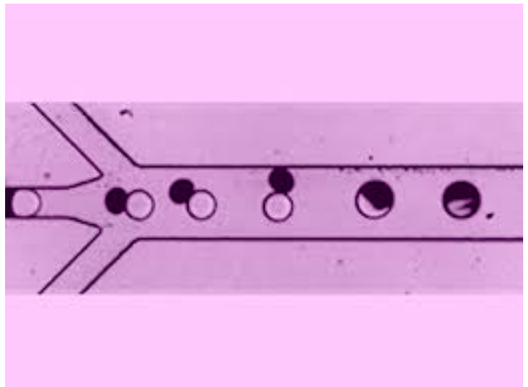
Choosing a single cell technology

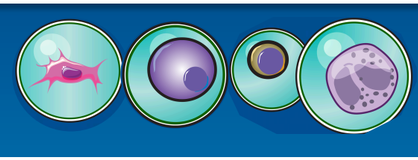
Choosing a single cell technology



Drop-seq

Outline

- **Key parameters that define technologies**
  - Cost, scale, data quality (sensitivity)
  - Input requirements, potential biases
- **Brief overview of technologies and tradeoffs**
  - SMART-Seq, CelSeq, Fluidigm, Droplet-based methods
- **Impossibly hard questions**:
  - How many cells do you need to sequence to make I've discovered everything?
  - How deep do I need to sequence per cell?
  - Which technology is the best one for all possible experiments?
  - Should I always sequence more cells at high depth or fewer cells at low depth?

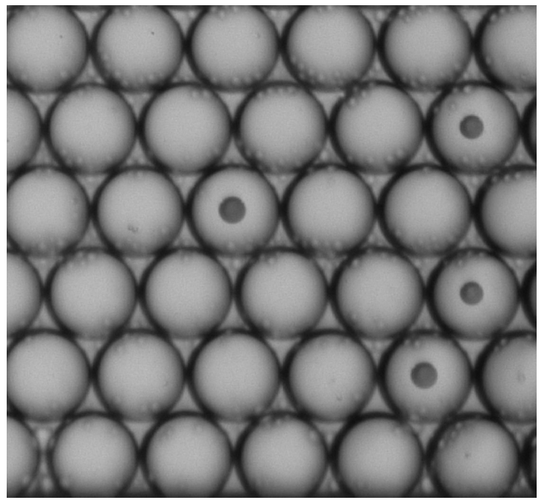


$$$$$$

Extreme A

Fluidigm C1 : 96-cell chip
~$35.00/cell

Extreme B



DropSeq/inDrop
$0.05/cell

Middle ground

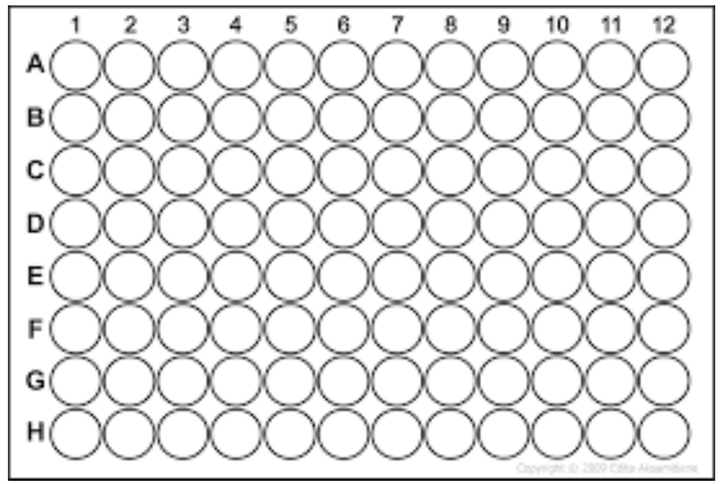


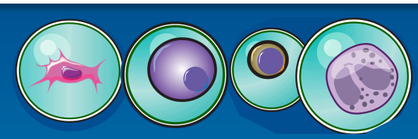
Plate based methods
$3-6/cell

Key considerations for cost

- **Pooled or individual library preparation?**
  - Pooled methods do not grow linearly in cost, but need up-front investment
  - No current protocols support pooled library prep and full-length transcripts

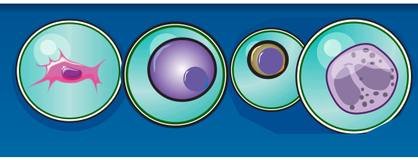
Key considerations for cost

- **Pooled or individual library preparation?**
  - Pooled methods do not grow linearly in cost, but need up-front investment
  - No current protocols support pooled library prep and full-length transcripts
- **Commercial or home-brew?**
  - ‘Do-it-yourself’ strategies exist for almost all approaches
  - Ease of use costs $$



Key considerations for cost

- **Pooled or individual library preparation?**
  - Pooled methods do not grow linearly in cost, but need up-front investment
  - No current protocols support pooled library prep and full-length transcripts
- **Commercial or home-brew?**
  - ‘Do-it-yourself’ strategies exist for almost all approaches
  - Ease of use costs $$
- **Which is squeezing your budget? Library prep or sequencing?**
  - Sequencing costs are user-defined, and can become overwhelming for large numbers of cells



Scale (how many cells per run?)

Extreme A

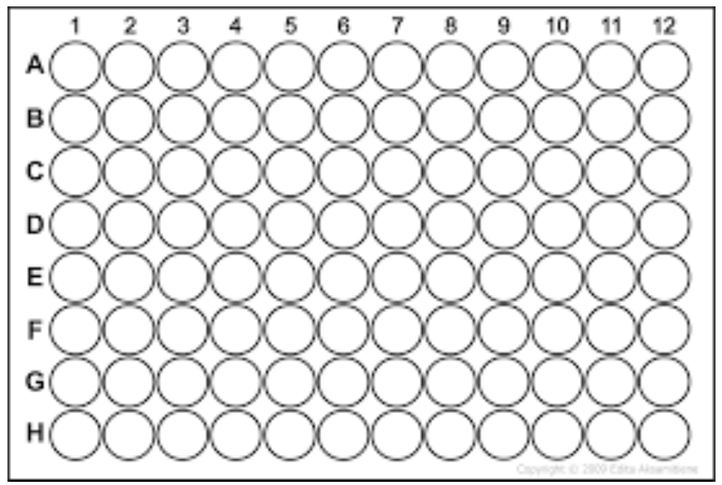


Plate-based methods
One cell at a time
Can be automated

Extreme B

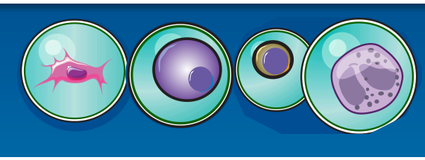
Chromium system (10x)
48,000 cells/run

Middle ground

Fluidigm C1 (96 well chip)
48-96 cells/run

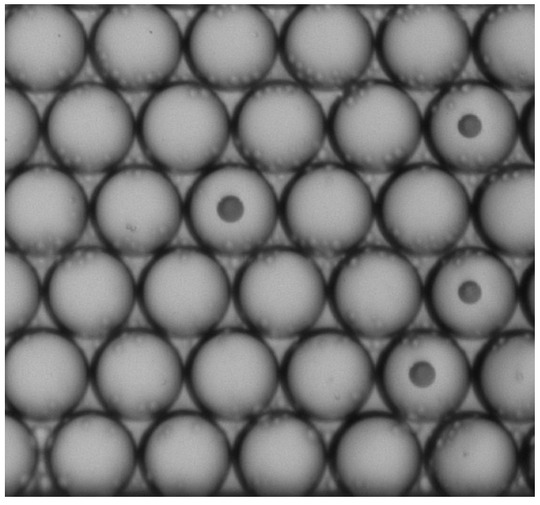
Key considerations for scale

- **Methods for parallelization**
  - Automation for plate-based systems
  - Droplets are massively-parallel, offer the largest scale right now
- **Protocol length?**
  - Can vary dramatically between techniques, especially for library prep
  - Nextera is one of the fastest, but other techniques require have much longer protocols (2-3 days)
- **Individual attention**
  - Pooling reduces costs, but makes it impossible to ‘zero-in’ on a cell of interest



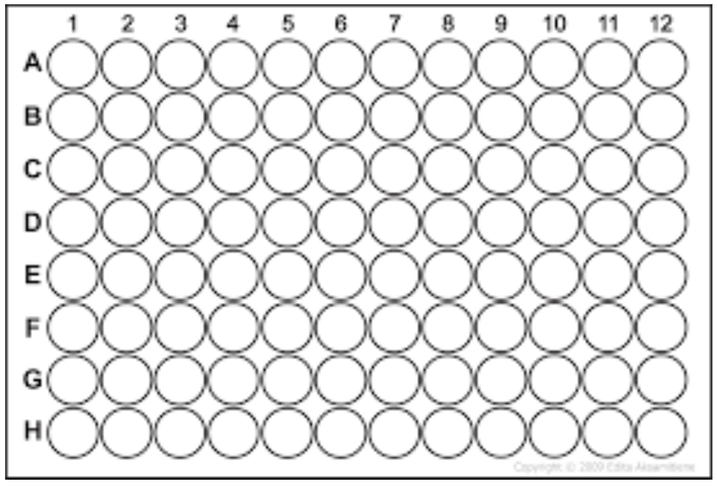
Sensitivity (genes/cell)

Extreme A



DropSeq/inDrop/10x
Cell lines : ~5kgenes/cell
Primary : ~1-3k genes/cell

Extreme B



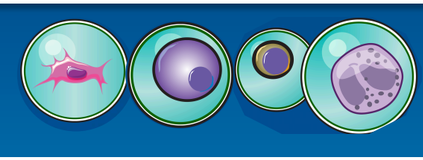
Smart-Seq2 and CelSeq2
Cell lines : ~7-10kgenes/cell
Primary: ~2-6k genes/cell

Extreme B’

Fluidigm C1 (96-cell chip)
Cell lines: ~6-9k genes/cell
Primary: ~1-5k genes/cell

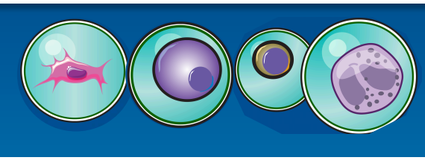
Key considerations for sensitivity

- **Molecular biology**
  - Optimization of lysis, RT
  - Minimize purifications and material loss prior to amplification
- **Pre-amplification**
  - Unevenness inamplification (i.e. GC bias) can dampen sensitivity
  - Overamplification can mask lowly expressed genes
- **Cell size and RNA content is the greatest determinant of data quality**



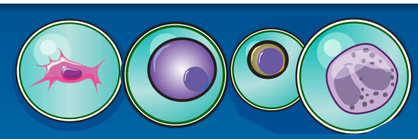
Other important considerations for data quality

- **Read 'efficiency'**
  - Many reads are discarded: don't align, aren't assigned to QC-passing cells, etc.
  - What % of reads are actually useful for calculating gene expression?
- **Unique molecular identifiers**
  - Random sequences attached during RT, enable collapsing of PCR duplicates
  - Sacrifices full-length data and (potentially) sensitivity
- **Evenness of coverage**
  - Especially for pooled protocols, do a subset of cells soak up all the reads?



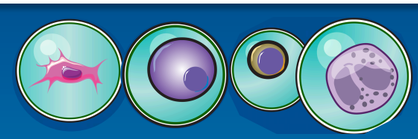
Strengths and tradeoffs : Plate-based approaches

- **Strengths:**
  - Optimized sensitivity, reasonable price, capable of automation
  - **Smart-Seq2:** Full-length transcripts, PCR-based, ~6-8hr protocol
  - **CelSeq2:** 3’ end counting, pooled linear amplification,UMI, 2-3day protocol
- **Weaknesses:**
  - Laborious, lack the scale of droplet-based methods
- **Ideal use-cases**
  - Deep and sensitive characterization of ~hundreds-thousands of cells
  - Unique advantages for time-course and index-sorting experiments



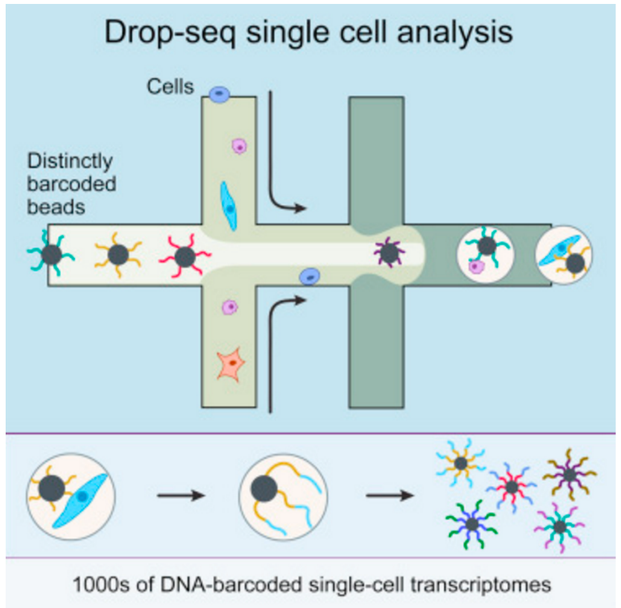
Strengths and tradeoffs : Droplet methods

- **Strengths:**
  - Transformative scale, low cost.
  - Parallelized approach dramatically reduces batch effects, data is UMI-based
  - Minimal equipment setup (no sorter or automation),
- **Weaknesses:**
  - Sensitivity and coverage, particularly for primary cells
  - Cannot visualize or profile (i.e. index-sorting) sequenced cells
- **Ideal use-cases**
  - Unbiased discovery of rare populations (<1%)

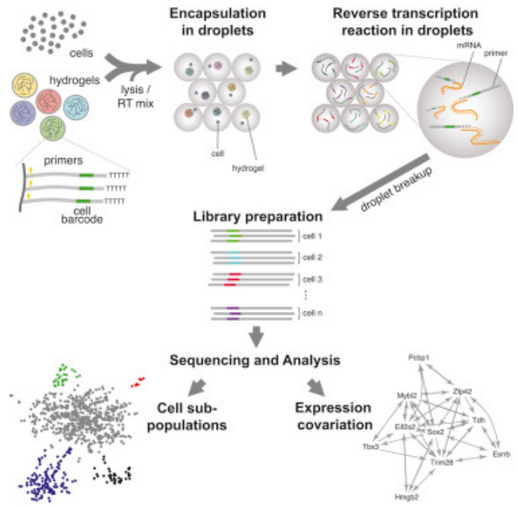


Strengths and tradeoffs : Droplet methods

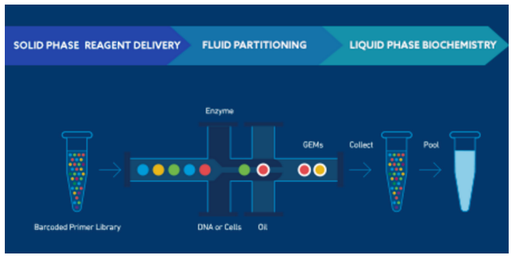
Drop-seq

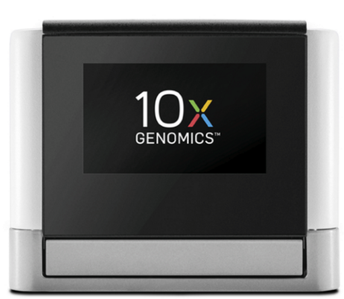


inDrop



10X Chromium

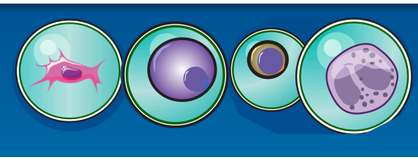




Similarities outnumber the differences, but there are differences

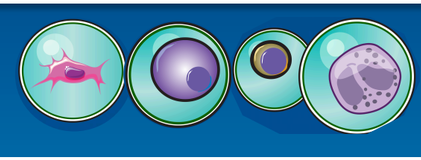
Strengths and tradeoffs : Fluidigm C1

- **Strengths:**
  - Fully automated workflow up until library preparation, full-length sequencing
  - Possible to visualize cell prior to sequencing
- **Weaknesses:**
  - High costs (equipment and microfluidic chips), capture is biased by cell size
  - Time-course experiments require multiple machines
- **Ideal use-cases**
  - Linking visual phenotypes with gene expression



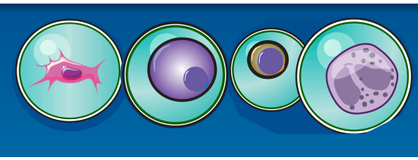
How many cells do I need to sequence?

- **It depends!**



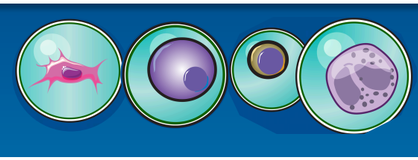
How many cells do I need to sequence?

- **It depends!**
  - Strong analogy to Human Genetics : How many people do I need to sequence?
    - » Common disease or driven by rare variants?
    - » What is the effect size of each variant?
  - Single cell RNA-seq analogy
    - » How rare is the cell type of interest?
    - » Does it have highly expressed markers?



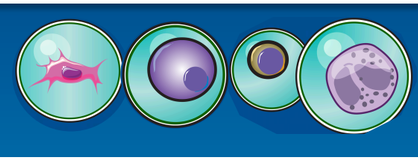
How many cells do I need to sequence?

- **It depends!**
  - Strong analogy to Human Genetics : How many people do I need to sequence?
    - » Common disease or driven by rare variants?
    - » What is the effect size of each variant?
  - Single cell RNA-seq analogy
    - » How rare is the cell type of interest?
    - » Does it have highly expressed markers?
- **Rahul's rule of thumb:**
  - Aim to profile 20-50 of each expected cell type/state (50-100 for droplet-based data)
  - Check out www.satijalab.org/howmanycells (great for budget justifications!)



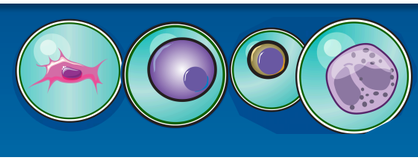
How deep do I need to sequence?

- **It depends!**
  - Are you able to pool correlated genes across cells?
    - » Discovery of cell types, reconstructing regulatory networks
    - » Can exploit gene-gene correlations to gain power
    - » Could use much low-coverage sequencing
  - Are you studying heterogeneity of single genes within a population?
    - » Would require deeper sequencing and more sensitive protocols



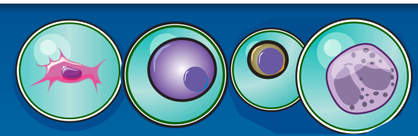
Which technology is best for my experiment?

- **It depends!**
  - Balance **biological questions** with **experimental design**
  - How much **prior knowledge** do you have about cellular heterogeneity
  - Are you searching for **rare populations** or 50/50 splits?
  - Would it be beneficial to have **protein surface marker** data for your cells?
  - Do you need to sample across **multiple experimental conditions simultaneously?**
  - Are cell or sample number limiting?

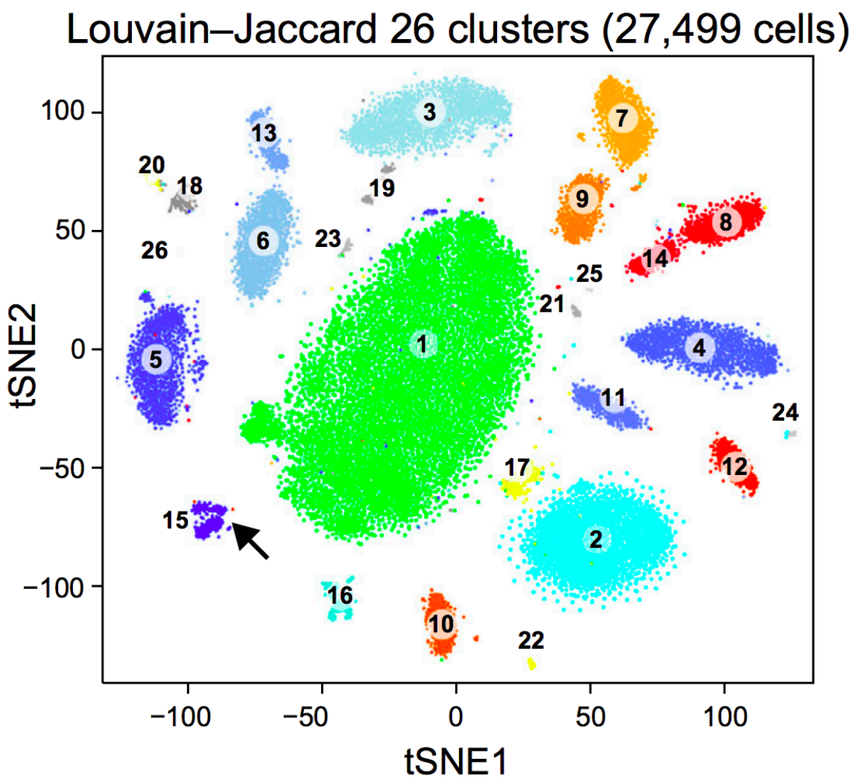


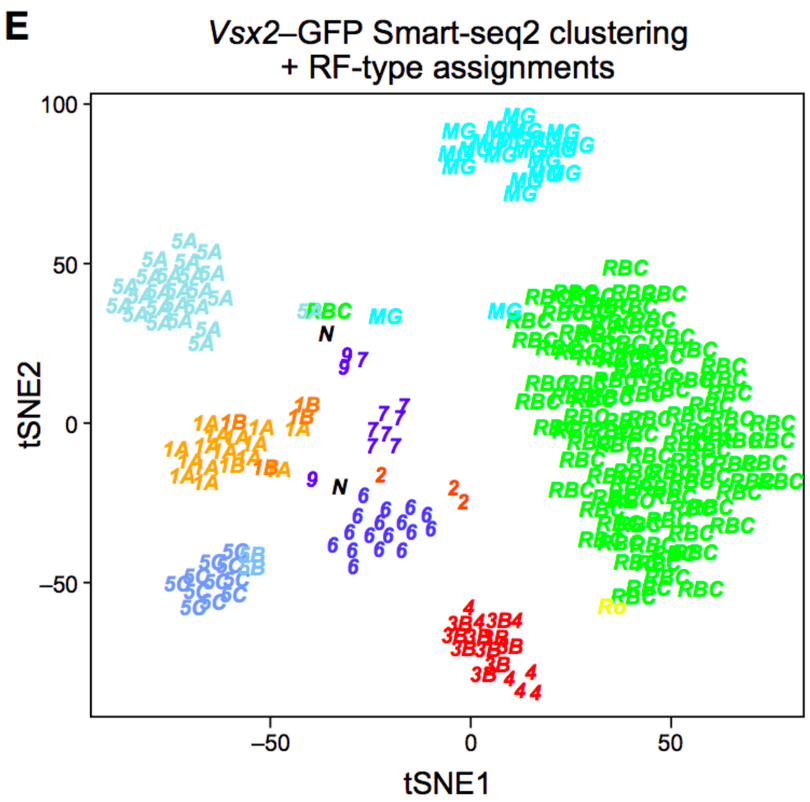
Breadth or Depth?

- **It depends!**
- The argument for more cells at low coverage
  - Discovering rare populations requires huge datasets
  - Cell types are defined by highly expressed markers, networks are defined by highly expressed targets
  - Profiling more cells augments statistical power to subdivide abundant groups
- The argument for fewer cells at high coverage
  - Subtle subdivisions may be defined by lowly expressed transcripts, that are not in the top 1-2k most highly expressed genes



Breadth or Depth?





Shekhar et al., Cell 2016

Rule of thumb : More cells >> More genes/cell