Healthy Infant Metabolome is Robust to Normal Variations in Gut Microbiome

NERIC 2019

Quang Nguyen

Filtering, Transformation and Normalization

We normalize taxonomic data to proportions by dividing individual counts by sample-specific sequencing depths. We first imputed all 0 values with a unit pseudocount and then transform the data using center log-ratio transformation implemented in the function `clr` from the package `compositions` in `R`. For untargeted metabolomic data, we transform the data into relative abundances similar to our taxonomic profiles, and then subsequently use the arcsine square root transform to ensure approximate homoscedasticity when applying linear models. For targeted metabolomic data sets, since the data is already in exact concentration format, we simply perform a log transformation, accounting for 0 values by imputing with the smallest non-zero concentration times 0.5. Models are fitted to the transformed data and the resulting predictions are back-transformed to preserve the coverage of predicted metabolite compositions.

Model Fitting and evaluation

We fit our machine learning models using the `caret` package in `R`. Scaling and centering was performed on the predictor matrix prior to model fitting. We evaluate our models using 10-fold nested cross validation, where within each outer training set we nest a cross-validation procedure for parameter tuning which should prevent overfitting.

Models we compared include elastic net (EN), random forest (RF), support vector machines with radial basis function kernel (SVM), sparse partial least squares (SPLS) as well as multivariate extensions of EN, SPLS and RF. These models were chosen based on prelminiary fit, as well as from existing literature.

We evaluate our models using four different criterion: Spearman's rho, Root mean squared error (RMSE), Root relative squared error (RRSE) and Predictive R-squared (R2). In order to contextualize our values, we simulated null distributions of our test statistic by permuting both the predictor and outcome matrices using the function `randomizeMatrix` in the package `picante`. The procedure performs permutations while perserving the general structure of the dataset.

Additional graphics

Taxa and metabolite relative abundances

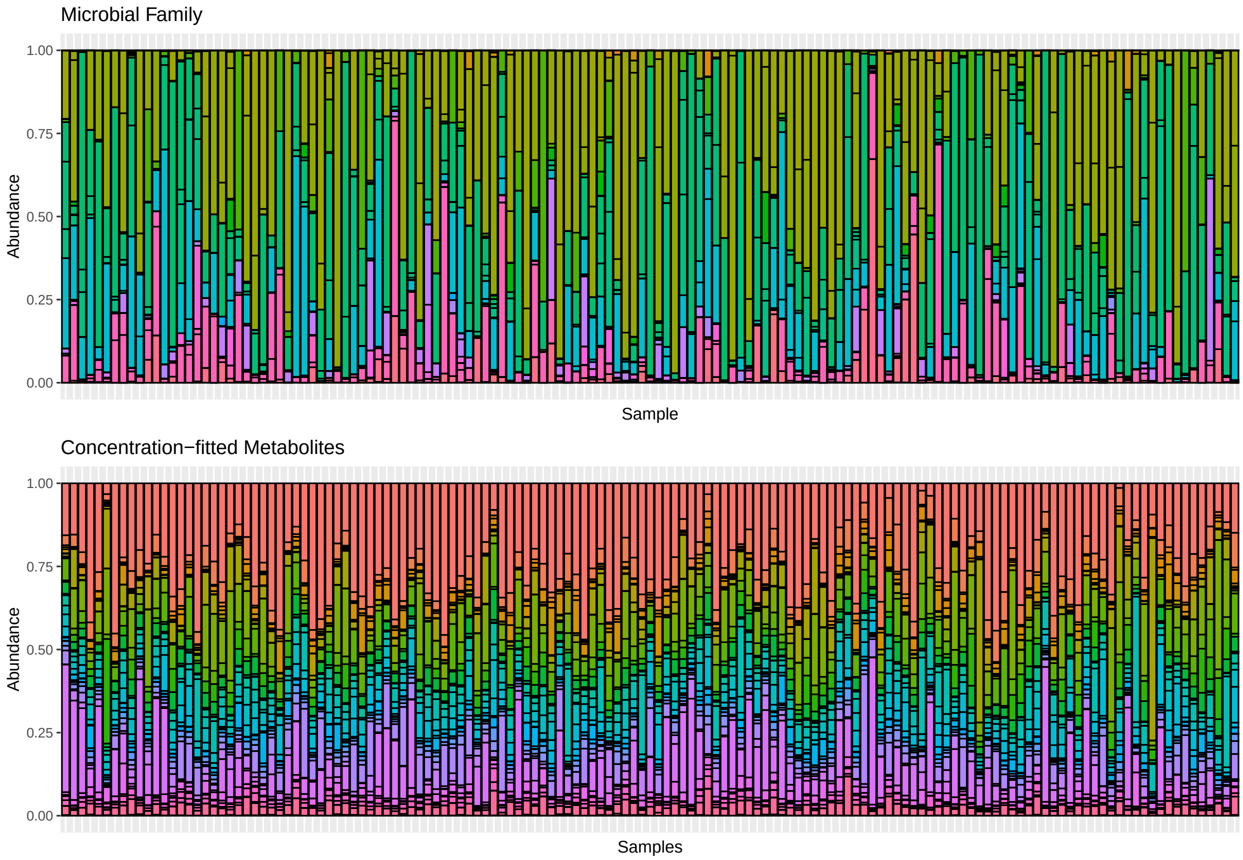

Figure 1: **Figure 1.** Relative abundances of bacterial families (top panel) and targeted metabolites (bottom panel) for 6W samples

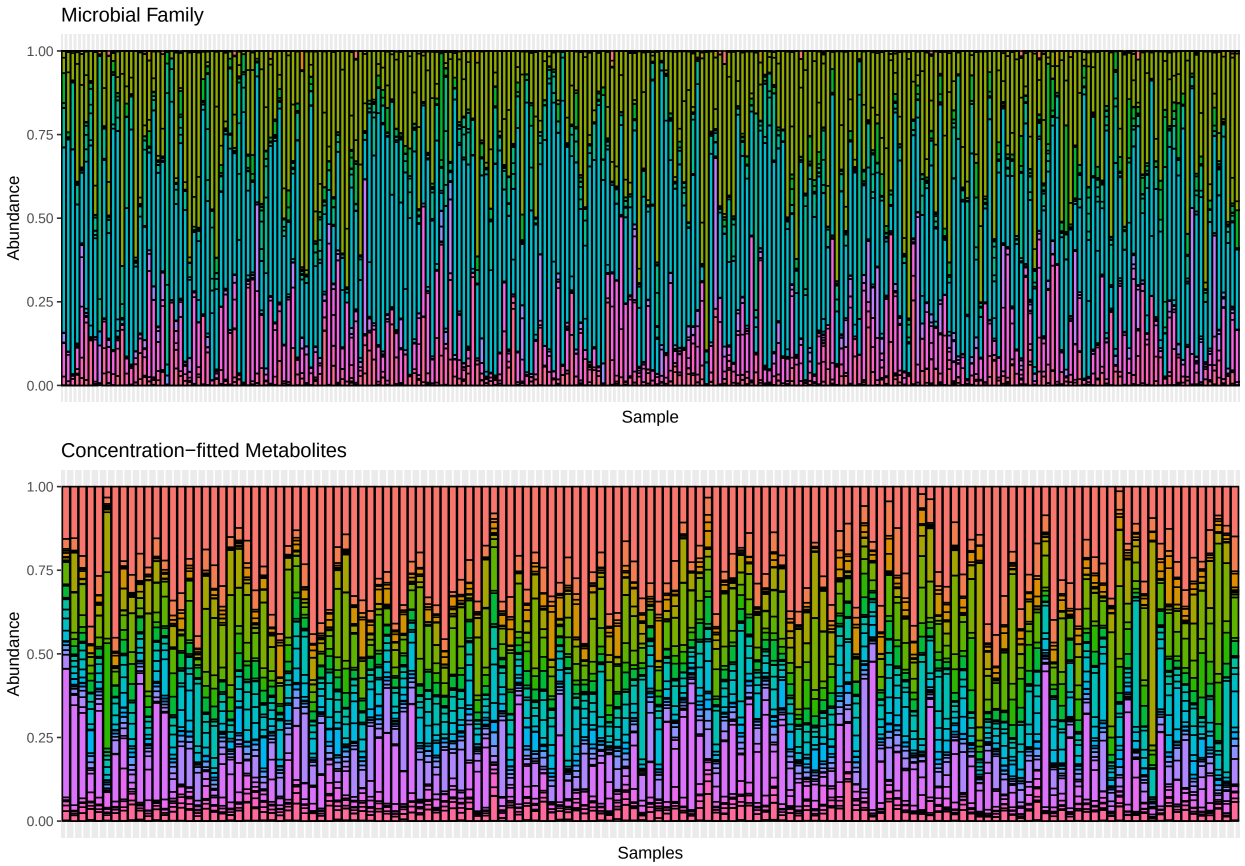

Figure 2: **Figure 2.** Relative abundances of bacterial families (top panel) and targeted metabolites (bottom panel) for 12M samples