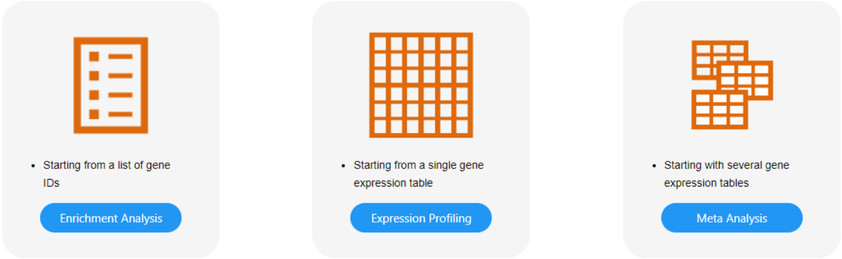
ExpressAnalyst - Tutorial
Starting from multiple tables

-- Comprehensive platform for gene expression and meta-analysis

Intro to ExpressAnalyst

- Web platform for the analysis of gene expression data and meta-analysis
  - Previously part of NetworkAnalyst
- Designed for bench researchers rather than specialized bioinformaticians
- Integrates data processing, statistical analysis and data visualization to support:
  - Data comparisons
  - Biological interpretation
  - Hypothesis generation



- Starting from a list of gene IDs

- Starting from a single gene expression table

- Starting with several gene expression tables

Gene list

Single matrix

Meta-analysis

Computer and browser requirements

- A modern web browser with JavaScript enabled
- Supported browsers include Chrome, Safari, Firefox, and Internet Explorer 9+
- For best performance and visualization, use:
    - Latest version of Google Chrome
    - A computer with at least 4GB of physical RAM
    - A 15-inch screen or bigger (larger is better)
- Browser must be WebGL enabled for 3D scatter visualization
- 50MB limit for data upload
    - ~300 samples for gene expression data with 20 000 genes

Goals for this tutorial

- Meta-analysis is a quantitative synthesis of results from multiple studies that test similar hypotheses
- Gene expression meta-analysis aims to identify molecular signatures and shared functional enrichment results to increase understanding of biological processes
- Requires advanced statistics and visualization strategies
- The goal of this tutorial is to complete a meta-analysis of expression profiles from 3 different studies:
    - Perform meta-analysis statistical tests
    - Visualize results in interactive heatmaps, Venn diagrams, and 3D PCA plots

Appropriate datasets

- The two main steps of a meta-analysis are:
  - Systematic literature review to identify studies that test the same hypothesis
  - Rigorous statistical analysis of the datasets using established methods
- NetworkAnalyst provides a platform for the second step
- For the meta-analysis to be a success, appropriate datasets should be used:
  - Study designs should compare the same experimental factors
  - Gene expression platforms should be comparable (i.e. studies should not be spread over > 10 years)
  - Relative similarity of host factors (i.e. species, tissue, sex, age etc.)

Data format

The data file can be tab delimited (.tab) or comma delimited (.csv)

Sample names

Meta-data

Needs to be consistent across datasets and also Only supports case-control Design (two factors)

#NAME	Sample1	Sample2	Sample3	Sample4	Sampl5	Sampl6	Sample7	Sample8
#CLASS	case	case	case	case	control	control	control	control
Gene1	-3.06	-2.25	-1.15	-6.64	0.4	1.08	1.22	1.02
Gene2	-1.36	-0.67	-0.17	-0.97	-2.32	-5.06	0.28	1.32
Gene3	1.61	-0.27	0.71	-0.62	0.14		0.11	0.98
Gene4	0.93	1.29	-0.23	-0.74	-2	-1.25	1.07	1.27

…

Gene/probe ids

https://www.expressanalyst.ca/ExpressAnalyst/resources/data/test/estrogen.txt

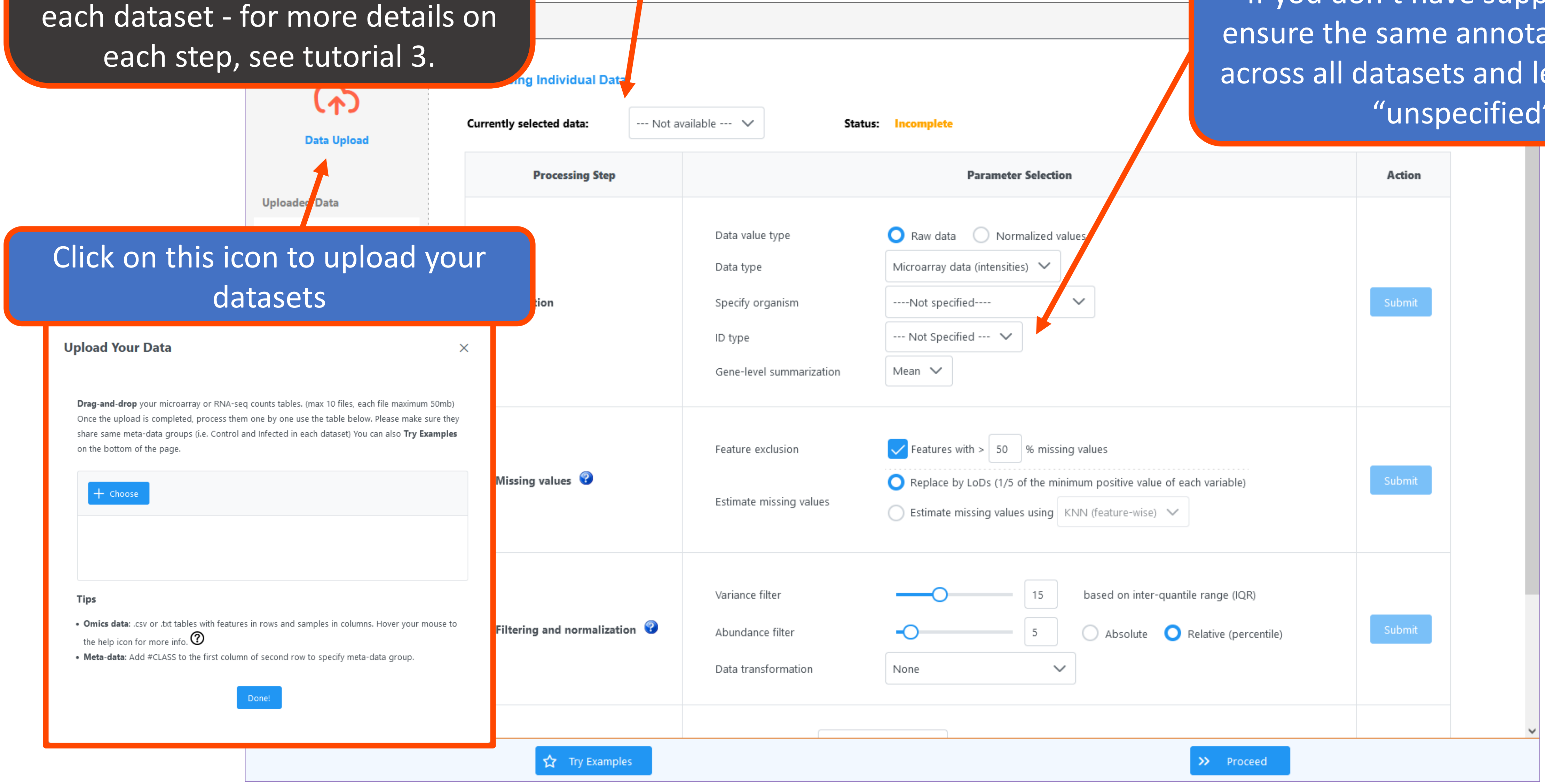
Upload data

The first step is to upload and process all of your individual datasets. This repeats the steps of a single gene expression table for each dataset - for more details on each step, see tutorial 3.

Select between different uploaded datasets using this dropdown

If you don’t have supported IDs, ensure the same annotation is used across all datasets and leave ID type “unspecified”

Click on this icon to upload your datasets



Data Upload

Uploaded Data

ing Individual Data

Currently selected data: --- Not available ---

Status: Incomplete

Processing Step	Parameter Selection	Action
ion	Data value type: ● Raw data ○ Normalized values Data type: Microarray data (intensities) Specify organism: ----Not specified---- ID type: --- Not Specified --- Gene-level summarization: Mean	Submit
Missing values	Feature exclusion: ☑ Features with > 50 % missing values Estimate missing values: ● Replace by LoDs (1/5 of the minimum positive value of each variable) ○ Estimate missing values using KNN (feature-wise)	Submit
Filtering and normalization	Variance filter: 15 based on inter-quantile range (IQR) Abundance filter: 5 ○ Absolute ● Relative (percentile) Data transformation: None	Submit

Try Examples

Proceed

Upload Your Data

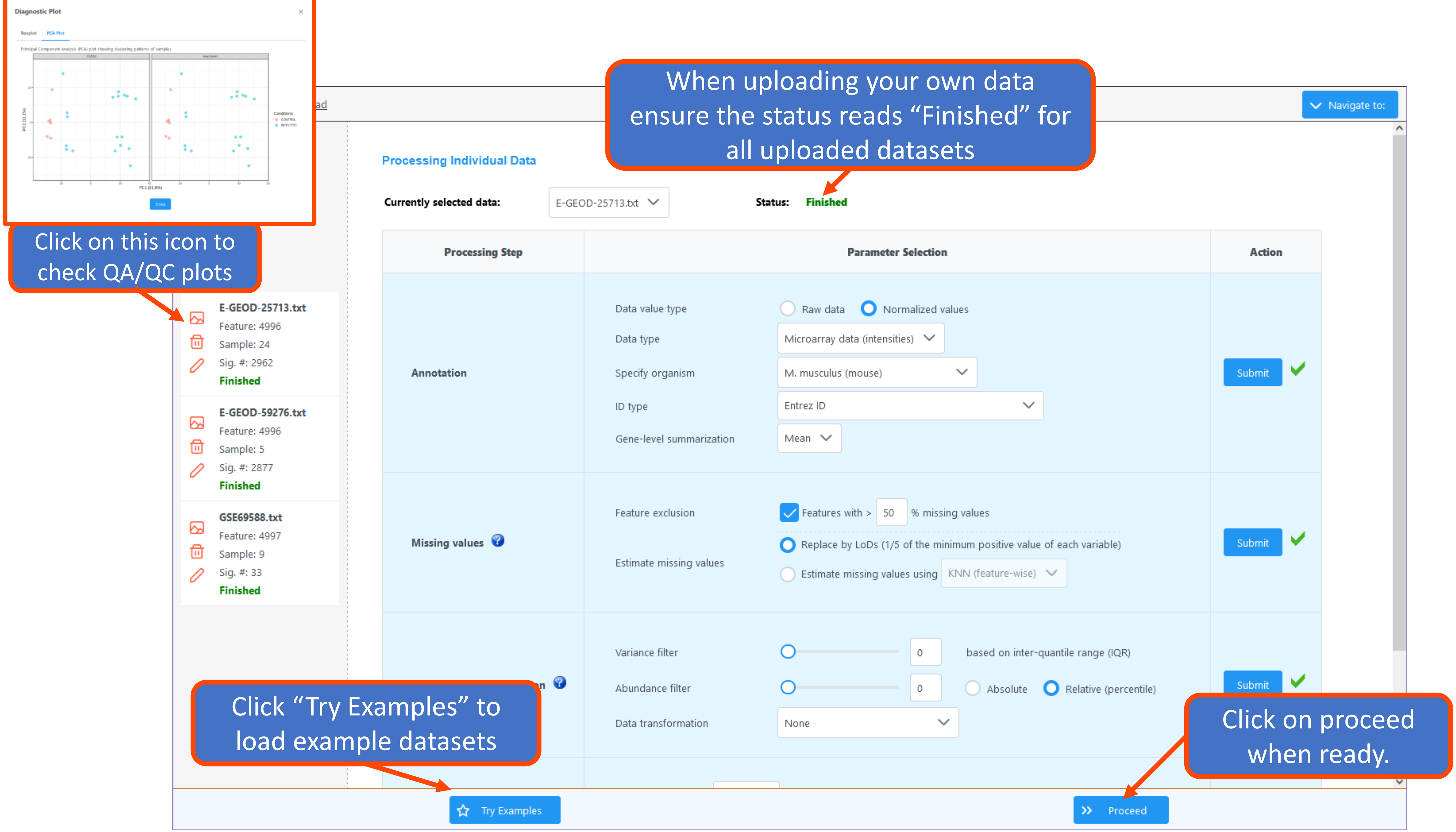
Drag-and-drop your microarray or RNA-seq counts tables. (max 10 files, each file maximum 50mb) Once the upload is completed, process them one by one use the table below. Please make sure they share same meta-data groups (i.e. Control and Infected in each dataset) You can also **Try Examples** on the bottom of the page.

+ Choose

Tips

- **Omics data**: .csv or .txt tables with features in rows and samples in columns. Hover your mouse to the help icon for more info. ?
- **Meta-data**: Add #CLASS to the first column of second row to specify meta-data group.

Done!



Principal Component Analysis (PCA) plot showing clustering patterns of samples

When uploading your own data ensure the status reads “Finished” for all uploaded datasets

Click on this icon to check QA/QC plots

Click “Try Examples” to load example datasets

Click on proceed when ready.

Processing Individual Data

Currently selected data: E-GEOD-25713.txt

Status: Finished

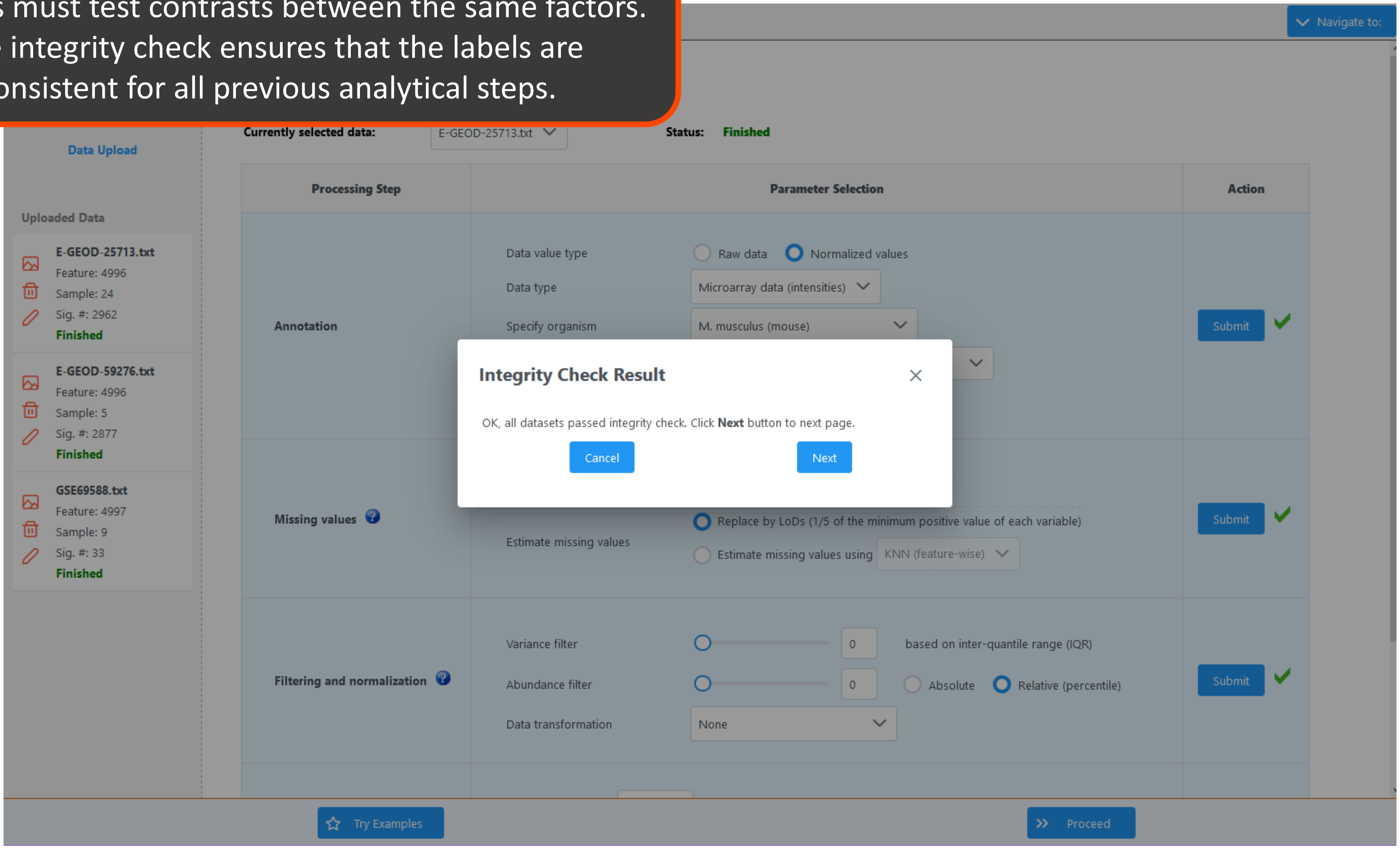
E-GEOD-25713.txt
Feature: 4996
Sample: 24
Sig. #: 2962
Finished

E-GEOD-59276.txt
Feature: 4996
Sample: 5
Sig. #: 2877
Finished

GSE69588.txt
Feature: 4997
Sample: 9
Sig. #: 33
Finished

Processing Step	Parameter Selection	Action
Annotation	Data value type: ○ Raw data ● Normalized values Data type: Microarray data (intensities) Specify organism: M. musculus (mouse) ID type: Entrez ID Gene-level summarization: Mean	Submit ✓
Missing values	Feature exclusion: ☑ Features with > 50 % missing values Estimate missing values: ● Replace by LoDs (1/5 of the minimum positive value of each variable) ○ Estimate missing values using KNN (feature-wise)	Submit ✓
n	Variance filter: 0 based on inter-quantile range (IQR) Abundance filter: 0 ○ Absolute ● Relative (percentile) Data transformation: None	Submit ✓

For a meta-analysis to be done properly, the individual analyses must test contrasts between the same factors. The integrity check ensures that the labels are consistent for all previous analytical steps.



Currently selected data: E-GEOD-25713.txt   Status: Finished

E-GEOD-25713.txt
Feature: 4996
Sample: 24
Sig. #: 2962
Finished

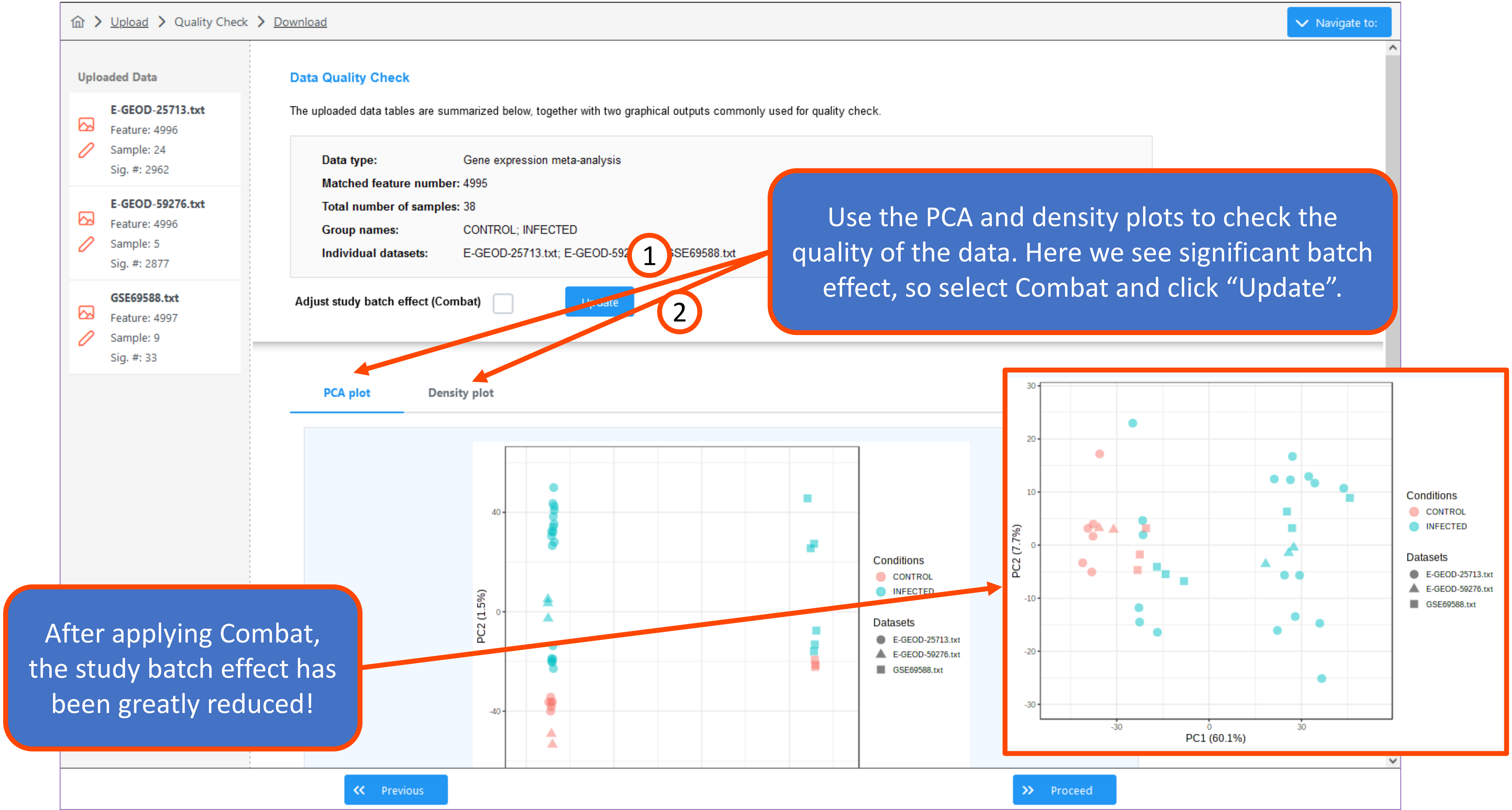
E-GEOD-59276.txt
Feature: 4996
Sample: 5
Sig. #: 2877
Finished

GSE69588.txt
Feature: 4997
Sample: 9
Sig. #: 33
Finished

Processing Step	Parameter Selection	Action
Annotation	Data value type: ○ Raw data ● Normalized values Data type: Microarray data (intensities) Specify organism: M. musculus (mouse)	Submit ✓
Missing values	Estimate missing values: ● Replace by LoDs (1/5 of the minimum positive value of each variable) ○ Estimate missing values using KNN (feature-wise)	Submit ✓
Filtering and normalization	Variance filter: 0 based on inter-quantile range (IQR) Abundance filter: 0 ○ Absolute ● Relative (percentile) Data transformation: None	Submit ✓

Integrity Check Result

OK, all datasets passed integrity check. Click **Next** button to next page.



Upload > Quality Check > Download

Data Quality Check

The uploaded data tables are summarized below, together with two graphical outputs commonly used for quality check.

Data type:	Gene expression meta-analysis
Matched feature number:	4995
Total number of samples:	38
Group names:	CONTROL; INFECTED
Individual datasets:	E-GEOD-25713.txt; E-GEOD-592… GSE69588.txt

Adjust study batch effect (Combat) □ Update

Uploaded Data

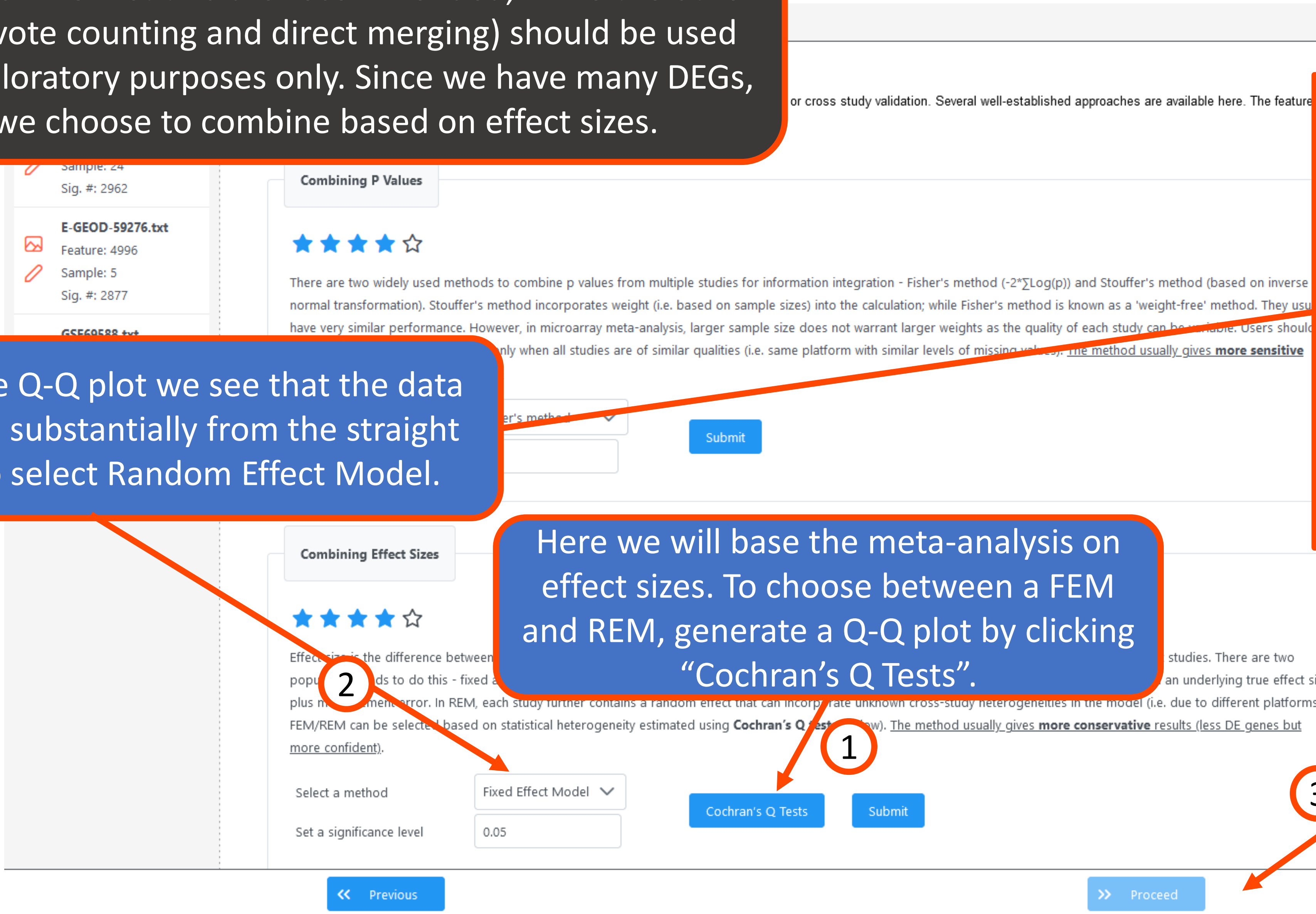
E-GEOD-25713.txt	Feature: 4996	Sample: 24	Sig. #: 2962
E-GEOD-59276.txt	Feature: 4996	Sample: 5	Sig. #: 2877
GSE69588.txt	Feature: 4997	Sample: 9	Sig. #: 33

Use the PCA and density plots to check the quality of the data. Here we see significant batch effect, so select Combat and click “Update”.

After applying Combat, the study batch effect has been greatly reduced!

Gene-level meta-analysis

ExpressAnalyst has four approaches for gene-level meta-analysis. The first two are recommended, while the other two (vote counting and direct merging) should be used for exploratory purposes only. Since we have many DEGs, we choose to combine based on effect sizes.



or cross study validation. Several well-established approaches are available here. The feature

Combining P Values

★★★★☆

There are two widely used methods to combine p values from multiple studies for information integration - Fisher's method (-2*∑Log(p)) and Stouffer's method (based on inverse normal transformation). Stouffer's method incorporates weight (i.e. based on sample sizes) into the calculation; while Fisher's method is known as a 'weight-free' method. They usu have very similar performance. However, in microarray meta-analysis, larger sample size does not warrant larger weights as the quality of each study can be variable. Users should nly when all studies are of similar qualities (i.e. same platform with similar levels of missing values). The method usually gives **more sensitive**

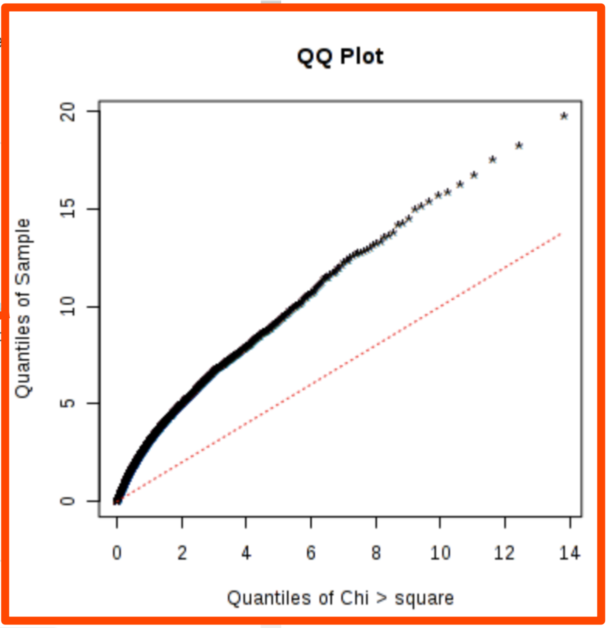
Combining Effect Sizes

★★★★☆

Effect size is the difference between studies. There are two popu ds to do this - fixed a an underlying true effect size plus m ment error. In REM, each study further contains a random effect that can incorporate unknown cross-study heterogeneities in the model (i.e. due to different platforms). FEM/REM can be selected based on statistical heterogeneity estimated using **Cochran's Q test** w). The method usually gives **more conservative** results (less DE genes but more confident).

Select a method: Fixed Effect Model

Set a significance level: 0.05

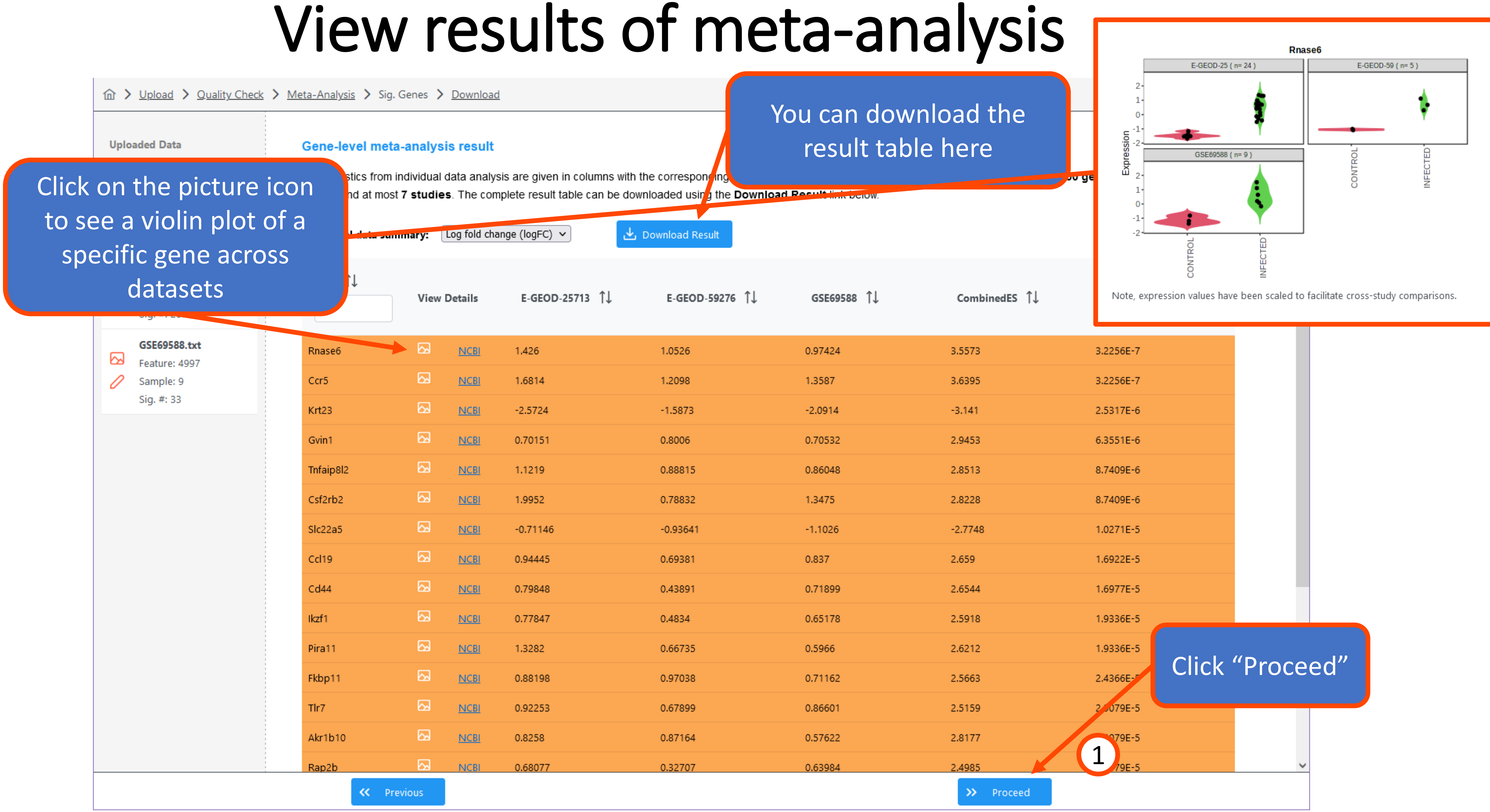


From the Q-Q plot we see that the data deviates substantially from the straight line, so select Random Effect Model.

Here we will base the meta-analysis on effect sizes. To choose between a FEM and REM, generate a Q-Q plot by clicking "Cochran's Q Tests".

Click "Proceed"

View results of meta-analysis



GSE69588.txt
Feature: 4997
Sample: 9
Sig. #: 33

Gene-level meta-analysis result

...stics from individual data analysis are given in columns with the correspon ing ... nd at most **7 studies**. The complete result table can be downloaded using the **Download Result** link below.

...summary: Log fold change (logFC)

Download Result

Note, expression values have been scaled to facilitate cross-study comparisons.

	View Details		E-GEOD-25713	E-GEOD-59276	GSE69588	CombinedES	
Rnase6	🖼	NCBI	1.426	1.0526	0.97424	3.5573	3.2256E-7
Ccr5	🖼	NCBI	1.6814	1.2098	1.3587	3.6395	3.2256E-7
Krt23	🖼	NCBI	-2.5724	-1.5873	-2.0914	-3.141	2.5317E-6
Gvin1	🖼	NCBI	0.70151	0.8006	0.70532	2.9453	6.3551E-6
Tnfaip8l2	🖼	NCBI	1.1219	0.88815	0.86048	2.8513	8.7409E-6
Csf2rb2	🖼	NCBI	1.9952	0.78832	1.3475	2.8228	8.7409E-6
Slc22a5	🖼	NCBI	-0.71146	-0.93641	-1.1026	-2.7748	1.0271E-5
Ccl19	🖼	NCBI	0.94445	0.69381	0.837	2.659	1.6922E-5
Cd44	🖼	NCBI	0.79848	0.43891	0.71899	2.6544	1.6977E-5
Ikzf1	🖼	NCBI	0.77847	0.4834	0.65178	2.5918	1.9336E-5
Pira11	🖼	NCBI	1.3282	0.66735	0.5966	2.6212	1.9336E-5
Fkbp11	🖼	NCBI	0.88198	0.97038	0.71162	2.5663	2.4366E-
Tlr7	🖼	NCBI	0.92253	0.67899	0.86601	2.5159	2.0079E-5
Akr1b10	🖼	NCBI	0.8258	0.87164	0.57622	2.8177	079E-5
Rap2b	🖼	NCBI	0.68077	0.32707	0.63984	2.4985	79E-5

Click on the picture icon to see a violin plot of a specific gene across datasets

You can download the result table here

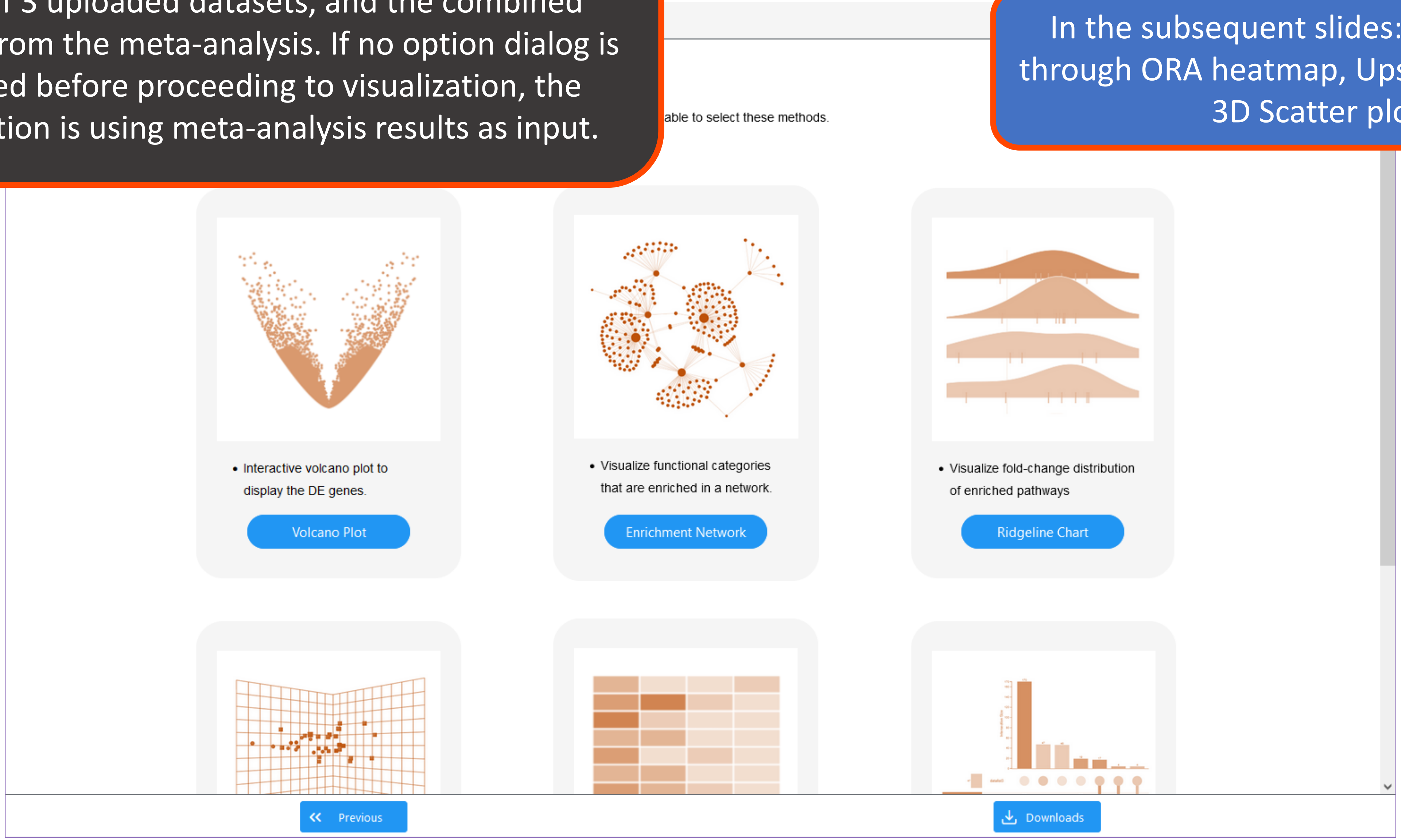
Click “Proceed”

1

Analysis overview

For the visual analytics of the meta-analysis results, there are up to 4 datasets to work with: the significant genes of 3 uploaded datasets, and the combined statistics from the meta-analysis. If no option dialog is displayed before proceeding to visualization, the visualization is using meta-analysis results as input.

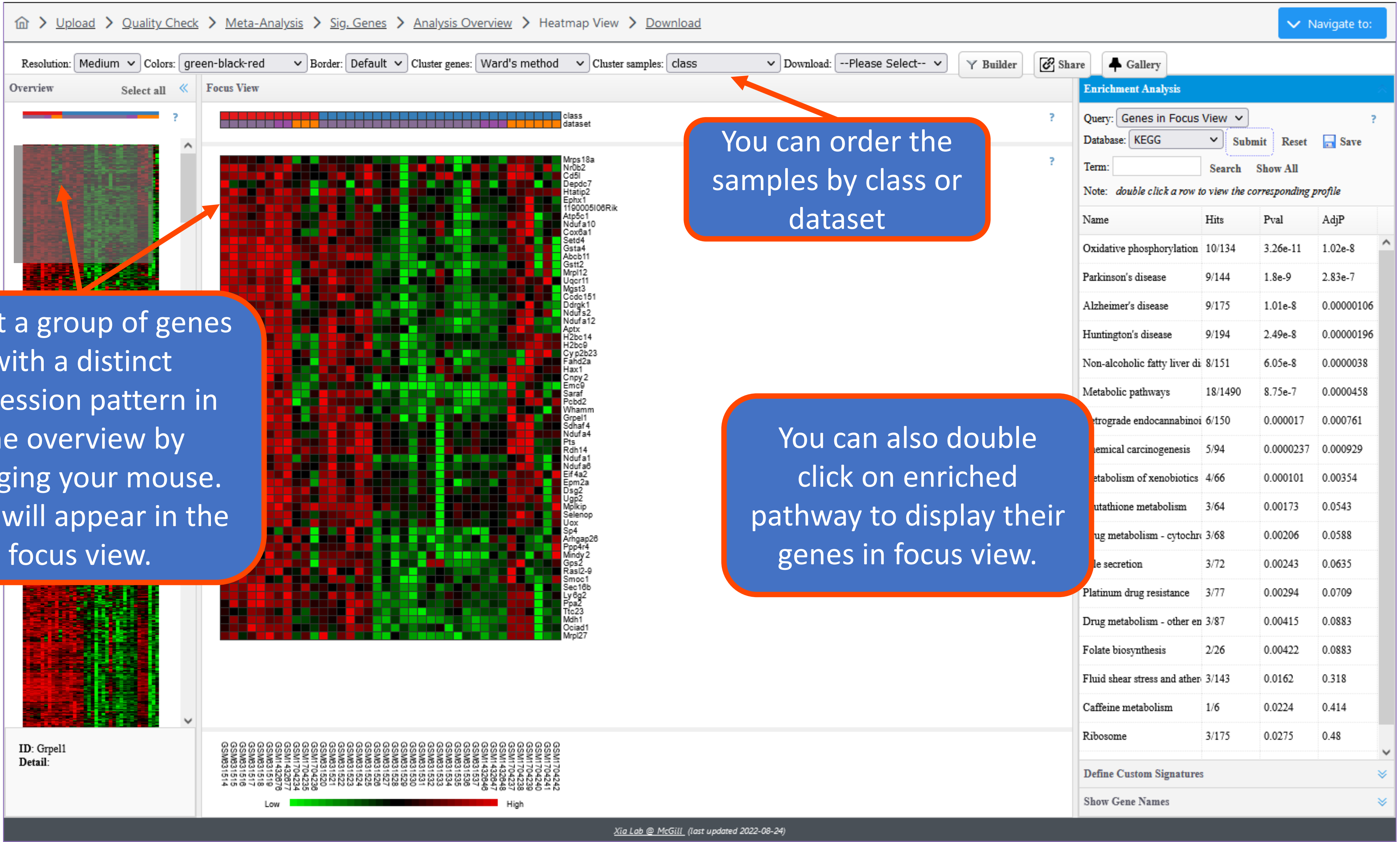
In the subsequent slides: we are going through ORA heatmap, Upset diagram and 3D Scatter plot.



able to select these methods.

- Interactive volcano plot to display the DE genes.
- Visualize functional categories that are enriched in a network.
- Visualize fold-change distribution of enriched pathways

ORA heatmap



Resolution: Medium | Colors: green-black-red | Border: Default | Cluster genes: Ward's method | Cluster samples: class | Download: --Please Select--

ID: Grpel1
Detail:

Query: Genes in Focus View
Database: KEGG
Term:
Note: *double click a row to view the corresponding profile*

Name	Hits	Pval	AdjP
Oxidative phosphorylation	10/134	3.26e-11	1.02e-8
Parkinson's disease	9/144	1.8e-9	2.83e-7
Alzheimer's disease	9/175	1.01e-8	0.00000106
Huntington's disease	9/194	2.49e-8	0.00000196
Non-alcoholic fatty liver di	8/151	6.05e-8	0.0000038
Metabolic pathways	18/1490	8.75e-7	0.0000458
etrograde endocannabinoi	6/150	0.000017	0.000761
emical carcinogenesis	5/94	0.0000237	0.000929
etabolism of xenobiotics	4/66	0.000101	0.00354
utathione metabolism	3/64	0.00173	0.0543
ug metabolism - cytochr	3/68	0.00206	0.0588
le secretion	3/72	0.00243	0.0635
Platinum drug resistance	3/77	0.00294	0.0709
Drug metabolism - other en	3/87	0.00415	0.0883
Folate biosynthesis	2/26	0.00422	0.0883
Fluid shear stress and ather	3/143	0.0162	0.318
Caffeine metabolism	1/6	0.0224	0.414
Ribosome	3/175	0.0275	0.48

Xia Lab @ McGill (last updated 2022-08-24)

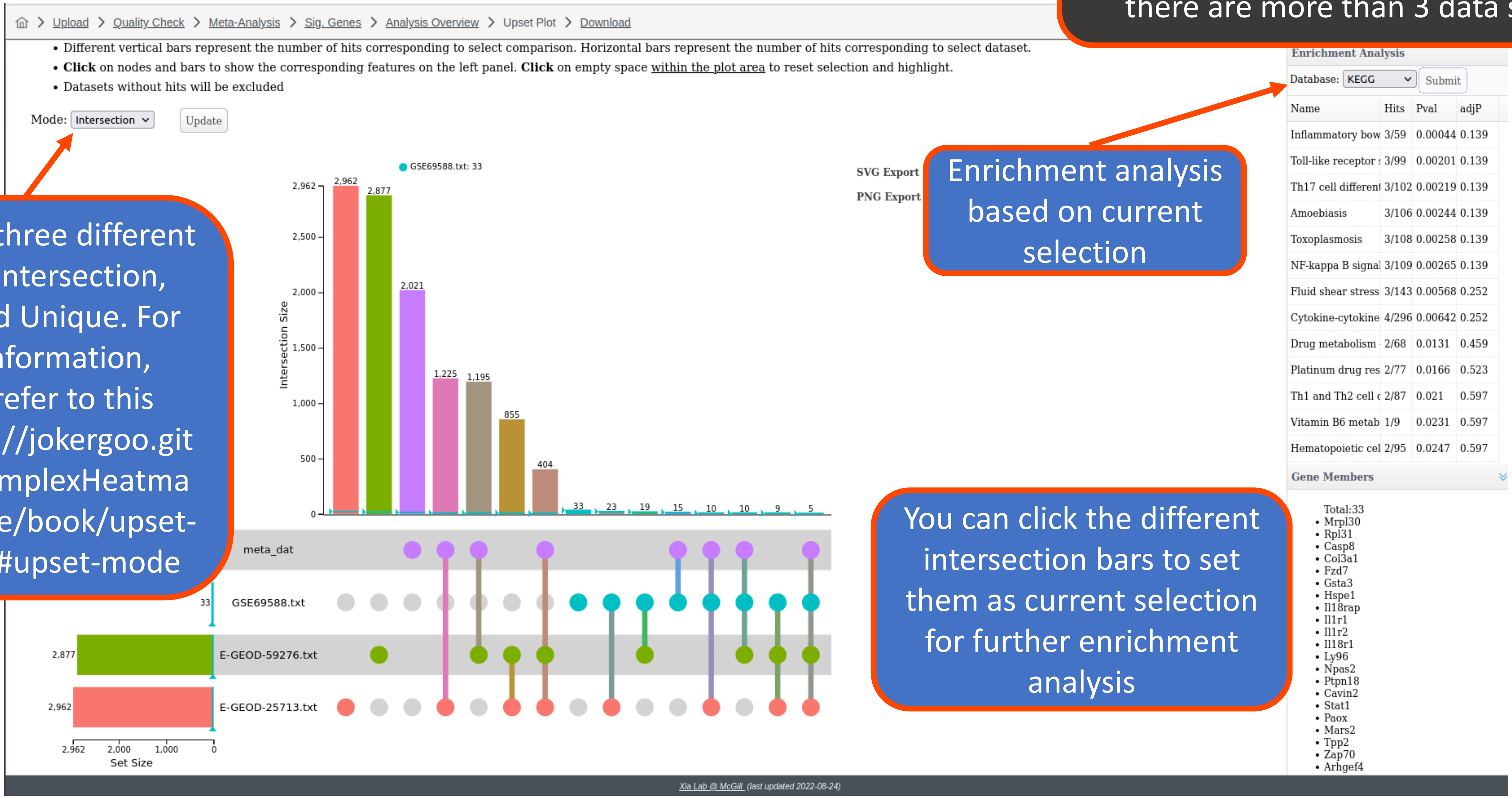
You can order the samples by class or dataset

Select a group of genes with a distinct expression pattern in the overview by dragging your mouse. They will appear in the focus view.

You can also double click on enriched pathway to display their genes in focus view.

Upset diagram

Upset diagram is an alternative to Venn diagram, its strength lies on its ability to show intersecting data when there are more than 3 data sets.



- Different vertical bars represent the number of hits corresponding to select comparison. Horizontal bars represent the number of hits corresponding to select dataset.
- **Click** on nodes and bars to show the corresponding features on the left panel. **Click** on empty space within the plot area to reset selection and highlight.
- Datasets without hits will be excluded

Mode: Intersection

Enrichment Analysis

Database: KEGG Submit

Name	Hits	Pval	adjP
Inflammatory bow	3/59	0.00044	0.139
Toll-like receptor :	3/99	0.00201	0.139
Th17 cell different	3/102	0.00219	0.139
Amoebiasis	3/106	0.00244	0.139
Toxoplasmosis	3/108	0.00258	0.139
NF-kappa B signal	3/109	0.00265	0.139
Fluid shear stress	3/143	0.00568	0.252
Cytokine-cytokine	4/296	0.00642	0.252
Drug metabolism ·	2/68	0.0131	0.459
Platinum drug res	2/77	0.0166	0.523
Th1 and Th2 cell c	2/87	0.021	0.597
Vitamin B6 metab	1/9	0.0231	0.597
Hematopoietic cel	2/95	0.0247	0.597

Gene Members

Total:33

- Mrpl30
- Rpl31
- Casp8
- Col3a1
- Fzd7
- Gsta3
- Hspe1
- Il18rap
- Il1r1
- Il1r2
- Il18r1
- Ly96
- Npas2
- Ptpn18
- Cavin2
- Stat1
- Paox
- Mars2
- Tpp2
- Zap70
- Arhgef4

Xia Lab @ McGill (last updated 2022-08-24)

There are three different modes: Intersection, Union and Unique. For more information, please refer to this link: https://jokergoo.github.io/ComplexHeatmap-reference/book/upset-plot.html#upset-mode

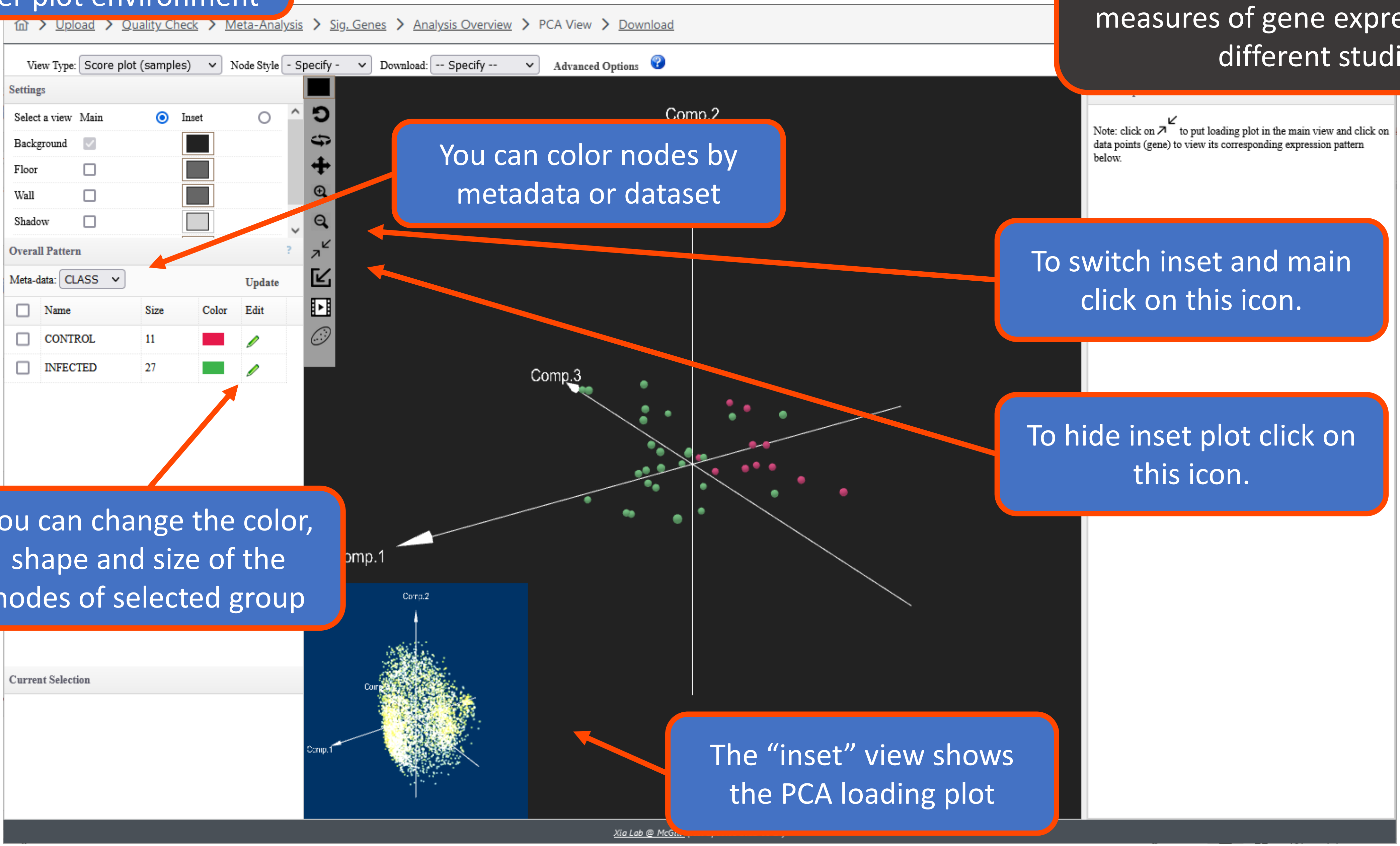
Enrichment analysis based on current selection

You can click the different intersection bars to set them as current selection for further enrichment analysis

3D Scatter Plot

The settings panel is useful for customizing the scatter plot environment

3D PCA plots are useful for visualizing the variance in whole-transcriptome measures of gene expression across different studies.



🏠 > Upload > Quality Check > Meta-Analysis > Sig. Genes > Analysis Overview > PCA View > Download

View Type: Score plot (samples) Node Style - Specify - Download: -- Specify -- Advanced Options

Settings

Select a view Main ● Inset ○

Background ☑

Floor □

Wall □

Shadow □

Overall Pattern

Meta-data: CLASS Update

□	Name	Size	Color	Edit
□	CONTROL	11	🟥	✎
□	INFECTED	27	🟩	✎

Current Selection

Note: click on ↗ to put loading plot in the main view and click on data points (gene) to view its corresponding expression pattern below.

You can color nodes by metadata or dataset

To switch inset and main click on this icon.

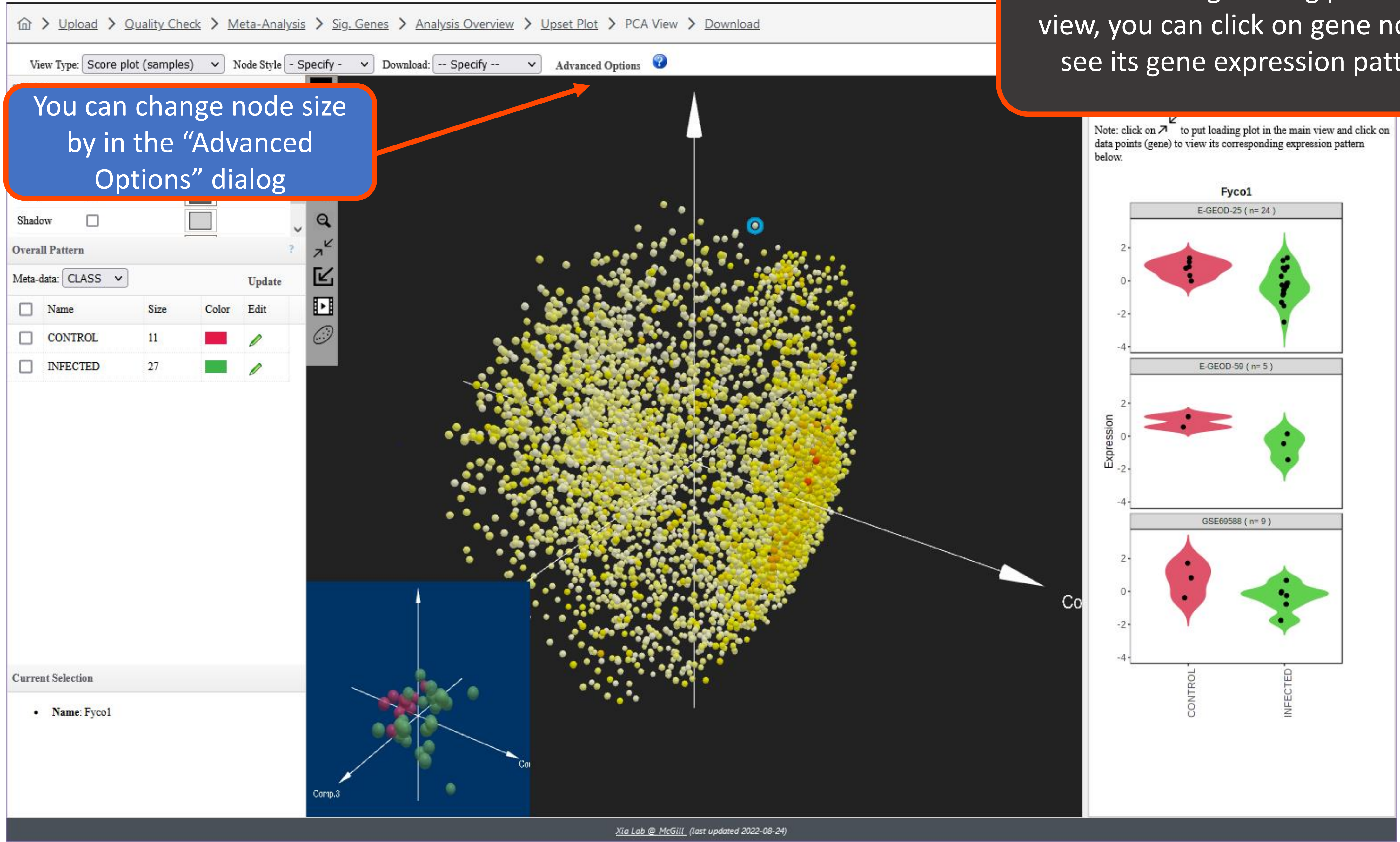
To hide inset plot click on this icon.

You can change the color, shape and size of the nodes of selected group

The “inset” view shows the PCA loading plot

After switching loading plot to main view, you can click on gene node to see its gene expression pattern.

You can change node size by in the “Advanced Options” dialog



⌂ > Upload > Quality Check > Meta-Analysis > Sig. Genes > Analysis Overview > Upset Plot > PCA View > Download

View Type: Score plot (samples) | Node Style - Specify - | Download: -- Specify -- | Advanced Options

Shadow □

Overall Pattern

Meta-data: CLASS | Update

□	Name	Size	Color	Edit
□	CONTROL	11	🟥	✎
□	INFECTED	27	🟩	✎

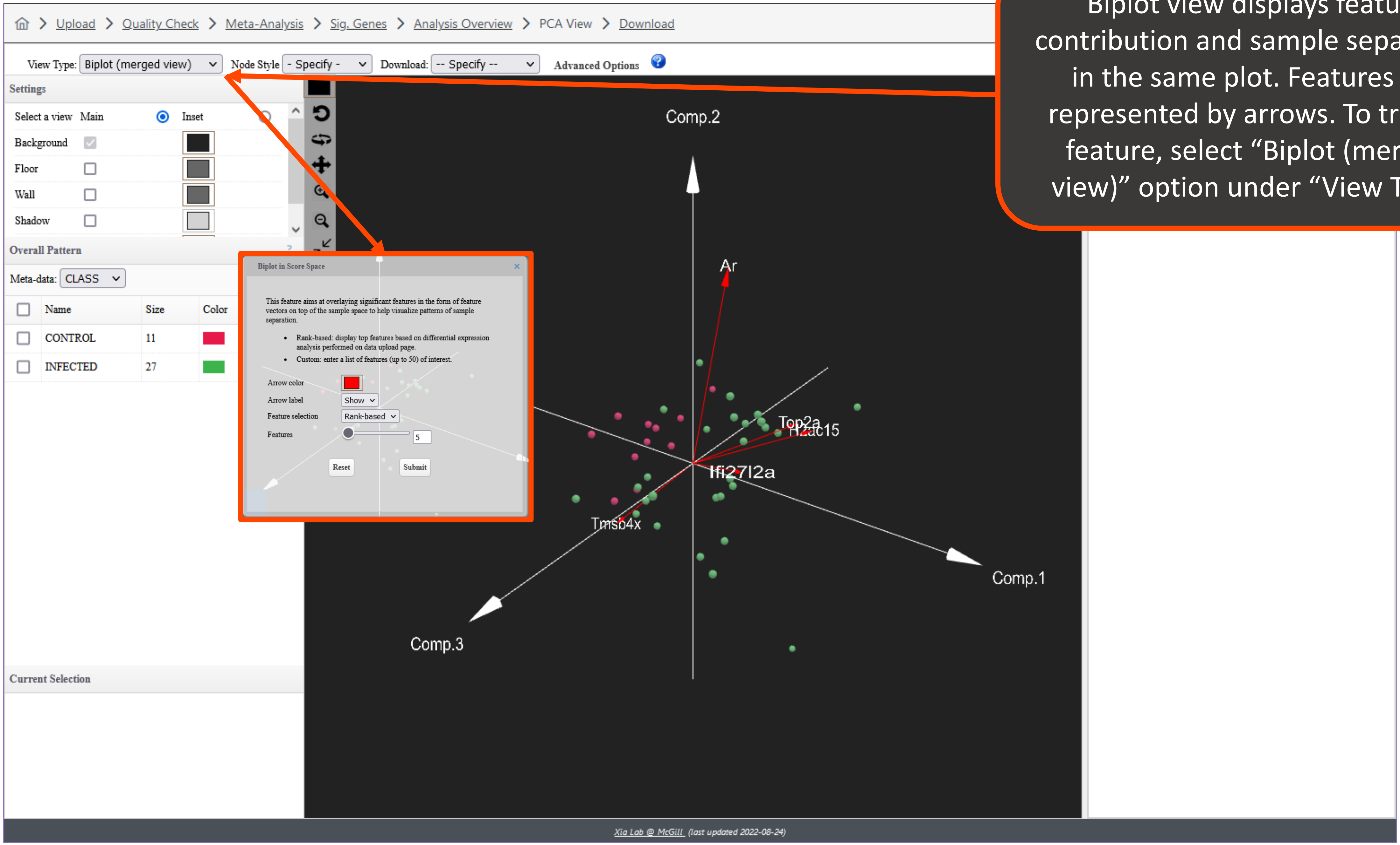
Current Selection

- **Name**: Fyco1

Note: click on ↗↙ to put loading plot in the main view and click on data points (gene) to view its corresponding expression pattern below.

Xia Lab @ McGill (last updated 2022-08-24)

Biplot view displays feature contribution and sample separation in the same plot. Features are represented by arrows. To try this feature, select “Biplot (merged view)” option under “View Type”



Upload > Quality Check > Meta-Analysis > Sig. Genes > Analysis Overview > PCA View > Download

View Type: Biplot (merged view)   Node Style: - Specify -   Download: -- Specify --   Advanced Options

Settings

Select a view: Main ● Inset ○

Background	☑	⬛
Floor	□	⬛
Wall	□	⬛
Shadow	□	⬜

Overall Pattern

Meta-data: CLASS

□	Name	Size	Color
□	CONTROL	11	🟥
□	INFECTED	27	🟩

Current Selection

Biplot in Score Space

This feature aims at overlaying significant features in the form of feature vectors on top of the sample space to help visualize patterns of sample separation.

- Rank-based: display top features based on differential expression analysis performed on data upload page.
- Custom: enter a list of features (up to 50) of interest.

Arrow color: 🟥
Arrow label: Show
Feature selection: Rank-based
Features: 5

Xia Lab @ McGill (last updated 2022-08-24)

The End

For more information, visit Tutorials, Resources

and Contact pages on www.expressanalyst.ca

Also visit our forum for FAQs on www.omicsforum.ca