Microbiome Data Analysis with MicrobiomeAnalyst

User ID: guest11745102577535067247

May 25, 2020

1 Background

Metagenome-wide association studies have already found strong association of microbes with host health and disease. It also identifies a large number of microbes differentially regulated in various conditions. However, computational methods for analyzing such differentially regulated microbes from microbiome study are limited. TSEA or Taxon Set Enrichment Analysis is a way to identify biologically or ecologically meaningful patterns by analyzing them with context to pre-defined taxon set (microbes sharing some common trait) from a given list of significant features or microbes. approaches. In conventional approach, microbes are evaluated individually for their significance under conditions of study. Those microbes that have passed certain significance level are then combined to see if any meaningful patterns can be discerned. In contrast, TSEA directly investigates if a set of functionally related microbes without the need to preselect compounds based on some arbitary cut-off threshold. It has the potential to identify subtle but consistent changes among a group of related microbes, which may go undetected with the conventional approaches.

Essentially, TSEA is a microbiome version of the popular GSEA (Gene Set Enrichment Analysis) software with its own collection of taxon set libraries as well as an implementation of user-friendly web-interfaces. GSEA is widely used in genomics data analysis and has proven to be a powerful alternative to conventional approaches. For more information, please refer to the original paper by Subramanian A, and a nice review paper by Nam D, Kim SY [1].

2 TSEA Overview

Taxon set enrichment analysis consists of 4 steps - data input, data processing, data analysis, and results download. Based on the taxonomic resolution of microbes to analyze, three types of taxon sets are created and supported in MicrobiomeAnalyst Different taxon sets are selected based on different input types. Users can also perform taxon name mapping to higher taxonomic level and between a variety of microbes names and major database identifiers.

3 Data Input

There are three enrichment analysis algorithms offered by TSEA. Accordingly, three different types of data inputs are required by these three approaches:

- A list of microbes names characterized at any possible taxonomic level - entered as a one column data (*Mixed-level taxa*);
- A list of microbes names characterized at any species level - entered as a one column data (*Species-level taxa*);

[1] Dougu Nam Seon-Young Kim *Gene-set approach for expression pattern analysis*, Breif. in Bioinformatics 2008

- A list of microbes names (Binomial Nomenclature Name/GOLD ID/NCBI Taxonomy ID) characterized at any strain level - entered as a one column data (*Strain-level taxa)*)

4 Selection of Taxon Set Library

Depending upon type of input list, Taxon set library will be selected. There are four built-in libraries offered by TSEA:

- Mixed-level Taxon sets associated with human genetic variations - used with mixed-level taxa (*currently contains 1520 entries*);
- Mixed-level taxon sets associated with human diseases - used with mixed-level taxa (*currently contains 39 entries*);
- Species-level Taxon sets associated with human physiology, development, life styles etc. - used with species-level taxa) (*currently contains 170 entries*)
- Strain-level Taxon sets based on their shared phenotypic traits or ecological niches - used with strain-level taxa (*currently contains 100 entries*)

You have provided a list of microbes annotated at mixed-level. Mixed-level taxon sets associated with Human genetic variations will be used for performing enrichment analysis.

5 Data Processing

The first step is to match the user's entered microbes name with the microbes contained in the Taxon set library. All the microbes name should be in universally accepted format (Binomial Nomenclature). TSEA also provides conversion between microbe names and other database identifiers such as NCBI Taxonomy ID and GOLD ID. The unmapped name will be indicated by a - *or empty cell* and will be removed from further analysis.

Table 1: Result from Taxa Name Mapping

	Query	Match	Species	Genus	NCBI_Taxonomy_ID
1	Bacteroides	Bacteroides	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
2	Bacteroides uniformis	Bacteroides uniformis	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
3	Faecalibacterium prausnitzii	Faecalibacterium prausnitzii	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
4	Prevotella copri	Prevotella copri	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
5	Ruminococcus	Ruminococcus	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
6	Bacteroides eggerthii	Bacteroides eggerthii	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B
7	Prevotella stercorea not recognized		-	-	-
8	Blautia	Blautia	-	-	<a href=https://www.ncbi.nlm.nih.gov/Taxonomy/B

6 Enrichment Analysis

Over Representation Analysis (ORA) is performed when a list of taxa or microbes is provided. The list of microbes can be obtained through differential abundance testing, or from biomarker analysis or from a clustering algorithm performed using MDP to investigate if some biologically meaningful patterns can be identified.

ORA was implemented using the *hypergeometric test* to evaluate whether a particular Taxon set is represented more than expected by chance within the given compound list. One-tailed p values are provided after adjusting for multiple testing. **Table 2** below provides the detail about enriched taxon set.

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
RFT1	3.00	0.08	2.00	0.00	1.00	1.00
LINGO2	1.00	0.03	1.00	0.03	1.00	1.00
RP11-521D12.1; SLC9A2	1.00	0.03	1.00	0.03	1.00	1.00
ADCK4	1.00	0.03	1.00	0.03	1.00	1.00
CACNA1E	1.00	0.03	1.00	0.03	1.00	1.00
PNPLA7	1.00	0.03	1.00	0.03	1.00	1.00
SLC45A1	1.00	0.03	1.00	0.03	1.00	1.00
TFR2	1.00	0.03	1.00	0.03	1.00	1.00
UTP14C	1.00	0.03	1.00	0.03	1.00	1.00
ADAD2	2.00	0.05	1.00	0.05	1.00	1.00
CH25H	2.00	0.05	1.00	0.05	1.00	1.00
FRMPD1	2.00	0.05	1.00	0.05	1.00	1.00
GSDMB	2.00	0.05	1.00	0.05	1.00	1.00
LGI4	2.00	0.05	1.00	0.05	1.00	1.00
SLC17A9	2.00	0.05	1.00	0.05	1.00	1.00
BCL11A	3.00	0.08	1.00	0.07	1.00	1.00
SEC14L6	3.00	0.08	1.00	0.07	1.00	1.00
SMTNL2	3.00	0.08	1.00	0.07	1.00	1.00
SRA1	3.00	0.08	1.00	0.07	1.00	1.00
ADAMTS14	4.00	0.10	1.00	0.10	1.00	1.00
ALOX15B	4.00	0.10	1.00	0.10	1.00	1.00
C19orf45	4.00	0.10	1.00	0.10	1.00	1.00
CACNA1H	4.00	0.10	1.00	0.10	1.00	1.00
COCH	4.00	0.10	1.00	0.10	1.00	1.00
CSPG4	4.00	0.10	1.00	0.10	1.00	1.00
JPH3	4.00	0.10	1.00	0.10	1.00	1.00
PDZRN3	4.00	0.10	1.00	0.10	1.00	1.00
SYTL1	4.00	0.10	1.00	0.10	1.00	1.00
ZNF280A	4.00	0.10	1.00	0.10	1.00	1.00
ADAMTS4	5.00	0.13	1.00	0.12	1.00	1.00
C10orf11	5.00	0.13	1.00	0.12	1.00	1.00
CDH20	5.00	0.13	1.00	0.12	1.00	1.00
LRFN2	5.00	0.13	1.00	0.12	1.00	1.00
MASTL	5.00	0.13	1.00	0.12	1.00	1.00
SH3RF2	5.00	0.13	1.00	0.12	1.00	1.00
AZI1	6.00	0.15	1.00	0.14	1.00	1.00
FLNB	6.00	0.15	1.00	0.14	1.00	1.00
PLXNC1	6.00	0.15	1.00	0.14	1.00	1.00
SOX8	6.00	0.15	1.00	0.14	1.00	1.00
KCNIP4	7.00	0.18	1.00	0.17	1.00	1.00
MANBA	7.00	0.18	1.00	0.17	1.00	1.00
NLRP13	7.00	0.18	1.00	0.17	1.00	1.00
PIWIL4	7.00	0.18	1.00	0.17	1.00	1.00
PRKCDBP	7.00	0.18	1.00	0.17	1.00	1.00
AGBL2	8.00	0.20	1.00	0.19	1.00	1.00
ATP7B	8.00	0.20	1.00	0.19	1.00	1.00
SLC5A8	8.00	0.20	1.00	0.19	1.00	1.00
AGT	9.00	0.23	1.00	0.21	1.00	1.00
CHPT1	9.00	0.23	1.00	0.21	1.00	1.00
HAPLN3	9.00	0.23	1.00	0.21	1.00	1.00
LAMB1	9.00	0.23	1.00	0.21	1.00	1.00
NTNG2	9.00	0.23	1.00	0.21	1.00	1.00
PNPLA6	9.00	0.23	1.00	0.21	1.00	1.00
USP53	10.00	0.25	1.00	0.23	1.00	1.00
EHD2	11.00	0.28	1.00	0.25	1.00	1.00

The report was generated on Mon May 25 04:53:37 2020 with R version 3.6.3 (2020-02-29).