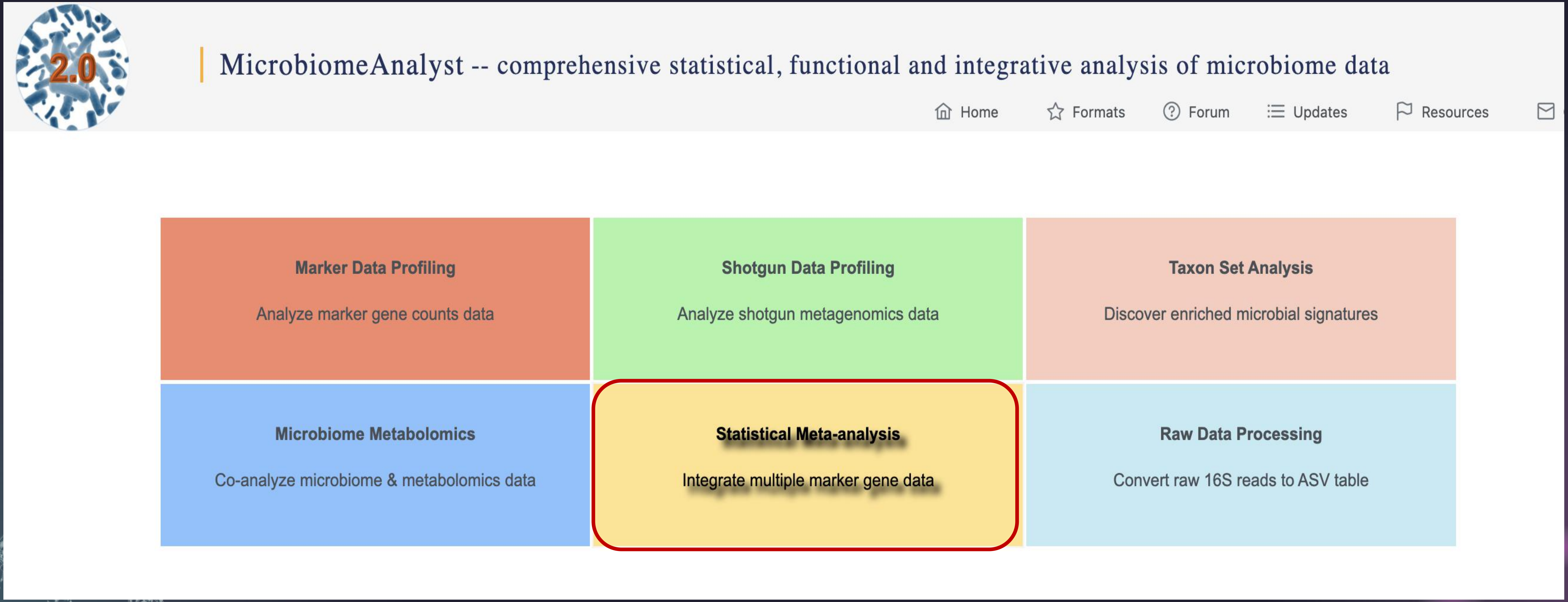
MicrobiomeAnalyst 2.0

Comprehensive statistical, functional and integrative analysis of microbiome data

xialab@mcgill 2023-Mar-03

Tutorial for Statistical Meta-analysis



MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

Marker Data Profiling

Analyze marker gene counts data

Shotgun Data Profiling

Analyze shotgun metagenomics data

Taxon Set Analysis

Discover enriched microbial signatures

Microbiome Metabolomics

Co-analyze microbiome & metabolomics data

Statistical Meta-analysis

Integrate multiple marker gene data

Raw Data Processing

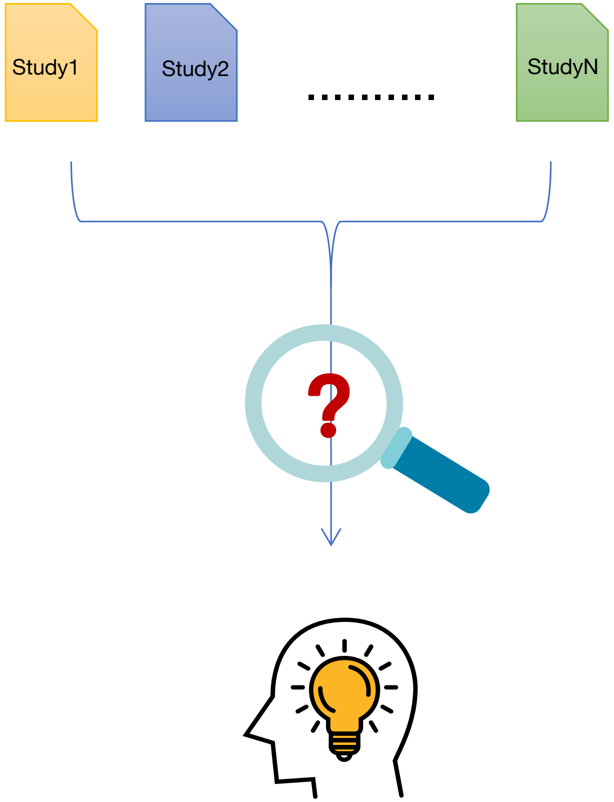
Convert raw 16S reads to ASV table

Motivation

- Increasing microbiome studies result in tremendous data designed for understanding different experimental variables, such as diseases and environment pressure, associated with changes in microbial community structure.

- However, it remains a major challenge to achieve reproducible features across different microbiome studies due to the variation in experimental design, analysis methods and quantitative assessment.

- There is an unmet need for analytical tools that provide rigorous statistical analysis dedicated to mine available data against same hypothesis and obtain consistent interpretation across different studies.



Overview

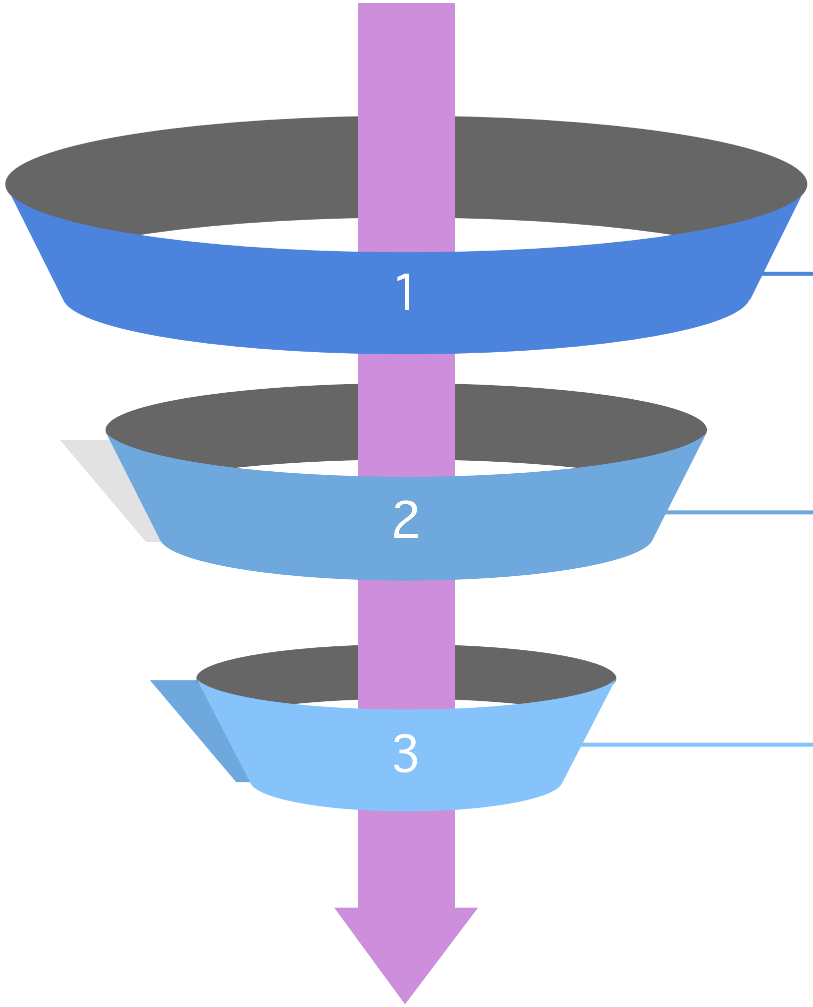
Goal: To provide a framework for integrating multiple maker gene studies to help identify robust and reproducible features from multiple microbiome studies.

Strategy and Approach:

- The MMUPhin method is employed to alleviates batch effects in the joint analysis of microbial profiles. It adjust for differences in technical or experimental variation between studies by considering batch/study effects which can significantly increases the comparability of different microbiome studies.

- Three analysis tracks are offered for user to explore the consistent pattern and potential biomarkers – visual exploration, diversity meta-analysis, and biomarker meta-analysis.

Datasets selection



Study designs should compare the same experimental factors

16S sequecing 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: data table

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

Sample names

OTU ids

#NAME	Sample1	Sample2	Sample3	Sample4	Sampl5	Sampl6	Sample7	Sampl8
OTU1	-3.06	-2.25	-1.15	-6.64	0.4	1.08	1.22	1.02
OTU2	-1.36	-0.67	-0.17	-0.97	-2.32	-5.06	0.28	1.32
OTU3	1.61	-0.27	0.71	-0.62	0.14		0.11	0.98
OTU4	0.93	1.29	-0.23	-0.74	-2	-1.25	1.07	1.27

…

Please take a look at these example data tables:

https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data1.csv
https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data2.csv
https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data3.csv

Data format: meta-data table

Primary meta-data

Sample names

#NAME	study_condition	age
SID31004	CRC	64
SID31009	control	68
SID31021	control	60
SID31071	control	68
SID31112	control	66
SID31129	control	73
SID31159	CRC	73

…

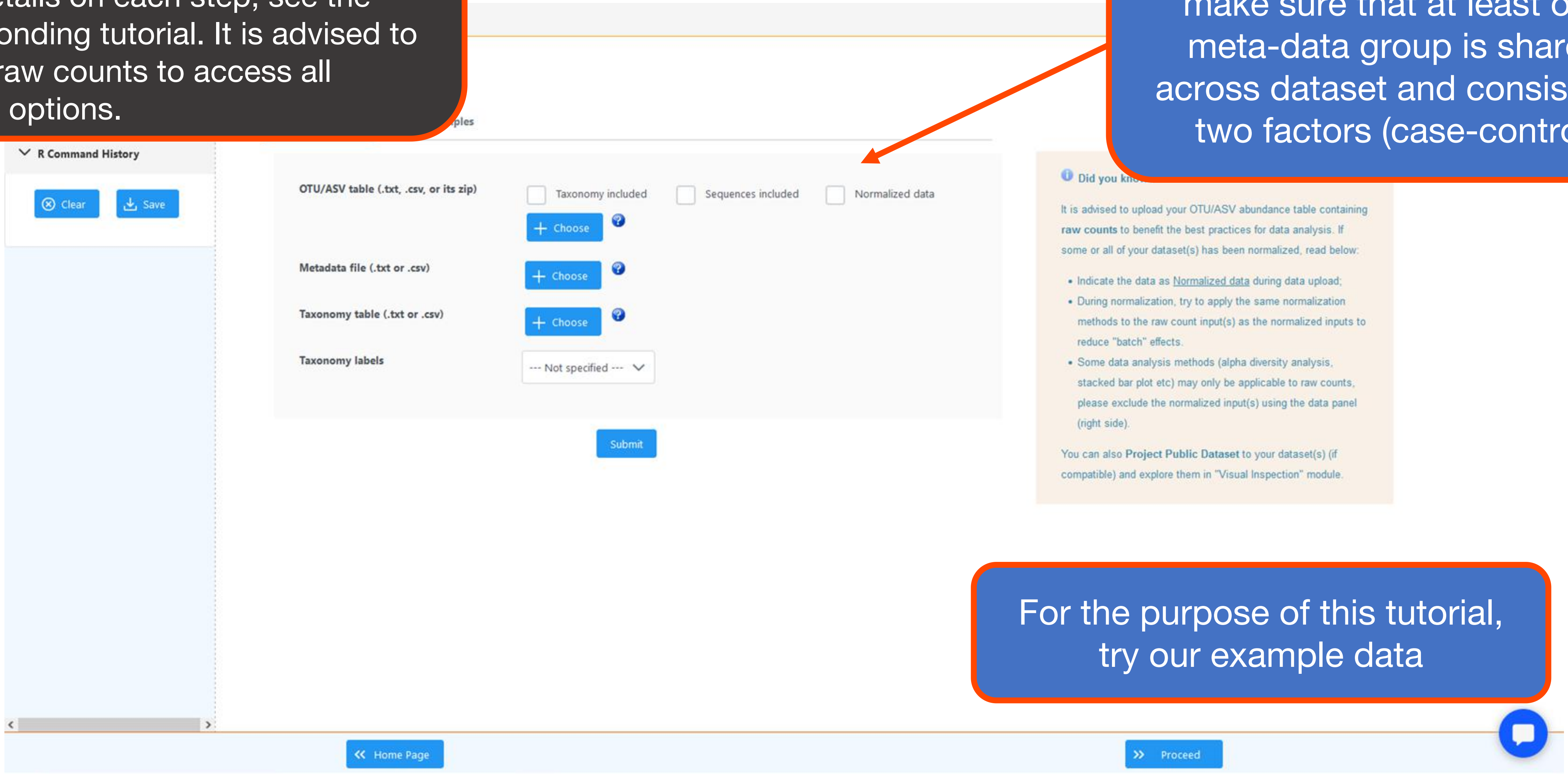
The primary meta-data needs to be consistent across datasets.
Only supports case-control Design (two factors)

https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data1_meta.csv
https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data2_meta.csv
https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/resources/data/metaanal/data3_meta.csv

Upload data

The first step is to upload and process all your individual datasets. This repeats the steps of a single marker data profiling for each dataset - for more details on each step, see the corresponding tutorial. It is advised to upload raw counts to access all analysis options.

Upload your dataset 1 by 1, make sure that at least one meta-data group is shared across dataset and consists of two factors (case-control)



OTU/ASV table (.txt, .csv, or its zip) □ Taxonomy included □ Sequences included □ Normalized data

+ Choose

Metadata file (.txt or .csv) + Choose

Taxonomy table (.txt or .csv) + Choose

Taxonomy labels --- Not specified ---

Submit

Did you kno

It is advised to upload your OTU/ASV abundance table containing **raw counts** to benefit the best practices for data analysis. If some or all of your dataset(s) has been normalized, read below:

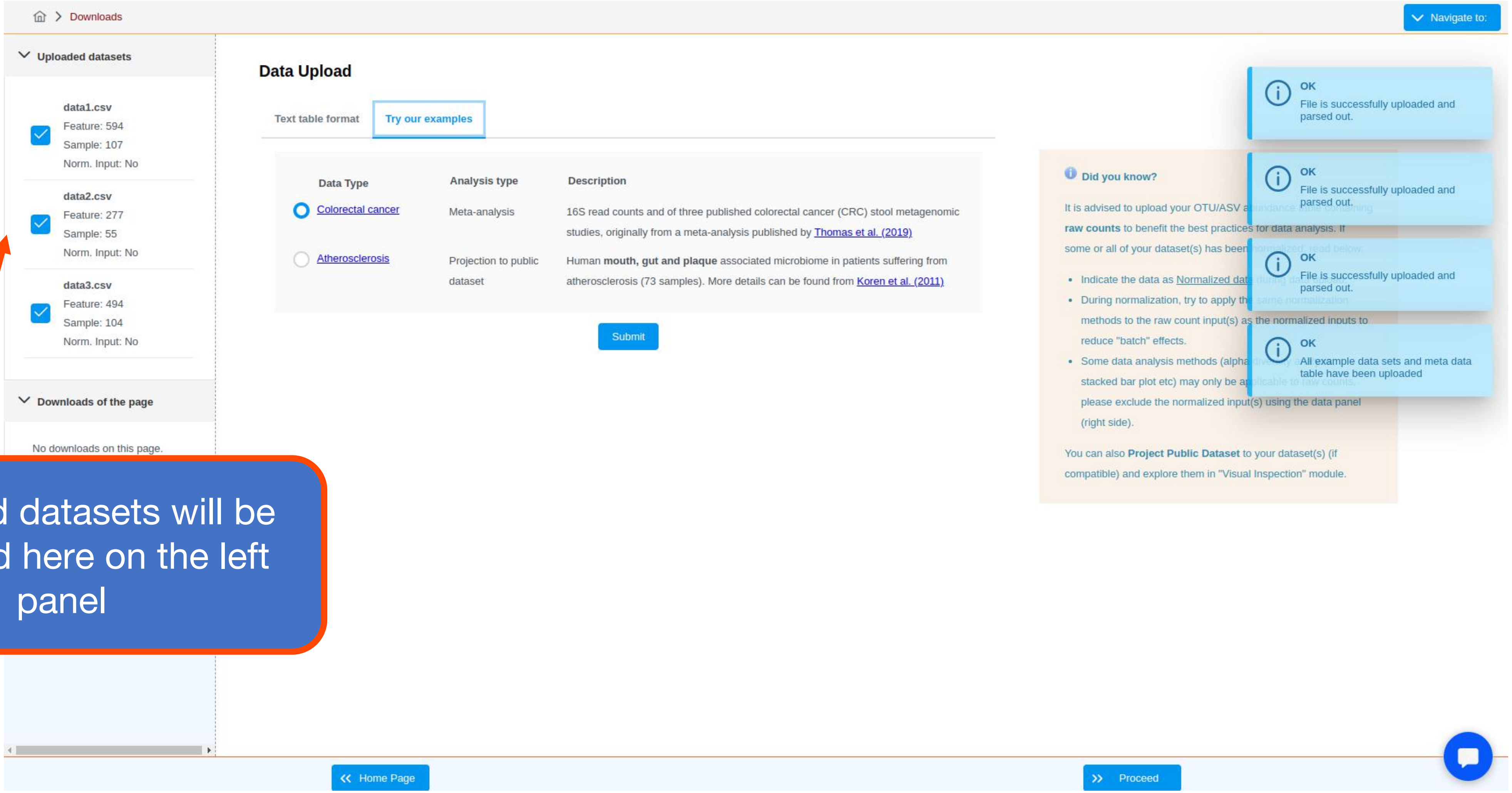
- Indicate the data as Normalized data during data upload;
- During normalization, try to apply the same normalization methods to the raw count input(s) as the normalized inputs to reduce "batch" effects.
- Some data analysis methods (alpha diversity analysis, stacked bar plot etc) may only be applicable to raw counts, please exclude the normalized input(s) using the data panel (right side).

You can also **Project Public Dataset** to your dataset(s) (if compatible) and explore them in "Visual Inspection" module.

For the purpose of this tutorial, try our example data

The example datasets come from stool samples of three 16S colorectal cancer studies; the datasets have been trimmed for testing purposes.

Example data



Uploaded datasets

☑ data1.csv
Feature: 594
Sample: 107
Norm. Input: No

☑ data2.csv
Feature: 277
Sample: 55
Norm. Input: No

☑ data3.csv
Feature: 494
Sample: 104
Norm. Input: No

Downloads of the page

No downloads on this page.

Data Upload

Data Type	Analysis type	Description
● Colorectal cancer	Meta-analysis	16S read counts and of three published colorectal cancer (CRC) stool metagenomic studies, originally from a meta-analysis published by Thomas et al. (2019)
○ Atherosclerosis	Projection to public dataset	Human **mouth, gut and plaque** associated microbiome in patients suffering from atherosclerosis (73 samples). More details can be found from Koren et al. (2011)

Did you know?

It is advised to upload your OTU/ASV abundance table containing **raw counts** to benefit the best practices for data analysis. If some or all of your dataset(s) has been normalized, read below:

- Indicate the data as Normalized data during data upload.
- During normalization, try to apply the same normalization methods to the raw count input(s) as the normalized inputs to reduce "batch" effects.
- Some data analysis methods (alpha diversity analysis, stacked bar plot etc) may only be applicable to raw counts, please exclude the normalized input(s) using the data panel (right side).

You can also **Project Public Dataset** to your dataset(s) (if compatible) and explore them in "Visual Inspection" module.

OK
File is successfully uploaded and parsed out.

OK
File is successfully uploaded and parsed out.

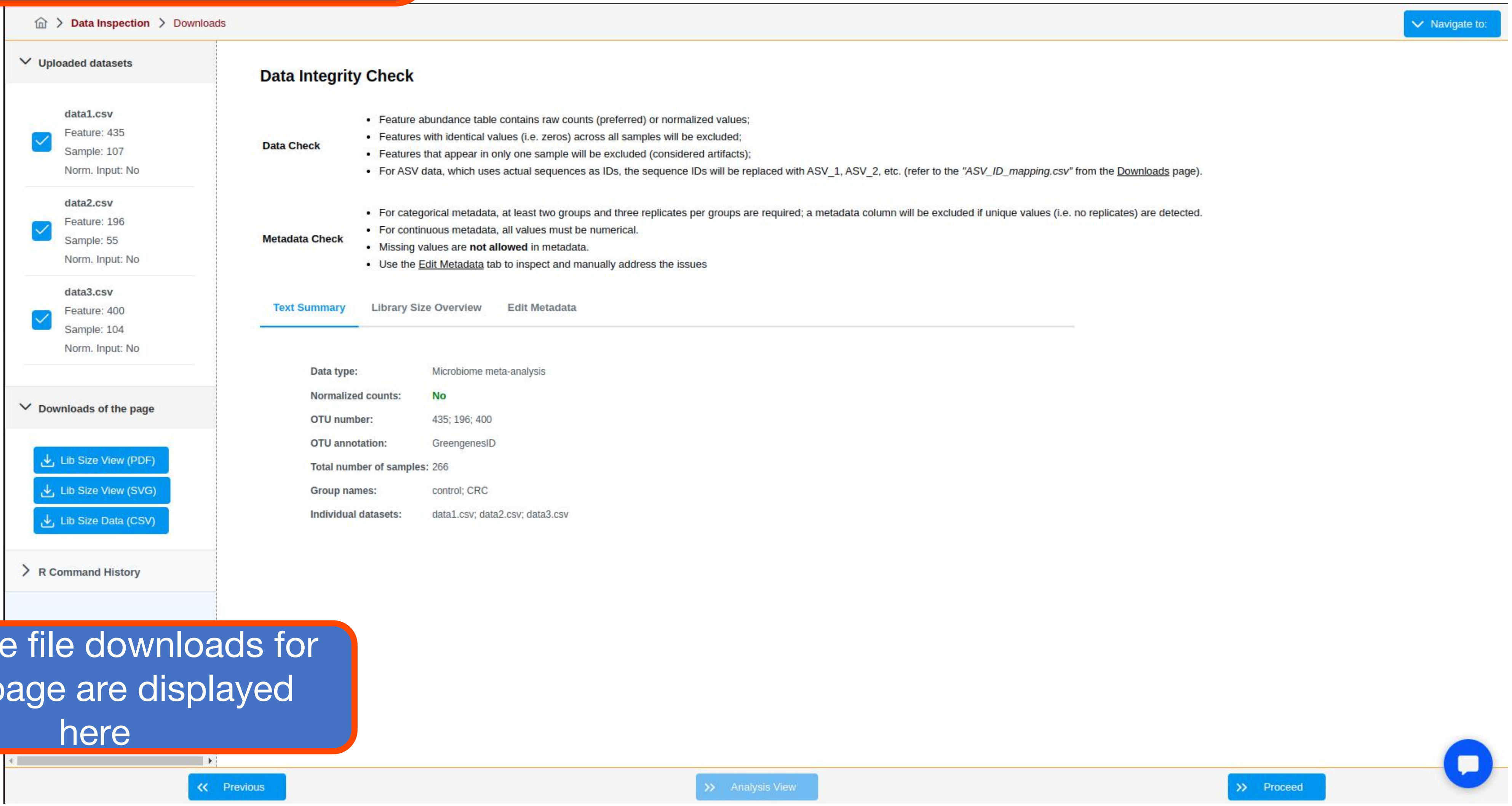
OK
File is successfully uploaded and parsed out.

OK
All example data sets and meta data table have been uploaded

Uploaded datasets will be displayed here on the left panel

This page provides general text summary and library size graphical overview on the uploaded datasets

Data Summary



data1.csv
Feature: 435
Sample: 107
Norm. Input: No

data2.csv
Feature: 196
Sample: 55
Norm. Input: No

data3.csv
Feature: 400
Sample: 104
Norm. Input: No

Data Integrity Check

Data Check

- Feature abundance table contains raw counts (preferred) or normalized values;
- Features with identical values (i.e. zeros) across all samples will be excluded;
- Features that appear in only one sample will be excluded (considered artifacts);
- For ASV data, which uses actual sequences as IDs, the sequence IDs will be replaced with ASV_1, ASV_2, etc. (refer to the *"ASV_ID_mapping.csv"* from the Downloads page).

Metadata Check

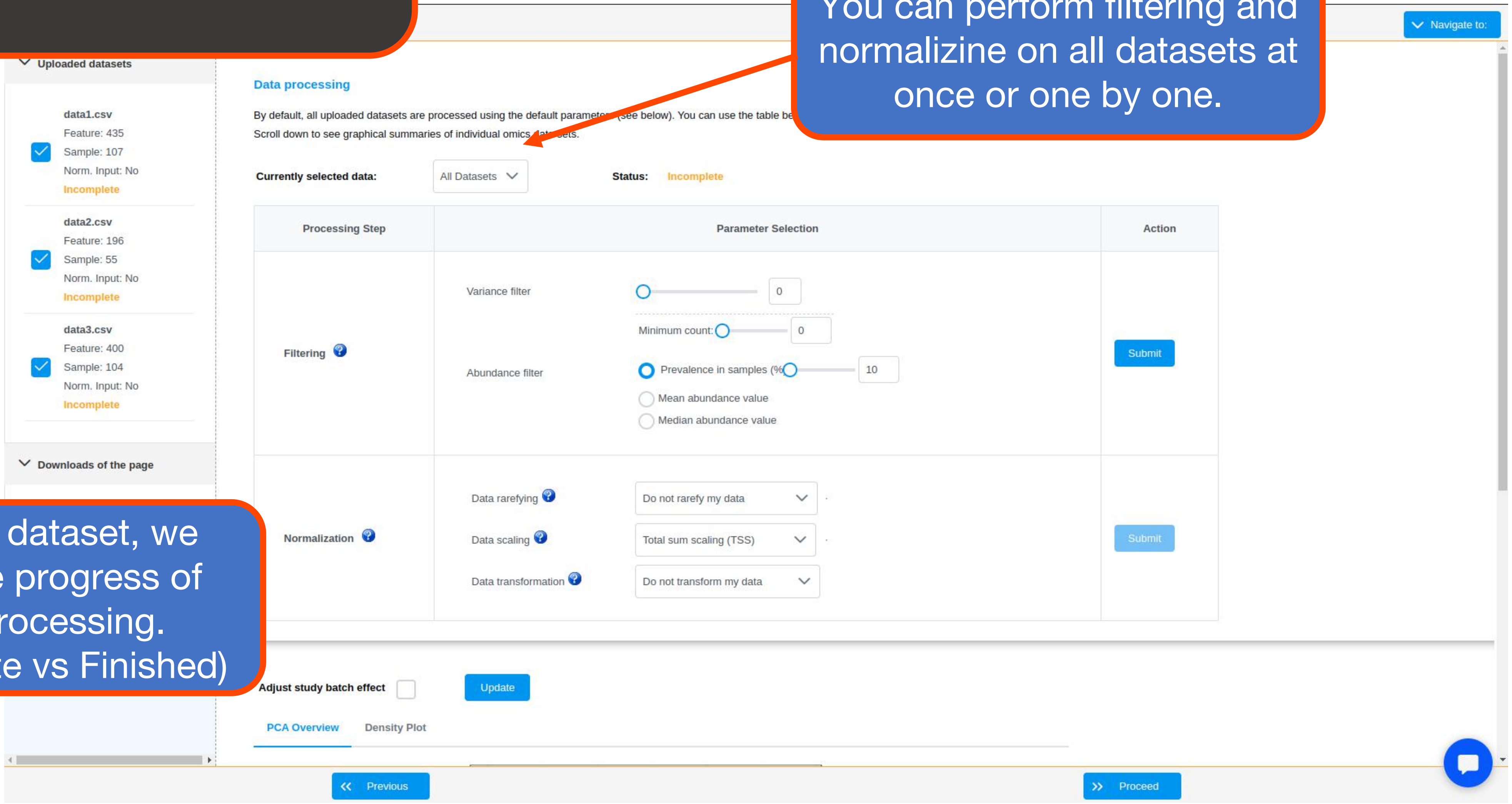
- For categorical metadata, at least two groups and three replicates per groups are required; a metadata column will be excluded if unique values (i.e. no replicates) are detected.
- For continuous metadata, all values must be numerical.
- Missing values are **not allowed** in metadata.
- Use the Edit Metadata tab to inspect and manually address the issues

Data type:	Microbiome meta-analysis
Normalized counts:	**No**
OTU number:	435; 196; 400
OTU annotation:	GreengenesID
Total number of samples:	266
Group names:	control; CRC
Individual datasets:	data1.csv; data2.csv; data3.csv

Available file downloads for each page are displayed here

Data processing

Data processing page offers the same filtering and normalization options available for single gene marker profiling with the addition of batch effect correction to remove study-specific bias



data1.csv
Feature: 435
Sample: 107
Norm. Input: No
Incomplete

data2.csv
Feature: 196
Sample: 55
Norm. Input: No
Incomplete

data3.csv
Feature: 400
Sample: 104
Norm. Input: No
Incomplete

Data processing

By default, all uploaded datasets are processed using the default parameters (see below). You can use the table be

Scroll down to see graphical summaries of individual omics datasets.

Currently selected data: All Datasets

Status: Incomplete

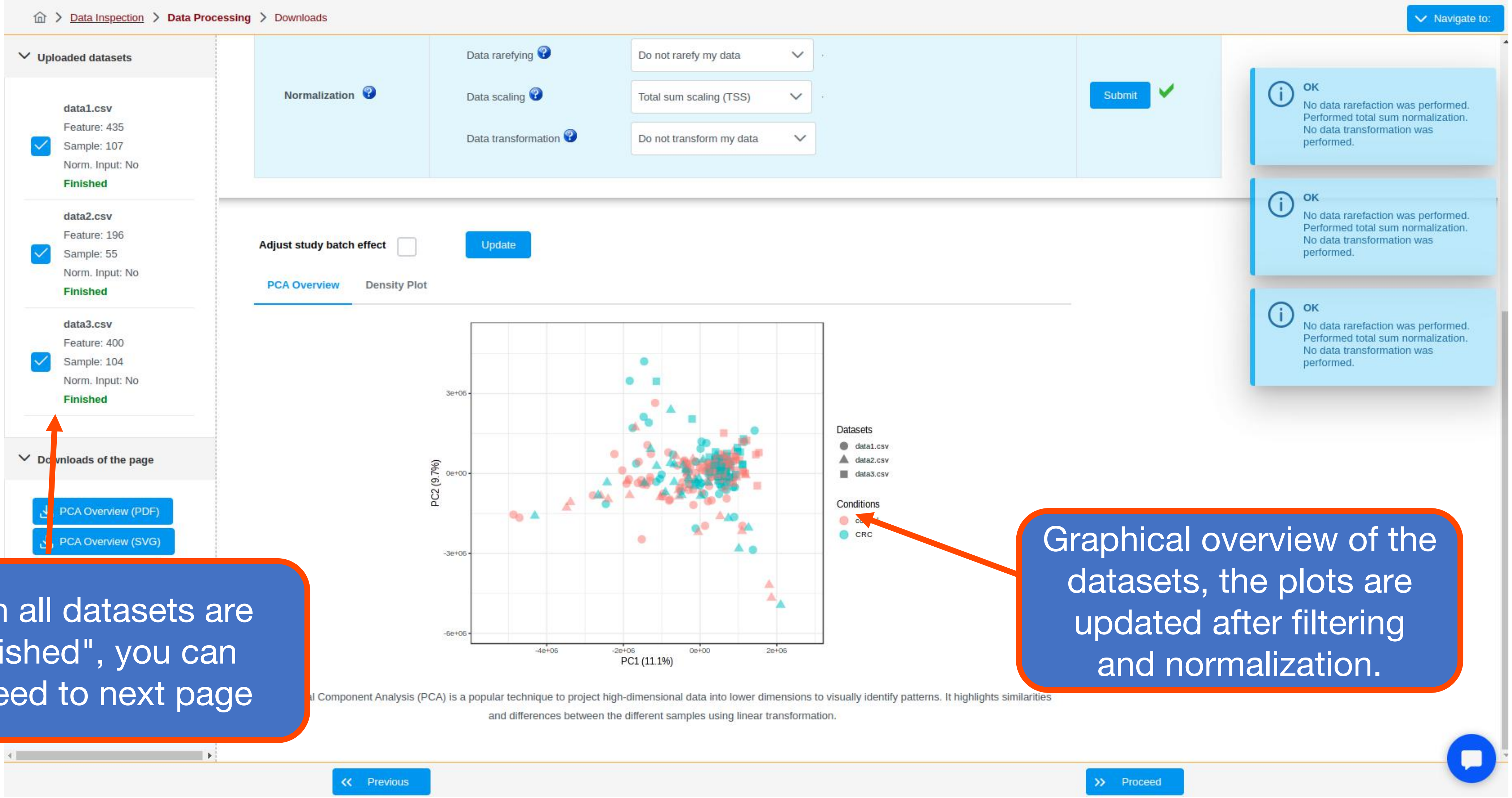
Processing Step	Parameter Selection	Action
Filtering	Variance filter: 0 Abundance filter: Minimum count: 0 ● Prevalence in samples (%) 10 ○ Mean abundance value ○ Median abundance value	Submit
Normalization	Data rarefying: Do not rarefy my data Data scaling: Total sum scaling (TSS) Data transformation: Do not transform my data	Submit

Adjust study batch effect □ Update

You can perform filtering and normalizine on all datasets at once or one by one.

On each dataset, we show the progress of data processing. (Incomplete vs Finished)

Data processing



Data Inspection > **Data Processing** > Downloads

Uploaded datasets

☑ **data1.csv**
Feature: 435
Sample: 107
Norm. Input: No
Finished

☑ **data2.csv**
Feature: 196
Sample: 55
Norm. Input: No
Finished

☑ **data3.csv**
Feature: 400
Sample: 104
Norm. Input: No
Finished

Downloads of the page

Normalization

Data rarefying: Do not rarefy my data

Data scaling: Total sum scaling (TSS)

Data transformation: Do not transform my data

OK
No data rarefaction was performed. Performed total sum normalization. No data transformation was performed.

OK
No data rarefaction was performed. Performed total sum normalization. No data transformation was performed.

OK
No data rarefaction was performed. Performed total sum normalization. No data transformation was performed.

Adjust study batch effect □

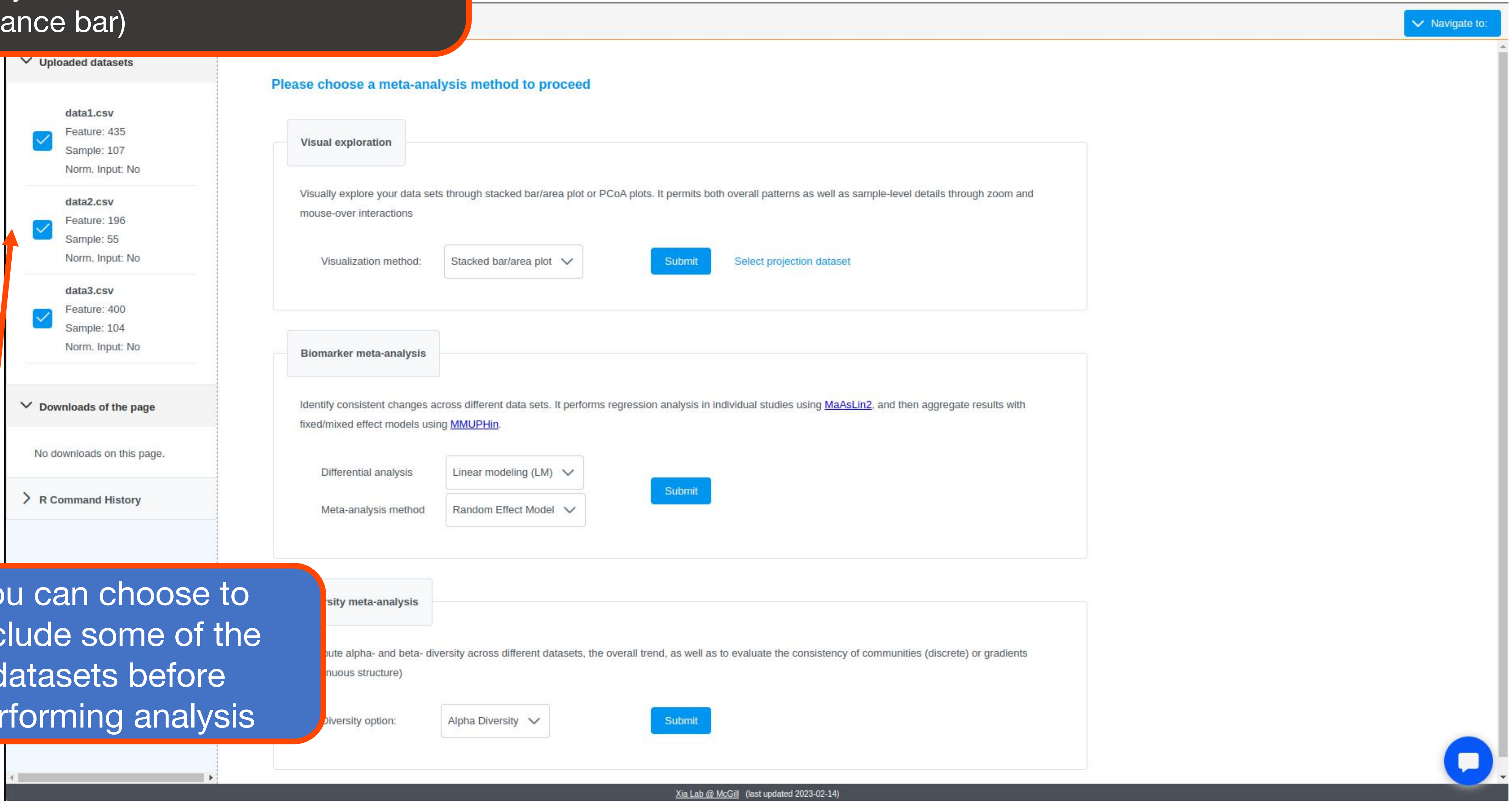
l Component Analysis (PCA) is a popular technique to project high-dimensional data into lower dimensions to visually identify patterns. It highlights similarities and differences between the different samples using linear transformation.

When all datasets are "Finished", you can proceed to next page

Graphical overview of the datasets, the plots are updated after filtering and normalization.

Methods Selection

Note that some methods can only be performed on counts data (i.e. biomarker meta-analysis, alpha diversity, stacked area/taxa abundance bar)



☑ data1.csv
Feature: 435
Sample: 107
Norm. Input: No

☑ data2.csv
Feature: 196
Sample: 55
Norm. Input: No

☑ data3.csv
Feature: 400
Sample: 104
Norm. Input: No

No downloads on this page.

Please choose a meta-analysis method to proceed

Visual exploration

Visually explore your data sets through stacked bar/area plot or PCoA plots. It permits both overall patterns as well as sample-level details through zoom and mouse-over interactions

Visualization method: Stacked bar/area plot

Biomarker meta-analysis

Identify consistent changes across different data sets. It performs regression analysis in individual studies using MaAsLin2, and then aggregate results with fixed/mixed effect models using MMUPHin.

Differential analysis: Linear modeling (LM)

Meta-analysis method: Random Effect Model

sity meta-analysis

ute alpha- and beta- diversity across different datasets, the overall trend, as well as to evaluate the consistency of communities (discrete) or gradients nuous structure)

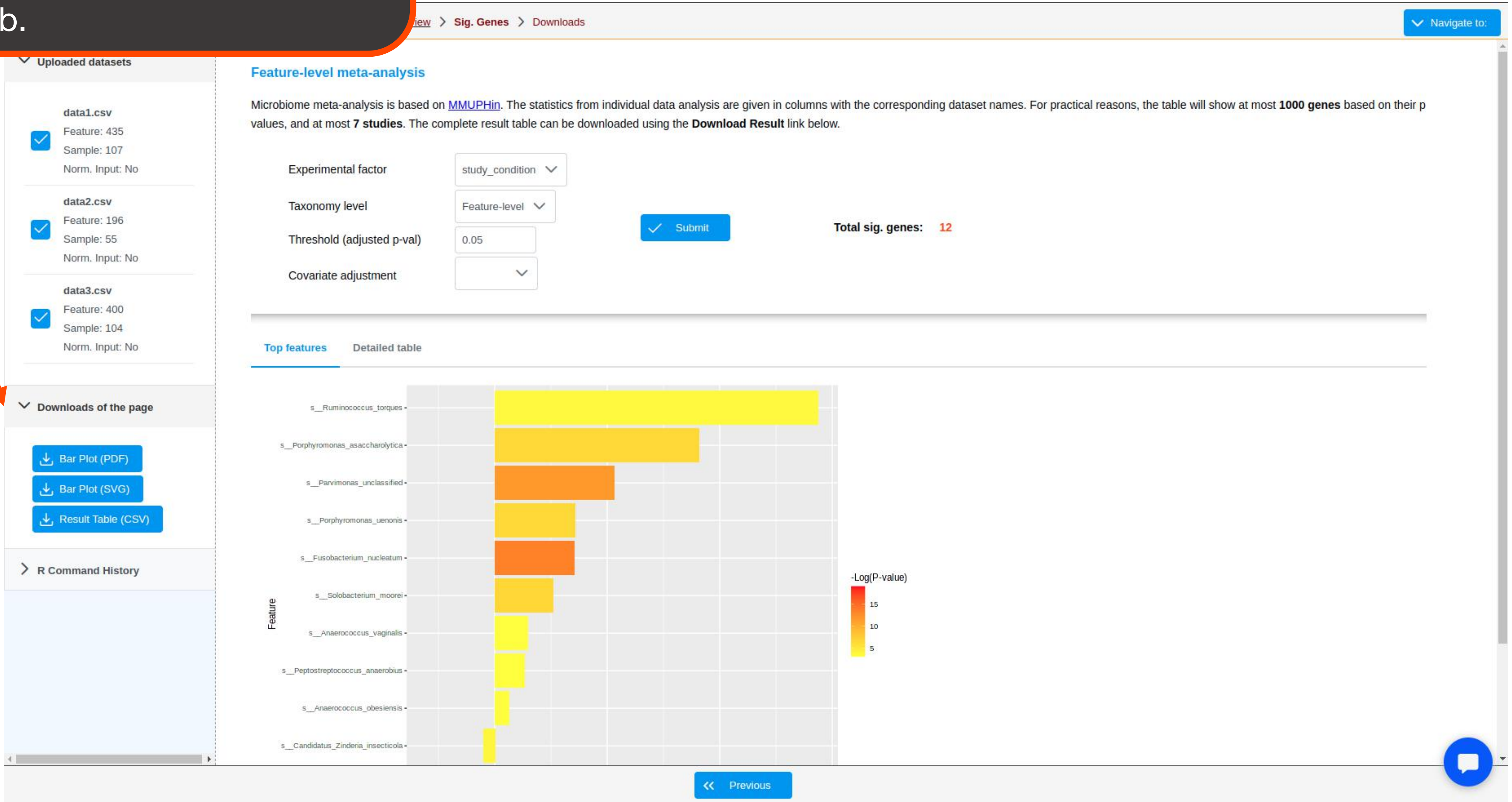
Diversity option: Alpha Diversity

Xia Lab @ McGill (last updated 2023-02-14)

You can choose to exclude some of the datasets before performing analysis

Biomarker meta-analysis

The graphical overview displays a maximum of top significant features. Detailed table contains the results for all features. You can also download the result table in the "Downloads of the page" tab.



iew > Sig. Genes > Downloads

Navigate to:

Uploaded datasets

☑ data1.csv
Feature: 435
Sample: 107
Norm. Input: No

☑ data2.csv
Feature: 196
Sample: 55
Norm. Input: No

☑ data3.csv
Feature: 400
Sample: 104
Norm. Input: No

Downloads of the page

Bar Plot (PDF)
Bar Plot (SVG)
Result Table (CSV)

R Command History

Feature-level meta-analysis

Microbiome meta-analysis is based on MMUPHin. The statistics from individual data analysis are given in columns with the corresponding dataset names. For practical reasons, the table will show at most **1000 genes** based on their p values, and at most **7 studies**. The complete result table can be downloaded using the **Download Result** link below.

Experimental factor: study_condition
Taxonomy level: Feature-level
Threshold (adjusted p-val): 0.05
Covariate adjustment:

Submit

Total sig. genes: 12

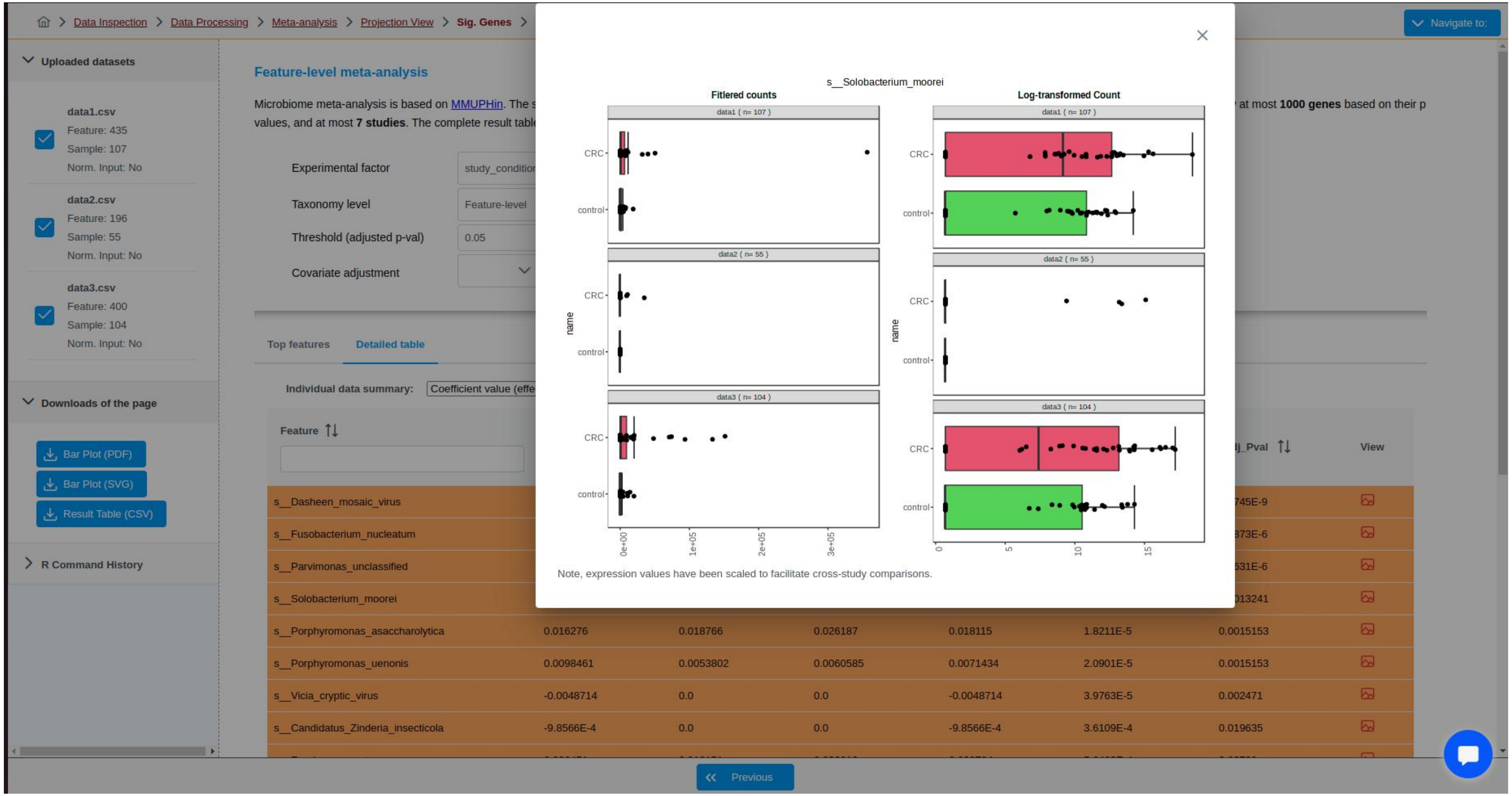
Top features | Detailed table

s__Ruminococcus_torques
s__Porphyromonas_asaccharolytica
s__Parvimonas_unclassified
s__Porphyromonas_uenonis
s__Fusobacterium_nucleatum
s__Solobacterium_moorei
s__Anaerococcus_vaginalis
s__Peptostreptococcus_anaerobius
s__Anaerococcus_obesiensis
s__Candidatus_Zinderia_insecticola
Feature
-Log(P-value)
15
10
5

Previous

You can visualize overall abundance profiles of individual feature using our Detailed table, under "View" column

Biomarker meta-analysis



Data Inspection > Data Processing > Meta-analysis > Projection View > Sig. Genes >

Navigate to:

Uploaded datasets

☑ data1.csv
Feature: 435
Sample: 107
Norm. Input: No

☑ data2.csv
Feature: 196
Sample: 55
Norm. Input: No

☑ data3.csv
Feature: 400
Sample: 104
Norm. Input: No

Downloads of the page

Bar Plot (PDF)
Bar Plot (SVG)
Result Table (CSV)

R Command History

Feature-level meta-analysis

Microbiome meta-analysis is based on MMUPHin. The s … at most **1000 genes** based on their p values, and at most **7 studies**. The complete result tabl

Experimental factor: study_conditio

Taxonomy level: Feature-level

Threshold (adjusted p-val): 0.05

Covariate adjustment:

Top features | Detailed table

Individual data summary: Coefficient value (effe

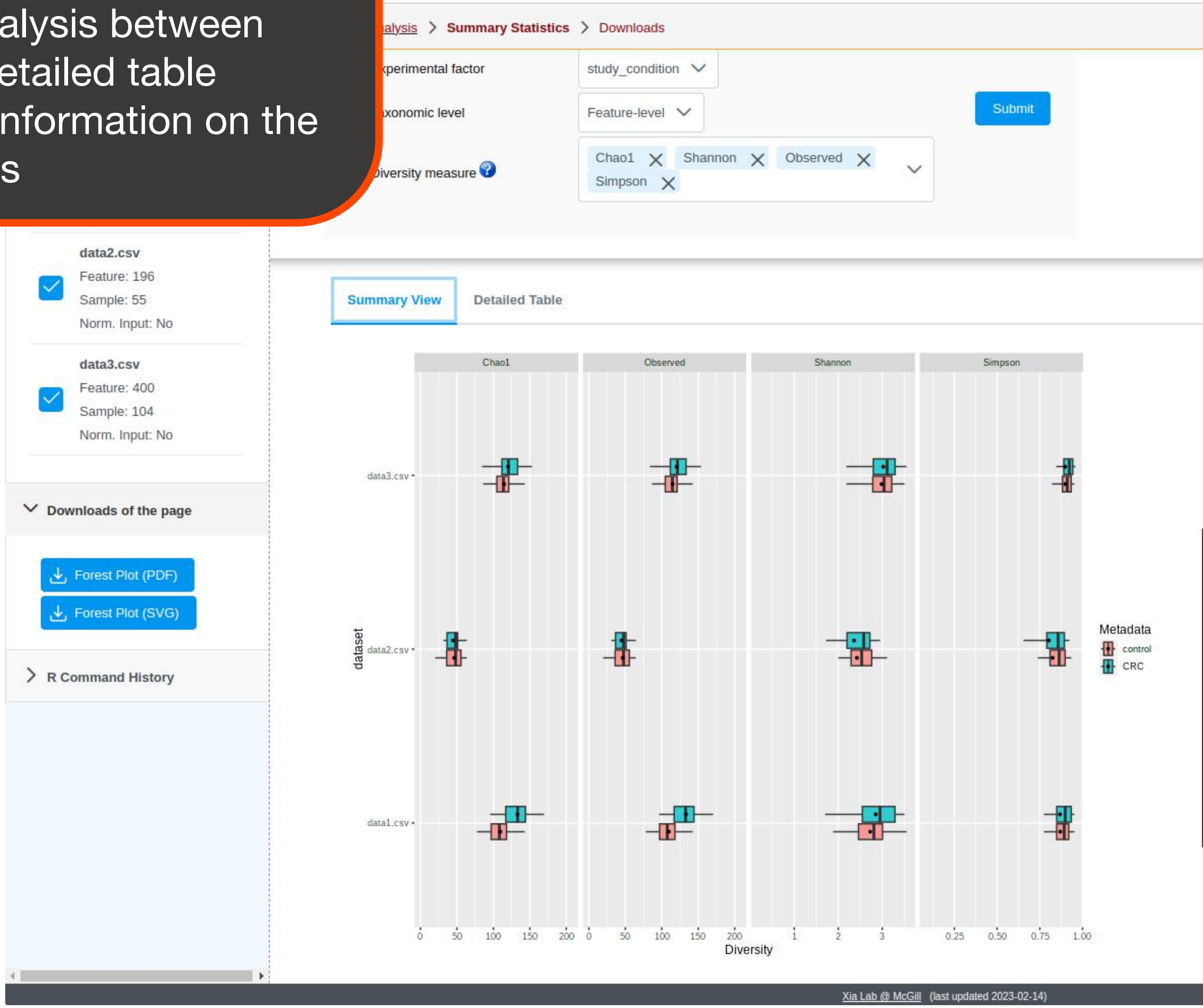
Feature ↑↓					…j_Pval ↑↓	View
s__Dasheen_mosaic_virus					…745E-9	
s__Fusobacterium_nucleatum					…373E-6	
s__Parvimonas_unclassified					…531E-6	
s__Solobacterium_moorei					…013241	
s__Porphyromonas_asaccharolytica	0.016276	0.018766	0.026187	0.018115	1.8211E-5	0.0015153
s__Porphyromonas_uenonis	0.0098461	0.0053802	0.0060585	0.0071434	2.0901E-5	0.0015153
s__Vicia_cryptic_virus	-0.0048714	0.0	0.0	-0.0048714	3.9763E-5	0.002471
s__Candidatus_Zinderia_insecticola	-9.8566E-4	0.0	0.0	-9.8566E-4	3.6109E-4	0.019635

Previous

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

This module offers two graphical representation: 1) box plot displays the distribution of diversity metrics; 2) log2 ratio view displays results from comparative analysis between case-control. Detailed table provides more information on the statistical results

Alpha diversity analysis



alysis > Summary Statistics > Downloads

xperimental factor: study_condition

axonomic level: Feature-level

Diversity measure: Chao1 ✕ Shannon ✕ Observed ✕ Simpson ✕

Submit

data2.csv
Feature: 196
Sample: 55
Norm. Input: No

data3.csv
Feature: 400
Sample: 104
Norm. Input: No

Downloads of the page

Forest Plot (PDF)

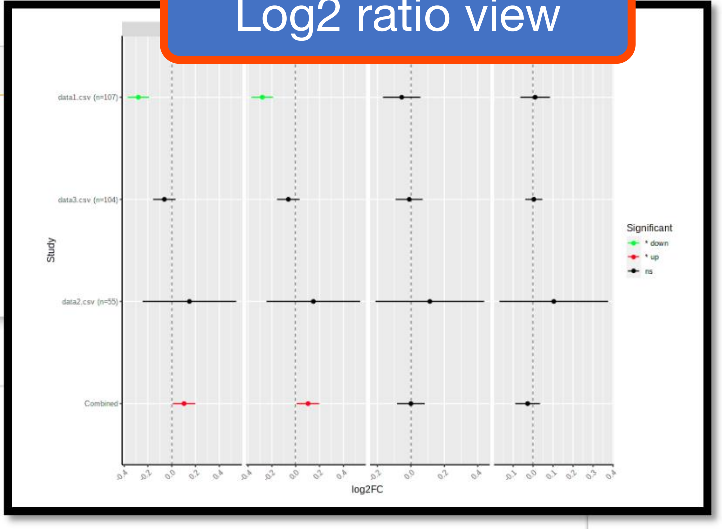
Forest Plot (SVG)

R Command History

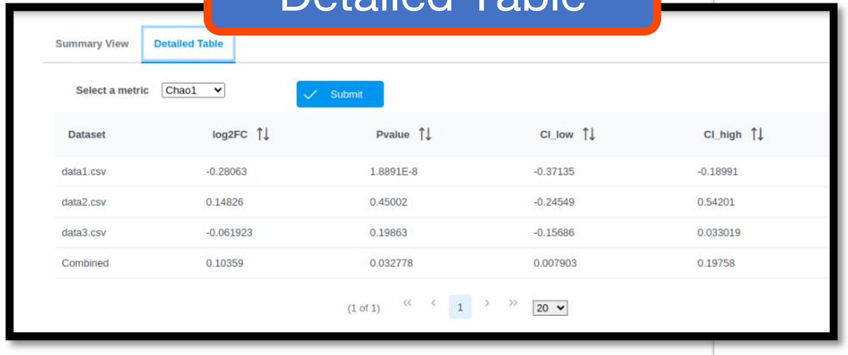
Summary View | Detailed Table

Xia Lab @ McGill (last updated 2023-02-14)

Log2 ratio view



Detailed Table



Summary View | Detailed Table

Select a metric: Chao1

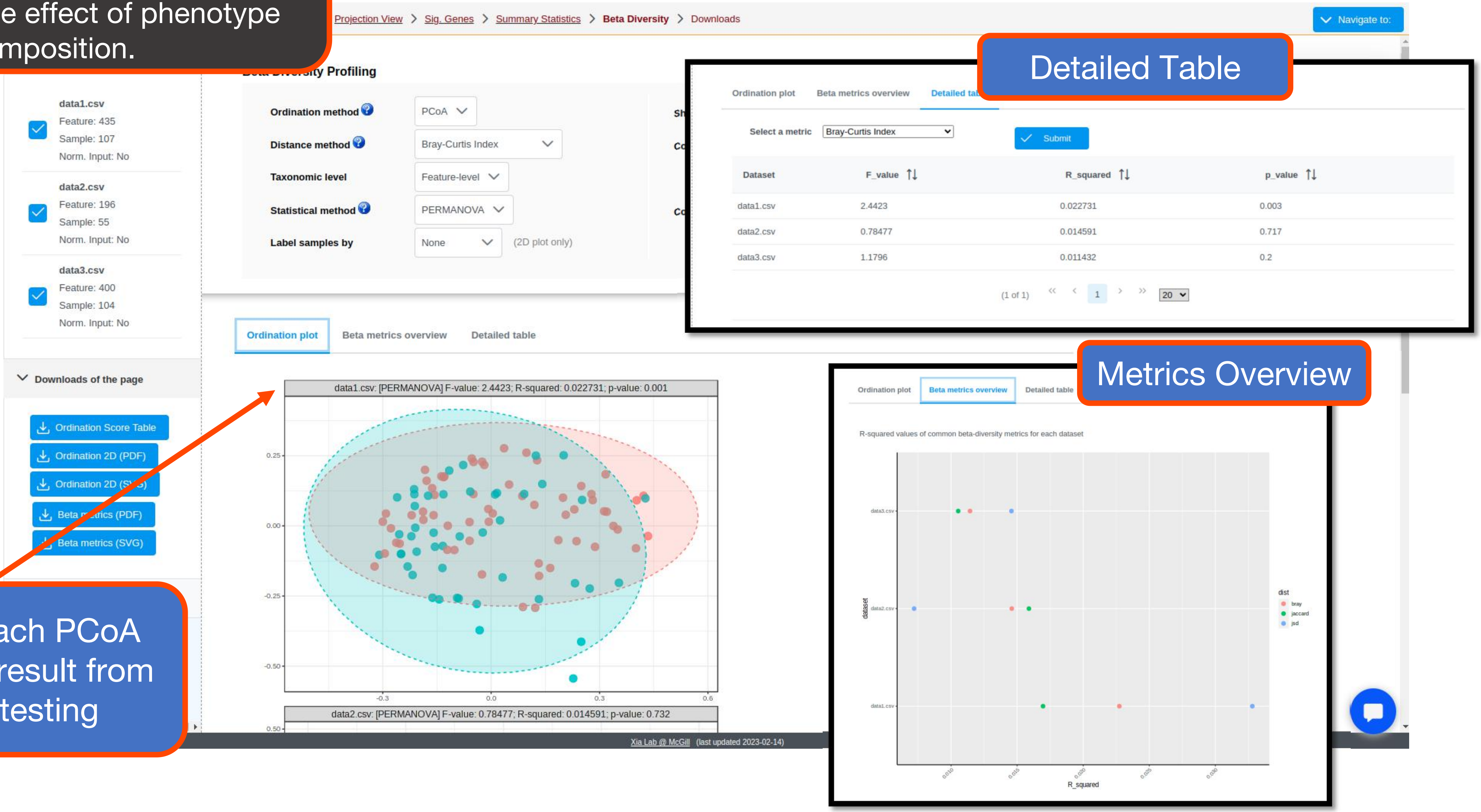
✓ Submit

Dataset	log2FC ⇅	Pvalue ⇅	CI_low ⇅	CI_high ⇅
data1.csv	-0.28063	1.8891E-8	-0.37135	-0.18991
data2.csv	0.14826	0.45002	-0.24549	0.54201
data3.csv	-0.061923	0.19863	-0.15686	0.033019
Combined	0.10359	0.032778	0.007903	0.19758

(1 of 1) « ‹ 1 › » 20

This module applies PCoA of beta diversity distance matrices along with statistical testing to measure significance on the effect of phenotype on community composition.

Beta diversity analysis



Projection View > Sig. Genes > Summary Statistics > Beta Diversity > Downloads
Navigate to:

data1.csv
Feature: 435
Sample: 107
Norm. Input: No

data2.csv
Feature: 196
Sample: 55
Norm. Input: No

data3.csv
Feature: 400
Sample: 104
Norm. Input: No

Beta Diversity Profiling

Ordination method: PCoA
Distance method: Bray-Curtis Index
Taxonomic level: Feature-level
Statistical method: PERMANOVA
Label samples by: None (2D plot only)

data1.csv: [PERMANOVA] F-value: 2.4423; R-squared: 0.022731; p-value: 0.001

data2.csv: [PERMANOVA] F-value: 0.78477; R-squared: 0.014591; p-value: 0.732

Detailed Table

Select a metric: Bray-Curtis Index

Dataset	F_value	R_squared	p_value
data1.csv	2.4423	0.022731	0.003
data2.csv	0.78477	0.014591	0.717
data3.csv	1.1796	0.011432	0.2

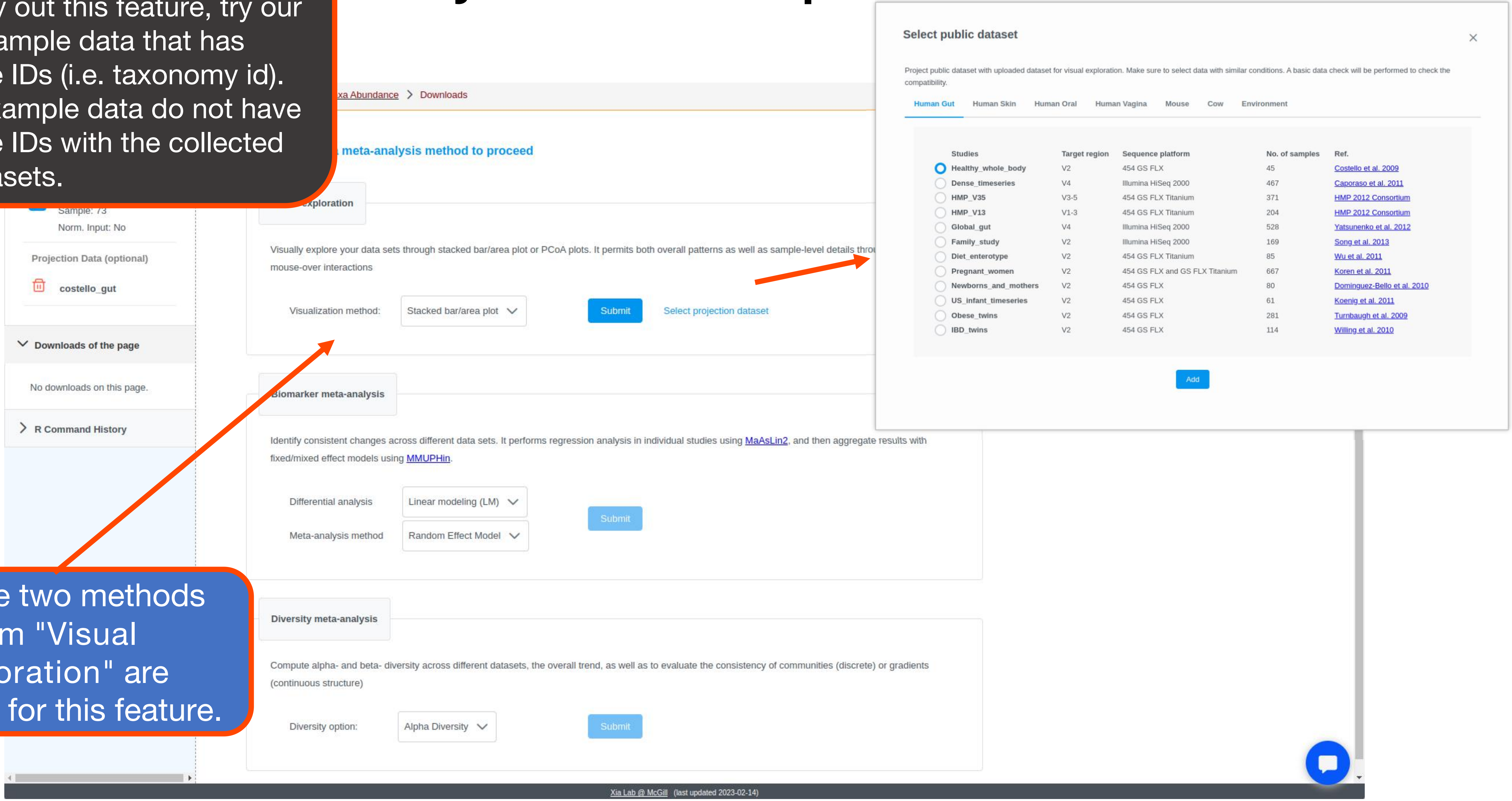
Metrics Overview

R-squared values of common beta-diversity metrics for each dataset

The title of each PCoA plot contains result from statistical testing

Projection to public dataset

The "Projection to public dataset" module has been merged here. To try out this feature, try our second example data that has compatible IDs (i.e. taxonomy id). Our first example data do not have compatible IDs with the collected public datasets.



Visually explore your data sets through stacked bar/area plot or PCoA plots. It permits both overall patterns as well as sample-level details throu mouse-over interactions

Visualization method: Stacked bar/area plot

Identify consistent changes across different data sets. It performs regression analysis in individual studies using MaAsLin2, and then aggregate results with fixed/mixed effect models using MMUPHin.

Differential analysis: Linear modeling (LM)

Meta-analysis method: Random Effect Model

Compute alpha- and beta- diversity across different datasets, the overall trend, as well as to evaluate the consistency of communities (discrete) or gradients (continuous structure)

Diversity option: Alpha Diversity

Select public dataset

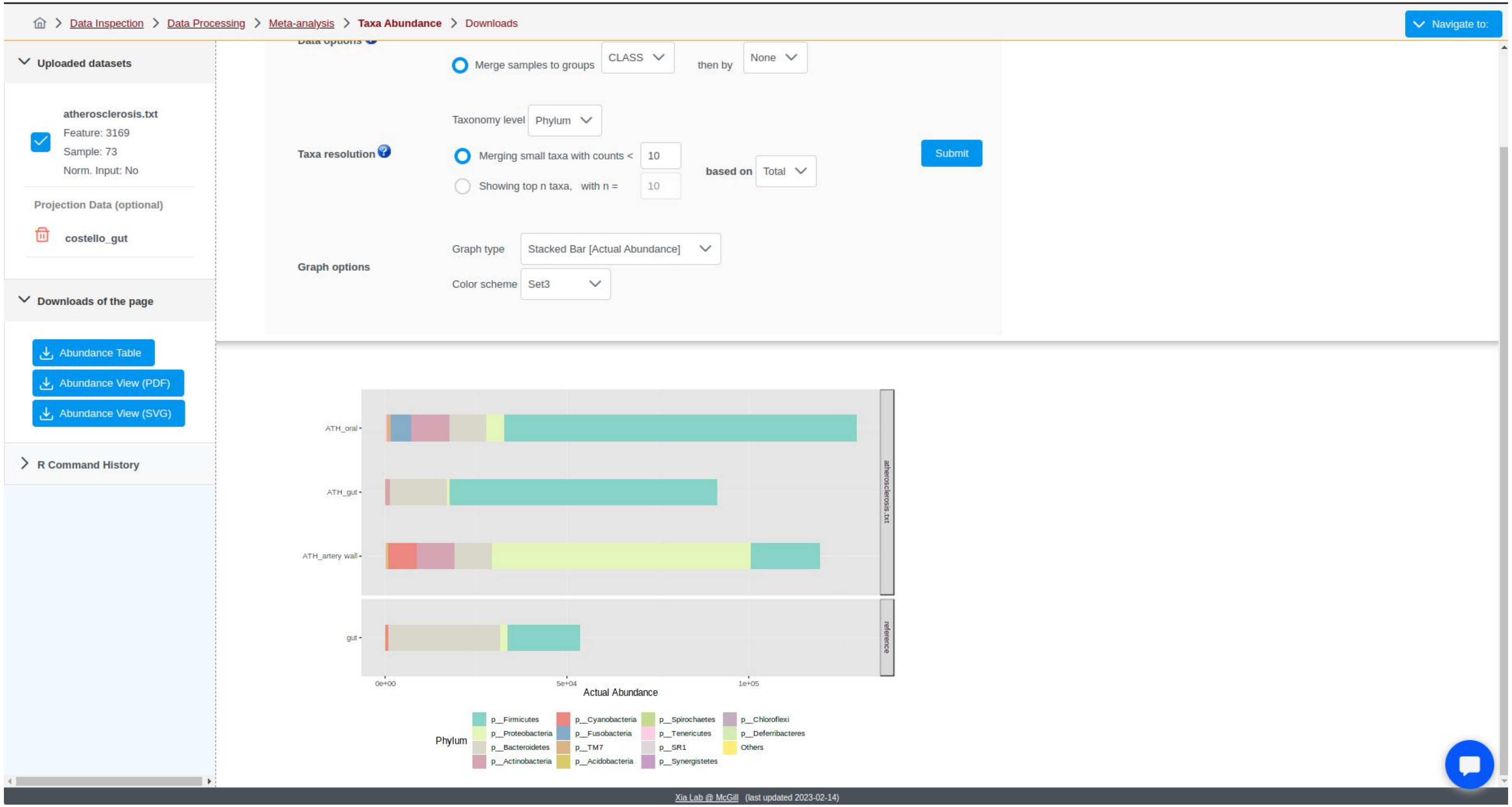
Project public dataset with uploaded dataset for visual exploration. Make sure to select data with similar conditions. A basic data check will be performed to check the compatibility.

	Studies	Target region	Sequence platform	No. of samples	Ref.
●	Healthy_whole_body	V2	454 GS FLX	45	Costello et al. 2009
○	Dense_timeseries	V4	Illumina HiSeq 2000	467	Caporaso et al. 2011
○	HMP_V35	V3-5	454 GS FLX Titanium	371	HMP 2012 Consortium
○	HMP_V13	V1-3	454 GS FLX Titanium	204	HMP 2012 Consortium
○	Global_gut	V4	Illumina HiSeq 2000	528	Yatsunenko et al. 2012
○	Family_study	V2	Illumina HiSeq 2000	169	Song et al. 2013
○	Diet_enterotype	V2	454 GS FLX Titanium	85	Wu et al. 2011
○	Pregnant_women	V2	454 GS FLX and GS FLX Titanium	667	Koren et al. 2011
○	Newborns_and_mothers	V2	454 GS FLX	80	Dominguez-Bello et al. 2010
○	US_infant_timeseries	V2	454 GS FLX	61	Koenig et al. 2011
○	Obese_twins	V2	454 GS FLX	281	Turnbaugh et al. 2009
○	IBD_twins	V2	454 GS FLX	114	Willing et al. 2010

Only the two methods from "Visual Exploration" are available for this feature.

You can compare taxa abundance of your uplodaded data with reference data using "Stacked bar/area plot".

Stacked bar plot



Uploaded datasets

atherosclerosis.txt
Feature: 3169
Sample: 73
Norm. Input: No

Projection Data (optional)

costello_gut

Data options

● Merge samples to groups: CLASS, then by: None

Taxa resolution

Taxonomy level: Phylum

● Merging small taxa with counts < 10, based on: Total

○ Showing top n taxa, with n = 10

Graph options

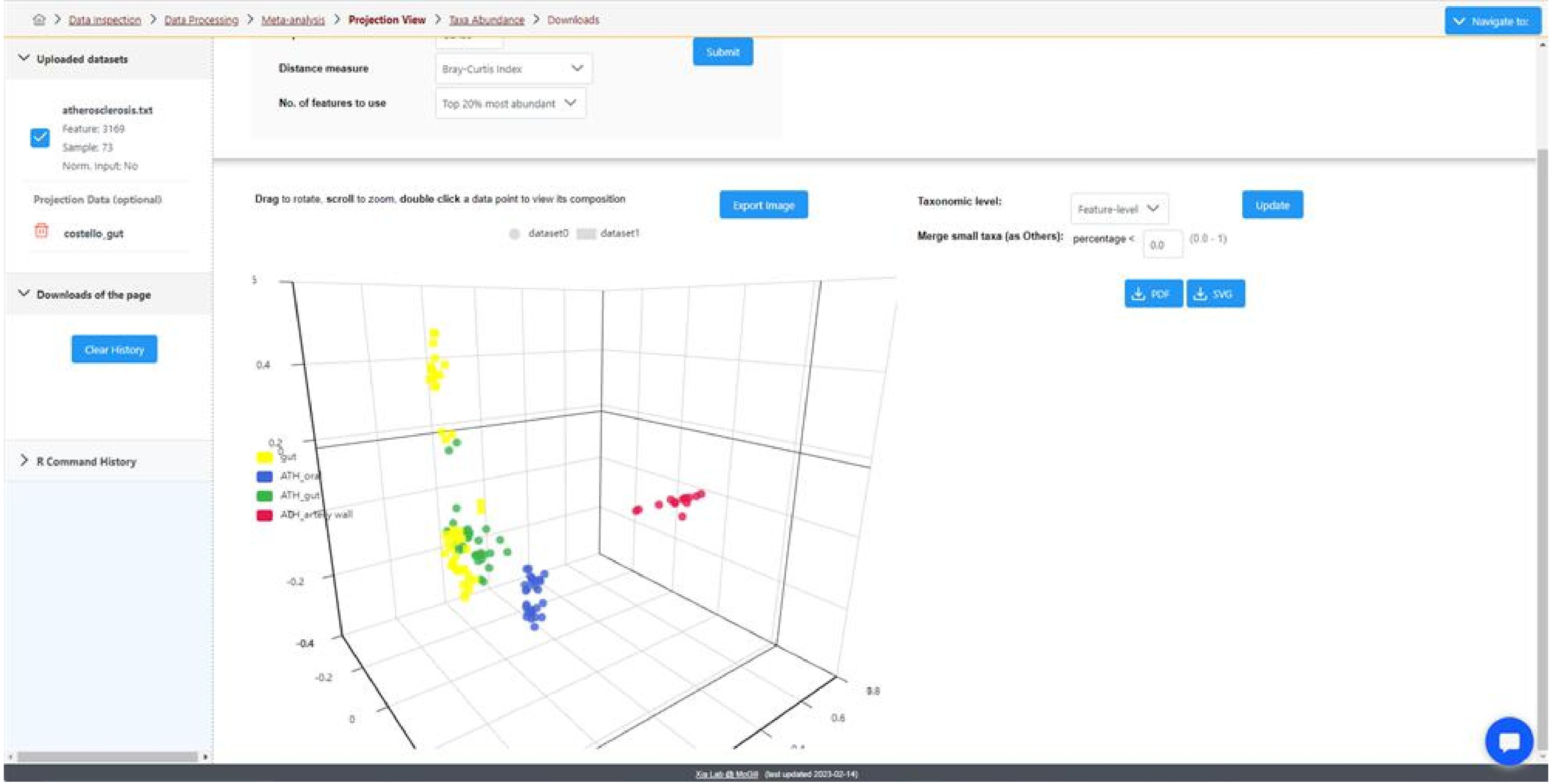
Graph type: Stacked Bar [Actual Abundance]

Color scheme: Set3

Xia Lab @ McGill (last updated 2023-02-14)

Similarly, you can visualize beta-diversity community composition with reference dataset in PCoA space.

PCoA projection



Uploaded datasets

☑ atherosclerosis.txt
Feature: 3169
Sample: 73
Norm. input: No

Projection Data (optional)

costello_gut

Distance measure: Bray-Curtis Index

No. of features to use: Top 20% most abundant

Drag to rotate, scroll to zoom, double click a data point to view its composition

Taxonomic level: Feature-level

Merge small taxa (as Others): percentage < 0.0 (0.0 - 1)

Xia Lab @ McGill (last updated 2023-02-14)

The End

For more information, visit Tutorials, Resources
and Contact pages on www.microbiomeanalyst.ca
Also visit our forum for FAQs on www.omicsforum.ca