What is MicrobiomeAnalyst?

- "a user-friendly tool that integrates recent progress in statistics and visualization techniques, coupled with novel knowledge bases, to enable comprehensive analysis of common data outputs produced from microbiome studies" [1]
- 4 Modules
  - Marker Data Profiling (MDP)
  - Shotgun Data Profiling (SDP)
  - Taxon Set Analysis (TSA)
  - Projection with Public Data (PPD)

Module Details

- Marker Data Profiling (MDP): Facilitates analysis of user provided 16S rRNA marker gene survey data
- Shotgun Data Profiling (SDP): Contains functions for analyzing user provided metagenomics or metatranscriptomics data
- Taxon Set Enrichment Analysis (TSEA): Allows the user to test for biologically or ecologically meaningful patterns from a list of taxa of interest
- Projection with Public Data (PPD): Enables users to visually juxtapose their data with public datasets in search of novel patterns or biological insights

Data Input - MDP, PPD

- 16S rRNA maker gene survey data
- Inputs
  - Taxon or OTU abundance table
  - Metadata file
  - Phylogenetic tree (optional)
- Formats
  - BIOM
  - mothur
  - .txt
  - .csv

Data Input - SDP, TSEA

- SDP
  - Same as MDP, PPD
  - Also requires gene IDs
    - KEGG Orthology (KO)
    - Enzyme Commission (EC)
    - Cluster of Orthologous Groups (COG)
    - Other
- TSEA
  - List of taxa with a taxa type

Data Filtering

By default, features are:
- excluded if they have all zeros or only appear in a single sample (a.k.a. "minimally cleaned")
  - reserved for alpha diversity analyses to understand individual sample diversity
- filtered based on abundance levels and sample prevalence

Users can filter features that are:
- low count (via minimum count cutoff based on their mean/median values)
- low variance (based on their IQRs, SDs or CVs)

Data Normalization

- Rarefying
- Scaling
  - Total sum scaling (TSS)
  - Cumulative sum scaling (CSS)
  - Upper-quartile normalization (UQ)
- Transformation
  - Relative log expression (RLE)
  - Trimmed mean of M-values (TMM)
  - Centered log ratio (CLR)

Data Analysis

- Visual exploration
- Community profiling
- Clustering & correlation
- Comparison & classification
- Functional prediction
- We will be going over the first two in the Demo
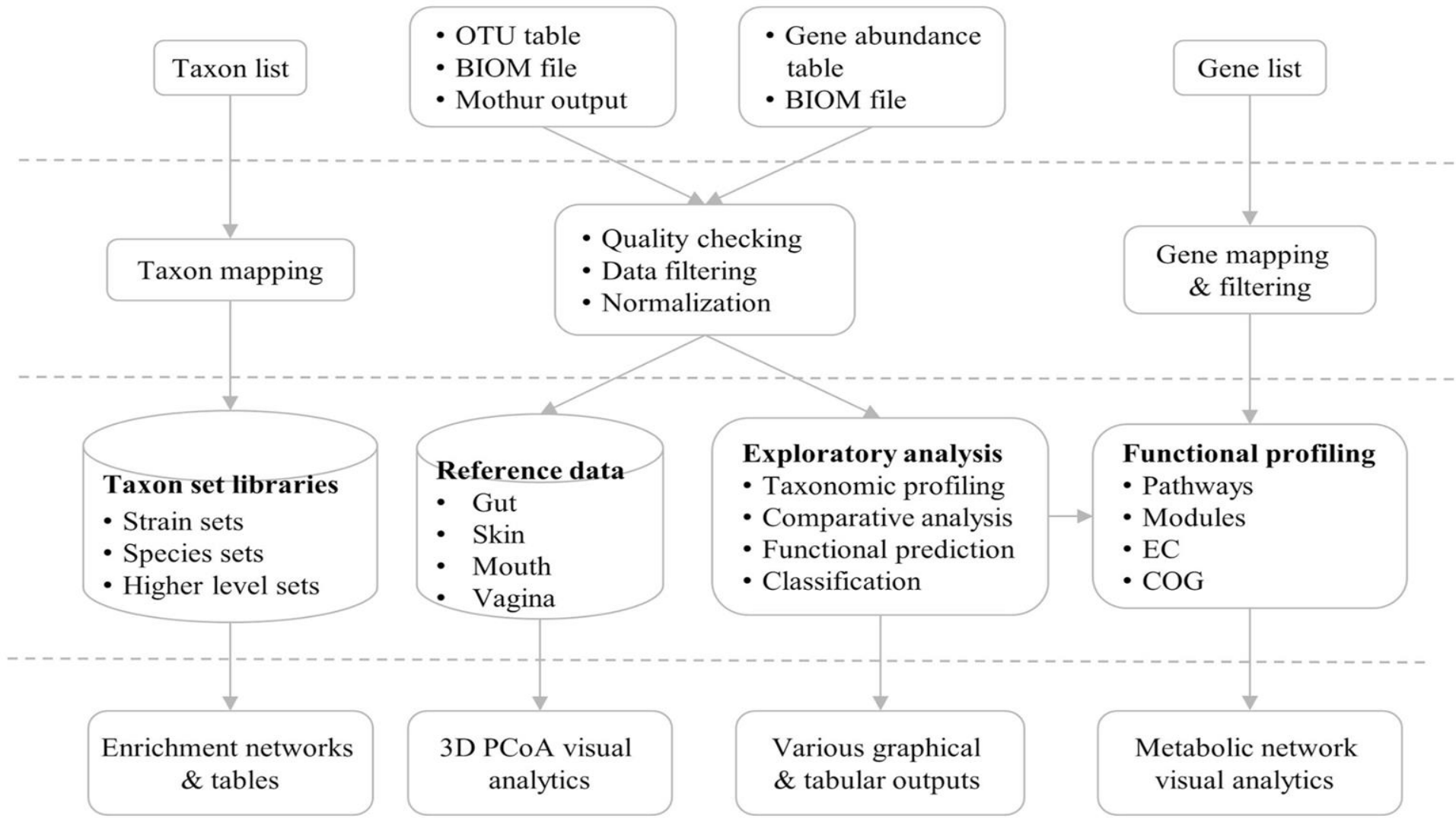
Clustering & Correlation

- Heatmap clustering
  - Allows for hierarchical clustering which is visualized through the use of a heatmap
- Dendrogram analysis
  - Creates a tree diagram that shows the hierarchical clustering of the data
- Correlation network (SparCC)
  - Can use different correlation algorithms (by default it uses SparCC) to analyze any correlations that exist in the data provided
- Pattern search
  - By entering a pattern (profile or specific feature) the user can perform a search

Comparison & Classification

- Univariate analysis
  - Explore variables in a data set separately
- metagenomeSeq
  - Determines differentially abundant features
- RNA-seq methods
  - Sequencing technique for identifying info about RNA in a sample
- LEfSe
  - Built for microbiome data to extract features for biomarker discovery
- Random Forest
  - Generic non-parametric machine learning algorithm

Functional Prediction

- PICRUSt (Greengenes)
    - Predict the functional composition of a metagenome
    - Uses marker gene data and database of reference genomes
- Tax4Fun (SILVA)
    - Predict functional potential based on minimum 16SrRNA sequence similarity
    - Works with data from SILVAngs web server or QIIME against SILVA database
    - Takes more time to complete (~2 mins)

Demo

https://www.microbiomeanalyst.ca/

Resources

https://academic.oup.com/nar/article/45/W1/W180/3760191