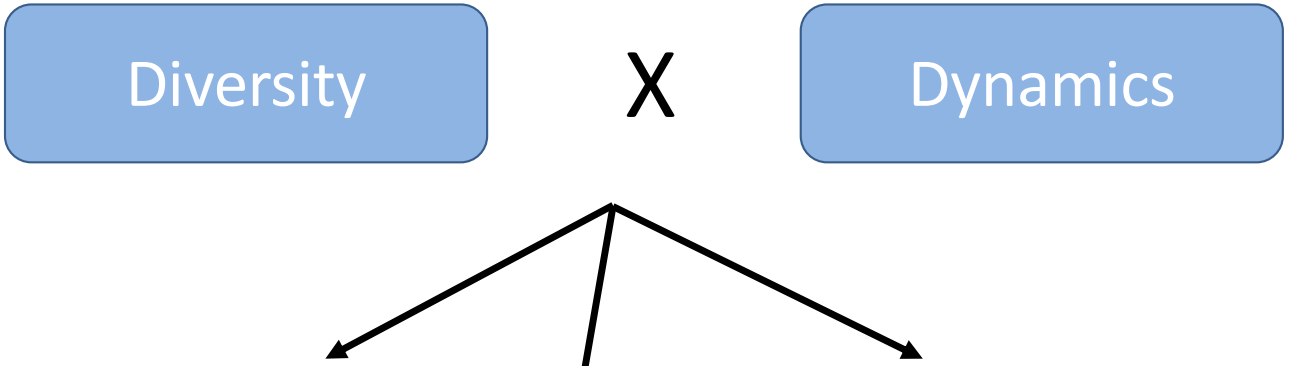
UCDAVIS Bioinformatics Core

Single Cell Transcriptomics

Jie (Jessie) Li
Bioinformatics Core
Genome Center
University of California, Davis
jjsli@ucdavis.edu; bioinformatics.core@ucdavis.edu

UCDAVIS Bioinformatics Core

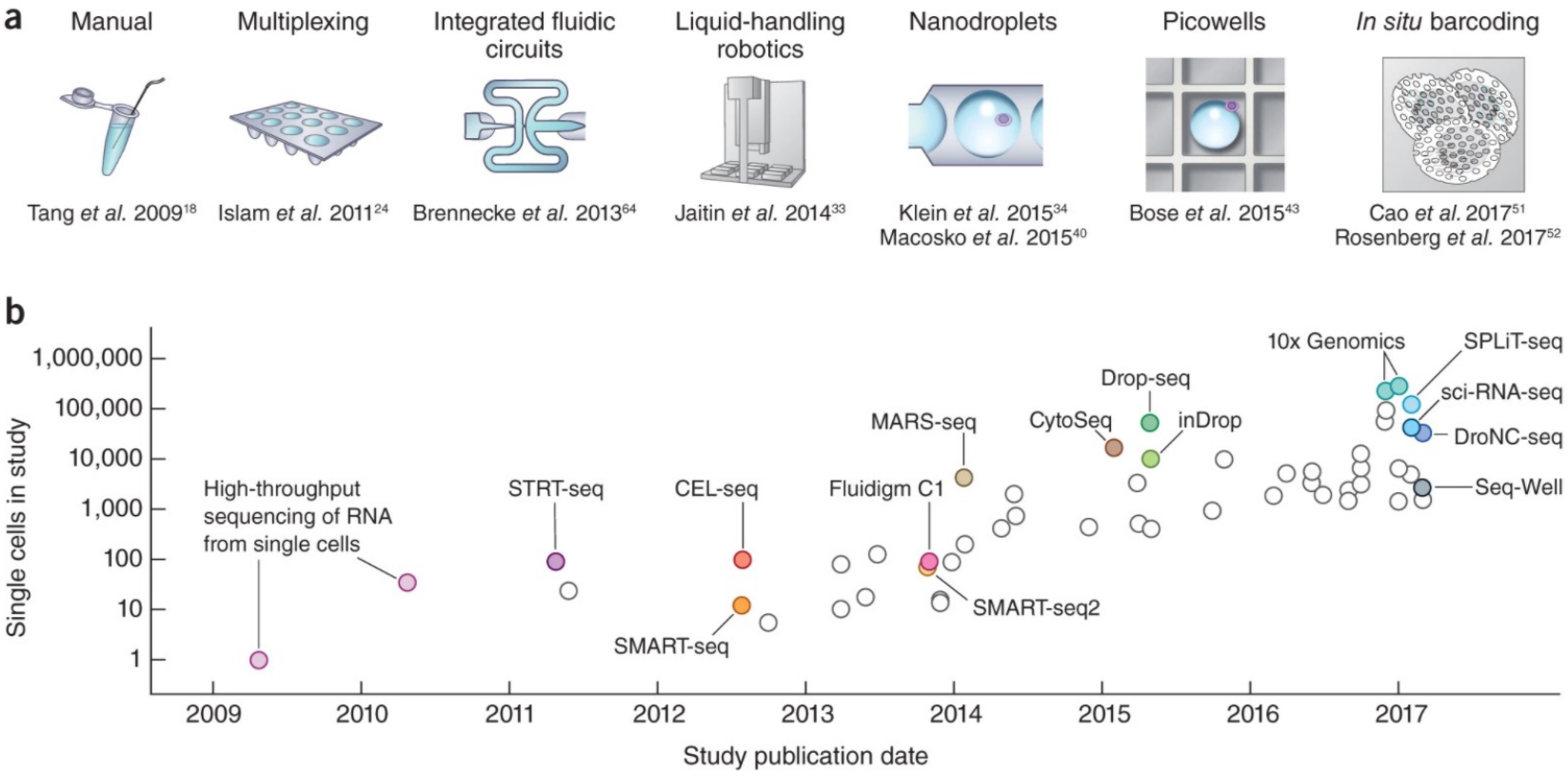
High resolution biology



- Immunology
- Neurology
- Oncology
- Developmental biology
- Stem cell biology

UCDAVIS Bioinformatics Core

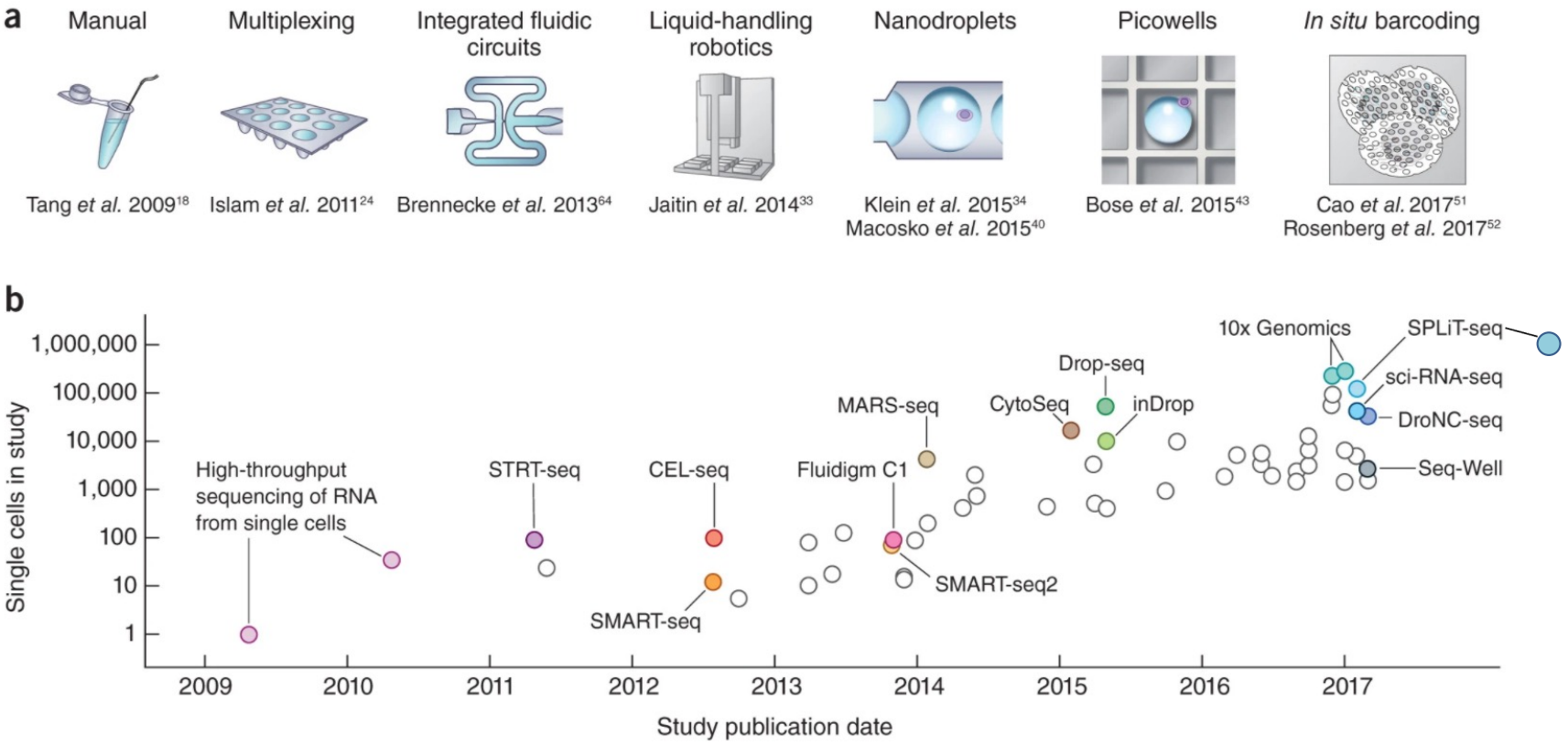
Single cell technologies



Svensson, etc., 2018, Nature Protocols https://www.nature.com/articles/nprot.2017.149

UCDAVIS Bioinformatics Core

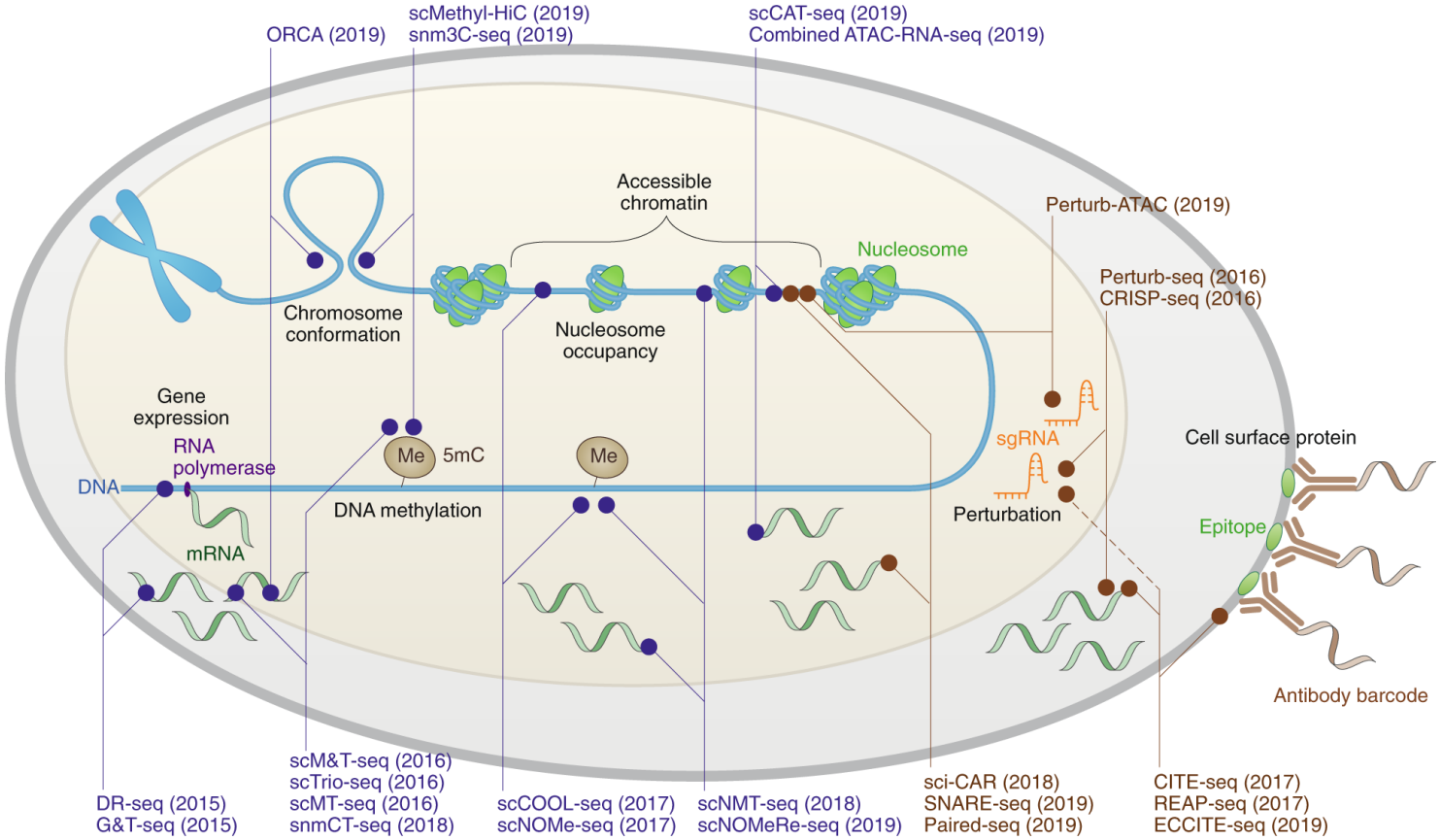
Single cell technologies



Svensson, etc., 2018, Nature Protocols https://www.nature.com/articles/nprot.2017.149

UCDAVIS Bioinformatics Core

High resolution biology

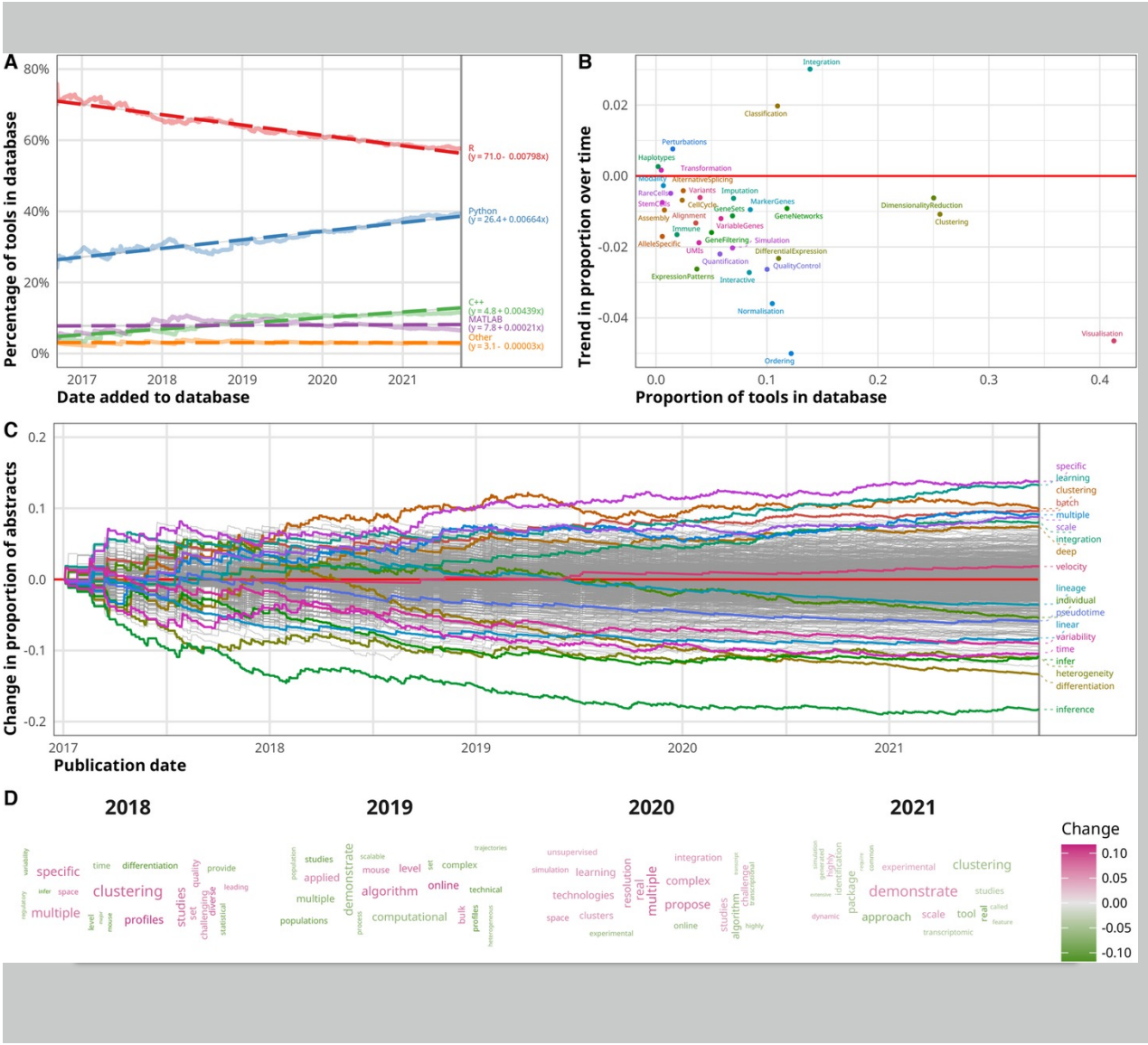


Zhu, etc., *Nature Methods*, 2020, https://www.nature.com/articles/s41592-019-0691-5

UCDAVIS Bioinformatics Core

High resolution biology

- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02519-4



UCDAVIS Bioinformatics Core

Designing Experiments

Beginning with the question of interest (and working backwards)

- The final step of an analysis is comparisons between sample/conditions, which means the application of a model to each gene in your dataset.

  Traditional statistical considerations and basic principals of statistical design of experiments apply.

  - **Control** for effects of outside variables, avoid/consider possible biases, avoid confounding variables in sample preparation.
  - **Randomization** of samples, plots, etc.
  - **Replication** is essential (triplicates are THE minimum)
- You should know your final (DE) model and comparison contrasts before beginning your experiment.

How many cells to target?

- The number of cells to target can be estimated based on:
  - The expected heterogeneity of all cells in a sample
  - The minimum frequency expected of a particular cell type within the sample, and
  - The minimum number of cells of each type desired in the resulting data set.
- With this information, a negative binomial distribution can be used to estimate the number of cells likely to capture at least a set number of cells from your rarest cell type.
- For example, if we sequence a mixture of ~10 cell types where the frequency of the rarest cell type is ~0.03, then we would need to sequence ~2250 cells to have a 90% chance of capturing at least 50 of those rare cells.

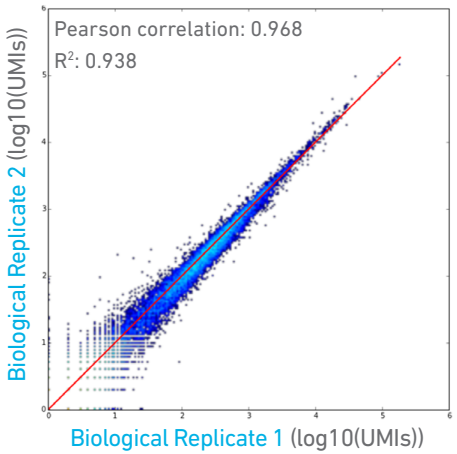
www.satijalab.org/howmanycells

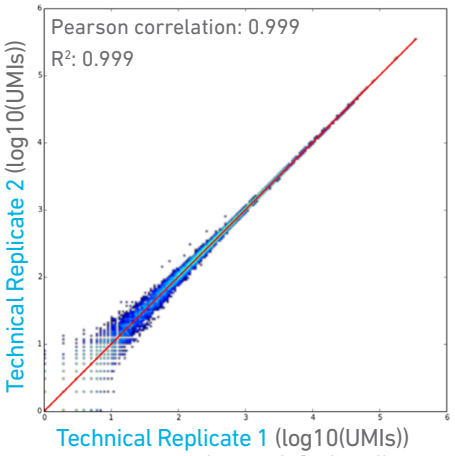
UCDAVIS Bioinformatics Core

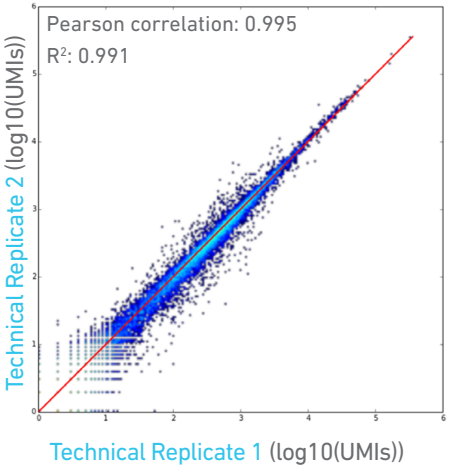
General rules for preparing samples

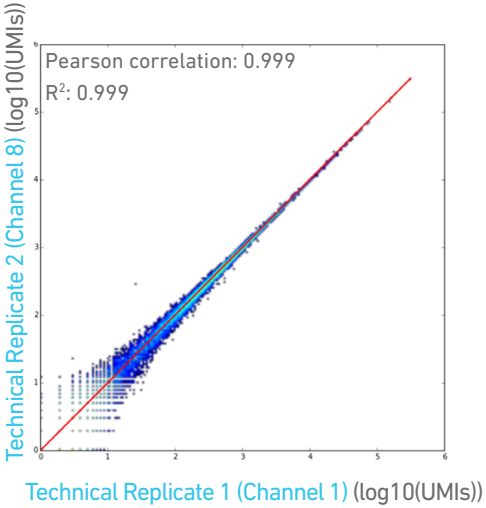
- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- If cells clumps or cell debris are observed, filter cells using a cell strainer with an appropriate pore size
- Determine the cell concentration using a Countess® II Automated Cell Counter or other cell counting device
- Initial cell count depends on the target, however, expect at least 50% loss in the final stages and loss during cleanup

UCDAVIS Bioinformatics Core









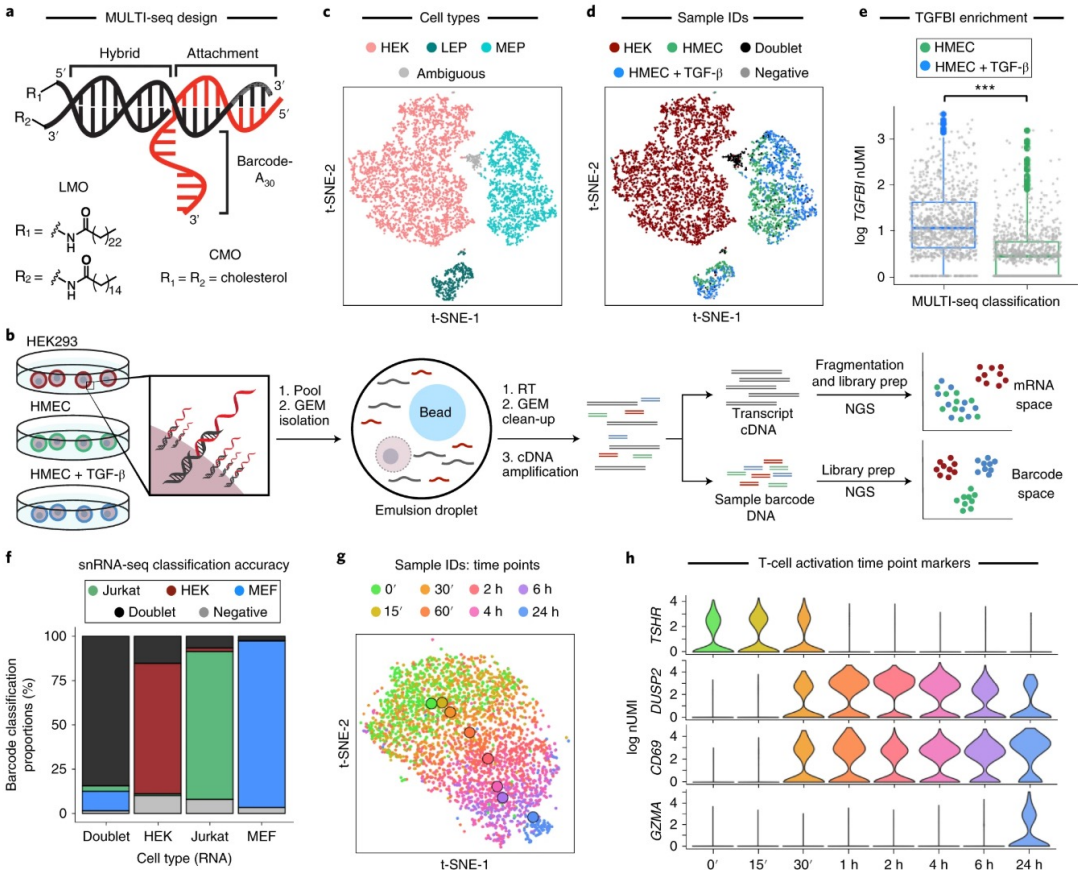
https://cdn.10xgenomics.com/image/upload/v1660261285/support-documents/CG000170_TechNote_BiologicalandTechnicalVariationinSingleCell3_GeneExpressionExperiments_RevA_.pdf

UCDAVIS Bioinformatics Core

MULTIseq - https://www.nature.com/articles/s41592-019-0433-8

https://satijalab.org/costpercell

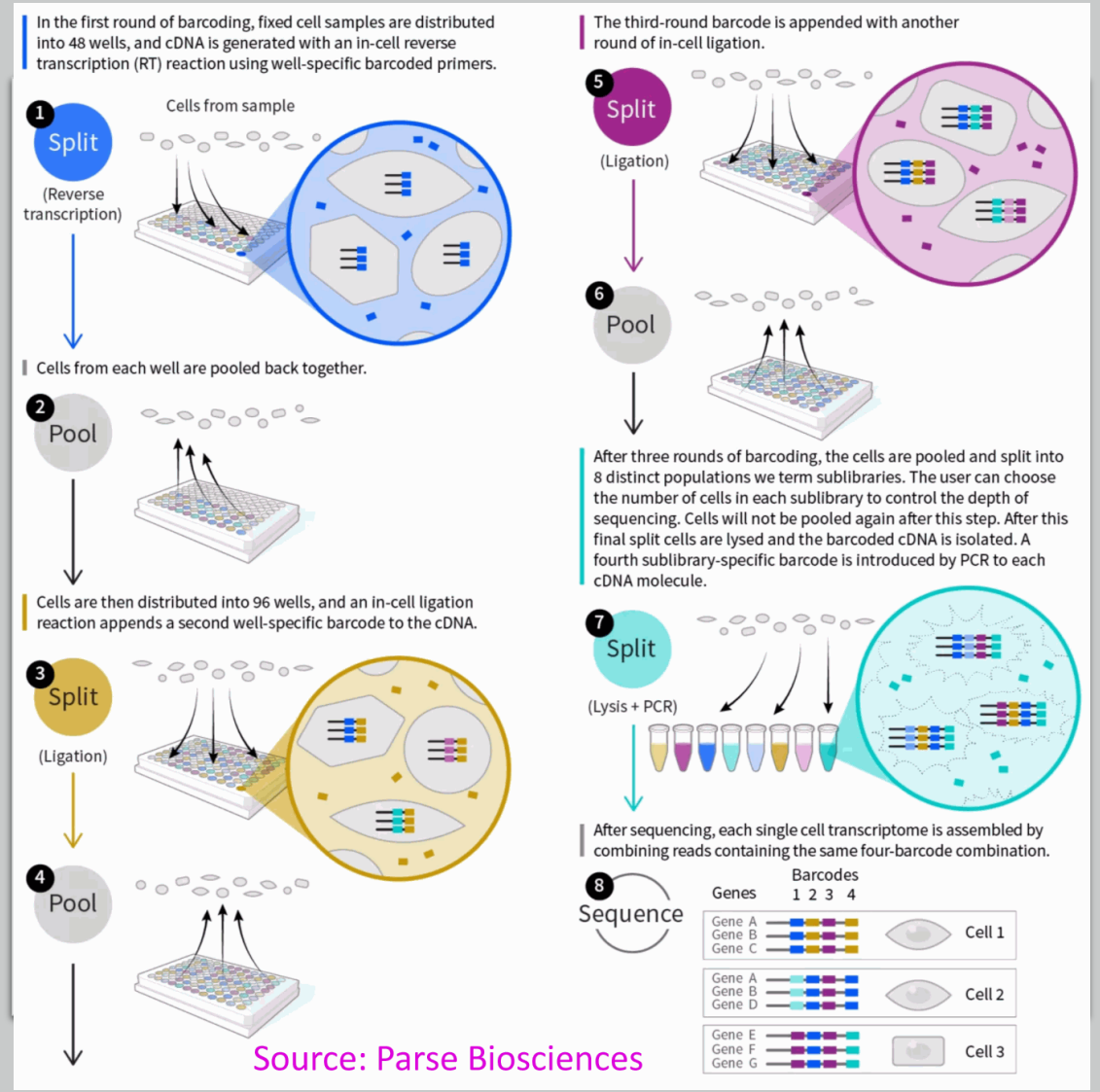
Multiplexing – cell hashing



UCDAVIS Bioinformatics Core

Multiplexing – cell hashing

- Parse Biosciences



Source: Parse Biosciences

Sequencing Depth

- Coverage is determined differently for “Counting” based experiments (RNAseq, amplicons, etc.) where an expected number of reads per cell is typically more suitable.
- The first and most basic question is how many reads per cell will I get Factors to consider are (per lane):
  1. Number of reads being sequenced
  2. Number of cells being sequenced (estimates)
  3. Expected percentage of usable data

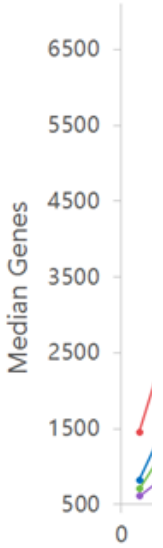
$$\frac{reads}{cell} = \frac{reads.sequenced \ * 0.8}{cells.pooled}$$

- Read length, or SE vs PE, does not factor into sequencing depth.

UCDAVIS Bioinformatics Core

Sequencing - Characterization of transcripts, or differential gene expression

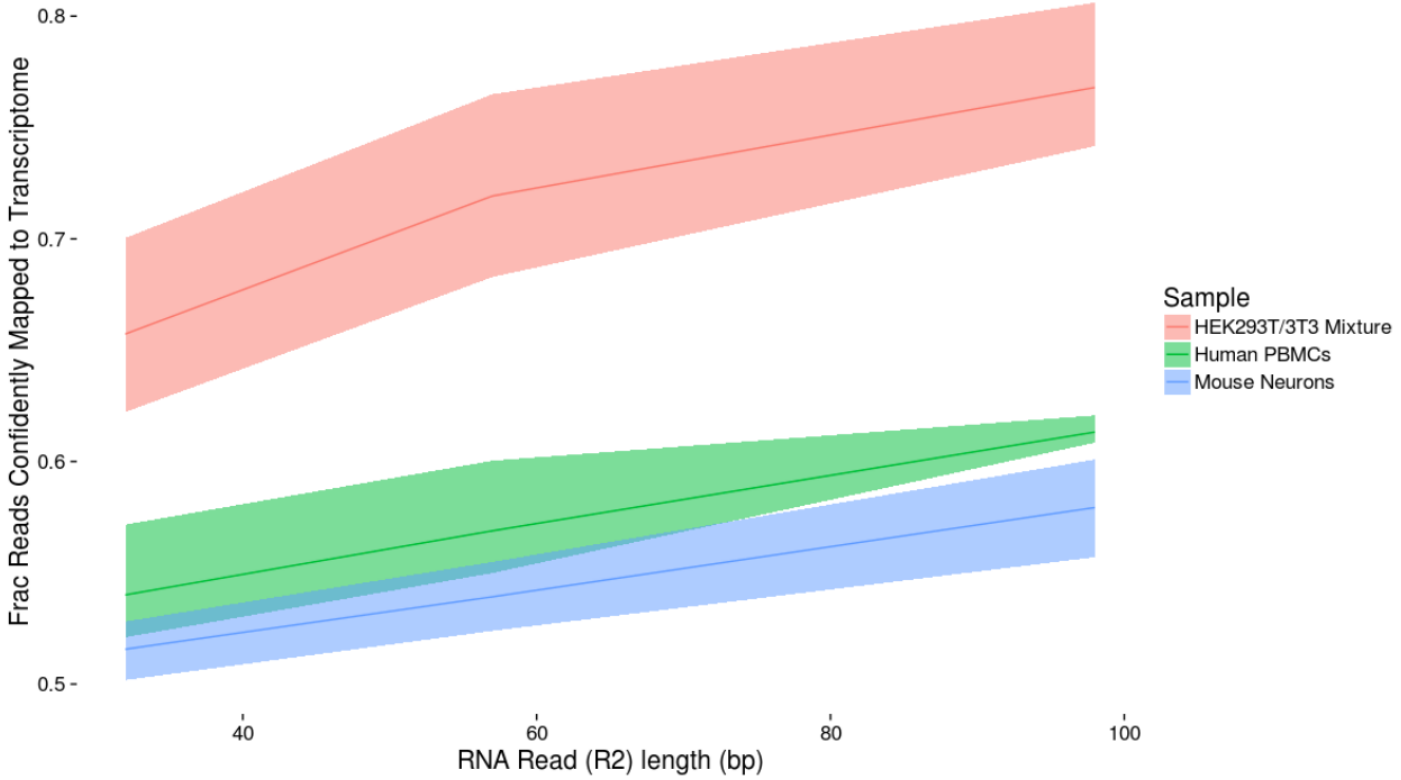
- Read length needed depends on likelihood of mapping uniqueness, but generally longer is better and paired-end is better than single-end (except when its not) (75bp or greater is best).
- Complexity of sample: the higher the complexity, the higher the depth.
- Interest in detecting genes expressed at low levels: the lower the level, the higher the depth.
- The fold change you want to be able to detect (smaller fold change requires more replicates and higher depth).
- Detection of novel transcripts, or quantification of isoforms (full-length libraries) requires >> sequencing depth. [NON 3' based methods]



The amount of sequencing needed for a given experiment is best determined by the goals of the experiment and the nature of the sample.

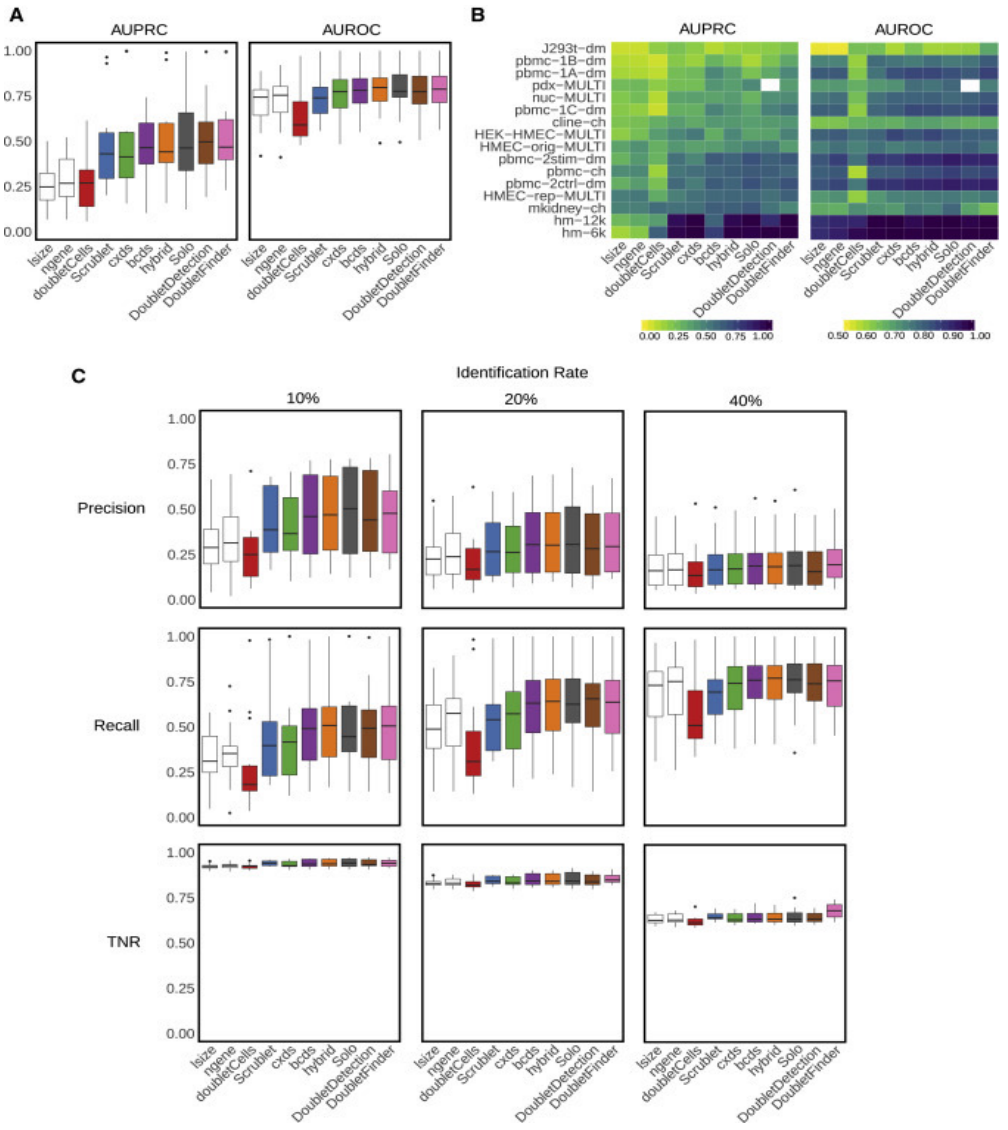
UCDAVIS Bioinformatics Core

Read length matters (10x slide)



UCDAVIS Bioinformatics Core

Doublet detection



Xi, etc., Cell Systems, 2021, https://www.sciencedirect.com/science/article/pii/S2405471220304592

UCDAVIS Bioinformatics Core

Sequencing, V3

Validated on
- Novaseq
- HiSeq 4000
- HiSeq 2500 Rapid Run
- NextSeq
- MiSeq

Recommendation

- 20,000* raw reads per cell is the recommended sequencing depth for 'typical' samples.
- Given variability in cell counting/loading, extra sequencing may be required if the cell count is higher than anticipated.

*Adjust sequencing depth for the required performance or application. The Sequencing Saturation metric and curve in the Cell Ranger run summary can be used to optimize sequencing depth for specific sample types.

sequencing run, with 3 reads, V3 kits

Sequence Read	Minimum Length	Read Description
Read 1	28bp (16bp bc, 12bp UMI)	barcode and UMI
I7 Index	8bp	Sample Index Read
Read2	100bp	Transcript Tag

**Shorter transcript reads may lead to reduced transcriptome alignment rates. Cell barcode, UMI and Sample index reads must not be shorter than indicated. Any read can be longer than recommended.

@ full capacity 10,000 cells per sample and 20K reads per cell = 200M reads or ~0.5 lanes/sample

UCDAVIS Bioinformatics Core

Cost Estimation

- Cell Isolation
- Library preparation (Per sample/pool)
- Sequencing (Number of lanes)
- Bioinformatics

  General rule is to estimate the same dollar amount as data generation, i.e. double your budget

http://dnatech.genomecenter.ucdavis.edu/prices/

https://bioinformatics.ucdavis.edu/rates

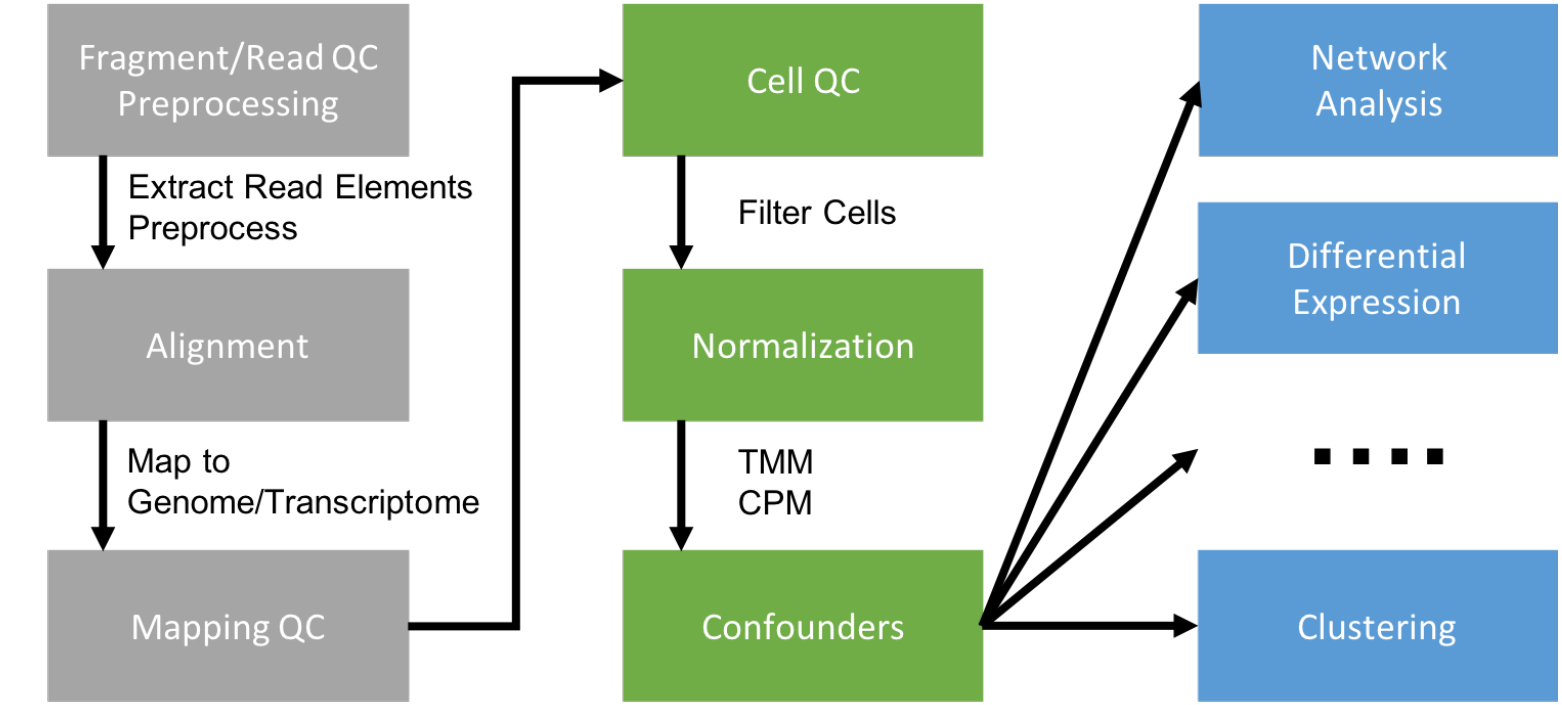
UCDAVIS Bioinformatics Core

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

UCDAVIS Bioinformatics Core

Single Cell RNASeq Analysis



UCDAVIS Bioinformatics Core

Genomics and Bioinformatics

Following data science principles, 2 stages in bioinformatics

- **Data reduction**

  Sequence data (raw data) to summarized form.

  * Command line, shell scripting, and programming.
  * Requires an understanding of the technology, molecular biology.
  * Removing technical noise from data.

- **Data analysis**

  Summarized data to biological interpretation

  * R/Python statistical programming
  * Requires an understanding of the biological question, statistics.

UCDAVIS Bioinformatics Core

Summary:

Data science (Bioinformatics) is both a **science** and an **art**.

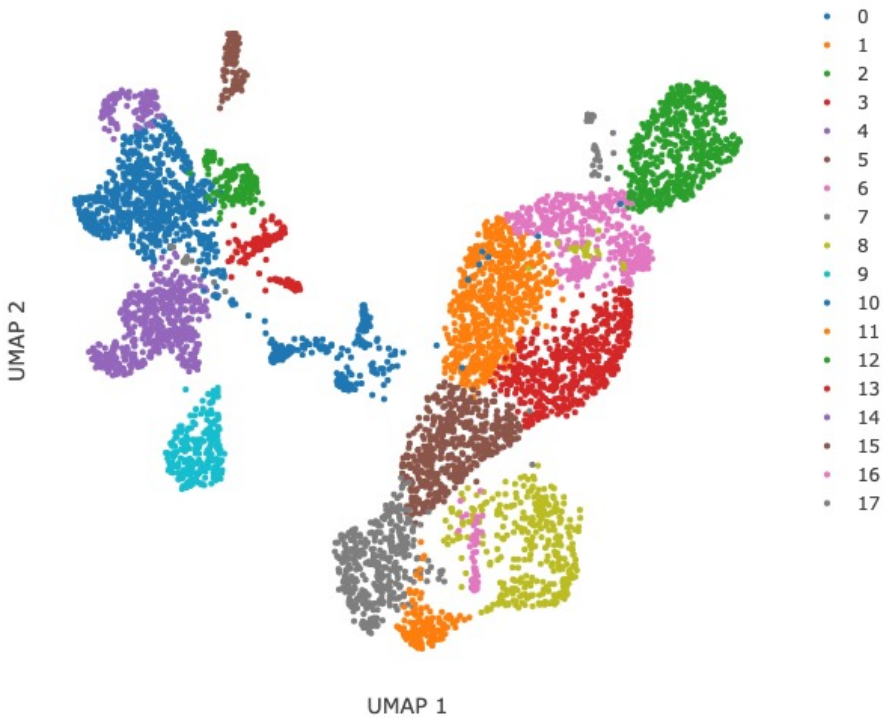
Spend the time (and money) planning and producing **good quality, accurate and sufficient data.**

Get to know to the data, develop and test expectations, explore and identify patterns.

Result, **spend much less time** (and less money) extracting biological significance and results with fewer failures and reproducible research.

UCDAVIS Bioinformatics Core

Example 1: multiplex 11 samples, >5K cells



UCDAVIS Bioinformatics Core

Example 2: multiplex 6 samples, >100K cells