Single Cell Sequencing: Powerful Applications and Practical Considerations

NIH Biowulf 20th Anniversary Seminar Series

July 24th, 2019

Michael Kelly, PhD
NCI-CCR Single Cell Analysis Facility
Cancer Research Technology Program
Frederick National Lab for Cancer Research

Overview

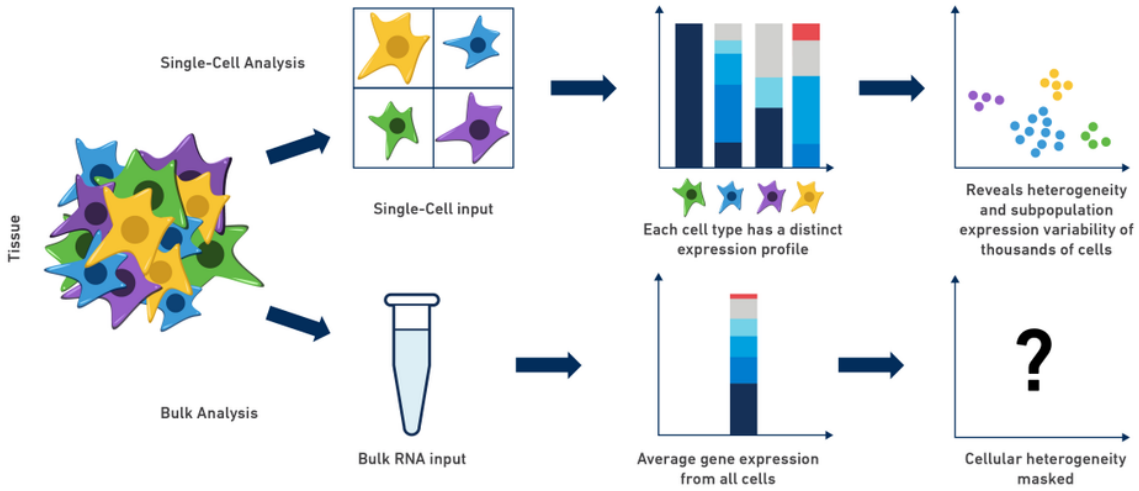
Outline:

- Brief introduction to single cell why and how
- Evolution of single cell methods and some biological answers using those methods
- Add-on modalities to scRNA-Seq (VDJ and feature barcoding)
- Sample preparation methods and considerations
- Concerns of batch effects and experimental design considerations
- Analysis tools for multiple experience levels

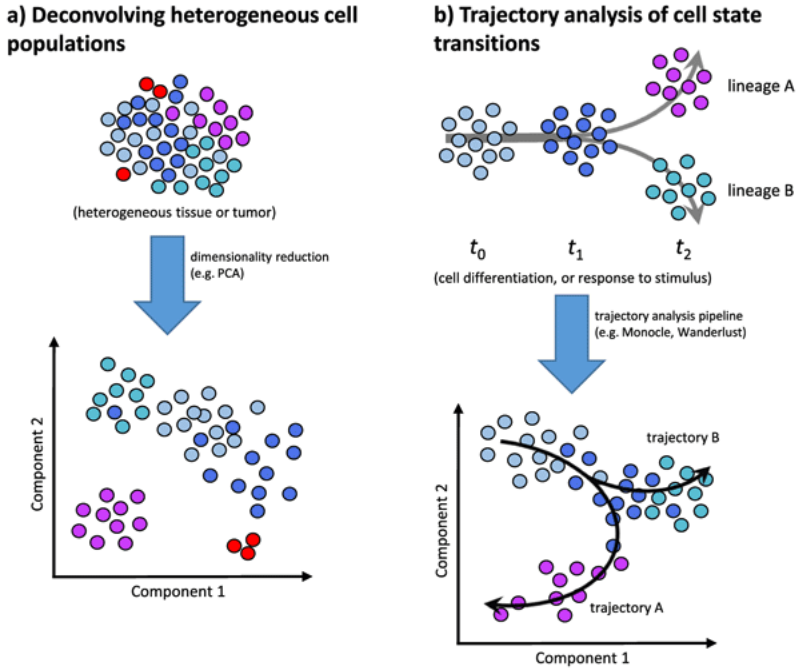
Single cell sequencing is now part of the standard repertoire of biological research techniques

Single cell sequencing avoids caveat of bulk averaging & allows inference of dynamic processes

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute



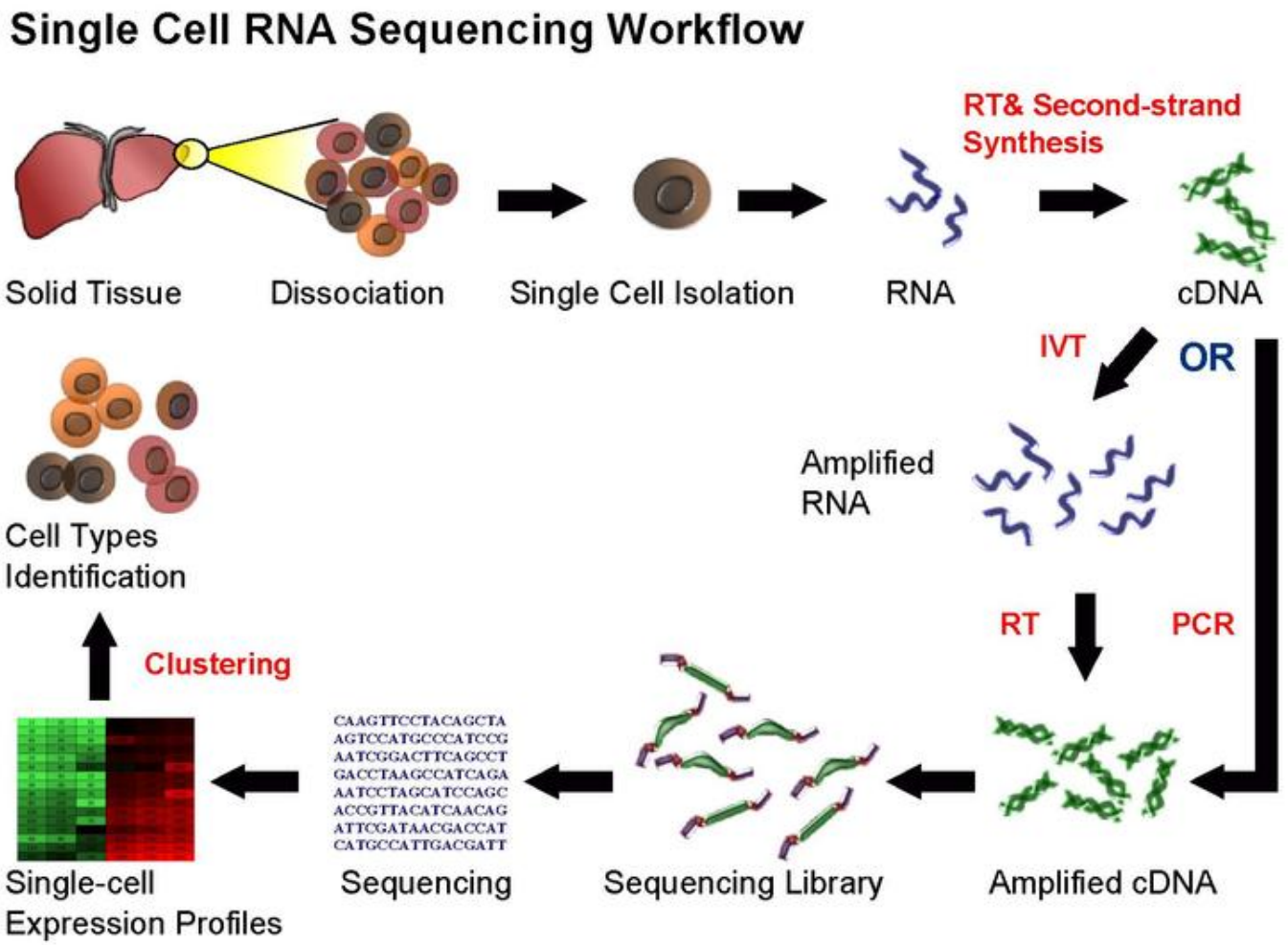
From 10x Genomics Community Website



Liu & Trapnell Review – F1000Research 2016

- Survey cells present in a biological system and identify gene signatures associated with cell types
- Compare cell number and/or phenotypic differences between conditions (i.e. healthy vs disease)
- Model dynamic changes representing biological processes
- ***Interrogate potential mechanisms at cellular resolution in health and disease***

Generalized workflow of generating single cell RNA-Seq data

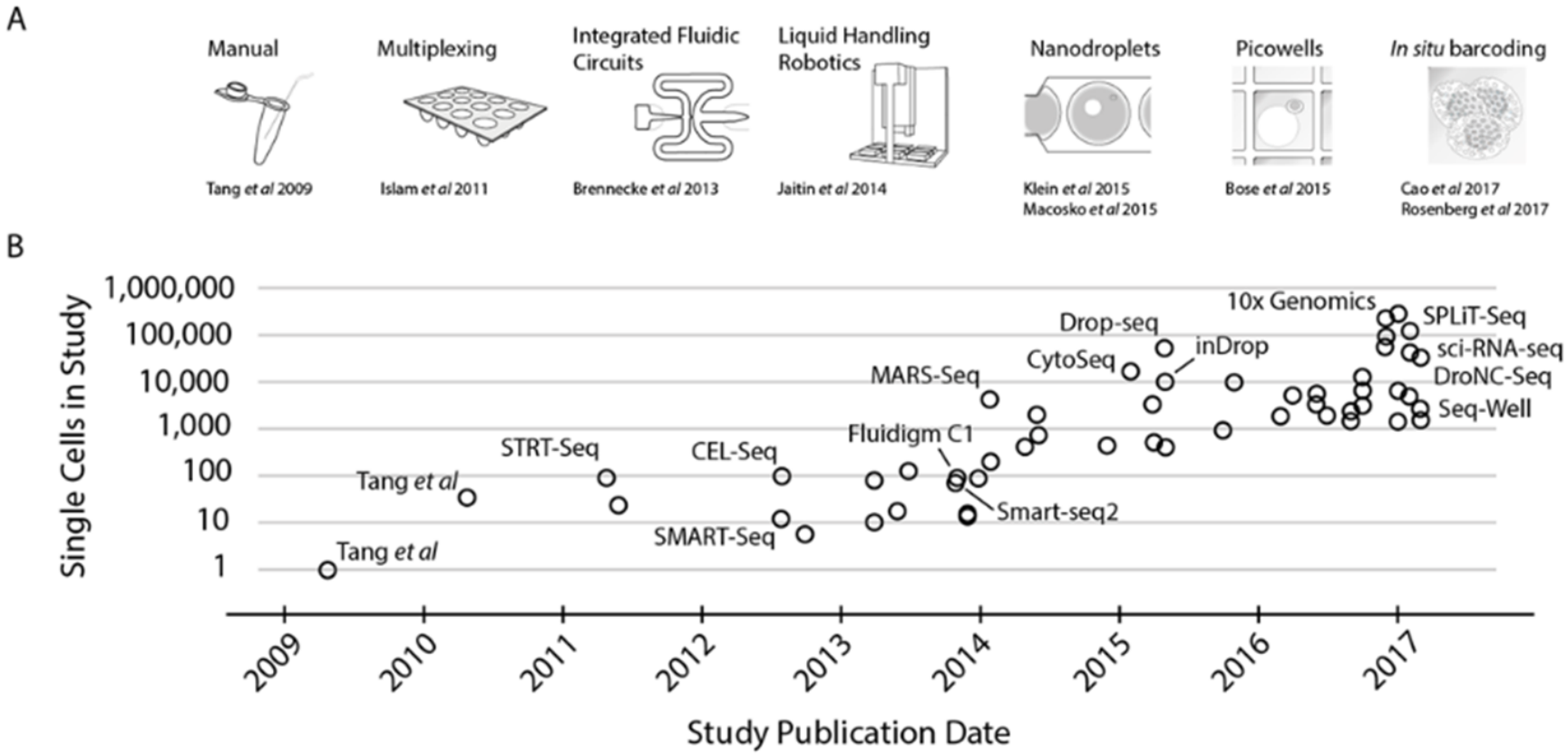


- Partition single cells
- Convert mRNA into cDNA
- Amplify cDNA
- Generate sequencing library
- Sequence
- Data analysis with identification of what transcripts are expressed by each cell profiled

https://hemberg-lab.github.io/scRNA.seq.course/

The evolution of single cell to higher throughput methods

Single cell sequencing has become easier and ability of cells per sample number has increased

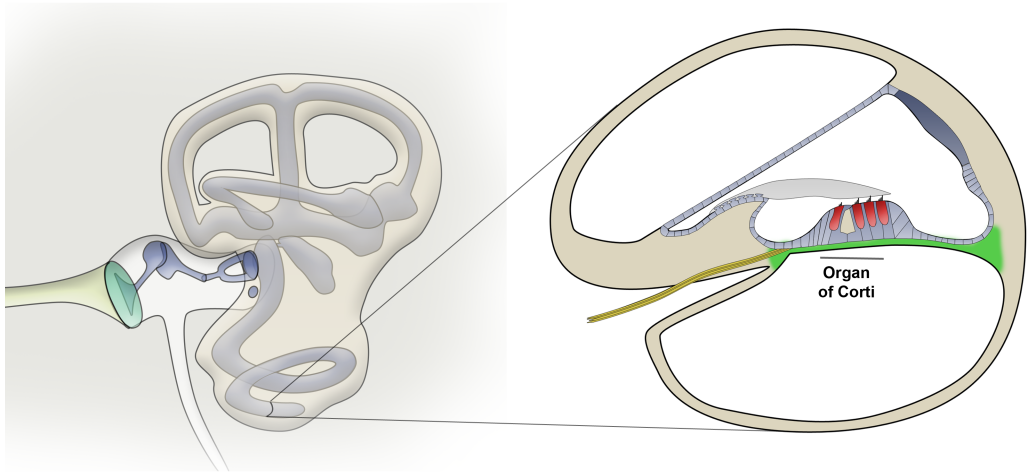


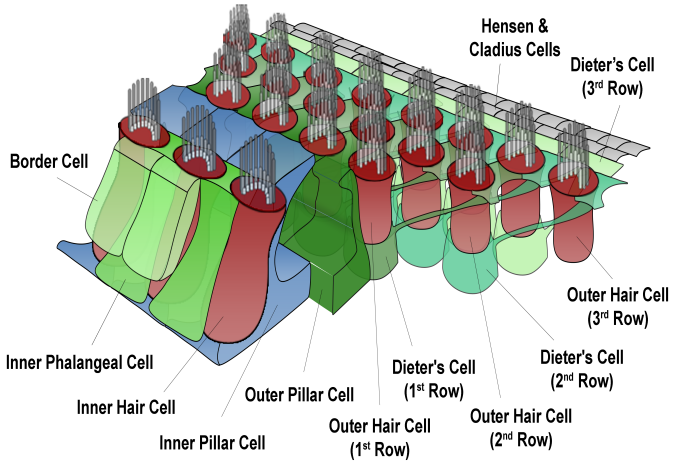
- First single cell whole transcriptome single cell RNA-Seq used manually picking of cells (2009)
- More widely adopted in 2012/2013 with Fluidigm C1 platform and SMARTer chemistry
- Huge increase in throughput with droplet based methods in 2015 (Drop-Seq / InDrops)
- Third generation of methods may see additional increase in throughput / decrease in cost (sciRNA-Seq / SPLiT-Seq / Seq-Well) ~2017/2018

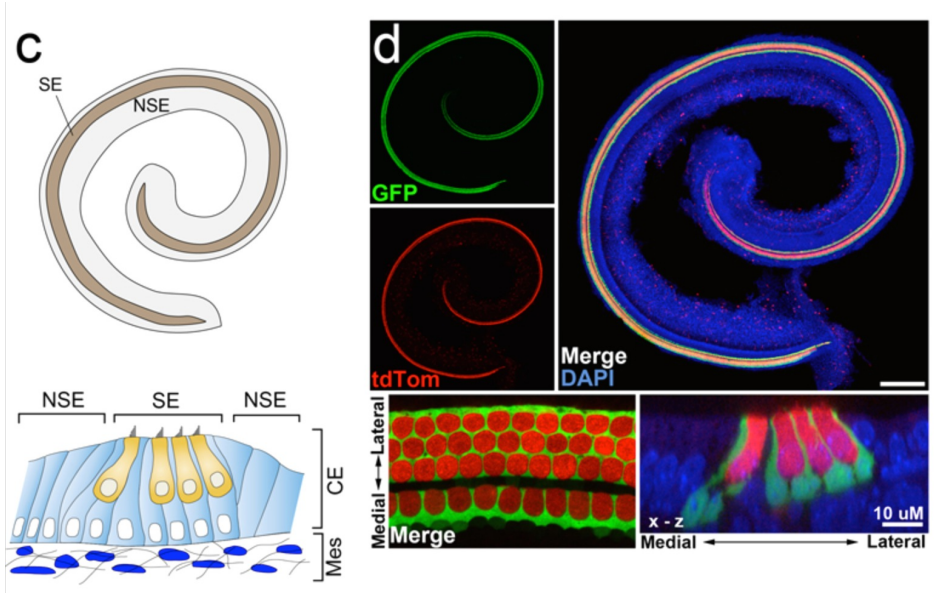
Svensson et al. 2018

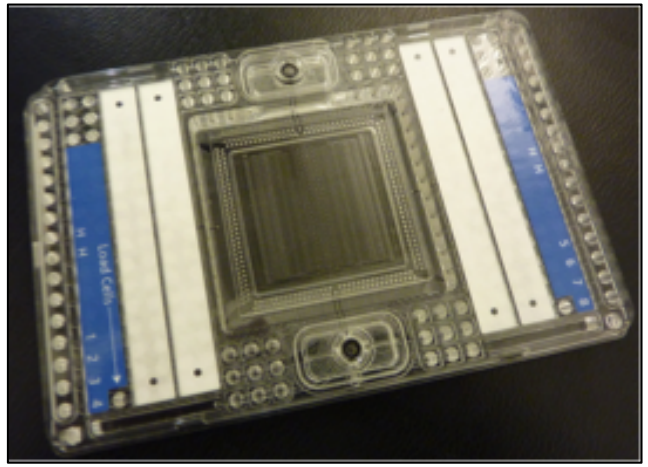
Profiling unique cells types in the mouse cochlea with single cell RNA-Seq - Fluidigm's C1 makes it accessible

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute



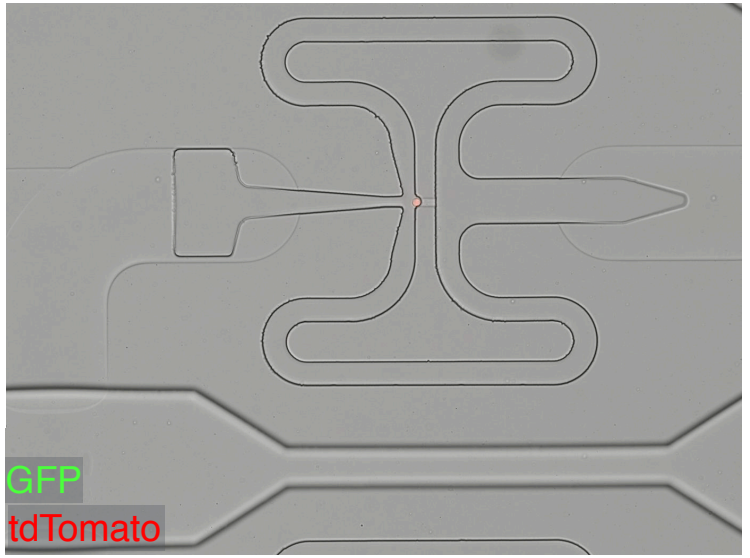


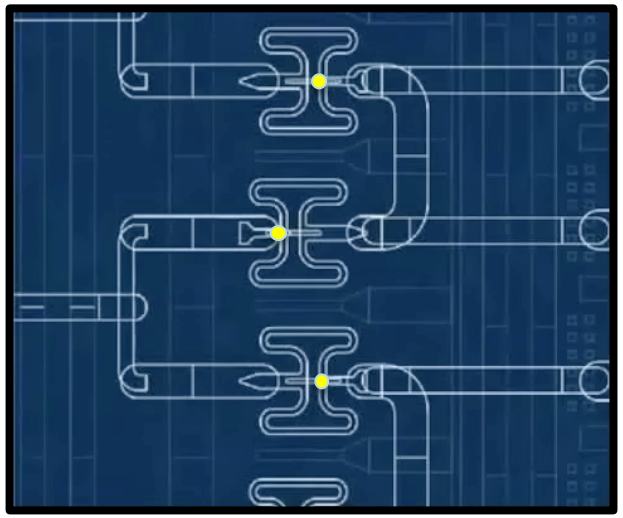


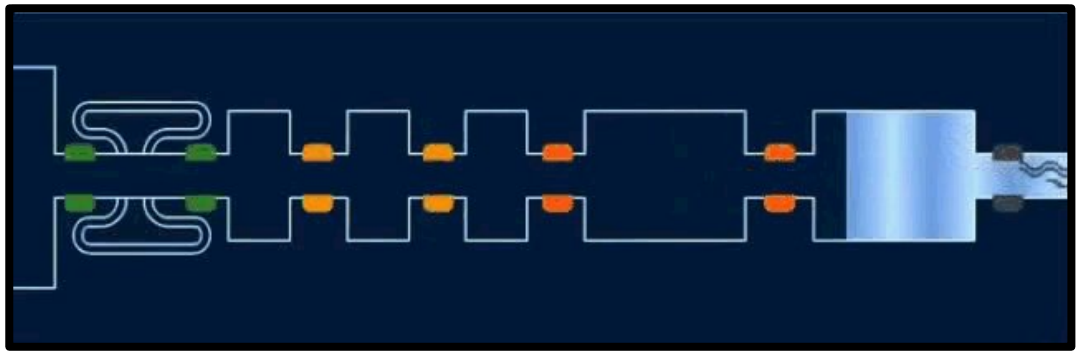


Fluidigim C1 - Cells are captured and imaged on the microfluidic chip before Lysis, RT and PCR

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

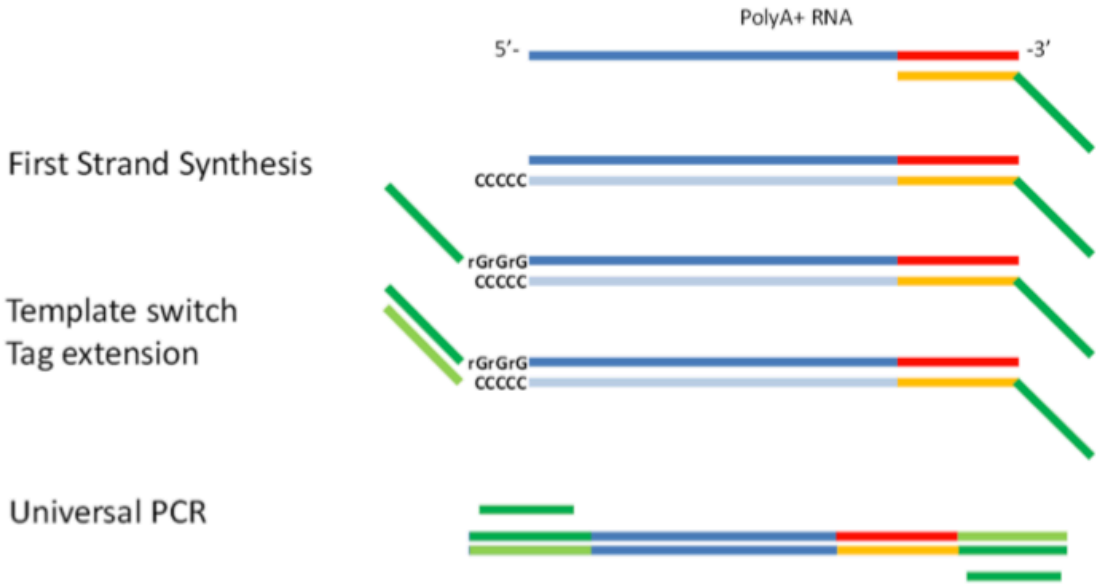


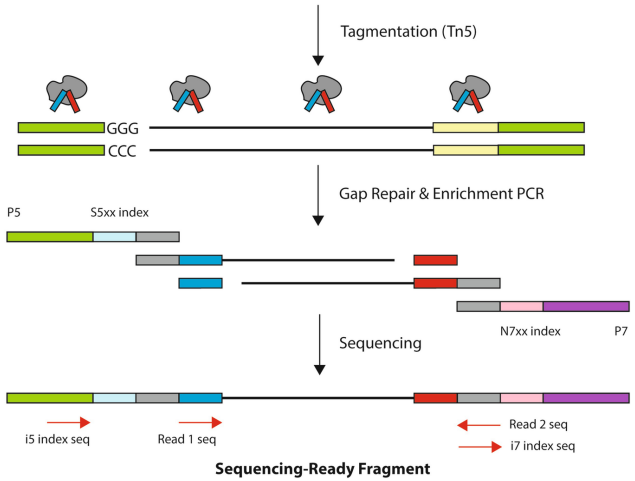




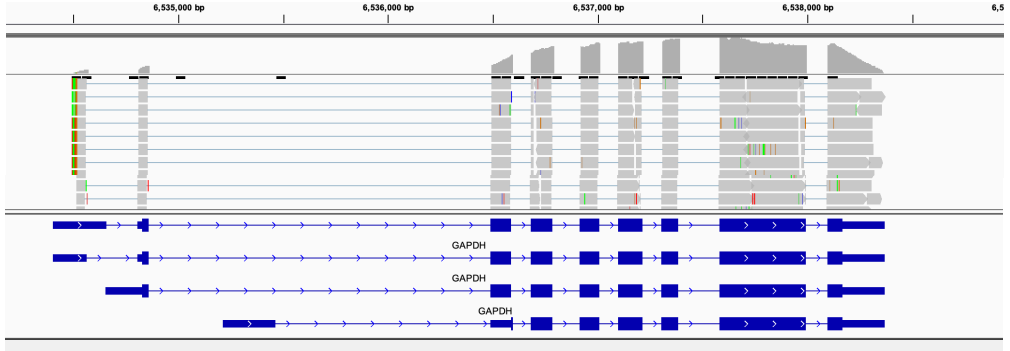
Output is amplified single cell cDNA in 96-well format

SMART-Seq – full length cDNA generated with template switching is then fragmented and indexed for multiplexed sequencing

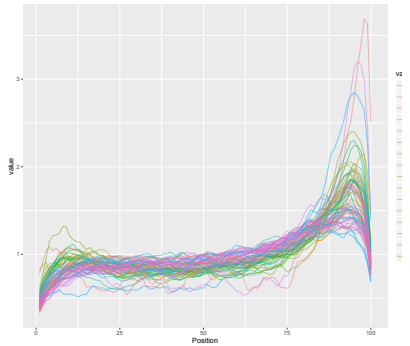




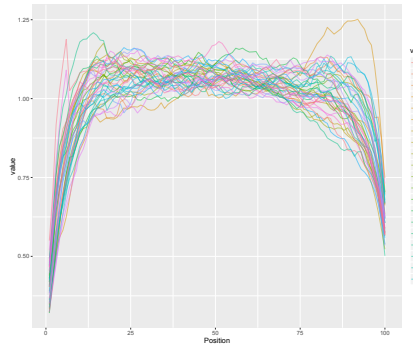
For first scRNA-Seq study, data was processed and reprocessed *many* times using Biowulf…



5' to 3' Coverage Plots



3' biased coverage

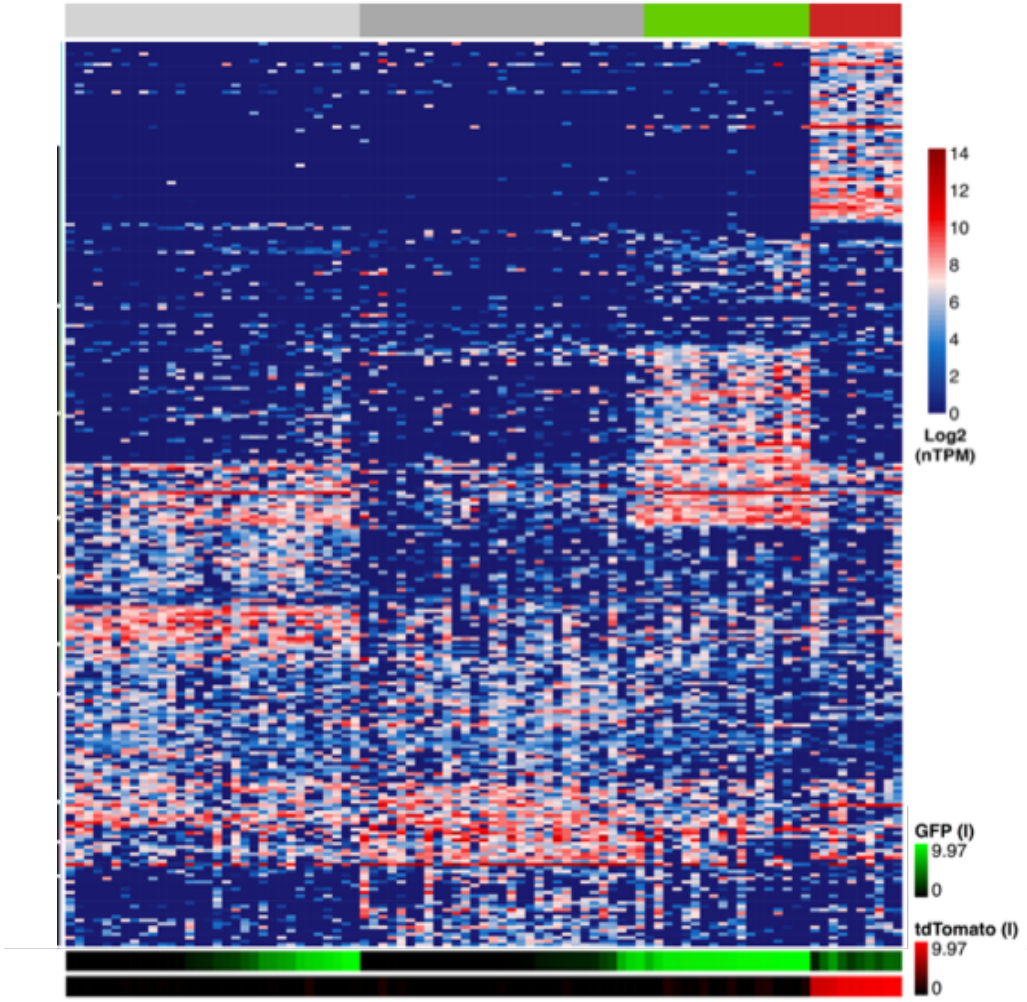


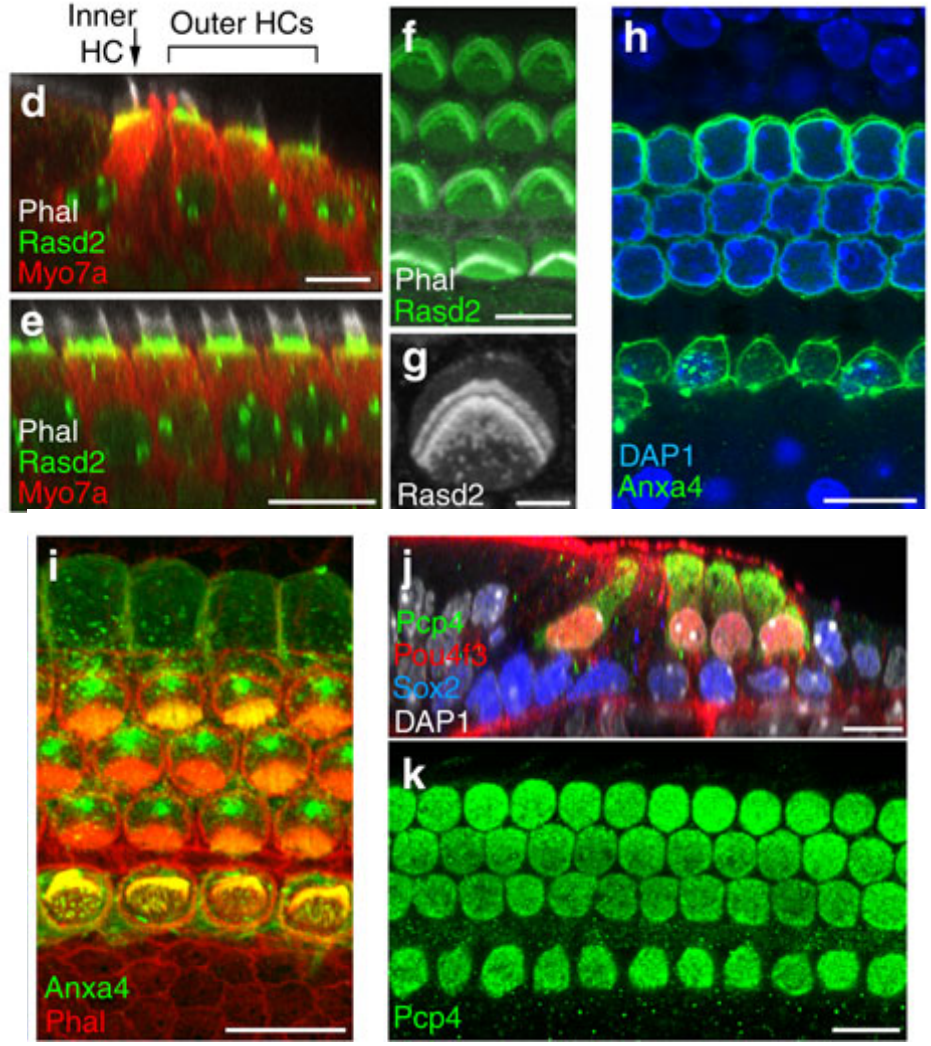
good coverage

P1 single cell transcriptional analysis shows distinct profiles of cochlear hair cells and supporting cells

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

113 single cell libraries!

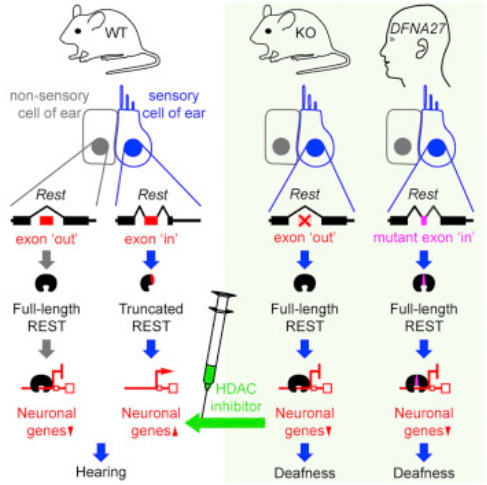




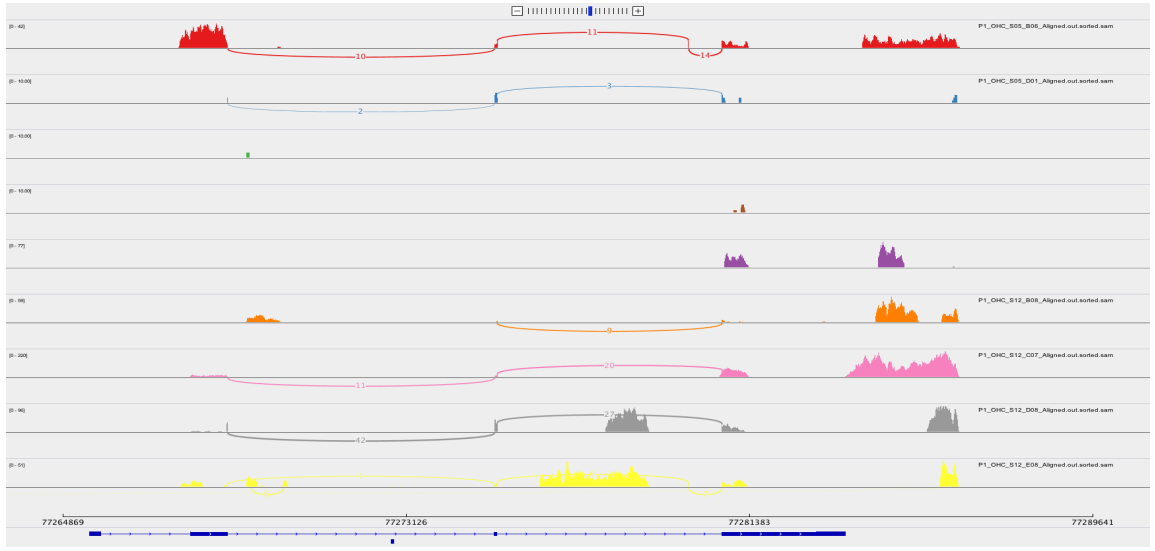
Burns & Kelly et al. Nat Comm 2015

Work done as a fellow in Kelley lab (NIDCD)

Full-length scRNA-Seq allows discrimination of isoforms… depends on expression level and 5-3’ coverage



Mutation in mice and humans in intron of REST leads to mis-splicing and deafness. Cell type specific phenotype. Should be able to detect exon usage in our existing single cell data…

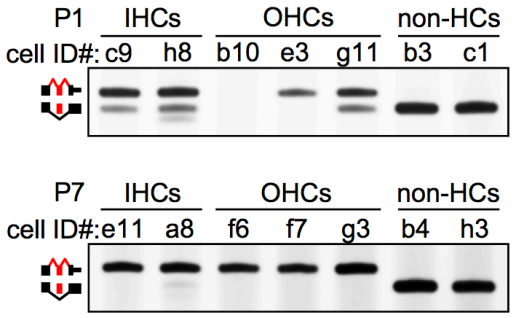


C1

Biomark HD

Whole transcriptome scRNA-Seq data was too sparse and had 3’ bias – not reliable enough.

Solution was to generate specific-target amplified cDNA where cell identity was determined and splicing in single cells confirmed by PCR

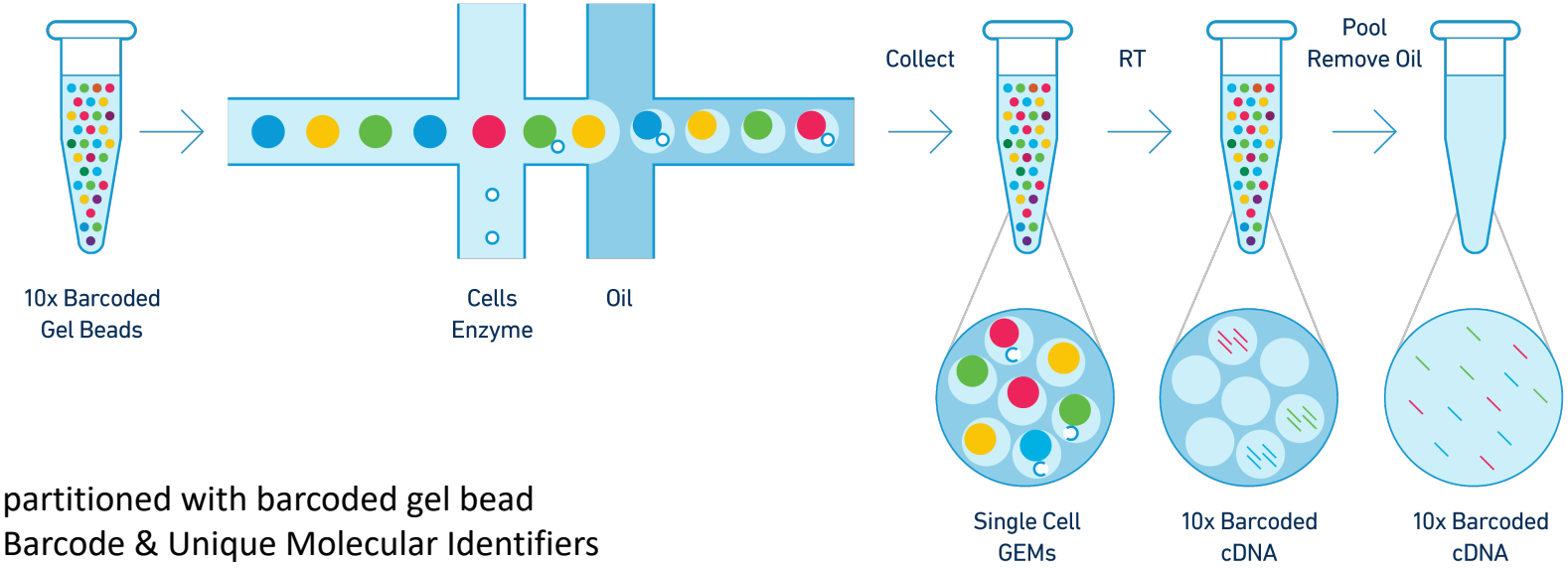


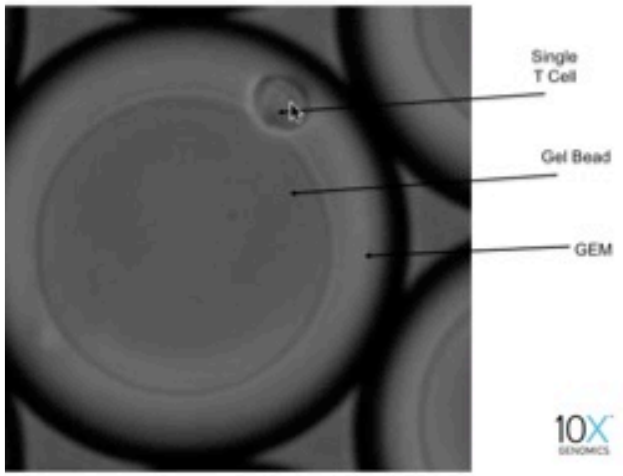
Nakano et al. Cell 2018

Collaborative work with Banfi Lab (Univ of Iowa) done as a fellow in Kelley lab (NIDCD)

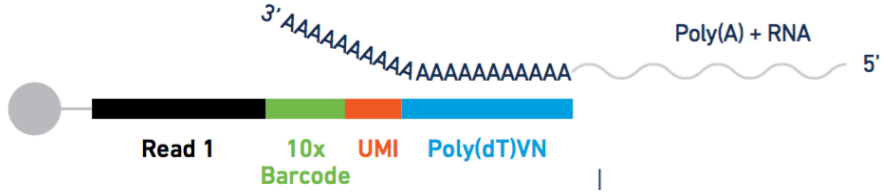
Droplet-based sc-Seq allows high-throughput profiling with unprecedented ease *(for good samples)*

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute





- Cell partitioned with barcoded gel bead
- Cell Barcode & Unique Molecular Identifiers (UMI) on end of cDNA
- ~2000 median genes detected on average with ~50,000 mean sequencing reads per cells
- Gene-level counts data



FACs-enrichment of limited target cells to assay transcriptional effect of transcription factor overexpression

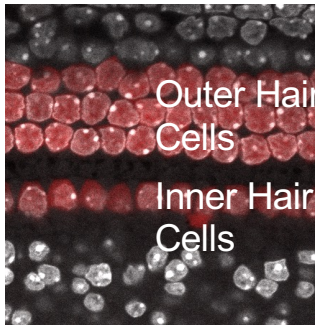
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

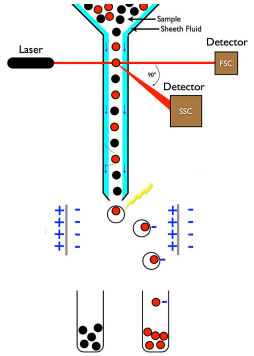
Injected w/ Anc80-EGFP or Anc80-Ikzf2

P1 → P8

Harvested, Hair Cells Enriched

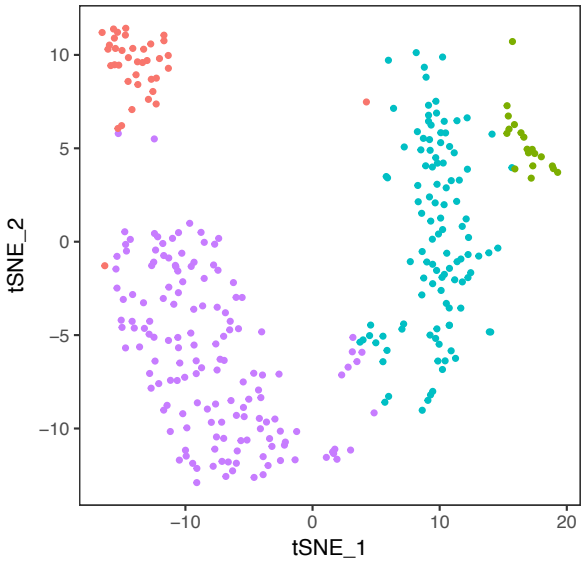
FACS Enrichment of TdTomato+ Hair Cells





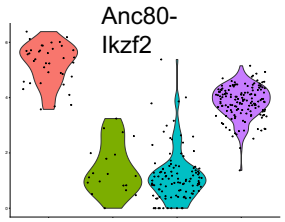
Modified from Wikimedia.org

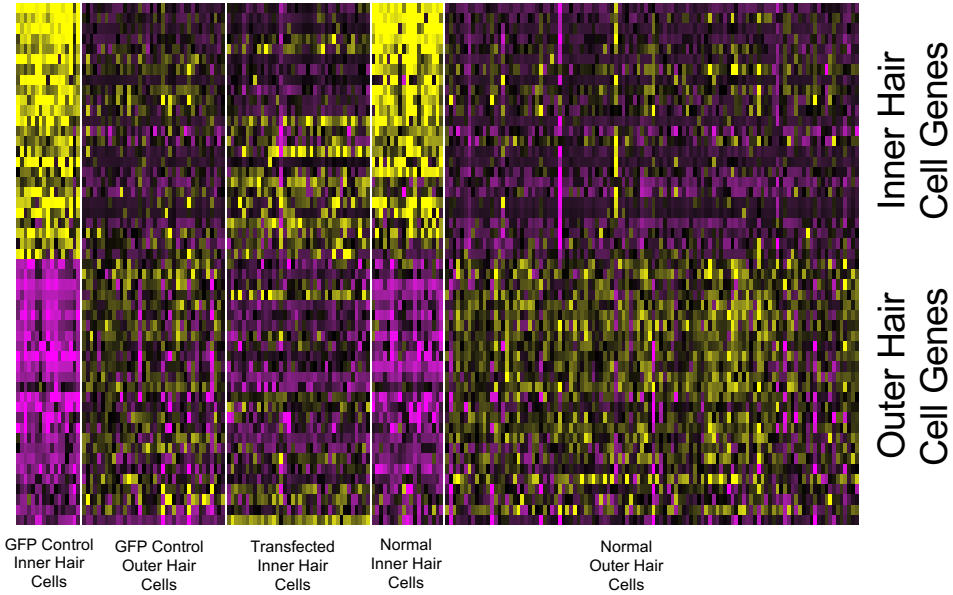
10X Capture



Anc80-Ikzf2 Transfected Cochlear Single Hair Cells

- Unknown – High Ikzf2
- Inner Hair Cells
- Outer Hair Cell Group 1
- Outer Hair Cell Group 2

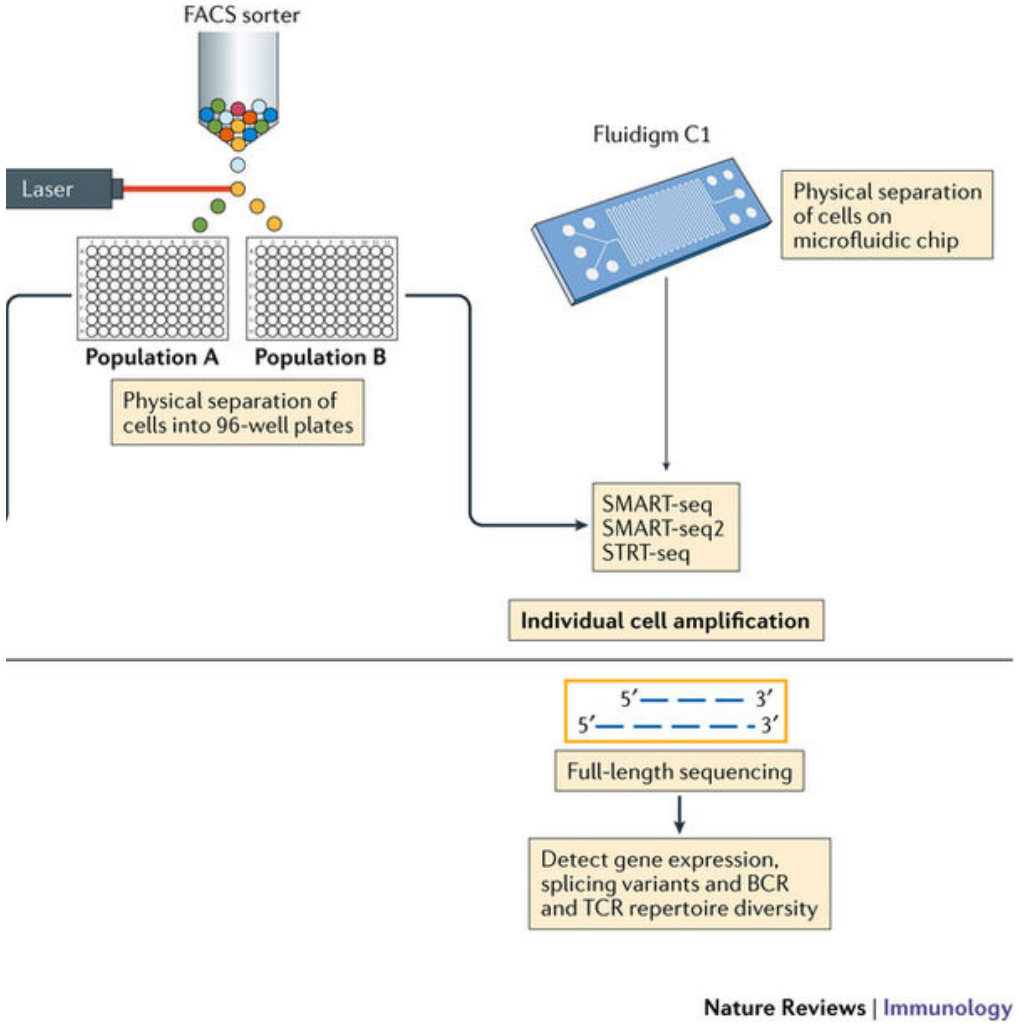




Chessum & Matern et al. Nature 2018

Collaborative work with Hertzano Lab (UMBC) done as a fellow in Kelley lab (NIDCD)

Plate-based sc-Seq still has its utility and can be readily implemented with FACs capable of plate-sorting

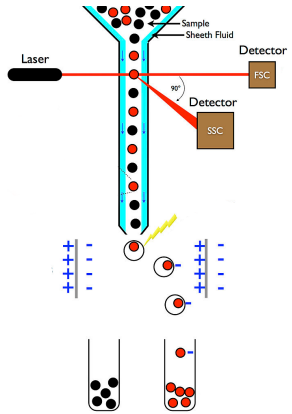


Papalexi & Satija (2018)

- Cells sorted directly into lysis in plates (no time for transcriptional changes due to sorting)
- Both 'homebrew' (i.e. SMART-Seq2) and commercial reagents exist (i.e. Takara SMARTer v4, Qiagen 3' UPX)
- Allows for full-length transcript detection (isoform-level analysis possible)
- In general, greater gene detection sensitivity expected
- Higher cost per cell than droplet-based methods
- Generally more hands-on time for processing libraries (depending on sample pooling stage)

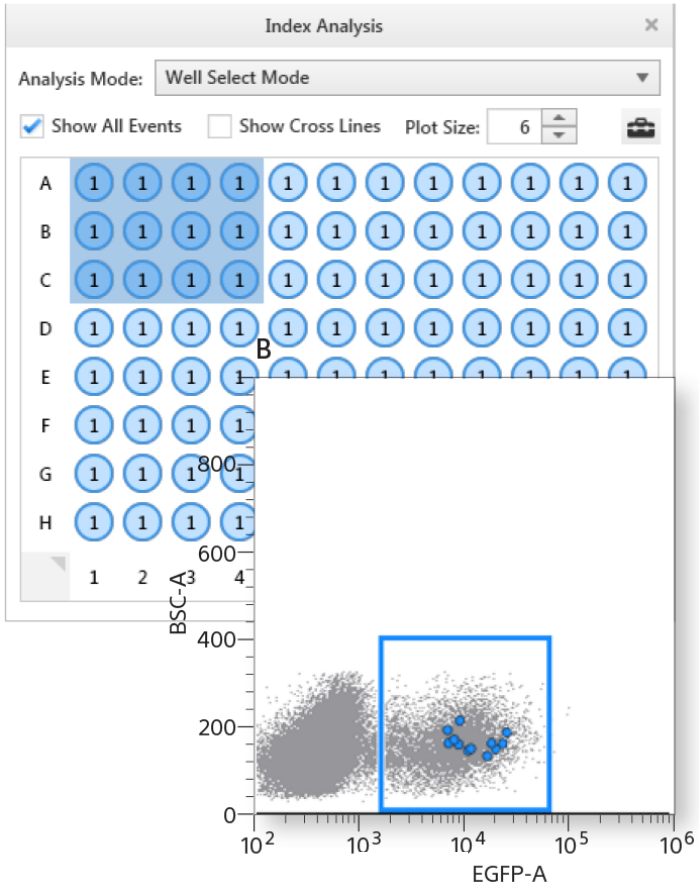
Index sorting single cells for plate-based SMART-Seq2 single cell RNA-Seq analysis of target cells

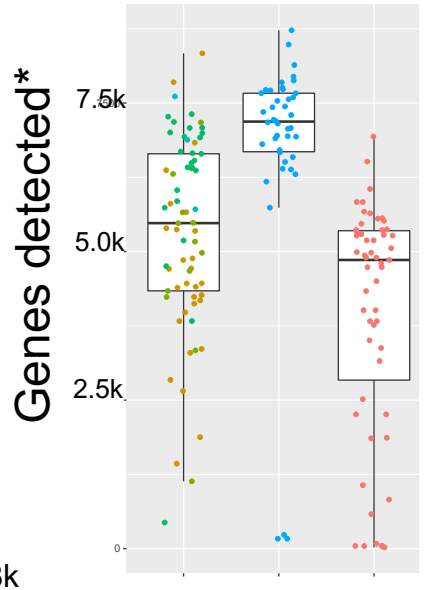
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute



Modified from Wikimedia.org

Index sorting – keep track of phenotypic parameters





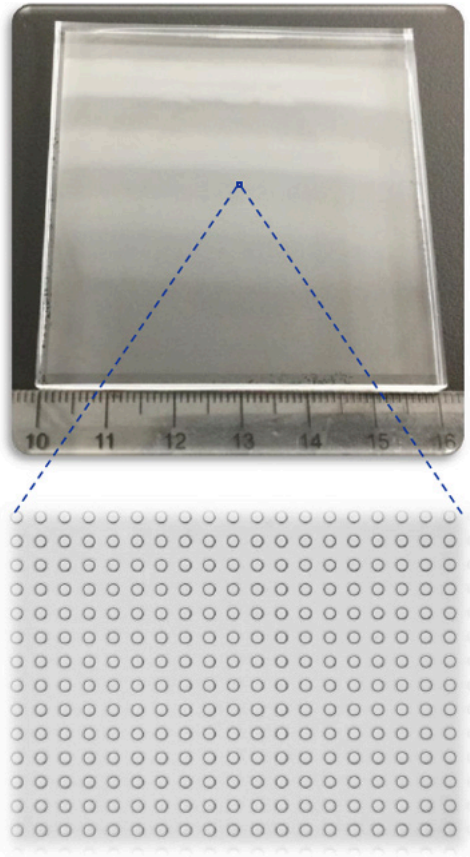
*58k Ensembl Reference

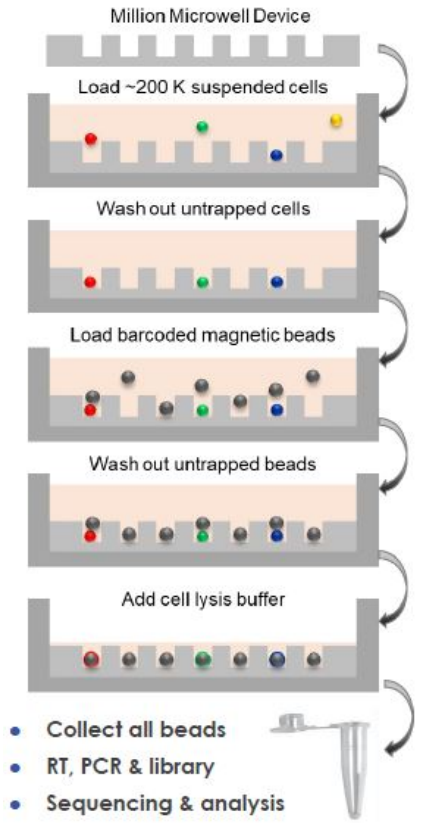
- Full-length coverage & better gene detection
- Low throughput* & UMIs are not standard – not readily accessible with Illumina short read format

RT Enzyme
SS: SuperScript
PS: PrimeScript

Higher throughput than droplet-based single cell?

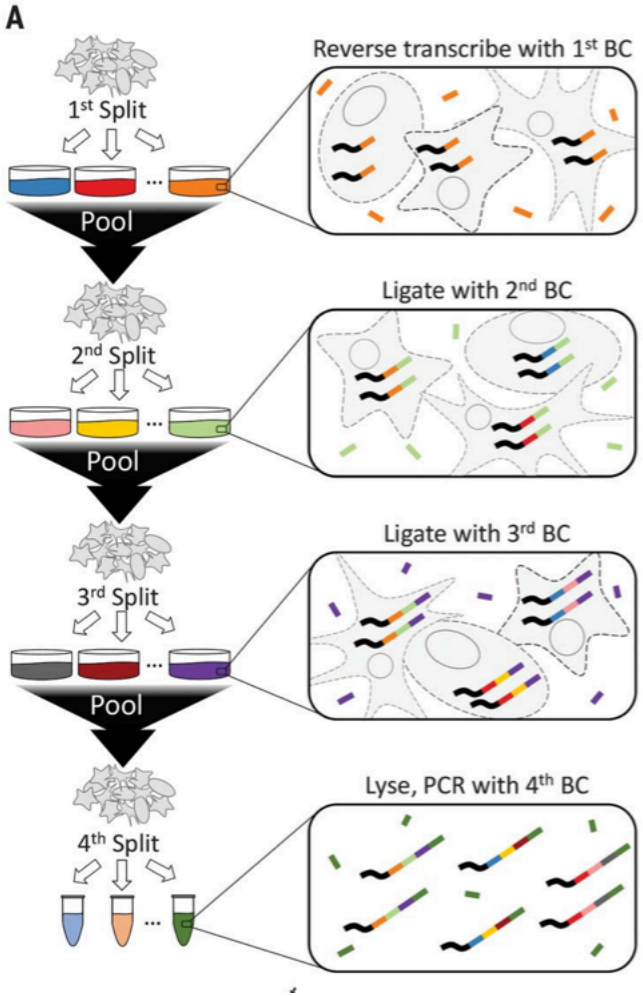
Microwell barcoding





https://www.tebu-bio.com/

Combinatorial Indexing



Rosenberg & Roco Science 2018

Beyond single cell RNA-Seq (scRNA-Seq+)

Number of accessible modalities available to interrogate at single cell resolution has increased

Gene Expression Profiling & Add-on modalities (scRNA-Seq+)

- **scRNA-Seq gene expression profiling**
- **VDJ sequencing for TCR or BCR expression**
- **Cell surface protein measurements with barcode-conjugated antibodies**
- **Expressed barcodes to track clonal relationships**
- **Functional genomics with CRISPR-based perturbations**

Chromatin accessibility

- **scATAC-Seq**

Genomic structural variation

- **scCNV**

Number of accessible modalities available to interrogate at single cell resolution has increased

Gene Expression Profiling & Add-on modalities (scRNA-Seq+)

- **scRNA-Seq gene expression profiling**
- **VDJ sequencing for TCR or BCR expression**
- **Cell surface protein measurements with barcode-conjugated antibodies**
- **Expressed barcodes to track clonal relationships**
- **Functional genomics with CRISPR-based perturbations**

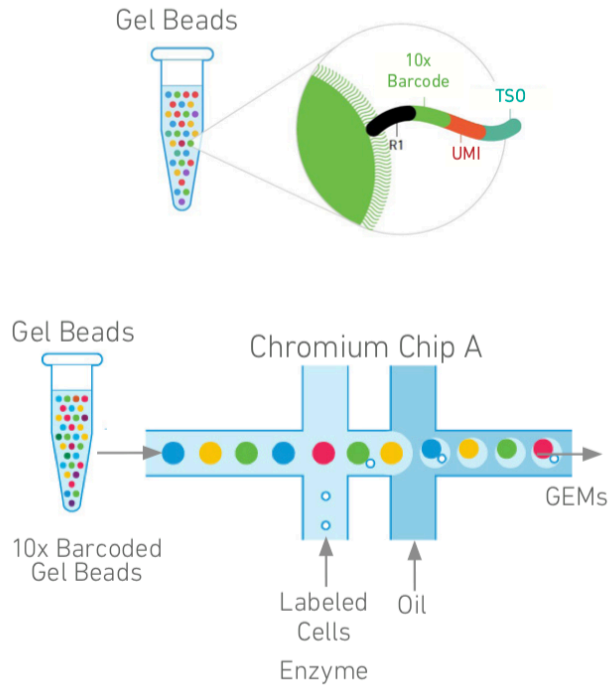
Chromatin accessibility

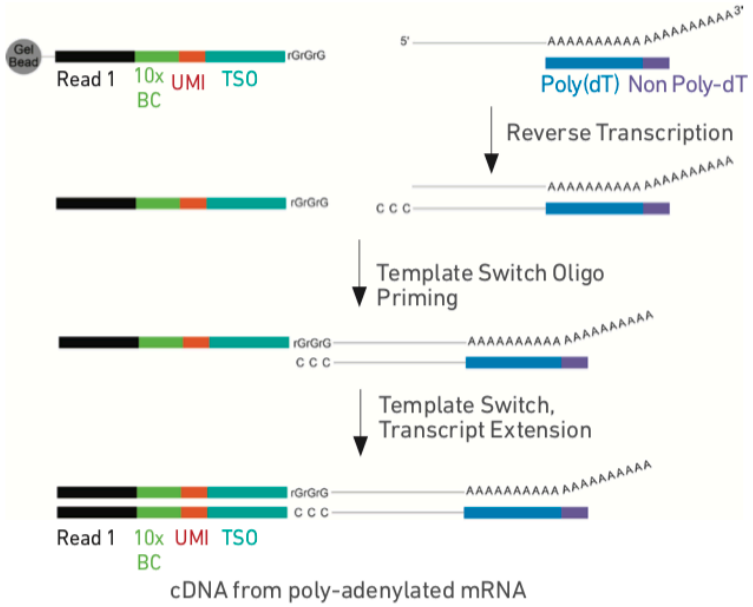
- **scATAC-Seq**

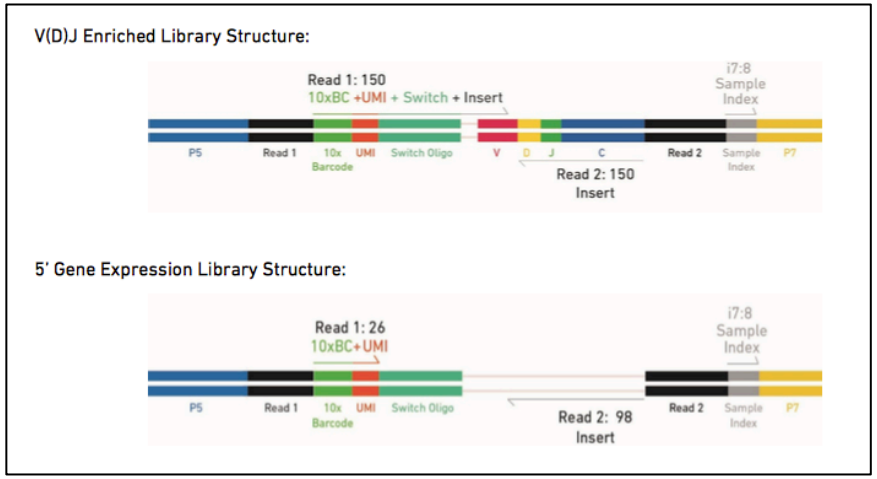
Genomic structural variation

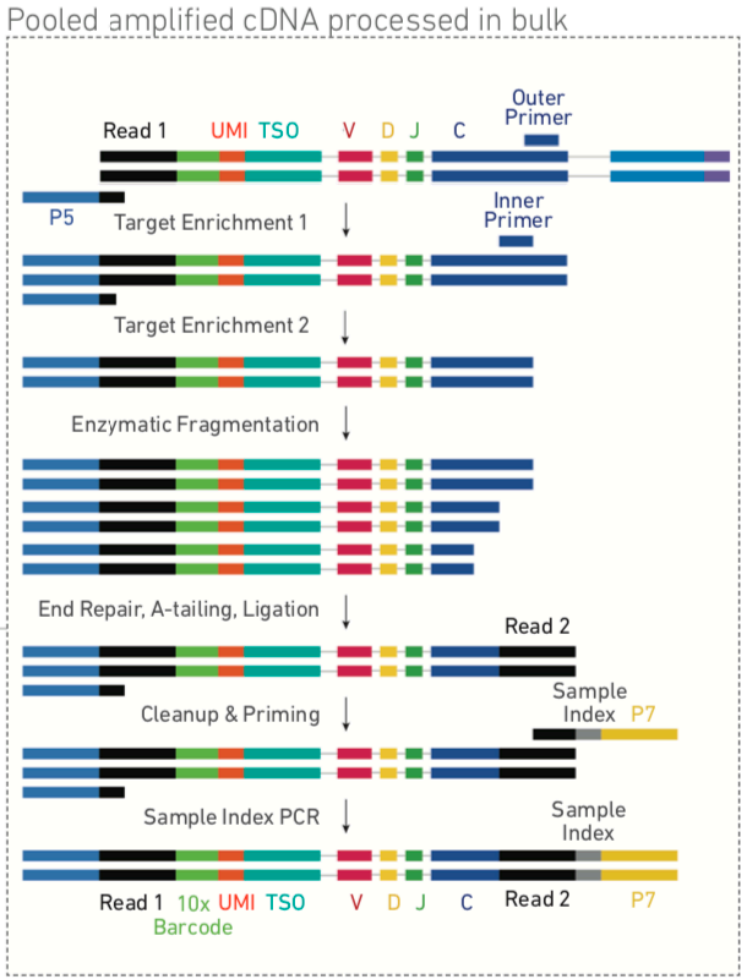
- **scCNV**

5’ end barcoding single cell allows gene expression and VDJ (T-cell and B-cell receptor) sequencing from the same cells



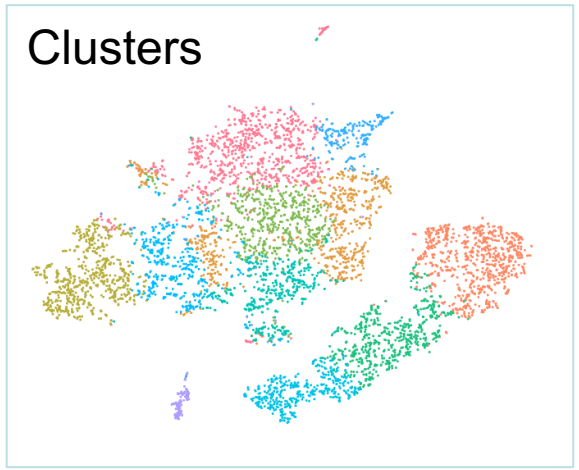


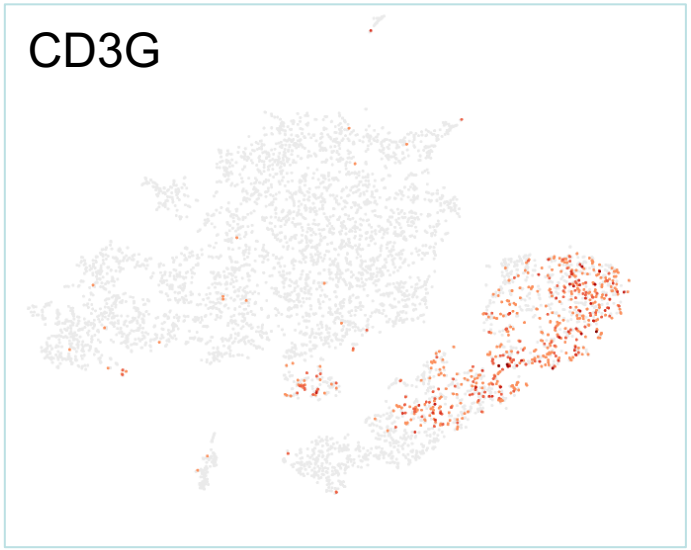


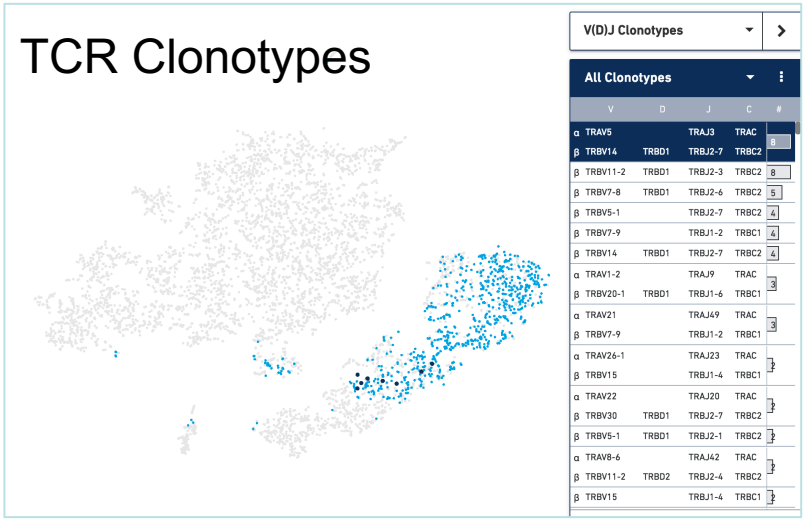


Specific target genes can be amplified from the single cell cDNA library and sequenced – linked by cell barcodes

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

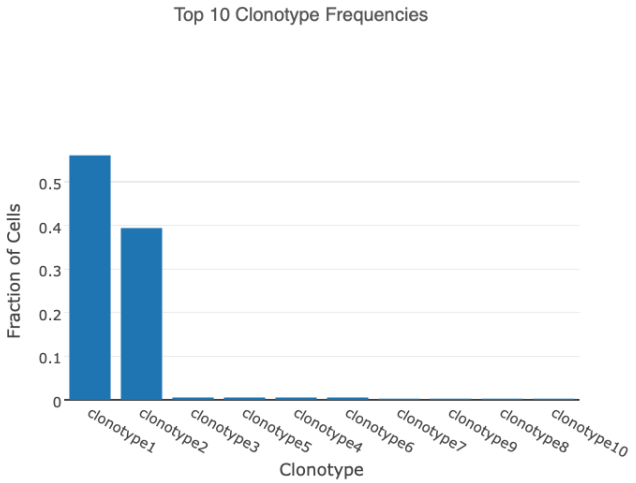






Applications:

-Determine genes associated with an expanded clone (malignancy)
-Determine phenotype of cells activated and responding to target (tumor killing T-cells)
-Determine the TCR of an expanded T-cell clone effective at responding to a target

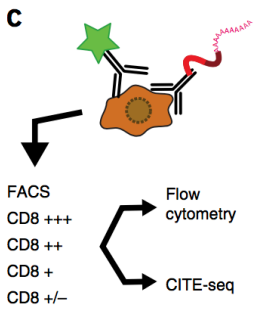


Notes:
-Higher chance of getting full VDJ from high viability samples
-VDJ sequencing requires longer sequencing reads = more expensive
–Analysis of a diverse repertoire takes some additional work

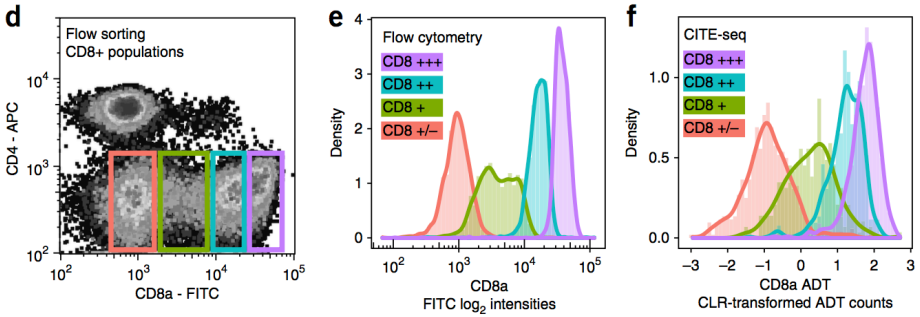
Feature barcoding for cell surface protein measurement in parallel to gene expression profiling

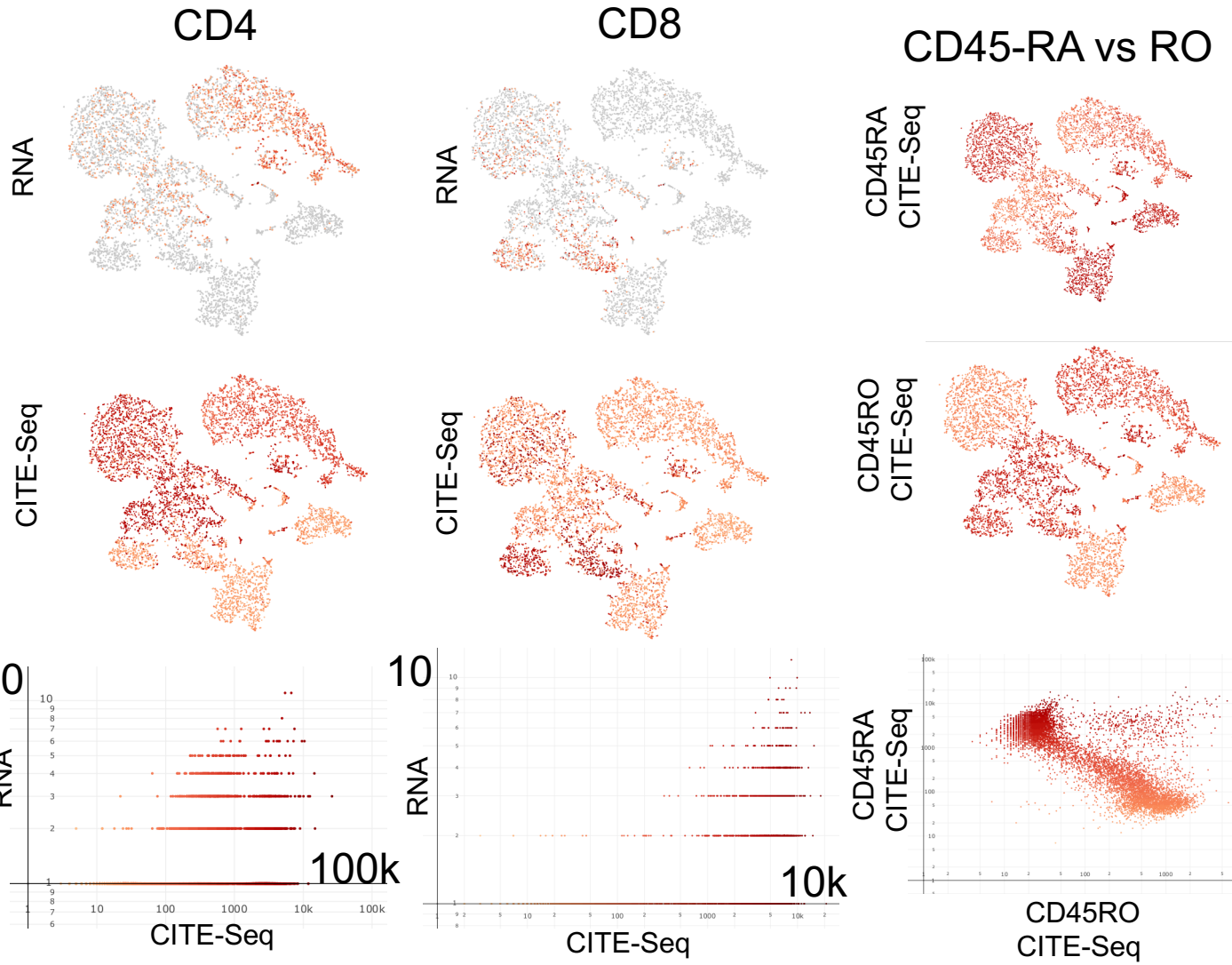
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

'CITE-Seq' = 'AbSeq' =
'Antibody Feature Barcoding'



Poly-adenylated antibody (or other feature label) labels cells of interest – similar to FACS





10x Genomics Example Dataset (10k PBMC 3' v3 with Feature Barcodes)

Sample preparation methods

scRNA-Seq capture input of a high viability cell preparation is important for good data

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

Single Cell Suspension for Optimal Performance

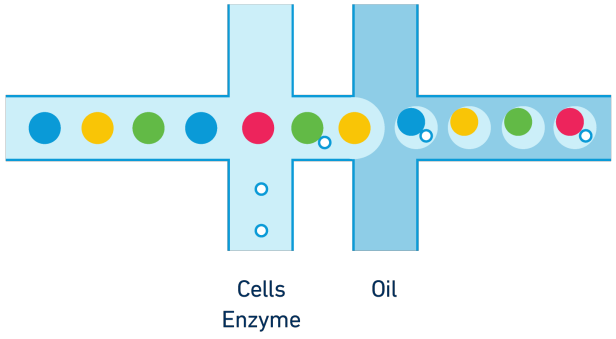
10x Genomics® Single Cell Protocols require a suspension of viable single cells as input. Minimizing the presence of cellular aggregates, dead cells, non-cellular nucleic acids and potential inhibitors of reverse transcription is critical to obtaining high quality data.

1.3. Factors Influencing Cell Recovery

To recover the expected number of cells, it is critical to maximize viability, minimize the cell preparation time, accurately measure the input cell concentration and pipette the correct volume into each reaction.

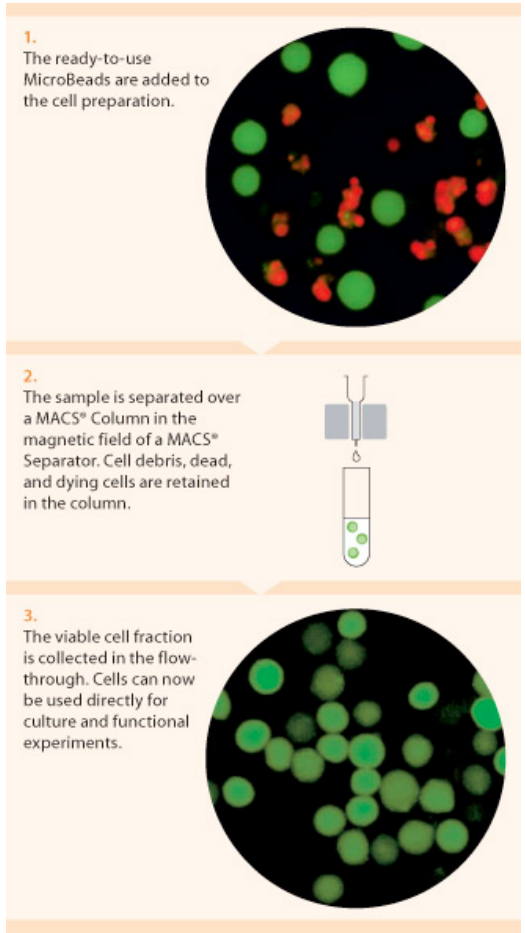
Ideally, input cell suspensions should contain more than 90% viable cells. Non-viable and dying cells generally contain less and more fragmented RNA that may not be efficiently captured by 10x Genomics Single Cell Solutions. The presence of a high fraction of non-viable cells in the input suspension may therefore decrease the apparent efficiency of cell partitioning and recovery.

- **10X Genomics recommends loading 90% viable cells or higher**
- **Often the underlying causes of below target # of cells**
- **Can contribute a “background signal” – ambient RNA**

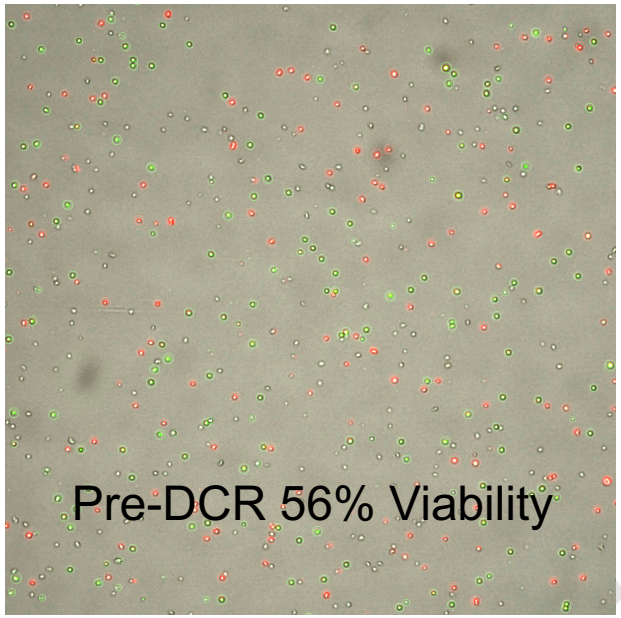


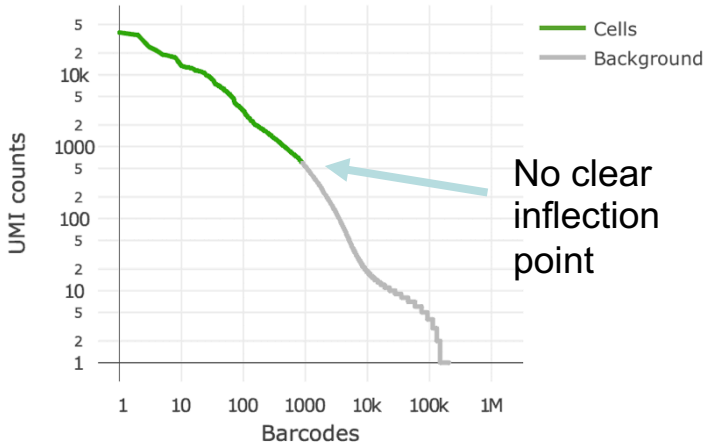
Effects of low cell viability and ways to improve you input population

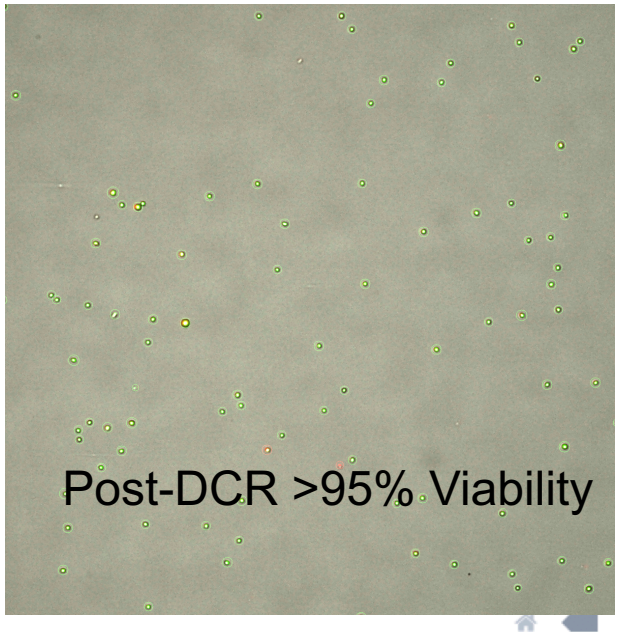
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

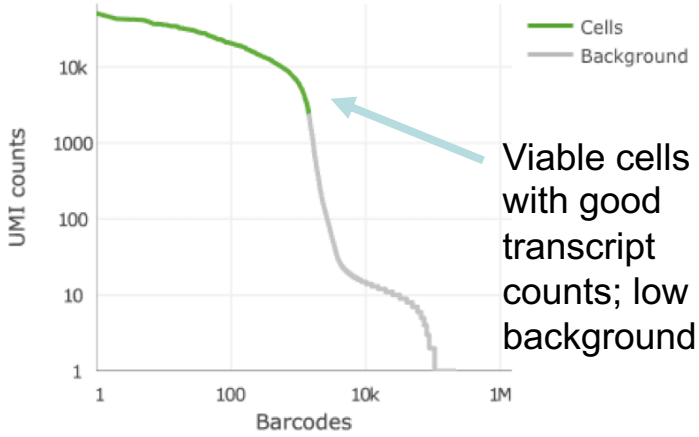


Miltenyi Dead Cell Removal

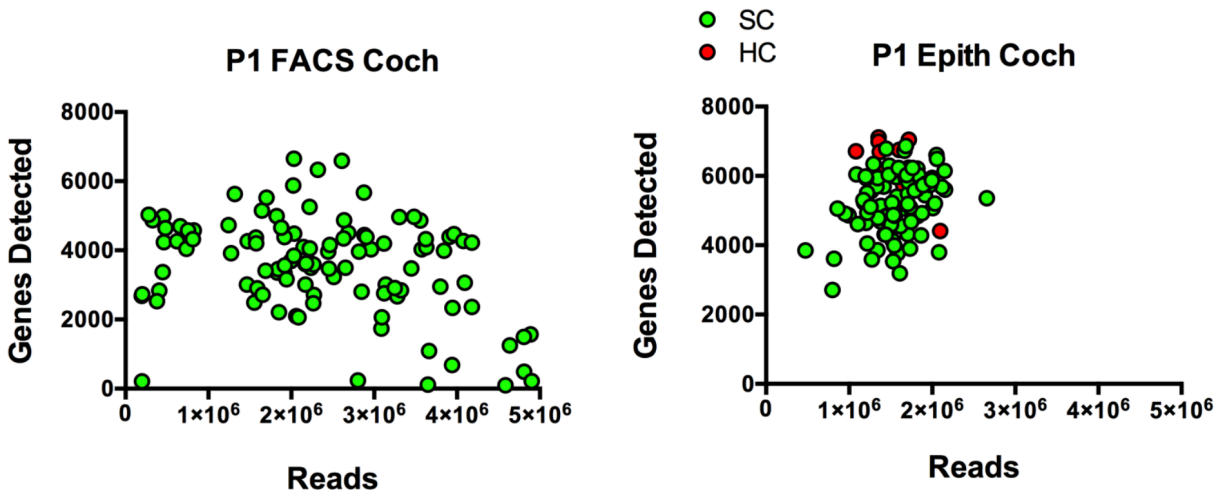


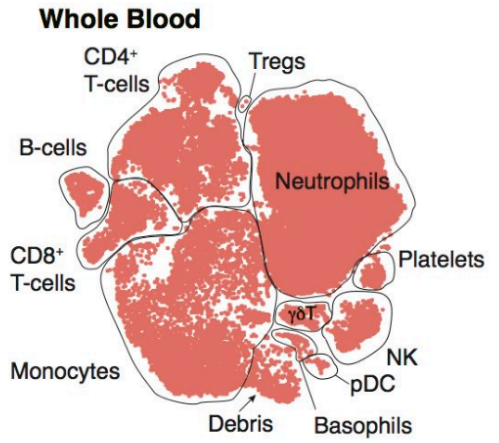


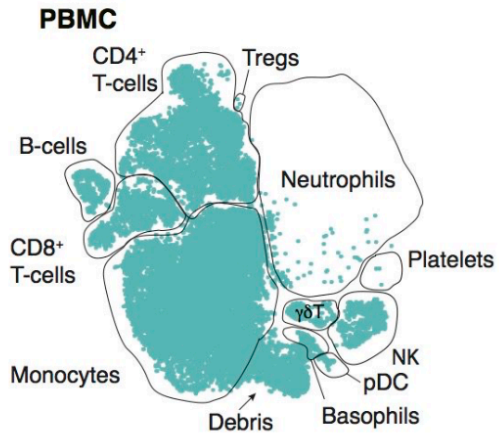




Any preparation / purification step has the potential to change or bias you single cell assays





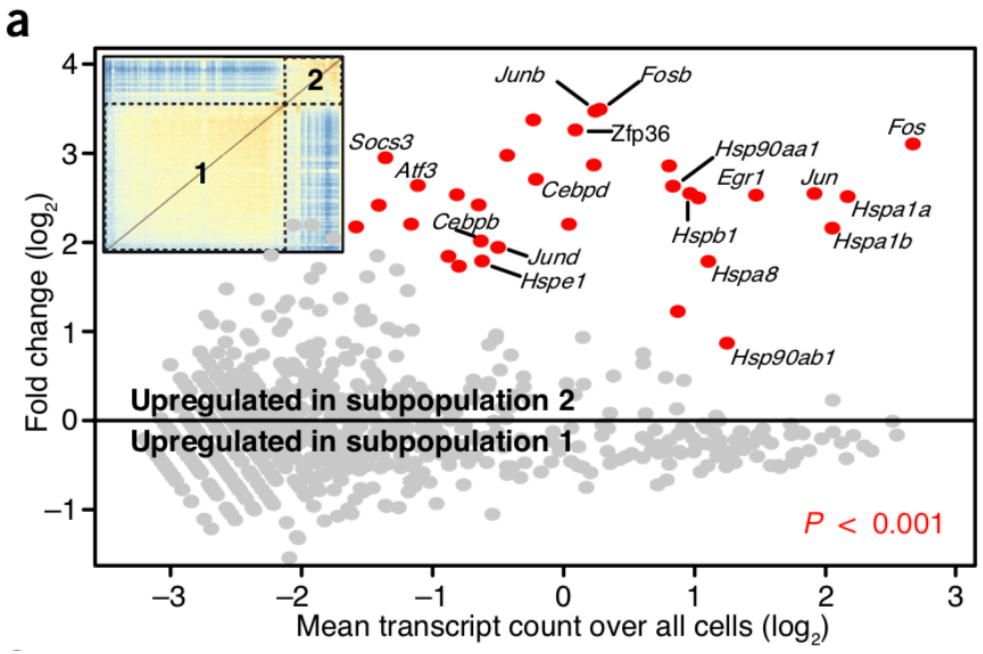


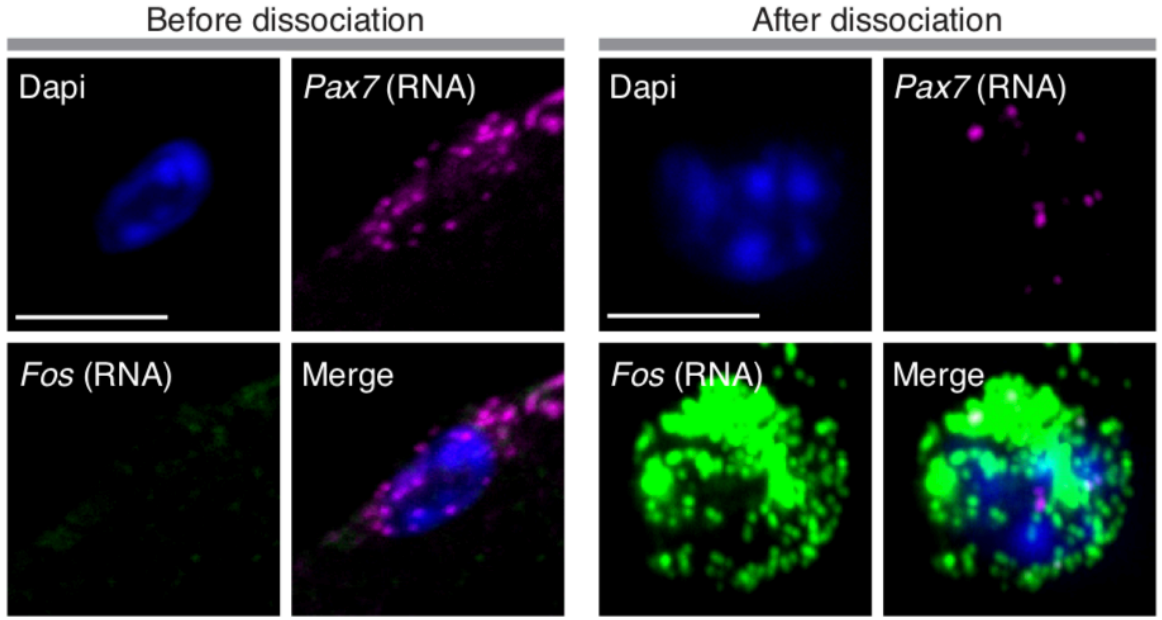
- Removal of dead cells may preferentially remove certain cell types (more fragile) – important to know cell composition of tissue
- FACs enrichment is powerful and allows targeting cell types of interest, but may increase cell stress / decrease RNA content
- Even standard cell preps (PBMC extraction from whole blood) can drastically bias cell type representation in datasets
- What effect does the dissociation protocol have on the physiological state of cells in the tissue?

Brodin et al (2019)

Are you profiling normal biology or the effect of tissue dissociation in your scRNA-Seq study?

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute





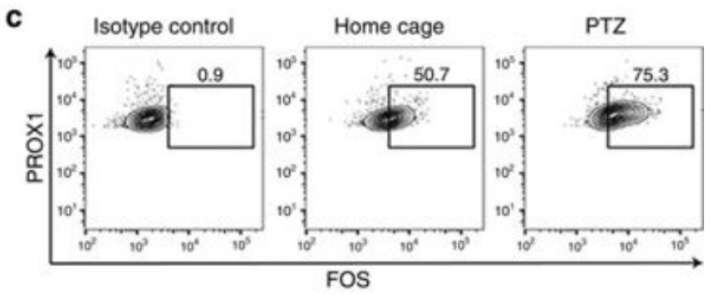
van den Brink et al Nat Methods 2017 PMID 28960196

- **Effects of dissociation and / or sorting and enrichment may have a significant effect on transcriptional state of a cell**
- **How far from the biologically relevant transcriptional profile are we?**
- **Some methods may help reduce this effect (single nuclei profiling, fixed cells, actinomysin D, cold-active protease, etc.)**

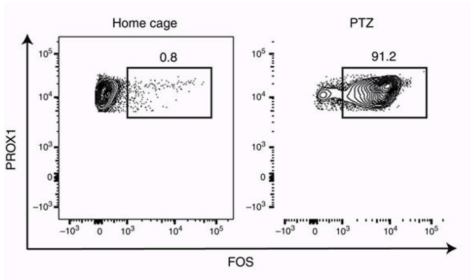
Single nuclei RNA-Seq allows fast-access to cellular transcriptomes with reduced dissociation-associated noise

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

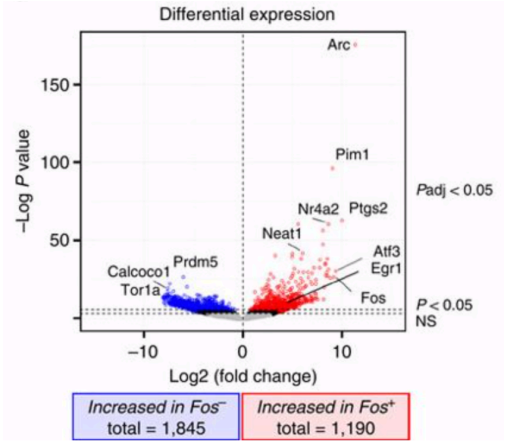
Whole Cell



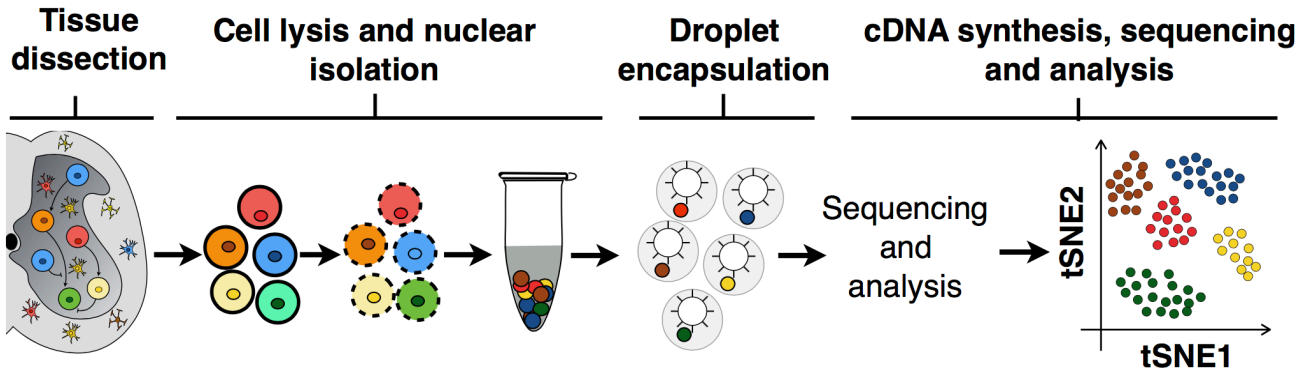
Single Nuclei

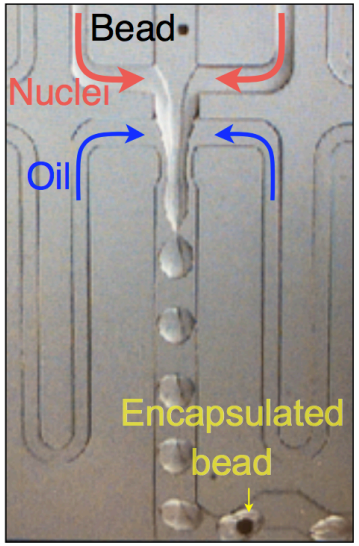


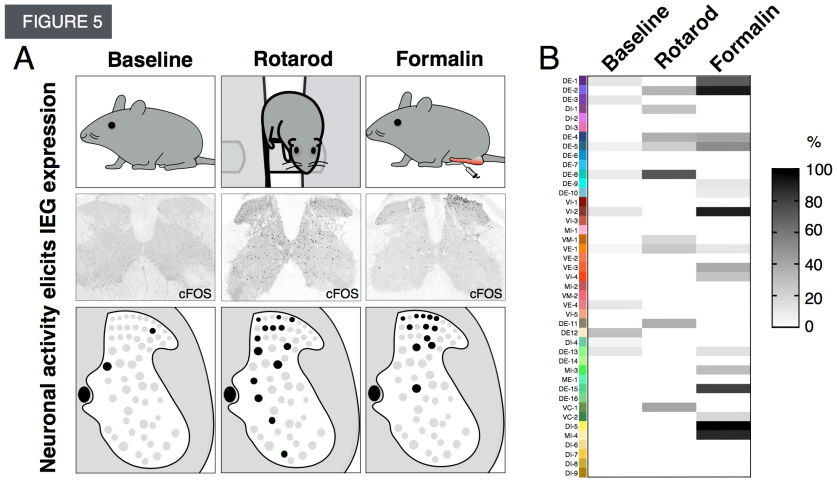
Single Nuclei – Active Neurons Marked by Fos



Lacar et al 2016





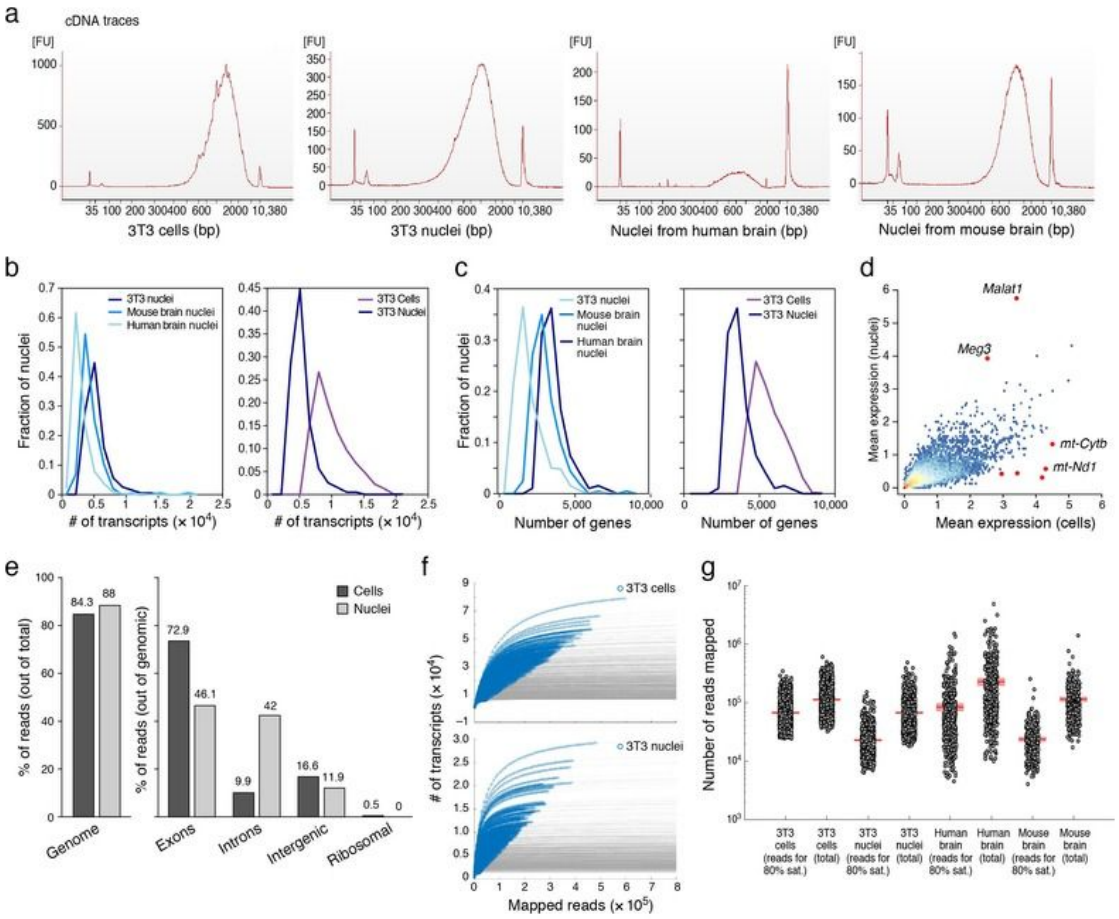


Sathyamurthy et al. Cell Reports 2018

Collaborative work with Levine Lab (NINDS) done as a fellow in Kelley lab (NIDCD)

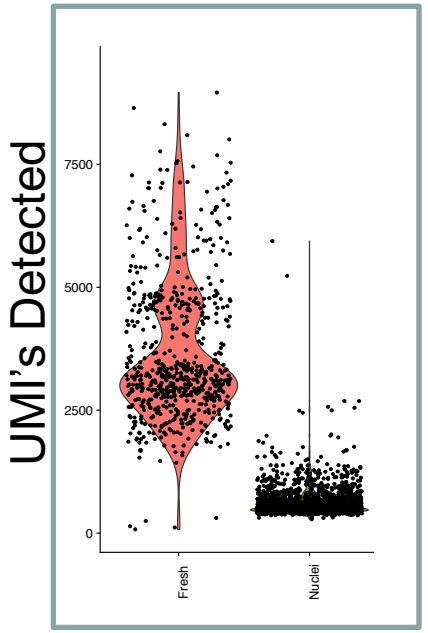
Nuclei have different RNA – don't expect exact equivalence to whole single cell RNA-Seq – still a powerful technique

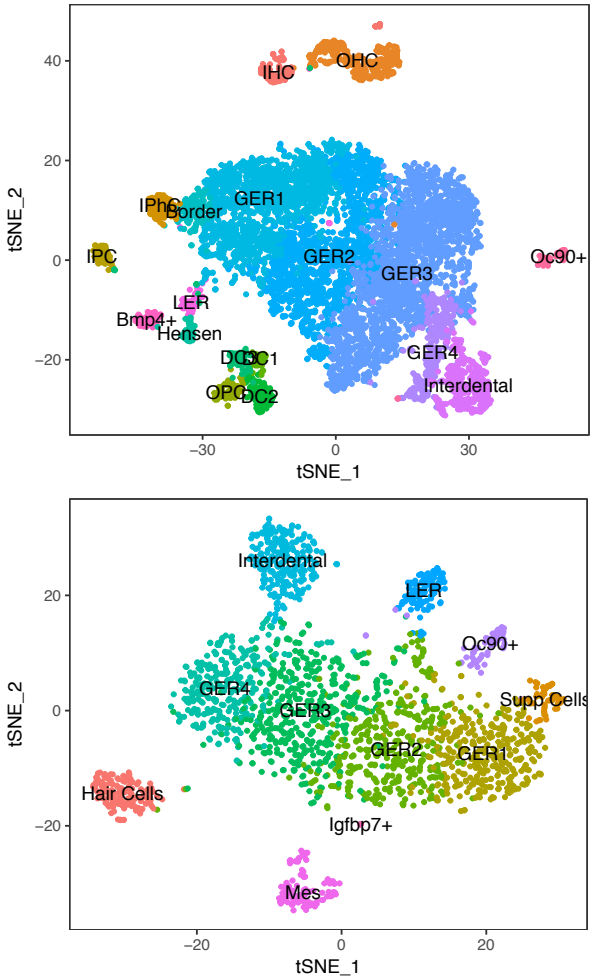
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute



Habib et al 2017 Nature Methods "DroNc-seq"

Evaluating sensitivity and effectiveness with cochlear epithelial cells





Single nuclei have less RNA content (also may be less accessible), and RNA has more introns – still appears to be representative and good for cell type identifications

The single nuclei sequencing option

Advantages of single nuclei

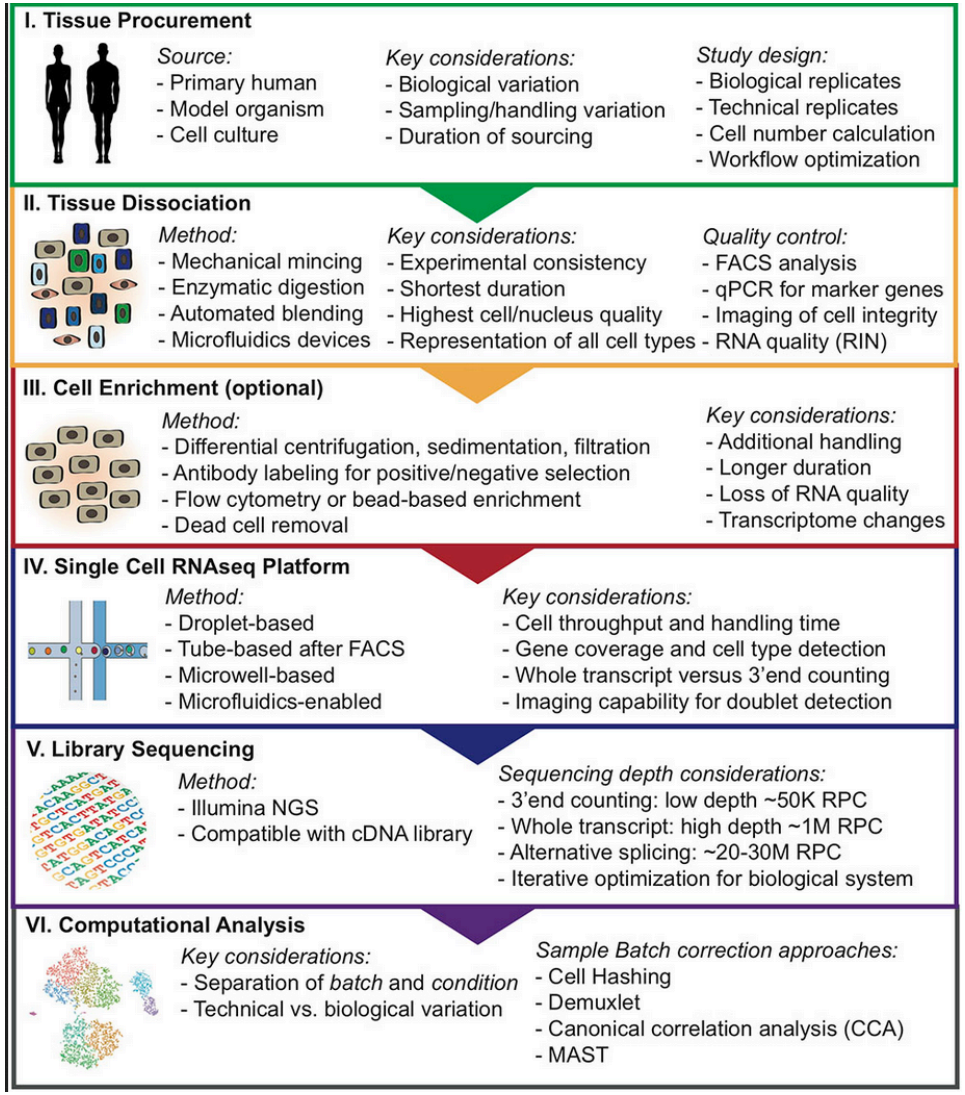
- Fast and relatively easy to extract
- Option for difficult-to-dissociate tissues (highly intercalated cells, strongly-adherent cells, fragile cells, or from solid fresh-frozen samples)
- May be more agnostic to variability in selection bias – better survey of cell types?
- Transcriptional artifacts of dissociation can be largely avoided

Caveats of single nuclei

- Lower RNA content
- Higher proportion of introns retained in transcripts
- Largely not compatible with barcoded antibody technology (more later)
- Some optimization may be required for each tissue
- Less ability to quality control input sample (low viability cells may still have nuclei, but little RNA)

Preparing for batch effects

Many potential sources of variation in single cell – some are what we are looking for; some are unintended



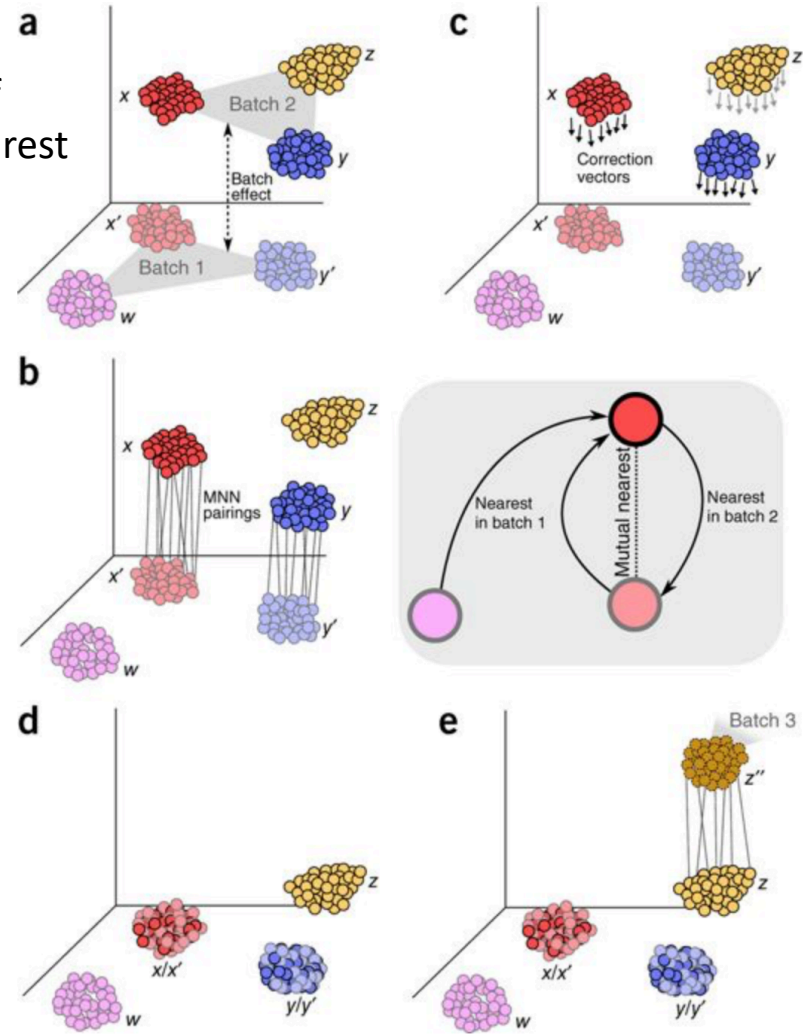
Nguyen et al., 2018

Some general thoughts on variation single cell:

- Inaccuracy in biological measurements have always been a part of biological science
- Biological and technical replicates are helpful, but there can be practical limitations (sample access, cost, etc.)
- Being able to attribute variation to biological or technical aspects of experiment can increase sensitivity to biological signal
- Managing variation across larger studies (particularly human patient longitudinal studies) can be particularly challenging
- Over-correction of 'batch effects' may be just as detrimental to the dataset / conclusions – don't smooth out the biological signal you set out to find
- Standard exogenous RNA spike-ins are generally not useful in high-throughput scRNA-Seq (such as droplet-based scRNA-Seq)

Most data alignment tools are aided (or require) shared cell identities between groups

Principle of mutual nearest neighbor alignment

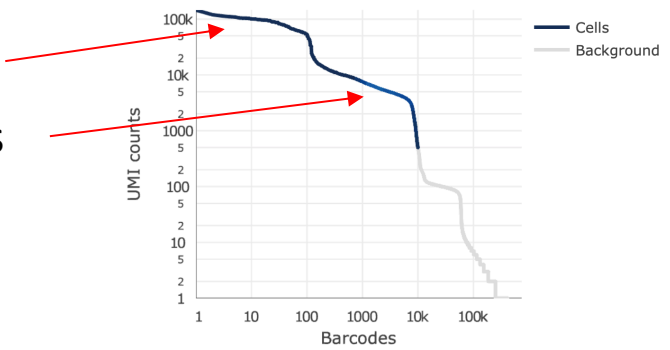


Haghverdi et al 2018

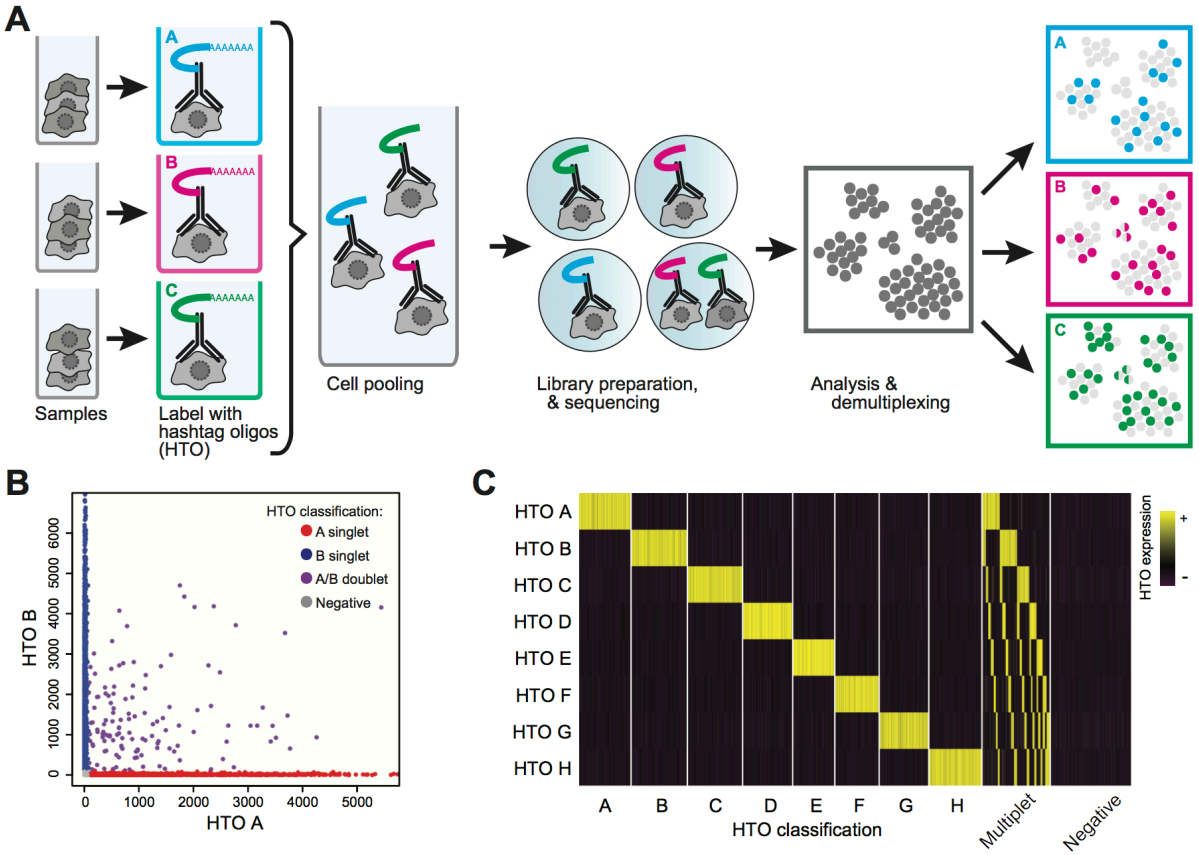
- Samples that inherently have shared cell types between datasets work well
- Even “contaminating cells” in a dataset can help
- Some groups are adding exogenous cells as cellular “spike-ins” for checking or adjusting data alignment
  - *Suggestions from NCI CCR Collaborative Bioinformatics Group*
- If considering, use ‘spike-ins’ cautiously – important to not introduce additional perturbations into the biological system being studied & some reads will go to ‘spike-ins’

Spiked-in cell line

Target primary cells



Feature barcoding for sample multiplexing with gene expression profiling can also reduce sample variation



Benefits of Sample Multiplexing:

- Reduce technical variation
- Reduce cost (compared to multiple capture lanes)
- Increase detection of doublets
- Increase number of cells in capture lanes ('super-loading')

Considerations of Sample Multiplexing:

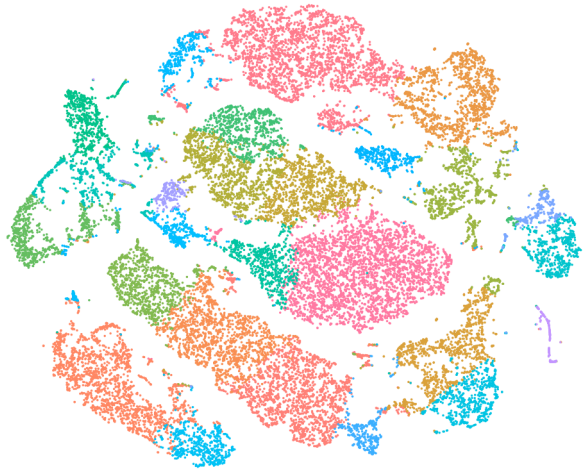
- What happens if you can't resolve barcodes?
- Extra sample handling & time
- Need more cells because cells are lost during post-staining washing steps
- Surface protein-based barcoded antibodies not compatible with single nuclei preps

Note: Both Seurat v3 and FlowJo-SeqGeq have methods for handling cell hashing features

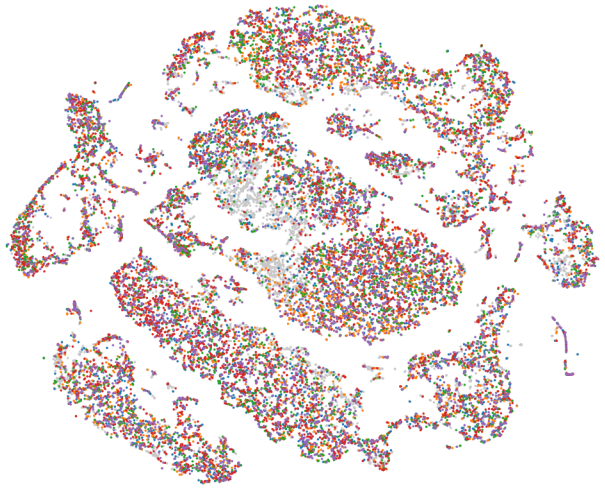
Feature barcoding for sample multiplexing with gene expression profiling

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

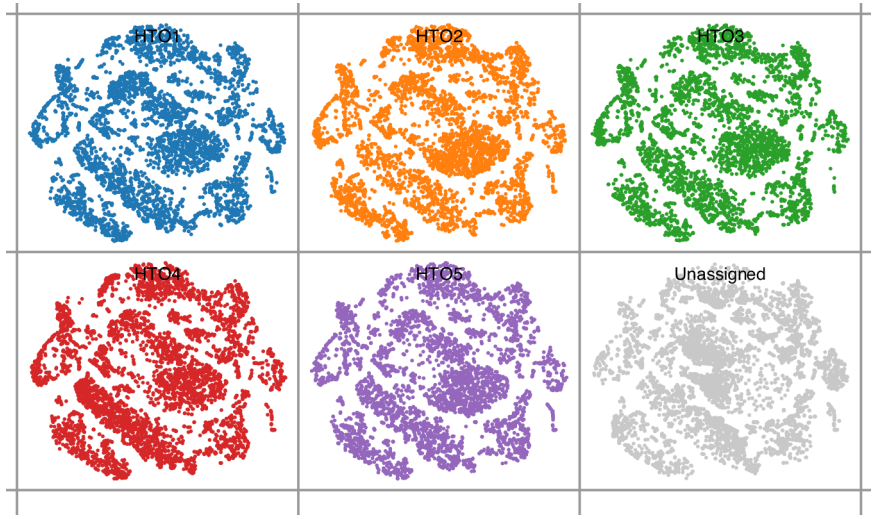
Gene Expression tSNE
(color = graph-based clusters)



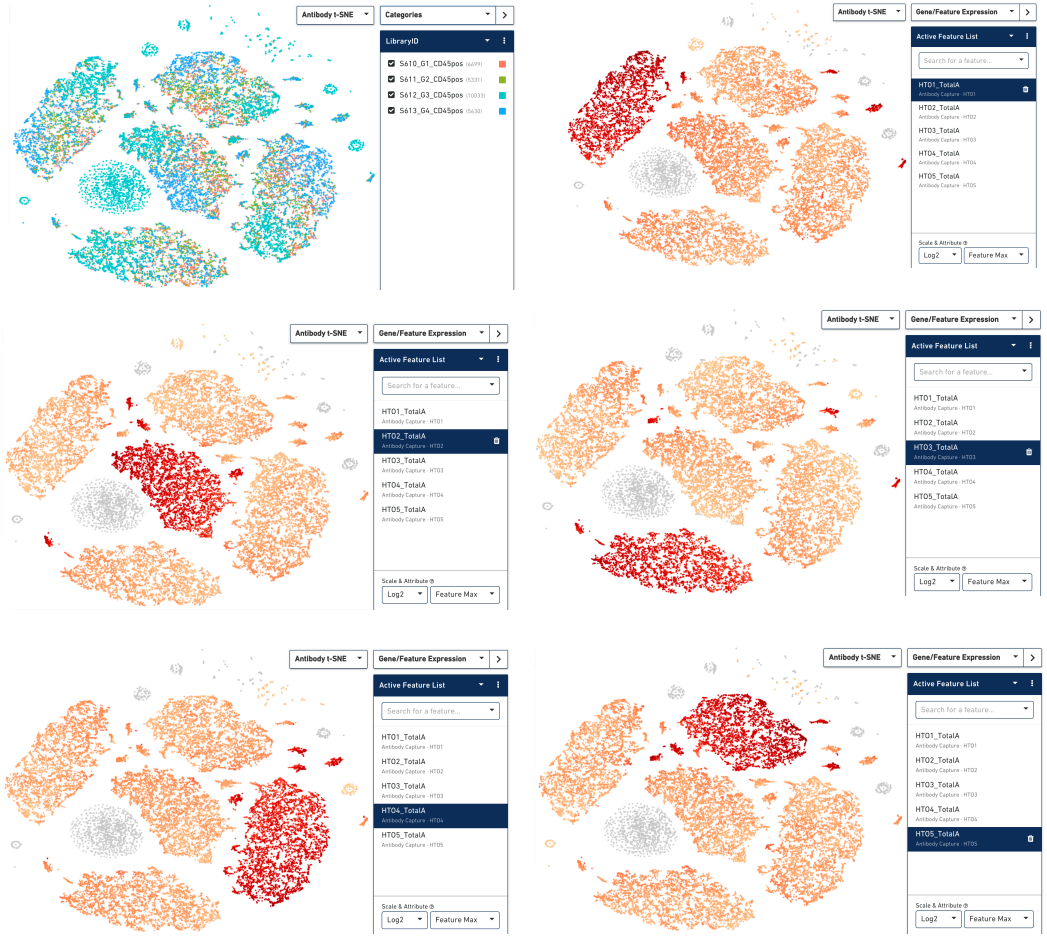
Gene Expression tSNE
(color = hashed biological replicate)



Gene Expression tSNE
(Split by Biological Replicate)



tSNE Projection from Antibody Labels Counts



Data generated by SCAF in support of Soumen Roy (Giorgio Trinchieri's Lab)

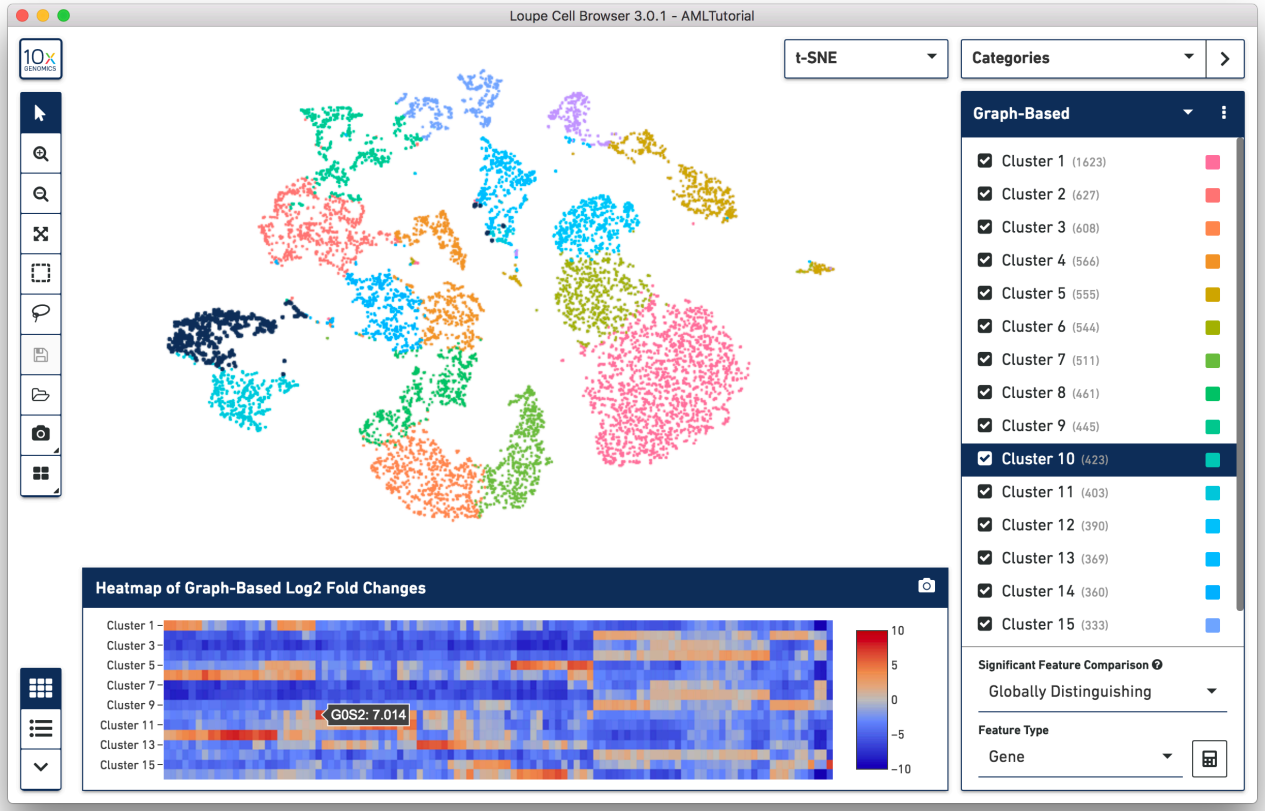
Accessible analysis tools

Commercial Software for scRNA-Seq Analysis

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

Loupe Cell Browser (free from 10x Genomics)

- Most straightforward, but still relatively powerful
- Interrogate data for specific gene expression, and run custom differential gene expression
- Only takes data from 10x Genomics platform
- Limited in re-analysis that can be done



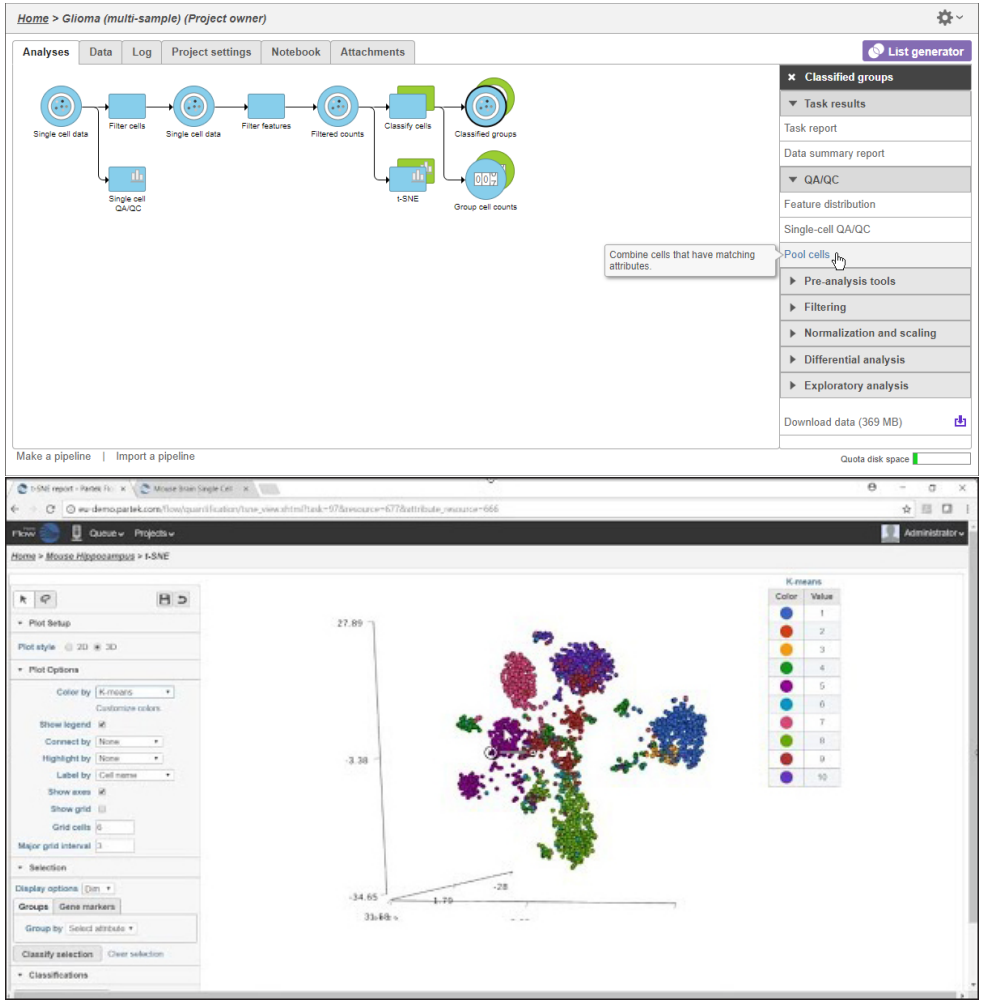
Commercial Software for scRNA-Seq Analysis

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

Partek Flow Single Cell

(14-day free trial available)

- **Relatively intuitive workflow. Control of various parameters**
- **Somewhat familiar for previous users of Partek Flow**
- **Some advanced secondary analysis currently limited**
- **Hosted on the NIH HPC systems**



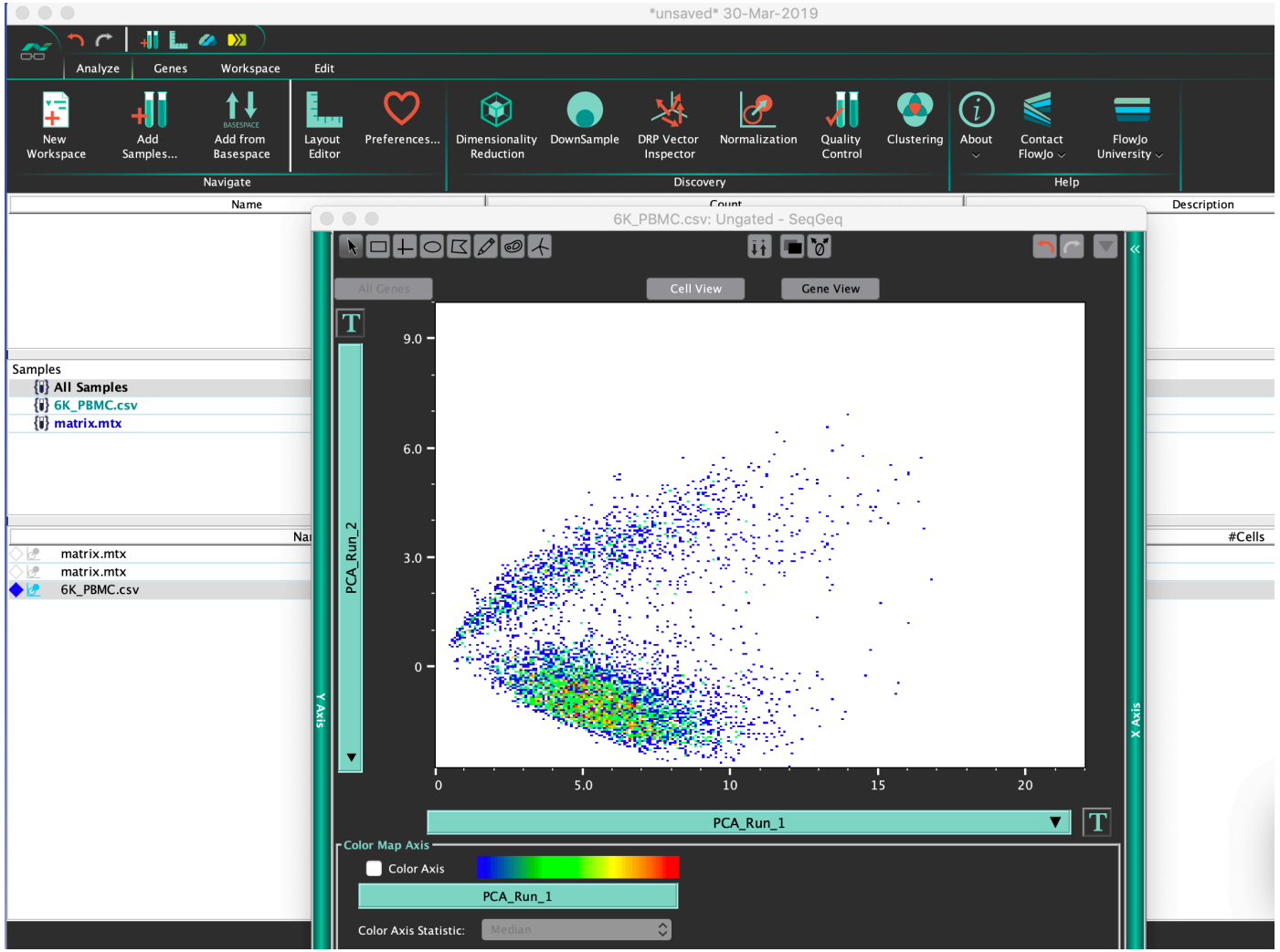
Commercial Software for scRNA-Seq Analysis

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

FlowJo SeqGeq

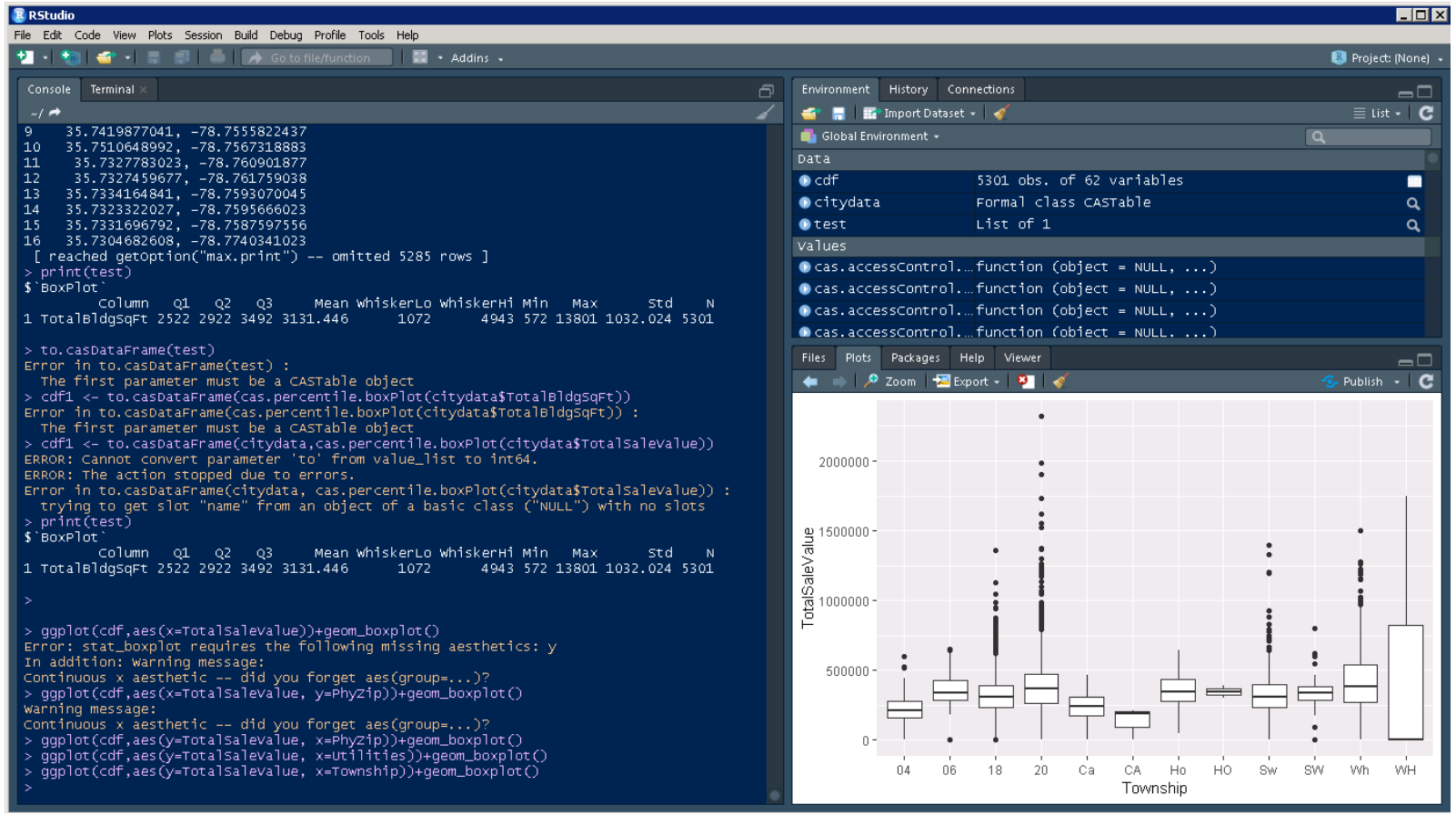
(60-day free trial available)

- Familiar to those who have used FlowJo
- Some ”gating” and workflows similar to FACs analysis
- Still need to learn order of steps (some reading in FlowJo University or a training)
- Some powerful add-in’s available – batch correction, trajectory analysis, various visualization tools
- Integration of VDJ & CITE-Seq



The advantages of R (Rstudio)

- Efficient analysis
- Full control of analysis parameters
- Decent documentation for dominant analysis packages (Seurat, Monocle, etc.)
- Most update to date innovative analysis tools
- Does have a learning curve; but you can use these tool far beyond just sc-Seq analysis
- Can easily run R on Biowulf, including with GUI using NoMachine – especially helpful for large datasets that require large memory usage



Some useful resources

Useful Resource Links

- **Hemberg Lab Online, Single Cell Self-Paced "Course"**
  - https://scrnaseq-course.cog.sanger.ac.uk/website/index.html
- **10X Genomics Demonstrated Protocols, Videos and Datasets**
  - https://support.10xgenomics.com/single-cell-gene-expression
- **Sean Davis's "Awesome Single Cell" Page – collection of analysis tools, database collections, papers and other resources**
  - https://github.com/seandavi/awesome-single-cell
- **NIH Single Cell Genomics Scientific Interest Group & User Group – seminars, invited speakers, discussion groups**
  - https://nih-irp-singlecell.github.io
- **Many NIH Institutes have core facilities that can support single cell sequencing**

Public scRNA-Seq Datasets

- **10x Genomics**
  - 3’ Gene Expression Datasets (including Surface Protein Feature Barcodes):

    https://support.10xgenomics.com/single-cell-gene-expression/datasets
    - Human, Mouse, and species-mixed
    - PBMCs, Tumor, T-cells, Brain (whole cell, nuclei, and methanol-fixed)

  - 5’ Gene Expression with TCR / BCR (and with Feature Barcodes):

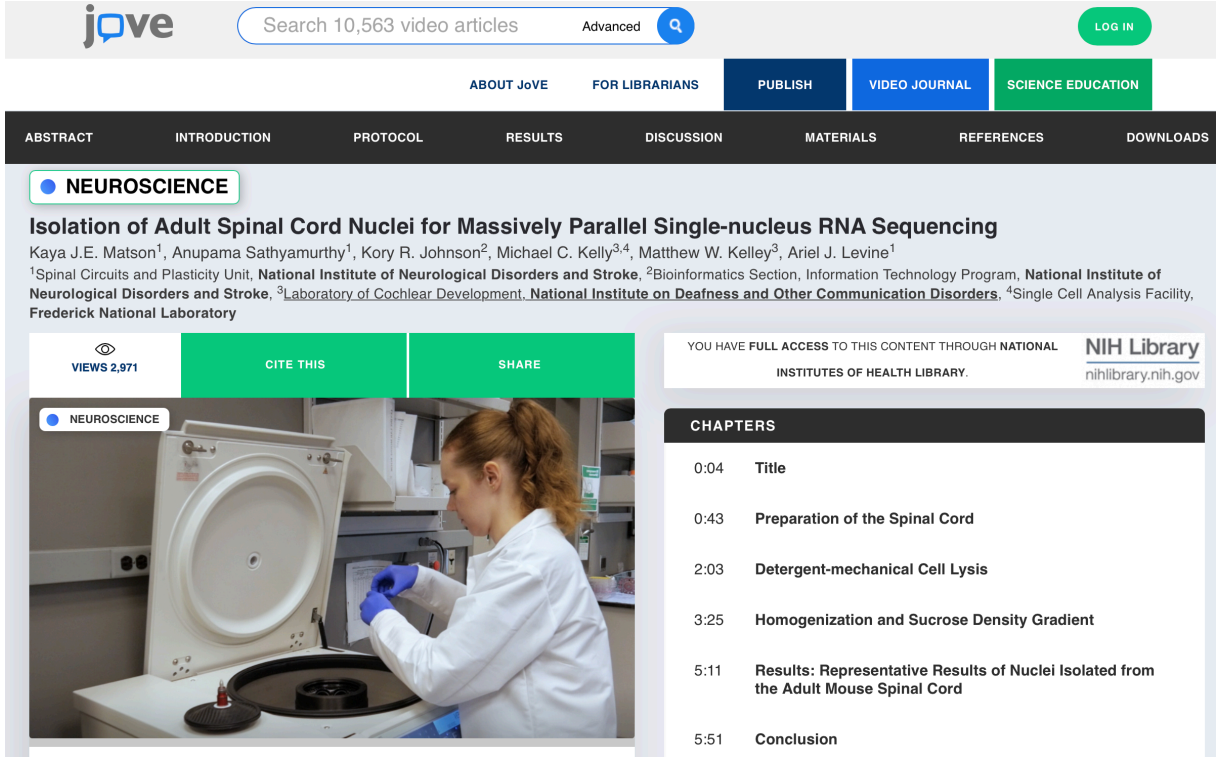
    https://support.10xgenomics.com/single-cell-vdj/datasets
    - Human and Mouse

- **BROAD Single Cell Portal**
  - https://portals.broadinstitute.org/single_cell

Getting started with single nuclei sequencing

- **10x Genomics Demonstrated Protocols**
  - For various tissue types
  - If running scATAC-Seq, important to start with their protocols and buffers

- **JoVE (from Levine Lab – NINDS)**
  - Visual walkthrough of method
  - Not all tissues may be the same



Recap

- Single cell sequencing is now accessible for many labs and research goals
- Important to choose the right sample preparation method for best data
- Multiple modalities can now be collected alongside RNA-Seq data
- Experimental design can be aided by sample multiplexing techniques
- Relatively user-friendly analysis tools can make complex data types accessible for individuals with varying degrees of comfort with bioinformatics

A Special Thanks to HPC / Biowulf Staff

Thank you for…

- **Providing fast & reliable computational resources**
- **Extensive package and software support, with timely and responsive updates**
- **Constantly improving resources, speeds, and connections**
- **Fostering a great environment of learning and support**
  - User guides, software support pages, and direct one-on-one support
  - Allows trial and error without cloud core-hour charges – great for learning and for setting up new methods
- **Being truly invested partners in cutting-edge science**

Remember to acknowledge use of HPC systems in your publications!

This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov)

Thank you!

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- **NCI-CCR SCAF Team**
  - Zach Rae
  - Maria Hernandez, PhD
  - Allison Ruchinskas
  - Kimia (Ezzat) Dadkhah, PhD
- **Frederick National Lab CRTP Genomics Groups & ABCS**
  - Sequencing Facility & SF-IFX
  - Genomic Technology Lab
- **Other CCR Cores**
  - Bld 37 Genomics Core
  - CCBR
  - Collaborative Protein Group
  - Bld 37 Flow Cytometry Core

Laboratory of Cochlear Development (NIDCD)

- Matt Kelley (PI), Joe Burns, Betsy Driver, Weise Chang, Joseph Mays

NIDCD Genomics and Computational Biology Core

- Rob Morell & Erich Boger

Collaborators

- Levine lab (NINDS), Friedman lab (NIDCD), Banfi lab (Univ of Iowa), Hertzano lab (UMBC)

NCI-CCR Leadership, Investigators, and Staff & Trainees

- Investigators that we've worked using these technologies

NIH Single Cell Community

- https://nih-irp-singlecell.github.io

Analysis Package Developers

Enjoy your long weekend of HPC downtime!

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

A reminder that an extended downtime of the Biowulf Cluster will start this **Thursday July 25 at 7 am** and last until **Monday July 29 at 10 pm**. A node reservation is in effect leading up to the maintenance window which prevents any HPC cluster jobs being scheduled that would overlap with the downtime period. Any running jobs will be terminated this Thursday at 7 am.

During this five day maintenance window all HPC services including the following will be unavailable to users:

- Biowulf login node & cluster
- Helix
- Helixdrive
- Mascot
- Sciware

In addition, the HPC Globus endpoint will be unavailable.

As the next step in the expansion of the NIH HPC cluster, this extended downtime will allow for the recabling of the HPCnetwork fabric to support additional compute nodes and storage as well as allow for the upgrading of the HPC storage systems.

Thank you for your patience as we continue to grow our computing resources.

##
Please contact staff@hpc.nih.gov with any questions about the NIH HPC Systems