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).
  + 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.
* ‘Stratified’ files lists each functional term alongside the individual taxa that contribute to its read count.
  + ‘Avg Proportion’ tab: for each feature/taxon combination, the average proportion (n=276) of the functional reads deriving from the given taxon are provided along with the standard deviation and minimum/maximum proportions observed.
  + “Min 33% of Fun Term”: The number of times a taxon has a mean proportion of at least 0.33 was tallied and compared to the total number of functional terms assessed (i.e. all differentially abundant terms). Only taxa with n>0 tallies were tabulated.
* “Taxonomy\_Abundance\_and\_Prevalence”: This denotes the prevalence and relative/CLR-transformed mean abundances for each microbe in the dataset. Values were calculated separately for PD (n=176) and controls (n=100). Log 2 fold changes are provided for relative abundance measures, and the difference in CLR abundances is analogously provided for CLR-transformed abundances.

## Diversity of Functional Data

These files include univariate and multivariate statistics that correspond to the Shannon alpha diversity and Bray-Curtis beta diversity metrics calculated for each functional dataset.

## 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. “Multivar Species” tabs are identical to “Multivar”, but are filtered to include the metagenomic explanatory variable only (no covariables).

## Network Analysis

* Network Analysis Species-Level Data: This lists the betweenness centrality, closeness centrality, and the distances of each microbe/microbial connection.
* 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)”). A list of all species included in the network analysis is also provided.

## Progression

This contains three types of data. Disease metrics include MDS-UPDRS scores (totals for each part and overall), Hoehn & Yahr, and levodopa equivalent dose. For all analyses, **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). Progression rates are then reported as the mean (standard deviation) change per year for each disease metric/progression tertile combination. **“Prog Rates (Table S2)”** is equivalent to Table S2 and 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”** 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.

## Validation of Statistical Tests

This folder contains all supplementary tests performed in order to validate the use of statistical tools such as Wilcoxon rank-sum tests, linear regression, and linear mixed models.

# 2 – Data Dictionaries

All terms used in the supplementary data are defined below.

## Differential Abundance

“\_Stats” files:

* Species: the taxonomic species or the functional term being assessed as the outcome variable.
* DA: the differential abundance tool.
* Variable: the explanatory variable. StatusPD refers to disease status, depth is sequencing depth, SexMale is self-reported biological sex, and laxatives is laxative use (yes/no), where ‘yes’ is the level corresponding to the statistics.
* p\_val and q\_val: P and corrected P (Q) values
* estimate, se: the effect size and standard error of each test
* association\_dir: the association direction of the correlation. If the estimate is negative, so is the association direction, and vice versa.

“Stratified” files:

* Avg Proportion:
  + feature: the functional term being assessed.
  + Taxon: the taxonomic group that the functional reads are assigned to
  + Proportion (mean, sd, min, max): The proportion of functional reads attributable to the given taxon (n=276).
* Min 33% of Fun Term:
  + Taxon: the taxonomic group that the functional reads are assigned to
  + Denominator column: within all differentially abundant functional terms in the dataset (denominator), the taxon contributed an average of >33% of the reads for n functional terms.

Taxonomy\_Abundance\_and\_Prevalence:

* Rel\_Ab\_Ctrl/PD: the mean relative abundance for Ctrl or PD
* Log 2 FC (PD/Ctrl): log2(Rel\_Ab\_PD/Rel\_Ab\_Ctrl)
* CLR\_Ab\_Ctrl/PD: the mean CLR-transformed abundance for Ctrl or PD
* CLR Change (PD-Ctrl): CLR\_Ab\_PD – CLR\_Ab\_Ctrl
* Prevalence\_Ctrl/PD: the proportion of Ctrl or PD samples that have measurable levels of the given taxon.

## Enrichment Analysis

* Dir: associaton direction of the enrichment. Negative = depletion, Positive = enrichment.
* Level: denotes the hierarchal level being tested. COG only has one level. EC has three levels (ex. EC1 vs EC1.1 vs EC1.1.1), and KO has four. L1 corresponds to the broadest hierarchal level. Testing of KO categories was restricted to the level 1 KEGG Orthology module, so L1 was not tested for this dataset.
* Test: the functional category being tested
* pval, qval: P and corrected P (Q) values
* estimate: the effect size of each test, as reported by the Fisher’s exact test output
* total\_sig, total\_ns: the total number of enriched or depleted functional terms within the entire dataset, and the total number of non-enriched or depleted functional terms within the entire dataset
* pct\_total: 100\*total\_sig/(total\_sig+total\_ns)
* pwy\_sig, pwy\_ns: the total number of enriched or depleted functional terms within the functional category of interest, and the total number of non-enriched or depleted functional terms within the functional category of interest
* pct\_pwy: 100\*pwy\_sig/(pwy\_sig+pwy\_ns)
* Excel tabs in KEGG file: KEGG contains enrichment analyses pertaining to the KEGG Orthology module, the main module within the BRITE hierarchy. Other tabs refer to other BRITE hierarchies that were assessed due to significance within the KEGG Orthology module (see manuscript).

## Metabolite Associations

Univar tab:

* Group: All, PD, or Ctrl (group being tested)
* Method: the statistical test performed (all are spearman)
* Variable: the metabolite tested
* Taxon: the bacterial or functional term being tested
* Rho: the rho value of the spearman test
* Pval, qval: P and corrected P (Q) values

Multivar:

* Group: All, PD, or Ctrl (group being tested)
* Taxon: the bacterial or functional term being tested
* Response: the metabolite tested
* Explanatory: the variable corresponding to the statistical output
  + Df$laxatives is laxative use (yes/no), df$bristol is Bristol stool rating, df$SexMale is self-reported biological sex, df$depth is sequencing depth, and (Intercept) is the linear model y intercept.
* Estimate, Std. Error, t value: the effect size, standard error and t value of each linear model
* Pval, qval: P and corrected P (Q) values

Multivar Species:

* Same as Multivar, but Explanatory is filtered to only include Microbe (no covariables).

## Network Analysis

Network Analysis Species-Level Data:

* Betweenness, Closeness: betweenness/closeness centrality measures for each taxon
* Taxon-Taxon distances:
  + V, To: Origin and destination taxa within a given connection. Note that the dataset contains forward and reverse connections (i.e. V, To switched), but not self connections.
  + Distance: the network distance between V and To (sum of node lengths). Forward and reverse connections have identical distances.
  + Connected: If a path cannot be drawn between two microbes, the distance is infinite and Connected=No.
  + Comparison: V + To in a single string.

Network Analysis Statistics:

* Excel tab “Species Included”: this tab lists all of the species included in the network analysis.
* Other tabs: statistics corresponding to each component of the network analysis.
  + Fig. 1b, No Figure (Prop. 0-edge nodes): the chi square output statistics.
    - Statistic: the chi square statistic
    - P.value: p value
    - Parameter: degrees of freedom
    - Method: statistical test (chi square)
  + Fig. 1c-e, S2: the Wilcoxon output statistics.
    - Statistic.w: the value of the test statistic
    - P.value: p value
    - Null.value.location shift: the location parameter mu
    - Alternative: two-sided test
    - Method: Statistical method applied (Wilcoxon)
    - Data.name: description of the comparison being made
    - Conf.int1 and 2: the 95% confidence interval of the estimate (below)
    - estimate.(pseudo)median: estimate of the location parameter

## Progression

0. Progression of Disease Severity and Medication:

* Avg Values Per Visit:
  + Visit: Visit number
  + Months in Study: number of months since baseline visit
  + Remaining columns: mean (standard deviation) of disease severity metrics per visit number
* Change Per Year & Tertile:
  + Progression: denotes tertiles of MDS-UPDRS 4 progression rates
  + Remaining columns: mean (standard deviation) of change in disease severity metrics per year
* Prog Rates (Table S2):
  + Variable: disease severity metric
  + Change/Month, Std. Error, Degrees Freedom, t value: the effect size, standard error, degrees of freedom, and t value of each mixed model
  + P value, Q value: significance metrics

[…]Disease Progression Stats:

* Effect: denotes whether the variable is a fixed or random effect
* Group: denotes whether the row corresponds to a formula variable or the residual error (only applicable to random variables, see ‘All’ tabs for complete stats)
* Term: the explanatory variable in question
* Estimate, std.error, statistic, df: the effect size, standard error, test statistic, and degrees of freedom of each mixed model
* P.value, q.value: p value, q value
* Variable: disease severity metric (outcome variable)
* Species: the taxonomic species/functional term being assessed in each model

Baseline Severity Stats:

* Univar:
  + Method: the statistical method used (Spearman)
  + Variable: the disease severity metric being tested
  + Taxon: the bacterial species/functional term being tested
  + Rho: the rho value of the spearman test
  + Pval, qval: P and corrected P (Q) values
* Multivar:
  + Taxon: the bacterial species/functional term being tested
  + Response: the disease severity metric being tested
  + Explantory: explanatory variable corresponding to the statistics
    - df$laxatives is laxative use (yes/no), df$bristol is Bristol stool rating, df$SexMale is self-reported biological sex, df$depth is sequencing depth, and (Intercept) is the linear model y intercept. df$Disease.duration is the years since PD onset.
  + Estimate, Std. Error, t value: the effect size, standard error, and t value of each mixed model
  + Pval, qval: P and corrected P (Q) values
* Multivar Species: same as Multivar, Explanatory is filtered to include only the microbe-related statistics

## Random Forest

* Feature: denotes the bacterial species/functional term included in the model
* Importance: denotes the importance score for each feature, where higher values indicate that the feature contributed more to the model
* Data: model of origin

## Diversity of Functional Data

Shannon Diversity Univariate:

* Statistic: Wilcoxon test statistic
* P.value, qval: p and corrected p (Q) values
* Method: statistical test (Wilcoxon test)
* Alternative: two-tailed
* Variable: the explanatory variable
* Type: functional dataset tested

Shannon Diversity Multivariate:

* Type: functional dataset tested
* Variable: the explanatory variable
* Estimate, Std. Error, t value: the effect size, standard error, and t value of each mixed model
* Pr(>|t|), qval: P value, corrected P (Q) value

Bray-Curtis Univariate & Multivariate:

* Type: functional dataset tested
* Variable: the explanatory variable
* DF, SumOfSqs, R2, F: degrees of freedom, sum of squares, R squared value, and F statistic of the PERMANOVA
* Pr(>|t|), qval: P value, corrected P (Q) value

## Validation of Statistical Tests

* Figure 1: Validation of Wilcoxon signed-rank tests. For each species, the difference in each metric was calculated between groups and plotted using histograms to ensure that the distribution of points was reasonably symmetric. (Note: normality is not required)
* Figure 3 – Variable Normality: Density plots of all variables included in the linear models. Individual taxon/pathways were not visualized due to the large number of features, and were assumed to follow zero-inflated negative binomial distributions. Sequencing depth and Bristol rating were not transformed, while the metabolites were log-transformed in order to optimize normality.
* Figure 3 – Other: Validation of linear regression models between p-cresol/phenylacetylglutamine and sequencing data. Page 1 shows scatterplots of each association, where sequencing data are CLR transformed and metabolites abundances are autoscaled. Page 2 shows the relationship between fitted model values and residuals, where points should be spread evenly on either side of the red line with no apparent pattern. Page 3 shows Q-Q plots of each association, where successfully normalized data will produce points that closely follow the red line with minimal deviation.
* Figure 5 – Progression Normality: Density plots of all variables included in the linear mixed/regression models. The top plots present raw progression marker data, while the bottom plots represent square root-transformed data.
* Figure 4 – Baseline: these plots represent the linear regression models used to test the correlation between baseline disease severity and sequencing data. The same three plots are used here as described above: scatterplots, Fitted values vs. Residuals, and Q-Q plots.
* Figure 4 – Long Multivar: Validation plots of the linear mixed models. Plots include:
  + Scatterplots of time vs. progression metric, divided into tertiles of the feature of interest. Coloured lines represent individual trajectories, and the black line (shaded region) represents the line of best fit (95% confidence interval).
  + Residual plot (fitted values vs residuals of model). Points should be spread evenly on either side of the red line with no apparent pattern.
  + Q-Q plots of residuals, random intercepts, and random slopes. Successfully normalized data will produce points that closely follow the red line with minimal deviation.
* Figure S5: The presented plots are identical to those described above (Figure 4 – Long Multivar), except that there is no microbial feature of interest and the scatterplots are therefore not split across multiple panels.