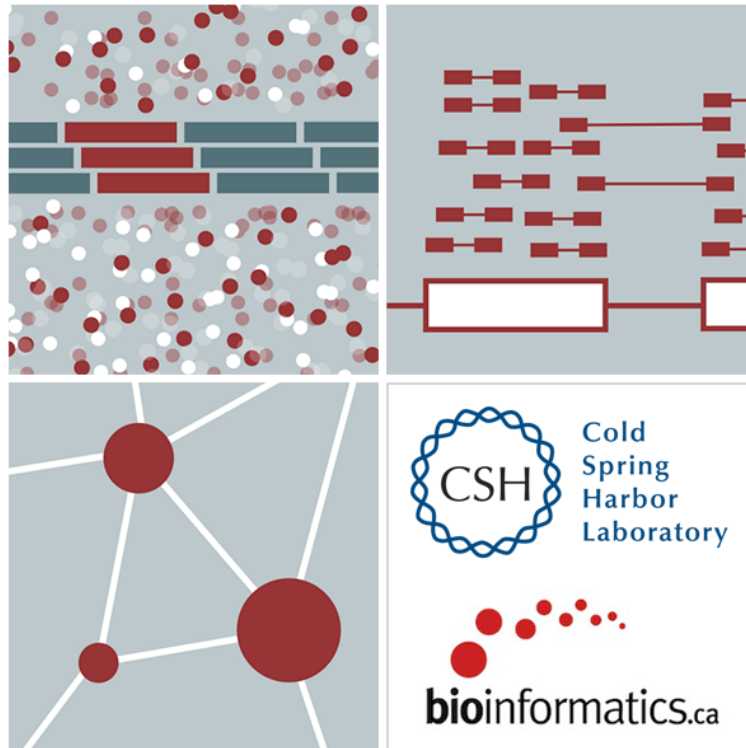


RNA-Seq Module 11

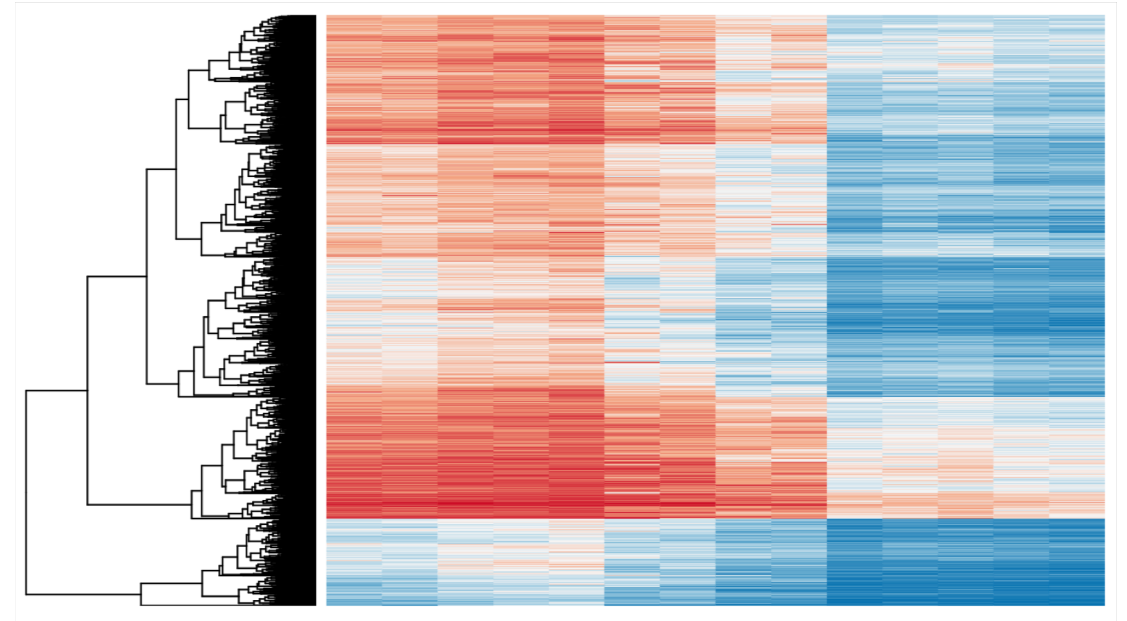
Differential Expression

Kelsy Cotto, Malachi Griffith, Chris Miller, Peter Ronning



Differential Expression

- Tying gene expression back to genotype/phenotype
- What genes/transcripts are being expressed at higher/lower levels in different groups of samples?
- Examples:
 - Wild-type vs gene KO mouse
 - Tumor vs Normal tissue
 - Good responders vs Poor responders



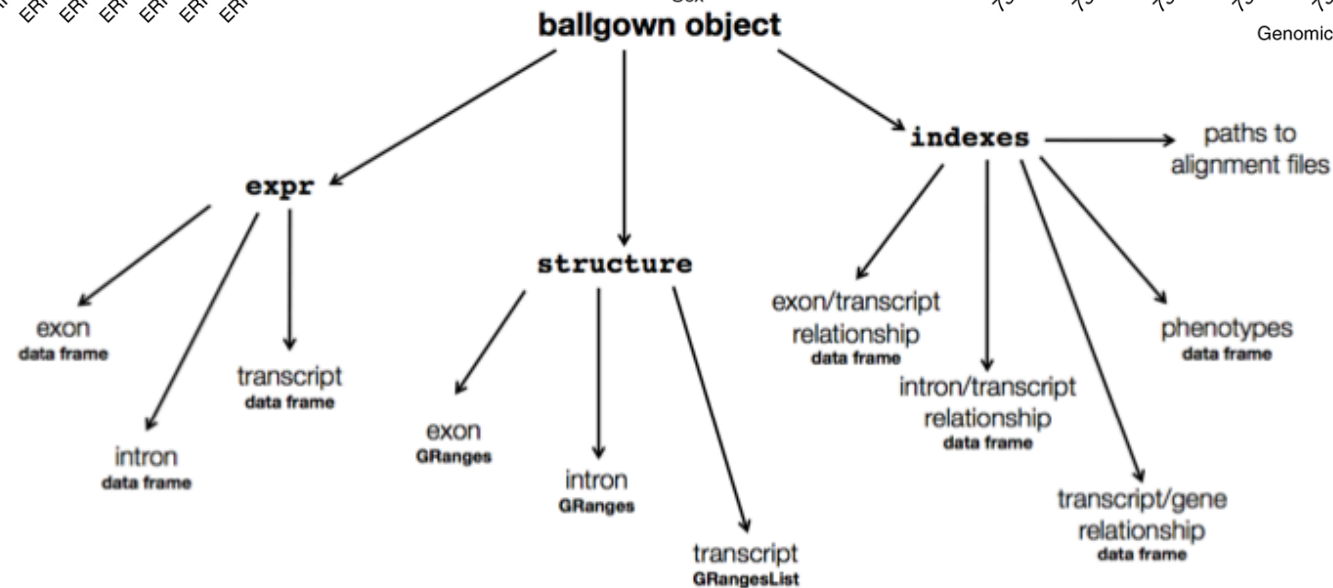
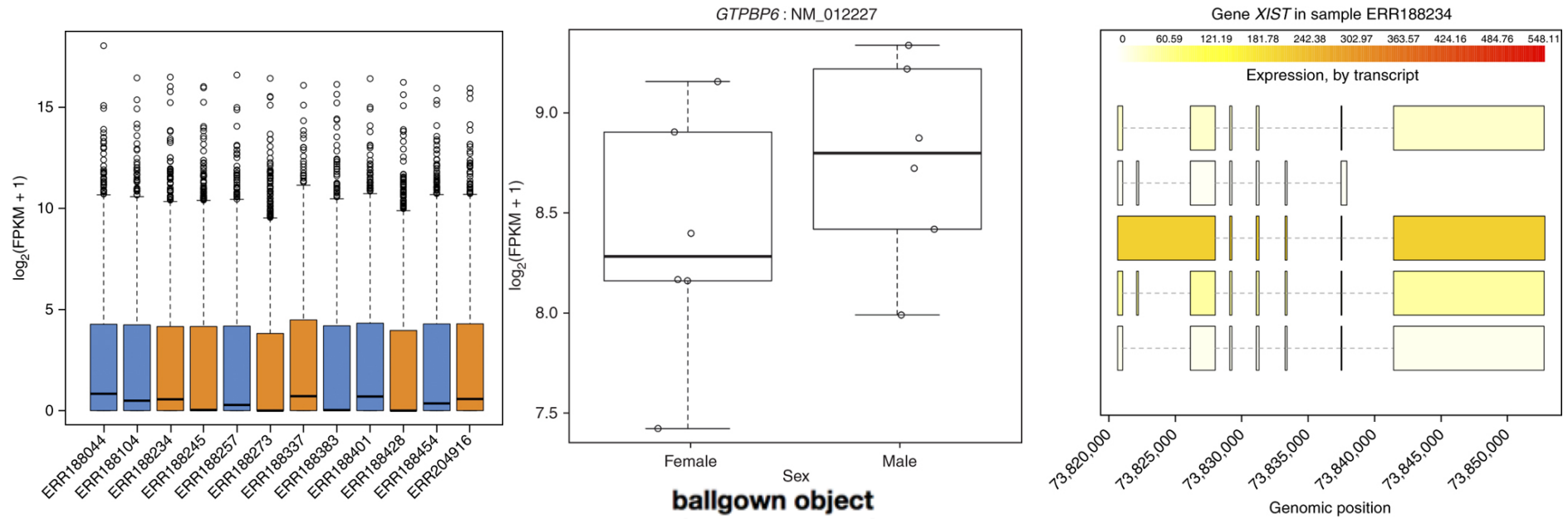
Differential Expression with Ballgown

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.
- We adjust for multiple testing by reporting q-values:
 - $q < 0.05$ the false discovery rate should be controlled at $\sim 5\%$.

[Frazee et al. \(2014\)](#)

Ballgown for Visualization with R



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 cufflinks

- HTSeq (htseq-count)

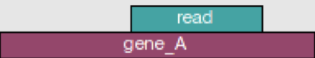
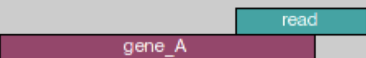



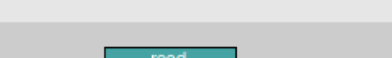

- <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

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

- 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

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

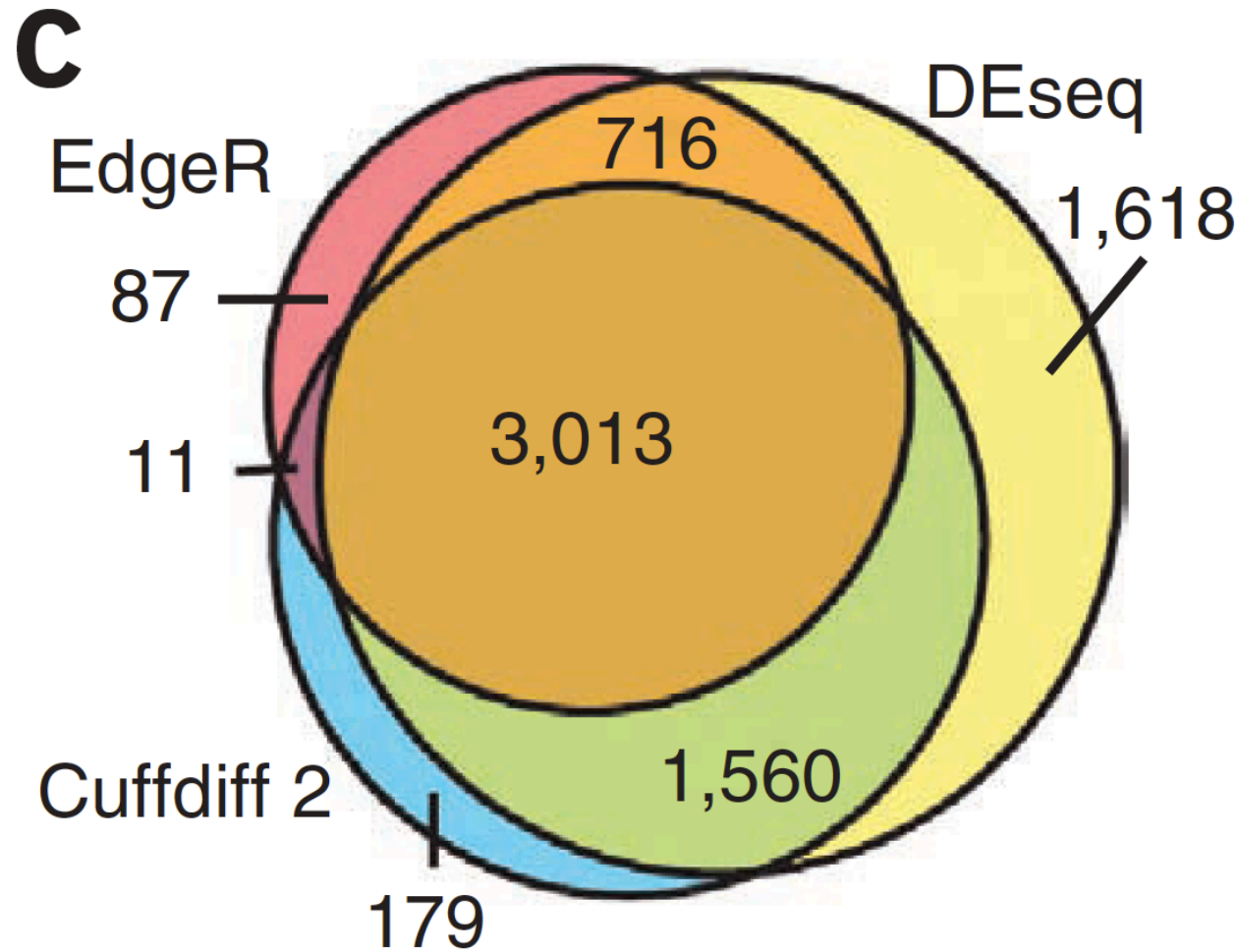
Alternative differential expression methods

- 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/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
 - 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, 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 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>

https://genviz.org/module%204/0003/12/31/Expression_Profiling_and_Visualization/

HISAT2/StringTie/Ballgown RNA-seq Pipeline

