Single Cell Transcriptomics scRNAseq

Matthew L. Settles
Genome Center Bioinformatics Core
University of California, Davis
settles@ucdavis.edu; bioinformatics.core@ucdavis.edu

Purpose

The sequencing of the transcriptomes of single-cells, or single-cell RNA-sequencing, has now become the dominant technology for the identification of novel cell types and for the study of stochastic gene expression.

Single-cell transcriptomics determines what genes (and in what relative quantity) are being expressed in each cell.

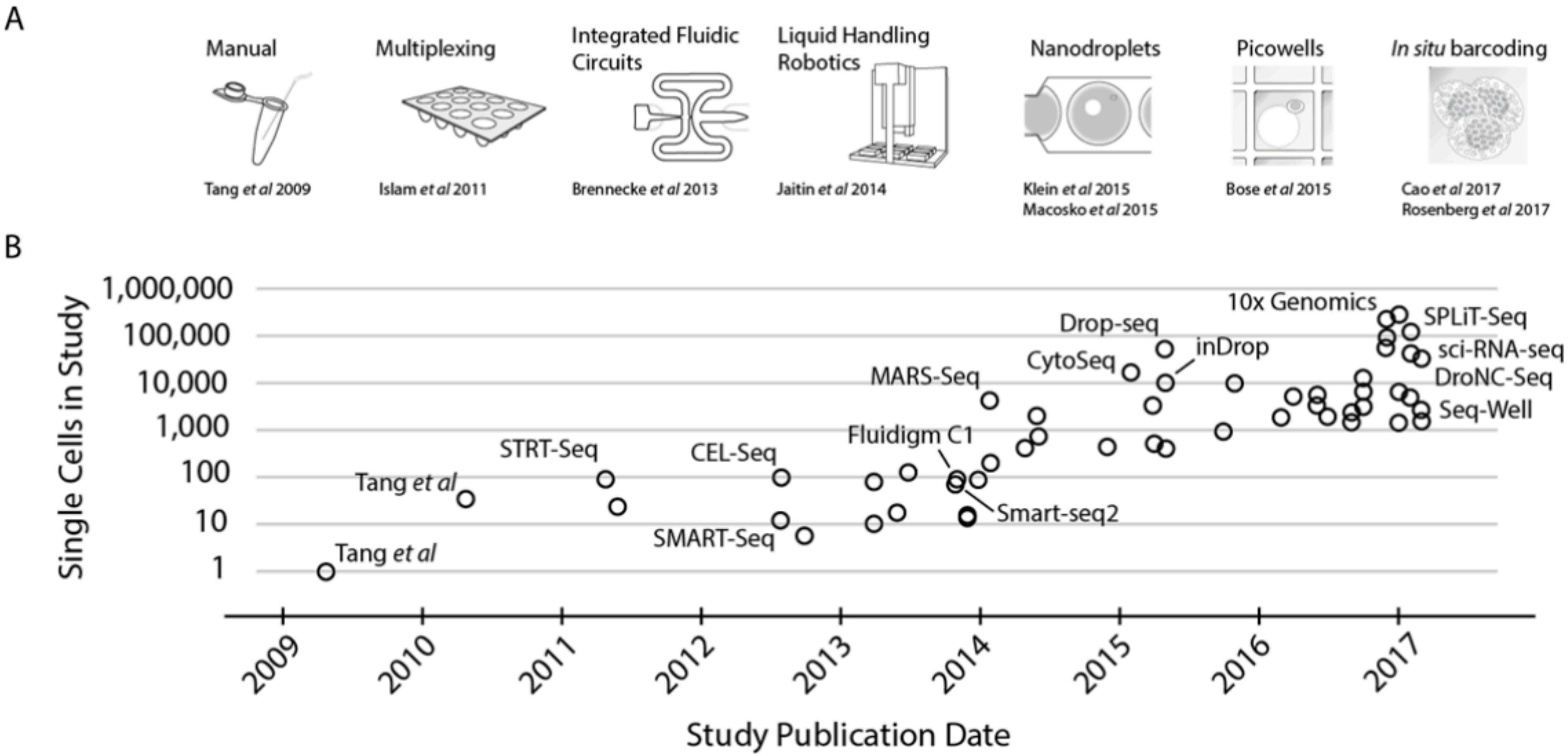
Major reasons to conduct single cell analysis

Bulk RNAseq, where you measure the ’average’ expression of all constituent cells, is sometimes insufficient for some experimental questions.

- Gene dynamics - what changes in gene expression effect different cell characteristics, such as during differentiation
- RNA splicing – cell to cell variation in alternative splicing
- Cell typing - genes expressed in a cell are used to identify types of cells. The main goal in cell typing is to find a way to determine the identity of cells that don't have known genetic markers.
- Spatial Transcriptomics – isolation of cells with known spatial location.

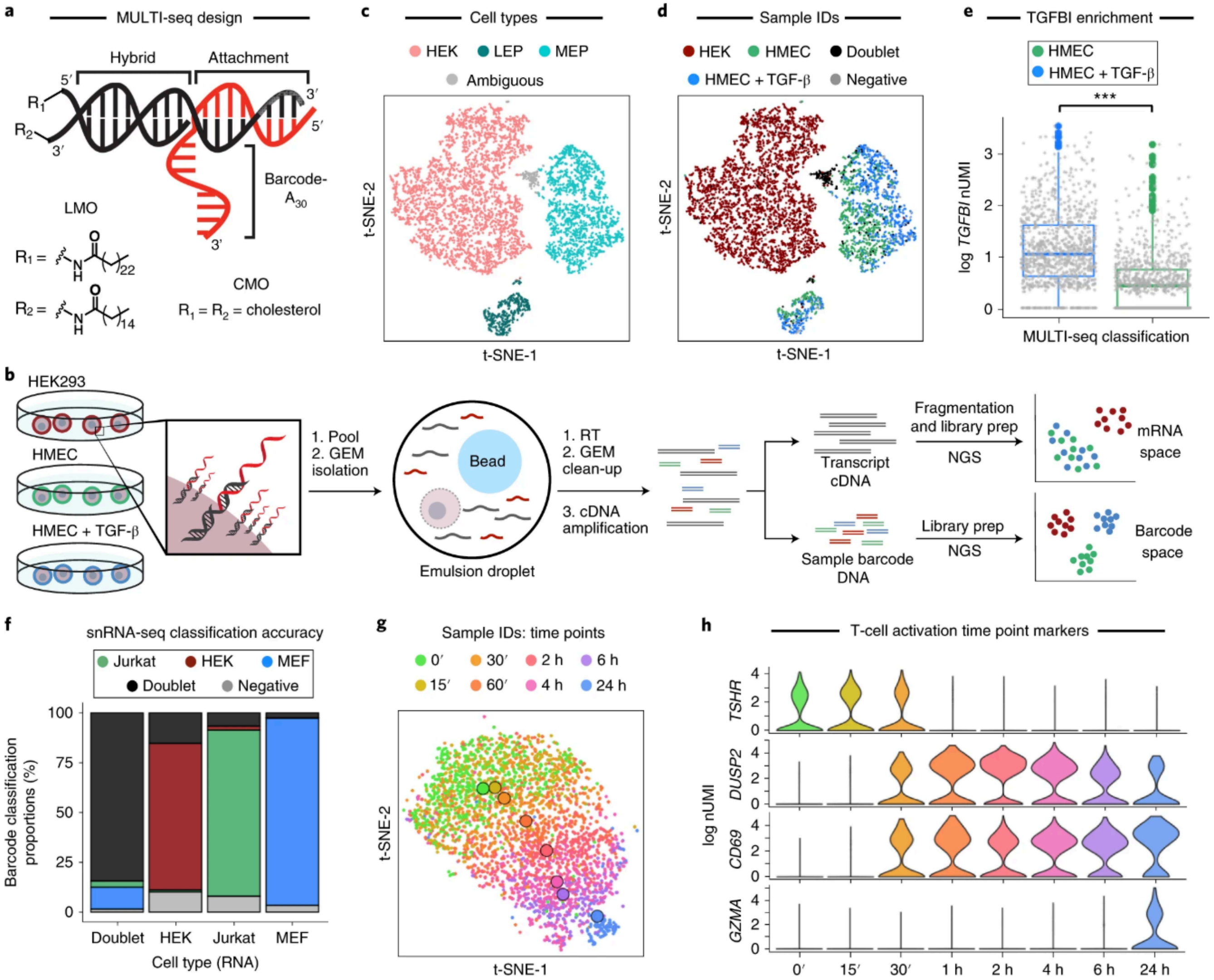
Exponential scaling of single-cell RNAseq in the last decade

https://arxiv.org/abs/1704.01379

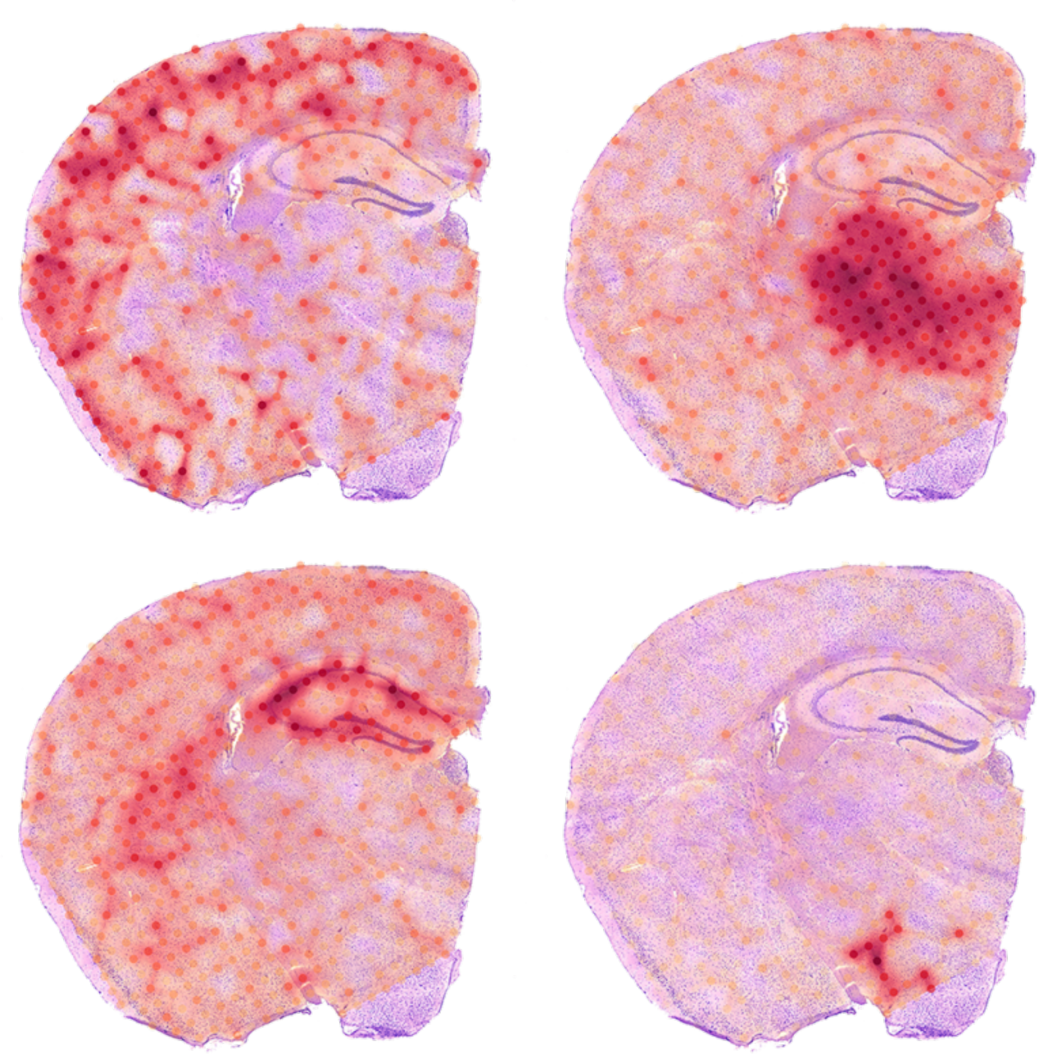


https://www.nature.com/articles/s41592-019-0433-8

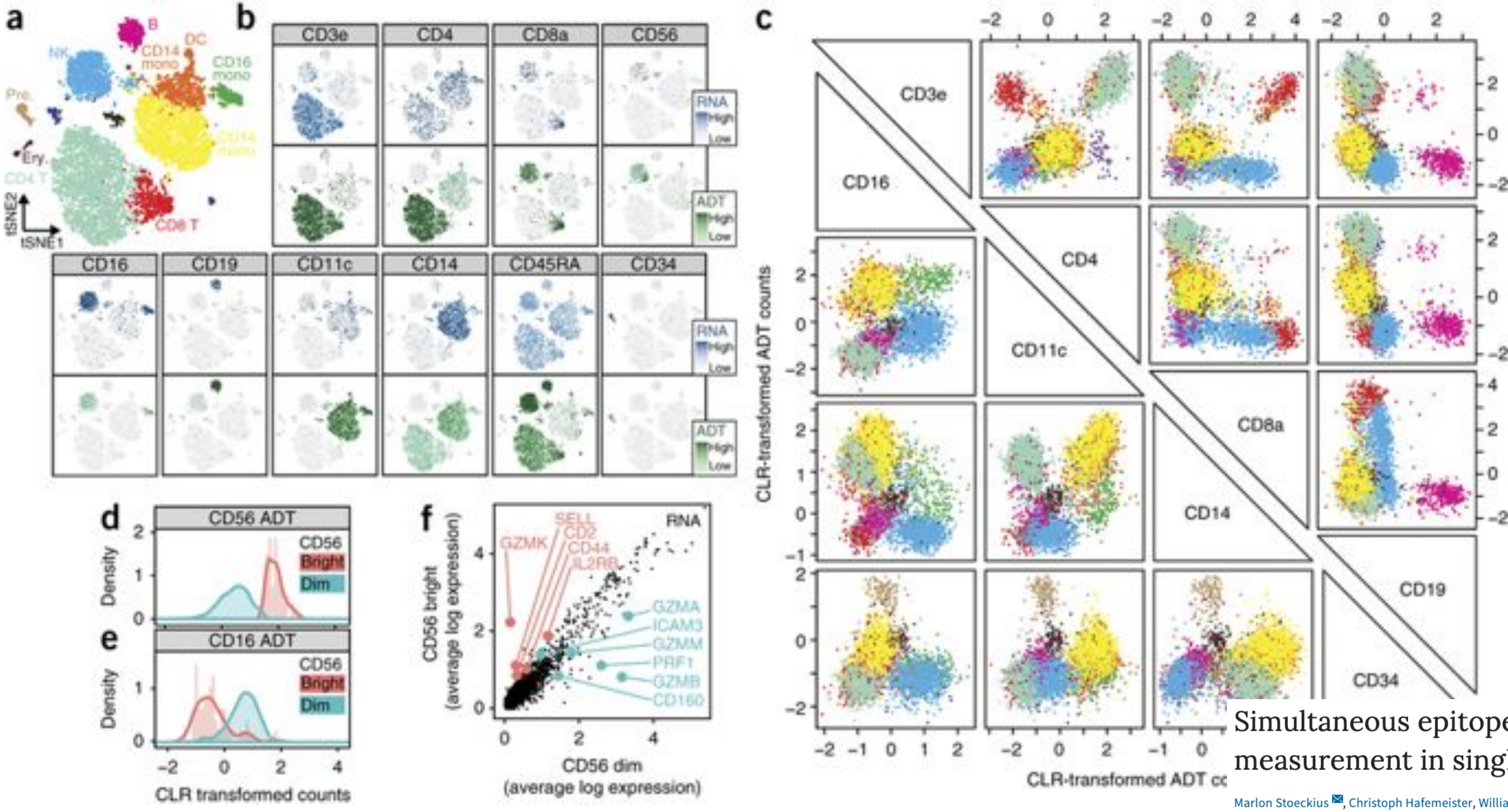
Multiplexing (MULTIseq)



Spatial Transcriptomics (and single cell)



Cite-seq and Epitope/transcriptome integration



Simultaneous epitope and transcriptome measurement in single cells

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija & Peter Smibert

Designing Experiments

Beginning with the question of interest (and working backwards)

- The final step of a DE analysis is the application of a linear 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 ~2200 cells to have a 90% chance of capturing at least 50 of those rare cells.

www.satijalab.org/howmanycells

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
- ~~Quality should be established using Fragment analysis traces (pseudo-gel images, RNA RIN > 7.0)~~
- ~~DNA/RNA should not be degraded~~
  - ~~260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable~~
- ~~Quantity should be determined with a Fluorometer, such as a Qubit.~~

Comparison to RNA-seq libraries

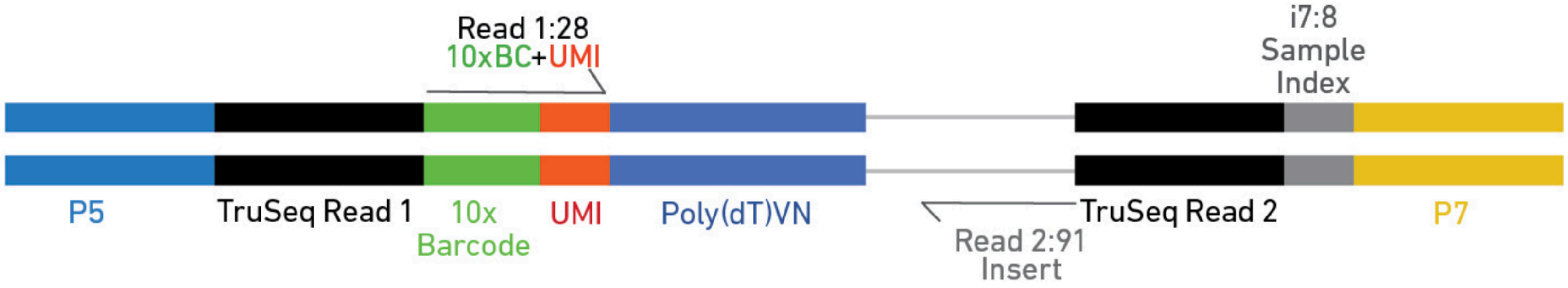
Considerations

- QA/QC of ~~RNA samples~~ Cells [Consistency across samples is most important.]

  'Cleanliness' of cells and accurate cell counts
- What is the RNA of interest [polyA extraction is pretty universal]
- Library Preparation
  - Stranded Vs. Unstranded [Standard is pretty universal]
- Size Selection/Cleanup [Target kit recommendations]
  - Final QA [Consistency across samples remains most important.]

Elements of a Library

- Library Barcode (Sample Index) - Used to pool multiple samples on one sequencing lane
- Cell Barcode (10x Barcode) – Used to identify the cell the read came from
- Unique Molecular Index (UMI) – Used to identify reads that arise during PCR replication
- Sequencing Reads – Used to identify the gene a read came from



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.

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, >> complexity -> the >> depth.
- Interest in measuring genes expressed at low levels, << level -> the >> depth.
- The fold change you want to be able to detect (< fold change more replicates and more 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.

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

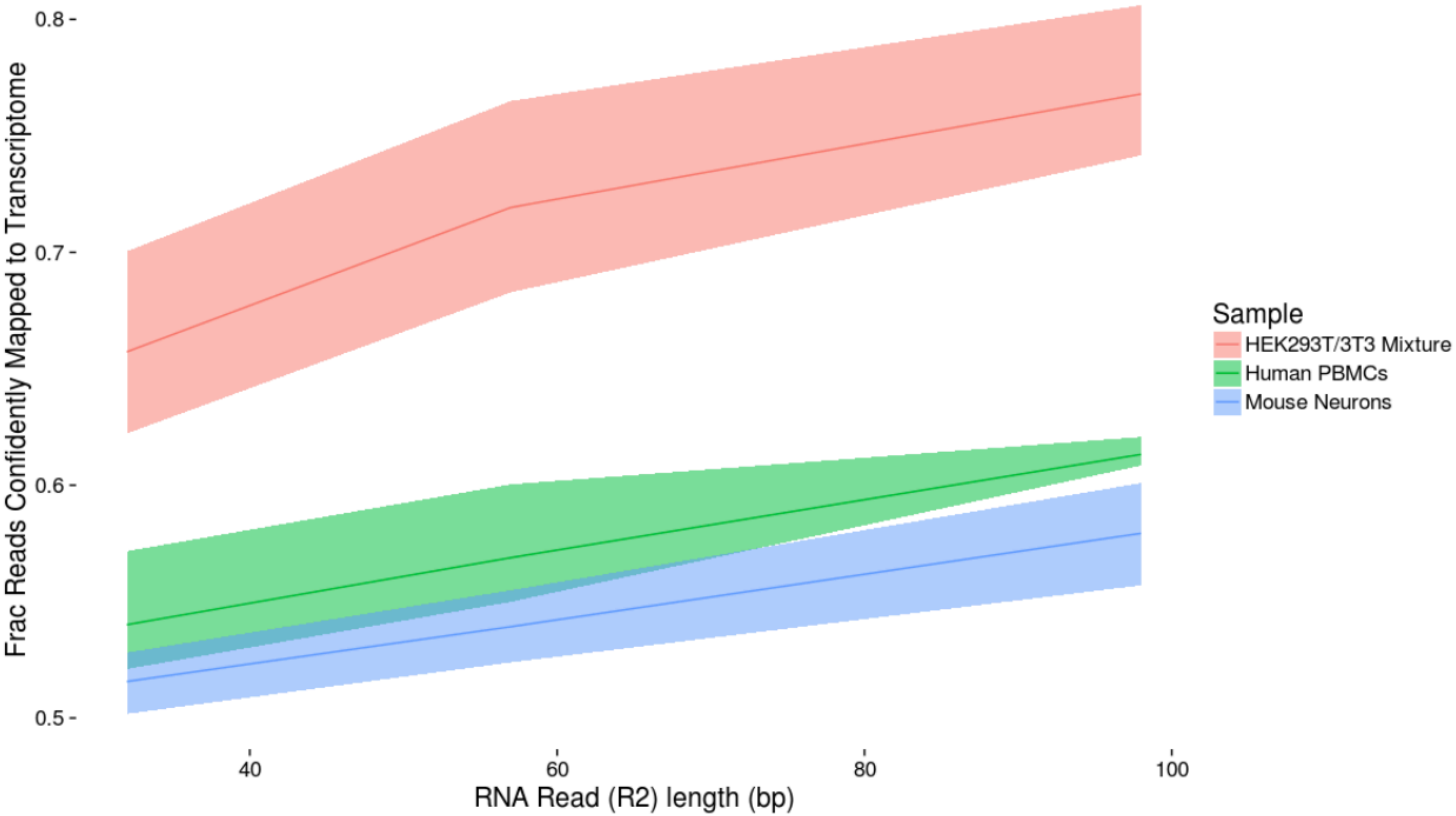
Sequence Read	Minimum Length	Read Description
Read 1	28bp (16bp bc, 12bp UMI)	barcode and UMI
I7 Index	8bp	Sample Index Read
Read2	100bp	Transcript Tag

**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.

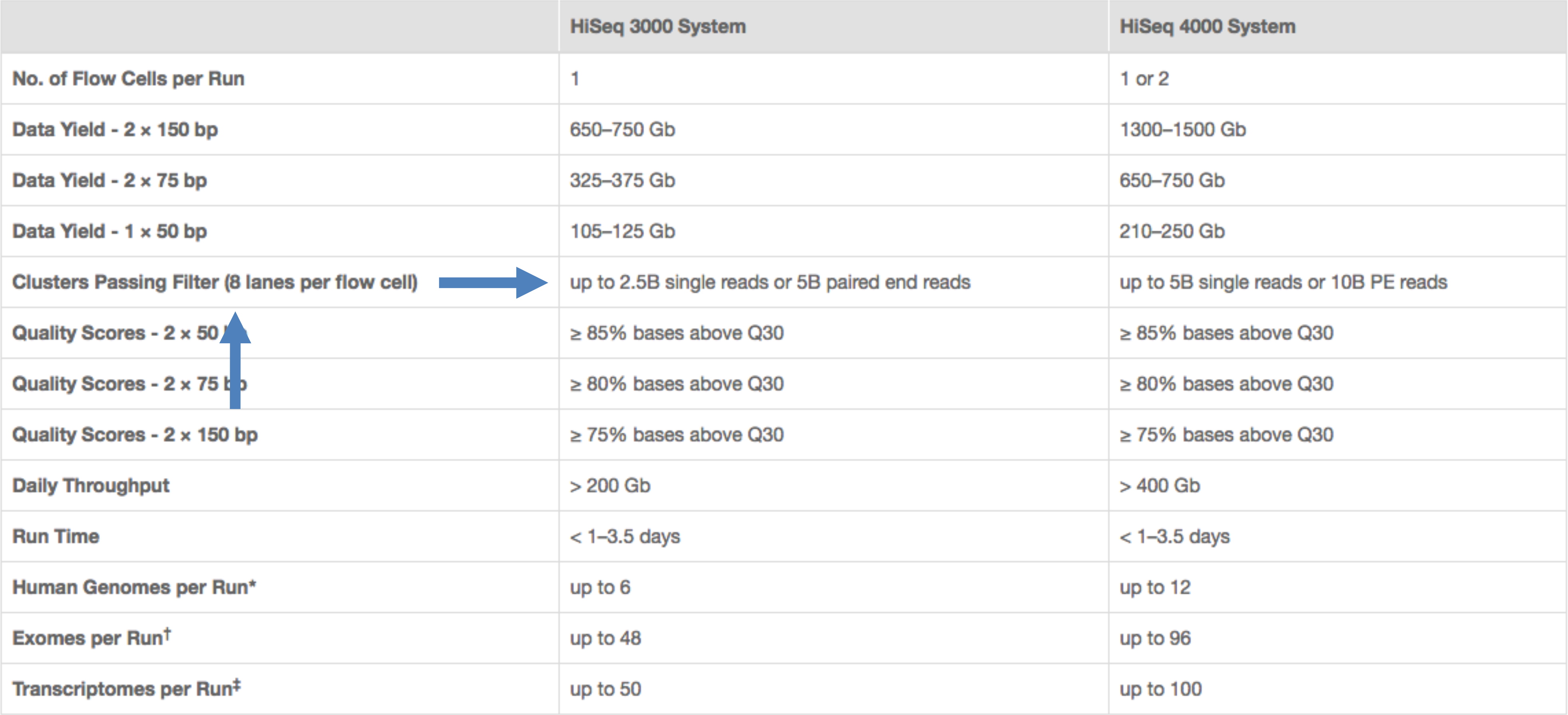
@ full capacity 10,000 cells per sample and 20K reads per cell = 200M reads or ~0.5 lanes/sample

UCDAVIS Bioinformatics Core

Read length matters (10x slide)



Illumina sequencing



	HiSeq 3000 System	HiSeq 4000 System
No. of Flow Cells per Run	1	1 or 2
Data Yield - 2 × 150 bp	650–750 Gb	1300–1500 Gb
Data Yield - 2 × 75 bp	325–375 Gb	650–750 Gb
Data Yield - 1 × 50 bp	105–125 Gb	210–250 Gb
Clusters Passing Filter (8 lanes per flow cell)	up to 2.5B single reads or 5B paired end reads	up to 5B single reads or 10B PE reads
Quality Scores - 2 × 50 bp	≥ 85% bases above Q30	≥ 85% bases above Q30
Quality Scores - 2 × 75 bp	≥ 80% bases above Q30	≥ 80% bases above Q30
Quality Scores - 2 × 150 bp	≥ 75% bases above Q30	≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run*	up to 6	up to 12
Exomes per Run†	up to 48	up to 96
Transcriptomes per Run‡	up to 50	up to 100

http://www.illumina.com/systems/hiseq-3000-4000/specifications.html

Cost Estimation

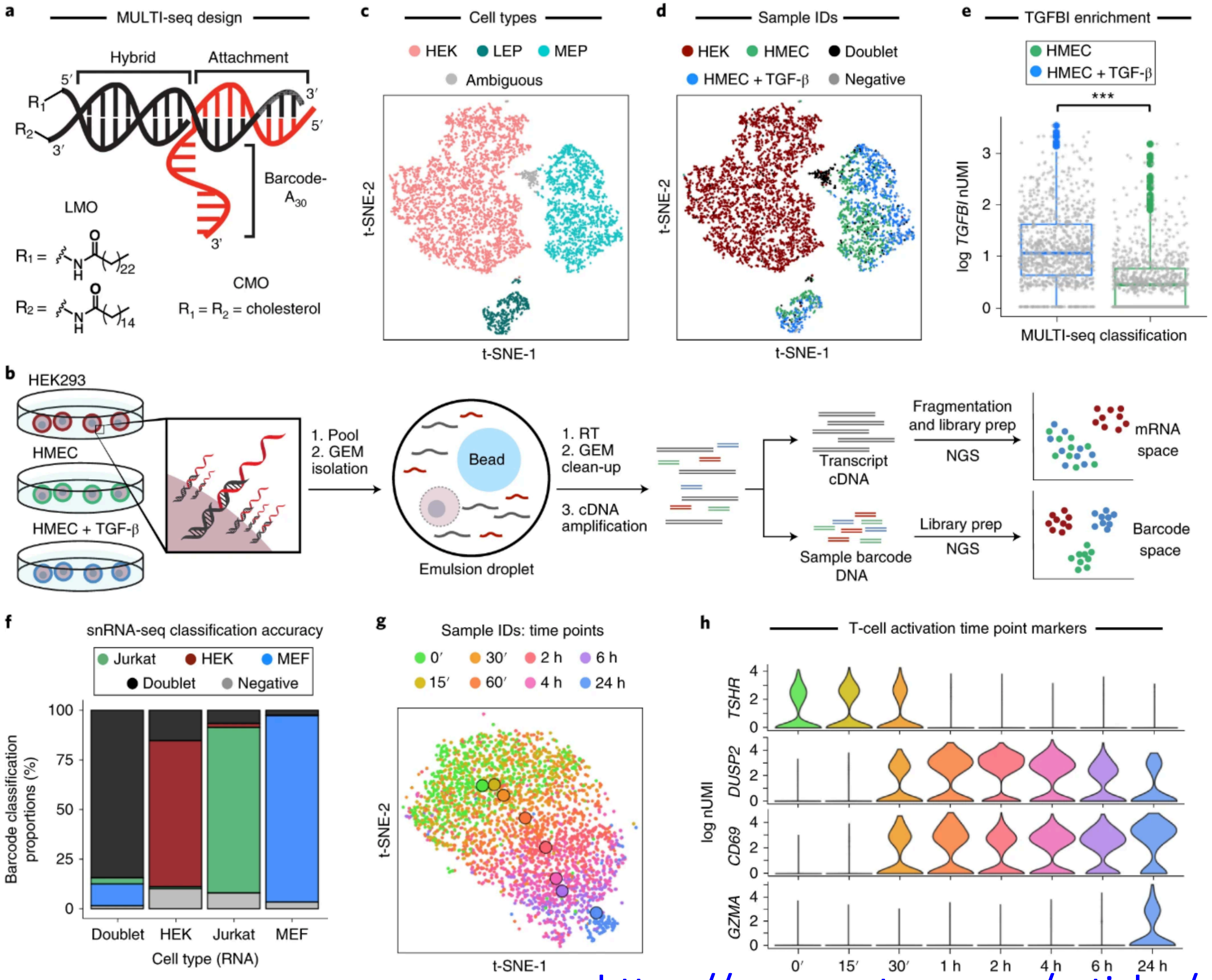
- Cell Isolation
- Library preparation (Per sample/pool)
- Sequencing (Number of lanes)
- Bioinformatics

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

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

http://bioinformatics.ucdavis.edu/services-2/

Multiplexing (MULTIseq)



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

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

The Bottom Line:
In Genomics

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.