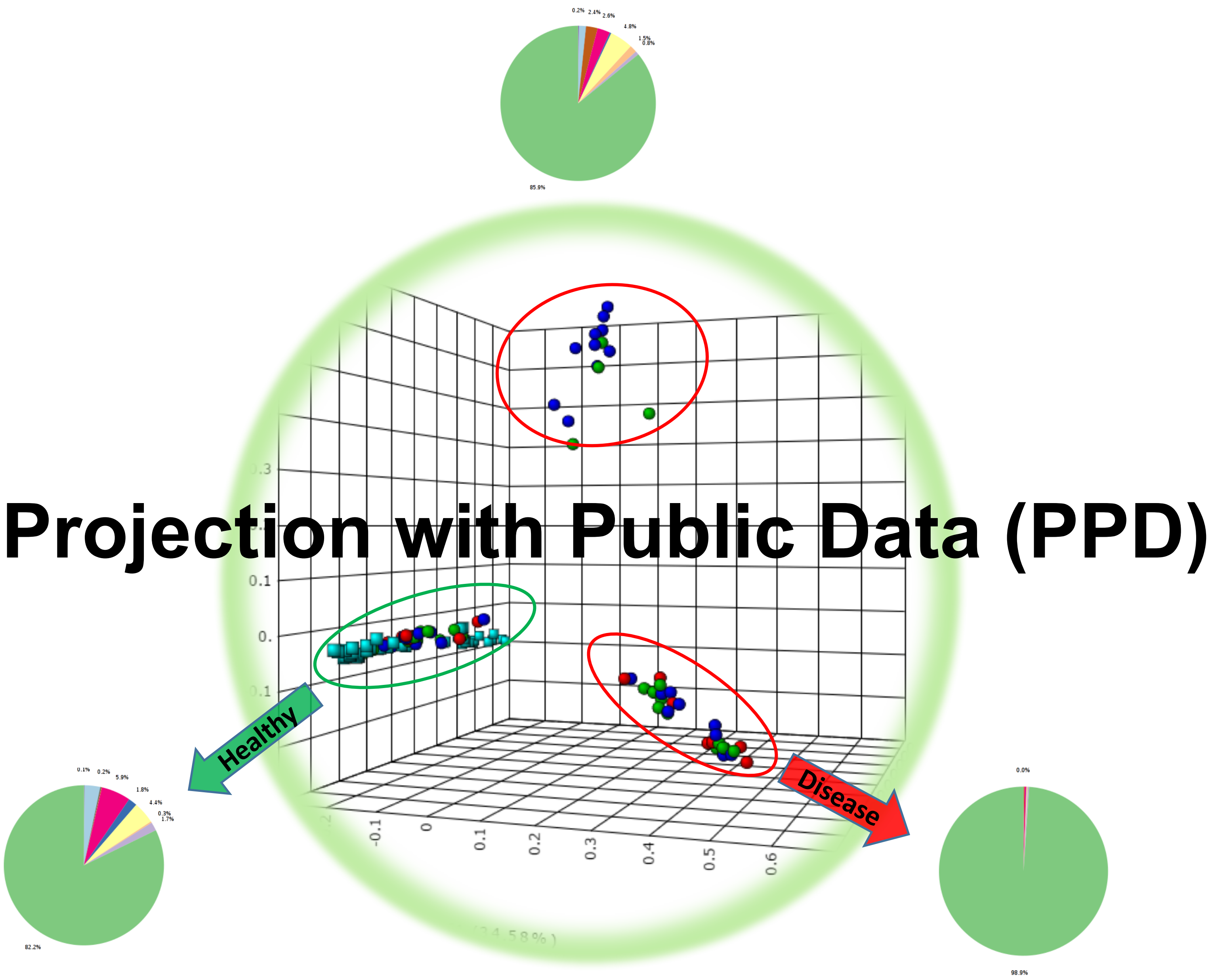
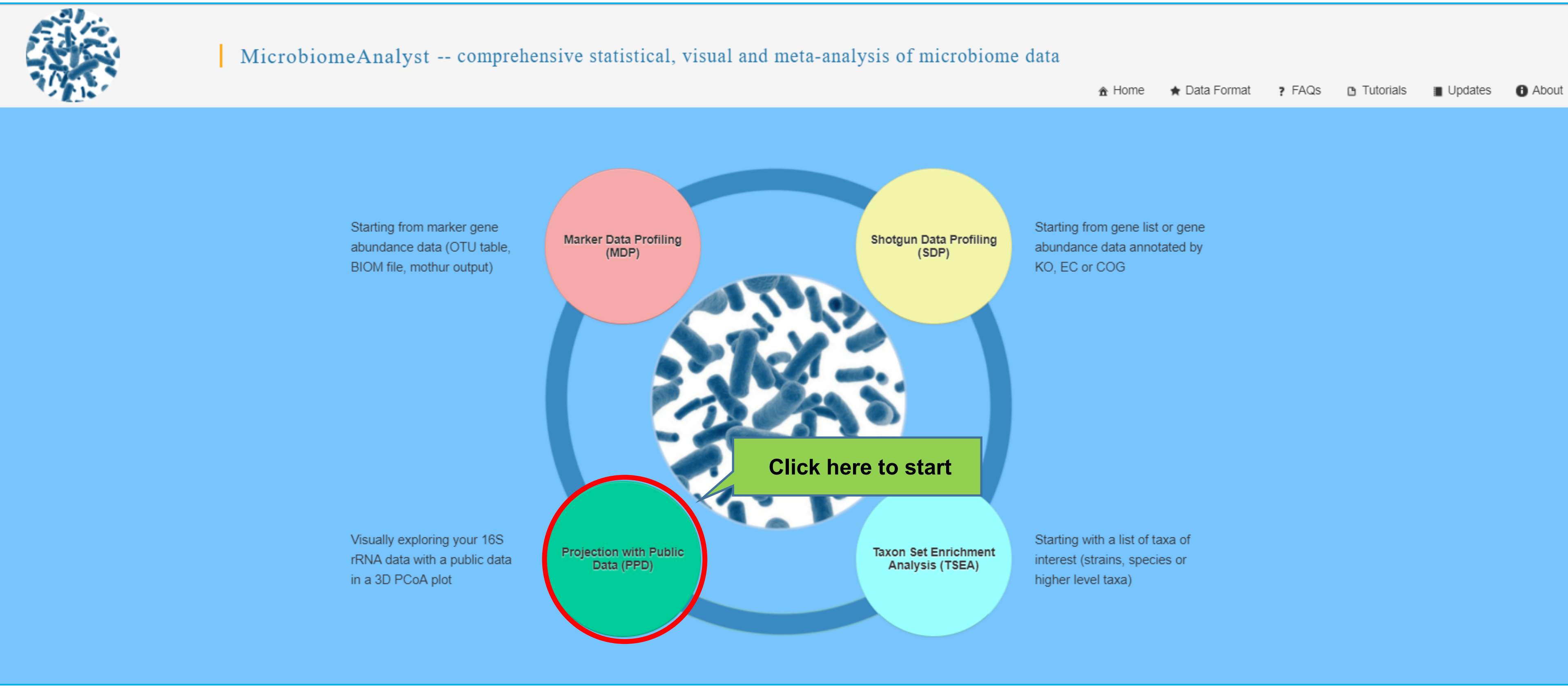
Projection with Public Data (PPD)



Goal

- To compare users’ 16S rRNA data with published datasets by processing and normalization them together, and projecting into 3D PCoA plot for visual comparative analysis



MicrobiomeAnalyst -- comprehensive statistical, visual and meta-analysis of microbiome data

Home
Data Format
FAQs
Tutorials
Updates
About
Marker Data Profiling (MDP)
Shotgun Data Profiling (SDP)
Projection with Public Data (PPD)
Taxon Set Enrichment Analysis (TSEA)

Starting from marker gene abundance data (OTU table, BIOM file, mothur output)

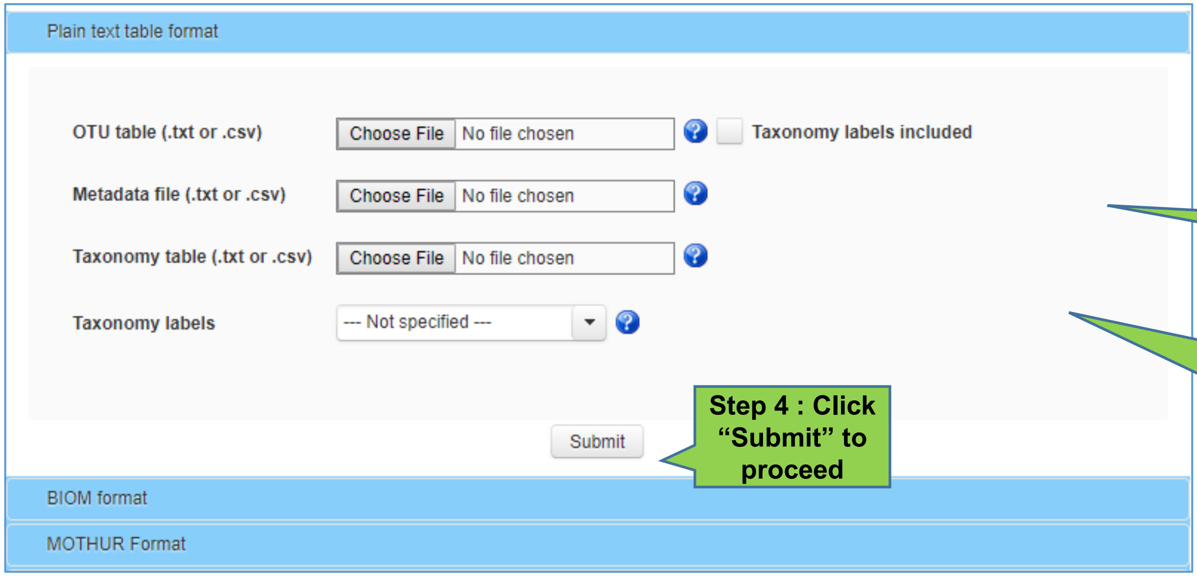
Starting from gene list or gene abundance data annotated by KO, EC or COG

Visually exploring your 16S rRNA data with a public data in a 3D PCoA plot

Starting with a list of taxa of interest (strains, species or higher level taxa)

Click here to start

1. Data Upload



Plain text table format

OTU table (.txt or .csv) | Choose File | No file chosen | □ Taxonomy labels included

Metadata file (.txt or .csv) | Choose File | No file chosen

Taxonomy table (.txt or .csv) | Choose File | No file chosen

Taxonomy labels | --- Not specified ---

Step 1: Upload your taxonomic profile data

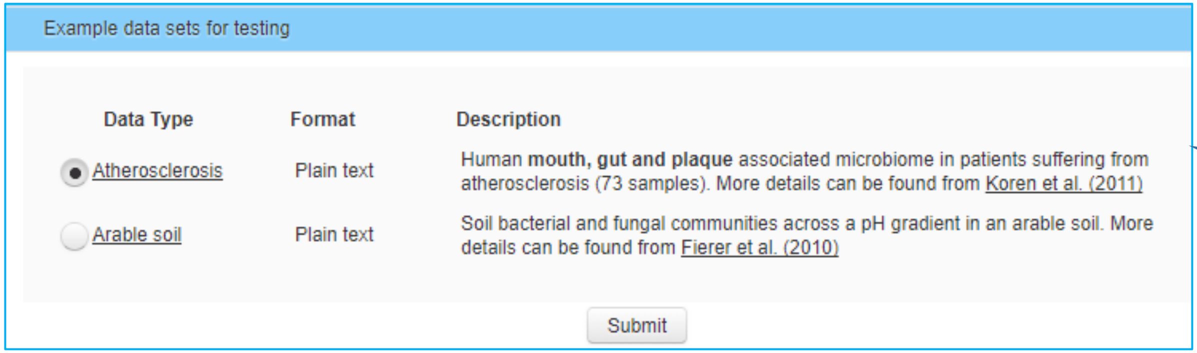
- 3 formats supported

Step 2: Upload your metadata file

Step 3: Upload taxonomic information (if not present within OUT table) and specify taxonomy annotation

- 3 IDs supported

Step 4 : Click "Submit" to proceed



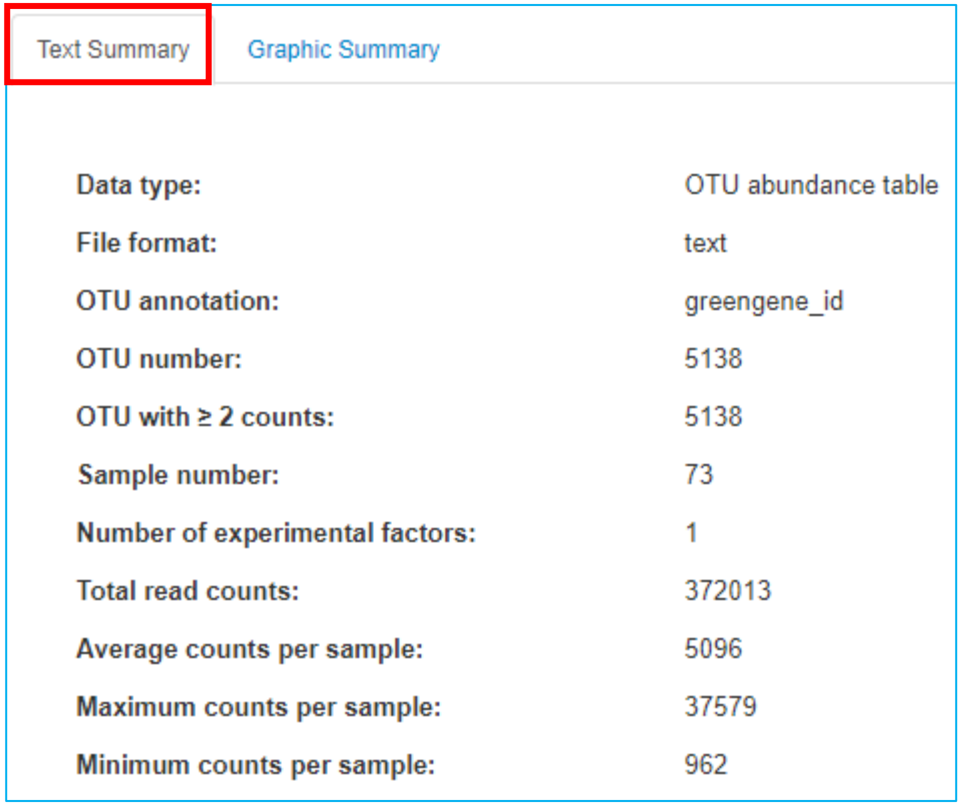
Example data sets for testing

Data Type	Format	Description
● Atherosclerosis	Plain text	Human **mouth, gut and plaque** associated microbiome in patients suffering from atherosclerosis (73 samples). More details can be found from Koren et al. (2011)
○ Arable soil	Plain text	Soil bacterial and fungal communities across a pH gradient in an arable soil. More details can be found from Fierer et al. (2010)

You can try our example also

Note: **Please check "Format" section of MicrobiomeAnalyst web server for details about each format.**

2. a) Data Integrity Check

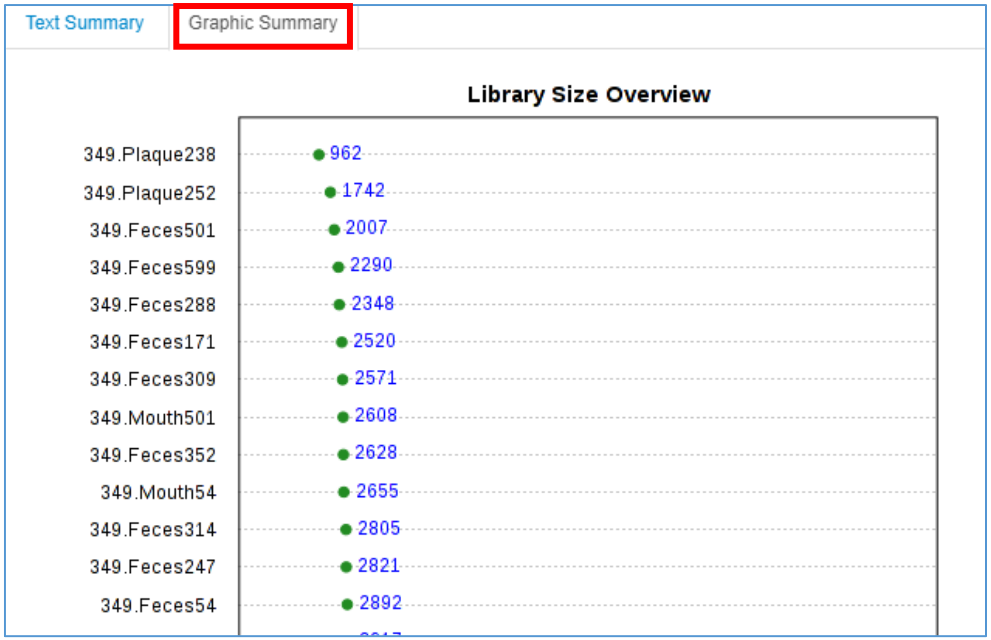


Data type:	OTU abundance table
File format:	text
OTU annotation:	greengene_id
OTU number:	5138
OTU with ≥ 2 counts:	5138
Sample number:	73
Number of experimental factors:	1
Total read counts:	372013
Average counts per sample:	5096
Maximum counts per sample:	37579
Minimum counts per sample:	962

Click “Proceed” if all information is correct.

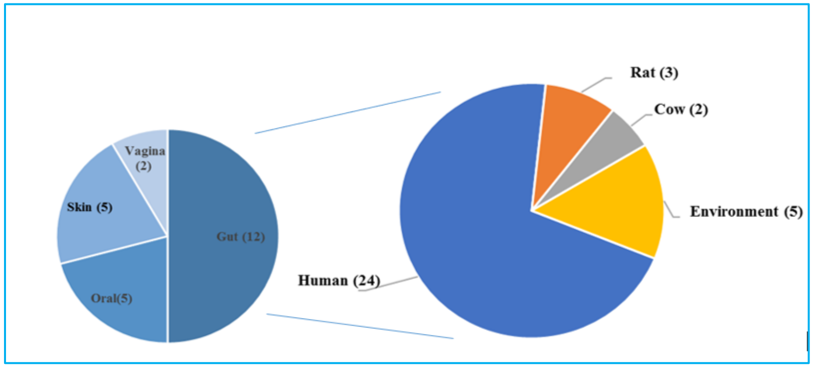
- Provides the information summary of users uploaded data and also check whether all files necessary for analysis are present or not.

2. b) Graphic Summary



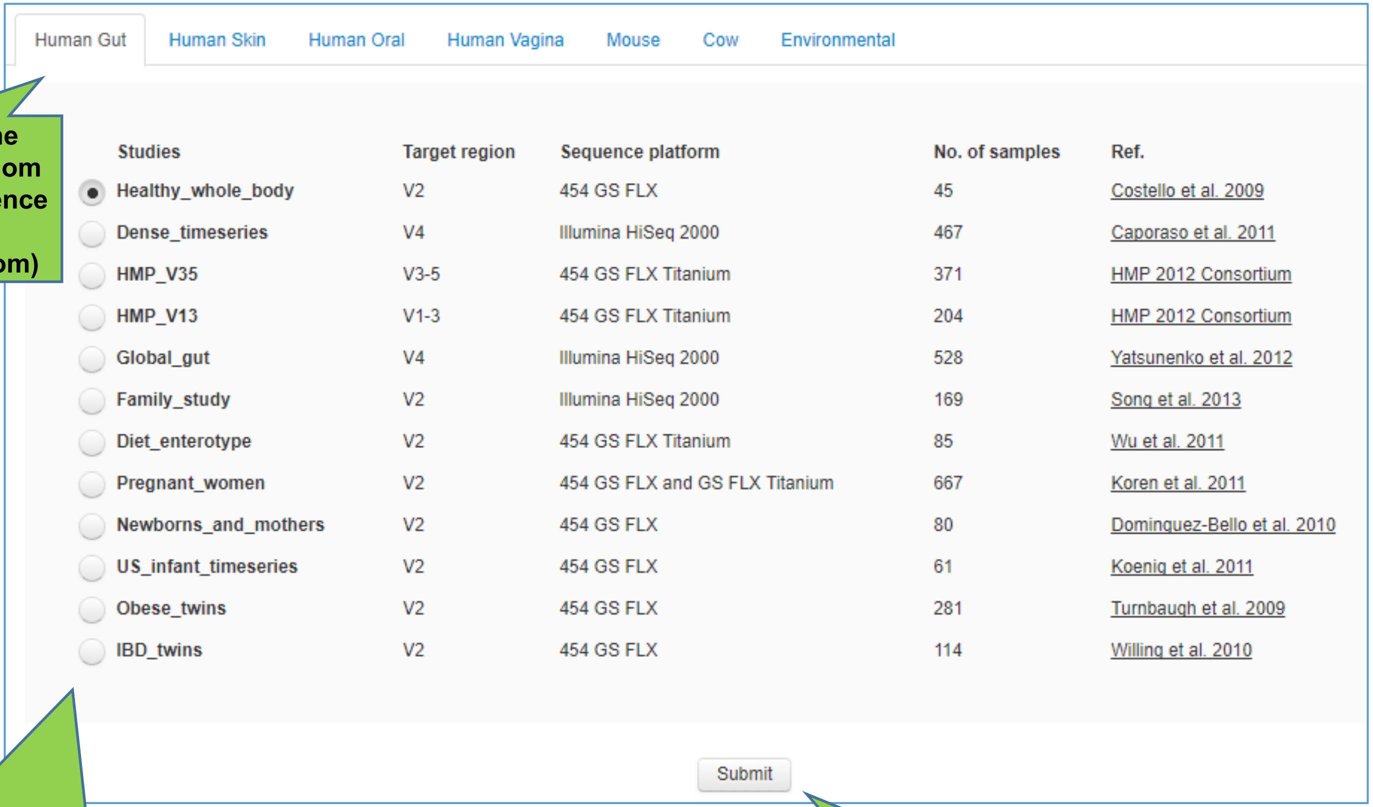
- Provides user the information about library size or total number of reads present in of each sample and help in identifying the potential outliers due to undersampling or sequencing errors.

3. Reference study selection



- In MicrobiomeAnalyst, human reference studies data has been downloaded from publicly accessible database (**QIITA**).
- These reference datasets has been primarily separated based on **the biom or site of sample collection**. Other than human, three others biom (including cow, mouse and environment) related datasets are available.
- Further, the datasets collected from human are partitioned based on **sampling body sites (4 body sites: gut ,oral etc.)** , **sequencing platform** (3 platforms: Illumina HiSeq 2000 , 454 GS FLX etc.), target region (4 regions : V1, V2-V3 etc.) and **factors shaping microbiota**. (8 factors: **diet**, **host genetics**, **cultural traditions and geography**, **age**, **pregnancy** etc.)
- Currently, we have around **34** datasets derived from **~20** large-scale studies.
- Users should try to match experimental protocols used in their data generation with the reference data for meaningful comparison.

3. Reference study selection



	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

Step 1 : Select the sampling site or biom of interest of reference study (studies from 4 biom)

Step 2 : Select the reference study you want to compare with your data. (Reference data: Greengenes id annotated)

Step 3 : Click "Submit" to proceed

4. Data processing

This is the most critical step in which:

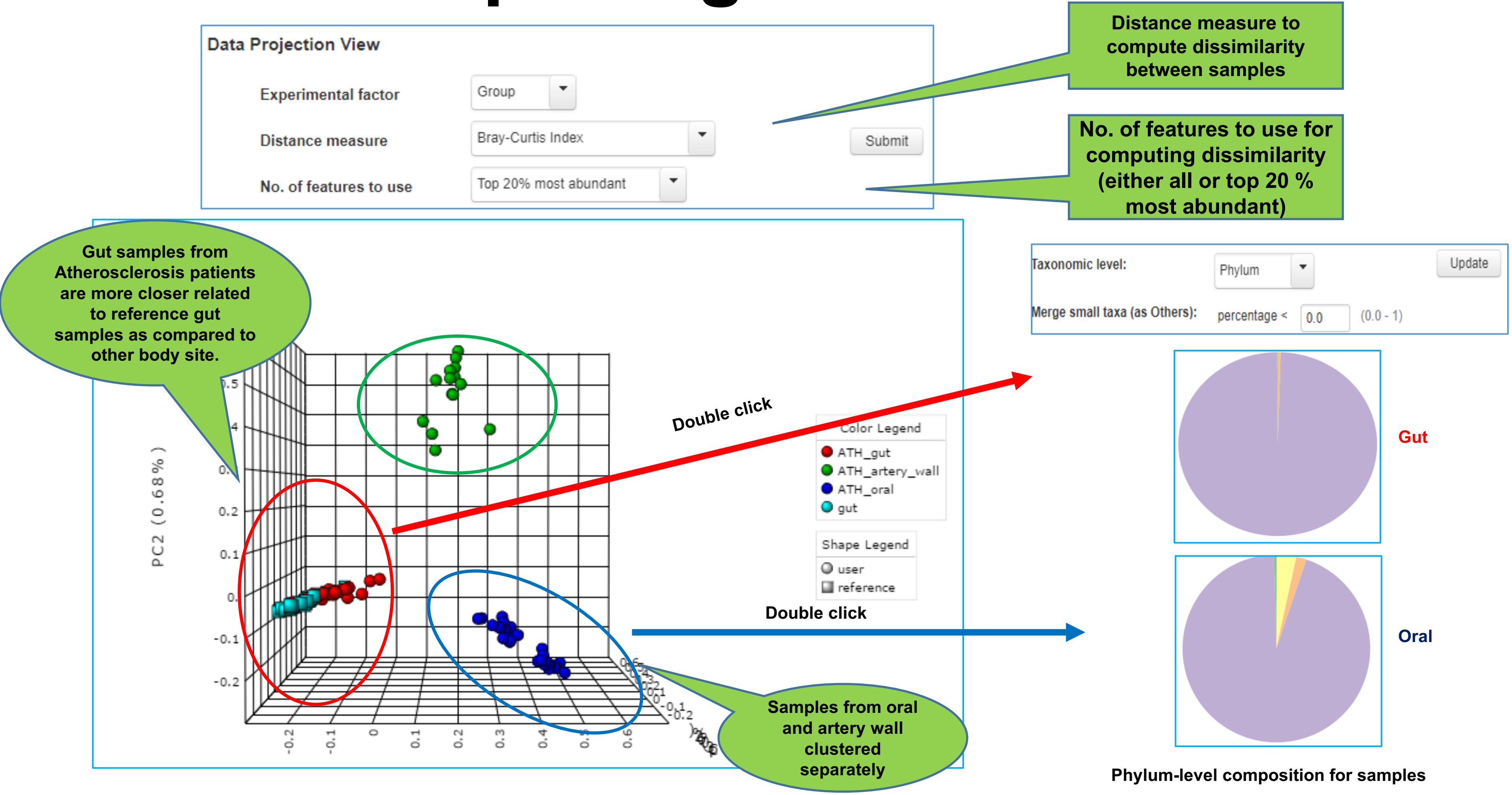
- User uploaded data will be merge user with selected reference data on the basis of common features (taxonomic labels). Even though, reference data are annotated with **Greengenes** database , user can also upload **SILVA** annotated data. (internal mapping will be performed on the basis of common **GenBank** id)

- There has to be significant features overlap (20%) between user and reference data for meaningful comparison. (otherwise, you can’t proceed)

- Merged data is processed and normalized together.

- User can choose from multiple distance matrices as well as number of features to keep for performing **PCoA** comparative analysis.

5. Visual exploring the result



Data Projection View

Experimental factor: Group

Distance measure: Bray-Curtis Index

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

Submit

Distance measure to compute dissimilarity between samples

No. of features to use for computing dissimilarity (either all or top 20 % most abundant)

Taxonomic level: Phylum

Update

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

Gut samples from Atherosclerosis patients are more closer related to reference gut samples as compared to other body site.

Double click

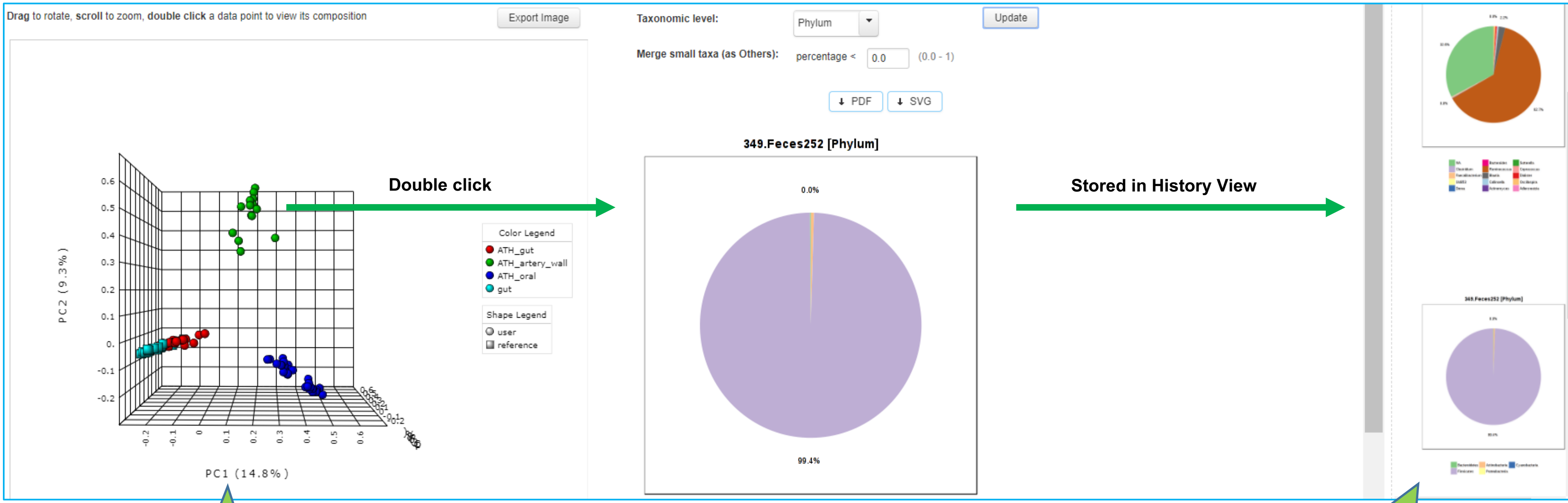
Double click

Samples from oral and artery wall clustered separately

Phylum-level composition for samples

- Result is visualized as **interactive 3D** plot in which user can found some interesting patterns in their data in comparison with reference (e.g. **healthy** vs **disease** samples clustering separately or **body sites** clustered separately)
- Users can double click individual data points (sample) to see the corresponding distributions of the taxa that underlying these difference at all possible taxonomic levels.

5. Visual exploring the result



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

Taxonomic level: Phylum

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

349.Feces252 [Phylum]

Double click

Stored in History View

3 D PCoA plot

current selected or double clicked sample composition

History View
all the images of the samples that have been inspected.

==END==