

Statistical Analysis User Guide

The “Statistical Analysis” tool enables users to quantitatively compare counts of functional genes between groups of genomes or metagenomes. To get started, user must select and save genomes or metagenomes of interest as discrete “Genome Sets” in workspace. Genome sets can be created from the genome cart using “Upload & Export & Save” (see below). *Please refer to the [IMG workspace user Guide](#) for more details on other workspace functions and options.*

Genome Cart

Only a maximum of 20000 genomes can be in cart.

24 genome(s) in cart

Genomes in Cart | **Upload & Export & Save**

Upload Genome Cart

The Genome Cart is used for genomes already in IMG. Only previously exported IMG genomes or IMG genomes saved as a genome set to the workspace, can be uploaded. [To upload private genomes](#), you must submit your data to IMG through the [submission site](#).

You may upload a genome cart from a tab-delimited file.
Uploading a genome cart will add the genomes to the list of selected genomes.
The file should have a column header 'taxon_oid'.
(This file may initially be obtained by exporting genomes from [Genome Browser](#) to Excel or by using the [Export Genomes](#) section below)

File to upload:
 No file chosen

Export Genomes

You may select genomes from the cart to export.

Download selected genomes via [JGI Portal](#)

Save Genomes to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save **selected genomes** to [My Workspace](#).
(Special characters in file name will be removed and spaces converted to _)

Save to File name:

Append to the following genome set:

Replacing the following genome set:
Actinos_edited_4James ▼

Save Selected to Workspace

Using the Statistical Analysis tool in Workspace:

Go to Workspace> Genome Sets> Tab over to “Statistical Analysis”

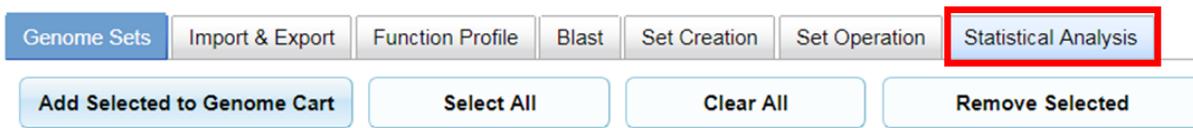
Genome Workspace List

Genome Cart

You have [24 Genome\(s\)](#) in your cart.

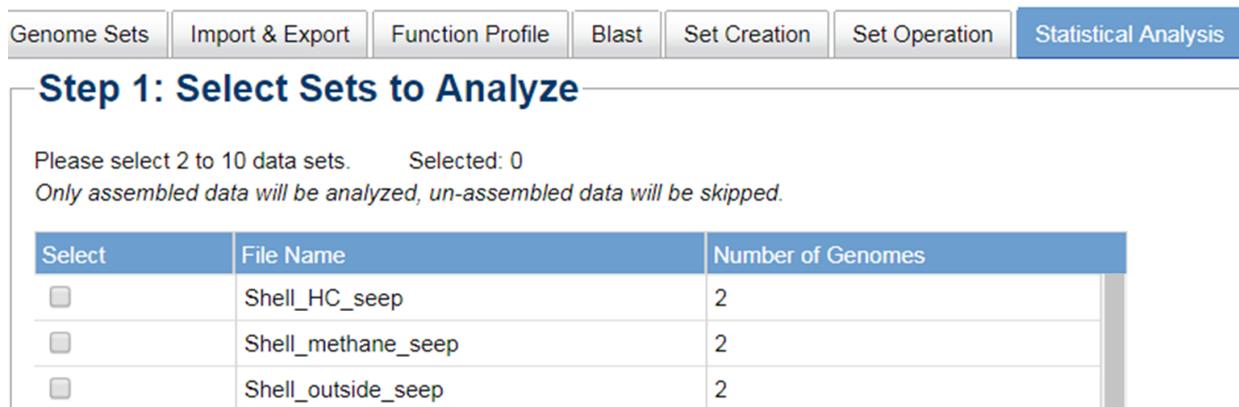
Genome Sets List

Set(s) or file(s) shared from all groups are displayed. Please go to [Preferences](#) to change the sharing display options.



The screenshot shows a navigation bar with tabs: Genome Sets, Import & Export, Function Profile, Blast, Set Creation, Set Operation, and Statistical Analysis. The 'Statistical Analysis' tab is highlighted with a red box. Below the navigation bar are four buttons: Add Selected to Genome Cart, Select All, Clear All, and Remove Selected.

Step 1: Choose Genome Sets to compare. A maximum of 10 sets or groups may be selected.



The screenshot shows the 'Statistical Analysis' tab selected in the navigation bar. Below the navigation bar is a section titled 'Step 1: Select Sets to Analyze'. It contains the text: 'Please select 2 to 10 data sets. Selected: 0' and 'Only assembled data will be analyzed, un-assembled data will be skipped.' Below this text is a table with three columns: Select, File Name, and Number of Genomes.

Select	File Name	Number of Genomes
<input type="checkbox"/>	Shell_HC_seep	2
<input type="checkbox"/>	Shell_methane_seep	2
<input type="checkbox"/>	Shell_outside_seep	2

Step 2: Choose a feature for statistical comparison – for example, choosing “Pfam” means counts of genes assigned to Pfams will be compared across groups. Choosing “Pfam Category” means gene counts of all Pfams belonging to a “category” are summed (to aggregate) before comparisons are performed. For metagenomes-based analyses only, an indirect comparison of taxonomic composition is possible. For this feature, the numbers of genes assigned to a lineage (based on predicted lineage of the scaffold that the genes reside on) are compared across the metagenome sets. *IMG scaffold lineage is predicted based on the last common ancestor of LAST hits (against IMG-NR isolates database) of the genes on the scaffold, where at least 30% of the genes have LAST hits against the database.* Again, for metagenomes only, step 2b (Measurement) allows user to select “estimated gene copies” so that gene counts reflect copy number based on their abundance in the sample – specifically, the gene count is multiplied by average read depth of the scaffold that the gene resides on. CAUTION: user must verify that read depth information is available for all their chosen samples in all groups.

Step 2: Select Features

a) Feature Type

By Function

COG

PFam

KO

By Function Category

COG Category

PFam Category

KEGG Modules

By Taxonomy (Only Metagenomes)

Class

Family

Genus

or

or

b) Measurement

Gene Count

Estimated Gene Copies (If no reads data default is 1)

Step 3: The default statistical method is determined by the number of sets chosen for comparison as well as the number of individual genomes or metagenomes within those sets. The decision tree for choice of default statistical test as well as general attributes of each test is detailed in the help document (click the question mark icon). The user may pulldown and select an alternate test if they disagree with the default test or wish to compare results from multiple tests. Our recommendation is to use the results from more than one statistical method to draw inferences.

Step 3: Select Statistical Method

System Selects (default recommend) ▾
System Selects (default recommend)
Mann-Whitney (2 sets 5+ samples)
Fisher's Exact (2 sets)
Welch's T-test (2 sets)
Anova (3+ sets)
Kruskal-Wallis (3+ sets 10+ samples)

Relative (default)

Absolute



Step 4: Choose output display options. To control for false positives, we introduce a P-value adjustment for multiple corrections using a routine Benjamini-Hochberg method to control false discovery rate (FDR). We further delineate functions with an adjusted FDR P-value of <0.05 as “significant” if this option is chosen.

Step 4: Choose Display Options

Show all rows

Show only rows with at least one non-zero gene count (filters input data)

Show only rows with significant hits (filters display results with adjusted Pvalue ≤ 0.05)

Step 5: Enter a meaningful job name and hit “Run Analysis”

Save as a new job with name:

Replace the selected job:

Run Analysis

User will be alerted by email upon job completion. Alternatively, user can check on the status of their submission by navigating to Workspace > My Jobs.

The results table displayed in the user interface will include only up to 1000 rows and is displayed in ascending order of the FDR adjusted P-value (last column). Table columns are feature ID, feature name, mean gene count for each genome set submitted for analysis, standard error of mean for each set, P-value and FDR error-corrected P-value. This table can be downloaded using “Export”. The “Download Full Results” produces a table with all rows and all columns, and additionally includes normalized and raw gene counts for every genome in every set. Finally, user can select features of interest and add to the Function Cart for downstream analyses.

Results of Analysis

Only the first 1000 rows are shown filter by lowest adjusted Pvalue
 Workspace job name: Butyr_eno_posVSneg
 Features: pfam
 Measurement: geneCount
 Statistical method: MWTest relative
 Display Options: hits
 Sets: Butyrvibrio_eno_neg Butyrvibrio_eno_pos

[Add Selected to Function Cart](#) [Select All](#) [Clear All](#)

[Download Full Results](#)

Filter column: MWTest adjPval Filter text: [Apply](#) [?](#)

[Export](#) Page 1 of 2 << first < prev 1 2 next > last >> 100

[Column Selector](#) [Select Page](#) [Deselect Page](#)

Select	Feature	Description	Mean Butyrn	Mean Butyrvit	StdErr Butyrv	StdErr Butyr	MWTest UStat	MWTest Pval	MWTest adjPval
<input type="checkbox"/>	pfam00113	Enolase, C-terminal TIM barrel domain	0	0.0254341	0	0.00082085	0	5.682e-11	1.615e-07
<input type="checkbox"/>	pfam03952	Enolase, N-terminal domain	0	0.0230331	0	0.00123742	24	5.795e-10	8.235e-07
<input type="checkbox"/>	pfam01098	Cell cycle protein	0.0757263	0.111211	0.0017927	0.00862976	99	1.343e-06	0.00127233
<input type="checkbox"/>	pfam00005	ABC transporter	2.23718	2.51696	0.0279941	0.0456065	113	5.897e-06	0.00376739
<input type="checkbox"/>	pfam00580	UvrD/REP helicase N-terminal domain	0.158502	0.112193	0.00669482	0.00586356	604	7.954e-06	0.00376739
<input type="checkbox"/>	pfam09363	XFP C-terminal domain	0	0.0150458	0	0.00231314	144	7.766e-06	0.00376739
<input type="checkbox"/>	pfam02597	ThiS family	0	0.0143317	0	0.0023551	156	1.702e-05	0.00439672
<input type="checkbox"/>	pfam03894	D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase	0	0.0143317	0	0.0023551	156	1.702e-05	0.00439672
<input type="checkbox"/>	pfam05690	Thiazole biosynthesis protein ThiG	0	0.0143317	0	0.0023551	156	1.702e-05	0.00439672
<input type="checkbox"/>	pfam09364	XFP N-terminal domain	0	0.0143317	0	0.0023551	156	1.702e-05	0.00439672
<input type="checkbox"/>	pfam16173	Domain of unknown function (DUF4874)	0	0.0141962	0	0.00233531	156	1.702e-05	0.00439672
<input type="checkbox"/>	pfam02645	Uncharacterised protein, DegV family COG1307	0.165223	0.121155	0.0075514	0.00645851	595	1.884e-05	0.00446277