Introduction to Next-Generation Sequencing Technologies

Kris Holton

HMS Research Computing

Genetics 303qc

HMS Research Computing

- Manage O2 High Performance Compute Cluster
- Research Computing Consultants
  - Planning experiments
  - Analysis
  - Scaling/scripting
- User Training
  - O2 for New/Orchestra users
  - R/Python/Perl/Matlab
  - Parallel Computing/Git and Github
  - NGS
  - Biostatistics

rchelp@hms.harvard.edu

Topics for today

- Sequencers + Technology
    - HiSeq/MiSeq/NextSeq/IonTorrent/PacBio/NanoPore/Fluidigm
- NGS Branches
    - DNA/ChIP/ATAC/Exome/RNA/miRNA/SingleCell/Drop/inDrop/10x/CLiP/Ribo/16s
- Library Prep
- Analysis
    - Options
    - Software
- Experimental Design
- Data Deposition

Sequencing Core

- Two Illumina cBot stations
- One Illumina HiSeq 2500 sequencer
- Three Illumina MiSeq sequencers
- Four Illumina NextSeq 500 sequencer
- Single-cell: Fluidigm C1
- Library prep service: IntegenX Apollo
- Shearing: Covaris S2
- QC: Agilent TapeStation, BioAnalyzer

Illumina HiSeq 2500

- Up to 2 x 250 reads (paired end)
- Rapid Run or High Output
- Single or Dual Flow Cell
- Flow Cell: 8 lanes
- Up to 1TB/run

Illumina MiSeq

- Targeted, small genome
- 2 x 300 reads (paired-end)
- 15GB output/run
- Single flow cell
- Single lane
- Multiplex: up to 384 samples/run

Illumina NextSeq 500

- 2 x 150 reads (paired end)
- High Output/Mid Output
- Up to 120GB/run
- Single flow cell
- 4 lanes/flow cell

SBS: Sequencing By Synthesis

- Video!

Ion Torrent

- Semiconductor chip
- Adding dNTP: release pyrophosphate + H^+
- Add single nucleotide, measure proton release
- 400 base read length
- Homopolymers

PacBio

- SMRT technology: Single Molecule, Real-Time
- Long read lengths (circularized)
- Zero-mode waveguides (illuminate well)
- Labelled flurophores
- “Movie” of sequence-by-synthesis

Oxford NanoPore

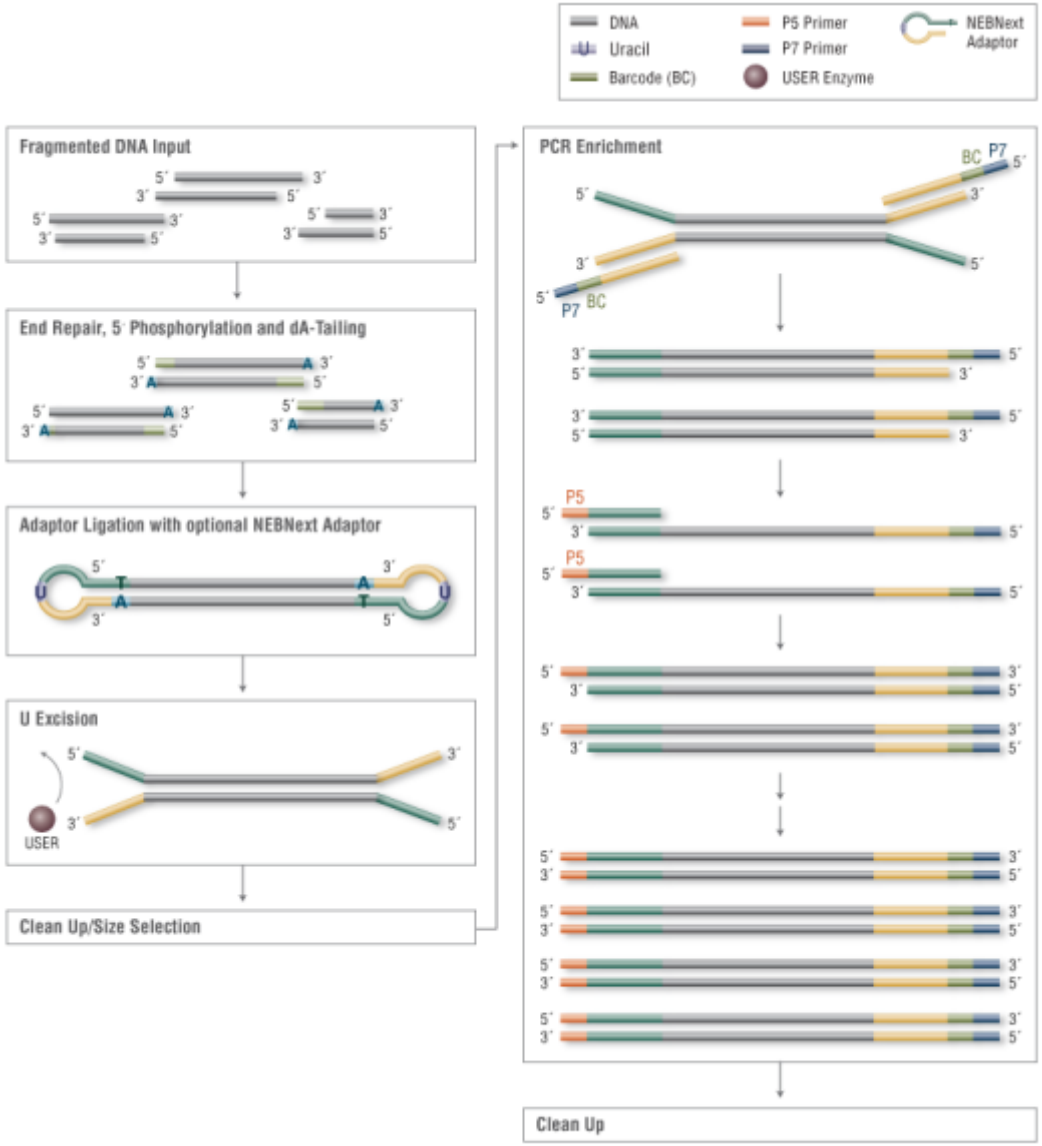
- Biological or synthetic pores
- Measure change in current through pore
- Long read lengths (200KB)
- Real Time
- Portable (USB)
- Application: any type of molecule

Fluidigm C1

- Single cell isolation
- Integrated fluidic circuit
- Stain captured cells/visualize for viability, cell surface markers, reporter genes
- Lyse for ‘seq

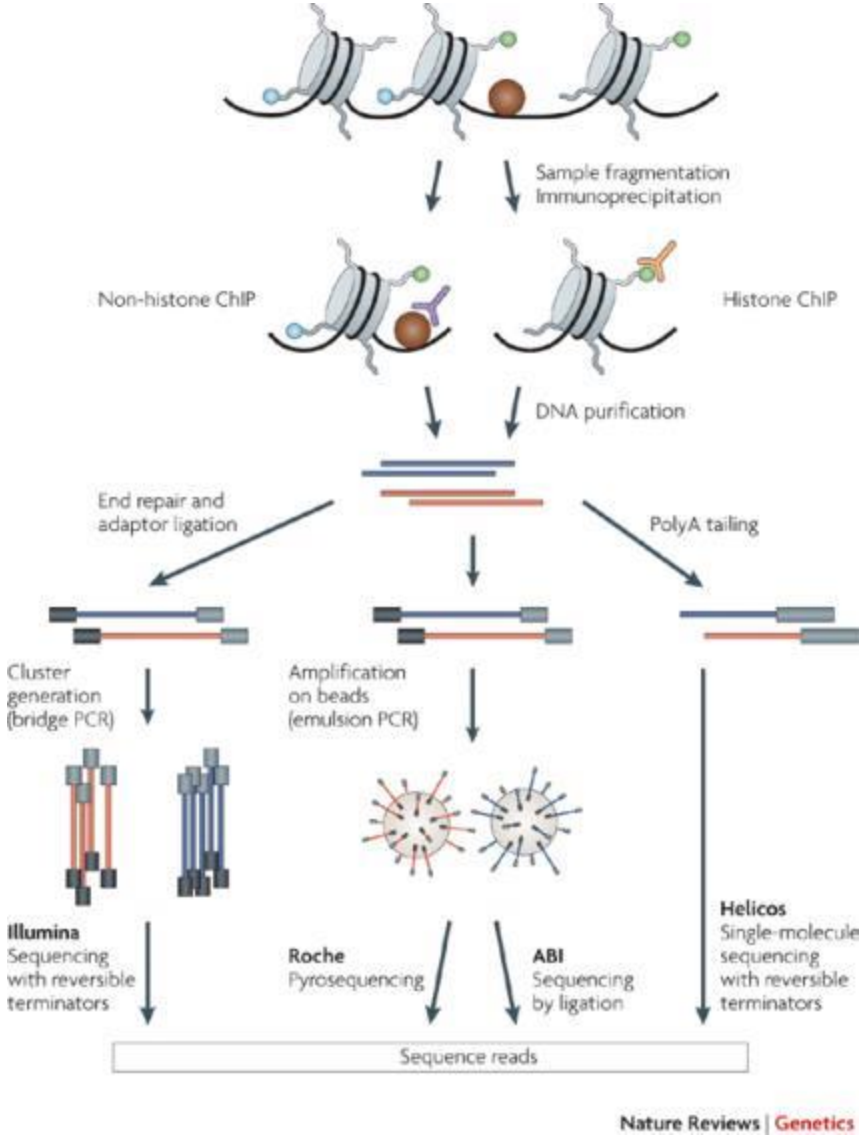
NGS Technology Variations

DNA-seq



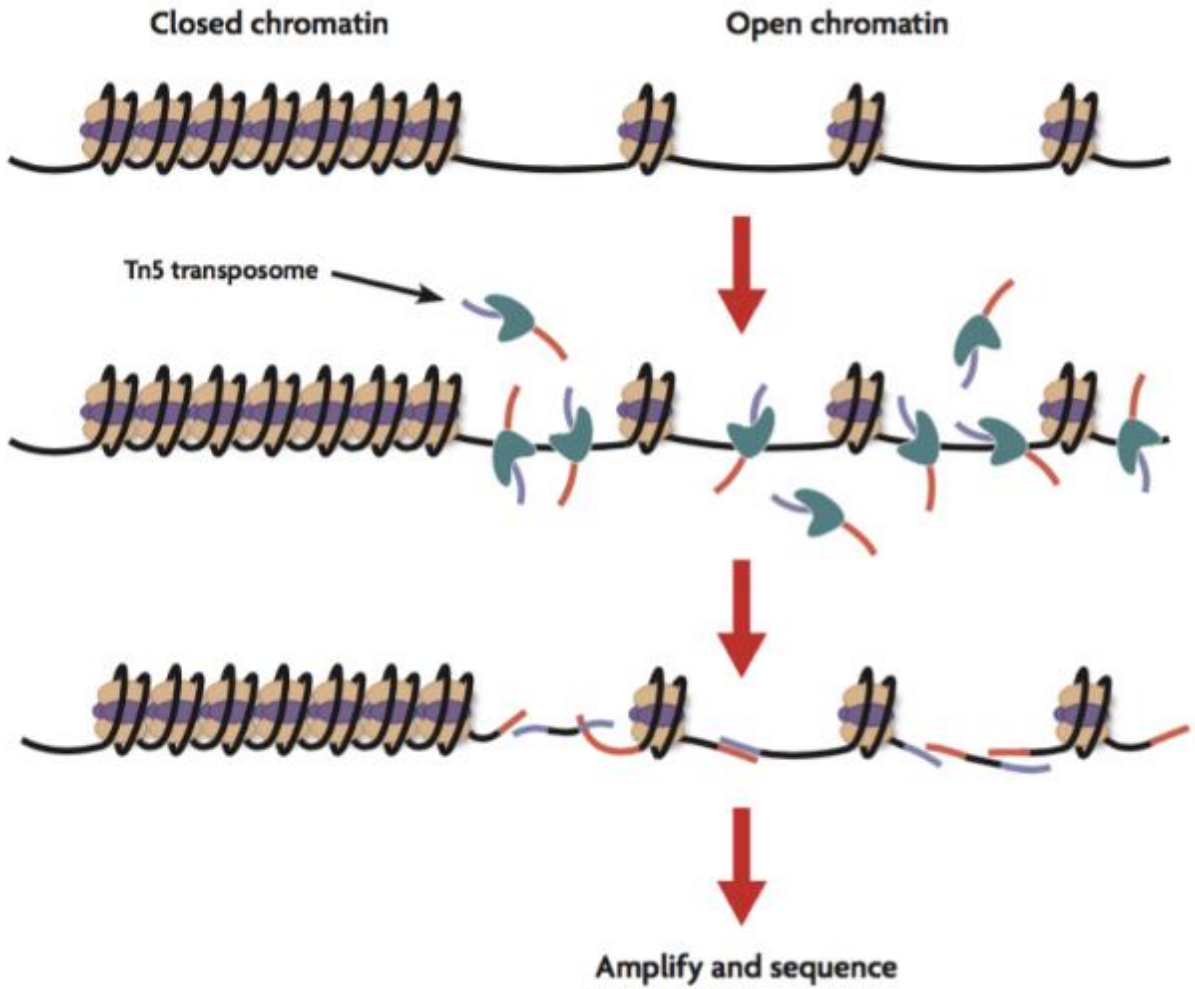
New England Biolabs

ChIP-seq



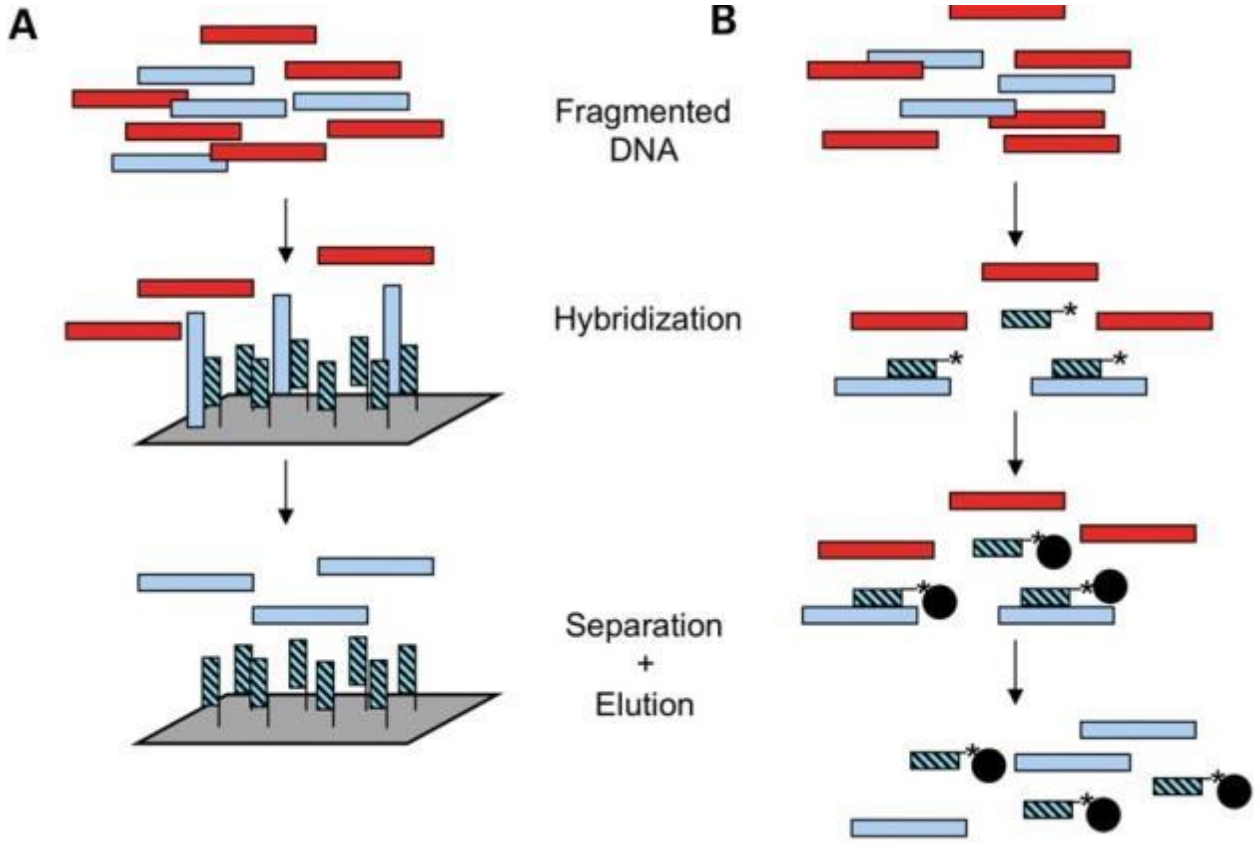
Peter J. Park, Nature 2009

ATAC-seq



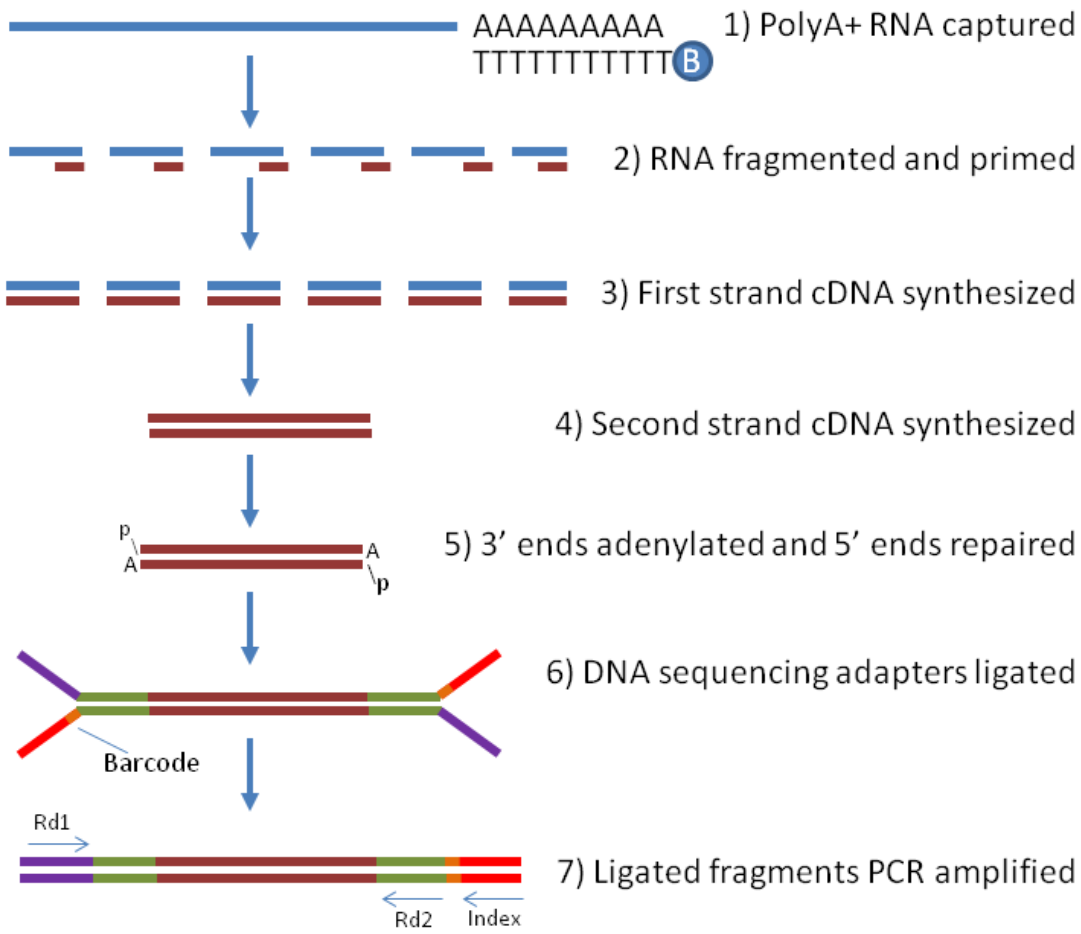
www.activemotif.com

Exome Sequencing - Capture



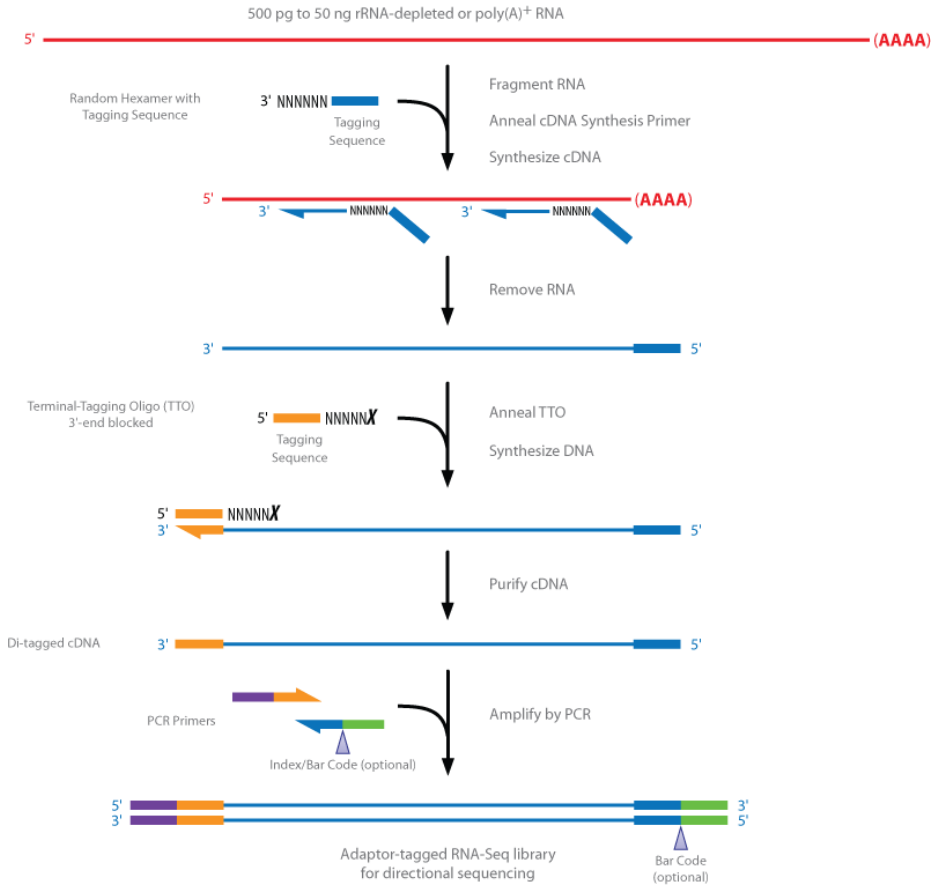
Teer & Mullikin, Human Molecular Genetics 2010

RNA-seq



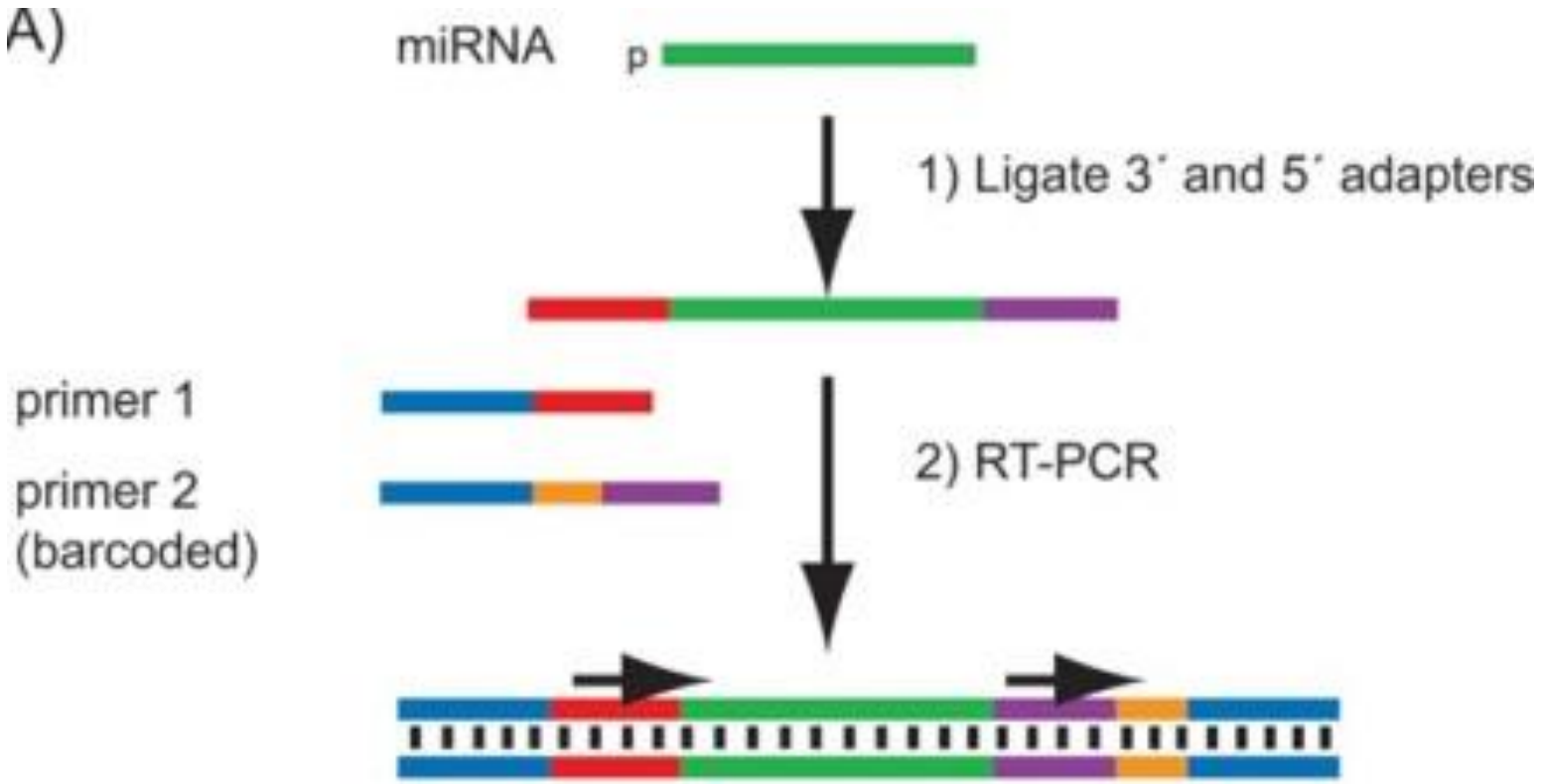
Labome

RNA-seq: strand-specific



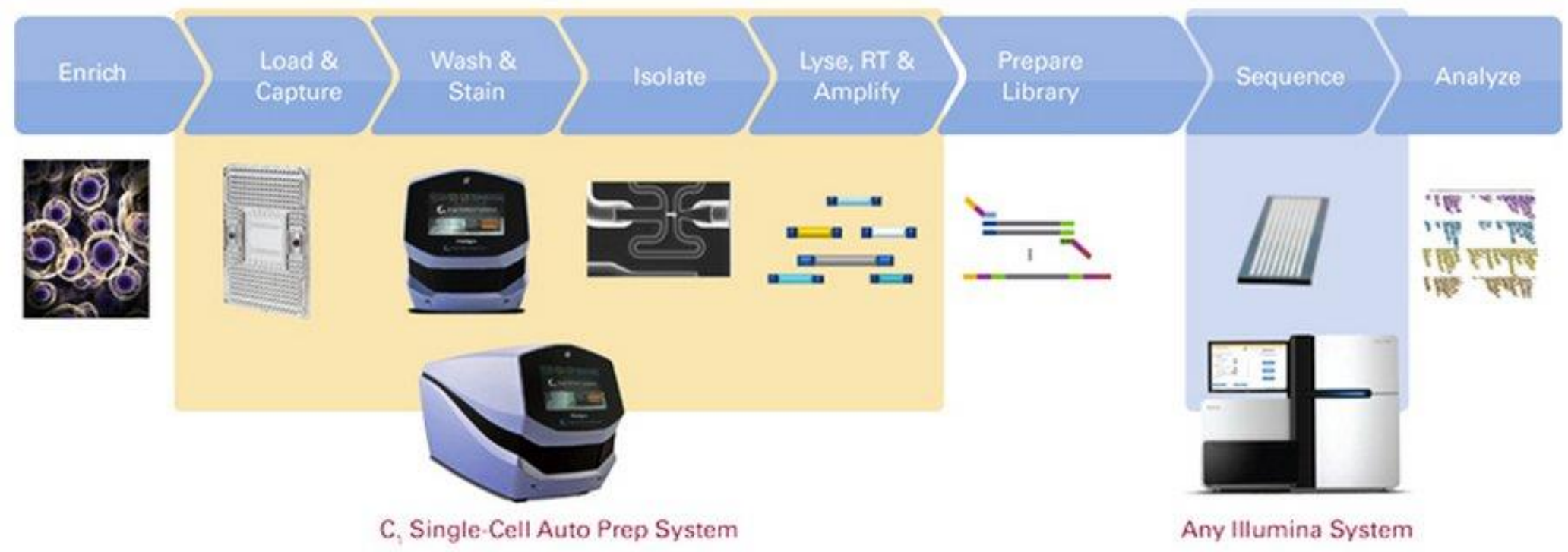
Illumina

miRNA-seq

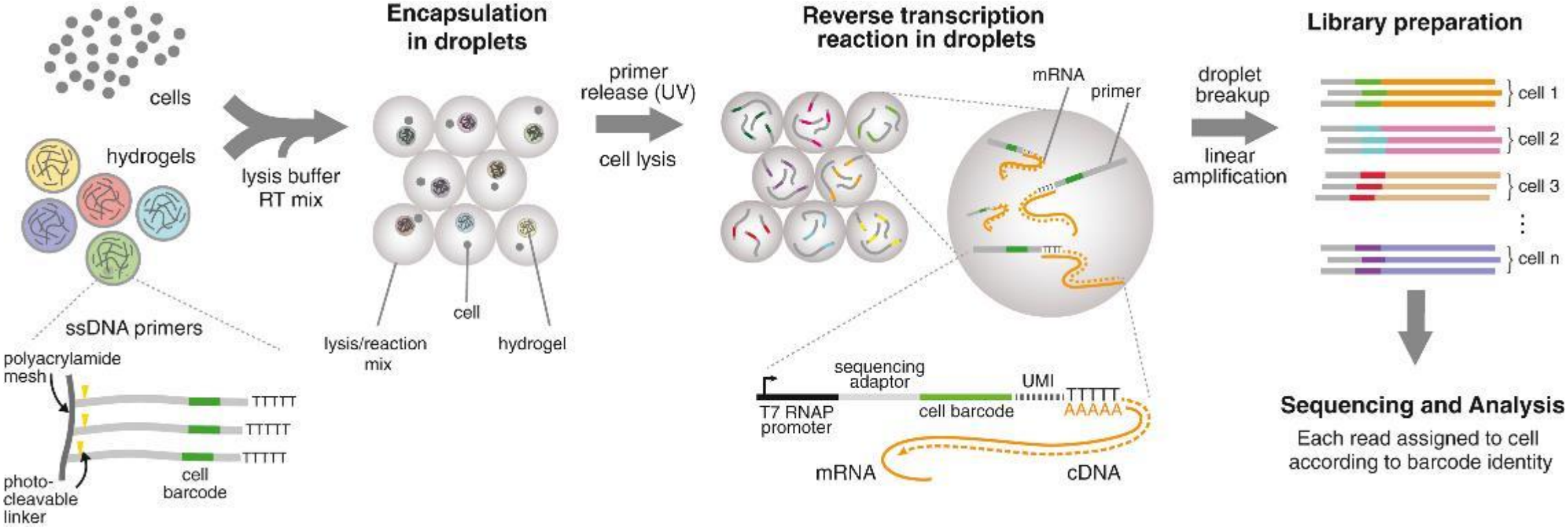


Head et al Biotechniques 2014

Single Cell RNA-seq

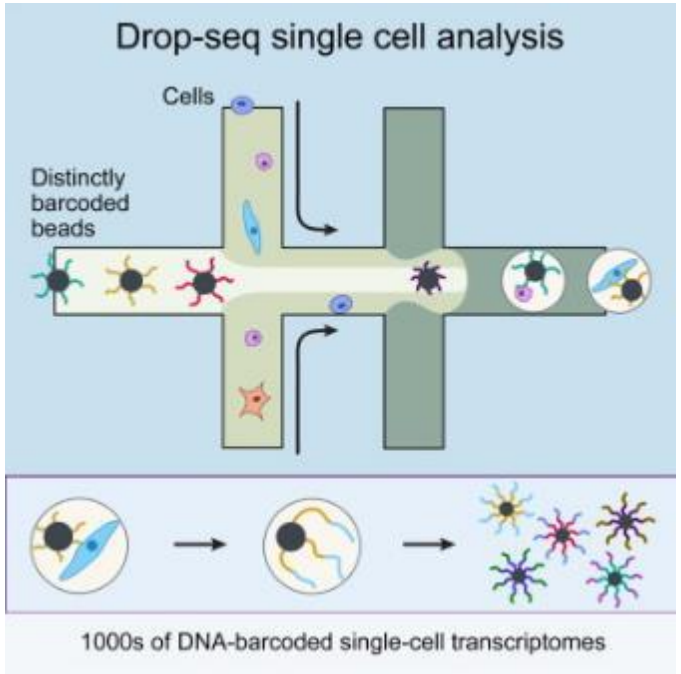


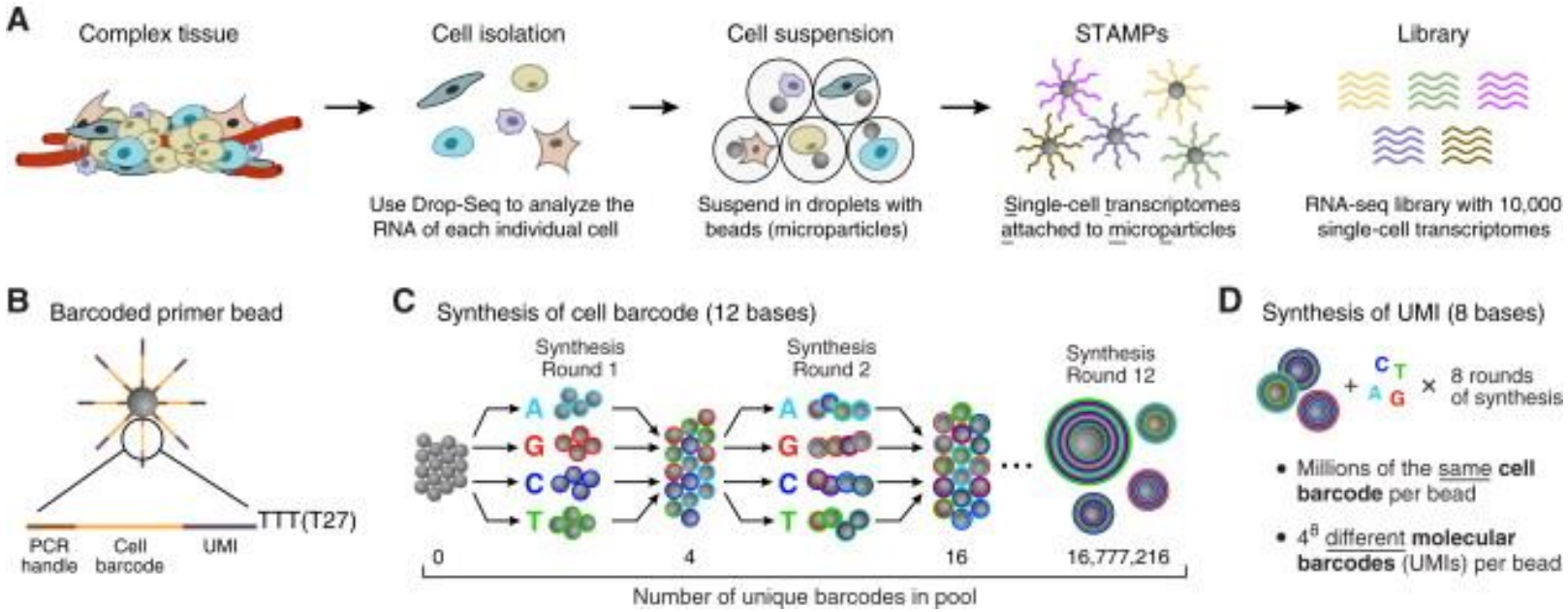
inDrop



Klein et al Cell 2015

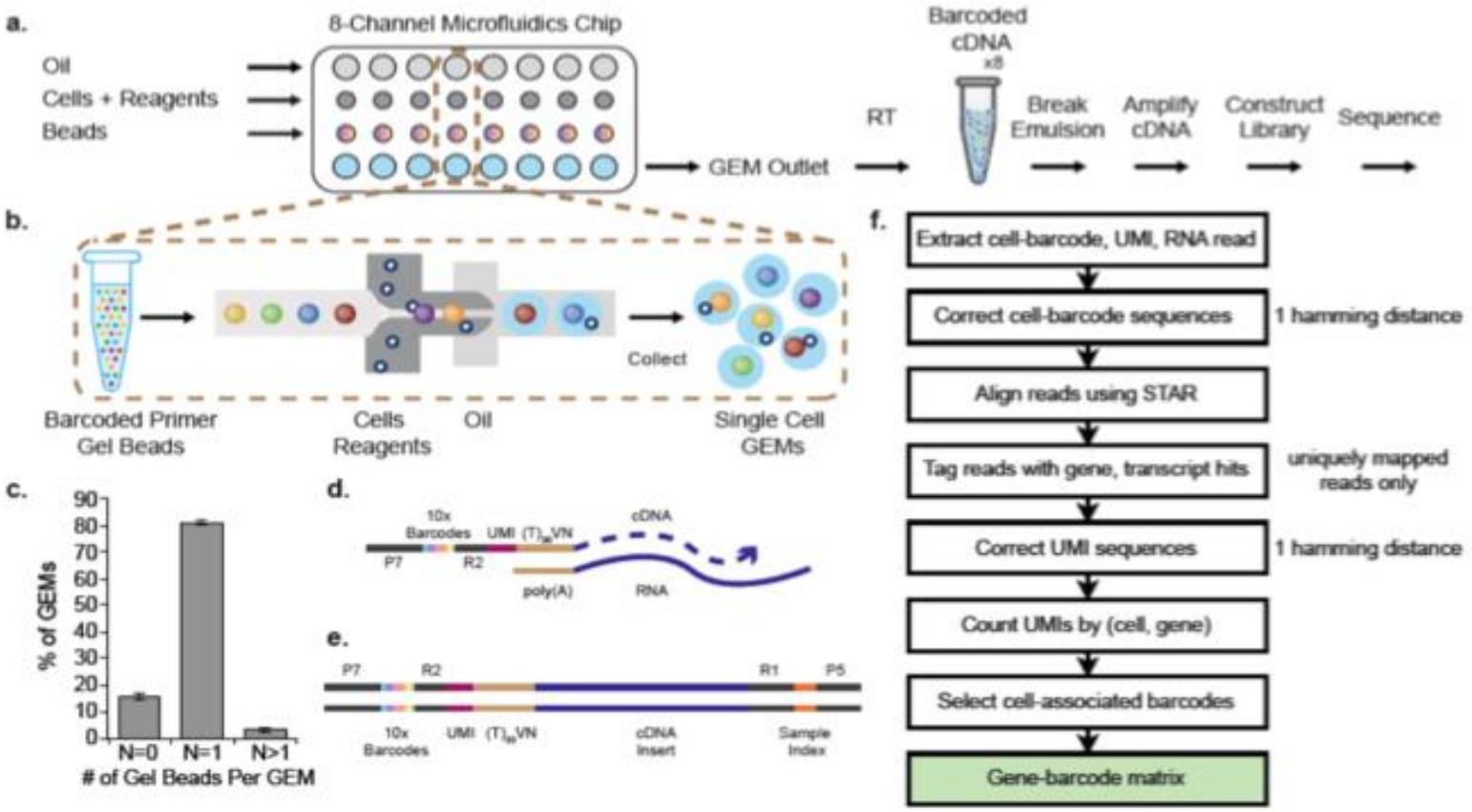
Drop-seq





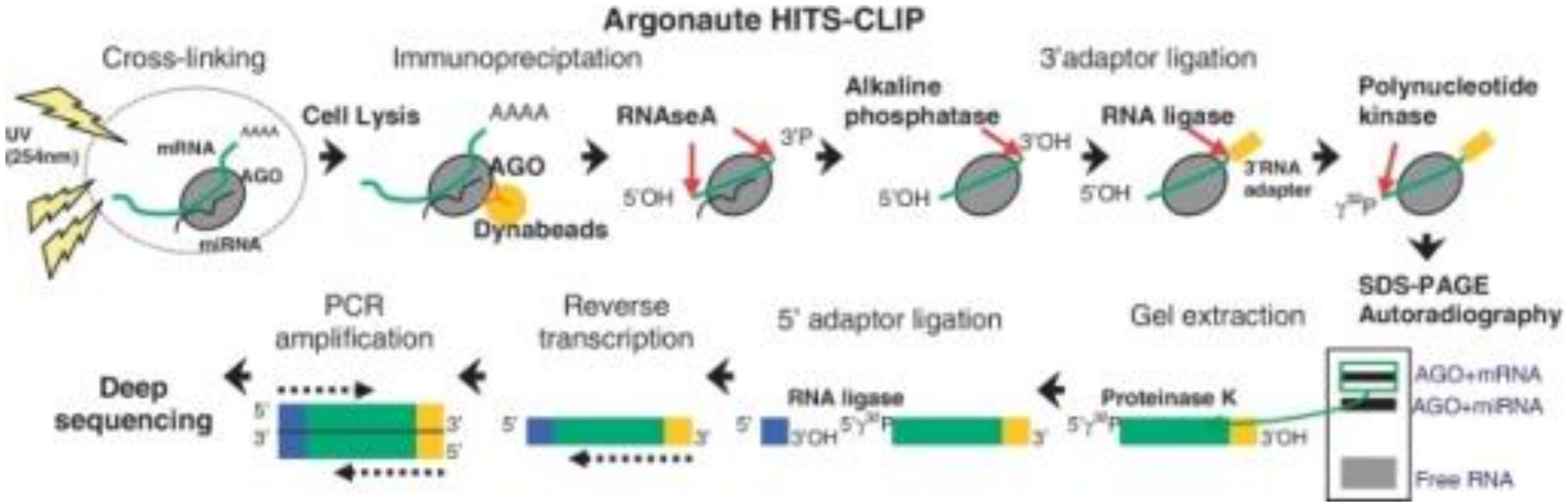
Macosko et al., Cell, 2015

10x Genomics Single Cell



10x Genomics

HITS-CLIP

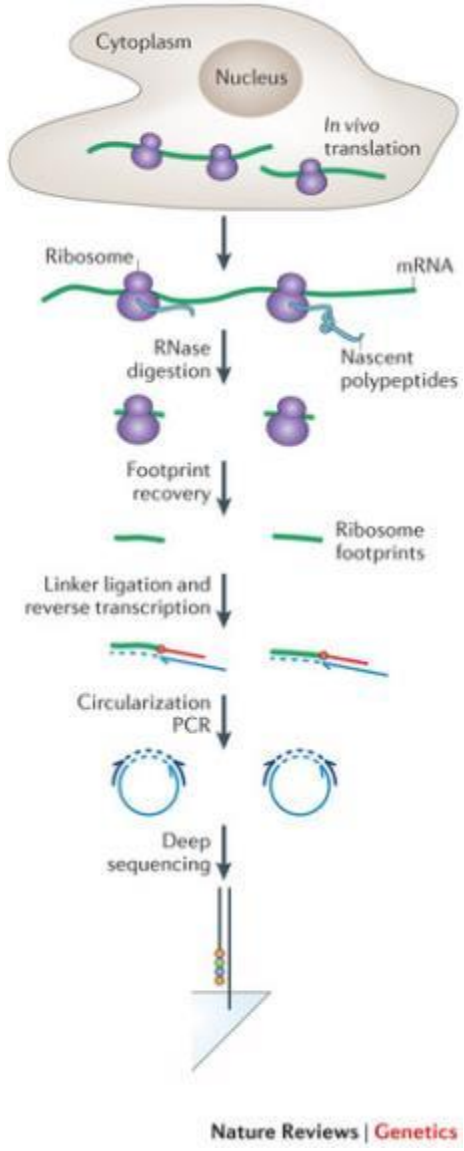


Thomson et al Nucleic Acids Research 2011

CLIP-seq Approaches

- HITS-CLIP: UV crosslinking + IP
- PAR-CLIP: photoreactive ribonucleoside + UV crosslink + IP
- iCLIP: 3’ exonuclease to crosslink

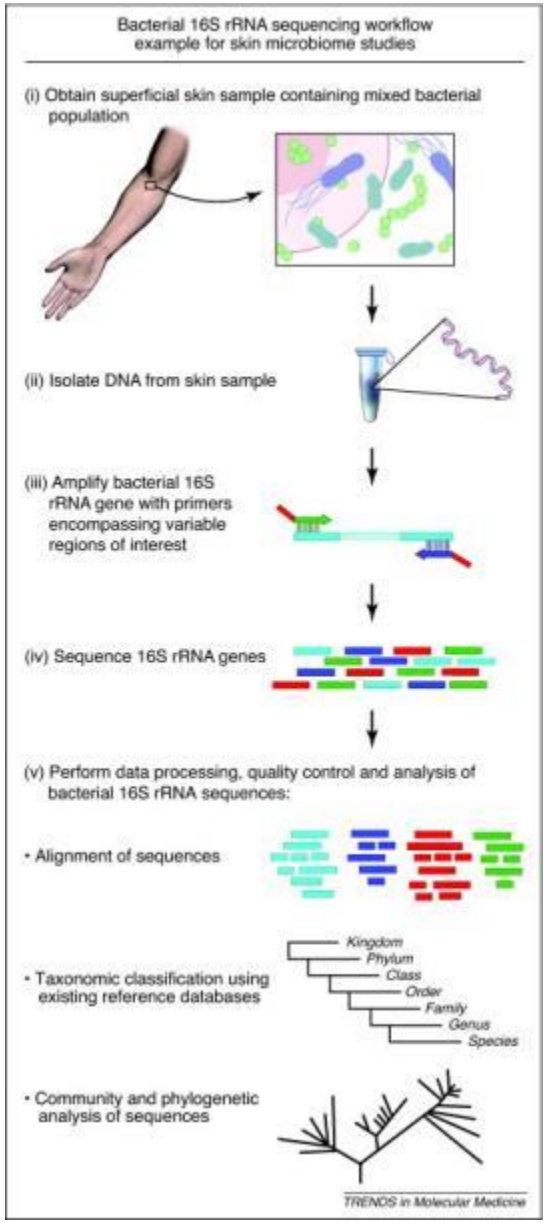
Ribo-seq



Ingolia, Nature 2014

16s Amplicon Sequencing

- Microbiome: study phylogeny and taxonomy
- Based on rRNA
- Ideal for MiSeq



Kong Science 2011

Library Prep

Service vs DIY Approach

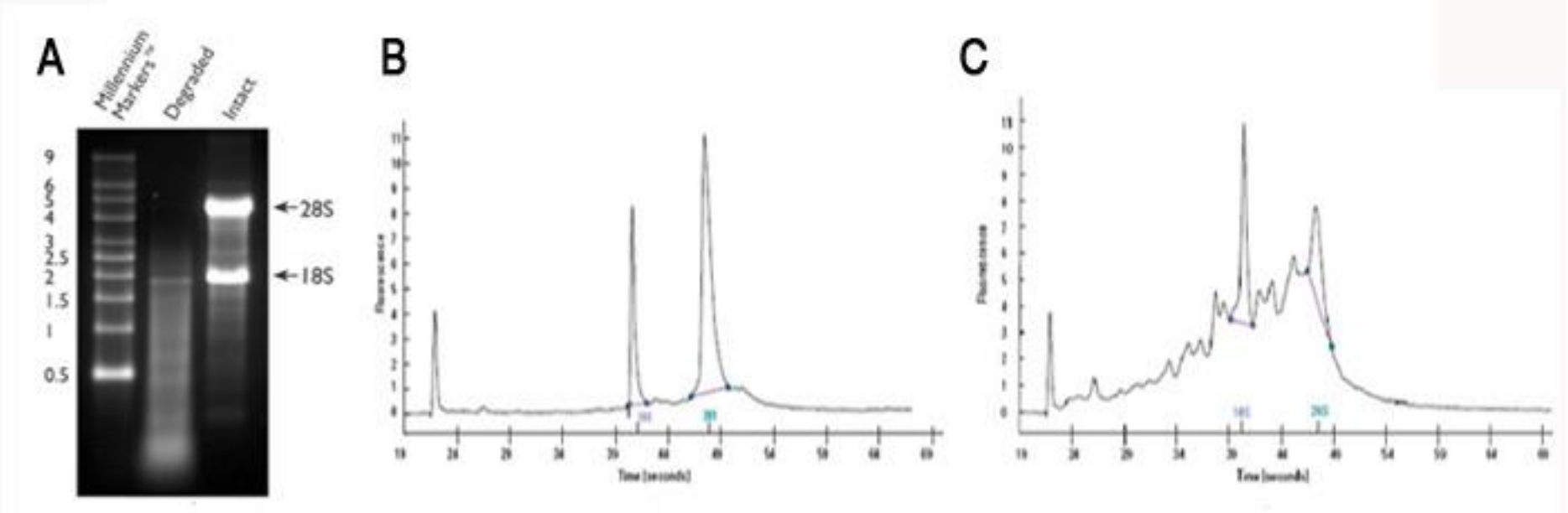
Library Prep: Biopolymers

- Bring DNA or RNA
- Apollo Wafergen 324 Robot
- Covaris S2
- Hamilton Star Plus Robot
- MJ Research Tetrad DNA Engine Thermal Cycler
- Qiagen Qiagility Robot

Library Prep: Isolation

- Mechanical
- Organic
- Solid-phase
- QC check: TapeStation, BioAnalyzer, Qubit

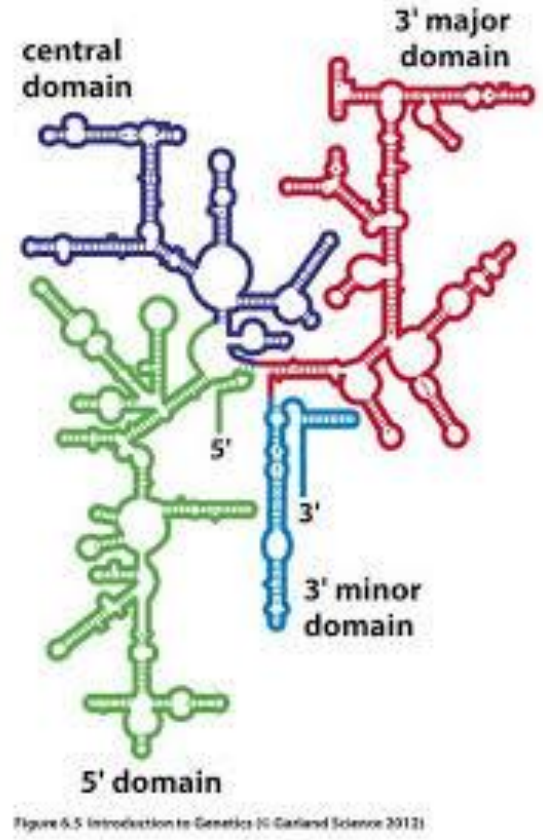
Library Prep: RNA QC



RNA-seqlopedia

RNA Target Enrichment

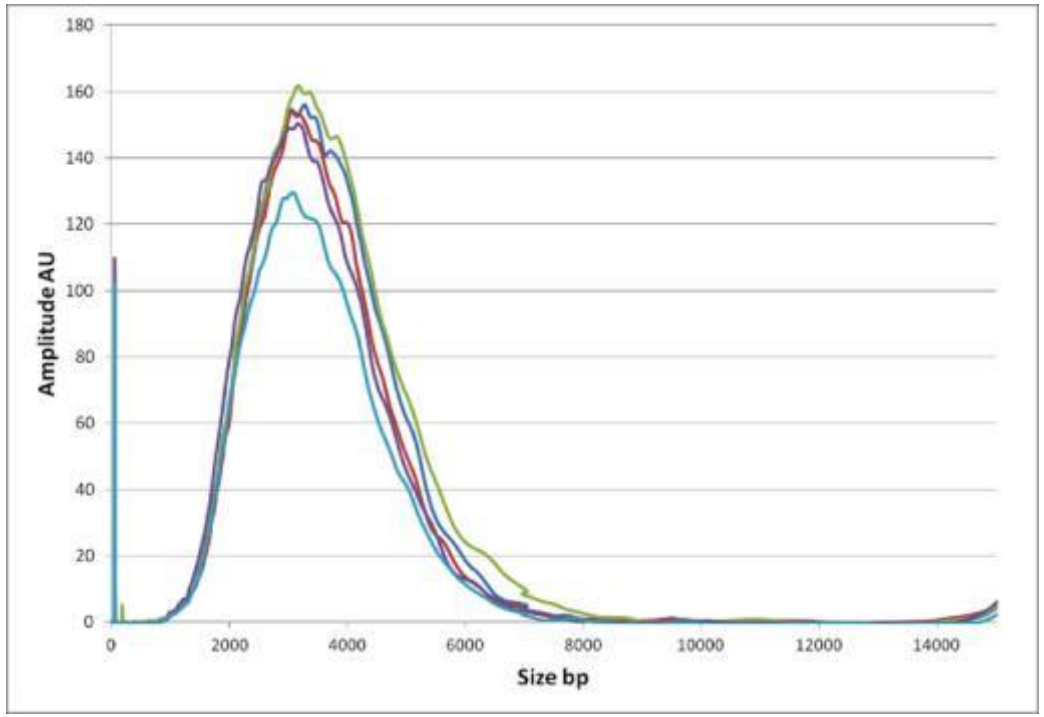
- Get rid of rRNA!
- oligo-DT beads (pull down poly-A tail RNA only
- rRNA depletion by hybridization
- Ribominus, Ribo-Zero, GeneRead kits



www.mun.ca

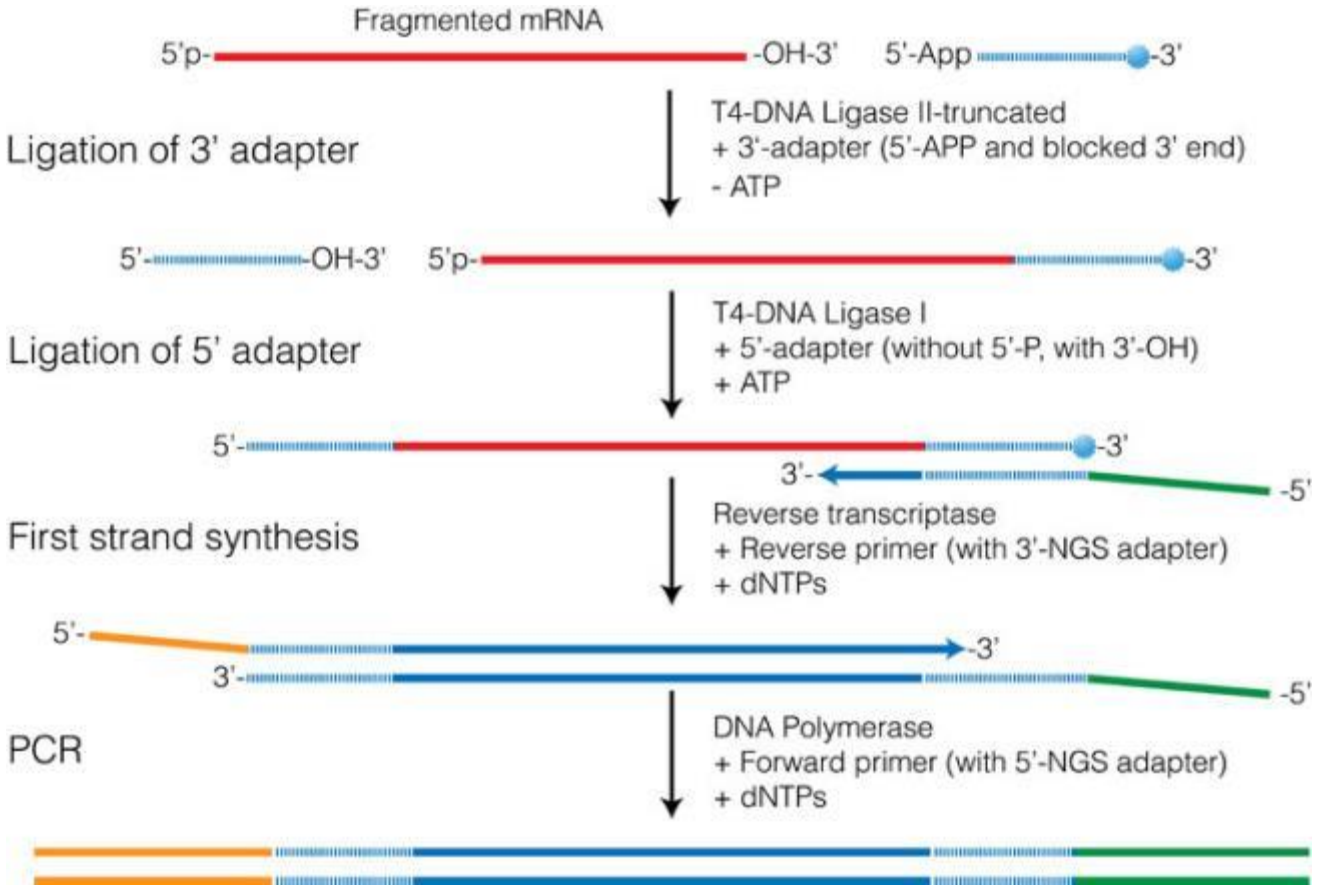
Fragmentation: DIY

- Covaris hydroshearing (available at BioPolymers): uniform distribution
- Heat
- Ribonuclease
- Sonication



Covaris

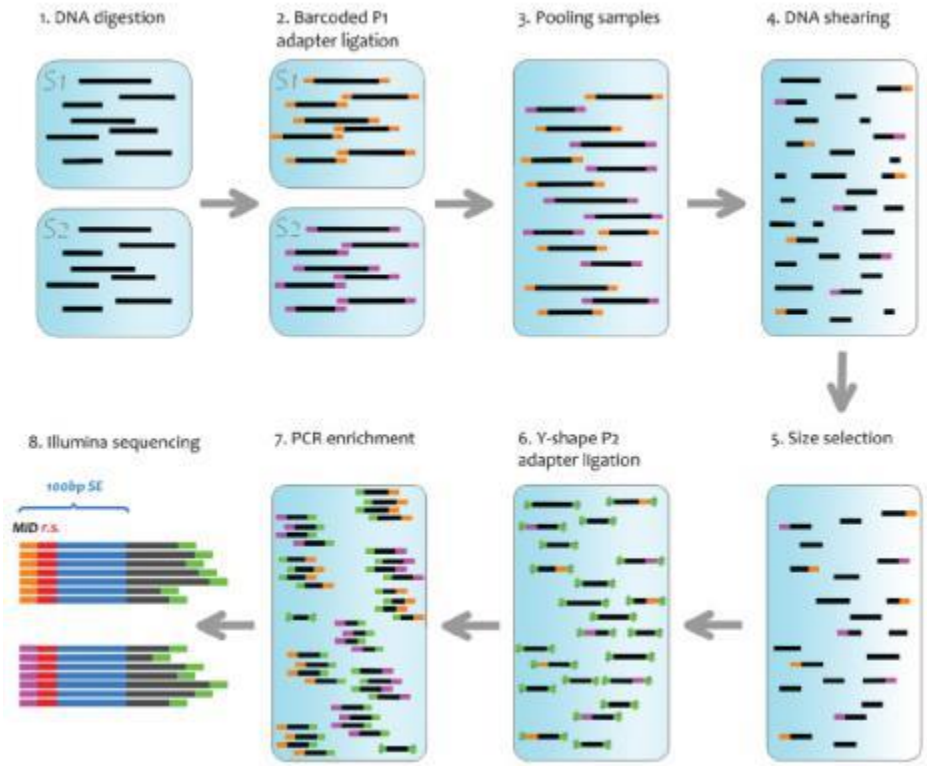
RNA-seq library prep



RNA-seqlopedia

Multiplexing

- Run more than 1 sample per lane in a flowcell
- Attach barcodes with unique sequence IDs
- Separate .fastq files created for each barcode
- Purchase sets from Biopolymers



IDT DNA

Analysis

Analysis Options

- HMS RC/O2 HPC environment
  - User Training courses
  - Consulting on individual experiments, from design to analysis
  - DIY
  - Pipelines
  - Free!

- HCBC:
  - User Training courses (fee)
  - Consult (fee), comprehensive analysis

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Galaxy

- Graphical, web-based tool to analyze NGS
- Front-end for popular tools like “Tuxedo” family
- Create own cloud instance or use public servers
- Limited in how much data can be uploaded
- Not scalable

High Performance Computing for NGS

- Spread computation over multiple cores with a large amount of allocated memory
- Long runtimes
- Large storage allocations
- Some algorithms are linux-specific builds
- Allows maximum customization of options
- Automation of workflows
- “Set it & forget it”

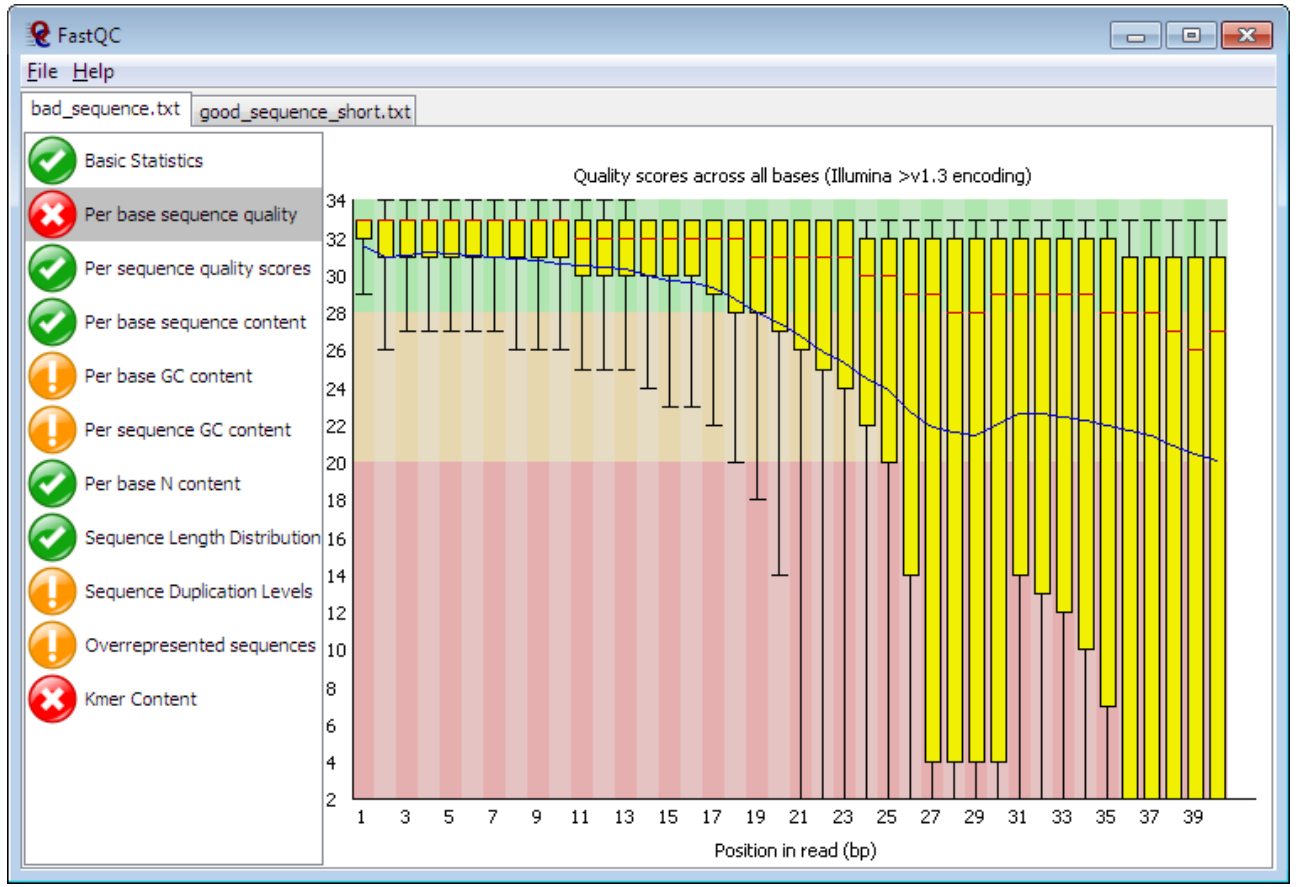
Analysis: Getting Started

Quality Control, Trimming

Quality Report: FastQC

- Check the quality of sequence, identify issues
- Quality score of bases along read length
- Presence of barcode, adapter, repetitive sequence, kmers
- GC content

FastQC: Poor Sequence



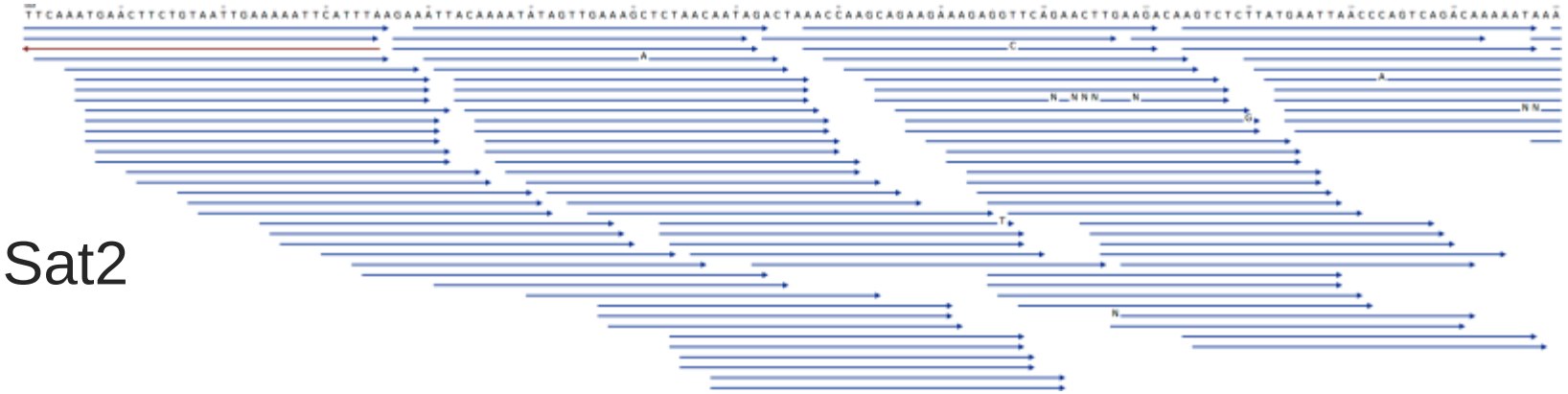
Trimming: Adapter/Barcode Removal

- Sequences won't align well with too much of these!
- Dynamic (based on sequence/quality) or blunt (remove X from 5', Y from 3')
- O2 Options:
  - Cutadapt
  - Trimmomatic
- PCR Duplicates
  - Picard MarkDuplicates
  - Samtools rmdup

Alignment

Aligners

- Create BAM/SAM alignment file
- bwa mem
- blat
- Bowtie1/2
- Tophat2->HiSat2
- Novoalign
- STAR
- Kallisto



seqan.readthedocs.org

What is an alignment index file?

- Algorithm-specific way to parse a genome
- Created from a .fasta file of the genome or transcriptome
- O2: /n/groups/shared_databases
  - BWA
  - Blat
  - Bowtie1, 2
  - Hisat2
  - Novoalign
  - STAR
  - RSEM
  - kallisto

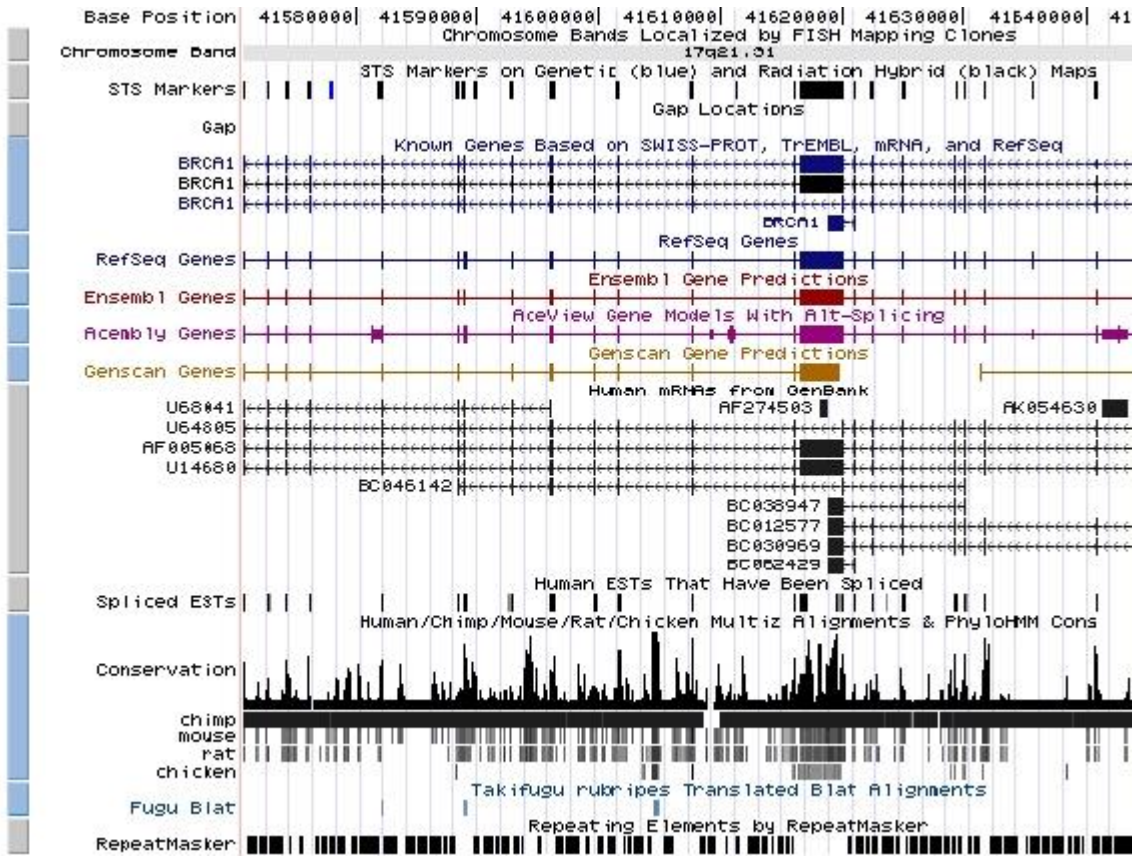
Alignment Considerations

- Number of substitutions/deletions/additions
- Gap length
- Quality
- Unique mapping of reads
- Maximum number of mappings
- Splicing/isoforms

Genome Visualization: IGV

Broad – IGV

Genome Visualization: UCSC



UCSC

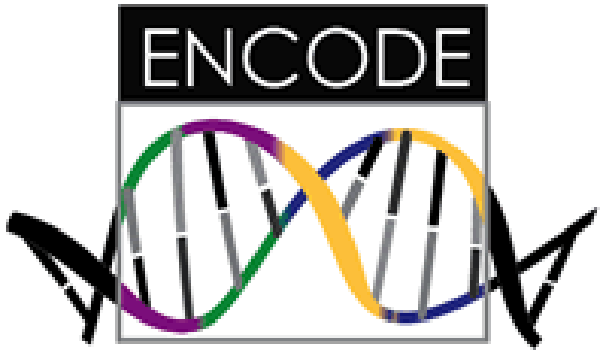
Analysis: After Alignment

DNA/Exome: Variant Callers/CNV

- Genome Analysis Tool Kit (GATK)
- VarScan2
- MuTect
- Breakdancer
- CONTRA
- CNVnator
- Annotate: ANNOVAR

Peak Callers : ChIP, ATAC

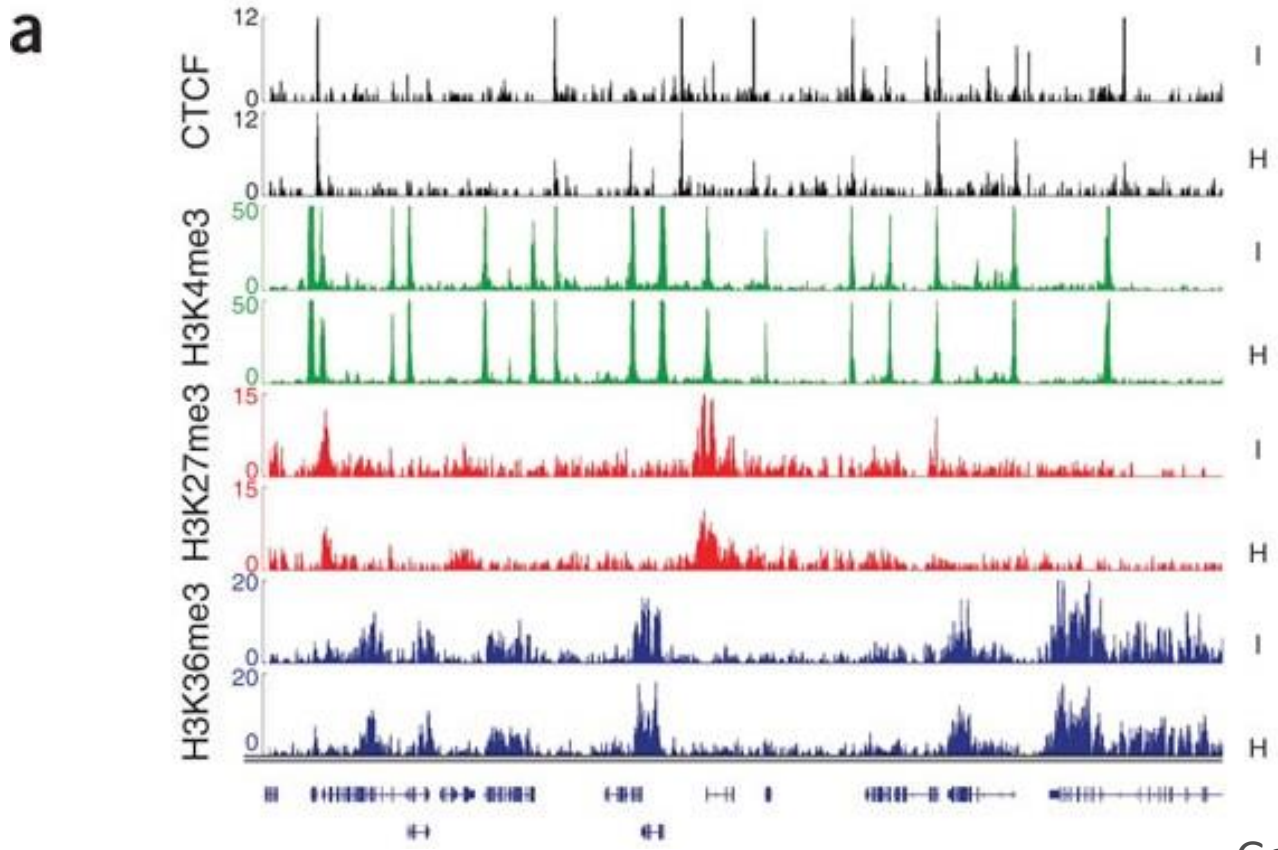
- SPP (R)
- GEM
- PeakSeq
- MACS2
- Differential: DiffBind (R)



Peak Callers: CLIP-seq

- PARalyzer
- dCLIP
- CIMS

Peak Visualization



Goren et al Nature Methods 2010

Motif Analysis

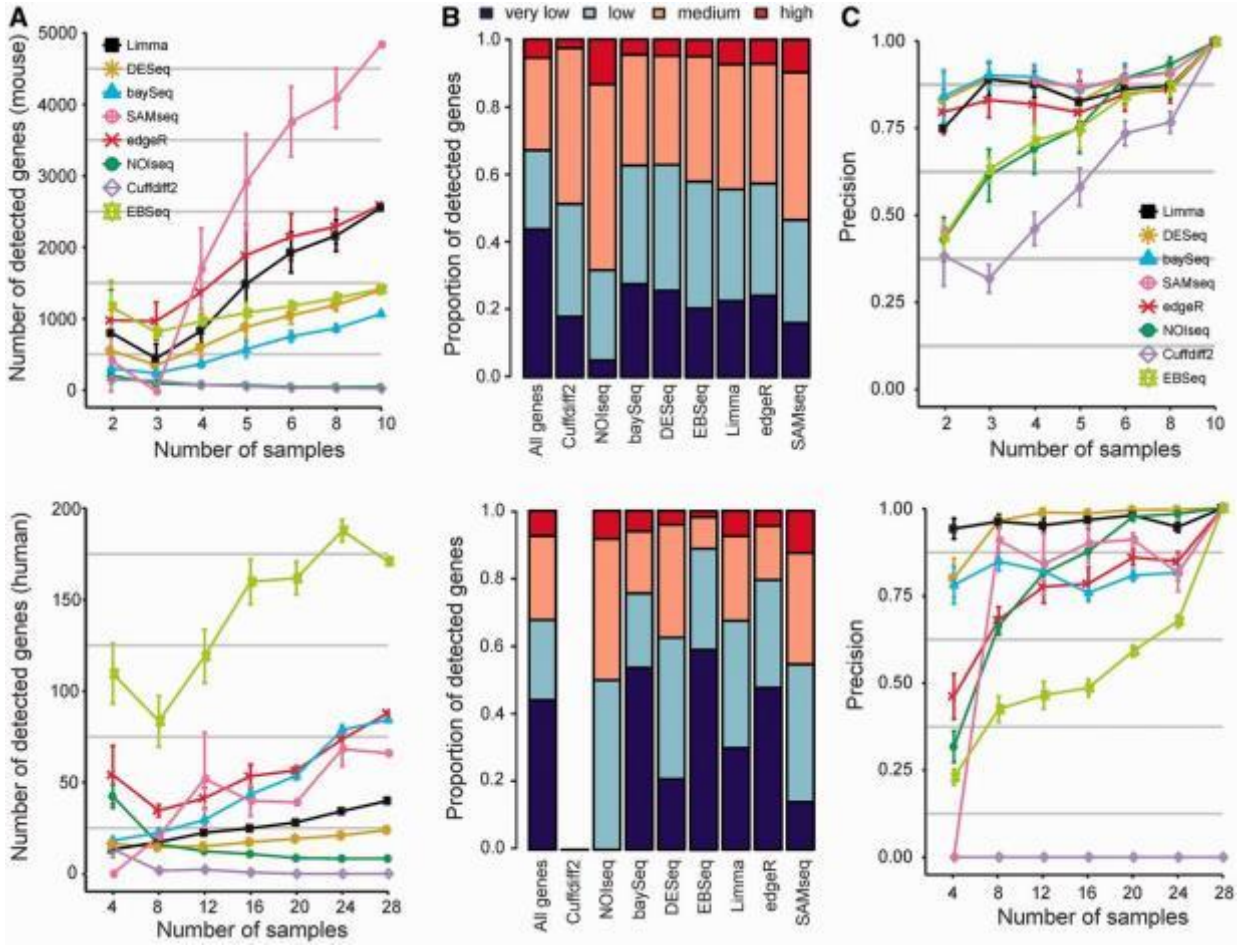
- HOMER
- MEME/MAST
- de Novo & Known Motifs

MEME
Multiple Em for Motif Elicitation
MAST
Motif Alignment & Search Tool

Differential Expression Analysis

- DESeq2 (R - counts)
- edgeR (R - counts)
- baySeq (R - counts)
- EBSeq (R - counts)
- CuffDiff (Tuxedo suite –RC Pipeline)
- RSEM (RC Pipeline)

Comparing DE algorithms



Seyednasrollah et al Briefings in Bioinformatics 2013

Single Cell

- scde (R)
- Seurat (R)
- Pagoda (R)
- MAST (R)
- Monocle (R)

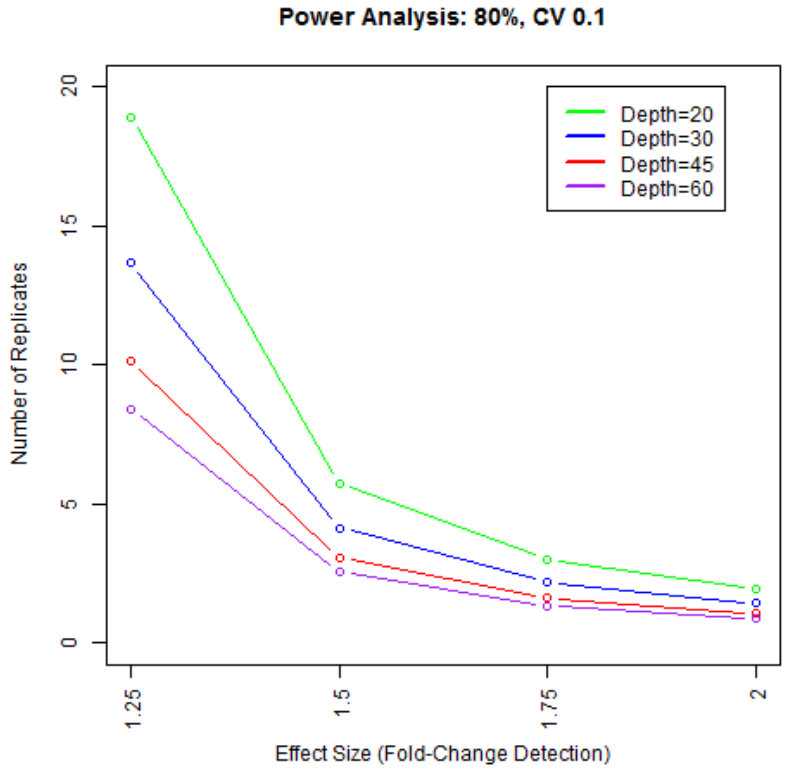
Functional Enrichment Analysis

- GOSeq (R)
  - Control for Gene Length
  - Query GO and KEGG
- Metacore (Countway)
  - Pathway, Drug-rich vocabulary
- Ingenuity (Countway)

Considerations

Power Calculation

- Number of replicates needed
- Sequencing depth
- Statistical power determines ability to draw conclusions (refute the Null Hypothesis of “no difference”)



Experimental Design

- Control variables
- Biological Replicates
- Cell prep: treatments & days matter
- Mice: age, sex, isolate location, date of isolation, date of library prep
- Talk to RC/HCBC: one conversation can save $$$ & headache!

Data Deposition

- GEO (Gene Expression Omnibus)
- Upload as SRA
- Funding source may require data deposit

Bild et al PLOS Biology 2014

- Don’t be a jailer!

For further questions

- http://rc.hms.harvard.edu
- rchelp@hms.harvard.edu
- Office Hours: Wed 1-3p Gordon Hall 500