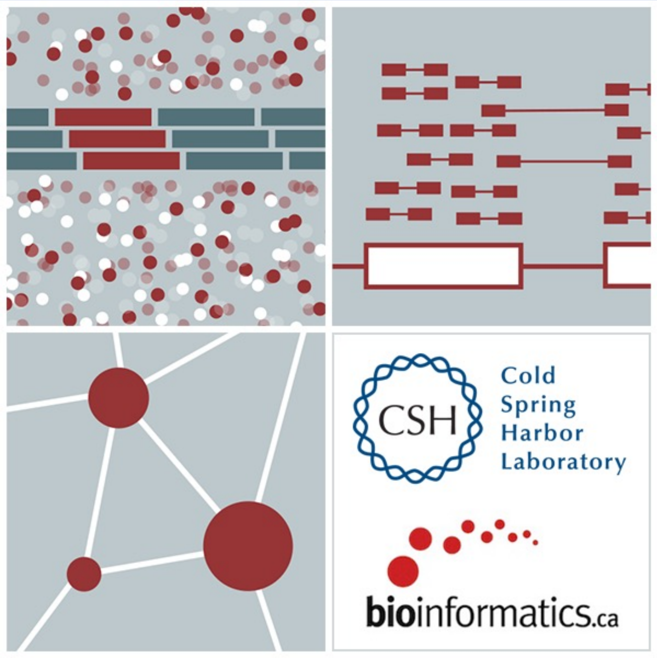
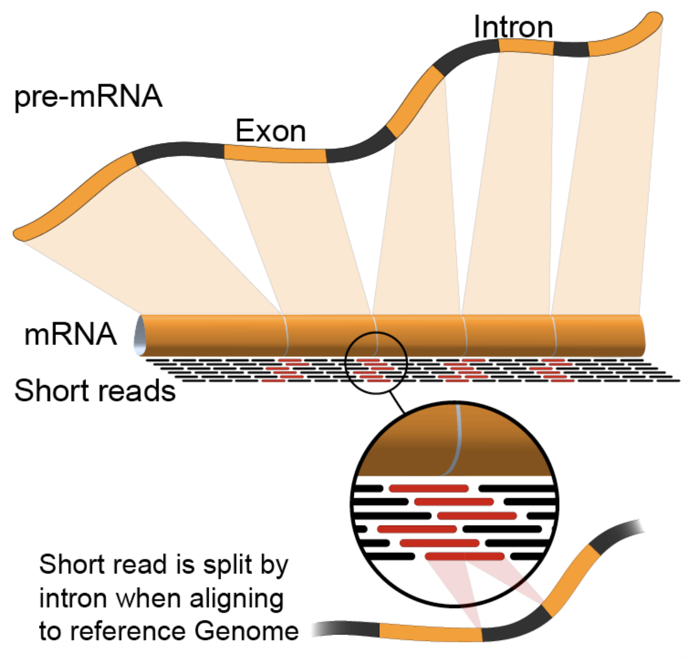
Cold Spring Harbor Laboratory

RNA-Seq Module 3
Abundance Estimation and Differential Expression

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

Advanced Sequencing Technologies & Bioinformatics Analysis November 5-19, 2023



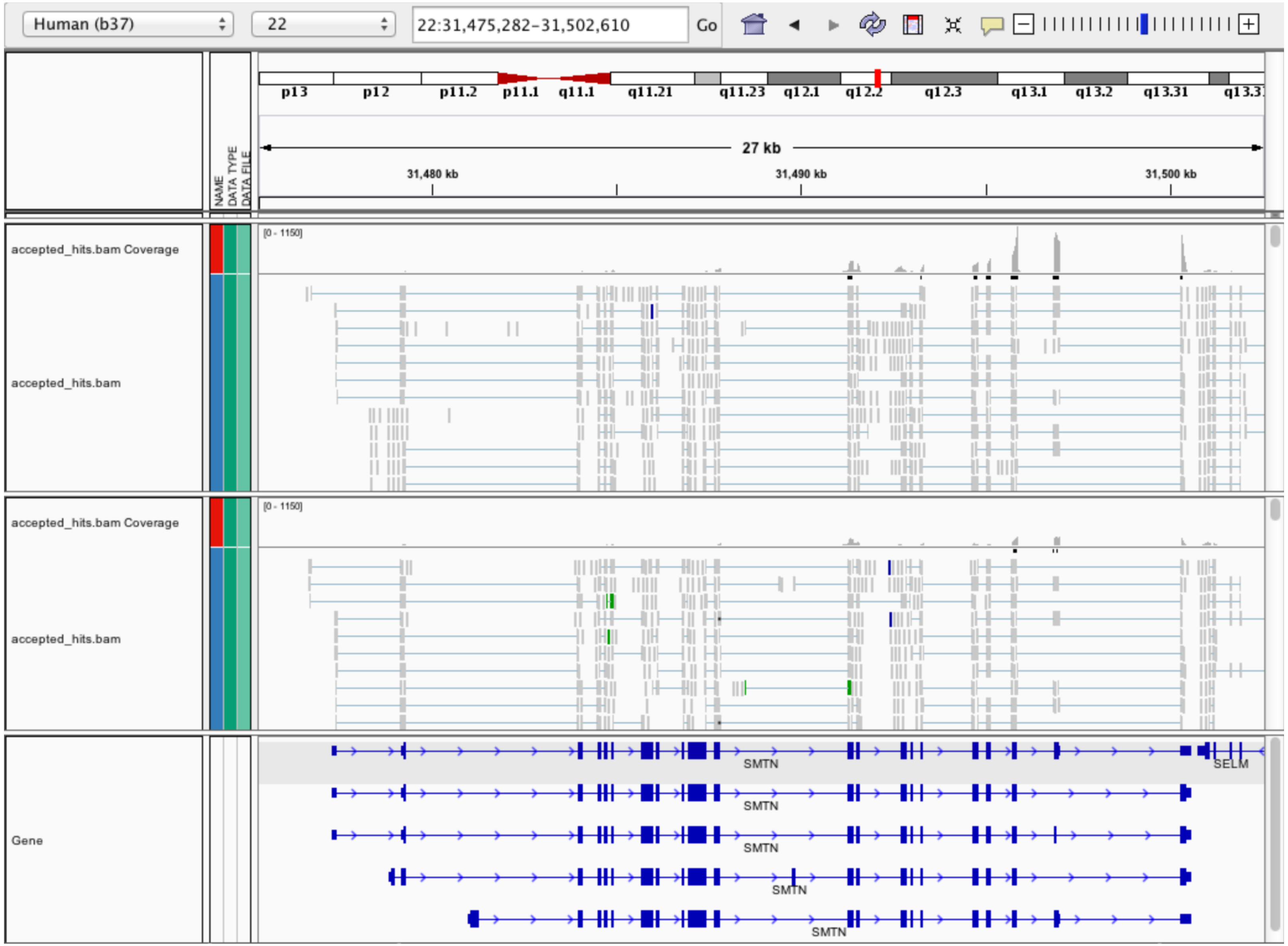


Washington University in St. Louis
School of Medicine

Learning Objectives of Module 3

- Review basic concepts and popular metrics of abundance estimation
- Review StringTie estimation approach and options
- Discuss raw read count approaches
- Review differential expression analysis approaches and caveats

Expression estimation for known genes and transcripts



3’ bias

Down-regulated

What is FPKM (RPKM)?

- RPKM: **Reads** Per Kilobase of transcript per Million mapped reads.
- FPKM: **Fragments** Per Kilobase of transcript per Million mapped reads.

- No essential difference - Just a terminology change to better describe paired-end reads!

What is FPKM?

- Why not just count reads in my RNAseq data? → **Fragments**
- The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - # fragments is biased towards larger genes → **Per Kilobase of transcript**
  - # fragments is related to total library depth → **per Million mapped reads.**

What is FPKM?

- FPKM attempts to normalize for gene size and library depth
    - remember – RPKM is essentially the same!

- C = number of mappable fragments for a gene (transcript)
- N = total number of mappable fragments in the library
- L = number of base pairs in the gene (transcript)
    - FPKM = (C / (N x L)) x 1,000 x 1,000,000
    - FPKM = (1,000,000,000 x C) / (N x L)
    - FPKM = (C / (N / 1,000,000)) / (L/1000)

- More reading:
    - http://www.biostars.org/p/11378/
    - http://www.biostars.org/p/68126/

How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:

FPKM

1) Determine total fragment count, divide by 1,000,000 (per Million)

2) Divide each gene/transcript fragment count by #1 (Fragments Per Million)

3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)

TPM

1) Divide each gene/transcript fragment count by length of the transcript in kilobases (Fragments Per Kilobase)

2) Sum all FPK values for the sample and divide by 1,000,000 (per Million)

3) Divide #1 by #2 (TPM)

- The sum of all TPMs in each sample is the same. Easier to compare across samples!

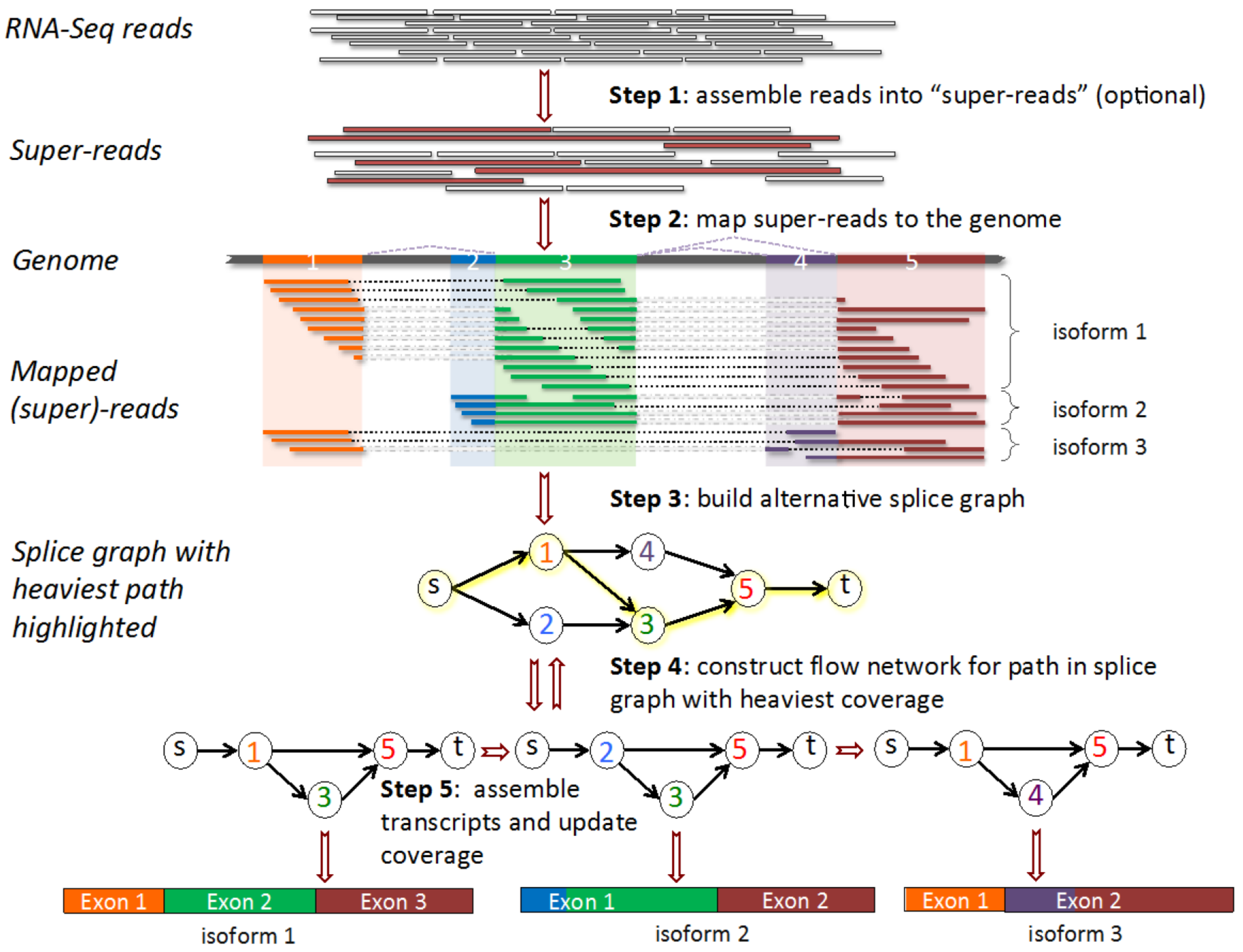
- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

How does StringTie work?

- Align reads to the genome, optionally assemble super-reads and re-align
- Group reads into clusters

Infer isoforms:

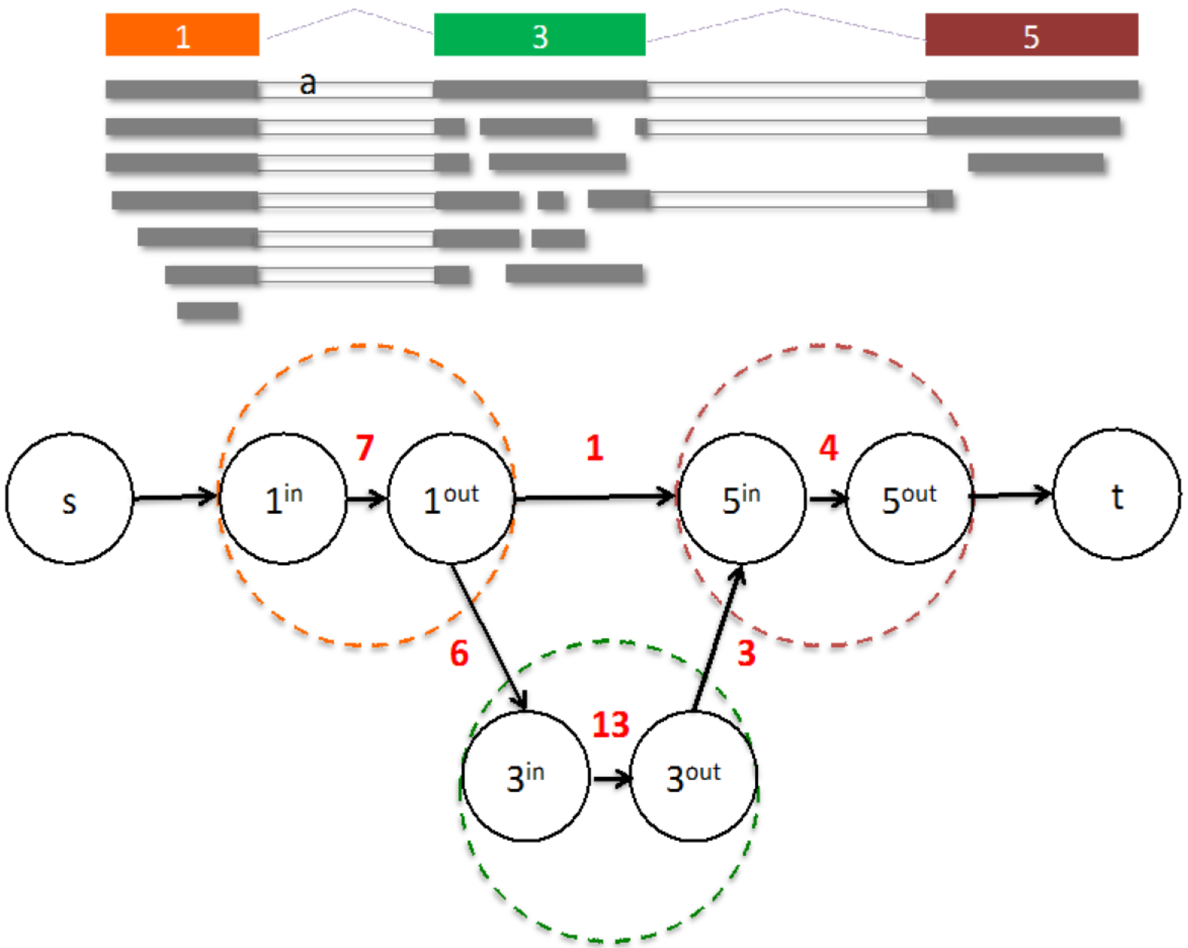
- Build alternative splice graph (ASG)
- Iteratively extract the heaviest path from a splice graph
- construct a flow network
- compute maximum flow to estimate abundance
- update the splice graph by removing reads that were assigned by the flow algorithm
- This process repeats until all reads have been assigned.



RNA-Seq reads
Step 1: assemble reads into "super-reads" (optional)
Super-reads
Step 2: map super-reads to the genome
Genome
1
2
3
4
5
Mapped (super)-reads
isoform 1
isoform 2
isoform 3
Step 3: build alternative splice graph
Splice graph with heaviest path highlighted
s
1
2
3
4
5
t
Step 4: construct flow network for path in splice graph with heaviest coverage
Step 5: assemble transcripts and update coverage
Exon 1
Exon 2
Exon 3
isoform 1
Exon 1
Exon 2
isoform 2
Exon 1
Exon 2
isoform 3

Pertea et al. Nature Biotechnology, 2015

From flow network for each transcript, maximum flow is used to assemble transcript and estimate abundance



StringTie uses basic graph theory (splice graph), custom heuristics (heaviest path), more graph theory (flow network) and optimization theory (maximum flow). See StringTie paper for definitions and math.

StringTie Modes

- Expression estimation mode (“Reference Only”)
  - “–G $GTF_File” AND “–e” option
  - no "novel" transcript assemblies (isoforms)
  - read alignments not overlapping reference transcripts ignored
  - Faster, especially when given limited set of reference transcripts
    - Avoids complicated steps of clustering and building alternative splice graph from scratch, skipping straight to abundance estimation
- “Reference guided mode”
  - “–G $GTF_File” WITHOUT “–e” option
  - Both known and novel transcript assemblies
- “De novo” mode
  - NEITHER “–G $GTF_File” NOR “–e” option
  - Novel transcript assemblies only

Pertea et al. Nature Protocols, 2016

StringTie -merge

- Merge together all gene structures from all samples
  - Some samples may only partially represent a gene structure

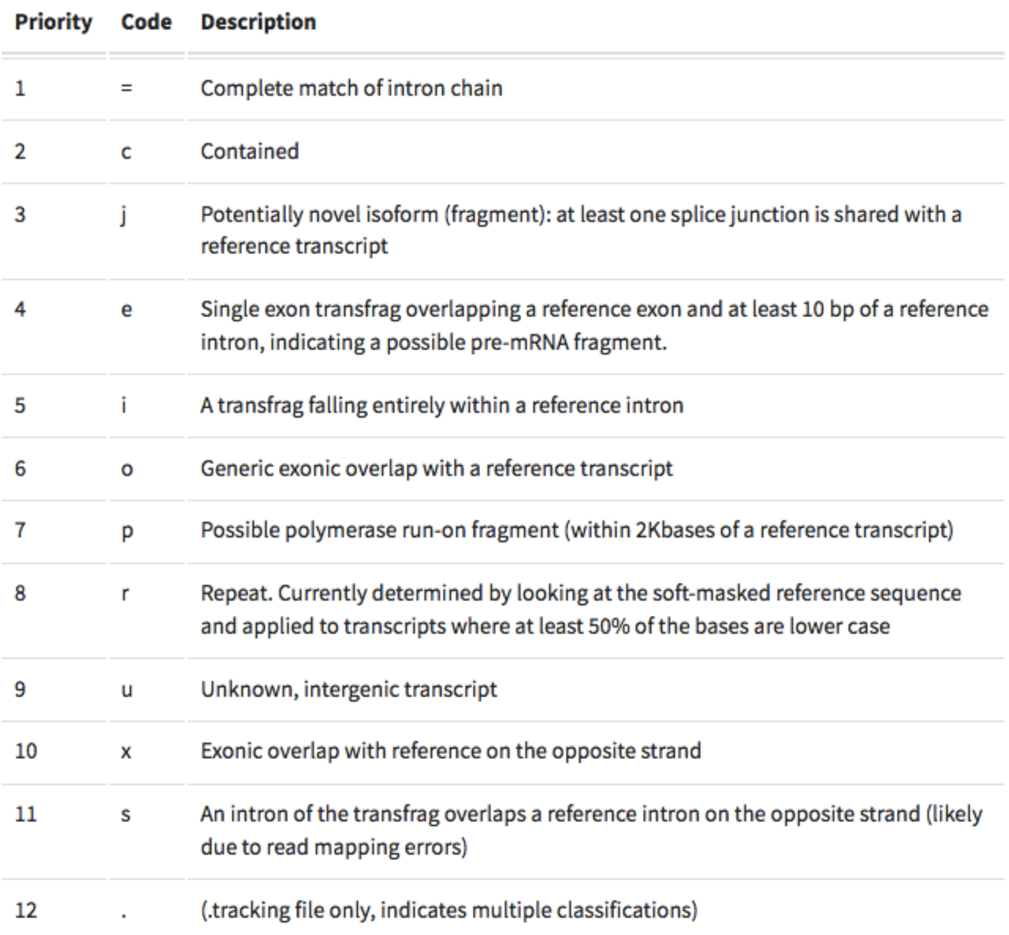
- Incorporates known transcripts with assembled, potentially novel transcripts

- For de novo or reference guided mode, we will rerun StringTie with the merged transcript assembly.

Pertea et al. Nature Protocols, 2016

gffcompare

- gffcompare will compare a merged transcript GTF with known annotation, also in GTF/GFF3 format
- http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#cuffcompare-output-files



Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

Alternatives to FPKM

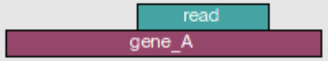
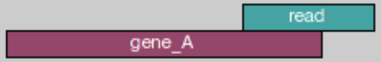
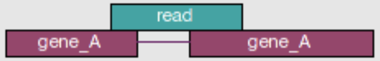
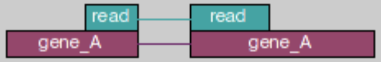
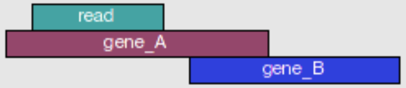
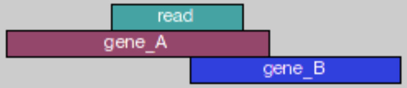
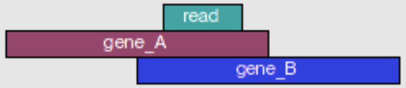
- Raw read counts for differential expression analysis
  - Assign reads/fragments to defined genes/transcripts, get “raw counts”
    - Transcript structures could still be defined by something like Stringtie

- HTSeq (htseq-count)
  - https://htseq.readthedocs.io/

```
htseq-count --mode intersection-strict --stranded no --minaqual 1 --type exon --idattr transcript_id
accepted_hits.sam chr22.gff > transcript_read_counts_table.tsv
```

- Caveats of ‘transcript’ analysis by htseq-count:
    - Designed for genes - ambiguous reads from overlapping transcripts may not be handled!
    - http://seqanswers.com/forums/showthread.php?t=18068

HTSeq-count basically counts reads supporting a feature (exon, gene) by assessing overlapping coordinates

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Note, if gene_A and gene_B on opposite strands, sequence data is stranded, and correct HTSeq parameter set then this read may not be ambiguous

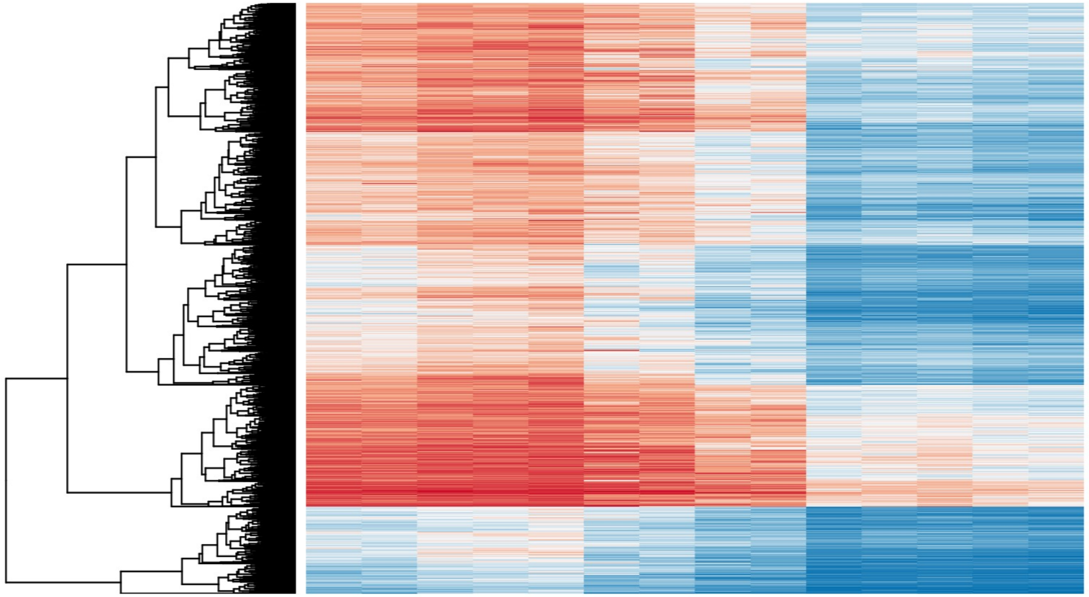
Whether a read is counted depends on the nature of overlap and “mode” selected

Differential Expression

- Tying gene expression back to genotype/phenotype

- What genes/transcripts are being expressed at higher/lower levels in different groups of samples?
  - Are these differences ‘significant’, accounting for variance/noise?

- Examples (used in course):
  - UHR cells vs HBR brain
  - Tumor vs Normal tissue
  - Wild-type vs gene KO cells



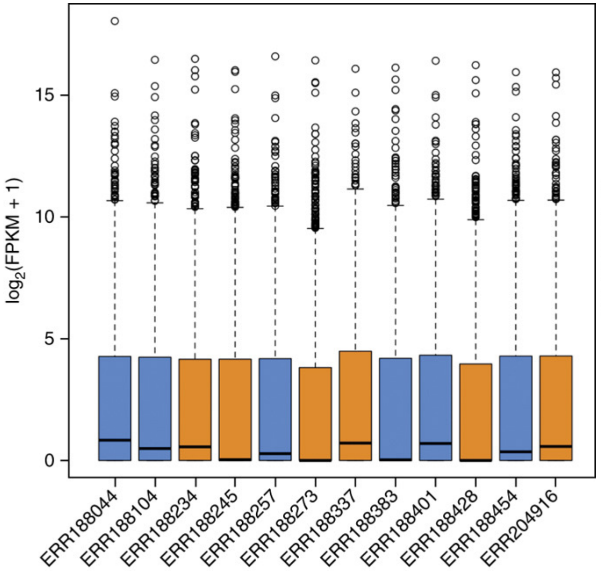
Differential Expression with Ballgown

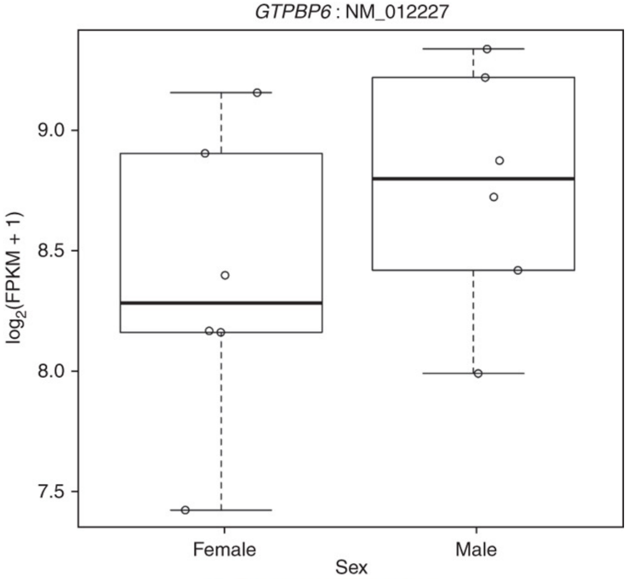
Parametric F-test comparing nested linear models

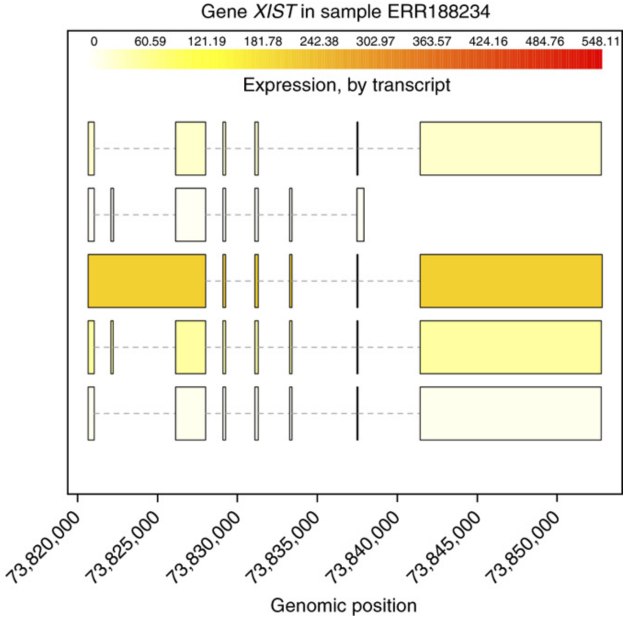
- Two models are fit to each feature, using expression as the outcome
  - one including the covariate of interest (e.g., case/control status or time) and one not including that covariate.
- An F statistic and p-value are calculated using the fits of the two models.
  - A significant p-value means the model including the covariate of interest fits significantly better than the model without that covariate, indicating differential expression.
- Adjust for multiple testing by reporting q-values:
  - q < 0.05 the false discovery rate should be controlled at ~5%.

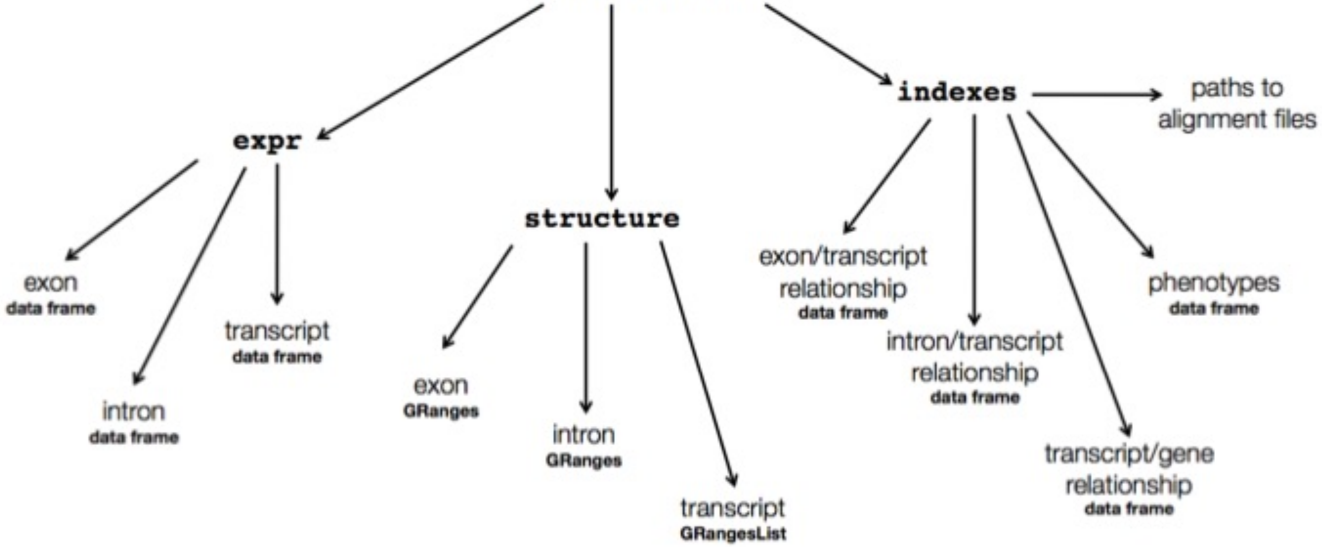
Frazee et al. (2014)

Ballgown for Visualization with R









ballgown object

expr

- exon data frame
- intron data frame
- transcript data frame

structure

- exon GRanges
- intron GRanges
- transcript GRangesList

indexes

- paths to alignment files
- exon/transcript relationship data frame
- intron/transcript relationship data frame
- transcript/gene relationship data frame
- phenotypes data frame

Alternative differential expression methods

- Raw count approaches
  - DESeq2 - http://www-huber.embl.de/users/anders/DESeq/
  - edgeR - http://www.bioconductor.org/packages/release/bioc/html/edgeR.html
  - Others…

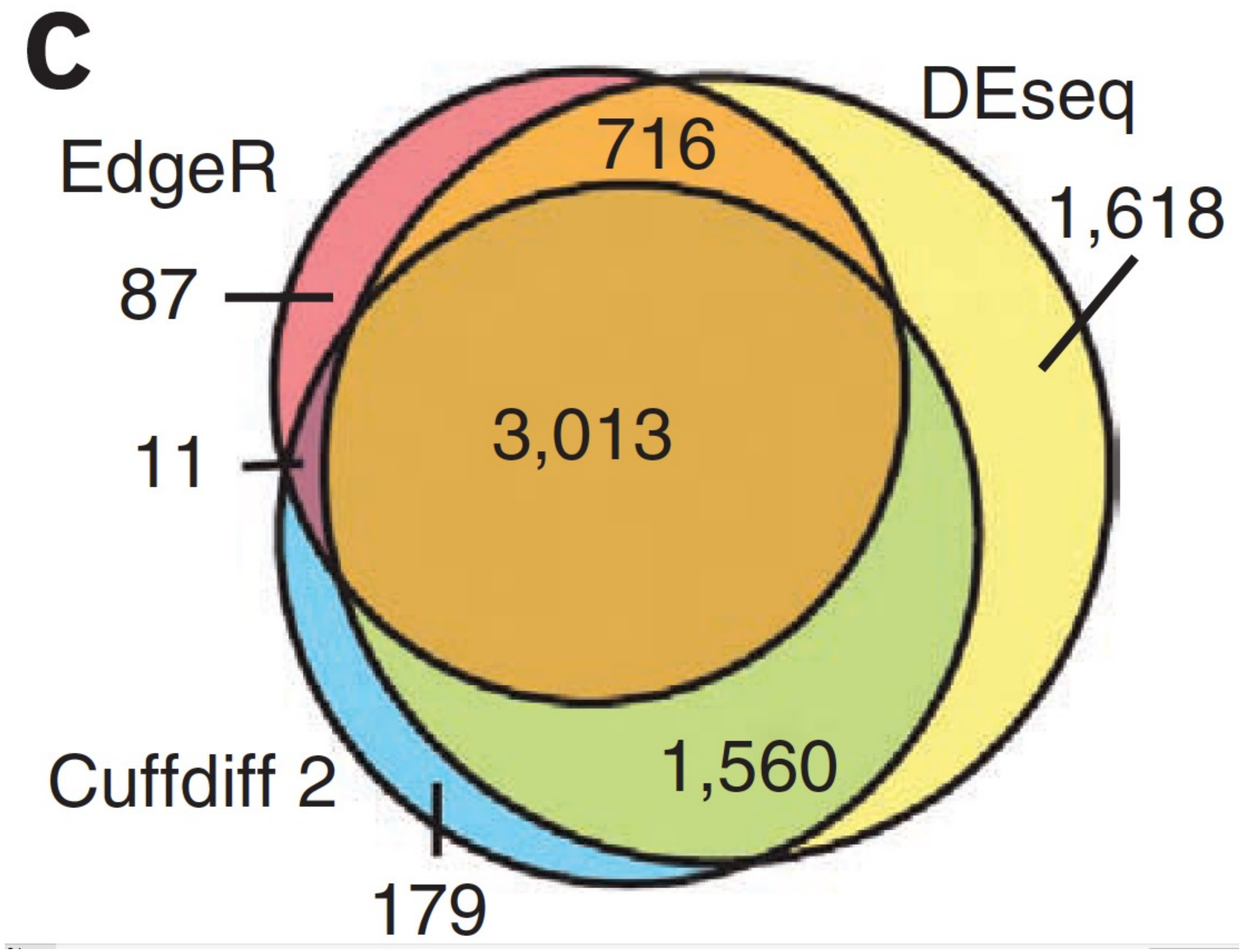
‘FPKM/TPM’ expression estimates vs. ‘raw’ counts

- Which should I use?
  - Long running debate, but the general consensus:

- FPKM/TPM
  - When you want to leverage benefits of tuxedo suite
    - Isoform deconvolution
  - Good for visualization (e.g., heatmaps)
  - Calculating fold changes, etc.

- Counts
  - “More robust" statistical methods for differential expression
    - Stringtie/Ballgown approach is also robust
  - Accommodates more sophisticated experimental designs with appropriate statistical tests

Multiple approaches advisable



Lessons learned from microarray days

- Hansen et al. “Sequencing Technology Does Not Eliminate Biological Variability.” Nature Biotechnology 29, no. 7 (2011): 572–573.

- Power analysis for RNA-seq experiments
  - http://scotty.genetics.utah.edu/

- RNA-seq need for biological replicates
  - http://www.biostars.org/p/1161/

- RNA-seq study design
  - http://www.biostars.org/p/68885/

Multiple testing correction

- As more attributes are compared, differences due solely to chance become more likely!

- Well known from array studies
  - 10,000s genes/transcripts
  - 100,000s exons

- With RNA-seq, more of a problem than ever
  - All the complexity of the transcriptome gives huge numbers of potential features
    - Genes, transcripts, exons, junctions, retained introns, microRNAs, lncRNAs, etc

- Bioconductor multtest
  - http://www.bioconductor.org/packages/release/bioc/html/multtest.html

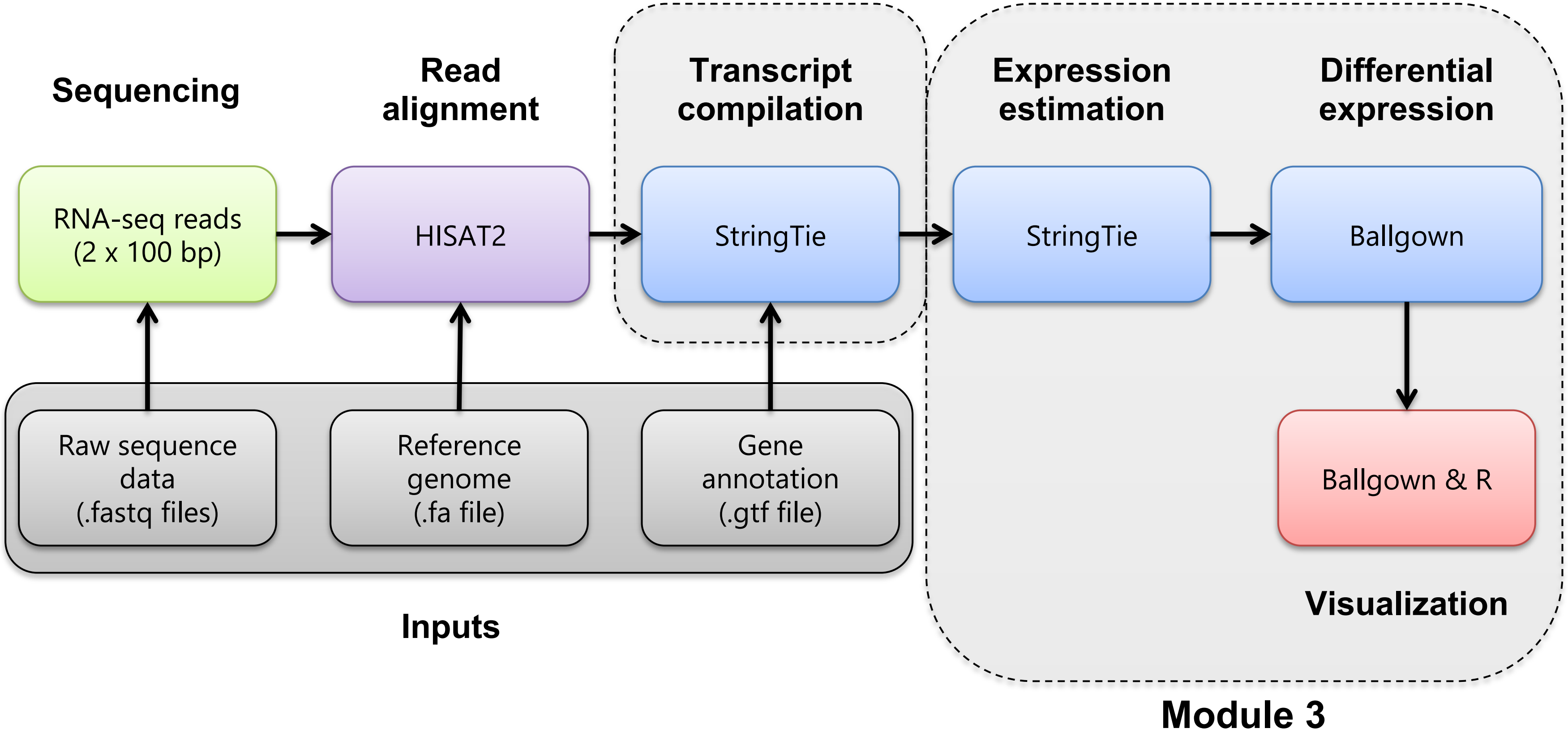
Downstream interpretation of expression analysis

- Topic for an entire course
- Expression estimates and differential expression lists from StringTie, Ballgown or other alternatives can be fed into many analysis pipelines
- See supplemental R tutorial for how to format expression data and start manipulating in R

- Clustering/Heatmaps
  - Provided by Ballgown
  - For more customized analysis various R packages exist:
    - hclust, heatmap.2, plotrix, ggplot2, etc.
- Classification
  - For RNA-seq data we still rarely have sufficient sample size and clinical details but this is changing
    - Weka is a good learning tool
    - RandomForests R package (biostar tutorial being developed)
- Pathway analysis
  - GSEA, IPA, Cytoscape, many R/BioConductor packages: http://www.bioconductor.org/help/search/index.html?q=pathway

https://genviz.org/module-04-expression/0004/01/01/Expression_Profiling_and_Visualization/

HISAT2/StringTie/Ballgown RNA-seq Pipeline



Sequencing

Read alignment

Transcript compilation

Expression estimation

Differential expression

RNA-seq reads (2 x 100 bp)

HISAT2

StringTie

StringTie

Ballgown

Raw sequence data (.fastq files)

Reference genome (.fa file)

Gene annotation (.gtf file)

Ballgown & R

Visualization

Inputs

Module 3

We are on a Coffee Break & Networking Session