

Canadian Bioinformatics Workshops

www.bioinformatics.ca
bioinformaticsdotca.github.io



CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Canonical URL: <https://creativecommons.org/licenses/by-sa/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

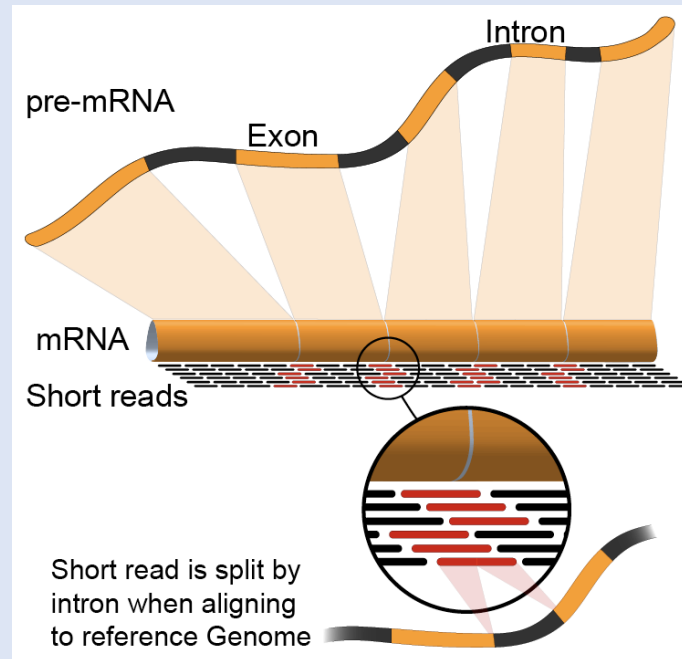
Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

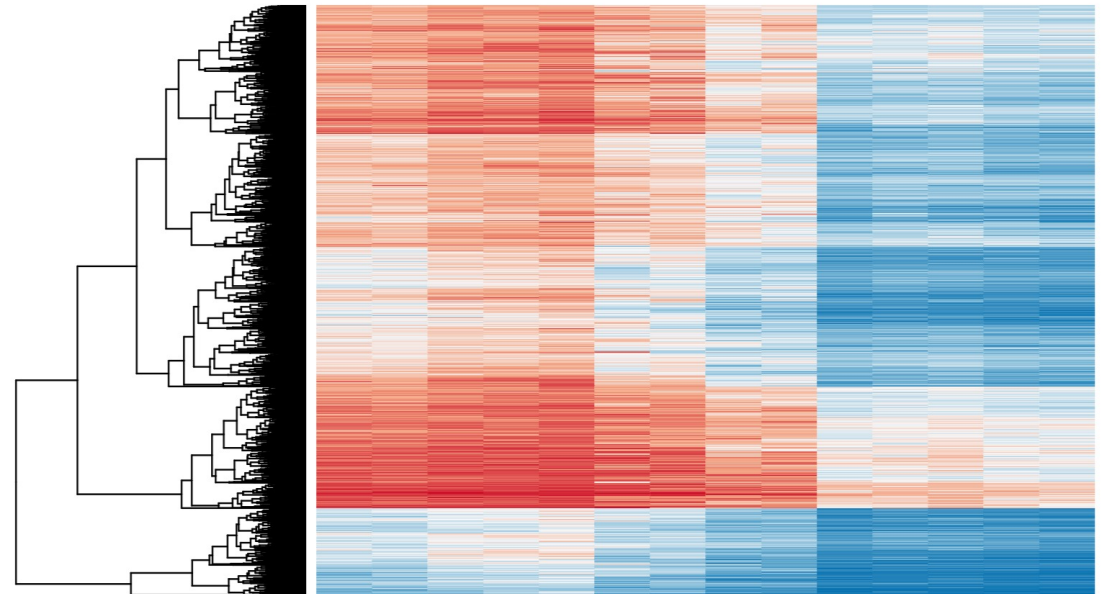
RNA-Seq Module 3: Differential Expression

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian
RNA-seq Analysis 2024. June 17-19, 2024



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



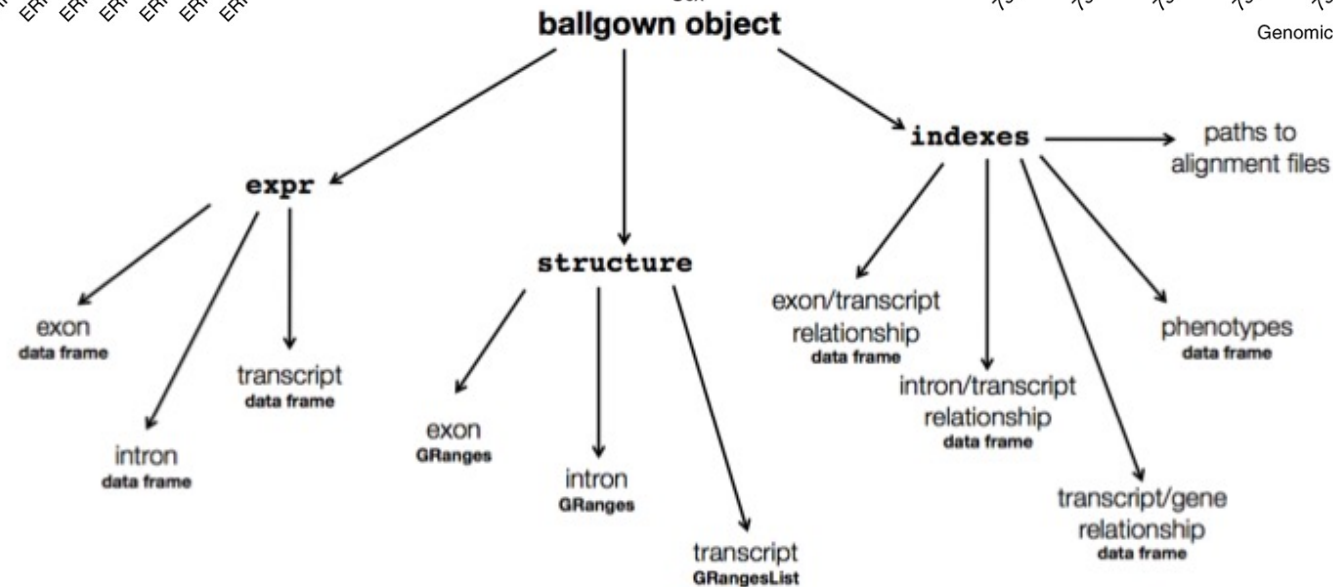
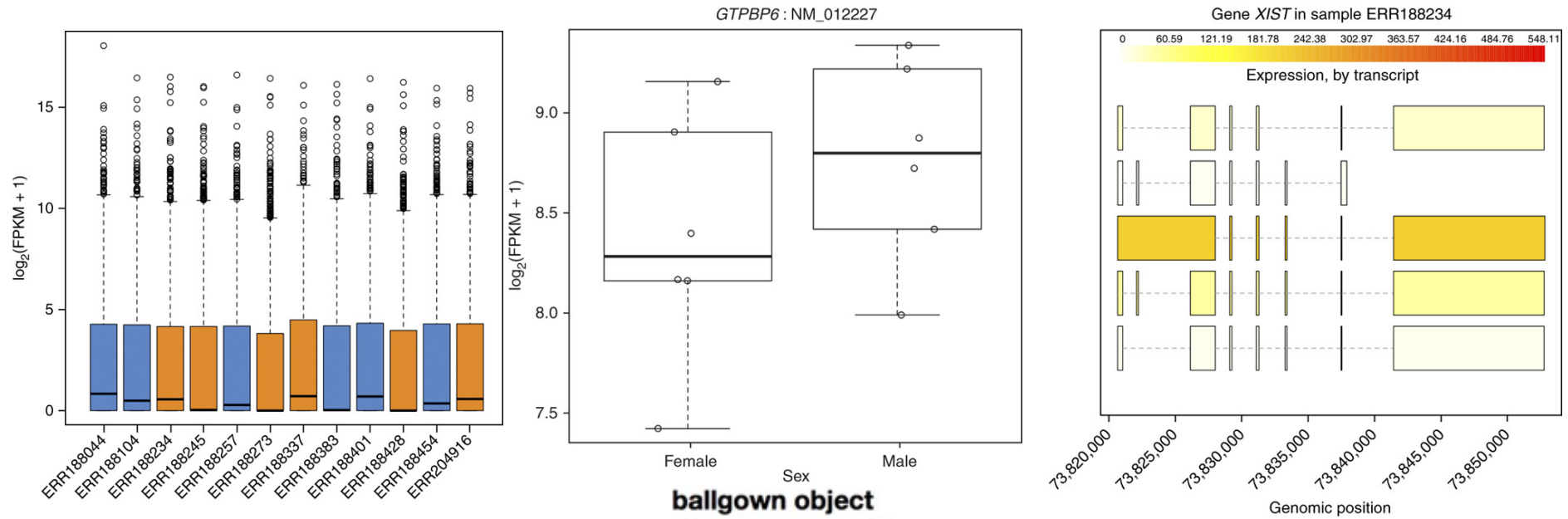
Differential Expression with Ballgown

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.
- 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



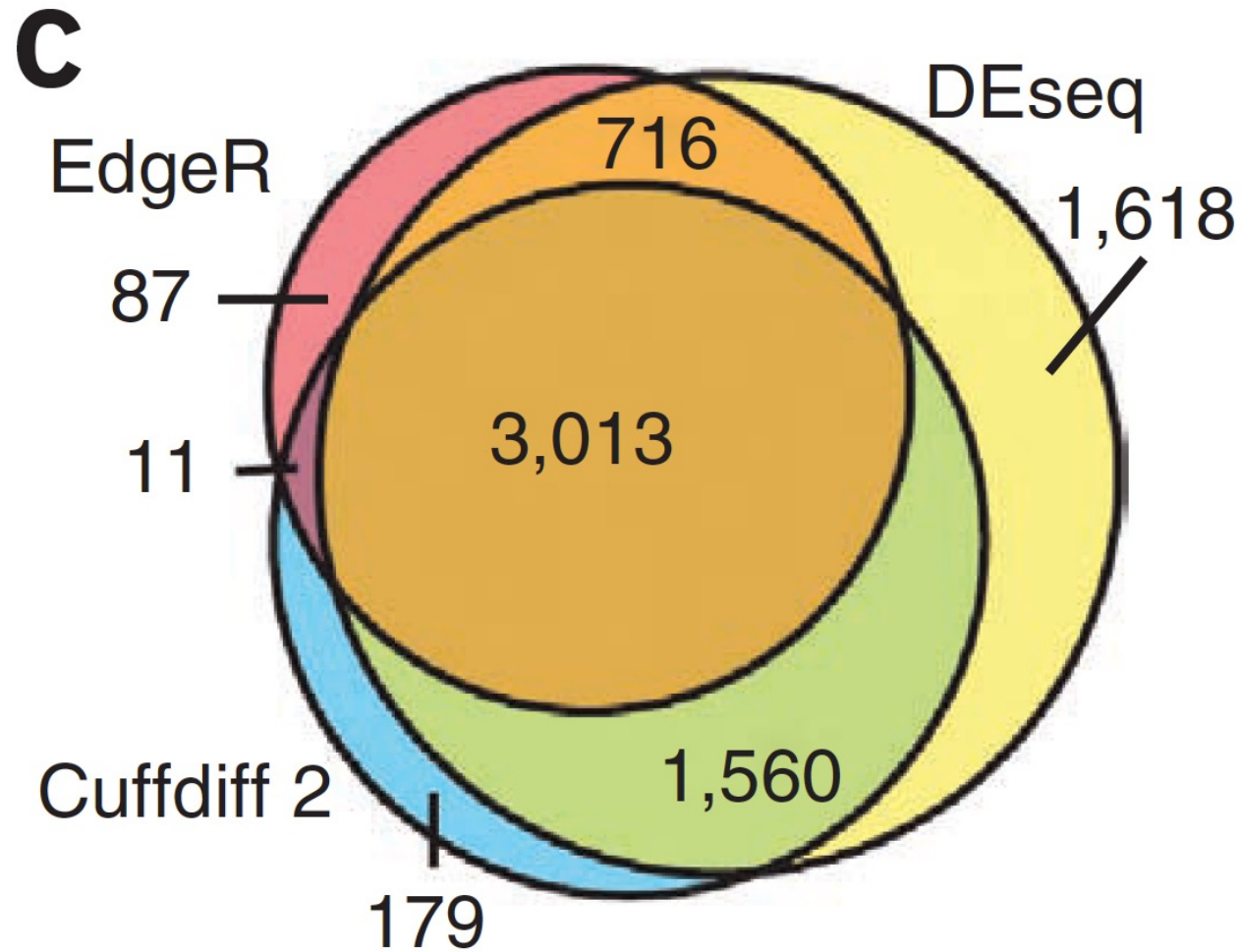
Alternative differential expression methods

- 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://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

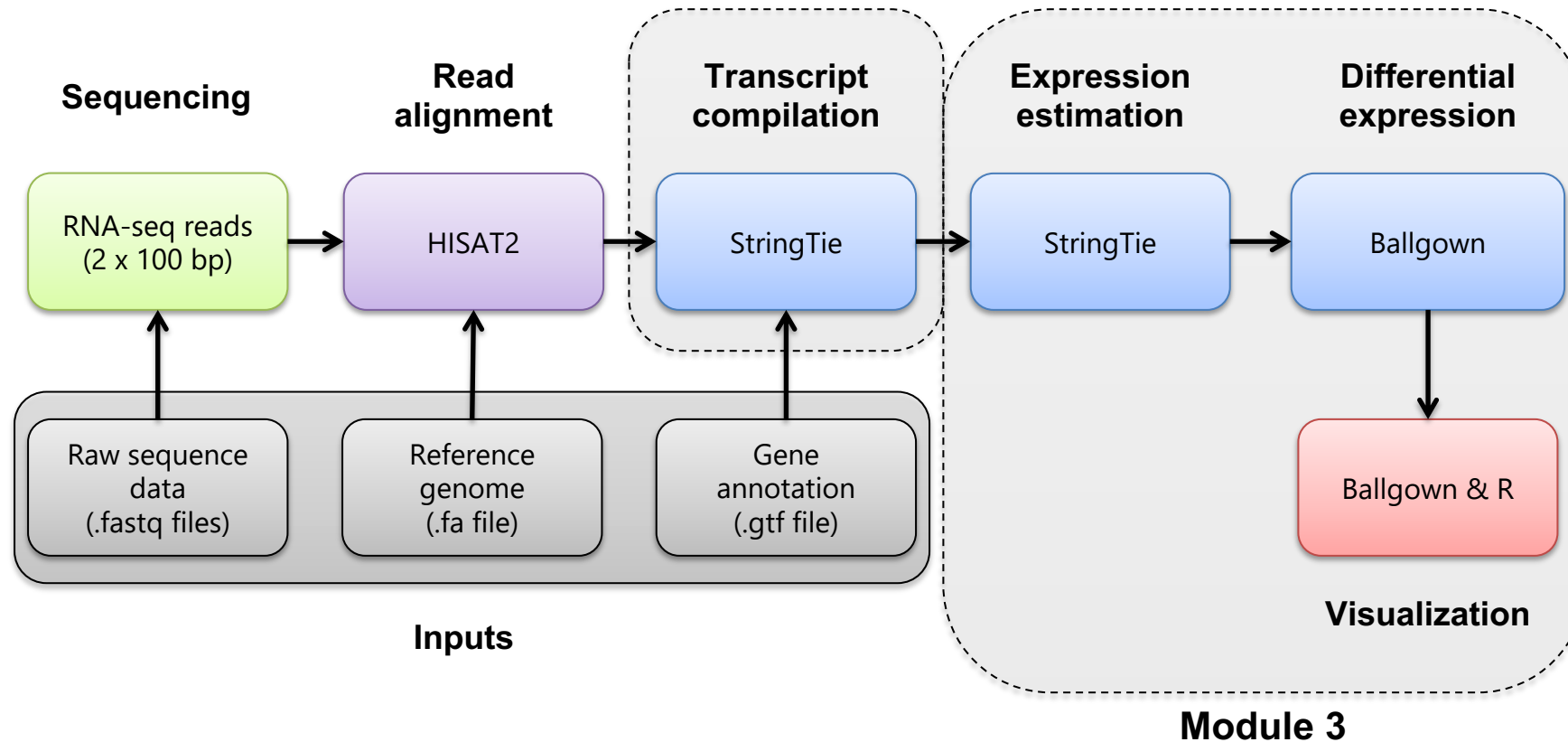
- As more attributes are compared, differences due solely to chance become more likely!
- Well known from array studies
 - 10,000s genes/transcripts
 - 100,000s exons
- With RNA-seq, more of a problem than ever
 - All the complexity of the transcriptome gives huge numbers of potential features
 - Genes, 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



We are on a Coffee Break & Networking Session

Workshop Sponsors:



Canadian Centre for
Computational
Genomics



HPC4Health

