

# Single Cell Forum

July 23, 2020

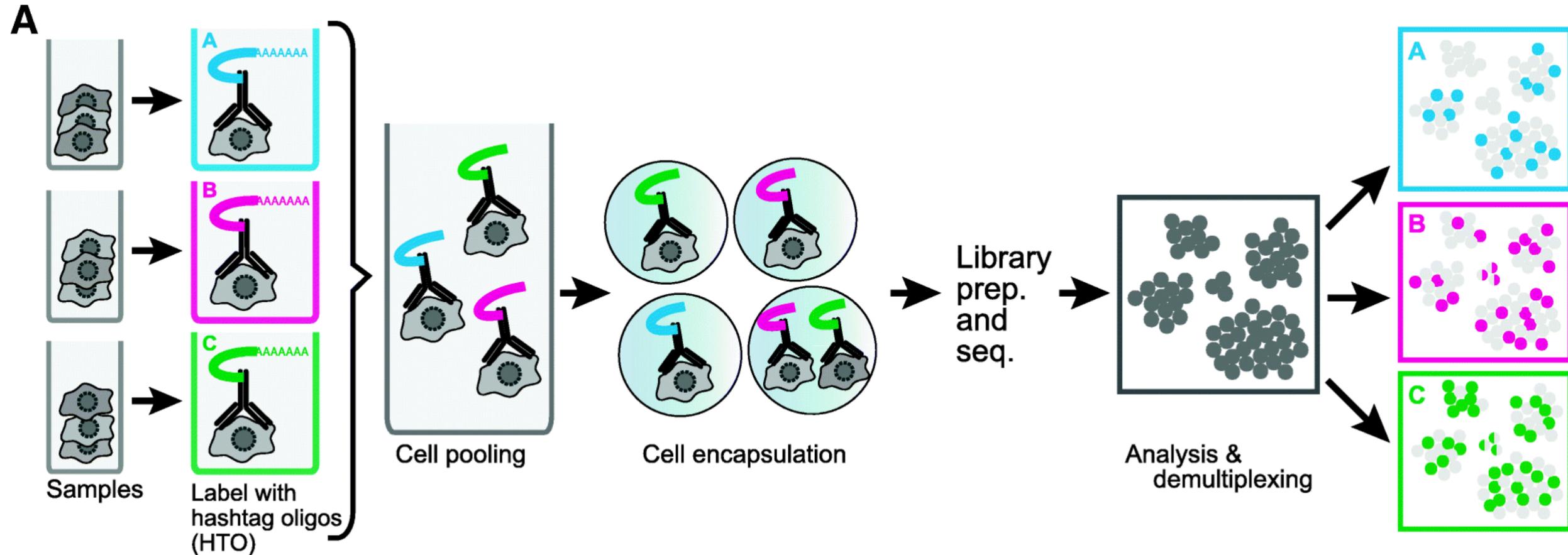
# Multiplexing Samples with Natural Barcodes

- ‘Natural barcode’ for demultiplexing: variant-based deconvolution of samples
- Probable prerequisites:
  - Genotype of each sample determined *a priori*
  - Given that distributions of reads is biased to 3’ and 5’ ends of transcript – though with surprisingly long tail, would need fairly deep sequencing to permit deconvolution of cDNA
- Resolution
  - Frequency of SNPs is roughly 1 per 1,000 bp
  - Given 100 bp read length, a generous – but naïve - estimate is that 10% of reads would capture a SNP; remainder of reads bland
  - Estimate naïve for many reasons, one of which already mentioned: actual distributions of reads is biased to 3’ and 5’ ends of transcript
  - If genotypes of samples are close, frequency of distinguishing SNPs much lower

# Multiplexing Samples with Natural Barcodes

- Possible work arounds:
  - Construct single cell libraries separately but with different sample IDs; multiplex labeled libraries for sequencing [achieve some economy from lower sequencing depth requirements]
  - Use cell hashing with barcoded antibodies before library construction [Permits both multiplexing as well as doublet detection – works nicely with recent enhancements in Seurat]

# Multiplexing Samples with Natural Barcodes



M. Stoeckius et al. 'Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics' *Genome Biology* **19**(2018)

# Assessing validity of clustering

- Optimizing experimental design
  - Cell surface protein expression [CITE-seq / REAP-seq] - if characteristic for particular cell type - can be very helpful. Beware the background.
  - Cell hashing to label (sub) samples can preserve information you might have about a sample
- Optimizing subsequent analysis (To be discussed in more detail in future event)
  - Assessing for stability of clusters with respect to parameter changes
  - Awareness that fewer/more clusters may arise as we convolve/resolve subsets (e.g. different subsets of immune cells)
  - Good statistical methods to subtract out background from CITE-seq data
  - Transferring labels/annotations from existing annotated datasets

# How many cells to sequence?

- Different technologies have different capabilities
- Roughly speaking, in terms of numbers of cells
  - SMART-seq < Droplet based methods < Microwell methods < Combinatorial Indexing methods
  - Trade off: lower throughput methods retrieve more genes
- Restricting attention to one popular technology 10X scRNA-seq
  - About 10,000 cells is [current] ceiling
  - Trade off: more cells implies potentially more doublets

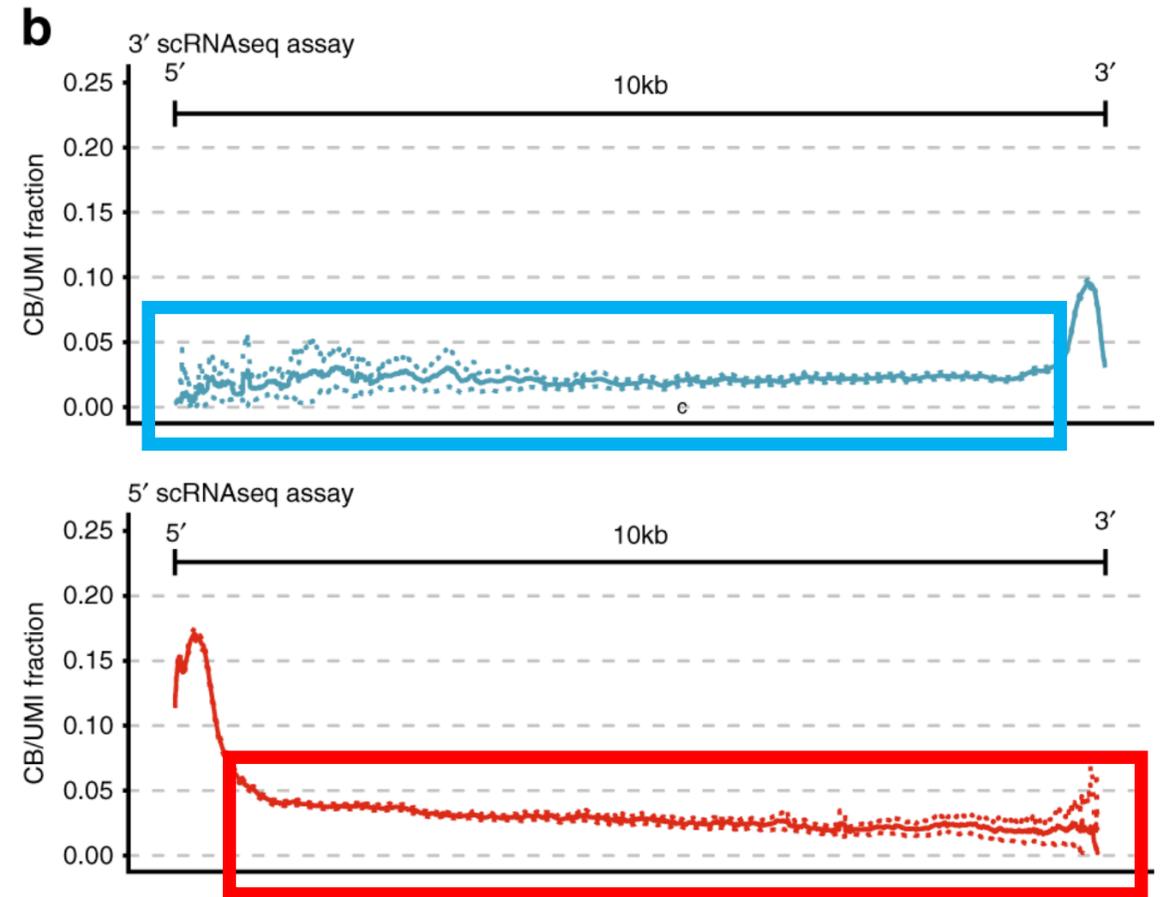
# How many cells to sequence?

- Doublets can be partially mitigated
  - Experimentally via cell hashing
  - Bioinformatically [to be discussed next time]

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~800	~500
~0.8%	~1,600	~1,000
~1.6%	~3,200	~2,000
~2.3%	~4,800	~3,000
~3.1%	~6,400	~4,000
~3.9%	~8,000	~5,000
~4.6%	~9,600	~6,000
~5.4%	~11,200	~7,000
~6.1%	~12,800	~8,000
~6.9%	~14,400	~9,000
~7.6%	~16,000	~10,000

# Sequencing Depth

- Depending on goals of experiment require different sequencing depths
- Sequencing to greater depth, enables recovery of
  - Less highly expressed genes
  - Mutations (for characterization of clones/subclones)
  - Splice junctions (for information on alternative splicing)
  - Fusion genes



A. Petti et al. 'A general Approach for detecting expressed mutations in AML cells using single cell RNA-sequencing'. *Nature Communications* **10**(2019)

# Sequencing Depth

- 10X recommendation for single cell 3' v3.1 Gene Expression: 20,000 read pairs/cell (Older recommendation for v2 was 50,000 read pairs/cell)
- Parenthetically: There are variations in read depth that vary by cell. Modern bioinformatics analysis packages (e.g. Seurat and others) mitigate much of this