Applying Single Cell genomics to your research:
Discussion of experimental and computational frameworks

Alexandra-Chloé Villani, PhD

Society of Immunotherapy of Cancer Meeting

Workshop on Single Cell Techniques in Immunology and Cancer Immunotherapy

November 9th 2017

BROAD INSTITUTE

- **No disclosure**

Outline

1. Introduction
   A. Relevance & advances in single cell sequencing
   B. Overview single cell assay
2. Cell isolation for single cell readout
3. scRNAseq protocols
4. Other single cell readouts & multi-omics
5. Analysis overview
6. Technical challenges
7. Experimental design & common questions
8. Application & validation

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[3], Schraga Schwartz[3], Brian Fowler[4], Suzanne Weaver[4], Jing Wang[4], Xiaohui Wang[4], Ruihua Ding[1,2], Raktima Raychowdhury[3], Nir Friedman[5], Nir Hacohen[3,6], Hongkun Park[1,2,3], Andrew P. May[4] & Aviv Regev[3,7]

Science

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin,[1]* Ephraim Kenigsberg,[2,3]* Hadas Keren-Shaul,[1]* Naama Elefant,[1] Franziska Paul,[1] Irina Zaretsky,[1] Alexander Mildner,[1] Nadav Cohen,[2,3] Steffen Jung,[1] Amos Tanay,[2,3]†‡ Ido Amit[1]†‡

Distinct myeloid progenitor–differentiation pathways identified through single-cell RNA sequencing

Roy Drissen[1,2], Natalija Buza-Vidas[2], Petter Woll[1,3], Supat Thongjuea[1], Adriana Gambardella[1,2], Alice Giustacchini[1,3], Elena Mancini[4], Alya Zriwil[5], Michael Lutteropp[1,3], Amit Grover[1,2,4], Adam Mead[1,3], Ewa Sitnicka[5], Sten Eirik W Jacobsen[1,3,6] & Claus Nerlov[1,2,4,6]

Cell

An Immune Atlas of Clear Cell Renal Cell Carcinoma

Stéphane Chevrier,[1,15] Jacob Harrison Levine,[2,15] Vito Riccardo Tomaso Zanotelli,[1,3] Karina Silina,[4] Daniel Schulz,[1] Marina Bacac,[5] Carola Hermine Ries,[6] Laurie Ailles,[7,8] Michael Alexander Spencer Jewett,[8] Holger Moch,[9] Maries van den Broek,[4] Christian Beisel,[10] Michael Beda Stadler,[11,12] Craig Gedye,[13] Bernhard Reis,[14] Dana Pe'er,[2] and Bernd Bodenmiller[1,16,*]

Cell

Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses

Yonit Lavin,[1,2,3] Soma Kobayashi,[1,2,3,14] Andrew Leader,[1,2,3,14] El-ad David Amir,[2,3,9] Naama Elefant,[10] Camille Bigenwald,[1,2,3] Romain Remark,[1,2,3,13] Robert Sweeney,[6,7] Christian D. Becker,[4] Jacob H. Levine,[11] Klaus Meinhof,[4] Andrew Chow,[1,2,3] Seunghee Kim-Shulze,[2,3,9] Andrea Wolf,[6] Chiara Medaglia,[10] Hanjie Li,[10] Julie A. Rytlewski,[12] Ryan O. Emerson,[12] Alexander Solovyov,[1,3,5,8] Benjamin D. Greenbaum,[1,3,5,8] Catherine Sanders,[12] Marissa Vignali,[12] Mary Beth Beasley,[8] Raja Flores,[6] Sacha Gnjatic,[2,3,5,9] Dana Pe'er,[11] Adeeb Rahman,[2,3,7,9] Ido Amit,[10] and Miriam Merad[1,2,3,9,15,*]

LETTER

doi:10.1038/nature20105

Single-cell RNA-seq identifies a PD-1^{hi} ILC progenitor and defines its development pathway

Yong Yu[1]*, Jason C. H. Tsang[1,2,3]*, Cui Wang[1,4]*, Simon Clare[1], Juexuan Wang[1], Xi Chen[1], Cordelia Brandt[1], Leanne Kane[1], Lia S. Campos[1], Liming Lu[5], Gabrielle T. Belz[6,7], Andrew N. J. McKenzie[8], Sarah A. Teichmann[1,9], Gordon Dougan[1,10] & Pentao Liu[1]

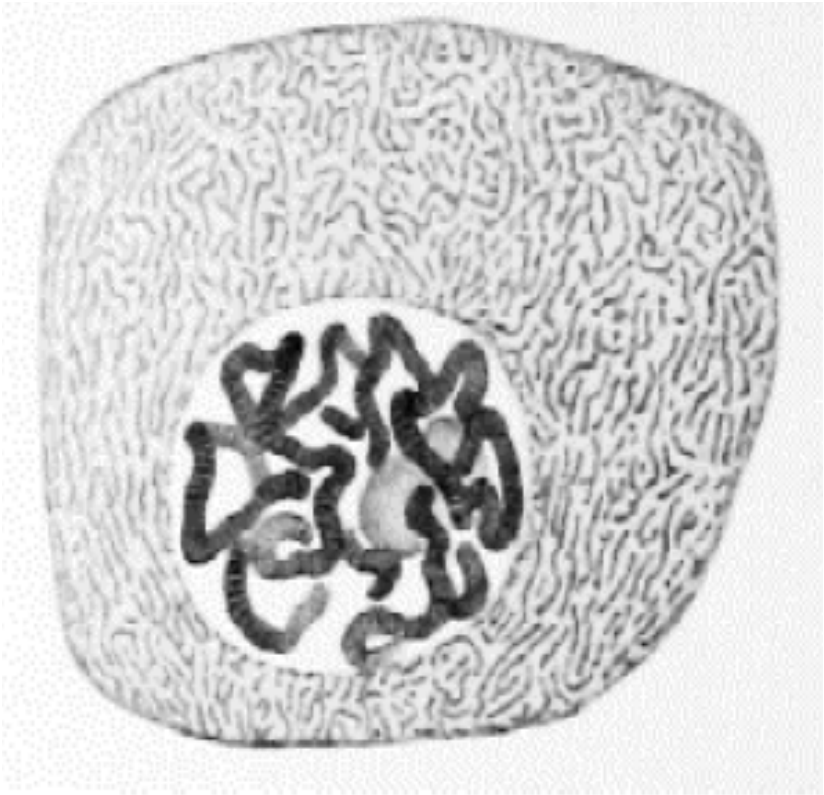
RESEARCH ARTICLES

CANCER GENOMICS

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

Itay Tirosh,[1]* Benjamin Izar,[1,2,3]*†‡ Sanjay M. Prakadan,[1,4,5,6] Marc H. Wadsworth II,[1,4,5,6] Daniel Treacy,[1] John J. Trombetta,[1] Asaf Rotem,[1,2,3] Christopher Rodman,[1] Christine Lian,[7] George Murphy,[7] Mohammad Fallahi-Sichani,[8] Ken Dutton-Regester,[1,2,9] Jia-Ren Lin,[10] Ofir Cohen,[1] Parin Shah,[2] Diana Lu,[1] Alex S. Genshaft,[1,4,5,6] Travis K. Hughes,[1,4,6,11] Carly G. K. Ziegler,[1,4,6,11] Samuel W. Kazer,[1,4,5,6] Aleth Gaillard,[1,4,5,6] Kellie E. Kolb,[1,4,5,6] Alexandra-Chloé Villani,[1] Cory M. Johannessen,[1] Aleksandr Y. Andreev,[1] Eliezer M. Van Allen,[1,2,3] Monica Bertagnolli,[12,13] Peter K. Sorger,[8,10,14] Ryan J. Sullivan,[15] Keith T. Flaherty,[15] Dennie T. Frederick,[15] Judit Jané-Valbuena,[1] Charles H. Yoon,[12,13]† Orit Rozenblatt-Rosen,[1]† Alex K. Shalek,[1,4,5,6,11,16]† Aviv Regev,[1,17,18]†‡ Levi A. Garraway[1,2,3,14]†‡

Analysis at single cell level is an old concept!

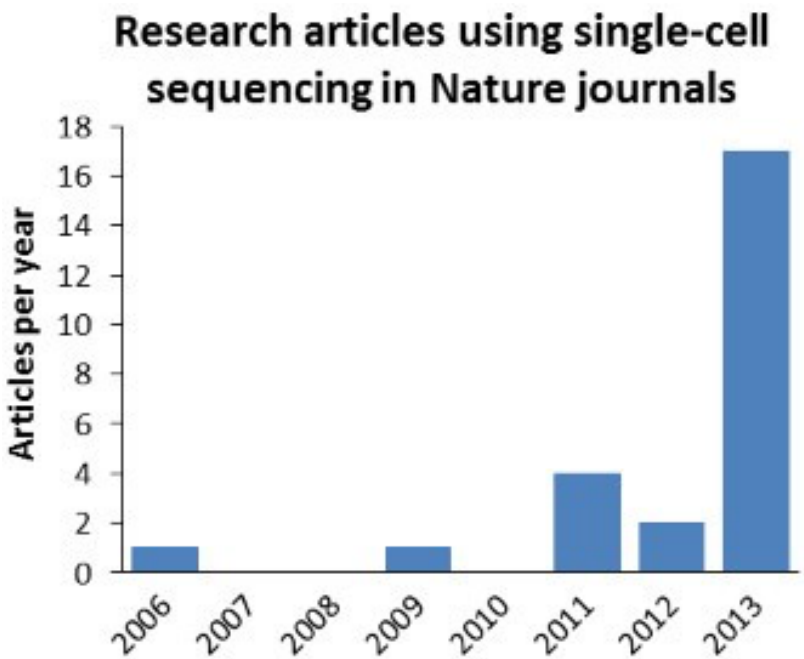


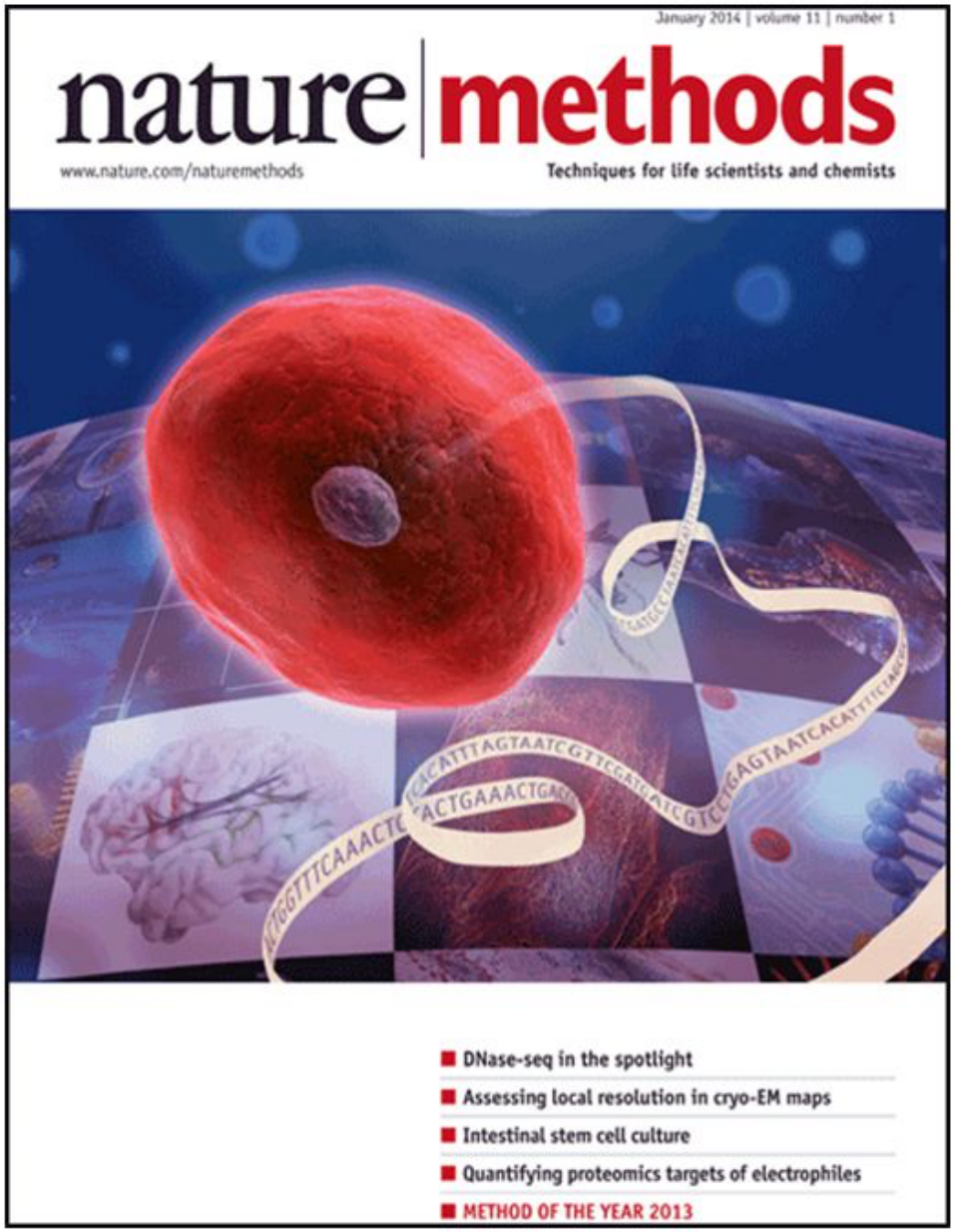
A single-cell genome image of polytene chromosomes from insects from 1882 monograph by Flemming

Blainey et al, 2014

Methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine.

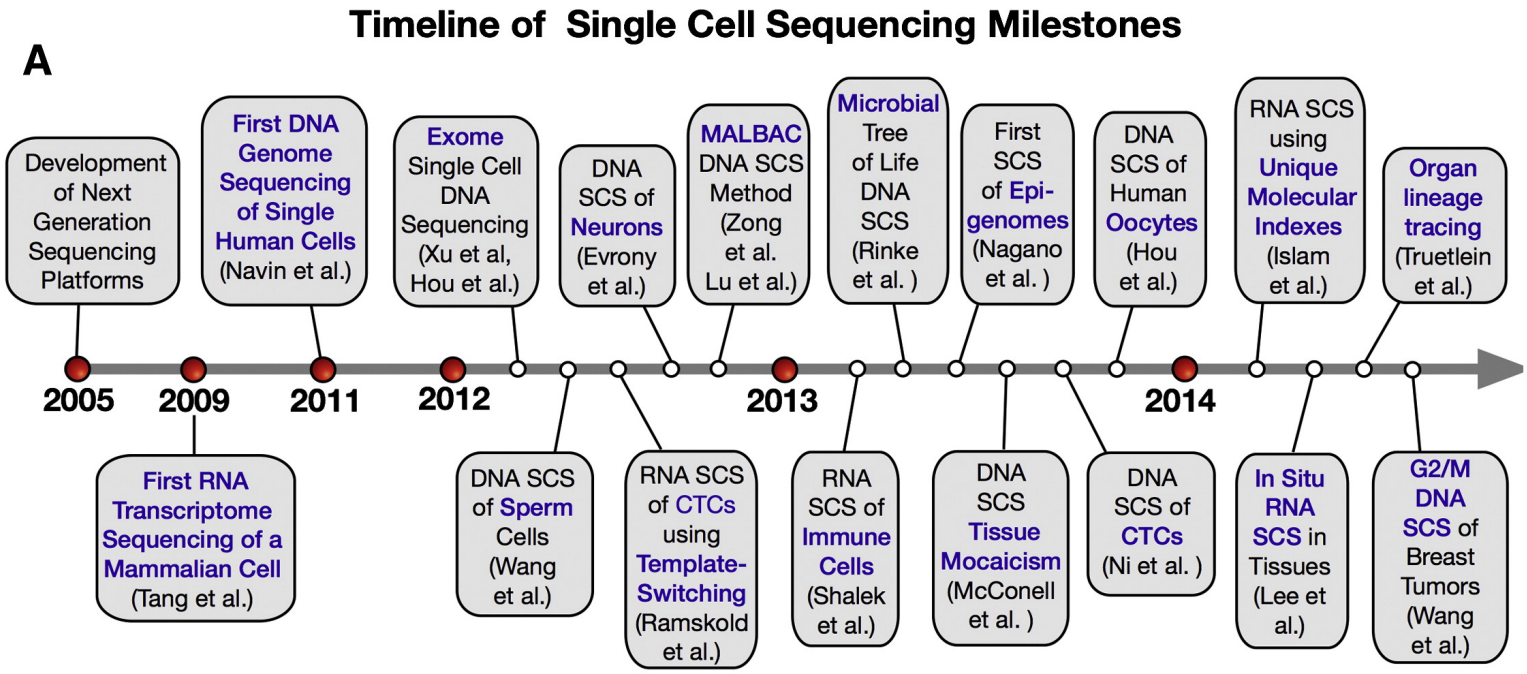
--- Nature Methods

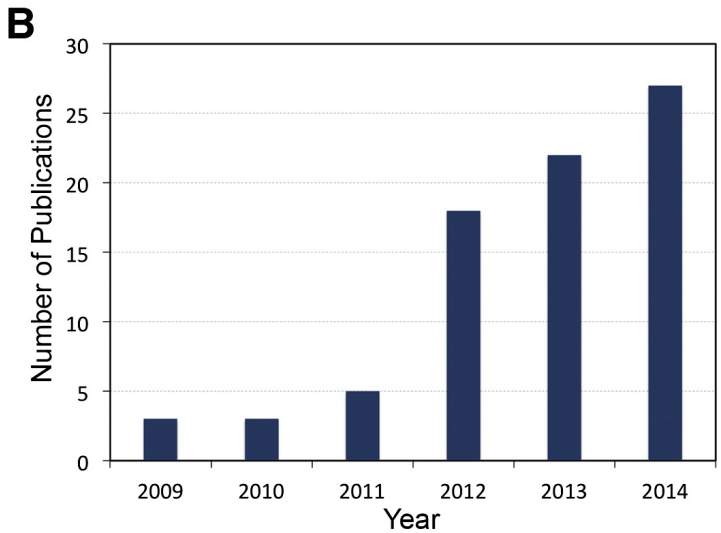


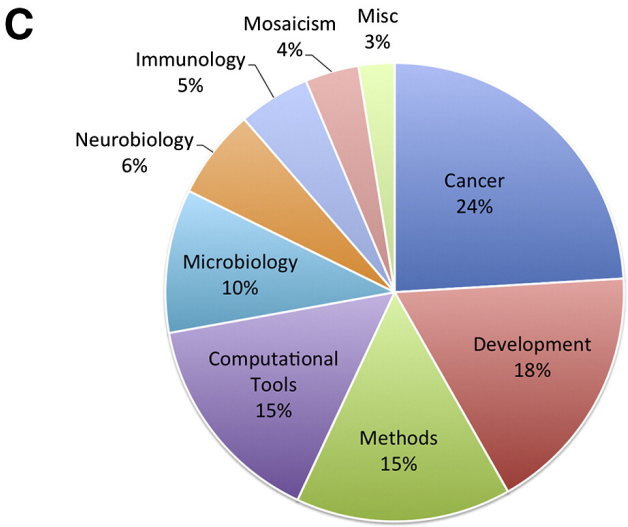


Significant increase in publications and data in the last 2 years

Advances & Application of single cell sequencing

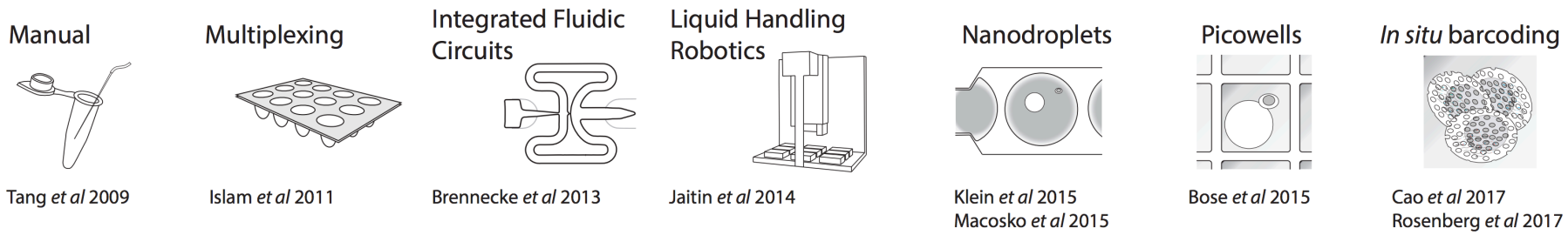


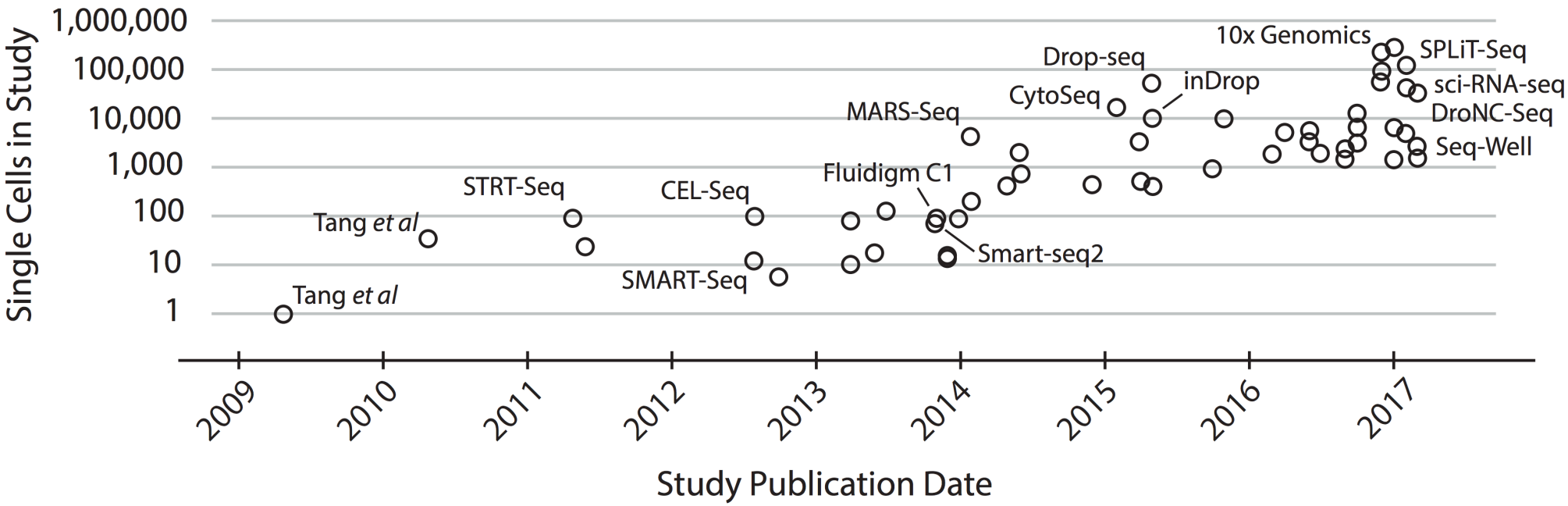




Wang et al. 2015

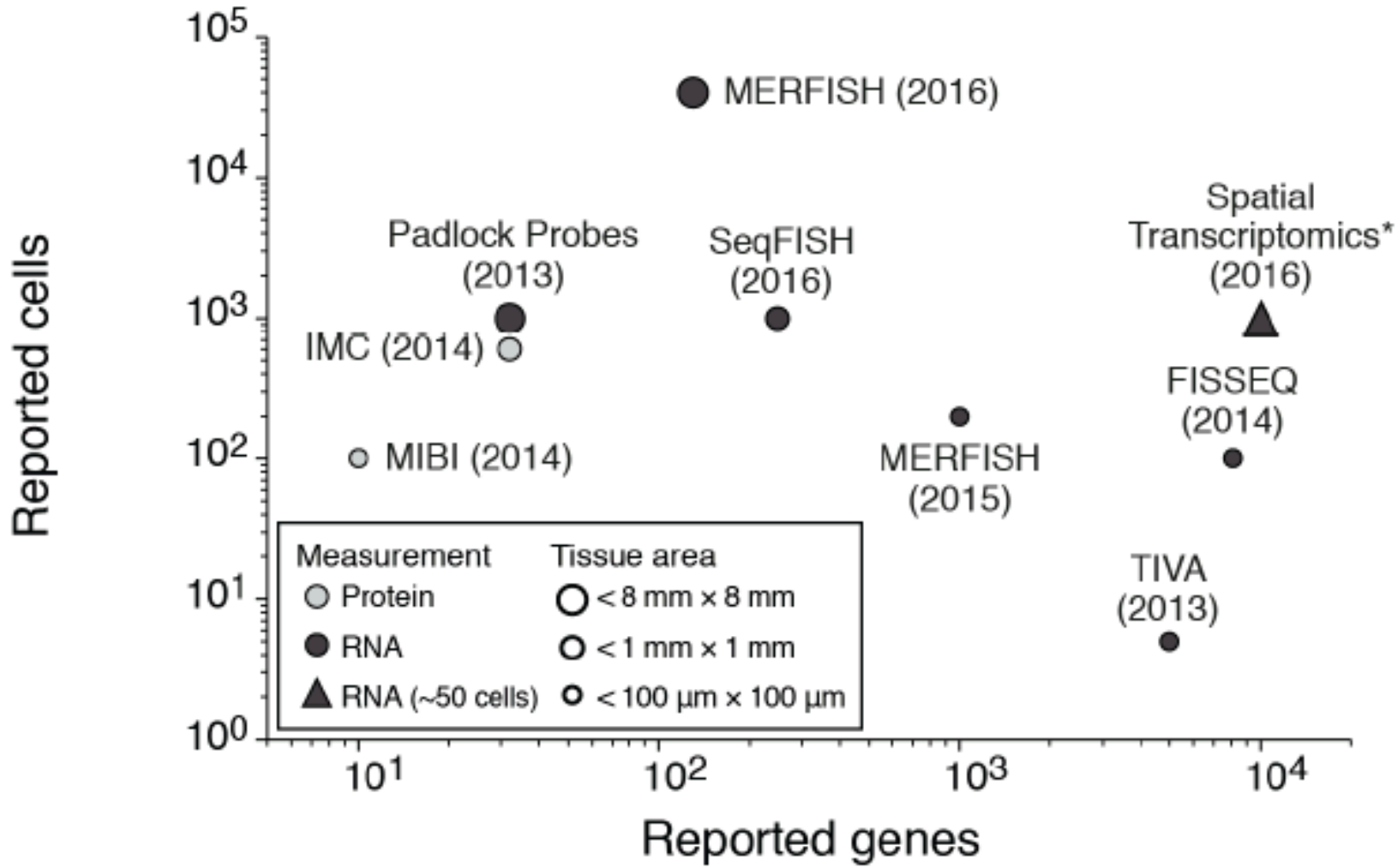
Technological advances are empowering scalability & additional dimensionalities



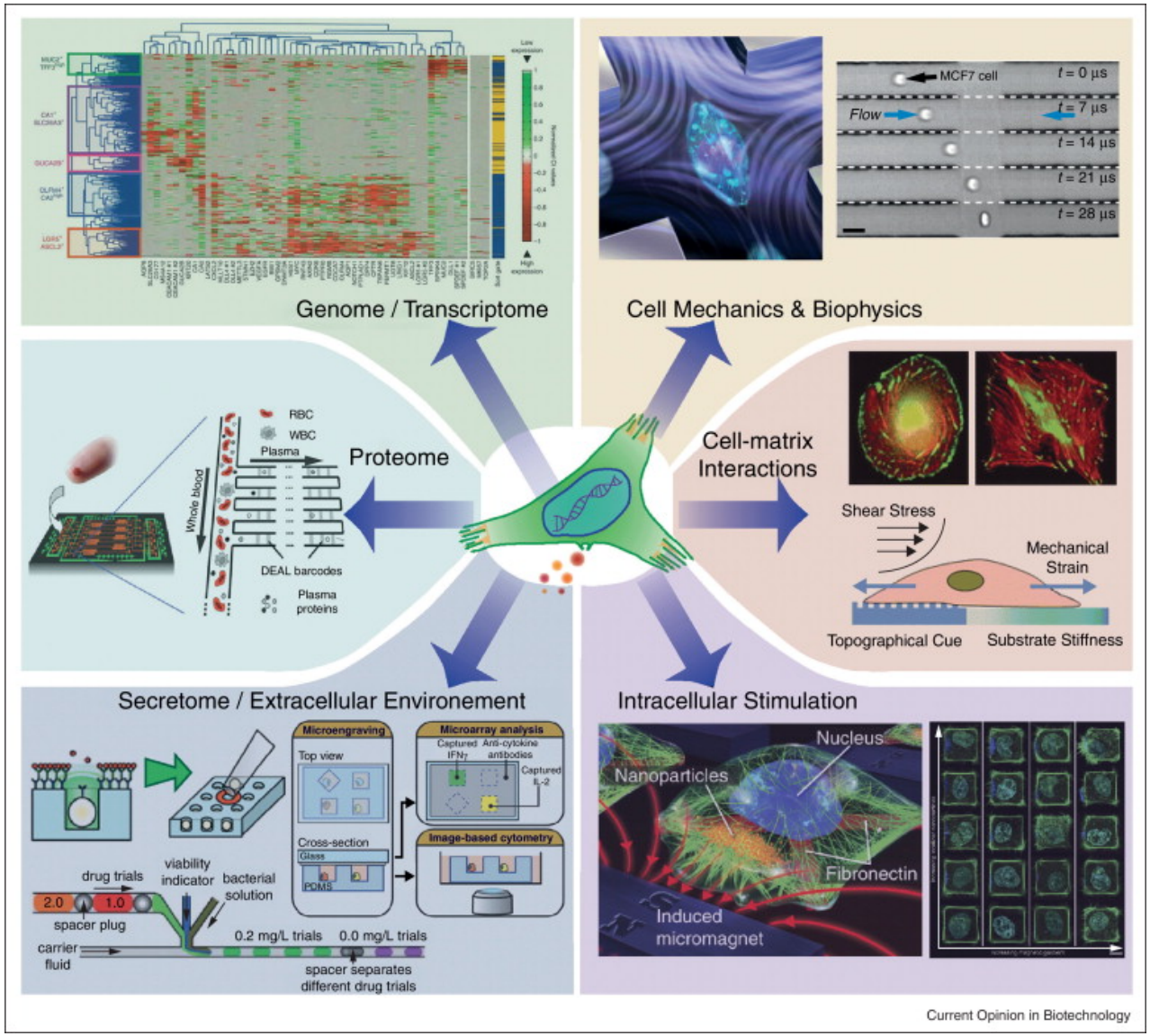


HCA White Paper. 2017

Technological advances are empowering scalability & additional dimensionalities



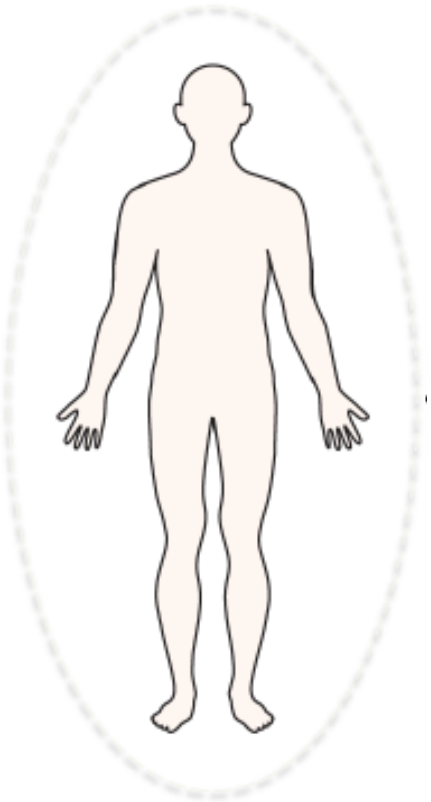
HCA White Paper. 2017



"Single-cell approaches stand poised to revolutionize our capacity to understand the scale of genomic, epigenomic, and transcriptomic diversity that occurs during the lifetime of an individual organism."
Machaulay & Voet 2014

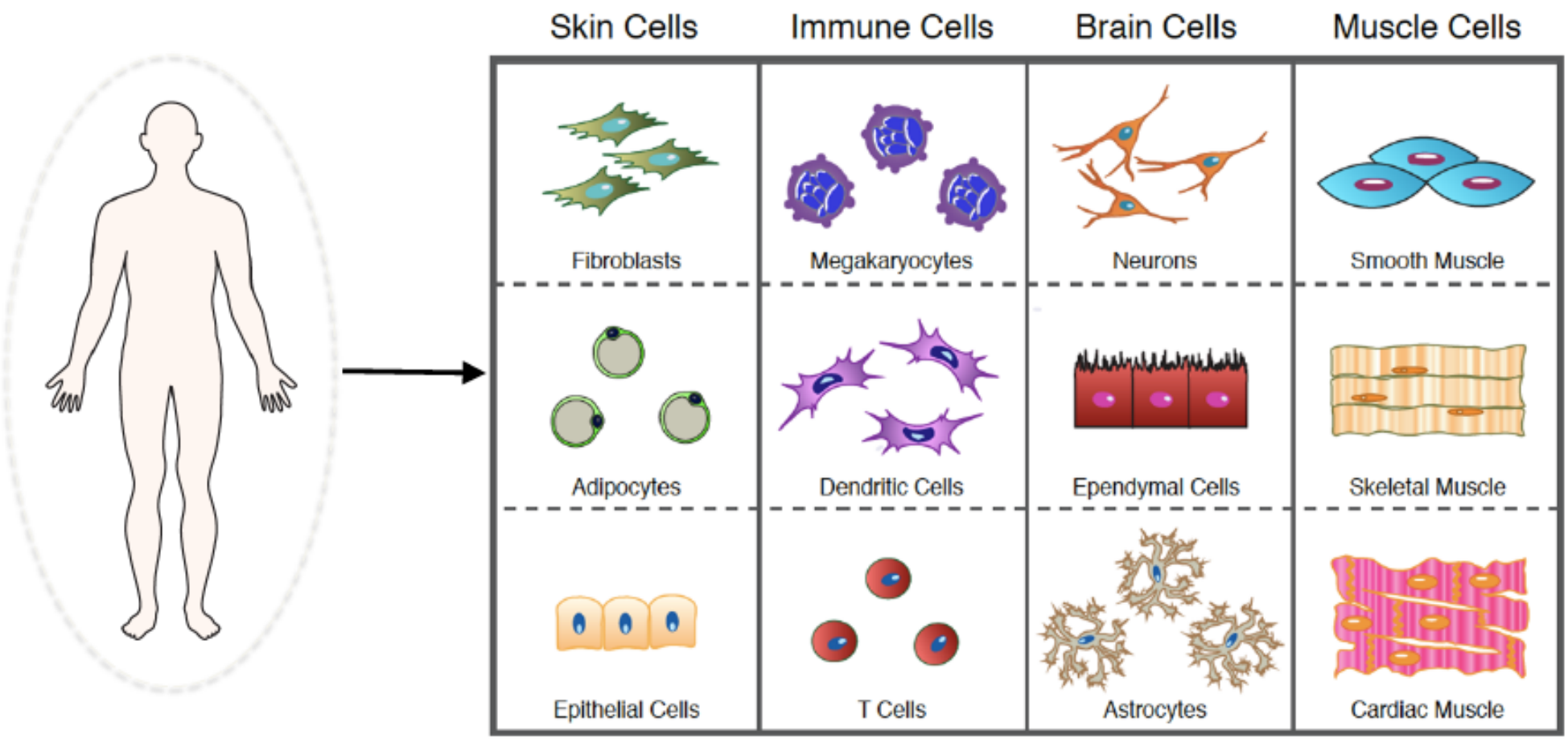
Weaver, 2014

Do we really know cells defining the human system?



- ~ 30 trillion cells
- Text book → ~ 300 'major' cell types?
- Science → ~ 100 subtypes of immune cells!

Cells define our core constituents



How do we define and classify cell type?

How do we define and classify cell types?

molecular markers
morphology
spatial localization
physical properties
functions
developmental origins
transcription factor dependency
growth factor dependency
chromatin states
biochemical states
…

Limitations of current cell type/state definitions

- **Purity**: Defined cell types may not be pure using the historically defined markers

- **Species**: The more well-defined mouse cell types may not directly translate to human

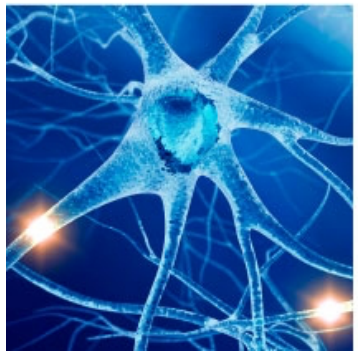
- **Variations**: An immune response induces new and unexpected states

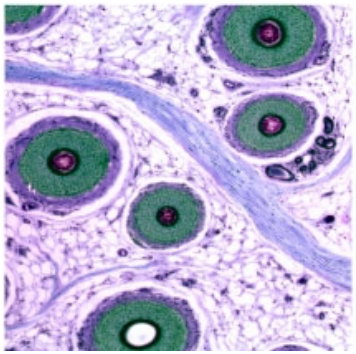
→ Do existing 'standard' set of surface markers truly define distinct immune cell types?

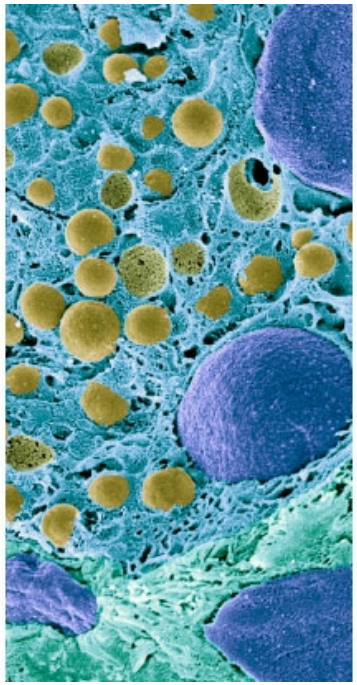
→ Are there more cell subsets that are not currently appreciated?

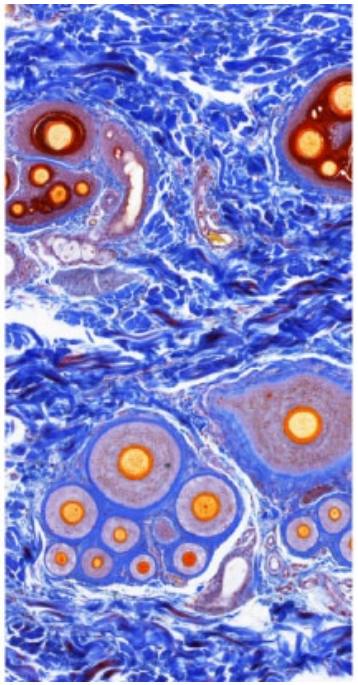
Solution: Leveraging the power of single cell profiling to generate map *de novo* & integrate legacy knowledge

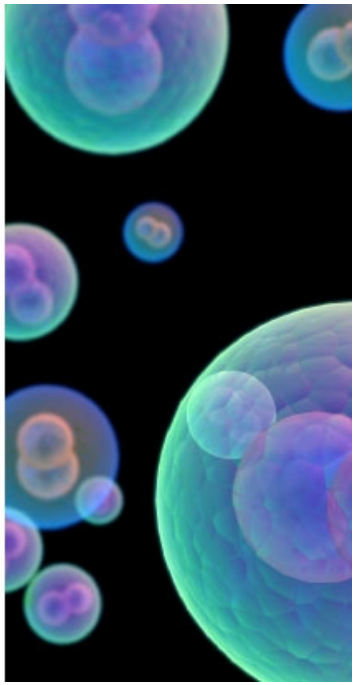
HUMAN
CELL
ATLAS







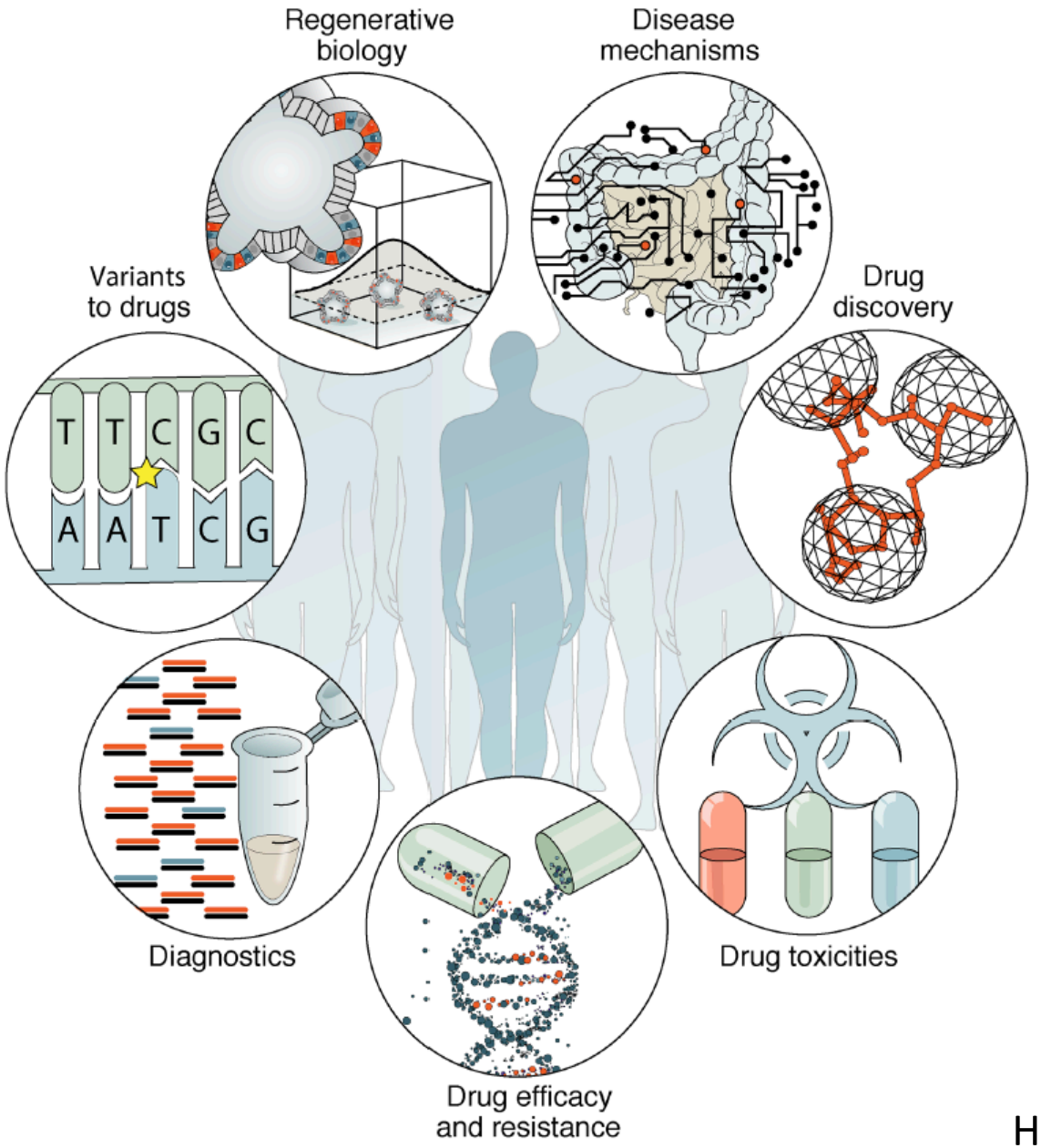




Mission: To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease

https://www.humancellatlas.org

Redefining the human system at single cell resolution has tremendous potential for biology & medicine



HCA White Paper. 2017

What can we learn from single cell

- **Taxonomy & Census** → data-driven molecular definition of cell types & dissection of tissue heterogeneity
- **Anatomy & Physiology** → spatial structure of tissue
- **Pathology** → defining disease cells and associated ecosystem
- **Physiology** → dissection of temporal changes, responses to challenges (e.g. drug treatment)
- **Developmental biology** → cell fate / lineage mapping
- **Molecular mechanisms** → cellular circuitry

First critical step → cell isolation

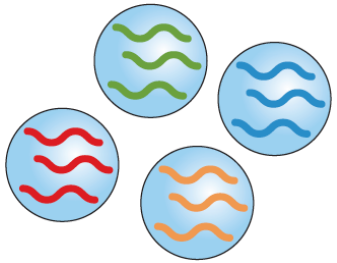
Cell Isolation

Amplification method

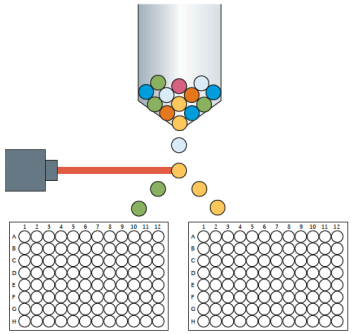
Sequencing method

Applications

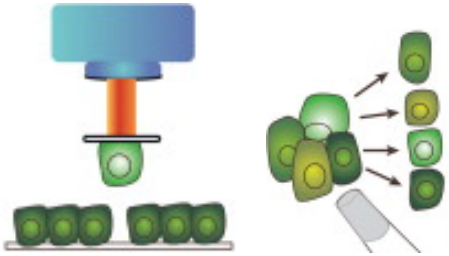
Trapping cells in droplets



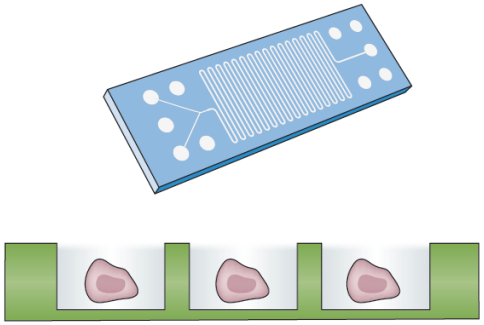
FACS / cell sorter



Microdissection & micromanipulation



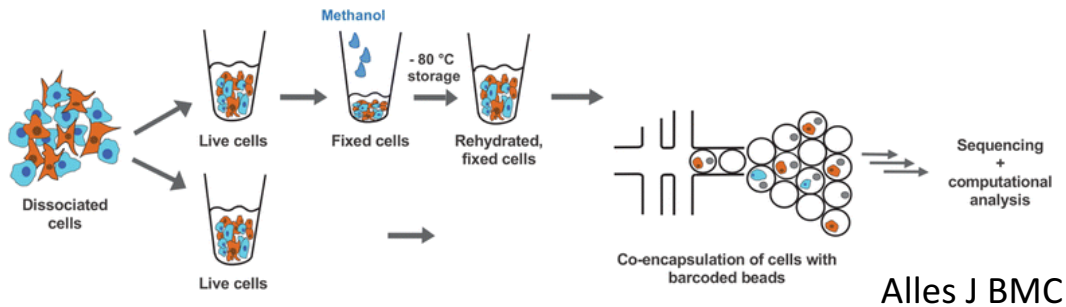
Microfluidics & microwells



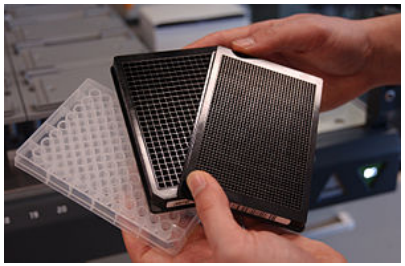
Adapted from Papalexi E et al. Nat Rev Immunol 2017

Common considerations for sample collection & dissociation

- **Fresh vs. Frozen** → cells vs. nuclei (e.g. considering multi-sites study?)
- **Cell dissociation optimization**
  - Minimizing leakage and RNA degradation
  - Need to optimize for every tissue → e.g readouts: FACS & bulk sequencing
  - Challenging dissociation? Consider LCM & nuclei sequencing
- **Enrichment strategy**
  - Even the sampling to enrich for rare cells (e.g. profiling human blood)
  - Separate immune from non-immune cells (sorting or bead/column)
  - Profiling uniquely T and B cells for TCR & BCR
- **Cell death & RBC removal**
  - Live/death & CD235a marker-based depletion by FACS
  - Magnetic bead depletion-based
  - Column-based (e.g. MACS) depletion→ some cell types get caught in columns
- **Work to limit RNA degradation** (fixation protocol work in some case)

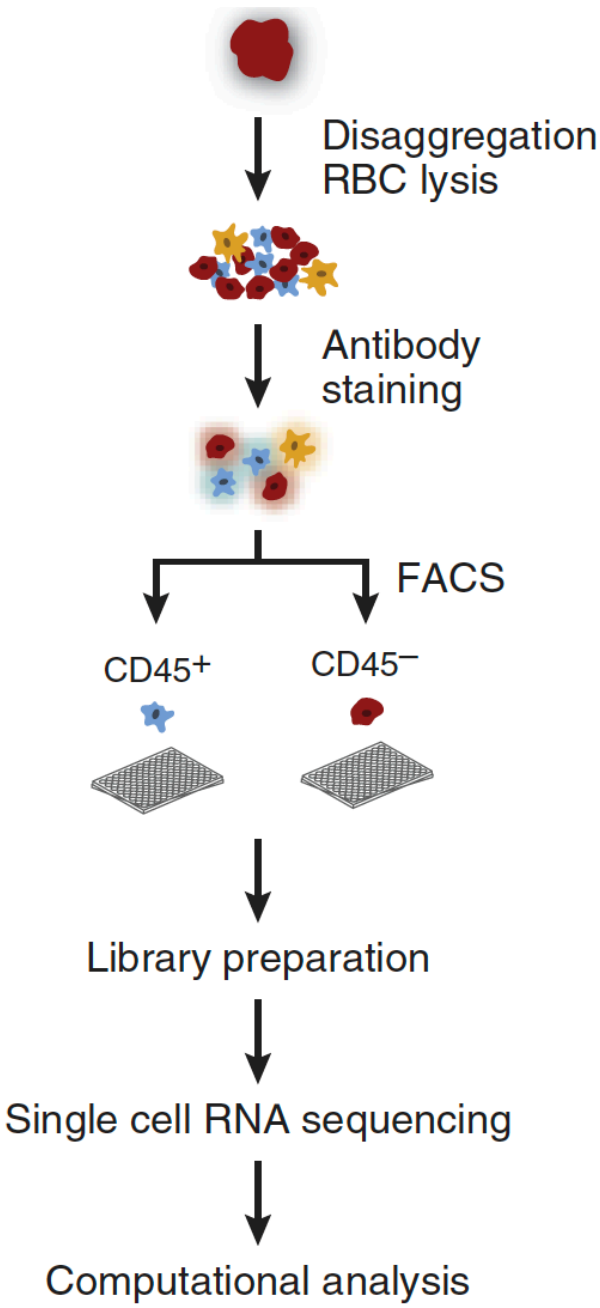


Alles J BMC Biol 2017



FACS isolation

- **Advantages:**
  - Sorting based on specific cell phenotype
  - Archiving potential
  - Full-length cDNA readout possible
- **Disadvantages:**
  - Larger amount of cell required
  - Occasional isolation of more than one cells
  - Putative damage of cells (epithelial cells)
  - Labor intensive & more costly
- **Know your cells, are they sticky, are they big?**
  - Select an appropriate sized nozzle
- **Don't sort too quickly (1-2k cells per second or lower)**
  - The slower the more time cells sit in lysis after sorting
  - 10 minutes max in lysis (some say 30 minutes)
- **Calibrate speed of instrument with beads**
  - Check alignment every 5-6 plates
- **Afterwards spin down to make sure cells are in lysis buffer**
  - Flash freeze on dry ice and move to -80C (use very adherent seals for archiving)

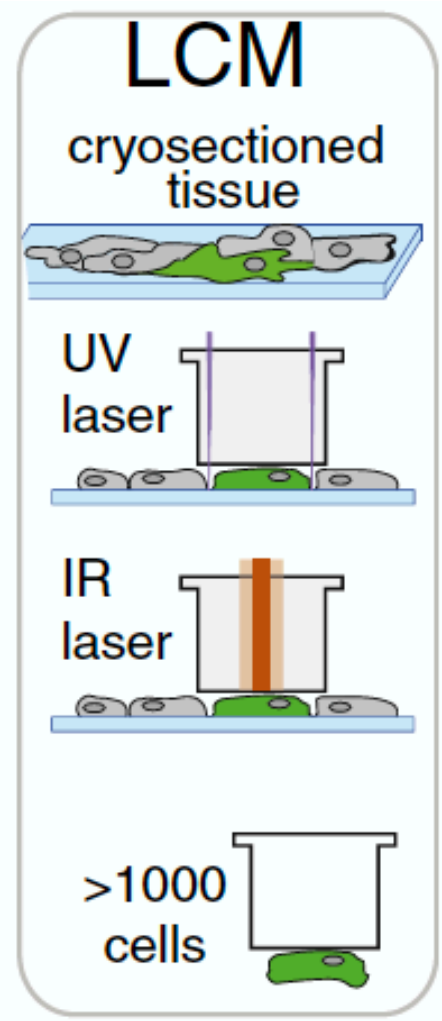


Tirosh et al. Science 2016

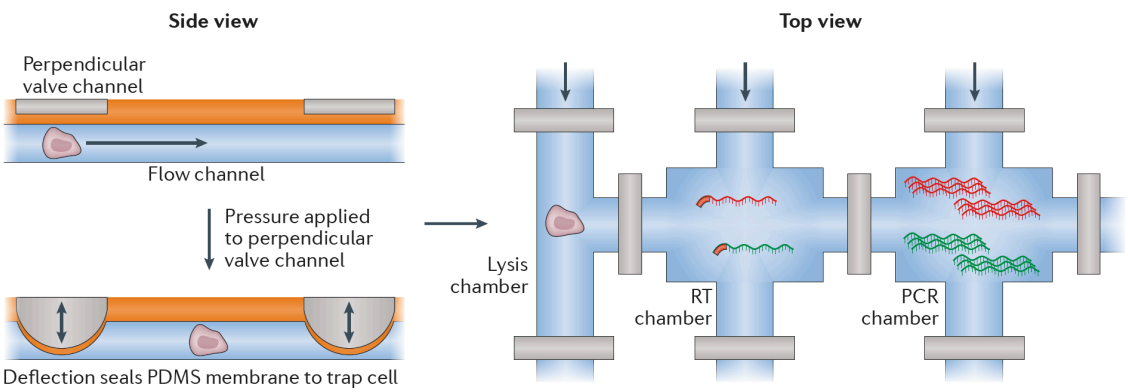
Micromanipulation & LCM

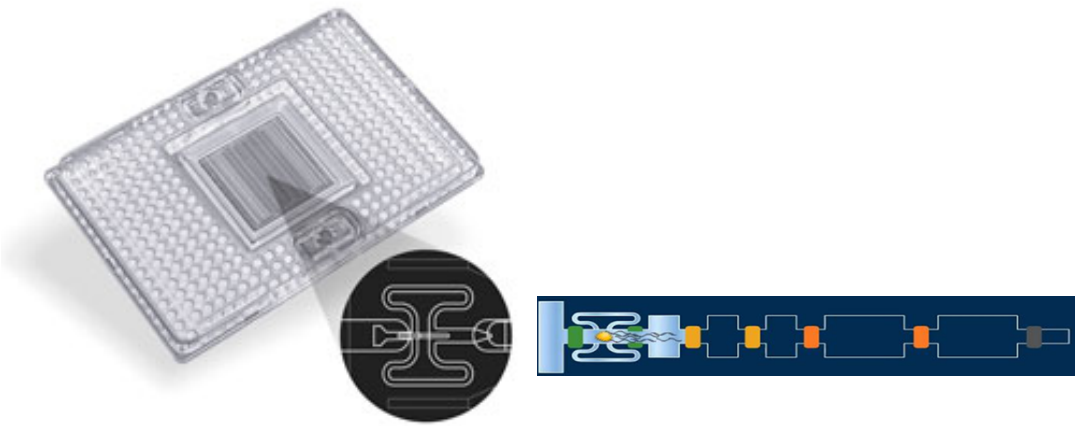
- **Advantages:**
  - Visual confirmation
  - Applicable when only few cells are available
  - Retain topological information of the cell
  - Permits isolation of a cell from fixed tissue or cryosection

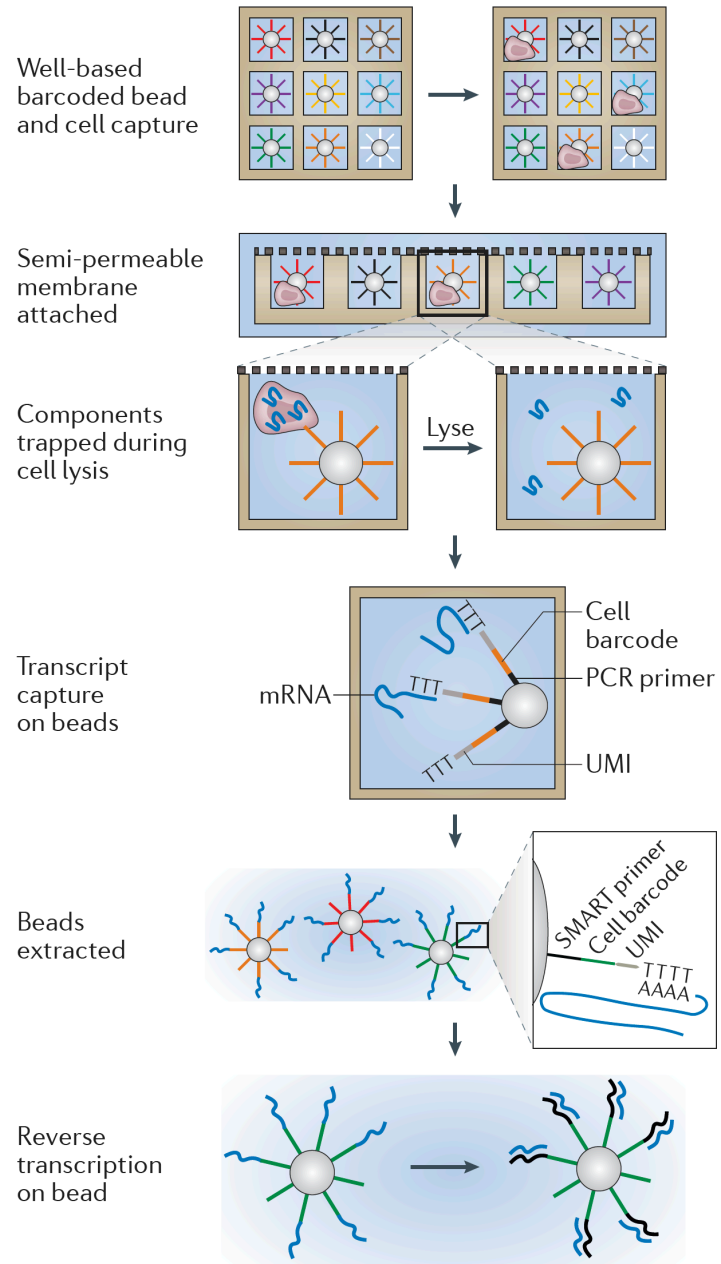
- **Disadvantages:**
  - Low throughput
  - Lengthy process → RNA degradation
  - Operator bias
  - Contamination of other cells
  - Potential loss of cellular material (LCM)



Microfluidics & Micro-wells

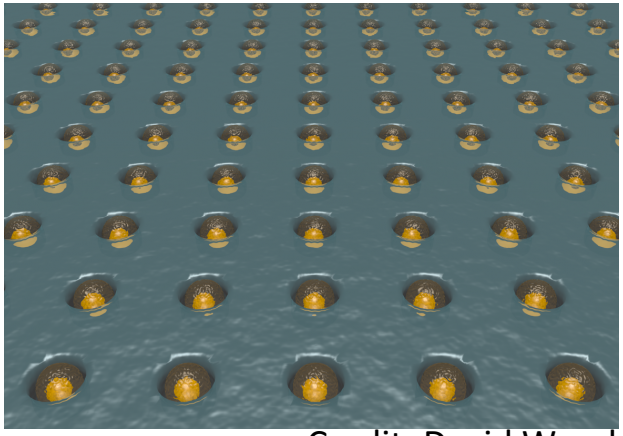






Prakadan et al. Nat Rev Genet 2017

Microfluidics & Micro-wells



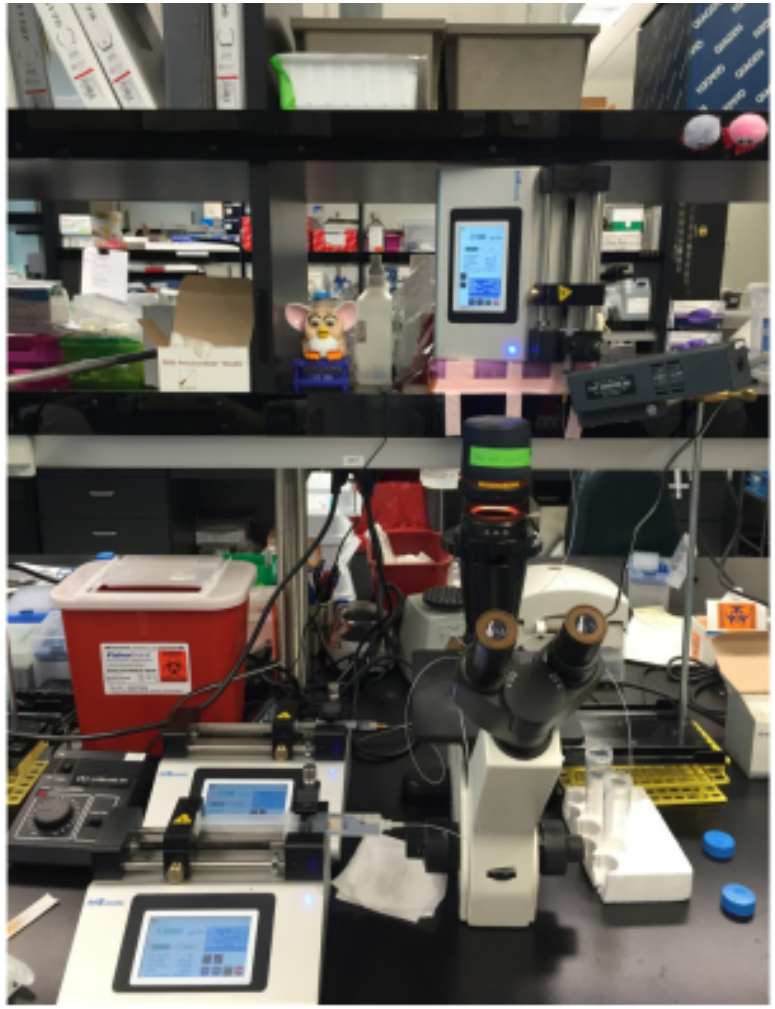
Credit: David Wood

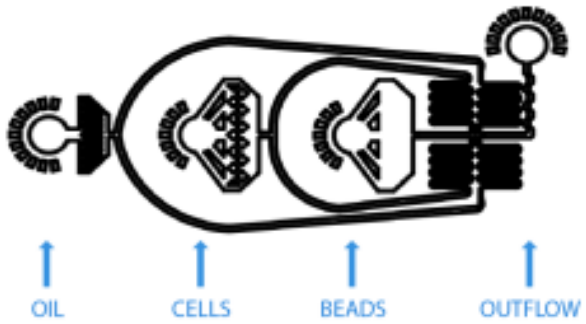
- **Advantages:**
  - Highly standardized nanoliter reaction (lower reagent cost/cell)
  - Less operator bias in cell isolation and enzymatic reactions
  - Automated higher throughput cell isolation with visual confirmation

- **Disadvantages:**
  - Putative loss of cells → capture efficiency lower than if sorting in plates in some cases
  - Cannot select specific cells (unlike cell sorting)
  - Bias driven by cell size and adherence (fixed size devices)
  - Bias driven by cell type frequency (will capture mostly abundant types)
  - In some case still need to enrich first and cells sit around longer before lysis

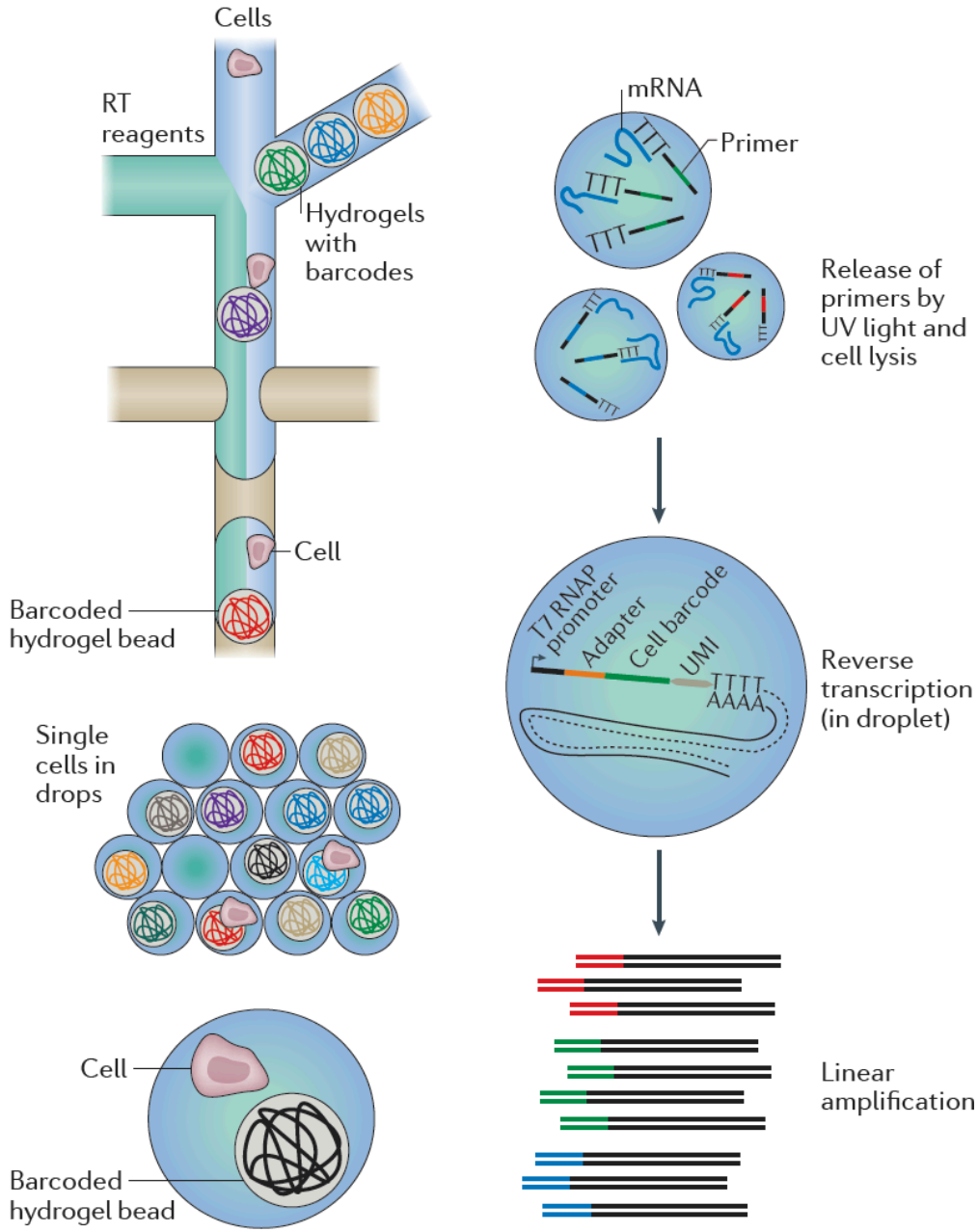
Emulsion-based / Droplets

DropSeq setup



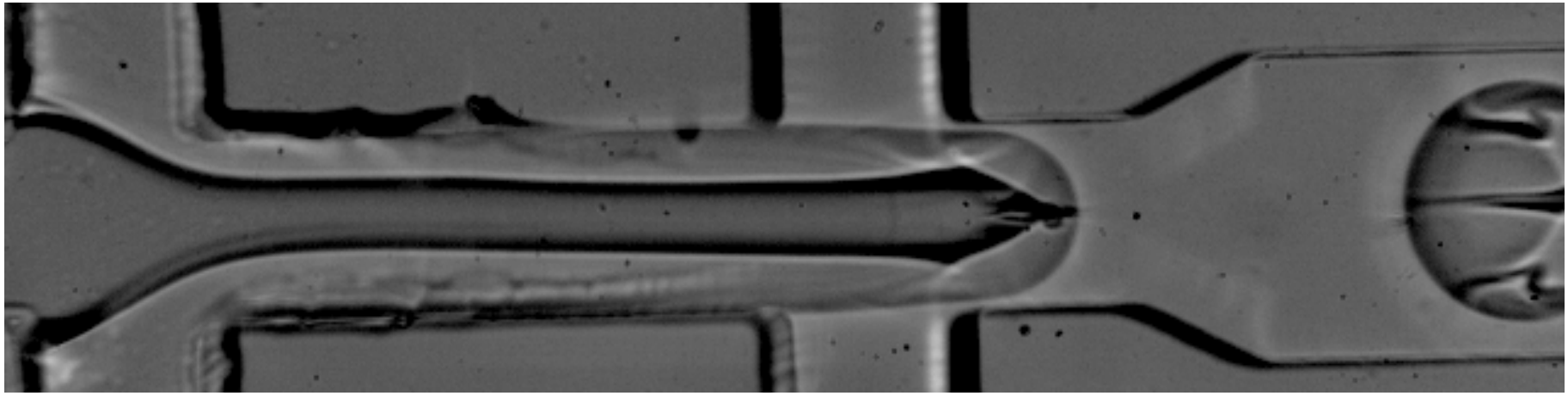


Emulsion-based / Droplets



Prakadan et al. Nat Rev Genet 2017

Drop-seq – Overview



Emulsion-based / Droplets

- **Advantages:**
  - Very scalable → thousands of cells per experiment
  - Smaller volumes → higher detection & better reproducibility
  - Smaller volumes → cheaper reagent cost
  - Sequencing cost become bottleneck → often shallow sequencing

- **Disadvantages:**
  - High cell input required (DropSeq) though low cell capture
  - Variable quality of beads → can increase cost
  - Need to be familiar with microfluidics (unless opt for commercial option like 10X)
  - Droplet-based assays can have leaky RNA (unlike plate → compartmentalization)
  - Capture less transcripts than plate-based (lower resolution)
  - Only 3' end readout

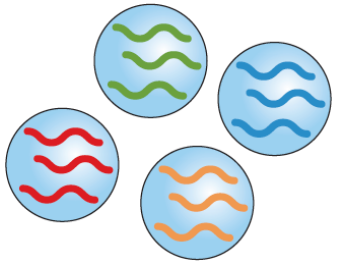
- **Some pointers:**
  - Before library generation wash off any medium (inhibits library generation)
  - Adding PBS & BSA (0.05-0.01%) can help protect the cells
  - Filter all reagents with micron strainer before loading on microfluidic
  - Some purchased devices come with hydrophobic coating
    → Can deteriorate (2 months at best) → recoating works

Credit: Monika Kowalczyk & Tim Tickle

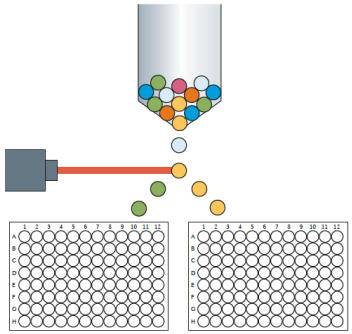
Selecting scRNAseq protocol

Cell Isolation

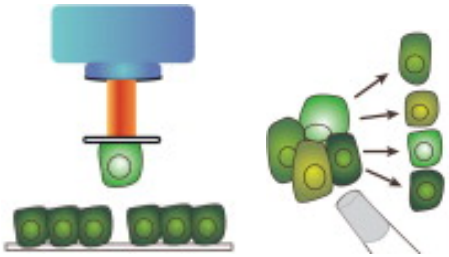
Trapping cells in droplets



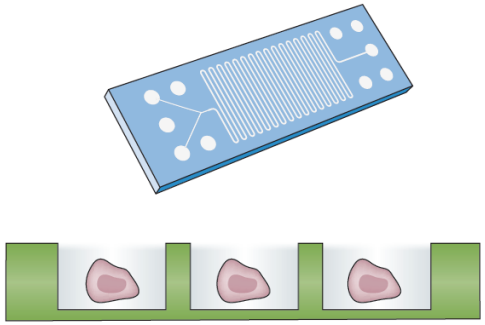
FACS / cell sorter



Microdissection & micromanipulation



Microfluidics & microwells

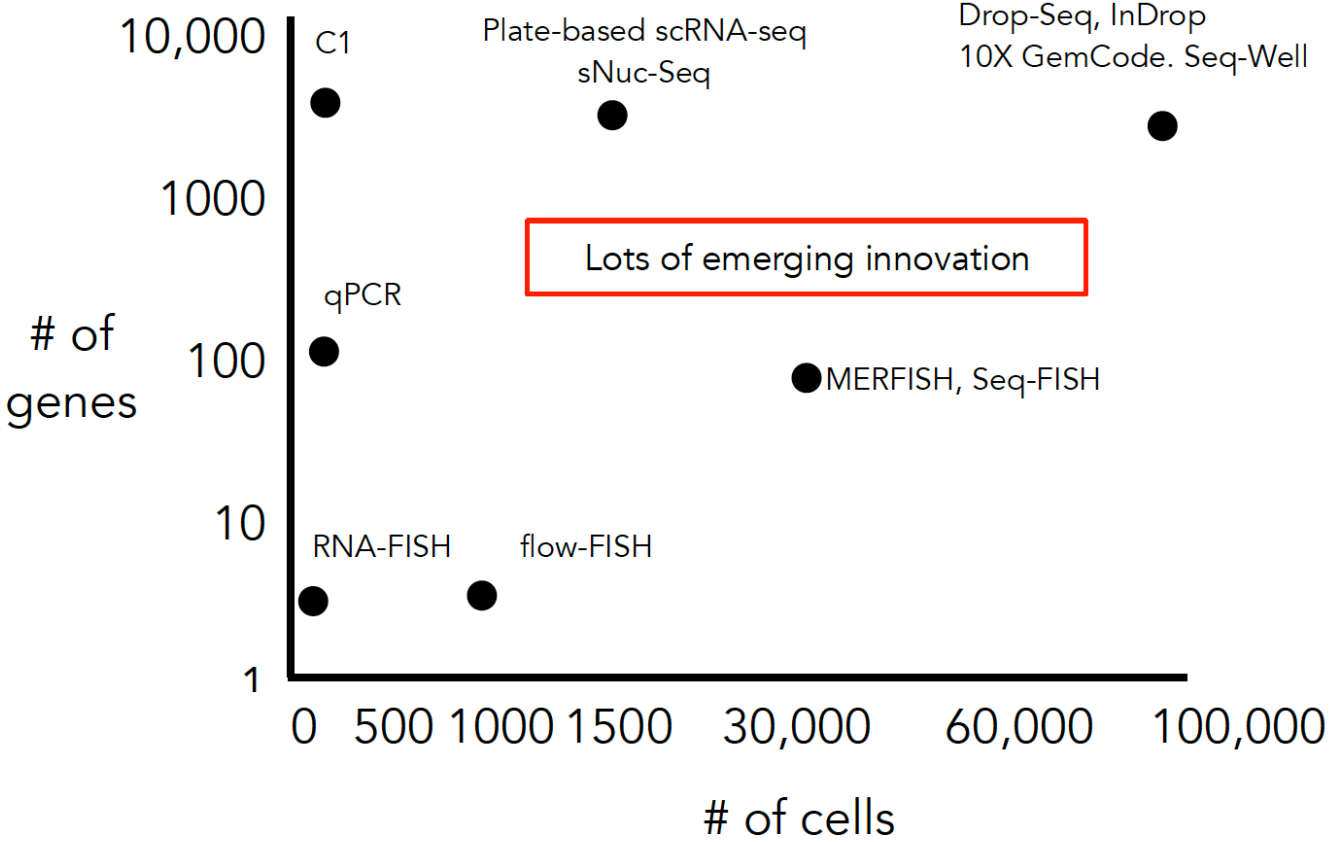


Amplification method

Sequencing method

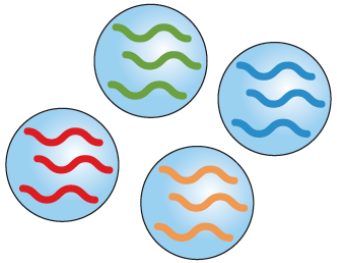
Applications

Tradeoff between scale & resolution

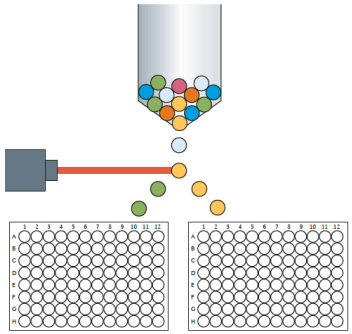


Cell Isolation

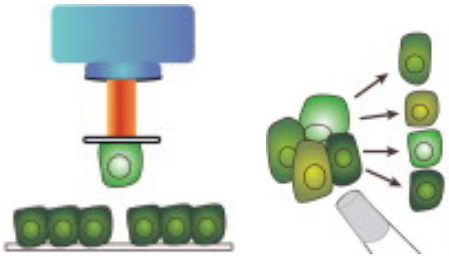
Trapping cells in droplets



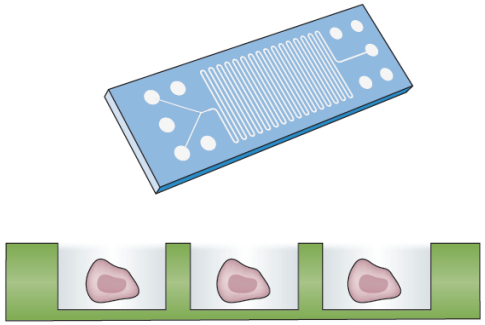
FACS / cell sorter



Microdissection & micromanipulation



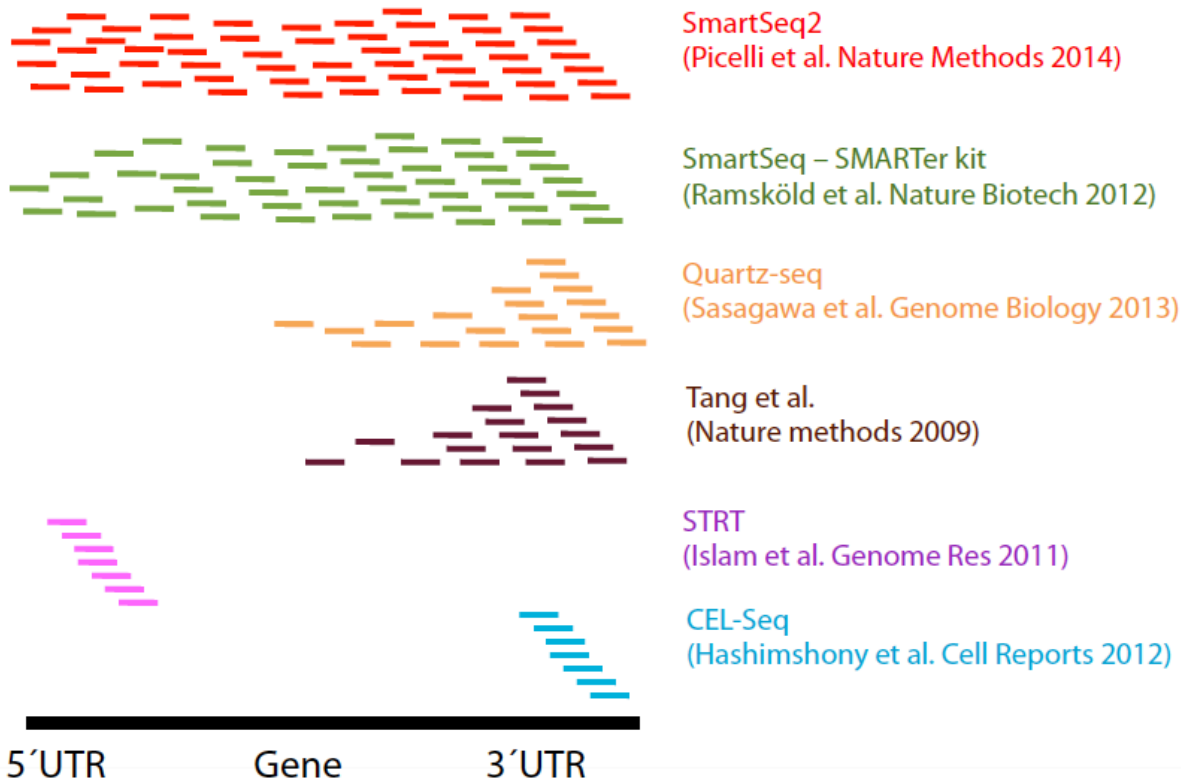
Microfluidics & microwells



Amplification method

Sequencing method

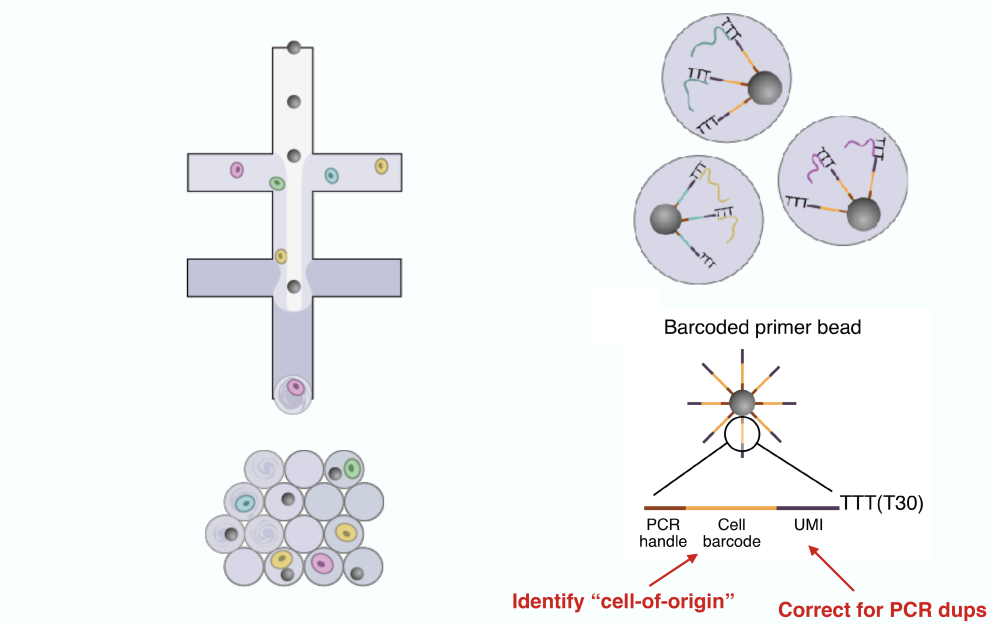
Applications



Unique molecular identifies (UMIs) and cellular barcodes

- **Cellular barcodes**
  - Introduced at RT step with one unique sequence per cell
  - Enables pooling many libraries into one tube for subsequent step (reduces cost & technical errors)

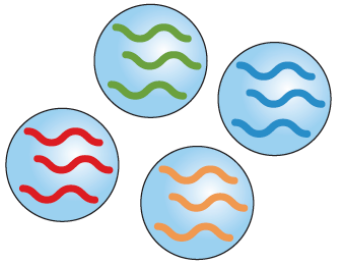
- **UMIs**
  - Introduce random sequences at the beginning of each sequence
  - Reduces effect of amplification bias by removing PCR duplicate



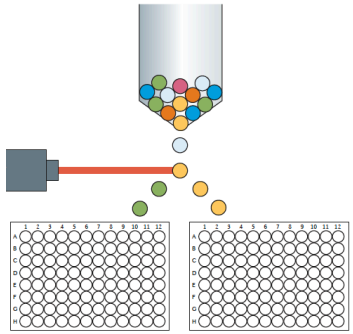
Macosko*, Basu et al *Cell* 2015

Cell Isolation

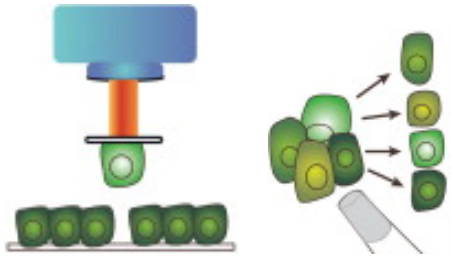
Trapping cells in droplets



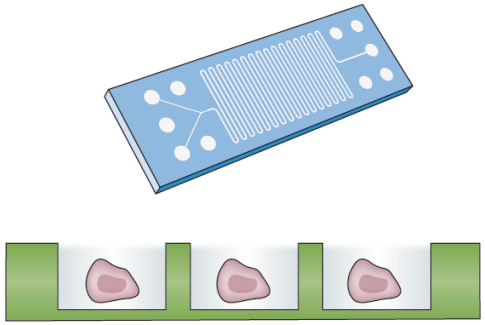
FACS / cell sorter



Microdissection & micromanipulation



Microfluidics & microwells



Amplification method

InDrops
DropSeq
10X

Cel-Seq
MARS-Seq
SCRB-Seq
CEL-Seq2

SMART-seq
SMART-Seq2
STRT-Seq

Pooled PCR amplification

Individual cell amplification

Sequencing method

3' Sequencing

Full-length sequencing

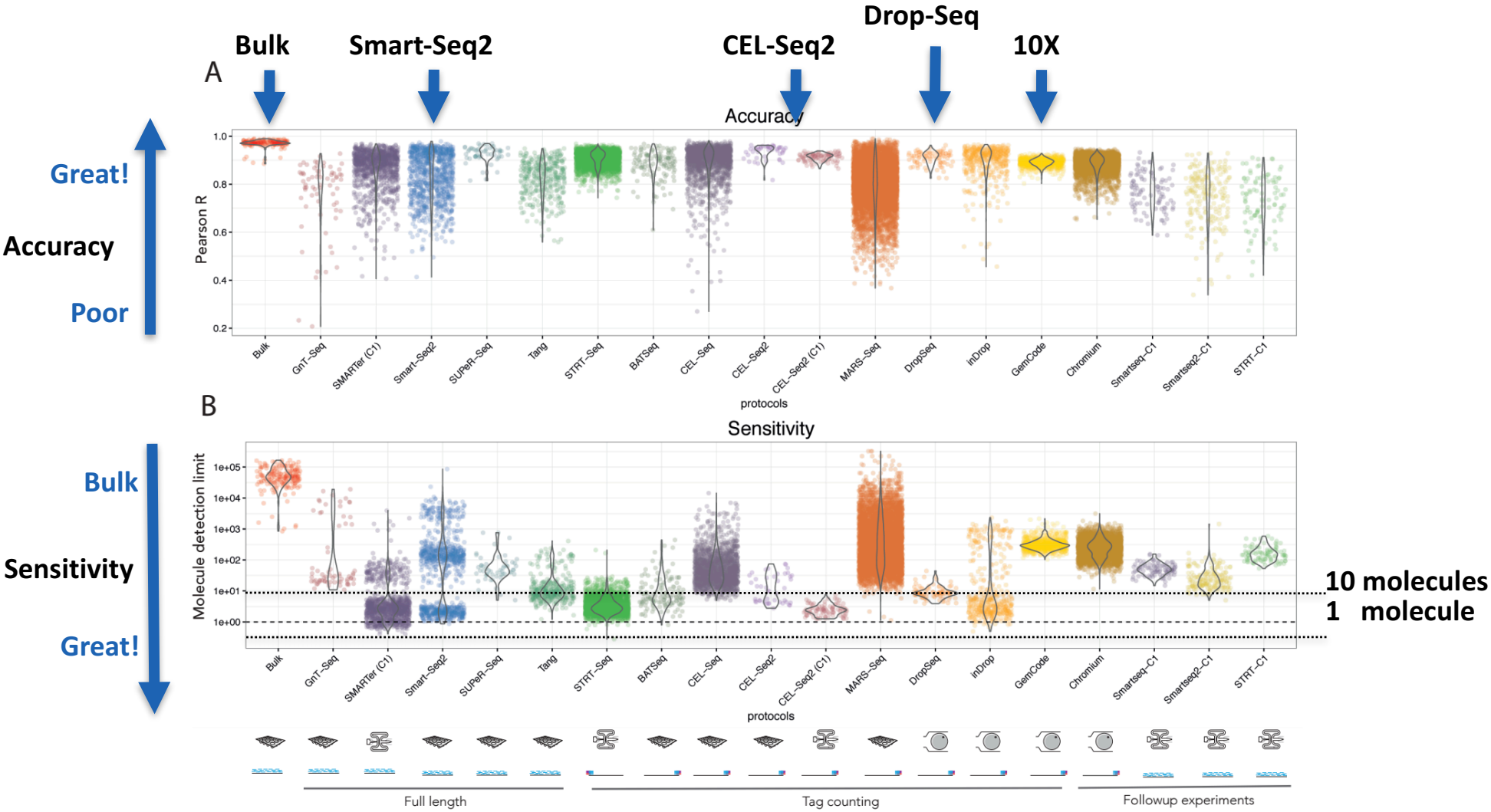
Applications

- **Detect Gene expression**

- **Detect gene expression**
- **Splicing variants**
- **Allelic variations**
- **TCR & BCR repertoire diversity**

Adapted from Papalexi E et al. Nat Rev Immunol 2017

Sensitivity and Specificity

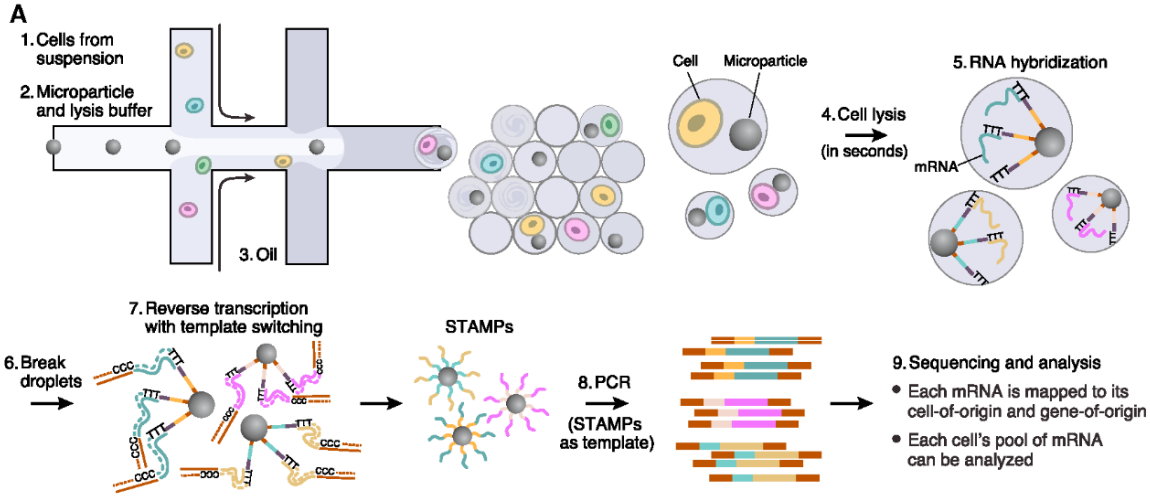


→ All better than bulk
→ Many between 1 and 10 molecule detection
→ Sensitivity dependent on sequencing depth → can sequence more!
→ Sensitivity = critical when studying lowly expressed genes

Svensson V et al. Nat Methods 2017
Tim tickle

Scalability – Massively parallel scRNAseq approaches

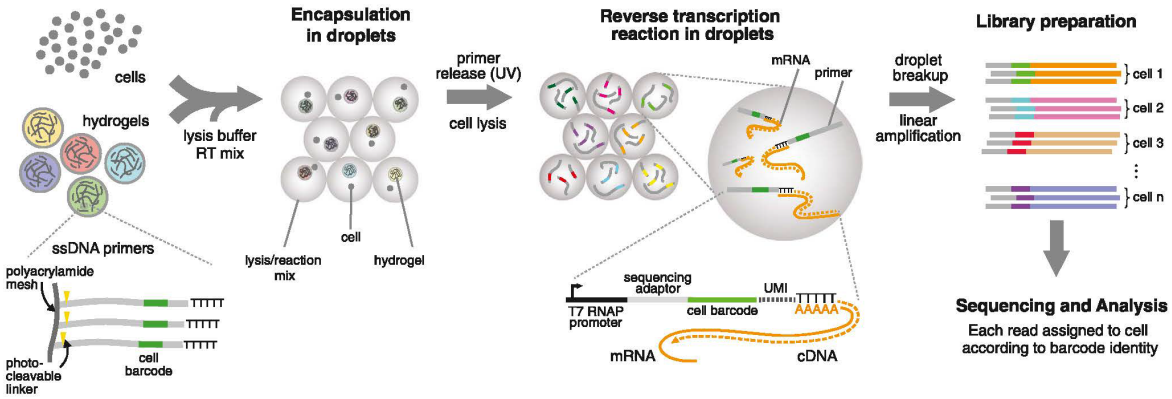
DropSeq



Macasko et al. Cell 2015

- → Cell lysed in the drop & hybridize to primers attached to beads
- → STAMP: single cell transcriptomes attached to microparticles
- → Droplets are broken & RT/template switching occurs on pool

InDrops



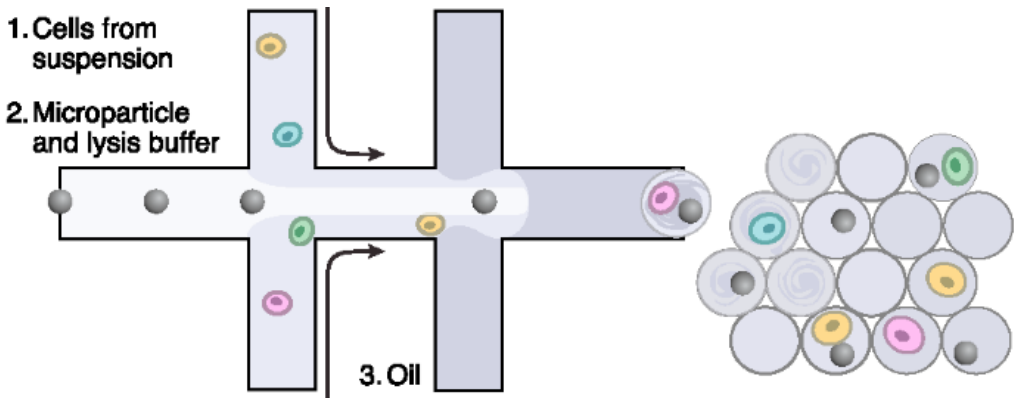
Klein et al. Cell 2015

- → Lysis and reverse transcription occurs in the droplet
- → Samples are frozen after RT as RNA:DNA in gel

Adapted from Boswell S.
https://iccb.med.harvard.edu/single-cell-core

DropSeq vs. InDrops

DropSeq

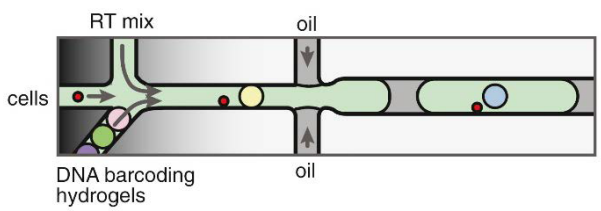


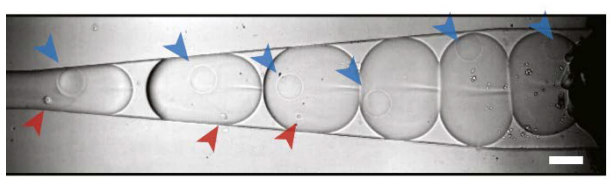
→ 1/10 droplets contain microparticle

→ 1/10 droplets contain cell

→ 1/20 droplets contain both cell and microparticle

InDrops





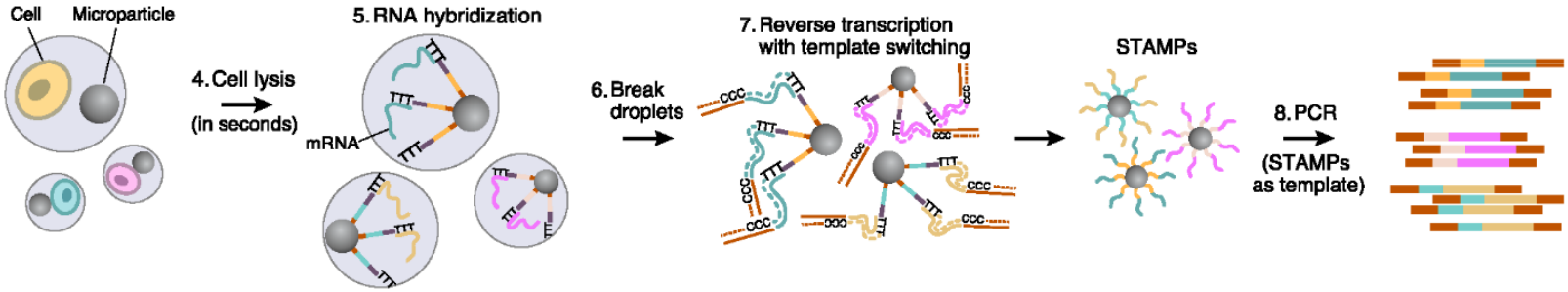
→ Match speed of bead injection with speed of droplet generation

→ Nearly every droplet loaded with one barcode

Adapted from Boswell
https://iccb.med.harvard.edu/single-cell-core

DropSeq vs. InDrops

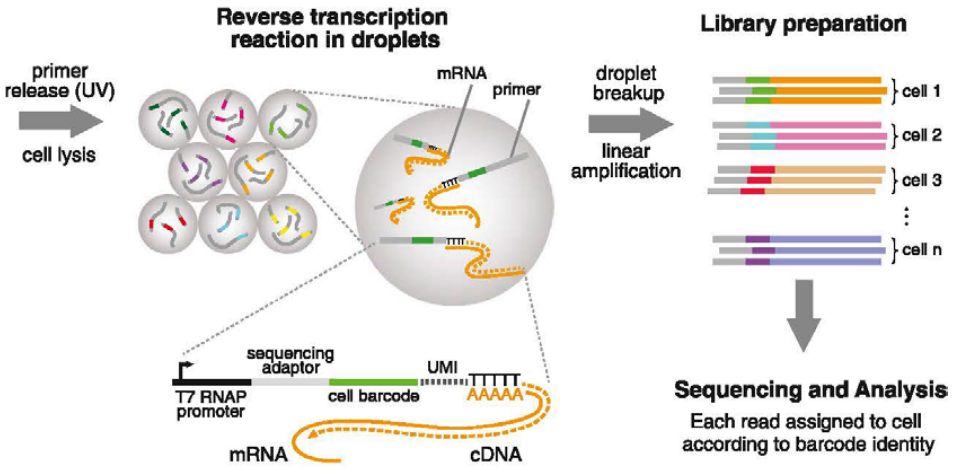
DropSeq



→ Smart-Seq: RT/template-switching to tagmentation

→ Immediate lysis

InDrops



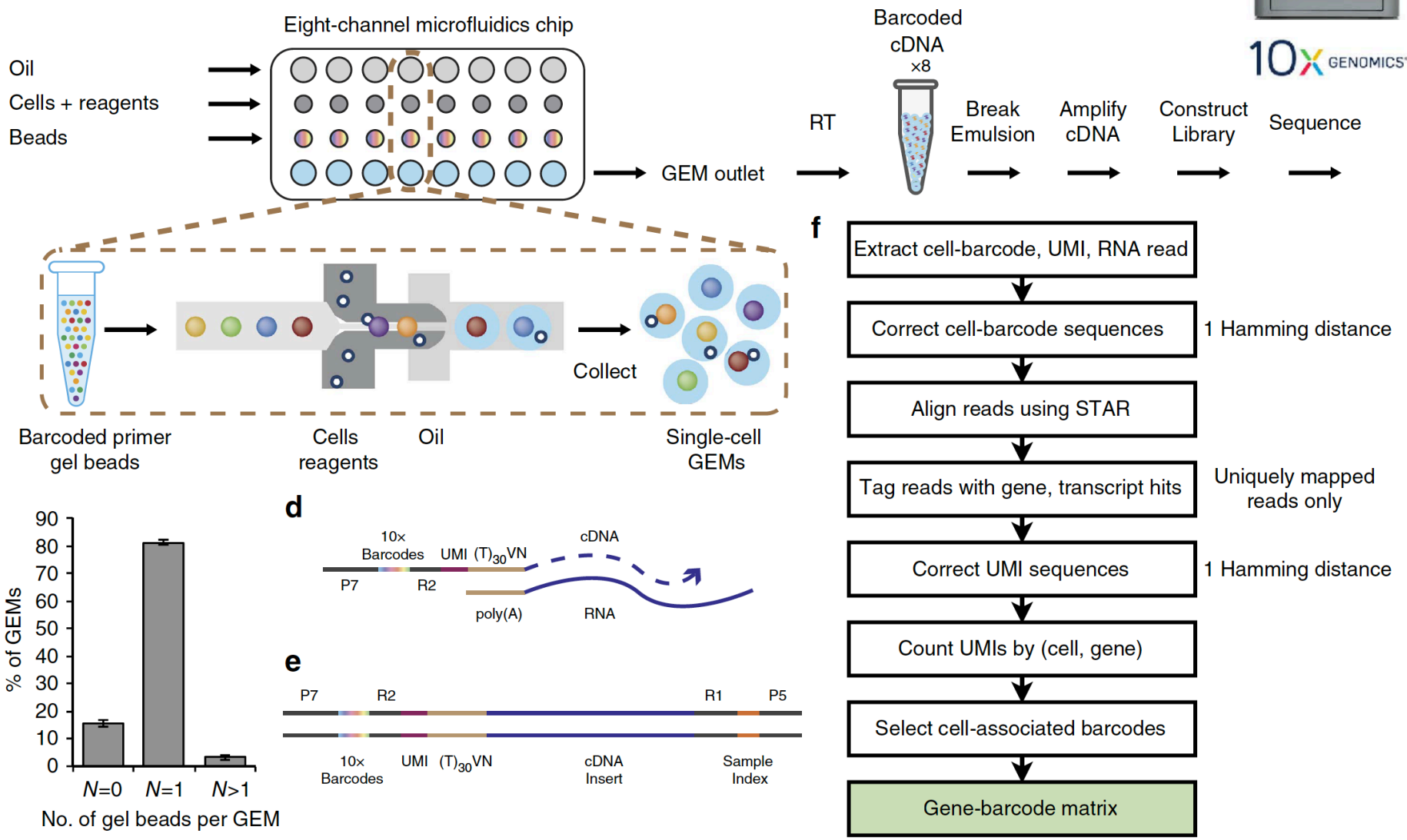
→ CEL-Seq: RT/second strand synthesis to IVT and RNA fragmenting

→ Gentile lysis that may not be completed until sample collection finished

Adapted from Boswell
https://iccb.med.harvard.edu/single-cell-core

10X Genomics 3' mRNA sequencing

10X GENOMICS



A higher throughput "plug & play" version of InDrops

Zheng et al. Nat Comm 2017

inDrops, DropSeq, 10X Genomics 3' mRNA sequencing

	Capture Efficiency*	Doublet Rate	Number of samples at once	Library prep
InDrops	50-90%	3%	1	CEL-Seq
10X	50-60%	3%	8	CEL-Seq
DropSeq	5-10%	10%	1	Smart-Seq

** Capture efficiency is of the cells that reach the device*

→ InDrops and 10X are very similar technologies

→ InDrops & DropSeq → more labor intensive but customizable & cheaper; need some expertise in handling microfluidics

→ 10X → more scalable (8 samples in parallem), "plug & play", comes with standardized pipeline, but much more expensive (upfront cost $25k)

→ DropSeq requires 100,000 cells as input vs. 7,000 cells for 10X

→ Number of transcripts detected varies between approaches (also tissue dependent)

→ Cost per library varies greatly!

Your biological question will dictate which method(s) to pursue

- **Different scRNAseq have pros and cons**

- **Needing scalability**
  - Do you know which cell type you want to study?
  - Looking to generate cell census?
  - Are you trying to map very rare cell subsets for which you do not know markers?
  - Dissecting tissue (healthy/disease) ecosystem?
  - Mapping response to treatment (pre vs. post), not knowing which cells would be affected?

- **Needing hither sensitivity and/or full-length transcripts**
  - Predicting binding specificity of TCR receptors?
  - Interested in studying a particular population, potentially rare?
  - Want to map at higher resolution the transcriptome of signaling components/less abundant transcripts to dissect particular biology / pathway?
  - Interested in mapping allelic expression, x-chromosome inactivation, or spliced isoforms?

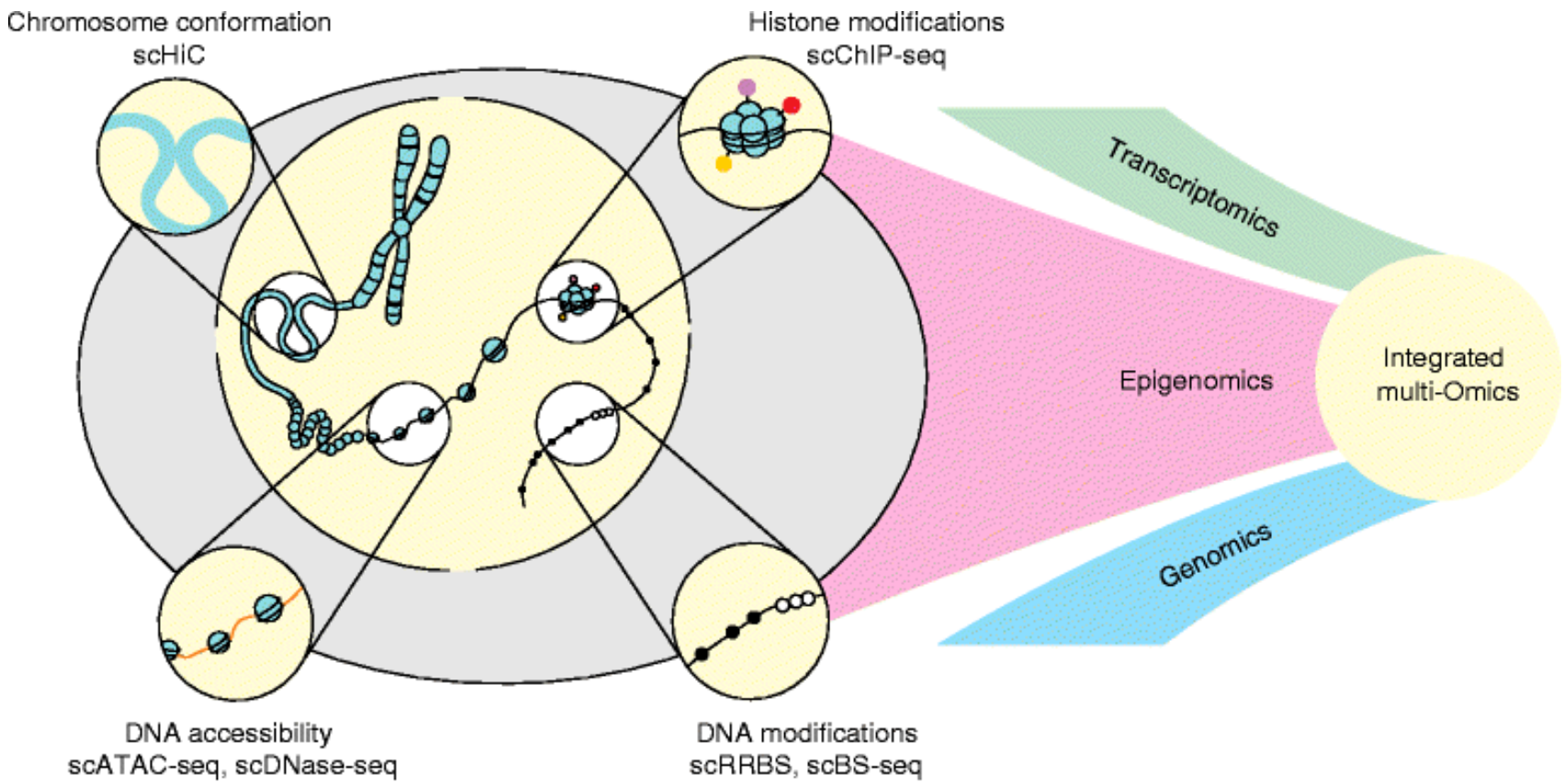
- **Interested into lineage tracing?** Consider combining DNA/RNA seq and ATACseq

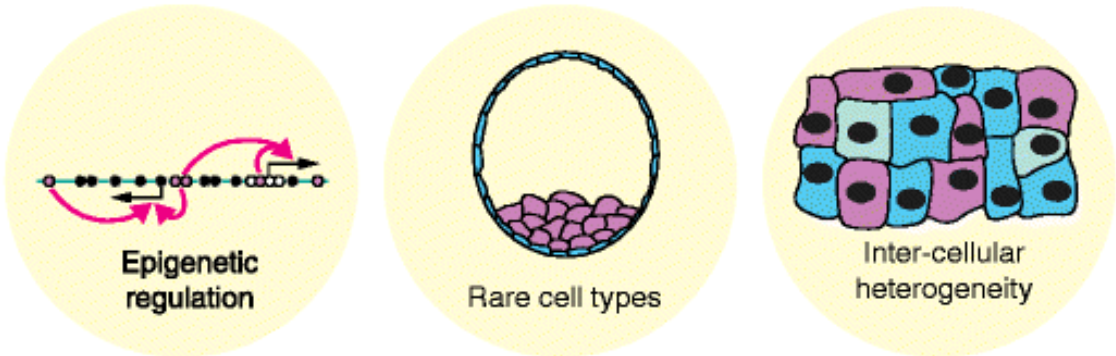
- **Trying to decipher interacting partners** → spatial omics

- Consider **combining different approaches** in your study design!

Other single cell readouts & multi-omics approaches

Single cell epigenomics





Clak et al. 2017

Technological evolving landscape – stay tuned!

Split & pool barcoding (not relying on microfluidics)

- SPLiT-Seq (Rosenberg et al. 2017)
- SCI-Seq (Cao et al. 2017)

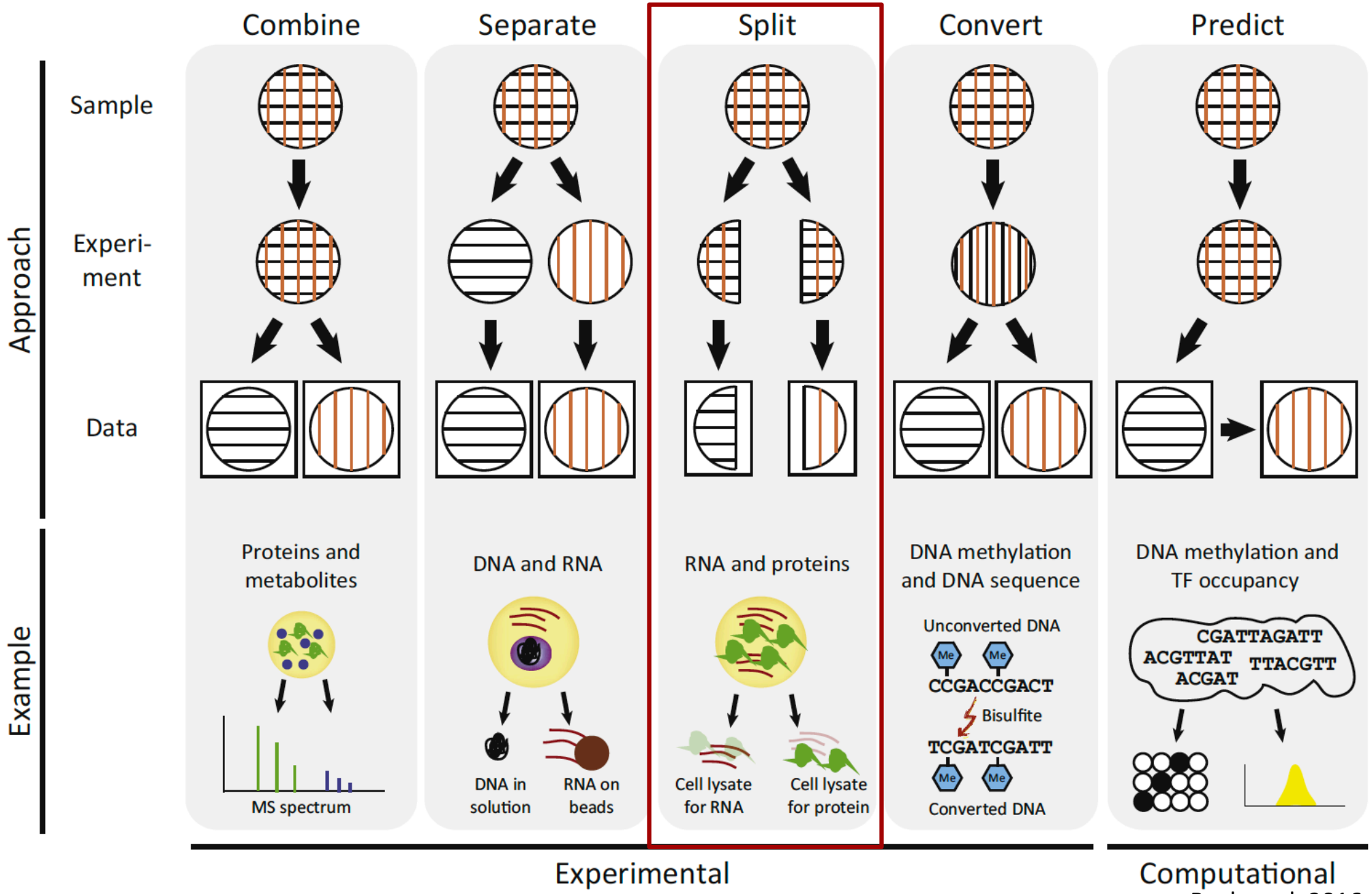
Spatial 'Omics"

- Multiplex FISH (Seq-FISH, MERFISH)
- *In situ* RNA-seq (e.g. FISSEQ)

Multi-omics

- DNA + RNA (G +T)
- RNA + protein (T + P)
- Epigenome + RNA

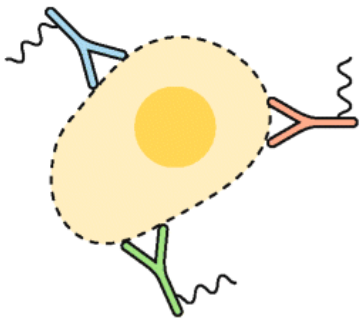
Multi-omics strategies



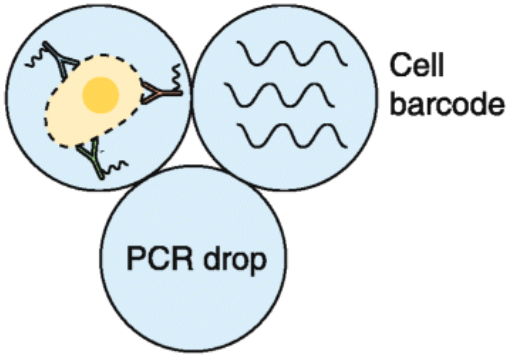
Bock et al. 2016

RNA-Seq & cell-surface proteomics in a drop!

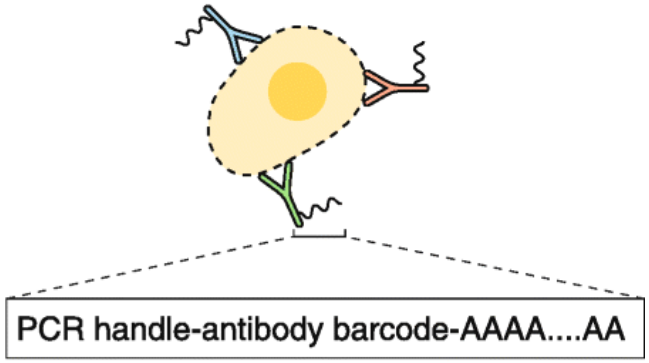
ABCD Ullal et al. 2014



Abseq Payam et al. 2017



CITE-Seq Stoeckius et al. 2017



	ABCD	Abseq	CITE-Seq
Cell surface proteins	✓	✓	✓
Intracellular proteins	✓	✗	✗
Multiplexed	✗	✓	✓
RNA-seq	✗	✗	✓
Sequencing technology	Nanostring	Illumina	Illumina

Baron et al. 2017

Analysis – I have generated some scRNAseq data … what are the next steps

Starting with the gene expression matrix

"Raw data"

```
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
AAATTATGACGAAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
TTGCCGTGGAGTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
GTTAAACGTACCGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CGTTAGATGGCATCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTGCCGTGGAGTCGTGAGGG............TTCCAAGG
CGTTAGATGGCACCTGTGTA............TGGTACGT
GTTAAACGTACCATCCGGTG............TTAAACCG
...........
...........
...........
```

(Hundreds of millions of reads)

"Processed" data

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

- Sequences derived from different scRNAseq assays are complex and vary
- Different pipelines are needed to address different sequence formats
- Common steps include:
  - Aligning
  - QC
  - Read counting

Credit: Karthik Shekhar

Starting with the gene expression matrix

"Raw data"

```
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
AAATTATGACGAAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
TTGCCGTGGAGTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
GTTAAACGTACCGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CGTTAGATGGCATCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTGCCGTGGAGTCGTGAGGG.............TTCCAAGG
CGTTAGATGGCACCTGTGTA.............TGGTACGT
GTTAAACGTACCATCCGGTG.............TTAAACCG
...........
...........
...........
```

(Hundreds of millions of reads)

Qualifications

- Full length vs. 3' vs 5'
- Poly A vs. Random priming
- Strand-specific vs non-specific
- UMI vs. non-UMI

"Processed" data

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

Credit: Karthik Shekhar

Starting with the gene expression matrix

"Raw data"

```
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
AAATTATGACGAAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
TTGCCGTGGAGTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
GTTAAACGTACCGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CGTTAGATGGCATCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTGCCGTGGAGTCGTGAGGG.............TTCCAAGG
CGTTAGATGGCACCTGTGTA.............TGGTACGT
GTTAAACGTACCATCCGGTG.............TTAAACCG
...........
...........
...........
```

(Hundreds of millions of reads)

Qualifications

- Full length vs. 3' vs 5'
- Poly A vs. Random priming
- Strand-specific vs non-specific
- UMI vs. non-UMI

- Quality filtering
- Cell barcode stratification
- Alignment
- Multimapping reads/intronic reads
- Quantification / UMI collapse

"Processed" data

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

Credit: Karthik Shekhar

Starting with the gene expression matrix

"Raw data"

```
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
AAATTATGACGAAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
TTGCCGTGGAGTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
GTTAAACGTACCGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CGTTAGATGGCATCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTGCCGTGGAGTCGTGAGGG.............TTCCAAGG
CGTTAGATGGCACCTGTGTA.............TGGTACGT
GTTAAACGTACCATCCGGTG.............TTAAACCG
...........
...........
...........
```

(Hundreds of millions of reads)

Qualifications

- Full length vs. 3' vs 5'
- Poly A vs. Random priming
- Strand-specific vs non-specific
- UMI vs. non-UMI

- Quality filtering
- Cell barcode stratification
- Alignment
- Multimapping reads/intronic reads
- Quantification / UMI collapse

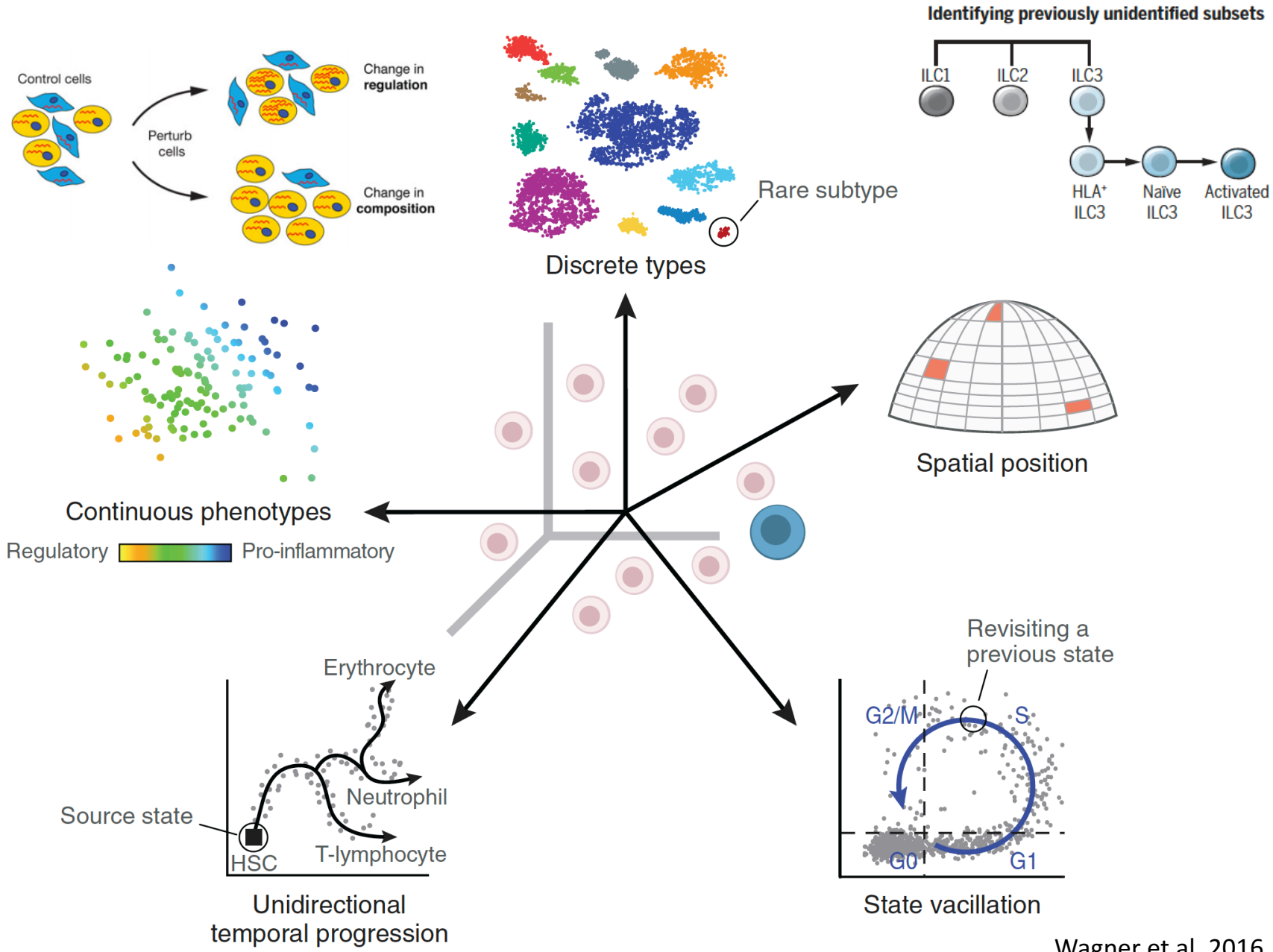
"Processed" data

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

Once I have my gene expression matrix, what's next?

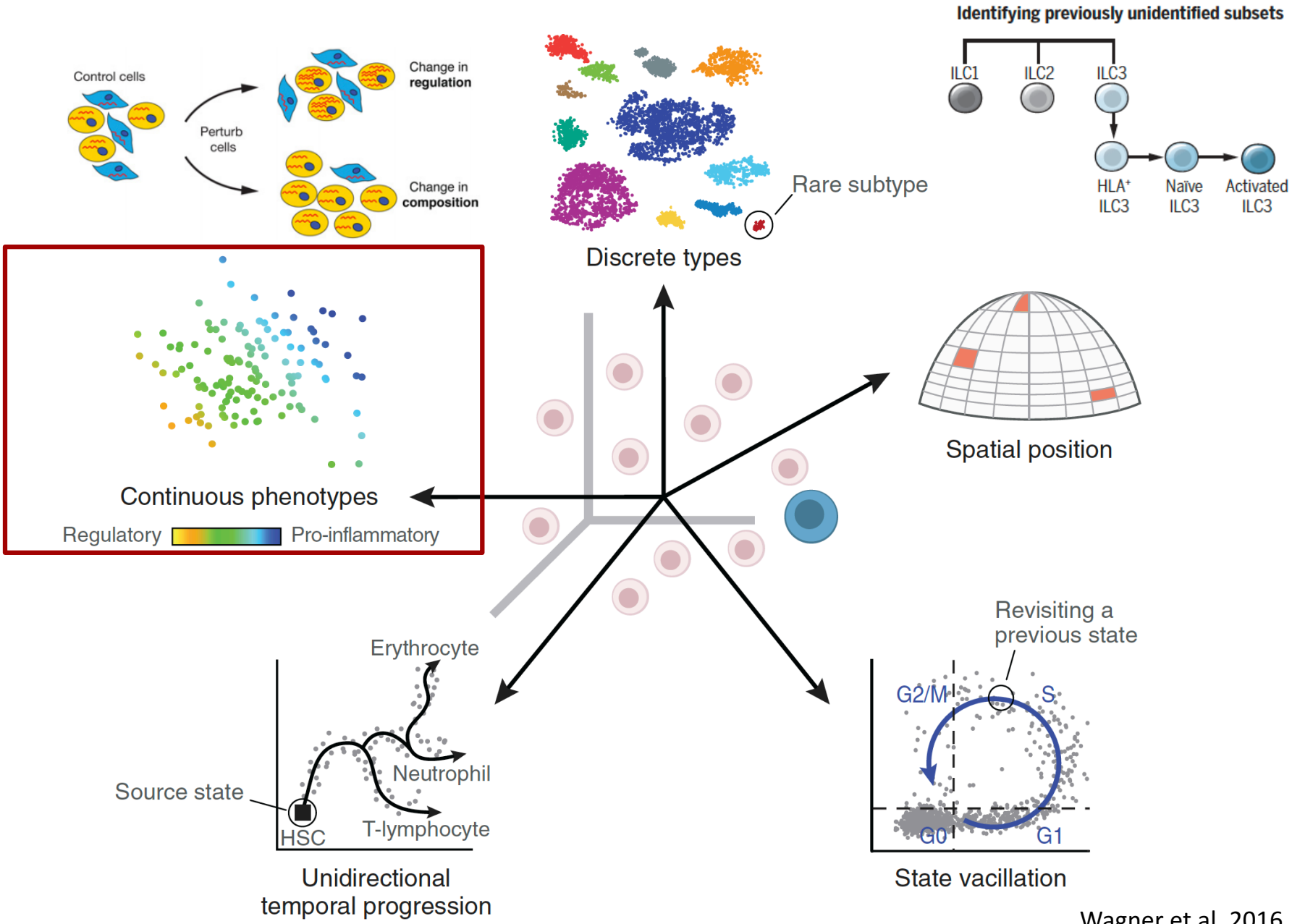
Credit: Karthik Shekhar

Inference – from data to biology



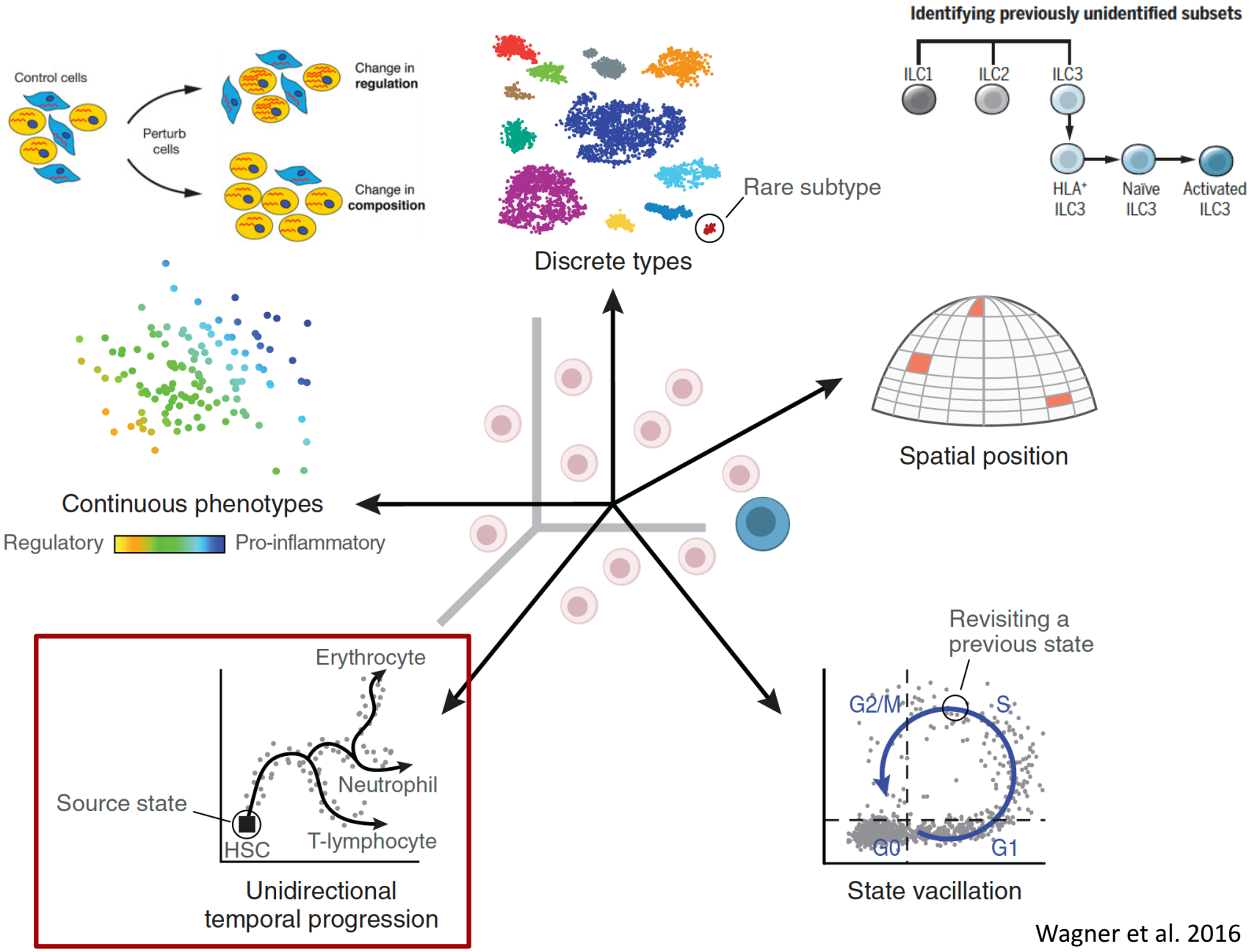
Wagner et al. 2016
Stubbington et al. 2017

Inference – from data to biology



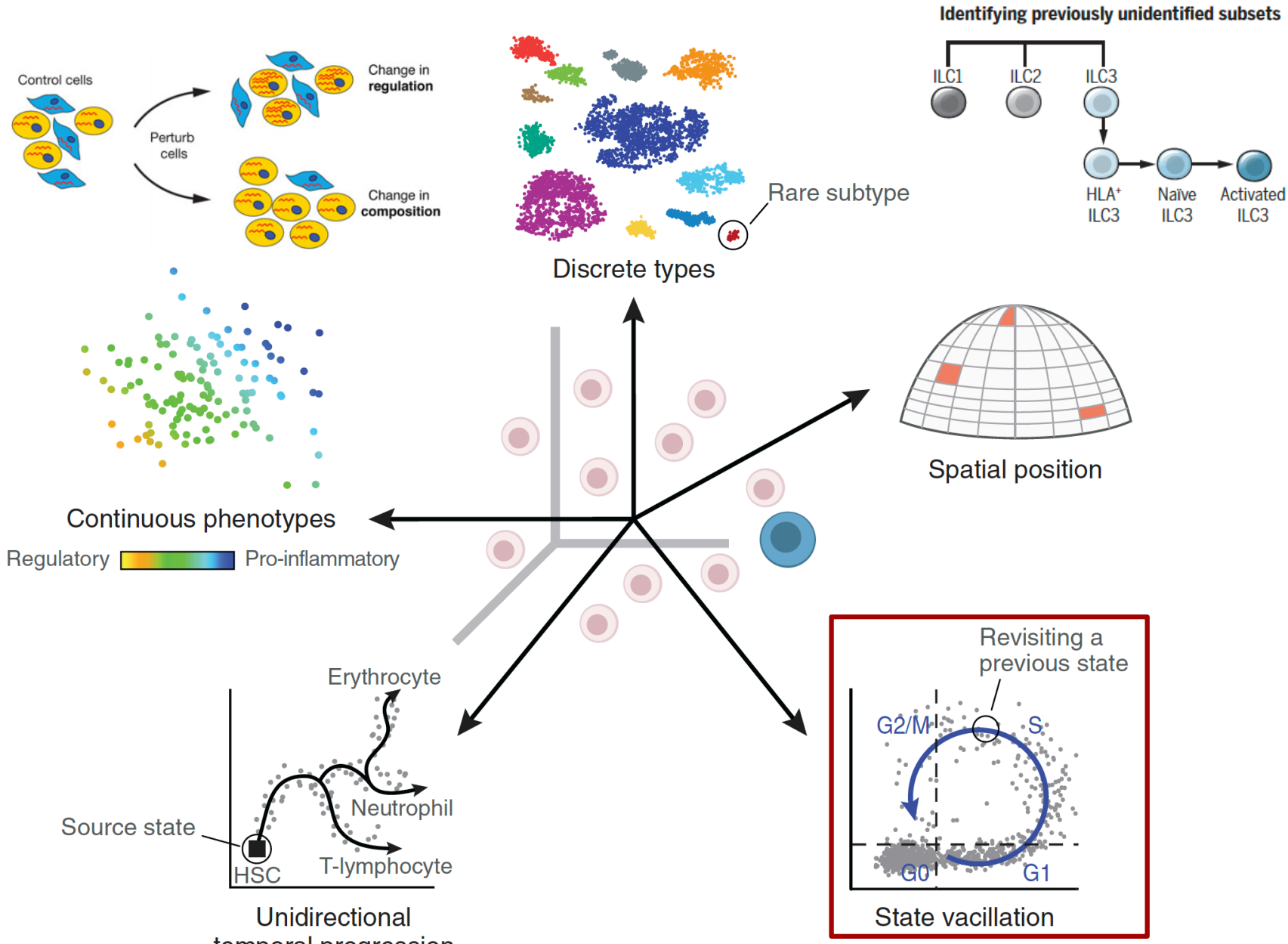
Wagner et al. 2016
Stubbington et al. 2017

Inference – from data to biology



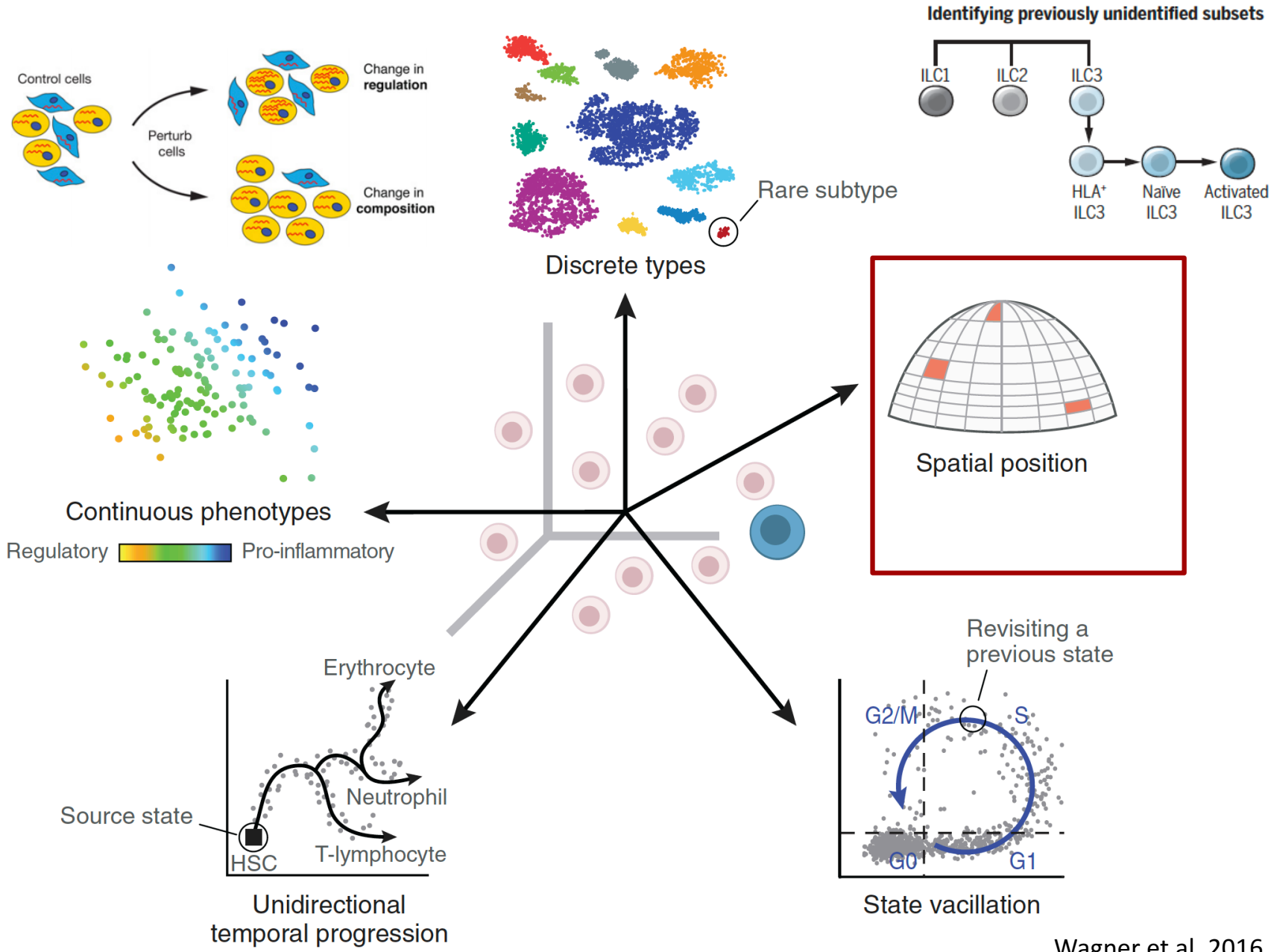
Wagner et al. 2016
Stubbington et al. 2017

Inference – from data to biology



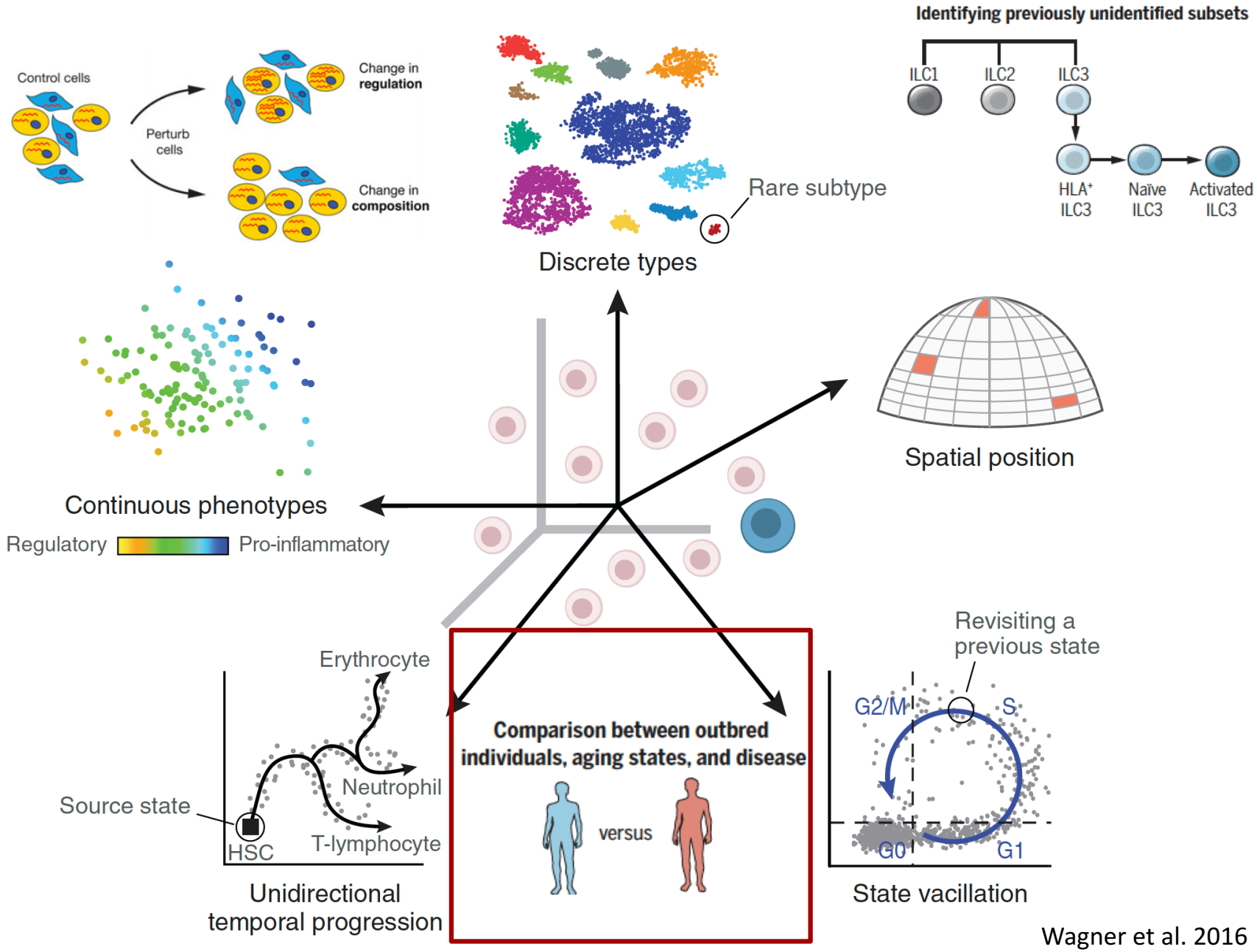
Wagner et al. 2016
Stubbington et al. 2017

Inference – from data to biology



Wagner et al. 2016
Stubbington et al. 2017

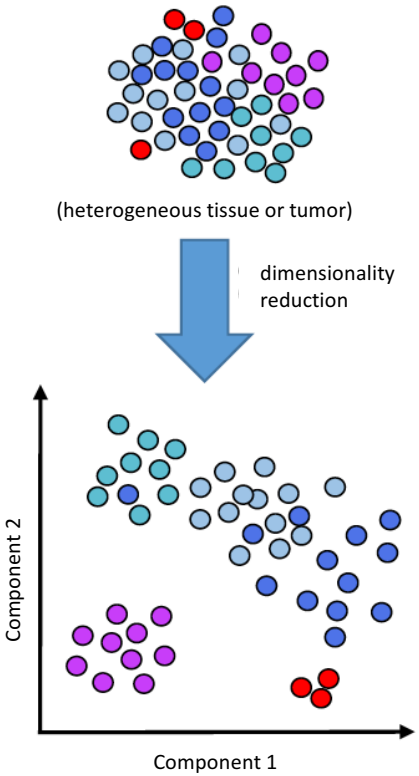
Inference – from data to biology



Wagner et al. 2016
Stubbington et al. 2017

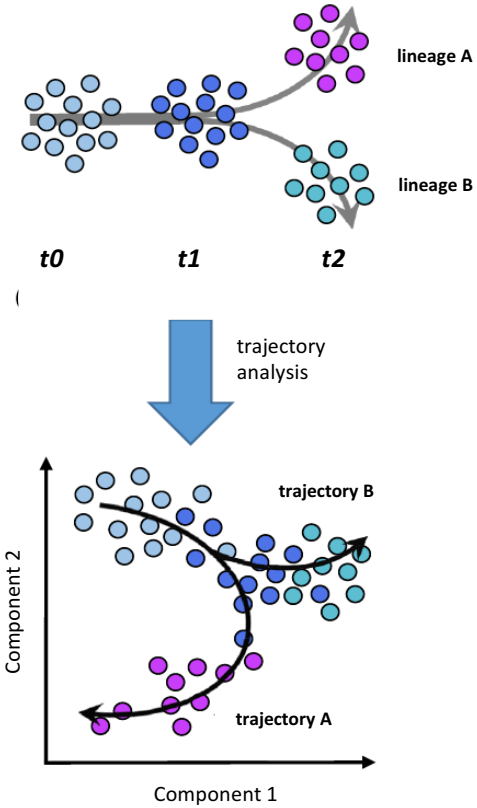
The type of biological questions will dictate analyses to be undertaken

(1) Deconvolution of heterogeneous population



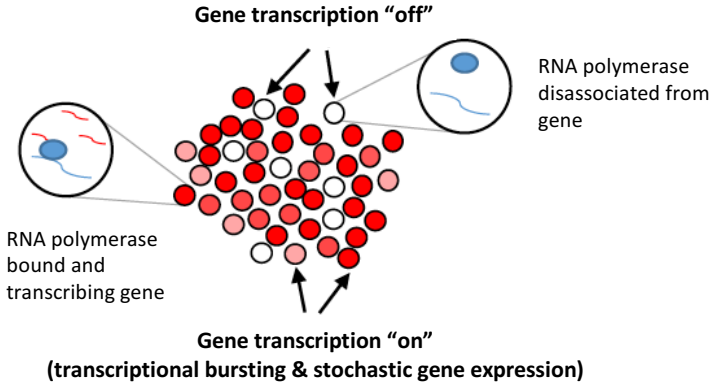
→ Cell atlas
→ Diseased vs. healthy
→ Pre- vs. post-therapy

(2) Trajectory analysis

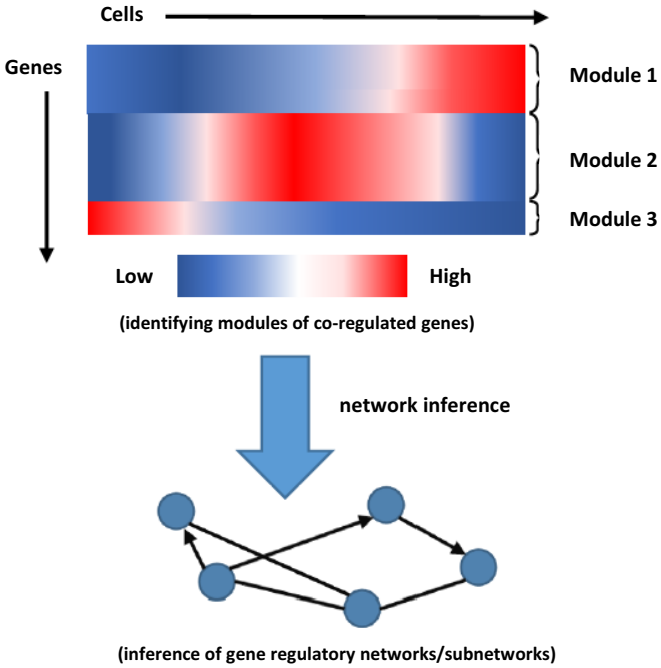


→ Cell state transition:
  – cell differentiation
  – response to stimulus
→ Development

(3) Dissecting transcription mechanics

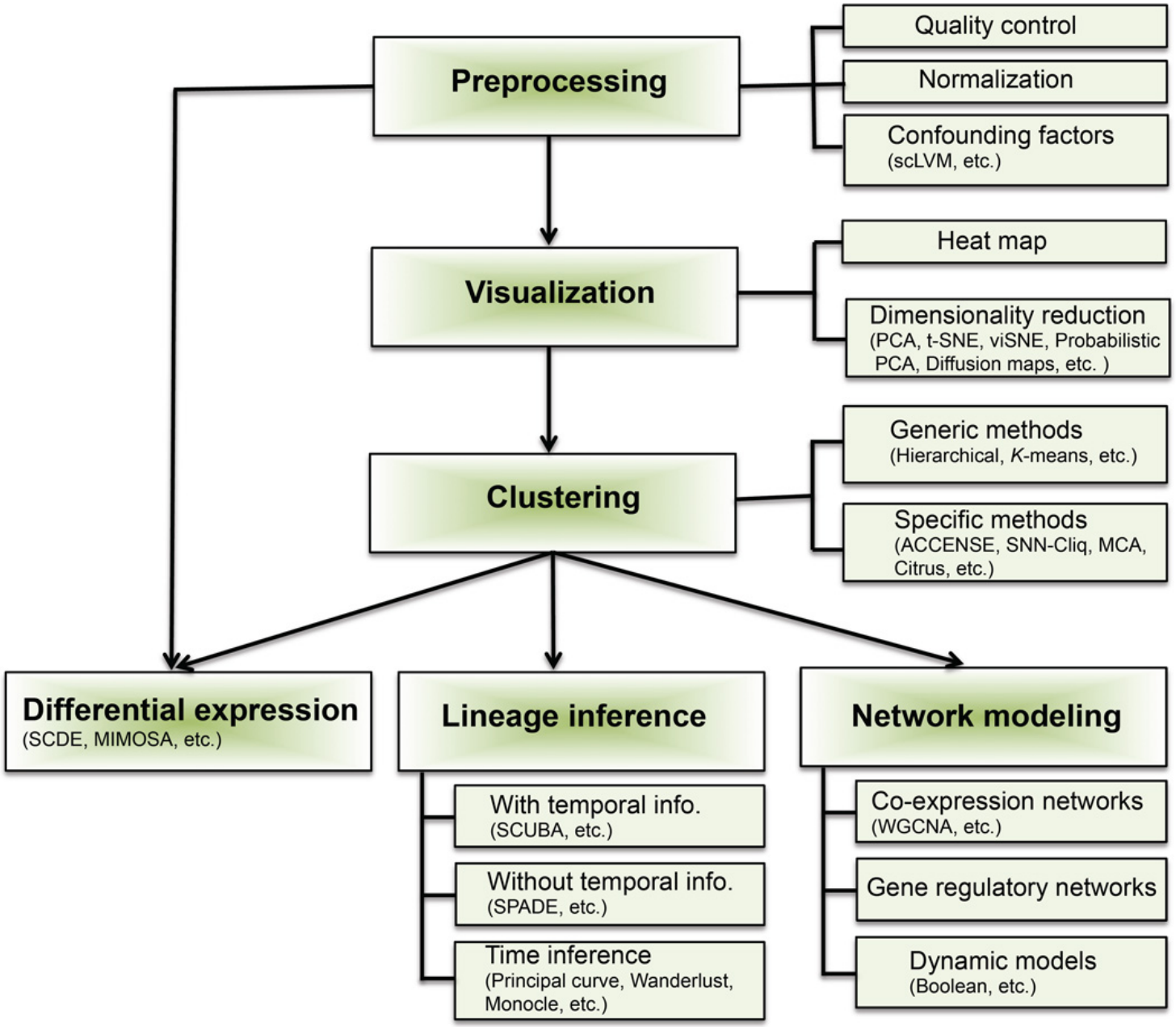


(4) Network Inference



Adapted from Trapnell 2016

Typical chart for scRNAseq analysis

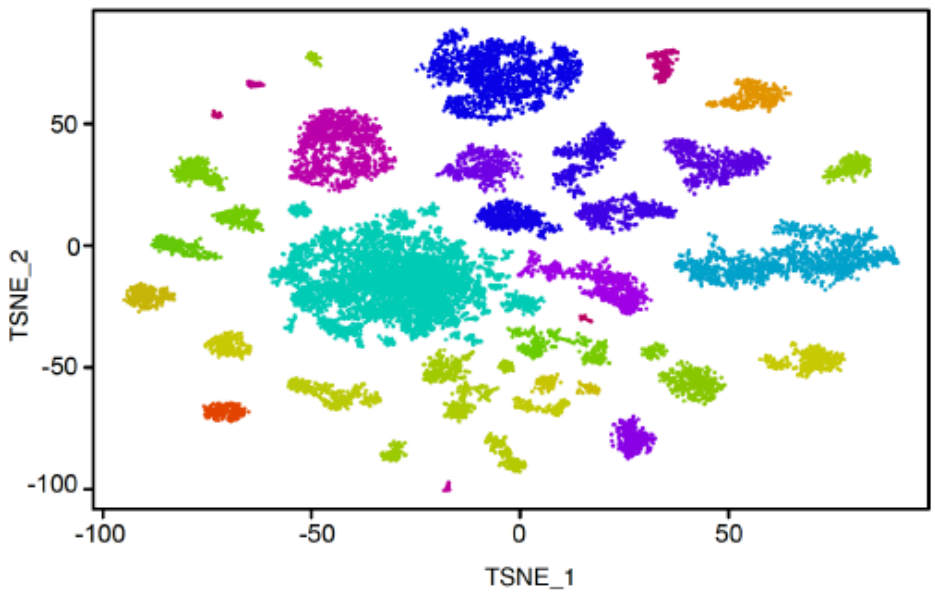


Saadatpour et al. 2015

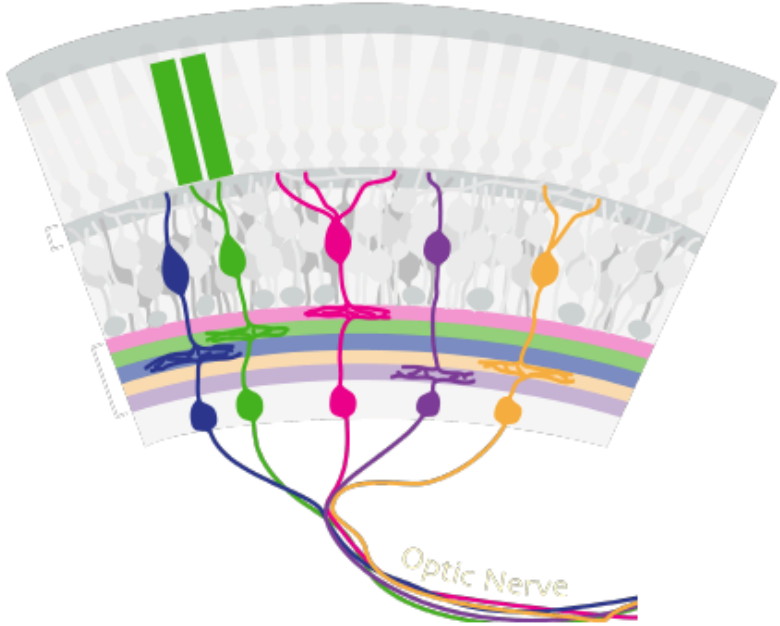
Unsupervised clustering for cell type discovery

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

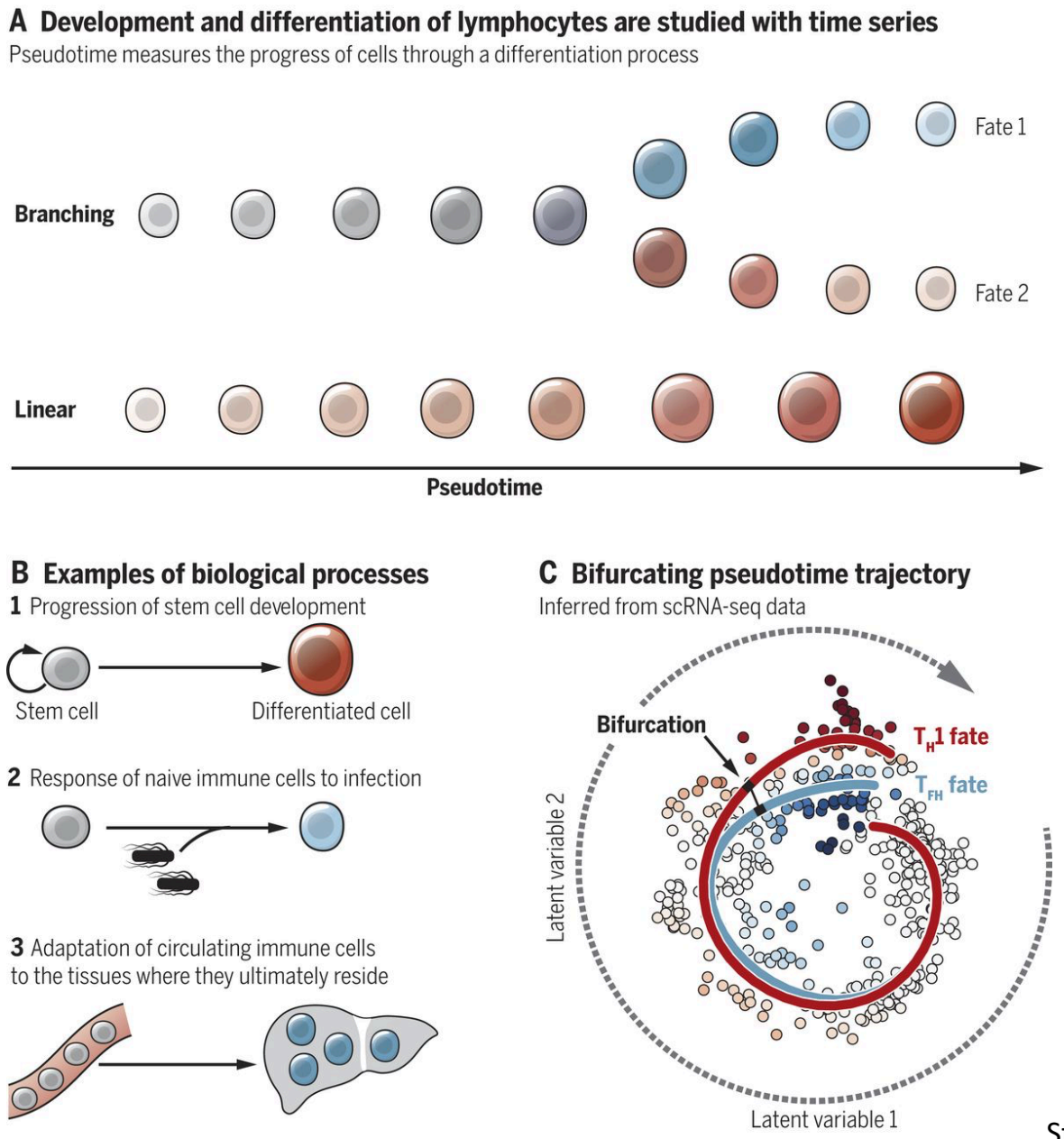
Macosko et al., *Cell*, 2015



Proof of Principle



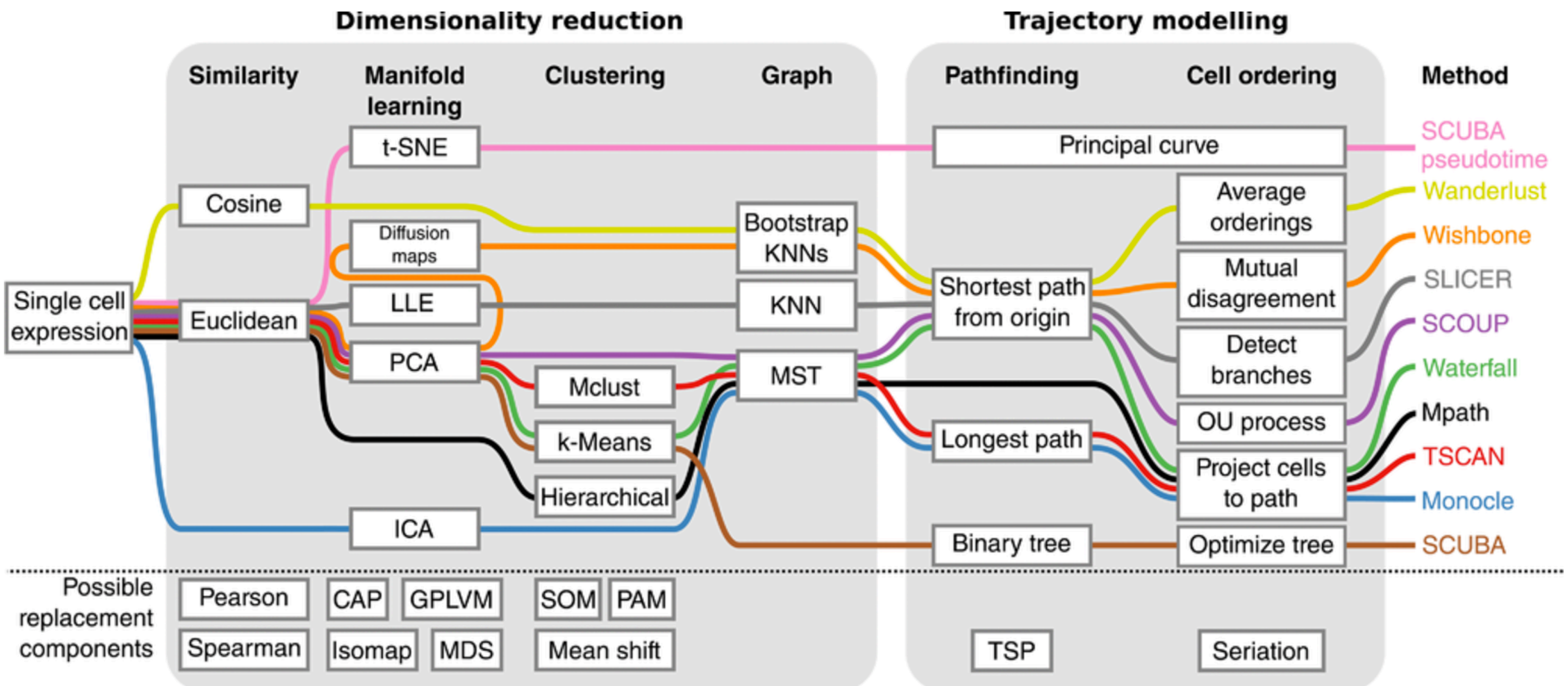
Single cell trajectory analyses



Stubbington et al. 2017

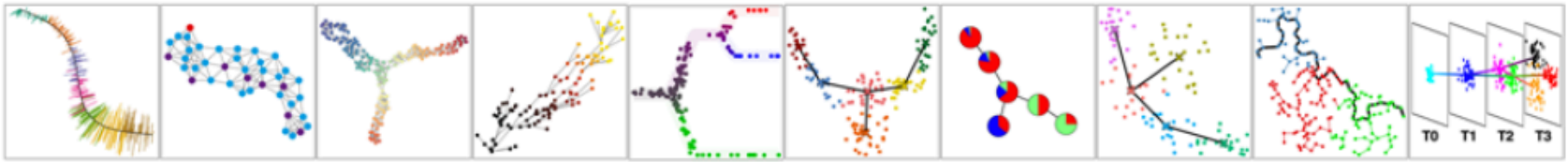
Single cell trajectory analyses

Simplified representation of dataset



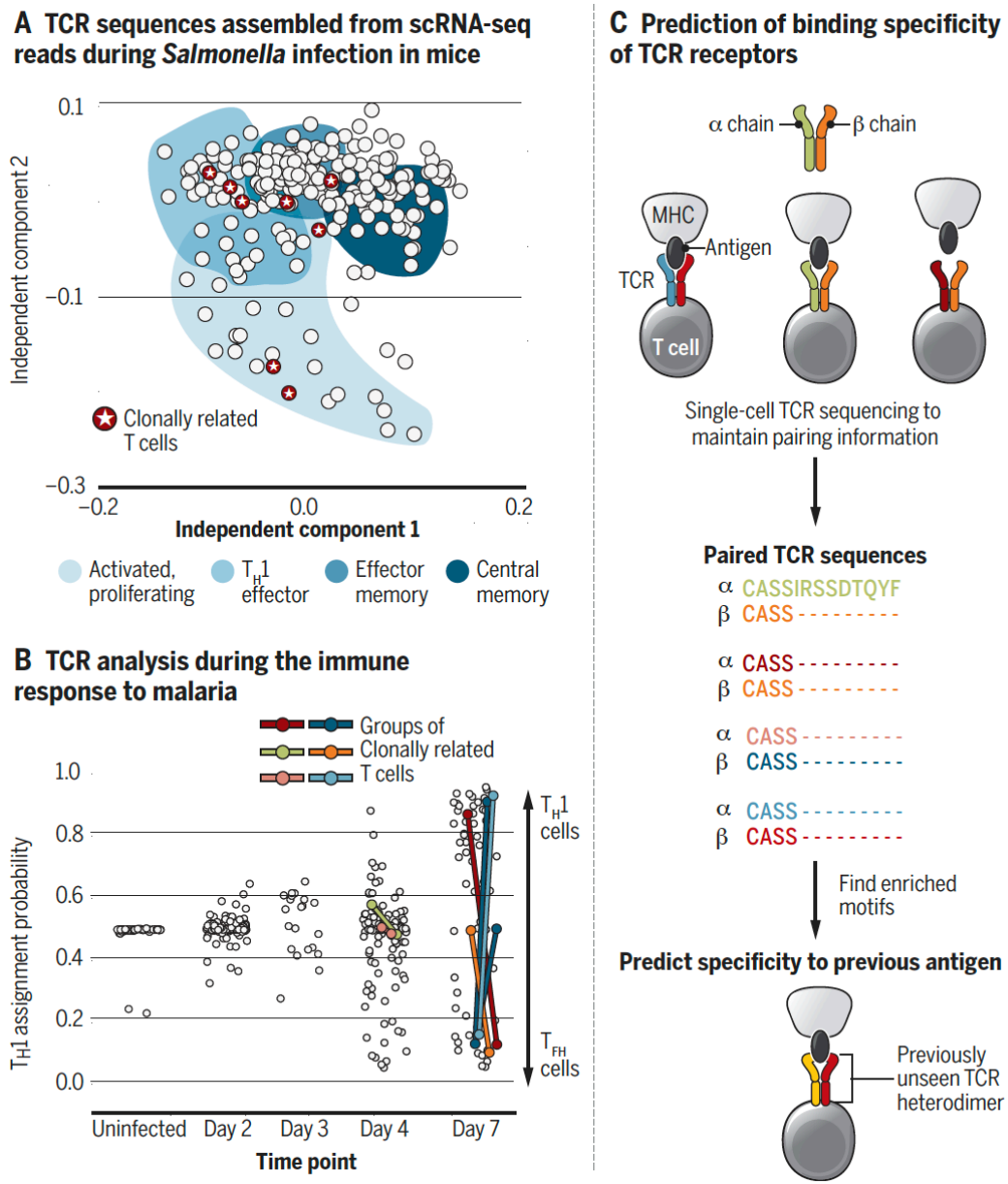
Cannoodt et al. 2016

Single cell trajectory analyses

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	–	+	–	–
Scalability w.r.t. cells	–	–	±	±	–	±	+	+	–	±
Scalability w.r.t. genes	+	+	+	+	–	+	±	±	±	+
Code and documentation	–	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	–	±	–	+	+	+

Cannoodt et al. 2016

Revealing T clone distributions between transcriptional state by analyzing TCR (requires full-length or custom primers)



Stubbington et al. 2017

Examples of additional analyses

- Pathway and Geneset OverDispersion Analysis (PAGODA; Fan et al. Nat. Methods 2016)
- Alternative splicing
- Allelic expression
- Copy-number variation
- N.B. : alternative splicing and allelic expression require full length methods
  - Can draw conclusions with certainty only for highly expressed genes with good coverage
  - Take into consideration the drop-out rate → a unique splice form/allele in a single cell may be the results of detection issue

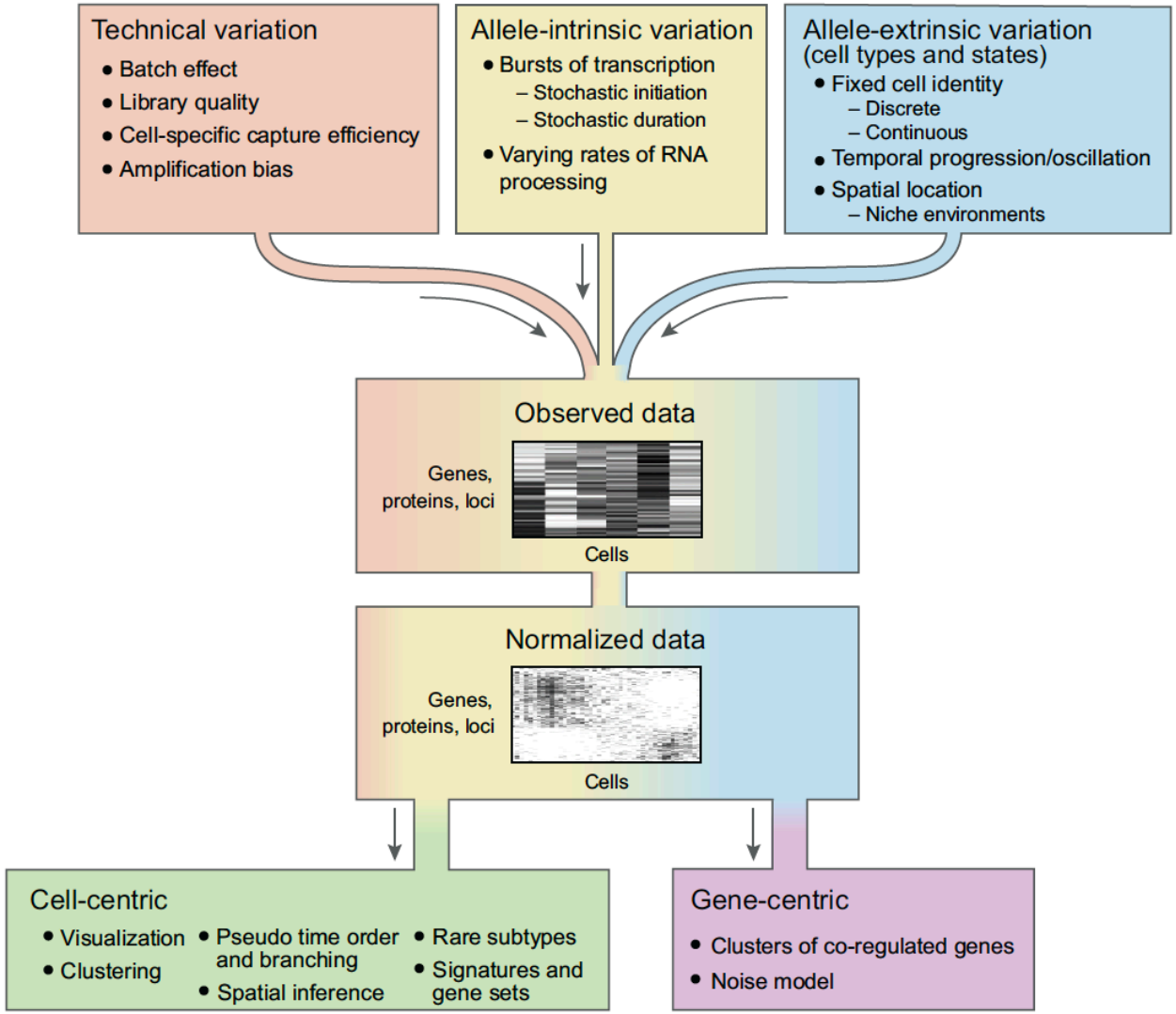
List of references for methods & tutorial

- Thank you to Sean Davis for the "Awesome single cell" compilation of software packages (and the people developing these methods) for analysis, including RNA-seq, ATAC-seq, etc.
  - https://github.com/seandavi/awesome-single-cell

- Examples of tutorials to get started:
  - Seurat (v2.0) - Guided Clustering Tutorial: http://satijalab.org/seurat/pbmc3k_tutorial.html
  - Sanger, Hemberg Lab scRNA-seq course materials: http://hemberg-lab.github.io/scRNA.seq.course
  - Harvard Single Cell Workshop (hosted by Peter Kharchenko): http://hms-dbmi.github.io/scw/

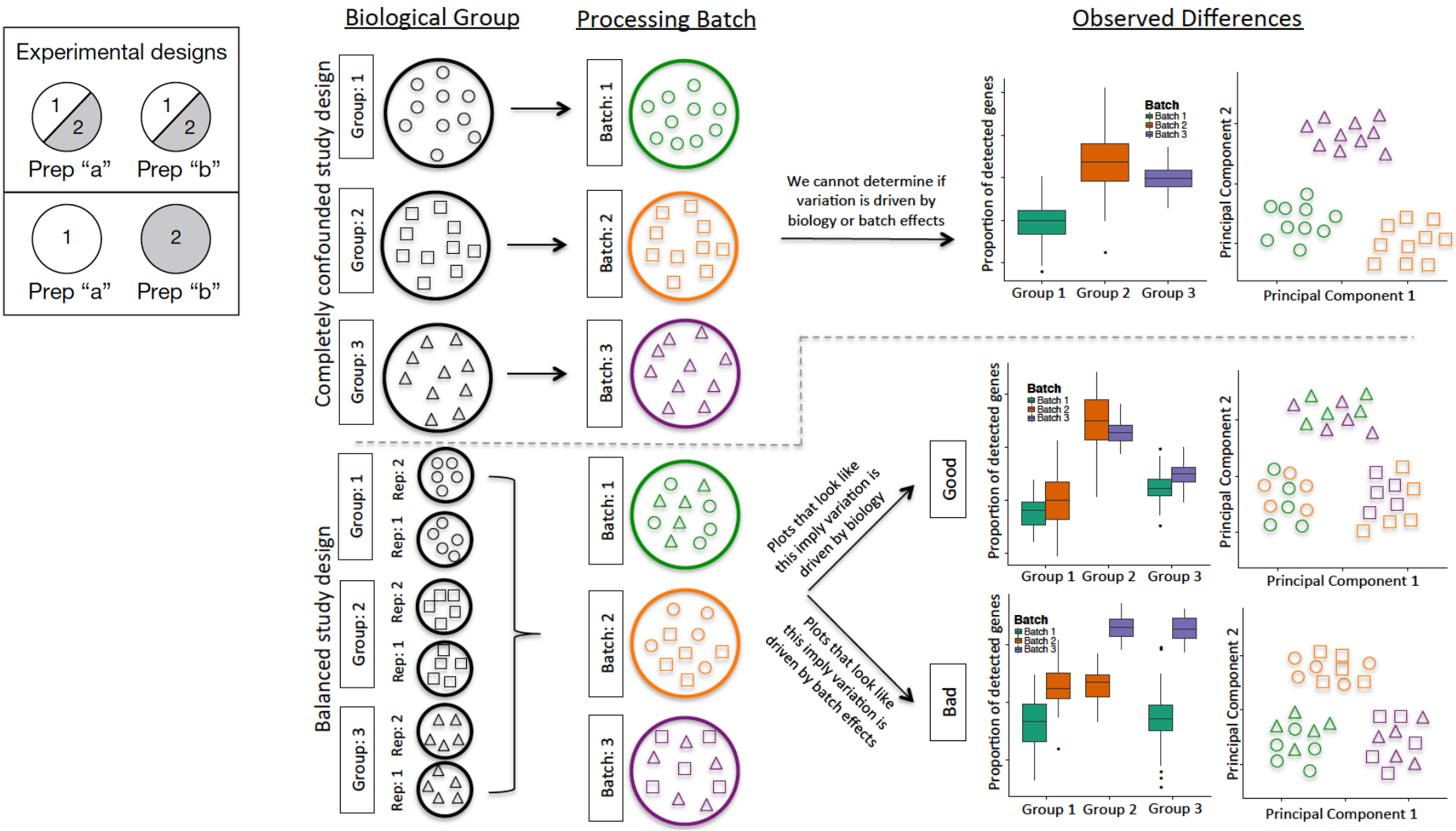
Technical challenges in scRNAseq

Biological and technical factors driving gene expression readout



Wagner et al. Nat Biotech 2016

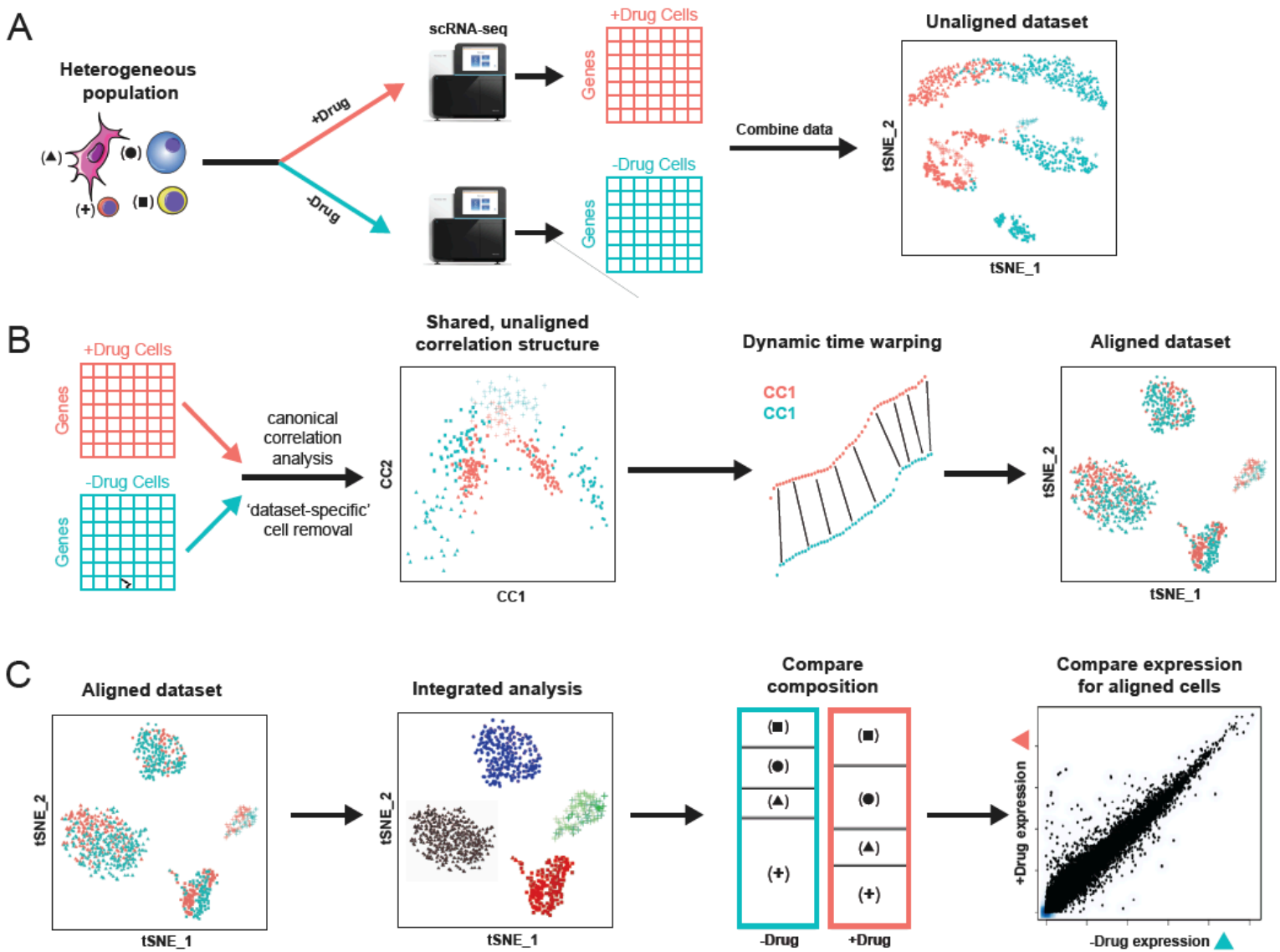
Technical confounders in scRNAseq: Batch effect



http://hemberg-lab.github.io/scRNA.seq.course

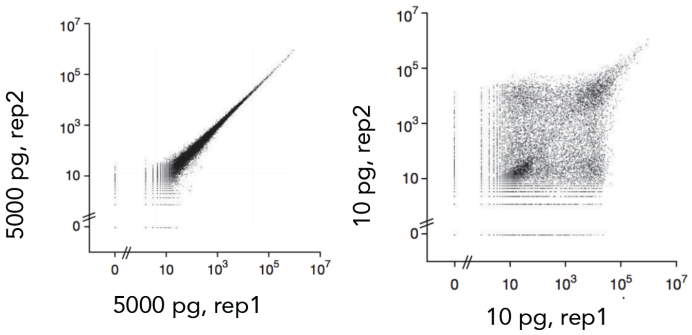
Integrated analysis of single cell transcriptomic data across conditions, technologies, and species

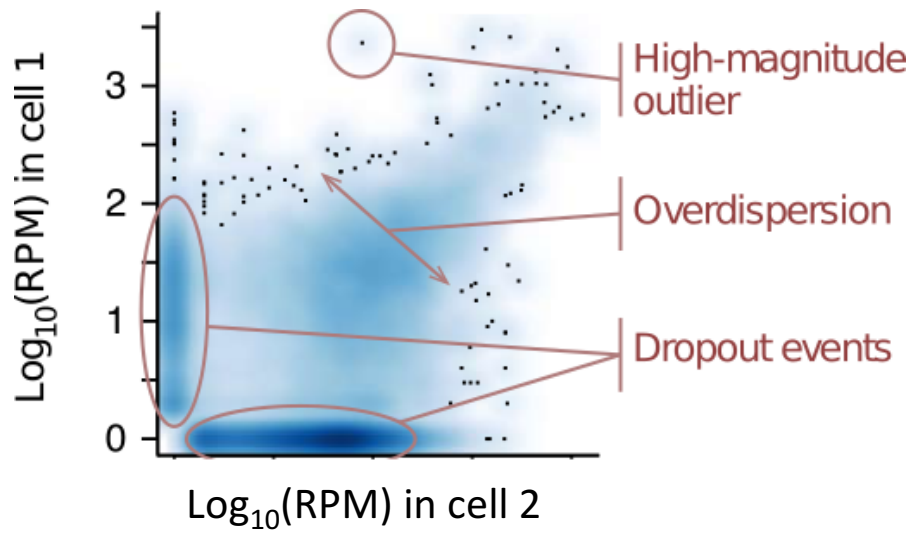
Andrew Butler[1,2] and Rahul Satija[1,2,#]



Technical confounders in scRNAseq: Dropouts

(inefficient mRNA capture → sparse data / ~10% non-zero values)





- **Zero inflation**
  - Drop-out event during reverse-transcription
  - Genes with more expression have less zeros
  - Complexity varies
- **Transcription stochasticity**
  - Transcription bursting
  - Coordinated transcription of multigene networks
  - Over-dispersed counts
- **Higher Resolution**
  - More sources of signal

BRIEF COMMUNICATIONS

Bayesian approach to single-cell differential expression analysis

Peter V Kharchenko[1–3], Lev Silberstein[3–5] & David T Scadden[3–5]

Credit: Tim Tickle

Technical confounders in scRNAseq: Dropouts

(inefficient mRNA capture → sparse data / ~10% non-zero values)

Solution: imputing missing data!

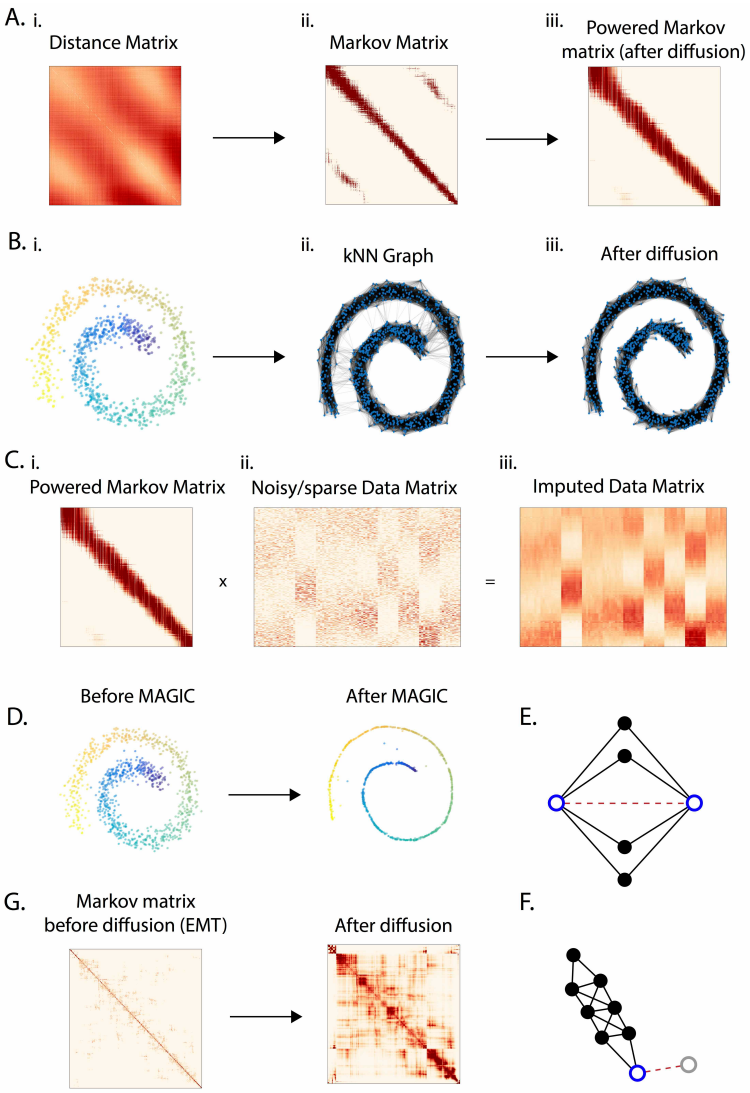
bioRxiv preprint first posted online Feb. 25, 2017; doi: http://dx.doi.org/10.1101/111591.

MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data

David van Dijk[1], Juozas Nainys[2,4], Roshan Sharma[1,3], Pooja Kaithail[1,4] Ambrose J. Carr[1,4], Kevin R. Moon[5,6], Linas Mazutis[1,2], Guy Wolf[5], Smita Krishnaswamy[6*], Dana Pe'er[1*]

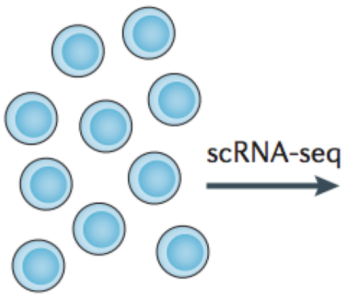
MAGIC = *Markov Affinity-based Graph Imputation of Cells*

→ *Method for imputing missing values & restoring* structure in the data



Other technical confounders in scRNAseq

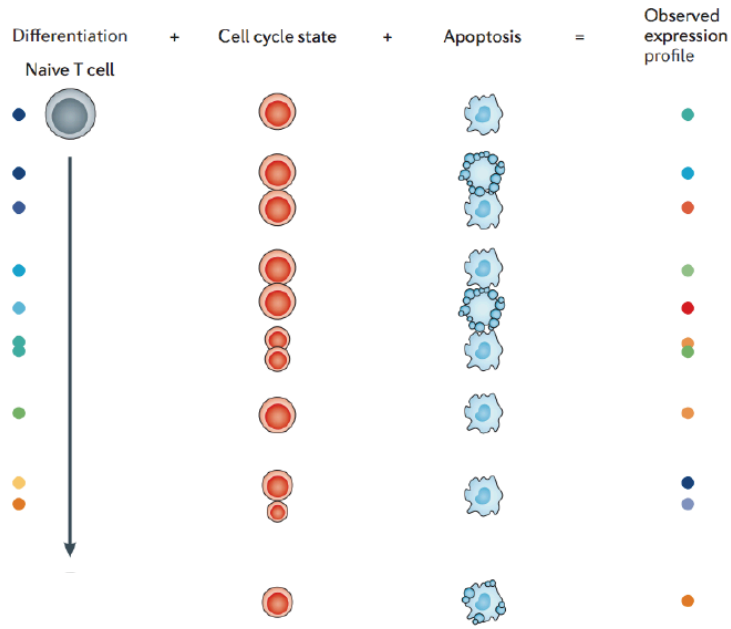
1- Variation in cell size and quality



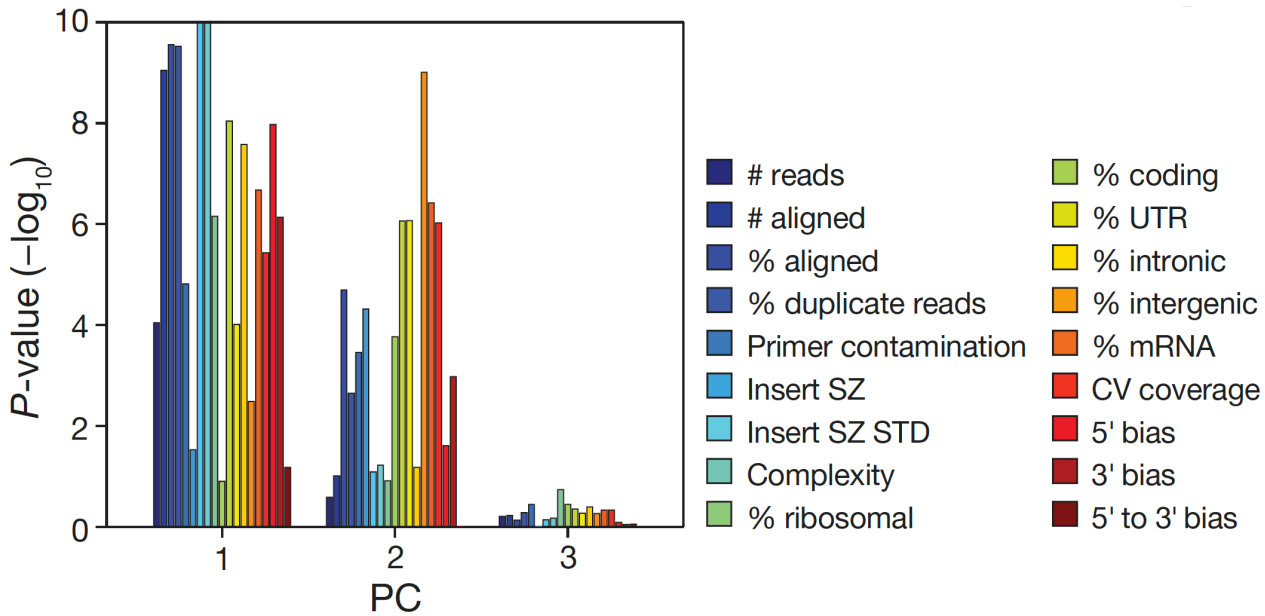
Read counts

	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			
Spike 1	103	180	
Spike 2	13	19	
...			

2- Observed gene expression is a convolution



3- Variation dominated by "technical factors



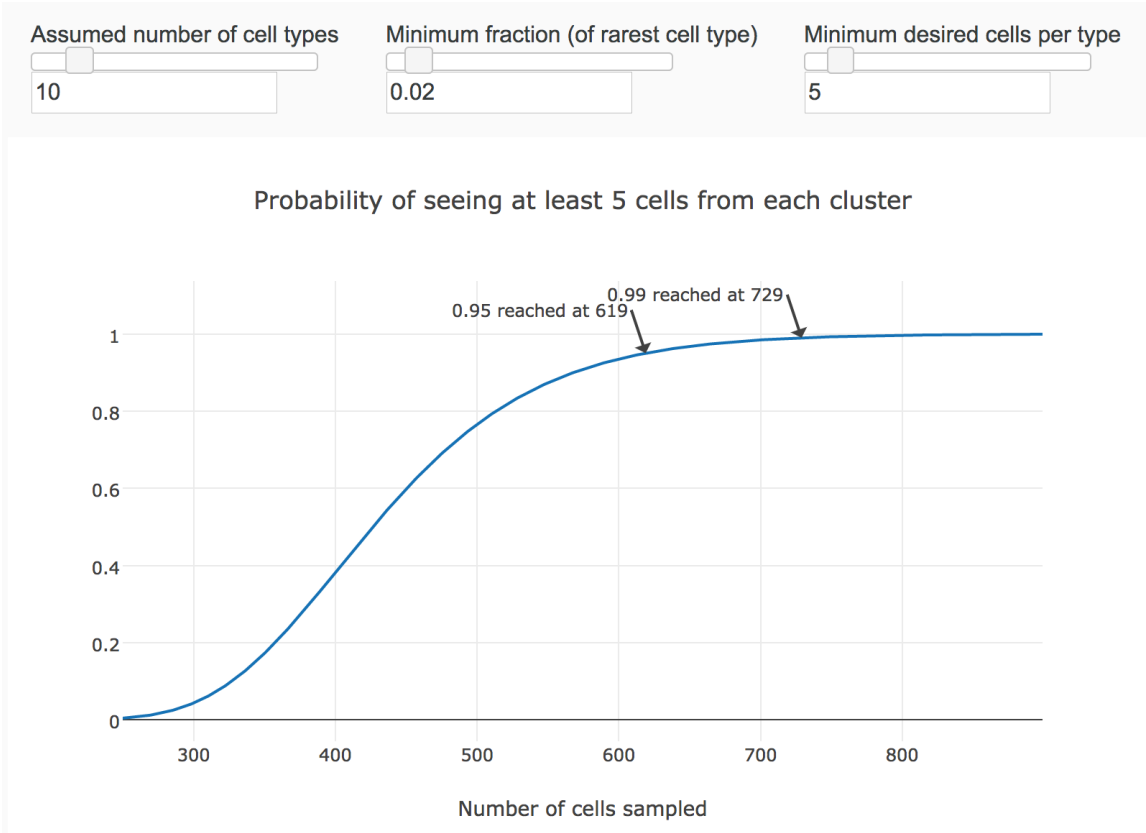
Buettner et al. 2015
Wagner et al. 2017

Experimental design & common questions

How many cells should I be profiling?

- Can change depending on the variability of the biology and the expectation of finding rare populations.

- Satija lab online tool – satijalab.org/howmanycells



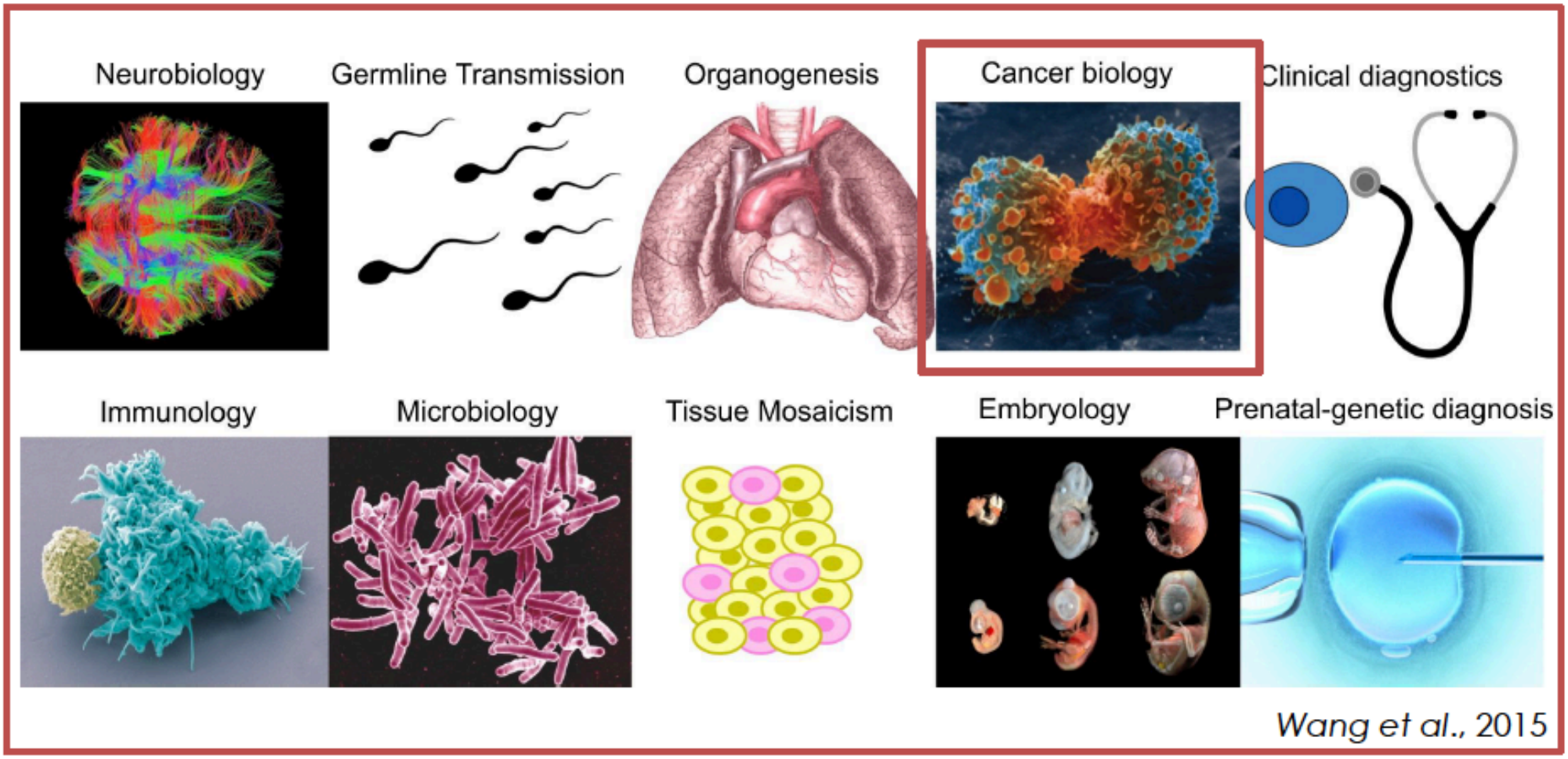
Cell number & Read depth

- For initial pilot study → aim for around 25-30 cells from each type
  - Sample with minor cell types < 5% will require sequencing at least 400 cells
  - Cell preselection/enrichment may be necessary, but unbiased cell selection is always preferred

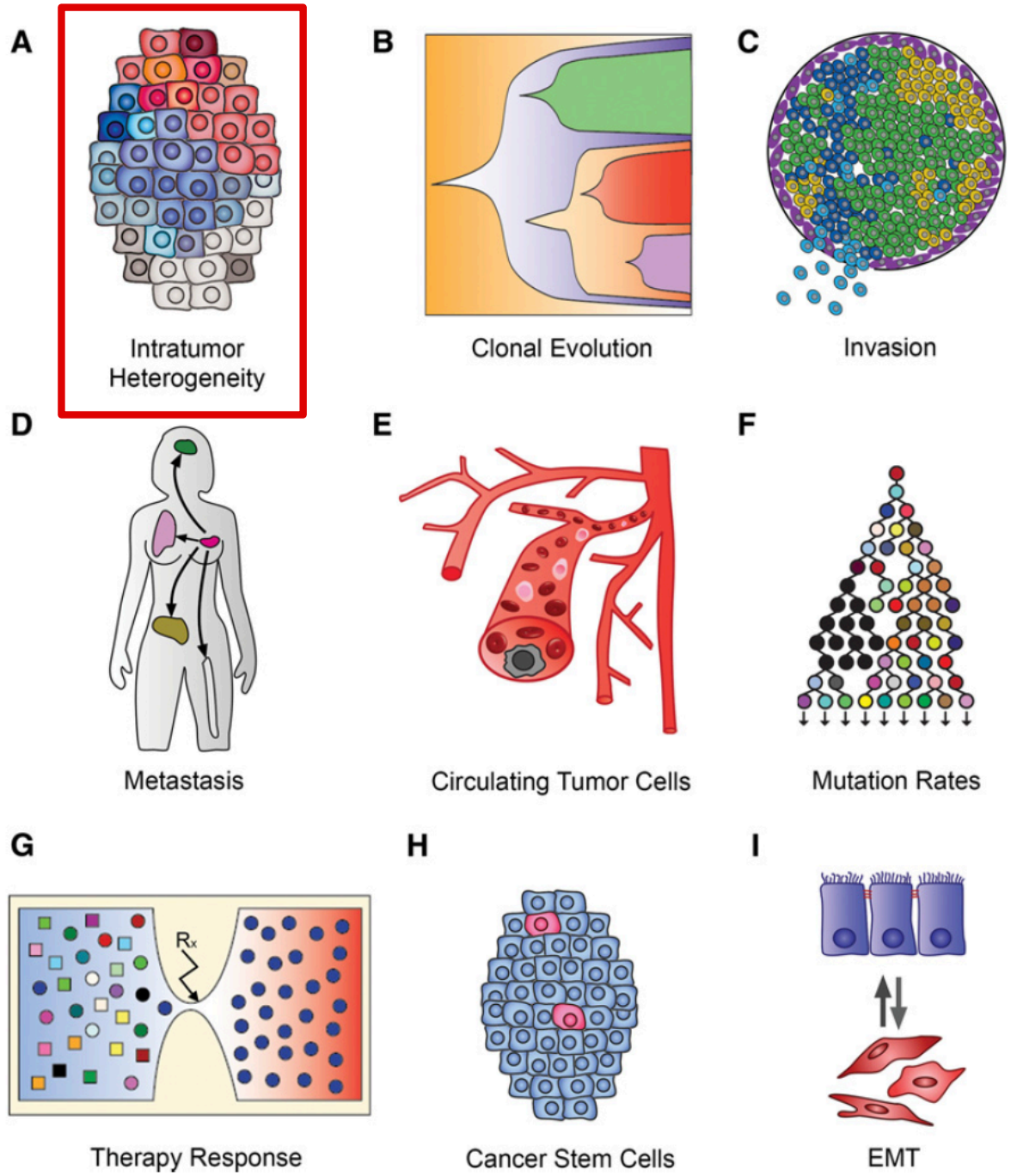
- To study gene expression only, sequencing depth doesn't always have to be deep (depends on questions)
  - Multiplexing hundreds of samples on one sequencing lane is common
  - Cell clustering & cell-type identification benefits from large number of cells and doesn't always require high sequencing depth (~100,000 reads per cell)
  - Gene detection starts saturating from 1 million reads per cell
  - Transcription factor detection (regulatory networks) require higher read depth and most sensitive protocols

Applications

Applications – Cancer biology



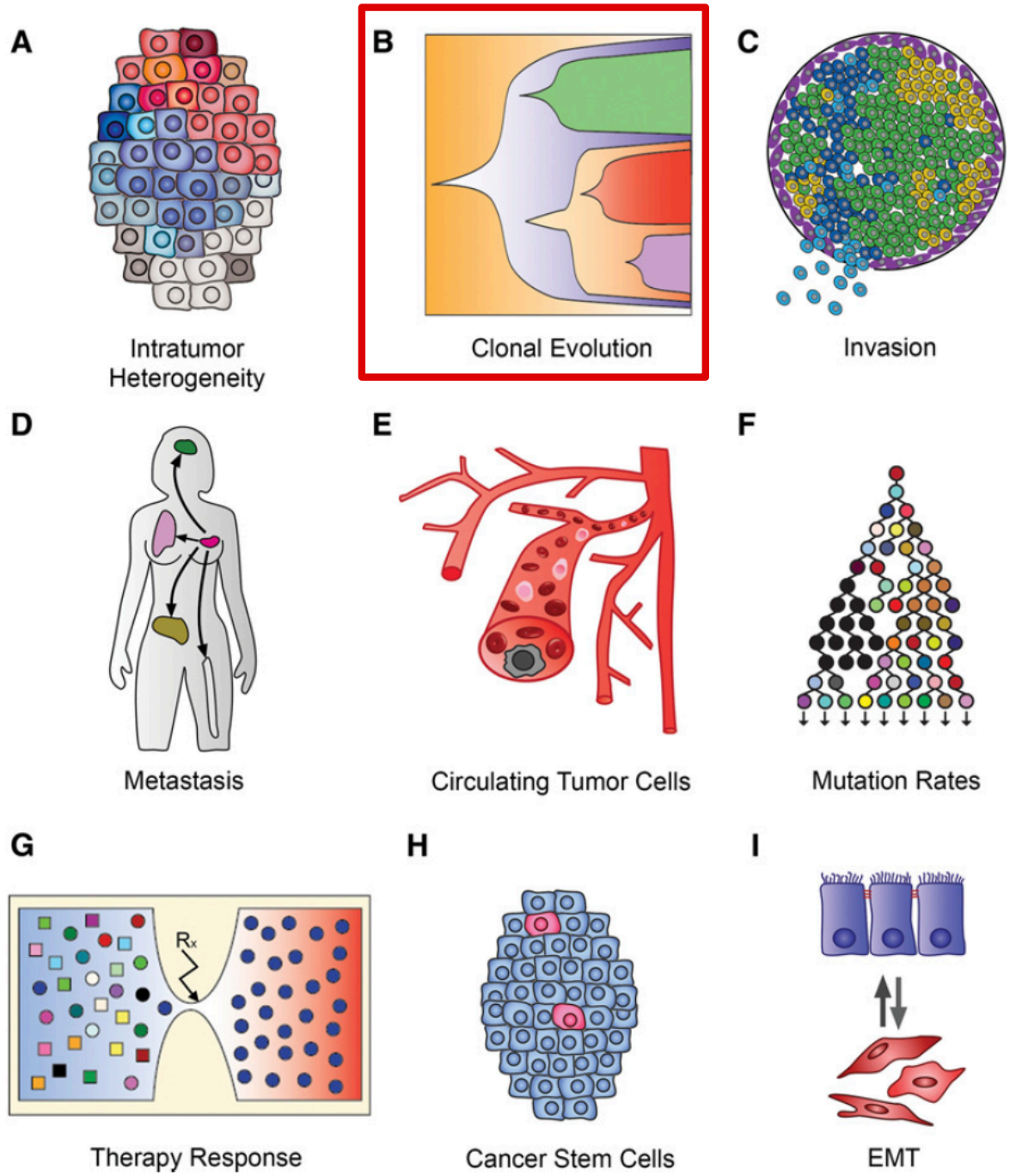
A- Resolving intratumoral heterogeneity & dissecting microenvironment



Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

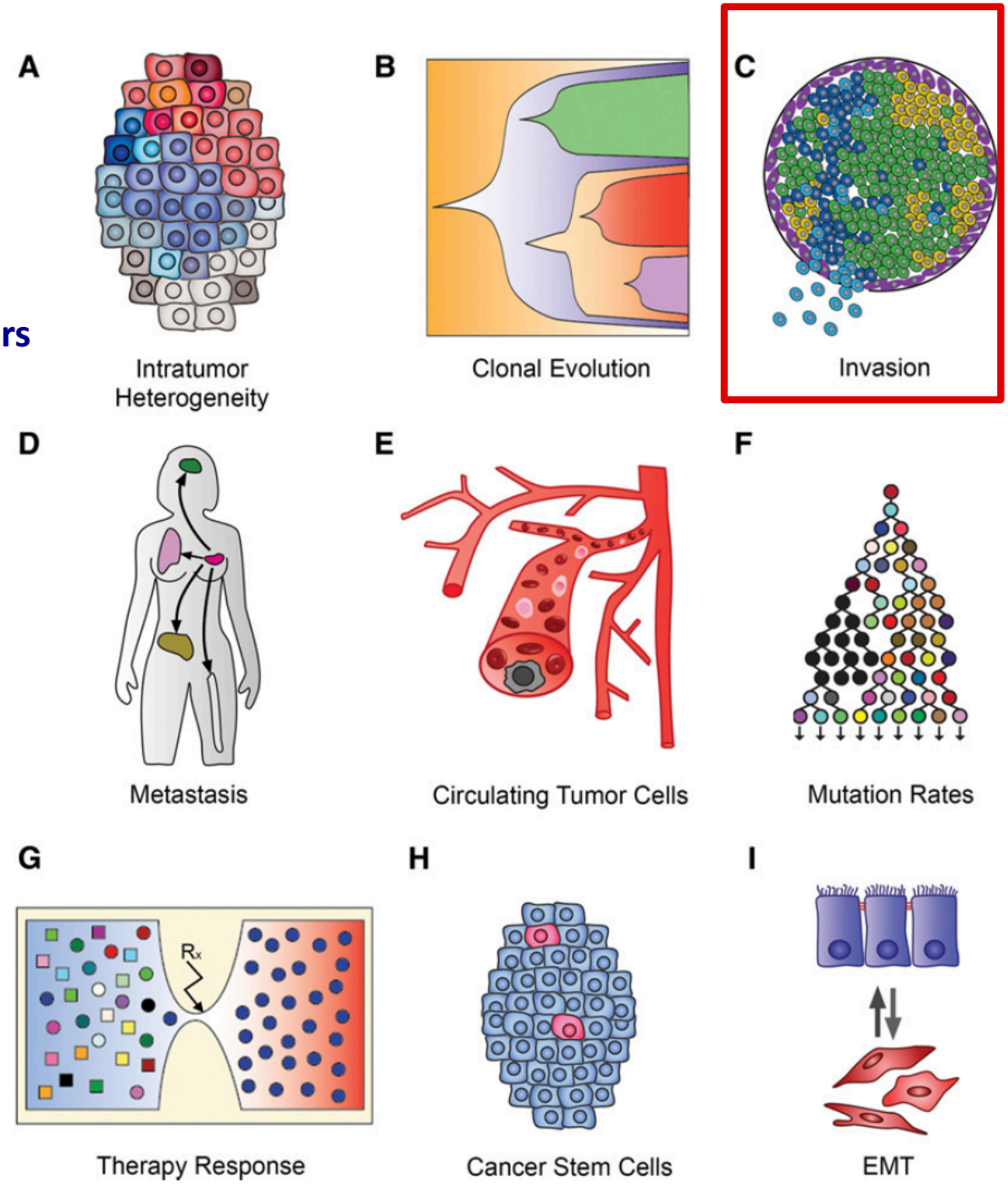


Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers



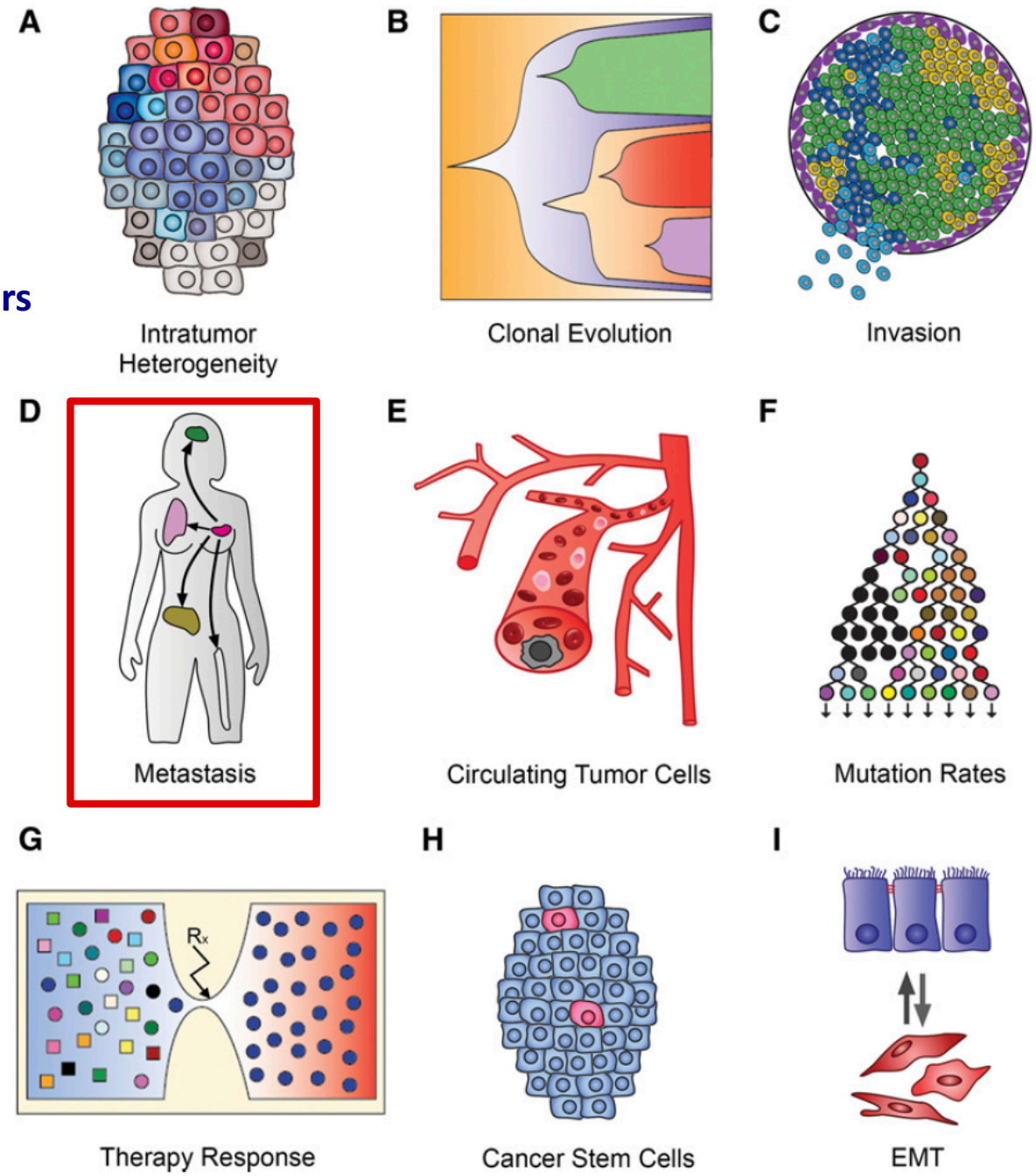
Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers

D- Tracking metastatic dissemination



Navin, 2015

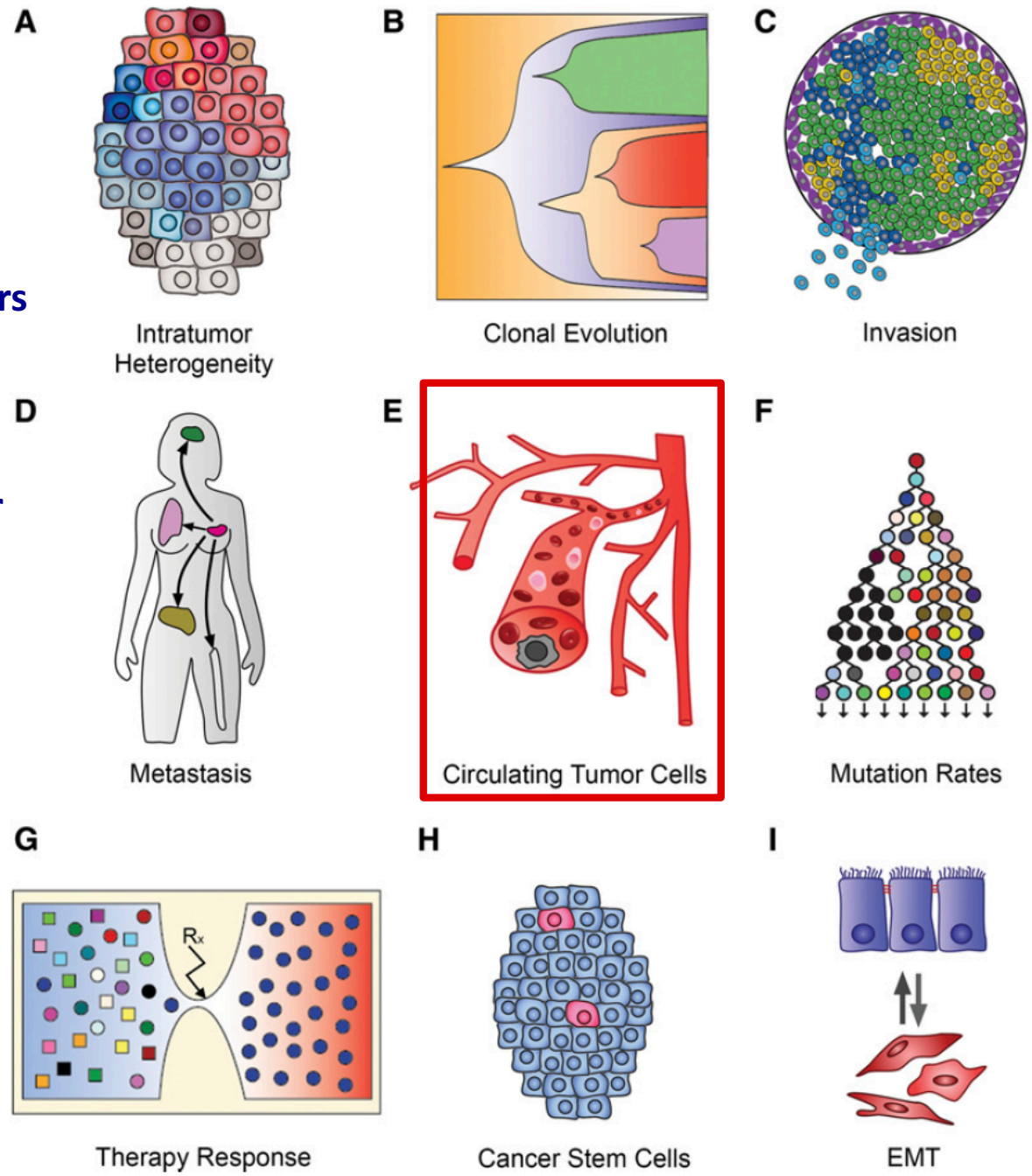
A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers

D- Tracking metastatic dissemination

E- Genomic profiling of circulating tumor cells



Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

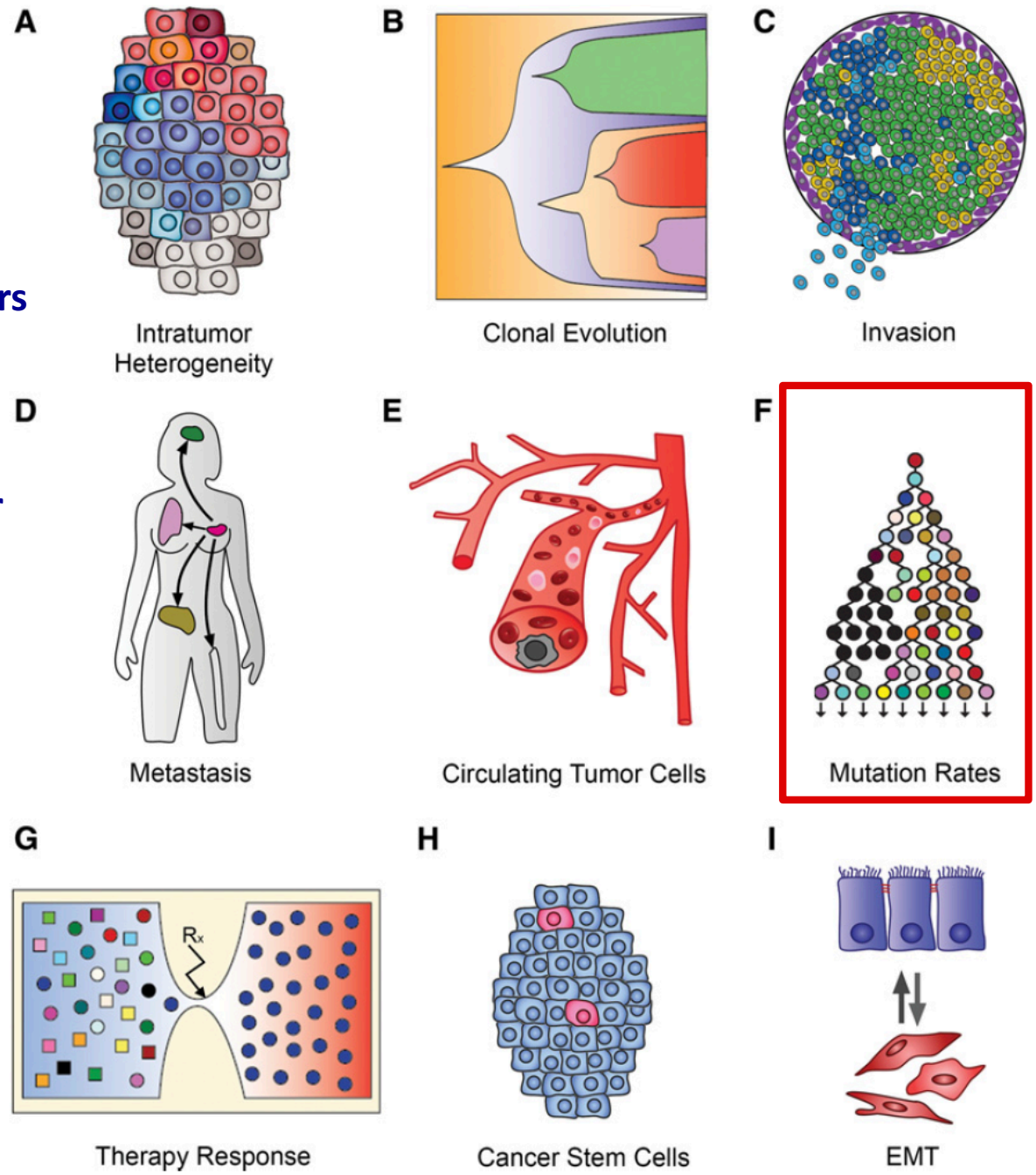
B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers

D- Tracking metastatic dissemination

E- Genomic profiling of circulating tumor cells

F- Studying mutation rare and mutated phenotypes



Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

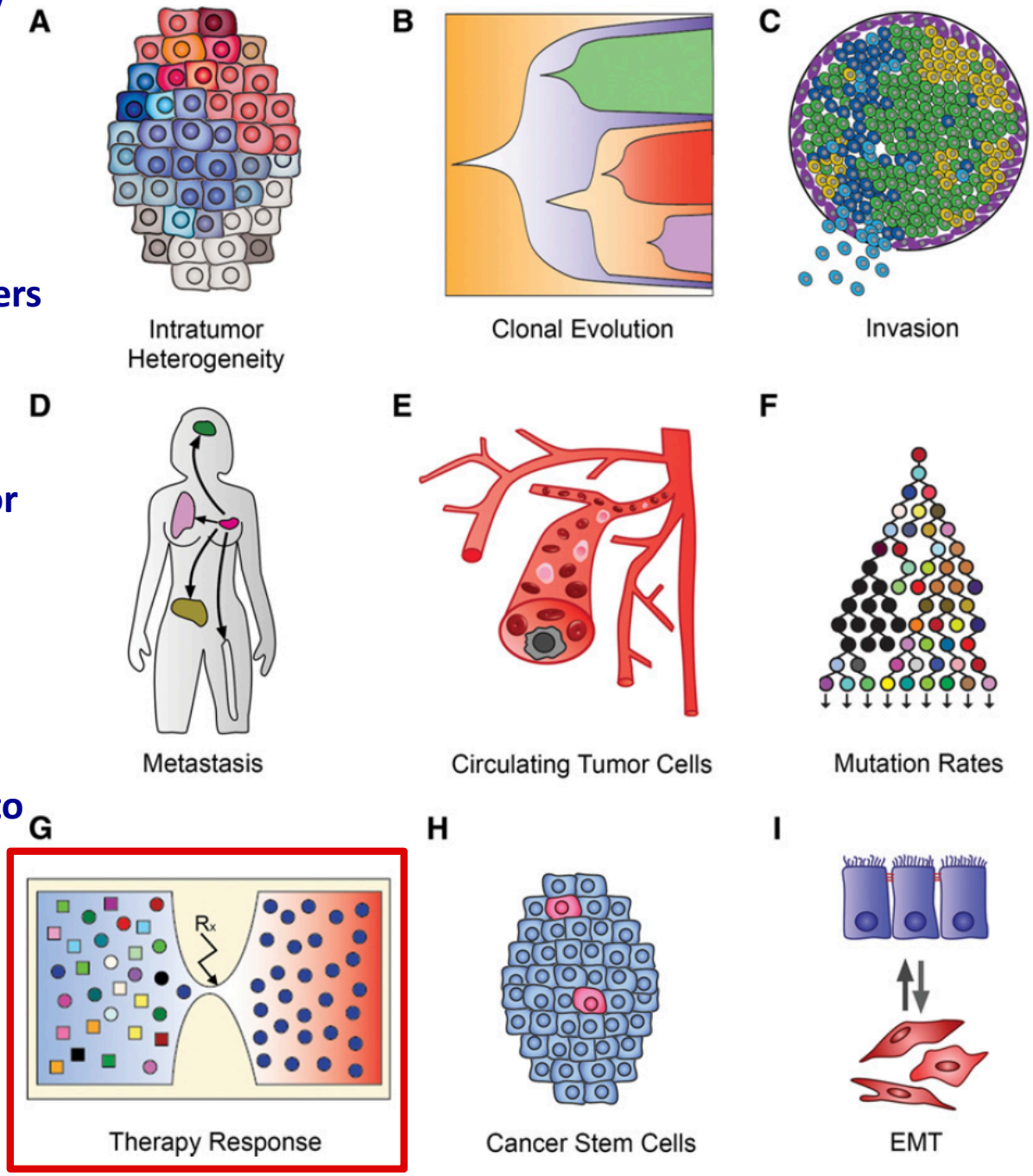
C- Studying invasion in early stage cancers

D- Tracking metastatic dissemination

E- Genomic profiling of circulating tumor cells

F- Studying mutation rare and mutated phenotypes

G- Understanding resistance evolution to therapy



Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers

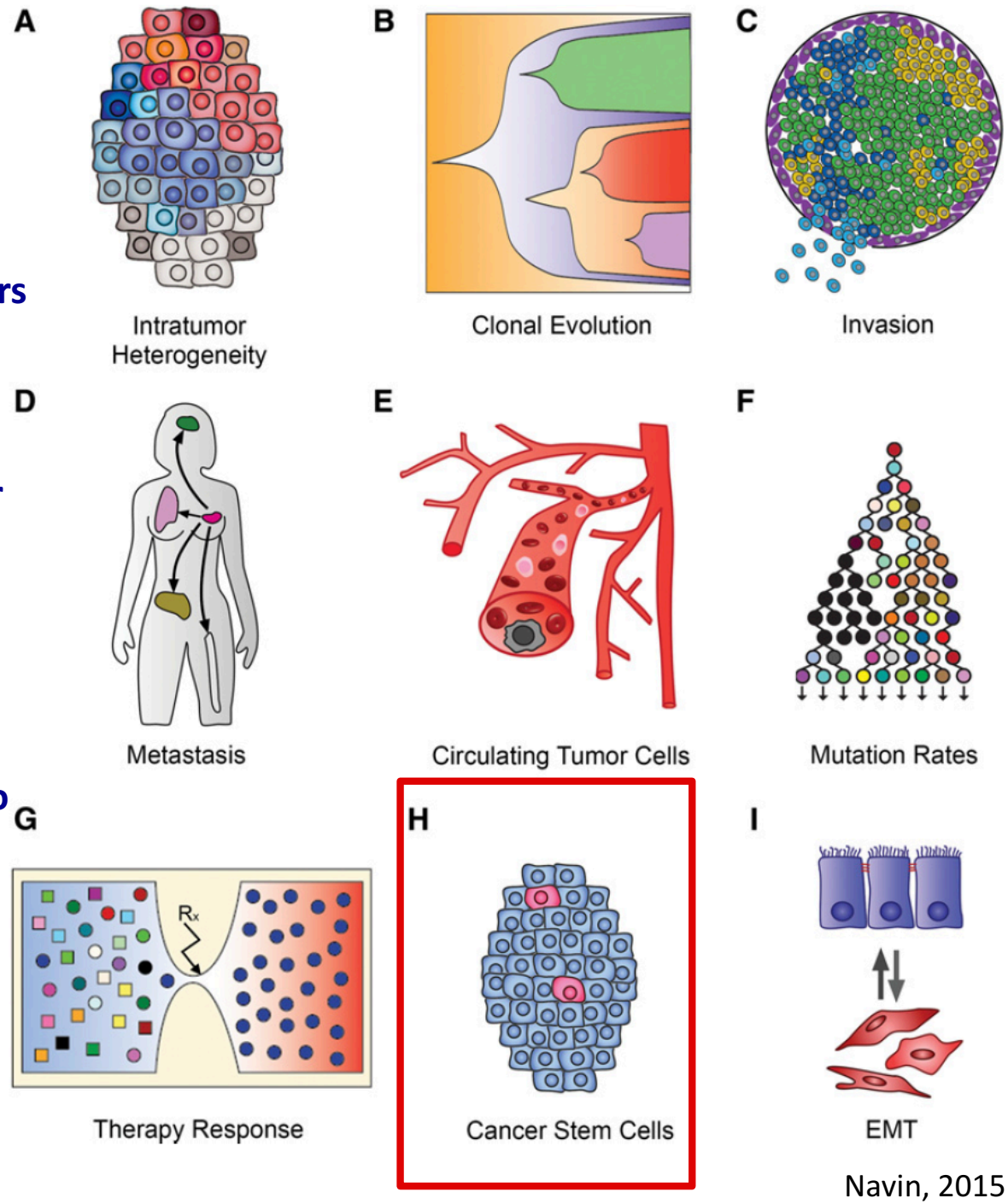
D- Tracking metastatic dissemination

E- Genomic profiling of circulating tumor cells

F- Studying mutation rare and mutated phenotypes

G- Understanding resistance evolution to therapy

H- Understanding cancer stem cell & cell hierarchies



Navin, 2015

A- Resolving intratumoral heterogeneity & dissecting microenvironment

B- Investigating clonal evolution in primary tumors

C- Studying invasion in early stage cancers

D- Tracking metastatic dissemination

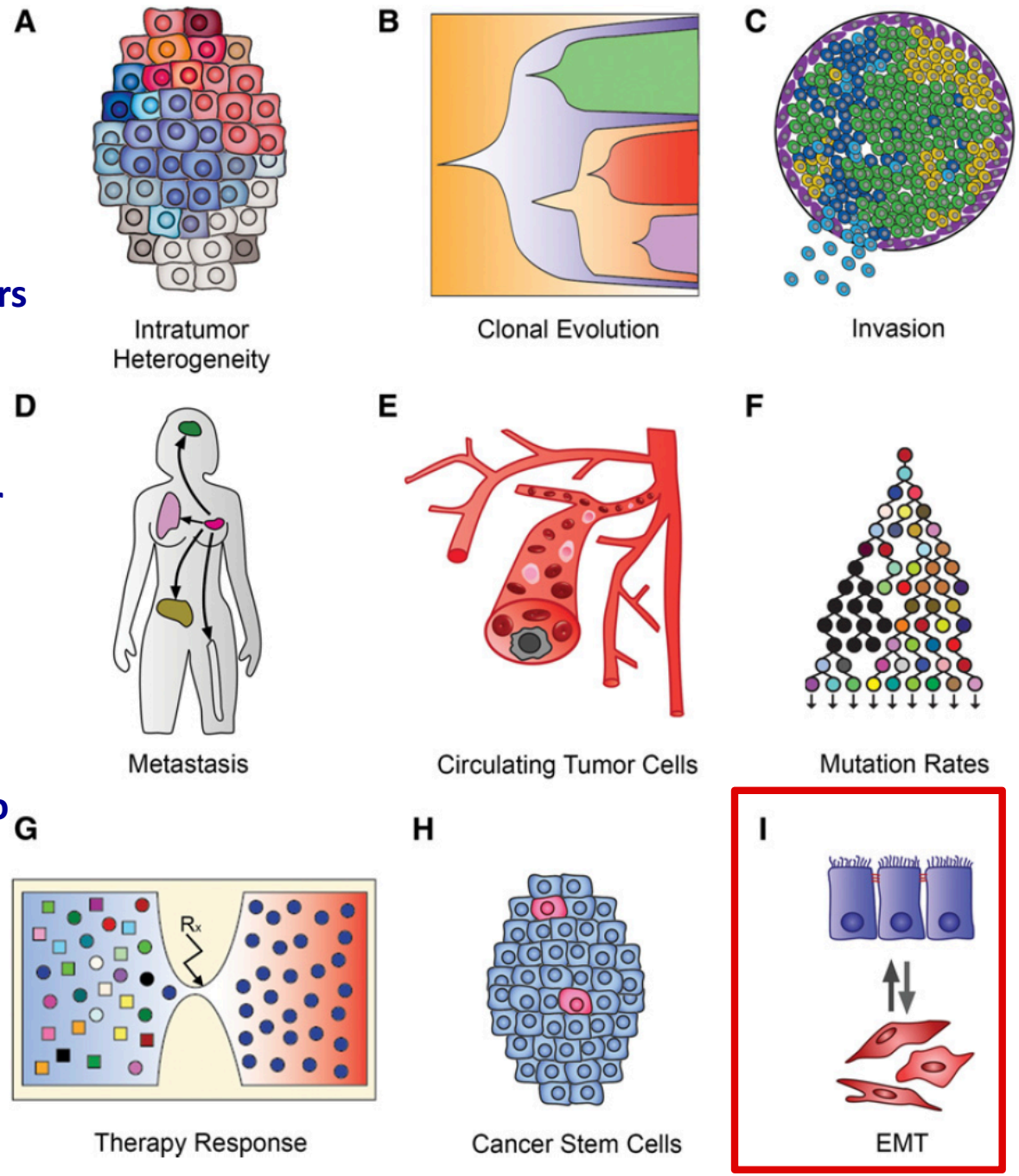
E- Genomic profiling of circulating tumor cells

F- Studying mutation rare and mutated phenotypes

G- Understanding resistance evolution to therapy

H- Understanding cancer stem cell & cell hierarchies

I- Studying cell plasticity and epithelial-to-mesenchymal transition



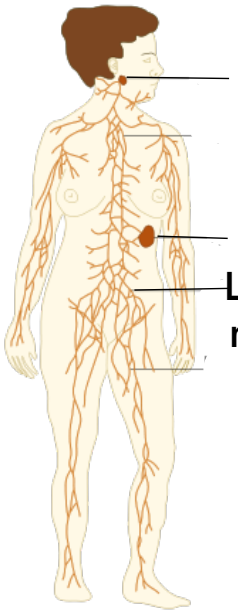
Navin, 2015

Strategies for census and validation

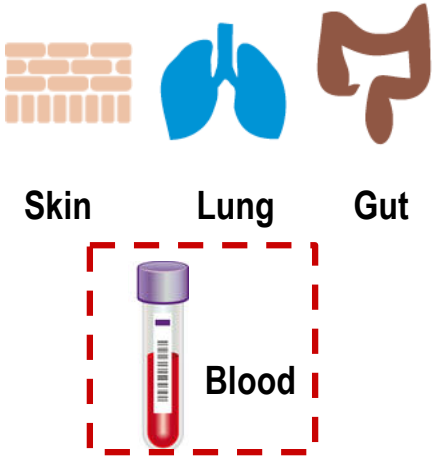
Phase I: Generating unbiased DC map

Healthy tissue to be profiled

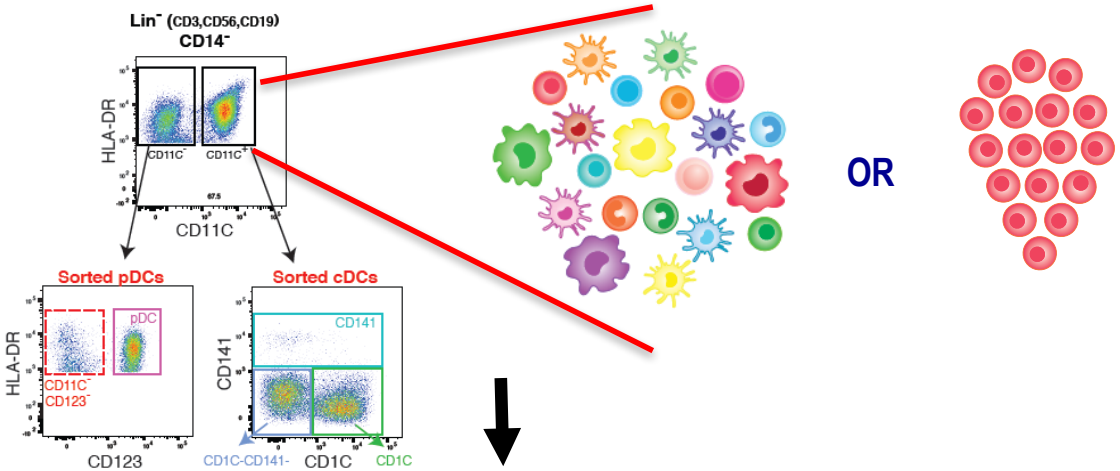
Lymphoid Organs



Non-Lymphoid "Barrier" Organs



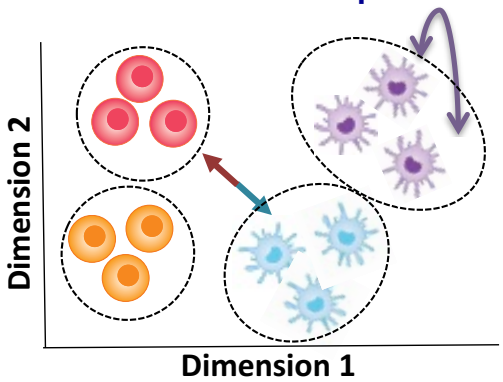
Sample dissociation, enrichment



Single cell RNA-Sequencing
→ Plate-based (e.g. SS2, Cel-Seq2, SCRB-Seq)
→ Droplet-based (e.g.10X, DropSeq, InDrop)

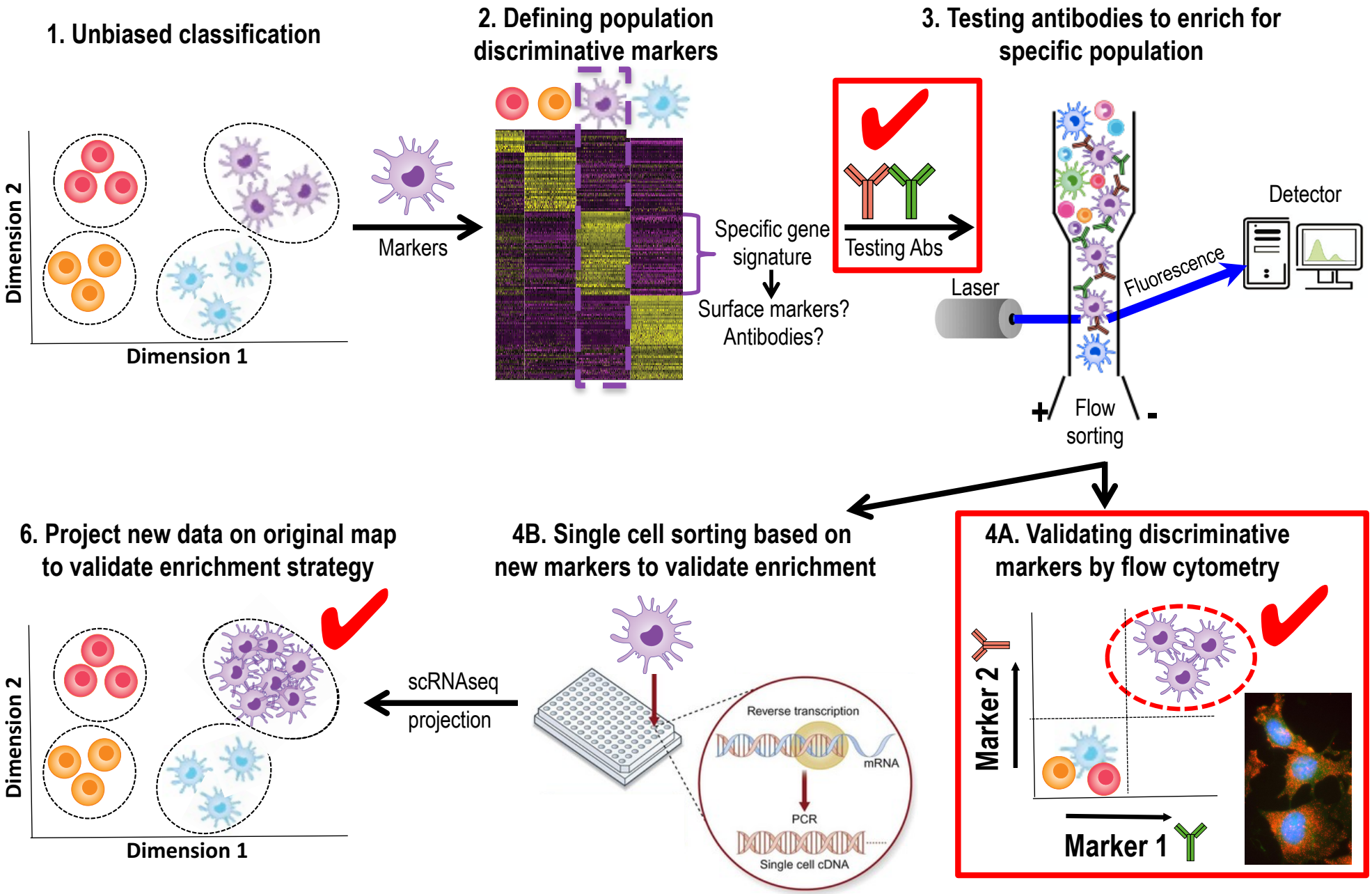
Expression profile clustering

Cell subsets map



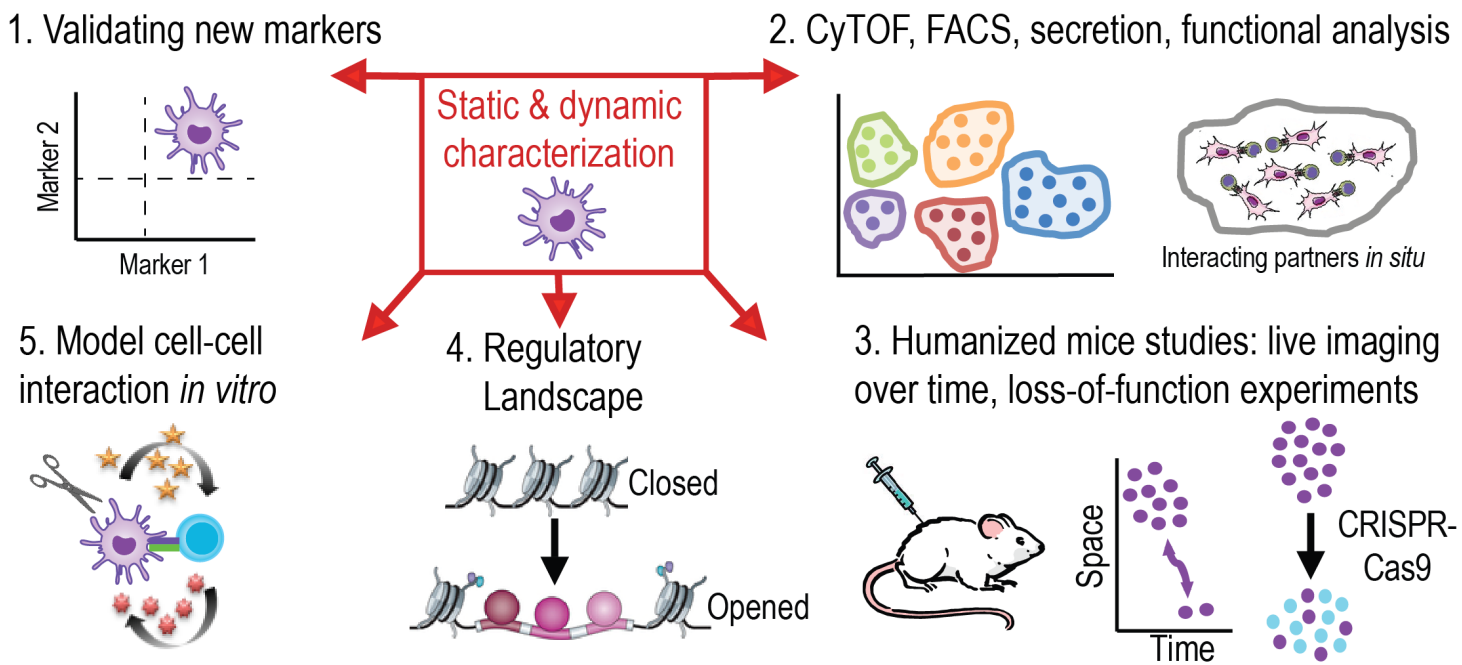
- ✓ Cell type identification
- ✓ Deconvolution of population structure
- ✓ Identification of markers
- ✓ Variability of transcription
- ✓ Regulatory network inference

Phase II: Enriching for new predicted cell populations – developing & validating reagents and isolations strategies

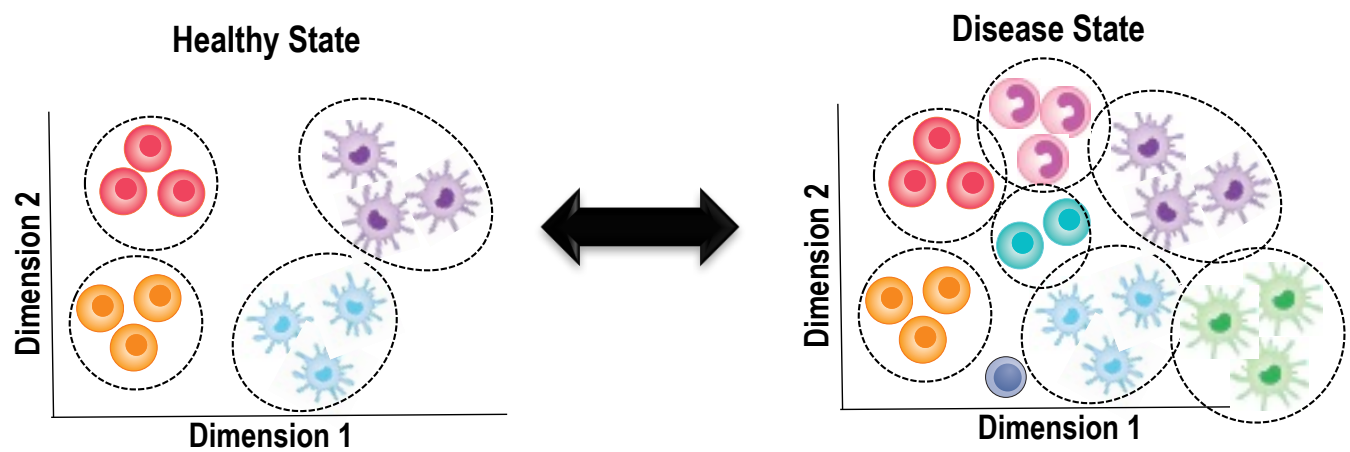


Phase III: Functionally defining uniqueness of predicted new cell population in health and disease

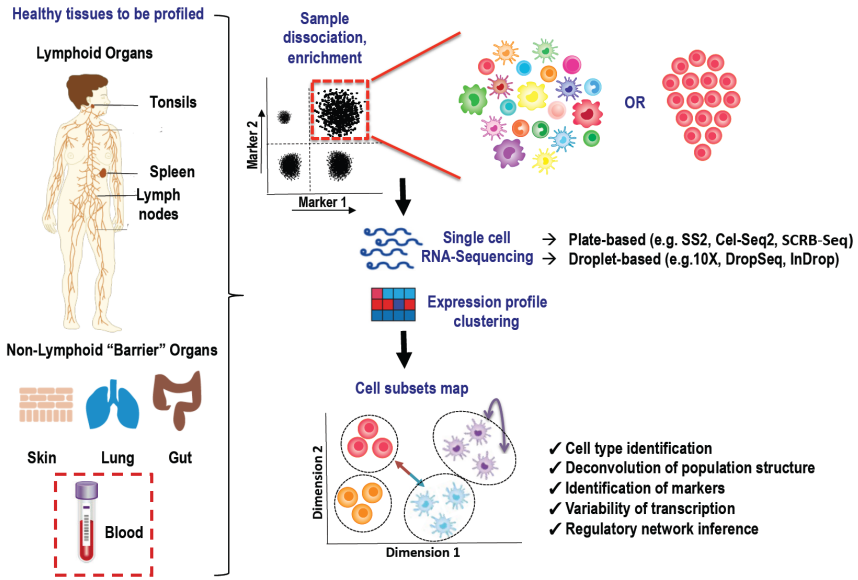
A- Functional Study & Characterization



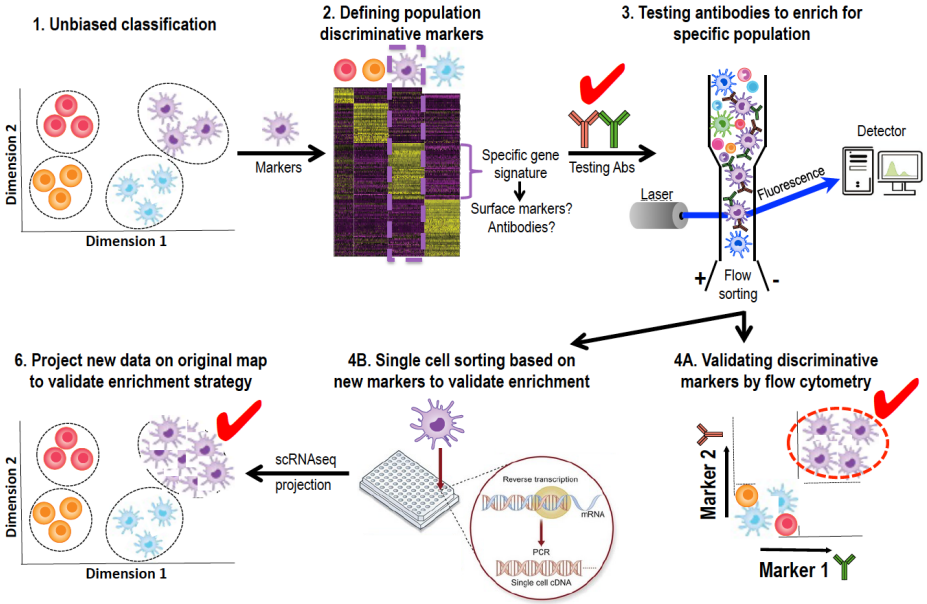
B- Mapping Disease



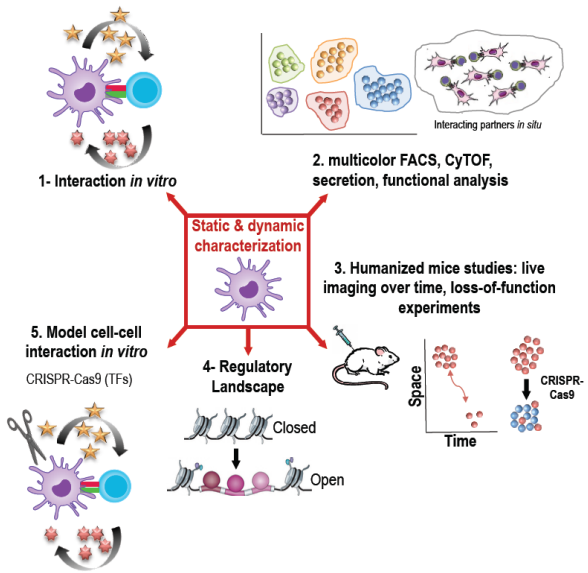
1. unbiased DC map



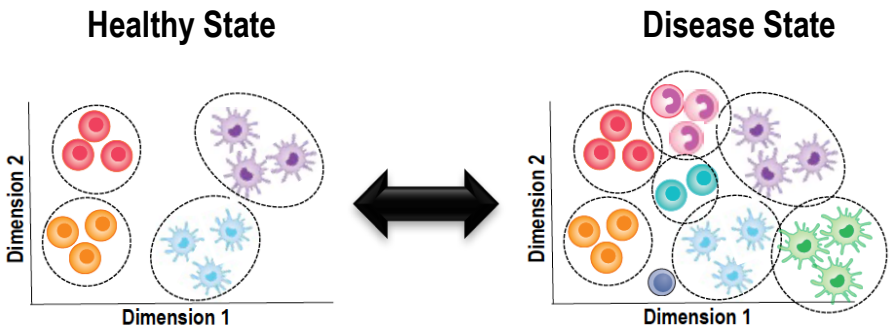
2. Identifying & validating new markers and gating strategies



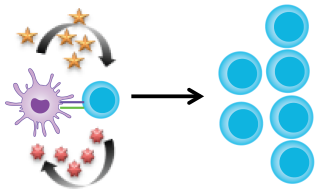
3. Functional characterization



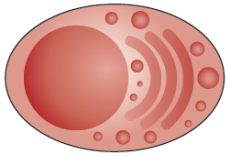
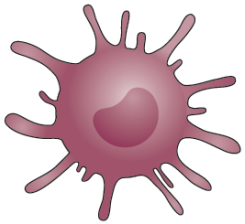
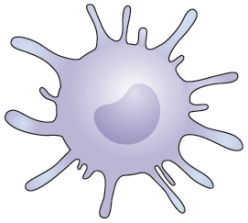
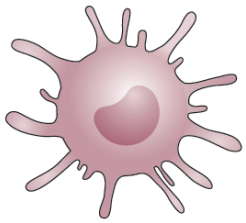
4. Mapping & studying disorders



Dendritic Cells (DCs) & Monocytes



- DCs ≈ 1-3% & monocytes ≈ 10-25% in blood
- DCs function in pathogen sensing, antigen presentation, T cell activation
- Monocytes role in phagocytosis, cytokine production, macrophage source
- Involved is several auto-immune diseases & cancers; therapeutic target
- Several subtypes have been defined:

pDC	CD141^{+}	CD1c^{+}	CD1c^{-} CD141^{-}	CD14^{+} CD16^{lo}	CD14^{+} CD16^{+}	CD16^{+} CD14^{lo}
						
15-20% of DC	3-5% of DC	19-25% of DC	50-70% of DC	75-80% of mono	2-5% of mono	10-15% of mono
Interferon production	Antigen presentation to CD8^{+} T cells	Inflammation: Ag presentation to CD4^{+} T cells				

Answering key questions to discover & characterize all blood dendritic cell (DC) & monocyte subsets

1) How many subsets can be found in blood?
2) Do they have the expected markers?
3) Can we identify better markers?
4) Is there heterogeneity within the major subsets?
5) Are there previously uncharacterized subsets?
6) Can these subsets be used to map cells in human disease?

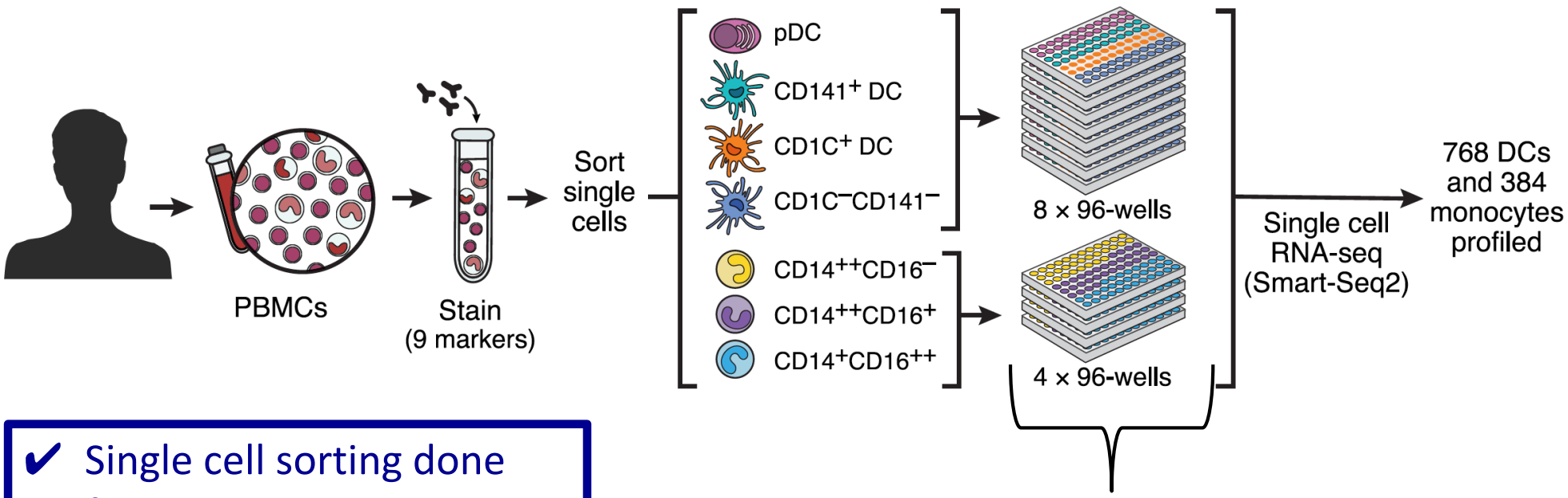
Villani *et al. Science* 2017 Apr 21;356(6335).

Answering key questions to discover & characterize all blood dendritic cell (DC) & monocyte subsets

1) How many subsets can be found in blood?
2) Do they have the expected markers?
3) Can we identify better markers?
4) Is there heterogeneity within the major subsets?
5) **Are there previously uncharacterized subsets?**
6) Can these subsets be used to map cells in human disease?

Villani *et al. Science* 2017 Apr 21;356(6335).

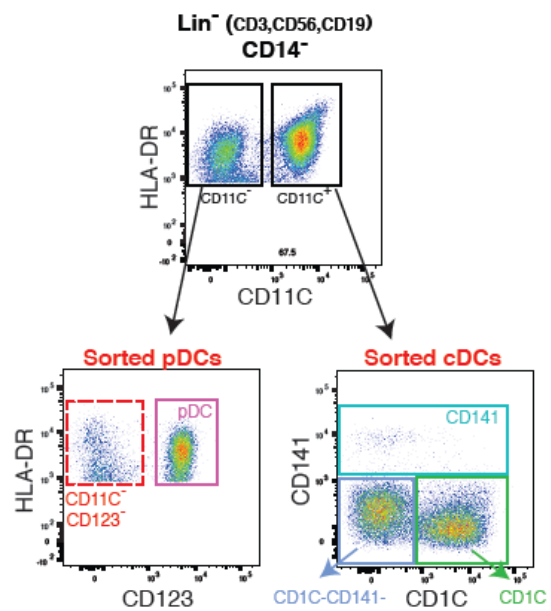
How should we discover DC subsets?



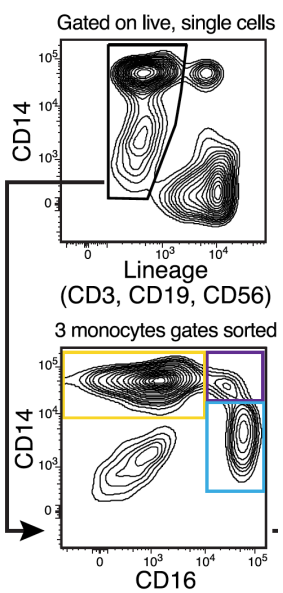
- ✔ Single cell sorting done from a constant source
- ✔ Cell sorting with optimized panel of markers
- ✔ Deep Sequencing (1-2M reads/cells)

Adaptive sampling strategy:

DCs:

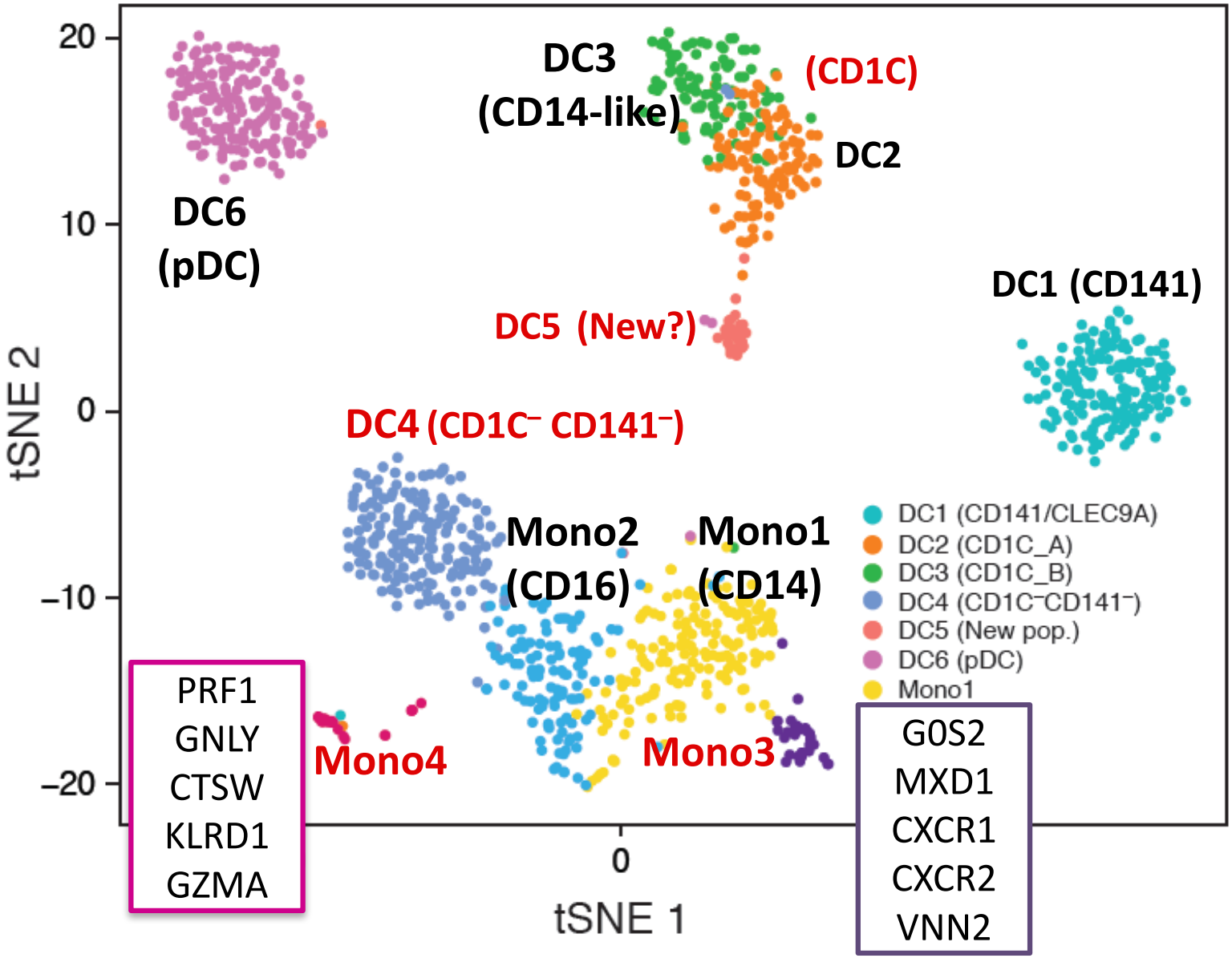


Monocytes:



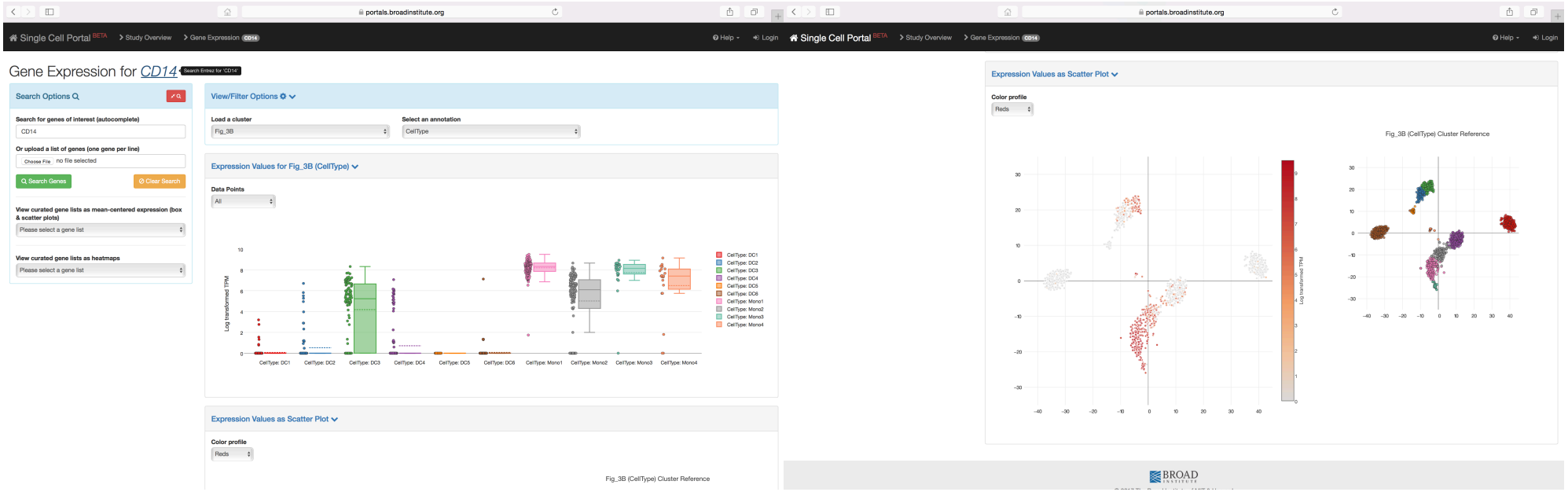
How many subsets can be found in blood?

Observed 6 DC & 4 Monocyte clusters in blood



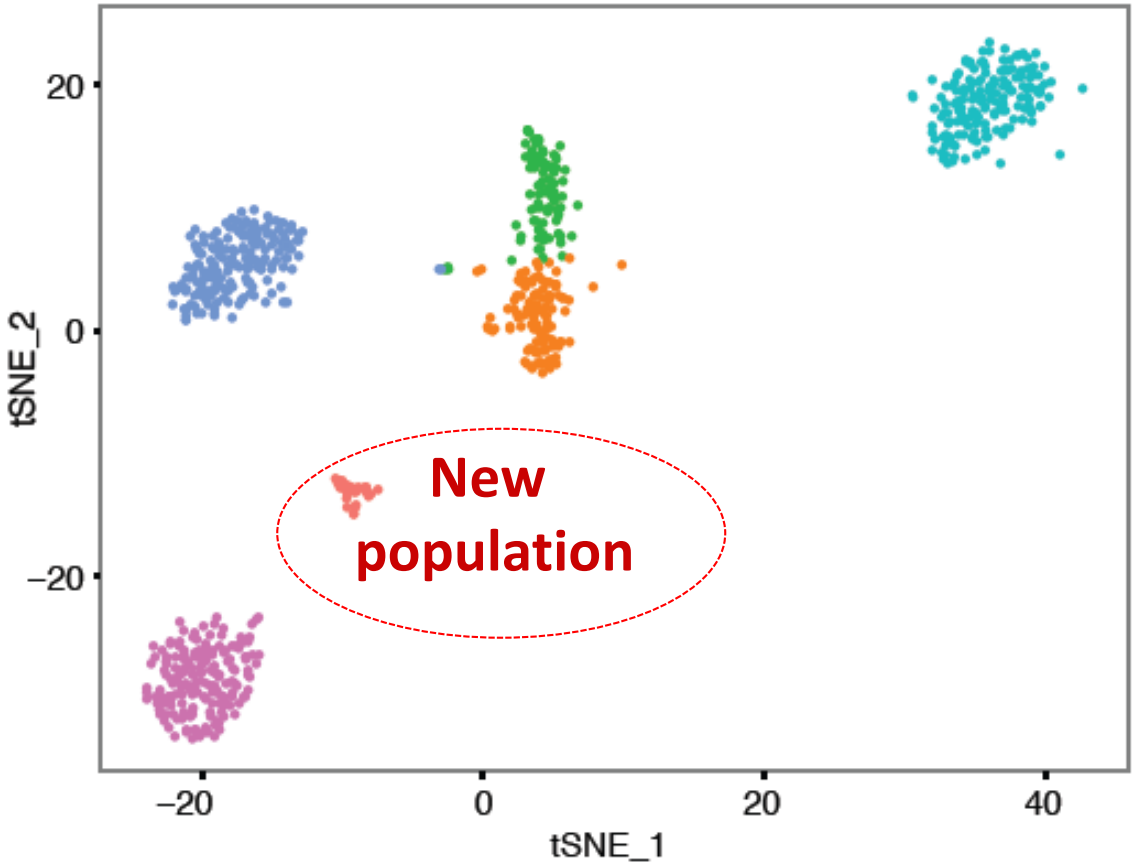
Data uploaded to single cell portal

https://portals.broadinstitute.org/single_cell/study/atlas-of-human-blood-dendritic-cells-and-monocytes



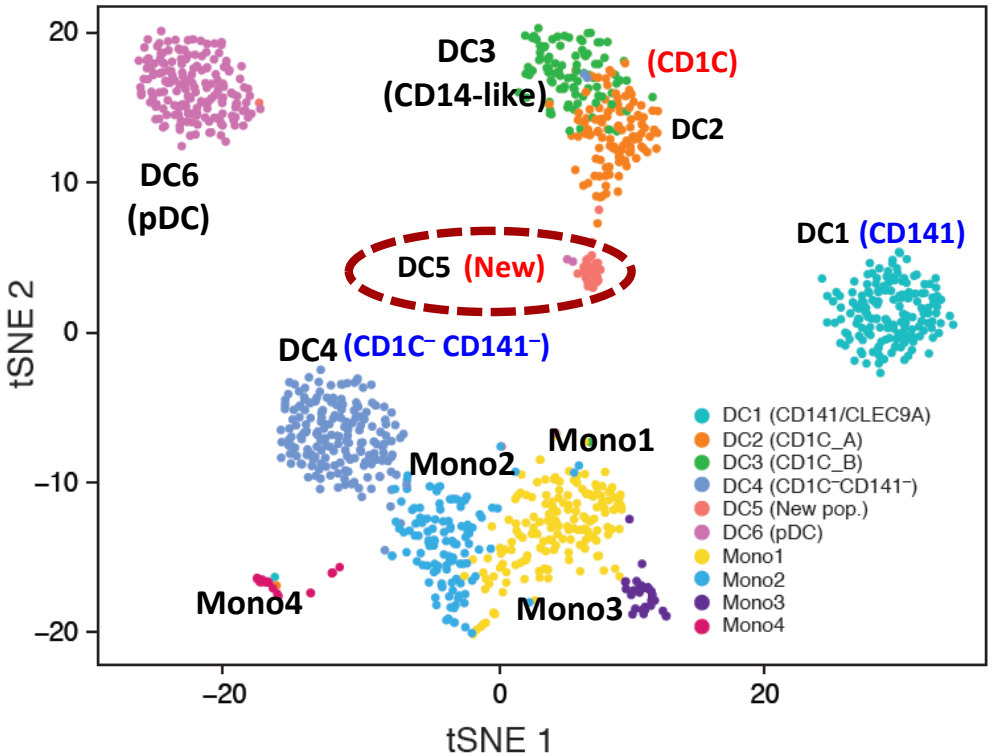
Tim Tickle

What is the uncharacterized DC subset?

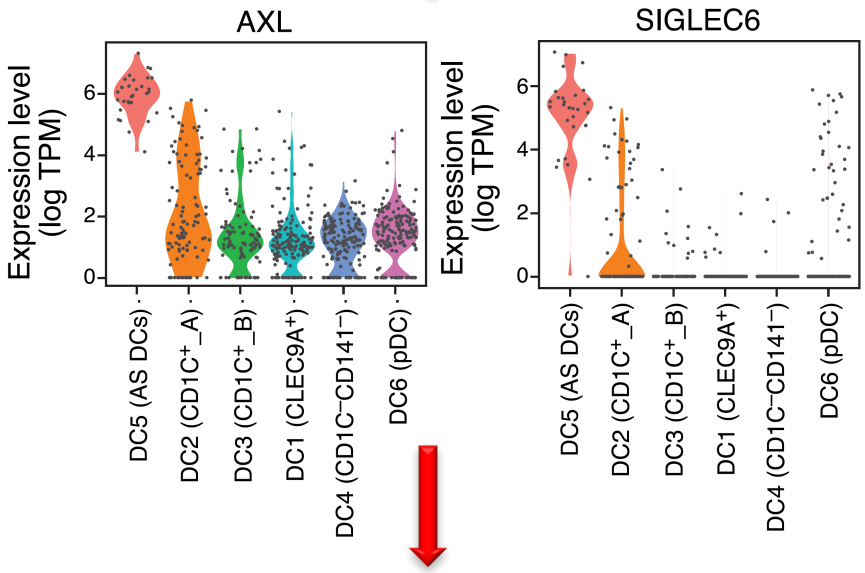


Automatic multi-dimensional classifier predicts the presence of rare new DC subset

DC5 Challenge: rare (≈ 0.06% of PBMCs)



Unique signature



New sorting panel

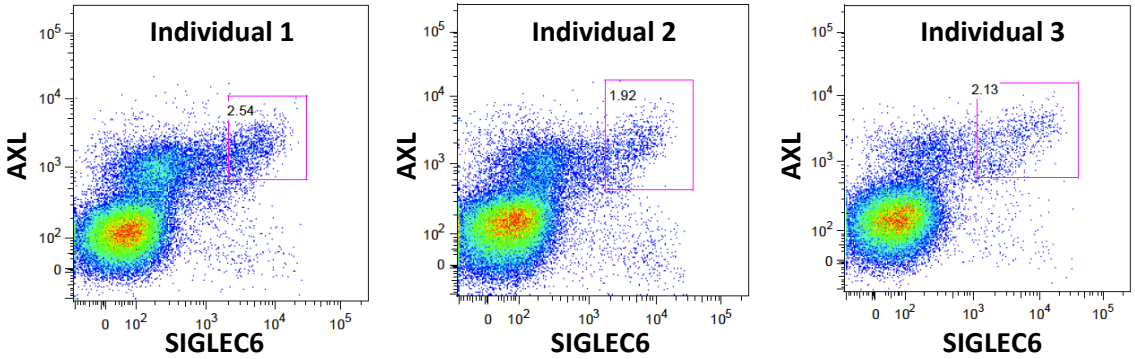
Approach:

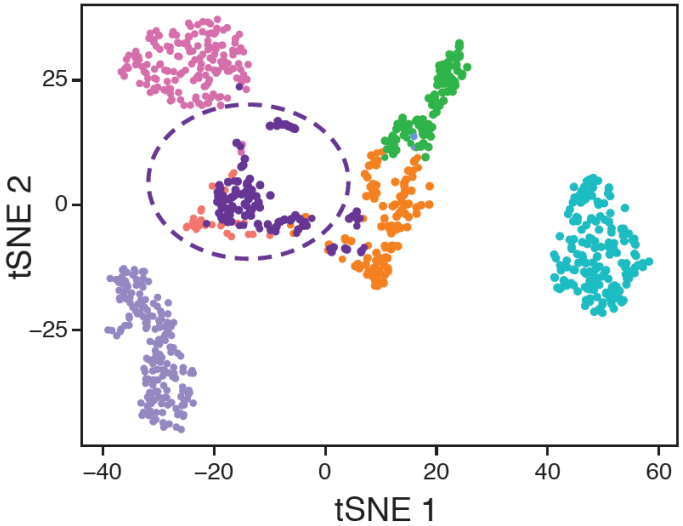
(1) Find markers specific for population

(2) Develop new sorting panel

(3) Profile cells from additional individuals

Validation of DC5 population existence by flow cytometry & scRNAseq of prospectively isolated cells





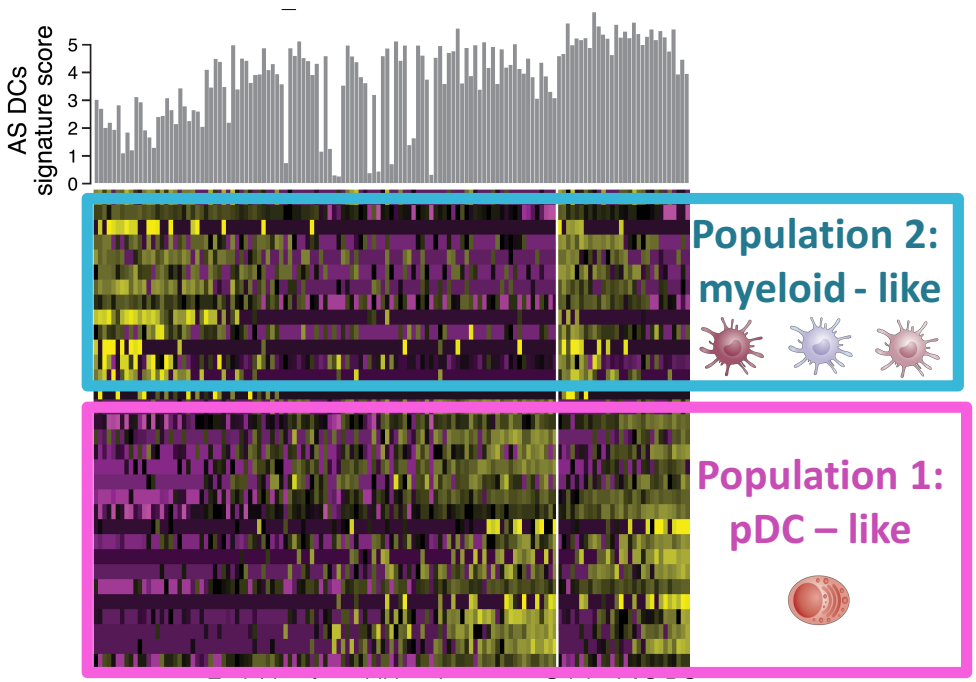
→ New DC population observed in **ALL 10 patients**

→ **Represents 1.9-3.2% of the DC / 0.04-0.064% of PBMCs**

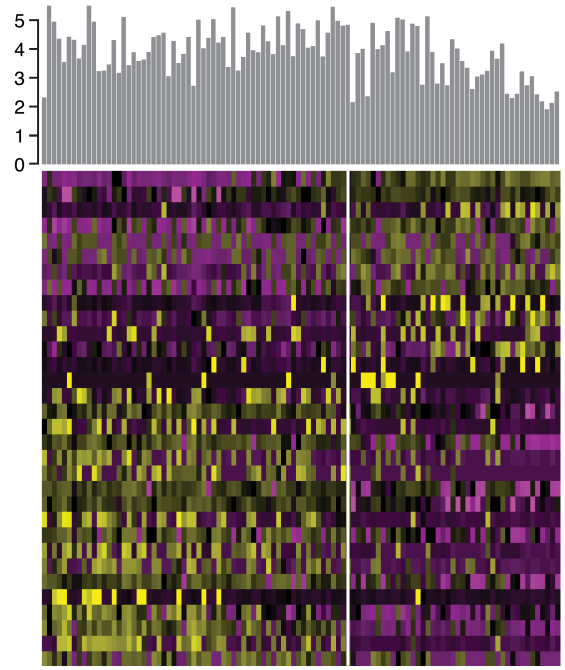
- **What do they look like?**
  - Transcriptionally – what's distinct and common
  - Morphology
- **What are its communication capabilities with other cells**
  - Receptors, secreted factors
  - Co-culture with other cell types
- **Who are the direct interacting partners**
  - *In-situ* co-localization staining
- **Where are these cells in DC gating strategy?**

DC5 new population falls along continuum with 2 clear extremes
Successful enrichment of both subsets & validation in 10 healthy individuals

DC5 falls across 2 extremes



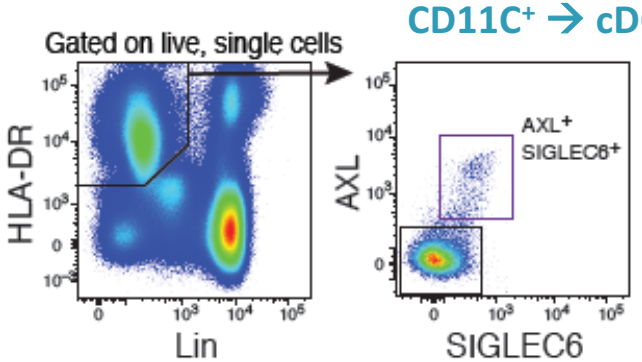
B- Validation of 2 putative subsets



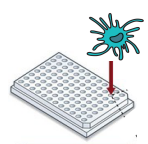
A- NEW panel to enrich for both putative populations

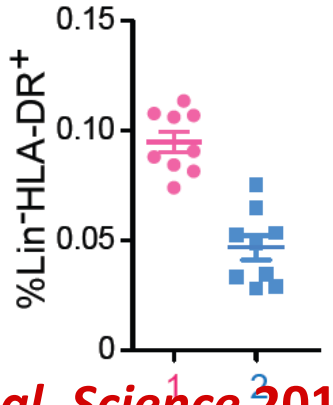
CD123⁺ → pDC

CD11C⁺ → cDC



C- Validation across 10 individuals





- → Pop. 1 & 2 share unique signature
- → Pop. 1 shares signature with pDC
- → Pop. 2 shares signature with cDC

Villani *et al. Science* 2017 Apr 21;356(6335).

Concluding thoughts

- Single cell genomics methods are becoming an essential tool for dissecting biology at an unprecedented resolution
- Single cell multi-omics will empower new definition of cell types/states and tissue
- Being able to track live cells over time will be truly transformative
- Scale will continue growing and price will come down →empowering translational efforts!
- New analyses techniques and framework are needed to handle such large dataset

A Word of Caution

"**Tempering some of the enthusiasm are myriad challenges inherent to the process**, from the isolation of cells, to amplification of their genomes or transcriptomes, to making sense of the data. Cost is also a consideration leaving good reason to carefully select situations that justify going to the single-cell level."

Bottom Line:

Single cell transcriptomics is not the solution to answering every biological question!

Kelly Rae Chi. *Nat Methods* 2014: 13-7

Acknowledgements

Rahul Satija

Nir Hacohen

Aviv Regev

Hacohen Group
Siranush Sarkizova
Weibo Li
All lab members

Christophe Benoist
David Puyraimond-Zemmour

Philip De Jager
Alina Von Korff
Laura Glick

Muzlifah Haniffa
Gary Reynolds
James Fletcher
Laura Jardine
Andrew Filby

Ragon Institute:
Marcus Altfeld
Morgane Griesbeck
Ryan Park
Michael Waring
Adam Chicoine

Regev Group
Karthik Shekhar
Orit Rozenblatt-Rosen

Dana-Farber Cancer Institute
Andrew A. Lane
Suzan Lazo-Kallanian

Olink Proteomics
Ida Grundberg
Emil Nilsson

KLARMAN CELL OBSERVATORY
AT BROAD INSTITUTE

Questions: cvillani@broadinstitute.org

Harvard Medical School
MGH Single Cell Genomics Research Program

Villani Lab: postdoc positions available

Contact: cvillani@broadinstitute.org