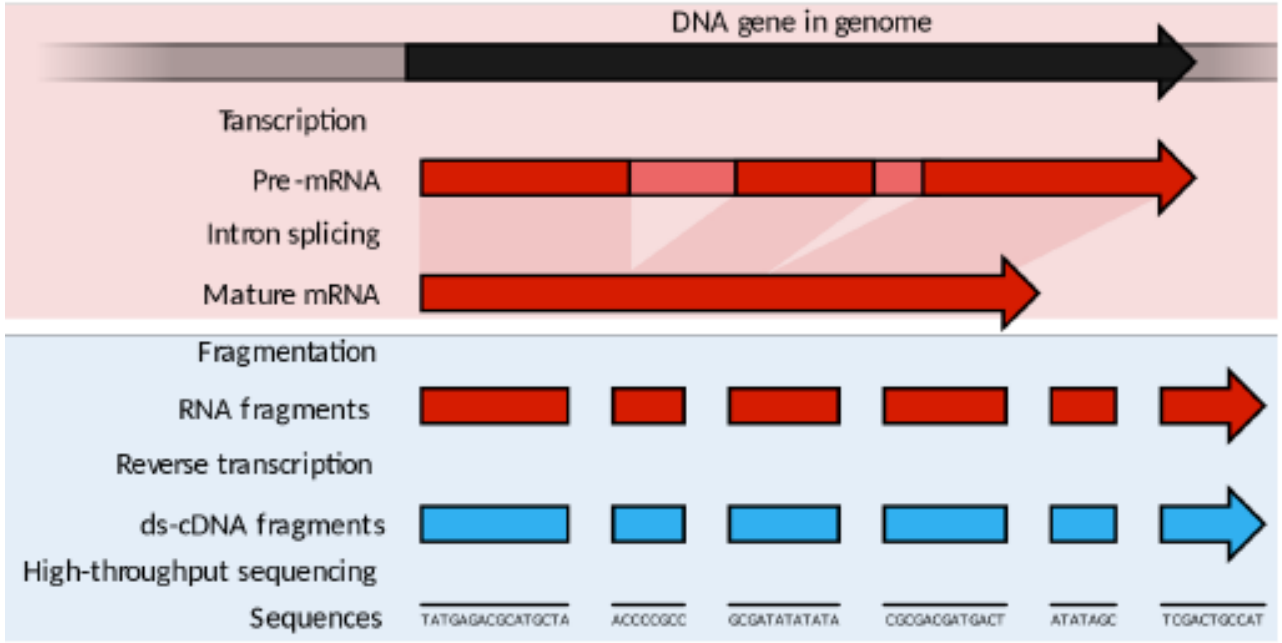
RNASeq & Analysis



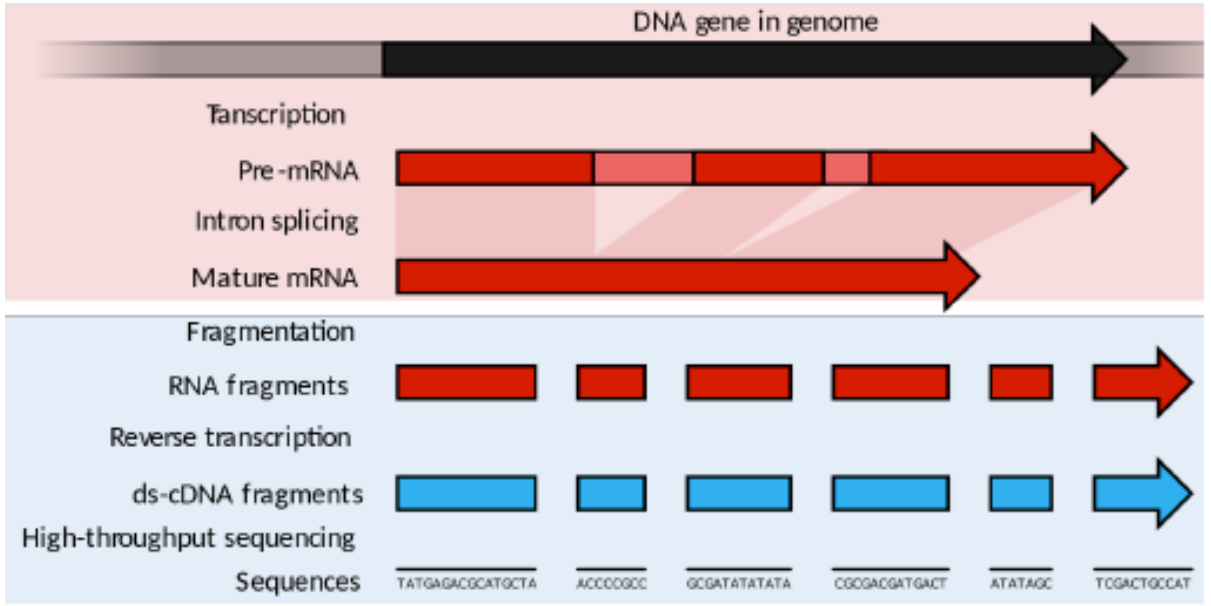
BIOL 435/535: Bioinformatics

4/19/2022

Sequencing RNA vs. DNA

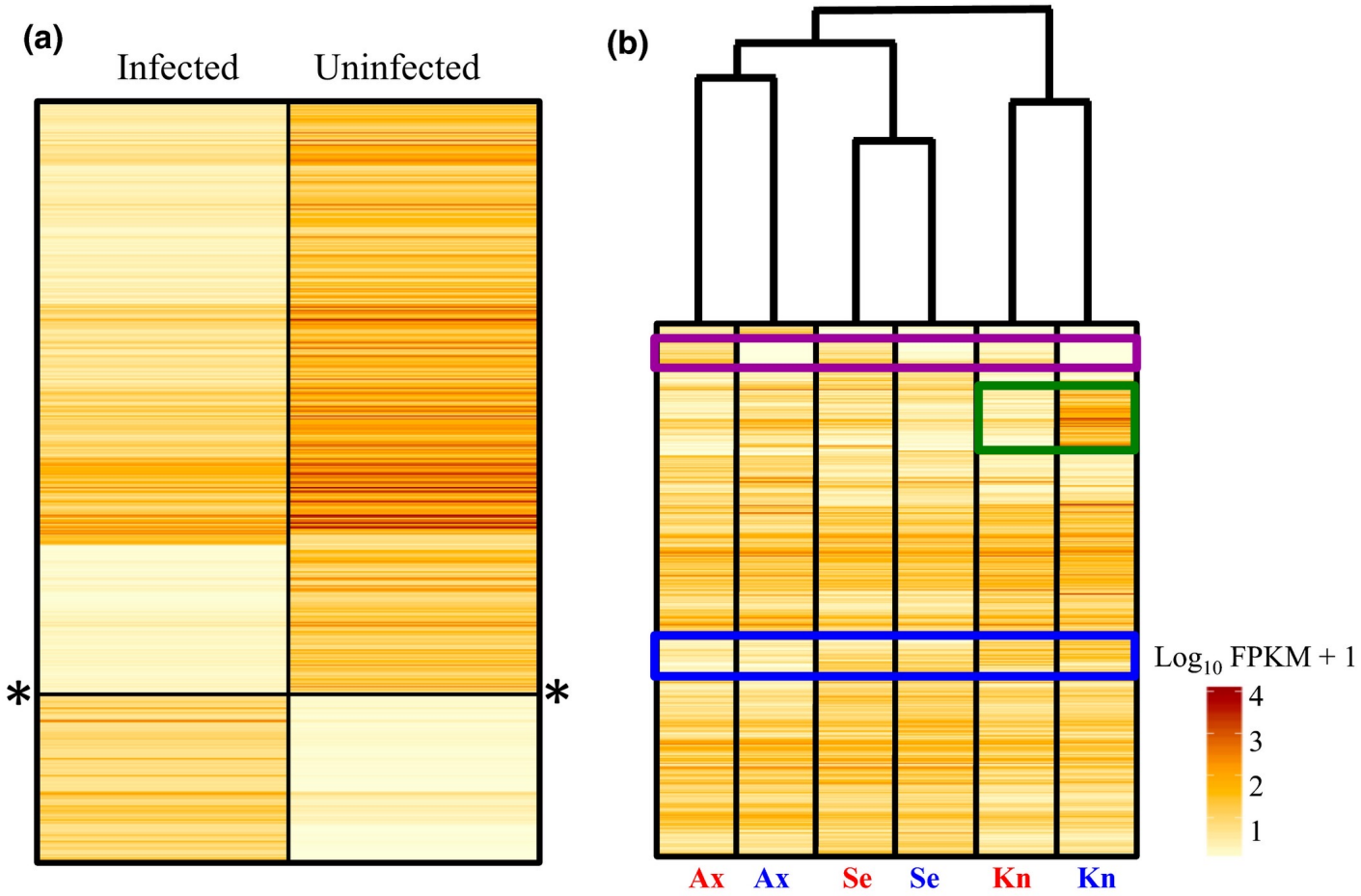
Important differences

- Single-stranded
- Variable copy number
- Post transcriptionally modified
  - Intron splicing
  - RNA editing
  - Polyadenylation (Poly-A tail)



RNAseq Applications

- Quantify gene expression



Bankers et al., 2017 Mol Ecol

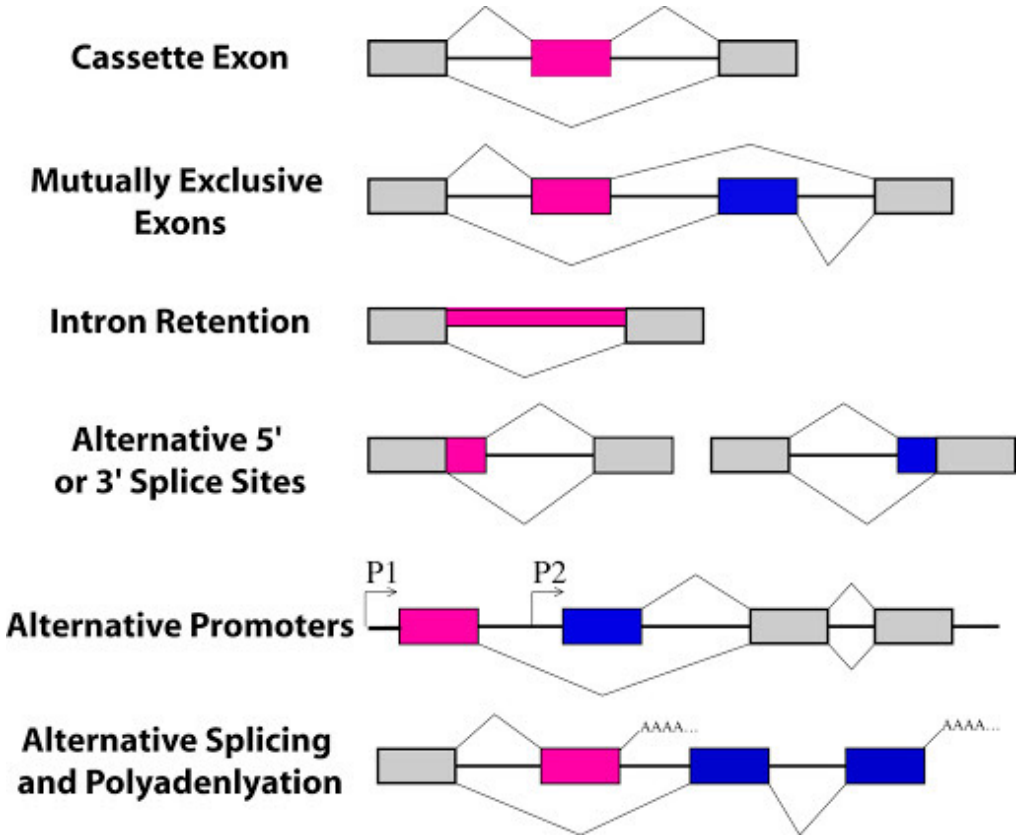
RNAseq Applications

- Quantify gene expression
- Determine gene sequences in non-model species (de novo transcriptome assembly)

McElroy et al., 2017 GBE

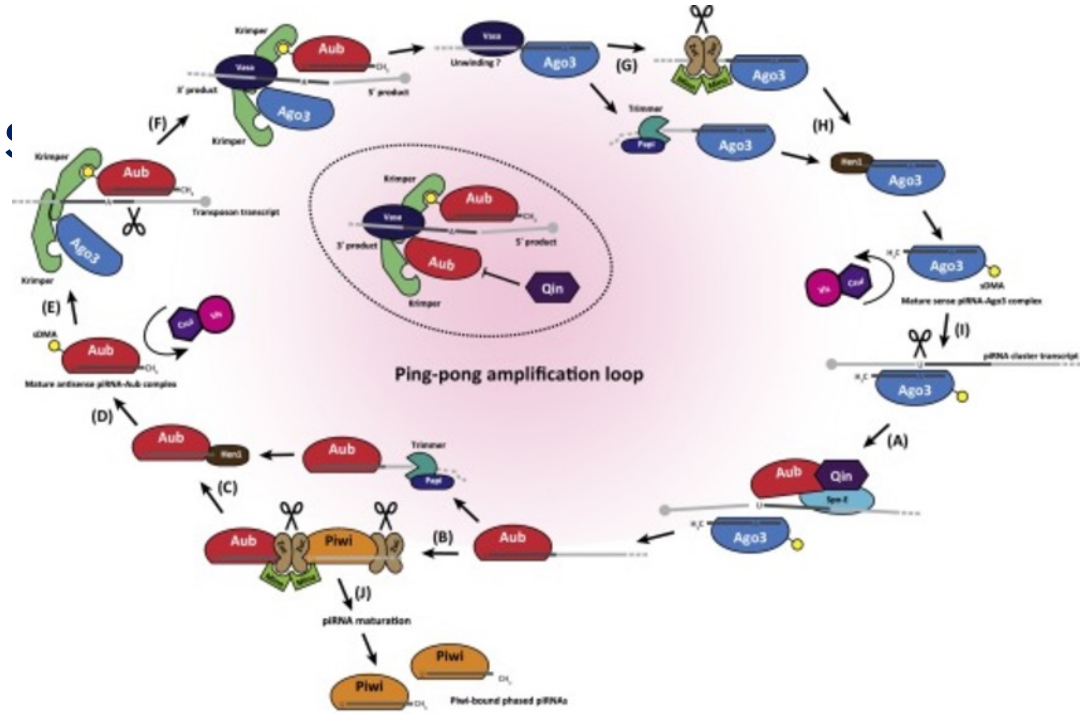
RNAseq Applications

- Quantify gene expression
- Determine gene sequences in non-model species (de novo transcriptome assembly)
- Identify post-transcriptional modification (alternative) splicing, RNA editing



RNAseq Applications

- Quantify gene expression
- Determine gene sequences in non-model species (de novo transcriptome assembly)
- Identify post-transcriptional modifications (alternative) splicing, RNA editing
- Identify functional/transcribed elements (non-coding RNAs)



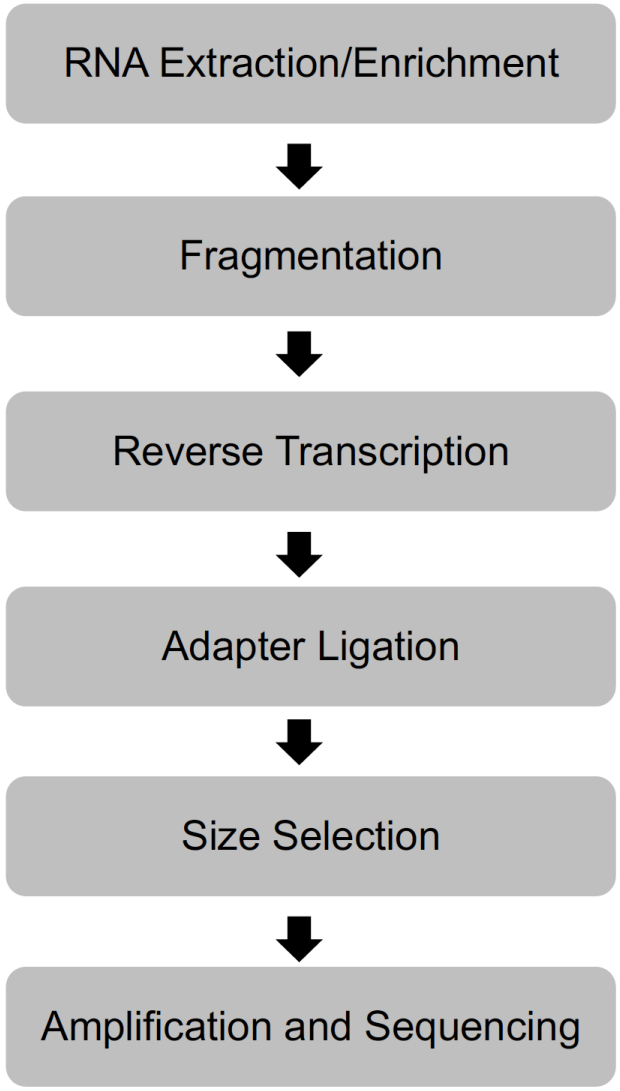
Trends in Biochemical Sciences

RNAseq Applications

- Quantify gene expression
- Determine gene sequences in non-model species (de novo transcriptome assembly)
- Identify post-transcriptional modifications: (alternative) splicing, RNA editing
- Identify functional/transcribed elements (non-coding RNAs)
- Identify targets or footprints of RNA binding proteins

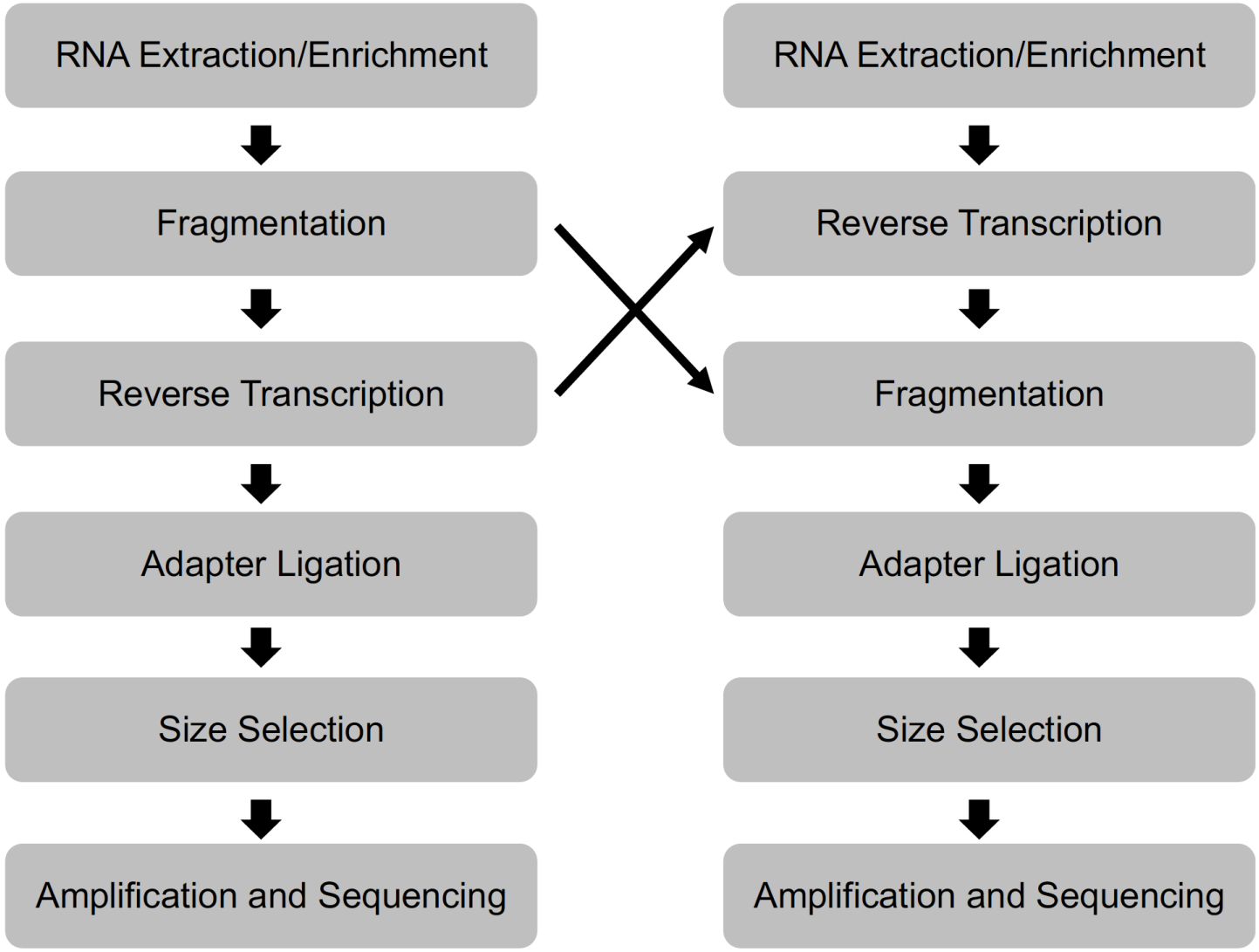
Slide courtesy of Jessica Warren

RNAseq Protocols



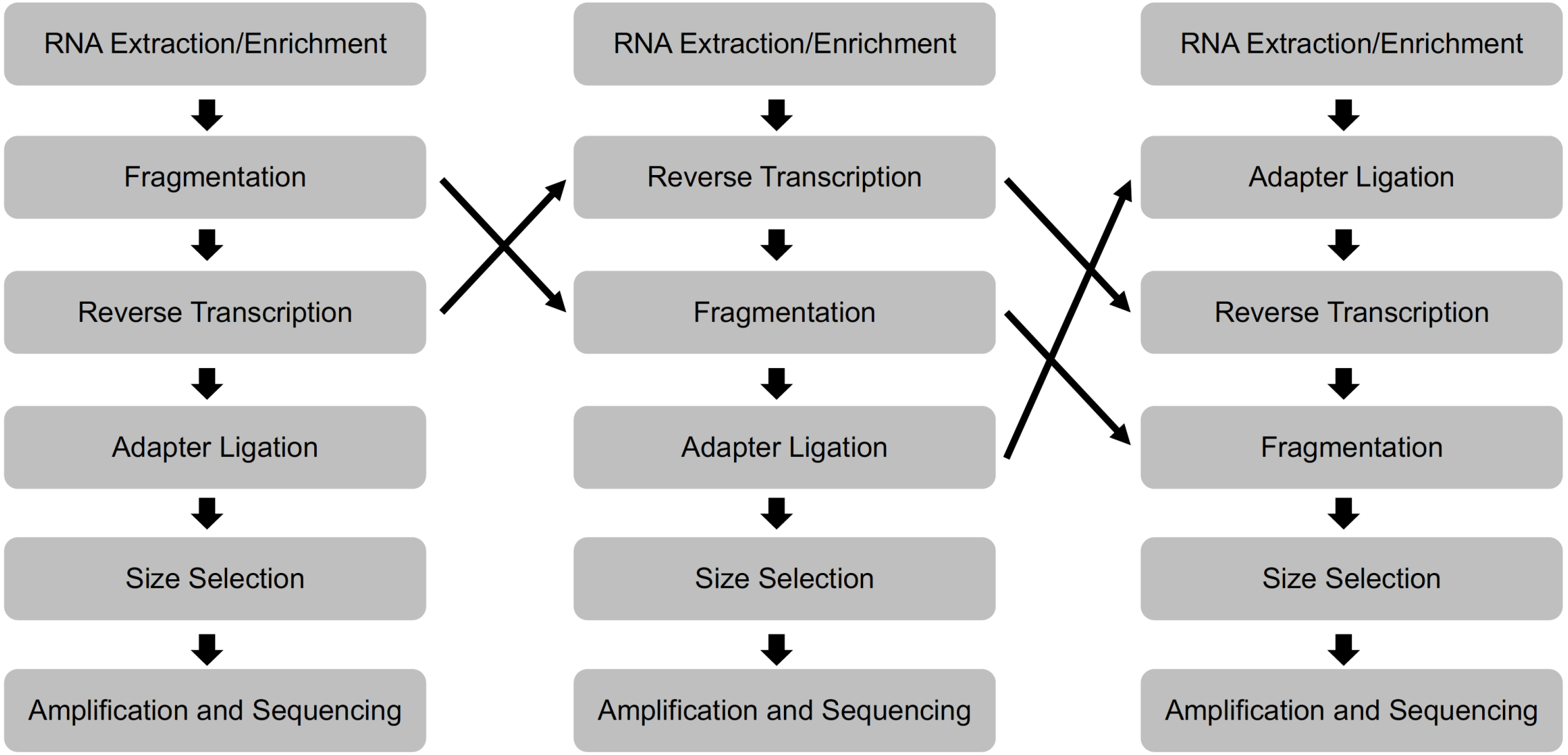
Slide courtesy of Jessica Warren

RNAseq Protocols



Slide courtesy of Jessica Warren

RNAseq Protocols



Slide courtesy of Jessica Warren

RNA Extraction

Silica-membrane kits

- Does NOT retain RNAs < 200 nt
- More expensive
- Lower yield?
- Higher purity?

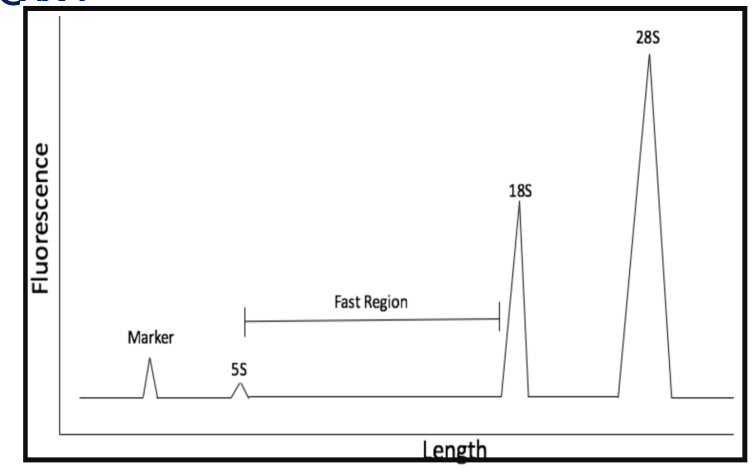
TRIzol

- Does retain RNAs < 200 nt
- Less expensive
- Higher yield?
- Lower purity?

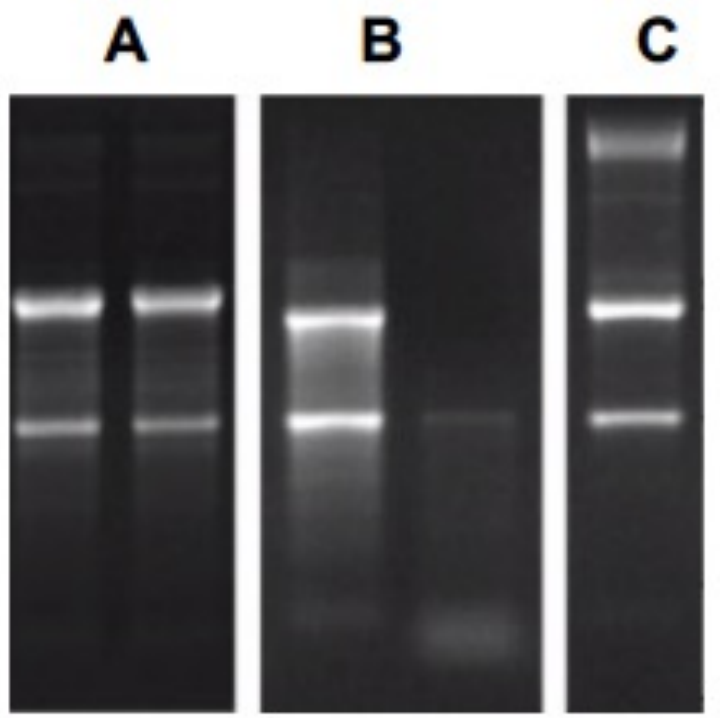
Slide courtesy of Jessica Warren

RNA Quantity, Quality, and RIN

- DNA contamination (DNase treatment)
- Measure concentration (NanoDrop, Qubit/fluorometry, Agilent BioAnalyzer/TapeStation)
- Verify purity (NanoDrop, Qubit/ fluorometry, agarose gel, PCR)
- Verify size/integrity (agarose gel, Agilent BioAnalyzer/TapeStation)



An idealized TapeStation trace showing a RIN Score of 10



A. Relatively non-degraded RNA showing a 2:1 28S and 18S bands

B. Two degraded RNA samples, the left being severely degraded.

C. DNA contamination

Slide courtesy of Jessica Warren

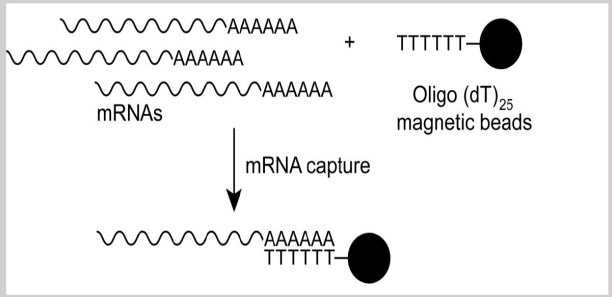
RNA molecules differ in abundance

~80% of total RNA is rRNA
~15% is tRNA
~4% is mRNA
~1% is other ncRNA

Ribosomal RNA (rRNA) Depletion

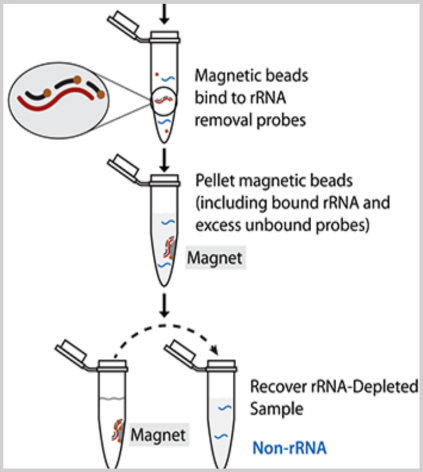
Positive Selection

- Oligo-dT magnetic beads to capture polyA+ transcripts
- Inexpensive
- mRNAs only
- Biased representation (e.g., organellar RNAs, transcript maturation, AT-rich sequences, 3' ends)



Negative Selection

- Targeted probes and magnetic beads for rRNA
- More costly
- Everything except rRNA
- Taxon-specific (not available for all organisms)



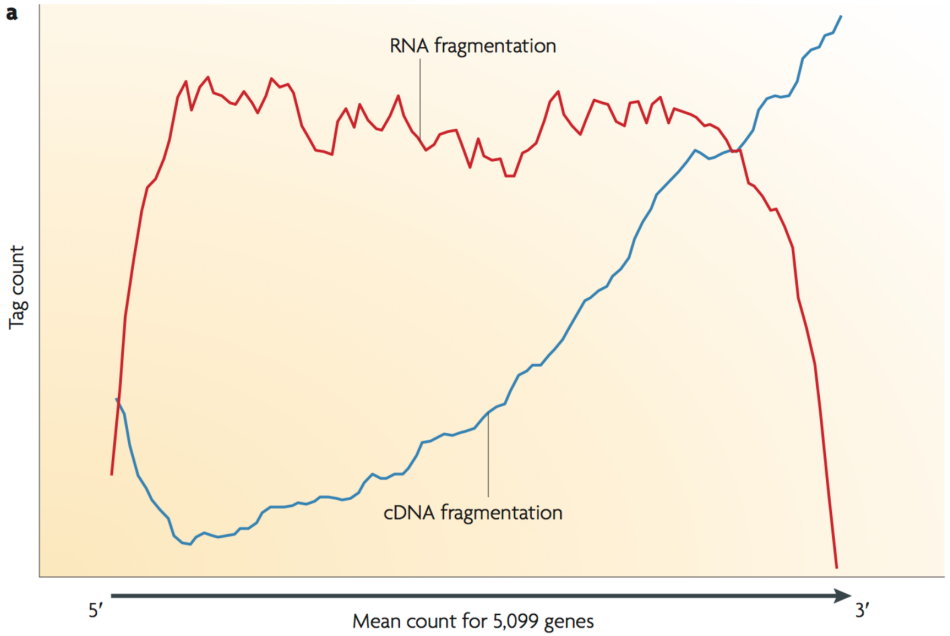
Slide courtesy of Jessica Warren

Fragmentation

Size limitation on Illumina sequencers is <600 bp.

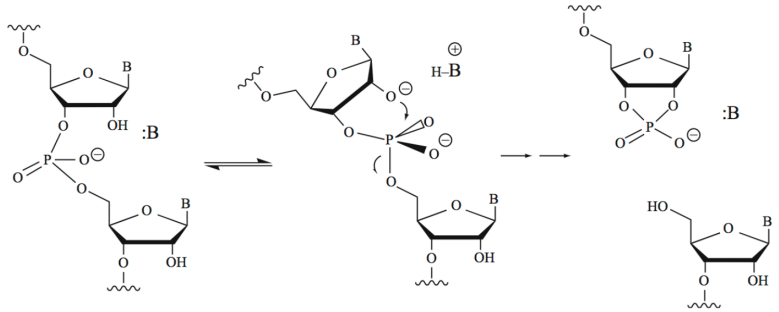
The average size of an animal mRNA is 2.5 kb.

Fragment at the cDNA or RNA stage?



Fragmenting RNA rather than cDNA results in less bias in coverage

Fragmentation methods



Chemical fragmentation of RNA with heat and salts (divalent cations)

- Quick, cheap, and easy
- Fine-tunable: adjust incubation time

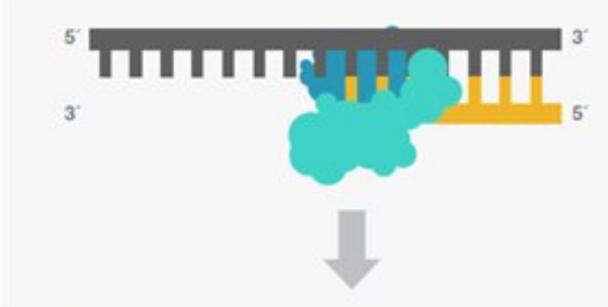
Enzymatic fragmentation (RNase III)

- Quick, cheap, and easy
- Has a preference for double-stranded RNA, can result in sequencing bias

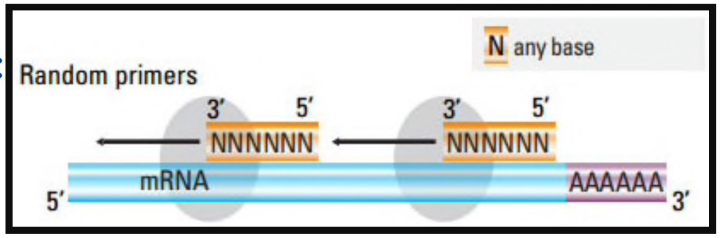
Small RNA-seq → No fragmentation necessary!

Slide courtesy of Jessica Warren

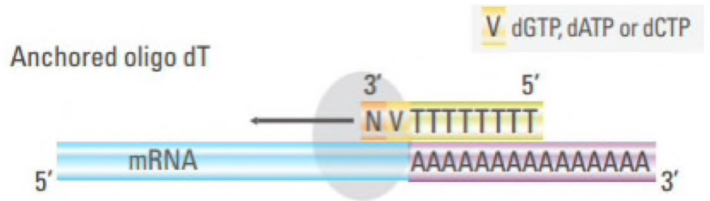
Reverse Transcription (RNA → cDNA)



- Random primers (random hexamer): (preferable to oligo-dT primers



- Creates DNA that is complementary to RNA transcript à cDNA



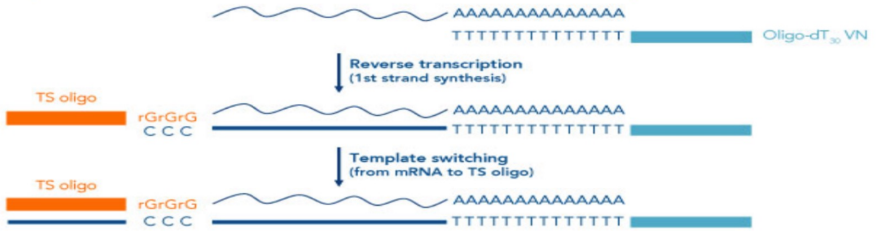
- First- and second-strand synthesis

Specialized RT methods

- Not So Random (NSR) primers. Avoids rRNA but species-specific, common in prokaryotic studies with small inputs

https://www.thermofisher.com/us/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/rt-education/reverse-transcriptase-attributes.html

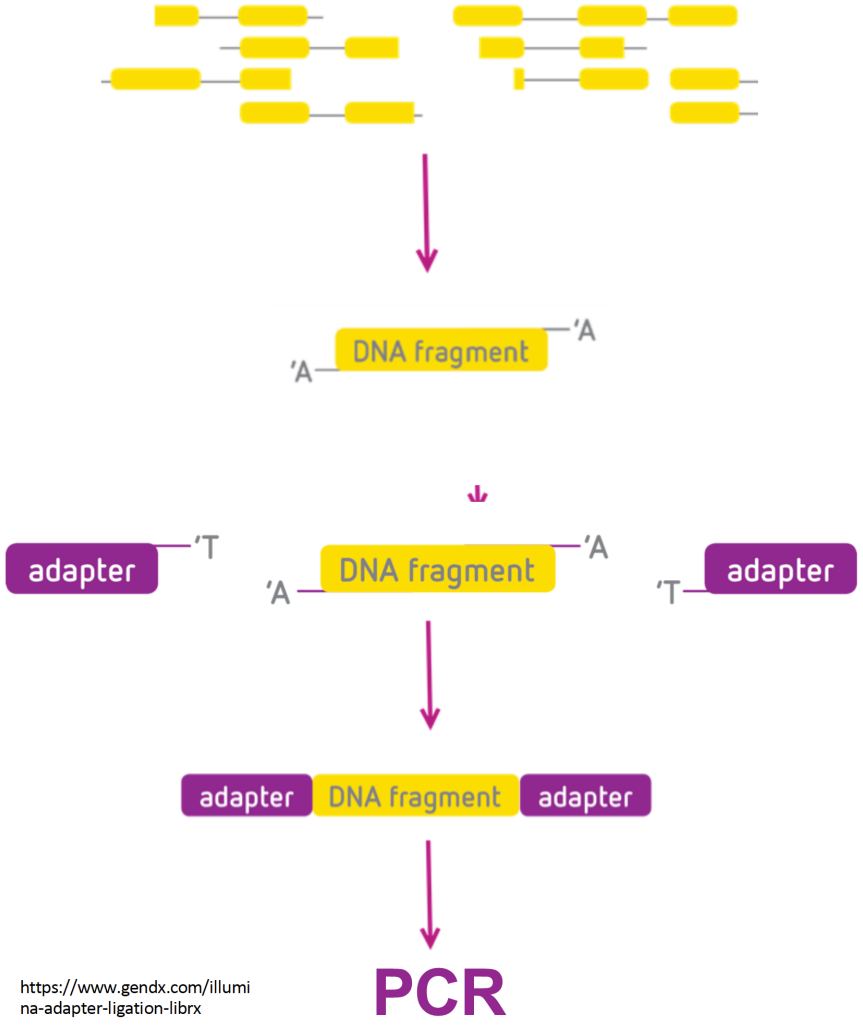
- Template-switching RT:



https://www.idtdna.com/pages/education/decoded/article/use-of-template-switching-oligos-(ts-oligos-tsos)-for-efficient-cdna-library-construction

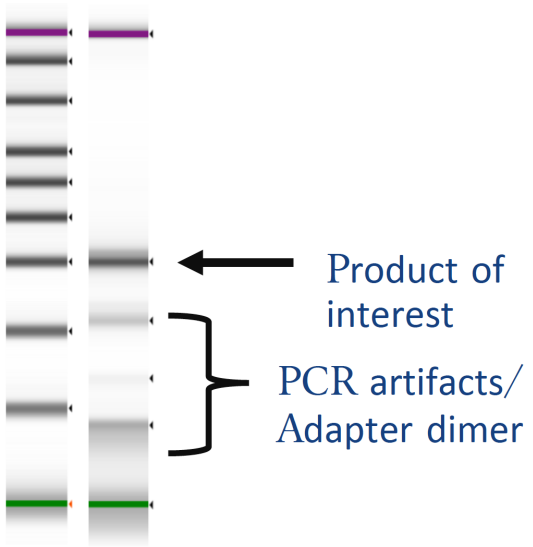
Slide courtesy of Jessica Warren

Adapter ligation, Amplification, Size Selection, Sequencing



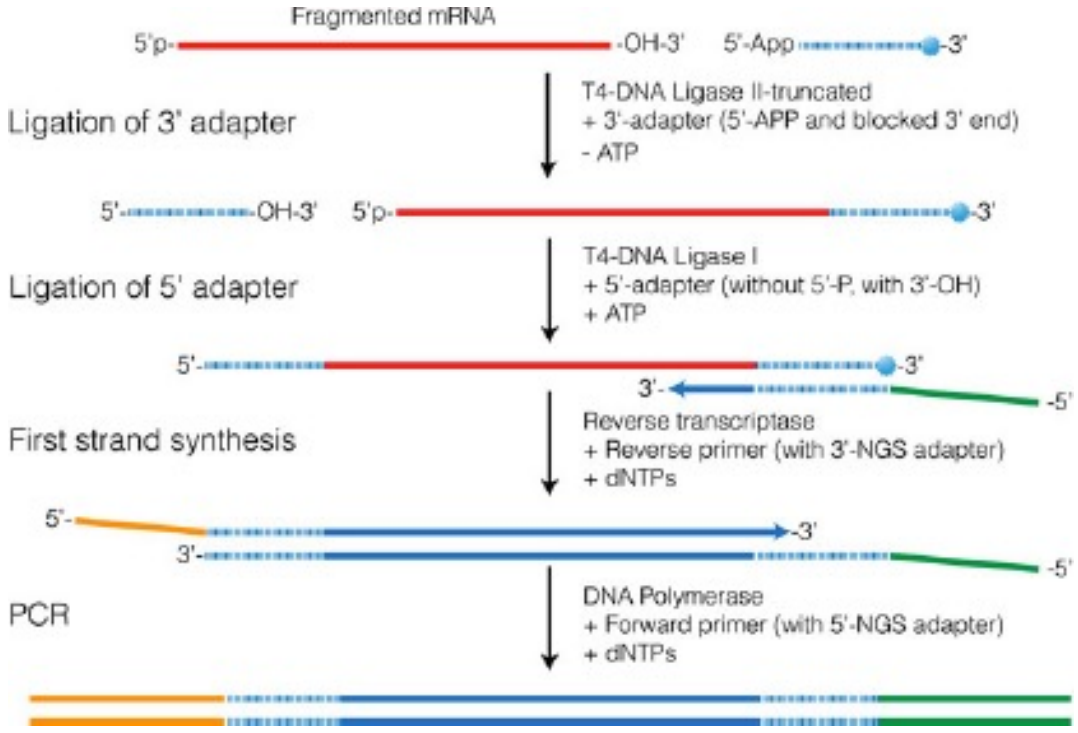
https://www.gendx.com/illumina-adapter-ligation-librx

- The ends of the DNA fragments are repaired to create blunt ends and then a dA-tail is added.
- The adapters are then ligated to the DNA fragment.
- PCR amplification: Normally 8–12 cycles are used during PCR.
- Size selection can be done to avoid sequencing PCR artifacts and adapter dimers.



Slide courtesy of Jessica Warren

Small RNA Library Construction



Adapters are ligated at the RNA stage

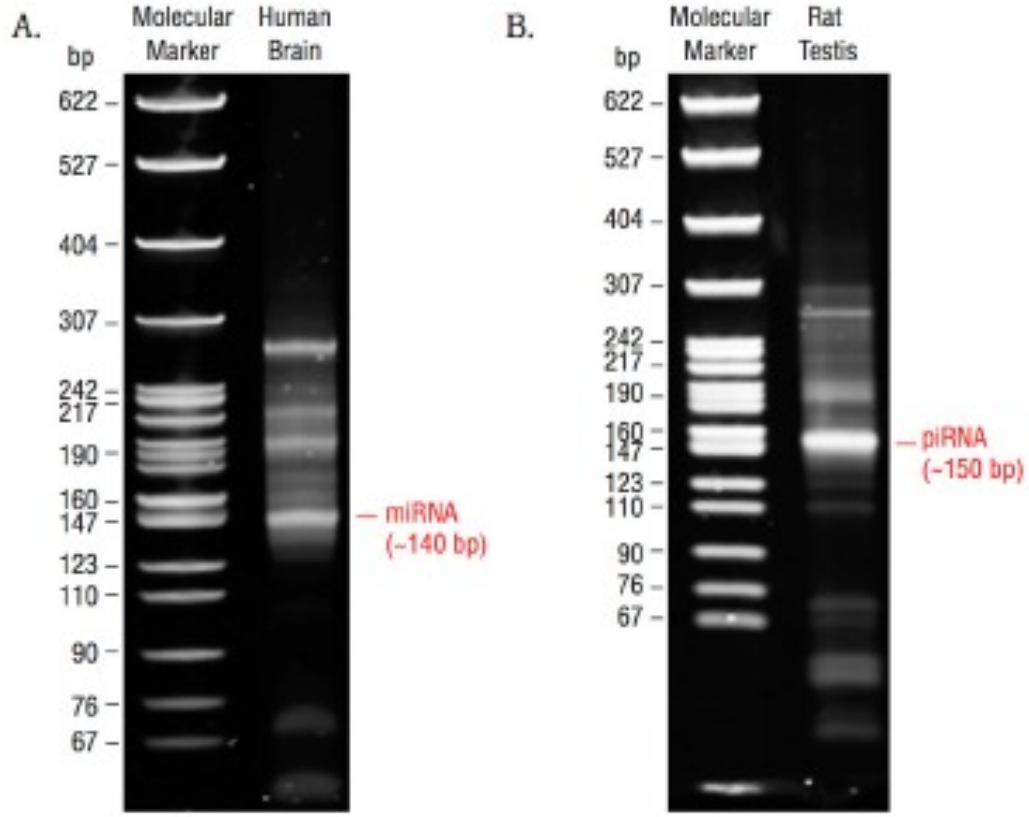
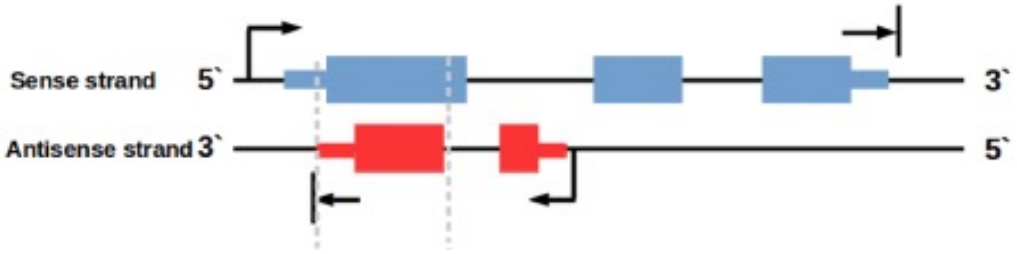


Figure 3: Shows typical results from Human Brain (A) and Rat Testis (B) Total RNA libraries. The 140 and 150 bp bands correspond to miRNAs (21 nt) and piRNAs (30 nt), respectively.

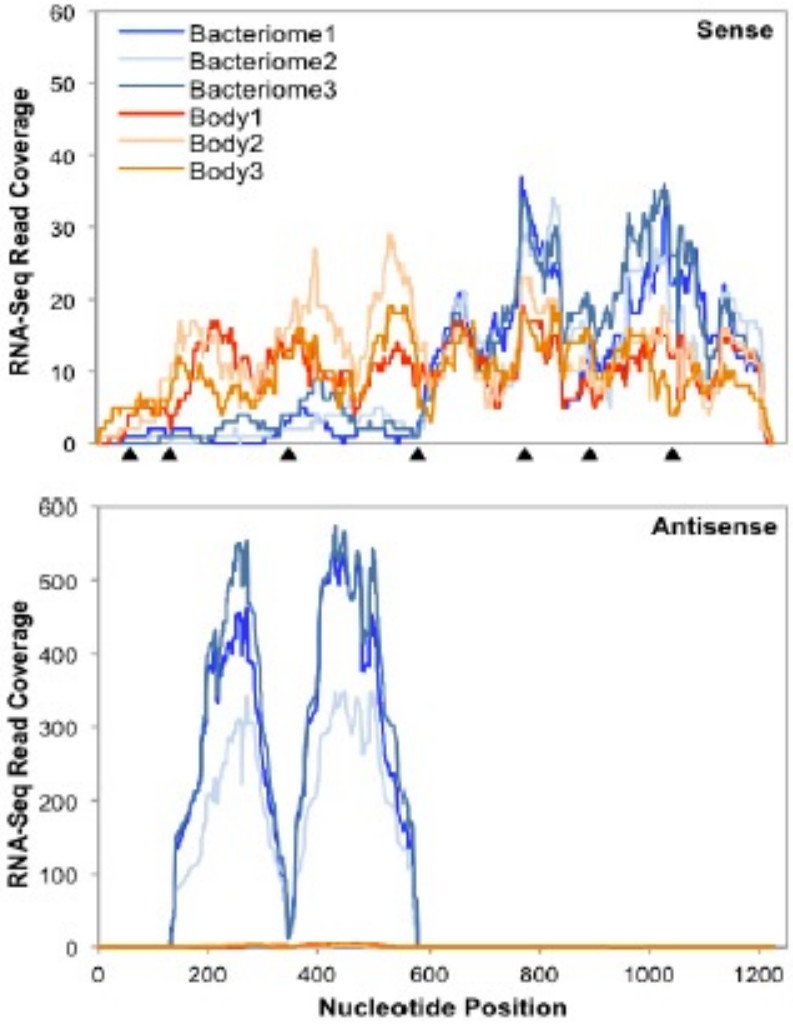
Slide courtesy of Jessica Warren

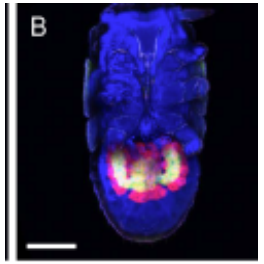
Strand-specific RNA sequencing (Do it!)

Sense vs. antisense matters!



Which strand is transcribed?





PMCID: PMC5730177

https://www.ecseq.com/

Slide courtesy of Jessica Warren

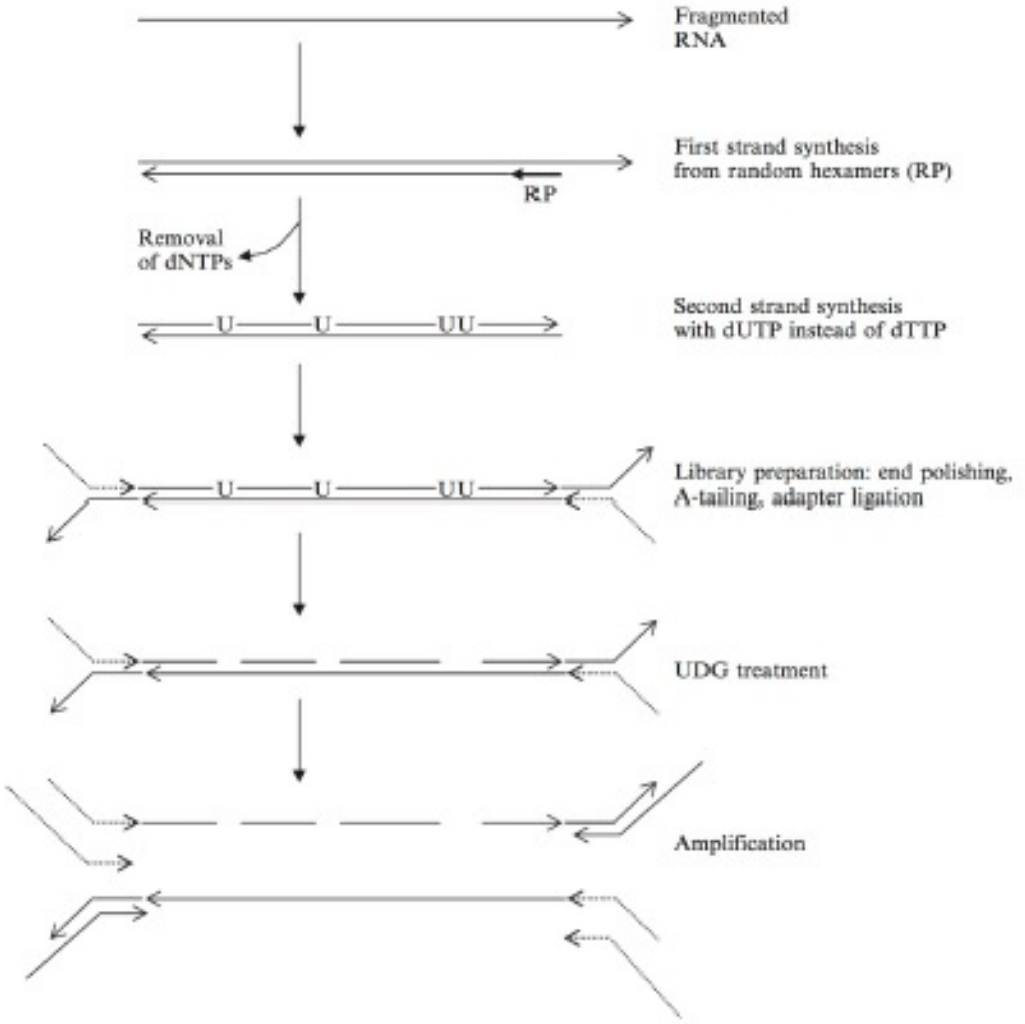
Strand-specific RNA sequencing (Do it!)

- Multiple alternative methods using RNA adapter ligation
  Levin et al. 2010 Nat Methods. 7:709–715

- Template switching, 3′ adapter sequence is added to the cDNA molecule, and first strand cDNA molecule can be PCR amplified directly without second strand synthesis.

- The “strand-marking” dUTP/UDG method has emerged as the standard.

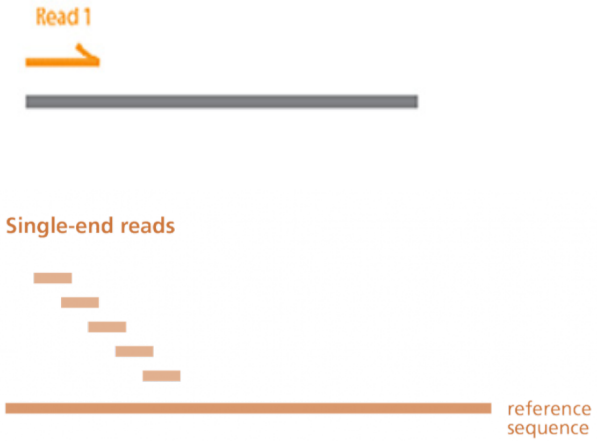
- If using an RNA-seq kit, check to make sure it directional/strand specific.



Slide courtesy of Jessica Warren

Paired-end vs. single-end sequencing

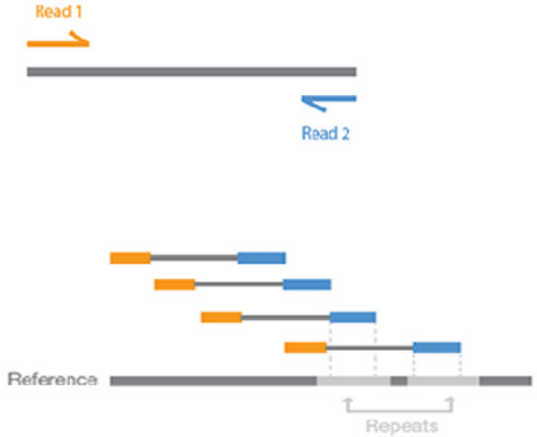
Single-end reads



Gene expression: # Independent reads

More independent reads per dollar with single-read sequencing.

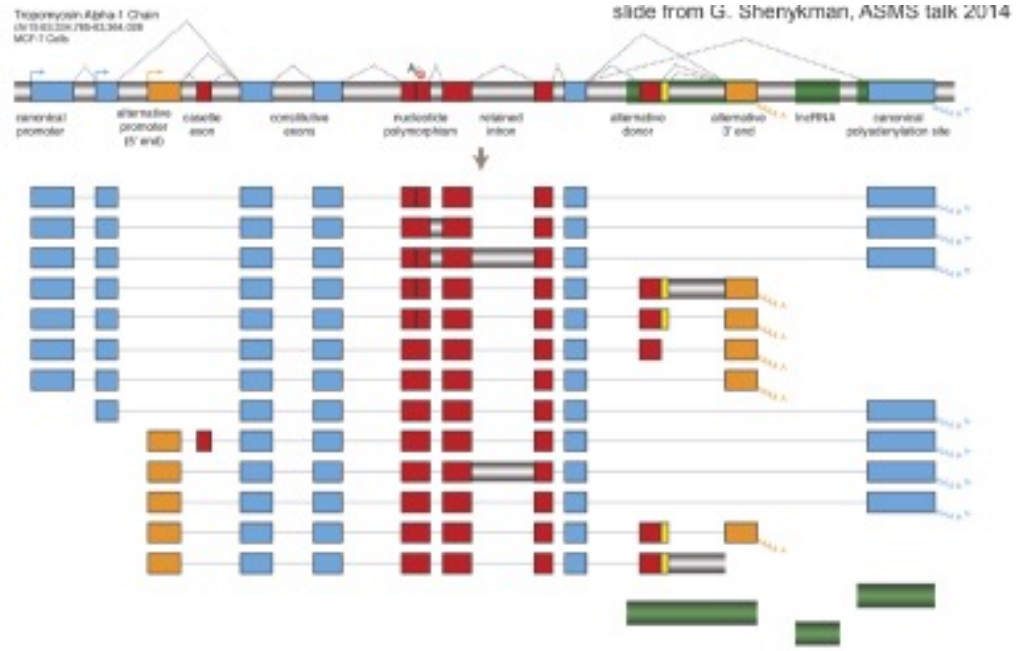
Paired-end reads

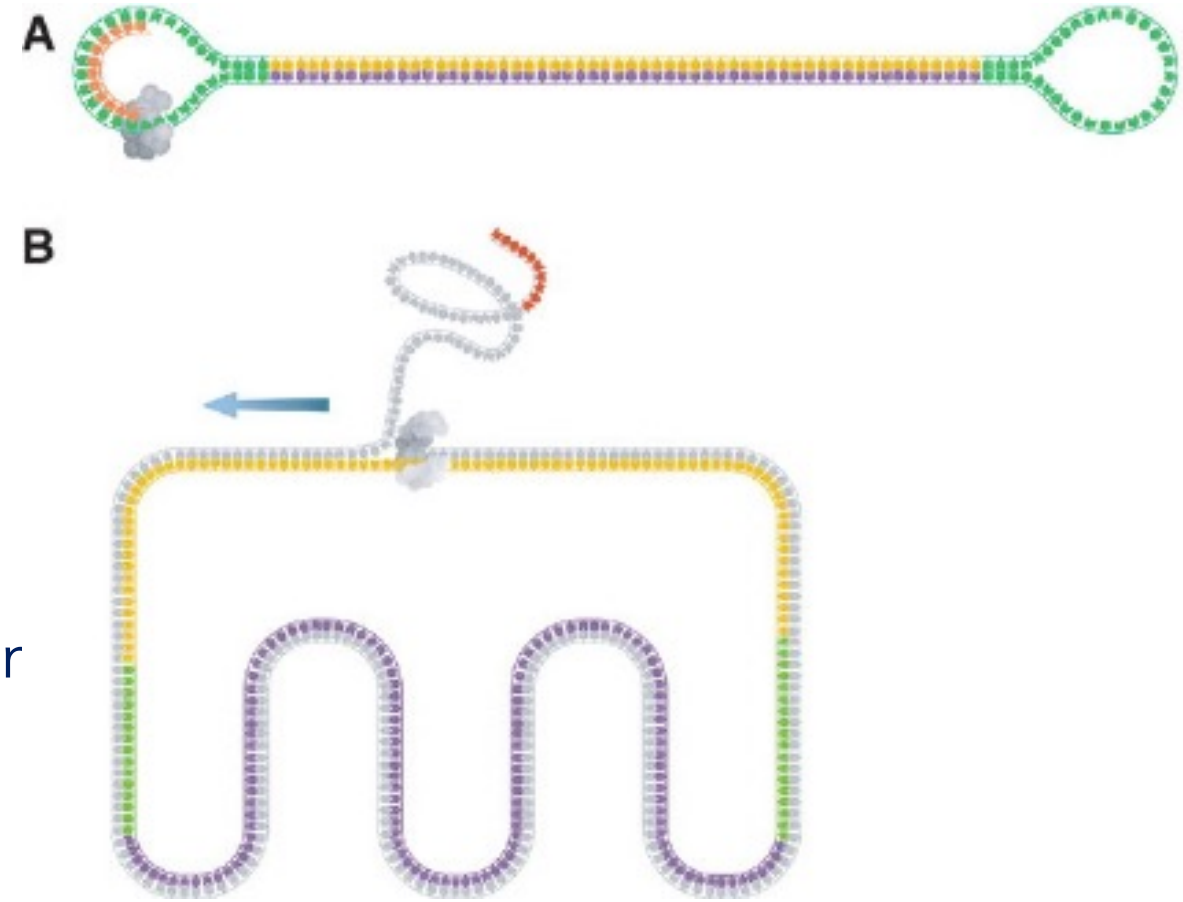


- De novo transcriptome assembly
- Mapping to repetitive regions

Slide courtesy of Jessica Warren

Single-molecule long-read sequencing (PacBio/Nanopore)





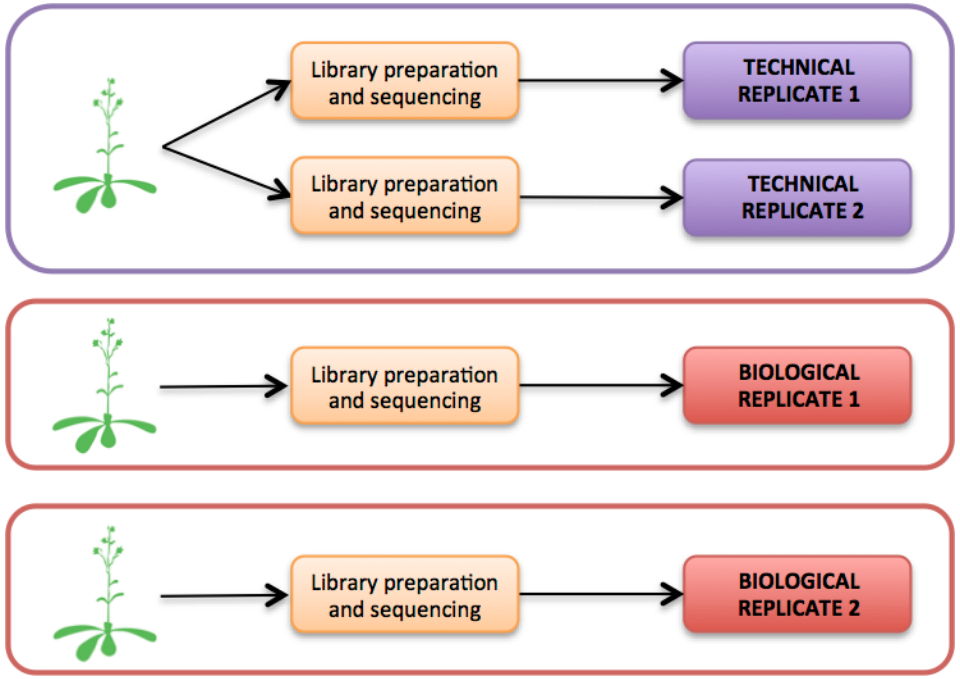
Generate full-length cDNA sequences — r assembly required — to characterize transcript isoform

Determine relative timing of different splicing and/or RNA editing events

Slide courtesy of Jessica Warren

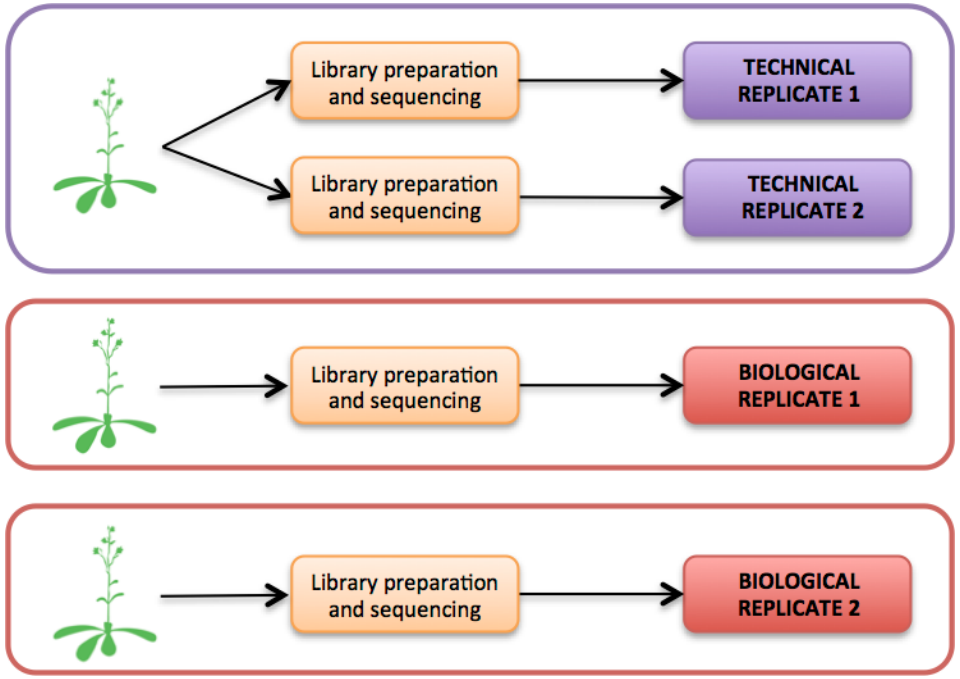
Replication in RNAseq experiments

Gene expression is tremendously variable, so your sampling scheme needs to reflect that



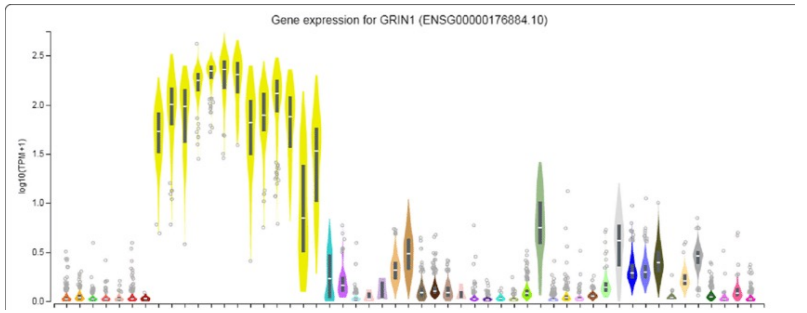
Replication in RNAseq experiments

Gene expression is tremendously variable, so your sampling scheme needs to reflect that

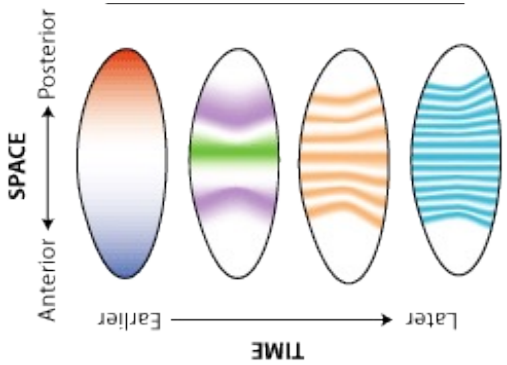


Maximize apples-to-apples comparisons

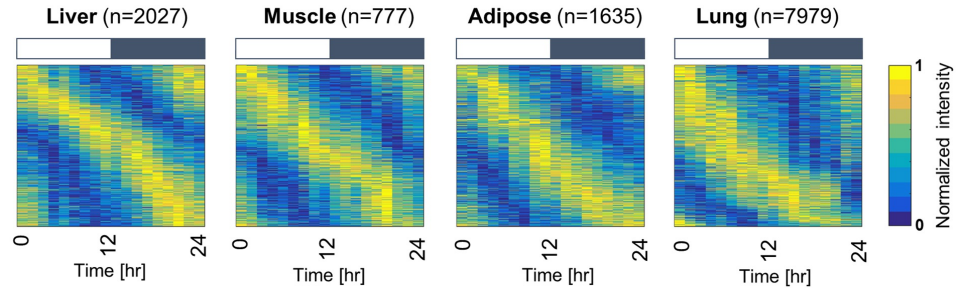
Tissue



Development

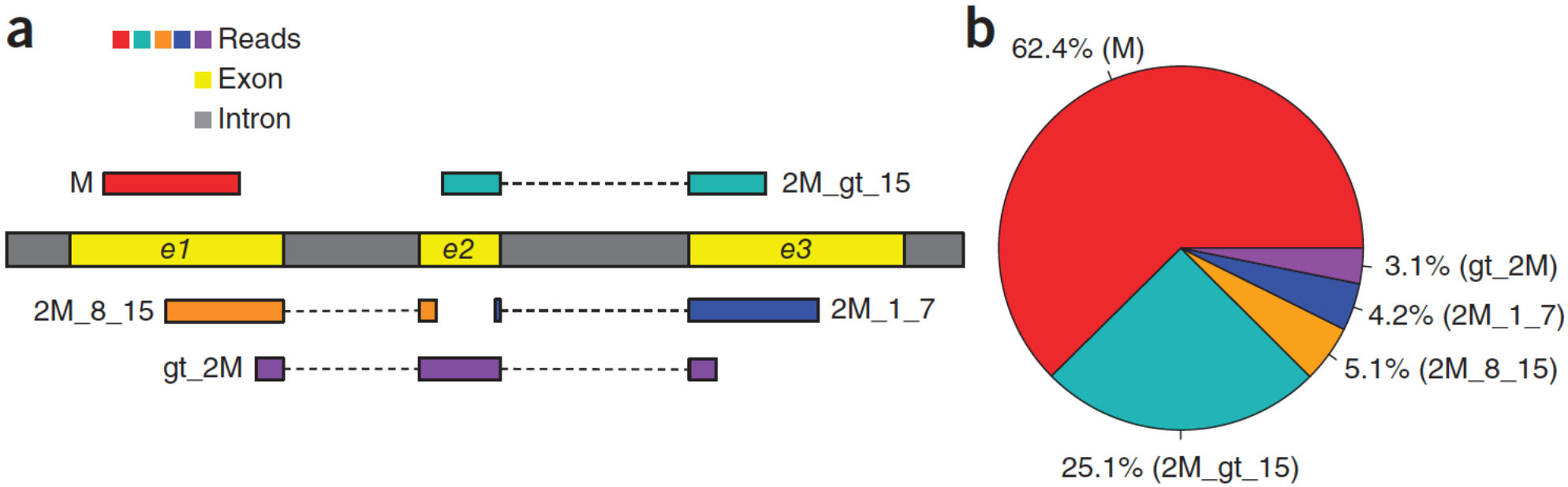


Time

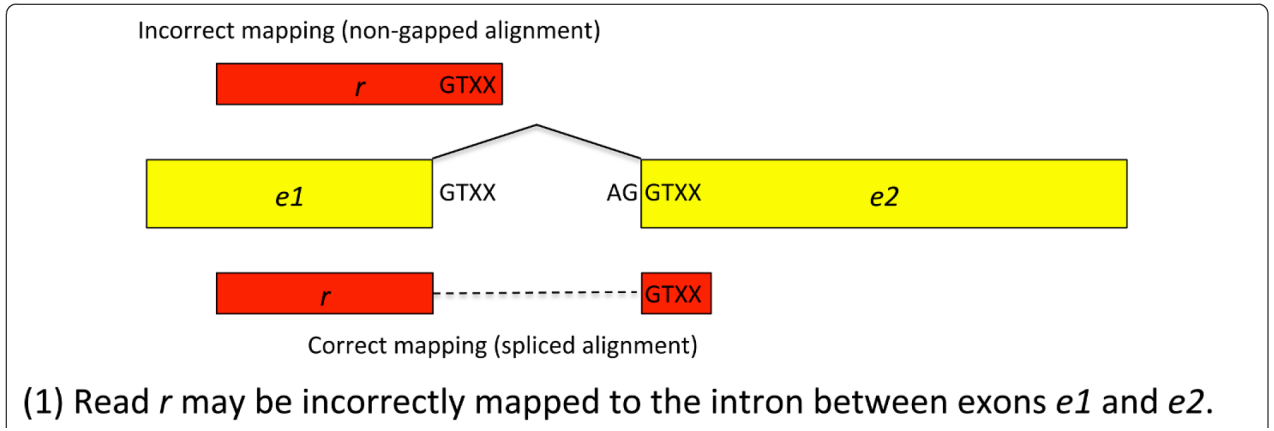


Splice-aware alignment

RNAseq reads can align to genomes in multiple ways



Aligning RNA to DNA is quite challenging!



(1) Read *r* may be incorrectly mapped to the intron between exons *e1* and *e2*.

Kim et al., 2013

Aligning RNA to DNA is quite challenging!

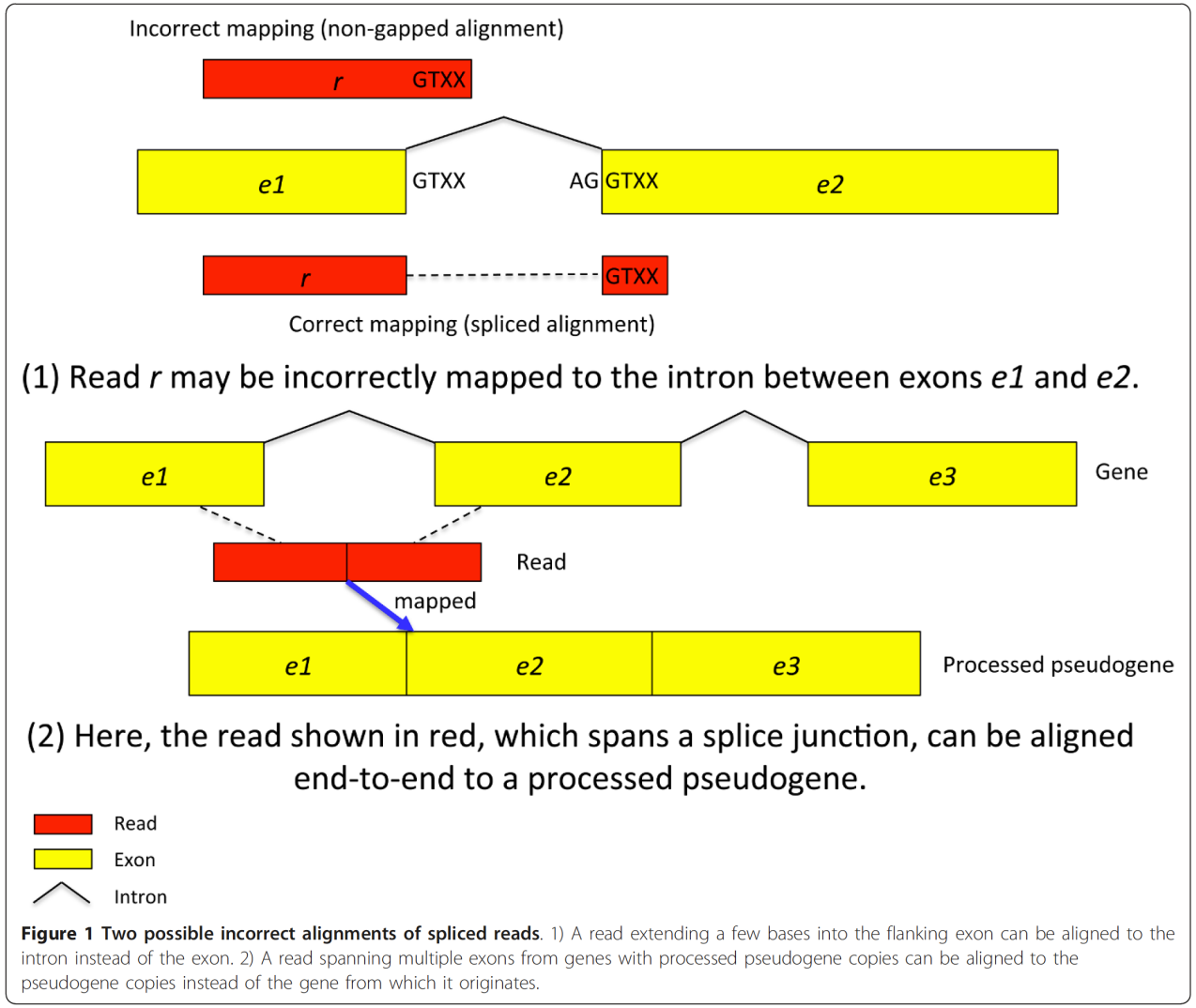


Figure 1 Two possible incorrect alignments of spliced reads. 1) A read extending a few bases into the flanking exon can be aligned to the intron instead of the exon. 2) A read spanning multiple exons from genes with processed pseudogene copies can be aligned to the pseudogene copies instead of the gene from which it originates.

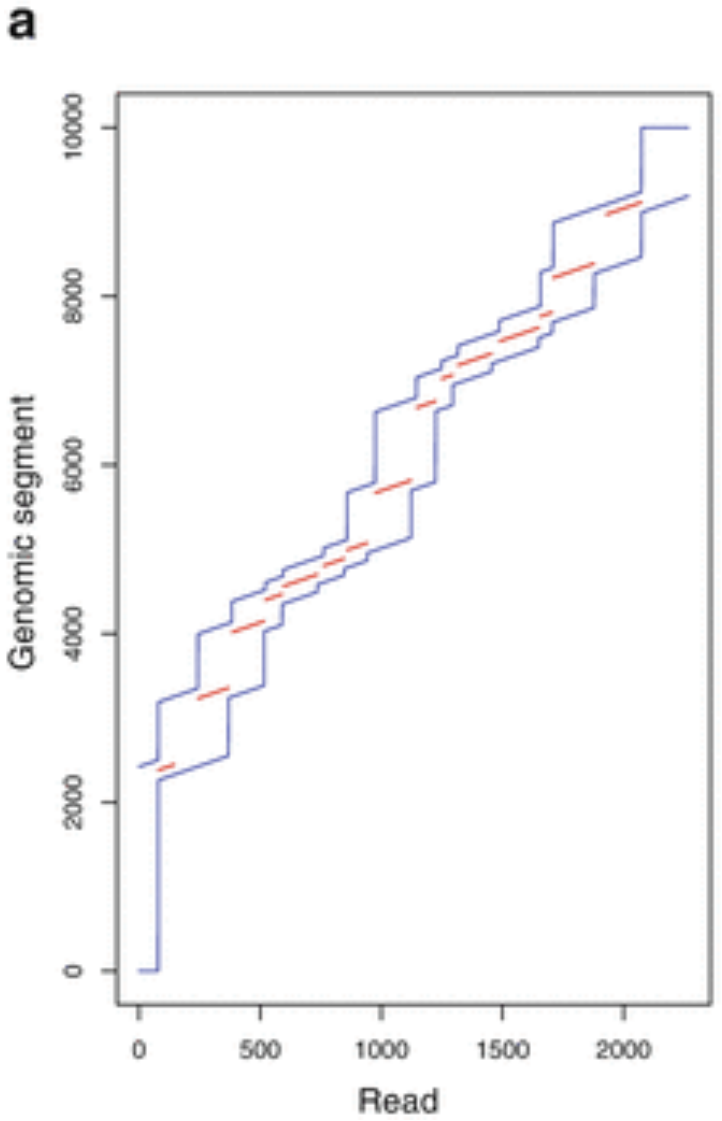
Kim et al., 2013

How to do it? GSNAP

GSNAP – Breaks genome into small pieces

- Map reads to genome using 12-mers
- Spanning set analysis
- Complete set analysis
- Segment combination

- Good for super-short reads (30-75bp), spanning at most one intron

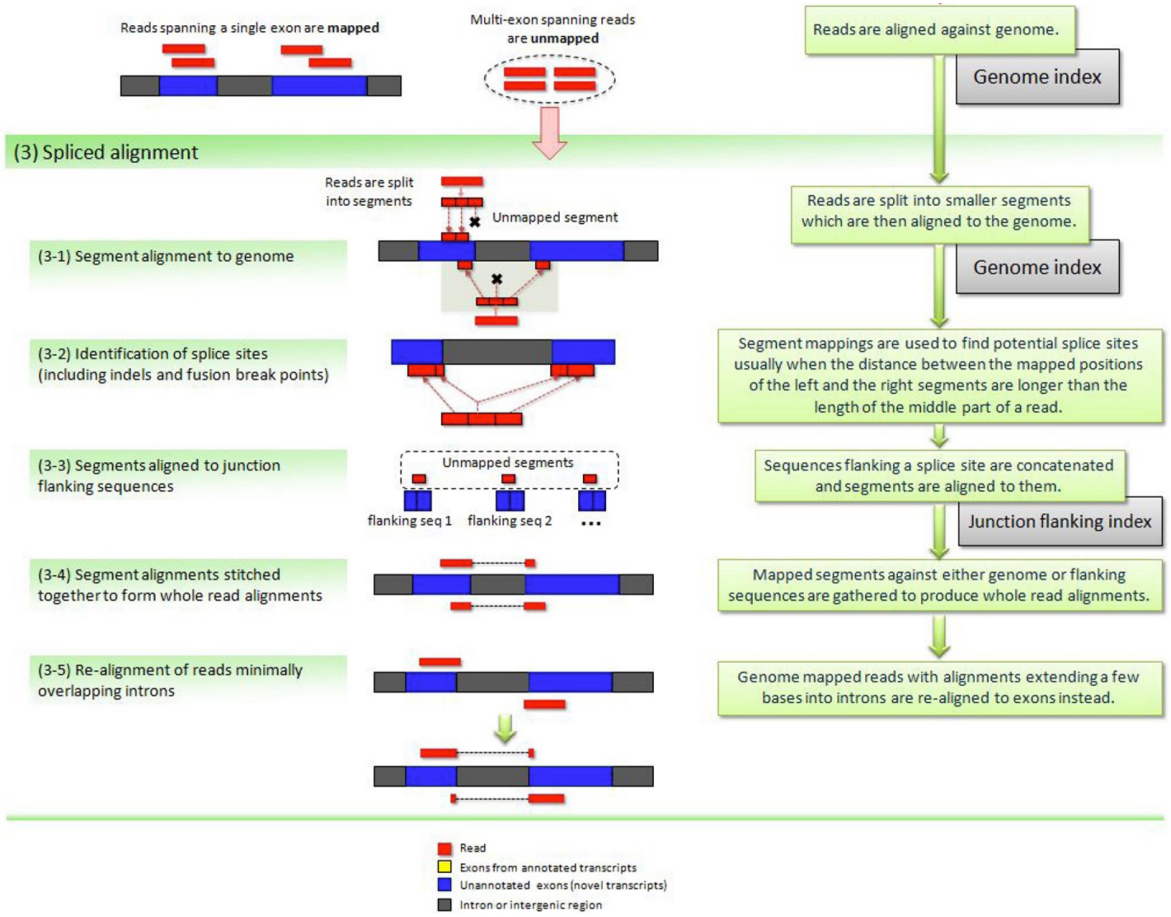


Wu et al., 2016

How to do it? Tophat

Tophat splits the reads in half

- Map reads to genome using BWT
- Split reads that map but not perfectly (i.e., clipped alignments)
- Re-map split reads
- Repeat iteratively until splice junctions identified
- Stich reads back together across identified splice junctions
- Look for minimally overlapping reads
- Good for large datasets, not most accurate



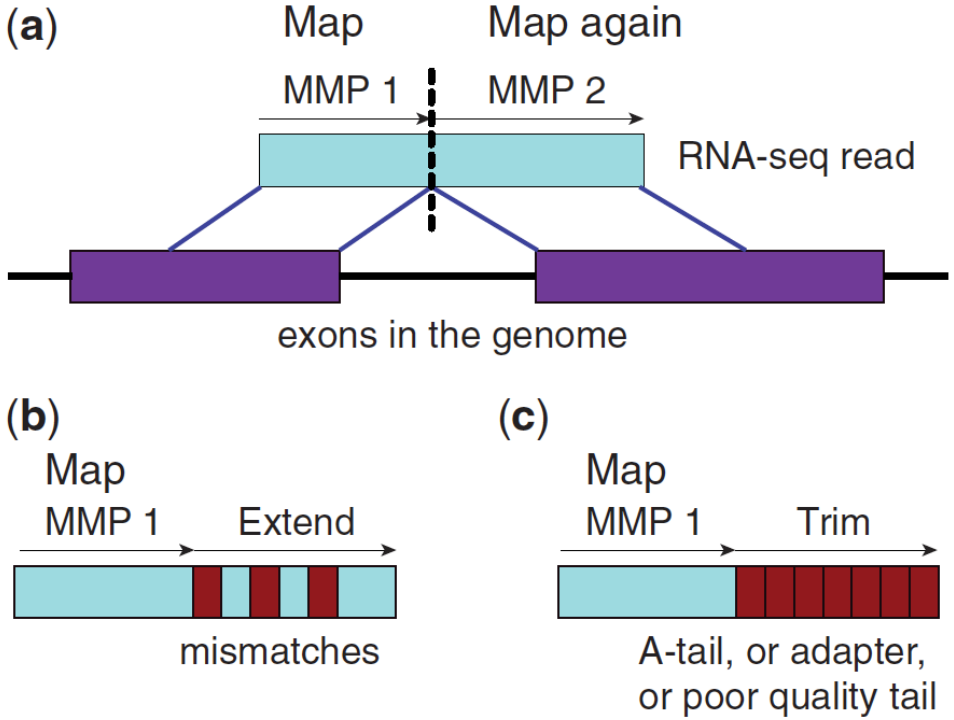
Kim et al., 2013

How to do it? STAR

STAR aligns iteratively

- Map reads to genome using suffix arrays
- Identify Maximal Mappable Prefix (MMP)
- Map un-mapped portion of read
- Stich reads back together

- High memory requirement, super fast, very accurate



Dobin et al., 2013

How to do it? HISAT

HISAT –Breaks the reference into small overlapping pieces

- Construct genome index with small, overlapping fragments (24kb in length)
- Map reads to fragments using BWT to identify candidate regions
- Extend alignment base by base against the whole genome

- Low memory requirement, super fast, very accurate

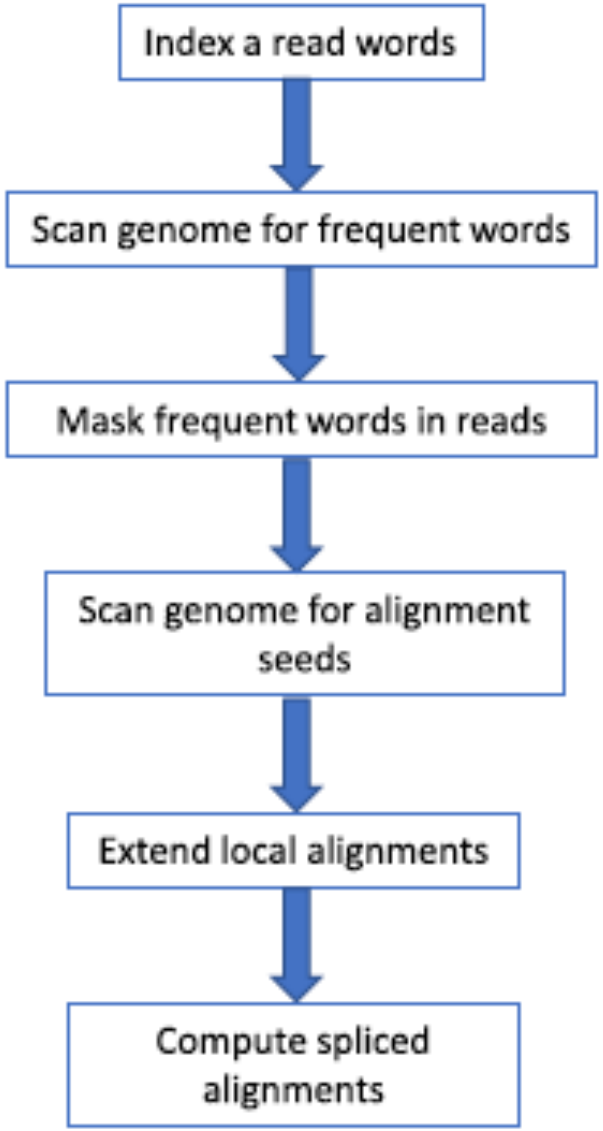
Kim et al., 2015

How to do it? Magi-BLAST

Magic-BLAST

- Uses BLAST-principles to perform the search
- Word search
- Seed extension
- Compute spliced alignments

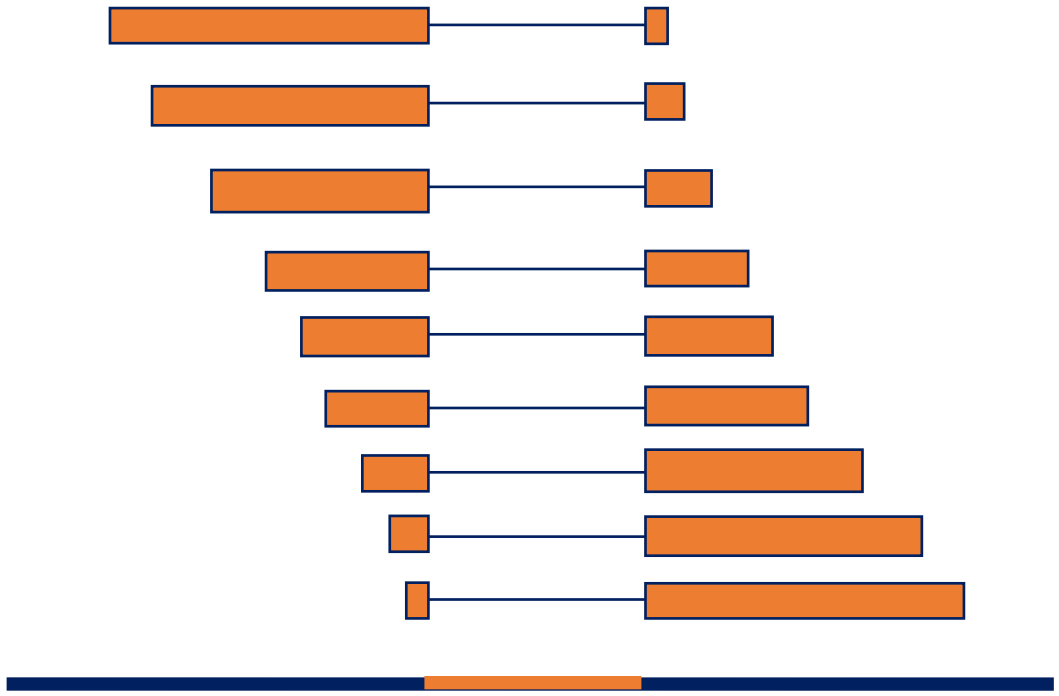
- Claims to be good for long reads



Boratyn et al., 2019

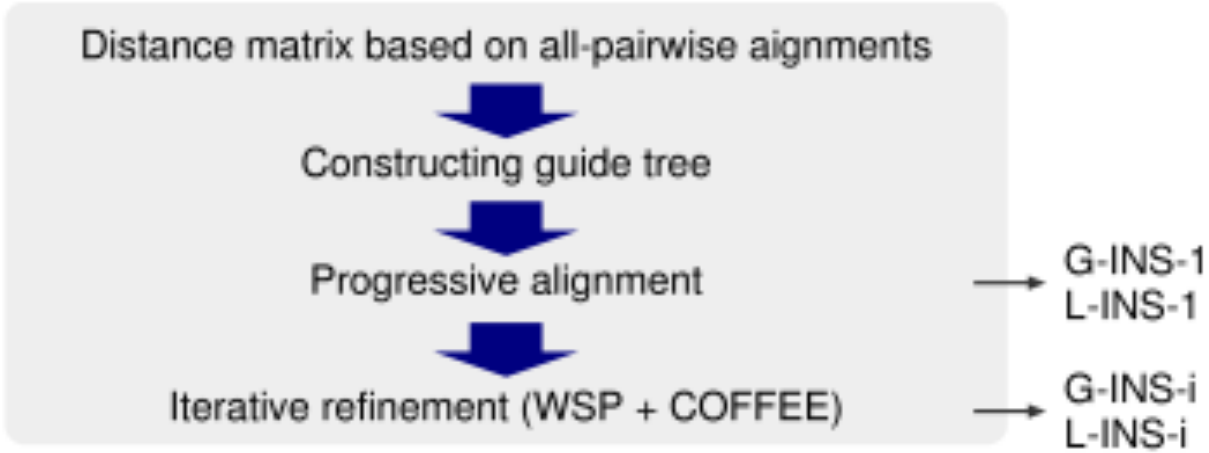
Spliced alignment CIGAR strings allow for intron identification

145M2000N5M
135M2000N15M
120M2000N30M
100M2000N50M
75M2000N75M
55M2000N95M
40M2000N110M
30M2000N120M
25M2000N125M



Multiple Sequence Alignment accounting for splicing

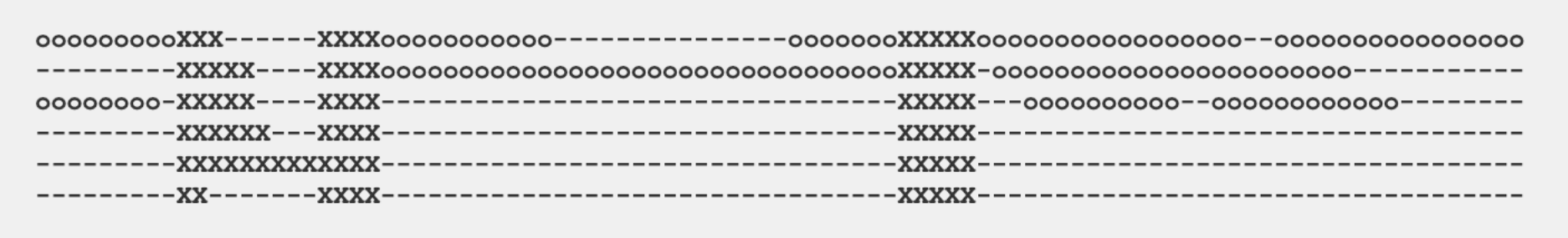
MAFFT



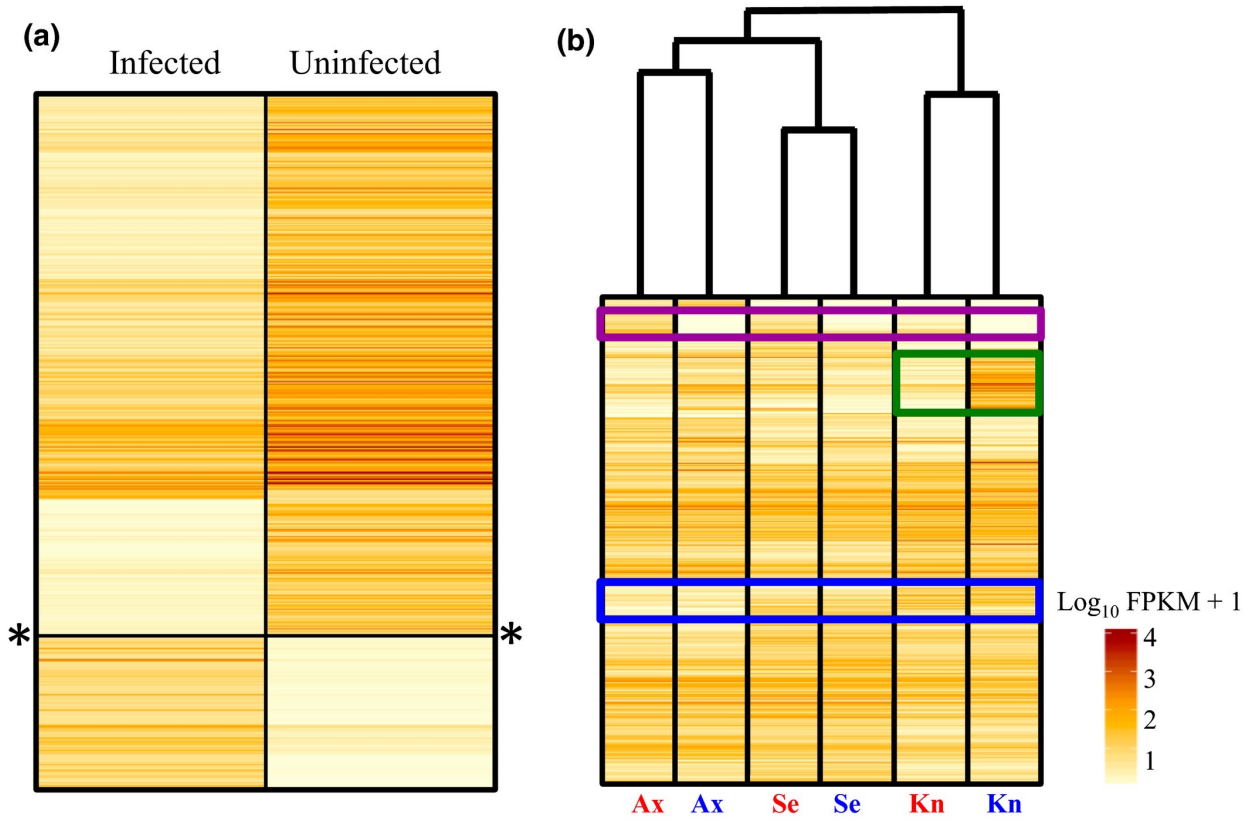

```
ooooooooXXX------XXXXoooooooooo--------------ooooooXXXXXooooooooooooooooo--ooooooooooooooo
---------XXXXX----XXXXoooooooooooooooooooooooooooooXXXXX-ooooooooooooooooooooo-----------
ooooooo-XXXXX----XXXX---------------------------------XXXXX---ooooooooo--ooooooooooo--------
---------XXXXXX---XXXX---------------------------------XXXXX-----------------------------------
---------XXXXXXXXXXXX---------------------------------XXXXX-----------------------------------
---------XX------XXXX---------------------------------XXXXX-----------------------------------
```

Multiple Sequence Alignment accounting for splicing

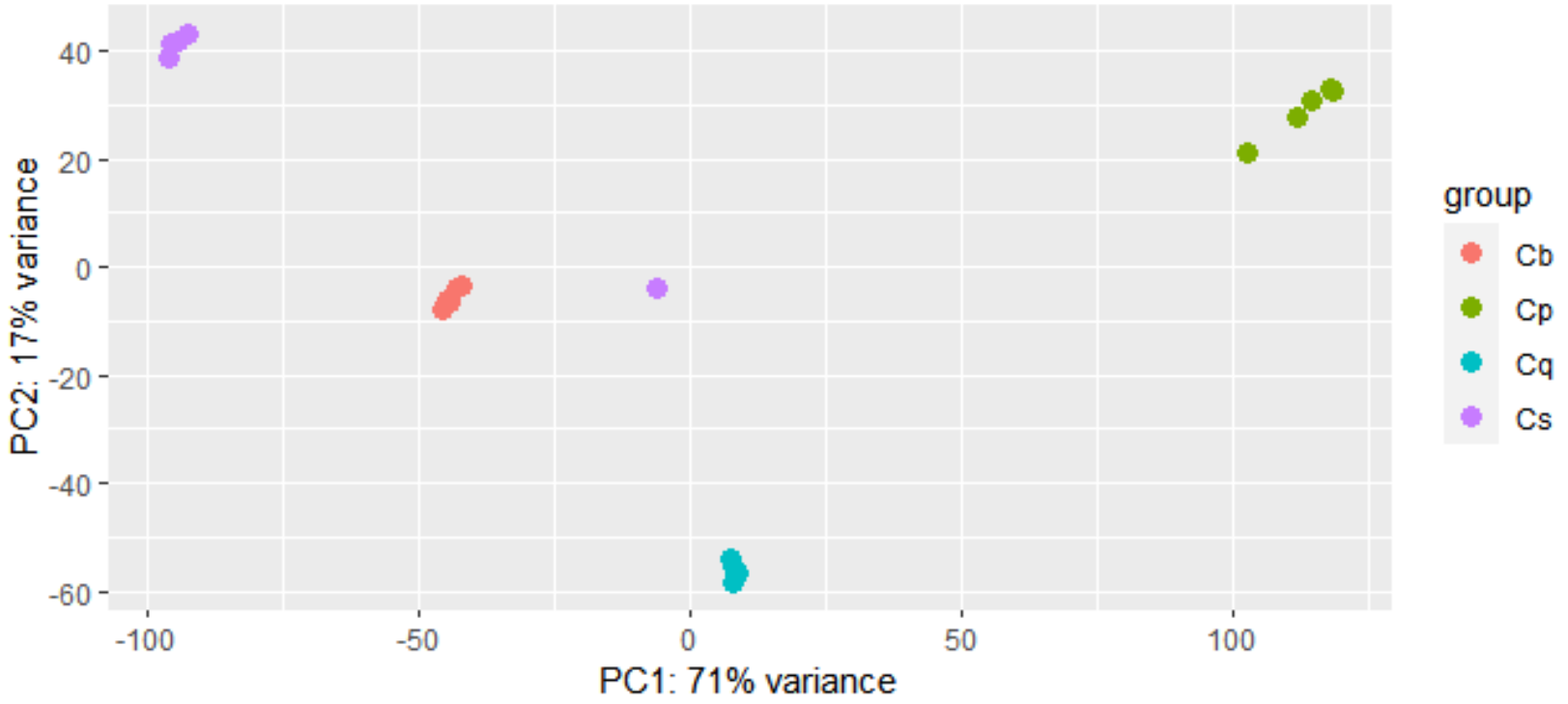
Identifying isoforms/alleles/paralogs from transcripts requires splice-aware MSA



Quantifying Expression



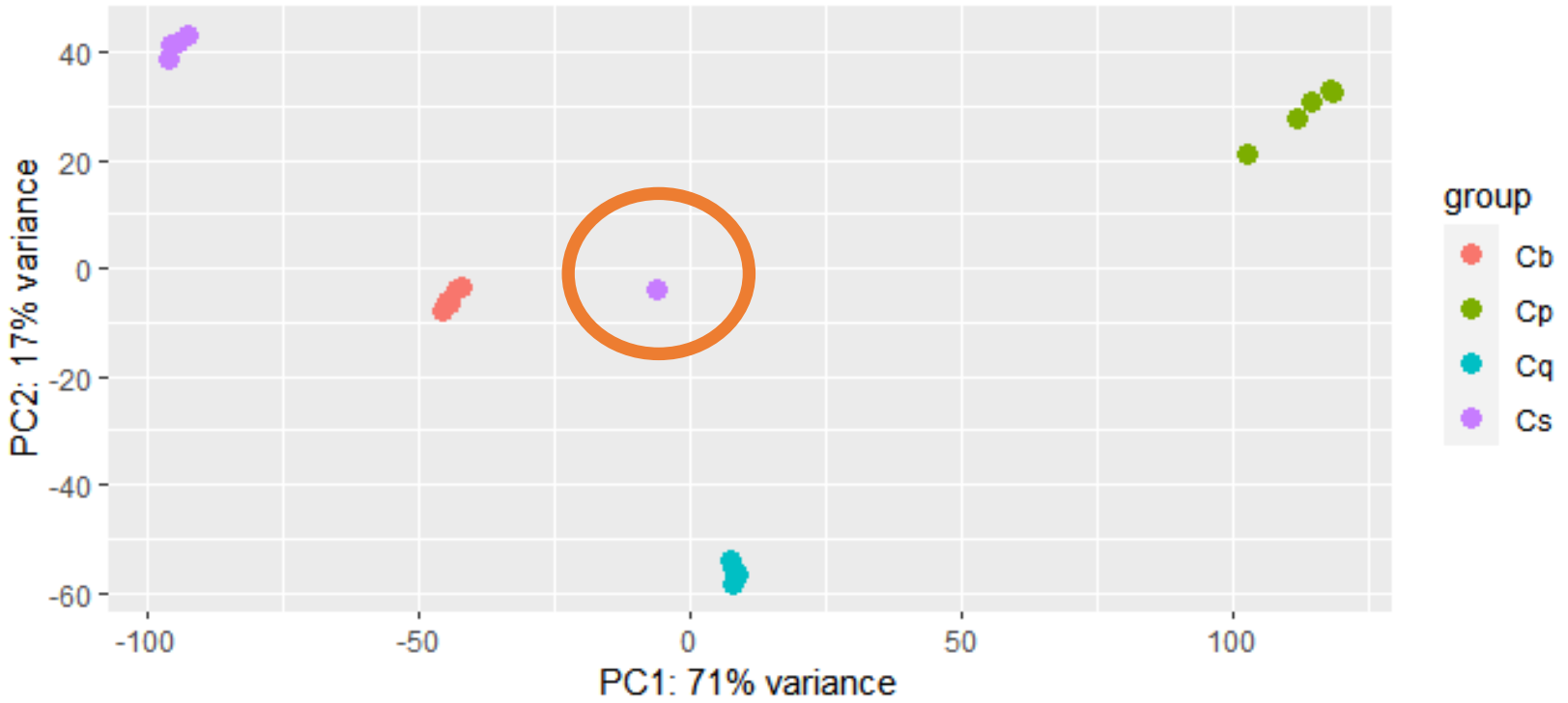
Visualize RNAseq data with PCA



Principle Components (PC) are linear representations of the data. Each successive PC is orthogonal to prior PCs, and allows for dimensionality reduction

pcaExplorer – Marini & Binder 2019. BMC Bioinfo.

Visualize RNAseq data with PCA



PCA can be used to check data for potential outliers/ bad sequencing runs

pcaExplorer – Marini & Binder 2019. BMC Bioinfo.

How to quantify gene expression?

1) Reads per kilobase mapped – RPKM (FPKM)
2) Transcripts per million – TPM

Reads per kilobase mapped (RPKM)

Need to account for the fact that longer transcripts get sequenced more often (all else being equal) than short transcripts

$$RPKM = \frac{C10^9}{NL}$$

C = Number of reads mapped to a transcript

N = Total number of reads

L = Length of the transcript

Mortazavi et al. 2008. Nat Methods

Fragments per kilobase mapped (FPKM)

Paired reads do not represent independent data points, so you need to combine them to get proper estimates

$$FPKM = \frac{C10^9}{FL}$$

C = Number of fragments mapped to a transcript

F = Total number of fragments = # read pairs

L = Length of the transcript

Mortazavi et al. 2008. Nat Methods

RPKM practice

1. Gene A is of length 4500 bp, and Gene B is 2400 bp. Both genes had 6000 single-end reads map to them. Calculate RPKM for both genes, assuming 70 Million total reads. Which gene is expressed at a higher level?

$$RPKM = \frac{C10^9}{NL}$$

2. In a second experiment, 80 Million paired-end reads were produced (i.e., 40 Million fragments). Of those, 5000 fragments mapped to Gene A and 2500 fragments mapped to Gene B. Calculate FPKM. Which gene is more highly expressed?

$$FPKM = \frac{C10^9}{FL}$$

Transcripts per million (TPM – fraction of transcripts)

- Expressed transcript length may vary between samples, making RPKM problematic for across-sample or across-species comparisons of the same gene
- RPKM assumes uniform sampling of reads across the transcript (i.e., average is a good measure of the central tendency)

$$TPM = 10^6 * \frac{\frac{C_i}{l_i}}{\sum_{j=0}^{n} \frac{C_j}{l_j}}$$

C = Number of reads/fragments mapped to a transcript

l = length of a transcript

TPM practice

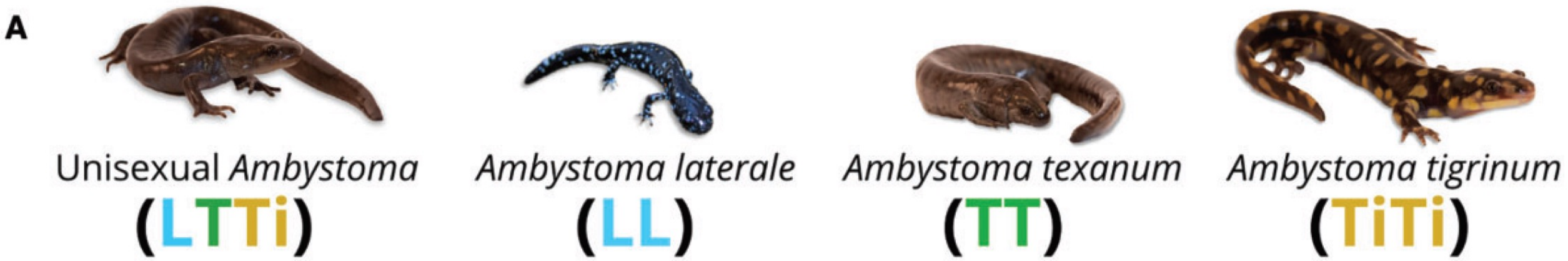
$$TPM = 10^6 * \frac{\frac{C_i}{l_i}}{\sum_{j=0}^{n} \frac{C_j}{l_j}}$$

1. Gene A is of length 4500 bp, and Gene B is 2400 bp. Both genes had 6000 single-end reads map to them. Calculate TPM for both genes, assuming 70 Million total mapped reads and a total transcriptome length of 90,000,000 bp. Which gene is expressed at a higher level?

2. In a second experiment, 80 Million paired-end reads mapped (i.e., 40 Million fragments mapped). Of those, 5000 fragments mapped to Gene A and 2500 fragments mapped to Gene B. Calculate TPM. Which gene is more highly expressed?

Packages to calculate Gene Expression

- RSEM
- edgeR
- DESeq



Statistical significance of differentially expressed transcripts

Volcano plots can help visualize differential expression