DESCRIPTION OF DATA OBJECTS AND CODES USED IN THE CURATED METAGENOMIC DATA ANALYSIS

Description of RData workspace:

The following table lists the objects contained in the CMDAnalysis.RData object

|  |  |  |
| --- | --- | --- |
| Sl. No | Object Name | Description |
| 1 | combined\_species\_profile\_with\_age\_country\_final | Matrix containing the species abundance for the samples |
| 2 | core\_species | List of 836 species occurring with a minimum abundance of 0.01 across at least 5% of the samples |
| 3 | combined\_species\_profile\_with\_age\_country\_log1 | Log transformed abundance of species with base 10, zero abundances replaced by 0.000001 |
| 4 | combined\_short\_metadata\_2, combined\_short\_metadata\_final | Various metadata files |
| 5 | Young, Middle, Elderly | List of individuals classified as Young (20-39), Middle (40-59) and Elderly (60 or greater) |
| 6 | SelectControls | Selected Controls (as described in the Methods section of the manuscript) |
| 7 | IBDIndividuals, CRCIndividuals, AdenomaIndividuals, T2DIndividuals and CirrhosisIndividuals | List of individuals belonging to different disease categories |
| 8 | FranzosaEA\_2018Individuals, NielsenHB\_2014Individuals, KarlssonFH\_2013Individuals, QinJ\_2012Individuals, QinN\_2014Individuals, VogtmannE\_2016Individuals, ZelleG\_2014Individuals, FengQ\_2015Individuals | List of individuals belonging to each study cohort |
| 9 | AllEUIndividuals, NorthAmericaIndividuals, AsiaIndividuals | List of Individuals belonging to each continent |
| 10 | em\_anthro\_with\_strata | Anthropometric Metadata for the ELDERMET subjects |
| 11 | EMCommunityIndividuals, EMLongstayIndividuals | List of Eldermet subjects belonging to the Community (Community + Daycare) and the Longstay (Rehab + Longstay) groups |

**List of Programs**

|  |  |  |
| --- | --- | --- |
| Sl. | Program Name | Function |
| 1 | GroupDataBins.R | Group samples into country, continent and continent level bins |
| 2 | AdonisAcrossDiseaseIntAgeGroup.R | Performs PERMANOVA analysis on the effect of age-group specific disease signatures after taking care of all confounders |
| 3 | IterativeRFAcrossAgeGroups.R | Iterative random forest analysis trained on one age-group and tested on the same or different age-group |
| 4 | SpecificMarkerList.R | Identifies Top 85 percentile markers for each category |
| 5 | AgeGroupSpecificMarkerIdentification.R | Age group specific markers identified using RF |
| 6 | IBD\_LM.R, Adenoma\_LM.R, CRC\_LM.R, T2D\_LM.R and Cirrhosis\_LM.R | Linear model-based validation of age-differential markers using Linear Models |
| 7 | EffSizeCalculator.R | Calculates the cohens’ D between disease and healthy for each scenario |
| 8 | DirectionalityComputeFinal.R, DirectionalityComputeCountryCohort.R, DirectionalityComputeStudyCohort.R | Programs for identifying the significantly different features along with their directionality by grouping samples to continent level (Final), country level (Country) and Study level (Study) bins |
| 9 | GenericDiseaseClassification.R | Program for running the Generic Disease Classification analysis |
| 10 | function\_library.R | Contains the library of in-house developed functions as well as the libraries required |