



Introduction to Single Cell Genomics

Michael Kelly, Ph.D.

NIDCD

A Primer on Single Cell Genomics at CCR

Nov 14, 2017

Objectives

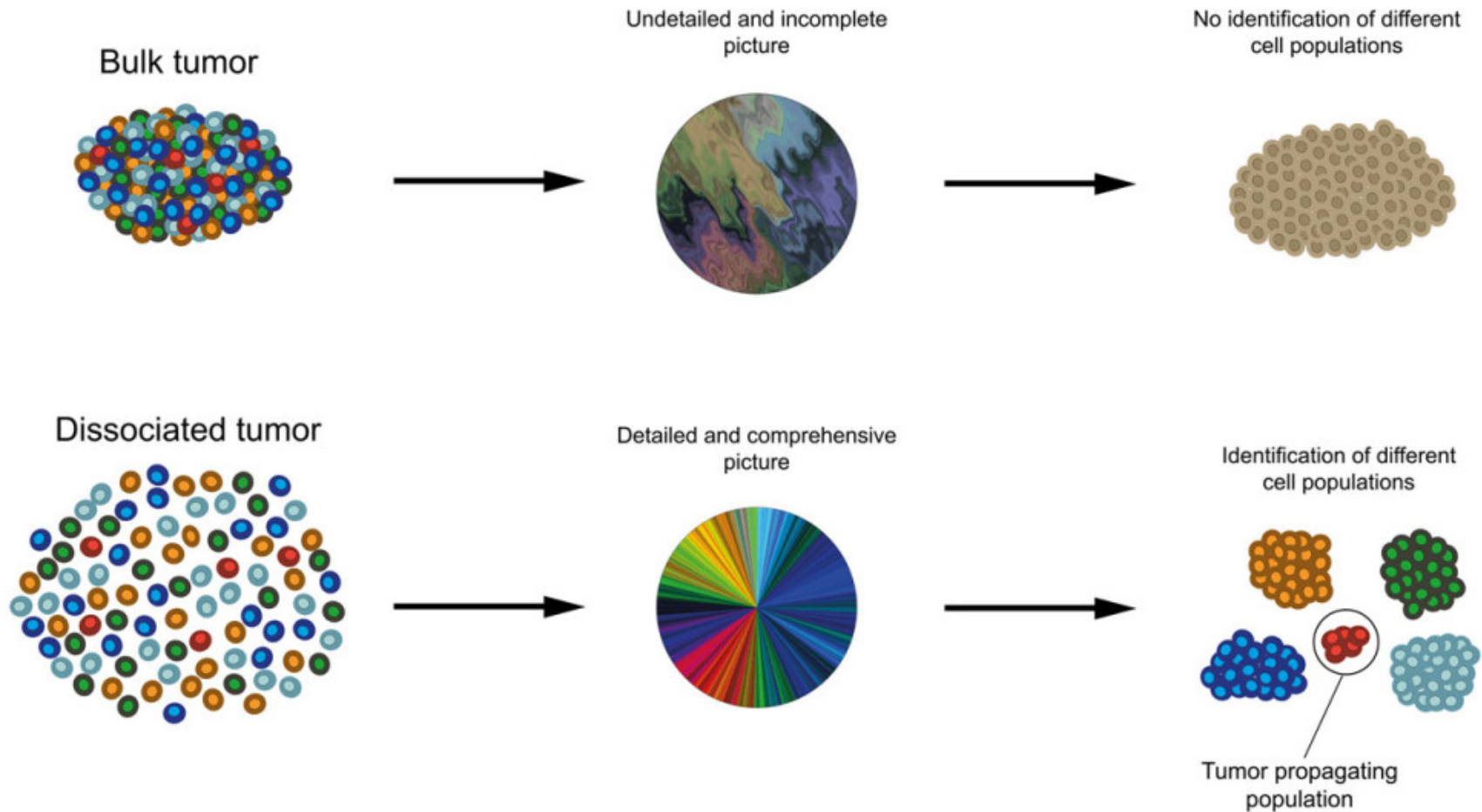
- Understand some of the key concepts in the methods and analysis of single cell genomics data
- Understand some of the current limitations
- Appreciate important experimental design considerations, including platform selection
- Be introduced to some of the “established” and emerging single cell genomics applications

Outline

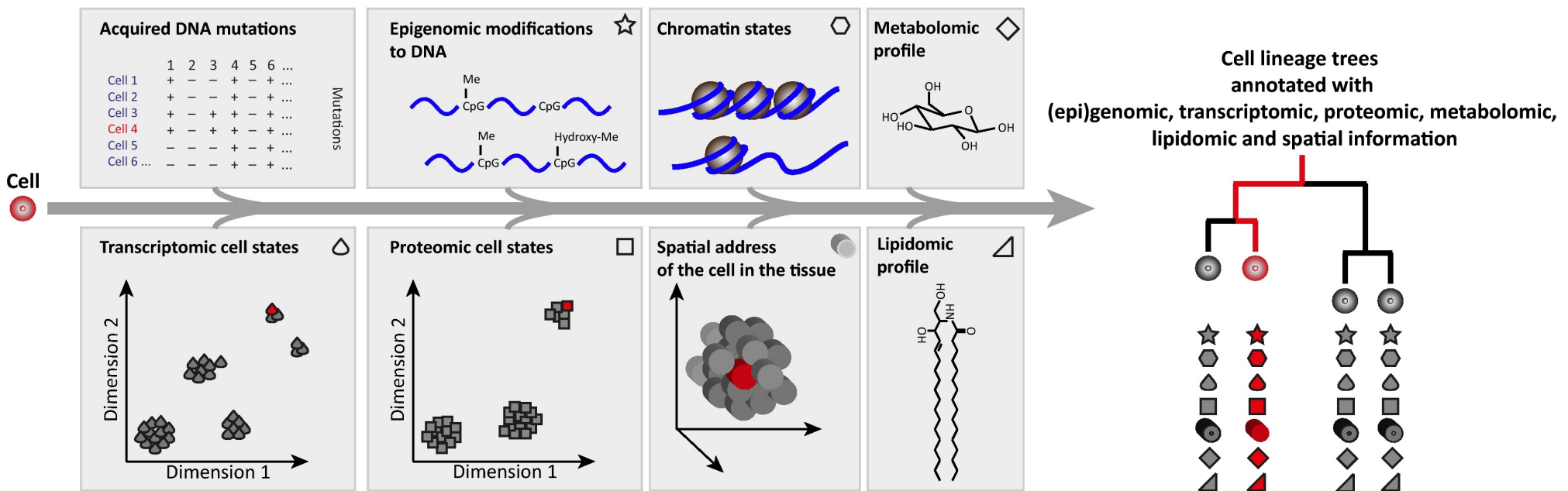
- Key Concepts in Single Cell Genomics
- Example Single Cell RNA-Seq Workflows
- Experimental Design & Platform Selection
- Single Cell Genomic Applications

Key Concepts in Single Cell Genomics

Single Cell Genomics – Avoiding the Caveat of Averaging



Single Cell "Genomics" Not Limited to Transcriptome



Trends in Genetics

Macaulay et al 2017 doi.org/10.1016/j.tig.2016.12.003

Genomics

Transcriptome
Genome
Epigenome

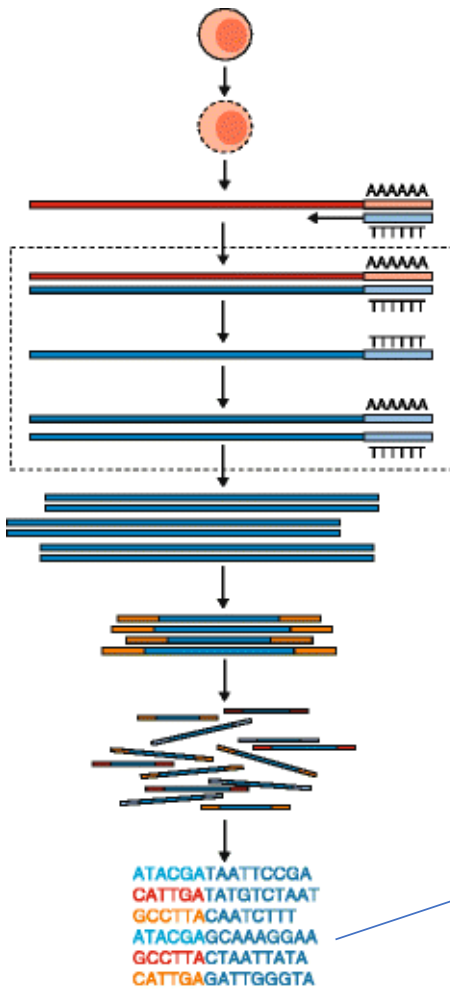
Other 'Omics

Metabolome
Proteome
Lipidome

Meta-Information

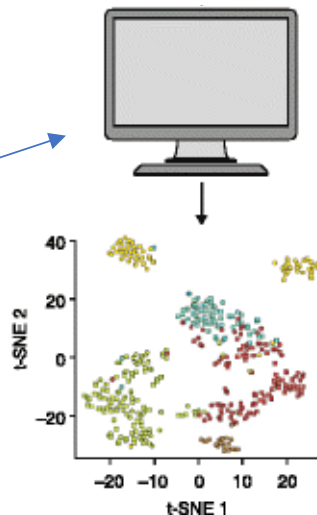
Lineage
History / Status
Spatial

Generalized Workflow for Single Cell RNA-Seq



- ① Isolate single cells from a tissue sample (including micro-dissection and manipulation, flow cytometric cell-sorting, microfluidic platforms, and droplet-based methods)
- ② Single cell lysis in a way that preserves cellular mRNA
- ③ mRNA molecule capture using poly(T) sequence primers that bind to mRNA poly(A) tails
- ④ Convert poly(T)-primed mRNA into cDNA using reverse transcription
- ⑤ cDNA amplification (usually by PCR or by *in vitro* transcription)
- ⑥ cDNA sequencing library preparation (insert 'index' nucleotide barcodes to identify each library)
- ⑦ Pool cDNA sequencing libraries
- ⑧ Sequence libraries (via Next Generation Sequencing)

1 - Partition Single Cells
 2 - Barcode & Sequence
 3 - Analyze & Interpret

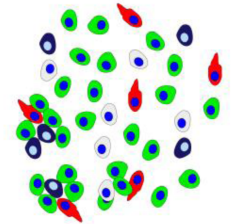


- ⑧ Use bioinformatic methods to perform quality control and to assess technical variability in the scRNA-seq data
- ⑨ Use bioinformatic and/or computational methods to interpret robust data biologically

What's different about Single Cell data?

Requires the partitioning of single cells

- Assigning information from one cell versus another is usually done via a barcoding strategy, which occurs when the cell is partitioned
- Isolating single cells is not trivial from some tissues – they can either be difficult to dissociate and/or fragile



Single cell data is "sparse"

- Low amount of starting material and less-than-ideal conditions for sensitivity
- Single cell RNA-Seq might give you 500 genes expressed in a single cell
- Analysis methods take some of this into consideration, and may differ from analysis of bulk datasets

Single cell data is "noisy"

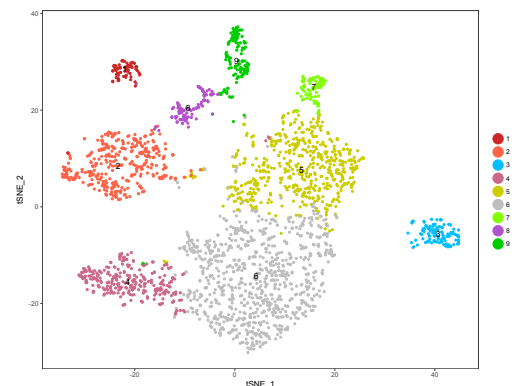
- Prone to technical noise and variation, making measurement of biological signal tricky
- Lots PCR and molecular biology wizardry at work – technical bias may arise
- Even when a molecule is present, it's detection is not guaranteed (low-abundance molecules are especially prone to these "drop-outs")
- Better to rely on correlated sets of genes rather than single genes for analysis

Datasets are flexible / usually require specialized analysis

- Standard control versus treatment type testing often only part of the analysis
- Differential expression may start with defining which samples to compare
 - May require identification of outlier samples, normalization, and clustering of data
 - Ability to select samples in each comparison groups makes data very flexible

Example scRNA-Seq Dataset

Cell # 1 ...	Dataset	20
Xkr4
Gm1992
Gm37381
Rp1
Rp1.1
Sox17 1
Gm37323
Mrp115 1 2
Lyp1a1	1 2 2
Gm37988 1



Example of Single Cell RNA-Seq Workflows

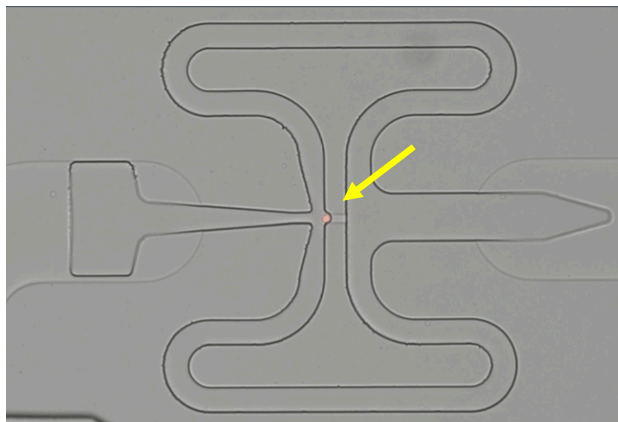
- 1 - Partition Single Cells
- 2 - Barcode & Sequence
- 3 - Analyze & Interpret

Cell isolation and handling

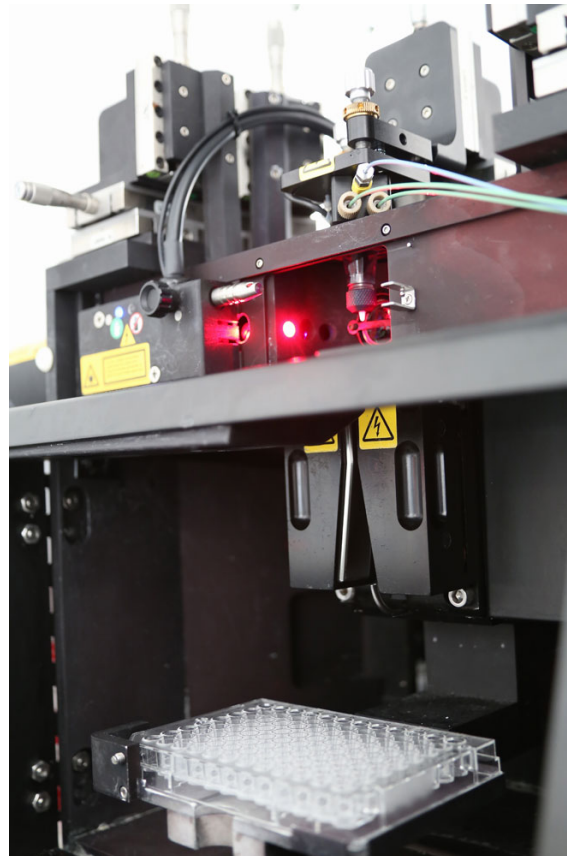
- Ideally want to measure the native biological state
 - Minimize transcriptional drift, degradation, etc. during isolation process
 - Preserve viability and diversity of cell types
- Do you need to test and optimize dissociation processes?
 - Selection of enzymes, incubation times, etc.
 - Effects on cell viability?
- Do you need to enrich for target cell types?
 - FACS, MACs, other?
 - What effect will this additional processing have on what you are looking to assay?
- What if fresh samples cannot be obtained, or the tissue cannot be efficiently dissociated?
 - Some preservation / fixation methods have been demonstrated
 - Isolating nuclei instead of whole cells may be an option

Partitioning Single Cells – Some Common Examples

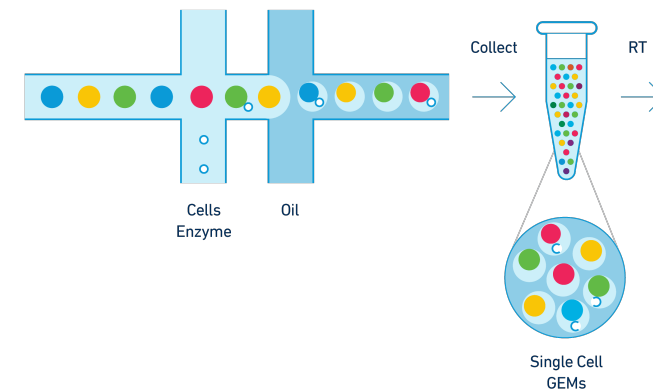
Microfluidic
Capture



FACs

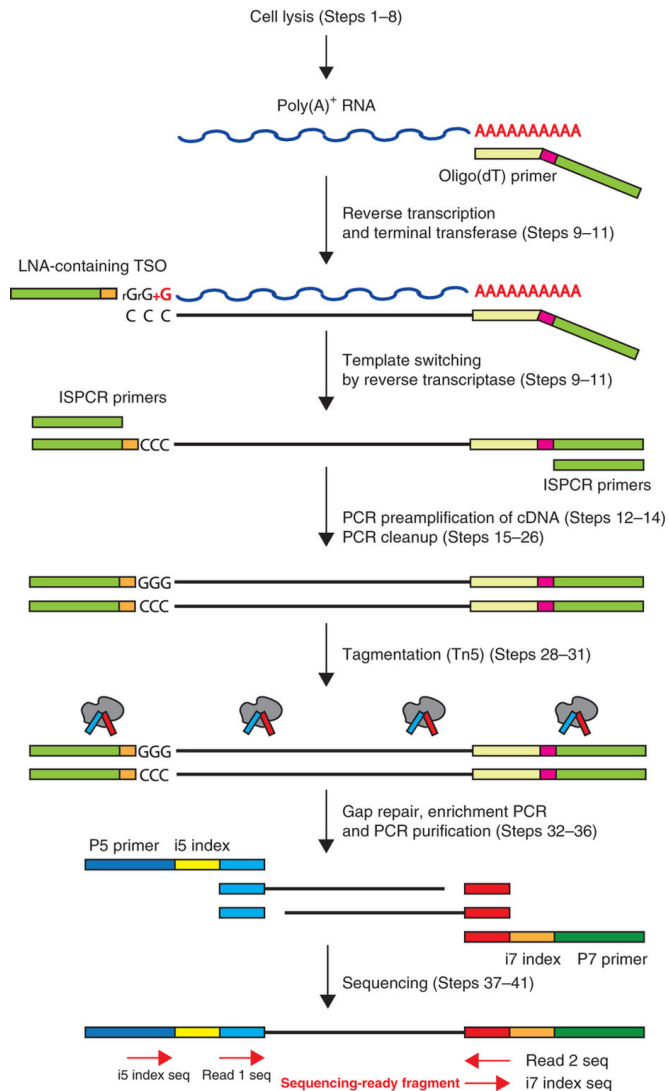


Droplet Based
Methods



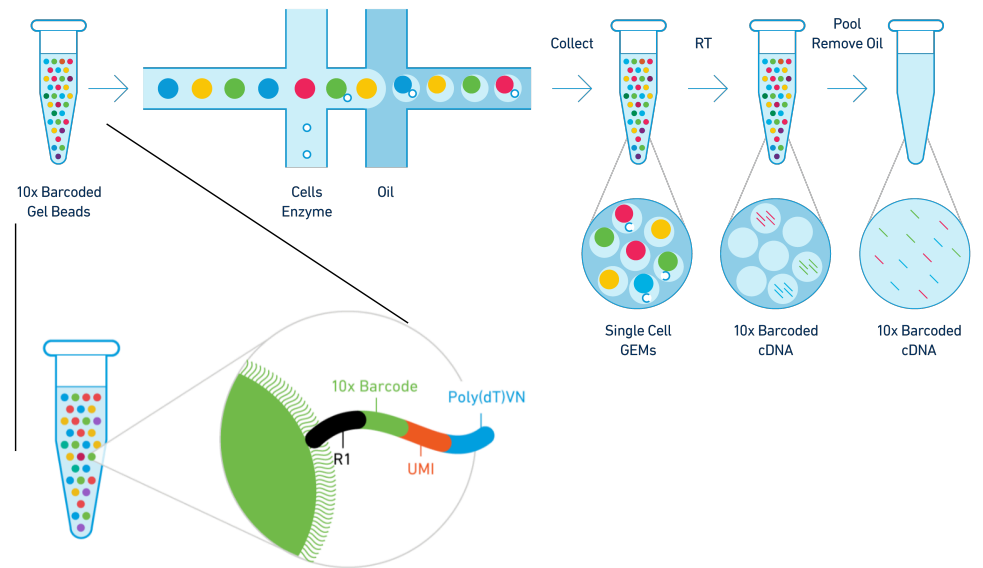
Capturing mRNA and Adding Barcodes

Single Cell Per Well Protocols



Picelli et al 2014

Droplet-Based Protocols

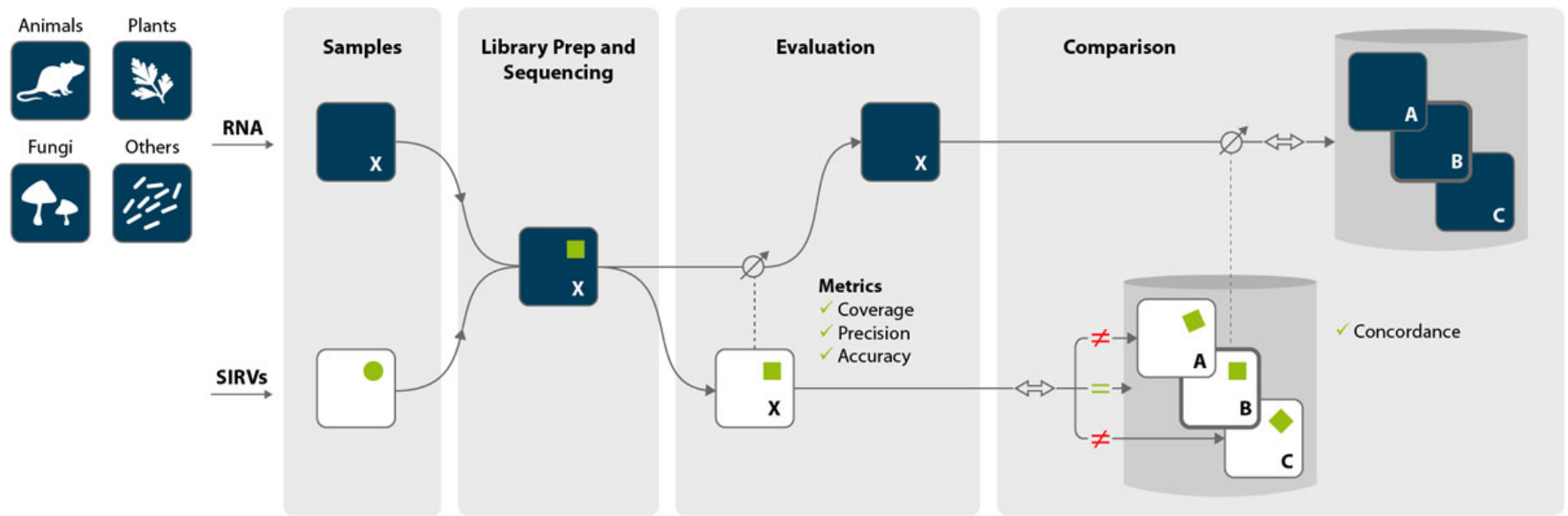


Transcriptional profiling of individual cells



From 10X Genomics Promotional Material

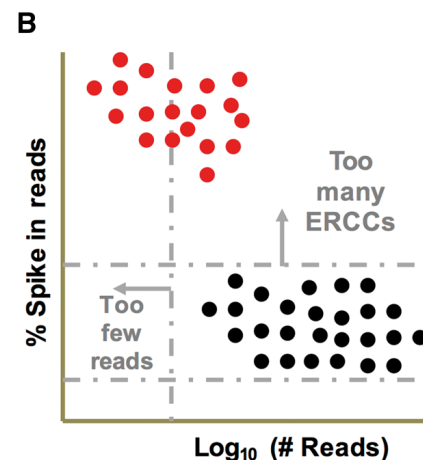
A quick intro to spike-in's and UMI's: Exogenous spike-in's provide a known reference concentration for comparison / adjustment



<https://www.lexogen.com/sirvs/>

ERCCs: up to 96 synthetic RNA molecules in known molar ratios and lengths

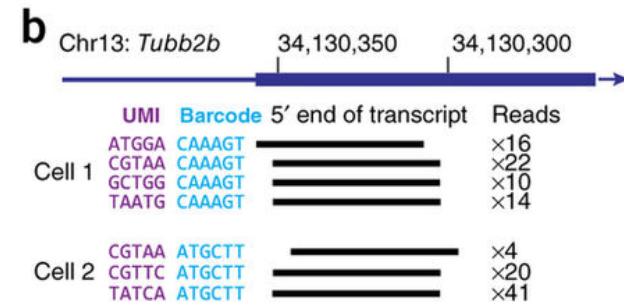
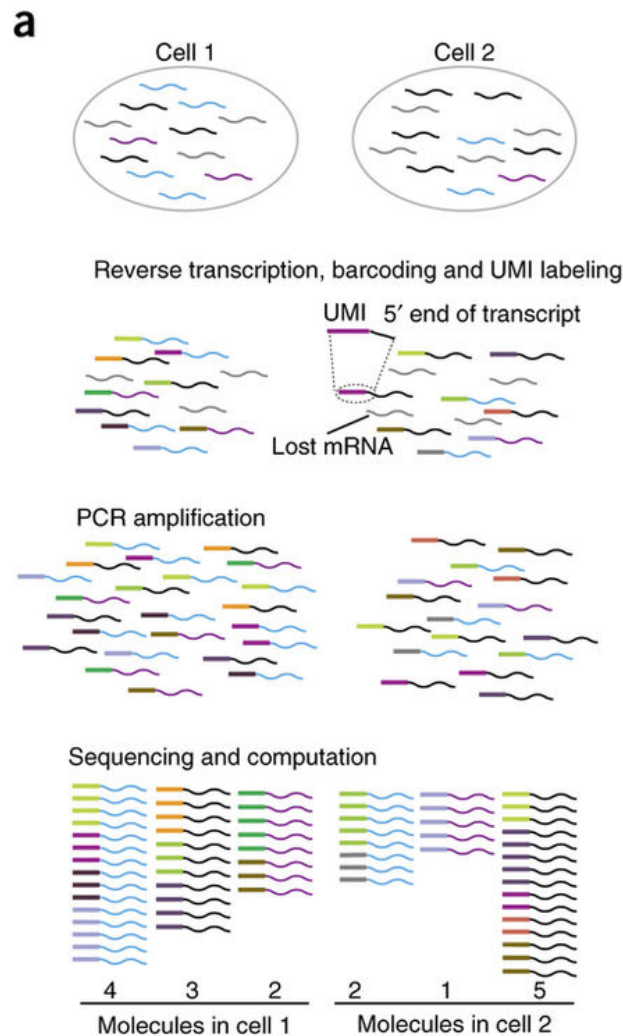
SIRVs: synthetic RNA molecules that model alternatively splicing and other variation. Vary in length and complexity.



Baran-Gale et al 2017

- Spike-in concentration needs to be tuned to RNA content of your samples
 - Balance detection with sequencing cost
- The utility of spike in's may be limited for droplet based methods

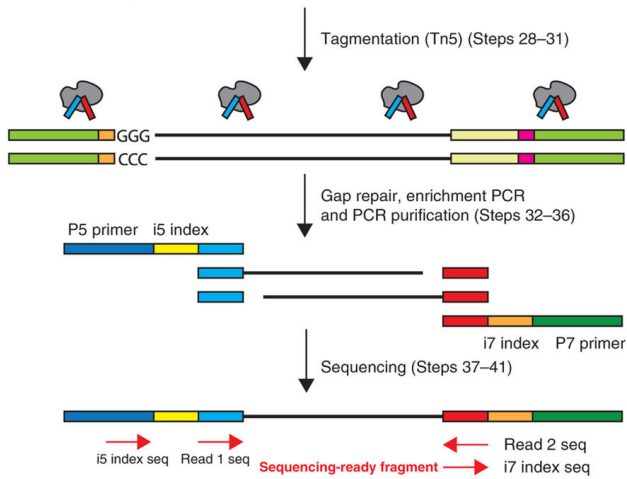
A quick intro to spike-in's and UMI's: Molecular indices allow tracking of how many original molecules existed



- Multiple reads for the same gene within the same cell can be collapsed to a count of one if they have the same UMI barcode
- Increased diversity of tags when cell barcode and target identity included.
- Unique molecular identifiers are currently only possible with 5' or 3' end methods

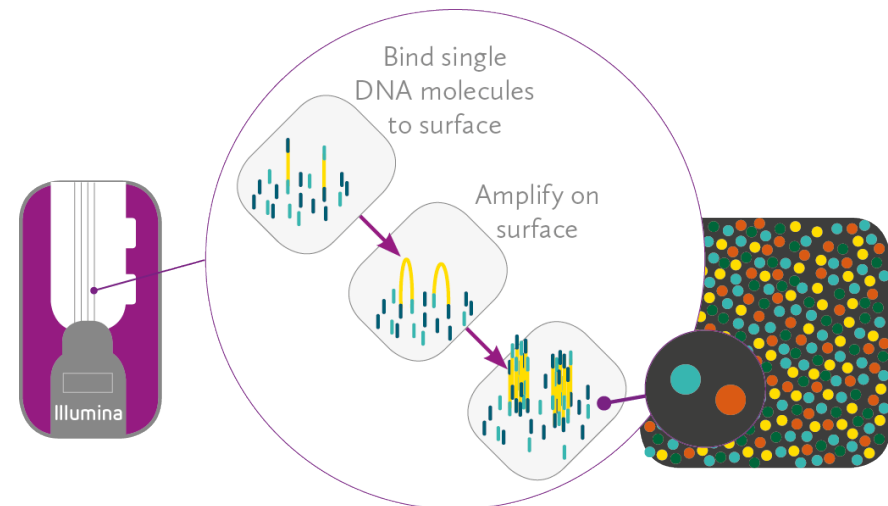
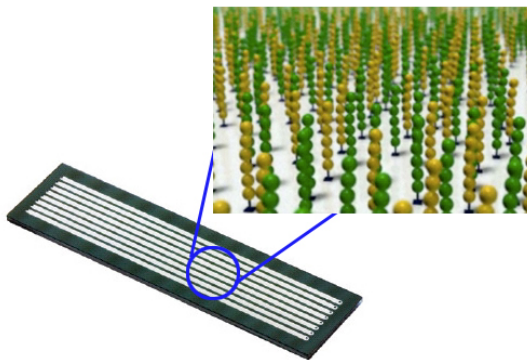
Library Prep & Sequencing

Full-length Libraries




3' End Libraries

Final Library Structure:



Data Analysis

Part I: Processing & Alignment

- 
- Demultiplex individual samples using cell barcodes
 - Single cell-per-well protocols generally use Illumina indices
 - Droplet-based systems use custom scripts to extract cell barcodes
 - Trimming and alignment
 - Removal of adapter sequences and low-quality information
 - Alignment of reads to reference genome with transcript coordinates
 - Full-length libraries can handle some multi-mapping; 5 or 3' end libraries usually on utilize non-ambiguously mapped reads
 - Assessment of alignment metrics
 - Percentage of reads mapped
 - Percentage exon vs intron vs intergenic
 - For full-length: gene body coverage and detection of splice sites

Input:


Raw sequencing files

Output:

Gene expression matrix

Data Analysis

Part II: Dimensionality reduction, clustering and differential expression testing

- 
- Initial QC and filtering
 - Outlier identification
 - Thresholding based on read depth, UMI counts, and/or genes detected
 - Cross-sample normalization
 - Adjustment for library size, etc.
 - Variance thresholding and stabilization
 - Selection of variable genes (non-“housekeepers”)
 - Dispersion (variance over mean) threshold often used
 - Data transformed to reduce statistical weight of huge expression values (e.g. log-transformation)
 - Dimensionality reduction
 - Principle component analysis (or similar) to look for structure in data
 - Define relationships between individual samples
 - Clustering (hierarchical, k-means, graph-based)
 - Trajectory modeling
 - Differential expression testing

- 1 - Partition Single Cells
- 2 - Barcode & Sequence
- 3 - Analyze & Interpret

Input:
Gene expression matrix
Output:
Analyzed data

Glossary of terms

Cell Barcode: sequence tag associated with all molecules from a single cell sample that allows tracking of individual transcriptomes

Unique Molecular Index (UMI): A unique sequence tag for every transcript molecule

Sensitivity: Ability to detect specific molecules, if present. Usually reported as number of UMI counts (transcripts) or genes detected.

Reads vs Counts: Reads are reported by the sequencer. Counts are the enumeration of observed molecules, which can be estimates based on transcript models or transcript counts with UMIs. Multiple reads of a gene with the same UMI can be a single count.

Multiplexing: Combining samples together for more efficient handling and analysis. De-multiplexed via cell barcodes.

Spike-in's: Exogenous synthetic molecules of known composition and concentration added to the initial reaction to compare to molecules from the cell. Allows for determination of sensitivity and a conversion of relative data to more absolute values.

Full-length vs 3' Only: For single cell RNA-Seq, referring to whether full length transcript information is assayed, or only the 3' end of the molecule – giving gene-only level information.

Dimensionality Reduction: Decreasing the complexity of the dataset by evaluating correlated structure between genes and grouping as a “meta-genes” to help interpret highly-multidimensional data.

Experimental Design & Platform Selection Considerations

<https://btep.ccr.cancer.gov/november-2017-single-cell-rna-seq-mind-read-starting-adventure/>

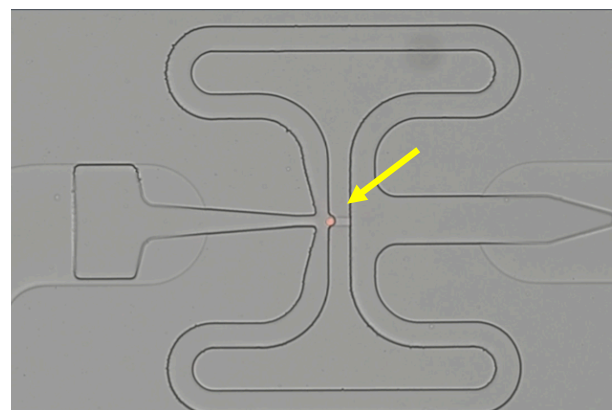
Initial Experimental Design Questions

- Why you need single cell resolution?
 - Single cell has technical limitations and extra cost
 - Assaying a heterogeneous population, a dynamic process, or surveying a tissue or system with diverse cell types?
- What do you expect to get from your data?
 - Knowing what analysis and comparisons you want to make will help make sure you include the right samples and controls
- Who will analyze the data?
 - A strong partnership between the biological subject matter expert and someone with bioinformatic expertise will increase the chances of project success
 - Bioinformatic consultation at project outset often helps in improving design

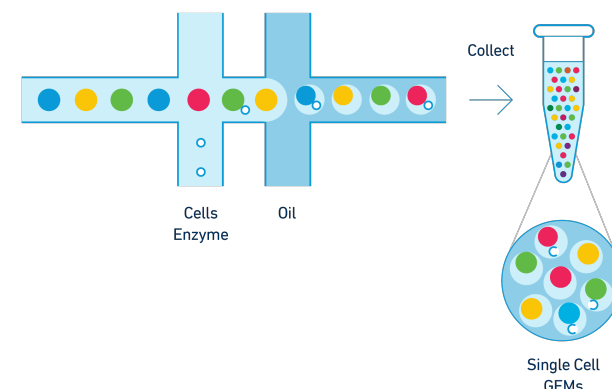
Experimental Design <-> Platform

Common Considerations:

- Cost per cell
- Throughput
- Efficiency of capture
- Full-length or 3'-Only Protocol
- Sensitivity
- Linking to other modality
- Multiple conditions in parallel?



*Fluidigm C1:
Higher cost, low-throughput, full-length, with ability to image*



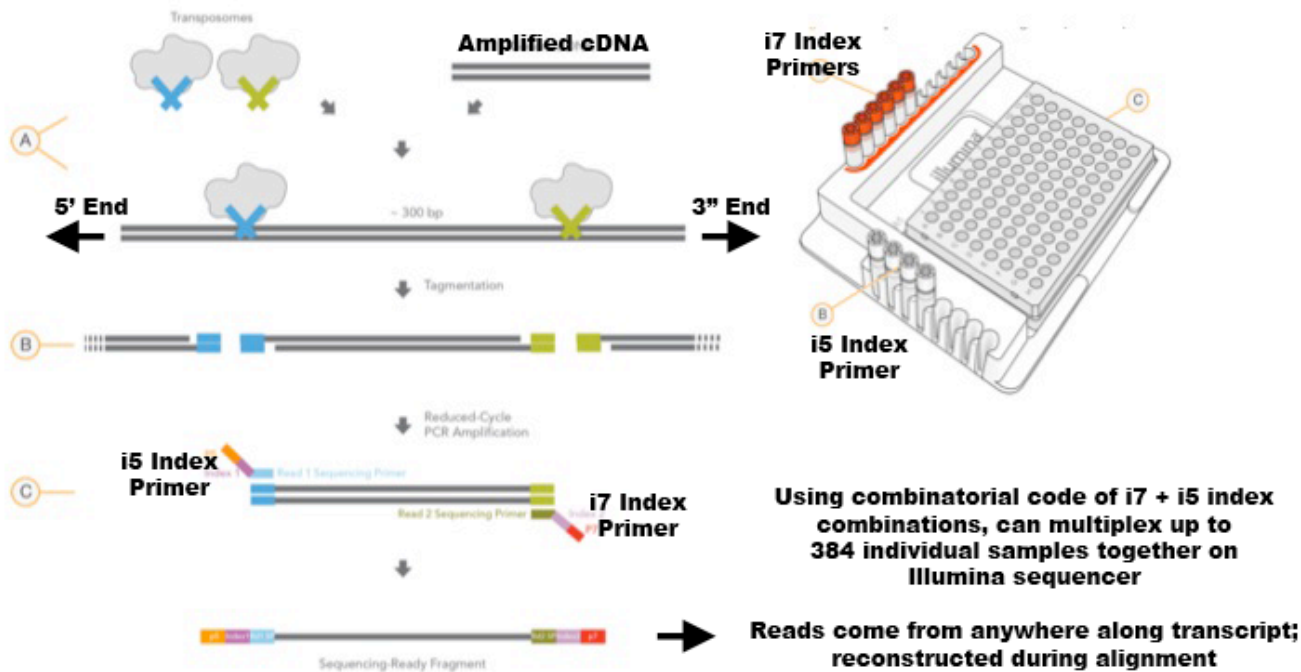
*10X Chromium:
Lower cost, high-throughput, 3'-end, up to 8 samples in parallel*

Method	\$ system	\$ per cells	No. cells	Doublets	Transcript type	UMIs	Capture Efficiency
DROP-seq	\$50000	\$0.65	up to 50000	0.36-11.3%	3' mRNA	Yes	~2%
Fluidigm C1	\$150,000	\$1.5-10	96, 800 (10k?)	10-23%	mRNA	No	~10%
10X Genomics	\$125,000	\$0.20-1.00	1000-6000	1-5%	3' mRNA	Yes	65%
Wafergen	\$200,000	\$1.5-2.5	~1800	1-5%?	3' mRNA	Yes	?

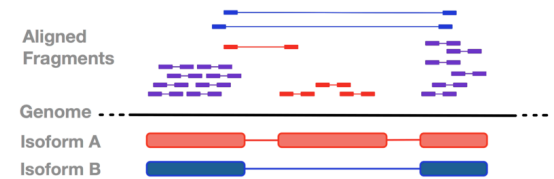
Modified from core-genomics.blogspot.com

Single cell-per-well methods allow full-length scRNA-seq on Illumina NGS sequencing platforms

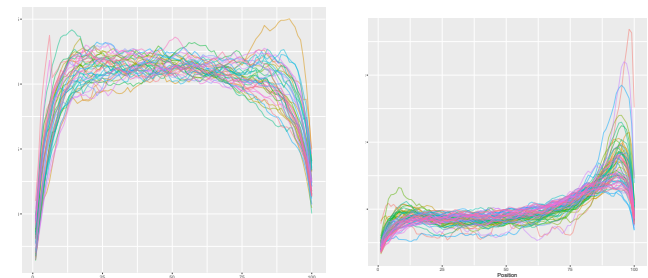
Nextera "tagmentation" library prep



Sample 1: i7=N708; i5=S510
Sample 2: i7=N712; i5=S511
Sample 3: i7=N708; i5=S512



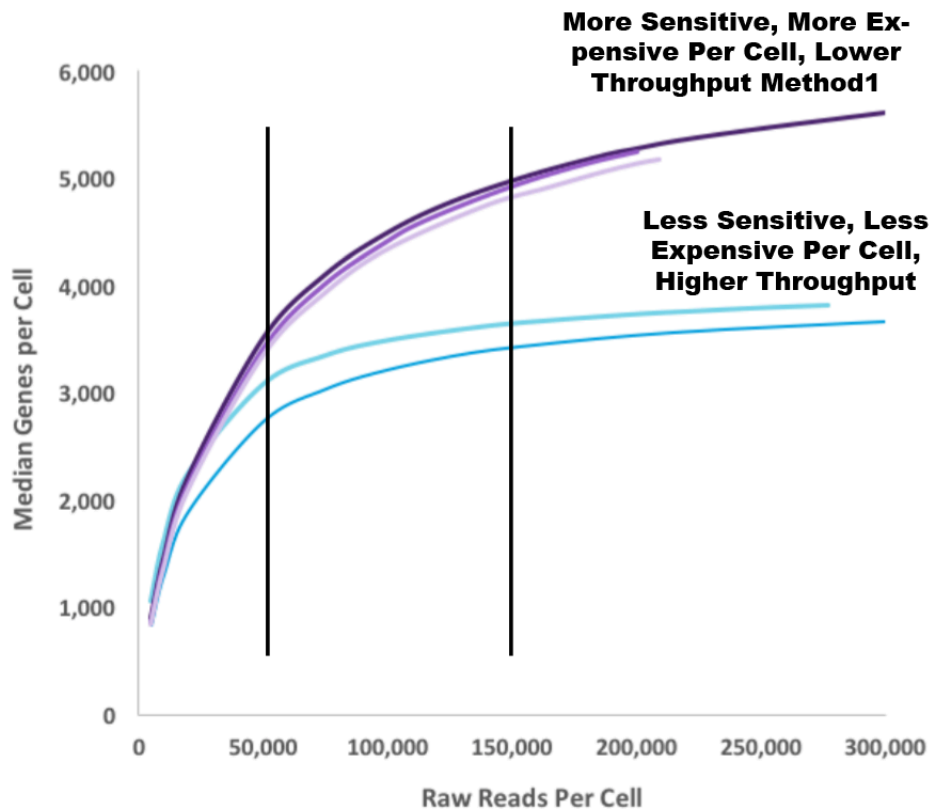
5-3' Transcript Coverage



Good

Bad

Which is better – more cells or greater depth?

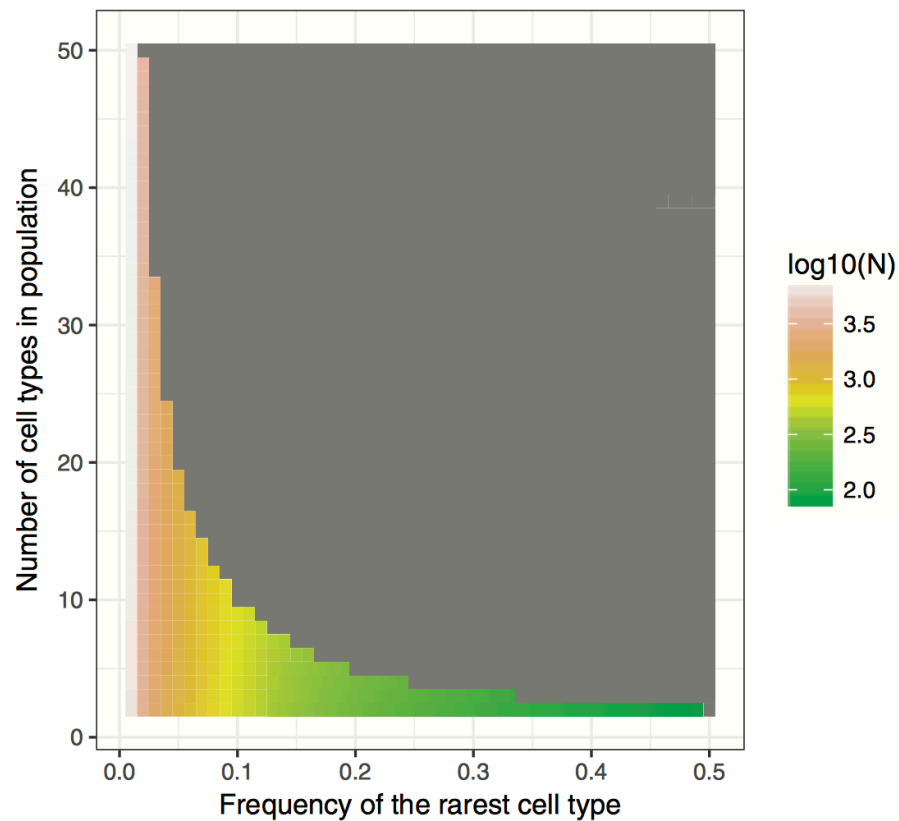


Modified from 10X Genomics material

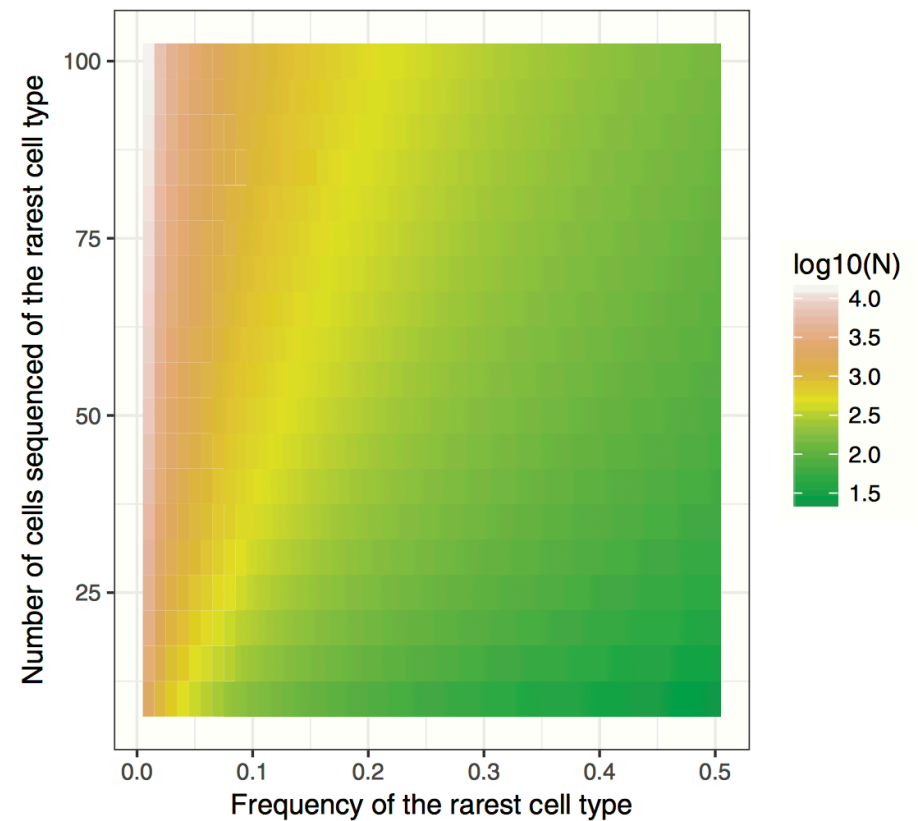
- More information can be gained by sequencing to greater depth – especially using sensitive methods
 - More genes detected; fewer “drop-outs”
 - Better isoform discrimination (when full-length libraries sequenced)
- More independent observation (more cells) is better for cell identity classification – averages out noise
- Classic scientific non-answer: it depends on what you are looking for
 - Broad survey of cell types or dynamics processes best modeled by higher-throughput data
 - Investigation of presumably low-expressed (or specific isoforms) requires greater depth

How Many Cell for Rare Cell Populations?

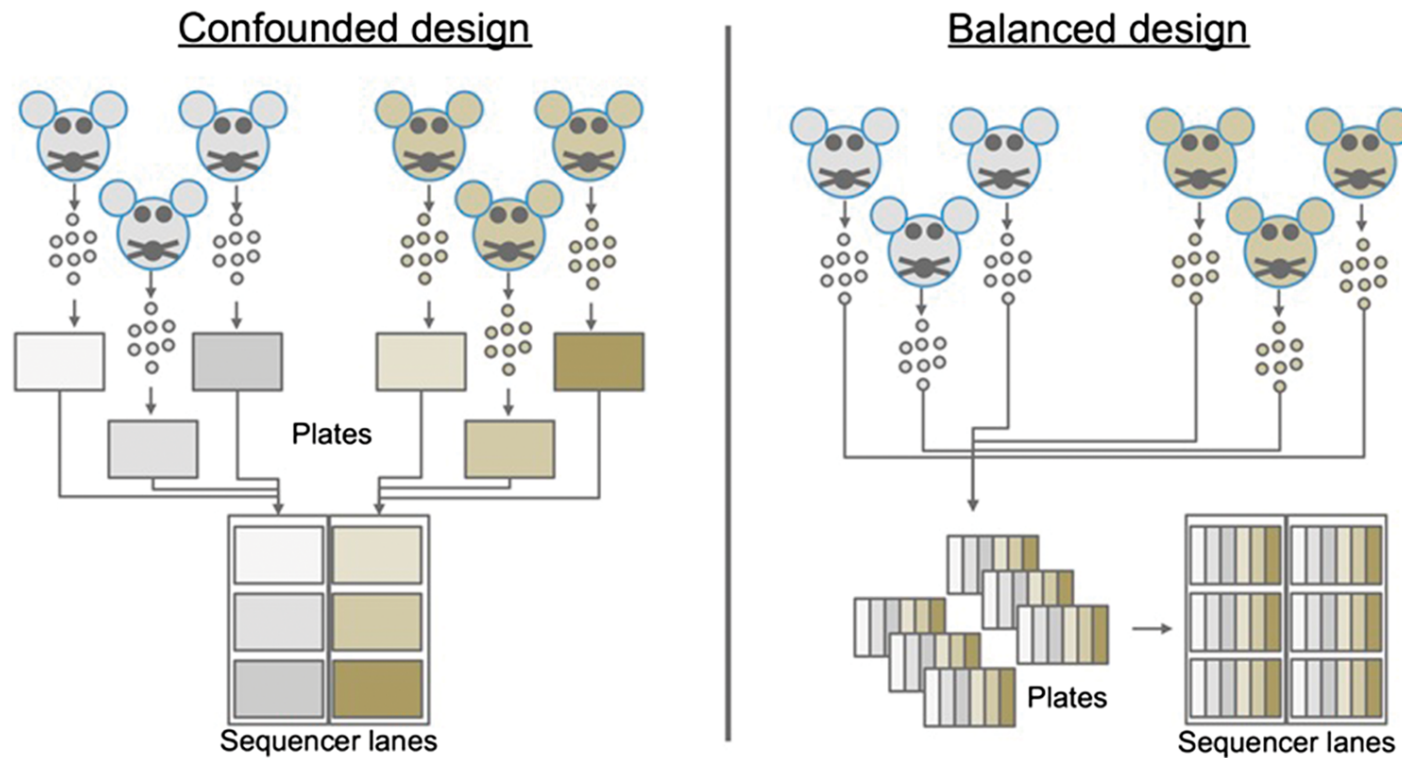
A Number of cells required to have 90% chance to sequence at least 50 cells of each type



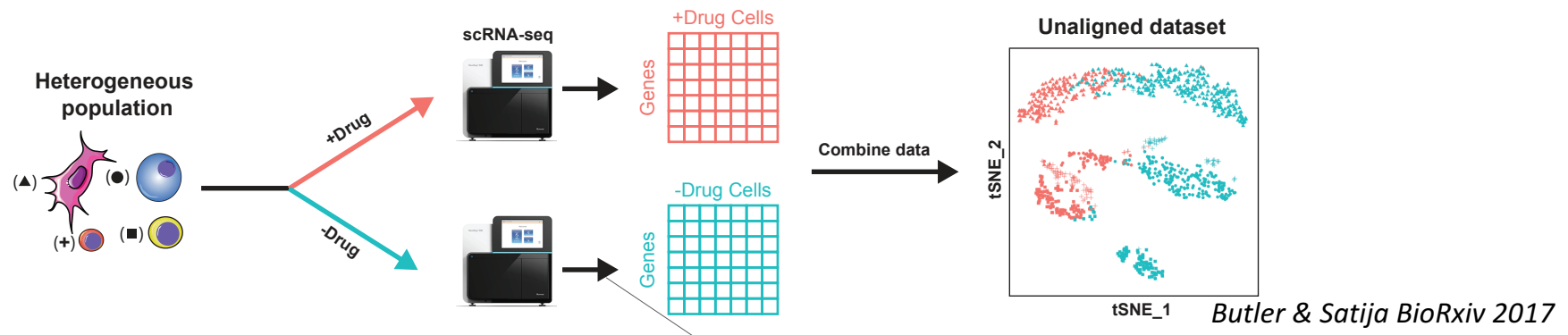
B Number of cells required to have 90% chance to sequence at least Y cells of each type



Control for Batch Effects in Design

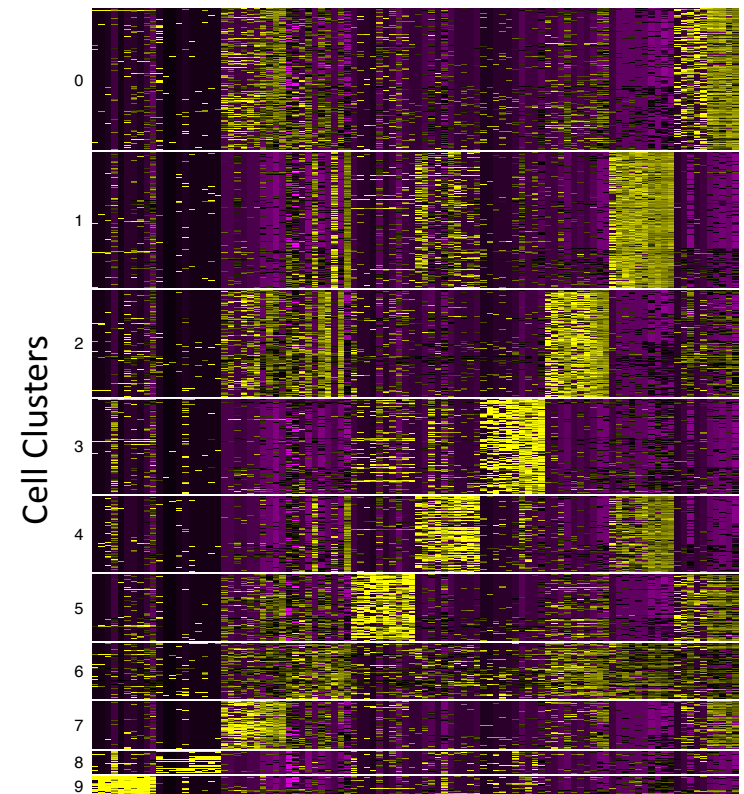
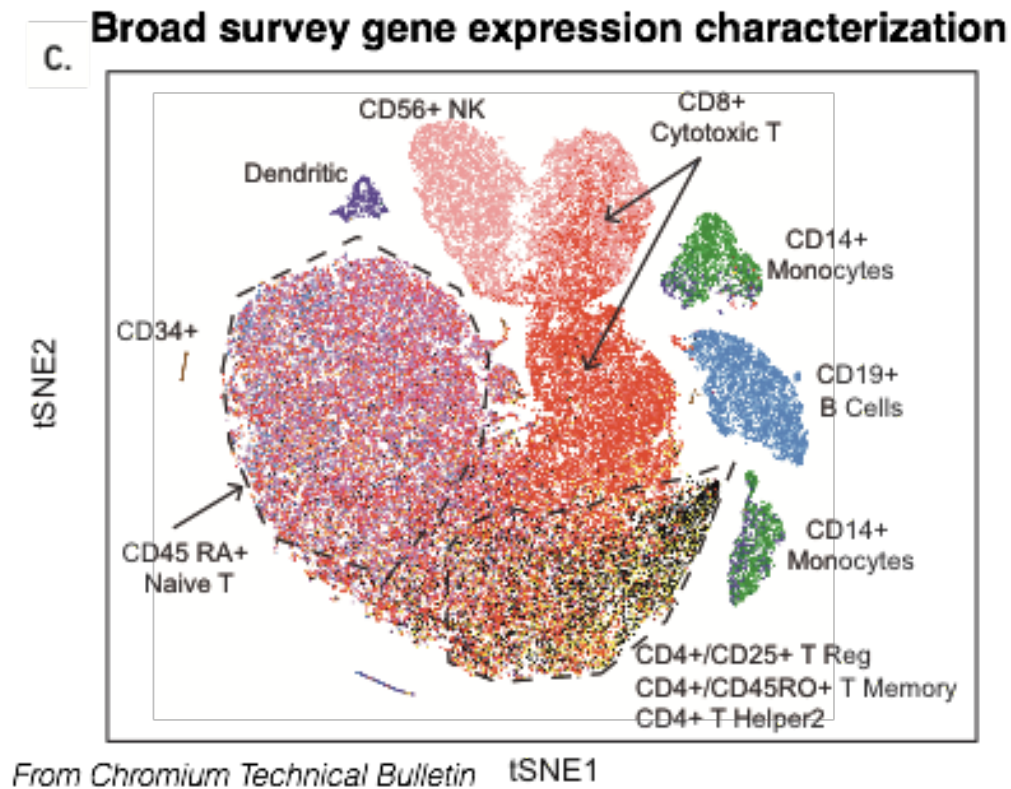


Baran-Gale et al 2017 (doi.org/10.1093/bfpg/elx035)



Single Cell Genomic Applications

Example of scRNA-Seq Analysis: Unbiased Identification of New Cell Types and Markers



Marker Genes for Each Cluster

Example of scRNA-Seq Analysis: Unbiased survey of cell ratio and transcriptional phenotypes changes

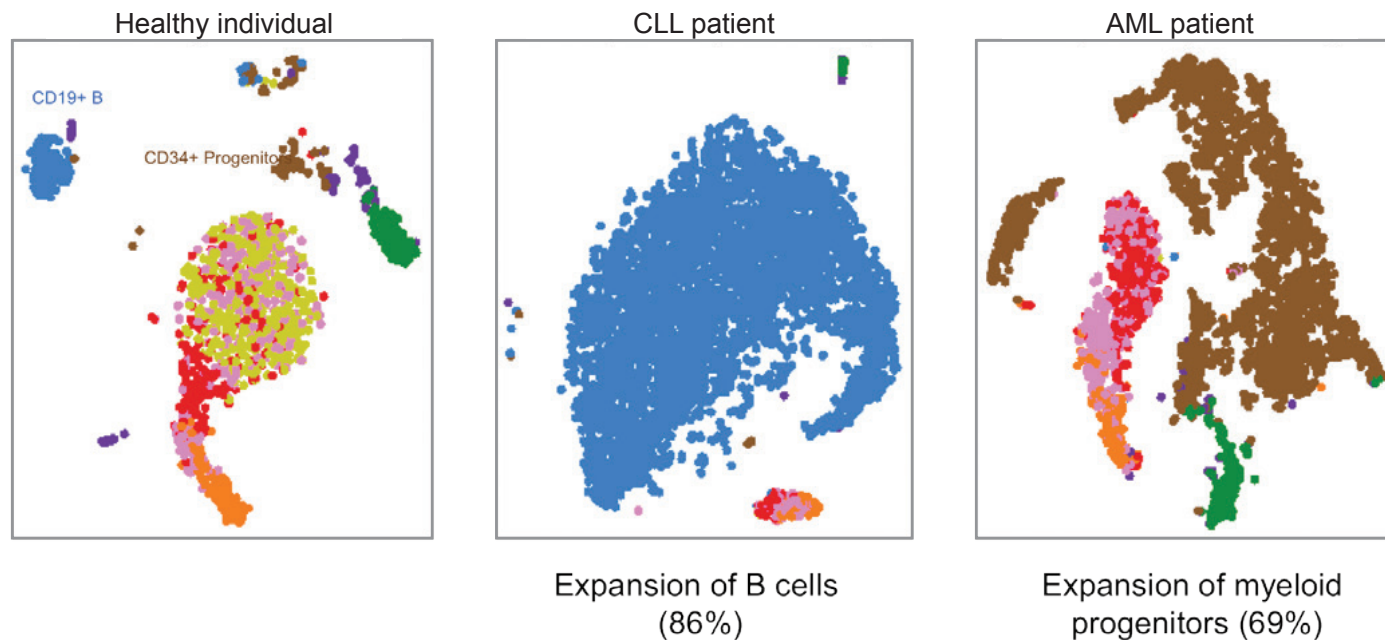
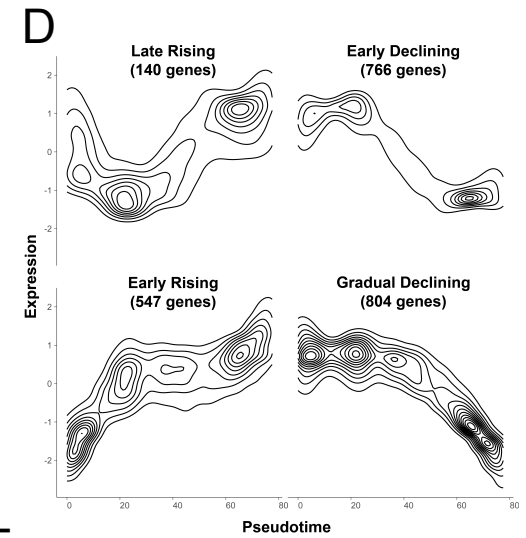
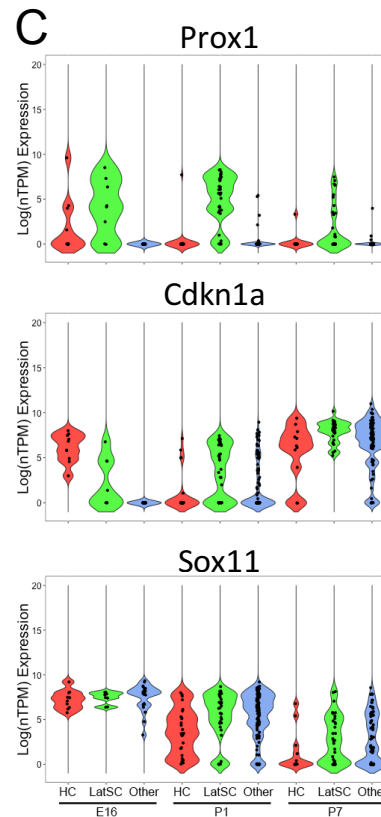
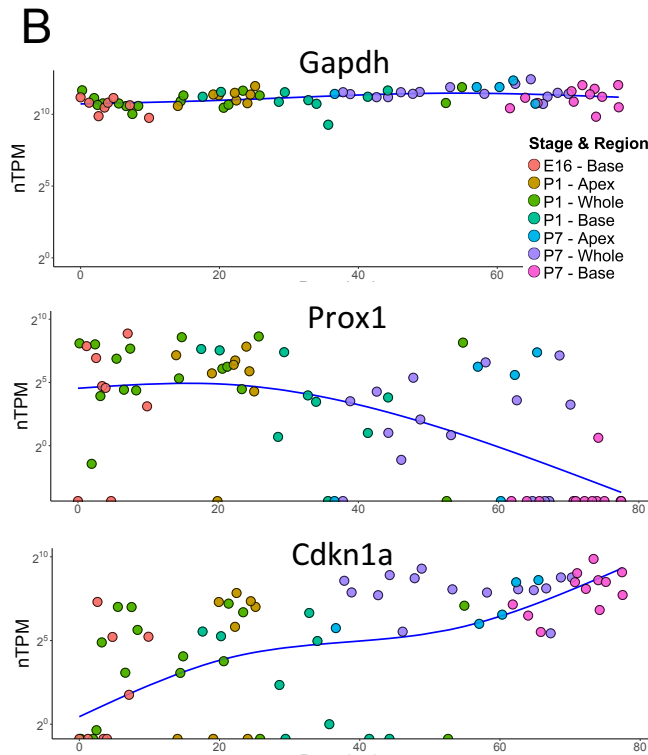


Figure 5. Single cell profiling from healthy and malignant tumor cell samples. Single cell profiling of BMMCs from healthy, CLL and AML patients. ~30,000 reads/cell in this experiment.

Example of scRNA-Seq Analysis: Developmental Trajectory Analysis

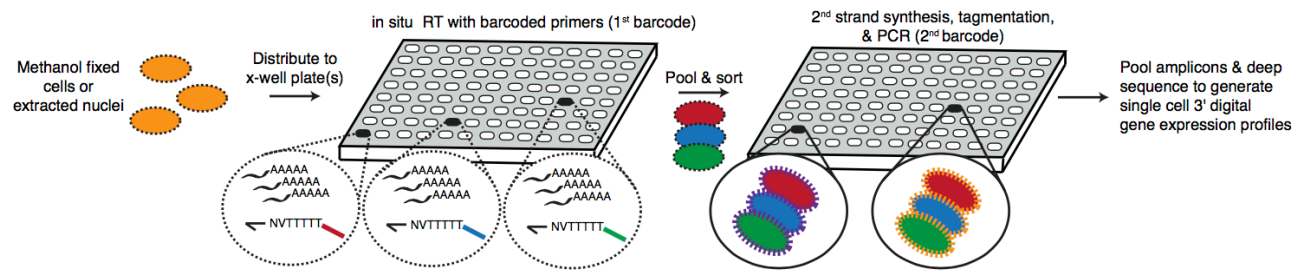


E

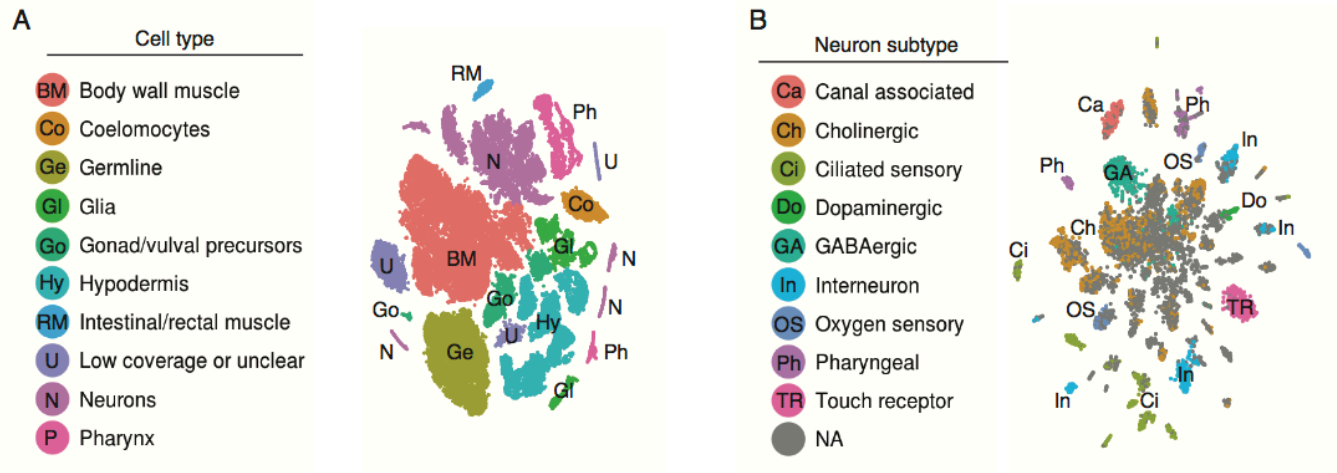
Rising		Declining	
Gene	ANOVA Rank (p-value)	Gene	ANOVA Rank (p-value)
Plekhb1	1 st (~0)	Fn1	1 st (4.41x10 ⁻¹¹)
Enho	2 nd (5.22x10 ⁻¹³)	Chst15	2 nd (6.91x10 ⁻¹⁰)
Sdc4	3 rd (2.10x10 ⁻¹²)	Epha7	3 rd (1.53x10 ⁻⁸)
Cdkn1a	7 th (9.71x10 ⁻¹⁰)	Sox11	17 th (4.23x10 ⁻⁰⁶)
Car14	13 th (2.00x10 ⁻⁸)	Prox1	759 th (0.01)

Pseudotime ordering of single cell samples according to differentiation stage (using Monocle package)

Increased throughput with combinatorial indexing



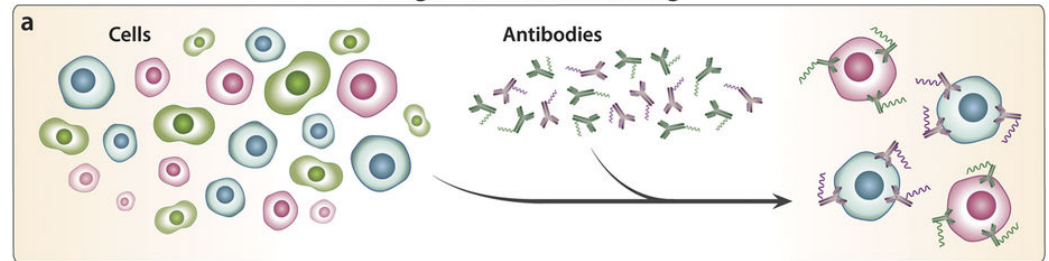
- Whole organism scale assays with combinatorial indexing
- Makes broad classification surveying possible
- Increased sampling allows greater resolution of dynamic processes and sparse data (epigenetics)



Encoding multiple modalities

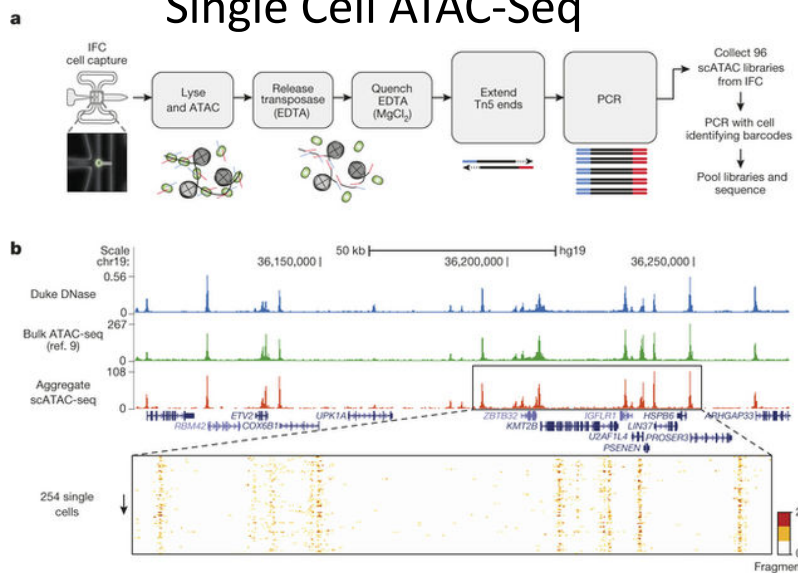
- Protein or Reporters
- Genomic / Epigenome
- Spatial Location
- History (such as activity)

Single Cell Ab-Seq Single-Cell Protein Profiling



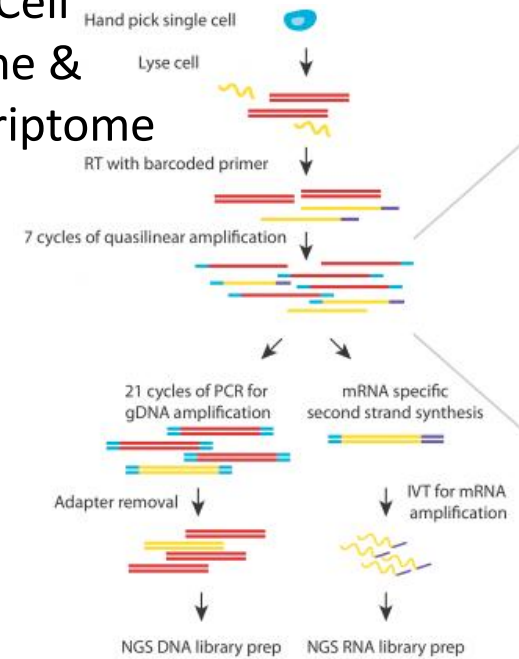
Shahi et al 2017

Single Cell ATAC-Seq



Buenrostro et al 2015

Single Cell Genome & Transcriptome



Dey et al 2015

Encoding lineage information

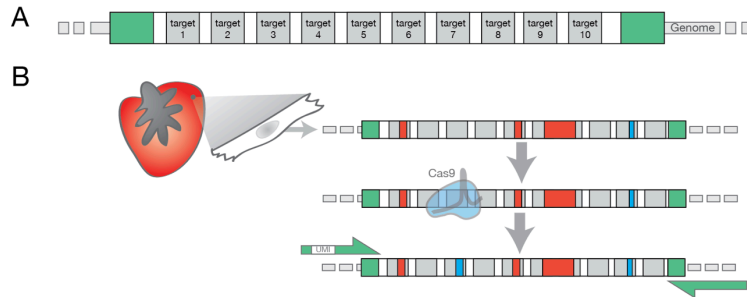
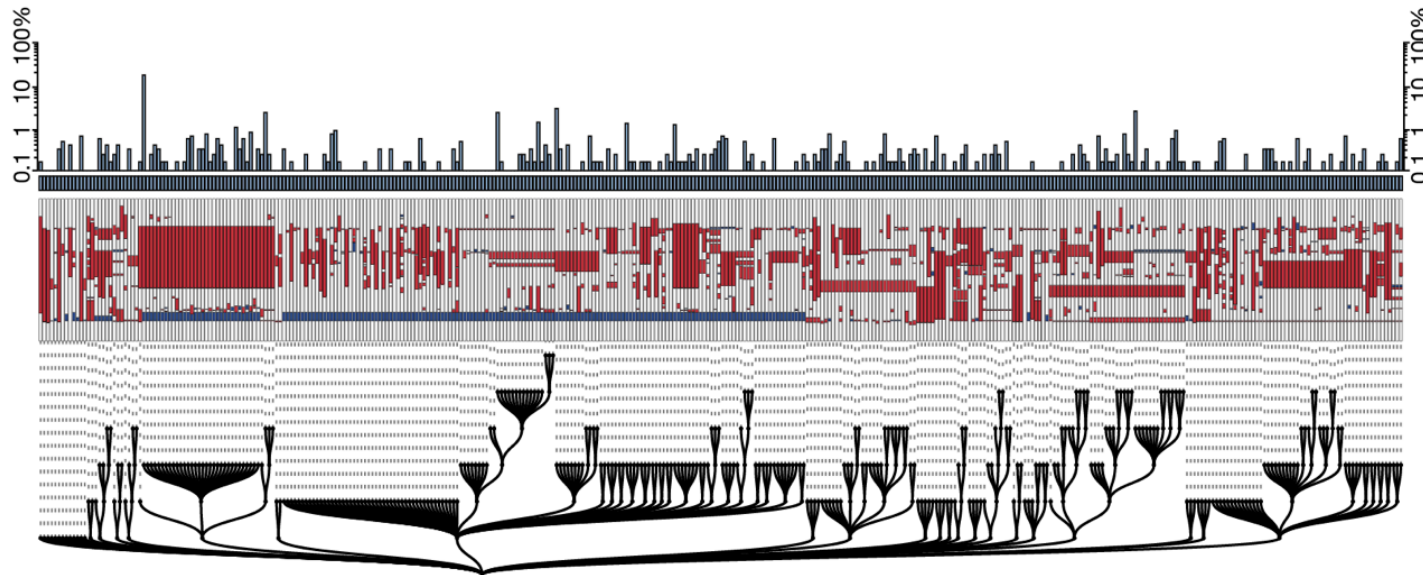


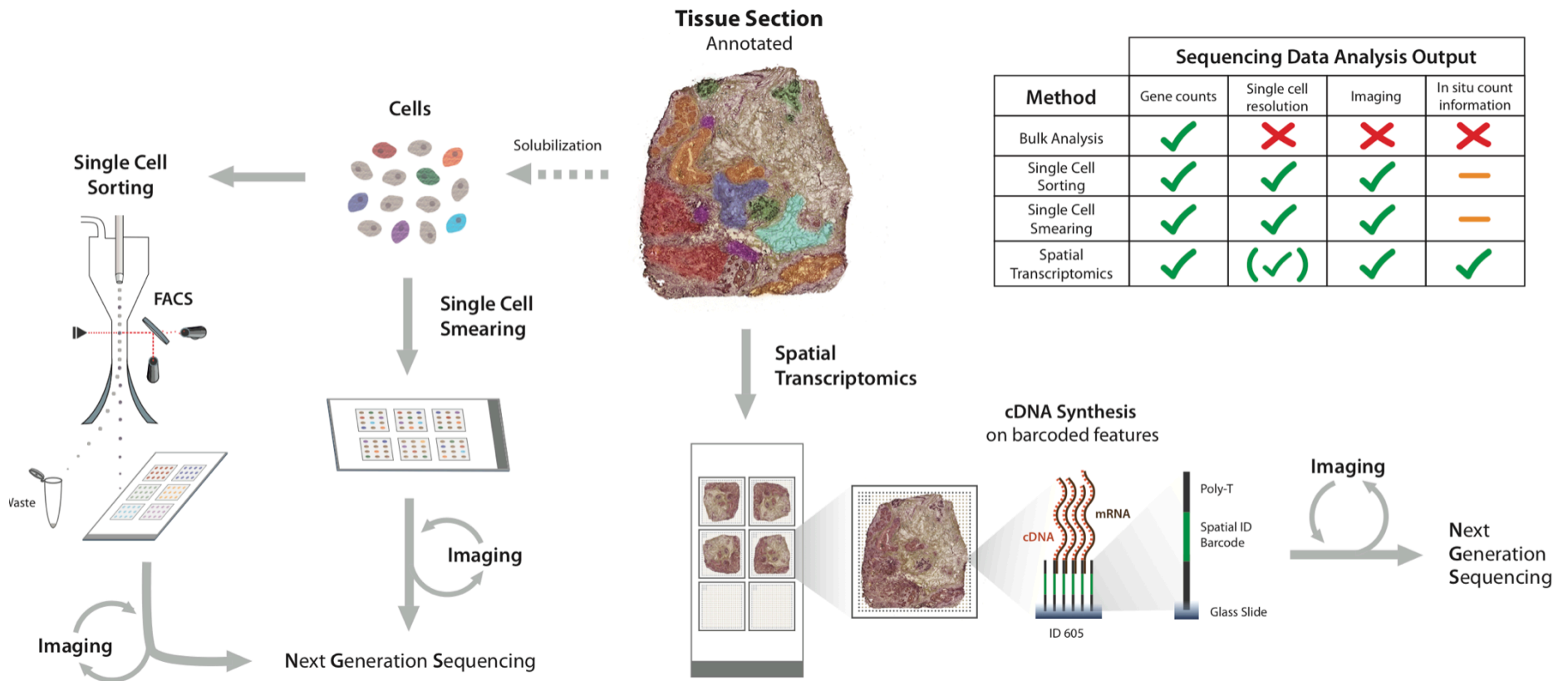
Figure 1. A GESTALT barcode. (A) A barcode with ten Cas9 target sites (gray bars), as well as flanking primer sequences (green) is introduced into the genome of interest. (B) A GESTALT barcode from a single cardiomyocyte of the heart. The barcode of this cell has already acquired deletions (red) and insertions (blue) in an ancestor cell, edits which are shared with other related cells. During this cell's lifetime Cas9 introduces an additional insertion (target 3), a mark that will be passed onto all progeny cells. The pattern of shared edits between many thousands of cells can be used to infer lineage.

- Accumulation of CRISPR mutations allow lineage reconstruction
- Theoretically can be linked to other information (such as transcriptome) for cell identity



3. Reconstruction of the alleles from a single zebrafish embryo using the V6 GESTALT barcode. Adapted from [1].

Spatial Information



Method	Sequencing Data Analysis Output			
	Gene counts	Single cell resolution	Imaging	In situ count information
Bulk Analysis	✓	✗	✗	✗
Single Cell Sorting	✓	✓	✓	—
Single Cell Smearing	✓	✓	✓	—
Spatial Transcriptomics	✓	(✓)	✓	✓

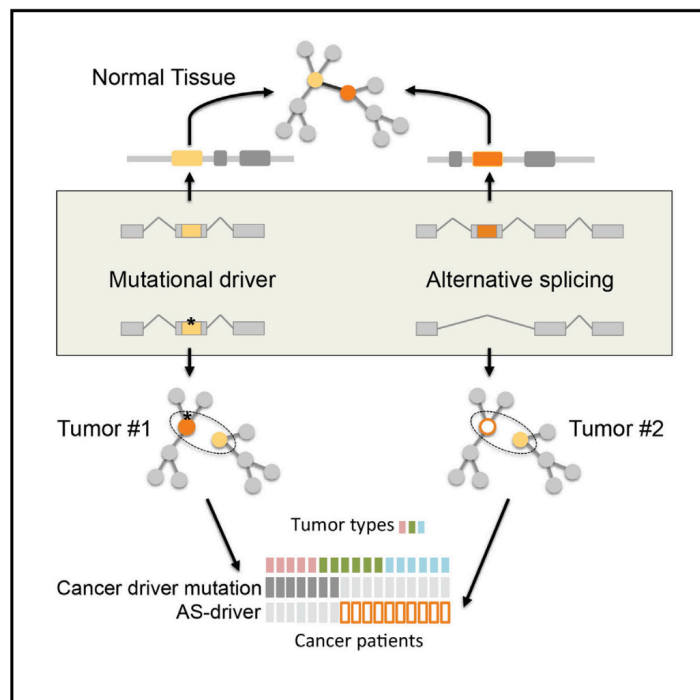
Single Cell Isoform Detection

Resource

Cell Reports

The Functional Impact of Alternative Splicing in Cancer

Graphical Abstract



Authors

Héctor Climente-González,
Eduard Porta-Pardo, Adam Godzik,
Eduardo Eyras

Correspondence

eduardo.eyras@upf.edu

In Brief

Climente-González et al. show that alternative splicing (AS) changes in tumors are linked to a significant loss of functional domain families that are also frequently mutated in cancer. These domain losses happen independently of somatic mutations and lead to the remodeling of complexes and protein-protein interactions in cancer.

- In the push for more cells and broader surveys, many platforms utilize gene-only level counting
- Transcript isoforms are important
- Single cell per well protocols that allow “full-length” are generally low throughput, expensive, and lack UMIs

Improved analysis methods

- Infrastructure and data structures to handle large multidimensional datasets
- Established workflows and best practices
- Modeling of dynamics, etc.
- Inclusion / handling of multi—modal datasets
- Discovery of fundamental transcriptional controls and gene regulatory networks
- Functional genomics at single cell resolution

Objectives

- Understand some of the key concepts in the methods and analysis of single cell genomics data
- Understand some of the current limitations
- Appreciate important experimental design considerations, including platform selection
- Be introduced to some of the “established” and emerging single cell genomics applications

Are there single cell RNA-Seq datasets to play with before collecting my own?

- Most single cell publications have data deposited in GEO
 - Can download raw data and usually processed expression matrices
- Some commercial platforms provide example datasets to view and analysis
 - 10X Genomics (<https://support.10xgenomics.com/single-cell/datasets>)
- Some analysis package developers provide example datasets
 - Seurat (http://satijalab.org/seurat/get_started.html)
- Some data can also be viewed in web-based portals

Some final thoughts

- Manage expectations
- Experimental design and platform selection is important
- Proper data analysis will take some time and your bioinformatician will be your best friend
- Don't assume bulk RNA-Seq analysis tools are appropriate for scRNA-Seq data

How do I keep up?

- *Who to watch:*
- *Regev, Linnarsson, Satija, Trapnell, Teichmann, Kharchenko and Shendure Labs*
- *Twitter: @scell_papers*

- *NIH Single Cell User Group*
 - *<http://nih-irp-singlecell.github.io>*