



Cold  
Spring  
Harbor  
Laboratory

# Advanced Sequencing Technologies & Applications

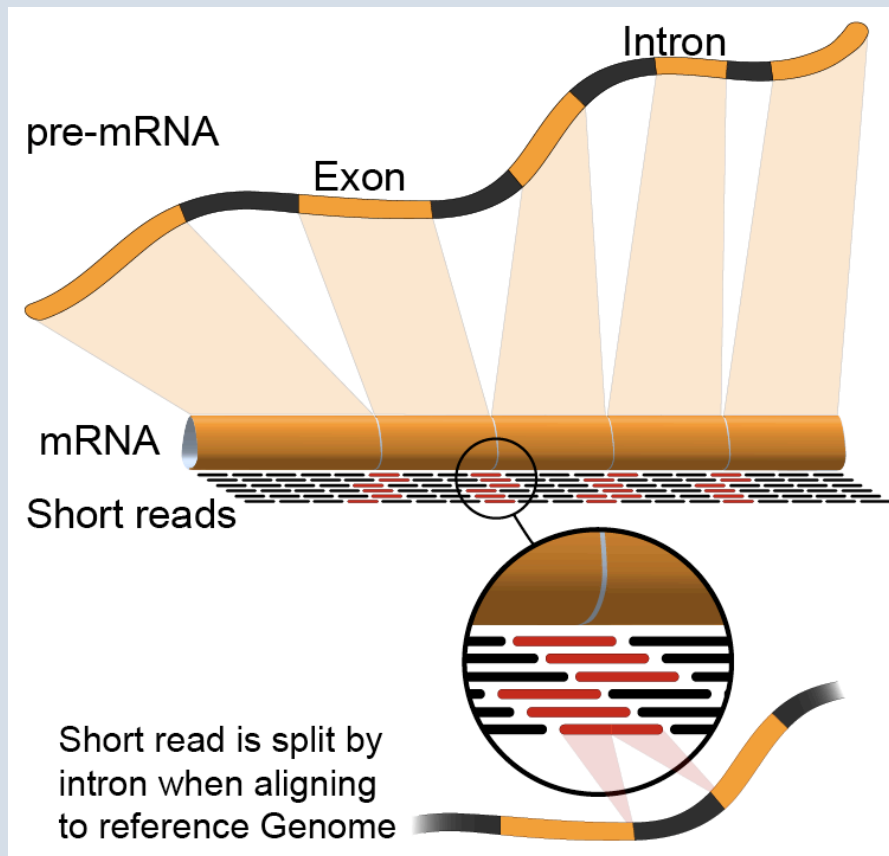
<http://meetings.cshl.edu/courses.html>



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 - 20, 2016



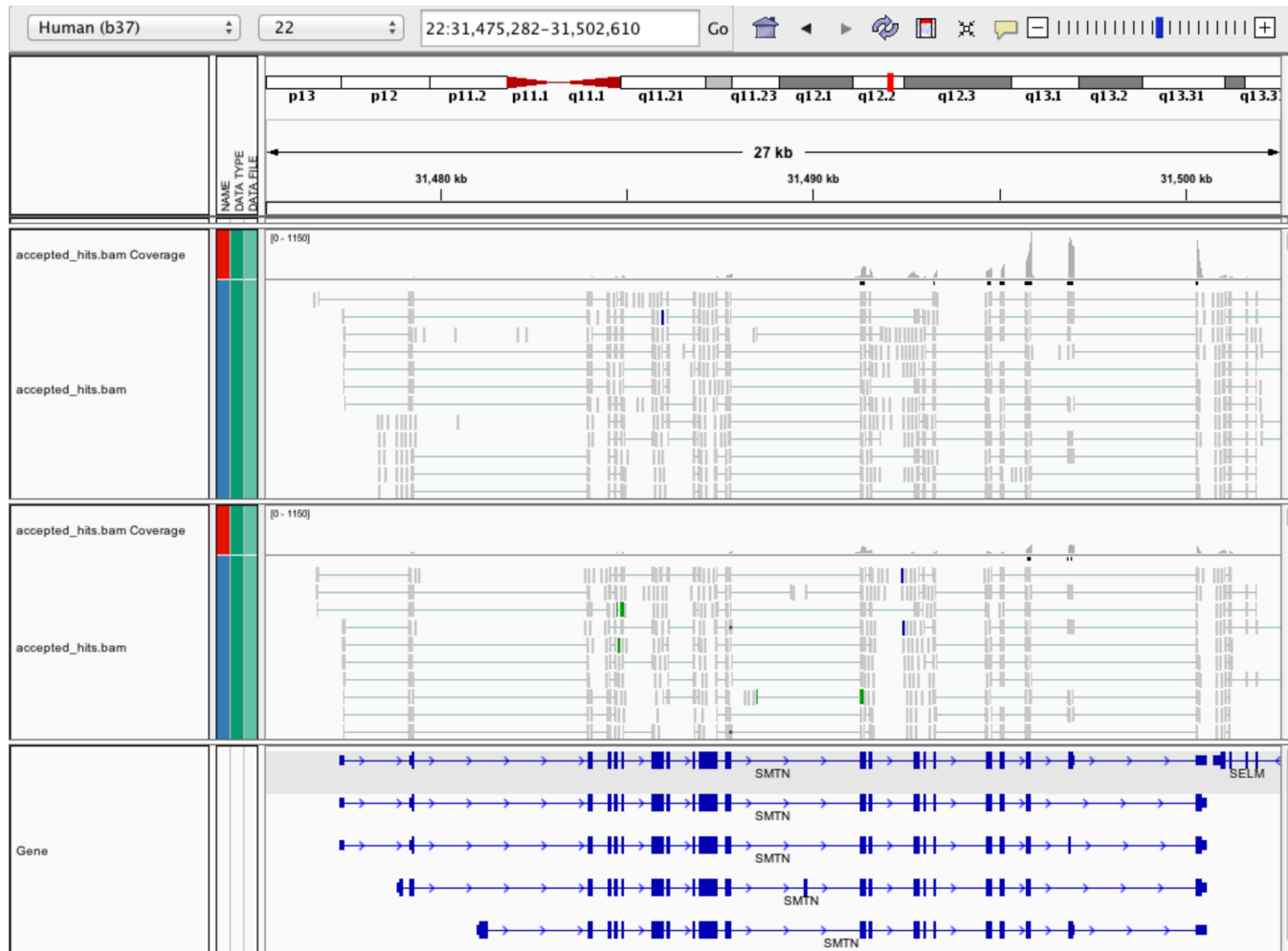
# Learning objectives of the course

- 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  
→

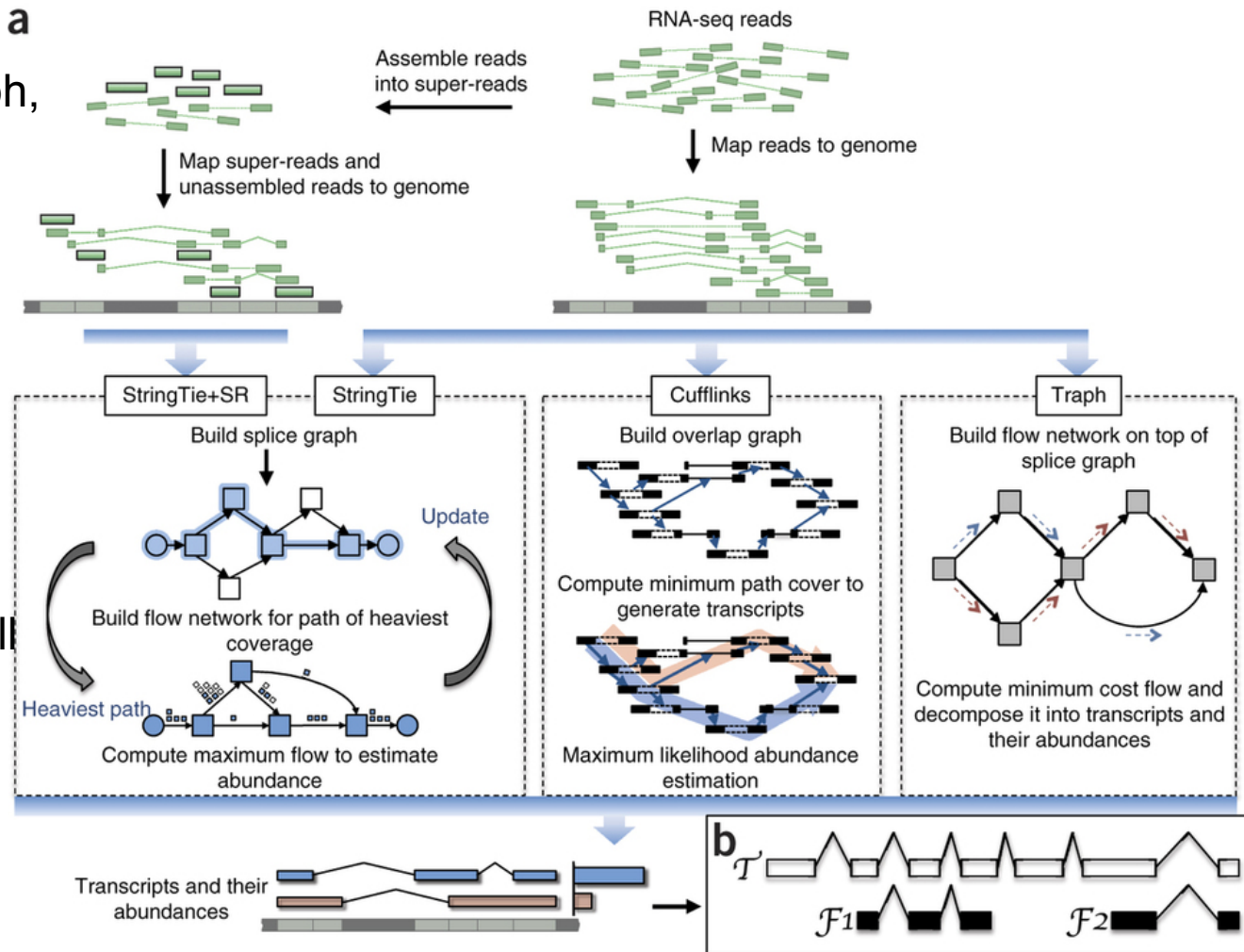
↓  
Down-regulated

# 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 (or RPKM) attempt to normalize for gene size and library depth
- $$\text{RPKM (or FPKM)} = (10^9 * C) / (N * L)$$
  - 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 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.
- Annotated transcript  $T$  for which read data covers only the fragments  $F1$  and  $F2$ .



# StringTie –merge & gffcompare for

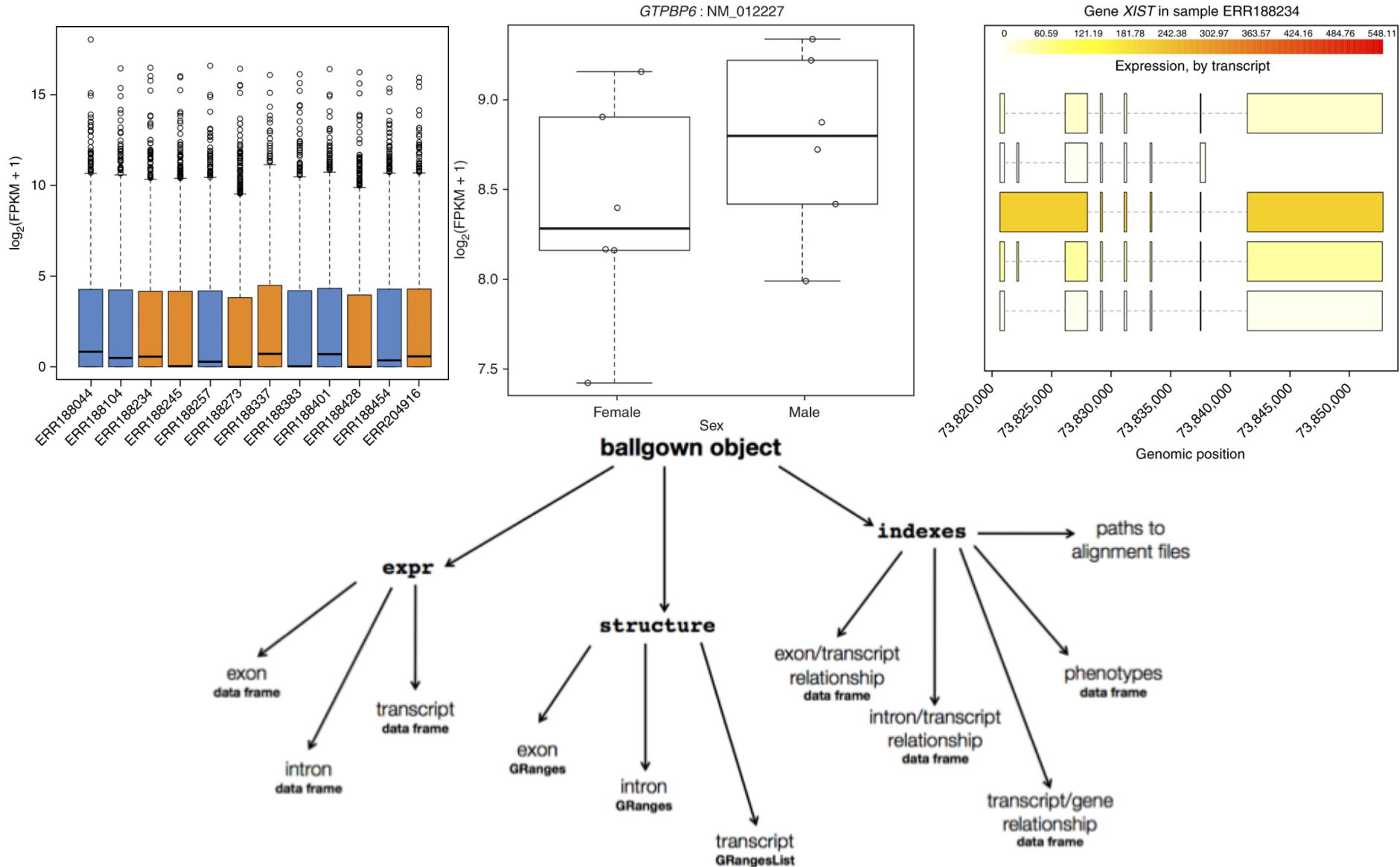
- --merge allows for the incorporation of known transcripts with assembled, potentially novel transcripts
- Gffcompare will compare a merged transcript GTF with known annotation, also in GTF/GFF3 format
  - [INSERT class codes]



# 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%.

# 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 --minqual 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>

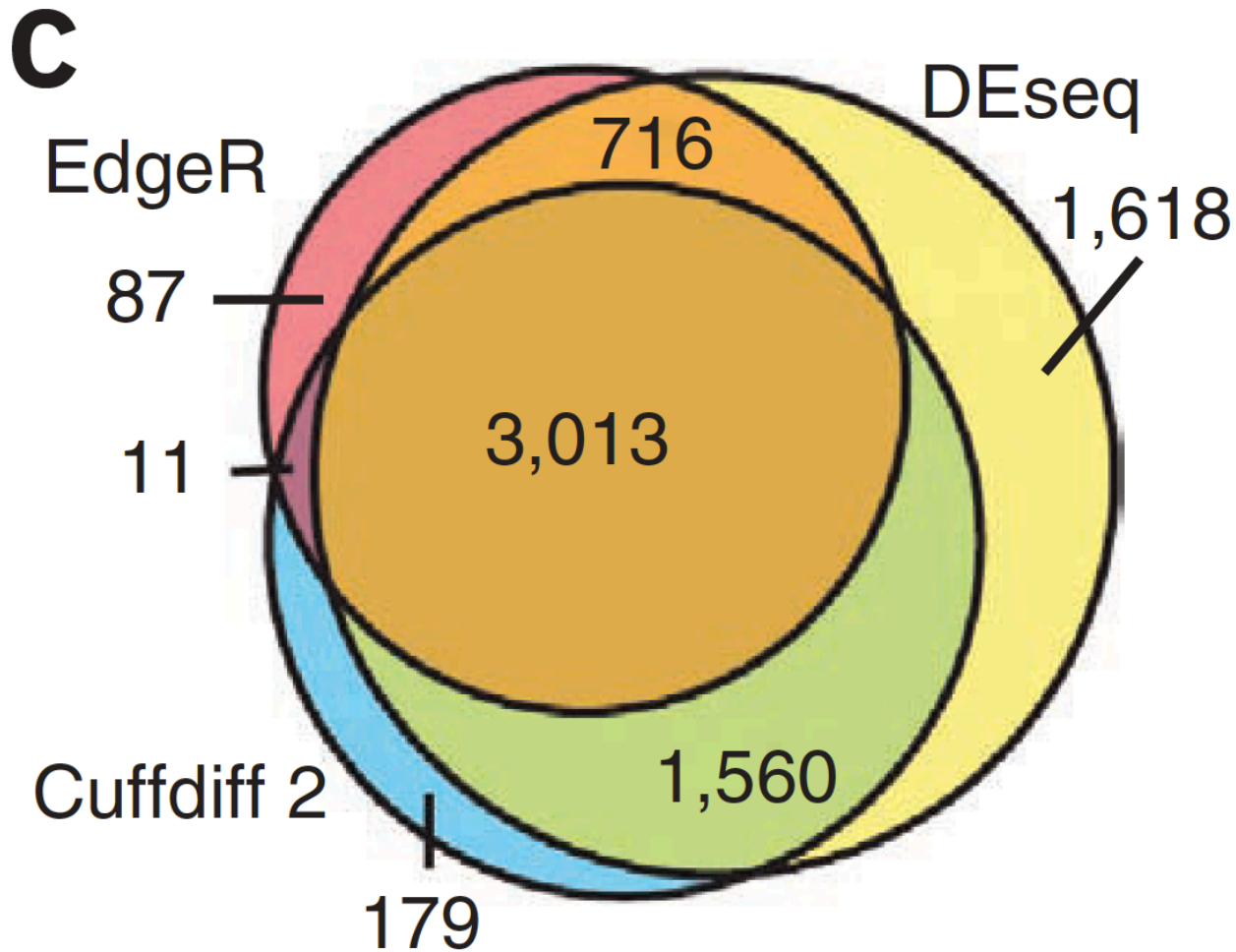
# **‘FPKM’ expression estimates vs. ‘raw’ counts**

- Which should I use?
- FPKM
  - When you want to leverage benefits of tuxedo suite
  - 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

# Alternative differential expression methods

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

# 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>

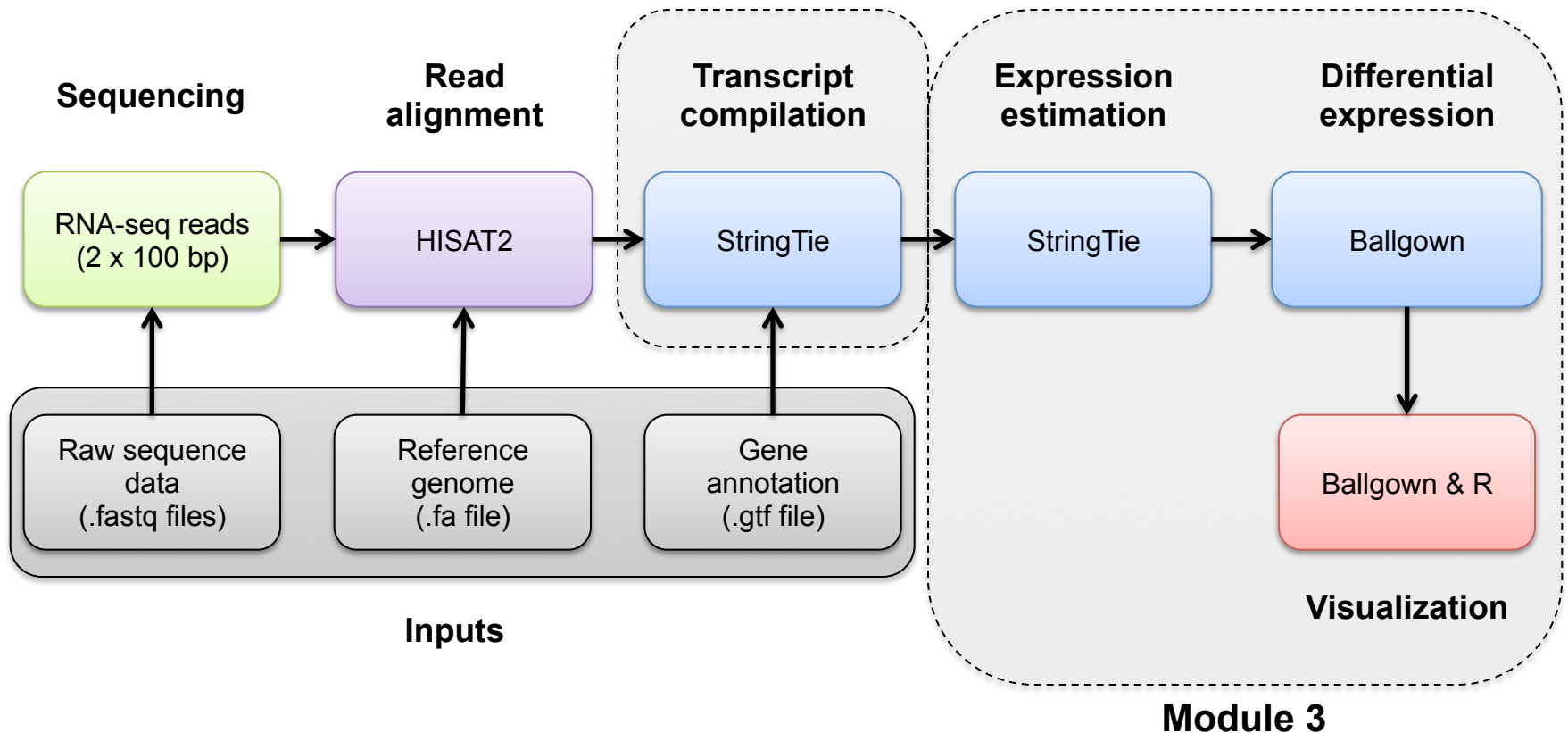


# Downstream interpretation of expression analysis

- 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



Break