Genix

User Manual

Version 1.0

Gene Expression Analysis with GeniE

GeniE provides comprehensive analysis of TempO-Seq data using novel, validated and robust bioinformatics methods. Upload raw Tempo-Seq data and instantaneously obtain in-depth analysis including cluster plots, principal component plots, differentially expressed gene list, and pathway analysis.

In addition, GeniE offers extrapolated (predicted) full transcriptome from TempO-Seq measurements performed on a much smaller set of genes, such as the S1500+ platform.

GeniE solves the problem of predicting gene expression of whole transcriptomes from measurements performed on a limited number of genes.

The typical transcriptome is comprised of tens of thousands of genes. Measuring gene expression for the entire genome is infeasible for large scale studies involving hundreds or thousands of samples due to analytical limitations, as well as economic constraints. Fortunately, knowledge of the expression levels of a carefully chosen collection of genes provides sufficient information to accurately extrapolate estimates of the expression levels of the remaining genes. The methods implemented in GeniE create extrapolation models based on the deeply interdependent nature of gene expression throughout the transcriptome. We have found that using the gene expression signal from about 5-15% of the representative genes for a given species, GeniE can accurately predict gene expression for the rest of the transcriptome.

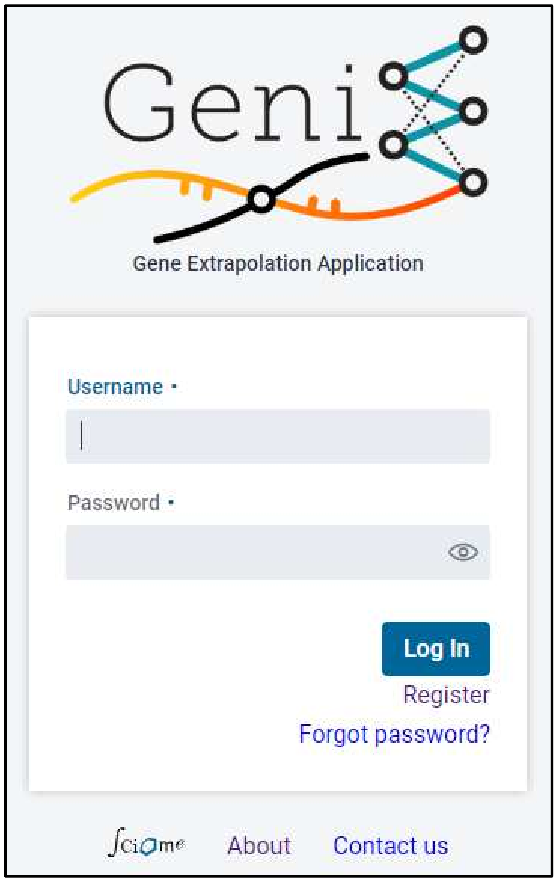
For example, you could input a file with the gene expression for 1,000-3,000 genes in 5,000 samples for the human genome. GeniE will perform an algorithmic Whole Transcriptome Extrapolation aided by a large well curated knowledge base. GeniE will compute the gene expression values for the rest of the 15,000-20,000 genes in the human transcriptome.

GeniE provides an easy access interface for submitting a user’s gene expression data set to Sciome’s transcriptome extrapolation engine. The expression signal in a simple tab-delimited format is uploaded to the extrapolation server.

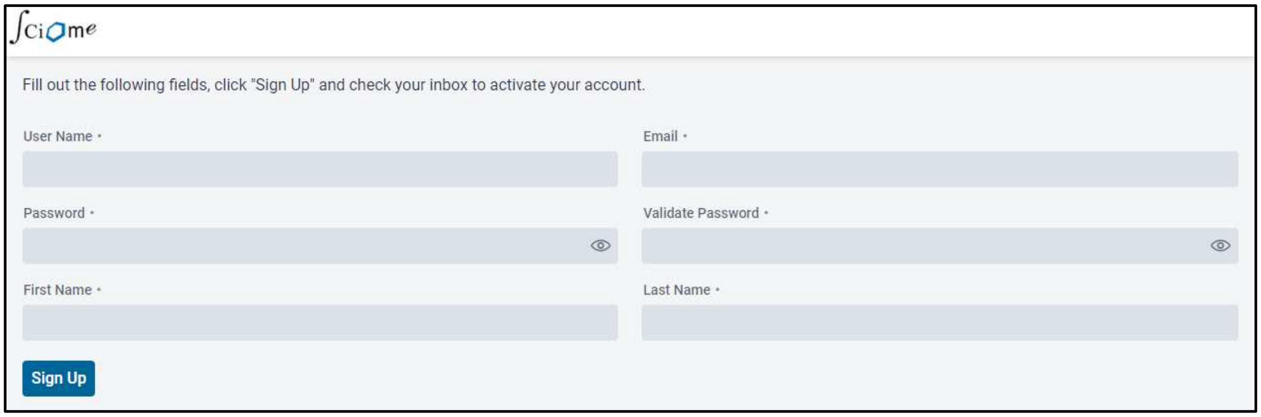
For the price of extrapolation, GeniE will also provide you with QC plots and downstream analysis (DEG and DEP results for as many contrasts as you want).

Login

The login page can be found at https://apps.sciome.com/genie/login.



Users must have Sciome account to login/register to GeniE. For a first time user, please click “Register” to be brought to the registration page at https://user.sciome.com/registration.

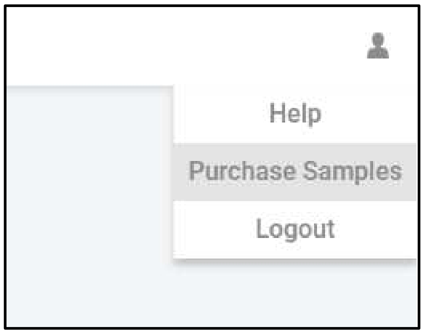

Fill out the following fields, click "Sign Up" and check your inbox to activate your account.

After you sign up you will be emailed a link to activate your account. After activation, you can login to the GeniE page with your Sciome account credentials.

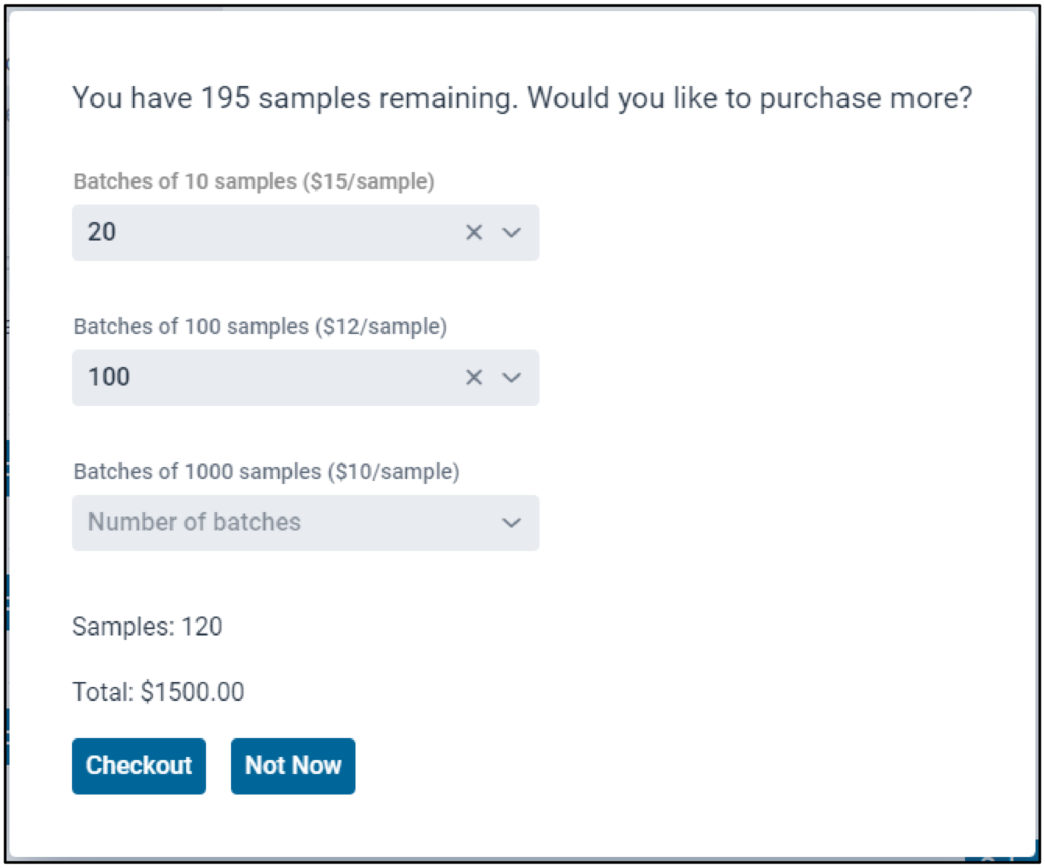
Purchasing Samples

In order to use GeniE, you will need to pre-purchase samples. Each purchased sample allows you to run one sample through the full GeniE pipeline (normalization, extrapolation, QC plots, and downstream analysis). A typical run will require multiple samples for a given study.

To purchase samples, click on the person icon in the upper right corner of your screen, and select "Purchase Samples."



Pricing per sample decreases as you make larger bulk purchases.



You have 195 samples remaining. Would you like to purchase more?

Batches of 10 samples ($15/sample)

20

Batches of 100 samples ($12/sample)

100

Batches of 1000 samples ($10/sample)

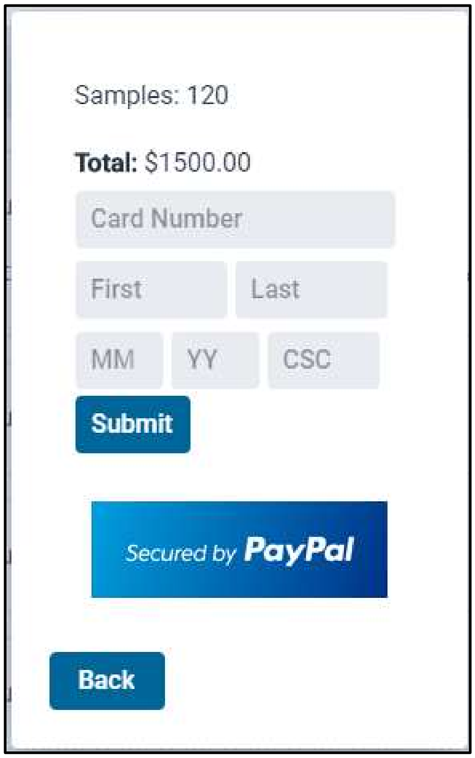
Number of batches

Samples: 120

Total: $1500.00

Use the dropdown menus to select the number of samples you would like to purchase, and click “Checkout.” The minimum number of samples is 10.

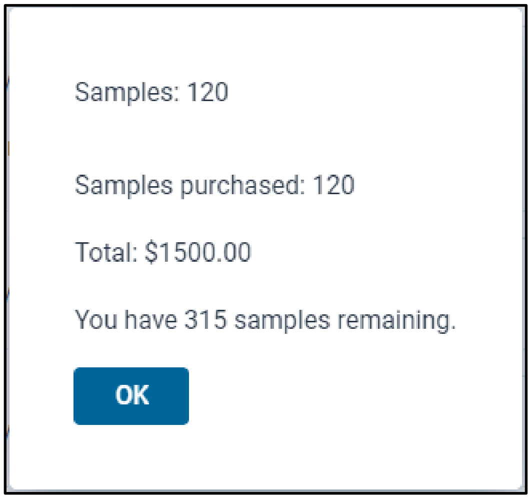
Please enter your credit card information and “Submit” to the PayPal service.



Samples: 120

Total: $1500.00

After successful processing, you will see a window confirming the purchase and letting you know how many samples you have remaining.



Samples: 120

Samples purchased: 120

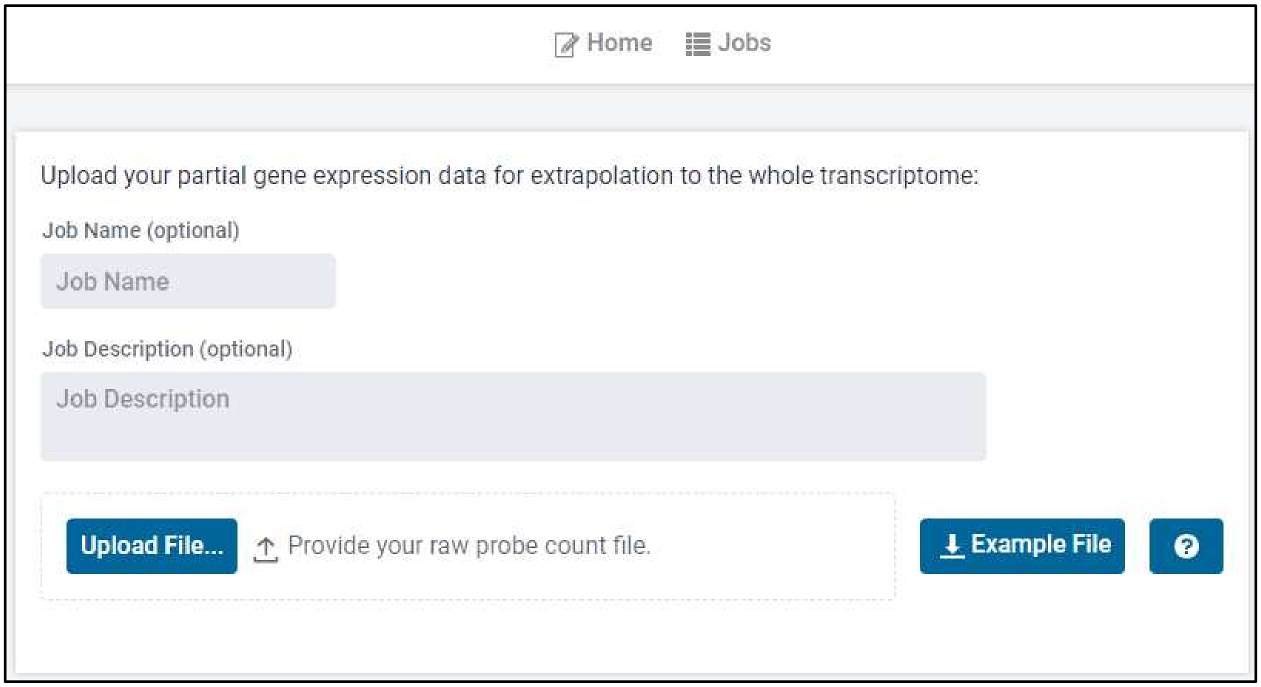
Total: $1500.00

You have 315 samples remaining.

You will also receive an email to the account used for GeniE login.

Starting your first submission

At first login a user will be brought to the “Home” page to begin a submission.



Upload your partial gene expression data for extrapolation to the whole transcriptome:

Job Name (optional)

Job Name

Job Description (optional)

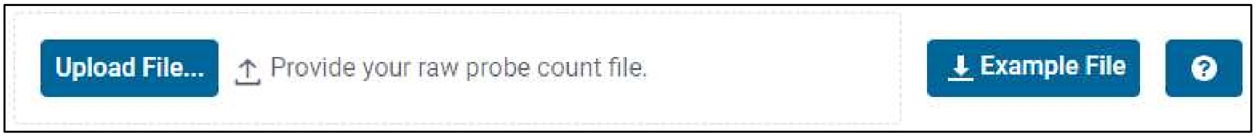
Job Description

Provide your raw probe count file.

Providing a “Job Name” and “Job Description” are optional. If the “Job Name” field is empty, a random string will be assigned to name your job at submission.

Description of the input data files

Input Raw Probe Count File (probes/genes by samples)



The input data file is a flat tab-delimited text file with the gene expression signal (unnormalized counts for each probe) for all the genes in all the samples. For example, Table 1 shows the first few rows of an input raw probe count file with 6 samples. Each row has a unique row identifier. The complete version of this Example data can be downloaded by clicking the “Example File” button.

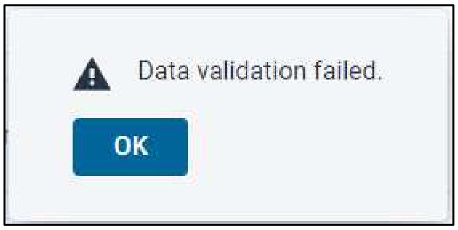
← Samples →

↑ Probes ↓

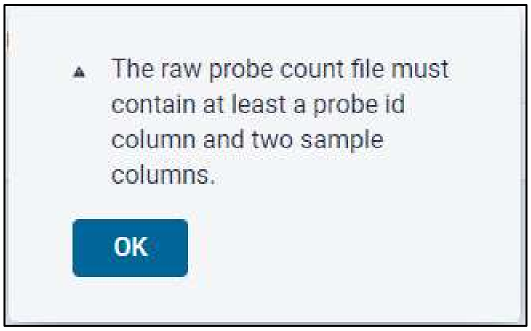
Probe name	G01	D01	F01	D02	F02	G02
A2M_12371	159	129	92	74	114	86
AARS_3	900	553	554	269	355	453
AASDHPPT_21936	293	206	89	95	107	127
AATF_26432	171	148	113	53	111	75
ABCA6_22690	16	5	13	9	18	7
ABCA6_28368	37	38	31	19	46	16

Table 1: Format of input raw probe count file (probes/genes by samples)

1. It is a probe by sample file with one row for each probe and one column for each sample in the experiment.
2. The file would contain the gene expression signal or read counts for each probe/gene in each sample.
3. The file must have exactly one header row with unique sample identifiers for each sample/column (G01, D01, etc.). Non-unique sample names will result in an error.

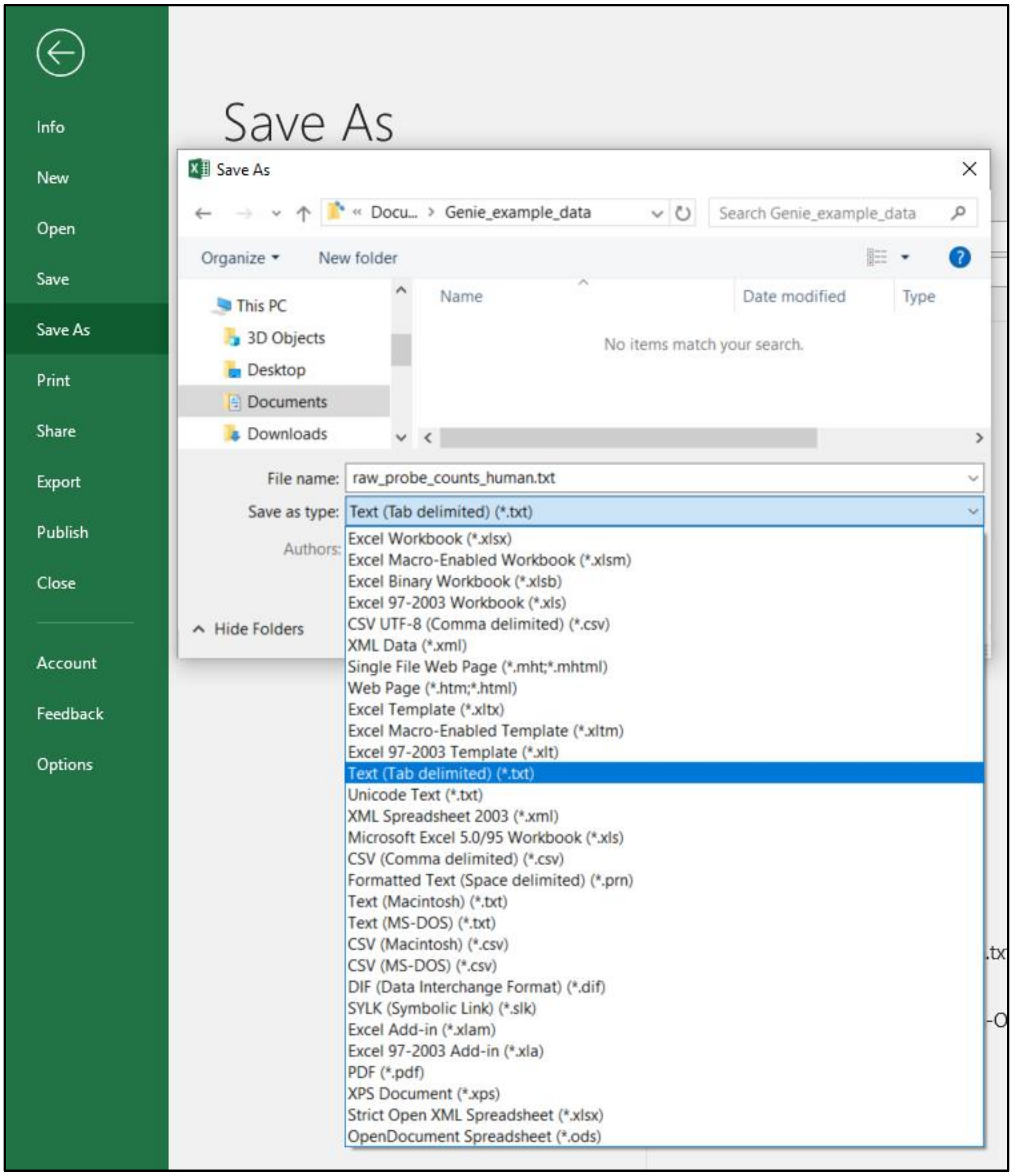

Data validation failed.

4. The file must have exactly one header “Probe name” column with unique probe identifiers for each row/probe (A2M_12371, AARS_3, etc.). Non-unique probes will also result in a data validation error.
5. The file can have more than one row for a gene as long as the full probe names are distinct (**ABCA6**_22690 and **ABCA6**_28368).
6. A minimum of 2 samples are required for the software to work. A file with only one sample will result in an error.



The raw probe count file must contain at least a probe id column and two sample columns.

When using Excel to create/format a raw probe count file, please "Save As" using type "Text (Tab delimited) (*.txt)" to ensure proper formatting as input for the GeniE software.

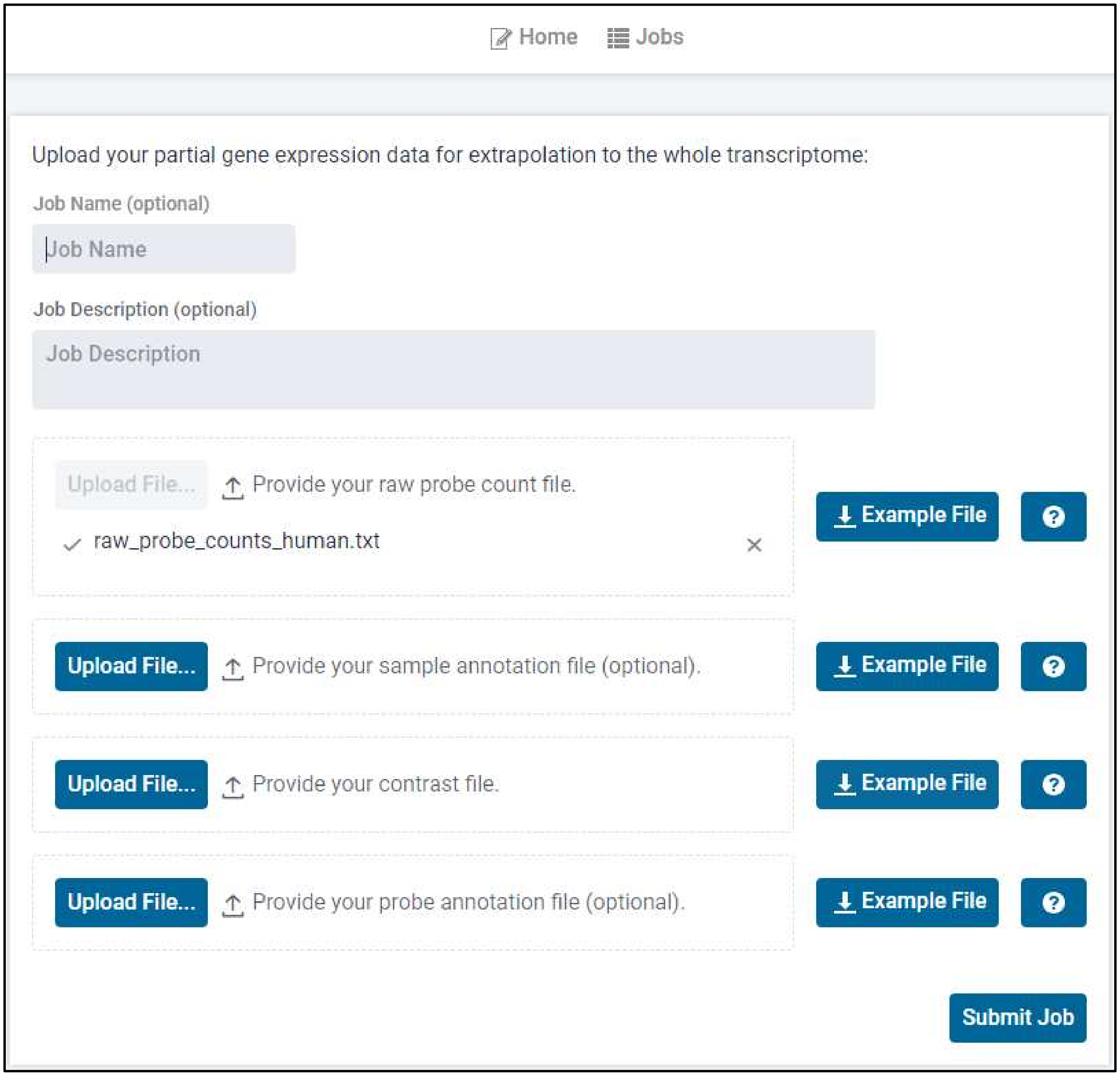


File name: raw_probe_counts_human.txt

Save as type: Text (Tab delimited) (*.txt)

Info
New
Open
Save
Save As
Print
Share
Export
Publish
Close
Account
Feedback
Options
Save As
Save As
« Docu... › Genie_example_data
Search Genie_example_data
Organize ▾
New folder
This PC
3D Objects
Desktop
Documents
Downloads
Name
Date modified
Type
No items match your search.
Authors:
Hide Folders
Excel Workbook (*.xlsx)
Excel Macro-Enabled Workbook (*.xlsm)
Excel Binary Workbook (*.xlsb)
Excel 97-2003 Workbook (*.xls)
CSV UTF-8 (Comma delimited) (*.csv)
XML Data (*.xml)
Single File Web Page (*.mht;*.mhtml)
Web Page (*.htm;*.html)
Excel Template (*.xltx)
Excel Macro-Enabled Template (*.xltm)
Excel 97-2003 Template (*.xlt)
Text (Tab delimited) (*.txt)
Unicode Text (*.txt)
XML Spreadsheet 2003 (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv)
Formatted Text (Space delimited) (*.prn)
Text (Macintosh) (*.txt)
Text (MS-DOS) (*.txt)
CSV (Macintosh) (*.csv)
CSV (MS-DOS) (*.csv)
DIF (Data Interchange Format) (*.dif)
SYLK (Symbolic Link) (*.slk)
Excel Add-in (*.xlam)
Excel 97-2003 Add-in (*.xla)
PDF (*.pdf)
XPS Document (*.xps)
Strict Open XML Spreadsheet (*.xlsx)
OpenDocument Spreadsheet (*.ods)

To import the text file to GeniE, click “Upload File…” and select the appropriate file.



Upload your partial gene expression data for extrapolation to the whole transcriptome:

Job Name (optional)

Job Name

Job Description (optional)

Job Description

Provide your raw probe count file.

✓ raw_probe_counts_human.txt ×

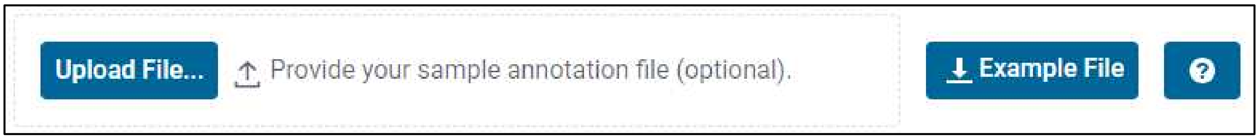
Provide your sample annotation file (optional).

Provide your contrast file.

Provide your probe annotation file (optional).

If you do not wish to run downstream analyses, often no other file is required to run GeniE. At this point the user can press “Submit Job” or add any of the optional data files.

Sample annotation file (optional)



The user can optionally provide a sample annotation file. There are two use cases that require sample annotations:

1. To use sample metadata to color and split the data by some variable(s) in the output plots.
2. To specify sample groups that will be used to determine contrasts for downstream analysis comparisons.

When to include a sample annotation file?

- For plotting features: Consider a typical dose response study with multiple chemicals and multiple doses for each chemical, the user could split data by each chemical for easy visualization and interpretation. Also, the plot for each chemical can be colored based on the doses to see if there is a pattern based on the doses.
- For downstream analysis: When there are any comparisons for which you want differential expression analysis or pathway analysis performed.

Please see section "Description of output files" subsection "Output plots" to plots for the simple example shown below using various options.

The sample annotation file is a flat tab-delimited text file with the sample metadata for all the samples in the raw probe count file. For example, Table 2 shows an input sample annotation file for the gene expression data in Table 1. This Example data can be downloaded by clicking the "Example File" button.

← Required Columns → ← Optional Column(s) →

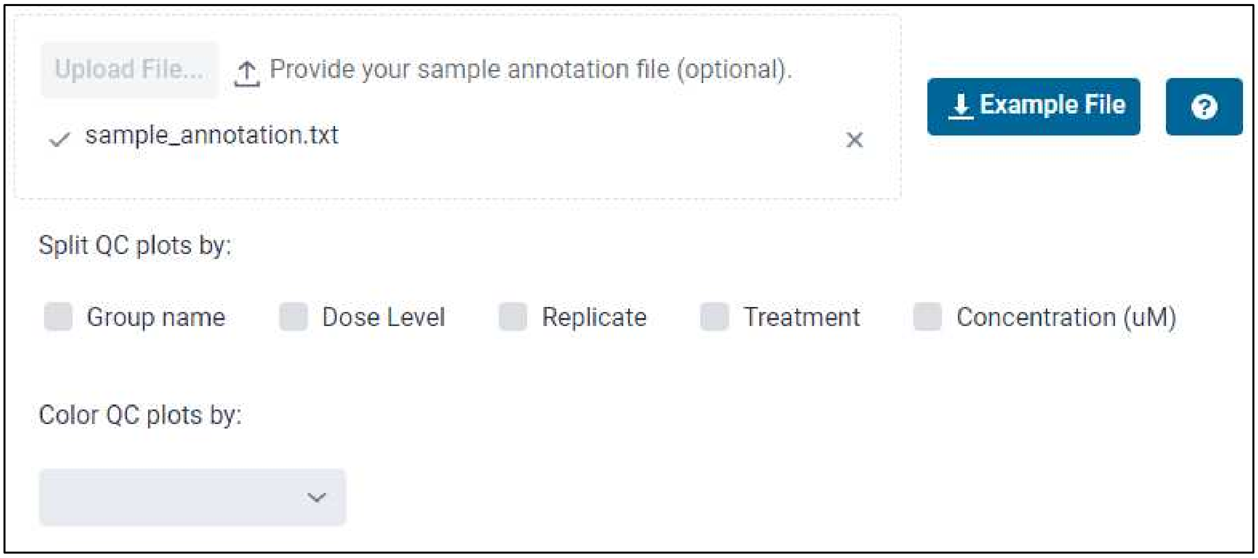
Sample name	**Group name**	Dose Level	Replicate	Treatment	Concentration (uM)	Cell	Dose Type	Organism
G01	DMSO	CONTROL	1	DMSO	0.2			
D01	A_LOW	LOW	1	Treatment A	330			
F01	A_HIGH	HIGH	1	Treatment A	3300			
D02	B_LOW	LOW	1	Treatment B	10			
F02	B_HIGH	HIGH	1	Treatment B	100			
G02	UNTREATED	CONTROL	1	Untreated Control	0			
G03	DMSO	CONTROL	2	DMSO	0.2			
D03	A_LOW	LOW	2	Treatment A	330			
F03	A_HIGH	HIGH	2	Treatment A	3300			
D04	B_LOW	LOW	2	Treatment B	10			
F04	B_HIGH	HIGH	2	Treatment B	100			
G04	UNTREATED	CONTROL	2	Untreated Control	0			

G05	DMSO	CONTROL	3	DMSO	0.2			
D05	A_LOW	LOW	3	Treatment A	330			
F05	A_HIGH	HIGH	3	Treatment A	3300			
D06	B_LOW	LOW	3	Treatment B	10			
F06	B_HIGH	HIGH	3	Treatment B	100			
G06	UNTREATED	CONTROL	3	Untreated Control	0			

Table 2: Format of input sample annotation file

1. The file must have exactly one header column called “Sample name” with unique sample identifiers for each sample (G01, D01, etc.) that match the column names in the raw probe count file.
2. If the user would like to have GeniE perform downstream analysis (differential gene expression and pathway analysis), then there is a second required column. It must be called “Group name” listing categorical groups. Each sample can fall into only one group. At least two groups must be listed.
3. The file can have any additional number of descriptive columns that can be used for the software’s plotting features.

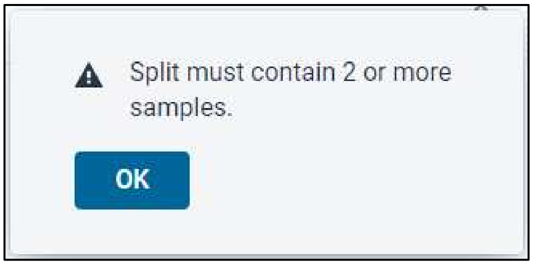
After uploading the file, the menu will expand with options.



The user can optionally choose column(s) to use to split the plots by clicking to place a check mark in the appropriate box(es) in the “Split QC plots by” subsection.

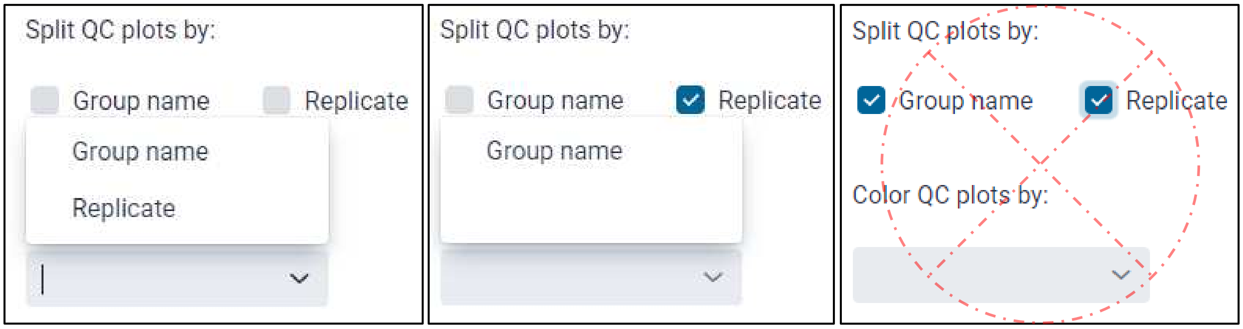
The QC plots can be split by as many sample columns as the user wants. However, at least one column must remain for the “Color QC plots by” input.

Sample columns with (1) the same value for all samples or (2) with a unique value for every sample will not appear as options for “Split QC plots by” because (1) does not split the data at all and (2) would create a new plot for every single sample. Neither of these are useful or desired. If the user selects any combination of column names for “Split QC plots by” that results in any group containing only one sample, the box will be immediately unchecked and a warning will display:

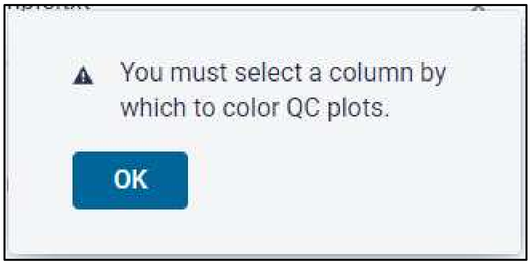


Split must contain 2 or more samples.

In the drop-down menu, select the "Color QC plots by" input. The drop-down menu will only contain columns that have not already been checked for plot splitting (examples shown below for simplified a simplified version of the example sample annotation file with only columns "Group name" and "Replicate" describing the samples).

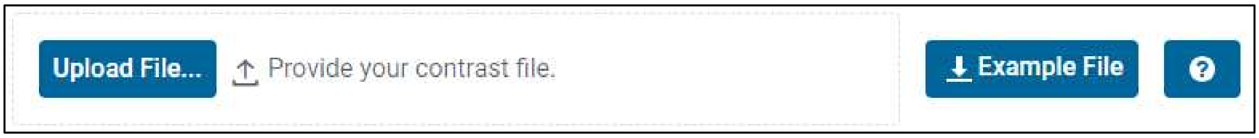


A "Color QC plots by" input must be selected when supplying a sample annotation file. Not doing so would result in an error.



You must select a column by which to color QC plots.

Contrast file (necessary for downstream analysis)



In order to tell GeniE to make differentially expressed gene (DEG) and differentially enriched pathway (DEP) calls, a contrast file must be supplied.

The contrast file is a flat tab-delimited text file describing the comparisons the user wishes to analyze and indicating sample groups for this comparison. For example, Table 3 shows an input sample annotation file for the samples in Table 2. This Example data can be downloaded by clicking the "Example File" button.

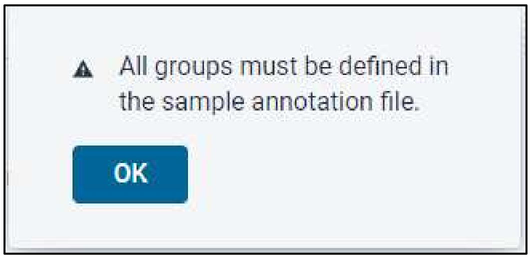
← Required Columns → ← Optional Column(s) →

↑ Contrasts ↓

Contrast name	Group names (A)	Group names (B)	Description	Cell	Organism
A_HIGH_vs_DMSO	A_HIGH	DMSO	Treatment A: 3000uM vs DMSO		
A_vs_DMSO	A_LOW+A_HIGH	DMSO	Treatment A vs DMSO		
B_HIGH_vs_UNTREATED	B_HIGH	UNTREATED	Treatment B: 100uM vs Untreated		

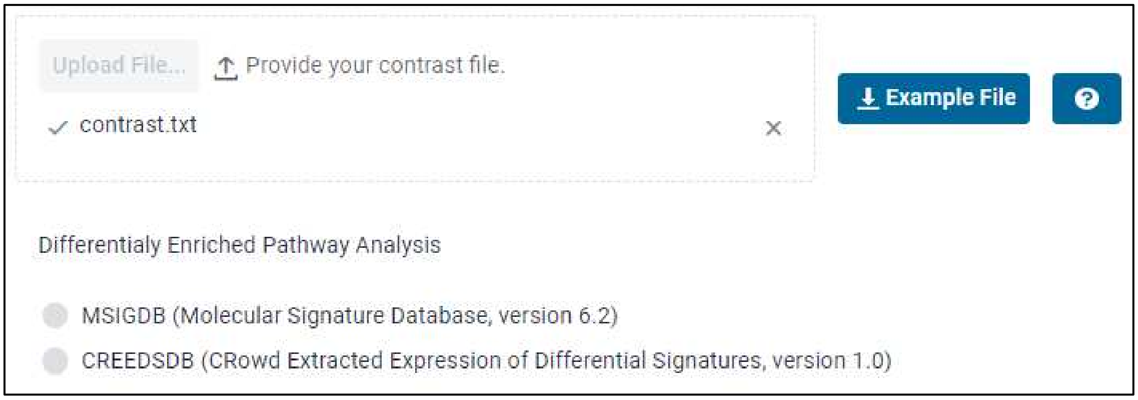
Table 3: Format of contrast file

1. The file must have exactly one header column called "Contrast name" designating the comparisons to make for DEG and DEP calls. A contrast can have any name desired by the user, but it must contain only alphanumeric and underscore "_" characters.
2. Two additional columns are required: "Group names (A)" and "Group names (B)".
    a. For each contrast (row), the group(s) listed in A will be in the numerator for fold change calculations.
    b. For each contrast (row), the group(s) listed in B will be the denominator for fold change calculations (often a control group).
    c. The plus "+" operator can be used (without any blank space) to include multiple groups of samples in the numerator or denominator.
    d. All groups must appear in the sample annotation file. Any group names that do not match will result in an error.


▲ All groups must be defined in the sample annotation file.

3. The file can have any additional number of descriptive columns that will not be used by the software.

After uploading the file, the menu will expand with database options.

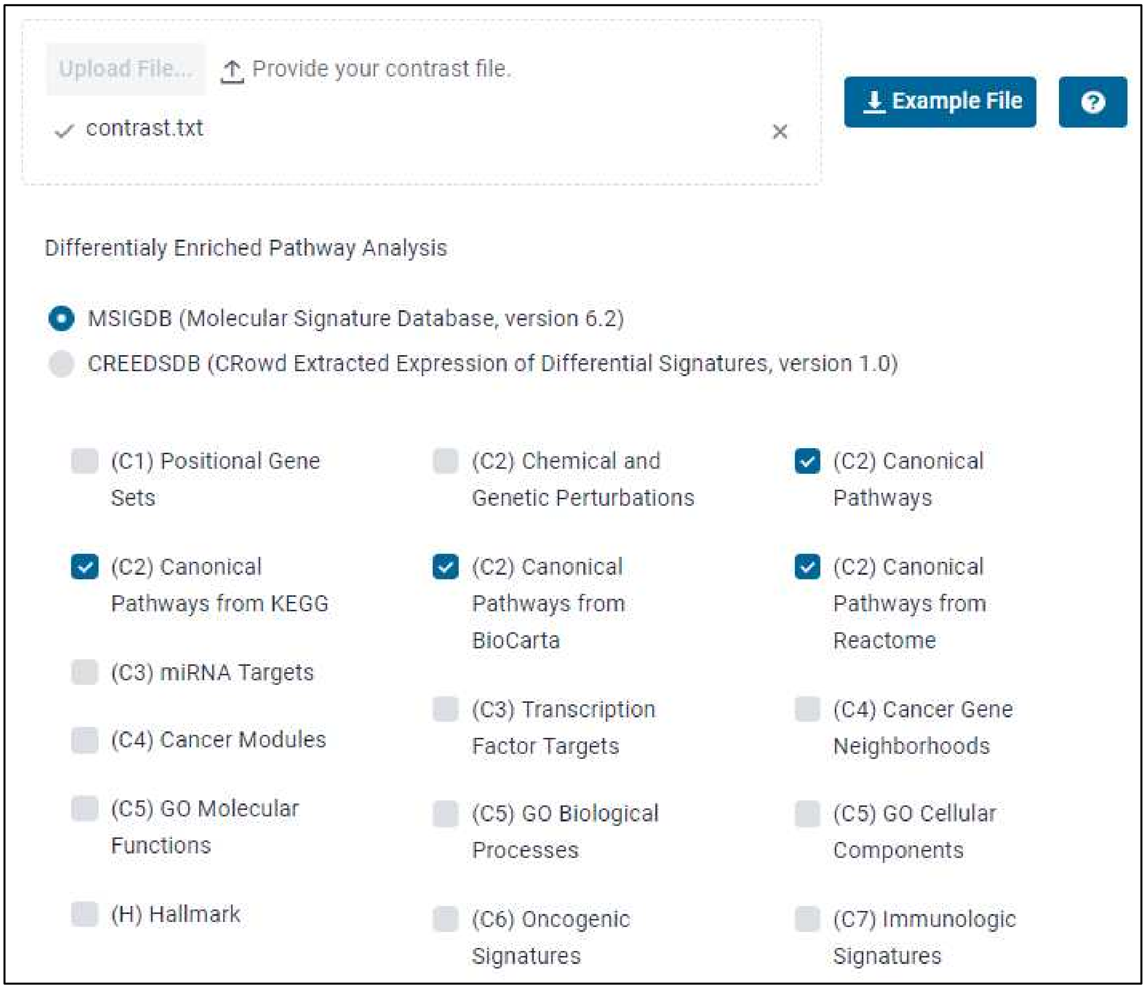


✓ contrast.txt ×

Differentialy Enriched Pathway Analysis

○ MSIGDB (Molecular Signature Database, version 6.2)
○ CREEDSDB (CRowd Extracted Expression of Differential Signatures, version 1.0)

Select "MSIGDB (Molecular Signatures Database, version 6.2)" to use signatures from MSigDB to perform pathway analysis. These are the standard signatures for gene set enrichment analysis (GSEA). By default, the C2-CP canonical pathways will be selected. Check out their website or publication to determine which signatures are most appropriate for your data.



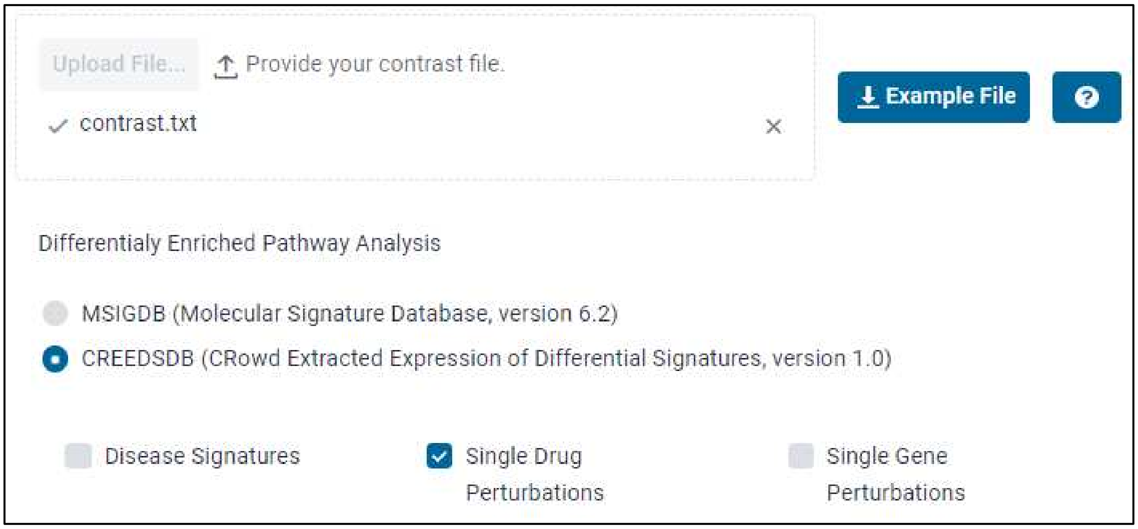
✓ contrast.txt ×

Differentialy Enriched Pathway Analysis

● MSIGDB (Molecular Signature Database, version 6.2)
○ CREEDSDB (CRowd Extracted Expression of Differential Signatures, version 1.0)

☐ (C1) Positional Gene Sets
☐ (C2) Chemical and Genetic Perturbations
☑ (C2) Canonical Pathways
☑ (C2) Canonical Pathways from KEGG
☑ (C2) Canonical Pathways from BioCarta
☑ (C2) Canonical Pathways from Reactome
☐ (C3) miRNA Targets
☐ (C3) Transcription Factor Targets
☐ (C4) Cancer Gene Neighborhoods
☐ (C4) Cancer Modules
☐ (C5) GO Molecular Functions
☐ (C5) GO Biological Processes
☐ (C5) GO Cellular Components
☐ (H) Hallmark
☐ (C6) Oncogenic Signatures
☐ (C7) Immunologic Signatures

Select “CREEDSDB (CRowd Extracted Expression of Differential Signatures, version 1.0)” to use the CREEDS signatures for pathway analysis. These gene sets are directional, so using CREEDS will perform an analysis that accounts for the up- or down- direction of DEGs in your data when making inferences about the enrichment of each signature. By default, the “Single Drug Perturbations” subcategory will be selected. Check out their website or publication to determine which signatures are most appropriate for your data.



✓ contrast.txt ×

Differentialy Enriched Pathway Analysis

○ MSIGDB (Molecular Signature Database, version 6.2)

● CREEDSDB (CRowd Extracted Expression of Differential Signatures, version 1.0)

□ Disease Signatures ☑ Single Drug Perturbations □ Single Gene Perturbations

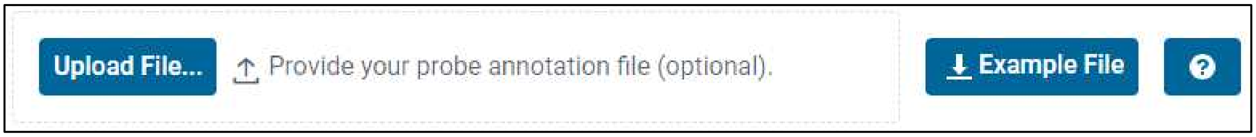
You can only choose to use MSigDB or CREEDS in a single run of GeniE but can include number of subcategories from your selected option. To generate output from both MSigDB and CREEDS two separate runs of your samples through the software would be required.

The calculations to determine p-values, etc. for each of these pathway/signature database options are described here.

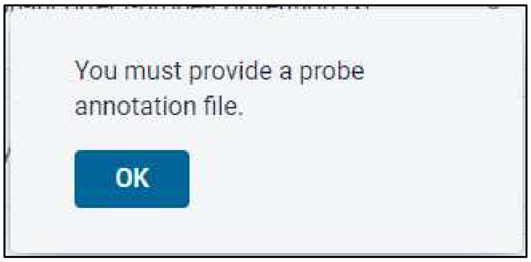
Probe annotation file (optional)

For each unique probe the software needs to know the corresponding gene. Currently GeniE will compare the probe names from the raw probe count file to available manifests to automatically determine the correct manifest for the correct species (human, rat, or mouse) to assign probes to genes.

The user can optionally provide gene annotations for their probes in a separate file. In most cases this is not required. Please refer to the FAQ for more information.



If your probes are named using a different convention than the traditional BioSpyder S1500+ platforms or many probes do not match up to an available manifest, the software will alert you that a probe annotation file is required after you upload your raw probe count file.


You must provide a probe annotation file.

If this pop-up does not appear, your probes should be correctly assigned to genes without uploading a probe annotation file.

Table 4 shows a probe annotation file corresponding to the Table 1 probes. This Example data can be downloaded by clicking the "Example File" button.

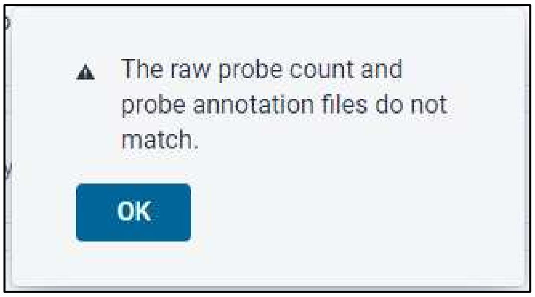
← Required Columns → Optional Column(s)

Probe name	Gene_Symbol	Entrez_Gene_ID	Probe sequence
A2M_12371	A2M	2	CTTCAGGTTCAACCAACAGAGGCTTGATGACTGTGTCTTTCCTTCCGTGT
AARS_3	AARS	16	TGCAGACACATCCTTGCCACCACCTTTACCGTCCATCAAGCCTGACACCT
AASDHPPT_21936	AASDHPPT	60496	TCATCAGACGACCAGCCATGGCTGCCTTAGCGTCCCGGGCAAAGACGAAC
AATF_26432	AATF	26574	GGCTCAGGTTTGCCAAGAACTCGATAGACAGAGCGCTTGGTCTGTGTCCT
ABCA6_22690	ABCA6	23460	GGCTCAGCACAAAACACAACTTTCTCGTGATTCCTGCTGTTAATTTCTGC
ABCA6_28368	ABCA6	23460	GGTTAAAGTTATGCTTCACTGCTTCTAATTTGTGAAAGGTCTGTGATAGA

Table 4: Format of input probe annotation file

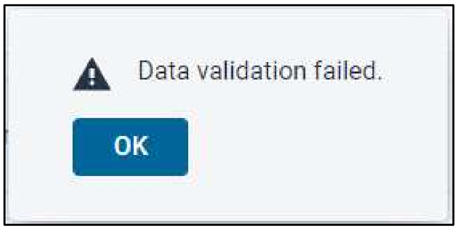
The probe annotation file provides more details about each row in the raw probe count file.

1. The probe annotation file **must** have all the same unique probe names as the gene expression signal file in a header column named “Probe name.” The user will get an error message if there are any discrepancies.



The raw probe count and probe annotation files do not match.

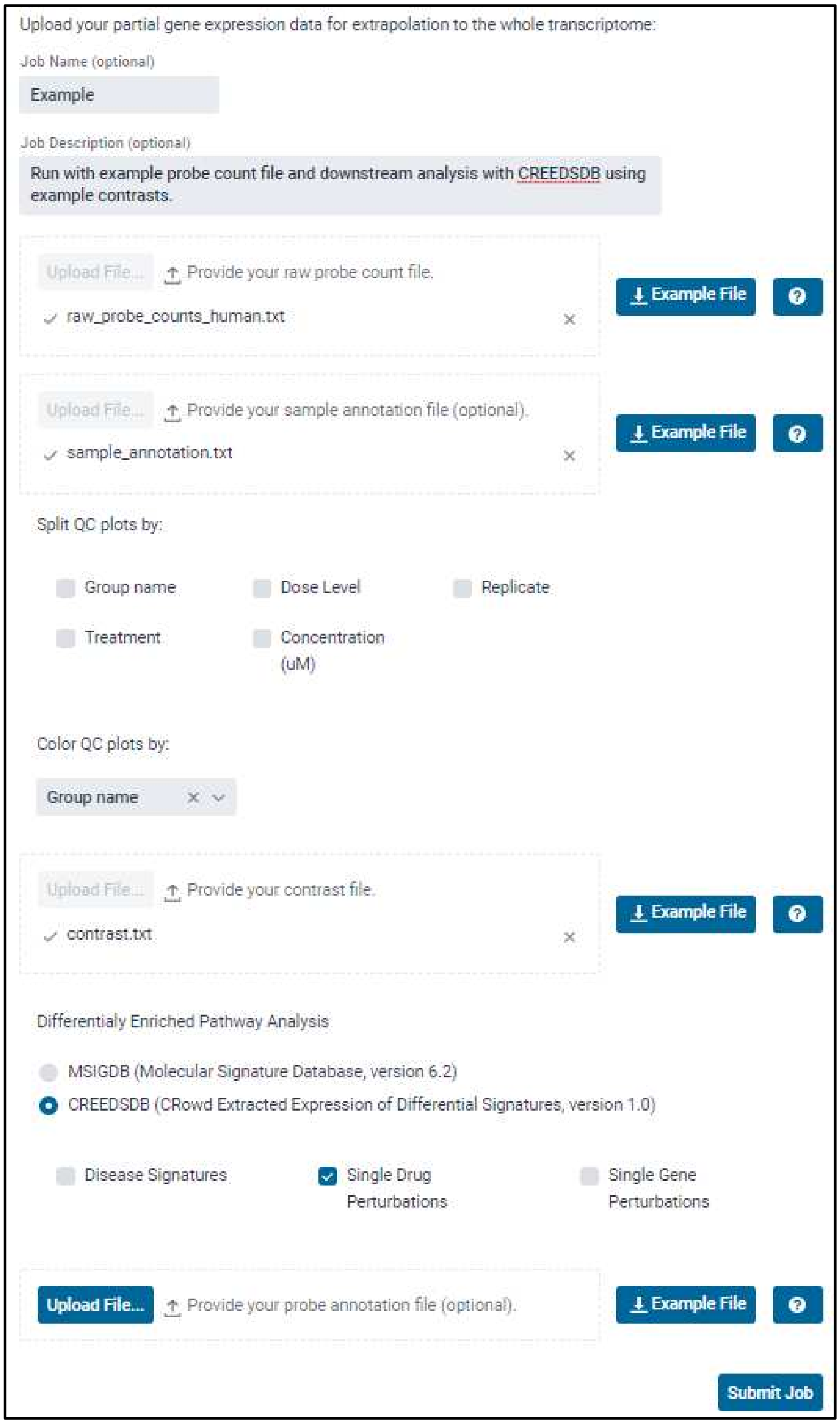
2. For each unique probe, it **must** have the gene symbol and Entrez ID. These columns must be named “Gene_Symbol” and “Entrez_Gene_ID” or the user would get an error message.



Data validation failed.

3. The probe annotation file could have additional optional information about the gene such as the probe sequence, gene description, etc. These additional columns are ignored by the tool.

Submitting a job



Upload your partial gene expression data for extrapolation to the whole transcriptome:

Job Name (optional)

Example

Job Description (optional)

Run with example probe count file and downstream analysis with CREEDSDB using example contrasts.

Provide your raw probe count file.

✓ raw_probe_counts_human.txt ×

Provide your sample annotation file (optional).

✓ sample_annotation.txt ×

Split QC plots by:

□ Group name □ Dose Level □ Replicate

□ Treatment □ Concentration (uM)

Color QC plots by:

Group name × ⌄

Provide your contrast file.

✓ contrast.txt ×

Differentialy Enriched Pathway Analysis

○ MSIGDB (Molecular Signature Database, version 6.2)

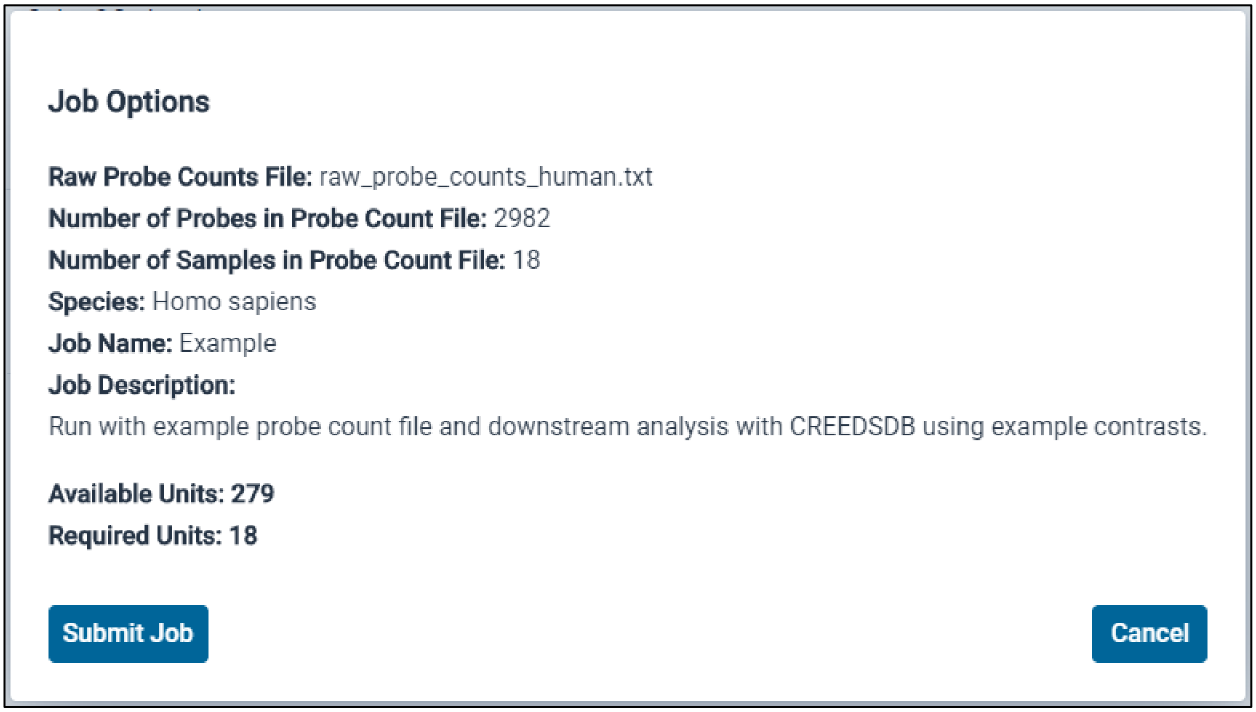
● CREEDSDB (CRowd Extracted Expression of Differential Signatures, version 1.0)

□ Disease Signatures ☑ Single Drug Perturbations □ Single Gene Perturbations

Provide your probe annotation file (optional).

After importing the raw probe count file and any other optional files and selections, the user may click on “Submit Job” to begin the submission.

A “Job Options” box will appear describing the run the user is about to submit.



Job Options

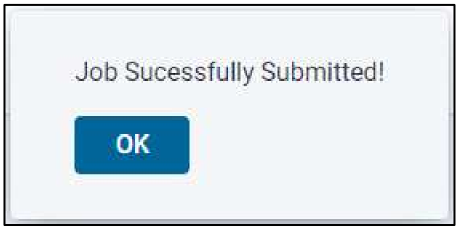
Raw Probe Counts File: raw_probe_counts_human.txt
Number of Probes in Probe Count File: 2982
Number of Samples in Probe Count File: 18
Species: Homo sapiens
Job Name: Example
Job Description:
Run with example probe count file and downstream analysis with CREEDSDB using example contrasts.

Available Units: 279
Required Units: 18

In this example, the software determined the correct species and will use the corresponding manifest file to annotate probes to genes (since no probe annotation file was uploaded). If the “Species” (“Homo sapiens” in this example) is not correct for the input expression data, please check the data files!

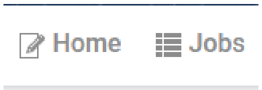
The “Job Options” dialog also displays how many samples will be required to run the job, reflected in the “Required Units” portion. This number will be equivalent to the number of samples in the inputted raw probe count file. If the “Required Units” is higher than the “Available Units” please follow the directions in the “Purchasing Samples” section to purchase more samples in order to run the job.

Click the red “Submit Job” text to finish submitting the job. If the uploaded data has been formatted correctly the following message should appear.



Job Sucessfully Submitted!

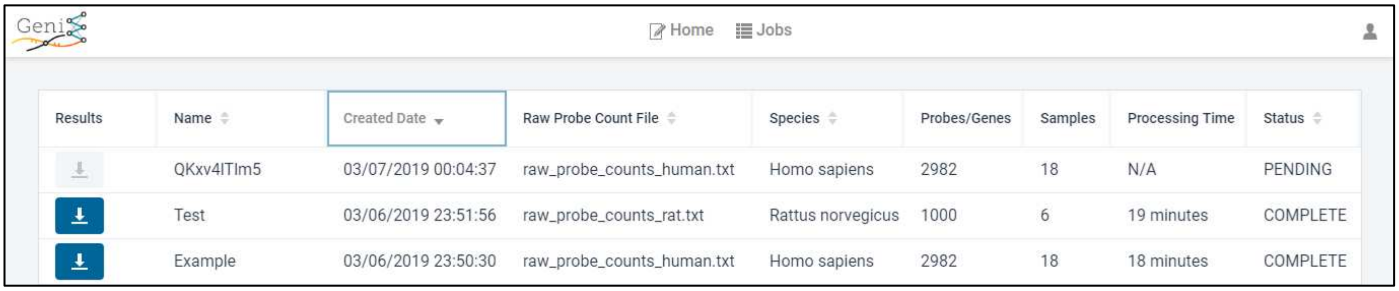
Shortcut Icons



There are two main icons on the menu bar (on the top). Each one leads to a different page.

Click the "Home" icon to go the GeniE home page to start a new analysis and submit a new job.

Click the "Jobs" icon to view/download all the previous/current jobs, ran by user.

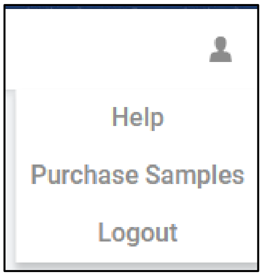


Results	Name	Created Date	Raw Probe Count File	Species	Probes/Genes	Samples	Processing Time	Status
⤓	QKxv4ITIm5	03/07/2019 00:04:37	raw_probe_counts_human.txt	Homo sapiens	2982	18	N/A	PENDING
⤓	Test	03/06/2019 23:51:56	raw_probe_counts_rat.txt	Rattus norvegicus	1000	6	19 minutes	COMPLETE
⤓	Example	03/06/2019 23:50:30	raw_probe_counts_human.txt	Homo sapiens	2982	18	18 minutes	COMPLETE

Contents:

- Results: An icon allows the user to download the output if the job has completed.
- Name: The name that the user supplied or automatically generated name for the submission.
- Created Date: Date and time that the job was submitted.
- Raw Probe Count File: Name of the input raw probe count file.
- Species: Name of the species determined by the software. In most use cases this will be determined based on the input probe names. If the user inputted an optional probe annotation file and the probes were named differently than in the BioSpyder platforms, then this field may say "Unknown."
- Probes/Genes: Number of rows in the input raw probe count file.
- Samples: Number of samples in the input raw probe count file.
- Processing Time: The time it took for a completed job to run, or N/A if the job is still running or failed.
- Status: COMPLETE for a completed job, PENDING for a job that is still running, FAILED for a job that ended with an error.

The jobs can be sorted by any of the columns with up and down arrows. The above example was sorted by descending "Created Date."



The user menu is in the top right-hand corner of the screen. Clicking on this icon opens 3 additional options.

Click “Help” to see the HTML User Manual. “Question mark” symbols within sections of the Home page will also direct you to specific sections of the User Manual.

Click “Purchase Samples” to purchase samples for runs of the GeniE software.

Click “Logout” to logout of GeniE.

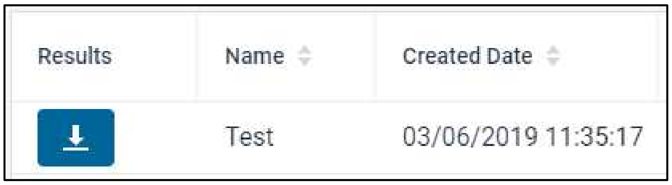
Email Notification

After submitting the job, user will get an email (associated with their login account) about submission of the job. The email will be titled “GeniE job [Job Name] submitted” containing the following:

- Raw Probe Counts File Name: Name of the input raw probe count file.
- Sample Annotation File (if provided): Contains metadata about each sample.
- Sample Contrast File Name (if provided): Contains contrasts defining comparisons for downstream analysis.
- Species: Name of the species determined by the software. In most use cases this will be determined based on the input probe names. If the user inputted an optional probe annotation file and the probes were named differently than in the BioSpyder platforms, then this field may say “Unknown.”
- Job Name: The name that the user supplied or automatically generated name for the submission.
- Job Description: The description provided by the user with the submission, blank if none was supplied.
- Submission Date: The date and time that the job was submitted.
- Information about the number of units (i.e. samples) the user has remaining.

After completion of the job, the user will get another email about successful completion or failure of the job with the reason for a job failure. The email will reiterate all of the above job information in addition to the completion date and time.

The user must login to GeniE tool and go to the “Jobs” page to download the output. After navigating to “Jobs”, the user can view all their past jobs’ results and download them as output (txt files and plots) as a zipped file.



Results	Name	Created Date
⬇	Test	03/06/2019 11:35:17

By Clicking on “download” icon, user can download the output in zipped format.

Description of output files

Output data

The output is downloaded as a zipped file called "[JobName]-Deliverables.zip" that contains a Deliverables folder with a descriptive README file and the following subfolders:

- Probe-Level-Normalized-Array
    - Output data at the probe-level (ie., level of the input data) with log2 transformed TPM normalized expression values for each sample for each input probe.
- Post-Extrapolation-Normalized-Array
    - Output data for extrapolated genes in addition to input genes (input probes summarized at the gene-level) with normalized expression values for each sample for each unique gene (based on Entrez ID) in the transcriptome for that species.
- Pre-Extrapolation-QC-Plots
    - Output plots (see next subsection for details).
- Post-Extrapolation-QC-Plots
    - Output plots (see next subsection for details).
- Differentially-Expressed-Genes-Analysis
    - DEG results if a contrast file was uploaded.
- Differentially-Enriched-Pathways-Analysis
    - DEP results if a contrast file was uploaded.

Each "Array" folder contains the following files:

- SAMPLE_ANNOTATIONS.TXT
    - File containing one row per input sample with user-provided metadata (if supplied) and additional sample details computed by the software.
- PROBE_ANNOTATIONS.TXT
    - File containing one row per probe or gene with additional descriptive information including Entrez IDs and a Measured vs. Extrapolated indicator.
- PROBE_LEVEL_EXPRESSION.TXT or WHOLE_TRANSCRIPTOME_EXPRESSION.TXT
    - File containing one row per probe or gene and one column per sample. This file reports log2 transformed normalized expression.

Number of rows of data for each file type:

- SAMPLE_ANNOTATIONS.TXT
    - Equivalent to the number of input samples.
- PROBE_ANNOTATIONS.TXT and *_EXPRESSION.TXT
    - In Probe-Level-Normalized-Array:
        - Equivalent to the number of input probes (ie., number of rows in the input raw probe count file).
    - In Post-Extrapolation-Normalized-Array:

- The number of input genes (where multiple probes may be assigned to the same gene) plus the number of extrapolated genes. The number of input genes will be less than or equal to the number of input probes, and the input genes will be listed first in the output data files.

For the Example job, probe-level files will contain the 2,982 input probes. These probes map to 2,821 unique genes. Post-extrapolation files will contain expression for all 25,701 genes in the human whole transcriptome (2,821 input genes + 22,880 extrapolated genes).

Output downstream analysis results

If a contrast file was supplied, then there will be two additional folders with the results of DEG and DEP analyses. Please see here for a description of the methods, calculations of the output values, and citations.

The “Differentially-Expressed-Genes-Analysis” folder contains the following files:

- Gene-Annotation.txt
- Filenames-Key.txt
    - This includes the information from the input contrast file.
- [Contrast name].txt
    - One txt file per contrast will be outputted. Each file will contain the following columns:
        - NCBI_Entrez_Gene_Identifier
        - Symbol
        - Status
        - Foldchange
        - Test Statistic
        - Distribution P-value
        - Local Permutation based P-value
        - Global Permutation based P-value
        - False Detection Rate (FDR)
        - Family-wise Error Rate (FWER)

The gene-level analysis performs t-tests for all genes. We recommend using a combination of foldchange, uncorrected, and corrected p-value thresholds to determine DEGs. Therefore, all of these metrics are provided to the user.

The “Differentially-Enriched-Pathways-Analysis” folder contains the following files:

- Gene-Annotation.txt
- Filenames-Key.txt
    - This includes the information from the input contrast file.
- [Contrast name].txt
    - One txt file per contrast will be outputted. Each file will contain the following columns:
        - STANDARD_NAME
        - SYSTEMATIC_NAME
        - Enrichment Score (ES)
        - ES P-value
        - Normalized Enrichment Score (NES)
        - NES P-value
        - NES False Discovery Rate (FDR)
        - NES Family-wise Error Rate (FWER)

The pathway analysis is performed using the GSEA method for MSigDB signatures or a modified version for directional CREEDS signatures, depending on user-input options when uploading the contrast file. We recommend sorting gene signatures using NES and NES P-values to determine significance.

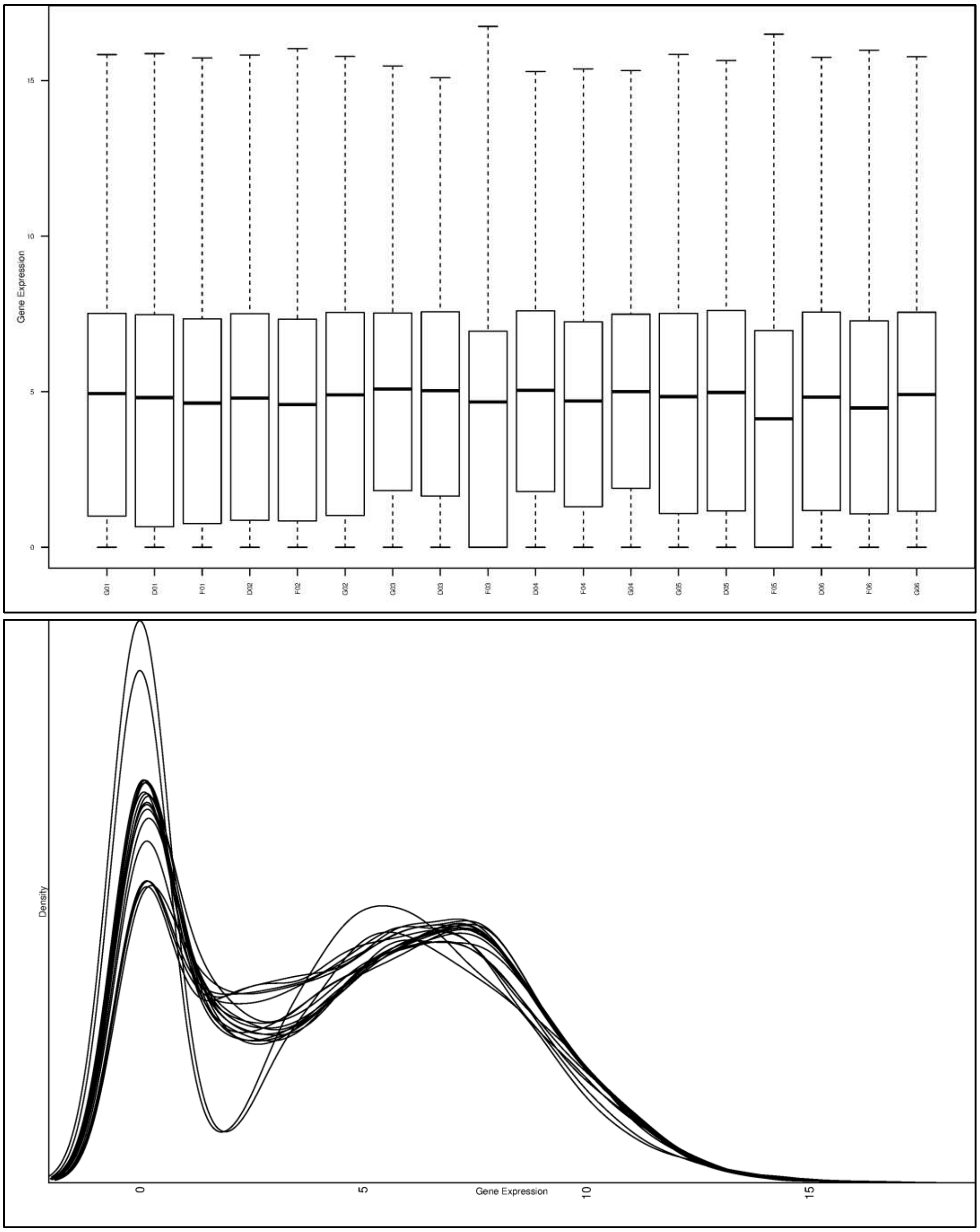
Output plots

If a sample annotation file was not imported, the “QC-Plots” folders will have only the following component:

Density-Plots.rtf

This file type can be opened with Microsoft Word or similar programs. The density plots will be black and white with one box/line per sample in each plot in each rtf file. The file will also contain the run date and time on each page.

Pre-Extrapolation: Density Plots

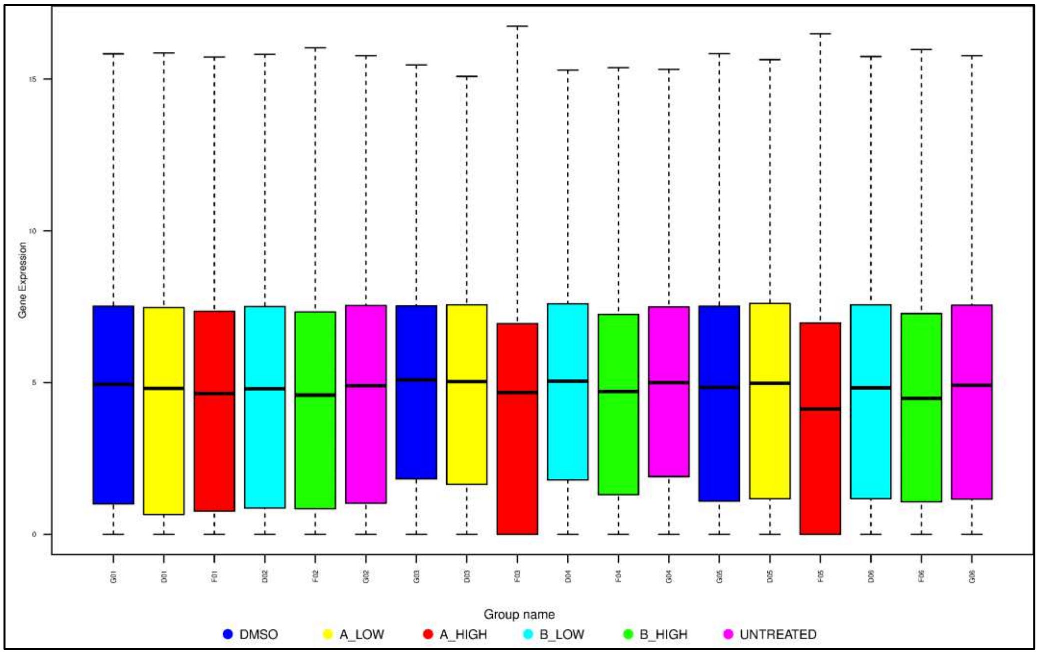


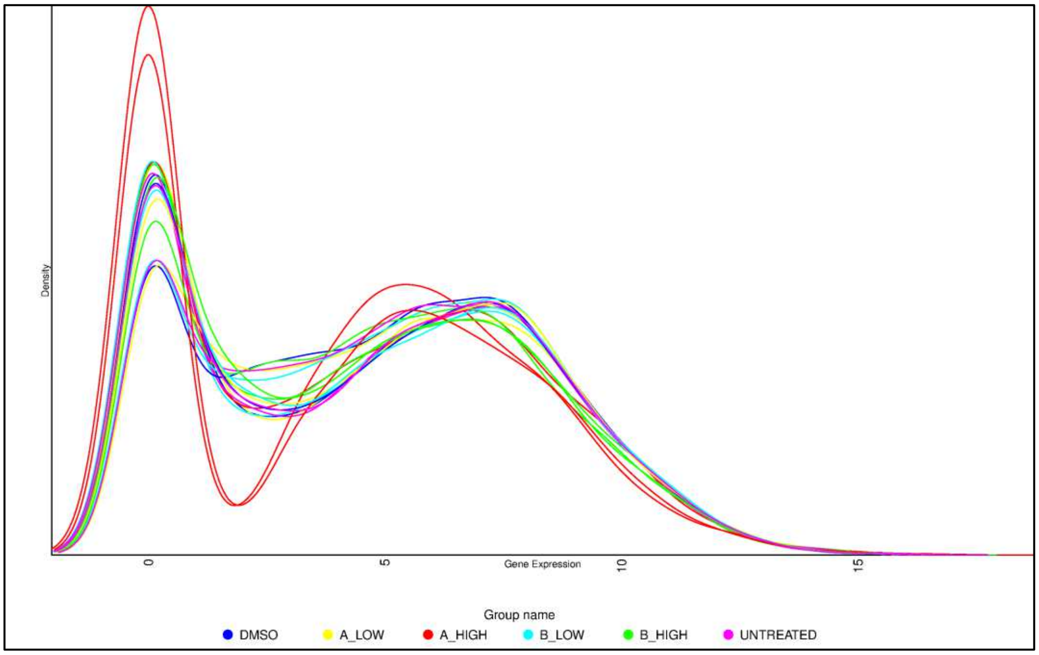
If a sample annotation file was supplied, the “QC-Plots” folders will have pre- and post-extrapolation versions of the following:

- Density-Plots.rtf
- Cluster-Plot.rtf
- Scatter-Plots.rtf
- PCA-Plots folder

The density plots will be colored by the “Color QC plots by” input column. Resulting pre-extrapolation density plots for the Example with “Color QC plots by” Group name:

Pre-Extrapolation: Density Plots

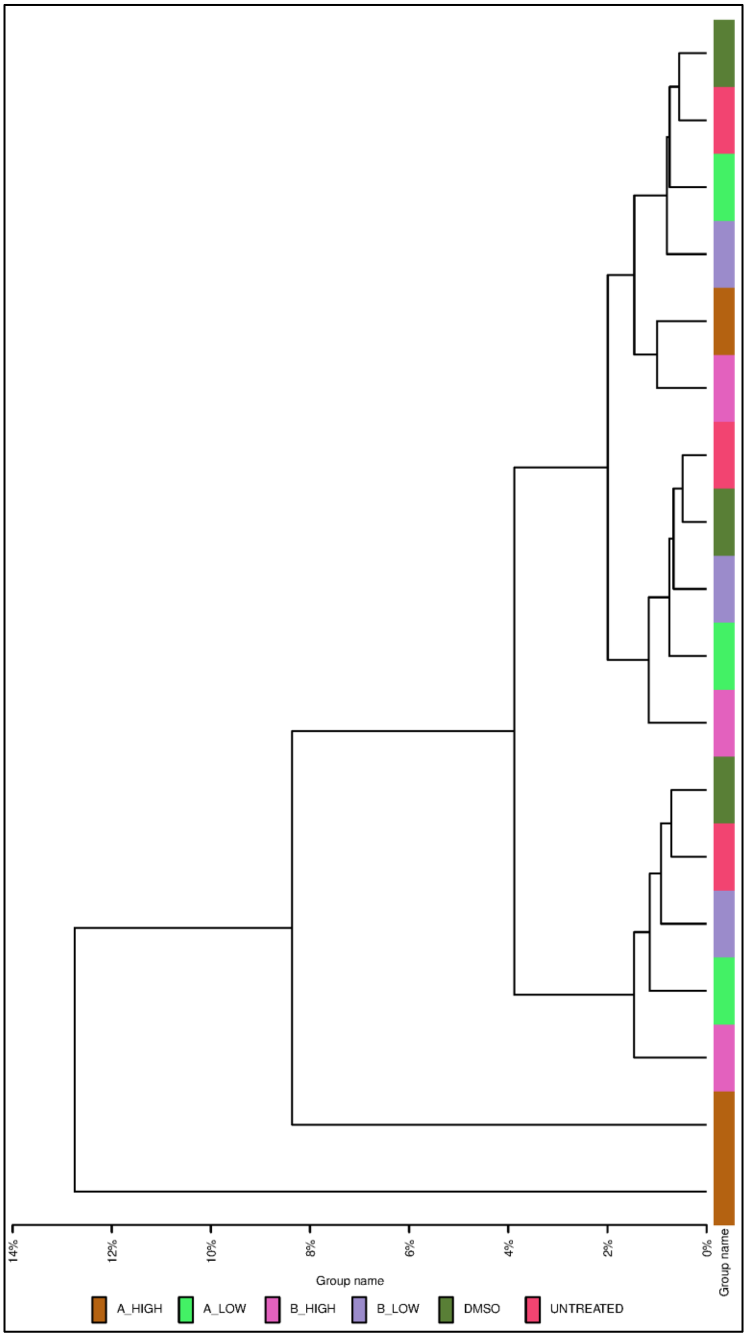




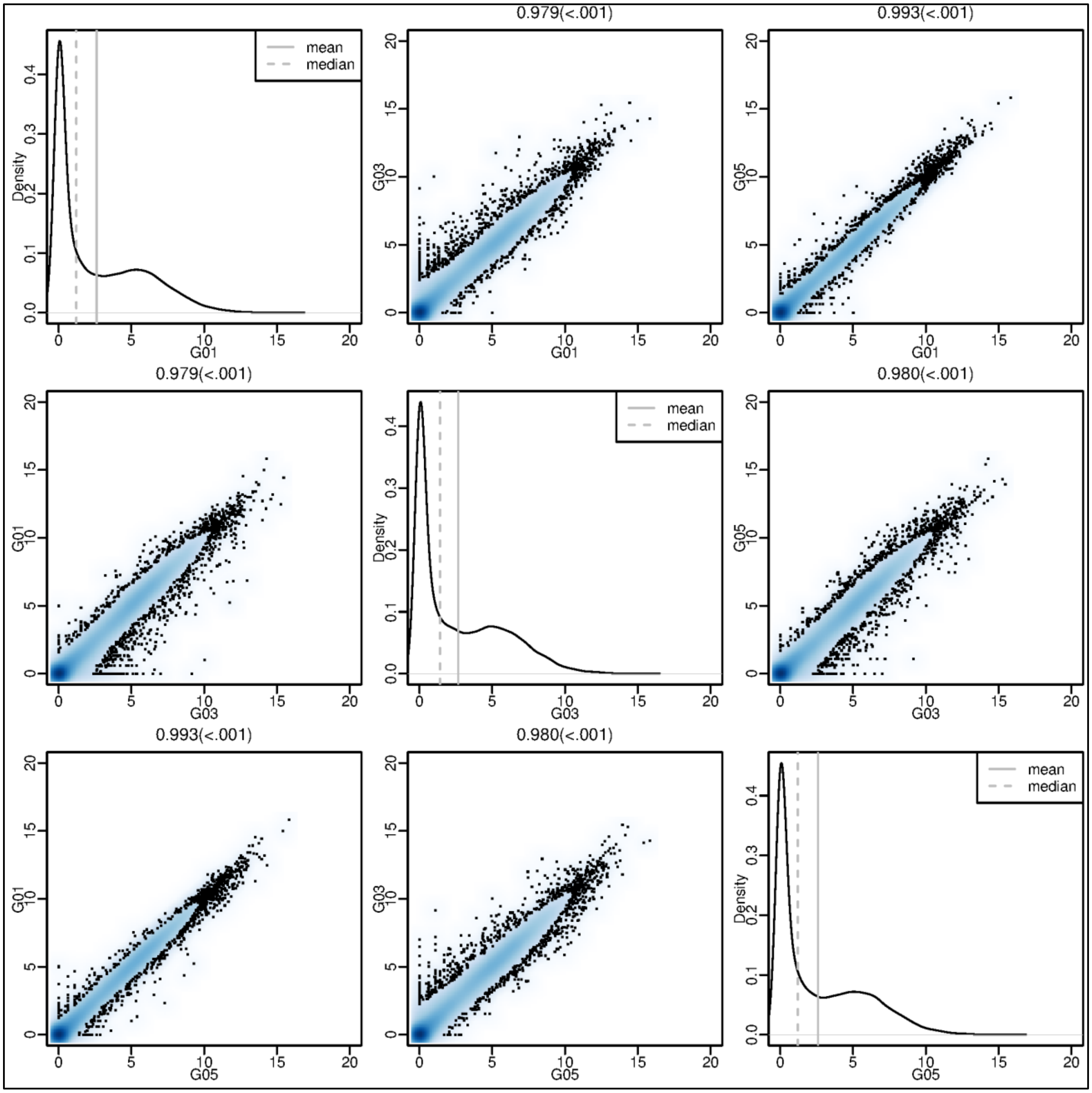
If any boxes were checked for “Split QC plots by” then there will be one of each type of plot per category of the combined splitting variable(s). For example, if we had split by “Treatment” there would be 4 resulting versions of each plots (one for each of DMSO, Treatment A, Treatment B, and Untreated Control). If we had split by “Dose Level” and “Replicate” there would be 9 plots (3 Dose Levels x 3 Replicates).

Here are examples of the other output plots using the Example data with “Color QC plots by” Group name and no splitting:

Post-Extrapolation: Cluster Plots



Post-Extrapolation Scatter Plots
Group name=DMSO
Pearson's correlation coefficients

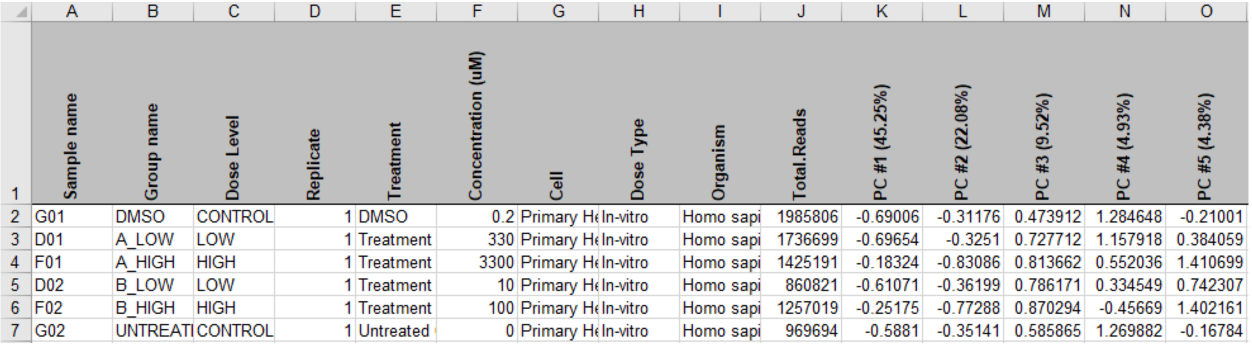


NOTE 1: The scatter plots are not colored by the "Color QC plots by" factor (Group name), but rather split by the "Color QC plots by" factor. There would be additional plots output for "Group name=A_HIGH" and all other groups using the same color scheme here.

Contents of the PCA-Plots folders:
Examples shown for Post-Extrapolation using "Color QC plots by" Group name.

- Batch_[number]-Principal-Componenets.xml

- Portion of Example XML file opened with Excel:



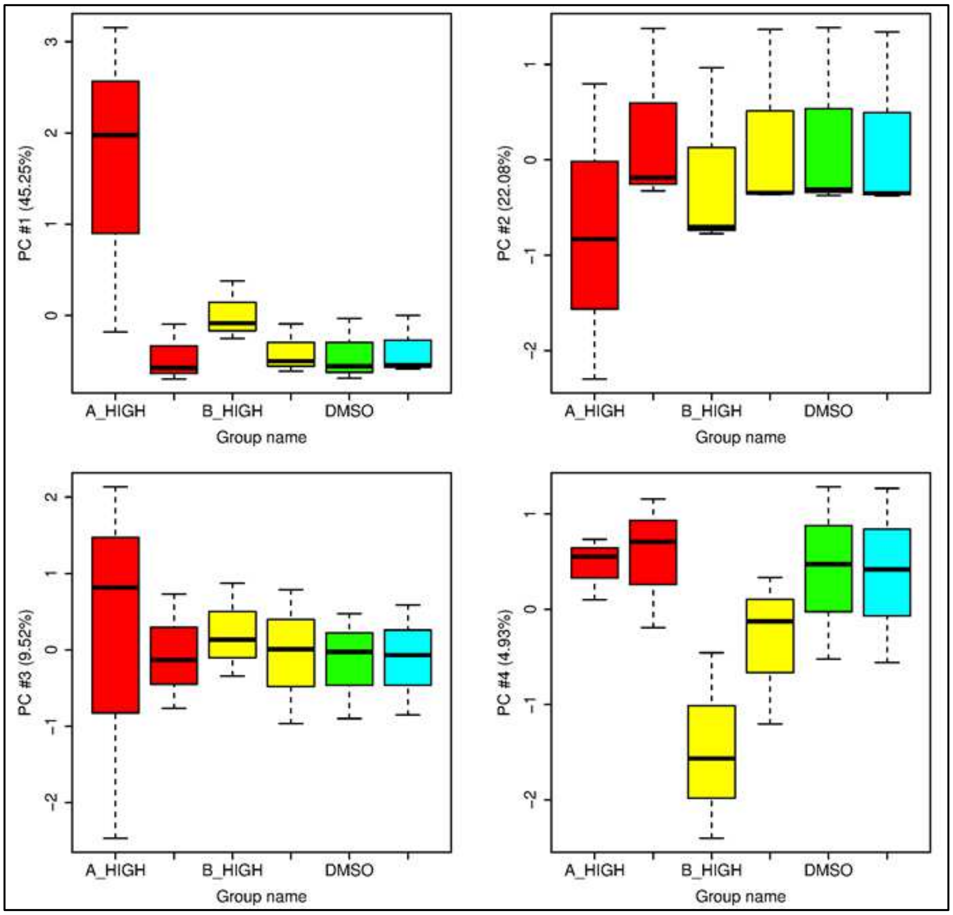
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample name	Group name	Dose Level	Replicate	Treatment	Concentration (uM)	Cell	Dose Type	Organism	Total.Reads	PC #1 (45.25%)	PC #2 (22.08%)	PC #3 (9.52%)	PC #4 (4.93%)	PC #5 (4.38%)
2	G01	DMSO	CONTROL	1	DMSO	0.2	Primary He	In-vitro	Homo sapi	1985806	-0.69006	-0.31176	0.473912	1.284648	-0.21001
3	D01	A_LOW	LOW	1	Treatment	330	Primary He	In-vitro	Homo sapi	1736699	-0.69654	-0.3251	0.727712	1.157918	0.384059
4	F01	A_HIGH	HIGH	1	Treatment	3300	Primary He	In-vitro	Homo sapi	1425191	-0.18324	-0.83086	0.813662	0.552036	1.410699
5	D02	B_LOW	LOW	1	Treatment	10	Primary He	In-vitro	Homo sapi	860821	-0.61071	-0.36199	0.786171	0.334549	0.742307
6	F02	B_HIGH	HIGH	1	Treatment	100	Primary He	In-vitro	Homo sapi	1257019	-0.25175	-0.77288	0.870294	-0.45669	1.402161
7	G02	UNTREATI	CONTROL	1	Untreated	0	Primary He	In-vitro	Homo sapi	969694	-0.5881	-0.35141	0.585865	1.269882	-0.16784

  - Note the complete file contains 18 PCs because there are 18 samples in the dataset and no “Split QC plots by” factor was used.

- There will be one Principal Components xml file per batch (where a batch is determined by the combination of “Split QC plots by” factors).
  - If no splitting is performed, then there will be one file.
  - If splitting is performed then, there will be more files, each with a number of PCs equal to the number of samples in the splitting batch.

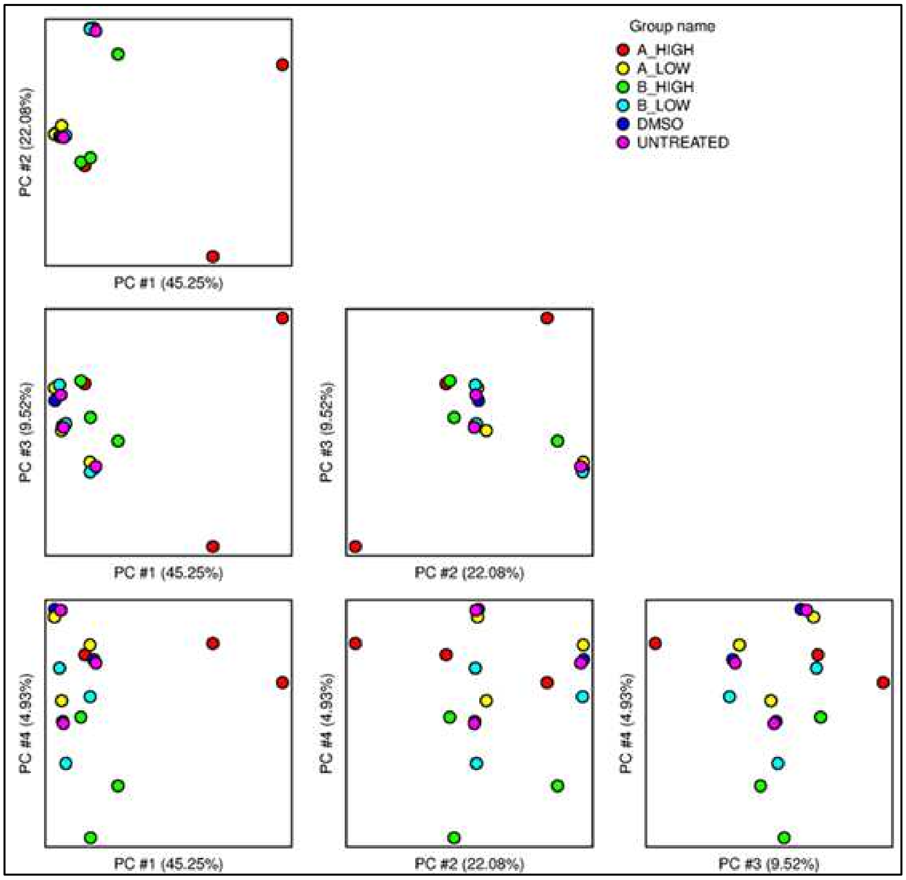
- Box-Whisker-Plots.rtf

Post Extrapolation: Principal Component Analysis
Principal Components



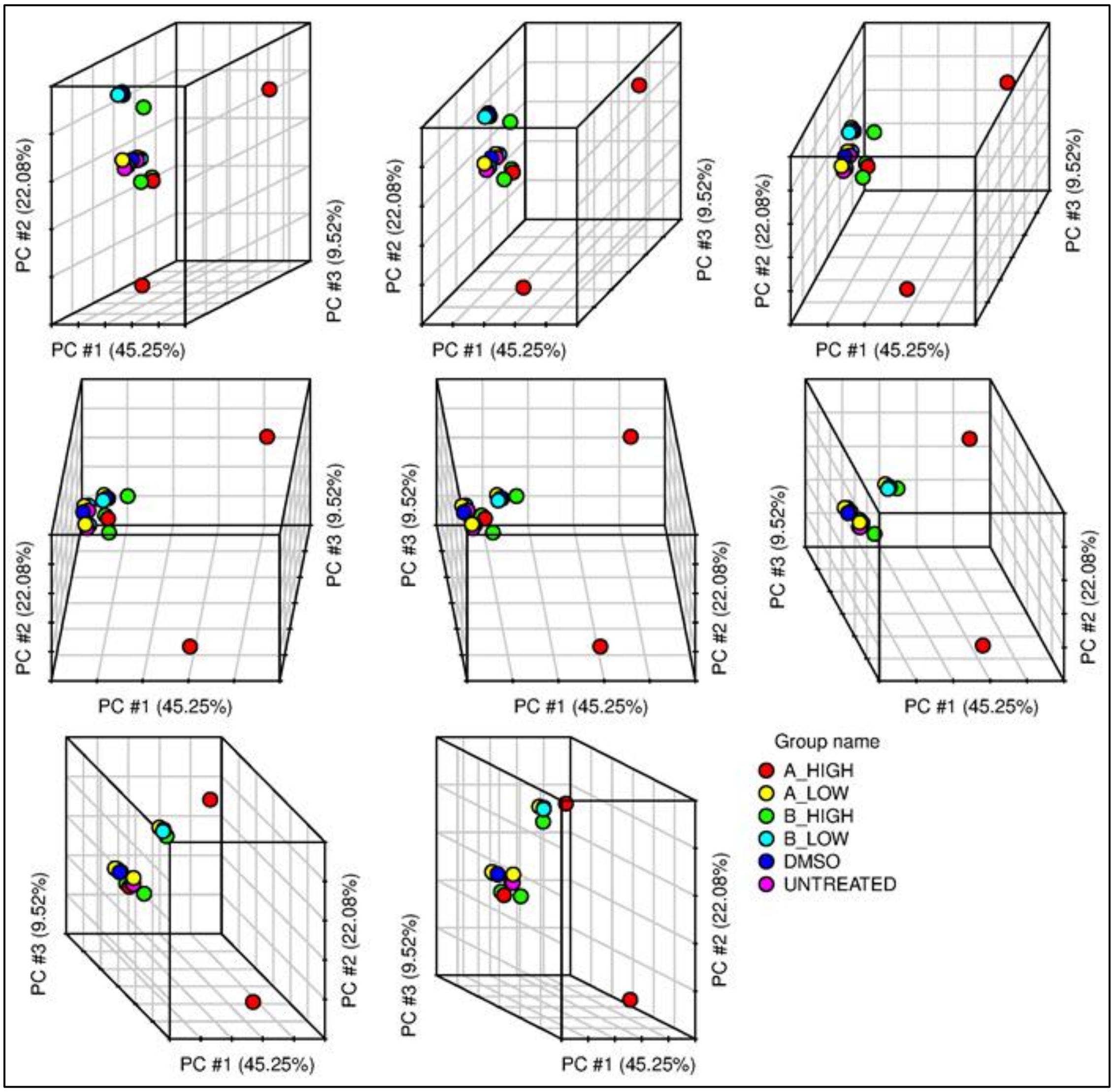
- Scatter-Plots-2D.rtf

Post Extrapolation: Principal Component Analysis
Principal Components



- Scatter-Plots-3D.rtf

Post Extrapolation: Principal Component Analysis
Principal Components



FAQ

1. **How does GeniE extrapolation work, and is the extrapolation method published?**

   A. Sciome's GeniE extrapolator was built using a training data set constructed from a large and diverse set of microarray experiments using Principle Component Regression [1] and is documented with the development of the human S1500+ in a recent publication [2]. Since this publication, additional RNA-seq training data [3] has been added, the extrapolator has been updated to output normalized gene expression, and the extrapolator has been trained on mouse and rat microarray and RNA-seq data. A new manuscript is in preparation discussing these updates as well as methodological improvements and the utility of the tool.

   [1] Jolliffe IT (1982). A Note on the Use of Principal Components in Regression. *Appl. Stat;* 31(3):300.

   [2] Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, Collins JB, Gerhold DL, Judson RS, Karmaus AL, Maull EA, Mendrick DL, Merrick BA, Sipes NS, Svoboda D, Paules RS (2018). A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLoS One*; 13-2, e0191105.

   [3] Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. "Massive mining of publicly available RNA-seq data from human and mouse." *Nature Communications*; 9(1366), doi:10.1038/s41467-018-03751-6.

2. **What species are currently supported?**

   A. GeniE currently supports human (*Homo sapiens*), rat (*Rattus norvegicus*), and mouse (*Mus musculus*). As additional S1500+ platforms are developed, we plan to expand species supported within GeniE.

3. **How many genes will be in the extrapolated results?**

   A. Currently GeniE can extrapolate to a total of 25,690 human genes, 16,682 rat genes, and 27,549 mouse genes.

4. **How long does it take to run?**

   A. For an example project consisting of 18 samples measured with the human S1500+ platform which includes 2,982 probes (corresponding to 2,821 genes), the job was processed in 11 minutes without downstream analysis and 18 minutes (MSIGDB) or 39 minutes (CREEDS) with DEG and DEP analysis for 3 contrasts. Please note, compute time may vary if large concurrent jobs are running on the server for a larger number of GeniE users.

5. **What if some of my probes do not match the manifest?**

   A. If any input probe ID is not in the species manifest (and no probe annotation file has been uploaded by the user), then the convention in the software is to assume the probe is written as [GeneSymbol]_[Numbers]. The part preceding the "_" in the input probe name will be assumed to be the name of the gene that the probe corresponds to.
      - If that name matches a known gene symbol, then the corresponding Entrez ID will be assigned to that probe and will be used in the extrapolation algorithm.
      - If that name does not match a known gene symbol, then the input data will be shown in "PROBE_LEVEL_EXPRESSION.TXT" but will not be used for extrapolation nor will it appear in "WHOLE_TRANSCRIPTOME_EXPRESSION.TXT".

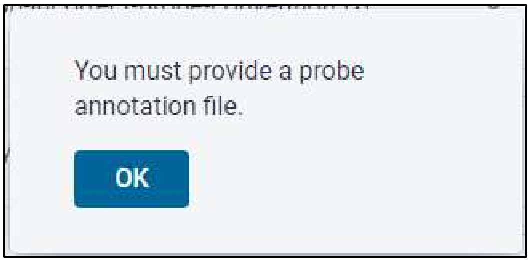
6. **If the species is listed as "Unknown" will GeniE still use the correct species training data for the extrapolation algorithm?**

   A. Yes! The initial validation check may not determine the species in cases where a probe annotation file was required, and many probe IDs do not match an S1500 species manifest. However, internally the software will be able to determine the species based on Entrez IDs for the genes in the probe annotation file if it has been correctly created by the user. This will lead to the correct training data being used and extrapolation of the species-specific genes.

7. **In what situations should I upload a probe annotation file?**

   A. There are two typical scenarios where a probe annotation file may be necessary:
      1) If you have any additional probes that are not found in the manifest (for example, probes specifically requested for your research that are not typically on the S1500 platform).
      2) If your input raw probe count file does not use the probe naming convention [GeneSymbol]_[Numbers].

   Failure to supply a probe annotation file in Case 1 will lead to omission of your additional genes in the extrapolation algorithm if they are not named with the above convention. In Case 2, after uploading your raw probe count file the software will immediately tell you that a probe annotation file is required.



You must provide a probe annotation file.

8. **What types of metadata belong in the sample annotation file?**

    A. Any metadata that you would like the output QC plots to be split or colored by should be included. This can range from things like cell line, exposure group (chemical, dose, duration), cytotoxicity data or anything else. The sample annotation file can contain additional rows that will not be used for data splitting and plot coloring, so there is no need to remove extra rows before uploading to GeniE.