Supplementary Material for:

Gut Microbiota in Adolescents and the Association with Fatty Liver: the EPOCH Study

Supplemental Figures

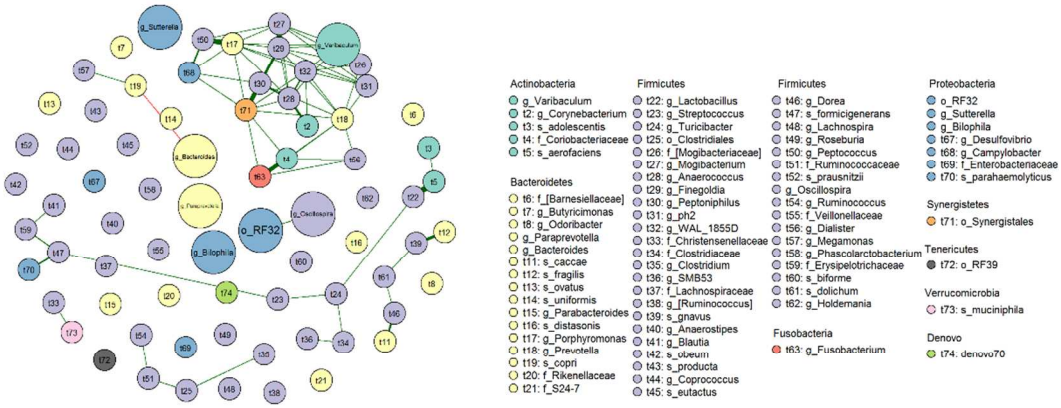
Legends

Supplemental Figure S1. This network depicts all the gut microbiota taxa present in at least 10 adolescents in the EPOCH cohort and their interrelationships. Connected taxa are correlated; positive correlations are shown with green lines and negative correlations with red lines. The width of the connecting lines represents the strength of association. The taxa that were selected by random forests as most relevant to the outcome of hepatic fat fraction are shown with larger circles. Colors correspond to phylum-level taxa, as shown in the legend.

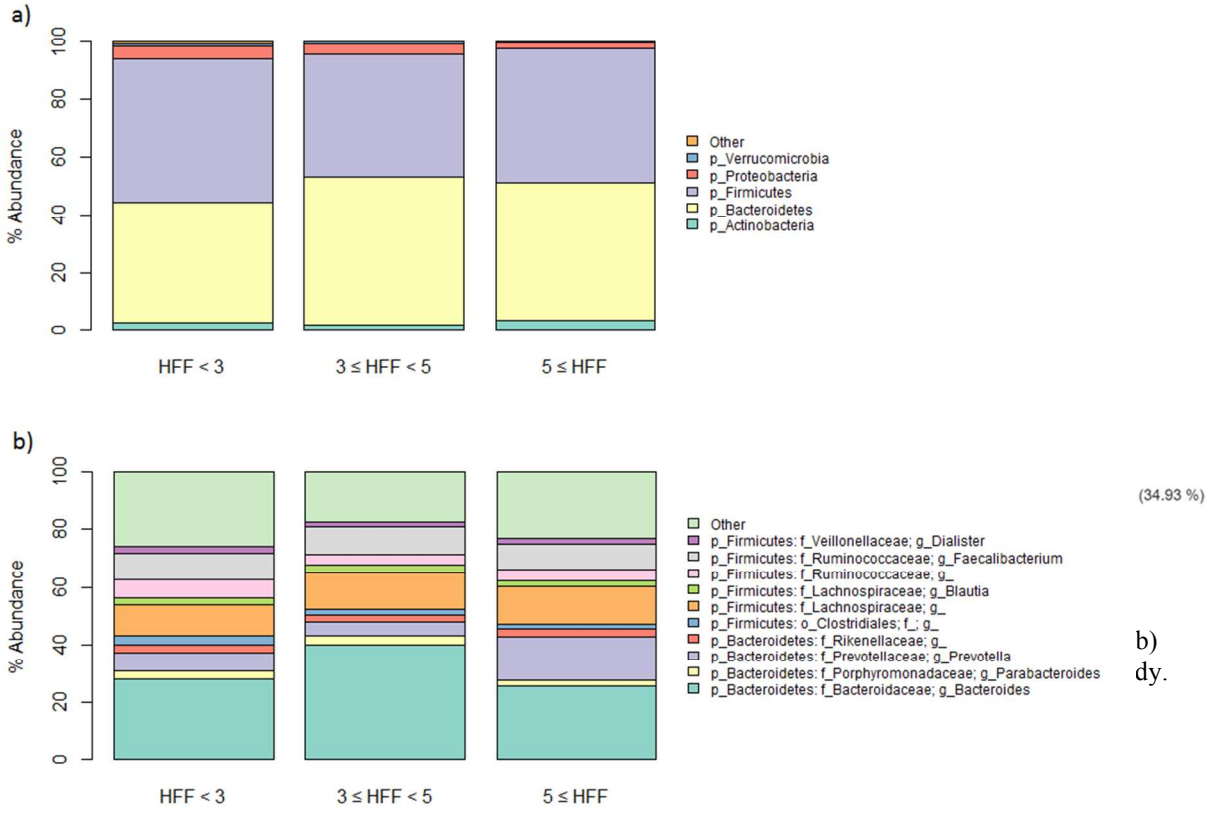
Supplemental Figure S2. Percent abundance of the most abundant a) phylum-level and b) genus-level taxa of adolescents in the EPOCH study by HFF group: low (<3), moderate (3-5) and high / NAFLD (≥5).

Supplemental Figure S3. These partial plots show the adjusted relationship between a) taxa, b) dietary components, c) demographics and comorbidities, and d) the combined set of features selected by the random forests as important to predict HFF against the predicted values of HFF ($\hat{y}$). Since random forests do not provide regression coefficients, these plots are used to understand the direction of the relationships between each predictor and the outcome, while the other selected predictors are held constant, as in multiple regression. A linear regression would fit only a linear relationship between the predictors and the outcome, but random forests allow for any type of relationship, including complex interactions. The general trends shown in these pictures are summarized in Table 2. The following abbreviations are used: pmufa_ii is % monounsaturated fats, ptotfat_ii is % total fat, and pcarbo_ii is % carbohydrates; bmiz is BMI z-score, delmeth is delivery method, CS is cesarean delivery and V is vaginal delivery.

Supplemental Figure S4. These heat plots show the 2-way inter-relationships between a) taxa, b) dietary components, c) demographics and comorbidities, and d) the combined set of features selected by the random forests as important to predict HFF, with the predicted values of HFF represented in color: red indicates high values off HFF, yellow corresponds with low values. Since random forests do not provide regression coefficients and may include complex interactions between the features, these plots are used to understand these interactions. For example, the plot of *Oscillospira* and *Bilophila* suggests an interaction: the previous figure shows that *Bilophila* has a positive relationship with HFF (higher abundance generally corresponds to higher HFF), but this plot shows that high abundance of *Bilophila* corresponds with high HFF only when *Oscillospira* abundance is ≤20%. The presence of interactions suggested by these plots are noted in Error! Reference source not found.. The following abbreviations are used: pmufa_ii is % monounsaturated fats, ptotfat_ii is % total fat, and pcarbo_ii is % carbohydrates; bmiz is BMI z-score, delmeth is delivery method, CS is cesarean delivery and V is vaginal delivery.

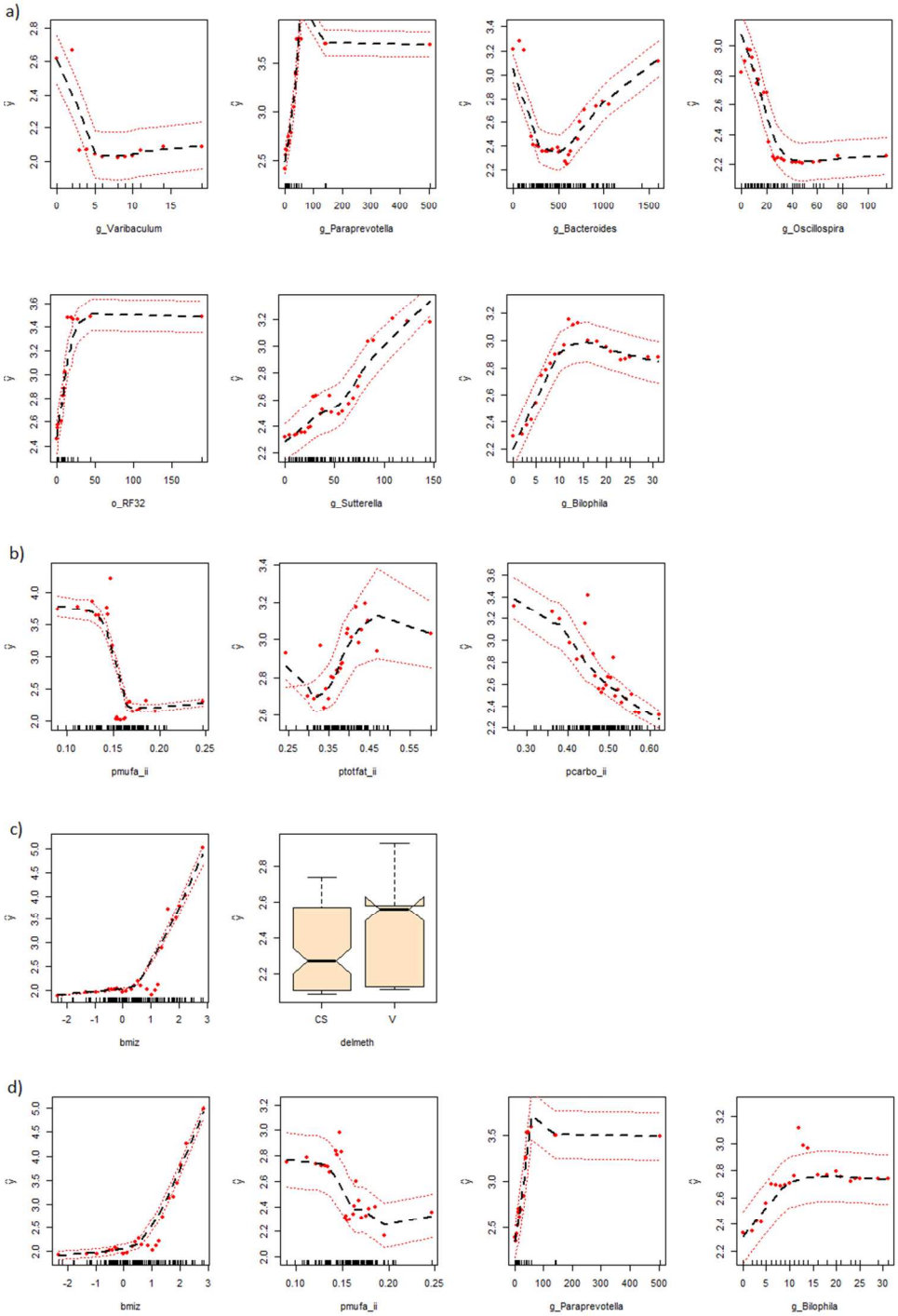


Supplemental Figure S1. This network depicts all the gut microbiota taxa present in at least 10 adolescents in the EPOCH cohort and their interrelationships. Connected taxa are correlated; positive correlations are shown with green lines and negative correlations with red lines. The width of the connecting lines represents the strength of association. The taxa that were selected by random forests as most relevant to the outcome of hepatic fat fraction are shown with larger circles. Colors correspond to phylum-level taxa, as shown in the legend.



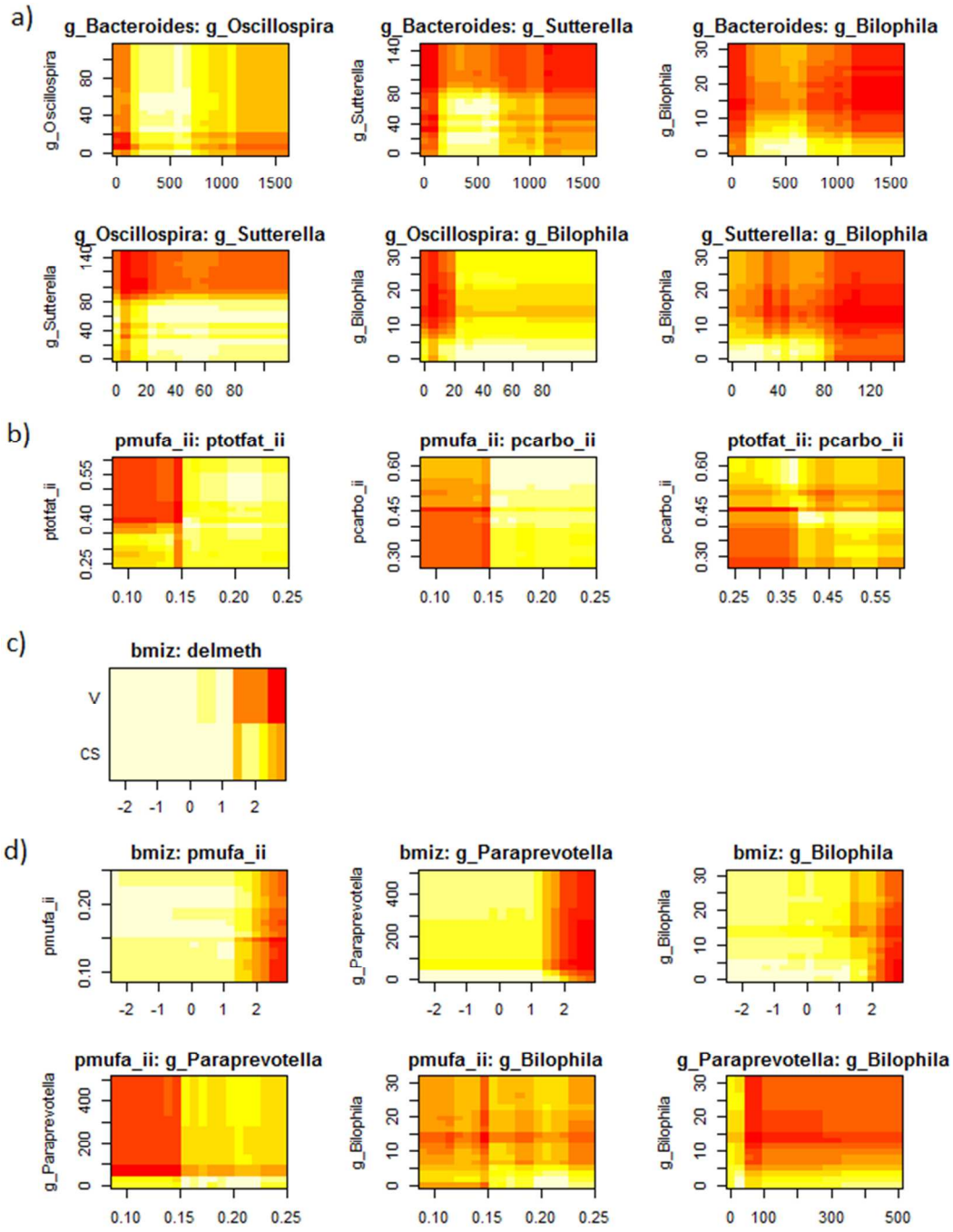
b)
dy.

Supplemental Figure S2. Percent abundance of the most abundant a) phylum-level and b) genus-level taxa of adolescents in the EPOCH study by HFF group: low (<3), moderate (3-5) and high / NAFLD (≥5).



Supplemental Figure S3. These partial plots show the adjusted relationship between a) taxa, b) dietary components, c) demographics and comorbidities, and d) the combined set of features selected by the random forests as important to predict HFF against the predicted values of HFF ($\hat{y}$). Since random forests do not provide regression coefficients, these plots are used to understand the direction of the relationships between each predictor and the outcome, while the other selected predictors are held constant, as in multiple regression. A linear regression would fit only a linear relationship between the predictors and the outcome, but random forests allow for any type of relationship, including complex interactions. The general trends shown in these pictures are summarized in Error! Reference source not found.. The following abbreviations are used: pmufa_ii is % monounsaturated fats, ptotfat_ii is % total fat, and

pcarbo_ii is % carbohydrates; bmiz is BMI z-score, delmeth is delivery method, CS is cesarean delivery and V is vaginal delivery.



Supplemental Figure S4. These heat plots show the 2-way inter-relationships between a) taxa, b) dietary components, c) demographics and comorbidities, and d) the combined set of features selected by the random forests as important to predict HFF, with the predicted values of HFF represented in color: red indicates high values off HFF, yellow corresponds with low values. Since random forests do not provide regression coefficients and may include complex interactions between the features, these plots are used to understand these interactions. For example, the plot of *Oscillospira* and *Bilophila* suggests an interaction: the previous figure shows that *Bilophila* has a positive relationship with HFF (higher abundance generally corresponds to higher HFF), but this plot shows that high abundance of *Bilophila* corresponds with high HFF only when *Oscillospira* abundance is $\leq$20%. The presence

of interactions suggested by these plots are noted in Error! Reference source not found.. The following abbreviations are used: pmufa_ii is % monounsaturated fats, ptotfat_ii is % total fat, and pcarbo_ii is % carbohydrates; bmiz is BMI z-score, delmeth is delivery method, CS is cesarean delivery and V is vaginal delivery.