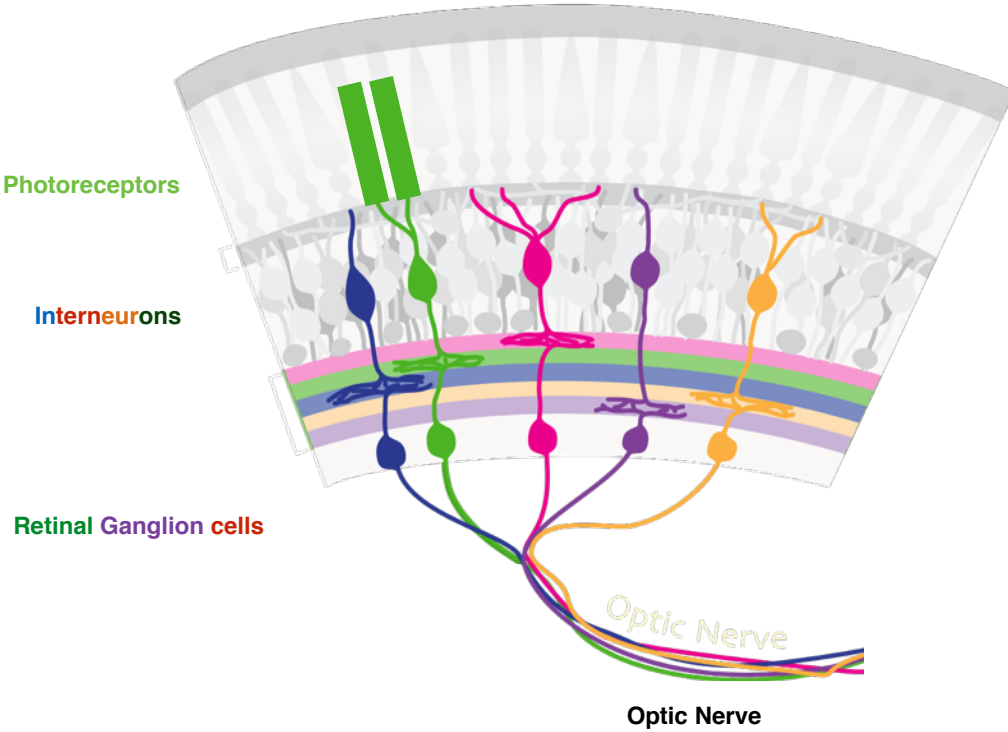


Census and taxonomy

- A complete list of molecularly specified cell types in a tissue
- What's a cell type? - how do we harmonize different aspects of cell identity

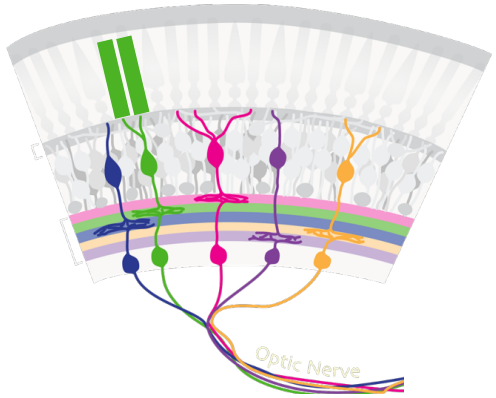
The retina: Light sensitive tissue in the eye that processes vision





Towards a complete census of the mouse retina

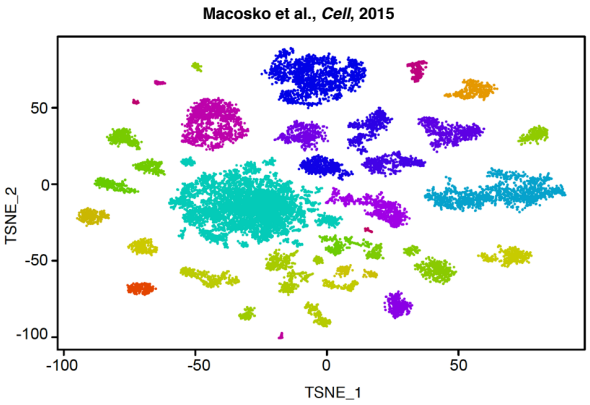
Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches



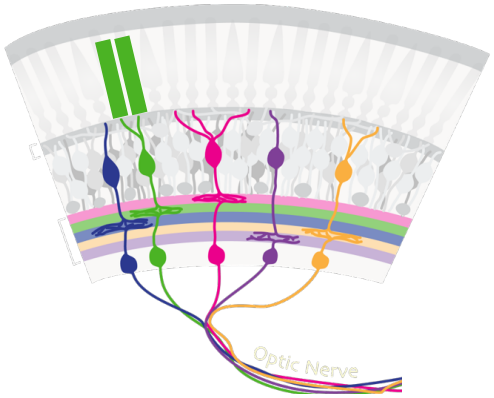
Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

Full Retina Drop-seq (45000 cells, 39 types)

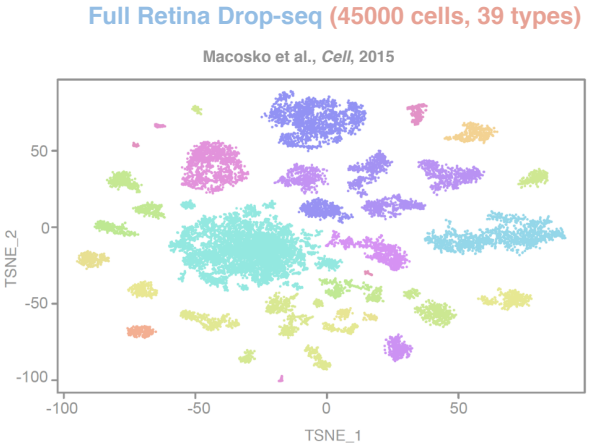


Proof of Principle

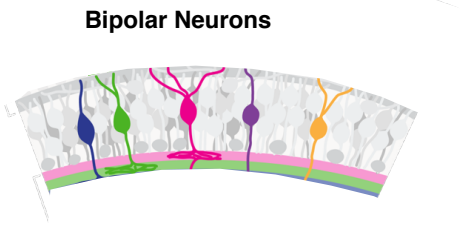


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches



Proof of Principle

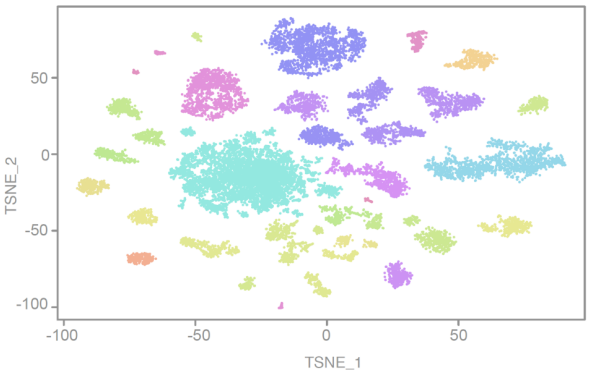


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

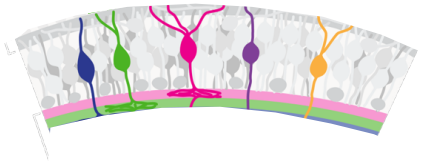
Full Retina Drop-seq (45000 cells, 39 types)

Macosko et al., *Cell*, 2015



Proof of Principle

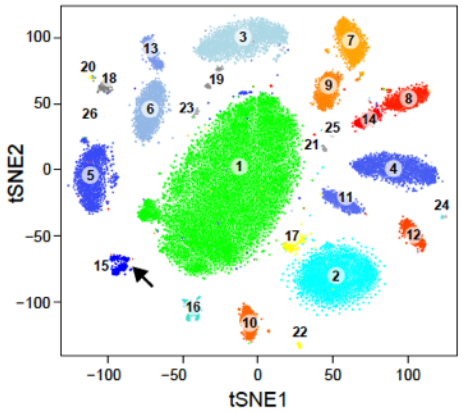
Bipolar Neurons



Comprehensive classification of Bipolar Neurons (28,000 cells, 15 types)

Faster, more robust graph based clustering

Shekhar et al. *Cell*, 2016

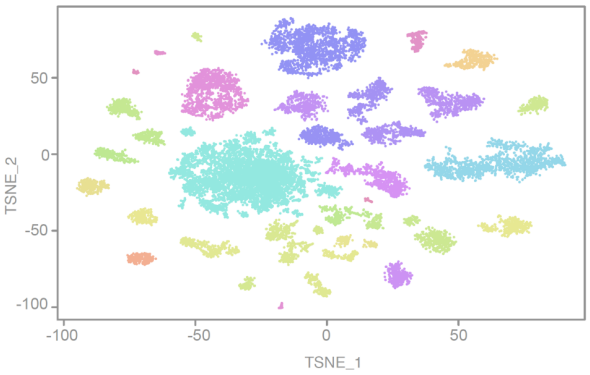


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

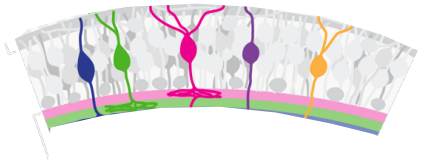
Full Retina Drop-seq (45000 cells, 39 types)

Macosko et al., *Cell*, 2015



Proof of Principle

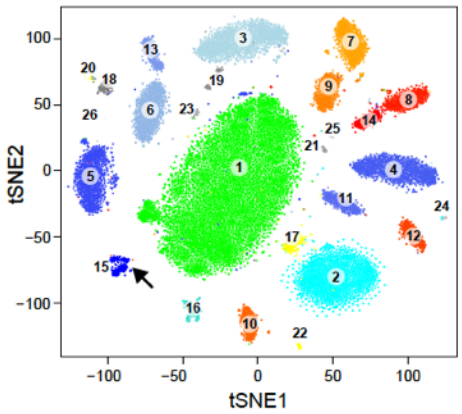
Bipolar Neurons



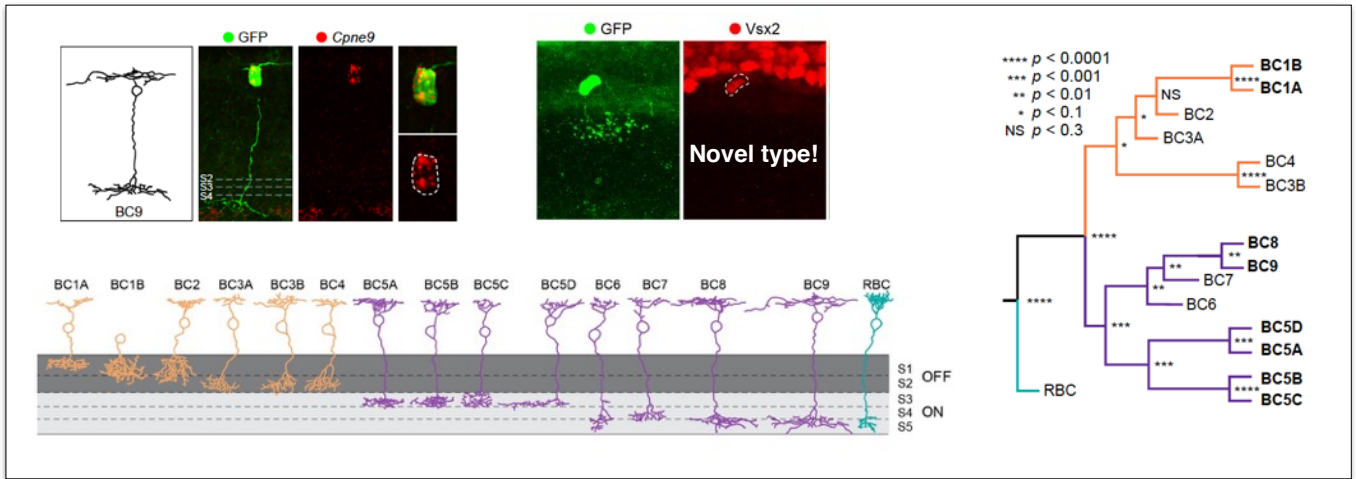
Comprehensive classification of Bipolar Neurons (28,000 cells, 15 types)

Faster, more robust graph based clustering

Shekhar et al. *Cell*, 2016



Matched molecular expression to morphology for 15 bipolar types

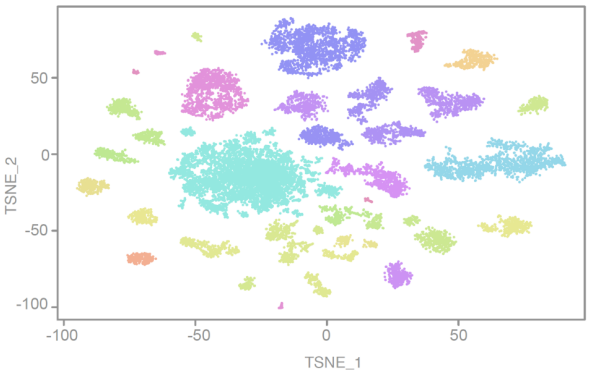


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

Full Retina Drop-seq (45000 cells, 39 types)

Macosko et al., *Cell*, 2015



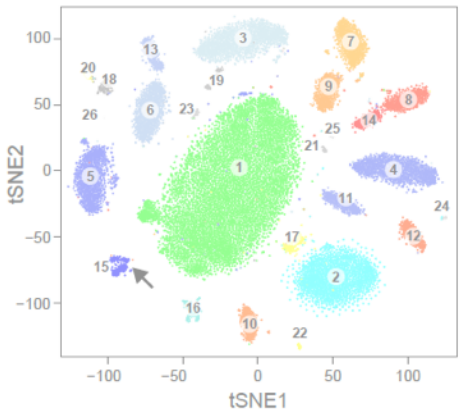
Proof of Principle

Retinal Ganglion Cells

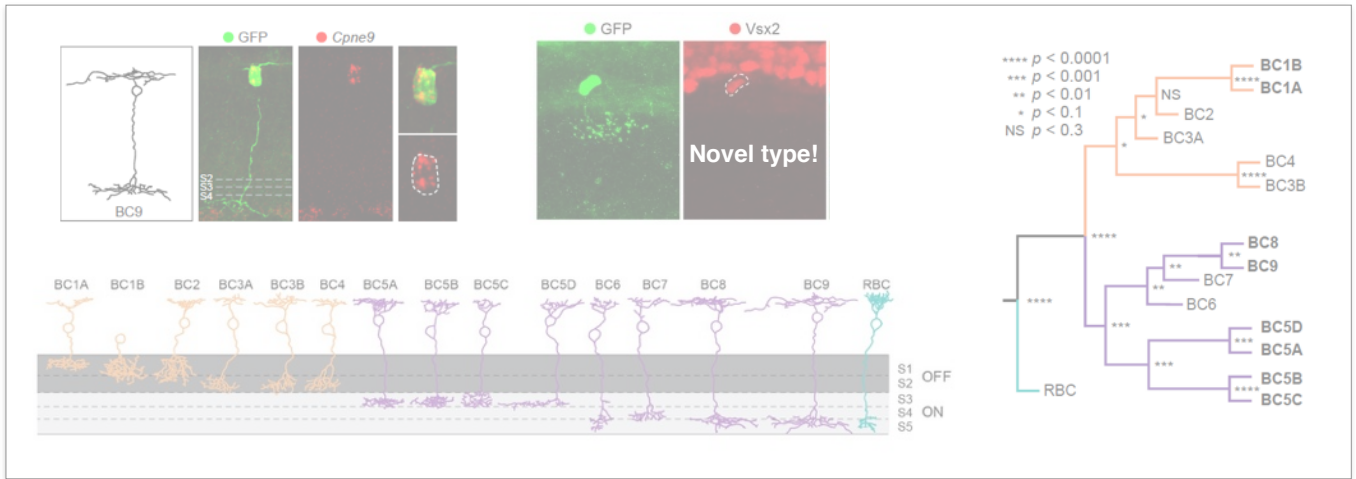
Comprehensive classification of Bipolar Neurons (28,000 cells, 15 types)

Faster, more robust graph based clustering

Shekhar et al. *Cell*, 2016



Matched molecular expression to morphology for 15 bipolar types

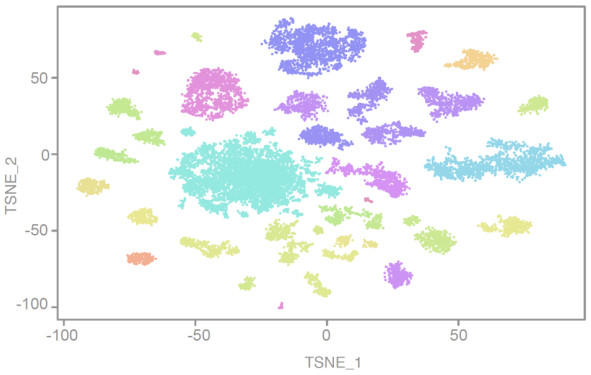


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

Full Retina Drop-seq (45000 cells, 39 types)

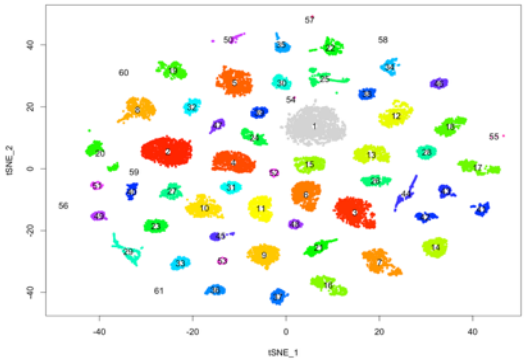
Macosko et al., *Cell*, 2015



Proof of Principle

Retinal Ganglion Cells

Unpublished



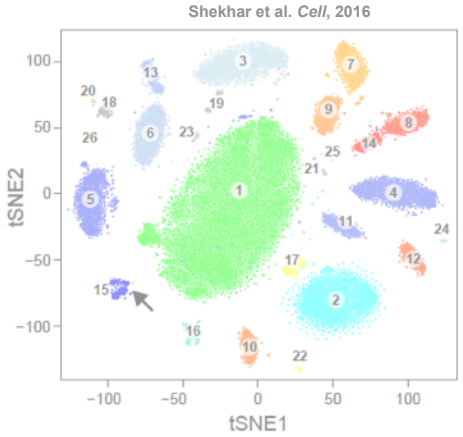
Classification of Ganglion cells (42,000 cells, ~45 types)

Kernel based clustering for higher sensitivity

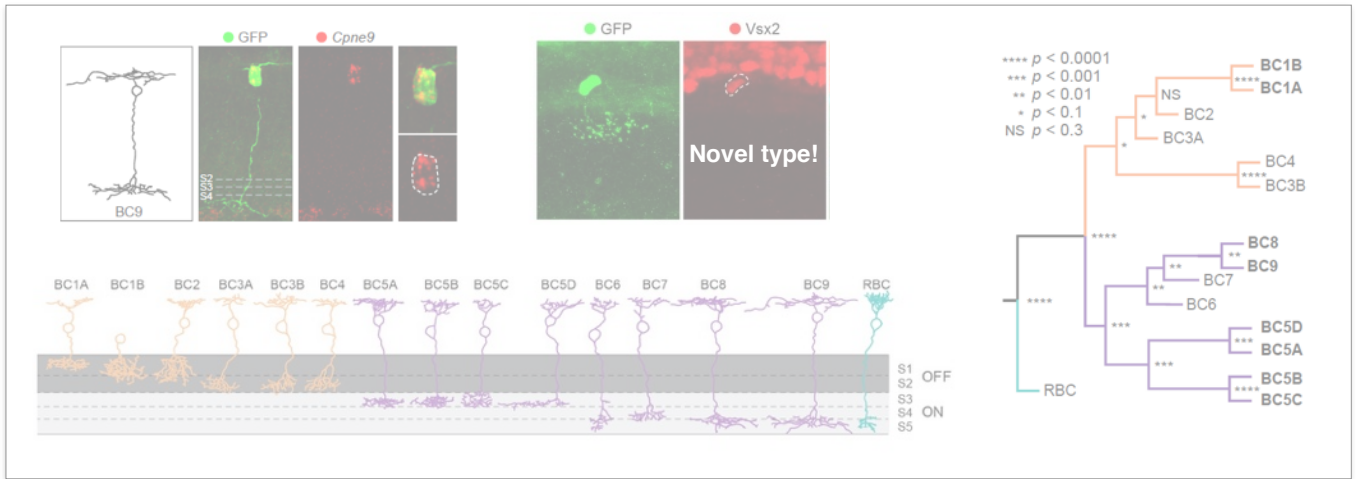
Comprehensive classification of Bipolar Neurons (28,000 cells, 15 types)

Faster, more robust graph based clustering

Shekhar et al. *Cell*, 2016



Matched molecular expression to morphology for 15 bipolar types

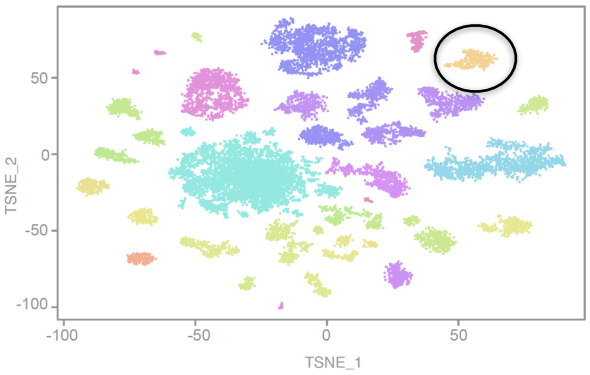


Towards a complete census of the mouse retina

Challenge : Highly diverse (~150 types), with widely varying frequencies
Advantage : Vast array of molecular and genetic tools to validate new approaches

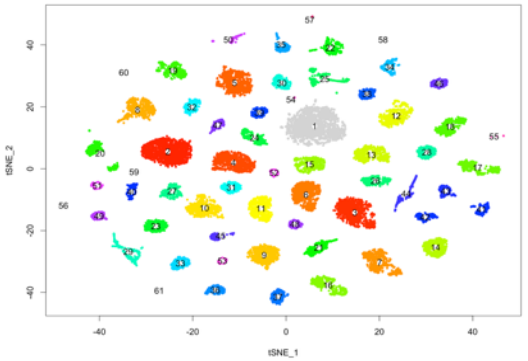
Full Retina Drop-seq (45000 cells, 39 types)

Macosko et al., *Cell*, 2015



Proof of Principle

Unpublished



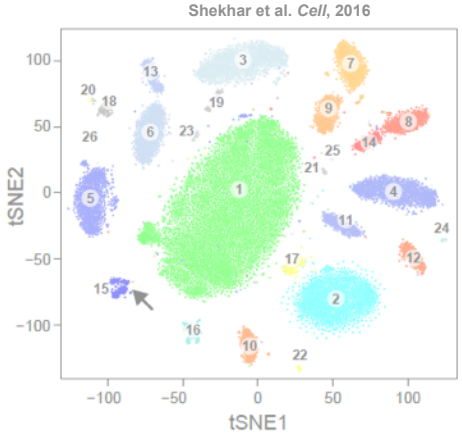
Classification of Ganglion cells (42,000 cells, ~45 types)

Kernel based clustering for higher sensitivity

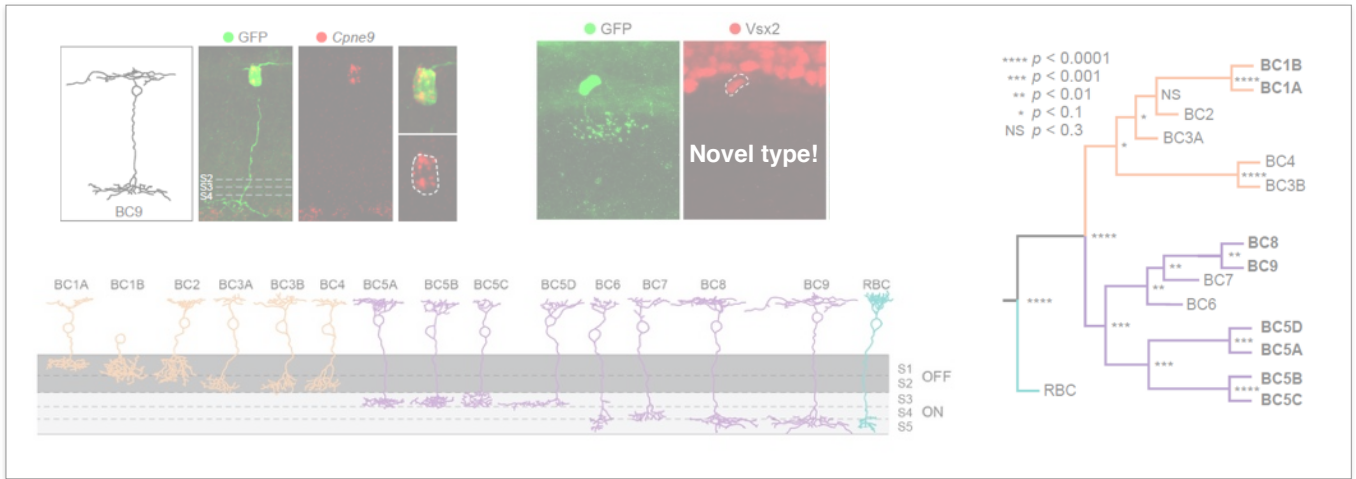
Comprehensive classification of Bipolar Neurons (28,000 cells, 15 types)

Faster, more robust graph based clustering

Shekhar et al. *Cell*, 2016



Matched molecular expression to morphology for 15 bipolar types



Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

n cells

m genes

Cell:	1	2	···	*N*
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

n cells

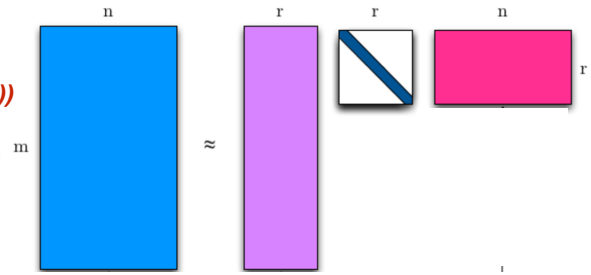
	Cell: 1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

m genes

*O(mn*log(r))*

low rank approximation using randomized SVD

Halko, Martinsson and Tropp, 2009



Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

n cells

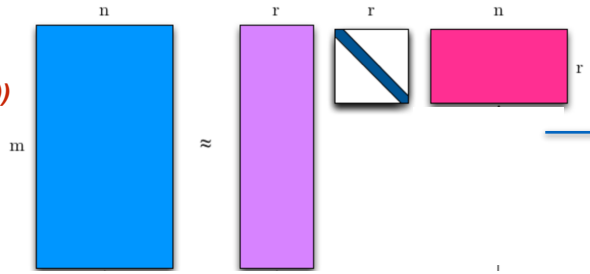
	Cell: 1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

m genes

*O(mn*log(r))*

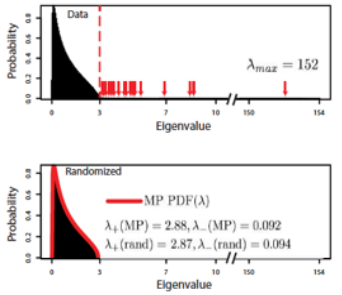
low rank approximation using randomized SVD

Halko, Martinsson and Tropp, 2009



Estimate matrix rank (permutation, MP law)

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}\mathbf{1}_{[\lambda_-,\lambda_+]}\,dx$$



Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

n cells

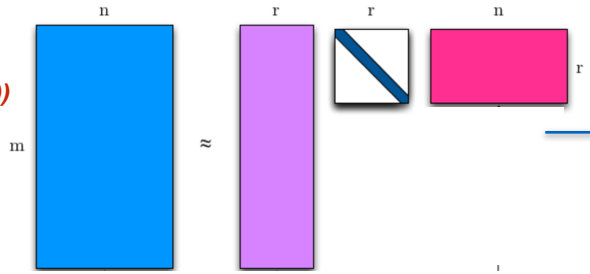
m genes

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

*O(mn*log(r))*

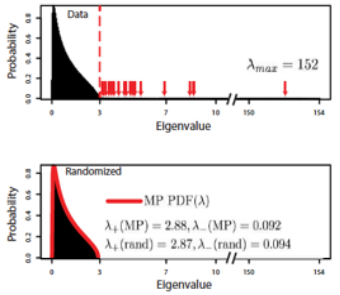
low rank approximation using randomized SVD

Halko, Martinsson and Tropp, 2009

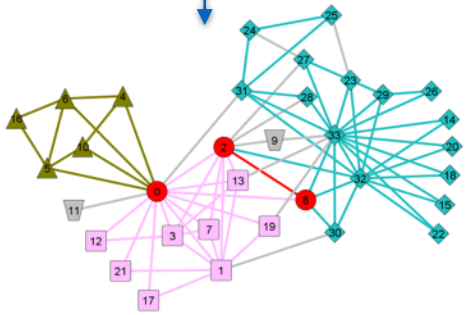


Estimate matrix rank (permutation, MP law)

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}\mathbf{1}_{[\lambda_-,\lambda_+]}\,dx$$



*O(n) - O(n*log(n))*



***k*-NN graph on the low rank representation using approximate search / LSH**

Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

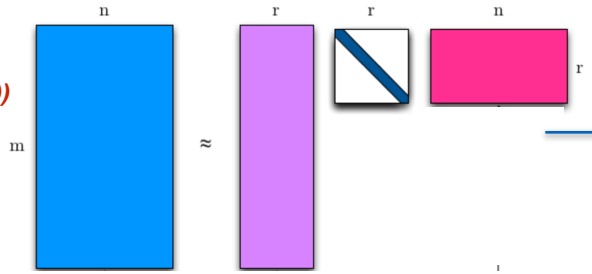
n cells

	Cell:	1	2	···	N
m genes	GENE 1	1	2		14
	GENE 2	4	27		8
	GENE 3	0	0		1
	⋮	⋮	⋮		⋮
	GENE M	6	2		0

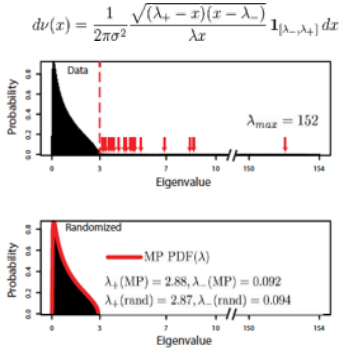
*O(mn*log(r))*

low rank approximation using randomized SVD

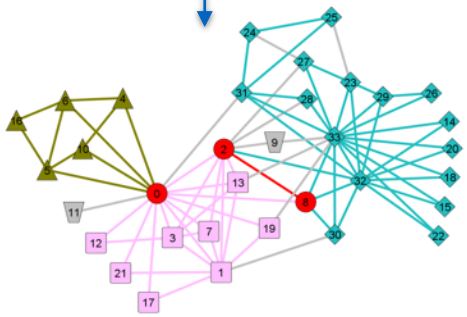
Halko, Martinsson and Tropp, 2009



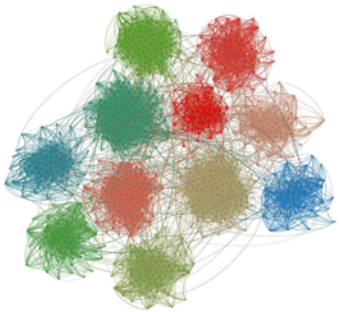
Estimate matrix rank (permutation, MP law)



*O(n) - O(n*log(n))*



*O(n) - O(n*log(n))*



Community detection using graph clustering

Blondel et al., J. Stat. Phys., 2008
Rosvall and Bergstrom, PNAS, 2008

***k*-NN graph on the low rank representation using approximate search / LSH**

Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

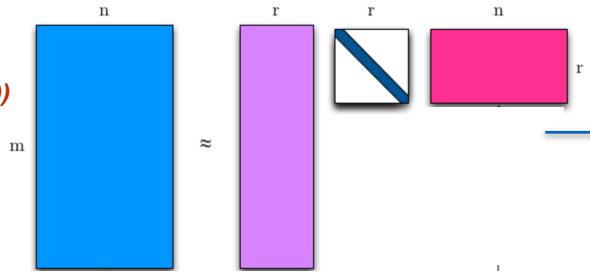
n cells

	Cell:	1	2	···	N
m genes	GENE 1	1	2		14
	GENE 2	4	27		8
	GENE 3	0	0		1
	⋮	⋮	⋮		⋮
	GENE M	6	2		0

*O(mn*log(r))*

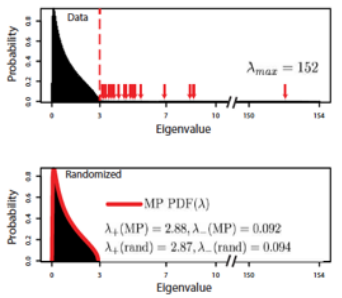
low rank approximation using randomized SVD

Halko, Martinsson and Tropp, 2009



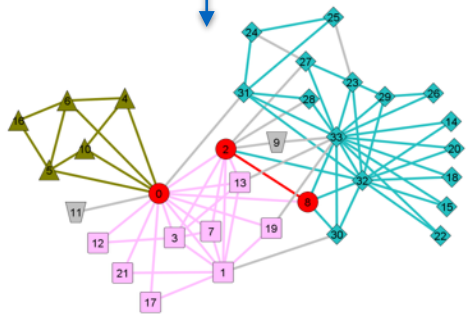
Estimate matrix rank (permutation, MP law)

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}\mathbf{1}_{[\lambda_-,\lambda_+]}\,dx$$



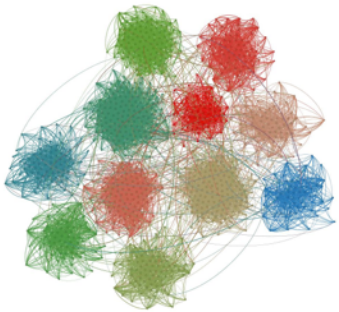
*O(n) - O(n*log(n))*

Robustness checks by bootstrapping, consensus clustering



***k*-NN graph on the low rank representation using approximate search / LSH**

*O(n) - O(n*log(n))*



Community detection using graph clustering

Blondel et al., J. Stat. Phys., 2008
Rosvall and Bergstrom, PNAS, 2008

Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Normalized, transformed sparse expression matrix

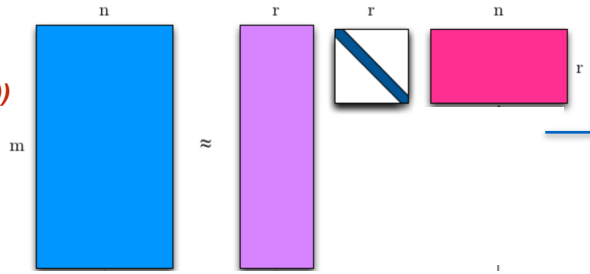
n cells

	Cell: 1	2	···	N
m genes				
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

*O(mn*log(r))*

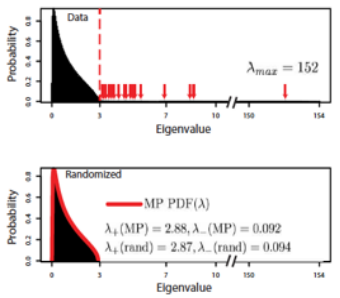
low rank approximation using randomized SVD

Halko, Martinsson and Tropp, 2009



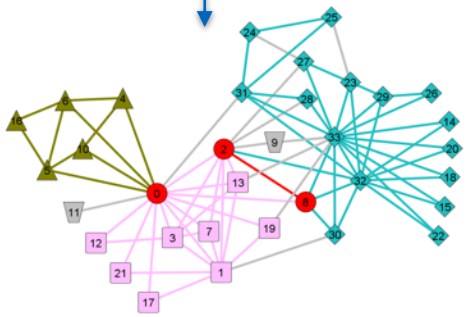
Estimate matrix rank (permutation, MP law)

$$d\nu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}\mathbf{1}_{[\lambda_-,\lambda_+]}\,dx$$



*O(n) - O(n*log(n))*

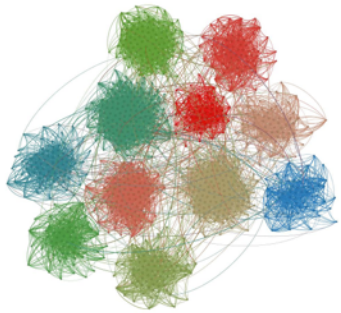
k-NN graph on the low rank representation using approximate search / LSH



*O(n) - O(n*log(n))*

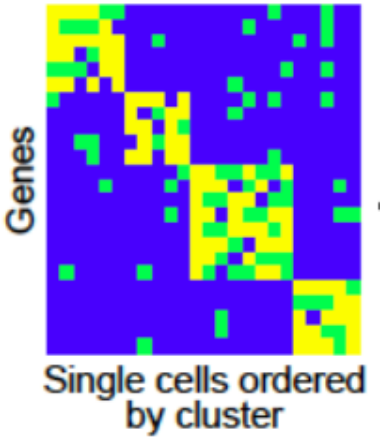
Community detection using graph clustering

Blondel et al., J. Stat. Phys., 2008
Rosvall and Bergstrom, PNAS, 2008



Robustness checks by bootstrapping, consensus clustering

Differential expression / GO / GSEA / visualization



Computational Approach

No dearth of clustering and dimensionality reduction methods, but need approaches that are scalable and reproducible

Cost of PCA/clustering for 50,000 cell dataset on 2.6 GHz Mac

- **Earlier**: 32 GB RAM, 6 hrs computing time
- **Now**: 4 GB RAM, <10 mins computing time

What can we learn?

1. Census and taxonomy

2. Anatomy and Physiology

3. Cancer

4. Development

Anatomy and Physiology: Goals and Challenges

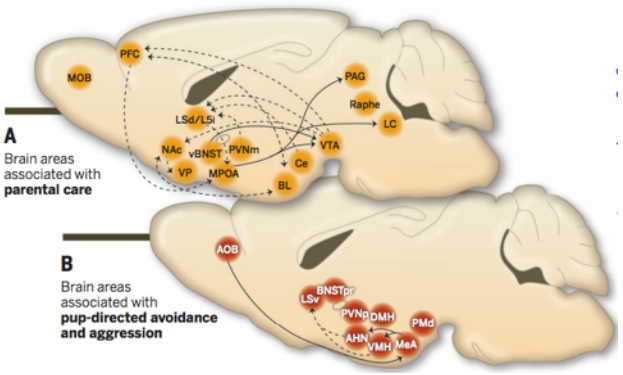
- The spatial location of a cell, and who it interacts with is critical for its function
- Also important for mapping causality

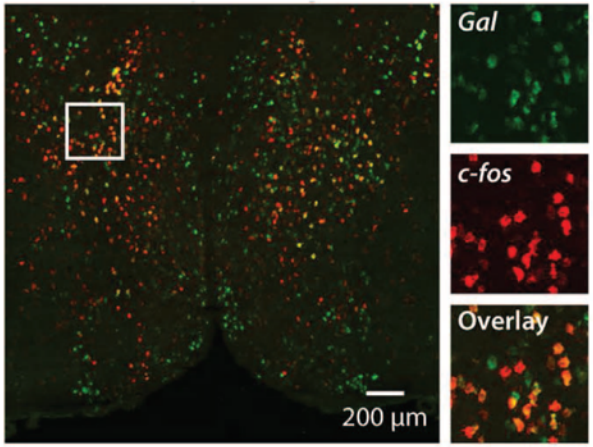
Dulac et al., 2014

Anatomy and Physiology: Goals and Challenges

- The spatial location of a cell, and who it interacts with is critical for its function
- Also important for mapping causality

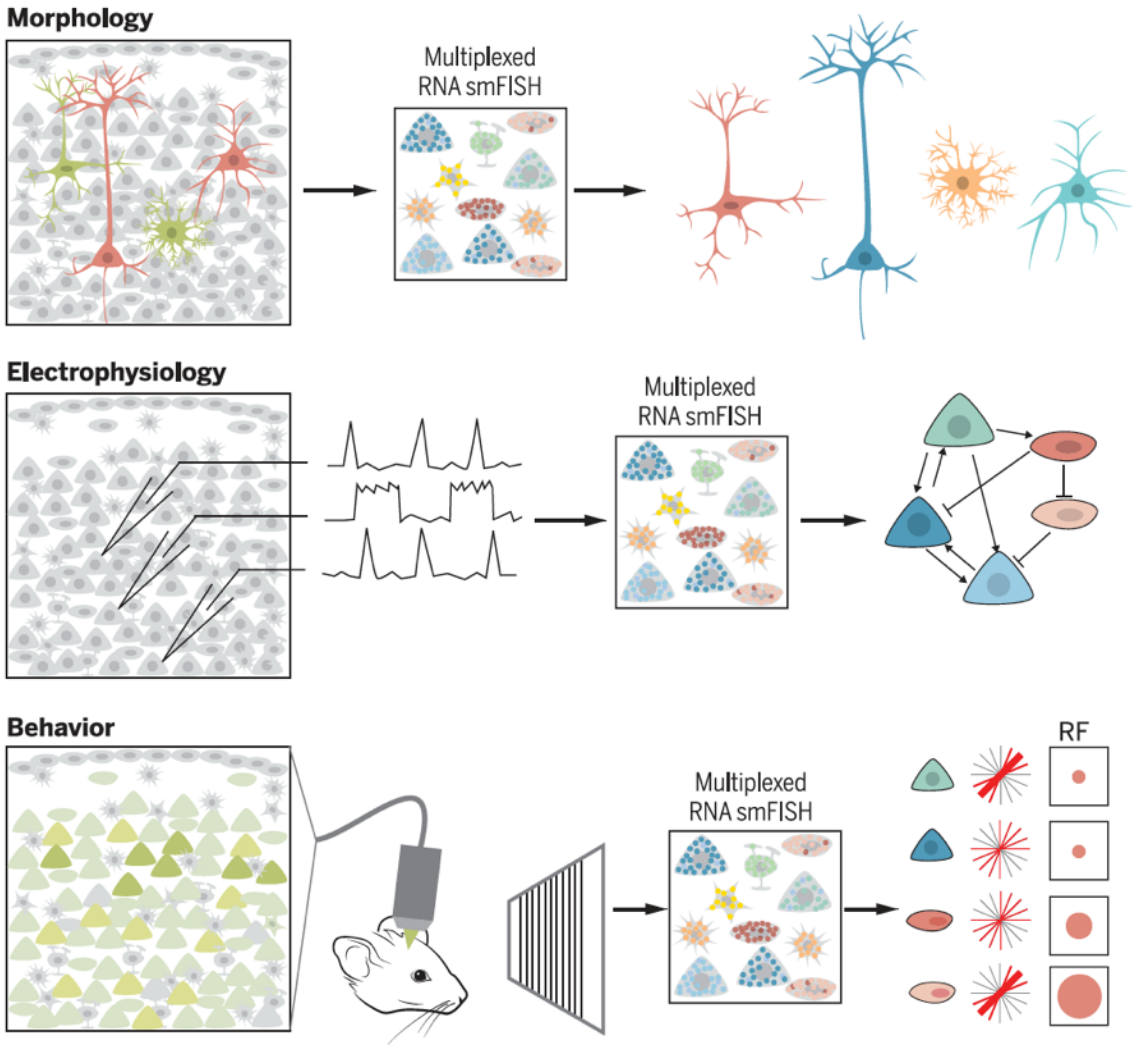
Gal+ neurons in the MPOA underlie parental behavior





Dulac et al., 2014

Spatially resolved transcriptomics



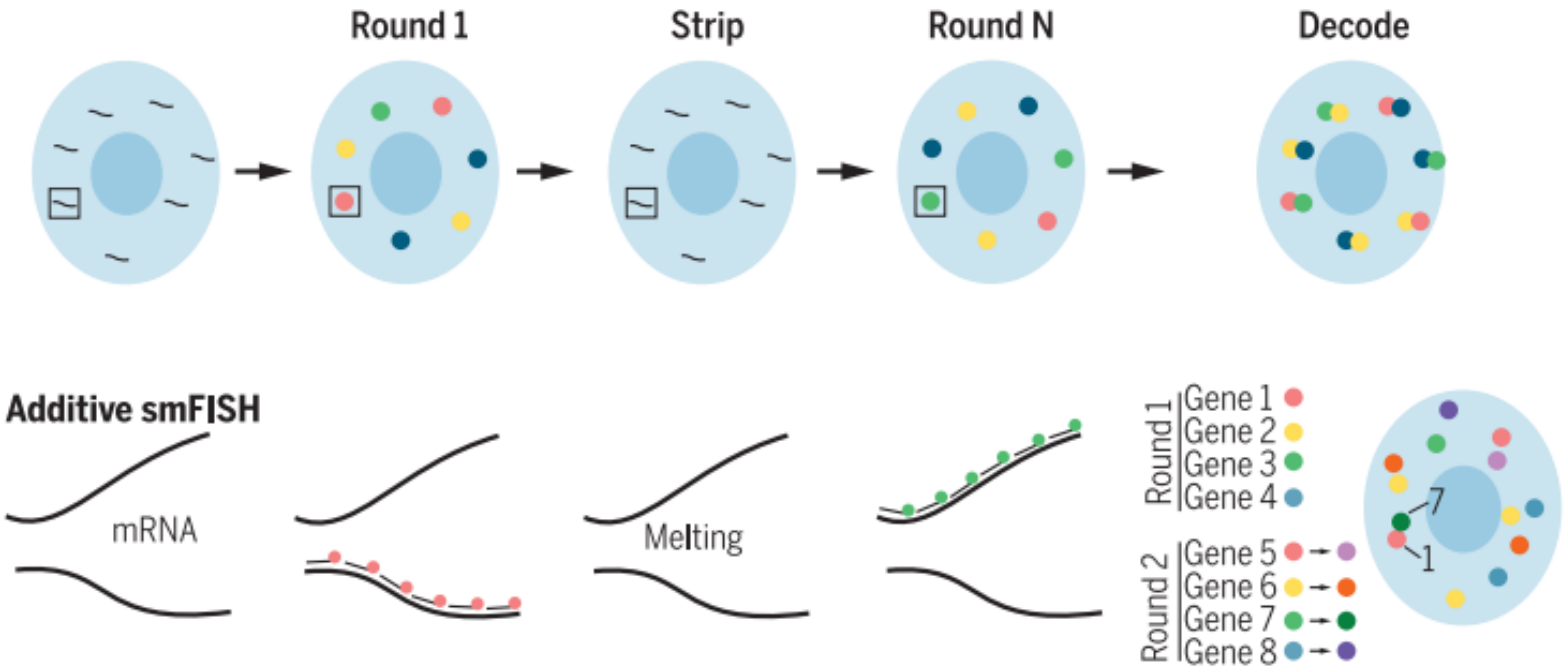
Location + Expression
+ Morphology

Location + Expression
+ Electrophysiology

Location + Expression
+ Electrophysiology

Lein et al., Science, 2017

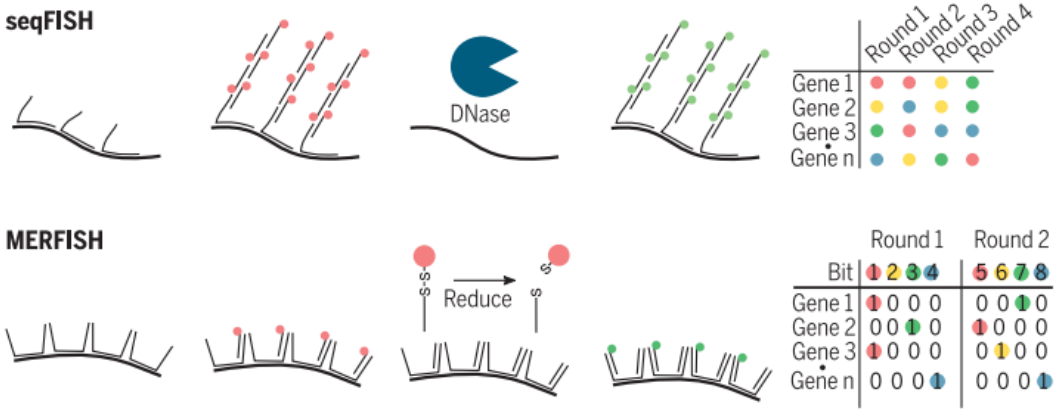
How does spatial transcriptomics work?



At each stage, a mutually exclusive subset of genes are measured. In this scheme, errors cannot be corrected.

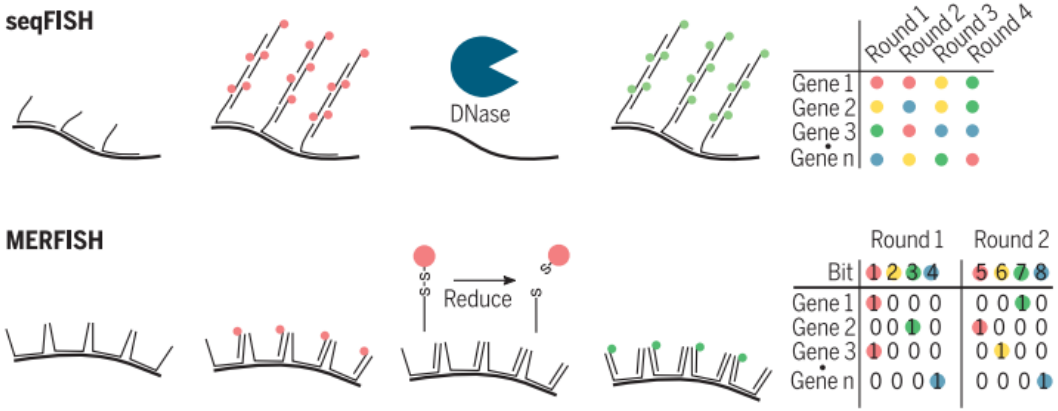
Two approaches to spatially resolve transcripts

Robust, error correcting multiplexed FISH



Two approaches to spatially resolve transcripts

Robust, error correcting multiplexed FISH

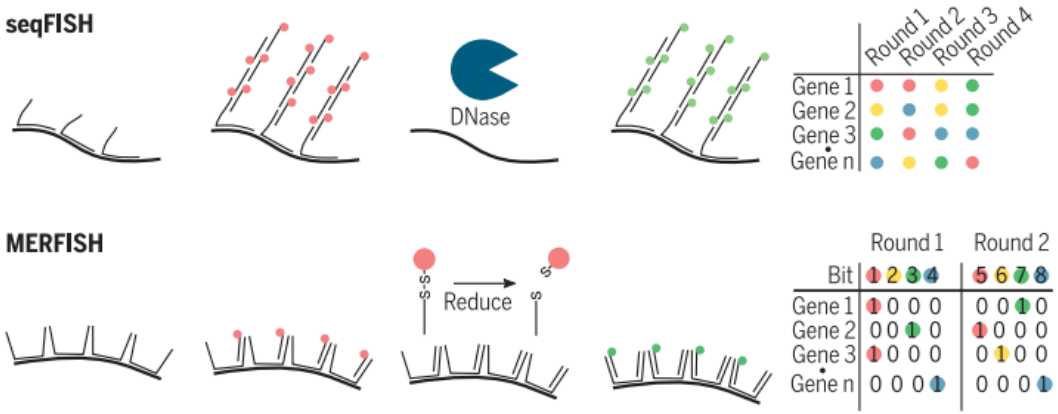


✓ ~100% sensitivity

- Probe dependent
- Has not been demonstrated in tissue sections

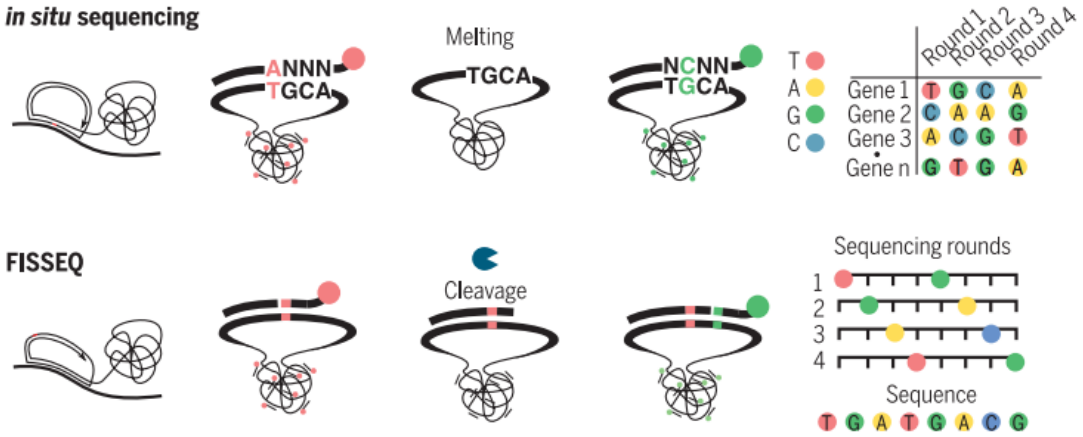
Two approaches to spatially resolve transcripts

Robust, error correcting multiplexed FISH



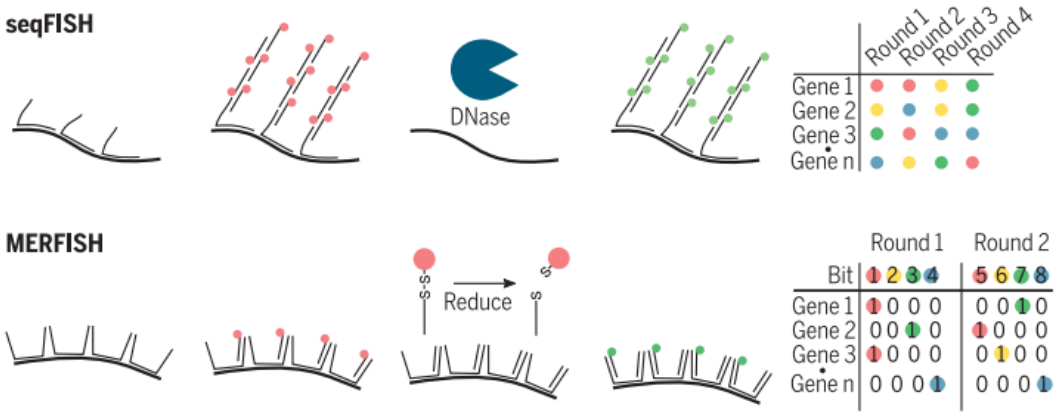
- ✓ ~100% sensitivity
- Probe dependent
- Has not been demonstrated in tissue sections

Sequencing RNA (*in situ* sequencing, FISSEQ)



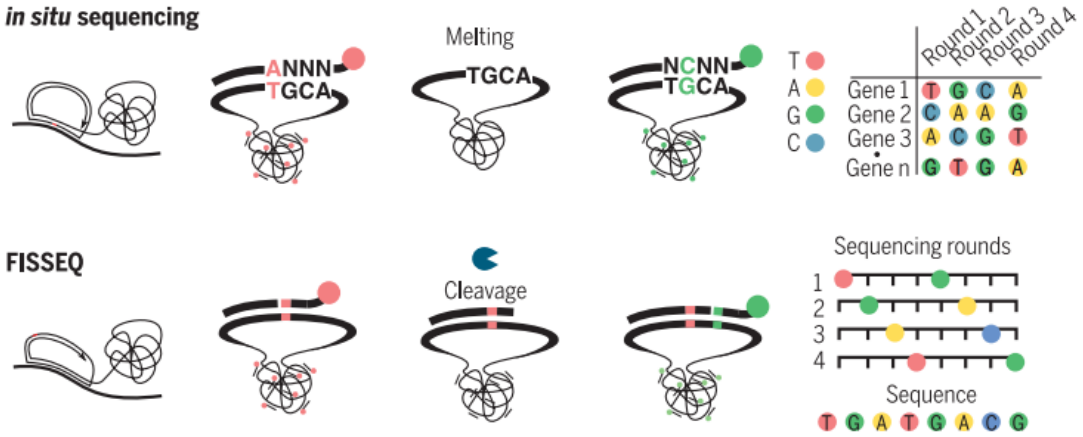
Two approaches to spatially resolve transcripts

Robust, error correcting multiplexed FISH



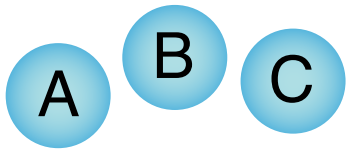
- ✓ ~100% sensitivity
- Probe dependent
- Has not been demonstrated in tissue sections

Sequencing RNA (*in situ* sequencing, FISSEQ)

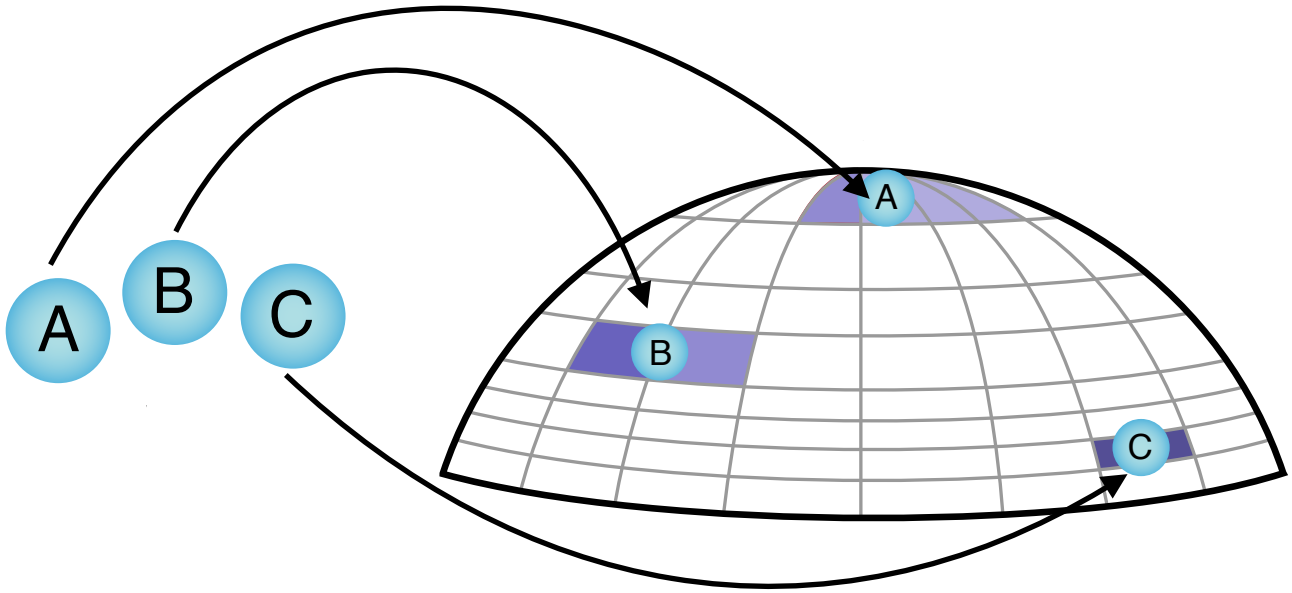


- ✓ Unbiased
- ✓ Works in tissue sections
- < 1% sensitivity

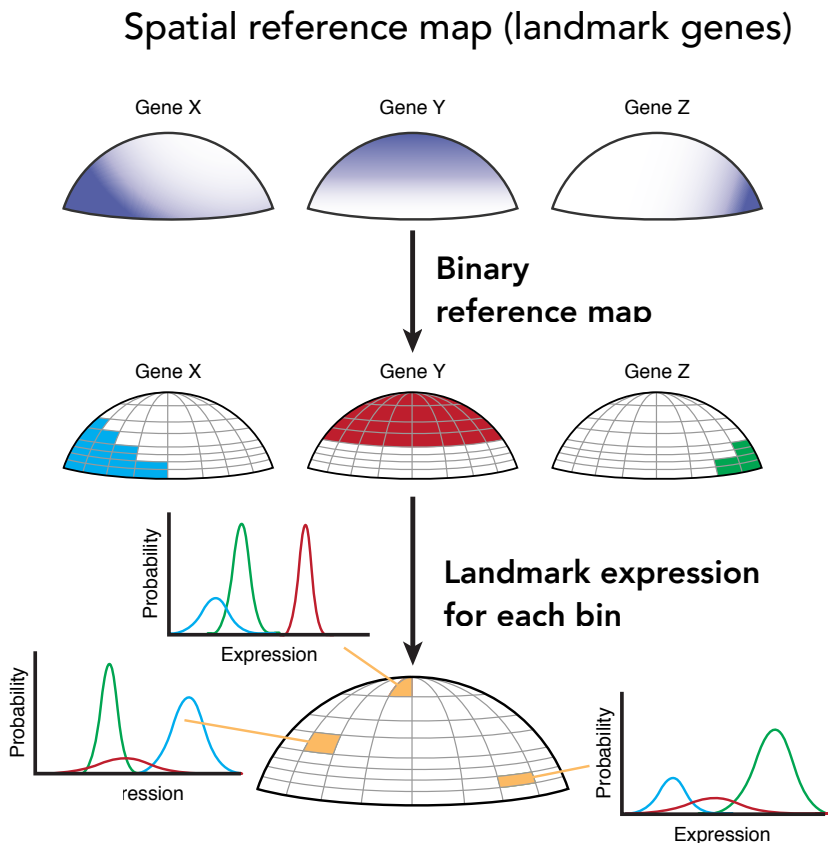
Seurat : Where did the cells come from?



Seurat : Where did the cells come from?

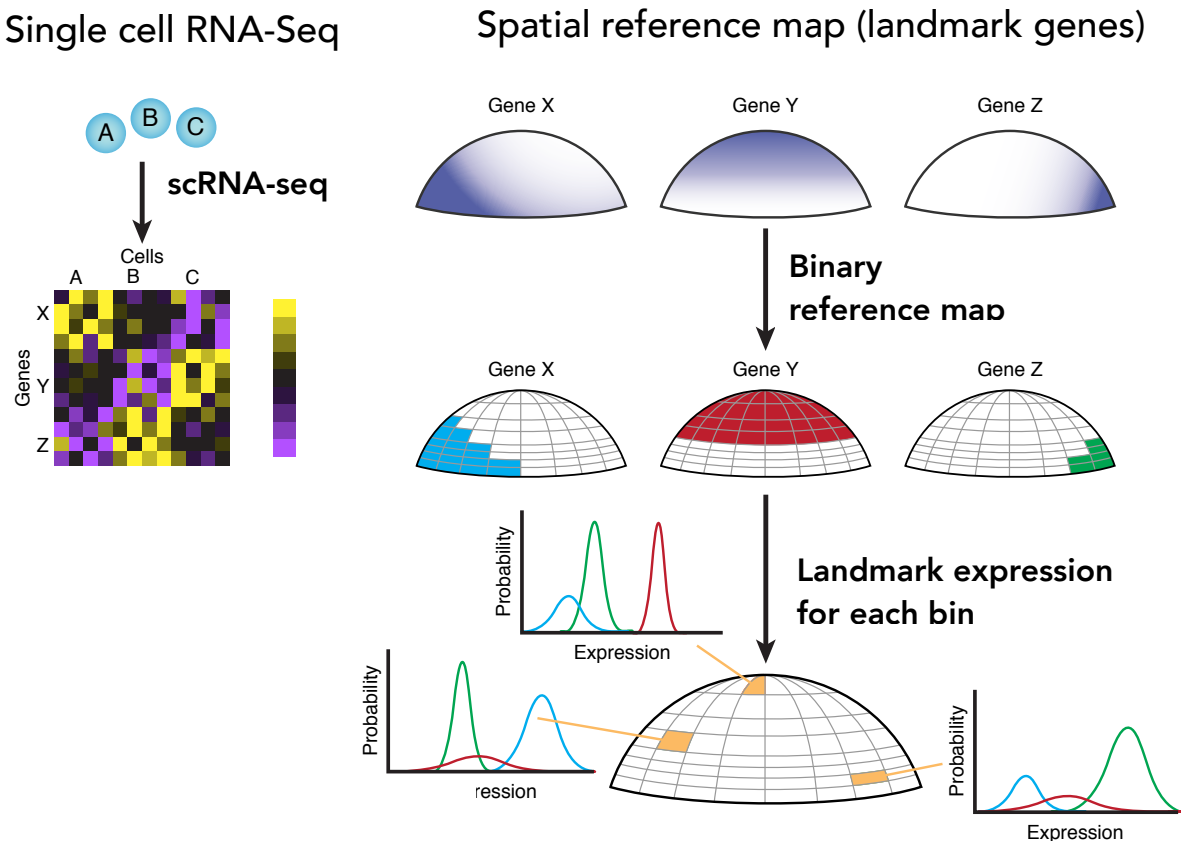


Seurat: Semi-supervised inference of spatial location from gene expression



Satija*, Ferrell* et al., Nature Biotechnology, 2016
also Junker et al., Cell, 2014
Achim et al., Nature Biotechnology, 2014

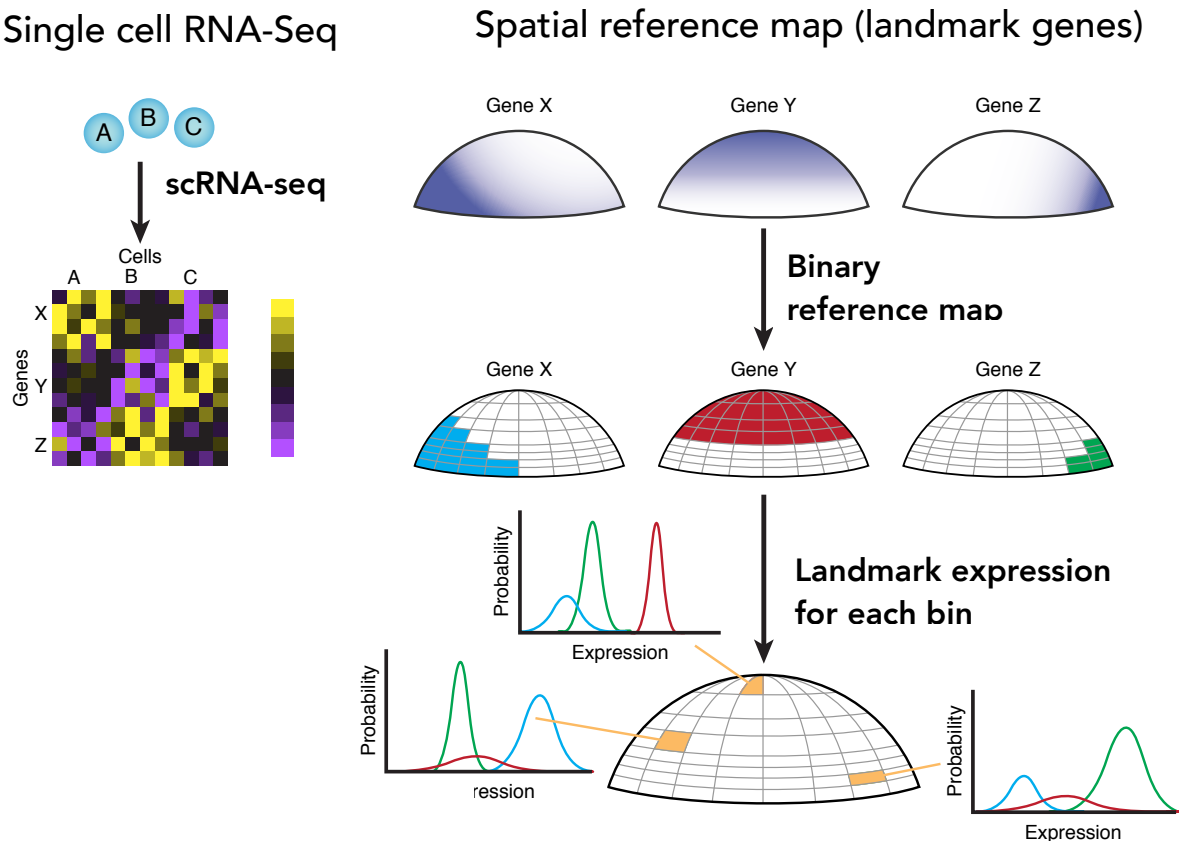
Seurat: Semi-supervised inference of spatial location from gene expression



Satija*, Ferrell* et al., Nature Biotechnology, 2016
also Junker et al., Cell, 2014
Achim et al., Nature Biotechnology, 2014

Seurat: Semi-supervised inference of spatial location from gene expression

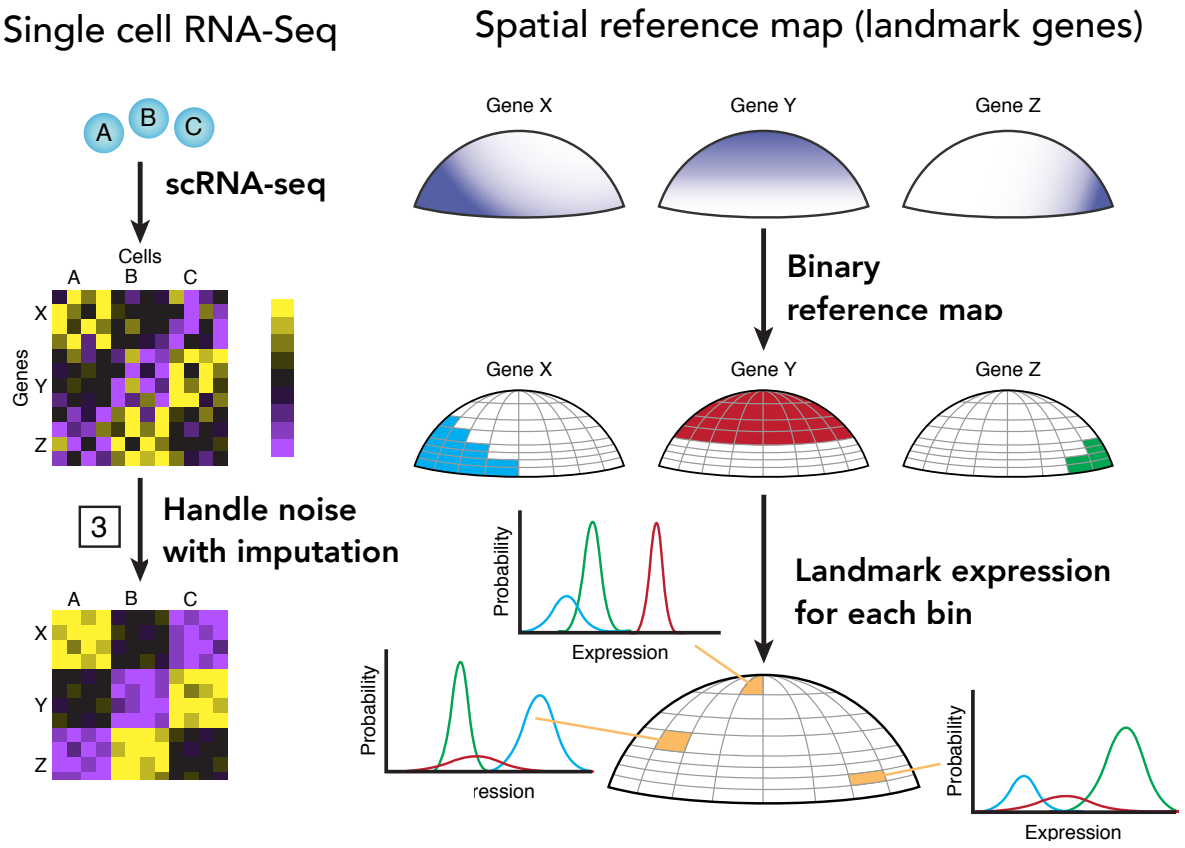
Dropouts!



Satija*, Ferrell* et al., Nature Biotechnology, 2016
also Junker et al., Cell, 2014
Achim et al., Nature Biotechnology, 2014

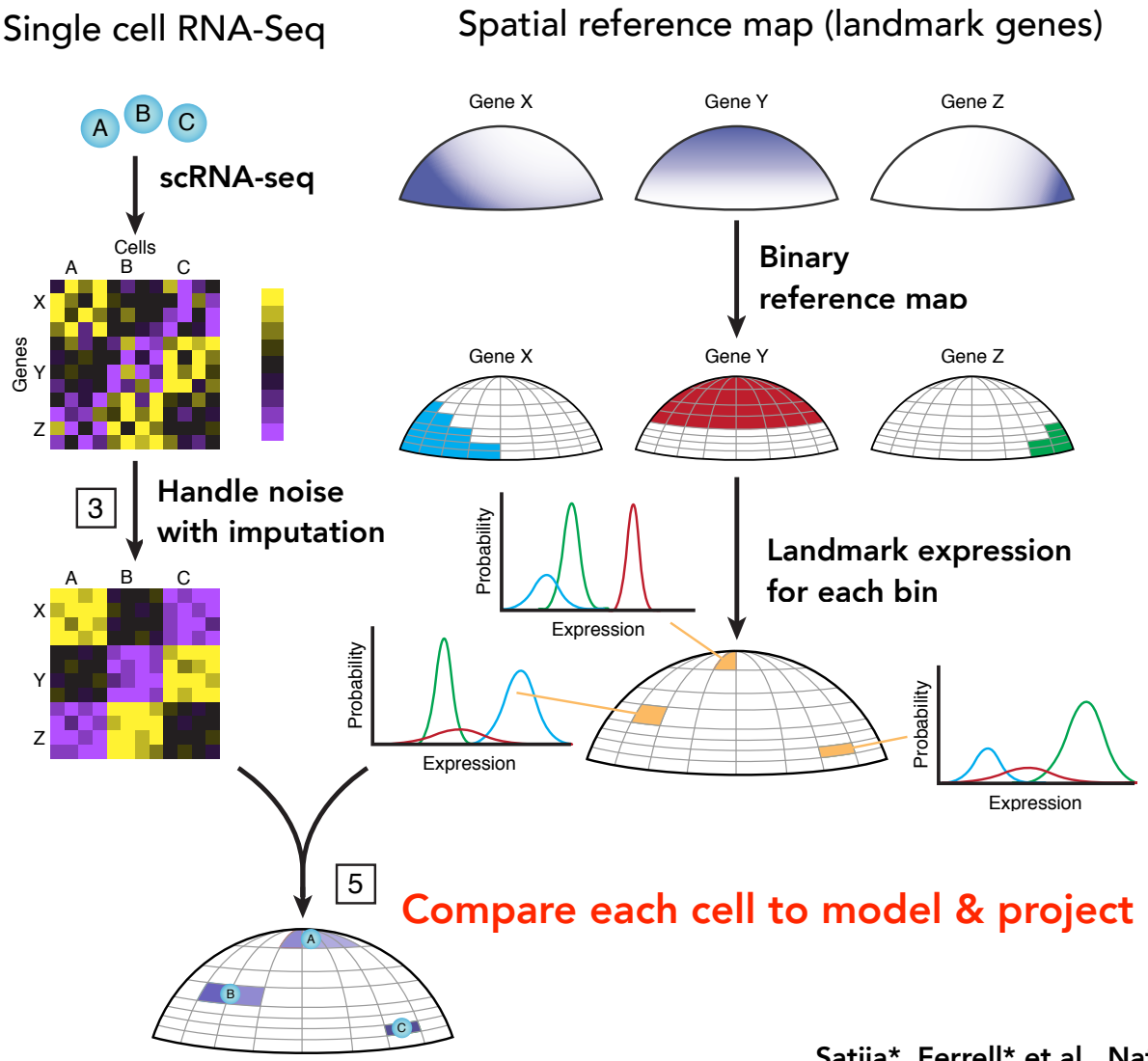
Seurat: Semi-supervised inference of spatial location from gene expression

Dropouts!



Satija*, Ferrell* et al., Nature Biotechnology, 2016
also Junker et al., Cell, 2014
Achim et al., Nature Biotechnology, 2014

Seurat: Semi-supervised inference of spatial location from gene expression

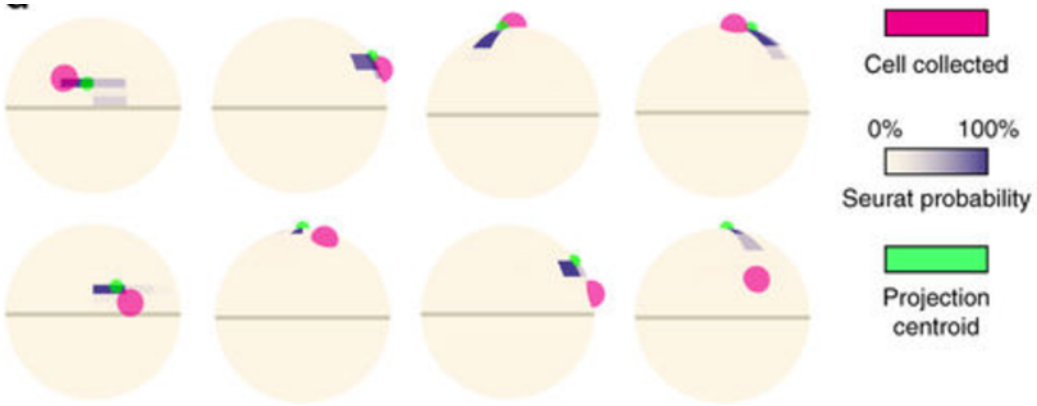


Dropouts!

Satija*, Ferrell* et al., Nature Biotechnology, 2016
also Junker et al., Cell, 2014
Achim et al., Nature Biotechnology, 2014

Spatial positioning of zebrafish embryo cells

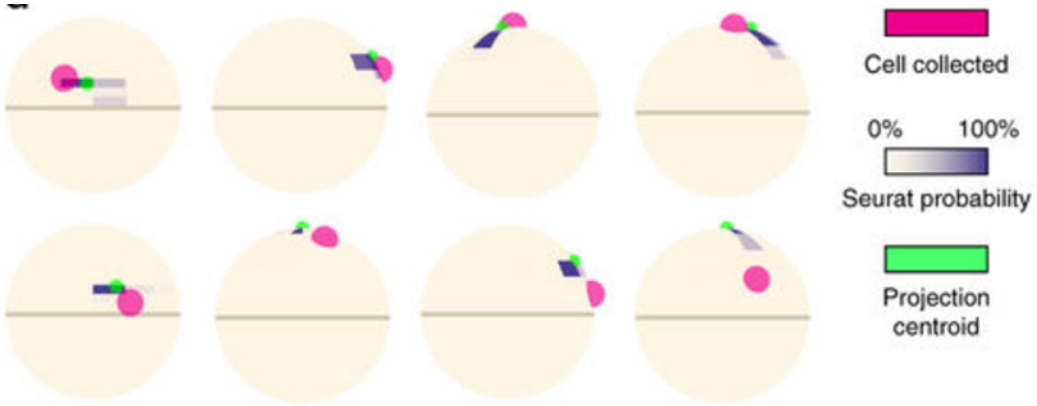
Accurate identification of the spatial position of handpicked cells



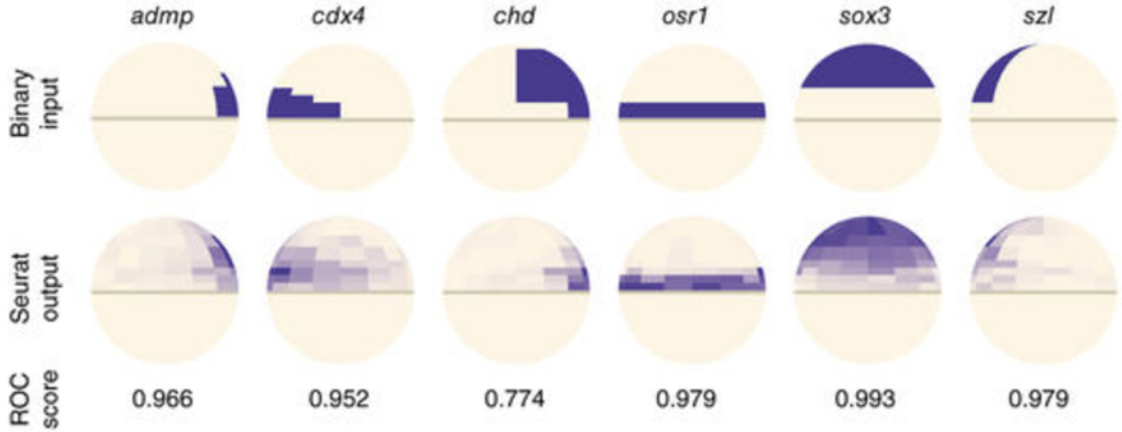
Satija*, Ferrell* et al., Nature Biotechnology, 2015

Spatial positioning of zebrafish embryo cells

Accurate identification of the spatial position of handpicked cells



Predicted spatial expression patterns of new genes



Satija*, Ferrell* et al., Nature Biotechnology, 2015

What can we learn?

1. Census and taxonomy

2. Anatomy and Physiology

3. Cancer

4. Development

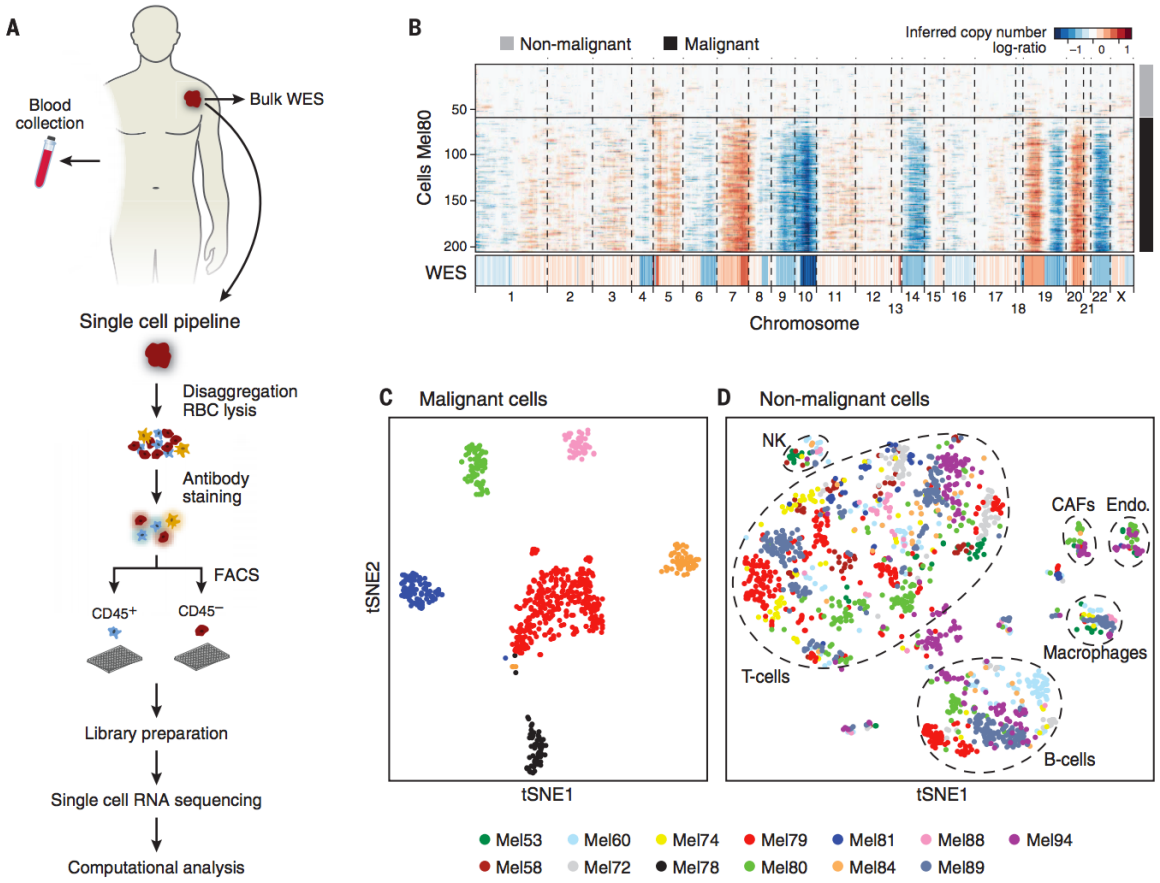
Cancers are multi-cellular ecosystems

- Defining the spectrum of malignant states (cancer stem cells)
- Understanding the role of the microenvironment (immune cells, stromal cells)
- Inter-patient heterogeneity, clonal evolution

Cancers are multi-cellular ecosystems

- Defining the spectrum of malignant states (cancer stem cells)
- Understanding the role of the microenvironment (immune cells, stromal cells)
- Inter-patient heterogeneity, clonal evolution

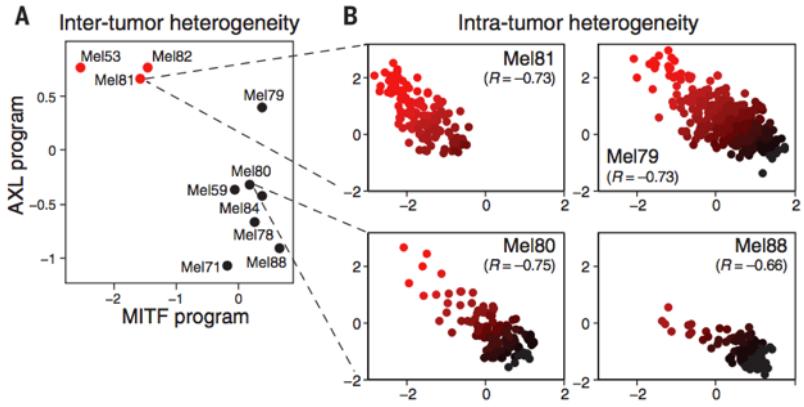
Metastatic Melanoma



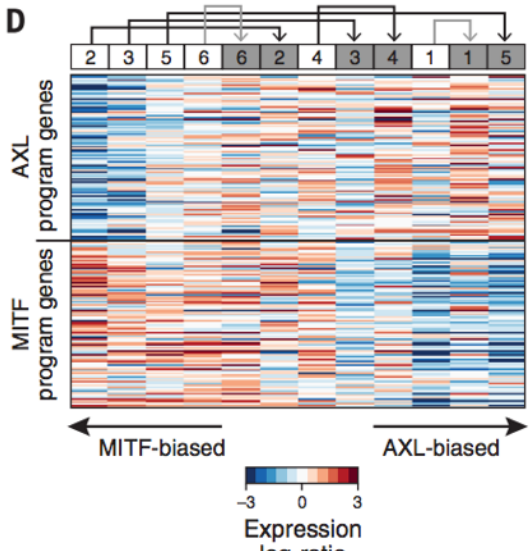
Tirosh et al., Science, 2016

Analyzing the microenvironment

Relative expression of AXL (treatment resistant)
vs. MITF (treatment sensitive) program in patients

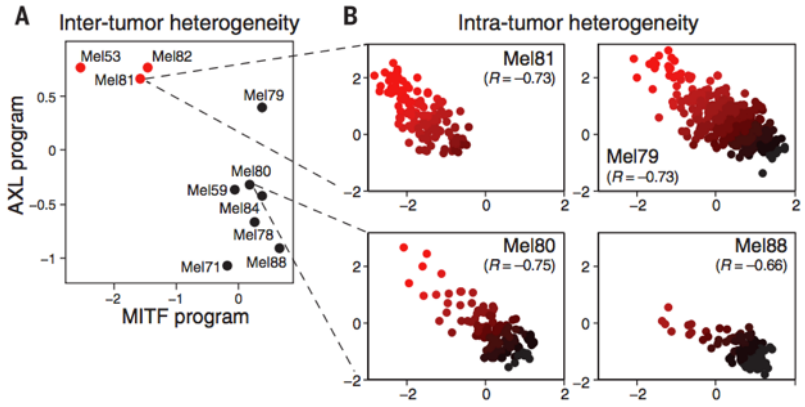


Cells in the post relapse patients are AXL high

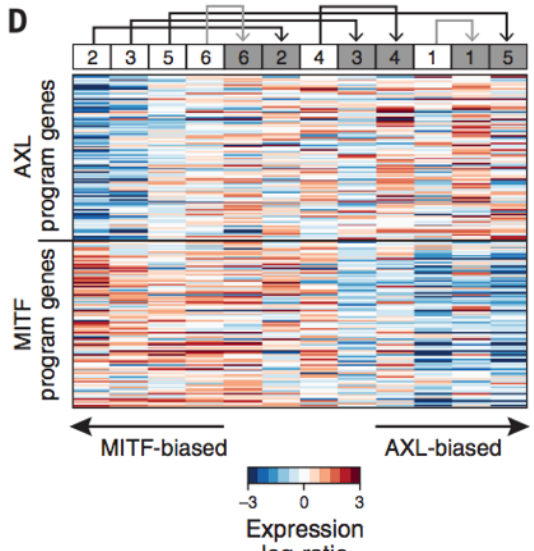


Analyzing the microenvironment

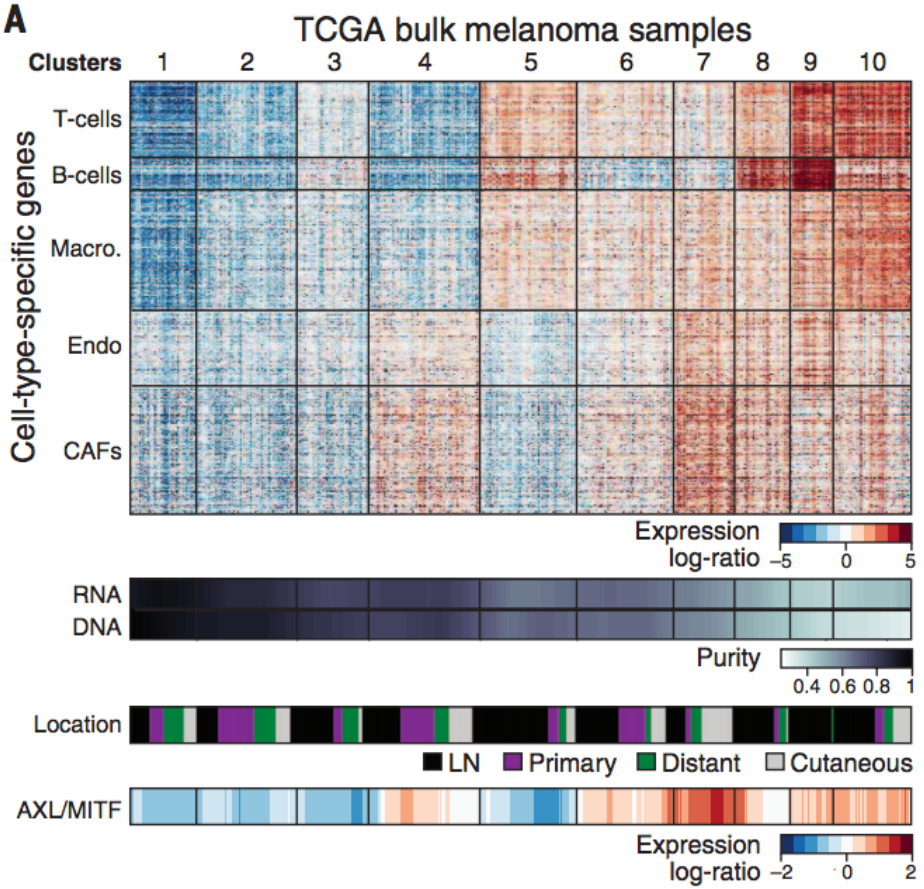
Relative expression of AXL (treatment resistant) vs. MITF (treatment sensitive) program in patients



Cells in the post relapse patients are AXL high



Bulk deconvolution : Single-cell guided analysis of TCGA samples suggests that AXL high tumors are distinguished by abundance of cancer-associated fibroblasts



What can we learn?

1. Census and taxonomy

2. Anatomy and Physiology

3. Cancer

4. Development

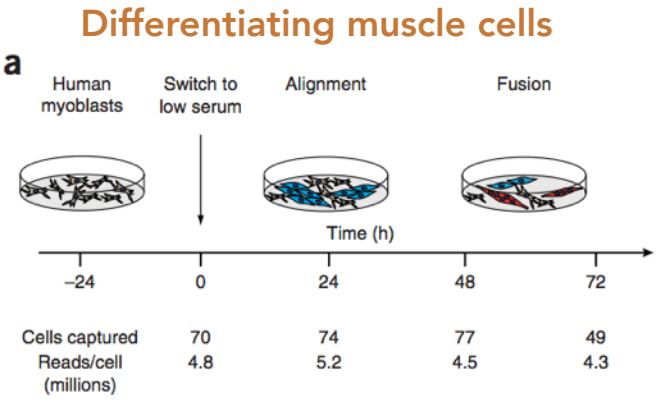
Learning lineages from single cell data

- Single-cell measurements are inherently destructive, so we cannot follow the transcriptome of the same cell over time
- Lineage has to be inferred by connecting the dots …
- Steady-state (**hematopoesis**) vs. non-steady state (**neural development**)

Trapnell et al., Nat. Biotech, 2014

Learning lineages from single cell data

- Single-cell measurements are inherently destructive, so we cannot follow the transcriptome of the same cell over time
- Lineage has to be inferred by connecting the dots …
- Steady-state (**hematopoesis**) vs. non-steady state (**neural development**)

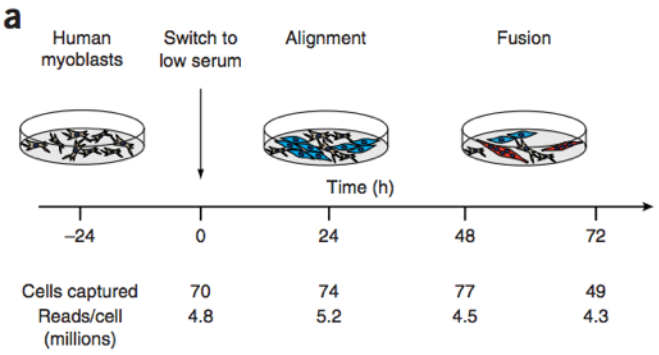


Trapnell et al., Nat. Biotech, 2014

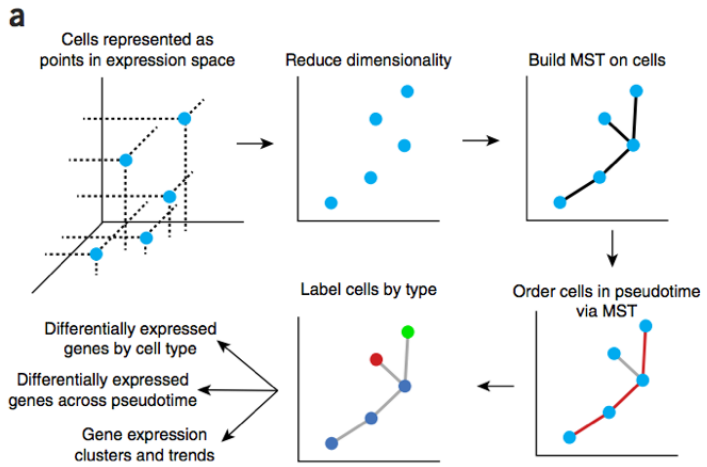
Learning lineages from single cell data

- Single-cell measurements are inherently destructive, so we cannot follow the transcriptome of the same cell over time
- Lineage has to be inferred by connecting the dots …
- Steady-state (**hematopoesis**) vs. non-steady state (**neural development**)

Differentiating muscle cells



Monocle : Building trajectories in space

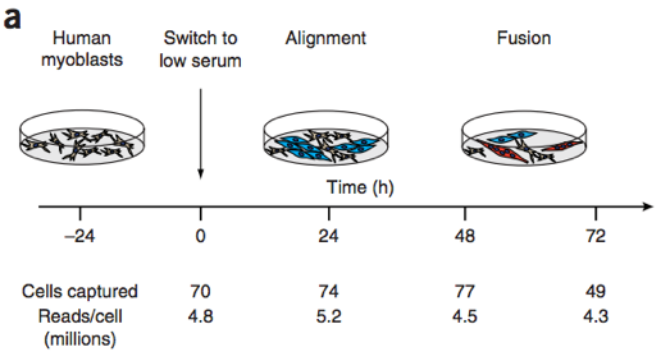


Trapnell et al., Nat. Biotech, 2014

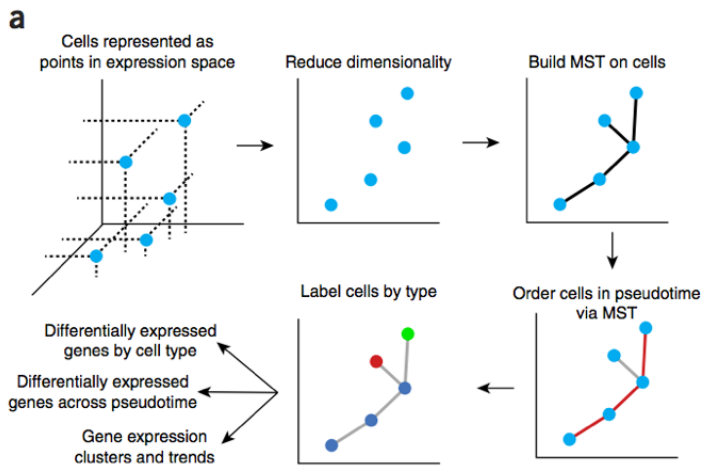
Learning lineages from single cell data

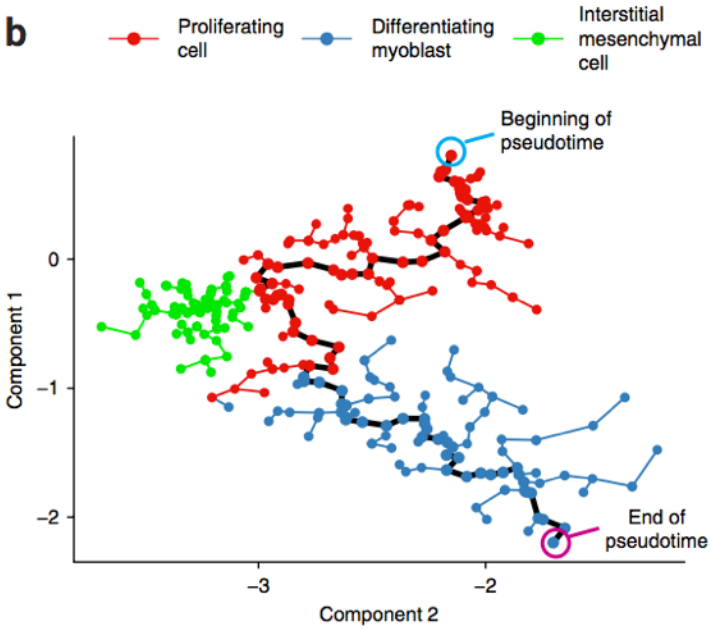
- Single-cell measurements are inherently destructive, so we cannot follow the transcriptome of the same cell over time
- Lineage has to be inferred by connecting the dots …
- Steady-state (**hematopoesis**) vs. non-steady state (**neural development**)

Differentiating muscle cells



Monocle : Building trajectories in space

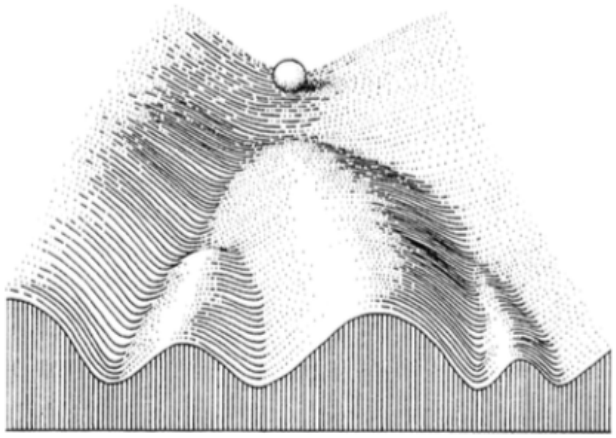




Trapnell et al., Nat. Biotech, 2014

Time as an explicit variable

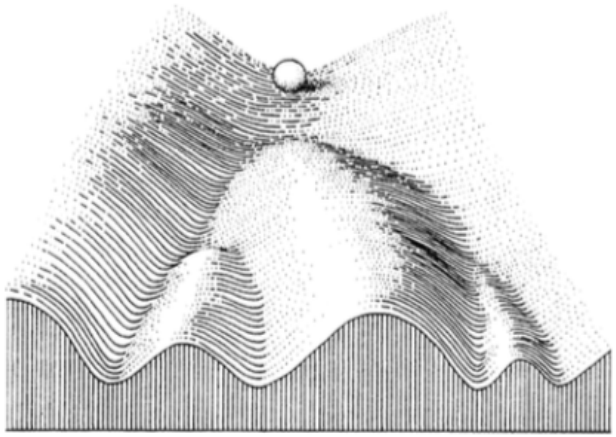
Waddington's landscape

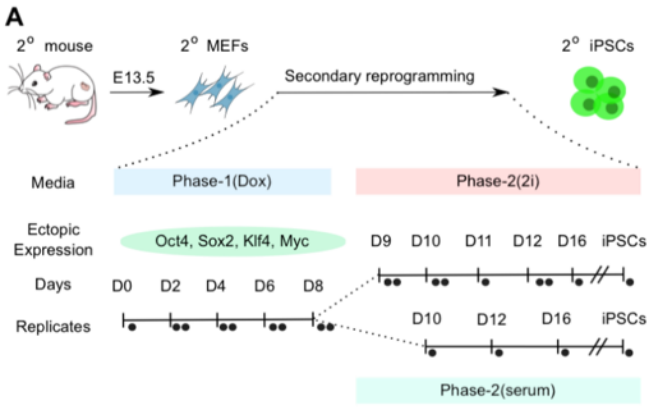


Scheibinger, Shu, Tabaka, Cleary et al., 2017 (collaboration with Lander lab)

Time as an explicit variable

Waddington's landscape





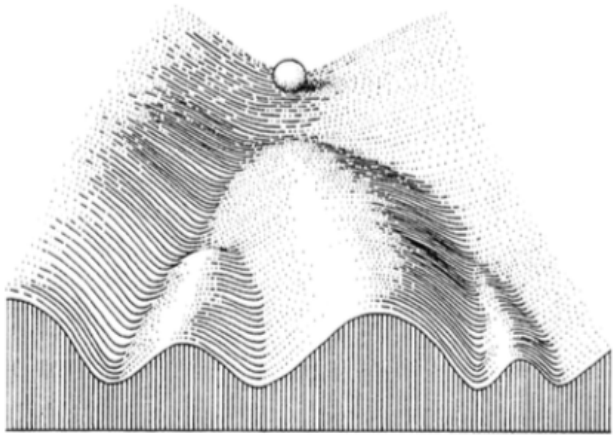
B

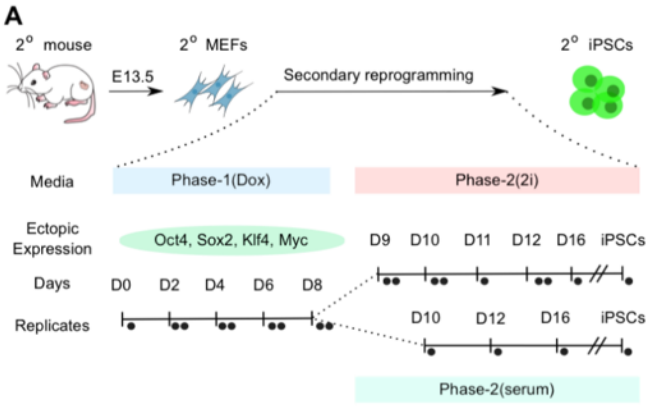
	Sample	Cell Number
Phase-1(Dox)	D0	4,060
	D2-1	2,890
	D2-2	2,729
	D4-1	2,882
	D4-2	3,962
	D6-1	3,198
	D6-2	3,168
	D8-1	2,142
	D8-2	2,625
Phase-2 (2i)	D9-1	2,441
	D9-2	2,174
	D10-1	2,878
	D10-2	2,619
	D11	1,529
	D12-1	5,139
	D12-2	2,155
	D16	4,500
	iPSCs	2,916
Phase-2 (serum)	D10	2,088
	D12	2,895
	D16	3,703
	iPSCs	3,088
	Total	65,781

Scheibinger, Shu, Tabaka, Cleary et al., 2017 (collaboration with Lander lab)

Time as an explicit variable

Waddington's landscape

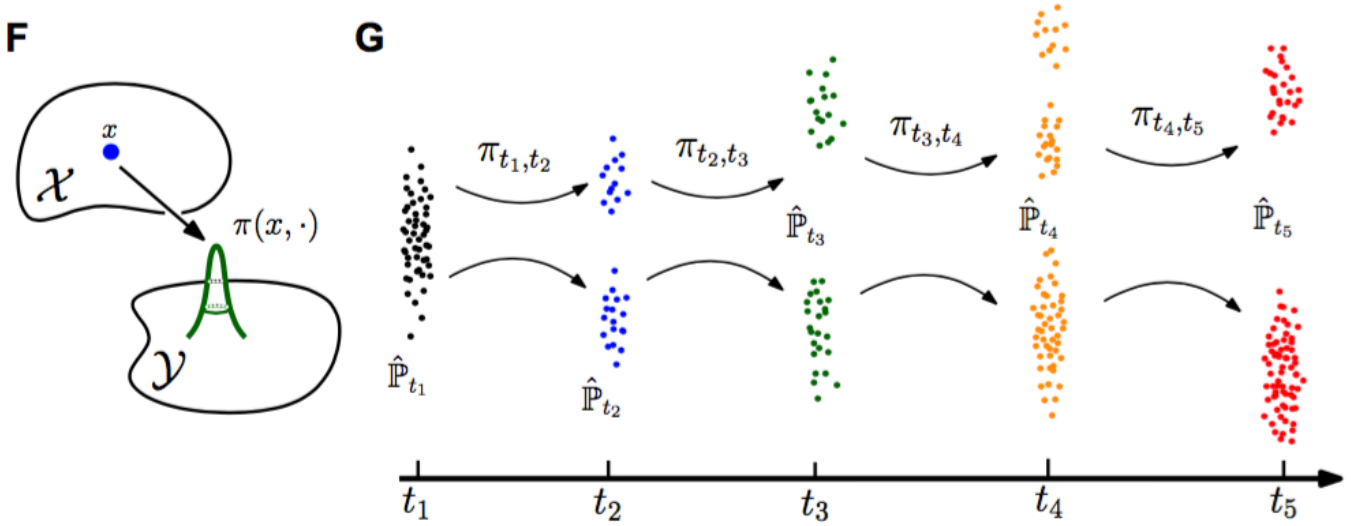




B

	Sample	Cell Number
Phase-1(Dox)	D0	4,060
	D2-1	2,890
	D2-2	2,729
	D4-1	2,882
	D4-2	3,962
	D6-1	3,198
	D6-2	3,168
	D8-1	2,142
	D8-2	2,625
Phase-2 (2i)	D9-1	2,441
	D9-2	2,174
	D10-1	2,878
	D10-2	2,619
	D11	1,529
	D12-1	5,139
	D12-2	2,155
	D16	4,500
	iPSCs	2,916
Phase-2 (serum)	D10	2,088
	D12	2,895
	D16	3,703
	iPSCs	3,088
	Total	65,781

How does a distribution of cell states at time t1 change at time t2?

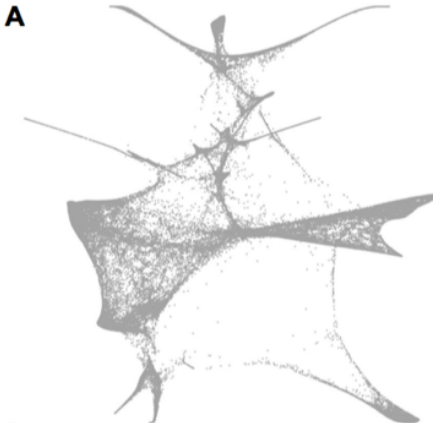


Scheibinger, Shu, Tabaka, Cleary et al., 2017 (collaboration with Lander lab)

Optimal Transport : Results

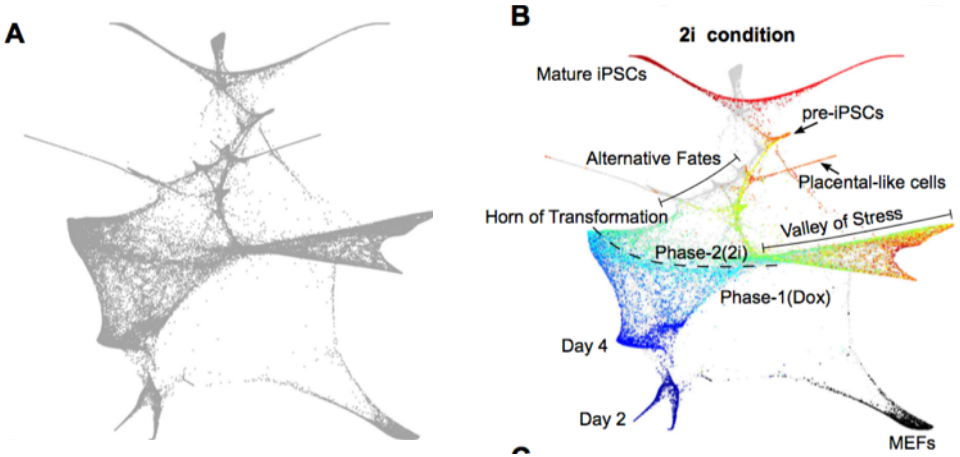
Scheibinger, Shu, Tabaka, Cleary et al., 2017

Optimal Transport : Results



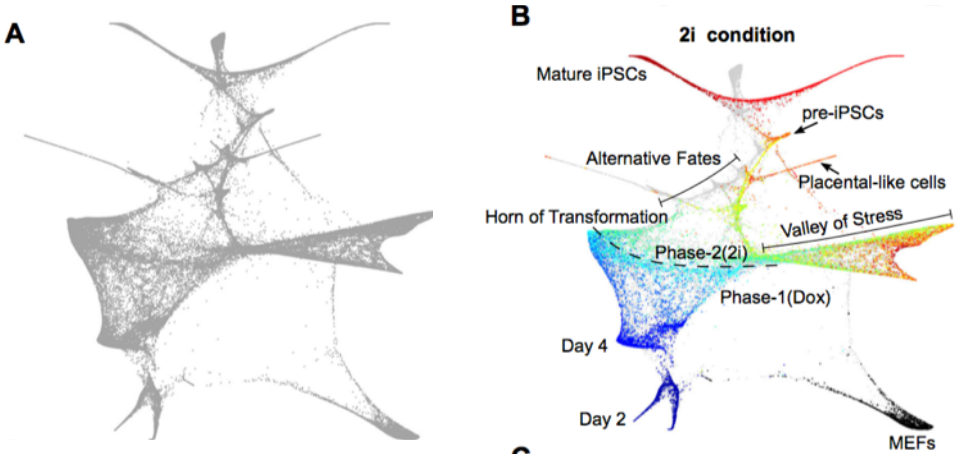
Scheibinger, Shu, Tabaka, Cleary et al., 2017

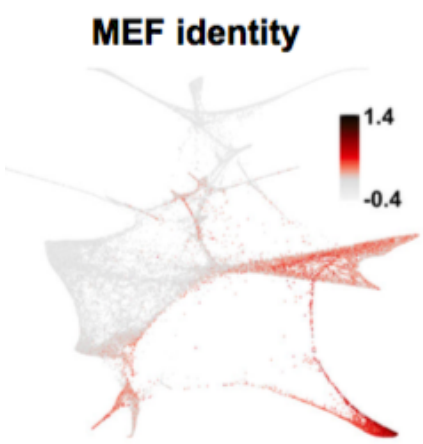
Optimal Transport : Results

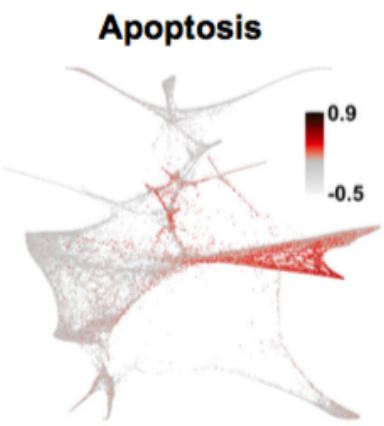


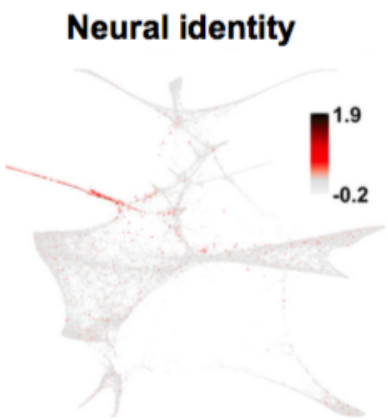
Scheibinger, Shu, Tabaka, Cleary et al., 2017

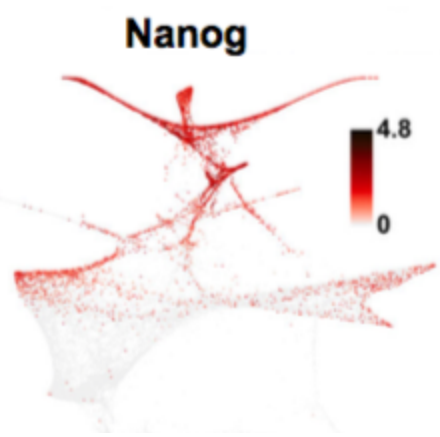
Optimal Transport : Results





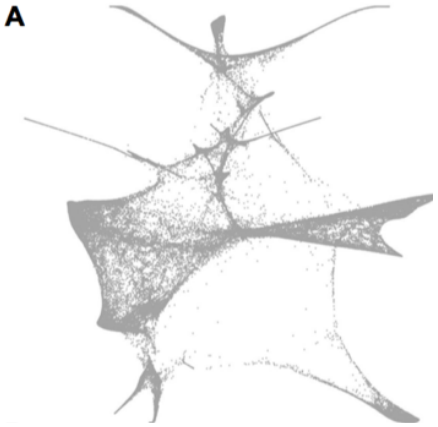


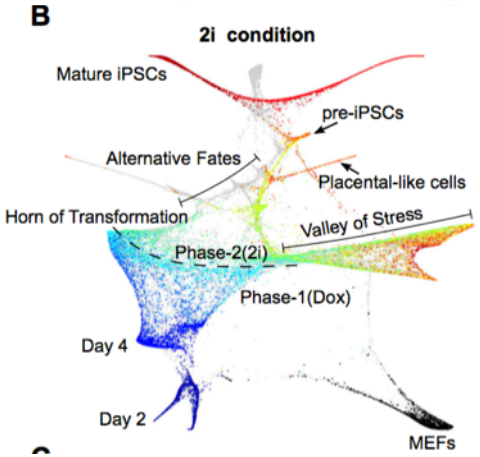


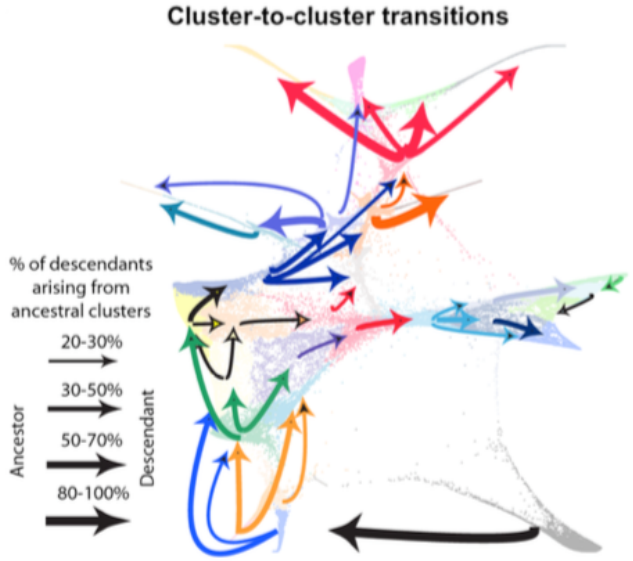


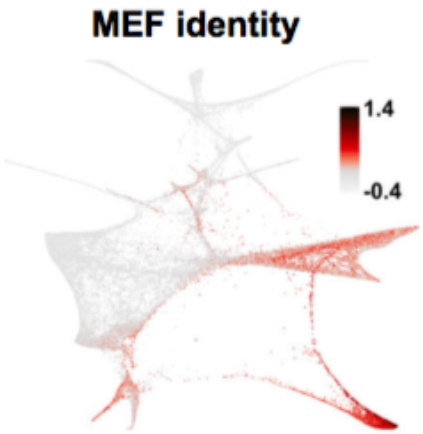
Scheibinger, Shu, Tabaka, Cleary et al., 2017

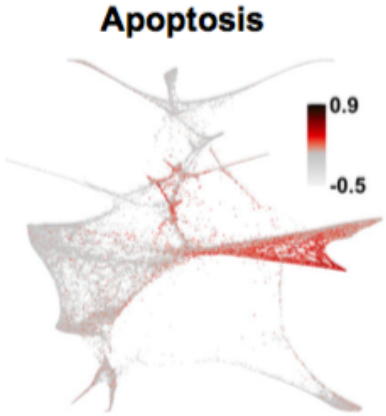
Optimal Transport : Results

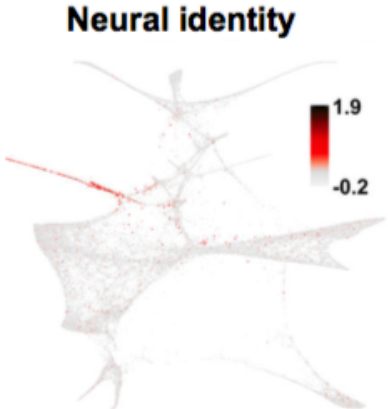


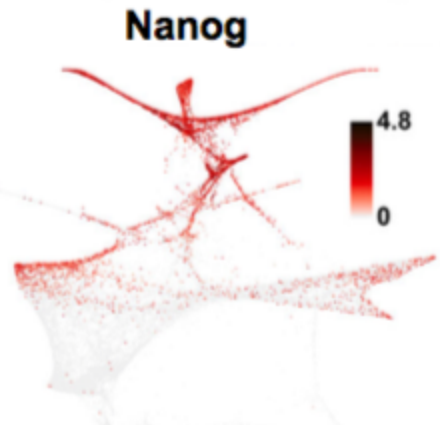












Scheibinger, Shu, Tabaka, Cleary et al., 2017

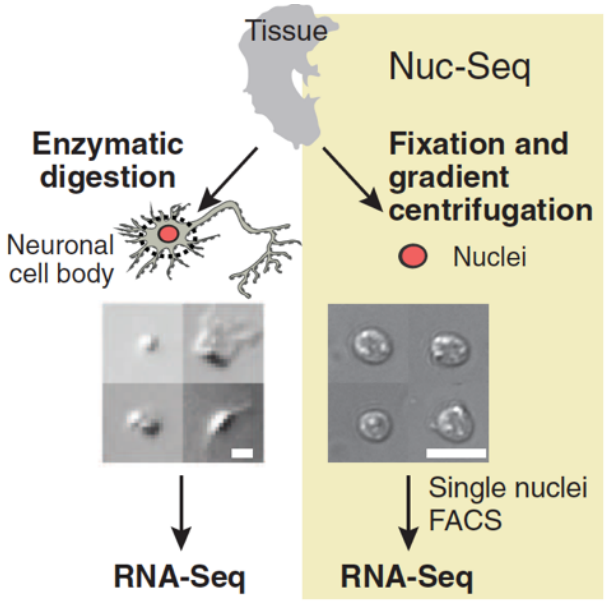
Emerging Technologies

- Single-nucleus seq
- Combi-seq
- Perturb-seq

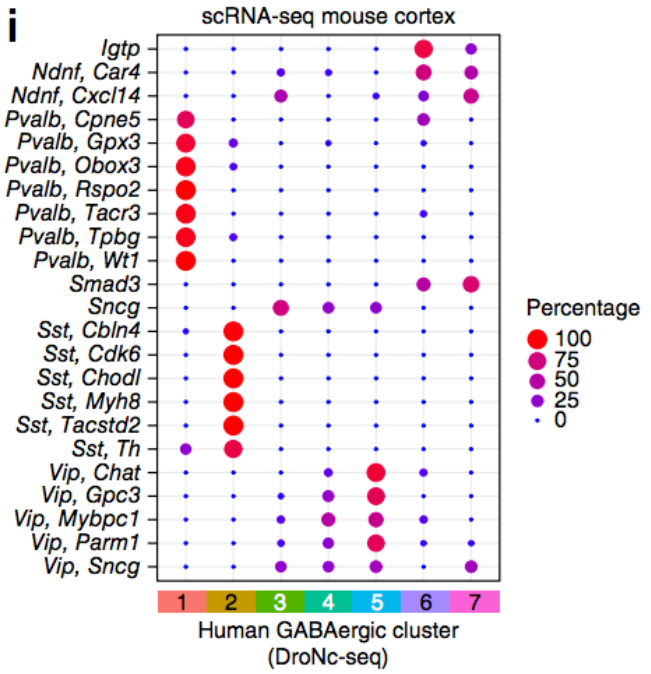
When scRNA-seq is not possible

- Many tissues are difficult to enzymatically dissociate - this is especially true for fixed or flash-frozen human post-mortem samples
- **The solution : Nuclei-seq!**

Nuclei can be isolated by gentle centrifugation



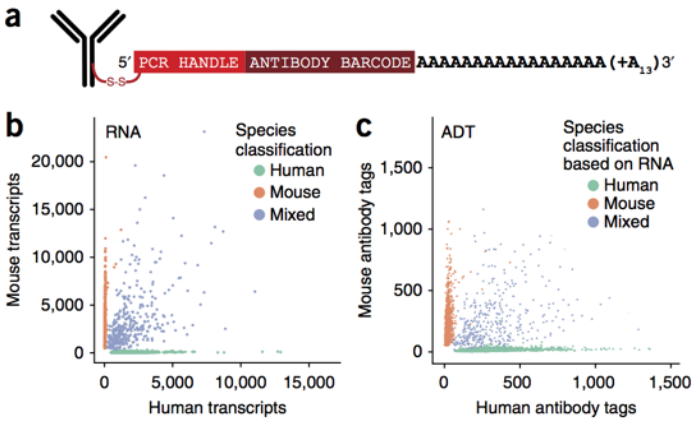
Molecularly distinct cell-types can be detected through nuclei

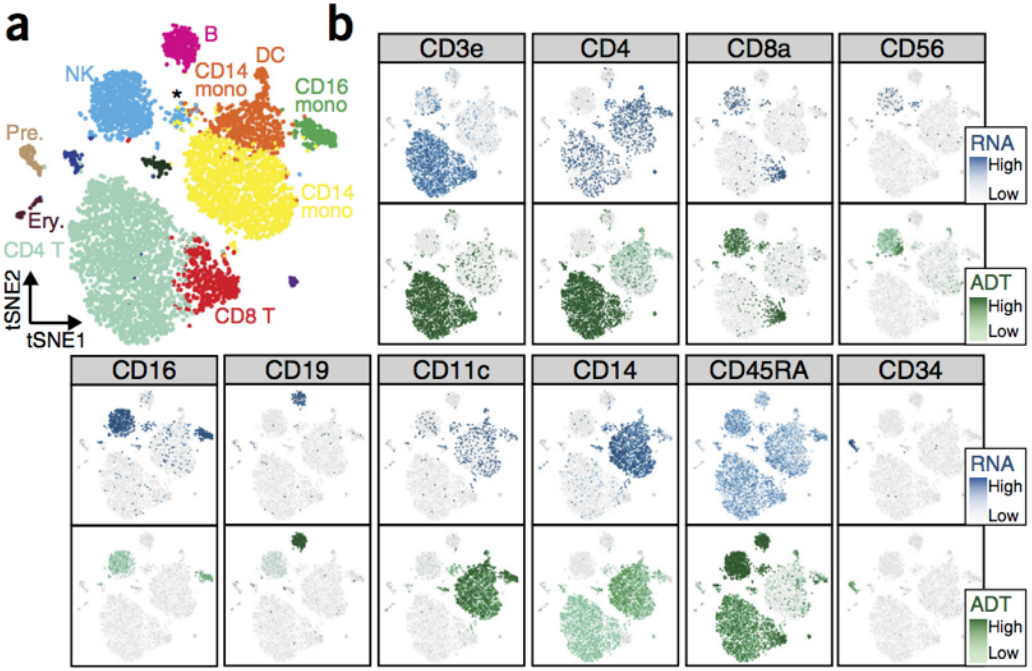


Habib et al., 2015, 2017

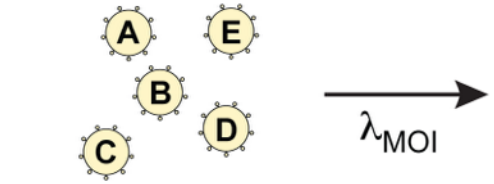
Simultaneous protein and RNA measurement in single cells

- Many tissues are difficult to enzymatically dissociate - this is especially true for fixed or flash-frozen human post-mortem samples





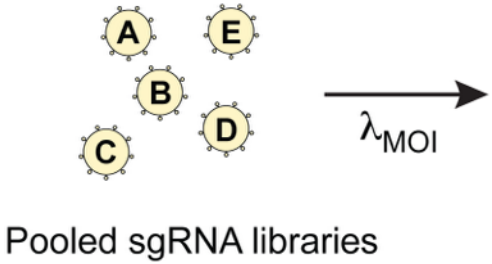
Perturb-Seq: Pooled CRISPR screens with scRNA-seq



Pooled sgRNA libraries

Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

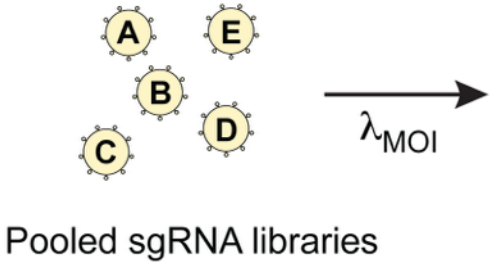
Perturb-Seq: Pooled CRISPR screens with scRNA-seq



I. Pooled sgRNA library,
polydenylated guide barcode

Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

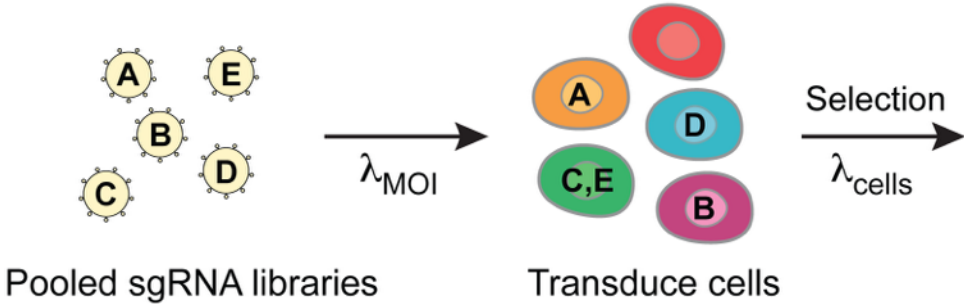
Perturb-Seq: Pooled CRISPR screens with scRNA-seq



I. Pooled sgRNA library,
polydenylated guide barcode

Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

Perturb-Seq: Pooled CRISPR screens with scRNA-seq

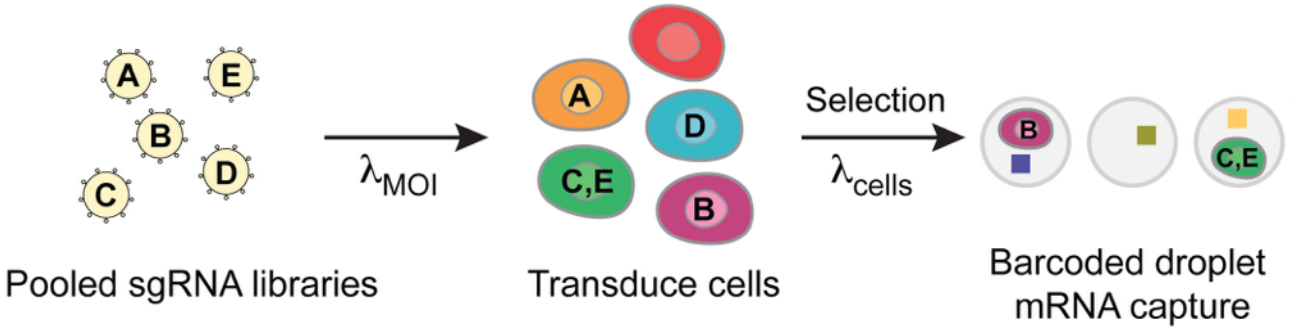


I. Pooled sgRNA library, polydenylated guide barcode

II. Transduce cells (low or high MOI)

Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

Perturb-Seq: Pooled CRISPR screens with scRNA-seq



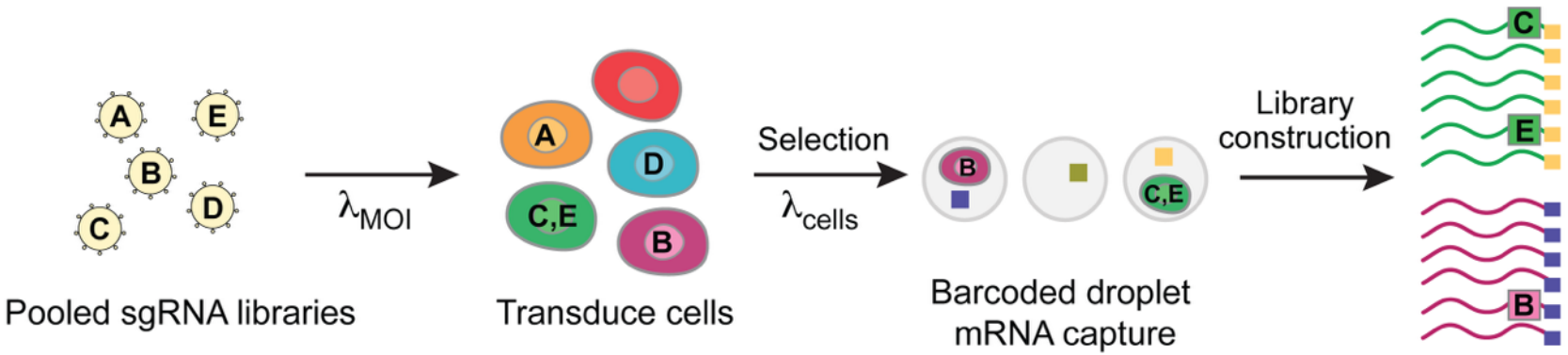
I. Pooled sgRNA library, polydenylated guide barcode

II. Transduce cells (low or high MOI)

III. Capture cells, barcode mRNA

Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

Perturb-Seq: Pooled CRISPR screens with scRNA-seq



I. Pooled sgRNA library, polydenylated guide barcode

II. Transduce cells (low or high MOI)

III. Capture cells, barcode mRNA

IV. Match guide barcode to transcriptome

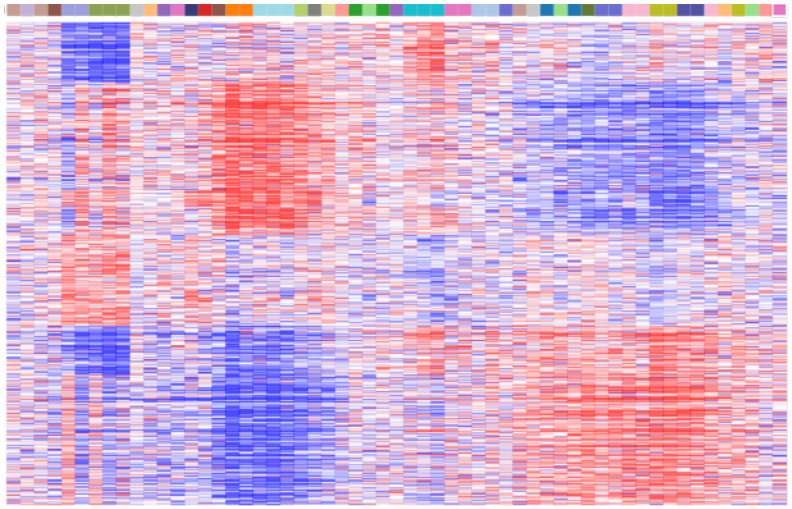
Dixit*, Parnas* et al., *Cell* 2016; Adamson*, Norman* et al., *Cell* 2016

Regulatory response of Dendritic Cells to LPS stimulation

Guides

(color code: target gene)

Genes



Regulatory coefficient

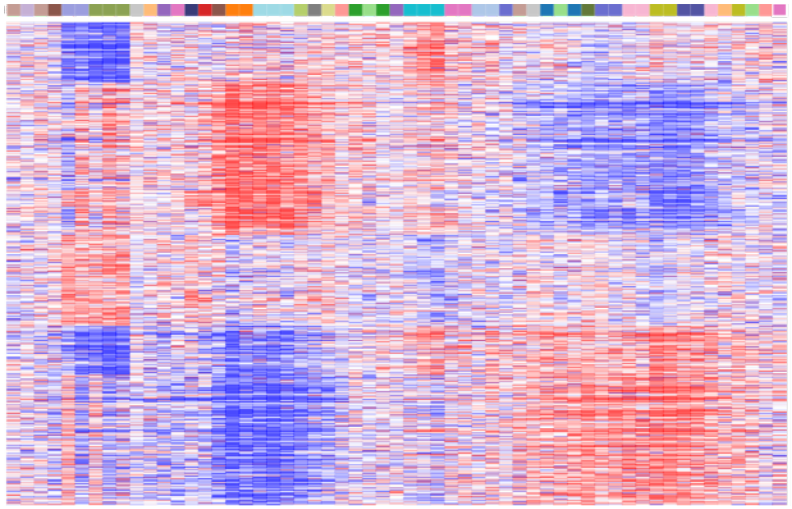
min max

Regulatory response of Dendritic Cells to LPS stimulation

Guides
(color code: target gene)

Grouping of guides for the same gene

Genes



Regulatory coefficient
min max

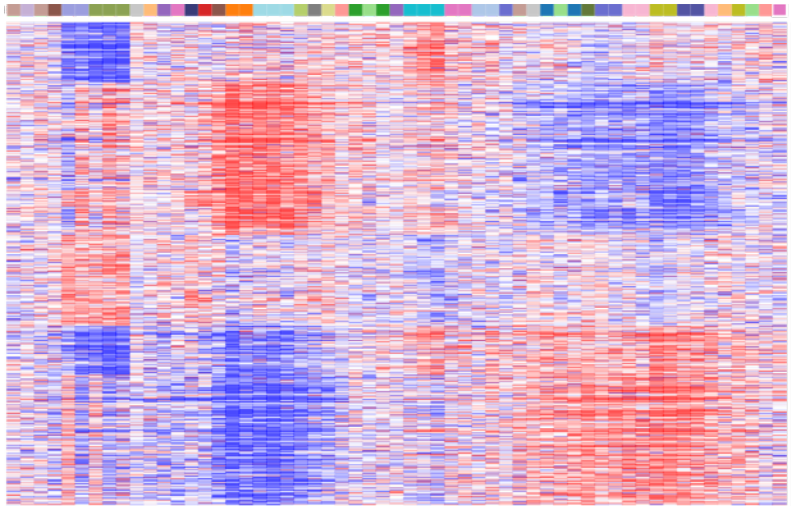
Regulatory response of Dendritic Cells to LPS stimulation

Guides

(color code: target gene)

PROGRAMS

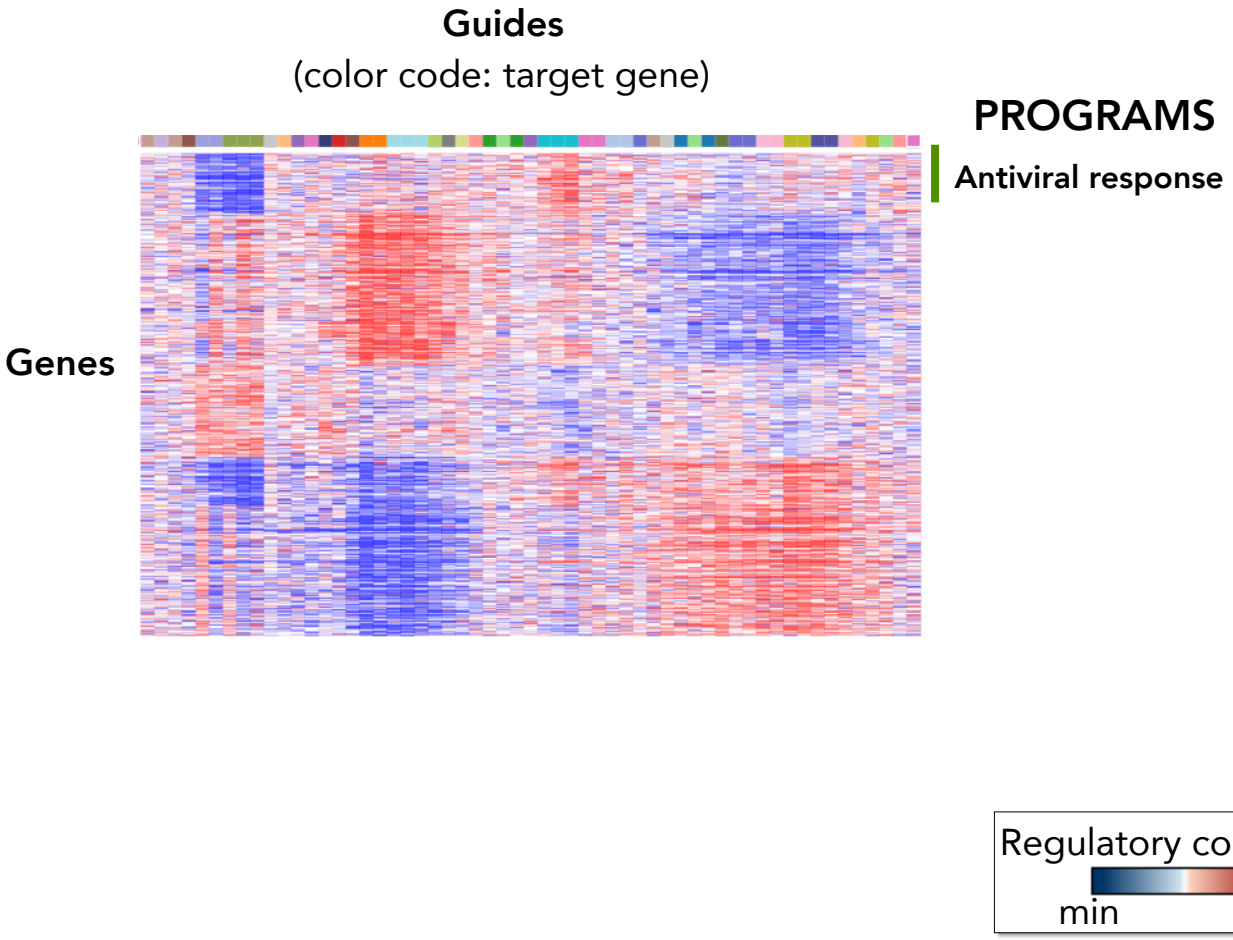
Genes



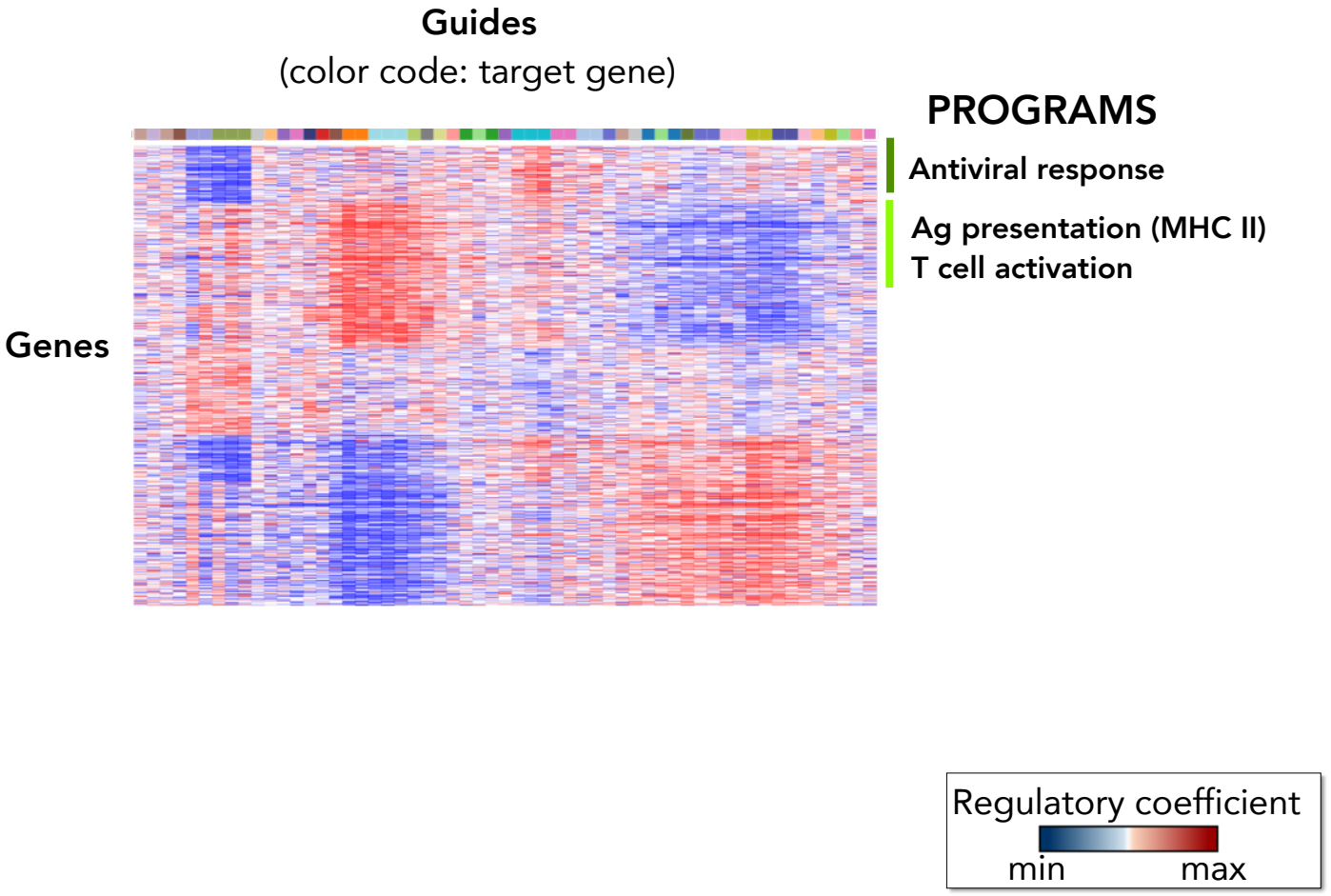
Regulatory coefficient

min max

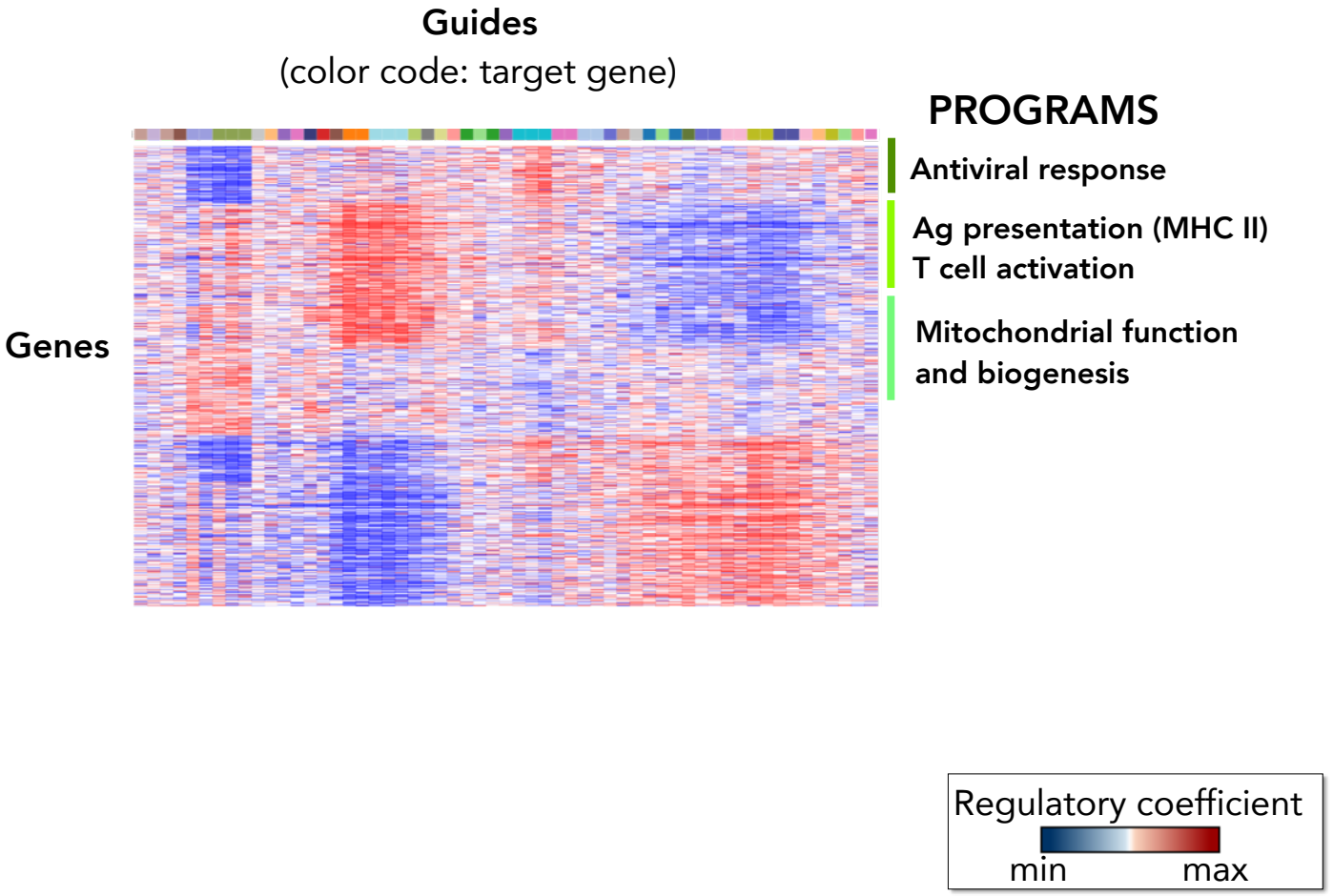
Regulatory response of Dendritic Cells to LPS stimulation



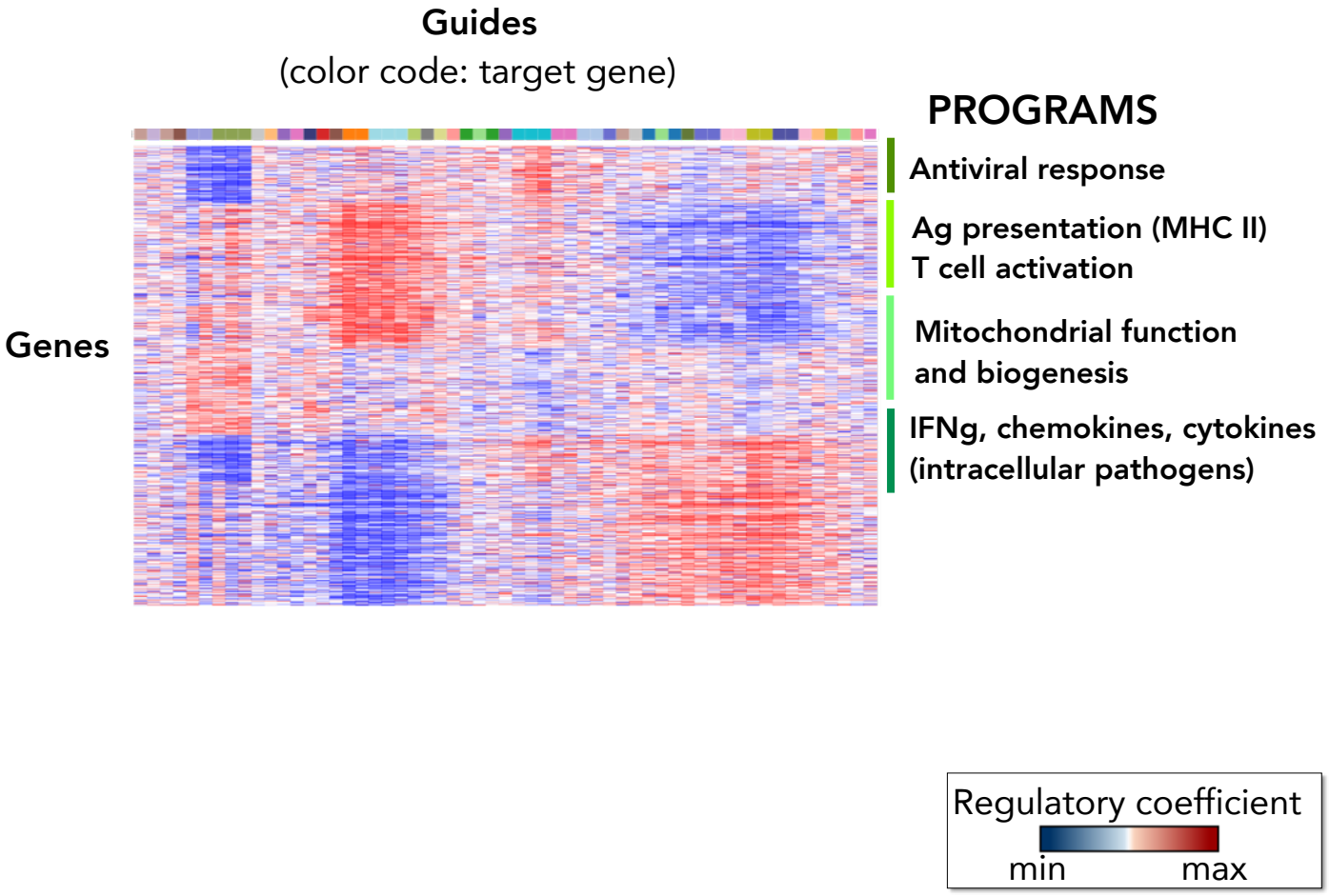
Regulatory response of Dendritic Cells to LPS stimulation



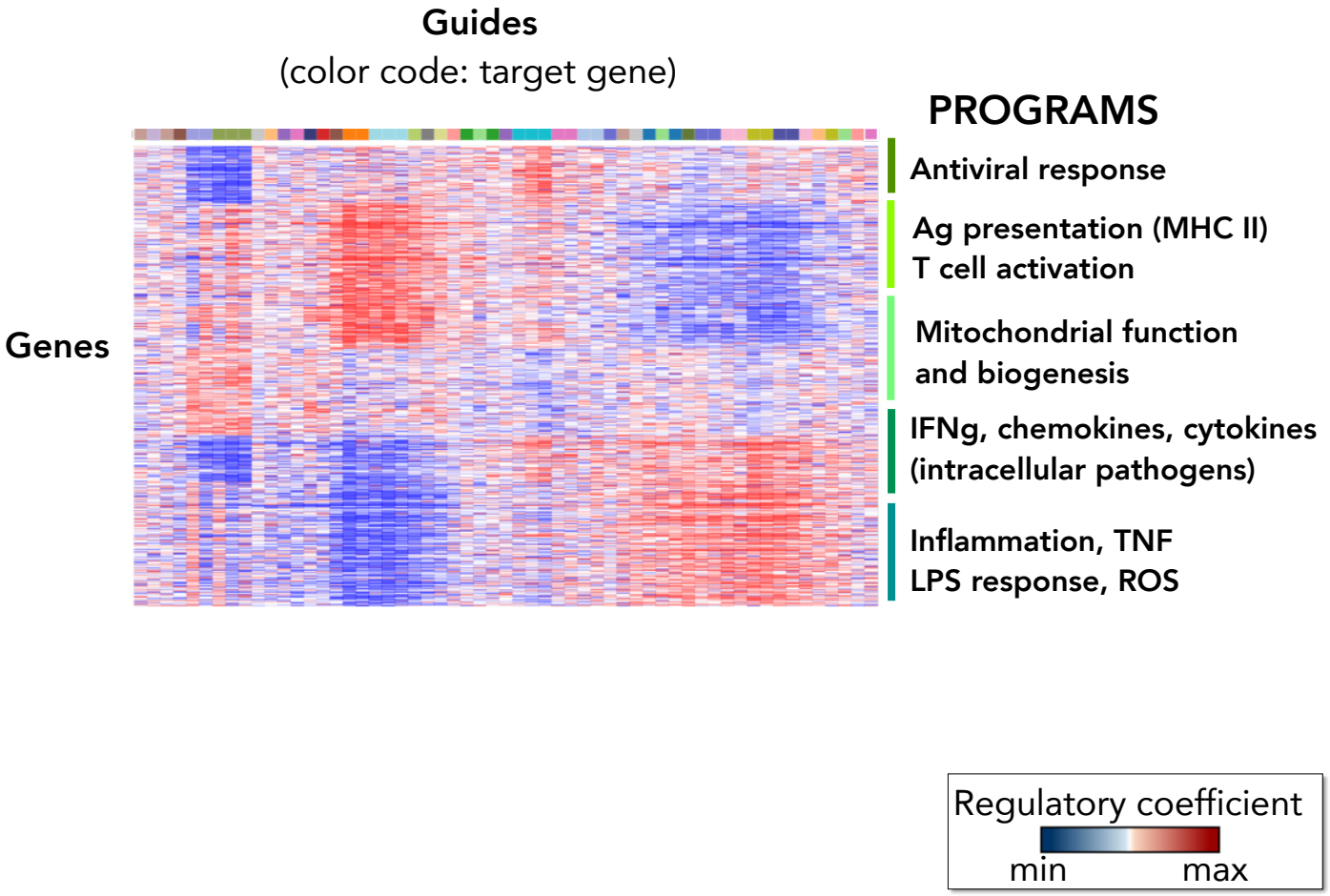
Regulatory response of Dendritic Cells to LPS stimulation



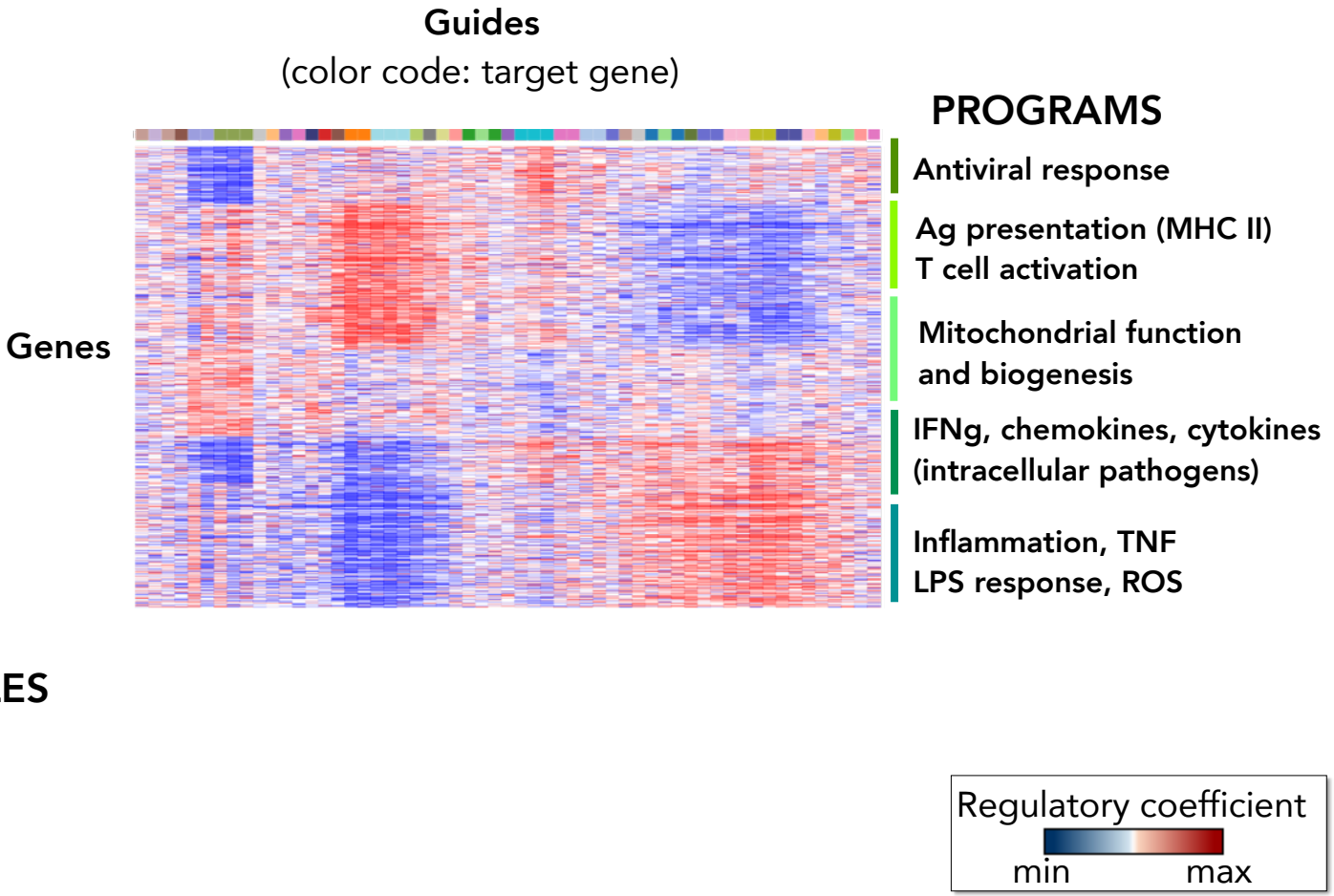
Regulatory response of Dendritic Cells to LPS stimulation



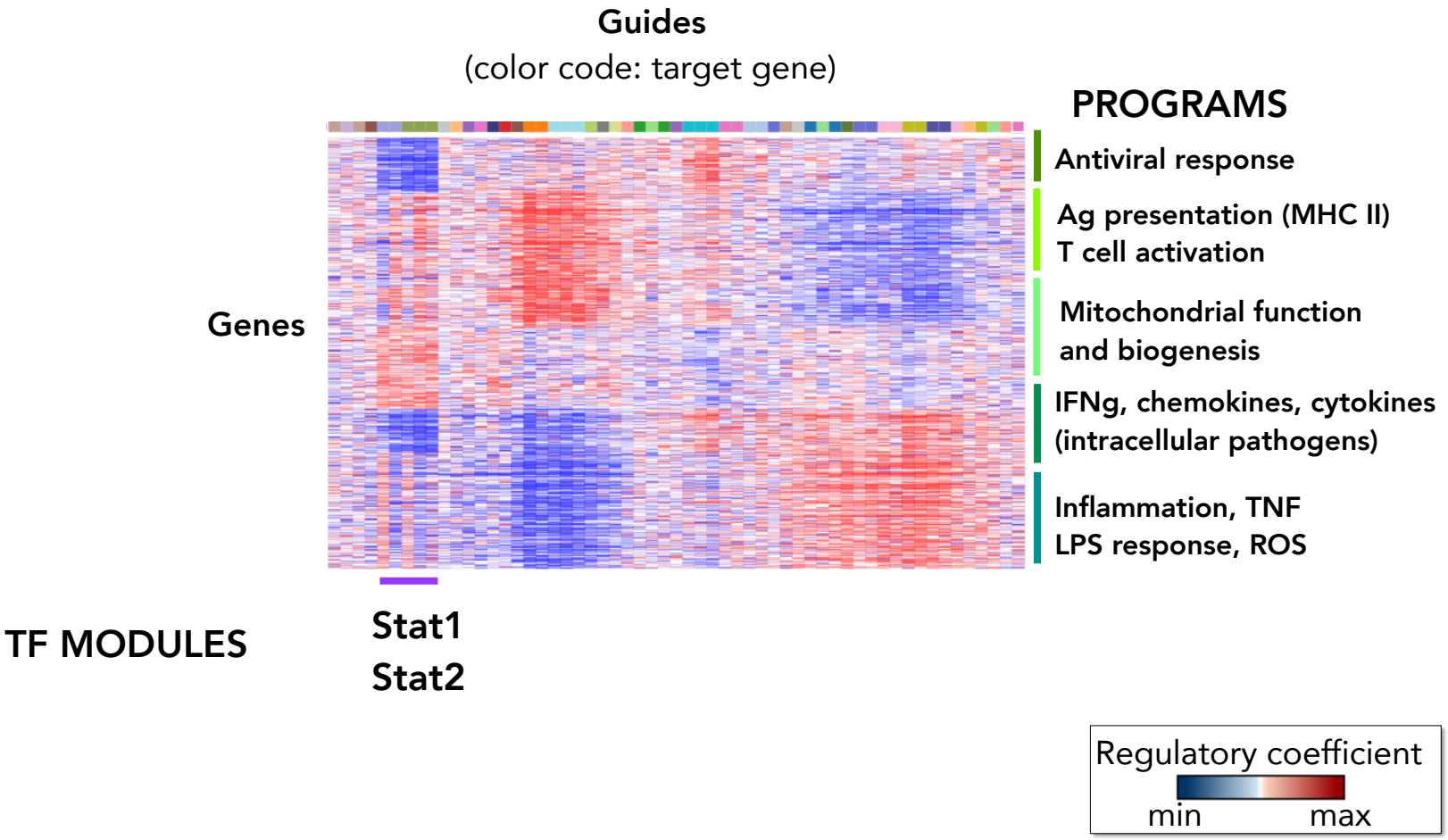
Regulatory response of Dendritic Cells to LPS stimulation



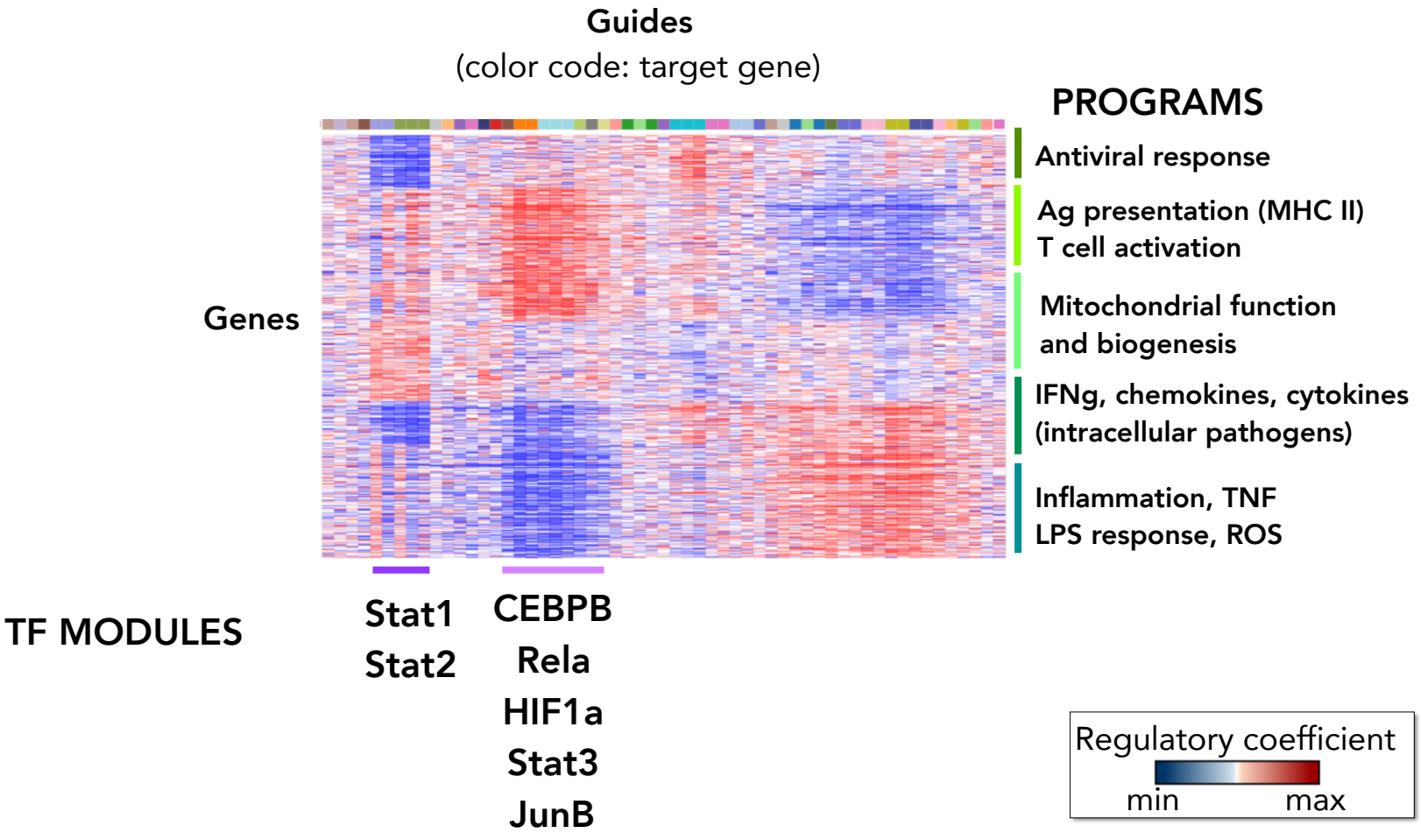
Regulatory response of Dendritic Cells to LPS stimulation



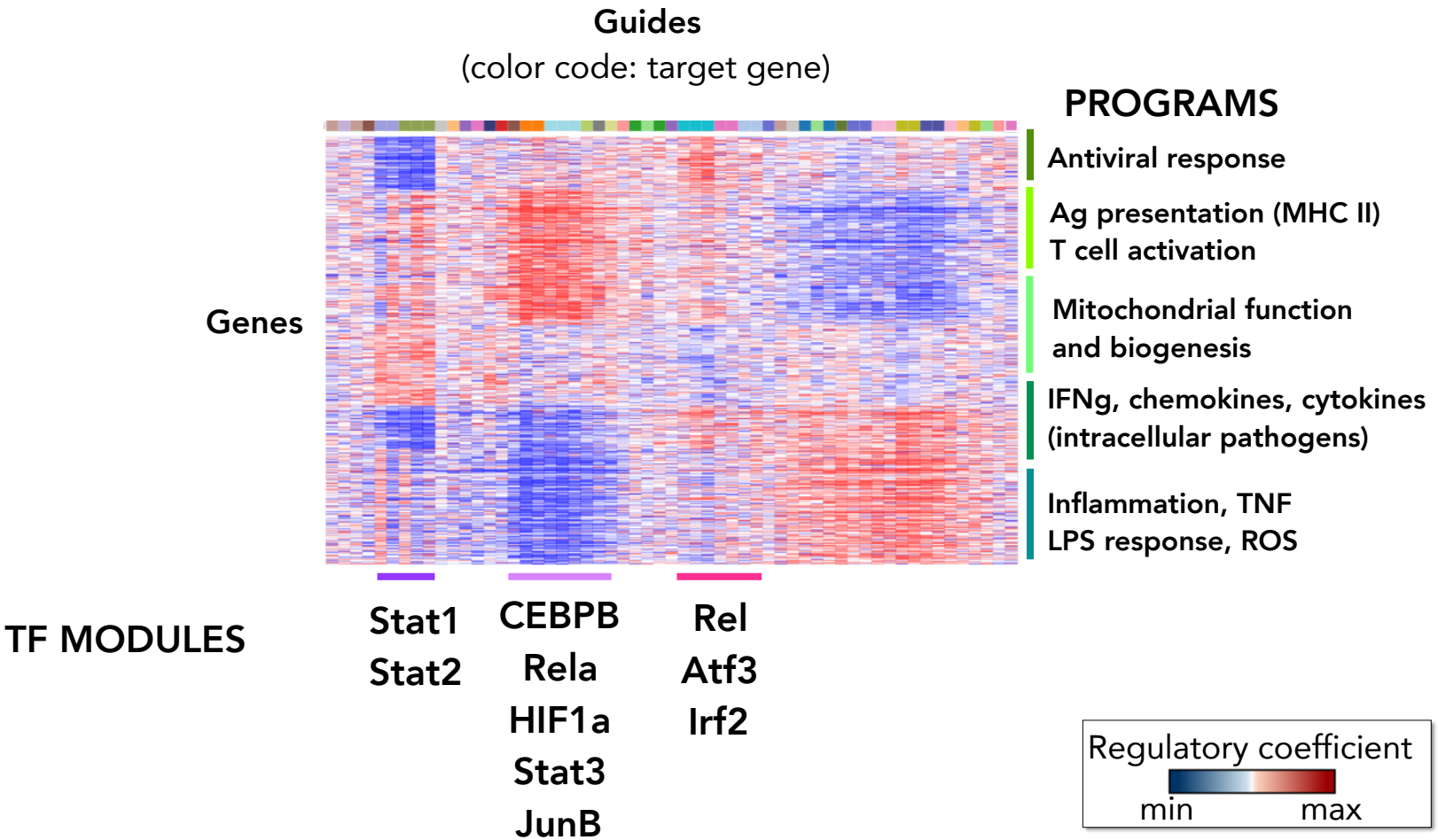
Regulatory response of Dendritic Cells to LPS stimulation



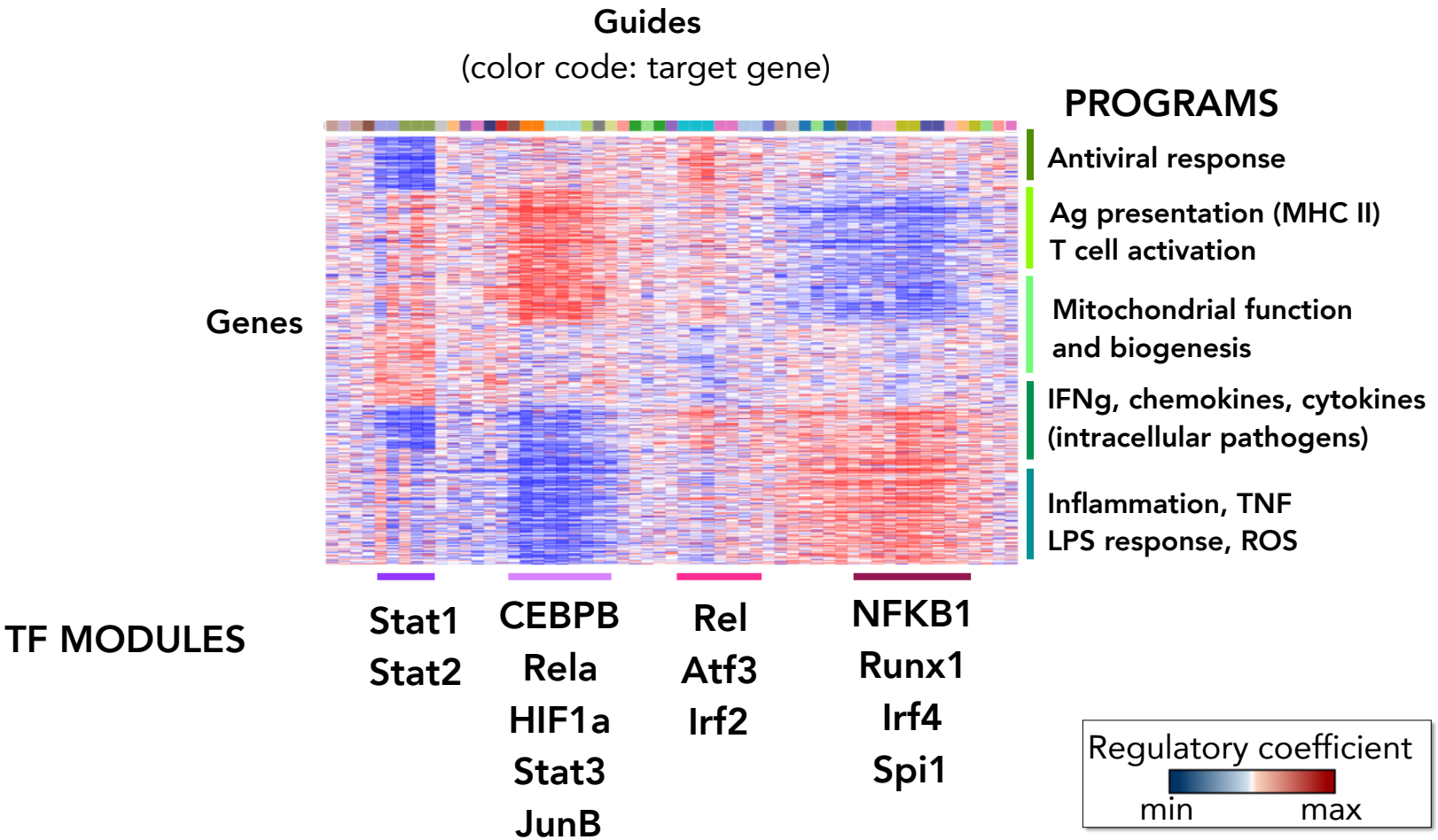
Regulatory response of Dendritic Cells to LPS stimulation



Regulatory response of Dendritic Cells to LPS stimulation



Regulatory response of Dendritic Cells to LPS stimulation



Acknowledgments

Klarman cell observatory

Aviv Regev
Orit Rozenblatt-Rosen
Regevlab members

Collaborators

Evan Macosko
Sylvain Lapan
Irene Whitney
Yirong Peng
Steve McCarroll
Connie Cepko
Joshua Sanes

Seurat

Rahul Satija
Jeff Farrell
Alex Schier

Acknowledgments

Klarman cell observatory

Aviv Regev

Orit Rozenblatt-Rosen

Regevlab members

Colla

Evan

Sylva

Irene

Yiron

Steve McCarroll

Connie Cepko

Joshua Sanes

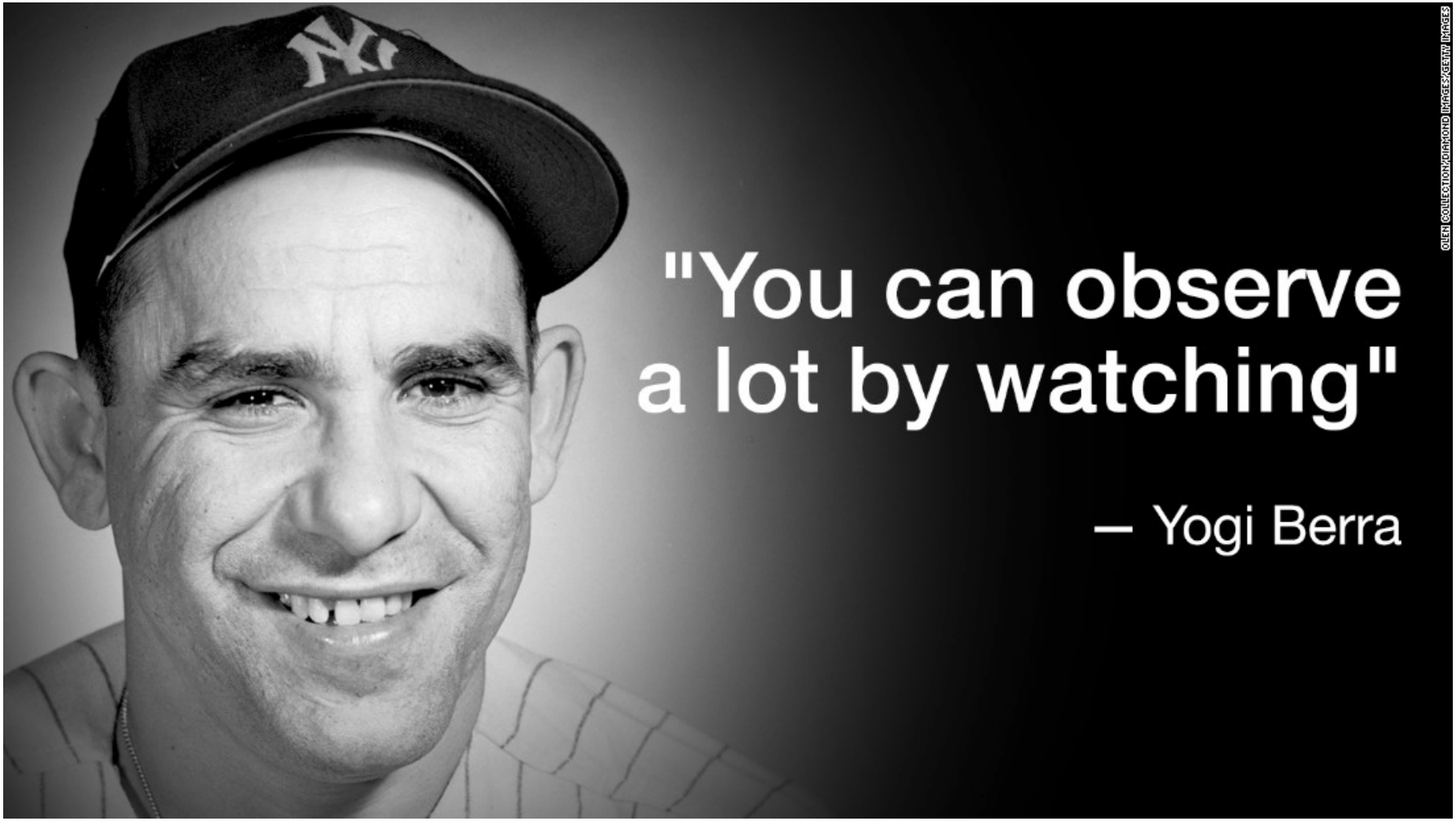
Scott, Brian, Dag et al. for putting this great meeting together, and for having me over !

Seurat

Rahul Satija

Jeff Farrell

Alex Schier


"You can observe
a lot by watching"
— Yogi Berra
GLEN COULLESTON/DIAMOND IMAGES/GETTY IMAGES