



Cold
Spring
Harbor
Laboratory

Advanced Sequencing Technologies & Applications

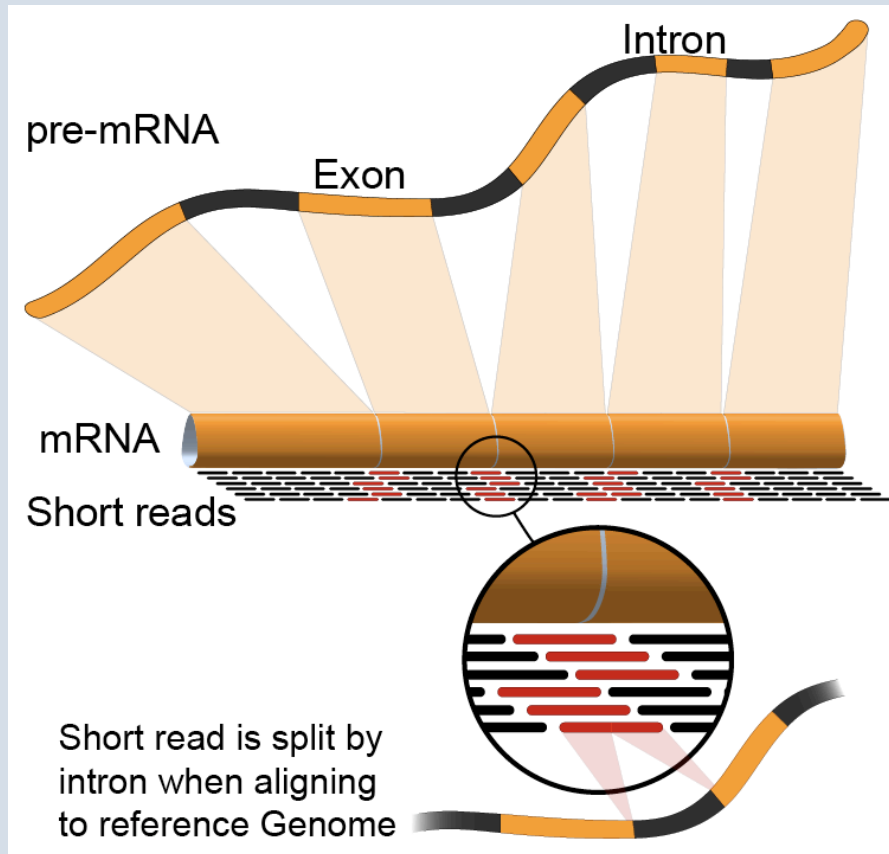
<http://meetings.cshl.edu/courses.html>



Cold
Spring
Harbor
Laboratory

Module 4 Expression and Differential Expression (lecture)

Malachi Griffith, Obi Griffith, Jason Walker, Ben Ainscough
Advanced Sequencing Technologies & Applications
November 11-23, 2014



Learning objectives of the course

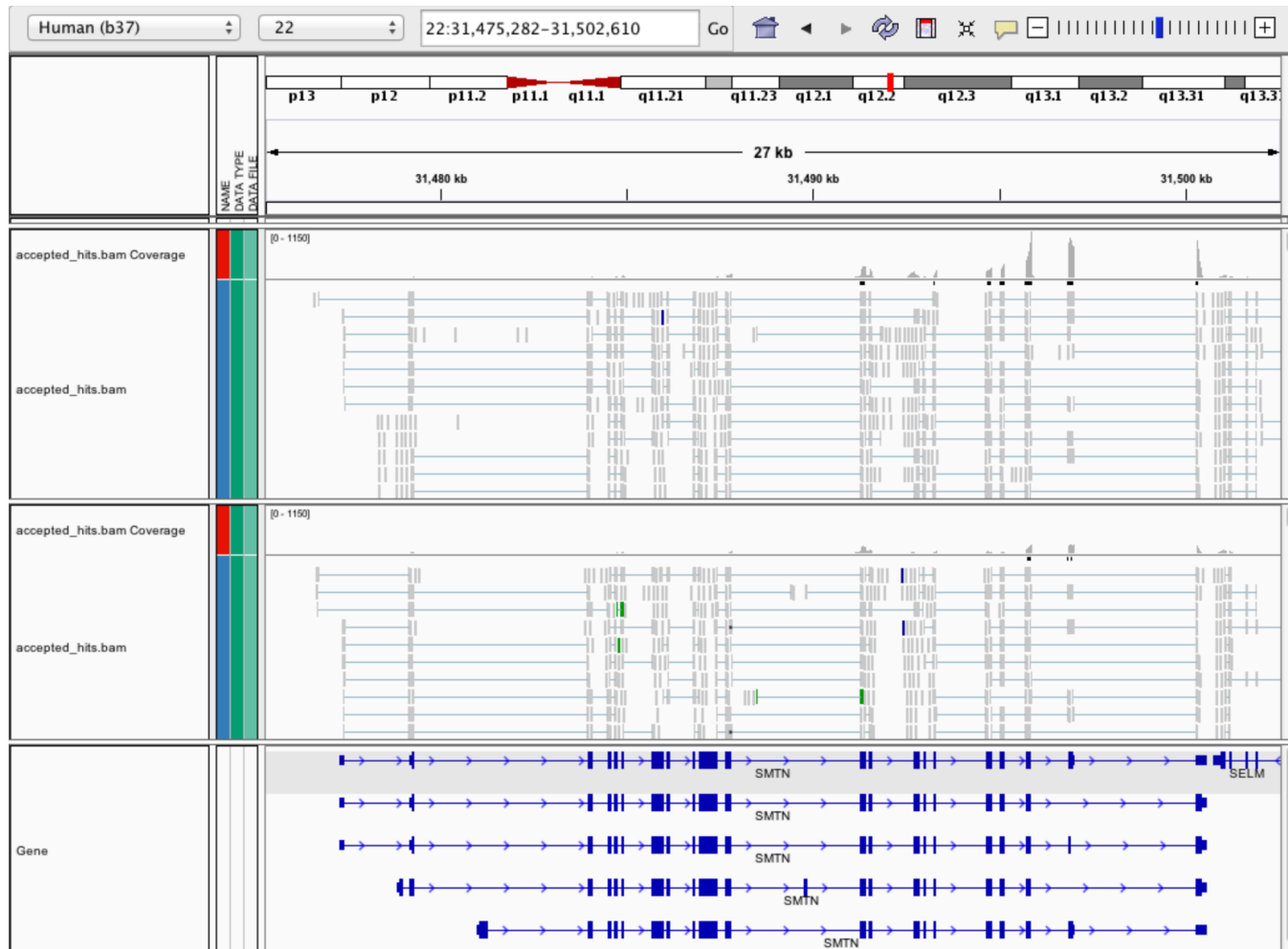
- Module 1: Introduction to cloud computing
- Module 2: Introduction to RNA sequencing
- Module 3: RNA-seq alignment and visualization
- **Module 4: Expression and Differential Expression**
- Module 5: Isoform discovery and alternative expression

- 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

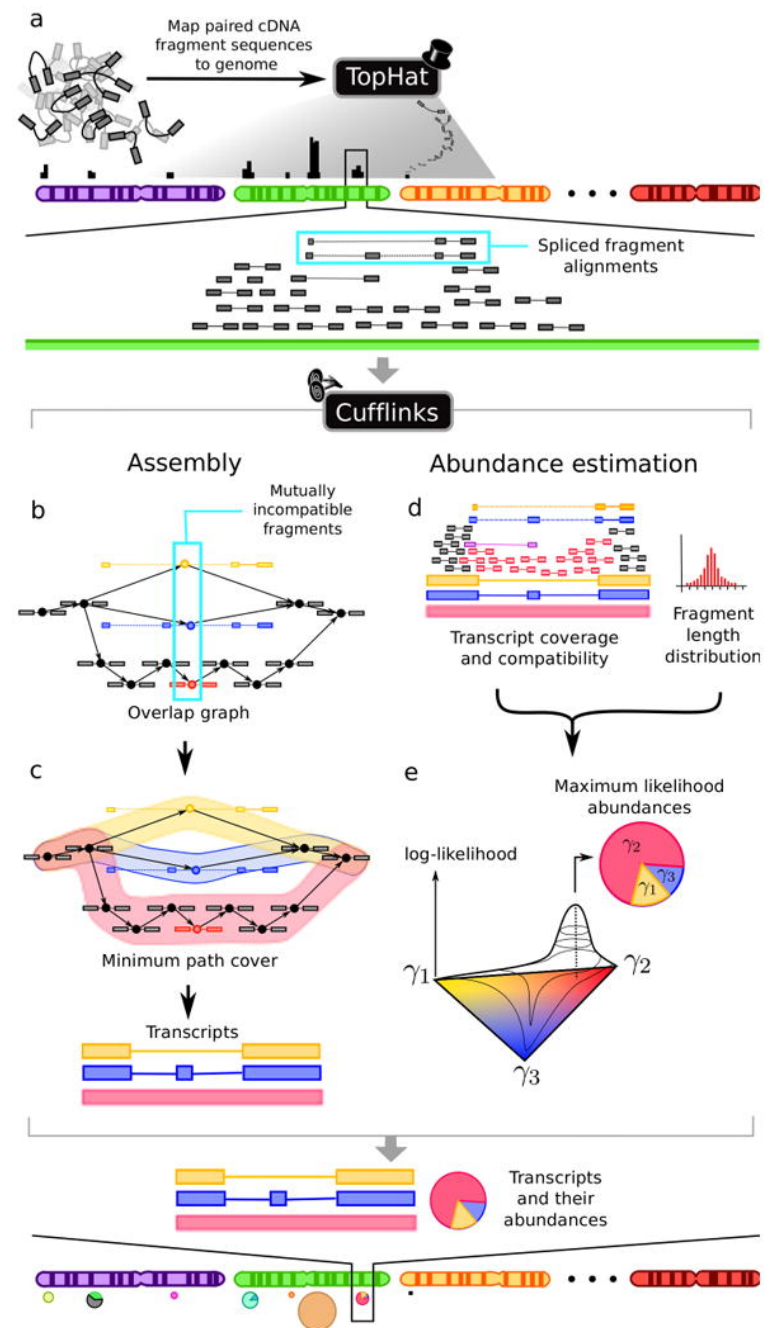
What is FPKM (RPKM)

- 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
- $RPKM/FPKM = (10^9 * C) / (N * L)$
 - 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 cufflinks work?

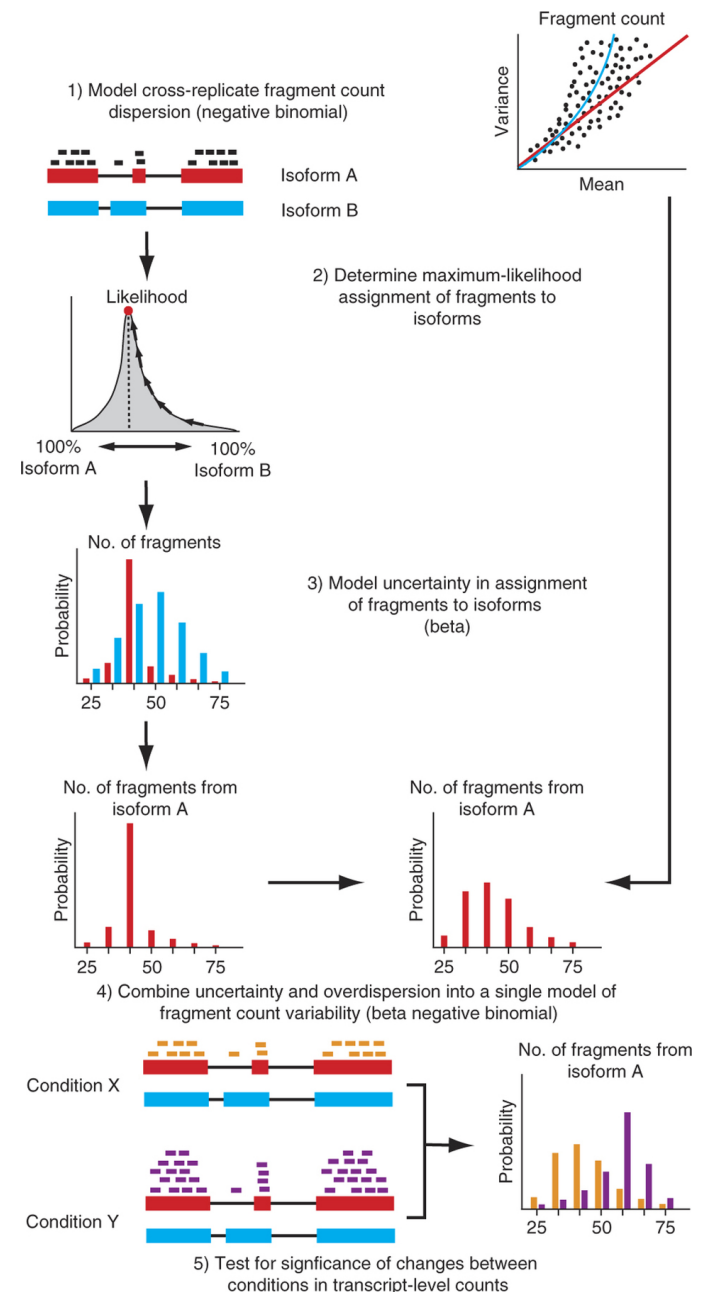
- Overlapping 'bundles' of fragment alignments are assembled, fragments are connected in an overlap graph, transcript isoforms are inferred from the minimum paths required to cover the graph
- Abundance of each isoform is estimated with a maximum likelihood probabilistic model
 - makes use of information such as fragment length distribution

<http://cufflinks.cbc.umd.edu/howitworks.html>



How does cuffdiff work?

- The variability in fragment count for each gene across replicates is modeled.
- The fragment count for each isoform is estimated in each replicate (as before), along with a measure of uncertainty in this estimate arising from ambiguously mapped reads
 - transcripts with more shared exons and few uniquely assigned fragments will have greater uncertainty
- The algorithm combines estimates of uncertainty and cross-replicate variability under a beta negative binomial model of fragment count variability to estimate count variances for each transcript in each library
- These variance estimates are used during statistical testing to report significantly differentially expressed genes and transcripts.



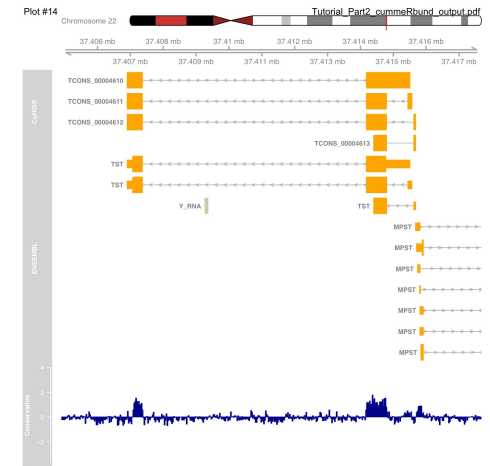
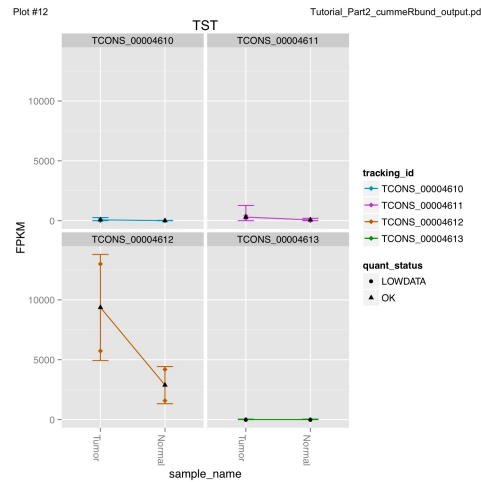
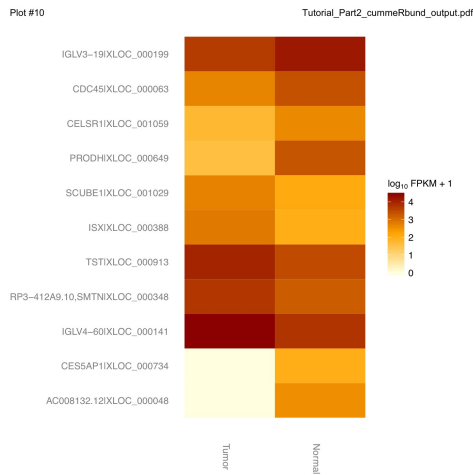
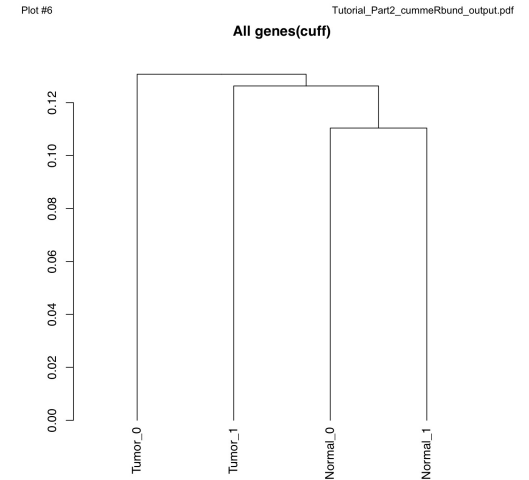
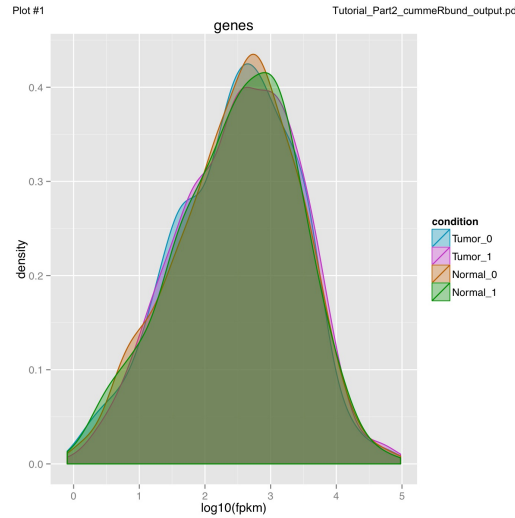
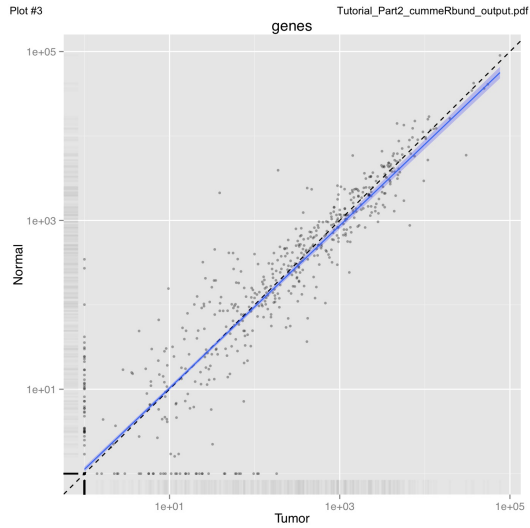
Why is cuffmerge necessary?

- Cuffmerge
 - Allows merge of several Cufflinks assemblies together
 - Necessary because even with replicates cufflinks will not necessarily assemble the same numbers and structures of transcripts
 - Filters a number of transfrags that are probably artifacts.
 - Optional: provide reference GTF to merge novel isoforms and known isoforms and maximize overall assembly quality.
 - Make an assembly GTF file suitable for use with Cuffdiff
 - Compare apples to apples

What do we get from cummeRbund?

- Automatically generates many of the commonly used data visualizations
- Distribution plots
- Overall correlations plots
- MA plots
- Volcano plots
- Clustering, PCA and MDS plots to assess global relationships between conditions
- Heatmaps
- Gene/transcript-level plots showing transcript structures and expression levels

What do we get from cummeRbund?



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>

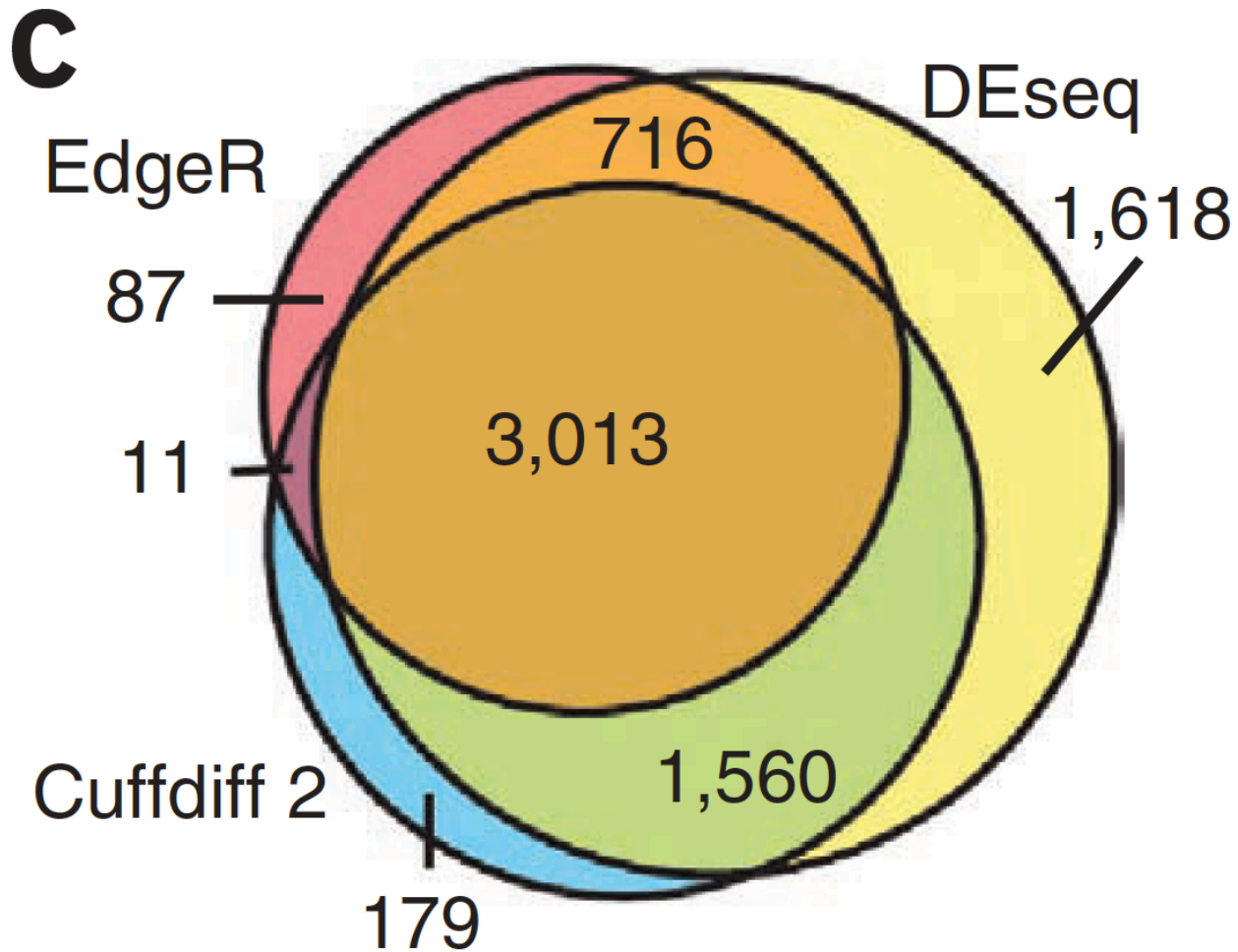
'FPKM' expression estimates vs. 'raw' counts

- 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

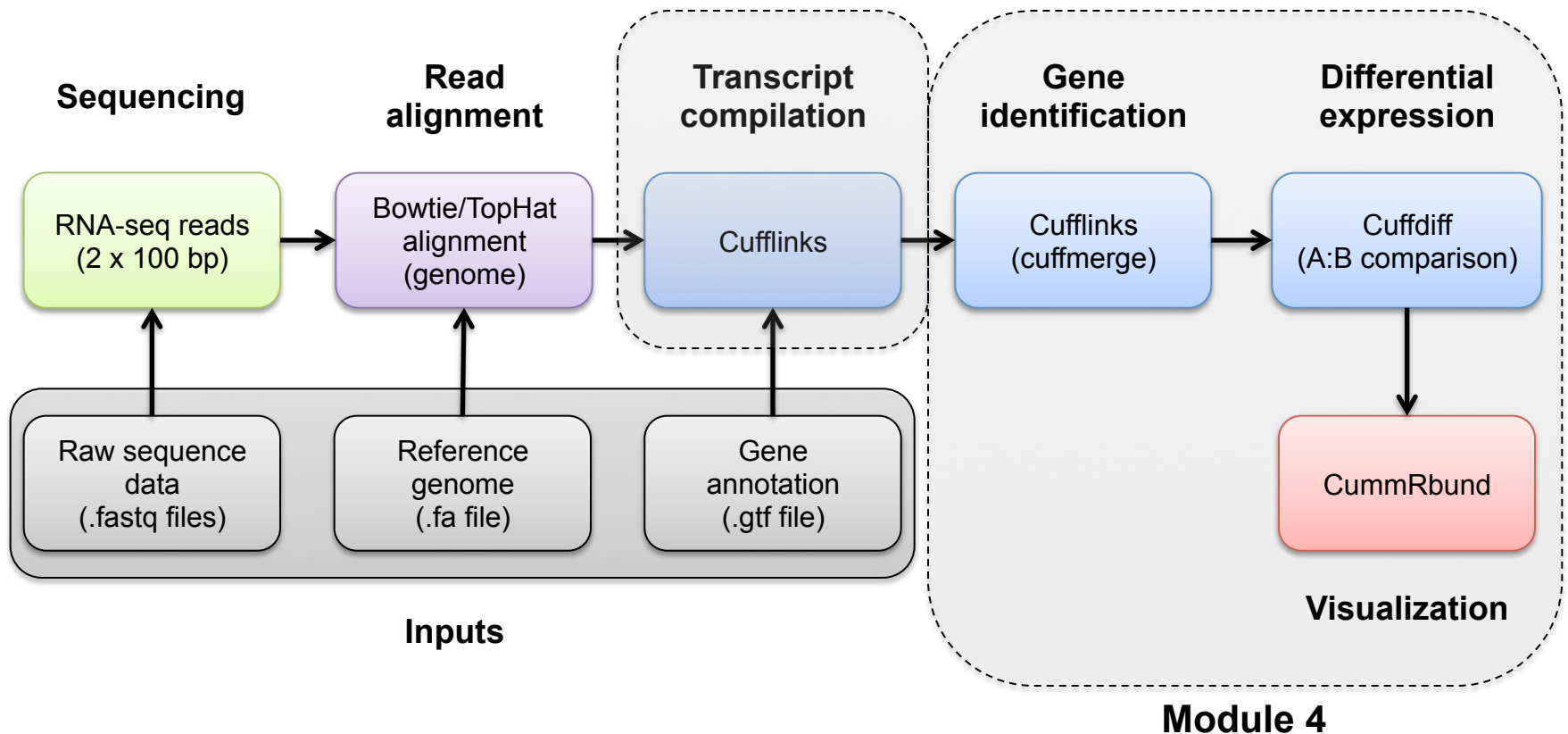
- As more attributes are compared, it becomes more likely that the treatment and control groups will appear to differ on at least one attribute by random chance alone.
- Well known from array studies
 - 10,000s genes/transcripts
 - 100,000s exons
- With RNA-seq, more of a problem than ever
 - All the complexity of the transcriptome
 - Almost infinite number of potential features
 - Genes, transcripts, exons, junctions, retained introns, microRNAs, lncRNAs, etc, 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 cufflinks/cuffdiff (or alternative) can be fed into many analysis pipelines
- See supplemental R tutorial for how to format cufflinks 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
 - David
 - IPA
 - Cytoscape
 - Many R/BioConductor packages: <http://www.bioconductor.org/help/search/index.html?q=pathway>

Introduction to tutorial (Module 4)

Bowtie/TopHat/Cufflinks/Cuffdiff RNA-seq Pipeline



Break