Description of Supplementary Data

This document describes the contents of each file included in the Supplementary Data folder of the paper by Metcalfe-Roach et. al (2023). Key terms are also defined.

# Part 1 – Folder/file descriptions

The Supplementary Data folder contains six folders and one file. Folders contain all data relevant to a specific analysis. Folders and the file are as follows, in alphabetical order:

## Differential Abundance

Contains the results of all differential abundance tests for taxonomic and functional datasets. All tests examined correlations between metagenomic data and disease status (PD vs Ctrl).

* ‘\_Stats’ files list the individual test statistics for each term and differential abundance tool (ANCOM-BC, ALDEx2, MaAsLin2).
* ‘Stratified’ files list the individual linear model statistics for each term within the taxon-stratified functional datasets
* All Excel files include univariate and multivariate results, with the exception of taxonomy – as four taxonomic levels were assessed, models were separated into separate files to avoid cluttering.

## Enrichment Analysis

Contains the full statistics of the enrichment analyses performed on each dataset. In short, for each dataset, terms were grouped into functional categories. The proportion of terms that were either enriched or depleted (uncorrected p value < 0.10) within the category was compared to the proportion of enriched or depleted terms within the entire dataset. This is calculated using one-sided Fisher’s exact tests.

## Metabolite Associations

This contains all correlations between taxonomic or functional data and the major metabolites of interest, p-cresol and phenylacetylglutamine. Correlations with C reactive protein serum abundances are also included within the same datasets. The multivariate statistics were used to create Figure 3 and Figure S4.

## Network Analysis

* Network Analysis Statistics: this file contains all statistics represented in Figure 1, as well as statistics depicting that the number of taxa which had zero connections was not different between PD and control networks (final tab, “No Figure (Prop. 0)”).

## Progression

This contains three types of data. Disease metrics include MDS-UPDRS scores (totals for each Part and overall) and levodopa equivalent dose. For all analyses represented, **only those with longitudinal data** were included:

* **“0. Progression of Disease Severity and Medication”:** this file describes the overall symptom progression of all participants with PD over the course of the study. Tab 1, **“Avg Values Per Visit**”, denotes the mean (standard deviation) of each disease metric for each follow up visit. Tab 2, **“Change Per Year & Tertile”,** divides people into Slow, Med, and Fast progression groups based on MDS-UPDRS Part 4 progression (see Random Forest section of paper). **“Prog Rates – Mixed Model”** uses mixed models to calculate the univariate progression rate of each metric over time. Estimates represent the mean change per month of followup. **“Scaled Prog Rates – Mixed Model”** is identical to the previous tab, but all metrics were scaled to a 0-1 range prior to analysis to allow for the direct comparison of estimates. Larger estimates indicate faster progression relative to other metrics.
* **“Disease Progression Stats”:** Mixed model statistics that compare metagenomic abundances to the rate of PD progression.
* **“Baseline Severity Stats”:** Linear model statistics that compare metagenomic abundances to baseline disease metrics.

## Random Forest

This folder contains the output weights of all random forest analyses, in the form of importance values for each variable within each random forest. Larger importance values indicate that the variable of interest contributed more to the random forest model, and is thus considered to be more relevant to the outcome variable.

## Diversity of Functional Data (singular file)

Only one file was generated for diversity analysis, and so no folder was generated. This file includes univariate and multivariate statistics that correspond to the Shannon alpha diversity and Bray-Curtis beta diversity metrics calculated for each functional dataset.