



Cold  
Spring  
Harbor  
Laboratory

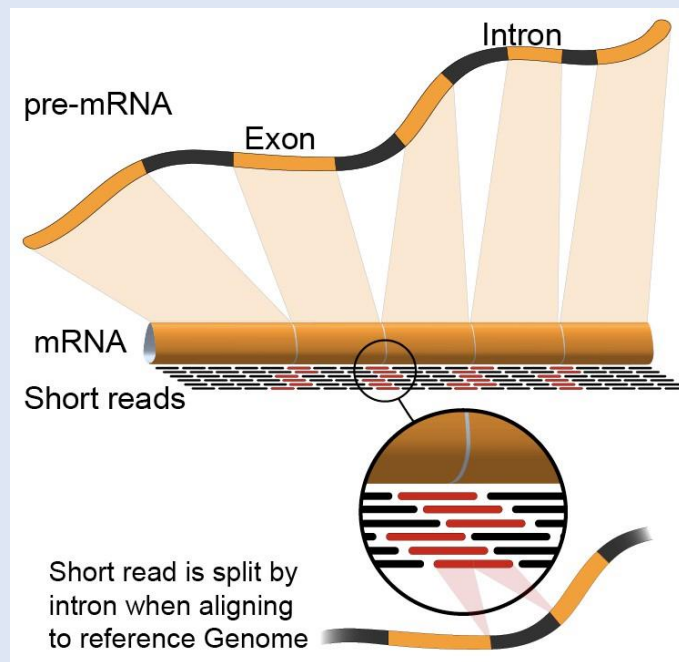
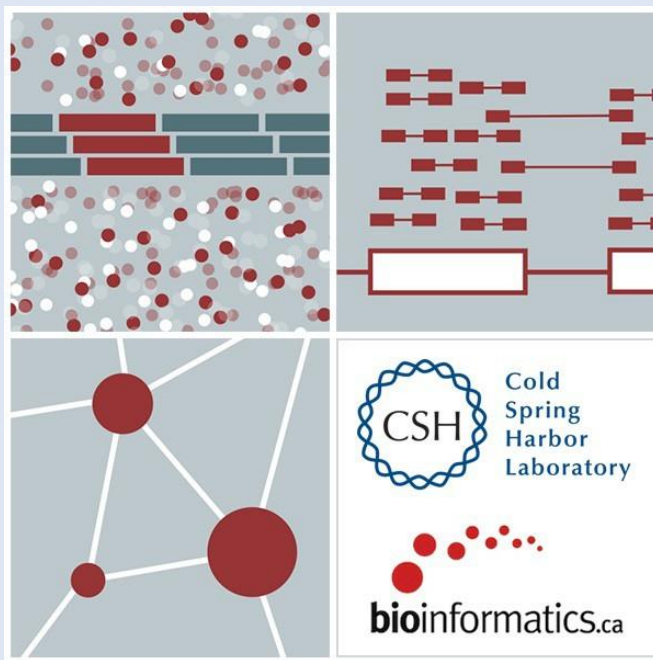
# RNA-Seq Module 3

## Differential Expression

Mariam Khanfar

Bfx workshop\_Nov 4, 2024

Slides adapted from CSHL SeqTec course RNA-seq lecture series by Obi & Malachi Griffith



 Washington University in St. Louis  
SCHOOL OF MEDICINE

# Learning Objectives of Module 3

- Expression estimation for known genes and transcripts
- FPKM/TPM expression normalized vs. raw counts
- Differential expression methods

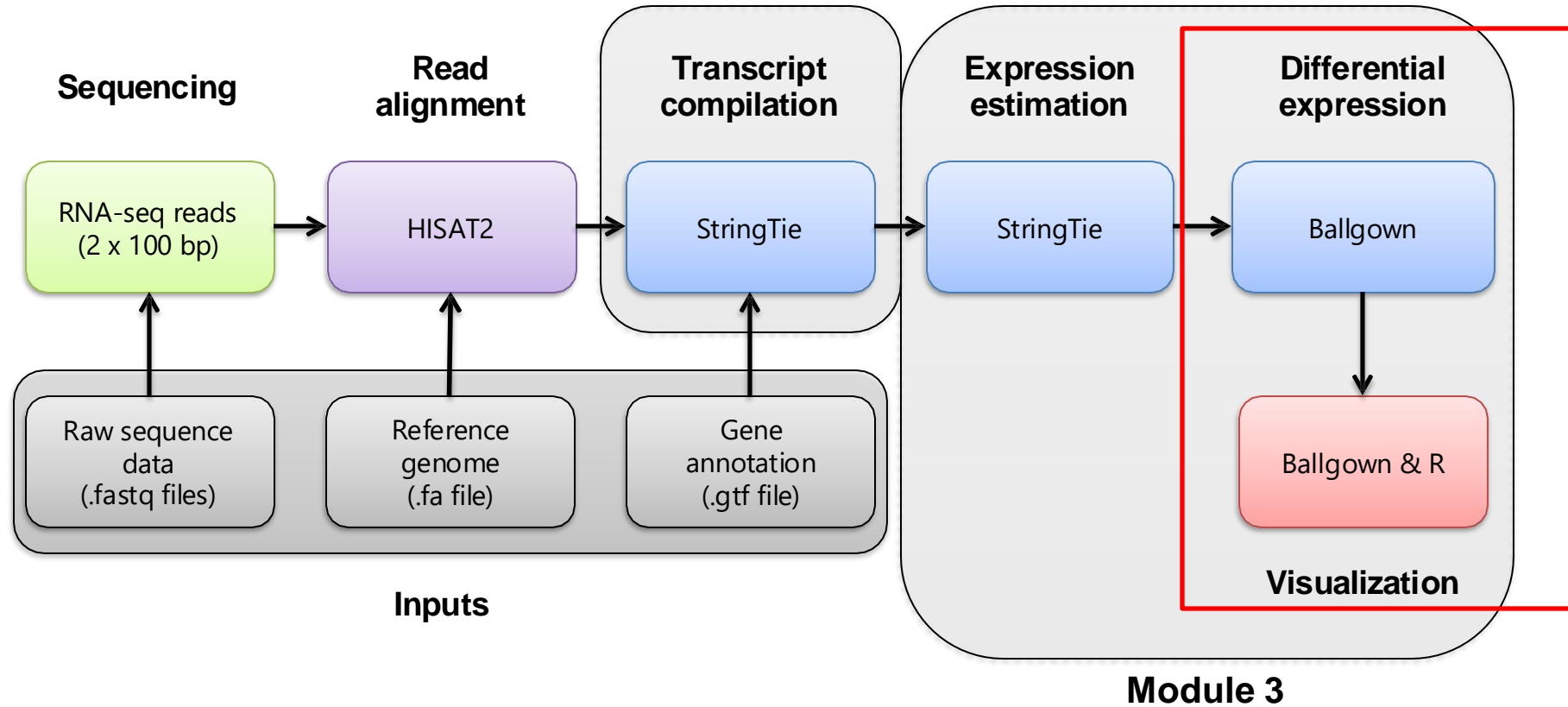
# Learning Objectives of Module 3

- ~~Expression estimation for known genes and transcripts~~
- ~~FPKM/TPM expression normalized vs. raw counts~~
- Differential expression methods → Ballgown (Stringtie)

# To-do

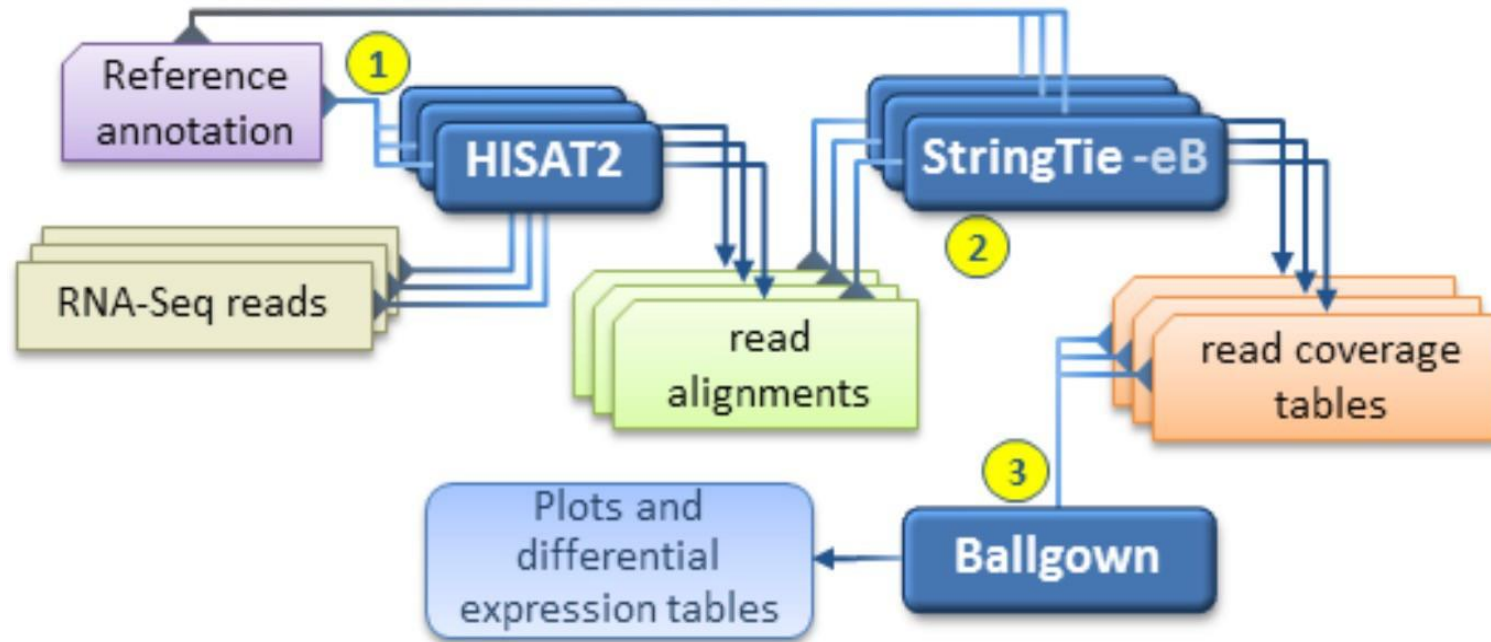
- Open docker desktop app in the background. Then in terminal, type:  
`$ docker pull griffithlab/rnabio:0.0.1`
- Make sure to:
  1. Switch User `su` to the `ubuntu` user → `su ubuntu`
  2. `source ~/.bashrc`
  3. Set the environment variable →  
`export RNA_HOME=~/.workspace/rnaseq`
- Prepare input data (if you are stuck, download from:  
<http://genomedata.org/rnaseq-tutorial/results/cshl2022/rnaseq.tar.gz>)

# HISAT2/StringTie/Balgon RNA-seq Pipeline



Last week: Expression estimation (StringTie, htseq count)  
This week: Differential expression (Ballgown, edgeR)

# Stringtie\_Expression Estimate



This is the workflow we used in last week's exercise:  
StringTie -G and -e

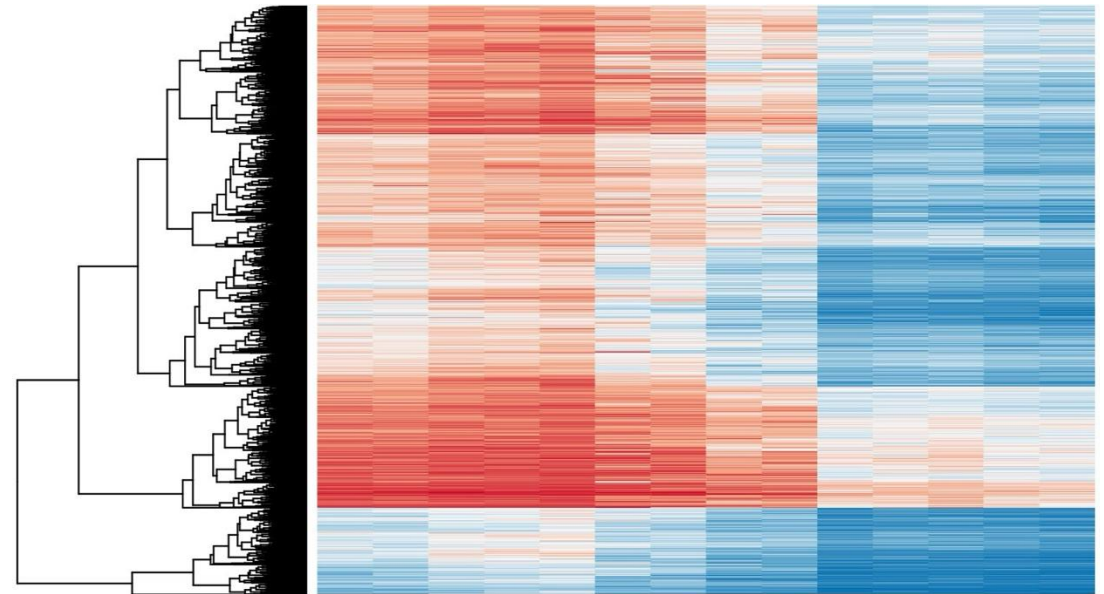
### Expression estimation mode (“Reference Only”)

# Summary

- Normalized counts account for sequencing depth and gene length biases
  - RPKM ~ single-end sequencing, FPKM ~ paired-end sequencing
  - The sum of all TPMs in each sample is the same. Easier to compare across samples!
- Abundance estimation tool that calculates normalized count (FPKM, TPM): StringTie
- Abundance estimation tool that calculates raw count: HTseq

# 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





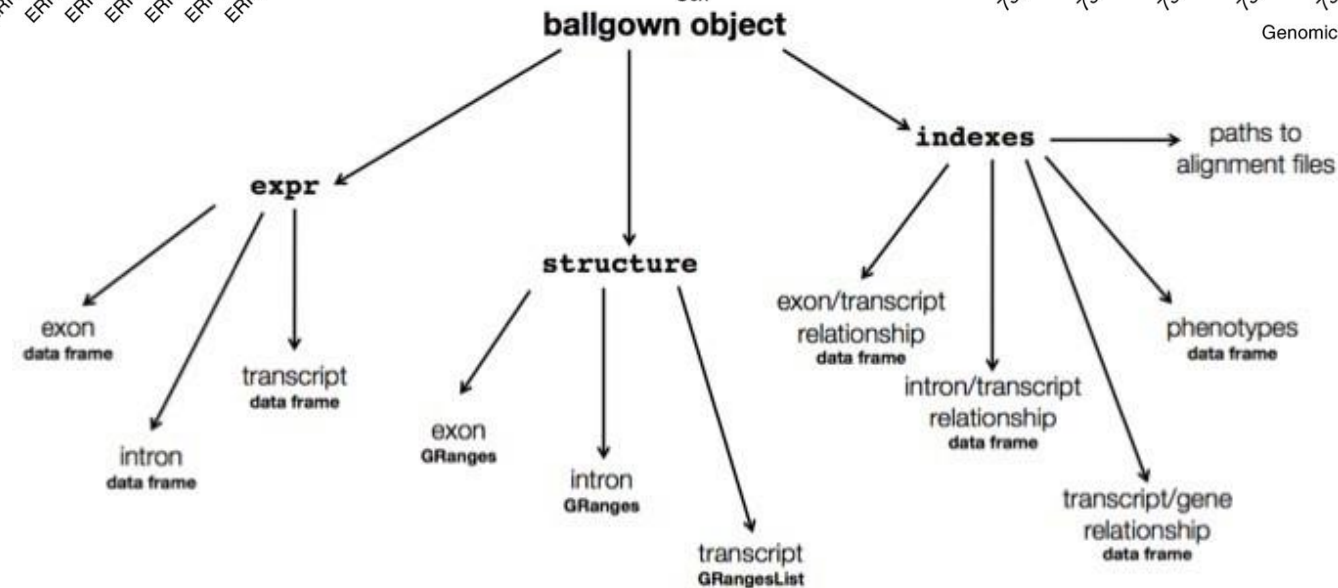
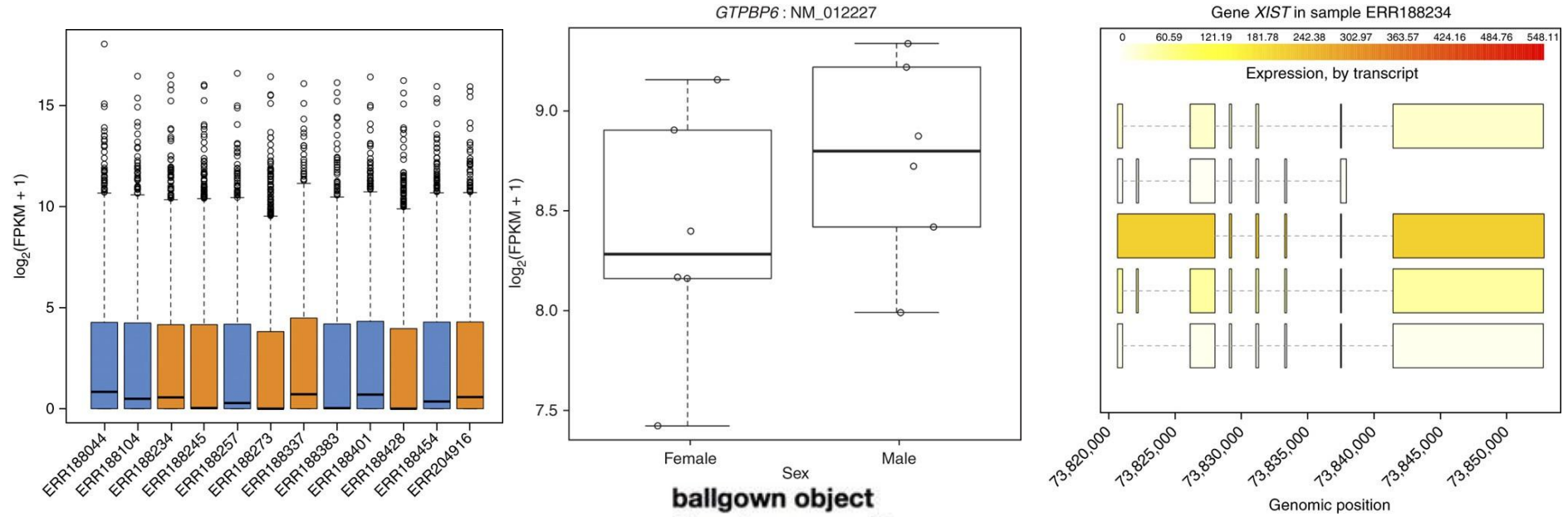
# Differential Expression with Ballgown

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.
- 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



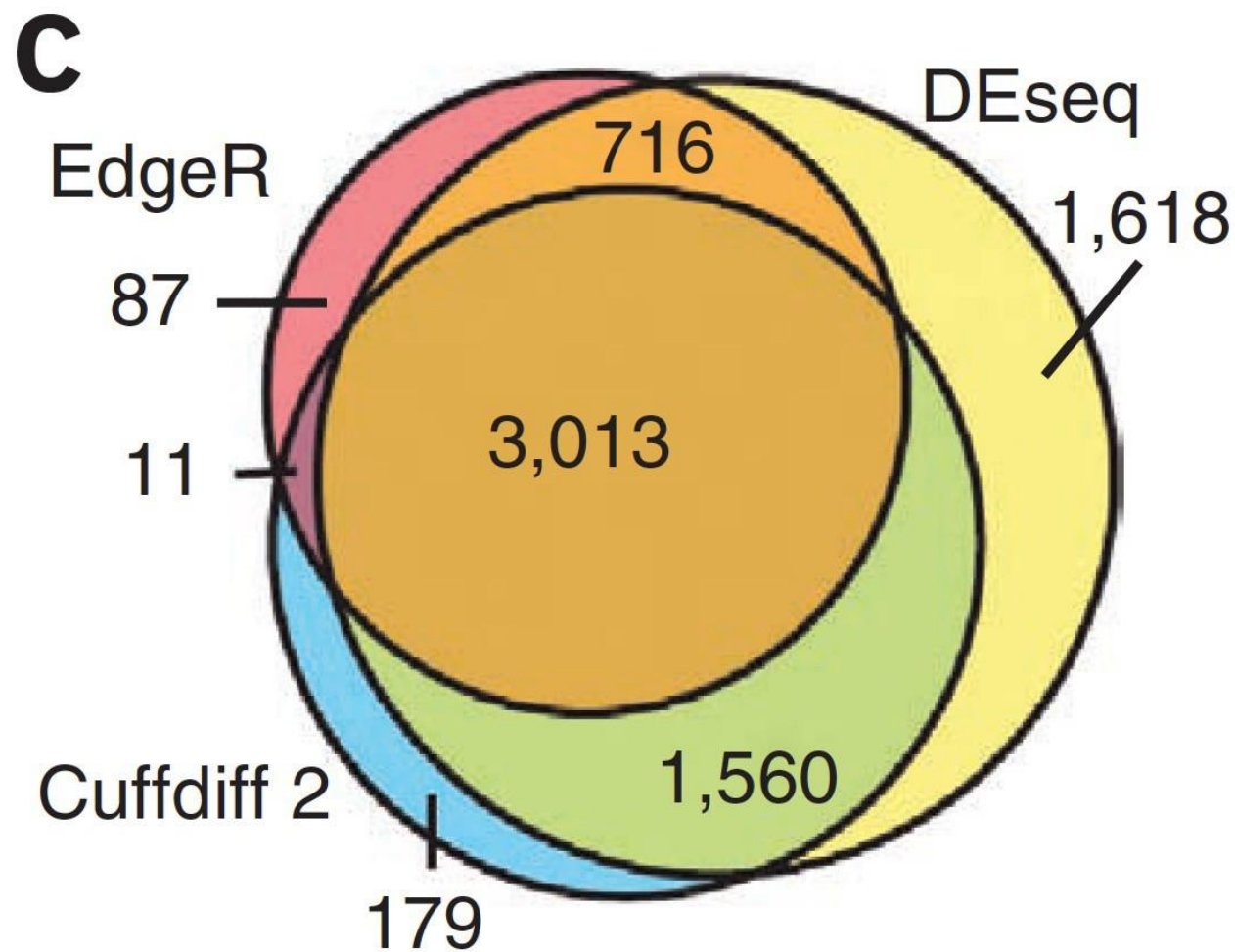
# Alternative differential expression methods

- 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
    - Stringtie/Ballgown approach is also robust
  - 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>

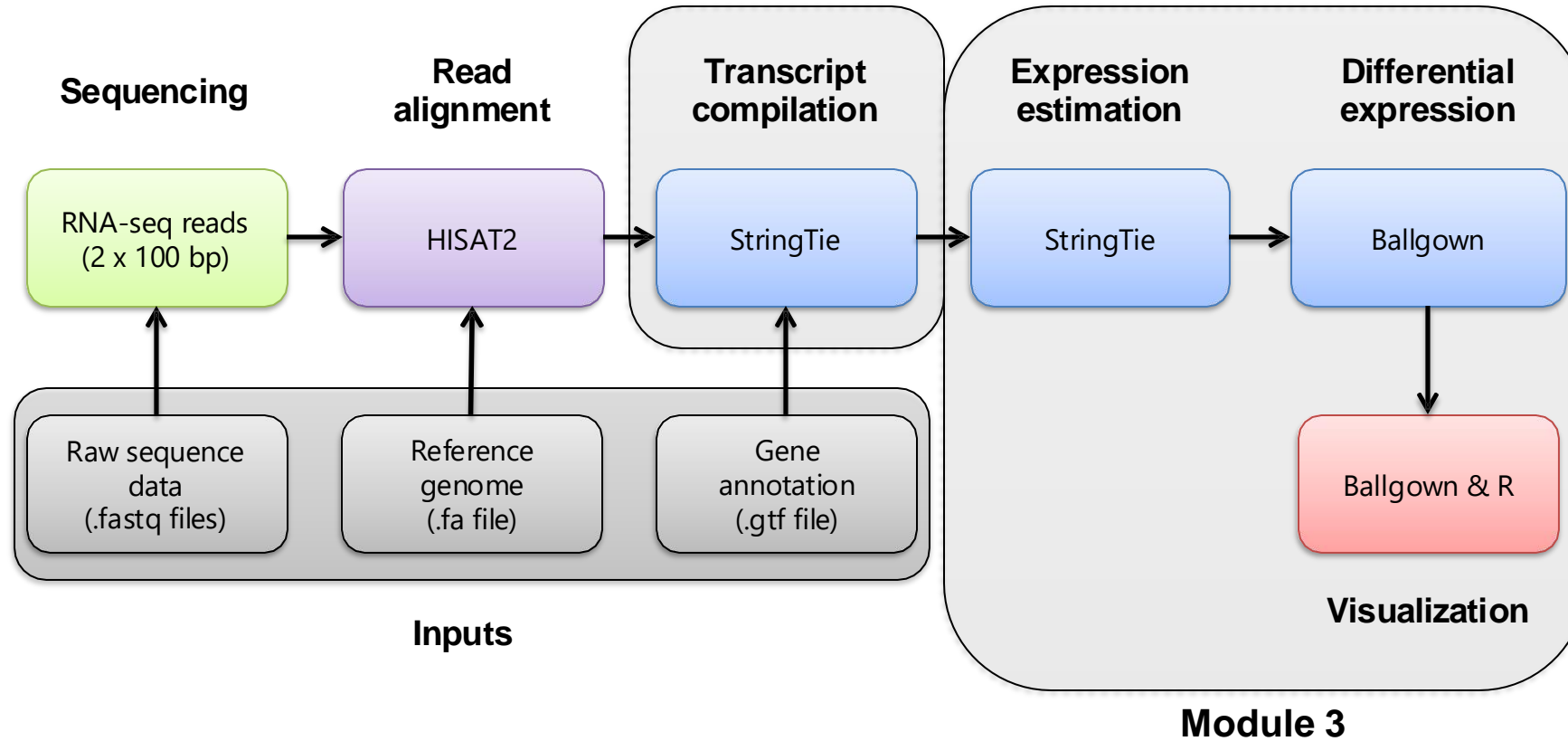
[https://genviz.org/module-04-expression/0004/01/01/Expression\\_Profiling\\_and\\_Visualization/](https://genviz.org/module-04-expression/0004/01/01/Expression_Profiling_and_Visualization/)



# RNA Input Data








- Universal Human Reference (UHR) and Human Brain Reference (HBR)
- In addition, a spike-in control was used (ERCC ExFold RNA Spike-In Control Mixes) to each sample.
  - UHR + ERCC Spike-In Mix1, Replicate 1
  - UHR + ERCC Spike-In Mix1, Replicate 2
  - UHR + ERCC Spike-In Mix1, Replicate 3
  - HBR + ERCC Spike-In Mix2, Replicate 1
  - HBR + ERCC Spike-In Mix2, Replicate 2
  - HBR + ERCC Spike-In Mix2, Replicate 3

# HISAT2/StringTie/Balgon RNA-seq Pipeline



# Stringtie outputs

- Stringtie gives 3 metrics for expression levels: coverage, FPKM, TPM ; for 2 types : transcript and gene.
- Focus on the 'transcript.gtf' and 'gene\_abundance.tsv'

 e_data.ctab	CTAB File	<b>e_data.ctab:</b> Contains information about exon-level expression
 e2t.ctab	CTAB File	<b>i_data.ctab:</b> Contains information about intron-level expression (used less frequently in expression analysis)
 gene_abundances	TSV File	<b>t_data.ctab:</b> Contains information about transcript-level expression
 i_data.ctab	CTAB File	<b>e2t.ctab, i2t.ctab:</b> These 'edge' files map relationships between transcripts and exons/introns
 i2t.ctab	CTAB File	<b>gene_abundances.tsv:</b> A tab-separated file containing gene-level expression estimates, typically including FPKM values
 t_data.ctab	CTAB File	<b>transcripts.gtf:</b> The assembled transcripts file in GTF format that includes the estimated transcript structures and their expression levels
 transcripts.gtf	GTF File	

# Assignment: Ballgown DE Analysis

- Run R and load required libraries

```
ubuntu@c6c4b48da477:~/workspace/rnaseq/de/ballgown/ref_only$ pwd
/home/ubuntu/workspace/rnaseq/de/ballgown/ref_only
ubuntu@c6c4b48da477:~/workspace/rnaseq/de/ballgown/ref_only$ ls
ubuntu@c6c4b48da477:~/workspace/rnaseq/de/ballgown/ref_only$ R

R version 4.0.0 (2020-04-24) -- "Arbor Day"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(ballgown)

Attaching package: 'ballgown'

The following object is masked from 'package:base':

    structure

> library(genefilter)
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:ballgown':

    contains, expr, last

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> library(devtools)
Loading required package: usethis
> |
```

# Create phenotype data needed for ballgown analysis

```
> getwd()
[1] "/workspace/rnaseq/de/ballgown/ref_only"
> results="/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/"
> results
[1] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/"
> path=paste(results,ids,sep="")
> pheno_data=data.frame(ids,type,path)
> path
[1] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep1"
[2] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep2"
[3] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep3"
[4] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep1"
[5] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep2"
[6] "/home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep3"
> pheno_data
  ids type
1 UHR_Rep1 UHR
2 UHR_Rep2 UHR
3 UHR_Rep3 UHR
4 HBR_Rep1 HBR
5 HBR_Rep2 HBR
6 HBR_Rep3 HBR

                                path
1 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep1
2 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep2
3 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/UHR_Rep3
4 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep1
5 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep2
6 /home/ubuntu/workspace/rnaseq/expression/stringtie/ref_only/HBR_Rep3
```

# Create the ballgown data structure

```
> bg = ballgown(samples = as.vector(pheno_data$path), pData = pheno_data)
Mon Dec 11 03:12:48 2023
Mon Dec 11 03:12:48 2023: Reading linking tables
Mon Dec 11 03:12:48 2023: Reading intron data files
Mon Dec 11 03:12:49 2023: Merging intron data
Mon Dec 11 03:12:49 2023: Reading exon data files
Mon Dec 11 03:12:51 2023: Merging exon data
Mon Dec 11 03:12:51 2023: Reading transcript data files
Mon Dec 11 03:12:52 2023: Merging transcript data
Wrapping up the results
```

```
> bg
ballgown instance with 4564 transcripts and 6 samples
```

Verify your object

# Attributes

- Extract all transcript-level expression data from the bg object. Then extract unique gene and unique transcript IDs.

```
> head(bg_table)
  t_id chr strand   start     end      t_name num_exons length
1    1  22     - 10736171 10736283 ENST000000615943         1    113
2    2  22     - 10939388 10961338 ENST000000635667         9    749
3    3  22     - 11065974 11067346 ENST000000623473         2     54
4    4  22     + 11066501 11068089 ENST000000624155         2    120
5    5  22     + 11124337 11125705 ENST000000422332         2   1241
6    6  22     - 11249809 11249959 ENST000000612732         1    151

  gene_id gene_name cov.UHR_Rep1 FPKM.UHR_Rep1 cov.UHR_Rep2
1 ENSG00000277248      U2              0              0              0
2 ENSG00000283047   FRG1FP              0              0              0
3 ENSG00000280363 CU104787.1              0              0              0
4 ENSG00000279973    BAGE5              0              0              0
5 ENSG00000226444  ACTR3BP6              0              0              0
6 ENSG00000276871   5_8S_rRNA              0              0              0

  FPKM.UHR_Rep2 cov.UHR_Rep3 FPKM.UHR_Rep3 cov.HBR_Rep1 FPKM.HBR_Rep1
1              0              0              0              0              0
2              0              0              0              0              0
3              0              0              0              0              0
4              0              0              0              0              0
5              0              0              0              0              0
6              0              0              0              0              0

  cov.HBR_Rep2 FPKM.HBR_Rep2 cov.HBR_Rep3 FPKM.HBR_Rep3
1              0              0              0              0
2              0              0              0              0
3              0              0              0              0
4              0              0              0              0
5              0              0              0              0
6              0              0              0              0
```

```
> head(bg_gene_names)
  gene_id gene_name
1 ENSG00000277248      U2
2 ENSG00000283047   FRG1FP
3 ENSG00000280363 CU104787.1
4 ENSG00000279973    BAGE5
5 ENSG00000226444  ACTR3BP6
6 ENSG00000276871   5_8S_rRNA
```

```
> head(bg_transcript_names)
  t_id      t_name
1    1 ENST000000615943
2    2 ENST000000635667
3    3 ENST000000623473
4    4 ENST000000624155
5    5 ENST000000422332
6    6 ENST000000612732
```

# Perform DE analysis without filtering

```
results_transcripts = stattest(bg, feature="transcript", covariate="type",  
getFC=TRUE, meas="FPKM")
```

```
results_transcripts = merge(results_transcripts, bg_transcript_names, by.x=c("id"),  
by.y=c("t_id"))
```

```
> head(results_transcripts)
  id feature      fc      pval      qval      t_name
1  1 transcript 1.0000000      NaN      NaN ENST00000615943
2 10 transcript 1.0000000      NaN      NaN ENST00000448473
3 100 transcript 1.0000000      NaN      NaN ENST00000517943
4 1000 transcript 1.0000000      NaN      NaN ENST00000403807
5 1001 transcript 1.0000000      NaN      NaN ENST00000302273
6 1002 transcript 0.8876485 0.8829198 0.955367 ENST00000624350
> head(results_genes)
      id feature      fc      pval      qval gene_name
1 ENSG00000008735    gene 0.01383563 0.0002270410 0.003574835 MAPK8IP2
2 ENSG00000015475    gene 1.58883098 0.0054844568 0.026638790      BID
3 ENSG00000025708    gene 1.39579593 0.3992876858 0.596233639      TYMP
4 ENSG00000025770    gene 1.46572045 0.0316457273 0.103699543    NCAPH2
5 ENSG00000040608    gene 0.10280538 0.0004183902 0.005360246      RTN4R
6 ENSG00000054611    gene 1.09845192 0.1808088853 0.352059592    TBC1D22A
```



# Filter low-abundance genes

`subset()`: This function subsets the `bg` object to include only those transcripts with a variance across the samples greater than 1.

This step is designed to remove transcripts that do not show much change across your conditions, under the assumption that they are not likely to be biologically interesting

`bg_filt = subset (bg, "rowVars(expr(bg)) > 1", genomesubset=TRUE)`

where `genomesubset=TRUE` → ensures that when you subset the transcripts, the associated genomic features (like exons and introns) are also appropriately subsetted.

```
> nrow(bg_table)
[1] 4564
> nrow(bg_filt_table)
[1] 2924
> nrow(bg_gene_names)
[1] 1410
> nrow(bg_filt_gene_names)
[1] 830
> nrow(bg_transcript_names)
[1] 4564
> nrow(bg_filt_transcript_names)
[1] 2924
```

# Perform DE analysis using the filtered data and identify significant genes

```
results_transcripts = statstest(bg_filt, feature="transcript", covariate="type", getFC=TRUE, meas="FPKM") <- for the filtered one...
```

```
sig_transcripts = subset(results_transcripts, results_transcripts$pval<0.05)
```

```
sig_genes = subset(results_genes, results_genes$pval<0.05)
```

```
> head(sig_transcripts)
```

	id	feature	fc	pval	qval	t_name
13	1035	transcript	32.66479	0.0006079048	0.03192359	ENST00000302097
14	1036	transcript	517.14266	0.0134847913	0.18801647	ENST00000398743
16	1038	transcript	1314.56328	0.0060651295	0.11368230	ENST00000398741
17	1039	transcript	19.19580	0.0227337966	0.24896487	ENST00000543184
18	1040	transcript	930.65593	0.0175994799	0.21084735	ENST00000405655
19	1041	transcript	60.37137	0.0115773205	0.17271472	ENST00000402697

```
> head(sig_genes)
```

	id	feature	fc	pval	qval	gene_name
1	ENSG00000008735	gene	0.01411366	0.0001930115	0.003076761	MAPK8IP2
2	ENSG00000015475	gene	1.57305207	0.0056168528	0.025899932	BID
4	ENSG00000025770	gene	1.47840611	0.0248558798	0.079347616	NCAPH2
5	ENSG00000040608	gene	0.10382267	0.0004159533	0.004810006	RTN4R
8	ENSG00000069998	gene	2.57657413	0.0068230552	0.029342673	CECR5
9	ENSG00000070010	gene	2.18390717	0.0009926846	0.007772907	UFD1L

# Expected output in

*\$RNA\_HOME/de/ballgown/ref\_only/*

```
ubuntu@c6c4b48da477:~/workspace/rnaseq/de/ballgown/ref_only$ pwd
/home/ubuntu/workspace/rnaseq/de/ballgown/ref_only
ubuntu@c6c4b48da477:~/workspace/rnaseq/de/ballgown/ref_only$ ls -alt
total 2868
-rw-rw-r-- 1 ubuntu ubuntu 26648 Dec 11 04:00 UHR_vs_HBR_gene_results_sig.tsv
drwxrwxr-x 1 ubuntu ubuntu 4096 Dec 11 04:00 .
-rw-rw-r-- 1 ubuntu ubuntu 38495 Dec 11 04:00 UHR_vs_HBR_transcript_results_sig.tsv
-rw-rw-r-- 1 ubuntu ubuntu 69678 Dec 11 04:00 UHR_vs_HBR_gene_results_filtered.tsv
-rw-rw-r-- 1 ubuntu ubuntu 249590 Dec 11 03:59 UHR_vs_HBR_transcript_results_filtered.tsv
-rw-rw-r-- 1 ubuntu ubuntu 95058 Dec 11 03:48 UHR_vs_HBR_gene_results.tsv
-rw-rw-r-- 1 ubuntu ubuntu 325014 Dec 11 03:48 UHR_vs_HBR_transcript_results.tsv
-rw-rw-r-- 1 ubuntu ubuntu 2114225 Dec 11 03:42 bg.rda
drwxrwxr-x 1 ubuntu ubuntu 4096 Dec 11 03:02 ..
```

# Purpose of ERCC in RNAs and RNA-Seq Analysis

**.Normalization:** control for variations in RNA input, reverse transcription efficiency, PCR amplification biases, and sequencing depth

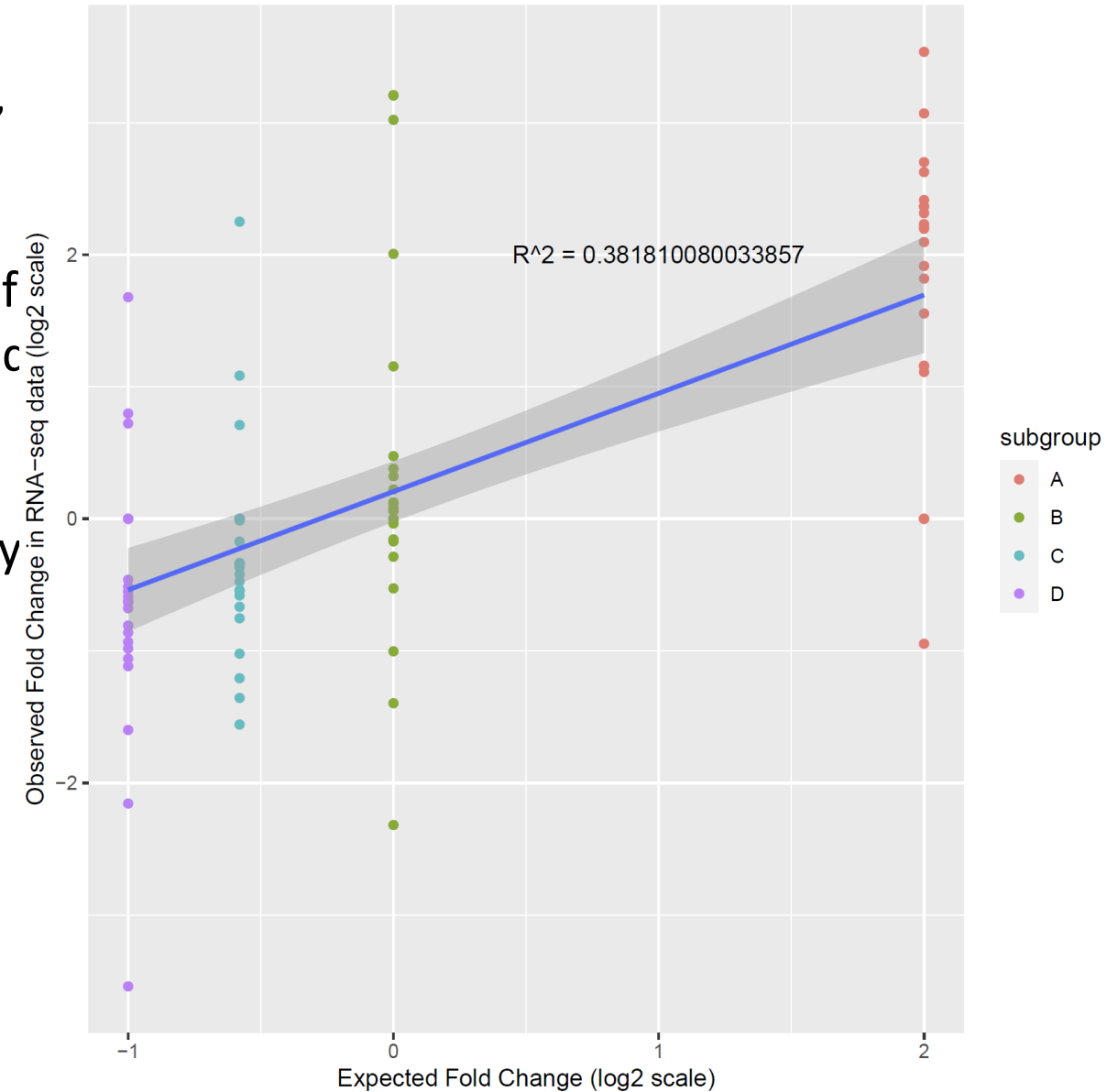
**.Validation:** provide a way to assess the accuracy of RNA-Seq measurements by comparing the observed fold changes to the known, expected fold changes

**.Quality Control:** used to check the overall performance of the RNA-Seq workflow, from library preparation to sequencing and data analysis

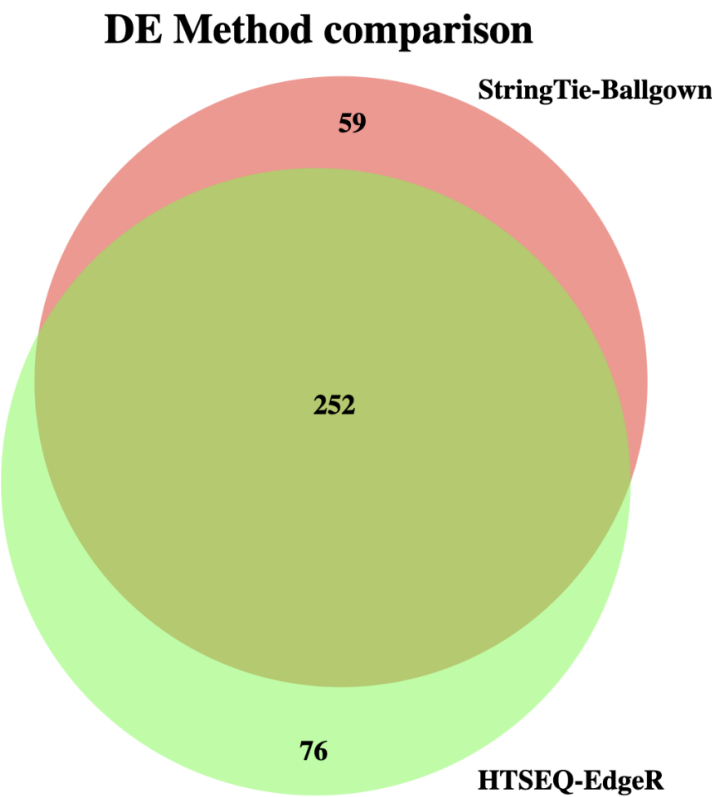
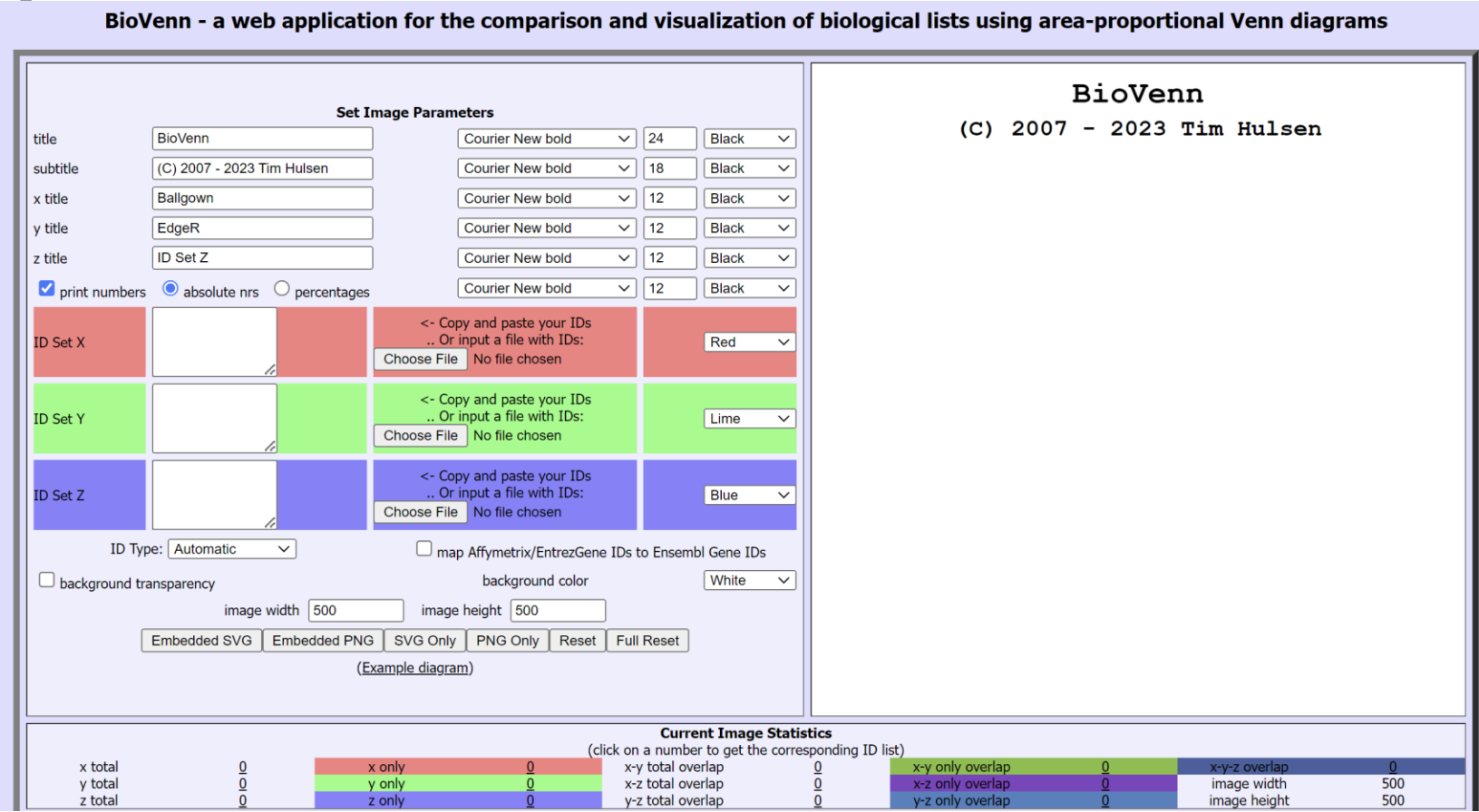
Table 1 Transcript molar ratios in ERCC Spike-In Mixes

Subgroup	Mix 1:Mix 2 <sup>†</sup>
A	4.00
B	1.00
C	0.67
D	0.50

<sup>†</sup> Applies only to Spike-In Mix 1 and Mix 2 with same manufacturing lot number.



# DE Method Comparison ( Venn Diagram (DE genes from StringTie/Ballgown vs HTSeq/EdgeR)



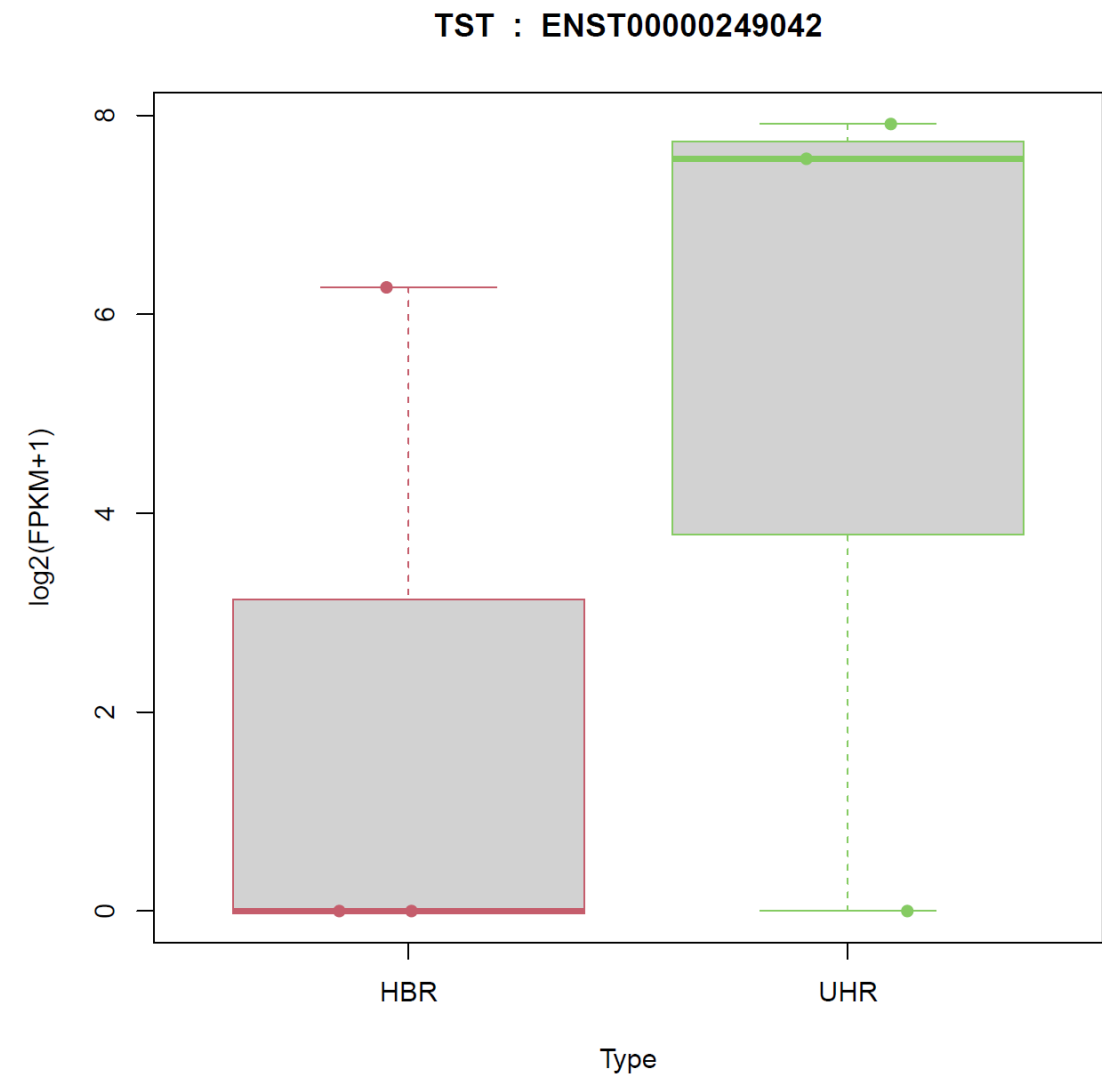
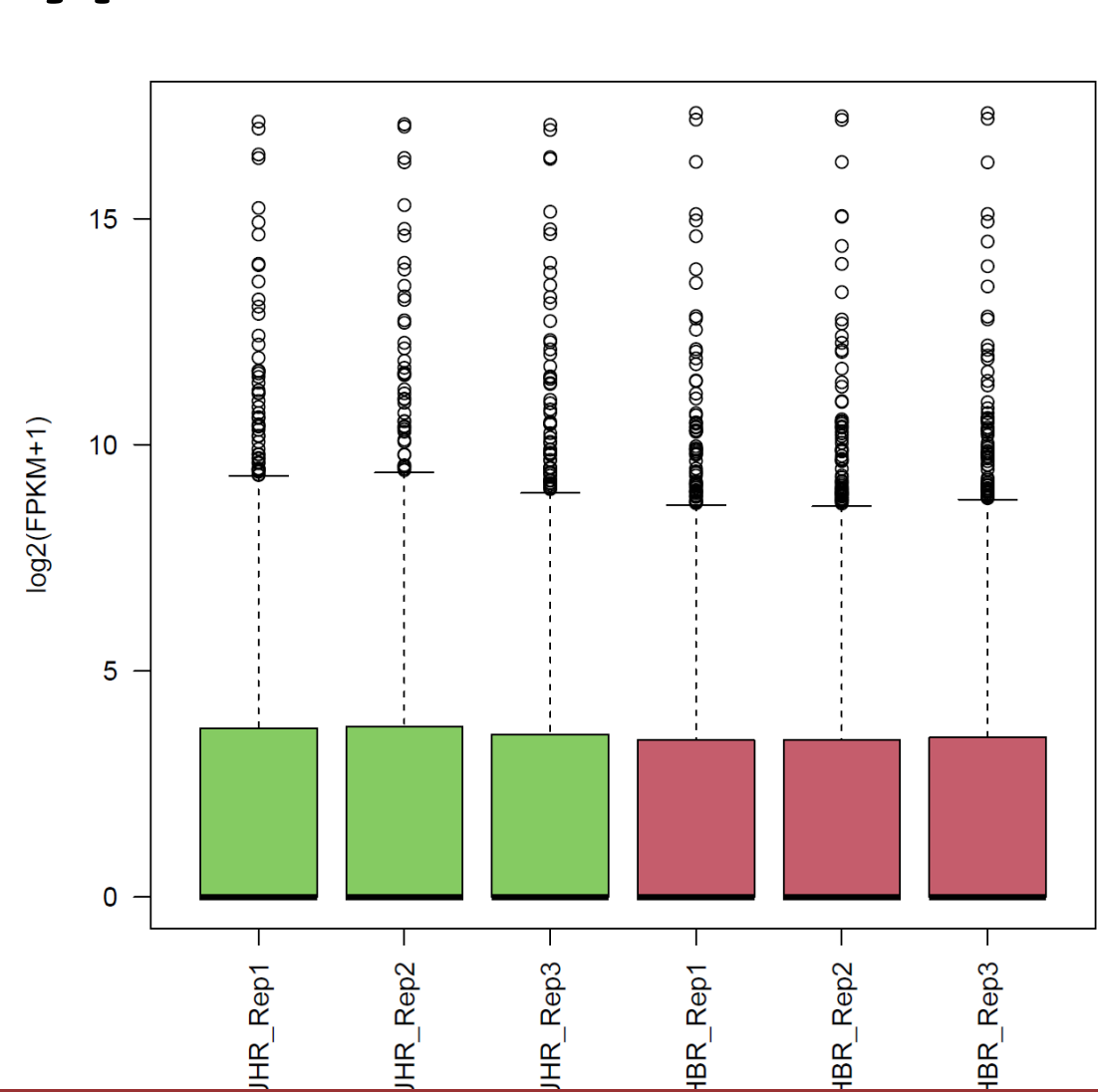
# DE Visualization

[https://rnabio.org/module-03-expression/0003/04/01/DE\\_Visualization\\_Ballgown/](https://rnabio.org/module-03-expression/0003/04/01/DE_Visualization_Ballgown/)

# DE\_Visualization

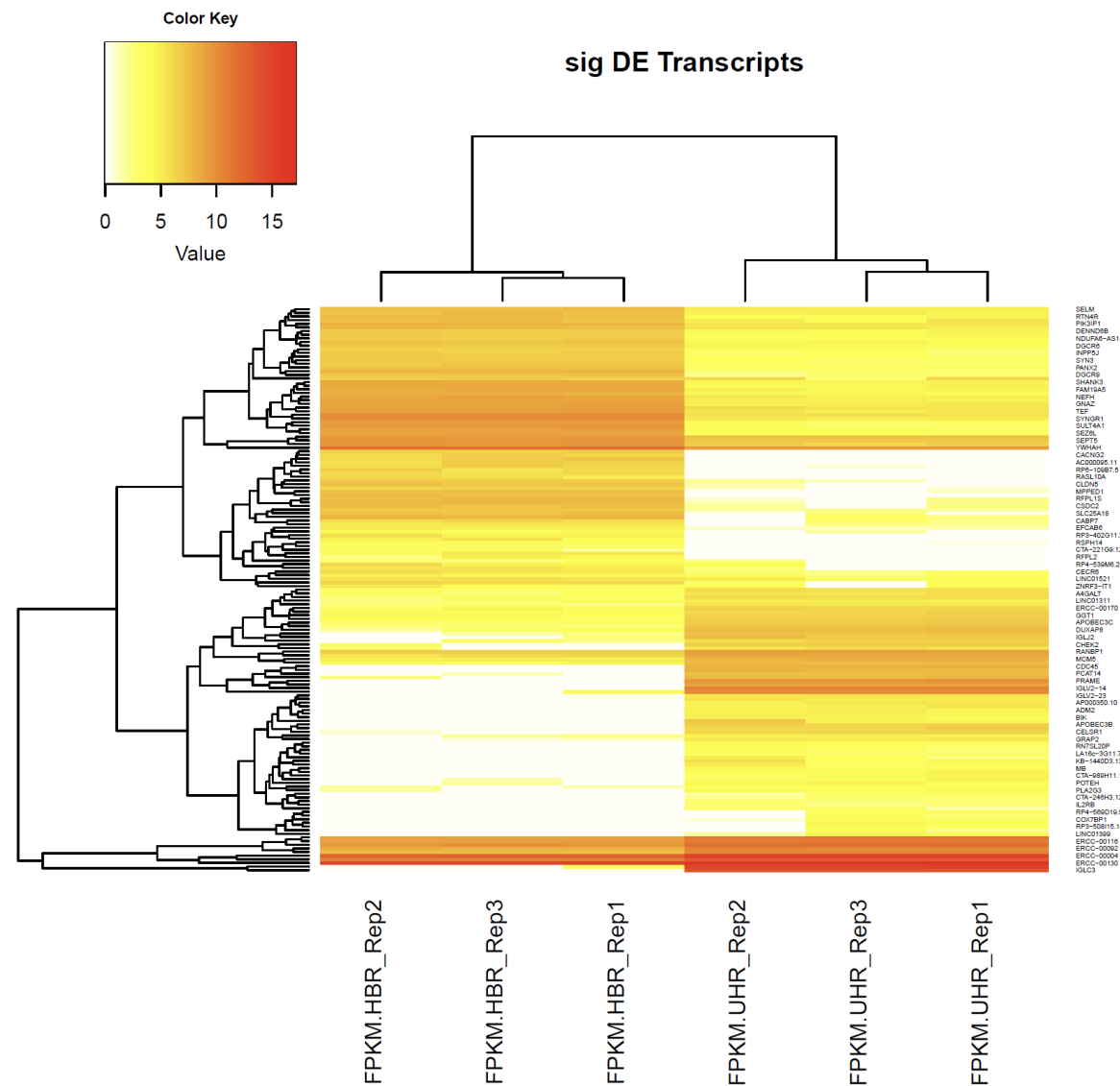
- Phenotype data is loaded along with the Ballgown object (bg.rda), which contains the results of your differential expression analysis
- FPKM values are extracted, log-transformed, and prepared for plotting  
`fpkm = log2(fpkm+1)`

# FPKM values for each sample/ across different types





# DE\_Heatmap



# DE\_Volcano Plot

UHR vs HBR



