UCDAVIS Bioinformatics Core

Single Cell Transcriptomics scRNAseq

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

UCDAVIS Bioinformatics Core

The **mission** of the Bioinformatics Core facility is to facilitate outstanding omics- scale research through these activities:

Data Analysis

The Bioinformatics Core promotes experimental design, advanced computation and informatics analysis of ‘omics’ scale datasets that drives research forward.

Research Computing

Maintain and make available high-performance computing hardware and software necessary for todays data-intensive bioinformatic analyses.

Training

The Core helps to educate the next generation of bioinformaticians through highly acclaimed training workshops, seminars and through direct participation in research activities.

UCDAVIS Bioinformatics Core

UC Davis Bioinformatics Core in the Genome Center

Core Facility Manager
Dr. Matthew Settles

Faculty Advisor
Dr. Ian Korf

Data Analysis Group

Genomics Bioinformatics

Dr. Joseph Fass
Dr. Monica Britton
Nikhil Joshi

Proteomics Bioinformatics

Metabolomics Bioinformatics

Dr. Jessie Li

Biostatistics

Dr. Blythe Durbin-Johnson

Undergraduate Assistants

Research Computing Group

System Administration

Michael Casper Lewis
Richard Feltstykket

Database/Web Programming

Adam Schaal

Undergraduate Assistant

UCDAVIS Bioinformatics Core

Contacts

- Website: http://bioinformatics.ucdavis.edu/
- Computing Issues, including but not limited to

  User account questions, equipment failure/malfunction, software install, software failures (not related to use)

  helpdesk@genomecenter.ucdavis.edu
- Bioinformatics related questions, including but not limited to

  bioinformatic methods questions, software use, data questions

  bioinformatics.core@ucdavis.edu
- DNA Technologies and Expression Analysis Core

  dnatech@ucdavis.edu

- Mailing lists: http://bioinformatics.ucdavis.edu/contact-us/

Goals

- Purpose – Experiments Conducted with Single Cell RNA Analysis
- Technologies and Libraries
- Bioinformatics as a Data Science
- Experimental Design
- Single Cell Analysis
  - Preprocessing
  - Mapping
  - Analysis
- Brief discussion on preparation of cell by Diana (DNA Tech Core)

UCDAVIS Bioinformatics Core

Purpose

Single-cell RNA Sequencing

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.

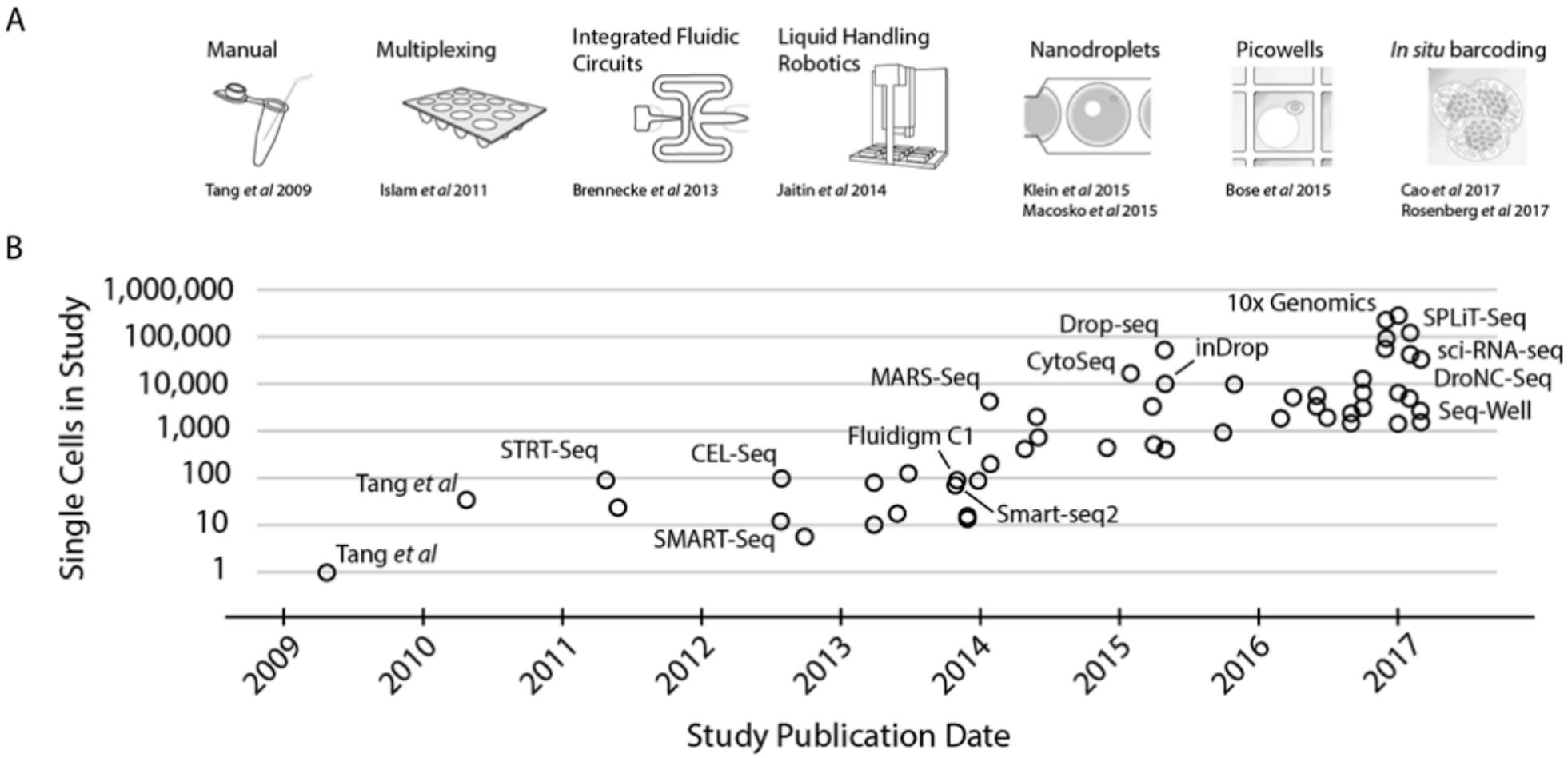
UCDAVIS Bioinformatics Core

Technology and Libraries

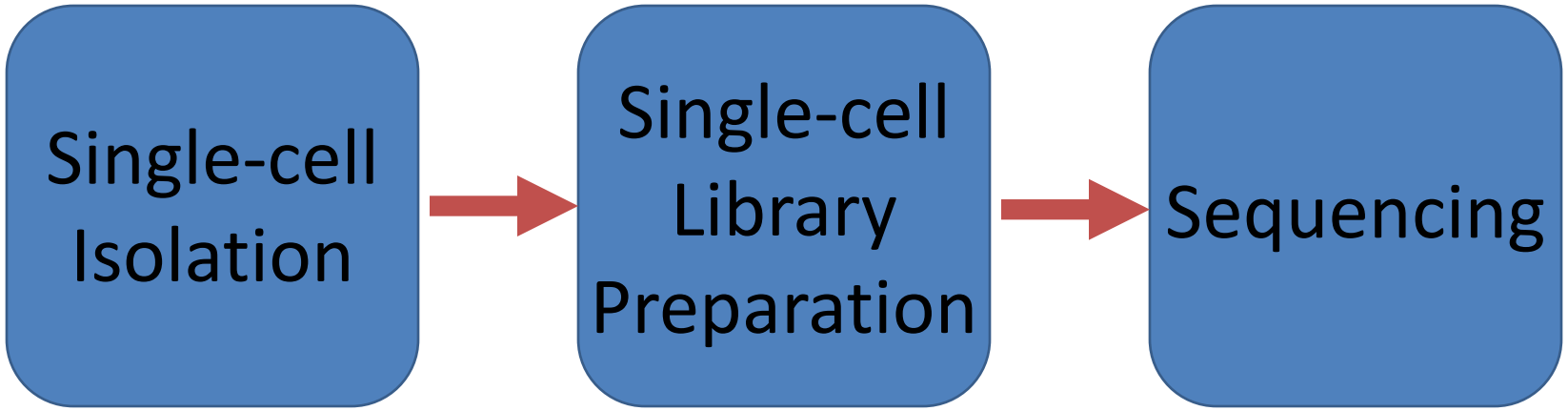
Single-cell RNA Sequencing

Exponential scaling of single-cell RNAseq in the last decade

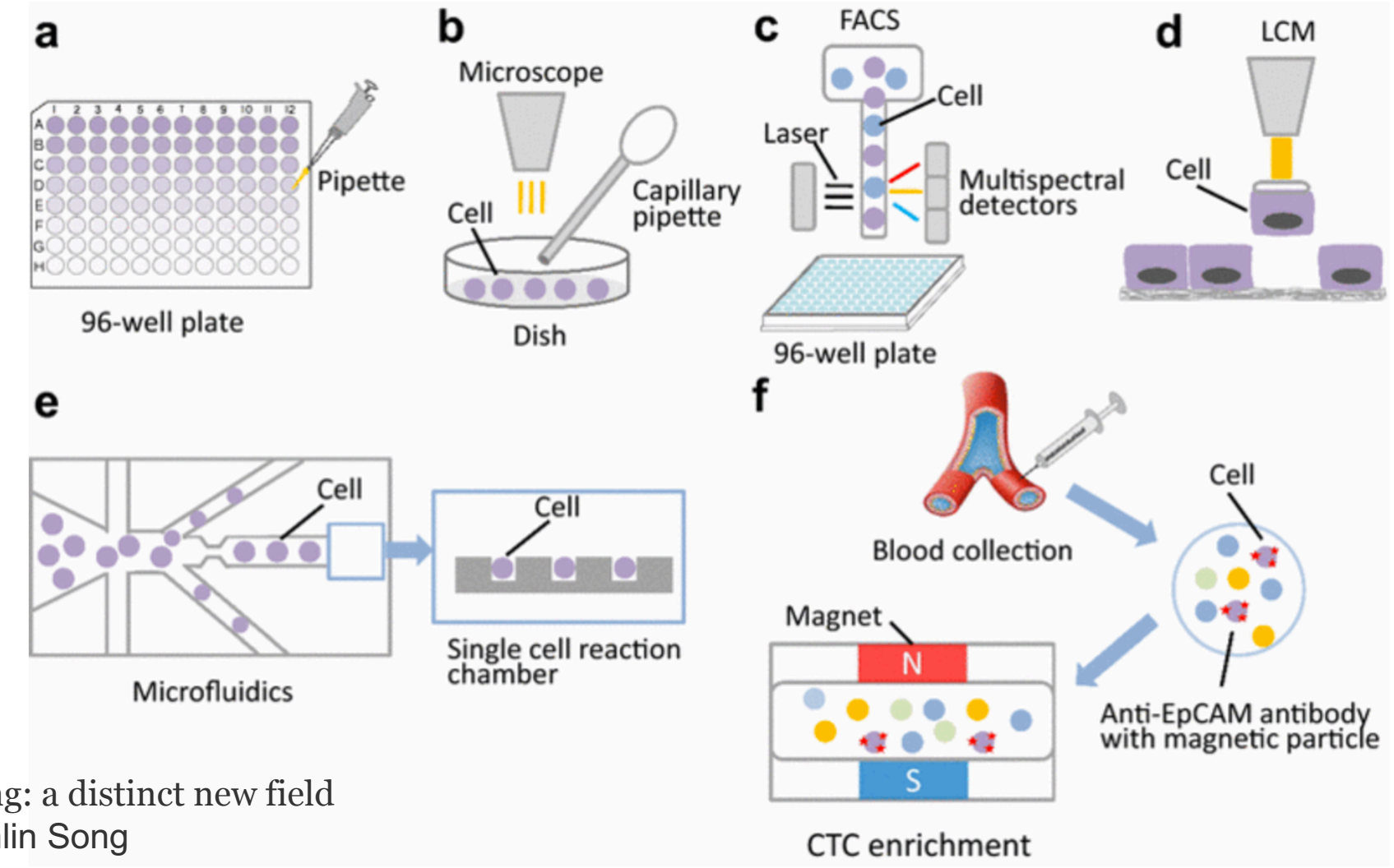
https://arxiv.org/abs/1704.01379



Generating Single-cell Data



Single-cell isolation methods



Single cell sequencing: a distinct new field
Jian Wang and Yuanlin Song

UCDAVIS Bioinformatics Core

Molecular Cell

Comparative Analysis of Single-Cell RNA Sequencing Methods

Authors

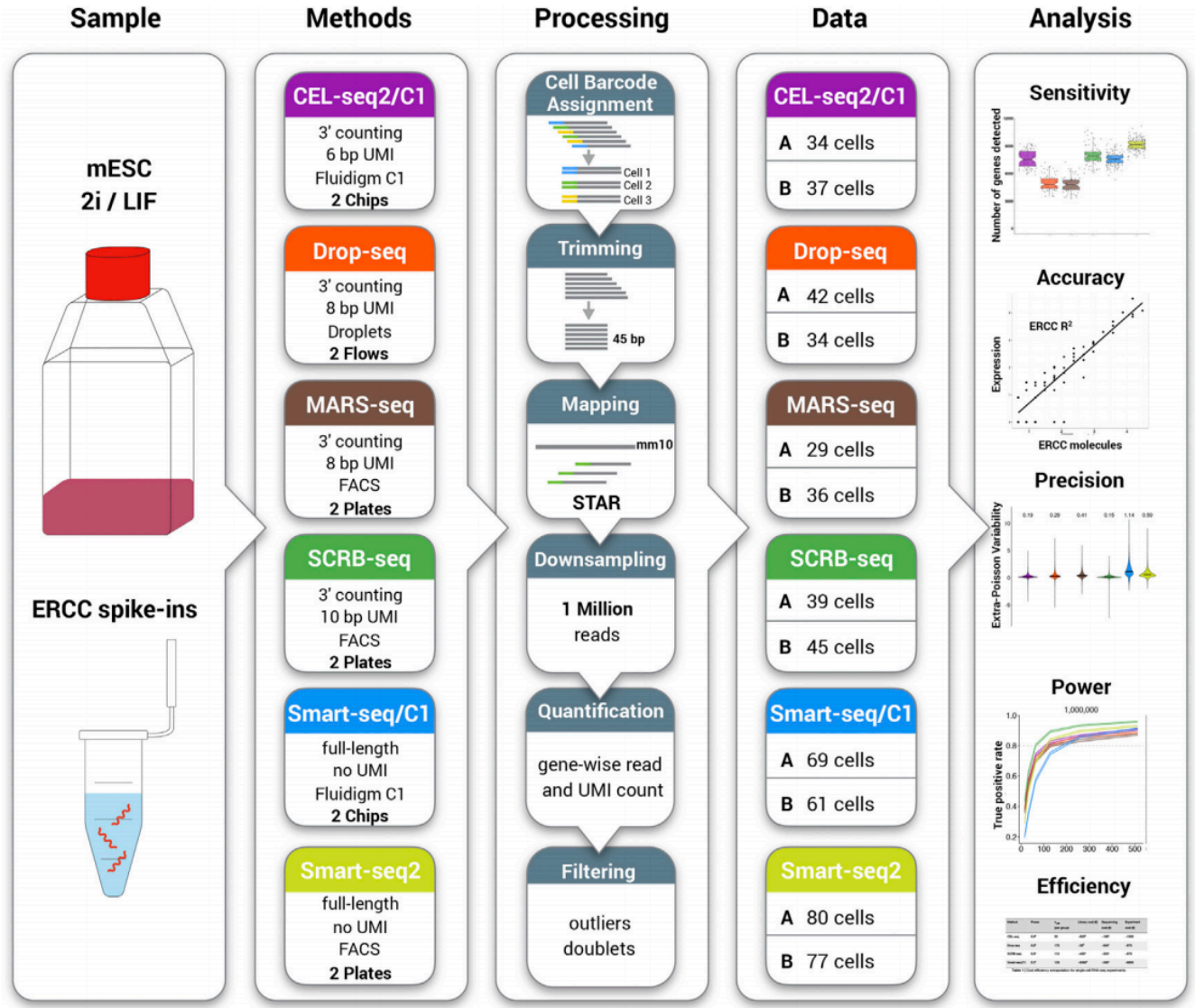
Christoph Ziegenhain, Beate Vieth, Swati Parekh, ..., Holger Heyn, Ines Hellmann, Wolfgang Enard

Commercial Platforms

- Fluidigm C1
- Wafergen ICELL8
- 10X Genomics Chromium

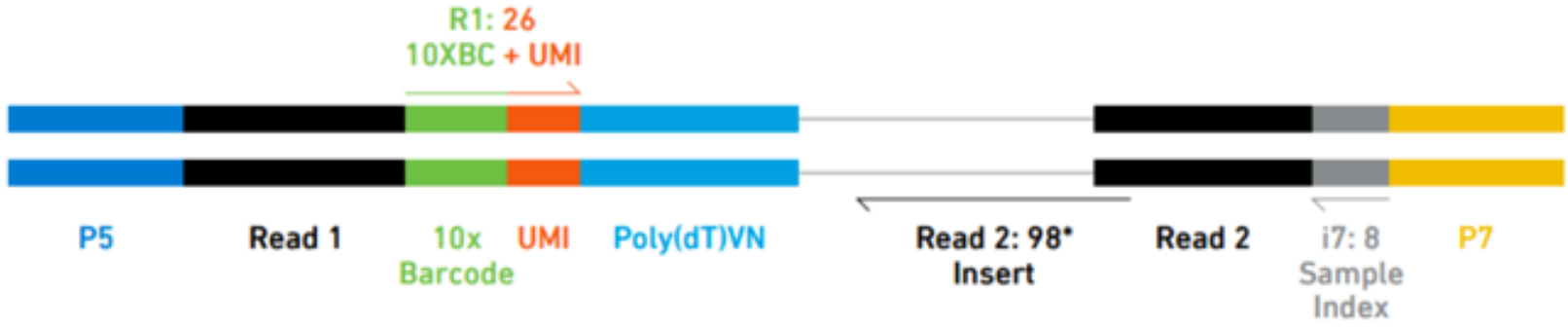
Commercial Kits

- BD Precise kits



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



3’ counting vs full-length

- 3’ counting techniques
  - 1 read per transcript
  - Based on polyA
  - Expression analysis only
  - Fewer reads per cell needed (~60K reads/cell)
  - Less noise in expression patterns
- Full-length
  - Based on polyA
  - Expression analysis
  - Splicing information
  - The more information desirec beyond expression, the higher the reads needed per cell (~60K reads/cell to 10M reads/cell)

Single-Cell with 10x genomics

10x Chromium Box

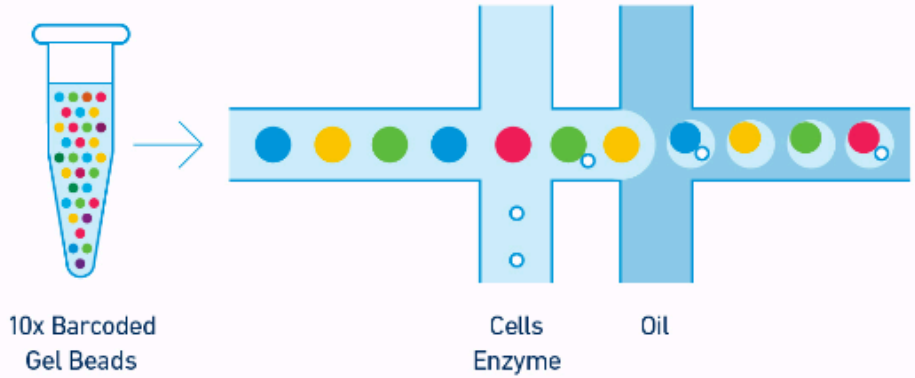
Basic Stats

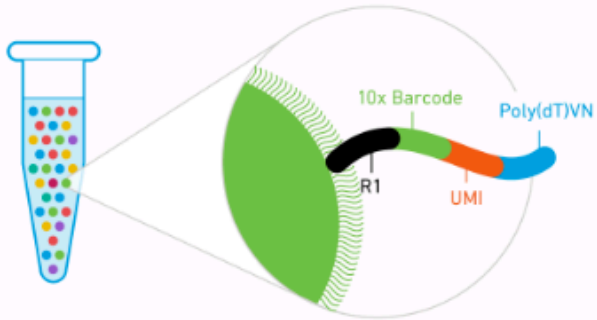
- Up to 8 channels processed in parallel
- 500 to 10,000 (V2) cells per channel
- 10 minute run time per chip
- Up to 30 um cell diameter tested
- ~50 % cell processing efficiency

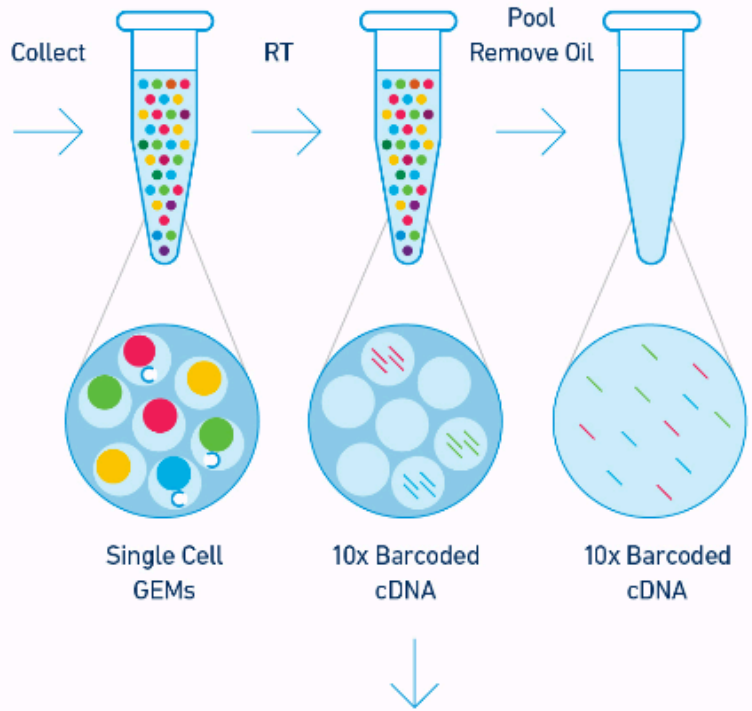
Number of cells	Expected Doublet Rate (%)
1,200	~1.2
3,000	~2.9
6,000	~5.7

Number of cells	Expected Doublet Rate (%)
500	~0.4
1,000	~0.8
3,000	~2.3
5,000	~3.9
10,000	~7.6

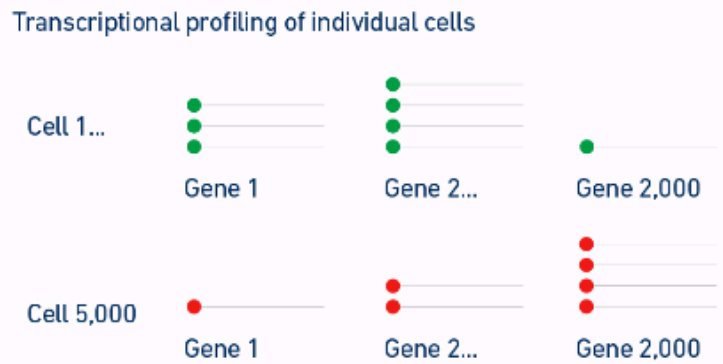
User controlled trade off between cell numbers and doublet rate





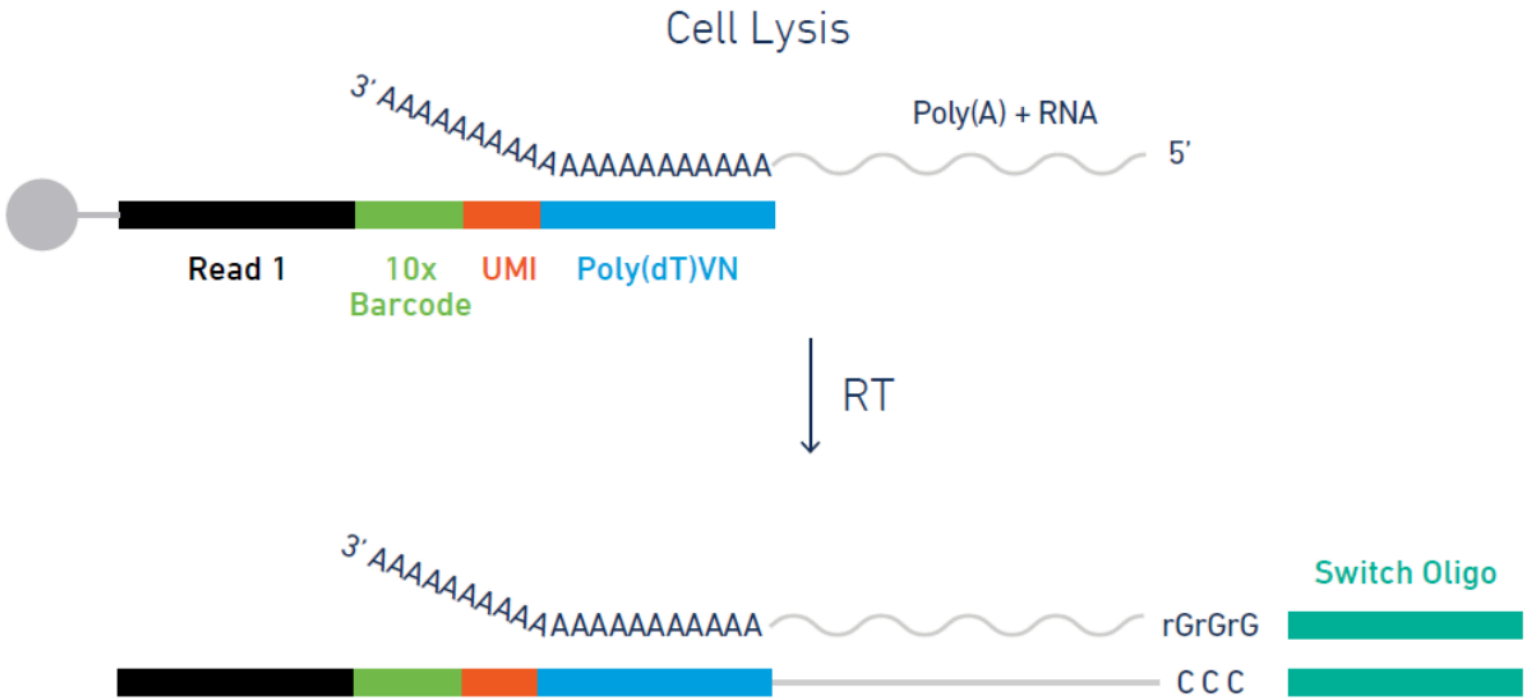


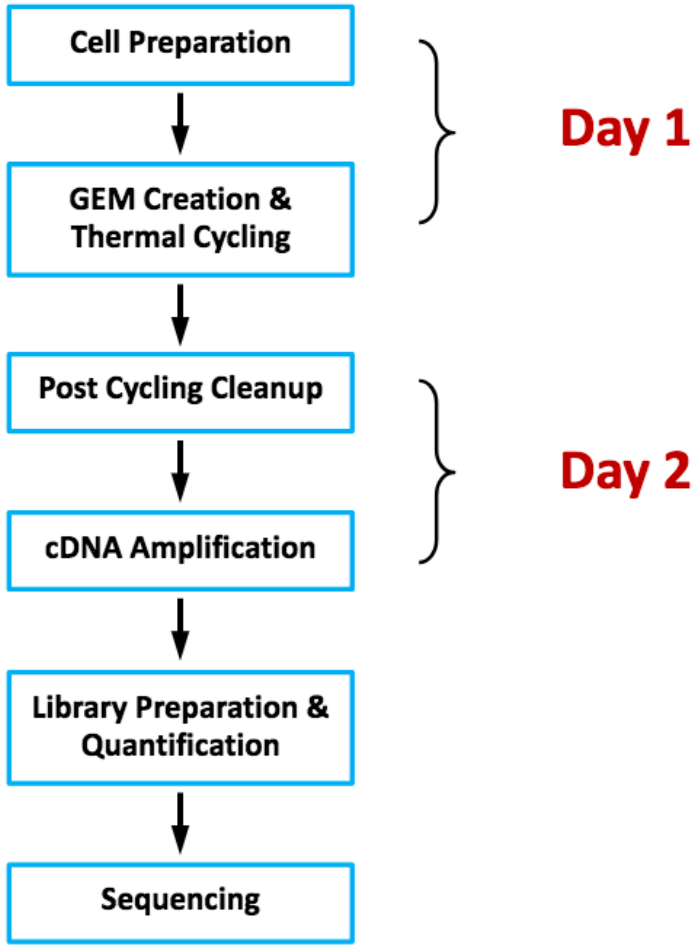
- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell



UCDAVIS Bioinformatics Core

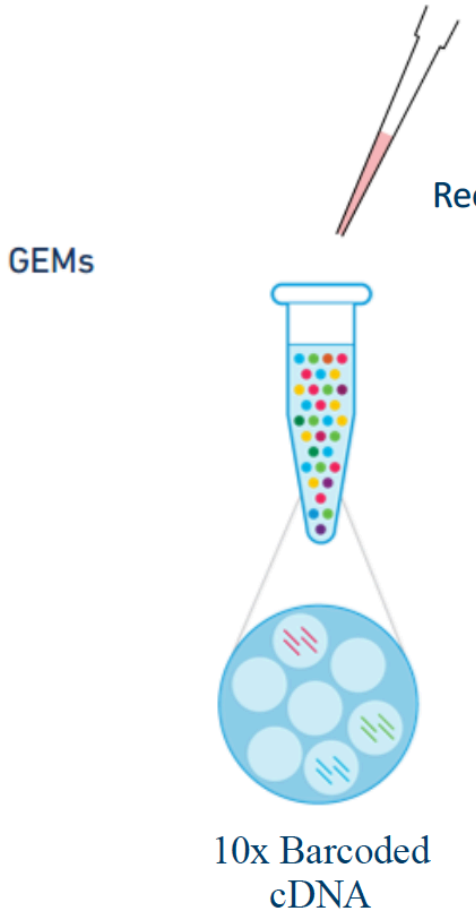
Basically a TAGseq protocol per cell 3’ expression



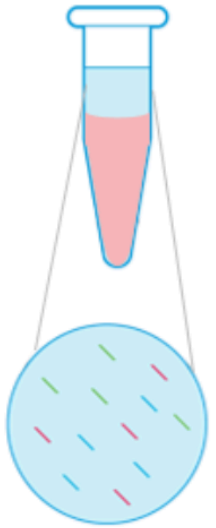


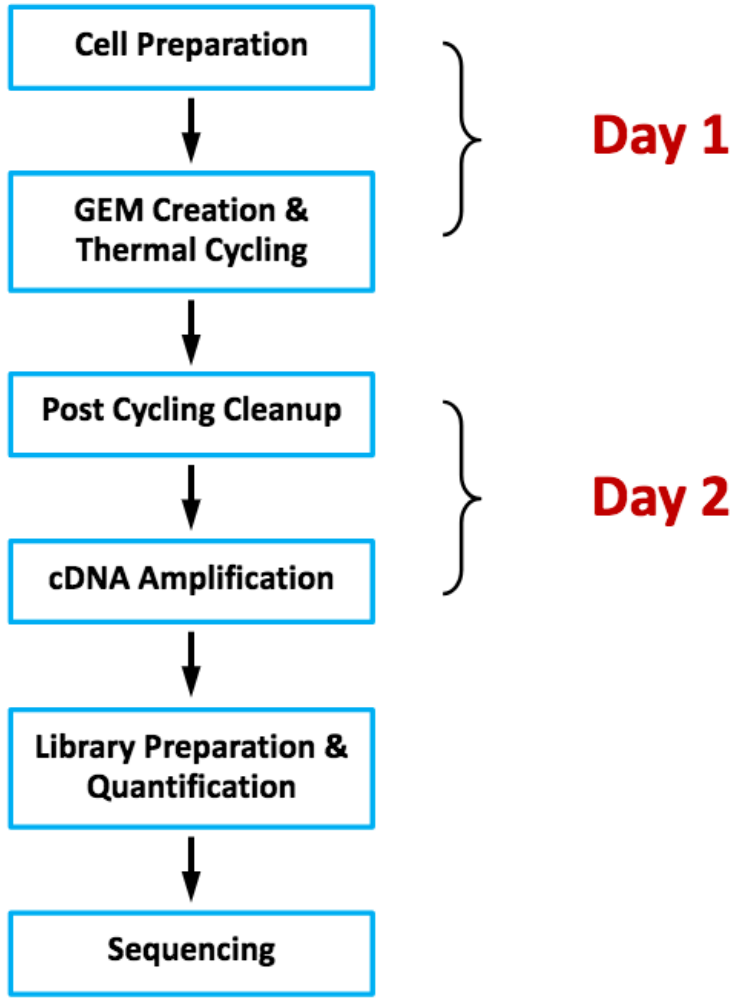
UCDAVIS Bioinformatics Core

Basically a TAGseq protocol per cell 3’ expression



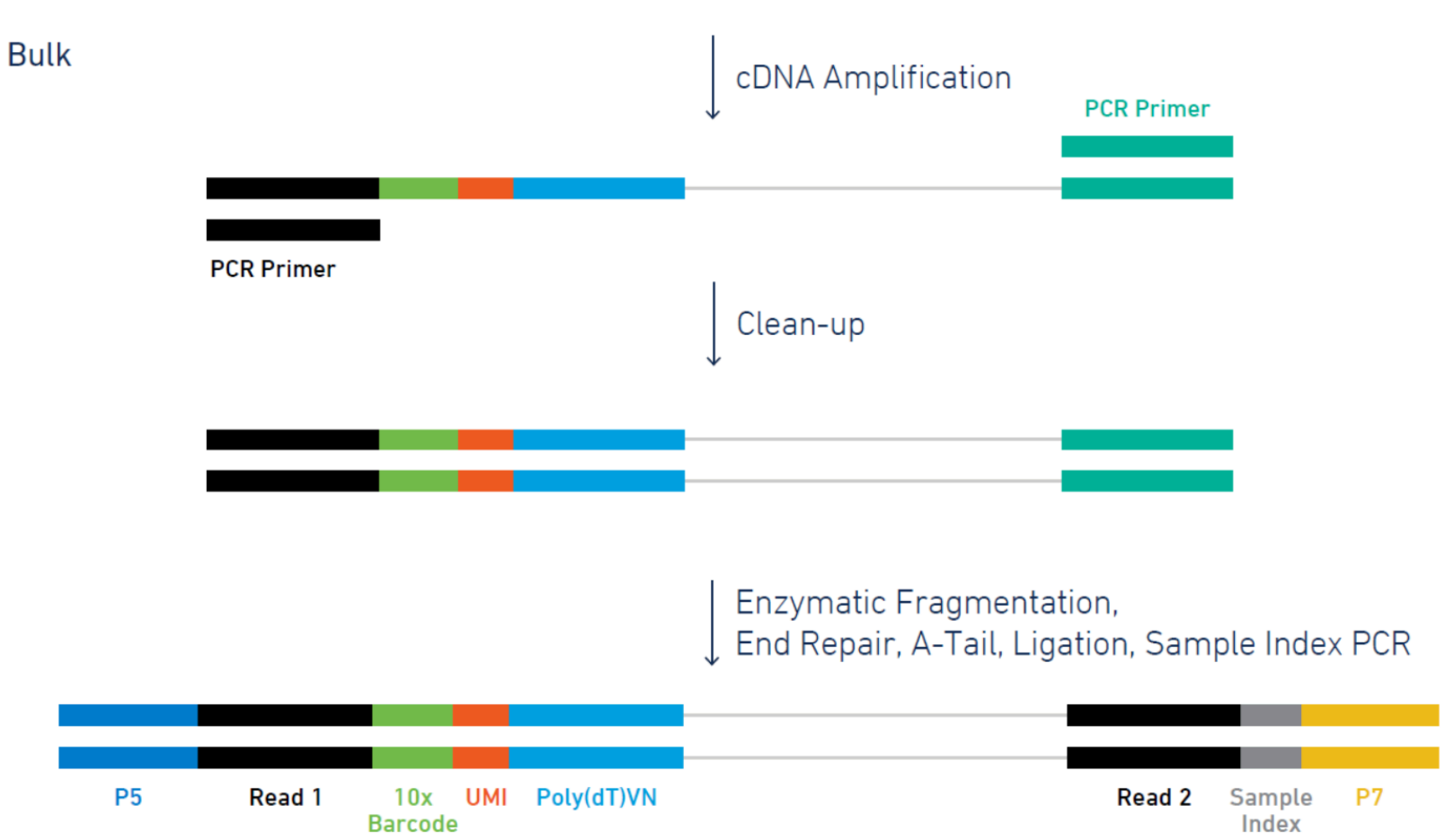
GEMs broken with Recovery Agent
Bulk cDNA used in remaining workflow steps

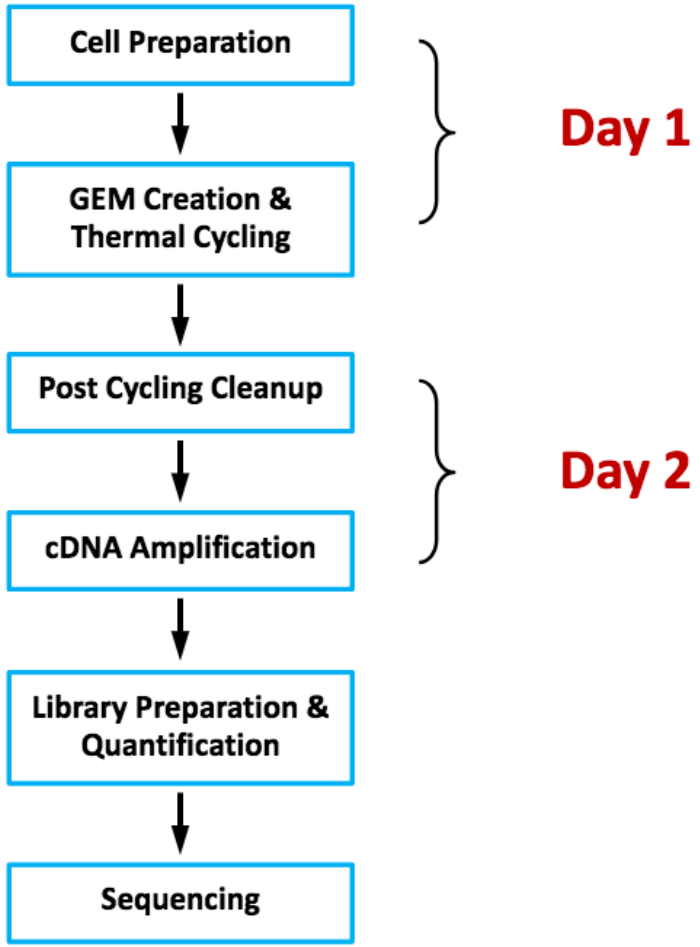




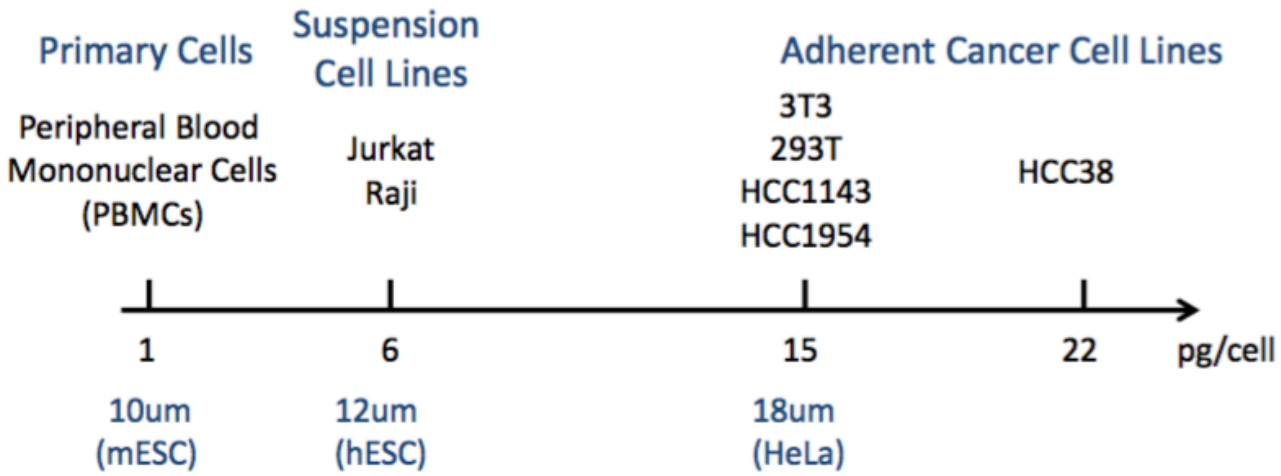
UCDAVIS Bioinformatics Core

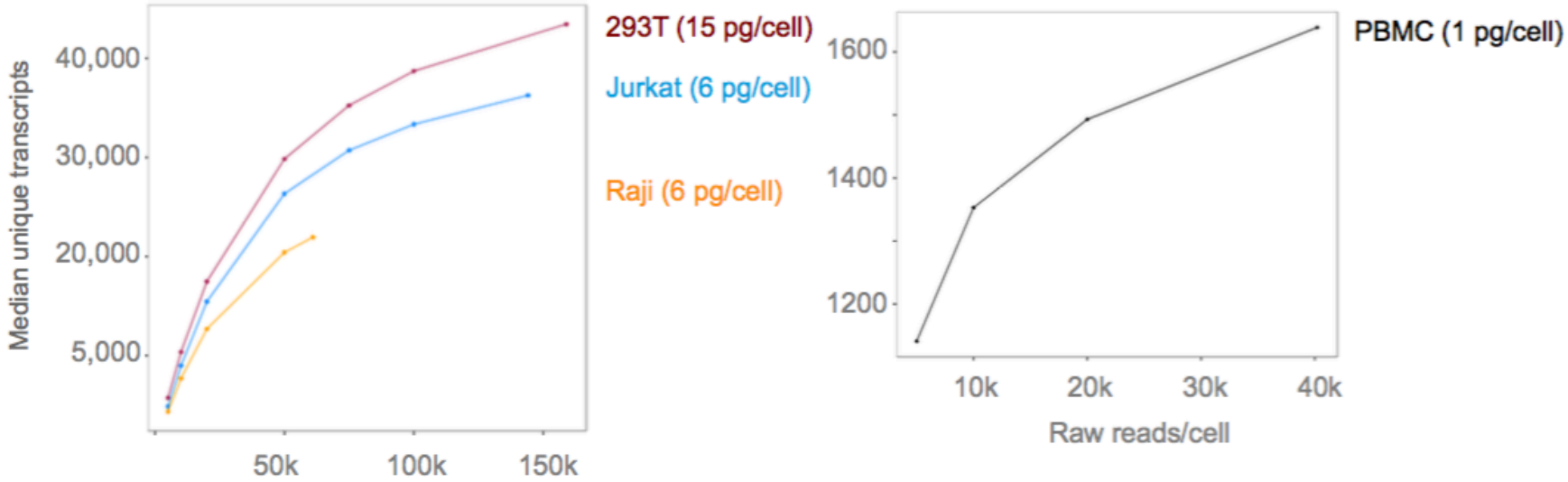
Basically a TAGseq protocol per cell 3’ expression





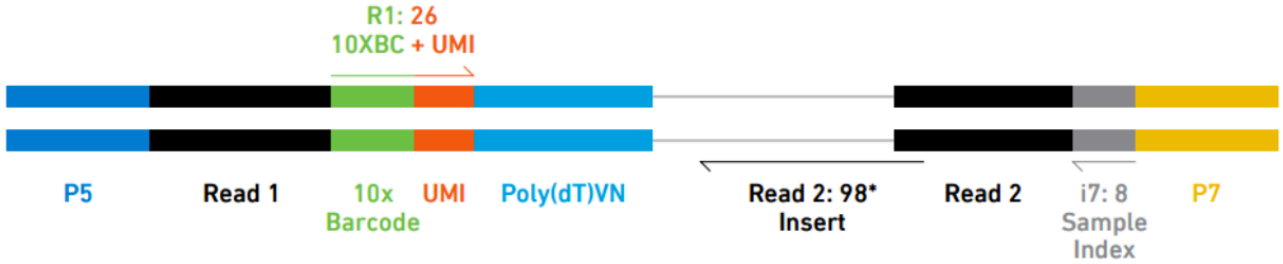
Cells of differing sizes and complexity





UCDAVIS Bioinformatics Core

Sequencing, V2



Recommendation

- 50,000 raw reads per cell is the recommended sequencing depth for 'typical' samples.
- 30,000 raw reads per cell is sufficient for RNA-poor cell types such as PBMCs.
- Given variability in cell counting/loading, extra sequencing may be required if the cell count is higher than anticipated.

Validated on

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

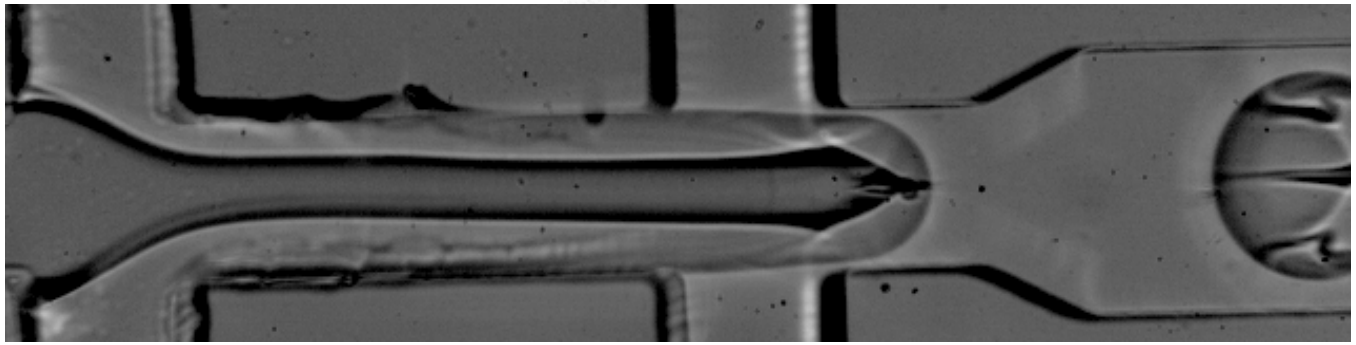
Typical sequencing run, with 3 reads, V2 kits

Sequence Read	Recommended Length	Read Description
Read 1	100bp (16bp bc, 10bp UMI)	10 barcode and UMI
I7 Index	8bp	Sample Index Read
Read2	100bp	Transcript Tag

@ full capacity 10,000 cells per sample and 50K reads per cell = 500M reads or ~1.25 lane/sample

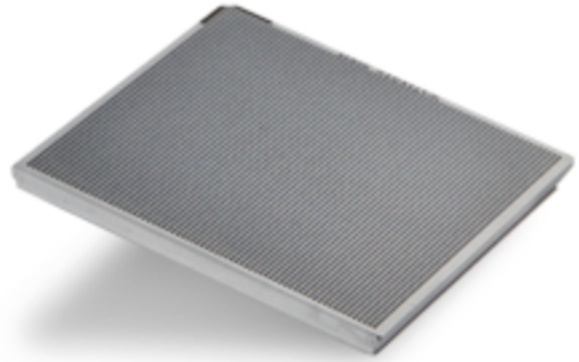
UCDAVIS Bioinformatics Core

Single-Cell with Drop-seq



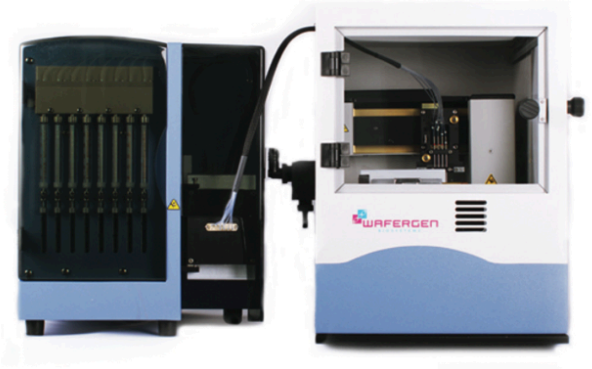
UCDAVIS Bioinformatics Core

Single-Cell with Wafergen

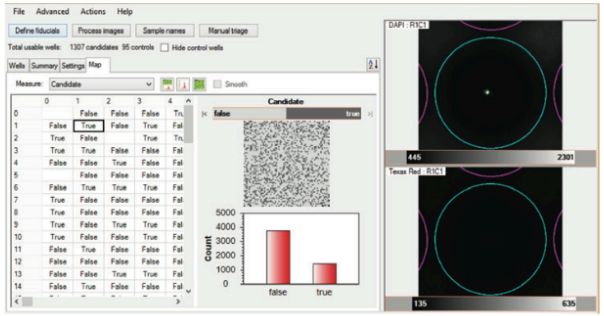


ICELL8 Chips and Reagents

Imaging Station



MultiSample NanoDispenser



CellSelect Software

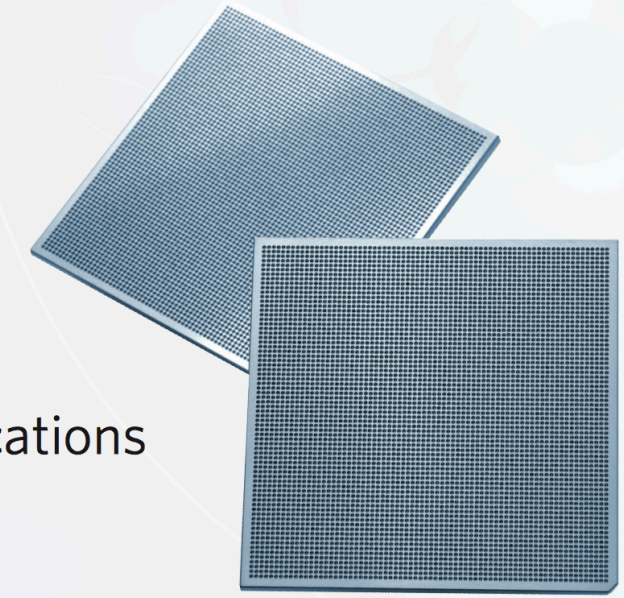
UCDAVIS Bioinformatics Core

Single-Cell with Wafergen

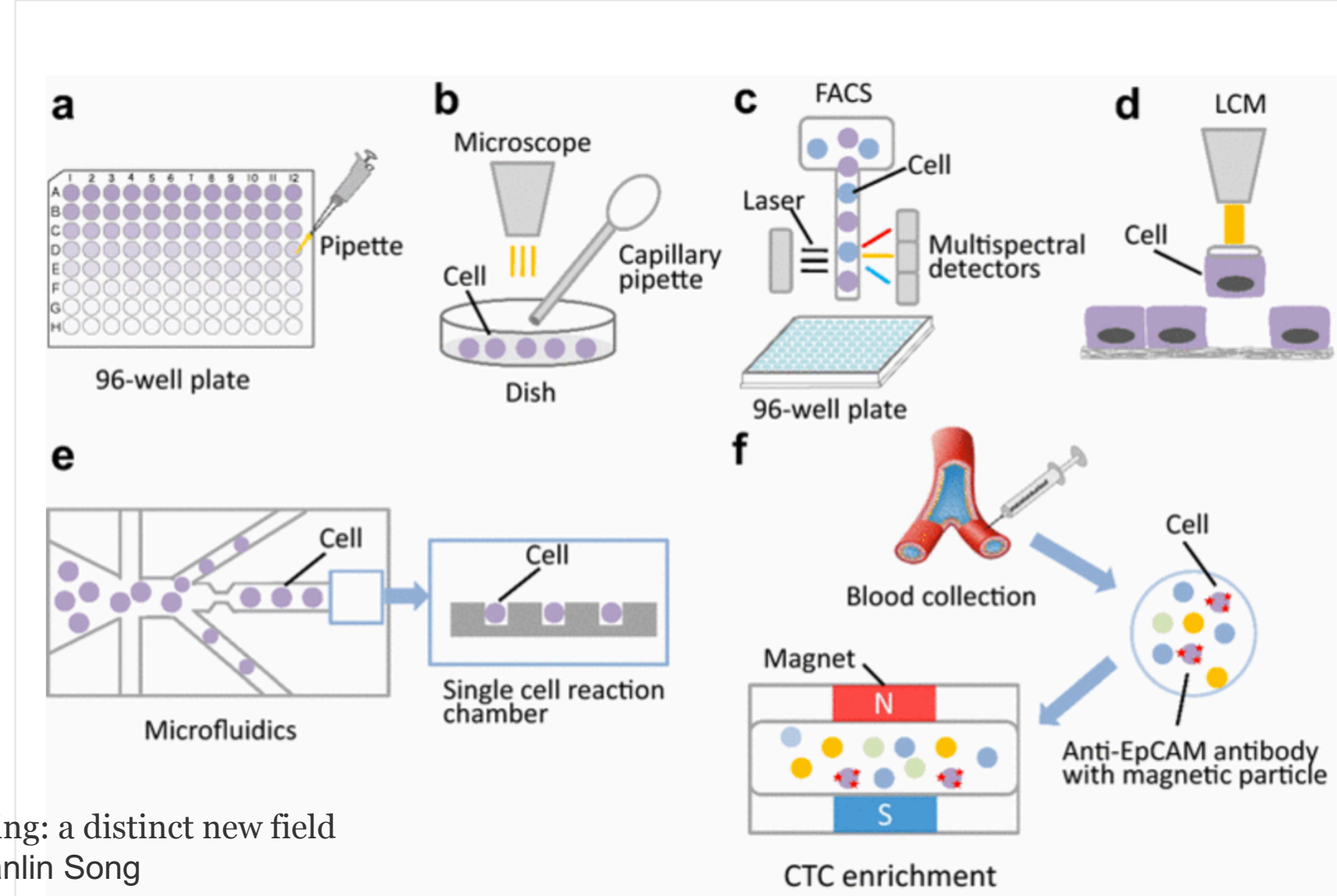
WAFERGEN BIOSYSTEMS

REVOLUTIONARY NEW SINGLE-CELL PLATFORM

1. Isolate up-to 1,800 cells per chip
2. Evaluate cells from 5-100 µm per sample
3. Select specific cells for downstream applications
4. Discover unique populations of cells



Which to Choose?



Single cell sequencing: a distinct new field
Jian Wang and Yuanlin Song

UCDAVIS Bioinformatics Core

Experimental Design

Single-Cell RNA Sequencing

Treating Bioinformatics as a Data Science

Data science done well looks easy and that's a big problem for data scientists

simplystatistics.org
March 3, 2015 by Jeff Leek

Bad Data Science (Bioinformatics) also looks easy

Treating Bioinformatics as a Data Science

1. Define the question of interest — Keep the end in mind
2. Get the data
3. Clean the data
4. Explore the data — 80% of work
5. Fit statistical models — Keep it simple
6. Communicate the results
7. Make your analysis reproducible — Remember this is Science validate

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.

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.]

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

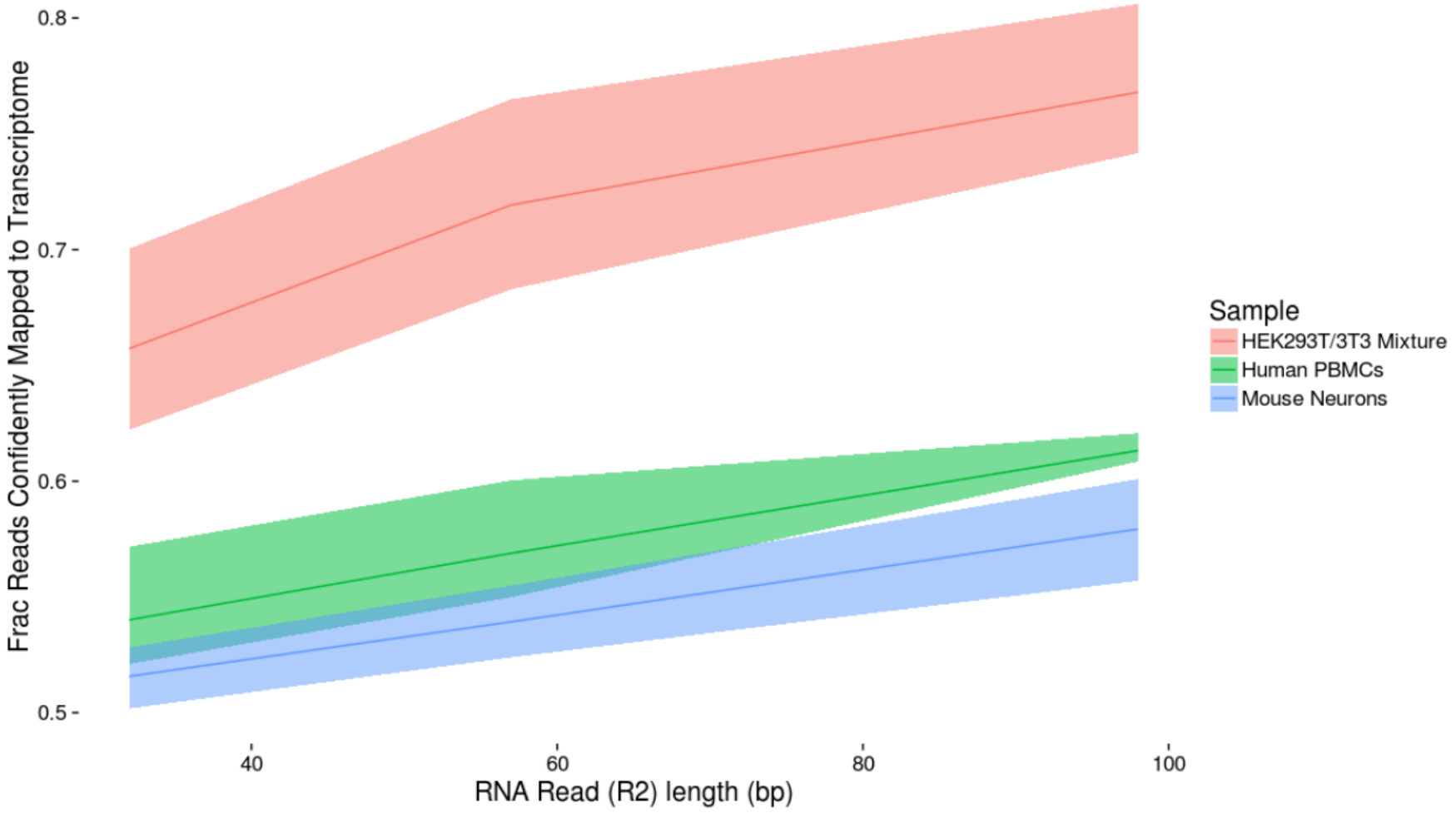
Factors to consider are:

- 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 requires >> sequencing depth.

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)



Illumina sequencing

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

2500

MiSeq

	HISEQ 3000 SYSTEM	HISEQ 4000 SYSTEM
No. of Flow Cells per Run →	1	1 or 2
Data Yield: ← 2 × 150 bp 2 × 75 bp 1 × 50 bp	 650-750 Gb 325-375 Gb 105-125 Gb	 1300-1500 Gb 650-750 Gb 210-250 Gb
Clusters Passing Filter (Single Reads) (8 lanes per flow cell) →	2.1-2.5 billion	4.3-5 billion
Quality Scores: ↑ 2 × 50 bp 2 × 75 bp 2 × 150 bp	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30	 ≥ 85% bases above Q30 ≥ 80% 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

Cost Estimation

- Cell Isolation
- Library preparation (Per sample/cell)
- Sequencing (Number of lanes)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

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

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

Cost Estimation

- 12 Samples
  - QA Bioanalyzer = $98 for all 12 samples
  - Library Preparation (ribo-depletion) = $383/sample = $4,596
- Sequencing = $2346 per lane
  - 2.1 - 2.5 Billion reads per run / 8 lanes = Approximately 300M reads per lane
  - Multiplied by a 0.8 buffer equals 240M expected good reads
  - Divided by 12 samples in the lane = 20M reads per sample per lane.
  - Target 30M reads means 2 lanes of sequencing $2346 x 2 = $4692
- Bioinformatics, simple pairwise comparison design, DE only $2000
  - This is the most basic analysis, for in depth collaborative analysis double sequencing budget.

Total = $98 + $4596 + $4692 + $2000 = $11,386

Approximately $950 per sample @ 40M reads per sample

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

UCDAVIS Bioinformatics Core

Single cell Analysis

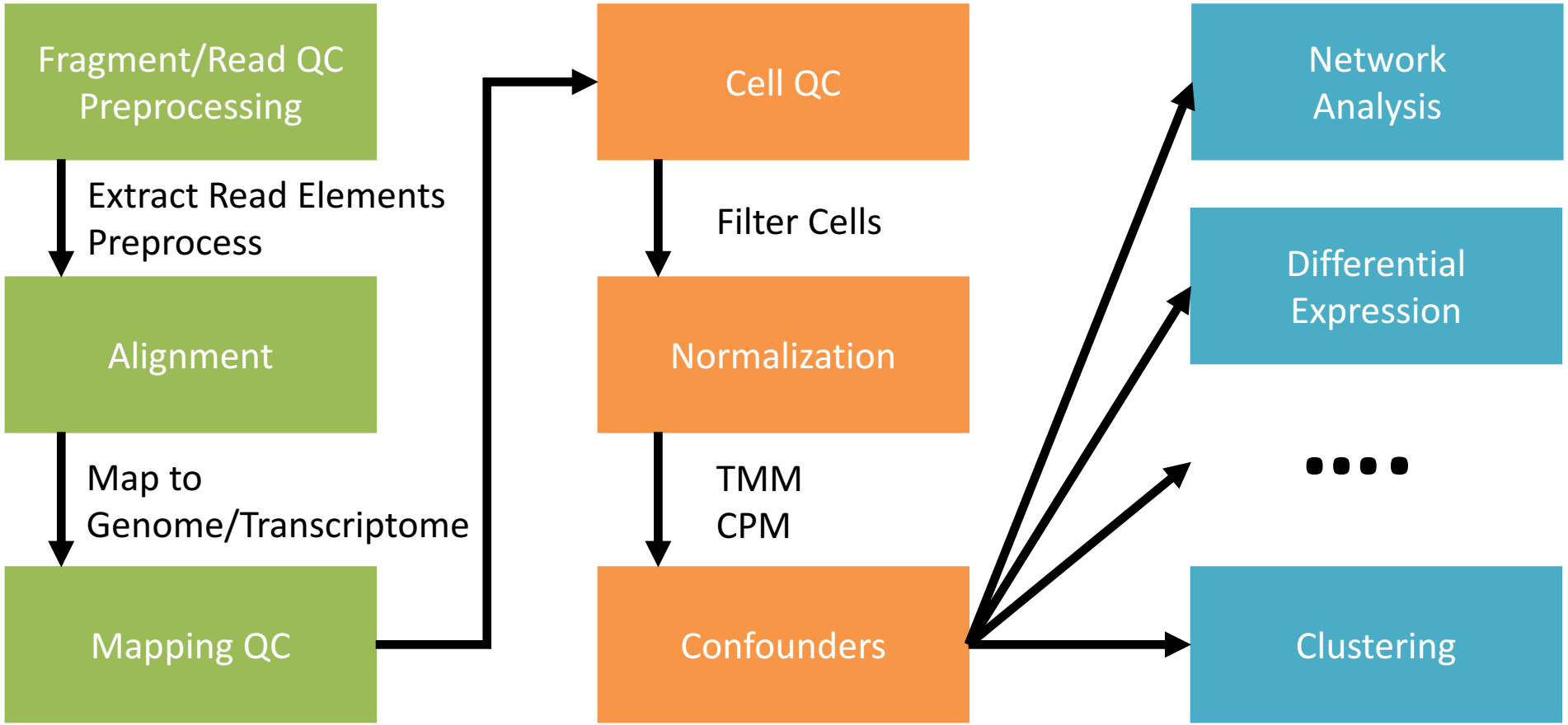
Single-cell RNA Sequencing

Prerequisites

- Access to a multi-core (24 cpu or greater), ‘high’ memory 64Gb or greater Linux server.
- Familiarity with the ’command line’ and at least one programming language.
- Basic knowledge of how to install software
- Basic knowledge of R (or equivalent) and statistical programming
- Basic knowledge of Statistics and model building

UCDAVIS Bioinformatics Core

Analysis Pipeline



Amount of reads mapping to rRNA/tRNAs
Proportion of uniquely mapping reads
Multimappers, Etc.

Batch Effects

Cell Barcode and UMI filtering

- Cell barcodes
  - Must be on static list of known cell barcode sequences
  - May be 1 mismatch away from the list if the mismatch occurs at a low-quality position (the barcode is then corrected).

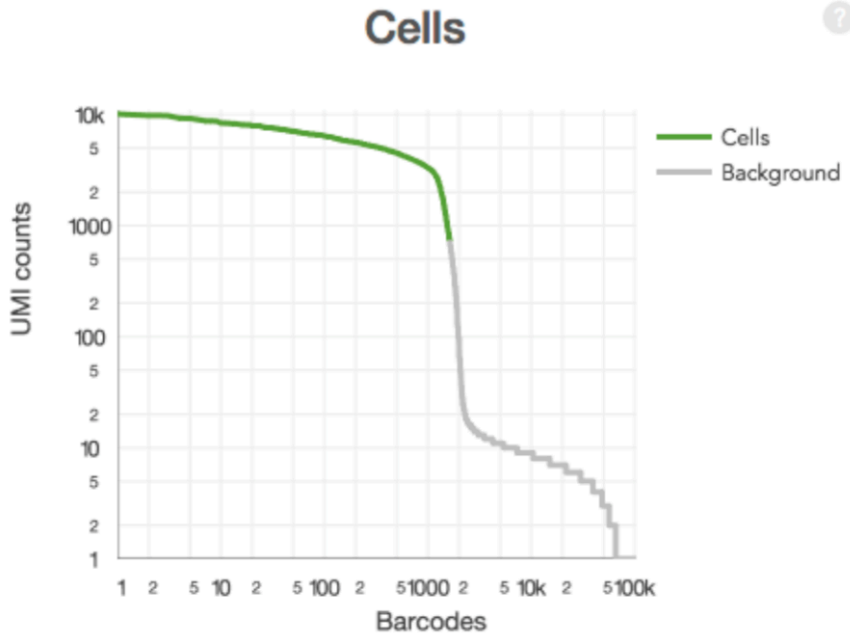
- UMIs (Unique Molecular Index)
  - Must not be a homopolymer, e.g. AAAAAAAAAA
  - Must not contain N
  - Must not contain bases with base quality < 10
  - UMIs that are 1 mismatch away from a higher-count UMI are corrected to that UMI if they share a cell barcode and gene.

Marking Duplicates

- Using only the confidently mapped reads with valid barcodes and UMIs,
  - Correct the UMIs
    - UMIs are corrected to more abundant UMIs that are one mismatch away in sequence.
  - Record which reads are duplicates of the same RNA molecule
  - Count only the unique UMIs as unique RNA molecules
  - These UMI counts form an **unfiltered gene-barcode matrix.**

Filtering Cells (the 10x way)

- Select barcodes that likely contain cells
  - Sum UMI counts for each barcode
  - Select barcodes with total UMI count >10% of the 99th percentile of the expected recovered cells.
- Produces a **filtered gene-barcode matrix**.



Downstream Analysis – offered by 10x

- Clustering analysis results
  - For each ‘K’ (number of clusters desired): – Which cells go into which clusters – Differentially expressed genes across cluster
- Principle Component Analysis (PCA) results
  - How much each gene contributes to the lower-dimensional space
  - PCA projection coordinates of cells
- t-SNE analysis results
  - The coordinates of each cell in 2-d space
- R package to visualize
  - 10x genomics

R and Seurat for analysis

- Seurat - Most Complete R package for scRNAseq analysis
  - Horrible in terms of design

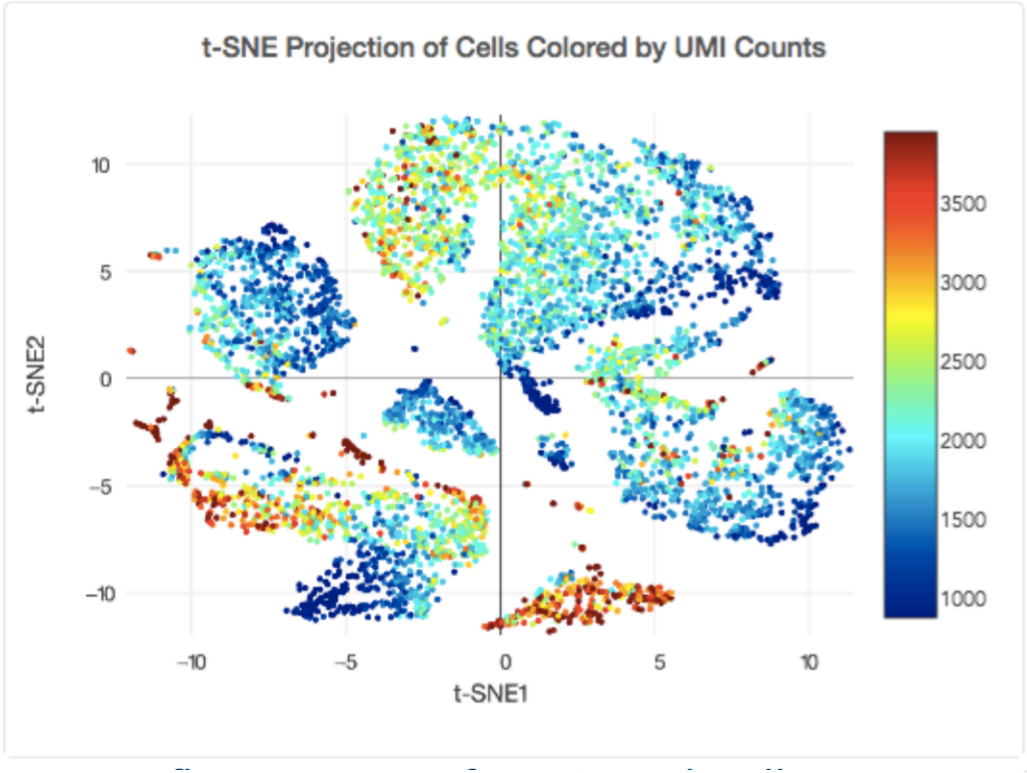
1. Unsupervised clustering and discovery of cell types and states
2. Spatial reconstruction of single cell data
3. Integrated analysis of single cell RNA-seq across conditions, technologies, and species

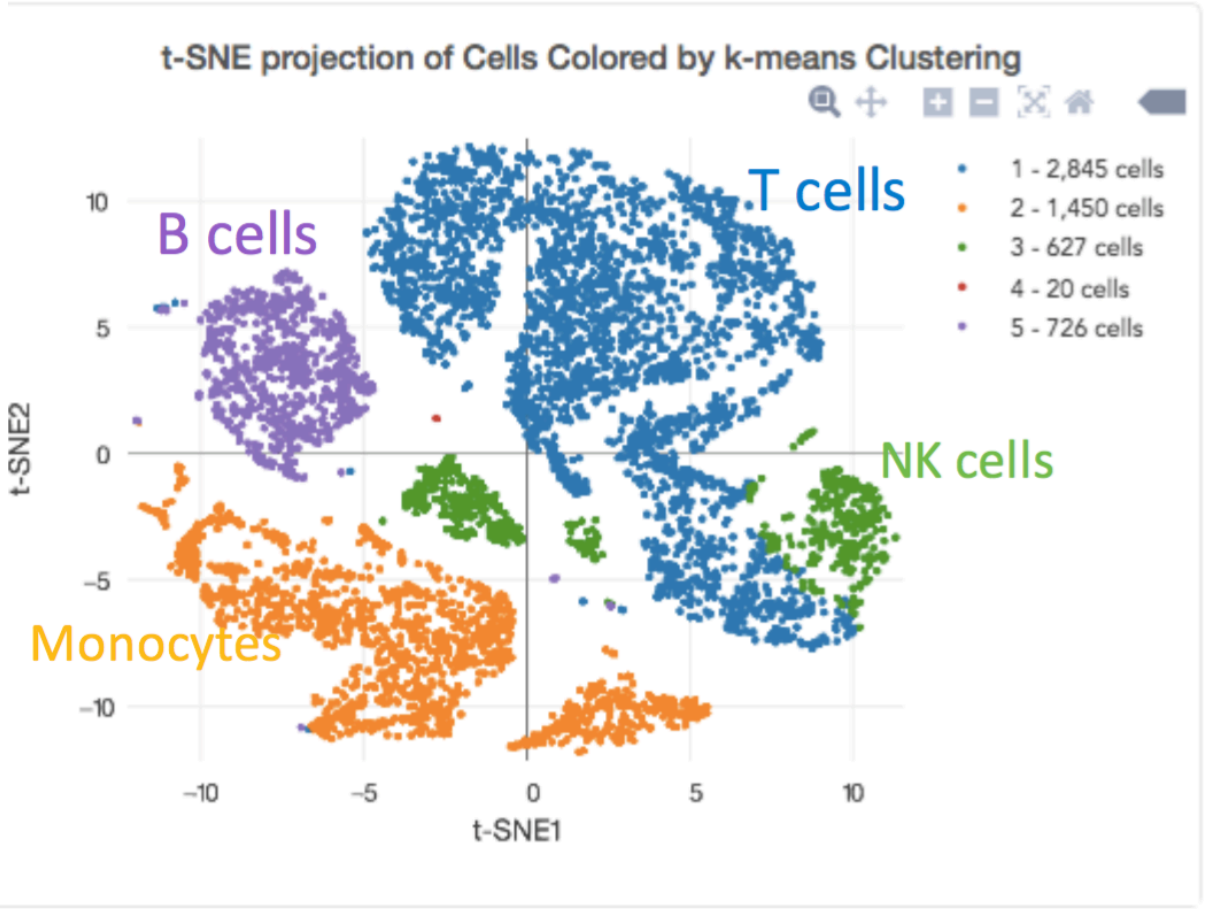
- “Easy to Use by both dry-lab and wet-lab researchers” - NOT

Seurat https://github.com/satijalab/seurat

UCDAVIS Bioinformatics Core

Clustering Analysis (TSNE)

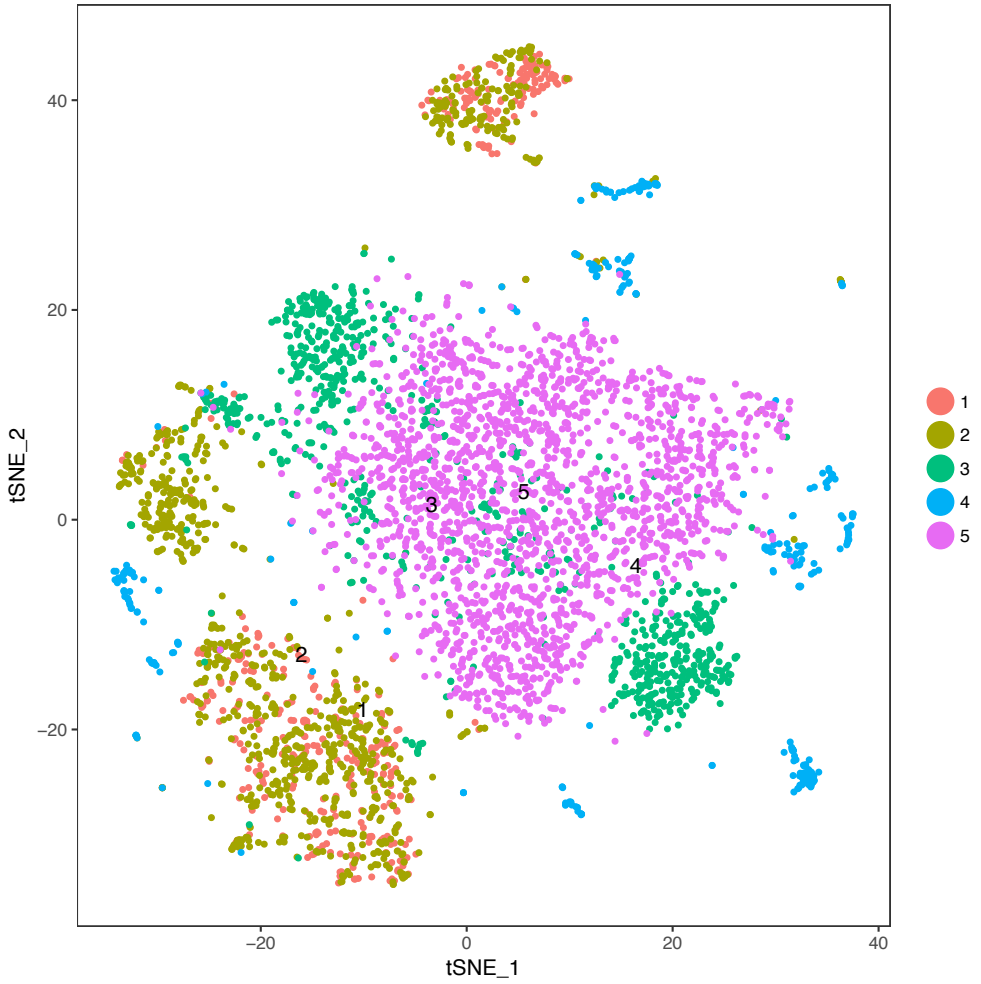


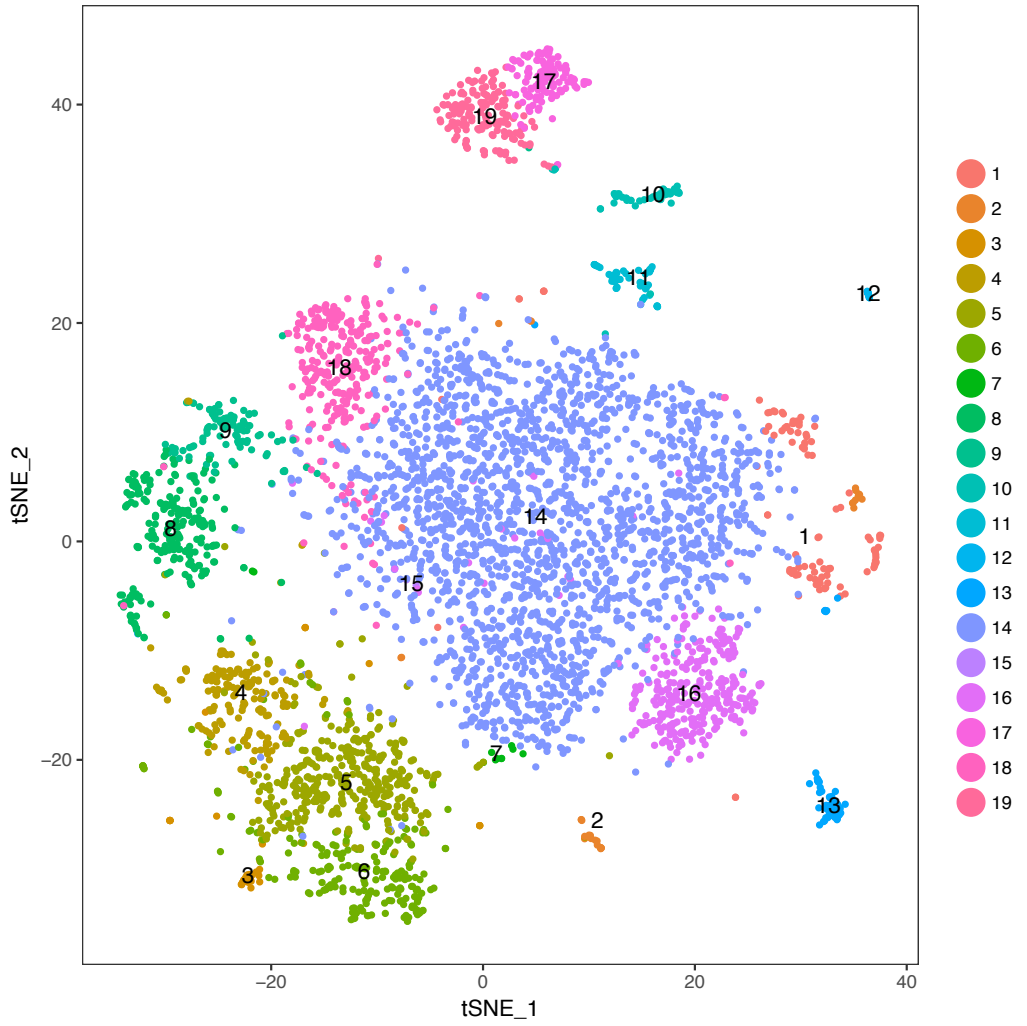


Mix of guided and unguided techniques

UCDAVIS Bioinformatics Core

Sample to Sample variability warning

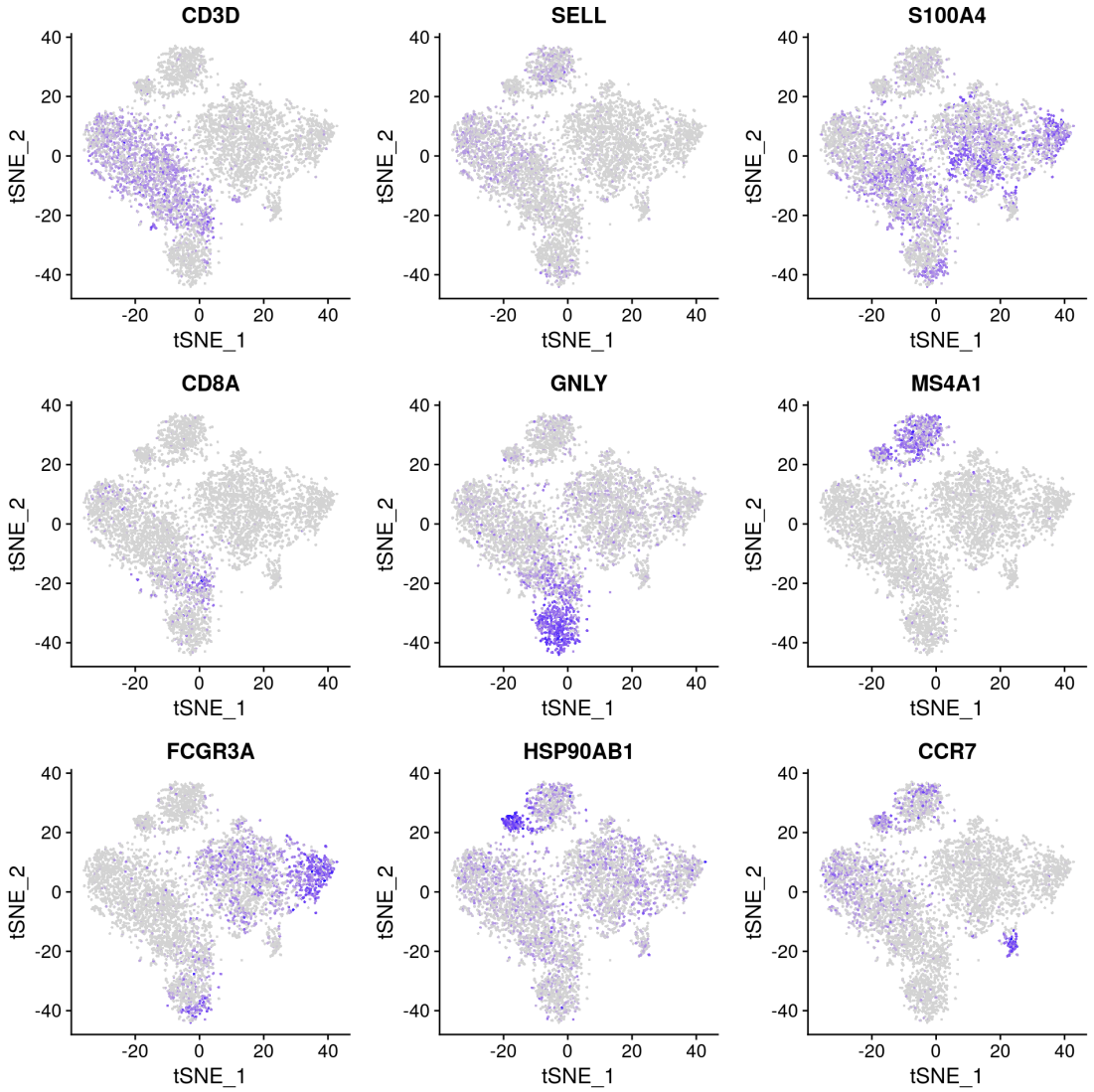




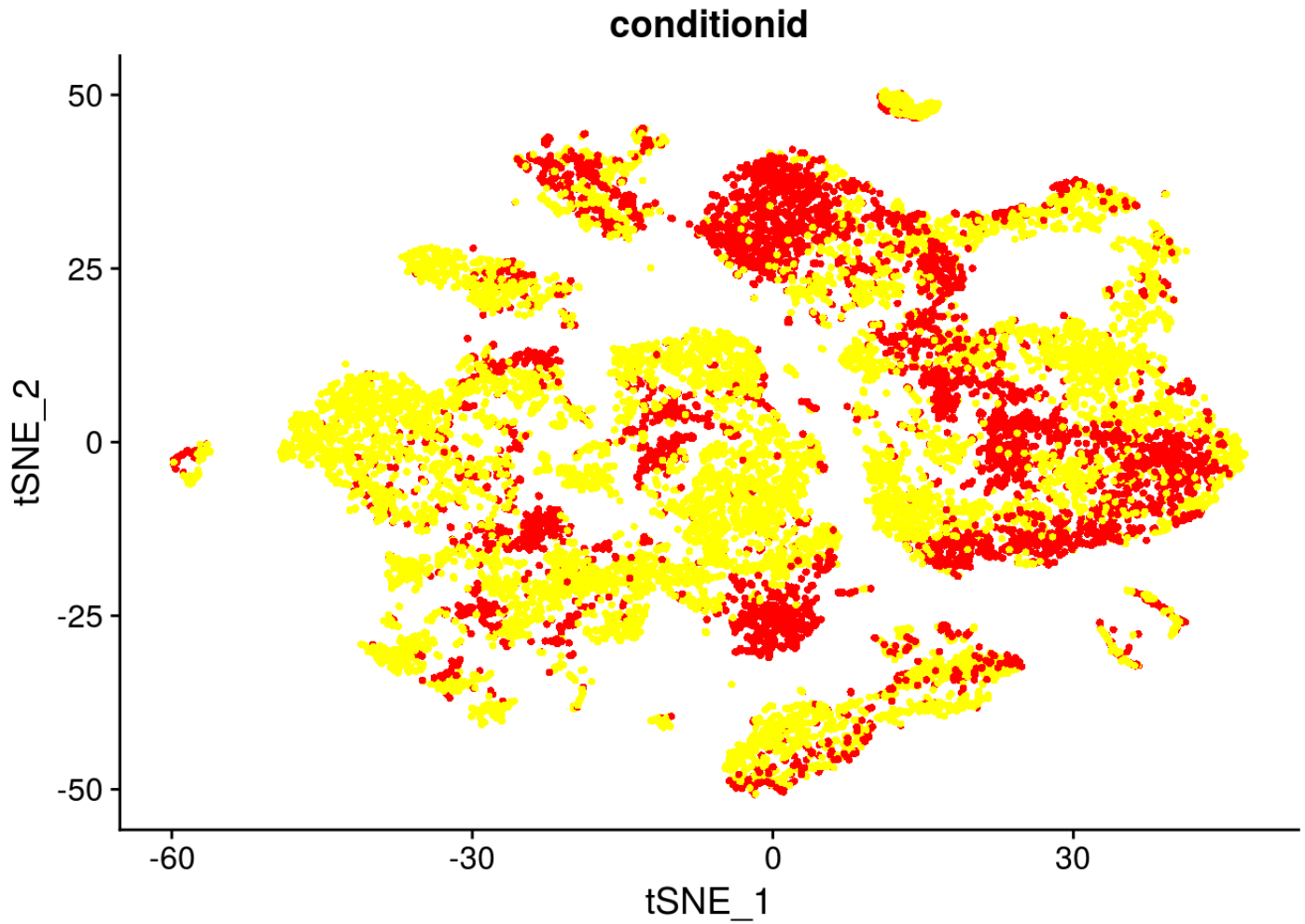
Marker Identification

Cluster markers can be identified as:

"genes differentially expressed in a cluster relative to all other clusters"



Comparison across experimental conditions



UCDAVIS Bioinformatics Core

3-day Single Cell RNA Seq Workshops

- Dec 18 - Dec 20, 2017

  Bioinformatics: Single Cell RNA-Seq Workshop @ UC Davis
- Jan 10 - Jan 12, 2018

  Bioinformatics: Single Cell RNA-Seq Workshop @ UC Berkeley
- Mar 14 - Mar 16, 2018

  Bioinformatics: Single Cell RNA-Seq Workshop @ UCSF

https://registration.genomecenter.ucdavis.edu/