Cold Spring Harbor Laboratory

Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html

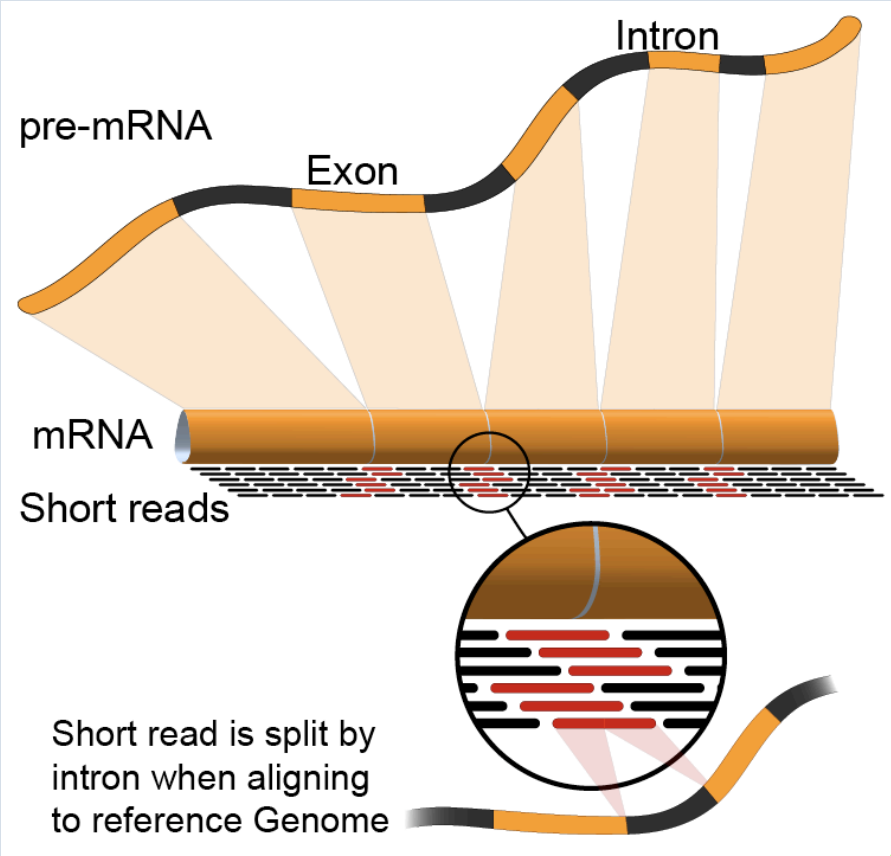
Cold Spring Harbor Laboratory

RNA-Seq Module 3
Expression and Differential Expression (lecture)

Malachi Griffith, Obi Griffith, Jason Walker, Alex Wagner

Advanced Sequencing Technologies & Applications

November 7 - 18, 2017



THE GENOME INSTITUTE at Washington University

Learning objectives of the course

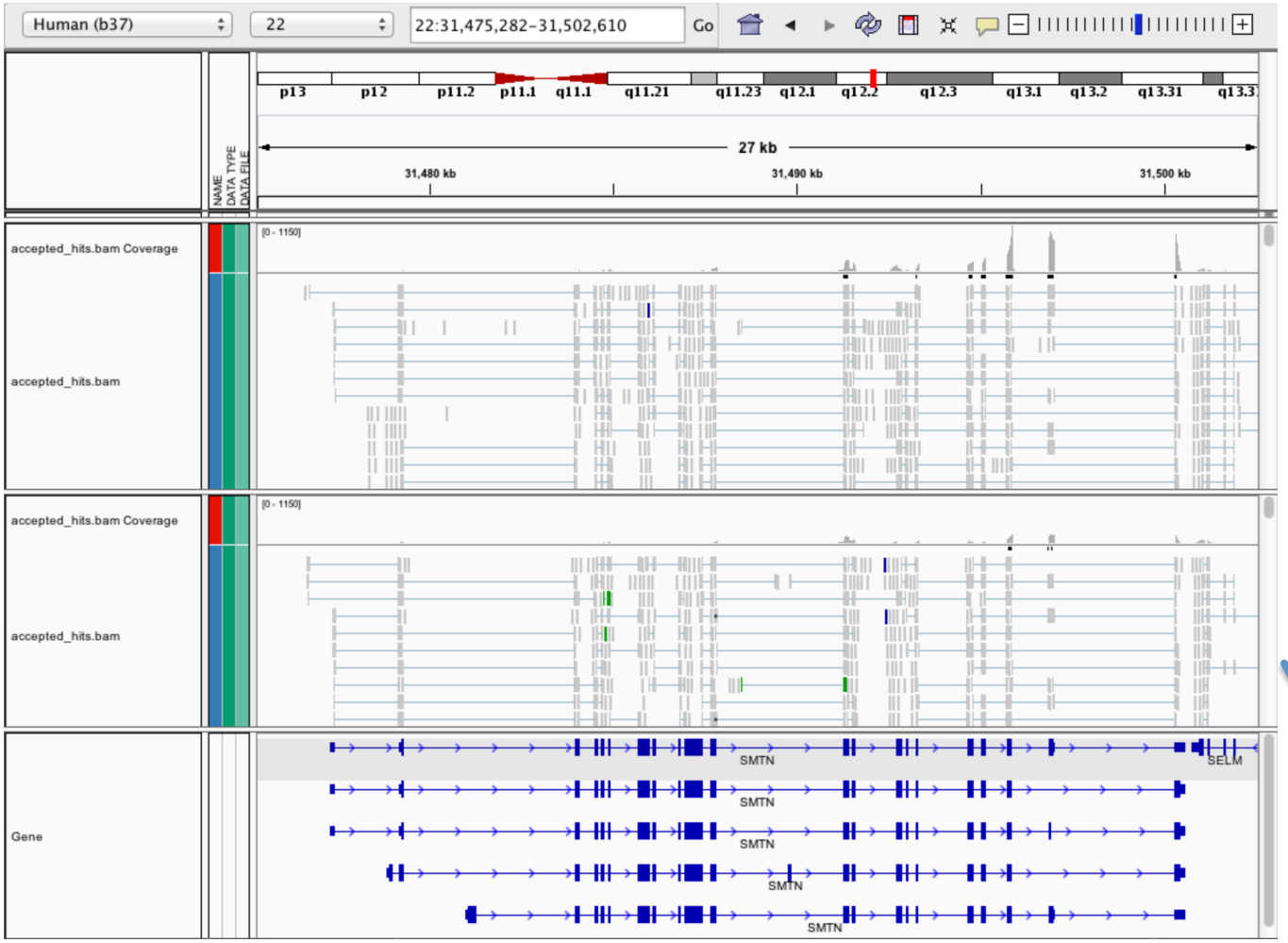
- Module 1: Introduction to RNA Sequencing
- Module 2: Alignment and Visualization
- **Module 3: Expression and Differential Expression**
- Module 4: Isoform Discovery and Alternative Expression

- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a ‘reasonable’ amount of time with modest computer resources
  - Self contained, self explanatory, portable

Learning Objectives of Module

- Expression estimation for known genes and transcripts
- 'FPKM' expression estimates vs. 'raw' counts
- Differential expression methods
- Downstream interpretation of expression and differential estimates
  - multiple testing, clustering, heatmaps, classification, pathway analysis, etc.

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.
- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - The number of fragments is also biased towards larger genes
  - The total number of fragments is related to total library depth
- FPKM (RPKM) attempt to normalize for gene size and library depth

- FPKM (RPKM) = (10^9 * C) / (N * L)
  - C = number of mappable reads/fragments for a gene/transcript/exon/etc
  - N = total number of mappable reads/fragments in the library
  - L = number of base pairs in the gene/transcript/exon/etc

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

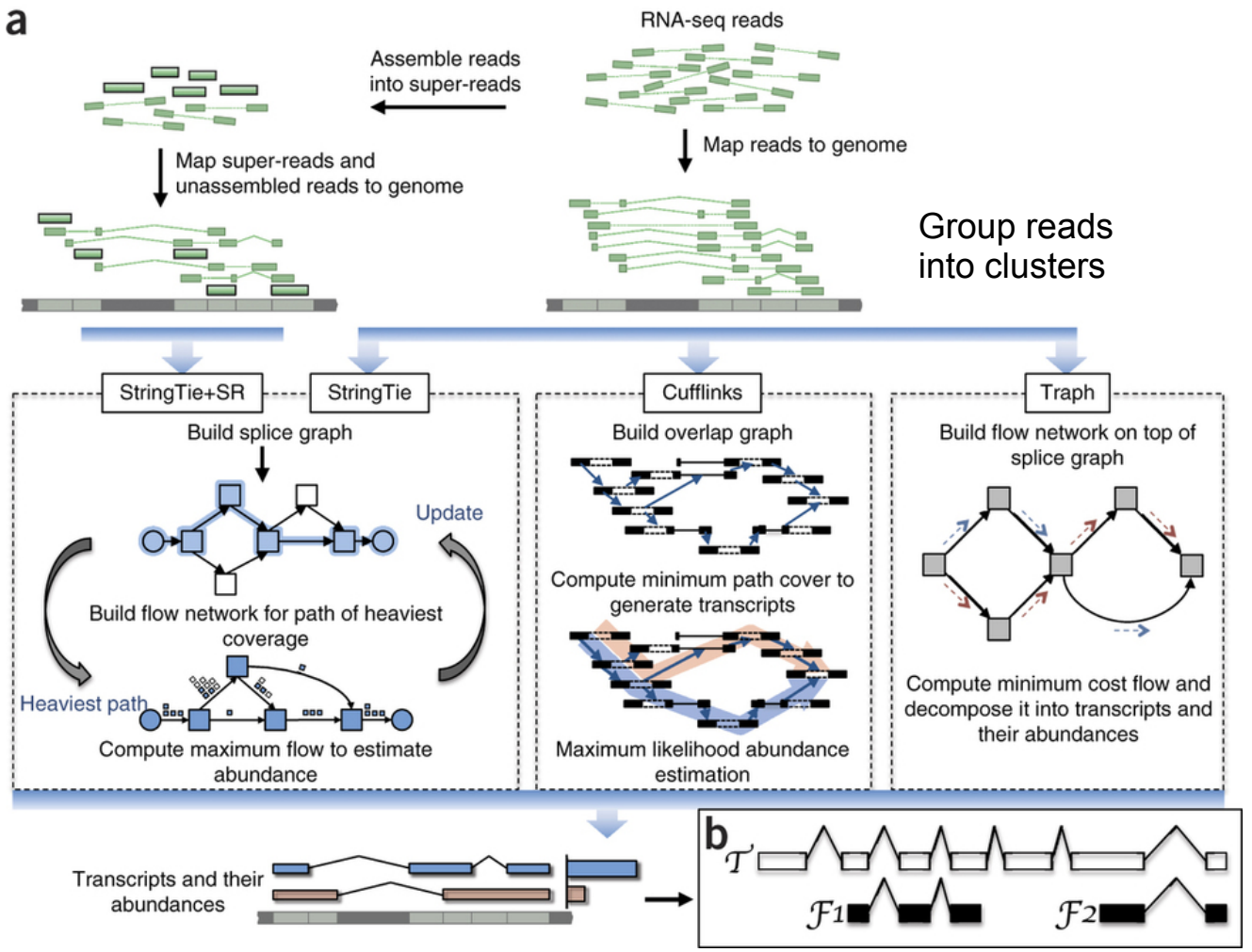
How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:
  - FPKM
    - 1) Determine total sample fragment count and divide by 1,000,000
      - "per million" scaling factor
    - 2) Divide each gene/transcript fragment count by #1
      - fragments per million, FPM
    - 3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)
  - TPM
    - 1) Divide each fragment count by length of each transcript in kilobases
      - fragments per kilobase, FPK
    - 2) Sum all FPK values for the sample and divide by 1,000,000
      - "per million" scaling factor
    - 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

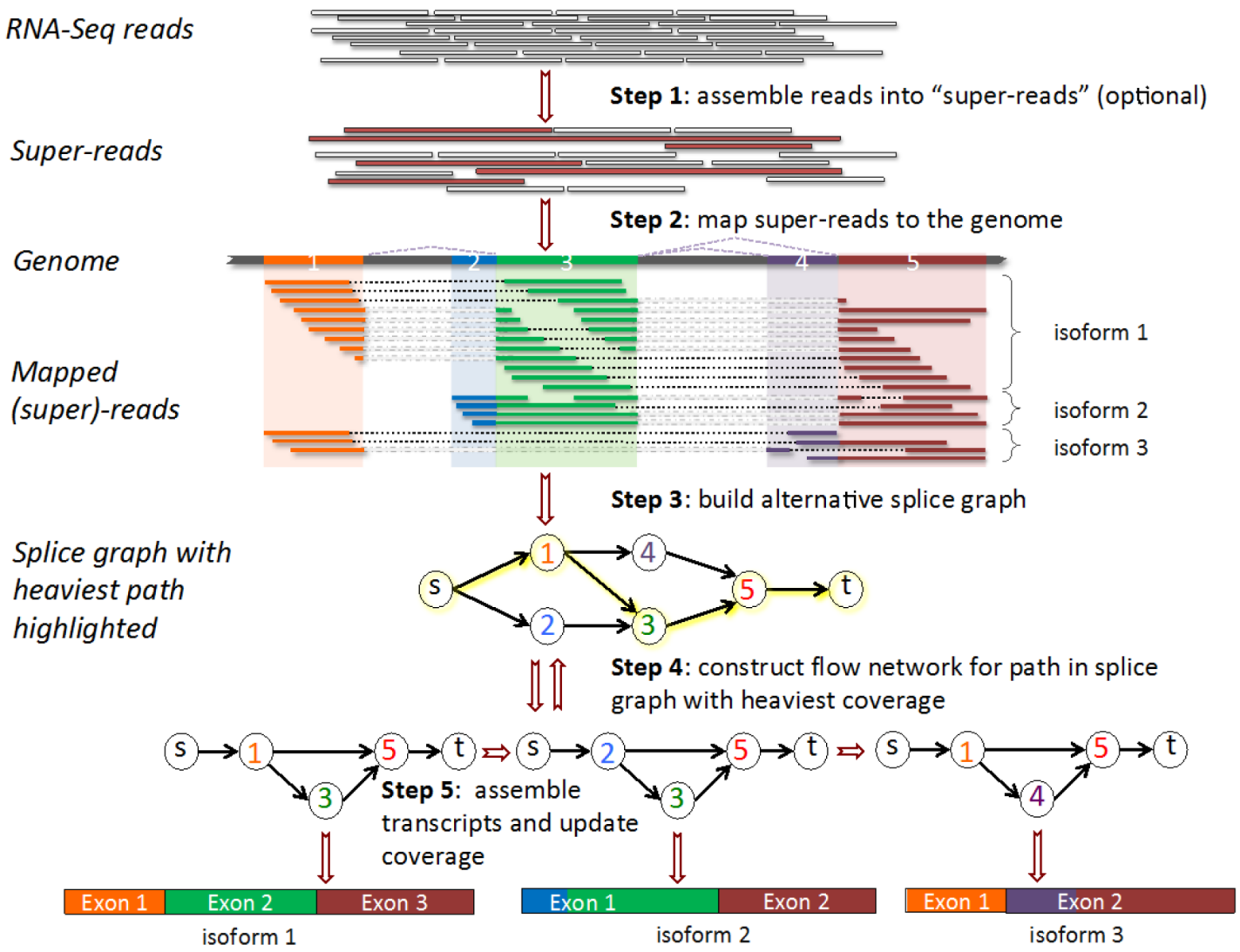
How does StringTie work?

StringTie iteratively extracts the heaviest path from a splice graph, constructs a flow network, computes maximum flow to estimate abundance, and then updates the splice graph by removing reads that were assigned by the flow algorithm. This process repeats until all reads have been assigned.

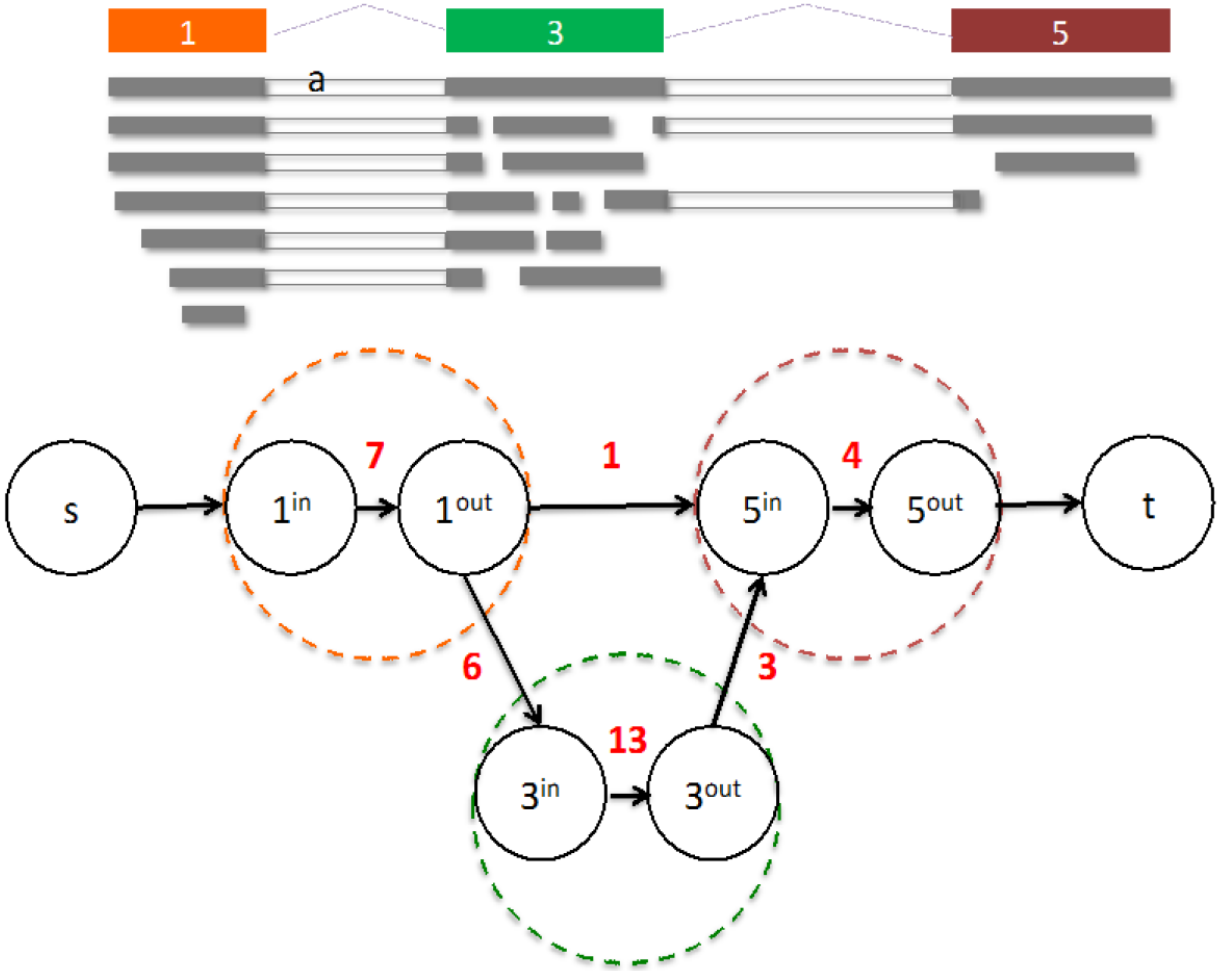


Pertea et al. Nature Biotechnology, 2015

Construct splice graph, identify path with heaviest coverage, construct flow network, assemble transcript, remove reads and repeat



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 defintions and math.

StringTie -merge

- Merge together all gene structures from all samples
  - Some samples may only partially represent a gene structure
- Allows for the incorporation of 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

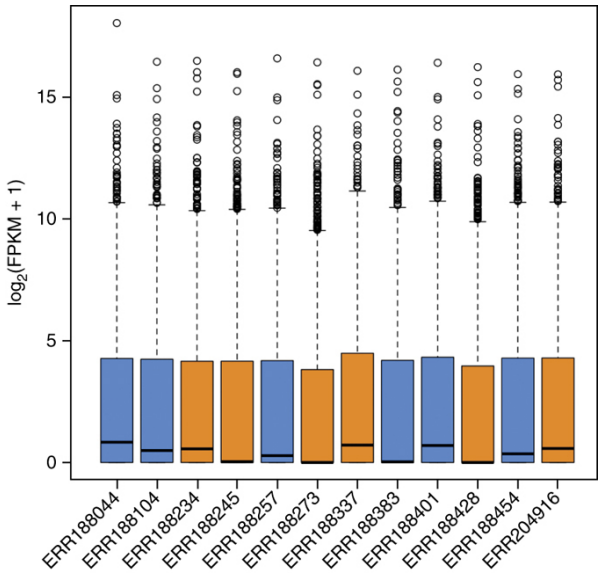
Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible 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)

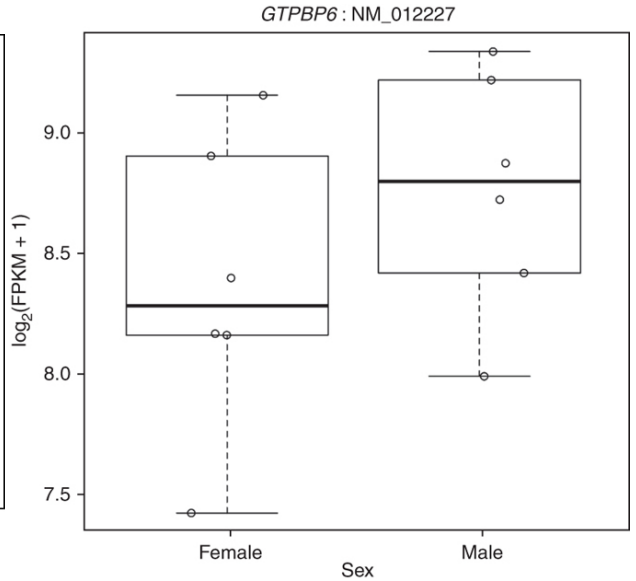
Ballgown for Differential Expression

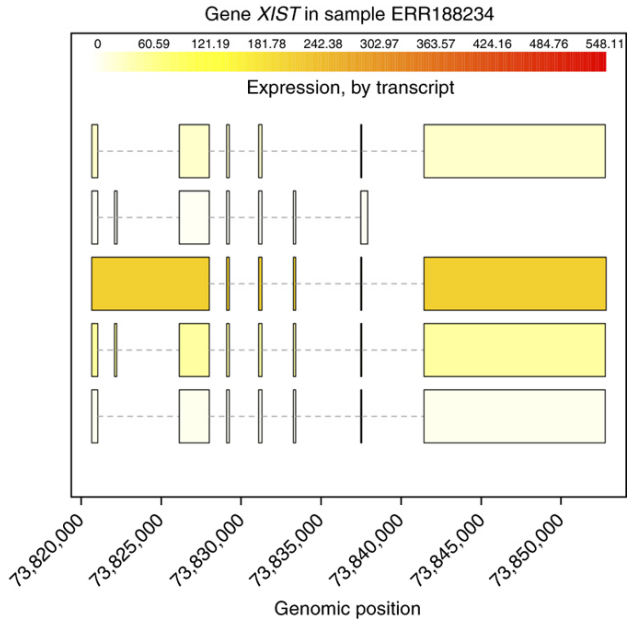
- Parametric F-test comparing nested linear models
- Two models are fit to each feature, using expression as the outcome
  - one including the covariate of interest (e.g., case/control status or time) and one not including that covariate.
- An F statistic and p-value are calculated using the fits of 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.
- We adjust for multiple testing by reporting q-values:
  - $q < 0.05$ the false discovery rate should be controlled at ~5%.

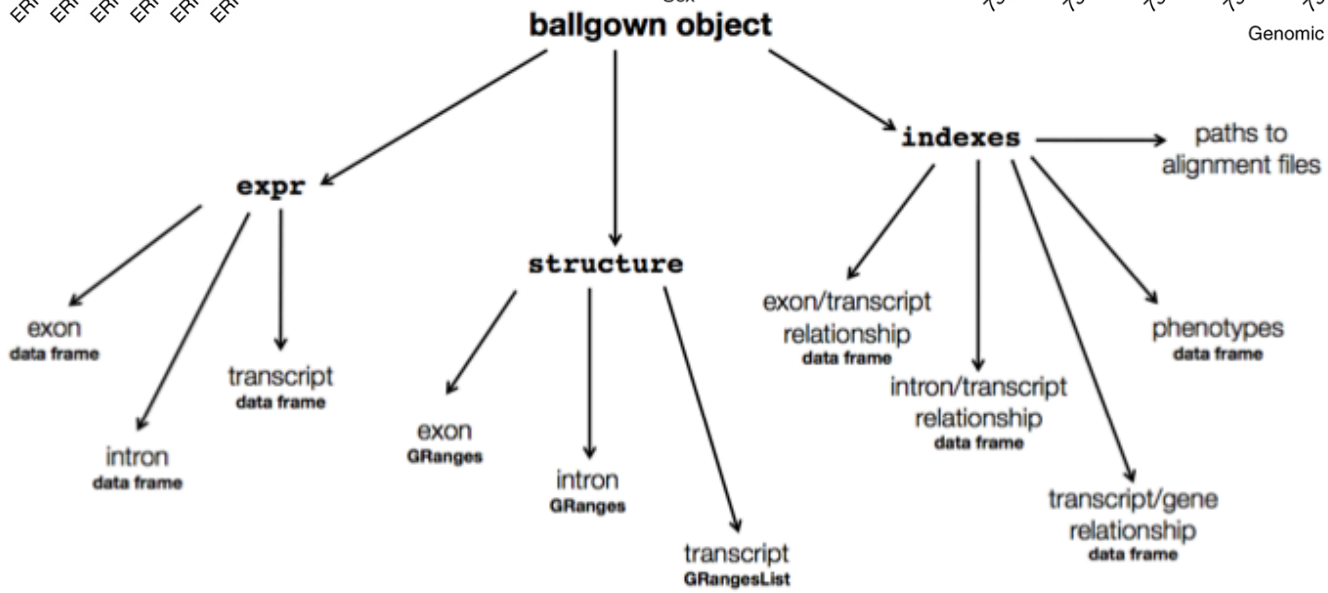
Frazee et al. (2014)

Ballgown for Visualization with R





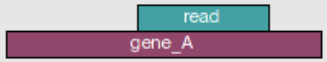
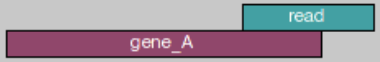
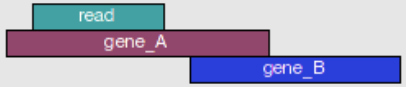
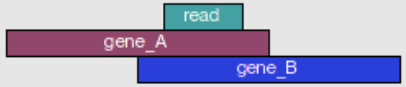




Alternatives to FPKM

- Raw read counts as an alternate for differential expression analysis
  - Instead of calculating FPKM, simply assign reads/fragments to a defined set of genes/transcripts and determine "raw counts"
    - Transcript structures could still be defined by something like cufflinks
- HTSeq (htseq-count)
  - http://www-huber.embl.de/users/anders/HTSeq/doc/count.html
  - htseq-count --mode intersection-strict --stranded no --minaqual 1 --type exon --idattr transcript_id accepted_hits.sam chr22.gff > transcript_read_counts_table.tsv
  - Important caveat of 'transcript' analysis by htseq-count:
    - 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

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

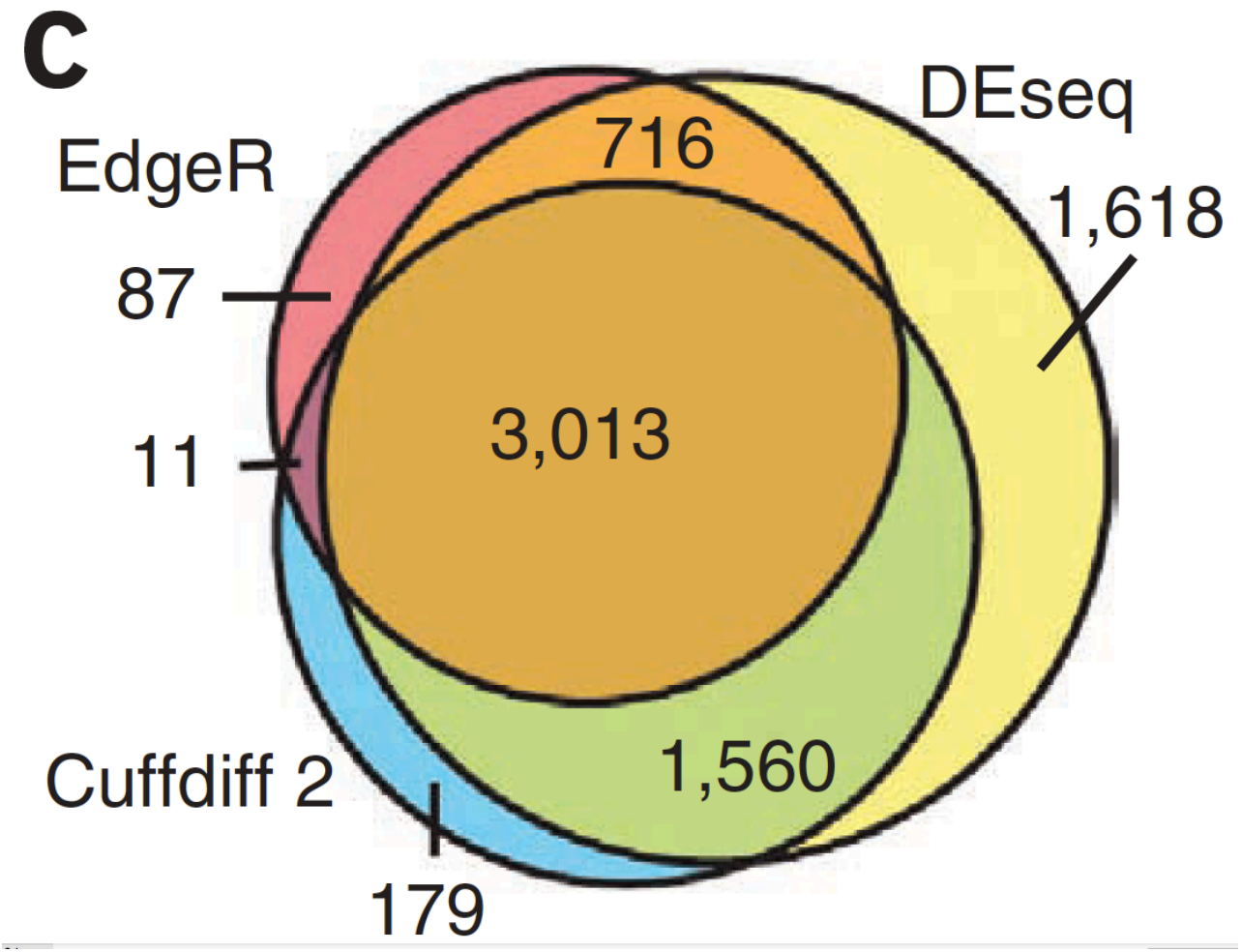
Alternative differential expression methods

- 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’ expression estimates vs. ‘raw’ counts

- Which should I use?
  - Long running debate with countless blogs and analyses arguing the advantages of each. The general consensus:
- FPKM
  - 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
  - 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://euler.bc.edu/marthlab/scotty/scotty.php
- 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

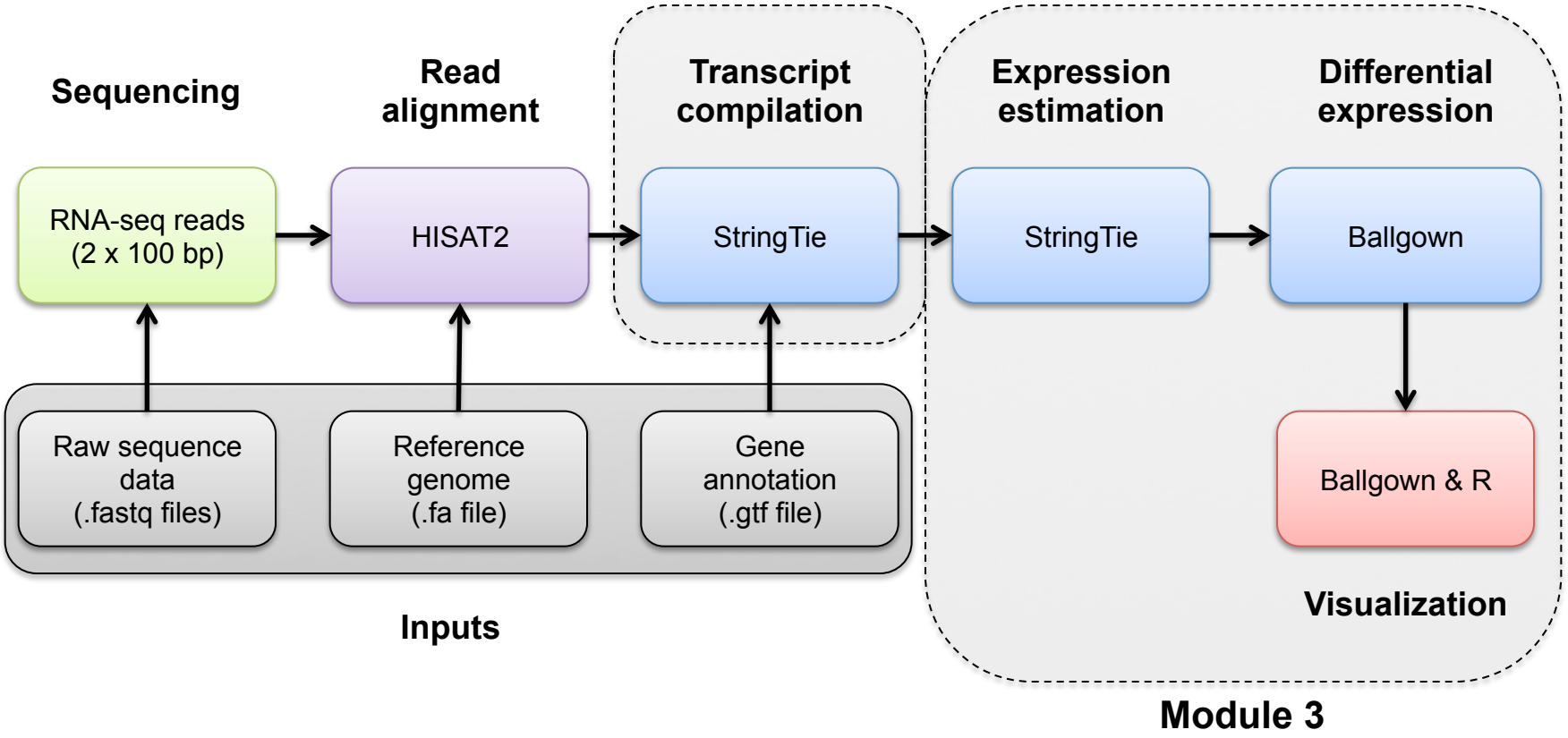
- As more attributes are compared, it becomes more likely that the treatment and control groups will appear to differ on at least one attribute by random chance alone.
- Well known from array studies
  - 10,000s genes/transcripts
  - 100,000s exons
- With RNA-seq, more of a problem than ever
  - All the complexity of the transcriptome
  - Almost infinite number 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 cummeRbund
  - 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
  - IPA
  - Cytoscape
  - Many R/BioConductor packages: http://www.bioconductor.org/help/search/index.html?q=pathway

Introduction to tutorial (Module 3)

HISAT2/StringTie/Ballgown RNA-seq Pipeline



We are on a Coffee Break & Networking Session