File Usage Documentation

# Basic File Rules

- Headers are case sensitive.  
 - All files should be CSV or TSV format.  
 - All listed required columns must be present for the file to be accepted.  
 - Any ID-type columns must not have duplicate values.  
 - All sample columns must include a header that represents the sample and must start with a `.` character so that they can be identified by the application.

# File Type Summaries and Examples

## 1. Metadata File

Maps each sample to a row of variables that describe the sample.

Required Columns:  
- sample : Contains the sample names, which should match the sample names in other uploaded files. (The application automatically prepends a `.` to sample names if missing.)

Example:

|  |  |
| --- | --- |
| **sample** | **condition** |
| .sample1 | treatment |
| .sample2 | control |
| .sample3 | treatment |

## 2. Raw Count Table

A matrix of counts representing the number of times each gene occurred within each sample.

Required Columns:  
- gene\_id : Unique identifier for each gene.  
- gene\_name : Name of the gene (recommended to represent actual genes).  
- Sample columns : Contain raw count data.

Example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| gene\_id | gene\_name | .sample1 | .sample2 | .sample3 |
| G1 | NME1 | 100 | 200 | 150 |
| G2 | AK1 | 50 | 75 | 60 |
| G3 | AK6 | 400 | 410 | 420 |

## 3. Differential Expression Analysis Results Output

The output/results from a previous DESeq2 experiment. Not all the columns that are returned from DESeq2 are required to use this file. DESeq2 does not append the sample counts to the results, rather this must be done manually before uploading with the DESeq results and the normalized counts matrix given by DESeq2.

Required Columns:  
- gene\_id : Unique identifier for each gene.  
- gene\_name : Name of the gene.  
- log2FoldChange : Log2 fold change value.  
- padj : Adjusted p-value.

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| gene\_id | gene\_name | log2FoldChange | padj | .sample1 | .sample2 |
| G13 | DCK | -1.2 | 0.05 | 120 | 100 |
| G65 | ADSS2 | 2.3 | 0.01 | 210 | 230 |
| G4 | ADSL | 1.8 | 0.003 | 400 | 450 |

## 4. Abundance Data File

A matrix of normalized counts for each gene and sample to be used in the pathway analysis portion of the application. The normalized counts can be retrieved from a DESeq2 experiment and modified to the required format.

Required Columns:  
- gene\_name : Name of the gene (must represent actual genes).  
- Sample columns : Contain normalized count data.

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| gene\_name | .sample1 | .sample2 | .sample3 |
| ENPP3 | 1.2 | 1.1 | 1.3 |
| APRT | 1.6 | 1.5 | 1.4 |
| RRM2 | 2.1 | 2.2 | 2.0 |

## 5. Previous Pathway Analysis Results Output

The output/results from a previous pathfindR pathway analysis experiment. Cluster is a required column, so you must use pathfindR to cluster the results before uploading to the application.

Required Columns:  
- ID : Unique identifier for each pathway.  
- Term\_Description : Name of the pathway.  
- Fold\_Enrichment : Fold enrichment value.  
- occurrence : Integer.  
- support : Float.  
- lowest\_p : Lowest p-value.  
- highest\_p : Highest p-value.  
- non\_Signif\_Snw\_Genes : Comma-separated list of genes.  
- Up\_regulated : Comma-separated list of upregulated genes.  
- Down\_regulated : Comma-separated list of downregulated genes.  
- all\_pathway\_genes : Comma-separated list of all genes in the pathway.  
- num\_genes\_in\_path : Number of genes in the pathway.  
- Cluster : Cluster ID (integer).  
- Status : `Representative` or `Member`.

Example:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Term\_Description | Fold\_Enrichment | occurrence | support | lowest\_p | highest\_p | Cluster | Status |
| P01 | Prion Disease | 2.5 | 5 | 0.8 | 0.0001 | 0.1 | 1 | Representative |
| P02 | Huntington Disease | 1.8 | 3 | 0.5 | 0.00003 | 0.02 | 2 | Member |

CONT…

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| non\_Signif\_Snw\_Genes | Up\_regulated | Down\_regulated | all\_pathway\_genes | num\_genes\_in\_path |
| GMPS, CMPK1 | TK1, DUT | TYMS, CTPS1 | GMPS, CMPK1, TK1, DUT, TYMS, CTPS1 | 6 |
| UCK2, NT5C3B | HPRT1, IMPDH2 |  | UCK2, NT5C3B, HPRT1, IMPDH2 | 4 |

# When To Use Each File

The application can be experienced in its entirety with the use of just the two following files: Metadata and Raw Counts Table.

1. Metadata files are required with any use of the application

2. Raw Count Table is used when a new Differential Gene Expression Analysis needs to be done. The application can then directly run the results into the pathway analysis portion of the application.

The next three files are useful if an experiment has already been done outside of the application with DESeq2 or pathfindR and the user just wants access to the visuals the application provides. This allows more flexibility if the user wishes to do advanced experimental designs, formulas, or custom filtering of the results, that is not accessible within the application. The downside however is that some features may become unavailable due to missing information.

3. Differential Expression Analysis Results Output is used in conjunction with a Metadata file to skip the process of running DESeq2 and to just visualize results that have already been acquired. After uploading these two files, you may also directly pass them into the pathway analysis portion of the application without uploading anything else.

4 & 5. The Abundance data and Pathway Analysis Results Output may be used in conjunction with Metadata to bypass the rest of the application and just visualize and explore previously acquired results from a pathway analysis. An abundance data file is not required to use this method; however some features may not be available without it.