

Single Cell Transcriptomics

Jie (Jessie) Li

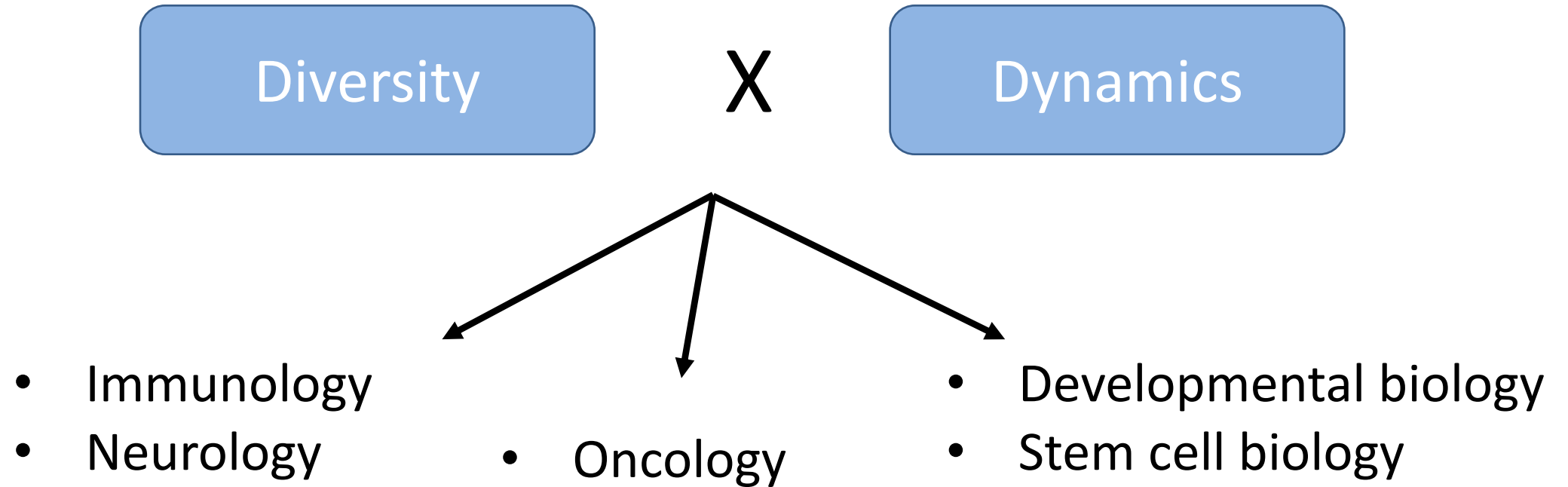
Bioinformatics Core

Genome Center

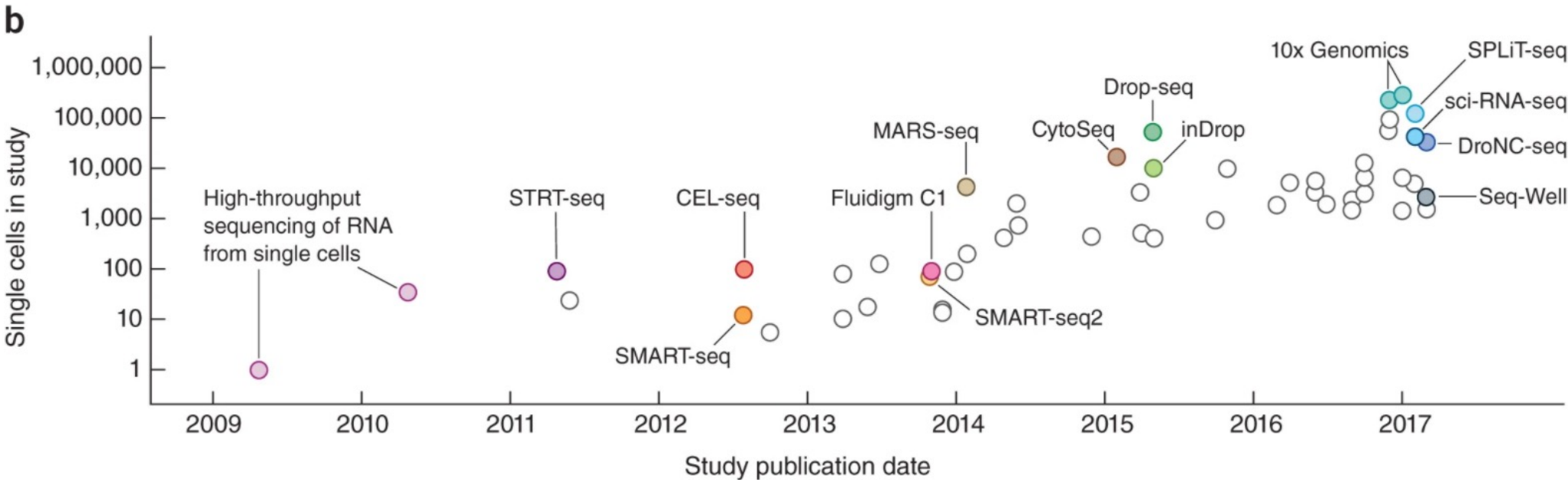
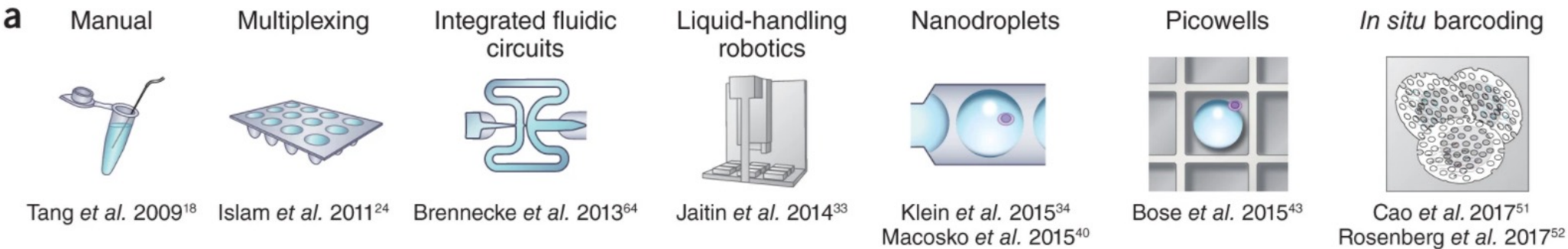
University of California, Davis

jjli@ucdavis.edu; bioinformatics.core@ucdavis.edu

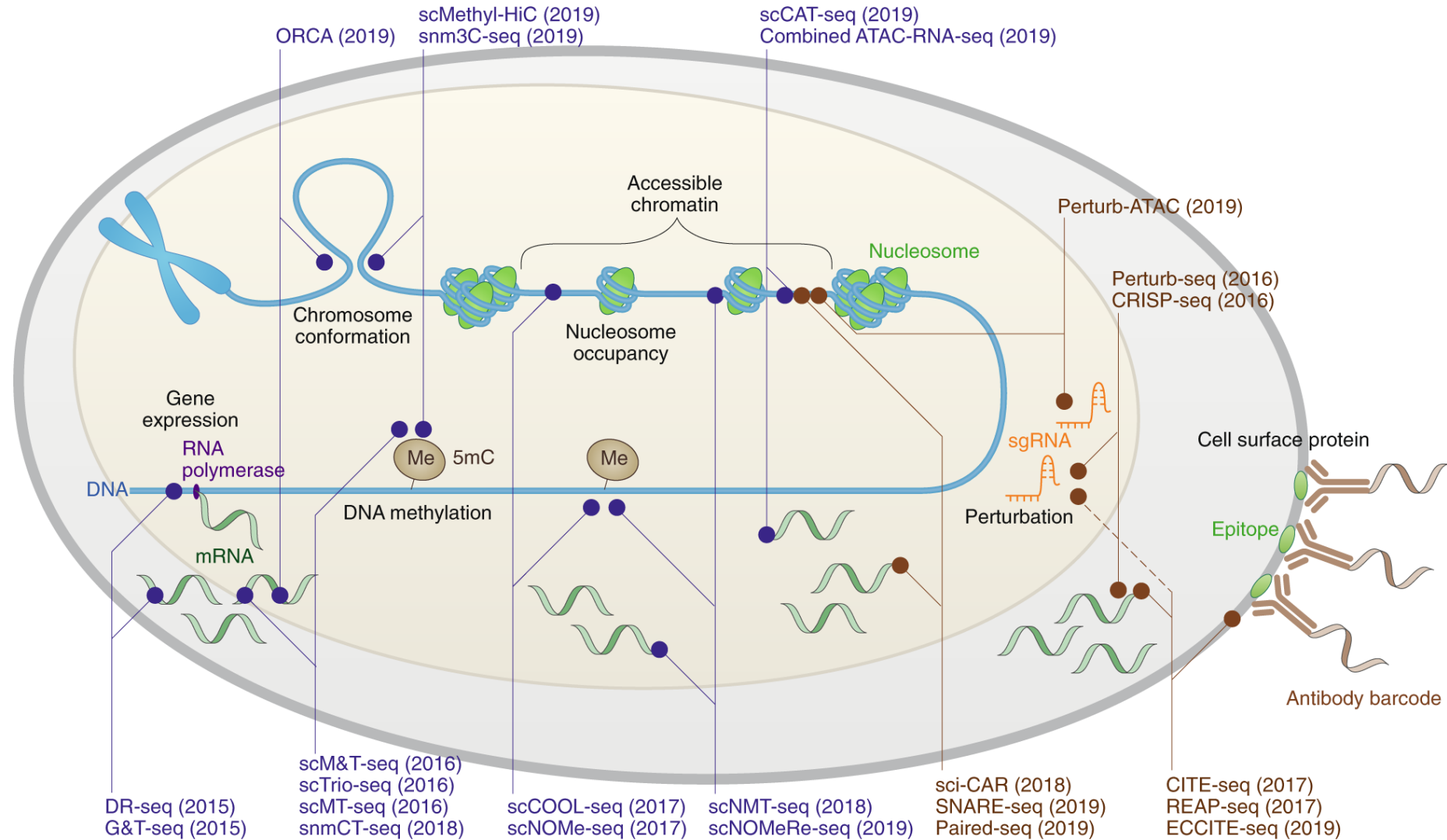
High resolution biology



Single cell technologies

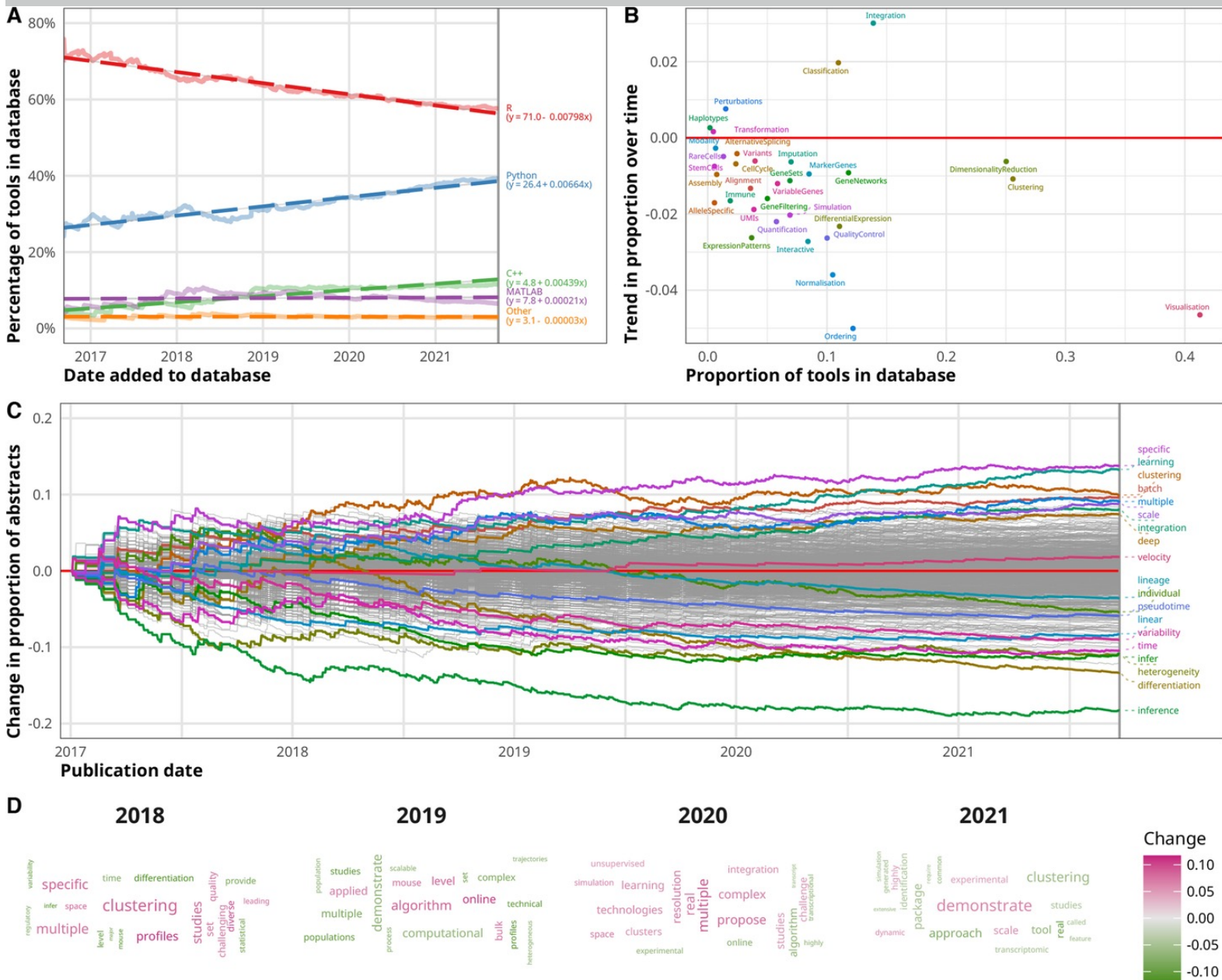


High resolution biology



High resolution biology

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



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.

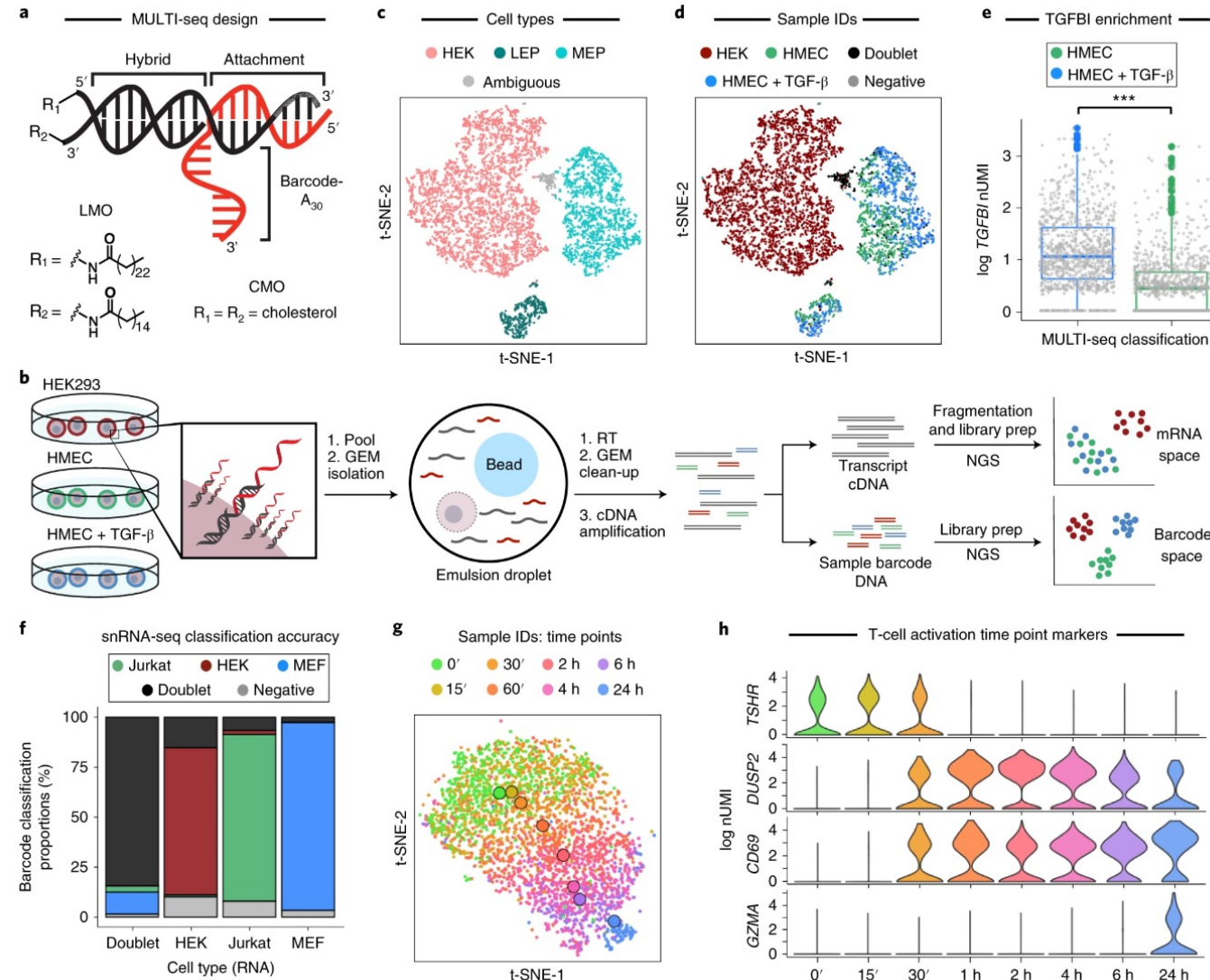
How many cells to target?

- 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.

General rules for preparing samples

- Prepare more samples than 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

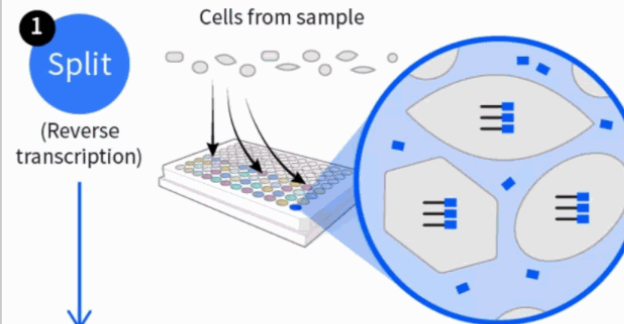
Multiplexing – cell hashing



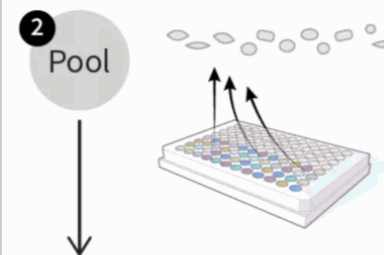
Multiplexing – cell hashing

- Parse Biosciences Mega kit – 1M cells

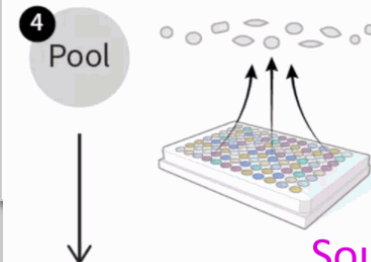
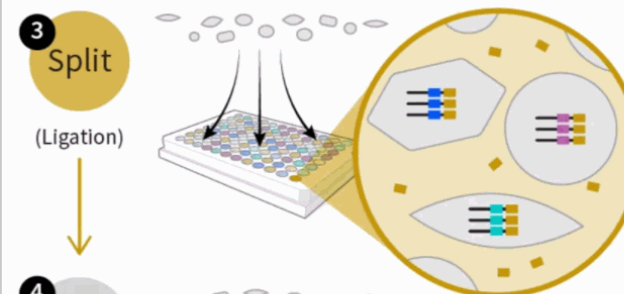
In the first round of barcoding, fixed cell samples are distributed into 48 wells, and cDNA is generated with an in-cell reverse transcription (RT) reaction using well-specific barcoded primers.



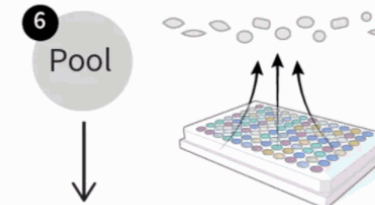
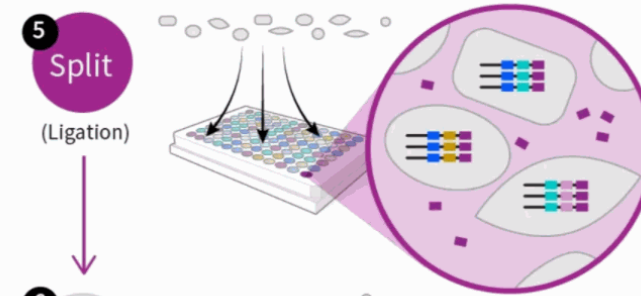
Cells from each well are pooled back together.



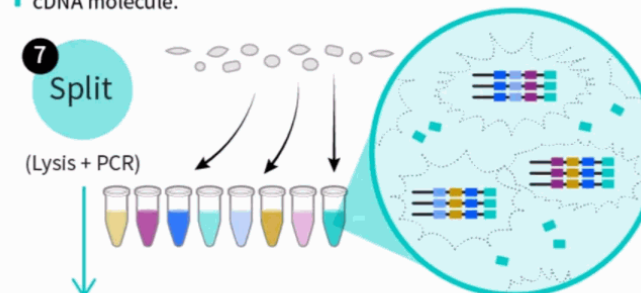
Cells are then distributed into 96 wells, and an in-cell ligation reaction appends a second well-specific barcode to the cDNA.



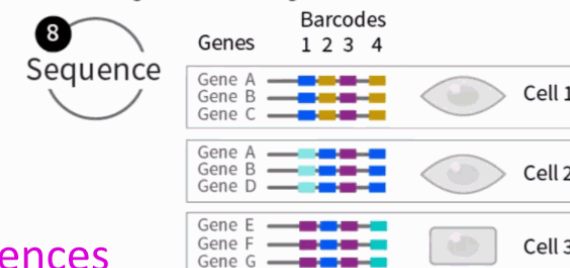
The third-round barcode is appended with another round of in-cell ligation.



After three rounds of barcoding, the cells are pooled and split into 8 distinct populations we term sublibraries. The user can choose the number of cells in each sublibrary to control the depth of sequencing. Cells will not be pooled again after this step. After this final split cells are lysed and the barcoded cDNA is isolated. A fourth sublibrary-specific barcode is introduced by PCR to each cDNA molecule.



After sequencing, each single cell transcriptome is assembled by combining reads containing the same four-barcode combination.



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

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

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

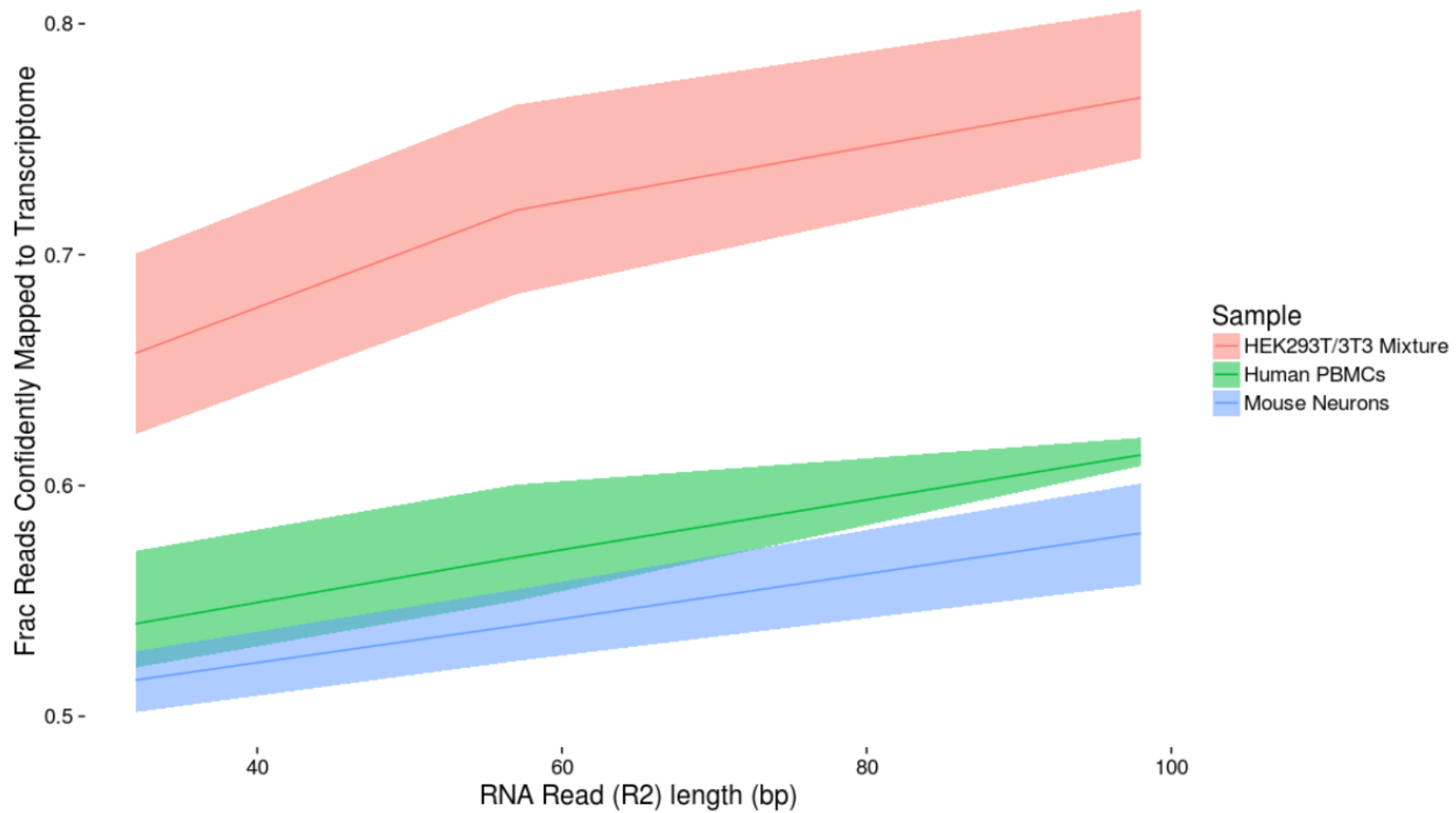
Sequencing - Characterization of transcripts, or differential gene expression

Factors to consider are:

- 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

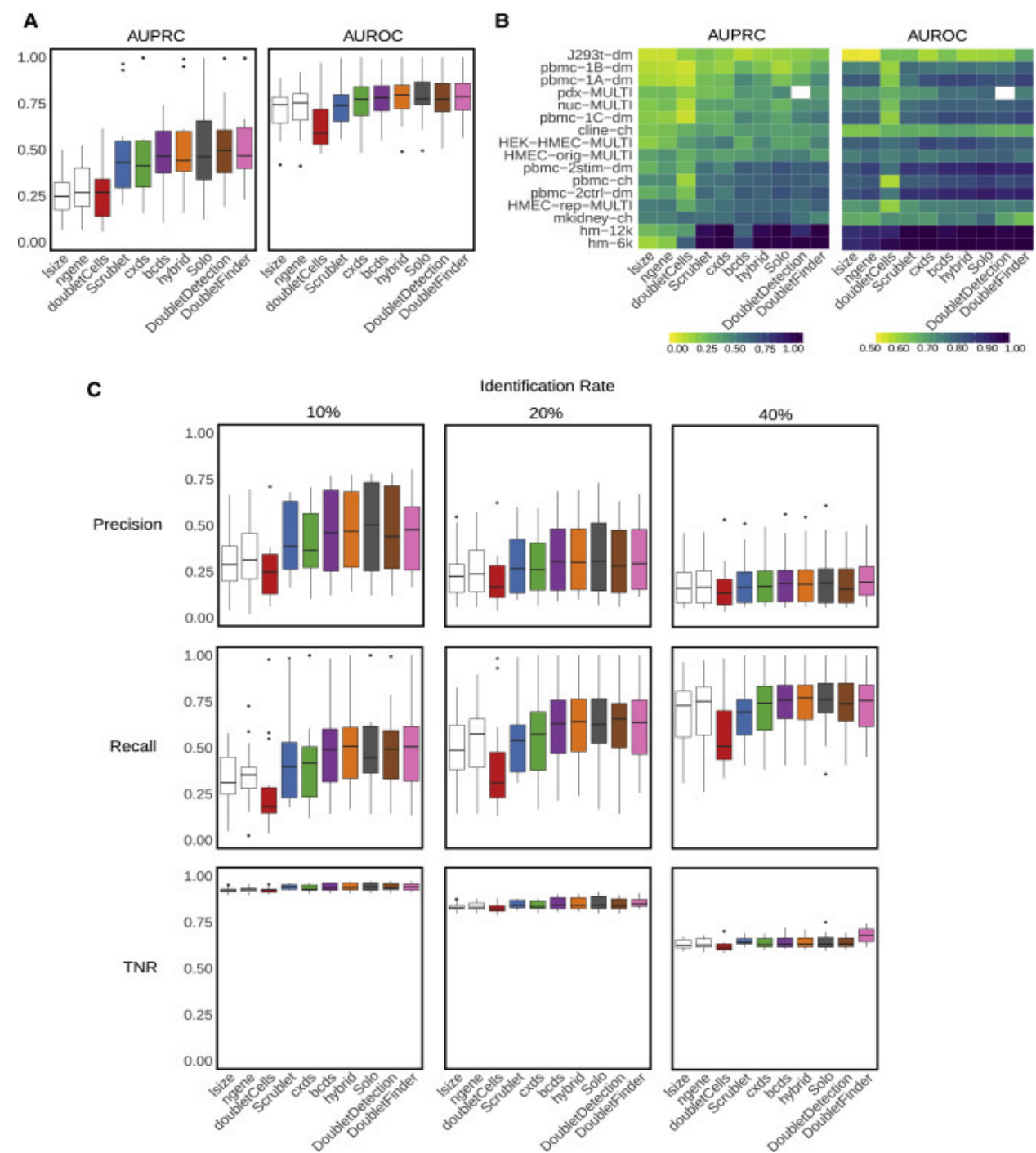
The amount of sequencing needed for a given experiment is best determined by the goals of the experiment and the nature of the sample.

Read length matters (10x slide)



Doublet detection

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



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

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

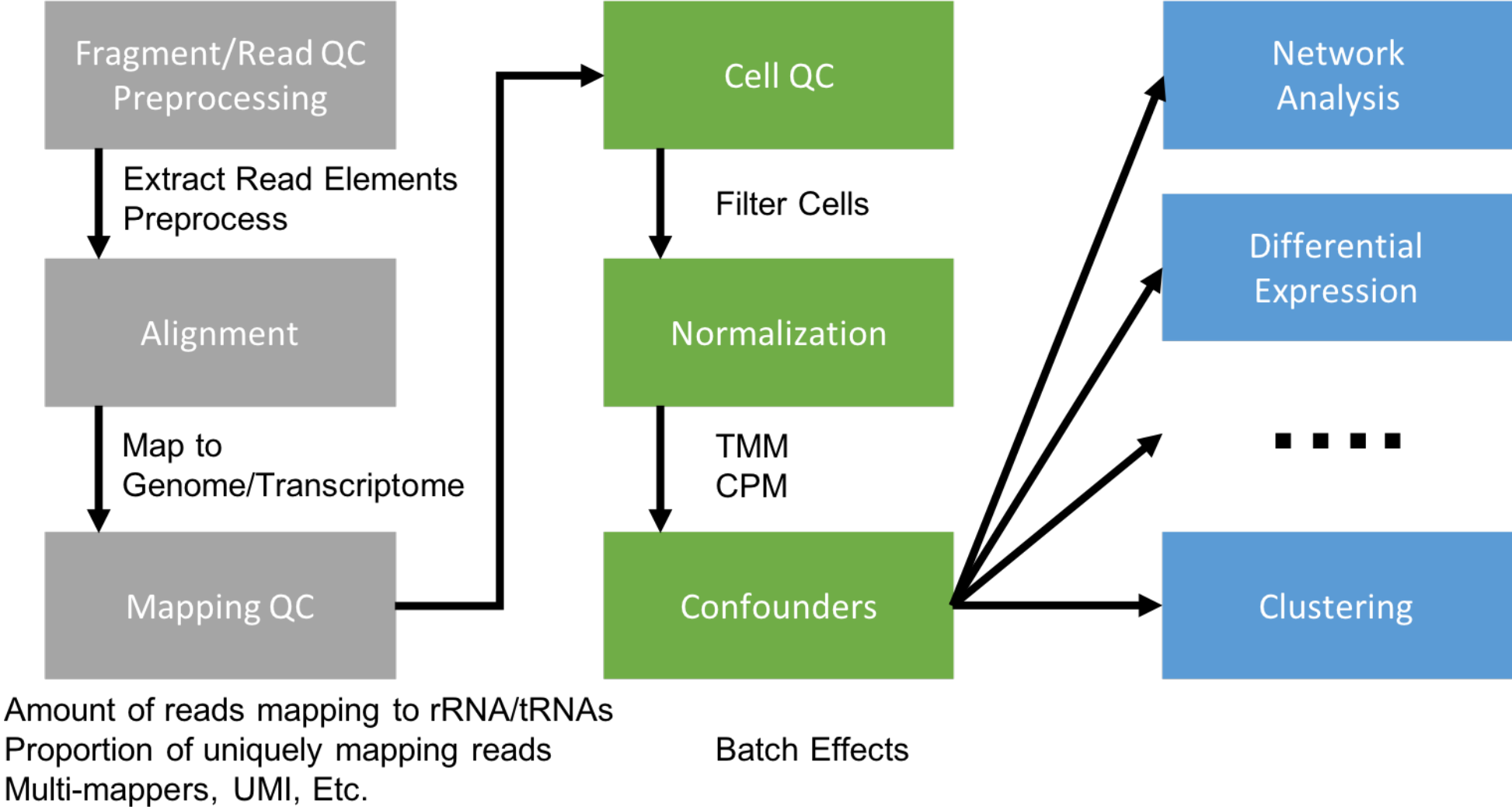
<http://dnatech.genomecenter.ucdavis.edu/prices/>

<https://bioinformatics.ucdavis.edu/rates>

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

Single Cell RNASeq Analysis



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.

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 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.