Sequencing Technologies
(with an incredibly narrow focus on RNA-seq)

Morgan Sammons, Assistant Professor of Biological Sciences

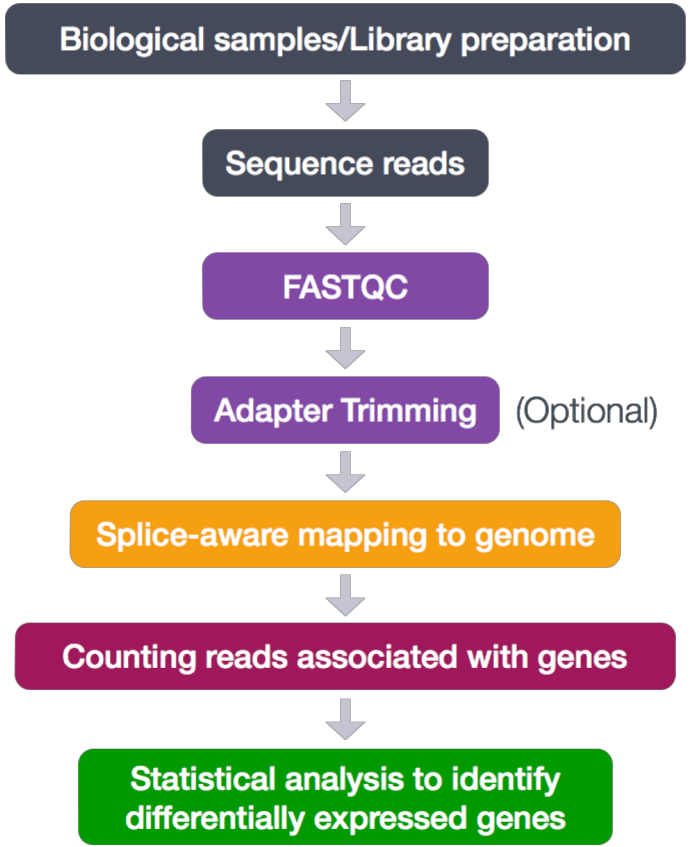
Ryan Meng, Bioinformatics Support Specialist

Nicholas Schiraldi, Data Analytics Specialist, ITS

UNIVERSITY AT ALBANY
State University of New York

Thank you to illumina, sponsor of the Sequencing Technologies Workshop

illumina®


Biological samples/Library preparation
Sequence reads
FASTQC
Adapter Trimming
(Optional)
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

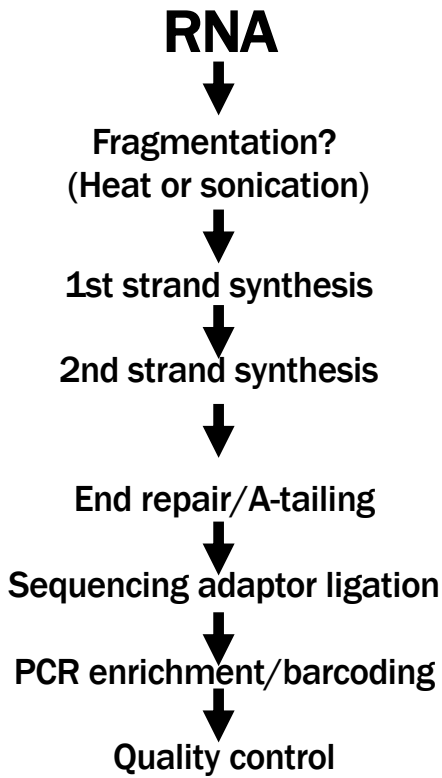
If you can isolate a nucleic acid, you can *sequence* it

Garbage in, Garbage out

Remember: Good starting material will more likely result in “good” data.

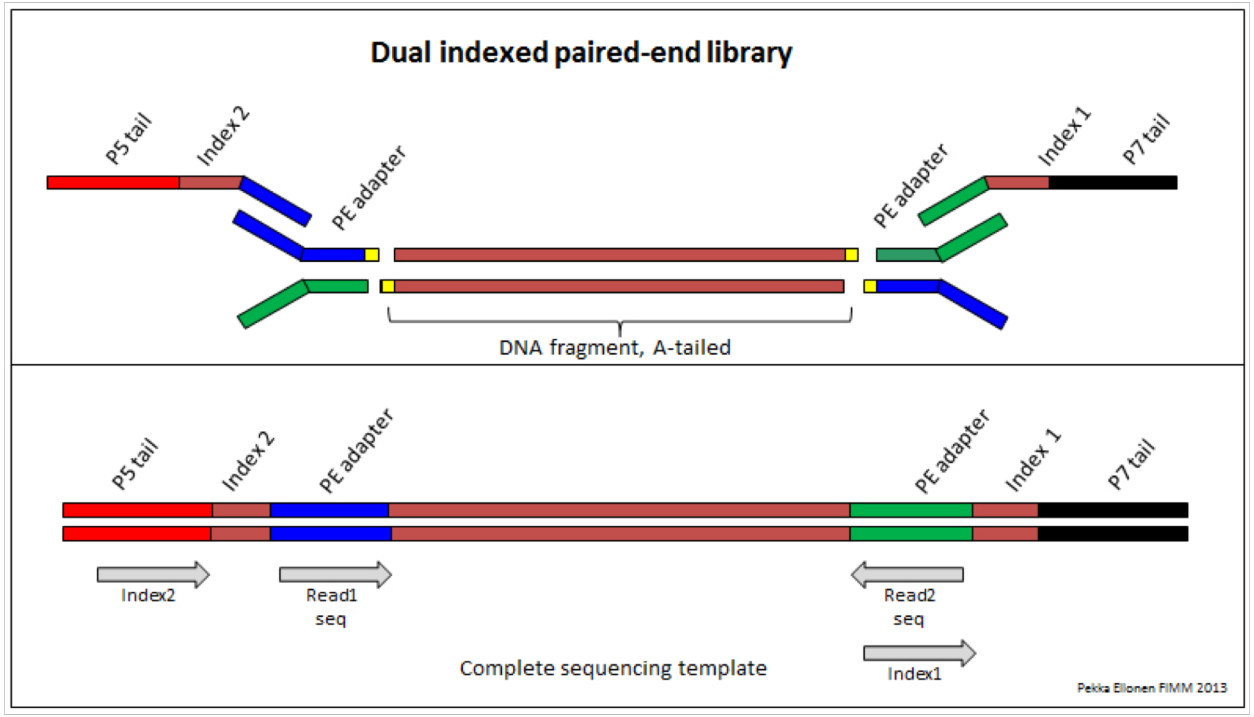
Remember: Data that do not fit expected results may reflect less-than-optimal starting material.

Workflow for sequencing libraries*



*for Illumina sequencers

What is a sequencing library?



Making sequencing libraries for:

RNA-Seq

Small RNA-Seq

~~**ChIP-Seq**~~

~~**DNA-seq**~~

RNA-Seq - Advantages

RNA isolation is straightforward

Low sample requirements (as low as 10pg…1 cell)

Unbiased view of the transcriptome (no prior knowledge)

Robust data analysis/statistical pipelines available

Mature technology

RNA-Seq - Disadvantages

"Relatively" expensive

High knowledge barrier to entry (many many many tools/software packages

EVERYTHING is observed (no more willful ignorance)

Validation?

Depletion and Enrichment strategies

rRNA depletion

More complex transcriptome
lncRNA, miRNA, tRNA, eRNA…

Expensive (>$50/sample)

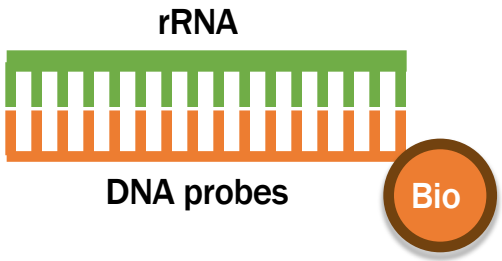
PolyA Enrichment

Less complex transcriptome
Only mRNA (-/+ a few things)

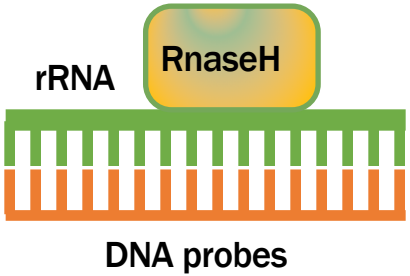
Cheap (≈$3/sample)

rRNA depletion strategies for RNA-seq

Affinity-mediated depletion



RNAse-mediated degradation

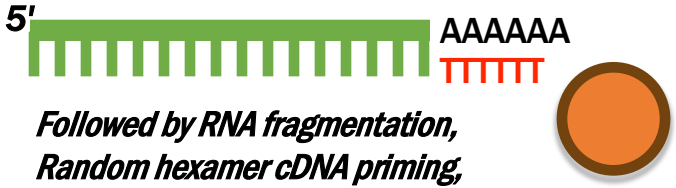


5’ phosphate-dependent exonuclease

Also relevant to cells/tissues with high levels of other transcripts (globin in RBC)

Strategies for polyA-enriched RNA-seq

Isolation of polyA with anchored poly dT beads



Followed by RNA fragmentation, Random hexamer cDNA priming, RNAseH-mediated 2nd strand

Most Vendors (Illumina, NEB, etc)

Full-length, 1st strand synthesis from a total RNA sample using 5' template switching chemistry

Followed by sonication and DNA library prep or tagmentation

Clontech SMRT-seq kits, others

3’ Tag Counting and Alternate polyA Site Identification

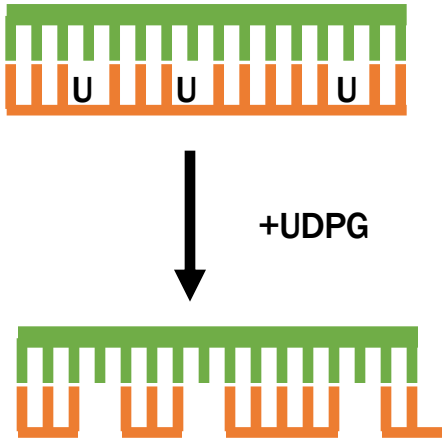
Priming the cDNA reaction using an anchored polyT primer

Digest RNA
Prime 2nd Strand Reaction with Random N-mers

Fairly straightforward data analysis
Inexpensive
Easy to multiplex
Miss out on data….

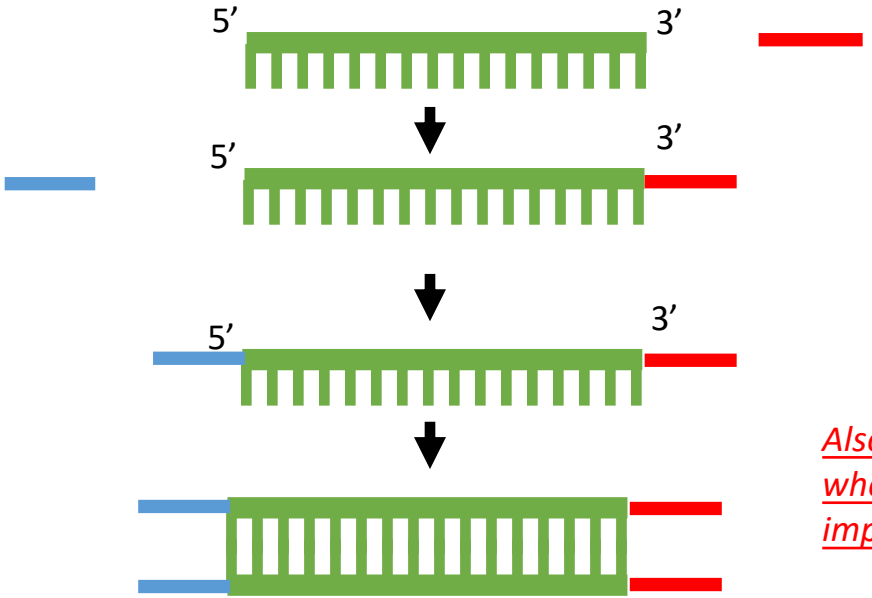
Methods for determining RNA strandedness

Defining RNA strandedness during 2nd strand synthesis using dUTP



Defining RNA strandedness through ligation

Adaptors of a specific, known sequence are added to the 5' and 3' end of the RNA (or one of the cDNA strands)



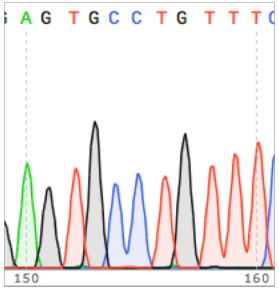
Also, usual method for small RNA-seq where random or specific priming is impractical

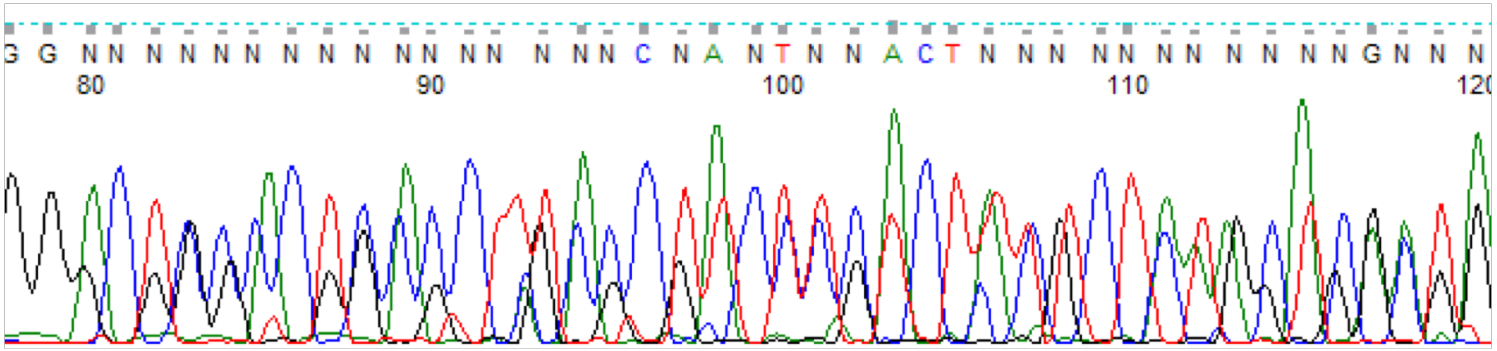
dsDNA can now be used to prepare a sequencing library

Important Sequencer-specific information could have been added earlier during synthesis steps or at this point through standard DNA library prep processes

Internal barcodes, unique molecular identifiers (UMI), or other information can be added during 1st strand, 2nd strand, or at this step.

```
>FastaSequence
CTCAATTCGTGTCTGAACNTTTGAACATCTTCCTTGGNAGCCTTGACCTTT
```





Structure of a FASTQ File

```
1 @SRR636633.1 HISEQ2:193:D0CW2ABXX:3:1101:1122:2244 length=51
2 CTCAATTCGTGTCTGAACNTTTGAACATCTTCCTTGGNAGCCTTGACCTTT
3 +
4 @@@DDDDDD=DAC?FHII#3AAGBFGGHFGHII>BGG#1?BGHGGIIIIHB
```

1 – sequence identifier and optional information. Always starts with @
2 – This is your sequence
3 – starts with a +; optional information is repeat of line 1
4 – encodes the quality values for the sequence in line 2

PHRED Scores/Quality Values

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

The higher the Q value, the lower the likelihood of an incorrect basecall

Paired end sequencing data

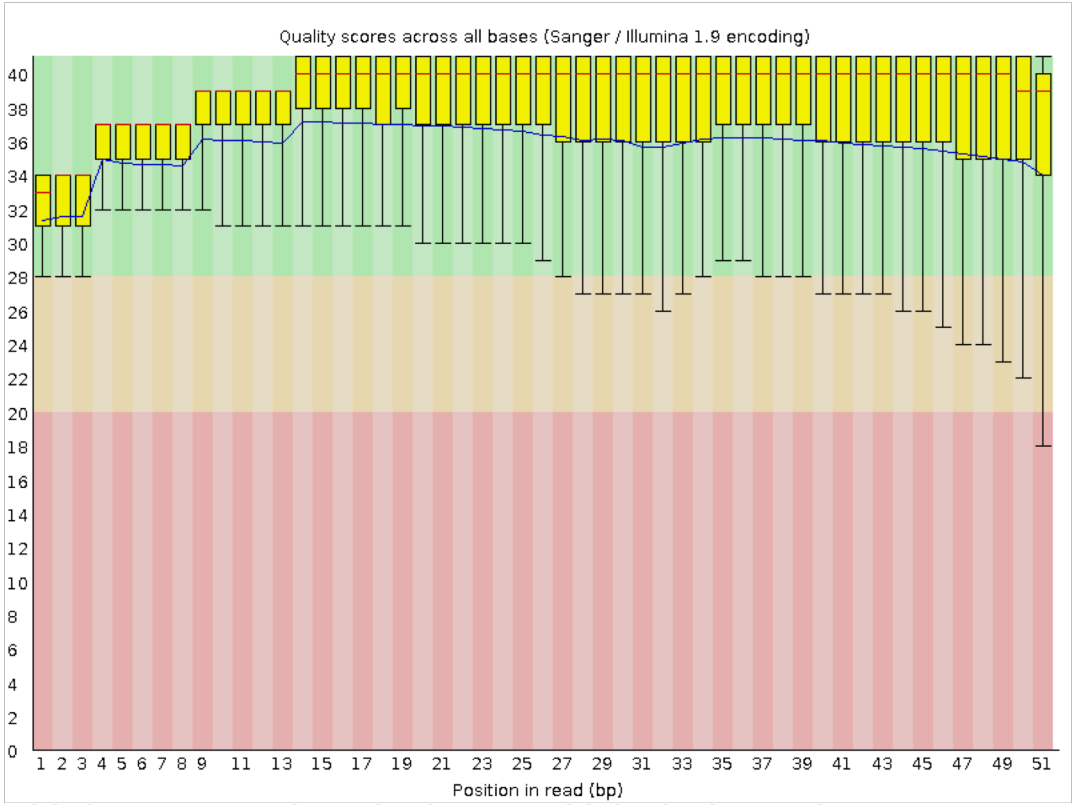
FASTQ R1	FASTQ R2
Read 1	Read 1 Mate
Read 2	Read 2 Mate
Read 3	Read 3 Mate
Read 4	Read 4 Mate
Read 5	Read 5 Mate
Read N	Read N Mate

Interleaved PE FASTQ

Read 1
Read 1 Mate
Read 2
Read 2 Mate
Read 3
Read 3 Mate
Read 4
Read 4 Mate

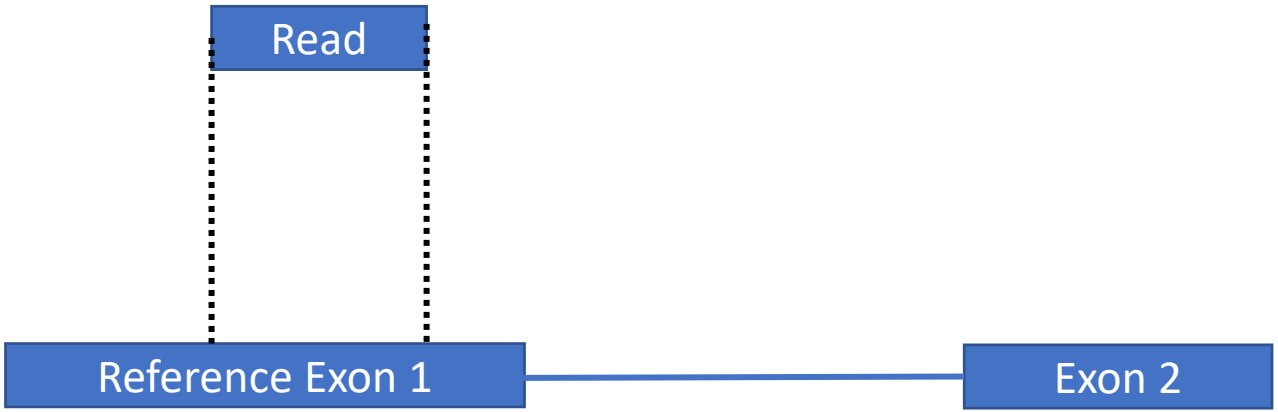
Assessing Quality of Sequencing Data

Fastqc Example

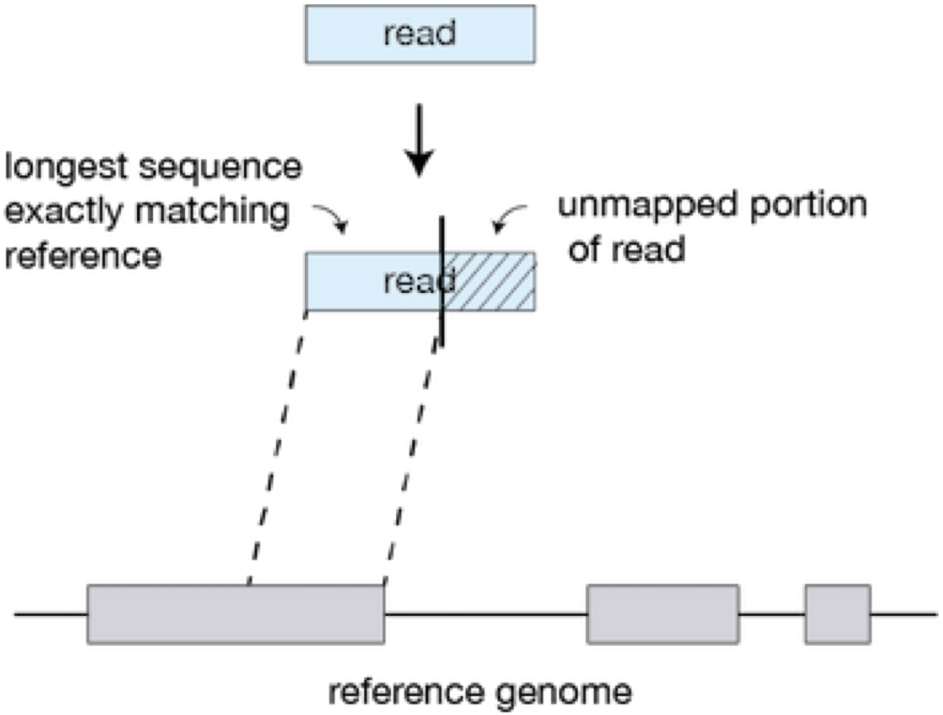


Important point: this is not assessing whether your biological question was answered…just whether the sequencing went well

Mapping to a Reference Genome

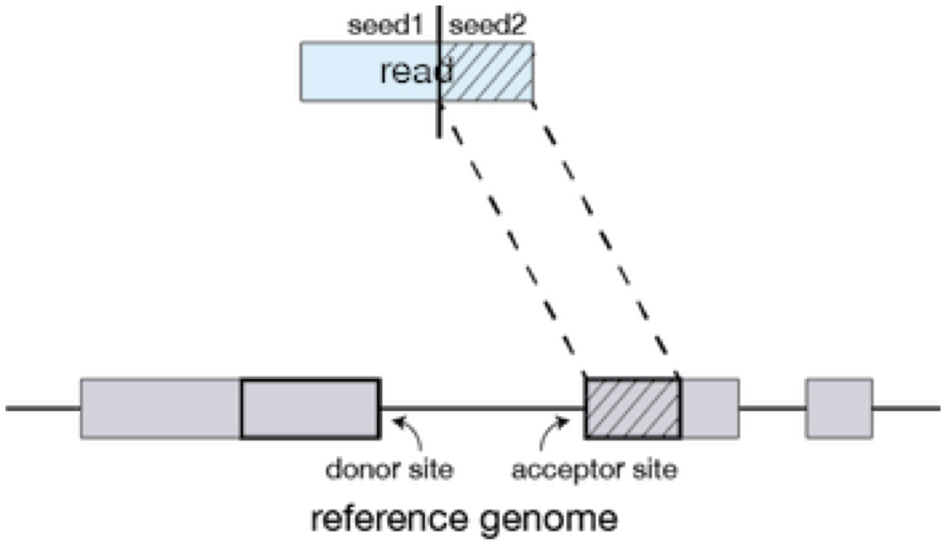


Splice-aware Alignments



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

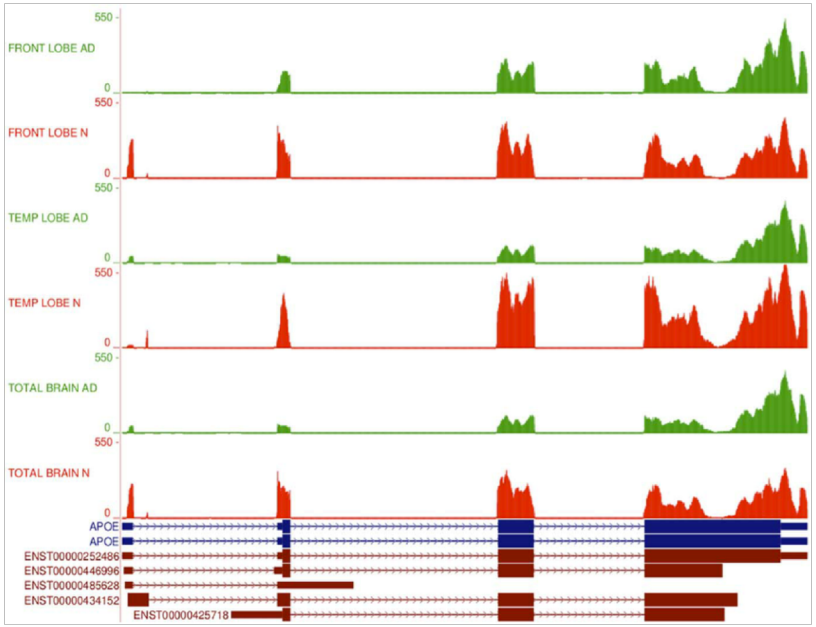
Splice-aware Alignments



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

DATA ANALYSIS

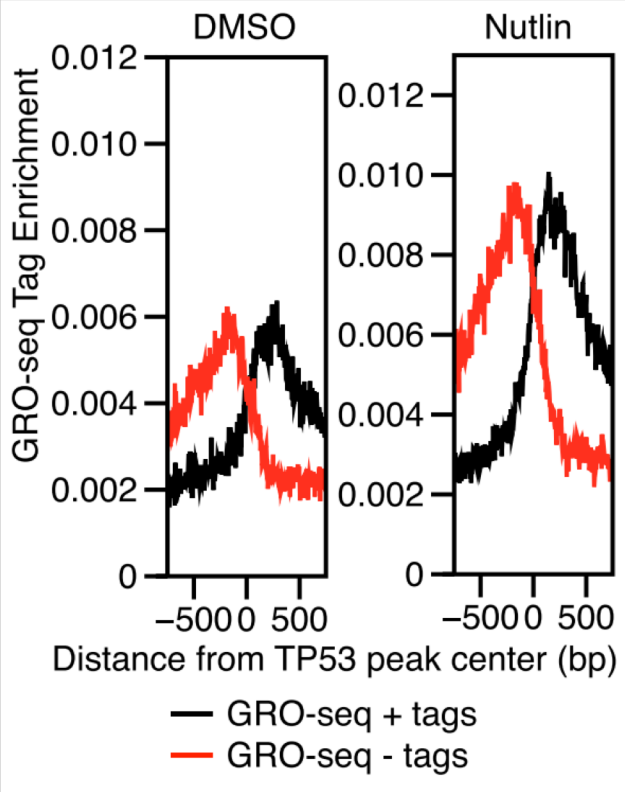
Visualizing read density across a gene

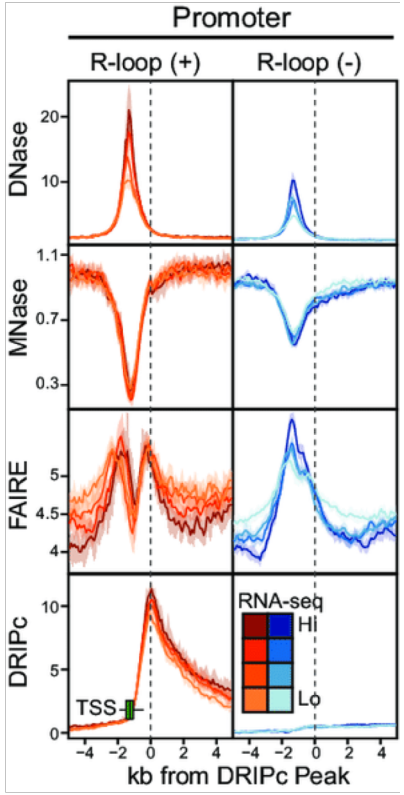


BAM can be loaded directly into **IGV**

BAM can be processed into a more lightweight file (bedGraph, bigwig) for viewing in **UCSC** or **WashU** Genome Browsers or **IGV**

”Metaplot”-style graphs





Calculating RPKM and TPM

You may want to have some value that you attribute to a particular gene or mRNA isoform.

Our analysis provides RAW counts, or how many reads map to that locus.

This is an unnormalized, raw value.

Larger genes will have more reads mapped than smaller genes just because of the size.

Also, library depth matters. If Sample 1 has 10e6 reads and Sample 2 has 20e7, we expect same gene in Sample 2 to see 2X reads mapped by default.

We can solve these size normalization problems a few different ways.

Calculating RPKM and TPM

RPKM

Reads per Kilobase per Million

$$\frac{\text{Total \# of Reads Mapped in Sample}}{1{,}000{,}000} = \text{Scaling Factor}$$

$$\frac{\text{Total \# of Reads Mapped to Gene X}}{\text{Scaling Factor}} = \text{RPM}$$

$$\frac{\text{RPM}}{\text{Length of Gene, in kb}} = \text{RPKM}$$

TPM

Transcripts per million

$$\text{RPK} = \frac{\text{Total \# of Reads Mapped to Gene X}}{\text{Length of Gene, in kb}}$$

$$\text{Scaling Factor} = \frac{\text{Add all RPK values for sample}}{1{,}000{,}000}$$

$$\text{TPM} = \frac{\text{Individual Gene RPK}}{\text{Scaling Factor}}$$

Other fantastic tools

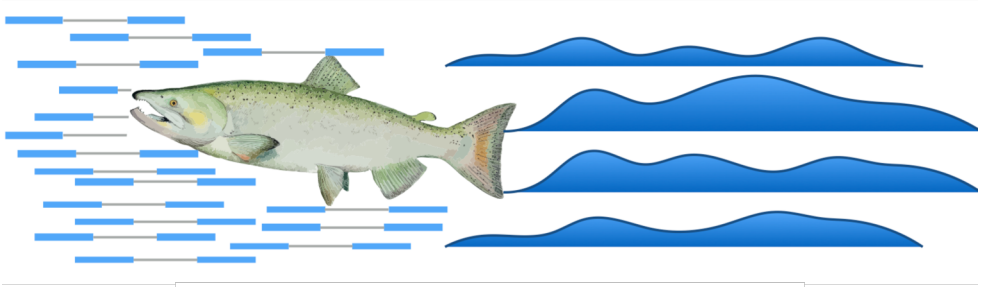
Galaxy

- Graphical User Interface versions of all of these tools we used today, plus MANY others
- No working at the command line
- But, computation/analysis is done on a distant server. Queue times can be slow. Data must be transferred (time = money). Not HIPAA compliant
- Local HPC admins can set up a version, but please be nice to them!

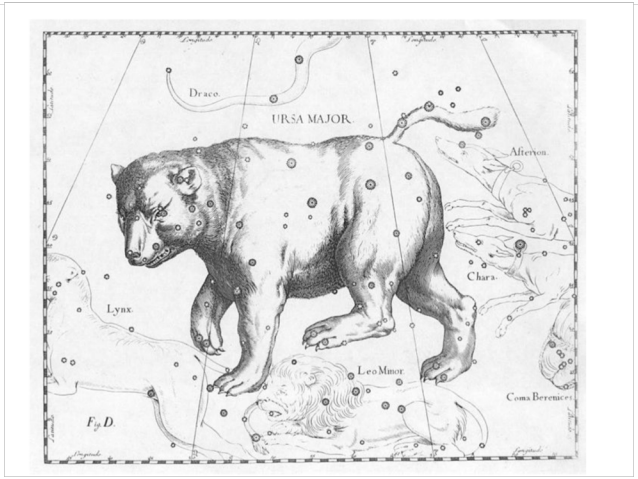
Illumina BaseSpace

- Graphical User Interface for Many RNA-seq and other computational biology tools
- Integrated with your Illumina Sequencer (less data transfer)
- Really straightforward
- But, it does cost $$$, but it's pretty nice and you don't need any infrastructure to run

Quantifying RNA Expression using Alignment-free Methods



salmon



kallisto

SAM/BAM File Format