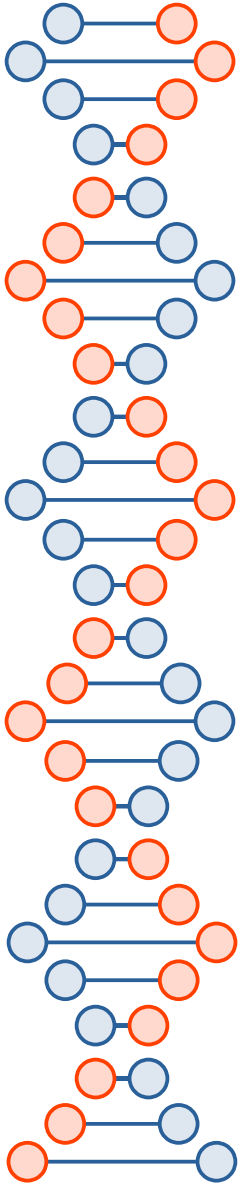


Tutorial

Analyzing Differentially Expressed Genes and Differential Binding Sites

Presented by: Eric Arezza

Prerequisite

- Familiarity with processed files (.bam, .bed)
  - This tutorial proceeds from outputs of the previously presented ngs_processing_pipeline.py
  - Scripts found in lab Github *Downstream_Analysis/DifferentialGeneExpression/*
- Familiarity with programming in R

Preface

Main DEG analysis tools used here include:

- DESeq2

  "**D**ifferential **e**xpression analysis for **seq**uence count data"

  https://bioconductor.org/packages/release/bioc/html/DESeq2.html

- edgeR

  "**E**mpirical Analysis of **D**igital **G**ene **E**xpression Data in **R**"

  https://bioconductor.org/packages/release/bioc/html/edgeR.html

- DiffBind

  "**Diff**erential **Bind**ing Analysis of ChIP-Seq Peak Data"

  https://bioconductor.org/packages/release/bioc/html/DiffBind.html

  - Bundles DESeq2 and edgeR when performing analysis

Preface

Prior to running any “analyze_degs” scripts, some files need to be manually prepared – tables that define the comparisons to be made between samples

- Sample info (.csv) for RNA-Seq analysis
- DiffBind samplesheet (.csv) for peaks analysis

Learning Goals

DEG Analysis:

1) **Bulk RNA-Seq Analysis**:
   - Count matrix creation and sample sheet file preparation
   - Normalization and DESeq2 + edgeR
   - Computation of DEGs
   - Annotations and figures

2) DiffBind Peaks Analysis:
   - Sample sheet file preparation
   - Consensus peaksets
   - Occupancy analysis
   - Affinity analysis

Bulk RNA-Seq Analysis – Preface

- At least 2 replicates required for each sample
- Alignment files (.bam + .bai) **with** duplicates should be used for RNA-Seq
  - Retains natural transcript expression without bias
    - Shorter transcripts and highly expressed genes would be falsely reduced otherwise

RNA-Seq Analysis – Count Matrix Creation

A count matrix must first be generated from the alignment files

Common tools:

- htseq-count
- **featureCounts** (usable in R script, also a standalone program)
  - Found in *Rsubread* package, install first to use featureCounts https://bioconductor.org/packages/release/bioc/html/Rsubread.html

RNA-Seq Analysis – Count Matrix Creation

Using featureCounts in R:

1) Copy all **.bam and .bai** files (*Mapped.MAPQ10*) into a folder, **bams/**
2) Run the following lines:

Lookup input options for your data

```
bam_files <- list.files(path="bams/", full.names=TRUE)[c(TRUE, FALSE)]

bamcounts <- featureCounts(bam_files, annot.inbuilt="mm10", countMultiMappingReads=FALSE, ignoreDup=FALSE, isPairedEnd=TRUE, strandSpecific=0, nthreads=4, verbose=TRUE)

rownames(bamcounts$counts) <- mapIds(org.Mm.eg.db, keys=rownames(bamcounts$counts), column="SYMBOL", keytype="ENTREZID")

bamcounts$counts <- bamcounts$counts[!(is.na(rownames(bamcounts$counts))), ]

for (n in names(bamcounts)){

  write.table(bamcounts[[n]], file=paste(getwd(), "/", n, ".csv", sep=""), sep=",", quote=F, col.names=NA)

}
```

RNA-Seq Analysis – Count Matrix Creation

Alternatively, run the ***get_rnaseq_counts.R*** script

See input options and defaults for use.

- Custom featureCounts options will require modification/manually performing the code shown in the previous slide

RNA-Seq Analysis – Count Matrix

Genes (rows) x Samples (columns)

e.g. counts.csv

Modify column names as desired*

	WT_1	WT_2	WT_A	WT_B	KO_1	KO_2	KO_A	KO_B
Gene1	0	0	3	8	20	20	22	23
Gene2	10	11	12	5	10	15	16	13
Gene3	14	13	12	11	0	0	2	0
Gene4	0	2	2	1	30	30	37	23
Gene5	400	230	150	300	50	100	55	66

RNA-Seq Analysis – Sample Info File

Sample info file (.csv)

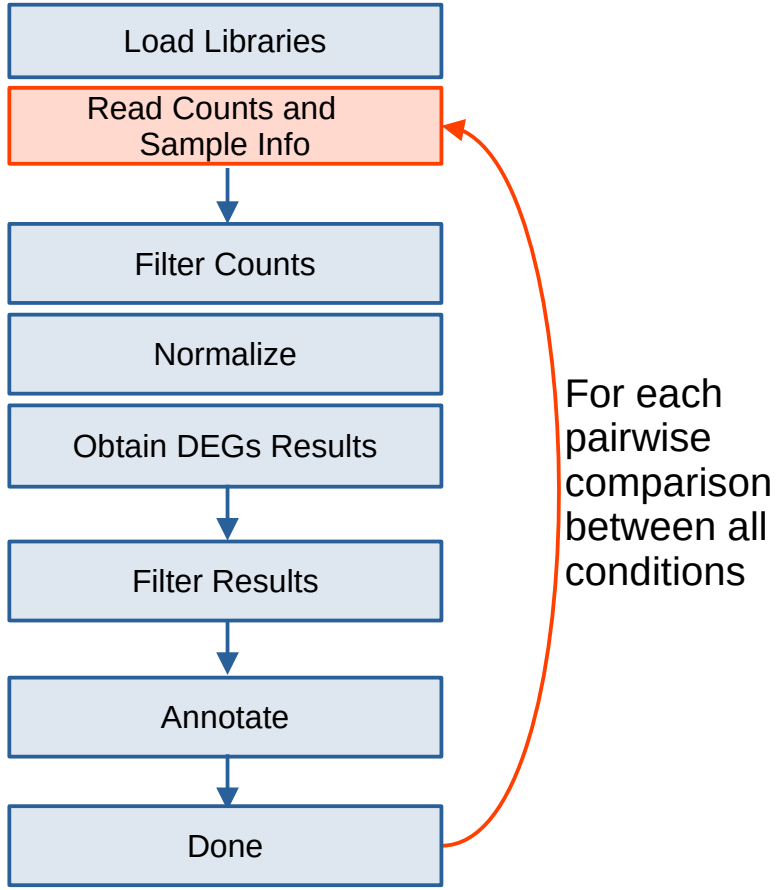
Sample	Condition	Group
WT_1	WT_1-2	1
WT_2	WT_1-2	1
WT_A	WT_A-B	2
WT_B	WT_A-B	2
KO_1	KO_1-2	3
KO_2	KO_1-2	3
KO_A	KO_A-B	4
KO_B	KO_A-B	4

Sample info names **MUST** match column names in counts matrix

Count matrix (.csv)

	WT_1	WT_2	WT_A	WT_B	KO_1	KO_2	KO_A	KO_B
Gene1	0	0	3	8	20	20	22	23
Gene2	10	11	12	5	10	15	16	13
Gene3	14	13	12	11	0	0	2	0
Gene4	0	2	2	1	30	30	37	23
Gene5	400	230	150	300	50	100	55	66

RNA-Seq Analysis – Overview



RNA-Seq Analysis – Setup

Required files:

1) Sample info file (.csv)
   - Manually created by you
   - Defines conditions/replicates for comparison
2) Count matrix (.csv)
   - Can generate within script (commented section)
     - Requires .bam files for all samples

Currently supports mm10, hg38, and rn6 assembly:

- Genome annotation (.gtf) file required if other assembly needed and not built-in with functions

RNA-Seq Analysis – Options

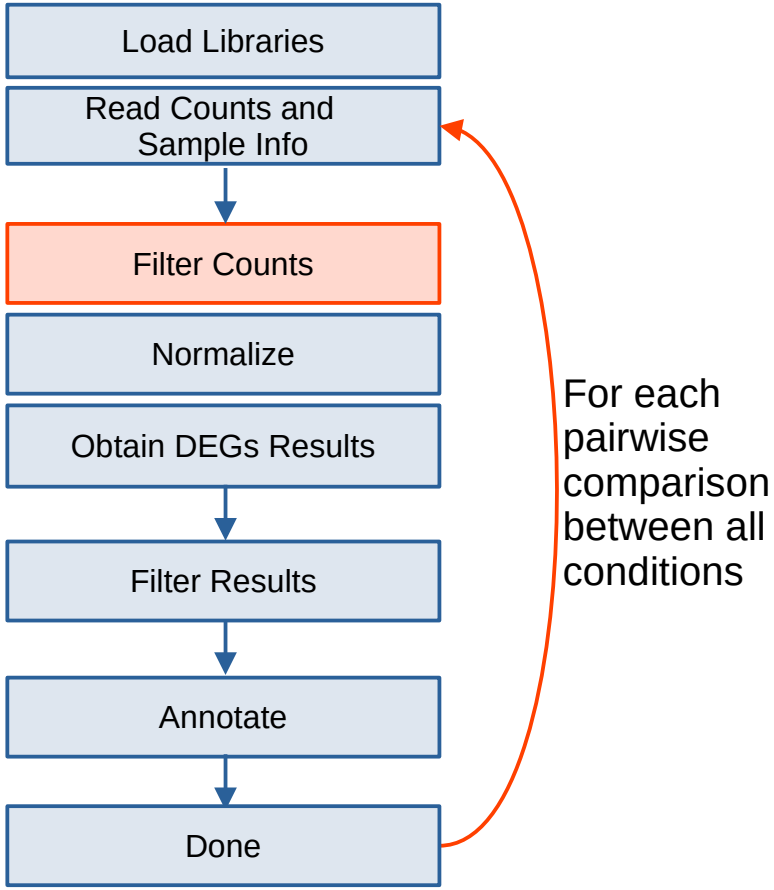
Rscript analyze_rnaseq_degs_DESeq2.R

and/or

Rscript analyze_rnaseq_degs_edgeR.R

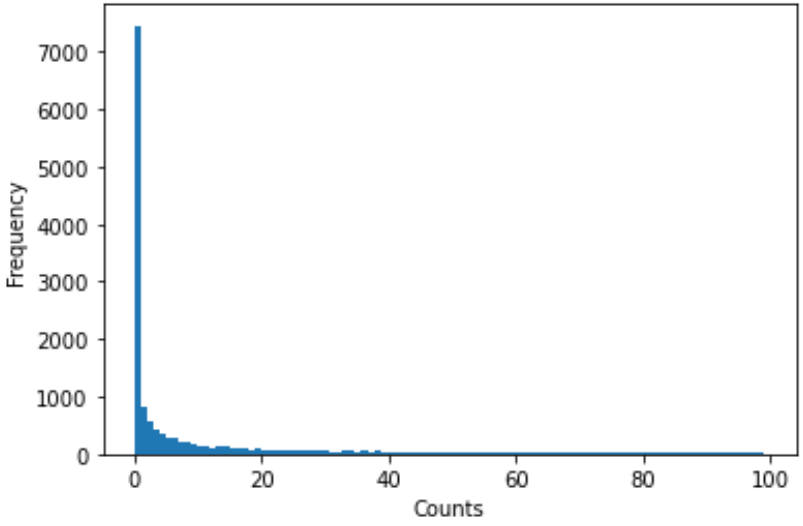
- --countsfile (required)
- --sampleinfo (required)
- --organism (default: mouse)
- --result_dir (default: DEG_Analysis/)
- --filter (default: FALSE)
- --min_count (default: 1)
- --min_basemean (default: 10)
- --lfc (log2foldchange, default: 0.585)
- --pvalue (default: 0.05)

RNA-Seq Analysis – Overview

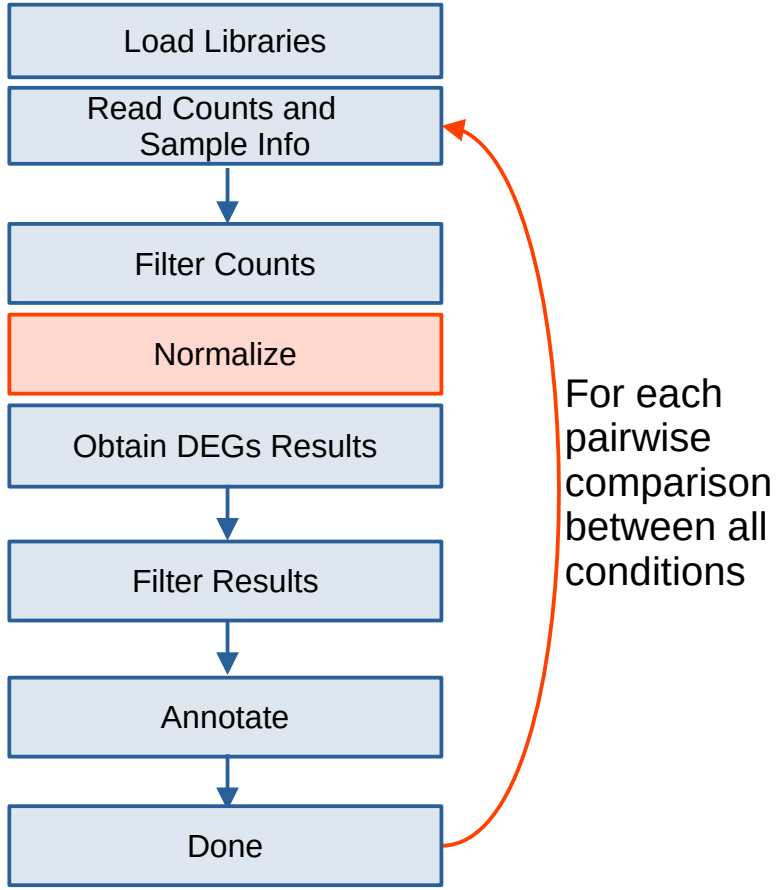


RNA-Seq – Filter Counts

- Remove genes with 0 counts across samples
  - Removes irrelevant genes
  - Reduces biases (variance+means) in computing DEGs
- Most genes have little-to-no expression

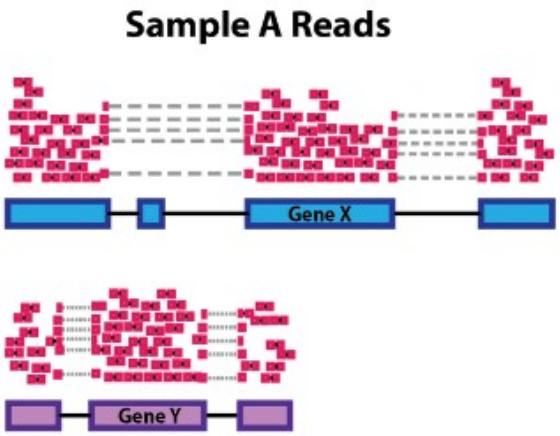


RNA-Seq Analysis – Overview



RNA-Seq – Normalizing Counts

(Gene length bias)



Example *within-sample* comparison:

- Comparing gene X and gene Y
- Sequenced to same depth
- Gene X has more reads mapped due to gene length
- May appear that gene X is enriched more than gene Y

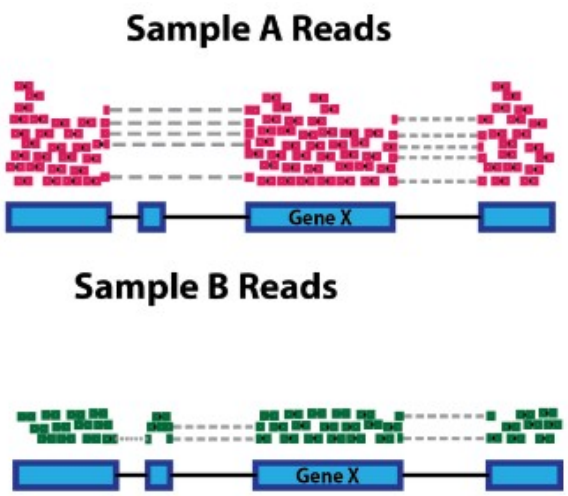
Normalize using TPM:

- Counts per length of transcript (kb) per million mapped reads

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Normalizing Counts

(Sequence depth bias)



Example *between-sample* comparison:

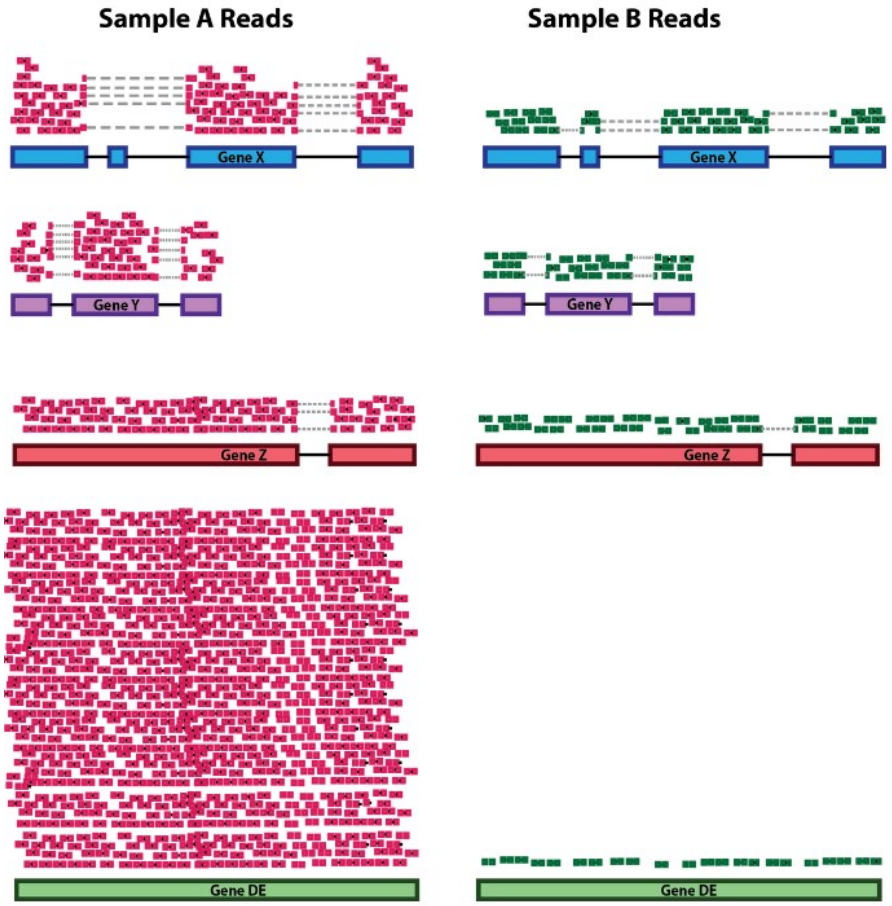
- Comparing gene X for Sample A and Sample B
- Sample A sequenced ~2x deeper than Sample B
- Appears that treatment for Sample A enriches gene X expression

Normalize using CPM/TPM:

- Divides gene counts by total number of reads

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Normalizing Counts



Example *between-sample* (DEG Analysis):

- Dividing gene counts by total reads for each sample
  - Gene X, Y, Z in Sample A divided by larger value due to DE gene
  - X, Y, Z would appear to be expressed less in Sample A
- CPM/TPM normalization not exactly appropriate here

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Normalizing Counts

DESeq2

- Median of ratios

edgeR

- Trimmed mean of M values (TMM)

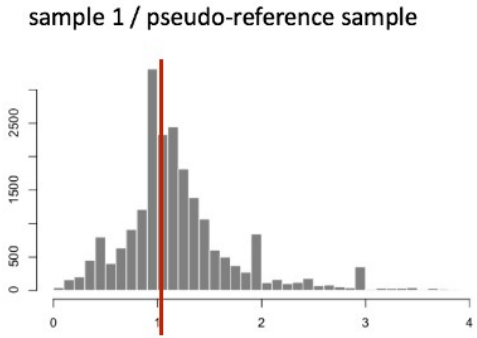
Since comparing DEGs between samples (gene-to-gene, same genome), these assume gene length is constant. Thus, does not (no need to) account for gene length.

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Normalizing Counts

DESeq2

- Median of ratios



Gene	Sample A Counts	Sample B Counts	Pseudo-reference sample	Sample A/reference Ratio	Sample B/reference Ratio	Median of Ratios A	Median of Ratios B	Normalized Sample A	Normalized Sample B
ABC1	**1489**	**906**	sqrt(1489*906) =1161.48	1489/1161.48 =1.28	906/1161.48 =0.78	Median (1.28, 1.30, 1.39) =1.3	Median (0.78, 0.77, 0.72) =0.77	1489/1.3 =**1145.4**	906/0.77 =**1246.8**
DEF2	**22**	**13**	sqrt(22*13) =16.91	22/16.91 =1.30	13/16.91 =0.77			22/1.3 =**16.9**	13/0.77 =**16.9**
XYZ3	**793**	**410**	sqrt(793*410) =570.20	793/570.20 =1.39	410/570.20 =0.72	1.3 1.3 1.3 1.3	0.77 0.77 0.77 0.77	793/1.3 =**610**	410/0.77 =**532.5**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

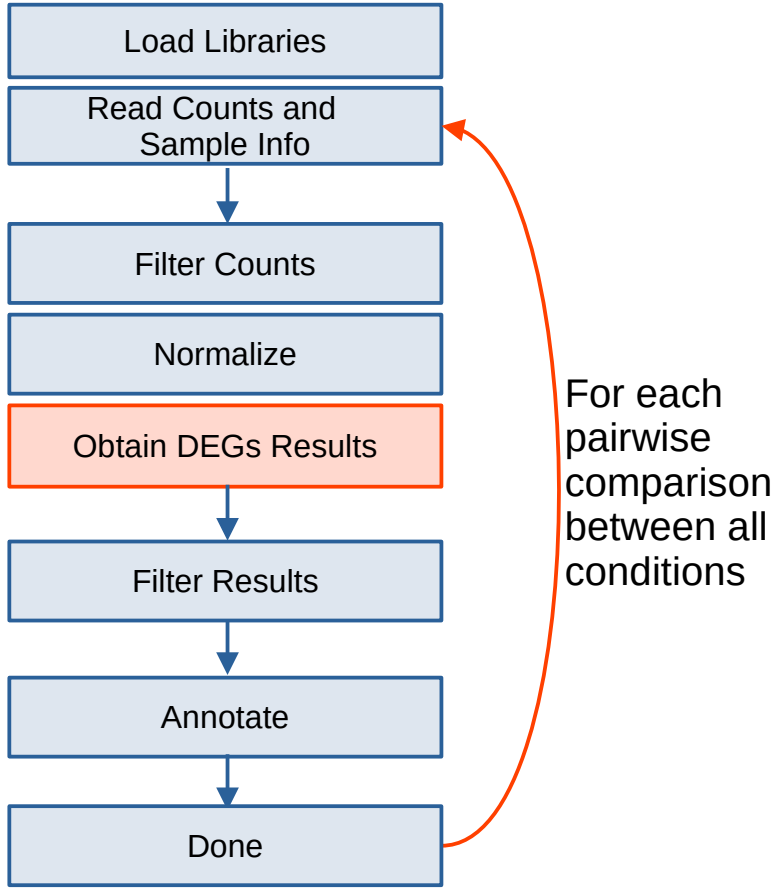
RNA-Seq – Normalizing Counts

edgeR

- Trimmed mean of M values (TMM)
  - Trimmed mean of log expression, assumes most genes are not differentially expressed
  - Removes extreme outliers and computes mean counts relative to library size
  - New scaling factor created for effective library size between samples

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq Analysis – Overview



RNA-Seq – Obtain DEGs

Differentially expressed genes (DEGs) are identified based on statistical metrics at defined thresholds

Metrics:

- $\log_2$(fold-change), $\log_{10}$(fold-change) *where fold-change = value2/value1*
- p-value, adjusted p-value a.k.a false-discovery rate (FDR) a.k.a q-value

Metric	Purpose	Example Value	Interpretation
$\log_2$(fold-change)	Magnitude of difference in gene expression	1	Sample2 expresses twice that of Sample1
adjusted p-value	Confidence in difference of gene expression	0.05	95% certainty that DEG isn't just random, only 5 false positive in 100 true

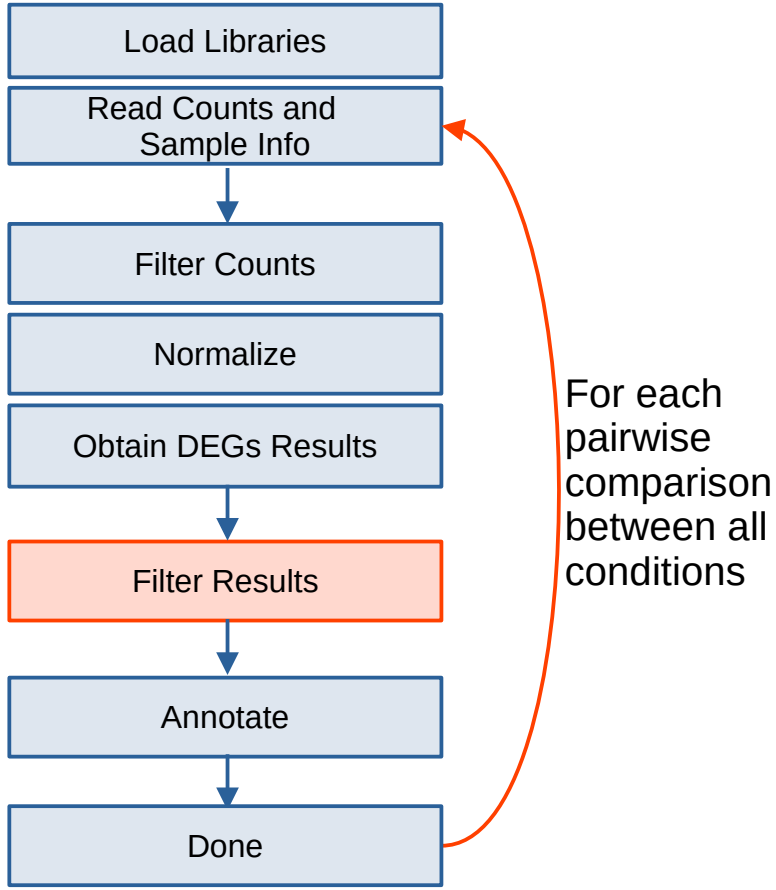
RNA-Seq – Obtain DEGs

With normalized counts:

- Compute baseMean of each gene (average of normalized counts across samples)
- Compute log2FoldChange of gene
  - $\log_2$(mean(condition 2) / mean(condition 1))
  - Standard error associated
- Compute p-value for each gene
  - Null hypothesis = no difference between conditions
  - Compute adjusted p-value (sort p-values, (rank/n)*FDR)

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq Analysis – Overview



RNA-Seq – Filter DEG Results

Magnitude:

- Up-regulated genes: logFoldChange > 1.5
- Down-regulated genes: logFoldChange < -1.5

Confidence:

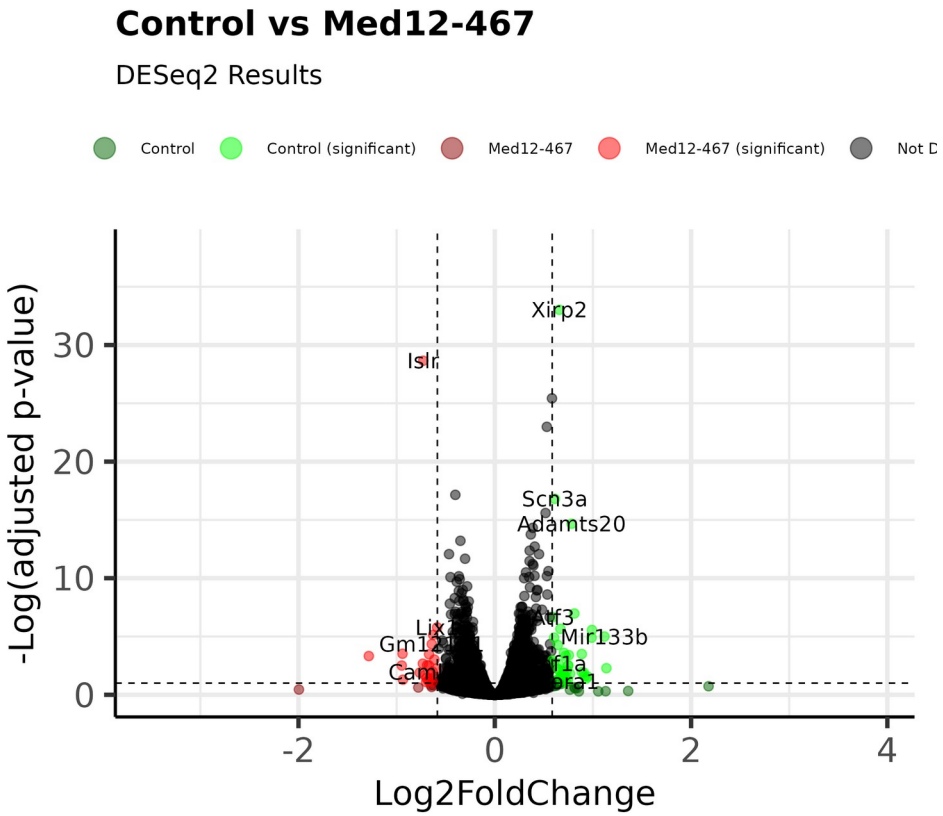
- Statistically significant: (adjusted) p-value <= 0.05

These values are chosen at your discretion as input parameters.

Note: p-value used as input in case not enough data points for genes to compute adjusted p-value.

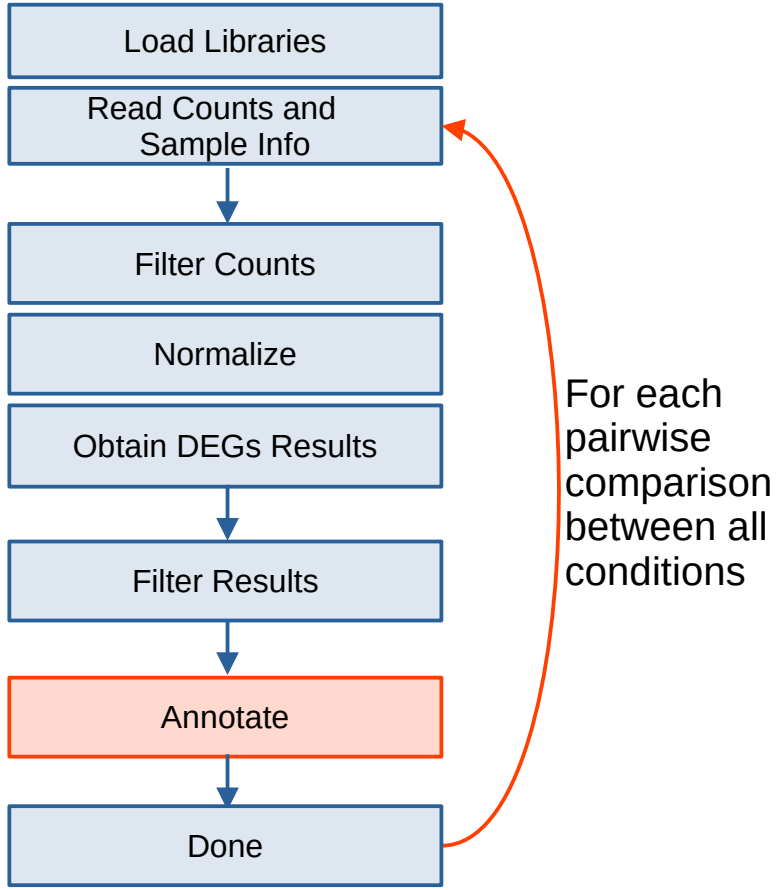
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Filter DEG Results



Total = 14104 genes
Control = 350 DEGs
Med12-467 = 139 DEGS

RNA-Seq Analysis – Overview



RNA-Seq – Annotate Results

1. Gene Ontology (GO):
   - Cellular Component (CC)
   - Molecular Function (MF)
   - Biological Process (BP)

2. Kyoto Encyclopedia of Genes and Genomes (KEGG):
   - High-level functional pathways associated with genes

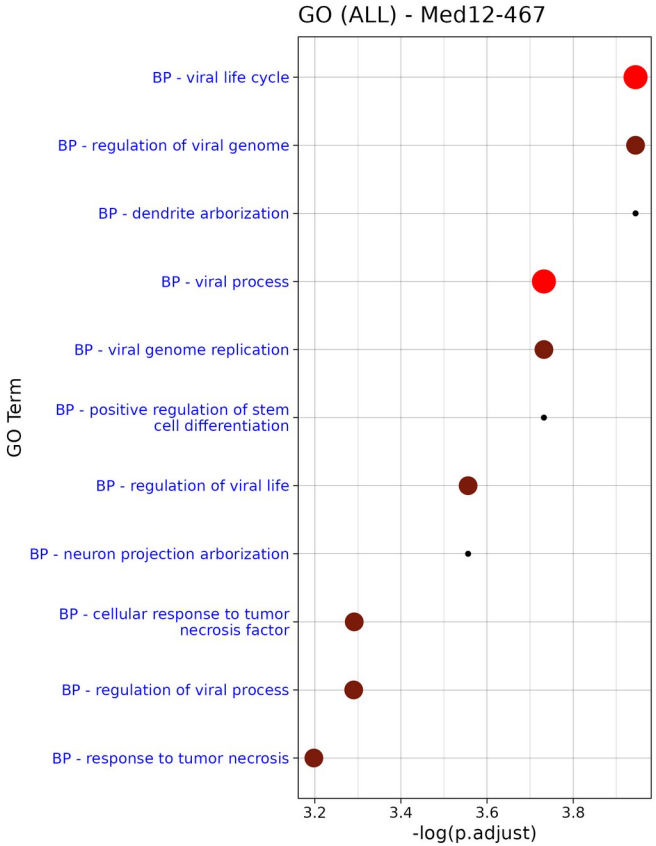
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

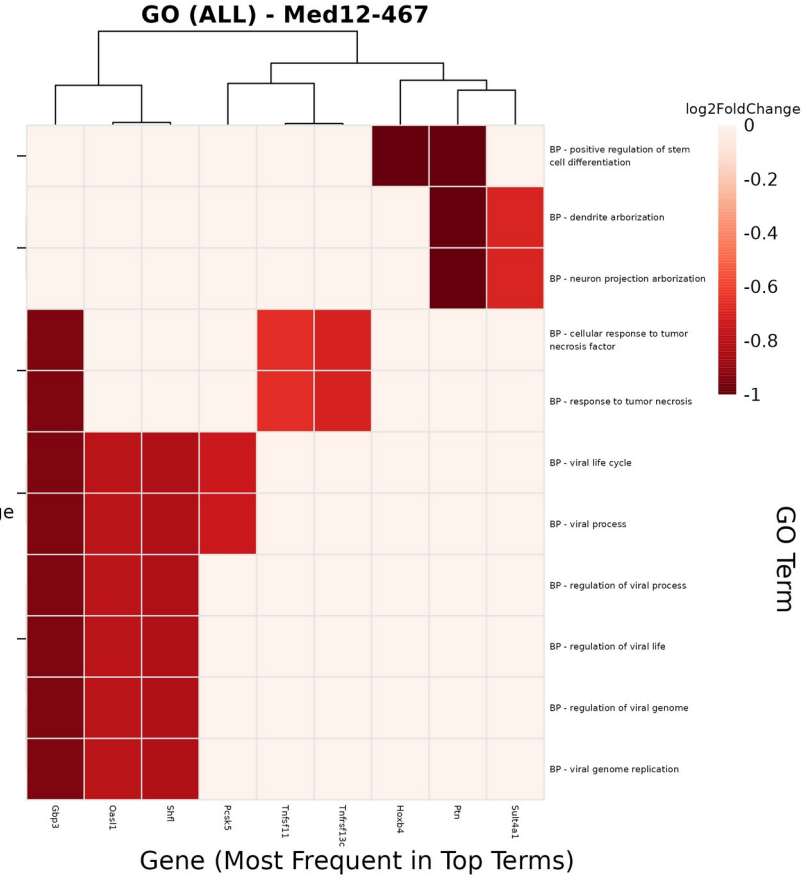
RNA-Seq – Annotate Results

- Take list of genes from DEG results
- Take database with functional information about genes
- Annotate DEG results with term of functional information
  - GeneRatio: genes in list with term / genes in list
  - BgRatio: all known genes with term / genes in database
  - p-value, padj, etc...

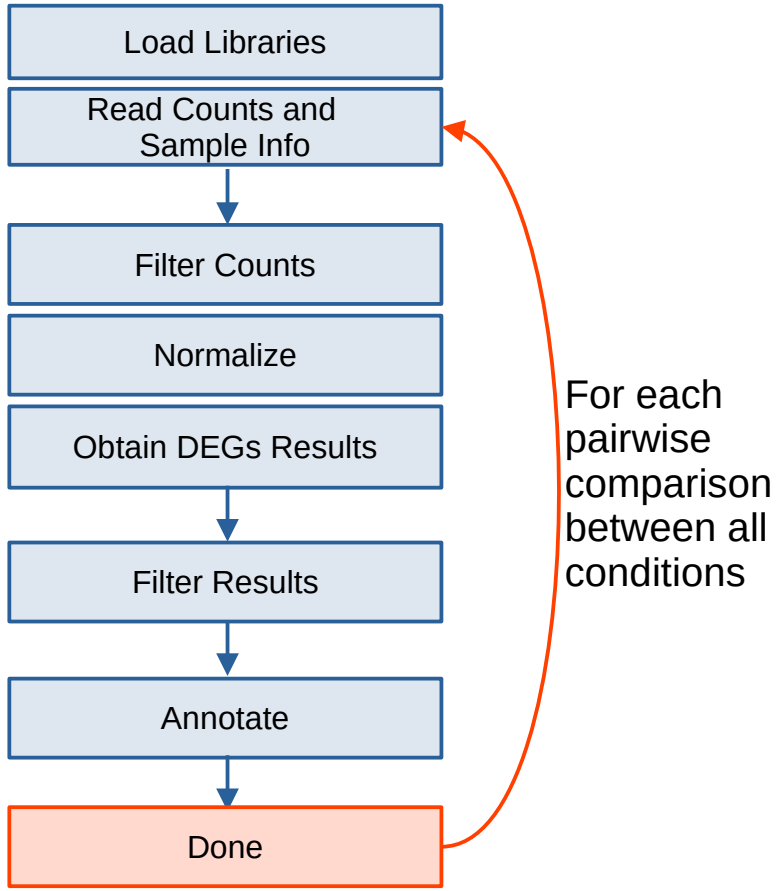
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

RNA-Seq – Annotate Results





RNA-Seq Analysis – Overview



Learning Goals

<u>DEG Analysis:</u>

1) Bulk RNA-Seq Analysis:
   - Count matrix creation and sample sheet file preparation
   - Normalization and DESeq2 + edgeR
   - Computation of DEGs
   - Annotations and figures
2) **DiffBind Peaks Analysis**:
   - Sample sheet file preparation
   - Consensus peaksets
   - Occupancy analysis
   - Affinity analysis

Peaks Analysis – Preface

- At least 2 replicates required for each sample
- Alignment files (.bam + .bai) **without** duplicates should be used for peaks analysis
  - However, program will automatically remove duplicates if .bam files with duplicates are provided
  - Retains library complexity without PCR duplication artifacts
- Similar to RNA-Seq, but relies on peaks files instead of a counts matrix

Peaks Analysis – Options

Rscript analyze_peaks_degs.R

--file (required)

--fragmentsizes (default: NULL)

--organism (default: mouse)

--result_dir (default: Peaks_Analysis/)

--database (default: ucsc)

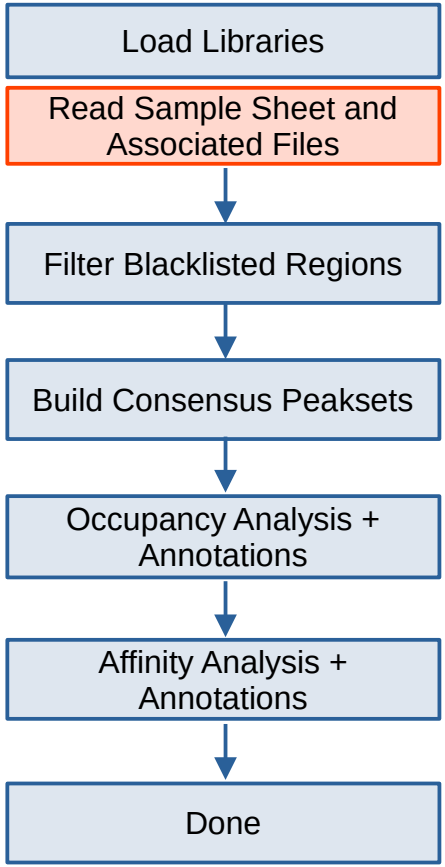
--min_count (default: 1)

--add_replicates (default: FALSE)

--lfc (log2foldchange, default: 0.585)

--fdr (default: 0.05)

Peaks Analysis – Overview



Peaks Analysis – Sample sheet

DiffBind sample sheet (.csv) – create manually

SampleID	Condition	Replicate	Peaks	PeakCaller	PeakFormat	bamReads	Factor
WT-1-SEACR	WT	1	WT-1.stringent.bed	bed	bed	WT-1.Mapped.MAPQ10.bam	SEACR
WT-1-MACS	WT	1	WT-1_peaks.narrowPeak	bed	bed	WT-1.Mapped.MAPQ10.bam	MACS
WT-1-GoPeaks	WT	1	WT-1_gopeaks_peaks.bed	bed	bed	WT-1.Mapped.MAPQ10.bam	GoPeaks
WT-2-SEACR	WT	2	WT-2.stringent.bed	bed	bed	WT-2.Mapped.MAPQ10.bam	SEACR
WT-2-MACS	WT	2	WT-2_peaks.narrowPeak	bed	bed	WT-2.Mapped.MAPQ10.bam	MACS
WT-2-GoPeaks	WT	2	WT-2_gopeaks_peaks.bed	bed	bed	WT-2.Mapped.MAPQ10.bam	GoPeaks
KO-1-SEACR	KO	1	KO-1.stringent.bed	bed	bed	KO-1.Mapped.MAPQ10.bam	SEACR
KO-1-MACS	KO	1	KO-1_peaks.narrowPeak	bed	bed	KO-1.Mapped.MAPQ10.bam	MACS
KO-1-GoPeaks	KO	1	KO-1_gopeaks_peaks.bed	bed	bed	KO-1.Mapped.MAPQ10.bam	GoPeaks
KO-2-SEACR	KO	2	KO-2.stringent.bed	bed	bed	KO-2.Mapped.MAPQ10.bam	SEACR
KO-2-MACS	KO	2	KO-2_peaks.narrowPeak	bed	bed	KO-2.Mapped.MAPQ10.bam	MACS
KO-2-GoPeaks	KO	2	KO-2_gopeaks_peaks.bed	bed	bed	KO-2.Mapped.MAPQ10.bam	GoPeaks

Peaks Analysis – File info

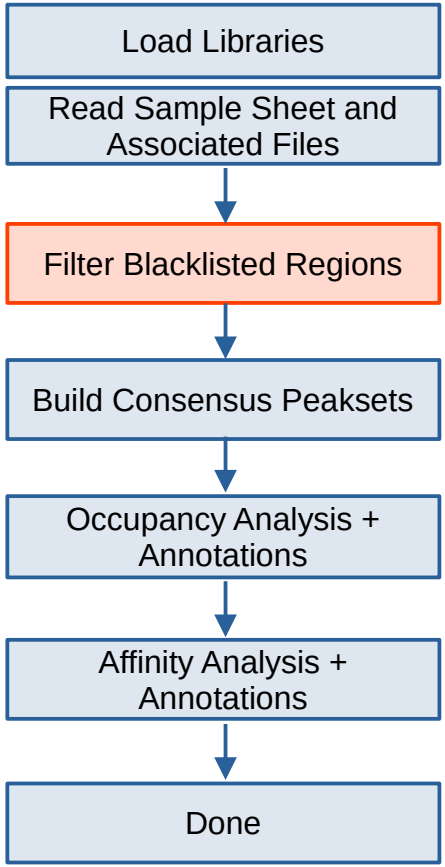
Peak regions are loaded from files defined in samplesheet.

E.g.

- 340,487 unique sites (ignoring overlapping intervals)

```
Raw peaksets:
> dbObj
12 Samples, 340487 sites in matrix:
                   ID  Factor Condition Replicate Intervals
1    Ac-TRF2-1-SEACR   SEACR   Ac-TRF2         1     74911
2     Ac-TRF2-1-MACS    MACS   Ac-TRF2         1      1402
3  Ac-TRF2-1-GoPeaks GoPeaks   Ac-TRF2         1        87
4    Ac-TRF2-2-SEACR   SEACR   Ac-TRF2         2     95329
5     Ac-TRF2-2-MACS    MACS   Ac-TRF2         2      2449
6  Ac-TRF2-2-GoPeaks GoPeaks   Ac-TRF2         2        94
7     Ac-IgG-1-SEACR   SEACR    Ac-IgG         1     92869
8      Ac-IgG-1-MACS    MACS    Ac-IgG         1      5162
9   Ac-IgG-1-GoPeaks GoPeaks    Ac-IgG         1       190
10    Ac-IgG-2-SEACR   SEACR    Ac-IgG         2    131290
11     Ac-IgG-2-MACS    MACS    Ac-IgG         2      6413
12  Ac-IgG-2-GoPeaks GoPeaks    Ac-IgG         2       146
```

Peaks Analysis – Overview

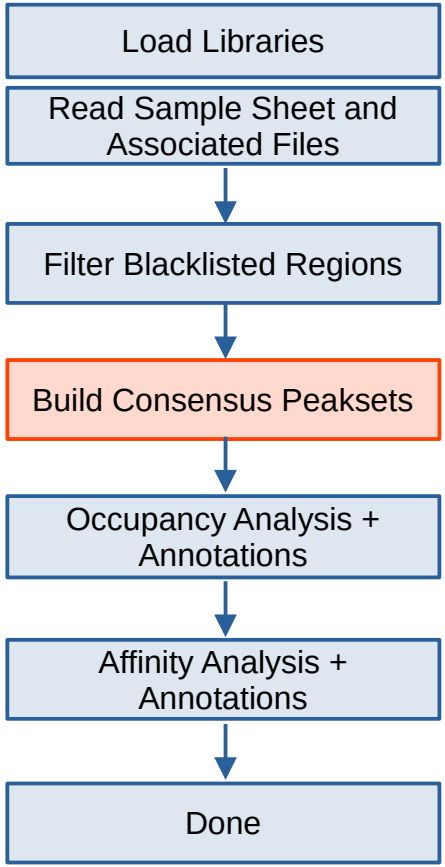


Peaks Analysis – Filter Blacklisted Regions

Remove sites known to be problematic artifacts in sequencing/alignment

```
> dbObj.noblacklist
12 Samples, 46401 sites in matrix (332718 total):
                  ID  Factor Condition Replicate Intervals
1    Ac-TRF2-1-SEACR   SEACR   Ac-TRF2         1     72388
2     Ac-TRF2-1-MACS    MACS   Ac-TRF2         1      1145
3  Ac-TRF2-1-GoPeaks GoPeaks   Ac-TRF2         1        15
4    Ac-TRF2-2-SEACR   SEACR   Ac-TRF2         2     92280
5     Ac-TRF2-2-MACS    MACS   Ac-TRF2         2      2124
6  Ac-TRF2-2-GoPeaks GoPeaks   Ac-TRF2         2        16
7     Ac-IgG-1-SEACR   SEACR    Ac-IgG         1     90005
8      Ac-IgG-1-MACS    MACS    Ac-IgG         1      4753
9   Ac-IgG-1-GoPeaks GoPeaks    Ac-IgG         1        87
10    Ac-IgG-2-SEACR   SEACR    Ac-IgG         2    127624
11     Ac-IgG-2-MACS    MACS    Ac-IgG         2      5875
12  Ac-IgG-2-GoPeaks GoPeaks    Ac-IgG         2        31
```

Peaks Analysis – Overview



Peaks Analysis – Consensus Peaksets

Reads processing pipeline uses 3 peak callers:

- MACS (.narrowPeak file)
- SEACR (.stringent.bed file)
- GoPeaks (gopeaks_peaks.bed file)

Consolidating resulting peak regions called by these tools provides more validation in each peak called (or not called).

Likewise, replicates also provide additional validation.

Peaks Analysis – Consensus Peaksets

True Peak

True No Peak

1/3 callers suggest peak

2/3 callers suggest peak

Peaks Analysis – Consensus Peaksets

Steps to build consensus:

1) Caller consensus:
   - Merge peaks found in 2/3 peak callers for each replicate
2) Replicate consensus:
   - Merge peaks found in at least 2 replicates' caller consensus or *(if very few sites resulted...)*
   - Combine peaks from all replicates' caller consensus

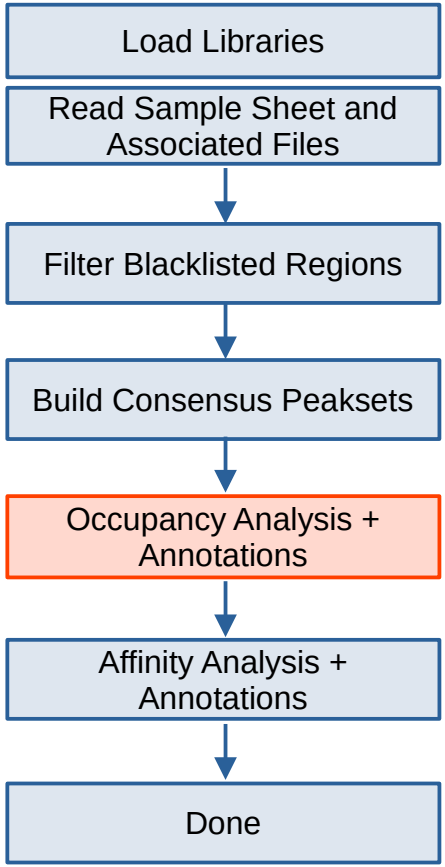
Peaks Analysis – Consensus Peaksets

Final consensus sites are used for occupancy + affinity analysis (differential binding analysis)

```
Consensus between peak callers...
> dbObj.caller_consensus
4 Samples, 12112 sites in matrix:
         ID              Factor Condition Replicate Intervals
1 Ac-TRF2:1 SEACR-MACS-GoPeaks   Ac-TRF2         1      1131
2 Ac-TRF2:2 SEACR-MACS-GoPeaks   Ac-TRF2         2      2108
3  Ac-IgG:1 SEACR-MACS-GoPeaks    Ac-IgG         1      4672
4  Ac-IgG:2 SEACR-MACS-GoPeaks    Ac-IgG         2      5798

Consensus between replicates...
> dbObj.consensus
2 Samples, 1065 sites in matrix:
       ID              Factor Condition Replicate Intervals
1 Ac-TRF2 SEACR-MACS-GoPeaks   Ac-TRF2       1-2       119
2  Ac-IgG SEACR-MACS-GoPeaks    Ac-IgG       1-2      1009
```

Peaks Analysis – Overview

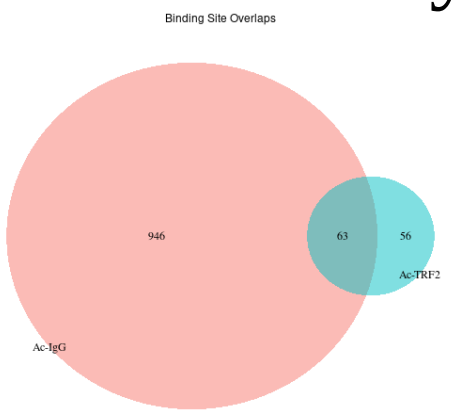


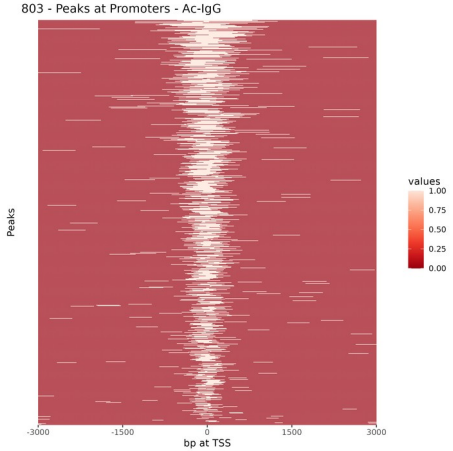
Peaks Analysis – Occupancy Analysis

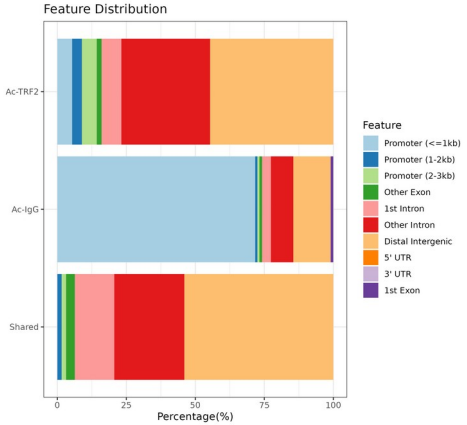
DiffBind's occupancy analysis:

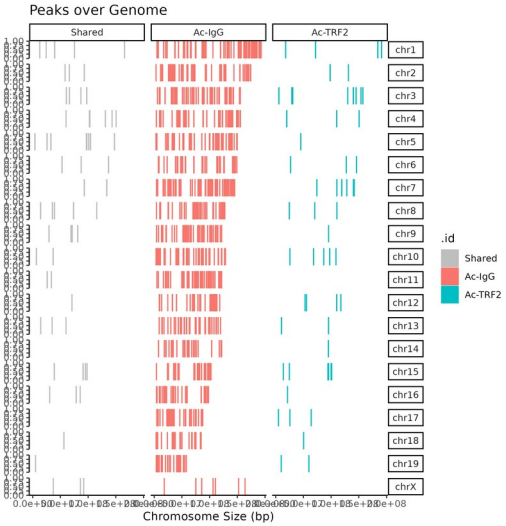
- Strictly considers binding site regions in peakset
- Does not consider information about numbers of mapped reads at each region

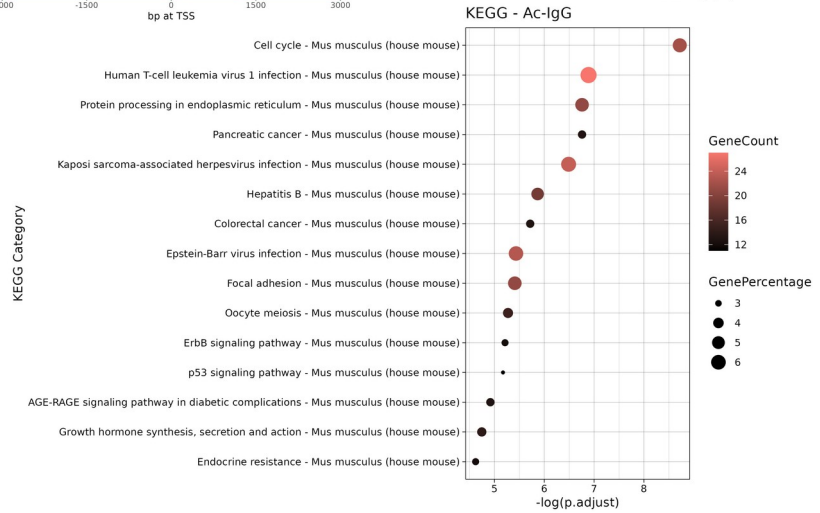
Peaks Analysis – Occupancy Analysis



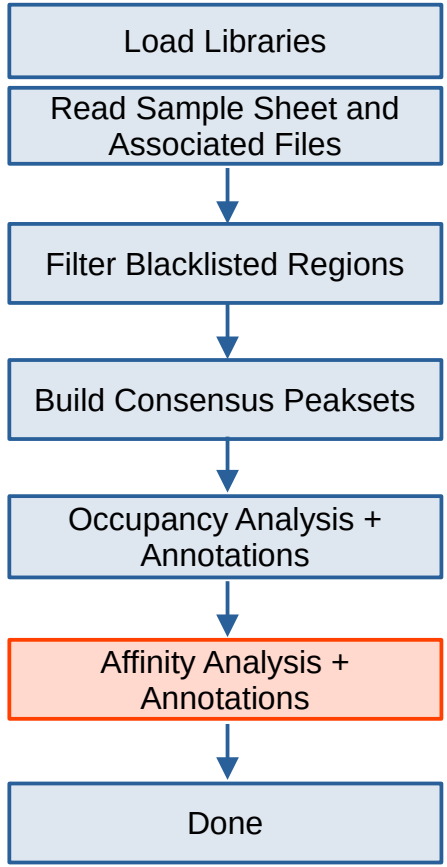








Peaks Analysis – Overview

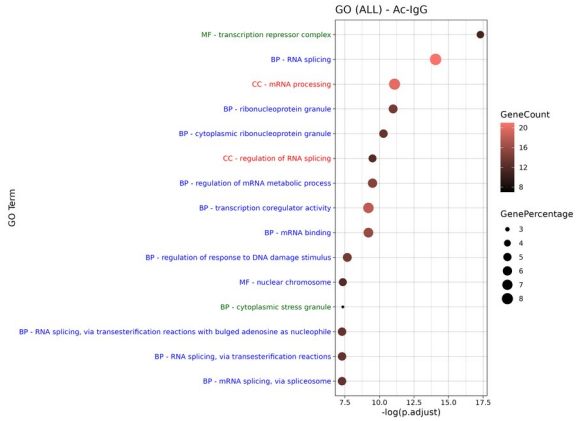


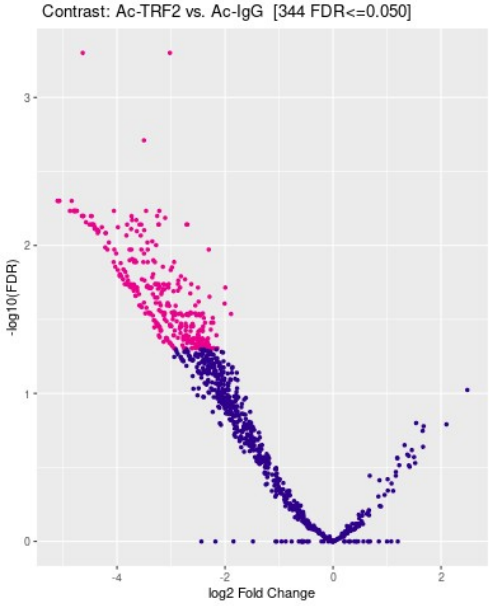
Peaks Analysis – Affinity Analysis

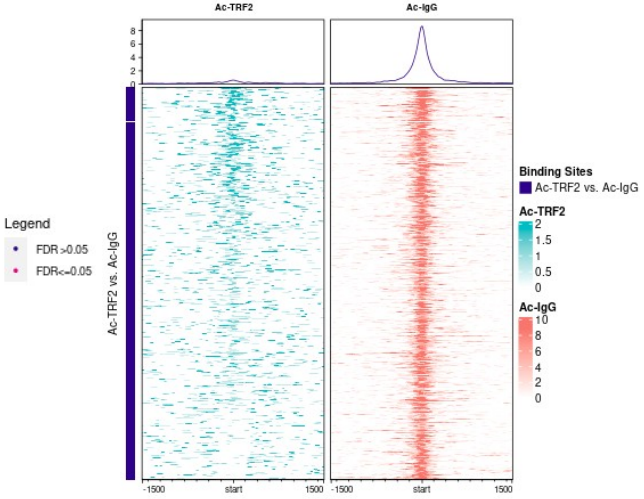
DiffBind's affinity analysis:

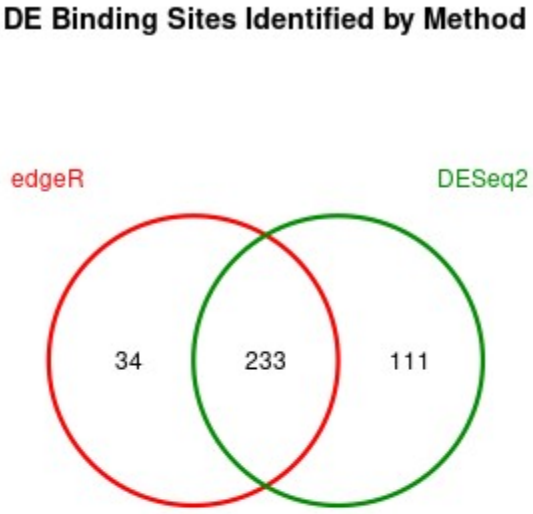
- Performs DESeq2 and edgeR on regions defined in consensus peaksets
- Output here is similar to the RNA-Seq scripts
  - Annotations
  - Plots
  - Differential binding analysis result files

Peaks Analysis – Affinity Analysis

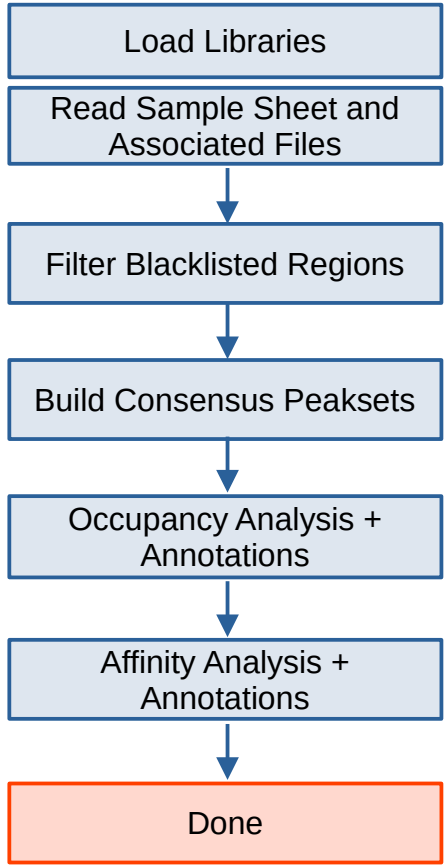








Peaks Analysis – Overview



Notes on Files

Conventions:

- “DESeq2” and “edgeR” in name for their respective outputs
- Pairwise comparisons, conditions are color-coded (mostly) for figures
  - Figures are generated from respective .tsv or .csv files
    - All detailed info is there, more genes may exist than in figure
  - Condition in filename for respective genes that are up-regulated compared to other condition
  - “DEG” in filename for all genes with logfoldchange > |1.5|

Troubleshooting

If certain files aren't generated:

- No DEGs were found
  - Try relaxing constraints
- Dataset may require different handling
  - Double-check counts files, sample info files, samplesheets, alignment files, peaks files, etc…
  - Try running step-by-step in Rstudio to see where problem exists
  - Contact earezza@ohri.ca

Trouble running script:

- Double-check R version and packages installed
- Double-check input options and relative filepaths

Questions?

Contact:
earezza@ohri.ca