Documentation: DESeq2 Differential Expression Analysis Output Table

The output table from DESeq2 provides detailed information about differentially expressed genes (DEGs) based on the statistical analysis of RNA-seq data. Below is an explanation of the key columns in the output table:

Key Columns in the DESeq2 Output Table (***LogRatios_*.tsv*** and its subset ***SignifDiffExpr_*.tsv*** in which are only the genes that have adjusted p-value <0.05)

1. `baseMean`:
    - **Description**: The average normalized counts across all samples, representing the overall expression level of the gene.
    - **Use**: Helps identify highly or lowly expressed genes.
    - Consider genes with BaseMean <50 as genes with low expression
2. `log2FoldChange`:
    - **Description**: The log2-transformed fold change in expression between two conditions (e.g., treatment vs. control).
    - **Use**: Indicates the magnitude and direction of change:
        - **Positive values**: Upregulation in the treatment group.
        - **Negative values**: Downregulation in the treatment group.
3. `lfcSE`:
    - **Description**: The standard error of the `log2FoldChange`.
    - **Use**: Provides an estimate of the uncertainty in the fold change.
4. `stat`:
    - **Description**: The Wald test statistic, calculated as `log2FoldChange / lfcSE`.
    - **Use**: Used to compute the p-value.
5. `pvalue`:
    - **Description**: The raw p-value from the Wald test, indicating the statistical significance of the differential expression.
    - **Use**: Identifies statistically significant DEGs (typically < 0.05).

6. `padj`:
    - **Description**: The adjusted p-value (e.g., Benjamini-Hochberg correction) to control for multiple testing.
    - **Use**: Reduces false positives; commonly used threshold is `padj < 0.05`.
7. `gene` **(optional)**:
    - **Description**: The gene identifier (e.g., gene name, Ensembl ID, or other annotation).
    - **Use**: Links DEGs to their functional annotation.

How to Interpret the Output

- **Significant DEGs**: Genes with `padj < 0.05` are considered significant.
- **Upregulated Genes**: Genes with `log2FoldChange > 0` and `padj < 0.05`.
- **Downregulated Genes**: Genes with `log2FoldChange < 0` and `padj < 0.05`.
- **Low-Expression Genes**: Genes with low `baseMean` values (50-100) may be filtered out as they are less reliable.

Example Output Table

gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
GeneA	500.12	2.34	0.15	15.60	1.2e-10	3.6e-08
GeneB	300.45	-1.78	0.20	-8.90	5.7e-07	1.2e-05
GeneC	50.23	0.45	0.30	1.50	0.13	0.25

Documentation: Functional enrichment with Webgestalt

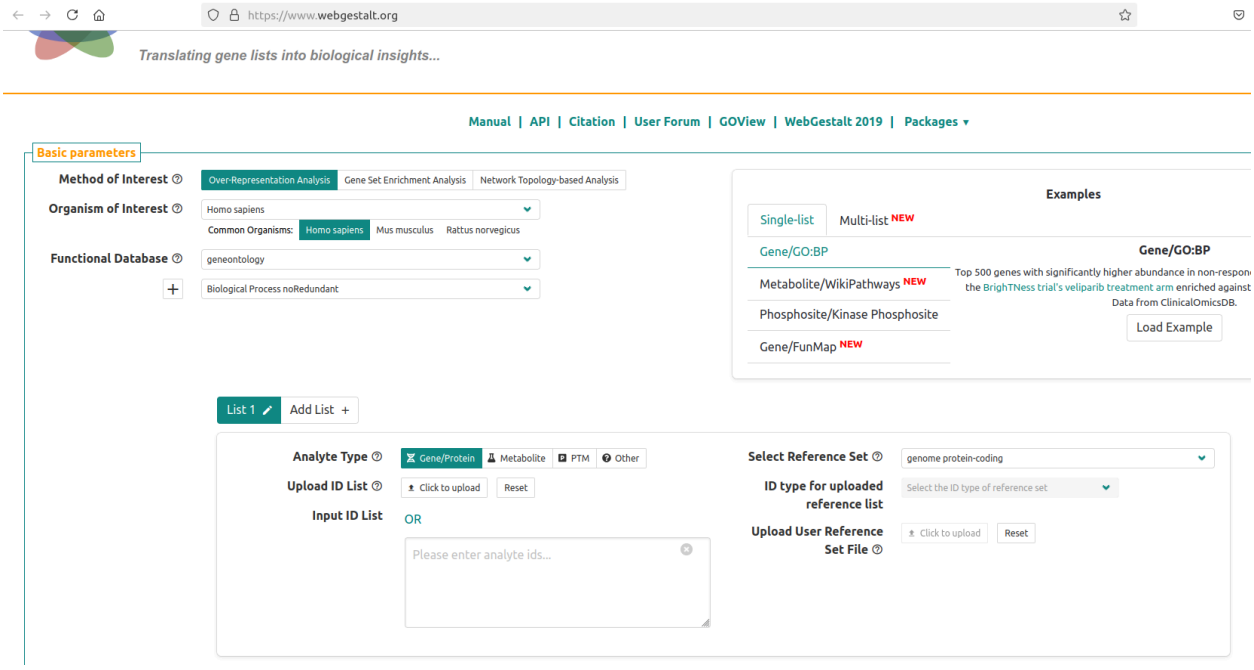
https://www.webgestalt.org/

Overrepresentation analysis

For Overrepresentation analysis, you click on the “Over Representation Analysis” as it was done below, select organism and database (most common: *geneontology -> Biological Processes nonredundant*)

In the **input ID list** box you add the ***column of the gene symbols*** in the **SignifDiffExpr_*.tsv** file.

ATTENTION!!! Unless otherwise specified, you select the genome protein_coding genes at “**Select Reference Set**”.

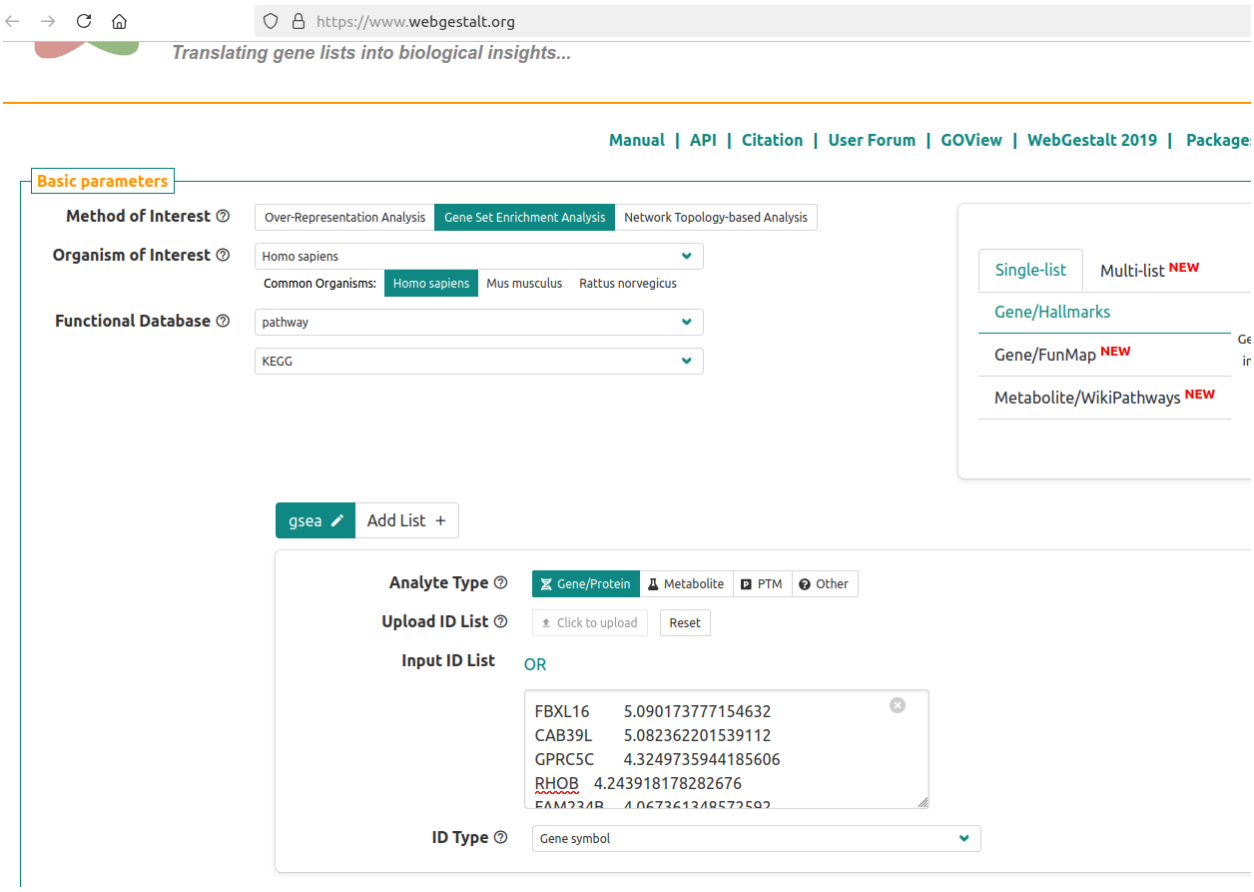


GSEA

For Gene set Enrichment analysis, you click on the “GeneSet Enrichment Analysis” as it was done below, select organism and database (most common: *geneontology -> Biological Processes nonredundant*).

Instructions for Preparing the Input File:

1. **Open the `LogRatios_*.tsv` File**:
    - Use **Excel** or **Google Sheets** (recommended, especially if the `log2FoldChange` values appear unusual, e.g., with multiple dots like `-3.989.094.943`).
2. **Sort the Data**:
    - Sort the `log2FoldChange` column in **descending order**, from the **most positive** to the **most negative** values.
    -
3. **Prepare the Input for WebGestalt**:
    - Use the **gene name** and `log2FoldChange` columns from the `LogRatios_*.tsv` file as the input for WebGestalt.
    - Ensure the format matches the example shown in the figure below.



Documentation: Additional files and explanations

Results folder contains:

Dea (differential expression analysis)

- **LogRatios and SignifDiffExpr files**: These files contain the results of differential expression analysis from DESeq2 for each comparison described above. The filenames indicate the reference condition—for example, *SignifDiffExpr_CON_PAT.tsv* means that a gene with a positive log2 fold change (log2FC) is upregulated in the Control sample group.

- **Table_input_gene_counts.tsv**: This file contains the raw gene count data used as input for DESeq2. The counts for each sample were generated using the *featureCounts* function from the Subread tool.

- **library_normalized_gene_counts.tsv**: This file contains gene count data normalized using DESeq2's size factors to account for differences in library sizes.

- **Plots:**
  - **PCA_plot** of the samples
  - **Volcanoplot** of DEGs (differentially expressed genes) for every comparison
  - **Heatmap** of DEGs for every comparison (**verify the sample grouping in the dendrogram!**)
  - **plots_deg folder:** violin plot for every DEGs, viewed for all the sample groups

Report_align_count.html to visualize the general statistics of the mapped reads (% Aligned, %aAssigned to feature)

Considerations for experimental validations of your DEGs

- **Prioritize DEGs with higher fold changes**: Small log2 fold changes (log2FC) may not be reliably detected by qPCR. Consider validating genes with **log2FC > 1**.
- **Consider expression levels**: Genes with very low expression across conditions (e.g., **BaseMean < 100**) may be more prone to variability and harder to validate.
- **Check for biological relevance**: Select DEGs based on functional significance, pathway involvement, or known associations with the phenotype of interest.