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

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 reason to conduct single cell analysis

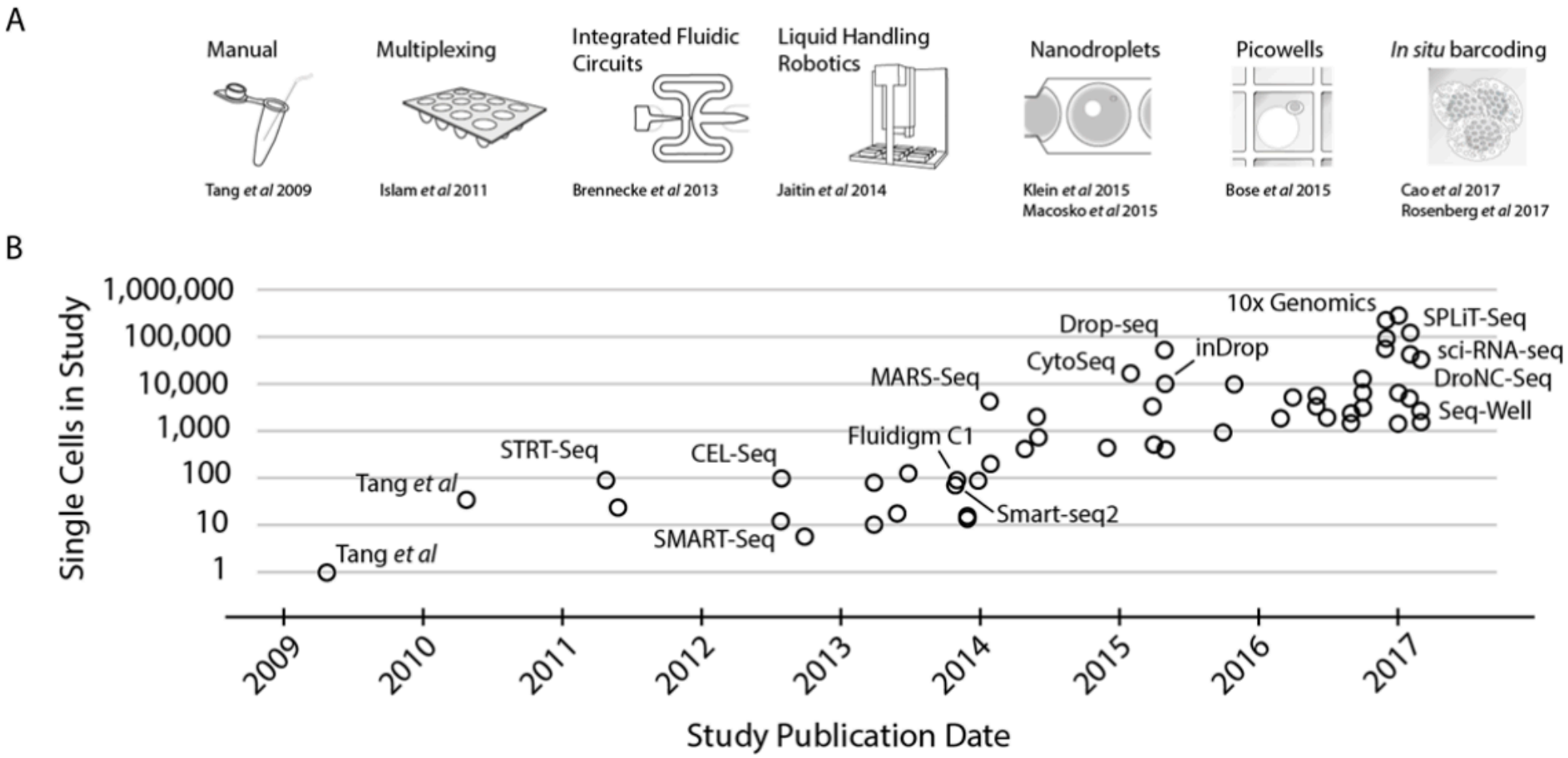
Bulk RNAseq, where you measure the 'average' expression of all constituent cells, is sometimes insufficient for some experimental questions.

- 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.
- Spacial Transcriptomics – isolation of cells with known spacial location.

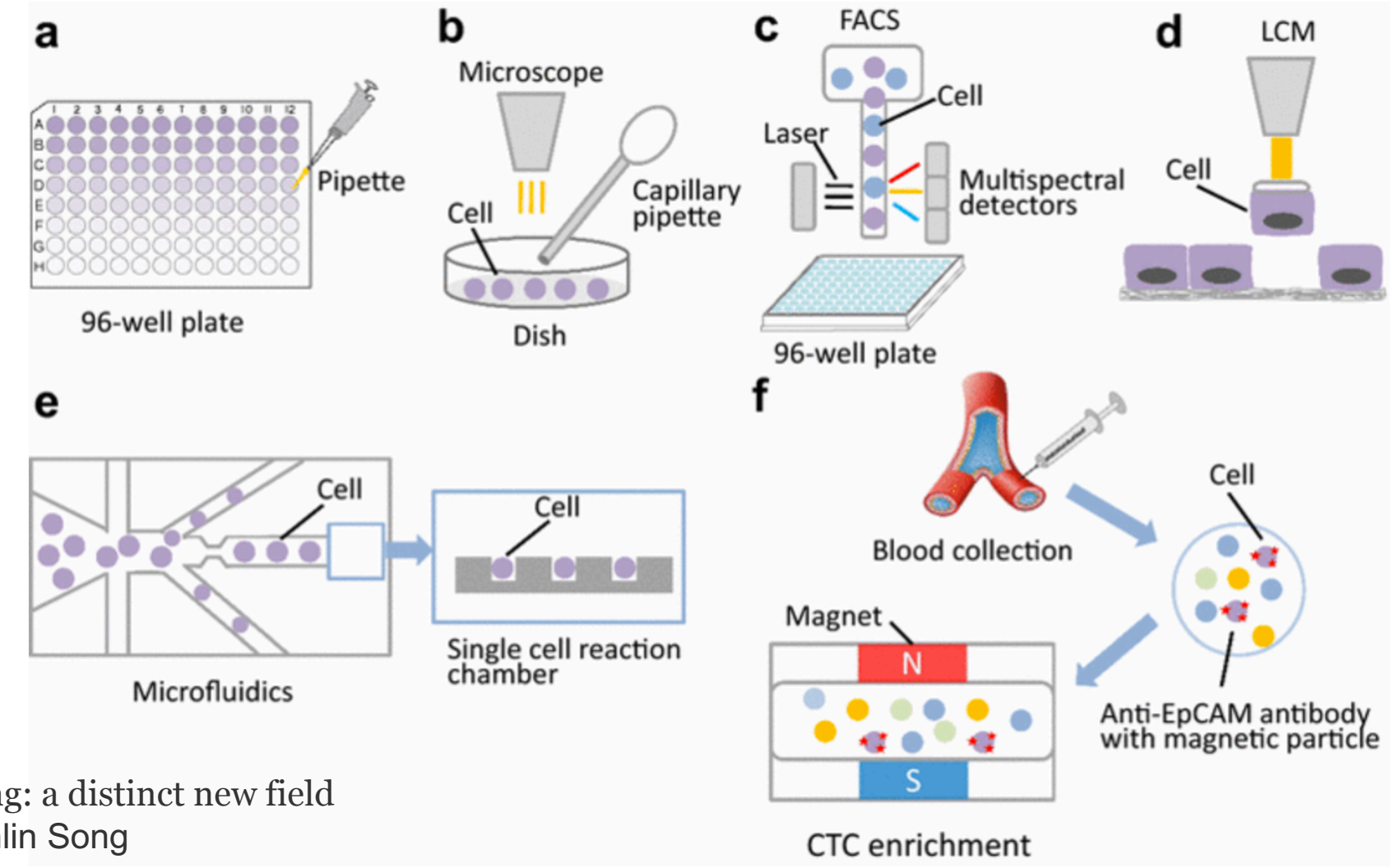
UCDAVIS Bioinformatics Core

Exponential scaling of single-cell RNAseq in the last decade

https://arxiv.org/abs/1704.01379



Single-cell isolation methods



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

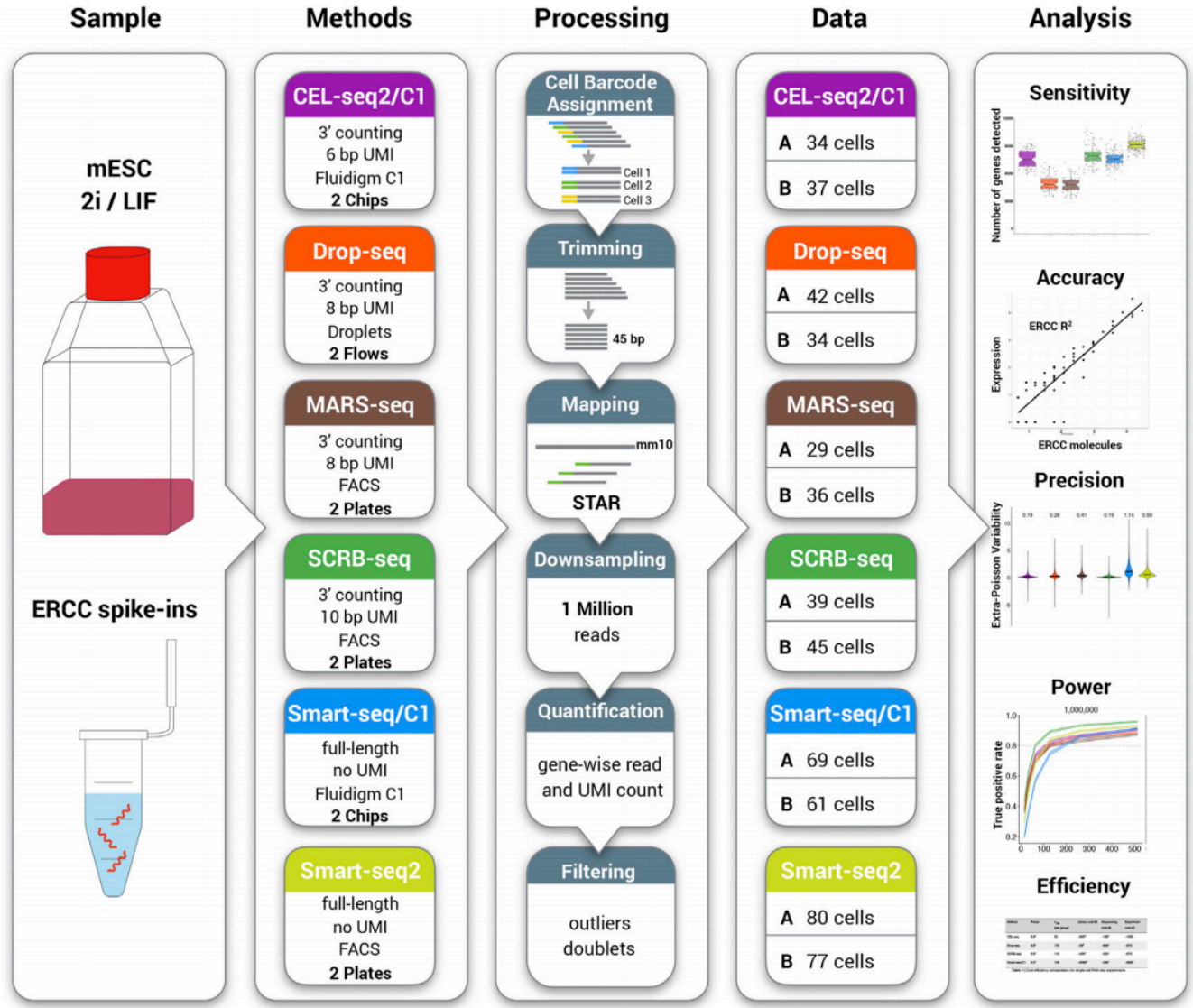
UCDAVIS Bioinformatics Core

Molecular Cell

Comparative Analysis of Single-Cell RNA Sequencing Methods

Authors

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



3’ counting vs whole transcript

- In 3’ counting techniques – 1 reads per transcript
  - Based on polyA
  - Expression analysis only
  - Fewer reads per cell needed (~60K reads/cell)
  - Less noise in expression patterns
- Whole transcript
  - Based on polyA
  - Expression analysis
  - Splicing information
  - More information need beyond expression, the higher the reads needed per cell (~60K reads/cell to 10M reads/cell)

UCDAVIS Bioinformatics Core

Single-Cell with 10x genomics

Gene expression profiling at scale with single cell resolution

10x Chromium Box

Basic Stats

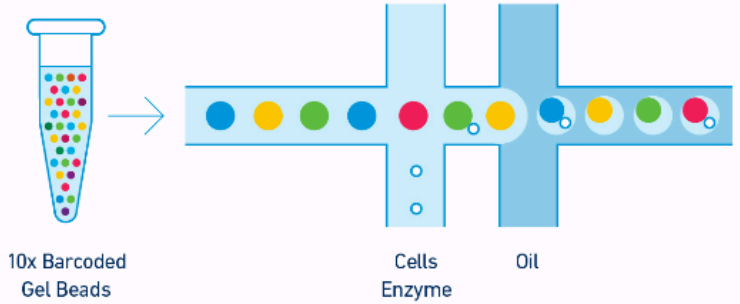
- Up to 8 channels processed in parallel
- 500 to 6,000 (V1) 10,000 (V2) cells per channel
- 10 minute run time per chip
- Up to 30 um cell diameter tested
- ~50 % cell processing efficiency

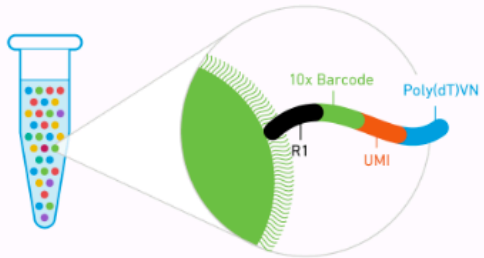
Number of cells	Expected Doublet Rate (%)
1,200	~1.2
3,000	~2.9
6,000	~5.7

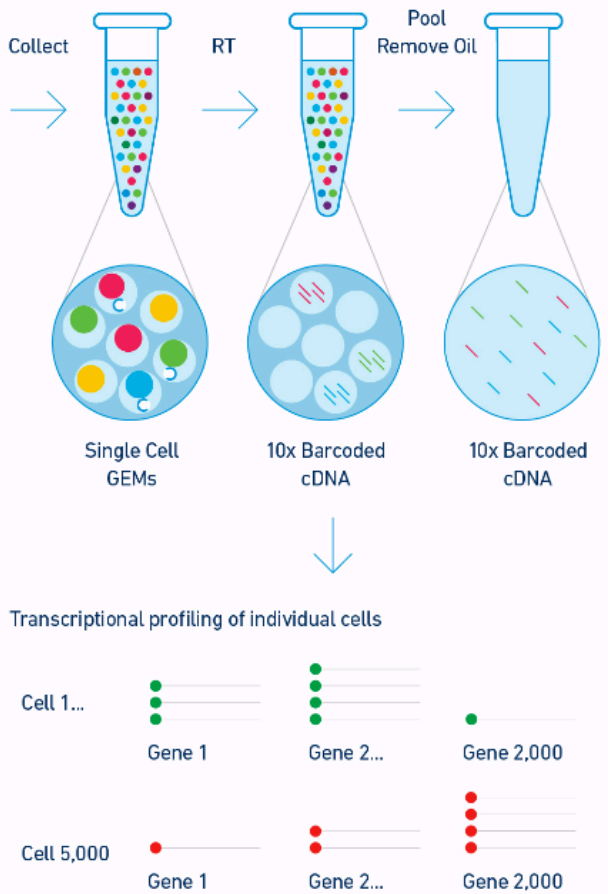
Number of cells	Expected Doublet Rate (%)
500	~0.4
1,000	~0.8
3,000	~2.3
5,000	~3.9
10,000	~7.6

User controlled trade off between cell numbers and doublet rate

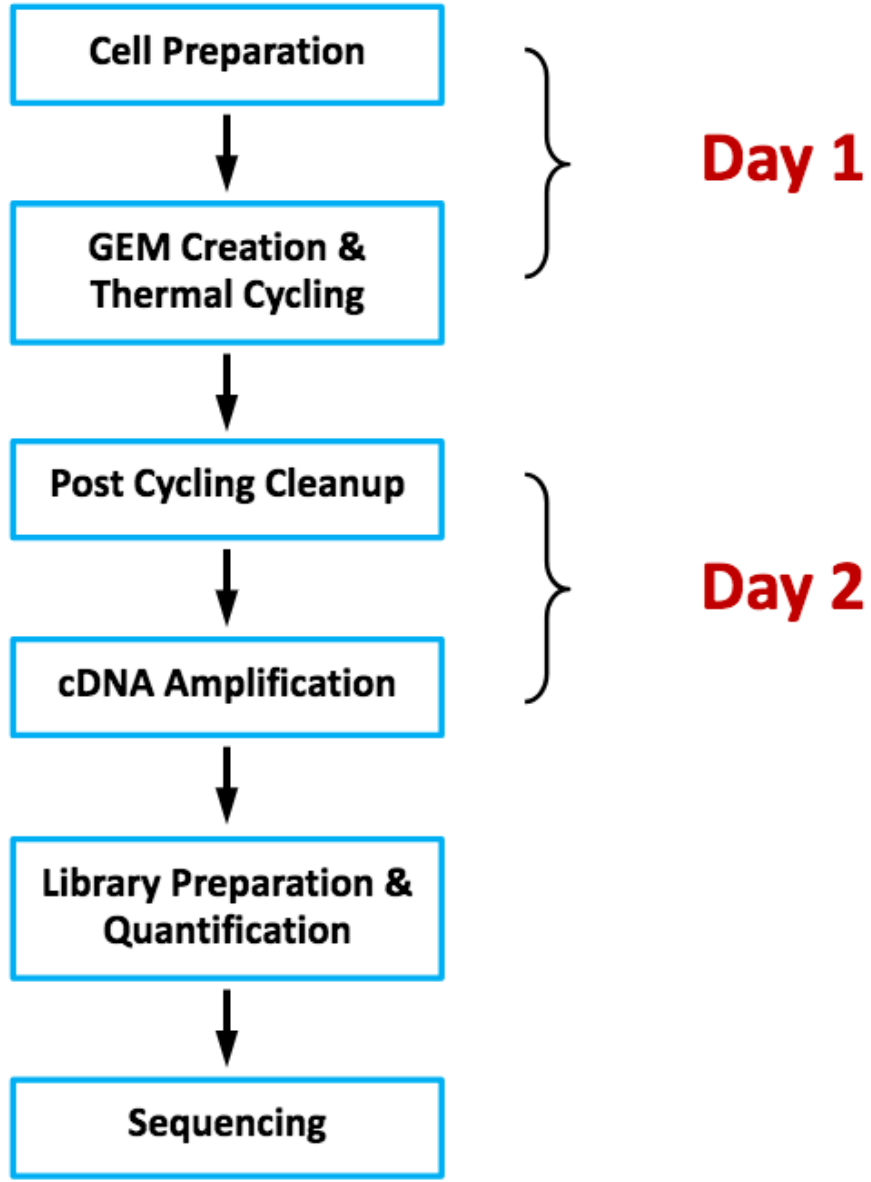
UCDAVIS Bioinformatics Core





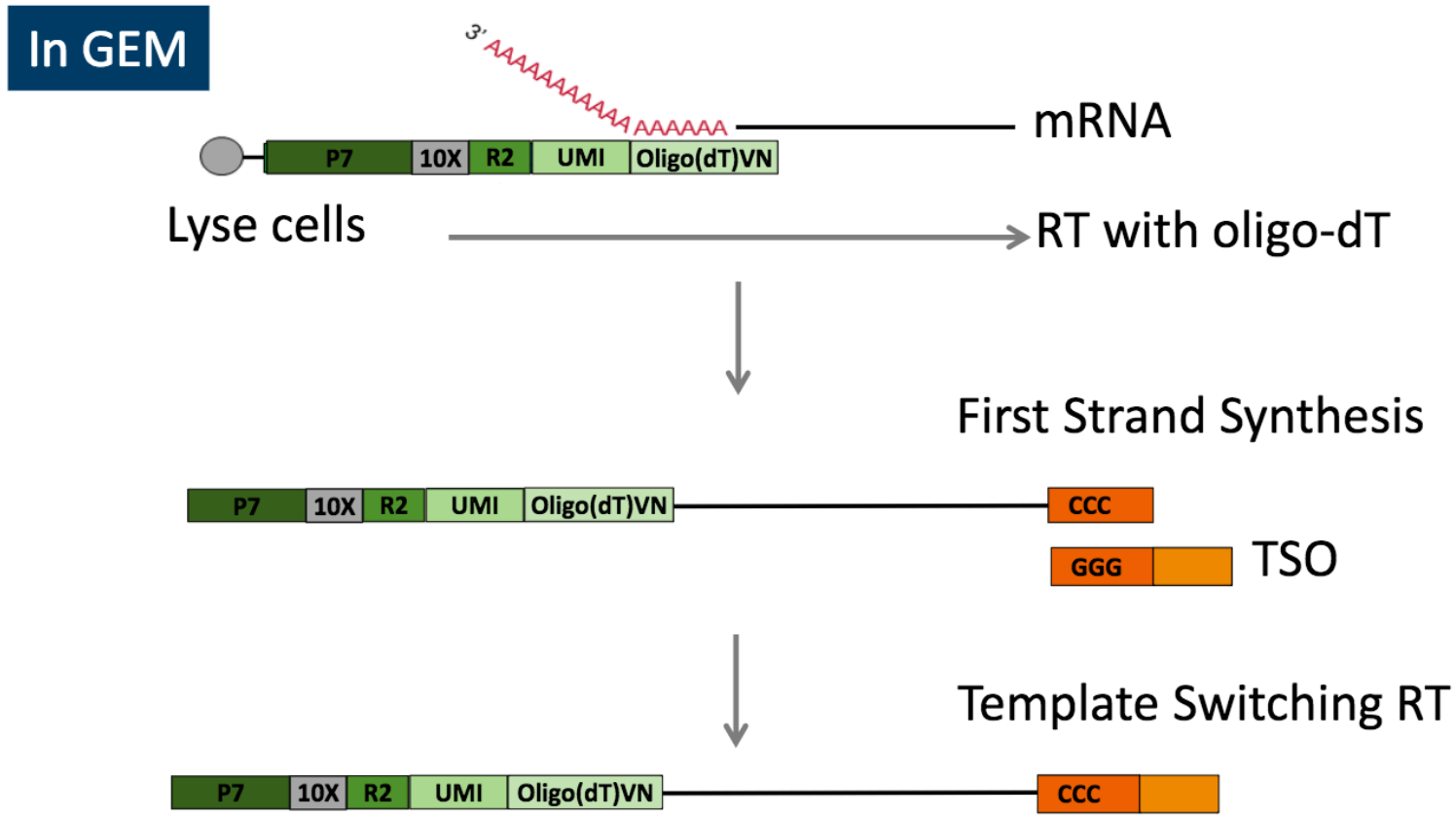


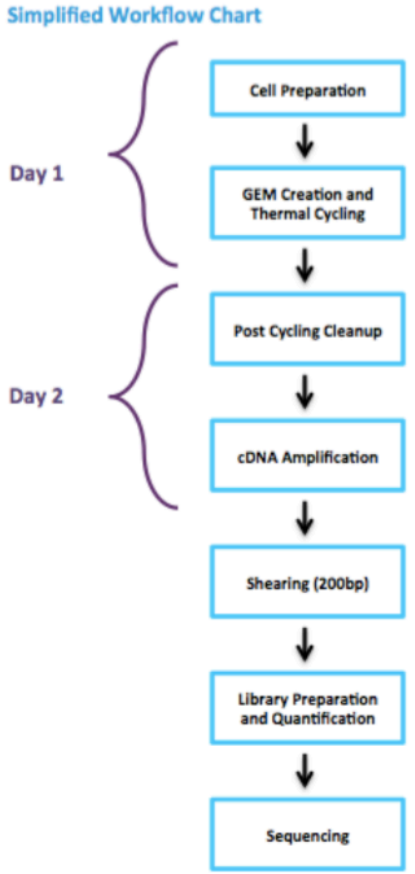
- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell



UCDAVIS Bioinformatics Core

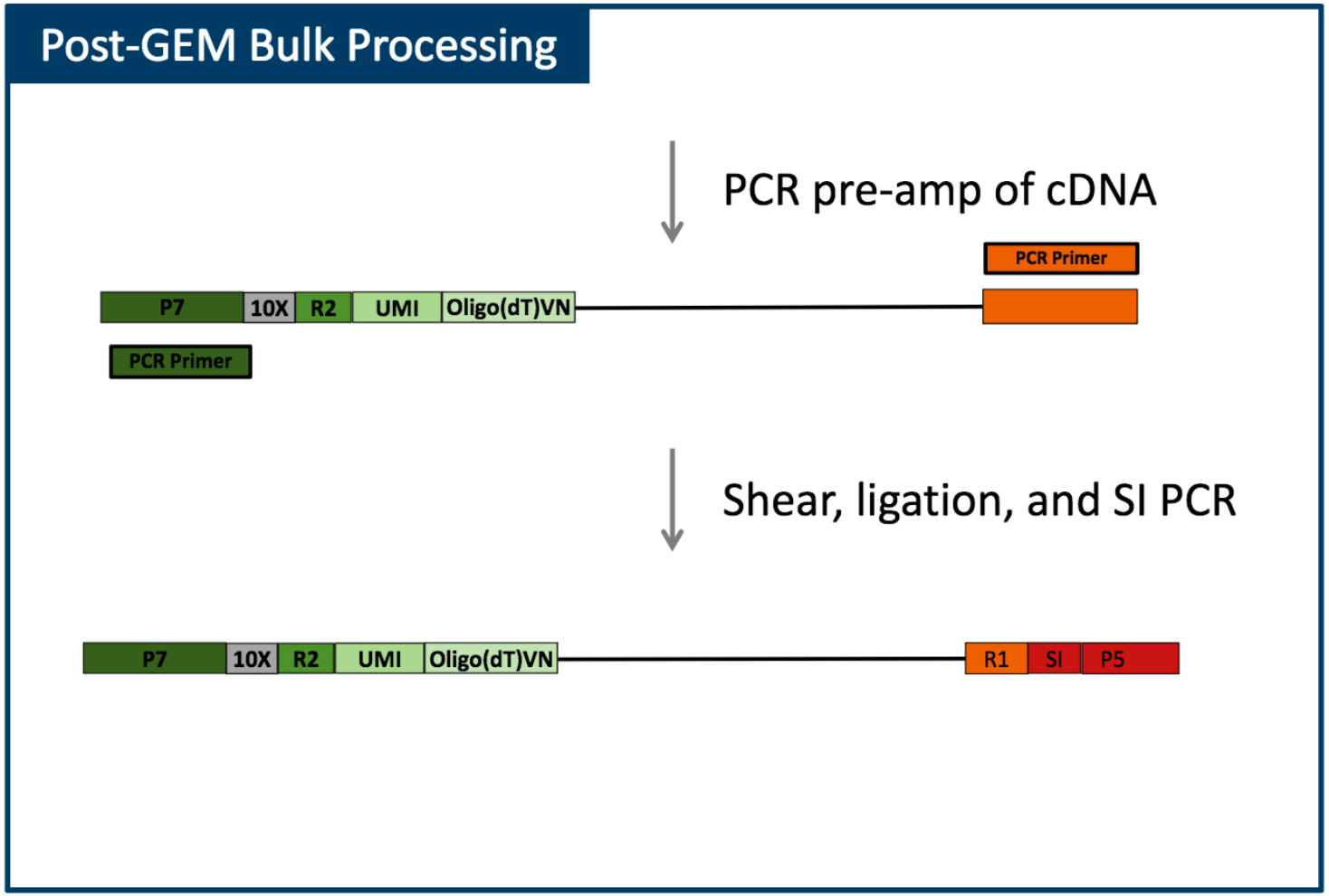
Basically a TAGseq protocol per cell 3’ expression

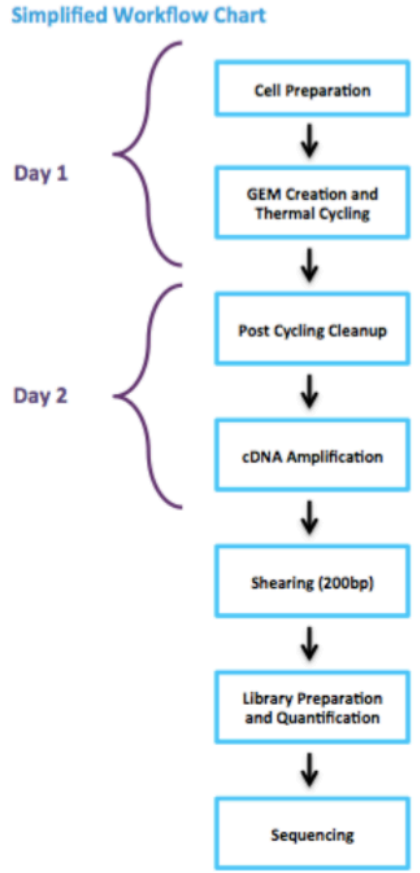




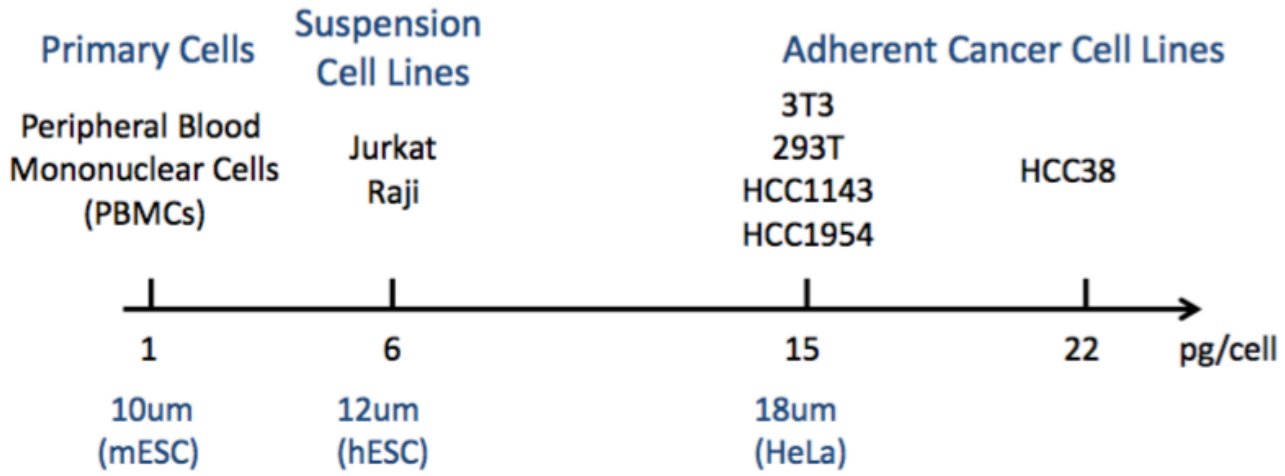
UCDAVIS Bioinformatics Core

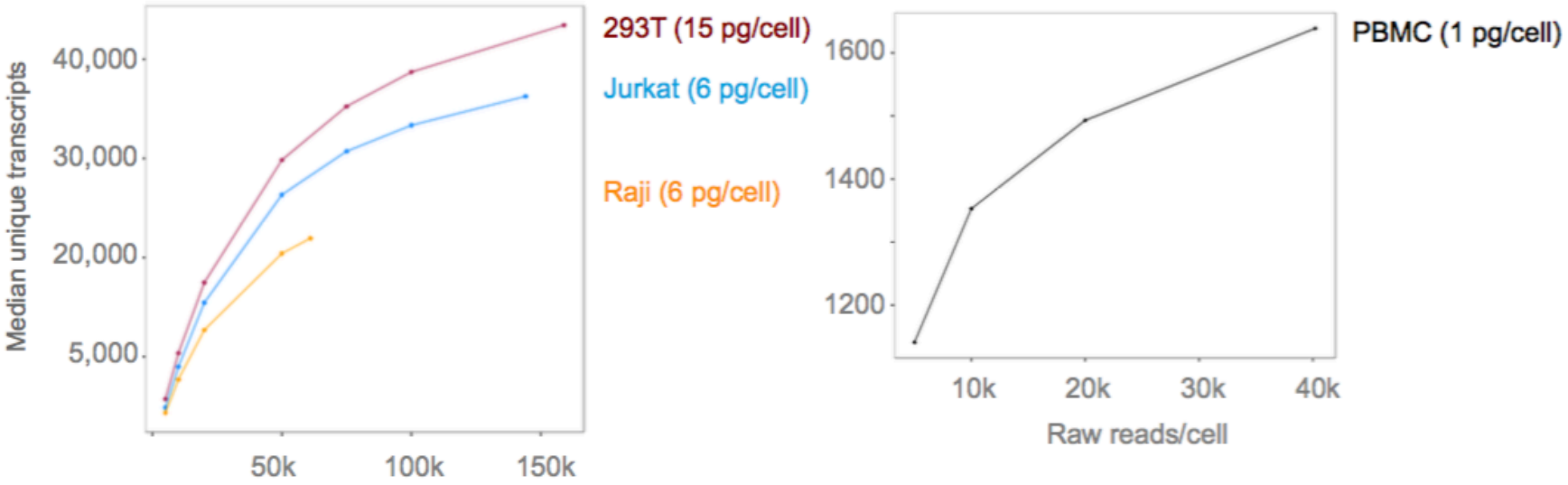
Post-GEM Bulk Processing





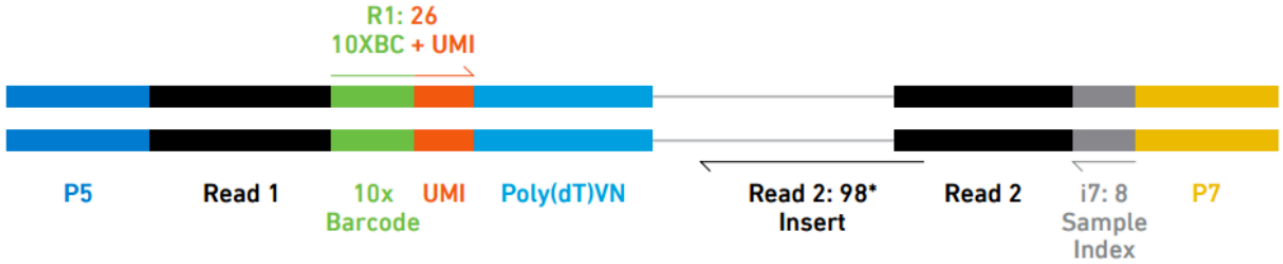
Cells of differing sizes and complexity





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

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

Custom sequencing run, with 3 reads, V2 kits

Sequence Read	Recommended Length	Read Description
Read 1	100bp	10 barcode and UMI
I7 Index	8bp	Sample Index Read
Read2	100bp	Transcript Tag

@ full capacity 10,000 cells per sample and 60K reads per cell = 500M reads or ~1.25 lane/sample

10x Software – System requirements

Local

- Run on single, standalone Linux system
- CentOS/RedHat 5.2+ or Ubuntu 8.04+
- 8+ cores, 64GB RAM

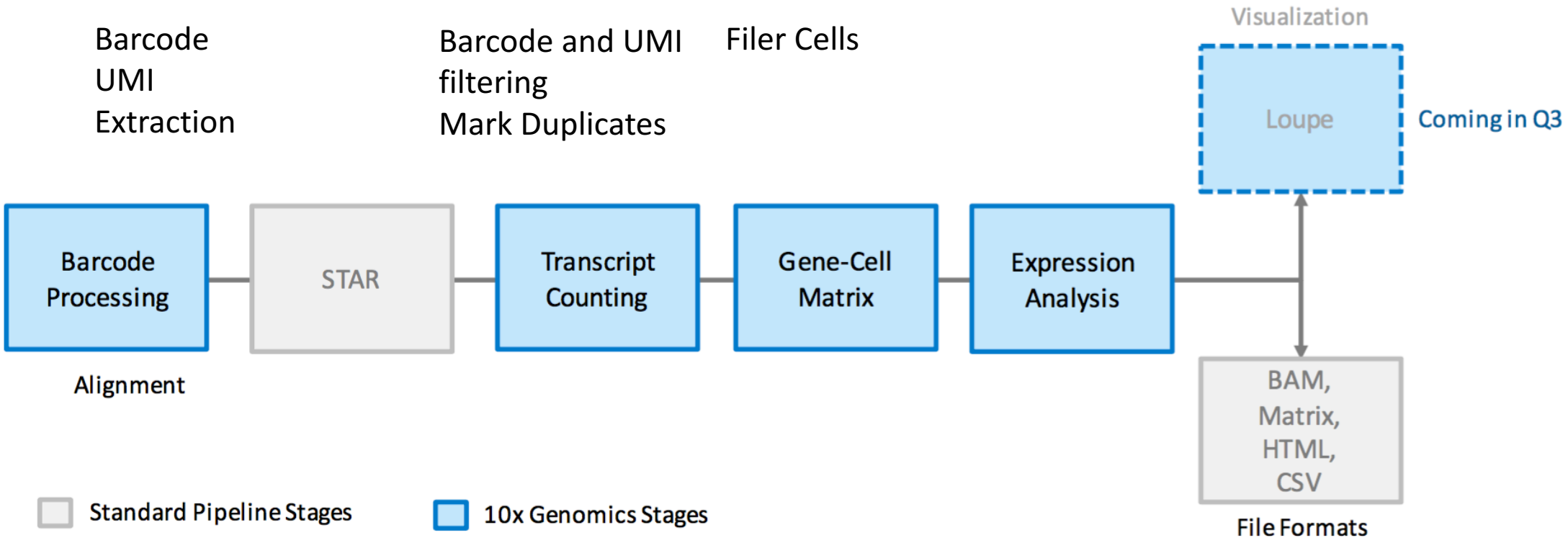
Cluster

- Run on SGE and LSF
- Each node must have 8+ cores and 8GB+ RAM/core
- Shared filesystem between nodes (e.g. NFS)

50 core-hours per 100M reads, 5000 cells, 40k reads/cell: 95 core-hours

UCDAVIS Bioinformatics Core

Analysis Workflow



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.**

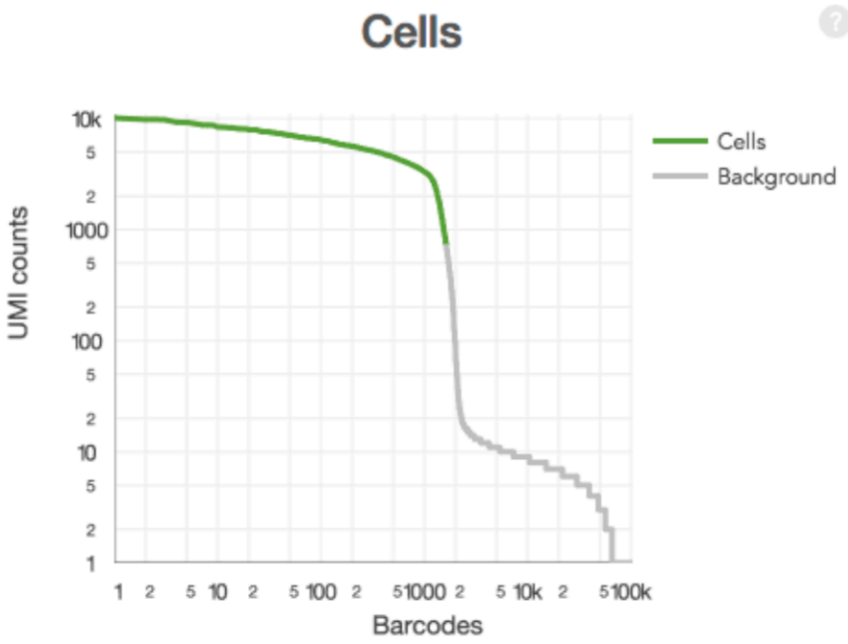
Filtering Cells

- Select GEMs 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

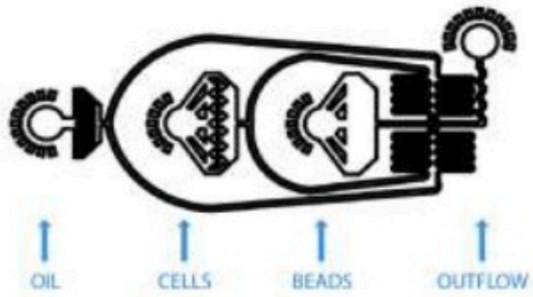
- Clustering analysis results
  - For each ‘K’ (number of clusters desired): – Which cells go into which clusters – Differentially expressed genes across cluster
- Principle Component Analysis (PCA) results
  - How much each gene contributes to the lower-dimensional space
  - PCA projection coordinates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

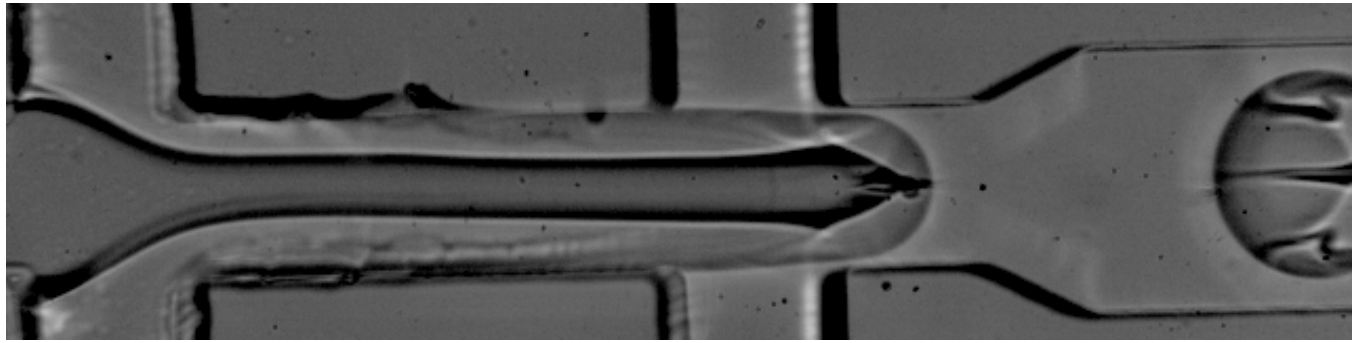
UCDAVIS Bioinformatics Core

Single-Cell with Drop-seq

Gene expression profiling at scale with single cell resolution

UCDAVIS Bioinformatics Core





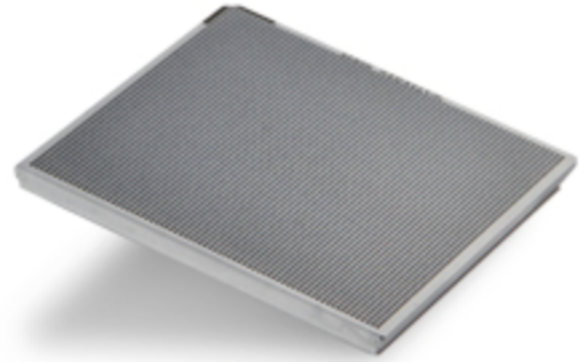
UCDAVIS Bioinformatics Core

Single-Cell with Wafergen

Gene expression profiling at scale with single cell resolution

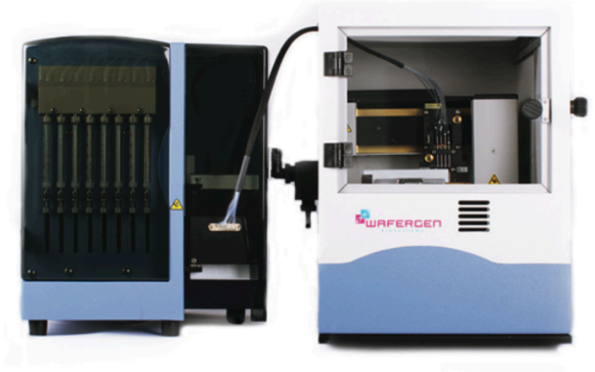
Single-cell sequencing technologies

WAFERGEN BIOSYSTEMS

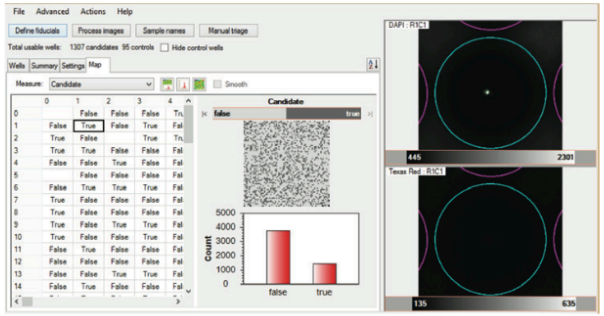


ICELL8 Chips and Reagents

Imaging Station



MultiSample NanoDispenser



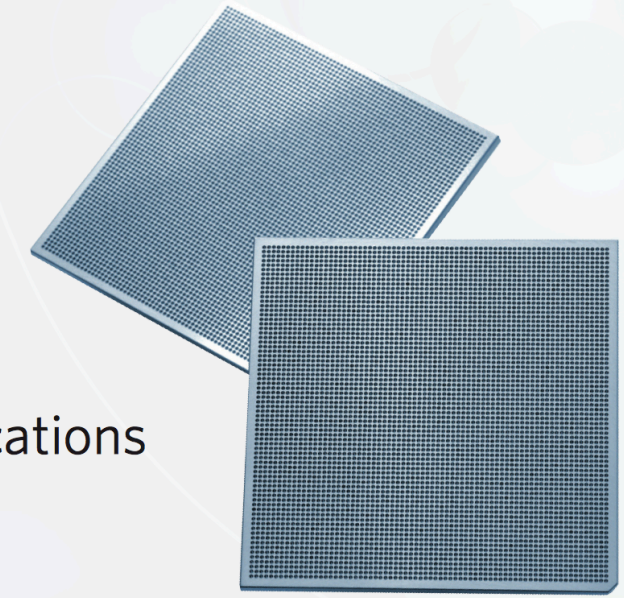
CellSelect Software

Single-cell sequencing technologies

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



R and Seurat for analysis

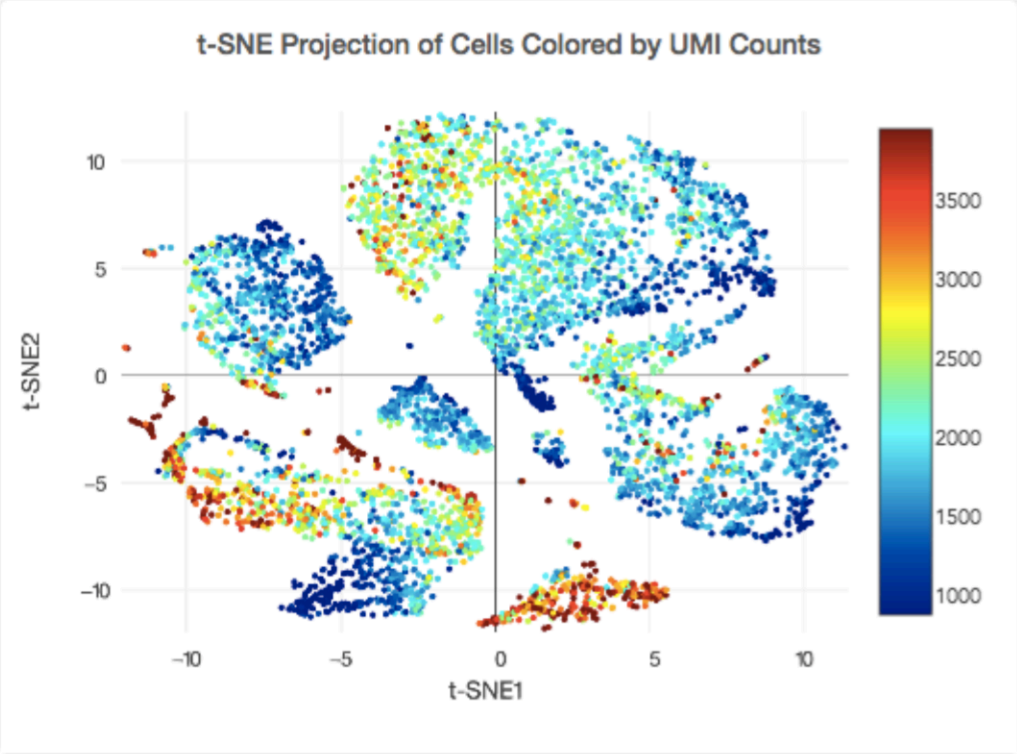
- Seurat - Most Complete R package for scRNAseq analysis
  - Horrible in terms of design

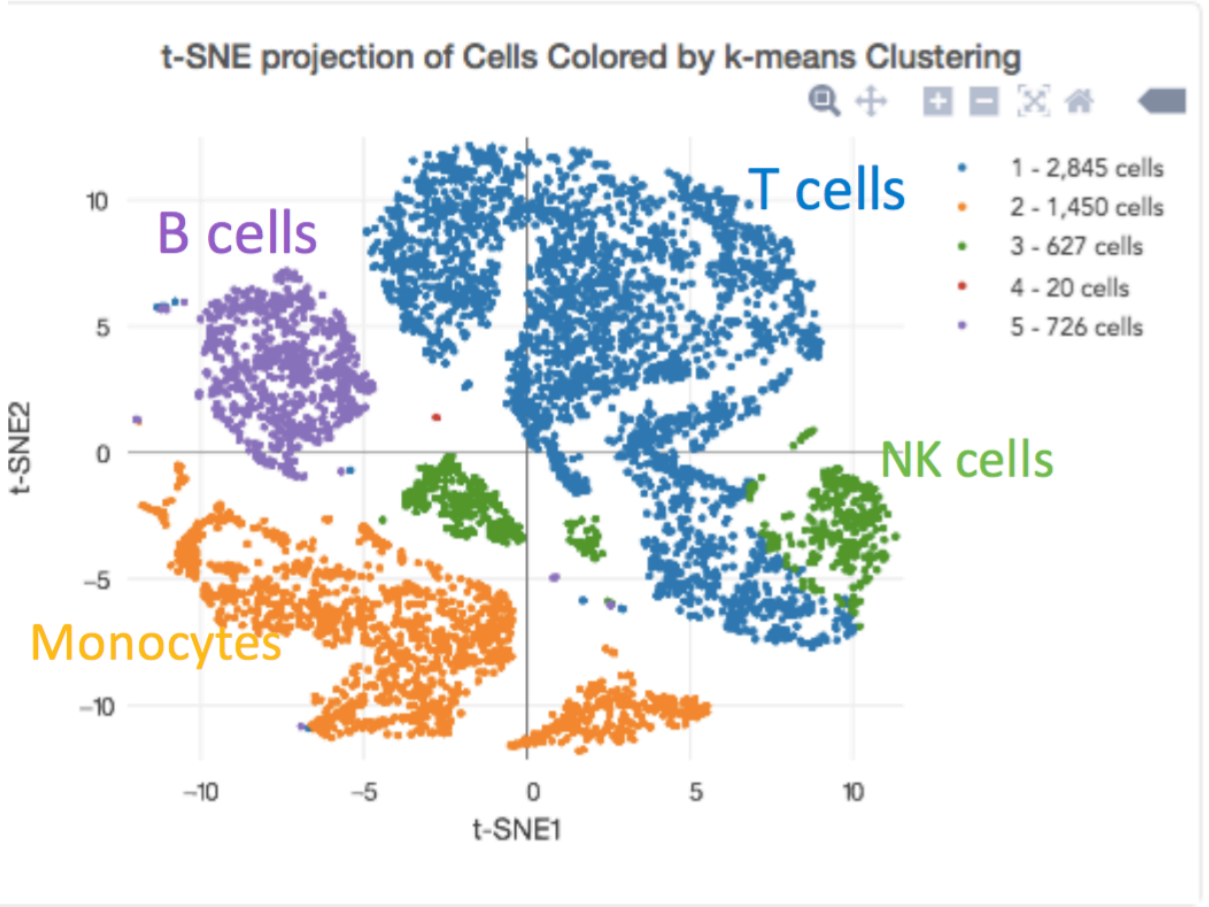
1. Unsupervised clustering and discovery of cell types and states
2. Spatial reconstruction of single cell data
3. Integrated analysis of single cell RNA-seq across conditions, technologies, and species

- “Easy to Use by both dry-lab and wet-lab researchers” - NOT
- Seurat https://github.com/satijalab/seurat

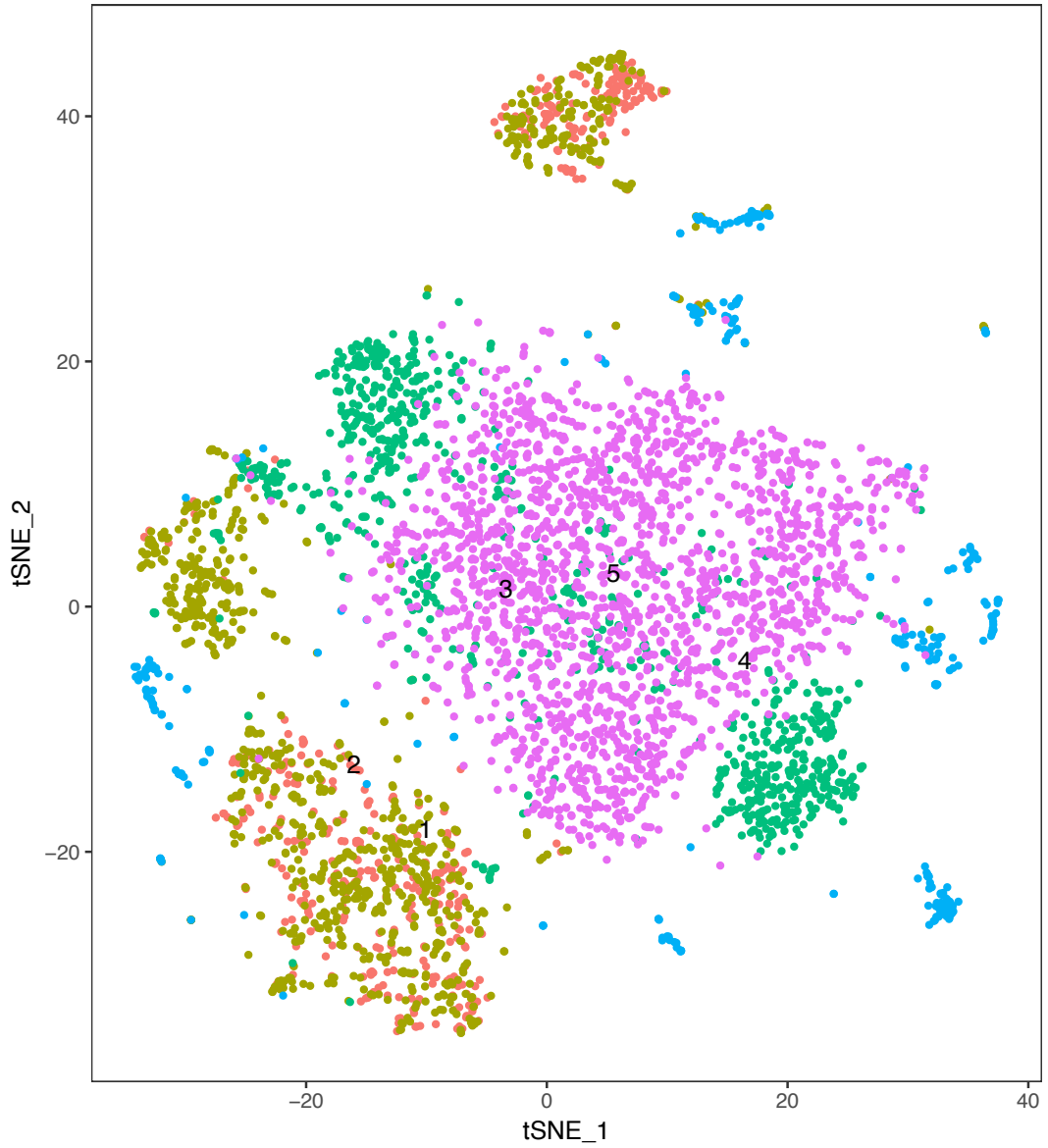
UCDAVIS Bioinformatics Core

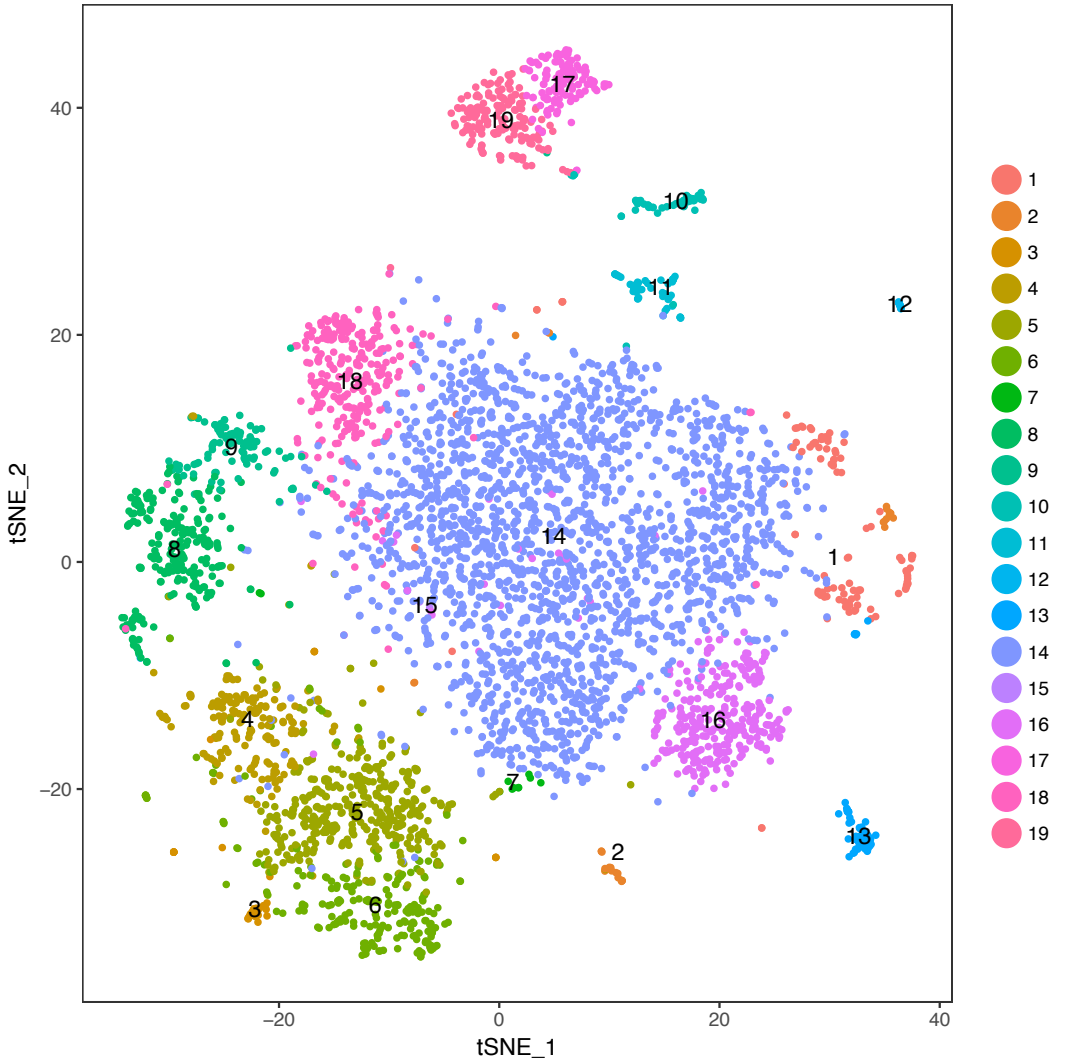
The classical clustering plot



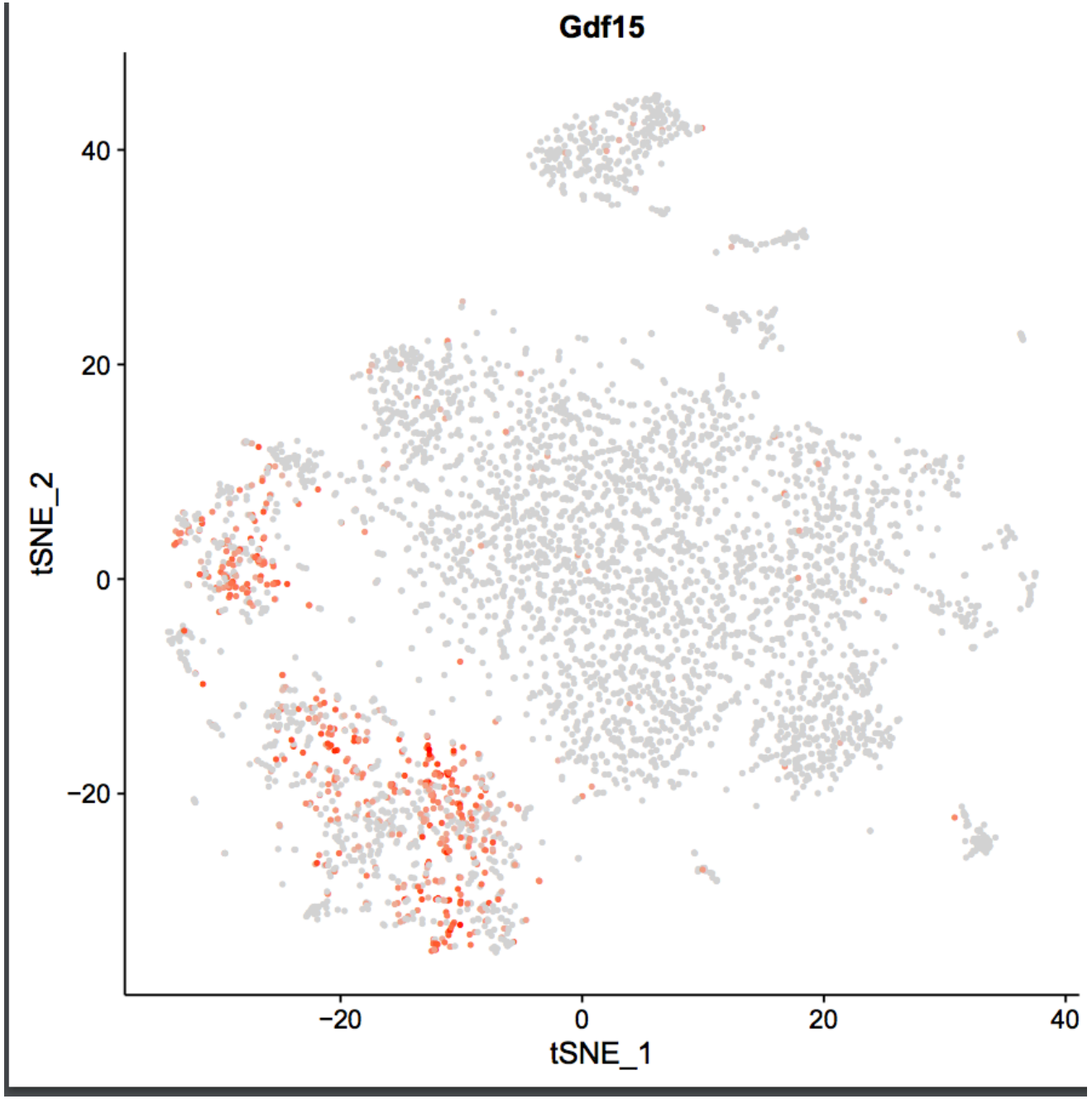


UCDAVIS Bioinformatics Core





UCDAVIS Bioinformatics Core



UCDAVIS Bioinformatics Core