Introduction to RNA-Seq

Monica Britton, Ph.D.
Bioinformatics Analyst

September 2014 Workshop

Overview of Today's Activities

Morning

- RNA-Seq Concepts, Terminology, and Work Flows
- Two-Condition Differential Expression (Single and Paired-End)
- Guest Speaker: Dr. Stephen Pearce

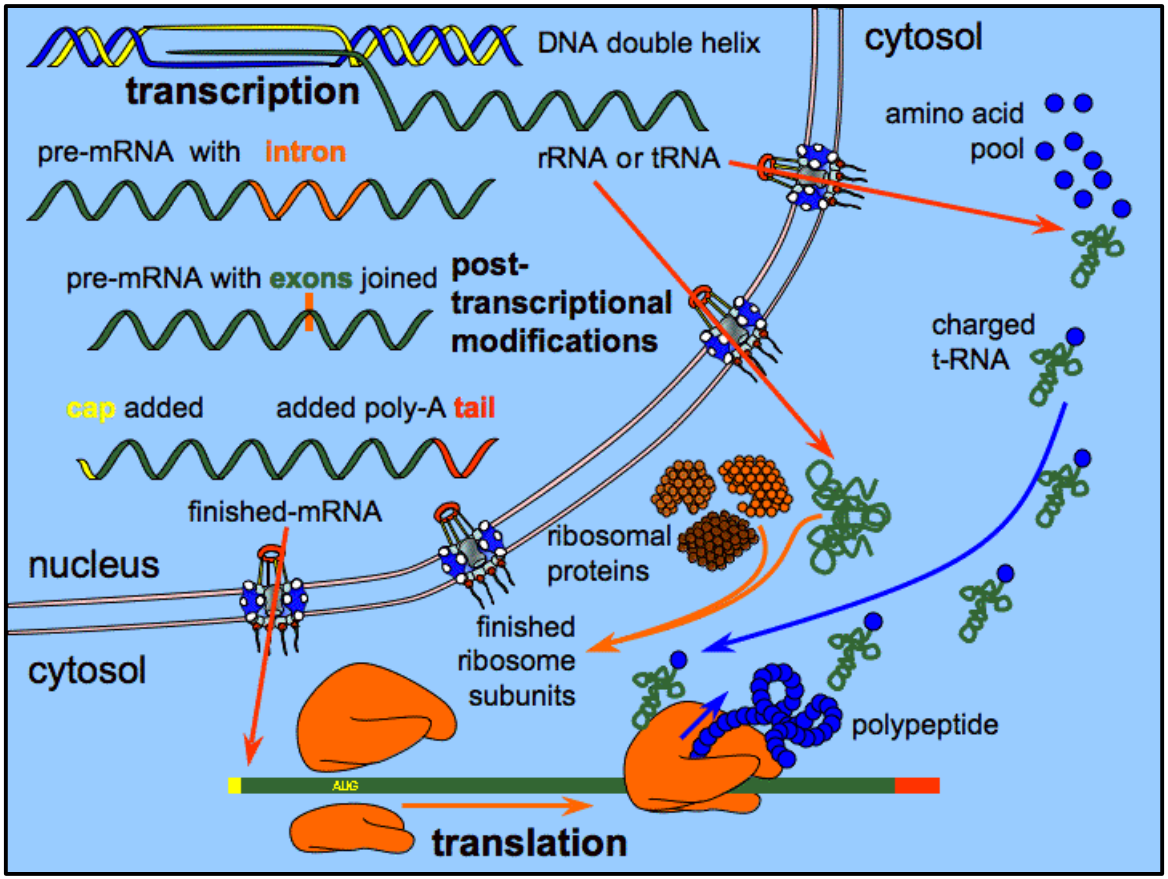
Afternoon

- Gene Construction with Paired-End Reads
- RNA-Seq Statistics (Blythe Durbin-Johnson)
- Alignment to a Reference Transcriptome

Now that you're adept at running bioinformatics software, you'll be doing the exercises "on your own". Don't worry if you can't finish them all today

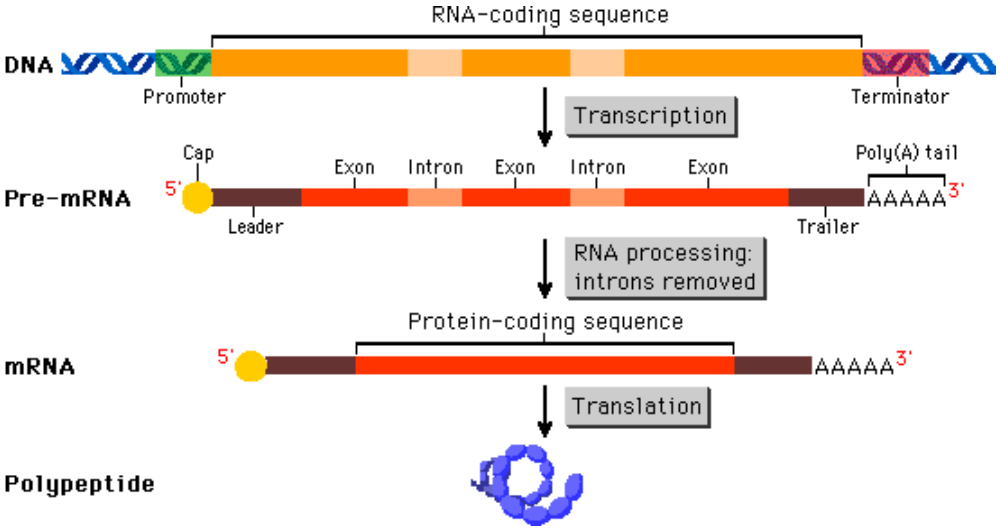
RNA Transcription and Processing

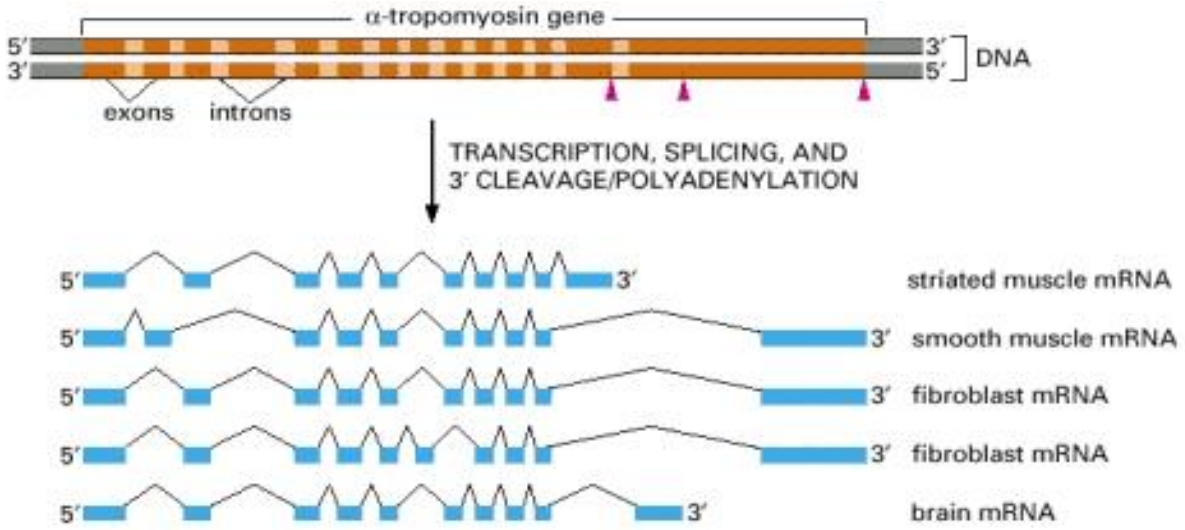
A cell contains many types of RNA (rRNA, tRNA, mRNA, miRNA, lncRNA, snoRNA, etc.) – Only ~2% is mRNA



Koning, Plant Physiology Information Website

Gene Structure and Alternative Splicing



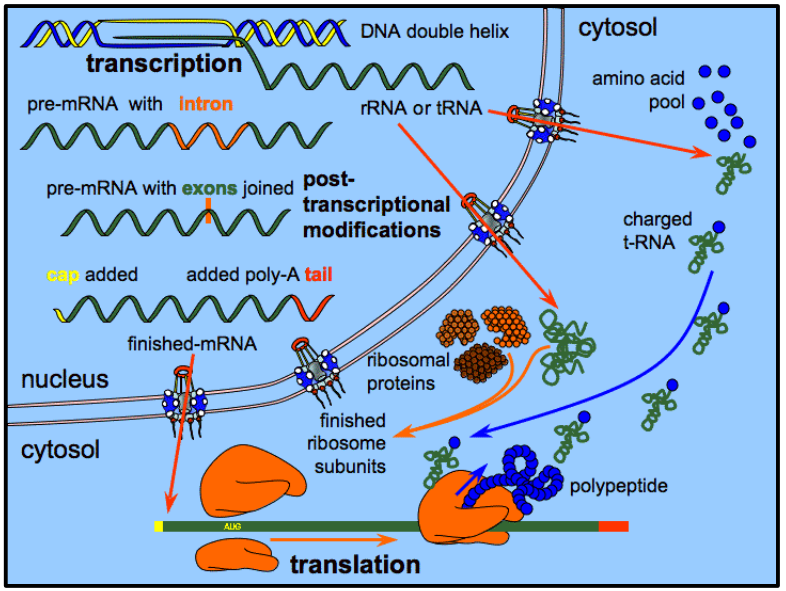


Molecular Biology of the Cell, 4th ed.

Some mRNA-Seq Applications

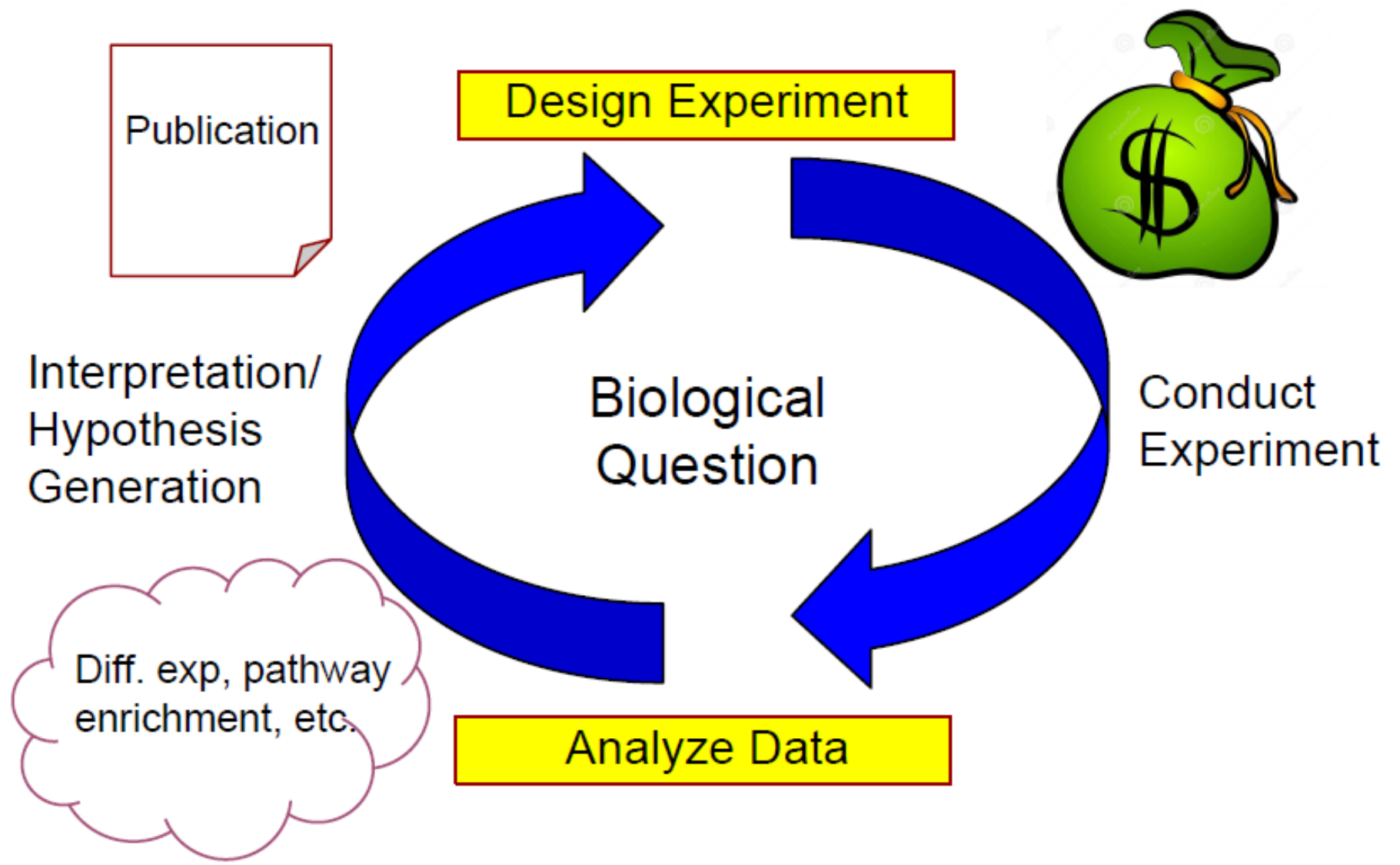
- Differential gene expression analysis
- Transcriptional profiling

Assumption:
Changes in transcription/ mRNA levels correlate with phenotype (protein expression)



- Identification of splice variants
- Novel gene identification
- Transcriptome assembly
- SNP finding
- RNA editing

The Circle of Research



Experimental Design

- What biological question am I trying to answer?
- What types of samples (tissue, timepoints, etc.)?
- How much sequence do I need?
- Length of read?
- Platform?
- Single-end or paired-end?
- Barcoding?
- Pooling?
- Biological replicates: how many?
- Technical replicates: how many?
- Protocol considerations?

What Is the Goal of the Experiment?

Many biological questions, such as…

"Characterize the differences between the wild-type and mutant"

are broad and open-ended.

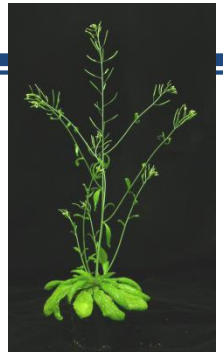
Such RNA-Seq experiments can be used to generate hypotheses, help form a more-focused question for the next experiment.

Make sure your experimental approach is suitable for the question you're asking. (You will not find mutations in non-transcribed regions with RNA-Seq.)

Influence of the Organism





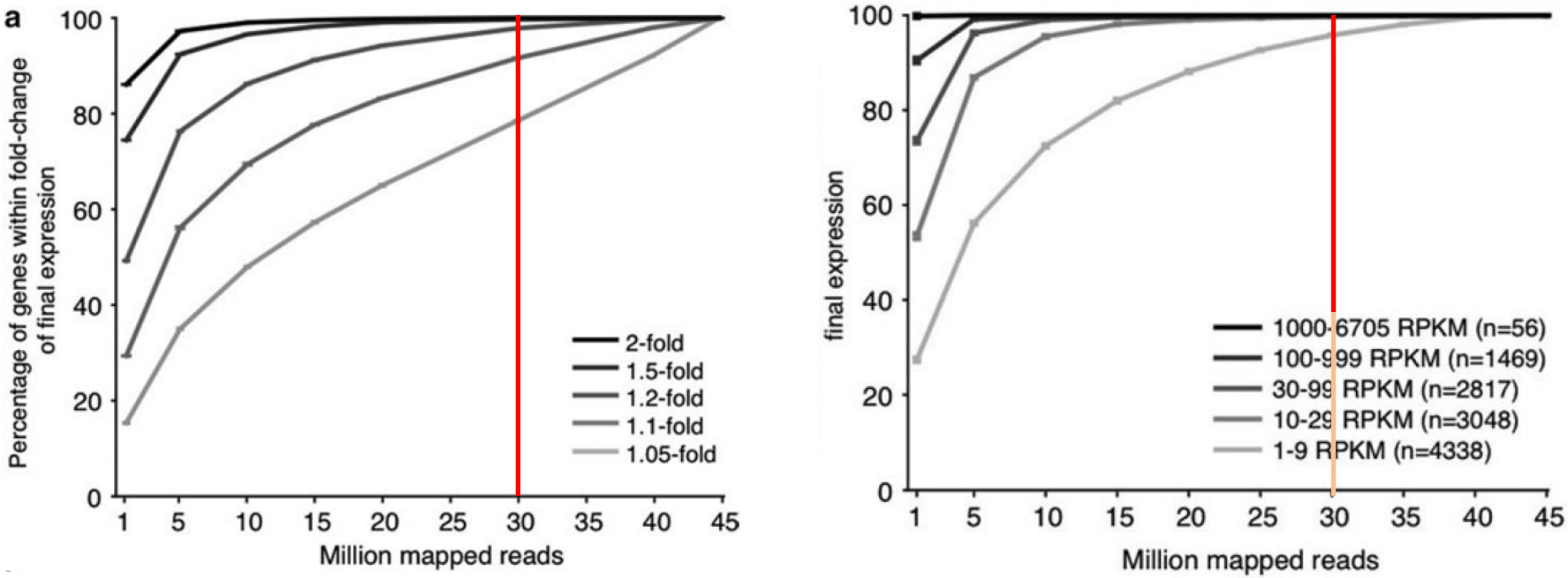



- **Novel** – little/no previous sequencing
- **Non-Model** – some sequence available (ESTs, Unigene set)
- **Genome-Sequenced** – draft genome
  - Thousands of scaffolds, maybe tens of chromosomes
  - Some annotation (*ab initio*, EST-based, etc.)
- **Model** – genome fully sequenced and annotated
  - Multiple genomes available for comparison
  - Well-annotated transcriptome based on experimental evidence
  - Genetic maps with markers available
  - Basic research can be conducted to verify annotations (mutants available)

Amount of Sequence

- Differential gene expression, reads/sample
  - Eukaryotes: 30+ million recommended
  - Bacteria: 10+ million recommended
- More sequence is needed to detect rare transcripts

Measures of Robustness of Expression Levels vs. Sequencing Depth

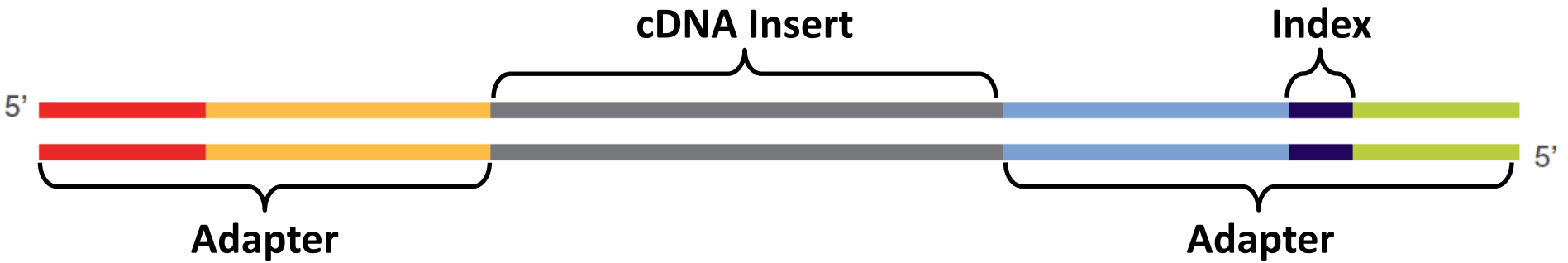


Ramskold, et al., 2012

Platform and Read Length Options

Read Length	Platform	Applications
40+ SE	Illumina (SOLiD)	Gene expression quantitation SNP-finding
40+ PE	Illumina Ion Proton	Better specificity for the above Splice variant identification
100+ PE	Illumina Ion Proton	All the above and: Differentiation within gene familes/paralogs Transcriptome assembly
200-300 400-600 400-800 5000 avg 10kb+?	Ion Torrent Sanger (454) PacBio (Oxford Nanopore)	Splice variant identification Transcriptome assembly Resolve haplotypes (phasing) Not recommended for gene expression quantitation

Multiplexing



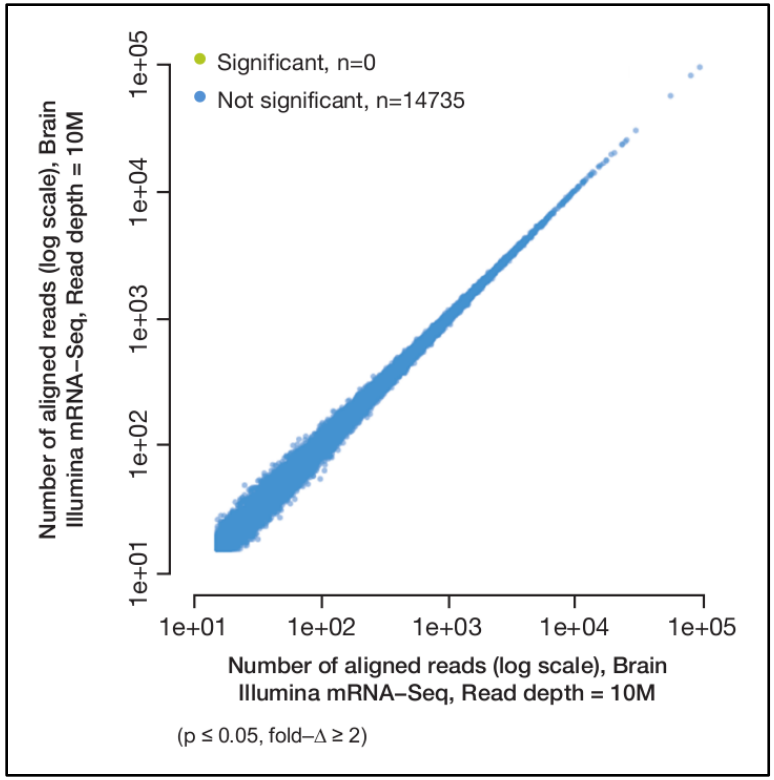
- Short (6-8 nt), unique barcodes (index) introduced as part of adapters
- Provide unique identifier for each sample
- Barcodes should be tolerant of 1-2 sequencing errors
- Barcodes allow deconvolution of samples
- Allows pooling samples to mitigate lane effects
- Allows sequencing capacity to be used efficiently
- Dual barcodes allow deep multiplexing (e.g., 96 samples)

Biological Replicates

- Allow measurement of variation between individuals/samples
- More are better (up to a point)
- Genetic Variation/Hetrozygosity:
  - Is each individual a different genotype?
  - Are individuals highly inbred or clonal?
  - Haploid or diploid or polyploid?
- Pooling with barcodes – each sample is a replicate
- Pooling without barcodes – each pool is a replicate
  - Validation on individual samples

Technical Replicates

- Account for variation in preparation
- Cost can be prohibitive
- Better to do more biological replicates
- Barcoding/pooling samples across multiple lanes
  - Recommended to even out lane effects
  - Allow data processing even if one lane fails



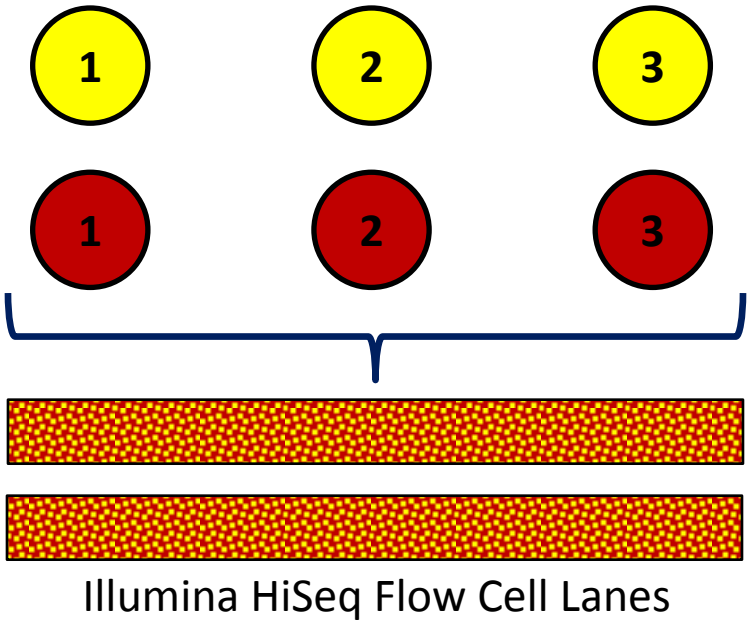
Example

- This experimental design has biological replicates and is multiplexed to mitigate lane effects
- Each sample will generate, on average, 50-60 million reads.

Control: 3 biological replicates

Treated: 3 biological replicates

Each sample is individually barcoded; all samples are pooled and run in two HiSeq lanes



Illumina HiSeq Flow Cell Lanes

mRNA-Seq Protocol Overview

TOTAL RNA

↓

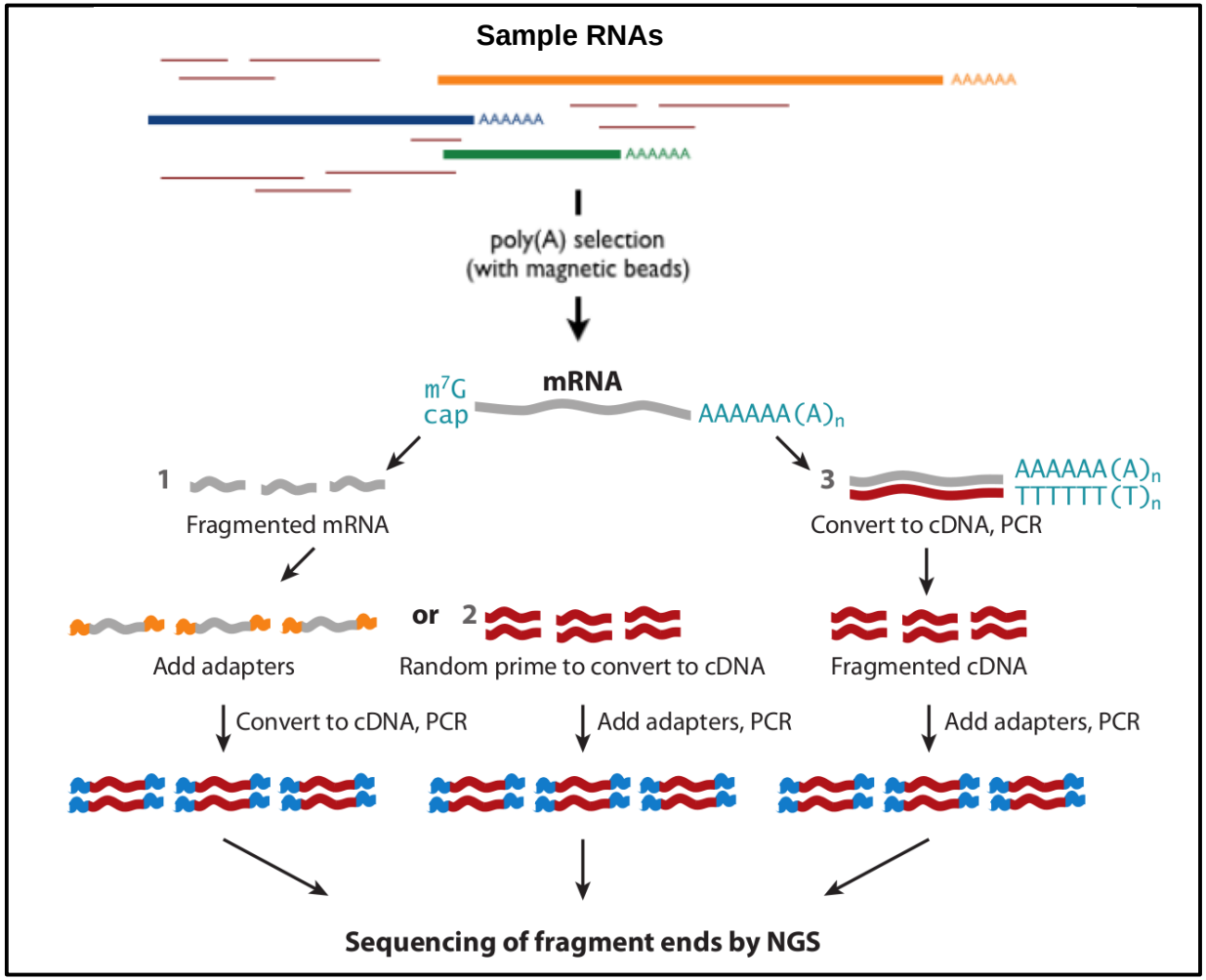
mRNA

↓

FRAGMENTED mRNA/cDNA

↓

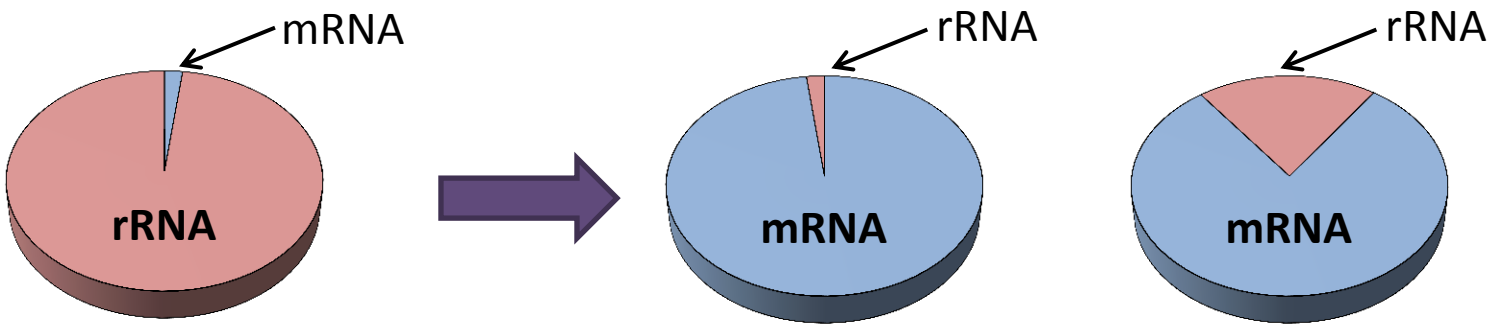
FINISHED LIBRARY



Adapted from Simon et al., 2009, Ann. Rev. Plant Biol. 60:305

RNA Processing

- PolyA Selection
  - Oligo-dT, often using magnetic beads
  - Isolates mRNA very efficiently ***unless total RNA is very dilute***
  - Can't be used to sequence non-polyA RNA
- rRNA Depletion
  - RiboZero, RiboMinus
  - Non-polyA RNAs preserved (non-coding, bacterial RNA, etc.)
  - Can be less effective at removing all rRNA



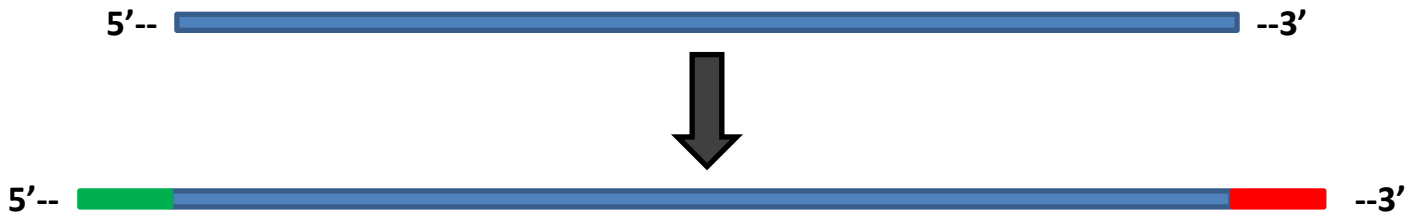
Total RNA sample

After PolyA isolation

After DSN rRNA depletion

Strand-Specific (Directional) RNA-Seq

- Preserves orientation of RNA after reverse transcription to cDNA



- Inform alignments to genome
  - Determine which genomic DNA strand is transcribed
  - Identify anti-sense transcription (e.g., lncRNAs)
  - Quantify expression levels more precisely
  - Demarcate coding sequences in microbes with overlapping genes
- Very useful in transcriptome assemblies
  - Allows precise construction of sense and anti-sense transcripts

Insert Size

Adapter
~60 bases

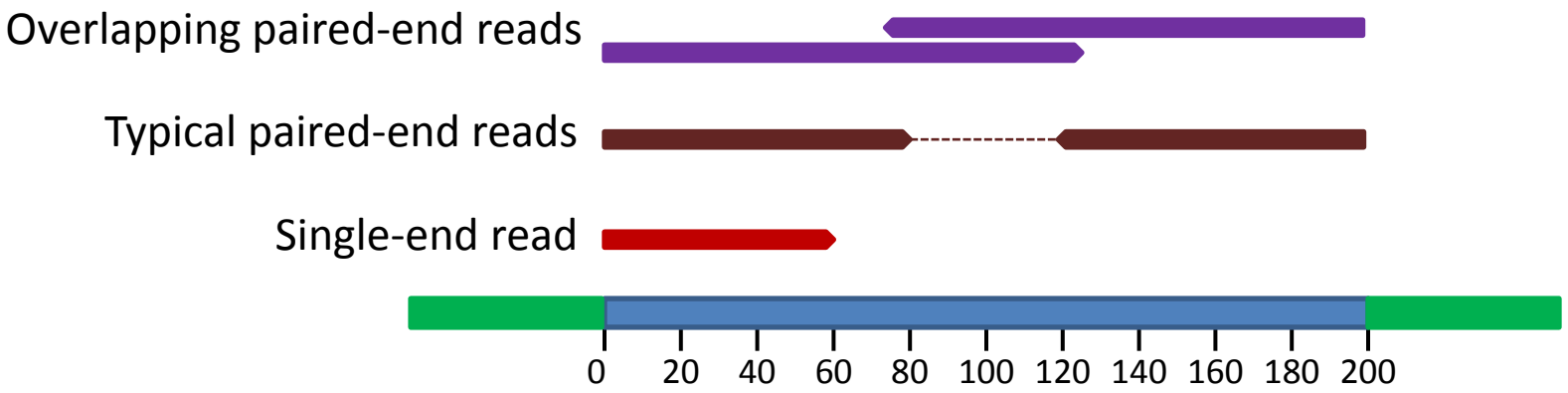
cDNA
100-800 bases
=**Insert Size**

Adapter
~60 bases

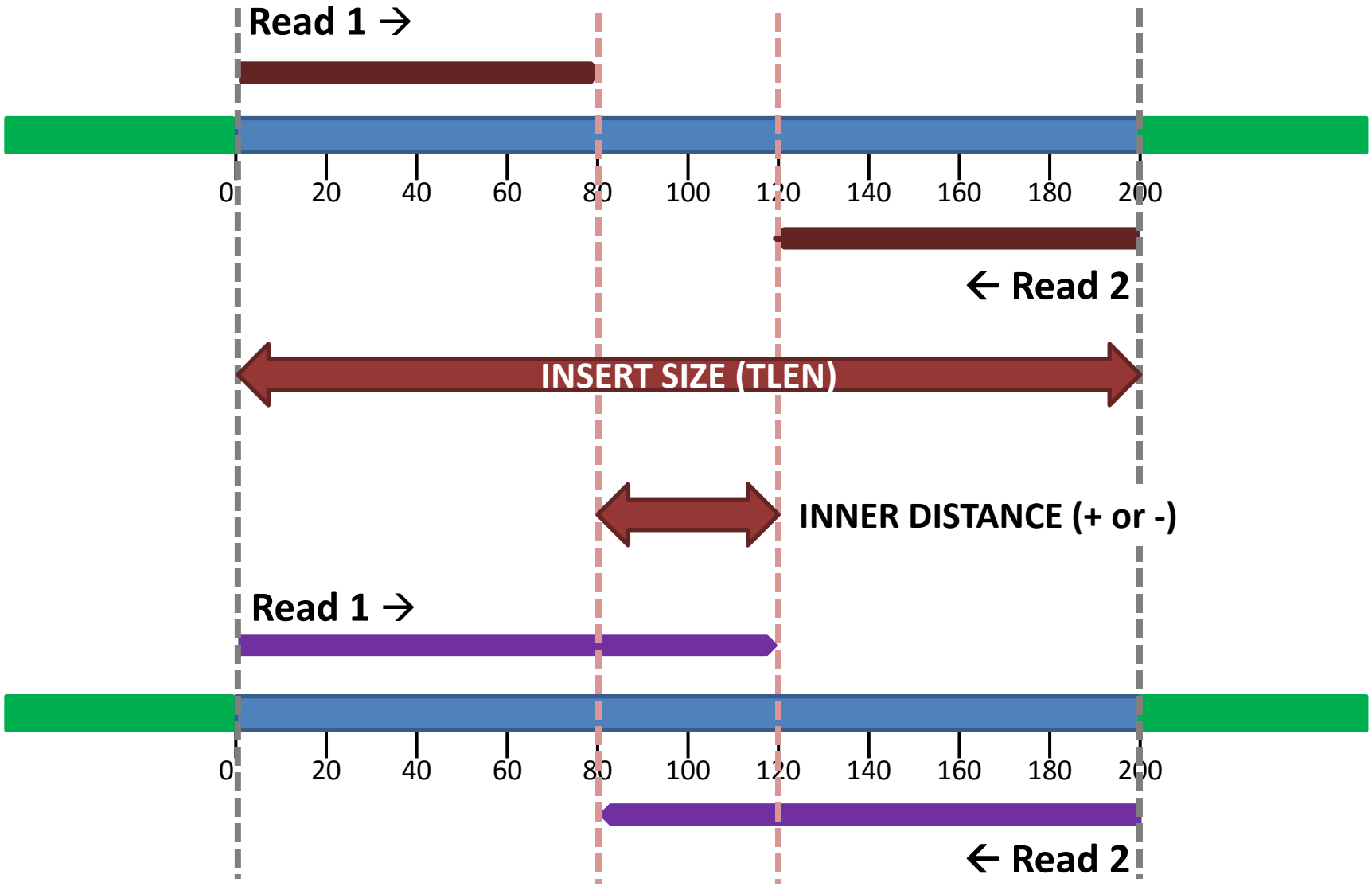
The “insert” is the cDNA (or RNA) ligated between the adapters.

Typical insert size is 160-200 bases, but can be larger.

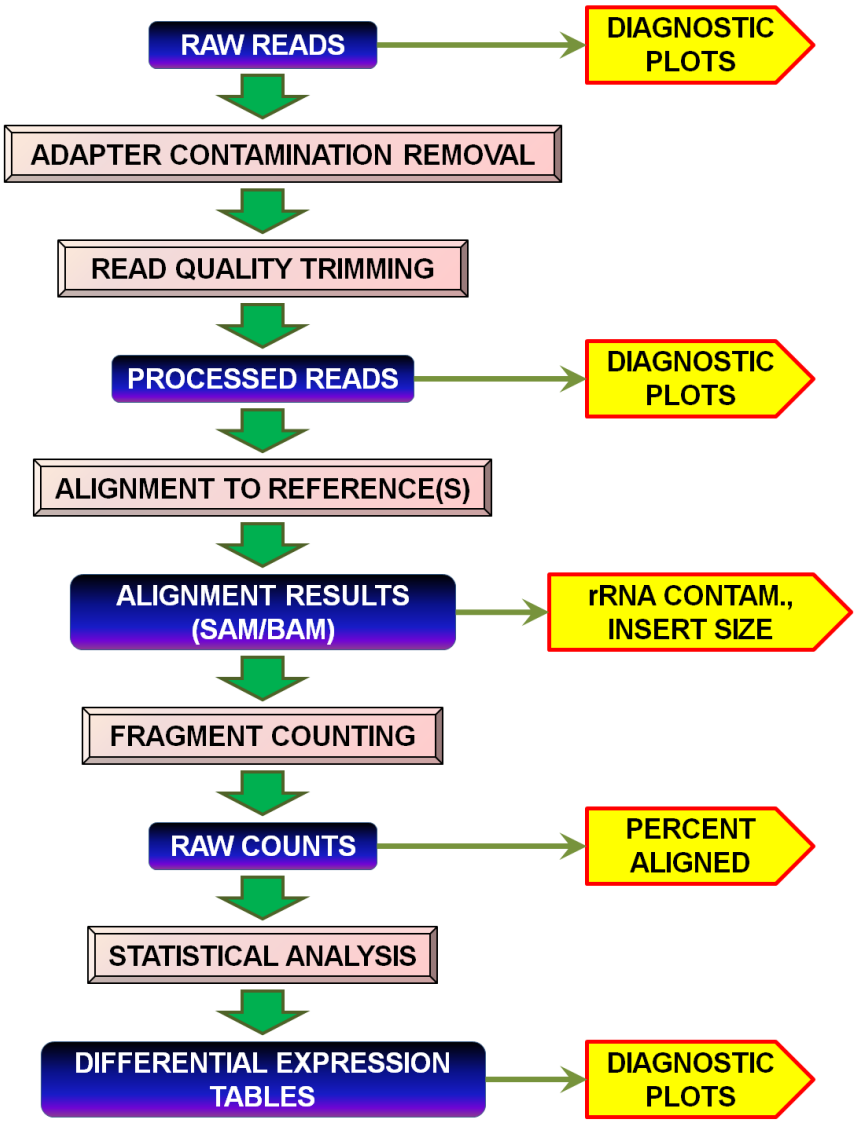
Insert size distribution depends on library prep method.



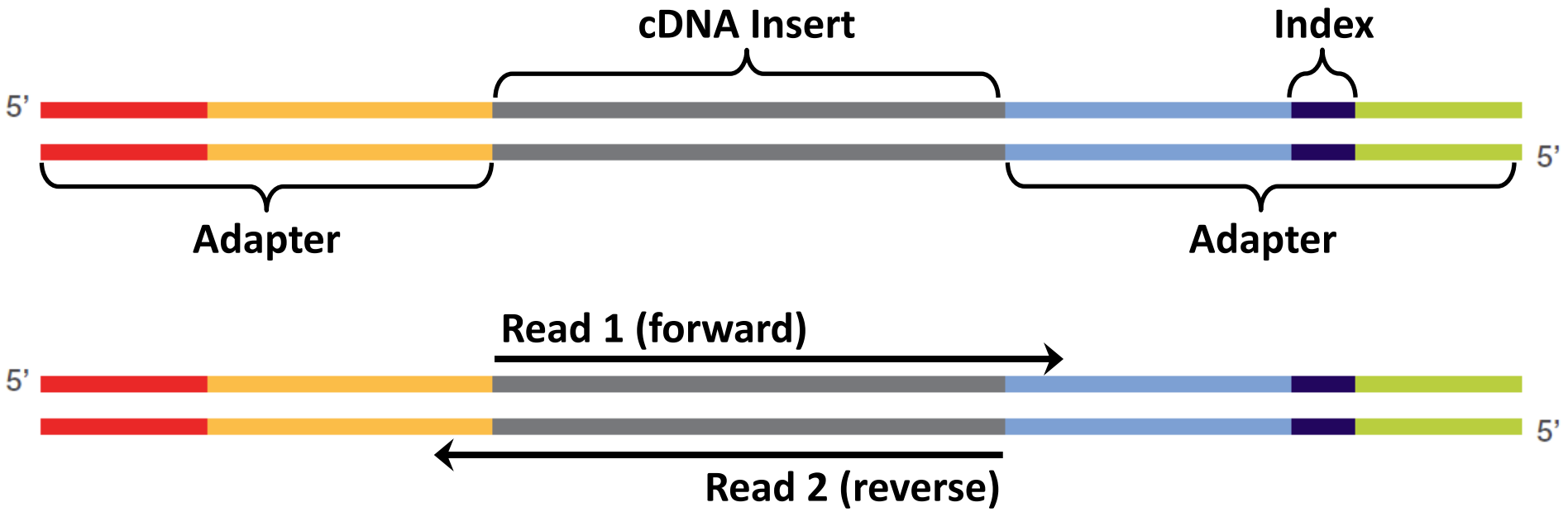
Paired-End Reads



Differential Gene Expression Generalized Workflow



Adapter Contamination



- cDNA inserts are a distribution of sizes
- There will be some read-through with adapter sequence at 3' end
- Removal of adapter contamination can improve fraction of reads that align to the reference
- Very important for de novo assemblies

Alignment - Choosing a Reference

- Fully sequenced and annotated genome
  - Provides exon information to find splice variants
- Predicted/validated transcriptome
  - Simple to use
  - Comprehensive for all but the most novel genes
- NCBI Unigene Sets
  - Often incomplete
  - Good for medium to highly expressed genes
- No Genome? No Problem!
  - Transcriptome assembly
  - Useful for organisms with little or no sequence available
  - But, expect some redundancy and collapsing of gene families

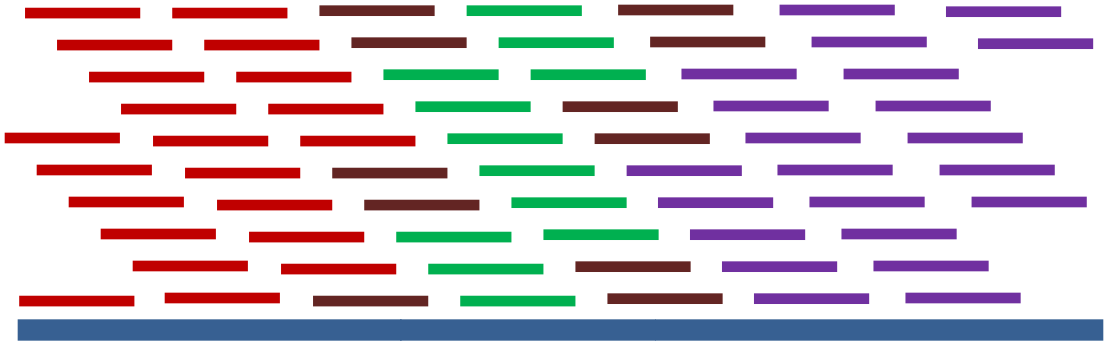
Read Alignment and Counting

- Align reads to genome or transcriptome (output sam/bam)
- Convert alignments to read-counts per gene
  - May need to parse genomic intervals from gene models
  - Output is table of raw counts per gene for each sample
- Simple Normalization
  - RPKM (Reads Per Kilobase per Million reads mapped)
  - FPKM (Fragments Per Kilobase per Million reads mapped)
    - Fragment = cDNA insert
    - Ideally, there are two mappable reads per fragment
- Statistical Analysis (Blythe's talk)
  - Compare expression between samples, tissues, etc.
  - Use appropriate statistical model for your experiment.

Read Alignment and Counting

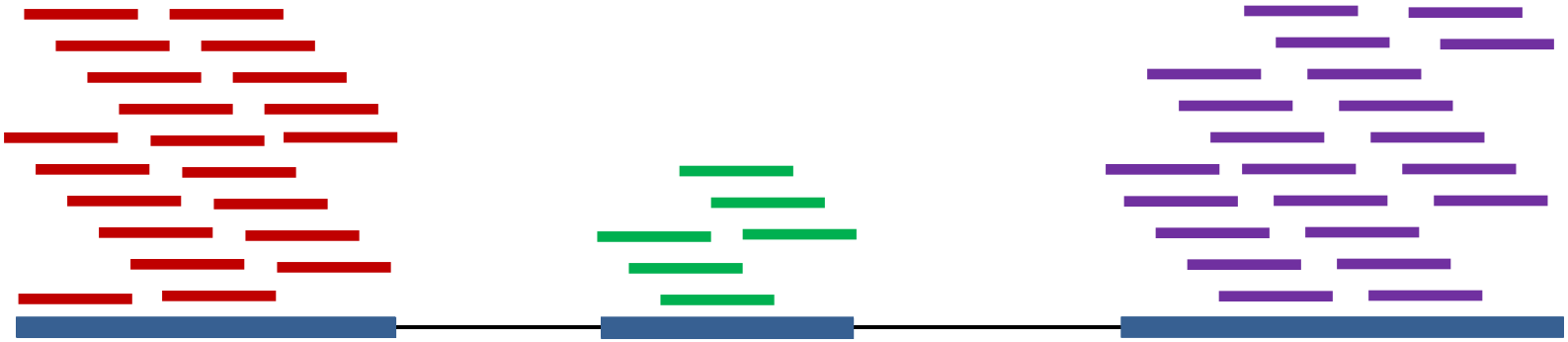
Alignment to Genome – one splice variant

Alignment to Transcriptome

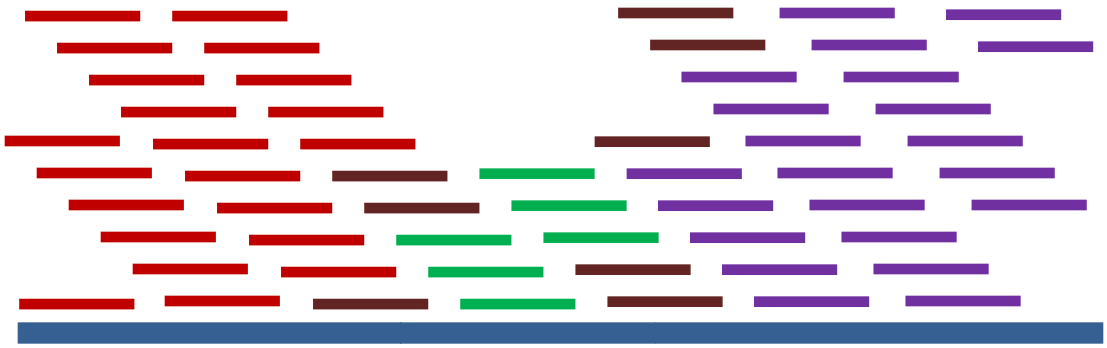


Read Alignment and Counting

Alignment to Genome – two splice variants

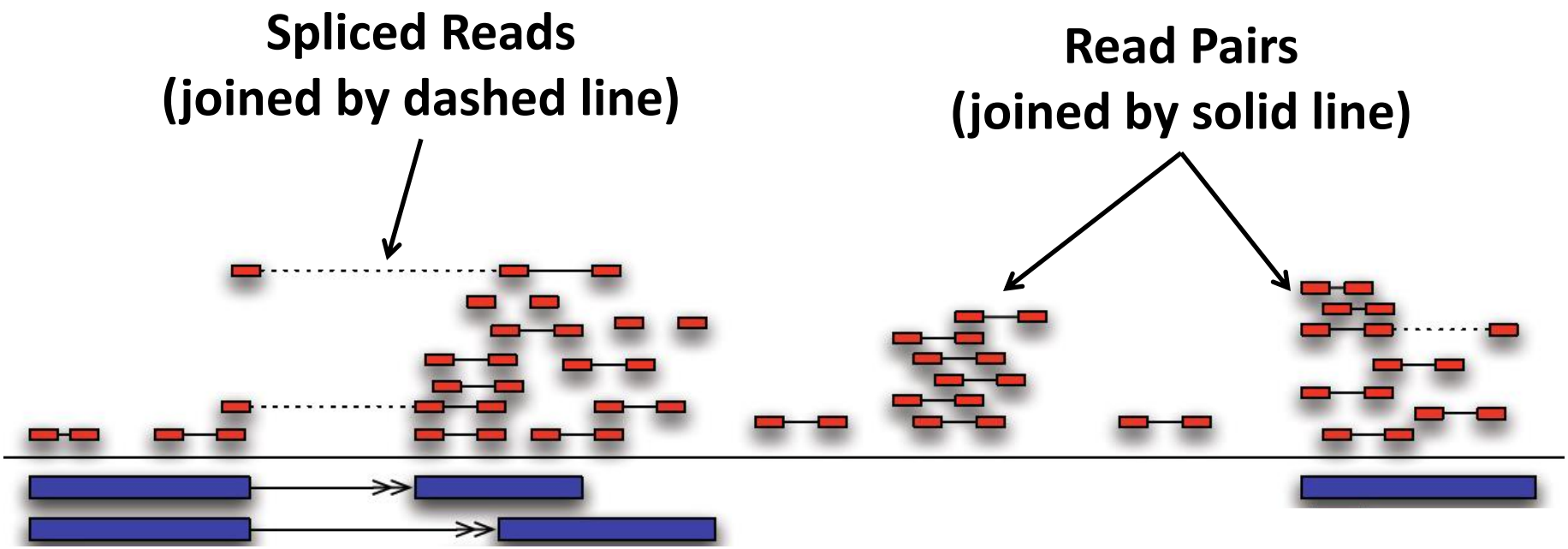


Alignment to Transcriptome (Gene Sequences)



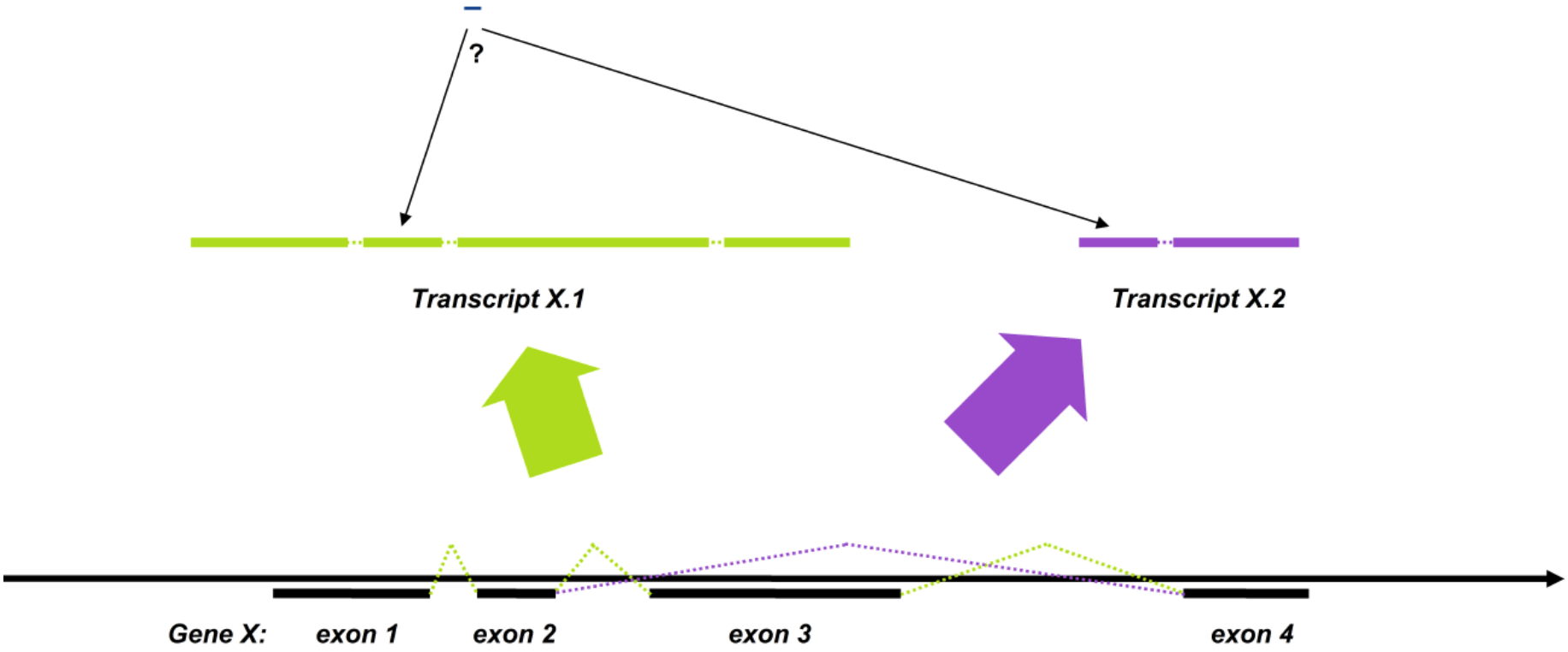
Splicing-Aware Alignment

A splicing-aware aligner will recognize the difference between a short insert and a read that aligns across exon-intron boundaries



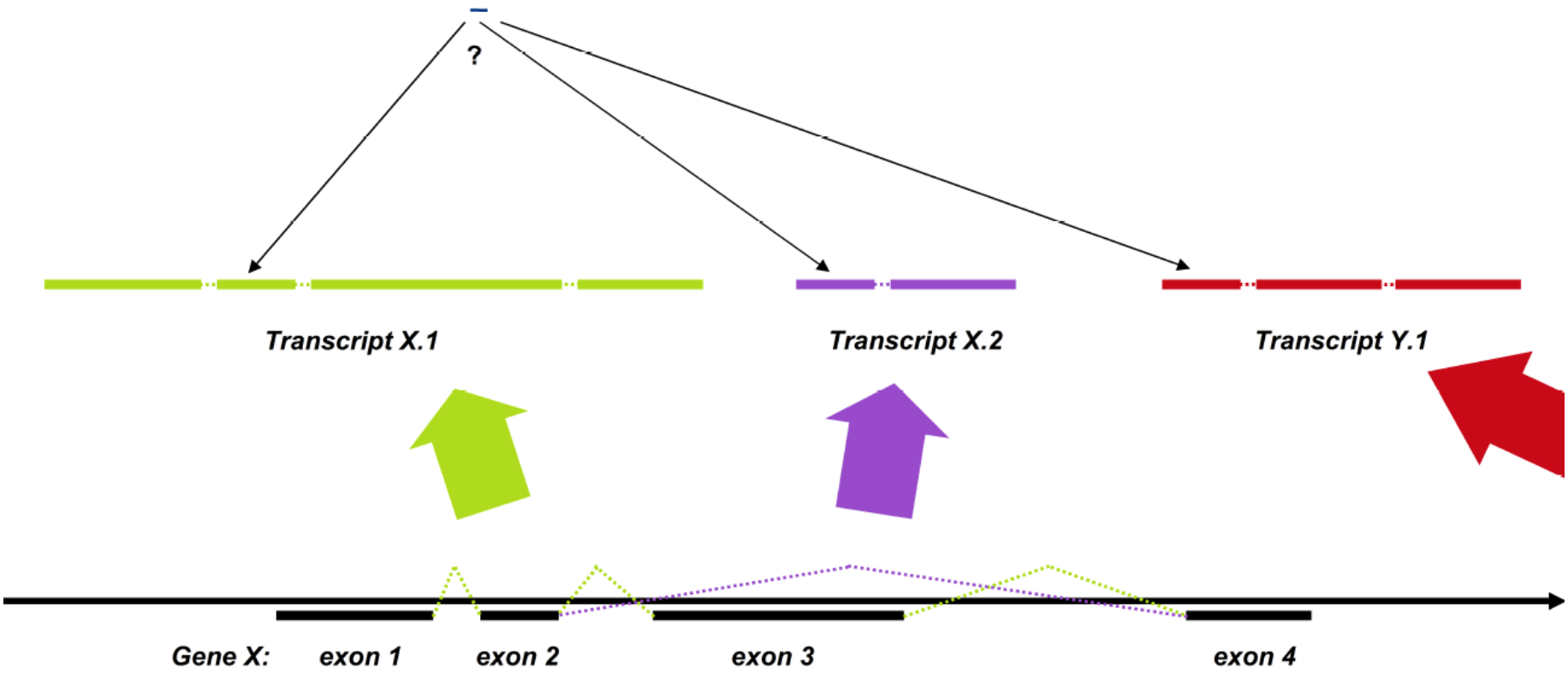
Transcript Reference vs. Genome Reference

Some reads will align uniquely to an exon in the genome. How can transcript abundance be determined?



Multiple Mapping within the Genome

Some reads will align to more than one location in the genome. Which gene/transcript should this read be assigned to?



Multiple-Mapping Reads ("Multireads")

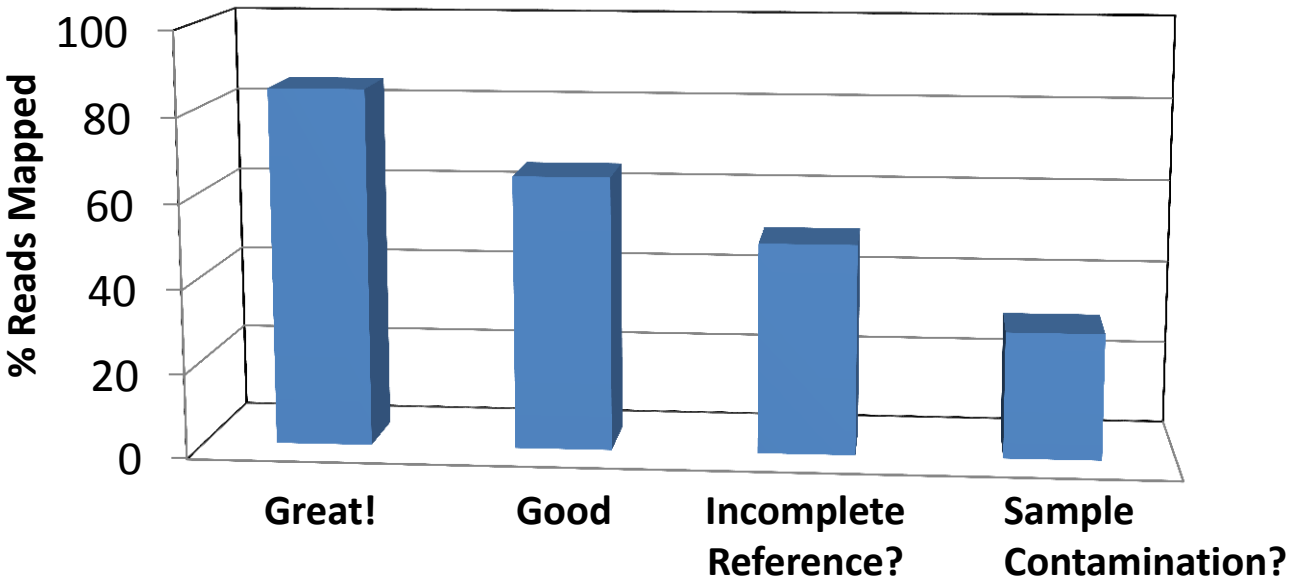
- Some reads will align to more than one place in the reference, because:
  - Shared exons (if reference is transcriptome)
  - Common domains, gene families
  - Paralogs, pseudogenes, etc.
- This can distort counts, leading to misleading expression levels
- If a read can't be uniquely mapped, how should it be counted?
- Should it be ignored (not counted at all)?
- Should it be randomly assigned to one location among all the locations to which it aligns equally well?
- This may depend on the question you're asking…
- …and also depends on the software you use.

Choosing an Aligner

- Transcriptome reference – BWA, Bowtie2
- Genome reference
  - Aligner must be splicing-aware to account for reads that cross intron-exon boundaries
  - TopHat (Bowtie)/TopHat2 (Bowtie2) (tophat.cbcb.umd.edu)
  - GSNAP (research-pub.gene.com/gmap/)
  - STAR (http://gingeraslab.cshl.edu/STAR/) – newest, fastest, uses most memory
- Each aligner has multiple parameters that can be tweaked, affecting read mapping results
- Most software is updated regularly, to improve performance and accommodate new technologies
- GET ON THE MAILING LISTS!

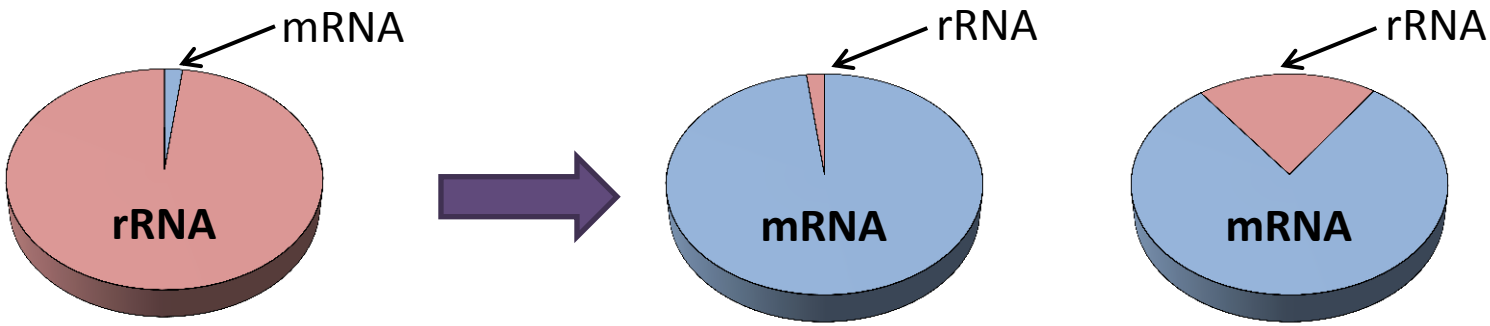
How well did your data align to the reference?

Calculate percentage of reads mapped per sample



RNA Quality Assessment

- rRNA contamination



Total RNA sample

After PolyA isolation

After DSN rRNA depletion

- Align reads to rRNA sequences from organism or relatives
- Generally, don't need to remove rRNA reads

Checking Your Results

- Key genes that may confirm sample ID
  - Knock-out or knock-down genes
  - Genes identified in previous research
- Specific genes of interest
  - Hypothesis testing
  - Important pathways
- Experimental validation (e.g., qRT-PCR)
  - Generally required for publication
  - The best way to determine if your analysis protocol accurately models your organism/experiment
  - Ideally, validation should be conducted on a different set of samples

Analysis Choices

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

RESEARCH ARTICLE Open Access

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

This paper compares eleven methods

This paper compares seven methods

Briefings in Bioinformatics Advance Access published September 17, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Analysis Choices

Method

Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci and **Doron Betel**

Genome Biology 2013, **14**:R95 doi:10.1186/gb-2013-14-9-r95

Published: 10 September 2013

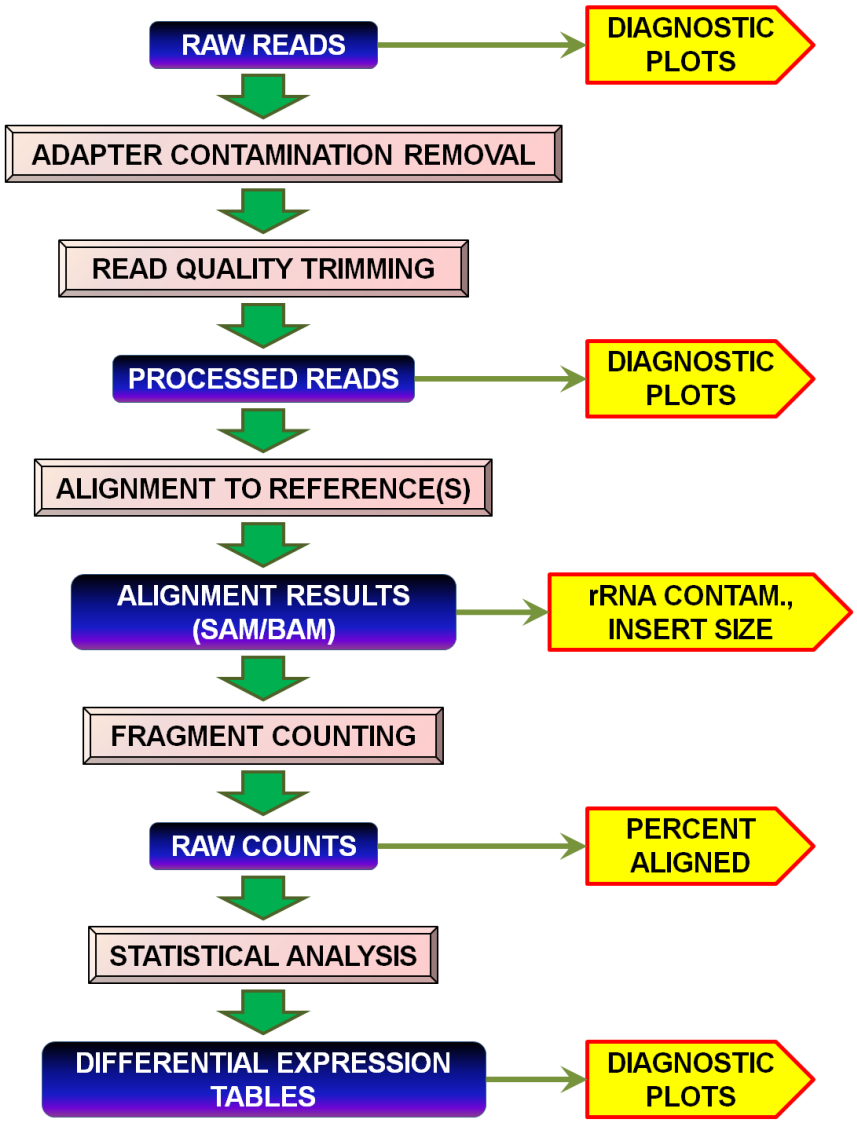
- Evaluated 6 differential gene expression analysis software packages (did not investigate differential isoform expression)
- **Increasing replicates is more important than increasing sequencing depth**
- Transcript length bias reduces the ability to find differential expression in shorter genes.
- limma and baySeq most closely model “reality”.
- limma and edgeR had the fewest number of false positives.
- BUT, 5 of 6 packages were out-of-date by publication date; at least two changed substantially, so this analysis might be different today (or next year)

Where to find some guidance?

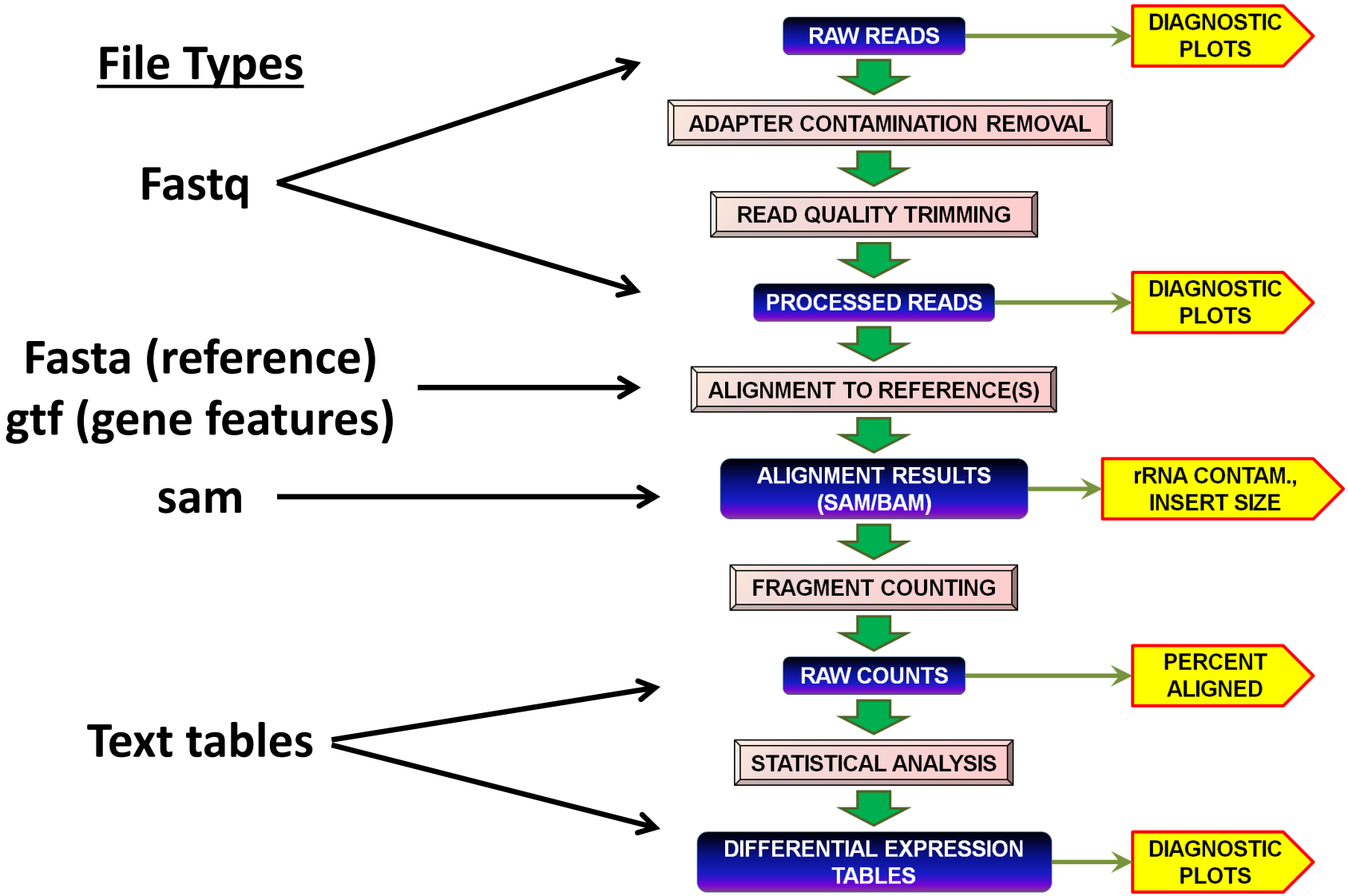
- The RNA Sequencing Quality Control (SEQC) Consortium and the Association of Bimolecular Resource Facilities (ABRF) conducted several systematic large-scale assessments
- RNA-Seq replicates run in ABRF study:
  - 15 lab sites
  - 4 protocols (polyA-select, ribo-depleted, size selected, degraded)
  - 5 platforms (HiSeq, Ion Torrent PGM & Proton, PacBio, 454)
- SEQC generated over 100 billion reads across three platforms
- More than 10Tb data available for analysis

Differential Gene Expression Generalized Workflow

- **Bioinformatics analyses are *in silico* experiments**
- The tools and parameters you choose will be influenced by factors including:
  - Available reference/annotation
  - Experimental design (e.g., pairwise vs. multi-factor)
- The "right" tools are the ones that best inform on your experiment
- **Don't just shop for methods that give you the answer you want**



Differential Gene Expression Generalized Workflow



The GTF (Gene Transfer Format) File

chr12	unknown	exon	4382902	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4383207	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	start_codon	4383207	4383209	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4409026	4409172	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4409026	4414522	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	stop_codon	4409173	4409175	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

chr12	unknown	exon	4382902	4383401	.
chr12	unknown	CDS	4383207	4383401	.
chr12	unknown	start_codon	4383207	4383209	.
chr12	unknown	CDS	4385171	4385386	.
chr12	unknown	exon	4385171	4385386	.
chr12	unknown	CDS	4387926	4388085	.
chr12	unknown	exon	4387926	4388085	.
chr12	unknown	CDS	4398008	4398156	.
chr12	unknown	exon	4398008	4398156	.
chr12	unknown	CDS	4409026	4409172	.
chr12	unknown	exon	4409026	4414522	.
chr12	unknown	stop_codon	4409173	4409175	.

The left columns list source, feature type, and genomic coordinates

The right column includes attributes, including gene ID, etc.

+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

Fields in the GTF File

```
chr12   unknown  CDS   3677872  3678014 .   +   2   gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";
```

Sequence Name (i.e., chromosome, scaffold, etc.)	**chr12**
Source (program that generated the gtf file or feature)	**unknown**
Feature (i.e., gene, exon, CDS, start codon, stop codon)	**CDS**
Start (starting location on sequence)	**3677872**
End (end position on sequence)	**3678014**
Score	**.**
Strand (+ or -)	**+**
Frame (0, 1, or 2: which is first base in codon, zero-based)	**2**

Attribute (";"-delimited list of tags with additional info)

This attribute provides info to Tophat/Cufflinks

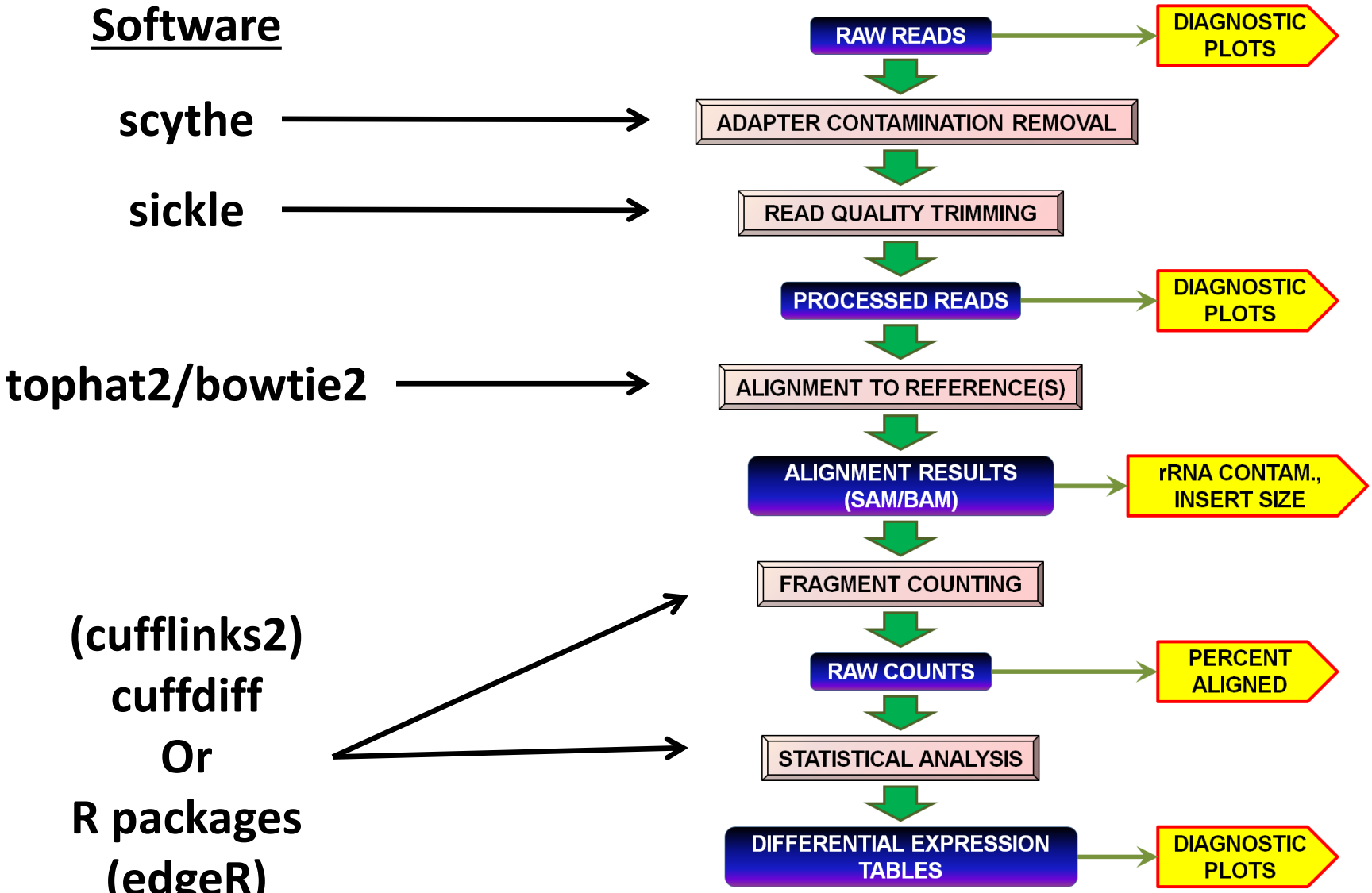
gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933";
transcript_id "NM_019854"; tss_id "TSS4368";

An Unusual GTF File

```
supercont1.164  VectorBase      gene            834185  840032  .  -  .  gene_id "AAEL005599";
supercont1.164  VectorBase      mRNA            834185  840032  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
supercont1.164  VectorBase      exon            839892  840032  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";
supercont1.164  VectorBase      exon            839622  839832  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";
supercont1.164  VectorBase      exon            839190  839447  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";
supercont1.164  VectorBase      exon            834185  834315  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";
supercont1.164  VectorBase      five_prime_utr  840015  840032  .  -  .  transcript_id "AAEL005599-RA";
supercont1.164  VectorBase      start_codon     840012  840014  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
supercont1.164  VectorBase      CDS             839892  840014  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";
supercont1.164  VectorBase      CDS             839622  839832  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";
supercont1.164  VectorBase      CDS             839190  839447  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";
supercont1.164  VectorBase      CDS             834185  834315  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";
supercont1.164  VectorBase      stop_codon      834185  834187  .  -  .  gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
supercont1.164  VectorBase      exon            1334853 1334924 .  -  .  gene_id "AAEL016379"; transcript_id "AAEL005599-RA"; exon_number "5 of 4";
supercont1.164  VectorBase      exon            1497675 1497746 .  -  .  gene_id "AAEL016380"; transcript_id "AAEL005599-RA"; exon_number "6 of 4";
```

```
gene_id "AAEL005599";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";
transcript_id "AAEL005599-RA";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "1 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "2 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "3 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA"; exon_number "4 of 4";
gene_id "AAEL005599"; transcript_id "AAEL005599-RA";
gene_id "AAEL016379"; transcript_id "AAEL005599-RA"; exon_number "5 of 4";
gene_id "AAEL016380"; transcript_id "AAEL005599-RA"; exon_number "6 of 4";
```

Differential Gene Expression Generalized Workflow



Today's Exercises – Differential Gene Expression

Today, we'll analyze the same human RNA-Seq data in a few ways:

1. Single-end reads ("tag counting" with well-annotated genomes)
2. Paired-end reads (finding novel transcripts in a genome with incomplete annotation)
3. Paired-end reads for gene expression when only a transcriptome is available (such as after a de novo transcriptome assembly)

And we'll be using a few different software:

1. Tophat to align spliced reads to a genome
2. Cuffdiff for differential expression of transcripts/genes from tophat alignments
3. htseq-count to generate raw counts tables for…
4. edgeR, which can also handle more complex experimental designs
5. bwa to align reads to a transcriptome reference
6. sam2counts.py to extract raw counts from bwa alignment

Today's Exercises – Differential Gene Expression

Let's get started!