2018 ABRF Meeting – Satellite Workshop 4
Bridging the Gap: Isolation to Translation (Single Cell RNA-Seq)
Sunday, April 22

Basics of RNA-Seq

(With a Focus on Application to Single Cell RNA-Seq)

Michael Kelly, PhD

Team Lead, NCI Single Cell Analysis Facility

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

Outline

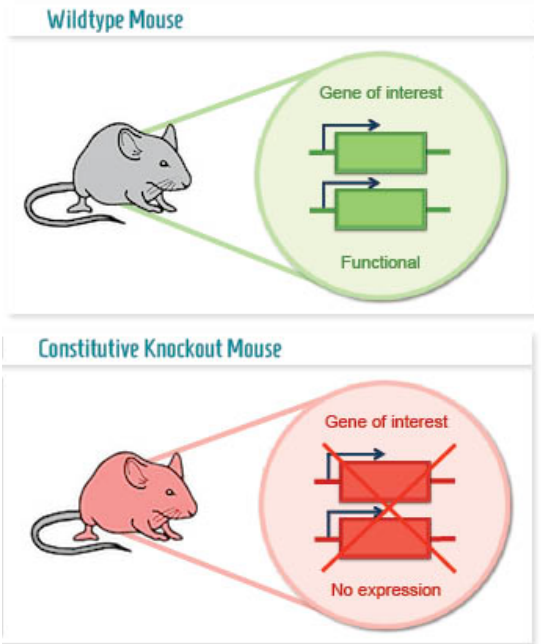
- **Why RNA-Seq?**
- **RNA: What are we assaying?**
- **RNA-Seq with Illumina ”Next-Generation” Sequencing (NGS)**
- **NGS RNA-Seq Data Processing**

- RNA-Seq at Single Cell Resolution (scRNA-Seq)

Why RNA-Seq?

- Assaying gene expression differences between conditions (as well as a variety of other experimental designs)

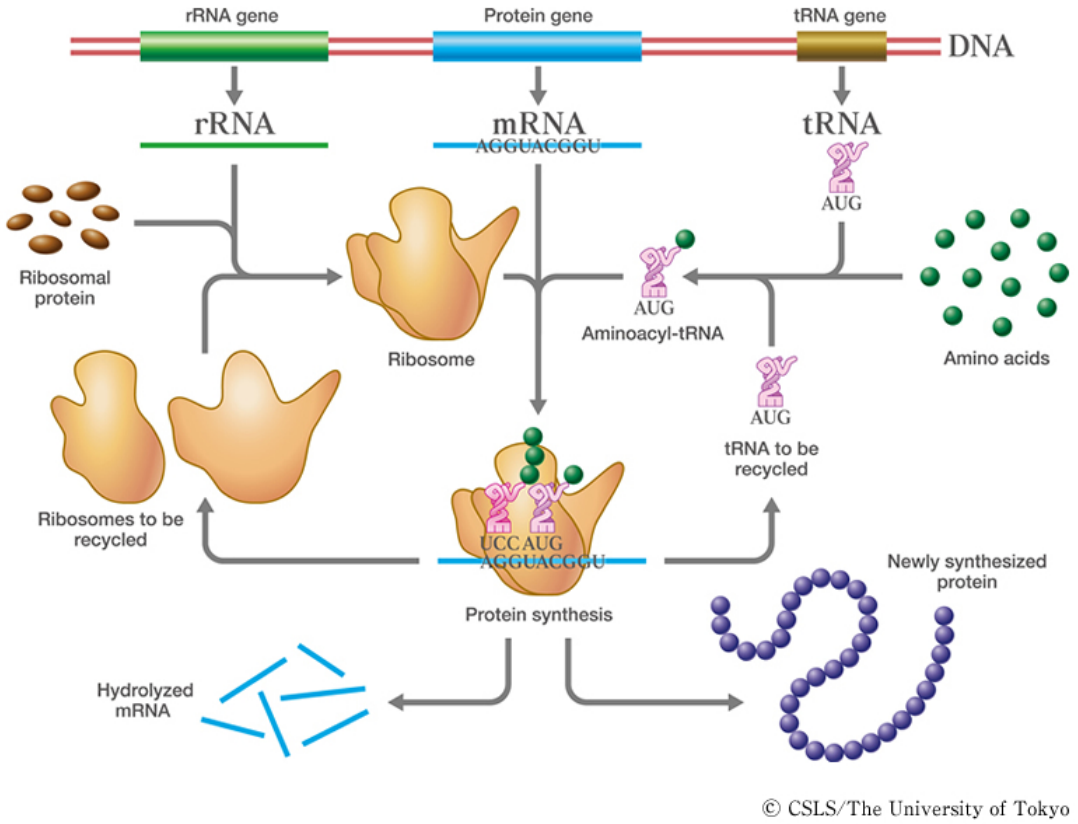
- Whole transcriptome = maximum “discovery”
- Increasingly “routine” methodology and analysis
- Sequencing costs have decreased
- Increased access to facilities and expertise



What other genes are affected by knocking out target gene “A”?

Diagram modified from Genoway.com

World of RNA – From Genome to Functional Protein

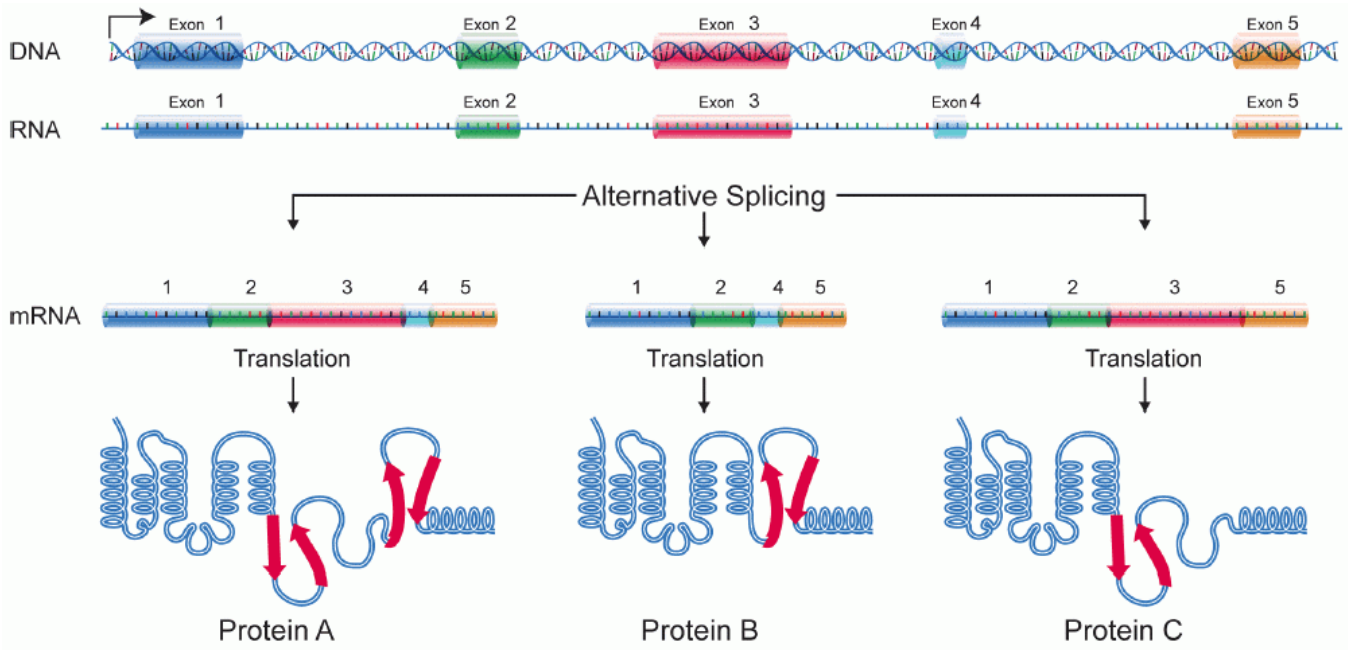


Only 1-3% of Total RNA is mRNA!

Other RNA's not shown, including miRNA, lncRNA, etc.

Alternative Splicing Increases RNA Transcript Complexity

- mRNA transcribed from gene locus, which in most cases, is made up of exons and introns
- Post-transcription, nascent transcripts have their exons spliced together
- Different sets of exons can be used; can result in alteration to protein coding sequence & function
- Splicing can vary across tissues, specific cells, developmental time & disease



Not all exons / transcript isoforms are well annotated in standard references…

Diagram from Wikipedia "Alternative Splicing Entry"

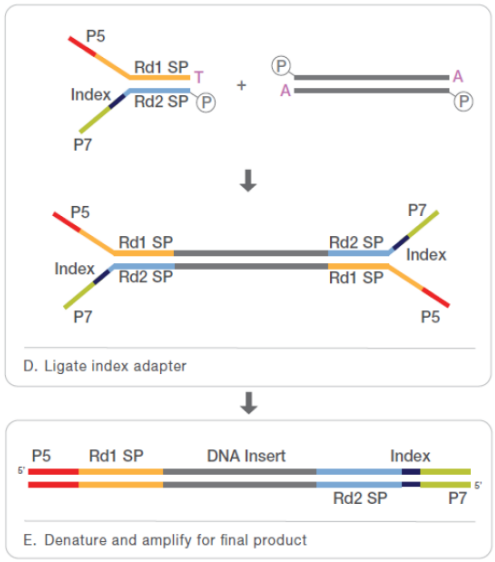
How to we sequence it?

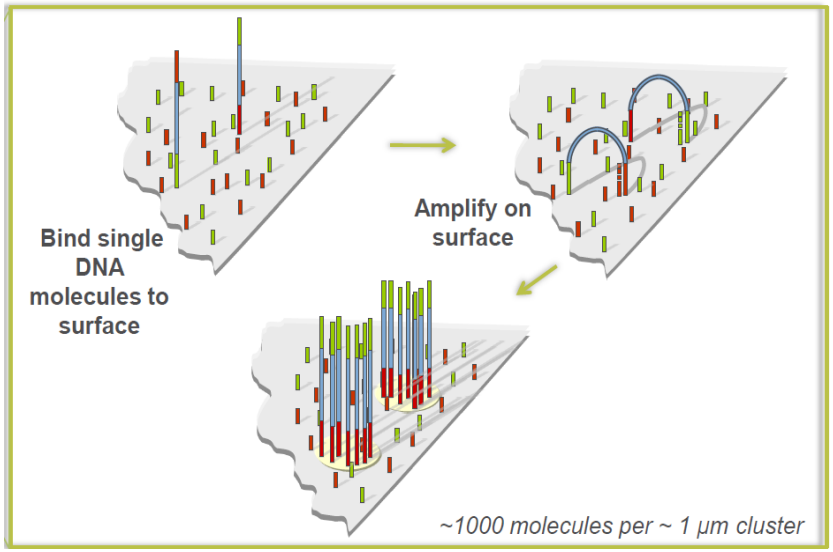
- Reverse transcribe RNA sequence into complementary DNA sequence
- Generation of second strand and amplification (if needed)
- Preparation of sequencing library by fragmenting full-length molecules and addition of adapters

Size limitation: Fragments ideally ~500 bp for Illumina sequencing

Diagram from Thermo Fisher Sci "Reverse Transcription"

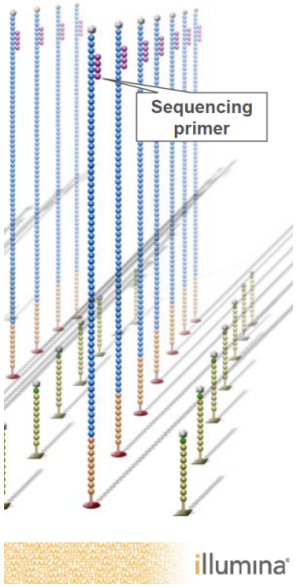
Illumina Next-Generation Sequencing (NGS) “Sequencing by Synthesis”

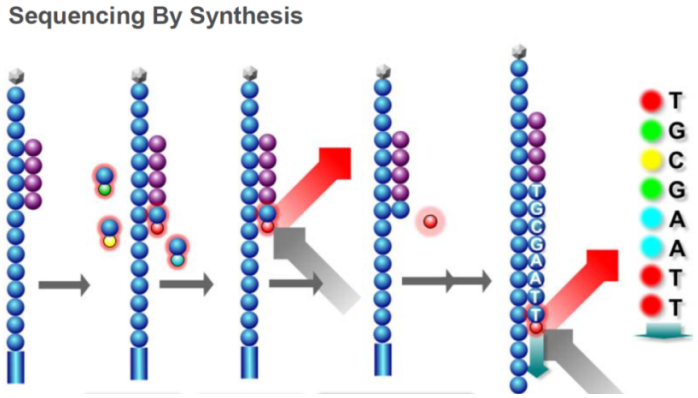




- Adapters on ends of cDNA molecules allow binding to sequencing “flow cell”
- Amplification of clonal “clusters” of cDNA fragments
- “Massively parallel sequencing”: hundreds of millions of fragments sequenced in parallel

Illumina Next-Generation Sequencing (NGS) “Sequencing by Synthesis”

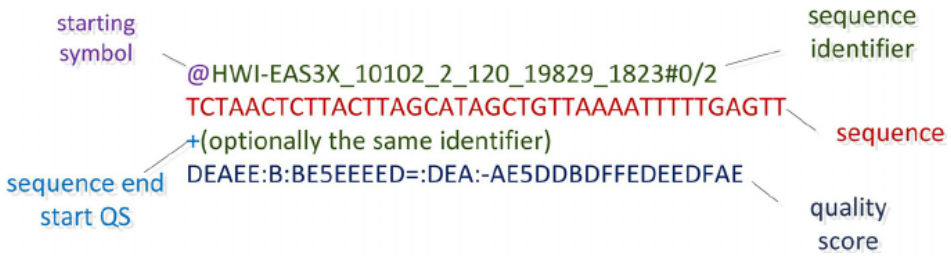




- Sequence read by fluorescent nucleotide incorporation during each “cycle”
- Each cluster dot will display a color associated with nucleotide (A, B, G, or T)
- Image processing -> conversion to Fastq output (sequence with quality score)

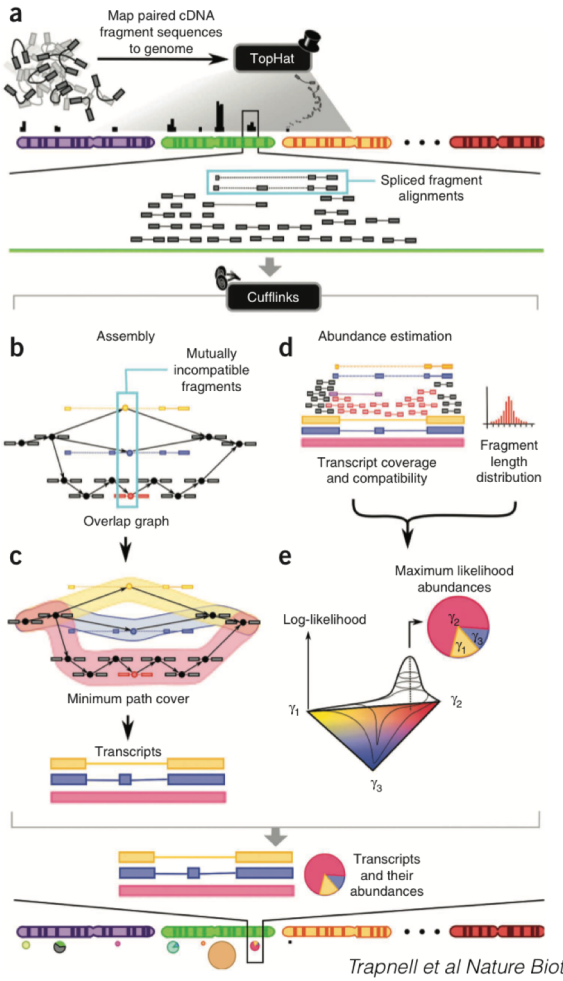
Aligning / Mapping Reads & Putting the Pieces Back Together

Example Fastq sequence record



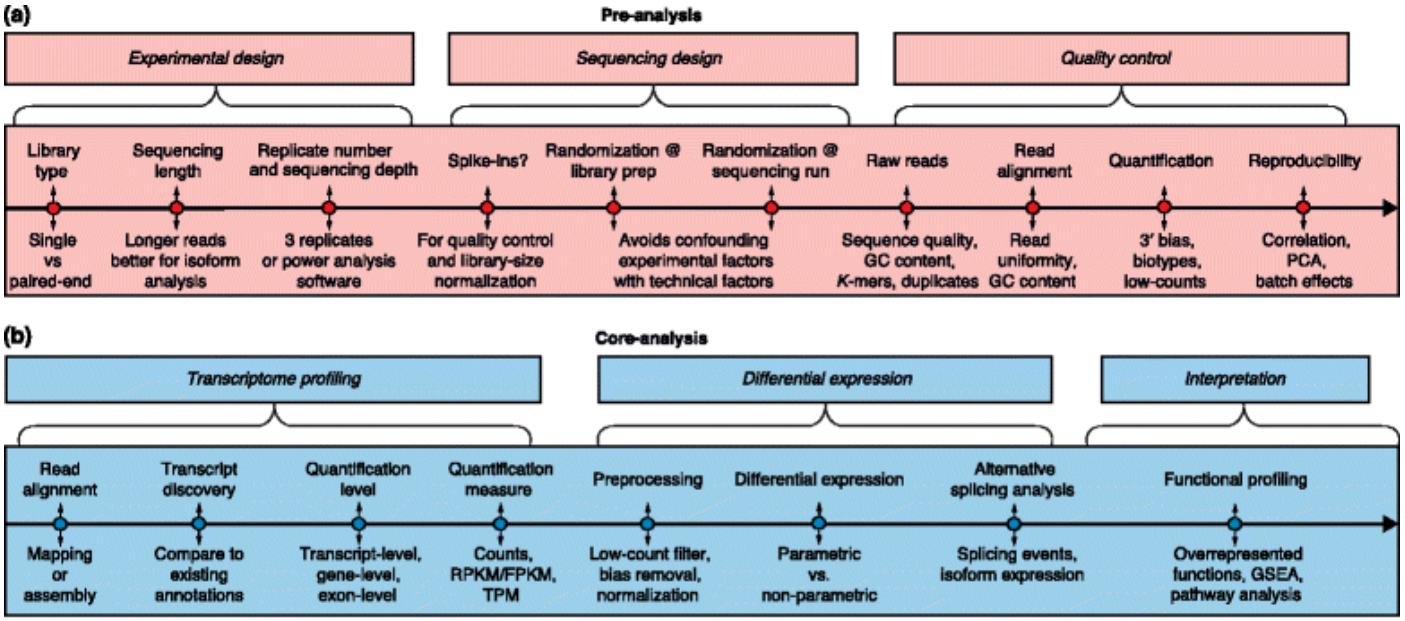
From Pavlopoulos et al (2013)

- Reads aligned to a reference genome (or transcriptome), while accounting for splicing at exon junctions
- Note that fragment length is usually longer than actual sequencing reads – end-sequencing to infer insert
- Transcript isoform inferred from exon coverage and spliced reads (aided with statistical algorithm to resolve ambiguity)
- Transcript abundance estimation should account for transcript size – *larger transcripts more likely to have more reads…*



Trapnell et al Nature Biotech 2010

Overview of RNA-Seq Analysis



Conesa et al (BMC Genome Biology 2016)

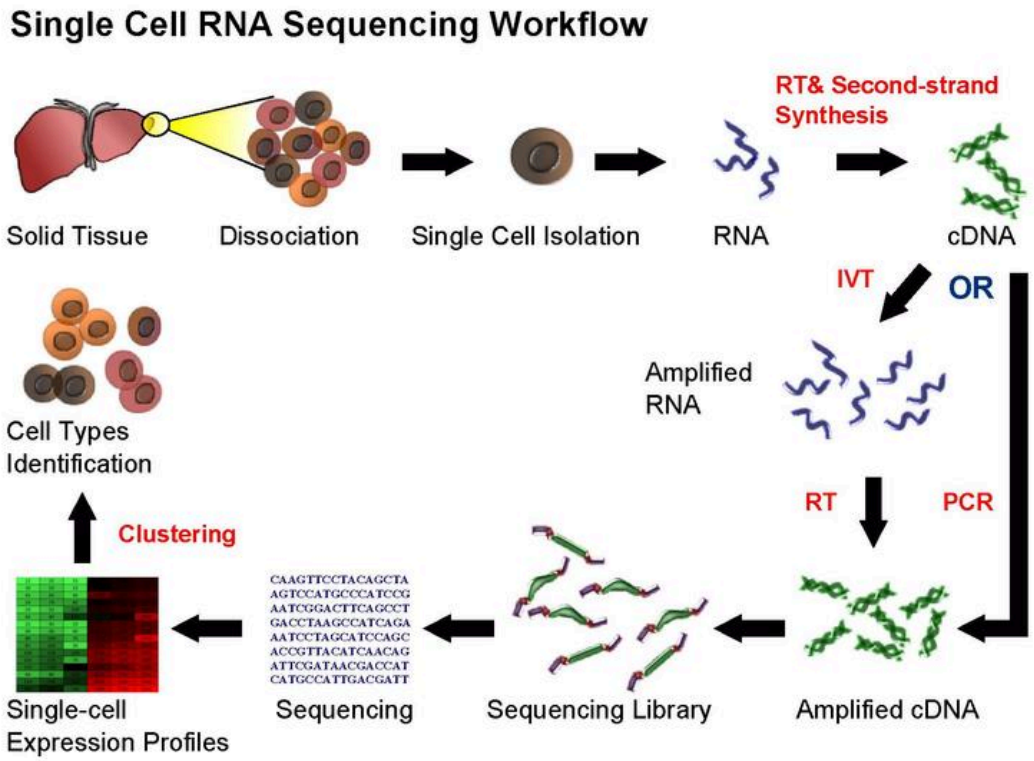
- Experimental design and quality control are crucially important – talk to your bioinformatician early and often!
- Many tools exist for primary data processing, differential expression (DE) testing, and functional / pathway testing
- RNA-Seq datasets can continue to be utilized beyond the scope of the original study – queried as expression database, integrate with other datasets, etc.

Outline

- Why RNA-Seq?
- RNA: What are we assaying?
- RNA-Seq with Illumina ”Next-Generation” Sequencing (NGS)
- NGS RNA-Seq Data Processing

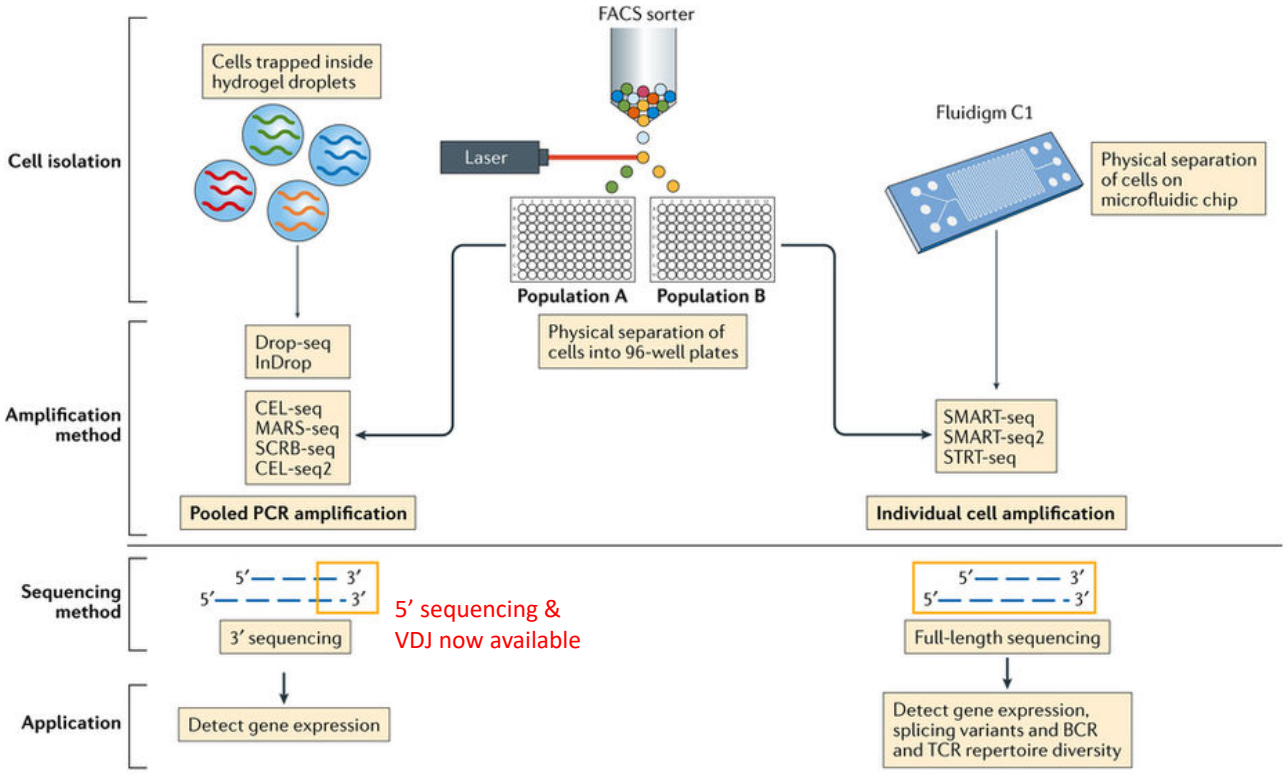
- **RNA-Seq at Single Cell Resolution (scRNA-Seq)**

RNA-Seq at Single Cell Resolution



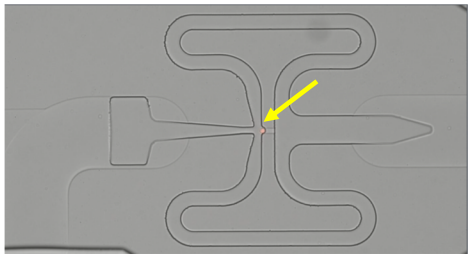
https://hemberg-lab.github.io/scRNA.seq.course/

Overview of Common Single Cell RNA-Seq Methods

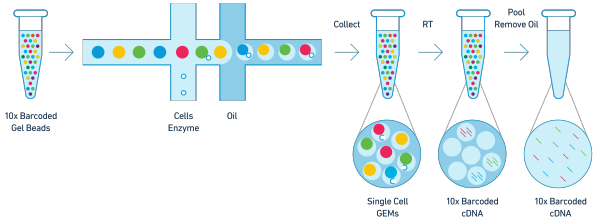


Papalexi & Satija (2018)

Fluidigm C1 96-site chip



10X Genomics Chromium



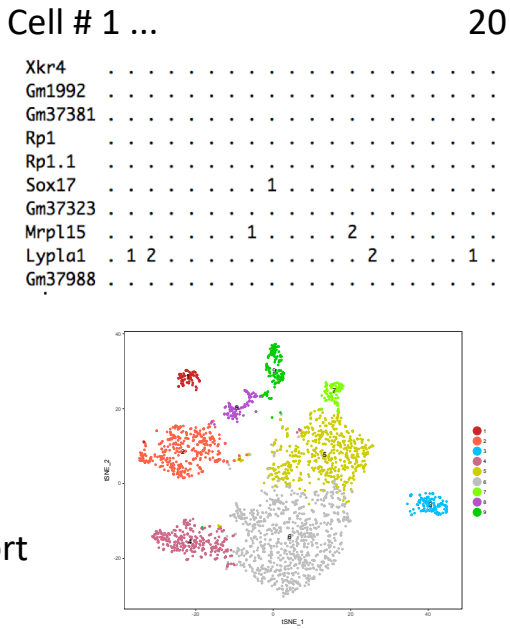
Is single cell RNA-seq just RNA-seq with more samples? Not really.

- scRNA-Seq is zero-heavy data
  - Depending on method, you could have 500 genes of 40,000 have non-zero values
  - Analysis is a combination of discrete and continuous math (10 vs 0, and 1000 vs 1)

- Differential expression usually starts 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

- Don't trust any one gene. Dimensionality reduction provide more reliable "meta-genes"
  - Both "drop-out events and noise/over-amplification can give the wrong impression
  - Biologically relevant principle components can represent "meta-genes" that can help sort out cell types

- Protocols are limited by the low-input amount of RNA
  - scRNA-Seq relies on quite a bit of PCR
  - Total RNA, stranded, specialized protocols or total RNA methods generally not supported
  - Reverse transcription usually happens in the presence of the lysate (not ideal conditions)



- Manage expectations
- Don't assume bulk RNA-Seq analysis tools are appropriate for scRNA-Seq data

Summary – RNA-Seq Basics

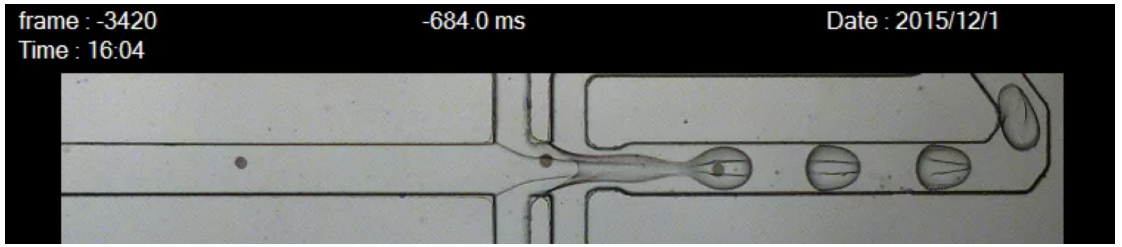
- RNA-Seq allows the systematic assaying of the RNA expression within biological systems with maximal discovery
- Selection of RNA-Seq protocol define which RNA molecules are assayed (i.e. oligo-dT methods only pick up poly-adenylated transcripts)
- Illumina NGS sequencing requires fragmentation of molecules for sequencing library preparation
- Informatic processing aligns reads to an annotated reference to determine transcript identity; transcript isoforms can be inferred from sequenced fragments by ratios of exon usage and splice sites
- Characterization of gene expression, or differences in gene expression can be determined by carefully controlled bioinformatic analysis

Summary – Single Cell RNA-Seq Basics

- More widely accessible protocols are limited to RNA-Seq of polyadenylated transcripts
- Sensitivity of detection is quite low and affected by many technical challenges – results in ”noisy” and zero-heavy data

- Technology, methods, and analysis methods have advanced, resulting in higher throughput, lower cost, and better feasibility as a increasingly “common” gene expression methodology
- Generates extremely flexible datasets, but still requires bioinformatics investment / expertise – great efforts being made to make the analysis more accessible

Acknowledgements



- National Institute on Deafness and Other Communication Disorders (NIDCD)
  - Laboratory of Cochlear Development (Matthew Kelley's Lab)
  - Genomics & Computational Biology Core
    - Robert Morell, PhD & Erich Boger, PhD
- NCI Center for Cancer Research
  - Office of Science Technology Resources
  - NCI Genomics Core
    - Val Bliskovsky, PhD & Liz Conner, PhD
- Frederick National Lab - Cancer Technology Research Program
  - Sequencing Facility & Genomic Technology Lab
- Developers & Maintainers of Open-Source Methods and Software