# RNA-Seq Data Analysis 2

Lauren Vanderlinden

BIOS 6660

Spring 2019

# Overview From Last Time

- Data visualization is important!
- Learned about various normalization methods
- Able to run a GLM in DESeq2

```
> strain.adjForTissue = results(dds.cov)
> head(strain.adjForTissue)
log2 fold change (MLE): strain WhiteEyed vs Sevenless
Wald test p-value: strain WhiteEyed vs Sevenless
DataFrame with 6 rows and 6 columns
                   baseMean      log2FoldChange             lfcSE                stat
                  <numeric>           <numeric>         <numeric>           <numeric>
FBgn0000003 699.145888361084   0.481065789920314 0.315813565099327    1.52325879279129
FBgn0000008 2954.45567305398   0.180576506548756 0.173719729223854    1.03947034315294
FBgn0000014 8.21095639029486   -1.51479593858499 0.982522214724989   -1.54174217730943
FBgn0000015 3.79786939660886 -0.0561051226257818  1.37892407109896 -0.0406876084054923
FBgn0000017  5671.9724202071   0.227839409937837 0.152736766573332    1.49171293231775
FBgn0000018 113.196237762821   0.264910424695025  0.19763036163017     1.3404338407818
                     pvalue              padj
                  <numeric>         <numeric>
FBgn0000003 0.127693972189392 0.353716203604252
FBgn0000008 0.298586043835595 0.567851795440525
FBgn0000014 0.123136258247327 0.346438222750232
FBgn0000015 0.967544940475883 0.987751006971347
FBgn0000017 0.135774417382166 0.365765984149215
FBgn0000018 0.180104339899211 0.430512887063497
```

# Independent Filtering

**Independent filtering increases detection power for high-throughput experiments**

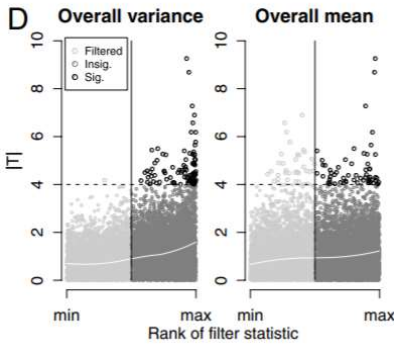
Richard Bourgon[a], Robert Gentleman[b], and Wolfgang Huber[c,1]

[a]European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [b]Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990; and [c]European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 22, 2010 (received for review December 3, 2009)

- DESeq2 default method in their analysis is to use this independent filtering
- Example of 2x2 design:
  - 15,127 genes
  - 4,681 filtered out prior to multiple testing correction

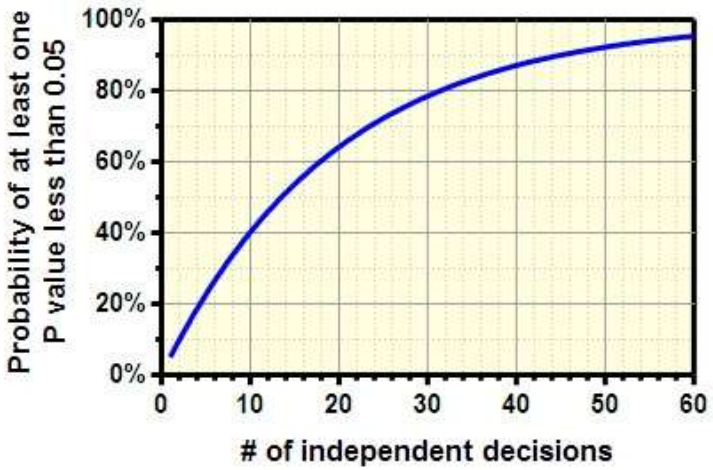
- Suppose to optimize the number of adjusted p-value less than a critical alpha level (default = 0.1). In the DESeq2 manual it says the mean (regardless of group) of the normalized counts is used as a filter, but the paper mentions also filtering on overall variance as if you used just the mean you would loose a lot of significant results.



# Multiple Testing Issue

- Performing the same statistical model on every feature in your dataset.
  - 20K genes, then you have 20K tests
- If you test several independent null hypotheses and leave the threshold at 0.05 for each comparison, the chance of obtaining at least one "statistically significant" result is greater than 5% (even if all null hypotheses are true).
  - 10,000 tests and leave alpha = 0.05 you would expect 500 false positive results (Yikes!)
  - Normally False Positives aren't an issue, but with omics data this is a huge issue

| | Null True | Alternative True |
|---|---|---|
| Tested Non-Significant | Correct Call | False Negative |
| Tested Significant | **False Positive** | Correct Call |




Source: wiki

# Multiple Testing Correction Methods

- Bonferroni
  - Take the p-value and multiple by n total tests
  - Most conservative and considered too conservative

$$p.bonferroni = \min(n * pvalue, 1)$$

- Šidák
  - Single adjustment as well

$$p.sidak = 1 - (1 - pvalue)^{n}$$

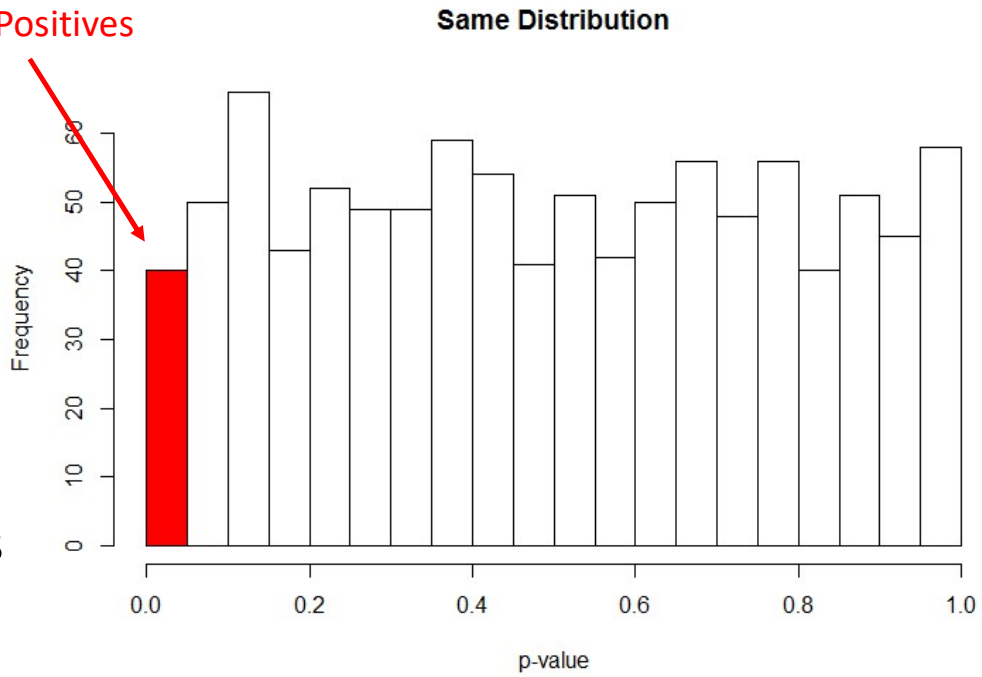
- False Discovery Rate (FDR)
  - Benjamini-Hochberg Method
  - Step-wise method

$$p.FDR = \min(pvalue * \frac{n}{n - rank}, FDR_{\text{previousRank}})$$

- All methods "adjust" p-values (i.e. make them larger)
- Assuming all tests are independent
  - We know this is not true in a biological system

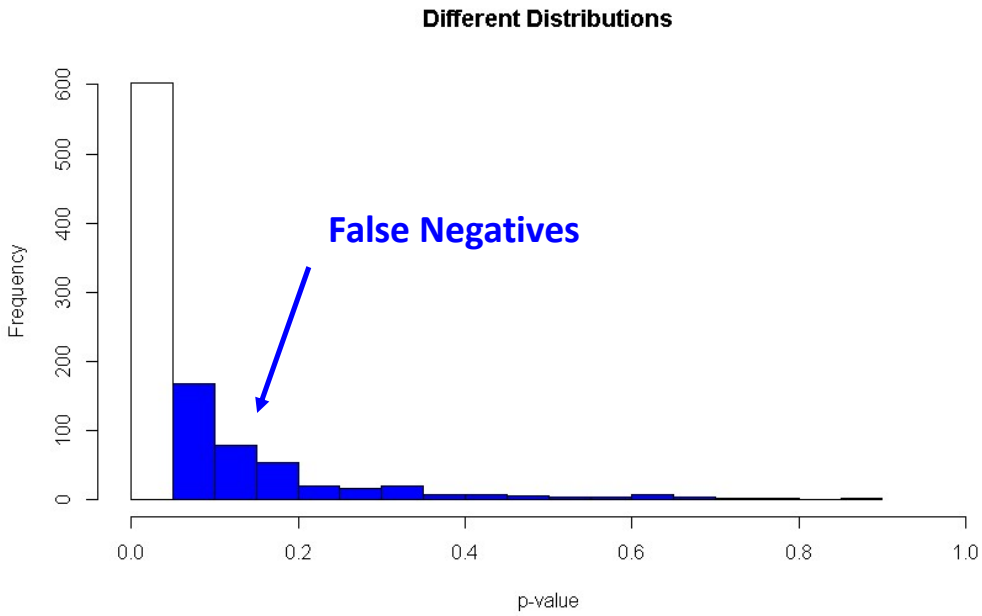
# Distribution of p-values in Null Hypothesis

- Basic 2-group comparison
  - T-test
- Took random 4 samples from same distribution
  - Normal (mean = 1, sd = 1)
- Looking at 1,000 tests
- Histogram of p-values:  Looks very uniform
- Equal probability a test p-value falls into any 1 of these bins
- Expect 50 false positives by chance
  - My example got 40 false positives



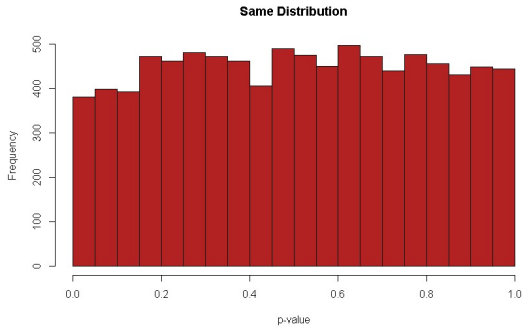
# Distribution of p-values from different distributions

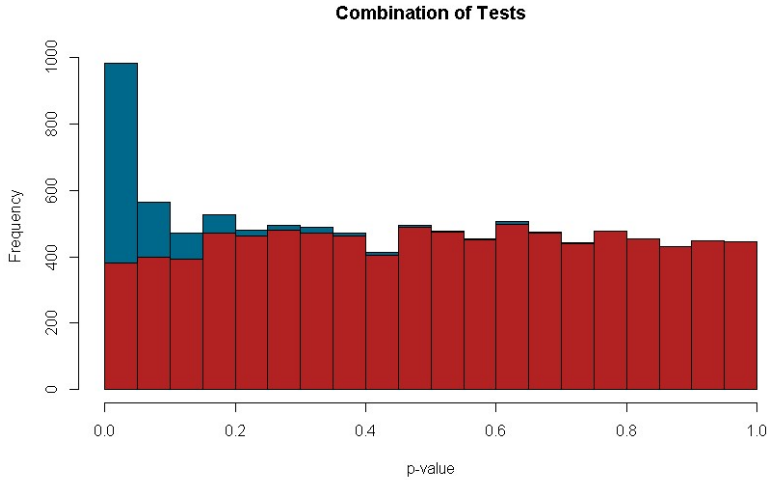
- Basic 2-group comparison
  - T-test
- Took random 4 samples from 2 different distributions
  - Normal (mean = 1, sd = 1)
  - Normal (mean = 3, sd = 1)
- Looking at 1,000 tests
- Histogram of p-values:  Heavily skewed towards 0
  - Can change how we would expect this by power

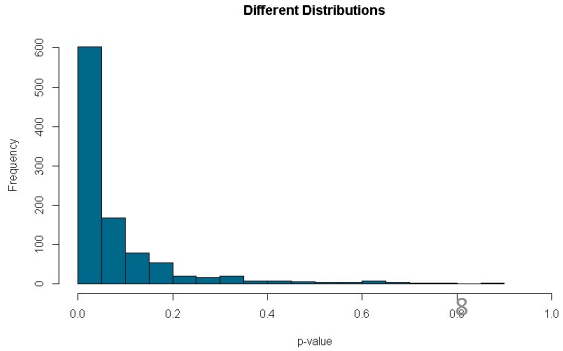


# Realistic Distribution of p-values

- In a study, you will have some features effected and some not
- For example, you are testing 10,000 genes in a treated vs control study
  - 9,000 genes are not effected by treatment
    - 9,000 p-values coming from same distribution
  - 1,000 genes are effected by treatment
    - 1,000 p-values coming from 2 different distributions

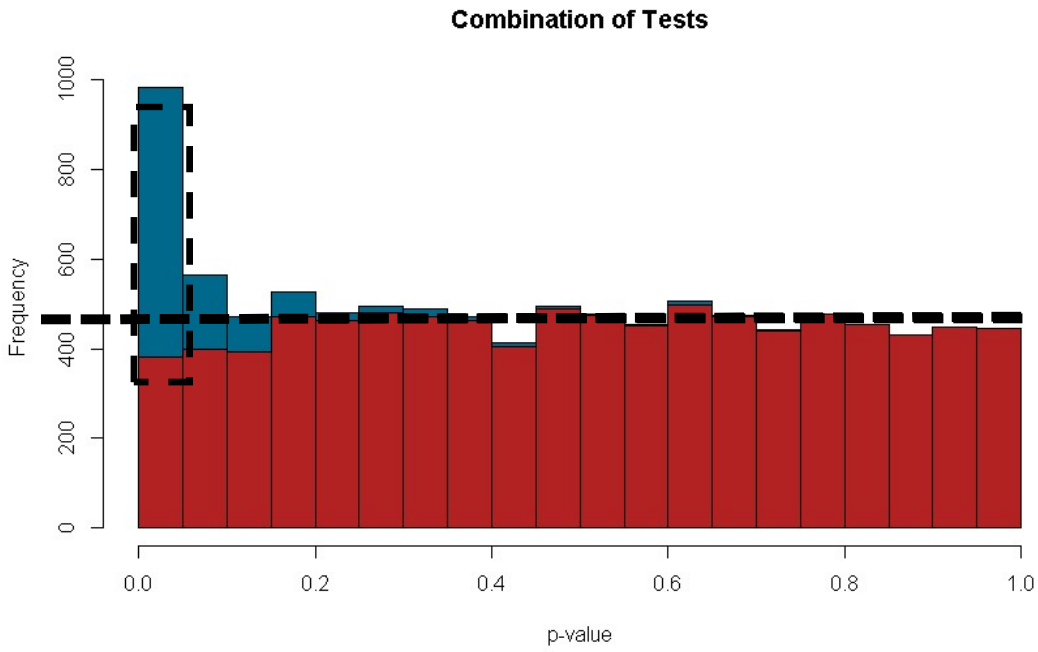


# FDR Estimates Uniform Background

- Tries to estimate a uniform background based on high p-values
  - For our example here, about 450
- Extend this line out to the lower p-values
  - At the <0.05 bin, we have about 550 above the dotted line
- How to sort out significant bin?
  - Take the smallest 550 p-values
  - Makes sense because of skewness seen in the true different p-value histogram
  - Still will expect 5% false positives because not all true positive will be extremely small



# Mathematical Model

- Rank p-values (say 10 from uniform distribution of p-values)
- Largest p-value is the same
- Moving down in descending rank, FDR the smallest option of
  - $pvalue * \frac{\#\ Tests}{\#\ Tests - rank}$
  - FDR for previous rank (rank + 1)

- For the 9th rank p-value:

0.81 * (10/9) = 0.90

Since 0.90 < 0.91, the FDR = 0.90

- Keep moving down the rank….

| p-value | 0.01 | 0.11 | 0.21 | 0.31 | 0.41 | 0.51 | 0.61 | 0.71 | 0.81 | 0.91 |
|---|---|---|---|---|---|---|---|---|---|---|
| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| FDR | 0.10 | 0.55 | 0.70 | 0.77 | 0.82 | 0.85 | 0.87 | 0.89 | 0.90 | 0.91 |

# Challenges

- Even though less conservative than Bonferroni, still considering all tests are independent from one another. In biological system we know this is not the case.
- Makes power calculations difficult. Most times in grants we say "using an alpha level of 5e-06, which is a 0.05 p-value Bonferroni adjusted for 10K tests…" and mention that this is a conservative estimate of power since we will be doing FDR, not Bonferroni in the analysis plan.

| p-value | 0.01 | 0.11 | 0.21 | 0.31 | 0.41 | 0.51 | 0.61 | 0.71 | 0.81 | 0.91 |
|---|---|---|---|---|---|---|---|---|---|---|
| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| FDR | 0.10 | 0.55 | 0.70 | 0.77 | 0.82 | 0.85 | 0.87 | 0.89 | 0.90 | 0.91 |
| Bonferroni | 0.10 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

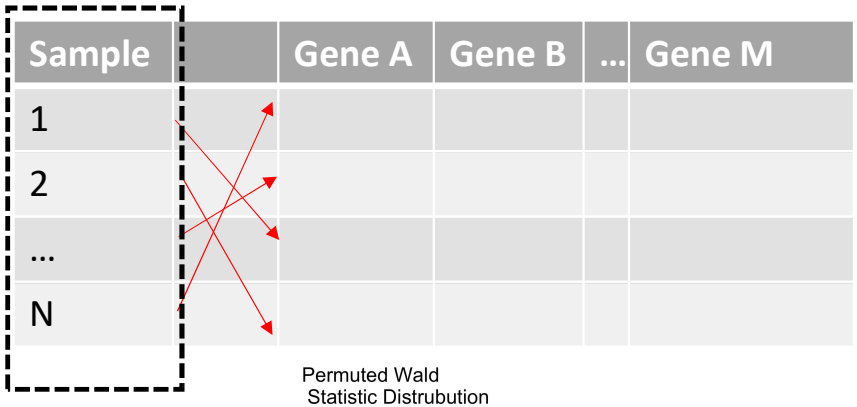
# DESeq2 Multiple Testing Options

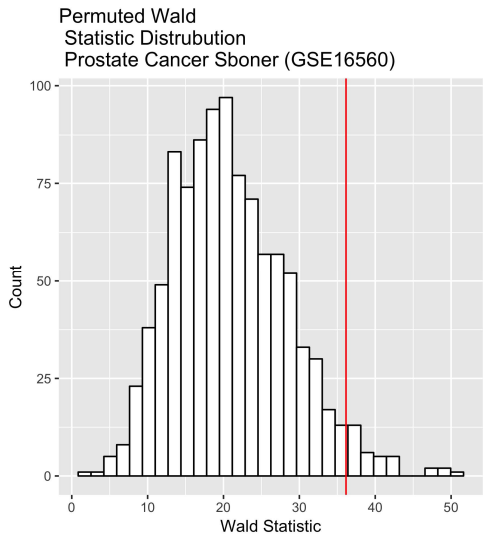
- Default adjustment is Benjamini Hochberg (aka “BH”)
- Can use pAdjustMethod
- Options to change to: c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none")
  - Note: fdr and BH are the same method

```
> results.bh = results(dds.inter)
> results.bh[1:2, c(1,6)]


DataFrame with 2 rows and 2 columns
                   baseMean                padj
                  <numeric>           <numeric>
FBgn0000003 699.145888361084  0.694121945967225
FBgn0000008 2954.45567305398 0.0276409235420397
> results.bon = results(dds.inter, pAdjustMethod="bonferroni")
> results.bon[1:2, c(1,6)]


DataFrame with 2 rows and 2 columns
                   baseMean      padj
                  <numeric> <numeric>
FBgn0000003 699.145888361084         1
FBgn0000008 2954.45567305398         1
```

# Permutation P-values

- Computationally INTENSIVE
- Generate empirical distribution of test statistics when the null hypothesis is true
- Randomly assign phenotypes to each sample, breaking your true relationship
- Perform tests on all permutations
  - Determine how many times your test statistic > permuted test statistic

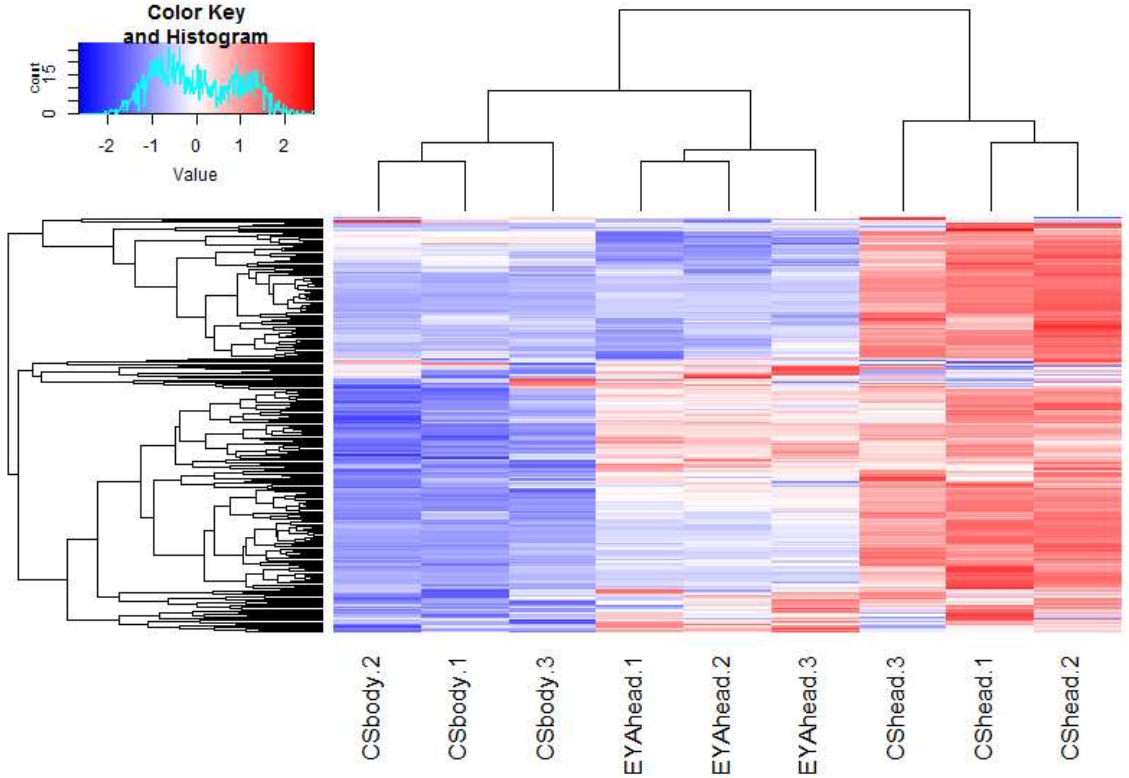


# Selecting Candidates

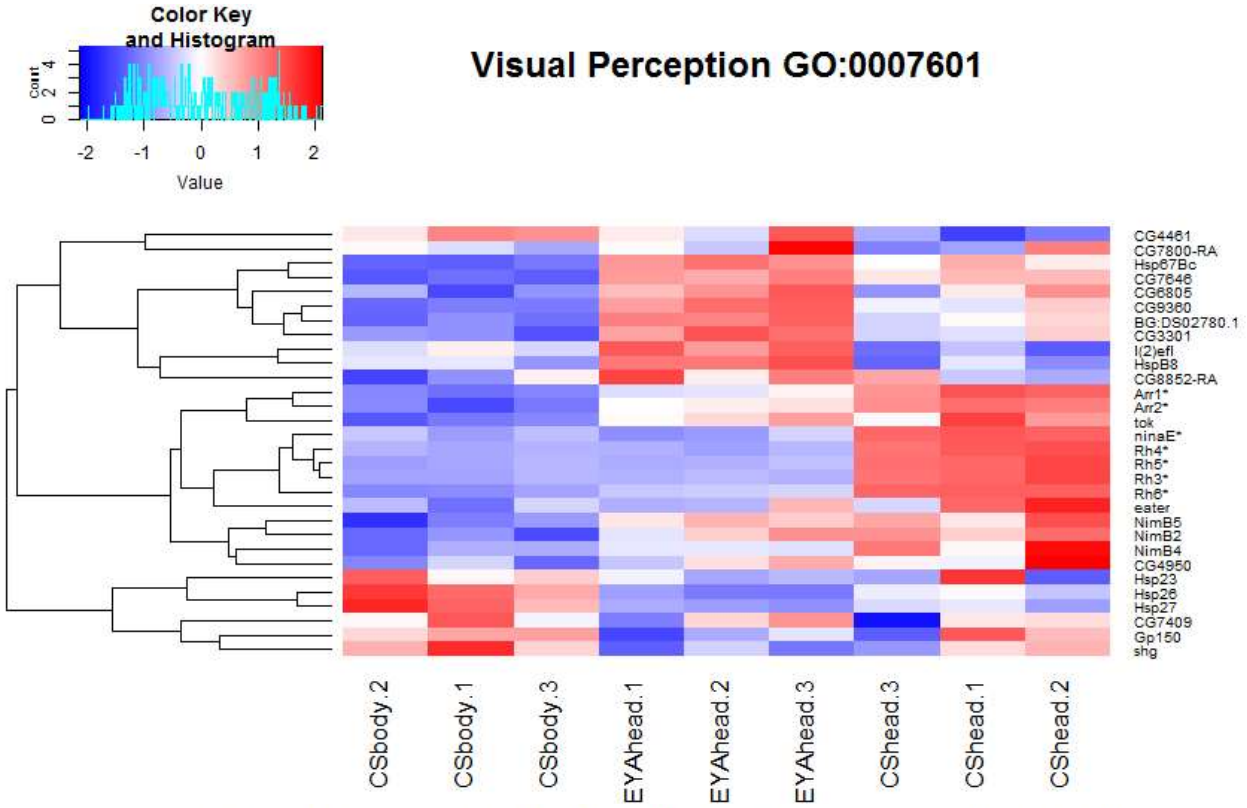
- Traditionally p-value < 0.05 is gold
- Omics has a little more wiggle room because of the multiple testing
  - Have seen < 0.1 considered significant
- How many candidates do you want?
  - Follow up validation studies
  - Enrichment analysis (need at bare minimum 20, ideally 100 or so)

# Visualizing Results – Heatmap Example 1




- This is showing all our candidates from this experiment.
- Good to see large trends
- Too many to see individual gene names

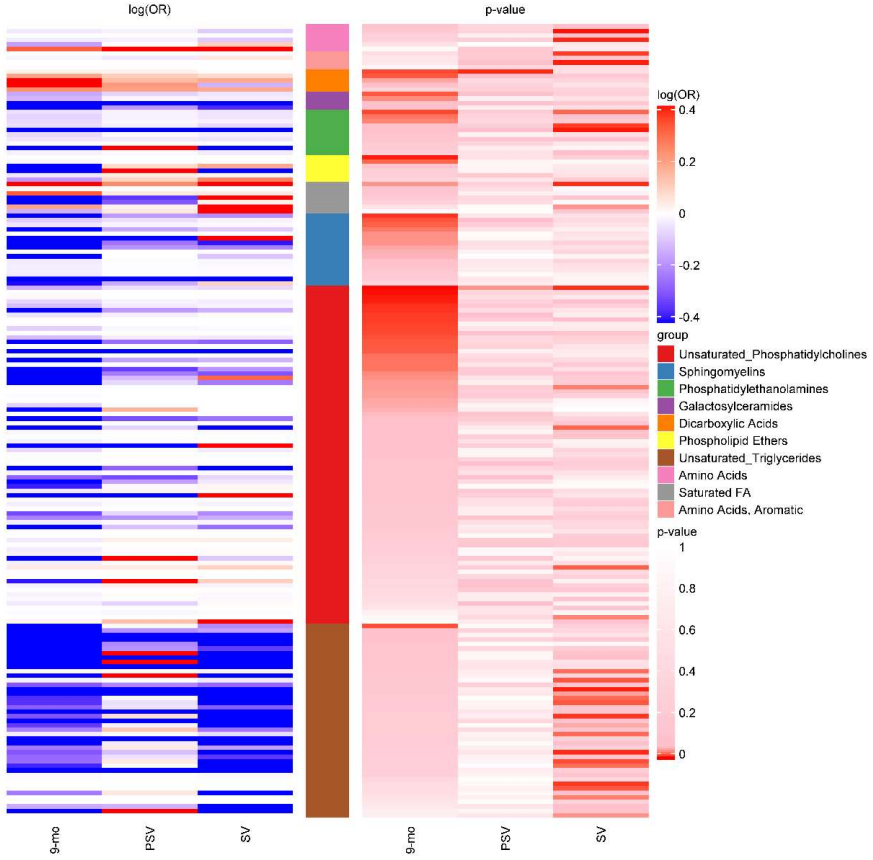
# Visualizing Results – Heatmap Example 2




- Investigator interested in this specific group of 30 genes
- Want to see the overall trend within them

```
zscore = t(apply(expr.want, 1, function(a) (a - mean(a))/sd(a)))
my_palette <- colorRampPalette(c("blue", "white", "red"))(n = 299)
heatmap.2(zscore, dendrogram="both", trace="none", col=my_palette, labRow=FALSE, margins = c(7, 5))
```

# Visualizing Results – Heatmap Example 2




complexHeatmap package in R

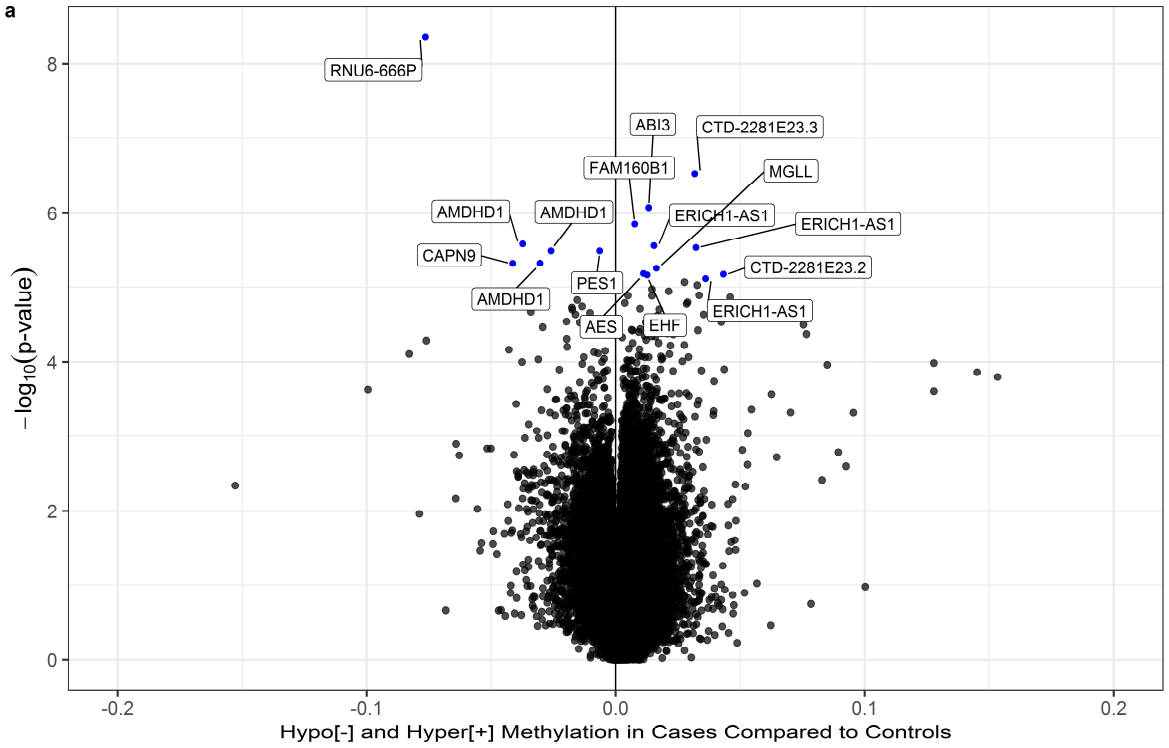
Allows you to add multiple heatmaps in 1 figure

Easy group coloring

# complexHeatmap Code

```
########## Make Figure #################
colors=c(brewer.pal(9, "Set1"), brewer.pal(9, "Paired"))
#OR.sig.forHeat is a matrix of OR estimates where each row is a feature & each column a time point
or_heat = Heatmap(as.matrix(log(OR.sig.forHeat)), name="log(OR)",
        row_title_gp = gpar(fontsize=8),
        row_order = c(1:nrow(OR.sig.forHeat)),
        cluster_columns = FALSE,
        cluster_rows = FALSE,
        show_column_dend = FALSE,
        show_row_dend = TRUE,
        show_row_names = FALSE,
        row_names_gp = gpar(fontsize = 8),
        col = colorRamp2(c(log(1/1.5),0,log(1.5)), c("blue", "white", "red")),
        column_names_gp = gpar(fontsize=8),
        heatmap_legend_param = list(color_bar = "continuous",
          legend_height = unit(5, "cm"),
          title_gp=gpar(fontsize=8)),
          column_title = "log(OR)"
        )

ha_row = rowAnnotation(df = data.frame(group = OR.sig$ClusterLabel),
    col = list(group = c("Unsaturated_Phosphatidylcholines" = colors[1], "Sphingomyelins" = colors[2], "Phosphatidylethanolamines"
= colors[3], "Galactosylceramides" = colors[4], "Dicarboxylic Acids" = colors[5], "Phospholipid Ethers" = colors[6],
"Unsaturated_Triglycerides"=colors[7], "Amino Acids"=colors[8], "Saturated FA" = colors[9], "Amino Acids, Aromatic"=colors[14])),
width = unit(1, "cm"))

pval_heat = Heatmap(as.matrix(pval.sig.forHeat), name="p-value",
        row_title_gp = gpar(fontsize=8),
        row_order = c(1:nrow(OR.sig.forHeat)),
        cluster_columns = FALSE,
        cluster_rows = FALSE,
        show_column_dend = FALSE,
        show_row_dend = TRUE,
        show_row_names = FALSE,
        row_names_gp = gpar(fontsize = 8),
        col = colorRamp2(c(0, 0.05, 1), c("darkgreen", "darkolivegreen2", "white")),
        column_names_gp = gpar(fontsize=8),
        heatmap_legend_param = list(color_bar = "continuous",
          legend_height = unit(5, "cm"),
          title_gp=gpar(fontsize=8)),
          column_title = "p-value"
 )
ht_list = or_heat + ha_row + pval_heat
draw(ht_list)

png(file="Y:/LaurenV_random/Norris/data/metabolomicsQC/TEDDY/heatmap.20190227.png", res=600, width=480*10, height=480*10)
draw(ht_list)
dev.off()
```

# Visualizing Results – Volcano Plot




- Here you are plotting an effect size on x-axis & -log10(p-value) on y-axis
- Good if you have an interpretable effect size
  - Fold Change
  - Odds Ratio
  - Difference between 2 groups
- Good to visualize both effect size and significance

# Volcano Plot Code

```
library(ggpubr)

library(dplyr)
library(ggplot2) # best plotting package
library(ggrepel) # ggplot2 addon for better labels
```

```
pdf(file="Y:/LaurenV_random/Norris/data/T1Dpaper/sesameModeling/longVolcanoPlot.pdf")
volcplot <- ggplot(as.data.frame(subset(MetaResults.withAnno, (candidate == 0))), aes(x=Beta.diff.meta, y=log10pval))
 +
  geom_point(size = 2, alpha=0.7, na.rm=T) +
  geom_point(data=as.data.frame(subset(MetaResults.withAnno, (candidate == 1))), aes(x=Beta.diff.meta, y=log10pval),
  shape=16, size = 2, alpha=1, na.rm=T, color="blue", fill="blue") +
  theme_bw(base_size=16) +
  theme(legend.title=element_blank()) +
  theme(legend.position = "none") +
  theme(text=element_text(family="sans")) +
  xlab("Hypo[-] and Hyper[+] Methylation in Cases Compared to Controls") +
  ylab(expression(-log[10]("p-value"))) +
  geom_vline(xintercept = 0, colour = "black") +
  scale_colour_gradient(low = "black", high = "black", guide = FALSE) +
  scale_x_continuous(limits = c(-0.2, 0.2)) +
  geom_label_repel(data = subset(MetaResults.withAnno, (candidate == 1)),
                   aes(label = closestGene.gene_name),
                   family = "sans",
                   color = "black",
                   box.padding = unit(0.5, "lines"),
                   point.padding = unit(0.5, "lines"),
                   force = 1)
volcplot
dev.off()
```

Thank you Randi Johnson for the ggrepel code!

# Enrichment Analysis

- Aka Pathway analysis, over-representation analysis, network analysis
- Are my candidates a part of any known biological pathway or system?
- Databases:
  - Gene Ontology (GO)
  - KEGG Pathways
  - Panther Pathways
  - DSigDB (drug database)



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

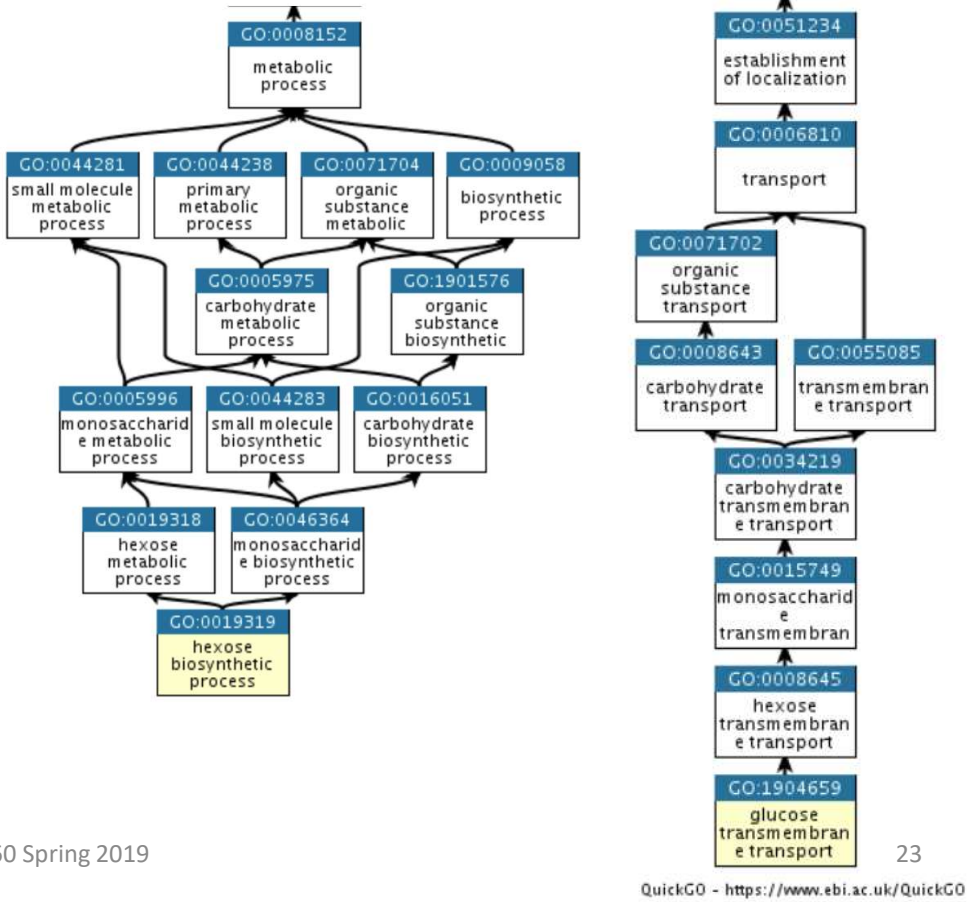

# Enrichment Methods

- Static
  - 2x2 table of candidates and background genes
- Fluid
  - Gene Set Enrichment Analysis (GSEA)
  - include p-values and effect sizes and move along different cut-off points of significance to determine enrichment

| | **Candidates** | **Genome (background)** |
|---|---|---|
| **In Pathway** | | |
| **Not in Pathway** | | |

# Gene Ontology Hierarchy

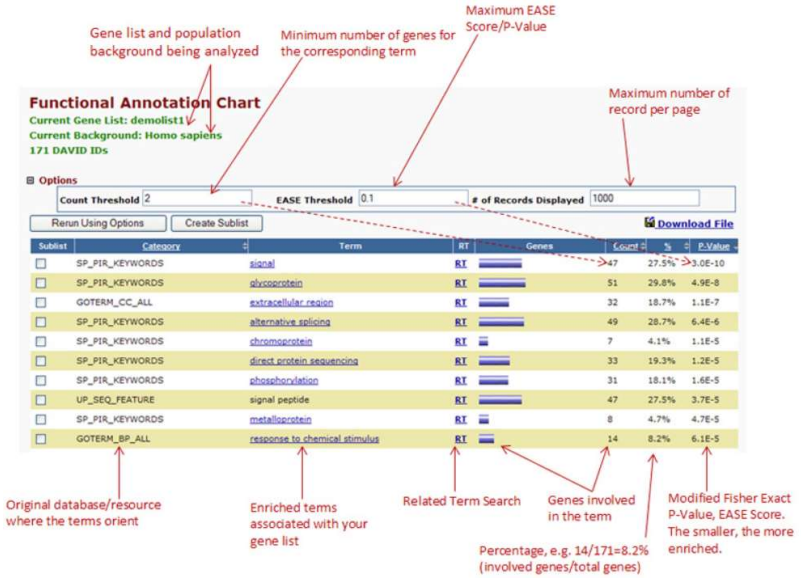
- 3 main roots
1. cellular component
2. biological process
3. molecular function
- Parent terms are those higher up
- Child terms are the most specific
- Many times report most specific child term possible



# DAVID

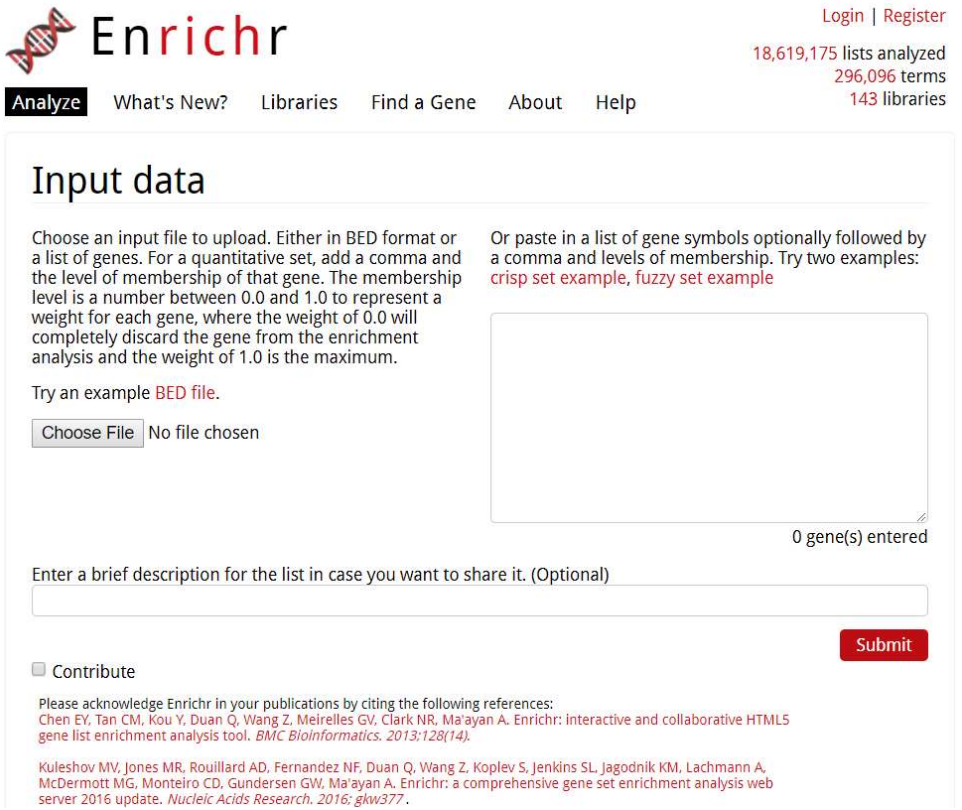


- One of the first tools
- No longer maintains databases ☹



# EnrichR

- Put in a gene-list
  - can use their web interface or within R
- Stats: Z-score permutation background correction on a Fisher's Exact Test
- Pro: easy, good databases
- Con: only human (R code) and doesn't allow you to upload a unique background




FlyEnrichr YeastEnrichr WormEnrichr FishEnrichr

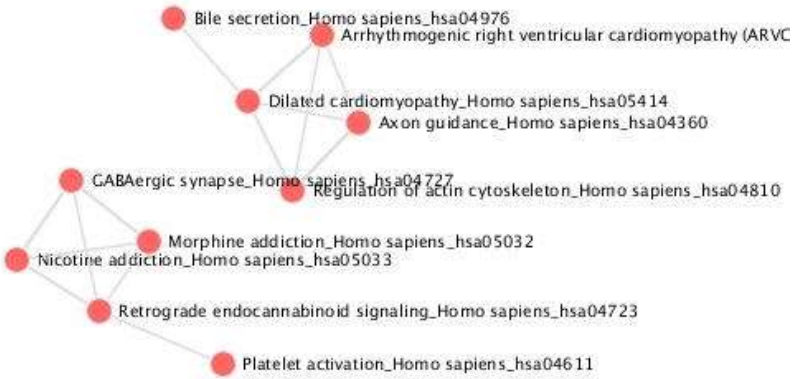
# EnrichR

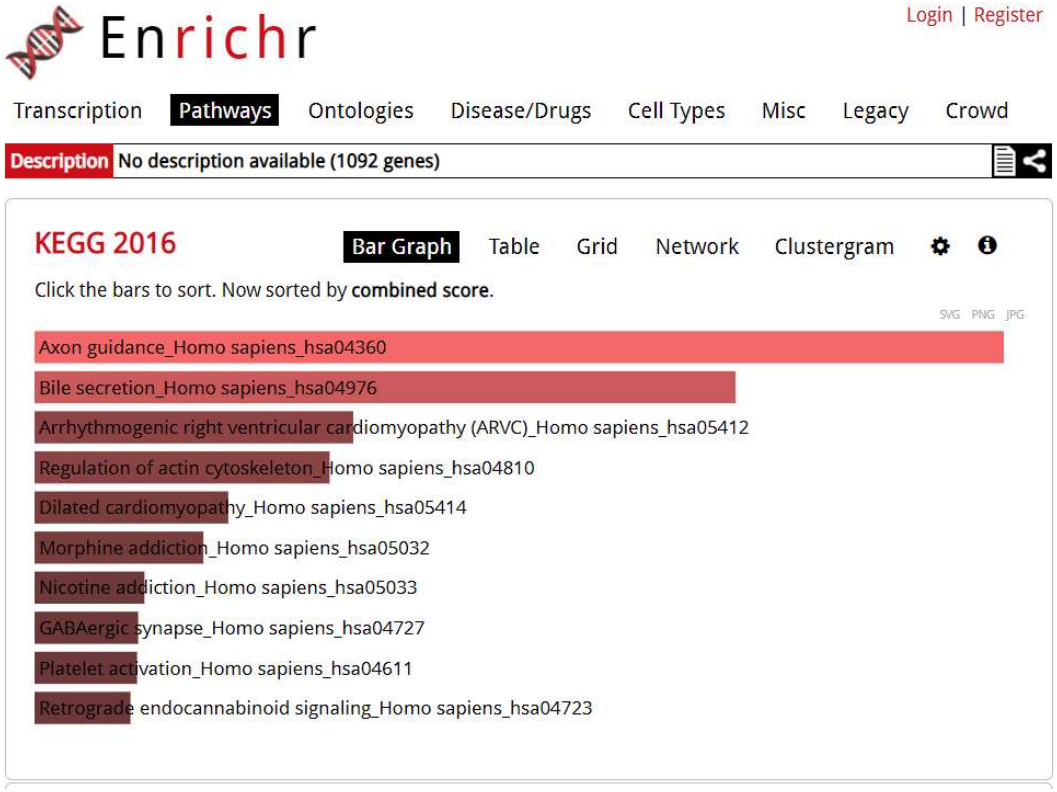
```
dbs <- c("GO_Molecular_Function_2017b", "GO_Biological_Process_2017b", "KEGG_2016")
bin1.enrich <- enrichr(bin1.humanGeneSymbol, dbs)
```

```
> summary(bin1.enrich)
                          Length Class      Mode
GO_Molecular_Function_2017b 9      data.frame list
GO_Biological_Process_2017b 9      data.frame list
KEGG_2016                 9      data.frame list
> head(bin1.enrich$GO_Biological_Process_2017b)
                                                        Term Overlap      P.value Adjusted.P.value  Old.P.value
1                          ion transmembrane transport (GO:0034220)  47/372 7.371718e-08     1.639470e-04 9.940321e-06
2                     ephrin receptor signaling pathway (GO:0048013)  30/151 6.040388e-10     4.030147e-06 1.634416e-07
3                                  axon guidance (GO:0007411)  29/182 1.963853e-07     1.871832e-04 1.263547e-05
4 enzyme active site formation via O4'-phospho-L-tyrosine (GO:0018334)   15/46 1.074412e-08     3.584239e-05 1.393328e-06
5                        anion transmembrane transport (GO:0098656)  30/194 2.382866e-07     1.987310e-04 1.505995e-05
6                       cation transmembrane transport (GO:0098655)  30/200 4.671549e-07     2.703373e-04 2.524068e-05
  Old.Adjusted.P.value   Z.score Combined.Score
1          0.014914636 -4.478661       73.55318
2          0.001090482 -3.359573       71.31495
3          0.014914636 -3.711552       57.31819
4          0.004648142 -2.967579       54.45184
5          0.014914636 -3.442261       52.49376
6          0.015309621 -3.528267       51.43015


                                          Genes
1 ABCB1;ABCB7;ABCB4;AQP8;ABCB5;CLCNKB;SLC5A12;CLCNKA;NEDD4L;CLCN2;CLCN1;SLC5A5;LASP1;PSMB2;ANO10;SLC26A11;ATP6V0E2;GABRA2;SLC36A1;SLC12A2
;SLC12A3;GABRA1;ABCC4;SLC36A2;ATP6V0E1;GABRA6;GABRA5;GABRA4;ABCA3;SLC12A1;GABRA3;ATP1B3;ANO7;ATP1B2;ATP1B1;GABRG3;GABRG2;ATP4B;PSMA6;SLC5
A8;PEX3;ATP13A4;ATP13A5;ATP13A2;ABCG4;KCNK3;ABCG1
2
                      BLK;EPHB6;SRC;MYL12A;EFNB2;EFNB1;NCSTN;EFNB3;FYN;EPHB2;EPHB1;EPHB4;GIT1;AP2M1;EPHB3;LYN;EPHA5;PSENEN;EPHA4;EPHA7;YES
1;EPHA6;EPHA8;PTPN11;PTK2;FGR;HCK;LCK;RASA1;EPHA3
3
                   SRC;NTN1;NTN3;MAPK1;NCAM1;FYN;EPHB2;EPHB1;MAPK3;EPHB3;NTNG1;EPHA5;NTNG2;LINGO1;LINGO4;SEMA6A;LRRN2;EPHA8;RYK;SIAH2;SIAH
1;WNT7A;PTPN11;PTK2;LRFN3;LRFN2;LRFN4;LRFN1;FGFR2
4
```

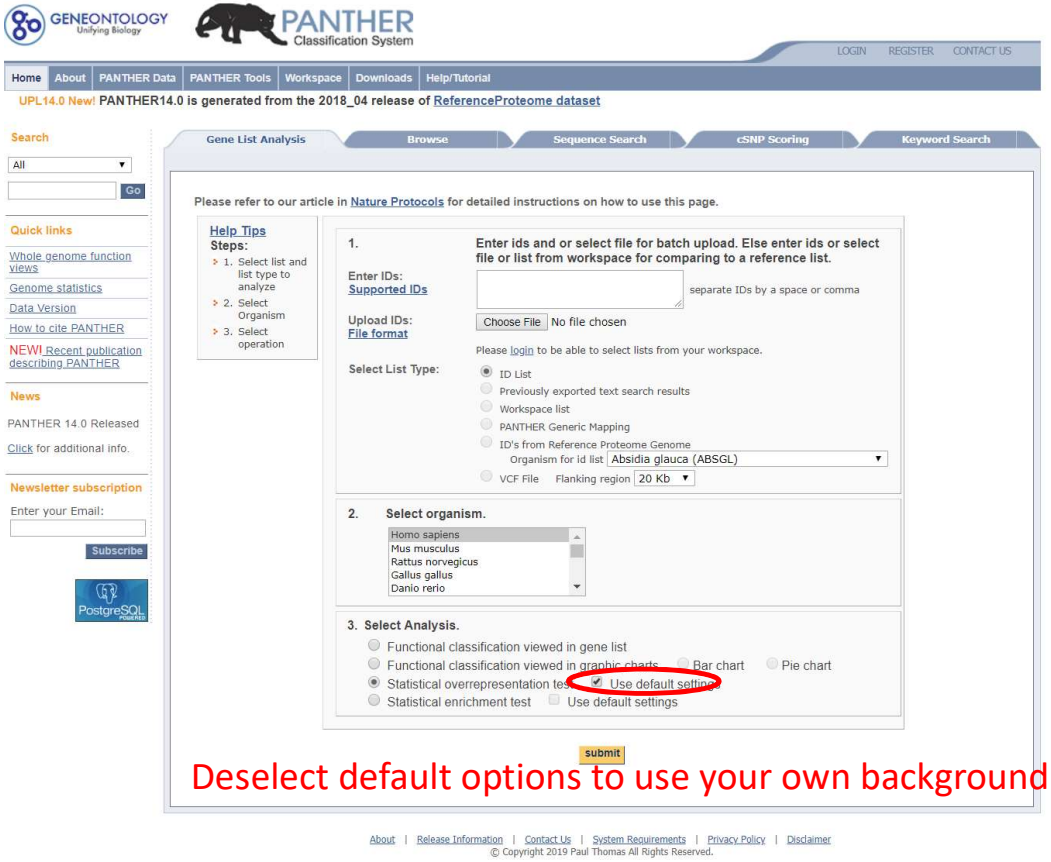
# EnrichR Web

- The network shown is based on enriched pathways
- Shows what genes are overlapping what pathways

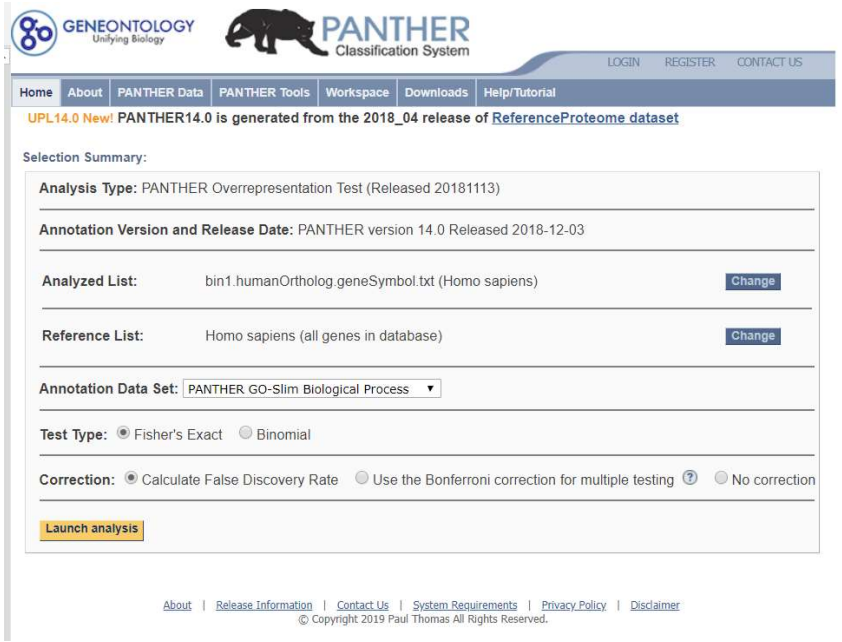


# Panther



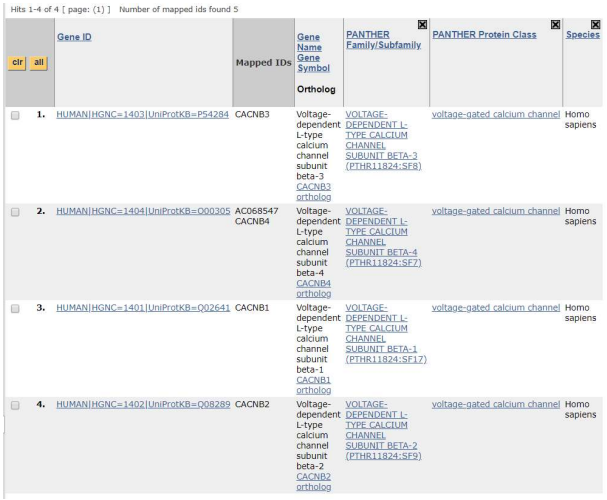

Deselect default options to use your own background




- Pro: Can define your own background
- Con: Not available in R (or any other language)

# Panther Output

- Like how it indents and color codes based on GO hierarchy
- Child is on top



Hits 1-4 of 4 [ page: (1) ] Number of mapped ids found 5

| clr all | Gene ID | Mapped IDs | Gene Name Gene Symbol Ortholog | PANTHER Family/Subfamily | PANTHER Protein Class | Species |
|---|---|---|---|---|---|---|
| 1. | HUMAN\|HGNC=1403\|UniProtKB=P54284 | CACNB3 | Voltage-dependent L-type calcium channel subunit beta-3 CACNB3 ortholog | VOLTAGE-DEPENDENT L-TYPE CALCIUM CHANNEL SUBUNIT BETA-3 (PTHR11824:SF8) | voltage-gated calcium channel | Homo sapiens |
| 2. | HUMAN\|HGNC=1404\|UniProtKB=O00305 | AC068547 CACNB4 | Voltage-dependent L-type calcium channel subunit beta-4 CACNB4 ortholog | VOLTAGE-DEPENDENT L-TYPE CALCIUM CHANNEL SUBUNIT BETA-4 (PTHR11824:SF7) | voltage-gated calcium channel | Homo sapiens |
| 3. | HUMAN\|HGNC=1401\|UniProtKB=Q02641 | CACNB1 | Voltage-dependent L-type calcium channel subunit beta-1 CACNB1 ortholog | VOLTAGE-DEPENDENT L-TYPE CALCIUM CHANNEL SUBUNIT BETA-1 (PTHR11824:SF17) | voltage-gated calcium channel | Homo sapiens |
| 4. | HUMAN\|HGNC=1402\|UniProtKB=Q08289 | CACNB2 | Voltage-dependent L-type calcium channel subunit beta-2 CACNB2 ortholog | VOLTAGE-DEPENDENT L-TYPE CALCIUM CHANNEL SUBUNIT BETA-2 (PTHR11824:SF9) | voltage-gated calcium channel | Homo sapiens |

| | Reference list | bin1.humanOrtholog.geneSymbol.txt |
|---|---|---|
| Mapped IDs: | 20996 out of 20996 | 1087 out of 1154 |
| Unmapped IDs: | 0 | 5 |
| Multiple mapping information: | 0 | 86 |

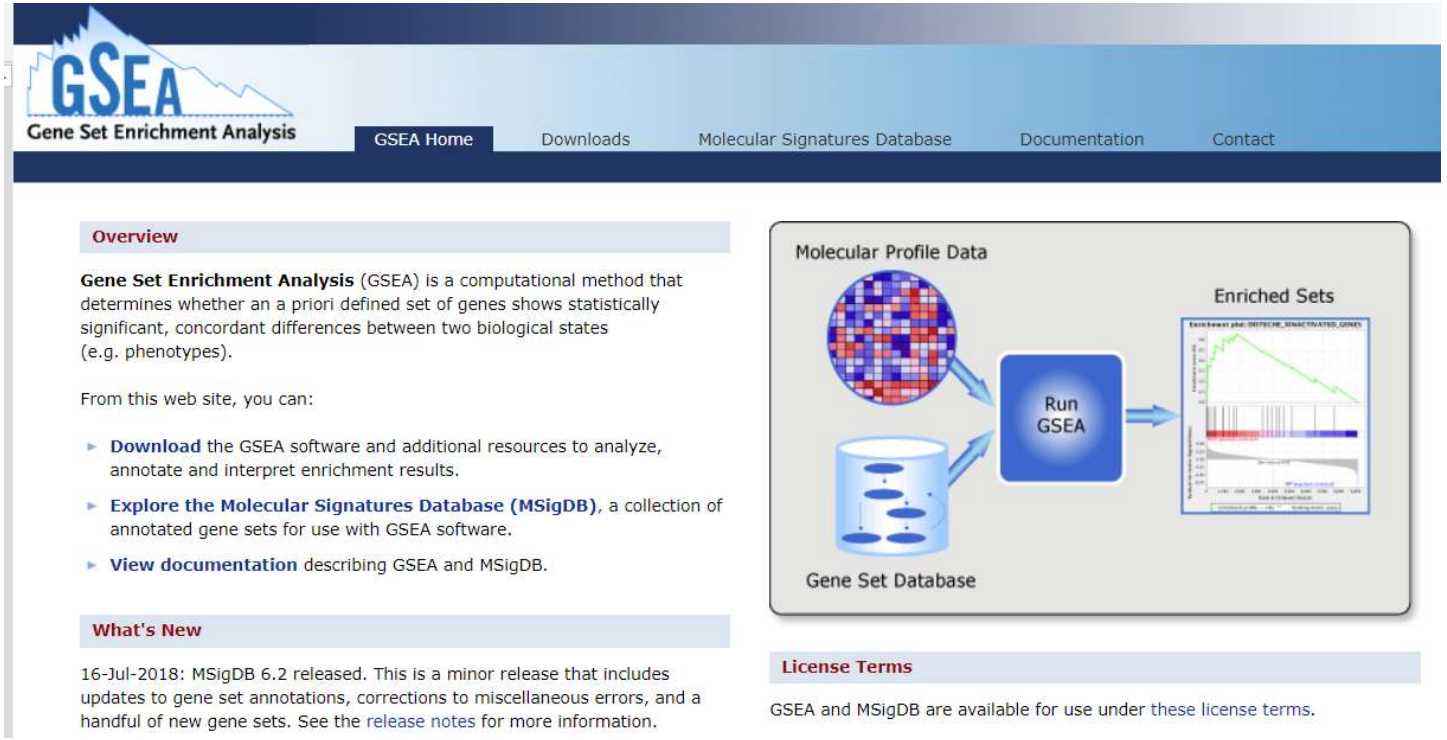
Export results View: -- Please select a chart to display --

Displaying only results for FDR P < 0.05, click here to display all results

| | Homo sapiens (REF) | bin1.humanOrtholog.geneSymbol.txt (▽ Hierarchy NEW!) | | | | | |
|---|---|---|---|---|---|---|---|
| PANTHER GO-Slim Biological Process | # | # | expected | Fold Enrichment | +/- | raw P value | FDR |
| multi-organism cellular process | 3 | 3 | .16 | 18.19 | + | 2.50E-03 | 3.96E-02 |
| regulation of calcium ion transmembrane transport | 6 | 4 | .33 | 12.13 | + | 1.19E-03 | 2.32E-02 |
| ↳biological regulation | 4097 | 274 | 225.18 | 1.22 | + | 6.20E-04 | 1.40E-02 |
| regulation of dendrite morphogenesis | 8 | 5 | .44 | 11.37 | + | 3.45E-04 | 8.55E-03 |
| ↳regulation of dendrite development | 13 | 5 | .71 | 7.00 | + | 1.84E-03 | 3.14E-02 |
| ↳regulation of cellular component organization | 274 | 31 | 15.06 | 2.06 | + | 3.79E-04 | 9.27E-03 |
| ↳regulation of cellular process | 2460 | 187 | 135.21 | 1.38 | + | 1.12E-05 | 7.69E-04 |
| ↳cell differentiation | 350 | 37 | 19.24 | 1.92 | + | 4.47E-04 | 1.08E-02 |
| ↳cellular developmental process | 506 | 46 | 27.81 | 1.65 | + | 1.78E-03 | 3.08E-02 |
| ↳developmental process | 1035 | 90 | 56.89 | 1.58 | + | 4.53E-05 | 2.19E-03 |
| ↳system development | 85 | 13 | 4.67 | 2.78 | + | 1.70E-03 | 3.04E-02 |
| ↳multicellular organismal process | 695 | 87 | 38.20 | 2.28 | + | 2.12E-11 | 9.49E-09 |
| cellular response to light stimulus | 17 | 10 | .93 | 10.70 | + | 5.30E-07 | 9.47E-05 |
| ↳cellular response to radiation | 24 | 10 | 1.32 | 7.58 | + | 5.90E-06 | 4.58E-04 |
| ↳response to abiotic stimulus | 87 | 13 | 4.78 | 2.72 | + | 2.05E-03 | 3.42E-02 |
| ↳response to stimulus | 1479 | 54 | 81.29 | .66 | - | 1.55E-03 | 2.83E-02 |
| ↳cellular response to abiotic stimulus | 28 | 10 | 1.54 | 6.50 | + | 1.76E-05 | 1.12E-03 |
| response to nutrient | 7 | 4 | .38 | 10.40 | + | 1.80E-03 | 3.09E-02 |
| septin ring organization | 17 | 9 | .93 | 9.63 | + | 3.83E-06 | 3.60E-04 |
| ↳septin cytoskeleton organization | 17 | 9 | .93 | 9.63 | + | 3.83E-06 | 3.42E-04 |
| embryonic organ development | 12 | 6 | .66 | 9.10 | + | 2.13E-04 | 5.86E-03 |
| ↳animal organ development | 112 | 22 | 6.16 | 3.57 | + | 1.62E-06 | 1.93E-04 |
| myofibril assembly | 20 | 10 | 1.10 | 9.10 | + | 1.64E-06 | 1.83E-04 |
| ↳actomyosin structure organization | 36 | 12 | 1.98 | 6.06 | + | 4.60E-06 | 3.73E-04 |
| ↳actin cytoskeleton organization | 89 | 20 | 4.89 | 4.09 | + | 7.91E-07 | 1.29E-04 |
| ↳actin filament-based process | 99 | 21 | 5.44 | 3.86 | + | 9.54E-07 | 1.31E-04 |

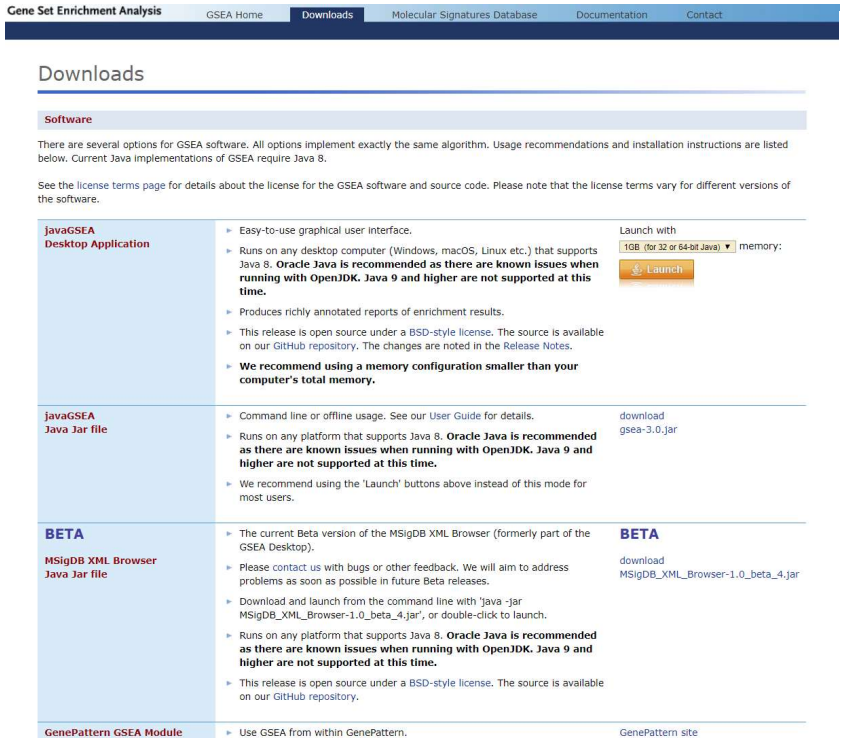
# GSEA

- Very popular
- Order your results and go along to find what level pathways are enriched
- Not easy to use in R




http://software.broadinstitute.org/gsea/index.jsp

# GSEA Options



Gene Set Enrichment Analysis | GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

## Downloads

**Software**

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 8.

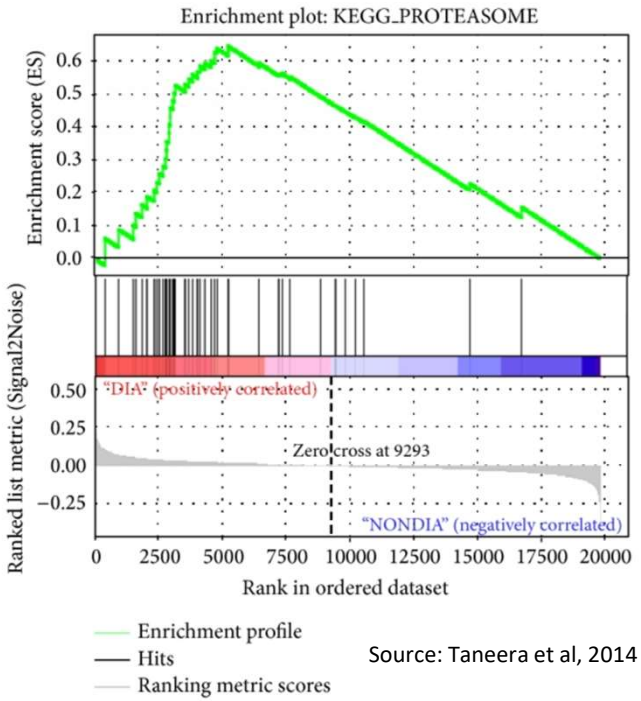
See the license terms page for details about the license for the GSEA software and source code. Please note that the license terms vary for different versions of the software.

| | | |
|---|---|---|
| **javaGSEA Desktop Application** | ▸ Easy-to-use graphical user interface.<br>▸ Runs on any desktop computer (Windows, macOS, Linux etc.) that supports Java 8. **Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.**<br>▸ Produces richly annotated reports of enrichment results.<br>▸ This release is open source under a BSD-style license. The source is available on our GitHub repository. The changes are noted in the Release Notes.<br>▸ **We recommend using a memory configuration smaller than your computer's total memory.** | Launch with 1GB (for 32 or 64-bit Java) memory: Launch |
| **javaGSEA Java Jar file** | ▸ Command line or offline usage. See our User Guide for details.<br>▸ Runs on any platform that supports Java 8. **Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.**<br>▸ We recommend using the 'Launch' buttons above instead of this mode for most users. | download gsea-3.0.jar |
| **BETA**<br>**MSigDB XML Browser Java Jar file** | ▸ The current Beta version of the MSigDB XML Browser (formerly part of the GSEA Desktop).<br>▸ Please contact us with bugs or other feedback. We will aim to address problems as soon as possible in future Beta releases.<br>▸ Download and launch from the command line with 'java -jar MSigDB_XML_Browser-1.0_beta_4.jar', or double-click to launch.<br>▸ Runs on any platform that supports Java 8. **Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.**<br>▸ This release is open source under a BSD-style license. The source is available on our GitHub repository. | **BETA** download MSigDB_XML_Browser-1.0_beta_4.jar |
| **GenePattern GSEA Module** | ▸ Use GSEA from within GenePattern. | GenePattern site |

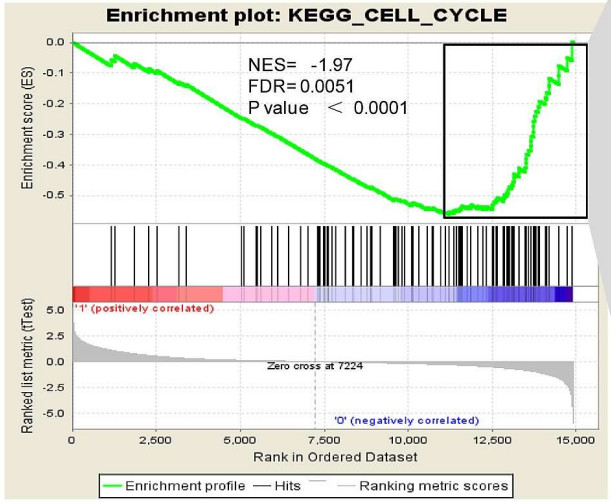
| | | |
|---|---|---|
| **GenePattern GSEA Module** | ▸ Use GSEA from within GenePattern.<br>▸ Use GSEA in concert with a large suite of other analytics found in GenePattern (a powerful and flexible analysis platform developed at the Broad Institute). | GenePattern site |
| **R-GSEA R Script** | ▸ Usage from within the R programming environment.<br>▸ **Note that this script has not been updated since 2005 and may not work as-is with modern R distributions. It is provided for reference purposes only and is no longer being maintained or supported.**<br>▸ Click here to learn more about the R-GSEA script | download GSEA-P-R.1.0.zip |
| **javaGSEA v2.2.4 Java Jar file** | ▸ This is the last release of the GSEA Desktop based on the older version of the code, made available for reference purposes. **Note that this version is no longer being maintained or supported** (as of the GSEA v3.0 release, July 2017). Note that the core GSEA algorithm has not changed.<br>▸ Command line or offline usage. See our User Guide for details.<br>▸ Runs on any platform that supports Java 7 or 8. Java 8 is recommended. **Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.** | download gsea2-2.2.4.jar |

I would use the Desktop application
Does depend on java

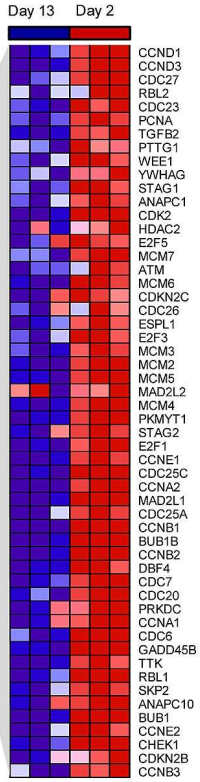
# GSEA output



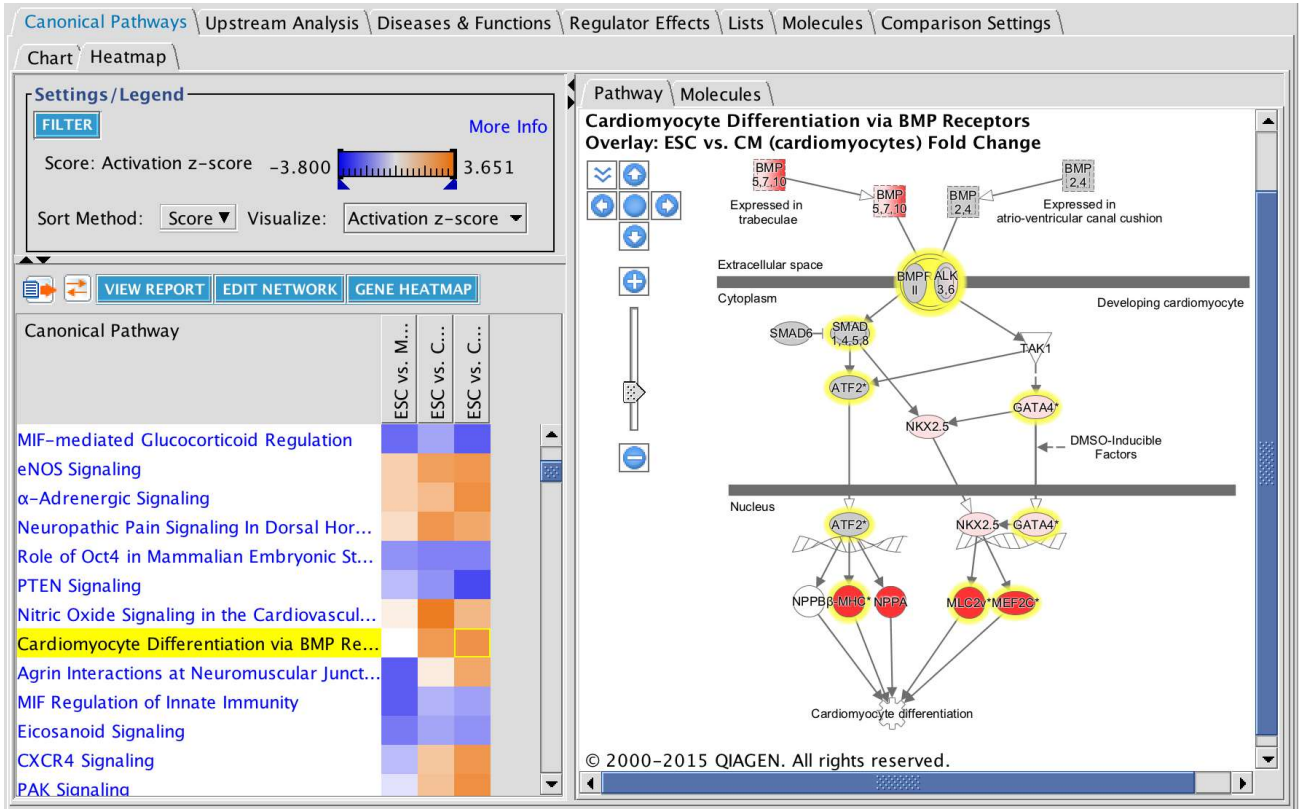

Source: Taneera et al, 2014




Source: Broad Institute



# Ingenuity Pathway Analysis (IPA)




Source: Qiagen

Major Con: Costs a lot of money

# Transformations

- What if you need to do a more complex model and need to take your data outside of the DESeq2 or EDASeq packages?
  - Mixed modeling
  - Network analysis (WGCNA)
- NOTE: YOU CARE IF RNA-EXPRESSION IS OUTCOME IN MODEL
- Types of transformations:
  - Voom
  - VST
  - Regularized log (rlog)
  - Log2(counts + 1)

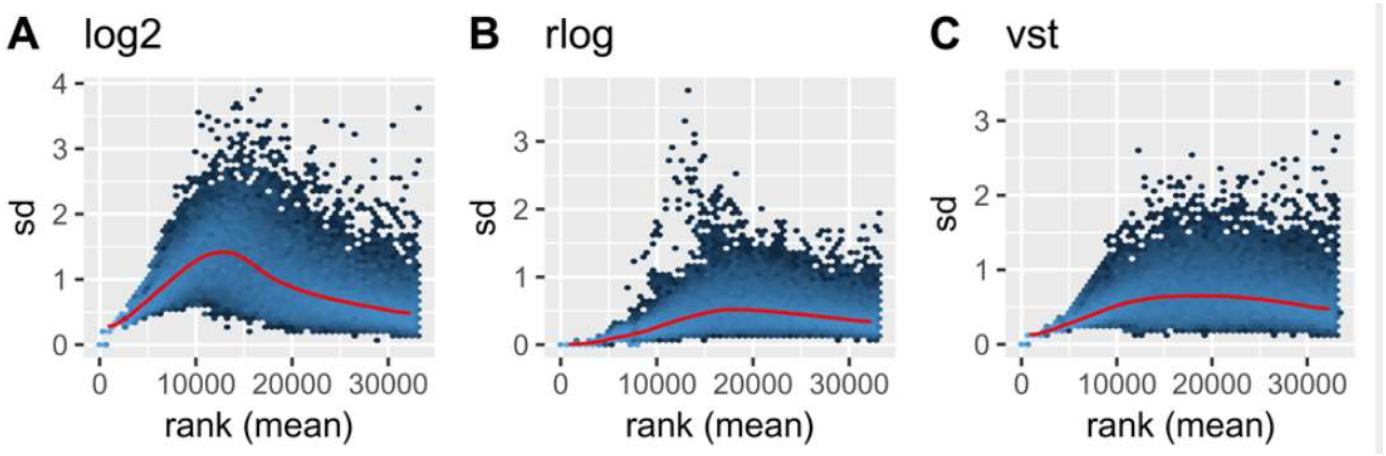
Specifically made for sequencing data

# Voom

- Mean-variance modelling at the observational level
- limma package in R
  - Go-to package for microarrays
- Trying to get data to look like microarray data so you can use their same pipeline
- log(CPM + 0.5)
- “The lmFit function is used to fit row-wise linear models. The lowess function is then used to fit a trend to the square-root-standard-deviations as a function of average logCPM. The trend line is then used to predict the variance of each logCPM value as a function of its fitted value, and the inverse variances become the estimated precision weights.” – limma user manual
- VERY BAD FOR LOW COUNTS, NEED LOTS OF DABG FILTERING

# VST & rlog

- Variance Stabilizing Transformation & regularized log transformation
- Fit a dispersion-mean relationship and then transforms the counts
- Goal is to: 1. reduced extreme outliers and 2. have constant variance along the range of mean values



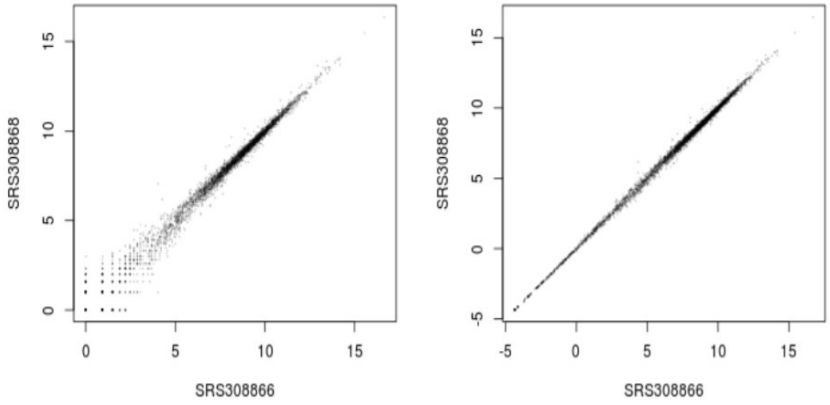

Source: Steinbaugh et al, 2017

# VST & rlog

- Available through DESeq2
- rlog is the newer version
- Transforms the count data to the log2 scale in a way which minimizes differences between samples for rows with small counts, and which normalizes with respect to library size
- rlog is more robust in the case when the size factors vary widely

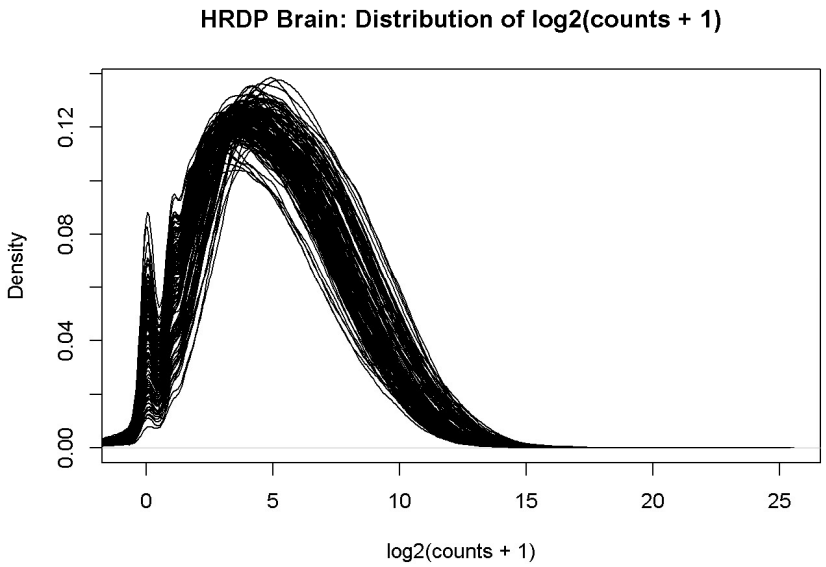
- Main point of rlog is to remove the dependence of the variance on the mean, particularly the high variance of the logarithm of count data when the mean is low.

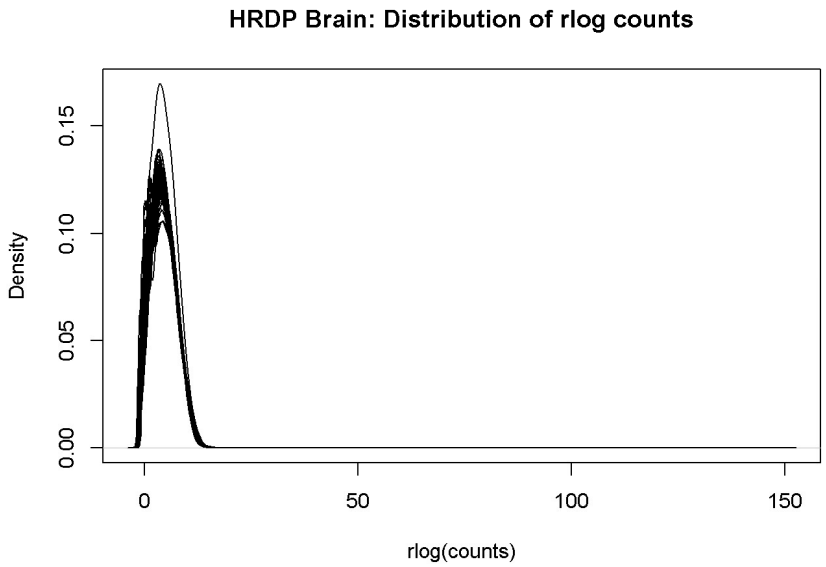
log2 rlog




source: Beginners Guide To DESeq2

# rlog distribution




Recommend using rlog BUT this is a very computationally intensive step if you have lots of samples
VST runs really fast

# References


Bonferroni, C. E., (1936) Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze

Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Journal of the Royal Statistical Society, Series B. 57 (1): 289–300.

Taneera J, Storm P, Groop L. *Downregulation of type II diabetes mellitus and maturity onset diabetes of young pathways in human pancreatic islets from hyperglycemic donors.* J Diabetes Res. 2014;2014:237535.

Michael J. Steinbaugh , Lorena Pantano , Rory D. Kirchner , Victor Barrera , Brad A. Chapman , Mary E. Piper , Meeta Mistry , Radhika S. Khetani , Kayleigh D. Rutherford , Oliver Hofmann , John N. Hutchinson , Shannan Ho Sui. *bcbioRNASeq: R package for bcbio RNA-seq analysis* F1000Research 2017, 6:1976.

https://seqqc.wordpress.com/2015/02/16/should-you-transform-rna-seq-data-log-vst-voom/