Readme for HMPdistance(Level, Data, Distance\_measure)

PURPOSE

This function computes the mean distance dstool of a microbiome sample (or samples) to a reference set of 92 stool samples from HMP (Human Microbiome Project) and the mean distance dnasal to a reference set of 74 HMP nasal samples.

ARGUMENTS

Level

Level defines the taxonomic level for which the distances are computed. Legitimate values of Level are ‘L2.phylum’, ‘L3.order’, ‘L4.class’, ‘L5.family’ and ‘L6.genus’.

Distance\_measure

Distance\_measure = Bray-Curtis or Pearson. Default Bray-Curtis. Legitimate argument values are ‘bc’ or ‘corr’. These are not metrics and should perhaps be called dissimilarity measures instead. If  are two  composition vectors, their Bray-Curtis distance is  The Pearson distance is , because the means  each equal 

Data

Data is a matrix in R. There are *n* rows in Data, each corresponding to a different microbiome sample. There are *p* columns in Data corresponding to different taxa at a given taxonomic level. In a given row, each column contains the relative abundance of its taxon. At the genus level, Greengenes 13.8 (GG13.8) includes *p=*2929 different genus sequences, including sequences with missing elements. For example, one such sequence is k\_\_Bacteria. p\_\_Actinobacteria.c\_\_Acidimicrobiia. o\_\_Acidimicrobiales. f\_\_Acidimicrobiaceae.g\_\_Ferrimicrobium which gives the kingdom (Bacteria or Archaea), phylum, class, order, family and genus. Another “genus” in GG13.8 is k\_\_Bacteria. p\_\_Acidobacteria. c\_\_TM1.o\_.f\_.g\_ which has missing information at the order, family and genus levels but is a unique genus sequence in GG13.8. GG13.8 does not include all possible missingness patterns, but only those that have been found in the collection of OTUs in GG13.8. Likewise, at the phylum level, GG13.8 includes *p*=91 kingdom/phyla. A list of the 2929 sequences at the genus level and the sequences at the family (*p*=1116), order (*p*=664), class(*p*=319) and kingdom/phylum (*p*=91) are given in order in the attached file GG\_13\_8.xlsx.

If one uses the GG13.8 closed reference library to classify OTUs in a given sample, the pipeline will discard OTUs that do not match one of the 99,322 OTUs in GG13.8. At a given taxonomic level (e.g. phylum or genus), the relative abundance each of the remaining OTUs will be placed in one of the taxa (sequences) in GG13.8. For each sample, the row sum will be 1.0. Data will be in a file called d.new.filename = new data csv filename (together with path if needed) or rdata file. This is an *n* x *p* matrix of relative abundances in the new dataset, as described above. The matrix can contain other information in the columns, such as sample ID, etc. The columns that are not relative abundance columns should be specified in a vector of indices passed to d.new.ix.col.not.rel.abu .

If the user does not employ GG13.8, the following conventions could be used.

1. Suppose the user knows only the relative abundances for the kingdom/phylum level and wants to compute distances at the kingdom/phylum level. Then the user can enter these sequences using exactly the spelling and order in the kingdom/phylum columns of GG\_13\_8.xlsx . No information on order, class, family and genus are entered (not even missingness codes). If the kingdom is known but the kingdom/phylum is not among the 91 patterns in GG13.8, including the missing patterns k\_Archaea.p\_ and k\_Bacteria.p\_, then the program will recognize that the phylum is not known and its relative abundance will be put in k\_Kingdom.p\_ together with the relative abundance from OTUs with truly missing phylum.
2. Suppose the user has used a different system to establish the taxonomy for the OTUs down to the genus level. In order to use this algorithm to compute distances at the genus level, the user must:
   1. Use the naming conventions as in GG13.8 (also shown in GG\_13\_8\_taxonomy.xls) for each level. If a particular genus is not among the 2929 sequences in GG13.8, then write the genus name anyway. The program will recognize that is not in GG13.8.
   2. Any correctly spelled genus sequence that is not in GG13.8 will be assigned missing values up to the lowest level at which the sequence does match a sequence in GG13.8. Let .K.P.O.C.F.G denote a particular sequence in GG13.8 and let .K.P.O.C.F\*.G\* denote a sequence to be classified at the genus level but which itself is not in GG13.8 because of mismatches at the family and genus level, but which matches at the kingdom, phylum, order and class levels. It will be assigned K.P.O.C.f\_.g\_ if that is a sequence in GG13.8. It might be that this particular missingness pattern is not in GG13.8, in which case it will be assigned .K.P.O.c\_.f\_.f\_. If this sequence is not found in GG13.8, the next higher level of missingness will be searched. Eventually be such a match will be found because GG13.8 contains .K.p\_.o\_.c\_.f\_.g\_.
   3. For computing distances at the phylum level, if a correctly spelled sequence is not among the 91 kingdom/phyla in GG13.8 but its kingdom is known, it will be assigned .K.p\_ .
   4. Similar comments apply to other intermediate taxonomic levels.

OUTPUT

The program returns a n samples x 2 matrix of distances to stool and nasal (ie. with columns dstool, dnasal).

EXAMPLES

The following example shows how to call the program to compute distances for two samples. The example illustrates performance for a user of GG13.8.

In the paper, the mean distance for each sample of Baxter’s project was used for the discussion. The 16S data was processed using GG 13\_8 database independently from HMP set.

Here is the function to generate distance to HMP stool and HMP nasal at a given taxonomy level.

HMPdistance(tax.level ='l2.phylum',  
 d.new.filename = ""HMP\_L2\_030317\_phylum.csv" ",   
measure = 'bc', #specifies Bray-Curtis distance  
 d.new.ix.col.not.rel.abu = 1:4, #columns that do not contain relative abundance data  
 print.details = F)

Results for the first 6 samples:

dist.to.hmp.stool dist.to.hmp.nasal

1 0.2205655 0.7363895  
2 0.1886405 0.7537370  
3 0.4300298 0.6236202  
4 0.8891262 0.5393743  
5 0.8002390 0.2495285  
6 0.1935302 0.7418604

Reanalysis for Pearson at phylum level for the dataset from Baxter et al.

HMPdistance(tax.level ='l2.phylum',

d.new.filename = "Baxter\_L2\_rel\_abundance.csv", # filename of the dataset from Baxter et al  
measure = 'corr' , d.new.ix.col.not.rel.abu = 1:4, print.details = F)

dist.to.hmp.stool dist.to.hmp.nasal

1 0.2014029 0.6639146   
2 0.2353937 0.6156050  
3 0.4965863 0.6233213  
4 0.5338425 0.5765289  
5 0.1386081 0.6831418  
6 0.5358267 0.4076855

Notes for running functions:

1. For input file d.new.filename, is either a csv file or an rdata file.
2. d.new.filename should only contain relative abundances at the taxonomic level specified. For example if Level= ‘L3.class’, there are 319 GG13.8 columns for taxa, and these taxa sequences include kingdom.plylum.order.class but not family or genus data.
3. If file d.new.filename contains additional columns with metadata but not relative abundance data, the vector of indices of those columns should be passed to d.new.ix.col.not.rel.abu as illustrated in the examples. If for example columns 1:4 and 50:60 contain metadata, set d.new.ix.col.not.rel.abu = c(1:4,50:60)
4. For Bray-Curtis distance, use measure = “bc”. For Pearson distance, use measure = “corr”. Default is Bray-Curtis distance.
5. All arguments except for filenames and directory paths are case-insensitive.