Introduction to Next-Generation Sequencing Technologies

Kris Holton

HMS Research Computing

Spring 2016

HMS Research Computing

- Manage Orchestra High Performance Compute Cluster
- Research Computing Consultants
  - Planning experiments
  - Analysis
  - Scaling/scripting
- User Training
  - HPC/Linux
  - R/Python/Perl/Matlab
  - NGS
  - Biostatistics

rchelp@hms.harvard.edu

Topics for today

- Sequencers + Technology
    - HiSeq/MiSeq/NextSeq/IonTorrent/EdgeSeq/Fluidigm
- NGS Branches
    - DNA/ChIP/Exome/RNA/miRNA/SingleCell/CLIP/Ribo/16s
- Library Prep
- Analysis
    - Options
    - File Formats
    - Software
- Experimental Design
- Data Deposition

Sequencing Core

- Two Illumina cBot stations
- Two Illumina HiSeq 2500 sequencers
- Two Illumina MiSeq sequencers
- One Illumina NextSeq 500 sequencer
- Single-cell: Fluidigm C1
- HTG Edge-Seq
- 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

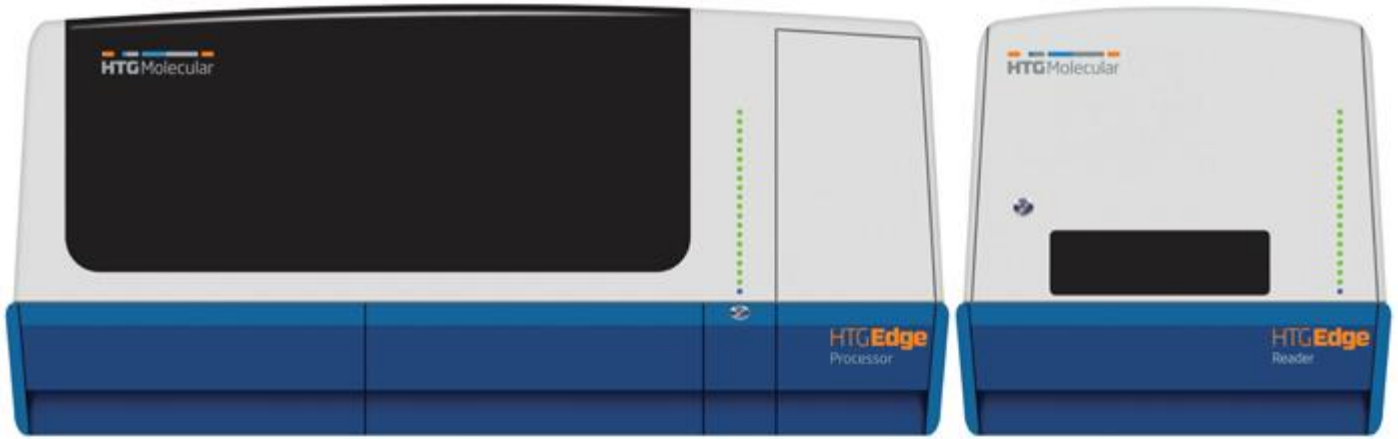
- Video!

Ion Torrent

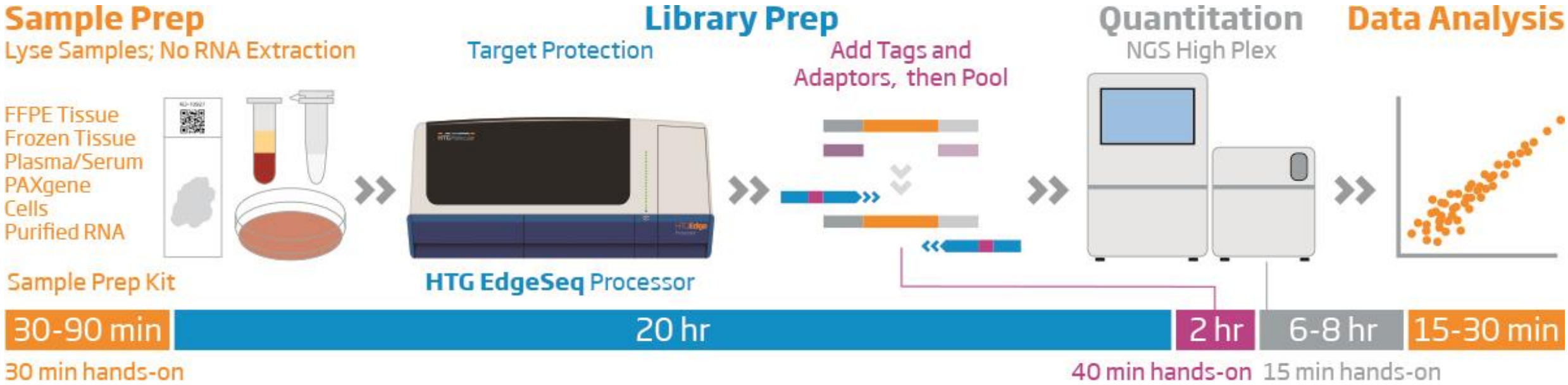
- Semiconductor chip
- Adding dNTP: release pyrophosphate + H^+
- Add single nucleotide, measure proton release
- 400 base read length

HTG EdgeSeq

- Extraction-free chemistry
- miRNA, mRNA, fusions, DNA
- FFPE, plasma



HTG EdgeSeq: Limited Sample



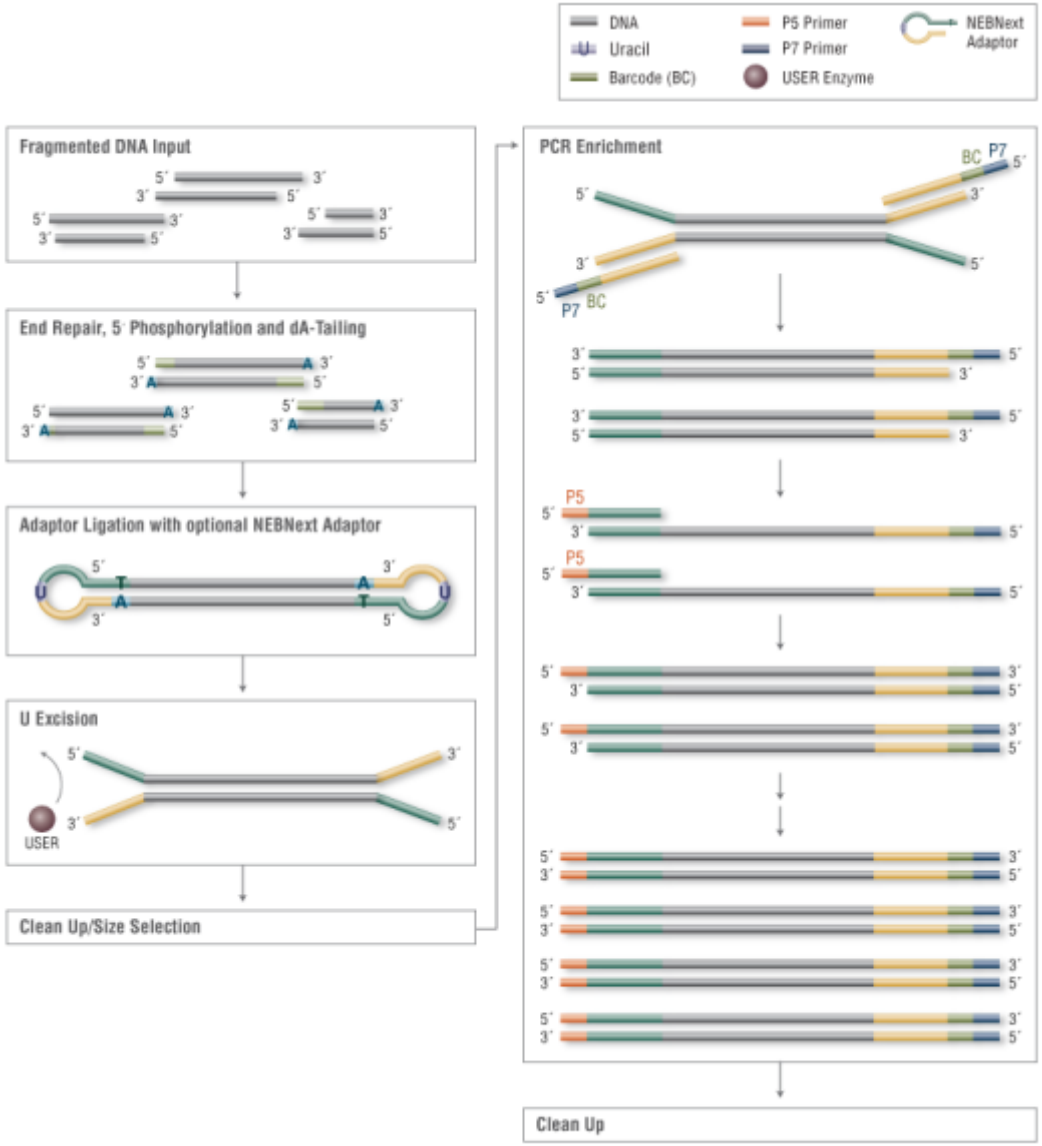
HTG

Fluidigm C1

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

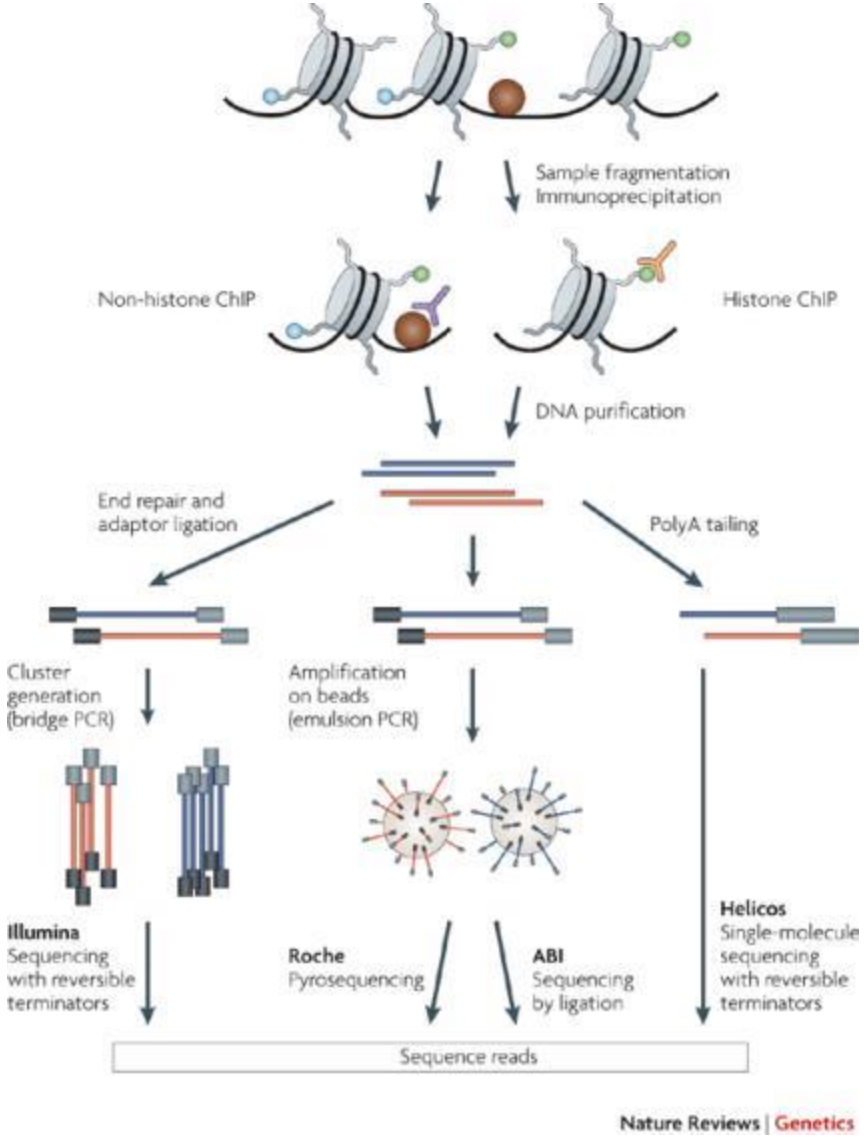
NGS Technology Variations

DNA-seq



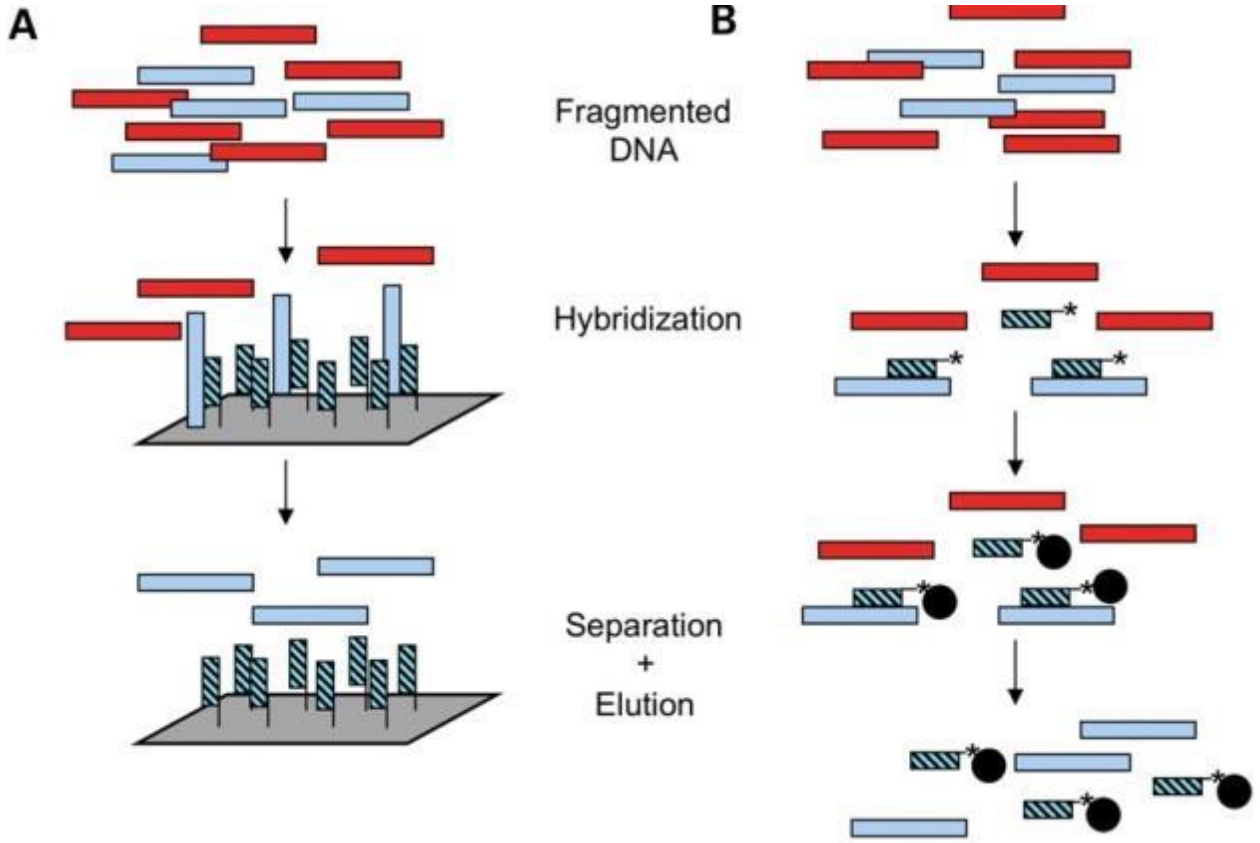
New England Biolabs

ChIP-seq



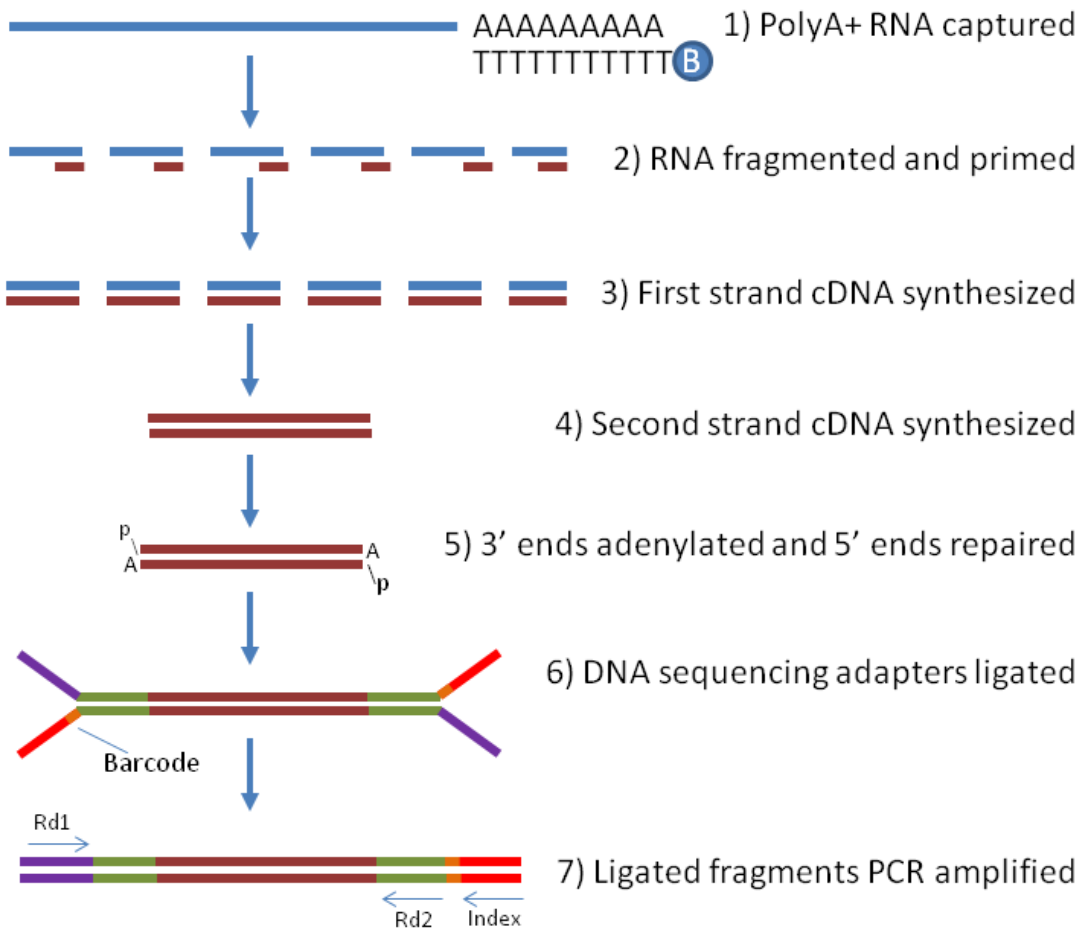
Peter J. Park, Nature 2009

Exome Sequencing - Capture



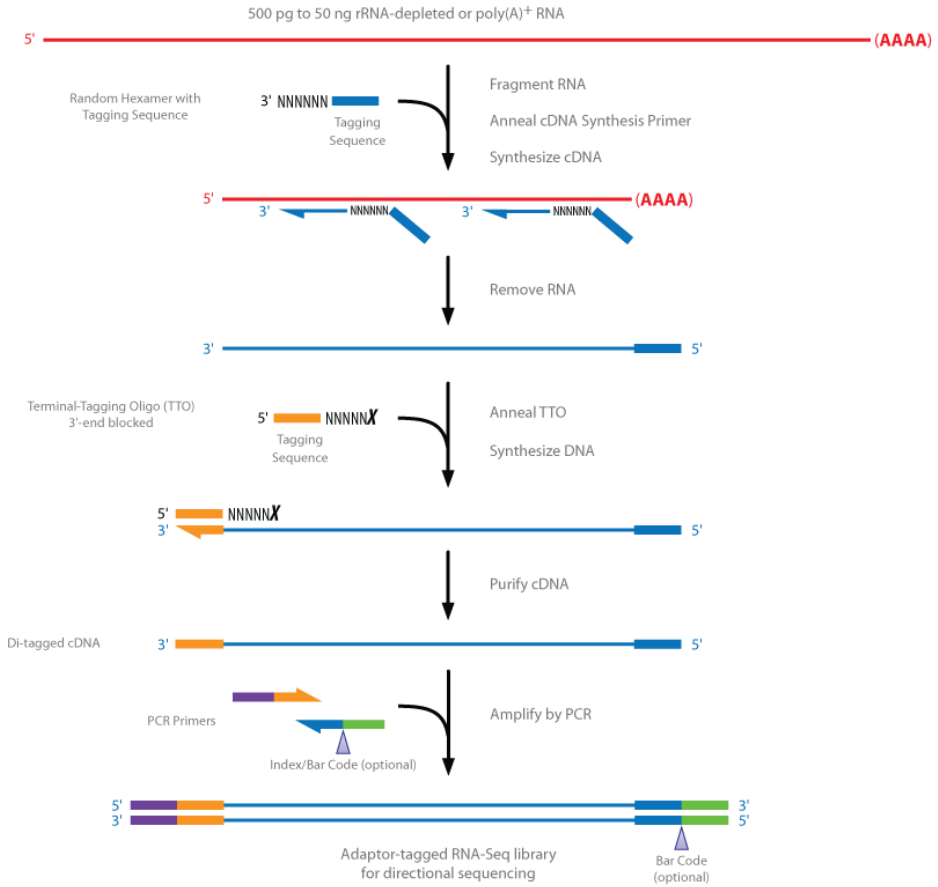
Teer & Mullikin, Human Molecular Genetics 2010

RNA-seq



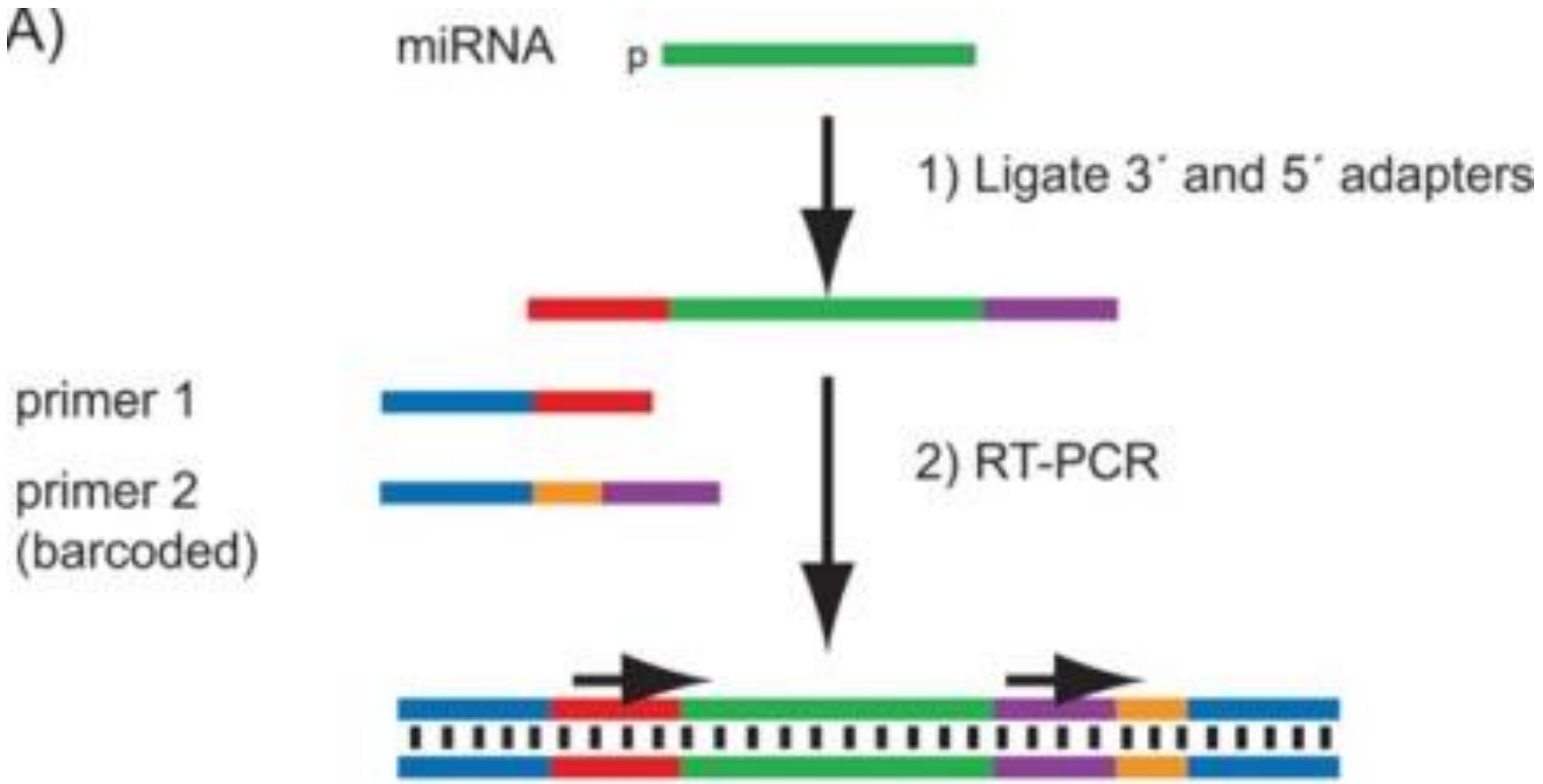
Labome

RNA-seq: strand-specific



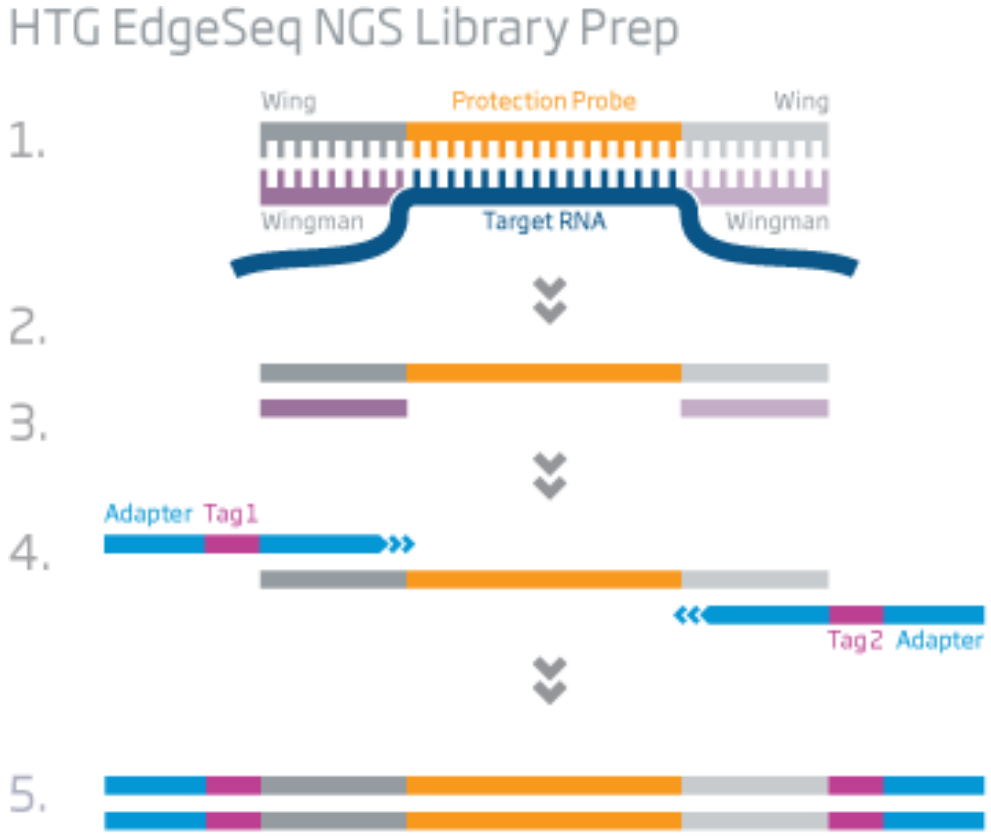
Illumina

miRNA-seq



Head et al Biotechniques 2014

HTG EdgeSeq Library Prep



1. Protection probes with wings hybridize to wingmen and target RNA

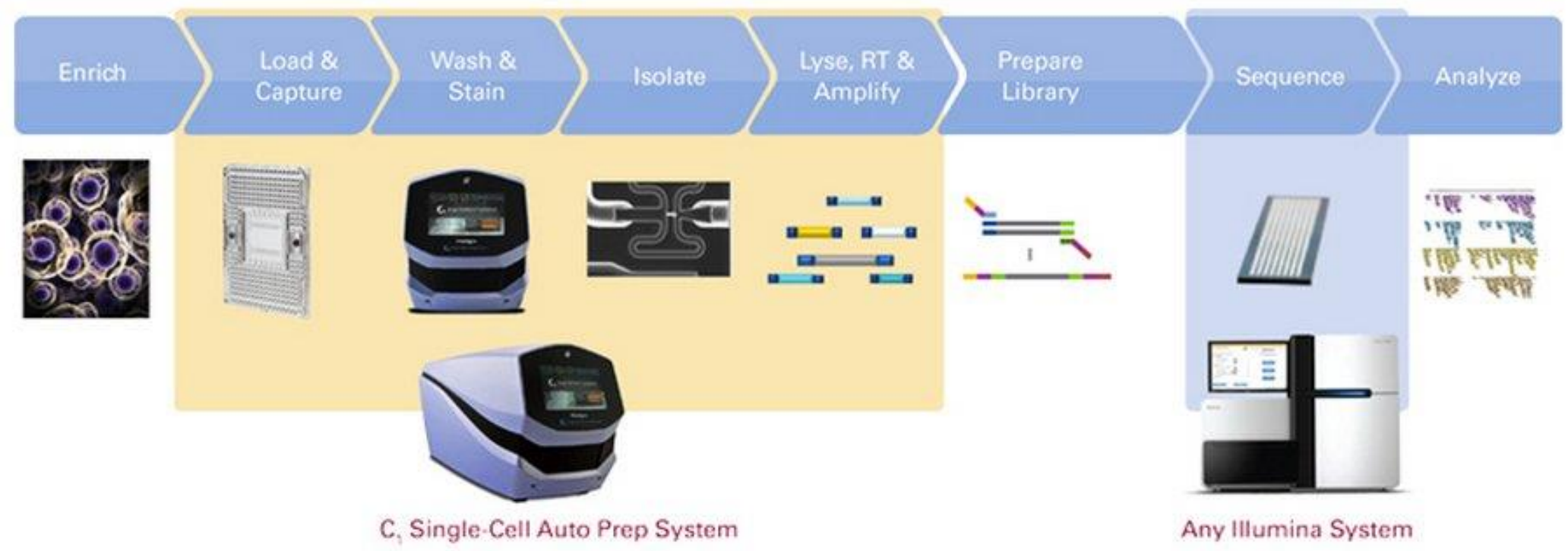
2. S1 nuclease digests unbound RNA and probes

3. Base hydrolysis eliminates target RNA

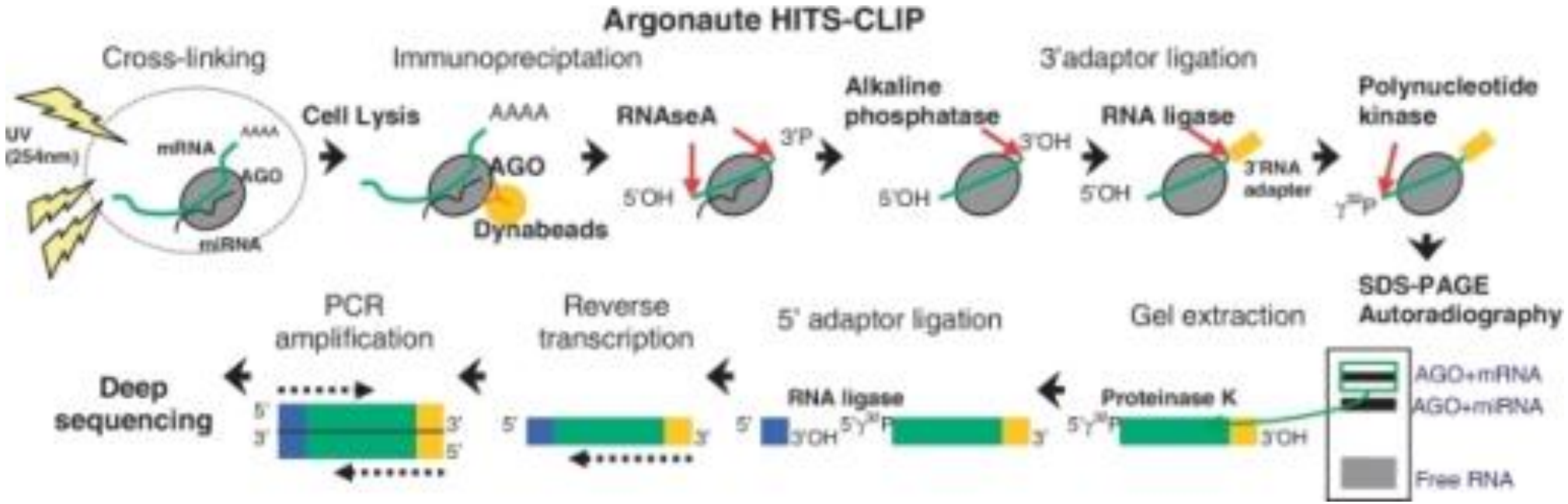
4. PCR step adds adapters and tags

5. Sequence using any NGS system

Single Cell RNA-seq



HITS-CLIP

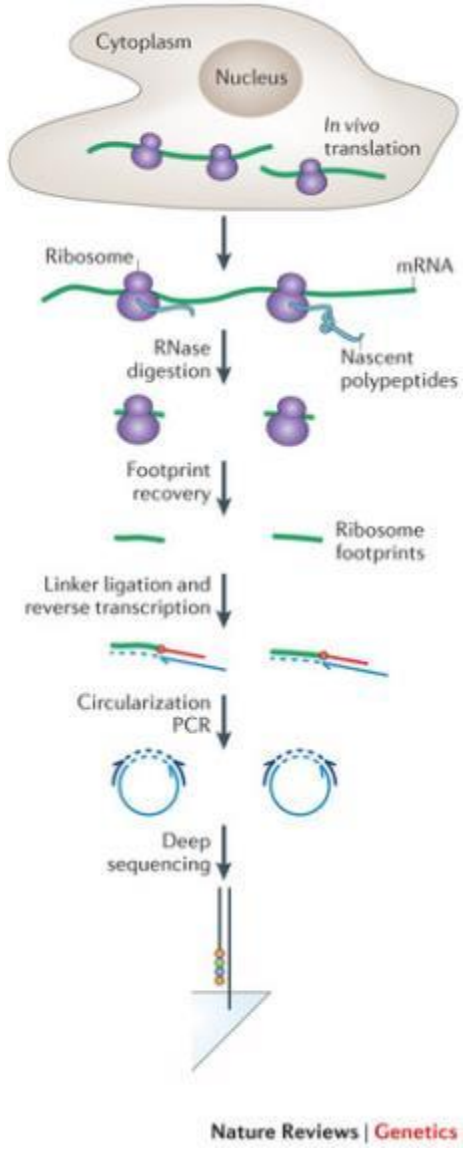


Thomson et al Nucleic Acids Research 2011

CLIP-seq Approaches

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

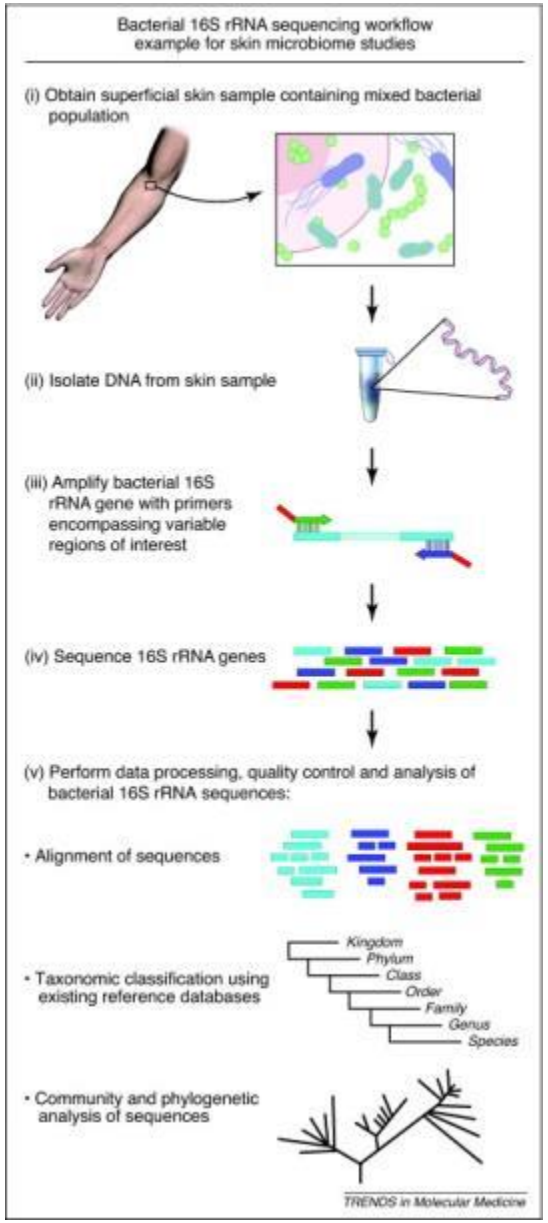
Ribo-seq



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

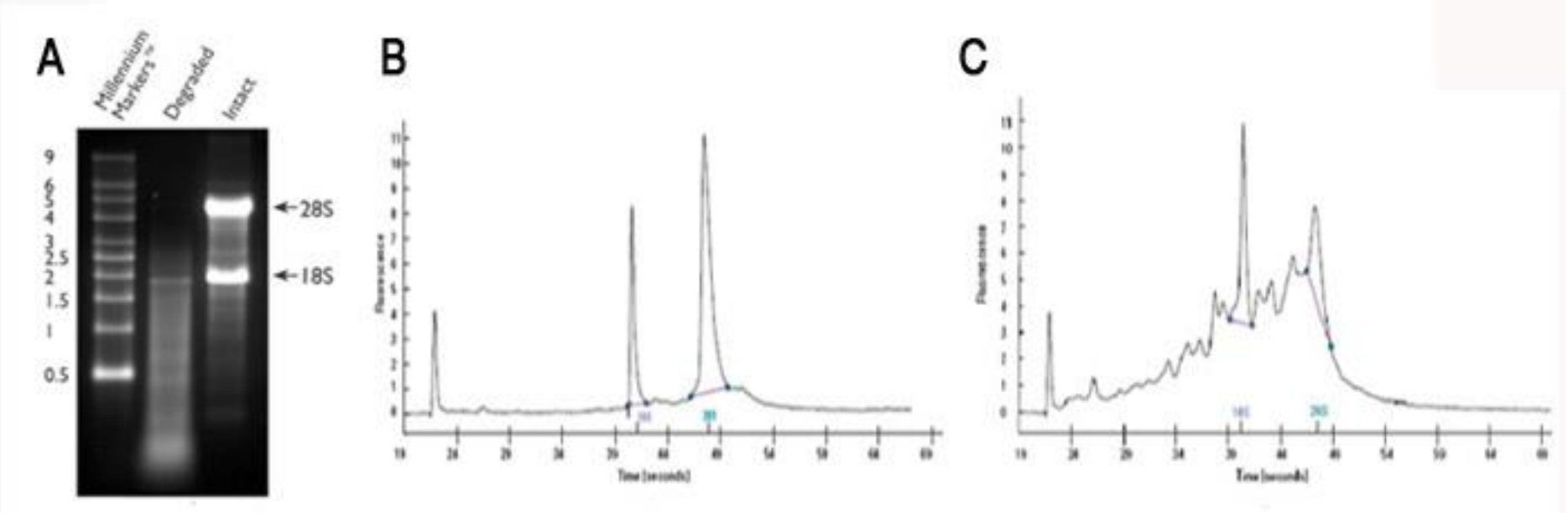
Library Prep: Biopolymers

- 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

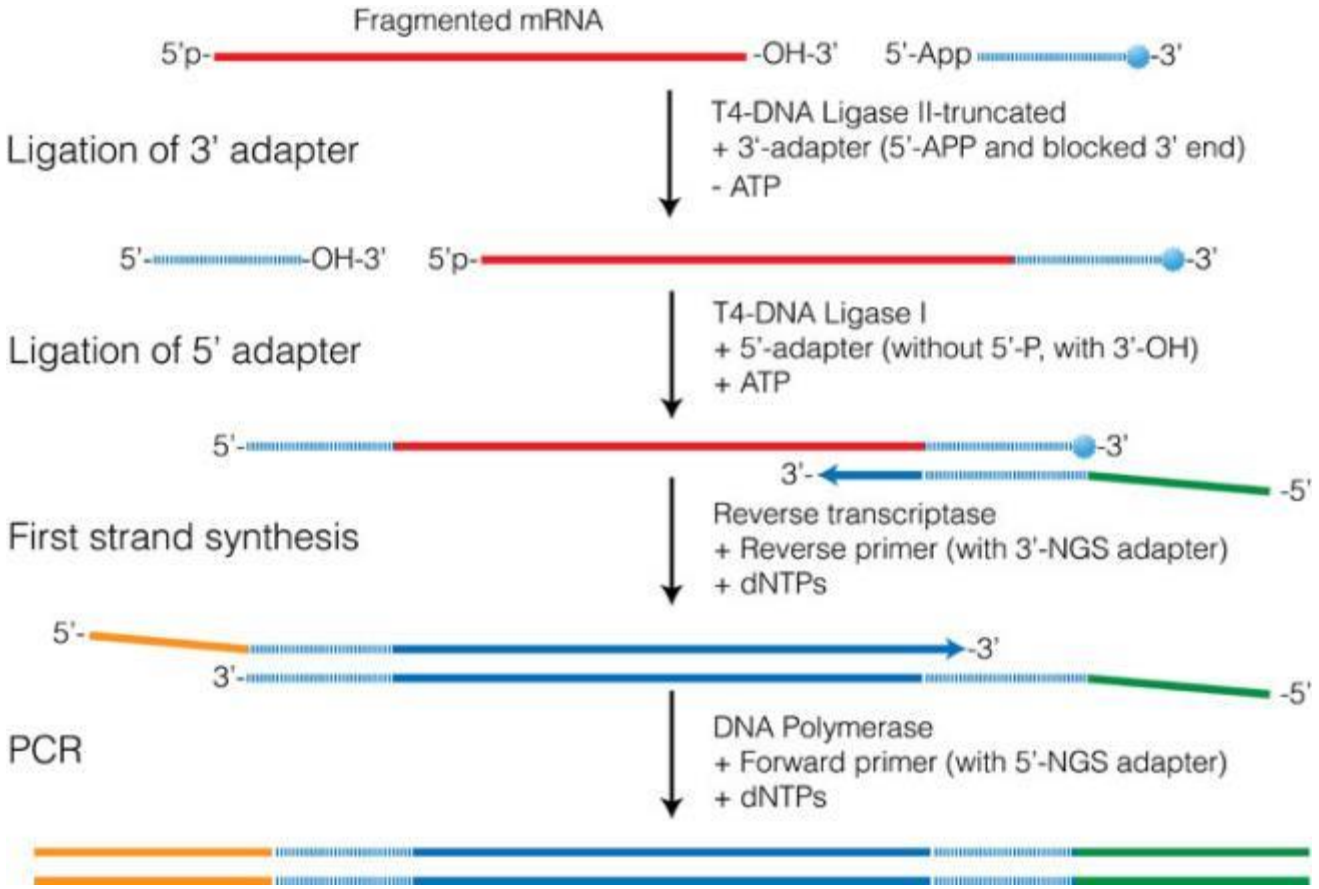
RNA Target Enrichment

- Get rid of rRNA!
- oligo-DT
- rRNA depletion by hybridization: Ribominus, Ribo-Zero, GeneRead

Fragmentation: DIY

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

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

Coverage, Power, Sample Size

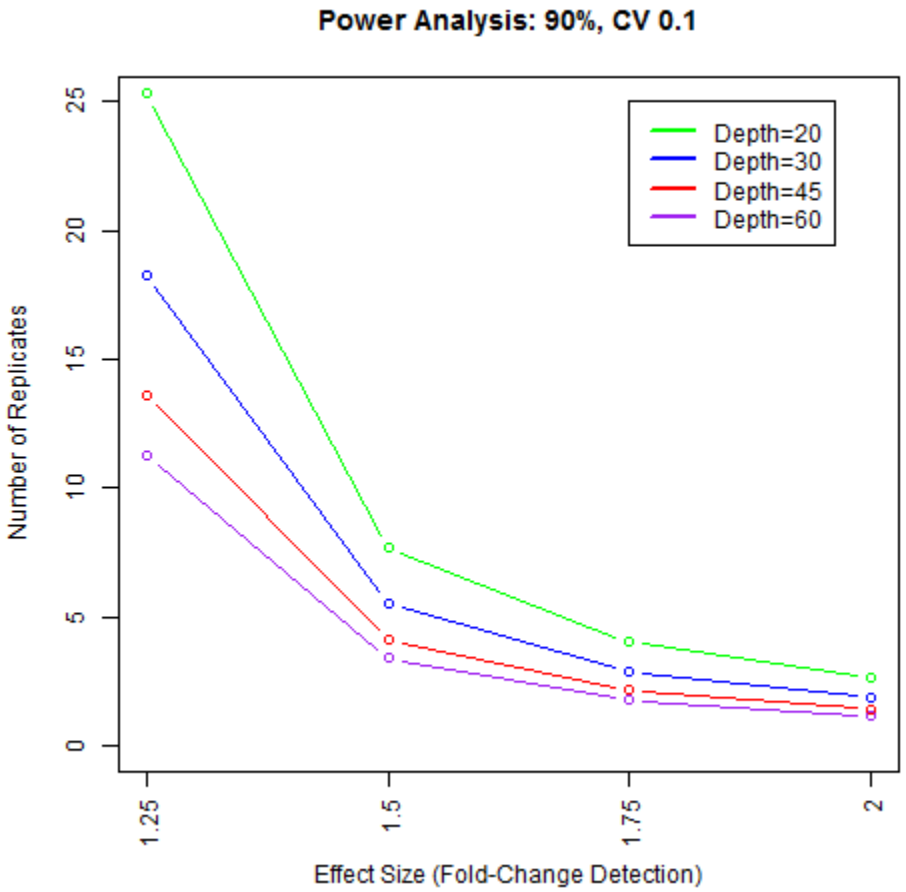
Coverage (Depth)

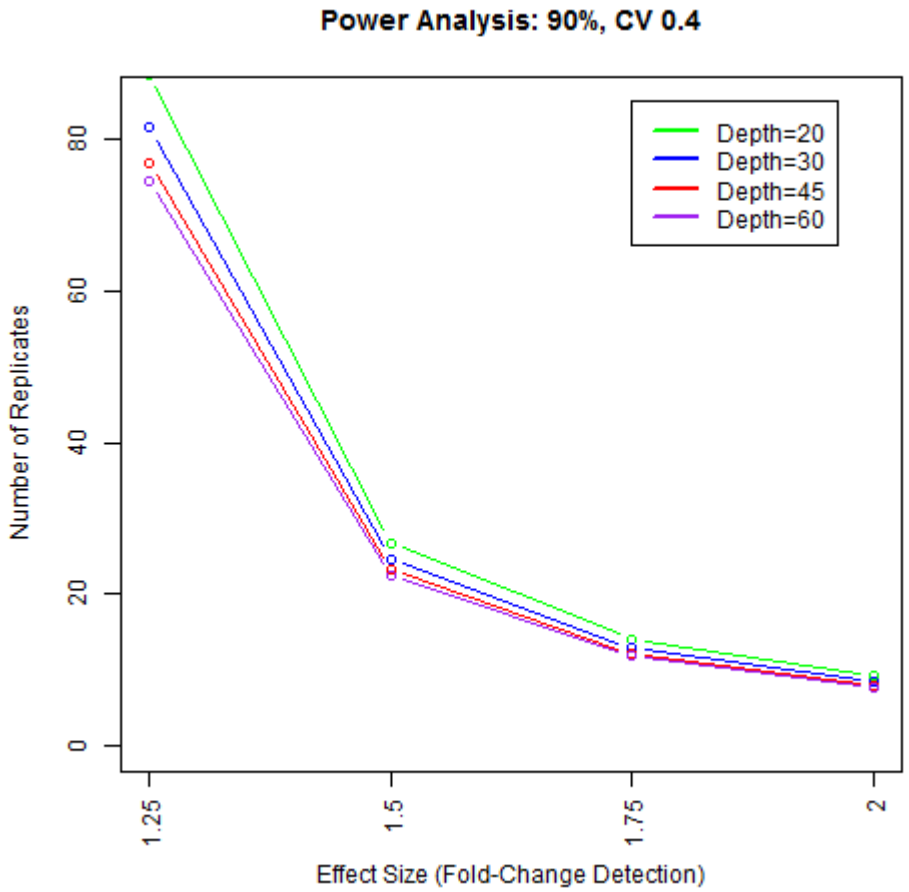
- How many times a position in the genome is sequenced
- Coverage = (Read Length * Number of Reads) / Size of Genome
- C=LN/G
- Toy example: 10X coverage, 100bp PE, Human (3 * 10^9): how many reads?
- 10 = 100 (x) / 3*10^9
- 300 million reads = 150 million read pairs

Power Analysis

- Power: sensitivity
- Coverage (and reads required) to achieve statistical significance rely on:
  - Number of Biological Replicates (per condition)
  - Effect Size trying to detect
  - Coefficient of Variation of Replicates

Power Analysis: 90%





Analysis

Analysis Options

- HMS RC/Orchestra 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

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”

Nomenclature

What is a FASTQ File?

- FASTA with Quality: Reads!
- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Wikipedia

What is a Phred Score?

- ASCII encoding of the quality of each base position
- Sanger: score 0-93 using ASCII 33-126
- Solexa/Illumina score 1.0: -5-62 using ASCII 59-126
- Illumina 1.3: score 0-62 using ASCII 64-126
- Illumina 1.8: return to Sanger (Phred + 33)

Wikipedia

What is a SAM/BAM file?

- Read Mapping!

1. QNAME Query template/pair NAME
2. FLAG bitwise FLAG
3. RNAME Reference sequence NAME
4. POS 1-based leftmost POSition/coordinate of clipped sequence
5. MAPQ MAPping Quality (Phred-scaled)
6. CIGAR extended CIGAR string
7. MRNM Mate Reference sequence NaMe ('=' if same as RNAME)
8. MPOS 1-based Mate POSistion
9. LEN inferred Template LENgth (insert size)
10. SEQ query SEQuence on the same strand as the reference
11. QUAL query QUALity (ASCII-33 gives the Phred base quality)
12. OPT variable OPTional fields in the format TAG:VTYPE:VALUE

Samtools

What is a CIGAR string?

- How the read maps to the reference!
- Number of bases that match/mismatch/insertions/deletions

Reference: C C A T A C T G A A C T G A C T A A C

Read: A C T A G A A T G G C T

POS: 5

CIGAR: 3M1I3M1D5M

Wikipedia example

What is a BED file?

- Coordinates file! Can be visualized!

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.
4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

http://genome.uscs.edu

What is a GFF/GTF file?

- Annotation File!

1. **seqname** - name of the chromosome or scaffold
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Ensembl.org

What is a VCF file?

- Variant Call File!

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

File manipulation tools

- SAMtools
- BAMtools
- BEDtools
- VCFtools
- Picard

Coordinates!

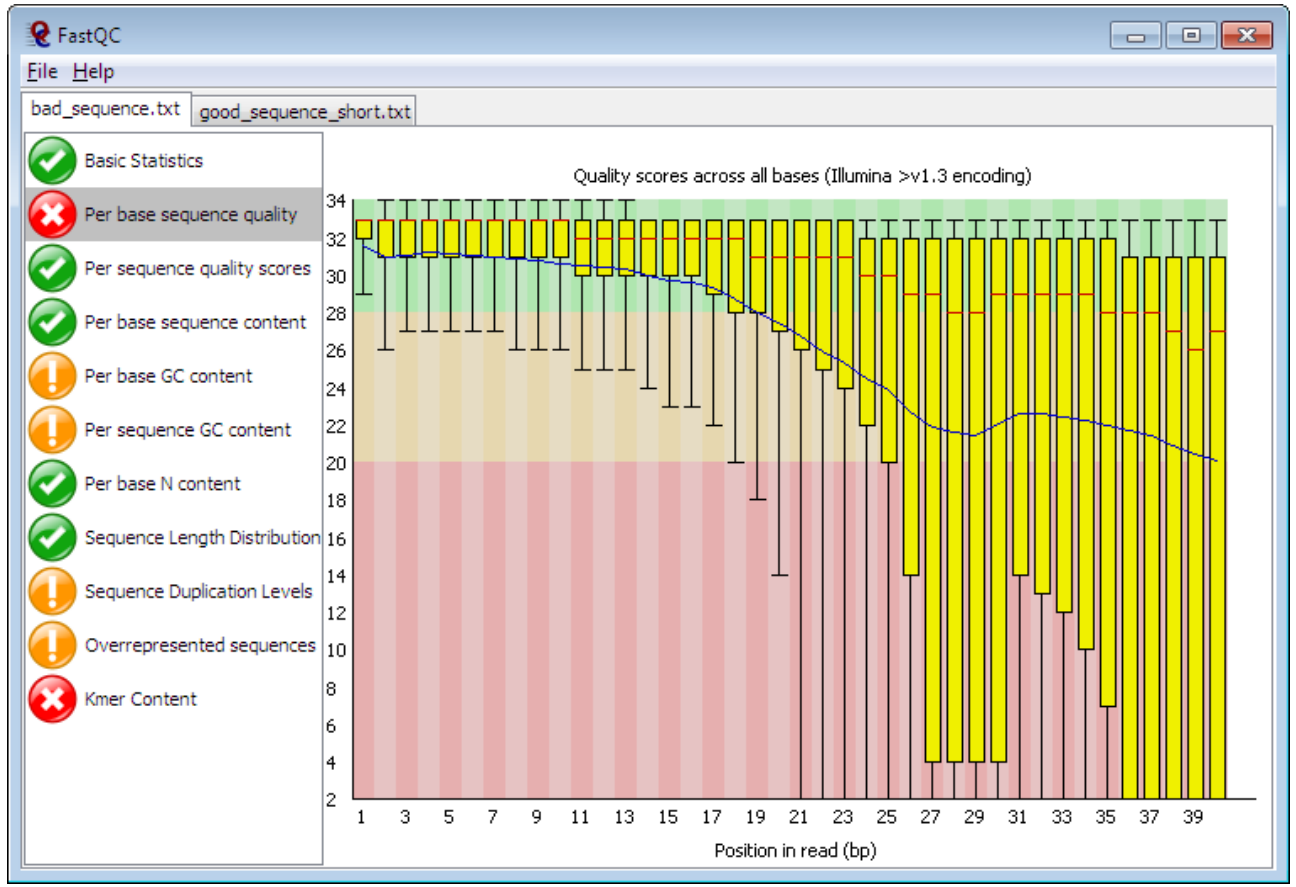
Analysis: Getting Started

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, kmer

FastQC: Poor Sequence



Trimming: Low Quality/Adapter/Barcode Removal

- Sequences won't align with these!
- Dynamic (based on sequence) or blunt (remove X from 5', Y from 3')
- Orchestra Options:
  - Clipper
  - Cutadapt
  - Fastx-trimmer
  - Flexbar
  - Trimmomatic
  - Trim galore
- PCR Duplicates

Alignment

Aligners

- Create BAM/SAM alignment file
- bwa
- Bowtie1
- Bowtie2
- Tophat2
- Novoalign
- STAR

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
- Orchestra: /groups/shared_databases
  - BWA
  - Bowtie1
  - Bowtie2
  - Novoalign
  - STAR

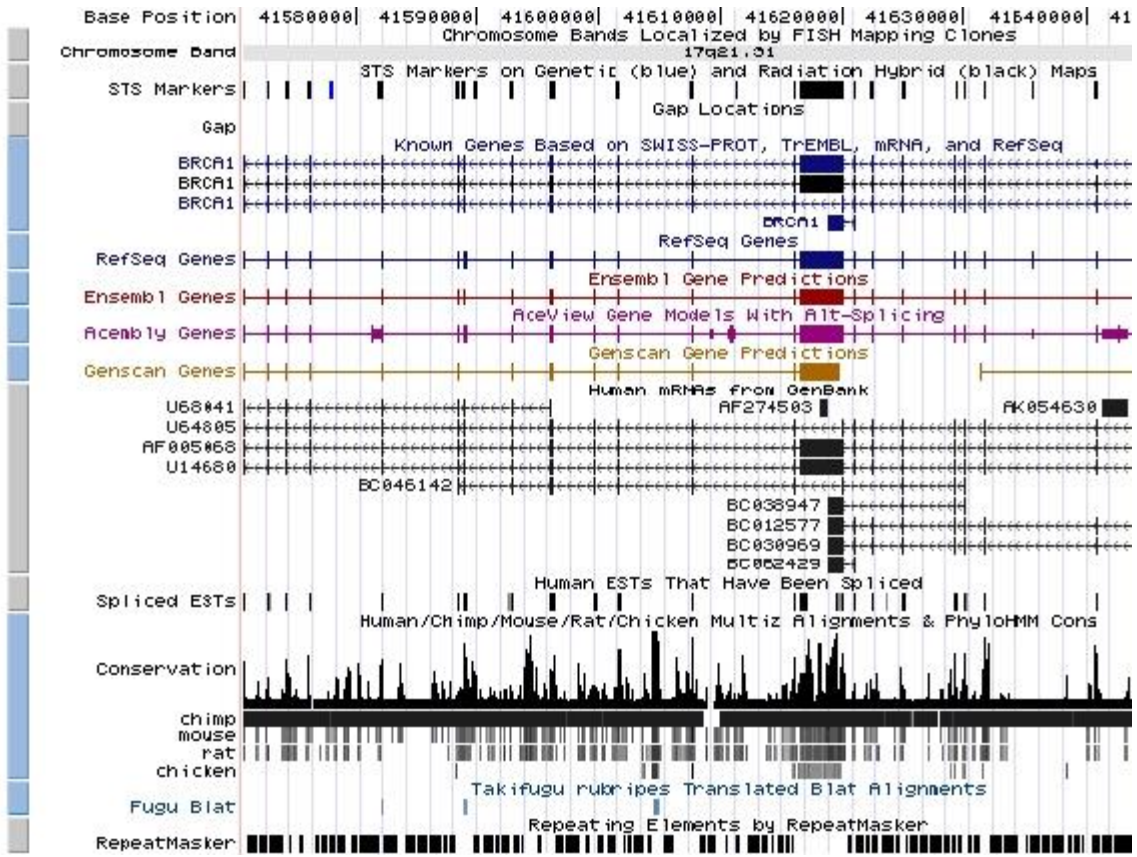
Alignment Considerations

- 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

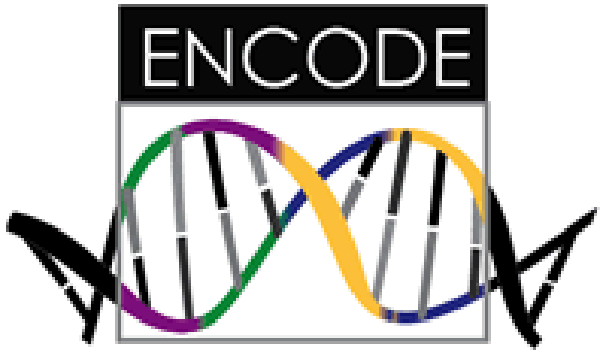
Analysis: After Alignment

DNA/Exome: Variant Callers

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

Peak Callers : ChIP-seq

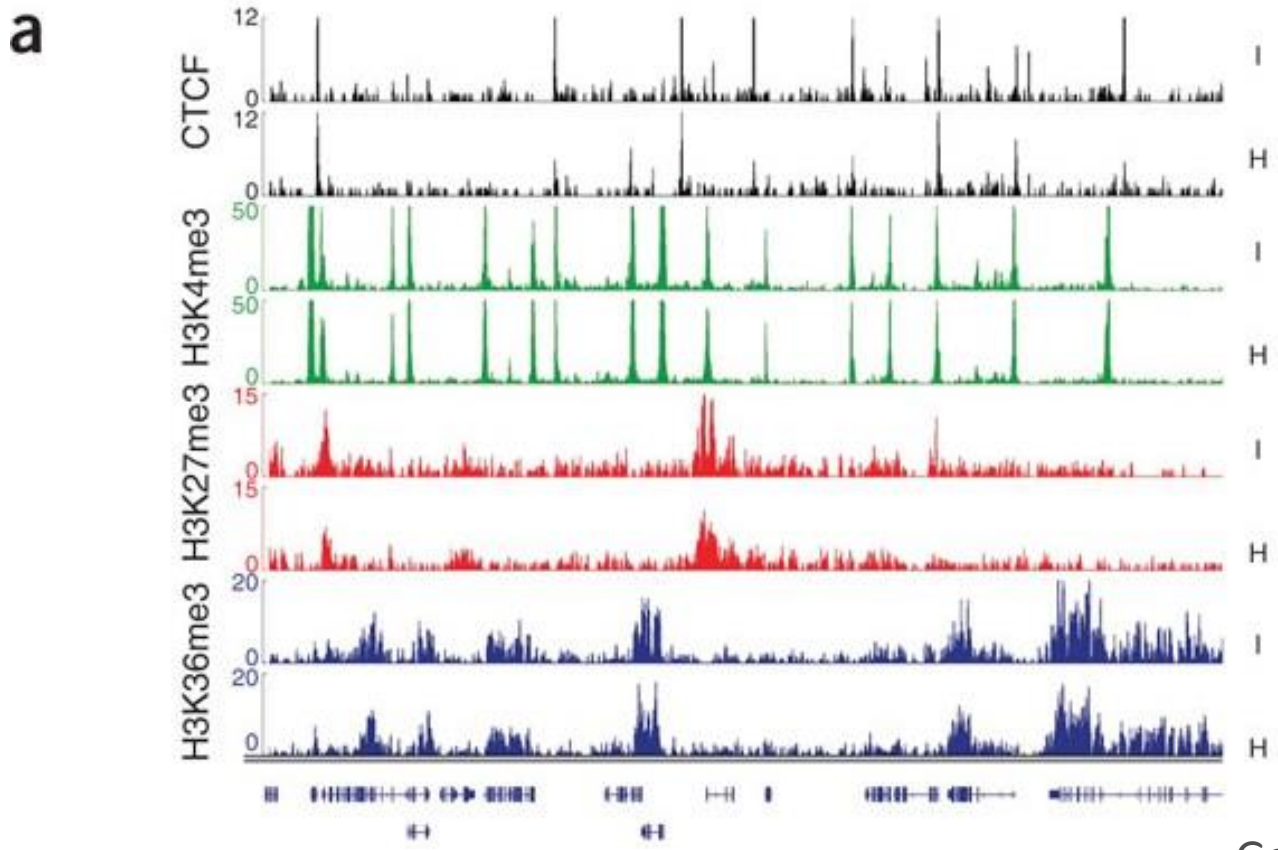
- SPP (R)
- GEM
- PeakSeq
- MACS2



Peak Callers: CLIP-seq

- PARalyzer
- dCLIP
- CIMS

Peak Visualization



Goren et al Nature Methods 2010

Motif Analysis

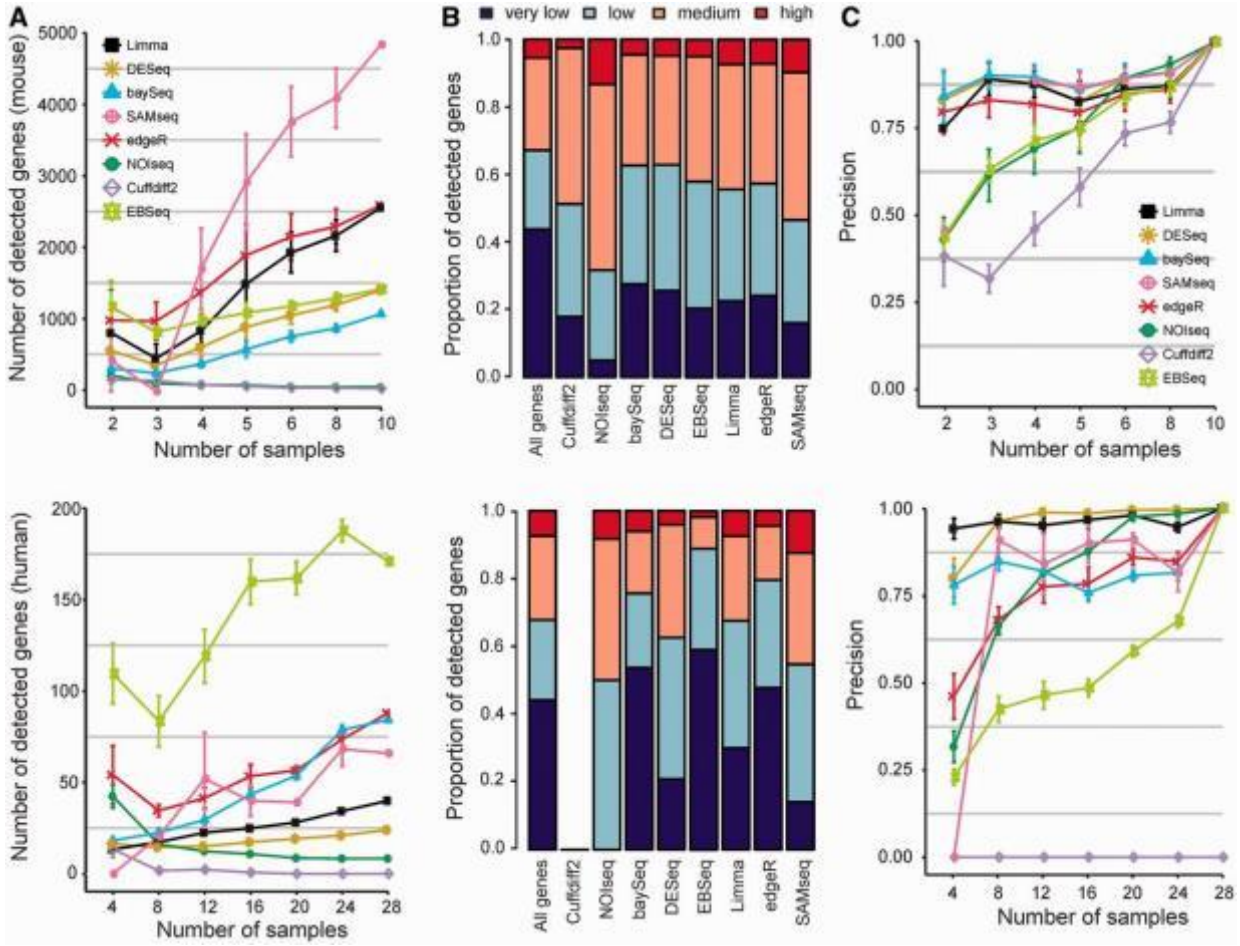
- HOMER
- MEME/MAST
- de Novo & Known Motifs

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

Differential Expression Analysis

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

Comparing DE algorithms



Seyednasrollah et al Briefings in Bioinformatics 2013

Functional Enrichment Analysis

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

Considerations

Experimental Design

- Sample size: Power Calculation!
  - Number of replicates needed, at what sequencing depth, to achieve statistical power
- Control variables
- 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

- Don’t be a jailer!

Bild et al PLOS Biology 2014

For further questions

- http://rc.hms.harvard.edu
- rchelp@hms.harvard.edu
- Office Hours: Wed 1-3p Gordon Hall 500
- Survey: http://hmsrc.me/introngs2016-survey2